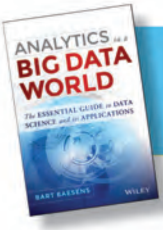
ANALYTICS *in a* BIG DATA WORLD

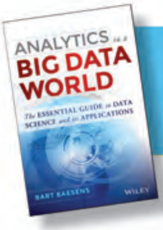
The ESSENTIAL GUIDE *to* DATA SCIENCE *and its* APPLICATIONS

BART BAESENS

WILEY



Contents



CHAPTER 1

Big Data and Analytics

Data are everywhere. IBM projects that every day we generate 2.5 quintillion bytes of data.[1] In relative terms, this means 90 percent of the data in the world has been created in the last two years. Gartner projects that by 2015, 85 percent of Fortune 500 organizations will be unable to exploit big data for competitive advantage and about 4.4 million jobs will be created around big data.[2] Although these estimates should not be interpreted in an absolute sense, they are a strong indication of the ubiquity of big data and the strong need for analytical skills and resources because, as the data piles up, managing and analyzing these data resources in the most optimal way become critical success factors in creating competitive advantage and strategic leverage.

Figure 1.1 shows the results of a KDnuggets[3] poll conducted during April 2013 about the largest data sets analyzed. The total number of respondents was 322 and the numbers per category are indicated between brackets. The median was estimated to be in the 40 to 50 gigabyte (GB) range, which was about double the median answer for a similar poll run in 2012 (20 to 40 GB). This clearly shows the quick increase in size of data that analysts are working on. A further regional breakdown of the poll showed that U.S. data miners lead other regions in big data, with about 28% of them working with terabyte (TB) size databases.

A main obstacle to fully harnessing the power of big data using analytics is the lack of skilled resources and "data scientist" talent required to

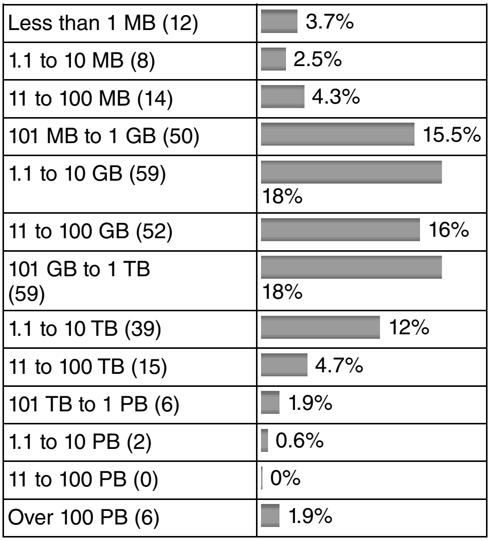


Figure 1.1 Results from a KDnuggets Poll about Largest Data Sets Analyzed
Source: www.kdnuggets.com/polls/2013/largest-dataset-analyzed-data-mined-2013.html.

exploit big data. In another poll ran by KDnuggets in July 2013, a strong need emerged for analytics/big data/data mining/data science education.[4] It is the purpose of this book to try and fill this gap by providing a concise and focused overview of analytics for the business practitioner.

EXAMPLE APPLICATIONS

Analytics is everywhere and strongly embedded into our daily lives. As I am writing this part, I was the subject of various analytical models today. When I checked my physical mailbox this morning, I found a catalogue sent to me most probably as a result of a response modeling analytical exercise that indicated that, given my characteristics and previous purchase behavior, I am likely to buy one or more products from it. Today, I was the subject of a behavioral scoring model of my financial institution. This is a model that will look at, among other things, my checking account balance from the past 12 months and my credit payments during that period, together with other kinds of information available to my bank, to predict whether I will default on my loan during the next year. My bank needs to know this for provisioning purposes. Also today, my telephone services provider analyzed my calling behavior and my account information to predict whether I will churn during the next three months. As I logged on to my Facebook page, the social ads appearing there were based on analyzing all information (posts, pictures, my friends and their behavior, etc.) available to Facebook. My Twitter posts will be analyzed (possibly in real time) by social media analytics to understand both the subject of my tweets and the sentiment of them. As I checked out in the supermarket, my loyalty card was scanned first, followed by all my purchases. This will be used by my supermarket to analyze my market basket, which will help it decide on product bundling, next best offer, improving shelf organization, and so forth. As I made the payment with my credit card, my credit card provider used a fraud detection model to see whether it was a legitimate transaction. When I receive my credit card statement later, it will be accompanied by various vouchers that are the result of an analytical customer segmentation exercise to better understand my expense behavior.

