Chapter Eight

Predictive Analytics: Helping to Make Sense of Big Data

Richard Sherlund of Barclays says the challenge for CIOs in dealing with big data is what to do with all this information.[1]

One rule of thumb today is that almost anything that a typical person can do with less than one second of mental thought we can either now or in the very near future automate.[2]

VLADGRIN/Getty Images

LEARNING OBJECTIVES

After studying this chapter, you should be able to:

- Define data mining.
- Explain the categories of tools available in data mining.
- Relate common data mining terminology to standard statistical terminology.
- Explain the relationship between correlation and data mining.
- Examine the common diagnostic tests used in data mining.

[1] "Why CIOs Aren't Prepared for Big Data," *Wall Street Journal*, March 7, 2017, (https://www.wsj.com/articles/why-cios-arent-prepared-for-big-data-1488855666).

[2] "How Artificial Intelligence Will Change Everything," *Wall Street Journal*, March 7, 2017, (https://www.wsj.com/articles/how-artificial-intelligence-will-change-everything-1488856320).

Applying Analytics in Financial Institutions' Fight Against Fraud

Using data along with other cutting-edge tools can help organizations make better decisions and step up efforts to monitor fraudulent transactions.

Forty years ago, banking fraud might have involved simply forging an account holder's signature on a withdrawal slip. Now the speed and intricacy of the schemes are mind-boggling: a student bank account (with details obtained by a crime gang) receives a payment of £10,000. Within minutes, the funds have been cycled through dozens of accounts before being forwarded to an international account, where the trail suddenly goes cold. No alarm bells go off. No inquiries are made to the bank. The fraud is only discovered much later, at which point the money and the fraudsters are long gone.

Around the world, fraud is an ever-increasing risk for businesses of all stripes. The *2015/16 Global Fraud Report* by Kroll and the Economist Intelligence Unit found that 75 percent of companies surveyed had been victims of fraud in the past year, an increase of 14 percentage points from three years earlier. And, perhaps unsurprisingly, fraud is a particularly serious issue for financial institutions. The Association for Financial Professionals' 2016 Payments Fraud and Control Survey found that 73 percent of finance professionals reported an attempted or actual payments fraud in 2015.

As prevalent as the fraud problem is for financial institutions, it can be difficult to address. Factors that contribute to the challenge include the sheer volume of transactions handled by most institutions versus the relatively small number of fraudulent transactions, the speed with which technology allows fraudsters to operate, poor or incomplete data, and the lack of information sharing among financial institutions. All too often, banks lack the technology and capabilities to implement the necessary safeguards, responding to a primarily digital problem in an analog way—for example, phone calls attempting to piece together the path of a rapid series of money transfers.

For financial institutions, data and analytics can speed the decision cycles used to observe, orient, decide, and act in fighting fraud. Since the best insights are often at the margins of where industries or data sets overlap, it's necessary to pose targeted questions and develop solutions from a variety of information sources. By combining proprietary data sets with industry benchmarks and government information, financial institutions can use artificial intelligence, machine learning, and analytics in the fight against financial fraud. Financial executives should move now to adopt appropriate processes, develop and acquire the necessary talent, and create the right culture to integrate analytics into their fraud-detection efforts.

DEFINING THE ROLE OF ANALYTICS IN ADDRESSING THE CHALLENGES OF FINANCIAL FRAUD

A vast amount of data flows through financial-services organizations, so the ability to harness those data and analyze them effectively could transform the industry's fraud-detection efforts and provide a host of other benefits. Coupling these rich data sets with appropriate analytical models provides a way to harvest the information needed to identify and prevent fraud more effectively. In some cases, an institution's data can be combined with other fraud markers necessary to provide a data set for training the analytics models used to detect possible incidents of fraud.

For financial institutions and government agencies looking to fight fraud, then, the goal should be to aggregate the existing data needed to support more timely detection and to couple those data with the expertise needed to create and apply the most effective fraud-detection models. Doing so successfully can not only produce financial savings but also protect the company's reputation and maintain public confidence. A recent example demonstrates how applying analytics to fraud detection can provide immediate and significant benefits.

A NEW MODEL DETECTS AN UNPRECEDENTED VOLUME OF INVOICE REDIRECTION

Imagine receiving an email from your CEO requesting an update to the payment details of a key supplier. Coming from a trusted source, you might carry out the task without question. But in doing so, you would become an unknowing accomplice to CEO fraud. In this crime,

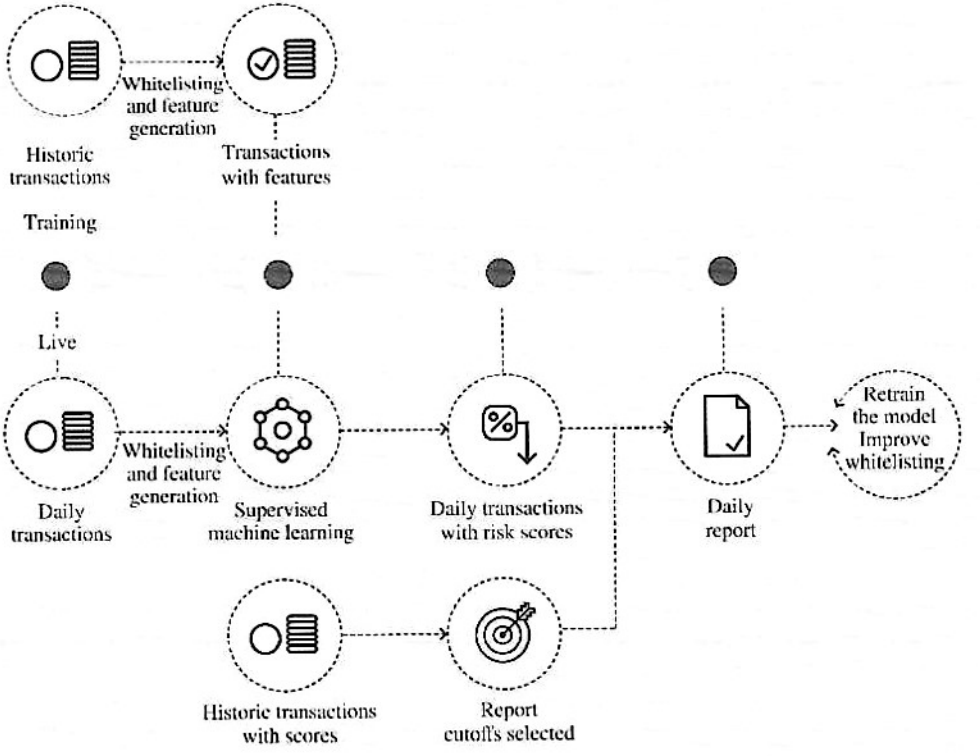
imposters gain access to business email accounts and use them to convince unsuspecting employees to send funds to bogus accounts. CEO fraud has jumped 270 percent from 2015 through Q3 2016 and has led to losses of more than $2.3 billion over the past few years.

Most banks have manual fraud-detection procedures or rules-based solutions, but their effectiveness is limited. The task is especially challenging for invoice redirection, where banks must spot bogus accounts that look very much like the real thing. It's truly like looking for fraud needles in the banking-transaction haystack. In such cases, banks have no way of knowing whether they are paying a legitimate account.

Assembling the data needed to train an analytics model that can accurately identify potential invoice redirection can be a potent weapon in the fight against fraud. QuantumBlack worked with a major bank looking to reduce invoice-redirection fraud—some tens of millions of dollars in value in such invoice redirections from 2010 to 2015—leveraging one of the largest data sets in its country of operation. The goal was to develop a tool that could provide daily reports of suspicious transactions and identify more than 80 percent of fraud cases in both value and incidence.

To score every one of the millions of daily transactions for fraud risk, QuantumBlack built a supervised machine learning model (Figure 8.1). But while the model needed a sufficiently large data set to learn to detect fraud, the number of potentially fraudulent transactions on any given day is so small that waiting for the natural operational work flow to generate the needed number would have taken too long. In response, the QuantumBlack team decoupled the training process from the day-to-day operation and created a partially synthetic data set to train the model.

FIGURE 8.1 A Supervised Machine Learning Model Helped Monitor Transactions for Fraud.

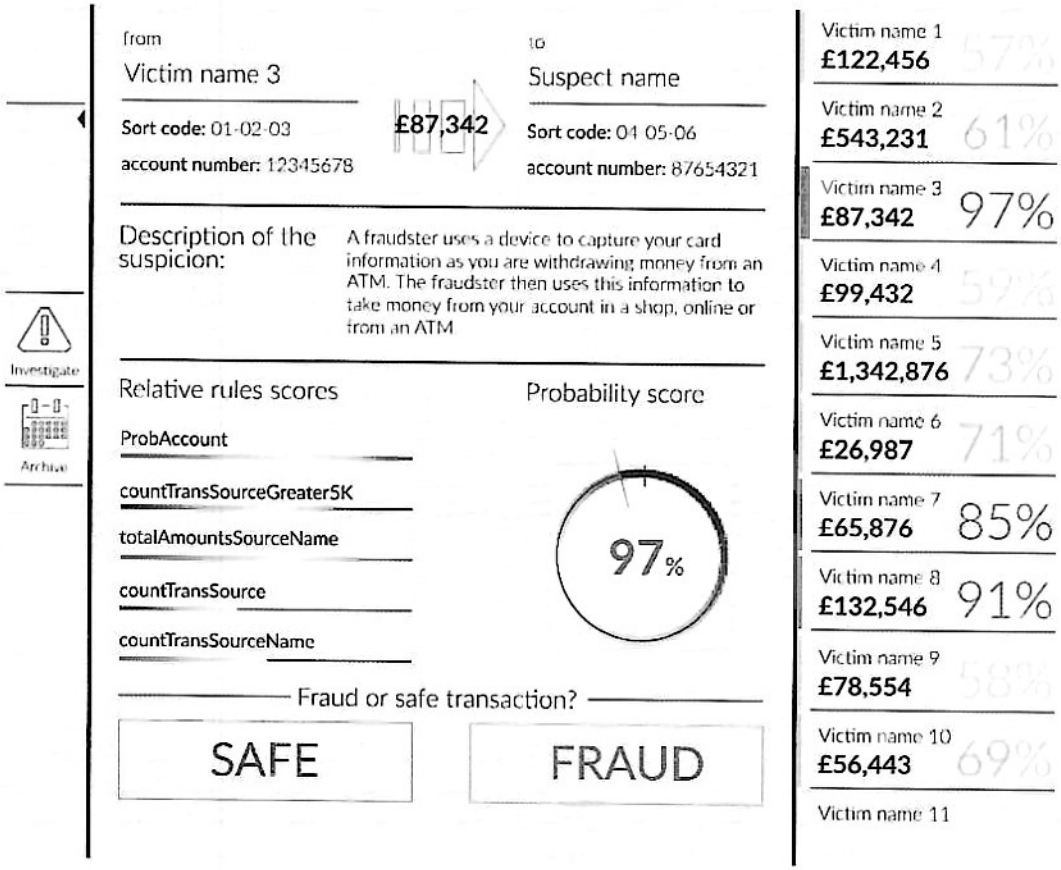


Our team worked closely with the client's data engineering team to ensure computational performance, database best practices, and legal compliance. The curated data sets successfully trained the model to determine which transactions are safe and which are potentially fraudulent.

In actual use, most daily transactions can be immediately categorized as nonfraudulent. The remaining few thousand transactions are run through the machine learning model, which provides a risk score indicating which transactions are most suspicious and which can be assumed safe. By using analytics to combine the value and risk probability of each transaction, the model can instantly rank transactions by risk score. The risk score is computed taking into consideration two different transaction patterns: one between the source and the destination account and one that covers relationships established at the destination account.

