Creating Value with Data Analytics in Marketing

Mastering Data Science

SECOND EDITION

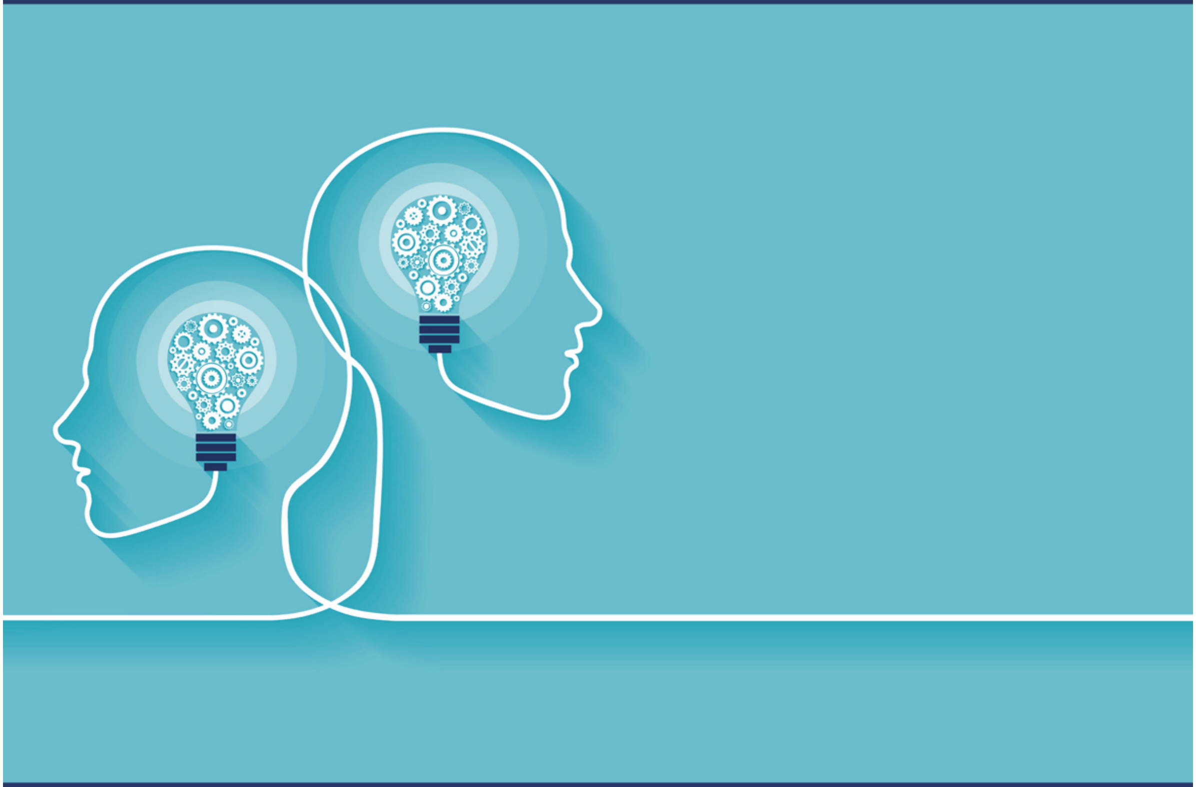
Peter C. Verhoef, Edwin Kooge, Natasha Walk and Jaap E. Wieringa

ROUTLEDGE

Creating Value with Data Analytics in Marketing

Mastering Data Science

SECOND EDITION



Peter C. Verhoef, Edwin Kooge, Natasha Walk and Jaap E. Wieringa

Creating Value with Data Analytics in Marketing

This book is a refreshingly practical yet theoretically sound roadmap to leveraging data analytics and data science. The vast amount of data generated about us and our world is useless without plans and strategies that are designed to cope with its size and complexity, and which enable organizations to leverage the information to create value in marketing.

Creating Value with Data Analytics in Marketing provides a nuanced view of big data developments and data science, arguing that big data is not a revolution but an evolution of the increasing availability of data that has been observed in recent times. Building on the authors' extensive academic and practical knowledge, this book aims to provide managers and analysts with strategic directions and practical analytical solutions on how to create value from existing and new big data. The second edition of this bestselling text has been fully updated in line with developments in the field and includes a selection of new, international cases and examples, exercises, techniques and methodologies.

Tying data and analytics to specific goals and processes for implementation makes this essential reading for advanced undergraduate and postgraduate students and specialists of data analytics, marketing research, marketing management and customer relationship management.

Online resources include chapter-by-chapter lecture slides and data sets and corresponding R code for selected chapters.

Peter C. Verhoef is the Dean of the Faculty of Economics and Business and Professor of Marketing at the University of Groningen, the Netherlands.

Edwin Kooge is a co-founder of BAYZ, a consulting agency focusing on teaching companies to master analytics. He is a data science expert, results-focused business advisor and entrepreneur with more than 25 years' experience in analytics.

Natasha Walk is a co-founder of BAYZ, a consulting agency focusing on teaching companies to master analytics. She is a pragmatic data science expert focusing on talent development in data science with more than 25 years' experience in analytics.

Jaap E. Wieringa is Professor of Research Methods in Business at the Department of Marketing at the University of Groningen, the Netherlands, and is Research Director of the Customer Insights Center.

Creating Value with Data Analytics in Marketing

Mastering Data Science

SECOND EDITION

Peter C. Verhoef, Edwin Kooge, Natasha Walk and Jaap E. Wieringa

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

Second edition published 2022
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

and by Routledge
605 Third Avenue, New York, NY 10158

Routledge is an imprint of the Taylor & Francis Group, an informa business

First edition published by Routledge 2016

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloguing-in-Publication Data
Names: Verhoef, Peter C., author. | Kooge, Edwin, author. | Walk, Natasha, author. | Wieringa, Jaap E., author.
Title: Creating value with data analytics in marketing: mastering data science / Peter C. Verhoef, Edwin Kooge, Natasha Walk, Jaap E. Wieringa.
Description: Second edition. | New York, NY: Routledge, 2022. |
Previous edition: Creating value with big data analytics. | Includes bibliographical references and index.
Subjects: LCSH: Consumer profiling. | Big data. | Marketing—Data processing.
Classification: LCC HF5415.32 .V475 2022 (print) | LCC HF5415.32 (ebook) | DDC 658.8/3—dc23
LC record available at https://lccn.loc.gov/2021020151
LC ebook record available at https://lccn.loc.gov/2021020152

ISBN: 978-0-367-81978-1 (hbk)
ISBN: 978-0-367-81979-8 (pbk)
ISBN: 978-1-003-01116-3 (ebk)

DOI: 10.4324/9781003011163

Typeset in Berling
by codeMantra

Access the Support Material: www.routledge.com/9780367819798

To: Petra, Anne Mieke, Maurice and Nathalie

Contents

Figures

Tables

Preface

When we started our careers in marketing analytics, it was a small discipline which attracted only minor attention from the boards of companies. Analytics was mainly developed in firms having a strong direct marketing focus, such as Readers Digest. Beyond that, research agencies were trying to develop analytical solutions for more brand-oriented companies. During our careers this situation has dramatically changed. Analytics have become a major discipline in many firms and scientific evidence strongly supports the performance impact of a strong analytics department. Successful examples in leading firms provide only more support for having a strong analytical function. Marketing has become more data-driven in the past decade!

This development has only become more prominent with the arrival of “big data” in the last decade. Big data has induced the development of data science within firms. CEOs of banks, retailers, telecom providers, etc. now consider data science a very important discipline. As a consequence, current and future employees have to develop data science skills and learn how data science can be applied in marketing. This has become extremely important with the growth of machine learning and artificial intelligence.

In 2016 Peter Verhoef, Edwin Kooge and Natasha Walk published the first edition of our book *Creating Value with Big Data Analytics*, focusing on big data. We received many positive reactions to the first edition of that book and the publisher urged us to write a second edition. However, since then data science has become more prominent. In particular, analytics has changed and machine learning techniques have become common and accepted. We have therefore changed the title to: *Creating Value with Data Analytics in Marketing: Mastering Data Science*. We have also asked Jaap Wieringa to join our team and we were very happy that he was willing to do so. We believe this new edition will help our readers to master data science. We have therefore added assignments per chapter to help them apply the theory discussed. In particular, the data chapters and the analysis chapters have been updated, incorporating the new developments in data and data analytics including machine learning. It is our real aim that readers should carry out and practice the discussed analytics. We have therefore provided data and the program codes in the open software package R. The new edition is also visually improved, which makes it more attractive.

Writing this new edition was a great way to share our knowledge on data science and specifically data analytics, which we have further developed in recent years. Focusing on new developments, such as machine learning, also induced us to go beyond our existing knowledge and to explain sophisticated new analytical techniques in an accessible way. This new direction was very valuable for all of us. It was a pleasure to work together as a team despite the challenges faced during the COVID-19 pandemic, which led us to work mainly digitally together.

This book could not have been written without the great contribution of Alieke Engel, who helped us when writing the book with a lot of practical support. We also want to thank our students in both (under)graduate and executive programs, who during the years have forced us to explain data science and analytics in an accessible way. Finally, we want to thank our families who supported us in the writing process, which was once again a time consuming and an intensive effort.

Peter Verhoef, Edwin Kooge, Natasha Walk, Jaap Wieringa

CHAPTER 1

Data science and big data

DOI: 10.4324/9781003011163-1

1.1 INTRODUCTION

One of the most important developments in the last decade is the increasing prevalence of data. This is frequently referred to as big data. One of the main underlying drivers of this explosion is the increasing digitalization of our society, business, and marketing. One can hardly imagine that consumers around the globe nowadays could live without smartphones, tablets, Facebook, Instagram, and Twitter. Marketing is probably one of the business disciplines most affected by new developments in technology. In recent decades, technological developments such as increasing data-storage capacity, increasing analytical capacity, the growth of online etc., have dramatically changed specific aspects of marketing. More specifically, we have seen the development of Customer Relationship Management (Kumar & Reinartz, 2005). This arrival of CRM posed challenges for marketing and raised issues on how to analyze and use all the available customer data to create loyal and valuable customers (Verhoef & Lemon, 2013). With the omnipresence of even more data and other types of data, such as text and unstructured data, firms consider this an even more important problem (Leeflang et al., 2014). The explosion of data has led to the current strong focus on data science and analytics in today's business. Machine learning, algorithms and artificial intelligence have become important buzzwords. Investigations by the European Parliament show that the market value of (big) data analytics is around 116 billion Euros in the US and 54 million in the EU (see Figure 1.1) highlighting the great importance of big data for today's economy. In addition around six million people are employed in this industry in the EU.

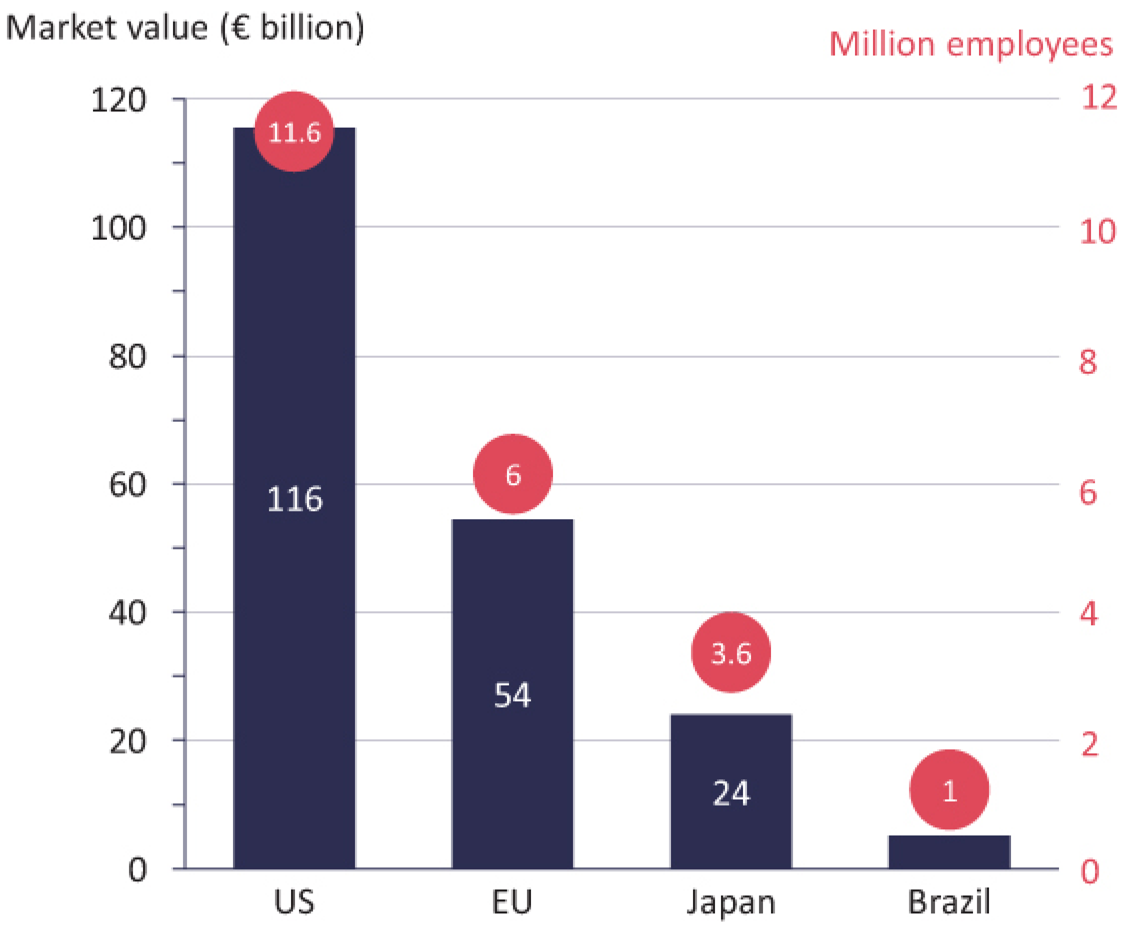


FIGURE 1.1 Big data employment and market value in the EU and other major economies

Source: https://epthinktank.eu/2016/09/29/economic-impact-of-big-data/big_data_employment

1.2 EXPLOSION OF DATA

Data have been around for decades. However, 30 to 40 years ago, these data were usually available on an aggregate level yearly or monthly. With developments such as scanning technologies, weekly data became the norm. In the 1990s, firms started to invest in large customer databases which produced records for millions of customers in which information on purchase behavior, marketing contacts, and other customer characteristics was stored (Rigby, Reichheld & Schefter, 2002). The arrivals of the Internet and, more recently, social media have led to a further explosion of data, and daily or even real-time data have become available for multiple firms. It is clear that getting value from these data is very important.

The Internet has become one of the most important marketplaces for transactions of goods and services. For example, online consumer spending across the globe has now reached around 3.53 Trillion USD (Statista, 2020).[1] Besides B2C and B2B-markets, online C2C markets have grown in importance, such as LuLu, eBay and YouTube. Amazon is now a dominant online platform and retailer, as is Alibaba in China, with very strong growth in market capitalization. Twitter users send half a million tweets every minute.[2] Companies are also increasingly investing in social media. In the USA firms have spent around 34 Billion USD on social media advertising, and this will continue to grow in the coming years (Hootsuite, 2020).[3] Managers invest in social media to create brand fans, which tends to have positive effects on word-of-mouth recommendation and brand loyalty (De Vries, Gensler & Leeflang, 2012; Uptal & Durham, 2010). There are 3.5 billion searches on Google every day, growing by around 10% every year. The use of social media also creates a tremendous increase in customer insights, including how consumers interact with each other and the products and services they consume. Specifically, blogs, product reviews, discussion groups, product ratings, etc. are new and important sources of information (Mayzlin & Yoganarasimhan, 2012; Onishi & Manchanda, 2012). The increasing use of online media, including mobile technology, also allows firms to follow customers in their customer journeys (Lemon & Verhoef, 2016).

1.3 DATA SCIENCE BECOMES THE NORM

In 2020 it is estimated that there are around 40 trillion gigabytes of data. Big data has become the norm and firms understand that they might be able to compete more effectively in a digital environment by analyzing these data (e.g., Davenport & Harris, 2007; Verhoef et al., 2021). There are several popular examples of firms analyzing these data, such as IBM, Tesco, Capital One, Amazon, Google, Netflix, Zalando, etc. Data science is often used as an overall term that focuses on activities within a firm to extract value from data using analytics. In future years this can only become more important, given strong developments in artificial intelligence (e.g. Huang & Rust, 2018), in which data, analytics, and algorithms are very important.

However, there are also problems with data science. Many firms still struggle to implement an effective data science strategy. Moreover, the increasing presence of data and data science has stirred up heated discussion and public concern on privacy issues. These discussions and concerns have become even more prevalent as a consequence of Edward Snowden, who leaked documents that uncovered the existence of numerous global surveillance programs, many of them run by the NSA and Five Eyes with the cooperation of telecommunication companies and European governments.[4] Firms continue to underestimate the privacy concerns of customers and societal organizations. For example, when Dutch based ING bank announced that they will use payment information to provide customers with personalized offers and advice, there were strong reactions on (social) media.

The problems with creating value from data science mainly arise due to a lack of knowledge and skills on how to analyze and use the (big) data. In addition, firms might overestimate the benefits of big data and data science (Meer, 2013). One important danger is that firms start too big and start thinking too ambitiously, while actually lacking good quality knowledge of the basics and challenges of good data analysis in regard to already existing data, such as CRM-data and survey data, and how this can contribute to business performance. Firms start up large-scale big data projects with complex data mining and computer science techniques and software programs, without a proper definition of the objectives of these projects and the underlying statistical techniques. As a consequence, firms invest heavily in data science but may well face a negative return on these investments.

1.4 OBJECTIVES

So, given the growing importance of data science, its economic value and the problems firms face in capitalizing on these opportunities, we believe there is an urgent need to provide greater knowledge and to develop data science skills. By writing this book we aim to provide readers with this guidance. We specifically have the following objectives.

The main objectives of the book are threefold. The first is to learn how data science can be used to create value. For that reason, we discuss the increasing presence of data and relevance of data science, and also important value concepts. We consider privacy and data security issues, which are essential for effective use of data science in a responsible manner. As a second objective, we aim to show how specific analytical approaches are required, how value can be extracted from these data and to develop new growth opportunities among new and existing customers. Our third objective is to discuss organizational solutions for development and organization of the marketing analytical function within firms that will create value from data science.

1.5 OUR APPROACH

We believe in the potential power of data science. With this book we aim to teach readers how to use data science and analytics to create value. Building on extensive academic and practical knowledge of multiple issues surrounding analytics, we have written a book that aims to provide managers and analysts with strategic directions, practical data- and analytical solutions on how to create value from existing and new big data. This book has two specific target groups. First, it is targeted at students who want to gain knowledge on data science and develop their skills in courses on data science. We include examples in the different chapters and also assignments at the end of each chapter. The chapters on data analytics provide examples based on available data sets and codes in R. This allows readers to actively carry out the analytical techniques discussed. The additional data are available at www.masteringdatascience.eu. Second, this book targets managers and data science users who are interested in how data science can be used in marketing to create value and in the analytical and other skills that are required to implement data science initiatives.

1.6 OVERVIEW OF CHAPTERS

In Chapter 2 we discuss our main data science value creation model that will be used as a guidance for the following chapters. Next, we have a chapter on the value objectives and metrics that are important in measuring value creation (Chapter 3). Subsequently, we have three chapters on data. Specifically, we focus on data as an asset (Chapter 4), data storage and integration (Chapter 5) and privacy and data-security (Chapter 6). Subsequent chapters focus on analytic strategies and (new) analytic techniques, beginning with a general chapter on data analytics (Chapter 7). Next, we continue with a chapter on data exploration techniques (Chapter 8) and a chapter on data modelling techniques (Chapter 9). Having analyzed the data, the results have to be communicated and visualized, which we discuss in Chapter 10. We end with two chapters on the implementation of data science, specifically focusing on how analysis can be used to create value in Chapter 11, and the capabilities and skills a firm needs to develop in Chapter 12.

NOTES

1. Weblink: https://www.statista.com/topics/871/online-shopping/
2. Weblink: https://techjury.net/blog/big-data-statistics/#gref
3. Weblink: https://blog.hootsuite.com/social-media-advertising-stats/
4. See: https://en.wikipedia.org/wiki/Global_surveillance_disclosures_(2013%E2%80%93present)

REFERENCES

Davenport, T. & Harris, J. (2007). *Competing on Analytics – The New Science of Winning*. Boston, MA: Harvard Business School Press.

De Vries, L., Gensler, S., & Leeflang, P. S. H. (2012). Popularity of brand posts on brand fan pages: An investigation of the effects of social media marketing. *Journal of Interactive Marketing*, *26*(2), 83–91.

Huang, M. H. & Rust, R. T. (2018). Artificial intelligence in service. *Journal of Service Research*, *21*(2), 155–172.

Kumar, V. & Reinartz, W. (2005). *Customer Relationship Management: A Databased Approach*. Chichester: John Wiley and Sons.

Leeflang, P. S. H., Verhoef, P. C., Dahlström, P., & Freundt, T. (2014). Challenges and solutions for marketing in a digital era. *European Management Journal*, *32*(1), 1–12.

Lemon, K. N. & Verhoef, P. C. (2016). Understanding customer experience throughout the customer journey. *Journal of Marketing*, *80*(6), 69–96.

Mayzlin, D. & Yoganarasimhan, H. (2012). Link to success: How blogs build an audience by promoting rivals. *Management Science*, *58*(9), 1651–1668.

Meer, D. (2013). The ABCs of analytics. *Strategy Business*, *70*, 6–8.

Onishi, H. & Manchanda, P. (2012). Marketing activity, blogging and sales. *International Journal of Research in Marketing*, *29*(3), 221–234.

Rigby, D. K., Reichheld, F. F., & Schefter, P. (2002). Avoid the four perils of CRM. *Harvard Business Review*, *82*(11), 101−109.

Uptal, M. D. & Durham, E. (2010). One cafe chain's Facebook experiment. *Harvard Business Review*, *88*(3), 26–26.

Verhoef, P. C., Broekhuizen, T., Bart, Y., Bhattacharya, A., Dong, J. Q., Fabian, N., & Haenlein, M. (2021). Digital transformation: A multidisciplinary reflection and research agenda. *Journal of Business Research*, *122*, 889–901.

Verhoef, P. C., & Lemon, K. N. (2013). Successful customer value management: Key lessons and emerging trends. *European Management Journal*, *31*(1), 1–16.

CHAPTER 2

Creating value with data science

DOI: 10.4324/9781003011163-2

2.1 INTRODUCTION

In past years, firms have heavily invested in big data and data science. Data scientist is now a normal role within a company. This phenomenon is present in all sectors of the economy including telecom, (online) retailing, and financial services. Firms have a strong belief that the science of analyzing data can lead to a competitive advantage and can create new business opportunities.

In this chapter, we lay out the foundations for a sound value-creating data science strategy. We specifically discuss how data science can create value and what elements are required to create value.

2.2 DATA SCIENCE VALUE CREATION MODEL

One of the biggest challenges of data science is how firms can create value with data and data science. We have developed the data value creation model to show how this value creation occurs (see Figure 2.1). This model has five elements:

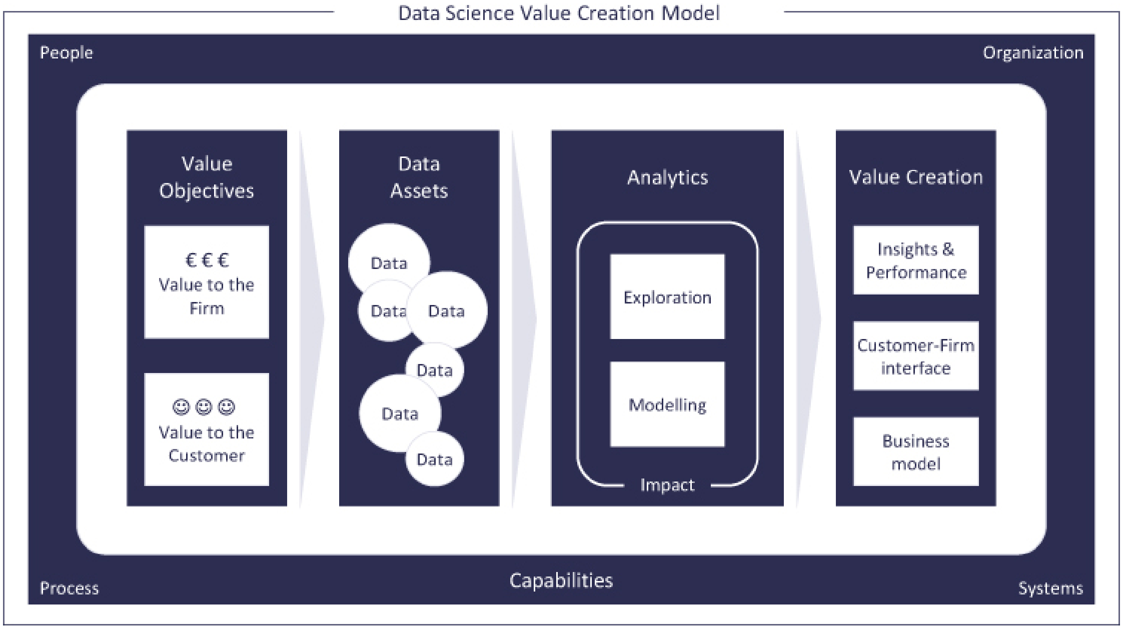


FIGURE 2.1 Data science value creation model

1. Value objectives
2. Data assets
3. Analytics
4. Value creation
5. Capabilities

The model starts with value objectives that have to be set before developing a data science strategy. The core data science strategy elements are data assets and analytics, which then should lead to value creation. The data science strategy should be enabled by data science capabilities.

2.3 VALUE CREATION OBJECTIVES

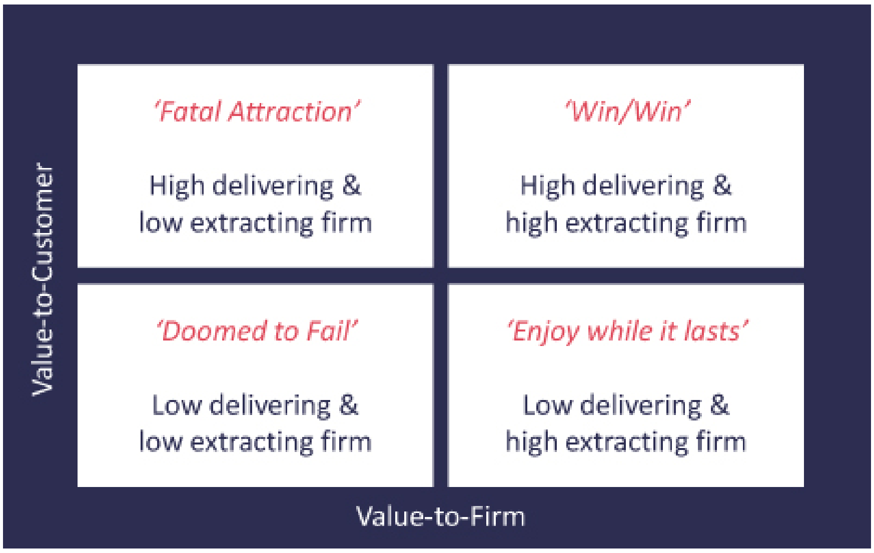
Value creation should be the ultimate objective of every data science strategy. However, value creation is one of those terms that is easily written down without a complete understanding of the topic. Importantly, we consider value from two perspectives:

1. Value to the customer (V2C)
2. Value to the firm (V2F)

These two perspectives are not novel. In fact, the classical definitions of marketing put forward in basic marketing textbooks (e.g., Kotler & Armstrong, 2014) emphasize that marketing should focus on creating superior value for customers (through high quality, attractive brand propositions and striving for appropriate relationship), and that firms can capture value from customers in return for this value creation. This is sometimes also referred to as “value delivery” and “value extraction.” Value extraction from customers is considered to be a direct consequence of value delivery. Value extraction occurs in paid price premiums, higher loyalty rates (lower churn), higher revenues per customer, and stronger customer advocacy (Reichheld, 1996; Srivastava, Tasadduq & Fahey, 1998).

2.3.1 Balance between V2F and V2C

Firms can be classified on two value dimensions (see Figure 2.2). A high value delivery and high value extraction strategy is considered as a win-win strategy. It is usually seen as the best strategy for firms. Despite this, we frequently observe that firms tend to outperform on a single value dimension (upper-left and lower-right cells). This can have dramatic consequences. Frequently, firms tend to focus solely on value extraction: examples of such firms can be found in multiple sectors. A historically dramatic example is the banking industry. There has been a strong focus on shareholders’ value within banks, inducing them to focus less on customers and the delivery of value to customers. The crisis in 2008, with many banks facing difficult problems, showed that this sole focus on value extraction can have severe consequences for firms and for society (Verhoef, 2012). Firms in other industries sometimes solely focus on V2F. For example, if telecom (telco) firms focused on creating shareholder value, marketing management had a strong focus on customer lifetime value creation through communication tactics and contractual offers (i.e. moving from one-year to two-year contracts, minute rounding, or other short-term pricing tactics). Identifying potential churn candidates was an important element of that strategy. In terms of value delivery to shareholders, this strategy was rewarding (Hansen, Ibara & Peyer, 2013). However, the presumed lack of investment in service quality and innovation was considered a weakness in the strategies of many telco firms. Nowadays, telco firms like Vodafone strongly focus on a good customer experience.



Value-to-Customer

'Fatal Attraction'
High delivering & low extracting firm

'Win/Win'
High delivering & high extracting firm

'Doomed to Fail'
Low delivering & low extracting firm

'Enjoy while it lasts'
Low delivering & high extracting firm

Value-to-Firm

FIGURE 2.2 Value-to-Customer vs. Value-to-Firm

Source: Adapted from: Reinartz, 2011; Wiesel *et al*., 2011

A mismatch between delivered customer value and extracted firm value can also occur (upper-left cell). These firms are rather attractive for customers, but they fail to extract more value through, for example, achieved higher loyalty and higher price premiums. Many start-up online firms struggle here; they provide much value in terms of free services, and lower prices etc., but find it difficult to keep customers and/or ask for fees for the services they provide. Empirical research, however, shows that while firms are commonly in the downward "Enjoy it while it lasts cell," the number of firms in the upper-left "Fatal Attraction" cell is rather limited (Bouma *et al*., 2010). We do, however, find a number of examples of firms in the "Doomed-to-Fail" cell, where they provide low customer value and are unable to extract sufficient value. Firms in this cell are in a dangerous position as their value proposition and delivery require strong investments while, due to their inability to extract value, they may lack long-term resources for those investments. Consider retailer Toys'R US with its decreasing attractiveness to customers and decreasing sales leading to bankruptcy because of strong online competition from, for example, Amazon.

2.3.2 V2S: Extending value creation

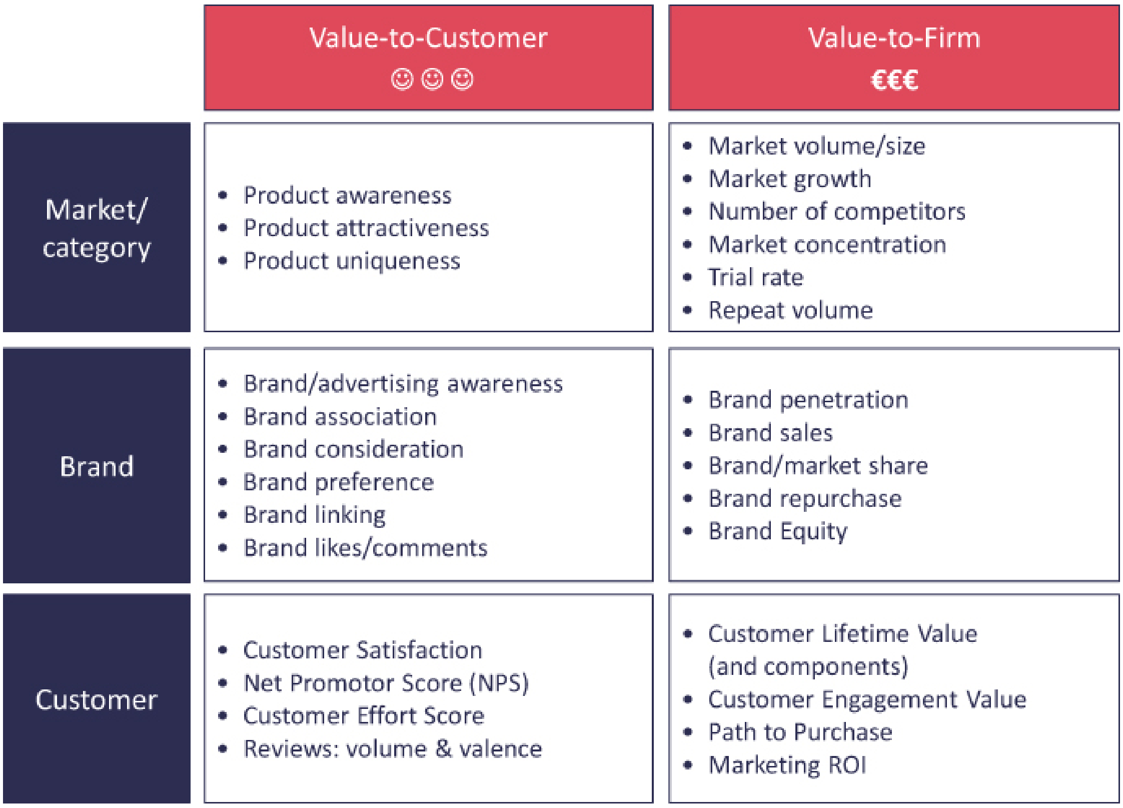
The above value concepts are sometimes extended. Especially in an era where firms are considered to also fulfill a societal role and there is an increased focus on, for example, corporate social responsibility, a focus solely on V2C and V2F is not sufficient (Korschun, Bhattacharya & Swain, 2014; Porter & Kramer, 2011). In fact, banks have not only been criticized for insufficient focus on customers, but also for insufficient consideration of society as a whole (Verhoef, 2012). One could therefore suggest extending the value concept by also taking into account value delivery to society (V2S). This could be done in many ways. Some firms, such as Unilever, consider sustainability as one of the core elements in their corporate strategy and aim to show that in their business operation, including brand propositions. However, Procter & Gamble uses a more tactical approach, with specific activities at the brand level, such as dental education programs in Hispanic neighborhoods in the US, to show their involvement with local communities.

V2S can partially be reflected in delivered value to customers. Rust, Zeithaml and Lemon (2000), for example, consider brand ethics as an integral part of the delivered value of brands. Similarly, corporate social responsibility is considered as a driver of customer satisfaction (Korschun, Bhattacharya & Swain, 2014).

2.3.3 Metrics for V2F and V2C

Metrics have become very important due to increasing attention given to accountability within firms and the resulting consequences for marketing departments (Verhoef & Leeflang, 2009). Metrics are measuring systems that quantify trends, dynamics, or characteristics (Farris *et al*., 2010). There are numerous metrics in marketing that can be tracked. Farris *et al*. (2010) discuss more than 50 metrics that every executive should master. We classify metrics into V2C and V2F metrics. V2F metrics are usually much more transaction-oriented and focus on concrete market outcomes that can be related to monetary consequences for the firm. V2C metrics typically focus on the evaluation of value by customers.

Beyond the distinction between V2C and V2F metrics, we also distinguish between metrics at the market, brand, and customer levels (see Figure 2.3). V2C metrics at the market level include issues such as product awareness and penetration of new products and services. Brand level V2C metrics focus on the brand evaluations and brand knowledge of customers. For example, brand awareness would be a typical V2C metric but so are brand consideration and brand attitudes. Some of these brand attitudes, such as brand uniqueness and brand innovativeness, are considered as input merely for attitudinally based brand equity measures. At the customer level, typical metrics would be customer satisfaction and relationship quality measures. Sometimes these metrics are referred to as customer feedback metrics (e.g., De Haan, Verhoef & Wiesel, 2014). A very popular V2C metric is the Promotor Score. One might argue that operational measures, such as the number of complaints, or the number of reported problems with the product or service, can be considered as V2C metrics. Although these metrics are typically not evaluations of customers and could be mainly considered as input for customers’ perceived value, they could be very valuable measures reflecting the delivered value to customers (e.g., Gijsenberg, Van Heerde & Verhoef, 2015). In this era of increasing data availability these metrics have become more available and they should definitely be considered in an extended V2C value creation analysis.



	Value-to-Customer ☺☺☺	Value-to-Firm €€€
Market/ category	• Product awareness • Product attractiveness • Product uniqueness	• Market volume/size • Market growth • Number of competitors • Market concentration • Trial rate • Repeat volume
Brand	• Brand/advertising awareness • Brand association • Brand consideration • Brand preference • Brand linking • Brand likes/comments	• Brand penetration • Brand sales • Brand/market share • Brand repurchase • Brand Equity
Customer	• Customer Satisfaction • Net Promotor Score (NPS) • Customer Effort Score • Reviews: volume & valence	• Customer Lifetime Value (and components) • Customer Engagement Value • Path to Purchase • Marketing ROI

FIGURE 2.3 Classification of V2C and V2F metrics

Typical V2F metrics at the market level are market volume, category sales, market size, and number of customers. These V2F metrics are generally not so firm specific. At the brand level, one would measure brand or market share and brand sales, and also brand equity, which is a more monetary evaluation of the brands' value. Another measure that can be used here is revenue or price premium (Ailawadi, Lehmann & Neslin, 2003). At the customer level, Customer Lifetime Value is a customer metric that has received enormous attention in the last decade. It can be considered as a key V2F customer metric that really tries to capture the monetary value generated by an average customer over the whole of his or her relationship with firms. This measure can be extended by also considering Customer Engagement Value (Kumar *et al*., 2010), that may include outcomes, such as referrals and actual word-of-mouth (e.g., Bijmolt *et al*., 2010). An in-depth discussion of V2F and V2C metrics can be found in Chapter 3.

2.4 DATA ASSETS

Assets are usually considered as resource endowments that a firm has accumulated over time. These assets can be both tangible (i.e., plant) and intangible (i.e., brands, customer relationships). In the past, customer databases were considered as an important asset for firms (Srivastava, Tasadduq & Fahey, 1998). These databases could be used, for example, to create stronger relationships with customers, and to achieve higher loyalty and more efficient and effective (cross)-selling techniques. Nowadays, data itself seems no longer unusual. One could argue that data are no longer so valuable, as data are omnipresent, can be collected in multiple ways, and are frequently publicly available to many firms (i.e., data on online reviews). In principle, we sympathize with this view. However, we also observe that within firms there is a lack of knowledge about the simple existence of data within the firm itself and outside the firm. Moreover, due to privacy regulation (i.e. GDPR) data could actually become scarce again.

Big data itself has also changed the data landscape. Big data has specific characteristics known as the 3Vs of big data, posing specific challenges for researchers and managers (Leeflang *et al*., 2014; Taylor *et al*., 2014):

- Increasing data **V**olume
- Increasing data **V**elocity
- Increasing data **V**ariety

Increasing data volume implies that databases become very large, where the analysis of data of millions of customers with hundreds of characteristics is no longer an exception (e.g., Reimer, Rutz & Pauwels, 2014). Data is also arriving more quickly, which encourages faster analysis and faster action (Leeflang *et al*., 2014). We have moved from yearly data to monthly data, to weekly data, to daily data, and now even to data per hour/minute or even real-time data. Finally, data is becoming more complex as it arrives in different formats. In the past, numerical data was standard. Nowadays, there is also more unstructured data, such as text and audio data, and also video data through, for example, YouTube. Other examples include data on Facebook postings, GPS data from mobile devices, etc. The three Vs have been extended to 5 Vs with Veracity and Value being added. Veracity refers to the messiness and trustworthiness of data. With the increasing availability of data, not all data is as reliable as one would like. Hence, data quality can be low. For example, it is known that customer reviews are frequently manipulated. Value refers to the value that is captured from analyzing and using the data. Although we clearly do acknowledge that value should be captured (see our data value creation model), it is not a specific characteristic of big data. That specific characteristic is changing analytics. We will discuss the different sources and types of (big) data in Chapter 4.

Furthermore, the changing nature of data also leads to more data sources that are no longer available in one single database and that have different characteristics. One challenge firms face is integrating the different sources of data. Data integration is thus an important topic which we discuss extensively in Chapter 5.

The increasing availability and use of data have also induced a stronger concern about privacy and data security. In 2016 the EU implemented regulation on data and privacy, General Data Protection Regulation (GDPR), with important consequences for data science. Within a data science strategy, there ample attention should be given to privacy and data security. We therefore discuss this topic in Chapter 6.

2.5 DATA ANALYTICS

Reading a textbook about data science, one would probably expect data analytics to merit immediate attention. However, analytics not embedded in the organization and without the relevant data, culture, and systems will have limited impact and value creating potential.

The presence of data provides huge opportunities for analytical teams. One of the easiest approaches is probably just to start analyzing and digging into the available data. By digging into the data, one might gain very interesting insights which can guide marketing decisions. The most famous example in this respect is that of UK-based retailer Tesco. When they analyzed data on their loyalty card, they discovered that consumers buying diapers also frequently buy beer and chips (Humby, Hunt & Phillips, 2008). Although such an example can be inspiring, we posit that before starting any analytical exercise one should clearly understand the benefits and disadvantages of the specific analysis strategy, as well as have sufficient knowledge of analytics.

Sufficient knowledge and human intelligence around analytics are extremely important. In today's data science a new set of techniques are available, specifically focusing on machine learning and artificial intelligence. Machine learning and artificial intelligence are container terms, including both new and existing techniques. In this book, we will discuss both the existing and new techniques. We make a distinction therein between what we call data exploration techniques and data modeling techniques. Data exploration techniques are generally more descriptive and focus on gaining sufficient insights on how, for example, sales develop, or customer behavior is changing over time, and how satisfaction differs between different customer segments. It also focuses on finding underlying structures in the data, as well as detecting customer segments. Data modeling focuses on detecting (causal) relationships between independent and dependent variables. Within marketing analytics one then frequently focuses on the impact of marketing instruments on marketing metrics, such as the impact of offline advertising on online sales or the impact of a service improvement on customer satisfaction. We will discuss general analytical strategies in Chapter 7, while data exploration techniques in data modeling will be discussed in Chapters 8 and 9. In those chapters, we look in more detail at machine learning.

2.5.1 The power of visualization and storytelling

Statistically trained analysts will find it easy to understand numerical outcomes of analyses. However, for many other people interpreting numbers remains a challenge. Presentation of analyses and discussion of outcomes are therefore crucial tasks in analyzing the data. One way to have more impact is to visualize the data because humans generally look for structures, anomalies, trends, and relationships. Visualization supports this by presenting the data in various forms and with different interactions. This can provide a powerful qualitative overview of data and analytical results. It can also show important relationships in the data (Grinstein & Ward, 2002). Visualization is frequently used in online dashboards. For example, the Dutch government provides descriptive data with many visuals on the development of the COVID-19 pandemic on an online dashboard. We believe that visualization is a very important analytical capability whose importance is frequently neglected despite its ability to enable researchers to have more impact on daily marketing decisions, through its enhancement of the accessibility of analytical results, especially for right-brain trained marketing executives. Nevertheless, one should also be very careful. Visualization can lead to an oversimplification of results (i.e., by providing a scatter plot of a spurious correlation) or can easily distort results with some scaling tricks on graphical axes.

Next to visualization, data science results should also be accurately communicated to decision-makers. Decision-makers should understand a problem and know what step is required next. As mentioned, visualization can aid this communication, but it also requires a clear storyline in which the decision-maker is taken through the problem, the outcomes of the analysis and the resulting actions. We will discuss both visualization and storytelling in Chapter 10.

2.6 VALUE CREATION

We consider two ways in which data science can create value for customers and firms. First, data science or data analytics can create important new insights that improve marketing decision-making. The focus of data science is then on the marketing function. For example, data analytics can demonstrate how firms can improve customer satisfaction through improving specific features of the service experience. Having these insights, marketing budgets can be allocated more effectively. Instead of relying on intuition, brand managers can, for example, invest in a positioning strategy that effectively differentiates brands from competitors.

A second value-creation route of data analytics is the development of more effective marketing campaigns and, more specifically, more effective targeting of campaigns by selecting the right customers through personalization. This value-creation route thus focuses on the customer-firm interface, which is becoming more prominent in digital environments. The improvement of actions and campaigns is mainly relevant in CRM and online environments. It mainly has to do with whom to target, when to target, and with what message. It has been shown that through effective selection of customers, the ROI of campaigns can be improved (e.g., Bult & Wansbeek, 1995). We also observe that customization of messages and offers, specifically in an online environment, can be very valuable (Ansari & Mela, 2003). Another development in an online environment is real-time behavioral targeting which is being used to adapt and personalize online environments and advertising to the specific considered needs of the customer. Value creation from data science will be discussed in Chapter 11.

2.7 DATA ANALYTICS CAPABILITIES

Still, the value of data lies not in the mere presence of the data and analytics, but in the underlying capabilities that are able to exploit these data. We consider capabilities as the *glue* that enables a range of assets to be exploited to create value (Day, 1994). For example, using different data sources on customer experiences, one could learn how to improve these experiences, thereby also building on the qualitative input of key-customers (a relational asset) that may further improve the customers' experience.

Capabilities surround data science concern:

1. People
2. Process
3. Systems
4. Organization.

People

To exploit data people are very important. Without the right set of skilled analysts, it is not sensible to develop a data science strategy. Having intelligence departments with the right capabilities is of essential importance (Verhoef & Lemon, 2013). This is one of the biggest challenges for firms (Leeflang *et al.*, 2014). Firms now have many data scientists, but such people are difficult to find. As a consequence, firms have also chosen to educate data scientists in-house through, for example, specific internal programs and academies (Verhoef & Lemon, 2013). In addition, creating value with data science is not solely the responsibility of data scientists. The benefits of data science will only become truly scalable by ensuring that every employee can contribute to the data use and analysis to create value for the organization and its customers. Therefore, more organizations invest in developing the basic skills of all employees within all layers of the organization, from decision-makers to users at such a level so that they understand the possibilities and impact of using data and analytics for their own work.

Process

Processes in smart data analytics mainly concern the way firms have organized data-input and storage, the accessibility of data to analytical teams, and the cooperation between analytic teams and (marketing) management. The first two processes are relevant for smooth and real-time data-accessibility. Importantly, these processes also involve how firms deal with privacy and data security issues and legal issues with regard to data usage. The other aspect of processes concerns how marketing and analytical teams communicate. The analytical cycle starts with a business challenge and has the ultimate goal of finding a solution for this. The most important hurdle, however, is the cooperation between the different departments and the analytical department. These departments can sometimes function as separate silos that do not understand each other. On the one hand, marketing should clearly communicate the management problems and challenges it faces and how analytics could be helpful in solving them. On the other hand, analytical teams should be able to effectively communicate their findings through insightful reports and marketing dashboards. Moreover, in an era where data analytics can create value, analytical teams should be able to effectively communicate data-based value creating solutions to management. These processes will probably develop naturally but it might also be important to define processes up-front so that, for example, marketing is expected to talk to their analytical teams when a marketing problem (i.e., a decrease in loyalty) is observed.

Systems

Concerning systems, we strongly emphasize the importance of data integration and providing an integrated data-ecosystem that allows the firm to analyze data from multiple sources. We continue to see that, within firms, data are collected in different systems or databases that are not sufficiently linked. This data-integration requires specific data management skills and software. Data-integration becomes even more difficult when firms are operating in multiple channels or in multiple countries where different systems are being used (Neslin *et al*., 2006). A key-question for firms is to what extent data should be integrated, as the marginal returns on data integration might decline (Neslin *et al*., 2006). An important trend with systems, due to the size of big data, is that cloud solutions have been developed. Similarly, we observe several new trends in available analytical software. One of the major trends is the development of open source “packages”, such as R and Python, which can be used for free. Although this involves a lot of programming, the programs are widely shared between communities of users, which means that these packages become more easily accessible.

Organization

Besides having skilled people, firms also need to devote attention to how data analytics can be organized internally. One crucial question in this respect is whether analytics or intelligence departments can have a real impact on daily business. We have observed several models for how the analytical function is embedded within firms. Typically, data science functions are located in a separate staff department that provides the marketing and sales line functions with the outcomes of their analyses, either on request or self-initiated. However, in order to have a stronger impact, firms may also deliberately choose to integrate the intelligence department in the marketing/sales department. The underlying idea is that this will induce a stronger use of analytics within marketing decision-making (Hagen *et al.*, 2013). It may, however, also reduce the independence of the data science department with negative consequences, such as a lack of innovation and insufficiently thought through analyses. A disadvantage of such an organization might also be that analytical knowledge is not used optimally within the organization as it is fragmented over multiple departments and/or functions.

2.7.1 The role of culture

One of the most prevalent issues in benefiting from data science is the internal culture and related processes. Traditionally, marketing has been a function that tended to rely on intuition and gut feeling. Simply having a good idea is, however, no longer good enough in many firms (De Swaan Arons, Van den Driest & Weed, 2014). In fact, there is an increasing trend towards more data-driven or fact-based decision-making, partially explained by a stronger emphasis on marketing accountability (Verhoef & Leeflang, 2009). Data science can only survive within firms that embrace this trend and indeed are open to relying more on analytics and their resulting insights and models in providing ideas for innovation and showing the effectiveness of specific marketing actions etc. Support from top management is essential to create a strong data science function that has something meaningful to say within the organization. It has often been shown to be a driver of the adoption and use of marketing decision support systems and data-based marketing (e.g., Van Bruggen & Wieringa, 2010). In addition, marketing departments have to change their decision-making styles and gain more knowledge on analytics and how they can be used to make smarter marketing decisions. One specific challenge is how the analytical left-brain culture can be combined with a more creative/intuitive right brain culture (De Swaan Arons, Van den Driest & Weed, 2014; Leeflang *et al*., 2014). We discuss the four capabilities and culture in Chapter 12.

2.8 DATA SCIENCE VALUE CREATION MODEL AS GUIDANCE FOR THIS BOOK

As set out above, the different elements of the data science value creation model are discussed in different chapters. We have extended the model in Figure 2.1 with the inclusion of the relevant chapters (see Figure 2.4), as referred to previously. This figure summarizes how the chapters relate to each element in our data science value creation model. In each chapter we start with this model to show where the chapter fits within the model, which guides our discussion on how data science can be used for creating value in marketing.

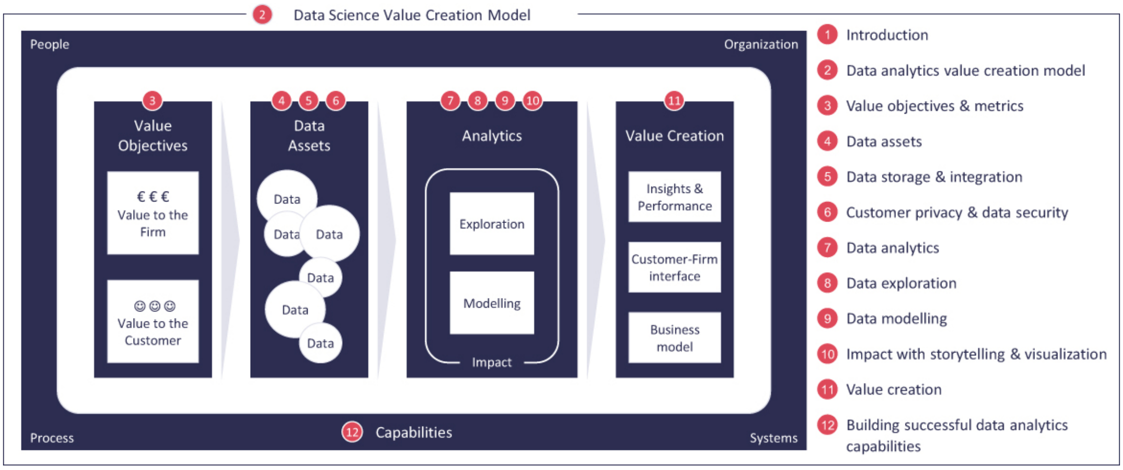


FIGURE 2.4 Data analytics value creation model

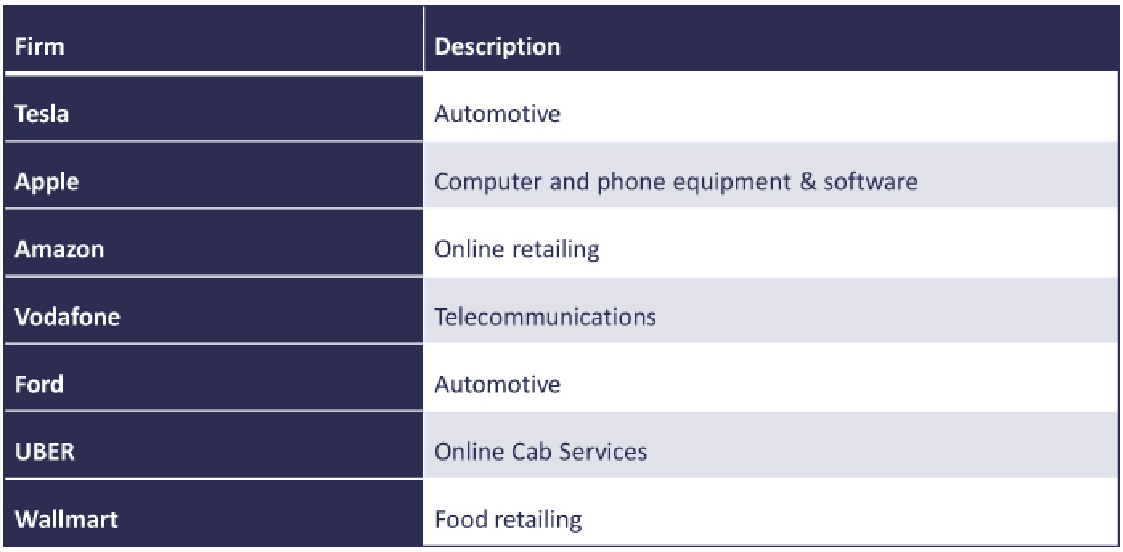
2.9 CONCLUSIONS

In this chapter, we discussed the data science value creation model in marketing. This model is essential to understand the value-creation potential of data analytics. Data assets are an important first step in data science strategy. Data analytics can create marketing insights and models that can subsequently improve marketing decision-making and improve the success of actions and campaigns. We considered value as a multi-dimensional construct consisting of value creation to customers (V2C) and value creation to firms (V2F) suggesting that firms should define clear value objectives before starting up a data science project or strategy. Data science can result in both more V2C and more V2F. The capabilities involve people, processes, systems and the organization.

ASSIGNMENT 2.1: V2C AND V2F COMPANY CLASSIFICATION

With data science, value can be created in two ways; value to the organization (V2F) and value to the customer (V2C). As argued previously, successful organizations strive to realize both V2F and V2C. One way to visualize how companies succeed in achieving this is with a matrix (see Figure 2.2). The vertical axis outlines the created V2C and the horizontal axis shows the created V2F. Combining these two axes leads to four quadrants, with the top right quadrant displaying where lots of V2F and lots of V2C is created, in other words the Win/Win quadrant.

Questions:



Firm	Description
Tesla	Automotive
Apple	Computer and phone equipment & software
Amazon	Online retailing
Vodafone	Telecommunications
Ford	Automotive
UBER	Online Cab Services
Wallmart	Food retailing

FIGURE 2.5 List of different industries

1. Below we list companies from different industries (see Figure 2.5). Indicate for each company in which quadrant you would categorize it and explain your reasoning.
2. For each of the V2F axis and the V2C axis, give two examples of measured values/ variables with which you can determine V2F and V2C per company.
3. Indicate for at least two companies (from the list above) that you have placed in two different quadrants, which measurements for V2F and V2C are crucial for their position in their quadrant.
4. Define which strategy/ policies you would suggest to a company in order for the company to move towards quadrant 4 if they are currently in quadrants 1, 2, or 3.

REFERENCES

Ailawadi, K., Lehmann, D., & Neslin, S. (2003). Revenue premium as an outcome measure of brand equity. *Journal of Marketing*, *67*(4), 1–17.

Ansari, A., & Mela, C. (2003, May). E-Customization. *Journal of Marketing Research*, *40*(2), 131–145.

Bijmolt, T., Leeflang, P., Block, F., Eisenbeiss, M., Hardie, B., Lemmens, A., & Saffert, P. (2010, Augustus). Analytics for customer engagement. *Journal of Service Research*, *13*(3), 341–356.

Bouma, J., Bugel, M., Verhoef, P., Alleman, T., Wiesel, T., & Wesselius, T. (2010, April). Dutch customer performance index: het nieuwe meten van klantprestaties. *Tijdschrift voor Marketing*, 58–63.

Bult, J., & Wansbeek, T. (1995). Optimal selection for direct mail. *Marketing Science*, *14*(4), 378–394.

Day, G. (1994, October). The capabilities of market-driven organizations. *Journal of Marketing*, *58*(4), 37–52.

De Haan, E., Verhoef, P., & Wiesel, T. (2014, February). The predictive ability of different customer feedback metrics for retention. (K. Pauwels, ed.) *International Journal of Research in Marketing*, *32*(2), 195–206.

De Swaan Arons, M., Van Den Driest, F., & Weed, K. (2014, July/Augustus). The ultimate marketing machine. *Harvard Business Review*, *92*(7/8), 54–63.

Farris, P., Bendle, N., Pfeifer, P., & Reibstein, D. (2010). Marketing metrics: The definitive guide to measuring marketing performance. *Journal of Research and Management*, *6*(1), 18–23.

Gijsenberg, M., Van Heerde, H., & Verhoef, P. (2015). Losses loom longer than gains: Modeling the impact of service crises on customer satisfaction over time. *Journal of Marketing Research*, *52*(2), 642–656.

Grinstein, G. & Ward, M. (2002). Introduction to data visualization. In: U. Fayyad, G. Grinstein, & A. Wierse (eds) *Information Visualization in Data Mining and Knowlegde Discovery* (pp. 21–46). USA: Morgan Kaufmann Publishers.

Hagen, C., Kahn, K, Ciobo, J., Miller, J., Wall, D., Evans, H. & Yady, Y (2013). *Big data and the creative destruction of today's business models*, Holland Management Review, 148(4), 25–37.

Hansen, M.T., Ibarra, H., & Peyer, U, (2013). *The best-performing CEOs in the world*, Harvard Business Review, 91 (1/2), 81–97.

Hoekstra, A., Mekonnen, M., Chapagain, A., Mathews, R., & Richter, B. (2012, February). Global monthly water scarcity: Blue water footprints versus blue water availability. *PloS One*, *7*(2), e32688.

Humby, C., Hunt, T., & Phillips, T. (2008, November). *Scoring Points: How Tesco Is Winning Customer Loyalty* (2nd edition). London: Kogan Page Limited.

Korschun, D., Bhattacharya, C., & Swain, S. (2014, May). Corporate social responsibility, customer orientation, and the job performance of frontline employees. *Journal of Marketing*, *78*(3), 20–37.

Kotler, P., & Armstrong, G. (2014). *Principles of Marketing*. Pearson Education. Harlow, United Kingdom

Kumar, V., Aksoy, L., Donkers, B., Venkatesan, R., Wiesel, T., & Tillmanns, S. (2010, Augustus). Undervalued or overvalued customers: Capturing total customer engagement value. *Journal of Service Research*, *13*(3), 297–310.

Leeflang, P., Verhoef, P., Dahlström, P., & Freundt, T. (2014, February). Challenges and solutions for marketing in a digital era. *European Management Journal*, *32*(1), 1–12.

Neslin, S., Grewal, D., Leghorn, R., Shankar, V., Teerling, M., Thomas, J., & Verhoef, P. (2006, November). Challenges and opportunities in multichannel customer management. *Journal of Service Research*, *9*(2), 95–112.

Porter, M. & Kramer, R. (2011, January/February). Creating shared value. *Harvard Business Review*, *89*(1/2), 62–77.

Reichheld, F. (1996). *The Loyalty Effect: The Hidden Force Behind Growth, Profits, and Lasting Value*. Boston: Harvard Business School Press.

Reimer, K., Rutz, O., & Pauwels, K. (2014, November). How online consumer segments differ in long-term marketing effectiveness. *Journal of Interactive Marketing*, *28*(4), 271–284.

Reinartz, W. (2011), *Presentation on customer management on Dutch customer performance awards*. Zeist.

Rust, R., Zeithaml, V., & Lemon, K. (2000). *Driving Customer Equity: How Customer Lifetime Value Is Reshaping Corporate Strategy*. New York: The Free Press.

Srivastava, R., Tasadduq, A., & Fahey, L. (1998, January). Market-based assets and shareholder value: A framework for analysis. *Journal of Marketing*, *62*(1), 2–18.

Taylor, L., Cowls, J., Schroeder, R., & Meyer, E. (2014, December). Big data and positive change in the developing world. *Policy & Internet*, *6*(4), 418–444.

Van Bruggen, G., & Wierenga, B. (2010). *Marketing Decision Making and Decision Support: Challenges and Perspectives for Successful Marketing Management Support Systems*. Boston: Now.

Verhoef, P. (2012). Multichannel customer management strategy. In V. Shankar & G. S. Carpenter (eds) *Handbook of Marketing Strategy* (pp. 135–150). Edgar Elgar Publishers, Cheltenham.

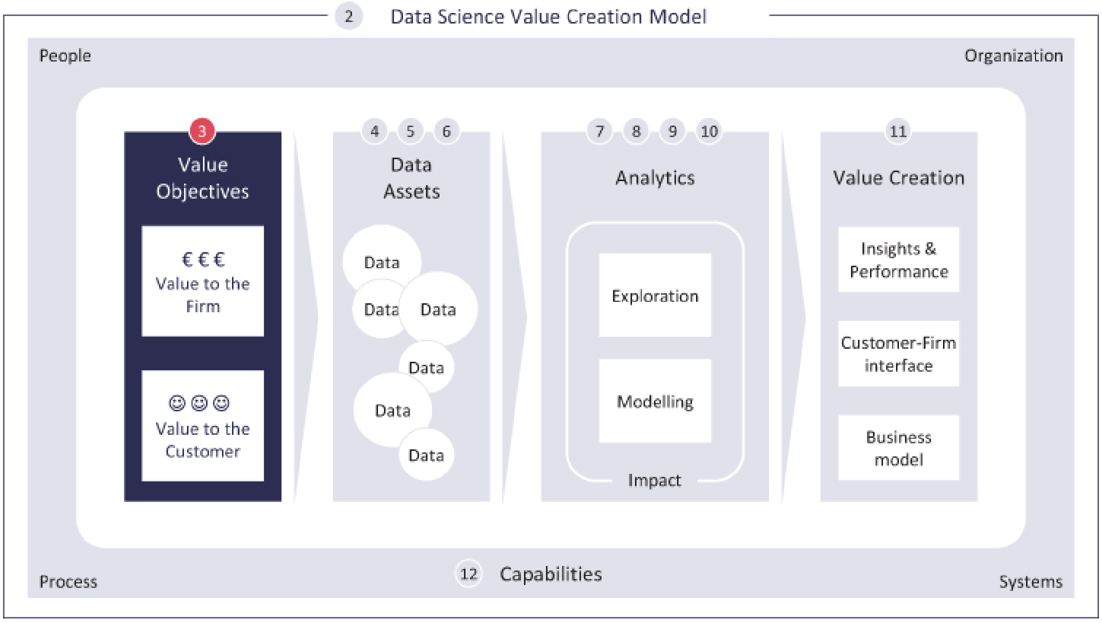
Verhoef, P. & Leeflang, P. (2009, March). Understanding the marketing department's influence within the firm. *Journal of Marketing*, *73*(2), 14–37.

Verhoef, P., & Lemon, K. (2013, February). Successful customer value management: Key lessons and emerging trends. *European Management Journal*, *13*(1), 1–15.

Wiesel, T., Alleman, T., Bouma, J.T., Bügel, M.S., de Haan, E., Hoving-Wesselius, T. & Teunter, L. (2011). *Customer performance impact; interessante relaties tussen DCPI, NPS en omzet, Customer Insights Center Rapport CIC-2011-02*, University of Groningen.

CHAPTER 3
Value objectives and metrics

DOI: 10.4324/9781003011163-3



3.1 INTRODUCTION

Before starting with a data science strategy, it should be clear which value is being created. As discussed in Chapter 2, we distinguish between V2C and V2F. Both V2C and V2F involve metrics which can be measured and collected over time. Given the focus on data science where one typically also aims to understand how value can be influenced, we focus on these metrics in this chapter. We start with V2C metrics and next discuss V2F metrics. We distinguish between market, brand, and customer metrics. Moreover, we will discuss standard metrics and new big data metrics.

3.2 V2C METRICS

Value-to-customer (V2C) metrics focus on the delivered value to customers. Sometimes these metrics also refer to “share of heart” or “share of mind.” In essence, these metrics indeed focus on what a firm achieves in a customer's mind and whether it results in positive cognitive and affective responses. These metrics in themselves do not reflect any value beyond what customers know and feel. However, they can indeed be linked to value-to-firm (V2F) metrics and extensive research has shown substantial effects of different V2C metrics on V2F metrics. Next to V2C metrics, we will also pay limited attention to value-to-society (V2S) metrics, such as corporate social responsibility.

3.2.1 Market metrics

V2C market metrics are mainly relevant in the early phases of a product life cycle, as different firms aim to communicate the value and relevance of newly introduced products and services. The important framework is the adoption model as proposed by Rogers (1995): he suggested that new products can be evaluated based on several dimensions: relative advantage, complexity, compatibility, observability, and trialability. In a broader sense, metrics could focus on the knowledge of products (**product awareness**) and, he believes, on the value offered by products (**product attractiveness and product uniqueness**). These metrics are typically measured using surveys of potential customers that have extensive scales. In Table 3.1 we give an example of how these constructs are being measured in relation to an online grocery channel. The validity of these dimensions has been frequently demonstrated, and indeed customer perceptions of these advantages predict usage intentions relating to new product innovations (e.g., Arts, Frambach & Bijmolt, 2011; Verhoef & Langerak, 2001). Another frequently used model in this respect is the so-called "technology acceptance model" (TAM). This model suggests there are two main attitudes to be considered for new technologies: ease of use and usefulness (e.g., Davis, 1989; Davis, Bagozzi & Warshaw, 1989).

TABLE 3.1 Example of items used to measure Rogers' adoption drivers (Adapted from: Verhoef & Langerak, 2001)

Perceived relative advantage
Electronic shopping is less exciting
Using electronic shopping saves much time
Using electronic shopping makes me less dependent on opening hours
Perceived compatibility
Electronic shopping suits me
Electronic shopping requires few adaptations in my personal life
Electronic shopping creates few problems for me
Perceives complexity
Electronic shopping is complex because I cannot feel and see the products

With electronic shopping it is hard to find the needed products
With electronic shopping it is difficult to order products
With electronic shopping it is problematic to compare products
Electronic shopping is complex
Intention to adopt electronic grocery shopping
Please indicate on the response scale from 0 to 10 to what extent you intend to use electronic shopping to obtain your groceries in the near future

3.2.2 New big data market metrics

Big data developments and specifically online conversations on products and product usage, may give firms a deeper understanding of how customers view and use products. Of specific use here could be statistics on the use of different search terms on search engines, such as Google and Yahoo. These search terms may show initial interest in products and brands.

A tool used for this is Google trends. In Figure 3.1 we show the search results for "tablet" as a product over time. As one can observe, the number of search requests for tablets increased until 2014 and then decreased, and one can observe some peaks. New updates of, for example, the iPad could explain these peaks.

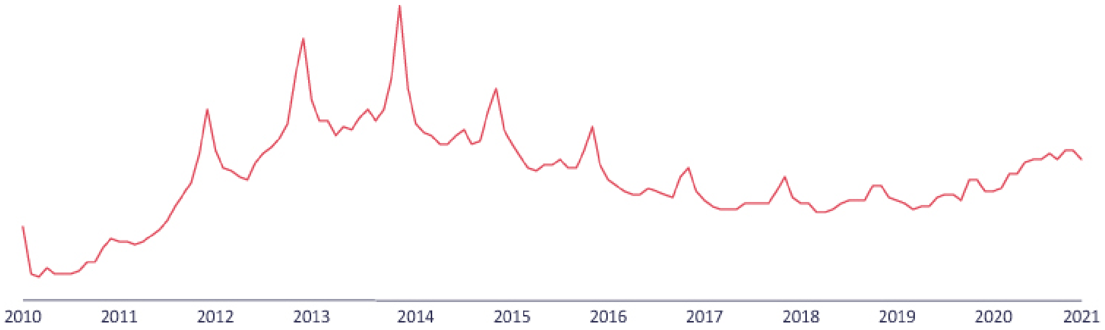


FIGURE 3.1 Search results on tablet worldwide

Source: Created using Google trends in 2021

Similar figures can be derived from multiple, and also more generic, search terms. In Figure 3.2 we show search results for data science and big data. Here you clearly see that the interest in big data is declining, whereas there is a strong and increasing interest in data science. Importantly, Google trends can also show interest across, for example, different geographical markets and specific cities across the globe.

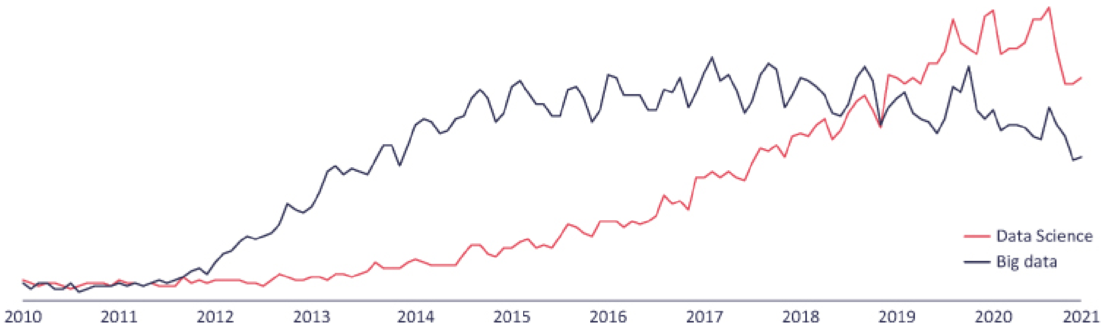


FIGURE 3.2 Search interest in data science and big data

Source: Created using Google trends in 2021

3.3 BRAND METRICS

V2C brand metrics are frequently collected on a continuous basis. For many firms it is very important to continuously measure indicators of their brand performance and, related to that, track the outcomes of advertising campaigns. Many research and advertising agencies around the globe, such as Young & Rubicam, have developed standard brand performance measurements. Importantly, brand metrics are collected among all customers in the market place, as firms aim to measure the position of brands relative to competing brands. This contrasts with customer metrics, which are typically measured among existing customers of a firm or brand.

Traditional brand performance measures can be structured around the sales funnel from being aware to final purchase of the brand and subsequent resulting loyalty. Brand metrics can also be classified based on their focus. Broadly, one could distinguish between cognitive brand metrics focusing on customers' knowledge of a brand and affective measures focusing on customers' feelings and emotions towards a brand (Hanssens *et al*., 2014).

A typical cognitive brand metric is brand awareness, which measures whether customers know the brand. **Brand awareness** can be unaided or spontaneous, reflecting top-of-mind awareness and aided brand awareness (see Figure 3.3 for a trend-line on these two-brand metrics). Especially for strong brands, such as Coca Cola and Apple, brand awareness metrics are close to 100% and do not vary much over time. The added value of this brand awareness metric for these brands could be relatively small. For unknown brands, tracking brand awareness can provide very useful information. Beyond brand awareness, firms also frequently measure **advertising awareness**, which focuses on whether consumers are aware of, usually, a brand's recent advertising campaign.

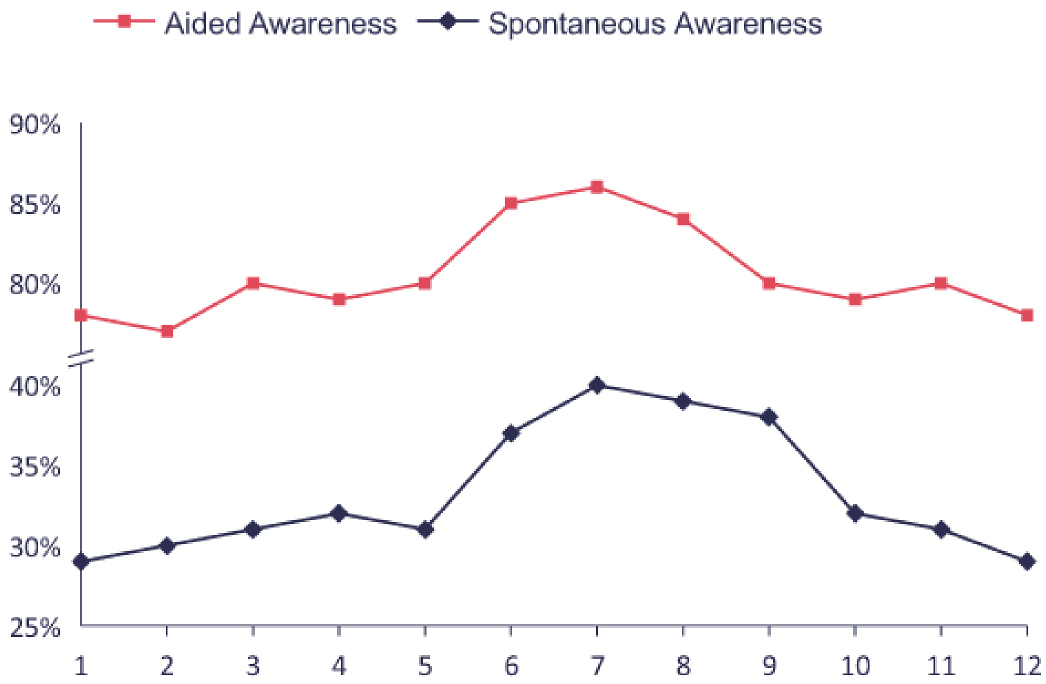


FIGURE 3.3 Example of tracking aided and spontaneous awareness through time

At a deeper level, customers gain more knowledge about the brand. This may be reflected in specific **brand associations**, such as whether it is an innovative brand or a high-quality brand. These associations are likely to vary more between brands and over time, as brands have developed specific positioning strategies. In contrast with brand awareness metrics, these metrics are also likely to change more over time (e.g.,Hunneman, Verhoef & Sloot, 2015; Mizik & Jacobson, 2009).

Brand metrics that are closer to the final purchase concern brand consideration and brand preference or brand liking. **Brand consideration** metrics focus on whether a brand is in the set of brands that customers consider buying. Brand consideration is traditionally considered as a necessary condition for a brand to be purchased. However, brands can easily enter a consideration set when they are, for example, in a promotion or there is effective in-store communication (e.g., Baxendale, Macdonald & Wilson, 2015; Van Nierop *et al.*, 2011).

Brand preference measures whether customers prefer a specific brand over competing brands. High brand preference levels for brands should typically lead to higher market shares for brands. One could argue that a brand preference measure is so close to behavior that it might actually be a V2F metric. In Figure 3.4, we provide an example of the brand preference for multiple smartphone brands and how these preference measures vary between different market segments. As one can observe, the Apple iPhone is the most preferred brand, but this preference varies between segments.

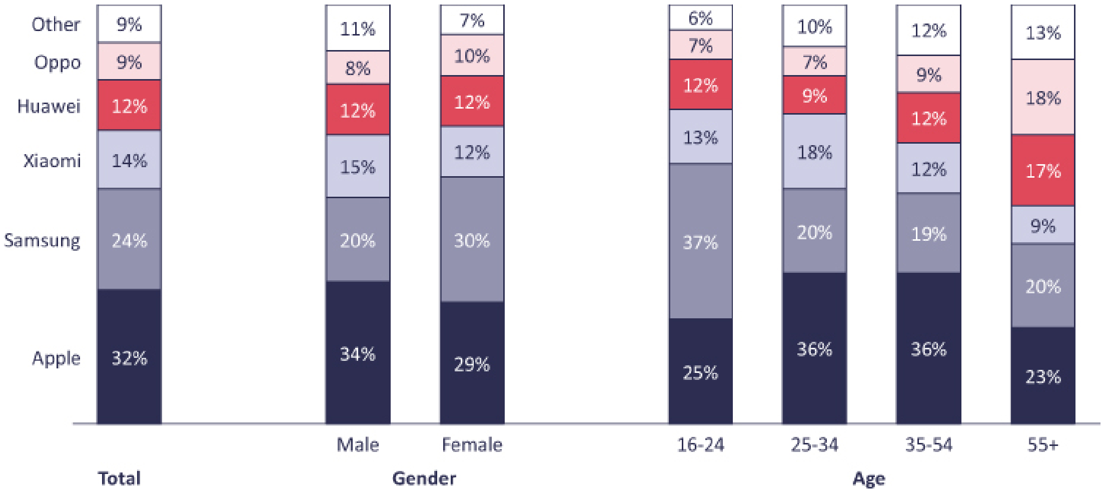


FIGURE 3.4 Fictive example of brand preference of smartphone users, de-averaged to gender & age

An alternative measure is **brand linking**, which is also a metric focusing on the effective attraction of a brand. It is typically measured with a question which customers answer by stating their liking of a brand on a scale (i.e., 1= not all like the brand, 7= "like enormously") (Hanssens *et al*., 2014).

3.3.1 Brand-Asset Valuator®

One of the most influential V2C brand measurement systems is the one developed and used by Young & Rubicam. They developed the Brand-Asset Valuator® (BAV). This brand-asset valuator focuses on multiple dimensions of the brand (see Figure 3.5).

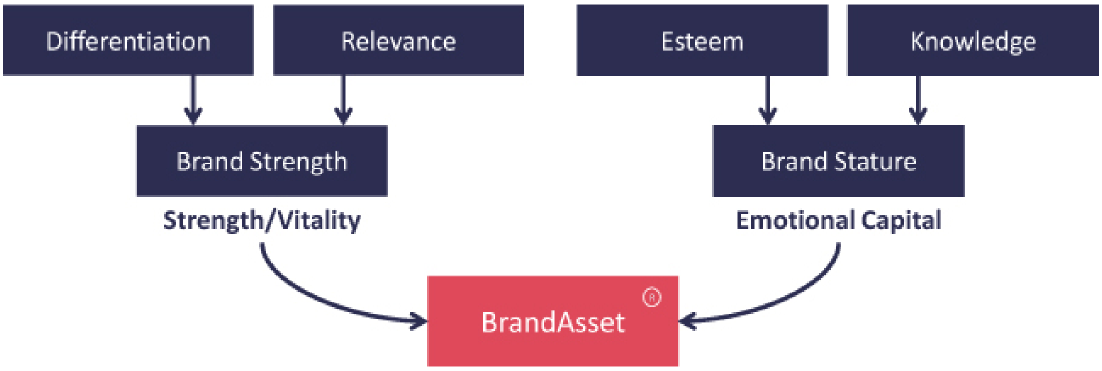


FIGURE 3.5 Brand-Asset Valuator ® model

Source: Adapted from Young & Rubicam

The BAV distinguishes between brand strength and brand stature. Each of these measures can then be subdivided into underlying metrics: differentiation and relevance for brand strength and esteem and knowledge for brand stature. In some updated models of BAV, brand energy is also measured (Mizik &

Jacobson, 2009). The measures and definitions for each of these dimensions, including brand energy, are provided in Table 3.2.

TABLE 3.2 Definitions of BAV® components (Adapted from Mizik & Jacobson, 2009, p. 16)

BAV Pillar	***Underlying Perceptual Metrics***	***Survey Scale***	***BAV Data***	***Meaning and Role of the Pillar***
Differentiation	1. Unique 2. Distinctive	Yes/no Yes/no	% responding “yes” % responding “yes”	Perceived distinctiveness of the brand. Defines the brand and reflects its ability to stand out from competition. Is the “engine of the brand train; … if the engine stops, so will the train.”
Relevance	1 Relevant to me	1–7 scale	Average score	Personal relevance and appropriateness and perceived importance of the brand. Drives market penetration and is a source of brand’s staying power.
Esteem	1. Personal 2. Regard 3. Leader 4. High quality 5. Reliable	1–7 scale Yes/no Yes/no Yes/no	Average score % responding “yes” % responding “yes” % responding “yes”	Level of regard consumers hold for the brand and valence of consumer attitude. Reflects how well the brand its promises.
Knowledge	1. Familiarity with the brand	1–7 scale	Average score	Awareness and understanding of the brand identity. Captures consumer intimacy with the brand. Results from brand-related (marketing) communications and personal experiences with the brand.

BAV Pillar	*Underlying Perceptual Metrics*	*Survey Scale*	*BAV Data*	*Meaning and Role of the Pillar*
Energy (new pillar)	1. Innovative 2. Dynamic	Yes/no Yes/no	% responding "yes" % responding "yes"	Brand's ability to meet consumers' needs in the future and to adapt and respond to changing tastes and needs. Indicated future orientation and capabilities of the brand.

3.3.2 Do brand metrics matter?

Collecting the brand metrics discussed above provides firms with early information on the future health of their brands. For example, changes in brand consideration may signal that somewhere in the near future brand market share could decline. As noted, it also provides information on the competitive positioning of the brand compared with other brands. This information may help firms to redefine positioning strategies and a communicated USP. For example, if the price image of supermarkets is changing, supermarkets may want to change their retail mix in such a way that this price image is improved (e.g., Hunneman, Verhoef & Sloot, 2015; Van Heerde, Gijsbrechts & Pauwels, 2008).

Hanssens *et al.* (2014) suggest that good V2C brand metrics should do well on three criteria, V2C brand metrics should:

1. have the potential for growth;
2. have some stickiness;
3. be responsive to marketing efforts.

The first criterion refers to whether the metric can indeed change and grow over time. This frequently does not apply for brand awareness metrics, as a natural ceiling may be reached. The stickiness of the metric focuses on the fact that the metric does not change too much over time. There should be some staying power, which may result from inertia or lock-in. The responsiveness to marketing efforts refers to marketing's ability to "move the needle" on the V2C metric. If a V2C metric does not respond to changes in the marketing mix, it is probably not a very effective metric.

3.3.3 New big data brand metrics

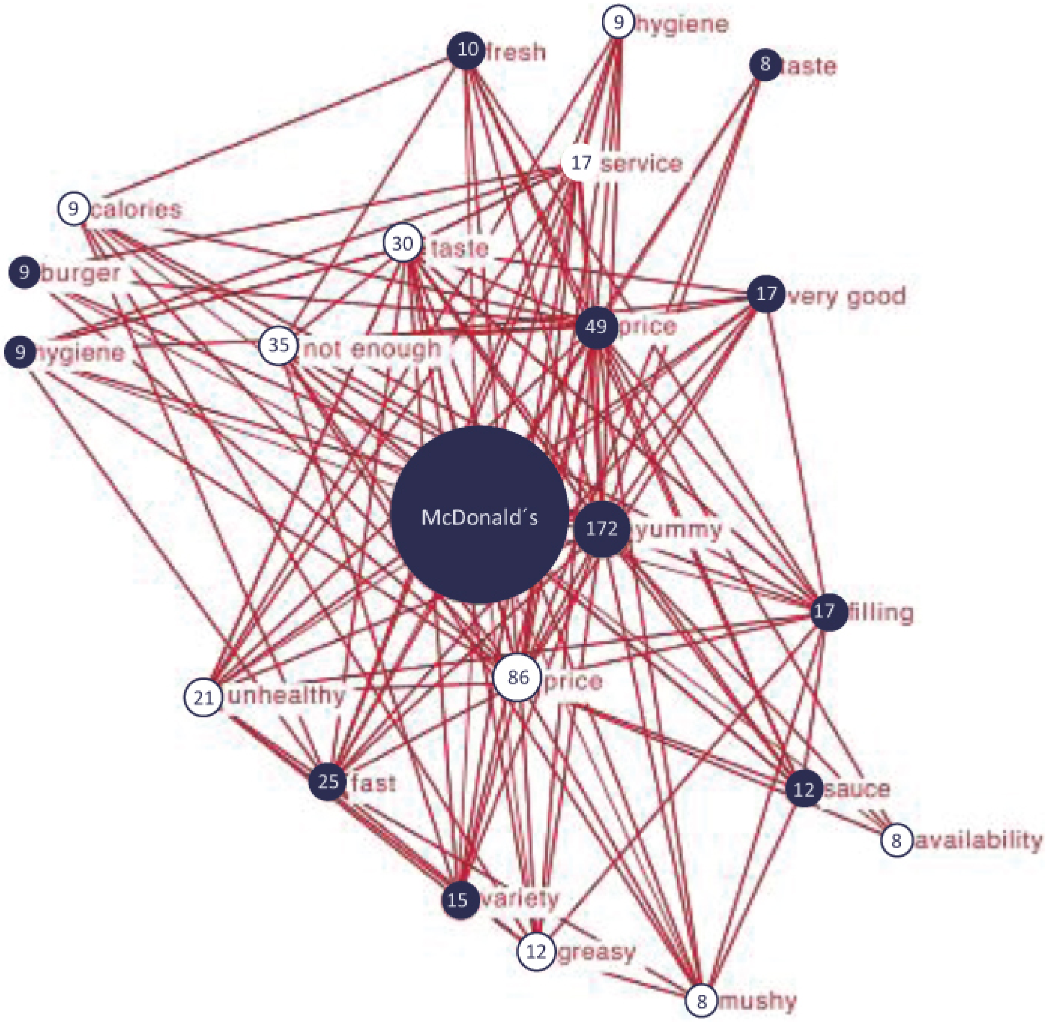
The metrics discussed so far are rather traditional and have been around for years. Only recently have we started to understand the actual impact of these metrics on brand performance outcomes. Being able to combine different data sources, together with a continuous collection of brand metrics data, has allowed researchers to do this.

Particularly as a result of online and social media developments, where customers share their opinions about brands and may also indicate their liking of brands, new sources for data on brands have been developed. This is sometimes referred to as "user-generated content" (UGC). Importantly, this UGC can be collected and analyzed to create brand metrics. We consider the following specific new big data metrics:

- Digital brand association networks
- Summarized digital brand metrics
- Social media brand metrics

3.3.4 Digital brand association networks

In expressing their opinions about brands, customers may share different views in writing about them. For example, they may share somewhere on a blog that Apple is considered innovative, whereas Samsung is a real Asian brand. Similarly, ideas on brands like Nike and Adidas can be shared. Researchers have developed methods to collect these data and to analyze them, thereby considering the valence of the words. In Figure 3.6 an association network for McDonalds is shown based on digital data. As one can observe from this network, the main association for McDonald's is yummy; however, negative associations are service, taste, and portion size (not enough).



Note: White (grey) circles respresent favorable (unfavorable) brand associations. Number in circles represent normalized, weighted degree centrality (per mill)

FIGURE 3.6 Association network of McDonald's based on online data

Source: Adapted from: Gensler *et al*., 2015

3.3.5 Digital brand metrics

With digital brand associations, we are mainly interested in actual associations, providing managers with more knowledge on what customers think about a brand. Research agencies have now developed methods to assess how positive or negative customers' views are when they discuss brands online. Using text analytics, the valence of words is then assessed. Based on dictionaries, such as the dictionary of affect, the negativity and positivity of these words can be assessed. Whereas in the past this was done manually (e.g., Antonides, Verhoef & De Hoog, 2004), it is now done automatically using text analysis programs. Based on this positive and negative valence for brands, a valence score can be calculated. These valence scores are related to the sales of brands. Correlations with shareholder value metrics have also been reported (e.g., Onish & Manchanda, 2012; Tellis & Johnson, 2007). These digital summary indices are also referred to as electronic word-of-mouth, or **eWOM** (Trusov, Bucklin & Pauwels, 2009).

3.3.6 Social media brand metrics

Not only do customers discuss brands online in social media, but the brands themselves also actively use social media for promotional purposes. Customers can visit brand pages on, for example, Facebook. Customers can react to them by 'liking' content and providing comments. The numbers of **brand likes** and brand comments are considered to be two relevant social media brand metrics. The number of brand likes may be an indicator of brand preference among customers. The number of comments may indicate some brand involvement. Importantly, the content of the social media marketing campaigns affects both metrics. For example, brands get more likes when they include a contest in the campaign and also a video (De Vries, Gensler & Leeflang, 2012). Unsurprisingly, more comments are received when a question is asked in the campaign. Overall, one could doubt the value of these metrics as real V2C brand metrics. Still, the number of likes and comments can be substantial, although it varies a lot between brands and industries (see Figure 3.7). To some extent, they mainly reflect reactions to the social media presence of a specific brand, while brand knowledge and attitudes are based on multiple interactions in multiple channels and touchpoints.

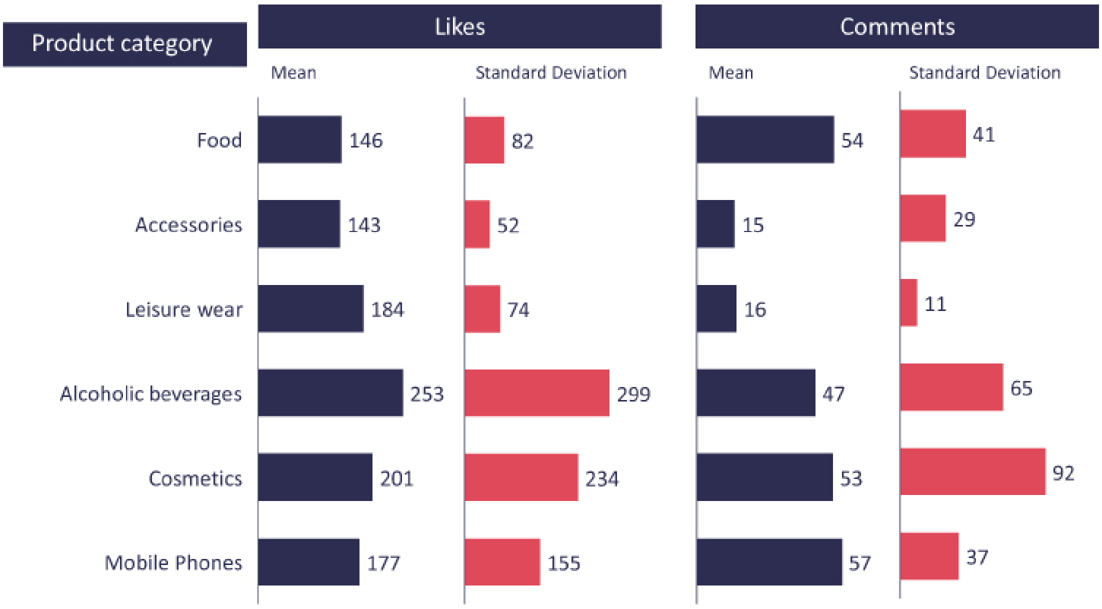


FIGURE 3.7 Average number of likes and comments per product category

Source: Adapted from: De Vries, Gensler & Leeflang, 2012, p. 87

3.4 CUSTOMER METRICS

V2C customer metrics are frequently labeled as customer feedback metrics by marketing researchers. Customer feedback metrics (CFMs) have become very popular for measuring the customer experience through the customer journey. Firms are aiming to improve the customer experience across multiple touchpoints and are seeking ways to measuring this experience (Lemon & Verhoef, 2016). Targets are set on these metrics and there is continuous measurement. In some cases, specific feedback mechanisms are built in. Such a mechanism is the so-called customer feedback loop. In this loop, customers give feedback to a firm on a specific service event and subsequently they are called back when they score low on this metric and asked about what is behind their low score. Firms then try to solve these issues, which hopefully results in higher performance. All these kinds of systems result in series of customer feedback metrics from thousands of different customers over time. The most popular metrics are the net promoter score (NPS), customer satisfaction, and the customer effort score (CES). These metrics can be measured in different ways. In Table 3.3 the exact questions and operationalizations are discussed:

TABLE 3.3 Overview of different CFMs (Adapted from: De Haan, Verhoef & Wiesel, 2015)

CFM	*Measurement*
1. Customer satisfaction	"All in all, how satisfied or unsatisfied are you with [company X]?" (1=very unsatisfied, 7=very satisfied).
2. Top-2-box customer satisfaction	A dummy at the customer level indicating if the customer has given a score of 6 or 7 on the customer satisfaction question. At the firm (industry) level this is the proportion of customers of that firm (industry) that gave a score of 6 or 7.
3. Net promoter score (NPS)	"How likely is it that you would recommend [company X] to a friend or colleague?" (0=very unlikely, 10=very likely). Respondents who gave a score of 0–6 are "detractors," those who gave a 7 or 8 are "passives," and those who gave a 9 or 10 are "promoters." Subtracting the proportion of promoters from the proportion of detractors provides the NPS at the firm level (Reichheld, 2003).
4. NPS-value	The untransformed NPS score (on the 0–10 range) provided by the customer.
5. Customer effort score (CES)	"Did you try to contact [company X] with any kind of request?" (yes/no) If yes, the following question is asked: "How much effort did you personally have to make to handle your request?" (1=very low effort, 5=very high effort).

We distinguish between these metrics on two dimensions. The first dimension is introduced by Bolton, Lemon and Verhoef (2004), who focus on the time span of measures and distinguish between more backward-looking and more forward-looking metrics. Forward-looking CFMs focus on what customers plan to do in the future and may signal something about the future performance of the relationship. The NPS introduced by Reichheld (2003) is an example of a forward-looking CFM since it considers the willingness to recommend a firm in the future, which may also signal a customer's future relationship with the firm. Backward-looking metrics focus on the past and current performance of a company towards customers.

The CES is a typical backward-looking CFM, as it measures perceived service performance from a past specific experience (Dixon, Freeman & Toman, 2010). It is based on a single question ("How much effort did you personally have to put forth to handle your request?") and measured on a five-point scale. Dixon, Freeman and Toman (2010) suggest that the CES is a better predictor of repurchase (intentions) and increased spending than the NPS or customer satisfaction.

Finally, customer satisfaction focuses more on the overall evaluation of the interactions between the customer and the firm over time and tends to have a more present focus (Verhoef, 2003), although it may also be based on past experiences.

Our second dimension is about how the measurement scale of the CFM is used. There are some who advocate not looking at the mean value of the scale, but at the proportion of people responding very positively and/or very negatively. An example of this is "top-2-box customer satisfaction," which measures the proportion of customers filling in the two highest-scoring points for their overall[1]customer satisfaction (Morgan & Rego, 2006). The calculation underlying the official NPS also distinguishes between very positive, moderate, and negative responses (Reichheld, 2003). Transformations can theoretically be defended, as it has been shown that customers mainly focus on extreme experiences and therefore the effects of CFMs can be rather non-linear (e.g. Streukens & De Ruyter, 2004; Van Doorn & Verhoef, 2008). Moreover, service marketing experts advise firms to creating extremely satisfied customers and that customers should have high scores on the CFM scales (Oliver, Rust & Varki, 1997). Firms can, however, also choose not to use a transformation, and instead make use of the full scale, e.g. the 0–10 scale of the NPS. If we combine the two dimensions, we end up with the three-by-two classification matrix as provided in Table 3.4.

TABLE 3.4 Conceptualization of studied CFMs (Adapted from: De Haan, Verhoef & Wiesel, 2015)

		Pre-defined Data

		Past Focus	*Pre-defined Data* *Present Focus*	*Future Focus*
		Past Focus	***Present Focus***	***Future Focus***
Part of the scale used	Full scale	Customer Effort Score	Customer Satisfaction	NPS Value
	Focus on extremes		Top-2-Box Customer Satisfaction	Official NPS

3.4.1 Is there a silver metric?

Proponents of different metrics propose that their metrics have the best performance for future growth and customer retention. For example, Reichheld (2003) strongly advocated the NPS, while Dixon, Freeman and Toman (2010) believed and showed that CES performed very strongly in predicting customer loyalty, surpassing the performance of other competing metrics. Unsurprisingly, academics questioned these claims and have investigated the actual quality of the metrics. Typically, they compared the performance of different metrics in predicting future business growth and customer loyalty. Initial studies were not so positive about the performance of the NPS and tended to prefer customer satisfaction (e.g. Keiningham *et al*., 2007). However, more follow-up studies show smaller differences between satisfaction and NPS (e.g., Van Doorn, Leeflang & Tijs, 2013). De Haan, Verhoef and Wiesel (2015) found no clear winner. NPS and customer satisfaction scored equally well. The CES, however, performed very poorly. They also observed that there might be some benefits in combining metrics, suggesting that at least firms should monitor multiple metrics (i.e. NPS and satisfaction). This would imply making use of dashboards involving multiple metrics. Based on existing knowledge and practical insights we have the following recommendations for firms:

- Rely on overall metrics with a **present or future** focus
- It is very valuable to analyze the development of the **top scores** of metrics (e.g., top-2-boxes)
- Do **not immediately adopt** new metrics promising superior performance, but carefully consider the actual performance
- Measure **multiple metrics** and report in a dashboard

3.4.2 Other theoretical relationship metrics

Especially within the CRM literature, other metrics or customer attitudes have been discussed and have gained attention. These metrics mainly focus on the quality of the customer relationship. Customer satisfaction can be an indicator of this quality, although this attitude usually focuses more on the cognitive side of the relationship, being the mere evaluation of delivered products and services (Bolton, Lemon & Verhoef, 2004). The most specific metric that has been proposed is commitment, defined as the enduring desire to continue a valued relationship (Moorman, Zaltman & Desphandé, 1992). In general, this attitude reflects a more emotional evaluation and also considers the future development of the relationship. Researchers have also considered several forms of commitment (e.g., Verhoef, Franses & Hoekstra, 2002).

Another customer metric that gained attention is customer **trust**. Trust is defined as the customers' confidence in the quality and reliability of the services provided (Garbarino & Johnson, 1999). Authors have also proposed that one should distinguish between reliability and benevolence. Reliability focuses on the fact that a firm acts on its promises, while benevolence considers the fact that a firm not only cares about its own interests but also the customer's interest and acts on that (Geyskens *et al*., 1996).

3.4.3 Customer equity drivers[2]

Within customer management, the customer equity model as developed by Rust, Zeithaml and Lemon (2000) has been influential. They consider customer equity—the net present value of all future and current customers—as a very important outcome variable in customer-centric firms. Conceptually they consider three drivers:

- value equity (VE), defined as a customers' objective assessment of the utility of services based on perceptions of "what is given up" for "what is received." VE reflects the outcome of customers' comparisons between their expectations and firms' performance;
- brand equity (BE), which reflects customers' subjective and intangible assessment of the brand image; and
- relationship equity (RE), defined as customers' assessments of their interactions with the firm. This factor depends on customers' relationships with sales and service persons, loyalty programs, customer communities/networks, and so forth. Positive RE provides relatively more financial and social benefits to customers (Rust, Zeithaml & Lemon, 2000). This enhances feelings of reciprocity and benevolence, which should positively influence loyalty (Selnes & Gønhaug, 2000).

These drivers can be measured with attitude-like questions in which customers evaluate the performance of firms and brands on these dimensions. There is sufficient evidence that these drivers contribute to customer loyalty.

3.4.4 New big data customer metrics

Big data customer V2C metrics mainly arise from online interactions, but also from internal data sources, which are frequently neglected.

3.4.4.1 Internal data sources

Within CRM databases, the focus has been strongly on collecting transaction data. However, firms have many interactions with customers, which can be used as indicators for V2C. In general, we advocate paying more attention to this information. For example, firms can collect customer **complaints**. A complaint is a V2C indicator and can probably predict switching behavior. We note that, according to the service recovery paradox, firms that act on complaints, can increase customer loyalty (Van Doorn & Verhoef, 2008). Personal contact with customers can also be an indicator of V2C. Personal contacts (i.e., e-mail-conversations, call center contacts) can be analyzed and the positivity or negativity of these contacts can be assessed (Antonides, Verhoef & De Hoog, 2004).

Other internal data sources can provide information on the delivered value to customers and specifically use **internal operations data**. These data can be customer-specific or measured at a more aggregate level. An example of customer-specific data involves the resolution time of a service problem for a specific customer. Bolton, Lemon and Verhoef (2008) show that operational data can predict the probability of contract renewal and the level of service upgrading. The delivered service performance can also be measured at a more aggregate level. For example, for railway firms, the percentage of trains with a specific delay or the percentage of unsuccessful connections between trains can be an internal measure for the actually delivered value (Gijsenberg, Van Heerde & Verhoef, 2015). One important issue is that these internal measures do not reflect the actual perception of the delivered value. Perceptions frequently arise from a comparison of the expected service level and the delivered service level. Expectations are, however, frequently based on the past and thus having time series data on the delivered performance may create a more accurate understanding (Gijsenberg, Van Heerde & Verhoef, 2015).

3.4.4.2 Online sources

The most important customer metric is customer reviews. Reviews of firms and specific products are placed online and frequently involve relatively independent websites, such as Tripadvisor or Zoover. Additionally, online retailers allow consumers to provide reviews of sold products. An online survey of 2,005 American shoppers showed that 65% of potential shoppers selected a brand in their consideration set that was initially not in that set, as a result of online reviews (Weber Shandwick, 2012). This suggests that reviews can be very powerful and influential in steering customer choices online. As expected, the effect of reviews on sales is significant and substantial, with a mean elasticity of

.35 increase for review volume and .65 for review valence (Floyd *et al*., 2014). "**Review volume"** is a measure of the number of reviews and "**review valence"** is a measure of the positivity of the review. There is, however, considerable variation in the effectiveness of reviews. Reviews have stronger effects for products with high customer involvement, while third-party reviews and critic reviews have a stronger effect than normal customer reviews.

Note that in an online context the differences between brand metrics and customer metrics become blurred. Customer reviews can be used as input for the creation of summary brand metrics. Furthermore, customer reviews are considered as eWOM. Despite this, it is quite clear that reviews are becoming very important. They affect sales as the customers include these reviews in their purchase decision. If, for example, a hotel has poor reviews, fewer customers will be likely to book that hotel.

3.5 V2S METRICS[3]

The V2C concept can also be extended to focus on the societal value of firms. Many firms are developing strategies and initiatives to, for example, improve their sustainability. Such initiatives are significant elements of the corporate strategy of many (multi) national corporations (Beard *et al*., 2011) Porter & Kramer (2011) assert that businesses can be a positive force for good in the world and that this approach is in the interests of all firms' stakeholders.

Firms also collect metrics to measure their performance on these societal strategies. These metrics are typically not the responsibility of marketing but are mainly administered by staff departments responsible for sustainability and/or corporate reputation. Within the management and marketing literature, there is one metric receiving considerable attention: **Corporate Social Responsibility** (CSR).

CSR is a firm's commitment to ensure societal and stakeholder well-being through discretionary business practices and contributions of corporate resources (Du, Bhattacharya & Sen, 2010; Kotler & Lee, 2005; Luo & Bhattacharya, 2006). As a broad concept, CSR can include business practices as diverse as cash donations to charity, the equitable treatment of workers, and an environmentally friendly production policy. Frequently CSR is seen as a kind of perception measure similar to satisfaction and involves seeking opinions on statements such as "this company emphasizes the importance of its social responsibilities" and "this company provides an evident social contribution" (Du, Bhattacharya & Sen, 2007).

3.6 SHOULD FIRMS COLLECT ALL V2C METRICS?

A reader of this in-depth chapter might wonder whether firms should collect all the metrics mentioned. Our answer to this is clearly: No! Firms should focus on a limited number of metrics and include them in a marketing dashboard. Managers should then strive to influence these metrics with marketing strategies. We specifically observe that many firms are not following this path. They collect a plethora of V2C metrics in different layers in the organization. For example, a European service company collected satisfaction data, data on NPS, brand-metrics, digital sentiment metrics, and corporate reputation metrics. Satisfaction and NPS data are mainly under the ownership of the marketing research department. Marketing management and communication are mainly interested in brand metrics, whereas corporate strategy owns the corporate reputation metrics. The use of multiple measures results in strong discussions on which metrics to use, instead of how they could be influential and how their use could give value to the firm. When these metrics are analyzed it frequently becomes clear that there are strong correlations between many of them. An in-depth comparison of each of these metrics using specific criteria developed by Ailawadi, Neslin and Lehmann (2003) (see Table 3.5) on what good metrics are resulted in the decision to focus on satisfaction as the most important outcome metric.

TABLE 3.5 Criteria for good metrics (Adapted from: Ailawadi, Neslin & Lehmann, 2003)

1. Theory based
2. Complete
3. Diagnostic
4. Future potential
5. Objective
6. Based on existing data
7. A single number
8. Intuitive and trustworthy for top management
9. Robust and reliable
10. Validated with other outcome measures

3.7 VALUE TO FIRM METRICS

Value-to-firm metrics focus on the value delivered by customers to firms. These metrics are typically behavioral and are frequently financial in nature. As with the V2C metrics, firms can choose from a plethora of V2F metrics. Some of them, such as brand sales and market share, have been around for a long time, while other metrics, such as Customer Lifetime Value are relatively novel. Big data developments have created new V2F metrics. Especially in digital environments, many new metrics have been developed and are now being used in digital marketing. The challenge for firms is how to interpret and use these new metrics and link them to existing metrics (Leeflang *et al*., 2014).

3.8 MARKET METRICS

At the market level, firms will mainly be interested in metrics that show the attractiveness of the market. These market size metrics are relevant for firms when they make strategic market entry decisions and in strategic portfolio decisions. Based on these types of analyses a company such as Royal Dutch Phillips can make strategic decisions: it decided to retract from the electronics and lighting market to focus on health. Market metrics are also very important when considering new product sales. We, therefore, distinguish between two types of market metrics:

1. Market attractiveness metrics
2. New product sales metrics

3.8.1 Market attractiveness metrics

Typical market metrics being used are:

– Market size: The total market demand in terms of number of (potential) customers (i.e., target population), units, or $ sales
– Market growth: The annual growth of the market
– Number of competitors
– Market Concentration: This metric measures whether the market is dominated by a few players or whether many players have relatively equal positions in the market. The so-called Herfindahl Index is used to measure this. It is defined for a specific market with J firms as:

$$Herfindahl\ Index = \sum_{j=1}^{J} Market\ Share_j^2$$

and can vary between 0 and 1, with 1 representing a very concentrated market with a monopolist having 100% of the market.

One problem that firms face is how to define the market and competition. This is becoming rather difficult, because of market fragmentation. For example, strong retailers such as Tesco not only sell fast-moving consumer goods, but also telecom subscriptions and financial services. Similarly, new players are entering the market in this digital era. Google can, for example, become a competitor of online retailers when they it actively starts Google Shop, whereas banks are confronted with the growth of new payment forms such as Paypal and Bitcoin and are also worried about Google getting into banking. In Figure 3.8 we provide an example of the worldwide sales of smartphones.

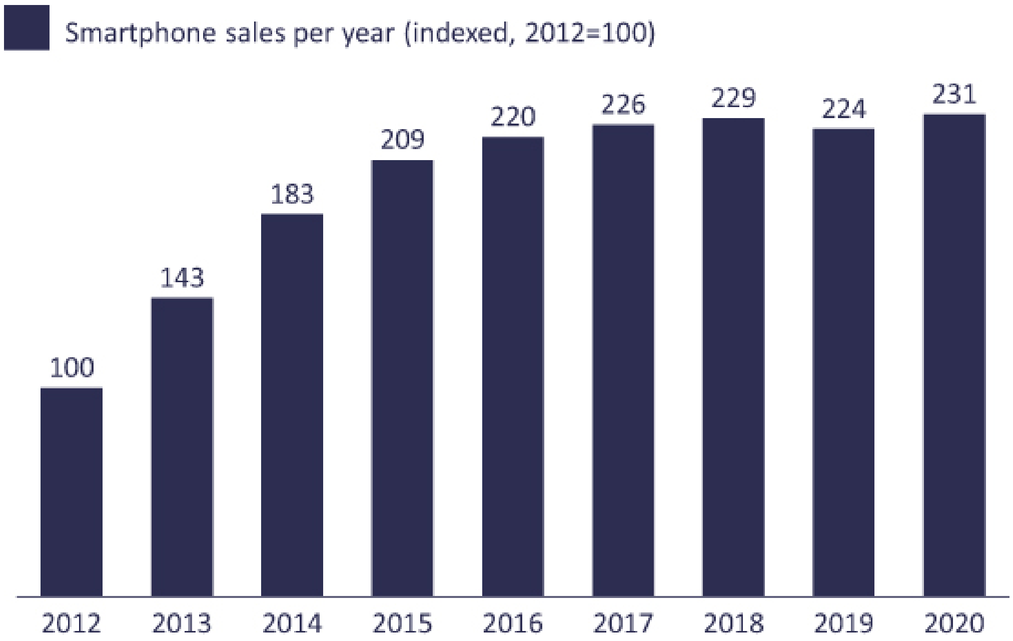


FIGURE 3.8 Indexed worldwide smartphone sales

Source: Adapted from Statista

3.8.2 New product sales metrics

For forecasting new product sales it is considered important to use several metrics (e.g., Farris *et al*., 2006):

– Trial rate: The number of first-time new product users as a percentage of the target population
– Repeat volume: Number of repeat buyers multiplied by the number of products they buy in each purchase, multiplied by the number of times they purchase per period
– Penetration rate: Number of repeat users plus the number of new trials, divided by the market population.

Based on these figures, volume projections of the market can be made. A very simple equation for sales volume is for example:

Sales Volume = Penetration Rate * Purchase Frequency * Units Purchased

Note that, so far, we describe the above metrics at the product level. However, similar metrics can be calculated for (new) brands.[4]

Brand Metrics

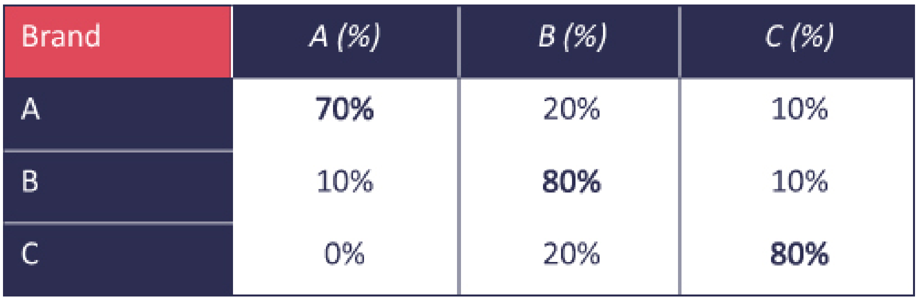
We consider two types of V2F brand metrics:

1. Brand market performance metrics
2. Brand valuation metrics

3.8.3 Brand market performance metrics

Brand market performance metrics focus on the actual performance of brands in the market. Traditional brand metrics are well known and are very frequently measured: **brand penetration, brand sales, and market share**. Brand sales can be measured internally. Brand penetration and market share data are more difficult to measure as information on the whole market is required. In some markets, such as the FMCG (fast-moving consumer goods) market, these metrics are measured daily or weekly using scanning technologies by market research firms such as GfK, AC Nielsen, and IRI. In many other markets, including financial services and telecom, data on market shares are less frequently measured. One specific issue is that in these markets, data collection on actual purchases is not well organized or is fairly difficult to execute. For example, when measuring market shares in the insurance industry, customers have to report accurately on their ownership of insurances and the firms they purchased from. It is questionable whether this can be done accurately.

Next to these more aggregate brand market performance metrics, firms also frequently measure brand loyalty metrics. These metrics could be based on stated intentions, such as **brand repurchase intentions**. This can be measured using scales, such as a Juster scale (0 = will absolutely not repurchase, 10 will absolutely repurchase) (Juster, 1969). However, one could also use scales in which customers have to divide 100 points across brands in the marketplace—called a constant sum scale (Rust, Lemon & Zeithaml, 2004). This more accurately reflects that customers often buy multiple brands and are frequently not loyal to one specific brand. This purchasing behavior of multiple brands has been referred to as "polygamous loyalty" (e.g., Dowling & Uncles, 1997). Brand loyalty metrics concern the **brand repurchase rate**, which is the percentage of customers purchasing a specific brand who will repurchase the brand on the next purchase occasion as well. It is important to note that a no-purchase in this next purchase occasion can be followed by a repurchase of the brand in a subsequent period. Using this information, a so-called "switching matrix" can be formed (see Figure 3.9). In this matrix, one observes the repurchase rates of a brand and the switching probabilities to other brands in the market. In this specific example, the switching probability from brand A to B is 20%, whereas the probability of customers switching back to brand A in the subsequent purchase occasion is 10%. Notably, this switching matrix can not only be based on actual repurchase rates but can also use the constant sum scale of the division of 100 points across brands (Rust, Lemon & Zeithaml, 2004).



Brand	*A (%)*	*B (%)*	*C (%)*
A	**70%**	20%	10%
B	10%	**80%**	10%
C	0%	20%	**80%**

FIGURE 3.9 Example of a brand switching matrix

Within marketing science, the relationship between market share and brand repurchase rates has been studied, and the general finding is that there is a strong relationship between them. In particular, the school around former London-based marketing professor Andrew Ehrenberg, with followers such as Byron Sharp, has aimed to demonstrate that this empirical relationship can be shown in many markets. One of the most important implications of this is that brands with a high market share have a high repurchase rate and that brands with low market shares have a low repurchase rate. There are some exceptions, such as a niche brand targeted at a specific market segment, with loyal customers. One of the conclusions Ehrenberg and colleagues draw is that firms should not believe too strongly in creating loyal customers (Sharp 2010). They are therefore very critical of loyalty strategies, such as loyalty programs (Dowling & Uncles, 1997; Sharp, 2010). However, although one could draw this conclusion, these analyses only provide a current status quo. There is sufficient evidence that loyalty strategies can create a higher repurchase rate for brands (e.g., Leenheer *et al.*, 2007).

As well as looking at the revenue side of brands, brand investment can also be measured. These brand investments could involve, for example, advertising costs. The results of these advertising costs are measured with so-called "gross rating points" (GRPs). This is calculated as the percentage of the target market reached multiplied by the exposure frequency of the advertisement.

3.8.4 Brand evaluation metrics

As brands are very important assets for firms, there has been strong attention on how to financially evaluate them; in particular, financially oriented brand equity (BE) metrics have been developed. These metrics differ from the V2C brand equity metrics that typically focus on customer-based BE and focus only on awareness and attitudinal measures, such as brand preference. Probably the best-known financially oriented BE metric is the one developed by Interbrand. Each year they develop this to calculate the BE of global brands such as Google, Amazon, Coca-Cola, Apple, and BMW and publish a top 100. The **Interbrand** measure[5] is based on three pillars: financial performance of the brand, role of the brand in purchase decisions, and brand strength, which is the ability of the brand to create brand loyalty (www.interbrand.com). The exact methodology of this metric is not fully shared and therefore represents a black box for many researchers. Beyond these well-known metrics, academic researchers have also proposed other metrics to evaluate BE, focusing on shareholder value, brand preference measurement, and the price premium paid.

Interestingly, studies report strong correlations of brand equity metrics with product-market metrics. For example, Ailawadi *et al*. (2004) show that their revenue premium measure focusing on the additional revenue a brand can achieve above a product without a brand (typically a private label) (see Figure 3.10), is relatively strongly correlated with measures such as market share. One could probably debate the additional value of brand equity metrics above frequently used product market metrics. One problem with these metrics is that they are frequently short-term oriented, while BE is more long-term. Financial BE metrics should be diagnostically about a brand's long-term health; this goes beyond sales and could take in factors such as the attraction power of brands, price premiums, and financial value.

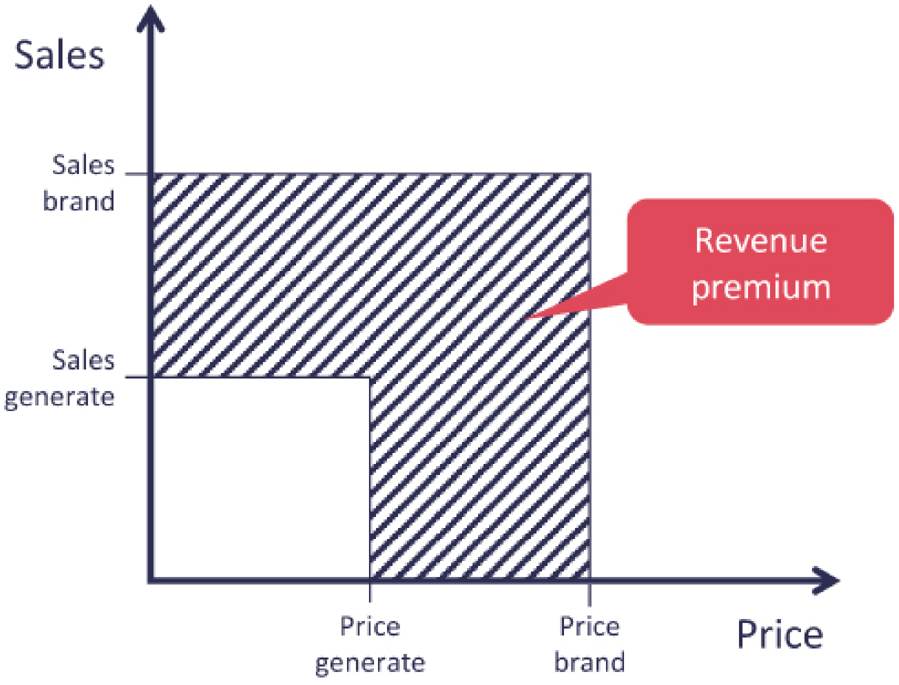


FIGURE 3.10 Brand Revenue Premium

Source: Adapted from Ailawadi, Lehmann & Neslin, 2004

3.9 CUSTOMER METRICS

Customer metrics focus on the behavior of individual customers in their relationship with the firm. We adopt the so-called "relationship lifecycle" as an underlying model of many of these metrics. In this relationship lifecycle, the trajectory of an individual customer moving from being a prospective to an actual customer, and finally, a defecting customer is considered (see Figure 3.11). Based on the relationship lifecycle concept one can distinguish three types of metrics:

1. Customer acquisition metrics, which focus on the first phase of attracting customers
2. Customer development metrics, which focus on how the relationship develops after acquiring the customers
3. Customer value metrics, which consider the financial value of the customer during the relationship and can involve both the acquisition phase and relationship development phase.

