Business Understanding

If you don’t know where you’re going, you’ll probably end up somewhere else.

—Laurence Peter

Having a clear vision of what you want to accomplish is the cornerstone of successful data mining. Unfortunately, there is a common misconception of data mining as a black box into which you drop data and then “something interesting” magically appears. Successful data mining, however, rests on two principles: a solid foundation of business knowledge and a focused attack on a specific problem.

The CRISP-DM process model includes a series of tasks that help you develop the business understanding necessary to make your data mining efforts pay off. These tasks are:

- Determining business objectives
- Assessing the situation
- Determining data mining goals
- Producing a project plan

Completing these tasks enables you to understand where you want to go, where you are right now, and how the business problem translates into a data mining problem, and it also gives you a concrete plan to get the job done.

Determining Business Objectives

The first task is to develop a thorough understanding, from a business perspective, of what you want to accomplish. Keep in mind that successful data mining is driven primarily by business rather than by data or technology. Often, you will have many competing objectives and constraints that must be properly balanced. If you try to take
on the whole world at once, you are likely to fail. On the other hand, if you plan carefully and focus on a specific, achievable, measurable goal, you will be well on your way to success.

**Background**

The first step is to understand the background of the problem. By *background*, we mean the details of the organization's business situation at the start of the project. This information helps refine the business goal and identify resources that can be used to reach that goal.

**Example:** *CRM*. The company is a consumer electronics retailer and would like to improve relationships with customers. Improving relationships with customers is a very broad goal, one that is achieved in manageable steps. There are three possibilities for improving the value of your customer relationships: decrease the cost of initiating the relationship, increase the amount and/or quality of business with the customer, or increase the length of the relationship. For now, the company will focus on the second possibility.

Increasing the amount and quality of business with customers is still too general, so the company needs to narrow the goal further. Cross-selling presents an important way to increase the amount of business with a customer. When done properly, cross-selling benefits both customers and the company. Customers have more of their needs met in one place, and the company sees increased revenue generated through additional sales. So, for this example, the company will consider how to use cross-selling to increase business volume with existing customers.

**Example:** *Web mining*. The company runs an e-commerce Web site selling legwear (socks, nylons, etc.). The company would like to make the site friendlier and easier to use so that visitors enjoy their experience more and are more likely to become repeat customers.

**Example:** *Fraud detection*. The agency is concerned about identifying potentially fraudulent grant applications and would like to find a way to use available information (from the application itself and other sources) to select suspicious applications for further investigation.

*(see last page for a description of these examples)*
Business Objectives

Business objectives are detailed descriptions of what you want to accomplish using data mining. This always includes a primary objective and often includes secondary objectives as well. Be specific in defining business objectives. You must pin down precisely what you want to see happen in your business. The path to defining business objectives includes describing the problem to be solved, specifying all business questions, detailing any related business requirements, and outlining the expected benefits from data mining.

Example: CRM. The electronics retailer wants to use cross-selling to increase customer value and loyalty. This needs to be narrowed further to derive a useful business objective. The company starts by focusing on a new product line and would like to launch a marketing campaign to let customers know about the new offerings. However, the company has a limited budget, so the campaign will focus on the customers who are most likely to respond. Therefore, the business objective will be to identify the customers who are most likely to respond to the marketing campaign.

Example: Web mining. The objective is to be able to identify factors that increase the probability of a surfer leaving the site. There could be any number of factors involved here, including pages that take too long to load, problems finding the product of choice, requiring too many clicks to get to the desired page, or other unknown factors.

Business Success Criteria

This piece of the puzzle focuses on how you will measure the results of your efforts and what kind of performance will be regarded as successful. Again, each criterion should be defined from the business perspective, for example—“increase sales of new credit card accounts by X%” rather than “build a neural network with X% accuracy.”

