DATA MINING FOR BUSINESS ANALYTICS

CONCEPTS, TECHNIQUES, AND APPLICATIONS IN R

Galit Shmueli • Peter C. Bruce
Inbal Yahav • Nitin R. Patel
Kenneth C. Lichtendahl Jr.

with website

WILEY

DATA MINING
FOR BUSINESS ANALYTICS

Concepts, Techniques, and Applications in R

GALIT SHMUELI
PETER C. BRUCE
INBAL YAHAV
NITIN R. PATEL
KENNETH C. LICHTENDAHL, JR.

WILEY

This edition first published 2018

Library of Congress Cataloging-in-Publication Data applied for

Hardback: 9781118879368

Cover Design: Wiley
Cover Image: © Achim Mittler, Frankfurt am Main/Gettyimages

Set in 11.5/14.5pt BemboStd by Aptara Inc., New Delhi, India
Printed in the United States of America.

Contents

PART II DATA EXPLORATION AND DIMENSION REDUCTION

PART III PERFORMANCE EVALUATION

PART IV PREDICTION AND CLASSIFICATION METHODS

PART V MINING RELATIONSHIPS AMONG RECORDS

Part VI FORECASTING TIME SERIES

PART VIII CASES

Foreword by Gareth James

The field of statistics has existed in one form or another for 200 years, and by the second half of the 20th century had evolved into a well-respected and essential academic discipline. However, its prominence expanded rapidly in the 1990s with the explosion of new, and enormous, data sources. For the first part of this century, much of this attention was focused on biological applications, in particular, genetics data generated as a result of the sequencing of the human genome. However, the last decade has seen a dramatic increase in the availability of data in the business disciplines, and a corresponding interest in business-related statistical applications.

The impact has been profound. Ten years ago, when I was able to attract a full class of MBA students to my new statistical learning elective, my colleagues were astonished because our department struggled to fill most electives. Today, we offer a Masters in Business Analytics, which is the largest specialized masters program in the school and has application volume rivaling those of our MBA programs. Our department's faculty size and course offerings have increased dramatically, yet the MBA students are still complaining that the classes are all full. Google's chief economist, Hal Varian, was indeed correct in 2009 when he stated that "the sexy job in the next 10 years will be statisticians."

This demand is driven by a simple, but undeniable, fact. Business analytics solutions have produced significant and measurable improvements in business performance, on multiple dimensions and in numerous settings, and as a result, there is a tremendous demand for individuals with the requisite skill set. However, training students in these skills is challenging given that, in addition to the obvious required knowledge of statistical methods, they need to understand business-related issues, possess strong communication skills, and be comfortable dealing with multiple computational packages. Most statistics texts concentrate on abstract training in classical methods, without much emphasis on practical, let alone business, applications.

This book has by far the most comprehensive review of business analytics methods that I have ever seen, covering everything from classical approaches such as linear and logistic regression, through to modern methods like neural networks, bagging and boosting, and even much more business specific procedures such as social network analysis and text mining. If not the bible, it is at the least a definitive manual on the subject. However, just as important as the list of topics, is the way that they are all presented in an applied fashion using business applications. Indeed the last chapter is entirely dedicated to 10 separate cases where business analytics approaches can be applied.

In this latest edition, the authors have added an important new dimension in the form of the R software package. Easily the most widely used and influential open source statistical software, R has become the go-to tool for such purposes. With literally hundreds of freely available add-on packages, R can be used for almost any business analytics related problem. The book provides detailed descriptions and code involving applications of R in numerous business settings, ensuring that the reader will actually be able to apply their knowledge to real-life problems.

We recently introduced a business analytics course into our required MBA core curriculum and I intend to make heavy use of this book in developing the syllabus. I'm confident that it will be an indispensable tool for any such course.

GARETH JAMES

Marshall School of Business, University of Southern California, 2017

Foreword by Ravi Bapna

Data is the new gold—and mining this gold to create business value in today's context of a highly networked and digital society requires a skillset that we haven't traditionally delivered in business or statistics or engineering programs on their own. For those businesses and organizations that feel overwhelmed by today's Big Data, the phrase *you ain't seen nothing yet* comes to mind. Yesterday's three major sources of Big Data—the 20+ years of investment in enterprise systems (ERP, CRM, SCM, ...), the 3 billion plus people on the online social grid, and the close to 5 billion people carrying increasingly sophisticated mobile devices—are going to be dwarfed by tomorrow's smarter physical ecosystems fueled by the Internet of Things (IoT) movement.

