Introduction

Predictive analytics is an area that has been growing in popularity in recent years. However, data mining, of which predictive analytics is a subset, has already reached a steady state in its popularity. In spite of this recent growth and popularity, the underlying science is at least 40 to 50 years old. Engineers and scientists have been using predictive models since at least the first moon project. Humans have always been forward-looking creatures and predictive sciences are a reflection of this curious nature.

So who uses predictive analytics and data mining today? Who are the biggest consumers? A third of the applications are centered on marketing (Rexer, 2013). This involves activities such as customer segmentation and profiling, customer acquisition, customer churn, and customer lifetime value management. Another third of the applications are driven by the banking, financial services and insurance (BFSI) industry, which uses data mining and predictive analytics for activities such as fraud detection and risk analysis. Finally the remaining third of applications are spread among various industries ranging from manufacturing to technology/Internet, medical-pharmaceutical, government, and academia. The activities range from traditional sales forecasting to product recommendations to election sentiment modeling.

While scientific and engineering applications of predictive modeling are based on applying principles of physics or chemistry to develop models, the kind of predictive models we describe in this book are built on empirical knowledge, more specifically, historical data. As our ability to collect, store, and process data has increased in sync with Moore's Law, which implies that computing hardware capabilities double every two years, data mining has found increasing applications in many diverse fields. However, researchers in the area of marketing pioneered much of the early work. Olivia Parr Rud, in her *Data Mining Cookbook* (Parr Rud, 2001) describes an interesting anecdote on how back in the early 1990s building a logistic regression model took about 27 hours. More importantly, the process of predictive analytics had to be carefully orchestrated because a good chunk of model building work is data preparation. So she had

Predictive Analytics and Data Mining. http://dx.doi.org/10.1016/B978-0-12-801460-8.00001-X

to spend a whole week getting her data prepped, and finally submitted the model to run on her PC with a 600MB hard disk over the weekend (while praying that there would be no crashes)! Technology has come a long way in less than 20 years. Today we can run logistic regression models involving hundreds of predictors with hundreds of thousands of records (samples) in a matter of minutes on a laptop computer.

The process of data mining, however, has not changed since those early days and is not likely to change much in the foreseeable future. To get meaningful results from any data, we will still need to spend a majority of effort preparing, cleaning, scrubbing, or standardizing the data before our algorithms can begin to crunch them. But what may change is the automation available to do this. While today this process is iterative and requires analysts' awareness of best practices, very soon we may have smart algorithms doing this for us. This will allow us to focus on the most important aspect of predictive analytics: interpreting the results of the analysis to make decisions. This will also increase the reach of data mining to a broader cross section of analysts and business users.

So what constitutes data mining? Are there a core set of procedures and principles one must master? Finally, how are the two terms—predictive analytics and data mining—different? Before we provide more formal definitions in the next section, it is interesting to look into the experiences of today's data miners based on current surveys (Rexer, 2013). It turns out that a vast majority of data mining practitioners today use a handful of very powerful techniques to accomplish their objectives: decision trees (Chapter 4), regression models (Chapter 5), and clustering (Chapter 7). It turns out that even here an 80/20 rule applies: a majority of the data mining activity can be accomplished using relatively few techniques. However, as with all 80/20 rules, the long tail, which is made up of a large number of less-used techniques, is where the value lies, and for your needs, the best approach may be a relatively obscure technique or a combination of several not so commonly used procedures. Thus it will pay off to learn data mining and predictive analytics in a systematic way, and that is what this book will help you do.

1.1 WHAT DATA MINING IS

Data mining, in simple terms, is finding useful patterns in the data. Being a buzzword, there are a wide variety of definitions and criteria for data mining. Data mining is also referred to as knowledge discovery, machine learning, and predictive analytics. However, each term has a slightly different connotation depending upon the context. In this chapter, we attempt to provide a general overview of data mining and point out its important features, purpose, taxonomy, and common methods.

Data mining starts with *data*, which can range from a simple array of a few numeric observations to a complex matrix of millions of observations with thousands of variables. The act of data mining uses some specialized computational *methods* to discover meaningful and useful structures in the data. These computational methods have been derived from the fields of statistics, machine learning, and artificial intelligence. The discipline of data mining coexists and is closely associated with a number of related areas such as database systems, data cleansing, visualization, exploratory data analysis, and performance evaluation. We can further define data mining by investigating some its key features and motivation.