Each stated business goal should have at least one criterion, and each criterion should correspond to one of the original business goals. If you specify a success criterion that doesn’t directly correspond to one of the stated business goals, you may need to reconsider whether the criterion is relevant for this data mining project. If you find that you do consider the criterion in question to be a good measure of success, you may need to redefine the business goals to better match your definition of success.

The success criteria can range from the very concrete, such as “sales will be $X for the next quarter,” to the subjective, such as “we will understand why our customers are
buying from our competitors." The more objective the criteria are, the easier it will be to assess your results later. However, subjective criteria can also be used to evaluate data mining results as long as you identify exactly who will make the final judgment about the criteria at this early stage of the process.

**Example:** CRM. The aim is to entice existing customers to make purchases from the company’s new product line. Since this is the first foray into e-CRM/data mining, the company will start with a small project and will consider the project a success if the profit increase due to data mining is larger than the cost of the mining. That would constitute a positive return on investment for this project, and it would also demonstrate the value of data mining, providing a good justification for broadening the scope of its data mining efforts.

**Example:** Fraud detection. The success criterion here is less distinct, because the agency doesn’t have historical data on fraudulent claims. The agency won’t be able to make purely predictive models in this situation but can create models that highlight anomalous grant applications. That should allow the agency to focus its investigative resources on the cases most likely to reveal irregularities.

### Assessing the Situation

Now that you know where you want to end up, it’s time to take stock of where you are now in order to chart a course that will get you to your destination. This includes identifying available resources, constraints, assumptions, and other factors that will determine how to proceed. This information will help you formulate a realistic and achievable plan applying data mining.

### Inventory of Resources

You need to know what resources you will have at your disposal during the data mining process, including the following:

- **Personnel.** Business and data experts, technical support staff, and data mining personnel.

- **Data.** Fixed extracts and/or live access to data warehouses, data marts, or operational databases.
- **Computing resources.** Hardware for performing data analysis and related activities.
- **Software.** Programs for accessing, cleaning, and analyzing data, and applications for reporting and deploying the results.

The availability of these resources will affect important aspects of your project, such as the scope and time line. Having a clear picture of what you will have at your disposal will allow you to plan accordingly and increase your odds of success.

**Example: Web mining.** For personnel, the company has a business analyst with substantial retail experience (50% time), a database analyst (50% time), and a data mining consulting firm (2 analysts, 100% time). For data, the company has a data warehouse of information derived from Web logs, registration information for many customers, and order (purchase) information. The company also has some purchased demographics that have been merged with the operational data. For computing resources, the data warehouse resides on a multiprocessor UNIX system. The company also has a dedicated, high-speed, Windows XP workstation that can be completely devoted to the data mining project. The data mining consultants will also provide some hardware support for their activities. For software, in addition to the RDBMS used to host the data, the company will use the Clementine data mining workbench from SPSS Inc. A copy of SPSS Base software is available for certain tasks that are more easily done there than in Clementine.

**Requirements, Assumptions, and Constraints**

In addition to understanding your data mining assets (resources), you need to understand the liabilities associated with the data mining project. These are factors that can potentially affect the schedule for the project and the quality of results, and these factors can also have security and legal ramifications.

In any large project like data mining, you will have to base your planning on certain assumptions. They may concern the data or other resources for the project, or they may concern the business model underlying the entire project. Data assumptions are relatively simple to test during the data mining process, but business assumptions can be more complicated to understand and assess. Some business assumptions may prove to be untestable, but they can still have an impact on the results of the project. Explicitly outlining these assumptions will help you to understand some of the key determinants of success for the project and will also help you to avoid potential pitfalls.
Chapter 3

In planning the project, take the different constraints into account. Constraints come in various forms: limitations on available resources, legal and ethical restrictions, and conflicting business requirements. (An example of the latter might be “take every opportunity to expose potential customers to our products,” and “don’t drive customers away with overbearing marketing tactics.”)

**Example:** CRM. The fundamental requirement is that the company must identify a target group of customers who will be likely to respond to a new product promotion.

