

CHAPTER 1

A Value-Centric Perspective Towards Analytics

INTRODUCTION

In this first chapter, we set the scene for what is ahead by broadly introducing profit-driven business analytics. The value-centric perspective toward analytics proposed in this book will be positioned and contrasted with a traditional statistical perspective. The implications of adopting a value-centric perspective toward the use of analytics in business are significant: a mind shift is needed both from managers and data scientists in developing, implementing, and operating analytical models. This, however, calls for deep insight into the underlying principles of advanced analytical approaches. Providing such insight is our general objective in writing this book and, more specifically:

- We aim to provide the reader with a structured overview of state-of-the art analytics for business applications.
- We want to assist the reader in gaining a deeper practical understanding of the inner workings and underlying principles of these approaches from a practitioner's perspective.

- We wish to advance managerial thinking on the use of advanced analytics by offering insight into how these approaches may either generate significant added value or lower operational costs by increasing the efficiency of business processes.
- We seek to prosper and facilitate the use of analytical approaches that are customized to needs and requirements in a business context.

As such, we envision that our book will facilitate organizations stepping up to a next level in the adoption of analytics for decision making by embracing the advanced methods introduced in the subsequent chapters of this book. Doing so requires an investment in terms of acquiring and developing knowledge and skills but, as is demonstrated throughout the book, also generates increased profits. An interesting feature of the approaches discussed in this book is that they have often been developed at the intersection of academia and business, by academics and practitioners joining forces for tuning a multitude of approaches to the particular needs and problem characteristics encountered and shared across diverse business settings.

Most of these approaches emerged only after the millennium, which should not be surprising. Since the millennium, we have witnessed a continuous and pace-gaining development and an expanding adoption of information, network, and database technologies. Key technological evolutions include the massive growth and success of the World Wide Web and Internet services, the introduction of smart phones, the standardization of enterprise resource planning systems, and many other applications of information technology. This dramatic change of scene has prospered the development of analytics for business applications as a rapidly growing and thriving branch of science and industry.

To achieve the stated objectives, we have chosen to adopt a pragmatic approach in explaining techniques and concepts. We do not focus on providing extensive mathematical proof or detailed algorithms. Instead, we pinpoint the crucial insights and underlying reasoning, as well as the advantages and disadvantages, related to the practical use of the discussed approaches in a business setting. For this, we ground our discourse on solid academic research expertise as well as on many years of practical experience in elaborating industrial analytics projects in close collaboration with data science professionals. Throughout the book, a plethora of illustrative examples and case studies are discussed. Example datasets, code, and implementations

are provided on the book's companion website, www.profit-analytics.com, to further support the adoption of the discussed approaches.

In this chapter, we first introduce business analytics. Next, the profit-driven perspective toward business analytics that will be elaborated in this book is presented. We then introduce the subsequent chapters of this book and how the approaches introduced in these chapters allow us to adopt a value-centric approach for maximizing profitability and, as such, to increase the return on investment of big data and analytics. Next, the analytics process model is discussed, detailing the subsequent steps in elaborating an analytics project within an organization. Finally, the chapter concludes by characterizing the ideal profile of a business data scientist.

Business Analytics

Data is the new oil is a popular quote pinpointing the increasing value of data and—to our liking—accurately characterizes data as raw material. Data are to be seen as an input or basic resource needing further processing before actually being of use. In a subsequent section in this chapter, we introduce the analytics process model that describes the iterative chain of processing steps involved in turning *data* into *information* or *decisions*, which is quite similar actually to an oil refinery process. Note the subtle but significant difference between the words *data* and *information* in the sentence above. Whereas data fundamentally can be defined to be a sequence of zeroes and ones, information essentially is the same but implies in addition a certain utility or value to the *end user* or *recipient*. So, whether data are information depends on whether the data have utility to the recipient. Typically, for raw data to be information, the data first need to be processed, aggregated, summarized, and compared. In summary, data typically need to be analyzed, and insight, understanding, or knowledge should be added for data to become useful.

Applying basic operations on a dataset may already provide useful insight and support the end user or recipient in decision making. These basic operations mainly involve selection and aggregation. Both selection and aggregation may be performed in many ways, leading to a plentitude of indicators or statistics that can be distilled from raw data. The following illustration elaborates a number of sales indicators in a retail setting.

Providing insight by customized reporting is exactly what the field of **business intelligence (BI)** is about. Typically, visualizations are also adopted to represent indicators and their evolution in time, in easy-to-interpret ways. Visualizations provide support by facilitating

EXAMPLE

For managerial purposes, a retailer requires the development of real-time sales reports. Such a report may include a wide variety of indicators that summarize raw sales data. Raw sales data, in fact, concern transactional data that can be extracted from the online transaction processing (OLTP) system that is operated by the retailer. Some example indicators and the required selection and aggregation operations for calculating these statistics are:

- *Total amount of revenues generated over the last 24 hours:* Select all transactions over the last 24 hours and sum the paid amounts, with *paid* meaning the price net of promotional offers.
- *Average paid amount in online store over the last seven days:* Select all online transactions over the last seven days and calculate the average paid amount;
- *Fraction of returning customers within one month:* Select all transactions over the last month and select customer IDs that appear more than once; count the number of IDs.

Remark that calculating these indicators involves basic selection operations on characteristics or dimensions of transactions stored in the database, as well as basic aggregation operations such as sum, count, and average, among others.

the user's ability to acquire understanding and insight in the blink of an eye. Personalized dashboards, for instance, are widely adopted in the industry and are very popular with managers to monitor and keep track of business performance. A formal definition of business intelligence is provided by Gartner (<http://www.gartner.com/it-glossary>):

Business intelligence is an umbrella term that includes the applications, infrastructure and tools, and best practices that enable access to and analysis of information to improve and optimize decisions and performance.

Note that this definition explicitly mentions the required infrastructure and best practices as an essential component of BI, which is typically also provided as part of the package or solution offered by BI vendors and consultants. More advanced analysis of data may further support users and optimize decision making. This is exactly where analytics comes into play. *Analytics* is a catch-all term covering a wide variety of what are essentially data-processing techniques.