1.1.1 Extracting Meaningful Patterns

Knowledge discovery in databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns or relationships in the data to make important decisions (Fayyad et al., 1996) The term "nontrivial process" distinguishes data mining from straightforward statistical computations such as calculating the mean or standard deviation. Data mining involves inference and iteration of many different hypotheses. One of the key aspects of data mining is the process of *generalization* of patterns from the data set. The generalization should be valid not just for the data set used to observe the pattern, but also for the new unknown data. Data mining is also a process with defined steps, each with a set of tasks. The term "novel" indicates that data mining is usually involved in finding previously unknown patterns in the data. The ultimate objective of data mining is to find potentially useful conclusions that can be acted upon by the users of the analysis.

1.1.2 Building Representative Models

In statistics, a model is the representation of a relationship between variables in the data. It describes how one or more variables in the data are related to other variables. Modeling is a process in which a representative abstraction is built from the observed data set. For example, we can develop a model based on credit score, income level, and requested loan amount, to determine the interest rate of the loan. For this task, we need previously known observational data with the credit score, income level, loan amount, and interest rate. Figure 1.1 shows the inputs and output of the model. Once the representative model is created, we can use it to predict the value of the interest rate, based on all the input values (credit score, income level, and loan amount).

In the context of predictive analytics, data mining is the process of building the representative model that fits the observational data. This model serves two purposes: on the one hand it predicts the output (interest rate) based on the input variables (credit score, income level, and loan amount), and on the other hand we can use it to understand the relationship between the output variable and all the input variables. For example, does income level really matter in

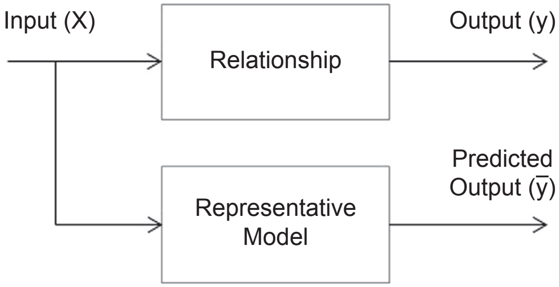

FIGURE 1.1
Representative model for Predictive Analytics.

determining the loan interest rate? Does income level matter more than credit score? What happens when income levels double or if credit score drops by 10 points? Model building in the context of data mining can be used in both predictive and explanatory applications.

1.1.3 Combination of Statistics, Machine Learning, and Computing

In the pursuit of extracting useful and relevant information from large data sets, data mining derives computational techniques from the disciplines of statistics, artificial intelligence, machine learning, database theories, and pattern recognition. Algorithms used in data mining originated from these disciplines, but have since evolved to adopt more diverse techniques such as parallel computing, evolutionary computing, linguistics, and behavioral studies. One of the key ingredients of successful data mining is substantial prior knowledge about the data and the business processes that generate the data, known as *subject matter expertise*. Like many quantitative frameworks, data mining is an iterative process in which the practitioner gains more information about the patterns and relationships from data in each cycle. The art of data mining combines the knowledge of statistics, subject matter expertise, database technologies, and machine learning techniques to extract meaningful and useful information from the data. Data mining also typically operates on large data sets that need to be stored, processed, and computed. This is where database techniques along with parallel and distributed computing techniques play an important role in data mining.

1.1.4 Algorithms

We can also define data mining as a process of discovering previously unknown patterns in the data using *automatic iterative methods*. Algorithms are iterative step-by-step procedure to transform inputs to output. The application of sophisticated algorithms for extracting useful patterns from the data differentiates data mining from traditional data analysis techniques. Most of these algorithms were developed in recent decades and have been borrowed from the fields of machine learning and artificial intelligence. However, some of the algorithms are based on the foundations of Bayesian probabilistic theories and regression analysis, originated hundreds of years ago. These iterative algorithms automate the process of searching for an optimal solution for a given data problem. Based on the data problem, data mining is classified into tasks such as classification, association analysis, clustering, and regression. Each data mining task uses specific algorithms like decision trees, neural networks, k-nearest neighbors, k-means clustering, among others. With increased research on data mining, the number of such algorithms is increasing, but a few classic algorithms remain foundational to many data mining applications.

