PREDICTIVE ANALYTICS AND MACHINE LEARNING FOR MANAGERS

Business Question
Data Work
Model: Methods, Specification, Assumptions
Analysis, Evaluation, Model Selection
Reporting, Interpretation, Storytelling, Insights

In Press - Please do not cite, copy or share without permission of the author

J. ALBERTO ESPINOSA

“Throughout my professional experience building data science practices for the Federal Government and commercial industries, I have realized the need for better and faster data-driven decision-making capabilities. Today, managers and business leaders need to become data literate, understand the power of analytics, and develop their analytical skills. In this book, Professor Espinosa provides a robust analytic roadmap for business leaders, providing practitioners with a better understanding of how advanced data analytics can improve their businesses. The first half of this book is an excellent compendium of statistical concepts to help professionals understand and implement advanced, complex analytical models. The second half of the book explains what business leaders are asking today—how to make better business decisions using data and machine learning algorithms to create business value. The scripts and code presented in this book will enable managers to understand and experiment with various predictive analytics and machine learning methods.”

ROD FONTECILLA
Partner and Chief Innovation Officer
Technology Solutions
Guidehouse

“Professor Espinosa’s book is a must-read for analysts and managers interested in learning how to use data analytics for decision-making and business problem-solving. Through this book, one learns how to frame a business analytics question, how to identify the right predictors and model, and how to interpret results. Professor Espinosa has been teaching predictive analytics for a decade, and his book has a good balance of technical and managerial insights. He does a wonderful job of explaining fundamental terms in a concise and understandable manner. Further, his deep experience is exemplified in the book, which will help business professionals and analysts understand the analytics lifecycle from a managerial perspective. The accompanying GitHub site appendices provide useful scripts and examples illustrated in the book, which will enhance the learning of the technical aspects presented. This book is a comprehensive and valuable guide for analysts and managers.”

WAI FONG BOH

President’s Chair and Professor of Information Systems
Deputy Dean of Nanyang Business School
Nanyang Technological University in Singapore

Title: Predictive Analytics and Machine Learning for Managers
First Edition: February 2023
ISBN paperback: 979-8-9876543-1-6
ISBN ebook: 979-8-9876543-0-9
Published by: Jibe4Fun Press
Published in the United States of America

Cover & Book Design: Alison Rayner

Trademarks
All brand names and product names referred to in this book are registered trademarks and unregistered trade names of their owners. There is no implied endorsement of any of them.

Disclaimers
This publication aims to provide accurate and reliable information regarding the subject matter covered. However, neither the publisher nor the author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

PREDICTIVE ANALYTICS AND MACHINE LEARNING FOR MANAGERS

J. Alberto Espinosa, Ph.D.

○○○

Dr. Espinosa is a Professor of Information Technology and Analytics (IT&A) at the Kogod School of Business at American University, Washington, DC. He holds Ph.D. and Master of Science degrees in Information Systems from the Tepper School of Business at Carnegie Mellon University, an MBA from Texas Tech University, and a Mechanical Engineering degree from Pontificia Universidad Catòlica del Peru. He is the architect of Kogod's MS Analytics program (for both campus and online delivery) and of the undergraduate programs in Information Technology and Business Analytics. In addition to this book, he has co-authored two books, *I'm Working While They're Sleeping: Time Zone Separation Challenges and Solutions* (https://www.amazon.com/dp/0983992509) and *Obtaining Value from Big Data for Service Systems: Volume I: Big Data Management* (https://www.amazon.com/dp/B07SG16RJT) and *Volume II: Big Data Technology* (https://www.amazon.com/dp/B07SKY6QMY). His research focusses on coordination and performance in technical projects across global boundaries, particularly distance and time-separation (i.e., time zones and schedule shifts). His current research focus is on the visual and quantitative representation and analysis of team knowledge using social network analytics. Dr. Espinosa is a multi-method researcher, but most of his work involves field studies with technical organizations, using quantitative methods. His work has been published in leading scholarly journals, including *Management Science*, *Organization Science*, *Information Systems Research*, *The Journal of Management Information Systems*, *IEEE Transactions on Software Engineering*, *IEEE Transactions on Engineering Management*, *Communications of the ACM*, *Human Factors*, *Information, Technology and People*, and *Software Process: Improvement and Practice*. Dr. Espinosa's work has also been presented and featured at leading academic conferences. He teaches predictive analytics, social and organizational network analytics, R programming for analytics, information technology foundations and business process analysis, and programming for business applications. He also has several years of working experience, first as a design engineer for oil and mining projects, and later as a senior manager, VP, and CFO with international organizations directly supporting, supervising, and formulating policy for finance, human resources, global IT, and data management and analytics applications to support geographically distributed work in Africa, Latin America, and Eastern Europe.

○○○

From The Author

I was motivated to write this book after architecting the Kogod School of Business' MS program and undergraduate specialization in Business Analytics with my colleagues, and teaching and practicing analytics for over a decade. My goal was to provide strong but understandable conceptual foundations and practical material for graduate students and managers, describing how to frame a business question, identify various model specification (i.e., feature engineering) and model methods (explainable and black box), select the optimal model based on the bias, variance, and cross-validation testing, and interpret results with meaningful storytelling for clients and managers. The book contains two components: (1) the main text with two sections—one with conceptual, mathematical, and managerial foundations, the other about advanced predictive modeling methods based on machine learning; and (2) an appendix companion with annotated R Markdown code with hands-on applications, posted in GitHub.

○○○

This book is dedicated to my wife, Delphine Clegg,
who has supported me on this book project
and in life for many years.

This book was edited by Andrew Erickson and Delphine Clegg.

Andrew was one of my top students and an awesome teaching assistant for my Predictive Analytics course. He has also been a writer for *American*, the American University magazine, and is now a business analytics professional. Andrew reviewed the book for effective communication, comprehension, clarity, and overall quality of the material.

Delphine is a freelance editor with years of experience in editorial and communications work. She was the final editor of the book. She reviewed all the writing in detail and did an outstanding job ensuring consistency of content and style.

Alison Rayner designed this book. She laid out and presented the book's content masterfully, including the front and back covers, for digital and print formats.

TABLE OF CONTENTS

○ ○ ○

OVERVIEW

Note About the R Code Companion for This Book: Appendices A1 to A11 provide the respective R scripts, code and programming notes associated with each chapter and are available at https://github.com/jibe4fun/paml4m/tree/R-Code. The book is devoted to conceptual issues and the appendices cover hands-on modeling with R. I plan to include appendices with scripts and code using the Python language at some point in the future.

> *An analyst refused to report his/her partner's stolen credit card. The analyst told a friend in confidence that the refusal was because a predictive model showed with statistical confidence that the thief spent less money than the partner.*

This humorous story illustrates an important point about this book, MD^3, which stands for Models Don't Make Decisions, Managers Do. A model can tell us what the best quantitative solution may be, but it is not necessarily what a rational human would decide. Nevertheless, some predictive models based on machine learning are written to automate such decisions (e.g., recommender systems, self-driving vehicles). It is important, therefore, to understand the goals of the particular predictive model—interpretability, inference, and/or predictive accuracy.

There have been many articles predicting large shortages of professionals with deep data science skills. But the predicted shortages are of a magnitude larger for managers with analytical skills. Analytics and data science today permeate every aspect of business and organizational work and managers are expected to know how to do basic analytic work and how to be good consumers of analytics reports. As such, the goal of this book is to provide knowledge and skills for the analytical manager. I place a strong emphasis on the conceptual foundations of predictive analytics and machine learning. My intent is to provide the analytical manager with the necessary knowledge to: specify business questions of interest and translate them into equivalent analytics questions; define the analytics goals for a project; select and specify the appropriate model to answer the business questions and fulfill the analytics goals; build the model using R open source software; test competing models or tune model parameters using machine learning and cross-validation methods; and interpret results and extract meaning from the data to tell the *business story* behind the analysis.

I have been teaching predictive analytics and machine learning at a business school for several years, and I have been applying these methods in my own research for about two decades. When looking for an appropriate textbook for my course, I read about a dozen books on predictive analytics and machine learning. Most of these books had this in common: (1) they start with very basic and simple material; (2) at some point the material turns very technical and cryptic for the average beginner to follow; (3) they focus on describing predictive models, rather than on understanding how to select the appropriate modeling method and model specification; and (4) they tend to cover data science and statistical aspects of these models, rather than their business application and interpretation. It is difficult to find a book on predictive analytics and machine learning that is both readable for managers and technically deep. Most books are one or the other. This book attempts to fill these gaps. And while a technical background is not necessary in order to understand the content of this book, some basic understanding of statistics and software programming may be helpful. I recommend that readers brush up on basic concepts like frequency distributions, descriptive statistics, correlation analysis, and linear regression, and also learn the basics of statistical programming languages like R. Again, I cover these topics in detail so they are not prerequisites to reading this book, but a basic understanding of these topics will facilitate your understanding of the material.

We often hear politicians and people in the media people saying things like: "we have to follow the data;" "decisions should be based on data;" "models are only as good as their assumptions;" etc. What do they really mean by "following the data"? Data is usually full of anomalies, imperfections, and missing elements, so simply following the data will not always provide an answer to our questions. What does it mean for a model to be correct or incorrect? Models can be tested for statistical fit and accuracy, and nothing is ever certain. All models have levels of statistical confidence and accuracy, but no model can be correct 100% of the time and there is never a guarantee that any model will predict accurately with new data. A perfect example of all this is with weather forecasting in which the data changes continually, so predictions become more and more uncertain when the time horizon is long into the future. Any weather or pandemic prediction expert will agree that as the time horizon for predictions increases, the predictive accuracy of the models diminishes substantially. Another way of saying this is that the confidence in our prediction diminishes sharply. One way to illustrate this notion is to pay attention to hurricane forecasting models. Meteorologists usually show predicted hurricane paths from multiple predictive models (e.g., Global Forecasting System, European Model, Canadian Model, etc.), color coded for TV viewing. When the models agree, there is some degree of certainty as to where the hurricane will make landfall. When models disagree, landfall predictions become very uncertain (i.e., the predictive accuracy confidence goes down). Predictive models for business are no different. For any analytics questions, there will be a wide range of modeling methods, data transformations and model specifications to choose from, plus various ensemble models that aggregate the results from various models. The main goal of this book is to help the reader navigate through the various modeling options and provide one with the ability to test them and compare them in order to select the optimal model and specification to meet the analytics goals.