In its broadest sense, analytics strongly overlaps with data science, statistics, and related fields such as artificial intelligence (AI) and machine learning. Analytics, to us, is a toolbox containing a variety of instruments and methodologies allowing users to analyze data for a diverse range of well-specified purposes. Table 1.1 identifies a number of categories of analytical tools that cover diverse intended uses or, in other words, allow users to complete a diverse range of tasks.

A first main group of tasks identified in Table 1.1 concerns prediction. Based on observed variables, the aim is to accurately estimate or predict an unobserved value. The applicable subtype of predictive analytics depends on the type of target variable, which we intend to model as a function of a set of predictor variables. When the target variable is categorical in nature, meaning the variable can only take a limited number of possible values (e.g., churning or not, fraudster or not, defaulter or not), then we have a classification problem. When the task concerns the estimation of a continuous target variable (e.g., sales amount, customer lifetime value, credit loss), which can take any value over a certain range of possible values, we are dealing with regression. Survival analysis and forecasting explicitly account for the time dimension by either predicting the timing of events (e.g., churn, fraud, default) or the evolution of a target variable in time (e.g., churn rates, fraud rates, default rates). Table 1.2 provides simplified example datasets and analytical models for each type of predictive analytics for illustrative purposes.

The second main group of analytics comprises descriptive analytics that, rather than predicting a target variable, aim at identifying specific types of patterns. Clustering or segmentation aims at grouping **entities** (e.g., customers, transactions, employees, etc.) that are similar in nature. The objective of association analysis is to find groups of **events** that frequently co-occur and therefore appear to be associated. The basic **observations** that are being analyzed in this problem setting consist of variable groups of events; for instance, transactions involving various products that are being bought by a customer at a certain moment in time. The aim of sequence analysis

Table 1.1 Categories of Analytics from a Task-Oriented Perspective

Predictive Analytics	Descriptive Analytics
Classification	Clustering
Regression	Association analysis
Survival analysis	Sequence analysis
Forecasting	

Table 1.2 Example Datasets and Predictive Analytical Models

Example dataset

ID	Recency	Frequency	Monetary	Churn
C1	26	4.2	126	Yes
C2	37	2.1	59	No
C3	2	8.5	256	No
C4	18	6.2	89	No
C5	46	1.1	37	Yes
...	...	...	...	...

Classification

Decision tree classification model:

```
graph TD; A["Frequency < 5"] -- Yes --> B["Recency > 25"]; A -- No --> C["Churn = No"]; B -- Yes --> D["Churn = Yes"]; B -- No --> E["Churn = No"];
```

Regression

ID	Recency	Frequency	Monetary	CLV
C1	26	4.2	126	3,817
C2	37	2.1	59	4,31
C3	2	8.5	256	2,187
C4	18	6.2	89	543
C5	46	1.1	37	1,548
...	...	...	...	...

Linear regression model:

$$\text{CLV} = 260 + 11 \cdot \text{Recency} + 6.1 \cdot \text{Frequency} + 3.4 \cdot \text{Monetary}$$

Survival analysis

ID	Recency	Churn or Censored	Time of churn or Censoring
C1	26	Churn	181
C2	37	Censored	253
C3	2	Censored	37
C4	18	Censored	172
C5	46	Churn	98
...	...	...	...

General parametric survival analysis model:

$$\log(T) = 13 + 5.3 \cdot \text{Recency}$$

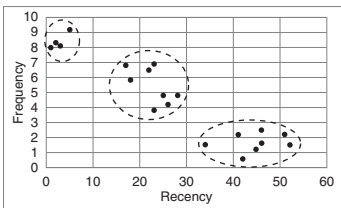
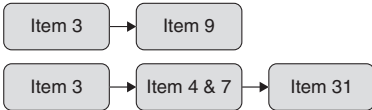
Forecasting

Timestamp	Demand
January	513
February	652
March	435
April	578
May	601
...	...

Weighted moving average forecasting model:

$$\text{Demand}_t = 0.4 \cdot \text{Demand}_{t-1} + 0.3 \cdot \text{Demand}_{t-2} + 0.2 \cdot \text{Demand}_{t-3} + 0.1 \cdot \text{Demand}_{t-4}$$

Table 1.3 Example Datasets and Descriptive Analytical Models

Data			Descriptive analytical model																					
Clustering																								
<table><thead><tr><th>ID</th><th>Recency</th><th>Frequency</th></tr></thead><tbody><tr><td>C1</td><td>26</td><td>4.2</td></tr><tr><td>C2</td><td>37</td><td>2.1</td></tr><tr><td>C3</td><td>2</td><td>8.5</td></tr><tr><td>C4</td><td>18</td><td>6.2</td></tr><tr><td>C5</td><td>46</td><td>1.1</td></tr><tr><td>...</td><td>...</td><td>...</td></tr></tbody></table>			ID	Recency	Frequency	C1	26	4.2	C2	37	2.1	C3	2	8.5	C4	18	6.2	C5	46	1.1	...	...	...	K-means clustering with $K = 3$: 
ID	Recency	Frequency																						
C1	26	4.2																						
C2	37	2.1																						
C3	2	8.5																						
C4	18	6.2																						
C5	46	1.1																						
...	...	...																						
Association analysis																								
<table><thead><tr><th>ID</th><th>Items</th></tr></thead><tbody><tr><td>T1</td><td>beer, pizza, diapers, baby food</td></tr><tr><td>T2</td><td>coke, beer, diapers</td></tr><tr><td>T3</td><td>crisps, diapers, baby food</td></tr><tr><td>T4</td><td>chocolates, diapers, pizza, apples</td></tr><tr><td>T5</td><td>tomatoes, water, oranges, beer</td></tr><tr><td>...</td><td>...</td></tr></tbody></table>			ID	Items	T1	beer, pizza, diapers, baby food	T2	coke, beer, diapers	T3	crisps, diapers, baby food	T4	chocolates, diapers, pizza, apples	T5	tomatoes, water, oranges, beer	...	...	Association rules: If baby food And diapers Then beer If coke And pizza Then crisps ...							
ID	Items																							
T1	beer, pizza, diapers, baby food																							
T2	coke, beer, diapers																							
T3	crisps, diapers, baby food																							
T4	chocolates, diapers, pizza, apples																							
T5	tomatoes, water, oranges, beer																							
...	...																							
Sequence analysis																								
<table><thead><tr><th>ID</th><th>Sequential items</th></tr></thead><tbody><tr><td>C1</td><td><{3},{9}></td></tr><tr><td>C2</td><td><{1 2},{3},{4 6 7}></td></tr><tr><td>C3</td><td><{3 5 7}></td></tr><tr><td>C4</td><td><{3},{4 7},{9}></td></tr><tr><td>C5</td><td><{9}></td></tr><tr><td>...</td><td>...</td></tr></tbody></table>			ID	Sequential items	C1	<{3},{9}>	C2	<{1 2},{3},{4 6 7}>	C3	<{3 5 7}>	C4	<{3},{4 7},{9}>	C5	<{9}>	...	...	Sequence rules:  ...							
ID	Sequential items																							
C1	<{3},{9}>																							
C2	<{1 2},{3},{4 6 7}>																							
C3	<{3 5 7}>																							
C4	<{3},{4 7},{9}>																							
C5	<{9}>																							
...	...																							