Assumptions include the following: it is possible to find a model that will generate accurate (enough) predictions from the available data; the company will be able to devise a marketing strategy that can effectively take advantage of the knowledge gained through data mining; and economic conditions in the near future will be similar to the recent past, so that the model, which is based on past data, will still make sense in the deployment time frame.

The company’s main constraints are a three-month implementation time frame that coincides with the launch of the new product line and, of course, resource limitations.

---

**Risks and Contingencies**

In any project, there are risks. There is the potential for events beyond your control to disrupt the process. Those events can have a huge impact on the schedule of your project and the quality of the results. However, forewarned is forearmed; by taking the time to list the relevant risks and outlining contingencies to handle them, you can avert disaster. Obviously, not all adverse events can be predicted, and some risks may not allow an adequate contingency plan. But if you can cover the majority of potential problems, you will minimize the likelihood of an unexpected setback.

In a data mining project of any size, risk comes in many forms:

- **Business risks**, such as changes in your market or in the general business climate.
- **Organizational risks**, such as staff turnover or political shifts.
- **Financial risks**, such as diversion or loss of resources in mid-project.
- **Technical risks**, such as equipment breakdowns or incompatibility between computing platforms or software packages.
- **Data risks**, such as poor data quality or access problems.
Each type of risk presents unique challenges. For each risk, consider the conditions under which the project would be compromised. As much as possible, be sure to have a contingency plan to cover each risk. In some cases, one contingency plan may suffice to cover several risks.

**Example:** **Web mining.** With control over most aspects of the project, the company can do a lot to minimize risks. Data is backed up both on- and off-site, and the company has service contracts on computer equipment to have it repaired or replaced within 24 hours. The company can’t control the business environment, but evaluating various economic indicators has revealed that any major change that would invalidate the project is unlikely to happen in the relevant time frame. The time frame is short enough that organizational risks should be small. One possible wild card is the data mining consultants. If the consultants turn out to be unable to complete the task, the company will need to find someone else to do the technical work. This would create a setback to the schedule, but the company has made arrangements with the consulting agency to get a replacement up to speed quickly, if necessary. Since the company is running the data warehouse internally, it has a pretty good sense of what it contains and the condition of the data. The analysts know that the data are not perfect, but they are confident that overall the data are in decent shape. In addition, they have the source data from operational system archives. If they find severe problems in the warehouse, they can probably reconstruct the data to fix those problems, though this would clearly involve a significant cost in terms of delays and increased resource requirements.

---

**Terminology**

Every field has its own terminology or jargon. When team members from different backgrounds try to discuss a project using these specialized vocabularies, it can lead to a virtual Tower of Babel.

To help your data mining project flow as smoothly as possible, it’s important to outline the terminology that will be used throughout the project. Particularly important is an understanding of how the terminology for describing the business problem and the terminology for specifying the data mining procedure mesh together. At this point, you should outline the relevant terminology from both perspectives and make sure that everyone on the team understands the meaning of each term. It is worth putting these definitions in writing, as well. A formal glossary gives team members a reference to guide them through ambiguities that may arise during the course of the project.
Chapter 3

Compiling such a glossary may seem like a luxury that can be skipped in the name of expediency, but it is a wise investment. If you don't give this matter the attention it deserves, your project will very likely suffer from confusion and wasted effort later.

Example: **Web mining.** During various planning meetings, a designated party keeps a list of jargon and buzzwords. Part of the meeting follow-up is to agree on definitions for any ambiguous items. The designated party makes this continuously updated glossary available to everyone on the project via a central document on the company intranet.

**Costs and Benefits**

As with any other investment, the key to understanding its place in your business plan is a detailed cost-benefit analysis. To make your data mining project pay off, you need to understand what the project will cost and what you stand to gain by implementing it.