The idea that we can use sensors to connect physical objects such as homes, automobiles, roads, even garbage bins and streetlights, to digitally optimized systems of governance goes hand in glove with bigger data and the need for deeper analytical capabilities. We are not far away from a smart refrigerator sensing that you are short on, say, eggs, populating your grocery store's mobile app's shopping list, and arranging a Task Rabbit to do a grocery run for you. Or the refrigerator negotiating a deal with an Uber driver to deliver an evening meal to you. Nor are we far away from sensors embedded in roads and vehicles that can compute traffic congestion, track roadway wear and tear, record vehicle use and factor these into dynamic usage-based pricing, insurance rates, and even taxation. This brave new world is going to be fueled by analytics and the ability to harness data for competitive advantage.

Business Analytics is an emerging discipline that is going to help us ride this new wave. This new Business Analytics discipline requires individuals who are grounded in the fundamentals of business such that they know the right questions to ask, who have the ability to harness, store, and optimally process vast datasets from a variety of structured and unstructured sources, and who can then use an array of techniques from machine learning and statistics to uncover new insights for decision-making. Such individuals are a rare commodity today, but their creation has been the focus of this book for a decade now. This book's forte is that it relies on explaining the core set of concepts required for today's business analytics professionals using real-world data-rich cases in a hands-on manner, without sacrificing academic rigor. It provides a modern day foundation for Business Analytics, the notion of linking the x's to the y's of interest in a predictive sense. I say this with the confidence of someone who was probably the first adopter of the zeroth edition of this book (Spring 2006 at the Indian School of Business).

I can't say enough about the long-awaited R edition. R is my go-to platform for analytics these days. It's also used by a wide variety of instructors in our MS-Business Analytics program. The open-innovation paradigm used by R is one key part of the analytics perfect storm, the other components being the advances in computing and the business appetite for data-driven decision-making.

I look forward to using the book in multiple fora, in executive education, in MBA classrooms, in MS-Business Analytics programs, and in Data Science bootcamps. I trust you will too!

RAVI BAPNA

Carlson School of Management, University of Minnesota, 2017

Preface to the R Edition

This textbook first appeared in early 2007 and has been used by numerous students and practitioners and in many courses, ranging from dedicated data mining classes to more general business analytics courses (including our own experience teaching this material both online and in person for more than 10 years). The first edition, based on the Excel add-in XLMiner, was followed by two more XLMiner editions, a JMP edition, and now this R edition, with its companion website, www.dataminingbook.com.

This new R edition, which relies on the free and open-source R software, presents output from R, as well as the code used to produce that output, including specification of a variety of packages and functions. Unlike computer-science or statistics-oriented textbooks, the focus in this book is on data mining concepts, and how to implement the associated algorithms in R. We assume a basic facility with R.

For this R edition, two new co-authors stepped on board—Inbal Yahav and Casey Lichtendahl—bringing both expertise teaching business analytics courses using R and data mining consulting experience in business and government. Such practical experience is important, since the open-source nature of R software makes available a plethora of approaches, packages, and functions available for data mining. Given the main goal of this book—to introduce data mining concepts using R software for illustration—our challenge was to choose an R code cocktail that supports highlighting the important concepts. In addition to providing R code and output, this edition also incorporates updates and new material based on feedback from instructors teaching MBA, undergraduate, diploma, and executive courses, and from their students as well.

One update, compared to the first two editions of the book, is the title: we now use *Business Analytics* in place of *Business Intelligence*. This reflects the change in terminology since the second edition: Business Intelligence today refers mainly to reporting and data visualization ("what is happening now"), while Business Analytics has taken over the "advanced analytics," which include predictive analytics and data mining. In this new edition, we therefore use the updated terms.