1.2 WHAT DATA MINING IS *NOT*

While data mining covers a wide set of techniques, applications, and disciplines, not all analytical and discovery methods are considered data mining processes. Data mining is usually applied, though not limited to, large data sets. Data mining also goes through a defined process of exploration, preprocessing, modeling, evaluation, and knowledge extraction. Here are some commonly used data discovery techniques that are not considered data mining, even if they operate on large data sets:

- **Descriptive statistics:** Computing mean, standard deviation, and other descriptive statistics quantify the aggregate structure of a data set. This is essential information to understand any data set, but calculating these statistics is not considered a data mining technique. However, they are used in the exploration stage of the data mining process.
- **Exploratory visualization:** The process of expressing data in visual coordinates enables users to find patterns and relationships in the data and comprehend large data sets. Similar to descriptive statistics, they are integral in the preprocessing and postprocessing steps in data mining.
- **Dimensional slicing:** Business intelligence and online analytical processing (OLAP) applications, which are prevalent in business settings, mainly provide information on the data through dimensional slicing, filtering ,and pivoting. OLAP analysis is enabled by a unique database schema design where the data is organized as *dimensions* (e.g., Products, Region, Date) and quantitative facts or *measures* (e.g., Revenue, Quantity). With a well-defined database structure, it is easy to slice the yearly revenue by products or combination of region and products. While these techniques are extremely useful and may provide patterns in data (e.g., Candy sales decline after Halloween in the United States), this is considered information retrieval and not data mining.
- **Hypothesis testing:** In confirmatory data analysis, experimental data is collected to evaluate whether a hypothesis has enough evidence to support it or not. There are many types of statistical testing and

they have a wide variety of business applications (e.g., A/B testing in marketing). In general, data mining is a process where many hypotheses are generated and tested based on observational data. Since the data mining algorithms are iterative, we can refine the solution in each step.

- **Queries:** Information retrieval systems, like web search engines, use data mining techniques like clustering to index vast repositories of data. But the act of querying and rendering of the result is not considered a data mining process. Query retrieval from databases and slicing and dicing of data are not generally considered data mining (Tan et al., 2005).

All of the above techniques are used in the steps of a data mining process and are used in conjunction with the term "data mining." It is important for the practitioner to know what makes up a complete data mining process. We will discuss the specific steps of a data mining process in the next chapter.

1.3 THE CASE FOR DATA MINING

In the past few decades, we have seen a massive accumulation of data with the advancement of information technology, connected networks and businesses it enables. This trend is also coupled with steep decline in the cost of data storage and data processing. The applications built on these advancements like online businesses, social networking, and mobile technologies unleash a large amount of complex, heterogeneous data that are waiting to be analyzed. Traditional analysis techniques like dimensional slicing, hypothesis testing, and descriptive statistics can only get us so far in information discovery. We need a paradigm to manage massive volume of data, explore the interrelationships of thousands of variables, and deploy machine learning algorithms to deduce optimal insights from the data set. We need a set of frameworks, tools, and techniques to intelligently assist humans to process all these data and extract valuable information (Piatetsky-Shapiro et al., 1996). Data Mining is one such paradigm that can handle large volumes with multiple attributes and deploy complex algorithms to search for patterns from the data. Let's explore each key motivation for using data mining techniques.

1.3.1 Volume

The sheer volume of data captured by organizations is exponentially increasing. The rapid decline in storage costs and advancements in capturing every transaction and event, combined with the business need to extract all possible leverage using data, creates a strong motivation to store more data than ever. A study by IDC Corporation in 2012 reported that the volume of recorded digital data by 2012 reached 2.8 zettabytes, and less than 1% of the data are currently analyzed (Reinsel, December 2012). As data becomes more granular, the need for using large volume data to extract information increases. A rapid increase in the volume of data exposes the limitations of current analysis methodologies. In a few implementations, the time to create generalization models is quite critical and data volume plays a major part in determining the time to development and deployment.