To summarize, the relevance, importance, and impact of analytics are now bigger than ever before and, given that more and more data are being collected and that there is strategic value in knowing what is hidden in data, analytics will continue to grow. Without claiming to be exhaustive, Table 1.1 presents some examples of how analytics is applied in various settings.

Table 1.1 Example Analytics Applications

Marketing	Risk Management	Government	Web	Logistics	Other
Response modeling	Credit risk modeling	Tax avoidance	Web analytics	Demand forecasting	Text analytics
Net lift modeling	Market risk modeling	Social security fraud	Social media analytics	Supply chain analytics	Business process analytics
Retention modeling	Operational risk modeling	Money laundering	Multivariate testing		
Market basket analysis	Fraud detection	Terrorism detection			
Recommender systems					
Customer segmentation					

It is the purpose of this book to discuss the underlying techniques and key challenges to work out the applications shown in Table 1.1 using analytics. Some of these applications will be discussed in further detail in Chapter 8.

BASIC NOMENCLATURE

In order to start doing analytics, some basic vocabulary needs to be defined. A first important concept here concerns the basic unit of analysis. Customers can be considered from various perspectives. Customer lifetime value (CLV) can be measured for either individual customers or at the household level. Another alternative is to look at account behavior. For example, consider a credit scoring exercise for which the aim is to predict whether the applicant will default on a particular mortgage loan account. The analysis can also be done at the transaction level. For example, in insurance fraud detection, one usually performs the analysis at insurance claim level. Also, in web analytics, the basic unit of analysis is usually a web visit or session.

It is also important to note that customers can play different roles. For example, parents can buy goods for their kids, such that there is a clear distinction between the payer and the end user. In a banking setting, a customer can be primary account owner, secondary account owner, main debtor of the credit, codebtor, guarantor, and so on. It is very important to clearly distinguish between those different roles when defining and/or aggregating data for the analytics exercise.

Finally, in case of predictive analytics, the target variable needs to be appropriately defined. For example, when is a customer considered to be a churner or not, a fraudster or not, a responder or not, or how should the CLV be appropriately defined?

ANALYTICS PROCESS MODEL

Figure 1.2 gives a high-level overview of the analytics process model.[5] As a first step, a thorough definition of the business problem to be solved with analytics is needed. Next, all source data need to be identified that could be of potential interest. This is a very important step, as data is 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. All data will then be gathered in a staging area, which could be, for example, a data mart or data warehouse. Some basic exploratory analysis can be considered here using, for example, online analytical processing (OLAP) facilities for multidimensional data 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, and so forth. In the analytics step, an analytical model will be estimated on the preprocessed and transformed data. Different types of analytics can be considered here (e.g., to do churn prediction, fraud detection, customer segmentation, market basket analysis). Finally, once the model has been built, it will be interpreted and evaluated by the business experts. Usually, many trivial patterns will be detected by the model. For example, in a market basket analysis setting, one may find that spaghetti and spaghetti sauce are often purchased together. These patterns are interesting because they provide some validation of the model. But of course, the key issue here is to find the unexpected yet interesting and actionable patterns (sometimes also referred to as *knowledge diamonds*) that can provide added value in the business setting. 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). It is important to consider here how to represent the model output in a user-friendly way, how to integrate it with other applications (e.g., 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 1.2 is iterative in nature, in the sense that one may have to go back to previous steps during the exercise. For example, during the analytics step, the need for additional data may be identified, which may necessitate additional cleaning, transformation, and so forth. Also, the most time consuming step is the data selection and preprocessing step; this usually takes around 80% of the total efforts needed to build an analytical model.

