IBM SPSS Modeler CRISP-DM Guide

Note

Before you use this information and the product it supports, read the information in "Notices" on page 39.

Product Information

This edition applies to version 18, release 4, modification 0 of IBM® SPSS® Modeler and to all subsequent releases and modifications until otherwise indicated in new editions.

© Copyright International Business Machines Corporation .
US Government Users Restricted Rights – Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

Preface ..... **vii**

Chapter 1. Introduction to CRISP-DM **1**
CRISP-DM Help Overview 1
CRISP-DM in IBM SPSS Modeler 2
Additional resources 3

Chapter 2. Business Understanding **5**
Business Understanding Overview 5
Determining Business Objectives 5
E-Retail Example--Finding Business Objectives 5
Compiling the Business Background 5
Defining Business Objectives 6
Business Success Criteria 6
Assessing the Situation 7
E-Retail Example--Assessing the Situation 7
Resource Inventory 7
Requirements, Assumptions, and Constraints 8
Risks and Contingencies 8
Terminology 8
Cost/Benefit Analysis 9
Determining Data Mining Goals 9
Data Mining Goals 9
E-Retail Example--Data Mining Goals 9
Data Mining Success Criteria 10
Producing a Project Plan 10
Writing the Project Plan 10
Sample Project Plan 10
Assessing Tools and Techniques 11
Ready for the next step? 11

Chapter 3. Data Understanding **13**
Data Understanding Overview 13
Collecting Initial Data 13
E-Retail Example--Initial Data Collection 13
Writing a Data Collection Report 14
Describing Data 14
E-Retail Example--Describing Data 14
Writing a Data Description Report 14
Exploring Data 15
E-Retail Example--Exploring Data 15
Writing a Data Exploration Report 15
Verifying Data Quality 16
E-Retail Example--Verifying Data Quality 16
Writing a Data Quality Report 16
Ready for the next step? 17

Chapter 4. Data Preparation **19**
Data Preparation Overview 19
Selecting Data 19

E-Retail Example--Selecting Data.... 19
Including or Excluding Data.... 19
Cleaning Data.... 20
E-Retail Example--Cleaning Data.... 20
Writing a Data Cleaning Report.... 20
Constructing New Data.... 21
E-Retail Example--Constructing Data.... 21
Deriving Attributes.... 21
Integrating Data.... 21
E-Retail Example--Integrating Data.... 22
Integration Tasks.... 22
Formatting Data.... 22
Ready for modeling?.... 23

Chapter 5. Modeling.... 25
Modeling Overview.... 25
Selecting Modeling Techniques.... 25
E-Retail Example--Modeling Techniques.... 25
Choosing the Right Modeling Techniques.... 25
Modeling Assumptions.... 26
Generating a Test Design.... 26
Writing a Test Design.... 26
E-Retail Example--Test Design.... 26
Building the Models.... 27
E-Retail Example--Model Building.... 27
Parameter Settings.... 27
Running the Models.... 27
Model description.... 28
Assessing the Model.... 28
Comprehensive Model Assessment.... 28
E-Retail Example--Model Assessment.... 28
Keeping Track of Revised Parameters.... 29
Ready for the next step?.... 29

Chapter 6. Evaluation.... 31
Evaluation Overview.... 31
Evaluating the Results.... 31
E-Retail Example--Evaluating Results.... 31
Review Process.... 32
E-Retail Example--Review Report.... 32
Determining the Next Steps.... 32
E-Retail Example--Next Steps.... 33

Chapter 7. Deployment.... 35
Deployment Overview.... 35
Planning for Deployment.... 35
E-Retail Example--Deployment Planning.... 35
Planning Monitoring and Maintenance.... 36
E-Retail Example--Monitoring and Maintenance.... 36
Producing a Final Report.... 36
Preparing a Final Presentation.... 37
E-Retail Example--Final Report.... 37
Conducting a Final Project Review.... 37
E-Retail Example--Final Review.... 38

Notices.... 39
Trademarks.... 40

Terms and conditions for product documentation..... 40

Index..... 43

Preface

IBM SPSS Modeler is the IBM Corp. enterprise-strength data mining workbench. SPSS Modeler helps organizations to improve customer and citizen relationships through an in-depth understanding of data. Organizations use the insight gained from SPSS Modeler to retain profitable customers, identify cross-selling opportunities, attract new customers, detect fraud, reduce risk, and improve government service delivery.

SPSS Modeler's visual interface invites users to apply their specific business expertise, which leads to more powerful predictive models and shortens time-to-solution. SPSS Modeler offers many modeling techniques, such as prediction, classification, segmentation, and association detection algorithms. Once models are created, IBM SPSS Modeler Solution Publisher enables their delivery enterprise-wide to decision makers or to a database.

About IBM Business Analytics

IBM Business Analytics software delivers complete, consistent and accurate information that decision-makers trust to improve business performance. A comprehensive portfolio of business intelligence, predictive analytics, financial performance and strategy management, and analytic applications provides clear, immediate and actionable insights into current performance and the ability to predict future outcomes. Combined with rich industry solutions, proven practices and professional services, organizations of every size can drive the highest productivity, confidently automate decisions and deliver better results.

As part of this portfolio, IBM SPSS Predictive Analytics software helps organizations predict future events and proactively act upon that insight to drive better business outcomes. Commercial, government and academic customers worldwide rely on IBM SPSS technology as a competitive advantage in attracting, retaining and growing customers, while reducing fraud and mitigating risk. By incorporating IBM SPSS software into their daily operations, organizations become predictive enterprises - able to direct and automate decisions to meet business goals and achieve measurable competitive advantage. For further information or to reach a representative, visit http://www.ibm.com/spss.

Technical support

Technical support is available to maintenance customers. Customers may contact Technical Support for assistance in using IBM products or for installation help for one of the supported hardware environments. To reach Technical Support, see the IBM web site at http://www.ibm.com/support. Be prepared to identify yourself, your organization, and your support agreement when requesting assistance.

Chapter 1. Introduction to CRISP-DM

CRISP-DM Help Overview

CRISP-DM, which stands for Cross-Industry Standard Process for Data Mining, is an industry-proven way to guide your data mining efforts.

- As a **methodology**, it includes descriptions of the typical phases of a project, the tasks involved with each phase, and an explanation of the relationships between these tasks.
- As a **process model**, CRISP-DM provides an overview of the data mining life cycle.

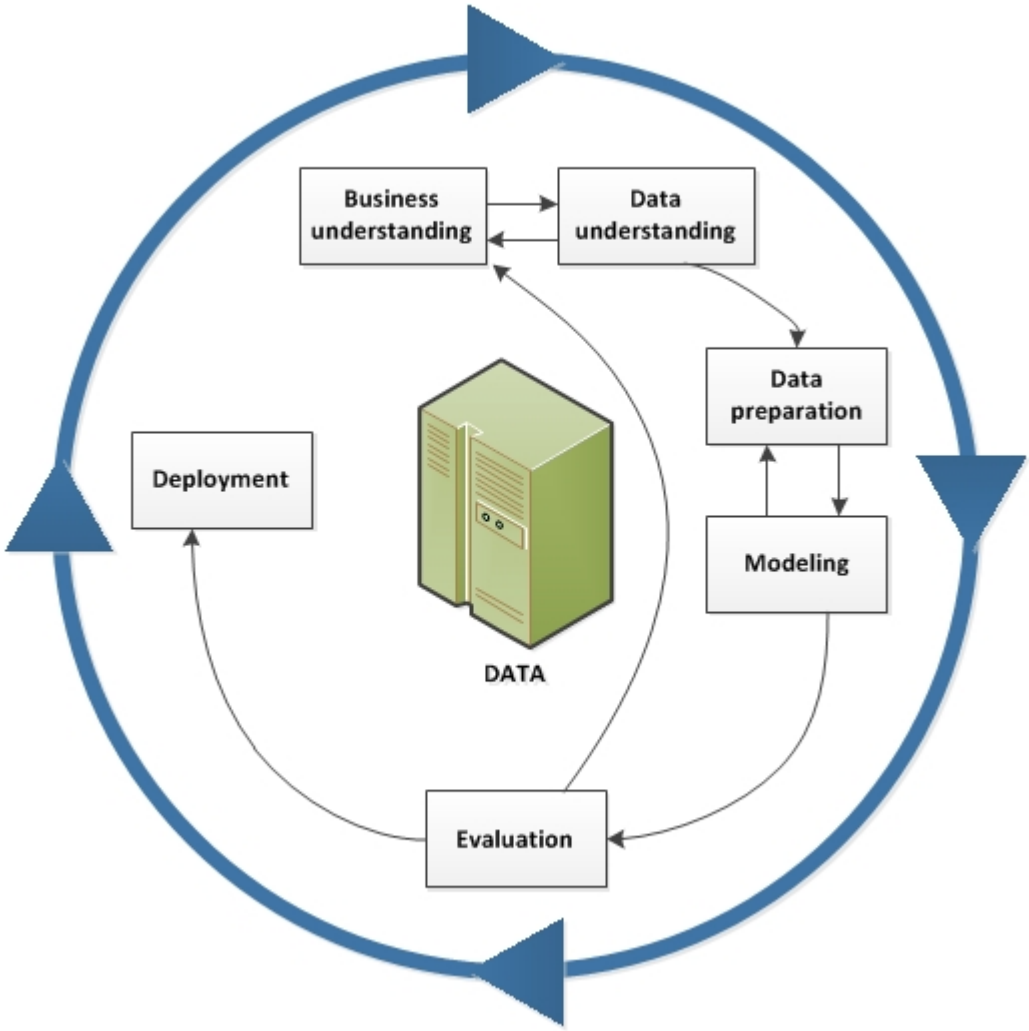

Figure 1. The data mining life cycle

The life cycle model consists of six phases with arrows indicating the most important and frequent dependencies between phases. The sequence of the phases is not strict. In fact, most projects move back and forth between phases as necessary.

