Chapter 8

Digging into Your Data

In This Chapter

- Focusing on specific problems
- Building on business knowledge
- Appreciating the advantages of your own data

Many organizations possess a mountain of data that's been collected in the course of routine business, and they're adding new data each day. As a data miner, you'll use this internal data as your primary natural resource.

This chapter focuses on framing a problem and finding relevant data within your existing resources. If you have more data on hand than you know what to do with, you're in the very situation that data mining was created to address. But on the other hand, if your data resources seem skimpy, don't worry. The ideas in this chapter still apply to you. Make the most of whatever you have! (See more about expanding your data resources in Chapters 9, 10, and 11.)

Focusing on a Problem

A data-mining project begins when you identify a specific business issue to investigate. The narrower and better-defined the question, the more effectively it can be answered. The more clearly the question is defined, the more clearly the data requirements can be understood, as well as the limitations of the answer. If you're faced with an issue that is very broad (such as "Why are we not selling enough?"), it helps to first break the question down into manageable bits. You don't have to cover the whole topic at once; just take one narrow part of the big problem and start with that.

Take, for example, one retailer's initial question, "How much repeat business are we getting?" That sounds at first like a simple, straightforward question, but it's actually a broad question that encompasses many smaller, more specific questions, like these:

- How many new customers come back?
- How many second-time customers return a third time?
- Do customers who first buy suits come back for shoes?
- Do those who make a small purchase return for larger purchases?

Answering these questions doesn't require a lot more than counting. It's just not that difficult to calculate how many customers return a second or third time, if you have some way to identify individuals. That's easy for online stores, where shoppers can be tracked by an account login or email address. Traditional retailers can identify customers by house credit cards or loyalty cards, although not every customer uses those.

From here, as the retailer got more familiar with data mining and the potential of predictive analytics and data-mining tools, his questions became more sophisticated and action oriented:

- How does the amount spent in the first visit relate to long-term spending?
- What behaviors or characteristics are indicators of high future spending? If so, what are they?
- Would additional information (for example, demographic data) improve our ability to predict a customer's spending behavior?

The object of data mining is to move beyond simply knowing what has already happened and understand how you may influence what will happen in the future.

Building on business knowledge

The most fundamental data needed for any data-mining project isn't the kind of data that is stored in electronic files. It's the business knowledge that you and others on your team have accumulated from your own experience and training. You don't have to be the foremost expert in the field you're investigating, but you do have to understand the basics of the business. You need to know the definitions of the fields in the data, and a little about how the data is collected and what flaws might occur in the data. If you know more, so much the better.

One client's unrealistic expectations

Some people have unrealistic expectations about data mining. To put it simply, they think it is magic, and expect results that only magic could provide. But it's really a down-to-earth practical process that helps you leverage your business knowledge and a few good tools to quickly extract useful information from data, so that you can use that information to address a specific business problem.

Here's a real-life example of an unrealistic expectation for data mining, and how that got in the way of the data-mining process. A large insurance company sent me a data sample and asked for results. But the insurance company didn't state any business concern that needed to be addressed. In fact, it didn't even label any of the data. I looked at the data file and found that it was nothing but unmarked columns and rows of data. The insurance company's staff thought that was all I needed. Situations like this are not rare, and they pose a challenge that data miners must patiently address. I contacted the client to explain that I could do nothing without knowing what the variables in the data were. I explained that, no, data mining really did not work that way, made a case for the client to provide the missing information, and then waited for it to be assembled and sent me. All of that took time. Even then, the client wasn't willing to reveal any specific business problem to address, so I had to guess. I suspected that claims processing would be an important cost and customer satisfaction issue, and confirmed that with an industry expert, which took more time. When the work was done (a focused analysis of the time required to process claims, identifying offices that processed claims more quickly than others, with the aim of investigating practices used by those offices to use as models for process improvement elsewhere), I wasn't sure that the client would value the results, because the client had never expressed any real interest in putting the information into action.

Think how much faster and better the process would have been if the client had begun with realistic expectations and teamed with me to discuss business issues from the start.

Managing Scope

Asking questions and exploring data can be fun. Now that you are a data miner, you'll find that you can ask and answer questions that were previously beyond your reach. Finding answers is motivating. You'll think of yet more questions. Perhaps you'll discover something so cool that you'll want to tell everyone about it.