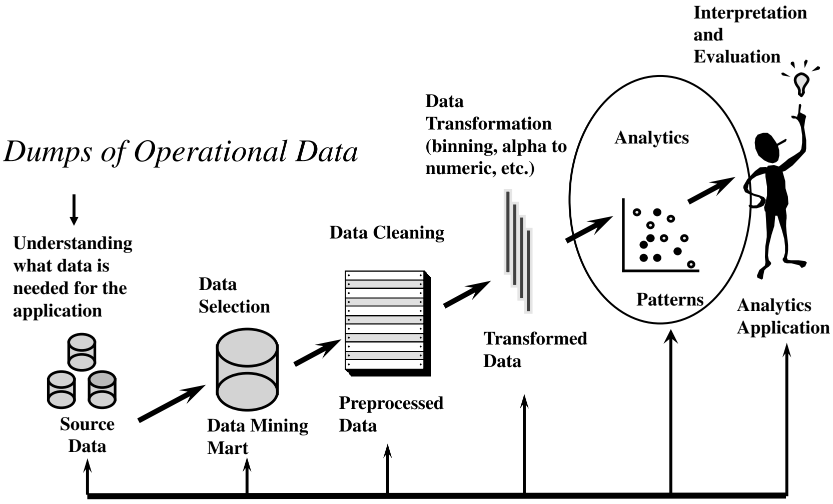

Interpretation and Evaluation
Dumps of Operational Data
Data Transformation (binning, alpha to numeric, etc.)
Analytics
Understanding what data is needed for the application
Data Cleaning
Data Selection
Patterns
Analytics Application
Transformed Data
Preprocessed Data
Source Data
Data Mining Mart

Figure 1.2 The Analytics Process Model

JOB PROFILES INVOLVED

Analytics is essentially a multidisciplinary exercise in which many different job profiles need to collaborate together. In what follows, we will discuss the most important job profiles.

The database or data warehouse administrator (DBA) 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. Because 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 example, be a credit portfolio manager, fraud detection expert, brand manager, or e-commerce manager. This person has extensive business experience and business common sense, which is very valuable. 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 becoming more and more important given that not all data can be used in an analytical model because of privacy, discrimination, and so forth. For example, in credit risk modeling, one can typically not discriminate good and bad customers based upon gender, national 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 (both national and, for example, European). A key challenge here is that privacy and other regulation highly vary depending on the geographical region. Hence, the legal expert should have good knowledge about what data can be used when, and what regulation applies in what location.

The data scientist, data miner, or data analyst is the person responsible for doing the actual analytics. This person should possess a thorough understanding of all techniques involved and know how to implement them using the appropriate software. A good data scientist should also have good communication and presentation skills to report the analytical findings back to the other parties involved.

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 and campaign management, and so on.

ANALYTICS

Analytics is a term that is often used interchangeably with *data science, data mining, knowledge discovery,* and others. The distinction between all those is not clear cut. All of these terms essentially refer to extracting useful business patterns or mathematical decision models from a preprocessed data set. Different underlying techniques can be used for this purpose, stemming from a variety of different disciplines, such as:

- Statistics (e.g., linear and logistic regression)
- Machine learning (e.g., decision trees)

- Biology (e.g., neural networks, genetic algorithms, swarm intelligence)
- Kernel methods (e.g., support vector machines)

Basically, a distinction can be made between predictive and descriptive analytics. In predictive analytics, a target variable is typically available, which can either be categorical (e.g., churn or not, fraud or not) or continuous (e.g., customer lifetime value, loss given default). In descriptive analytics, no such target variable is available. Common examples here are association rules, sequence rules, and clustering. Figure 1.3 provides an example of a decision tree in a classification predictive analytics setting for predicting churn.

More than ever before, analytical models steer the strategic risk decisions of companies. For example, in a bank setting, the minimum equity and provisions a financial institution holds are directly determined by, among other things, credit risk analytics, market risk analytics, operational risk analytics, fraud analytics, and insurance risk analytics. In this setting, analytical model errors directly affect profitability, solvency, shareholder value, the macroeconomy, and society as a whole. Hence, it is of the utmost importance that analytical

Customer	Age	Recency	Frequency	Monetary	Churn
John	35	5	6	100	Yes
Sophie	18	10	2	150	No
Victor	38	28	8	20	No
Laura	44	12	4	280	Yes

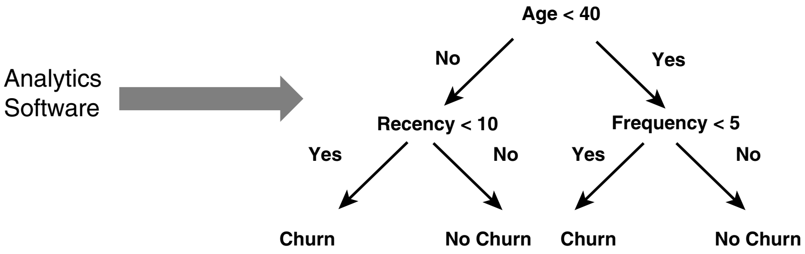


Figure 1.3 Example of Classification Predictive Analytics

models are developed in the most optimal way, taking into account various requirements that will be discussed in what follows.