1.3.2 Dimensions

The three characteristics of the Big Data phenomenon are high volume, high velocity, and high variety. Variety of data relates to multiple types of values (numerical, categorical), formats of data (audio files, video files), and application of data (location coordinates, graph data). Every single record or data point contains multiple attributes or variables to provide context for the record. For example, every user record of an ecommerce site can contain attributes such as products viewed, products purchased, user demographics, frequency of purchase, click stream, etc. Determining what is the most effective offer an ecommerce user will respond to can involve computing information along all these attributes. Each attribute can be thought as a dimension in the data space. The user record has multiple attributes and can be visualized in multidimensional space. Addition of each dimension increases the complexity of analysis techniques.

A simple linear regression model that has one input dimension is relatively easier to build than multiple linear regression models with multiple dimensions. As the dimensional space of the data increases, we need an adaptable framework that can work well with multiple data types and multiple attributes. In the case of text mining, a document or article becomes a data point with each unique word as a dimension. Text mining yields a data set where the number of attributes ranges from a few hundred to hundreds of thousands of attributes.

1.3.3 Complex Questions

As more complex data are available for analysis, the complexity of information that needs to get extracted from the data is increasing as well. If we need to find the natural clusters in a data set with hundreds of dimensions, traditional analysis like hypothesis testing techniques cannot be used in a scalable fashion. We need to leverage machine-learning algorithms to automate searching in the vast search space.

Traditional statistical analysis approaches a data analysis problem by assuming a stochastic model to predict a response variable based on a set of input variables. Linear regression and logistic regression analysis are classic examples of this technique where the parameters of the model are estimated from the data. These hypothesis-driven techniques were highly successful in modeling simple relationships between response and input variables. However, there is a significant need to extract nuggets of information from large, complex data sets, where the use of traditional statistical data analysis techniques is limited (Breiman, 2001)

Machine learning approach the problem of modeling by trying to find an algorithmic model that can better predict the output from input variables. The algorithms are usually recursive and in each cycle estimate the output and "learn" from the predictive errors of previous steps. This route of modeling greatly assists in exploratory analysis since the approach here is not validating a hypothesis but generating a multitude of hypotheses for a given problem. In the context of the data problems we face today, we need to deploy both techniques. John Tuckey, in his article "We need both exploratory and confirmatory," stresses the importance of both exploratory and confirmatory analysis techniques (Tuckey, 1980). In this book, we discuss a range of data mining techniques, from traditional statistical modeling techniques like regressions to machine-learning algorithms.

1.4 TYPES OF DATA MINING

Data mining problems can be broadly categorized into *supervised* or *unsupervised* learning models. Supervised or directed data mining tries to infer a function or relationship based on labeled training data and uses this function to map new unlabeled data. Supervised techniques predict the value of the output variables based on a set of input variables. To do this, a model is developed from a *training* data set where the values of input and output are previously known. The model generalizes the relationship between the input and output variables and uses it to predict for the data set where only input variables are known. The output variable that is being predicted is also called a class label or target variable. Supervised data mining needs a sufficient number of labeled records to learn the model from the data. Unsupervised or undirected data mining uncovers hidden patterns in unlabeled data. In unsupervised data mining, there are no output variables to predict. The objective of this class of data mining techniques is to find patterns in data based on the relationship between data points themselves. An application can employ both supervised and unsupervised learners.

Data mining problems can also be grouped into classification, regression, association analysis, anomaly detection, time series, and text mining tasks (Figure 1.2). This book is organized around these data mining tasks. We present an overview of the types of data mining in this chapter and will provide an in-depth discussion of concepts and step-by-step implementations of many important techniques in the following chapters.

Classification and *regression* techniques predict a target variable based on input variables. The prediction is based on a generalized model built from a previously known data set. In regression tasks, the output variable is numeric (e.g., the mortgage interest rate on a loan). Classification tasks predict output variables, which are categorical or polynomial (e.g., the yes or no decision to approve a loan). *Clustering* is the process of identifying the natural groupings in the data set. For example, clustering is helpful in finding natural clusters in customer data sets, which can be used for market segmentation. Since this is unsupervised data mining, it is up to the end user to investigate why these clusters are formed in the data and generalize the uniqueness of each cluster. In retail analytics, it is common to identify pairs of items that are purchased together, so that specific items can be bundled or placed next to each other. This task is called market basket analysis or *association analysis*, which is commonly used in recommendation engines.