The CRISP-DM model is flexible and can be customized easily. For example, if your organization aims to detect money laundering, it is likely that you will sift through large amounts of data without a specific modeling goal. Instead of modeling, your work will focus on data exploration and visualization to uncover suspicious patterns in financial data. CRISP-DM allows you to create a data mining model that fits your particular needs.

In such a situation, the modeling, evaluation, and deployment phases might be less relevant than the data understanding and preparation phases. However, it is still important to consider some of the questions raised during these later phases for long-term planning and future data mining goals.

CRISP-DM in IBM SPSS Modeler

IBM SPSS Modeler incorporates the CRISP-DM methodology in two ways to provide unique support for effective data mining.

- The CRISP-DM project tool helps you organize project streams, output, and annotations according to the phases of a typical data mining project. You can produce reports at any time during the project based on the notes for streams and CRISP-DM phases.
- Help for CRISP-DM guides you through the process of conducting a data mining project. The help system includes tasks lists for each step as well as examples of how CRISP-DM works in the real world. You can access CRISP-DM Help by choosing **CRISP-DM Help** from the main window Help menu.

CRISP-DM Project Tool

The CRISP-DM project tool provides a structured approach to data mining that can help ensure your project's success. It is essentially an extension of the standard IBM SPSS Modeler project tool. In fact, you can toggle between the CRISP-DM view and the standard Classes view to see your streams and output organized by type or by phases of CRISP-DM.

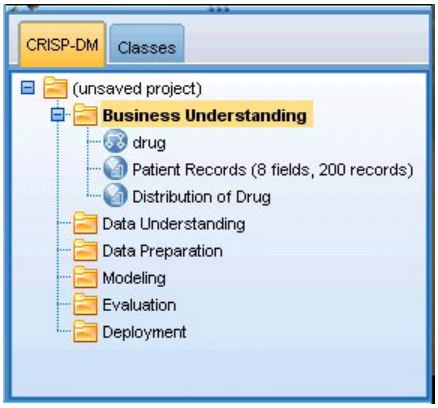

Figure 2. CRISP-DM project tool

Using the CRISP-DM view of the project tool, you can:

- Organize a project's streams and output according to data mining phases.
- Take notes on your organization's goals for each phase.
- Create custom tooltips for each phase.
- Take notes on the conclusions drawn from a particular graph or model.
- Generate an HTML report or update for distribution to the project team.

Help for CRISP-DM

IBM SPSS Modeler offers an online guide for the non-proprietary CRISP-DM process model. The guide is organized by project phases and provides the following support:

- An overview and task list for each phase of CRISP-DM
- Help on producing reports for various milestones
- Real-world examples illustrating how a project team can use CRISP-DM to light the way for data mining
- Links to additional resources on CRISP-DM

You can access CRISP-DM Help by choosing **CRISP-DM Help** from the main window Help menu.

Additional resources

In addition to IBM SPSS Modeler support for CRISP-DM, there are several ways to expand your understanding of data mining processes.

- Read the CRISP-DM manual, created by the CRISP-DM consortium and supplied with this release.
- Read *Data Mining with Confidence*, copyright 2002 by SPSS Inc., ISBN 1-56827-287-1.

Chapter 2. Business Understanding

Business Understanding Overview

Even before working in IBM SPSS Modeler, you should take the time to explore what your organization expects to gain from data mining. Try to involve as many key people as possible in these discussions and document the results. The final step of this CRISP-DM phase discusses how to produce a project plan using the information gathered here.

Although this research may seem dispensable, it's not. Getting to know the business reasons for your data mining effort helps to ensure that everyone is on the same page before expending valuable resources.

Determining Business Objectives

Your first task is to try to gain as much insight as possible into the business goals for data mining. This may not be as easy as it seems, but you can minimize later risk by clarifying problems, goals, and resources.

The CRISP-DM methodology provides a structured way for you to accomplish this.

Task List

- Start gathering background information about the current business situation.
- Document specific business objectives decided upon by key decision makers.
- Agree upon criteria used to determine data mining success from a business perspective.

E-Retail Example--Finding Business Objectives

A Web-Mining Scenario Using CRISP-DM

As more companies make the transition to selling over the Web, an established computer/electronics e-retailer is facing increasing competition from newer sites. Faced with the reality that Web stores are cropping up as fast (or faster!) than customers are migrating to the Web, the company must find ways to remain profitable despite the rising costs of customer acquisition. One proposed solution is to cultivate existing customer relationships in order to maximize the value of each of the company's current customers.

Thus, a study is commissioned with the following objectives:

- Improve cross-sales by making better recommendations.
- Increase customer loyalty with a more personalized service.

Tentatively, the study will be judged a success if:

- Cross-sales increase by 10%.
- Customers spend more time and see more pages on the site per visit.
- The study finishes on time and under budget.

Compiling the Business Background

Understanding your organization's business situation helps you know what you're working with in terms of:

- Available resources (personnel and material)
- Problems
- Goals

You'll need to do a bit of research on the current business situation in order to find real answers to questions that can impact the outcome of the data mining project.

Task 1--Determine Organizational Structure

- Develop organizational charts to illustrate corporate divisions, departments, and project groups. Be sure to include managers' names and responsibilities.
- Identify key individuals in the organization.
- Identify an internal sponsor who will provide financial support and/or domain expertise.
- Determine whether there is a steering committee and procure a list of members.
- Identify business units that will be affected by the data mining project.

Task 2--Describe Problem Area

- Identify the problem area, such as marketing, customer care, or business development.
- Describe the problem in general terms.
- Clarify the prerequisites of the project. What are the motivations behind the project? Does the business already use data mining?
- Check on the status of the data mining project within the business group. Has the effort been approved, or does data mining need to be "advertised" as a key technology for the business group?
- If necessary, prepare informational presentations on data mining to your organization.

Task 3--Describe Current Solution

- Describe any solutions currently used to address the business problem.
- Describe the advantages and disadvantages of the current solution. Also, address the level of acceptance this solution has had within the organization.

Defining Business Objectives

This is where things get specific. As a result of your research and meetings, you should construct a concrete primary objective agreed upon by the project sponsors and other business units affected by the results. This goal will eventually be translated from something as nebulous as "reducing customer churn" to specific data mining objectives that will guide your analytics.

Task List

Be sure to take notes on the following points for later incorporation into the project plan. Remember to keep goals realistic.

- Describe the problem you want to solve using data mining.
- Specify all business questions as precisely as possible.
- Determine any other business requirements (such as not losing any existing customers while increasing cross-sell opportunities).
- Specify expected benefits in business terms (such as reducing churn among high-value customers by 10%).

Business Success Criteria

The goal ahead may be clear, but will you know once you're there? It's important to define the nature of business success for your data mining project before proceeding further. Success criteria fall into two categories:

- **Objective.** These criteria can be as simple as a specific increase in the accuracy of audits or an agreed-upon reduction in churn.
- **Subjective.** Subjective criteria such as "discover clusters of effective treatments" are more difficult to pin down, but you can agree upon who makes the final decision.

Task List

- As precisely as possible, document the success criteria for this project.
- Make sure each business objective has a correlative criterion for success.
- Align the arbiters of the subjective measurements of success. If possible, take notes on their expectations.

Assessing the Situation

Now that you have a clearly defined goal, it's time to make an assessment of where you are right now. This step involves asking questions such as:

- What sort of data are available for analysis?
- Do you have the personnel needed to complete the project?
- What are the biggest risk factors involved?
- Do you have a contingency plan for each risk?

E-Retail Example--Assessing the Situation

A Web-Mining Scenario Using CRISP-DM

This is the electronics e-retailer's first attempt at Web mining, and the company has decided to consult a data mining specialist to help in getting started. One of the first tasks the consultant faces is to assess the company's resources for data mining.

Personnel. It's clear that there is in-house expertise with managing server logs and product and purchase databases, but little experience in data warehousing and data cleaning for analysis. Thus, a database specialist may also be consulted. Since the company hopes the results of the study will become part of a continuing Web-mining process, management must also consider whether any positions created during the current effort will be permanent ones.

Data. Since this is an established company, there is plenty of Web log and purchase data to draw from. In fact, for this initial study, the company will restrict the analysis to customers who have "registered" on the site. If successful, the program can be expanded.

Risks. Aside from the monetary outlays for the consultants and the time spent by employees on the study, there is not a great deal of immediate risk in this venture. However, time is always important, so this initial project is scheduled for a single financial quarter.

Also, there is not a lot of extra cash flow at the moment, so it is imperative that the study come in under budget. If either of these goals should be in danger, the business managers have suggested that the project's scope should be reduced.

Resource Inventory

Taking an accurate inventory of your resources is indispensable. You can save a lot of time and headaches by taking a real look at hardware, data sources, and personnel issues.

Task 1--Research Hardware Resources

- What hardware do you need to support?