This R edition includes the material that was recently added in the third edition of the original (XLMiner-based) book:

- Social network analysis
- Text mining
- Ensembles
- Uplift modeling
- Collaborative filtering

Since the appearance of the (XLMiner-based) second edition, the landscape of the courses using the textbook has greatly expanded: whereas initially, the book was used mainly in semester-long elective MBA-level courses, it is now used in a variety of courses in Business Analytics degrees and certificate programs, ranging from undergraduate programs, to post-graduate and executive education programs. Courses in such programs also vary in their duration and coverage. In many cases, this textbook is used across multiple courses. The book is designed to continue supporting the general "Predictive Analytics" or "Data Mining" course as well as supporting a set of courses in dedicated business analytics programs.

A general "Business Analytics," "Predictive Analytics," or "Data Mining" course, common in MBA and undergraduate programs as a one-semester elective, would cover Parts I–III, and choose a subset of methods from Parts IV and V. Instructors can choose to use cases as team assignments, class discussions, or projects. For a two-semester course, Part VI might be considered, and we recommend introducing the new Part VII (Data Analytics).

For a set of courses in a dedicated business analytics program, here are a few courses that have been using our book:

Predictive Analytics: Supervised Learning In a dedicated Business Analytics program, the topic of Predictive Analytics is typically instructed across a set of courses. The first course would cover Parts I–IV and instructors typically choose a subset of methods from Part IV according to the course length. We recommend including the new Chapter 13 in such a course, as well as the new "Part VII: Data Analytics."

Predictive Analytics: Unsupervised Learning This course introduces data exploration and visualization, dimension reduction, mining relationships, and clustering (Parts III and V). If this course follows the Predictive Analytics: Supervised Learning course, then it is useful to examine examples and approaches that integrate unsupervised and supervised learning, such as the new part on "Data Analytics."

Forecasting Analytics A dedicated course on time series forecasting would rely on Part VI.

Advanced Analytics A course that integrates the learnings from Predictive Analytics (supervised and unsupervised learning). Such a course can focus on Part VII: Data Analytics, where social network analytics and text mining are introduced. Some instructors choose to use the Cases (Chapter 21) in such a course.

In all courses, we strongly recommend including a project component, where data are either collected by students according to their interest or provided by the instructor (e.g., from the many data mining competition datasets available). From our experience and other instructors' experience, such projects enhance the learning and provide students with an excellent opportunity to understand the strengths of data mining and the challenges that arise in the process.

1

Introduction

1.1 WHAT IS BUSINESS ANALYTICS?

Business Analytics (BA) is the practice and art of bringing quantitative data to bear on decision-making. The term means different things to different organizations.

Consider the role of analytics in helping newspapers survive the transition to a digital world. One tabloid newspaper with a working-class readership in Britain had launched a web version of the paper, and did tests on its home page to determine which images produced more hits: cats, dogs, or monkeys. This simple application, for this company, was considered analytics. By contrast, the *Washington Post* has a highly influential audience that is of interest to big defense contractors: it is perhaps the only newspaper where you routinely see advertisements for aircraft carriers. In the digital environment, the *Post* can track readers by time of day, location, and user subscription information. In this fashion, the display of the aircraft carrier advertisement in the online paper may be focused on a very small group of individuals—say, the members of the House and Senate Armed Services Committees who will be voting on the Pentagon's budget.

Business Analytics, or more generically, *analytics*, include a range of data analysis methods. Many powerful applications involve little more than counting, rule-checking, and basic arithmetic. For some organizations, this is what is meant by analytics.

The next level of business analytics, now termed *Business Intelligence* (BI), refers to data visualization and reporting for understanding "what happened and what is happening." This is done by use of charts, tables, and dashboards to display, examine, and explore data. BI, which earlier consisted mainly of generating static reports, has evolved into more user-friendly and effective tools and practices, such as creating interactive dashboards that allow the user not only to access real-time data, but also to directly interact with it. Effective dashboards are those that tie directly into company data, and give managers a tool to quickly see what might not readily be apparent in a large complex database. One such tool for industrial operations managers displays customer orders in a single two-dimensional display, using color and bubble size as added variables, showing customer name, type of product, size of order, and length of time to produce.

Business Analytics now typically includes BI as well as sophisticated data analysis methods, such as statistical models and data mining algorithms used for exploring data, quantifying and explaining relationships between measurements, and predicting new records. Methods like regression models are used to describe and quantify "on average" relationships (e.g., between advertising and sales), to predict new records (e.g., whether a new patient will react positively to a medication), and to forecast future values (e.g., next week's web traffic).