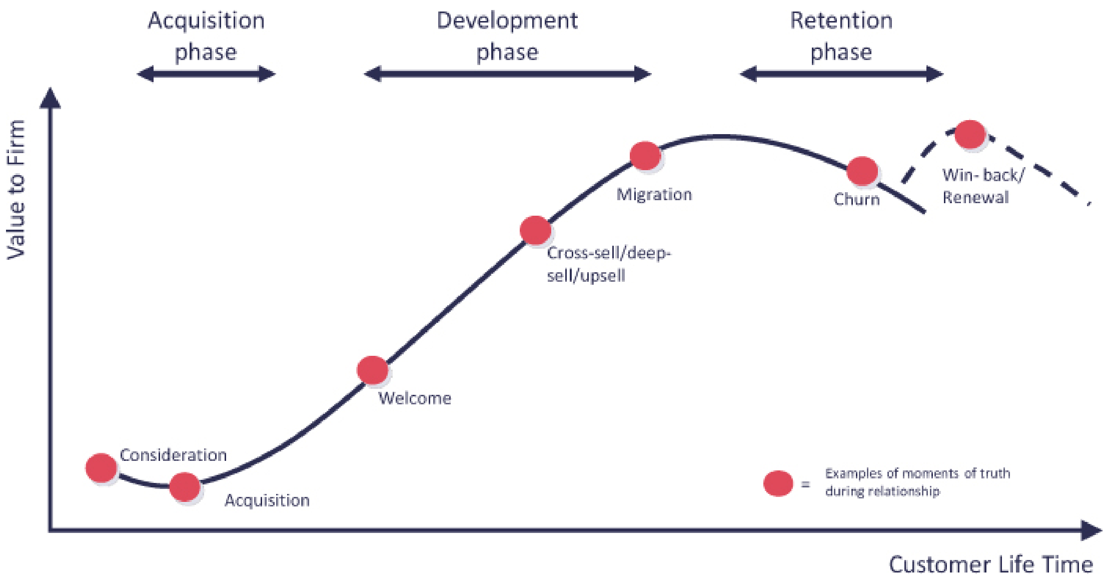


FIGURE 3.11 Relationship lifecycle concept

3.9.1 Customer acquisition metrics

Customer acquisition metrics consider how customers respond to acquisition actions. From a customer perspective, we mainly consider individual customer acquisition techniques, such as direct marketing efforts, telemarketing etc. On a more aggregate level, firms will be interested in the **number of customers acquired** through these different channels. From a cost perspective, they would like to know the **acquisition costs (per customer)**. Importantly, we also suggest considering the value of the acquired customers. Firms constantly fall into the trap of acquiring many customers with attractive price offers who are likely to be unprofitable and will switch frequently after one year (Lewis, 2006). When evaluating acquisition campaigns two specific metrics are frequently considered:

- Response rate: the number of customers responding to a campaign divided by the number of total customers approached by a campaign
- Conversion rate: the number of customers acquired divided by the number of customers responding to a campaign.

3.9.2 Customer development metrics

Customer development metrics concern multiple dimensions of relationship development (e.g., Bolton, Lemon & Verhoef, 2004). Specifically, we consider three types of metrics:

1. Relationship continuation or length metrics
2. Relationship expansion metrics
3. Relationship costs and risk metrics.

3.9.2.1 Relationship length metrics

These metrics focus on the continuation of the relationship. Specific metrics concern:

– Churn or customer defection: The percentage of customers quitting the relationship with the firm
– Customer retention: The percentage of customers continuing the relationship with the firm (1—customer churn/defection)
– Customer lifetime/relationship length: The (expected) length of the relationship between the customer and the firm
Purchase Frequency: The number of times a customer purchases from a company in a specific time period
– Recency: Time since last purchase.

One specific metric you could consider here is the win-back percentage of defected customers. Win-back metrics become prevalent when firms actively approach churned customers to become customers again.

3.9.2.2 Relationship expansion metrics

Relationship expansion concerns metrics that focus on the growth of the relationship. This expansion can involve multiple metrics:

– Average number of products/services sold per customer
– Cross-buying rate: The percentage of customers purchasing additional products or services from a company
– Upgrading rate: The percentage of customers upgrading their products or services to a higher level (i.e., upgraded service contract)
– Adoption rate: The percentage of customers purchasing the newly introduced product or service
– Customer share: The number of products or services purchased by a customer from a firm divided by the number of products or services purchased by a customer in that specific product category

– Share of wallet: The money spent by a customer at a firm divided by the money spent by a customer in that specific product category.

The first four measures can be gained from the customer database. The share metrics provide information on the relative loyalty position of a firm for a specific customer. In order to calculate these metrics, additional data on the purchase behavior of customers in a specific category is required (Verhoef, 2003). Higher cross-buying rates should usually also be reflected in higher values for the share metrics. Importantly, recent research suggests that cross-buying can have positive profit consequences, but not all customers with large cross-buying percentages are profitable (Shah *et al.*, 2012).

3.9.2.3 Relationship costs and risk metrics

During the relationship, firms also invest in customer relationship. Investments focus on developing the relationship and can, for example, include the costs of a loyalty program. We discuss these costs in more depth in the next section on CLV. Costs mainly involve the **costs to serve** an individual. Importantly, these costs may vary strongly between customers. One specific risk that firms frequently face is that customers might not pay. In this regard measuring the **debt risk** of the customer base and individual customers is of importance, especially for firms delivering products or services before payment is received (e.g., utilities, telecom). Debt risk can be modeled and predicted (L'Hoest-Snoeck, Van Nierop & Verhoef, 2015). For retailers, **the number of product returns** has also become an important cost-related metric. Returns occur when customers order a product that does not deliver what they hoped for. Frequently, customers can return these products for free and firms are obliged to repay the price paid. Reducing return rates has become one of the top priorities for online retailers, such as Zalando, in order to improve profitability. Overall, Shah *et al.* (2012) suggest that firms should consider metrics that reveal the so-called "adverse behavior" of customers, which may involve debt risk, large service costs, and product returns. Interestingly, customers show a pattern in these adverse behaviors, and thus in targeting policies firms can choose not to make these customers attractive offers (Shah *et al.*, 2012).

3.9.3 Customer value metrics

The past metrics focus on behavior during the relationship lifecycle. Customer value metrics consider the resulting financial value of customers. Value metrics can be categorized based on the forward-looking nature of these metrics. Non-forward-looking value metrics focus on the current status of a customer and involve **customer revenue or the monetary value of a customer, customer margin, and customer profitability**. Forward-looking metrics consider the expected value of customers in the future. The most prominent metric in this regard is **Customer Lifetime Value**. This metric has gained so much attention in the literature—and has been accepted in practice—that we devote an extensive discussion to this metric in this chapter.

3.9.3.1 Customer Lifetime Value

Customer lifetime value (CLV) is a metric that has taken off with the strong development of customer relationship management. This metric is typically used to evaluate the value of customers. CLV is frequently defined as "*the present value of the future cash flows attributed to the customer during his/her entire relationship with the company*" (Farris *et al.*, 2010). In this definition, CLV assesses the total value delivered by a customer. However, firms are frequently more interested in the future value of customers. CLV can then be considered as a forward-looking customer-centric metric, based on assumptions and predictions, and can be defined as the net present expected value of a customer. In other words, the value created in the past is not taken into account when calculating the value of a customer or a customer base. However, past customer value can be a predictor of future value (e.g., Donkers, Verhoef & De Jong, 2007). CLV is a very important metric, because (when calculated properly) it can be the link between the marketing and the financial department and create the common platform to bring these worlds together. Furthermore, there is sufficient evidence to show that, when calculated properly, CLV can be a good indicator for firm valuation (e.g., Gupta, Lehmann & Stuart, 2004), especially in those industries/companies where customers are the biggest asset and markets are rather stable. A very simple definition of CLV for customer i at time t=0 and with 'd' as the discount rate is:

$$CLV_{i,t} = \sum_{t=0}^{T} \frac{Margin_{i,t}}{(1+d)^{t}}$$

In this definition, we assume that each year a specific margin is earned per customer. These margins are summated over time until a chosen endpoint T. Typically, periods of 3–5 years are used for CLV-calculation (Rust, Zeithaml & Lemon, 2000). A discount rate is used to make the future earnings present. This discount rate is set in cooperation with the finance department. A high discount rate implies that future earnings contribute less to CLV and may signal that firms value these future earnings less. This implies a more short-term orientation.

3.9.3.2 Calculating CLV

For calculating the actual CLV, based on our CLV model and its components, we will need a formula that incorporates all elements so that we can calculate a CLV per individual customer that can be expressed in a euro/dollar amount. At a minimum level, firms need to have data on customer margins and expected lifetime. Beyond that, especially when calculating the value of a new customer, data on investments in acquisition or retention should be taken into account. The extensive literature on CLV has proposed several more extended CLV models (e.g., Berger & Nasr, 1999; L'Hoest-Snoeck, Van Nierop & Verhoef, 2015; Venkatesan & Kumar, 2004). We specifically discuss a more extensive version of the simple equation in which we include the retention rate (r) to account for the expected lifetime of individual customers, thereby taking into account: that earnings from customers may drop because customers churn, and that firms invest in acquisition.

$$CLV_{i,t} = -\text{Acquistion/Retention Investment}_i + \sum_{t=0}^{T} \frac{(r_{i,t})^t {*}\text{Margin}_{i,t}}{(1+d)^t}$$

An easy way to calculate CLV is to take T to infinity. After some mathematical computations, the following simple formula is achieved for the calculation of the CLV for customer i:

$$CLV_i = -\text{Acquisition/Retention Investment}_i + \text{margin}_i \left(\frac{r_i}{1+d-r_i} \right)$$

Each of the above components models can be used to forecast the individual retention rates and margins. For example, retention rates can be predicted using a logistic regression model. More ambitious versions of the above models can be developed; one way might be to consider margin growth through, for example, reduced costs or increased revenues from cross-buying or upgrading.

3.9.3.3 Getting started with CLV: Be pragmatic

In building a CLV model, one should realize that it is quite impossible to build a full-blown CLV model from scratch. In reality, a phased approach is much more realistic and can help in processing new insights on the dynamics of the CLV model. Defining the different phases in a pragmatic way on a project basis can help the organization in building experience with CLV at an early stage and diminish the probability of failure. Starting with a model that is too complex can result in disappointment, losing organizational acceptance, wrong results, and delays in delivery. It is good to know that there are many pragmatic approaches and rules-of-thumb available to realize a first CLV model for quick results.

Let's start with the CLV formula:

$$CLV_i = -\text{Acquisition/Retention Investment}_i + \text{margin}_i \left(\frac{r_i}{1+d-r_i} \right)$$

- The first component, the (monthly) margin (revenues minus costs) could be simplified by just taking a percentage of the revenue. This could be further refined by specifying this margin percentage per product, assuming that per product a cost allocation model delivers this percentage, and differentiating this per product. The next step in margin calculation could be to take the largest cost buckets and try to allocate them based on consumer behavior.
- A good proxy of the lifetime of the customer, based on the retention rate (r) in the formula above for a contractual setting, can be obtained by assuming that the lifetime equals the duration of the first contract. In a non-contractual setting, one could take the buying frequency per year (or another timeframe) as an indicator for lifetime. It might be clear that the above proxies are still a rather rough proxy of the real lifetime, with only a little differentiation per customer. So, the next step could be to build a first basic churn prediction model to allocate customers into churn buckets and assigning a churn (in a contractual setting) or inactivity (in a non-contractual) setting probability per segment.
- The easiest way to estimate the investments in acquisition or retention is to take the total investments and divide this by the total number of transactions for new or renewing customers. This could be refined by splitting these total costs in acquisition and retention investment and dividing this by either the number of acquisitions or the number of retentions. The next step could be specifying the investments per channel and/or product.

To show how CLV can be calculated we present a CLV case for a large energy company (see Case 3.1), that tried to model CLV over the customer base to assess the importance of every element of the CLV formula. We also suggest taking a look at the numerous calculators that can be used as a guide in building your own CLV model. Specifically, we would like to mention the calculator developed by Harvard Business School.[6] Even the basic version is a very nice example of how to calculate CLV.

CASE 3.1: CASE CLV AT ENERGY COMPANY[7]

Situation

The Dutch energy market has been liberalized and customers can now switch between energy companies. This has resulted in voluntary churn rates moving from 0% to much higher levels. New providers with a price focus have entered the market. As a consequence, the number of customers of the energy company has decreased and revenues are under pressure. The company realized that competition has changed and that they need to compete for customers and customer value.

Complication

Although churn seems to be the problem, that may not be the case. Customer value at energy companies is not only driven by churn but also involves energy and other product usage, service costs, payment issues etc. The question is which value-drivers they should try to influence and how they should do so. So far, the value components have not been well understood. This complication requires a strong conceptualization of the value drivers and gaining the right data.

Key message

The energy firms should not only look at revenues and retention. When managing customers and optimizing customer value, they should also thoroughly consider how they can reduce inbound service costs and payment enforcement costs.

Data and model used

The CRM database of the firm was relatively rich and embedded the information on product possession, churn, margins, payment method, payment problems etc. We were able to analyze data from 0.9 million customers. Before setting up the econometric model, we created a conceptual model on drivers of customer lifetime value. Based on this, logit models were estimated explaining each of the drivers and predicting the occurrence of an event. These predictions were used to predict the CLV. Next, the outcomes of the models were used to understand how specific drivers can be influenced and which specific actions could increase CLV.

Results

Conceptually, five drivers of CLV were identified (see Figure 3.12):

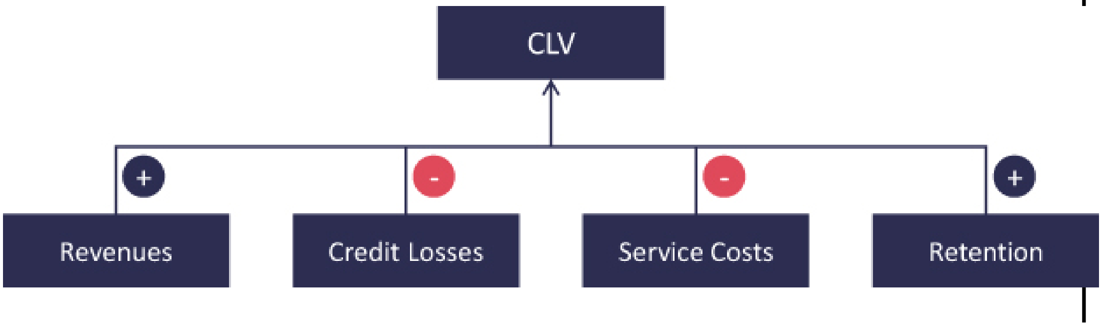


FIGURE 3.12 Value drivers for an energy company

– Retention (1—churn)
– Revenues (electricity and gas)
– Credit losses (not paying)
– Service costs (including payment enforcement costs and inbound calls).

Predictions were used to understand the contribution of each of the components and this showed that revenues and service costs are the most important value drivers in terms of their contribution to CLV. Retention makes only a limited contribution (see Figure 3.13), whereas revenues per customer, bad debt (including payment enforcement), and service cost strongly contributed to CLV.

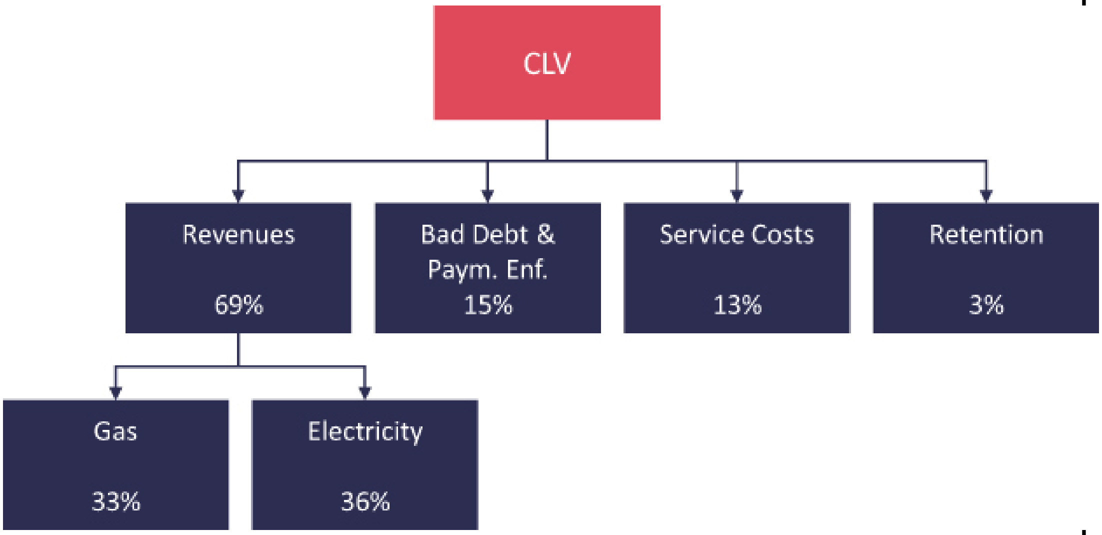


FIGURE 3.13 Contribution of each of the value drivers to CLV

The next question asked was whether one could influence the drivers, i.e. through specific actions, with an idea of their potential success. An immediate reaction is that one would probably suggest selling more energy. However, energy usage is largely defined by factors such as household size and the weather. Moreover, from a sustainability perspective, energy firms are implementing measures to lower energy usage, for example through promoting energy-saving light bulbs.

To define success probability, the researcher talked with several marketing employees to understand which actions would be more likely to be successful. This resulted in a qualitative assessment of success probability. Next, the numbers of customers that could be affected and which value-gain can be achieved were assessed. Using the simulation results, it was found that lowering service contacts and stimulating a lower level of payment enforcement in particular had the largest potential value impact (see Figure 3.14).

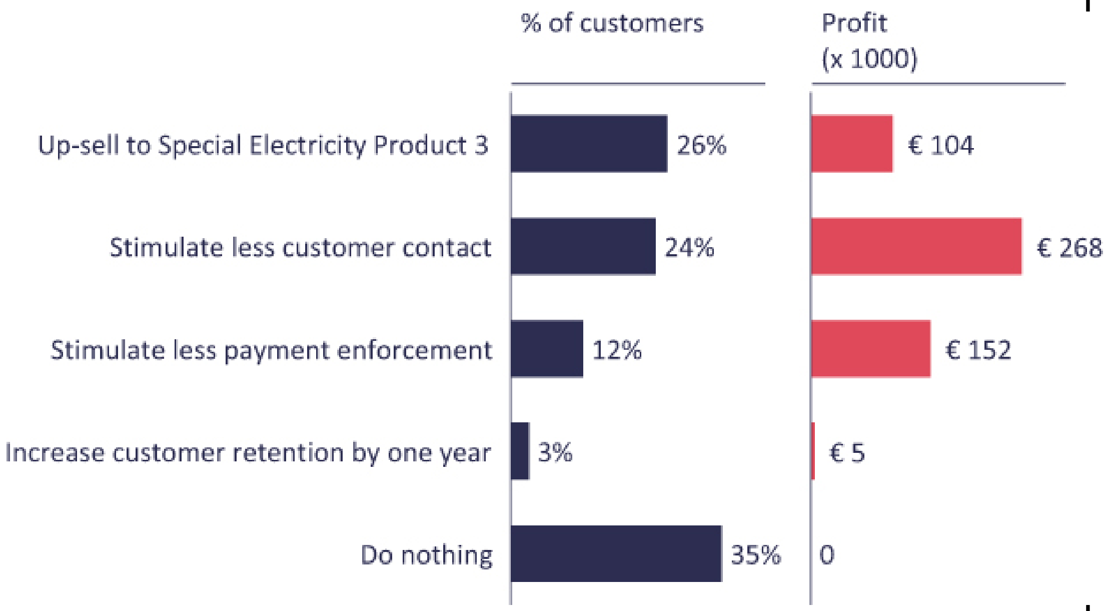


FIGURE 3.14 Impact of different value driver improvements on CLV

Additional insights

The analysis also showed that the value drivers differ significantly per value-segment. Service costs and bad debt become especially important in the low-value segment. These customers could become more valuable if service costs were lower.

Success factors

The model results helped the company base its marketing actions on customer value drivers. In summary:

– The analysis was executed in-house by a senior data analyst, who was very well aware of scientific studies on CLV and was able to apply sophisticated models.
– The simulation of the model results involved both actual model results and qualitative assessments of success.
– There was a rich CRM database available: all the required data were available in that database and it was not necessary to collect data from multiple data silos. As such we were able to analyze a very large sample of customers.

3.9.4 Customer equity

Customer Equity is closely related to CLV. The metric, as successfully proposed by Rust, Zeithaml and Lemon (2000), is the summation of all CLVs of current and future customers of a firm. As such it is broader than CLV because it considers all customers and also future customers. Hence, it considers both the value of existing relationships resulting from customer loyalty and the ability of the firm to attract new valuable customers. To achieve this, Rust *et al*. use a switching matrix approach, in which customers can switch between suppliers. In subsequent work, Rust, Lemon and Zeithaml (2004) showed that firms can calculate the consequences of investments in drivers that increase the value delivered to customers (reflected in customer perceptions on customer equity) by considering acquisition and retention consequences and subsequent effects on CLV and customer equity. By comparing the investments in value creation (e.g., increasing leg space in airplanes) with the customer equity changes, a marketing ROI can be calculated (see Figure 3.15).

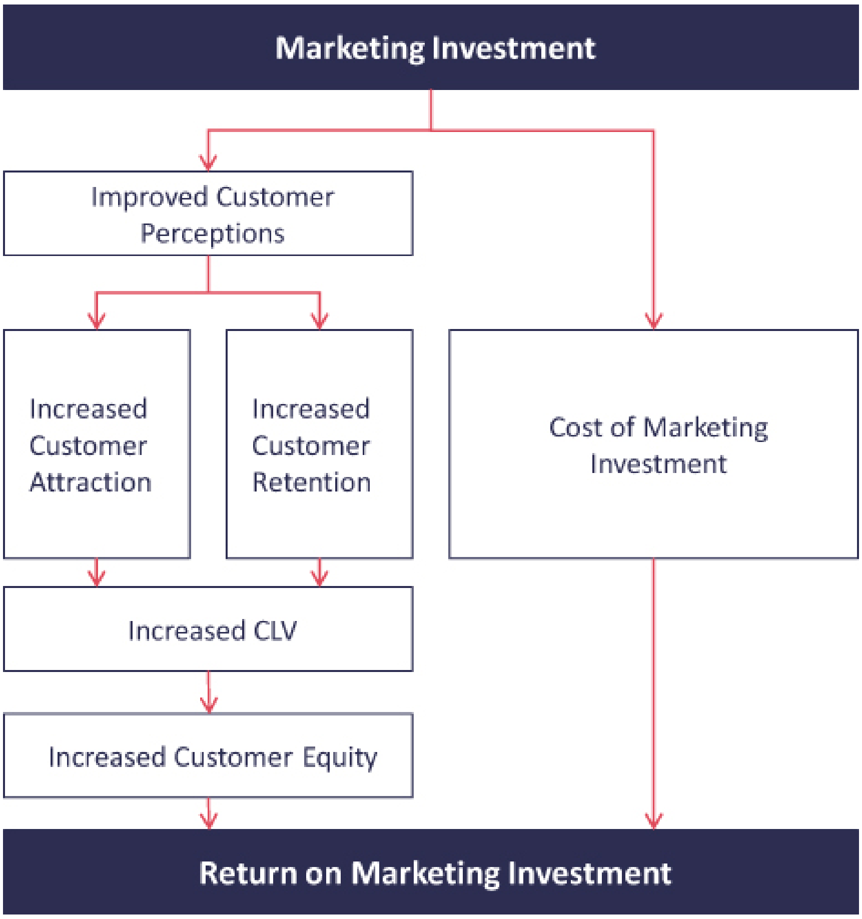

Marketing Investment
Improved Customer Perceptions
Increased Customer Attraction
Increased Customer Retention
Cost of Marketing Investment
Increased CLV
Increased Customer Equity
Return on Marketing Investment

FIGURE 3.15 Customer equity ROI model

Source: Adapted from: Rust, Lemon and & Zeithaml, 2004

3.9.5 New big data metrics

As a result of the development of big data, we can consider two important new areas of development that require additional metrics:

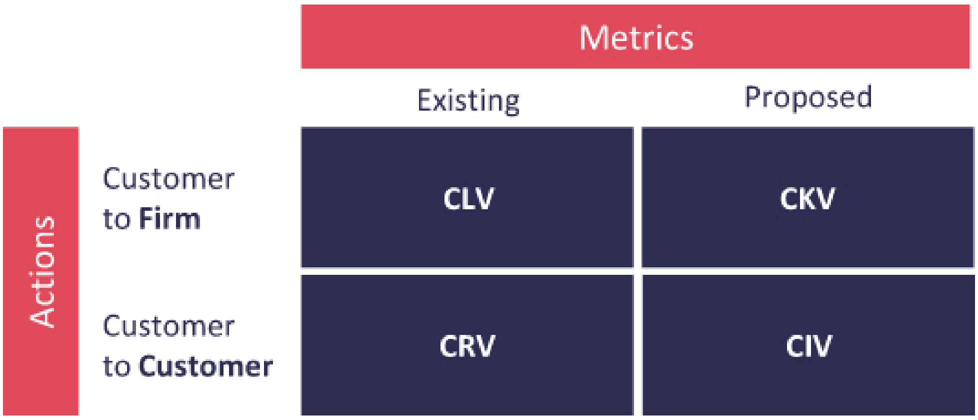
- Customer engagement
- Customer journey.

3.9.5.1 Customer engagement

The increasing presence of social media has provoked attention to the non-transactional behavior of customers. Customers not only add value with their purchase behavior, but may also add value by sharing their experiences online, influencing other customers, and providing input through co-creation. This non-transactional behavior is frequently referred to as "customer engagement behavior" (van Doorn *et al*., 2010). This behavior results in a possible need for additional metrics, such as:

- The number of referrals per customer
- The number of relationships of a customer with other customers
- The number of ideas (e.g., new products, service improvements) of customers provided to a firm
- The influential power of customers (i.e., measured by opinion leadership or social network variables; see for example Risselada, Verhoef & Bijmolt, 2015).

One problem with these metrics is that they are frequently difficult to measure as they may involve social network information and/ or self-reports on influence. We will reflect on this in more depth in the chapter on analytics, where we discuss social network analytics. Firms have, however, been able to collect data in their databases on referrals and potentially on a number of ideas (e.g., through measuring complaints). The importance of customer engagement also implies an extension of the CLV concept. Specifically, Kumar *et al*. (2010) introduced the concept of "customer engagement value." Within this concept, three new additional value components are introduced: customer referral value (CRV), customer influence value (CIV), and customer knowledge value (CKV) (see Figure 3.16). Kumar *et al*. distinguish between customer-to-customer (C2C) and customer-to-firm (C2F) values. CRV and CIV are C2C values, whereas CLV and CKV are C2F values. CRV has been operationalized and measured using actual referral behavior in work by Kumar and colleagues (2010, 2013). Interestingly, they show that customers with a medium CLV have the highest CRV. CIV is more difficult to measure because of the need for network data. Kumar *et al*. (2013) measured CIV for the social media campaign of an Indian ice-cream retailer, Hokey Pokey. They calculated both CLV and CIV and summing these metrics they calculated the ROI of the social media campaign.



		Metrics	
		Existing	Proposed
Actions	Customer to Firm	CLV	CKV
	Customer to Customer	CRV	CIV

FIGURE 3.16 Customer Engagement Value: extending CLV

Source: Adapted from Kumar *et al*., 2010

3.9.5.2 Customer journey metrics

The digital revolution has led to a new omnichannel environment (e.g., Verhoef, Kannan & Inman, 2015). Customers are now developing their own path to purchase. They browse and search online, switch between offline and online channels, use multiple devices, etc. and are still being influenced by traditional advertising. In sum, different customers face brands in different touchpoints that affect customers in different ways (e.g., Baxendale, Macdonald & Wilson, 2015). This new development also results in a mix of new brand and customer metrics. At the brand level, we may, for example, observe the number of click-throughs on a banner ad. For customers, firms may observe conversion rates. However, for prospective customers these metrics are less easy to measure. For firms it is still important to understand this path to purchases and to measure specific outcomes during this path to purchase for both customers and prospective customers. (e.g., Li & Kannan, 2014; Verhoef, Kannan & Inman, 2015). This results in different new digital metrics:

– Number of website visits: When paying other websites to show your advertisement and/or banner the associated cost is measured against the costs per 1,000 views/eyeballs (CPM= cost per mile)
– Click through rates: The percentage of customers viewing an online ad, search engine outcome, etc., clicking on the ad to visit the referred website; the financial metric to represent the associated costs when, for example, showing an online advertisement via Google— cost per click (CPC)
– Purchase conversion rate: The number of purchases after a website visit, divided by the number of website visits. The financial metric to show the costs of advertisements or online activities is cost per order/transaction (CPO)
– Average order size: The average order size of each purchase
– Costs of each unique touchpoint

– Channel switching: The migration patterns of a customer to other channels (especially relevant, when firms aim to migrate customers to a low-costs channel (e.g., Gensler, Verhoef & Böhm, 2012; Trampe, Konuş & Verhoef, 2014)

– Research shopping percentage: Percentage of customers searching in one channel and purchasing in another channel (e.g., Verhoef, Neslin & Vroomen, 2007). One specific form of this is the showrooming percentage, where the search channel is the store and the purchase channel is online (Rapp *et al*., 2015).

One specific concern managers face is linking investments in different (digital) acquisition channels to purchase outcomes. This is not so trivial, as firms may be influenced by multiple channels or touchpoints in their purchase decision and specific channels (e.g., search engines) are by definition closer to the purchase decision than other channels (e.g., advertising). Firms require strong attribution models to quantify the contribution of every channel.

3.10 MARKETING ROI

One final metric that firms are interested in is the ROI of marketing investments. We have already briefly mentioned this in our discussion of the customer equity metric since ROI is directly related to current and future CLV. The ROI on marketing investments is calculated as the additional CLV divided by the marketing investments. Currently, marketing ROI is most of the time only very narrowly measured, measuring direct effects on sales of measurable efforts that can be directly attributed to these sales.

Ideally, marketing ROI should cover all possible marketing activities that can be deployed and not just the (direct) marketing activities addressing new and existing customers. Furthermore, the measure to calculate ROI, should not be based on sales volume, but ideally on created CLV, and it should also take into account the extent to which marketing activities contribute to V2C metrics (see Figure 3.17). In fact, a marketing ROI calculation should also encompass, for example, investments made in ATL campaigns, specific brand activities, and other marketing mix elements. But few cases successfully show marketing ROI as a holistic (i.e. analyzing CLV and V2C effects of the whole marketing mix) approach.

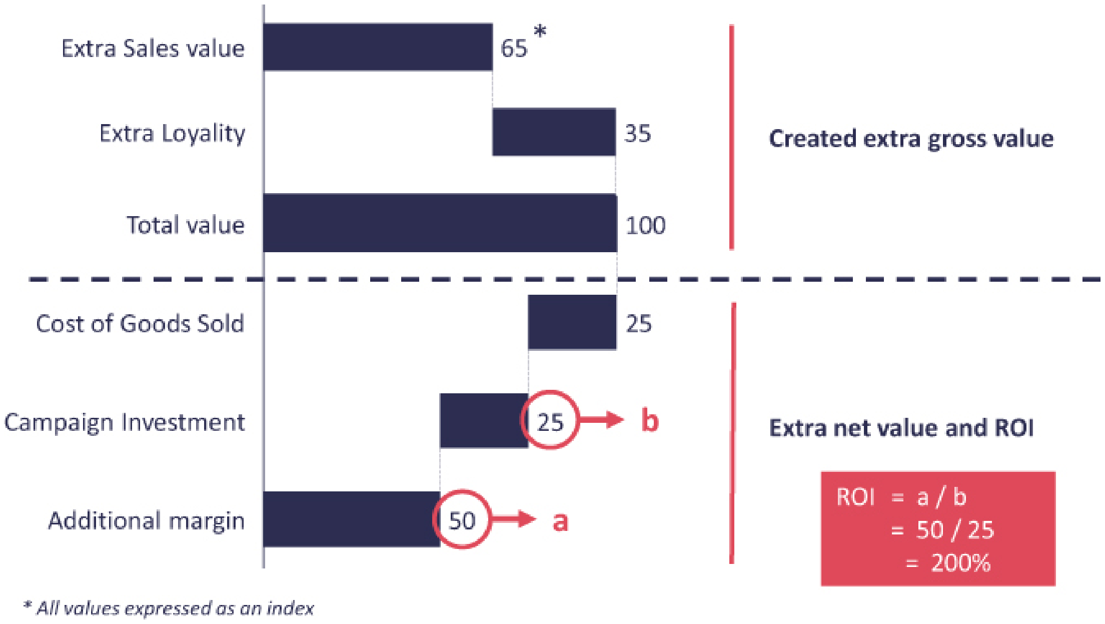


FIGURE 3.17 Example of ROI calculation

There are two major reasons why this is not common practice:

- The relationship between, for example, an ATL campaign and the extra CLV or other V2F metrics is difficult to assess
- Large scale complexity of integrating all data sources (see Chapter 5)

3.11 CONCLUSIONS

In this chapter, we discussed V2C and V2F metrics at the market, brand, and customer levels. It is clear that there is a plethora of metrics from which firms can choose. We aimed to provide an overview of frequently used metrics but clearly acknowledge that one can easily come up with many other relevant (or irrelevant) metrics. For very interested readers we refer to Farris *et al*. (2006), who provide a very extensive discussion of multiple metrics for multiple areas of marketing. Importantly, we also discussed some new big data metrics. Specifically, we focused on, for example, customer engagement metrics and customer journey metrics, as we believe that major developments are occurring here.

ASSIGNMENT 3.1: CLV HEALTH INSURANCE COMPANY

A health insurance company evaluates its recent acquisition campaign. These campaigns are mainly carried out in November and December so consumers can switch at the beginning of each subsequent calendar year. Thus, in the preceding period many insurers do a lot of advertising. These advertising campaigns are, however, quite costly and raise the question of whether they are profitable. The ROI on these campaigns could potentially be negative given the high costs. To calculate this ROI, the insurer wants to know the CLV of a newly recruited customer.

The average monthly profit margin on a customer's health insurance is approximately five euros. The retention rate in year 1 is 75%. After that, the retention rate goes up and in year 2 it is 90%. The insurer applies a discount rate of 9%. The costs of acquisition must be recouped in four years. Consumers pay their insurance premium per month.

Questions:

1. Calculate the new customer's CLV. Clearly indicate which assumptions you make and which formulae you use.
2. The health insurer is considering increasing the discount rate to 13%. What could be an underlying reason for increasing the discount rate? How much does the value of a new customer's CLV increase or decrease?
3. An average of 300 euros is spent to acquire a new customer. What is the ROI on the acquisition-campaign?
4. By how much would the ROI increase if the health insurer is able to acquire customers who would bring in six euros per month and the retention rate in year 1 rises to 80%?
5. The health insurer finds the outcome of the CLV calculation to be quite low. According to an analyst, other customer value components should be taken into account. Which customer value components should be considered?

ASSIGNMENT 3.2: METRICS DUTCH SUPERMARKETS

Background

Albert Heijn has been the market leader for many years in the Netherlands, holding over 30% of the market share. Competition for Albert Heijn is, however, increasing. The family company Jumbo, for example, formulated a clear growth strategy. This growth strategy is realized on the one hand through autonomous growth, which means that existing stores attract more customers, and on the other hand through the acquisition of other supermarket chains, such as the Super de Boer stores.

Besides the competition, Albert Heijn also experiences a lot of pressure from discounters and cheaper supermarkets, including Aldi, Lidl, Plus, C1000, and Dirk. Supermarket Plus mainly competes on service. Lidl is increasingly positioning itself as a discounter with high-quality products. For several years now they have won the prize for supermarket with the best fruit and vegetable product range. Especially in times of economic downturn, these chains are more popular with Dutch consumers.

In 2012 the Netherlands was still in the middle of a recession and Albert Heijn was faced with the challenge of maintaining its market leadership position. The main question was: how should they do that? Should they focus on increasing revenue per customer or increase the number of loyal customers? What should they change for customers? For example, should they improve their price-value reputation? It is known that Dutch consumers often find Albert Heijn to be expensive.

In 2011 and 2012, a major study investigated how supermarkets, among others, score on V2C and V2F components. Customers were questioned about their opinion on the relationship with their supermarket, the perceived value proposition, the attractiveness of the price, and the perceived brand value on a seven-point scale (1 = low, 7 = high). For the price variable, the measurement changed between 2011 and 2012. In 2011 customers were asked to assess the price (1 = high, 7 = low), while in 2012 customers were asked to assess the competitiveness of the price (1 = not competitive, 7 = very competitive). In addition, loyalty (0–100%) and revenue per customer (expressed on a scale of 0–100) were measured (see Table 3.6 for scores).

TABLE 3.6 Scores of various measured V2C and V2F metrics of supermarket chains

Measured Variables	***Year***	***Supermarket Chains***						
		Albert Heijn	***Jumbo***	***Lidl***	***Aldi***	***C1000***	***Plus***	***Dirk***
Relationship quality	2011	3,8	3,5	3,6	3,4	3,6	3,5	3,7
	2012	3,8	3,8	3,4	3,2	3,4	3,3	3,6
Price-value	2011	5,5	5,5	5,6	5,2	5,1	5,1	5,3
	2012	5,6	5,6	5,4	5,2	5,2	5,0	5,4
Price	2011	3,2	3,9	4,3	4,2	3,6	3,2	3,9
	2012	4,2	5,2	5,3	4,9	4,4	3,7	5,0
Brand	2011	5,3	5,1	4,8	4,4	4,7	4,5	4,6
	2012	5,5	5,3	4,6	4,2	4,6	4,4	4,6
Loyalty	2011	52,5	32,9	28,9	27,9	29,1	28,7	36,1
	2012	43,3	33,0	25,9	23,7	29,9	26,1	31,0
Revenue	2011	43,5	38,9	36,9	33,7	39,2	36,8	44,2
	2012	40,0	41,9	33,7	32,6	36,6	34,6	34,6

With regression analysis, researchers investigated the importance of different V2C metrics for explaining loyalty. This provided insight into how to influence loyalty. The standardized coefficients for 2011 and 2012 are as follows.

TABLE 3.7 Influence of standardized regression coefficients of V2C metrics on loyalty

Measured Variable	2011	2012
Relationship quality	0,23	0,17
Price-value	0,33	0,28
Price	0,02	0,08
Brand	0,21	0,08

Questions:

1. What are the main changes in the V2C and V2F for the different supermarkets when comparing 2012 with 2011?
2. Jumbo is an important competitor for Albert Heijn. How has Albert Heijn's position compared to Jumbo changed from 2011 to 2012? What does that mean for Albert Heijn's competitive position?
3. To what extent do you see Lidl in 2012 as an increasingly important competitor to Albert Heijn based on the numbers presented? What other data would you like to collect in order to make a statement about this?
4. Which factors will Albert Heijn need to improve in order to increase loyalty again?
5. What marketing advice would you give Albert Heijn in 2012 to maintain and strengthen their market leadership based on a thorough study of the figures presented? Note that Albert Heijn suspects that Jumbo also wants to acquire C1000 and that Lidl only wants to profile itself more in terms of service.

NOTES

1. This part of the text is derived from De Haan, Verhoef and Wiesel (2015).
2. This section is based on Ou *et al*. (2014).
3. This section is partially based on van Onrust *et al*. (2014).
4. We refer to Farris *et al*. (2006) for a detailed discussion on some of these metrics and how they can be used to calculate sales volumes.
5. Other agencies have also developed brand equity metrics. The INTERBRAND metric is, however, most known and influential, and we have therefore included this metric in this chapter and ignore some other metrics of commercial agencies.
6. See http://hbswk.hbs.edu/archive/1436.html.
7. This case is based on L'Hoest-Snoeck, Van Nierop and Verhoef (2015), for more details we refer to that study. We thank Sietske L'Hoest-Snoeck for

sharing insights and some internally used pictures.

REFERENCES

Ailawadi, K. L., Lehmann, D. R., & Neslin, S. A. (2004). Revenue premium as an outcome measure of brand equity. *Journal of Marketing*, *67*(4), 1–17.

Ailawadi, K. L., Neslin, S. A., & Lehmann, D. R. (2003). Revenue premium as an outcome measure of brand equity. *Journal of Marketing*, *67*(4), 1–17.

Antonides, G., Verhoef, P. C., & De Hoog, A. N. (2004). Service encounters as a sequence of events the importance of peak experiences. *Journal of Service Research*, *7*(1), 53–64.

Arts, J., Frambach, R. T., & Bijmolt, T. H. A. (2011). Generalizations on consumer innovation adoption: A meta-analysis on drivers of intention and behavior. *International Journal of Research in Marketing*, *28*(2), 134–144.

Baxendale, S., Macdonald, E. K., & Wilson, H. N. (2015). The impact of different touchpoints on brand consideration. *Journal of Retailing*, *91*(2), 235–253.

Beard, A., Hornik, R., Wang, H., Ennes, M., Rush, E., & Presnal, S. (2011). It's hard to be good. *Harvard Business Review*, *89*(11), 88–96.

Berger, P. D., & Nasr, N. I. (1999). Customer lifetime value: Marketing models and applications. *Journal of Interactive Marketing*, *12*(1), 17–30.

Bolton, R. N., Lemon, K. N., & Verhoef, P. C. (2004). The theoretical underpinnings of customer asset management: A framework and propositions for future research. *Journal of the Academy of Marketing Science*, *32*(3), 271–292.

Bolton, R. N., Lemon, K. N., & Verhoef, P. C. (2008). Expanding business-to-business customer relationships: Modeling the customer's upgrade decision. *Journal of Marketing*, *72*(1), 46–64.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–340.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, *35*(8), 982–1003.

de Haan, E., Verhoef, P. C., & Wiesel, T. (2015). The predictive ability of different customer feedback metrics for retention. *International Journal of Research in Marketing*, *32*(2), 195–206.

De Vries, L., Gensler, S., & Leeflang, P. S. H. (2012). Popularity of brand posts on brand fan pages: An investigation of the effects of social media marketing. *Journal of Interactive Marketing*, *26*(2), 83–91.

Dixon, M., Freeman, K., & Toman, N. (2010). Stop trying to delight your customers. *Harvard Business Review*, *88*(7/8), 116–122.

Donkers, B., Verhoef, P. C., & De Jong, M. G. (2007). Modeling CLV: A test of competing models in the insurance industry. *Quantitative Marketing and Economics*, *5*, 163–190.

Dowling, G. R., & Uncles, M. (1997). Do customer loyalty programs really work? *Sloan Management Review*, *38*(2), 71–82.

Du, S., Bhattacharya, C. B., & Sen, S. (2007). Reaping relational rewards from corporate social responsibility: The role of competitive positioning. *International Journal of Research in Marketing*, *24*(3), 224–241.

Du, S., Bhattacharya, C. B., & Sen, S. (2010). Maximizing business returns to corporate social responsibility (CSR): The role of CSR communication. *International Journal of Management Reviews*, *12*(1), 8–19.

Farris, P. W., Bendle, N. T., Pfeifer, P. E., & Reibstein, D. J. (2006). *Marketing Metrics: Fifty+ Metrics Every Marketer Should Know*. Philedelphia: Wharton School Publishing.

Farris, P. W., Bendle, N. T., Pfeifer, P. E., & Reibstein, D. J. (2010). *Marketing Metrics: The Definitive Guide to Measuring Marketing Performance*. Upper Saddle River, New Jersey: Pearson Education.

Floyd, K., Freling, R., Alhoqail, S., Cho, H. Y., & Freling, T. (2014). How online product reviews affect retail sales: A meta-analysis. *Journal of Retailing*, *9*(2), 217–232.

Garbarino, E., & Johnson, M. (1999). The different roles of satisfaction, trust, and commitment in customer relationships. *Journal of Marketing*, *63*(2), 70–87.

Gensler, S., Verhoef, P. C., & Böhm, M. (2012). Understanding consumer's multichannel choices across the different stages of the buying process. *Marketing Letters*, *23*(4), 987–1003.

Gensler, S., Völckner, F., Egger, M., Fischbach, K., & Schoder, D. (2015). Listen to your customers: Insights into brand image using online consumer-generated product reviews. *International Journal of Electronic* Commerce*,* *20*(1), 112–141.

Geyskens, I., Steenkamp, J. B. E. M., Scheer, L. K., & Kumar, N. (1996). The effects of trust and interdependence on relationship commitment: A transatlantic study. *International Journal of Research in Marketing*, *13*(4), 303–317.

Gijsenberg, M. J., Van Heerde, H. J., & Verhoef, P. C. (2015). Losses loom longer than gains: Modeling the impact of service crises on customer satisfaction over time. *Journal of Marketing Research*, *52*(5), 642–656.

Gupta, S., Lehmann, D. R., & Stuart, J. A. (2004). Valuing customers. *Journal of Marketing Research*, *41*(1), 7.

Hanssens, D. M., Pauwels, K. H., Srinivasan, S., Vanhuele, M., & Yildirim, G. (2014). Consumer attitude metrics for guiding marketing mix decisions. *Marketing Science*, *33*(4), 534–550.

Hunneman, A., Verhoef, P. C., & Sloot, L. M. (2015). The impact of consumer confidence on store satisfaction and share of wallet formation.

Journal of Retailing, *91*(3), 516–532.

Juster, F. T. (1969). Consumer anticipations and models of durable goods demand. In Mincer J (1969). *Economic Forecasts and Expectations*. National Bureau of Economic Research, New York.

Keiningham, T. L., Cooil, B., Aksoy, L., Andreassen, T. W., & Weiner, J. (2007). The value of different customer satisfaction and loyalty metrics in predicting customer retention, recommendation, and share-of-wallet. *Managing Service Quality*, *17*(4), 361–384.

Kotler, P., & Lee, N. (2005). *Corporate Social Responsibility – Doing the Most Good for Your Company and Your Cause*. New Jersey: John Wiley and Sons.

Kumar, V., Bhaskaran, V., Mirchandani, R., & Shah, M. (2013). Creating a measurable social media marketing strategy for Hokey Pokey: Increasing the value and ROI of intangibles & tangibles. *Marketing Science*, *32*(2), 194–212.

Kumar, V., Donkers, A. C., Venkatesan, R., Wiesel, T., & Tillmanns, S. (2010). Undervalued or overvalued customers: Capturing total customer engagement value. *Journal of Service Research*, *13*(3), 297–310.

Leeflang, P. S. H., Verhoef, P. C., Dahlström, P., & Freundt, T. (2014). Challenges and solutions for marketing in a digital era. *European Management Journal*, *32*(1), 1–12.

Leenheer, J., Bijmolt, T. H. A., Van Heerde, H. J., & Smidts, A. (2007). Do loyalty programs really enhance behavioral loyalty? An empirical analysis accounting for self-selecting members. *International Journal of Research in Marketing*, *24*(1), 31–47.

Lemon, K. N., & Verhoef, P. C. (2016). Understanding customer experience throughout the customer journey. *Journal of Marketing*, *80*(6), 69–96.

Lewis, M. (2006). Customer acquisition promotions and customer asset value. *Journal of Marketing Research*, *43*(2), 195–203.

L'Hoest-Snoeck, S., Van Nierop, J. E. M., & Verhoef, P. C. (2015). Customer value modelling in the energy market and a practical application for marketing decision making. *International Journal of Electronic Customer Relationship Management*, *9*(1), 1–32.

Li, H. A., & Kannan, P. K. (2014). Attributing conversions in a multichannel online marketing environment: An empirical model and a field experiment. *Journal of Marketing Research*, *51*(1), 40–56.

Luo, X., & Bhattacharya, C. (2009). The debate over doing good: Corporate social performance, strategic marketing levers, and firm-idiosyncratic risk. *Journal of Marketing*, *73*(6), 198–213.

Mizik, N., & Jacobson, R. (2009). Valuing branded businesses. *Journal of Marketing*, *73*(6), 137–153.

Moorman, C., Zaltman, G., & Deshpande, R. (1992). Relationship between providers and users of market research: The dynamics of trust within and between organizations. *Journal of Marketing Research*, *29*(3), 314–328.

Morgan, N. A., & Rego, L. L. (2006). The value of different customer satisfaction and loyalty metrics in predicting business performance. *Marketing Science*, *25*(5), 426–439.

Oliver, R. L., Rust, R. T., & Varki, S. (1997). Customer delight: Foundations, findings, and managerial insight. *Journal of Retailing*, *73*(3), 311–336.

Onish, H., & Manchanda, P. (2012). Marketing activity, blogging and sales. *International Journal of Research in Marketing*, *29*(3), 221–234.

Onrust, M., Verhoef, P. C., Van Doorn, J., & Bügel, M. S. (2014). When doing good leads to increased customer loyalty: Why weak firms can benefit from CSR. *Working paper*, University of Groningen.

Ou, Y. C., De Vries, L., Wiesel, T., & Verhoef, P. C. (2014). The role of consumer confidence in creating customer loyalty. *Journal of Service Research*, *17*(3), 229–354.

Porter, M. E., & Kramer, M. R. (2011). Creating shared value. *Harvard Business Review*, *89*(1/2), 62–77.

Rapp, A., Baker, T. L., Bachrach, D. G., Ogilvie, J., & Beitelspacher, L. S. (2015). Perceived customer showrooming behavior and the effect on retail salesperson self-efficacy and performance. *Journal of Retailing*, *91*(2), 358–369.

Reichheld, F. F. (2003). The one number you need to grow. *Harvard Business Review*, *81*(12), 46–54.

Risselada, H., Verhoef, P. C., & Bijmolt, T. H. A. (2015). Indicators of opinion leadership in customer networks: Self reports and degree centrality. *Marketing Letters*, *27*(3), 449–460.

Rogers, E. M. (1995). *The Diffusion of Innovations*. New York: The Free Press.

Rust, R. T., Lemon, K. N., & Zeithaml, V. A. (2004). Return on marketing: Using customer equity to focus marketing strategy. *Journal of Marketing*, *68*(1), 109–127.

Rust, R. T., Zeithaml, V. A., & Lemon, K. N. (2000). *Driving Customer Equity: How Customer Lifetime Value Is Reshaping Corporate Strategy*. New York: The Free Press.

Selnes, F., & Gønhaug, K. (2000). Effects of supplier reliability and benevolence in business marketing. *Journal of Business Research*, *49*(3), 259–271.

Shah, D., Kumar, V., Qu, Y., & Chen, S. (2012). Unprofitable cross-buying: Evidence from consumer and business markets. *Journal of Marketing*, *76*(3), 78–95.

Sharp, B. (2010). *How Brands Grow*. Australia & New Zealand: Oxford University Press.

Streukens, S., & De Ruyter, K. (2004). Reconsidering nonlinearity and asymmetry in customer satisfaction and customer loyalty models: An empirical study in three retail service settings. *Marketing Letters*, *15*(2/3), 99–111.

Tellis, G. J., & Johnson, J. (2007). The value of quality. *Marketing Science*, *26*(6), 758–773.

Trampe, D., Konuş, U., & Verhoef, P. C. (2014). Customer responses to channel migration strategies toward the e-channel. *Journal of Interactive Marketing*, *28*(4), 257–270.

Trusov, M., Bucklin, R. E., & Pauwels, K. (2009). Effects of word-of-mouth versus traditional marketing: Findings from an Internet social networking site. *Journal of Marketing*, *73*(5), 90–102.

Van Doorn, J., Leeflang, P. S. H., & Tijs, M. (2013). Satisfaction as a predictor of future performance: A replication. *International Journal of Research in Marketing*, *30*(3), 314–318.

Van Doorn, J., Lemon, K. N., Mittal, V., Naβ, S., Pick, D., Pirner, P., & Verhoef, P. C. (2010). Customer engagement behavior: Theoretical foundations and research directions. *Journal of Service Research*, *13*(3), 253–266.

Van Doorn, J., & Verhoef, P. C. (2008). Critical incidents and the impact of satisfaction on customer share. *Journal of Marketing*, *72*(4), 123–142.

Van Heerde, H. J., Gijsbrechts, E., & Pauwels, K. (2008). Winners and losers of a major price war. *Journal of Marketing Research*, *45*(5), 499–518.

Van Nierop, E., Leeflang, P. S. H., Teerling, M. L., & Huizingh, E. K. R. (2011). The impact of introducing and using an informational website on offline customer buying behavior. *International Journal of Research in Marketing*, *28*(2), 155–165.

Venkatesan, R., & Kumar, V. (2004). A customer lifetime value framework for customer selection and resource allocation strategy. *Journal of Marketing*, *68*(4), 106–215.

Verhoef, P. C. (2003). Understanding the effect of customer relationship management efforts on customer retention and customer share development. *Journal of Marketing*, *67*(4), 30–45.

Verhoef, P. C., Franses, P. H., & Hoekstra, J. C. (2002). The effect of relational constructs on customer referrals and number of services purchased from a multiservice provider: Does age of relationship matter? *Journal of the Academy of Marketing Science*, *30*(3), 202–216.

Verhoef, P. C., Kannan, P. K., & Inman, J. (2015). From multi-channel retailing to omni-channel retailing: Introduction to the special issue on multi-channel retailing. *Journal of Retailing*, *91*(2), 174–181.

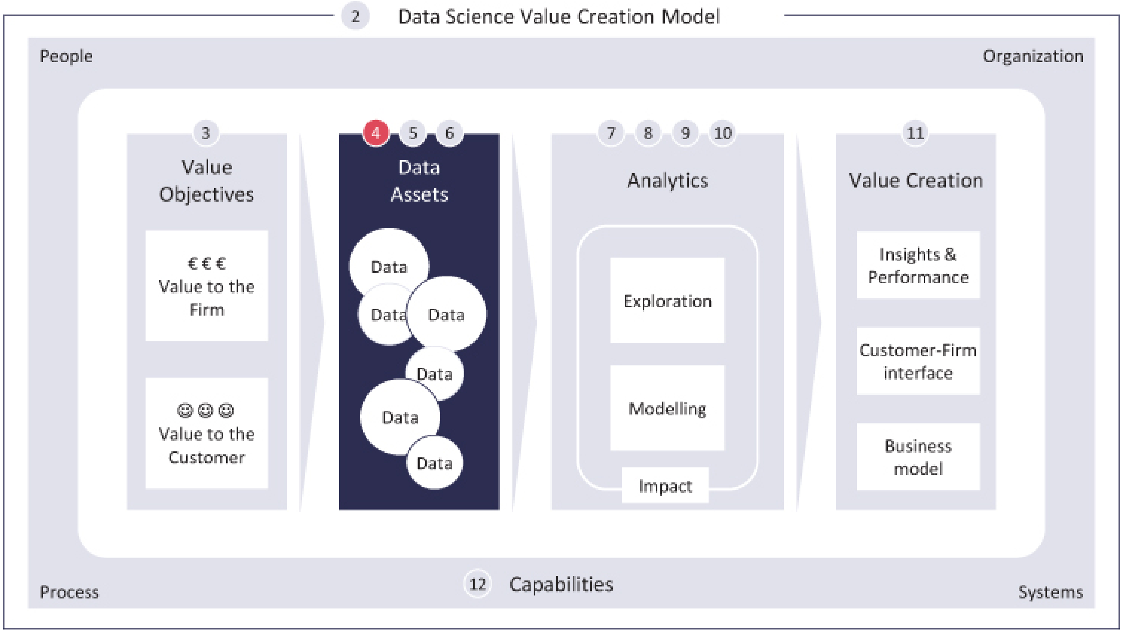
Verhoef, P. C., & Langerak, F. (2001). Possible determinants of consumers' adoption of electronic grocery shopping in the Netherlands. *Journal of Retailing and Consumer Services*, *8*(5), 275–285.

Verhoef, P. C., Neslin, S. A., & Vroomen, B. (2007). Browsing versus buying: Determinants of customer search and buy decisions in a multichannel environment. *International Journal of Marketing Research*, *24*(2), 129–148.

Weber Shandwick/KRC Research (2012). Buy It, Try It, Rate It. Retrieved from http://www.webershandwick.com/uploads/news/files/ReviewsSurveyReportFINAL.pdf

CHAPTER 4
Data assets

DOI: 10.4324/9781003011163-4

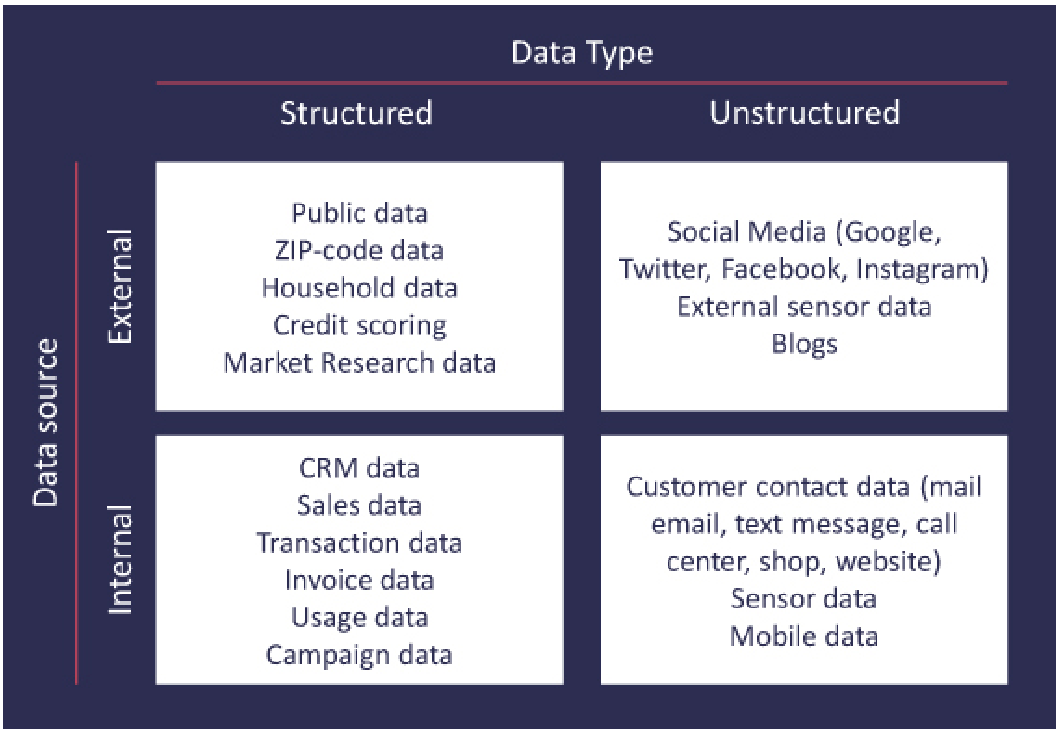


4.1 INTRODUCTION

In this world of big data, everything starts by having to deal with the overload and variety of available data, to help realize the ambitious objectives of value creation using data science. That's why data is a very important part of this book. Earlier, we explained why we talk about "big" data (by using the three Vs model). One of the three Vs was Variety. This represents the fact that data nowadays comes from more and more sources, and that data can be found everywhere. In this chapter, we will elaborate on this variety of data sources. We will discuss what kind of data sources can be distinguished and how all these data can be categorized, as well as how data should be processed and stored. Specific attention will be paid to data quality, data cleansing, and missing values. We will follow the same structure as in the other chapters, meaning that we will discuss market data sources, product and brand data sources, and data sources with customer data—all with a specific focus on the new data sources that have become available.

4.2 DATA SOURCES AND THE DIFFERENT TYPES OF DATA

Firms have many different data sources at their disposal to fill their commercial data environment. By commercial data environment, we mean the technical infrastructure where all data is stored, processed, and accessed, coming from all kinds of sources to be used for commercial steering. In this environment, we see nowadays a combination of "traditional" data warehousing technology combined with modern technology like data lakes that come with big data tooling, such as Hadoop, MapReduce, etc. This combination of technologies facilitates dealing with the challenges posed by the existence of different types of data sources, and dealing with specific challenges in data volume and variety. Data warehouses and data lakes will be discussed in more detail in Chapters 5 and 12. Within the data sources, we distinguish between internal versus external and structured versus unstructured data sources. The different combinations of these types of data sources are shown in Figure 4.1.



		Data Type	
		Structured	Unstructured
Data source	External	Public data ZIP-code data Household data Credit scoring Market Research data	Social Media (Google, Twitter, Facebook, Instagram) External sensor data Blogs
	Internal	CRM data Sales data Transaction data Invoice data Usage data Campaign data	Customer contact data (mail email, text message, call center, shop, website) Sensor data Mobile data

FIGURE 4.1 Two dimensions of data: data source versus data type

Source: Adapted from Nair and Narayayan (2012)

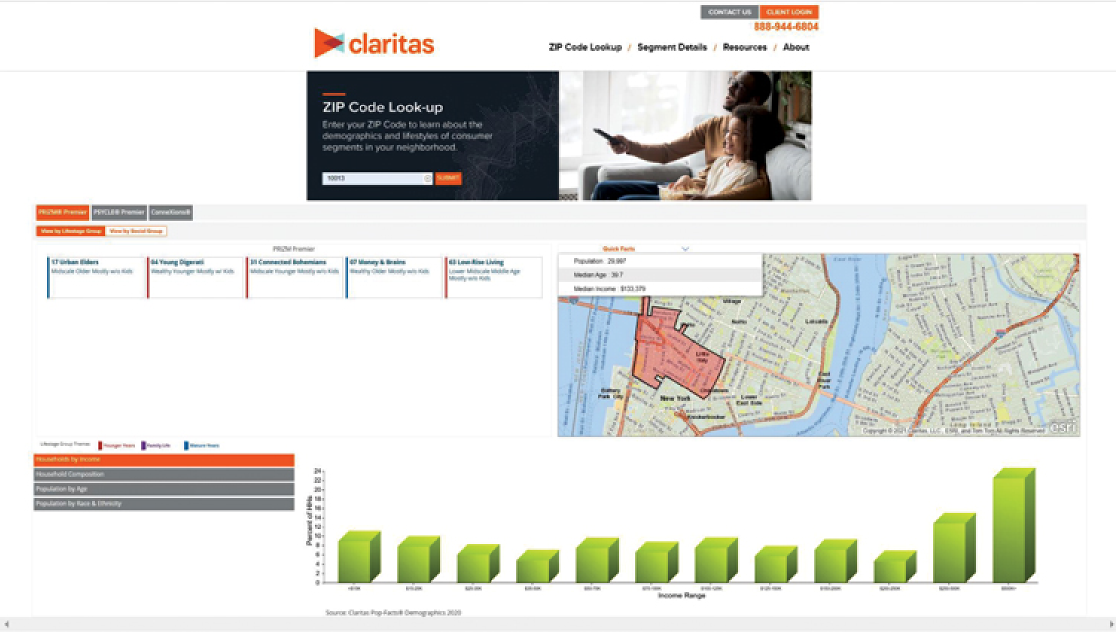
A first important distinction in these data sources is between external and internal data sources.

4.2.1 External data sources versus internal data sources

External data sources are not present within the firm and are frequently purchased from (or collected by) external data vendors and added to the commercial data environment. Important examples of these data are zip-code or household datasets. In zip-code datasets, information is provided on the characteristics of households living in that zip-code, such as average income level, education level, average house-price, etc. This information can be linked to individual customers through the zip-code of the customer. More and more vendors are transforming their zip-code datasets into household datasets, by enriching zip-code datasets with information obtained from omnibus questionnaires and publicly available datasets for commercial use (such as information on the sale of houses, including house prices). This transformation helps to create more powerful profiles of households, instead of having to make use of extrapolations for all households in a zip-code. The external information described above can be considered as more generic data that is useful for different purposes and in a wide variety of industries.

Other external information may be more industry-specific, such as information regarding the financial creditworthiness of a customer. Credit card firms, including American Express, Mastercard, and Visa, have accurate information about a consumer's creditworthiness from analyzing the use of their credit card. An increasing number of firms (e.g., Mastercard) are starting to exploit their datasets to be used as an external data source for checks of creditworthiness, for example with telecom providers.

Important suppliers of external consumer data are Experian, Acxiom, and Claritas. External B2B data are provided by firms such as Graydon and Altares Dun & Bradstreet. External consumer data are frequently used for target marketing purposes and profiling current customers in terms of socio-demographics, lifestyle, and psychographic variables. Suppliers such as Claritas have developed detailed segmentation systems, in which each household belongs to a certain psychographic-lifestyle segment, such as Urban elders, Young digerati, or Connect bohemians, etc. (see Figure 4.2).



17 Urban Elders
Midscale Older Mostly w/o Kids

04 Young Digerati
Wealthy Younger Mostly w/ Kids

31 Connected Bohemians
Midscale Younger Mostly w/o Kids

07 Money & Brains
Wealthy Older Mostly w/o Kids

63 Low-Rise Living
Lower Midscale Middle Age Mostly w/o Kids

Population : 29,997

Median Age : 39.7

Median Income : $133,379

FIGURE 4.2 Example of Claritas information for a New York zip-code

Source: https://claritas360.claritas.com/mybestsegments/ (accessed February 18, 2021)[1]

Another kind of external data is marketing research data, mostly collected for firms by their market research agencies. These agencies often have large panel sets, with respondents in place to perform quantitative market research on topics such as brand performance, campaign evaluation, media effectiveness, and product innovation. Market research can also be realized by using data on the firm's customers, such as customer satisfaction or NPS, usually collected by external marketing research suppliers. Depending on privacy regulations, customer permission, and/or the specific research professional code, market research data can be included in the commercial data environment as well.

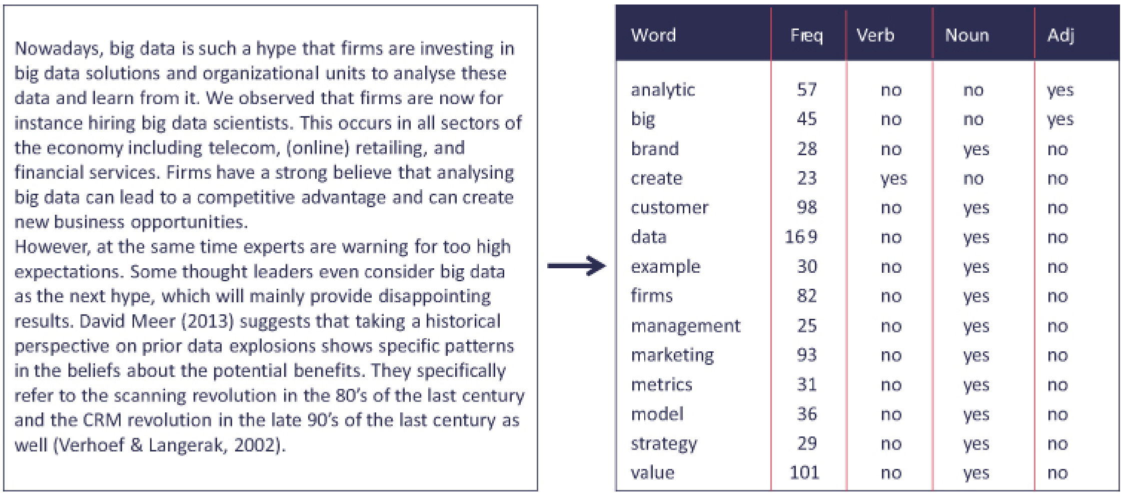
Furthermore, firms may collect competitive intelligence data on competitive actions. These data are usually not internally present and can be collected by marketing intelligence departments. Research has shown that competitive actions, such as competitive advertising, have an impact on customer behavior (Prins & Verhoef, 2007). The last type of external data to be addressed is open or public data (see also 4.2.3). This type of data is shared by governments or organizations for free (or almost free) and is publicly available. It opens up completely new opportunities to gain unique insights.

Another type of fast-growing external data comes from social media. These data come from parties like Facebook, Twitter, LinkedIn, Instagram, TikTok, and other social media platforms, all of which have huge user bases so that, often, a substantial proportion of a firm's customers participate in these social media. Linking these data to the commercial data environment is quite a challenge and not yet a common practice. This is mainly due to the highly unstructured way such data are being created. Firms are collecting these data by using platforms like a module in Salesforce (formerly known as Radian6) for social media monitoring but are still struggling to interpret the data and integrate these platforms with their commercial environments.

Internal data sources are already present within the firm and may include point-of-sales data, transaction data, invoice data, contact data, and usage data. These internal data sources are gathered and stored in data warehouses or data lakes. Internal data can be considered very powerful. The possession of information from internal data is important if customer behavior is to be described, understood, and predicted. However, these data sources need a lot of data preparation to create valuable information. Within customer value management (CVM) this is the most frequently stored data (Verhoef *et al.*, 2002). In addition, many models aiming to optimize the customer value only use internal data (e.g., Venkatesan & Kumar, 2004).

4.2.2 Structured versus unstructured data

An alternative way to structure data sources is to make a distinction between structured and unstructured data sources. Structured data are data that come in a fixed format, based on a detailed record and variable structure, good labeling of values in the database, and high data quality. The invoice data mentioned earlier (internal data) and zip-code data (external data) are good examples of highly structured data. On the other side of the coin, we have unstructured data. These data are often very bulky in size, without a fixed format, containing lots of free format text and often needing a good deal of data interpretation and data reduction to create useful information. Examples of these data sources are data from customer contact (internal data), where customers and their questions or remarks are often registered in free format text, and social media data (external data) contained in Twitter messages, Facebook comments, etc. Another important source that is becoming more and more relevant (and bulky) is all kinds of video and image data that is being collected (via YouTube for example). Figure 4.3 shows an example of how unstructured data can be transformed into structured data.



Nowadays, big data is such a hype that firms are investing in big data solutions and organizational units to analyse these data and learn from it. We observed that firms are now for instance hiring big data scientists. This occurs in all sectors of the economy including telecom, (online) retailing, and financial services. Firms have a strong believe that analysing big data can lead to a competitive advantage and can create new business opportunities.

However, at the same time experts are warning for too high expectations. Some thought leaders even consider big data as the next hype, which will mainly provide disappointing results. David Meer (2013) suggests that taking a historical perspective on prior data explosions shows specific patterns in the beliefs about the potential benefits. They specifically refer to the scanning revolution in the 80's of the last century and the CRM revolution in the late 90's of the last century as well (Verhoef & Langerak, 2002).

Word	Freq	Verb	Noun	Adj
analytic	57	no	no	yes
big	45	no	no	yes
brand	28	no	yes	no
create	23	yes	no	no
customer	98	no	yes	no
data	169	no	yes	no
example	30	no	yes	no
firms	82	no	yes	no
management	25	no	yes	no
marketing	93	no	yes	no
metrics	31	no	yes	no
model	36	no	yes	no
strategy	29	no	yes	no
value	101	no	yes	no

FIGURE 4.3 Illustration of structured and unstructured data

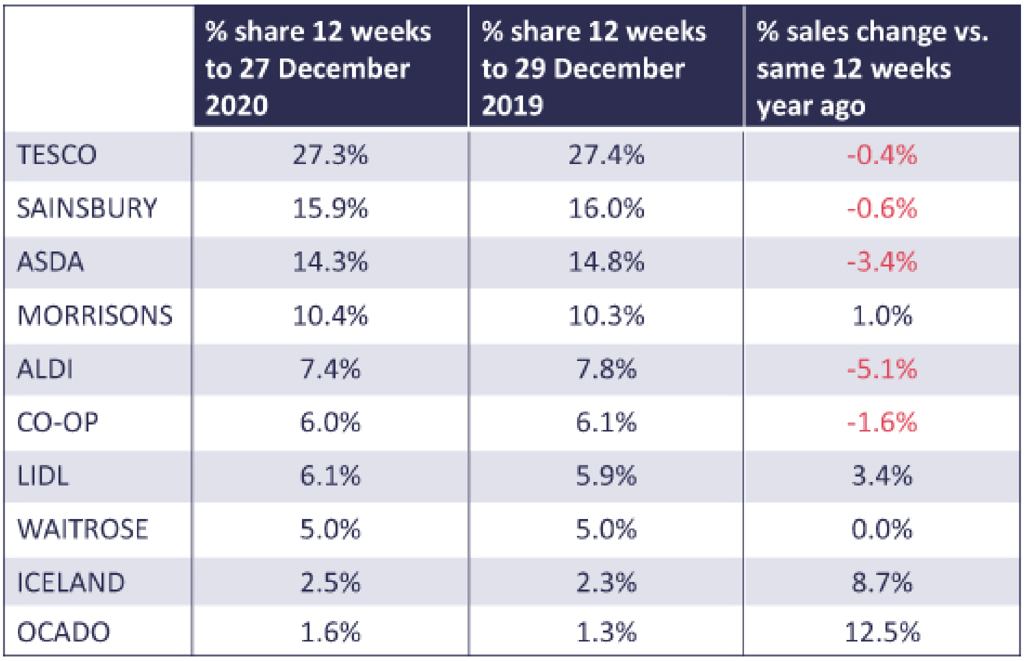
A special comment is necessary about mobile data. Mobile data (in Figure 4.1 depicted as internal, unstructured data) took off with the increase of smartphone and tablet penetration. This kind of data is unique since it offers organizations the possibility not only to monitor what customers are doing but also where they are doing it—thereby giving big opportunities to all kinds of location-based services. Installed apps give the customer access to products and services that offer all kinds of service and transaction functionalities. Since the usage of data generated by these apps is often not yet integrated within the architecture of the commercial data environment, and because the data from these apps are often unstructured, there is still pioneering work to be done in this area.

4.2.3 Market data

Another lens for discussing data sources is our breakdown by market, brand/product, and customer. This is not only useful from the perspective of the big data challenge but it also creates awareness about the differences between the data sources for the market, product/brand, and customers, differences that arise due to scope, detail, and power of the data source at hand. We define two types of market data:

1. Market data on the supply side: Data that describe and explain metrics like market size, market volume, market share, market development, media spending, etc. for all players in the market, ideally with the possibility to be shown per competitor in the market. This type of data is very useful for assessing market potential in terms of money and opportunities to capture value.
2. Market data on the demand side: Data that describe the consumers in the market, their buying and spending behavior, their socio-demographics (e.g., age, household size, income), and their needs.

Market data on the supply side is often collected by agencies that aggregate, for example, sales figures from different suppliers to create a market overview. Sometimes this can be an agency that operates for the whole industry. For example, employment agencies in the Netherlands all deliver their vacancies per four-week period to ABU, an organization that collects this data for the whole industry and delivers it back at an aggregated level. Another example is Kantar, which collates the household grocery purchasing habits of 30,000 demographically representative households in Great Britain and creates detailed sales figures based on this (see Figure 4.4).



	% share 12 weeks to 27 December 2020	% share 12 weeks to 29 December 2019	% sales change vs. same 12 weeks year ago
TESCO	27.3%	27.4%	-0.4%
SAINSBURY	15.9%	16.0%	-0.6%
ASDA	14.3%	14.8%	-3.4%
MORRISONS	10.4%	10.3%	1.0%
ALDI	7.4%	7.8%	-5.1%
CO-OP	6.0%	6.1%	-1.6%
LIDL	6.1%	5.9%	3.4%
WAITROSE	5.0%	5.0%	0.0%
ICELAND	2.5%	2.3%	8.7%
OCADO	1.6%	1.3%	12.5%

FIGURE 4.4 Example of market data on the supply side for UK supermarkets

Source: Adapted from Kantar, https://www.kantarworldpanel.com/grocery-market-share/great-britain/snapshot/27.12.20/29.12.19 (accessed February 18, 2021)[2]

Market data on the demand side is mainly collected by research agencies. An example is Ipsos, which creates, for instance, an overview of the insurance market by asking consumers for details of their insurance documents. The main difference with supply-side data is the method of data collection. Demand-side data are built up from the users or buyers of products or services, often at the household level, as far as it concerns consumers. We can consider the external data that is provided by vendors like Experian (see above) as an example of market data since these are built up from the consumer/demand-side perspective at the household or zip-code level. Market data on the demand side are not only data on product usage or product-buying and socio-demographics, but they can also consist of "softer data," describing consumer needs and values as they exist in the market. These datasets categorize consumers in segments like "cosmopolitans" or "post-materialists" as used by, for example, Motivaction, a Dutch research agency. These groups display specific needs with respect to a certain industry or product category. In Figure 4.5 the mentality segmentation of Motivaction is displayed as an example of market data on the demand side.

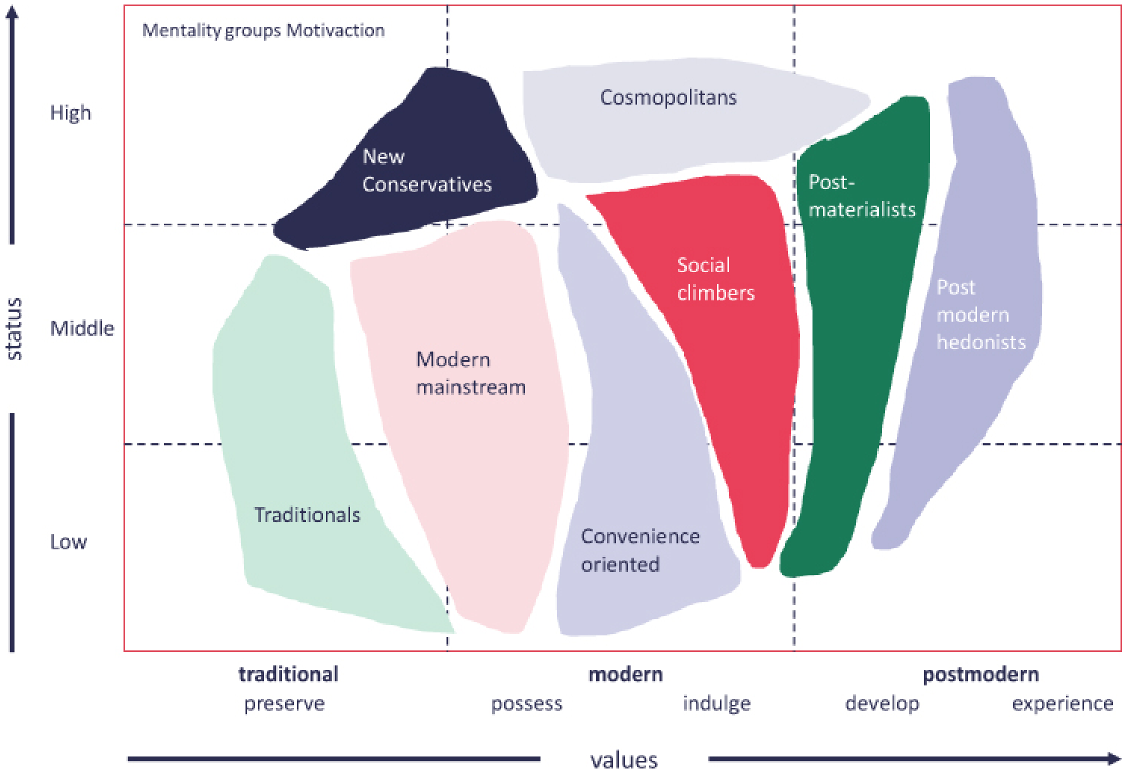


FIGURE 4.5 Example of market data on the demand side

Source: Motivaction, https://www.motivaction.nl/en/mentality/mentality-segmentation (accessed February 18, 2021)[3]

By definition market data have the broadest scope, as they aim to represent the total market for an industry or product/service. Since collecting and updating this kind of data can be very costly, however, the data may sometimes not be as detailed as desired and in some cases might even be outdated. This of course limits the power of use of these kinds of data sources.

4.2.4 Big data influence on market data

Alternatives for traditional market data are increasingly available, especially due to developments like open data projects, where governments and organizations share their data using the enormous power of digitalization and online data. As an example, analyzing Google location data collected by mobile devices during the Covid-19 crisis, showing the steep decline in the number of visits to offices and schools and in usage of public transportation demonstrates the effectiveness of government policies, showing how peoples' movements have changed as presented in Google Covid-19 Community Mobility Reports.

Another example is data collected by all kinds of comparison websites. They all have a good view of market volumes and transactions, especially since the online share of transactions in all industries is still rising very fast (mainly in the orientation phase of the buying process) and possible biases are reducing.

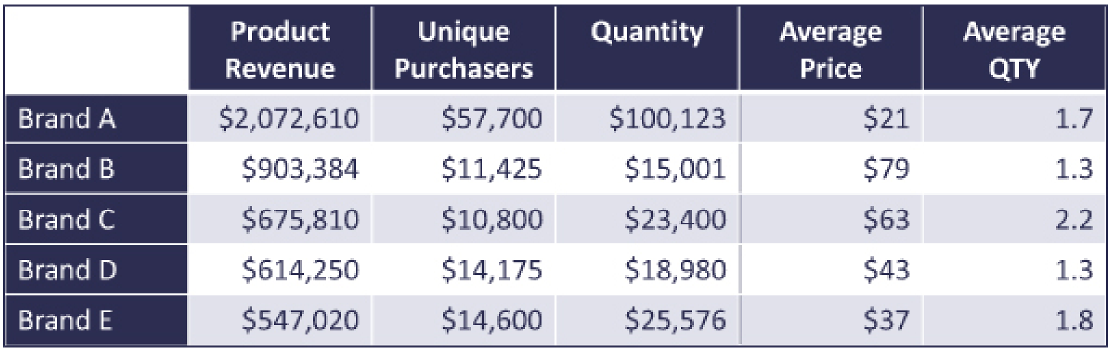
4.2.5 Brand data

For brand data, we again make a distinction between:

1. Brand data on the supply side: Data that describe the volume and market share of specific brands/products
2. Brand data on the demand side: Data that describe and explains how (potential) customers judge a certain brand.

Brand data have a limited scope in that they focus only on brands and their performance. Collecting these kinds of data, especially for the demand side, is a costly effort. To collect these data, research is necessary, and this brings up discussions about the quality and validity of results.

For brand data on the supply side, there are no specific data sources built up. Either these data are a subset or cross-section of the market data mentioned above (for example, to calculate the market share of a specific brand) or they are extracted from the systems that store the sales per brand for the organization (see discussion of customer data later in this chapter). In Figure 4.6 we show an example of what this type of data might look like.



	Product Revenue	Unique Purchasers	Quantity	Average Price	Average QTY
Brand A	$2,072,610	$57,700	$100,123	$21	1.7
Brand B	$903,384	$11,425	$15,001	$79	1.3
Brand C	$675,810	$10,800	$23,400	$63	2.2
Brand D	$614,250	$14,175	$18,980	$43	1.3
Brand E	$547,020	$14,600	$25,576	$37	1.8

FIGURE 4.6 Illustration of brand supply data extracted from internal systems

Brand data on the demand side (see the example in Figure 4.7) focuses on the different steps in the customer orientation process, as measured by brand funnels with steps such as brand awareness, brand consideration, and brand preference (see Chapter 3). These kinds of data are collected by conducting (online) market research that is performed in longitudinal studies.

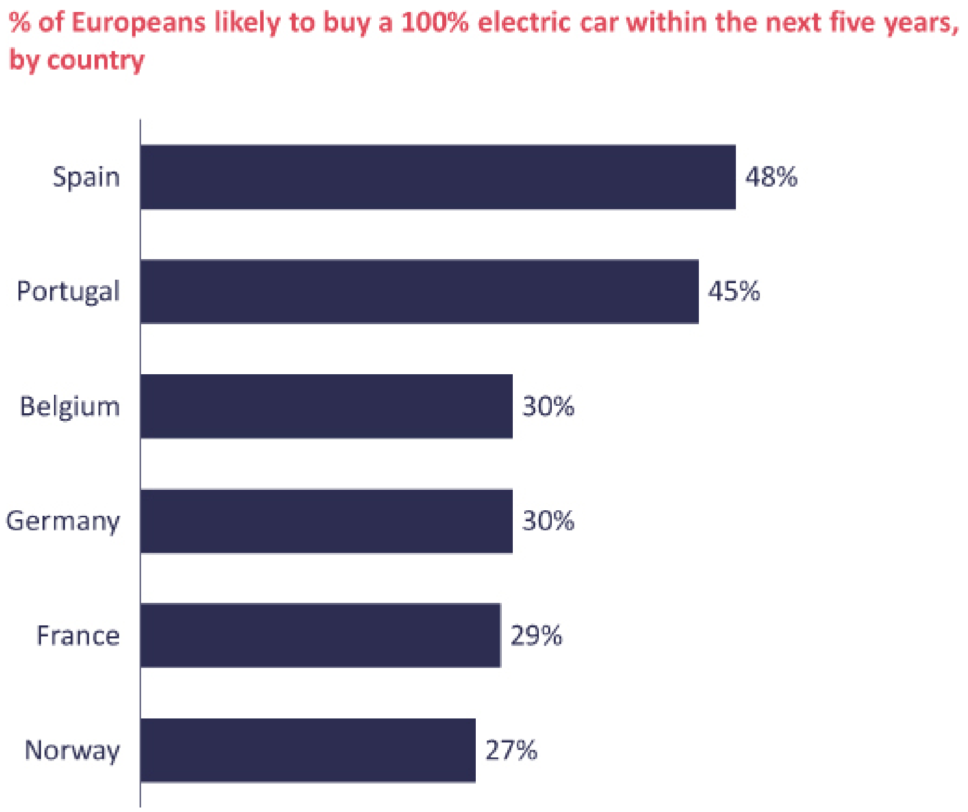


FIGURE 4.7 Illustration of brand demand based on market research

Source: Adapted from Statista https://www.statista.com/statistics/1154513/intention-purchase-hybrid-car-electric-europe/ (accessed February 18, 2021)[4]

4.2.6 Big data influence on brand data

In these days of big data, brand data can also be collected by looking into social media to measure brand sentiment, not to replace but to add to brand data collected by research. Reviews and ratings can also be an important big data source for measuring and collecting brand data. To prevent bias, we suggest that these kinds of new data sources are analyzed in combination with traditional sources, instead of completely replacing them.

4.2.7 Customer data

Although the distinction between supply and demand data can also be made for customer data, there is as much customer data available on the supply side as there is on the demand side:

1. Customer data on the supply side: Data that describes the (historical) product and services used by the customer during the relationship with the organization
2. Customer data on the demand side: Data that describes the expectations, satisfaction, and interactions of the customer with the organization.

Customer data is different from market and brand data in several ways: customer data is often very detailed and accurate (especially since billing data is the life-blood of the organization) and by definition only covers customers (and sometimes prospects and ex-customers). So it has a narrower scope and has become (from the moment organizations were successful in extracting and storing this data) very powerful in the commercial steering process.

Customer data on the supply side are mainly stored in the CRM systems of the organization (see Figure 4.8 for an example of a relational marketing database environment). In a contractual setting, the billing data are an important source for identifying the customer and assessing the financial value of the customer, as well as helping to analyze product holding and usage. In a non-contractual setting where customer identification is not always possible these data are often limited to data on the product level, stating usage and buying of products, including repeat purchase.

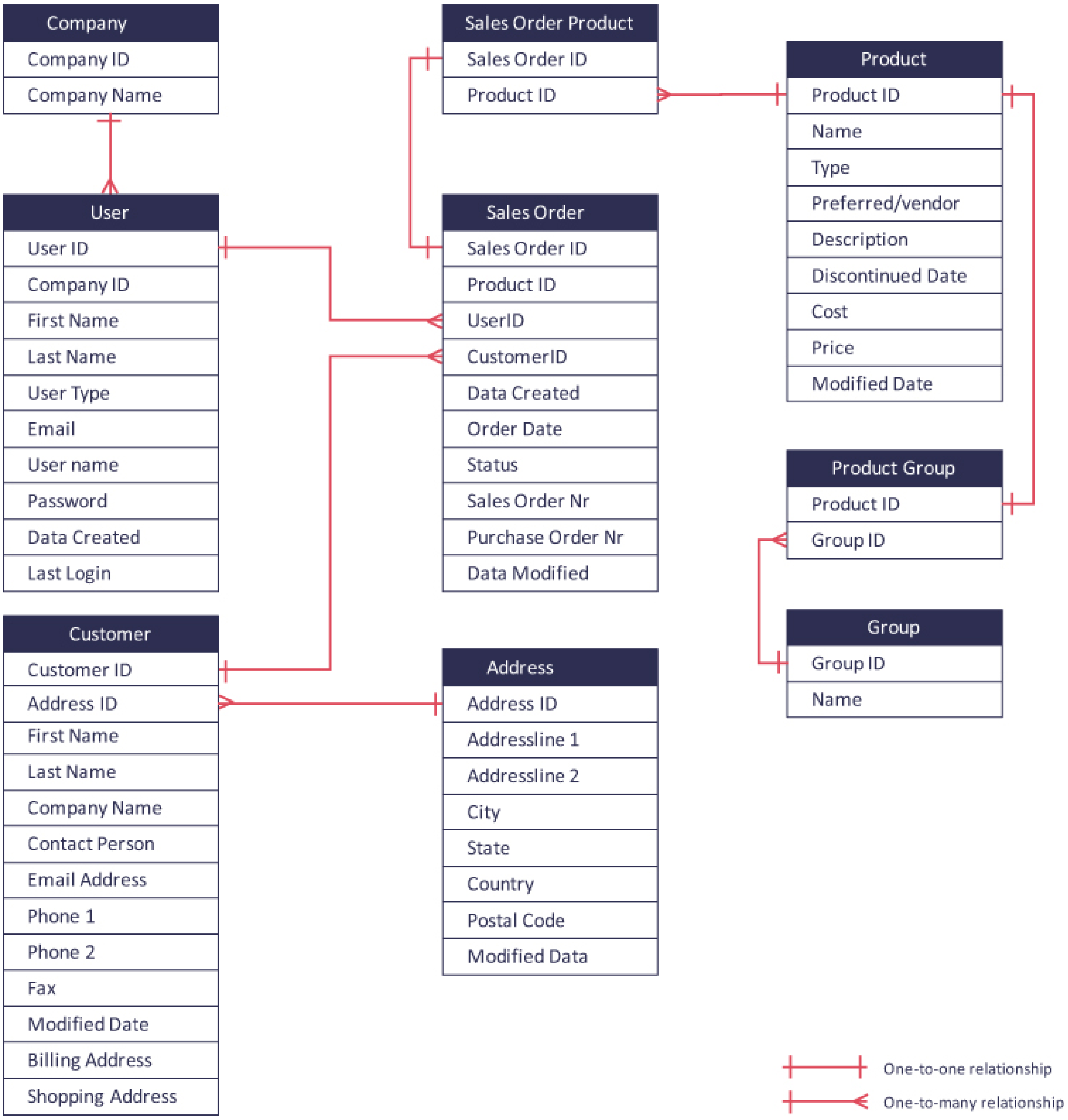


FIGURE 4.8 Illustration of a data model of customer supply data

Customer data from the demand side (see Figure 4.9) comes from market research on the customer base, researching metrics such as NPS or customer satisfaction, or from interactions between the customer and the organization during the customer journey, for example in a call center, shop, or on a website. Because these data are often (at least partly) unstructured, interpretation and analysis of them are necessary to transform them into information and knowledge.

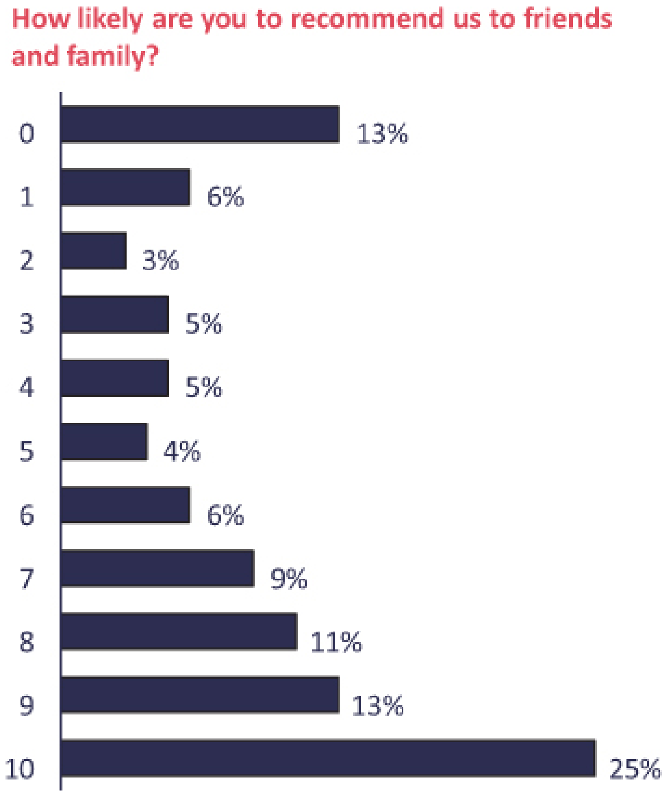


FIGURE 4.9 Illustration of customer demand data (NPS)

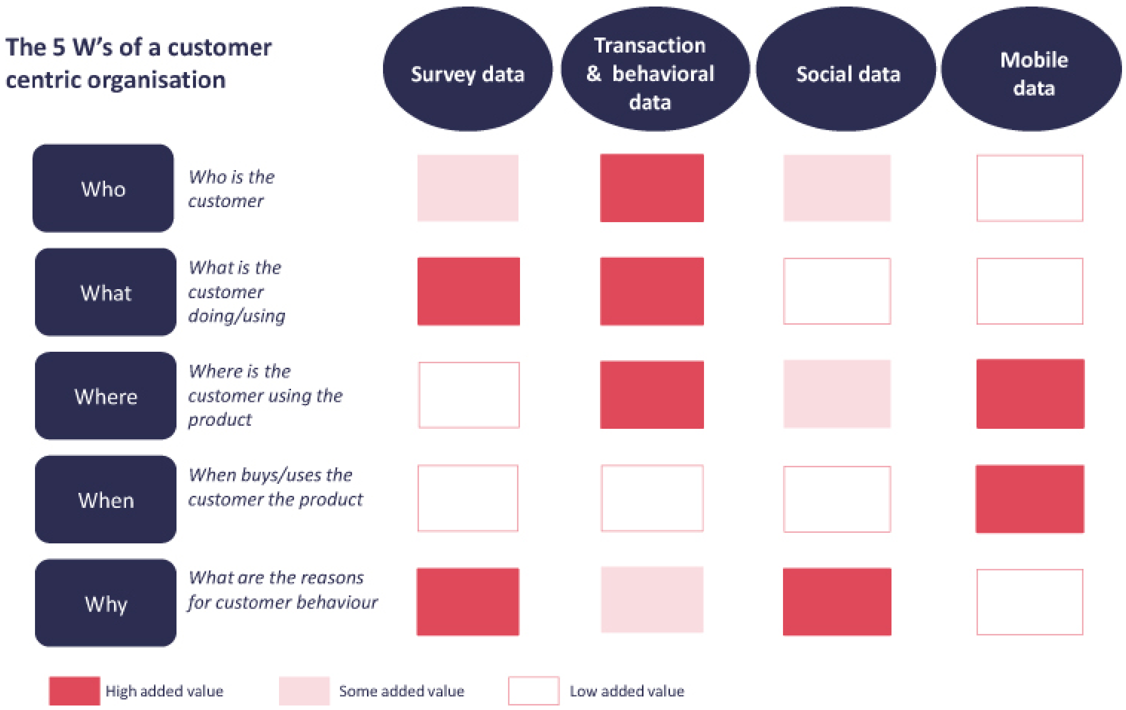
4.2.8 Big data influence on customer data

The influence of big data on customer data can be found in the large internal, unstructured datasets that are being built up and that are more and more within the scope of the data analyst. Also, the online presence of organizations designed to help them to serve their customers is something that is creating big data—for example, in the "my [organization name]" environment that many organizations create to inform their customers about their billing, products, etc., not to mention all kinds of "self-service" options. This is influencing customer data on the supply side as well as customer data on the demand side.

4.3 USING THE DIFFERENT DATA SOURCES IN THE ERA OF BIG DATA

To assess and to make maximum use of the added value of every data source, we use the 5 "Ws" (see Figure 4.10):

- Who is the customer?
- What is the customer doing/using?
- Where does the customer use or buy the product?
- When does the customer use or buy the product?
- Why does the customer use or buy the product?



The 5 W's of a customer centric organisation		Survey data	Transaction & behavioral data	Social data	Mobile data
Who	*Who is the customer*	▨	🟥	▨	□
What	*What is the customer doing/using*	🟥	🟥	□	□
Where	*Where is the customer using the product*	□	🟥	▨	🟥
When	*When buys/uses the customer the product*	□	□	□	🟥
Why	*What are the reasons for customer behaviour*	🟥	▨	🟥	□

🟥 High added value ▨ Some added value □ Low added value

FIGURE 4.10 The 5 "Ws" model for assessment of data sources

From a customer centric perspective, these are the questions that always pop up. It shows that every data source has its specific strength in answering certain "W" questions, depicted by the color in Figure 4.10. So survey data, for example, is very effective in answering the "Why" of customer behavior as well as "What" the customer is doing or intending to do. However, these "W" questions are often only answered from a single data source perspective. This can cause problems because the answers to the different "Ws" might not be consistent, or they may even be contradictory. To solve this, we think that the data sources that might come from surveys, transactions, social media, or mobiles need to be combined.

Combining these data sources is, however, challenging from a technical, statistical, and legal perspective. From a technical perspective, because some data sources might be anonymous or do not have a unique key to link the sources. From a statistical perspective, because not all data sources have the same coverage of the population (for example, a population sample or only data on customers and not on the total market). To deal with this, extrapolation techniques will be necessary. From a legal perspective this is also challenging, because combining sources on the individual level might create conflicts with privacy policies, a professional code (such as that of market researchers), or legislation on privacy. In Chapter 5 we will go into more depth on how to solve merging datasets from different origins, with different coverage and/or legal hurdles for merging, using techniques such as data fusion.

4.4 DATA QUALITY AND DATA CLEANSING

4.4.1 Data quality

Data quality is a very important and relevant topic when data is being used for servicing customers or when it is fundament for crucial business decisions. If the data is not correct, using it can have serious consequences for the organization. At the same time, many organizations struggle with data quality, realizing that they have serious issues with it that they should be able to manage. For a good understanding of data quality, there are several dimensions to consider. These include:

- Completeness of data
- Accuracy of data
- Consistency of data.

Completeness of data refers to whether all available data are present for all customers. For example, a data acquisition channel may be present for only a subset of customers (e.g., Verhoef & Donkers, 2005). This is also called the problem of missing values/ missing observations (see the last part of this chapter for a deeper understanding of this problem). Reasons for a lack of completeness of data can be data incomplete registration for forms or input fields that have been added at a later stage, where this data is missing for older records. System migrations, where for example two customer datasets are merged, may also result in incomplete data.

Accuracy of data refers to whether the data factually represents the information as available in the “real world”. The main explanation for a lack of accuracy of data is human error. But data that is outdated can also be considered inaccurate.

Mistakes frequently occur, especially concerning customer descriptors, such as name and address. These mistakes may arise as customers write down unclear names etc. on forms, perhaps on purpose, or when typographical errors mean that a data entry is not done correctly.

Having up to date data means data is updated frequently enough. A database that is not up to date can easily contain mistakes on all kinds of variables. For example, if customers have moved to another address, the address in the database will be wrong if it has not been updated. Or if the integration of databases is not done frequently, a recent product purchase or a recent defection might not have been included yet, leading to unreliable information.

Mistakes can potentially have strong negative reputational consequences for a firm. A firm may continue to send mail to someone who has recently passed away. Moreover, data being out of date may also cause wrong predictions to be made on future customer value, which might lead to less than optimal strategies.

The last dimension of data quality in this book is the consistency of data. There is always a risk that these data sources are contradictory when compared, especially when multiple sources provide information for customer datasets. Even within data sources there is a risk of inconsistency. For example, when a 25-year-old appears to be living in a seniors' home. This can be an important sign of inconsistency. Outliers in the data, i.e., extreme and rare values are also examples of problems with the consistency of the data.

Data quality is thus important for shaping performance in CVM strategies (Zablah, Bellenger, & Johnston, 2004). However, the question is whether firms should have an extremely high level of data quality with perfectly complete data, no data mistakes, and 100% up-to-date data. Neslin *et al*. (2006) propose that there might an optimum level (see Figure 4.11). Achieving good data quality comes at a cost. However, these costs rise in a non-linear fashion when higher levels of data quality are achieved, while the return in terms of increased customer value from this data quality may actually decrease, also in a non-linear fashion. This implies that firms should assess the optimum level of data quality and not pursue a 100% score.

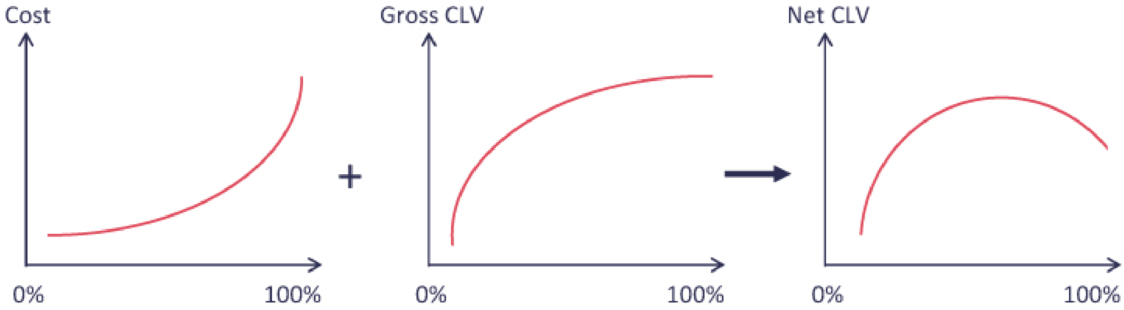


FIGURE 4.11 Net benefits of investing in data quality

Source: Adapted from Neslin *et al*. (2006)

4.4.2 Data cleansing

Data cleansing is the procedure for cleaning up data. It is the act of detecting and correcting problems in the data and identifying corrupt or inaccurate records from a record set, a table, or database, and from there on replacing modifying and/or deleting these incomplete, corrupt, or inaccurate parts of 'dirty' data. Proper reporting of data cleansing is key! Data cleansing can occur within a single set of records, or between multiple sets of data that need to be combined. The goal is not just cleaning up the data, but also bringing consistency to different sets of data that have been merged from separate sources. Examples of data cleansing are resolving typos or spelling errors, properly labeling mislabeled data, or completing incomplete or missing entries. Ideally, the majority of data cleansing is fully automated, with only a very small amount requiring manual assessment and correction.

There are several steps to take to solve problems with data quality that is technically correct. The first step is parsing. Parsing locates and identifies individual data elements in the source file(s) and then isolates these data elements in the target files. Splitting the house number from the address field into a 'street' field and a 'house number' field is an example of parsing. The next step is correcting the individual parsed data components, using algorithms and secondary data sources (reference data). Suppose, for example, that in the house number field alphanumeric values are identified, then you would like to identify and delete and/or split off these values. After correcting the data, the standardizing step follows. This means applying conversion routines to transform data into its preferred (and consistent) format using both standard and customized business rules (Boehmke, 2016). In this example, this could mean standardizing the format of the writing of the street name, by using 'st.' instead of 'street' or 'st'. After this step comes the matching. This means searching and matching records within and across the parsed, corrected, and standardized data, based on predefined business rules to eliminate duplications. The last step is consolidating all the efforts into one single representation.

4.4.3 Missing values and data fusion

As already pointed out, data will never be perfect. One of the major problems data professionals may face is missing values in the data. For example, there is no information on gender and age for some customers. One easy way to deal with these missing data is to throw away observations with missing values. This may, however, cause sample problems, especially when these missing values occur very frequently in a non-random fashion. For example, an organization that has customers in different geographical regions may find that, mainly for customers in a Northern geographic area, socio-demographic is missing because customers were not asked to provide these data. Missing values may also have an underlying reason. For example, customers may deliberately not provide data because they distrust an organization or are not very loyal customers (Vroomen *et al*., 2005). In the latter case, missing values can potentially be predictors of behavior. For example, one might assume that customers with missing values are more likely to defect, as they distrust the firm. Thus, in general, one should carefully analyze the reason for multiple missing values and consider whether these missing values occur in a random fashion or not. To solve the problem of missing values, there are different methods available. There are the so-called "naive methods," where one could delete observations with missing values or replace the missing values with, for example, the mean value. However, there are also more advanced methods such as “likelihood methods” (Dempster, Laird, & Rubin, 1977), where an estimate of the missing value is determined or “imputation methods,” where imputation means “filling in the data” (Van Buuren, 2012).

A related problem is that many firms have specific data for only a subset of the sample. For example, suppose satisfaction data are only present for 10% of the customer database. For the other 90%, no satisfaction data are available. One might like to know the satisfaction scores for those customers as well. In the same vein, one could have a share of wallet data for a sample of the customer base. Fortunately, methods have been developed by which the information in the sample can be used to estimate the missing data in the rest of the customer base. These are called data fusion techniques (Kamakura & Wedel, 1997). A rather simple form of these techniques is reported in Donkers and Verhoef (2001). They used a regression model to predict the total number of insurance products a customer purchases. This includes purchases at both the focal firm and outside it—i.e., with competitors. For more information on advanced data fusion techniques, we refer to Kamakura and Wedel (2003) and Kamakura *et al*. (2005).

4.5 CONCLUSIONS

In this chapter, we started by making the distinction between different data sources, grouping them in different ways: structured versus unstructured and internal versus external. We stressed that big data are not just about the external unstructured sources, but also have to deal with the other sources. In line with the other chapters in this book, we distinguished three types of data: market data, brand/product data, and customer data. For every type, there is a demand-side type of data and a supply-side type of data. Further, we discussed the added value of every source of data, by analyzing the different “Ws” (Who, What, When, Why, Where). The description of possible issues with data quality and missing values should help in tackling practical issues around working with data.

The assignment for this chapter is combined with the assignment of Chapter 5 and you will find it at the end of Chapter 5.

NOTES

1. Weblink: https://claritas360.claritas.com/mybestsegments/ (accessed February 18, 2021).
2. See https://www.kantarworldpanel.com/grocery-market-share/great-britain/snapshot/27.12.20/29.12.19

3. For more information on mentality groups: https://www.motivaction.nl/en/mentality/mentality-segmentation (accessed February 18, 2021).
4. See https://www.statista.com/statistics/1154513/intention-purchase-hybrid-car-electric-europe/ (accessed February 18, 2021).

REFERENCES

Boehmke, B. C. (2016). Data Wrangling with R. *Use R!* Cham, Switzerland: Springer.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38.

Donkers, B. & Verhoef, P. C. (2001). Predicting customer potential value: An application in the insurance industry. *Decision Support Systems*, *32*(2), 189–199.

Kamakura, W. A. & Wedel, M. (1997). Statistical data fusion for cross-tabulation. *Journal of Marketing Research*, *3*(4), 485–498.

Kamakura, W. A. & Wedel, M. (2003). List augmentation with model based multiple imputation: A case study using a mixed-outcome factor model. *Statistica Neerlandica*, *57*(1), 46–57.

Kamakura, W., Mela, C. F., Ansari, A., Bodapati, A., Fader, P., Iyengar, R., Naik, P., Neslin, S. A., Sun, B., Verhoef, P. C., Wedel, M., & Wilcox, R. (2005). Choice models and customer relationship management. *Marketing Letters*, *16*(3/4), 279–291.

Nair, R. & Narayayan, A. (2012). Getting results from big data: A capabilities-driven approach to the strategic use of unstructured information, Booz & Hamilton. Retrieved from www.strategyand.pwc.com/media/file/Strategyand_Getting-Results-from-Big-Data.pdf (accessed November 26, 2015).

Neslin, S. A., Grewal, D., Leghorn, R., Shankar, V., Teeling, M. L., Thomas, J. S., & Verhoef, P. C. (2006). Challenges and opportunities in multichannel customer management. *Journal of Service Research*, *9*(2), 95–112.

Prins, R. & Verhoef, P. C. (2007). Marketing communication drivers of adoption timing of a new e-service among existing customers. *Journal of Marketing*, *71*(2), 169–183.

Van Buuren, S. (2012). *Flexible Imputation of Missing Data*. New York: Chapman and Hall/CRC.

Venkatesan, R. & Kumar, V. (2004). A customer lifetime value framework for customer selection and resource allocation strategy. *Journal of Marketing*, *68*(4), 106–215.

Verhoef, P. C. & Donkers, B. (2005). The effect of acquisition channels on customer loyalty and cross-buying. *Journal of Interactive Marketing*, *19*(2), 31–43.

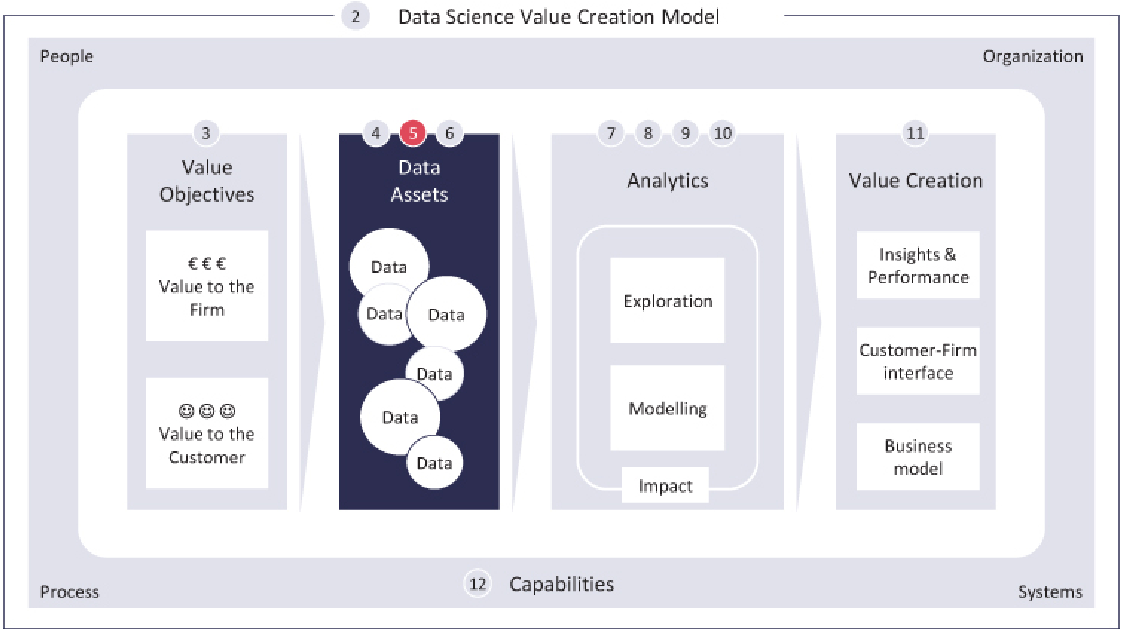
Verhoef, P. C., Spring, P. N., Hoekstra, J. C., & Leeflang, P. S. H. (2002). The commercial use of segmentation and predictive modelling techniques for database marketing in the Netherlands. *Decision Support System*, *34*(4), 471–481.

Vroomen, B., Donkers, B., Verhoef, P. C., & Franses, P. H. (2005). Selecting profitable customers for complex services on the Internet. *Journal of Service Research*, *8*(1), 37–47.

Zablah, A. R., Bellenger, D. N., & Johnston, W. J. (2004). An evaluation of divergent perspectives on customer relationship management: Towards a common understanding of an emerging phenomenon. *Industrial Marketing Management*, *33*(6), 475–489.

CHAPTER 5
Data storing and integration

DOI: 10.4324/9781003011163-5



5.1 INTRODUCTION

Although many people believe the biggest challenges in the big data era seem to lie in collecting data, we believe the real challenge lies in the storage and integration of all kinds of data sources to realize successful data value creation. Many of these data sources are not meant to be stored or built up for the purpose of integration with other data sources. In addition, the data sources (integrated or not) often contain data variables that need further processing to create useful information. In this chapter, we will discuss traditional data storing as well as data storing in the age of big data and the different steps that are necessary for data integration. The centralized storage environment for the different commercially relevant data sources from different departments or data warehouses/data marts is the integrated environment that we call the commercial data environment. Furthermore, we discuss in this chapter the process of creating marketing variables out of the different available data items within the commercial data environment.

5.2 STORING AND INTEGRATING DATA SOURCES IN DATA WAREHOUSES

The data warehouse is traditionally considered to be the central element of the commercial data environment. Data sources from inside and outside the firm provide data to the data warehouse. Within the data warehouse, the first step of data integration takes place.

5.2.1 Storing data in the data warehouse

Through a data gathering system, the data are transformed into an electronic medium. The data warehouse functions to:

1. Prepare the data for storage
2. Store the data
3. Describe the data
4. Manage and control the data.

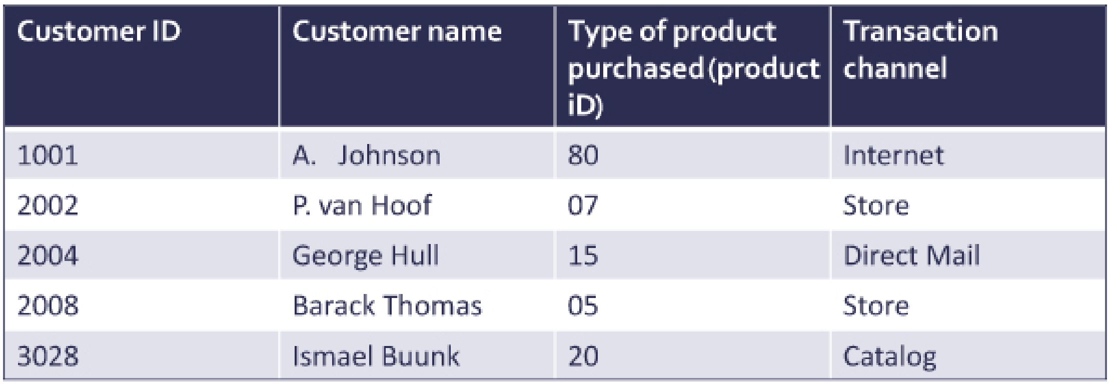
From the data warehouse, analytical databases or data marts are provided to customer intelligence specialists, data scientists, and database analysts (Zikmund, McLeod, & Gilbert, 2003).

Within a firm, several databases with relevant commercial data are usually present. Firms may have databases on point-of-sales information (e.g., store sales, sales through salespersons, products, invoicing, customer contacts, etc.). In a data warehouse for commercial purposes, all these data sources are integrated into one large customer database. The central focus of this database is the customer. However, it is not only customer information that should be retrieved from this database, but also other information, such as on salespersons' performance.

An important element of good databases is that the data are arranged in such a way that they can easily be retrieved by users. The structure of the database is called a data model, and this defines how the data elements are stored and which inter-relationships exist between these elements in a standardized way.

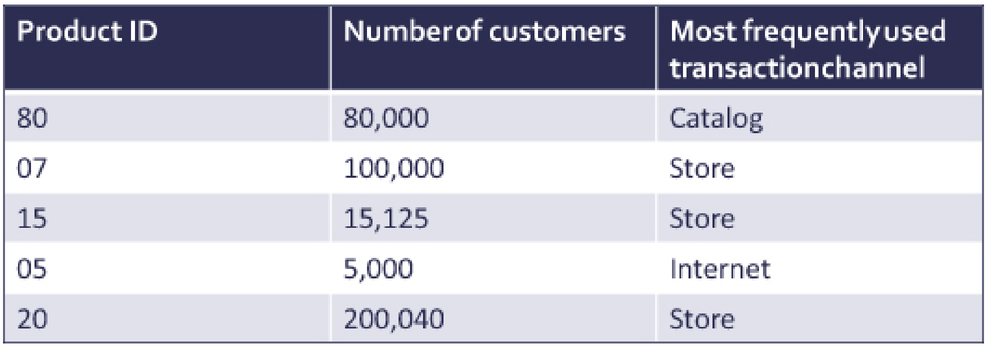
The desired database structure may, however, depend on the user. A sales manager is probably most interested in the performance of the sales reps, while a product manager most likely wants to know the performance of the products. A customer manager is most likely interested in customer metrics, such as customer profitability. To overcome this hurdle, relational database structures are currently standard. Relational databases (like the example in Figure 4.8) use different key-variables which link several databases to each other. For example, in an insurance context, the customer ID and the product ID are usually key variables. In a B2B context, the customer ID can again be a key variable, while the sales rep ID can also be a key variable. In a sales table, the salesperson ID would be the leading primary key variable, while in a customer table the customer ID is the leading primary key variable.

Figure 5.1 shows an example of the structure of a customer table. In this example, the customer ID is the leading primary key variable. For CVM purposes the customer, as the primary key variable, is of essential importance. The customer name field provides further information on that customer. The type of product is a second key variable, and the transaction channel is another. One can easily add more key variables, such as membership of a loyalty program, time, etc. From this table, one can in principle derive a database with the other primary key variables as leading key variables. For example, if one wanted to take the type of product as a key variable, a table could be created through aggregation and transformation procedures which shows per product the number of customers and the most frequently used transaction channels (see Figure 5.2)



Customer ID	Customer name	Type of product purchased (product iD)	Transaction channel
1001	A. Johnson	80	Internet
2002	P. van Hoof	07	Store
2004	George Hull	15	Direct Mail
2008	Barack Thomas	05	Store
3028	Ismael Buunk	20	Catalog

FIGURE 5.1 Example of simple data table with customer as central element



Product ID	Number of customers	Most frequently used transactionchannel
80	80,000	Catalog
07	100,000	Store
15	15,125	Store
05	5,000	Internet
20	200,040	Store

FIGURE 5.2 Example of product data table derived from customer database

Having multiple key variables, a database or table can be treated in a multi-dimensional way. For example, one can have two dimensions per customer, by knowing which product in which year (time) they purchased. Working with more dimensions quickly becomes more complicated. Customer information users generally find analyses at a high dimensional level difficult to understand, although software vendors have developed multi-dimensional databases in an effort to overcome this (Zikmund, McLeod, & Gilbert, 2003).