When estimating costs, remember to factor in items such as developing the solution and ongoing maintenance for any implemented solution. When estimating benefits, remember to include less tangible benefits such as increased customer satisfaction or improved relations with the community.

Example: **CRM.** For this marketing campaign, the cost per contact is estimated to be about $0.30 per person, plus fixed costs of about $5,000 for copy writing, design, mail handling, etc. The average profit per purchase in the new product line is expected to be about $20. The cost of the data mining project is estimated at $15,000. So, to justify the effort, the company needs to attract about 500 more buyers than it would have without data mining, given the same number of mailers sent out.

Example: **Fraud detection.** The costs involved here will be more modest, because the data sets involved are simpler and smaller than those used in the other examples. In addition, only in-house data analysis expertise will be used—no high-priced consultants. The fixed costs will be approximately $5,000–$7,000 for the data mining project. In this case, there is an additional cost of investigation for each claim identified as potentially fraudulent. However, the grants awarded by the agency generally range between $10,000 and $300,000 with some as large as $600,000. Thus, if a single fraudulent claim can be identified and rejected as a result of data mining, the effort will be worthwhile.
Determining Data Mining Goals

Now that you have developed a thorough understanding of the business problem to be solved and the general business context, the next step is to identify how data mining will help you solve that problem. Specifically, you need to identify what information you will need to obtain by mining your data. That information is your data mining goal. Here, you are mapping your business goal to a data mining goal. By achieving the data mining goal, you produce the key element for achieving your business goal.

Data Mining Goals

A data mining goal describes the requirements for solving your business problem in technical terms rather than business terms. This brings the problem to a concrete level, a level that brings to bear the powerful tools of data mining.

Example: CRM. The data mining goal is to build a model that predicts the purchase of a new product in response to a marketing campaign based on available customer data. The model has to be accurate enough to save time and effort by allowing the company to target the promotion to the right customers. It must also be general enough to apply to current and future customer profiles for a reasonable time window, in this case at least six months (the time for the data mining project plus three months of actual use to realize the return).

Example: Web mining. The data mining goal here is to produce a model (or models) that allow the analysts to identify factors related to Web site visitors leaving the site. These factors should give them insight into how to improve the site so that visitors stay longer and become customers.

Data Mining Success Criteria

When you define business objectives, you also define success criteria as a way to test how well those objectives are met. You must follow a similar pattern with data mining goals. Once you’ve specified the goals, you need to determine how you will assess your data mining results. In the case of data mining goals, the success criteria should be specified in technical terms, such as a certain degree of predictive accuracy.
Chapter 3

Once again, your job will be easier if you stipulate objective criteria for judging your success. However, you can also accommodate more subjective success criteria if you specify up front who will be responsible for making the judgment and the characteristics that will be used to formulate that judgment.

Finally, in evaluating your data mining success, remember to include deployment issues. A great data mining model is not much use if there’s no practical way to apply the model to your daily business routine. You don’t get any return on your investment until you capitalize on that new knowledge.

Example: CRM. The success criteria correspond to the data mining goals outlined above. To ensure that its model is general, the company will build the model on one set of customer data and test it on a different set of known data. This will test how well the model will perform on new data. In order to have adequate predictions, the accuracy of the model must be good enough to let the company increase the response rate to the desired level without requiring a corresponding increase in the number of contacts. The point is not to simply make more contacts but to make better contacts.

Producing a Project Plan

At this point, you have all the pieces necessary to construct a plan for your data mining project. This is the last task in developing your business understanding for the data mining project, and it provides the context for the rest of the work in your project.

Project Plan

Here, you will make a detailed road map for the project. The project plan should include all identified goals and selected techniques in a coherent procedure that directly addresses the business problem. The CRISP-DM process model provides a skeleton for this plan, but you will need to fill in the specifics at each point in the process, including duration, required resources, output, and dependencies. Also include details about where the risks and associated contingency plans fit into the process.
Example: CRM. The following table presents an outline of the project plan for the data mining project.