is similar to association analysis but concerns the detection of events that frequently occur sequentially, rather than simultaneously as in association analysis. As such, sequence analysis explicitly accounts for the time dimension. Table 1.3 provides simplified examples of datasets and analytical models for each type of descriptive analytics.

Note that Tables 1.1 through 1.3 identify and illustrate categories of approaches that are able to complete a specific task from a *technical* rather than an *applied* perspective. These different types of analytics can be applied in quite diverse business and nonbusiness settings and consequently lead to many specialized applications. For instance, predictive analytics and, more specifically, classification techniques may be applied for detecting fraudulent credit-card transactions, for predicting customer churn, for assessing loan applications, and so forth. From an application perspective, this leads to various groups of analytics such as, respectively, fraud analytics, customer or marketing analytics, and credit risk analytics. A wide range of business applications of analytics across industries and business departments is discussed in detail in Chapter 3.

With respect to Table 1.1, it needs to be noted that these different types of analytics apply to **structured data**. An example of a structured dataset is shown in Table 1.4. The rows in such a dataset are typically called observations, instances, records, or lines, and represent or collect information on *basic entities* such as customers, transactions, accounts, or citizens. The columns are typically referred to as (explanatory or predictor) variables, characteristics, attributes, predictors, inputs, dimensions, effects, or features. The columns contain information on a particular entity as represented by a row in the table. In Table 1.4, the second column represents the age of a customer, the third column the postal code, and so on. In this book we consistently use the terms **observation** and **variable** (and sometimes more specifically, explanatory, predictor, or target variable).

Because of the structure that is present in the dataset in Table 1.4 and the well-defined meaning of rows and columns, it is much easier to analyze such a structured dataset compared to analyzing unstructured data such as text, video, or networks, to name a few. Specialized techniques exist that facilitate analysis of unstructured data—for instance, text analytics with applications such as sentiment analysis, video analytics that can be applied for face recognition and incident detection, and network analytics with applications such as community

Table 1.4 Structured Dataset

Customer	Age	Income	Gender	Duration	Churn
John	30	1,800	Male	620	Yes
Sarah	25	1,400	Female	12	No
Sophie	52	2,600	Female	830	No
David	42	2,200	Male	90	Yes

mining and relational learning (see Chapter 2). Given the rough estimate that over 90% of all data are unstructured, clearly there is a large potential for these types of analytics to be applied in business.

However, due to the inherent complexity of analyzing unstructured data, as well as because of the often-significant development costs that only appear to pay off in settings where adopting these techniques significantly adds to the easier-to-apply **structured analytics**, currently we see relatively few applications in business being developed and implemented. In this book, we therefore focus on analytics for analyzing structured data, and more specifically the subset listed in Table 1.1. For unstructured analytics, one may refer to the specialized literature (Elder IV and Thomas 2012; Chakraborty, Murali, and Satish 2013; Coussement 2014; Verbeke, Martens and Baesens 2014; Baesens, Van Vlasselaer, and Verbeke 2015).

PROFIT-DRIVEN BUSINESS ANALYTICS

The premise of this book is that analytics is to be adopted in business for *better decision making*—“better” meaning *optimal* in terms of maximizing the net profits, returns, payoff, or value resulting from the decisions that are made based on insights obtained from data by applying analytics. The incurred returns may stem from a gain in efficiency, lower costs or losses, and additional sales, among others. The decision level at which analytics is typically adopted is the operational level, where many customized decisions are to be made that are similar and granular in nature. High-level, ad hoc decision making at strategic and tactical levels in organizations also may benefit from analytics, but expectedly to a much lesser extent.

The decisions involved in developing a business strategy are highly complex in nature and do not match the elementary tasks enlisted in Table 1.1. A higher-level AI would be required for such purpose, which is not yet at our disposal. At the operational level, however, there are many *simple* decisions to be made, which exactly match with the tasks listed in Table 1.1. This is not surprising, since these approaches have often been developed with a specific application in mind. In Table 1.5, we provide a selection of example applications, most of which will be elaborated on in detail in Chapter 3.

Analytics facilitates optimization of the fine granular decision-making activities listed in Table 1.5, leading to lower costs or losses and higher revenues and profits. The level of optimization depends on the accuracy and validity of the predictions, estimates, or patterns derived from the data. Additionally, as we stress in this book, the quality