In this book I begin by discussing general principles and statistical and linear regression concepts, which then become the building blocks to develop more advanced and complex models. To this end, the book is divided into two main sections: **Section 1 (Chapters 1-5) – Predictive Analytics Basics**; and **Section 2 (Chapters 6-11) – Advanced Models and Machine Learning**.

In **Section 1**, **Chapter 1**, I introduce predictive analytics and machine learning from a business perspective. I discuss key foundational aspects of predictive modeling, such as the analytics life cycle, general model categories (i.e., quantitative vs. classification), classic modeling tensions (e.g., bias vs. variance, explainability vs. accuracy, etc.). In **Chapter 2**, I provide an overview of the basic statistical foundations necessary to follow the rest of the book. Most of this material is based on descriptive analytics, used to understand the data before building any predictive models. In **Chapter 3**, I introduce the most basic quantitative models (i.e., regression and regression trees) and classification models (i.e., binomial logistic regression and classification trees). In addition, I discuss the basic assumptions or preconditions for these models, which is important for two reasons—models are built on mathematical assumptions, which need to be tested before using the model; and regardless of the testing results, these models tend to be the most unbiased, thus serving as useful benchmarks to evaluate other models. Most advanced models are departures or derivatives of these basic models, so it really helps to understand them and their respective assumptions well before using them. In **Chapter 4**, I discuss the importance of data pre-processing. It is estimated that about 80% of the work in an analytics project is extracting and preparing the data for analysis. Some aspects of data pre-processing are necessary (e.g., curation, missing data, cleansing anomalies, etc.) and some are done to improve model fit and accuracy (e.g., transformations, interactions, sub-grouping, etc.). The final chapter of **Section 1**, **Chapter 5**, is where I discuss variable selection. Selecting the predictors for a model is one of the most important initial steps when building predictive models. Predictors must be rooted in an understanding of the business domain of the analytics problem. You cannot really undertake the task of healthcare or marketing analytics without understanding the healthcare and marketing domains. At the same time, predictors must have a solid statistical foundation to be incorporated and retained in models. At the end of this chapter, readers should have a basic understanding of the fundamental principles of predictive analytics and machine learning. The combination of data pre-processing (chapter 4) and variable selection (chapter 5) is often referred to as *feature engineering*.

In **Section 2**, I discuss more advanced predictive modeling methods. We switch into high gear when I introduce the concepts of machine learning (ML) and cross-validation (CV) in **Chapter 6**. ML is about training models with data and testing the models for predictive accuracy. As more data comes in, the algorithms learn without the need to be reprogrammed. Because accuracy is critical to model method and specification selection, and because this accuracy will change as new data arrives, CV is a central aspect of machine learning. CV is used for many things, including evaluating single models, tuning models, comparing models and training models. One important predictive modeling tension I discuss in depth is bias vs. variance. Bias is about the effects reported by a model departing from the true effects, which is not good for interpretability. Variance refers to whether we get consistent results when we compare model results with multiple resamples of the data. Stable (low variance) models will yield consistent results across resamples. Conversely, unstable (high variance) will yield widely dissimilar results across resamples, making the model unreliable. Bias and variance represent one of the most fundamental tradeoffs in predictive modeling. Smaller and simpler models tend to have more bias—generally caused by omitted predictors and less variance, whereas larger and more complex models have less bias, but this comes at the expense of increased variance. CV testing helps find the optimal model size and complexity that minimizes the combined effect of bias and variance.

The topic of dimensionality is covered in **Chapter 7**. Complex business problems often require complex predictive models. As the number of predictors grows large and as the model becomes more complex, the issue of dimensionality will manifest itself in the form of increased variance. But there are effective ways to address this problem, which are covered in this chapter. In **Chapter 8**, I discuss non-linearity, which is when the predictors have a non-linear relationship (e.g., quadratic, cubic, interactive) with the outcome variable. Classification methods are discussed in more depth in **Chapter 9**. We depart from the binomial classification model (e.g., yes or no, approve or decline, positive or negative diagnostic, etc.) and look at multinomial classification models (i.e., more than one categorical outcome, e.g., green, amber, or red traffic light recognition). I also discuss classification accuracy evaluation methods based on key concepts like the confusion matrix and ROC curves. In **Chapter 10**, I discuss decision trees, both quantitative and classification, in more depth. This chapter also covers more advanced tree methods like bootstrap aggregation, random forest, and boosted trees. Finally, **Chapter 11** provides an introduction to deep learning predictive models, with a focus on neural networks.

SECTION 1

PREDICTIVE ANALYTICS BASICS

CHAPTER 1

INTRODUCTION TO PREDICTIVE ANALYTICS

○○○

Introduction to Predictive Analytics

What do people mean when they say that we have to "follow the science"? While I agree that decisions must be informed by science, science is never exact. Science is grounded in data and analytical models, both of which are often imperfect. There are good data, bad data, raw data, data with missing or inconsistent values, etc. There are also issues with sample size, sampling methods, measurement, probability distributions, collinearity, etc. Furthermore, often data needs to be pre-processed before it can be used, and bad pre-processing can lead to bad data. Models also have issues, which is why they are tested for compliance to their method assumptions, validated for accuracy and evaluated against competing models. In addition, effects reported by models are influenced by issues like confidence intervals, statistical significance, likelihood of predicting correctly, margins of error, etc. Following the science means understanding all these complex data and modeling issues and rendering optimal models that can help us interpret the effects on outcomes and make sound predictions. But when models do not yield these desirable results, we cannot really blame the models, but the people who built them. So, while data and modeling are essential to decision-making, it is also important that we understand the scope, generalizability, and limitations of these models, regardless of whether we are the analysts who build these models or the managers who consume the resulting analytics reports. Most analysts are well-intentioned and intelligent individuals, but predictive modeling is not trivial. Models are subject to algorithmic issues, model mathematical assumptions (i.e., conditions under which the model can be used), specification (i.e., which predictors to include and in what form), modeling method (e.g., quantitative, classification, regression based, tree based, etc.). And then all these things need to be applicable to the specific business question or model being analyzed. So, in sum, it is not about following the science, but about understanding the math, data, and methods in order to be able to provide answers to specific business questions with some degree of confidence. Figure 1.1 illustrates the various aspects that interrelate to form an appropriate predictive model.

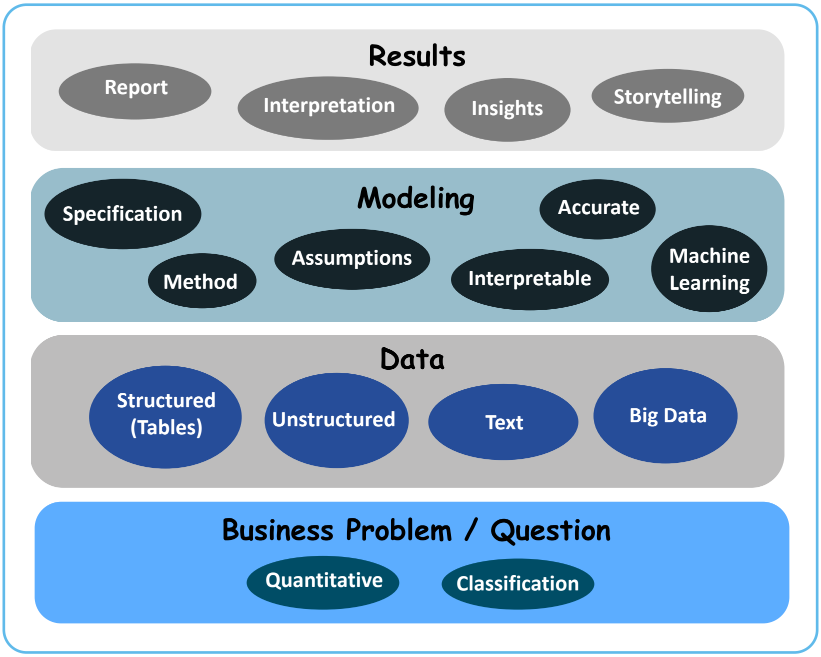


Figure 1.1 Factors Influencing a Predictive Model

“In God we trust, others must provide data.” According to Quote Investigator (https://quoteinvestigator.com/2017/12/29/god-data/), the first known use of this quote was by a professor of pathology in 1978 who stated in a congressional hearing that he needed good scientific data before he could provide an opinion about whether smoking was hazardous to non-smokers. In this age of big data and analytics, decisions backed by data rather than by intuition or opinion is now the norm. Providing data is now a necessary component of a business professional’s job. These days when you meet with a client or manager, you can no longer use your expertise alone to make a convincing case. Your audience will demand to see evidence from data to back up your claims. Thus, it is important for all business professionals to be able to understand and process data to support their arguments.

Ronald Coase, a British economist and Nobel Prize winner, once said in the 1970s, “If you torture the data, it will confess to anything.” Essentially this means that skilled statisticians have the ability to manipulate statistics to support their conclusions. Others have argued that the actual quote was “if you torture the data long enough, nature will confess,” meaning that if you manipulate the data for too long, the truth will eventually emerge. In either case, we need to be mindful of being honest in our predictive modeling and not torture the data to fit a story. The truth will eventually come out. And, by being savvy about data analytics, one can probably figure out when others are torturing the data.

After reading this the book you should be able to look through the data—along with some descriptive statistics, plots, and correlation data—of an analytics report and evaluate the soundness of the models used. Following your review, you should be able to assess whether someone is manipulating the results and be able to determine whether or not the results reported are accurate.