The result is that the bank now has a tool that significantly improves its capability to detect high-value fraudulent transactions (Figure 8.2). The live product notifies the bank of an average of 35 high-risk transactions a day out of the several million processed, allowing the bank's fraud team to focus on the transactions that truly demand closer investigation. The investigation results are then used to continue training the

FIGURE 8.2 The Tool Helped a Bank Improve its Fraud-Detection Capability.



machine learning model on both new fraudulent cases as well as new relationships validated as safe.

The predictive model identifies more than 85 percent of fraud cases in value and incidents on the day the transaction is processed, allowing the bank to halt transactions before close of business and recover the funds. Within the first few weeks of live-scoring transactions, the model detected approximately $100,000 in fraudulent transactions. Other banks have expressed interest in the product, which is just the first step of applying analytics and modeling to the financial fraud-detection space.

WORKING TOGETHER TO CRAFT PRACTICAL SOLUTIONS

These use cases reinforce the opportunities for financial institutions to wield analytics to implement real solutions to fraud. The projects often involve bringing multiple players to the table to assemble the data needed to train the models that will identify fraud; those combined efforts are handsomely rewarded through a significant reduction in fraud losses and increased public confidence in financial institutions.

To benefit from the opportunities that data analytics present to fight fraud, executives of financial institutions could implement a framework centered on four key areas:

- *Empower the organization with targeted tools and capabilities.* On top of advanced-analytics solutions, ensure that people can get results out of analytics by providing the training needed to help them understand the results and the markers of fraud. A key element will be creating a culture of vigilance and data-driven decisions. In some cases, it will be necessary to bring in new talent.
- *Redesign processes for speed and efficiency.* Determine how the organization will apply or alter its processes to improve fraud detection, possibly involving changes to the information that's reported or using new tools to obtain better information. An audit to identify data sources and measure data quality could be part of this phase.
- *Mobilize the entire enterprise through effective communications.* Craft a story around the fraud-detection effort and the new advanced-analytics capabilities, how they will be deployed, and their expected benefits. More important, make clear how each individual member of the organization has to change the way he or she operates to deploy those capabilities in day-to-day tasks. Use internal channels to share the story across the organization.
- *Activate the C-suite.* Drive change from the top down. Executives should be involved in analytics initiatives and be vocal advocates for integrating data-driven decision making into all facets of the organization.

Finally, institutions should determine whether to build their own internal data-science capability or work with an outside organization to close any gaps in analytics skills.

USING ANALYTICS TO FIGHT FRAUD

Fraud is a significant problem for all types of financial institutions, but analytics offers the potential to identify fraud cases more quickly and frequently, sometimes even before the fraudulent act occurs. Fortunately, financial institutions already collect a tremendous amount of data that can be used to help fight fraud. The data sets don't have to be perfect to be useful, but a good first step for most organizations is to assess existing data and their quality and determine what other useful data might be collected.

To benefit from the fraud-fighting potential of data analytics, financial institutions must commit to developing the necessary skills and creating the appropriate culture. Given the potentially sizable rewards of reduced fraud losses and maintaining public trust, that commitment should be one all organizations are willing to make.

This article was first published on the QuantumBlack website.

McKinsey Analytics, McKinsey & Company, April 2017

About the authors

Jacomo Corbo is the chief data scientist at QuantumBlack, Chris Wigley is the chief commercial officer at QuantumBlack and a partner in McKinsey's London office, and Carlo Giovine is a manager at QuantumBlack and a consultant in McKinsey's London office.

INTRODUCTION[3]

It became popular to declare oneself dead in order to avoid paying taxes.

A first step in the origin of predictive analytics goes back to the City of London in the 17th century. In 1665 and 1666, the bubonic plague swept through England. The "Great Plague," as it was called, is said to have killed 100,000 individuals. That amounted to about a fifth of the entire population of London at the time. Probably because the plague affected so many inhabitants of London and over such a short period of time, records of deaths fell a bit behind. In fact, it became popular to declare oneself dead in order to avoid paying taxes (even though one was very much alive). The king, of course, would not be in favor of such a practice as the crown was cheated out of taxes due. To prevent this practice, the king instituted the requirement of a death certificate that would include some basic information. As part of this new bureaucratic death certificate process, the king received a report that summarized the recent mortality details. In the preparation of the report (titled the *Bills of Mortality*), patterns were discovered; this may be the first instance recorded in which data collected resulted in the recognition of patterns abstracted from data and not simply the result of using one's eyes to observe a physical pattern in nature (e.g., the stars in the night sky, leaf structure in plants, or the in and out action of sea tides). We have already touched on mortality in Chapter 3; Benjamin Gompertz studied fruit flies to create the model we know today as the Gompertz Curve. That same pattern was exhibited in the London plague data. Recall the pattern of exponential death that showed up as an S-curve in the Gompertz model for new product forecasting. The pattern Gompertz wrote about is a special case of the generalized logistic function we use in present-day data mining.

So, the *Bills of Mortality* may be the first recorded instance of what we now refer to as data mining or predictive analytics. The scribes in London who collected the information were not looking for patterns (other than tax evaders), but unexpected patterns became evident in the numbers as they were collected (think of the process as an early version of streaming data). Data mining today is quite different from what most of us know as standard statistical techniques (sometimes called "frequentist" statistics). In most forecasting situations up to this point in the text, a particular model has been imposed on the data to produce the forecasts; that particular model has been chosen by the forecaster. In most business situations, we assume that our data will, or could, exhibit the patterns we have found to be common in most business data: trend, seasonality, and cyclicality. We have then chosen a model that we believe will represent the data well, say a Holt-Winter's smoothing model; we may believe this model is appropriate because we had visually observed all previous data to include these patterns. In a word, we selected the model and that model was capable of only recognizing certain types of patterns. If we thought "events" were an important pattern in our data, we likely chose an event model with an appropriate underlying model meant to capture the remainder of the pattern. For most of the history of forecasting, including the use of both time series models and demand planning models, this has been

[3] The author would like to thank Professor Eamonn Keogh of the Department of Computer Science & Engineering at the University of California, Riverside; Keogh has provided web readers with excellent examples and explanations of data mining tools and issues.

accepted practice. More recently, our access to large volumes of data and our access to many different types of data, along with some new algorithms, have changed the way data scientists approach prediction. While it is true that prediction is still the goal of data and text mining, the data itself and the tools used to manipulate it have changed in the last decade. The fact that the field of "analytics" is often referred to as "predictive analytics" is a good indicator that the goal of our efforts has not changed: analytics is, like forecasting proper, in the business of making accurate predictions.

If we fast-forward 200 years from the Great Plague, we could observe another early precursor to modern data mining. Florence Nightingale is thought of as the first modern nurse; she was a trailblazing nurse during the Crimean War in the mid-1850s. But Florence Nightingale was also a statistician and quite a skilled one. She was the first woman to be elected as a Fellow to the prestigious Royal Statistical Society in London (called the Statistical Society of London at the time).[4]

During the war, in addition to caring for wounded and infirm soldiers, Nightingale kept meticulous records and displayed them graphically to discern the patterns she thought so evident. She advocated for the uniform collection of medical statistics. Her "Nightingale Rose" (Figure 8.3) diagram, which depicted the collected Crimean War statistics, clearly showed that most soldiers did not die from wounds incurred in battle but rather from preventable diseases mitigated by conditions that could have been alleviated. While she is truly the mother of modern nursing, Nightingale is also one of the first data miners to extract useful information from abstract data. The roots of data mining go far back to the Great Plague and the Crimean War.

The data mining approach is today an altogether different way of viewing the world and the data in it compared to the methods we have employed in previous chapters. IBM is fond of using the term *Big Data* to represent this altogether distinct view of the data we are working with and the techniques used to produce knowledge from that data. But the term *Big Data* does not belong to just IBM; it is also a term commonly used in describe data mining in general.

BIG DATA

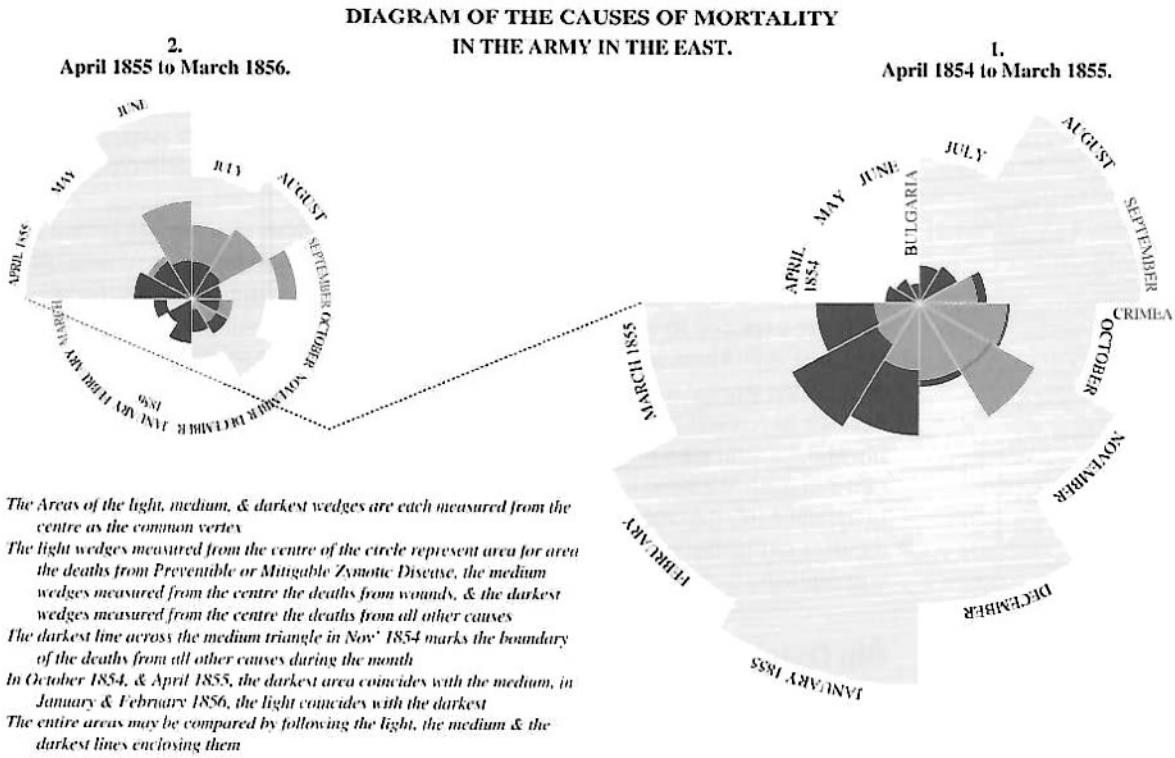
Distilling actionable knowledge from the mass of data is a bit like trying to drink from a firehose.

The sheer volume of the data sets we work with in predictive analytics are often, but not always, much larger than what we have used with time series models and demand planning models in previous chapters. While the size of many of the demonstration data sets we will use in the next few chapters are not appreciably larger than we have used in previous chapters, the algorithms we employ are often designed specifically to work with large data sets with reasonable calculation times. The reason for the emphasis on volume in predictive analytics is clear: while we used to have little data to work with and some firms were lax about saving everything that could be considered data, today firms have so

[4] Helen Joyce, "Florence Nightingale: A Lady with More Than a Lamp," *Significance*, December 2008, pp. 181–182.

FIGURE 8.3 **Nightingale Rose (*www.hugh-small.co.uk*).**

Source: Nightingale, Florence, *Notes on Matters Affecting the Health, Efficiency, and Hospital Administration of the British Army. Founded Chiefly on the Experience of the Late War. Presented by Request to the Secretary of State for War*. London: John W. Parker, 1853.