Table 1.5 Examples of Business Decisions Matching Analytics

Decision Making with Predictive Analytics	
Classification	Credit officers have to screen loan applications and decide on whether to accept or reject an application based on the involved risk. Based on historical data on the performance of past loan applications, a classification model may learn to distinguish <i>good</i> from <i>bad</i> loan applications using a number of well-chosen characteristics of the application as well as of the applicant. Analytics and, more specifically, classification techniques allow us to optimize the loan-granting process by more accurately assessing risk and reducing bad loan losses (Van Gestel and Baesens 2009; Verbraken et al. 2014). Similar applications of decision making based on classification techniques, which are discussed in more detail in Chapter 3 of this book, include customer churn prediction, response modeling, and fraud detection.
Regression	Regression models allow us to estimate a continuous target value and in practice are being adopted, for instance, to estimate customer lifetime value. Having an indication on the future worth in terms of revenues or profits a customer will generate is important to allow customization of marketing efforts, for pricing, etc. As is discussed in detail in Chapter 3, analyzing historical customer data allows estimating the future net value of current customers using a regression model. Similar applications involve loss given default modeling as is discussed in Chapter 3, as well as the estimation of software development costs (Dejaeger et al. 2012).
Survival analysis	Survival analysis is being adopted in predictive maintenance applications for estimating when a machine component will fail. Such knowledge allows us to optimize decisions related to machine maintenance—for instance, to optimally plan when to replace a vital component. This decision requires striking a balance between the cost of machine failure during operations and the cost of the component, which is preferred to be operated as long as possible before replacing it (Widodo and Yang 2011). Alternative business applications of survival analysis involve the prediction of time to churn and time to default where, compared to classification, the focus is on predicting <i>when</i> the event will occur rather than <i>whether</i> the event will occur.
Forecasting	A typical application of forecasting involves demand forecasting, which allows us to optimize production planning and supply chain management decisions. For instance, a power supplier needs to be able to balance electricity production and demand by the consumers and for this purpose adopts forecasting or time-series modeling techniques. These approaches allow an accurate prediction of the short-term evolution of demand based on historical demand patterns (Hyndman et al. 2008).

Table 1.5 (Continued)

Decision Making with Descriptive Analytics	
Clustering	Clustering is applied in credit-card fraud detection to block suspicious transactions in real time or to select suspicious transactions for investigation in near-real time. Clustering facilitates automated decision making by comparing a new transaction to clusters or groups of historical nonfraudulent transactions and by labeling it as suspicious when it differs too much from these groups (Baesens et al. 2015). Clustering can also be used for identifying groups of similar customers, which facilitates the customization of marketing campaigns.
Association analysis Sequence analysis	Association analysis is often applied for detecting patterns within transactional data in terms of products that are often purchased together. Sequence analysis, on the other hand, allows the detection of which products are often bought subsequently. Knowledge of such associations allows smarter decisions to be made about which products to advertise, to bundle, to place together in a store, etc. (Agrawal and Srikant 1994).

of data-driven decision making depends on the extent to which the actual use of the predictions, estimates, or patterns is accounted for in developing and applying analytical approaches. We argue that the actual goal, which in a business setting is to generate profits, should be central when applying analytics in order to further increase the return on analytics. For this, we need to adopt what we call *profit-driven analytics*. These are adapted techniques specifically configured for use in a business context.

EXAMPLE

The following example highlights the tangible difference between a statistical approach to analytics and a profit-driven approach. Table 1.5 already indicated the use of analytics and, more specifically, classification techniques for predicting which customers are about to churn. Having such knowledge allows us to decide which customers are to be targeted in a retention campaign, thereby increasing the efficiency and returns of that campaign when compared to randomly or intuitively selecting customers. By offering a financial incentive to customers that are likely to churn—for instance, a temporary reduction of the monthly fee—they may be retained. Actively retaining customers has been shown by various studies to be much cheaper than acquiring new customers to replace those who defect (Athanasopoulos 2000; Bhattacharya 1998).

It needs to be noted, however, that not every customer generates the same amount of revenues and therefore represents the same value to a company. Hence, it is much more important to detect churn for the most valuable customers. In a basic customer churn prediction setup, which adopts what we call a statistical perspective, no differentiation is made between high-value and low-value customers when learning a classification model to detect future churn. However, when analyzing data and learning a classification model, it should be taken into account that missing a high-value churning customer is much costlier than missing a low-value churning customer. The aim of this would be to steer or tune the resulting predictive model so it accounts for value, and consequently for its actual end-use in a business context.

An additional difference between the statistical and business perspectives toward adopting classification and regression modeling concerns the difference between, respectively, *explaining* and *predicting* (Breiman 2001; Shmueli and Koppius 2011). The aim of estimating a model may be either of these two goals:

1. To establish the relation or detect dependencies between characteristics or independent variables and an observed dependent target variable(s) or outcome value.
2. To *estimate* or *predict* the unobserved or future value of the target variable as a function of the independent variables.

For instance, in a medical setting, the purpose of analyzing data may be to establish the impact of smoking behavior on the life expectancy of an individual. A regression model may be estimated that *explains* the observed age at death of a number of subjects in terms of characteristics such as gender and number of years that the subject smoked. Such a model will establish or quantify the impact or relation between each characteristic and the observed outcome, and allows for testing the statistical significance of the impact and measuring the uncertainty of the result (Cao 2016; Peto, Whitlock, and Jha 2010).

A clear distinction exists with estimating a regression model for, as an example, software effort prediction, as introduced in Table 1.5. In such applications where the aim is mainly to predict, essentially we are not interested in what drivers *explain* how much effort it will take to develop new software, although this may be a useful side result. Instead we mainly wish to predict as accurately as possible the

effort that will be required for completing a project. Since the model’s main use will be to produce an estimate allowing cost projection and planning, it is the exactness or accuracy of the prediction and the size of the errors that matters, rather than the exact relation between the effort and characteristics of the project.

Typically, in a business setting, the aim is to predict in order to facilitate improved or automated decision making. Explaining, as indicated for the case of software effort prediction, may have use as well since useful insights may be derived. For instance, from the predictive model, it may be found what the exact impact is of including more or less senior and junior programmers in a project team on the required effort to complete the project, allowing the team composition to be optimized as a function of project characteristics.

In this book, several versatile and powerful profit-driven approaches are discussed. These approaches facilitate the adoption of a value-centric business perspective toward analytics in order to boost the returns. Table 1.6 provides an overview of the structure of the book. First, we lay the foundation by providing a general introduction to analytics in Chapter 2, and by discussing the most important and popular business applications in detail in Chapter 3.

Chapter 4 discusses approaches toward uplift modeling, which in essence is about distilling or estimating the net effect of a decision and then contrasting the expected result for alternative scenarios. This allows, for instance, the optimization of marketing efforts by customizing the contact channel and the format of the incentive for the response to the campaign to be maximal in terms of returns being generated. Standard analytical approaches may be adopted to develop uplift models. However, specialized approaches tuned toward the particular problem characteristics of uplift modeling have also been developed, and they are discussed in Chapter 4.