Anomaly or outlier detection identifies the data points that are significantly different from other data points in the data set. Credit card transaction fraud detection is one of the most prolific applications of anomaly detection. *Time series forecasting* can be either a special use of regression modeling (where models predict the future value of a variable based on the past value of the same variable) or a sophisticated averaging or smoothing technique (for example, daily weather prediction based on the past few years of daily data).

Text Mining is a data mining application where the input data is text, which can be in the form of documents, messages, emails, or web pages. To aid the

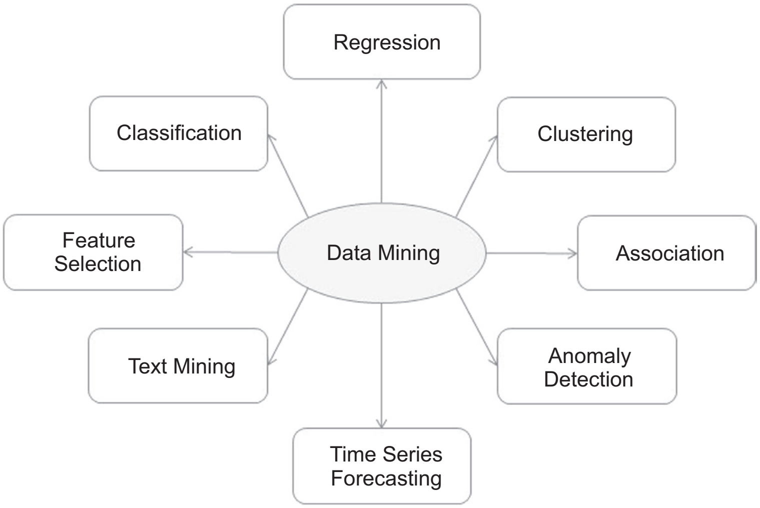

FIGURE 1.2
Data mining tasks.

data mining on text data, the text files are converted into document vectors where each unique word is considered an attribute. Once the text file is converted to document vectors, standard data mining tasks such as classification, clustering, etc. can be applied on text files. The *Feature selection* is a process in which attributes in a data set is reduced to a few attributes that really matter.

A complete data mining application can contain elements of both supervised and unsupervised techniques. Unsupervised techniques provide an increased understanding of the data set and hence are sometimes called descriptive data mining. As an example of how both unsupervised and supervised data mining can be combined in an application, consider the following scenario. In marketing analytics, clustering can be used to find the natural clusters in customer records. Each customer is assigned a cluster label at the end of the clustering process. A labeled customer data set can now be used to develop a model that assigns a cluster label for any new customer record with a supervised classification technique.

1.5 DATA MINING ALGORITHMS

An algorithm is a logical step-by-step procedure for solving a problem. In data mining, it is the blueprint for how a particular data problem is solved. Many of the algorithms are recursive, where a set of steps are repeated many times until a limiting condition is met. Some algorithms also contain a random variable as an input, and are aptly called *randomized algorithms*. A data mining classification task can be solved using many different approaches or algorithms such as decision trees, artificial neural networks, k-nearest neighbors (k-NN), and even some regression algorithms. The choice of which algorithm to use depends on the type of data set, objective of the data mining, structure of the data, presence of outliers, available computational power, number of records, number of attributes, and so on. It is up to the data mining practitioner to make a decision about what algorithm(s) to use by evaluating the performance of multiple algorithms. There have been hundreds of algorithms developed in the last few decades to solve data mining problems. In the next few chapters, we will discuss the inner workings of the most important and diverse data mining algorithms and their implementations.

Data mining algorithms can be implemented by custom-developed computer programs in almost any computer language. This obviously is a time-consuming task. In order for us to focus our time on data and algorithms, we can leverage data mining tools or statistical programing tools, like R, RapidMiner, SAS Enterprise Miner, IBM SPSS, etc., which can implement these algorithms with ease. These data mining tools offer a library of algorithms as functions, which can be interfaced through programming code or configuration through graphical user interfaces. Table 1.1 provides a summary of data mining tasks with commonly used algorithmic techniques and example use cases.