The Areas of the light, medium, & darkest wedges are each measured from the centre as the common vertex
The light wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic Disease, the medium wedges measured from the centre the deaths from wounds, & the darkest wedges measured from the centre the deaths from all other causes
The darkest line across the medium triangle in Nov' 1854 marks the boundary of the deaths from all other causes during the month
In October 1854, & April 1855, the darkest area coincides with the medium, in January & February 1856, the light coincides with the darkest
The entire areas may be compared by following the light, the medium & the darkest lines enclosing them

much data that distilling actionable knowledge from the mass of data is a bit like trying to drink from a firehose. The data scientist's job is to isolate the important patterns in this mass of data so that a firm can take actions to their benefit.

Analytics

The term *analytics* is used in a number of different ways; your authors use it to encompass all three areas of prediction: time series analysis, demand planning, and data/text mining. However, the term *analytics* is sometimes used to refer simply to the latest forms of prediction: data mining and text mining and their many variants. Data mining refers to the tools and techniques that are used in the large scale, or big data, arena. In the physical world, we have used tools such as the telescope or the microscope to see the characteristics of objects we were unable to examine with the naked eye. More recently, the radio telescope (a nonoptical instrument that is the combination of an antenna system and a radio receiver) and the electron microscope (which uses a beam of accelerated electrons for illumination) have allowed us to examine physical data in dimensions we could only previously

hypothesize about. In much the same way, data mining tools allow us to examine big data in order to make sense of what was previously unable to be seen; we can even discover patterns that were previously unknowable. Hal Varian, a Google data scientist, has termed this new area of statistics "the sexy profession."[5] What Varian has seen in the past decade is a fruitful relationship between statisticians and computer scientists working in machine learning. Together, those two groups are finding new ways to add value to data; when actionable information can be distilled from data, businesses and their customers stand to benefit. According to Varian, data mining applied to huge amounts of largely Internet-acquired data sets presents a whole set of new powers to the data scientist.

With data mining, the models we have used in previous chapters are turned a bit on their head. In data mining, we don't know what pattern or family of patterns may fit a particular set of data. It is not as simple as suggesting that specific patterns are expected to exist in the data, such as trend, seasonality, and cyclicality. We don't even know sometimes what it is we are attempting to predict or explain. This seems strange to a traditional forecaster; it's not the method of attacking the data we have been pursuing throughout our examination of time series forecasting and demand planning. To begin data mining and using big data, we require a new mindset. We need to be open to finding relationships and patterns we never imagined existed in the data we are about to examine. To use data mining is to let the data tell us the story (rather than to impose a model on the data that we feel will replicate the actual patterns in the data).

Big Data and Its Characteristics

Data mining traditionally uses very large data sets, oftentimes far larger than the data sets we are used to using in most business forecasting situations. But big data is not simply defined by the size of the data set. Think of big data as having four characteristics. Each of the characteristics begins with the letter V.

Volume
Velocity
Variety
Value

"Volume" of course refers to the size of the data set. And while some of the data sets used in analytics are quite large, not all of them need be so large in order to produce good usable results. What has changed in the last decade or so has been the "datafication"[6] of almost everything. Text has been turned into data, and we will examine how that takes place and how to analyze textual data in a later chapter. But voice (i.e., speech) has also been turned into data; "speech to text" tools are on

[5] "Hal Varian and the Sexy Profession," *Significance*, March 2011, pp. 32–34.

[6] "Datafication" is a term that has not yet made it into any dictionary we are aware of, but we first saw it used in Victor Mayer-Schonberger and Kenneth Cukier, *Big Data: A Revolution That Will Transform How We Live, Work and Think*. New York: Houghton Miffin Harcourt, 2013, p. 77.

almost every computer, tablet, and phone. And once that speech is text,it is subject to datafication. You are probably familiar with the device called the Amazon Echo; it is a digital assistant that allows you to give it voice commands and it responds with information gleaned from the Internet. You "wake up" the Echo by calling on Alexa, the disjointed voice that answers your questions, provides directions, and responds to a very wide set of commands. Notice the announcement Amazon carries on its website that explains how to set up your Echo device (Figure 8.4):

FIGURE 8.4 Amazon's Disclosure of Alexa Properties http://alexa.amazon.com/spa/index.html#welcome

Source: "Alexa Terms of Use," Amazon. June 23, 2017.

Amazon's Disclosure of Alexa Properties

> Amazon processes and retains audio and other information in the cloud to provide and improve our services, and may exchange information with third party serivces to fulfill your requests. Learn more. Alexa also allows purchasing by voice using your default payment and shipping settings. You can require a speakable confirmation code, turn purchasing off, and see product and order details in your Alexa app. Learn more.

Alexa responds to her name; when it is voiced, Alexa begins processing the next question or comment and gives a response. That means that the Echo's microphone must be on all the time; it's always listening for the "Alexa" wake-up word. Notice that Amazon in their statement says they retain audio in the cloud. Does that mean that everything you say, whether preceded by the codeword "Alexa" or not, is recorded by Amazon? The answer is probably yes, and even more importantly, that information is shared with other firms (i.e., the "third party services" mentioned). In return for Amazon's datafication of everything you utter, you do get some value. Your questions are answered; you can check the latest baseball scores; unfamiliar words can be defined; music can be chosen and played; the latest news can be read to you from a chosen source (you can choose the "spin" you like).

How long does Amazon keep this data? What form do they keep the data in? How do they share it with "third party services?" We don't know the answer to those questions, but we can certainly see that if they collect and store this data for every Alexa user, it amounts to a large volume of data. But Amazon is not alone in collecting "datafied" information. "Google Now," "Siri," and "Cortana" all act as digital assistants, and presumably they all collect data and likely share that data like the Echo. That is a lot of data!

That brings us to our second big data characteristic; the velocity or the rate at which the data arrives is also big in one sense. Each of those digital assistants operates in real time; we expect the assistant to be instantly available and to perform its task quickly. If more than a few seconds' elapses before we get our information, we are disappointed. Some of the algorithms we use in analytics will have to operate very quickly indeed if we are to provide usable results in real time. When you swipe a credit card or use Apple Pay or Android Pay, a data mining algorithm is used to check whether the

transaction initiated is legitimate; that algorithm must work quickly unless you expect to wait for more than the few seconds it usually takes to spit out your receipt indicating a completed (and likely legitimate) transaction. We did not mention the speed of any of the forecasting algorithms because it is rarely an issue, but the speed of a data mining algorithm may be relevant, especially with streaming data.

Unstructured data does not have a predefined data model.

The third characteristic of big data is variety. In the preceding chapters, we have dealt with numbers arranged in a database-like format, usually an Excel spreadsheet. But the data available now is more than just numbers. Social media posts, video, and audio are all data; they are unstructured data. Our databases and Excel spreadsheets were uniformly structured; the items were numbers, and they were arranged in a particular pattern within the database. Unstructured data does not have a predefined data model, nor is it organized in a particular manner. Social media posts are a good example of unstructured data. The posts are not exclusively numbers but also contain text. The posts are not arranged in a rigid format; posters may say anything they want in any order, often mixing words and phrases with numbers and punctuation. But the social media posts are data; they can be analyzed. In the chapter on text mining, we will learn exactly how that is done. For now, it is enough to know that data may vary in the characteristic of variety.

Finally, the fourth characteristic of what we call big data is value. Not everything that qualifies as data is of value to a given firm. But a great deal of what may not have been considered data in the past is of value to the firm today. It is the task of a data scientist to sort out the value that might be gleaned from what appears to be otherwise uninteresting "stuff." Consider the case of Luis von Ahn. You may have seen van Ahn on a *Nova* special on public television or read about him in *Wired* magazine. Shortly after graduating from college, he created something you have seen and used many times: he is the inventor of the CAPTCHA box (Figure 8.5). CAPTCHA stands for Completely Automated Public Turing Test to Tell Computers and Humans Apart. It is the little box that pops up sometimes when first entering a website and ostensibly is meant to ensure that the entrant is really a human and not a spambot or computer. This was just the first invention von Ahn became famous for creating. Have you ever used the free program called Duolingo on the Internet? It's a language learning device that doubles as a proficiency exam for various languages (more than 20 languages at last count). Millions of people have learned or brushed up on a foreign language with Duolingo. Von Ahn is quite clever in some of his creations; he is able to create value in a seemingly innocuous manner.

In a Ted Talk Luis von Ahn discusses the CAPTCHA boxes and the reCAPTCHA boxes as well as Duolingo. You might want to watch this Ted Talk at: https://www.ted

FIGURE 8.5
An Example of One of van Ahn's ReCAPTCHA Boxes.

Source: Google

.com/talks/luis_von_ahn_massive_scale_online_collaboration#t-314305. You would also find exploring https://www.duolingo.com interesting as an example of a useful application of these concepts. It may even help you in a foreign language class.

"Datafication"

Take the case of the CAPTCHA box. The successor to the original version is marketed by von Ahn's successor as ReCAPTCHA. In the newer ReCAPTCHA, the user upon attempting to enter a website is faced with a box containing two random words that are taken from a computer-scanned document. The user is requested to type those words in the provided box. Doing so correctly allows the user to pass through to the website. The entire process takes only moments of your time. But how does von Ahn gain or create value from providing this free service to websites that wish to exclude spambots? Google explains the value creation this way:

> "reCAPTCHA improves our knowledge of the physical world by creating CAPTCHAs out of text visible on Street View imagery. As people verify the text in these CAPTCHAs, this information is used to make Google Maps more precise and complete. So if you're a Google Maps user, your experience (and everyone else's) will be even better."[7]

By taking a few moments of your time, von Ahn is provided with the value of your expertise to help refine a character recognition algorithm. That algorithm will help to "datify" documents and decipher unclear words in digitized text, no matter where it exists. If you have ever filled in a ReCAPTCHA box, pat yourself on the back for helping with the effort to make clear what was previously unclear!

A great deal of what we have previously considered as "stuff" with little or no value to the firm has become data with some inherent value. Comments and reviews were once considered interesting and viewed one at a time provided some, but very little, value to a firm. Ingested by the hundreds of thousands, they become valuable data from which insights may be derived. Companies that excel in "datafication" such as Amazon and Google benefit in very tangible ways when they can predict more ably than their competitors. It could be reasonably argued that all data will become valuable at some point. Data storage costs have plummeted, and our ability to draw predictions from data increase with every new data mining algorithm. The data we have now may not be valuable, but if we keep it long enough, there are likely some valuable insights to be mined from it. Data has the curious characteristic that when it is used and creates value, the process does not diminish the value that is left. The same data may be used over and over: economists say a good such as data has the property of "nonrivalry" so that one person's use of the good to create value does not diminish the value another can extract from the data. Most material goods do not have that characteristic, and so it is worth noting that data is different. The first use of a particular set of data by a data scientist may be quite different from the purpose another researcher may employ, and yet both may find the data to add value to the firm. Data's full value is rarely extracted with its first use.

Data has the property of "nonrivalry" so that one person's use of the good to create value does not diminish the value another can extract from the data.

[7] "Creation of Value," Google. https://www.google.com/recaptcha/intro/#creation-of-value

Not only is the data itself somewhat different in analytics, the tools or algorithms we use are also somewhat different than standard business forecasting tools of time series and demand planning; some of the data mining tools will seem familiar, but they may be used in different ways than we have used them in previous chapters. The premise of data mining is that there is a great deal of information locked up in any database; it's up to us to use appropriate tools to unlock the information hidden within.

Business forecasting is explicit in the sense that we used specific models to estimate and forecast known patterns (e.g., seasonality, trend, cyclicality, the effects of advertising, etc.). Data mining, on the other hand, involves the extraction of implicit (often unknown) intelligence or valuable information from data. We need to be able to process very large quantities of data to find patterns and regularities that we did not know existed beforehand. Some of what we find will be quite useless and uninteresting (at the moment), perhaps only coincidences. But, from time to time, we will be able to find true gems in the mounds of data; the objective of this chapter is to introduce a variety of data mining methods for you to consider. Some of these methods are simple and meant only to introduce you to how the basic concept of data mining works. We will leave the more commercially used tools for the following chapters.