Task 2--Identify Data Sources and Knowledge Stores

- Which data sources are available for data mining? Take note of data types and formats.
- How are the data stored? Do you have live access to data warehouses or operational databases?
- Do you plan to purchase external data, such as demographic information?
- Are there any security issues preventing access to required data?

Task 3--Identify Personnel Resources

- Do you have access to business and data experts?
- Have you identified database administrators and other support staff that may be needed?

Once you have asked these questions, include a list of contacts and resources for the phase report.

Requirements, Assumptions, and Constraints

Your efforts are more likely to pay off if you make an honest assessment of liabilities to the project. Making these concerns as explicit as possible will help to avert future problems.

Task 1--Determine Requirements

The fundamental requirement is the business goal discussed earlier, but consider the following:

- Are there security and legal restrictions on the data or project results?
- Is everyone aligned on the project scheduling requirements?
- Are there requirements on results deployment (for example, publishing to the Web or reading scores into a database)?

Task 2--Clarify Assumptions

- Are there economic factors that might affect the project (for example, consulting fees or competitive products)?
- Are there data quality assumptions?
- How does the project sponsor/management team expect to view the results? In other words, do they want to understand the model itself or simply view the results?

Task 3--Verify Constraints

- Do you have all passwords required for data access?
- Have you verified all legal constraints on data usage?
- Are all financial constraints covered in the project budget?

Risks and Contingencies

It is also wise to consider possible risks over the course of the project. Types of risks include:

- Scheduling (What if the project takes longer than anticipated?)
- Financial (What if the project sponsor encounters budgetary problems?)
- Data (What if the data are of poor quality or coverage?)
- Results (What if the initial results are less dramatic than expected?)

After you have considered the various risks, come up with a contingency plan to help avert disaster.

Task List

- Document each possible risk.
- Document a contingency plan for each risk.

Terminology

To ensure that business and data mining teams are "speaking the same language," you should consider compiling a glossary of technical terms and buzzwords that need clarification. For example, if "churn" for your business has a particular and unique meaning, it is worth explicitly stating that for the benefit of the whole team. Likewise, the team may benefit from clarification of the usage of a gains chart.

Task List

- Keep a list of terms or jargon confusing to team members. Include both business and data mining terminology.
- Consider publishing the list on the intranet or in other project documentation.

Cost/Benefit Analysis

This step answers the question, **What is your bottom line?** As part of the final assessment, it's critical to compare the costs of the project with the potential benefits of success.

Task List

Include in your analysis estimated costs for:

- Data collection and any external data used
- Results deployment
- Operating costs

Then, take into account the benefits of:

- The primary objective being met
- Additional insights generated from data exploration
- Possible benefits from better data understanding

Determining Data Mining Goals

Now that the business goal is clear, it's time to translate it into a data mining reality. For example, the business objective to "reduce churn" can be translated into a data mining goal that includes:

- Identifying high-value customers based on recent purchase data
- Building a model using available customer data to predict the likelihood of churn for each customer
- Assigning each customer a rank based on both churn propensity and customer value

These data mining goals, if met, can then be used by the business to reduce churn among the most valuable customers.

As you can see, business and technology must work hand-in-hand for effective data mining. Read on for specific tips on how to determine data mining goals.

Data Mining Goals

As you work with business and data analysts to define a technical solution to the business problem, remember to keep things concrete.

Task List

- Describe the **type** of data mining problem, such as clustering, prediction, or classification.
- Document technical goals using specific units of time, such as predictions with a three-month validity.
- If possible, provide actual numbers for desired outcomes, such as producing churn scores for 80% of existing customers.

E-Retail Example--Data Mining Goals

A Web-Mining Scenario Using CRISP-DM

With the help of its data mining consultant, the e-retailer has been able to translate the company's business objectives into data mining terms. The goals for the initial study to be completed this quarter are:

- Use historical information about previous purchases to generate a model that links "related" items. When users look at an item description, provide links to other items in the related group (**market basket analysis**).
- Use Web logs to determine what different customers are trying to find, and then redesign the site to highlight these items. Each different customer "type" will see a different main page for the site (**profiling**).

- Use Web logs to try to predict where a person is going next, given where he or she came from and has been on your site (**sequence analysis**).

Data Mining Success Criteria

Success must also be defined in technical terms to keep your data mining efforts on track. Use the data mining goal determined earlier to formulate benchmarks for success. IBM SPSS Modeler provides tools such as the Evaluation node and the Analysis node to help you analyze the accuracy and validity of your results.

Task List

- Describe the methods for model assessment (for example, accuracy, performance, etc.).
- Define benchmarks for evaluating success. Provide specific numbers.
- Define subjective measurements as best you can and determine the arbiter of success.
- Consider whether the successful deployment of model results is part of data mining success. Start planning now for deployment.

Producing a Project Plan

At this point, you're ready to produce a plan for the data mining project. The questions you have asked so far and the business and data mining goals you have formulated will form the basis for this road map.

Writing the Project Plan

The project plan is the master document for all of your data mining work. If done well, it can inform everyone associated with the project of the goals, resources, risks, and schedule for all phases of data mining. You may want to publish the plan, as well as documentation gathered throughout this phase, to your company's intranet.

Task List

When creating the plan, be sure you've answered the following questions:

- Have you discussed the project tasks and proposed plan with everyone involved?
- Are time estimates included for all phases or tasks?
- Have you included the effort and resources needed to deploy the results or business solution?
- Are decision points and review requests highlighted in the plan?
- Have you marked phases where multiple iterations typically occur, such as modeling?

Sample Project Plan

The overview plan for the study is as shown in the table below.

Table 1. Sample project plan overview

Phase	Time	Resources	Risks
Business understanding	1 week	All analysts	Economic change
Data understanding	3 weeks	All analysts	Data problems, technology problems
Data preparation	5 weeks	Data mining consultant, some database analyst time	Data problems, technology problems
Modeling	2 weeks	Data mining consultant, some database analyst time	Technology problems, inability to find adequate model

Table 1. Sample project plan overview (continued)			
Phase	**Time**	**Resources**	**Risks**
Evaluation	1 week	All analysts	Economic change, inability to implement results
Deployment	1 week	Data mining consultant, some database analyst time	Economic change, inability to implement results

Assessing Tools and Techniques

Since you've already chosen to use IBM SPSS Modeler as your tool for data mining success, you can use this step to research which data mining techniques are most appropriate for your business needs. IBM SPSS Modeler offers a full range of tools for each phase of data mining. To decide when to use the various techniques, consult the modeling section of the online Help.

Ready for the next step?

Before exploring data and beginning work in IBM SPSS Modeler, be sure you have answered the following questions.

From a business perspective:

- What does your business hope to gain from this project?
- How will you define the successful completion of our efforts?
- Do you have the budget and resources needed to reach our goals?
- Do you have access to all the data needed for this project?
- Have you and your team discussed the risks and contingencies associated with this project?
- Do the results of your cost/benefit analysis make this project worthwhile?

After you've answered the above questions, did you translate those answers into a data mining goal?

From a data mining perspective:

- How specifically can data mining help you meet your business goals?
- Do you have an idea about which data mining techniques might produce the best results?
- How will you know when your results are accurate or effective enough? (*Have we set a measurement of data mining success?*)
- How will the modeling results be deployed? Have you considered deployment in your project plan?
- Does the project plan include all phases of CRISP-DM?
- Are risks and dependencies called out in the plan?

If you can answer "yes" to the above questions, then you're ready to take a closer look at the data.

Chapter 3. Data Understanding

Data Understanding Overview

The data understanding phase of CRISP-DM involves taking a closer look at the data available for mining. This step is critical in avoiding unexpected problems during the next phase--data preparation--which is typically the longest part of a project.

Data understanding involves accessing the data and exploring it using tables and graphics that can be organized in IBM SPSS Modeler using the CRISP-DM project tool. This enables you to determine the quality of the data and describe the results of these steps in the project documentation.

Collecting Initial Data

At this point in CRISP-DM, you're ready to access data and bring it into IBM SPSS Modeler. Data come from a variety of sources, such as:

- **Existing data.** This includes a wide variety of data, such as transactional data, survey data, Web logs, etc. Consider whether the existing data are enough to meet your needs.
- **Purchased data.** Does your organization use supplemental data, such as demographics? If not, consider whether it may be needed.
- **Additional data.** If the above sources don't meet your needs, you may need to conduct surveys or begin additional tracking to supplement the existing data stores.

Task List

Take a look at the data in IBM SPSS Modeler and consider the following questions. Be sure to take notes on your findings. See the topic “Writing a Data Collection Report” on page 14 for more information.

- Which attributes (columns) from the database seem most promising?
- Which attributes seem irrelevant and can be excluded?
- Is there enough data to draw generalizable conclusions or make accurate predictions?
- Are there too many attributes for your modeling method of choice?
- Are you merging various data sources? If so, are there areas that might pose a problem when merging?
- Have you considered how missing values are handled in each of your data sources?

E-Retail Example--Initial Data Collection

A Web-Mining Scenario Using CRISP-DM

The e-retailer in this example uses several important data sources, including:

Web logs. The raw access logs contain all of the information on how customers navigate the Web site. References to image files and other non-informative entries in the Web logs will need to be removed as part of the data preparation process.

Purchase data. When a customer submits an order, all of the information pertinent to that order is saved. The orders in the purchase database need to be mapped to the corresponding sessions in the Web logs.