Table 1.1 Data Mining Tasks and Examples

Tasks	Description	Algorithms	Examples
Classification	Predict if a data point belongs to one of the predefined classes. The prediction will be based on learning from a known data set.	Decision trees, neural networks, Bayesian models, induction rules, k-nearest neighbors	Assigning voters into known buckets by political parties, e.g., soccer moms Bucketing new customers into one of the known customer groups
Regression	Predict the numeric target label of a data point. The prediction will be based on learning from a known data set.	Linear regression, logistic regression	Predicting unemployment rate for next year Estimating insurance premium
Anomaly detection	Predict if a data point is an outlier compared to other data points in the data set.	Distance based, density based, local outlier factor (LOF)	Fraud transaction detection in credit cards Network intrusion detection
Time series	Predict the value of the target variable for a future time frame based on historical values.	Exponential smoothing, autoregressive integrated moving average (ARIMA), regression	Sales forecasting, production forecasting, virtually any growth phenomenon that needs to be extrapolated
Clustering	Identify natural clusters within the data set based on inherit properties within the data set.	k-means, density-based clustering (e.g., density-based spatial clustering of applications with noise [DBSCAN])	Finding customer segments in a company based on transaction, web, and customer call data
Association analysis	Identify relationships within an item set based on transaction data.	Frequent Pattern Growth (FP-Growth) algorithm, Apriori algorithm	Find cross-selling opportunities for a retailer based on transaction purchase history

1.6 ROADMAP FOR UPCOMING CHAPTERS

It's time to explore data mining and predictive analytics techniques in more detail. In the next couple of chapters, we provide an overview of the data mining process and data exploration techniques. The following chapters present the main body of this book: the concepts behind each predictive analytics or descriptive data mining algorithm and a practical use case (or two) for each. You don't have to read the chapters in a sequence. We have organized this book in such a way that you can directly start reading about the data mining tasks and algorithms you are most interested in. Within each chapter focused on a technique (e.g., decision tree, k-means clustering), we start with a general overview, and then present the concepts and the logic of the algorithm and how it works in plain language. Later we show how the algorithm can be implemented using RapidMiner. RapidMiner is a widely known and used software tool for data mining and predictive analytics (Piatetsky, 2014) and we have chosen it particularly for ease of implementation using GUI and it is a open source data mining tool. We conclude each chapter with some closing thoughts and list further reading materials and references. Here is a roadmap of the book.

1.6.1 Getting Started with Data Mining

Successfully uncovering patterns in a data set is an iterative process. Chapter 2 Data Mining Process provides a framework to solve data mining problems. A five-step process outlined in this chapter provides guidelines on gathering subject matter expertise; exploring the data with statistics and visualization; building a model using data mining algorithms; testing the model and deploying in production environment; and finally reflecting on new knowledge gained in the cycle.

A simple data exploration either visually or with the help of basic statistical analysis can sometimes answer seemingly tough questions meant for data mining. Chapter 3 Data Exploration covers some of the basic tools used in knowledge discovery before deploying data mining techniques. These practical tools increase one's understanding of the data and are quite essential in understanding the results of data mining process.

1.6.2 An Interlude...

Before we dive into the key data mining techniques and algorithms, we want to point out two specific things regarding how you can implement Data Mining algorithms while reading this book. We believe learning the concepts and implementation immediately after enhances the learning experience. All of the predictive modeling and data mining algorithms explained in the following chapters are implemented in RapidMiner. First, we recommend that you download the free version of RapidMiner software from http://www.rapidminer.com (if you have not done so already) and second, review the first couple of sections of Chapter 13 Getting Started with RapidMiner to familiarize yourself with the features of the tool, its basic operations, and the user interface functionality. Acclimating with RapidMiner will be helpful while using the algorithms that are discussed in the following chapters. This chapter is set at the end of the book because some of the later sections in the chapter build upon the material presented in the chapters on algorithms; however the first few sections are a good starting point for someone who is not yet familiar with the tool.

Each chapter has a data set we use to describe the concept of a particular data mining task and in most cases the same data set is used for implementation. Step-by-step instructions on practicing data mining on the data set are covered in every algorithm that is discussed in the upcoming chapters. All the implementations discussed in the book are available at the companion website of the book at www.LearnPredictiveAnalytics.com.
Though not required, we encourage you to access these files to aid your learning. You can download the data set, complete RapidMiner processes (*.rmp files), and many more relevant electronic files from this website.