If you wish to work with your own data (or that provided with this text), we recommend the Analytic Solver® Data Mining software.[8] Everyone capable of using an Excel spreadsheet will find Analytic Solver® Data Mining an excellent introduction into actually using most of the algorithms used by data miners. Just as with time series models and demand planning models, the best method for mastering predictive analytics is to work through the examples and algorithms yourself with the aid of proficient software tools such as Analytic Solver®.

DATA MINING

Not long ago one of the most pressing problems for a forecaster was the lack of data collected intelligently by businesses; forecasters were limited to few pieces of data and only limited observations on the data that existed. Computing power was also limited, but the real shortage was a lack of data. Today, however, we are overwhelmed with data. It is collected at grocery store checkout counters, while inventory moves through a warehouse, when users click a button on the World Wide Web, and every time a credit card is swiped. The rate of data collection is not abating; it seems to be increasing with no clear end in sight. The presence of

[8] Analytic Solver® Data Mining is an Excel add-in that is part of an integrated analytics solver from FrontlineSolvers (http://www.solver.com/). Both student and full versions of the software are available from FrontlineSolvers. It provides an excellent way to learn about data mining by applying the algorithms to medium-sized data sets within an Excel setting. A version of the tool is available for instructors who adopt this text to make available to their students. Instructors should contact FrontlineSolvers (http://www.solver.com/) for more information. The Analytic Solver® Data Mining software is provided with a complete manual (see the "help" menu) titled "Analytic Solver Data Mining User Guide." This user manual contains complete step-by-step examples of the procedures presented in this text.

large cheap storage devices means that it is easy to simply keep every piece of data produced. It may even be prudent to keep every piece of data produced; while not valuable for insights now, that could change in the future. The pressing problem now is not the generation of the data, but the attempt to understand it.

The job of a data scientist is to make sense of the mounds of data we now have available by probing the data for patterns. The single most important reason for the current fascination with data mining is due to the large volumes of data currently obtainable for analysis; there is a need for business professionals to convert data into useful information by "mining" it for the presence of patterns. You should not be at all startled by the emphasis on patterns; this entire text has been about patterns of one sort or another. Indeed, humans have looked for patterns in almost every endeavor undertaken by humankind. Early humans looked for patterns in the night sky, for patterns in the movement of the stars and planets, and to predict the best times of the year to plant crops. Modern humans still search for patterns in early election returns, in global temperature changes, and in sales data for new products. Over the last few decades, there has been an evolution from data processing to what we call data mining today. In the 1960s, businesses customarily collected data and processed it using database management techniques that allowed indexing, organization, and some query activity. Online transaction processing (OLTP) became routine, and the rapid retrieval of stored data was made easier by more efficient storage devices and quicker and more capable computing.

Database Management

Man has looked for patterns in almost every endeavor undertaken by mankind.

Database management advanced rapidly to include very sophisticated query systems (SQL or Structured Query Language is one commonly used example); it became routine not only in business situations but also in scientific inquiry. Databases began to grow at previously unheard of rates and for even routine activities. The volume of data in all the world's databases has been estimated recently to double in less than every two years.[9] That torrent of data would seem to call for analysis in order to make sense of the patterns locked within. Firms now routinely have what are called data warehouses and data marts. "*Data warehouse*" is the term used to describe a firm's central repository of integrated historical data; it is the "memory" of the firm, collective information on every relevant aspect of what has happened in the past. A "*data mart*," on the other hand, is a subset of a data warehouse; it routinely holds information that is specialized and has been grouped or chosen specifically to help companies make better decision on future actions.

Data Mining Versus Database Management

Data mining as a term for many years had an altogether different connotation than it enjoys today; instead of being an analysis that finds useful patterns in data, it carried the meaning that the researcher was imposing a model on data, whether

[9] "Extracting Value from Chaos," study sponsored by EMC (June 2011). The multimedia content may be viewed at https://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf.

Data mining is the extraction of useful information from large, often unstructured databases.

it fit or not. It was a notably derogatory term. When someone was called a "data miner," it was meant to be an insulting term applied to a person who tortured data until it "told" the preconceived story the researcher wanted to tell. The term *data mining* today denotes the analysis of databases, data warehouses, and data marts that already exist for the purpose of discovering new patterns or to answer some pressing question. Data mining is the extraction of useful information from large, often unstructured databases; it is about extracting knowledge or information from large amounts of data.[10] Data mining has come to be referenced by a few similar terms; in most cases, they all refer to much the same set of techniques that we will refer to as data mining in this chapter:

- Machine (or supervised) learning
- Business intelligence/analytics/analysis
- Data-driven discovery
- Knowledge discovery in databases (KDD)

Data mining is, however, quite separate from database management. Keogh points out that in database management, queries are well defined; we even have a language to write these queries (Structured Query Language or SQL, pronounced as "sequel"). A query in database management might take the form of "find all the customers in South Bend" or "find all the customers that have missed a recent payment."

Data mining, however, uses very different queries; they tend to be less structured and are sometimes quite vague. For example: "Find all the customers that are likely to purchase recreational vehicle insurance in the next six months" or "group all the customers with similar buying habits." In one sense, data mining is like statistical forecasting in that we are forward-looking in an attempt to obtain information about future likely events and drive better decision-making.

Many companies are data rich, but some of those same companies are information poor; data mining is the set of algorithms and techniques that can aid firms in making sense of the mountains of data they likely already have available. These available databases may be about customer profiles and the choices those customers have made in the past. There are possible patterns of behavior displayed in the data, but the sheer amount of the data will mask the underlying patterns and even an expert researcher, testing for patterns she believes will be exhibited in the data, will miss a great deal of the information locked within. Some of those underlying patterns may be interesting but unusable to a firm for informing future decisions, but some patterns may be predictive in ways that are very worthwhile to firms. If, for example, you "know" which of your customers are likely to switch their supplier in the near future, you may be able to prevent the customers from jumping ship and going with your competitor; it's always less costly to keep existing customers than to enlist new ones.[11] The evidence shows

[10] D. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*. Cambridge, MA: MIT Press, 2001.

[11] Amy Gallo, "The Value of Keeping the Right Customers," *Harvard Business Review*, October 29, 2014 (https://hbr.org/2014/10/the-value-of-keeping-the-right-customers).

that it is from five to 25 times more expensive to obtain a new customer. If you were to "know" which customers were likely to default on their loans in the near future you might be able to take pre-emptive measures to forestall the defaults or you might be less likely to try and enlist such individuals in the future. If you "know" (i.e., are able to predict) the characteristics of potential customers that are quite likely to continually purchase your product or service, you would be better able to direct your advertising and promotional efforts than if you were to blanket the market with advertising and promotions; a well-targeted approach is usually better than an unknowing "shotgun" approach. Data mining can help to define the appropriate target.

Patterns in Data Mining

What types of patterns can be mined? The answer is quite different from the patterns we expected to find in data with time series forecasting methods such as the Holt Winters' smoothing model. When a forecaster applies a Holt Winters' smoothing model to time series data, the expectation is that the data contain some level variation, some trend, and some seasonal variability. Experience with business data has taught us to expect those pattern types in virtually all business time series data.

Data mining, however, does not prespecify the patterns to be expected. In a sense, there is no preconception of what will be found in the data with most data mining techniques. We are simultaneously searching for several different kinds of patterns in parallel. At the same time, we are measuring the certainty or trustworthiness associated with the patterns we discover in somewhat the same vein as we do in standard business forecasting.

THE TOOLS OF ANALYTICS

Shmueli, Patel, and Bruce use a taxonomy of analytics tools that is useful for seeing the big picture. There are basically four categories of analytics tools or techniques; they represent the four very general types of patterns we would like to search across:

1) Prediction;
2) Classification;
3) Clustering (sometimes called segmentation); and
4) Association.

Prediction tools are most like the methods we have covered in previous chapters that dealt with time series models and demand planning; these tools most often attempt to predict the value of a numeric variable (e.g., sales or shipments). We might, for example, be attempting to predict the value of a piece of residential property or the amount that an individual might contribute yearly to a particular charity. The variable we are attempting to predict in these instances could be a continuous variable. But the variable to be predicted could also be a

categorical variable. For example, we might wish to predict whether an individual will contribute to a particular cause rather than how much they might contribute or whether an individual will make a certain purchase this year rather than how much they likely will spend on the purchase. Prediction then involves two types of variables: continuous and categorical.

Classification tools are the most commonly used methods in data mining. Classification tools distinguish between data classes or concepts; the purpose is to create a model that allows us to predict a class of objects whose label is unknown to us. For instance, when you present your credit card for a purchase in a retail store, the business must determine whether the impending transaction is a legitimate one (i.e., you are who you say you are, you have requisite purchasing power, etc.). Businesses do not find it profitable to hand over merchandise to everyone who presents a piece of plastic at the checkout counter. For this reason, there is a short delay between when your card is swiped and when the receipt begins to print. In those few seconds, a data mining algorithm's rules have been applied to your situation and a determination has been made; either the transaction is in the classification of "legitimate," or it is in the classification of "illegitimate."

Another example of classification involves banking. When you apply for a bank loan, your credit score is calculated based upon some personal characteristics, financial characteristics, and personal characteristics such as age and family status. That calculated credit score alerts the bank to the risk associated with making a loan to you; the bank is attempting to classify you as a creditworthy customer or an individual they would prefer sought a loan elsewhere.

Sometimes the classifications we are trying to make do not involve numbers at all; the U.S. Post Office must read the destination you place on a letter (including the street name, city, and the zip code). If you have handwritten the information on the envelope, it could be read by a human, but the Postal Service has automated the process by allowing a data mining algorithm to take the scan of the address and recognize (i.e., classify) the alphabetic characters and numbers, even though a human has handwritten them. Even if the algorithm has never seen your particular handwriting, it is able to recognize the characters; you receive mail every day that has been subjected to such a process.

Clustering (segmentation) analysis tools analyze data objects without consulting a known class label. The classes of the objects are not input by the user, it is the function of the clustering technique to define and attach the class labels. The clustering technique generates the labels. Clustering techniques group objects based upon maximizing the intraclass similarity and/or minimizing the interclass similarity. Whether the clusters unearthed by the techniques are useful to the business manager is subjective; some clusters will be interesting but not useful in a business setting, while others will be quite informative and will also be able to be exploited to advantage. Universities use cluster analysis to identify students with special needs; using the characteristics of a given student's background, the university is able to cluster the

student with others who require some specialized attention or above-average support in order to be successful.

Association rules discovery is sometimes called "affinity analysis." It is the discovery of rule attribute characteristics that often occur together in a given data set. If you have been handed coupons at a grocery store checkout counter, your purchasing patterns have probably been subjected to association rules discovery; Netflix will recommend movies you might like based upon movies you have watched and rated in the past. In each instance, an association rules discovery has taken place. Amazon will offer items that you have not selected but that they believe you may wish to purchase as a result of their affinity analysis.

STATISTICAL FORECASTING AND DATA MINING

Data mining allow the data itself to reveal the patterns within, rather than imposing the patterns on the data at the outset.

In time series forecasting and demand planning, we sought verification of previously held hypotheses; that is, we "knew" which patterns existed in the time series data we tried to forecast and we applied appropriate statistical models (e.g., Holt Winter's or regression) to accurately estimate those patterns. When an electric power company looks at their load demand, they expect that past patterns, such as trend, seasonality, and cyclicality, will replicate themselves in the future. Thus, the electric utility might reasonably use a regression with independent variables such as the previous day's temperature, the hour of the day, the day of the week, and the month of the year as a model to forecast future electric usage. Data mining, however, seeks the discovery of new knowledge from the data; it does not seek to merely verify the previously chosen hypotheses regarding the types of patterns in the data but seeks to discover new facts or rules from the data itself. Data mining allow the data itself to reveal the patterns within, rather than imposing the patterns on the data at the outset.