<table>
<thead>
<tr>
<th>Phase</th>
<th>Time</th>
<th>Resources</th>
<th>Risks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business understanding</td>
<td>1 week</td>
<td>All analysts</td>
<td>Economic change</td>
</tr>
<tr>
<td>Data understanding</td>
<td>3 weeks</td>
<td>All analysts</td>
<td>Data problems, technology problems</td>
</tr>
<tr>
<td>Data preparation</td>
<td>5 weeks</td>
<td>Data mining consultants, some database analyst time</td>
<td>Data problems, technology problems</td>
</tr>
<tr>
<td>Modeling</td>
<td>2 weeks</td>
<td>Data mining consultants, possibly some database analyst time</td>
<td>Technology problems, inability to find adequate model</td>
</tr>
<tr>
<td>Evaluation</td>
<td>1 week</td>
<td>All analysts</td>
<td>Economic change, inability to implement results</td>
</tr>
<tr>
<td>Deployment</td>
<td>1 week</td>
<td>Data mining consultants, database analyst</td>
<td>Economic change, inability to implement results</td>
</tr>
</tbody>
</table>

Initial Assessment of Tools and Techniques

At this planning stage, you also need to start thinking about what tools you will adopt for your data mining efforts. Your choice of tools will influence the rest of the data mining process, so it is important to choose a flexible suite of tools that provides the methods you need for the data preparation, modeling, evaluation, and deployment phases of the process. You will find a brief summary of data mining methods in Chapter 9.

For the examples this book, we will use Clementine from SPSS Inc., a comprehensive data mining workbench that provides data access and manipulation tools, a variety of modeling methods, and deployable model results that can be easily incorporated into regular business operations.
Summary

This chapter outlined the importance of establishing a solid business understanding of the data mining project. The critical tasks and associated results of this phase include:

Determining business objectives. Defining the goals from a business perspective.
Task results:
- Background
- Business objectives
- Business success criteria

Assessing the situation. Determining where things stand at the beginning of the project.
Task results:
- Inventory of resources
- Requirements, assumptions, and constraints
- Risks and contingencies
- Terminology
- Costs and benefits

Determining data mining goals. Translating business goals into terms that data mining can directly address.
Task results:
- Data mining goals
- Data mining success criteria

Producing a project plan. Creating a blueprint for the remainder of the project.
Task results:
- Project plan
- Initial assessment of tools and techniques
Chapter 2

Examples

In the remainder of this book, we will illustrate concepts by showing how they might apply to specific data mining projects. Most of the examples will be taken from one of the following scenarios.

Example 1: Customer Relationship Management

The first example scenario is a retail company selling consumer electronics. The company wants to use its data on customers, products, and transactions to find ways to improve customer satisfaction and decrease costs. The data for this example are fictitious.

Example 2: Public Sector Fraud Detection

The second example scenario is a government agency charged with administering assistance grants to farmers in need. The agency is primarily interested in reducing fraud in grant applications. The data for this example are fictitious.

Example 3: Web Log Analysis

The third example scenario is a company specializing in e-commerce, which wants to find ways to maximize the return from their online business-to-consumer (B2C) operations. They plan to do this by mining their Web logs along with other data, such as customer registration data and purchased demographics. The data for this example are actual clickstream and demographic data from a now defunct e-commerce company called Gazelle.com that specialized in selling legwear (socks, pantyhose, etc.). These data were used in the KDD-Cup 2000 data mining competition and are made available through the kindness of Blue Martini Software. (Customer identification information has been removed from the data.) For more information, see the KDD-Cup 2000 organizers' report (Kohavi et al., 2000), or visit the Web site at http://www.ecn.purdue.edu/kddcup. Note that the data mining process illustrated here with these data is not meant to represent analysis that was actually performed by Gazelle.com analysts or their associates; it merely represents a possible approach to mining data of this kind.