Table 1.6 Outline of the Book

Book Structure
Chapter 1: A Value-Centric Perspective Towards Analytics
Chapter 2: Analytical Techniques
Chapter 3: Business Applications
Chapter 4: Uplift Modeling
Chapter 5: Profit-Driven Analytical Techniques
Chapter 6: Profit-Driven Model Evaluation and Implementation
Chapter 7: Economic Impact

As such, Chapter 4 forms a bridge to Chapter 5 of the book, which concentrates on various advanced analytical approaches that can be adopted for developing profit-driven models by allowing us to account for profit when learning or applying a predictive or descriptive model. Profit-driven predictive analytics for classification and regression are discussed in the first part of Chapter 5, whereas the second part focuses on descriptive analytics and introduces profit-oriented segmentation and association analysis.

Chapter 6 subsequently focuses on approaches that are tuned toward a business-oriented evaluation of predictive models—for example, in terms of profits. Note that traditional statistical measures, when applied to customer churn prediction models, for instance, do not differentiate among incorrectly predicted or classified customers, whereas it definitely makes sense from a business point of view to account for the value of the customers when evaluating a model. For instance, incorrectly predicting a customer who is about to churn with a high value represents a higher loss or cost than not detecting a customer with a low value who is about to churn. Both, however, are accounted for equally by nonbusiness and, more specifically, non-profit-oriented evaluation measures. Both Chapters 4 and 6 allow using *standard* analytical approaches as discussed in Chapter 2, with the aim to maximize profitability by adopting, respectively, a profit-centric setup or profit-driven evaluation. The particular business application of the model will appear to be an important factor to account for in maximizing profitability.

Finally, Chapter 7 concludes the book by adopting a broader perspective toward the use of analytics in an organization by looking into the economic impact, as well as by zooming into some practical concerns related to the development, implementation, and operation of analytics within an organization.

ANALYTICS PROCESS MODEL

Figure 1.1 provides a high-level overview of the analytics process model (Hand, Mannila, and Smyth 2001; Tan, Steinbach, and Kumar 2005; Han and Kamber 2011; Baesens 2014). This model defines the subsequent steps in the development, implementation, and operation of analytics within an organization.

As a first step, a thorough definition of the business problem to be addressed is needed. The objective of applying analytics needs to be unambiguously defined. Some examples are: customer segmentation

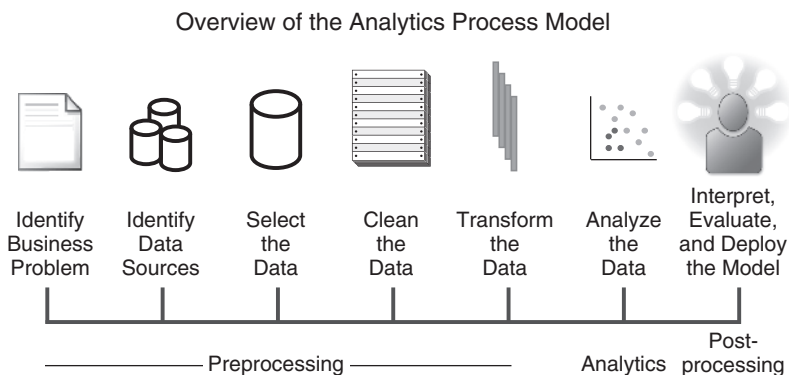


Figure 1.1 The analytics process model.
(Baesens 2014)

of a mortgage portfolio, retention modeling for a postpaid Telco subscription, or fraud detection for credit-cards. Defining the perimeter of the analytical modeling exercise requires a close collaboration between the data scientists and business experts. Both parties need to agree on a set of key concepts; these may include how we define a customer, transaction, churn, or fraud. Whereas this may seem self-evident, it appears to be a crucial success factor to make sure a common understanding of the goal and some key concepts is agreed on by all involved stakeholders.

Next, all source data that could be of potential interest need to be identified. This is a very important step as data are the key ingredient to any analytical exercise and the selection of data will have a deterministic impact on the analytical models that will be built in a subsequent step. The golden rule here is: the more data, the better! The analytical model itself will later decide which data are relevant and which are not for the task at hand. All data will then be gathered and consolidated in a staging area which could be, for example, a data warehouse, data mart, or even a simple spreadsheet file. Some basic exploratory data analysis can then be considered using for instance OLAP facilities for multidimensional analysis (e.g., roll-up, drill down, slicing and dicing). This will be followed by a data-cleaning step to get rid of all inconsistencies such as missing values, outliers and duplicate data. Additional transformations may also be considered such as binning, alphanumeric to numeric coding, geographical aggregation, to name a few, as well as deriving additional characteristics that are typically called features

from the raw data. A simple example concerns the derivation of the age from the birth date; yet more complex examples are provided in Chapter 3.

In the analytics step, an analytical model will be estimated on the preprocessed and transformed data. Depending on the business objective and the exact task at hand, a particular analytical technique will be selected and implemented by the data scientist. In Table 1.1, an overview was provided of various tasks and types of analytics. Alternatively, one may consider the various types of analytics listed in Table 1.1 to be the basic building blocks or solution components that a data scientist employs to solve the problem at hand. In other words, the business problem needs to be reformulated in terms of the available tools enumerated in Table 1.1.

Finally, once the results are obtained, they will be interpreted and evaluated by the business experts. Results may be clusters, rules, patterns, or relations, among others, all of which will be called analytical models resulting from applying analytics. Trivial patterns (e.g., an association rule is found stating that spaghetti and spaghetti sauce are often purchased together) that may be detected by the analytical model are interesting as they help to validate the model. But of course, the key issue is to find the unknown yet interesting and actionable patterns (sometimes also referred to as knowledge diamonds) that can provide new insights into your data that can then be translated into new profit opportunities. Before putting the resulting model or patterns into operation, an important evaluation step is to consider the actual returns or profits that will be generated, and to compare these to a relevant base scenario such as a do-nothing decision or a change-nothing decision. In the next section, an overview of various evaluation criteria is provided; these are discussed to validate analytical models.