TERMINOLOGY IN DATA MINING: SPEAK LIKE A DATA MINER

The terminology used in data mining is a bit different than that used in statistical forecasting models; while the terms are different, their meanings are quite similar.

Data Mining Terminology	Statistical Terminology
Output variable = Target variable	Dependent variable
Algorithm	Forecasting Model
Attribute = Feature	Explanatory variable
Record	Observation
Score	Forecast

Courtesy of Eamonn Keogh.

The subject of our interest in time series forecasting models and demand planning models was termed the dependent variable or the Y-variable. It referred to the value we were forecasting. In a data mining world, the corresponding term would be the "target." In some data mining algorithms, we specify the target variable, but in a few algorithms, there is no analog of the dependent variable and thus no target. Those variables that contributed toward forecasting the dependent variable were called explanatory variables in our previous models. Their analog in data mining would be "attributes." In forecasting proper, we spoke of using models (e.g., Holt Winters' model) but in data mining, the proper term is *algorithm*. Our data sets in forecasting were made up of numbers of observations, but in data mining, we refer to "records." Finally, our goal in using forecasting models was to make forecasts. In data mining, that process is called "scoring."

CORRELATION

One term that means the same whether speaking either about data mining, time series forecasting, or demand planning is *correlation*. As explained in Chapter 2, correlation is the degree of association between two variables or two data sets, whether causal or not. The Pearson product-moment correlation coefficient usually designated by ρ (rho) for a population and r for a sample, measures not only the direction of association but also its magnitude.

Data mining thrives in an environment that is not data starved, but rather data rich.

With forecasting data, we worked hard to reduce error rates in our models and spent a great deal of effort in sampling, testing samples for bias, and ensuring that our data represented the true population we were ultimately representing. With the move toward using big data, we are going to worry much less about whether our data in the remainder of the text represents the true population; in some cases, it is the population (or close to it). Data mining thrives on an environment that is not data starved but rather data rich. The quantity of data is now so vastly abundant compared to just a few decades ago that the tools we use to gain information from it must recognize the change in the situation. Individual data points are such a small part of the big picture that we will spend much less time worrying about whether they will bias the final result. In some cases, even incorrect values and outright mistakes in individual observations may be relatively unimportant in determining the final results. Much like the "wisdom of crowds"[12] eliminates errors because of compensating mistakes, big data provides a vast mesh of data in which a few mistakes will not affect the outcomes predicted by the preponderance of the data.

There is nothing inherently messy or mistake prone in big data; it's just that the data scientist may be a bit less troubled by mistakes or messiness because of the

[12] James Surowiecki, *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. New York: Knopf Doubleday Publishing Group, 2004.

sheer volume of the data. Logic would say that spending great effort and expense to correct mistakes in the big data may make little economic sense unless the results would appreciably change. The mantra of the data scientist may be "more data is better than less." With more data and lots more than we have used in any analysis up to this point, we can put aside the interest in larger and more random samples and live with some mistakes and missing observations. Conventional forecasting based on frequentist statistics is increasingly at odds with reality. The age of big data has arrived and along with it data of varying types (numbers, text, video, audio, etc.) and varying quality. Not only will this big data not fit into neatly defined categories, but the answers we seek may not even be visible before we begin the analysis. The data we analyzed and predicted in previous chapters was often created with SQL; it was by definition "structured" to begin with and rigid. Our shift in this and the remaining chapters is to accept data of varying types, sizes, and messiness. The plummeting costs of data storage and processing power are pushing the industry toward computationally intensive predictions.

Early Uses of Analytics

Everyone reading this has experienced suggestions from a website, in their morning email or from an in-store reminder app like those used by McDonald's, Burger King, and Walgreens. Sometimes these suggestions seem to be reading your mind, and at other times they just feel intrusive. They don't have to be on target all the time; they just have to be predictive some of the time (and they are). One of the early developers of the "recommendation engine," for want of a better term, is Greg Linden. He's written a blog about the early (1997) days he spent at Amazon.[13] In a rundown brick structure near Pike's Place Market in Seattle, he started working for the small Internet book seller (i.e., Amazon); in 1997, the concept of online book sales, or sales of anything online, was quite novel. Nobody knew if this would work. Did people really want to purchase anything online?

Linden started as a coder assigned to projects that placed him in contact with the early data collected by the small online company; he worked first on discounts. We would probably call it dynamic pricing these days. At some point, Linden began the early work on recommendations with BookMatcher, but he soon created an altogether different recommendation engine that operated in real time and was scalable. Based on what people bought, the engine started making recommendations of other things they might like to buy. What was originally done by humans at Amazon actually worked in real time; unfortunately, most of those humans lost their jobs when it became clear that the data driven recommendation engine could do a better job (and probably at lower cost). "According to Sucharita Mulpuru, a Forrester analyst, Amazon's conversion to sales of on-site recommendations could be as high as 60 percent in some cases based on the performance of other e-commerce sites."[14] So, if you are expecting fewer recommendations in

[13] http://glinden.blogspot.com/2006/05/early-amazon-end.html.

[14] Mangalindan, J. P. "Amazon's Recommendation Secret," *Fortune*, July 30, 2012 (http://fortune.com/2012/07/30/amazons-recommendation-secret/).

the future, you are probably out of luck; the recommendation systems work and increase the profits of the companies that use them. Some of us actually appreciate the service from time to time.

Correlation becomes a sledgehammer in the hands of a data scientist who has access to big data.

But how does any recommendation engine work? At the core of these algorithms is the concept of correlation, the same correlation we covered in Chapter 2. Correlation is certainly a useful tool when working with samples of data, but correlation becomes a sledgehammer in the hands of a data scientist who has access to big data. According to Linden, these engines are often based on a clustering algorithm that, in part, has its basis in correlation. Correlation is not causation. How can it be used for making suggestions?

The "mechanism of action" that would tell us exactly why someone might purchase something would be very valuable to know and allow us to build recommendations. But in the real world, we do not need to know the mechanism of action; we only need to know when two things are related. With big data and correlation, there is no certainty, just probability. As our calculated probabilities become better than guesses, our recommendations become more and more useful. With large amounts of data and correlation, we can find links that are useful; Amazon does it every day to great advantage. In 2004, Hurricane Frances was headed toward the Florida coast and Walmart was tracking the storm. A week before the hurricane hit, Walmart's CIO requested information on what had happened at Walmart stores when other hurricanes had occurred. She was looking for correlations. What correlated with a hurricane? The shopping history from Walmart stores indicated that Pop-Tart sales spiked up to seven times the normal rate when a hurricane was imminent. Walmart started predicting what to do based on correlations. The decisions they made to change inventories certainly benefited Walmart, but they also provided a welcome service to individuals in the affected area.[15] With 3,600 Walmart locations and over 100 million customers per week in those stores in 2004, Walmart had plenty of data with which to make correlation calculations.

Walmart probably didn't have a hypothesis that described why they were seeing spikes in Pop-Tart sales; they just needed the correlation to create a clear case for corporate action. When we examined regression models, we were building causal models; we assumed that changes in our independent variables were causing the changes we observed in the dependent variable. We had good reason to believe that these causal models were accurate; economists, for example, call the causal relationship between price and sales "the law of demand." The models we will describe in the coming chapters, however, do not imply causality and often do not rely on any hypothesis such as

[15] Constance L. Hays, "What Wal-Mart Knows About Customers' Habits," *The New York Times*, November 14, 2004 (http://www.nytimes.com/2004/11/14/business/yourmoney/what-walmart-knows-about-customers-habits.html?_r=0).

the law of demand. We have been very concerned with exactly what variables to include as independent variables in a regression; the AIC and BIC were described as tools that would help in the selection of appropriate independent variables. But with big data available, it may not be necessary to spend as much time on selecting which variables to examine. It is correlation that is going to be the basis of our predictions.

THE "STEPS" IN A DATA MINING PROCESS

One of the important commercial analytics software companies, SAS, has an approach they teach users that are embedded in their software as a series of steps. The tabs in SAS Enterprise Miner are labeled *SEMMA*, which is an acronym for *Sample*, *Explore*, *Modify*, *Model*, and *Assess*.[16]

The *explore* activity involves all those activities we referred to as data cleansing in earlier chapters. Missing data must be handled intelligently. The collected data may contain mistakes such as values outside reasonable ranges. People represented in the data may be referenced in a number of ways; Joseph Smith might also be listed as Joe Smith and J. Smith. Are they the same person? We don't wish to count them three times as separate individuals if we can correctly combine the records and attribute all the characteristics to a single individual. Measurements in the data are important; some of the techniques are topographical, and consistency may be important. Time periods, if used, should be understood. The explore activity may involve examining the data graphically, creating summary statistics of the attributes, or running an "audit" of the values in the data. In some cases, the data may need modification, but that will first require a clear understanding of the data as it is.

Modify may involve creating, selecting, or transforming data. Part of the modification may be eliminating rare variables, eliminating records with missing variables, replacing missing values, or modifying the data set attributes in some manner.

The *model* phase of this methodology is what most probably think data mining is all about. In the model phase, your goal will be to determine the data mining task required (classification, prediction, clustering, etc.), select an appropriate algorithm, and set whatever parameters are necessary to execute the process. Understanding exactly what the algorithm does is an important part of making the correct selection; the following chapters will detail the processes used by the most important data mining (and text mining) algorithms. It is in the model phase that we fit the predictive analytics algorithm by training the model with a portion of the data.

The final step is to *assess* the results obtained. It may involve making a choice among a number of different algorithms chosen as candidates for the analysis.

[16] SAS Institute. *Applied Analytics Using SAS Enterprise Miner*. Carey, NC.: SAS Institute, Inc., 2011, pp. 1-7–1-9.

Testing the final model is performed by data scientists in a manner that mimics our use of a "holdout" or "holdback" in traditional forecasting. Because most of the data we use in data mining is not time series data, it is possible to make many partitions of the data; some we use for training the algorithm, and others we use for validating the results.

THE DATA ITSELF

Beginning a project in analytics will always involve some preliminary steps before we can begin the more interesting portion of the analysis; these beginning steps are not optional and will be performed each time we begin to analyze new data. Data scientists always perform their data analysis on less than the total number of records that are available. This will allow the validation of whatever algorithm is chosen. Different algorithms will also have varying limitations on what types of variables they are able to handle and the numbers of records and attributes. The software used and the computing power may also be a consideration on the amount and type of data that can be handled.

The types of variables used differ from traditional forecasting because there are more types available in unstructured data. There are several ways of classifying variables. Variables can be numeric or text (character). The data can be continuous (taking on any real numeric value), integer values, or categorical (taking on one of a limited number of values such as "1," "2," or "3"). Categorical variables can also be either numeric or text (loan, mortgage, CD, or checking account). Continuous variables like those we have commonly used in forecasting can be handled by most data mining algorithms. The Naive Bayes algorithm will be unusual because it will exclusively accept categorical variables. Because much of what follows involves classification as a goal, some continuous variables will be transformed into categorical variables (data mining software includes routines to handle this process). Unordered categorical variables, however, cannot be used as is. They will commonly be decomposed into a series of dummy binary variables. To do this, you should remember the "iron law of dummy variables," which states that the maximum number of dummy variables must be one less than the states of nature. For example, a single variable that can have possible values of male or female would be split into two dummy variables of which only one would be used in analysis (either one could be used):

Male – 1/0

Female – 1/0

Only one of these variables would be used to represent an individual's gender (just as we only used three dummy variables to represent four quarters of the year). Creating dummy variables from categorical variables is usually done in the software; in text mining, it will become part of the dimension reduction process that is the hallmark of text mining.