Readers familiar with earlier editions of this book may have noticed that the book title has changed from *Data Mining for Business Intelligence* to *Data Mining for Business Analytics* in this edition. The change reflects the more recent term BA, which overtook the earlier term BI to denote advanced analytics. Today, BI is used to refer to data visualization and reporting.

WHO USES PREDICTIVE ANALYTICS?

The widespread adoption of predictive analytics, coupled with the accelerating availability of data, has increased organizations' capabilities throughout the economy. A few examples:

Credit scoring: One long-established use of predictive modeling techniques for business prediction is credit scoring. A credit score is not some arbitrary judgment of credit-worthiness; it is based mainly on a predictive model that uses prior data to predict repayment behavior.

Future purchases: A more recent (and controversial) example is Target's use of predictive modeling to classify sales prospects as "pregnant" or "not-pregnant." Those classified as pregnant could then be sent sales promotions at an early stage of pregnancy, giving Target a head start on a significant purchase stream.

Tax evasion: The US Internal Revenue Service found it was 25 times more likely to find tax evasion when enforcement activity was based on predictive models, allowing agents to focus on the most-likely tax cheats (Siegel, 2013).

The Business Analytics toolkit also includes statistical experiments, the most common of which is known to marketers as A–B testing. These are often used for pricing decisions:

- Orbitz, the travel site, found that it could price hotel options higher for Mac users than Windows users.
- Staples online store found it could charge more for staplers if a customer lived far from a Staples store.

Beware the organizational setting where analytics is a solution in search of a problem: A manager, knowing that business analytics and data mining are hot areas, decides that her organization must deploy them too, to capture that hidden value that must be lurking somewhere. Successful use of analytics and data mining requires both an understanding of the business context where value is to be captured, and an understanding of exactly what the data mining methods do.

1.2 What Is Data Mining?

In this book, data mining refers to business analytics methods that go beyond counts, descriptive techniques, reporting, and methods based on business rules. While we do introduce data visualization, which is commonly the first step into more advanced analytics, the book focuses mostly on the more advanced data analytics tools. Specifically, it includes statistical and machine-learning methods that inform decision-making, often in an automated fashion. Prediction is typically an important component, often at the individual level. Rather than "what is the relationship between advertising and sales," we might be interested in "what specific advertisement, or recommended product, should be shown to a given online shopper at this moment?" Or we might be interested in clustering customers into different "personas" that receive different marketing treatment, then assigning each new prospect to one of these personas.

The era of Big Data has accelerated the use of data mining. Data mining methods, with their power and automaticity, have the ability to cope with huge amounts of data and extract value.

1.3 Data Mining and Related Terms

The field of analytics is growing rapidly, both in terms of the breadth of applications, and in terms of the number of organizations using advanced analytics. As a result, there is considerable overlap and inconsistency of definitions.

The term *data mining* itself means different things to different people. To the general public, it may have a general, somewhat hazy and pejorative meaning of digging through vast stores of (often personal) data in search of something interesting. One major consulting firm has a "data mining department," but its responsibilities are in the area of studying and graphing past data in search of general trends. And, to confuse matters, their more advanced predictive models are the responsibility of an "advanced analytics department." Other terms that organizations use are *predictive analytics*, *predictive modeling*, and *machine learning.*

Data mining stands at the confluence of the fields of statistics and machine learning (also known as *artificial intelligence*). A variety of techniques for exploring data and building models have been around for a long time in the world of statistics: linear regression, logistic regression, discriminant analysis, and principal components analysis, for example. But the core tenets of classical statistics—computing is difficult and data are scarce—do not apply in data mining applications where both data and computing power are plentiful.

This gives rise to Daryl Pregibon's description of data mining as "statistics at scale and speed" (Pregibon, 1999). Another major difference between the fields of statistics and machine learning is the focus in statistics on inference from a sample to the population regarding an "average effect"—for example, "a $1 price increase will reduce average demand by 2 boxes." In contrast, the focus in machine learning is on predicting individual records—"the predicted demand for person i given a $1 price increase is 1 box, while for person j it is 3 boxes." The emphasis that classical statistics places on inference (determining whether a pattern or interesting result might have happened by chance in our sample) is absent from data mining.