Once the analytical model has been appropriately validated and approved, it can be put into production as an analytics application (e.g., decision support system, scoring engine). Important considerations here are how to represent the model output in a user-friendly way, how to integrate it with other applications (e.g., marketing campaign management tools, risk engines), and how to make sure the analytical model can be appropriately monitored and backtested on an ongoing basis.

It is important to note that the process model outlined in Figure 1.1 is iterative in nature in the sense that one may have to return to previous steps during the exercise. For instance, during the analytics

step, a need for additional data may be identified that will necessitate additional data selection, cleaning, and transformation. The most time-consuming step typically is the data selection and preprocessing step, which usually takes around 80% of the total efforts needed to build an analytical model.

ANALYTICAL MODEL EVALUATION

Before adopting an analytical model and making operational decisions based on the obtained clusters, rules, patterns, relations, or predictions, the model needs to be thoroughly evaluated. Depending on the exact type of output, the setting or business environment, and the particular usage characteristics, different aspects may need to be assessed during evaluation in order to ensure the model is *acceptable* for implementation.

A number of key characteristics of *successful* analytical models are defined and explained in Table 1.7. These broadly defined evaluation criteria may or may not apply, depending on the exact application setting, and will have to be further specified in practice.

Various challenges may occur when developing and implementing analytical models, possibly leading to difficulties in meeting the objectives as expressed by the key characteristics of successful analytical models discussed in Table 1.7. One such challenge may concern the dynamic nature of the relations or patterns retrieved from the data, impacting the usability and lifetime of the model. For instance, in a fraud detection setting, it is observed that fraudsters constantly try to out-beat detection and prevention systems by developing new strategies and methods (Baesens et al. 2015). Therefore, adaptive analytical models and detection and prevention systems are required in order to detect and resolve fraud as soon as possible. Closely monitoring the performance of the model in such a setting is an absolute must.

Another common challenge in a binary classification setting such as predicting customer churn concerns the imbalanced class distribution, meaning that one class or type of entity is much more prevalent than the other. When developing a customer churn prediction model typically many more nonchurners are present in the historical dataset than there are churners. Furthermore, the costs and benefits related to detecting or missing either class are often strongly imbalanced and may need to be accounted for to optimize decision making in the particular business context. In this book, various approaches are

Table 1.7 Key Characteristics of Successful Business Analytics Models

Accuracy	Refers to the predictive power or the correctness of the analytical model. Several statistical evaluation criteria exist and may be applied to assess this aspect, such as the hit rate, lift curve, or AUC. A number of profit-driven evaluation measures will be discussed in detail in Chapter 6. Accuracy may also refer to statistical significance, meaning that the patterns that have been found in the data have to be real, robust, and not the consequence of coincidence. In other words, we need to make sure that the model <i>generalizes</i> well (to other entities, to the future, etc.) and is not overfitted to the historical dataset that was used for deriving or estimating the model.
Interpretability	When a deeper understanding of the retrieved patterns is required—for instance, to validate the model before it is adopted for use—a model needs to be interpretable. This aspect involves a certain degree of subjectivism, since interpretability may depend on the user's knowledge or skills. The interpretability of a model depends on its format, which, in turn, is determined by the adopted analytical technique. Models that allow the user to understand the underlying reasons as to why the model arrives at a certain result are called white-box models, whereas complex incomprehensible mathematical models are often referred to as black-box models. White-box approaches include, for instance, decision trees and linear regression models, examples of which have been provided in Table 1.2. A typical example of a black-box approach concerns neural networks, which are discussed in Chapter 2. It may well be that in a business setting, black-box models are acceptable, although in most settings some level of understanding and in fact validation, which is facilitated by interpretability, is required for the management to have confidence and allow the effective operationalization of the model.
Operational efficiency	Operational efficiency refers to the time that is required to evaluate the model or, in other words, the time required to make a business decision based on the output of the model. When a decision needs to be made in real time or near-real time, for instance to signal possible credit-card fraud or to decide on a rate or banner to advertise on a website, operational efficiency is crucial and is a main concern during model performance assessment. Operational efficiency also entails the efforts needed to collect and preprocess the data, evaluate the model, monitor and back-test the model, and reestimate it when necessary.
Regulatory compliance	Depending on the context, there may be internal or organization-specific as well as external regulation and legislation that apply to the development and application of a model. Clearly, a model should be in line and comply with all applicable regulations and legislation—for instance, with respect to privacy or the use of cookies in a web browser.

Table 1.7 (Continued)

Economical cost	Developing and implementing an analytical model involves significant costs to an organization. The total cost includes, among others, the costs to gather, preprocess, and analyze the data, and the costs to put the resulting analytical models into production. In addition, the software costs, as well as human and computing resources, should be taken into account. Possibly also external data have to be purchased to enrich the available in-house data. On the other hand, benefits can be expected as a result of the adoption of the model. Clearly, it is important to perform a thorough cost-benefit analysis at the start of the project, and to gain insight into the constituent factors of the return-on-investment of building a more advanced system. The profitability of adopting analytics is the central theme of this book. The final chapter concludes by elaborating on the economic impact of analytics.
-----------------	---

discussed for dealing with these specific challenges. Other issues may arise as well, often requiring ingenuity and creativity to be solved. Hence, both are key characteristics of a good data scientist, as is discussed in the following section.

ANALYTICS TEAM

Profiles

The analytics process is essentially a multidisciplinary exercise where many different job profiles need to collaborate. First of all, there is the database or data warehouse administrator (DBA). The DBA ideally is aware of all the data available within the firm, the storage details and the data definitions. Hence, the DBA plays a crucial role in feeding the analytical modeling exercise with its key ingredient, which is data. Since analytics is an iterative exercise, the DBA may continue to play an important role as the modeling exercise proceeds.

Another very important profile is the business expert. This could, for instance, be a credit portfolio manager, brand manager, fraud investigator, or e-commerce manager. The business expert has extensive business experience and business common sense, which usually proves very valuable and crucial for success. It is precisely this knowledge that will help to steer the analytical modeling exercise and interpret its key findings. A key challenge here is that much of the

expert knowledge is tacit and may be hard to elicit at the start of the modeling exercise.