Overfitting

If we put too many attributes (or try to account for too many patterns) in a model, including some unrelated to the target, we are *overfitting*.

When we include more attributes in the process, there is always a risk of overfitting the data. What is overfitting? If we put too many attributes (or try to account for too many patterns) in a model, including some unrelated to the target, we are *overfitting*. If we put too few attributes (or try to account for too few patterns) in the algorithm, leaving out attributes (or patterns) that could help explain the target, we are *underfitting*. Data scientists are always trying to balance the possibility of overfitting against the possibility of underfitting.

Suppose we have some data that follow a cubic model:

$$Y = 10 - 8x + 4x^2 + 1x^3$$

The true model is shown in Figure 8.6 as the curved line that looks like a sinoid. The circles represent data points in our data set that have random variations about the true model. The straight line running from upper-left to lower-right is a trend line; note that it does a very poor job of describing the true model.

If we used a cubic function to estimate the true model, we would get a result like that in the left pane of Figure 8.7: the true model and the estimate are quite alike. Alternatively, we could connect up all the data points with a smooth but complicated polynomial function, one that explains all these data points perfectly and leaves no error (residuals). This can be seen in the right-hand pane in Figure 8.7. The model

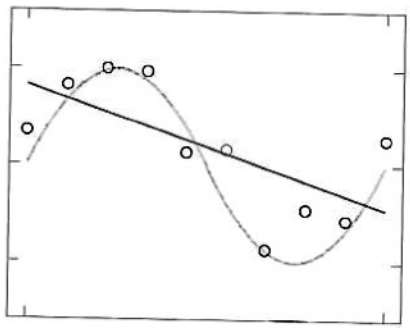


FIGURE 8.6
Fitting Example: The Curved Line Represents the True Underlying Model. The Straight Line Represents a Trend Model. The Circles Represent the Actual Sample Data Points.

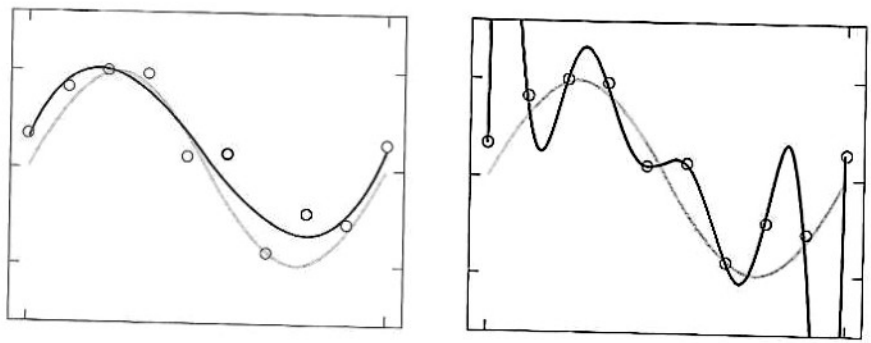


FIGURE 8.7
The True Model (Lighter Line) and a Reasonable Estimate (Darker Line) are Shown on the Left. The True Model and an Overfit Estimate (Dark Line) are Shown on the Right.

estimate goes through every data point, and the fit to the data is thus perfect. However, this estimate would do a very poor job indeed of predicting future iterations of the pattern; it has been overfit. It has fitted the noise as well as the actual pattern in the data and noise, because it is random, cannot be predicted well in a future period.

Accuracy and Fit (Again)

The primary goal in model construction is to describe relationships among attributes so that this description will do a reasonable job of predicting future outcome (target) values on the basis of future predictor (attribute) values. We certainly want the model estimated to do a good job in representing the data we now have (our known data set). But what we want more importantly is for the model to have predictive power outside the known data set. We want "accuracy" and not just "fit" to be a characteristic of the chosen model. But if we model the noise in the data (as well as the true signal), we end up explaining incorrectly some variation in the data that was nothing more than chance variation. We will be guilty of mislabeling the noise in the data as if it were part of the true signal. This is the classic definition of overfitting the model. In such a case, the misclassification rate will be extremely low (maybe even zero) with the known data, and thus the model will appear to be predicting well.

However, this low misclassification rate on the known data is misleading because it includes a representation of what in the known data was actually just noise (random variation unable to be predicted in future data). Some of the data mining models we will examine are so good at classification that they have a natural tendency to overfit the known data (much like ARIMA models). We will have to recognize this and correct for the tendency.

Some Other Data Considerations

Data mining differs from previous models we have used in that many more variables (i.e., attributes) may be used at one time. While we want the attributes to be correlated to the target, we should also be concerned when they are highly correlated with one another. The overfitting problem may be exacerbated by the inclusion of highly correlated variables. Some of the data mining algorithms are highly sensitive to attributes correlated one with another. It may also be costly to collect variables that simply "tell the same story" as variables we already have in the data. "Dimensionality" of a model is the number of attribute or predictor variables used by the model. A standard practice in data mining is finding ways to reduce dimensionality without sacrificing accuracy.

Data mining software often includes an "audit" routine to allow the data scientist to examine different aspects of the data that might otherwise be obscured in a tabular format. In XLMiner©, the user may examine any of the target or attribute variables in a number of dimensions. The Universal Bank data is a data set with approximately 5,000 customers of a bank with a broad range of attributes available for each customer. We might, for instance, be considering an analysis of what the characteristics are of an individual who holds a personal loan with the bank (Figure 8.8). An "audit" of the personal loan variable produces an answer.

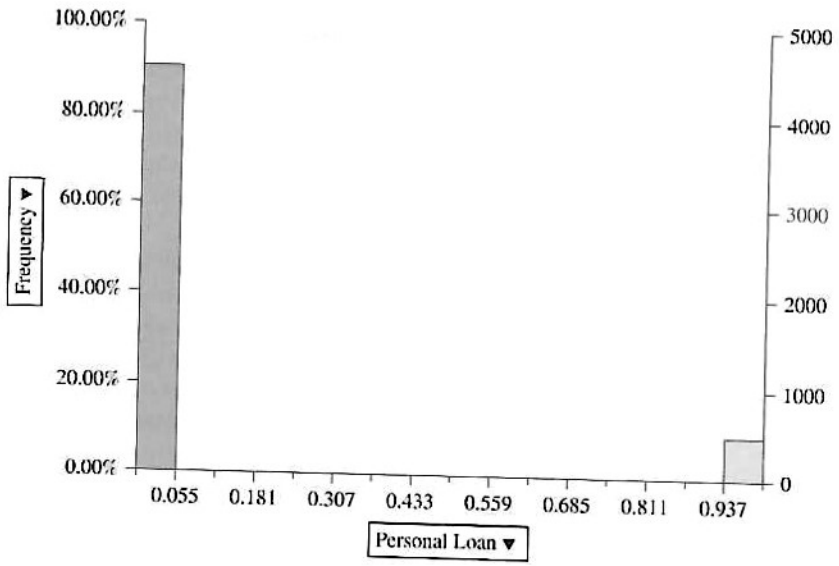


FIGURE 8.8
The Personal Loan Variable from the Universal Bank Data Set Produced Using the "Explore" Feature in XLMiner©.
(C8F8)

Source: Frontline Systems Inc.

The number of individuals who hold a loan with the bank is quite small; it is only about 10 percent of the customer base (about 500 individuals). This fact may be useful when we begin modeling bank customer behavior; if we are interested in predicting customers who will take out a loan, we should realize that only a small portion of the customer base fits into that category. In this particular situation, no missing data or possible outliers were detected, and so no action was necessary, but if either of these two instances occurs, some action will be necessary. Again, the data preparation of your analytics software will contain options for these situations.

The "Transform" drop-down menu in XLMiner© (Figure 8.9) will allow both the handling of missing data as well as the transformation of data that may contain outliers; locating possible outliers may be accomplished with the "Explore" drop-down menu.

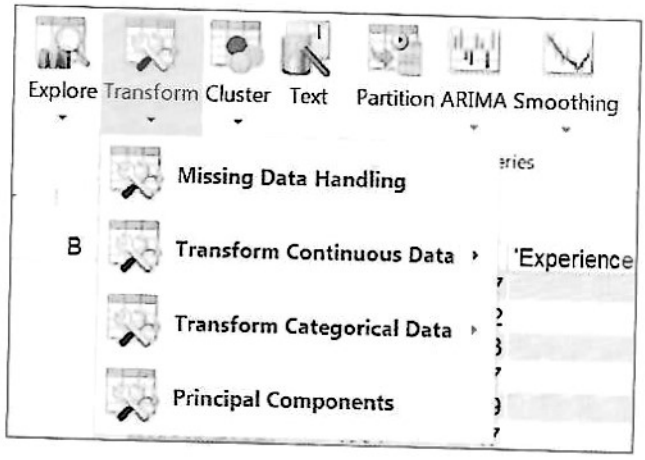


FIGURE 8.9
The Data Handling Section of XLMiner©.

Source: Frontline Systems Inc.

Partitioning and testing for accuracy are standard practice in analytics.

Sampling/Partitioning

In most cases, we will not use the entire data set to build a model; since our data is not time series, we have the ability to "partition" the data, using one partition to build the model and the other partition to test the model's accuracy. This partitioning and testing for accuracy are standard practice in analytics, and it has its roots in the "holdouts" or "holdbacks" we used when testing the accuracy (as opposed to fit) of standard forecasting models.

We can use our Universal Bank data to demonstrate this basic analytics procedure. The entire data set consists of 5,000 records; each record is an individual bank customer. If we build the model using the entire data set, we do not have any effective way of estimating how well the algorithm will work on classifying unseen data. Our goal here is to predict (classify in this case) whether an existing bank customer is likely to take out a personal loan; we will use all the attributes of bank customers to aid in the prediction.

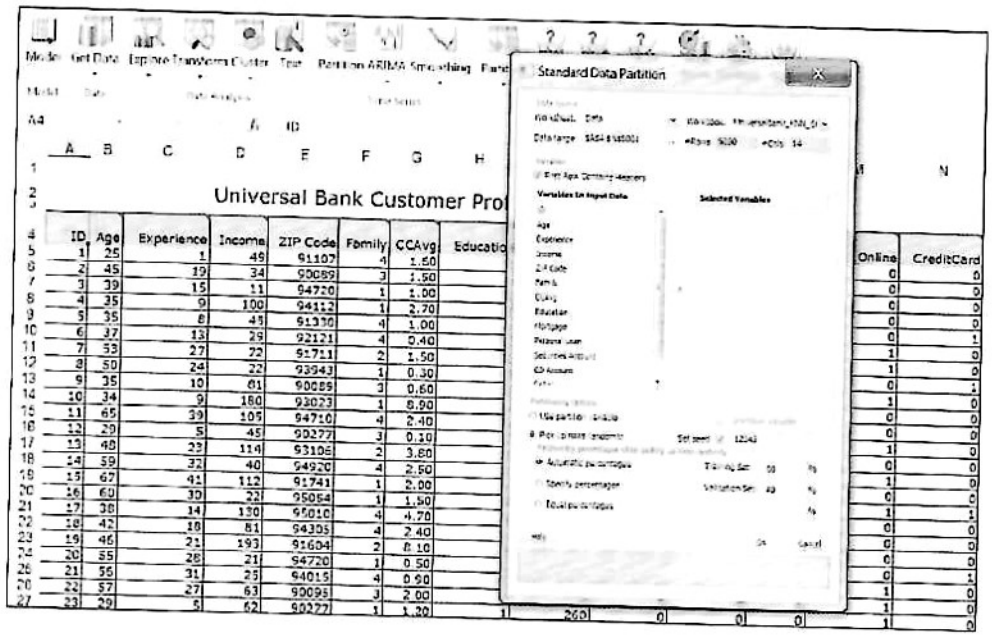
Instead of using the entire data set of 5,000 records, we partition the data into two parts. The partition we use to build the model is traditionally called the "Training" data, while the remainder of the data forms the partition we call the "Validation" data. Since the data is not time series data and the order of the data is not important, we may randomly create the partition. The important point here is why we create the partitions in the first place. The two partitions are created in order to allow the data scientist to gauge the accuracy of the created model. We are interested in comparing the performance of this model with others so that we may choose the one we believe will perform the best when used in actual practice. Why not just choose the model that works best (in terms of fit) on the entire data set? Why partition at all? The answer is that when we use the same data to both build and test the model, we introduce bias. The model chosen in this manner would suffer the chance that some features of the data happen to match the chosen model better than any other model, and that is the main reason this model was chosen. If we are using an algorithm that is prone to overfitting (and some data mining algorithms are), using the entire data set to build the model will almost certainly result in overfitting; the overfitting will result in great "fit" but quite poor "accuracy." Since ultimately it is accuracy we desire in our models, we had best partition and build as a matter of practice.