Product database. The product attributes may be useful when determining "related" products. The product information needs to be mapped to the corresponding orders.

Customer database. This database contains extra information collected from registered customers. The records are by no means complete, because many customers do not fill out questionnaires. The customer information needs to be mapped to the corresponding purchases and sessions in the Web logs.

At this moment, the company has no plans to purchase external databases or spend money conducting surveys because its analysts are busy managing the data they currently have. At some point, however, they may want to consider an extended deployment of data mining results, in which case purchasing additional demographic data for unregistered customers may be quite useful. It may also be useful to have demographic information to see how the e-retailer's customer base differs from the average Web shopper.

Writing a Data Collection Report

Using the material gathered in the previous step, you can begin to write a data collection report. Once complete, the report can be added to the project Web site or distributed to the team. It can also be combined with the reports prepared in the next steps--data description, exploration, and quality verification. These reports will guide your work throughout the data preparation phase.

Describing Data

There are many ways to describe data, but most descriptions focus on the quantity and quality of the data--how much data is available and the condition of the data. Listed below are some key characteristics to address when describing data.

- **Amount of data.** For most modeling techniques, there are trade-offs associated with data size. Large data sets can produce more accurate models, but they can also lengthen the processing time. Consider whether using a subset of data is a possibility. When taking notes for the final report, be sure to include size statistics for all data sets, and remember to consider both the number of records as well as fields (attributes) when describing data.
- **Value types.** Data can take a variety of formats, such as **numeric**, **categorical** (string), or **Boolean** (true/false). Paying attention to value type can head off problems during later modeling.
- **Coding schemes.** Frequently, values in the database are representations of characteristics such as gender or product type. For example, one data set may use *M* and *F* to represent *male* and *female*, while another may use the numeric values *1* and *2*. Note any conflicting schemes in the data report.

With this knowledge in hand, you are now ready to write the data description report and share your findings with a larger audience.

E-Retail Example--Describing Data

A Web-Mining Scenario Using CRISP-DM

There are many records and attributes to process in a Web-mining application. Even though the e-retailer conducting this data mining project has limited the initial study to the approximately 30,000 customers who have registered on the site, there are still millions of records in the Web logs.

Most of the value types in these data sources are symbolic, whether they are dates and times, Web pages accessed, or answers to multiple-choice questions from the registration questionnaire. Some of these variables will be used to create new variables that are numeric, such as number of Web pages visited and time spent at the Web site. The few existing numeric variables in the data sources include the number of each product ordered, the amount spent during a purchase, and product weight and dimension specifications from the product database.

There is little overlap in the coding schemes for the various data sources because the data sources contain very different attributes. The only variables that overlap are "keys," such as the customer IDs and product codes. These variables must have identical coding schemes from data source to data source; otherwise, it would be impossible to merge the data sources. Some additional data preparation will be necessary to recode these key fields for merging.

Writing a Data Description Report

To proceed effectively with your data mining project, consider the value of producing an accurate data description report using the following metrics:

Data Quantity

- What is the format of the data?
- Identify the method used to capture the data--for example, ODBC.
- How large is the database (in numbers of rows and columns)?

Data Quality

- Does the data include characteristics relevant to the business question?
- What data types are present (symbolic, numeric, etc.)?
- Did you compute basic statistics for the key attributes? What insight did this provide into the business question?
- Are you able to prioritize relevant attributes? If not, are business analysts available to provide further insight?

Exploring Data

Use this phase of CRISP-DM to explore the data with the tables, charts, and other visualization tools available in IBM SPSS Modeler. Such analyses can help to address the data mining goal constructed during the business understanding phase. They can also help to formulate hypotheses and shape the data transformation tasks that take place during data preparation.

E-Retail Example--Exploring Data

A Web-Mining Scenario Using CRISP-DM

Although CRISP-DM suggests conducting an initial exploration at this point, data exploration is difficult, if not impossible, on raw Web logs, as our e-retailer has found out. Typically, Web log data must be processed first in the data preparation phase to produce data that can be meaningfully explored. This departure from CRISP-DM underscores the fact that the process can and should be customized for your particular data mining needs. CRISP-DM is cyclical, and data miners typically move back and forth between phases.

Although Web logs must be processed before exploration, the other data sources available to the e-retailer are more amenable to exploration. Using the purchase database for exploration reveals interesting summaries about customers, such as how much they spend, how many items they buy per purchase, and where they come from. Summaries of the customer database will show the distribution of responses to the items on the registration questionnaire.

Exploration is also useful for looking for errors in the data. While most of the data sources are automatically generated, information in the product database was entered by hand. Some quick summaries of listed product dimensions will help to discover typos, such as "119-inch" (instead of "19-inch") monitor.

Writing a Data Exploration Report

As you create graphs and run statistics on the available data, start forming hypotheses about how the data can answer the technical and business goals.

Task List

Take notes on your findings for inclusion in the data exploration report. Be sure to answer the following questions:

- What sort of hypotheses have you formed about the data?
- Which attributes seem promising for further analysis?
- Have your explorations revealed new characteristics about the data?
- How have these explorations changed your initial hypothesis?
- Can you identify particular subsets of data for later use?

- Take another look at your data mining goals. Has this exploration altered the goals?

Verifying Data Quality

Data are rarely perfect. In fact, most data contain coding errors, missing values, or other types of inconsistencies that make analysis tricky at times. One way to avoid potential pitfalls is to conduct a thorough quality analysis of available data before modeling.

The reporting tools in IBM SPSS Modeler (such as the Data Audit, Table and other output nodes) can help you look for the following types of problems:

- **Missing data** include values that are blank or coded as a non-response (such as *$null$*, *?*, or *999*).
- **Data errors** are usually typographical errors made in entering the data.
- **Measurement errors** include data that are entered correctly but are based on an incorrect measurement scheme.
- **Coding inconsistencies** typically involve nonstandard units of measurement or value inconsistencies, such as the use of both *M* and *male* for gender.
- **Bad metadata** include mismatches between the apparent meaning of a field and the meaning stated in a field name or definition.

Be sure to take notes on such quality concerns. See the topic “Writing a Data Quality Report” on page 16 for more information.

E-Retail Example--Verifying Data Quality

A Web-Mining Scenario Using CRISP-DM

The verification of data quality is often accomplished during the course of the description and exploration processes. Some of the issues encountered by the e-retailer include:

Missing Data. The known missing data includes the unanswered questionnaires by some of the registered users. Without the extra information provided by the questionnaire, these customers may have to be left out of some of the subsequent models.

Data Errors. Most of the data sources are automatically generated, so this is not a great worry. Typographical errors in the product database can be found during the exploration process.

Measurement Errors. The greatest potential source for measurement error is the questionnaire. If any of the items are ill-advised or poorly worded, they may not provide the information the e-retailer hopes to obtain. Again, during the exploration process, it is important to pay special attention to items that have an unusual distribution of answers.

Writing a Data Quality Report

Based on your exploration and verification of data quality, you're now ready to prepare a report that will guide the next phase of CRISP-DM. See the topic “Verifying Data Quality” on page 16 for more information.

Task List

As discussed earlier, there are several types of data quality problems. Before moving to the next step, consider the following quality concerns and plan for a solution. Document all responses in the data quality report.

- Have you identified missing attributes and blank fields? If so, is there meaning behind such missing values?
- Are there spelling inconsistencies that may cause problems in later merges or transformations?
- Have you explored deviations to determine whether they are "noise" or phenomena worth analyzing further?

- Have you conducted a plausibility check for values? Take notes on any apparent conflicts (such as teenagers with high income levels).
- Have you considered excluding data that has no impact on your hypotheses?
- Are the data stored in flat files? If so, are the delimiters consistent among files? Does each record contain the same number of fields?

Ready for the next step?

Before preparing the data for modeling in IBM SPSS Modeler, consider the following points:

How well do you understand the data?

- Are all data sources clearly identified and accessed? Are you aware of any problems or restrictions?
- Have you identified key attributes from the available data?
- Did these attributes help you to formulate hypotheses?
- Have you noted the size of all data sources?
- Are you able to use a subset of data where appropriate?
- Have you computed basic statistics for each attribute of interest? Did meaningful information emerge?
- Did you use exploratory graphics to gain further insight into key attributes? Did this insight reshape any of your hypotheses?
- What are the data quality issues for this project? Do you have a plan to address these issues?
- Are the data preparation steps clear? For instance, do you know which data sources to merge and which attributes to filter or select?

Now that you're armed with both business and data understanding, it's time to use IBM SPSS Modeler to prepare your data for modeling.

Chapter 4. Data Preparation

Data Preparation Overview

Data preparation is one of the most important and often time-consuming aspects of data mining. In fact, it is estimated that data preparation usually takes 50-70% of a project's time and effort. Devoting adequate energy to the earlier business understanding and data understanding phases can minimize this overhead, but you still need to expend a good amount of effort preparing and packaging the data for mining.

Depending on your organization and its goals, data preparation typically involves the following tasks:

- Merging data sets and/or records
- Selecting a sample subset of data
- Aggregating records
- Deriving new attributes
- Sorting the data for modeling
- Removing or replacing blank or missing values
- Splitting into training and test data sets

Selecting Data

Based upon the initial data collection conducted in the previous CRISP-DM phase, you are ready to begin selecting the data relevant to your data mining goals. Generally, there are two ways to select data:

- **Selecting items (rows)** involves making decisions such as which accounts, products, or customers to include.
- **Selecting attributes or characteristics (columns)** involves making decisions about the use of characteristics such as transaction amount or household income.