It's all so exciting that it can easily get out of hand!

It's not just your own interests that can cause a project's scope to expand. As you work, you'll have discussions with coworkers, and they'll all have ideas and questions to inspire more exploration.

How a retailer got excited about data mining

Before people get enthusiastic about data mining, they're usually angry or frustrated about something else. As a data miner, your most satisfying moments, and your best opportunities to create loyal fans of your work, lie in addressing the problems that make managers lose sleep at night.

Earlier in this chapter, I told you about a retailer who began by asking simple questions such as, "How many first-time customers return for a second visit?" and gradually evolved into more sophisticated, action-oriented questions, such as, "What customer characteristics are associated with high levels of long-term spending?" What was on that retailer's mind when the process started? Just one thing: the lack of certain desired reports.

The retailer had invested in some very expensive software, with the goal of producing routine reports on a handful of simple metrics, such as the number of new customers returning for a second visit, but the software wasn't effective and the reports never materialized. Management was not happy.

So when the retailer went looking for a better solution, management wanted proof that the new solution would actually produce those reports. Just reports. Nobody asked for data mining. Nobody was thinking about asking questions that might give better guidance for action. Just give us our reports, please.

If you, as a data miner, heard about this retailer's situation, you might be tempted to shout, "Forget those old reports; data mining is better. You'll see that data mining is much more powerful than any report!" But that would be the wrong way to win over the retailer.

Here's an old adage: You must be equal before you can be better. So, if your manager or client wants something specific, you must first satisfy that want. You'll show that you are equal to the requirement. When you've done that, you will have earned the respect required to go forward, provide something additional and unexpected, and be . . . better.

And that's just what happened for this retailer. With the report completed, and the retailer's requirements satisfied, the data miner was free to dig a little deeper.

She had noticed that the data included some information obtained through the retailer's loyalty program, basic information about the customers, and details of their homes and interests. But this information was often left blank. She wondered, "Is it worthwhile to collect this information?" So, she quickly experimented with decision-tree models for predicting consumer spending (refer to Chapter 15 for more about decision-tree models). She tested combinations of behavioral data (what the customer bought) and demographic data (customer information collected with loyalty program registrations). She discovered that the loyalty program data was not only useful, but that by combining it with sales information for the customer's first purchase, she was also able to develop a surprisingly good prediction of the customer's long-term spending.

When the time came for a presentation, the data miner first presented the report that the retailer had requested at the start. Only when the retailer had reviewed and felt fully satisfied with that report did the data miner go on to show something more. And wow! The customer spending model was a great finale. The retailer became an instant fan of data mining.

You'll find no limit to the sources of inspiration available to the data miner. But a limit exists to your available time.

As you work, you must have specific goals in mind as well as a realistic plan for meeting them. The goals must be defined in business terms that suit your manager's or client's needs. Your plan is your assurance that you will produce something of value, not just something that you find interesting. Your plan is your guide for deciding which questions to address now and which questions must be set aside for later.

So focus on specific goals, refer to your plan, and don't let your project's scope expand or wander before you complete your goals. (Refer to Chapter 6 for a detailed discussion of planning for data mining.)

What if you've completed your project goals and still have time before your deadline? Fantastic! Now you have the opportunity to investigate one or more of the best new questions that have come to mind, and add a valuable extra to your final presentation.

Using Your Organization's Own Data

A data miner has nothing without data. And if you work in a large organization, you'll have hundreds, perhaps thousands, of existing data resources potentially available for data mining. Every activity generates records, and those records can become your raw material. Table 8-1 shows the variety of commonly collected data in a number of business activities.

Table 8-1 Data collected from common business activities

Business activity	*Data collected*
Research	Competitor product information experimental and test data
Manufacturing	Process data; procurement records production records inspection and test records
Marketing	Competitor marketing information and sales data campaign data marketing cost data
Sales	Sales activity sales data customer information

(continued)

Table 8-1 ***(continued)***

Business activity	*Data collected*
Fulfillment	Packaging records shipping records shipping complaints
Customer service	Customer interaction records product and service complaints service issues
Technical support	Support requests product problem reports design and other product suggestions
Training	Staff training records customer training records certification and other credentialing records
Accounting	Bills payments audit records taxes collected and paid

That's a pretty long list, yet it's really only a tiny sample of the activities and related data that's already waiting somewhere within your business.