In XLMiner©, there is a partitioning menu (Figure 8.10) that allows the construction of a standard data partition, including a training data partition and a validation partition; the defaults are set to choose 60 percent of the records for the training partition and 40 percent of the records for the validation partition. The records for each partition are randomly chosen with the random number generator either using a set "seed" number (so that the results may be reproduced exactly) or with a seed number generated by the software. The latter method is chosen in actual practice, but we will use the seed number 12345 so that the results may be exactly duplicated.

The validation portion of the data is not used to build the model and is only used after the model is built to test whether or not it works well on unseen data (i.e., the validation data partition). To test how well the model performs on the validation partition, data scientists use a number of measures (none of which we have used up to this point). The standard diagnostics we have used in past

FIGURE 8.10 **Data Partition Menu in XLMiner®.** (C8F8)

Source: Frontline Systems Inc.



chapters, such as MAPE and R^2, are measures of fit and may not be appropriate to data mining algorithms. If we are classifying a categorical variable, a misclassification rate is one useful measure of model performance. A "lift chart" that reorders the predicted classifications may provide even more information on performance. A "confusion matrix" (also called a coincidence matrix) will provide additional information about model performance.

Diagnostics (Evaluating Predictive Performance)

In the Universal Bank data, we are seeking to predict the state of a categorical variable titled "personal loan." The model will either predict an individual is likely to take out a personal loan given the individual's attributes (as detailed in the data) or the individual is unlikely to do so. The target (what we are predicting) consists of just two categories, and the model will decide which is more likely for every individual scored. In Chapter 9, we will detail the specifics of the model we might use, but for now, we will examine the results of the modeling process. We have obtained the data, cleaned it up to eliminate missing attribute values and adjust outliers, and partitioned the data into training and validation data partitions.

The Confusion Matrix and Misclassification Rate

We normally request diagnostic statistics on both partitions of the data, although the results for the validation partition (the unseen data) are most relevant to evaluating the model.

FIGURE 8.11 Universal Bank kNN Model Validation Partition Confusion Matrix and Error Report where n = 2,000 or 40 percent of the 5,000 Records. (C8F8)

Source: Frontline Systems Inc.

Validation: Classification Summary

Confusion Matrix

Actual\Predicted	0	1
0	1809	3
1	80	108

Error Report

Class	# Cases	# Errors	% Error
0	1812	3	0.165562914
1	188	80	42.55319149
Overall	2000	83	4.15

Figure 8.11 shows the validation partition confusion matrix; recall that this indicates model performance in classification on data that was not used to build the model. A matrix like this one is a standard output in data mining when a classification is the goal. The matrix gives results for the correct (i.e., true) classifications and the incorrect results or misclassifications. These are estimates, but if the model has a sufficiently large data set to examine, the results are likely to be accurate. We would expect the confusion matrix for the training data to provide better results; in this case, it does (see Figure 8.12). Note the figure in the far lower-right

FIGURE 8.12 Universal Bank kNN Model Training Partition Confusion Matrix and Error Report where n = 3,000 or 60 percent of the 5,000 Records. (C8F8)

Source: Frontline Systems Inc.

Training: Classification Summary

Confusion Matrix

Actual\Predicted	0	1
0	2704	4
1	98	194

Error Report

Class	# Cases	# Errors	% Error
0	2708	4	0.147710487
1	292	98	33.56164384
Overall	3000	102	3.4

of the Error Report; this is the misclassification rate on the entire partition. The misclassification rate for the training data partition is a very low 3.4 percent. The same rate for the validation partition is a somewhat higher 4.15 percent. Both are quite low, but the algorithm performs a bit better on the partition that was used to build the model, the training partition, and a bit worse on the validation partition. The estimated misclassification rate is calculated by taking the total records scored incorrectly (80 + 3 in the validation partition) and dividing it by the total number of records classified (80 + 3 + 1809 + 108 in the validation partition).

$$\frac{(80+3)}{(80+3+1809+108)} = 0.0415 \textit{ or } 4.15\%$$

The same calculation could be made for the training partition.

$$\frac{(98+4)}{(98+4+2704+194)} = 0.0340 \textit{ or } 3.4\%$$

If the data set is reasonably large, the misclassification estimates are probably reasonably accurate.

The Lift Chart

The lift chart and its resulting lift calculation is the standard for accuracy in data mining.

Another routinely used diagnostic device is to use a "lift chart." The lift chart and its resulting lift calculation is the standard for accuracy in data mining. A useful way to think of a lift chart is to examine the Universal Bank model that attempts to identify the likely individuals who will take out a personal loan by assigning each case a "probability of responding" score. The lift chart helps to determine how effectively the model can reorder the data set, placing those individuals who will take out a loan at the top of the list and those that are unlikely to do so at the bottom of the list, according to their scores. To construct a lift curve, we use the validation data set after it has been "scored" by appending to each case the estimated probability that it will belong to a given class. The data partition is then reordered from "most likely" to "least likely" to accept a loan.

It is convenient to look at the lift chart (sometimes called a cumulative gains chart when displayed in the default manner of XLMiner©), which summarizes the information into a graph. The graph is constructed with the cumulative number of cases (in descending order of probability) on the *x*-axis and the cumulative number of true positives on the *y*-axis, as shown in Figure 8.13. True positives are those observations from the important class (here "take out a personal loan") that are classified correctly. Figure 8.13 is the corresponding lift chart. The 45-degree line is a reference line. For any given number of cases (the *x*-axis value), the 45-degree line represents the expected number of "successes" we would predict if we did not have a model but simply selected the most common category. Remember that the most common category for this data set was "do not take out a personal loan." It provides a benchmark against which we can see the performance of the model. Because this reference line is drawn by assuming that we always choose the most prevalent category when scoring a new record, it is called the naive model, and we use it as a reference in the same manner we used a somewhat different naive model in previous chapters. The fact that the cumulative gains curve rises above the reference line indicates that the model estimated has "lift."

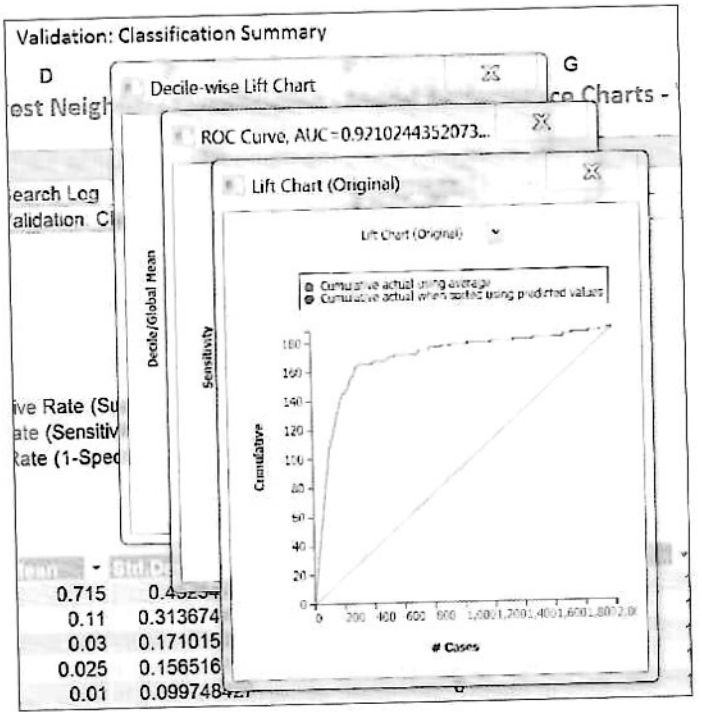


FIGURE 8.13
The Validation Partition Lift Chart (or Cumulative Gains Chart as Shown in the XLMiner® Default) for the Universal Bank Classification. (C8F8)

Source: Frontline Systems Inc.

The same information may be displayed using one of the alternative lift charts in XLMiner® called the decile-wise lift chart, as in Figure 8.14. The x-axis displays the 10 deciles in the validation partition, while the y-axis shows the lift associated with each decile. Note carefully that the data used to construct the chart is again reordered so that the most likely individuals to take out a loan (i.e., successes) are at the top of the data partition and the records representing individuals not likely to take out a personal loan are at the bottom of the data partition. The bars show the factor by which our model outperforms a naive model. Reading the first bar on the left, we see that taking the first 10 percent of the records that are ranked by the model as the most probable individuals to take out a loan yields almost eight times as many correct classifications as would a random selection of 10 percent of the records (i.e., the naive model). That "eight times" as successful as a naive model is the lift associated with this model.

XLMiner® has one additional way of displaying the same information; it is perhaps the most standard manner in which to display lift. Figure 8.15 shows the "true" lift chart with deciles on the x-axis. Most data mining software uses this standard type of display for lift. Remember again that we have always reordered the records from most likely to least likely before drawing the chart; failure to recognize this makes the chart impossible to interpret. Looking at the first decile in Figure 8.15 (0.1 on the x-axis) and following up to see the height of the fitted classifier at that point gives a number on the y-axis of between 8 and 9; this is the measure of lift for the first 10 percent of the records, or we could say it is the lift of the first decile. Note that we have accounted for most of the successes in the first decile; the lift curve in Figure 8.15 falls off quickly after the

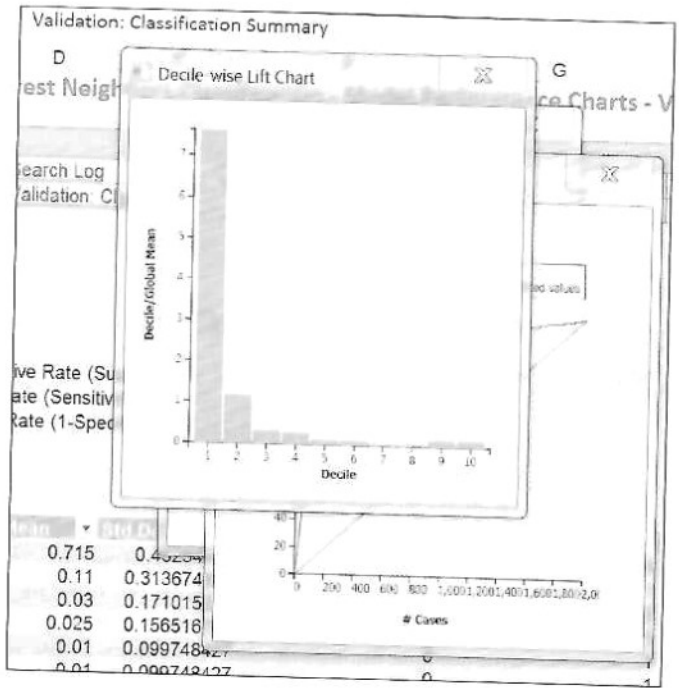


FIGURE 8.14
Decile-wise Lift Chart for the Universal Bank Classification. (C8F8)

Source: Frontline Systems Inc.