When properly executed, predictive analytics can be very powerful. Take for example one of my favorite books, *Moneyball* by Michael Lewis. Michael Lewis is a well-known financial analyst turned book author. He has written best sellers like *The Big Short* and others. Published in 2003, *Moneyball* is chock-full of interesting details and stories about how baseball statistics were used by the Oakland A’s to enhance on-field performance and cost-effectiveness. For example, they would sell a very expensive player worth millions of dollars and acquire a few rookie players, who individually were not as good as the player they just sold, but collectively those players had an aggregation of skills that were necessary for the team to replace that player. This is one of the first books to document the use of statistical analysis in sports management, and it sparked great interest in understanding big data and analytics.

Hopefully, the present book will not only open your eyes to the power of predictive analytics and machine learning but will also help you get started in this fascinating field. By the time you finish reading the book, you should be able to take a business problem or question and resolve or answer it through analytics. This book is not about mathematics, statistics, or algorithms. It is also not about programming in R or Python. You will learn these things, for sure. But the focus of this book is on learning how to apply the correct modeling method and specification to extract meaning from the data and answer a business question from the data. It is about how to define your business question, how to translate it into an analytics question, how to select the right model, how to evaluate or justify the right model, how to identify the appropriate predictors for the model, how to execute the model's analytics, and how to interpret the results so that you can convey your results to your client or manager.

Before we move on to the next section of this chapter, I want to make an important point, which is that you are not going to learn everything from this book. Predictive analytics and machine learning are very broad and rich fields, which keep changing every day. My goal is to not only provide sound insights into how one becomes a competent analytics professional, but more importantly, how one becomes a good "analytics learner." Once you understand the basic foundations of predictive modeling and learn a few advanced modeling methods presented in this book, you will have acquired sufficient knowledge on this topic, which will enable you to learn other advanced analytics concepts and methods on your own.

1.1 The Importance of Predictive Analytics

Humans are predictive analytics entities. We are always making predictions for everything, often using human learning, very similar to machine learning, to inform our decisions. For example, when a student enrolls in a university analytics program it is because, perhaps, the student is anticipating that this degree will lead to higher salaries, new employment opportunities, promotions, enhanced reputation, etc. This is a form of human predictive analytics, and many times we are not even thinking about it. Business managers are always making predictions and decisions, but of course in a more formal and rigorous manner based on data and tested models rather than on plain intuition and common sense.

However, predictive analytics is subject to manipulation or misinterpretation. For example, it is not uncommon for a politician or person in the media to make simplistic statements based on a single data point, a single predictor, or simple generalizations of anecdotal observations. A single data point is not statistical evidence, and a single variable predictor yields the most biased model you can have. Also, anecdotal observations do not qualify as statistical conclusions. The foregoing would be perfect examples of torturing the data to try to make a point, thus the importance of being knowledgeable and being able to distinguish a fake story from a real one.

Further, it is not uncommon, for example, for a president during a State of the Union address to mention the terrible or excellent experience of a guest sitting in the gallery while trying to make a point through a generalization of that person's situation. Again, one case is one data point, so it cannot constitute statistical confidence. One data point in analytics is useless and has no statistical power whatsoever; for this we need more data points and an appropriate model with appropriate controls.

An example of the importance of gathering more data points and using an appropriate model with appropriate controls can be illustrated when trying to answer the question about whether a gender pay gap exists in an organization. Some people will say, "Oh, yeah, there's a gender pay gap and we must fix this problem." Others will say that none exists and show some statistical analysis to prove it. So, depending on how one tortures the data, either could be proven. For example, if you compare the salary averages between males and females in a particular organization, you may find that males make higher salaries than females and conclude that there is a gender pay gap. If you make the same comparison by profession, perhaps you may still find a gap, but a smaller one because males tend to gravitate to certain professions while females tend to gravitate to certain others. But then if you factor in years of experience, you may find that males generally make higher salaries in an organization because, on average, they have more years of experience, or have been employed by the organization for more years, and therefore earn more. But then if you control for age, education, profession, etc., you may find different results. This example is hypothetical, but it illustrates how easily one can torture the data and manipulate the results. A correct model for this question would be to include gender as the focal predictor in the study and control for (i.e., account for) any other variables that may affect salaries.

As another example, what would be the answer to a question about how unemployment would be affected if the minimum wage was increased by one dollar? Some will believe that this increase in income would be beneficial to employees. Others will believe that the increase in labor cost would cause unemployment to go up. To answer this question appropriately, you would need to have a good dataset representative of most sectors of the economy and model minimum wage as a predictor of unemployment, but you would also need to control for (i.e., account for) any variable that may affect unemployment (e.g., S&P index, inflation, state of the economy, etc.).

Additionally, how might one answer, for example, a question related to healthcare? A patient goes to the doctor to get treatment for symptoms, and after reviewing the patient's vitals and some test results, the doctor diagnoses a particular illness. How does the doctor know how to make this diagnosis? Basically, doctors use a decision process similar to the decision process used in machine learning. Just as machine learning relies on past data to train models and make predictions, doctors retain lots of data in their memory about prior cases and based on the similarities and differences of the patient's vitals, they can make a diagnosis about the patient's illness. And just as machine learning algorithms learn and improve their predictive accuracy from data, a doctor becomes better at diagnosing illnesses from experience with similar cases. However, just like with machine learning, no prediction is exact or correct every time. Thus, the patient may need a second opinion. When using machine learning, we use re-sampling, re-testing, and trying different models to get another "opinion."

There are predictions one can make, for example, in answering a question like: Why are you reading this book? Perhaps it is just curiosity, but most likely you have learned about the importance of analytics, data science, or machine learning for business, and you predict that you will be better off if you learn this material. This is just a prediction, so it is not exact science. But whatever your reason, it will become part of your personal internal data, which you will refer to later to make future decisions. As an educator, I could try to predict which of my students may become CEOs of the companies they work for. This would be a classification prediction, that is, become a CEO (positive outcome) or not (negative outcome). Or I could try to predict your salary five years from now, which would be a quantitative prediction, rather than a classification prediction.

Finally, here is an interesting example of predictive analytics from a movie produced by John Malkovich titled *The Dancer Upstairs*. The movie is partly based on the capture of the leader of the Shining Path terrorist group in Peru, Abimael Guzman, who went into hiding in the mid-1970s and began his armed struggle in 1980. He was living on the second floor of a ballet studio (thus the movie title) when investigators identified a few insurgents in the neighborhood and determined that the output of garbage from this second-floor studio was more than could be attributed to a single person living there. In addition, at a nearby pharmacy, sales of a certain skin cream had surged (Guzman had a skin condition that required a special cream). The analysis of this data led to his capture in 1992.

Like these human prediction examples, businesses are constantly trying to make quantitative and classification predictions. They search for answers to such questions as: Does free WiFi in hotels lead to increased bookings? Does a certain TV ad increase sales? Is a given credit card transaction legitimate? Does a patient have a virus infection? Or, Is the virus infection likely to lead to hospitalization? To answer these questions, you need good data and a well-specified and properly tested model (or many).

1.2 Analytics and Its Cousins

There are so many terms used for this discipline today that they confuse most people, including myself. You may get a degree in data science only to find out that what an employer is looking for in a data scientist is something different. This is not necessarily the educator's fault, but it happens because the field is rich and broad—it consists of many related disciplines that are distinct, although they overlap somewhat. This is why for the purposes of this book I have labeled them "analytics cousins." These disciplines include disciplines like analytics, machine learning, data mining and data science, among others. In the next few sections I attempt to shed some light on the similarities and differences across these disciplines, but be aware that if you ask another educator, he/she may have a different view.

Analytics

Analytics is the scientific process of transforming data into insights for making good business decisions. This is how the Institution for Operations Research and Management Sciences (INFORMS) defines it. INFORMS is the premier association of analytics academics and professionals. Notice that the emphasis in INFORMS' definition is on managerial decision-making, which is what sets it apart from other related fields, or analytics cousins. The whole idea behind analytics is to extract meaning from the data to assist decision makers. Naturally, to do this, we need to tap into the power of mathematical and statistical analysis, data engineering, data mining, machine learning, and other cousins.

INFORMS also defines three main types of analytics—descriptive, predictive, and prescriptive. Predictive analytics, which is the subject of this book, requires some descriptive analytics beforehand, and prescriptive analytics relies on predictive and other decision models. Descriptive analytics is about extracting meaning from the existing data to develop familiarity and better understanding of the data available for the analysis. The idea is to plunge into the data to identify its characteristics and patterns through things like data mining, cluster analysis, descriptive statistics, correlation analysis and statistical distributions, among other things. Descriptive analytics can be either visual (e.g., plots, histograms, QQ-plots, boxplots, etc.) or quantitative.

Predictive Analytics

Predictive analytics is about using some of the data available in the dataset (i.e., predictors) to predict the values of another part of the dataset (i.e., outcomes). Predictive models can be of two varieties depending on the analytics question: quantitative or classification. Quantitative models are ones that predict an outcome quantity (e.g., sales, profits, enrollments, prices, event attendance). Classification models predict the likelihood of an outcome falling into a particular category. Binomial or binary classification models have two possible outcomes (e.g., spam vs. legitimate email, fraudulent vs. legitimate credit card transaction, loan default vs. no-default, etc.). Multinomial classification models predict the likelihood of an outcome falling in one of many categories (e.g., recognizing one of ten digits from handwritten zip codes).

Data Mining

Data mining is the analysis of data with the goal of discovering new, previously unknown patterns from existing relationships among data elements. We use data mining when we do not know anything about the data, and we want to explore it and identify new patterns. In some ways, you can think of data mining as a process of generating hypotheses you can later test with predictive models. What confuses some people is that analytics and data mining may rely on similar analytical methods. Some methods are developed from the analytics tradition and some from the data mining tradition but are often used in both. So, it is the actual purpose of the analysis that differentiates the two. When we know nothing or very little about the patterns in the data and want to uncover these patterns, we apply data mining. When we have a hypothesis about such patterns or have some business intuition about what predictors may affect an outcome, we apply analytics. In sum, data mining is about discovering patterns in the data, and analytics is about testing expected patterns and extracting meaning from the data.