5.2.2 The data model in a data warehouse

Before starting to build the data warehouse and integrate the data, it is important to design the data model. As pointed out above it is commonplace to design a relational database structure. In an operational environment the database structure looks like Figure 4.8, a so-called Entity Relationship diagram (ER diagram). However, this type of relational structure is often not suitable for supporting all kind of analyses and reports, where a query should not take too much time to generate the desired output and where efficiency is also desired in loading and updating the data. To overcome this, most data warehouses use a “star schema” (Chaudhuri & Dayal, 1997), to represent multidimensional data. In a star schema there is a single table for each dimension and a central “fact table.” The fact table points to each of the dimensions. Since star schemas do not support attributes that come with a hierarchy, a special variant of the star schema, called snowflake schemas has been developed. In Figure 5.3 the three different types of data models are presented.

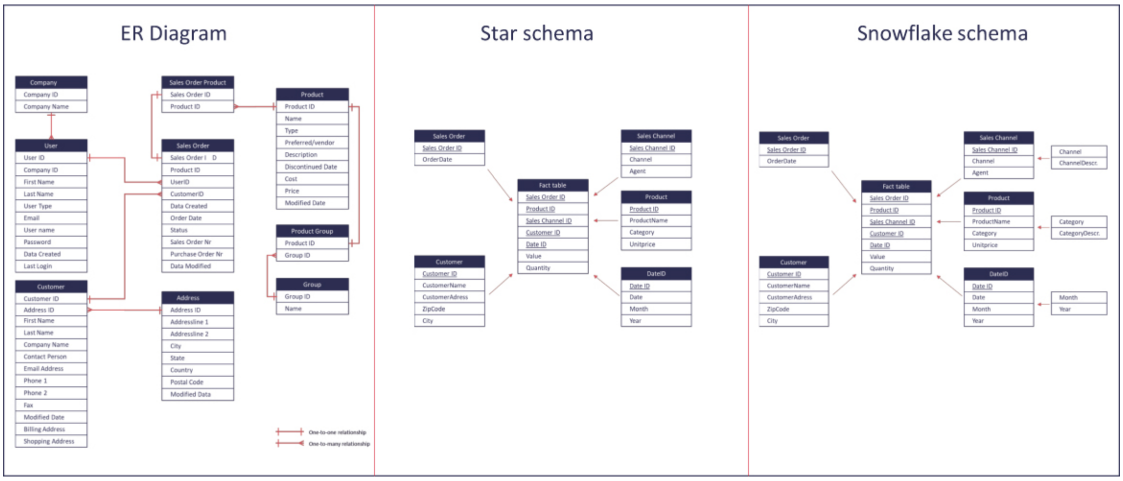


FIGURE 5.3 Example of an ER diagram compared to the star schema and the snowflake schema

5.2.3 Data integration into the data warehouse

The first step in the process of data integration is called ETL (see Figure 5.4); it is the process of extraction, transformation, and loading the input data sources into the data warehouse. This process of ETL can either be done by hand, or by using different types of programming (e.g., SQL), or using off-the-shelf software tools (e.g. Informatica).

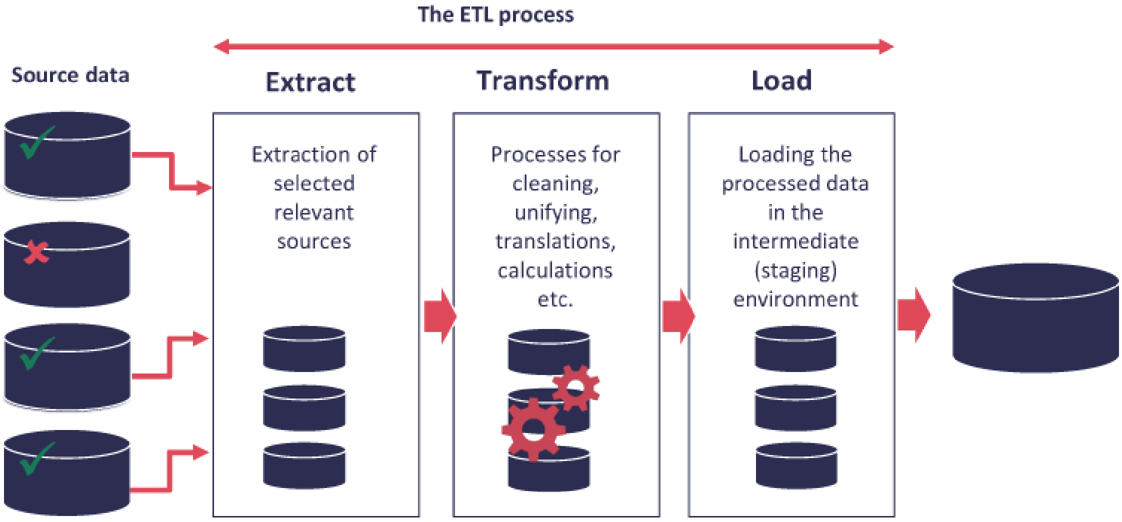

The ETL process
Source data
Extract
Transform
Load
Extraction of selected relevant sources
Processes for cleaning, unifying, translations, calculations etc.
Loading the processed data in the intermediate (staging) environment

FIGURE 5.4 The ETL process

5.2.3.1 Extraction

The crucial part of the extraction stage is the selection of the relevant data sources to be included in the data environment. Part of this process is validation: checking whether the data sources extracted are the right ones that are needed and were specified initially. This validation is necessary since time lags might occur in the input of data sources that might affect the further processing of the input data and even the possible relevance of the specific data source. Another aspect of the extraction step is how the data are extracted. This can be done via an automated query that runs at a specified fixed moment in time or it can be performed manually (not the best option because of possible human error and the amount of time it takes). However, in the first phase of setting up an integrated data environment and building up a sort of a "proof-of-concept" where the data still need to be extracted or are not yet fully stable, it may be preferable to perform several iterations by hand. The frequency of extraction of new datasets is also something that should be considered when organizing the loading process. Typically, for marketing information, a monthly, bi-weekly or weekly refreshment should be suitable (making a snapshot of the available data) although, especially in highly dynamic environments, a higher frequency could be considered. Real-time refreshment does not seem feasible from a technical perspective (there will always be a delay in data processing) and also from a business perspective. Analysts should ask whether the extra benefit gained from having more up-to-date data is worth the cost of collection (e.g., Neslin *et al.*, 2006). Those insights that are needed in real-time can be triggered in an operational environment by implementing business rules and/or algorithms to make maximum use of collected insights and can also be derived by real-time reporting on crucial KPIs.

5.2.3.2 Transformation

After extracting the data, the next step is transforming the different data sources to fit them into the commercial data environment by applying all kinds of business logic. The transformation of the data is often designed to make it easier to handle (more storage efficient and more robust in data quality). Typical transformations are selecting certain data attributes, replacing values in the data (e.g., “M” instead of “Male”), translating values from character into numeric (e.g., “1” instead of “active customer”), calculating values (revenue excluding VAT = revenue divided by 1.21), or extracting and splitting variables into different variables (e.g., taking out the house number and street name of the address field and putting it into two new variables). All steps in the transformation stage are designed to make the data easier to handle during the analysis, make storage easier, and make integration with other data sources more straightforward (by making the data sources uniform).

5.2.3.3 Loading

After transforming the data, they will be processed by loading them into the desired data environment. This will normally be done with an intermediate step called staging. During staging, all data sources are loaded in an intermediate environment before being loaded into the tables of the desired data environment. The benefit of this is that it makes it possible to do final checks before publishing the data into the environment where users start working with it or where reports are generated. The staging area also serves as a sort of a backup in case something goes wrong when loading into the final tables. Most of the time the data in the staging area is intermediate and will be erased when it is no longer needed.

There are two ways by which the data can be loaded into the final data environment. The first is by overwriting the earlier dataset in the commercial data environment. However, this could mean losing all kinds of historical data and is not the preferred option. Another way is to append the new data during the loading process to the existing data in the commercial data environment. This way historical data can be kept and new data points are added each time the data are refreshed.

5.3 STORING AND INTEGRATING DATA SOURCES IN DATA LAKES

In the era of big data, with the enormous increase in the volume and variety of data, new technologies have arisen to deal with these developments, especially regarding storing and processing unstructured data in so-called data lakes.

A recent development in the last few years, boosted by big data, is the rise of data lakes. An increasing number of organizations are adopting this technology to capture the opportunities of big data. A good definition for data lakes can be found at Gartner.[1]

> *A data lake is a concept consisting of a collection of storage instances of various data assets. These assets are stored in a near-exact, or even exact, copy of the source format and are in addition to the originating data stores.*

A different definition from Amazon (Amazon, 2021)[2] focuses on the unique features of data lakes:

> *A data lake is a centralized repository that allows you to store all your structured and unstructured data at any scale. You can store your data as-is, without having to first structure the data, and run different types of analytics—from dashboards and visualizations to big data processing, real-time analytics, and machine learning to guide better decisions.*

There are some big differences in data lakes compared to data warehouses. Below you will find an overview of the most important ones (Campbell, 2015; Raisinghani, 2019).

5.3.1 Data

What makes data lakes different from data warehouses is, in particular, how data lakes are capable of dealing with unstructured data. Big data has grown and developed through the increased diversity of data sources and it is mainly the unstructured sources that have exploded in number and size. Different from data warehouses, data lakes are built to store all data, without the explicit need to process it and they can do this without the typical ETL-process. This, however, means that the total volume of data in data lakes is much larger.

Furthermore, data lakes support all types and formats of data. Data is stored in its raw format and is only transformed when needed for use.

5.3.2 Data model

Another difference lies in the fact that data lakes do not require a predefined data model where extracted, transformed, and loaded data is stored. The organization stores the data as is and at a later stage, when specific data items are needed, the desired data model for a specific task or analysis is defined.

5.3.3 Users

The users of a data lake also differ from the typical users of a data warehouse. Due to the fixed structure of data warehouses, their users are more the BI people or business analysts whereas the data lakes are mainly being used by data scientists who explore different data sets.

5.3.4 Analytics

The difference in the types of users also translates into the analytics applied. Where data warehouses focus on standard reporting, on a fixed regular basis, data lakes are better suited to ad hoc questions, exploring data, and discovering unexpected relationships.

5.3.5 Price/quality

Although data lakes are a particular product of cloud working, companies can also build them themselves within their own organizations. There are also hybrid solutions or hybrid computing (partly cloud, partly on premises). Large providers of cloud solutions, with the potential to create data lakes are Amazon (with AWS), Microsoft (Azure), and Google (Google Cloud Storage). These companies are highly competitive in the rates they charge for data storage and offer a good price/quality ratio. On the other hand, most of the data warehouses are created and maintained within organizations. Due to their specific way of working, including ETL, backup facilities, technical support and necessary hardware, data warehouse solutions are costly and come at a much higher price than data lakes in the cloud. Yet, querying on data warehouse data is still often faster than querying on a data lake.

5.3.6 Flexibility

Whereas data warehouses cannot easily be changed to adapt to new requirements, data lakes are very flexible and can be changed much more easily, due to the raw format the data is stored in. As a result, data lakes can provide insights much faster than traditional data environments. This makes them very useful for data scientists who are exploring a wide set of perspectives in order to answer the questions they have.

5.3.7 Technology

The technology used for data lakes is also different (OvalEdge, 2021). In data warehouses where a SQL environment is dominant, we see that Hadoop with its Distributed File System (DFS) and Hadoop Clusters is almost synonymous with data lakes. In cloud solutions, Hadoop is linked to the proprietary storage environment of the cloud provider (like Amazon), whereas in data lakes on premise an open-source Hadoop File System is applied. Besides Hadoop, there are also “Apache Spark Clusters,” which offer a data lake solution in the cloud or in house. Compared to Hadoop this claims to be much faster, due to in-memory computing.

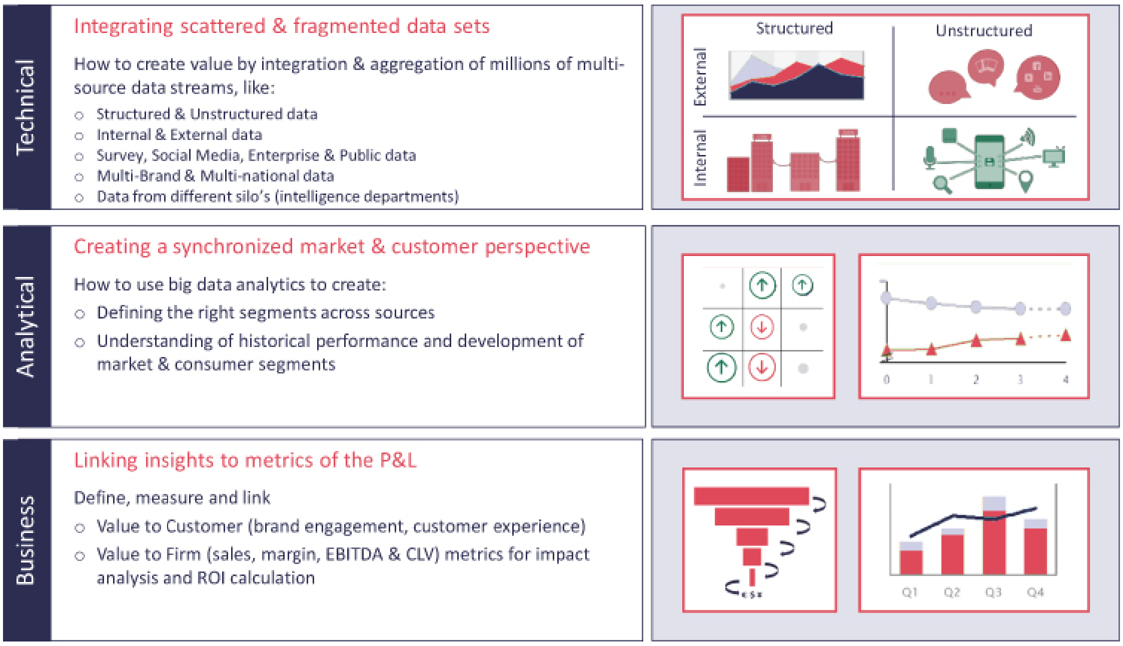
What can be concluded from the above is that data lakes offer unique features for working with large, diverse, and unstructured datasets and provide excellent opportunities for experimenting. This explains why an increasing number of organizations are choosing data lakes. Organizations starting from scratch might not consider data warehouses any longer. However, organizations that already have data warehouses in place may very well decide to have both data warehouses and data lakes (Blumberg et al., 2021). Due to the differences between data lakes and data warehouses, the latter might still serve a purpose, for example, in providing standardized reports for large parts of the organization, based on highly structured data.

5.4 CHALLENGES OF DATA INTEGRATION IN THE ERA OF BIG DATA

The ETL steps described when discussing traditional data storage are typical for a conventional data warehouse in a commercial setting. The data in a data warehouse for commercial purposes often have a customer/prospect-centric approach, where all data are being collected to link to individual customers or prospects at the individual or household level. Each table in the data warehouse contains a key data field that ultimately links them to the customer level. By combining all this data—ideally into one single table or flat file—from the data warehouse or data mart (a data warehouse or subset of a data warehouse for special purposes) with still rather raw data, the dataset needed for analysis becomes richer, especially if, as described above, new variables are created. When data is brought to the commercial data environment from data lakes, where there is no ETL process in place, *ad hoc* processing of the data, similar to ETL is needed. The next step is to deal with several data sources that can also be stored and combined in the commercial data environment but either contain information of only a subset of customers or were collected with another aggregation level in mind (as was the case for the market and product data described in Chapter 4). The technical challenge of dealing with the different aggregation levels, however, is not the only challenge in making good use of these combined sources that are not integrated. In this section, as well as discussing how these data sources with different aggregation levels can be technically integrated, we will also discuss other challenges. We intend to look at a new way of analyzing these sources to work out how to integrate them.

Non-integrated customer, market, and product data are typically used by different departments in the organization making analyses and reports on these data, each from the perspective of their own "silo." This often results in lots of confusion and annoyance with the end-users of the outputs. End-users make comments like: "We have tons of data, so why is it taking so much time to create the right insights when we need them?" or "Why do we have to gather our crucial marketing insights from so many different departments within the organization?" or "We are overloaded with reports and overviews, but they don't give us input on how to improve our business performance" or "Although I now understand what has happened, please tell me also how to act."

Analyzing these comments, we see there are three challenges organizations have to deal with to make maximum use of integrated data (see Figure 5.5):



Technical

Integrating scattered & fragmented data sets

How to create value by integration & aggregation of millions of multi-source data streams, like:

- Structured & Unstructured data
- Internal & External data
- Survey, Social Media, Enterprise & Public data
- Multi-Brand & Multi-national data
- Data from different silo's (intelligence departments)

Analytical

Creating a synchronized market & customer perspective

How to use big data analytics to create:

- Defining the right segments across sources
- Understanding of historical performance and development of market & consumer segments

Business

Linking insights to metrics of the P&L

Define, measure and link

- Value to Customer (brand engagement, customer experience)
- Value to Firm (sales, margin, EBITDA & CLV) metrics for impact analysis and ROI calculation

FIGURE 5.5 The challenges of data integration

- Technical challenges: scattered and fragmented datasets
- Analytical challenges: lack of a similar and synchronized customer/consumer perspective and consistent segmentation; explaining

the past instead of predicting the future; and how to analyze with datasets at different aggregation levels

- Business challenges: no link with metrics of the P&L to realize alignment and acceptance from the financial department.

5.4.1 The technical challenges of integrated data

The technical challenges of data integration are mainly caused by different aggregation levels of data sources, but also by the fact that, as in our example above, the customer data is likely to be stored in the CRM data warehouse, accessible by analysts from the customer intelligence department, while the market data is stored by the research agency (only distributing a report to its client) or within the market research department. So, there are two technical challenges, namely first the extraction and collection of relevant data sources and next combining these sources. The first challenge is to realize a centralized storage environment for the different data sources from the different departments; the integrated environment and the commercial data environment. The second challenge is to physically integrate the data sources, using relevant keys, to enable analysis across all these data sources.

The first technical challenge can easily be resolved by specifying the desired data extracts and downloading them into a dedicated storage environment. This can be a dedicated server or a stand-alone piece of hardware, like a powerful pc. Especially in the phase where a proof of concept is needed in a pilot for a big data initiative, the latter option can be an affordable alternative to a fully-fledged new commercial data environment.

The second challenge is the physical integration of the data sources. What are the keys that can be used to combine the different data sources into one single, flat file and that can deal with differences in the aggregation level of data sources. There are several ways of tackling this, each with its advantages and disadvantages. The different options are mainly driven by the aggregation level of the data source. The data source at the highest aggregation level can be considered "the weakest link," because it determines the aggregation level all other data sources have to align with when KPIs are analyzed and compared, coming from different data sources. In our example above concerning the insurance company, it is very likely that the market data typically analyzed at the market player level is the highest aggregated data source.

5.4.1.1 Integration at the individual level

The first option for technical integration of the data sources is what we call "integration at the individual level." This means that the individual (or read "household" for "individual") is identified in every data source and that we start linking the different data sources using the keys identifying the individual: the keys for combining two sources should correspond. Integrating the sources will usually start with the largest database at the individual level. The first step is thus integrating this database (often the customer database) with the next largest database. Let us assume in our example that we want to integrate customer data with online data. In the customer data, the individual can be identified by a customer ID, or for example a combination of address and birth date. The next step is to identify in our online data the right key for integrating individual online users to the customer base. In our online data, we will often find the data stored at the visit level. So, the first thing to do is to identify unique visitors. This should result in a unique visitor ID a combination of a cookie ID and an IP number (and for registered users who are customers the unique key could be the customer ID). The registered users should be integrated with the customer database by using the customer ID. The customers who were not online should be stored with the customer ID as the unique identifier and the non-customer website visitors should be stored with the unique visitor ID. Since integration at the individual level involves privacy issues and sensitive information, one should consider specific measures (for example pseudonymizing, as described in Chapter 6).

5.4.1.2 Integration at the intermediate level

The second option, integration at an intermediate aggregation level, uses a segmentation based on dimensions that can be identified in all sources to be integrated. The segmentation then becomes the common denominator to which all data should be aggregated for every time period. In our example of the insurance company, we could define the dimensions as age and income. Classifying each of these two dimensions in five classes would result in 25 segments to be identified in every data source per time period. If a data source does not have this information, this problem could be solved by first adding external data based on zip code or household level (e.g., from sources like Experian).

5.4.1.3 Integration at the time level

The third option for data integration is the least advanced. In this option, data will be aggregated to the time period that can be identified in the data sources and the time axis will be the dimension on which to compare the different data sources.

5.4.1.4 Mixture of integration options

The fourth option is perhaps the one that is used most—a combination of options 1, 2, and 3. In this option, every data source to be integrated will first be checked for available possible unique identifiers. Based on this assessment it will become clear which will be the best of the above three options to use for each data source.

5.4.2 The analytical challenges of integrated data

As well as the technical challenges to be addressed, we also see four analytical challenges. The first is the synchronization of the customer and consumer perspectives (as discussed in Chapter 4) to be used in the integrated dataset and in the segmentation to be used across and within the different datasets. Defining the customer level and the consumer level can be difficult. One important decision to be made is whether to define this at the household level or the individual level. Once that is decided the next problem is how to identify this level consistently across the different data sources. This asks for internal alignment within the organization and departments and clear business logic to apply the final choices. Assuming this to be the case, the next analytical challenge to be dealt with is the choice of the dimensions and calculation of the segmentation to be used. When using segmentation in the integrated data, either for the integration itself or for analyzing across segments across the data sources, a uniform way of segmenting the data is needed. Just as with the decision on what customer/consumer definition should be used, segmentation (for example based on crossing age groups with income groups) is also a topic that should be agreed, discussed, and applied in the right way within the organization. The third analytical challenge around integrated data is the type of insights to be generated. Since integrated data, combining customer, market, and brand input (see Chapter 4) should offer an extra benefit over reporting on the different data sources separately, we think the ultimate goal should not be trying to explain what has happened. The real benefit should be being able to make a prediction of what will happen next or give guidance ("prescription: what should we do?", see Chapters 7 and 9) by creating a holistic view of the marketing performance and identifying the levers that explain historical, current, and future performance. The fourth analytical challenge lies in how integrated data can be used for modeling purposes, especially since the aggregations for physical data integration can lose some of the modeling power of data. We will go into this in more detail in Chapters 7–9, when we discuss analytics.

5.4.3 The business challenges of integrated data

From a business perspective, we see two different challenges around integrated data. An important business challenge is defining the right key performance indicators (KPIs) within and between the different data sources. The KPIs to be defined should reflect the marketing performance of customers, brand, and market, and should indicate opportunities for improvement. The next challenge is to link the KPIs to the company's P&L.

5.4.3.1 Dealing with different data types

Data sources vary in terms of content, scaling, source, and presence within the commercial data environment. It is our strong belief that instead of focusing on the "V" (volume) of big data, the real challenge is addressing the "V" (variety) of data sources, especially as we believe that each source adds a specific dimension to the commercial data environment. Broadly speaking, we distinguish four data types (see also Figure 5.6):

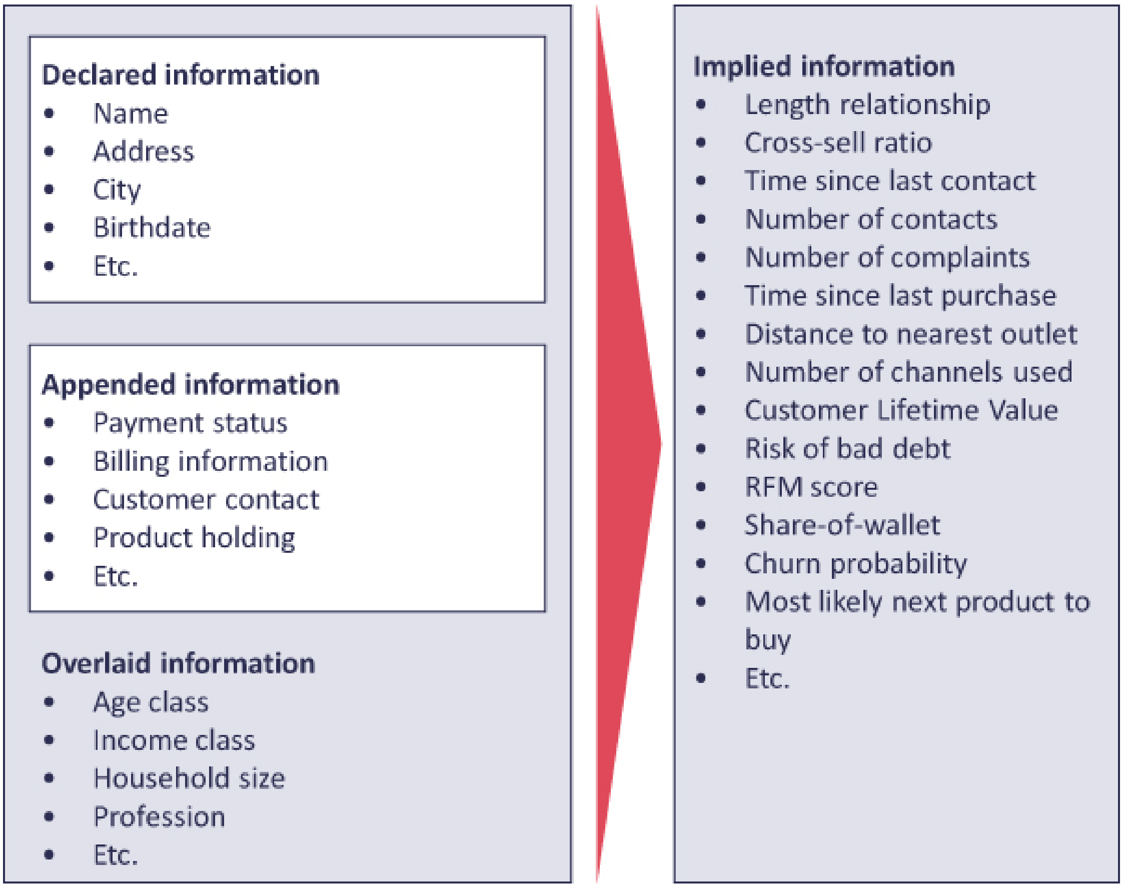


FIGURE 5.6 The different data types

- Declared data (customer descriptors)
- Appended data (transaction and billing data, customer contact data, marketing contact data, customer characteristics, customer attitudes, brand performance data)
- Overlaid data (zip code, household data, and research data like brand performance, customer attitudes, market, and competitor data)
- Implied data (segmentation, scoring models, share of wallet, recency frequency monetary value (RFM) classification, etc.).

5.4.3.2 Declared data: customer descriptors

Customer descriptors include all characteristics necessary to contact customers, such as name, address, zip code, phone number, and email address. With the help of this information, marketing campaigns can be targeted to individual customers and invoices can be sent to customers.

5.4.3.3 Appended data

Appended data refers to all data on customers' (financial) behavior and attitude towards the firm and can include variables such as the last moment of purchase, the type of product purchased, the monetary value of the purchase, transaction channel, and product returns. Another source of appended data is customer contact data (contact history), which results from customer-firm interactions that are initiated by the customer and may relate, for example, to information requests, complaints, clarifications on invoices, website visits, and contact channels (phone, email, etc.). Yet another source of appended data consists of marketing contact data that result from firm-initiated contacts and may include the number of mailings sent, the timing of mailings, and details about loyalty program membership. Customer characteristics are also part of appended data and cover additional information on the customer with regard to socio-demographics, psychographics, lifestyle, etc. Some of this data may come from internal sources. For example, health insurance requires customers to provide basis characteristics like their gender, age, household size, etc.

Other types of appended data that can come from other sources are customer attitude data, such as data on customer satisfaction and other attitudes, like commitment or net promoter score (NPS). These data are usually collected by carrying out surveys among samples of the customer base. As a result, this information is usually not present for all customers. Brand performance data (i.e., brand awareness, brand preference, etc.) are measured for customers and non-customers and are only available for a subset or sample of the customer base.

5.4.3.4 Overlaid data

External data suppliers also provide information, at various aggregation levels. A specific example of external profiling analysis is the incorporation of zip codes. External data providers, such as Acxiom and Experian, have specific zip code (or even household) information. Using this information, firms can gain insight into which zip codes (and thus local/regional areas) are home to a larger or smaller proportion of their customers. Along with this zip code/household information, these external data suppliers have also developed information on specific segments, such as "rural families" and "one-person households.” In Figure 5.7 we provide an overview of the segments used by Experian UK. Firms can use this information to further profile their customers and customer groups externally. For example, an online retailer may find that “rural families” are over-represented in their customer base, while “one-person households” are almost completely absent.

In Figure 5.8 we show a concrete example of a clothing retailer that has many stores in larger villages outside cities. This retailer wants to compare its clientele with the wider population. As can be observed from the analysis, “professional rewards” and “claimant cultures” have the highest over-representation in the customer base. The stores do not attract segments such as “industrial heritage.”

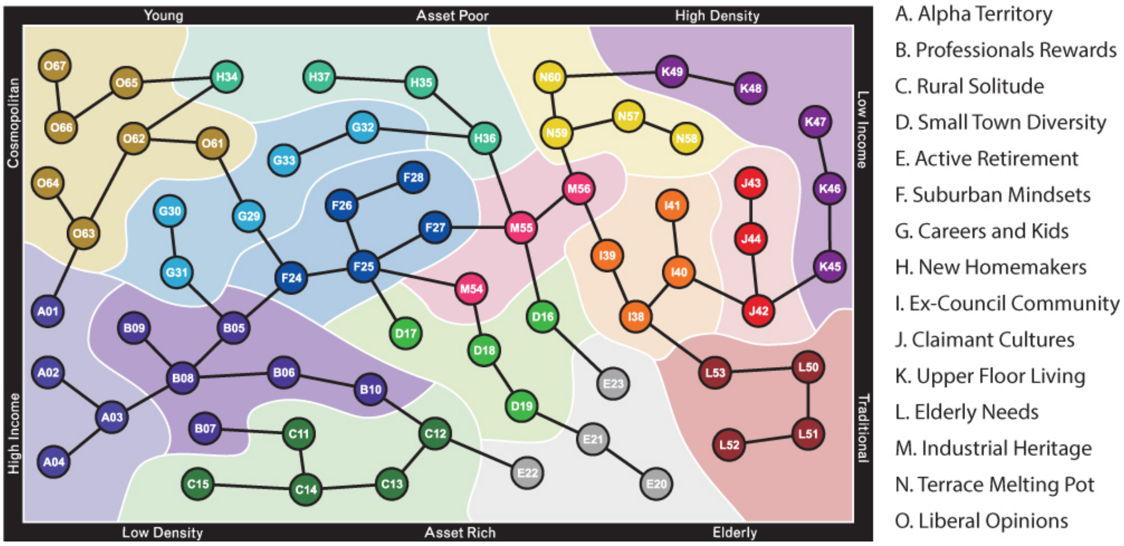
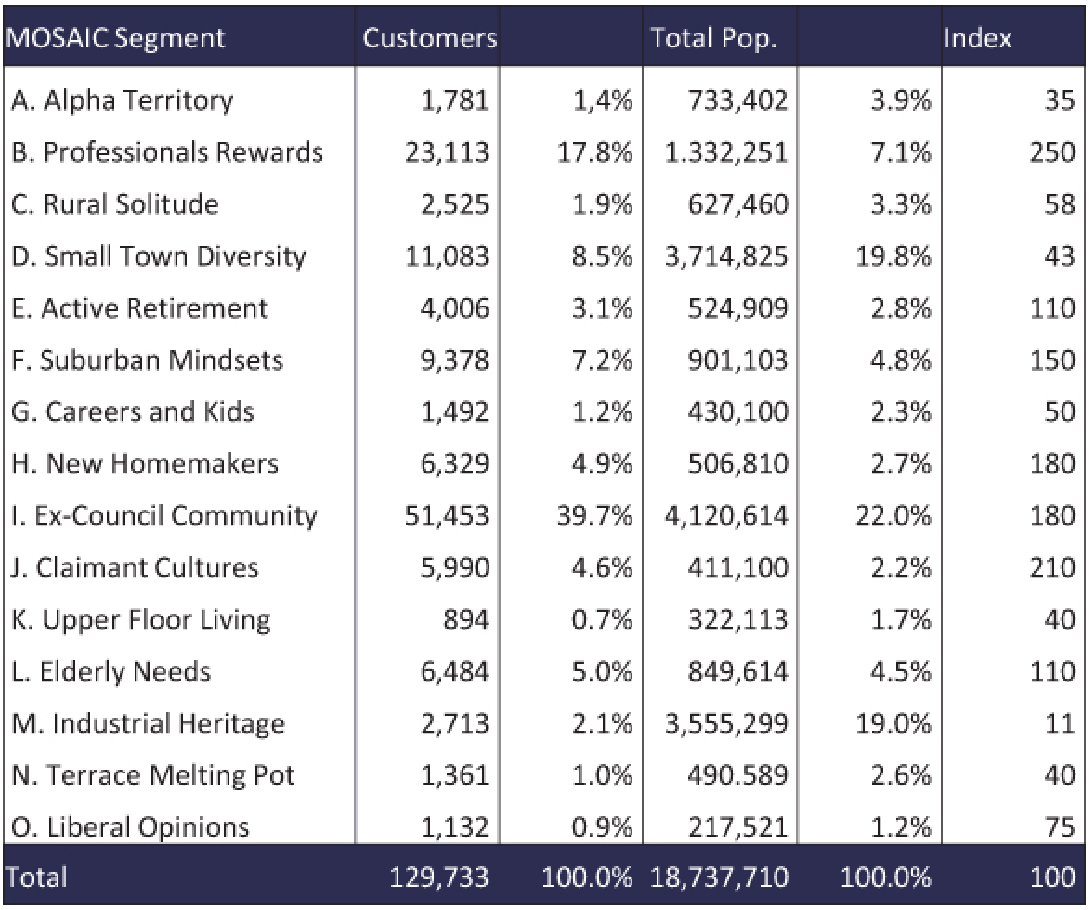


FIGURE 5.7 Overview of segmentation scheme used by Experian UK



MOSAIC Segment	Customers		Total Pop.		Index
A. Alpha Territory	1,781	1,4%	733,402	3.9%	35
B. Professionals Rewards	23,113	17.8%	1.332,251	7.1%	250
C. Rural Solitude	2,525	1.9%	627,460	3.3%	58
D. Small Town Diversity	11,083	8.5%	3,714,825	19.8%	43
E. Active Retirement	4,006	3.1%	524,909	2.8%	110
F. Suburban Mindsets	9,378	7.2%	901,103	4.8%	150
G. Careers and Kids	1,492	1.2%	430,100	2.3%	50
H. New Homemakers	6,329	4.9%	506,810	2.7%	180
I. Ex-Council Community	51,453	39.7%	4,120,614	22.0%	180
J. Claimant Cultures	5,990	4.6%	411,100	2.2%	210
K. Upper Floor Living	894	0.7%	322,113	1.7%	40
L. Elderly Needs	6,484	5.0%	849,614	4.5%	110
M. Industrial Heritage	2,713	2.1%	3,555,299	19.0%	11
N. Terrace Melting Pot	1,361	1.0%	490.589	2.6%	40
O. Liberal Opinions	1,132	0.9%	217,521	1.2%	75
Total	129,733	100.0%	18,737,710	100.0%	100

FIGURE 5.8 External profiling using zip code segmentation for clothing retailer

Source: Adapted from Experian, UK

Besides data from data providers on the zip code or address level, there are also providers of market data and competitor data. They provide data about market share, market volume, volume share, as well as information about the performance and market position of competitors. This can be considered overlaid data as well.

5.4.3.5 Implied data

Finally, firms can derive data by combining all these other data. Important derivative variables include share of wallet, propensity to buy scores, credit scoring, churn probability, and customer lifetime value (CLV). In fact, derived data can also be considered implicit information. The variables created are based on calculations, assumptions, and combinations of data sources. Here, the creativity and capabilities of data-scientists are crucial in creating a competitive edge in the use of data.

Verhoef *et al*. (2002) and a replication study (Verhoef, Hoekstra, & Van der Scheer, 2009) have investigated the presence of data types for Dutch firms.

Customer descriptors and transaction data are the types of data most often stored in customer databases. Between 2003 and 2008, a strong increase in the presence of all types of variables was observed (see Figure 5.9). This reflects the fact that firms have succeeded in setting up complete data warehouses which can now integrate information for all kinds of databases.

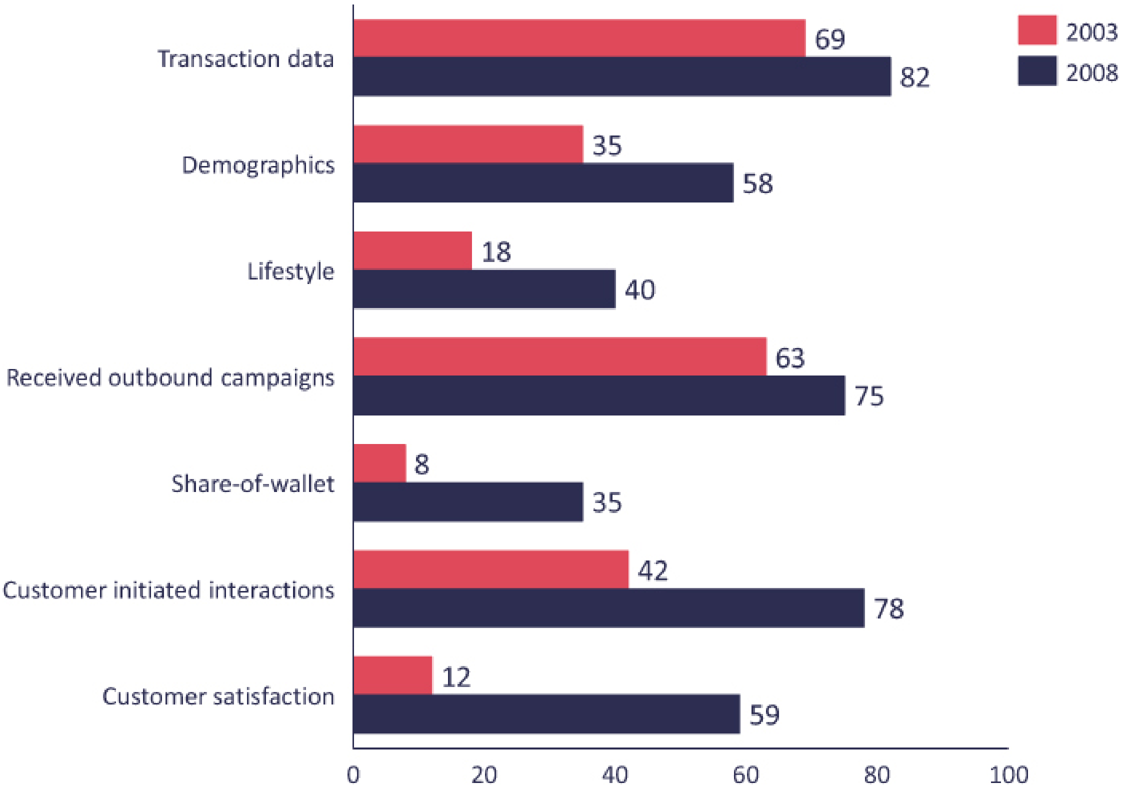


FIGURE 5.9 Presence of data types for Dutch firms

Sources: Verhoef et al. (2002) and Verhoef, Hoekstra, and Van der Scheer (2009)

An important element of the commercial data environment is the customer database. Today, firms have very large databases of customer data. These data come from a variety of sources. In the past, firms typically had multiple databases with different kinds of contents, which were used in business processes such as sending invoices and handling incoming phone calls. In addition, the data were usually structured around products rather than customers. These databases were not integrated. Instead, individual customer data were fragmented over many databases. As a consequence, firms did not have detailed individual customer information and could not get the full picture of what actually happened with each customer over time. Due to the integration of databases in many large-scale CRM projects, a single integrated customer database is now frequently available.

CASE 5.1: DATA INTEGRATION FOR AN INSURANCE COMPANY

In this case we will describe how a large insurance company, selling car, home and travel insurance, started with data integration as part of a bigger challenge. This exercise with data integration was meant as a proof-of-concept, to show the added value of the integrated data sets. The bigger challenge was to create a holistic marketing approach to monitor and to improve marketing performance. Management needed insights to explain the current performance of the organization and its competitive strength, including the levers for performance improvement and actionable initiatives.

In Chapter 11 we will revisit this case to describe the actual use of the integrated data, in what we call “the holistic marketing view,” realizing the desired improvement in marketing performance.

The complications in this data integration challenge were that data sources were scattered within and outside the organization, that they were not extracted and collected on a regular basis and that the data sources in view were not at the same aggregation level.

First, we will discuss the necessary sources in view and next we will describe how we integrated each data source and created the relevant analysis variables/KPIs.

The input data sources to be considered for integration were:

- Customer data
- Market performance data
- Brand performance data
- NPS Data
- Pricing data.

Customer data

The customer data is available at the individual customer level, with a customer-ID and the age of the customer, the products the customer bought, when the product was bought, where it was bought and the premium to be paid per product. Also, the CLV per customer is stored in this dataset. The customer data is enriched with external data from a zip code data provider.

Market performance data

We also have market data available per brand for all market players per moment in time (on a quarterly basis), indicating the average value per brand per product per customer and the number of customers per product per brand. The market data have been collected through market research via a random sample of the population of consumers with insurances, repeated on a quarterly basis. In this dataset basic demographics are also available.

Brand performance data

Each month, an external research agency measured the brand performance for a representative set of consumers via a panel. For the panel respondents the income was based on the external zip code data, while age was one of the questions in the survey. Typical measures collected were spontaneous brand awareness, brand consideration, brand preference and ad recall.

NPS data

NPS is measured on a monthly basis for a panel that consists of a random sample of the customer base. Here we measure the so called relational NPS. Since the panel consists of customers who have approved the storing of their data for analytical purposes and who are not anonymous, we can link this data with the customer data.

Pricing data

The pricing data is collected by storing data from a comparison site, where the prices and rankings of all insurance providers in the market are stored on a continuous basis. In this data age is also available. This data is enriched with zip code data as well.

All the data above was collected for the last two years, ideally at the monthly level and for some on a quarterly basis.

Since not all the datasets were at the same aggregation level, we had to define a common denominator. For the insurance business and then especially insurance for car, travel or home, income and age are strong determining dimensions for pricing and commercial decision making. That is why we chose to combine age (in six age classes: < 25, 25–34, 35–44, 45–54, 55–64 and >65) with yearly gross income (in five classes: unknown, <30k euros, 30–45k euros, 45–60k euros and > 60k euros). This resulted in 30 combinations.

In every source we had the availability of the dimensions age and income; for some we had to rely on zip code data to collect income data.

The next step was integrating the different sources with their history over the last two years. For the customer data we aggregated on the combined key of age-income classes, year-month, and acquisition channel. For every record of the combined key, we made the following variables:

- Total number of active customers (new + existing)
- Total number of churned customers
- Total number of new customers
- Total premium value active customers
- Total CLV of active customers
- Number of products of active customers
- Number of active customers per product (car, home and travel)
- Number of new customers per product (car, home and travel)
- Number of churned customers per product (car, home and travel)
- Total premium value per product of customers per active product.

For the Market Performance data, we aggregated on the combined key of age-income class and year-quarter and we used the included weight field in the sample data to represent in the aggregate the total market population. For every record of the combined (weighted) key we made the following variables:

- Number of households
- Number of customers per market player (for the top 10 largest)
- Number of customers per market player per product (for the top 10 largest)
- Total premium value
- Total premium value per market player (for the top 10 largest)
- Total premium value per product (for the top 10 largest).

From the NPS data we aggregated on the weighted key (to represent all customers) of age-income class and year-month. This resulted in the following variables:

- NPS per customer
- NPS per customer per product
- NPS per customer per acquisition channel.

The brand performance data was also aggregated on the weighted key (to represent the market) of age-income class and year-month. This resulted in the following variables:

- % spontaneous awareness per brand (top 10 players)
- % brand consideration per brand (top 10 players)

- % brand preference per brand (top 10 players)
- % ad recall per brand (top 10 players)
- % spontaneous awareness per brand per product (top 10 players)
- % brand consideration per brand per product (top 10 players)
- % brand preference per brand per product (top 10 players).

The pricing data was aggregated on the key “age-income class” per year-month. This resulted in the following variables for the top 10 companies:

- Company with highest ranking per product
- Company with second ranking per product
- Company with third ranking per product
- Etc.

Now that we have defined per source the KPIs to be combined as well as the unique key for integrating the data, we can actually combine all the data sources into one file on the key “age-income class year-month.” For the data sources that are built up per quarter we repeat the quarter value for every single month. This is the starting point for all the analyses on the relationship between the different KPIs as discussed in Chapter 11.

5.5 CONCLUSIONS

In this chapter, we discussed how organizations can organize their data storing, taking a more traditional approach by using data warehouses or adopting new ways of doing this (driven by big data), using data lakes or a combination of a data warehouse and a data lake. In particular, data lakes are very capable of dealing with large volumes of unstructured data. Furthermore, we discussed three data integration topics: the ETL process, the process of creating all kinds of new variables by combining datasets, and the various challenges surrounding the integration of data sources with different aggregation levels. While all three topics are critical to making big data integration a success, we consider the last topic most critical (and especially the technical challenge of dealing with different levels of aggregation) as it is driven by all developments around big data. Solving the challenges surrounding different levels of aggregation is the way to deal with the ‘V’ of variety in big data and is thus an important success factor in realizing the potential of big data.

ASSIGNMENT (CHAPTERS 4 AND 5): SUPERSTORE

The Online retailer SuperStore sells a multitude of products and brands. The brands in their collection are mainly premium brands. One of the most important categories is clothing. SuperStore is the market leader in this category online.

Currently, SuperStore has more than 1.5 million active customers. Active is defined as placing at least one order in the past six months.

SuperStore is currently busy with preparations for the coming summer season. In particular, the sale of fashionable swimwear has become an important spearhead. This year, they want to grow 25% in revenue here.

Up until now, the data environment has been developed to only a very limited extent. SuperStore has hired your team, with swimwear as a pilot, to demonstrate the added value of big data and analytics.

Questions:

1. Given the ambition of SuperStore to grow in revenue by 25% next season, which analysis questions can you come up with, using the 5 W questions (Who/What/When/Where/Why)?
2. Which data sources/tables would you like to collect (at least 1 per quadrant of the matrix internal/external and structured/unstructured) to build an analysis environment for SuperStore (e.g., think of a data source/table like "orders")? Use Figure 4.1.
3. Which variables would you at least want to collect per source/table given above (e.g., think of a variable like 'order date')? State this as exhaustively as possible.
4. One of the sources that is already being used is weather data from the KNMI (Royal Dutch Meteorological Institute). Name at least two examples of possible applications of KNMI data, including weather data.
5. Indicate for each analysis question from question 1 which data sources you will use for this.

   The second part of this assignment focuses on data integration. Use the following principles:

   > The analysis file must be built at customer level, in order to analyze in SPSS or R.
   >
   > The ID/key in this analysis file is the customer number and the final analysis file is unique to this.

6. Indicate for each source mentioned above (from question 2) which edits to this source are required in order to be able to add it to your analysis file. In addition, make a schematic representation (also called a data model) demonstrating how the data sources relate to each other.
7. Indicate for the named variables in question 3 which edits are required on these variables. Run the analyses in your analysis file. (For example, you

can edit the order date field from the orders table to find out the first order date of a customer, compare it with the current date, and thus calculate the time in days since the last order. Finally, this time in days can be classified into a number of categories.)

NOTES

1. Weblink: https://www.gartner.com/en/information-technology/glossary?glossarykeyword=data%20lakes
2. Weblink: https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/

REFERENCES

Amazon. (2021). *What is a data lake?* Amazon. Retrieved from https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/.

Blumberg, S., Machado, J., Soller, H., & Tavakoli, A. (2021). *Breaking through Data-architecture Gridlock to Scale AI*. McKinsey & Company.

Campbell, C. (2015). *Top Five Differences between Data Lakes and Data Warehouses*. Blue Granite. Retrieved from https://www.bluegranite.com/blog/bid/402596/top-five-differences-between-data-lakes-and-data-warehouses.

Chaudhuri, S., & Dayal, U. (1997). *An overview of data warehousing and OLAP technology*. *ACM Sigmod record*. 26(1), 65–74.

Neslin, S. A., Grewal, D., Leghorn, R., Shankar, V., Teeling, M. L., Thomas, J. S., & Verhoef, P. C. (2006). Challenges and opportunities in multichannel customer management. *Journal of Service Research*, *9*(2), 95–112.

OvalEdge. (2021). *Choosing the Technology Stack for a Data Lake*. OvalEdge. Retrieved from https://www.ovaledge.com/technology-stack-data-lake.

Raisinghani, J. (2019). *Data Lake vs Data Warehouse vs Data Mart*. The Holistics Blog. Retrieved from https://www.holistics.io/blog/data-lake-vs-data-warehouse-vs-data-mart/.

Verhoef, P. C., Hoekstra, J. C., & Van der Scheer, H. R. (2009). *Competing on analytics: Status quo van customer intelligence in Nederland*. Report of Customer Insights Center (RUGCIC-2009-02), University of Groningen.

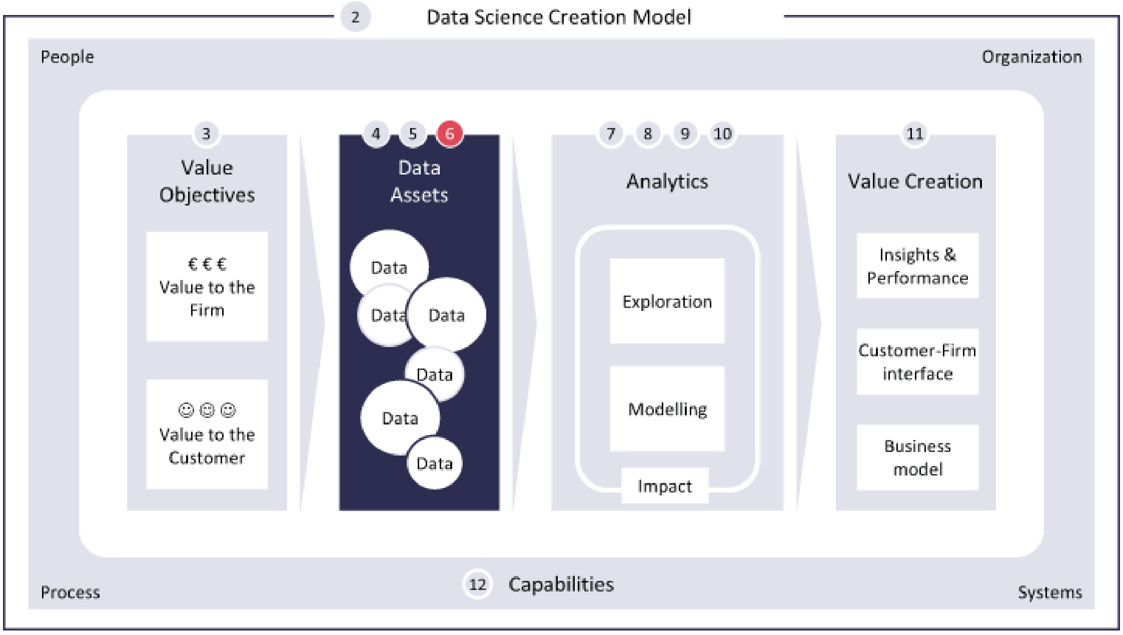
Verhoef, P. C., Spring, P. N., Hoekstra, J. C., & Leeflang, P. S. H. (2002). The commercial use of segmentation and predictive modelling techniques for database marketing in the Netherlands. *Decision Support System*, *34*(4), 471–481.

Zikmund, W. G., McLeod, Jr. R., & Gilbert, F. W. (2003). *Customer Relationship Management: Integrating Marketing Strategy and Information Technology*. New Jersey: Wiley & Sons.

CHAPTER 6

Customer privacy and data security

DOI: 10.4324/9781003011163-6



6.1 INTRODUCTION

So far in this book, we have mainly discussed the value opportunities of data science. Indeed, these value opportunities can be considerable. However, we are in an era when data firms are confronted with concerns about the storage and usage of data, specifically customer data. If customers have used digital and mobile devices, they will probably never be anonymous again. Their behavior is likely to be traceable online, but also offline. For example, if customers with a mobile device enter a store, retailers using WiFi-based tracking tools can follow them in the store and monitor how they shop. On a more global level, there are continuous debates about customers' data being analyzed by governments. This has become high profile news as a result of documents leaked by Edward Snowden, which revealed numerous global surveillance programs, many of them apparently run by the US National Security Agency (NSA) with the cooperation of telecommunication companies and European governments. This has raised a high level of concern globally about the information privacy of individual global citizens. Can firms like Google, Facebook, Microsoft, and Amazon be trusted regarding their privacy policies? And is the information contained in emails sent in Gmail, Microsoft Outlook, or Yahoo observable and available for analysis by governments?

Data science has thus created a stronger debate on privacy. The German Prime Minister Angela Merkel also considers privacy a big issue but believes that despite the rising privacy concern, we must be able to use big data to our advantage. Unfortunately, this is not so straightforward. For example, the Dutch Bank ING Retail announced a test that involved sharing payment data with other firms, intending to improve the value delivered to customers through providing attractive and relevant offers to ING customers. This resulted in strong reactions from customers, stakeholders, and politicians. Even the president of the Dutch National Bank actively communicated that he was not in favor of this big data initiative. ING had to retract the initiative and reduce their big data ambitions substantially. Another well-known case is that of Facebook when they worked together with Cambridge Analytica in the US elections of 2016, using data on around 87 million Facebook customers. When the use and analysis of Facebook data became widely known by the public it resulted in a big uproar, and Facebook lost $ 119 billion in market value (similar to the market value of McDonald's). Cambridge Analytica was terminated.

These two cases emphasize the challenge firms face with respect to privacy and big data, although the prevalence of privacy issues may vary between firms and sectors. It is, however, not only privacy that is an issue. Data security also deserves close attention, as data could become accessible to other parties with malicious or criminal motives through hacking of computer systems and unsecured data transport.

In this chapter, we will mainly focus on privacy. We will discuss what privacy entails and how customers consider privacy issues. We will also discuss the role of governments and, more specifically, legislation. In addition, we will elaborate on big data privacy policies. The last sections of this chapter consider data security.

6.2 WHY IS PRIVACY A BIG ISSUE?[1]

As discussed above, big data has stirred up the privacy discussion. Privacy has been an issue for decades, but big data development has put privacy back on the agenda of top management and governments. According to Jones (2003), big data are considered to be a threat to privacy for several reasons:

- Big data are permanently available
- Big data involve large volumes of data
- Big data on customers are collected invisibly
- There is no good assessment of the privacy sensitivity of data
- Due to the large volume of data, it is no longer accessible and comprehensible for customers
- Data from multiple sources are being merged
- Customers perceive a lack of control over big data collection.

If the privacy debate, particularly in Western societies, receives growing attention and if more restrictive policies are developed, the consequences for big data analytics and value creation are likely to be considerable. In 2012 The Boston Consulting Group[2] had already calculated that the so-called digital identity of customers had a value of approximately 330 billion Euros for firms (Boston Consulting Group 2012). This figure has only increased since then. However, these calculations clearly show that there is much at stake for both firms and customers. Firms can become more restricted in what they can do with big data, which may, for example, result in less efficient and effective advertising. Customers can be worse off, because they get less attractive personalized offers and personalized services may be reduced. Despite this, governments are rightfully worried about privacy issues and the use of data. There is a societal trade-off between data- and privacy requirements and the associated benefits from a lack of privacy.

6.3 WHAT IS PRIVACY?

Privacy is a concept that has a strong philosophical background. In essence, it implies the right to be left alone (Warren & Brandeis, 1890). This almost suggests that one should be able to live anonymously without being disturbed by anyone from any institution. This is of course rather unrealistic for the vast majority of consumers. Yet it clearly suggests that individuals should be able to make the trade-off between seclusion and interaction (Westin, 1967). This discussion is still rather philosophical. Bringing it into the context of data, Goodwin (1991) suggests that an individual should be conscious of which data and information are being shared and to what extent she or he controls this sharing. Two concepts are crucial here: consciousness and control.

A concept that is frequently discussed is “privacy concern.” A privacy concern usually focuses on six aspects of data:

1. Data collection: “Too much data and information is being collected”
2. Data usage: “Data is being used for other purposes than serving the consumer”

3. Data mistakes: “Mistakes in data can have negative consequences”
4. Data infringement: “Unauthorized access and usage of data”
5. Data control: “Insufficient control over own data”
6. Data consciousness: “Not sufficiently informed on data policies.”

Research on privacy has chiefly considered privacy concerns as the main topic of investigation and looked at drivers of privacy concerns and the consequences of privacy concerns. We will elaborate on this in the next section.

6.4 CUSTOMERS’ PRIVACY CONCERN

Customers may thus be concerned about the privacy consequences of big data. Concerns and fears are mainly related to the re-usage of data, reputation losses due to the use of data, and the wrong interpretation of data and information. Research also suggests that customers are unaware of usage and collection of information and they seem to be worried that they do not have sufficient knowledge on data usage (Hong & Thong, 2013). However, market research has also demonstrated that there are different segments of customers with different views on privacy concerns (Ackerman, Cranor, & Reagle, 1999; Fletcher, 2003):

- Fundamentalist: Fundamentally against the collection and usage of data by firms and governments
- Pragmatist: Willing to share information as long as benefits are received
- Unconcerned: Do not consider sharing information as a problem and see it as part of daily life.

Recently, Schumacher *et al*. (2020) executed a global segmentation study on willingness to share data. They also show similar segments. However, they show in addition two groups of countries. In many EU countries and Australia the privacy concerned segments are dominant (blue colored countries), while in countries in Asia and North and South America, the segment with a strong willingness to share data is dominant (red colored countries) (see Figure 6.1).

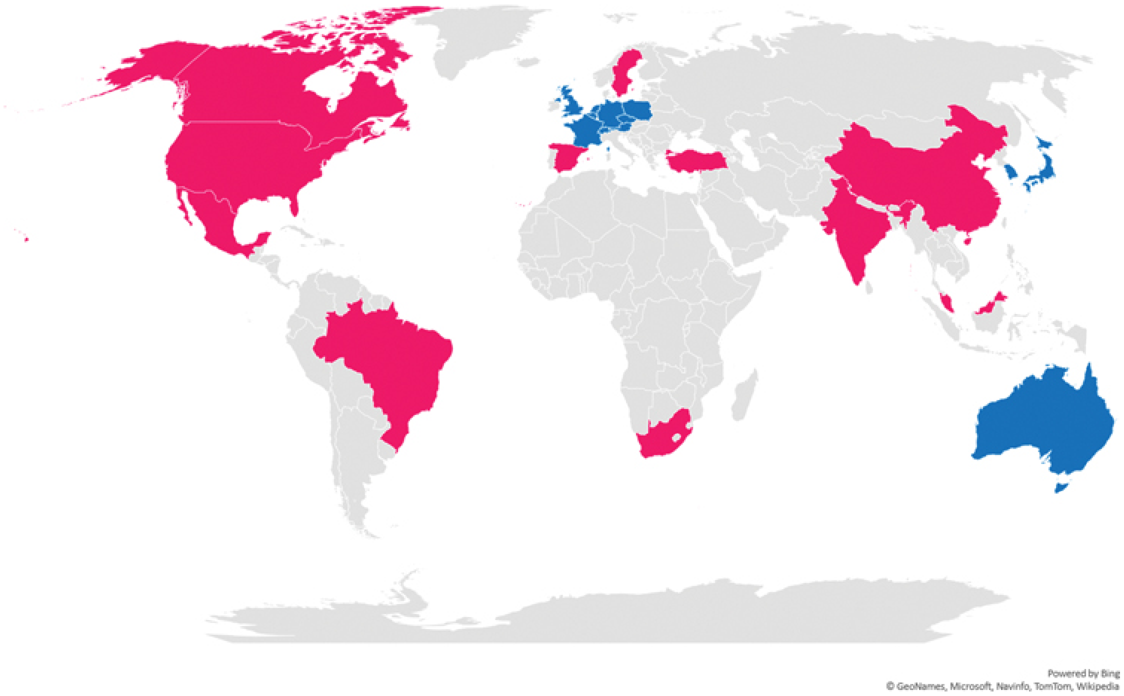


FIGURE 6.1 Global segmentation of willingness to share data

Source: Schumacher *et al*. (2020)

Substantial research has been devoted to specific drivers of privacy concerns. Older consumers are more concerned about their privacy, and females and consumers with low education levels are generally more apprehensive about it. Consumers who have experienced a privacy violation are also more concerned (Beke, Eggers, & Verhoef, 2018).

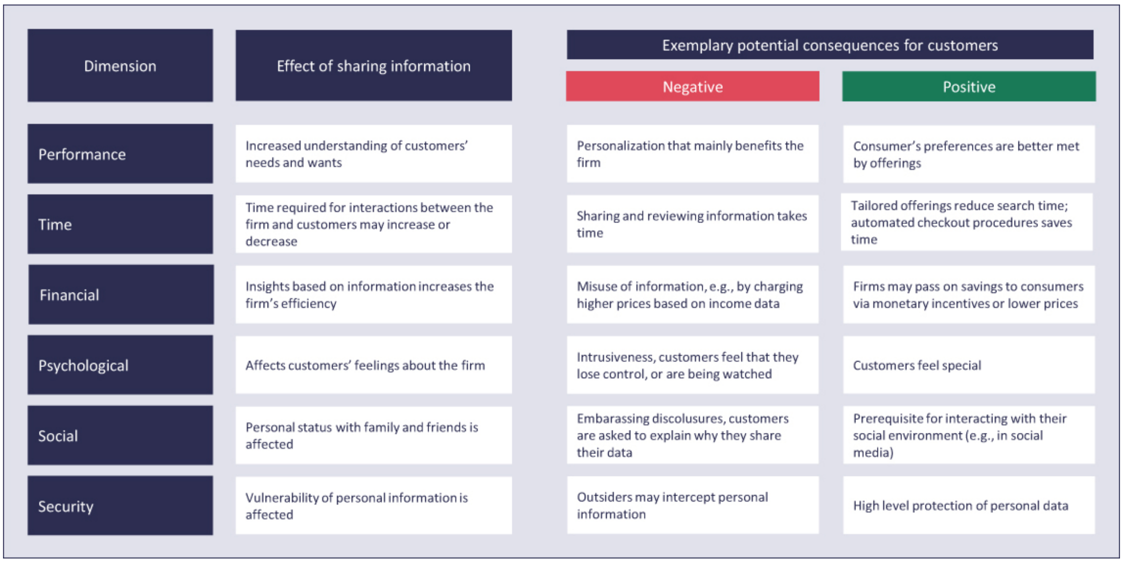
6.5 PRIVACY PARADOX AND PRIVACY CALCULUS

An important question is whether privacy concern also leads to less sharing of information by customers. The relationship between privacy concerns and actual data sharing behavior is not so obvious, and the correlation between privacy concern and behavior is low. Only when customers have very strong concerns about their privacy do they change their behavior (Van Doorn, Verhoef, & Bijmolt, 2007). It can be argued that customers do not behave very consistently. Customers seem to be worried about their privacy but billions of customers around the globe constantly share very personal information on Facebook and leave digital traces behind online. The strong disconnect between privacy concern and actual behavior is referred to as the "privacy paradox."

A concept that is important in this respect is the privacy calculus. Consumers look beyond negative outcomes (concerns), and also take positive outcomes of the collection, storage, and use of personal information into account. The privacy calculus suggests that consumers determine for themselves whether they regard the consequences of the collection, storage, and use of personal information to be beneficial (providing benefits) or detrimental (incurring costs or risks) in a specific situation (Beke, Eggers, & Verhoef, 2018). The benefits and costs to be considered when consumers use a privacy calculus are (Beke *et al*., 2021):

– Performance
– Time
– Financial
– Psychological
– Social
– Security.

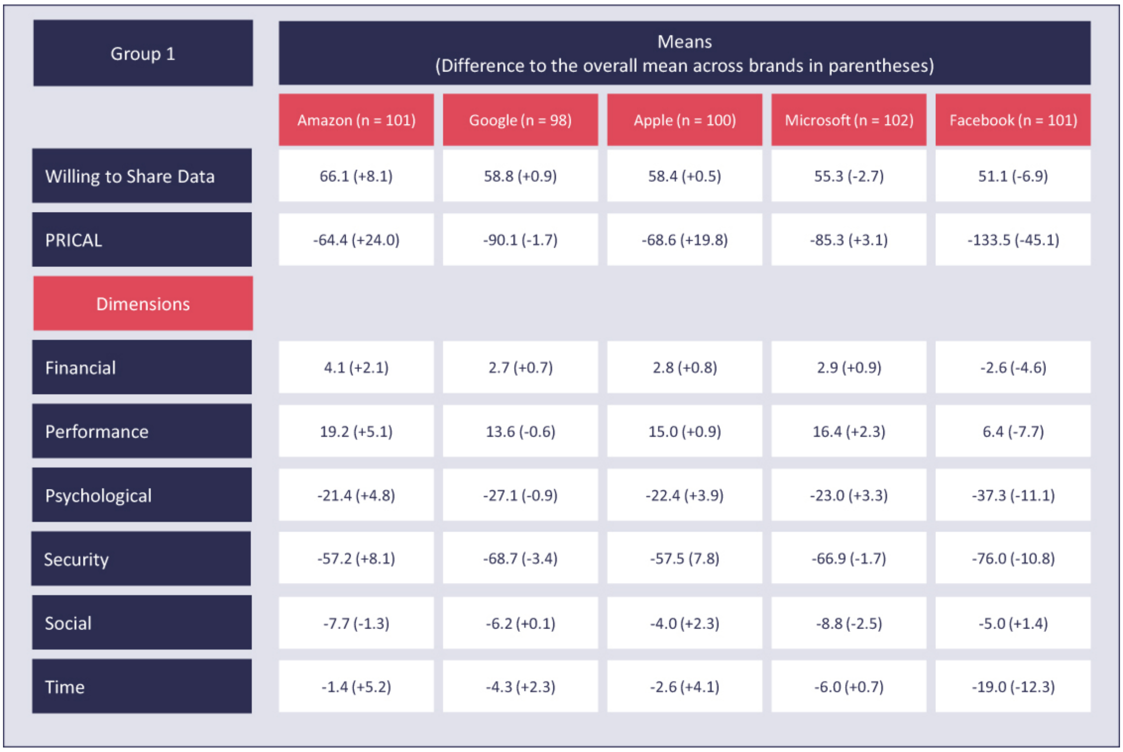
These dimensions are described in Figure 6.2, in which the effect of data sharing is shown as well as the benefits and costs for consumers. The dimensions have been measured using the PRICAL index for major tech firms, such as Google and Amazon (Beke *et al*., 2021). In this index, the value of the dimension and the valence is measured. There are clear differences in how these firms score on the total privacy calculus and the underlying dimensions. Facebook scores relatively low, while Amazon scores high (see Figure 6.3).



Dimension	Effect of sharing information	Exemplary potential consequences for customers	
		Negative	Positive
Performance	Increased understanding of customers' needs and wants	Personalization that mainly benefits the firm	Consumer's preferences are better met by offerings
Time	Time required for interactions between the firm and customers may increase or decrease	Sharing and reviewing information takes time	Tailored offerings reduce search time; automated checkout procedures saves time
Financial	Insights based on information increases the firm's efficiency	Misuse of information, e.g., by charging higher prices based on income data	Firms may pass on savings to consumers via monetary incentives or lower prices
Psychological	Affects customers' feelings about the firm	Intrusiveness, customers feel that they lose control, or are being watched	Customers feel special
Social	Personal status with family and friends is affected	Embarassing discolusures, customers are asked to explain why they share their data	Prerequisite for interacting with their social environment (e.g., in social media)
Security	Vulnerability of personal information is affected	Outsiders may intercept personal information	High level protection of personal data

FIGURE 6.2 Dimensions of privacy calculus

Source: Beke *et al*. (2021)



Group 1	Means (Difference to the overall mean across brands in parentheses)				
	Amazon (n = 101)	Google (n = 98)	Apple (n = 100)	Microsoft (n = 102)	Facebook (n = 101)
Willing to Share Data	66.1 (+8.1)	58.8 (+0.9)	58.4 (+0.5)	55.3 (-2.7)	51.1 (-6.9)
PRICAL	-64.4 (+24.0)	-90.1 (-1.7)	-68.6 (+19.8)	-85.3 (+3.1)	-133.5 (-45.1)
Dimensions					
Financial	4.1 (+2.1)	2.7 (+0.7)	2.8 (+0.8)	2.9 (+0.9)	-2.6 (-4.6)
Performance	19.2 (+5.1)	13.6 (-0.6)	15.0 (+0.9)	16.4 (+2.3)	6.4 (-7.7)
Psychological	-21.4 (+4.8)	-27.1 (-0.9)	-22.4 (+3.9)	-23.0 (+3.3)	-37.3 (-11.1)
Security	-57.2 (+8.1)	-68.7 (-3.4)	-57.5 (7.8)	-66.9 (-1.7)	-76.0 (-10.8)
Social	-7.7 (-1.3)	-6.2 (+0.1)	-4.0 (+2.3)	-8.8 (-2.5)	-5.0 (+1.4)
Time	-1.4 (+5.2)	-4.3 (+2.3)	-2.6 (+4.1)	-6.0 (+0.7)	-19.0 (-12.3)

FIGURE 6.3 Score on privacy calculus measure PRICAL and underlying dimensions for major tech firms

Source: Beke *et al*. (2021)

6.6 GOVERNMENTS AND PRIVACY LEGISLATION

Governments across the globe develop their own legislation. There is a great deal of fragmentation on privacy regulation. The strictest laws are found in the EU and Canada. The USA and Australia have less stringent privacy legislation. In emerging countries, i.e., Brazil, Russia, India, and China (known as BRIC), legislation is relatively limited (see Figure 6.4).

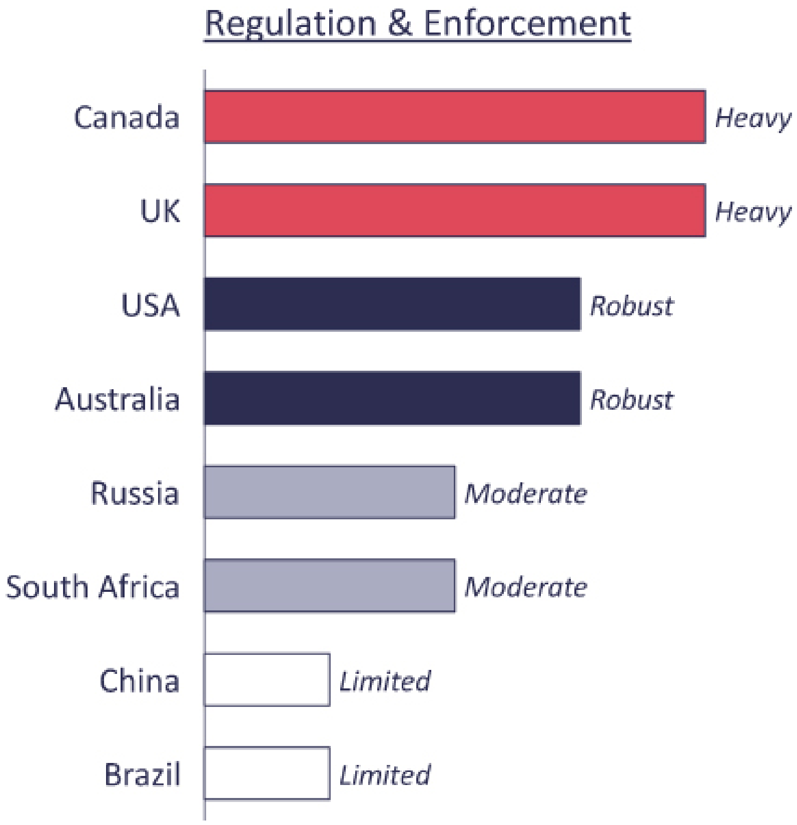


FIGURE 6.4 Data protection laws around the globe

Source: Adapted from D. L. A. Piper, Data protection laws of the world[3]

Within the European Union, the member states developed Data Protection Directives, which led to EU General Data Protection Regulation (GDPR). The EU has two main objectives with GDPR:

1. Protect the data and strengthen the privacy rights of EU citizens
2. Give EU citizens control of their data.

GDPR affects all firms and institutions that hold data on EU citizens. This means that, for example, US firms like Amazon have to take GDPR legislation into account when doing business with EU customers. There is an official EU institution that oversees the implementation of this legislation by firms. The institution has the authority to administer official warnings, audits, and ultimately fines with a maximum of 100 million Euros or 5% of the company's worldwide turnover. Indeed, the EU has started up privacy trajectories with important players, such as Google. A few principles in the EU legislation should be emphasized:

1. Right to access: How is personal data processed (i.e., purpose, type of data, storage period);

2. Right to rectification: Correction of inaccurate personal data, without any delay;
3. Right to erasure: Erase all personal data if not required any more, or if user withdraws consent;
4. Right to restriction of processing: In case the data accuracy is contested, unlawful or not required anymore;
5. Right to data portability: Customers have the right to view their collected personal information in a structured format; and
6. Right to object: Stop processing individual data on request, unless the controller demonstrates compelling reasons overriding the individual's interests and rights.

The legislation in the US is not as strong as it is in the EU. One of the largest differences lies in the approach. The EU is rather proactive in its legislation, whereas the US is rather passive. Beyond that, the legislation in the US differs between states. In Massachusetts and California, for example, privacy laws are stricter than in Georgia and Florida.[4] In the US the Federal Trade Committee (FTC) is responsible for general privacy laws. The FTC wrote a large privacy report in 2012 (FTC 2012). They concluded that industry efforts to address privacy through self-regulation "have been too slow, and up to now have failed to provide adequate and meaningful protection." The report recommended numerous actions. Some of them would bring US regulation more in line with that of the EU. For example, they recommend a privacy-by-design approach. Furthermore, they emphasized that customers should have a choice. Customers should be presented with choices about the collection and sharing of their data at the time and in the context in which they are making decisions—not after having to read long, complicated disclosures that are often difficult to find. FTC also recommends a "do not track" mechanism, which governs the collection of information about consumers' Internet activity to deliver targeted advertisements and for other purposes (FTC 2012).[5] So far, the report has only resulted in recommendations for firms, and these have still not been implemented in US legislation. In 2020, however, California, one of the largest states in the US where many tech firms reside, implemented the California Consumer Privacy Act.

The Act intends to provide California residents with the right to:

1. Know what personal data is being collected about them.
2. Know whether their personal data is sold or disclosed and to whom.
3. Say no to the sale of personal data.
4. Access their personal data.
5. Request a business to delete any personal information about a consumer.
6. Not be discriminated against for exercising their privacy rights.

6.6.1 Steps to comply with GDPR

Before considering the steps required to comply with the GDPR, it is important to understand the roles of different stakeholders within GDPR. GDPR distinguishes different roles:

- Data Subject: an individual, a resident of the EU, whose personal data are to be protected.
- Data Controller: an institution, business, or a person processing the personal data (e.g. e-commerce website).
- Data Processor: a subject (company, institution) processing data on behalf of the controller (e.g., Google, Facebook).
- Data Authority: a public institution monitoring implementation of the regulations in the specific EU member country.

Firms typically take the role of the data controller. Hence, they have a central role in the system (see Figure 6.5). In order to comply with the GDPR data controllers need to take several steps internally. Typical steps towards GDPR compliance are:

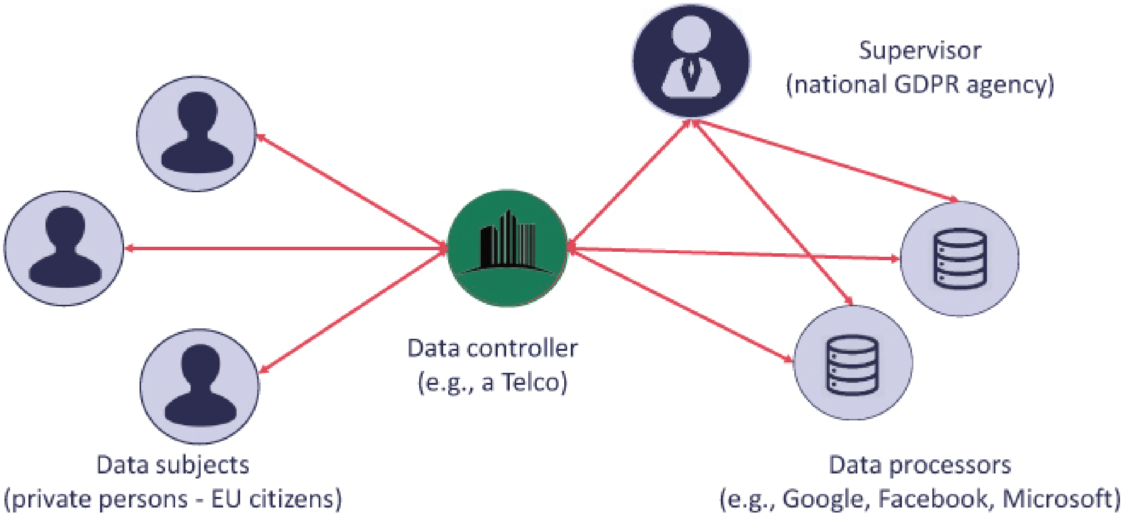


FIGURE 6.5 Different roles of stakeholders in GDPR

1. Analyze what information is collected and where data is stored (e.g., cookies, tracking pixels, emails, names, addresses …);
2. Check storage time (e.g., data still relevant after some time? If not, remove data)
3. Inform customers (e.g., how can they request, modify or delete their data?)
4. Monitor the access to personal clients' data.

Within GDPR the data controller has several responsibilities and must be aware of these responsibilities. A data controller should:

1. Audit data usage (what is collected, where is it stored):
2. Appoint a Data Protection Officer (DPO);
3. Check data processors; and
4. Monitor data breaches.

The DPO becomes an important role within firms under GDPR. A DPO is a person appointed by the data controller responsible for overseeing data protection practices.

The GDPR also clearly states responsibilities in terms of what firms should do in their communication to users. Firms should:

1. Communicate terms in plain language on:

   what data are collected; and the purpose of collecting data (e.g., billing, solve disputes, etc.)
2. Communicate the privacy policy
3. Communicate the cookie files policy
4. Ask for the consent of customers to use their data:

   for marketing purposes of company / trusted partners

   for remarketing by partners.

6.6.2 Going beyond legislation

Beyond data legislation, firms should be aware that the resulting actions derived from their data usage are also affected by laws. For example, if firms target personal characteristics, such as religion, gender, and race, this could result in discrimination. Although the firm may act in accordance with specific privacy legislation, its resulting policies could conflict with regulations on discrimination. Hence, firms should look beyond privacy legislation and also consider regulations related to specific marketing actions.

In this regard, firms and sector organizations can also self-impose measures going beyond government regulations. Examples are WebTrust and TRUSTe. WebTrust is developed by the American Institute of Certified Public Accountants and the Canadian Institute of Chartered Accountants. TRUSTe is founded by Electronic Frontier Foundation and CommerceNet Consortium, Inc. Firms signing up for TRUSTe adhere to TRUSTe’s privacy policies on disclosure, choice, access, and security. There is ongoing oversight and alternative dispute resolution processes for firms that sign up for TRUSTe.

6.7 PRIVACY AND ETHICS

On a higher level, one could debate whether firms should mainly focus on legislation and take that as the rule on how far they will go with data collection or whether they should take a broader perspective. The latter leads to a discussion on how firms make “moral” decisions. One could advocate an approach in which firms are allowed to collect data as long as the law allows them to do so. However, one could also adopt a more ethical perspective that goes beyond laws, where firms consider that they have an ethical responsibility. Tsalikis and Fritzsche (2013) provide an interesting overview of the business ethics literature, and review frameworks on how to implement morality in business decision making. De Bruin (2015) distinguishes two important dimensions in this respect: ethical decision making and moral intensity. Ethical decision making applied to privacy involves four stages:

1. Firms recognize that privacy decisions have a moral dimension
2. Firms form an ethical judgment concerning what ought to be done with regard to privacy
3. Firms establish the moral intention to act in conformity with what they have judged to be the right type of behavior with regard to privacy issues
4. Firms engage in that behavior.

The “moral intensity” of an ethical issue, such as privacy, refers to the magnitude of the consequences of the actions and the probability of it arising, as well as whether the consequences are concentrated on a group of people or dispersed among them. Importantly, it also depends on whether there is any social consensus about the fact that particular actions are good or evil (De Bruin, 2015). De Bruin (2015) explicitly states that, roughly speaking, when harmful consequences are likely to affect people in close proximity or a large number of people, and when the firm perceives this to be the case, the issue’s moral intensity is high. The moral intensity of an issue determines how firms proceed through each of the four ethical decision-making stages. If the moral intensity is high, then an issue is considered as a moral issue and elaborate ethical decision making is required.

The above discussion is rather theoretical. Taking a privacy perspective, in practice firms should put consideration of the moral intensity of privacy issues high on their agenda. We contend that the big data development has created a wider acknowledgement of privacy issues in society. It is probably too much to say that big data may result in wrong or harmful consequences. However, when data are accessible to criminals, “evil” things may happen to customers. Overall, we believe that privacy is an ethical issue that requires more attention than merely taking the law into account. This would also be in line with a more customer-centric approach, as firms caring for customers and the interests of customers should strongly consider their views on privacy and how they should deal with data. Moreover, recent discussions have shown that big data initiatives can potentially harm a firm's reputation. In sum, privacy issues surrounding big data should be a very important discussion topic within firms, and they should go through more intensive decision making than is required simply to follow the available legislation.

So, in general we recommend that firms should strongly consider the ethical and reputational consequences of their big data and privacy policies. Specifically, they should adopt stakeholder management and consider reactions from customers, the government (including politicians), and the media. This fits with the recently introduced notion of corporate digital responsibility (CDR). CDR is defined as a set of shared values and norms guiding an organization's operations with respect to four main processes related to digital technology and data. These processes are the creation of technology and data capture, operation and decision making, inspection and impact assessment, and refinement of technology and data (Lobschat *et al*., 2021). As with the idea of Corporate Social Responsibility (see Chapter 3), firms adopting CDR should make long-term financial gains despite the fact that initial investments are required to achieve a high level of CDR.

6.8 PRIVACY POLICIES

There has been extensive research on privacy and, more specifically, privacy policies. Based on this research we have multiple recommendations on how to deal with specific data and privacy issues:

- Only collect data that are relevant and congruent. Relevant data means data that are considered useful in servicing the customer. Congruent data is data that is related to the product or service provided. For example, for financial services, data on ownership of insurance products or financial transactions are congruent with the service. However, data on medical issues would not be considered congruent.
- From a privacy perspective the rule "more is less" holds. The more information is asked for, the less customers provide! Thus, firms should limit what they ask customers to provide.
- Give something back to customers. Data has value for firms as well as for customers. Rewards (monetary and non-monetary) can increase the customer's willingness to share information. However, this does not work for irrelevant information. Moreover, there are differences between customer segments.
- Be transparent on data usage. Providing a clear privacy statement positively influences data sharing. It may reduce customers' lack of awareness of data usage and reduce privacy concerns.

- Communicate the specific benefits of sharing data. If customers perceive the benefits of sharing data, they are more likely to share. Privacy mainly becomes a problem when the advantages of sharing data are not clear.
- Invest in brand trust. Customers are more likely to share information with firms they trust, as they believe the risks involved are lower. Trust particularly becomes an issue when customers are asked to share personal and sensitive information. Even trusted firms that incorrectly use the data provided may find that their use of data backfires on them.
- One final recommendation is that firms should give customers control! Customers who feel in control have more positive views on sharing information and share more information as well. This can strongly increase the effectiveness of marketing. Giving control to customers implies the use of opt-in and opt-out options and the use of permission-based marketing. Wieringa *et al*. (2021) discuss some initiatives that go even further where customers collect, store and control their own data, and only provide them to firms after they assess the benefits of doing so.

The power of giving control to customers is excellently demonstrated in a study by Catherine Tucker (2014) of MIT. She reports the results of a study among Facebook users, where a simple control button on privacy was added. Privacy became standard and friends were no longer visible to everyone. Moreover, opting-out for the firms' use of data was made more convenient. Tucker (2014) reports that the effectiveness of targeted and personalized advertising in particular, more than doubled (see Figure 6.6).

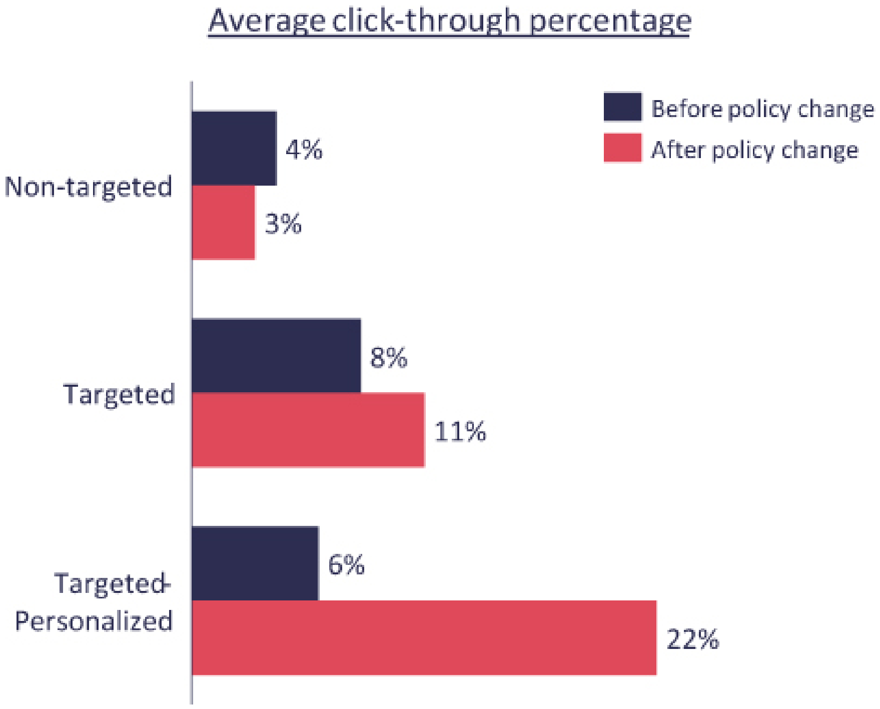


FIGURE 6.6 Effectiveness increase of Facebook advertising campaigns after addition of privacy button

Source: Adapted from Tucker (2014)

6.9 PRIVACY AND INTERNAL DATA ANALYTICS

Privacy also has consequences for how data is analyzed, as the analysis of data on an individual customer level might not be allowed due to legislation. On a general level firms can strive for data minimization and only store data that is necessary for gaining sufficient insights and predictions. However, firms may still need data but are probably not always allowed to use it. We consider the following specific solutions: (see Figure 6.7)

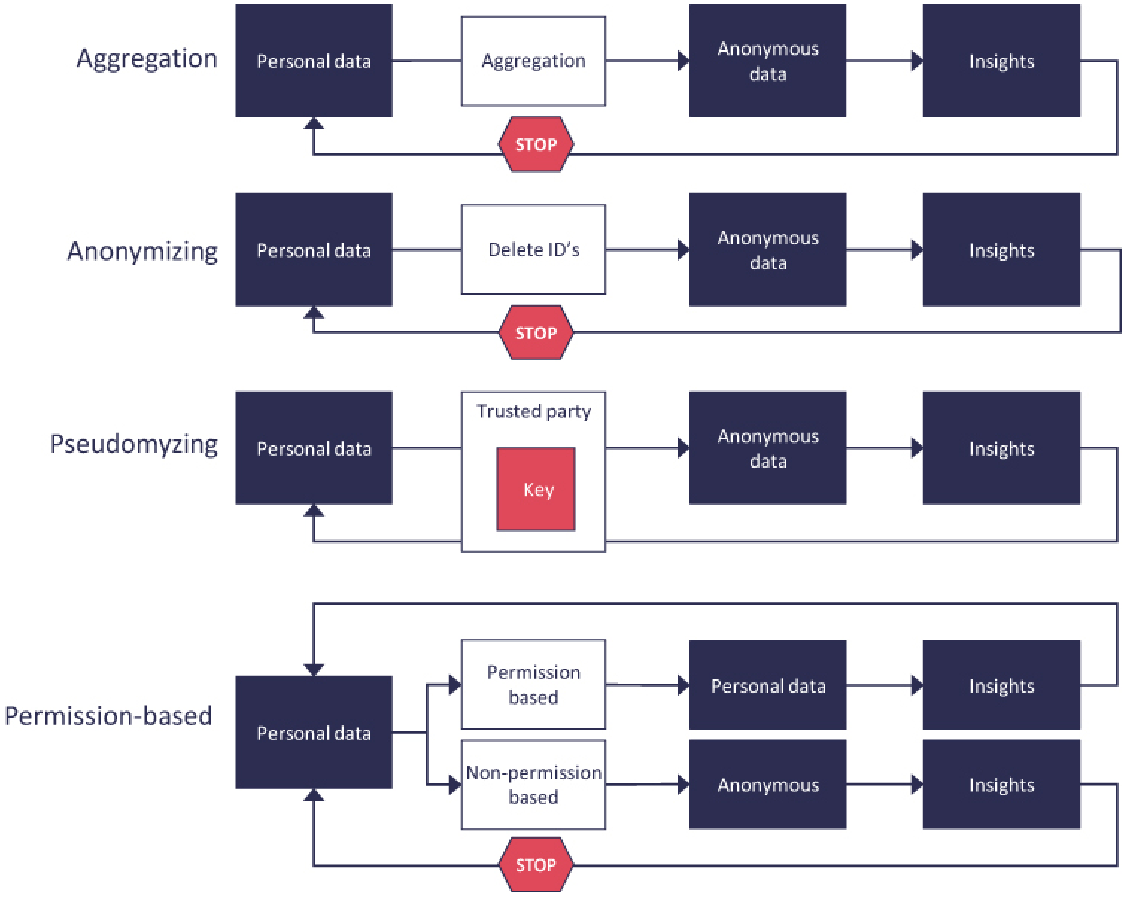


FIGURE 6.7 Different ways of handling privacy sensitive data

- If individual data are problematic to analyze from a privacy perspective, one could aim to analyze data on higher aggregation levels. For example, specific market segments could be studied, or aggregated analyses could be carried out. All these analyses can provide customer insights.
- When analyzing data, individual data can be anonymized. This is typically done when one only requires customer insights. Anonymizing

data is usually the standard in traditional market research. If one aims to include the results of the analysis in the database, the model results (e.g., churn model) can be used to create the model outcome in the data (e.g., churn probability).

- A specific form of anonymization is what we call pseudonymizing the data. With a specific key, data are anonymized. This anonymization is done by a trusted party (e.g., external IT firm, external law-firm). Only this trusted party knows the key. The analysts execute the analyses, and the results can be input into the normal customer database. Again, the external trusted party takes care of de-anonymizing the data.
- A final technique is linked to the already discussed permission-based marketing approach (Godin, 1999). A subsample of not anonymous data on customers who have given permission to analyze and use it is analyzed. For the customers granting permission, the results of the analyses can be included in the database. For the remaining part of the data, the outcomes of the analysis can be included by using model estimates to predict specific values. One concern here is that the permission-based sample will typically not be random and hence the analytical results could be biased due to some self-selection issues. Krafft, Arden and Verhoef (2017) show that expected personal relevance, provided entertainment, and consumer information control directly positively affect the probability of consumers granting permission, while pronounced registration cost, privacy concerns, and anticipated intrusiveness are shown to have negative effects on the likelihood of granting permission.

6.9.1 Model based solutions for privacy

Model-based solutions are currently being developed to tackle privacy issues in analytics. One such example is sophisticated model-based approaches for data protection, typically aiming to generate customer-level, “synthetic” data by mimicking an underlying data-generating process. The synthetic data-generating “engines” perform multiple imputation and bootstrap procedures to address gaps in data (Rubin, 1993), based on either a statistical (e.g., Bayesian) model that generates a posterior predictive distribution according to some protected, underlying probability distribution of the original data, or else some advanced machine or a deep learning approach (see also Chapter 9). Ponte and Wieringa (2021) employ Generative Adversarial Networks (GANs) in which two neural networks compete with each other to generate the data of non-existent customers. They show that analyzing these artificial data generates customer insights that are very close to the insights obtained when analyzing “real” data. Their procedure contributes to ensuring anonymity of customers.

Some new methods are also being developed that aim to benefit from historical data, using earlier analytical results. Holtrop *et al.* (2017) have developed the so-called GMOK (Generalized Mixture of Kalman Filters) model. In this model, they use state-space modeling techniques to predict customer churn. The general idea is that input from earlier churn models is used as input in the next model estimation. Firms then only have to store model output and do not have to store long histories of data to make better predictions. The results of the GMOK model are rather good, predicting 4.5 times more accurately than a random prediction, while the model also outperforms alternative models.

6.10 DATA SECURITY

A subject closely related to privacy is data security. Data infringement (the unauthorized access and usage of data) discussed in the previous section can seriously impact an individual's financial and/or personal safety and can also compromise the continuance of a firm with a security breach. Remember, for example, the hacking of LinkedIn passwords several years ago, or the Apple iCloud hack where nude pictures of celebrities were spread all over the Internet. Data breaches have been shown to negatively impact customer spending (Janakiraman, Lim, & Rishika, 2018). Moreover, data breaches in an industry not only affect the firm having the data breach but could also backfire on other firms in the same industry.[6] These examples illustrate that even renowned major companies can have data security problems and highlight the need for every firm working with privacy or company sensitive information to take the right measures to secure their data. A strong privacy and data security policy can also help firms to defend themselves against data breaches at other competing firms and even to benefit from them.

We distinguish between three elements of data security. The first is the people element: people that use or have access to the data. The second is the system element; dealing with the physical data storage and the environment where the data are stored. The third is the processes; the procedures and policies (including penalties) as defined by the organization, that define the rules for access, continuity, steering, and monitoring of security performance.

6.10.1 People

Unsurprisingly, people are often the weakest link in data security. Very often employees of the firm are (at least partly) responsible for security issues. Everybody knows examples of sensitive data on a USB stick being left in a public place or on a non-company computer, or email attachments being sent outside the organization to the wrong person. These are only examples of data security failures that are not intentional. Even more significant are the security threats from people both inside and outside the firm who have malicious or criminal intentions. This means that firms should protect themselves from human failure or misbehavior. They should screen their employees' credentials and make employees aware of the compliance policies in place—even let them (some of them at least) take an exam on the compliance and security rules of the organization. Another way of dealing with the risk is to grant access to only those data sources that someone needs to do his or her job.

6.10.2 Systems

By systems we mean the physical and technical environment where data are stored. Currently, the "cloud" is booming, and more and more firms are putting crucial data sources in the cloud—which means that it is sometimes not clear where the data are physically located. Especially due to legal implications (i.e., under which law the data should be treated) firms should be aware of where their data are at the geographical level. In many firms there are several systems for which data are stored, sometimes even redundant systems, meaning that all these systems should be in mind when considering the necessary measures for data security. Every system will have its own criteria on how critical the data in that system are for the daily operations of the firm, and what that implies for data security. Typical measures regarding systems include physical access to systems and computers (e.g., protocols for entering the rooms where the systems are located) or how often and when backups are made, and for how long it is acceptable that a system is "down."

6.10.3 Processes

We define three types of processes: (1) processes for access; (2) processes for continuity; and (3) processes for steering and monitoring. Processes for access define when and who has access to what data and for which purposes. This means defining, for each user, the different rights to work with or use data, and ensuring that every user has a strong username/password combination that will be updated at prescribed times. Processes for continuity define the firm's policies on how to deal with calamities, how and where to make a back-up, what fall-back options are available, and having a disaster recovery plan in place. Steering and monitoring processes make sure that a data security baseline is in place within the organization, defining the critical performance indicators, and including the standards for these indicators. Reporting on these indicators makes sure that the security measures are monitored and that the firm is aware of possible incidents, including the actions taken.

6.11 CONCLUSIONS

Privacy and security have become important issues for firms and in data science. Privacy, in a sense, directly links analytics to the customer. Hence it is very important when applying data science to take privacy issues into account. In this chapter, we have tried to give a comprehensive overview of what privacy entails and why it is important. We discussed some important privacy policies and regulations, such as GDPR. We also investigated how privacy impacts analytics and mentioned specific solutions to manage this. Finally, we discussed some important issues surrounding data security. In sum, privacy and security should be primary issues in data science. It is no longer only the firm's law department that should worry about this. Privacy and security impact marketing, marketing analytics, and even the board.

ASSIGNMENT 6.1: CURANI PET CARE

Veterinary care is becoming a big business. The total global market for veterinary clinics was around $67.3 billion in 2018. Within this market not only small independent veterinary clinics are active, but also major corporations, such as the CVS group and Mars. They have established large chains and operate globally. For these firms investing in customer data and building up data science capabilities can be promising; for example, a dog owner spends on average between $700 and $2,000 per year on routine veterinary care. The Veterinary Chain CurAni is now starting to collect more customer data on their customers and pets. In their billing database they already had data on the name of the customer, address, the type of pet(s) and name(s) of pet(s). They are now planning to collect more pet data, and also data on the customer. The pet data they are considering is the birthday of the pet and any health problems. From the customer they want to know more on activities with the pet (i.e. frequency of walks) and lifestyle data, so that they can personalize offers online and offline.

1. Would data on the pet be considered personal customer data?
2. CurAni is planning to ask for data from their customer on their pet(s): (1) birthday of the pet and (2) health problems of the pet.
    a. How would customers look at the privacy of these two data requests. You should use the PRICAL framework.
    b. To what extent would customers be willing to share these two sources of data?
3. In a next step they want to ask customers to share data on (1) activities with the pet, and (2) lifestyle data (i.e. hobbies)
    a. How would customers look at the privacy of these two data requests. You should use the PRICAL framework.

    b. To what extent would customers be willing to share these two sources of data?

4. CurAni is planning to send a birthday card for the pet using the pet data. Would customers of CurAni consider this intrusive?
5. What type of personalization actions would customers of CurAni potentially consider intrusive? Provide two examples and discuss why they are intrusive.

NOTES

1. This chapter is based on the report of the Customer Insights Center by Beke and Verhoef (2015).
2. See www.bcgperspectives.com/content/articles/digital_economy_consumer_insight_value_of_our_digital_identity/ (accessed September 18, 2015).
3. See www.dlapiperdataprotection.com (accessed September 17, 2015).
4. See: https://iapp.org/resources/article/state-comparison-table/
5. See www.ftc.gov/news-events/press-releases/2010/12/ftc-staff-issues-privacy-report-offers-framework-consumers (accessed September 17, 2015). The report can be downloaded from this website.
6. Weblink: https://hbr.org/2018/02/research-a-strong-privacy-policy-can-save-your-company-millions

REFERENCES

Ackerman, M. S., Cranor, L. F., & Reagle, J. (1999). *Privacy in e-commerce: Examining user scenarios and privacy preferences*. ACM Conference on Electronic Commerce, 1–8.

Beke, F. T., Eggers, F., & Verhoef, P. C. (2018). Consumer informational privacy: Current knowledge and research directions. *Foundations and Trends in Marketing*, *11*(1), 1–71.

Beke, F.T., Eggers, F., Verhoef, P.C., & Wieringa J.E. (2021). Consumers' privacy calculus: The PRICAL index development and validation.*International Journal of Research in Marketing*, forthcoming.

Beke, F. T. & Verhoef, P. C. (2015). *Privacy: Bedreigingen en kansen voor bedrijven en consumenten*. Report of Customer Insights Center (RUGCIC-2015-01), University of Groningen.

De Bruin, B. (2015). *Ethics and the Global Financial Crisis: Why Incompetence Is Worse Than Greed*. Cambridge: Cambridge University Press.

Fletcher, K. (2003). Consumer power and privacy: The changing nature of CRM. *International Journal of Advertising*, *22*(2), 249–272.

FTC (2012). *Protecting consumer privacy in an era of rapid change: Recommendations for businesses and policymakers*. Retrieved from FTC.gov. Retrieved September 18, 2015 from www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf

Godin, S. (1999). *Permission Marketing: Turning Strangers into Friends and Friends into Customers*. New York: Simon and Schuster.

Goodwin, C. (1991). Privacy: Recognition of a consumer right. *Journal of Public Policy & Marketing*, *10*(1), 149–166.

Holtrop, N., Wieringa, J. E., Gijsenberg, M. J., & Verhoef, P. C. (2017). No future without the past? Predicting churn in the face of customer privacy. *International Journal of Research in Marketing*, *34*(1), 154–172.

Hong, W., & Thong, J. Y. L. (2013). Internet privacy concerns: An integrated conceptualization and four empirical studies. *MIS Quarterly*, *37*(1), 275–298.

Janakiraman, R, Lim, J. H., & Rishika, R. (2018). The effect of a data breach announcement on customer behavior: Evidence from a multichannel retailer. *Journal of Marketing*, *82*(2), 85–105.

Jones, K. (2003). Privacy: What's different now? *Interdisciplinary Science Reviews*, *28*(4), 287–292.

Krafft, M., Arden, C. M., & Verhoef, P. C. (2017). Permission marketing and privacy concerns – Why do customers (not) grant permissions? *Journal of Interactive Marketing*, *39*, 39–54.

Lobschat, L., Müller B., Eggers, F., Brandimarte, L., Diefenbach, S., Kroschke, M., & Wirtz, J. (2021). Corporate digital responsibility. *Journal of Business Research*, *122*, 875–888.

Ponte, G., & Wieringa, J. E. (2021). Privacy-preserving generative adversarial networks to share data and derive marketing insights. *Working paper*, University of Groningen.

Rubin, D.B. (1993), Statistical disclosure limitation, *Journal of Official Statistics*, *9*(2), 461–468.

Schumacher, C., Eggers, F., Verhoef, P. C., & Maas, P. (2020). Understanding consumers' willingness to share personal information: A multinational segmentation analysis. *Working Paper*, University of St. Gallen.

The Boston Consulting Group. (2012). The value of our digital identity. Retrieved September 11, 2015 from www.slideshare.net/fred.zimny/boston-consulting-group-the-valueofourdigitalidentity

Tsalikis, J. & Fritzsche, D. J. (2013). Business ethics: A literature review with a focus on marketing ethics. *Citation Classics from the Journal of Business Ethics*, *8*(9), 337–404.

Tucker, C. (2014). Social networks, personalized advertising, and privacy controls. *Journal of Marketing Research*, *51*(5), 546–562.

Van Doorn, J., Verhoef, P. C., & Bijmolt, T. H. A. (2007). The importance of non-linear relationships between attitude and behaviour in policy research. *Journal of Consumer Policy*, *30*(2), 75–90.

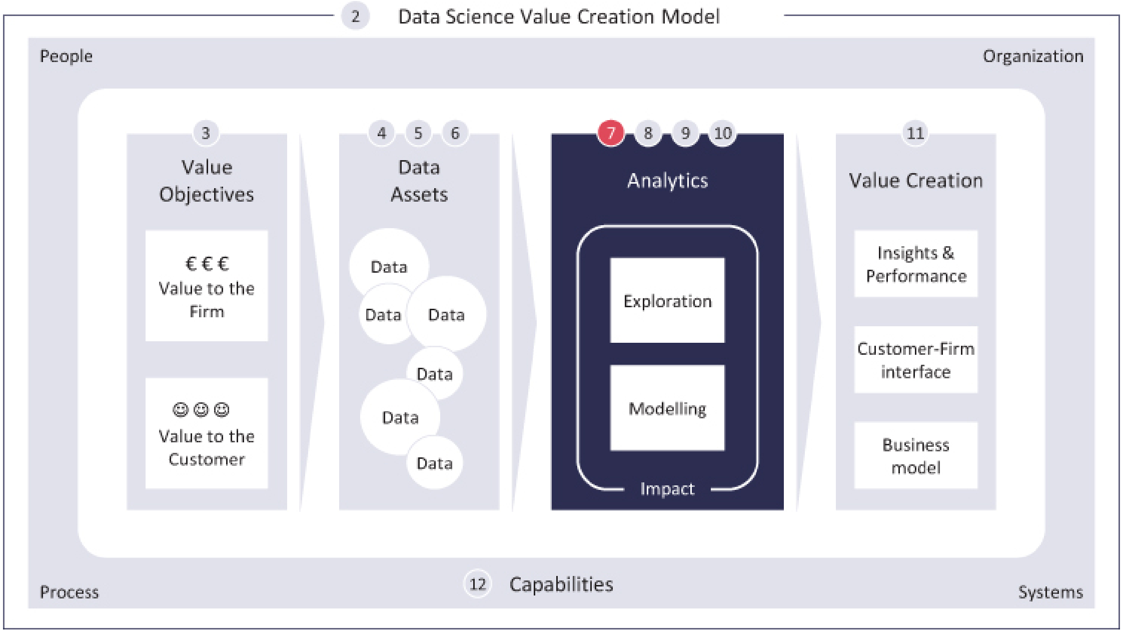
Warren, S. D., & Brandeis, L. D. (1890). The right to privacy. *Harvard Law Review*, *4*(5), 193–220.

Westin, A. F. (1967). *Privacy and Freedom*. New York: Atheneum.

Wieringa, J., Kannan, P. K., Ma, X., Reutterer, T., Risselada, H., & Skiera, B. (2021). Data analytics in a privacy-concerned world. *Journal of Business Research*, *122*, 915–925. Retrieved from https://doi.org/10.1016/j.jbusres.2019.05.005

CHAPTER 7
Data analytics

DOI: 10.4324/9781003011163-7



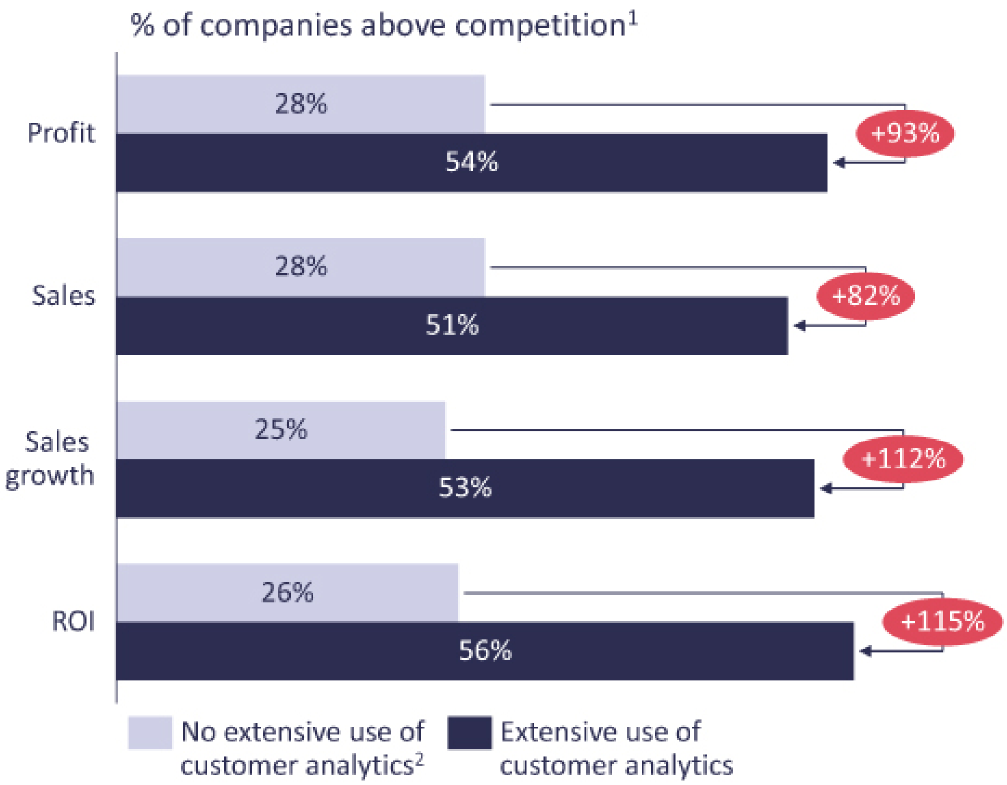
7.1 INTRODUCTION

Analytics is a major element of creating value from data. Statistical analytics of marketing data have been around for decades. The revolutions in scanner data and customer relationship management (CRM) have considerably increased the importance of analytics in marketing: it creates a strong market and customer insights and models that can be used for decision support, campaigns, and data-driven solutions. However, the emerging presence of big data and AI is changing analytics. Taking a more historical lens, we can observe certain developments in analytics. First, we will discuss the different strategies for analyzing data. Subsequently, we describe the role of analytics and general types of marketing analysis. We end with an understanding of the meaning of data science, Artificial Intelligence, Machine Learning en Deep Learning. This is followed by a discussion on how big data and AI is changing the working field of analytics. In this chapter we do not discuss details of specific analytical techniques—we do that in the in-depth Chapters 8 and 9. We then discuss how to have a greater impact with analytical results, through story-telling and visualization, in Chapter 10.

7.2 THE POWER OF ANALYTICS

In an era of big data, firms heavily rely on the analytical function. Davenport and Harris (2007) argue that firms can gain a competitive advantage if they build up strong and effective analytical capabilities: “Analytics is then defined as the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and actions” (Davenport & Harris, 2007). These analytical capabilities can be used in different kinds of functions, such as human resource management, logistics, finance, and marketing. The use of these capabilities may, for example, lead to less waste in all kinds of processes and a more effective analytical-based targeting of new and current customers, and may optimize certain decisions. This ultimately leads to higher revenues and lower costs.

A McKinsey study[1] shows how extensive users of customer analytics are significantly more likely to outperform the market (see Figure 7.1).



¹Based on “Please describe the performance of your firm/business unit in the following areas relative to your average competitor.” “Above competition” defined as 6 to 7 on 7 point scale: 1 = Well below competition, 7 = Well above competition.
²Based on “Please indicate how much you agree or disagree with the following statement: In our firm/business unit, we extensively use customer analytics.” Scale 1 to 7: 1 = Strongly disagree, 7 = Strongly degree, Comparison of low 2 vs top 2 box.

FIGURE 7.1 Extensive use of customer analytics drives corporate performance heavily

7.3 STRATEGIES FOR ANALYZING DATA

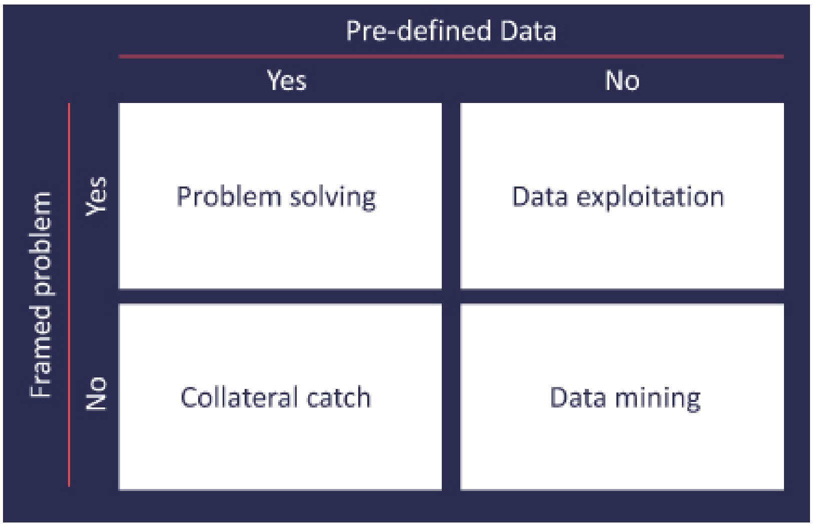
The presence of valuable data provides huge opportunities for analytical teams. One of the easiest ways of taking advantage of data is probably just to start up analyses and dig into the available data. By digging into the data one might gain very interesting insights which can guide marketing decisions. The most famous example in this respect is that of UK-based retailer Tesco: when analyzing data from their loyalty card, they discovered that consumers buying diapers also frequently buy beer and chips (Humby, Hunt, & Philips, 2008). The explanation for this unexpected correlation was that men were sent out on Friday evening to do bulky errands by car (such as buying packs of diapers) and then also bring home a crate of beer for the weekend for themselves.

Although such an example can be inspiring, we posit that before starting up an analytical exercise one should clearly understand the benefits and disadvantages of the specific analysis strategy—as well as that of other strategies.

In our two-by-two matrix, we distinguish between four basic analysis strategies (see Figure 7.2). We take into account two dimensions. First, analyses can be started based on whether or not a problem is pre-defined. A pre-defined problem can arise from:

- Marketing challenges (e.g. decreasing loyalty, eroding prices, lower acquisition rates)
- Marketing growth objectives (e.g. achieve sales growth of 20%, improve customer satisfaction).

Analytics can be done on the pre-defined data (e.g. a CRM database, customer transactions, or customer satisfaction scores), but one may also aim to look for available data or new data sources and combine these data based on emerging needs in the data analysis. This is the second dimension of our analysis strategy framework.



		Pre-defined Data	
		Yes	No
Framed problem	Yes	Problem solving	Data exploitation
	No	Collateral catch	Data mining

FIGURE 7.2 Data analysis strategies

Based on these two dimensions we distinguish four analytical strategies:

1. Problem solving
2. Data exploitation
3. Data mining
4. Collateral catch.

We will discuss the four strategies in the following subsections.

7.3.1 Problem solving

From a scientific perspective, a problem-solving analytical strategy is deductive. Usually, the analyst starts with a managerial problem or issue. Problems may include: “How can we increase the value of our customers?” or “How can we improve the net promoter score (NPS)?” or “Which pricing strategy should I use to attract more profitable customers?” After defining the problem hypotheses, assumptions could be defined, explicitly or implicitly, about potential solutions to the problem. For example, when studying drivers of NPS, analysts could develop a list of potential determinants of NPS, such as advertising, social media messages, the service experience, etc. The specific process, from defining the problem to developing a list of hypotheses, guides the selection of data to be analyzed (see Figure 7.3). Following up on the drivers of NPS, for example, researchers would probably look for available data that includes NPS and combine that with data on potential drivers. A big difference between this approach and the data mining approach is that researchers do not start with the data, but with the problem that has to be solved.

This fact-based way of working, which always starts with the business challenge, is discussed in detail in Chapter 11.

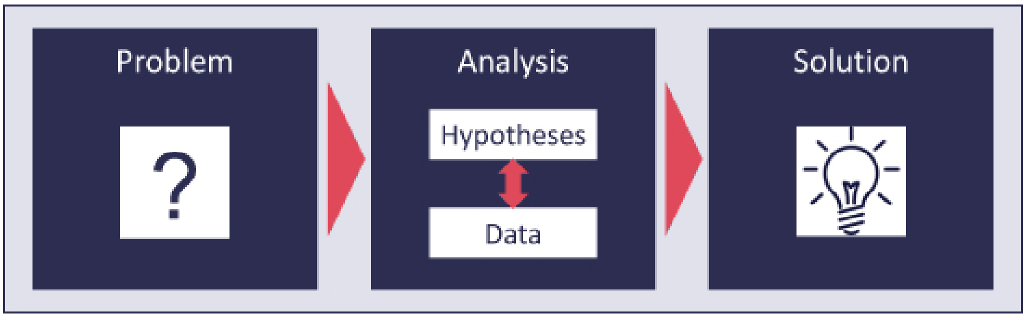


FIGURE 7.3 Problem-solving process

7.3.2 Data exploitation

There is also a pre-defined problem that needs to be solved in the data exploitation approach. For example, one objective could be to predict churn, as in the given example in Figure 7.4 which compares a model based on one data source with a model in which different data sources are combined and used to achieve the best possible prediction. The difference with the problem-solving approach is that the focus is more on data and especially on the use of new data sources. Using new data sources one could, for example, aim to find new potential predictors of churn. One potential pitfall with this approach is that analysts focus too much on the data and lack strong conceptualizations about why they find specific relationships in the data. Outcomes may easily be spurious correlations in which the association is based on an underlying, unobserved variable. The approach may also lead to undesirable outcomes, such as the famous example[2] in which a father of a teenager was offered baby products because his apparently pregnant daughter purchased baby products using his loyalty program. This is relevant to the privacy and behavioral targeting discussion put forward in Chapter 6. The approach can become overly data-driven instead of problem-driven, easily turning into a data-mining exercise. An advantage over the problem-solving approach, however, is that the data-exploitation approach is more flexible in terms of data usage. This may result in more innovative model solutions.