ANALYTICAL MODEL REQUIREMENTS

A good analytical model should satisfy several requirements, depending on the application area. A first critical success factor is business relevance. The analytical model should actually solve the business problem for which it was developed. It makes no sense to have a working analytical model that got sidetracked from the original problem statement. In order to achieve business relevance, it is of key importance that the business problem to be solved is appropriately defined, qualified, and agreed upon by all parties involved at the outset of the analysis.

A second criterion is statistical performance. The model should have statistical significance and predictive power. How this can be measured will depend upon the type of analytics considered. For example, in a classification setting (churn, fraud), the model should have good discrimination power. In a clustering setting, the clusters should be as homogenous as possible. In later chapters, we will extensively discuss various measures to quantify this.

Depending on the application, analytical models should also be interpretable and justifiable. *Interpretability* refers to understanding the patterns that the analytical model captures. This aspect has a certain degree of subjectivism, since interpretability may depend on the business user's knowledge. In many settings, however, it is considered to be a key requirement. For example, in credit risk modeling or medical diagnosis, interpretable models are absolutely needed to get good insight into the underlying data patterns. In other settings, such as response modeling and fraud detection, having interpretable models may be less of an issue. *Justifiability* refers to the degree to which a model corresponds to prior business knowledge and intuition.[6] For example, a model stating that a higher debt ratio results in more creditworthy clients may be interpretable, but is not justifiable because it contradicts basic financial intuition. Note that both interpretability and justifiability often need to be balanced against statistical performance. Often one will observe that high performing analytical models are incomprehensible and black box in nature. A popular example of this is neural networks, which are universal approximators and are high performing, but offer no insight into the underlying patterns in the data. On the contrary, linear regression models are very transparent and comprehensible, but offer only limited modeling power.

Analytical models should also be *operationally efficient*. This refers to the efforts needed to collect the data, preprocess it, evaluate the model, and feed its outputs to the business application (e.g., campaign management, capital calculation). Especially in a real-time online scoring environment (e.g., fraud detection) this may be a crucial characteristic. Operational efficiency also entails the efforts needed to monitor and backtest the model, and reestimate it when necessary.

Another key attention point is the *economic cost* needed to set up the analytical model. This includes the costs to gather and preprocess the data, the costs to analyze the data, and the costs to put the resulting analytical models into production. In addition, the software costs and human and computing resources should be taken into account here. It is important to do a thorough cost–benefit analysis at the start of the project.

Finally, analytical models should also comply with both local and international *regulation and legislation*. For example, in a credit risk setting, the Basel II and Basel III Capital Accords have been introduced to appropriately identify the types of data that can or cannot be used to build credit risk models. In an insurance setting, the Solvency II Accord plays a similar role. Given the importance of analytics nowadays, more and more regulation is being introduced relating to the development and use of the analytical models. In addition, in the context of privacy, many new regulatory developments are taking place at various levels. A popular example here concerns the use of cookies in a web analytics context.

NOTES

1. IBM, www.ibm.com/big-data/us/en, 2013.
2. www.gartner.com/technology/topics/big-data.jsp.
3. www.kdnuggets.com/polls/2013/largest-dataset-analyzed-data-mined-2013.html.
4. www.kdnuggets.com/polls/2013/analytics-data-science-education.html.

5. J. Han and M. Kamber, *Data Mining: Concepts and Techniques,* 2nd ed. (Morgan Kaufmann, Waltham, MA, US, 2006); D. J. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining* (MIT Press, Cambridge, Massachusetts, London, England, 2001); P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining* (Pearson, Upper Saddle River, New Jersey, US, 2006).
6. D. Martens, J. Vanthienen, W. Verbeke, and B. Baesens, "Performance of Classification Models from a User Perspective." Special issue, *Decision Support Systems* 51, no. 4 (2011): 782–793.