Analytics is also about answering business questions and solving business problems. While you may not have specific hypotheses to test, you must have some idea about certain relationships in the data you want to explore. For example, you may expect that advertising expenditures lead to increased sales. In analytics, you would build a model to test this hypothesis. In data mining, you would not do this. You would just plunge into the data and try to uncover patterns. But again, you may be using similar methods to analytics. For example, you may use cluster analysis to identify possible groupings of customers or deep neural networks to find previously unknown relationships in the data (i.e., data mining). But you could also use both approaches to evaluate a theory or hypothesis (i.e., analytics).

Business Intelligence (BI)

BI is another related analytics' cousin. In analytics, you first define the problem or question to be addressed. You then look at the data and find the best model to answer the question. And then find the best model specification to answer the question. When you are done with the analysis, you interpret your results. A lot of thinking goes into every step of the analytics process, and you make interpretations and decisions along the way. Once you understand how to do the analysis and figure out the most accurate and suitable model method, then you can implement your solution in a business intelligence platform, where you can automate the analysis–quickly and on demand in a BI platform, if you need to do a similar analysis on a routine basis. BI tools (e.g., MicroStrategy, PowerBI, Tableau) rely heavily on descriptive analytics and visual representations of the data, such as charts, plots, and dashboards.

Machine Learning (ML)

ML is a branch of artificial intelligence (AI). ML is about specifying models that use existing data, which is fed into algorithms that will process the data to yield an outcome. In other words, the model is "trained" with the data, and it is continually re-trained as new data arrives. In general, AI emulates human intelligence in a computer.

Logic-based AI is about programming software algorithms that can follow human reasoning when solving a problem. Rule-based AI or expert systems are about recording how experts would behave or decide under numerous what-if situations, recorded in a database called rule base. For example, if you wanted to program a system to estimate the cost of a software project, the traditional way was to write a program to use some characteristics of the software project (e.g., lines of code, complexity, functionality required, data sources needed, etc.) to estimate the effort required to build the software and the respective cost associated with that effort. Alternatively, you could develop an expert system based on rules elicited from software estimation experts in the subject stored in a rule base. The system would query the rule base and extract the decisions that an expert would have made to do the estimation. Predictions improve and become more accurate as you gather more and more data over time. In contrast, ML is AI that learns from the data with minimal programming. Once a model is developed, the effects and predictions of the model will change as the data changes. This has tremendous appeal for two reasons—ML models do not need to be updated as much as other AI algorithms when the data changes, and certain ML models (e.g., image recognition) can be used for multiple problems (e.g., handwritten digit recognition, tumor detection in graphic scans, red light recognition from a vehicle's camera, etc.).

Cross Validation (CV)

CV is a method tightly associated with ML, which aims to test the accuracy of the models. CV is about using a large portion of the data to fit (i.e., train) the model and the remaining part of the data to evaluate (i.e., test) the model. Training a model with part of the data and testing it with a different part of the data is central to ML, because it ensures that dimensionality and overfitting are minimized when the model is tested with data that was not used to train the model. CV is used extensively in ML to test individual models, tune individual models (i.e., by manipulating tuning parameters), compare models and model specifications for accuracy and in some cases, to perform the actual training of the models (e.g., deep learning). For example, neural networks are trained by approximation by assigning random weights to model inputs (i.e., predictors) and adjusting the weights recursively using CV until the optimal weights that yield the most accurate predictions are identified.

Unsupervised Learning

Unsupervised learning is when we apply ML without a specific analytical goal in mind. Unsupervised learning is almost like how a baby learns—he/she simply listens, observes the environment, and develops associations (e.g., cry → get fed). A baby learns to recognize people, learns to speak, etc. without a particular agenda in mind. There is no specific goal given to the baby. The baby will just learn by observation. This is called *unsupervised learning*. The environment provides clues for the learning process. Descriptive analytics and data mining are closely associated with unsupervised learning because they both aim to learn patterns from the data without a specific analytical goal in mind.

Supervised Learning

Supervised learning is about pursuing a learning goal, such as when you decide to go to college. You have an educational goal and decide to enroll in courses to fulfill that goal. That is supervised learning because you want something out of that experience (e.g., a career). The same is true for supervised learning in analytics; we call it predictive modeling. So, ML and predictive modeling, though conceptually distinct, are often used interchangeably. A model that is trained with (i.e., learns from data with the goal of making predictions) is both a supervised ML model and a predictive model. However, ML is more often associated with models that are tested and tuned with CV.

Data Science

Data science is what I consider to be the mother field of all other analytical disciplines. Data science encompasses everything else, but this field is very broad and rich, making it difficult to pin down. There are a lot of data science programs and courses out there with different orientations (e.g., statistical, computational, business). But a serious data scientist needs to have deep knowledge about just about everything having to do with data, including things like software programming, data algorithms, data structures, data engineering, database, data warehousing, ETL (extraction, translation, and loading), data pre-processing, data mining, statistics, etc. In addition, a data scientist needs to have deep knowledge of the core disciplines or business domains (e.g., healthcare, sports, finance, marketing) in which the data science is being conducted. As the name implies, data science also requires an understanding of the concept of scientific analysis.

Figure 1.2 shows these various disciplines and how they overlap with each other in some respects, but not in others. All these concepts have something in common and something unique.

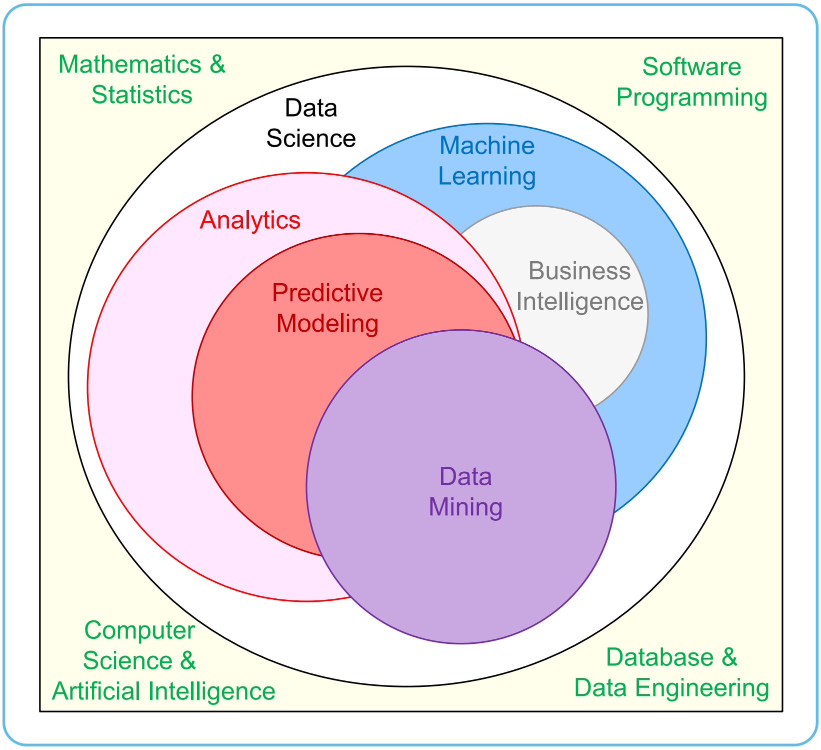


Figure 1.2 Analytics and It's Cousins

1.3 Key Tradeoffs

In many books and tutorials, the emphasis is on teaching how to specify and apply specific models. For example, if you want to train a neural network, you will find plenty of tutorials and sample R or Python courses for you to do that. It is amazing to me that something as complex as a deep neural network can be trained so easily with a few lines of R or Python code. But as with user friendly power tools, they can be quite destructive in the hands of a novice. Again, the main challenge in predictive analytics is not about how to train one model or another, but about selecting the most appropriate modeling method and model specification. This difficulty stems from the fact that in analytics there is rarely a clear winner because there are many tradeoffs involved when evaluating models. Models and specifications that may be good for one thing are not so good for other things. This is what model selection and model tuning is about—balancing these tradeoffs and making optimal (but never perfect) decisions. In the next sections I articulate two important tradeoffs that analysts face when specifying and training predictive models.

Bias vs. Variance

This is perhaps one of the most fundamental challenges analysts face. Bias occurs when the effect reported by a model deviates from the true effect. We can never know what a true effect is, but we can evaluate if a reported effect is biased. For example, a model that predicts the miles per gallon (mpg) of a vehicle with vehicle weight as the only predictor will probably find that weight (wt) has a negative effect on mpg. But do we really know if the decline in mpg is actually due to the vehicle's wt? Or is it perhaps due to some other factor omitted from the model that may correlate with wt, such as engine horsepower (hp)? More powerful engines tend to weigh more. So, is it wt, hp, or both that affect mpg? If we omit hp, the effect of wt will be biased because it will pick up some of the effect of the omitted predictor hp. This is known as *omitted variable bias*. Generally speaking, if any two predictors belong in a model but are both correlated, omitting one will cause the independent effect of the included predictor to be biased. Including both will reduce the bias because the effects of both predictors are considered. So, would this suggest that we are better off including lots of predictors? Not exactly! As we add more predictors, we increase the dimensionality of the model. Dimensionality comes with a number of problems such as multicollinearity, which we will discuss in depth later on. But the general consequence of adding predictors to the model is increased model variance. In a nutshell, a model has high variance if you drop a few observations and get very different results. A model has less variance (and hence is more reliable) if when we drop a few observations and re-train the model, we get consistently similar results. A good model should have low bias and low variance. But the problem is that bias increases and variance decreases with model simplicity (i.e., fewer predictors, less complex specifications). Conversely, as we add complexity to the models (e.g., more predictors, non-linear terms, etc.), the bias decreases but the model variance increases. This is known as the *bias vs. variance tradeoff*. Later in this book you will learn to apply CV methods to identify models and specifications that provide the optimal balance of bias vs. variance.