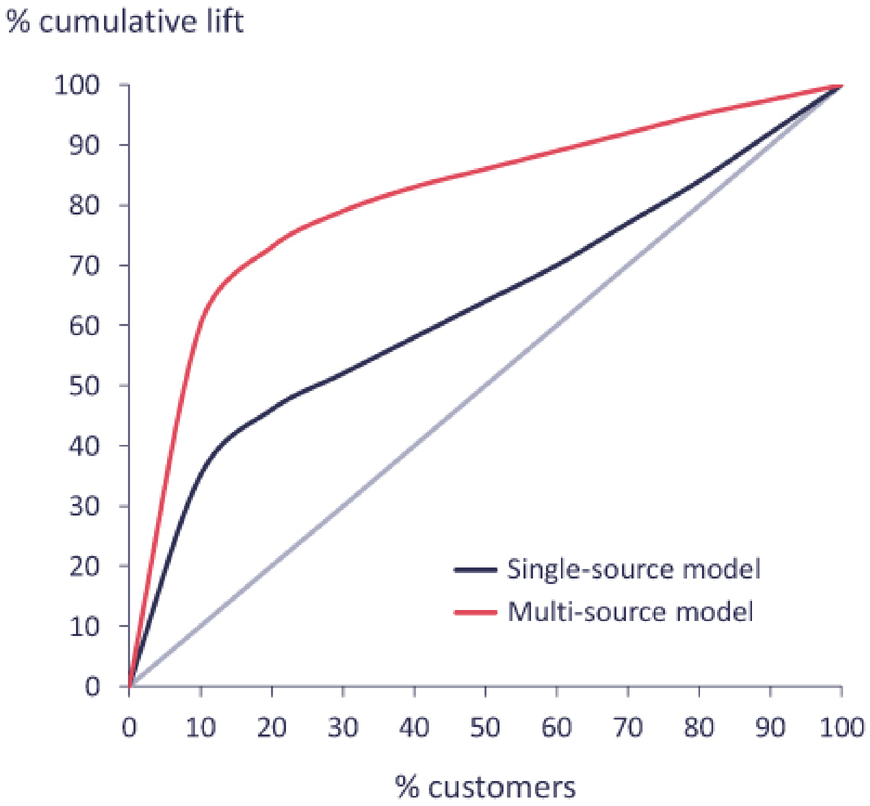


FIGURE 7.4 Model fit of a single-source versus multi-source model for churn prediction

7.3.3 Data mining

This is a much more exploratory analytical strategy. Typically, data are not predefined and a defined problem does not guide the analysis. In addition, no hypotheses are implicitly or explicitly stated. From a scientific perspective, it is a typical inductive analysis. The key belief is that the widely available data can provide valuable insights just by digging into them. In doing so, relevant new relationships can be discovered that can be potentially valuable. The classic data mining example, as mentioned earlier, is that when analyzing the Tesco Club Card data, analysts discovered that customers buying diapers also frequently purchased beer and chips on a Friday night (see Figure 7.5). This finding could be used to target promotions. The discovery of these patterns can create innovations. However, one major potential pitfall is that the analyses are unguided and may result in all kinds of associations that are difficult to interpret. Furthermore, as no specific problems are solved, many of these analyses are of little use, and may not offer any impact.

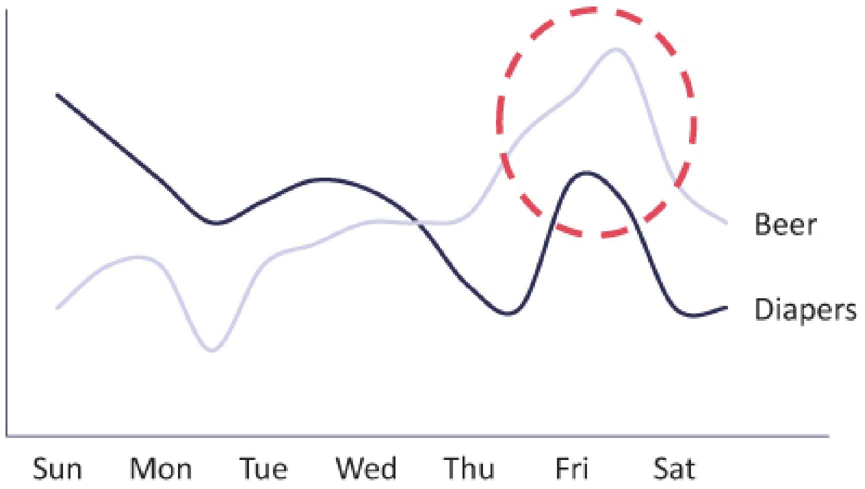


FIGURE 7.5 Beer & diaper case for Tesco

Source: Adapted from De Haan et al. (2015)

7.3.4 Collateral catch

The last analytical strategy is not explicit. When analyzing pre-defined data, analysts may sometimes discover new relationships, which can be very valuable. For example, when analyzing the impact of different touchpoints on conversion rates for an online retailer, we found differences in conversion rates for different devices (e.g. mobile, tablet, desktop), which we were not looking for in our analysis (see Figure 7.6). This induced us to study conversion rates in more depth, and specifically how device switching in the purchase funnel would impact conversion rates. These relationships are called "collateral catches" because they are not sought initially and only become available as a result of some unusual feature of the research, or through some exploratory analysis being executed next to the more structured problem-solving analysis. One thing we have learned is that analysts should be open to these collateral catches. They may provide a deeper understanding of the phenomenon studied. Furthermore, these catches may be valuable in guiding new directions for solving the defined problem, or in providing innovations for executing marketing.

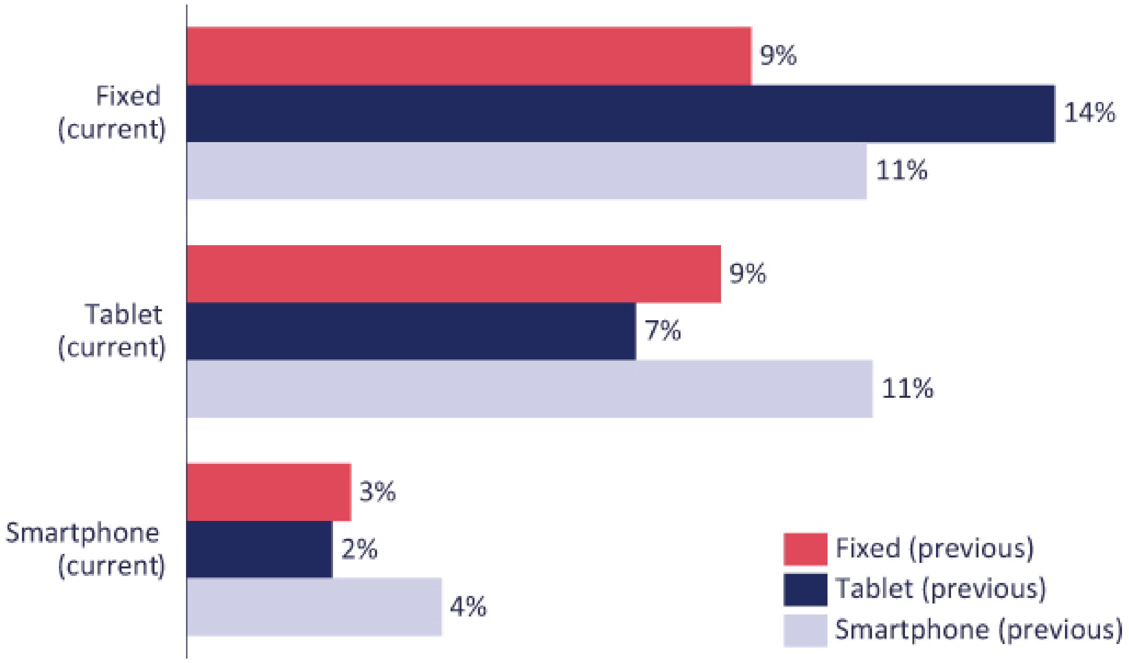


FIGURE 7.6 Different conversion rates after device switch

Source: Adapted from De Haan et al. (2015)

7.4 TYPES OF DATA ANALYTICS

Analytics is a broad concept and different types of analytics can be performed on data. Davenport and Harris (2007) distinguish several sophistication levels in analytics. A higher sophistication level should lead to a larger competitive advantage (see Figure 7.7).

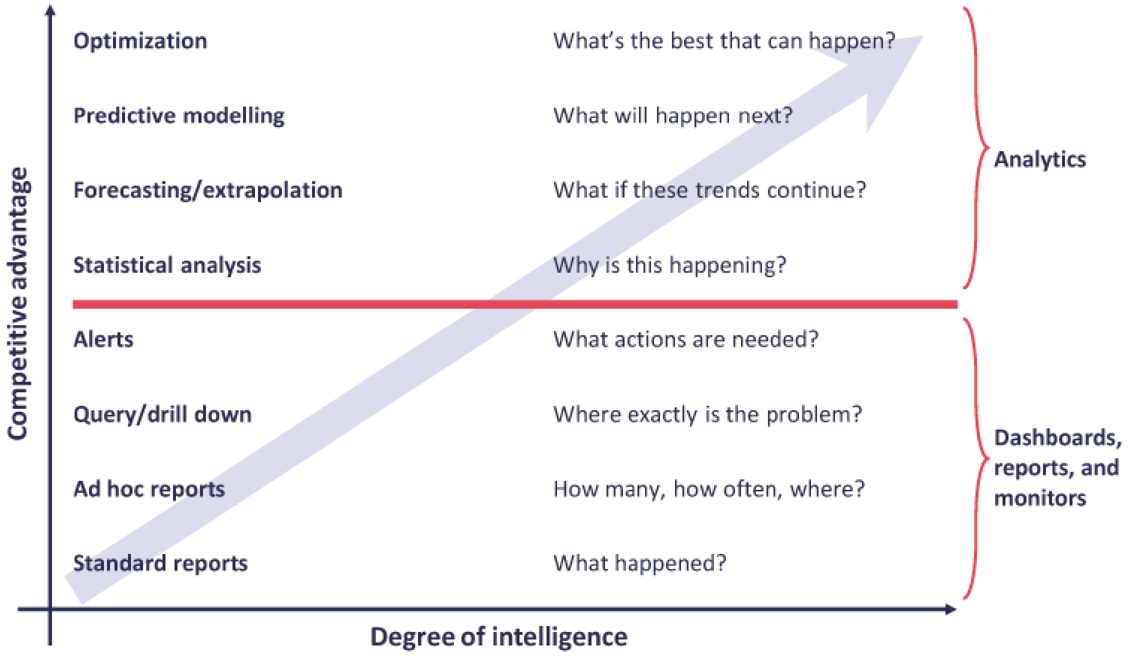


FIGURE 7.7 Different levels of statistical sophistication

Source: Adapted from Davenport and Harris (2007)

On a broad level, they distinguish between access to dashboard, reports and monitors, and analytics. Access reporting are frequently involves standardized tools—for example online analytical processing (OLAP)—within firms. This analysis focuses on gaining market and customer insights from descriptive statistical analyses. The main objectives are to learn more about markets and customers and to provide management with information on marketing and customer metrics, such as market shares, brand awareness, retention rates, and customer profitability. These types of analyses focus on what has happened (past) instead of what will or could happen (future). Usually, simple statistical techniques are used, such as calculating averages. Access and reporting are pretty standard and are often included in management dashboards provided by software suppliers such as Microsoft Power BI, Tableau, Qlik, etc. Analytics involves more sophisticated statistical techniques and answers more complicated business problems. It also focuses more on the future than on the past and is more prescriptive. Analyses here may answer questions such as “What is the optimal product assortment to offer in a store?” or “What would be the optimal price level?” Yet other questions could be “What is the optimal number of emails to be sent to a customer?” and “Through which channels should we contact a customer to optimize customer value?” According to Davenport and Harris (2007) firms using these types of analytics should be the winners in tomorrow’s business landscape.

The Gartner Analytic Ascendancy model (2012)[3] is widely used to illustrate the relationship between the different types of analytics. In Figure 7.8 we visualize this model and identify four different types of analytics, ranking these in terms of value and complexity. Here, we start with the simplest one and go further to the more sophisticated types.

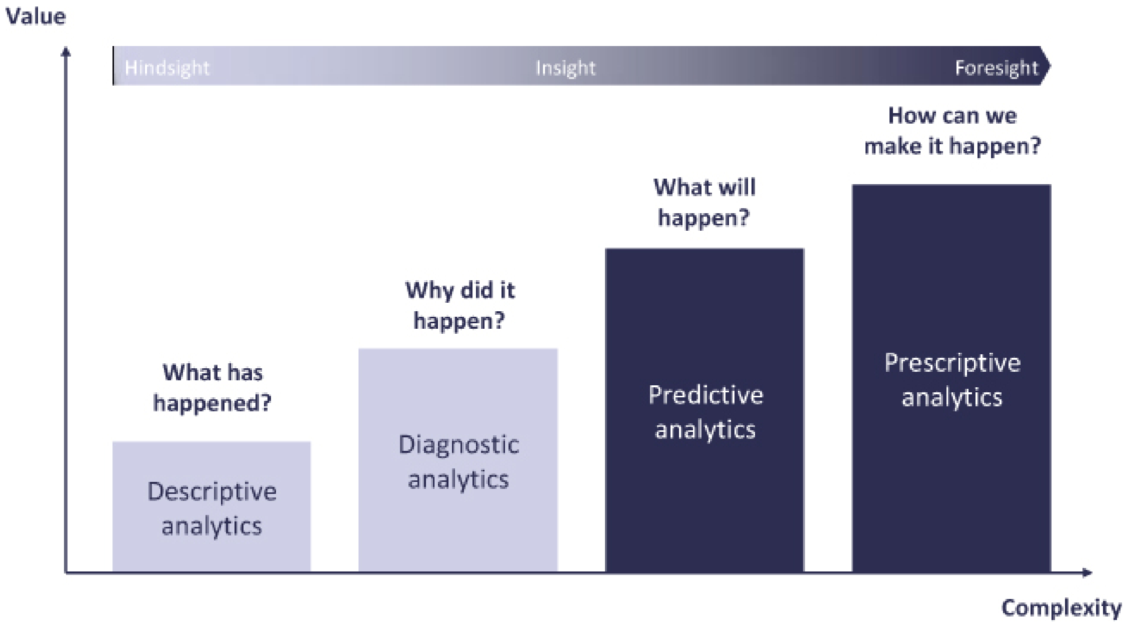


FIGURE 7.8 Different types of analytics

Source: Adapted from Maoz (2013)

7.4.1 Descriptive analytics

Descriptive analysis is the simplest form of analysis. However, its importance should not be underestimated because descriptive analysis is the first step in a research process and it provides the starting point for further analysis. The power of descriptive analysis is the fact that relatively simple insights, such as mean, median, mode, or minimum and maximum values, can be used to answer the question “What has happened?” For example, a web shop can gain insight into its average daily/weekly/monthly sales volume, or an insurance company can gain insight into the number of website visits (see example in Figure 7.9) or incoming customer service calls. However, these findings simply indicate what happened or that something is wrong or right, without explaining why it happened.

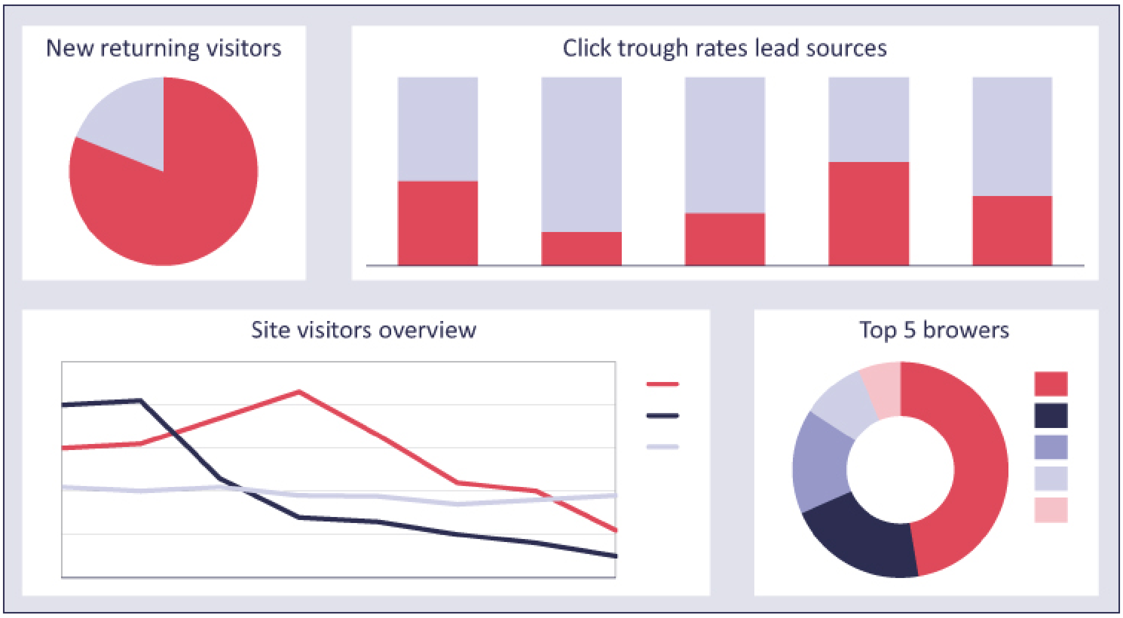


FIGURE 7.9 Example of a dashboard with descriptive analytics of website visits

7.4.2 Diagnostic analytics

An answer to the question of why something happened can be obtained through diagnostic analytics. To do this, a deeper analysis is necessary to examine the context and causality of a specific phenomenon. You do this by looking not only at averages but at differences, by examining patterns, looking back in time, and comparing historical data with other data. Some commonly used techniques are drill-down, correlations, migrations, trend and time series analysis, and pattern recognition.

Diagnostic analysis can provide answers to questions such as "How effective was our promotional campaign?" Or "Why has the outflow of our customers increased?" Or "Why are sales volumes lagging behind forecasts?" Or "What caused the fall of our revenues in the past year?" Figure 7.10 provides an example of a diagnosis in which the last question is answered.

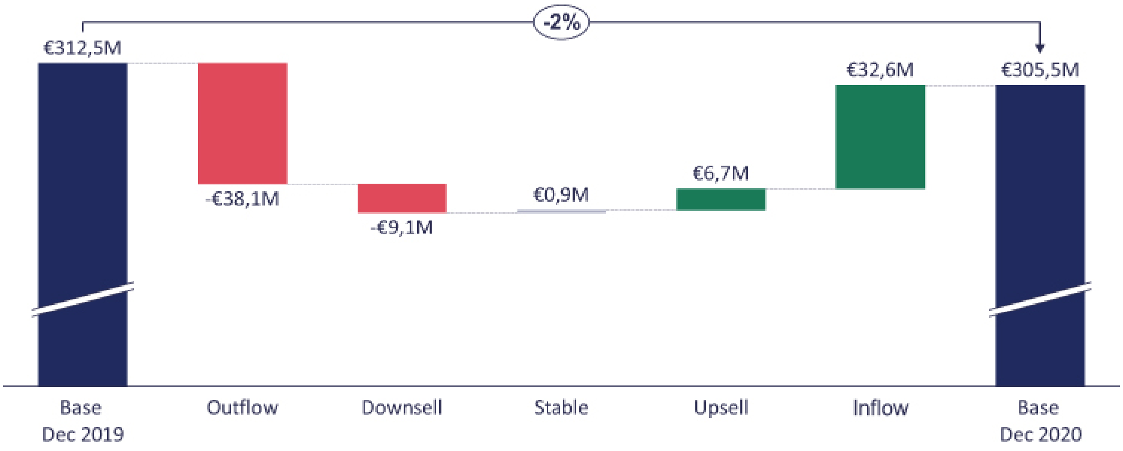


FIGURE 7.10 Example of diagnostic analytics to decompose the development of the total revenue of a telecom operator over time (in €M)

A good diagnosis leads to a fact-based substantiation of a business question. This can be used to determine whether it concerns a large or small challenge to solve a problem and creates a common understanding and support for tackling an issue. However, diagnostic analytics do not provide insights into how a problem can be solved.

7.4.3 Predictive analytics

Predictive analytics can tell what is likely to happen in the future. The findings of the descriptive and diagnostic analytics form the basis for more advanced analytics, with which, for example, future trends, response in customer behavior can be predicted. Commonly used methods and techniques are regression, decision trees, and classification, as well as techniques that are classified as machine learning and deep learning.

Predictive analytics can answer questions such as “What effect does the brand campaign have on sales volume?” Or “What is the price point that I should choose to maximize our market share and revenue?” Or “What stock should I hold to meet the expected demand next month?” Or “How should I optimize the channel mix to improve the net profit?” (See the example in Figure 7.11.

Note that predictive analyzes provide a probability of the occurrence of a specific event, so it is not guaranteed that an event will occur. The accuracy of the prediction strongly depends on the quality of the data and the stability of the situation. Additionally, it is important to continuously review and optimize prediction models.

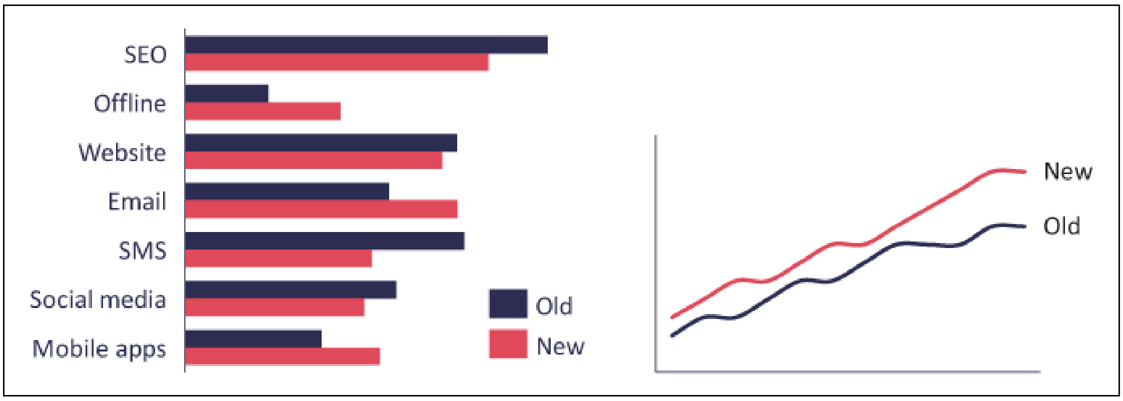


FIGURE 7.11 Example of predictive analysis to optimize the channel mix of an insurance company to improve the net profit

7.4.4 Prescriptive analytics

The last type of analysis answers the question of what action to take to prevent a future problem or to seize an opportunity. Prescriptive analytics builds on the results of predictive analytics to suggest favorable outcomes and determine what action to take to achieve a particular outcome. A well-known example of prescriptive analysis is the recommendation systems used by Netflix and Amazon. Another example is the personalization of website content based on historical search behavior or purchase history. Figure 7.12 shows a simplified display of prescriptive analytics like collaborative and content filtering that is used in recommendation and personalization systems.

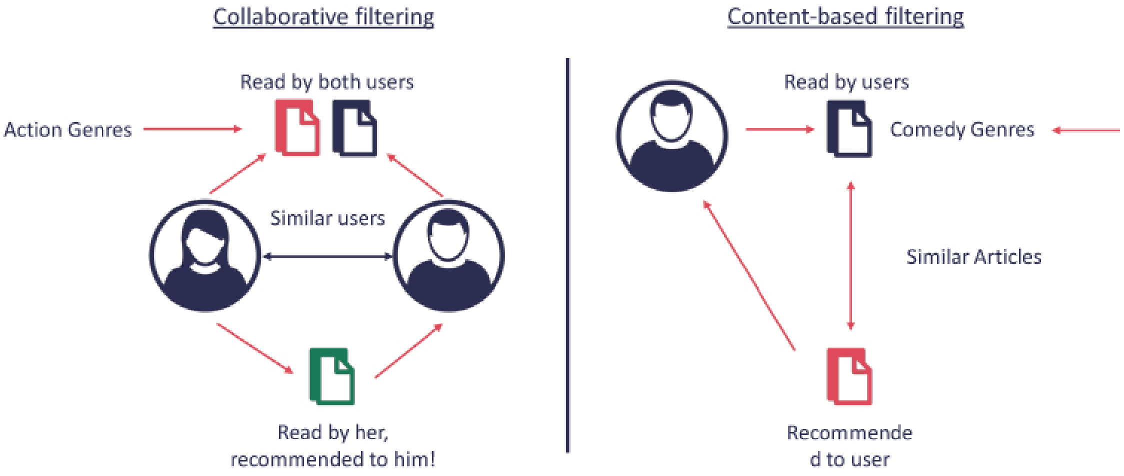


FIGURE 7.12 Example of simplified display of prescriptive analytics that are used in recommendation and personalization systems

Source: https://data-flair.training/blogs/data-science-at-netflix/[5]

Prescriptive analytics leads to business rules and algorithms to ensure that prescribed actions are performed automatically and in real-time. When prescriptive analytics uses advanced tools and technologies, like machine learning and AI, it places high demands on the IT environment.

Unlike the previous three types of analysis, prescriptive analysis often uses a feedback system in order to learn and to improve the relationship between prescribed actions and their results.

7.5 HOW BIG DATA AND AI CHANGE ANALYTICS

The growth of big data is changing how analyses are being executed. It is changing the scope of analytical questions to be answered as more data are available. Also the new and continuous developments in domains such as data science, machine learning, and Artificial Intelligence have impacted how we perform analytics nowadays. Although the disciplines have common ground, they are certainly not synonymous with each other.

We first discuss what these terms mean and how they differ from or relate to each other. Then we will examine how developments in these disciplines change the working field of analytics.

7.5.1 Definitions

The Venn-diagram shown on the next page (see Figure 7.13) visualizes overlapping analytics-related terminology. Big data has already been discussed in Chapter 4. Here we discuss the other terms one by one.[4]

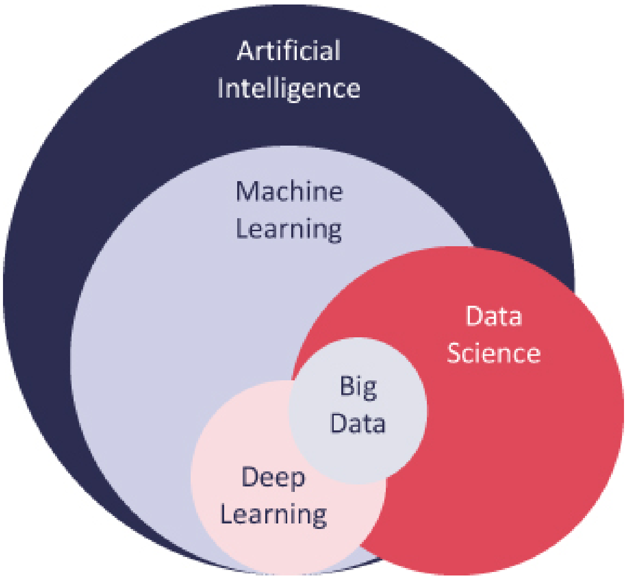


FIGURE 7.13 The Venn-diagram that visualizes overlapping analytic-related terminology

7.5.1.1 Data science

Data science is a multidisciplinary field aimed at gaining insights based on data, which can help an organization to make better decisions. Predictive analytics makes it possible to identify hidden patterns in data that you didn't even know existed. And data-driven insights (or prescriptive analytics) can be extremely valuable for targeted actions and campaigns. And with the availability of more and more amounts of data (big data), the value of the use of data science is increasing.

Many data science applications sound like AI applications. This is because data science overlaps the field of AI in many areas. A data scientist uses tools such as statistical modeling, visualization methods, and machine learning algorithms, which are also used to develop AI solutions.

7.5.1.2 Artificial Intelligence (AI)

The concept of AI has been around since ancient times. The foundation of modern AI was laid by classical philosophers who tried to describe human thinking/intelligence as a symbolic system. The programmable digital computer, the machine based on abstract mathematical reasoning, was first invented in the 1940s based on this knowledge.

The term "Artificial Intelligence" was introduced in 1955 by John McCarthy, a computer scientist. AI research at the time focused on neural networks, inspired by the way neurons work in the human brain.

But building artificially intelligent machines was not that easy; the limitations in the computing power of the computer also hindered the progress of AI. For decades, AI was restricted to research labs. AI took off in the early 2000s when major tech giants started building supercomputers and investing in AI.

AI is still under development and is considered a very broad term. It is difficult to give an unambiguous definition for AI. But you could think of AI as the power we can give a machine to enable it to:

- Understand / interpret data
- Learn from data
- Make "intelligent" decisions based on insights and patterns from data.

In that regard, an AI-controlled machine performs tasks by simulating human intelligence. Where AI is often able to perform more than is possible for a human being.

7.5.1.3 Machine learning (ML)

The term machine learning (ML) was coined by Arthur Samuel in 1959. ML is a subset of AI. It is used when you need machines to learn from huge amounts of data. The acquired knowledge is applied to a new set of data. ML allows a machine to learn from (or about) newer datasets without giving it explicit instructions.

Some of the most common methods of 'making machines learn' are:

- Supervised learning
- Non-supervised learning
- Reinforced machine learning.

Some methods pre-inform the machine of independent (input) and dependent (output) variables. The machine learns the relationship between these two types of variables by analyzing a set of data called the "training data set."

Once a machine is sufficiently “trained” or when an ML model is ready, it is applied to a new set of data called the “test data set.” The ML model does not enter production mode until it has been adequately tested for reliability and accuracy. In Chapter 9, we elaborate on the different methods and techniques of ML.

ML techniques have evolved a lot in recent years. Even programmers with no background in statistics or no training in AI can build, train, test, and implement ML models. Yet it remains important to know exactly how different ML algorithms work and how this leads to outcomes. That is why the use of a data scientist in deploying ML is indispensable.

7.5.1.4 Deep learning (DL)

DL is a subset of ML. In general, ML is suitable if the data set is relatively small.

DL is preferred when:

- The data has too many features
- The data set is huge
- Extremely high level of accuracy is required.

Deep learning is based on artificial neural networks. Deep learning enables computers to learn new things from large amounts of data. Compared to ML, DL can solve more complex problems, but it is more difficult to implement, requires specialized hardware to run, and requires more time to train the model. In Chapter 9 we take a closer look at how neural network models work and are applied.

7.5.2 Important changes in the analytical working field

In this section, we discuss our views on how big data and developments in AI, ML, and DL change the working field of analytics. This is not an exhaustive explanation but we discuss six important aspects here.

7.5.2.1 From analyzing from a single source to multi-source data

A very important development concerning data is that we are moving from single-source data to multiple sources of data. This has considerable consequences, as with multi-source data you have to connect data intelligently. This is not always obvious: connecting can seem easy but one needs a common variable that can be linked. Connecting data may also involve data connecting at multiple levels. For example, customer data can be linked to time-series data such as that arising from advertising; one then has data at different aggregation levels. The challenge here is that there is much variation in data between customers, but that the variation of time (i.e. in advertising or distribution) may be limited, while the number of data points at that level is also small compared to the number of data points at the customer level. This may also leads to the use of more complicated models, such as multi-level models.

7.5.2.2 From analyzing structured to unstructured data

Another important development is that data is becoming more unstructured, with text data being an important example. But social network data can also be included, especially at the customer level. Furthermore, the large amounts of data generated by the increased sharing of images and videos via social media are an example of unstructured data.

As a result, this data is complex and raw and often also generated in high volumes. To get started with unstructured data, new processes often have to be set up to unlock the data and many data operations may have to be carried out. For organizations it often takes a lot of time and effort to get started with unstructured data. In practice, we frequently see frustration and misunderstanding arise, while the added value of unstructured data is still being used sparingly.

7.5.2.3 From analyzing samples to analyzing the full population

The increasing volume of data in a big data era suggests that we can now analyze very large databases with millions of observations. We are moving from studying a small sample to studying the full population. The rising computer power of the last few years has facilitated this, while data storage capacity seems unlimited, with sufficient space in the cloud. It sounds very impressive and convincing when an analyst can report that millions of observations have been analyzed. The advantages of analyzing very large databases are, however, relatively small. In reality, the outcomes of an analysis of a truly random sample of observations of a population of millions should not differ substantially from the analysis of the full population. It is not the volume that is important, but whether the analyzed sample is representative of the target population. The volume of data is only important in getting more reliable answers at a higher significance level. An analyst should, however, be more interested in biases due to a wrong representation of customers, brands, etc. in their data, rather than in the volume of the data. This is not to say that the size of the analyzed sample is not relevant. Increases in lower sample size numbers (e.g. moving from 400 to 2,000 observations) can be especially valuable, as reliability may increase and specific econometric issues, such as collinear independent variables, may also become less of an issue. However, moving from 20,000 to 50,000 or 100,000 observations will become less rewarding for analytical purposes. In general, we provide the following simple rules for when larger samples become more valuable:

- If more variables are studied in an analysis and specifically when studying the effects of multiple variables on an outcome variable
- If one needs to study different sub-samples. The sub-sample should be of sufficient size to analyze.
- If there is lots of heterogeneity. For example, customers may differ strongly in how they behave, and how they respond to marketing
- If the studied variable occurs very rarely (e.g. conversion on an email campaign) and a sufficient number of data points is required to understand the drivers of this event
- If there is strong collinearity between the independent variables used to predict or explain a dependent variable, such as sales or churn.

Although the volume of available data is increasing, we observe that this is not the case for all types of data. At the individual customer level, data volumes have indeed become huge and can be enriched with many other data sources. Especially in the online environment, data can become massive. However, brand-level data on, for example, brand sales, brand preferences, and advertising are frequently still limited. For example, for a European public transport company, we analyzed the effects of advertising on traveled kilometers per month. While having three years of data this only resulted in 31 data points (Gijsenberg & Verhoef, 2015).

7.5.2.4 From significance to substantive and size effects

Analyzing large samples frequently reveals many highly significant effects. Small p-values seem the norm rather than the exception. However, in a big data era with these large samples, we argue that we should move away from significance and focus more on the size of these effects. With effect size, we look for whether the found effects are substantial or small. For example, there might be a difference of one year in the average age between switching and loyal customers (e.g. 43 vs. 42 years), which is highly significant in a big data analysis. However, one might question the size of this difference. Should a firm then focus more on younger customers to prevent switching, as these customers are less loyal? Similarly, the addition of variables in an explanatory model will certainly frequently increase the explained variance of the model.

The key focus should then be on the size of this additional explanatory power —is it substantial or is it only incremental? Further, it is not only the size of the effect that should matter but also the substantive and managerial meaning of a found effect. Going back to the loyalty example, one might question the managerial implication of such an effect. Should a firm develop different specific strategies for younger customers than they do for older customers, as young customers are more likely to switch? We have some specific recommendations for actions that an analyst should take when interpreting big data results:

- Focus on the size of the significant effects instead of significance only
- Visualize the found effects and consider the effect sizes
- Compare the found effects with existing benchmarks (e.g. when assessing the effectiveness of social media, one could compare it with the effect of traditional advertising)
- Develop marketing implications for found effects and challenge them.

7.5.2.5 From standard to machine learning/ computer science models

In recent decades computer scientists have developed various methods to analyze large amounts of data in a faster and smarter way. One of the best-known early examples was the so-called “neural networks.” During recent years, many new machine learning methods have been added, such as random forest, bagging, and boosting. These estimates are based on standard models, such as decision trees and regression, with a large number of subsamples, in the hope of make the -most accurate prediction.

Evidence of stronger data science modeling performance is mixed. In general, it seems difficult to come up with the best model for all applications. Donkers, Lemmens, and Verhoef (2014) study about 14 different databases and report that no method wins consistently. The likelihood that a method will perform best seems to depend heavily on the data being analyzed. They therefore suggest that the best way to get good results is to combine the different methods and take advantage of the power of each of them.

One of the main problems many people face with machine learning-based modeling is that it is often viewed as a “black box.” That is, it is not clear to analysts which of the input variables affect the output variable. Moreover, the direction of these effects is not clear, nor is how it is specified. Models in marketing generally used to have an econometric background and were therefore explainable and specified (e.g., Leeflang *et al*., 2015). An advantage of these models is that they are strongly based on predetermined hypotheses and effect expectations, while machine learning models can easily lead to data mining without a strong theoretical and practical basis. If algorithms lead to conclusions without an understanding of how they have done so, the conclusions cannot be explained. This creates a lot of ethical and practical problems. More and more examples are coming to light of such AI-models containing discriminatory effects that lead to undesirable outcomes. Organizations must increasingly be accountable by law and regulations for explaining why certain decisions are made. They can no longer hide behind mathematical models.

Some recommendations for analysts to follow when applying models are:

- Understand the backgrounds and pitfalls of (new) models before applying them
- Be inherently skeptical about the communicated performance of (new) models from software providers
- Test the performance of different methods on data being analyzed
- Select a method that can be communicated to marketing managers and has a good performance. That is, find the optimal balance between performance and managerial insightfulness
- Use visual aids to communicate findings from specifically computer-science-based methods.

7.5.2.6 From ad hoc models to real-time models

In an environment where data is readily available, we move to a position where models can be updated much more frequently. The need for model updates is clearly shown in extant research. Especially in more turbulent environments, models can easily become outdated and as a consequence, their findings are no longer valid and their prediction quality decreases. For example, in a telecom setting Risselada, Verhoef, and Bijmolt (2010) show that the predictive performance of models decreases substantially as time moves further away from the point at which the model was developed.

Luckily, data is now more readily available so that updating models is much easier. Especially in the online environment, data can be available in real-time. This provides opportunities for constantly updating them with new data in constantly changing (online) environments. These models can then be used in online targeting. This move to constantly updated models probably will not happen in offline environments. However, here also we advocate more model updating, as it is unlikely that model results hold good for a long time. The dynamics of today’s market, with frequently changing market environments and changing customer behaviors, require a constant updating of models. We have some recommendations:

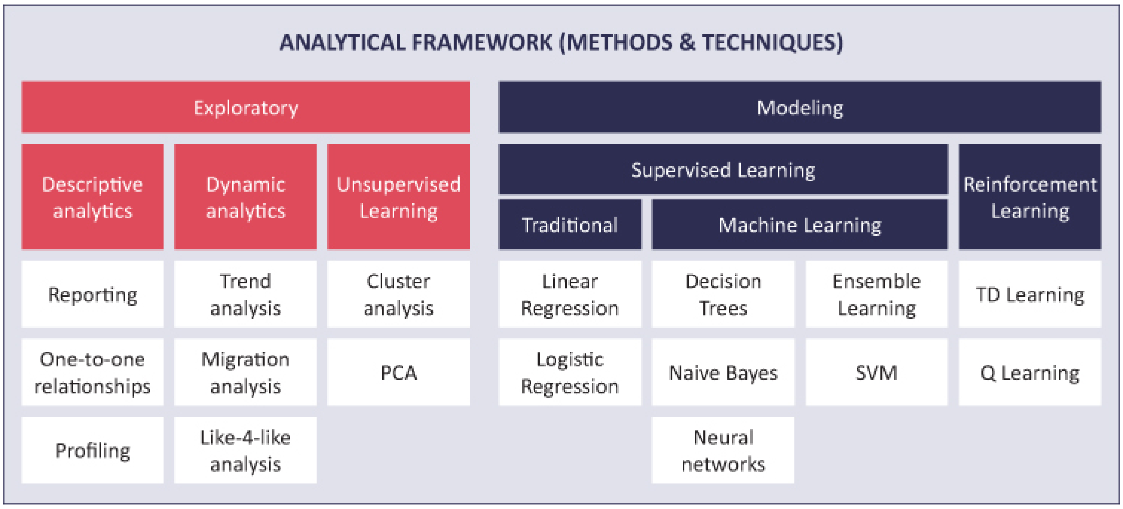
- Assess the stability of a developed model and the predictive performance of a model over time
- Do not expect your model-results, resulting insights, and predictive performance to have eternal life
- In online environments base models can be updated to real-time models for targeting purposes with real-time data as input.

7.6 ANALYTICAL METHODS AND TECHNIQUES

Data science is a broad discipline that allows for turning raw data into understanding, insight and knowledge to ultimately make better decisions, create successful campaigns, and develop data-driven solutions. In this book we want to provide a solid basis in the most important methods and techniques that can be used in the exploration phase and specifically for modeling data.

Chapter 8 focuses on methods and techniques that can be used for data exploration, rapid hypothesis generation and testing. Modeling data is often also an important part of the exploratory process, but in Chapter 9 we focus on methods and techniques to develop models that generate predictions.

In our analytical framework, we provide an overview of the methods and techniques (see Figure 7.14), which we discuss in Chapters 8 and 9, where we distinguish between exploratory analyses and models. Both types of analyses have subcategories, which include various methods and techniques. These are discussed in Chapters 8 and 9, respectively.



ANALYTICAL FRAMEWORK (METHODS & TECHNIQUES)

Exploratory			Modeling			
Descriptive analytics	Dynamic analytics	Unsupervised Learning	Supervised Learning			Reinforcement Learning
			Traditional	Machine Learning		
Reporting	Trend analysis	Cluster analysis	Linear Regression	Decision Trees	Ensemble Learning	TD Learning
One-to-one relationships	Migration analysis	PCA	Logistic Regression	Naive Bayes	SVM	Q Learning
Profiling	Like-4-like analysis			Neural networks		

FIGURE 7.14 Analytical framework with different types of analysis methods and techniques

The last step in data science is communication, an absolutely critical part of any data analysis project. It doesn't matter how good the insights and models are, they only have value if others also understand what their meaning is and what you can do with them. In Chapter 10 we pay detailed attention to how storytelling and visualization can create an impact for data and analytics.

7.7 CONCLUSIONS

In this chapter, we focused on analytics, more specifically, the power of analytics. We strongly believe that analytics can improve the quality of marketing decisions and produce smarter marketing decisions. We made a distinction between different strategies for analyzing data. Approaches that start with the nature of a framed problem are our preference. Analytics can focus on building insights, (predictive) models, and optimization models. We have clearly laid out different analysis types illustrated with practical examples.

Big data and AI are changing the working field of analytics. We have discussed these changes and specifically focused on some frequently stated developments.

NOTES

1. Source: Why customer analytics matter, May 26, 2016, article by Lars Fiedler, Till Großmaß, Marcus Roth, and Ole Jørgen Vetvik.
2. How Companies Learn Your Secrets, By Charles Duhigg February 16, 2012 New York Times.
3. Maoz, M. (2013), "How IT should deepen big data analysis to support customer-centricity", Gartner.
4. Adapted from https://medium.com/ai-in-plain-english/data-science-vs-artificial-intelligence-vs-machine-learning-vs-deep-learning-50d3718d51e5
5. Weblink: https://data-flair.training/blogs/data-science-at-netflix/ (Accessed: 2019)

REFERENCES

Davenport, T. & Harris, J. (2007). *Competing on Analytics – The New Science of Winning*. Boston: Harvard Business School Press.

De Haan, E., Kannan, P. K., Verhoef P. C., & Wiesel T. (2015). The impact of device switching on conversion rates. *Working Paper*, University of Groningen.

Donkers, B., Lemmens, A., & Verhoef, P. C. (2014). The predictive power of churn models. *Working Paper*, Erasmus University Rotterdam.

Gijsenberg, M. J., & Verhoef, P. C. (2019). Moving Forward: The Role of Marketing in Fostering Public Transport Usage. *Journal of Public Policy & Marketing*, *38*(3), 354–371.

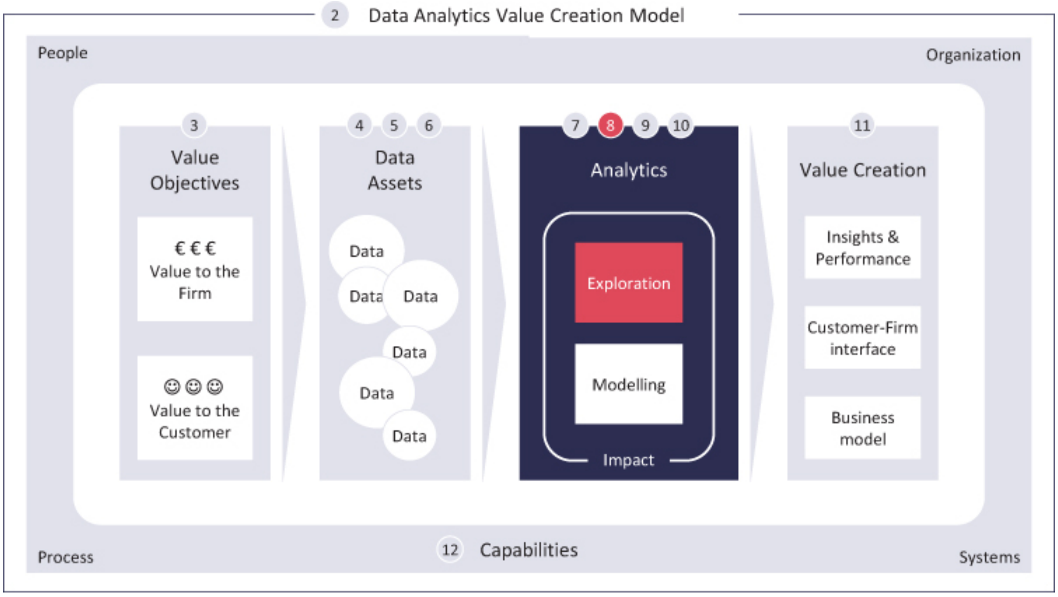
Humby, C., Hunt, T., & Phillips, T. (2008). *Scoring Points: How Tesco Is Winning Customer Loyalty*. Philadelphia: Kogan Page Publishers.

Leeflang, P., Wieringa, J. E., Bijmolt, T., & Pauwels, K. (2015). *Modeling markets: Analyzing marketing phenomena and improving marketing decision making*. New York: Springer-Verlag.

Risselada, H., Verhoef, P. C., & Bijmolt, T. H. A. (2010). Staying power of churn prediction models. *Journal of Interactive Marketing*, *24*(3), 198–208.

CHAPTER 8
Data exploration

DOI: 10.4324/9781003011163-8



8.1 INTRODUCTION

In this chapter, we discuss data exploration. The techniques discussed in this chapter can be used to examine new incoming data for patterns, regularities, and to get a "feel" for what issues may be relevant in the data. The techniques in this chapter are mainly data driven. That is, we do not assume a formal model based on psychological or economic theory to be driving the relationship between the variables in the analysis. Instead, we remain open-minded and "take what the data give us." In that respect, the purpose of this chapter is more descriptive and exploratory rather than testing of hypotheses, and we do not perform the analyses to confirm a theory with empirical results, but more to investigate empirical results to identify what insights the data can potentially generate. In Chapter 9, we take an approach where theory informs our models, and we want to confirm or reject hypothesized relationships that the theory proposes. The techniques discussed in this chapter can be divided into two broad categories: descriptive analyses and unsupervised learning.

In descriptive analyses, we are interested in sketching the overall picture of variables relevant to our data science project. That encompasses examining common metrics such as the mean, the range, or the shape, and presenting them (or combinations thereof) either as numbers or in a graphical way. Examples are tables with descriptive statistics, histograms, and boxplots. As we assume that the reader is familiar with such techniques, we will not discuss them here, but refer to books on basic statistics and marketing research (see e.g., Malhotra, 2019). We do not omit these techniques from this book because we believe that they are unimportant. On the contrary, in many cases, they offer very useful and quick insights into the data and may even provide answers that are useful for the question at hand. For example, when developing a storyline (see Section 10.3), overall statistics are generally very useful in providing evidence-based information about the situation.

However, descriptive analyses also involve initial exploration of the data, and that is what we want to devote our time and attention to in this chapter. Examples of such initial analyses are reporting, one-to-one analyses, profiling, trend analysis, and migration analysis, and these are discussed in Sections 8.2–8.5.

In Section 8.6 we turn to unsupervised learning as it provides us with a data-driven approach to identify patterns and structure in the data. In Section 8.6.1 we discuss cluster analysis, and in Section 8.6.2 we discuss principal components analysis. Figure 8.1 provides a schematic overview of the topics that are discussed in this chapter.

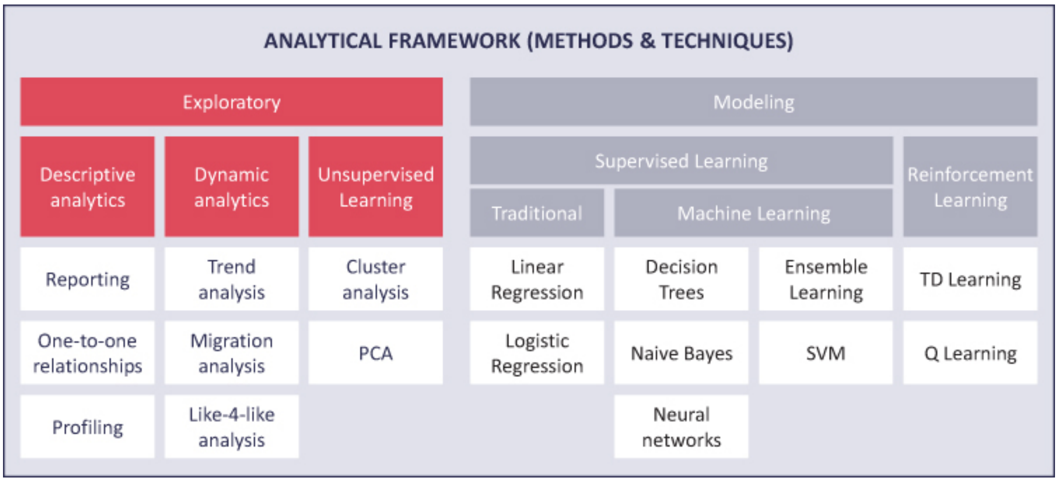
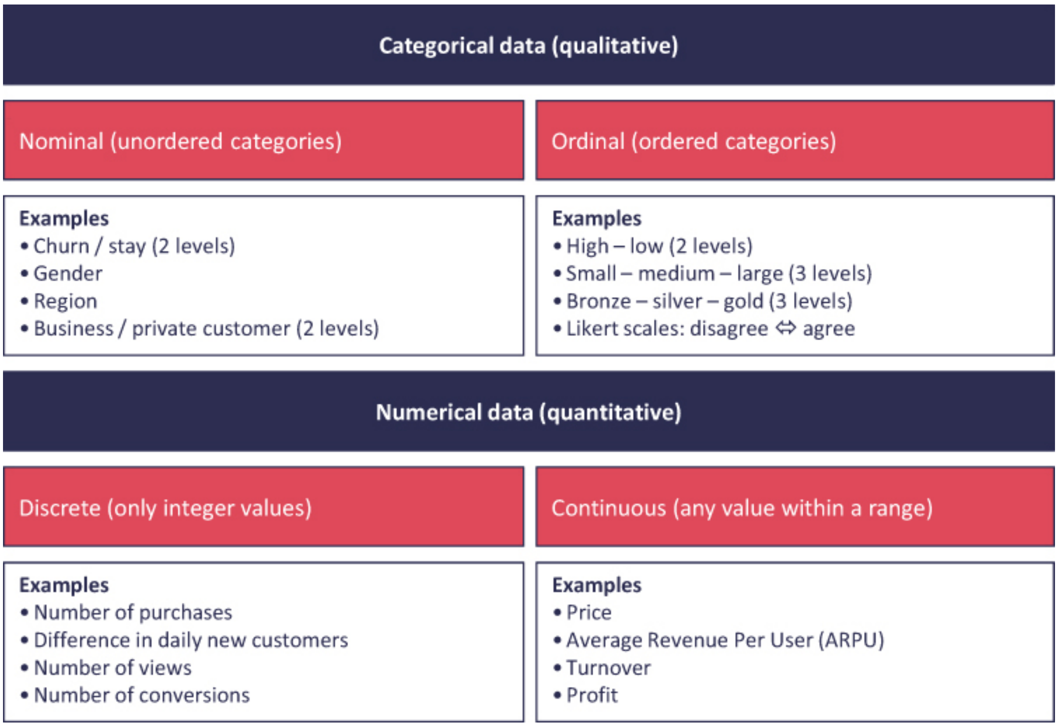


FIGURE 8.1 Analytical framework with different types of analysis methods and techniques

8.2 DESCRIPTIVE ANALYSES—REPORTING

With reporting, analysts aim to provide management information on some relevant descriptive statistics about specific KPIs, such as market share, customer satisfaction, and/or customer profitability. Reporting is especially important in marketing dashboards and specific tooling to facilitate this has been developed in business intelligence software such as Tableau, Qlik, and Microsoft Power BI. Averages are probably most frequently reported (e.g., average satisfaction score) and should be relatively straightforward in most cases. Some important points when reporting KPIs include:

– Be aware of the measurement scale, as averages are meant for variables that are measured on a metric scale. In Figure 8.2 we distinguish between numerical data and categorical data. For categorical data, averages cannot be calculated, and percentages or rates should be reported. For instance, churn can be measured with a simple yes/no question, and the churn rate can be used to summarize such a variable.



Categorical data (qualitative)

Nominal (unordered categories)

Examples
- Churn / stay (2 levels)
- Gender
- Region
- Business / private customer (2 levels)

Ordinal (ordered categories)

Examples
- High – low (2 levels)
- Small – medium – large (3 levels)
- Bronze – silver – gold (3 levels)
- Likert scales: disagree ⇔ agree

Numerical data (quantitative)

Discrete (only integer values)

Examples
- Number of purchases
- Difference in daily new customers
- Number of views
- Number of conversions

Continuous (any value within a range)

Examples
- Price
- Average Revenue Per User (ARPU)
- Turnover
- Profit

FIGURE 8.2 Types of data

– Averages can be misleading: just monitoring averages does not always provide sufficient insight, especially in cases with a high level of heterogeneity. Therefore, we strongly recommend considering the spread in the variables as well, or more generally to investigate their distribution. To describe the distribution, additional descriptive statistics, such as the standard deviation or the kurtosis, can be useful, which can be illustrated by graphical means such as boxplots or histograms. For example, a histogram can distinguish between two cases depicted in Figure 8.3, where the left panel indicates that an average satisfaction score of about 6 is a good summary of satisfaction. The average satisfaction score is also about 6 in the right panel of Figure 8.3. However, only reporting the average score conceals the fact that there are two segments of customers: apparently there are very dissatisfied customers (scoring around 4) and very satisfied customers (scoring around 8).

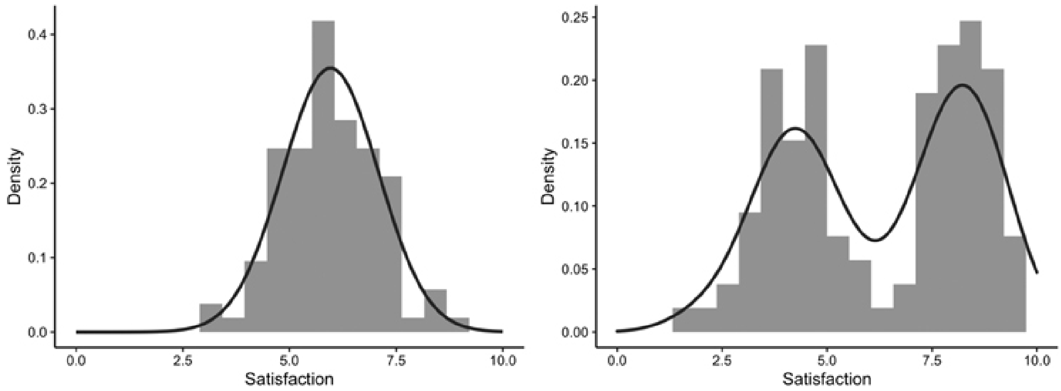


FIGURE 8.3 Different distributions with similar averages

– Focus on extremes: an average performance is often not enough to outperform the competition. Firms need to satisfy customers and have highly evaluated brands in order to compete successfully. Marketing metric reporting should therefore go beyond only reporting averages and also consider extreme responses. This is particularly important for customer feedback metrics, in which top2-boxes or specific transformations such as the net promoter score (NPS) can be very valuable (see also our discussion in Chapter 3).

– Report trends: reporting on the current status can be informative but managers will be more interested in the trend. Is the churn rate going down? Are customers becoming more satisfied? Is our brand image improving? Is the market growing? In this sense, specific trend metrics, such as growth rates, can be very informative descriptors. Static descriptors then become more dynamic and will raise specific questions (see Figure 8.4 for a sales trend). This will be further investigated in Section 8.5.

Finally, we note that reporting descriptive analyses can be very valuable as the first step in an analysis. The analyst gains more knowledge about the different data and their development over time. Moreover, a descriptive analysis can indicate whether there are specific mistakes in the data and/or specific outliers (abnormal observations). We therefore strongly recommend that analysts conduct some descriptive analyses before executing more complicated techniques.

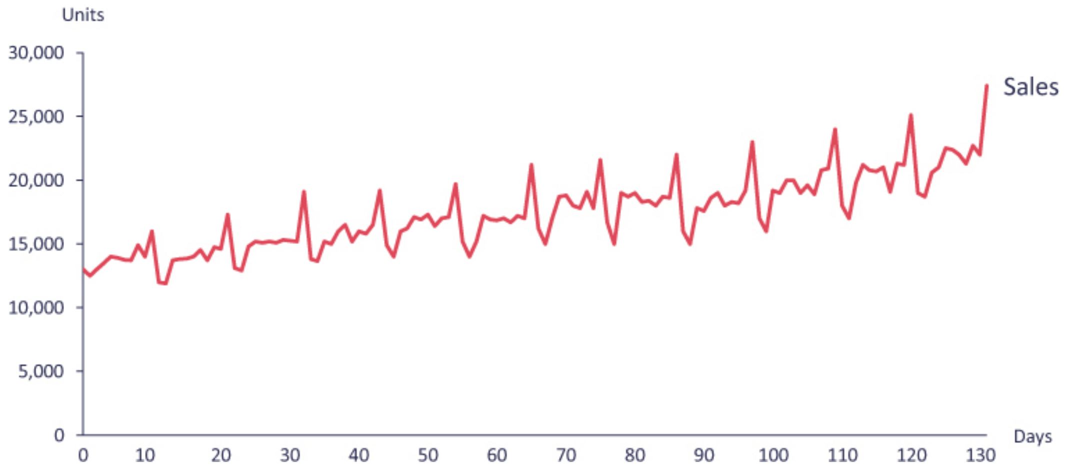
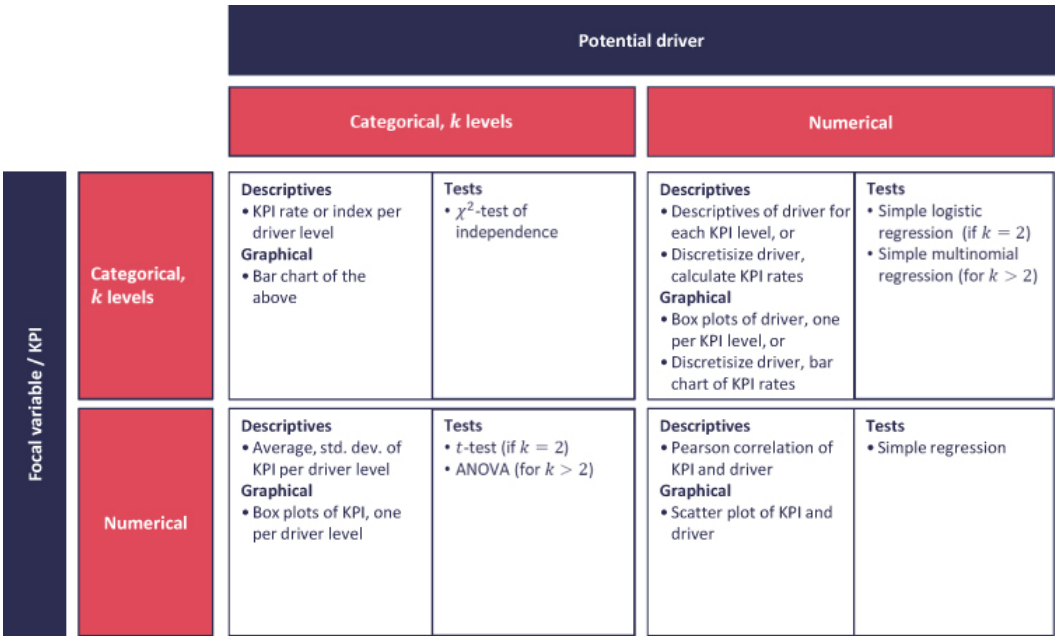


FIGURE 8.4 Example of time series for sales

8.3 DESCRIPTIVE ANALYSES—INVESTIGATING ONE-TO-ONE RELATIONSHIPS

As a second step in a descriptive analysis, managers might be interested in finding out where the variation in the KPI originates from. For example, in the right-hand side panel of Figure 8.3, we suggested that the variation in satisfaction may be due to the existence of two segments. To investigate this further, it is often enlightening to conduct a drill-down of the KPI according to some possible drivers, for example segments or age. Hence, in many situations, it is of interest to investigate whether the KPI is different for different levels of another variable. In Figure 8.5, we summarize possible initial analyses that can be conducted to investigate such one-to-one relationships. In line with the previous discussion, the measurement level of the variables should be taken into account when examining relationships. Since we have distinguished two types of data in Figure 8.2, and we focus on two variables (a KPI and a possible driver), we identify four main cells in Figure 8.5.



		Potential driver			
		Categorical, k levels		Numerical	
Focal variable / KPI	Categorical, k levels	**Descriptives** • KPI rate or index per driver level **Graphical** • Bar chart of the above	**Tests** • χ^2-test of independence	**Descriptives** • Descriptives of driver for each KPI level, or • Discretisize driver, calculate KPI rates **Graphical** • Box plots of driver, one per KPI level, or • Discretisize driver, bar chart of KPI rates	**Tests** • Simple logistic regression (if $k = 2$) • Simple multinomial regression (for $k > 2$)
	Numerical	**Descriptives** • Average, std. dev. of KPI per driver level **Graphical** • Box plots of KPI, one per driver level	**Tests** • t-test (if $k = 2$) • ANOVA (for $k > 2$)	**Descriptives** • Pearson correlation of KPI and driver **Graphical** • Scatter plot of KPI and driver	**Tests** • Simple regression

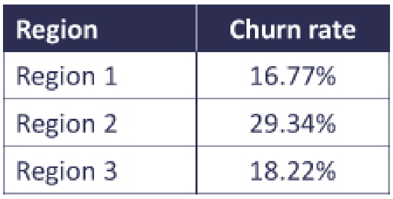
FIGURE 8.5 Investigating one-to-one relationships

Below, we discuss each of the four cells using data on 1,216 customers from a bank. We have two KPIs that we want to explore. The first is a churn variable, which is categorical as there are two values that this variable can attain. The second is the customer's balance, which is numerical. We have several possible drivers of these KPIs that we want to investigate. We turn to their analysis according to their measurement level below.

8.3.1 KPI categorical, driver categorical

As a first exploration of the data, we investigate whether the region where the customer lives affects the churn variable. As the region variable has three levels (Region 1, Region 2, and Region 3), we are in the upper-left quadrant of Figure 8.5.

Consequently, relevant descriptive statistics are the churn rate in each region, which can be found in Figure 8.6.



Region	Churn rate
Region 1	16.77%
Region 2	29.34%
Region 3	18.22%

FIGURE 8.6 Churn rate per region

In Figure 8.7, we illustrate these numbers graphically. The plot suggests that Region 2 has a higher churn rate than the other two regions, and the error bars suggest that this is a significant difference.

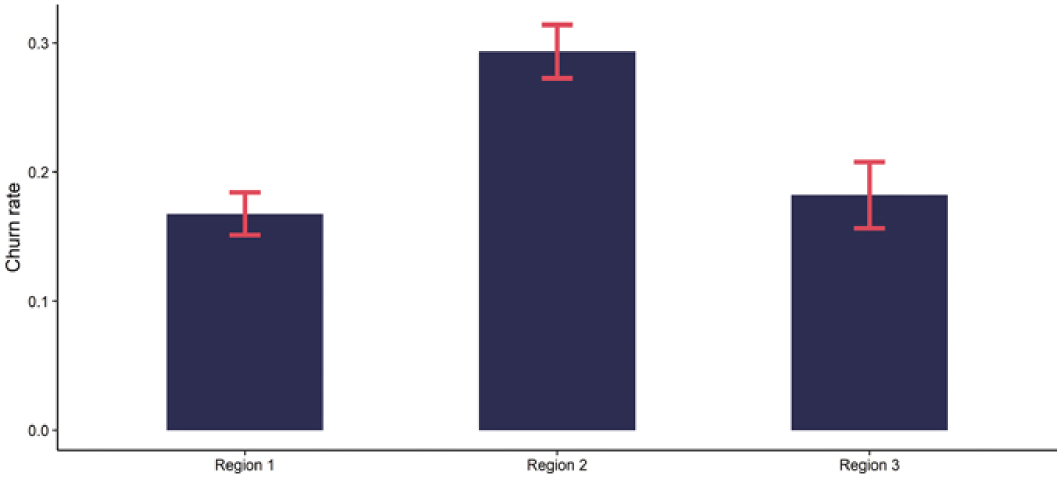
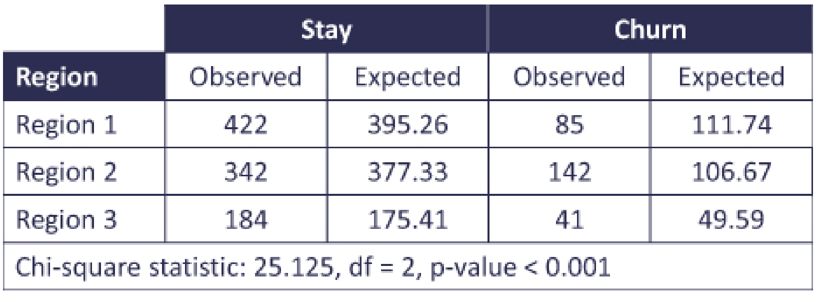


FIGURE 8.7 Bar chart of churn rates per region

The chi-square test of independence in Figure 8.8 confirms that the churn rates indeed differ significantly per region.

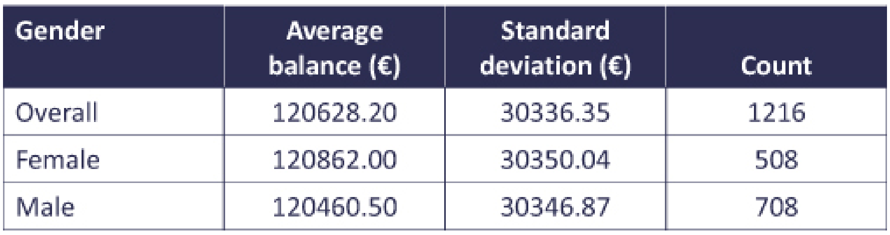


	Stay		Churn	
Region	Observed	Expected	Observed	Expected
Region 1	422	395.26	85	111.74
Region 2	342	377.33	142	106.67
Region 3	184	175.41	41	49.59
Chi-square statistic: 25.125, df = 2, p-value < 0.001				

FIGURE 8.8 Chi-square test for churn rate per region

8.3.2 KPI numerical, driver categorical

Turning to the investigation of a numerical KPI, we present the mean and the standard deviation of the balance variable in Figure 8.9 and break this down according to a different possible categorical driver than was used above: gender.



Gender	Average balance (€)	Standard deviation (€)	Count
Overall	120628.20	30336.35	1216
Female	120862.00	30350.04	508
Male	120460.50	30346.87	708

FIGURE 8.9 Descriptive statistics of the balance variable per gender category

Figure 8.10 depicts a box plot of the balance variable for each of the two levels of gender. In line with what the numbers in Figure 8.9 indicate, the red lines in the box plots show that the median balance of female customers is slightly higher than the median balance of male customers.

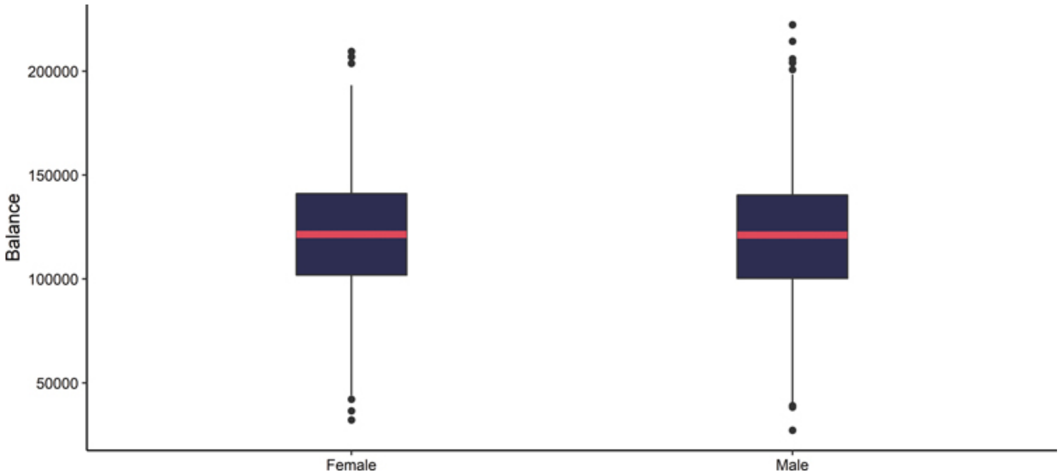
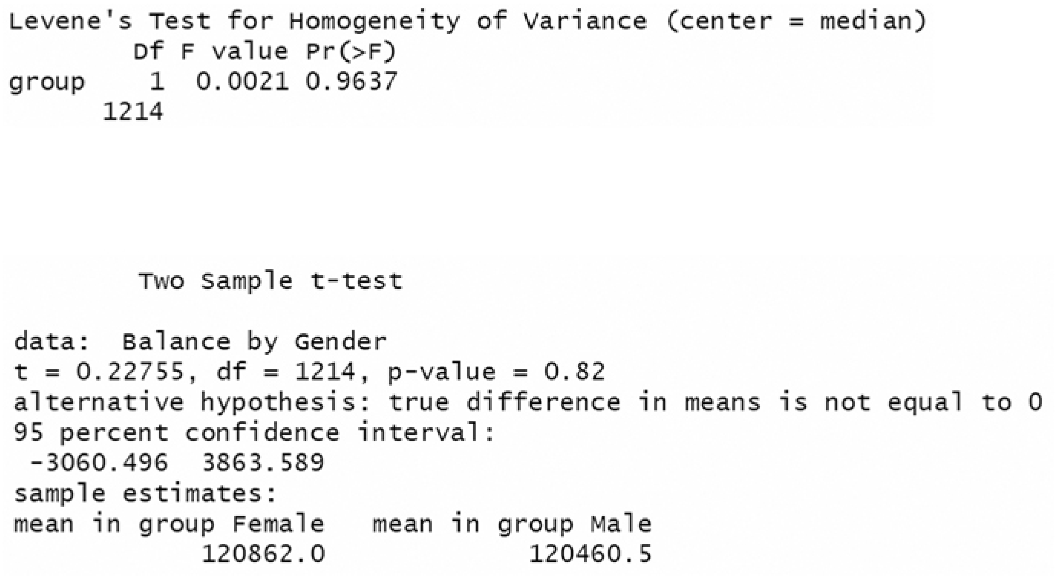


FIGURE 8.10 Box plots of the balance variable per gender category

In order to investigate whether this is a significant difference, we can conduct an independent-samples t-test, since our driver variable has only two levels. To this end, we first need to investigate whether the variance in the two groups is the same (see e.g., Malhotra, 2019). A Levene's test of homogeneity of variance does not reject the null hypothesis of equal variances, so we conduct the independent-samples t-test assuming that the variances are the same (see the R output in Figure 8.11).



```
Levene's Test for Homogeneity of Variance (center = median)
        Df F value Pr(>F)
group    1  0.0021 0.9637
      1214
```

```
        Two Sample t-test

data:  Balance by Gender
t = 0.22755, df = 1214, p-value = 0.82
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3060.496  3863.589
sample estimates:
mean in group Female   mean in group Male
            120862.0             120460.5
```

FIGURE 8.11 R output of the independent-samples t-test

Because the *p*-value is larger than 0.05 in the results of the independent-samples t-test (see Figure 8.11), we conclude that the balance does not differ significantly between male and female customers.

The analyses in Figure 8.11 depend critically on the assumption that the KPI is normally distributed. As all the data points in the Q-Q plot in Figure 8.12 are close to the straight line, normality seems to be a reasonable assumption for the balance variable (large deviations from the solid blue line in Figure 8.12 would be an indication for non-normality). This is confirmed by several normality tests in the R-script, which can be found in the accompanying online Appendix to this chapter.

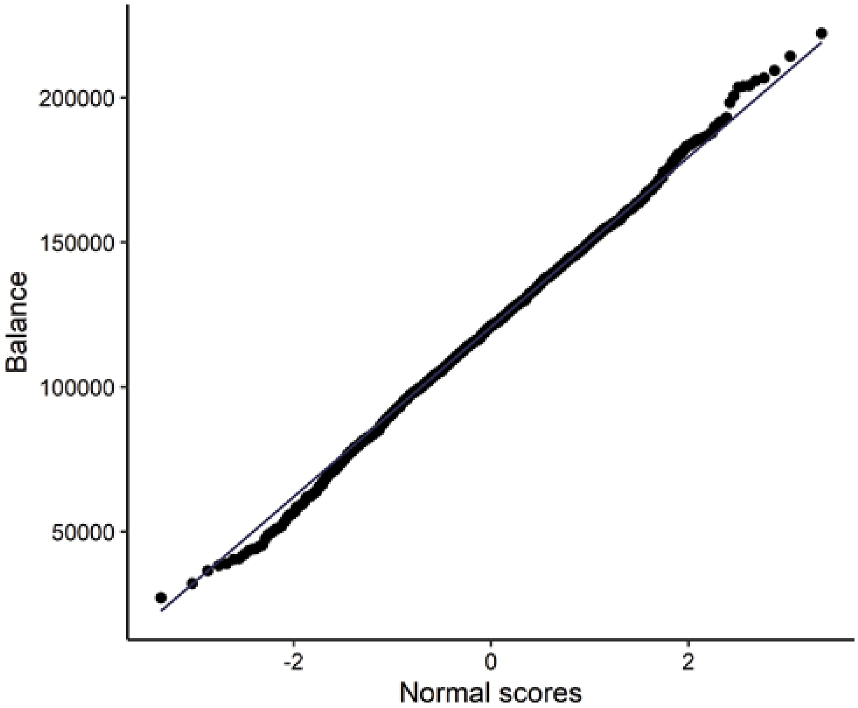


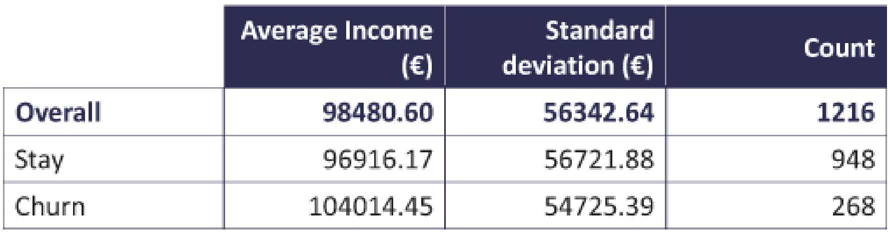
FIGURE 8.12 Q-Q plot of the balance variable

In case Figure 8.12 and the normality tests would have indicated that the normality assumption did not hold for the balance variable, the t-test would not have been appropriate. In that case, the Wilcoxon rank sum test could have been used since this is the non-parametric equivalent to an independent samples t-test and does not require normality of the variable that is being tested. Since the normality assumption appears not to be violated in this case, the outcome of the Wilcoxon rank sum test is very similar to the outcome of the independent samples t-test (see the R-script in the online Appendix to this chapter).

If we were interested in testing differences in balance for a categorical variable with more than two levels (e.g., the region variable), the t-test could not have been used either. Figure 8.5 indicates that an analysis of variance (ANOVA) would then have been appropriate. Similar to the t-test, ANOVA also relies on normality of the variable being tested. If a Q-Q plot or the normality tests indicate that this cannot be assumed, the Kruskal-Wallis test can be used as a non-parametric alternative. These analyses are not presented here but can be found in the R-script in the online Appendix to this chapter.

8.3.3 KPI categorical, driver numerical

Let us return to investigating churn, but let us now study the effect of income, a numerical variable. As a first descriptive analysis, we suggest calculating descriptive statistics for each KPI level to investigate how the distribution of the driver differs across the levels of the KPI. This is exemplified in Figure 8.13 where we present the mean and the standard deviation of income for churning customers and customers who decided to stay.



	Average Income (€)	Standard deviation (€)	Count
Overall	**98480.60**	**56342.64**	**1216**
Stay	96916.17	56721.88	948
Churn	104014.45	54725.39	268

FIGURE 8.13 Descriptive statistics of the income variable for churning and staying customers

The numbers in Figure 8.13 may be illustrated graphically by box plots of the driver variable for each KPI level (see Figure 8.14 which depicts two boxplots of the income variable, one for each level of churn).

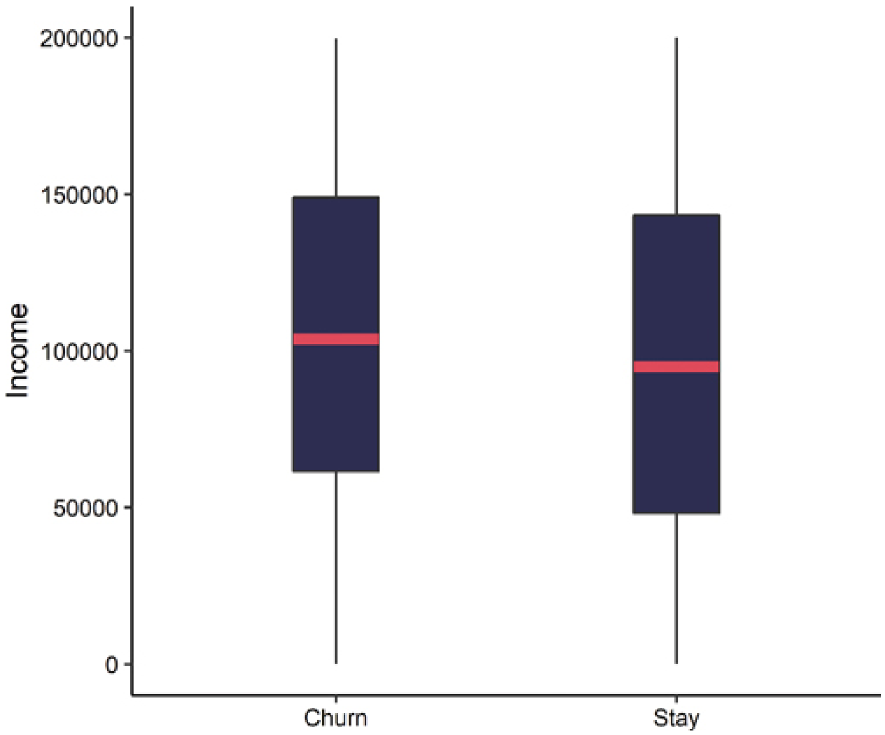


FIGURE 8.14 Box plots of income for each level of the churn variable

For some readers, Figure 8.14 may be confusing as it does not follow the convention of depicting the driver on the horizontal axis and the KPI on the vertical axis. To address this issue, horizontal box plots can be used. Alternatively, the driver variable can be discretized by dividing its values into a number of similar-sized groups, and rates can be determined within each of the groups. This is illustrated in Figure 8.15, which depicts churn rates for ten income groups where the groups are arranged from lower to higher incomes.

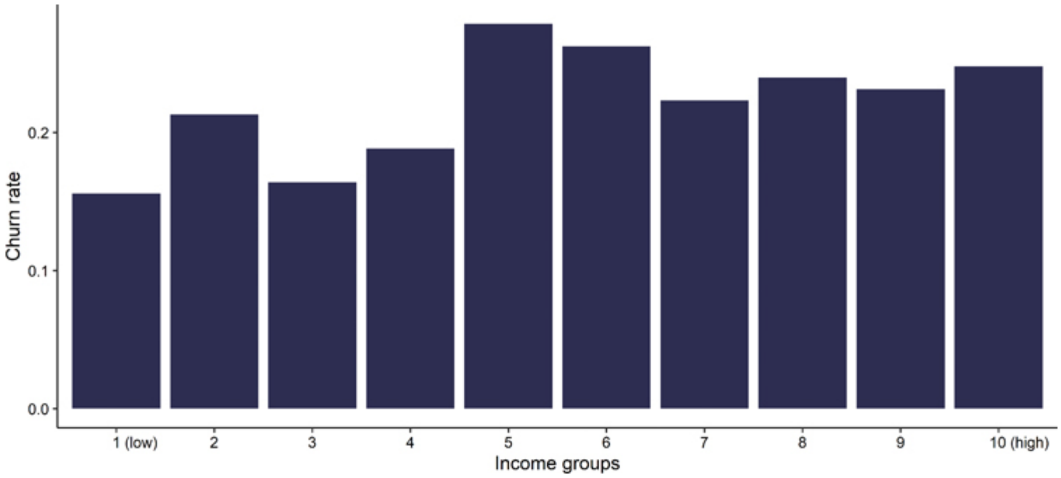
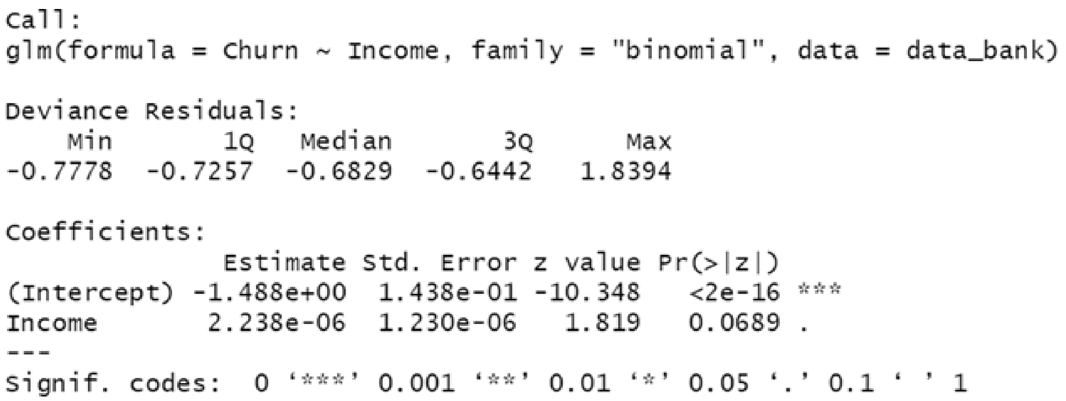


FIGURE 8.15 Bar chart of churn rate per income group

The bars in Figure 8.15 suggest that churn rates may be increasing with income. To investigate this relationship somewhat more formally, Figure 8.5 suggests conducting a simple logistic regression. We present the outcome of a logistic regression where churn is explained by the income variable in Figure 8.16 but defer a more detailed discussion of logistic regression to Section 9.4.



```
Call:
glm(formula = Churn ~ Income, family = "binomial", data = data_bank)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7778  -0.7257  -0.6829  -0.6442   1.8394

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.488e+00  1.438e-01 -10.348   <2e-16 ***
Income       2.238e-06  1.230e-06   1.819   0.0689 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

FIGURE 8.16 R output of a simple logistic regression of churn on income

The output of logistic regression in Figure 8.16 shows a positive coefficient for the income variable. This suggests a confirmation of the impression from Figures 8.14 and 8.15 that churn and income are positively related, but since the *p*-value of the coefficient exceeds 0.05, it fails to reach significance. As the *p*-value is close to 0.05 (and smaller than 0.1), this is sometimes referred to as income being marginally significant for churn.

8.3.4 KPI numerical, driver numerical

The final cell of Figure 8.5 provides guidance on what exploratory analyses to conduct when both the KPI and the potential driver are numerical variables. Figure 8.17 presents a first descriptive and graphical analysis of the relationship between the age of a customer and his or her balance. Both the correlation and the scatter plot suggest that this relationship is not very strong.

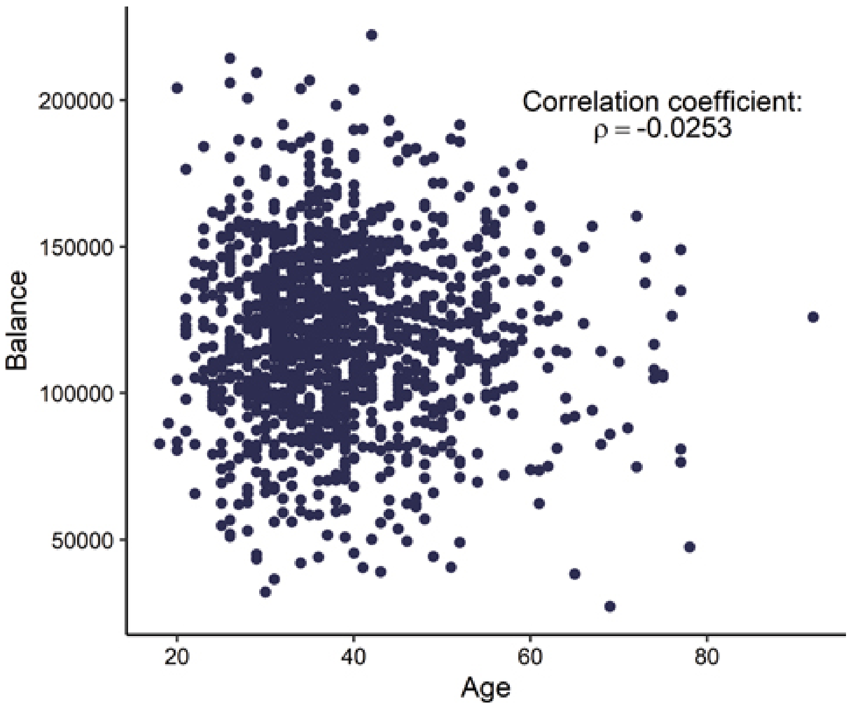
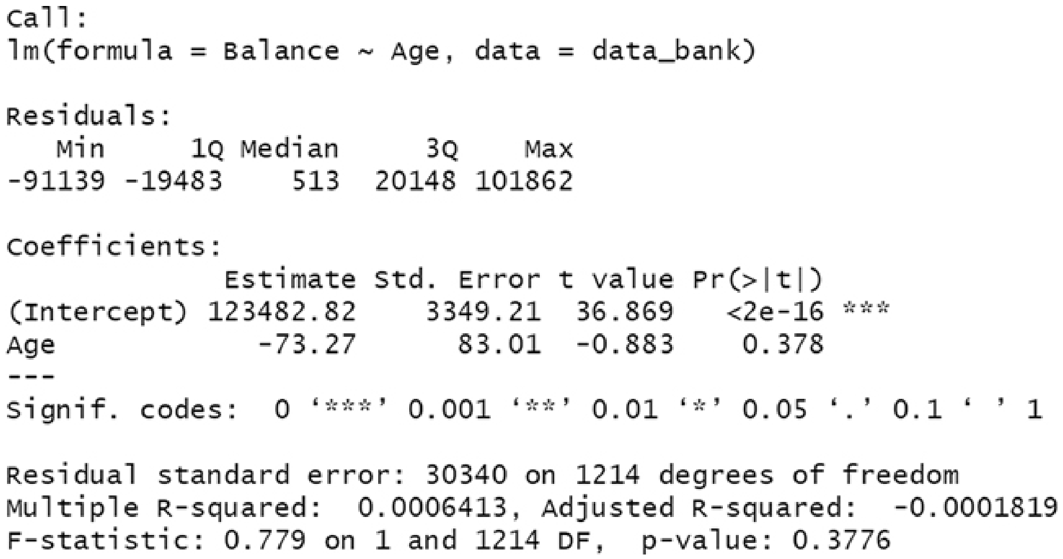


FIGURE 8.17 Correlation coefficient and scatter plot of balance and age

To test the significance of the relationship between a numerical KPI and a numerical potential driver, simple regression can be conducted. Figure 8.18 presents the R output of regressing balance on age in the bank example. We observe that the *p*-value for age exceeds 0.05. As we explain in further detail in Section 9.3, this indicates that the relationship between age and balance is not significant.



```
Call:
lm(formula = Balance ~ Age, data = data_bank)

Residuals:
   Min      1Q Median      3Q     Max 
-91139 -19483    513  20148 101862 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 123482.82    3349.21  36.869   <2e-16 ***
Age            -73.27      83.01  -0.883    0.378    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30340 on 1214 degrees of freedom
Multiple R-squared:  0.0006413, Adjusted R-squared:  -0.0001819 
F-statistic: 0.779 on 1 and 1214 DF,  p-value: 0.3776
```

FIGURE 8.18 R output of a simple regression of balance on age

8.3.4.1 General remarks

– Please note that the guidelines above are for exploratory analyses, where one driver is investigated at a time. In almost all situations, further analyses are warranted in which, typically, more than one driver is included. Such further analyses typically are not solely data-driven and usually involve a combination of theory and empirics and are discussed in Chapter 9.

– The analyses discussed in this section may provide guidance on which variables are significantly related to the KPI and should certainly be included in such further analyses.

– The reverse is not necessarily true, as the effect of the driver on the KPI might be small and concealed if the analysis does not control for the effect of other variables. Consequently, if there is a variable for which a relationship between a KPI and a driver cannot be established in an exploratory analysis, this should not be a reason to exclude such a driver from further analyses, especially if there are good theoretical reasons to include such a variable.

– Several analyses are special cases of the techniques that we discuss in this section. One important example is A/B testing, which is a popular (large-scale) experimentation method to test changes in a website's design, see Figure 8.19. For example, Booking.com could change specific wordings in their offers and observe how this improves conversion rates. A/B testing is nothing more than a randomized field experiment using a large number of visitors to the website. Typically, a new website is tested using a group of visitors and the results are compared with a control group in which the old version of the website is still current. One can extend this testing by using more complicated experimental designs, such as factorial designs. To analyze whether the experiment is fully randomized, ANOVA can be used to test for significant effects on outcomes. To account for potential customer specific effects, covariates such as age and loyalty can be included in ANCOVA. Note that, given the large scale of these experiments, not only should the significance of these effects be considered, but also the size of the found differences (see Chapter 7). Besides, in general A/B testing is not different from testing of other marketing tactics used for years, such as testing of different versions of direct mailings (e.g., David Sheppard Associates, 1999). The only difference is probably that A/B testing usually involves very large samples.

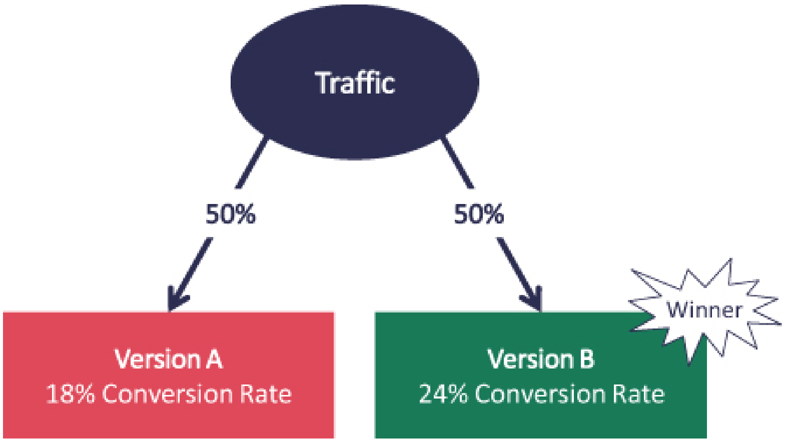


FIGURE 8.19 A/B testing

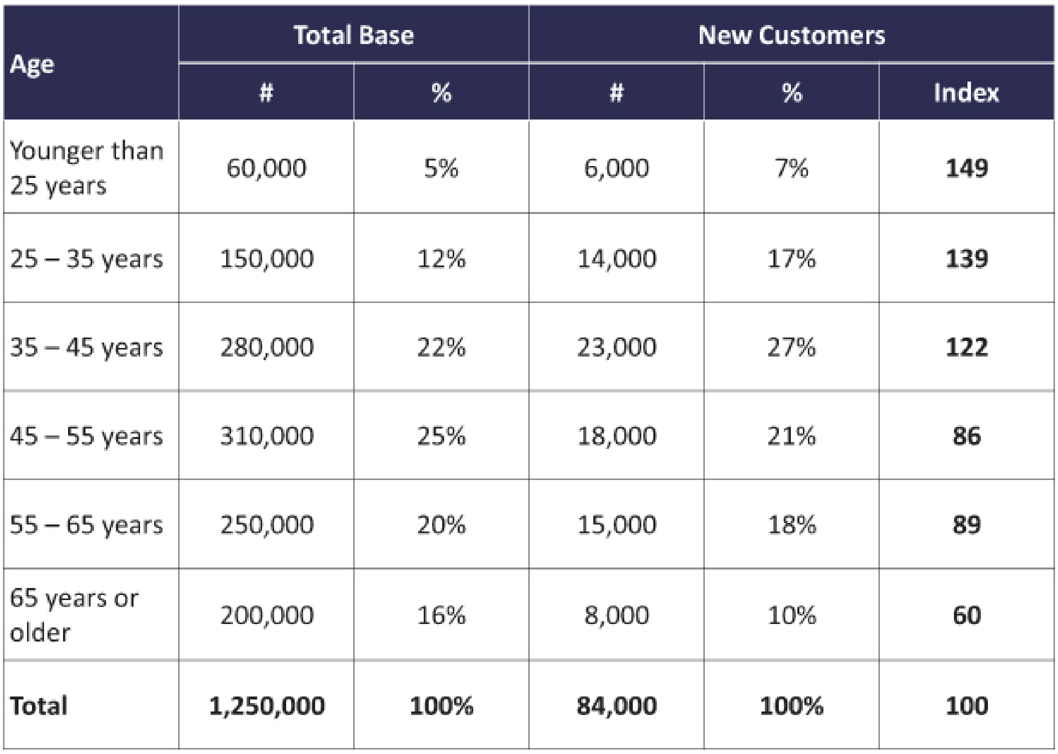
8.4 SPECIAL CASES OF ONE-TO-ONE EXPLORATORY ANALYSES

In this section, we discuss several specific examples of the analyses mentioned in the previous section.

8.4.1 Profiling and customer crossings

With profiling, analysts aim to provide insights into differences between brands or customer segments. For example, researchers may want to compare Apple users to Android users (discrete KPI) across levels of a discrete driver variable, such as education. Similarly, numerical KPIs, such as customer lifetime value (CLV) can be compared across different customer segments. These analyses may help firms to improve their understanding of the attractiveness of subgroups and segments, which in turn may be useful for tailoring propositions. In customer management, it can be useful to identify look-alikes for acquisition campaigns. To achieve this, so-called crossings should be made, where it is investigated whether brands or customer segments differ on important KPIs or customer characteristics. Figure 8.5 can be used to select appropriate techniques for such investigations, depending on the measurement scale of the KPI or the customer characteristic. The driver is then the categorical variable that indicates the brand or the segment to which a customer belongs.

One important issue is how to present the outcomes of customer crossings in a way that allows readers to quickly understand differences between customer groups. One useful way is to index the value for each group, where the index is the value of the descriptive (e.g., average) within a group divided by the average for the whole customer base multiplied by one hundred. In Figure 8.20 we show this indexation for profiling of new customers on age classification. As a kind of rule of thumb, analysts typically consider an index larger than 110 or smaller than 90 as "substantial." However, one should be careful with small cell sizes as in these instances the differences can be very large in terms of indices, while they are not actually significant. This can be tested with either a t-test (when there are two segments) or an ANOVA (when there are more than two segments), see Figure 8.5.

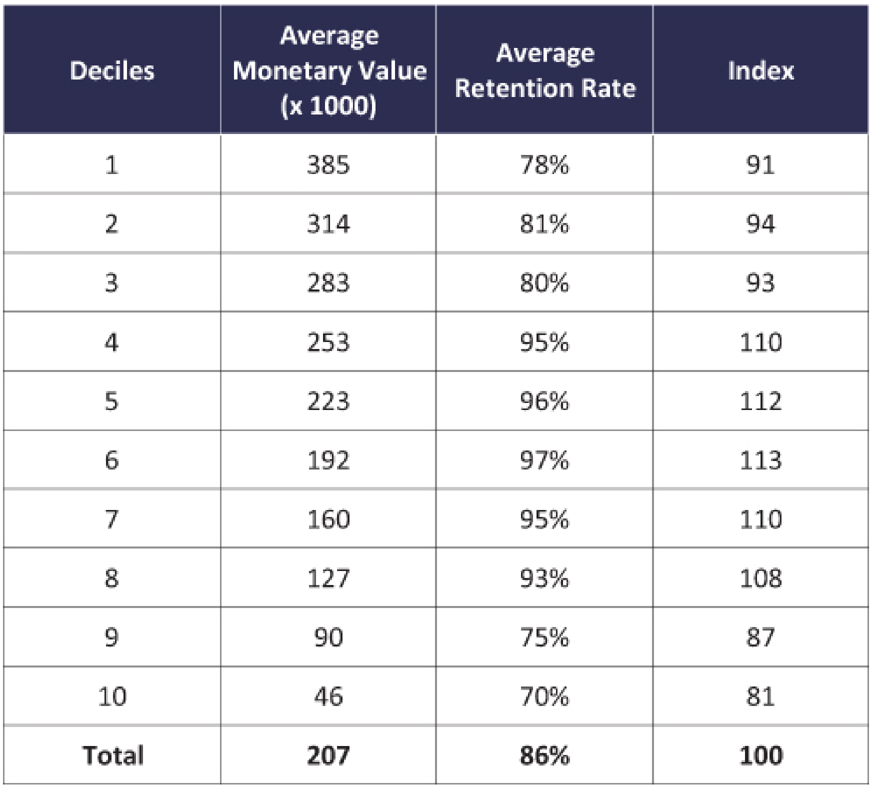


Age	Total Base		New Customers		
	#	%	#	%	Index
Younger than 25 years	60,000	5%	6,000	7%	**149**
25 – 35 years	150,000	12%	14,000	17%	**139**
35 – 45 years	280,000	22%	23,000	27%	**122**
45 – 55 years	310,000	25%	18,000	21%	**86**
55 – 65 years	250,000	20%	15,000	18%	**89**
65 years or older	200,000	16%	8,000	10%	**60**
Total	**1,250,000**	**100%**	**84,000**	**100%**	**100**

FIGURE 8.20 Profiling new customers on age classification

8.4.2 Decile analysis

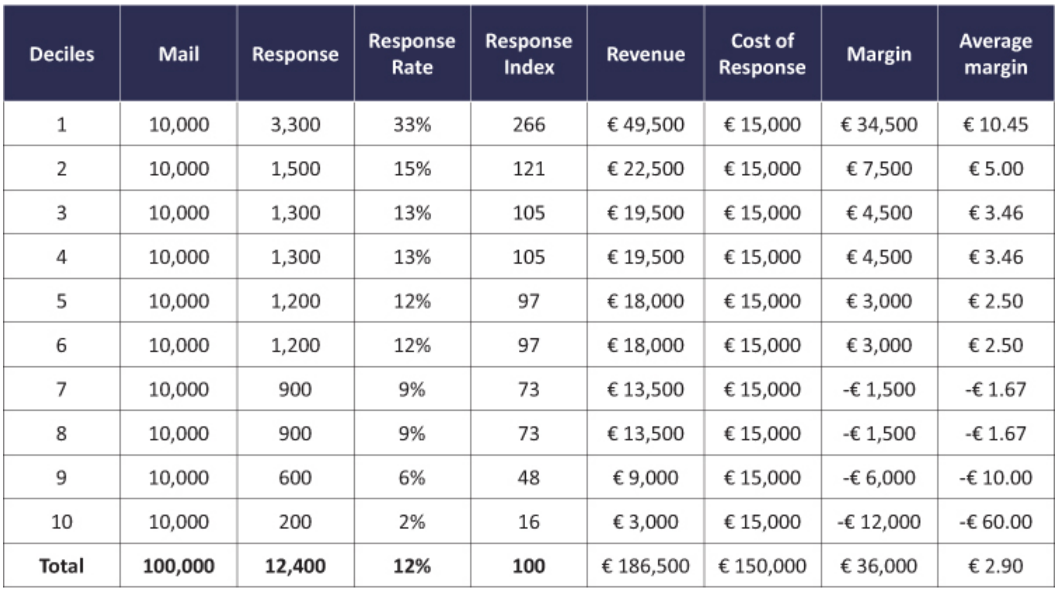
A frequently used technique to investigate differences between customer groups is decile analysis. With this method, customers are divided into ten equal-sized groups (deciles) according to their value on a numerical KPI or customer characteristic, such that each group consists of 10% of the customer database. To this end, the customers are ordered based on their value on the KPI, and those with the lowest score are classified as members of decile 10, whereas the 10% with the highest value are classified as members of decile 1. In a subsequent analysis, decile membership can be used as a categorical driver to describe other KPIs or customer characteristics such as customer revenues, margins, or responses to marketing activities. For example, for each customer profitability decile the average age can be calculated, or the average retention rate (see Figure 8.21). We carried out a decile analysis for a book club, based on the monetary value of each customer. In a subsequent analysis, we calculated the average retention rate in test mailing. As can be observed from Figure 8.21, the average retention rate is highest in decile 6, which does not have the highest monetary value. Unfortunately, the retention rate is relatively low for the most valuable segment.



Deciles	Average Monetary Value (x 1000)	Average Retention Rate	Index
1	385	78%	91
2	314	81%	94
3	283	80%	93
4	253	95%	110
5	223	96%	112
6	192	97%	113
7	160	95%	110
8	127	93%	108
9	90	75%	87
10	46	70%	81
Total	**207**	**86%**	**100**

FIGURE 8.21 Decile analysis for monetary value and retention rates

A decile analysis is often referred to as a gains chart analysis for customer response analysis. This gains chart is very similar to decile analysis. The deciles are ranked based on their response probability. Some subsequent calculations can be done to calculate the margin per decile and the deciles with a positive margin can be selected for targeting (see Figure 8.22).



Deciles	Mail	Response	Response Rate	Response Index	Revenue	Cost of Response	Margin	Average margin
1	10,000	3,300	33%	266	€ 49,500	€ 15,000	€ 34,500	€ 10.45
2	10,000	1,500	15%	121	€ 22,500	€ 15,000	€ 7,500	€ 5.00
3	10,000	1,300	13%	105	€ 19,500	€ 15,000	€ 4,500	€ 3.46
4	10,000	1,300	13%	105	€ 19,500	€ 15,000	€ 4,500	€ 3.46
5	10,000	1,200	12%	97	€ 18,000	€ 15,000	€ 3,000	€ 2.50
6	10,000	1,200	12%	97	€ 18,000	€ 15,000	€ 3,000	€ 2.50
7	10,000	900	9%	73	€ 13,500	€ 15,000	-€ 1,500	-€ 1.67
8	10,000	900	9%	73	€ 13,500	€ 15,000	-€ 1,500	-€ 1.67
9	10,000	600	6%	48	€ 9,000	€ 15,000	-€ 6,000	-€ 10.00
10	10,000	200	2%	16	€ 3,000	€ 15,000	-€ 12,000	-€ 60.00
Total	**100,000**	**12,400**	**12%**	**100**	€ 186,500	€ 150,000	€ 36,000	€ 2.90

FIGURE 8.22 Gain chart analysis for book club

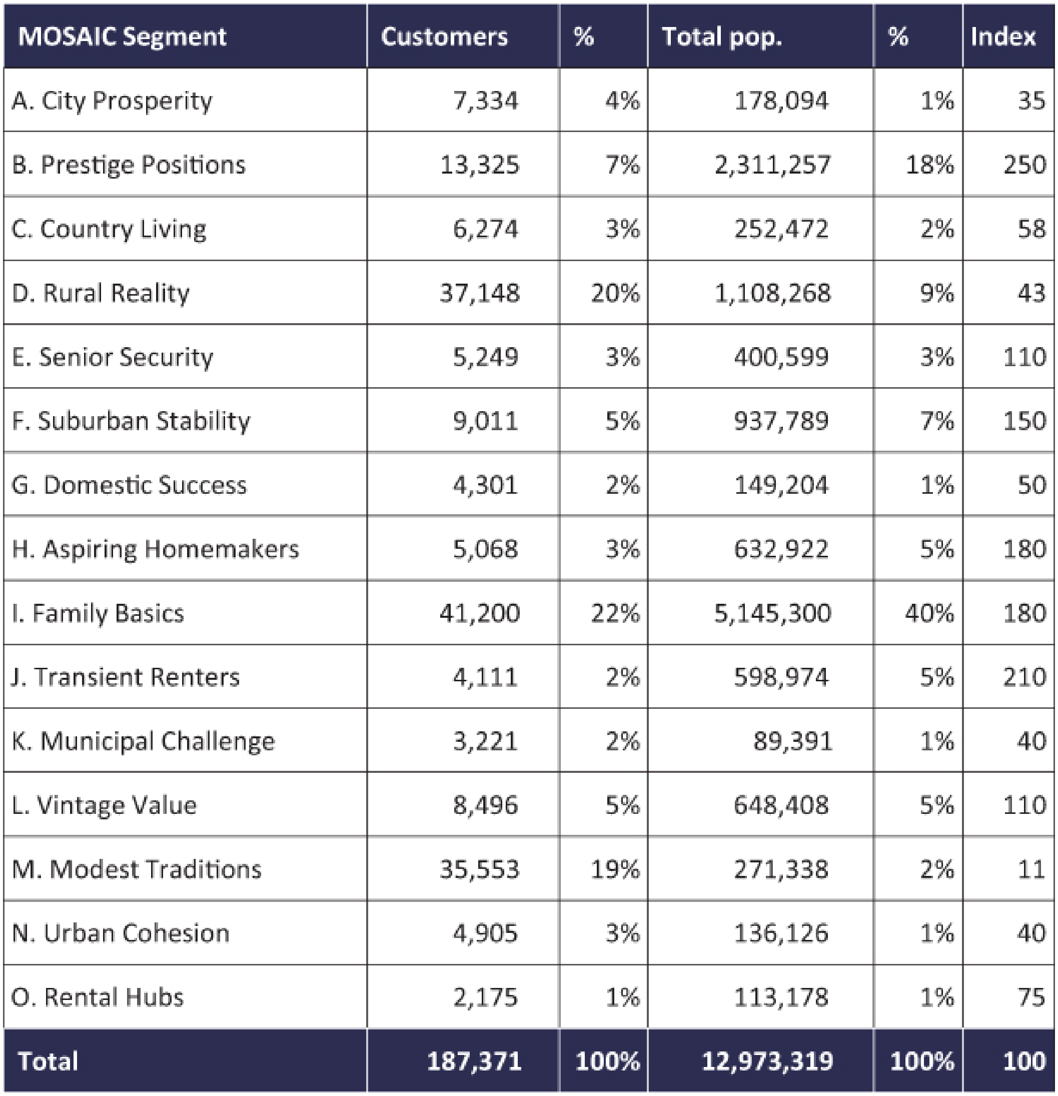
8.4.3 External profiling

So far, we have mainly discussed analyses where we compare customer segments with other segments. One could label this an internal customer profiling analysis. However, managers are also interested in the profile of their customers (segments) in comparison with the rest of the market. For this purpose, firms use external profiling analyses. In this analysis, characteristics of a firm's customers (groups) are compared with customer characteristics in the market or population. Again, indexing is frequently used to make these comparisons. Using these indices, one could for example say that customers of a private bank are two times as wealthy as the average bank customer.

8.4.4 Zip code analysis

One specific external profiling analysis is the incorporation of zip codes. External data providers, such as Acxiom and Experian, offer specific zip-code level information (see Chapter 4). With this information firms can gain an understanding about which zip codes and thus local/regional areas, are over- or under-represented in their customer base. Along with this zip code information, these external data suppliers have also developed specific segments, such as "rural families" and "single households." For example, an online retailer may find that rural families are over-represented in their customer base, while the single household is almost not represented at all. Again, indexing can be very useful. To calculate indices, the frequency percentage of zip code segments within the customer base/group are divided by the frequency percentage of these zip code segments within the population.

In Figure 8.23 we present an example of a clothing retailer. This company aims to compare its clientele with the population. As can be observed from the analysis, the Prestige Positions, Aspiring Homemakers, Family Basics, and Transient Renters are overrepresented in the customer base. This can be more formally investigated with a Chi-square test.



MOSAIC Segment	Customers	%	Total pop.	%	Index
A. City Prosperity	7,334	4%	178,094	1%	35
B. Prestige Positions	13,325	7%	2,311,257	18%	250
C. Country Living	6,274	3%	252,472	2%	58
D. Rural Reality	37,148	20%	1,108,268	9%	43
E. Senior Security	5,249	3%	400,599	3%	110
F. Suburban Stability	9,011	5%	937,789	7%	150
G. Domestic Success	4,301	2%	149,204	1%	50
H. Aspiring Homemakers	5,068	3%	632,922	5%	180
I. Family Basics	41,200	22%	5,145,300	40%	180
J. Transient Renters	4,111	2%	598,974	5%	210
K. Municipal Challenge	3,221	2%	89,391	1%	40
L. Vintage Value	8,496	5%	648,408	5%	110
M. Modest Traditions	35,553	19%	271,338	2%	11
N. Urban Cohesion	4,905	3%	136,126	1%	40
O. Rental Hubs	2,175	1%	113,178	1%	75
Total	187,371	100%	12,973,319	100%	100

FIGURE 8.23 External profiling for a clothing retailer using zip code segmentation

8.4.4.1 Some practical guidelines

There are some specific issues an analyst can encounter when running a profiling analysis:

- The number of variables can be large. Profiling then becomes rather difficult as the number of possible comparisons on specific variables becomes too large. We therefore strongly recommend either focusing on a limited number of pre-selected variables or reducing the number of profiling variables by using principal components analysis (PCA), as explained in Subsection 8.6.2 later in this chapter.
- One should be careful with over-interpreting differences between groups. One common mistake is that the analysis reveals that a firm is overrepresented in a specific lifestyle segment. Firms may then only want to target this segment. However, this segment can be rather small. So, one should look beyond the profiles, but also consider factors such as segment size.

- Differences in profile analyses can become easily significant, especially when analyzing large data. As already discussed in Chapter 7, one should focus not only on significance but also on the size of the found differences.
- Profiling is a univariate technique that is suited for analyzing two variables. In essence, these are just associations and one should be very careful in interpreting these associations as there can be spurious correlations. Causal inferences cannot be made!
- Crossings with continuous variables are not very insightful given the large number of possible cells. One way to overcome this is to create subgroups in the continuous variable. To this end, decile analysis can be used, but the analysis can also be conducted with fewer subgroups (e.g., quartiles). This facilitates an improved understanding of how a variable is associated with other variables.

8.5 DYNAMIC ANALYSES

As noted in Section 8.2, firms are frequently interested in how a market develops or how brand sales will grow. For them, this is important so that they can constantly monitor market attractiveness, and for planning purposes they aim to know what level of brand sales they can realistically expect. For this purpose, firms might be interested in identifying trends in relevant KPIs. Is the churn rate going down? Are customers becoming more satisfied? Does our brand image improve? Does the market grow? Investigating dynamics in KPIs can be considered as an important subset of the analyses that are represented in Figure 8.5, where the drivers are variables that are related to time. These can be categorical (e.g., years), but also numerical (e.g., number of days). Given the importance of dynamics in many analytical projects, we investigate dynamic analyses in this section.

8.5.1 Trend analysis

Trend analysis constitutes an important example of dynamic analysis, in which the development of a numerical KPI is analyzed over time. For example, for a new product, the sales figures might rise with the years. In line with the recommendation in Figure 8.5, the first step in trend analysis is to plot the KPI against the driver in a line plot (see Figure 8.24 for an example).

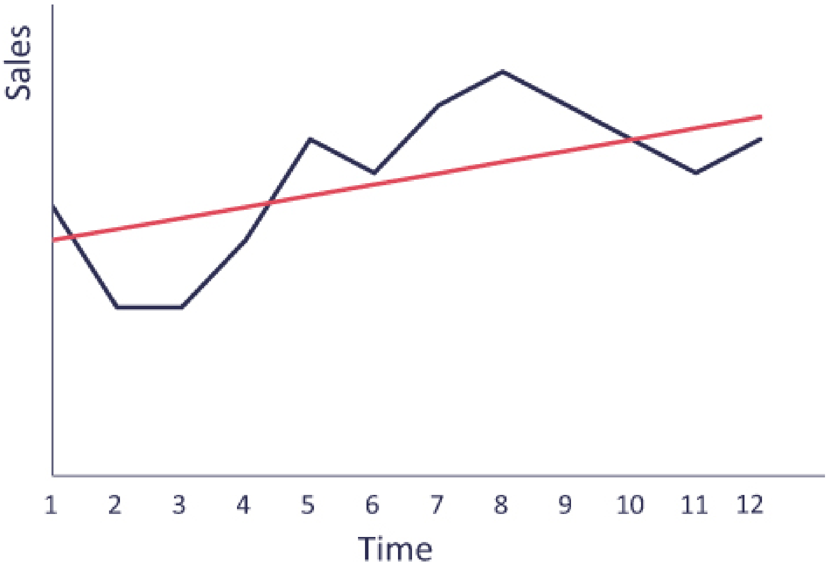
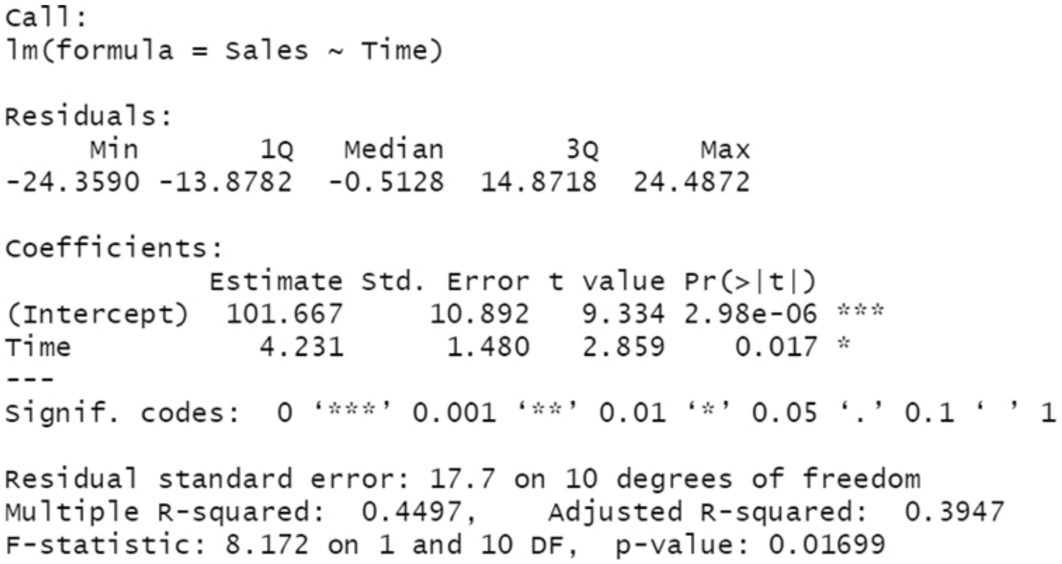


FIGURE 8.24 Trend in sales

This plotting immediately provides the analyst with information on the presence (or absence) of a trend in the data. It may also hint at specific trends. If it is a linear trend, there might be a straight-line development in sales. However, if it is nonlinear, sales might for example grow exponentially over time. As Figure 8.5 suggests, the presence of a trend can be identified using regression analysis.

Usually, an analyst will be mainly interested in the coefficient of the trend variable and its significance level. A positive and significant coefficient implies a positive linear trend, whereas a negative and significant coefficient implies a negative linear trend. In the R output that corresponds to Figure 8.24, we see in Figure 8.25 that the trend coefficient is indeed positive and significant, because the *p*-value associated with the time variable is smaller than 0.05. From the estimated coefficient for time we conclude that sales are expected to increase each month with 4.2 units.