In comparison to statistics, data mining deals with large datasets in an open-ended fashion, making it impossible to put the strict limits around the question being addressed that inference would require. As a result, the general approach to data mining is vulnerable to the danger of *overfitting*, where a model is fit so closely to the available sample of data that it describes not merely structural characteristics of the data, but random peculiarities as well. In engineering terms, the model is fitting the noise, not just the signal.

In this book, we use the term *machine learning* to refer to algorithms that learn directly from data, especially local patterns, often in layered or iterative fashion. In contrast, we use *statistical models* to refer to methods that apply global structure to the data. A simple example is a linear regression model (statistical) vs. a k-nearest-neighbors algorithm (machine learning). A given record would be treated by linear regression in accord with an overall linear equation that applies to *all* the records. In k-nearest neighbors, that record would be classified in accord with the values of a small number of nearby records.

Lastly, many practitioners, particularly those from the IT and computer science communities, use the term *machine learning* to refer to all the methods discussed in this book.

1.4 Big Data

Data mining and Big Data go hand in hand. *Big Data* is a relative term—data today are big by reference to the past, and to the methods and devices available to deal with them. The challenge Big Data presents is often characterized by the four V's—volume, velocity, variety, and veracity. *Volume* refers to the amount of data. *Velocity* refers to the flow rate—the speed at which it is being generated and changed. *Variety* refers to the different types of data being generated (currency, dates, numbers, text, etc.). *Veracity* refers to the fact that data is being generated by organic distributed processes (e.g., millions of people signing up for services or free downloads) and not subject to the controls or quality checks that apply to data collected for a study.

Most large organizations face both the challenge and the opportunity of Big Data because most routine data processes now generate data that can be stored and, possibly, analyzed. The scale can be visualized by comparing the data in a traditional statistical analysis (say, 15 variables and 5000 records) to the Walmart database. If you consider the traditional statistical study to be the size of a period at the end of a sentence, then the Walmart database is the size of a football field. And that probably does not include other data associated with Walmart—social media data, for example, which comes in the form of unstructured text.

If the analytical challenge is substantial, so can be the reward:

- OKCupid, the online dating site, uses statistical models with their data to predict what forms of message content are most likely to produce a response.
- Telenor, a Norwegian mobile phone service company, was able to reduce subscriber turnover 37% by using models to predict which customers were most likely to leave, and then lavishing attention on them.
- Allstate, the insurance company, tripled the accuracy of predicting injury liability in auto claims by incorporating more information about vehicle type.

The above examples are from Eric Siegel's book *Predictive Analytics* (2013, Wiley).

Some extremely valuable tasks were not even feasible before the era of Big Data. Consider web searches, the technology on which Google was built. In early days, a search for "Ricky Ricardo Little Red Riding Hood" would have yielded various links to the *I Love Lucy* TV show, other links to Ricardo's career as a band leader, and links to the children's story of Little Red Riding Hood. Only once the Google database had accumulated sufficient data (including records of what users clicked on) would the search yield, in the top position, links to the specific *I Love Lucy* episode in which Ricky enacts, in a comic mixture of Spanish and English, Little Red Riding Hood for his infant son.

1.5 Data Science

The ubiquity, size, value, and importance of Big Data has given rise to a new profession: the *data scientist*. *Data science* is a mix of skills in the areas of statistics, machine learning, math, programming, business, and IT. The term itself is thus broader than the other concepts we discussed above, and it is a rare individual who combines deep skills in all the constituent areas. In their book *Analyzing* *the Analyzers* (Harris et al., 2013), the authors describe the skill sets of most data scientists as resembling a 'T'—deep in one area (the vertical bar of the T), and shallower in other areas (the top of the T).

At a large data science conference session (Strata+Hadoop World, October 2014), most attendees felt that programming was an essential skill, though there was a sizable minority who felt otherwise. And, although Big Data is the motivating power behind the growth of data science, most data scientists do not actually spend most of their time working with terabyte-size or larger data.