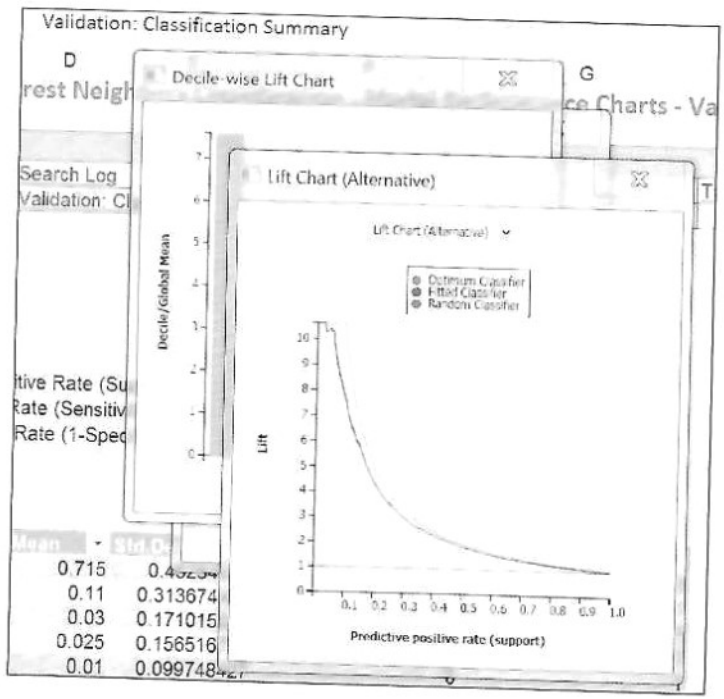


FIGURE 8.15
The "alternative" Lift Chart from XLMiner© for the Universal Bank Classification. This is the Standard Display Format used by Most Data Mining Software Packages. (C8F8)

Source: Frontline Systems Inc.

first decile. There were very few successes in the entire data set (we know that from our audit of the data), and most of those successes are accounted for in the first decile. This model appears to be highly predictive in categorizing bank customers in the first few deciles (the ones on the left of the chart). These first few deciles are of most interest to the data scientist. If we wanted to target customers for a personal loan campaign, we would only approach those near the top of our reordered list. The model has reordered the list by using the attributes considered in the analysis to categorize individuals as either successes or failures (i.e., "take out a loan" or "do not take out a loan").

THE RECEIVER OPERATING CURVE (ROC) AND AREA UNDER THE CURVE (AUC)

There is one final way of displaying essentially the same information. It is another method of explaining lift. It was developed during WWII by radar engineers (not data scientists). The engineers were using signals to detect hostile aircraft in the battlefield, and they wanted to better visualize the trade-off between correctly and incorrectly identifying an aircraft as a foe. The method they devised is called a receiver operating curve (ROC); Figure 8.16 shows the ROC for the Universal Bank classification model. The ROC provides an especially handy way to compare competing algorithms with a single number. It uses the same variable on the *y*-axis as the lift curve but

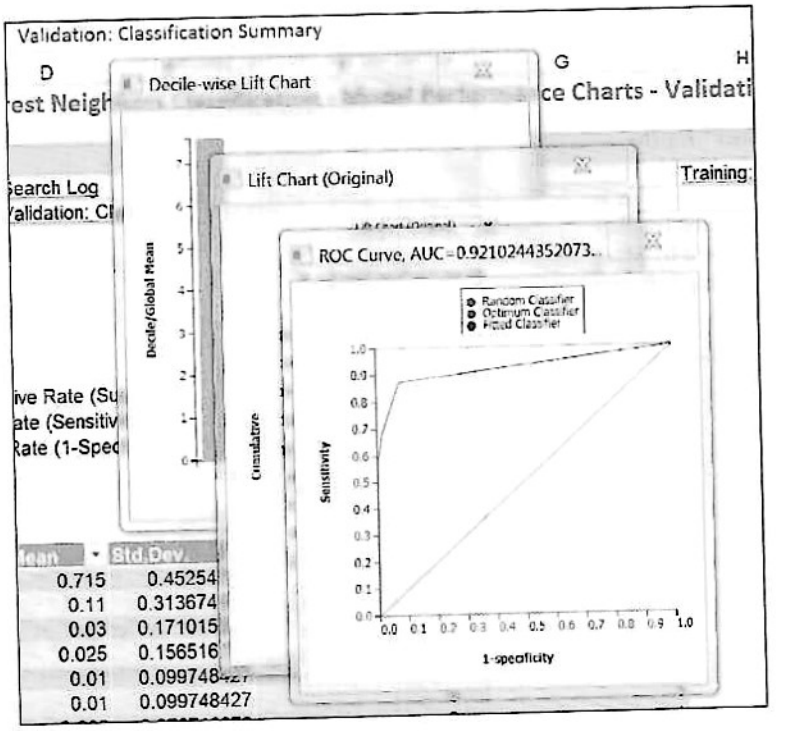


FIGURE 8.16
The Receiver Operating Curve for the Universal Bank Classification. (C8F8)

Source: Frontline Systems Inc.

expressed as a percentage of the maximum. The *x*-axis displays the false positives, also expressed as a percentage of the maximum.

Thus, the receiver operating curve is a plot of the true positive rate against the false positive rate. Another way of saying this is to say that the curve shows the trade-off between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity). Higher sensitivity will eventually find all the likely loan applicants, but it will also misclassify a large number of individuals as likely to take out a loan. Higher specificity, on the other hand, will eventually preclude incorrectly selecting any individuals as loan candidates, except ones who are certain to accept an offer, but it will also miss many likely loan candidates. Once again, recall that the data has been reordered according to the algorithm from most successful to least successful.

The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the algorithm. The curve allows you to quickly access the false positive rate that is associated with a true positive rate. The area under the ROC curve (designated as AUC or "Area Under the Curve" in XLMiner®) is a reflection of how good the test is at distinguishing (or "discriminating") between likely loan takers and those individuals not likely to take out a loan. The greater the area, the better the algorithm. Look at Figure 8.16, which shows a good test (that has a high sensitivity and specificity) and an AUC of 0.921. A worthless classification model would be described by the diagonal line in Figure 8.16.

What Is to Follow

In the following chapters, we will learn about and employ a number of classification algorithms. Why explain more than one classification algorithm? Each algorithm has its strong points and weaknesses. Just as with physical tools, it is best to use the appropriate tool for the particular job. Driving in a screw with a hammer is possible but not recommended. We will also learn about and employ clustering techniques; these algorithms are a form of unsupervised learning (because there is no target). Finally, we will learn about text mining; this will be a foray into unstructured data.

Suggested Readings

Davenport, Thomas H. *Big Data at Work: Dispelling the Myths, Uncovering the Opportunities*. Boston: Harvard Business Review Press, 2014.

Foreman, John W. *Data Smart*. Indianapolis, IN: John Wiley & Sons, Inc., 2014.

Hays, Constance L. "What Wal-Mart Knows About Customers' Habits." *The New York Times*, November 14, 2004.

Keating, Barry. "Analytics Off the Shelf." *Applied Marketing Analytics*, 2, no. 1 (Winter 2015–16), pp. 12–24.

Mangalindan, J. P. "Amazon's Recommendation Secret," *Fortune*, July 30, 2012.

Mayer-Schonberger, Viktor, and Kenneth Cukier. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. New York: Houghton Mifflin Harcourt, 2013.

Pierce, John R. *An Introduction to Information Theory: Symbols, Signals and Noise, Second Revised Edition*. New York: Dover Publications, 1980.

SAS Institute. *Applied Analytics Using SAS Enterprise Miner*. Carey, NC.: SAS Institute, Inc., 2011.

Shannon, Claude. "A Mathematical Theory of Communication," *Bell System Techincal Journal*, 27 (July, October 1948), pp. 379–423, 623–656.

Shmueli, Galit; Nitin R. Patel; and Peter C. Bruce. *Data Mining for Business Analytics*. Hoboken, NJ: John Wiley & Sons, Inc., 2016

Siegel, Eric. *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*. Hoboken, NJ: John Wiley & Sons, Inc., 2013.

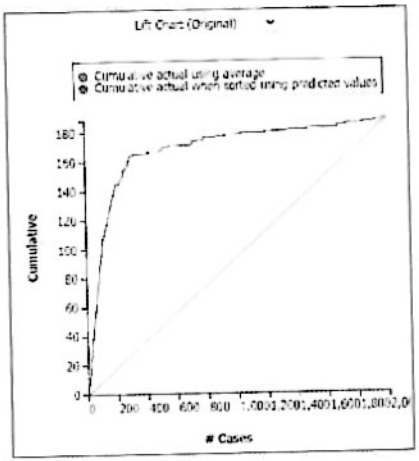
Silver, Nate. *The Signal and the Noise: Why So Many Predictions Fail - But Some Don't*. New York: Penguin Press, 2012.

Soni, Jimmy and Rob Goodman. *A Mind at Play: How Claude Shannon Invented the Information Age*. New York: Simon & Schuster, 2017.

Wendler, Tito; and Soren Grottrup. *Data Mining with SPSS Modeler*. Switzerland: Springer International Publishing, 2016.

Exercises

1. A classification model's misclassification rate on the validation data is a better measure of the model's predictive ability on new (unseen) data than its misclassification rate on the training data. Explain whether this statement is accurate and why that is so.
2. The first step in data mining procedures according to SAS and IBM/SPSS is to "sample" the data. Sampling here refers to dividing the data available for analysis into at least two parts: a training data set and a validation data set. Why do both SAS and IBM/SPSS recommend this as a first step? What are the risks of ignoring this procedural requirement?
3. How do "structured" and "unstructured" data differ? Which is the more prevalent form of data? How would the following be classified: numbers in an Excel spreadsheet, a thousand text files, a thousand video images, and a thousand audio files?
4. In the Universal Bank classification model estimated with XLMiner®, the software produced the validation data set lift chart shown.



Source: Frontline Systems Inc.

How is the naive model displayed in this diagram? What does the other line in the model represent?

5. Some data mining algorithms work so "well" that they have a tendency to overfit the training data. What does the term *overfit* mean, and what difficulties does overlooking it cause for the data scientist?

6. The validation data set confusion matrix for the Universal Bank data classification model is shown.

Validation: Classification Summary

Confusion Matrix

Actual\Predicted	0	1
0	1809	3
1	80	108

Source: Frontline Systems Inc.

How many records were in the validation data set? How many of these records were correctly classified by the algorithm? How many records were incorrectly classified? What is the "misclassification rate" for the entire validation data set? Would you predict that the misclassification rate for the training data set would be higher or lower on average than the rate you calculated for the entire validation data set?

7. Show the computation for the misclassification rate of this confusion matrix.

Confusion Matrix

Actual\Predicted	0	1
0	970	20
1	2	8

Source: Frontline Systems Inc.

8. In the Universal Bank data in this chapter, only 10 percent of the records represented customers who had taken out a personal loan (the target variable). If we were to score a new customer based upon the attributes we used in the algorithm, we would be accurate in the prediction about 90 percent of the time if we always scored the individual as "not accepting a personal loan" because that indeed is what most customers have done in the past. Why not accept being correct 90 percent of the time with this very simple decision rule?
9. Data has the characteristic of "nonrivalry." What is nonrivalry and why is it important to realize that data has this characteristic?
10. The lift chart and the confusion matrix are both standard diagnostic tools used to evaluate a data mining algorithm. Don't the two measures display the same information? Explain any differences between the two measures.