```
Call:
lm(formula = Sales ~ Time)

Residuals:
     Min       1Q   Median       3Q      Max 
-24.3590 -13.8782  -0.5128  14.8718  24.4872 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  101.667     10.892   9.334 2.98e-06 ***
Time           4.231      1.480   2.859    0.017 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.7 on 10 degrees of freedom
Multiple R-squared:  0.4497,	Adjusted R-squared:  0.3947 
F-statistic: 8.172 on 1 and 10 DF,  p-value: 0.01699
```

FIGURE 8.25 R output of a trend analysis of sales

As noted, the trend can be non-linear, something which can be deduced from a line plot of the KPI over time. To account for this, the model might include a quadratic or exponential trend if the sales grow more strongly over time, or a square root or log trend if the sales grow less strongly over time. We strongly recommend analysts to consider different trend operationalizations.

An important issue is that trends should not be confused with seasonality patterns. With seasonal patterns, sales follow a specific shape during the year, because demand depends on specific developments during the year (e.g., beer sales might go up during summer). To account for these effects, seasonal dummy variables (variables set at value 1 for a specific season) should be included in the models. These variables pick up the specific seasonal effect. For a more comprehensive treatment of seasonality and time series analysis we refer to Pauwels (2017).

8.5.2 Migration analysis

Migration analysis can be used to investigate the development of customers and brand or product usage over time. This migration analysis is frequently required to understand changes in aggregate sales figures over time. Changes in the number of transactions, the number of customers, turnover etc., are often reported, especially in financial reports, but many of the changes are not major and this tends to suggest that everything is stable. However, there are multiple behaviors hidden under the surface of these aggregate figures. Important value drivers may change: new customers are being acquired, customers may churn, there may be cross-selling and down-selling, and product and brand switches may occur (e.g. Verhoef, Van Doorn, & Dorotić, 2007). To capture value development over time, analysts need to gain insights into how customer status (e.g., churn) and behavior are changing over time. Although, on a year-by-year basis, changes may be limited, over a longer time period structural changes in the underlying value drivers (e.g., lower acquisition rates, continued down-sell), can have dramatic effects. In Figure 8.26 we give the example of a telecom provider.

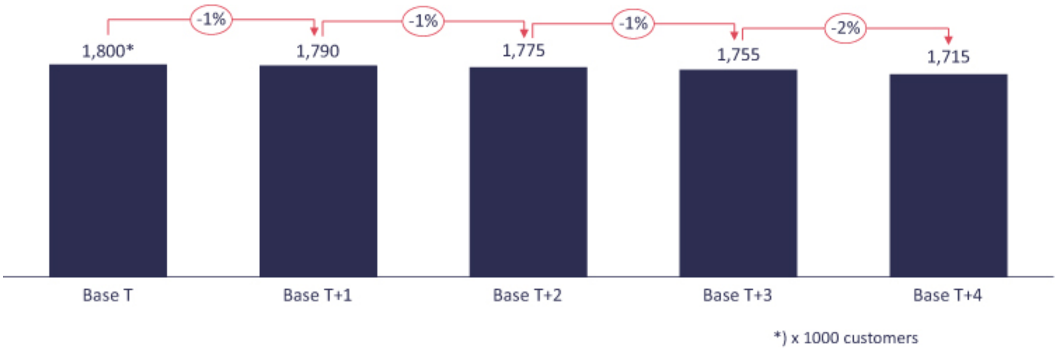


FIGURE 8.26 Development of the subscription base of a telecom provider

On a yearly basis, there are changes but they are not very dramatic. However, over five years, 5% of the customers are lost. The question is: what happened? A first analysis of the underlying value drivers, in which the base is decomposed into churners and newly acquired customers, shows that the number of acquired customers is not sufficient in size to overcome the loss of customers due to churn (see Figure 8.27).

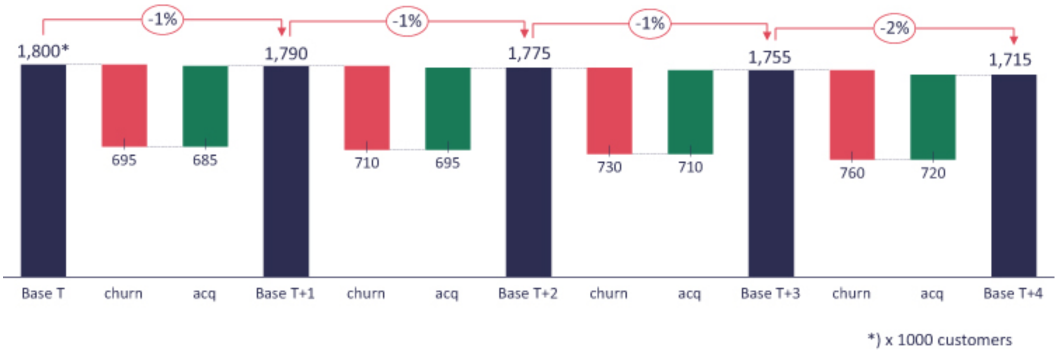
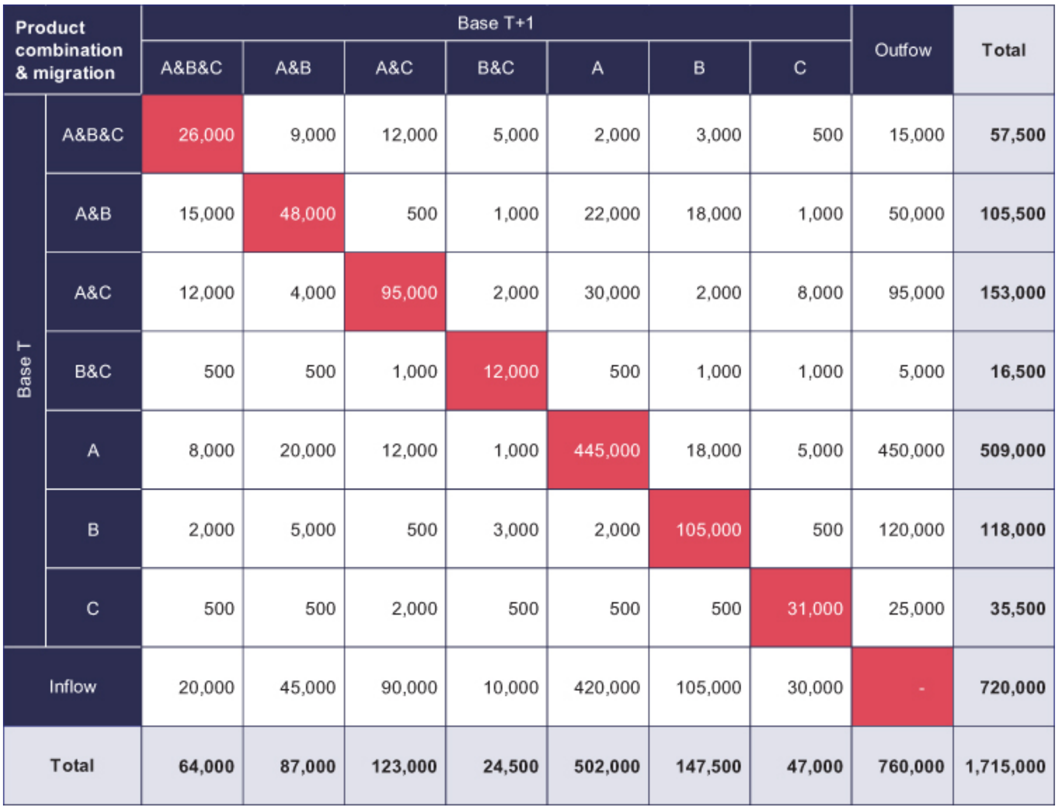


FIGURE 8.27 Decomposing subscription base in acquisition and churn

8.5.2.1 Migration matrix

A classical way to show migrations within the customer base is to use a migration matrix. These migration matrices show how customers change from one period (t) to the next period (t+1). In Figure 8.28, we show how the customers at time t, purchasing different services from a telecom firm, move in terms of their purchase behavior. For example, of the 72,500 customers purchasing products A, B, and C, 15,000 churned, while 26,000 customers kept the same product bundle. However, some customers also have a down sell. In this figure, we could also have chosen to show percentage values.



Product combination & migration		Base T+1							Outfow	Total
		A&B&C	A&B	A&C	B&C	A	B	C		
Base T	A&B&C	26,000	9,000	12,000	5,000	2,000	3,000	500	15,000	**57,500**
	A&B	15,000	48,000	500	1,000	22,000	18,000	1,000	50,000	**105,500**
	A&C	12,000	4,000	95,000	2,000	30,000	2,000	8,000	95,000	**153,000**
	B&C	500	500	1,000	12,000	500	1,000	1,000	5,000	**16,500**
	A	8,000	20,000	12,000	1,000	445,000	18,000	5,000	450,000	**509,000**
	B	2,000	5,000	500	3,000	2,000	105,000	500	120,000	**118,000**
	C	500	500	2,000	500	500	500	31,000	25,000	**35,500**
Inflow		20,000	45,000	90,000	10,000	420,000	105,000	30,000	-	**720,000**
Total		**64,000**	**87,000**	**123,000**	**24,500**	**502,000**	**147,500**	**47,000**	**760,000**	**1,715,000**

FIGURE 8.28 Migration matrix of customers of a telecom provider

Although the migration matrix has its advantages, it also has two disadvantages that makes it hard for managers to understand:

1. A large number of combinations in these migration matrices means that the matrix is difficult to interpret. Although Figure 8.28 seems insightful, it already has 64 combinations (8 x 8). If in these tables the percentages are also added, the number of figures to be processed becomes enormous.
2. Using a migration matrix only one KPI can be shown. For example, in Figure 8.28 we only show the number of customers in a group. However, one might also like to know the revenues and the CLV.

8.5.3 Like-4-like analysis

To overcome these issues, the so-called L4L analysis is frequently used, particularly in retailing, where management aims to understand how net turnover develops over time, while accounting for certain factors, such as opening a new store. Within customer management a L4L-analysis aims to insightfully show the customer flow accounting for different value drivers, such as cross-selling, up-selling, down-selling, etc. Further, it can combine volume and value KPIs. In Figure 8.29 we show an example of the value development decomposed by value driver.

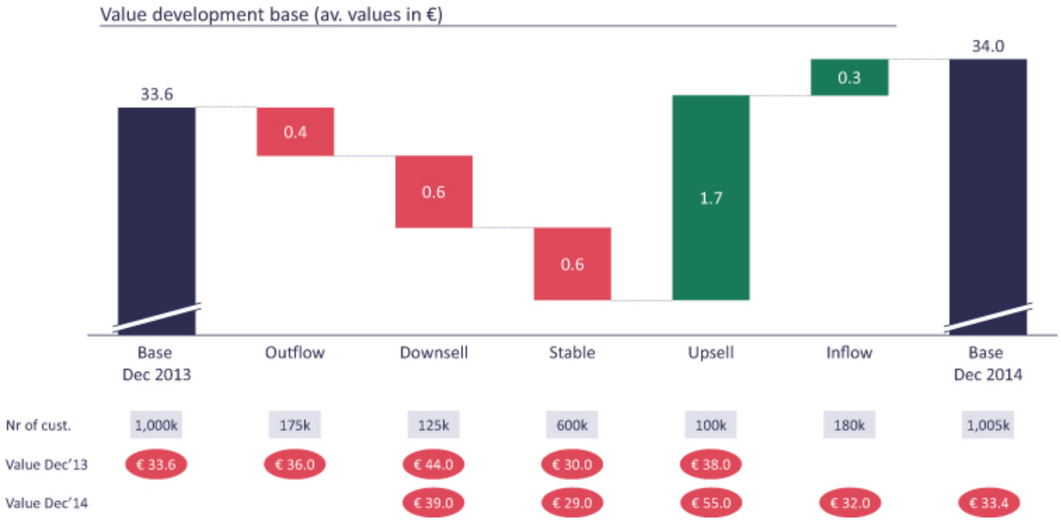
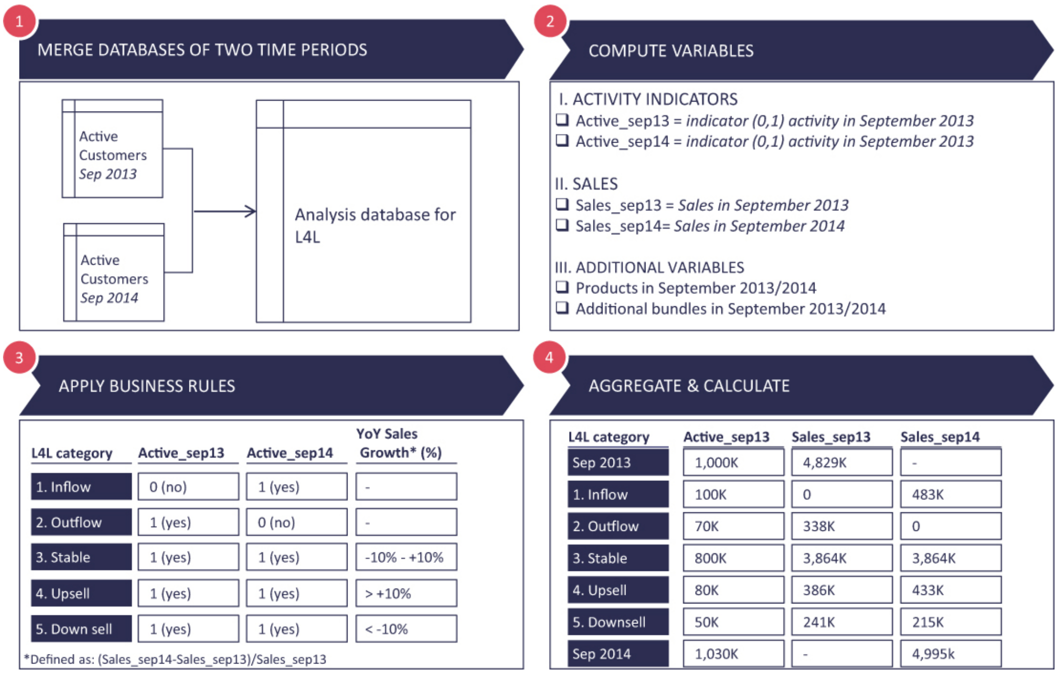


FIGURE 8.29 Like-4-like analysis for value development of the customer base of a phone operator

As can be observed from Figure 8.29, considerable value is created with up-selling customers, whereas value is lost through, for example, the outflow of customers (churn) and down-selling. Note that inflow also has a negative value effect on the customer base, probably because acquisition costs are rather high. Creating a L4L analysis seems rather easy, but it is more difficult than is frequently considered initially. We consider the following steps to create a L4L-analysis for period t and t+1:

1. Calculate the number of customers at period t
2. Determine the total value of customers at period t and period t+1 per value-driver. For outflow value is zero
3. at t+1 and for inflow value is zero at t
4. Determine the differences between the two time periods t and t+1
5. Determine the relative contribution in % of what each group has on the total value of the base
6. Determine the weighted value impact by multiplying the percentage by the value difference.

In Figure 8.30 we display the steps on how to carry out an L4L analysis, which clearly shows the implementation of the five steps discussed above.



1 MERGE DATABASES OF TWO TIME PERIODS

Active Customers *Sep 2013*

Active Customers *Sep 2014*

Analysis database for L4L

2 COMPUTE VARIABLES

I. ACTIVITY INDICATORS
❑ Active_sep13 = *indicator (0,1) activity in September 2013*
❑ Active_sep14 = *indicator (0,1) activity in September 2013*

II. SALES
❑ Sales_sep13 = *Sales in September 2013*
❑ Sales_sep14= *Sales in September 2014*

III. ADDITIONAL VARIABLES
❑ Products in September 2013/2014
❑ Additional bundles in September 2013/2014

3 APPLY BUSINESS RULES

L4L category	Active_sep13	Active_sep14	YoY Sales Growth* (%)
1. Inflow	0 (no)	1 (yes)	-
2. Outflow	1 (yes)	0 (no)	-
3. Stable	1 (yes)	1 (yes)	-10% - +10%
4. Upsell	1 (yes)	1 (yes)	> +10%
5. Down sell	1 (yes)	1 (yes)	< -10%

*Defined as: (Sales_sep14-Sales_sep13)/Sales_sep13

4 AGGREGATE & CALCULATE

L4L category	Active_sep13	Sales_sep13	Sales_sep14
Sep 2013	1,000K	4,829K	-
1. Inflow	100K	0	483K
2. Outflow	70K	338K	0
3. Stable	800K	3,864K	3,864K
4. Upsell	80K	386K	433K
5. Downsell	50K	241K	215K
Sep 2014	1,030K	-	4,995k

FIGURE 8.30 Steps for executing a L4L-analysis

The L4L-analysis provides insights into the changes of the value drivers over time. A next step is to execute more in-depth analyses, where an in-depth analysis is conducted per value driver. One example is the so-called cohort analysis, in which acquired customers (inflow) are followed over time. In Figure 8.31, we show the results of such a cohort-analysis, revealing that, for this firm especially, customers acquired in June develop less well over time.

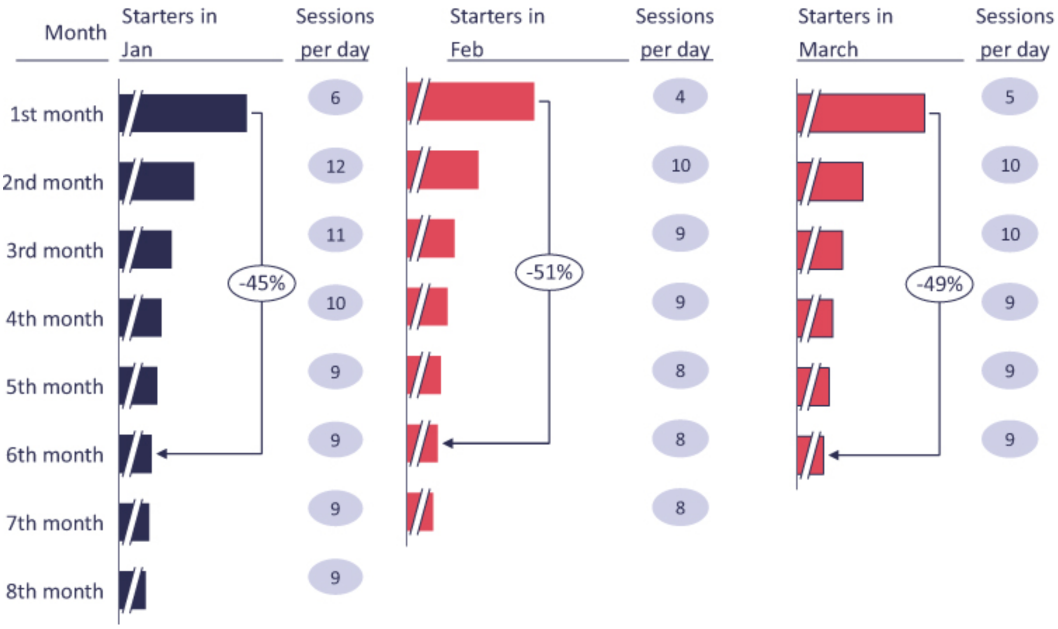


FIGURE 8.31 Example of cohort analysis

8.6 IDENTIFYING STRUCTURE IN THE DATA—UNSUPERVISED LEARNING

In the foregoing, we discussed reporting (Section 8.2) and investigating a KPI with respect to underlying variables (Sections 8.3–8.5). The techniques in the latter sections can be classified as dependence technique: all techniques that were discussed focus on the question of how the KPI depends on a driver variable (see also Figure 8.5). In this section, we focus on so-called interdependence techniques, where we are not interested in explaining one variable but in identifying patterns or structures in the data. We distinguish between two cases.

First, we focus on identifying groups of customers who are similar to each other, but dissimilar to customers in other groups. This aims to identify a structure in the data without trying to explain or predict a KPI. Once such a structure has been identified, the analyses that we discussed in Sections 8.3–8.5 can be used to further profile such segments. The technique to identify such groups of customers (or segments) is cluster analysis and this will be discussed in Section 8.6.1.

Second, we focus on identifying similarities or overlap between customer characteristics. In many cases, similar characteristics of a customer are measured using different variables. For example, satisfaction, trust, and commitment may all reflect a customer's perception of the quality of their relationship with the firm. Identifying such overlap does not aim to explain or predict a relevant KPI, but it can help in, for example, reducing the length of a satisfaction survey by eliminating redundant questions, which are already represented by other variables. The technique to identify the overlap between variables is principal components analysis and is discussed in Section 8.6.2.

In computer science terminology, both cluster analysis and factor analysis belong to the class of unsupervised machine learning techniques. That is because in both cases, there is no dependent variable that needs to be explained or predicted. As explained above, the focus here is on identifying patterns or structure in the data among all variables in the analyses.

8.6.1 Cluster analysis

Clustering techniques allow researchers to segment on multiple variables and to base the segmentation on statistical criteria, such as statistics on model fit (Wedel & Kamakura, 2000). Researchers have several methods at their disposal to execute a segmentation analysis. Methods available in statistical software include K-Means, two-step cluster analysis, and hierarchical cluster analysis. Researchers face several decisions in this analysis, of which deciding the number of segments is the most important. This decision can be made on statistical fit-indices and/or more subjective grounds, such as the interpretation of the several cluster solutions.

8.6.1.1 Execution of cluster analysis

For now, we assume that most analysts still rely on the available cluster methods in statistical packages, such as K-Means. We consider five important steps when executing such an analysis:

1. Selection of cluster variables
2. Data preparation
3. Running the analysis
4. Selecting the number of clusters
5. Profiling the clusters.

8.6.1.2 Selection of cluster variables

Typically, two general types of variables are distinguished: internal (segmentation) variables and external (profiling) variables. Segmentation variables are used for creating segments, whereas profiling variables are used to describe the segments. The selection of segmentation variables should ideally be based on some underlying idea of which segments could be present and, more specifically, on which bases one aims to segment the market. Segmentation can be done using, for example, socio-demographic variables, values, lifestyles, perceived benefits, brand and product usage, or customer profitability. If one aims to develop benefit segmentation, the segmentation variables should measure benefits of products.

8.6.1.3 Data preparation

For cluster analysis, there are two main problems. First, the number of segmentation variables is frequently large. Second, the measurement scale frequently does not fit. We will first discuss the second problem. Although cluster analysis is rather flexible, more continuous scales are preferred. Scales with multiple non-ordered categories create problems. However, binary variables (e.g., gender) can, if required, be included. The first problem leads to massive interpretation problems for clusters. We therefore frequently first use a data-reduction technique, such as PCA, to end up with a lower number of cluster variables (see Section 8.6.2). The component scores can subsequently be included in the cluster analysis. Note that one should be careful when using this technique because the derived principal components only explain limited variance in the underlying variables. The variables should then be included in the analysis.

8.6.1.4 Running the analysis

Many cluster methods are available within the standard software packages, and within these methods specific options are also available. An analyst should have sufficient knowledge about each of these methods and their options to make an informed choice. Typically, analysts also combine specific methods. For example, the hierarchical cluster analysis on a small sample is very useful as a first step in an analysis (hierarchical cluster analysis is less suited for large datasets and the selection of clusters is also more difficult with a larger number of data points).

Based on this hierarchical cluster analysis, the number or the range of clusters can be determined. These clusters and their average values can then be used as input into a K-means analysis, which can handle larger datasets more easily.

8.6.1.5 Selection of the number of clusters

A dendrogram can be used to select the number of clusters. A dendrogram shows how specific cases are combined into clusters (see Figure 8.32). Based on a subjective assessment of the dendrogram a range of cluster solutions can be considered. When running the subsequent K-means analysis, several solutions for this range of clusters can be achieved. A definite selection should then be based on segmentation criteria, such as the size of the clusters and their interpretability. Profiling the cluster solutions (see Section 8.3) can be helpful in interpreting the clusters.

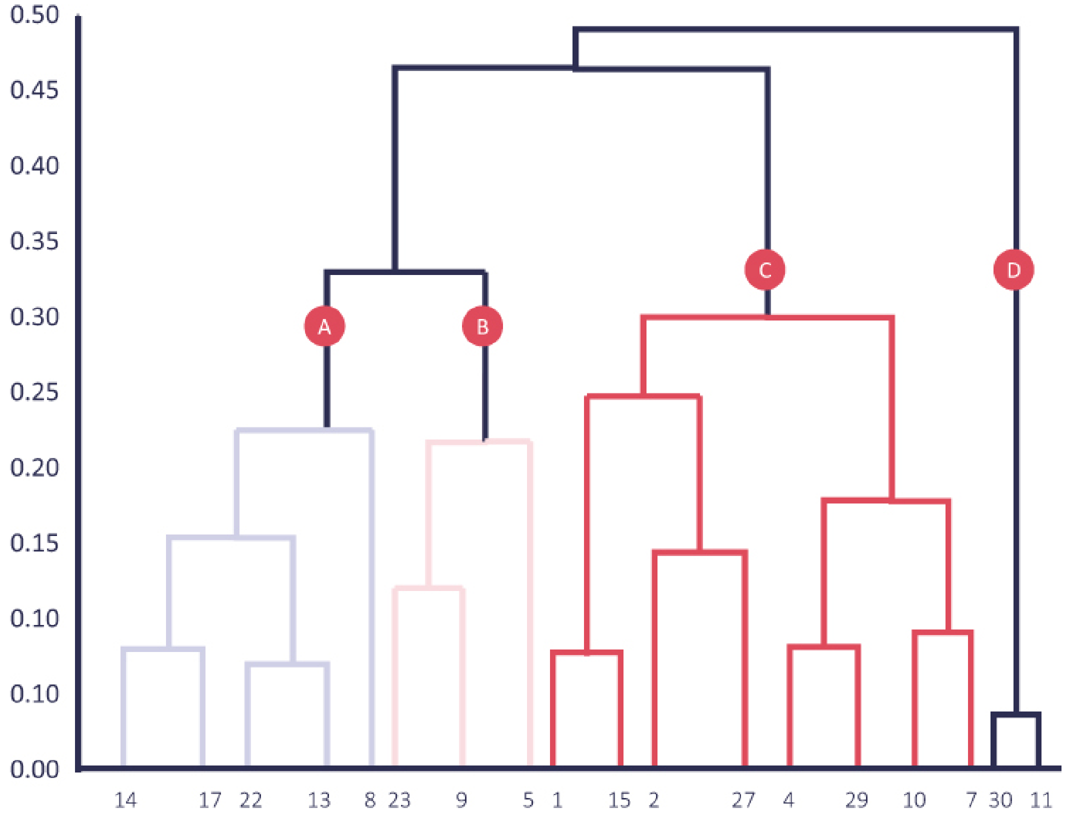


FIGURE 8.32 Example of a dendrogram

8.6.1.6 Profiling the clusters

The profiling of clusters can be done on the internal cluster variables and the external (profiling) variables not included in the cluster analysis. This helps to gain a further understanding of the found segments, as profiling variables such as socio-demographic characteristics, media usage, channel usage, and brand purchase behavior are being used. One way to profile the clusters is to use logistic regression or multinomial regression, which can show which variables explain specific cluster memberships. One can also use this technique to classify customers in a cluster.

Based on profiling, target segments can be chosen, and target-marketing strategies can be developed. A smart way to understand differences between segments is to use a scatterplot in which the segments are related to the most influential profile variables (see Figure 8.33).

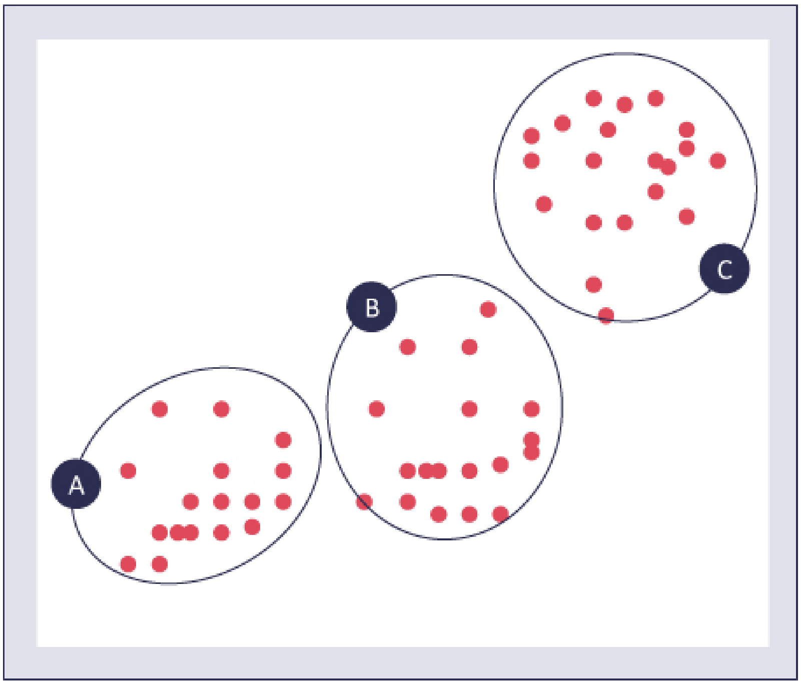


FIGURE 8.33 Visualization of clusters

8.6.2 Principal components analysis

The main use of principal components analysis (PCA) within the analyses we discuss is as a data-reduction technique. Analysts typically analyze a large number of variables which frequently strongly correlate. Although analyzing this large number of variables is in principle possible, it creates problems. The number may just become so large that interpreting the different outcomes for the different variables becomes too complex. Especially in regression-type models, the large number of variables can create problems in estimation due to multi-collinearity. As a consequence, the estimated coefficients and accompanying significance levels are no longer reliable.

One solution for this problem is to use PCA (also referred to as "factor analysis"). This technique reduces the number of variables into a lower number of uncorrelated principal components (PCs) or factors, which represent an underlying dimension. For example, if one has variables on desktop usage (denoted as x_1), mobile usage (denoted as x_2), and tablet usage (denoted as x_3), these variables could end up in one PC, representing the digital level of customers:

$$PC = w_1x_1 + w_2x_2 + w_3x_3$$

where w_1, w_2, w_3 are weights that determine to what extent the PC is related to each of the variables. Small weights imply that the PC does not represent that variable, and larger weights reveal that the corresponding variable is an important determinant of that PC. Multiple PCs can be extracted from a set of variables, but the number of PCs should be smaller than the number of variables, as the aim is dimension reduction. Hence, in the foregoing example, we extract two PCs at maximum:

$$\mathrm{PC}_1 = w_{11}x_1 + w_{12}x_2 + w_{13}x_3$$

$$\mathrm{PC}_2 = w_{21}x_1 + w_{22}x_2 + w_{23}x_3$$

In these equations, the weights now have a second index which indicates that the weights of a variable differ for the two components. For example, if w_{11} is large, and w_{12} and w_{13} are small, PC_1 represents fixed device usage, and if w_{21} is small, and w_{22} and w_{23} are large, PC_2 reflects mobile device usage.

The resulting PCs can, for example, be used to profile or as input for cluster analysis. The use of PCs in a regression model reduces multicollinearity problems (see Section 9.3) as the PCs are uncorrelated. PCA is not solely used for data reduction: it can also be used substantively in survey scale development and brand positioning research (e.g., Lilien, Rangaswamy, & De Bruyn, 2017).

The key challenge is to find the right number of interpretable PCs while preserving as much of the original data's variation as possible. We illustrate the steps to conduct a PCA using data that can be found on Kaggle.com.[1] In this example, we consider the satisfaction scores of 103,904 passengers on 14 dimensions of their experience with an airline. This is a large enough sample for PCA, exceeding the rule of thumb that the sample size should be at least 150 observations or at least five to ten observations per variable. The scores are measured on a 1–5 scale. The formal requirement for PCA is that the variables should be at least interval-scaled, but ordinal variables are very frequently used as well.

The 14 satisfaction dimensions are:

- Inflight Wi-Fi service
- Departure/arrival time convenient
- Ease of online booking
- Gate location
- Food and drink
- Online boarding
- Seat comfort
- Inflight entertainment
- On-board service
- Leg room service
- Baggage handling
- Check-in service
- Inflight service
- Cleanliness.

Dimension reduction is only possible if there is a significant correlation between the variables in PCA. Hence, the correlation matrix of the variables should not be close to an identity matrix, as that would indicate that all variables are uncorrelated and grouping them into PCs would not be sensible. In order to investigate this, the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy can be computed, and Bartlett's test can be conducted. The KMO measure indicates the size of partial correlations. If the partial correlations between the variables are too small the data is not suited for PCA. Guidelines for assessing the KMO measure are:

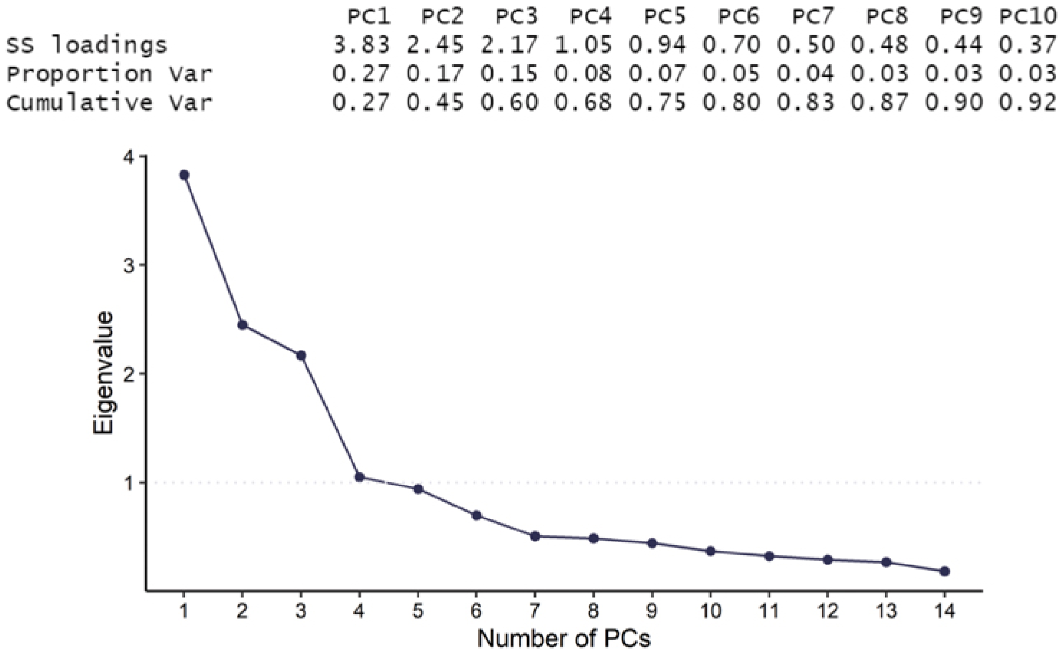
- – magnificent
- – fine
- – sufficient
- – mediocre
- – bad (perfectly uncorrelated).

In our airline example, KMO equals 0.78. Bartlett's test of sphericity tests whether the correlation matrix of the variables deviates significantly from the identity matrix. If the null hypothesis cannot be rejected this indicates that the variables are not sufficiently related and PCA is not applicable for these data. In our airline example, Bartlett's test is significant.

The next step is to select the number of PCs. To guide this decision, the following criteria can be used:

1. Retain PCs with eigenvalue larger than one
2. The selected PCs should jointly explain more than 60% of the variability in the data
3. Retain PCs that individually explain more than 5% of the variability in the data
4. Stop adding PCs when the slope of the scree plot is clearly leveling off (the "elbow")
5. The factors should have a clear interpretation.

PCs with eigenvalues larger than one each account for more variance than a single variable, so that retaining them contributes to the goal of dimension reduction. The second criterion ensures that dimension reduction does not result in the loss of too much of the total variability in the original data. The same logic holds for the third criterion, but focusses on retaining single PCs. The fourth criterion concerns the scree plot, which is a plot of the eigenvalues against the number of PCs that are extracted from the data. At the point where the slope levels off, adding more PCs does not contribute much to dimension reduction. In Figure 8.34 we present parts of the R output of a PCA analysis of the airline data.

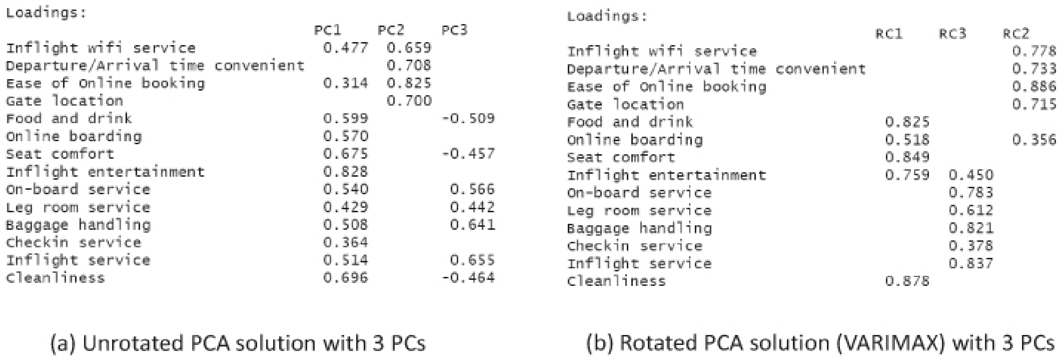


```
                     PC1  PC2  PC3  PC4  PC5  PC6  PC7  PC8  PC9 PC10
SS loadings         3.83 2.45 2.17 1.05 0.94 0.70 0.50 0.48 0.44 0.37
Proportion Var      0.27 0.17 0.15 0.08 0.07 0.05 0.04 0.03 0.03 0.03
Cumulative Var      0.27 0.45 0.60 0.68 0.75 0.80 0.83 0.87 0.90 0.92
```

FIGURE 8.34 R output of a PCA on airline data

The first line in Figure 8.34 presents the eigenvalues for the first ten PCs, ordered according to explained variance. Applying the first criterion leads to the conclusion that four PCs should be extracted. For the second criterion, we inspect the third line in Figure 8.34 and conclude that extracting three PCs would satisfy this requirement. The second line in Figure 8.34 should be inspected for the third criterion. There we see that up to and including PC6 all PCs explain at least 5%, meaning that six PCs should be retained according to this criterion. The scree plot indicates a sharp drop after PC3. Selecting three factors seems to agree fairly well with most criteria, except for criterion 2. Hence, we proceed and extract 3 PCs for the airline data.

In the left panel of Figure 8.35 we present the loadings table for applying PCA to the airline data with three PCs. The numbers in this table are called “Loadings” and can be interpreted as correlations. Hence, they indicate how strongly each extracted factor is related to the variables in the data. For readability, all loadings with an absolute value smaller than 0.3 are omitted.



```
Loadings:
                                  PC1    PC2    PC3   
Inflight wifi service              0.477  0.659       
Departure/Arrival time convenient         0.708       
Ease of Online booking             0.314  0.825       
Gate location                             0.700       
Food and drink                     0.599        -0.509
Online boarding                    0.570              
Seat comfort                       0.675        -0.457
Inflight entertainment             0.828              
On-board service                   0.540         0.566
Leg room service                   0.429         0.442
Baggage handling                   0.508         0.641
Checkin service                    0.364              
Inflight service                   0.514         0.655
Cleanliness                        0.696        -0.464
```

(a) Unrotated PCA solution with 3 PCs

```
Loadings:
                                  RC1    RC3    RC2   
Inflight wifi service                            0.778
Departure/Arrival time convenient                0.733
Ease of Online booking                           0.886
Gate location                                    0.715
Food and drink                     0.825              
Online boarding                    0.518         0.356
Seat comfort                       0.849              
Inflight entertainment             0.759  0.450       
On-board service                          0.783       
Leg room service                          0.612       
Baggage handling                          0.821       
Checkin service                           0.378       
Inflight service                          0.837       
Cleanliness                        0.878              
```

(b) Rotated PCA solution (VARIMAX) with 3 PCs

FIGURE 8.35 Loadings tables of a 3-PC solution for the airline data

The next step in a PCA is to interpret the extracted factors. In principle, this is deduced from the loadings table. The larger the loading, the more important the variable is for the PC. As a rule of thumb, loadings with an absolute value of 0.5 or higher are considered to be high loadings. If variables do not contribute to any PCs (only low loadings), they are excluded from the analysis. If variables have high loadings in multiple PCs it is typically hard to interpret the PCs. A solution is then to re-run the analysis and extract more PCs.

The loadings in panel (a) of Figure 8.35 show a large number of high loadings for PC1, making it hard to interpret this PC. This can be resolved by applying a rotation method (here: VARIMAX). The loadings for the rotated PCs (RCs) can be found in panel (b) of Figure 8.35. The high loadings for "Food and drink," "Online boarding," "Seat comfort," "Inflight entertainment," and "Cleanliness" suggest that PC1 can be interpreted as the comfort factor. Similarly, the high loadings for "Onboard service," "Leg room service," "Baggage Handling," and "Inflight Service" suggest that PC2 is the service factor. The high loadings for "In flight Wi-Fi service," "Departure/Arrival time convenient," "Ease of Online booking," and "Gate Location" suggest an interpretation of convenience for PC3.

As a last step in PCA, the PC scores can be saved in the database. Importantly, these scores have an average of 0 and a standard deviation of 1. The scores are difficult to interpret. A higher score on, for example, a PC measuring comfort only suggests that customers are more satisfied with the comfort-related aspects of their airline experience.

8.7 CONCLUSIONS

In this chapter we discussed several techniques for explorative analysis. We provided an overview of methods for reporting data and for investigating one-to-one relationships, and we discussed a number of unsupervised learning techniques that can be used to identify clusters in the data, or lower the dimensionality in the data. In the R script that accompanies this chapter the techniques are illustrated using numerical examples. For more detail on the methods discussed in this chapter we refer to textbooks on multivariate analysis (e.g. Malhotra, 2019).

ASSIGNMENT

Call.com is one of the largest companies in the telecom market in the Netherlands. Call.com is a supplier of telecommunication and ICT services and offers consumers landlines, mobile service, internet, and television.

In the mobile communications market, Call.com has two brands: Call.com and Hello. In total, Call.com has approximately 4.5 million customers with a mobile phone plan.

Call.com's revenue has grown by 3% in the past year. The growth in the number of customers, however, is minimal, namely 0.5% (see also Figure 8.36).

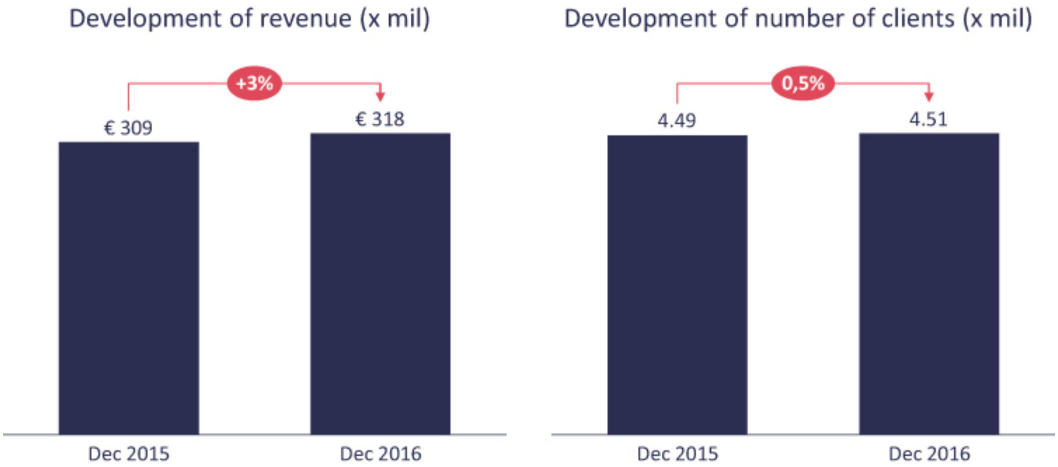


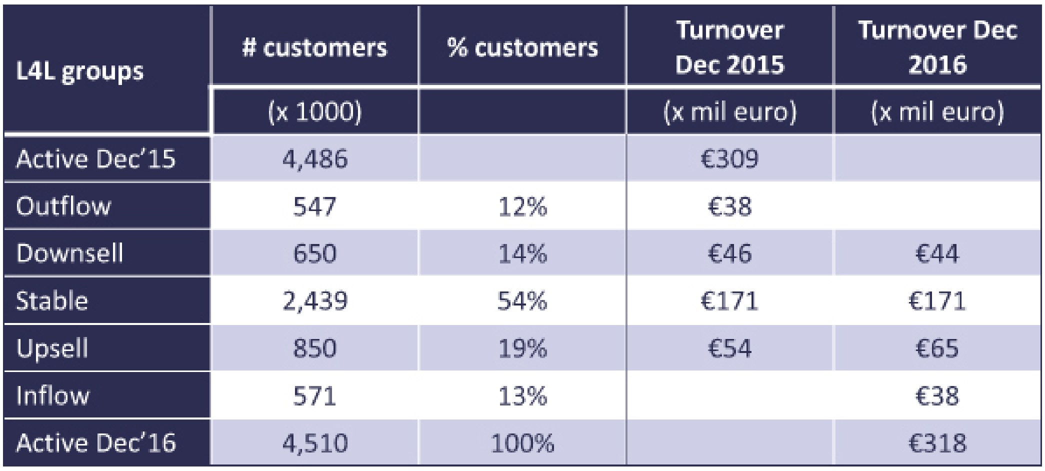
FIGURE 8.36 Development of revenue and number of clients Call.com 2015/2016

Part 1

Call.com wants to understand what causes the low customer growth and develop a strategy to ensure growth in the number of customers to at least the level of 3% per year.

Questions Part 1:

1. Identify possible reasons why Call.com has a higher growth in customer value than in the number of customers.
2. Based on the table in Figure 8.37, make a like-4-like analysis of the past year for Call.com and visualize this in a waterfall figure.
3. What conclusions can you draw based on the results of this analysis?
4. Sketch a realistic scenario for increasing the growth in the number of customers for the following year. In addition, calculate the impact this has on the total revenue and the average revenue per customer, based on the like-4-like analysis.



L4L groups	# customers	% customers	Turnover Dec 2015	Turnover Dec 2016
	(x 1000)		(x mil euro)	(x mil euro)
Active Dec'15	4,486		€309	
Outflow	547	12%	€38	
Downsell	650	14%	€46	€44
Stable	2,439	54%	€171	€171
Upsell	850	19%	€54	€65
Inflow	571	13%		€38
Active Dec'16	4,510	100%		€318

FIGURE 8.37 Input table for the like-4-like analysis

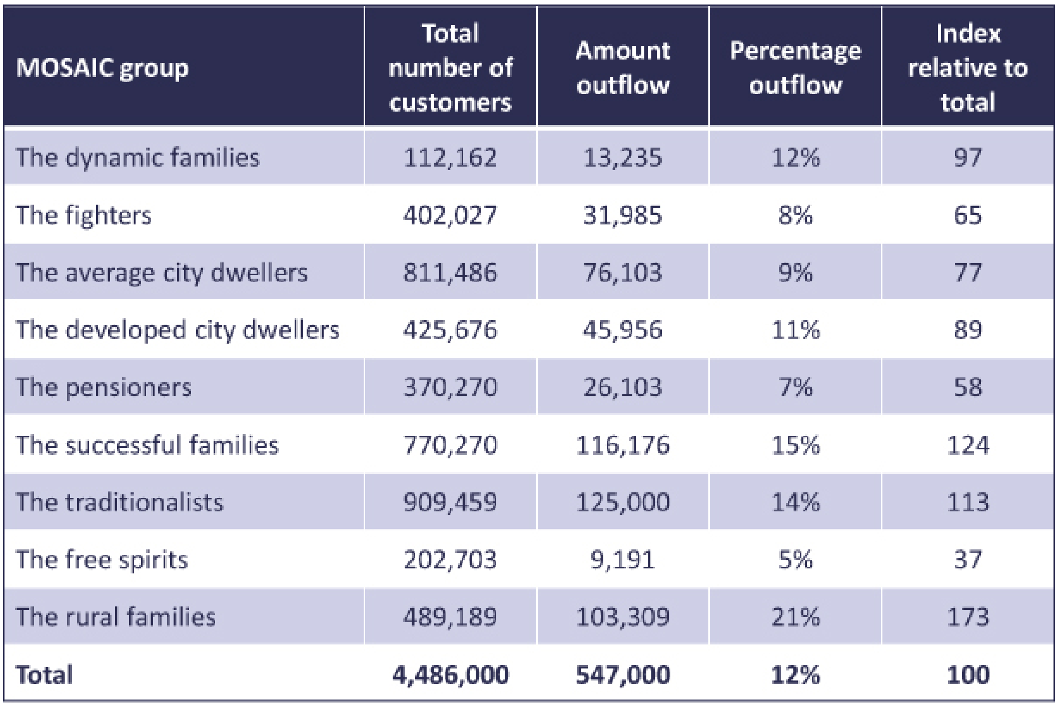
Part 2

One of the hypotheses explaining why the growth in the number of customers has slowed to only 0.5% is that Call.com has lost many customers in the youth segment. The Marketing Intelligence department has made a profile analysis of all customers who canceled their plan in the past year and compared this with the total customer base. Appendix 1 shows the output of this analysis.

Questions Part 2:

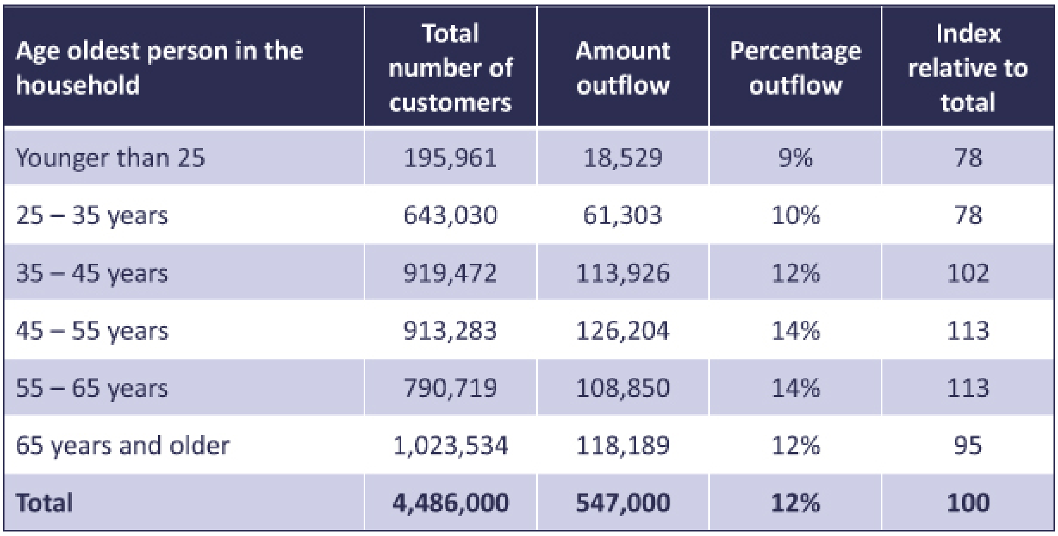
5. Why would you use a profile analysis to solve this question?
6. With what type of data was this profile analysis carried out?
7. Based on the results of the profile analysis, would you accept or reject the above hypothesis? Explain why.

Appendix 1. Profile analysis churners Call.com



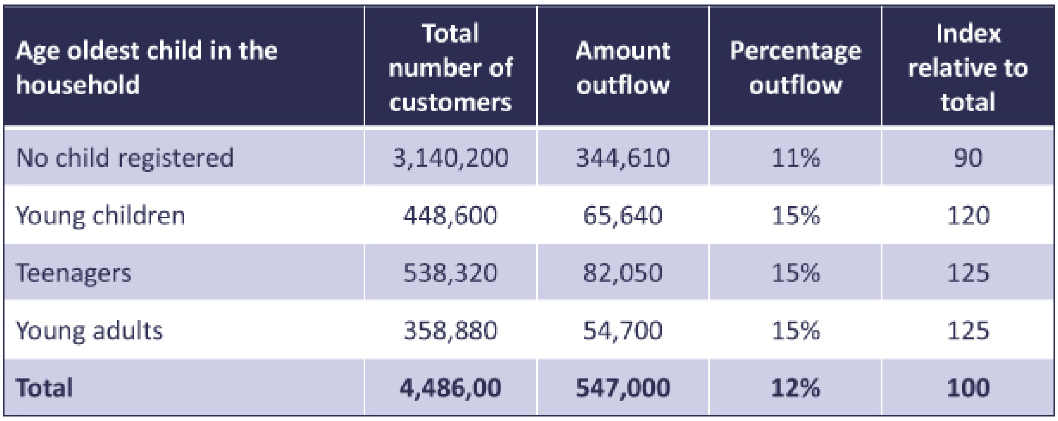
MOSAIC group	Total number of customers	Amount outflow	Percentage outflow	Index relative to total
The dynamic families	112,162	13,235	12%	97
The fighters	402,027	31,985	8%	65
The average city dwellers	811,486	76,103	9%	77
The developed city dwellers	425,676	45,956	11%	89
The pensioners	370,270	26,103	7%	58
The successful families	770,270	116,176	15%	124
The traditionalists	909,459	125,000	14%	113
The free spirits	202,703	9,191	5%	37
The rural families	489,189	103,309	21%	173
Total	**4,486,000**	**547,000**	**12%**	**100**

FIGURE 8.38 Assignment Chapter 8, Appendix (Part 1)



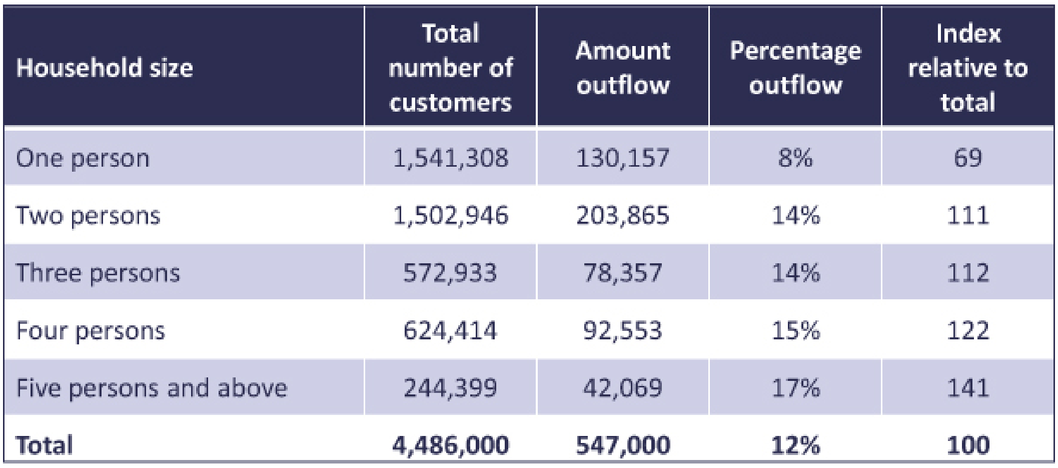
Age oldest person in the household	Total number of customers	Amount outflow	Percentage outflow	Index relative to total
Younger than 25	195,961	18,529	9%	78
25 – 35 years	643,030	61,303	10%	78
35 – 45 years	919,472	113,926	12%	102
45 – 55 years	913,283	126,204	14%	113
55 – 65 years	790,719	108,850	14%	113
65 years and older	1,023,534	118,189	12%	95
Total	**4,486,000**	**547,000**	**12%**	**100**

FIGURE 8.39 Assignment Chapter 8, Appendix (Part 2)



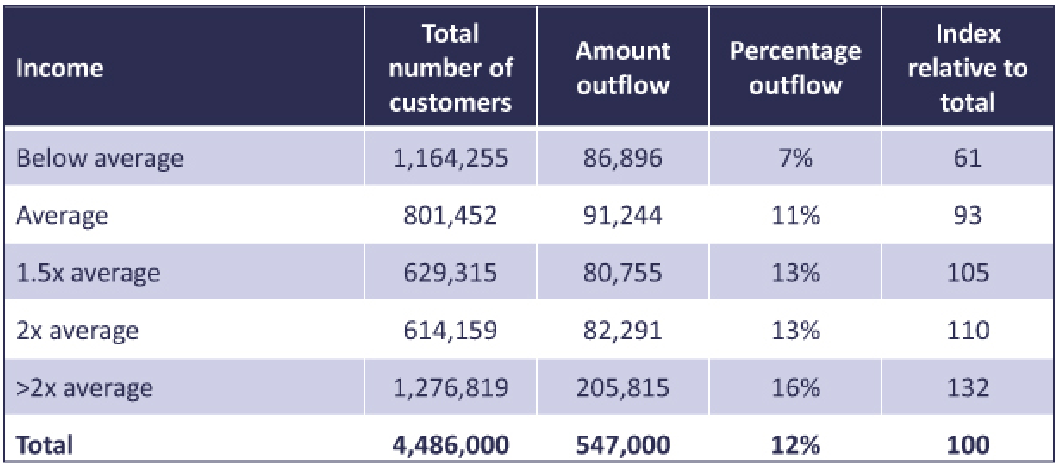
Age oldest child in the household	Total number of customers	Amount outflow	Percentage outflow	Index relative to total
No child registered	3,140,200	344,610	11%	90
Young children	448,600	65,640	15%	120
Teenagers	538,320	82,050	15%	125
Young adults	358,880	54,700	15%	125
Total	**4,486,00**	**547,000**	**12%**	**100**

FIGURE 8.40 Assignment Chapter 8, Appendix (Part 3)



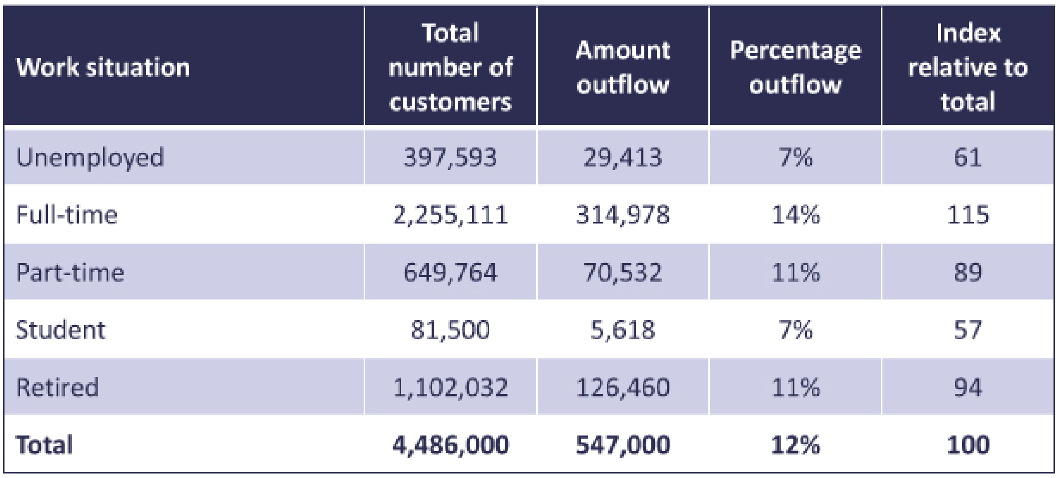
Household size	Total number of customers	Amount outflow	Percentage outflow	Index relative to total
One person	1,541,308	130,157	8%	69
Two persons	1,502,946	203,865	14%	111
Three persons	572,933	78,357	14%	112
Four persons	624,414	92,553	15%	122
Five persons and above	244,399	42,069	17%	141
Total	**4,486,000**	**547,000**	**12%**	**100**

FIGURE 8.41 Assignment Chapter 8, Appendix (Part 4)



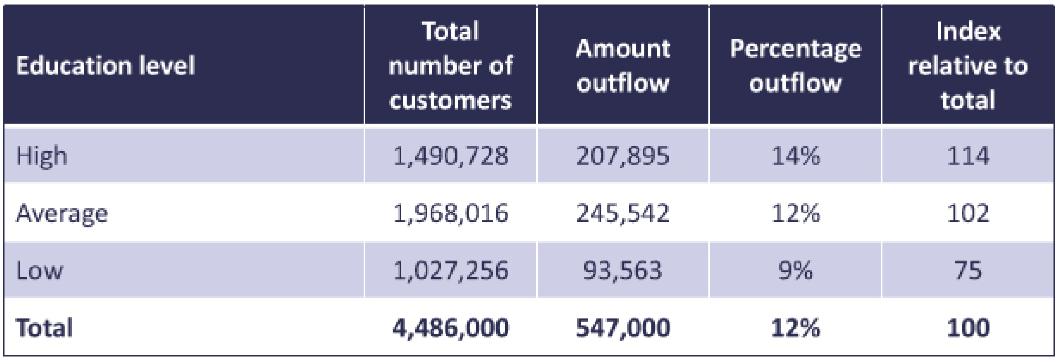
Income	Total number of customers	Amount outflow	Percentage outflow	Index relative to total
Below average	1,164,255	86,896	7%	61
Average	801,452	91,244	11%	93
1.5x average	629,315	80,755	13%	105
2x average	614,159	82,291	13%	110
>2x average	1,276,819	205,815	16%	132
Total	**4,486,000**	**547,000**	**12%**	**100**

FIGURE 8.42 Assignment Chapter 8, Appendix (Part 5)



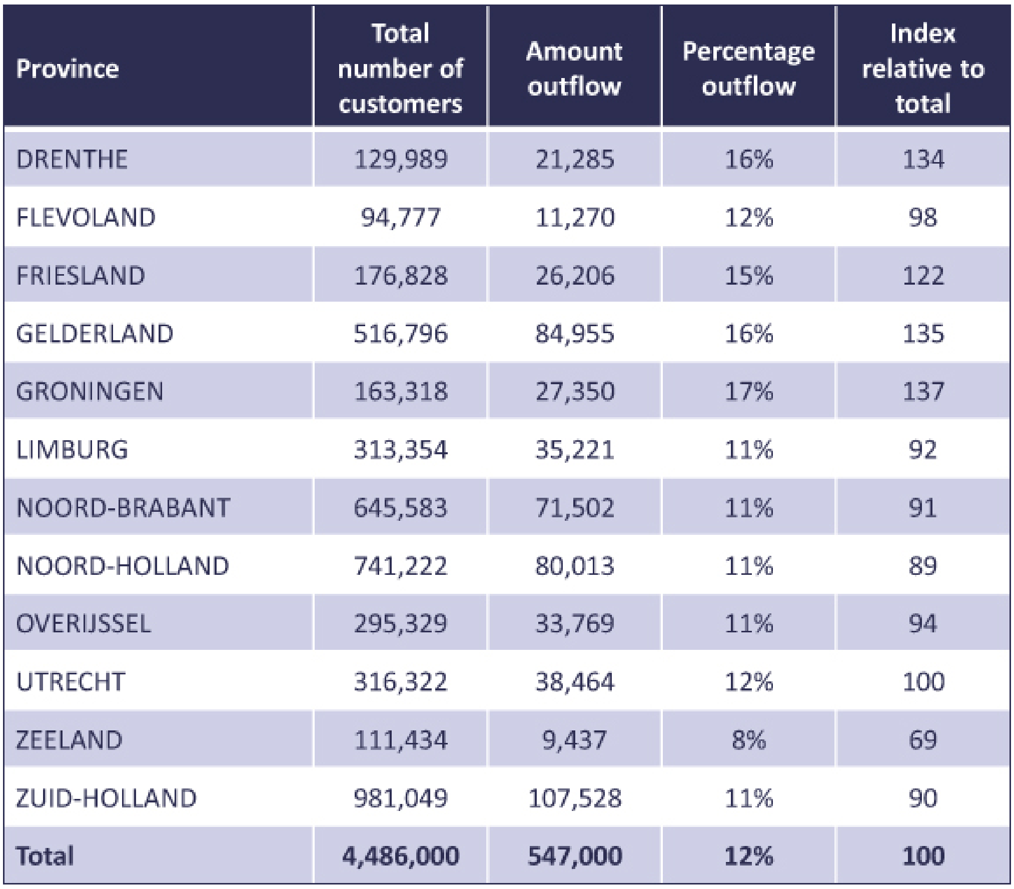
Work situation	Total number of customers	Amount outflow	Percentage outflow	Index relative to total
Unemployed	397,593	29,413	7%	61
Full-time	2,255,111	314,978	14%	115
Part-time	649,764	70,532	11%	89
Student	81,500	5,618	7%	57
Retired	1,102,032	126,460	11%	94
Total	**4,486,000**	**547,000**	**12%**	**100**

FIGURE 8.43 Assignment Chapter 8, Appendix (Part 6)



Education level	Total number of customers	Amount outflow	Percentage outflow	Index relative to total
High	1,490,728	207,895	14%	114
Average	1,968,016	245,542	12%	102
Low	1,027,256	93,563	9%	75
Total	**4,486,000**	**547,000**	**12%**	**100**

FIGURE 8.44 Assignment Chapter 8, Appendix (Part 7)



Province	Total number of customers	Amount outflow	Percentage outflow	Index relative to total
DRENTHE	129,989	21,285	16%	134
FLEVOLAND	94,777	11,270	12%	98
FRIESLAND	176,828	26,206	15%	122
GELDERLAND	516,796	84,955	16%	135
GRONINGEN	163,318	27,350	17%	137
LIMBURG	313,354	35,221	11%	92
NOORD-BRABANT	645,583	71,502	11%	91
NOORD-HOLLAND	741,222	80,013	11%	89
OVERIJSSEL	295,329	33,769	11%	94
UTRECHT	316,322	38,464	12%	100
ZEELAND	111,434	9,437	8%	69
ZUID-HOLLAND	981,049	107,528	11%	90
Total	**4,486,000**	**547,000**	**12%**	**100**

FIGURE 8.45 Assignment Chapter 8, Appendix (Part 8)

NOTE

1. Weblink: https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction, accessed on March 12, 2021.

REFERENCES

David Shepard Associates. (1999). *The New Direct Marketing: How to Implement a Profit-driven Database Marketing Strategy*. Europe: McGraw-Hill Education.

Lilien, G. L., Rangaswamy, A., & De Bruyn, A. (2017). *Principles of Marketing Engineering and Analytics* (3rd edition). State College: DecisionPro, Inc.

Malhotra, N. K. (2019). *Marketing Research: An Applied Orientation, Global Edition* (7th edition). Upper Saddle River, NJ: Pearson Education Limited.

Pauwels, K. H. (2017). Traditional time-series models. In: P. Leeflang, J. Wieringa, T. Bijmolt, K. Pauwels (eds) *Advanced Methods for Modeling Markets*. *International Series in Quantitative Marketing*. Cham: Springer. https://doi.org/10.1007/978-3-319-53469-5_3

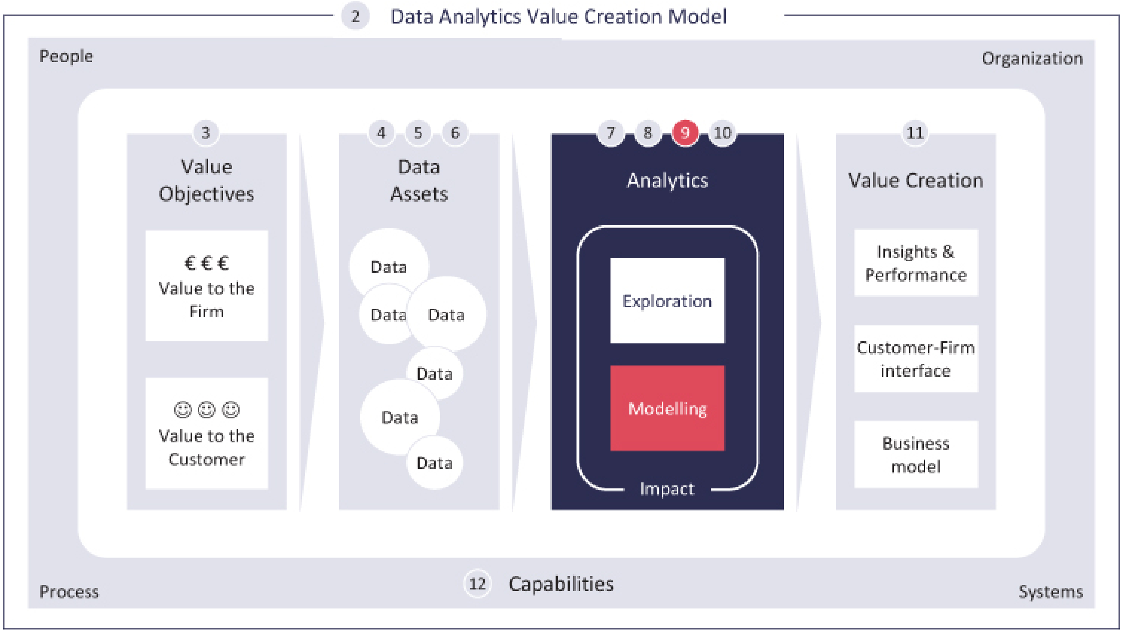
Verhoef, P. C., Van Doorn, J., & Dorotic, M. (2007). Customer value management: An overview and research agenda. *Marketing Journal of Research and Management*, *3*(2), 105–120.

Wedel, M. & Kamakura, W. A. (2000). *Market Segmentation: Conceptual and Methodological Foundations* (2nd edition). Springer US. https://doi.org/10.1007/978-1-4615-4651-1

CHAPTER 9

Data modeling

DOI: 10.4324/9781003011163-9



9.1 INTRODUCTION TO DATA MODELING

In Chapter 8 we discussed graphical techniques and tests that are suited for data exploration to obtain an overview of the characteristics of the individual variables, and to generate a first idea about the one-to-one relationships between a KPI and potential drivers. In this chapter, we take these insights further to investigate how a selection of multiple drivers jointly influences the KPI. We do this by creating models for the KPI.

While there can be several notions for the term “model,” we see a model as a representation of reality, to understand that reality better. Since a model is only a simplification of reality, it will never be able to capture all aspects of reality perfectly. This means that a model will never be complete or able to predict the KPI of a data science project without any errors. However, it helps us to identify systematic relationships between the KPI of a data science project and a set of drivers, which can be used to generate approximations of reality. And these can be very valuable when one needs to choose an action, make a decision, or generate predictions. For example, a churn model cannot be expected to generate perfect churn predictions for all individual customers. Yet, these predictions can be very helpful for efficient targeting of potential churning customers in a retention campaign.

Consequently, a model does not need to be perfect to be useful for a data science project. Aiming for the heroic goal of capturing all aspects of a complex business problem in a model is setting up for failure. At the same time, the better a model accounts for relevant phenomena that affect the KPI, the more useful that model will be. Finding a balance between these two opposing forces is one of the difficult challenges in building a "good" model and requires the input of all relevant stakeholders of a data science project.

Building a model forces the stakeholders of a data science project to abstract away from distracting details and only focus on key elements instead. This contributes to the generalizability of its outcomes and makes a model less *ad hoc* than one-time analyses. In turn, this leads to improved transparency and consistency in data-driven decision making. It also empowers a data scientist to present scenarios or predictions that are not limited to observed cases but are based on inter- or extrapolations of the data. However, such projections should remain within reasonable limits to be of practical value.

In contrast to the exploratory techniques that we discussed in Chapter 8, the models that we consider in this chapter are explicitly suited to accommodating the simultaneous effects of multiple drivers. This is relevant for many data science projects because it allows for ranking the drivers of the KPI according to their impact so that decisions can prioritize the most influential drivers. For example, when allocating a media spending budget, it is relevant to know what channels elicit the strongest consumer response. This requires modeling consumer response as a function of the expenditures on each of these channels.

Depending on the complexity of the data science project, models can be very simple, requiring the collection of a small number of variables, and using simple statistical techniques for analysis (e.g., a simple t-test). However, more intricate business issues may require extensive data collection and more sophisticated analyses.

A model can contribute to a data science project in several ways. It can be used for descriptive, predictive, and normative purposes. When the intended use of a model is descriptive, the model's purpose is to determine the separate effects of a collection of drivers on the KPI. For example, a model that identifies the different effects of the marketing variables of a number of competitors on own brand sales can be used to characterize the competitive landscape (see Section 9.3 for an example). Because models are typically calibrated with historical data, descriptive models are useful for answering the "Why did this happen?" question.

When a model is mainly used for predictive purposes, identifying separate effects of the variables in the model is less relevant. Instead, the model's ability to accurately predict future values of the KPI is the dominant performance criterion. Predictive models help answer the "What's next?" question. An appropriate answer to this question is relevant to many business problems. For example, predictions for the churn rate over the coming months may guide the decision whether or not to initiate a retention campaign.

A normative model is used to establish optimal levels of variables that are under the control of the decision maker. For example, a normative model may be used to determine the price level that maximizes profit. Normative models are used to answer the "What should I do?" question.

Since models are no more than analogies or a representation of the real-life business phenomenon that we are investigating in a data science project, they need a justification or an explanation to motivate why the analogy is reasonable (Derman, 2012). This justification can be based on theory, on data, or on both.

In Sections 9.2–9.4 we discuss models that rely on theory or existing knowledge in the initial stages of the development of a model and utilize data in later stages to estimate and validate the model. These models represent the more traditional econometrical way of building models in marketing analytics. We would like to stress that by labeling these techniques "traditional," we do not suggest that an up-to-date data scientist should shy away from using them. On the contrary, we believe that for many contemporary data science projects, these are still the first go-to techniques and they remain the most relevant workhorses in any data scientist's toolbox. We discuss the general steps of traditional model building in Section 9.2. The two main examples that fall into this category of models are regression models and logistic regression models. We discuss them in 9.3. and 9.4, respectively.

In Sections 9.5–9.11 we discuss approaches that build on data in the development stages of the model. These data-driven models originate for a large part from the computer science field but are gaining popularity in marketing analytics, mainly because of their excellent predictive performance. Somewhat broadly speaking these techniques are usually associated with "machine learning" and "artificial intelligence," but the boundaries of such classifications are blurry and not always consistent. In contrast to the models that are discussed in Sections 9.2–9.4, the models that correspond to this group can remain entirely data-driven, even up to deployment in a production environment. For example, a neural network that was trained to generate predictions of responses of individual consumers to an email campaign can be used for effective and profitable targeting decisions, without considering theory or asking the question of why customers differ in their response behavior. However, in most data science projects involving customers that we have seen, the urge to also understand the behavior of customers arises sooner or later in the project. Consequently, we take the stand that generating impact with purely data-driven methodology still requires extensive knowledge of the business context and data science projects are less likely to be effective when the analyses only rely on the expertise of specialists who are solely well-versed in methodological aspects.

In the second part of this chapter, we introduce data-driven techniques in Section 9.5. In subsequent sections, we discuss a number them in more detail, starting with decision trees in Section 9.6. Section 9.7 illustrates the use of decision trees in ensemble learning when we discuss bagging, random forests, and boosting. In Section 9.8 we discuss Naive Bayes classification models, and in Section 9.9 we discuss Support Vector Machines. Section 9.10 is devoted to neural networks. We conclude this chapter with a discussion of reinforcement learning in Section 9.11 and present our conclusions in Section 9.12. The right-hand side of the analytical framework depicted in Figure 9.1 shows a schematic overview of this chapter.

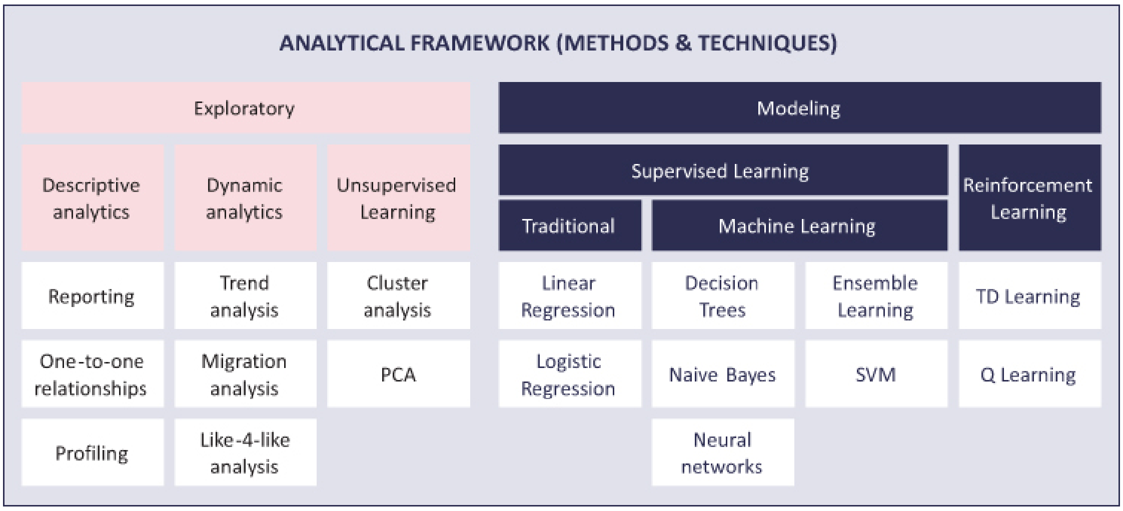


FIGURE 9.1 Analytical framework with different types of analysis methods and techniques

9.2 THEORY-DRIVEN MODELS

Traditional, theory-driven models have a long tradition in marketing analytics. Wedel and Kannan (2016) illustrate how these models were used to improve understanding of customers as early as the 1930s, and that methodologies and data availability have only improved since then. Leeflang and Wittink (1996) describe how these models were more and more usable in practical settings. Leeflang *et al*. (2015) provide a more elaborate overview of these models than we can do in this book, where we want to focus only on the key elements in the development of these models.

Any traditional model consists of several key elements. They are:

- dependent and independent variables
- disturbance term
- the mathematical form of relationship between variables

We illustrate these concepts using a simple formula in Equation 9.1.

$$y_t = \alpha + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t.$$

In Equation 9.1, y_t is the dependent variable, sometimes also (9.1) referred to as the criterion variable. This variable is explained by the joint effect of the terms on the right-hand side of the model. As we will discuss in Subsection 9.2.1, the selection of the dependent variable is of primary importance for the success of the modeling exercise. The x_{1t} and x_{2t} that we observe on the right-hand side of Equation 9.1 are the independent variables, also known as predictor variables. These are drivers that are suggested by theory or existing knowledge to influence y_t. The last term in Equation 9.1 is ε_t. This is the disturbance term or the residual.

The disturbance term reflects the notion that we mentioned in the introduction to this chapter: the explanation of the drivers is in most situations not perfect, and the remaining difference between the model outcome and the observed value for y_t is picked up by ε_t. These differences can originate from different sources: they can be due to random error, measurement error, missing variables, or specification error (functional form), and they can reflect behavior differences that are hard to model (between individuals and in time). These reasons for including an error term may create the impression that the error term is somewhat of an unwanted component of a model that is not worthy of our attention, and in many cases the residual term is ignored when discussing the outcomes of a model. However, this is unwarranted as the residual term governs many of the statistical properties and assumptions of a model. For this reason, the residual term is sometimes referred to as the stochastic part of a model, whereas the other terms on the right-hand side of the model are referred to as the systematic parts of a model. We return briefly to this issue when discussing the statistical validity of a model in Subsection 9.2.3.

The mathematical form of a model concerns the mathematical relationship that we assume for the joint effect of the independent variables and the error term on the dependent variable. In Equation 9.1 we specified a linear relationship between y_t and the two independent variables. The parameters α, β_1 and β_2 govern that relationship: α is the intercept and β_1 and β_2 are the response parameters of x_{1t} and x_{2t}, respectively, and indicate how strongly y_t responds to changes in any of the independent variables.

The steps for building a traditional theory-driven model are extensively described in Leeflang *et al.* (2015). They explain that building a model proceeds in four steps:

1. Specification: deriving $y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$
2. Estimation: obtaining estimates for β_0 and β_1
3. Validation: checking assumptions
4. Use

Building on their work, we describe these four steps briefly in Subsections 9.2.1–9.2.4.

9.2.1 Model specification

In the model specification step, the goal is to arrive at a mathematical formulation of the model. This involves making decisions for the elements of a model that were discussed above: selecting an appropriate dependent variable and the right independent variables, as well as the mathematical form that relates these variables. Furthermore, assumptions for the disturbance term need to be specified. This first step is guided by theory or other forms of existing knowledge so that, in principle, the model specification can be accomplished without using a single data point. For example, when building a churn model, theory might suggest which of the possible churn indicators it is most appropriate to use as a dependent variable and may also indicate which variables to use as independent variables.

In many practical cases, exploratory data analyses also inspire the model specification step. There is nothing wrong with using exploratory data analyses as input for model building, but care should be taken that the model is validated in step 3 with data that was not used in the earlier two steps (Wickham & Grolemund, 2017). To this end, the data is often split up into a training or calibration set and one or more hold-out sets for validation purposes.

The dependent variable should closely relate to a KPI that is central to the first steps in the analytical cycle and can be measured at the customer, brand, or market level. The selected level of aggregation typically has important consequences for the specification step. It impacts the type of model that needs to be specified but also affects the selection of independent variables. For example, when a customer-level binary churn indicator is selected as the dependent variable in an analytical project that investigates churn, the independent variables should also be measured at the customer level, and a logistic specification is a sensible choice for the mathematical form (see Section 9.4). However, if churn is measured at the brand level (e.g., number of defecting customers), the independent variables should also be measured at the brand level, thereby aggregating across customers, and a linear model is a suitable choice for the mathematical form (see Section 9.3).

We advise organizing a sufficient level of support and explicitly inviting input of all relevant stakeholders in the specification step of model development. A high level of involvement in the specification choices, such as the selection of the dependent variable and the independent variables, induces commitment to the data science project and most likely improves the quality of the model specification. Moreover, agreeing on such specification decisions with relevant stakeholders during the specification phase effectively excludes the possibility for them to criticize these decisions in a later stage when, for example, the analyses provide the opposite of the answer that was hoped for.

Little (1970) developed several criteria for a model to be a 'good' model. These can be used to guide the discussions with stakeholders on important specification decisions. A 'good' model should be:

1. Simple. A model should be easy to understand and easy to explain. To this end, models should be parsimonious, i.e., include only a small number of the most influential drivers of the dependent variable. This not only reduces the risk of overfitting (see Section 9.6), but parsimonious models sometimes also predict more accurately than more complex models. Predictions of more complex models can be less accurate because with more estimated parameters, more noise is added to the outcomes of the model as well. The simplicity of a model can be realized by restricting the number of independent variables, limiting the amount of (complex) interactions between variables, and selecting a simple functional form.
2. Evolutionary. A model should be able to accommodate advancing insight into the phenomenon that is studied in a data science project. For example, while running a data science project, it may emerge that a previously unused external data source is a strong predictor of customer purchase behavior. The model should allow for straightforward amendments so that this data source can be included in the analyses. Stimulating stakeholders to partake in the evolutionary improvements of the model can be very helpful in keeping them involved. It is good practice to start with a simple version of the model, jointly identify what relevant attributes are missing, and extend the model with variables that are reasonable measures for these attributes.
3. Complete. This criterion refers to the requirement that a model should capture all aspects that are relevant to the phenomenon that is being modeled. This is true for the focal variables for which hypotheses are formulated during the earlier stages of the analytical cycle, but we

strongly recommend also including those variables which are known to be highly influential, but for which no hypotheses are formulated. Not controlling for such variables may harm the faith that stakeholders have in the outcomes of the analyses. Moreover, it will also be more difficult to determine the correct size and significance of the effects of the focal variables that are included in the analyses. The estimated values of the effects of focal variables may be affected by omitting relevant control variables due to omitted variable biases. The significance of the focal variables is also impacted because the residuals may capture the effects of the omitted control variables, which will increase their variance. In turn, this makes it more difficult to establish significance for any variable in the model. It is hard to provide general guidelines for the completeness of the model, as this strongly depends on the model's purpose and context. Hence, completeness is a relative concept, and should again certainly be confirmed with important stakeholders. Fortunately, the insights that result from opportunity finding (see Section 11.6) provide a good checklist for whether all important aspects of the project are included in the model. A secondary high-over checklist that is helpful for a business project is to think about variables in the following categories:

   - variables that represent potential actions the company can take (e.g., use of marketing instruments)
   - variables that capture potential actions competitors can take
   - variables that capture environmental effects
   - variables that capture longitudinal or dynamic effects (e.g., delayed effects of advertising campaigns).

4. Adaptive. A model should be able to cope with new situations that present themselves over time. This is especially an important criterion for organizations that operate in turbulent markets. To some extent, adaptiveness is related to the evolutionary criterion and can be accomplished by updating model parameters when data on the new situation become available. Alternatively, this can be accomplished by adding or deleting variables when that provides an adequate reflection of the changing context of the model. It may be wise to make the setup of the model modular, especially when frequent adaptations of the model are needed.
5. Robust. The last criterion requires that a model should produce sensible outcomes under all circumstances. For example, a model that predicts market shares over time should always generate values between zero and one, irrespective of the values of the independent variables. Robustness can be accomplished by adding restrictions on the outcomes of the model, e.g., by ensuring that only values in the range 0–1 are predicted by a market-share model. Another possibility is to restrict parameter values, e.g., only allowing positive advertising effects. Other options are to

restrict interpolations to the observed data range, to select a correct mathematical form, and including meaningful interaction terms.

The discussion of the five criteria for a "good" model indicates that model specification is in most cases a balancing act: specifying a (too) simple model might interfere with the completeness or robustness of the model. Hence, it is important to keep stakeholders aligned and a model builder should be very open to any external input on the specification of the model.

Leeflang *et al*. (2015, Section 2.7) point out that the specification step should also include a consideration of how differences among entities should be accounted for. For example, if the scope of the data science project pertains to multiple regions, the question arises whether these regions can be considered similar so that one pooled model suffices, or whether separate models should be specified, one for each region. Another point to consider is how dynamic effects are incorporated in the model. Business variables rarely have only instantaneous effects, so that accommodating lagged effects is relevant in many practical examples. For example, for many marketing variables, notably advertising, it is well established that their effect is highly dynamic (see e.g., Pauwels, 2017), so that controlling for time dependency of their effects is warranted. Leeflang *et al*. (2015, Section 2.8) discuss several options for including dynamic effects in a marketing model.

9.2.2 Estimation

The result of step 1 of the model building process is a mathematical formula that provides a good representation of the phenomenon under study. Model 9.1 is a simple example of such a model. In step 2 of the model building process, we are concerned with obtaining estimates for the unknown parameters of the model. In terms of Model 9.1, this means that we would like to replace α, β_1 and β_2 by numerical values, which allows us to test hypotheses, make predictions, and so on. Estimation is concerned with finding appropriate values for the parameters, such that the estimated model fits the data that was collected to calibrate the model as closely as possible. Note that while (in principle) a model can be specified without using data, step 2 of the model building process explicitly requires data. Statisticians have developed many estimation procedures and which method is appropriate in what setting strongly depends on the model that was specified. For example, OLS is the methodology that is used for a linear regression model (see Section 9.3), whereas so-called Maximum Likelihood estimation is most commonly used to estimate logistic regression models (Section 9.4). For this reason, we will defer discussing the specifics of the estimation step and how to interpret the outcomes specifically for linear regression (Section 9.3) and specifically for logistic regression (Section 9.4).

9.2.3 Validation

In the third step of the model building process, the estimated model is validated. We distinguish three types of validation:

- Face validity
- Statistical validity
- Predictive validity (or testing).

Face validity is concerned with the question of whether the model produces sensible outcomes. Answering this question requires broad knowledge of the business context in which the model is applied. The model builder can rely on her or his own knowledge to assess face validity of a model, but we recommend consulting other sources as well. Specifically, it is advisable to compare the results to earlier findings within the organization or in the literature and to turn to stakeholders and ask them to comment on the model's results. But also, common sense should not be ignored when judging the face validity of a model. Face validity is closely related to the robustness criterion that we discussed in Subsection 9.2.1 but is a slightly broader concept. Assessing the robustness of a model is typically concerned with the question of whether it generates outcomes that are in line with the definition of the dependent variable. In Subsection 9.2.1 we gave the example of a model for market shares. For this dependent variable only values between zero and one agree with the definition of a market share. Such a robustness consideration is certainly a part of face validity, but face validity is also concerned with other issues regarding model outcomes, specifically the sign and the size of the parameter estimates. For example, in a brand sales model for a consumer packaged good where own price is included as one of the independent variables, a negative estimate for the price parameter would have high face validity, but a positive estimate would not. Moreover, given external knowledge that is presented to us in meta-analyses on the price of CPGs (see e.g., Bijmolt, Van Heerde, & Pieters, 2005), price elasticity estimates close to –2.6 have high face validity. Because face validity is highly context-specific, we refrain from giving further guidelines in this book. Instead, we call upon the data scientist to bring together sufficient knowledge from different sources to assess face validity of his or her model.

Statistical validity is related to the statistical soundness of the outcomes of the model. It refers to the overall significance of the model, which answers the question of whether the modeling exercise is useful at all, but also concerns the significance of individual drivers—are the estimates for the associated parameters systematically different from zero? Another important aspect of statistical validity is the fit of the model or the question of whether the model is sufficiently "close" to the data. A last important facet of statistical validity concerns the statistical assumptions that the estimation procedure relies on. Depending on the estimation procedure, there may be several such assumptions, and some may be more explicit than others. The assumption may concern all components of the model. Some assumptions may relate to the independent variables in a model (e.g., the multicollinearity assumption), but an important set of assumptions involves the residual term. This is because the residual term governs many of the statistical properties of a model, which we have already noted at the start of this section. An example of such an assumption is that a regression model requires the residual term to be normally distributed. Which aspects to consider when assessing the statistical validity of a model depends on what model was selected and what technique was used to estimate the model. We therefore discuss statistical validity separately for linear regression (Section 9.3) and logistic regression (Section 9.4).

Predictive validity is concerned with the question of how well the model can predict outside of the data that is used for estimation. To investigate this, the data that is available for the data science project is commonly split into two samples. The first (analysis or estimation) sample is used to estimate parameters (step 2 of the model building process), and to assess the face validity and statistical validity of the model outcomes. The second (holdout or validation) sample is used to investigate the last aspect of step 3 of the model building process: predictive validity. To this end, measures are calculated to judge the predictive validity. For the models discussed in Sections 9.3 and 9.4, we will discuss the most relevant predictive validity measures separately.

9.3 LINEAR REGRESSION

In our discussion of investigating one-to-one relationships in Section 8.3, we mentioned simple regression as a technique to establish relationships between a numerical KPI and a numerical driver. In multiple regression, we consider a numerical KPI, which we consider to be the dependent variable, written as y_t, and multiple potential drivers that are related to the KPI and which we expect from theory or earlier studies to explain the KPI. These variables are also known as independent variables, and the notation is x_{1t}, x_{2t}, $\ldots$, x_{kt}, where the first index indicates the number of the variable, and the second the observation of that variable. Regression is the go-to technique if, in the specification step, multiple independent variables were selected, where at least one of them is numerical. The variables which are not measured on a numerical scale can be taken into account by defining dummies for the levels of the categorical variables. We also suggest following this procedure when all independent variables are categorical. In such cases, it would be possible to apply multi-way ANOVA, but regression provides more interpretation options.

Regression assumes linear additive relationships between the independent variables and the dependent variable. This can be expressed as:

$$y_t = \alpha + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt} + \varepsilon_t. \tag{9.2}$$

This is an additive relationship because the effects of the independent variables cumulate and jointly affect the observations of the KPI. It is linear because each unit increase in x_t leads to the same expected increase in y_t.

The linearity of the relationship of the independent variables and the dependent variable seems to be a restrictive assumption because many business phenomena are intrinsically non-linear. For example, there may exist threshold levels that need to be exceeded before any effect is observed. This is illustrated in Figure 9.2, where the effect of advertising expenditures on sales is minimal if too little is spent on advertising. Only when the level of advertising exceeds A_0 the product is noticed, and sales increases with increasing advertising expenditures. Figure 9.2. also illustrates that advertising effectiveness may level off at very high levels of advertising expenditures (when advertising expenditures exceed A_1) for example, due to decreasing returns to scales. Despite these nonlinearities, assuming linearity may be reasonable for a dataset where the advertising expenditures are within the range of A_0 and A_1.

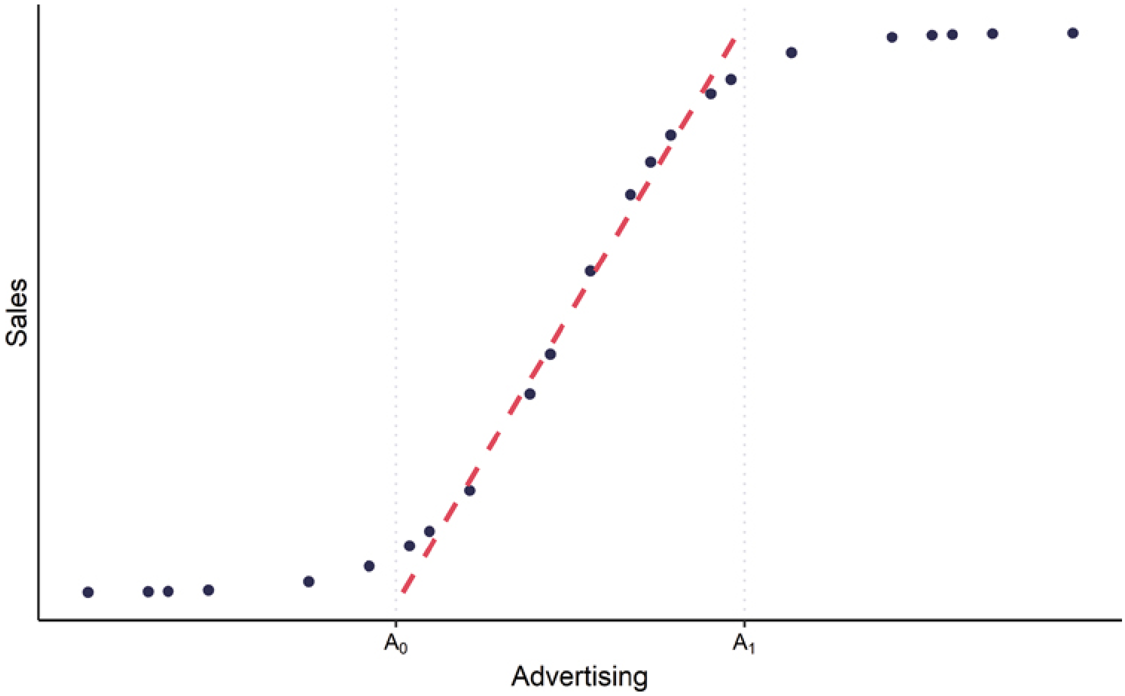


FIGURE 9.2 Linearity assumption

When it is unreasonable to assume a linear relationship for the range of observed variables one can use a transformed variable in Equation 9.2:

$$y_t = \alpha + \beta_1 x_{1t} + \beta_2 \sqrt{x_{2t}} + \cdots + \beta_k x_{kt} + \varepsilon_t.$$

In Equation 9.3, the relationship between x_{2t} and y_t is now curve (9.3) linear. Equation 9.3 is still an additive model, but now specifies a nonlinear relationship between x_{2t} and y_t . Other frequently used transformations involve a square, a logarithmic, or a reciprocal relationship. Linear models which incorporate non-linear effects via transformations of the x-variables are referred to as non-linear additive models. Once these transformed variables are included, estimation of non-linear additive models proceeds in a similar fashion to that for linear additive models.

There is a second set of models, multiplicative models, in which the relationships between the dependent and independent variable(s) are all very non-linear. Frequently, analysts will then apply a log-log transformation (i.e., both the dependent variable y and the independent variables x are log-transformed). The advantage of this method is that the estimated parameters for marketing variables, such as price and advertising, can be interpreted as elasticities.

Also, the additivity property of a linear additive model may be too restrictive for some applications. In many practical situations, the effect of one variable may depend on the level of another variable. In marketing, this is a frequently encountered phenomenon. For example, the effect of a price discount in a brand sales model may depend on whether this price discount was communicated. In such cases, where the effect of one x-variable on y depends on the value of another x-variable, we say that the model should accommodate an interaction effect of the two x-variables. Fortunately, such effects can be accommodated in a regression model by including the product of the variables. In Equation 9.4 we present a model where S_t, the sales of a brand at time t are explained by price at time t (denoted as P_t), feature advertising at time t (denoted as F_t) and the interaction of these two variables.

$$S_t = \alpha + \beta_1 P_t + \beta_2 F_t + \beta_3 P_t F_t + \varepsilon_t.$$

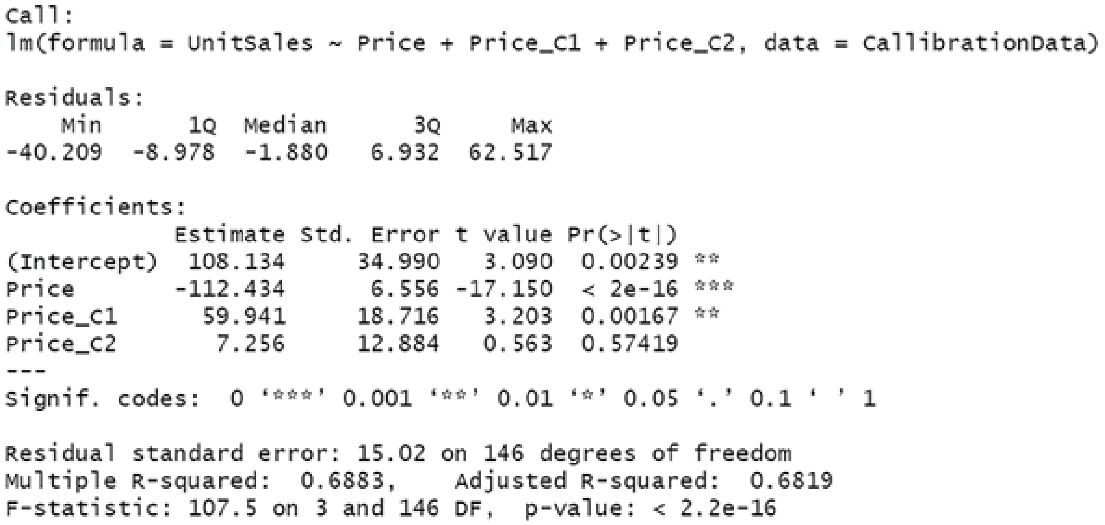
Let us illustrate the model building steps that were discussed in (9.4) Section 9.2 for the linear additive model with an example. Consider a supermarket owner who has collected data on sales of a canned tuna brand. She has available 150 weekly observations on unit sales and is interested in determining how sales respond to own-price changes and to price changes of two competing brands. To this end, she also recorded weekly prices of the canned tuna brand, and of two competing brands.

Step 1 of the model building process resulted in the specification of the following model:

$$\text{UnitSales}_t = \alpha + \beta_1 \text{Price}_t + \beta_2 \text{Price\_C1}_t + \beta_3 \text{Price\_C2}_t + \varepsilon_t,$$

(9.5)

where UnitSales_t represents the brand's unit sales and Price_t the price of the focal brand in week t. The variables Price_C1_t and Price_C2_t measure the weekly prices of competing brands 1 and 2 respectively. The standard technique for estimating a linear additive model (step 2) is ordinary least squares (OLS). The basic underlying idea of OLS is that the estimation method aims to minimize the difference between the estimated value of a dependent variable y and the true value of y (referred to as residual value or disturbance term ε). This implies that the objective of the estimation is to minimize the squared sum of all these differences. OLS is well implemented in many software packages, including R. The R output of estimating Equation 9.5 can be found in Figure 9.3.



```
Call:
lm(formula = UnitSales ~ Price + Price_C1 + Price_C2, data = CallibrationData)

Residuals:
    Min      1Q  Median      3Q     Max 
-40.209  -8.978  -1.880   6.932  62.517 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  108.134     34.990   3.090  0.00239 ** 
Price       -112.434      6.556 -17.150  < 2e-16 ***
Price_C1      59.941     18.716   3.203  0.00167 ** 
Price_C2       7.256     12.884   0.563  0.57419    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.02 on 146 degrees of freedom
Multiple R-squared:  0.6883,    Adjusted R-squared:  0.6819 
F-statistic: 107.5 on 3 and 146 DF,  p-value: < 2.2e-16
```

FIGURE 9.3 R output of estimating Equation 9.5

In step 3, the model is validated. Before we turn to the face validity of the outcomes, we first consider the statistical validity of the outcomes in Figure 9.3. For an effective and efficient application of OLS to estimate a linear additive model, several assumptions need to be fulfilled. Assumptions that need to be investigated are:

- The expected value of the disturbance term should be zero
- The residuals should be uncorrelated over time. This is only relevant for models that involve data that is observed over time. If successive residuals are correlated, this is referred to as autocorrelation
- The variance of the disturbance term should be constant (also referred to as homoscedasticity—violations are referred to as heteroscedasticity)
- The disturbance term should normality distributed
- The independent variables should be uncorrelated. When some of them are strongly correlated, this is referred to as multicollinearity
- The independent variables should all be uncorrelated with the disturbance term (also referred to as exogeneity). Deviations of this assumption are

referred to as endogeneity.

In Figure 9.4 we inspect the residuals of Equation 9.5 visually.

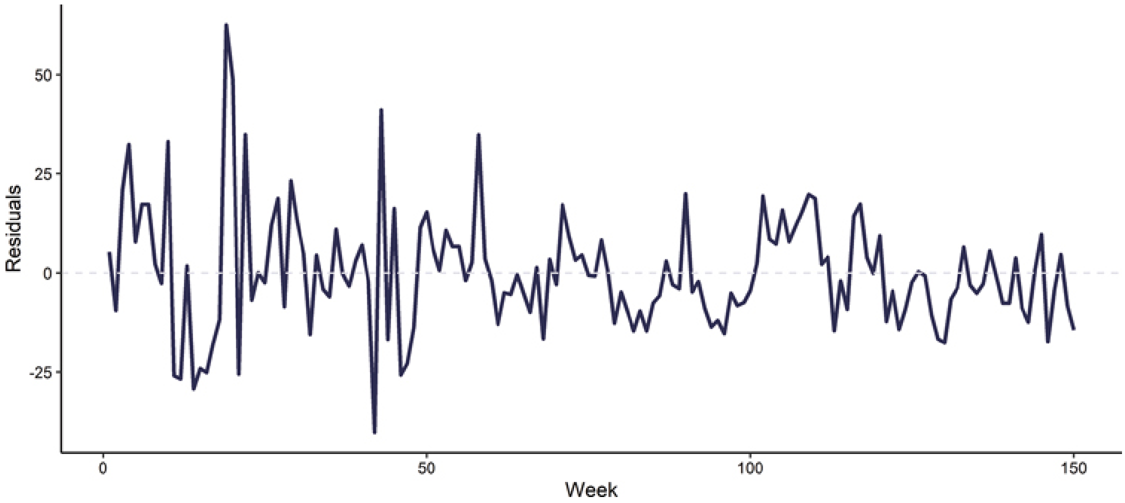
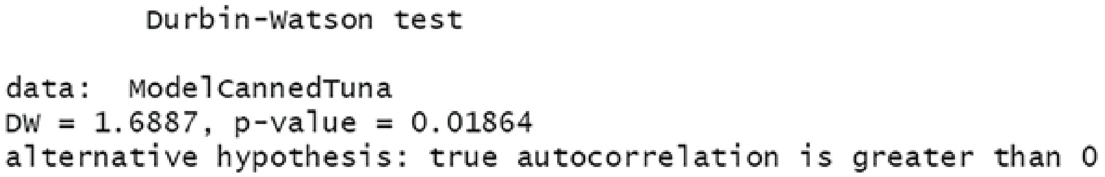


FIGURE 9.4 Plot of the residuals of Equation 9.5 over time

Concerning the second assumption for OLS, the pattern where we observe runs of positive residuals, followed by runs of negative residuals in Figure 9.4 is a possible indication of the presence of autocorrelation. We can investigate this issue further using the Durbin-Watson test. As the *p*-value for this test is smaller than 0.05 (see the R-output in Figure 9.5), we reject the null-hypothesis that autocorrelation is absent.

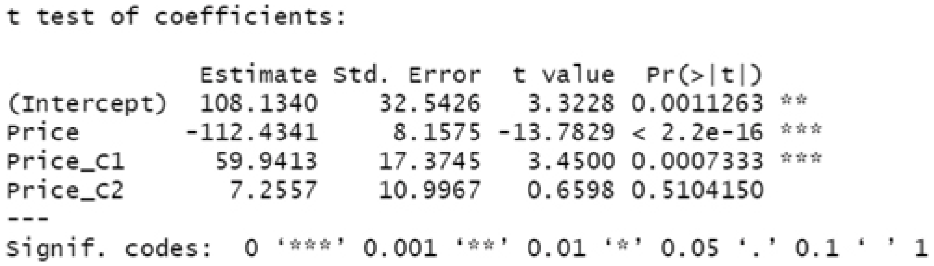


```
          Durbin-Watson test

data:  ModelCannedTuna
DW = 1.6887, p-value = 0.01864
alternative hypothesis: true autocorrelation is greater than 0
```

FIGURE 9.5 R Output of the Durbin Watson test for Equation 9.5

Moreover, the plot of the residuals in Figure 9.4 also hints at the possibility that the variance in the first half of the residuals is higher than in the second half (assumption 3). This is confirmed by a significant Levene’s test (see R script in the online Appendix). Consequently, the outcomes in Figure 9.3 appear to be affected by autocorrelation and heteroscedasticity in the residuals. To investigate to what extent this is the case, we can correct the estimated OLS standard errors using White robust standard errors. Figure 9.6 shows the corresponding R output.



```
t test of coefficients:

             Estimate Std. Error  t value  Pr(>|t|)    
(Intercept)  108.1340    32.5426   3.3228 0.0011263 ** 
Price       -112.4341     8.1575 -13.7829 < 2.2e-16 ***
Price_C1      59.9413    17.3745   3.4500 0.0007333 ***
Price_C2       7.2557    10.9967   0.6598 0.5104150    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

FIGURE 9.6 R Output of t-tests for Equation 9.5 based on White robust standard errors

The estimates for the parameters in the model are the same in Figures 9.3 and 9.6. This is because autocorrelation and heteroscedasticity do not affect the estimates themselves, but their standard error. This in turn, can lead to wrong conclusions about the significance of the corresponding estimates. Figure 9.6 shows that even after correcting for the presence of autocorrelation and heteroscedasticity, the intercept, the coefficients for Price_t, and Price_C1_t have p-values that are smaller than 0.05. This does not hold for Price_C2_t. Hence, we conclude that except for the coefficient of the price of the second competitor, all estimates for the parameters in Equation 9.5 deviate significantly from zero.

The normality assumption (assumption 4) can be investigated visually with the Q-Q plot of the residuals, see Figure 9.7.

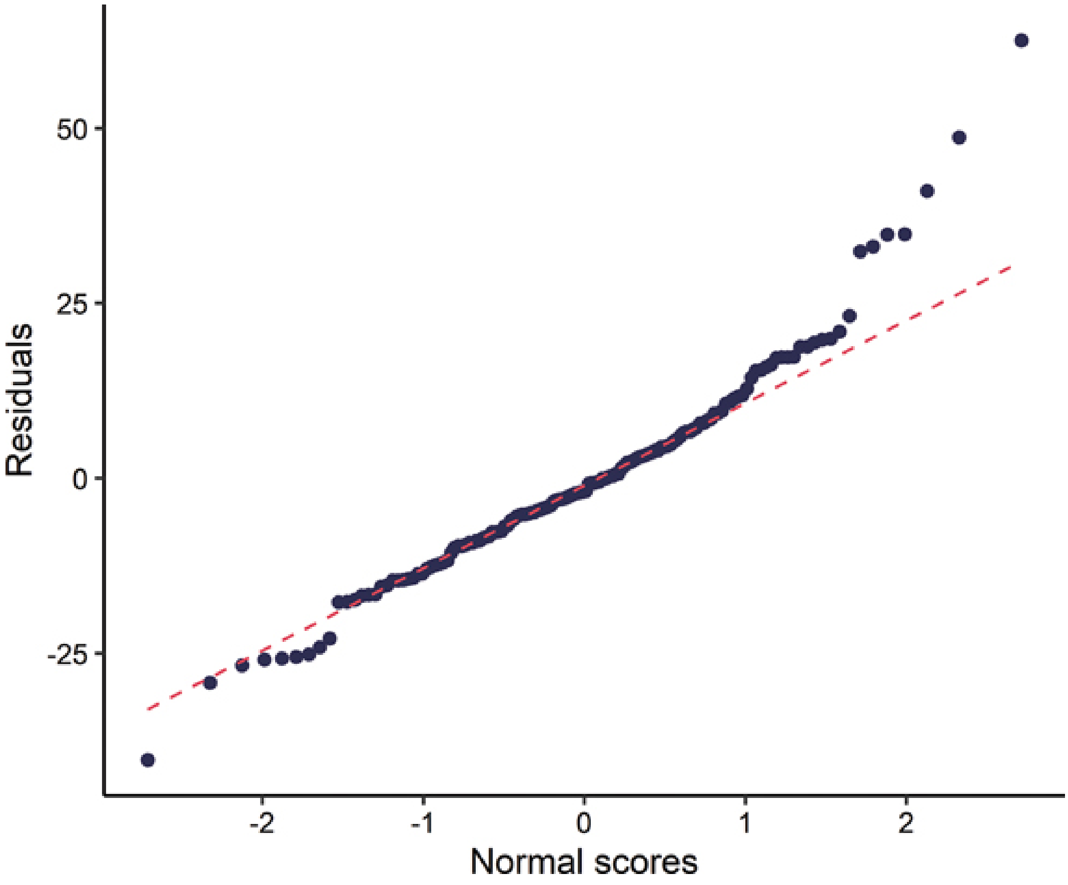


FIGURE 9.7 Q-Q plot of the residuals of Equation 9.5

As the Q-Q plot in Figure 9.6 illustrates, deviations from normality occur mostly at the right tail of the distribution of the residuals, and to a lesser extent at the left tail. Note that the large positive outliers could already have been observed in Figure 9.4. More formal normality tests confirm that the normality assumption for the residuals is indeed violated (see the R script that accompanies this chapter). Consequently, we cannot trust the *p*-values in the default output of OLS, because their calculation is based on the normality assumption. Figure 9.7 suggests that the deviation from normality is likely due to outliers in the residuals. Hence, we turn to robust estimation to find out whether the outliers affect our estimation results. The R script that accompanies this chapter shows that the conclusions regarding the significance of the parameters remain unchanged. As an alternative approach for determining appropriate *p*-values we could have resorted to bootstrapping.

To investigate potential issues with multicollinearity (violation of the fifth assumption of OLS), we can inspect a correlation matrix of the independent variables in Equation 9.5 (see Figure 9.8).

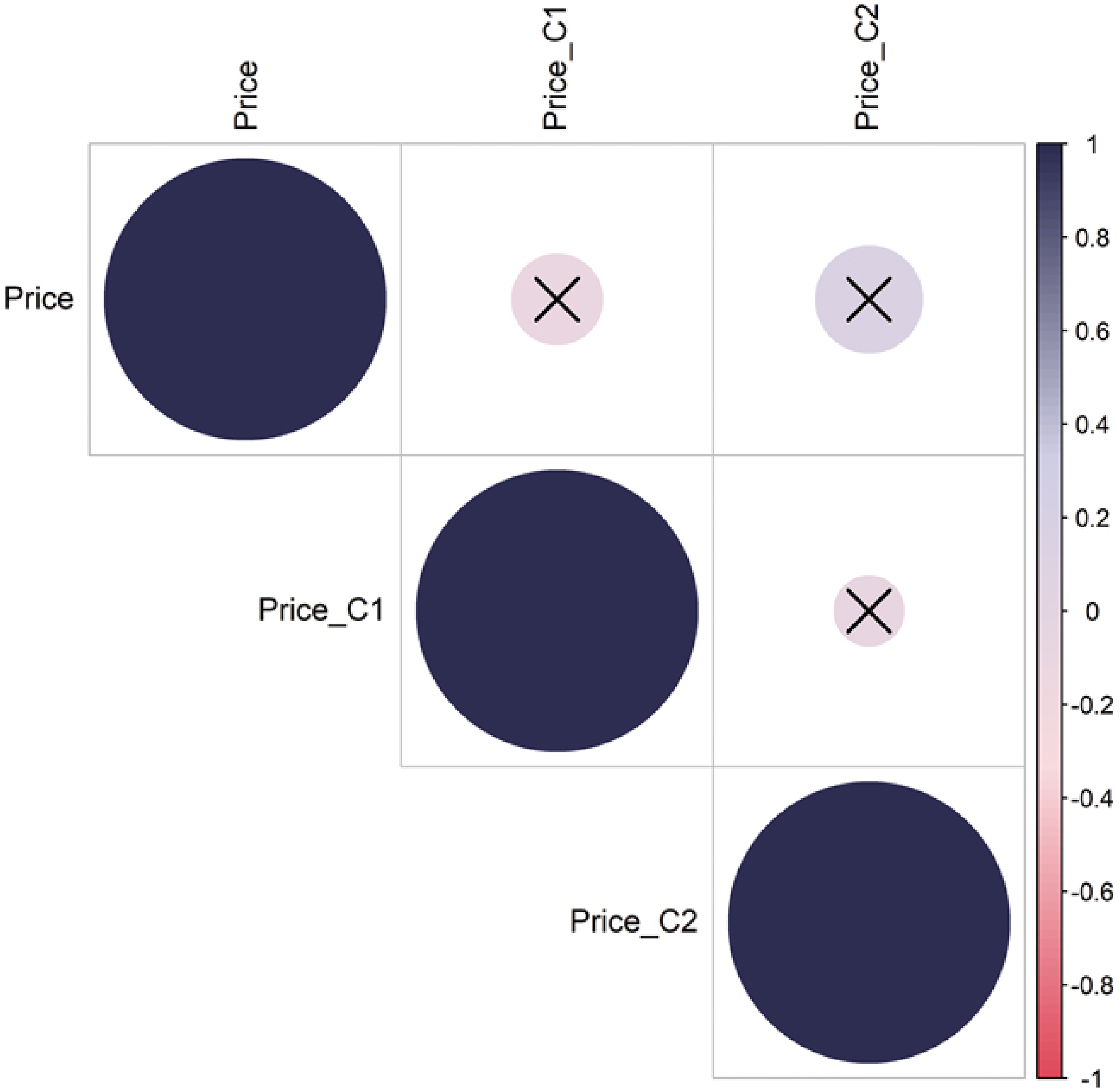
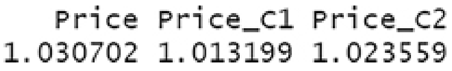


FIGURE 9.8 Correlation matrix of the independent variables in Equation 9.5

Figure 9.8 indicates that Price$_t$ and Price_C1$_t$ are slightly negatively correlated, whereas Price$_t$ and Price_C2$_t$ are slightly positively correlated, and Price_C1$_t$ and Price_C2$_t$ are weakly negatively correlated. However, the crosses indicate that none of these correlations reach significance. This signals that multicollinearity is not a major concern when estimating Equation 9.5. When multicollinearity would have been an issue, this would have inflated the variance of the estimated coefficients. So-called variance inflation factors (VIFs) indicate to what extent the variances are affected by correlations among the independent variables. Figure 9.9 presents the VIF values for the independent variables in Equation 9.5.



```
   Price Price_C1 Price_C2
1.030702 1.013199 1.023559
```

FIGURE 9.9 VIF values of the independent variables in Equation 9.5

VIF values that exceed five signal the presence of serious levels of multicollinearity, as they indicate that the variances of the corresponding estimates are inflated by a factor of more than five. The VIF values in Figure 9.9 do not exceed this threshold and are all close to the ideal value of one. We conclude that the results in Figure 9.3 are not affected by multicollinearity.

In the marketing science literature, the last OLS assumption (exogeneity) receives much attention (e.g., Rossi, 2014). Violations of this assumption could occur in the canned tuna example if the supermarket owner based the price of canned tuna on the expected effects and sales levels. Endogeneity can also be an important reason for concern in models for cross-sectional data, for example, due to self-selection. For example, customers become a member of a loyalty program because they believe they will buy more in the future (e.g., Leenheer *et al*., 2007). Endogeneity is problematic because it leads to biases in the parameter estimates. Consequently, for predictive purposes endogeneity seems less of an issue than for more descriptive purposes of a model (Ebbes, Papies & Van Heerde, 2011). We refer to Papies, Ebbes and Van Heerde (2017) for an excellent discussion of endogeneity issues.

Based on the foregoing, we can conclude that the own price and the price of the first competitor significantly affect the sales of the focal canned tuna brand. The intercept is also significantly different from zero.

Another important aspect of statistical validity of a linear model concerns its fit or explanatory power. The most important measure for this is R^2, which can have values between zero and one. An R^2 value of zero indicates that the model has no explanatory power and a value of one indicates a perfect fit of the model to the data. The R^2 measure is a relative measure that indicates what portion of the variation in the dependent variable is explained by the model. Its value depends on how well the regression line fits the data and the amount of dispersion in the values of the dependent variable. Although higher R^2 values are preferred because they indicate a higher fit between the model and the data, there are no widely accepted thresholds for "a good R^2." This is due to the fact that this strongly depends on the context. For some data, achieving an R^2 that barely exceeds 0.10 is a major achievement, whereas in other situations, even very simple models lead to R^2s that surpass 0.95. High values can be achieved with wrongly specified models, for example when part of the y-values is included in the independent variables (e.g., prices are set based on expected sales). Moreover, some data have such a large dispersion that it is just difficult to explain. High values can also be achieved by using more explanatory variables. To account for this effect, the so-called "adjusted R^2" can be calculated, which will be lower than the value of R^2, because a penalty is added.

The R^2 of the model for canned tuna sales equals 0.6883, which indicates that the model explains about 69% of the variation in unit sales of canned tuna.

The adjusted R^2 is slightly lower, which is due to the penalty that is added for the inclusion of each variable. The F-test at the bottom of Figure 9.3 has a p-value that is smaller than zero, which indicates that the model as a whole is significant.

After investigating the statistical validity of the outcomes in Figure 9.3, we now discuss face validity of the outcomes. As in many applications, interpreting the estimated value for the intercept is difficult, as it represents the expected value of the dependent variable when all independent variables are zero. In the tuna case example, the value of 108.134 should then be interpreted as the expected unit sales when the prices of all three brands are zero. Of course, this is not a very realistic situation, and the data is certainly not informative about this extreme case. Hence, in such cases, we refrain from interpreting the estimated value of the intercept, and just note that it is significantly different from zero.

Concerning the estimates for the coefficients of Price_t and Price_C1_t we note that these are both significant and that the signs are as expected. We expect a negative sign for own price, as sales are likely to go up when the price drops. For the price of competitor 1, the positive sign of the coefficient is also in line with what we would expect: when competitors raise their price, we expect our sales to increase. As the p-value associated with the price of competitor 2 is larger than 0.05, we conclude that the estimate for the corresponding parameter is statistically indistinguishable from zero. Hence, we do not interpret the coefficient for Price_C2_t that is shown in Figure 9.3 and conclude that competitor 2 does not significantly affect the sales level of the focal brand.

Finding a non-significant result is sometimes perceived as a disappointing outcome. But as the canned tuna example shows, this can have important consequences. Based on our analyses we conclude that brand 1 is an influential competitor for the focal brand, whereas we could not establish a significant effect for competitor 2. Such findings provide relevant input when developing a competitive strategy.

Regarding the size of the coefficients, we note that the own-price effect is stronger than the effect of competitor 1, which is in line with earlier findings (e.g., Horváth *et al.*, 2005) so that not only the signs of the estimated coefficients but also their sizes have high face validity. When the canned tuna brand lowers its price by ten cents, sales are expected to increase by 11.24 units. When competitor 1 lowers its price by 10 cents, sales of the focal brand are expected to decrease by 5.99 units.

The last element of the validation of the outcomes of the model in Equation 9.5 concerns predictive validity. To this end, the supermarket owner decided to collect 39 more weeks of data for all variables. With the model that was estimated on the 150 data points that she had been considering so far, she predicted unit sales for the new set of observations. The resulting predictions are presented in Figure 9.10.