Data of the terabyte or larger size would be involved at the deployment stage of a model. There are manifold challenges at that stage, most of them IT and programming issues related to data-handling and tying together different components of a system. Much work must precede that phase. It is that earlier piloting and prototyping phase on which this book focuses—developing the statistical and machine learning models that will eventually be plugged into a deployed system. What methods do you use with what sorts of data and problems? How do the methods work? What are their requirements, their strengths, their weaknesses? How do you assess their performance?

1.6 Why Are There So Many Different Methods?

As can be seen in this book or any other resource on data mining, there are many different methods for prediction and classification. You might ask yourself why they coexist, and whether some are better than others. The answer is that each method has advantages and disadvantages. The usefulness of a method can depend on factors such as the size of the dataset, the types of patterns that exist in the data, whether the data meet some underlying assumptions of the method, how noisy the data are, and the particular goal of the analysis. A small illustration is shown in Figure 1.1, where the goal is to find a combination of *household income level* and *household lot size* that separates buyers (solid circles) from

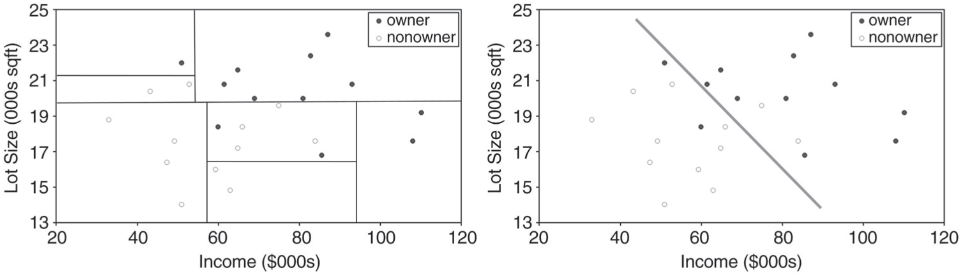

FIGURE 1.1 TWO METHODS FOR SEPARATING OWNERS FROM NONOWNERS

nonbuyers (hollow circles) of riding mowers. The first method (left panel) looks only for horizontal and vertical lines to separate buyers from nonbuyers, whereas the second method (right panel) looks for a single diagonal line.

Different methods can lead to different results, and their performance can vary. It is therefore customary in data mining to apply several different methods and select the one that appears most useful for the goal at hand.

1.7 Terminology and Notation

Because of the hybrid parentry of data mining, its practitioners often use multiple terms to refer to the same thing. For example, in the machine learning (artificial intelligence) field, the variable being predicted is the output variable or target variable. To a statistician, it is the dependent variable or the response. Here is a summary of terms used:

Algorithm A specific procedure used to implement a particular data mining technique: classification tree, discriminant analysis, and the like.

Attribute see **Predictor**.

Case see **Observation**.

Confidence A performance measure in association rules of the type "IF A and B are purchased, THEN C is also purchased." Confidence is the conditional probability that C will be purchased IF A and B are purchased.

Confidence also has a broader meaning in statistics (*confidence interval*), concerning the degree of error in an estimate that results from selecting one sample as opposed to another.

Dependent Variable see **Response**.

Estimation see **Prediction**.

Feature see **Predictor**.

Holdout Data (or **holdout set**) A sample of data not used in fitting a model, but instead used to assess the performance of that model. This book uses the terms *validation set* and *test set* instead of *holdout set*.

Input Variable see **Predictor**.

Model An algorithm as applied to a dataset, complete with its settings (many of the algorithms have parameters that the user can adjust).

Observation The unit of analysis on which the measurements are taken (a customer, a transaction, etc.), also called *instance*, *sample*, *example*, *case*, *record*, *pattern*, or *row*. In spreadsheets, each row typically represents a record; each column, a variable. Note that the use of the term "sample" here is different from its usual meaning in statistics, where it refers to a collection of observations.

Outcome Variable see **Response**.

Output Variable see **Response**.

***P* (*A* | *B*)** The conditional probability of event A occurring given that event B has occurred, read as "the probability that A will occur given that B has occurred."

Prediction The prediction of the numerical value of a continuous output variable; also called *estimation*.

Predictor A variable, usually denoted by X, used as an input into a predictive model, also called a *feature*, *input variable*, *independent variable*, or from a database perspective, a *field*.