1.6.3 The Main Event: Predictive Analytics and Data Mining Algorithms

Classification is the most widely used data mining task in businesses. As a predictive analytics task, the objective of a classification model is to predict a target variable that is binary (e.g., a loan decision) or categorical (e.g., a customer type) when a set of input variables are given (e.g., credit score, income level, etc.). The model does this by learning the generalized relationship between the predicted target variable with all other input attributes from a known data set. There are several ways to skin this cat. Each algorithm differs by how the relationship is extracted from the known data, called a "training" data set. Chapter 4 on classification addresses several of these methods.

- *Decision trees* approach the classification problem by partitioning the data into "purer" subsets based on the values of the input attributes. The attributes that help achieve the cleanest levels of such separation are considered significant in their influence on the target variable and end up at the root and closer-to-root levels of the tree. The output model is a tree framework than can be used for the prediction of new unlabeled data.
- *Rule induction* is a data mining process of deducing IF-THEN rules from a dataset or from decision trees. These symbolic decision rules explain an inherent relationship between the attributes and labels in the data set that can be easily understood by everyone.
- *Naïve Bayesian* algorithms provide a probabilistic way of building a model. This approach calculates the probability for each value of the class variable for given values of input variables. With the help of conditional probabilities, for a given unknown record, the model calculates the outcome of all values of target classes and comes up with a predicted winner.
- Why go through the trouble of extracting complex relationships from the data when we can just memorize entire training data set and pretend we have generalized the relationship? This is exactly what the *k-nearest neighbor* algorithm does, and it is therefore called a "lazy" learner where the entire training data set is memorized as the model.
- Neurons are the nerve cells that connect with each other to form a biological neural network. The working of these interconnected nerve cells inspired the solution of some complex data problems by the creation of *artificial neural networks*. The neural networks section provides a conceptual background of how a simple neural network works and how to implement one for any general prediction problem.

- *Support vector machines (SVMs)* were developed to address optical character recognition problems: how can we train an algorithm to detect boundaries between different patterns and thus identify characters? SVMs can therefore identify if a given data sample belongs within a boundary (in a particular class) or outside it (not in the class).
- *Ensemble learners* are "meta" models where the model is a combination of several different individual models. If certain conditions are met, ensemble learners can gain from the wisdom of crowds and greatly reduce the generalization error in data mining.

The simple mathematical equation y = ax + b is a linear regression model. Chapter 5 Regression Methods describes a class of predictive analytics techniques in which the target variable (e.g., interest rate or a target class) is *functionally* related to input variables.

- **Linear regression:** The simplest of all function fitting models is based on a linear equation, as mentioned above. Polynomial regression uses higher-order equations. No matter what type of equation is used, the goal is to represent the variable to be predicted in terms of other variables or attributes. Further, the predicted variable and the independent variables all have to be numeric for this to work. We explore the basics of building regression models and show how predictions can be made using such models.
- **Logistic regression:** It addresses the issue of predicting a target variable that may be binary or binomial (such as 1 or 0, yes or no) using predictors or attributes, which may be numeric.

Supervised data mining or predictive analytics predict the value of the target variables. In the next two chapters, we review two important *unsupervised* data mining tasks: Association analysis in Chapter 6 and Clustering in Chapter 7. Ever heard of the beer and diaper association in supermarkets? Apparently, a supermarket discovered that customers who buy diapers also tend to buy beer. While this may have been an urban legend, the observation has become a poster child for association analysis. Associating an item in a transaction with another item in the transaction to determine the most frequently occurring patterns is termed *association analysis*. This technique is about, for example, finding relationships between products in a supermarket based on purchase data, or finding related web pages in a website based on click stream data. This data mining application is widely used in retail, ecommerce, and media to creatively bundle products.

Clustering is the data mining task of identifying natural groups in the data. For an unsupervised data mining task, there is no target class variable to predict. After the clustering is performed, each record in the data set is associated with one or more cluster. Widely used in marketing segmentations and text mining, clustering can be performed by a wide range of algorithms. In Chapter 7, we will discuss three common algorithms with diverse identification approaches. The *k-means clustering* technique identifies a cluster based on a central prototype record. *DBSCAN* clustering partitions the data based on variation in the density of records in a data set. *Self-organizing maps (SOM)* create a two-dimensional grid where all the records related with each other are placed next to each other.