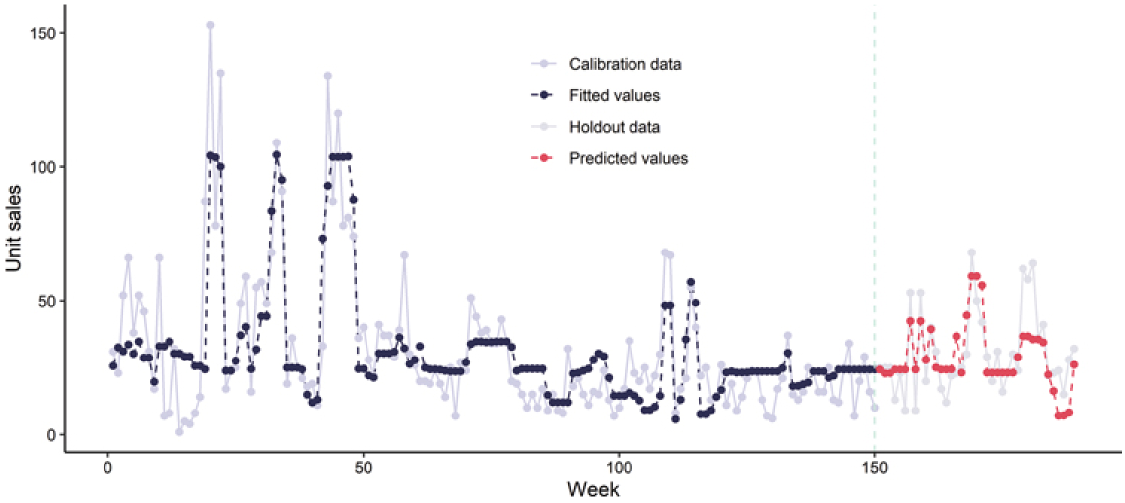


FIGURE 9.10 Fitted and predicted values for Equation 9.5

From Figure 9.10 we conclude that the model captures the variation in unit sales fairly well, both for the calibration data and for the holdout data. To compare predictive validity among various versions of the same model, several prediction metrics can be calculated:

- Average Prediction Error (APE), which is the simple average of the prediction residuals. This can be used to assess whether the model is over or under predicting in the holdout sample. If a formal test for this is desired, a one-sample t-test can be used to test APE against the value of zero. The APE for the canned tuna case equals 1.56, but this value does not significantly differ from zero.
- Average Squared Prediction Error (ASPE), also known as Mean Squared Error (MSE) determines the average of the squared prediction residuals. ASPE equals 125.62 for the canned tuna case. If the square root is taken from ASPE, the resulting value can be compared to the standard deviation of the residuals of the calibration sample. This can be found in the R output of Figure 9.3 (the “Residual standard error”) and equals 15.02. Since OLS minimizes the sum of squared residuals of the calibration sample, it is to be expected that RASPE is larger than standard deviation of the residuals of the calibration sample. That is surprisingly not the case for the canned tuna model, as RASPE is 11.21. This indicates that the predictions in the holdout sample are slightly more accurate than the fits in the calibration sample, which can happen occasionally.
- Mean Absolute Percentage Error (MAPE) is calculated as the mean of absolute values of the relative errors, where the prediction errors are divided by the true value. Compared to the other predictive validity metrics, MAPE has the attractive property that it is a relative measure that

does not depend on the measurement scale of the dependent variable. This makes MAPE a useful metric to compare a model's predictive validity across applications. This is less useful for the other metrics, and we advise to only compare them across different versions of the same model. MAPE for the canned tuna case equals 36.5%.

9.4 LOGISTIC REGRESSION

For many projects in customer management and branding research, a deeper understanding of the choices of individual customers is required. Consider for example the following questions:

- Will the customer purchase your product?
- Is the customer going to cancel her insurance?
- Will the customer respond to an offer via direct mail?
- Is the claim filed by a customer fraudulent?
- Will the customer adopt the new product that you are launching?
- Is the customer going to pay his next bill?

A common characteristic of these questions is that they have two possible outcomes: Yes or No. Consequently, many data science projects that involve decisions of individual customers have binary KPIs.

Analysts sometimes use a linear model that we discussed in the previous section to analyze data that can predict or explain these KPIs. The ordinary linear model, however, assumes that the dependent variable is continuous and therefore it is not perfectly suited for the prediction of these binary events.1 In our discussion of investigating one-to-one relationships in Section 8.3, we mentioned logistic regression as a technique to investigate the relationship between a binary KPI and one numerical driver. In this section, we extend that discussion to the situation where we want to study the effect of multiple independent variables on a binary KPI.

Let us code the outcome that we are interested in as 1, and the other outcome as 0. Logistic regression is concerned with modeling p, the probability that the dependent variable equals 1, as a function of a number of independent variables. However, similar to the argument that regression is not suited for a binary dependent variable, we cannot use the default linear model for p because there is no built-in restriction that its output values are bounded by zero and one. To overcome this issue, p is not modeled in logistic regression, but rather a transformation of p:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i,$$

where the index i indicates customer i. The right-hand side of (9.6) Equation 9.6 very much resembles the right-hand side of an ordinary linear regression model. The transformation that is applied to p at the left-hand side of the equation is a so-called logit transformation; this is also reflected in the name of the technique. If the values for the independent variables of customer i are known, the corresponding probability of observing the outcome that we are interested in can be calculated as:

$$\widehat{p}_i = \frac{1}{1 + e^{-(\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki})}}.$$

Equation 9.7 shows that the relationship between p and an (9.7) independent variable is no longer linear, but S-shaped. Figure 9.11 depicts the relationship between p an independent variable.

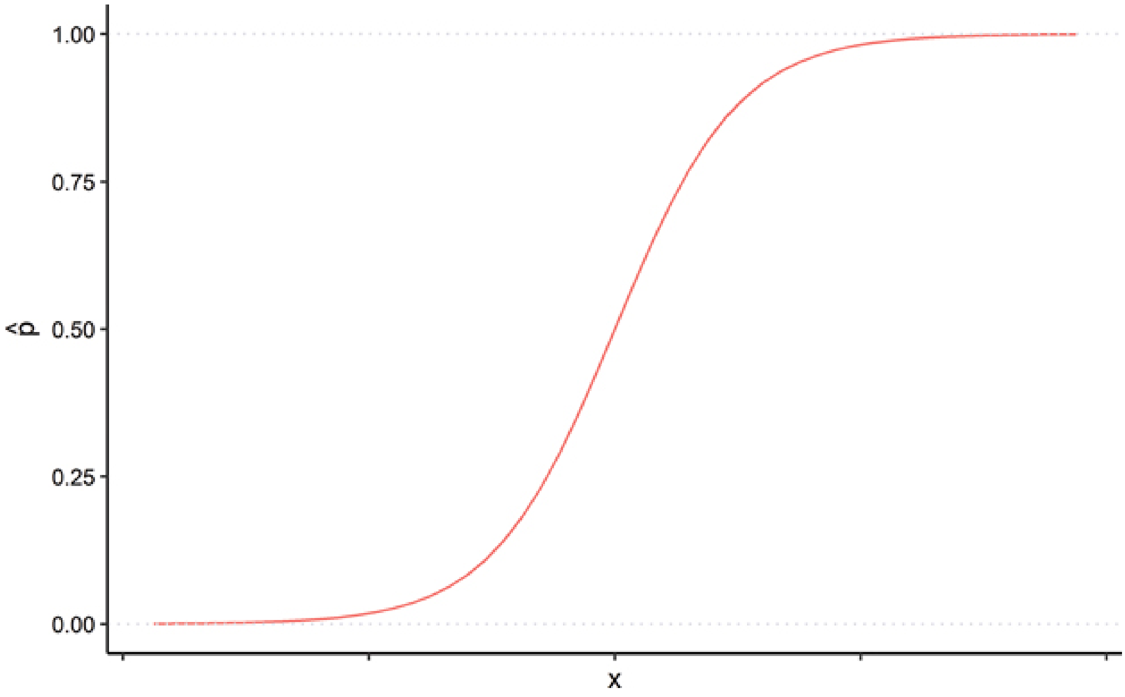


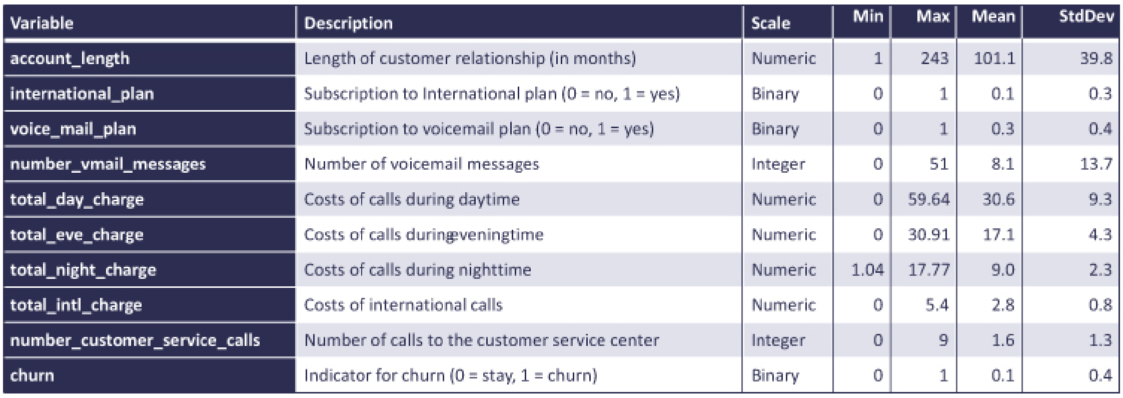
FIGURE 9.11 Relationship between an independent variable x and $\widehat{p}$

Figure 9.11 illustrates that, despite the resemblance between Equation (9.7) and a linear regression model, the relationships in a logistic regression are no longer linear. As we will see below, that also has consequences for the interpretation of the parameters estimates. They no longer represent constant slopes, as the slope is constantly changing in Figure 9.11.

Logistic regression is one of the most commonly used models to predict churn; together with decision trees (see Section 9.6) they accounted for 68% of the entries of a churn modeling contest in which both practitioners and academics participated (Neslin *et al.*, 2006). Neslin *et al.* (2006) concluded that logistic regression and tree approaches perform best in terms of prediction and that the differences between the models in predictive accuracy are “managerially meaningful.”

We will illustrate the steps of building a logistic regression model for churn of a telecom provider.

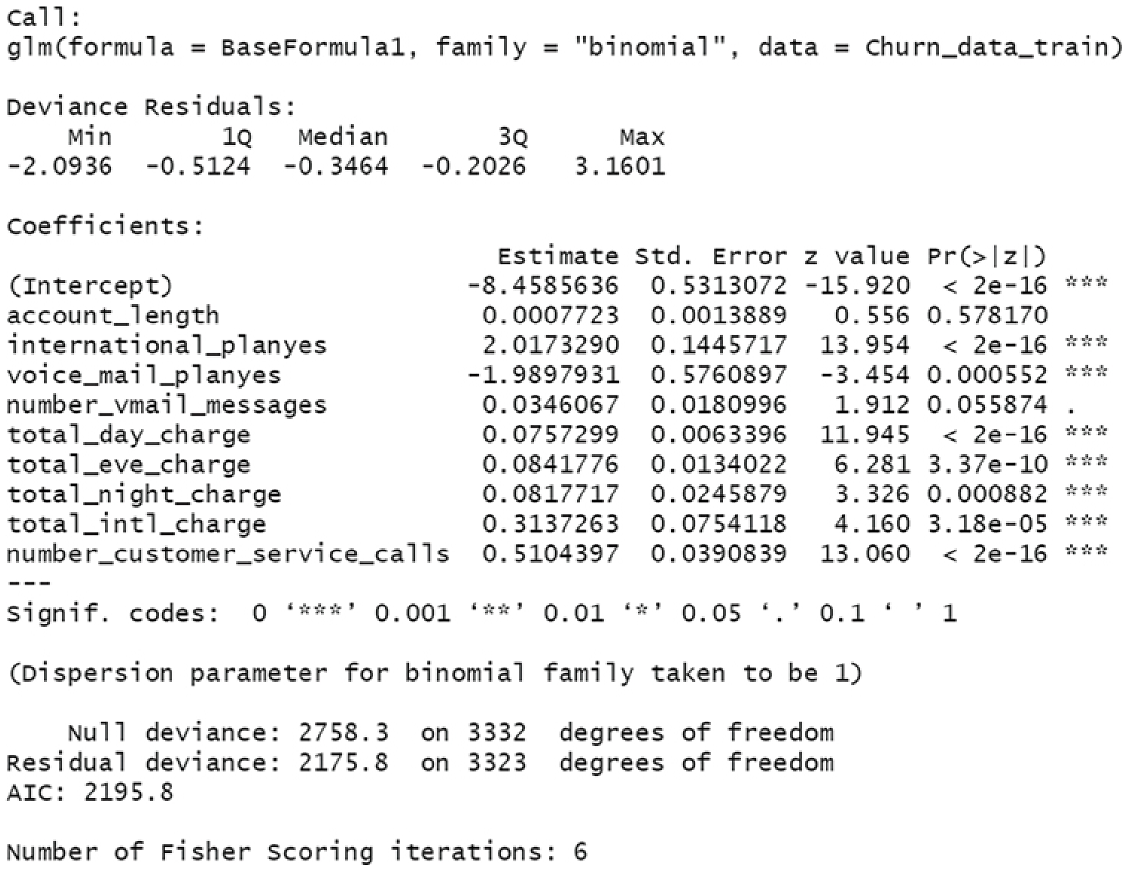
Suppose that during step 1 of building a churn model (the specification step) the telecom provider has selected the first nine variables that are listed in Figure 9.12 as independent variables to explain churn, the last variable in Figure 9.12.



Variable	Description	Scale	Min	Max	Mean	StdDev
account_length	Length of customer relationship (in months)	Numeric	1	243	101.1	39.8
international_plan	Subscription to International plan (0 = no, 1 = yes)	Binary	0	1	0.1	0.3
voice_mail_plan	Subscription to voicemail plan (0 = no, 1 = yes)	Binary	0	1	0.3	0.4
number_vmail_messages	Number of voicemail messages	Integer	0	51	8.1	13.7
total_day_charge	Costs of calls during daytime	Numeric	0	59.64	30.6	9.3
total_eve_charge	Costs of calls duringeveningtime	Numeric	0	30.91	17.1	4.3
total_night_charge	Costs of calls during nighttime	Numeric	1.04	17.77	9.0	2.3
total_intl_charge	Costs of international calls	Numeric	0	5.4	2.8	0.8
number_customer_service_calls	Number of calls to the customer service center	Integer	0	9	1.6	1.3
churn	Indicator for churn (0 = stay, 1 = churn)	Binary	0	1	0.1	0.4

FIGURE 9.12 Variables available for analyzing churn

The second step involves estimating the logistic regression equation. Since we are modeling the probability of churn in this case, we are not aiming to predict the binary outcomes of the y variable. Consequently, calculating residuals is not as straightforward as in ordinary regression, and the estimation method does not rely on minimizing the sum of squared residuals. Instead, the maximum likelihood principle is commonly used to obtain parameter estimates for each of the independent variables. The R output of estimating the churn model is presented in Figure 9.13.



```
Call:
glm(formula = BaseFormula1, family = "binomial", data = Churn_data_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0936  -0.5124  -0.3464  -0.2026   3.1601

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -8.4585636  0.5313072 -15.920  < 2e-16 ***
account_length                0.0007723  0.0013889   0.556 0.578170
international_planyes         2.0173290  0.1445717  13.954  < 2e-16 ***
voice_mail_planyes           -1.9897931  0.5760897  -3.454 0.000552 ***
number_vmail_messages         0.0346067  0.0180996   1.912 0.055874 .
total_day_charge              0.0757299  0.0063396  11.945  < 2e-16 ***
total_eve_charge              0.0841776  0.0134022   6.281 3.37e-10 ***
total_night_charge            0.0817717  0.0245879   3.326 0.000882 ***
total_intl_charge             0.3137263  0.0754118   4.160 3.18e-05 ***
number_customer_service_calls 0.5104397  0.0390839  13.060  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2758.3  on 3332  degrees of freedom
Residual deviance: 2175.8  on 3323  degrees of freedom
AIC: 2195.8

Number of Fisher Scoring iterations: 6
```

FIGURE 9.13 R output of estimating a logistic regression

Because of the similarities between ordinary regression and linear regression, the output in Figure 9.13 resembles the output in Figure 9.3. The central part of both figures shows us estimates of the coefficients for every independent variable, as well as tests to assess the significance of each variable.

Turning to step 3 of the model building process, we note that due to the issues with calculating residuals, it is less common to validate assumptions regarding the error term when estimating a logistic regression model. Other aspects that do not involve the error term, such as the fit of the model and multicollinearity, proceed similarly to what we have already seen in the previous section. Here we focus on the main differences compared to validating an ordinary regression model.

The deviance values printed at the bottom of Figure 9.13 can be used to conduct an overall test of the significance of the model. The “deviance” can be interpreted as a measure for the fit of the model, where a lower value indicates a higher fit. The null deviance is determined for a version of the model where all independent variables are excluded so that only the intercept is estimated. Fortunately, the residual deviance, which is calculated with all variables included in the model estimation, is lower. The difference between the two deviances is the test statistic for an overall test of a logistic regression model and follows a Chi-square distribution with degrees of freedom equal to the number of independent variables in the model. To test the overall significance of our logistic regression equation, we utilize a Chi-square distribution with 9 degrees of freedom to determine that the p -value corresponding to a difference of 582.52 in the deviances equals 0.000. Because this is lower than 0.05, we conclude that the churn model as a whole is significant.

The AIC value printed at the bottom of Figure 9.13 is related to the deviance measures but adds a penalty for each additional variable that is included in the model, much like the adjusted R^2 in ordinary regression. The value of Akaike's Information Criterion (AIC) is hard to interpret, but it can be useful for comparing different models for the same data: the model with the lowest AIC is the preferred model.

The output in Figure 9.13 does not report a value for R^2, which is also due to the absence of residuals as in ordinary regression. However, several statisticians have developed approximations to R^2 that can be interpreted in a similar fashion to what happens in ordinary regression. In the R script that accompanies this chapter, several of those pseudo-R^2s are calculated. McFadden's pseudo-R^2 equals 21.1%, Cox and Snell's pseudo-R^2 is 16.0%, and Nagelkerke's pseudo-R^2 equals 28.5%.

Turning to the individual variables in Figure 9.13, we see that all variables contribute significantly to churn, except for account_length and number_vmail_messages (their p -values exceed 0.05). Most of the significant coefficients are positive. Hence, we conclude that the churn probability generally increases when customers have an international plan, a higher number of voicemail messages, higher costs for calling during the day, evening, and night, higher costs for international calls, and more frequent interactions with the customer service desk. Only one coefficient is significantly negative. Consequently, we conclude that customers with a voicemail plan generally have a lower churn probability than customers without a voicemail plan. Note that for assessing the face validity of these outcomes, we can only interpret the sign of the coefficients, not the size. This is due to the fact that we cannot interpret the estimated parameter values as constant slopes as in ordinary regression.

Another representation of the effect of the independent variables on p in a logistic regression is to calculate e^{β} for each independent variable. The number that results for each independent variable can be interpreted as the factor by which the odds ratio changes when the value of the independent variable increases with one unit, where the odds ratio is defined as $p/(1-p)$. Figure 9.14 shows these factors for our churn case, along with their 95% confidence intervals.

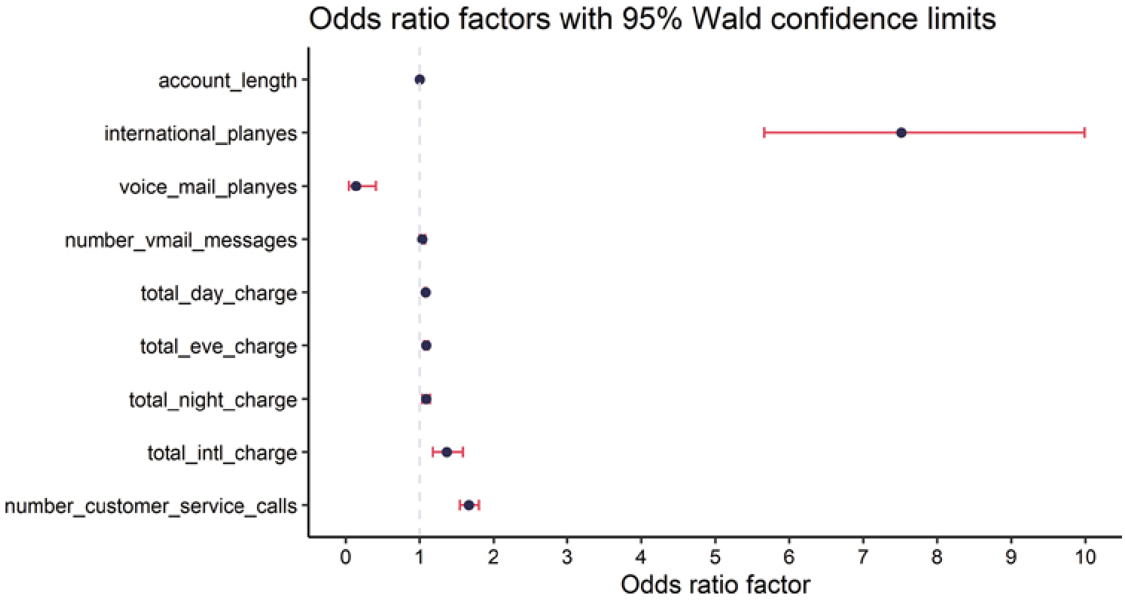


FIGURE 9.14 Odds ratio factors for the churn case

Since the values that are depicted in Figure 9.14 are factors, we compare them against a value of one. Odds ratio factors lower than one indicate a negative relationship between the corresponding independent variable and p, whereas values larger than one indicate a positive relationship.

Turning to an assessment of predictive validity, we use the estimated churn model to predict churn probabilities for a sample of holdout customers whose data was not used to calibrate the model. Comparing the churn probabilities to the observed switching behavior, we can evaluate the predictive validity of our churn model.

The literature proposes several metrics that can be used to assess the predictive performance of logistic models. The most important ones are:

- Hit rate
- Top-decile lift
- Gini coefficient.

9.4.1 Hit rate

The hit rate is defined as the percentage of customers that are correctly classified by the churn model. To that end, the predicted churn probabilities are used to classify customers with an estimated churn probability of 50% and higher as “churners,” and all customers with an estimated churn probability as “stayers.” The hit rate is then calculated as:

$$\text{hit rate} = \frac{\text{number of customers correctly classified}}{\text{number of customers}} {*}100\,\% \,.$$

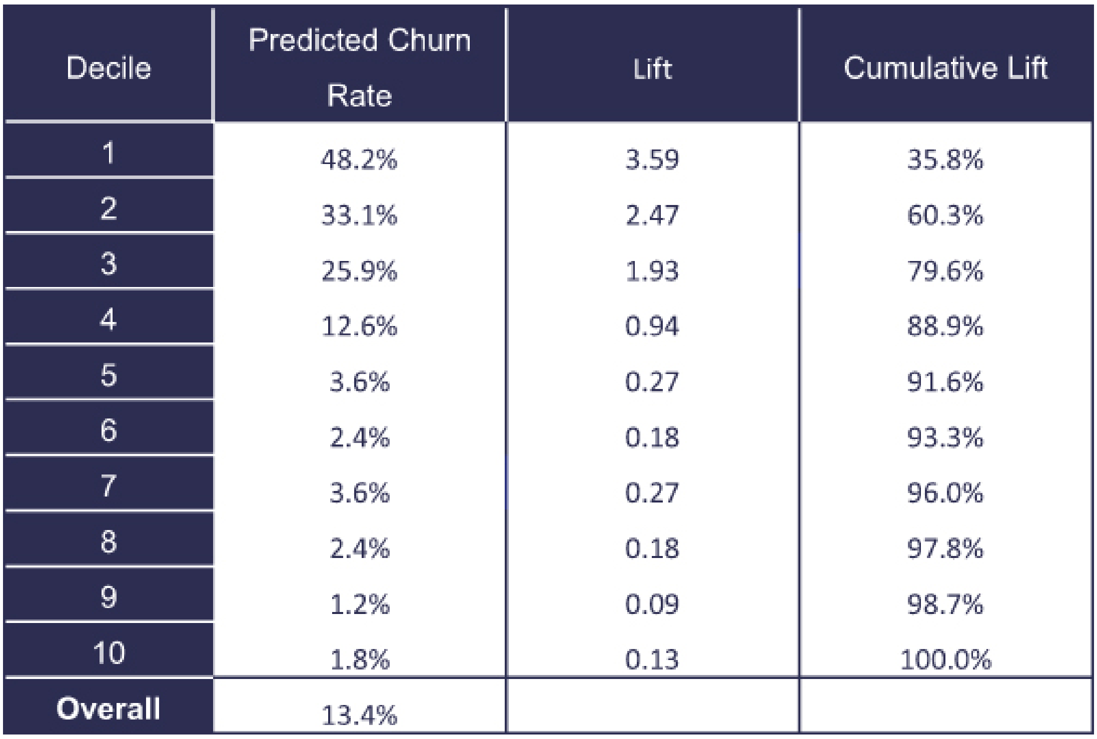
(9.8) For the churn model that we developed in this section, the hit rate equals 87%. The hit rate is a fairly easy metric to interpret. An important problem, however, is that the value of the hit rate depends on the ratio of zeros and ones in the data (in our example this is the fraction of churners). In cases where a rare event is modeled, this ratio is small, and it is relatively easy to obtain a good hit rate, by classifying all customers in the majority class. As a consequence, competing models should not be compared on hit rates alone.

9.4.2 Top-decile lift

"Lift" is the most common measure for model performance (Blattberg, Kim, & Neslin, 2008). Intuitively, lift measures to what extent the model can distinguish between the two classes that are modeled in a logistic regression. In our example, these are the churners and the non-churners. Values for lift are commonly determined per decile, where customers in the holdout sample are assigned to a decile based on their predicted (e.g. churn) probability that is calculated using the estimated logistic regression model. For each decile, the true fraction of ones is then calculated, and divided by the overall fraction of ones in the holdout sample:

$$\text{lift decile } j = \frac{\text{fraction of ones in decile } j}{\text{fraction of ones in entire holdout sample}}.$$

(9.9) In Figure 9.15 we calculate the lifts per decile for our churn model. The top-decile lift is the lift in decile 1. The top-decile lift is substantial, as the true churn rate in decile 1 is 48.2%, while the average churn rate is only 13.4% in the entire holdout sample. Based on the calculations of the lifts in each decile, the cumulative lift can be computed. The cumulative lift of the kth decile is defined by the percentage of all churners accounted for by the first k deciles (Blattberg, Kim, & Neslin, 2008: 319). In Figure 9.15 the top two deciles account for 60.3% of all churners in the holdout sample. The higher the cumulative lift in the lower deciles, the better the model.



Decile	Predicted Churn Rate	Lift	Cumulative Lift
1	48.2%	3.59	35.8%
2	33.1%	2.47	60.3%
3	25.9%	1.93	79.6%
4	12.6%	0.94	88.9%
5	3.6%	0.27	91.6%
6	2.4%	0.18	93.3%
7	3.6%	0.27	96.0%
8	2.4%	0.18	97.8%
9	1.2%	0.09	98.7%
10	1.8%	0.13	100.0%
Overall	13.4%		

FIGURE 9.15 Lift and cumulative lift per decile for the churn model

9.4.3 Gini coefficient

The Gini coefficient is related to the lift model performance measures, in that it is essentially the area between the model's cumulative lift curve and the lift curve that would result from random prediction. In a cumulative lift curve (or gains chart) the percentage of customers, ordered by the predicted probability of the outcome that we are interested in (churn in our example), is plotted along the x-axis, and the cumulative percentage of customers with the outcome that we are interested in is plotted along the y-axis (compare with cumulative lift). The cumulative lift curve for our churn case is shown in Figure 9.16. The dashed blue line represents a model with random predictions (no information). A larger distance between the curve and the dashed line means a stronger model performance.

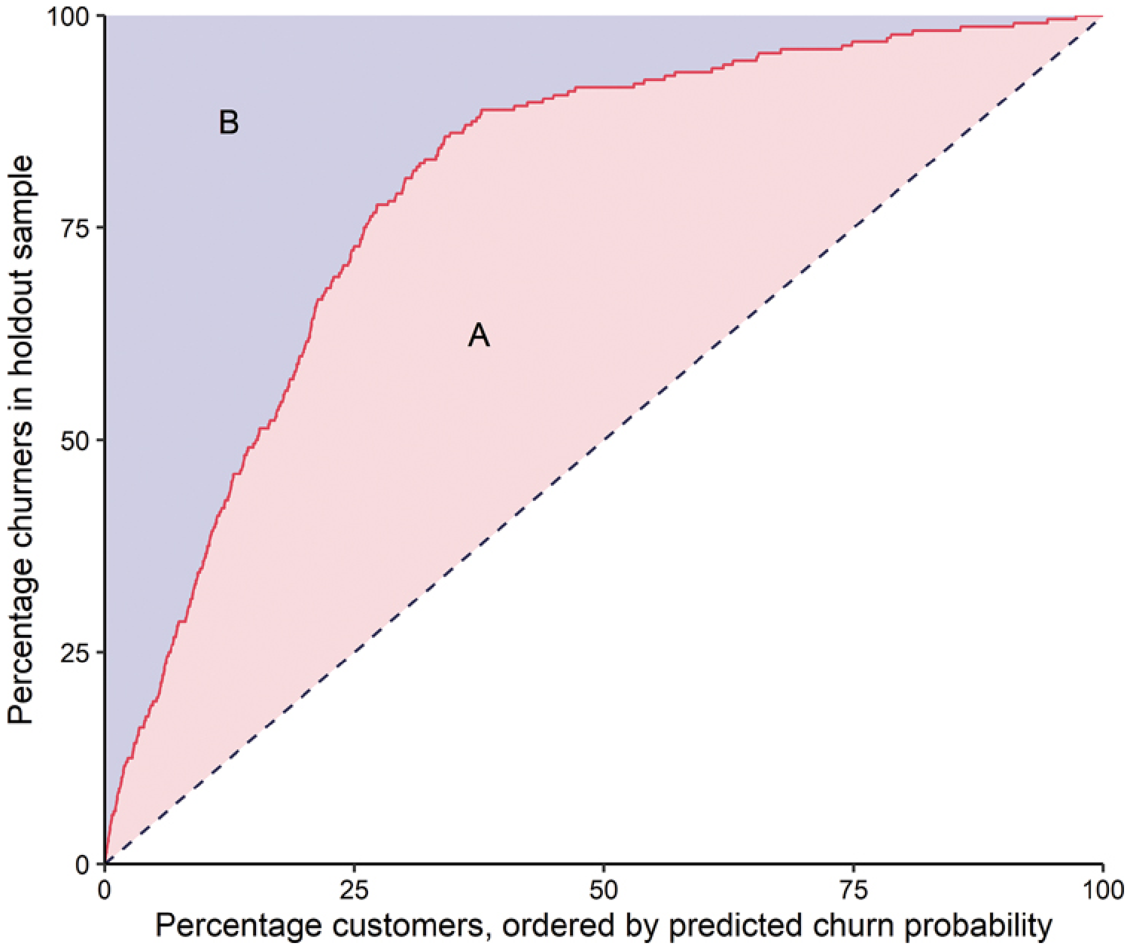


FIGURE 9.16 Cumulative lift curve for the churn model

The Gini coefficient can be derived from the cumulative lift curve in Figure 9.16. It is calculated by dividing A by the area above the dashed blue line (A+B). For the churn case, the Gini coefficient equals 0.71. By construction, the Gini coefficient can only attain values between 0 and 1. A value of 0 implies that the model predicts just as well as a random predictor. If the Gini coefficient equals 1, this is most likely due to the fact that there is only one customer and that this customer is classified as the customer with the highest probability of the outcome of interest. This is a very rare case. When comparing models, those with higher Gini coefficients are more preferred.

Before we turn to data-driven modeling, we note that in cases where rare events are modelled (such as conversion or churn), one of the two values of the dependent variable is typically under-represented in the data. This is known as class imbalance, and potentially leads to low prediction power, especially of the minority segment (Donkers, Franses, & Verhoef, 2003; Estabrooks, Jo, & Japkowicz, 2004; Marinakos & Daskalaki, 2017; Marqués, García, & Sánchez, 2013). In such cases, it is advisable to reduce the imbalance, but full class balance is not *per se* necessary and may not always be optimal. More balanced samples can be obtained in in different ways:

- Random under-sampling of the majority class: randomly select customers from the majority class until both classes have (about) the same number of observations
- Random over-sampling of the minority class: units are repeatedly sampled with replacement from the minority class until both classes have (about) the same number of observations
- Synthetic Minority Over-sampling TEchnique (SMOTE) is conceptually very similar to random oversampling but interpolates between units in the minority class to simulate “similar customers,” which results in no or fewer repeated observations (Chawla *et al.*, 2002).

In Section 9.6, under-sampling is illustrated in a direct-marketing example (see Figure 9.18).

9.5 DATA-DRIVEN MODELING AND MACHINE LEARNING

In Sections 9.2–9.4 we discussed theory-driven modeling, where we focused on linear regression and logistic regression. In those sections, the theory was guiding the specification of the model and data were used to calibrate and validate the model. In this section, we introduce data-driven modeling, where the entire process is more data-driven. Specifically, in contrast to theory-driven modeling, the characteristics of the data also guide the specification step in data-driven modeling.

The techniques that we will discuss in the upcoming sections[2] are commonly classified as machine learning methods. However, this may suggest an undue distinction from more traditional econometrical methods, some of which were discussed in the first part of this chapter. Machine learning is a catch-all collection of techniques and shares common methodologies with traditional econometrics such as (logistic) regression analysis, resampling, classification, and non-linear methods. However, the two fields have different origins (Kübler, Wieringa, & Pauwels, 2017). Traditional econometrical and statistical methods are largely applied in social sciences, economics and other related areas, whereas a large body of machine learning techniques were developed in computer sciences. As a consequence, the research philosophies of the two fields differ. Wasserman (2012) argues that the traditional methodology emphasizes formal statistical inferences (i.e., hypothesis testing, confidence intervals, etc.), whereas machine learning is more outcome-oriented and focuses on accurate prediction making. Therefore, the machine learning process differs from the model building steps that we discussed in Section 9.2.

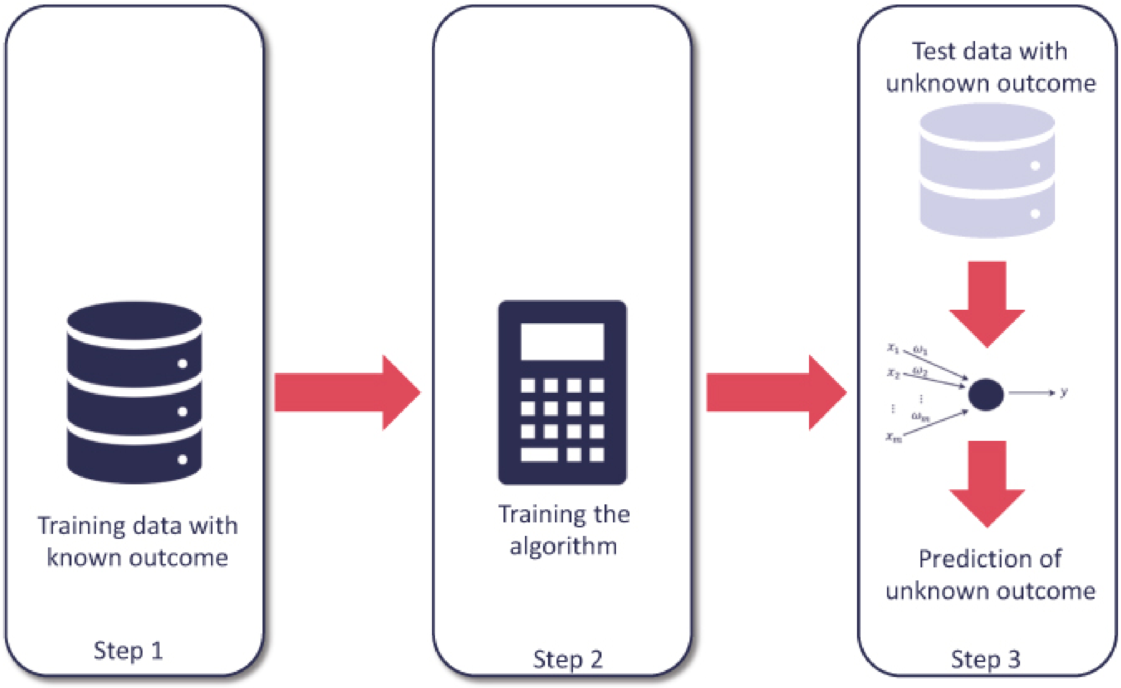

Training data with known outcome
Step 1
Training the algorithm
Step 2
Test data with unknown outcome
x_1
ω_1
x_2
ω_2
ω_m
x_m
y
Prediction of unknown outcome
Step 3

FIGURE 9.17 The supervised machine learning process

In Section 8.6 we discussed a number of unsupervised machine learning techniques and explained that they are applied in cases where there is no dependent variable that should be explained or predicted. In line with the earlier sections in this chapter, supervised machine learning methods are applied in the context where there is a dependent variable that needs to predicted or classified. Figure 9.17 depicts the steps of building a supervised machine learning model where data with known values for the dependent variable (in machine learning terminology: the output variable) are used to calibrate the system (or "train the algorithm"). In the final step, new data enters the system, and the "trained algorithm" predicts values for the output variable. The machine learning process is therefore very close to investigating predictive validity as discussed in the earlier sections in this chapter, but the terminology differs. The calibration data that we discussed earlier is known as training data in machine learning terminology and the holdout sample is commonly denoted as the test set. In addition, computer scientists also use alternative terminology for independent variables (inputs or features) and dependent variables (outputs). We conclude that, despite the different philosophies and terminology, traditional econometrics and machine learning have much in common and have similar DNA.

9.6 DECISION TREES

We start the discussion on data-driven modeling with decision trees, also known as tree models, which is a collective name for techniques based on flowcharts. A very attractive property of decision trees is that the flowchart provides a visual representation of decision rules that relate a KPI to one or more independent variables. Decision trees were introduced by Breiman *et al.* (1984), which makes them younger than (logistic) regression. Some of the methods to grow a decision tree are quite computer intensive and the increase in their popularity is partly due to computing power that has significantly increased since then.

But there are also several other reasons why the number of applications of decision trees is growing. They are widely available in statistical software, which makes it relatively straightforward to create a decision tree. The fact that the tree provides a graphical representation of the relations between independent variables and a KPI makes its output visually appealing and easy to explain to a non-technical audience. For example, outcomes of decision tree analyses on customer churn are often presented to boardroom members. Moreover, tree models have good predictive performance, especially in large data sets with many categorical variables, a setting that is quite common in customer data analyses. Tree models are also very flexible. We will see that they accommodate nonlinearity and interactions by default and do not require including additional terms to model such relationships.

Decision trees have a broad range of application areas in business. They are used in the field of operations research to model process-based environments, e.g., hospital logistics. Another area where decision trees play a major role is in the field of machine learning. Because of their data-driven nature and excellent performance, we see many very successful applications in predictive analytics. In marketing, their popularity has also increased over the years. There are many applications where decision trees aid the accurate classification of customers by credit risk, or predict churn.

Decision trees can be used in a broad range of statistical techniques. Based on the found associations between one independent binary variable (e.g., response) and usually multiple predictors, a tree structure with different customer segments can be drawn. In a statistical sense, one could argue that this technique involves both segmentation and prediction. However, in customer value management (CVM) practice it is mainly used for predictive purposes. Several names for these techniques are found in literature and software programs such as CHAID, Cart, and Answer Tree can be used.

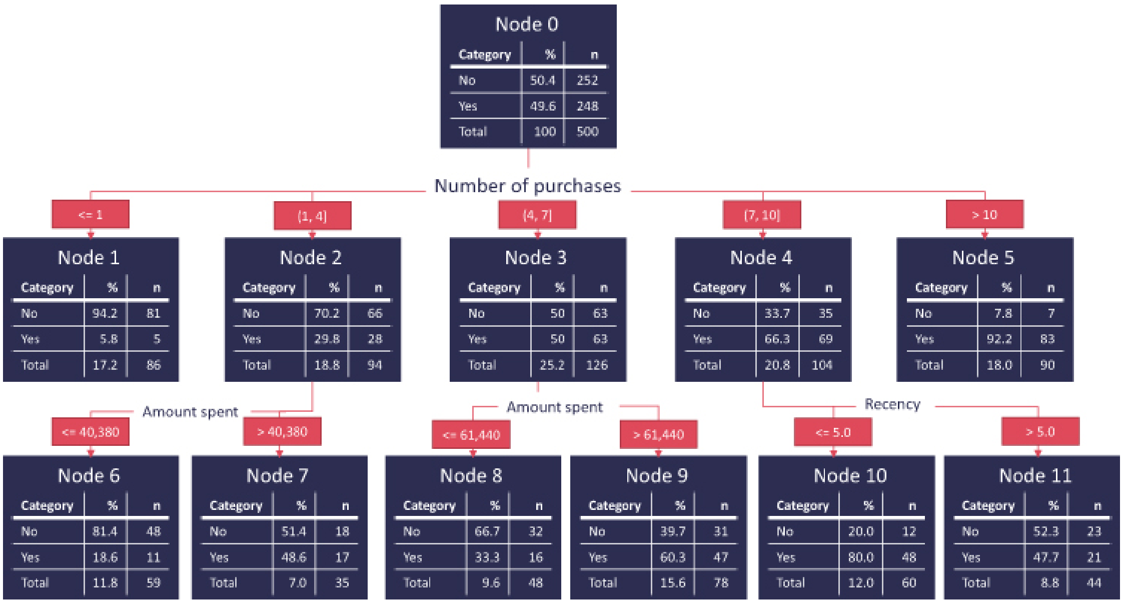


FIGURE 9.18 Example of a decision tree for the response to a test mailing

In Figure 9.18, an example of a tree model is presented. The tree should be read from top to bottom. The node at the top of the tree is called the root node and it indicates that we have 500 customers in analysis, of whom 50.4% did not respond to our test mailing and 49.6% did.[3] Subsequently, the root node is split into several child nodes, according to the number of customer purchases. This variable has a strong influence on the response to the test mailing, as customers with at most one purchase almost exclusively do not respond to the test mailing, whereas customers with more than ten purchases are almost certain to respond. The groups with a medium level of purchasing are less distinctive, and they are divided up according to splitting rules that are based on other variables. In line with tree terminology, the nodes at the end of the tree are called terminal leaves and we speak about "growing a tree."

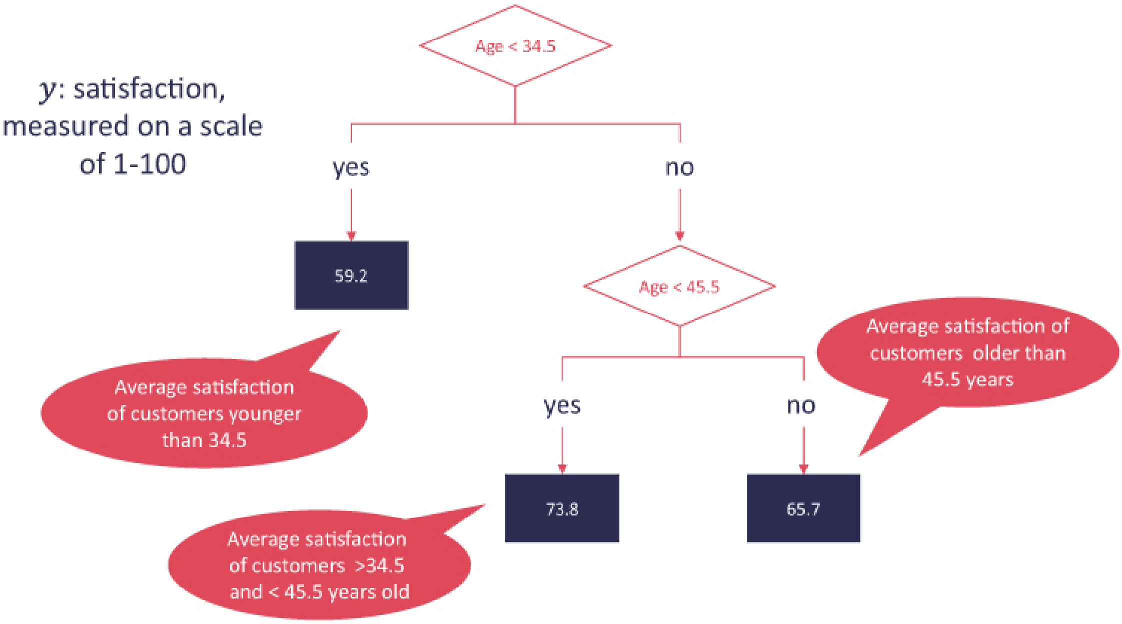
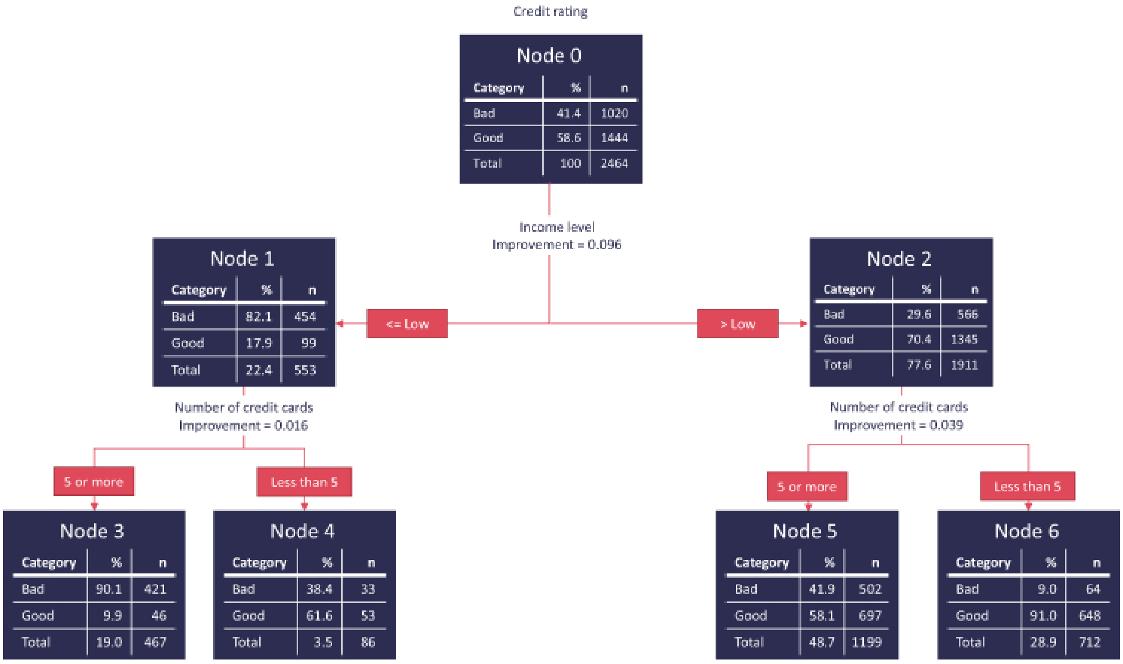

y: satisfaction, measured on a scale of 1-100
Age < 34.5
yes
no
59.2
Age < 45.5
Average satisfaction of customers younger than 34.5
yes
no
Average satisfaction of customers older than 45.5 years
73.8
65.7
Average satisfaction of customers >34.5 and < 45.5 years old

FIGURE 9.19 Nonlinearity in decision trees

Figure 9.19 illustrates that trees can also be used for numerical KPIs. The tree is quite simple and splits are based on age alone. The tree shows that customers younger than 34.5 years have an average satisfaction score of 59.2. Customers in the node that corresponds to the age group of 34.5–45.5 have an average satisfaction level of 73.8, and the satisfaction score of customers older than 45.5 is on average 65.7. If we read the three terminal leaves from left to right, we see that satisfaction first increases from 59.2 to 73.8 with the age of the customers, but later decreases to 65.7 as we move from the middle-aged group to the oldest group of customers. Importantly, this shows that a tree can accommodate the nonlinear effects of independent variables on the KPI.



Node 0

Category	%	n
Bad	41.4	1020
Good	58.6	1444
Total	100	2464

Node 1

Category	%	n
Bad	82.1	454
Good	17.9	99
Total	22.4	553

Node 2

Category	%	n
Bad	29.6	566
Good	70.4	1345
Total	77.6	1911

Node 3

Category	%	n
Bad	90.1	421
Good	9.9	46
Total	19.0	467

Node 4

Category	%	n
Bad	38.4	33
Good	61.6	53
Total	3.5	86

Node 5

Category	%	n
Bad	41.9	502
Good	58.1	697
Total	48.7	1199

Node 6

Category	%	n
Bad	9.0	64
Good	91.0	648
Total	28.9	712

FIGURE 9.20 Interaction in decision trees

Figure 9.20 shows yet another simple decision tree, now for a binary variable that represents credit rating. The first split in the tree is based on income and the two resulting groups are both split up according to the number of credit cards. Splitting on the same variable is not required by the decision tree; the algorithm could also have divided one of the two income groups using another variable. The data indicated that this resulted in the largest improvement of the fit of the tree model. We will elaborate more on splitting rules below.

Before discussing splitting rules in more detail, we would like to point out that the splits in the number of credit cards in both groups provide interesting insights. In the low-income group, having five or more credit cards is associated with an increase in the average bad credit rating of 90.1%–38.4% = 51.7%. In the higher income group, having more than five credit cards also relates to an increase in bad credit rating, but only with 41.9%–9.0% = 32.9%. Hence, we conclude that the effect of the number of credit cards on bad credit rating depends on income level. But this is exactly how we defined an interaction effect in Section 9.3! Consequently, decision trees not only accommodate nonlinearities by default, but also interaction effects. This is one of the reasons why decision trees have good predictive performance.

Over the years, several types of decision trees have been developed and they differ according to the decision rule that is used for deciding how to make splits in a tree. This involves selecting a variable for a split, but also selecting the threshold at which the split should occur. For example, the first split in Figure 9.20 is based on a split on the income variable. This could also have been a split on one of the other variables (e.g., the number of credit cards). Second, the split separates customers with a low income from customers with a medium or high-income level, whereas other splits on the same variable would also have been possible. For example, the algorithm could have distinguished between low-and-medium-income customers and high-income customers. Moreover, some splitting rules allow for splitting into more than two groups, so that the first split could also have resulted in three income groups. Broadly speaking, there are three popular types of splitting rules:

- Splitting rules that are based on statistical tests (Chi-square and F-tests)
- Splitting rules that are based on reducing diversity (Gini index)
- Splitting rules that are based on reducing entropy.

A popular example of the first type of splitting rule is known as CHAID, which short for CHi-squared Automatic Interaction Detector. If the dependent variable is categorical, Chi-square tests are conducted for all possible splits, and the split with the lowest *p*-value is chosen. If the dependent variable is numerical, the procedure is similar, but then F-tests are used. Further splits are not allowed if the splits do not result in significant differences in the child nodes. This type of decision rule is not restricted to binary splits and allows multi-way splits. Extensions of CHAID are exhaustive CHAID and Quick, Unbiased, and Efficient Statistical Trees (QUEST).

Splitting rules that aim for reducing diversity are based on the notion that a good split makes the child nodes more homogeneous than the parent node. For example, the split on the number of purchases in Figure 9.18 divided a very heterogeneous group of customers (about 50% responders and 50% non-responders) into groups that are much more homogenous (some groups are predominantly filled with responders, and others predominantly with non-responders). Classification and Regression Trees (CART) employ the Gini index as a "measure of impurity," and out of all possible splits, the split that reduces impurity the most is selected.

Splitting rules that aim to reduce entropy are based on the same underlying idea (reduce disorder in the system), but they use a different metric than the Gini-index for measuring non-homogeneity. C4.5 decision trees are a popular example of this and they are based on entropy to guide the splitting decisions.

There is no broad consensus about which selection rule is preferred over others. However, selecting a different method to grow a tree can strongly affect the shape and content of a tree. As a result of the ability to create multi-way splits, CHAID trees tend to be broader than other types of trees. Splitting into more than two child nodes sometimes resonates well with marketeers because this can align nicely with a segmentation approach. Tree methods that only allow for binary splits (such as CART) tend to produce "longer" trees, which are sometimes harder to explain. We advise using different splitting rules (tree models) and comparing the outcomes. If all trees roughly select the same variables and create similar splits, this confirms the robustness of the outcomes.

It is undesirable for a tree to have too many splits for two reasons. First, interpretability of a tree is hindered when the tree has a bewildering number of "branches." Second, a more complex tree is also more likely to suffer from overfitting. This is a problem that occurs if an unconstrained growth of the tree results in a near-perfect fit of the training data but performs poorly with the new data in the test set (see Section 9.6). The more specific our model is to our training data, the less it will be able to generalize to new data points. Kübler, Wieringa, and Pauwels (2017) argue that all machine learning models suffer from overfitting issues at some point and stress that it is the responsibility of the researcher to determine the right balance between the complexity of a model and its generalizability.

To avoid these two problems, there are several possibilities for controlling the size of the tree. For example, the maximum amount of sub splits can be controlled by restricting the depth of a tree and the creation of very small terminal leaves can be avoided by setting minimum requirements for the sizes of parent nodes and child nodes. Another approach is to first grow a large tree and 'prune' it later, for example using cost complexity pruning (Breiman *et al*., 1984).

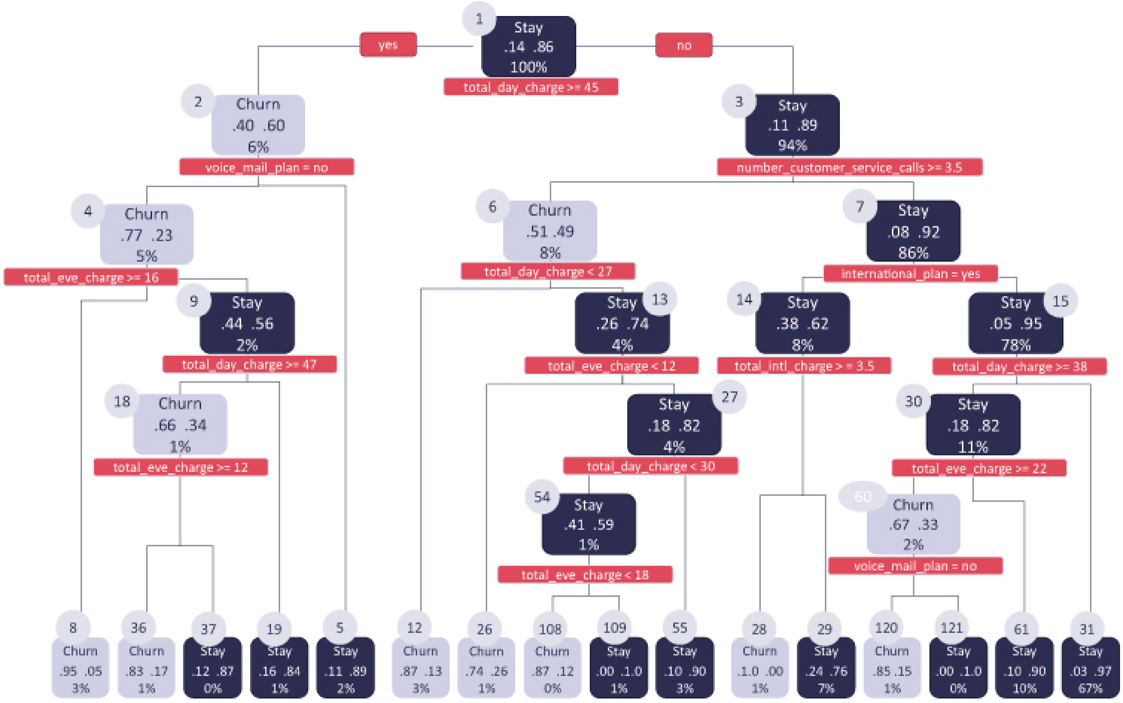


FIGURE 9.21 Tree model for the churn case

In Figure 9.21 we depict the R output of applying a tree model to the churn case that was discussed in Section 9.4. The tree identifies several high-risk customer groups: the customers with a high level of day charges and evening charges, and those without a voicemail plan. A large group of customers that have medium expenditures for day calls, that do not call the customer service desk frequently, and do not have an international plan are very likely to stay. Such findings can guide marketers when developing a retention campaign.

For evaluating the predictive validity of a tree model, we can utilize the same performance measures that were used for validating the logistic regression model. When we apply the fitted tree model to the holdout data, we find that the hit rate equals 93.1%, which is higher than the hit rate of logistic regression. Moreover, the top decile lift and the Gini coefficient are also much better, they increase to 5.65 and 0.74 respectively.

9.7 ENSEMBLE LEARNING MODELS: BAGGING, RANDOM FORESTS, BOOSTING

In computer science several models have also been developed that aim to improve predictions of the models discussed above. One problem with these models is that the results might be affected by the specific composition of a sample. Therefore, so-called aggregation or ensemble learning methods are used. The key idea behind these models is that improved predictions could be obtained by averaging the results of a large number of models. The idea behind aggregating multiple model results is that the quality of a single predictor might depend heavily on the specific sample (Breiman, 1996b) and this overfitting is not known beforehand. Averaging predictors of varying quality will result in more stable predictors (Breiman, 1996a, Malthouse & Derenthal, 2008). In this section we discuss a number of ensemble learning methods that are based on decision trees.

9.7.1 Bagging decision trees

An aggregation method that originated in the machine-learning field is bootstrap aggregation, or "bagging" (Lemmens & Croux, 2006). Bagging decision trees were first introduced into machine learning by Breimann (1996a). In the bagging procedure, a model is estimated on multiple bootstrap samples of the original training sample. To this end, a number (say, m) of datasets are created, consisting of random draws with replacement from the training set. Usually, these newly created datasets are of the same size as the full training set, so that some observations appear more than once in a bootstrapped dataset, while other observations may be missing. Each set is then used to train a decision tree.

A new data point in the test set is then used as input for each of the m trees that were trained in the previous step. Each tree generates a prediction for the output of the new data point, so that m predictions are obtained. All predictions for the output value of the new data point are then aggregated when assigning the final prediction. In case of a binary output variable, typically a majority vote is taken over the two classes, if the output variable is a numerical variable, the m predictions are averaged (Breiman, 1996a). The bagging procedure is illustrated in the left panel of Figure 9.22.

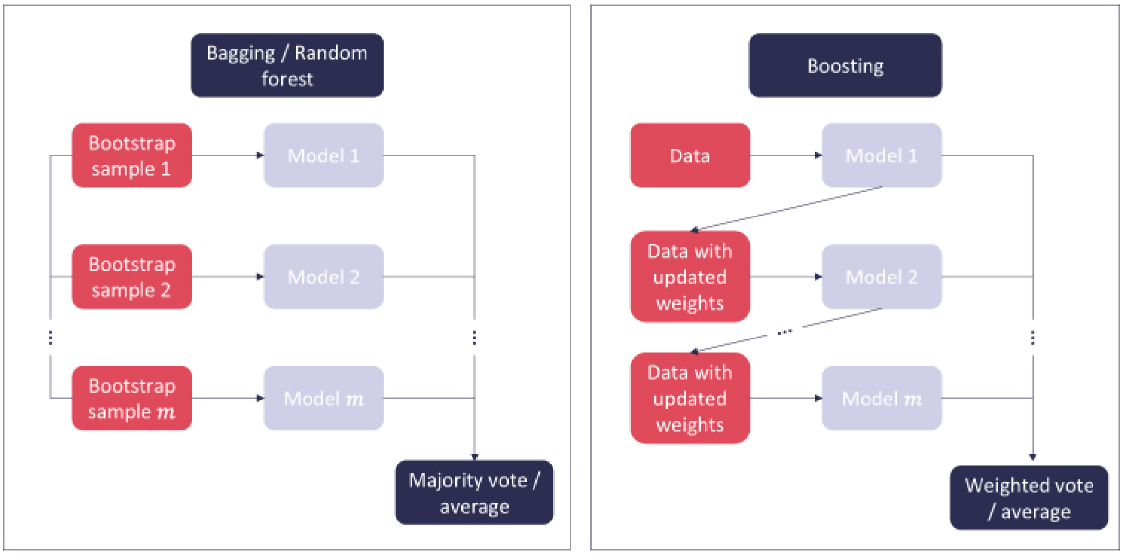


FIGURE 9.22 Illustration of ensemble learning methods based on decision trees

Bagging is a combination of bootstrapping and aggregation. The bagging procedure seems especially useful for improving the performance of decision trees. It can also be applied to other models, such as logistic regression models, but the latter type of model is not improved substantially by the bagging procedure as logistic regression model results are less affected by sample composition (Risselada, Verhoef, & Bijmolt, 2010). For more specific details on bagging, we refer to Lemmens and Croux (2006). Bagging is well implemented in R, for example in the library adabag.

When we apply bagging to the churn case of Section 9.4 (see the R script that accompanies this chapter), we obtain a hit rate of 93.3%, a top decile lift of 6.28, and a Gini coefficient of 0.69. We conclude that the bagging algorithm provides a significant improvement over logistic regression, and overall performs similar to a decision tree approach. It is, however, better able to separate churners from non-churners in our application.

9.7.2 Random forests

As explained above, the underlying idea of bagging decision trees is to improve generalizability of the predictions by aggregating across multiple decision trees. Each of these trees may individually be prone to overfitting of the training data but their aggregate result is not, unless the generated trees are correlated (Kübler, Wieringa, & Pauwels, 2017). Hence, the benefits of applying bagging are only realized if the m bootstrapped trees are uncorrelated. However, if some of the input variables are very strong predictors of the output variable, these will be selected in many of the m bootstrapped trees, causing them to become correlated. Ho (2002) introduces an adaptation of the bagging procedure to overcome this issue. At each split in training the m bootstrapped trees, he proposes to only allow splits based on a random subset of the input variables. Generating final predictions from the resulting set of trees is done similarly to bagging. Ensemble learning methods that are based on this idea are known as random forests. For more details, we refer to Skurichina (2002).

From the foregoing we conclude that random forests are closely related to bagging and, schematically, the procedures for obtaining predictions are very similar, as illustrated in the left panel of Figure 9.22. They differ in the way the models are derived from the m bootstrapped datasets (i.e., the difference is in the arrows that point from the bootstrap samples to the models in the left panel of Figure 9.22). With bagging, all input variables are used to train the models and in random forests the models are based on a random selection of the input variables.

When we apply the random forest adaptation to our churn case (see the R script that accompanies this chapter), we obtain a hit rate of 94.7%, a top decile lift of 6.77 and a Gini coefficient of 0.79, the highest scores on each of the criteria so far.

9.7.3 Boosting decision trees

Boosting is another example of an ensemble learning method for decision trees, in the sense that multiple trees are trained to achieve better predictive performance than can be obtained from individual trees. However, contrary to the bagging random forests, boosting does not train multiple trees in parallel and aggregates the results to generate predictions. Instead, a series of trees are trained sequentially on reweighted versions of the original training set.

In the first step of the algorithm, a decision tree is generated for the training data where all observations have the same weight. The misclassifications of this tree forms the basis for the second step, where each of the observations in the training data are reweighted. Contrary to expectations, the misclassified observations receive a higher weight than correctly classified observations. A second tree is then trained that takes the weights of the observations into account. The resulting misclassifications determine the weights for the third step, and so on. The idea is to increase the accuracy of subsequent steps by learning from the “mistakes” in the earlier steps. The final prediction is based on a weighted vote across the sequential predictions, where better predictions are weighted heavier. The boosting process is illustrated in the right panel of Figure 9.22.

Kübler, Wieringa, and Pauwels (2017) note that boosting is especially suitable in the case of different types of input variables and in the presence of missing data. In addition, boosting allows outliers and does not require any form of data transformation. For further reading we refer to Hastie, Tibshirani, and Friedman (2009) and Elith, Leathwick, and Hastie (2008).

When we apply boosting to our churn case (see the R script that accompanies this chapter), we obtain a hit rate of 94.1%, a top decile lift of 6.55, and a Gini coefficient of 0.77, so that boosting performs quite similarly to the random forest algorithm for our application.

The tree-based machine learning methods that we have discussed so far appear to have superior predictive power compared to logistic regression. However, they also have an important comparative disadvantage: they lack insight into the effects of the independent variables (input variables) on the dependent variable (output variable). This is especially true for ensemble learning methods, where multiple trees are involved in predicting values for the output variable. They are often criticized for being “a black box.” This criticism extends to the other machine learning methods that we will discuss in this chapter. Predictions from a machine learning model are seen as non-transparent because input data goes in, predictions for output variables come out, but the processes between input and output remain opaque. Linardatos, Papastefanopoulos, and Kotsiantis (2021) argue that this has made it problematic to adopt machine learning systems in many domains.

To overcome this issue, so-called feature importance scores can be calculated. Feature importance scoring refers to a class of techniques for assigning scores to input variables of a machine learning model that indicate the relative importance of each input variable when making a prediction of the output variable. Figure 9.23 shows importance scores for the tree-based methods that we have discussed so far. The most influential variable is total_day_charge in all tree methods, and total_evening_charge ranks second in most models. However, the graph also indicates that there are differences in the ranking of the variables in the different tree-based models.

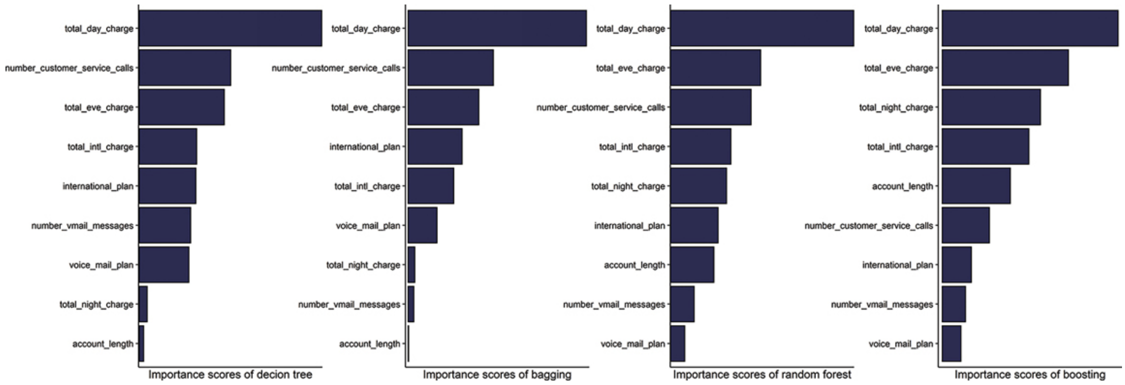


FIGURE 9.23 Importance scores for the tree-based models

Over recent years, many alternative measures have been developed to explain and interpret machine learning methods. The field of Explainable Artificial Intelligence (XAI) is an active area where methods that explain and interpret machine learning models are developed. A broad treatment of this field is outside the scope of this book. Instead, we refer to Linardatos, Papastefanopoulos, and Kotsiantis (2021) for a recent overview of this field.

9.8 NAIVE BAYES

Naive Bayes is a popular method for predicting the value of a categorical output variable, although its applications are not restricted to classification *per se*; the underlying idea can also be used to predict numerical values. Similar to tree-based methods, the Naive Bayes algorithm tries to build rules, based on Bayes' theorem, so that the (co)-occurrence of certain input values determines membership of the different classes of the output variable. The basic idea is that all input variables contribute independently to the probability of the occurrence of a value of the output variable. A drawback of this simple assumption is that it is somewhat naive (hence the name) as it is likely to be violated in many business applications. Despite this, the algorithm generally performs reasonably well and the advantage is that the algorithm trains faster than most other machine learning techniques and typically requires fewer training data points. Popular applications of Naive Bayes classifiers are email spam filters, customer class predictions, and collaborative filtering applications.

When we apply Naive Bayes to our churn case (see the R script that accompanies this chapter), we obtain a hit rate of 87.8%, a top decile lift of 4.17, and a Gini coefficient of 0.75. We conclude that for this application Naive Bayes is performing slightly better than logistic regression but is outperformed by the tree-based methods that were discussed in Sections 9.6 and 9.7.

9.9 SUPPORT VECTOR MACHINES (SVM)

Support vector machines also belong to the class of supervised machine learning techniques and are mostly used for classification problems (i.e. for classifying new data according to the categories of a nominal variable) but they are also suitable for numerical output (i.e. regression problems). Let us consider a classification setting where it is desired to separate two color classes according to two input variables, x_1 and x_2, as depicted in Figure 9.24. The simplest type of support vector machines use a linear approach to separate observations into homogenous classes. The left-hand side panel of Figure 9.24 shows that multiple straight lines separate the two color classes perfectly.

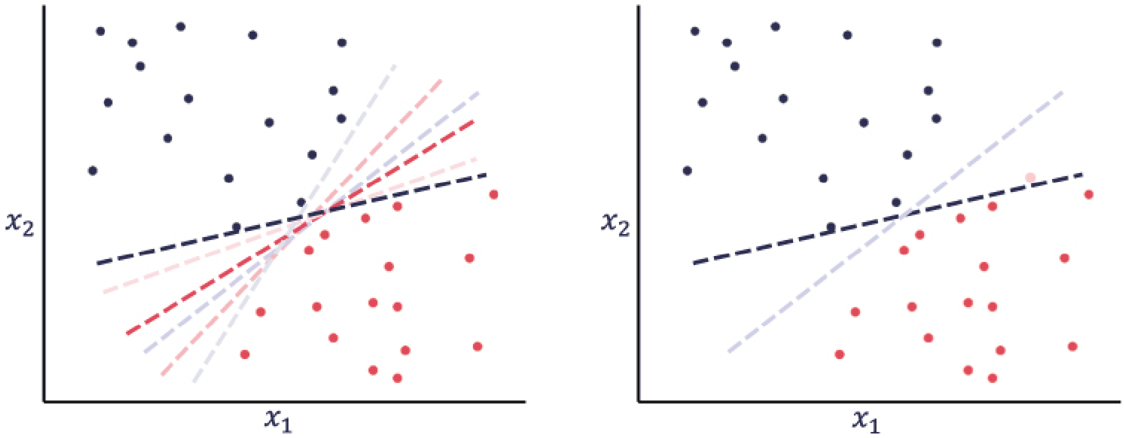


FIGURE 9.24 Different possible separation lines

Although all lines in the left-hand side panel of Figure 9.24 allow for an error-free separation of the two color classes, not all of them allow for optimal prediction for new incoming data. Suppose that a new data point of the red class is observed, which is represented by a pink dot in the right-hand side panel of Figure 9.24. Depending on the selection of one of the separation lines, the new observation is either correctly classified (light blue line) or wrongly classified (dark blue line).

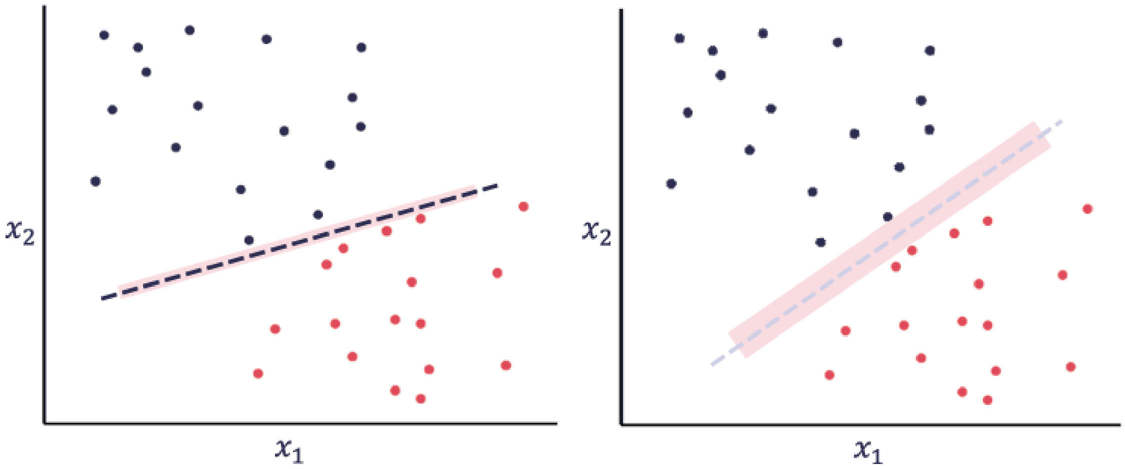


FIGURE 9.25 Selecting the optimal separation line

Kübler, Wieringa, and Pauwels (2017) argue that the difficulty with linear classification is to find a line that “best” separates the observed classes and allows the “best” possible prediction of new incoming data. As suggested by Figure 9.25, the linear classifier that separates the classes with the widest margin is expected to do well on new data. Therefore, support vector machines are sometimes called “large margin classifiers.” What is interesting about support vector machines is that that not all data points are used for determining the classifier. Instead, a relatively small portion of the data points determine the classifier and the associated margin. Those points are known as the support vectors (James *et al*., 2013).

In Figure 9.25, the two classes can be perfectly separated by a linear classifier. In practice, that is almost never the case. SVMs can deal with this via so-called soft-margin machines (Cortes & Vapnik, 1995). Moreover, support vectors can also allow for nonlinear classification via so-called kernel-based support vector machines (Boser, Guyon, & Vapnik, 1992). However, a discussion of such types of support vector machines is outside the scope of this book and we refer to James *et al.* (2013) for an excellent more detailed discussion.

When we apply a default support vector machine to our churn case (see the R script that accompanies this chapter for details), we obtain a hit rate of 94.1%, a top decile lift of 6.59, and a Gini coefficient of 0.78. We conclude that for this application the selected support vector machine is performing as well as the tree-based methods that were discussed in Sections 9.6 and 9.7.

9.10 NEURAL NETWORKS

Neural networks are among the most well-known types of machine learning algorithms. They are inspired by the structure and/or function of biological neural networks in a human brain (Kübler, Wieringa, & Pauwels, 2017). Mimicking the way a neuron in a human brain operates, a basic neural network assumes that the behavior of an output variable y is influenced by a set of input variables (or stimuli) $x_1, \ldots, x_m$. However, since not every input variable affects y equally strong, an artificial neuron in a neural network assumes that each input has its own weight: $\omega_1, \ldots, \omega_m$. How y responds to the weighed sum of the input variables is determined by a so-called activation function, as schematically depicted in Figure 9.26.

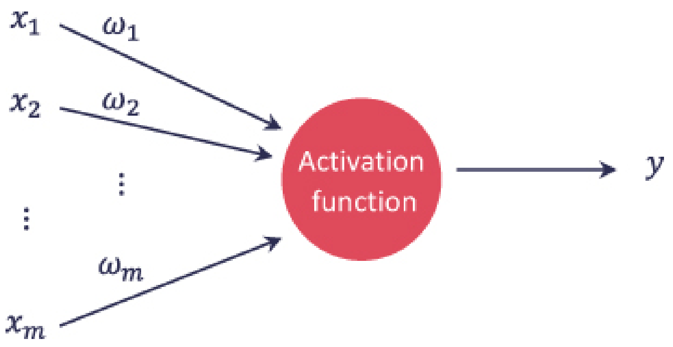


FIGURE 9.26 Schematic model of an artificial neuron

Kübler, Wieringa, and Pauwels (2017) say that artificial neurons can have different activation functions that transfer the weighted inputs to outputs. They present the threshold function as the most basic version, which is illustrated in panel (a) of Figure 9.27.

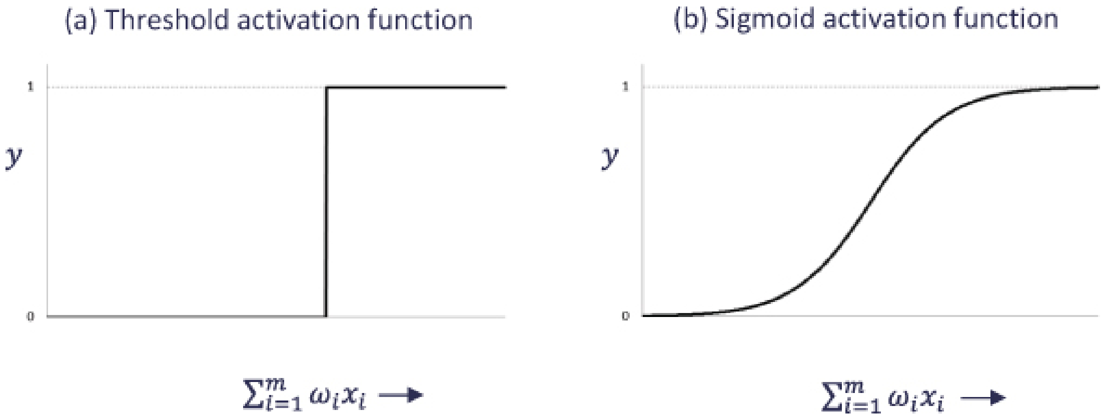


FIGURE 9.27 Two types of activation functions

When training a neural network with a threshold activation function, the algorithm learns values for the weights $\omega_1, \ldots, \omega_m$ and also for the threshold value where the output of the activation function changes from 0 to 1. Effective learning of the parameters involved is hindered by the step change in Figure 9.27(a), as a small change in one of the ω's or in the threshold value may result in a sudden 'jump' in y from 0 to 1. The learning process will be more efficient if small changes in the parameters lead to gradual changes in the output. To this end, so-called sigmoid neurons were developed, where the activation function is a logistic function of $\sum_{i=1}^{m} \omega_i x_i$, see Figure 9.27, panel (b).

The artificial neurons are the building blocks of a neural network and they are organized in layers, where each layer consists of one or more artificial neurons. Figure 9.28 illustrates that a neural network can be schematically depicted as an interconnected set of three types of layers. The neurons in the input layer are the first to process the information in the inputs. They "fire" their output to a second layer of neurons, which process this information as their inputs and pass their output on to the next layer. The last type of layer is the output layer that uses the incoming output of the preceding layer as input to make the final prediction.

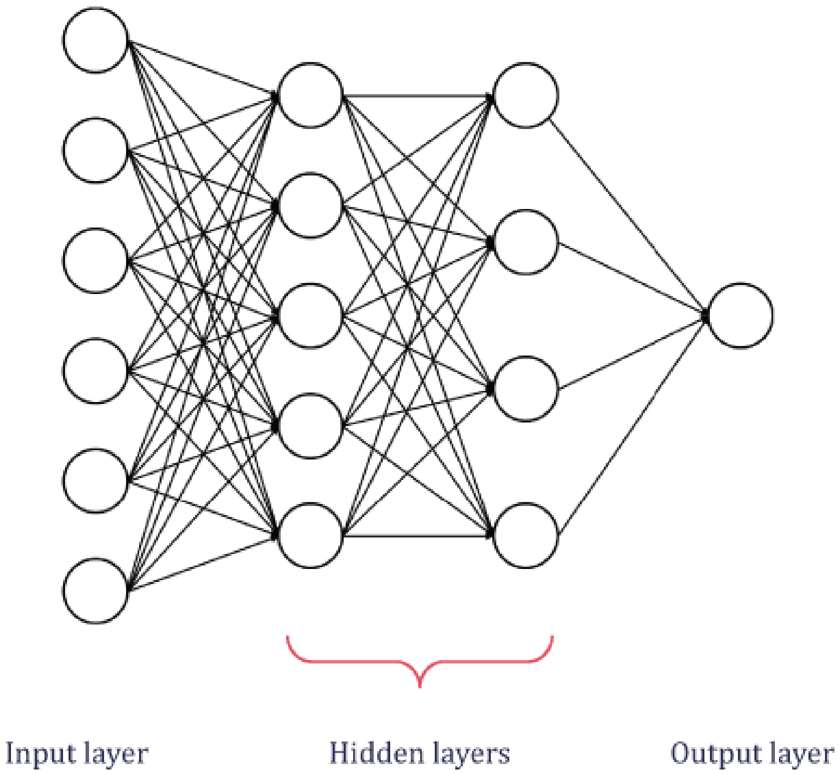


FIGURE 9.28 Structure of a neural network

Hidden layers reside in-between input and output layers and they are referred to as hidden because neither their inputs nor their outputs are directly observable. A neural network with more hidden layers is said to be a deeper neural network. As such models allow for much more complex processing of inputs, their predictive performance is often (much) better than that of shallower networks, e.g., those with one hidden layer. However, deep learning comes at the cost of additional training time.

There are different ways to set up the hidden layers. This makes it possible to adapt the neural network to specific characteristics of the data and the underlying data science project. For example, convolutional neural networks focus on image processing and recurrent neural networks are suited for modeling aspects of longitudinal or sequential data.

When applying an artificial neural network to a given problem, the researcher needs to decide on the number of layers, the number of nodes in each layer, the activation function, etc. Many approaches have been developed in the last decade and it is hard to give rules of thumb. For an overview, we refer to Nielsen (2015).

Apart from the appealing similarity between neural networks and the human brain, there is another important reason why neural networks receive much attention, and that is their so-called universality (Cybenko, 1989), which means that they can be used to approximate any continuous function to an arbitrary accuracy when the depth of the neural network is large enough.

However, Nguyen, Yosinski, and Clune (2015) show that some neural networks can easily be fooled.

When we apply a default neural network to our churn case (see the R script that accompanies this chapter for details), we obtain a hit rate of 94.6%, a top decile lift of 6.77, and a Gini coefficient of 0.79. We conclude that for this application the selected neural network ranks among the top-performing machine learning approaches.

9.11 REINFORCEMENT LEARNING

So far we have discussed two learning paradigms of machine learning: unsupervised learning (see Section 8.6) and supervised learning (see Sections 9.5–9.10). Kaplan and Haenlein (2019) distinguish a third paradigm: reinforcement learning, which they define as a learning process where the system receives an output variable to be maximized and a series of decisions that can be taken to impact the output. In reinforcement learning, the system aims to learn optimal behavior through trial-and-error interactions with a dynamic environment. All algorithms for reinforcement learning share the property that the feedback of the agent is restricted to a reward signal that indicates how well the agent is behaving. This makes it different from other machine learning approaches where a correct output value can be observed during training. In reinforcement learning there is no "correct" way to perform a task, yet there are rules the model has to follow. For example, a self-driving car should stay on the road and reach a given destination.

Kaplan and Haenlein (2019) illustrate the idea of reinforcement learning by means of an AI system that aims to learn to play Pac-Man simply by knowing that Pac-Man can move up, down, left and right and that the objective is to maximize the score obtained in the game. In much the same vein, Mnih *et al*. (2013) developed a single reinforcement network that learned to play seven different classic video games simply by analyzing raw pixel data, without knowing the rules of these games. The resulting system was able to outperform human experts on three of the games. Other well-known applications of reinforcement learning algorithms concern different computer-played board games (chess, Go), robotic hands, and self-driving cars.

De Bruyn *et al*. (2020) conceptualize reinforcement learning as follows. They define A as a set of actions an agent could take, S as the state of the environment, and R as the—long-term, and possibly discounted—reward received by the agent. Moreover, the problem is to approximate the function $f(S, A) \rightarrow R$ to select the best sequence of actions that leads to the highest stream of (discounted) rewards. The dynamic interaction of an agent with its environment is schematically depicted in Figure 9.29.

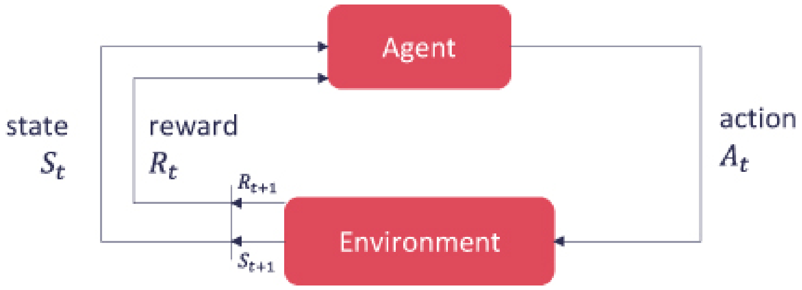


FIGURE 9.29 Schematic overview of a reinforcement learning process

De Bruyn *et al*. (2020) note that several interesting challenges arise in reinforcement learning, such as the necessity to balance exploration (to learn f in its entirety) and exploitation (to maximize rewards), or the difficulty posed by delayed rewards (where the agent may obtain the reward/penalty only at the very last step, such as winning/losing a game of chess, and the algorithm has to learn which actions led to that outcome). A full treatment of the area of reinforcement learning is outside the scope of this book. We refer to overviews by Sutton and Barto (2018) for a more detailed discussion, and to a review by Brei (2020) for interesting marketing applications of reinforcement learning.

We finalize this section by noting that reinforcement learning algorithms such as Temporal Difference learning, SARSA, and Q-learning benefit from the interactions between deep learning and machine learning, which means that reinforcement learning has the potential to become groundbreaking technology and the next step in AI development.

9.12 CONCLUSIONS

In this chapter we discussed theory-driven modeling and data-driven modeling. Within the first group of techniques, we elaborated on the steps of building a model guided by theory. We discussed two workhorses that belong to this group: linear regression and logistic regression. We applied both models to a practical case. The second group of models mainly concerns various machine learning techniques. We began with a discussion of tree models, since they are at the basis of several other techniques and are a popular multivariate technique for the following reasons:

- They are well-implemented in most statistical software
- They are very flexible as they, by default, accommodate nonlinearity and interactions
- There are no scaling restrictions for the predictors (x-variables)
- They are very suitable for datasets with many nominal variables
- They perform particularly well with large amounts of data
- They are easy to display graphically.

Tree methods can be a way to quickly see which variables are important because variables that are selected for early splits in a tree and repeating splits on the same variable indicate that such a variable is an important driver for the KPI under study. In the remainder of the chapter, we discussed a number of popular machine learning techniques and applied them to the same churn case that was analyzed with logistic regression. Figure 9.30 provides an overview of the performance of these techniques. We conclude that the machine learning methods offer a sizeable improvement in predictive performance compared to logistic regression, except for Naive Bayes.

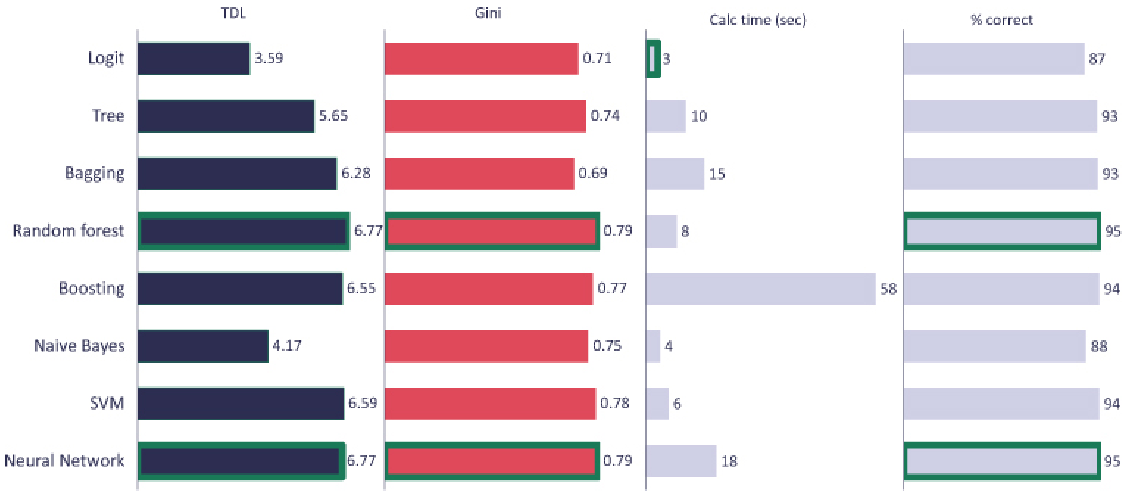


FIGURE 9.30 Performance of the machine learning methods

ASSIGNMENTS

VODAK is a mobile service provider in the Netherlands. The marketing manager would like to gain more insight into customer loyalty and what determines switching behavior. He also wants to be able to predict the value of customers. VODAK has access to a database containing more than a million customers. A sample of 7,001 customers was drawn from this database. Based on this data, you, as an analyst, will have to give the marketing manager more insights and make predictions. In the Appendix for this assignment, you will find a description of the data. You can download the database at www.masteringdatascience.eu.

Questions:

1. What is the average switching probability of a VODAK customer?
2. Determine, based on theory and intuition, which variables from the database influence the switching probability and what this influence will be (positive or negative)?
3. With simple descriptive analysis determine what differences exist between switchers and non-switchers.

4. Estimate a logistic regression to explain the switching probability. Develop a useful model and interpret the results. Do the results match your expectations and what are the differences? What is the predictive quality of the model? (The model can also be estimated with other techniques, such as decision trees and certain machine learning techniques. You can experiment.)
5. Predict the retention probability per customer based on the results of your model. Based on this also make a gains chart and deciles analysis. Specifically, look at how the deciles with the highest switch probability differ from those with a low switch probability.
6. Calculate the expected future Customer Lifetime Value per customer at a discount rate of 15% and a margin of 20% on revenue. How do you deal with the acquisition costs in this case?
7. Segment the customers based on the CLV and create profiles of the different segments.
8. How could VODAK increase the value of their customer base with a strategy targeting specific segments?
9. Data Description
    - Number of months since customer started
    - Age
    - Sex
    - Average revenue in euros per customer
    - Number of months since the start of the contract
    - Number of months till the end of the contract
    - Number of times the customer has switched phones
    - Age of the head of the household
    - Education level of the head of the household
    - Income of the head of the household
    - Household type
    - Work situation
    - Number of calls last month
    - Number of minutes used in the last month
    - Number of text messages sent in the last month
    - Average number of calls in the last six months
    - Average number of minutes used in the last six months
    - Average number of text messages in the last six months
    - Service contract
    - Acquisition costs
    - Switch (has the customer canceled the contract?).

A description of the values of the variables can be found in the txt file that accompanies the csv file (see www.masteringdatascience.eu).

NOTES

1. Although several authors dispute this argument (e.g., Wooldridge, 2010, Section 15.2).
2. Much of the discussion in these sections is based on Kübler, Wieringa, and Pauwels (2017).
3. Note that the unrealistically high response ratio in the data is a result of under-sampling of non-responding customers (see also Section 9.4).

REFERENCES

Bijmolt, T. H. A., Van Heerde, H. J., & Pieters, F. G. M. (2005). Determinants of price elasticity: New empirical generalizations. *Journal of Marketing Research*, *42*, 141–156.

Blattberg, R. C., Kim, B. D., & Neslin, S. A. (2008). *Database Marketing – Analyzing and Managing Customers*. New York: Springer Science & Business Media.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory—COLT, pp. 144–146.

Brei, V. (2020). Machine learning in marketing: Overview, learning strategies, applications, and future developments. *Foundations and Trends® in Marketing*, *14*, 173–236. https://doi.org/10.1561/1700000065

Breiman, L. (1996a). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Breiman, L. (1996b). The heuristics of instability in model selection. *The Annals of Statistics*, *24*(6), 2350–2383.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Monterey: Wadsworth & Brooks/Cole Advanced Books & Software.

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* (JAIR), *16*, 321–357. https://doi.org/10.1613/jair.953

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, *2*(4), 303–314. doi:10.1007/BF02551274

De Bruyn, A., Viswanathan, V., Beh, Y. S., Brock, J. K. U., & von Wangenheim, F. (2020). Artificial intelligence and marketing: Pitfalls and opportunities. *Journal of Interactive Marketing*, *51*, 91–105. https://doi.org/10.1016/j.intmar.2020.04.007

Derman, E. (2012). *Models.Behaving.Badly.: Why Confusing Illusion with Reality Can Lead to Disaster, on Wall Street and in Life*. New York: Free Press.

Donkers, B., Franses, P. H., & Verhoef, P. C. (2003). Using selective sampling for binary choice models. *Journal of Marketing Research*, *40*(4), 492–497.

Ebbes, P., Papies, D., & Van Heerde, H. J. (2011). The sense and non-sense of holdout sample validation in the presence of endogeneity. *Marketing Science*, *30*(6), 1115–1122.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*, 802–813.

Estabrooks, A., Jo, T., & Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*, *20*(1), 18–36. https://doi.org/10.1111/j.0824-7935.2004.t01-1-00228.x

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *Boosting and Additive Trees*. *The Elements of Statistical Learning* (2nd edition). New York: Springer.

Ho, T. K. (2002). A data complexity analysis of comparative advantages of decision forest constructors. *Pattern Analysis and Applications*, *5*, 102–112.

Horváth, C., Leeflang, P. S. H., Wieringa, J. E., & Wittink, D. R. (2005). Competitive reaction- and feedback effects based on VARX models of pooled store data. *International Journal of Research in Marketing*, *22*(4), 415–426.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (eds). (2013). *An Introduction to Statistical Learning: With Applications in R*. New York: Springer.

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, *62*(1), 15–25. https://doi.org/10.1016/j.bushor.2018.08.004

Kübler, R. V., Wieringa, J. E., & Pauwels, K. H. (2017). Machine learning and big data. *Advanced Methods for Modeling Markets*, 631–670. https://doi.org/10.1007/978-3-319-53469-5_19

Leeflang, P., Bijmolt, T., Pauwels, K., & Wieringa, J. (2015). Modeling Markets: Analyzing Marketing Phenomena and Improving Marketing Decision Making. Springer. https://doi.org/10.1007/978-1-4939-2086-0

Leeflang, P. S. H., & Wittink, D. R. (1996). Competitive reaction versus consumer response: Do managers overreact? *International Journal of Research in Marketing*, *13*, 103–119.

Leenheer, J., Van Heerde, H. J., Bijmolt, T. H. A., & Smidts, A. (2007). Do loyalty programs really enhance behavioral loyalty? An empirical analysis accounting for self-selecting members. *International Journal of Research in Marketing*, *24*, 31–47.

Lemmens, A., & Croux, C. (2006). Bagging and boosting classification trees to predict churn. *Journal of Marketing Research*, *43*(2), 276–286.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy*, *23*(1), 18. https://doi.org/10.3390/e23010018

Little, J. D. C. (1970). Models and managers: The concept of a decision calculus. *Management Science*, *16*, B466–B485.

Malthouse, E. C., & Derenthal, K. M. (2008). Improving predictive scoring models through model aggregation. *Journal of Interactive Marketing*, *22*(3), 51–68. https://doi.org/10.1002/dir.20117

Marinakos, G., & Daskalaki, S. (2017). Imbalanced customer classification for bank direct marketing. *Journal of Marketing Analytics*, *5*(1), 14–30.

Marqués, A. I., García, V., & Sánchez, J. S. (2013). On the suitability of resampling techniques for the class imbalance problem in credit scoring. *Journal of the Operational Research Society*, *64*(7), 1060–1070. https://doi.org/10.1057/jors.2012.120

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing Atari with Deep Reinforcement Learning. arXiv:1312.5602 [cs]. http://arxiv.org/abs/1312.5602

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research*, *43*(2), 204–211.

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 427–436.

Nielsen, M. A. (2015). *Neural Network and Deep Learning*. Determination Press. Retrieved from http://neuralnetworksanddeeplearning.com

Papies, D., Ebbes, P., & Van Heerde, H. J. (2017). Addressing endogeneity in marketing models. In: P. S. H. Leeflang, J. E. Wieringa, T. H. A. Bijmolt, & K. H. Pauwels (eds.) *Advanced Methods for Modeling Markets* (pp. 581–627). Springer International Publishing. https://doi.org/10.1007/978-3-319-53469-5_18

Pauwels, K. H. (2017). Traditional time-series models. In: P. S. H. Leeflang, J. E. Wieringa, T. H. A. Bijmolt, & K. H. Pauwels (eds.)

Advanced Methods for Modeling Markets (pp. 87–114). Springer International Publishing. https://doi.org/10.1007/978-3-319-53469-5_3

Risselada, H., Verhoef, P. C., & Bijmolt, T. H. A. (2010). Staying power of churn prediction models. *Journal of Interactive Marketing*, *24*(3), 198–208.

Rossi, P. E. (2014). Even the rich can make themselves poor: A critical examination of IV methods in marketing applications. *Marketing Science*, *33*(5), 655–672. https://doi.org/10.1287/mksc.2014.0860

Skurichina, M. (2002). Bagging, boosting and the random subspace method for linear classifiers. *Pattern Analysis and Applications*, *5*, 121–135.

Sutton, Richard S., & Barto, Andrew G. (2018). *Reinforcement Learning: An Introduction* (2nd edition). Cambridge, MA: MIT Press.

Wasserman, L. (2012). Statistics and Machine Learning, accessed online March 14, 2021 at https://normaldeviate.wordpress.com/2012/06/12/statistics-versus-machine-learning-5-2/

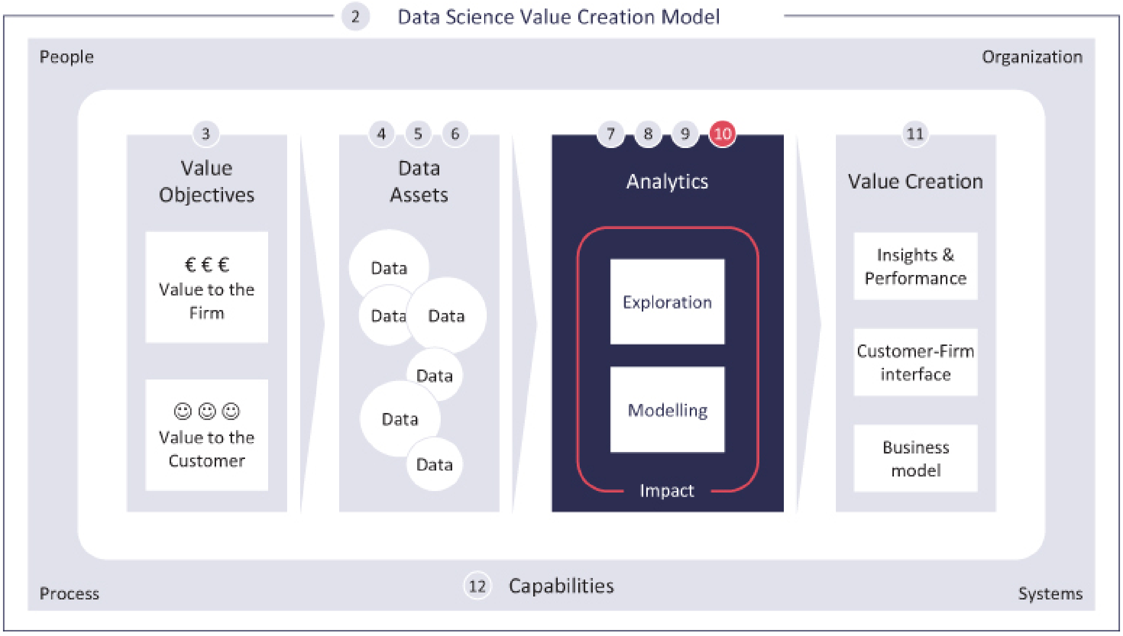
Wedel, M., & Kannan, P. K. (2016). Marketing analytics for data-rich environments. *Journal of Marketing*, *80*(6), 97–121. https://doi.org/10.1509/jm.15.0413

Wickham, H., & Grolemund, G. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data* (1st edition). O'Reilly Media. Retrieved from https://r4ds.had.co.nz/

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data* (2nd edition). Cambridge, MA: MIT Press.

CHAPTER 10
Creating impact with storytelling and visualization

DOI: 10.4324/9781003011163-10



10.1 INTRODUCTION

In the preceding in-depth chapters, we have discussed analytical techniques. For analysts to create impact with data analytics, the communication of the results is crucial. One of the main dangers of analysis is that a report is not even presented or ends up never-opened in the desks of managers and therefore never has any effect on management decisions. To have impact two issues are of crucial importance:

1. The presence of a clear storyline in which the message of the implications is concisely discussed.
2. The use of powerful visualization of the analytical results (i.e., effective use of visual aids).

The importance of these two issues is increasingly present. The growth in the availability of continuous digital information results in people having less time available to attend to communications. Consumers, as well as managers (being also consumers), continuously switch between multiple devices to read messages, emails, news apps, social media, etc. It is now very common that, in meetings, participants do not give their full attention to presentations due to distractions on their tablets or mobile phones. It is therefore very important that presentations of research results are sufficiently clear and attract attention. Further, today's overload of digital information means that managers have to find ways of filtering the right information and interpreting the results. This information overload is not new. Over the last two decades, people have been addressing the growing importance of information stress. This arises because of the growing divergence between what information is available and what we can process (see Figure 10.1). It can be viewed as a black hole between data and knowledge that starts to exist when information is not telling us what we want and should know. For a long time, managers did not understand what they did not know. However, now they do understand it and, as a consequence, they feel information stress (Wurman, 1989).

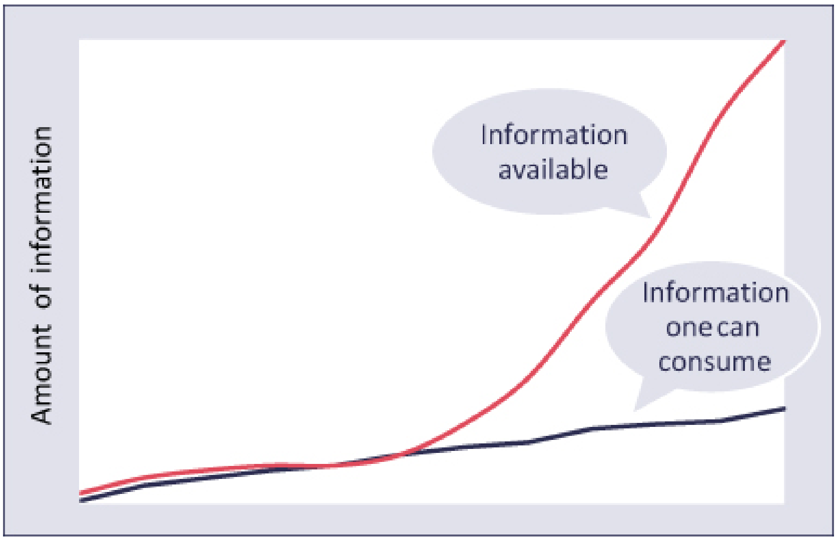


FIGURE 10.1 Information overload

This increasing overload is calling for solutions. One of the solutions has been to work with infographics. These infographics are a common method for making complex information more accessible to the reader. They exist in many forms. It might, however, be questionable whether infographics are effective. An infographic usually involves little structure and relays many facts in the form of text and figures. The graphs or pictures are mainly used as illustrations to make the information more attractive.

We strongly believe in the combination of data, storytelling, and visualization. These three elements should strengthen each other in such a way that the information has an impact (see Figure 10.2). If analysts, based on the strengths of their data and analytics, can tell a strong story and provide strong visualizations, they should have a strong impact. A kind of "sweet spot" is achieved, as strong visualizations and good storytelling, combined with excellent data and analytics, will be well received by managers in an era of information overload. To achieve this, multi-disciplinary skills are required. This is not easy, as frequently the analyst focuses on numbers and may be unskilled at communicating a strong story. We will discuss this general issue in more detail in Chapter 12.

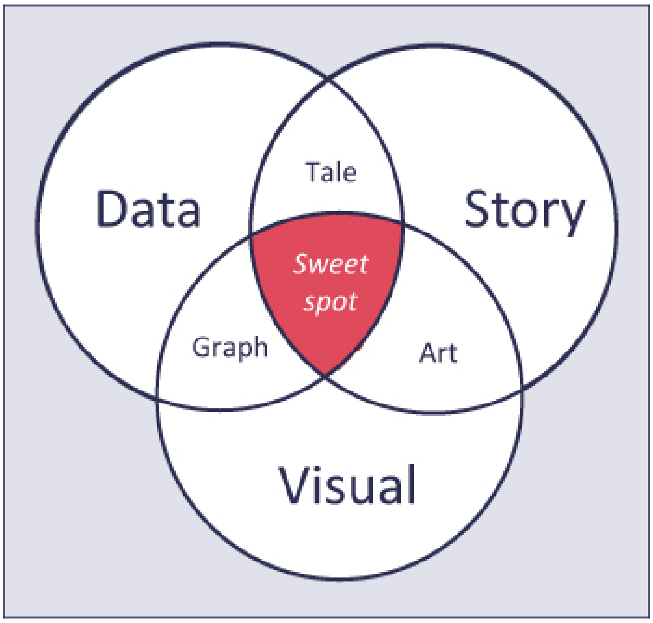


FIGURE 10.2 Sweet spot of data, story and visual

In this chapter, we discuss how to build up a good storyline in analytical reports and presentations and how strong visualization can be achieved. We also pay attention to how to recognize and avoid the misinterpretation of information by discussing some common misleading graphs. Before doing so, we first focus on why many analytical projects with strong data and analytics fail to have an impact.

10.2 FAILURE FACTORS FOR CREATING IMPACT

Many analysts have probably experienced carrying out a very nice study, but in the end found their work did not change marketing strategies or tactics. Why does this happen? It is not the analytical quality that is causing the problem. The numbers have been crunched in the right manner, the right research questions have been studied, but still impact is limited. This is probably the greatest frustration of many analysts. Having no impact will in the long run threaten the position of the analytical role within firms. Creating impact is also strongly required to create value with data analytics! It is therefore very important to understand why reports do not have an impact and what typically goes wrong. We have already emphasized the huge importance of storytelling and visualizations. Based on our experience in analytical functions in many firms, we can identify some specific issues that frequently go wrong and reduce the impact of analytical exercises:

- There is no structure to the report. Often they consist of different analyses that are not strongly related and as a consequence, many unrelated messages are communicated instead of a few strongly related messages.
- There are no strong and clear conclusions or messages. The findings are nice to know, but it is unclear what the manager should do differently after reading the report. One easily gets the "So what?" response.
- The reports include too many pages or slides. Moreover, the conclusions are only reported at the end of the report. The reader's attention has died by the end of the report and conclusions and implications are not read or mentally processed! The consequence is no impact.
- The main findings are good and are understood but they are combined with nice-to-know, irrelevant insights. These insights distract the reader and result in a diminished focus on the main message of the study.
- There are inconsistencies in the report. This creates a discussion on the content and may create confusion, reducing the perceived reliability of results.
- The report focuses too much on statistical details and the choices of specific methods. Although this is highly valued in scientific publications, managers have no strong interest in the details. Instead of reporting this in the main text, it can be provided as an appendix.
- The slides (or pages) are packed with messages and, as a result, look very crowded. The average processing capacity is limited, and people are easily distracted, which means they can usually process only a limited number of points.
- Analysts frequently only report a bunch of numbers instead of graphs. Numbers are harder to process than visual graphs.
- Graphs are used but they are too complex and provide too much information. As a consequence, it is a puzzle for managers to pick out the right information.

All these issues relate to weak communication. Improved communication can happen when analysts learn how to build a strong, focused story for their results and are able to visualize them in the right way.

10.3 STORYTELLING

One of the basic principles of a good report or presentation is that it has a core message. This core message should be introduced with a specific situation and complication and should subsequently be underpinned with arguments. This approach is based on the pyramid principle as advocated by Barbara Minto (2009).[1] At the end of the 1960s she worked as a consultant at McKinsey & Company, where she focused on the development of a method to help their advisors structure their presentations and reports and to lay out information in the most efficient way possible, based on how people with little time, especially executives, absorb information.

The pyramid refers to the principle that each advice should have a pyramid-like structure. First there is the core message, then the explanation. At the top of the pyramid is the advice, and below the top structure, in different points or paragraphs, is the motivation. If the motivation is divided into multiple subpoints or issues a new pyramid is created. Subsequently, one provides a powerful discussion (or description) of the complication to introduce the key-message. In our experience of giving advice to companies based on analytics, we have observed that this pyramid principle is very powerful. It really strengthens the impact of analytics. Schematically this results in the structure displayed in Figure 10.3.

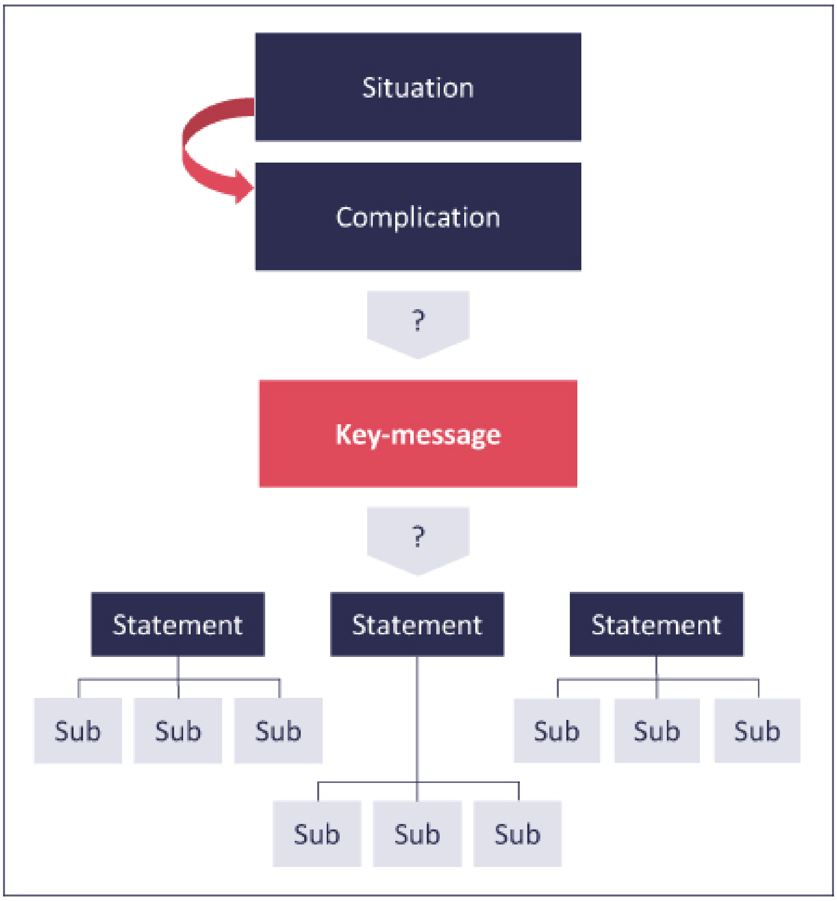


FIGURE 10.3 Building blocks for a clear storyline

To understand why this method can be powerful, we first consider what frequently happens when reporting. Normally, analysts start to discuss what they have done in a kind of chronological order. They want to show the manager their analytical road trip from problem statement to end results. The analysts only end their presentation or report with important conclusions. They also want to be as complete as possible and aim to tell every detail. This makes sense considering that this is how most analysts are trained in universities. When writing a thesis, they start with a problem statement, discuss the theory, the data collection, the analytical method, the results section, and end with important conclusions. Similar structures can be found in many scientific papers, and this may potentially explain the limited impact of scientific papers on practice (Roberts, Kayandé, & Stremersch, 2014). However, with this structure, the manager, with limited time and attention, only gets to the most important results at the end of the report or when the session is almost finished.

A report has much more impact when its core message and the context are directly understood. By "directly" we mean that this should be set out at the beginning of each report or presentation. The core message can then be underpinned with a limited number of arguments—one frequently uses seven as a kind of rule of thumb, a kind of magical number based on Miller's law. The cognitive psychologist George A. Miller of Princeton University has shown that there are severe limits in our capacity to process information (Miller, 1956). His work has been interpreted to mean that the average number of objects an average human can hold in memory is around 7. This strongly suggests limiting the volume of messages and arguments being discussed.

The above discussion clearly shows that there is a mismatch between how an analyst commonly presents an analysis and how it should be presented. An analyst frequently solves a problem with a bottom-up type of approach. However, effective communication suggests a top-down approach (see Figure 10.4). It is essential for analysts to understand this difference. When finishing a project and preparing the report and/or presentation they should move from the analytical to the effective communication mindset. We have observed that analysts typically find this difficult, given that they tend to focus on details and frequently forget the overall picture and the purpose of the analysis. It is therefore important to work in analytical teams where effective communication skills are embedded in the team (see also Chapter 11).

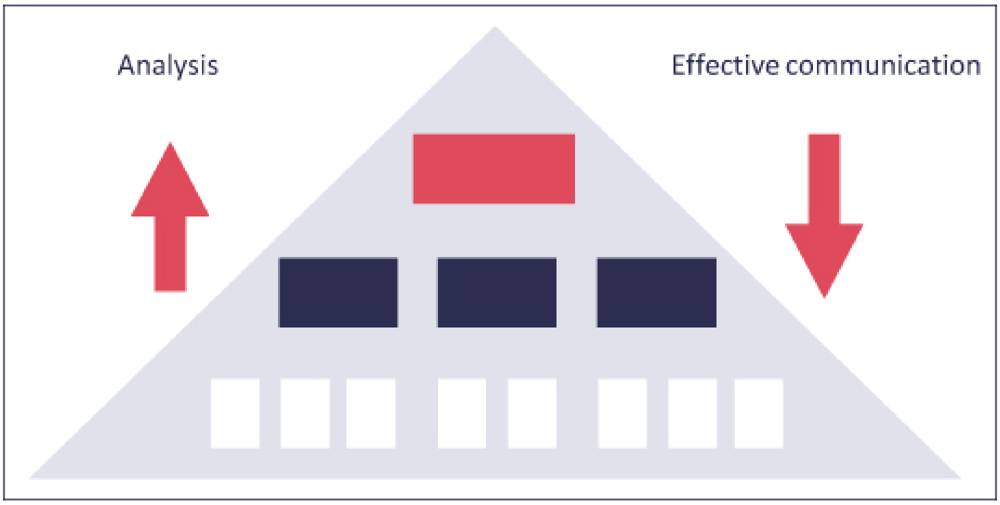


FIGURE 10.4 Analysis process vs. effective communication

The top-down approach of the pyramid principle offers an effective way of communicating the key message. This structure applies logical reasoning, because the main message is supported by coherent arguments, which ensure that the conclusion is logical and credible. The structure also offers the possibility of a pro-active Q&A dialogue, because the breakdown by arguments and sub-arguments provides answers to follow-up questions when a key message is introduced.

10.3.1 Checklist for a clear storyline

The above discussion probably seems rather intuitive, but how can its conclusions actually be implemented? We take the schema as shown in Figure 10.3 as a starting point and briefly point to some issues requiring attention.

10.3.1.1 Situation

When describing the initial situation, the following issues require consideration:

- Is the discussed situation not controversial? Does the description in itself raise specific questions and/or a debate? If the latter occurs it will be more difficult to discuss the core message.
- Does the audience recognize the described situation? If so, they will be more receptive.
- Is the situation description underpinned with figures and are these figures understood and believed in the organization? If the latter is not the case the situation description will be less effective. Nevertheless, figures showing specific problems in performance (e.g., decrease in net promoter

score (NPS), or increase in churn rates) are very important for showing the relevance of the report.
- Does the situation description create a complication and a specific research question?

10.3.1.2 Complication

The complication can be defined as describing the problem or challenge. This should be directly related to the situation description. There are some specific issues here to consider as well.

- Does the complication describe its potential impact on the organization? For example, in the case of decreasing churn rates, the impact could mean lower sales over time, decreasing market share, and lower profitability.
- Is the complication firmly underpinned with arguments and/or figures? Again, we advise focusing on figures that can be directly linked to performance consequences. This will create a stronger belief in the urgency and relevance of the study.

10.3.1.3 Message

When discussing the message, the following issues should be checked:

- Is there a single core message or are there multiple messages? We prefer to work with a single core message to ensure greater impact.
- Does the core message create some curiosity or questions? Curiosity will generate attention and a desire to listen.
- Does the core message provide an answer to the complication?

10.3.1.4 Underpinning the message

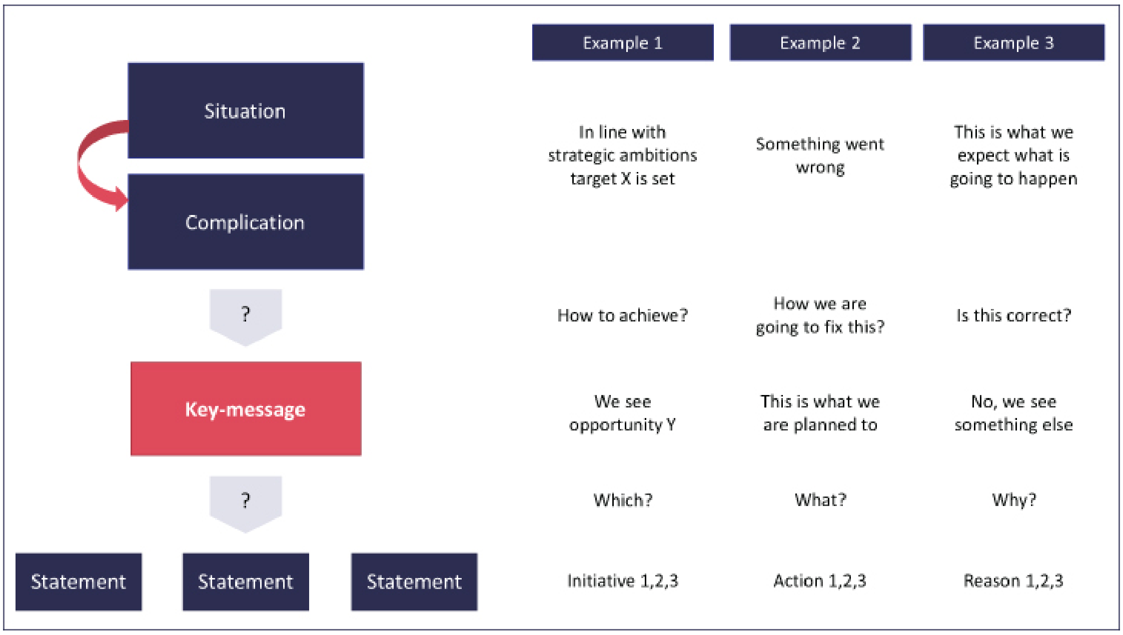
When providing arguments for the message, it is important to assess what, why, and how the argument is being used. Does the argument make sense and will it provide a strong underpinning for the message? Specific issues that require attention are:

- Do the arguments link with questions a reader will ask when reading the core message? It is thus very important to understand how managers will react and what questions will come up when the core message is being read.
- Are the arguments complete and mutually exclusive? A complete list of arguments will show that the analyst has seriously thought about the

conclusion offered. Mutually exclusive arguments mean that there is no overlap in the conclusions or in the opportunities you found.