E-Retail Example--Selecting Data

A Web-Mining Scenario Using CRISP-DM

Many of the e-retailer's decisions about which data to select have already been made in earlier phases of the data mining process.

Selecting items. The initial study will be limited to the (approximately) 30,000 customers who have registered on the site, so filters need to be set up to exclude purchases and Web logs of nonregistered customers. Other filters should be established to remove calls to image files and other non-informative entries in the Web logs.

Selecting attributes. The purchase database will contain sensitive information about the e-retailer's customers, so it is important to filter attributes such as the customer name, address, phone number, and credit card numbers.

Including or Excluding Data

As you decide upon subsets of data to include or exclude, be sure to document the rationale behind your decisions.

Questions to Consider

- Is a given attribute relevant to you data mining goals?
- Does the quality of a particular data set or attribute preclude the validity of your results?
- Can you salvage such data?

- Are there any constraints on using particular fields such as *gender* or *race*?

Are your decisions here different than the hypotheses formulated in the data understanding phase? If so, be sure to document your reasoning in the project report.

Cleaning Data

Cleaning your data involves taking a closer look at the problems in the data that you've chosen to include for analysis. There are several ways to clean data using the Record and Field Operation nodes in IBM SPSS Modeler.

Table 2. Cleaning data

Data Problem	Possible Solution
Missing data	Exclude rows or characteristics. Or, fill blanks with an estimated value.
Data errors	Use logic to manually discover errors and replace. Or, exclude characteristics.
Coding inconsistencies	Decide upon a single coding scheme, then convert and replace values.
Missing or bad metadata	Manually examine suspect fields and track down correct meaning.

The Data Quality Report prepared during the data understanding phase contains details about the types of problems particular to your data. You can use it as a starting point for data manipulation in IBM SPSS Modeler.

E-Retail Example--Cleaning Data

A Web-Mining Scenario Using CRISP-DM

The e-retailer uses the data cleaning process to address the problems noted in the data quality report.

Missing data. Customers who did not complete the online questionnaire may have to be left out of some of the models later on. These customers could be asked again to fill out the questionnaire, but this will take time and money that the e-retailer cannot afford to spend. What the e-retailer can do is model the purchasing differences between customers who do and do not answer the questionnaire. If these two sets of customers have similar purchasing habits, the missing questionnaires are less worrisome.

Data errors. Errors found during the exploration process can be corrected here. For the most part, though, proper data entry is enforced on the Web site before a customer submits a page to the back-end database.

Measurement errors. Poorly worded items on the questionnaire can greatly affect the quality of the data. As with missing questionnaires, this is a difficult problem because there may not be time or money available to collect answers to a new replacement question. For problematic items, the best solution may be to go back to the selection process and filter these items from further analyses.

Writing a Data Cleaning Report

Reporting your data-cleaning efforts is essential for tracking alterations to the data. Future data mining projects will benefit from having the details of your work readily available.

Task List

It's a good idea to consider the following questions when writing the report:

- What types of noise occurred in the data?
- What approaches did you use to remove the noise? Which techniques were successful?

- Are there any cases or attributes that could not be salvaged? Be sure to note data excluded due to noise.

Constructing New Data

It is frequently the case that you'll need to construct new data. For example, it may be useful to create a new column flagging the purchase of an extended warranty for each transaction. This new field, *purchased_warranty*, can easily be generated using a Set to Flag node in IBM SPSS Modeler.

There are two ways to construct new data:

- Deriving attributes (columns or characteristics)
- Generating records (rows)

IBM SPSS Modeler offers a multitude of ways to construct data using its Record and Field Operations nodes.

E-Retail Example--Constructing Data

A Web-Mining Scenario Using CRISP-DM

The processing of Web logs can create many new attributes. For the events recorded in the logs, the e-retailer will want to create timestamps, identify visitors and sessions, and note the page accessed and the type of activity the event represents. Some of these variables will be used to create more attributes, such as the time between events within a session.

Further attributes can be created as a result of a merge or other data restructuring. For example, when the event-per-row Web logs are "rolled up" so that each row is a session, new attributes recording the total number of actions, total time spent, and total purchases made during the session will be created. When the Web logs are merged with the customer database so that each row is a customer, new attributes recording the number of sessions, total number of actions, total time spent, and total purchases made by each customer will be created.

After constructing new data, the e-retailer goes through an exploration process to make sure that the data creation was performed correctly.

Deriving Attributes

In IBM SPSS Modeler, you can use the following Field Operations nodes to derive new attributes:

- Create new fields derived from existing ones using a **Derive node**.
- Create a flag field using a **Set to Flag node**.

Task List

- Consider the data requirements for modeling when deriving attributes. Does the modeling algorithm expect a particular type of data, such as numeric? If so, perform the necessary transformations.
- Do the data need be normalized before modeling?
- Can missing attributes be constructed using aggregation, averaging, or induction?
- Based upon your background knowledge, are there important facts (such as length of time spent at the Web site) that can be derived from existing fields?

Integrating Data

It is not uncommon to have multiple data sources for the same set of business questions. For example, you may have access to mortgage loan data as well as purchased demographic data for the same set of clients. If these data sets contain the same unique identifier (such as social security number), you can merge them in IBM SPSS Modeler using this key field.

There are two basic methods of integrating data:

- **Merging** data involves merging two data sets with similar records but different attributes. The data is merged using the same key identifier for each record (such as customer ID). The resulting data increases in columns or characteristics.
- **Appending** data involves integrating two or more data sets with similar attributes but different records. The data is integrated based upon a similar fields (such as product name or contract length).

E-Retail Example--Integrating Data

A Web-Mining Scenario Using CRISP-DM

With multiple data sources, there are many different ways in which the e-retailer can integrate data:

- **Adding customer and product attributes to event data.** In order to model Web log events using attributes from other databases, any customer ID, product number, and purchase order number associated with each event must be correctly identified and the corresponding attributes merged to the processed Web logs. Note that the merged file replicates customer and product information every time a customer or product is associated with an event.
- **Adding purchase and Web log information to customer data.** In order to model the value of a customer, their purchases and session information must be picked out of the appropriate databases, totaled, and merged with the customer database. This involves the creation of new attributes as discussed in the constructing data process.

After integrating databases, the e-retailer goes through an exploration process to make sure that the data merge was performed correctly.

Integration Tasks

Integrating data can become complex if you have not spent adequate time developing an understanding of your data. Give some thought to items and attributes that seem most relevant to the data mining goals and then get started integrating your data.

Task List

- Using Merge or Append nodes in IBM SPSS Modeler, integrate the data sets considered useful for modeling.
- Consider saving the resulting output before proceeding to modeling.
- After merging, data can be simplified by **aggregating** values. Aggregation means that new values are computed by summarizing information from multiple records and/or tables.
- You may also need to generate new records (such as the average deduction from several years of combined tax returns).

Formatting Data

As a final step before model building, it is helpful to check whether certain techniques require a particular format or order to the data. For example, it is not uncommon that a sequence algorithm requires the data to be presorted before running the model. Even if the model can perform the sorting for you, it may save processing time to use a Sort node prior to modeling.

Task List

Consider the following questions when formatting data:

- Which models do you plan to use?
- Do these models require a particular data format or order?

If changes are recommended, the processing tools in IBM SPSS Modeler can help you apply the necessary data manipulation.

Ready for modeling?

Before building models in IBM SPSS Modeler, be sure you have answered the following questions.

- Are all the data accessible from within IBM SPSS Modeler?
- Based upon your initial exploration and understanding, were you able to select relevant subsets of data?
- Have you cleaned the data effectively or removed unsalvageable items? Document any decisions in the final report.
- Are multiple data sets integrated properly? Were there any merging problems that should be documented?
- Have you researched the requirements of the modeling tools that you plan to use?
- Are there any formatting issues you can address before modeling? This includes both required formatting concerns as well as tasks that may reduce modeling time.

If you can answer the above questions, then you're ready for the crux of data mining--modeling.

Chapter 5. Modeling

Modeling Overview

This is the point at which your hard work begins to pay off. The data you spent time preparing are brought into the analysis tools in IBM SPSS Modeler, and the results begin to shed some light on the business problem posed during Business Understanding.

Modeling is usually conducted in multiple iterations. Typically, data miners run several models using the default parameters and then fine-tune the parameters or revert to the data preparation phase for manipulations required by their model of choice. It is rare for an organization's data mining question to be answered satisfactorily with a single model and a single execution. This is what makes data mining so interesting--there are many ways to look at a given problem, and IBM SPSS Modeler offers a wide variety of tools to help you do so.

Selecting Modeling Techniques

Although you may already have some idea about which types of modeling are most appropriate for your organization's needs, now is the time to make some firm decisions about which ones to use. Determining the most appropriate model will typically be based on the following considerations:

- **The data types available for mining.** For example, are the fields of interest categorical (symbolic)?
- **Your data mining goals.** Do you simply want to gain insight into transactional data stores and unearth interesting purchase patterns? Or do you need to produce a score indicating, for example, propensity to default on a student loan?
- **Specific modeling requirements.** Does the model require a particular data size or type? Do you need a model with easily presentable results?