Interpretable vs. Black Box Models

In some cases, we are interested in explaining what is causing a particular outcome to vary. For example, if you build a model to aid loan application decisions, you may want to use interpretable models. It is not just a matter of approving or declining a loan; you may need to explain to customers and regulators why a loan was declined. In this case, you will need a model to provide the effect of specific predictors like loan balance, credit score, age, income, etc. In other situations, you may be interested simply in making accurate predictions. For example, if you want to catch fraudulent credit card transactions or if you want to train a self-driving car to stop at a red light, you may not need an interpretable model. Interpretable models tend to be parametric but offer great insights into what are driving the observed effects. They are preferred when the analytic goals are interpretation and inference (i.e., testing hypothesized effects). Ordinary least squares (OLS) and logistic regression models are the most prominent examples of interpretable models for quantitative and classification predictions, respectively. These models need to be evaluated for good fit and are rich in coefficients and fit statistics to help select models and model specifications. Black box models, such as neural networks are generally non-parametric and are based on iterative calculations aimed at providing optimal predictive accuracy. They are preferred when the analytic goal is predictive accuracy and are typically evaluated and tuned through CV accuracy testing. Naturally, there are some models that are more interpretable, like principal components regression, boosted trees, and LASSO regression, which often yield good CV accuracy.

1.4 The Analytics Lifecycle

Analytics is not just modeling and analysis. It is a process aimed at solving problems and supporting managerial decisions. Also, while we often engage in individual analytics projects, most companies view analytics as an ongoing endeavor with a typical lifecycle. One of the most popular life-cycle frameworks for analytics work is the Cross-Industry Standard Process for Data Mining (CRISP-DM), which is an open standard process framework, illustrated in Figure 1.3.

The first step, CDM-1, is understanding the business. Analytics problems and questions are not analyzed in a vacuum. There must be a specific business problem that needs to be solved or a business question that needs to be answered. The business question then needs to be translated into an analytics question. The business and analytics question should be relatively similar, except for the following: (1) not every business question can be answered through analytics. We are only interested in business questions or problems that can be pursued with data; and (2) the business question is usually articulated in terms of a business case in plain English, whereas the analytics question gets to the specifics of the model to be trained. That is, the question should articulate whether we are pursuing a classification or a quantitative prediction and should also include the key predictors.

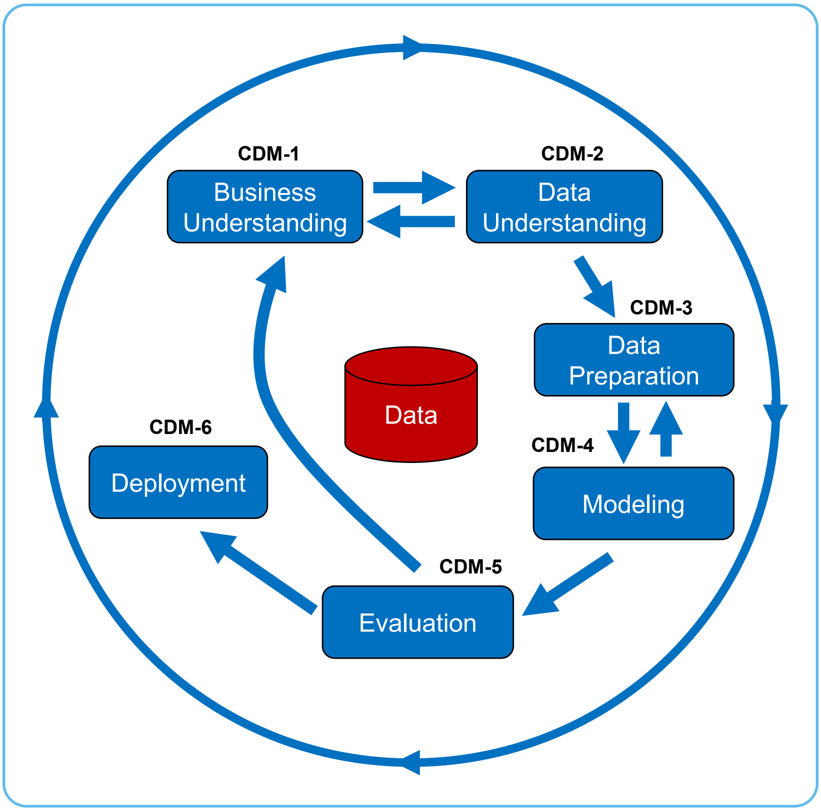


Figure 1.3 CRISP-DM Lifecycle Framework

The second step, CDM-2, is about understanding the data. This step involves doing descriptive analytics to understand the basic statistics, properties and trends of the data. This may involve visual analytics (e.g., plots, distributions, charts, etc.) or quantitative analysis (e.g., correlation, analysis of variance, cluster analysis, etc.). As the diagram above illustrates, after reviewing the data we usually loop back to refine our understanding of the business problem, loop forward again, and keep cycling as needed.

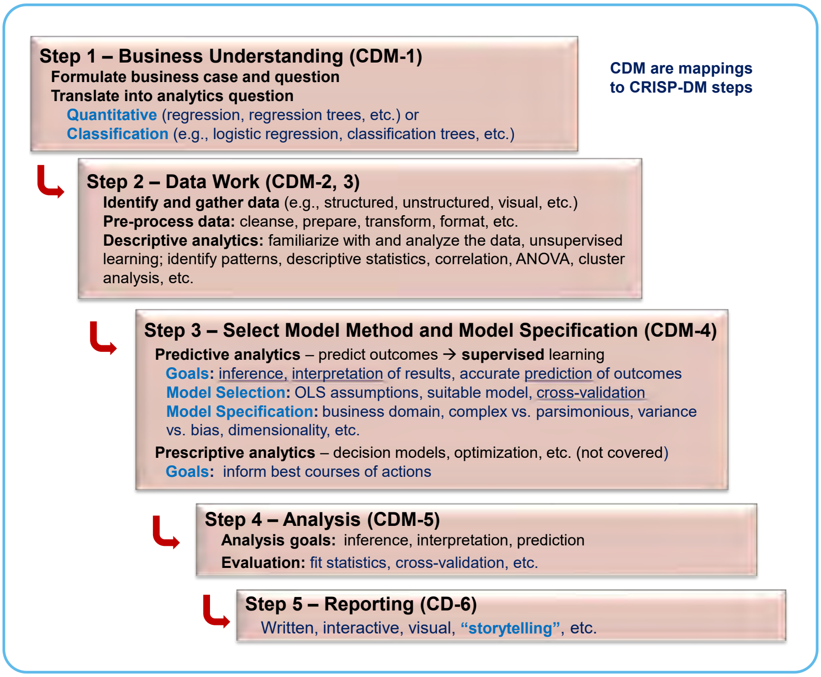
Also included in the second step, CDM-3 is about preparing the data. Experienced practitioners note that about 80% of the work in analytics projects involves data work. The data is never ready for analysis right off the bat. It usually requires a fair amount of extraction-translation-loading (ETL) work, data pre-processing, curation, correcting anomalies and inconsistencies, and dealing with missing data, among other things. In addition, it is often necessary to do some data transformations (e.g., logs, polynomials, merging tables, etc.), either to comply with model assumptions (e.g., linearity, independence of residuals) or to improve the statistical fit and accuracy of the model.

The third step is CDM-4: modeling. No serious predictive analytics project can be conducted with a single modeling method or specification. The analyst will need to try multiple modeling approaches and test various specifications. After modeling, you may need to loop back to CDM-3 and do more data work, loop forward to CDM-4, and continue modeling and cycling back until a stable model is identified.

The next step is CDM-5: evaluating the model. This involves selecting the model that has the best statistical properties, fit, interpretability and predictive accuracy, and one that provides the best answer to the focal business question(s) of the project. At this point, the process may loop back to CDM-1 and repeat the cycle to revise or refine the business question and do any additional data work and modeling required. If not, the process proceeds to final step, CDM-6: utilizing the final results. This step also involves a substantial amount of interpretation and meaningful storytelling and report writing.

I would like to note that this is not the only popular analytics framework. INFORMS also presented a Job Task Analysis (JTA), which has similar steps. The INFORMS Analytics Body of Knowledge (http://info.informs.org/analytics-body-of-knowledge) has a chapter that explains the JTA in detail, providing a very useful mapping comparison of the JTA task steps to the CRISP-DM steps. For this book, I have developed a framework that is consistent with the CRISP-DM and JTA but maps more closely to the activities covered in this book. Figure 1.4 illustrates this framework with the respective mappings to the CRISP-DM steps, which we will follow in the next few chapters.

The process depicted in Figure 1.4 fits this book's material better because, in addition to the CRISP-DM steps, it provides: more details about the business and analytics questions; distinguishes descriptive from predictive analytics; and considers analytics goals. The analytics goals



Step 1 – Business Understanding (CDM-1)
Formulate business case and question
Translate into analytics question
Quantitative (regression, regression trees, etc.) or
Classification (e.g., logistic regression, classification trees, etc.)

CDM are mappings to CRISP-DM steps

Step 2 – Data Work (CDM-2, 3)
Identify and gather data (e.g., structured, unstructured, visual, etc.)
Pre-process data: cleanse, prepare, transform, format, etc.
Descriptive analytics: familiarize with and analyze the data, unsupervised learning; identify patterns, descriptive statistics, correlation, ANOVA, cluster analysis, etc.

Step 3 – Select Model Method and Model Specification (CDM-4)
Predictive analytics – predict outcomes → **supervised** learning
Goals: inference, interpretation of results, accurate prediction of outcomes
Model Selection: OLS assumptions, suitable model, cross-validation
Model Specification: business domain, complex vs. parsimonious, variance vs. bias, dimensionality, etc.
Prescriptive analytics – decision models, optimization, etc. (not covered)
Goals: inform best courses of actions

Step 4 – Analysis (CDM-5)
Analysis goals: inference, interpretation, prediction
Evaluation: fit statistics, cross-validation, etc.

Step 5 – Reporting (CD-6)
Written, interactive, visual, **"storytelling"**, etc.