- There should be neither too few nor too many arguments. A general rule of thumb is that there should be a minimum of two arguments and a maximum of five.
- One should start with the most important and convincing argument and end with the least important one.
- Are the arguments compatible? For example, when offering strategic arguments, one should not use tactically based arguments. Or if arguments are based on facts, it is probably not wise to use sentiments as well.

In Figure 10.5 we give some examples of storylines that differ in their purpose. In example one we show how one can come up with business opportunities that achieve the business target.



	Example 1	Example 2	Example 3
Situation Complication	In line with strategic ambitions target X is set	Something went wrong	This is what we expect what is going to happen
?	How to achieve?	How we are going to fix this?	Is this correct?
Key-message	We see opportunity Y	This is what we are planned to	No, we see something else
?	Which?	What?	Why?
Statement	Initiative 1,2,3	Action 1,2,3	Reason 1,2,3

FIGURE 10.5 Examples of different storylines for different purposes

The pyramid principle is often applied in the creation of PowerPoint presentations, which are used to share analytical results. But we think it is important to note that the pyramid principle method is applicable in all kinds of (business) communication. This method of storytelling is also very effective in conveying the key-message in written reports, notes, emails, voice messages and video.

10.4 VISUALIZATION

Visualization is of utmost importance in creating impact with data analytics. The reason is rather simple: “A picture is worth a thousand words.” There are also some statistics underlying these claims. For example, if the information is transferred orally, only 10% of the recipients can remember that information after 72 hours; this percentage rises to 65% if the information is visualized. This is called the “picture superiority effect” (Paivio & Csapo, 1973). So, if you look at Figure 10.6 it is more likely that you will remember the right part (with the apple) than the left part.

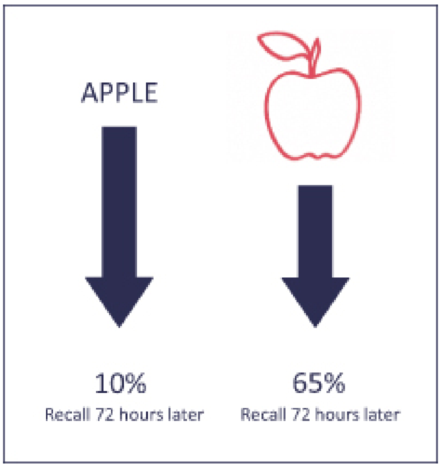
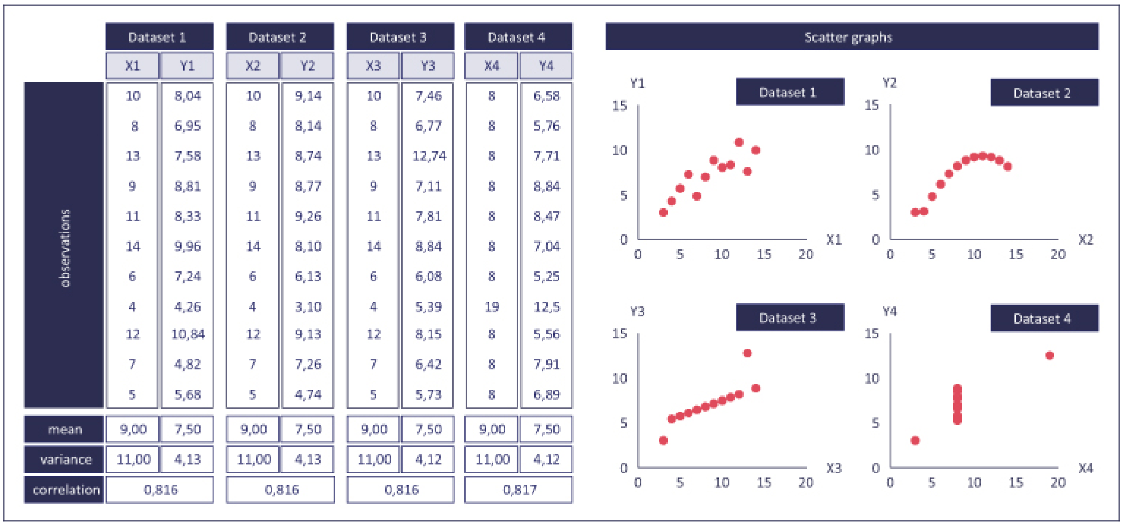


FIGURE 10.6 The picture superiority effect

It is much easier to understand and extract value from data when it is offered through data visualization rather than through looking at raw data or the simple data statistics. In 1973, the statistician Francis Anscombe demonstrated the importance of graphing data. Anscombe’s quartet shows how four sets of data with identical simple summary statistics can vary considerably when graphed (see Figure 10.7).



	Dataset 1		Dataset 2		Dataset 3		Dataset 4	
	X1	Y1	X2	Y2	X3	Y3	X4	Y4
observations	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,10	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,10	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
mean	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
variance	11,00	4,13	11,00	4,13	11,00	4,12	11,00	4,12
correlation	0,816		0,816		0,816		0,817	

FIGURE 10.7 Graph of Anscombe’s quartet data table

Source: Adapted from Anscombe (1973)

Visualization in analytics is used for many purposes. Generally, three objectives can be achieved by visualizing data:

- Exploration of data
- Understanding and making sense of the data
- Communicating the results of the analysis.

The first two objectives are generally parts of the analysis process. Before running all kinds of analyses, it is wise to explore the data with visuals to help make sense of them. This can lead to immediate valuable insights and the understanding of potential relationships in the data. The last objective is linked to the presentation of the results and creating impact. In some cases, visualization of data explorations can also be used in the presentation if it underpins the main message of the report. Visualizations in many forms can be used wherever results are presented, such as in reports, presentations, marketing dashboards, and websites.

In the next sections, we aim to provide some practical guidelines on how to effectively visualize when communicating the results. We specifically focus on choosing the right chart type. Furthermore, we will provide you with some practical tips and tricks to further improve the visualization. We also pay attention to how to recognize and avoid the most common ways information gets misinterpreted by discussing the most common misleading graphs.

10.4.1 Choosing the chart type

A common mistake when using a chart is to choose one type, for example a bar chart or a scatter plot, and assume that using it will be sufficient and self-explanatory in a report or presentation. However, this is frequently not the case. Choosing the right chart format to communicate analytical results should be done carefully. The problem that researchers face is that there are many graph types, styles, and methods for presenting data. This makes it difficult to choose the right format. To find the right chart type, it is important to know that there are four core types in visualizing data:

1. Showing a relationship between data points
2. Comparing data points
3. Showing the composition of data
4. Showing the distribution of data.

When choosing the right chart type it is first important to assess which graph type fits best with the message one aims to convey. In doing so, one has to consider the purpose of the graph. The above distinction in the various ways to visualize can be helpful in this respect. We will therefore discuss the different types of data visualization for each of these four types.

10.4.1.1 Relationship between data points

A graph displaying a relationship aims to show the association or correlation between two or more variables through the data presented, for example showing the relationship between in-store sales and holidays. The most common "relation charts" that do this are scatter plots and bubble charts. Yet you can also think about network charts when you want to show the relationship between objects (see Figure 10.8).

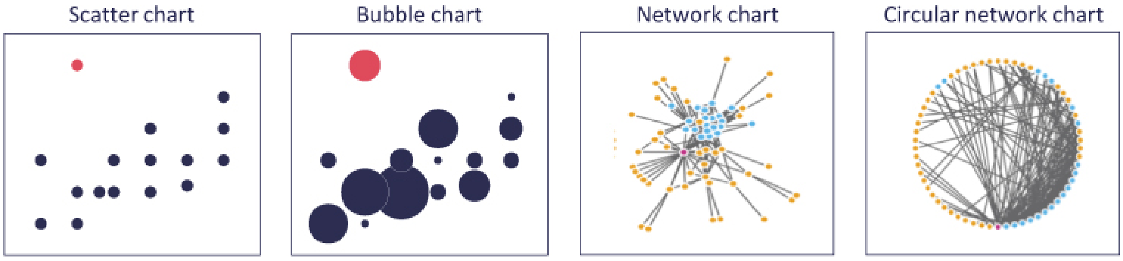


FIGURE 10.8 Relationship charts

- A scatter chart is used to show a relationship between two variables (X, Y) to determine if they tend to move in the same or different directions. An example might be plotting NPS (X) and retention (Y) for a sample of months.
- A bubble chart is an extension of the scatter chart, adding a third variable. This ends up being reflected in the size of the bubble. For example, when showing the relationship between NPS and retention, the size of the bubble might reflect the number of customers at a specific data point.
- A network chart typically shows the relationships between objects or individuals. This last type of visualization can be used in social network analysis.
- In a circular network chart, the network chart is extended by showing the position and the importance of the objects. One concern is that these network charts become less clear when many objects are involved and may even become unreadable.

10.4.1.2 Comparing data points

The basic idea of the comparison of data points is that one aims to compare scores for (a set of) variables across multiple subunits (e.g., groups, time). For example, one might aim to show the sales per category per quarter. Or one may aim to show the conversion rates for different websites over time. Again, different types of "comparison charts" can be chosen (see Figure 10.9).

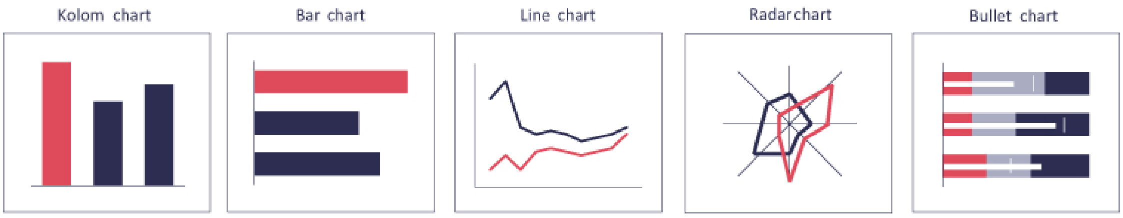


FIGURE 10.9 Comparison charts

If one has simple comparisons (e.g., sales per brand), then usually one of two rather similar chart types is used:

- A column chart is used when there is a limited number of subunits
- A bar chart is used when the number of subunits is larger, as more space is available in this graph to show the descriptions of the subunit.

When comparing measurements over time, it is easy enough to use a column graph if only a limited number of categories for only a few periods (e.g., four quarters) is being considered. However, usually, more categories and more time periods are considered. A line chart is especially useful if multiple time periods are being considered (e.g., sales development over the year by, say, weekly units).

For more complex comparisons, in which multiple measurements for multiple groups need to be compared, more complex graphs are required:

- A radar chart (also known as web chart, spider chart, star chart) is a graphical method of displaying multivariate data in the form of a two-dimensional chart of three or more quantitative variables represented on axes starting from the same point (e.g., the allocated budget versus actual spending of different departments, or scoring on product attributes of different designs). Sometimes it is hard to visually compare lengths of different spokes because radial distances are hard to judge, though concentric circles help as grid lines. Instead, one may use a simple line graph, particularly for time series.
- A bullet chart (developed by Stephen Few, 2006) builds upon the bar chart. The bullet graph provides a primary measure unit (e.g., year-to-year revenues) and compares this measure with other measurement units (e.g., target). It also shows the context in terms of ranges of performance—for example, "good," "average," and "weak."

10.4.1.3 Composition

When using a composition chart, the aim is to show how the data are being built up out of different subunits. In its most basic form, this results from a frequency table. For example, one might like to show the distribution of the origin of website visitors over different touchpoints (e.g., Google, Banners, Affiliates, Direct Load). There are different "composition charts" that can be used (see Figure 10.10).

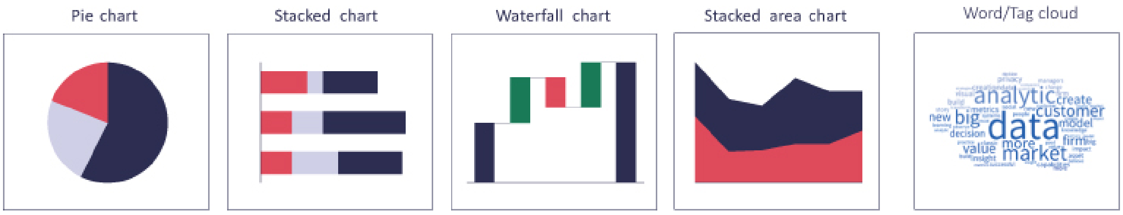


FIGURE 10.10 Composition charts

We would make the following observations about composition charts:

- The pie chart is very popular. It is very useful when a limited set of items or categories is being shown. A common mistake is to use it for many items. The pie chart quickly becomes unreadable and not informative! Some experts (e.g., Stephen Few) believe that one should never use a pie chart as, especially with multiple variables, they require a lot of space. Moreover, the pie charts might be difficult to interpret without providing exact figures.
- A stacked chart is a stapled column or bar chart. It can, for example, be used to show the distribution of sales per product per region, where the regions are shown in each bar and each bar represents a product.
- A waterfall chart is good for showing the breakdown of a variable in components. In contrast to the pie chart, this chart provides a good visualization of the size of each of the components. Moreover, it is also possible to show negative values, which is frequently impossible in many other graphs. We recommend using different colors for positive (e.g., green) and negative values (e.g., red). An example of such a chart is a breakdown of churn effects (i.e., total churn decomposed into inevitable churn, price churn, bad service churn, etc.). Extensions can also show developments over time. We frequently also use this chart to show the explanatory power of each variable in a regression equation.
- Tag/word clouds have become more popular with the increasing usage of unstructured data. Using outcomes of text mining exercises, the importance of each word can be visualized in a word cloud. The larger and the bolder the word the more frequently the word is observed. Generating these clouds is now straightforward using free online tools like Tagul, Wordle, or Tagcloud.

Some of the above charts can also be used to show changes over time. We have already mentioned the waterfall graph, but this can also be done with other graphs:

- In the stacked column chart, stapled columns per period can be linked with small lines to graphically show the time element.
- The stacked area chart is increasingly popular since this visualization can show more changes over time. In comparison with stacked column charts many more time units (even continuous, e.g., sales per week/day over a year) can be shown, thereby even showing the development of multiple stacked subunits (e.g., brands in a category). This plot can be confusing, especially when there are a number of peak periods and down periods in the data. The ratios in the data can then become unclear.

10.4.1.4 Distribution

As the name suggests, a distribution chart (see Figure 10.11) is used to display how data is distributed and to understand outliers and categories that are outside the norm. One could, for example, consider the distribution of age groups of customers, distribution of revenues across customers (is there an 80/20 rule?), or examine the power of a response prediction model.

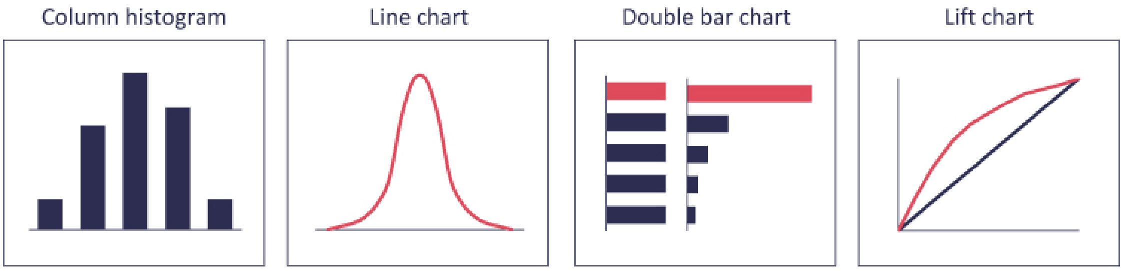


FIGURE 10.11 Distribution charts

The graphs used to show distribution are similar to those used for comparing variables. Graphs for a single variable are:

- A column histogram, which is a rather simple graph and is useful when there are a few categories per variable (e.g., age groups).
- A line histogram, which is similar to a column histogram but can handle many more categories (e.g., ages instead of age groups). In some cases, plots can be used, but in general, these are not recommended when there are large peaks and dips in the data (as discussed with the stacked area charts) because then they become hard to read.

If a researcher aims to display multiple variables and to show some distribution, the following graphs can be used:

- A double bar chart is useful if you want to compare, for example, the distribution of customers and distributions of revenues. This is also called a decile analysis.

- Lift and gain charts are a useful way to visualize how good a predictive model is. An example might be predicting direct mail response, where on the horizontal x-axis the number of customers is plotted, and the cumulative lift of the prediction model is on the vertical y-axis (see Chapter 8).

Andrew Abela (2008) has provided a nice decision process for the type of graph to be used (see Figure 10.12). Although not all the graphs that we discussed here are shown, this flow diagram can be very useful when searching for the right graph to use.

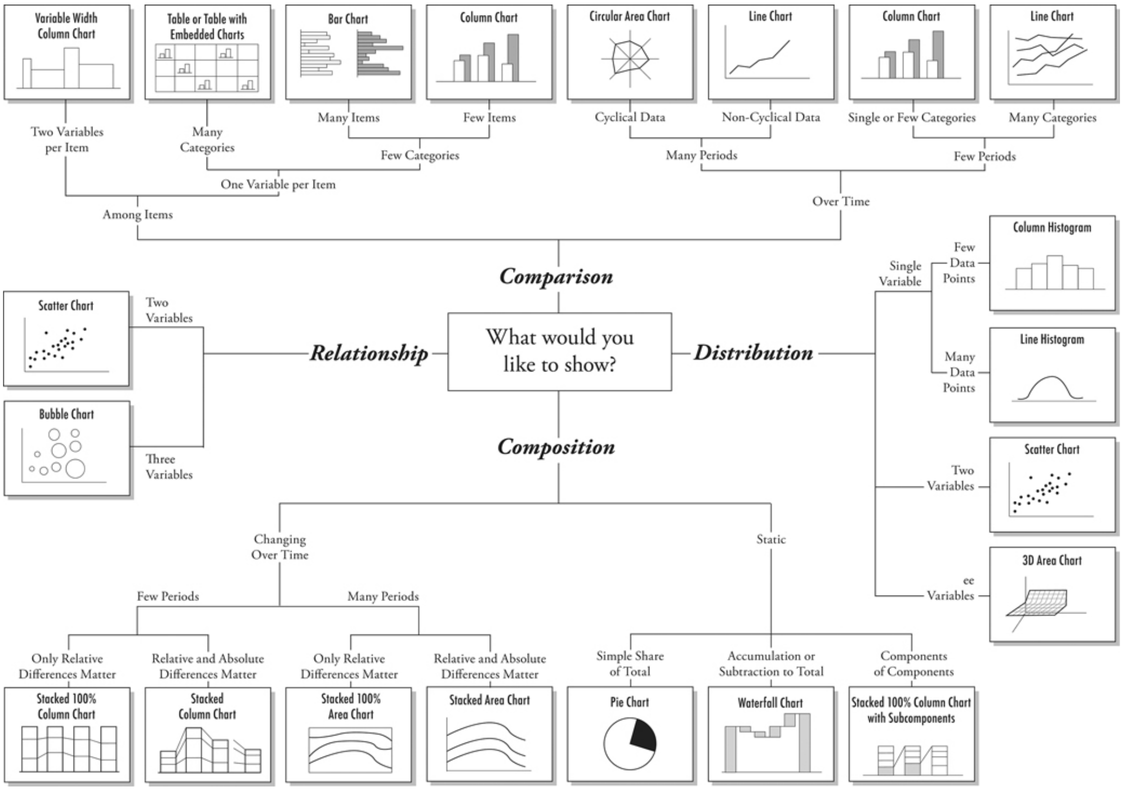


FIGURE 10.12 Chart suggestions–a thought starter

Source: Adapted from Abela (2008)

Finally, we stress that our overview of graphs is not exhaustive. We have aimed to discuss the most important and most frequently used graphs. Although care has to be taken when using these graphs, many of them can create more impact when incorporated into a report or presentation.

10.4.2 Graph design[2]

After choosing the graph type one should consider the design of the graph. This is also essential as the design will determine the attention the graph will get. Colin Ware, Director of the Data Visualization Research Lab at the

University of New Hampshire, terms the basic building blocks of the visualization process the "pre-attentive attributes" (Ware, 2008). These attributes immediately catch our eye when we look at a visualization. They can be perceived in less than ten milliseconds, even before we make a conscious effort to notice them. A list of pre-attentive attributes is given in Figure 10.13.

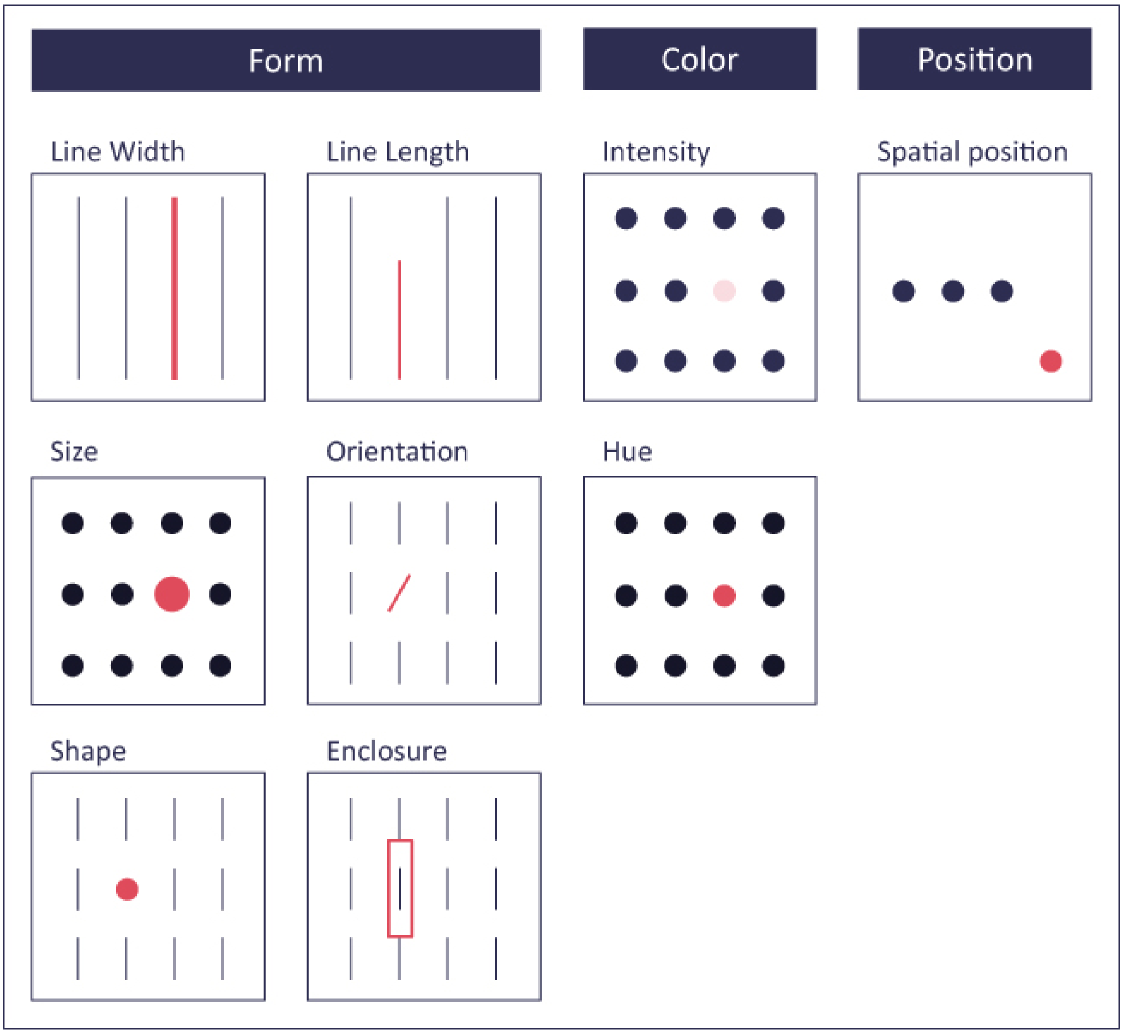


FIGURE 10.13 Pre-attentive attributes

Source: Adapted from Ware (2008)

These pre-attentive attributes can be useful when designing a graph, as they immediately identify a visual. These attributes are also the basis for patterns shown in a graph.

We have also some practical tips and tricks to further improve the visualization:

- Keep it simple! This is the golden rule. Always choose the simplest way to convey your information.

- Kill the grid lines unless they are absolutely necessary, or at least make them subtle so they do not distract from the information you're trying to present.
- Use color, size, and position to help the reader to see what is important. Color serves to highlight exceptions, not to enliven a dull dashboard.
- Your axes should be clearly labeled, and should have units on them where necessary, so no one has to guess or infer what you're trying to say.
- Use compelling headlines to describe the take-away message of the visualization.
- Remember, your goal is that anyone can pick up your chart, whether you're there to talk about it or not, and understand what information the data are trying to communicate.

10.4.3 Misleading graphs and other problems

So far, we've discussed which chart to choose and to shape in order to achieve the best visualization of analytical insights and conclusions. We also think it is important to pay attention to the interpretation of graphs, which are frequently presented in reports, presentations, news items, social media, etc. In practice, we see that with certain graph designs, there can generate misinterpretation, either intentionally or unintentionally. Misleading charts can be created deliberately to hinder the correct interpretation of data, or accidentally because they are too complex or poorly constructed. The latter also often occurs because default output of a visualization tool is simply assumed, without considering whether the representation leads to the correct conclusions. One of the first authors who wrote about misleading charts was Darrell Huff, publisher of the 1954 book How to Lie with Statistics.

In this section, we discuss seven common misleading graphs[3] that can lead to incorrect conclusions. We are going to discuss these one by one.

10.4.3.1 Truncated graphs

A truncated graph has a y-axis that does not start at zero. These graphs can give the impression of a major change where there is relatively little change. Truncated graphs are often used to highlight small differences or to save space. When used, the underlying numbers are always visually distorted, overestimating factual differences, which can lead to incorrect conclusions being drawn. In the example in Figure 10.14, the truncated graph implies that there are large differences in, for example, average sales per customer per region for an online retailer. In the representation of the original graph, these differences are relatively small.

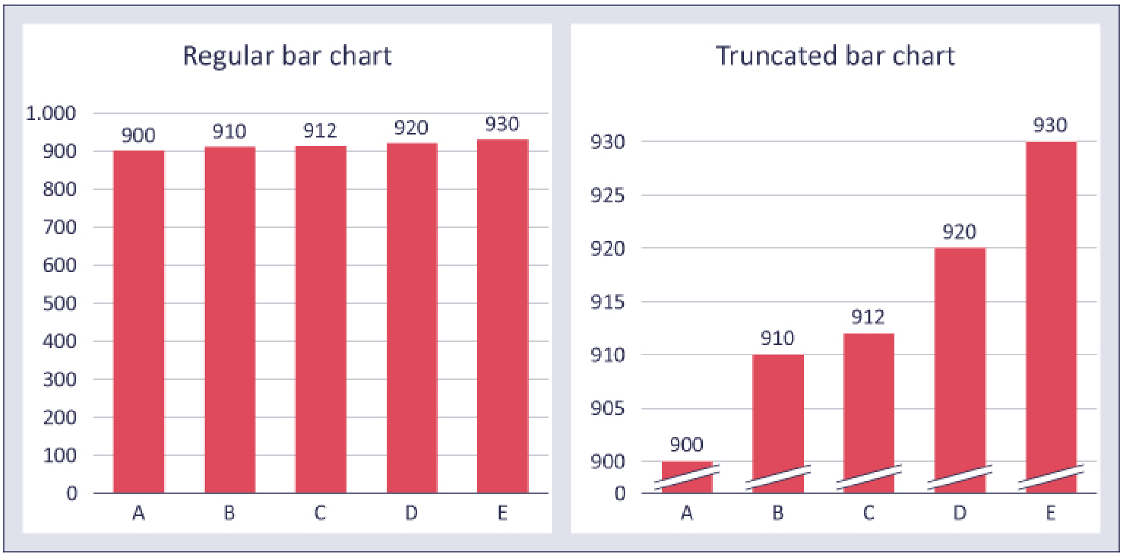


FIGURE 10.14 Comparison between a truncated and a regular bar chart

10.4.3.2 Adjusted axis

The choice of the maximum value of the y-axis affects how the graph is displayed. A higher maximum makes the chart appear to have less volatility, less growth, and a less steep line than a lower maximum. Figure 10.15 shows the influence of different choices for the maximum of the y-axis on the representation of increase in the number of visitors to a website. Compared to the original chart, the lower maximum implies a strong increase in the number of visitors over the last years. The graph with the higher maximum of the y-axis obviously suggests slower growth.

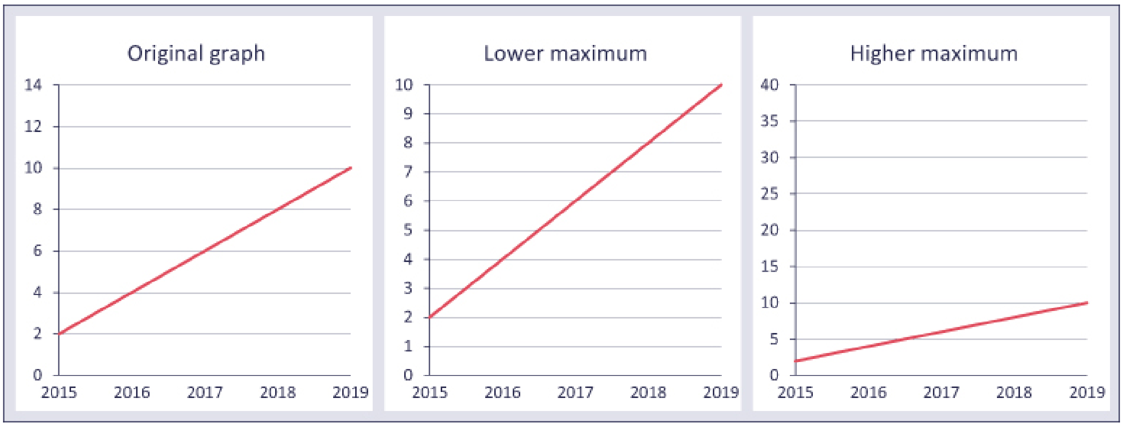


FIGURE 10.15 Different maximum value of y-axis

Another way to increase or decrease trends is to adjust the ratio of the chart dimensions. The width and height of graphs allow you to steer towards drawing certain conclusions. We often see this happen unintentionally because a graph is adjusted to the space available on a slide. In Figure 10.16 we see how big the differences are compared to the original graph by choosing half the width and twice the height and twice the width and half the height, respectively.

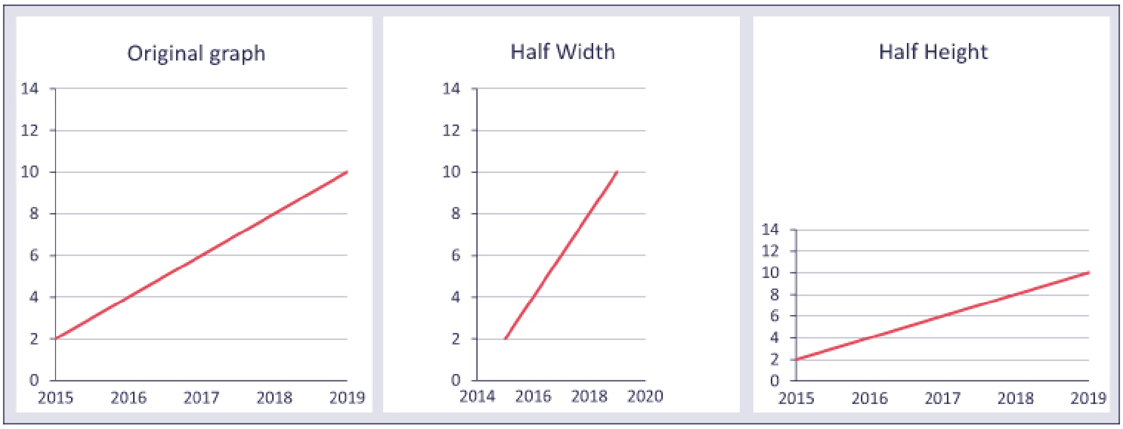


FIGURE 10.16 Adjusted height and width of the chart

10.4.3.3 Incorrect scaling

The use of pictograms in bar charts often leads to incorrect scaling. If these are scaled up uniformly in height and width, this results in a perceptually misleading comparison. The resulting scale makes the difference appear squared. The example in Figure 10.17 shows that in the incorrectly scaled pictogram bar chart, the image for B is nine times the size of A. In reality B is only three times the size of A, as shown in the right image.

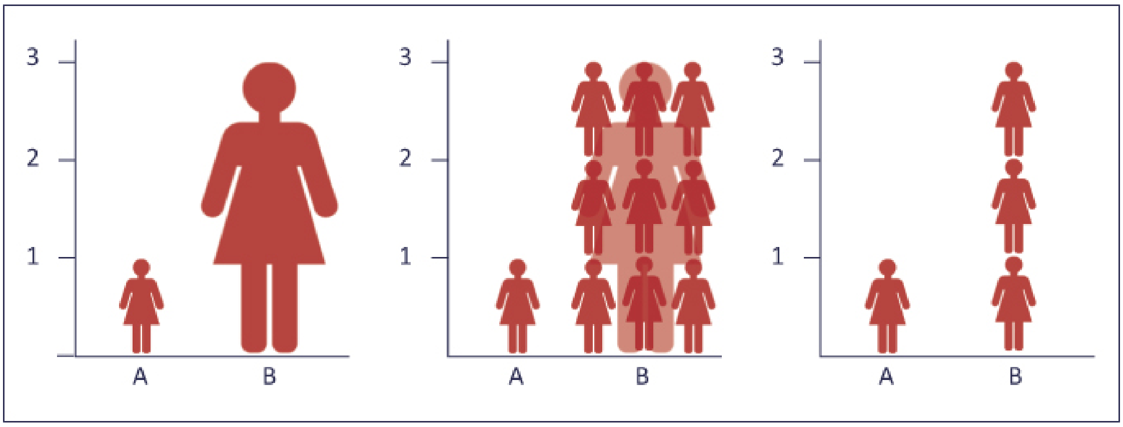


FIGURE 10.17 Incorrect scaling of pictograms

10.4.3.4 Logarithmic scaling

The advantage of a logarithmic scale is that you can plot observations that differ hugely in size, such as, for example, exponential growth. With a logarithmic scale, it is not the numerical value of a quantity itself that is given, but a logarithm of the ratio of the quantity to a reference value.

Misinterpretation of a logarithmic scale can cause vastly different values (such as 10 and 10,000) to appear close together at positions 1 and 4, if plotted on a logarithmic scale based on the 10th power. Misinterpretation of log scales can also cause relationships between quantities to appear linear when those relationships are in fact exponential. Figure 10.18 provides a comparison of linear and logarithmic scales for identical data. The top graph uses a linear scale, which clearly shows an exponential trend. The bottom graph, however, has a logarithmic scale, which generates a straight line. If the viewer of the chart is not aware of this, the chart would be interpreted as a linear trend.

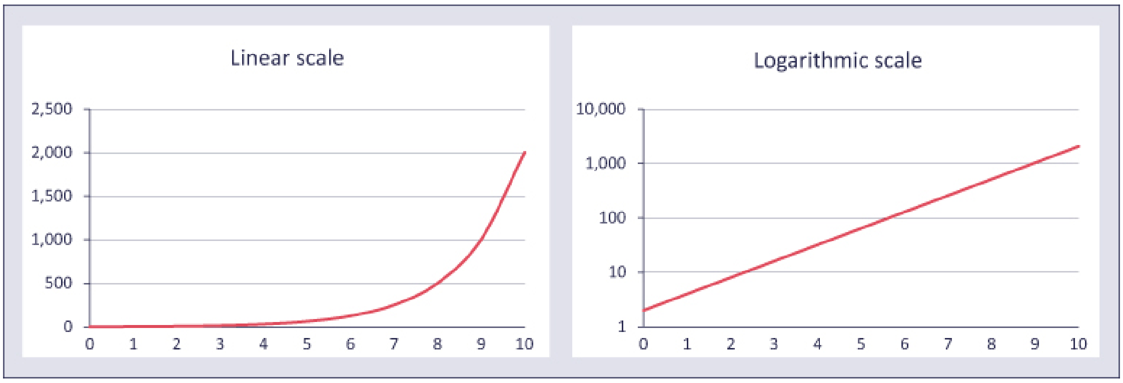


FIGURE 10.18 Comparison of linear and logarithmic scales for identical data

10.4.3.5 Omitting data

It sometimes happens that data points are removed from graphs because they are outliers and can distort an image. However, when essential data points are missing or are deliberately omitted, this can lead to false conclusions. Unfortunately, we often see in practice that data points that are not related to the conclusions you want to draw are omitted. For example, in financial reports that exclude negative interests that are not related to a positive outlook in order to create a more favorable visual impression. Figure 10.19 shows an example in which the trendline in the top graph shows a growth that is more linear with less variation than exists in reality, as shown in the bottom picture.

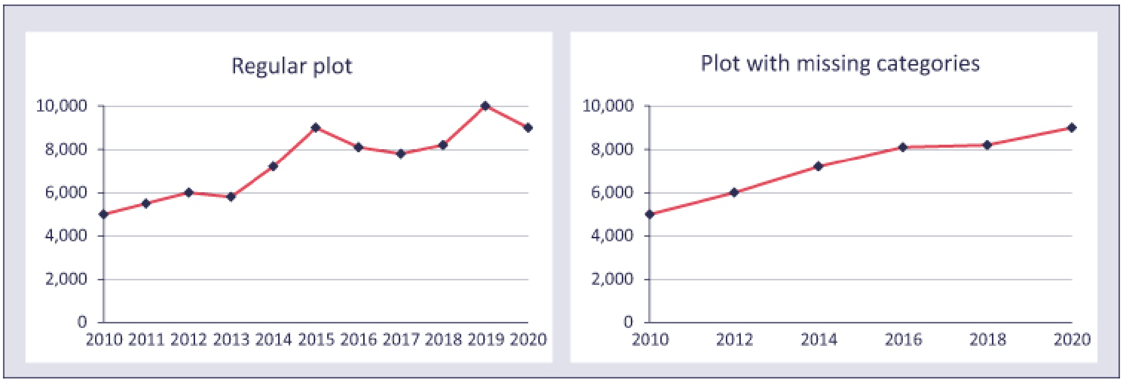


FIGURE 10.19 Effect of omitting data points

10.4.3.6 Simulated trends

We often add accents in graphs to help the reader to understand the conclusions. In the previous section we have already seen several design principles that you can use for this. Adding trend lines in time series or scatter plots is also a means for the reader to better identify effects or relationships. However, these trend lines can also be very manipulative when added in a case where, in reality, there are hardly any or no significant effects. Figure 10.20 gives an example of a scatter plot in which a positive relationship is suggested by a highly visible arrow. But if you take a closer look at the underlying scatter plot, you can already see by eye that the distribution of the datapoints shows there is hardly any correlation between the x and y values.

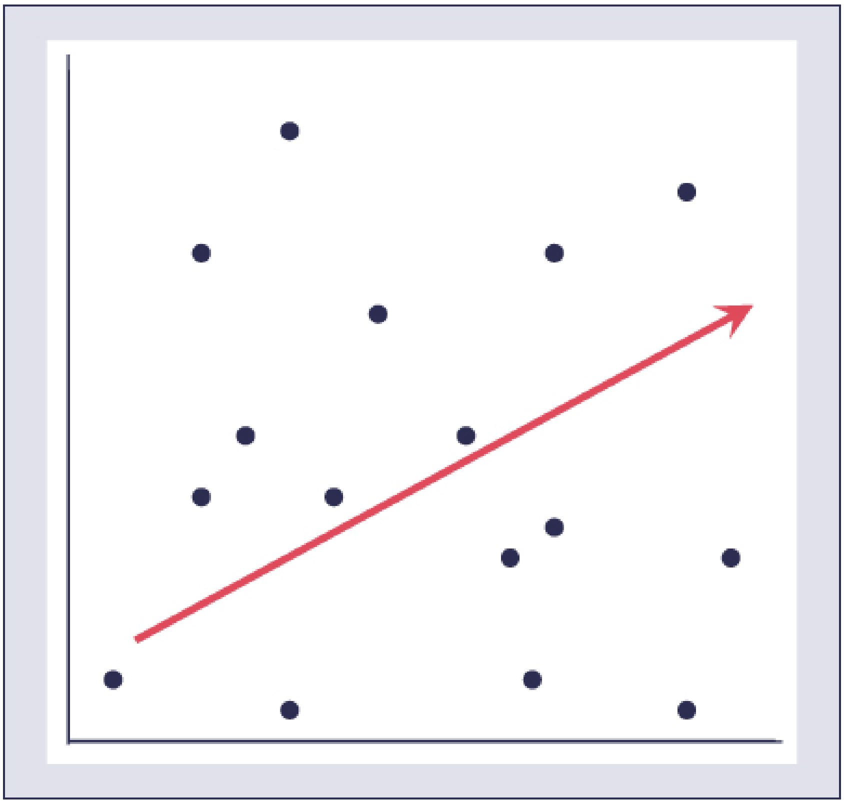


FIGURE 10.20 Simulating non-existing trends

10.4.3.7 Redundant 3D perspective

The use of an unnecessary third dimension which does not contain information, is strongly discouraged, as it may confuse the reader. 3D graphs can be unnecessarily confusing. The perspective information in the background can give the impression that it is less important than information in the front. If in bar charts the height of segments varies, interpretation becomes difficult because of the distorted 3D perspective effect. This is perhaps even more true for 3D circular diagrams. Often a 3D-look is given just for aesthetic reasons and not because it adds information. However, such a 3D effect is counter-productive because it not only makes interpretation more difficult, but also leads to misinterpretation.

The problem with a 3D pie chart is that the slices closer to the reader appear to be larger than the ones at the back. The angle at which 3D pie charts are presented, makes it difficult to judge the relative size.

Figure 10.21 compares 2D and 3D pie charts based on the same data. In the misleading 3D pie chart, item C appears to be at least as big as item A, when in reality it is less than half the size.

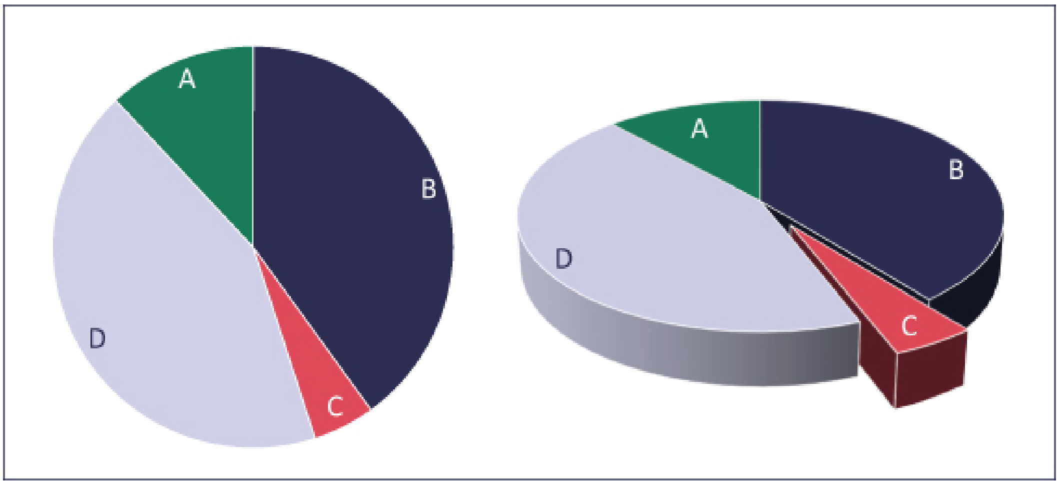


FIGURE 10.21 Comparison of 2D and 3D pie charts based on the same data

10.5 TRENDS IN VISUALIZATION

We conclude with a discussion of some trends in visualization. In practice and in line with the growth of big data development we observe a stronger focus on design. We also observe that more infographics are being used. David McCandless is taking infographics and data-visualization to the next level. In his book *Information is Beautiful* (2010) he visualizes captivating and intriguing patterns and connections across art, science, health, and pop. Analysts frequently lack the skill for this kind of work and professional design artists are being used. We also see an increasing use of text-based graphs, such as word-clouds.

We are probably aware of only some of the visualization trends owing to the technological advances that are becoming available to aid display of visual effects in a dynamic way. Just watch the way Professor Hans Rosling handles a presentation, commentating on a moving hologram that illustrates the health, wealth, and population of 200 countries over 200 years in less than a minute (Rosling, 2007). We also believe that the growing importance of video means that presentations will become more video-based.

10.6 CONCLUSIONS

In this chapter we have put forward the proposition that low impact is a general problem for many analytical studies. A very clear storyline and visualization are key ingredients for creating more impact. In sum, we have the following clear recommendations for analysts; if they are followed, the result should be a storyboard in which the storyline and the visuals are integrated.

- Start the presentation by creating a clear storyline.
- Write the storyline in full sentences:
  What is the situation/complication?
  What is the core message?
  How can this message be underpinned?
- Continue by drawing some initial slides and visuals. Do not use a computer; use a drawing board to stimulate creativity.
- Write out the headings of each slide in full sentences.
- Choose the right graphs to visualize the supporting insights.
- Using this basis, make a report/presentation.
- Use a critical colleague or friend to challenge the presentation.

Finally, be wary of misleading charts. They can lead to incorrect conclusions. To avoid misinterpretation, it is always important to take a critical look at the axes (truncated, choice of minimum and maximum value, use of deviating or irregular scaling), possible missing or omitted values, redundant dimensions, misleading accents and trend lines.

ASSIGNMENTS

Assignment 10.1: Vodafone press release

Carefully read the article below:

Source: Article about the Vodafone Group from the Financial Times, 22 July, 2016

"Vodafone has benefited from rebounding demand for its mobile services across Europe although the UK-based telecoms group continues to struggle in its home market.

Vodafone on Friday revealed that revenues from mobile services had grown in the first quarter at a higher rate than expected by analysts, sending its shares up more than 4% in morning trading to 235.1p per share.

Overall group service revenues grew 2.2% in the three months to the end of June, the telecoms group said on Friday, beating analysts' expectations. Group service revenues are the company's preferred measurement of sales success for its mobile operations. The metric takes into account access charges and roaming fees but excludes the sales of handsets.

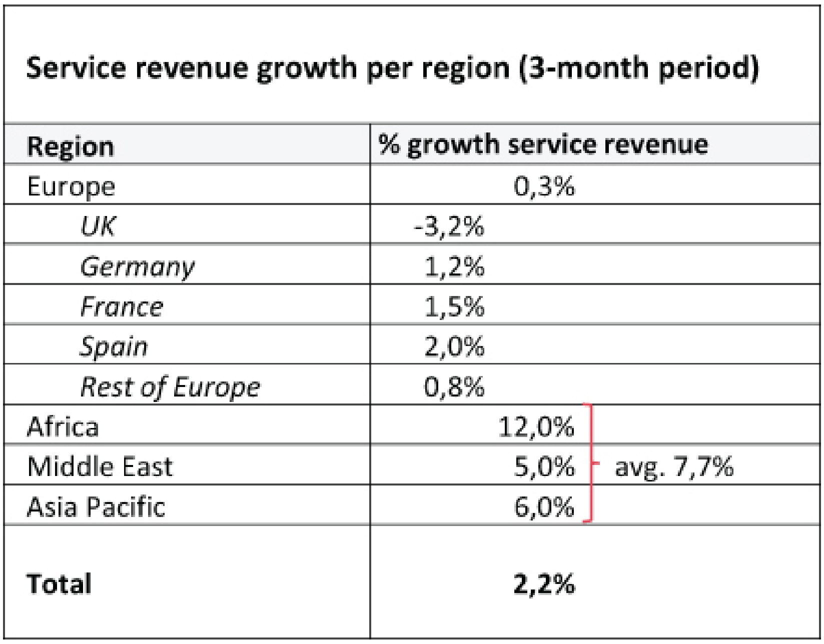
In Europe, Vodafone posted 0.3% service revenue growth over the three-month period, which continued the slow recovery from a five-year decline in the region that only ended last year.

Strong demand for mobile services in emerging markets remained the driver for growth for the group, however, with service revenues in Africa, the Middle East and Asia Pacific climbing 7.7%.

In the UK, however, Vodafone posted a 3.2% decline, which it blamed on the "impact of operational challenges" of a move to a new billing system. Vodafone has long struggled with customer complaints in the UK. In March it was revealed that the telecoms company was the most criticised pay-monthly mobile provider in the UK…"

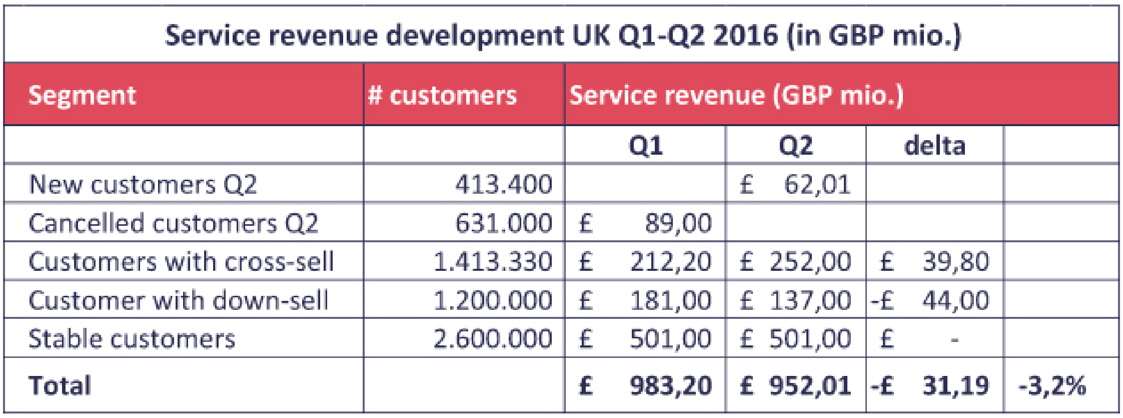
Questions:

1. What is an appropriate title demonstrating the key message of this article?
2. State at least two reasons for this core message.
3. Suppose you had access to the monthly turnover data for the Vodafone Group for 2016 With what kind of graph would you visualize this development of the turnover over time?
4. Based on the results in the UK, there seems to be a relationship between customer satisfaction and the turnover rate, what kind of graph would best visualize this?
5. Figure 10.22 below gives data to explain the 2.2% growth. Create a graph to visualize the differences by region. Show the differences within Europe. Justify your graph in words.
6. In the United Kingdom, complaining customers are the main cause of the negative turnover development. Figure 10.23 shows the development of the turnover during the second quarter of 2016. Make a graph to illustrate this development and explain it in words.
7. Based on the data/graphs, what conclusions can you draw about the causes of the negative turnover development in the UK? Do you think that this is well-substantiated in the article?



Service revenue growth per region (3-month period)		
Region	**% growth service revenue**	
Europe	0,3%	
UK	-3,2%	
Germany	1,2%	
France	1,5%	
Spain	2,0%	
Rest of Europe	0,8%	
Africa	12,0%	
Middle East	5,0%	avg. 7,7%
Asia Pacific	6,0%	
Total	**2,2%**	

FIGURE 10.22 Service revenue growth per region



Service revenue development UK Q1-Q2 2016 (in GBP mio.)					
Segment	**# customers**	**Service revenue (GBP mio.)**			
		Q1	**Q2**	**delta**	
New customers Q2	413.400		£ 62,01		
Cancelled customers Q2	631.000	£ 89,00			
Customers with cross-sell	1.413.330	£ 212,20	£ 252,00	£ 39,80	
Customer with down-sell	1.200.000	£ 181,00	£ 137,00	-£ 44,00	
Stable customers	2.600.000	£ 501,00	£ 501,00	£ -	
Total		**£ 983,20**	**£ 952,01**	**-£ 31,19**	**-3,2%**

FIGURE 10.23 Service revenue development UK Q1-Q2 2016

Assignment 10.2: Remove the errors from the graph

On August 16, 2018 (6 pm. to 8 pm.), there was an item from the bureau of economic policy analyses (CPB) in a news broadcast on Dutch television (NOS) about the latest economic prospects. One suggestion was that unemployment would continue to decline. The NOS made a graph for this. Unfortunately, it was not a good graph (see Figure 10.24). There were at least three major mistakes in it and more errors to mention. Name these mistakes and describe how you can correct them.

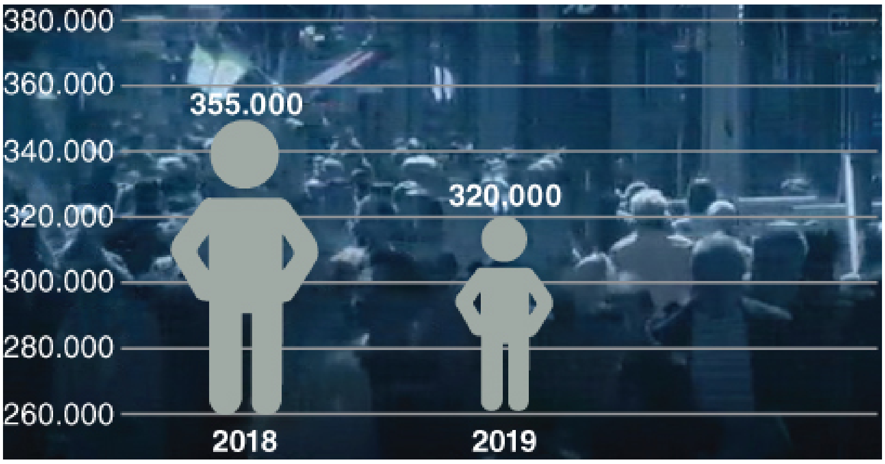


FIGURE 10.24 Displayed graph in a news broadcast of Dutch television (NOS)

Source: https://peilingpraktijken.nl/weblog/2018/08/deze-grafiek-van-de-nos-zitten-minstens-vier-fouten/

NOTES

1. There are other methods as well. However, we find that this method provides several advantages and focus on it in this chapter.
2. This section draws heavily on Colin Ware's study (2008).
3. Adapted from https://en.wikipedia.org/wiki/Misleading_graph

REFERENCES

Abela, A. (2008). *Advanced Presentations by Design: Creating Communication That Drives Action*. San Francisco: Pfeiffer.

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, *27*(1), 17–21.

Few, S. (2006). *Information Dashboard Design*. Beijing: O'Reilly.

Huff, D. (1954). *How to Lie with Statistics*. New York: Norton.

McCandless, D. (2010). *Information Is Beautiful*. London: HarperCollins.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.

Minto, B. (2009). *The Pyramid Principle: Logic in Writing and Thinking*. Edinburgh Gate, Harlow, Essex: Pearson Education.

Paivio, A. & Csapo, K. (1973). Picture superiority in free recall: Imagery or dual coding? *Cognitive Psychology*, *5*, 176–206.

Roberts, J. H., Kayandé, U., & Stremersch, S. (2014). From academic research to marketing practice: Exploring the marketing science value chain. *International Journal of Research in Marketing*, *3*(2), 127–140.

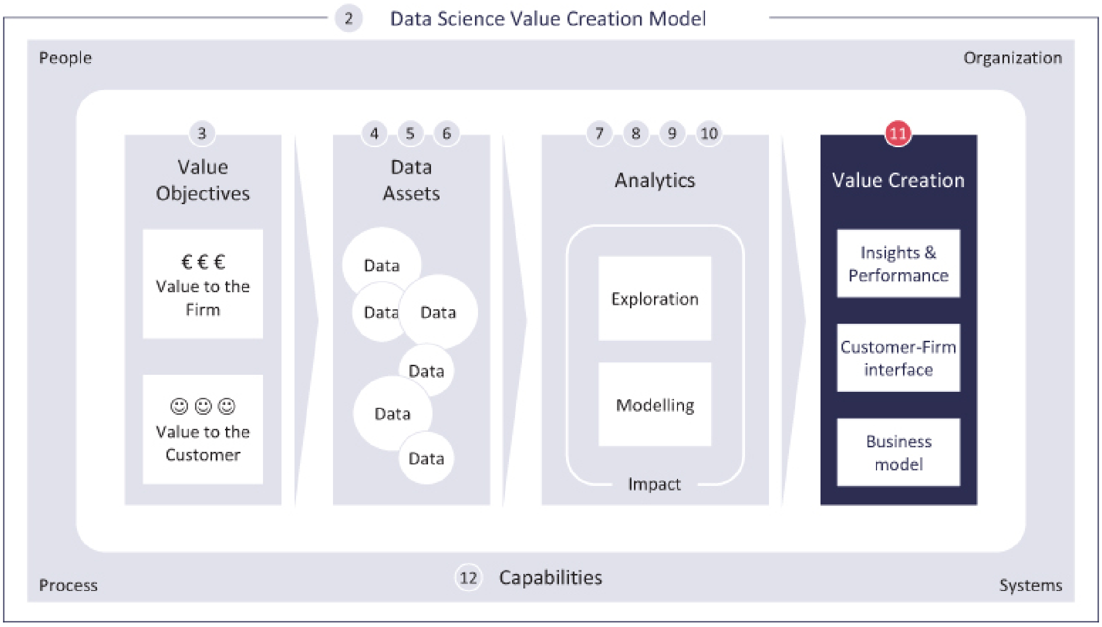
Rosling, H. (2007). New insights on property. Retrieved from TED.com September 11, 2015 www.ted.com/talks/hans_rosling_reveals_new_insights_on_poverty?language=en

Ware, C. (2008). *Visual Thinking: For Design*. Morgan Kaufmann Series in Interactive Technologies. Amsterdam: Elsevier.

Wurman, R. S. (1989). *Information Anxiety*. New York: Doubleday.

CHAPTER 11
Creating value with data science

DOI: 10.4324/9781003011163-11

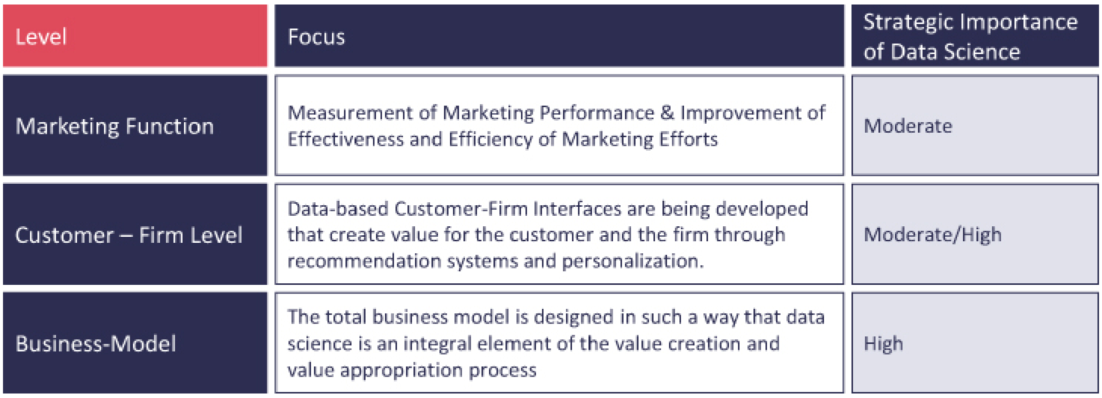


11.1 INTRODUCTION

So far, we have mainly discussed the different building blocks for creating value with data science. We have started with a discussion on what value entails, distinguishing between Value to the Customer (V2C) van Value to the Firm (V2F). Next, we focused on the important role of data and how to manage these data. Subsequently, we focused on analytics and specific data science methods, also emphasizing the importance of story-telling and visualization. An important next step is how to create value using data science and benefiting from the analytics carried out. In this chapter, our main aim is to describe how firms can do this. We begin with a conceptual classification of how firms can create value and then discuss several examples of how firms managed to create this value. Subsequently, we discuss the opportunity finding methodology that can be used to enable data science to have a strong value impact and organizational impact.

11.2 DATA SCIENCE VALUE CREATION

Firms should thus consider how they aim to create value with data science. Specifically, they should choose the strategic level within the firm at which data science plays an important role. We consider two levels (see Figure 11.1):



Level	Focus	Strategic Importance of Data Science
Marketing Function	Measurement of Marketing Performance & Improvement of Effectiveness and Efficiency of Marketing Efforts	Moderate
Customer – Firm Level	Data-based Customer-Firm Interfaces are being developed that create value for the customer and the firm through recommendation systems and personalization.	Moderate/High
Business-Model	The total business model is designed in such a way that data science is an integral element of the value creation and value appropriation process	High

FIGURE 11.1 Value-creation with data science on different strategic levels

1. Functional level of marketing
2. Customer—Firm Interface Level
3. Business Model Level.

11.3 VALUE CREATION AT MARKETING LEVEL

Data science applications at the marketing level have a long history. However, we also observe that due to strong developments in data, they have become more prominent in the last decade. We consider three main objectives of data science at the marketing function level:

1. Insights delivery
2. Marketing performance measurement
3. Effectiveness and efficient improvement of marketing efforts.

11.3.1 Insights delivery

Firms constantly require information and insights on market developments, changing brand reputation, and customer behavior. Data science plays an important role in delivering these insights. This can be done by just reporting data statistics on, for example, the adoption rate of newly introduced products by specific customer segments. Insights development could also use more advanced models. For example, in order to segment the market, segmentation analyses can be carried out that show the existence of specific market segments. The insights can be used to develop marketing tactics and strategies. For example, the segmentation analyses can be used as input for a strategic marketing plan where firms choose a segmentation and target specific segments with specific offerings. In general, these insights are very important in developing market and customer-focused strategies that can create a stronger performance for firms (Verhoef & Leeflang, 2009).

11.3.2 Marketing performance measurement

Firms thus require insights and data on how they perform and how to influence their performance. Not only do these data need to be collected constantly, but they should also be displayed in a marketing dashboard. Marketing dashboards have become rather popular and are defined by Pauwels *et al*. (2009: 3) as: "A relatively small collection of integrated key performance metrics and underlying performance drivers that reflects both short and long-term interests to be viewed in common throughout the organization." Pauwels *et al*. (2009) and Reibstein *et al*. (2005) propose five stages of dashboard development:

1. Selecting the key metrics
2. Populating the dashboard with data
3. Establishing relationships between the dashboard items
4. Forecasting and "what if" analysis
5. Connecting to financial consequences.

Observing these steps, we can generally derive two main objectives of a marketing dashboard. First, it reports the results on some key metrics (e.g., retention rates or NPS) on a regular (e.g., monthly) basis. Second, the results of models are included in such a way that managers can check what happens if they take specific actions. For example, they might be keen to know what happens with customer equity if they improve their service (Rust, Lemon, & Zeithaml, 2004). Dashboards are thus filled with data and final model results, which can be used to execute "what if" analyses.

Data science becomes more relevant when firms also aim to understand how their marketing efforts influence marketing KPIs and how these KPIs are linked, which is done in step 3 of dashboard development. For example, for a large German insurance company Lesscher, Lobschat, and Verhoef (2021) show that direct mailing affects both online search metrics (i.e., branded search, website visits) and also insurance sales. They conclude that direct mailings are still important for creating sales in a digitalizing world. To achieve these insights, they executed two experimental studies and collected online search data and sales data that are relevant in search (i.e., Google search data), consideration (i.e., website visits), and purchase. Furthermore, they use regression models (see Chapter 9) to estimate the effects of direct mailings on these metrics and the interrelationships between online search metrics and sales. In Figure 11.2 we show those interrelationships. Note, that in the second study the impact of display advertising and direct mailings is also studied. In marketing literature, there are many examples of how analytics can be used to assess the effectiveness of marketing efforts.

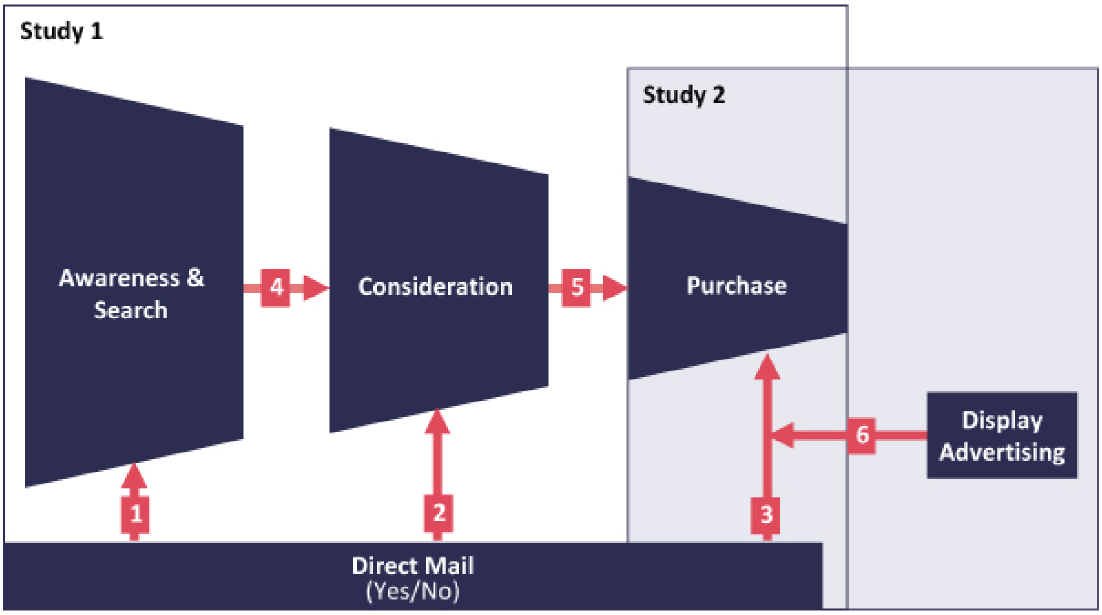


FIGURE 11.2 Effect of direct mailing in the purchase funnel from search to purchase

Source: Lesscher, Lobschat, and Verhoef (2021)

Within firms there is also a continuous focus on collecting the right KPIs and building marketing dashboards. We describe such an example for an insurance company in case 11.1. In this case, which builds on the earlier case in Chapter 5, we describe how a firm collects several brand KPIs for customer segments and how this information is used to increase the effectiveness of specific marketing investments. The main focus though, is on collecting the right KPIs and creating smart insights on the performance of different customer groups and brands. Furthermore, the firm was able to gain insight into how specific marketing investments affect their KPIs. The case clearly shows how firms can learn from insights and improve their marketing strategies.

CASE 11.1: MARKETING PERFORMANCE MEASUREMENT AT AN INSURANCE COMPANY

Situation

An insurance company with multiple brands has traditionally focused on a broad target group. Despite this broad focus, the customer base is not representative of the total market population, being overpopulated by elderly, high-income households. The insurance market is in decline and competition is fierce, with a move from traditional high-cost channels (via an intermediary or call center) to the lower-cost online channels (comparison sites, direct conversion on websites, etc.). Competitors are spending substantially more on media than they used to. This puts pressure on the market share. To compete better and to sustain or (even better) improve market share there is a need to substantially increase the effectiveness of current marketing investments. To realize this, a holistic marketing approach in combination with a granular perspective of the market and customers is desired. Furthermore, insights are needed into the performance per customer profile on brand, proposition, distribution, and pricing of the current marketing mix.

Complication

In realizing the above, three complications came up that made this a big challenge for the organization. The first complication was organizational: the responsibilities for the market, brand, pricing, and distribution were dealt with in different silos within the organization, not necessarily working closely together. Because of this, it was quite impossible to realize an aligned holistic and consistent marketing approach for the total marketing mix. The second challenge concerned the data sources needed to measure and improve performance. This was further discussed in Chapter 5. The third complication was the lack of a framework or integral segmentation that could serve as a common denominator for all the data sources and KPIs in scope. Within the organization several breakdowns of the market and the customer base were available; however, they were not aligned.

Key message

As a solution for the complications described above, an interactive, visually attractive big data dashboard was developed that was easy for both marketers and analysts to use (see Figure 11.3). This enabled a shift from separate measurements of the effect of marketing investments and value KPIs to detailed and dynamic measuring, interpreting, and forecasting of the marketing performance. In this way, an increase in performance and thus return on investment could be realized with a granular perspective on customer profiles

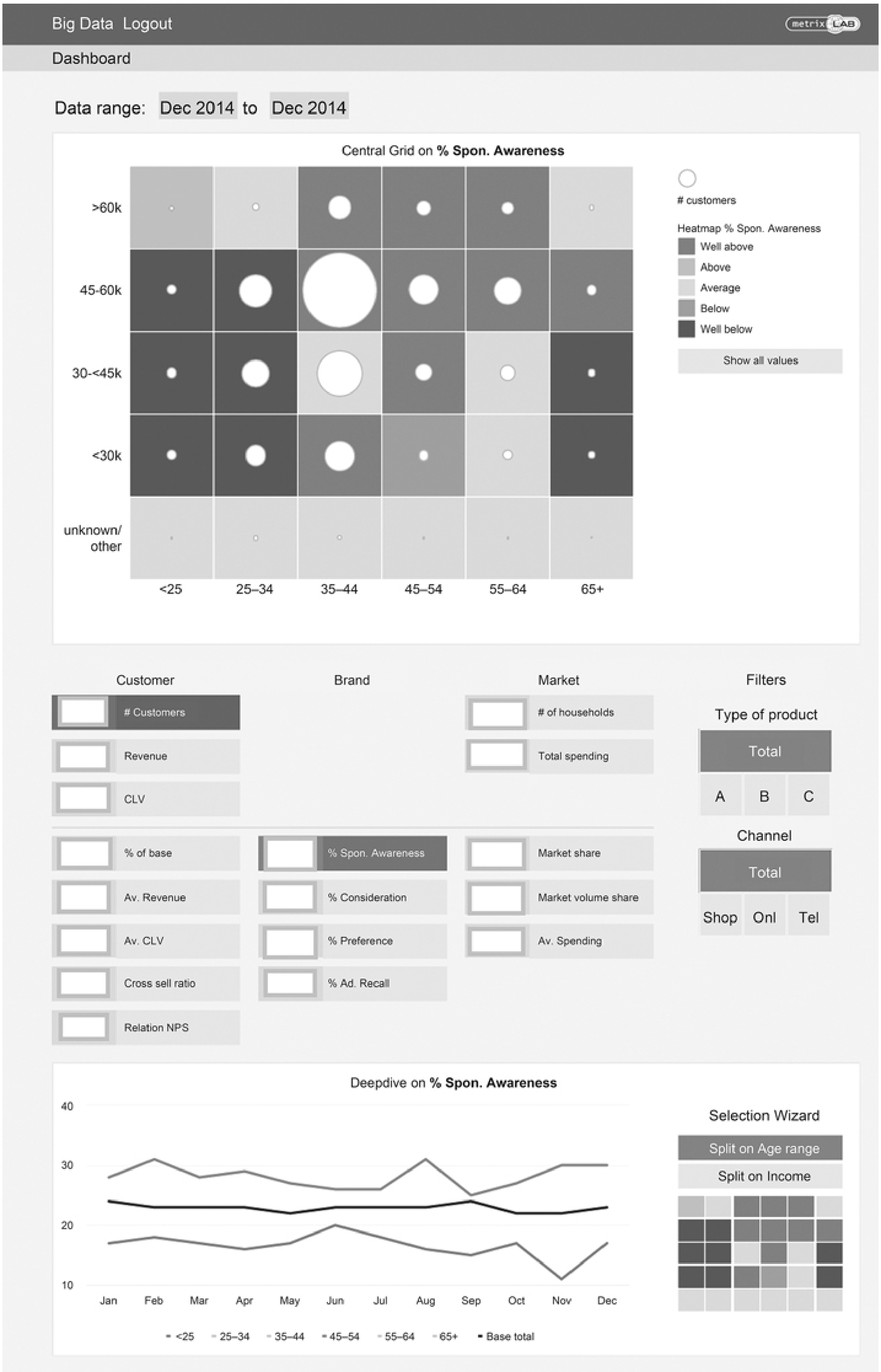

Data range: Dec 2014 to Dec 2014

FIGURE 11.3 The big data dashboard

The solution consisted of several elements that were crucial for success. In the dashboard, the central element was a heat map that showed a segmentation that was consistent for all KPIs in all data sources. The colors in the heat map were determined by the over or underrepresentation of a specific KPI for a specific segment in the heat map. The different KPIs were clustered around the customers, brand, and market. Furthermore, filters were added to make a deep dive possible on specific products and/or channels and/or time frames.

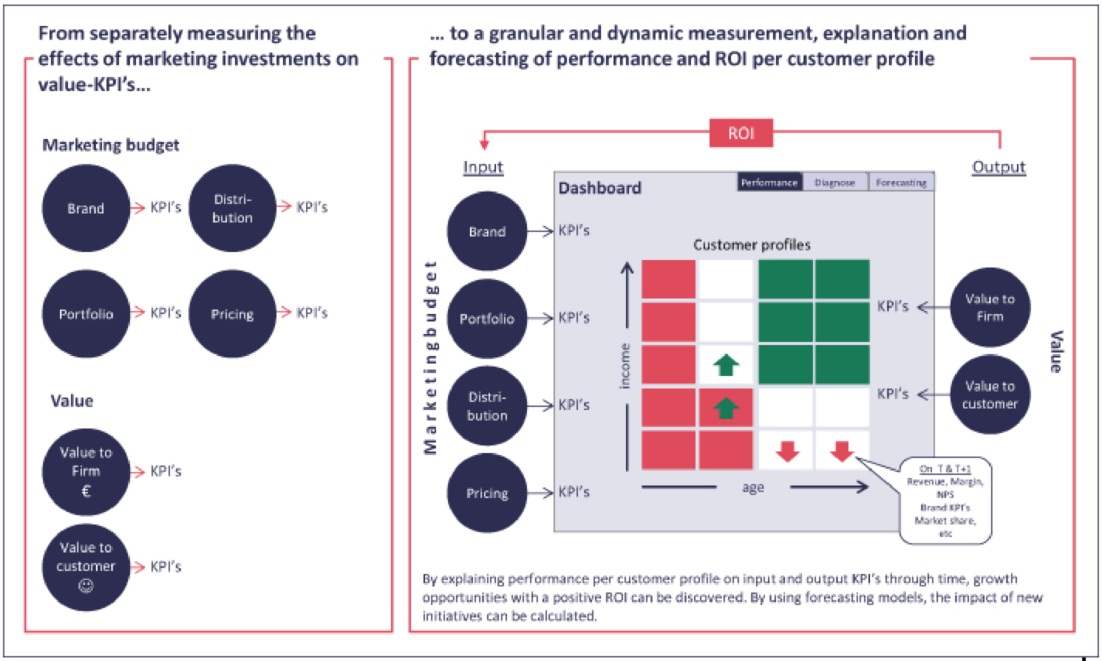
Results

The results created by realizing the big data dashboard approach were:

- Consensus on the set of KPIs to be measured, in order to measure the total marketing performance.
- A better understanding of total marketing performance and its effectiveness, per KPI and per segment, and also of the relationship between the KPIs.
- Substantial opportunities (multimillion euros) for initiatives to improve the marketing performance and marketing ROI.

Model used

At the start of the project, we developed a conceptual framework to visualize the shift to be made (see Figure 11.4). In this framework, we showed that, ideally, the effectiveness of the spending of the marketing budget would become visible in what we called the input KPIs. However, we also showed that the input KPIs are just an intermediary step. Linking the performance of the input KPIs to the defined output KPIs (measuring V2C and V2F metrics), by splitting them out across the segments in the heat map, should make the relationships visible.



From separately measuring the effects of marketing investments on value-KPI's...

... to a granular and dynamic measurement, explanation and forecasting of performance and ROI per customer profile

By explaining performance per customer profile on input and output KPI's through time, growth opportunities with a positive ROI can be discovered. By using forecasting models, the impact of new initiatives can be calculated.

FIGURE 11.4 The conceptual model for the holistic approach

Insights

Analyzing the performance, using the dashboard, in several interactive sessions with marketers and analysts showed that, for different reasons, different performance levels occurred in the segments of the heat map. This suggests that different strategies are required to improve the performance of specific customer profiles. One of the key insights was that performance with youngsters was low (low market share) due to branding issues. At the same time, we saw high in- and outflow from this segment, mainly due to aggressive campaigning targeted at this group in a certain time frame with only one specific product. The initiative aimed at this group included more selective targeting (only the potential loyal customers) and was at the same time more focused on offering a broader product range to create a more sustainable relationship. Another insight was that due to delays in updates of the pricing module (used for calculating a good offer to potential new customers) a lot of deals were lost, especially in the high-income family segment, as a result of over-conservative pricing by the outdated pricing module. Because the segment at stake is high value and high volume a slight improvement would result in a significant (euro) potential.

Improving marketing effectiveness and efficiency

Data science is also important in more effectively and efficiently using marketing instruments. Firms have to allocate their marketing budget among marketing instruments and touchpoints, and between customers. They should spend their market budget where it has the highest return. This can of course be done based on the judgments of the marketer, but data-based insights and model-based decisions should lead to better decisions. With the increasing amount of data, this has become possible for many firms. We distinguish between two levels of decision-making:

1. Marketing-instrument level
2. Customer (segment) level.

Marketing instrument level

At the marketing instrument level, firms decide how much to invest in a specific marketing instrument to create an effect on, for example, brand awareness or sales. For this purpose, analysts will build market response models looking at the effects of marketing instruments on, for example, sales (e.g., Leeflang *et al.*, 2015). Based on these insights one can assess, for example, advertising elasticity and how much should be invested in advertising to produce a specifically targeted increase in sales. One can also compare the effects of different marketing instruments. For example, Wiesel, Pauwels, and Arts (2011) show, using time-series models and a field experiment, that for a B2B firm selling office equipment, shifting the budget from firm-initiated offline activities (i.e., catalogs) to customer-initiated offline activities (i.e., Google search) improves profitability by 14%. Similarly, Konuş, Neslin, and Verhoef (2014) show that eliminating the catalog channel reduces customer revenues but increases profitability as a costly sales channel is dropped. These examples demonstrate that data science insights help to create value for firms, with increasing profitability. It should be noted, however, that these models tend to optimize V2F metrics, such as sales and ROI. Data science can also be used to understand drivers of V2C metrics, such as overall satisfaction and loyalty. For the Dutch railways, Verhoef, Heijnsbroek, and Bosma (2017) report the effects of different service attributes of trains on customer satisfaction, showing the importance of, for example, punctuality and availability of seating. Based on these insights the Dutch railways can decide which attributes to focus on to improve customer satisfaction.

In today's digital world data science is also becoming very important for assessing the impact of different digital touchpoints on the purchase. Firms want to know the impact of touchpoints that are being used during the customer journey on purchase outcomes. Firms will typically only have limited information on the total usage of touchpoints on this "path to purchase" and will typically only observe the ones that are used to visit the retailer's website. Consequently, the question is whether the sale can be attributed to this last-used touchpoint or to other touchpoints, which are generally not observed. It is essential to know this for the purpose of allocating scarce marketing budgets among touchpoints. It is also important to make deals with partners, such as Google or comparison websites. To assess the effects of customer touchpoints on conversion and sales, attribution modeling can be used. A logistic regression model (see Chapter 9) can be used to estimate the effect of each touchpoint on purchase. However, one concern here is that in some channels some customers have a higher inherent readiness to purchase. This would imply endogeneity of the touchpoints and one should correct for this. This is not straightforward, as more information is required and this information is frequently not available. In Case 11.2 we show how touchpoint usage on the last purchase occasion can function as a way to correct for the endogeneity in touchpoint usage. As expected, contributions to the purchase of specific touchpoints, such as search engines, are corrected downwards. Using these insights, a firm can choose to invest more in specific touchpoints, such as e-mails, given their higher effectiveness.

CASE 11.2: ATTRIBUTION MODELING AT AN ONLINE RETAILER[1]

Situation

An online retailer selling multiple products was using multiple acquisition touchpoints and media to attract customers to its website. They wanted to improve their allocation of budgets over these touchpoints and media. To do so, they needed to know how many customers are attracted per touchpoint/medium and the conversion rate per touchpoint/medium. They had data on the last used touchpoint/medium when customers entered the website, whether these customers made a purchase (conversion), and how large that purchase was (order size).

Complication

The main complication here is modeling the effectiveness of each touchpoint/medium. The retailer traditionally attributed the sale to the last used touchpoint/medium. This is also known as the last-click method. That means that if a customer arrives at the website through a search engine and then subsequently buys a product, the value of this purchase is attributed to the search engine. It is unlikely that this easy method is good, as customers use multiple touchpoints and media and might be influenced by any or all of them. Additionally, touchpoints are used in multiple phases of the purchase funnel. Hence the common belief is that the last-click method leads to exaggerated attribution values. The challenge is thus to achieve a better attribution method.

Key message

For this retailer, we developed a new, more accurate attribution method that could be used to improve the allocation of marketing resources over touchpoints and media. It turned out that a simpler method than originally assumed works as effectively as the more complicated model first developed.

Results

A more exact estimate of the true value of a touchpoint/medium can now be assessed than was possible using the traditional last-click method in which the sale is attributed to the last touchpoint used. This holds especially true for search engine advertising, search engine organic searches, and direct loads on the website. The value of an email is underestimated using last-click. An Excel tool was developed to implement the new attribution method on a daily basis.

Model used

We modeled the effect of used touchpoint/medium on conversion and order size. One main issue with attribution is that the touchpoint/medium used can be endogenous (see Chapter 8). To correct for this, we used instrumental variables and controlled for some background customer characteristics. The touchpoint medium used in a product category different from the one currently visited is used as the main instrumental variable (see Figure 11.5).

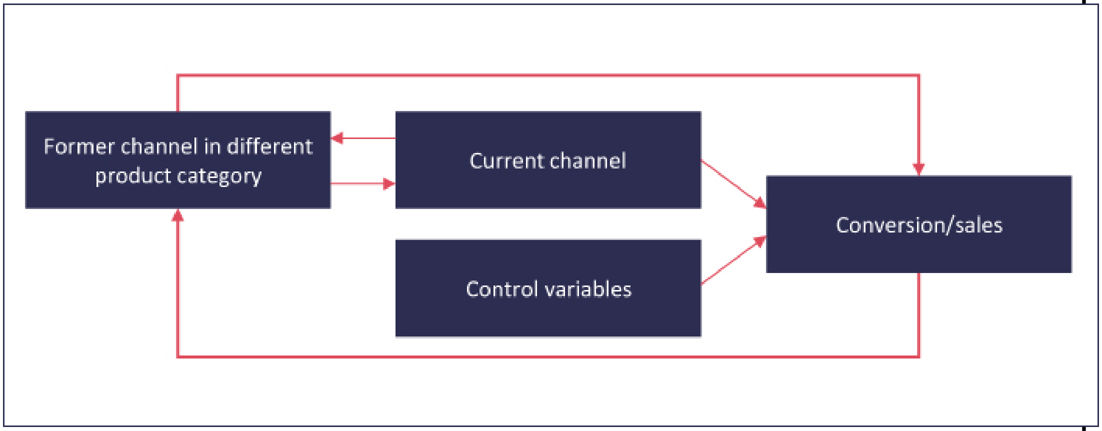


FIGURE 11.5 Visualization of model being used

Insights

The last-click method usually overestimates the value of touchpoints/media. We first compared the results of last-click with a model accounting for endogeneity. In general, the new model results in lower conversion rates and average order sizes per touchpoint/media. Only for email are effects stronger in the new model (see Figure 11.6)

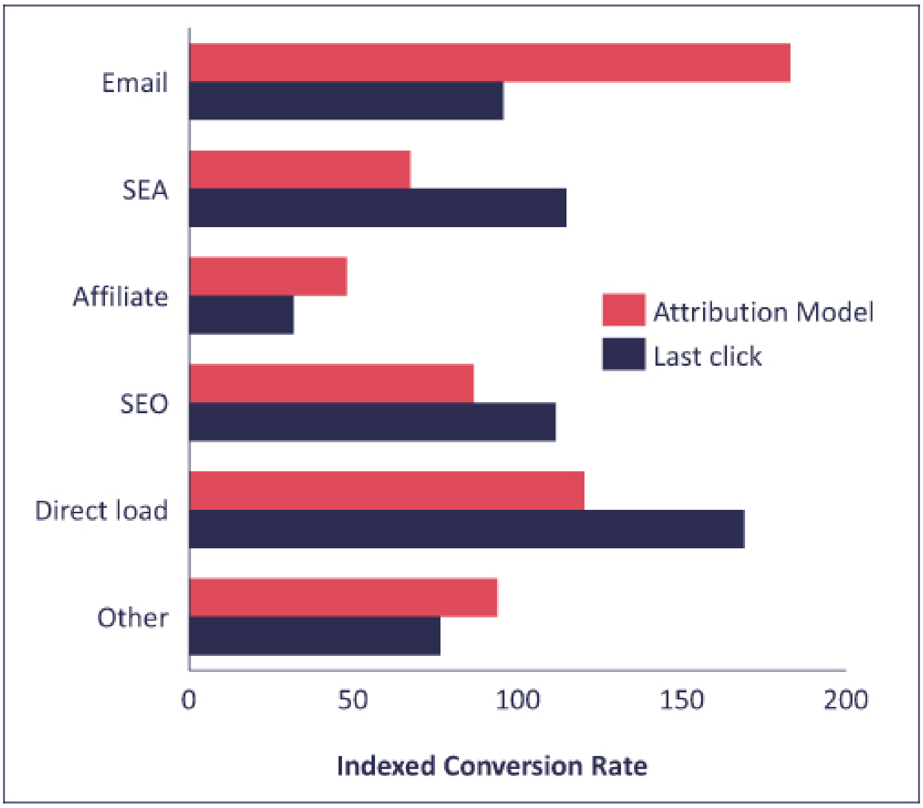


FIGURE 11.6 Comparison of effects for attribution model and last-click method

This study also achieved some collateral catches, for example:

- The sooner a customer gets back on the website the higher the conversion rate. This suggests strong opportunities for re-targeting of non-converted customers.
- Prior mobile sessions improve the conversion of consequent web sessions.[2]
- Conversion rates are highest for VIP customers.
- Young female customers tend to have a higher conversion rate.

Success factors

The development of the attribution model benefited from the following factors:

- The firm was able to deliver solid data.

- The models were estimated at the individual level instead of the aggregate level, which was done in a previous project. This firm is used to working at the individual level in their web-analytics and this way of modeling matches their way of working and thinking.
- There was openness to the use of complicated models, while at the same time simpler models were preferred if possible.

11.3.3 Customer level

At the customer level firms aim to allocate resources to customers that result in the highest return. This has been the traditional focus of direct or database marketing. Using statistical models these customers are selected to receive direct communication (i.e., direct mailings) because they are most likely to respond (Bult & Wansbeek, 1995). Within Customer Relationship Management the focus has shifted toward long-term value outcomes, such as Customer Lifetime Value (CLV). Firms should invest in these customers with marketing communications and relationship or loyalty programs that optimize CLV. Data science models that focus on the individual purchase behavior of customers, such as lifetime duration, are then used. Using these models, the CLV is predicted for each customer. Typically, the customer base is segmented based on the expected CLV from high-value to low-value customers and larger budgets are typically allocated to high-value customers. Firms will also strive to keep the high-value customers and focus their retention efforts on these customers instead of low-value customers. By doing so, they immediately create value for the firm. In marketing literature, V. Kumar (a marketing professor) has notably published numerous studies on predicting CLV and the optimization of the profitability of customers using models (e.g., Kumar & Reinartz, 2016). We also refer to the extensive discussion of the CLV metric in Chapter 3 and case 3.1 in the same chapter on how an energy company calculated CLV.

A next development for improving marketing at the customer level is to focus on the individual customer and personalizing marketing and offers for individual customers. This brings us to the next level in how data science can create value for firms: the customer-firm interface level.

11.4 VALUE CREATION AT CUSTOMER-FIRM INTERFACE LEVEL

The arrival and growth of the Internet and digital devices with apps have moved targeting to higher levels. Further developments in artificial intelligence, such as robots and the Internet of Things, will only create more opportunities. Using real-time behavioral data, firms aim to provide personalized offers to customers visiting the website or logging on to an app. Given the relatively low costs of approaching customers online, personalization strategies have become economically more attractive (e.g., Zhang & Wedel, 2009). Successful personalization can be a way to gain a competitive advantage, as it could result in more satisfied customers (V2C) and more effective marketing (V2F). As a consequence, closed-loop marketing (CLM) has become popular. CLM consists of a cycle in which customer information is continuously collected and updated, and advanced analytics are used to forecast customer behavior and to redesign and personalize products, services, and marketing efforts in short cycles (Chung & Wedel, 2014), as shown in Figure 11.7. This CLM is frequently an integral part of the customer-firm interface and distinguishes the offerings from firms stilly relying on mass marketing approaches where every customer or customer segment is offered the same products or services.

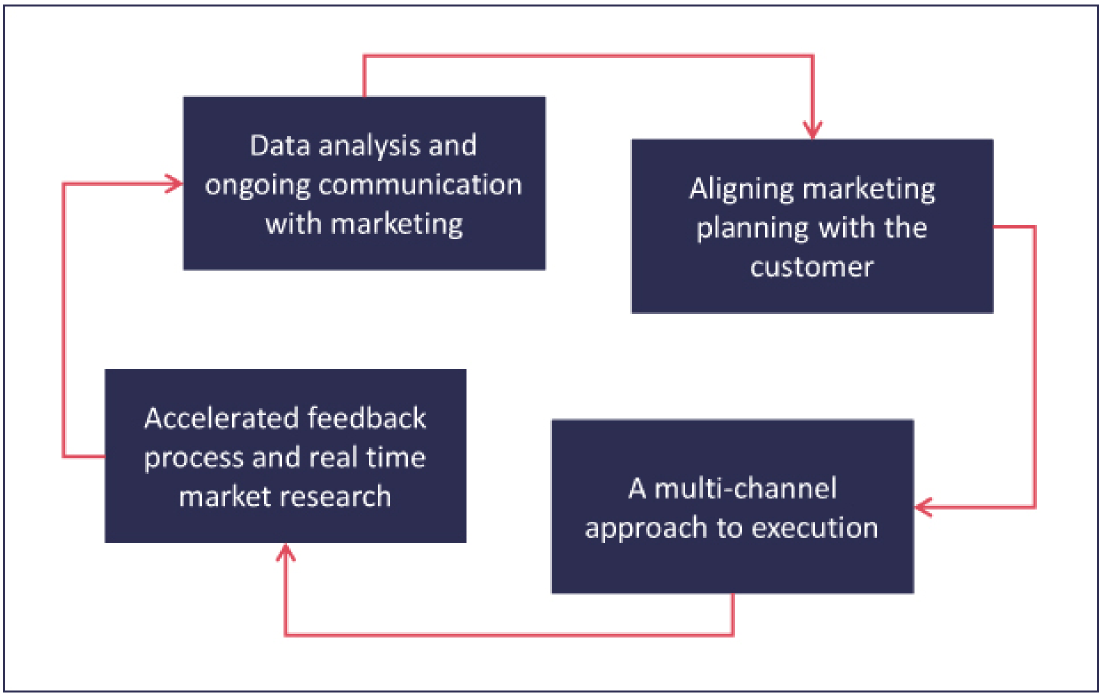

Data analysis and ongoing communication with marketing
Aligning marketing planning with the customer
A multi-channel approach to execution
Accelerated feedback process and real time market research

FIGURE 11.7 Closed-loop marketing process

Source: Adapted from Chung and Wedel (2014)

We distinguish between two major types of dynamic targeting approaches that differ with respect to the methods they use (Chung & Wedel, 2014):

- Recommendation systems
- Personalization systems.

11.4.1 Recommendation systems

Recommendation systems have been around since the start of Internet retailing and are used extensively by firms such as Amazon and Netflix. The key idea of recommendation systems is that, based on a customer's characteristics and characteristics of other customers, specific recommendations can be given to customers (e.g., "this book might be something for you"). Three types of recommendation systems can be distinguished (Chung & Wedel, 2014):

- Content filtering systems
- Collaborative filtering systems
- Hybrid forms of content and collaborative filtering systems.

Content filtering systems involve digital agents that produce recommendations based on the target customer's past preferences for products/services and the similarities between those products/services. Hence products or services are offered that are rather similar to the ones purchased before. For example, if one frequently buys fantasy books, the next recommendation will likely be a fantasy book as well. Collaborative filtering aims to make a recommendation using the preferences of other, similar customers. In practice, so-called memory-based systems are often used here. These systems use measures of similarity between customers' preferences or behaviors. These systems are simple to implement, easily scalable, and robust (Chung & Wedel, 2014). However, when billions of recommendations for millions of products are involved, more advanced technology, such as map reducing, could be needed. In case 11.3 we describe the case of an online retailer using this technology to achieve their business and marketing objectives with their recommendations. This case clearly shows the importance of having the right technology and that using data science models to personalize and customize creates value for the customer and the firm. Another technique often adopted is the nearest neighbor algorithm. Due to reported problems with the simpler memory-based systems, model-based systems have been developed in the marketing literature (e.g., Bodapati, 2008).

Ansari, Essegaier, and Kohli (2000) suggest that recommendation systems should not only be based on customers' revealed preferences and the revealed preferences of other customers but should also involve preferences for product attributes, expert judgments, and specific customer characteristics. Recommendation systems are therefore sometimes updated with preferences for attributes. For example, a combined algorithm using both conjoint utilities and behavioral patterns could be developed, thereby also taking into account the behavior of other, similar customers, for a website selling hotel breaks. In Figure 11.8 we provide an overview of this algorithm. Importantly, this algorithm also accounts for the reviews written about the hotel. This is an ongoing trend in recommendation agents. Many recommendation agents nowadays use reviews written by customers, as so many of them write such reviews. The popularity of social networks such as Facebook, where customers can *like* products and brands, stimulates the use of reviews in these agents as well.

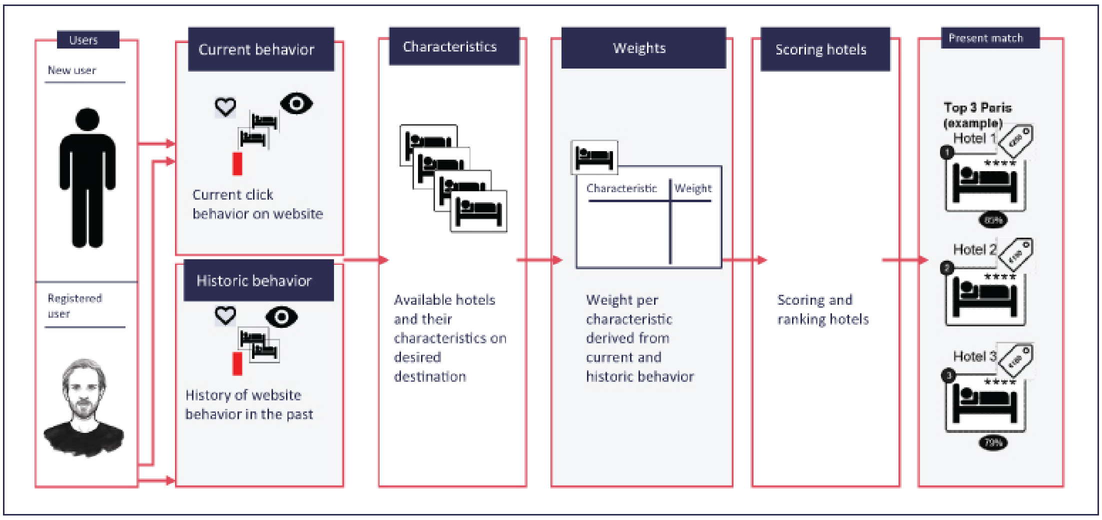

Users
New user
Registered user
Current behavior
Current click behavior on website
Historic behavior
History of website behavior in the past
Characteristics
Available hotels and their characteristics on desired destination
Weights
Characteristic
Weight
Weight per characteristic derived from current and historic behavior
Scoring hotels
Scoring and ranking hotels
Present match
Top 3 Paris (example)
Hotel 1
1
★★★★
85%
Hotel 2
2
★★★★
Hotel 3
3
★★★★
79%

FIGURE 11.8 Schematic overview of recommendation agent in hotel industry

CASE 11.3: IMPLEMENTATION OF BIG DATA ANALYTICS FOR RELEVANT PERSONALIZATION AT AN ONLINE RETAILER

Situation

The retailer is active in a growing market, with many strong competitors. To provide more value to their customers, they aim to inspire customers and provide more relevant recommendations to them when they visit the company's website. They strive to give customers fully automated suggestions of relevant product offers that may surprise them. They can already provide customized offers based on some product recommendation systems (see Chapter 4.2). However, they now aim to give more relevant personalized offers in different settings that make a difference and go beyond the "usual suspect" offers.

Complication

This retailer has millions of customers and offers an assortment of more than 8 million stock-keeping units (SKUs). Moreover, online customers search, look for, and purchase multiple products, either at the same time or sequentially. Overall, this leads to a very large number of customer/product interactions (2 billion per year) and even more product relations, in terms of searches and/or purchases in the same category or multiple categories, for the same brand, for different themes or occasions, etc. How to analyze these data is not obvious and requires lengthy computation times (i.e., 400 hours). Personalization based on all kinds of product relations can take so long to compute that it may not be effective. The challenge is to develop a scalable computation process, where instead of many weeks, a much shorter time period is required. This allows the retailer to come up with more real-time and relevant offers, which should lead to higher conversion rates.

Key message

The retailer implemented map-reducing technology through which, over two years, the computation time was reduced to one day and the click-through rate (CTR) was raised by 40%.

Approach

The retailer acquired limited hardware and started to use the open-source software Hadoop. They trained a team of data experts to use the software. They first did a pilot for a pre-sale set and used A/B testing (see Chapter 8) to test differences in CTRs between the old method and the new method. After the first positive results, they scaled up to other sets like sale and post-sale and sets around themes and product accessories.

Model used

To gain product recommendations and other added value propositions from customer interactions and behavior during a customer journey, an algorithm is necessary. A customer searches and/or purchases different products during one or more visits. These products, therefore, have a relationship (see Figure 11.9). For many customers and a product range of over 8 million products, this will result in billions and billions of product relations.

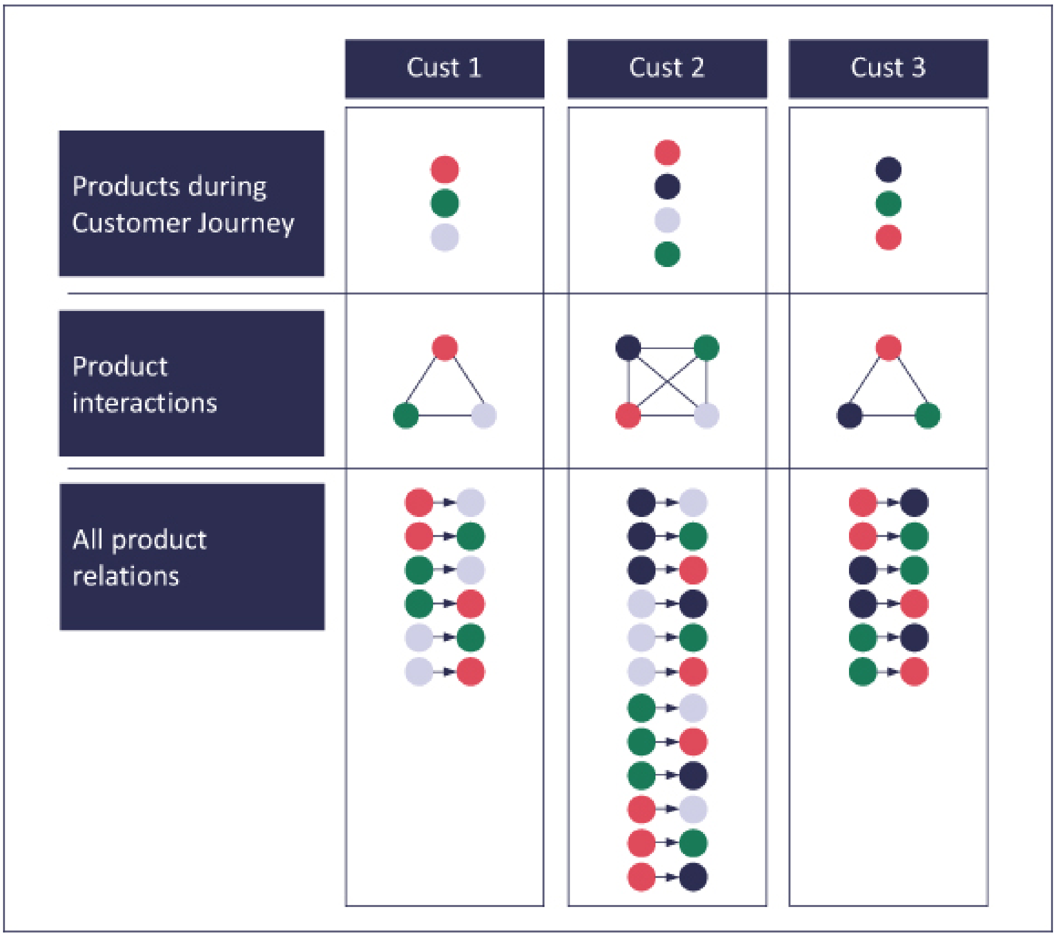


FIGURE 11.9 From search/purchase behavior to product combinations

To make sense of all these data, the retailer developed an algorithm by taking the following steps (see also Figure 11.10):

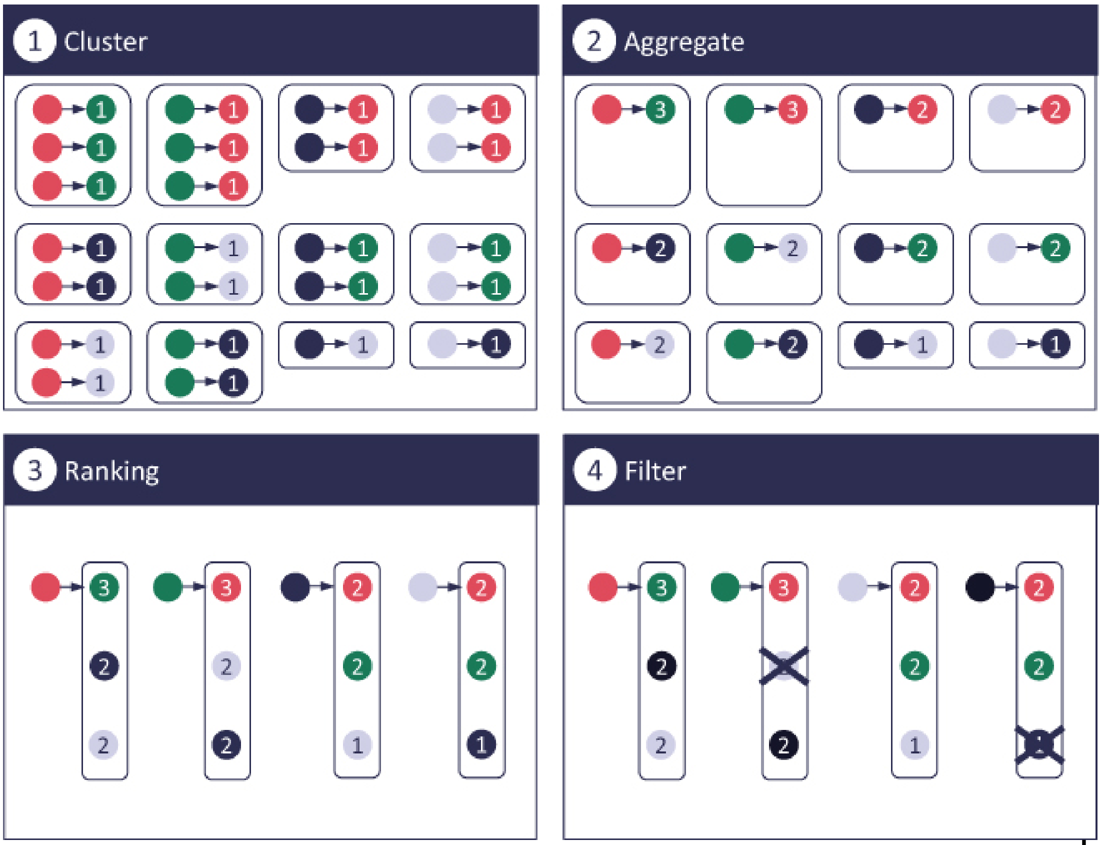


FIGURE 11.10 Algorithm for calculating product recommendations based on the product relation score

1. Cluster: the first step is to record all possible product relations and to cluster equal product relations.
2. Aggregate: by aggregating on unique product relations the number of times that each relation appears is calculated (the product relation score).
3. Rank: the next step is to rank each product by highest to lowest product relation scores.
4. Filter: the last step is to filter out undesirable relations. There are five sorts of filters:
    - Noise: very rare relations
    - Policy: unwanted relations such as medicines or eroticism
    - Practical: relations with products that are out of stock or out of range
    - Usual suspects: top five recommendations, which are already recommended in general
    - Derivatives: relations with product variants (for example, silver and gold iPhones).

Executing the algorithm over billions of product relations requires extensive computer time and working space. To manage this, the retailer used MapReduce programming technology (see Figure 11.11). The principles of

MapReduce are quite simple.[3] MapReduce can process large amounts of data in a short time because it splits a big task into subtasks. These subtasks are distributed across many computers, which can perform the subtasks simultaneously (distribution). This is done by using the features “map” and “reduce,” which are known from functional programming languages. Results are output files that are much smaller than the input files. After sending these back to the central server, the smaller output files are merged into an aggregated final file.

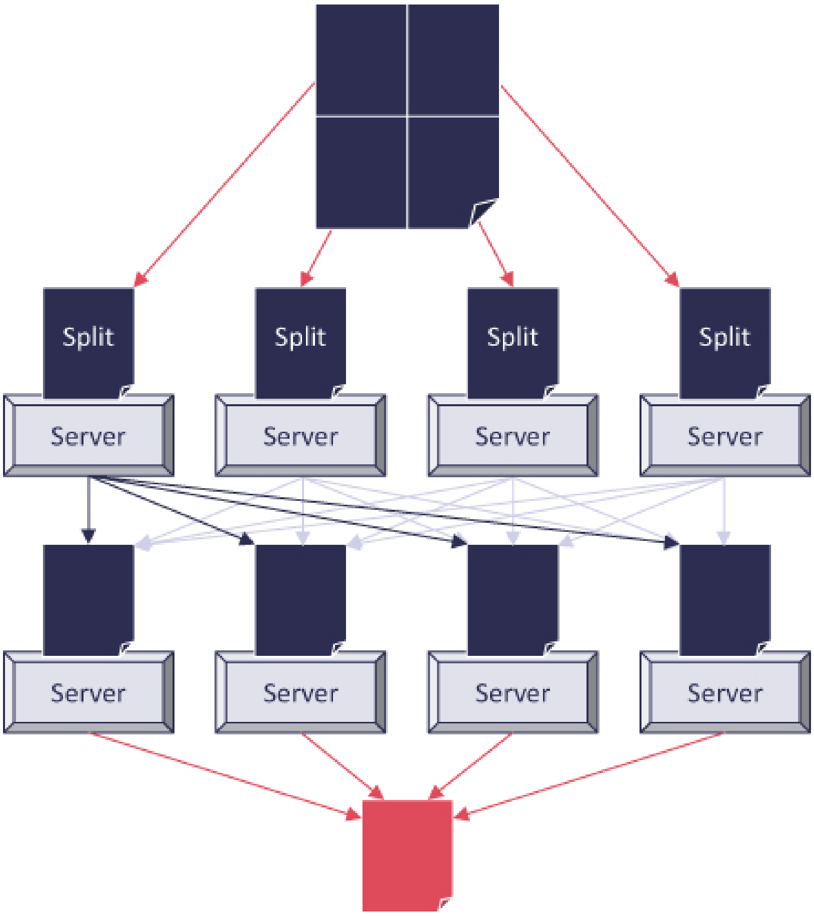


FIGURE 11.11 MapReduce programming model

Results

The CTR with the new method was significantly higher and almost doubled. This approach also led to cost reductions (reducing process time and having an in-house solution). Moreover, in general, the method created much more interesting offers for customers. An additional result was greater cooperation between marketing and IT, which may help develop new analytical projects in the future (see also Figure 11.12).

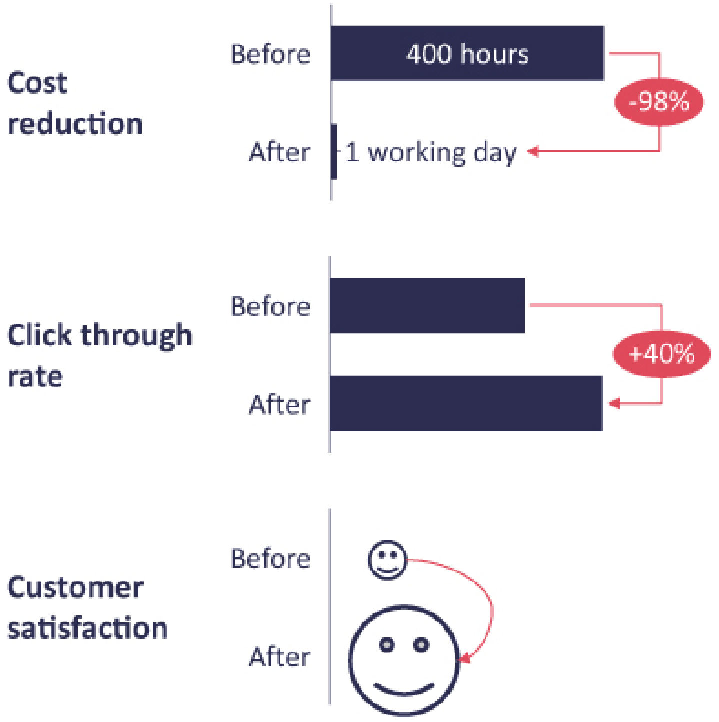


FIGURE 11.12 Results of personalization for V2C and V2F

Success factors

The keys to success for the approach can be summarized as follows:

- Marketing was responsible for the project. They did, however, work closely with IT and also adapted working methods as used by IT (e.g. agile working, scrum approaches). This stimulated a real cross-functional approach.
- By using a pilot, the analyst team within the firm could show that it worked and they could also learn from their mistakes. From the pilot, the project could be scaled up.
- The organization created a team dedicated to working on a project that was free to experiment with different solutions.
- The software used was free, while the hardware was standard and not complicated. This substantially reduced the costs, as neither advanced hardware nor software had to be purchased. In fact, the hardware consisted of four standard desktops, coupled to create a more powerful engine.

In sum, the business case was rather simple. Revenues were substantially improved by the higher CTRs and conversion rates, and there was a reduction of operational costs, all accomplished with a minimum of investment.

11.4.2 Personalization

The key difference between recommendation systems and personalization systems is that recommendation systems recommend existing products or services, whereas personalization systems adapt the offering to customers' needs. Note that this not only involves product or service offerings—how the website is displayed (looks and feels) to customers can also be personalized. In that sense, personalization systems are more customer-driven than recommendation systems. With personalization, the firm tailors the marketing mix to the customer, based on available customer information (Arora *et al.*, 2008). Examples of personalization can be found in many industries and for some—especially for information- based services—personalization has become key. For example, Netflix advises customers about series and films based on prior viewing behavior. Pandora suggests songs on such prior selections as well as on similarities between song attributes (Chung & Wedel, 2014).

For personalization systems, it is very important to learn consumer preferences. This can be done using a short survey on consumer preferences. Based on the results of this exercise, a personalized offer can be provided, and based on the outcome of this offer (i.e., accept or reject offer) a consumer's preferences can be updated (see CLM approach). For more mobile environments in particular, adaptive personalization systems (APS) have been developed. APS takes full advantage of unobtrusively obtained customer information to provide personalized services in real-time (Chung, Rust, & Wedel, 2009). These systems require very limited or no input from the customer and rely heavily on purchase data. Importantly, APS can learn from small pieces of information over time and pick-up changes in consumer preferences.

In the marketing literature, there are many applications of these APS, which vary in their use of statistical and econometric methods. One important issue in personalization is that one aims to account for customer heterogeneity in the used model. For other purposes, specific algorithms are used that frequently have their background in artificial intelligence. For example, Chung, Rust, and Wedel (2009) have developed an "adaptive personalization system" and illustrate its implementation for digital audio players. The proposed system automatically downloads personalized playlists of MP3 songs into a consumer's mobile digital audio device and requires little proactive user effort (i.e., no explicit indication of preferences or ratings for songs). They show that their system works in real-time and is scalable to the massive data typically encountered in personalization applications. Specifically, they develop a system consisting of (a) a personalized agenda, (b) an adaptive learning algorithm, (c) a collaborative customization model for all customers, and (d) a dynamic customization model for individual customers. For illustrative purposes, we provide their model in Figure 11.13. For more details we refer to their paper published in Marketing Science (Chung, Rust, & Wedel, 2009).

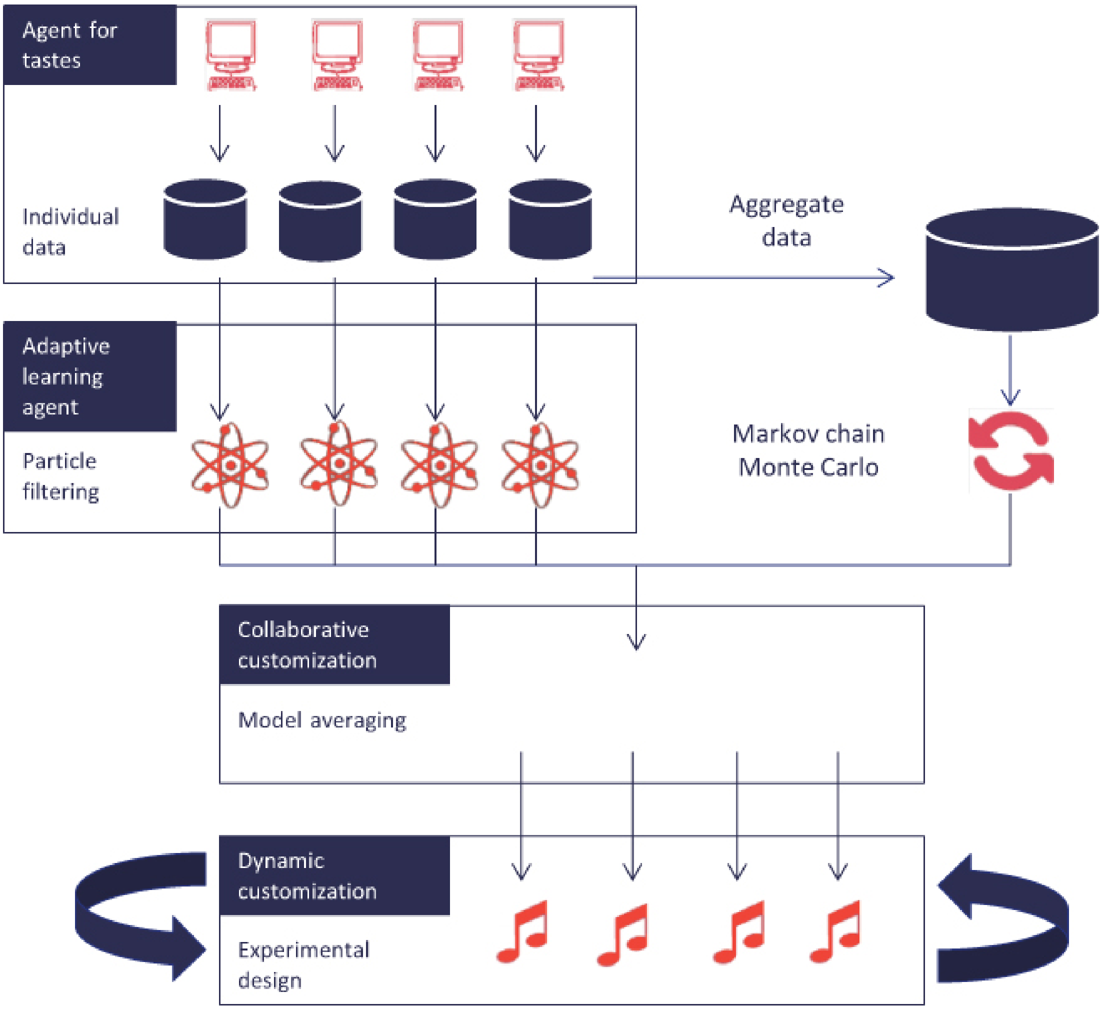


FIGURE 11.13 Flow diagram of the adaptive personalization system developed by Chung, Rust, and Wedel (2009)

Source: Adapted from Chung, Rust, and Wedel (2009)

11.4.3 Dark sides of personalization

In order for personalization to be effective, it should be executed well. This implies that the correct offers should be given to the correct customers. If offers are provided that are not in line with customer preferences customers can become dissatisfied. There is yet another danger from a customer viewpoint. If personalization becomes too close to customers, as models are able to predict so well, customers may feel that they are being monitored constantly. This can create feelings of intrusiveness and reactance leading to dissatisfaction (van Doorn & Hoekstra, 2013) (see also Chapter 6 on privacy). This suggests that models developed for personalization should find an optimum prediction accuracy.