Legal experts are gaining in importance since not all data can be used in an analytical model because of factors such as privacy and discrimination. For instance, in credit risk modeling, one typically cannot discriminate good and bad customers based on gender, beliefs, ethnic origin, or religion. In Web analytics, information is typically gathered by means of cookies, which are files that are stored on the user's browsing computer. However, when gathering information using cookies, users should be appropriately informed. This is subject to regulation at various levels (regional and national, and supranational, e.g., at the European level). A key challenge here is that privacy and other regulatory issues vary highly depending on the geographical region. Hence, the legal expert should have good knowledge about which data can be used when, and which regulation applies in which location.

The software tool vendors should also be mentioned as an important part of the analytics team. Different types of tool vendors can be distinguished here. Some vendors only provide tools to automate specific steps of the analytical modeling process (e.g., data preprocessing). Others sell software that covers the entire analytical modeling process. Some vendors also provide analytics-based solutions for specific application areas, such as risk management, marketing analytics, or campaign management.

The data scientist, modeler, or analyst is the person responsible for doing the actual analytics. The data scientist should possess a thorough understanding of all big data and analytical techniques involved and know how to implement them in a business setting using the appropriate technology. In the next section, we discuss the ideal profile of a data scientist.

Data Scientists

Whereas in a previous section we discussed the characteristics of a good analytical model, in this paragraph we elaborate on the key characteristics of a good data scientist from the perspective of the hiring manager. It is based on our consulting and research experience, having collaborated with many companies worldwide on the topic of big data and analytics.

A Data Scientist Should Have Solid Quantitative Skills

Obviously, a data scientist should have a thorough background in statistics, machine learning and/or data mining. The distinction between these various disciplines is becoming more and more blurred and is actually no longer that relevant. They all provide a set of quantitative techniques to analyze data and find business-relevant patterns within a particular context such as fraud detection or credit risk management. A data scientist should be aware of which technique can be applied, when, and how, and should not focus too much on the underlying mathematical (e.g., optimization) details but, rather, have a good understanding of what analytical problem a technique solves, and how its results should be interpreted. In this context, the education of engineers in computer science and/or business/industrial engineering should aim at an integrated, multidisciplinary view, with graduates formed in both the use of the techniques, and with the business acumen necessary to bring new endeavors to fruition. Also important is to spend enough time validating the analytical results obtained so as to avoid situations often referred to as data massage and/or data torture, whereby data are (intentionally) misrepresented and/or too much time is expended in discussing spurious correlations. When selecting the optimal quantitative technique, the data scientist should consider the specificities of the context and the business problem at hand. Key requirements for business models have been discussed in the previous section, and the data scientist should have a basic understanding of, and intuition for, all of those. Based on a combination of these requirements, the data scientist should be capable of selecting the best analytical technique to solve the particular business problem.

A Data Scientist Should Be a Good Programmer

As per definition, data scientists work with data. This involves plenty of activities such as sampling and preprocessing of data, model estimation, and post-processing (e.g., sensitivity analysis, model deployment, backtesting, model validation). Although many user-friendly software tools are on the market nowadays to automate and support these tasks, every analytical exercise requires tailored steps to tackle the specificities of a particular business problem and setting. In order to successfully perform these steps, programming needs to be done. Hence, a good

data scientist should possess sound programming skills in, for example, SAS, R, or Python, among others. The programming language itself is not that important, as long as the data scientist is familiar with the basic concepts of programming and knows how to use these to automate repetitive tasks or perform specific routines.

A Data Scientist Should Excel in Communication and Visualization Skills

Like it or not, analytics is a technical exercise. At this moment, there is a huge gap between the analytical models and the business users. To bridge this gap, communication and visualization facilities are key! Hence, a data scientist should know how to represent analytical models and their accompanying statistics and reports in user-friendly ways by using, for example, traffic light approaches, OLAP (online analytical processing) facilities, or if-then business rules, among others. A data scientist should be capable of communicating the right amount of information without getting lost in complex (e.g., statistical) details, which will inhibit a model's successful deployment. By doing so, business users will better understand the characteristics and behavior in their (big) data, which will improve their attitude toward and acceptance of the resulting analytical models. Educational institutions must learn to balance between theory and practice, since it is known that many academic degrees mold students who are skewed to either too much analytical or too much practical knowledge.

A Data Scientist Should Have a Solid Business Understanding

While this might seem obvious, we have witnessed (too) many data science projects that failed since the respective data scientist did not understand the business problem at hand. By *business* we refer to the respective application area. Several examples of such application areas have been introduced in Table 1.5. Each of those fields has its own particularities that are important for a data scientist to know and understand in order to be able to design and implement a customized solution. The more aligned the solution with the environment, the better its performance will be, as evaluated according to each of the dimensions or criteria discussed in Table 1.7.

A Data Scientist Should Be Creative!

A data scientist needs creativity on at least two levels. First, on a technical level, it is important to be creative with regard to feature selection,

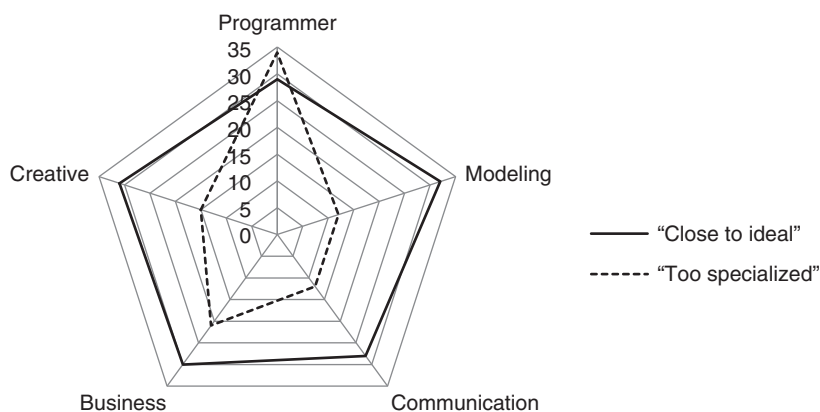


Figure 1.2 Profile of a data scientist.

data transformation and cleaning. These steps of the standard analytics process have to be adapted to each particular application and often the *right guess* could make a big difference. Second, big data and analytics is a fast-evolving field. New problems, technologies, and corresponding challenges pop up on an ongoing basis. Therefore, it is crucial that a data scientist keeps up with these new evolutions and technologies and has enough creativity to see how they can create new opportunities. Figure 1.2 summarizes the key characteristics and strengths constituting the ideal data scientist profile.