Figure 1.4 Predictive Analytics Process

can be inference—when we have a hypothesized effect (e.g., advertising effects increase sales) that we want to test, interpretation (e.g., what are the main predictors for loan approval), and/or prediction (e.g., accurate recognition of a tumor in a digital image). In a nutshell, the predictive analytics process we follow in this book has five steps: (1) formulate the business case and business question. Translate the business question into an analytics question, which should be formulated in a way that it can be answered through analytics. This means articulating if the question is about predicting a quantitative or classification outcome, and identifying the key predictors that may help answer the question; (2) perform data work by involving the necessary data ETL, curation, cleansing, pre-processing, transformations and descriptive analytics; (3) select a model, which starts by articulating the predictive modeling goals (i.e., inference, interpretation and/or prediction), and then exploring all feasible modeling methods (e.g., regression, decision trees, neural networks, etc.) and appropriate model specifications (i.e., variable selection, non-linear specifications, etc.); (4) evaluate all candidate models by analyzing model assumptions, fit statistics, and cross-validation accuracy testing, then select the best candidate and conduct the analysis; and (5) prepare the report with an emphasis on effective and clear storytelling.

1.5 Data Structures: Vectors, Matrices, and Data Frames

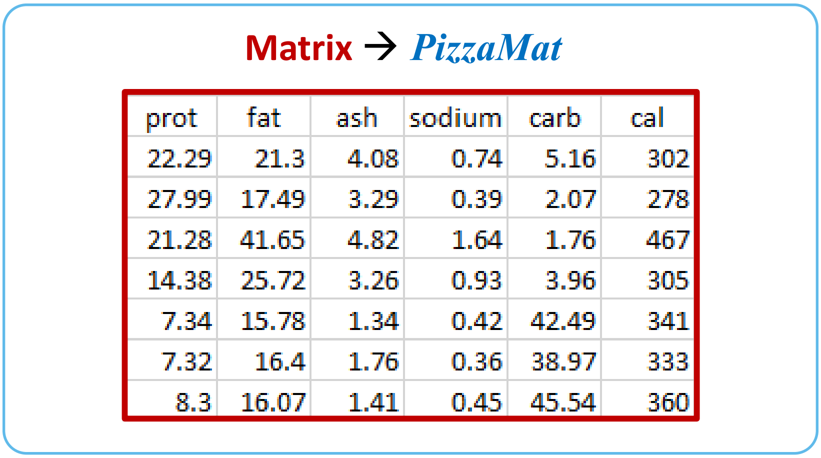
Data structures are essential to computational analytics. I discuss data structures in more detail in the R appendix to this chapter. In this section I describe three fundamental types of data structures used in predictive analytics.

Vectors

A vector is a collection of values of the same type. They can be a collection of numbers, text strings, dates, etc., but they cannot contain comingled data of different types. Vectors are of key importance in predictive analytics because they are ideal structures to store a variable's values and can be easily attached to or extracted from matrices and data frames, as explained below. For example, if we are building a model to predict a variable Y (e.g., sales) and we have one value of Y for each data observation, we can store all observations in a vector called Y (we use upper-case bold to denote vectors or matrices and regular font for single values).

Matrices

Matrices are like vectors but contain more than one dimension. For the purposes of this book, we can think of matrices as containing two dimensions—rows and columns. All values in a matrix must be of the same type. While you can store any type of data in a matrix, matrices are predominantly used in predictive modeling to store numerical values. Each row typically represents a data observation, and each column represents a variable in the dataset. Figure 1.5 provides an example of a data matrix containing nutritional information on pizza slices, with one row for each pizza brand and one column for each nutritional variable.



Matrix → *PizzaMat*

prot	fat	ash	sodium	carb	cal
22.29	21.3	4.08	0.74	5.16	302
27.99	17.49	3.29	0.39	2.07	278
21.28	41.65	4.82	1.64	1.76	467
14.38	25.72	3.26	0.93	3.96	305
7.34	15.78	1.34	0.42	42.49	341
7.32	16.4	1.76	0.36	38.97	333
8.3	16.07	1.41	0.45	45.54	360

Figure 1.5 Matrix Illustration

Data analysis software programs like R or Python are rich in matrix manipulation features because many statistical routines are computed using matrix algebra. To follow my explanation below, please note that I use uppercase bold to denote a vector or matrix and regular font for single values. For example, look at the matrix representation of the regression model in Figure 1.6. The vector Y represents a column of actual values for the outcome or response variable we want to make predictions for (e.g., sales). Each element in this vector represents the actual value of the outcome variable for that observation. X is a matrix containing one row for each data point or observation and one column for each predictor in the model. The goal of the predictive model is to find a vector of linear weights or coefficients β, containing one value for each predictor, representing the effect that the corresponding predictor has on the outcome variable. For example, if β_3 is 0.72, it means that when the predictor X_2 increases by 1 unit, the outcome variable increases by 0.72 units. If you multiply the matrix X times the vector β, it is the same as multiplying each predictor by its corresponding β weight effect, yielding the prediction vector $\hat{Y}$, which has the predicted or fitted values by the model for each data point. Naturally, the prediction is not perfect, so the model also yields the residuals or errors vector ε, which shows the difference between the actual values in Y and the predicted values in $\hat{Y}$.

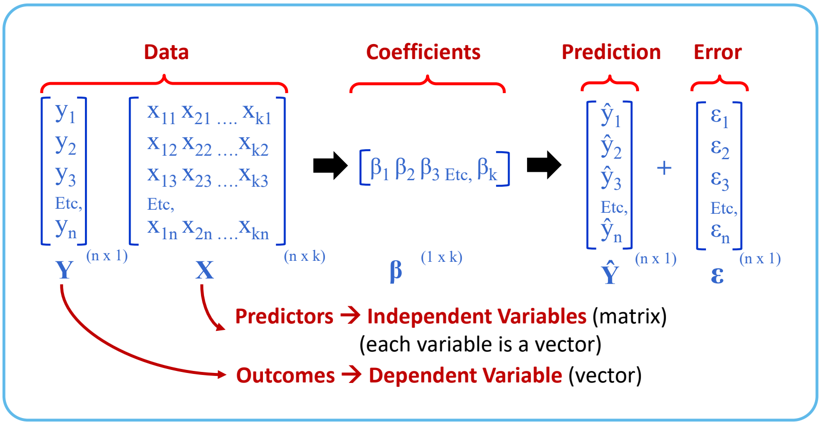

Data
Coefficients
Prediction
Error
y_1 y_2 y_3 Etc, y_n
x_{11} x_{21} .... x_{k1}
x_{12} x_{22} .... x_{k2}
x_{13} x_{23} .... x_{k3}
Etc,
x_{1n} x_{2n} .... x_{kn}
β_1 β_2 β_3 Etc, β_k
$\hat{y}_1$ $\hat{y}_2$ $\hat{y}_3$ Etc, $\hat{y}_n$
+
ε_1 ε_2 ε_3 Etc, ε_n
$\mathbf{Y}$ (n x 1)
$\mathbf{X}$ (n x k)
$\boldsymbol{\beta}$ (1 x k)
$\hat{\mathbf{Y}}$ (n x 1)
$\boldsymbol{\varepsilon}$ (n x 1)
Predictors → Independent Variables (matrix)
(each variable is a vector)
Outcomes → Dependent Variable (vector)

Figure 1.6 Matrix Representation of a Predictive Model

While we will not be working with matrices in this book, it is important to understand the basic concepts behind matrices because some functions in R and Python require that we use matrices. Some important statistical concepts are also best represented as matrices. Take, for example, two of the most fundamental statistical artifacts, the covariance and correlation matrices. Many statistical methods involve preparing and manipulating the covariance or correlation matrix. In fact, there is a whole family of advanced models called *covariance-based models*. Figure 1.7 below illustrates the covariance and correlation matrices and how they relate to each other.

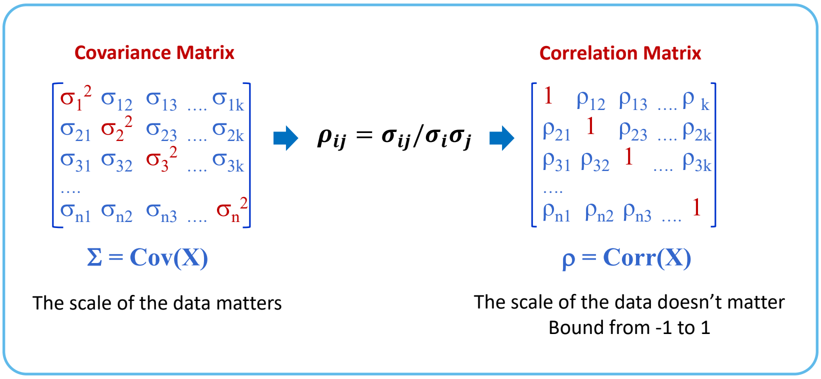
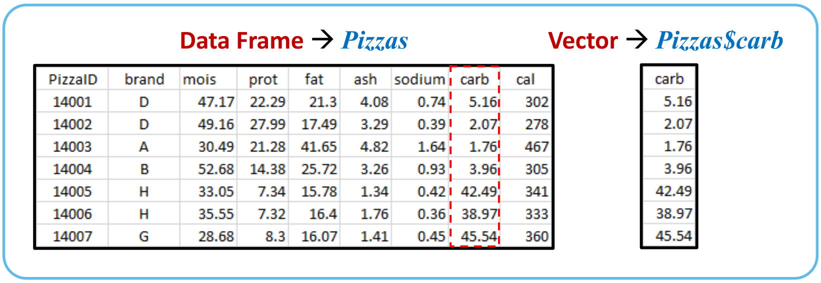

Covariance Matrix
σ_1^2 σ_{12} σ_{13} .... σ_{1k}
σ_{21} σ_2^2 σ_{23} .... σ_{2k}
σ_{31} σ_{32} σ_3^2 .... σ_{3k}
....
σ_{n1} σ_{n2} σ_{n3} .... σ_n^2
$\rho_{ij} = \sigma_{ij}/\sigma_i\sigma_j$
Correlation Matrix
1 ρ_{12} ρ_{13} .... ρ_{k}
ρ_{21} 1 ρ_{23} .... ρ_{2k}
ρ_{31} ρ_{32} 1 .... ρ_{3k}
....
ρ_{n1} ρ_{n2} ρ_{n3} 1
$\Sigma = \mathrm{Cov}(X)$
$\rho = \mathrm{Corr}(X)$
The scale of the data matters
The scale of the data doesn't matter
Bound from -1 to 1