There is also another risk of personalization from both a customer viewpoint and a merely societal viewpoint. Constantly personalizing implies that customers are confronted with similar offerings (i.e., similar series or music on

Netflix and Spotify, respectively), which creates a lower variety in choices and less exploratory behavior. This may narrow the choice set and customers who value variety (also called variety seekers) may become bored and dissatisfied.

From a societal point of view, the algorithms can lead to a bubble in which the personalized content constantly reconfirms preferences and opinions. In particular, social media like Twitter, Facebook, and Instagram are being accused of this. This bubble is referred to as the social media "filter bubble," and is defined by Eli Pariser, an internet activist, as a

> personal, unique universe of information that you live in online. And what's in your filter bubble depends on who you are, and it depends on what you do. But the thing is that you don't decide what gets in. And more importantly, you don't actually see what gets edited out (Gould, 2019).

This bubble could mean that customers only read content in line with their beliefs. This results in a stronger conviction of one's own opinions, less openness to other opinions and less empathy for the people that hold them. This can create strong tensions between groups within societies.[4]

11.5 DATA SCIENCE AS BUSINESS MODEL

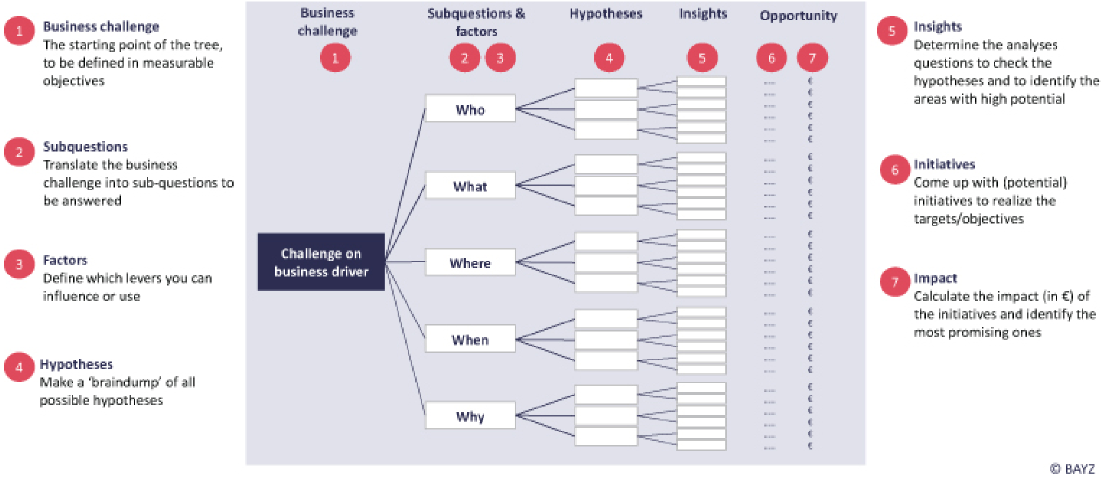
In this book, we focus on the value creation of data science for marketing. However, firms may also use data science in their total business model. Data science is then a central element of value creation and value appropriation for firms. This is mainly true for digital firms, such as Alphabet or Facebook. Nowadays many firms are implementing digital principles in their business models. This is referred to as digital transformation, which is defined as a change in how a firm employs digital technologies to develop a new digital business model that helps to create and appropriate more value for the firm (Verhoef *et al.*, 2021). In digital transformation, the capability to acquire and analyze big data for decision-making is crucial given that the functionality of digital technologies all relies on digital data. In comparison with the other levels of data science value creation, data science is affecting all business functions (not only marketing) and also goes beyond the customer-firm interface. For example, pure digital firms like Amazon and Booking.com constantly use analytics to tailor new offerings to customers as well as to optimize revenues with dynamic pricing and revenue management. Data science then becomes an integral part of business affecting, marketing, supply chain management, logistics, human resource management, and accounting and finance.[5] How this is realized is beyond the scope of this book and we refer to the evolving literature on digital transformation for more elaborate discussions on this topic.

11.6 OPPORTUNITY FINDING AS A METHODOLOGY TO CREATE MORE VALUE

The above discussion on value creation focusses on the strategic impact of data science within organizations. However, to have a value impact, data scientists should be aware of what to focus on. They could immediately start analyzing the data or set up all kinds of data science projects. More guidance is, however, required. They should focus on the opportunities that create value and have an impact on the organization. For this reason, we also discuss a specific methodology that can be used to detect opportunities: opportunity finding. This methodology has also been applied in some of the cases described in this chapter and is one of the success factors in the data science projects described.

Opportunity finding is a methodology for implementing two of the four analytical strategies, namely problem solving and data exploitation, as discussed in Chapter 7. It is very powerful since it helps in giving a fact based, data driven approach to business challenges. Opportunity finding is a structured way to identify solutions, following a top-down approach and resulting in an exhaustive breakdown of possible initiatives.

Opportunity finding is however not a goal in itself, but a way of working to identify new opportunities through a structured approach. It helps put a focus on the most valuable initiatives, combines business and intelligence expertise, and facilitates monitoring progress on objectives and targets for the defined business challenge. Opportunity finding keeps us from reinventing the wheel and at the same time leaves enough room for creativity. Typically, we identify seven steps in the process of opportunity finding (see also Figure 11.14)



1 **Business challenge**
The starting point of the tree, to be defined in measurable objectives

2 **Subquestions**
Translate the business challenge into sub-questions to be answered

3 **Factors**
Define which levers you can influence or use

4 **Hypotheses**
Make a 'braindump' of all possible hypotheses

5 **Insights**
Determine the analyses questions to check the hypotheses and to identify the areas with high potential

6 **Initiatives**
Come up with (potential) initiatives to realize the targets/objectives

7 **Impact**
Calculate the impact (in €) of the initiatives and identify the most promising ones

© BAYZ

FIGURE 11.14 The seven steps of opportunity finding

11.6.1 Step 1: The business challenge

Business challenges are often linked to the different phases of the customer lifecycle (like customer acquisition) or the phases of the customer journey and are quantified in terms of the necessary potential (the delta from current to desired status) to be realized, ideally expressed in a monetary value and at least measurable by one or more defined KPIs. A business challenge could be “We aim to reduce the customer churn from 8.5% to 7%, to achieve an extra EBIT of 2,5 mio. Euro’s.”

11.6.2 Step 2: The sub questions

The next step is to break down the business challenge with sub questions to be answered. For this we use the earlier discussed five W-questions (Who, What, Where, When and Why) that are typical for a customer centric organization (see also Chapter 4, Figure 4.10). Specifying the business challenge for every “W” helps provide a good diagnosis of the business challenge. In this example of churn, we could phrase the “Who” question as: “Who are the customer that show an above or below average churn percentage” and for the “What:” “What products have a higher or lower product churn, compared to others.” Although it is good to check all the five questions, not all five might be relevant for the business challenge. Typically, two or three W-questions jump out. These selected W-questions should be given highest priority and most attention moving forward.

11.6.3 Step 3: The factors

The factors behind the sub questions define the levers that can actually be used to start defining initiatives for the business challenge. The factors are in fact the variables or groups that you can identify in the data for a sub question. In our example of the “Who” for customer churn, we could define “age” as a possible factor that might be relevant in finding groups with a higher or lower churn percentage. Another angle could be to break down the customer base by value, creating value segments to explore whether this results in a difference in the churn percentage.

11.6.4 Step 4: Hypotheses

Since the process of defining all factors for all selected W-questions and then next analyzing all these factors is very time consuming, the step of defining hypotheses is very helpful and necessary in guiding the analysis process. This is a very important step since it is very helpful to make a “brain dump” of all hypotheses or ideas that might already exist within the business. Doing this is also very helpful in bringing the two worlds (business and data science) together. In our example a possible hypothesis could be “High value, young customers, show a high churn rate, due to cheaper alternatives and low switching barriers.” You might notice that in this hypothesis not only are two different factors within the “Who” combined (age and value), but also the

"Why" is present, namely the possible reason why customers are churning. This suggests that, when this hypothesis is accepted at a later stage, the final opportunity tree should be a combination of W-questions.

11.6.5 Step 5: Insights

In this step we determine the analyses questions to check the hypotheses and to identify areas with high potential. Potential is identified in the differences in performance we can uncover along the primary KPI of our business challenge, in this case the churn percentage. After this step we will have a list of validated or rejected hypotheses as well as a lot of valuable insights to guide the next step.

11.6.6 Step 6: Initiatives

Having broken down the initial business challenge and having the output of the hypotheses and analyses, it is noticeable that the boundaries for designing initiatives have become much sharper. Redrawing the opportunity tree, based on the relevant combinations of the W-questions into excluding branches, specified by the discriminating factors, will make it feasible to come up with a list of fact-based initiatives. Initiatives are then specified as ideas to realize (part of) the business challenge by addressing a target group via one or more channels at the right moment with a well targeted proposition. Mobilizing creativity within the organization to brainstorm within the defined boundaries should result in a list of initiatives that is long enough to realize the potential of the business challenge.

11.6.7 Step 7: Impact

In the last step the impact (in dollars or euros) of the initiatives is calculated to identify the most promising ones, but also to make sure that the initial potential of the business challenge is feasible. For this we sum up all the initiatives and compare the total against the business challenge. If there is a gap, the potential has been overestimated, or more effort should be made to come up with a longer list of initiatives. When implementing the initiatives, it will probably be necessary to prioritize them due to limited available resources. To do that, we suggest plotting the initiatives in a matrix with two dimensions: the necessary effort to realize an initiative and the potential value of every initiative. Then we can start with the low effort, high value initiatives.

11.7 CONCLUSION

In this chapter, we focused on the value creation opportunities of data science. We discussed three levels within a firm where data science can have a profound value creation: marketing function, customer-interface, and the business model. In our discussion, we focused on the first two levels, given the scope of this book. In this chapter, you also found multiple cases. These cases clearly illustrate the impact data science can have on value creation within firms. These cases are just examples and in business practice many other applications can be found. To help focus on the right data science project or initiatives, we also discussed opportunity finding methodology. In our experience, this methodology improves value impact and impact within the organization. The key message is that data science should not be executed for the sake of nice modeling or having extensive data available but should lead to value creation for customers and firms.

ASSIGNMENT 11.1: OPPORTUNITY FINDING FOR SURE.COM

Sure.com is a top 5 insurance company in the Netherlands, offering travel, car and legal insurances.

For the next year they have an ambition to grow in both premium value and number of policies.

Sure.com is looking for growth opportunities to achieve their targets.

Their growth ambition for the next year is:

+ 1% number insurance policies

+ 6% premium value.

You are asked by Sure.com to help them to define and to structure the business challenge.

Questions**:**

1. Have a look at Figures 11.15 to 11.18 below to get a better understanding of what has happened at Sure.com last year. What are your first impressions of what they should focus on?
2. Define the business challenge for Sure.com. How would you quantify the challenge?
3. Which KPIs do you think are relevant to monitor progress during the year?
4. Define the possible (relevant) sub questions, using the 5 Ws
5. Define per sub question at least two possible factors
6. Define per factor at least two hypotheses
7. Define per hypothesis at least one analysis question
8. Select the two most relevant W-questions, combine the defined factors and draw an opportunity tree with excluding branches
9. Define per branch at least two initiatives and make a rough estimation of the contribution (in value) per Initiative.

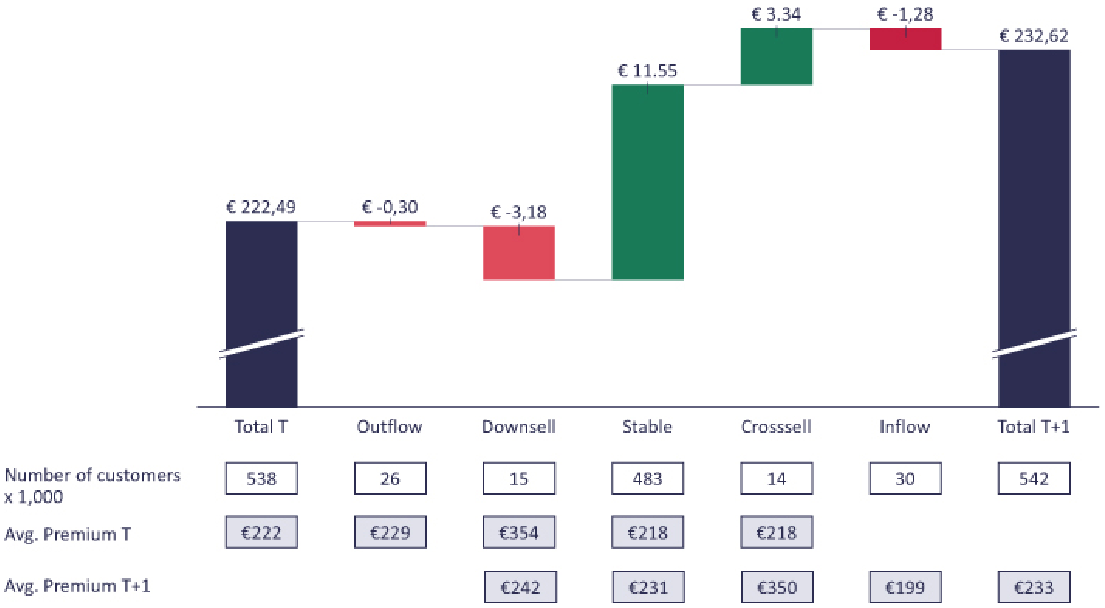
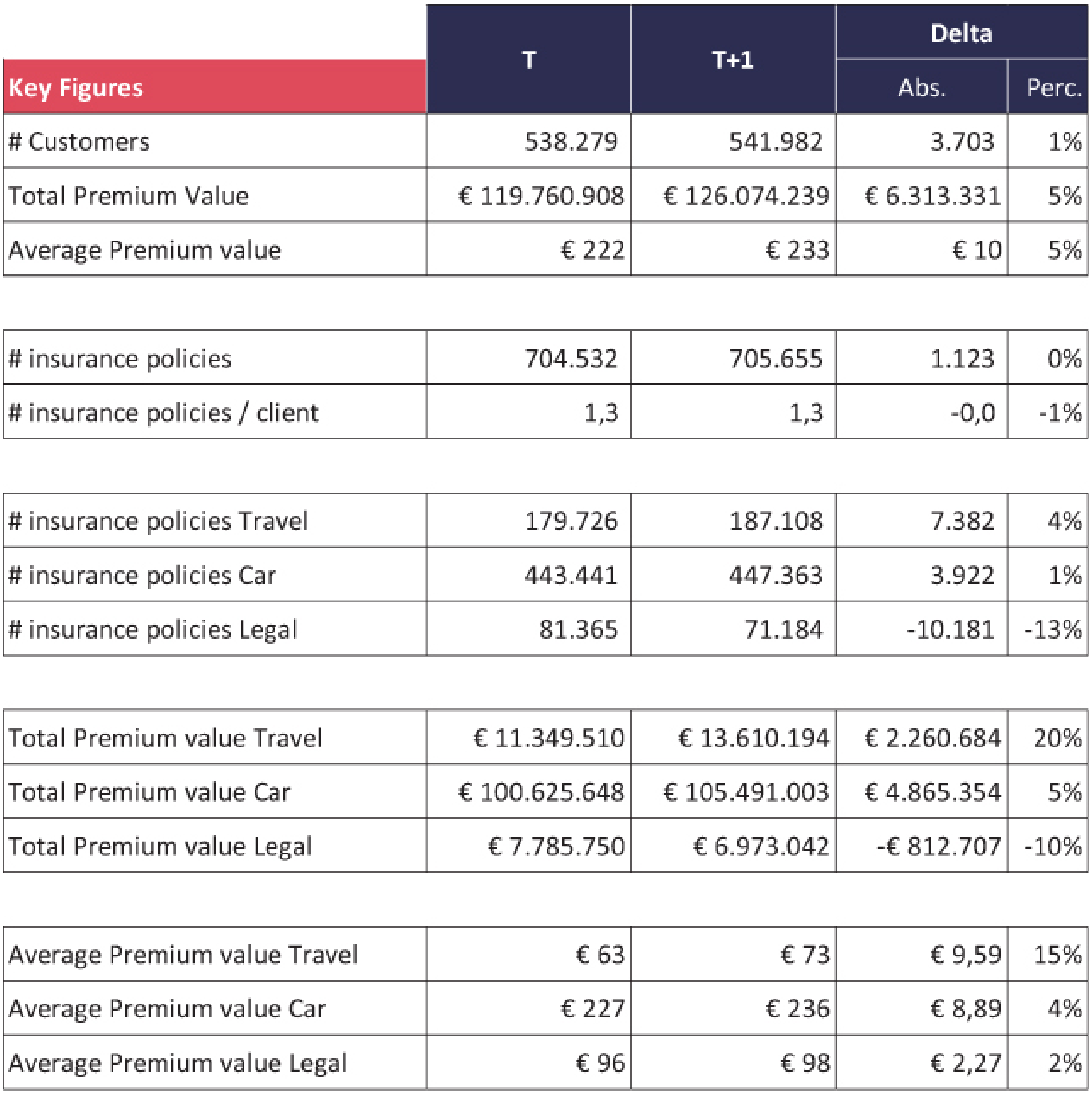
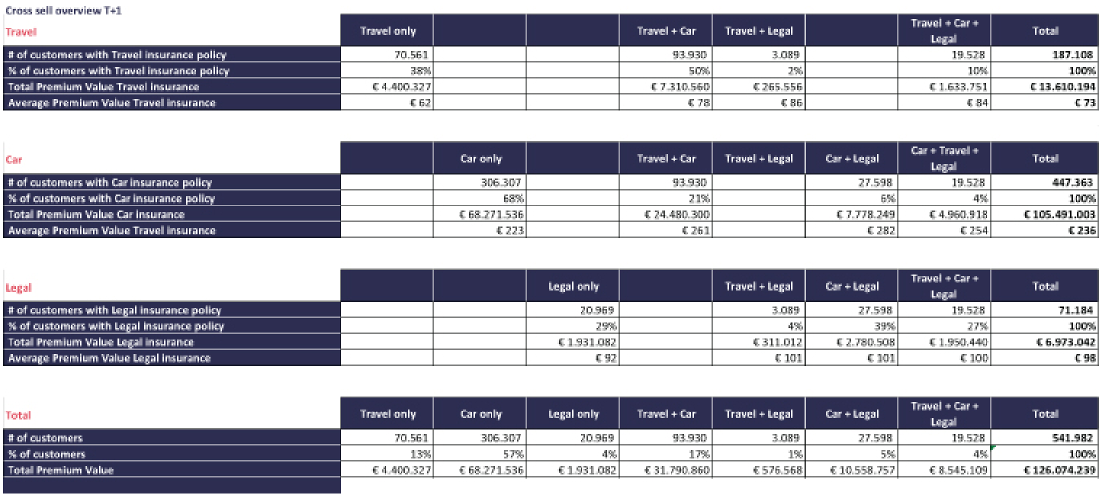


FIGURE 11.15 "Like-4-like" view on weighted average premium per customer



	T	T+1	Delta	
Key Figures			Abs.	Perc.
# Customers	538.279	541.982	3.703	1%
Total Premium Value	€ 119.760.908	€ 126.074.239	€ 6.313.331	5%
Average Premium value	€ 222	€ 233	€ 10	5%
# insurance policies	704.532	705.655	1.123	0%
# insurance policies / client	1,3	1,3	-0,0	-1%
# insurance policies Travel	179.726	187.108	7.382	4%
# insurance policies Car	443.441	447.363	3.922	1%
# insurance policies Legal	81.365	71.184	-10.181	-13%
Total Premium value Travel	€ 11.349.510	€ 13.610.194	€ 2.260.684	20%
Total Premium value Car	€ 100.625.648	€ 105.491.003	€ 4.865.354	5%
Total Premium value Legal	€ 7.785.750	€ 6.973.042	-€ 812.707	-10%
Average Premium value Travel	€ 63	€ 73	€ 9,59	15%
Average Premium value Car	€ 227	€ 236	€ 8,89	4%
Average Premium value Legal	€ 96	€ 98	€ 2,27	2%

FIGURE 11.16 Key figures



Cross sell overview T+1

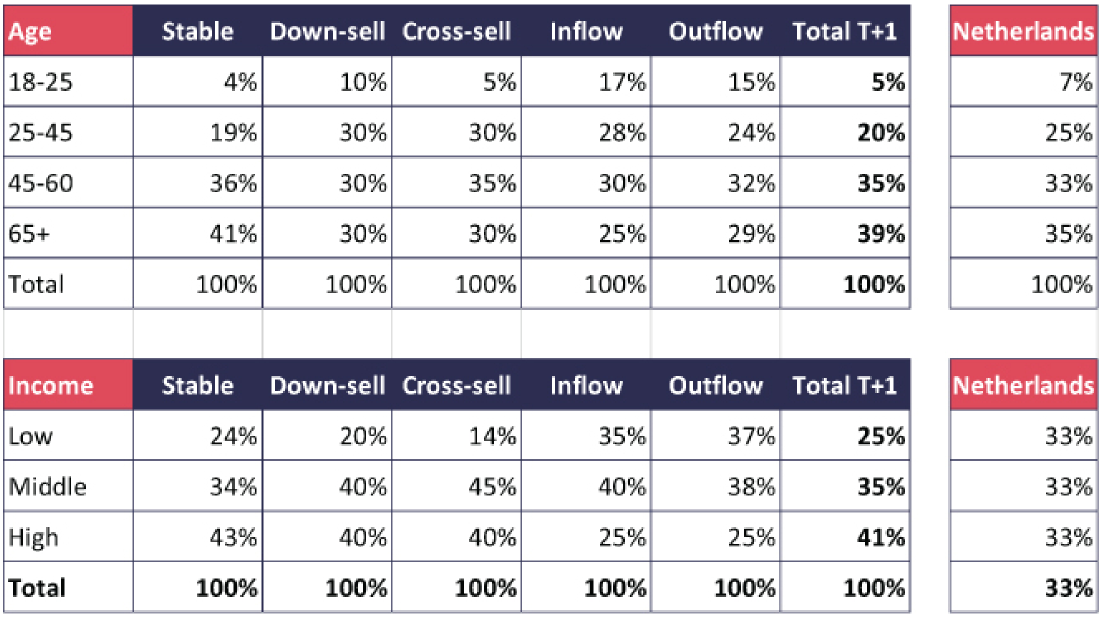
Travel	Travel only			Travel + Car	Travel + Legal		Travel + Car + Legal	Total
# of customers with Travel insurance policy	70.561			93.930	3.089		19.528	**187.108**
% of customers with Travel insurance policy	38%			50%	2%		10%	**100%**
Total Premium Value Travel insurance	€ 4.400.327			€ 7.310.560	€ 265.556		€ 1.633.751	**€ 13.610.194**
Average Premium Value Travel insurance	€ 62			€ 78	€ 86		€ 84	**€ 73**

Car		Car only		Travel + Car	Travel + Legal	Car + Legal	Car + Travel + Legal	Total
# of customers with Car insurance policy		306.307		93.930		27.598	19.528	**447.363**
% of customers with Car insurance policy		68%		21%		6%	4%	**100%**
Total Premium Value Car insurance		€ 68.271.536		€ 24.480.300		€ 7.778.249	€ 4.960.918	**€ 105.491.003**
Average Premium Value Travel insurance		€ 223		€ 261		€ 282	€ 254	**€ 236**

Legal			Legal only		Travel + Legal	Car + Legal	Travel + Car + Legal	Total
# of customers with Legal insurance policy			20.969		3.089	27.598	19.528	**71.184**
% of customers with Legal insurance policy			29%		4%	39%	27%	**100%**
Total Premium Value Legal insurance			€ 1.931.082		€ 311.012	€ 2.780.508	€ 1.950.440	**€ 6.973.042**
Average Premium Value Legal insurance			€ 92		€ 101	€ 101	€ 100	**€ 98**

Total	Travel only	Car only	Legal only	Travel + Car	Travel + Legal	Car + Legal	Travel + Car + Legal	Total
# of customers	70.561	306.307	20.969	93.930	3.089	27.598	19.528	**541.982**
% of customers	13%	57%	4%	17%	1%	5%	4%	**100%**
Total Premium Value	€ 4.400.327	€ 68.271.536	€ 1.931.082	€ 31.790.860	€ 576.568	€ 10.558.757	€ 8.545.109	**€ 126.074.239**

FIGURE 11.17 Cross-sell figures



Age	Stable	Down-sell	Cross-sell	Inflow	Outflow	Total T+1	Netherlands
18-25	4%	10%	5%	17%	15%	**5%**	7%
25-45	19%	30%	30%	28%	24%	**20%**	25%
45-60	36%	30%	35%	30%	32%	**35%**	33%
65+	41%	30%	30%	25%	29%	**39%**	35%
Total	100%	100%	100%	100%	100%	**100%**	100%

Income	Stable	Down-sell	Cross-sell	Inflow	Outflow	Total T+1	Netherlands
Low	24%	20%	14%	35%	37%	**25%**	33%
Middle	34%	40%	45%	40%	38%	**35%**	33%
High	43%	40%	40%	25%	25%	**41%**	33%
Total	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**33%**

FIGURE 11.18 Basic profile data

NOTES

1. This case was jointly devised with Evert de Haan and Thorsten Wiesel. We kindly thank them for allowing this case to be used in this book.
2. This is something we have investigated further in De Haan *et al*. (2018).
3. The name MapReduce originally referred to a proprietary Google technology but this has since been genericized. A popular open-source implementation based on this technology is Apache Hadoop.
4. Based on: https://www.nbcnews.com/better/lifestyle/problem-social-media-reinforcement-bubbles-what-you-can-do-about-ncna1063896

5. See for example: https://towardsdatascience.com/you-need-a-data-strategy-to-steer-your-digital-transformation-journey-7e87484a106b

REFERENCES

Ansari, A., Essegaier, S., & Kohli, R. (2000). Internet recommendation systems. *Journal of Marketing Research*, *37*(3), 363–375.

Arora, N., Dreze, X., Ghose, A., Hess, J. D., Iyengar, R., Jing, B., Joshi, Y., Kumar, V., Lurie, N., Neslin, S., Sajeesh, S., Su, M., Syam, N., Thomas, J., & Zhang, Z. J. (2008). Putting one-to-one marketing to work: Personalization, customization, and choice. *Marketing Letters*, *19*(3–4), 305–321.

Bodapati, A. V. (2008). Recommendation systems with purchase data. *Journal of Marketing Research*, *45*(1), 77–93.

Bult, J. R. & Wansbeek, T. (1995). Optimal selection for direct mail. *Marketing Science*, *14*(4), 378–394.

Chung, T. S., Rust, R. T., & Wedel, M. (2009). My mobile music: An adaptive personalization system for digital audio players. *Marketing Science*, *28*(1), 52–68.

Chung, T. S. & Wedel, M. (2014). Adaptive personalization of mobile information services. In: R. T. Rust, & M. H. Huang, (eds) *Handbook of Service Marketing Research* (pp. 395–412). Cheltenham: Edward Elgar.

de Haan, E., Kannan, P. K., Verhoef, P. C., & Wiesel, T. (2018). Device switching in online purchasing: Examining the strategic contingencies. *Journal of Marketing*, *82*(5), 1–19.

Gould, W.R. (2019), Are you in a social media bubble? Here is how to tell. NBC News, Oct. 19 2019 (https://www.nbcnews.com/better/lifestyle/problem-social-media-reinforcement-bubbles-what-you-can-do-about-ncna1063896).

Konuş, U., Neslin, S. A., & Verhoef, P. C. (2014). The effect of search channel elimination on purchase incidence, order size and channel choice. *International Journal of Research in Marketing*, *3*(1), 49–64.

Kumar, V. & Reinartz, W. (2016). Creating enduring customer value. *Journal of Marketing*, *80*(6), 36–68.

Leeflang, P., Bijmolt, T., Pauwels, K., & Wieringa, J. (2015). *Modeling Markets: Analyzing Marketing Phenomena and Improving Marketing Decision Making*. (International Series in Quantitative Marketing; Vol. 1). Berlin: Springer.

Lesscher, L., Lobschat, L., & Verhoef, P. (2021). Do offline and online go hand in hand? Cross-channel and synergy effects of direct mailing and display advertising. *International Journal of Research in Marketing*, forthcoming.

Pauwels, K., Ambler, T., Clark, B. H., LaPointe, P., Reibstein, D., Skiera, B., Wierenga, B., & Wiesel, T. (2009). Dashboards as a service why, what, how, and what research is needed? *Journal of Service Research*, *12*(2), 175–189.

Reibstein, D. J., Norton, D., Joshi, Y., & Farris, P. (2005). Marketing dashboard: A decision support system assessing marketing productivity. Marketing Science Conference. Atlanta.

Rust, R. T., Lemon, K. N., & Zeithaml, V. A. (2004). Return on marketing: Using customer equity to focus marketing strategy. *Journal of Marketing*, *68*(1), 109–127.

van Doorn, J. & Hoekstra, J. C. (2013). Customization of online advertising: The role of intrusiveness. *Marketing Letters*, *24*(4), 339–351.

Verhoef, P. C., Broekhuizen, T., Bart, Y., Bhattacharya, A., Dong, J. Q., Fabian, N., & Haenlein, M. (2021). Digital transformation: A multidisciplinary reflection and research agenda. *Journal of Business Research*, *122*, 889–890.

Verhoef, P. C., Heijnsbroek, M., & Bosma, J. (2017). Developing a service improvement system for the National Dutch Railways. *Interfaces*, *47*(6), 489–504.

Verhoef, P. C. & Leeflang, P. S. H. (2009). Understanding marketing department's influence within the firm. *Journal of Marketing*, *73*(2), 14–37.

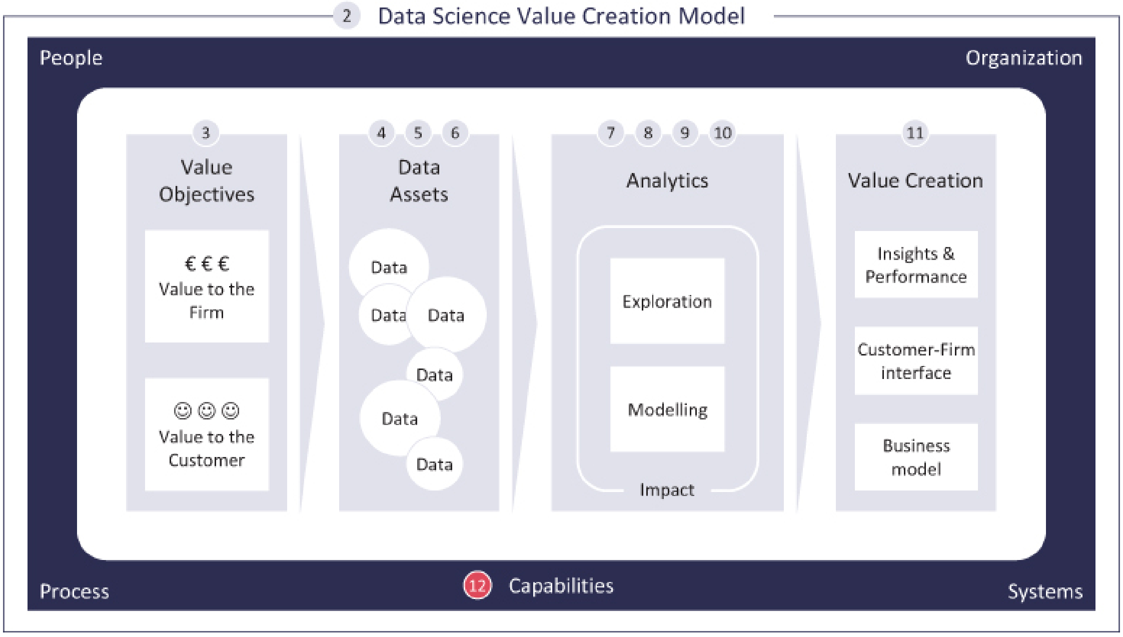
Wiesel, T., Pauwels, K., & Arts, J. (2011). Marketing's profit impact: Quantifying online and off-line funnel progression. *Marketing Science*, *30*(4), 604–611.

Zhang, J. & Wedel, M. (2009). The effectiveness of customized promotions in online and offline stores. *Journal of Marketing Research*, *46*(2), 190–206.

CHAPTER 12

Building successful data analytics capabilities

DOI: 10.4324/9781003011163-12



12.1 INTRODUCTION

Having (big) data is not sufficient of itself to develop a successful value-creating data strategy. Firms need to invest in capabilities that transform the organization into a fact-based and data-driven enterprise. At first blush, this seems easy to achieve. Firms should just buy some software, hire a few data science experts, and the big data and AI initiatives can take-off. Unsurprisingly, 25% of companies expect short-term spending on data management and analytics solutions to increase and, in the long run, 87% of CXOs say that becoming a more intelligent enterprise is their priority for the next five years (IDC,[1]2020). Current figures[2]show that the global big data and business analytics market was valued at 168.8 billion U.S. dollars in 2018 and is forecast to grow to 274.3 billion US dollars by 2022, with a five-year compound annual growth rate (CAGR) of 13.2%.

However, many of these investments will lead to serious disappointments. Only 32% of business executives say that they're able to create measurable value from data and only 27% of business executives say that their data and analytics projects produce actionable insights.[3] Another study[4] shows that only 24% of business decision-makers surveyed are fully confident in their ability to read, work with, analyze and argue with data. These numbers illustrate that many organizations struggle with the lack of capabilities and resources to use data and analytics to improve business performance. However, firms should invest in people, systems, processes, and the organization. Our experiences have shown that firms may face several hurdles to doing so. Firms may, for example, be confronted with old systems and databases that are difficult to replace. Importantly, there is also a shortage of analytical talent. SearchBusinessAnalytics[5] published that in August 2018, LinkedIn reported that there's a shortage of around 151,000 people with data science skills in the US, based on data from its platform. Combining this with a 15% discrepancy between job postings and job searches on Indeed, it's evident that demand for data scientists outstrips supply. Figure 12.1 illustrates that the gap between demand and supply of analytical talent continues to grow. An additional hurdle is changing the culture to one in which analytical solutions are considered as valuable input in marketing decision-making, instead of focusing solely on intuition.

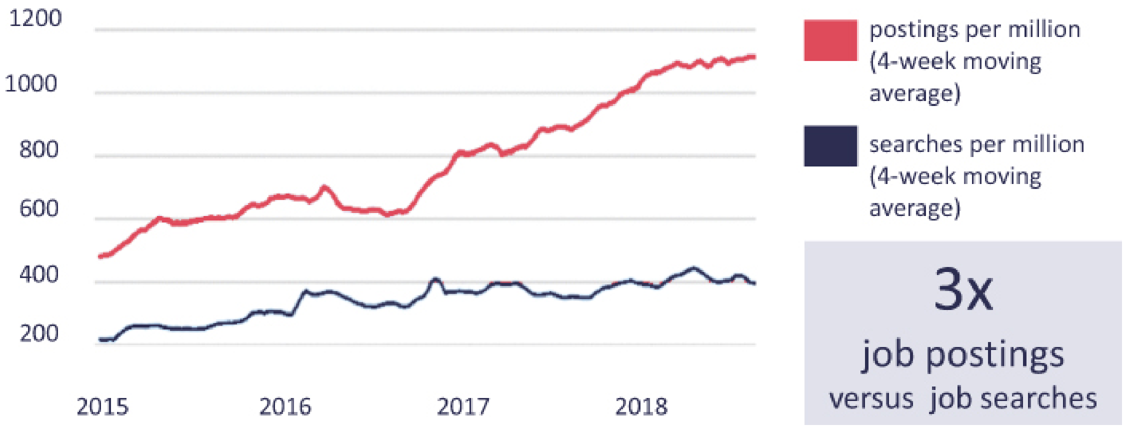


FIGURE 12.1 Growing gap between demand and supply of data scientists

In this chapter, we will discuss the main ingredients from which firms can successfully build an analytical competence, one that fully allows them to benefit from data opportunities. Based on our Data Science Value Creation Model, we will structure our discussion mainly around four building blocks of a successful analytical competence:

1. Process: creating a common fact driven way of working
2. People: recruiting, developing, and maintaining analytical talent
3. Systems: the platform and tools for an integrated data-ecosystem
4. Organization: taking the right role and place in the organization and establishing a data-driven culture for the highest impact.

First, however, we will start this chapter with our vision of what the transformation of an organization entails if it wishes to create strong analytical competences.

12.2 TRANSFORMATION TO CREATE SUCCESSFUL ANALYTICAL COMPETENCE

Over the past few decades, analytical departments have become more important within many firms. We observe this especially in retail, financial services, and service industries such as telecom. Numerous examples have been discussed in literature and can be seen in practice, in which organizations such as Tesco, Capital One, O'Hara's, and Vodafone have developed analytical functions to help them compete effectively (e.g., Davenport & Harris, 2007; Humby, Hunt, & Phillips, 2008; Verhoef & Lemon, 2013). As already discussed, there is also ample scientific empirical evidence that firms investing in these analytical competences can actually outperform their competition. Analytical functions, such as may be found in Marketing Intelligence (MI) or Customer Intelligence (CI) departments, perform an important role in creating these competences. In an era of big data and AI they will become even more important.

However, we also observe that analytical competences are not fully developed in many companies. In fact, you could argue that many companies are not well prepared internally for the supposed big data and AI revolution. In many companies, analytics departments are sticking to their traditional role, where they often only provide simple reports and customer selections on demand. They do not actively participate in value creation discussions due to a lack of analytical capabilities to provide strong market and customer insights. That type of analytical department mainly has a reactive supplier role.

12.2.1 Changing roles

In general, we have observed changing roles in how analytical departments operate in firms. These changing roles are displayed in Figure 12.2. CI departments frequently originate in a supplier role in which the credo is: "We deliver what is being asked for." In the next phase they can become a challenger, as they are also asked to provide input on specific marketing decisions based on their presumed market and customer knowledge. In the following phases, their function moves from an advisory role to an initiator role, and finally an orchestrating role. These three phases can be considered as top-class CI roles. In the advisor role, the CI function gets a stronger, say in marketing decision making, and is consulted as an important advisor whose input is valued and strongly taken into account. As an initiator, the CI function develops independent marketing proposals in close cooperation with the marketing department. The orchestrating role is observed very infrequently; it implies that CI is embedded within marketing decision-making and becoming a driving force responsible for orchestrating customer contacts over the multiple available channels.

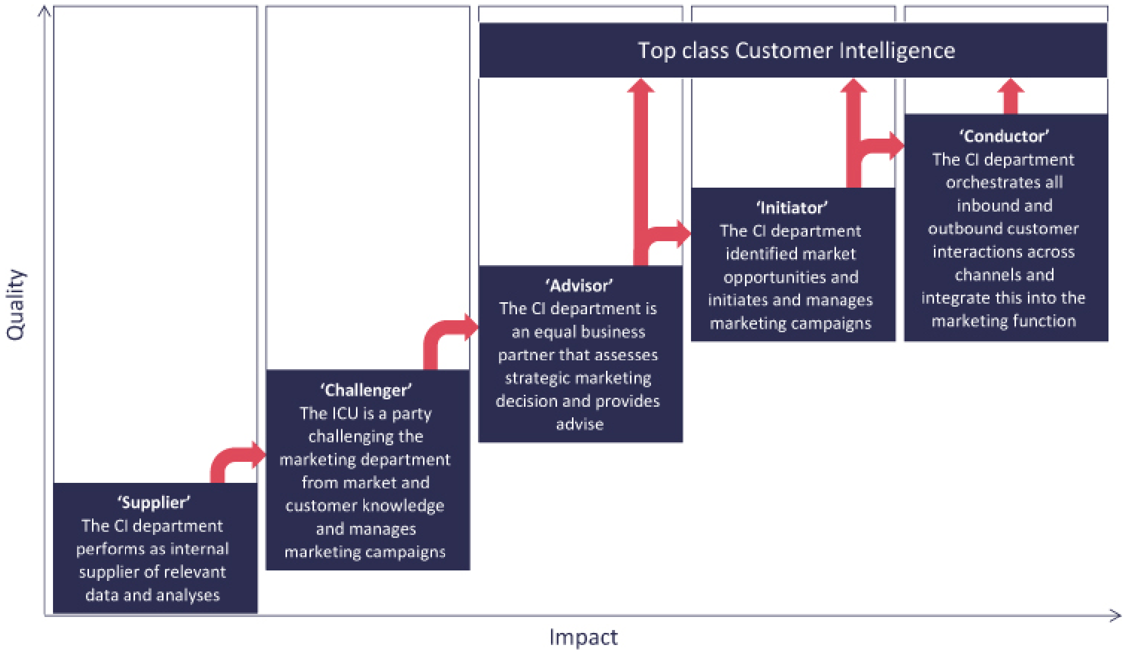

Top class Customer Intelligence
'Supplier'
The CI department performs as internal supplier of relevant data and analyses
'Challenger'
The ICU is a party challenging the marketing department from market and customer knowledge and manages marketing campaigns
'Advisor'
The CI department is an equal business partner that assesses strategic marketing decision and provides advise
'Initiator'
The CI department identified market opportunities and initiates and manages marketing campaigns
'Conductor'
The CI department orchestrates all inbound and outbound customer interactions across channels and integrate this into the marketing function
Quality
Impact

FIGURE 12.2 Changing role of the customer intelligence department

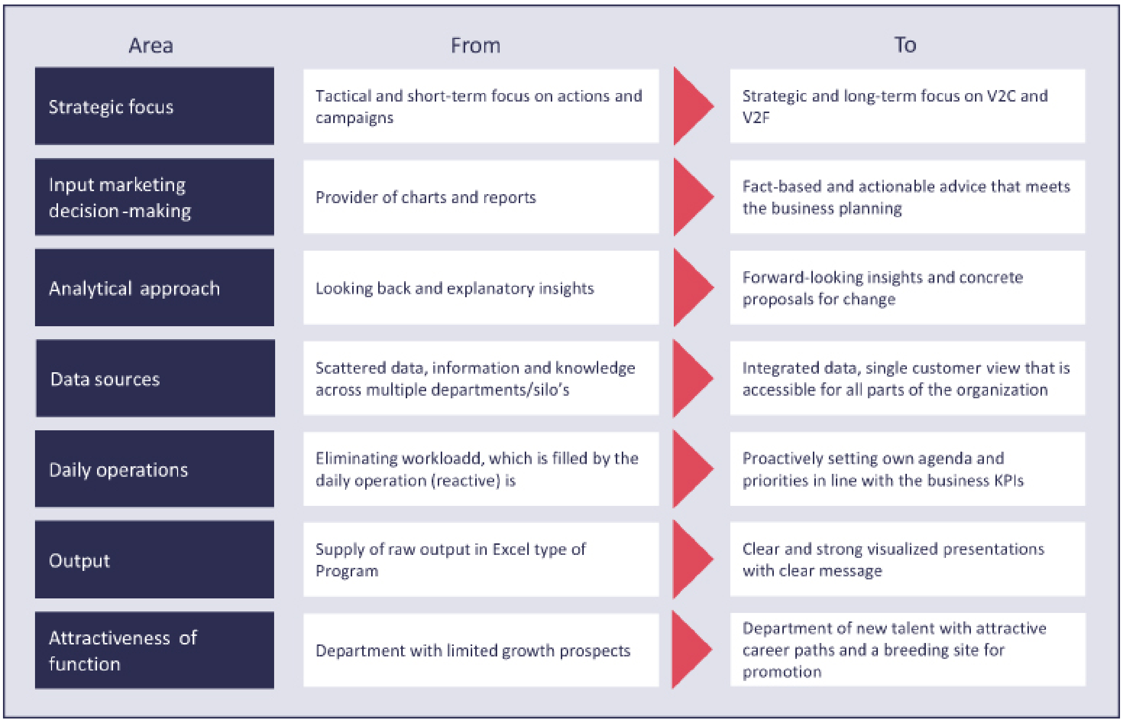
Source: Adapted from Verhoef, Hoekstra, and Van der Scheer (2009)

12.2.2 Changing focus

The changing role of the analytical department generates a change in focus, scope, and available capabilities within the department. Based on our years of experience with organizations that have made a strategic choice to bring the intelligence function to a higher level, we present an outline of the desired changes and resulting outcomes these organizations often envision and experience (see Figure 12.3). In sum, we observe that functions move from a tactical focus to a more strategic focus. As a consequence, their input in decision making is changing and they also use a different analytical approach.

Instead of looking back, a more forward-looking predictive approach is embraced. They also recognize that an integrated view of customer and market data is required. In doing so, they will be a driving force within the organization to stimulate data integration developments. In terms of daily operations, we observe a shift from reactive analyses to more proactive agenda-setting analyses. This induces a strong focus on clear reports and visualizations of results. Overall, these changes will result in a workplace that attracts young, talented, and ambitious analysts.

The main challenge for firms to achieve these transformations lies in developing an intelligence team with the right skills. We will elaborate on that in the following sections.



Area	From	To
Strategic focus	Tactical and short-term focus on actions and campaigns	Strategic and long-term focus on V2C and V2F
Input marketing decision-making	Provider of charts and reports	Fact-based and actionable advice that meets the business planning
Analytical approach	Looking back and explanatory insights	Forward-looking insights and concrete proposals for change
Data sources	Scattered data, information and knowledge across multiple departments/silo's	Integrated data, single customer view that is accessible for all parts of the organization
Daily operations	Eliminating workloadd, which is filled by the daily operation (reactive) is	Proactively setting own agenda and priorities in line with the business KPIs
Output	Supply of raw output in Excel type of Program	Clear and strong visualized presentations with clear message
Attractiveness of function	Department with limited growth prospects	Department of new talent with attractive career paths and a breeding site for promotion

FIGURE 12.3 Shifting focus of the analytical function

12.3 BUILDING BLOCK 1: PROCESS

The "process" focuses mainly on how, within firms, analytical projects are defined and executed. We have visualized this process in the analytic cycle. Within this cycle, we distinguish five important phases (see Figure 12.4).

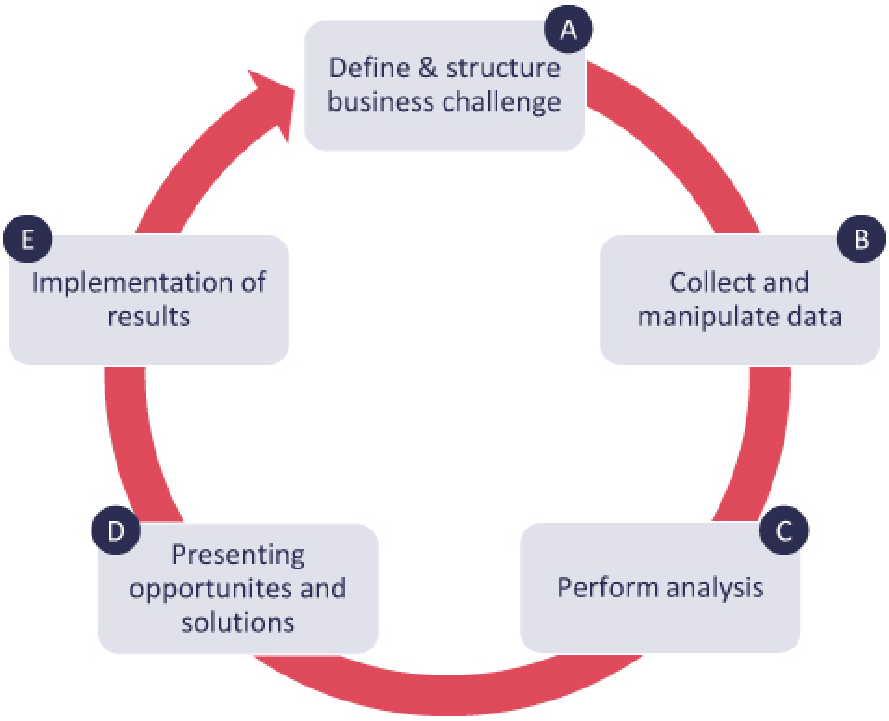

A
Define & structure business challenge
B
Collect and manipulate data
C
Perform analysis
D
Presenting opportunites and solutions
E
Implementation of results

FIGURE 12.4 Phases of the analytical cycle

The analytical cycle starts with a business challenge and has the ultimate goal of finding a solution for it. As outlined, this process seems rather straightforward and fits with our preferred analytical problem-solving analysis and data exploitation strategies from Chapter 7. However, many analytical projects fail and do not lead to concrete business applications. In 2015 Gartner[6] estimated that 60% of big data projects fail, and it seems that will not get better any time soon. In 2019 Gartner analyst Andrew White predicted that up until 2022, only 20% of analytic insights will deliver business outcomes.[7] The main reasons why the intended impact and added value of analytics are frequently limited include:

- Unclear starting point for the analysis
- Restricted availability of data and data quality issues
- Unclear and not impactful presentation of analytical insights and promising opportunities
- Initiatives cannot be implemented or get stuck in the Proof-of-Concept phase.

For each phase of the analytical cycle, we therefore provide a number of directions that help to increase the success of using data and analytics to solve problems and to create real value for the organization and its customers.

12.3.1 Define and structure the business challenge

The starting point of the analyses should always be related to a clear business challenge. This gives the right focus, the right priority for use of scarce resources (time, capacity, and budgets), and creates a common acknowledgment of the relevance and importance of the analyses carried out. To obtain a well-defined and to-the-point business question which drives the analyses, we believe the following points are crucial:

- Have a good initial discussion with all relevant owners of the problem within the firm
- Determine whether the project focuses on incremental improvement or optimization of current business and marketing practices or aims to achieve strategic changes and growth opportunities
- Determine the intended change in the KPI of interest and determine which definitions are used within the firm (see also Chapter 3)
- Validate and set measurable targets/objectives based on existing business plans, strategy papers, outlooks, and reports
- Get a grip on the problem by structuring and quantifying it based on existing knowledge and data.

For a more detailed discussion of the first step of the analytical cycle, we refer to Chapter 11, in which we discussed the approach of Opportunity Finding as a powerful method for fact-based problem solving.

12.3.2 Collect and manipulate data

After phase one, the process of collecting and processing the data you need to test hypotheses and perform the analysis starts. The challenges of this phase are often underestimated because data is not readily available. Relevant data is frequently contained in systems that are not easily accessible or must come from outside the organization. Additionally, poor data quality creates obstacles to getting started with data. Besides this, data preparation is also often very time-consuming, especially when it's stored in a raw or unstructured way. In Chapters 4 and 5 we discussed these issues and suggested several solutions. Here, we want to offer some recommendations on how you can still make progress despite these issues.

- Be creative in considering which sources inside and outside the organization can be used to gain new insights. In other words: look beyond the data sources you use every day.
- Start with the data that is available instead of putting lots of energy into unlocking new and unfamiliar sources in advance.
- Do not let data quality concerns take over and paralyze the process. There are solutions to overcome this (as discussed in Chapter 4). If necessary,

work with assumptions and error margins which can be validated later in the process.

- If sources are of crucial importance, but difficult to access, request budget, capacity, and priority to ensure that they become available.
- Consult the legal department about restrictions regarding the use of privacy-sensitive data to prevent finding out that the data cannot be used later on in the process.
- Manage expectations of other stakeholders about what can be achieved in the short and long term regarding the data.

12.3.3 Perform data analysis

In this phase, the analyses are executed to achieve new insights and models are developed to validate hypotheses and create solutions. Chapters 8 and 9 pay extensive attention to this. Concerning the process, there are some points to note:

- Make smart use of existing insights that have already provided answers to questions.
- Make a distinction between giving a 100% reliable outcome versus an outcome that is reliable enough to make a decision or to take action.
- Always validate results based on existing outlooks, business plans, strategy documents, and management reports.
- Make sure the right people are involved and informed during the analysis process.
- Provide a schedule and overview of the steps that you take.
- Ask for active support during the analysis phase, including time and capacity for coordination, consultation, and feedback.

12.3.4 Presenting opportunities and solutions

In this phase, it is very important to share analysis results with others in an impactful way and to ensure that decisions are made and follow-up steps are defined. In Chapter 10 we discussed the importance of a clear storyline and visualization for the presentation and interpretation of analytics. We also extensively discussed methods and guidelines on how to do this with impact. In addition to this, we would like to provide various tips and tricks to ensure that analysis results are properly communicated and that something is done with them.

- Do not underestimate the time it takes to translate analysis results into a piece of powerful advice. A common pitfall of analysts is to work until the last minute on the analyses and not allow enough time to make them presentable.

- Use a top-down approach (Pyramid Principle) to structure and present the results and conclusions of your analyses.
- Always write down the storyline on paper first to avoid spending unnecessary time making fancy PowerPoint slides, which may be unnecessary for presenting your advice later.
- Do not make the presentation of analysis results unnecessarily complex or detailed. Keep in mind that the audience must understand the meaning and impact at once.
- When presenting your advice, be clear about your pitch: what do you want to get out of it, when are you satisfied?
- Keep in mind that advice has an impact if:
  - – New insights, ideas, and initiatives are shared
  - – Solutions are presented
  - – Common goals are achieved
  - – Plans are made
  - – Commitment exists
  - – Decisions are made
  - – Clear follow-up actions are identified.
- When distributing documents, make sure that they are self-evident, and that no misrepresentation can arise.

12.3.5 Implementation of results

The final phase of the analysis cycle is to ensure that the outcomes of analyses lead to better decisions, successful actions and campaigns, and implementation of data-driven products, as discussed in Chapter 11. Depending on the desired outcome, an implementation based on analysis results can lead to relatively simple adjustment of policy or a translation into business rules (like price changes, campaign selection criteria, or budget allocation for different channels). The implementation of analysis results is much more comprehensive and complex if a new data-driven proposition or product has to be developed, whereby new models and algorithms have to be implemented. Often a new process is started in this phase, in which multi-disciplinary teams have to design and implement the solution. Many organizations opt for Scrum, an Agile way of working to ensure that an organization becomes more flexible and effective. Although this way of working is very relevant in this phase of the analytic cycle, it is beyond the scope of this book to go into detail. For more information about the Agile way-of-working, the different methods and techniques used, and best practices, we would like to refer to the work of Jeff Sutherland (2015). He is one of the creators of Scrum, a framework for developing, delivering, and sustaining complex products. Along with Ken

Schwaber, he wrote and maintains The Scrum Guide,[8] which contains the official definition of the framework (2020).

It is clear that the role of the analyst changes during the implementation phase. It does not, however, become less important. First of all, when models are put into production, in which case a translation of model parameters to a production environment must take place, the role of the data analyst is crucial. In addition, the analyst has an important role in testing, validating, and refining models based on the first results during the pilot phase and monitoring of the results in the rollout phase.

In this section, we would like to mention some guidelines to make this phase more successful. Note that this list is not exhaustive, because, as stated earlier, this phase affects many parts of the organization and can be very complex.

- Make sure that a convincing business case is made for an estimate of the expected investments and returns.
- Ensure commitment and involvement of all major stakeholders. This phase is not a handover to the IT department: both the business and the analysts must remain connected with the implementation.
- Start small with a pilot or Proof-of-Concepts or Minimal Viable Product (MVP) and scale up when it is successful.
- Use Agile or Scrum to achieve initial results in short strokes with a multidisciplinary team and to be able to make quick adjustments.
- Determine what the success factors are and make them measurable.

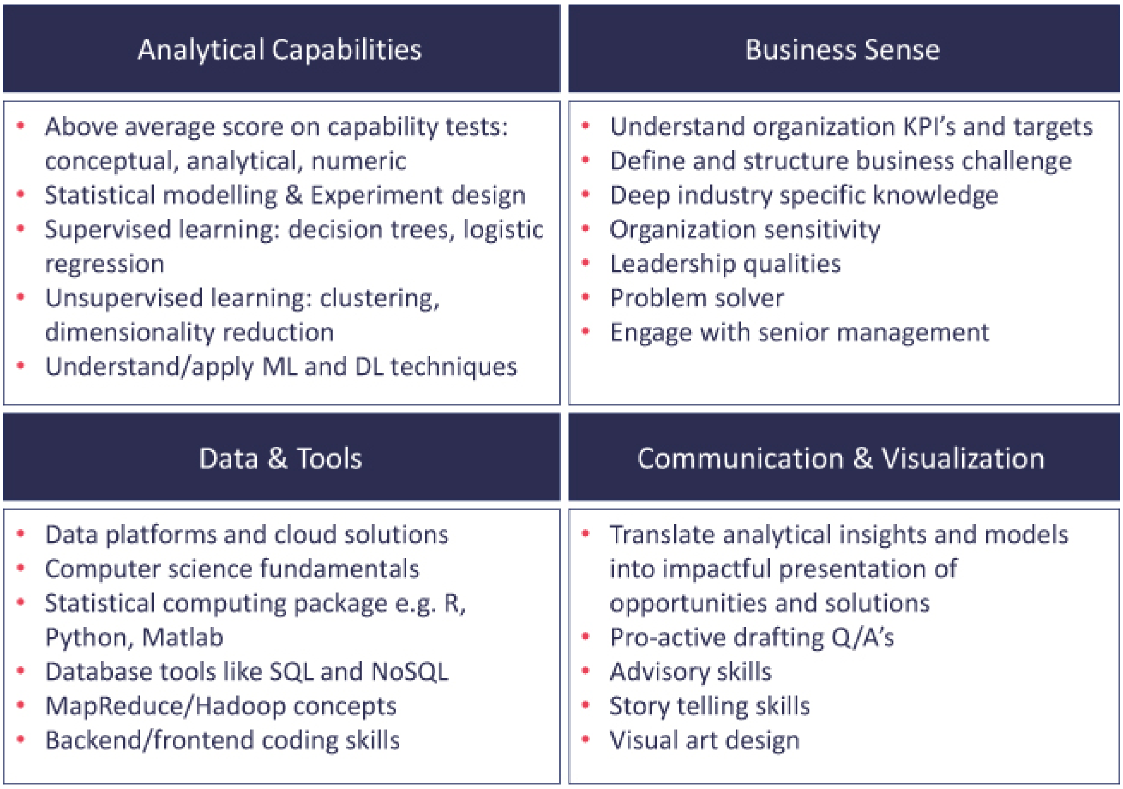
12.4 BUILDING BLOCK 2: PEOPLE

The transformation of the analytical departments also requires the hiring of employees who can effectively work in these departments. The biggest challenge is that it is difficult to find employees who fit in. There is a shortage of good analysts who can perform well in this highly demanding job. In addition to this market challenge, there are five people challenges:

1. What is the profile of the employees required?
2. How do you develop an excellent data analytics team?
3. How do I acquire the right analysts from the job market?
4. How do I keep the good analysts, given the shortage in the market?
5. How do I make analytical competence scalable, given the still growing demand?

12.4.1 Analyst profile

Although we mainly refer to the role of analyst, it almost feels like an old-fashioned term. In this era of big data and AI, firms are no longer looking for analysts, but instead for data scientists, data engineers, and data citizens. To get a better idea of why different roles exist within the analytical working field and how they differ from each other, we first want to outline the basic competencies that we believe a modern analyst should possess. These skills or individual capabilities can be divided into four areas (see Figure 12.5):



Analytical Capabilities

- Above average score on capability tests: conceptual, analytical, numeric
- Statistical modelling & Experiment design
- Supervised learning: decision trees, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Understand/apply ML and DL techniques

Business Sense

- Understand organization KPI's and targets
- Define and structure business challenge
- Deep industry specific knowledge
- Organization sensitivity
- Leadership qualities
- Problem solver
- Engage with senior management

Data & Tools

- Data platforms and cloud solutions
- Computer science fundamentals
- Statistical computing package e.g. R, Python, Matlab
- Database tools like SQL and NoSQL
- MapReduce/Hadoop concepts
- Backend/frontend coding skills

Communication & Visualization

- Translate analytical insights and models into impactful presentation of opportunities and solutions
- Pro-active drafting Q/A's
- Advisory skills
- Story telling skills
- Visual art design

FIGURE 12.5 Multi-disciplinary skills of a modern data analyst

- Analytic skills
- Data and tools
- Business sense
- Communication and visualization.

All the skills listed in Figure 12.6 are important. Unfortunately, these are often not skills that one person possesses. For example, a specialist with excellent analytical skills frequently lacks business sense and finds it very difficult to communicate effectively. This does not imply that one should not hire top analytical talent! These experts are needed, especially given today's strong big data challenges and the sophisticated analytical models required to solve them. Another strategy would be to go for someone scoring averagely on every dimension. However, it is unlikely that these people will move the organization forward, as they might have a lack of innovation in every dimension. In terms of building an analytical competence with the right people, firms should strive to develop a well-rounded team of professionals in which each of these capabilities is sufficiently present at a high level.

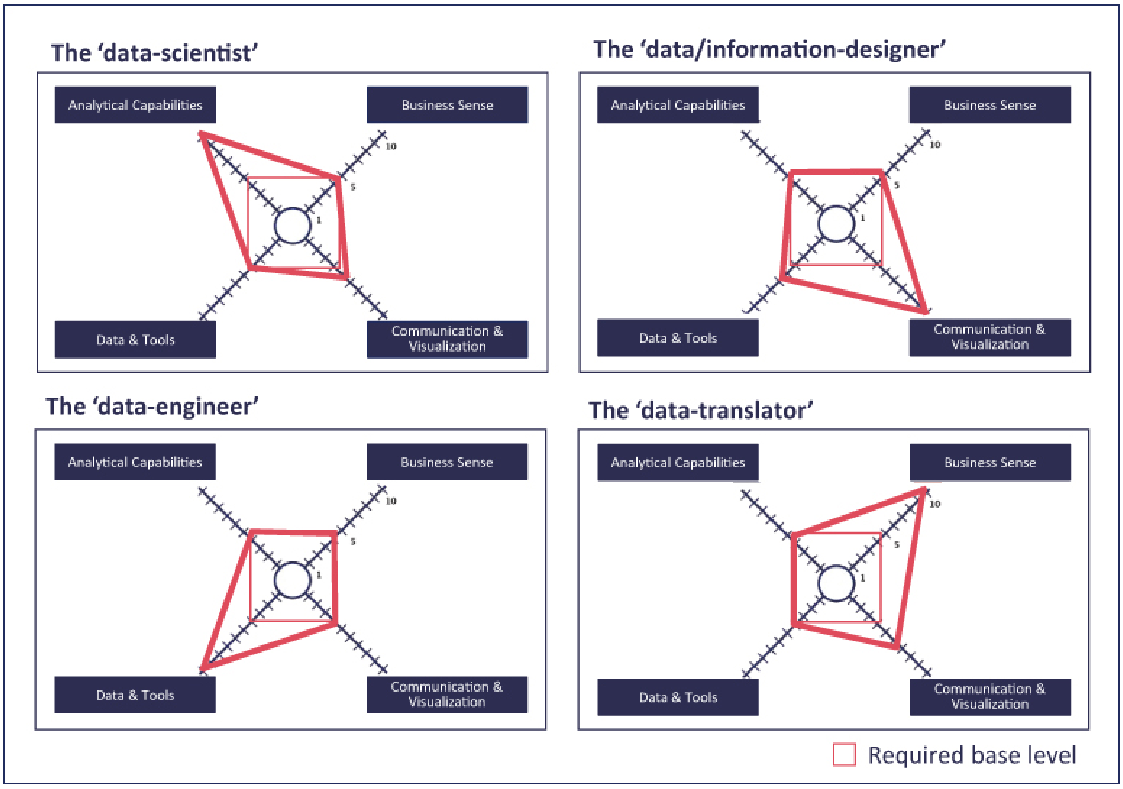


FIGURE 12.6 Typical profiles in working field of analytics

12.4.2 Team approach

It is important to build a team that contains each of these capabilities at a high level, and where employees can excel in their specific roles. Good collaboration between the different profiles arises from a common way-of-working method that embraces everyone and empowers each role (as we discussed in Section 12.3 Building Block 1: Process). This means that when facing business challenges, the optimal setup must always be chosen to make the most impact as a team. To help set up the team approach, it helps to map employees on their specific profiles and support them to excel in it. Based on the four capabilities, four common profiles[9] can be distinguished (see Figure 12.6):

1. The data scientist: has very strong analytical skills. Is creative in the optimal use of data and translating these into relevant and actionable insights. Knows how to develop the most reliable prediction models based on various statistical and advanced machine learning techniques.
2. The data engineer: has strong data & tools skills. Connects the data sources, packages the machine learning modules, and integrates with all other systems. Brings the front-end to life through a functional and

efficient user interface. Documents the code, maintains logs, and adopts software engineering standards.

3. The data/information designer: has strong communication skills. Makes the analytical insights and solutions functional and pleasing to use. The data designer develops mockups and detailed design prototypes. Makes the data insights consumable by identifying the right kind of charts, interactivity, and visual design to use. Is a master storyteller in understanding and presenting data.
4. The analytics translator: has strong business sense skills. Is a mediator between the decision-makers and data analysis experts. Has the task of managing the analytical cycle for various business challenges. Translates the business question into an analytical assignment and translates analytical insights and models into opportunities and actionable end-solutions for the business. The translator brings the data scientists, data engineers, and information designers together and empowers them to give their best. And is also responsible for change management and the adoption of the solution by business users.

The profiles described above are not intended to be exhaustive. In practice, we also see other profiles and job names that partly overlap with the above profiles or profiles that are based on combinations of skills. For example, the translator is also referred to as a data consultant. This is often a senior person who has progressed from a more specialized role as a data scientist. This is a typical profile in which a strong business sense is coupled with strong analytical skills.

It is important to note that building an analytical team is done gradually, as shown in Figure 12.7. Companies often start with a few independent data analysts and/or specialized consultancies. At a certain point the need for an in-house analytical team becomes apparent and a small team is installed This team must then be trained using specially developed internal programs or external programs from business schools and/or training agencies. At some point there must be sufficient knowledge and skills within the company to enable it to independently develop analytical competencies. At this stage, it becomes important to develop career paths for the analysts in the team.

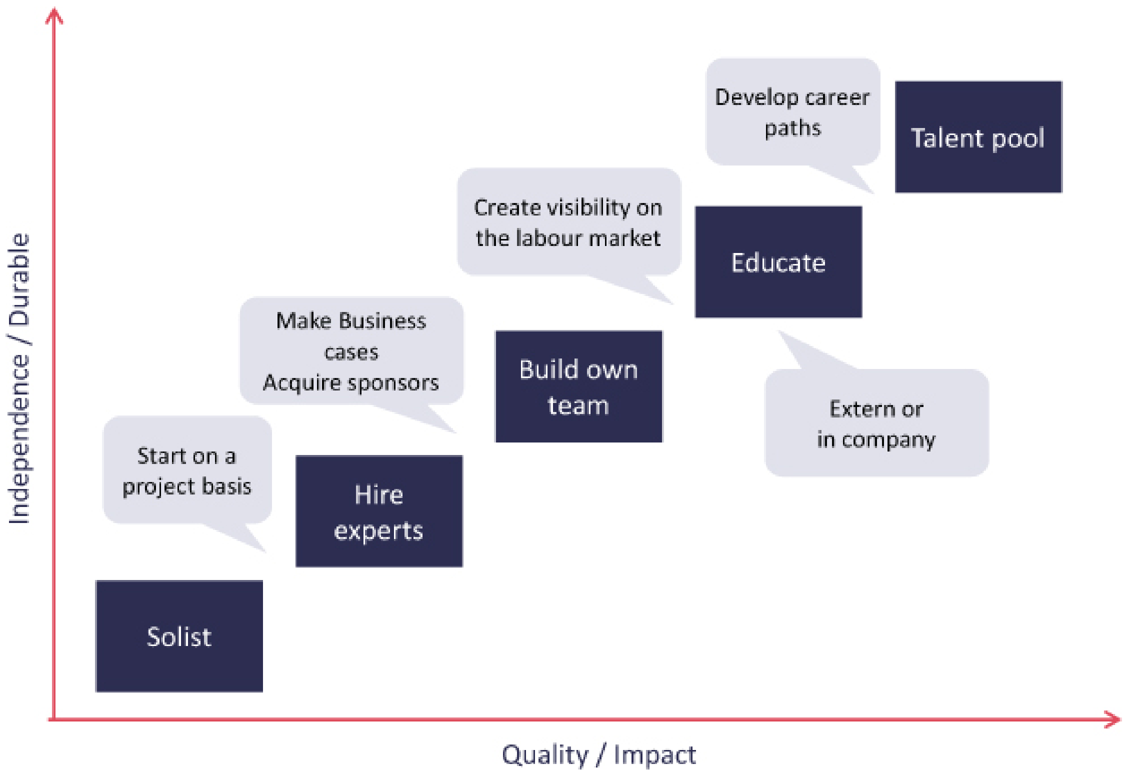


FIGURE 12.7 Stepwise development of analytical competence within an organization

12.4.3 Acquiring new talent

The supply of good analysts is much lower than the demand (see Figure 12.1). This is a challenge for companies looking to build analytical competence. One can easily develop the ideal department in terms of required profiles. Finding the right people is, however, far from easy. One of the problems is that MBA programs used not to be good at developing analytical skills, as they mainly focused on managerial skills. Fortunately, that has changed. In recent years, numerous university-level programs that focus on information technology, data science, AI, etc., have been introduced to the portfolio. We also see that well-trained students with strong econometric and analytical skills are increasingly choosing positions within marketing, commerce, and tech companies. Yet, the competition to attract these new talents is still great. That is why we make several suggestions which are useful for attracting top analytical talent:

- An obvious suggestion is that companies strive to build employee Brand Equity (BE) (Tavassoli, Sorescu, & Chandy, 2014). Young professionals prefer companies, such as Google, that are innovative and have a strong Brand Equity.
- An active training plan for young analysts can also help. Courses in which young analysts are trained on the job are considered a unique selling point.

- Reach out to universities and business schools where talented young people are trained. Companies could participate in analytic classes with data-based cases or organize hackathons, where they meet students and familiarize them with companies and their analytical job opportunities.

12.4.4 Talent retention

Given the shortage of talent and the investments required in the education of acquired talent, employee retention is of urgent importance for analytical roles. Furthermore, analysts build up tacit knowledge that it is very valuable to retain. Hence, analyst retention should be a top priority for firms building up analytical competence. Holtom *et al*. (2008) warn that in the industry, chronic shortages of qualified employees have driven up the costs of turnover much faster than the rate of inflation. This reenforces the importance of the loyalty of talented people. Loyal analysts are often not the top employees within the firm. So there is a danger that if firms do not invest in the loyalty of talented people, their MI or analytical department may slide down the ladder and the challenge to build up a fully- fledged analytical department will have to start all over again. We have seen this happen in many organizations. For example, within a service firm, at some point talented employees became unsatisfied and, as a consequence of a lack of support from top management and ongoing uncertainty, started looking around and found jobs in other industries. For talent retention, several issues have been deemed important, including salaries and atmosphere but, given the focus on young talent, the most important factor is personal development (e.g., De Vos & Meganck, 2009). We have the following suggestions for firms:

- Develop attractive future career opportunities within analytical departments, and potentially even outside the department as well. If former analytical employees become active in other departments (e.g., finance, marketing), they may become active ambassadors for big data and AI approaches within the firm.
- Create sufficient opportunities for personal development and freedom to innovate. Data-scientists are looking for ways to use data to solve problems and are less attracted by routine tasks such as campaign selections or updating marketing dashboards.
- A corporate vision on big data usage and building up the analytical competence is an important ingredient for creating a working environment in which talented people feel they can contribute to the success of the organization and career opportunities are sufficiently available. Organization literature refers to this: there needs to be sufficient organizational support (Holtom *et al*., 2008).

12.4.5 Scalable analytics

Analytics talent is a scarce and valuable resource for every company, but particularly for those with a desire to be data-driven: the demand for this is only growing. The recruitment of new talent and the increasing expansion of analytical competence teams is not only very challenging but also not the complete answer to meet the growing demand for analytical talent in organizations. Beyond attracting young talent, firms also have some other options at their disposal to work around the shortage of analytical talent and make the deployment of analytics scalable:

- One option is to train existing employees in the field of traditional marketing research and database marketing. However, our experience tells us that this is not as easy as it sounds. The knowledge of these employees is to some extent outdated in a big data and AI era and training them into a new work mentality with a stronger focus on business, communication, and visualization is not easy.
- A second option is to ensure effective work processes within organizations and to automate specific processes. This means less talent is needed. For example, customer selections can probably be automated. Reporting can also be automated, especially for continuous data collection.
- A third option to make analytics scalable is by ensuring that everyone can contribute to the use of data and analysis to create value for the organization and its customers. This is not solely the responsibility of analytical experts. There are more and more examples of companies that embrace this strategy. They invest in developing the basic skills of all employees within all layers of the organization, from decision-makers to users, at such a level that they understand the possibilities and impact of using data and analytics for their own work. These employees learn how the analytical cycle works, how to tackle fact-based problem solving and how they can easily use available data and analytical insights.

12.5 BUILDING BLOCK 3: SYSTEMS

One of the initial reactions of firms concerning big data is an immediate impulse to build large systems in which all big data are integrated, as well as analytical and support tools This approach is in line with what we have seen when firms were implementing customer relationship management. Firms invested millions in integrated software systems such as Oracle, Microsoft, Salesforce, etc. One of the major lessons learned was that firms should not focus on the technology but on what the technology can do for customers (Rigby, 2014; Verhoef & Lemon, 2013). This major lesson is easily forgotten in a big data and AI environment where every IT-oriented consulting and tooling firm is communicating the impressive opportunities of big data and AI. For firms, there are a vast number of options to choose from when it comes to providers and software solutions. In the 2020 edition of the marketing technology landscape supergraphic of chiefmartic.com,[10] about 8,000 vendors are represented across six categories. Chiefmartec produces this inventory of the number of vendors of martech solutions every year, and it has clearly taken off since 2011, with the number of vendors still exponentially growing (see Figure 12.8).

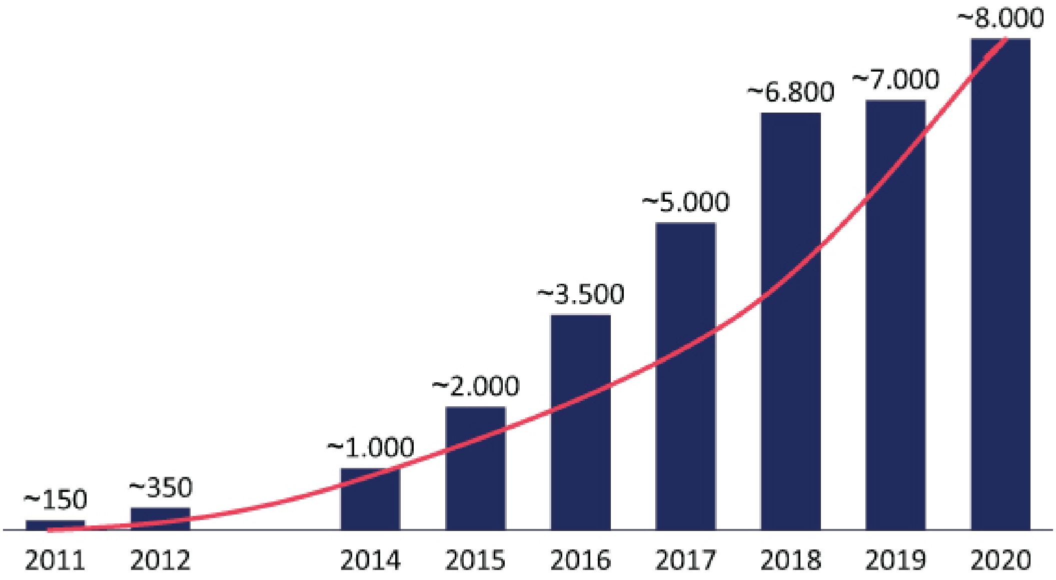


FIGURE 12.8 Number of vendors in marketing technology landscape represented in supergraphics of chiefmartec.com[11]

As we have already discussed in earlier chapters, sound expectations should drive business decisions on big data and AI. This also holds true for system investments. It is just as true to say that, with big data systems, technology should never drive decisions! For big data and AI marketing solutions, systems should be user and customer-driven instead of technology-driven. In that sense data scientists should have a strong say in the development of data systems and marketing should be an advocate of the customer in this process. Jointly with IT they should come up with workable and scalable solutions.

We observe many firms striving for a completely new big data system that is often not feasible and not necessary. Many firms already have good systems in place for different solutions (e.g., CRM or enterprise resource planning). The important thing about an analytical big data environment is that systems can be linked. We advocate an approach in which old systems and new big data solutions (e.g., Hadoop, data lakes) co-exist within an organization, loosely divided into four different “layers” (see Figure 12.9):

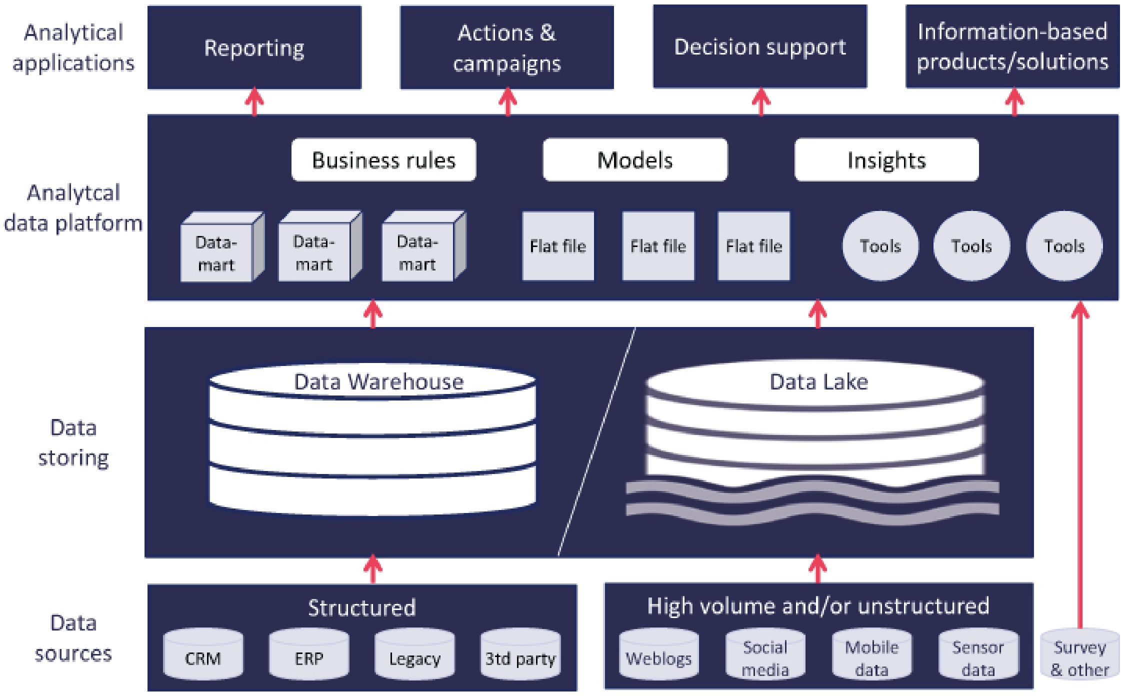


FIGURE 12.9 Different layers of a big data analytical system

1. Data sources
2. Data storing
3. Analytical data platform
4. Analytical applications.

It is very important to emphasize that the idea of using one big system is generally an illusion. Firms will use different systems in the different identified layers, but even within the layers (e.g., analytics), different software can be used depending on the objectives to be achieved.

12.5.1 Data sources

Within a firm, there are multiple data sources. Many of these data sources function on a stand-alone basis. For example, a firm can have a good billing data system, which functions well for billing purposes. Similarly, they could have databases in which operational data on customer service requests are stored. Before starting a big data system, firms should first make sure that the data in each of these internal databases is accurate and reliable. Data analyses are useless when inaccurate data are analyzed, despite datasets being large. In this regard, the saying "garbage in is garbage out" holds in a big data era as well. In many firms there have been efforts to integrate databases with a similar main focus. Specifically, firms have invested in CRM databases in which billing data, marketing data, customer service data etc., are integrated to create an overview of each independent customer. Integrated customer databases are of essential importance for customer-level analyses and are a key ingredient to successful predictions on churn, lifetime value, etc. Traditionally, these databases are then available in a data warehouse. Although these data warehouses could involve millions of data points depending on firm size, the number of customers, etc., they are relatively simple compared to the requirements of big data currently confronting firms.

As discussed, big data development leads to large data volumes, but probably more important are the different data sources with different data structures, which cannot be easily linked. For example, service interactions in a call center can be linked to individual customers. Interactions on websites with different devices, on the other hand, are more difficult to link to an individual customer (see Chapter 5). Similarly, unstructured data on websites, blogs, and social media cannot be fully linked to individual customers. Furthermore, aggregation levels of data may differ (e.g., brand level vs. customer level). It is impossible to store these data in an integrated data system. We strongly advise that in a big data strategy, firms should not strive to integrate all different data sources. They should only integrate data that can easily be integrated because there are strong identifiers between the data sources (e.g., customer ID for CRM database, Google Analytics ID for connecting the use of different devices). Instead, they should create a big data platform in which several databases are available and can easily be accessed by analysts based on their respective analytical questions.

12.5.2 Data storage

Several data sources can be saved in a large data warehouse. This data warehouse is typically internal to the firm. Many firms now also use the cloud to store data. We refer to Chapter 5 on the specifics of data storage.

However, due to the rise of big data (such as sensor data, social media and mobile data), it became apparent that it was difficult for data warehouses to work with unstructured data. Since organizations nowadays want to analyze, use, and manage a greater diversity (and amount) of unstructured data, new infrastructures have being devised to deal with this. We call this new form of data storage data lakes. The biggest difference between data warehouses and data lakes is that data warehouses consist of pre-structured data. This has the advantage that you get more straight-forward answers to questions such as the number of products sold, number of website visitors, and incoming telephone calls. Data lakes, on the other hand, are used to collect much larger amounts and different types of raw (unstructured) data. This data is stored and can then be analyzed for a variety of purposes and allows analysts to discover interrelationships between new data. A data lake provides a platform that makes vast amounts of data ready for business purposes, almost in real-time. This delivers faster results than the traditional data approach. A prerequisite, however, is that there must be sufficient knowledge about the quality and use of new big data sources.

In summary: Data lakes and data warehouses both meet a need, and complement each other. There will always be a need within organizations for both structured and unstructured data, where the structured data consists of selection and processing of unstructured data. By choosing only a data lake, an organization does not get the most out of the data, while with a data warehouse alone there is too little flexibility to be able to respond to changing needs from the business. A combination of the two makes the data more productive.

12.5.3 Analytical data platform

The big data analytical system involves a set of databases that should be accessible by analytical specialists. These databases can be the more traditional sort available in standard data warehouses, but for large volumes and/or unstructured data, specific big data solutions such as data lakes can be used. Firms should be mindful of security issues and should only provide authorization to access specific data to internally selected employees. We must emphasize that the analytical data system is not only a data platform. Using this analytical data system, a database specialist could aim to combine specific data sources through, for example, data fusion and SQL queries. For example, when studying the churn of customers, one could combine CRM data with data on social media (e.g., likes for specific customer segments). Or analysts, when doing longitudinal analyses of brand sales, could build a database consisting of weekly brand sales, weekly advertising efforts, and weekly social media likes.

These datasets should be structured in such a way that they can be analyzed in the program being used. In practice, the term “data mart” is used to refer to datasets derived from the data warehouse. A data mart is the access layer of the data warehouse environment that is used to get data out to the users. Data marts are typically small subsets of the data warehouse and are usually oriented to a specific business line or analytical team. Whereas data warehouses have an enterprise-wide depth, the information in data marts pertains to a single department. From this data mart flat files can be extracted which can be further integrated and processed to be used in analytical tools.

Analytical tools or packages focus on the statistical, econometric, and linguistic analysis of the data. Statistical packages have a full set of statistical analyses available that can be used for multiple analytical purposes. Analytical tools also have functionalities for further data processing, such as creating all kinds of new variables by classifying or combining the different input variables. In the past, firms typically chose a preferred analytical tool. This tool could be either (partially) specifically developed for the firm or a standard statistical software package, such as SPSS (nowadays part of IBM) or SAS.

These software packages have proved to be very useful and have become more user-friendly over the years, with the introduction of windows-interfaces, help functions, etc. However, big data development has changed the statistical analysis field dramatically. The traditional statistical packages are not yet fully suited for big data analytics. We specifically observe the following developments:

- Due to the presence of more unstructured data, new analysis techniques such as text mining and image recognition are being used more frequently.
- Big data development has attracted big data scientists, who are using and creating their own programs.
- Open-source packages, such as R and Phyton, that are being developed through a large knowledge base in the online community, can handle large databases in the cloud and estimate the more complicated models needed when analyzing big data.

A specific disadvantage of the use of analytical software (either standard or open source) is that the output is frequently not user-friendly. Therefore, data visualizations are indispensable for analyzing large amounts of information. Using visual elements such as charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. There are lots of tools with which you can visualize your data. Some well-known tools are: Tableau, Power BI, and Qlik. Besides this, Excel of course still offers the functionalities to create your own visualizations based on aggregated data. The appropriate tool to use depends on who the users are and whether they are used for an ad-hoc purpose or for structural information provision (for example by means of a dashboard). In addition, one tool has more functionalities than the other, is easier to use, and has a different cost. Therefore, firms often use a mixture of tools.

12.5.4 Analytical applications

The fourth layer of big data systems involves analytical applications—that is, reporting, actions and campaigns, decision support, and information-based products and solutions. We have already discussed applications of how analytics support marketing on a functional level and examples of applications in the customer-firm interface in Chapter 11. We will now focus on the link with operations in actions and campaigns.

12.5.4.1 The link with operations: actions and campaigns

In our discussion on the different layers of systems, we mainly took a data and analytical perspective. However, it is important to note that operational systems and processes are directly linked with analytical systems (see Figure 12.10). Operational systems involve, for example, software that links daily operations (e.g., a call center) to data and specific business rules or selection rules. Business rules and selection rules are the results of analytical exercises. These business rules could be that, for example, only customers that have a high customer lifetime value (CLV)—determined by the outcomes of analytics—get a reduction on the price they pay for a specific contract. In a call center, the agent using an operational system will be able to observe the customer information on a computer screen and using this system can then make offers to customers to renew a contract. Notably, although the interaction occurs in the operational environment, it is sourced back in databases. Building on the call center example, the interaction with a customer discussing the renewal of a contract can be included in the database, recording data on the offer, the outcome of the offer (renewal or not), and so on. In a big data analytical system, the data can be enriched with the unstructured data on the conversation between the agent and the customer (e.g., Verhoef, Antonides, & De Hoog, 2004). Selection rules typically involve a rule on how to select customers who will receive a specific targeted offer (e.g., an email campaign with a target promotion). These selection rules are very common and used in many sectors, including retailing and financial services. They are mainly used in outbound campaigns.

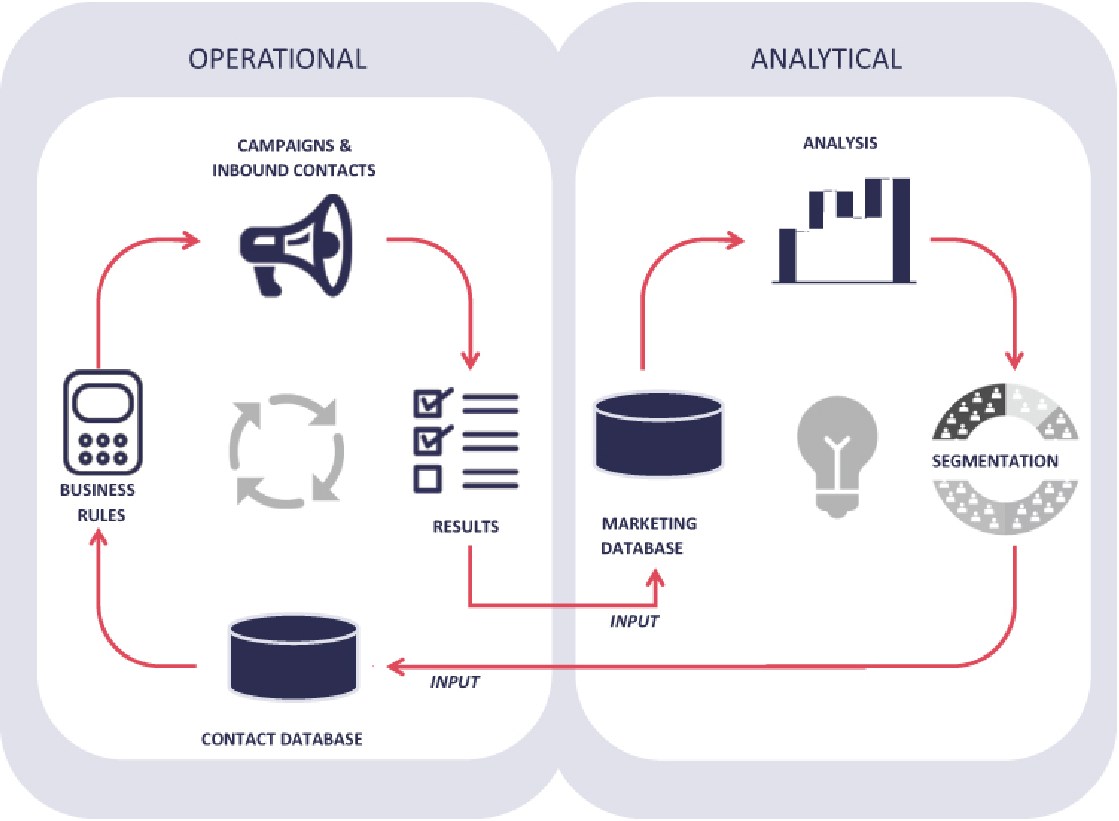


FIGURE 12.10 Linking data, analyses, actions and campaigns

In a digital environment, operational systems become very important as interactions on websites can be customized based on automated analytical exercises, which are themselves based on different business rules that consider the browsing behavior of customers at websites, the CLV of customers etc. The most famous example of this technique is probably that of firms like Amazon, which suggests products to customers that have been purchased by other customers who have a similar choice behavior (also called "collaborative filtering"). However, in current digital analytical environments, operations and analytics have become intertwined and the distinction between them has become blurred. With the millions of interactions occurring at websites, models can be developed that are constantly updated with relevant new information. The analyst first has to develop the "basic" model and its specific econometric specification. The estimation parameters can then be updated based on the constant digital interaction with customers on websites, apps, etc., and based on that, constantly relevant and targeted offers can be provided to customers visiting these digital channels. In Chapter 11 we discussed the example of an "adaptive personalization system," illustrating its implementation for digital audio players.

12.6 BUILDING BLOCK 4: ORGANIZATION

The organizational side of big data and AI development receives relatively less attention. There is a strong focus on attracting talented data analysts and tooling. The organization of the analytical competence within the organization is, however, important as well. We observe three specific challenges:

1. Centralization of the analytical function
2. Cooperation with other departments/functions
3. Presence of a data-driven culture.

12.6.1 Centralization or decentralization

Firms building up a data analytical competence face the challenge of where to locate this function. This is especially a challenge for firms active in multiple business units. For example, firms operating one business unit for the consumer market and another for the corporate market could have two analytical functions per business unit or one single analytical function serving both business units. Firms with global operations could have a global analytical function serving all separate country organizations or several functions serving local country organizations. Hagen *et al*. (2013) suggest that there are three models for organizing an analytical competence (see Figure 12.11).

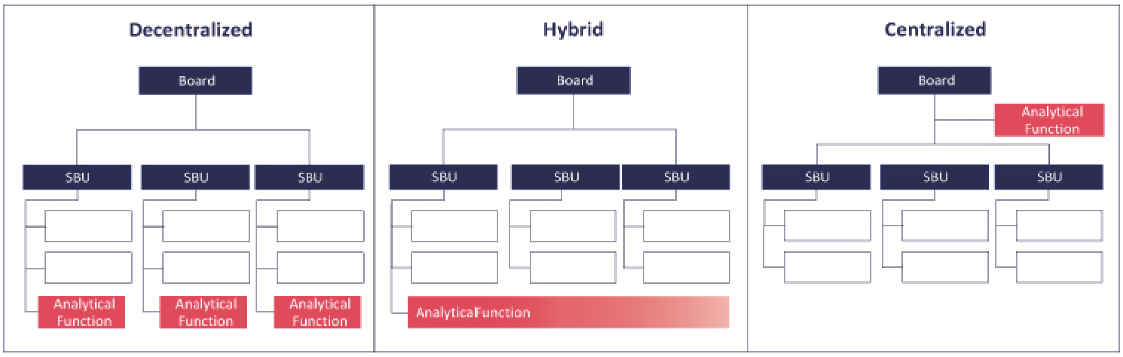


FIGURE 12.11 Organization models for the analytical function

Source: Adapted from Hagen *et al*. (2013)

In the first option a decentralized organization is chosen, where for each strategic business unit (SBU) an analytical function is set up. The advantages of this setup are that the analytical competence can be specifically developed for each SBU, serving the specific needs of each one. The function is also likely to have a strong impact on decision-making in each SBU. A strong disadvantage, however, is that functions are relatively small and inefficiencies occur. Specific knowledge and capabilities are developed in each SBU and there are no economies of scale. Further, these separate functions lack an overall strategic view. In the second hybrid model, a more decentralized function is developed under the responsibility of one SBU that serves that SBU as well as other SBUs. This may create more efficiencies and standardize solutions. However, one problem is that there will be competition among SBUs for the analytical capacity and it is likely that the responsible SBU will benefit most. In a final model, the analytical function becomes an independent centralized staff function serving multiple business units. As with the second model, this will lead to more standardization and less inefficiency. One big potential challenge with this organization format is that the function can become very independent and insufficiently connected with the different marketing departments for impactful analytical functions. This would indicate a need for analytical functions to be more closely linked to an SBU. However, there are some disadvantages to this as well, as it may, for example, lead to a lower overall analytical skill level.

In general, several rules can be used to help decide between a more decentralized versus a more centralized approach. Decentralization is preferred to centralization when:

- There are strong analytical skills within the firm
- There is a large number of analysts

- The analyst team is mature and independent
- There is a strong need for specialized knowledge in different SBUs
- Teams do not depend too much on data- and software suppliers.

12.6.2 Cooperation with other functions

One of the problems that analytical functions encounter is the level of cooperation they have with other departments and specifically the marketing department. Generally, cooperation between departments is considered beneficial. Problems can arise if departments function as separate silos that do not understand each other. Figure 12.12 shows how the MI/analytical department functions within the organization can have different views on the added value of the analytical function, and how this may result in reduced impact. It is therefore important to have a common understanding of the analysis process.

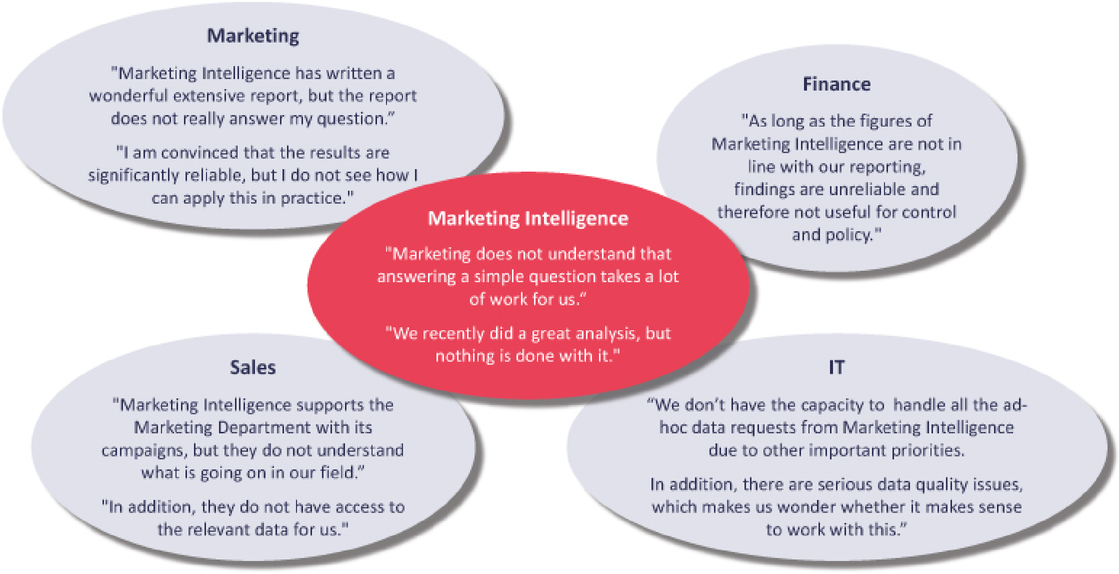

Marketing
"Marketing Intelligence has written a wonderful extensive report, but the report does not really answer my question."
"I am convinced that the results are significantly reliable, but I do not see how I can apply this in practice."
Finance
"As long as the figures of Marketing Intelligence are not in line with our reporting, findings are unreliable and therefore not useful for control and policy."
Marketing Intelligence
"Marketing does not understand that answering a simple question takes a lot of work for us."
"We recently did a great analysis, but nothing is done with it."
Sales
"Marketing Intelligence supports the Marketing Department with its campaigns, but they do not understand what is going on in our field."
"In addition, they do not have access to the relevant data for us."
IT
"We don't have the capacity to handle all the ad-hoc data requests from Marketing Intelligence due to other important priorities.
In addition, there are serious data quality issues, which makes us wonder whether it makes sense to work with this."

FIGURE 12.12 Views on analytical (Marketing Intelligence) function

A problem in cooperation can also arise if data scientists and, for example, marketing managers have different perceptions (Verhoef & Pennings, 2012) because of the different tasks they do and their different personalities. In general, analysts like to focus on the analysis itself and are less interested in its implications. Moreover, analysts will typically focus more on details of the methods applied. Marketing managers may move more easily to the next marketing action.

Inspired by the work of psychologist Carl Jung, we have profiled both marketing analysts and members of marketing departments on specific personality traits (see Figure 12.13). It is interesting to observe that there are indeed strong differences. Analysts are interested in in-depth discussions about details and strongly value personal relationships in their working environment. In contrast, senior marketing department members and managers are more entrepreneurial and focus more on outcomes and making decisions, taking a stronger leadership role. So each function must understand the different thought worlds and personal orientations of the other; this may overcome communication problems and lead to more effective co-operation.

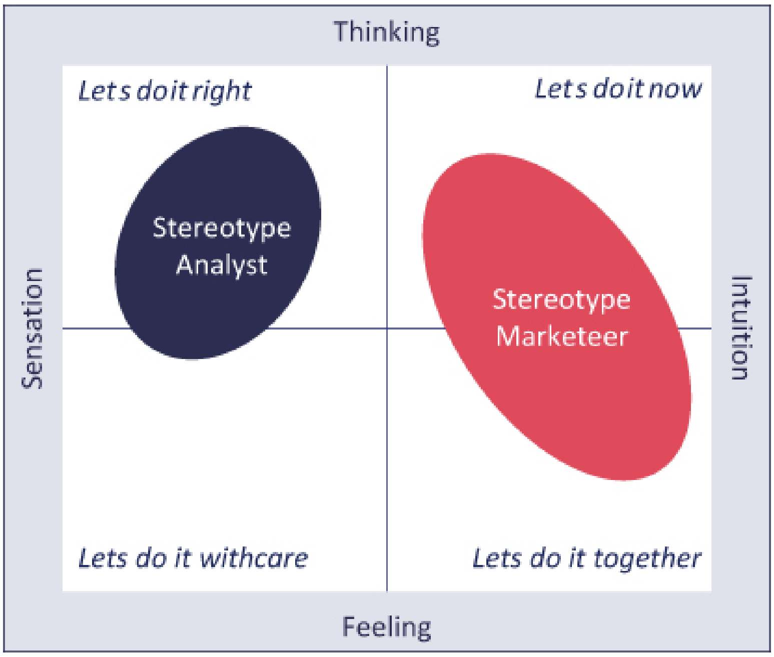


FIGURE 12.13 Different personality profiles of analysts and marketeers

12.6.3 Establishing a data-driven culture

A more overarching enabler of a successful implementation of big data analytics and data science within firms is that firms should transform in such a way that they rely more on data in their decision-making. You don't just organize this by making agreements, setting up processes for it, and choosing the right organization chart. This requires a data-driven culture, which in turn implies that all company employees should base their thinking and behavior more on established, data-driven facts, rather than relying on intuition or gut feeling. The latter is a frequent occurrence in marketing: marketing is frequently blamed for having nice creative ideas without understanding their implications for business and performance (Verhoef & Leeflang, 2011). As McGovern *et al*. (2004:74) state: "The marketing field is chockablock with creative thinkers, yet it's short on people who lean toward an analytic, left-brain approach to the discipline."

A more data-driven culture within marketing is also frequently referred to as "fact-based marketing," clearly suggesting the difference from primarily intuition-based marketing. A stronger focus on marketing accountability within firms will typically induce a stronger data-driven culture. Marketing accountability is generally considered as the extent to which marketing departments can show the effects of marketing actions on marketing and business performance metrics in their plans (ex-ante) or evaluations (ex-post).

For specific tactical tasks (e.g., mail selection, assortment optimization) there is sufficient evidence to demonstrate that the use of data and analytics improves performance, for example because scarce resources are now more efficiently distributed over different customer segments. The use of data and analytics is also increasingly seen as indispensable for more strategic long-term decisions. Analytics provides insights into market development, brand developments, new non-served segments, innovation opportunities, etc., which are essential as input in developing more long-term focused strategic marketing plans. Model outcomes and insights are then one of the inputs, but other considerations must also be taken into account, including management intuition. The research of Blattberg & Hoch (1990) shows that a combination of both model input and intuition leads to the best decisions.

Despite this evidence of the benefits of marketing accountability and the use of data and analytics in tactical and strategic decisions, establishing a data-driven culture is not easy because that affects more than just the marketing department. Developing a data-driven culture involves the entire organization and starts with top management.

Support from top management is essential in creating a strong analytical function that has something meaningful to say within the organization; it has often been shown to be a driver of the adoption and use of marketing decision support systems and data-based marketing (e.g., Verhoef & Hoekstra, 1999; Wierenga & Oude Ophuis, 1997). When culture changes are needed, support from top management is a prerequisite (e.g., Kirca, Jayachandran, & Bearden, 2005). There is a need to regulate investments (i.e., talent acquisition, talent retention, training, and systems) in the analytic functions and to establish a stronger data-driven culture. While this is well-known, support for a strong role for analytics cannot be taken for granted. However, there are several recommendations that can help to achieve this:

- Provide top-down support
  - The culture is largely shaped or stimulated from the boardroom. Management boards therefore also have an exemplary function in making decisions based on data and must also propagate this.
  - The increase in data offers opportunities to create innovation and growth by developing information-based solutions and products. Make this measurable and make sure top management is aware of it so that it is also more likely that they will provide support.

- Invest in human capital
    - – Investing in technology is a waste of money if people don't know how to use it. So make developing company-wide data & analytics skills a priority.
    - – Convince people that they must have these skills to contribute to the growth of the organization and their personal development.
    - – Provide a "learning budget," in the form of both money and time.
- Create confidence
    - – Create a corporate culture that encourages curiosity and challenges assumptions.
    - – Give employees the confidence and space for experimentation and failure.
    - – Ensure the barriers between experts and decision-makers are low, so that good ideas and solutions enter the boardroom more easily.
- Collaboration & knowledge sharing
    - – Encourage and facilitate collaboration in multi-disciplinary and cross-functional teams to share knowledge and learn from each other.
    - – Provide joint responsibility so that people are more willing to help each other and generate the enthusiasm to achieve a data-driven result together.
    - – Share best practices and make added value measurable.
    - – Ensure democratization of data: easy access to data leads to adoption.

12.7 CONCLUSIONS

In this chapter, we discussed the building of successful data analytical capabilities within firms. This is an important topic that is frequently neglected when discussing big data and AI opportunities. However, to create value with big data and AI, these capabilities are of essential importance. First, we considered the internal process for executing impactful analytics within firms. These analytics must start with a business question or a business need. Analytics for the sake of analytics does not create value. Subsequently, we discussed the important challenges of acquiring and keeping analytical talent and the development of an analytical function within firms. We provided an in-depth discussion of the different profiles looked for. We considered the systems required for applying successful analytics. Importantly, their design should be user led not technology led. We said that data analytical systems will consist of different layers. Finally, we elaborated on many organizational issues concerning organizational structure, cooperation between MI and analytical departments and other functions, the presence of a data-driven culture, and the requirement for top management support. Probably most important here is that firms require a strong data-driven culture which should be supported by top management and should be sufficiently embedded within the organizational structure to ensure successful cooperation between the analytical function and other (marketing) functions within the firm.

ASSIGNMENTS

Assignment 12.1: The multidisciplinary skills of the modern data analyst

Introduction

People are one of the most important building blocks for successful analytical competence. In other words, data analysts must be able to access, process, combine, analyze, and translate data into impactful recommendations and initiatives. Considering that the demand for these types of employees is many times greater than the supply, it is a major challenge for organizations to recruit and retain the right profiles.

Therefore, it is recommended that the skills required of a modern data analyst first be defined.

Subsequently, it is necessary to determine which competencies are already present in existing teams. From this point, it is possible to develop missing skills within teams and/or look for new analysts who can strengthen the team.

In Figure 12.6 the required skills of a data analyst are categorized in four sub-areas:

- Analytical skills
- Knowledge of data and tools
- Business sense or insight into how the business works
- Communication and visualization skills.

As an individual, it is almost impossible to score highly in all areas. In practice, we therefore see different typical profiles emerge. Examples of these profiles can be found in Figure 12.7.

Assignment

For this assignment, you have to form a team with four fellow students. Together you represent an analyst team from a medium-sized commercial organization.

Questions

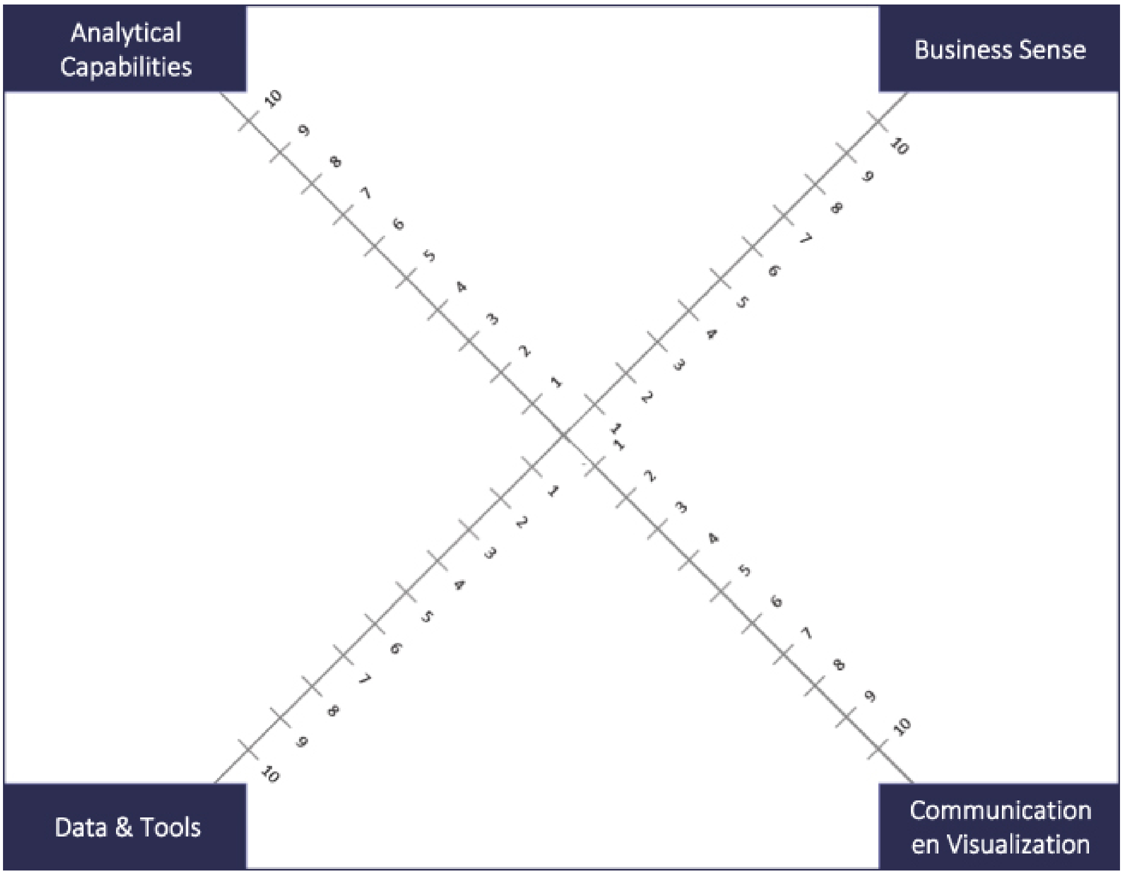


FIGURE 12.14 Fill-in graph to score yourself and your team members on a ten-point scale of the different sub-areas of a modern analyst

1. Score yourself on a ten-point scale in the different sub-areas in Figure 12.14 and explain your reasoning for each score per sub-area using concrete examples. Which "type" of analyst are you?
2. Now score the other team members in the different areas. If you are unsure, make an estimation.
3. Plot everyone's scores (including your own score) in the chart below and determine how balanced your team is.
4. What skills are missing from your team? What would you like to develop as a team?
5. Share your results and insights with the rest of the team. Write up your conclusions and development plan together.

Assignment 12.2: Data analytics function NL Insurance

Background

NL Insurance is active in the insurance market with various brands. The insurer tries to appeal to different target groups with these different brands. The main brands are:

- Vdirect (Dordrecht): this label is a well-known brand for consumers due to effective advertising. It serves both consumers and businesses but does so without intermediaries.
- NBVZ (Eindhoven): this label traditionally focuses mainly on consumers in the south of the Netherlands, but now also has national appeal. NBVZ also has no intermediaries.
- Vexperts (Groningen): Vexperts does have intermediaries and their products focus on consumers, as well as companies.
- Digipolis (Amsterdam): relatively new member of NL Insurance. Digipolis was set up separately with a focus on the digital channel. It is strongly based on a cooperative ideology. They do business directly with consumers and do not work with intermediaries.

NL Insurance's main office is situated in Hilversum in the center of the country. The different brands still have their offices in various locations in the country. The insurers often still work with their own systems, in which investments were mainly made in setting up Customer Relationship Management in the late 1990s and early 2000s. This is specifically the case for the brands which directly serve their customers. Think of systems, among others, from Oracle and Siebel. These systems worked well in executing a Customer Relationship Management strategy. With their digital focus, Digipolis has much more modern systems which are particularly relevant in a digital age and make more use of, for example, Cloud solutions and new analysis tools, such as R. The analysis function is still mainly organized in a decentralized manner, which means there are differences between the brands. Digipolis quickly realized that they should invest in this and hired a number of econometricians. These analysts have a strong numerical understanding and provide accurate predictions for, for example, the attribution of online channels and customer loyalty. Vdirect has always had a strong analysis function but is mainly concerned with CRM analysis, such as selecting good customers and calculating customer value. Big data is an unexplored area for them. In recent years, NBVZ has mainly cut back on the analysis function due to a strong emphasis on cost reduction, while Vexperts, with their strong focus on intermediaries, found the added value of the analysis function to be relatively low. Vexperts mainly employs analysts with a "you ask, we answer" mentality. The directors of the different brands therefore think differently about the importance of good marketing analysis.

Questions

1. To what extent is it important for a big insurance company to invest in data analytics capabilities?
2. How would you set up a business case to convince the management of a label that investment in data analytics capabilities is worth it? Discuss

which value components you want to have an impact on and how that will generate value for the company? Also, indicate which investments are specifically required.

3. Based on the information provided in the case, how would you assess the data analytics skills of NL Insurance and the underlying brands? Discuss the different skills.
4. The top management of NL Insurance is considering centralizing the data analytics analysis department in Hilversum. This department would then carry out the analyses group-wide and advise the marketing departments. They now ask for a recommendation from external advisers on whether this is a favorable organizational structure. What advice would you give to top management? Justify your choice.

NOTES

1. Source: IDC InfoBrief, Sponsored by Qlik, ‘Transformative Data Through Leadership Survey’, June 16, 2020.
2. Published by Statista Research Department, January 11, 2021.
3. Published by Accenture, ‘Closing data value gap’, August 29, 2019.
4. Published by Qlik.com, ‘How to drive data literacy within the enterprise’, 2018.
5. Published by SearchBusinessAnalytics, ‘Demand for data scientists is booming and will only increase’, January 21, 2019.
6. Published by Gartner, ‘Business Intelligence and Analytics Leaders Must Focus on Mindsets and Culture to Kick Start Advanced Analytics’, September 15, 2015.
7. Published by Gartner, Andrew White, “Our Top Data and Analytics Predictions for 2019,” January 3, 2019.
8. Ken Schwaber and Jeff Sutherland (2020). The Scrum Guide, This publication is offered for license under the Attribution Share-Alike license of Creative Commons, accessible at https://creativecommons.org/licenses/by-sa/4.0/legalcode.
9. Adapted from Forbes, Ganes Kesari, ‘The 5 roles that every data science team must hire’, November 24, 2020.
10. See https://chiefmartec.com/2020/04/marketing-technology-landscape-2020-martech- 5000/ (accessed April 4, 2022).
11. Published by Chiefmartec.com, ‘Marketing Technology Landscape Supergraphic (2020)’ April 22, 2020.

REFERENCES

Blattberg, R. C. & Hoch, S. J. (1990). Database models and managerial intuition: 50% model + 50% manager. *Management Science*, *36*(8), 887–899.

Davenport, T. & Harris, J. (2007). *Competing on Analytics – The New Science of Winning*. Boston, MA: Harvard Business School Press.

De Vos, A. & Meganck, A. (2009). What HR managers do versus what employees value: Exploring both parties' views on retention management from a psychological contract perspective. *Personnel Review*, *38*(1), 45–60.

Hagen, C., Khan, K., Ciobo, M., Miller, J., Wall, D., Evans, H., & Yaday, Y. (2013). Big data and the creative destruction of today's business models. *Holland Management Review*, *148*(4), 25–37.

Holtom, B. C., Mitchell, T. R., Lee, T. W., & Eberly, M. B. (2008). Turnover and retention research: A glance at the past, a closer review of the present, and a venture into the future. *The Academy of Management Annals*, *2*(1), 231–274.

Humby, C., Hunt, T., & Phillips, T. (2008). *Scoring Points: How Tesco Is Winning Customer Loyalty*. Philadelphia: Kogan Page Publishers.

Kirca, A. H., Jayachandran, S., & Bearden, W. O. (2005). Market orientation: A meta-analytic review and assessment of its antecedents and impact on performance. *Journal of Marketing*, *6*(2), 24–41.

McGovern, G. J., Court, D., Quelch, J. A., & Crawford, B. (2004). Bringing customers into the boardroom. *Harvard Business Review*, *82*(11), 70–80.

Pauwels, K. H., Ambler, T., Clark, B., LaPointe, P., Reibstein, D., Skiera, B., Wieranga, B., & Wiesel, T. (2009). Dashboards & marketing: Why, what, how and what research is needed? *Journal of Service Research*, *12*(2), 175–189.

Reibstein, D. J., Norton, D., Joshi, Y., & Farris, P. (2005). Marketing dashboard: A decision support system assessing marketing productivity. Marketing Science Conference. Atlanta.

Rigby, D. K. (2014). Digital-physical mashups. *Harvard Business Review*, *92*(9), 84–92.

Rust, R. T., Lemon, K. N., & Zeithaml, V. A. (2004). Return on marketing: Using customer equity to focus marketing strategy. *Journal of Marketing*, *68*(1), 109–127.

Sutherland, J. (2015). *Scrum: The Art of Doing Twice the Work in Half the Time*. New Orleans: Cornerstone.

Tavassoli, N. T., Sorescu, A., & Chandy, R. (2014). Employee-based brand equity: Why firms with strong brands pay their executives less. *Journal of Marketing Research*, *51*(6), 676–690.

Verhoef, P. C., Antonides, G., & De Hoog, A. N. (2004). Service encounters as a sequence of events: the importance of peak experiences.

Journal of Service Research, *7*(1), 53–64.

Verhoef, P. C. & Hoekstra, J. C. (1999). Status of database marketing in the Dutch fast moving consumer goods industry. *Journal of Market-Focused Management*, *3*(4), 313–331.

Verhoef, P. C., Hoekstra, J. C., & Van der Scheer, H., (2009). *Competing on Analytics: Status Quo Van Customer Intelligence in Nederland*. Groningen: Customer Insights Center.

Verhoef, P. C. & Leeflang, P. S. H. (2011). Accountability as a main ingredient of getting marketing back in the board room. *Marketing Review St. Gallen*, *28*(3), 26–31.

Verhoef, P. C. & Lemon, K. N. (2013). Successful customer value management: Key lessons and emerging trends. *European Management Journal*, *13*(1), 1–15.

Verhoef, P. C. & Pennings, J. M. (2012). The marketing finance Interface: An organizational perspective. In: S. Ganesan & S. Bharadwaj, (eds) *Handbook of Marketing and Finance* (pp. 225–243). Cheltenham: Edward Elgar Publishing.

Wierenga, B. & Oude Ophuis, P. A. M. (1997). Marketing decision support systems: Adoption, use and satisfaction. *International Journal of Research in Marketing*, *14*(3), 275–290.

Index

Note: **Bold** page numbers refer to tables and italic page numbers refer to figures.