For more information on the model types in IBM SPSS Modeler and their requirements, see the IBM SPSS Modeler documentation or online Help.

E-Retail Example--Modeling Techniques

The modeling techniques employed by the e-retailer are driven by the company's data mining goals:

Improved recommendations. At its simplest, this involves clustering purchase orders to determine which products are most often bought together. Customer data, and even visit records, can be added for richer results. The two-step or Kohonen network clustering techniques are suited for this type of modeling. Afterward, the clusters can be profiled using a C5.0 ruleset to determine which recommendations are most appropriate at any point during a customer's visit.

Improved site navigation. For now, the e-retailer will focus on identifying pages that are often used but require several clicks for the user to find. This entails applying a sequencing algorithm to the Web logs in order to generate the "unique paths" customers take through the Web site, and then specifically looking for sessions that have a lot of page visits without (or before) an action taken. Later, in a more in-depth analysis, clustering techniques can be used to identify different "types" of visits and visitors, and the site content can be organized and presented according to type.

Choosing the Right Modeling Techniques

Many modeling techniques are available in IBM SPSS Modeler. Frequently, data miners use more than one to approach the problem from a number of directions.

Task List

When deciding on which model(s) to use, consider whether the following issues have an impact on your choices:

- Does the model require the data to be split into test and training sets?
- Do you have enough data to produce reliable results for a given model?
- Does the model require a certain level of data quality? Can you meet this level with the current data?
- Are your data the proper type for a particular model? If not, can you make the necessary conversions using data manipulation nodes?

For more information on the model types in IBM SPSS Modeler and their requirements, see the IBM SPSS Modeler documentation or online Help.

Modeling Assumptions

As you begin to narrow down your modeling tools of choice, take notes on the decision-making process. Document any data assumptions as well as any data manipulations made to meet the model's requirements.

For example, both the Logistic Regression and Neural Net nodes require the data types to be fully **instantiated** (data types are known) before execution. This means you will need to add a Type node to the stream and execute it to run the data through before building and running a model. Similarly, predictive models, such as C5.0, may benefit from rebalancing the data when predicting rules for rare events. When making this type of prediction, you can often get better results by inserting a Balance node into the stream and feeding the more balanced subset into the model.

Be sure to document these types of decisions.

Generating a Test Design

As a final step before actually building the model, you should take a moment to consider again how the model's results will be tested. There are two parts to generating a comprehensive test design:

- Describing the criteria for "goodness" of a model
- Defining the data on which these criteria will be tested

A model's **goodness** can be measured in several ways. For supervised models, such as C5.0 and C&R Tree, measurements of goodness typically estimate the error rate of a particular model. For unsupervised models, such as Kohonen cluster nets, measurements may include criteria such as ease of interpretation, deployment, or required processing time.

Remember, model building is an iterative process. This means that you will typically test the results of several models before deciding on the ones to use and deploy.

Writing a Test Design

The test design is a description of the steps you will take to test the models produced. Because modeling is an iterative process, it is important to know when to stop adjusting parameters and try another method or model.

Task List

When creating a test design, consider the following questions:

- What data will be used to test the models? Have you partitioned the data into train/test sets? (This is a commonly used approach in modeling.)
- How might you measure the success of supervised models (such as C5.0)?
- How might you measure the success of unsupervised models (such as Kohonen cluster nets)?
- How many times are you willing to rerun a model with adjusted settings before attempting another type of model?

E-Retail Example--Test Design

A Web-Mining Scenario Using CRISP-DM

The criteria by which the models are assessed depend on the models under consideration and the data mining goals:

Improved recommendations. Until the improved recommendations are presented to live customers, there is no purely objective way to assess them. However, the e-retailer may require the rules that generate the recommendations to be simple enough to make sense from a business perspective. Likewise, the rules should be complex enough to generate different recommendations for different customers and sessions.

Improved site navigation. Given the evidence of what pages customers access on the Web site, the e-retailer can objectively assess the updated site design in terms of ease of access to important pages. However, as with the recommendations, it is difficult to assess in advance how well customers will adjust to the reorganized site. If time and finances allow, some usability testing may be in order.

Building the Models

At this point, you should be well prepared to build the models you've spent so long considering. Give yourself time and room to experiment with a number of different models before making final conclusions. Most data miners typically build several models and compare the results before deploying or integrating them.

In order to track your progress with a variety of models, be sure to keep notes on the settings and data used for each model. This will help you to discuss the results with others and retrace your steps if necessary. At the end of the model-building process, you'll have three pieces of information to use in data mining decisions:

- **Parameter settings** include the notes you take on parameters that produce the best results.
- The actual **models** produced.
- **Descriptions of model results**, including performance and data issues that occurred during the execution of the model and exploration of its results.

E-Retail Example--Model Building

A Web-Mining Scenario Using CRISP-DM

Improved recommendations. Clusterings are produced for varying levels of data integration, starting with just the purchase database and then including related customer and session information. For each level of integration, clusterings are produced under varying parameter settings for the two-step and Kohonen network algorithms. For each of these clusterings, a few C5.0 rulesets are generated with different parameter settings.

Improved site navigation. The Sequence modeling node is used to generate customer paths. The algorithm allows the specification of a minimum support criterion, which is useful for focusing on the most common customer paths. Various settings for the parameters are tried.

Parameter Settings

Most modeling techniques have a variety of parameters or settings that can be adjusted to control the modeling process. For example, decision trees can be controlled by adjusting tree depth, splits, and a number of other settings. Typically, most people build a model first using the default options and then refine parameters during subsequent sessions.

Once you have determined the parameters that produce the most accurate results, be sure to save the stream and generated model nodes. Also, taking notes on the optimal settings can help when you decide to automate or rebuild the model with new data.

Running the Models

In IBM SPSS Modeler, running models is a straightforward task. Once you've inserted the model node into the stream and edited any parameters, simply execute the model to produce viewable results. Results

appear in the Generated Models navigator on the right side of the workspace. You can right-click a model to browse the results. For most models, you can insert the generated model into the stream to further evaluate and deploy the results. Models can be also be saved in IBM SPSS Modeler for easy reuse.

Model description

When examining the results of a model, be sure to take notes on your modeling experience. You can store notes with the model itself using the node annotations dialog box or the project tool.

Task list

For each model, record information such as:

- Can you draw meaningful conclusions from this model?
- Are there new insights or unusual patterns revealed by the model?
- Were there execution problems for the model? How reasonable was the processing time?
- Did the model have difficulties with data quality issues, such as a high number of missing values?
- Were there any calculation inconsistencies that should be noted?

Assessing the Model

Now that you have a set of initial models, take a closer look at them to determine which are accurate or effective enough to be final. Final can mean several things, such as "ready to deploy" or "illustrating interesting patterns." Consulting the test plan that you created earlier can help to make this assessment from your organization's point of view.

Comprehensive Model Assessment

For each model under consideration, it is a good idea to make a methodical assessment based on the criteria generated in your test plan. Here is where you may add the generated model to the stream and use evaluation charts or analysis nodes to analyze the effectiveness of the results. You should also consider whether the results make logical sense or whether they are too simplistic for your business goals (for example, a sequence that reveals purchases such as wine > wine > wine).

Once you've made an assessment, rank the models in order based on both objective (model accuracy) and subjective (ease of use or interpretation of results) criteria.

Task List

- Using the data mining tools in IBM SPSS Modeler, such as evaluation charts, analysis nodes, or cross-validation charts, evaluate the results of your model.
- Conduct a review of the results based on your understanding of the business problem. Consult data analysts or other experts who may have insight into the relevance of particular results.
- Consider whether a model's results are easily deployable. Does your organization require that results be deployed over the Web or sent back to the data warehouse?
- Analyze the impact of results on your success criteria. Do they meet the goals established during the business understanding phase?

If you were able to address the above issues successfully and believe that the current models meet your goals, it's time to move on to a more thorough evaluation of the models and a final deployment. Otherwise, take what you've learned and rerun the models with adjusted parameter settings.

E-Retail Example--Model Assessment

A Web-Mining Scenario Using CRISP-DM

Improved recommendations. One of the Kohonen networks and a two-step clustering each give reasonable results, and the e-retailer finds it difficult to choose between them. In time, the company

hopes to use both, accepting the recommendations that the two techniques agree on and studying in greater detail the situations in which they differ. With a little effort and applied business knowledge, the e-retailer can develop further rules to resolve differences between the two techniques.

The e-retailer also finds that the results that include the session information are surprisingly good. There is evidence to suggest that recommendations could be tied to site navigation. A ruleset, defining where the customer is likely to go next, could be used in real time to affect the site content directly as the customer is browsing.

Improved site navigation. The Sequence model provides the e-retailer with a high level of confidence that certain customer paths can be predicted, producing results that suggest a manageable number of changes to the site design.

Keeping Track of Revised Parameters

Based on what you've learned during model assessment, it's time to have another look at the models. You have two options here:

- Adjust the parameters of existing models.
- Choose a different model to address your data mining problem.

In both cases, you'll be returning to the building models task and iterate until the results are successful. Don't worry about repeating this step. It is extremely common for data miners to evaluate and rerun models several times before finding one that meets their needs. This is a good argument for building several models at once and comparing the results before adjusting the parameters for each.

Ready for the next step?

Before moving on to a final evaluation of the models, consider whether your initial assessment was thorough enough.