Figure 1.7 Correlation and Covariance Matrices

We will discuss the concept of variance and covariance in more depth in the next chapter. For now, in simple terms, variance is about how much the values of a variable deviate from its mean (squared). For example, In Figure 1.6, $\sigma_3{}^2$ in the (diagonal of the) covariance matrix Σ represents the variance of predictor X_3 and σ_3, its standard deviation. A large covariance indicates that the values in X_3 vary widely with respect to X_3s mean. Covariance measures whether two variables tend to vary in the same direction or not. For example, σ_{23} in the (off-diagonal of the) covariance matrix Σ tells us whether X_2 and X_3 vary in the same direction. A high positive covariance indicates that when X_2 is high, X_3 tends to be high. A high negative covariance indicates that when X_2 is high, X_3 tends to be low. A covariance value close to 0 indicates that the two variables are unrelated. One of the big issues with covariance analysis is that covariances change substantially when you re-scale variables. For example, if you are analyzing the relationship between a vehicle's weight and its gas mileage, the covariance between these two variables will be very different if we use pounds and miles per gallon compared to if we use kilograms and kilometers per liter. To avoid this issue of scale invariance, many statistical methods rely on the correlation matrix, which has the standardized equivalents of the covariance values. If you divide each covariance value in Σ by the respective variables' standard deviations, you obtain the correlation matrix. This division causes the diagonal elements to be exactly 1 and the off-diagonal elements to be bound between -1 and +1. Thus, these values are not affected if you re-scale the variables. Like with covariance, correlation values close to -1, 0 or +1 indicate that the variables vary in opposite directions, are unrelated, or vary in the same direction, respectively.

Data Frames

Data frames are similar to matrices, except that each column can be of a different type, but the data within a column must be of the same type. For example, columns may contain numerical data (e.g., sales, advertising expenditures, etc.), dates (e.g., sale date), categorical data (e.g., state, customer type), and logical (e.g., true or false) or binary data (0 or 1). Since the data in each column must be of the same type, data frames are the most versatile data structures in predictive modeling. Notice in the matrix previously shown in Figure 1.5 that all the data is quantitative. Figure 1.8, however, illustrates the more complete data frame and how easy is to extract vectors from data frames or combine vectors into data frames. Vectors, matrices, and data frames are essential to descriptive and predictive analytics, and it is important to understand their differences because some R functions require the use of vectors, while others require the use of either matrices or data frames.



Data Frame → *Pizzas*

PizzaID	brand	mois	prot	fat	ash	sodium	carb	cal
14001	D	47.17	22.29	21.3	4.08	0.74	5.16	302
14002	D	49.16	27.99	17.49	3.29	0.39	2.07	278
14003	A	30.49	21.28	41.65	4.82	1.64	1.76	467
14004	B	52.68	14.38	25.72	3.26	0.93	3.96	305
14005	H	33.05	7.34	15.78	1.34	0.42	42.49	341
14006	H	35.55	7.32	16.4	1.76	0.36	38.97	333
14007	G	28.68	8.3	16.07	1.41	0.45	45.54	360

Vector → *Pizzas$carb*

carb
5.16
2.07
1.76
3.96
42.49
38.97
45.54

Figure 1.8 Data Frame and Vector Illustrations

1.6 Predictive Analytics Overview

Descriptive, Predictive, and Prescriptive Analytics

As we discussed earlier, INFORMS categorizes analytics into these three types. Descriptive analytics models aim to understand the data at hand and uncover trends in the data and can be visual or quantitative. Predictive analytics is the main subject of this book and section, and it aims to use part of the data to predict outcomes. Prescriptive analytics is about developing models to help managers make decisions and are based on management, quantitative and decision sciences.

Quantitative vs. Classification Prediction

Some models predict quantitative outcomes (e.g., sales, income, sale prices, amount of debt, etc.) whereas others predict whether a prediction will be categorized under a certain classification (e.g., loan default, fraudulent transaction, positive illness diagnosis, etc.). A quantitative model estimates an outcome based on the predictors supplied to the model. The evaluation of the fit of quantitative models is fairly straightforward, and it is usually based on some criteria that indicates the amount of error in the model predictions (e.g., R-Squared or the proportion of variance explained by the model; mean squared error). In contrast, a classification model predicts the likelihood that an outcome will fall into a particular category. The challenge with classification models is that we need to convert the likelihood into a specific classification prediction. This conversion needs to be based on some arbitrary classification threshold. For example, if the probability of a transaction to be fraudulent is 50% or more, we would classify the transaction as a fraud attempt. However, we can change this threshold to, say 70%, which will cause us to be more forgiving because a transaction will not be flagged as fraudulent until the likelihood reaches 70%. Conversely, if we want to be very conservative about preventing fraud, we could lower the threshold to, say 30%, so a small probability of fraud will trigger an alert. As we vary this classification threshold, the model's accuracy at predicting true positives (i.e., sensitivity, e.g., fraud) and accuracy at predicting true negatives (i.e., specificity, e.g., no-fraud) will move in opposite directions.

Parametric vs. Non-Parametric Models

A parametric model is one that is based on mathematical assumptions about certain parameters of the model. Parameters are important because they simplify an otherwise complex problem. For example, an OLS regression model requires that the residual errors be normally distributed and that the predictors be independent of each other. Why? Because assuming a normal distribution of the residuals and zero correlation between predictors simplifies the formula that solves the OLS regression model. It can also be demonstrated mathematically that the OLS model is the most unbiased model you can use, but only if its assumptions are met. Model assumptions and parameters also provide information that is useful to understand and interpret a model. For example, if we assume that a variable is normally distributed, it means that we assume that the variable has a mean and a variance that is equally distributed to the right and left of the mean, going from - ∞ to + ∞. Because the normal distribution is based on a mathematical gaussian curve, we know that +/- 2 time the standard deviation from the mean encompasses about 95% of the data, which is very helpful to compute confidence intervals for predictions and effect weights. However, if the data is not normally distributed, we cannot use any of this information because there is no parametric mathematical function to guide us. In contrast, a non-parametric model is not bound by any parameters, but this also means that there is no mathematical model to help you explain the model. Again, this is a tradeoff. Predictive models in general vary based on the extent to which the model relies on heavy parametric assumptions (e.g., OLS, logistic regression), no parametric assumptions (e.g., trees, k-nearest neighbors, neural networks), or somewhere in between (e.g., ridge regression, principal components regression, etc.).

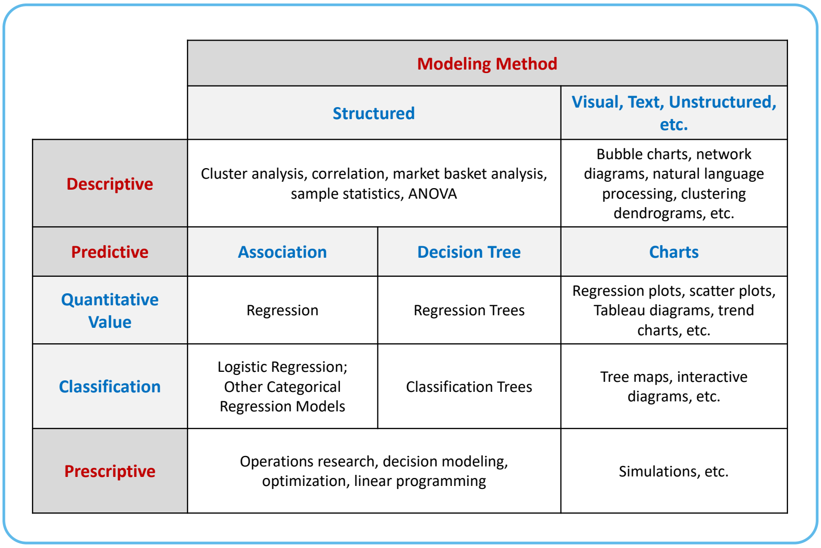
Association vs. Tree Methods

There are many modeling approaches. The most popular ones are either based on statistical association or decision trees. The most common methods based on statistical association are linear regression (quantitative), logistic regression (classification), ridge regression, LASSO regression, principal components regression, partial least squares, spline regressions (i.e., non-linear models), etc. Decision tree methods are either regression trees (quantitative) or classification trees. Within the decision tree methodology there are several tree modeling approaches like bootstrap aggregation, random forest, and boosted trees. In addition, there are many other modeling methods that fall somewhere in between, and each is based on some unique algorithmic approach, including: support vector machines, k-nearest neighbor, neural networks, etc.

Figure 1.9 provides a general framework for classifying analytics models.

1.7 Predictive Modeling Goals

Before selecting candidate predictive models, it is important to articulate the predictive modeling goals for the project. These goals should derive from the business and analytics questions and be applicable to both



	Modeling Method		
	Structured		Visual, Text, Unstructured, etc.
Descriptive	Cluster analysis, correlation, market basket analysis, sample statistics, ANOVA		Bubble charts, network diagrams, natural language processing, clustering dendrograms, etc.
Predictive	Association	Decision Tree	Charts
Quantitative Value	Regression	Regression Trees	Regression plots, scatter plots, Tableau diagrams, trend charts, etc.
Classification	Logistic Regression; Other Categorical Regression Models	Classification Trees	Tree maps, interactive diagrams, etc.
Prescriptive	Operations research, decision modeling, optimization, linear programming		Simulations, etc.

Figure 1.9 Analytics Models Framework

quantitative and classification models. Defining the predictive modeling goals will help narrow down the number of candidate models to consider. There can be many predictive modeling goals, but there are three specific types of goals that are very useful to guide the model selection process: interpretation, inference, and prediction.