How do we determine which algorithms work best for a given data set? Or for that matter how do we objectively quantify the performance of any algorithm on a data set? These questions are addressed in Chapter 8 Model Evaluation, which covers performance evaluation. We describe the most commonly used tools for evaluating classification models such as a confusion matrix, ROC curves, and lift charts.

1.6.4 Special Applications

Chapter 9 Text Mining provides a detailed look into the emerging area of text mining and text analytics. It starts with a background on the origins of text mining and provides the motivation for this fascinating topic using the example of IBM's Watson, the Jeopardy!-winning computer program that was built almost entirely using concepts from text and data mining. The chapter introduces some key concepts important in the area of text analytics such as term frequency–inverse document frequency (TF-IDF) scores. Finally it describes two hands-on case studies in which the reader is shown how to use RapidMiner to address problems like document clustering and automatic gender classification based on text content.

Forecasting is a very common application of time series analysis. Companies use sales forecasts, budget forecasts, or production forecasts in their planning cycles. Chapter 10 on Time Series Forecasting starts by pointing out the clear distinction between standard supervised predictive models and time series forecasting models. It provides a basic introduction to the different time series methods ranging from data-driven moving averages to exponential smoothing, and model-driven forecasts including polynomial regression and lag-series based ARIMA methods.

Chapter 11 on Anomaly Detection describes how outliers in data can be detected by combining multiple data mining tasks like classification, regression, and clustering. The fraud alert received from credit card companies is the result of an anomaly detection algorithm. The target variable to be predicted is whether a transaction is an outlier or not. Since clustering tasks identify outliers as a cluster, distance-based and density-based clustering techniques can be used in anomaly detection tasks.

In predictive analytics, the objective is to develop a representative model to generalize the relationship between input attributes and target attributes, so that we can predict the value or class of the target variables. Chapter 12 introduces a preprocessing step that is often critical for a successful predictive modeling exercise: *feature selection*. Feature selection is known by several alternative terms such as attribute weighting, dimension reduction, and so on. There are two main styles of feature selection: filtering the key attributes before modeling (filter style) or selecting the attributes during the process of modeling (wrapper style). We discuss a few filter-based methods such as principal component analysis (PCA), information gain, and chi-square, and a couple of wrapper-type methods like forward selection and backward elimination. Even in just one data mining algorithm, there are many different ways to tweak the parameters and even the sampling for training data set.

If you are not familiar with RapidMiner, the first few sections of Chapter 13 Getting Started with RapidMiner should provide a good overview, while the latter sections of this chapter discuss some of the commonly used productivity tools and techniques such as data transformation, missing value handling, and process optimizations using RapidMiner. As mentioned earlier, while each chapter is more or less independent, some of the concepts in Chapters 8 Model Evaluation and later build on the material from earlier chapters and for beginners we recommend going in order. However, if you are familiar with the standard terminology and with RapidMiner, you are not constrained to move in any fashion.

REFERENCES

Breiman, L. (2001). Statistical Modeling: Two Cultures. *Statistical Science*, *6*(3), 199–231.

Fayyad, U., Piatetsky-shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37–54.

Parr Rud, O. (2001). *Data Mining Cookbook*. New York: John Wiley and Sons.

Piatetsky, G. (2014). KDnuggets 15th Annual Analytics, Data Mining, Data Science Software Poll: RapidMiner Continues To Lead. Retrieved August 01, 2014, from http://www.kdnuggets.com/2014/06/kdnuggets-annual-software-poll-rapidminer-continues-lead.html.

Piatetsky-Shapiro, G., Brachman, R., Khabaza, T., Kloesgen, W., & Simoudis, E. (1996). *An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications*. KDD-96 Conference Proceedings.

Reinsel, J. G. (December 2012). *Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East* Sponsored by EMC Corporation. IDC iView.

Rexer, K. (2013). *2013 Data Miner Survey Summary Report*. Winchester, MA: Rexer Analytics. www.rexeranalytics.com.

Tan, P.-N., Michael, S., & Kumar, V. (2005). *Introduction to Data Mining*. Boston, MA: Addison-Wesley.

Tuckey, J. (1980). We need exploratory and Confirmatory. *The American Statistician*, *34*(1), 23–25.