Task List

- Are you able to understand the results of the models?
- Do the model results make sense to you from a purely logical perspective? Are there apparent inconsistencies that need further exploration?
- From your initial glance, do the results seem to address your organization's business question?
- Have you used analysis nodes and lift or gains charts to compare and evaluate model accuracy?
- Have you explored more than one type of model and compared the results?
- Are the results of your model deployable?

If the results of your data modeling seem accurate and relevant, it's time to conduct a more through evaluation before a final deployment.

Chapter 6. Evaluation

Evaluation Overview

At this point, you've completed most of your data mining project. You've also determined, in the Modeling phase, that the models built are technically correct and effective according to the **data mining success criteria** that you defined earlier.

Before continuing, however, you should evaluate the results of your efforts using the **business success criteria** established at the beginning of the project. This is the key to ensuring that your organization can make use of the results you've obtained. Two types of results are produced by data mining:

- The final **models** selected in the previous phase of CRISP-DM.
- Any conclusions or inferences drawn from the models themselves as well as from the data mining process. These are known as **findings**.

Evaluating the Results

At this stage, you formalize your assessment of whether or not the project results meet the business success criteria. This step requires a clear understanding of the stated business goals, so be sure to include key decision makers in the project assessment.

Task List

First, you need to document your assessment of whether the data mining results meet the business success criteria. Consider the following questions in your report:

- Are your results stated clearly and in a form that can be easily presented?
- Are there particularly novel or unique findings that should be highlighted?
- Can you rank the models and findings in order of their applicability to the business goals?
- In general, how well do these results answer your organization's business goals?
- What additional questions have your results raised? How might you phrase these questions in business terms?

After you have evaluated the results, compile a list of approved models for inclusion in the final report. This list should include models that satisfy both the data mining and business goals of your organization.

E-Retail Example--Evaluating Results

A Web-Mining Scenario Using CRISP-DM

The overall results of the e-retailer's first experience with data mining are fairly easy to communicate from a business perspective: the study produced what are hoped to be better product recommendations and an improved site design. The improved site design is based on the customer browsing sequences, which show the site features that customers want but require several steps to reach. The evidence that the product recommendations are better is more difficult to convey, because the decision rules can become complicated. To produce the final report, the analysts will try to identify some general trends in the rulesets that can be more easily explained.

Ranking the Models. Because several of the initial models seemed to make business sense, ranking within that group was based on statistical criteria, ease of interpretation, and diversity. Thus, the model gave different recommendations for different situations.

New Questions. The most important question to come out of the study is, How can the e-retailer find out more about his or her customers? The information in the customer database plays an important role in forming the clusters for recommendations. While special rules are available for making recommendations

to customers whose information is missing, the recommendations are more general in nature than those that can be made to registered customers.

Review Process

Effective methodologies usually include time for reflection on the successes and weaknesses of the process just completed. Data mining is no different. Part of CRISP-DM is learning from your experience so that future data mining projects will be more effective.

Task List

First, you should summarize the activities and decisions for each phase, including data preparation steps, model building, etc. Then for each phase, consider the following questions and make suggestions for improvement:

- Did this stage contribute to the value of the final results?
- Are there ways to streamline or improve this particular stage or operation?
- What were the failures or mistakes of this phase? How can they be avoided next time?
- Were there dead ends, such as particular models that proved fruitless? Are there ways to predict such dead ends so that efforts can be directed more productively?
- Were there any surprises (both good and bad) during this phase? In hindsight, is there an obvious way to predict such occurences?
- Are there alternative decisions or strategies that might have been used in a given phase? Note such alternatives for future data mining projects.

E-Retail Example--Review Report

A Web-Mining Scenario Using CRISP-DM

As a result of reviewing the process of the inital data mining project, the e-retailer has developed a greater appreciation of the interrelations between steps in the process. Initially reluctant to "backtrack" in the CRISP-DM process, the e-retailer now sees that the cyclic nature of the process increases its power. The process review has also led the e-retailer to understand that:

- A return to the exploration process is always warranted when something unusual appears in another phase of the CRISP-DM process.
- Data preparation, especially of Web logs, requires patience, since it can take a very long time.
- It is vital to stay focused on the business problem at hand, because once the data are ready for analysis, it's all too easy to start constructing models without regard to the bigger picture.
- Once the modeling phase is over, business understanding is even more important in deciding how to implement results and determine what further studies are warranted.

Determining the Next Steps

By now, you've produced results, evaluated your data mining experiences, and may be wondering, **Where to next?** This phase helps you to answer that question in light of your business goals for data mining. Essentially, you have two choices at this point:

- **Continue to the deployment phase.** The next phase will help you to incorporate the model results into your business process and produce a final report. Even if your data mining efforts were unsuccessful, you should use the deployment phase of CRISP-DM to create a final report for distribution to the project sponsor.
- **Go back and refine or replace your models.** If you find that your results are almost, but not quite, optimal, consider another round of modeling. You can take what you've learned in this phase and use it to refine the models and produce better results.

Your decision at this point involves the accuracy and relevancy of the modeling results. If the results address your data mining and business goals, then you are ready for the deployment phase. Whatever decision you make, be sure to document the evaluation process thoroughly.

E-Retail Example--Next Steps

A Web-Mining Scenario Using CRISP-DM

The e-retailer is fairly confident of both the accuracy and relevancy of the project results and so is continuing to the deployment phase.

At the same time, the project team is also ready to go back and augment some of the models to include predictive techniques. At this point, they're waiting for delivery of the final reports and a green light from the decision makers.

Chapter 7. Deployment

Deployment Overview

Deployment is the process of using your new insights to make improvements within your organization. This can mean a formal integration such as the implementation of a IBM SPSS Modeler model producing churn scores that are then read into a data warehouse. Alternatively, deployment can mean that you use the insights gained from data mining to elicit change in your organization. For example, perhaps you discovered alarming patterns in your data indicating a shift in behavior for customers over the age of 30. These results may not be formally integrated into your information systems, but they will undoubtedly be useful for planning and making marketing decisions.

In general, the deployment phase of CRISP-DM includes two types of activities:

- Planning and monitoring the deployment of results
- Completing wrap-up tasks such as producing a final report and conducting a project review

Depending on your organization's requirements, you may need to complete one or both of these steps.

Planning for Deployment

Although you may be anxious to share the fruits of your data mining efforts, take time to plan for a smooth and comprehensive deployment of results.

Task List

- The first step is to summarize your results--both models and findings. This helps you determine which models can be integrated within your database systems and which findings should be presented to your colleagues.
- For each deployable model, create a step-by-step plan for deployment and integration with your systems. Note any technical details such as database requirements for model output. For example, perhaps your system requires that modeling output be deployed in a tab-delimited format.
- For each conclusive finding, create a plan to disseminate this information to strategy makers.
- Are there alternative deployment plans for both types of results that are worth mentioning?
- Consider how the deployment will be monitored. For example, how will a model deployed using IBM SPSS Modeler Solution Publisher be updated? How will you decide when the model is no longer applicable?
- Identify any deployment problems and plan for contingencies. For example, decision makers may want more information on modeling results and may require that you provide further technical details.

E-Retail Example--Deployment Planning

A Web-Mining Scenario Using CRISP-DM

A successful deployment of the e-retailer's data mining results requires that the right information reaches the right people.

Decision makers. Decision makers need to be informed of the recommendations and proposed changes to the site, and provided with short explanations of how these changes will help. Assuming that they accept the results of the study, the people who will implement the changes need to be notified.

Web developers. People who maintain the Web site will have to incorporate the new recommendations and organization of site content. Inform them of what changes *could* happen because of future studies, so they can lay the groundwork now. Getting the team prepared for on-the-fly site construction based upon real-time sequence analysis might be helpful later.

Database experts. The people who maintain the customer, purchase, and product databases should be kept apprised of how the information from the databases is being used and what attributes may be added to the databases in future projects.

Above all, the project team needs to keep in touch with each of these groups to coordinate the deployment of results and planning for future projects.

Planning Monitoring and Maintenance

In a full-fledged deployment and integration of modeling results, your data mining work may be ongoing. For example, if a model is deployed to predict sequences of e-basket purchases, this model will likely need to be evaluated periodically to ensure its effectiveness and to make continuous improvements. Similarly, a model deployed to increase customer retention among high-value customers will likely need to be tweaked once a particular level of retention is reached. The model might then be modified and re-used to retain customers at a lower but still profitable level on the value pyramid.

Task List

Take notes on the following issues and be sure to include them in the final report.

- For each model or finding, which factors or influences (such as market value or seasonal variation) need to be tracked?
- How can the validity and accuracy of each model be measured and monitored?
- How will you determine when a model has "expired"? Give specifics on accuracy thresholds or expected changes in data, etc.
- What will occur when a model expires? Can you simply rebuild the model with newer data or make slight adjustments? Or will changes be pervasive enough as to require a new data mining project?
- Can this model be used for similar business issues once it has expired? This is where good documentation becomes critical for assessing the business purpose for each data mining project.

E-Retail Example--Monitoring and Maintenance

A Web-Mining Scenario Using CRISP-DM

The immediate task for monitoring is to determine whether the new site organization and improved recommendations actually work. That is, are users able to take more direct routes to the pages that they're looking for? Have cross-sales of recommended items increased? After a few weeks of monitoring, the e-retailer will be able to determine the success of the study.