Profile A set of measurements on an observation (e.g., the height, weight, and age of a person).

Record see **Observation**.

Response A variable, usually denoted by Y, which is the variable being predicted in supervised learning, also called *dependent variable*, *output variable*, *target variable*, or *outcome variable*.

Sample In the statistical community, "sample" means a collection of observations. In the machine learning community, "sample" means a single observation.

Score A predicted value or class. *Scoring new data* means using a model developed with training data to predict output values in new data.

Success Class The class of interest in a binary outcome (e.g., *purchasers* in the outcome *purchase/no purchase*).

Supervised Learning The process of providing an algorithm (logistic regression, regression tree, etc.) with records in which an output variable of interest is known and the algorithm "learns" how to predict this value with new records where the output is unknown.

Target see **Response**.

Test Data (or **test set**) The portion of the data used only at the end of the model building and selection process to assess how well the final model might perform on new data.

Training Data (or **training set**) The portion of the data used to fit a model.

Unsupervised Learning An analysis in which one attempts to learn patterns in the data other than predicting an output value of interest.

Validation Data (or **validation set**) The portion of the data used to assess how well the model fits, to adjust models, and to select the best model from among those that have been tried.

Variable Any measurement on the records, including both the input (X) variables and the output (Y) variable.

1.8 Road Maps to This Book

The book covers many of the widely used predictive and classification methods as well as other data mining tools. Figure 1.2 outlines data mining from a process perspective and where the topics in this book fit in. Chapter numbers are indicated beside the topic. Table 1.1 provides a different perspective: it organizes data mining procedures according to the type and structure of the data.

Order of Topics

The book is divided into five parts: Part I (Chapters 1–2) gives a general overview of data mining and its components. Part II (Chapters 3–4) focuses on the early stages of data exploration and dimension reduction.

Part III (Chapter 5) discusses performance evaluation. Although it contains only one chapter, we discuss a variety of topics, from predictive performance metrics to misclassification costs. The principles covered in this part are crucial for the proper evaluation and comparison of supervised learning methods.

Part IV includes eight chapters (Chapters 6–13), covering a variety of popular supervised learning methods (for classification and/or prediction). Within this part, the topics are generally organized according to the level of sophistication of the algorithms, their popularity, and ease of understanding. The final chapter introduces ensembles and combinations of methods.

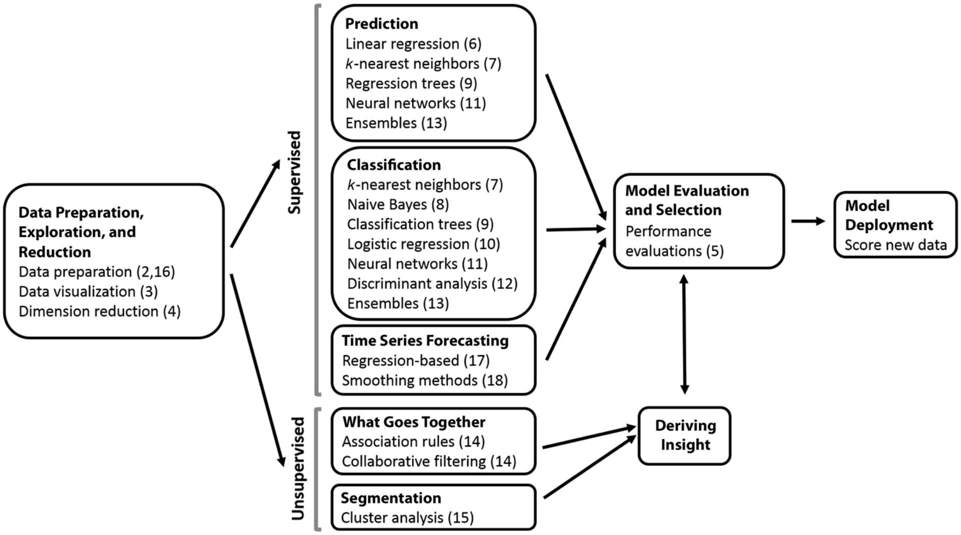

FIGURE 1.2 DATA MINING FROM A PROCESS PERSPECTIVE. NUMBERS IN PARENTHESES INDICATE CHAPTER NUMBERS