But knowing that data exists is not the same thing as being able to access and use it for data mining. For one thing, you'll need much more specific information about exactly what internal data is relevant to the specific business problem you're investigating. Who collects it? Who controls access? What variables (fields) are recorded, and for what range of time or activity? Where can you find documentation?

Appreciating your own data

You and your manager might choose from a number of options when selecting which project to tackle with data mining. You always have a choice of tools. But when it comes to data, you may have no choice at all: You use the data available to you or your company right now.

You may have doubts about this data. You are sure to know something about its flaws. And you may have heard about other organizations that have larger quantities of data or different types of data than your own. Nonetheless, your organization's internal data, the information collected in the course of everyday business, is your most valuable resource. It's the very best data that you can have for data mining. It is superior to all external sources in a number of ways:

- **Unique relevance:** The data pertains to your own business, with all its distinctive characteristics. It's about your own customers, your own products, your own business practices. Whatever you may discover in this data will clearly also be relevant to the business. Nobody will be able to reject your results with the *but our business is different* excuse.
- **Transparency:** You know (or you can find out) the sources of your own data. No mysteries should exist about the definitions of variables, the data collection methods, the time, the place, or the people involved.
- **Detail:** You'll have *raw data,* collected in the finest possible level of detail.
- **Range:** Your data resources cover the full scope of activity taking place in your business.
- **Competitive advantage:** Only you have your own internal data. It is not available to your current or your upcoming competitors.
- **Development potential:** You can build on your own data in ways that would not be possible with data from any outside source. If you want to integrate information from multiple sources, your data will contain the identifiers you need to do that. If you want to know more about customers, you have their names and contact information, and you can refer to other records, survey them, or even call and have a personal conversation. If you need more detailed or additional data, you may be able to change a data collection practice.

Another nice thing about your own data: You own it. Any data collection costs were covered by the business unit that generated the data in the first place. You'll pay no fees and have no licensing issues to consider when using and reusing the data. (You may face data storage and other data management issues, but that's true for any data source.)

Your own data resources will not be perfect in every way. You might discover that some data you'd like to use has not been collected, or has been discarded. You're bound to encounter some data quality problems. And, of course, internal data has limits — it tells you about your own organization, but not your competitors. Still, internal data will always be your primary and most valuable data resource.

Handling data with respect

Data mining, like any kind of data analysis or reporting, uses a lot of data, much more than most everyday business activities. When you access data and perform analysis, you must be careful to do so in ways that stay within your company's guidelines and that don't interfere with routine business processes.

Data resources can be just as precious, and just as private, as cash. Get off to the right start in data mining by treating data with respect and discovering proper practices for data management and governance that affect your work.

Failure to follow legal and good business practices for data governance can lead to serious trouble. It's important that data isn't accessed by people who should not use it, that records not be improperly changed or destroyed, and that new data you create be properly archived. Documentation is a necessity. Many legal and good business practice requirements will be relevant to your work in data mining. (See Chapter 3 to learn about teaming up with others to get guidance and help with data management.)

This may not be simple. You'll have to discover things about what data is available, how to get access, and how to handle the data properly so that you don't get in the way of others. In short, you'll have to get involved with new things and new people. And it will be worth it, because you'll get more done and broaden your own horizons as a result.

You'll have to find out new things, but you won't have to become a data governance expert. You can rely on the others in your organization who are experts in data governance and data management. Work with them constructively, and they will help you to stay within the law and to follow good data management practices.

Data miners and data governance professionals don't always play nicely together. Data miners have been known to resent controls on access to data and sometimes resort to elaborate schemes to avoid playing by the data access rules. Data governance experts don't always understand why data miners need to use so much data; they've been known to stall. Frustrating experiences in the past can affect the way that either group deals with the other.

So, as you start your career in data mining, make it a point to reach out to people in your Information Technology department. Talk with them about your data-mining work and discuss how it will benefit the organization. Ask about data governance issues that relate to your work, and show that you care about good data management. This show of respect is the way to start a positive and absolutely necessary working partnership between the data-mining and data governance teams. (Refer to Chapter 3 for more about teamwork.)