What can be handled automatically is the inclusion of new registered users. When customers register with the site, the current rulesets can be applied to their information to determine what recommendations they should be given.

Deciding when to update the rulesets for determining recommendations is a trickier task. Updating the rulesets is not an automatic process because cluster creation requires human input regarding the appropriateness of a given cluster solution.

As future projects generate more complex models, the need for and amount of monitoring will almost surely increase. When possible, the bulk of the monitoring should be automatic with regularly scheduled reports available for review. Alternatively, the creation of models that provide predictions on the fly may be a direction the company would like to take. This requires more sophistication from the team than the first data mining project.

Producing a Final Report

Writing a final report not only ties up loose ends in earlier documentation, it can also be used to communicate your results. While this may seem straightforward, it's important to present your results to the various people with a stake in the results. This can include both technical administrators who will be responsible for implementation of the modeling results as well as marketing and management sponsors who will make decisions based on your results.

Task List

First, consider the audience of your report. Are they technical developers or market-focused managers? You may need to create separate reports for each audience if their needs are disparate. In either case, your report should include most of the following points:

- A thorough description of the original business problem
- The process used to conduct data mining
- Costs of the project
- Notes on any deviations from the original project plan
- A summary of data mining results, both models and findings
- An overview of the proposed plan for deployment
- Recommendations for further data mining work, including interesting leads discovered during exploration and modeling

Preparing a Final Presentation

In addition to the project report, you may also need to present the project findings to a team of sponsors or related departments. If this is the case, you could use much of the same information in your report but presented from a broader perspective. The charts and graphs in IBM SPSS Modeler can easily be exported for this type of presentation.

E-Retail Example--Final Report

A Web-Mining Scenario Using CRISP-DM

The greatest deviation from the original project plan is also an interesting lead for further data mining work. The original plan called for finding out how to have customers spend more time and see more pages on the site per visit.

As it turns out, having a happy customer is not simply a matter of keeping them online. Frequency distributions of time spent per session, split on whether the session resulted in a purchase, found that the session times for most sessions resulting in purchases fall between the session times for two clusters of nonpurchase sessions.

Now that this is known, the issue is to find out whether these customers who spend a long time on the site without purchasing are just browsing or simply can't find what they're looking for. The next step is to find out how to deliver what they're looking for in order to encourage purchases.

Conducting a Final Project Review

This is the final step of the CRISP-DM methodology, and it offers you a chance to formulate your final impressions and collate the lessons learned during the data mining process.

Task List

You should conduct a brief interview with those significantly involved in the data mining process. Questions to consider during these interviews include the following:

- What are your overall impressions of the project?
- What did you learn during the process--both about data mining in general and the data available?
- Which parts of the project went well? Where did difficulties arise? Was there information that might have helped ease the confusion?

After the data mining results have been deployed, you might also interview those affected by the results such as customers or business partners. Your goal here should be to determine whether the project was worthwhile and offered the benefits it set out to create.

The results of these interviews can be summarized along with your own impressions of the project in a final report that should focus on the lessons learned from the experience of mining your stores of data.

E-Retail Example--Final Review

A Web-Mining Scenario Using CRISP-DM

Project member interviews. The e-retailer finds that project members most closely associated with the study from start to finish are for the most part enthusiastic about the results and look forward to future projects. The database group seems cautiously optimistic; while they appreciate the usefulness of the study, they point out the added burden on database resources. A consultant was available during the study, but going forward, another employee dedicated to database maintenance will be necessary as the scope of the project expands.

Customer interviews. Customer feedback has been largely positive so far. One issue that was not well thought out was the impact of the site design change on established customers. After a few years, the registered customers developed certain expectations about how the site is organized. Feedback from registered users is not quite as positive as from nonregistered customers, and a few greatly dislike the changes. The e-retailer needs to stay aware of this issue and carefully consider whether a change will bring in enough new customers to risk losing existing ones.

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

For license inquiries regarding double-byte (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBMproducts. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at "Copyright and trademark information" at www.ibm.com/legal/copytrade.shtml.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Terms and conditions for product documentation

Permissions for the use of these publications are granted subject to the following terms and conditions.

Applicability

These terms and conditions are in addition to any terms of use for the IBM website.

Personal use

You may reproduce these publications for your personal, noncommercial use provided that all proprietary notices are preserved. You may not distribute, display or make derivative work of these publications, or any portion thereof, without the express consent of IBM.

Commercial use

You may reproduce, distribute and display these publications solely within your enterprise provided that all proprietary notices are preserved. You may not make derivative works of these publications, or reproduce, distribute or display these publications or any portion thereof outside your enterprise, without the express consent of IBM.

Rights

Except as expressly granted in this permission, no other permissions, licenses or rights are granted, either express or implied, to the publications or any information, data, software or other intellectual property contained therein.

IBM reserves the right to withdraw the permissions granted herein whenever, in its discretion, the use of the publications is detrimental to its interest or, as determined by IBM, the above instructions are not being properly followed.

You may not download, export or re-export this information except in full compliance with all applicable laws and regulations, including all United States export laws and regulations.

IBM MAKES NO GUARANTEE ABOUT THE CONTENT OF THESE PUBLICATIONS. THE PUBLICATIONS ARE PROVIDED "AS-IS" AND WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, NON-INFRINGEMENT, AND FITNESS FOR A PARTICULAR PURPOSE.

Index

A

aggregating 22
algorithms 25
Append node 22
appending data 21
approved models 31
assessing
  current business situation 7
  models 28
  tools available 10, 11
attributes
  deriving 21
  selecting 19

B

background
  gathering information 5
blanks
  collecting data 13
  verifying data quality 16
books
  on CRISP-DM 3
Boolean values 14
business success
  evaluating results 31
business understanding 5

C

cleaning data 20
conclusions 31
constraints
  making a list 8
constructing data 21
cost/benefit analysis 9
CRISP-DM
  additional resources 3
  help 2
  in IBM SPSS Modeler 2
  overview 1
criteria
  for business success 6
  for data mining success 10

D

data
  attributes 13
  cleaning 20
  collecting 13
  collection report 14
  constructing new data 21
  describing 14
  examining quality 16
data *(continued)*
  excluding 19
  exploring 15
  flat files 16
  format 14
  formatting for modeling 22
  integrating 21
  merging 22
  missing values 16
  partitioning 26
  quality report 16
  selecting 19
  selecting attributes 19
  size statistics 14
  sorting 22
  types 13
  verifying quality 16
  visualization 15
data mining
  determining next steps 32
  review of process 32
  using CRISP-DM 1
data preparation 19
data understanding 13
defining
  project terminology 8
delimiters 16
deployment 35
Derive node 21

E

errors 20
evaluation
  determining next steps 32
  phase of CRISP-DM 31
examples
  business understanding phase 5, 7, 9, 10
  data preparation phase 19–22
  data understanding phase 13–16
  e-retail 22
  evaluation phase 31–33
  modeling phase 25–28
exploratory statistics 15

F

findings 31
flat files 16

G

goals
  adjusting 15
  setting business goals 5
  setting data mining goals 9

goodness 26

H

help
    CRISP-DM 2
HTML
    generating reports 2
hypothesis
    forming 15

M

maintenance 36
Merge node 22
merging data 13, 21, 22
metadata 16, 20
missing values 13, 16, 20, 21
model
    evaluating results 31
modeling
    assessment of output 28
    data requirements 22
    preparing data 19
    setting options 27
    techniques 25
    testing results 26
models
    building 27
    list of approved models 31
    parameters 27
    supervised 26
    types 27
    unsupervised 26
monitoring deployment 36

N

noise 16, 20
normalizing 21
numeric values 14

O

objectives
    setting business objectives 5
    tasks involved 6
options
    modeling 27
organization charts 5

P

parameters
    modeling 27, 29
partitioning 26
phase
    business understanding 5
    data preparation 19
    data understanding 13
    evaluation 31
    modeling 25
planning
planning *(continued)*
    deployment of results 35
    monitoring and maintenance 36
    writing the project plan 10
preparing data 19
presenting results 37
process
    review of data mining 32
project tool 2
projects
    conducting a final review 37
    conducting cost/benefit analysis 9
    inventory of resources 7
    listing requirements, assumptions, and constraints 8
    listing risks and contingencies 8
    writing the final report 36

Q

quality
    data examination 16
    data quality report 16

R

records
    generating 21
    selecting 19
reports
    data cleaning 20
    data collection 14
    data description 14
    data exploration 15
    data quality 16
    final project 36
    generating from the project tool 2
    project plan 10
requirements
    making a list 8
resources
    additional resources on CRISP-DM 3
    inventory of project resources 7
results
    evaluating 31
    presenting 37
reviewing
    data mining process 32
risks 8

S

selecting data 19
Set-to-Flag node 21
size
    data sets 14
sorting 22
statistics
    exploratory 15
success criteria
    from a business perspective 6
    from a data mining perspective 9
    in technical terms 10
supervised models 26

symbolic values 14

T

techniques
    modeling 25
terminology 8
tools
    assessment 10, 11
tooltips 2
train/test 26

U

understanding
    business needs 5
    data 13
    data mining goals 9
unsupervised models 26

V

visualization tools 15

W

Web mining
    e-retail 5, 7, 9, 19–22, 25–28, 31–33
writing
    data cleaning report 20
    data collection report 14
    data exploration report 15
    data quality report 16
    project plan 10

IBM®