CONCLUSION

Profit-driven business analytics is about analyzing data for making optimized operational business decisions. In this first chapter, we discussed how adopting a business perspective toward analytics diverges from a purely technical or statistical perspective. Adopting such a business perspective leads to a real need for approaches that allow data scientists to take into account the specificities of the business context. The objective of this book therefore is to provide an in-depth overview of selected sets of such approaches, which may serve a wide and diverse range of business purposes. The book adopts a practitioner's perspective in detailing how to practically apply and implement these approaches, with example datasets, code, and implementations provided on the book's companion website, www.profit-analytics.com.

REVIEW QUESTIONS

Multiple Choice Questions

Question 1

Which is not a possible evaluation criterion for assessing an analytical model?

- a. Interpretability
- b. Economical cost
- c. Operational efficiency
- d. All of the above are possible evaluation criteria.

Question 2

Which statement is false?

- a. Clustering is a type of predictive analytics.
- b. Forecasting in essence concerns regression in function of time.
- c. Association analysis is a type of descriptive analytics.
- d. Survival analysis in essence concerns predicting the timing of an event.

Question 3

Which statement is true?

- a. Customer lifetime value estimation is an example of classification.
- b. Demand estimation is an example of classification.
- c. Customer churn prediction concerns regression.
- d. Detecting fraudulent credit-card transactions concerns classification.

Question 4

Which is not a characteristic of a good data scientist? A good data scientist:

- a. Has a solid business understanding.
- b. Is creative.
- c. Has thorough knowledge on legal aspects of applying analytics.
- d. Excels in communication and visualization of results.

Question 5

Which statement is true?

- a. All analytical models are profit-driven when applied in a business setting.
- b. Only predictive analytics are profit-driven, whereas descriptive analytics are not.
- c. There is a difference between analyzing data for the purpose of explaining or predicting.
- d. Descriptive analytics aims to explain what is observed, whereas predictive analytics aims to predict as accurately as possible.

Open Questions

Question 1

Discuss the difference between a statistical perspective and a business perspective toward analytics.

Question 2

Discuss the difference between modeling to explain and to predict.

Question 3

List and discuss the key characteristics of an analytical model.

Question 4

List and discuss the ideal characteristics and skills of a data scientist.

Question 5

Draw the analytics process model and briefly discuss the subsequent steps.

REFERENCES

Agrawal, R., and R. Srikant. 1994, September. "Fast algorithms for mining association rules." *In Proceedings of the 20th international conference on very large data bases, VLDB* (Volume 1215, pp. 487–499).

- Athanassopoulos, A. 2000. "Customer Satisfaction Cues to Support Market Segmentation and Explain Switching Behavior." *Journal of Business Research* 47 (3): 191–207.
- Baesens, B. 2014. *Analytics in a Big Data World: The Essential Guide to Data Science and Its Applications*. Hoboken, NJ: John Wiley and Sons.
- Baesens, B., V. Van Vlasselaer, W. Verbeke. 2015. *Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection*. Hoboken, NJ: John Wiley and Sons.
- Bhattacharya, C. B. 1998. "When Customers Are Members: Customer Retention in Paid Membership Contexts." *Journal of the Academy of Marketing Science* 26 (1): 31–44.
- Breiman, L. 2001. "Statistical Modeling: The Two Cultures." *Statistical Science* 16 (3): 199–215.
- Cao, B. 2016. "Future Healthy Life Expectancy among Older Adults in the US: A Forecast Based on Cohort Smoking and Obesity History." *Population Health Metrics*, 14 (1), 1–14.
- Chakraborty, G., P. Murali, and G. Satish. 2013. *Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS*. SAS Institute.
- Coussement, K. 2014. "Improving Customer Retention Management through Cost-Sensitive Learning." *European Journal of Marketing* 48 (3/4): 477–495.
- Dejaeger, K., W. Verbeke, D. Martens, and B. Baesens. 2012. "Data Mining Techniques for Software Effort Estimation: A Comparative Study." *IEEE Transactions on Software Engineering* 38: 375–397.
- Elder IV, J., and H. Thomas. 2012. *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications*. Cambridge, MA: Academic Press.
- Han, J., and M. Kamber. 2011. *Data Mining: Concepts and Techniques*. Amsterdam: Elsevier.
- Hand, D. J., H. Mannila, and P. Smyth. 2001. *Principles of Data Mining*. Cambridge, MA: MIT Press.
- Hyndman, R. J., A. B. Koehler, J. K. Ord, and R. D. Snyder. 2008. "Forecasting with Exponential Smoothing." *Springer Series in Statistics*, 1–356.
- Peto, R., G. Whitlock, and P. Jha. 2010. "Effects of Obesity and Smoking on U.S. Life Expectancy." *The New England Journal of Medicine* 362 (9): 855–857.
- Shmueli, G., and O. R. Koppius. 2011. "Predictive Analytics in Information Systems Research." *MIS Quarterly* 35 (3): 553–572.

- Tan, P.-N., M. Steinbach, and V. Kumar. 2005. *Introduction to Data Mining*. Reading, MA: Addison Wesley.
- Van Gestel, T., and B. Baesens. 2009. *Credit Risk Management: Basic Concepts: Financial Risk Components, Rating Analysis, Models, Economic and Regulatory Capital*. Oxford: Oxford University Press.
- Verbeke, W., D. Martens, and B. Baesens. 2014. "Social Network Analysis for Customer Churn Prediction." *Applied Soft Computing* 14: 431–446.
- Verbraken, T., C. Bravo, R. Weber, and B. Baesens. 2014. "Development and Application of Consumer Credit Scoring Models Using Profit-Based Classification Measures." *European Journal of Operational Research* 238 (2): 505–513.
- Widodo, A., and B. S. Yang. 2011. "Machine Health Prognostics Using Survival Probability and Support Vector Machine." *Expert Systems with Applications* 38 (7): 8430–8437.