Interpretation

Interpretation is about extracting meaning from the predictive model to better understand not just whether certain predictors influence outcomes, but how and how much. For example, if you want to make managerial decisions to increase sales in a company, which operational aspects do you need to manipulate? Do you need to make changes to operations, the supply chain, advertising expenditures, product quality, etc.? To answer questions like these you need to have an interpretable model. I once heard a debate between a data scientist and a credit card company analyst. The data scientist's argument was that it was better to have a very accurate model that could predict good and bad loan applicants, so he did not care much for interpretability. The analyst replied that he needed an interpretable model because he was accountable to regulators and needed to have good explanations to give to applicants and regulators when applications were declined. When the predictive modeling goal is interpretation, parametric models are usually preferred. The most interpretable quantitative model is an OLS regression; for classification models it is a logistic regression model. The OLS model has many assumptions that need to be met, but when the OLS assumptions hold it is said to be BLUE—best (least variance) linear unbiased estimator. For this reason, I recommend all predictive modeling projects to start either with OLS (for quantitative outcomes) or logistic regression (for classification outcomes), not only because they are BLUE when all their model assumptions hold, but also because they are the most unbiased models so they can always be used as a benchmark to evaluate other models. The further a model departs from either of these models, the less we can rely on the interpretations of its effects on the outcome variable.

Inference

Inference is similar to interpretation in the sense that a model that is good for interpretability is also good for inference. Inference means that you have one or more hypotheses you want to test. For example, if you believe that increasing advertising expenditures will lead to more sales but you need to prove this to your client or manager, you can specify a model to test this hypothesis. A model that is interpretable will generally be suitable to test hypotheses like this. A hypothesis has two parts, H_0 or null hypothesis and H_A or alternative hypothesis. Your analysis needs to support one or the other. Generally speaking, H_0 is not what you are trying to prove, but what you are trying to reject, statistically. So, when you reject $H_{0,}$ you state that your analysis supports H_A. You will notice that this is a consistent approach throughout this book and in most standard tests. For example, in the sales example, you would formulate your hypotheses as follows:

H_0: Advertising expenditures have no effect on sales.
H_A: Advertising expenses have an effect on sales.

For H_A, you could be more specific and state that the effect is either positive or negative. An interpretable model based on statistical association like

OLS is perfect for testing hypotheses. The reason why we articulate H_0 as the null hypothesis we are interested in rejecting is because statistically, you can never prove a hypothesis. That is, you cannot accept it unconditionally because it is based on probabilities. You can only reject it or fail to reject it. We usually state a rigorous level of confidence when rejecting H_0 because this provides good statistical evidence that H_0 is not true, and therefore conclude that H_A must be true. We often use a threshold *significance* value of 5% (sometimes referred to as the α-value). The statistical significance level of a predictor is given by its p-value. Think of the statistical significance as the opposite of statistical confidence. If a particular regression predictor weight effect has a p-value < 0.05 (α-value), we say that this effect is significant at the 5% level. What this means is that we have a 95% confidence level that our results did not happen by chance and that we can therefore reject the null hypothesis of no (zero) effect. Conversely, if the regression weight for a particular predictor has a p-value > 0.05 (α-value), we say that the predictor is not significant and that we have no statistical confidence to reject the null hypothesis of no effect. That is, we conclude that there is no effect. This p-value < 0.05 is quite arbitrary and hotly debated in the literature, but it is also a generally accepted rule of thumb in predictive modeling. However, if you find an effect with a p-value < 0.10, it is not a bad thing because you still have 90% confidence that there is an effect, and it is entirely correct to say that the effect is significant at the $p < 0.10$ level. So, be thoughtful about this.

Prediction

Prediction is when you care more about being correct in your predictions, and interpretation is less important. Let us say you are trying to teach a vehicle how to drive. Well, you want that vehicle to stop at a red light. You want that vehicle to turn when you have the green arrow pointing to the right. You want the vehicle to stop when there is a pedestrian crossing. You do not care about interpretation. You just want the model to be correct. The same is true if you are developing a speech recognition machine learning algorithm in which your goal is to identify the spoken words correctly without having to provide interpretations. In many cases, the predictive modeling goals are more than one. For example, you may be interested in both good interpretability and predictive accuracy. In such cases, you would have to identify all feasible interpretable models as candidates and then select the most accurate of them. Accuracy is always tested using ML methods based on CV. The typical process involves training (i.e., fitting) the model with part of the data randomly selected and then making predictions with the rest of the data. Since you already have the data, you can compare how well your model predictions compare to the actual data. Many CV testing methods involve re-sampling many times, so that you train and test the model multiple times. This will give you more confidence that your model will predict well when new data arrives. For example, if you have data about individuals contracting a contagious illness and you want to predict what leads to infection, you could specify a classification model with positive vs. negative as a binomial outcome. You could then train your model with 80% of the data and test it with the remaining 20%. Because you have actual data for this 20%, you can evaluate how accurately your model is predicting. Again, if the goal is only predictive accuracy, then you can train and test other models (less interpretable, non-parametric, or black box). Typical applications of this type of model include detection of cybersecurity breaches or fraudulent transactions, detection of spam mail, recognition of handwritten zip codes, tumor detection in digital images, voice recognition, etc.

1.8 Modeling Method and Model Specification

Many analytics, data science, and machine learning books focus on teaching you how to fit (i.e., train) and test models. Some books go deep into the mathematics and statistics of the models and some stay more at a high level. There are several excellent books on these topics. If you are interested in learning, for example, how to train a deep convolutional neural network, you will find a wealth of materials in print and online. What distinguishes this book from others is that it contains far more information about how to select the right model and about how to specify a model correctly.

Modeling Method

In order to have the best possible model, you will need to try many models. Data scientists often try dozens of models before they write up their results. The first challenge is to select the most appropriate model from several candidate models based on the type of analytics question you seek to answer (i.e., quantitative or classification) and the predictive modeling goals for your project (i.e., interpretation, prediction and/or predictive accuracy). Interpretation and/or inference goals will lead you in the direction of more parametric and explainable models. If your goal is predictive accuracy, then any model is fair game, and the best modeling method should be one that yields the best CV test results.

Model Specification

Once you have several modeling methods identified, the next step is to formulate the specification of the model. This involves two things: (1) the specific predictors to include in the model, and (2) the type of transformations imposed on the data (e.g., logs, polynomials, interaction terms, etc.) either to meet model assumptions or to improve the model performance. The selection of predictors for a model can and should be done from two perspectives: business domain knowledge and statistical fit. The initial predictor selection should be driven by knowledge of the business domain of the analysis. Datasets can be overwhelmingly large, and it is not uncommon to find ourselves with hundreds, if not thousands, of predictors from which to choose. It would be a mistake to try initially every variable in dataset because you will most likely encounter all kinds of spurious correlations among predictors, redundant variables, and other anomalies. The first model specification should be driven by business domain familiarity. If you are building a model to predict the spread of an infectious disease or stock market performance, it would be extremely useful to consult with subject matter experts in healthcare and finance, respectively. This will narrow down the number of relevant predictors to a more manageable number.

But then you need to evaluate these predictors statistically. If the predictors are highly correlated, you will encounter issues with a persistent problem called *multi-collinearity*. If you have the freedom to include or remove predictors in the model, you can go through statistical methods for variable selection, such as best subsets or stepwise analysis. The general principle to follow is that it is okay to remove non-significant predictors, but you should retain significant predictors. As we discussed earlier, as you add predictors and complexity to a model, its bias is reduced, but its variance increases. At some point, the improvement in bias is offset by the increased variance in the model, which is a well-known problem known as *dimensionality*. One main contributor to dimensionality is multi-collinearity, which occurs when predictors are highly correlated with each other. If you have a model with a very large number of predictors, you will most likely encounter these problems. If removing predictors is not appropriate, you can try other methods like dimension reduction (e.g., principal components and partial least squares regression). You should also use business intuition and statistical testing to decide if there are appropriate transformations (e.g., logs, non-linear polynomials).

You may also rationalize and/or test the inclusion of interaction variables. Typically, the effect of predictors is additive. For example, if you are predicting miles per gallon in vehicles and you have engine size and origin (i.e., domestic vs. foreign), the effect engine size has on miles per gallon will be independent of origin and same thing for the effect of origin, and each predictor will contribute their respective effects in an additive fashion. However, if we suspect that the effect of engine weight may be stronger with foreign vehicles than with domestic ones, we could include a multiplicative term, which we call an *interaction term* to evaluate the *joint* effect of both predictors (i.e., how one affects the other).

Final Notes

My goal is to help you become a good predictive analyst by the time you finish reading this book. This book is not about programming in R or Python, and it is not about statistics or deep data science. You will learn a lot about these, no doubt. But this book is intended for the beginner analyst or analytical manager who wishes to dig deeper into predictive analytics and ML. While this book is for business professionals, it covers a fair amount of software programming, statistics, and mathematics. While not strictly required, the reader is encouraged to gain some basic familiarity in these areas before reading this book. More importantly, there is an insane number of modeling methods out there and many more coming out all the time. It would be impossible to cover all of them. So, the skills you will acquire by reading this book are a good understanding about the various model types out there and the possible model specifications to try, and a good sense for when to use these. In other words, this book will provide you with a basic and solid understanding of the various modeling approaches, which will enable you to research other models on your own, and help you become an effective analytics learner.

In the next chapter, I will provide an overview of important foundations of descriptive and regression modeling on which the rest of the book will be based.

“Today, managers and business leaders need to understand the power of data analytics. Professor Espinosa provides a robust roadmap to guide professionals interested in improving business outcomes with predictive analytics. He also provides an excellent compendium of statistical concepts to help professionals understand and implement advanced, complex analytical models, and discusses how to make better business decisions using data and machine learning.”

ROD FONTECILLA

Partner and Chief Innovation Officer
Technology Solutions
Guidehouse

“Professor Espinosa’s book is a must-read for analysts and managers interested in learning how to: use predictive analytics for decision-making, frame a business analytics question, identify relevant predictors, select the optimal model, and interpret results. He does a wonderful job explaining fundamental terms in an understandable manner. The GitHub appendices are a comprehensive guide for analysts and managers.”

WAI FONG BOH

President’s Chair and Professor of Information Systems
Deputy Dean of Nanyang Business School
Nanyang Technological University in Singapore