SPSS

SPSS Data Mining Tips

A handy guide to help you save time and money as you plan and execute your data mining projects

Table of contents

Introduction

Are you currently involved in a data mining project? Are you considering undertaking a data mining project for the first time? Regardless of your level of data mining experience, the *SPSS Data Mining Tips* guide will help you plan and execute your project.

Use the tips presented in this guide to save money, complete your project in a timely manner, and produce positive, measurable results.

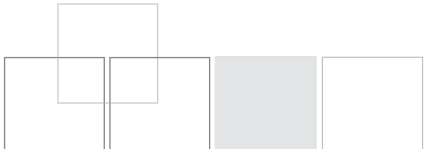
If you have questions about beginning or executing your data mining projects, please call your local SPSS office. We offer a variety of technology training and consulting programs to assist you. If you have any data mining suggestions or ideas, please send an e-mail to suggest@spss.com. And please visit us online at **www.spss.com**.

What is data mining?

Data mining solves a common paradox: The more customer data you have, the more difficult and time-consuming it is to effectively analyze and draw meaning from them. What should be a gold mine often lies unexplored due to a lack of personnel, time, or expertise. Data mining uses a clear business orientation and powerful analytic technologies to quickly and thoroughly explore mountains of data, pulling out the valuable, usable information—the business insight—that you need.

What types of data are used in data mining?
Depending on your data mining tool, your project can incorporate data from a wide range of sources. In fact, data mining projects often benefit from using several different types of data, each of which gives you additional insight into your area of study. Everything from transactional databases to survey data, textual documents, and online activity can add accuracy and depth to your results. Recent advances in analytics have led to two important new types of mining—text mining and Web mining. These two technologies open a rich vein of customer data in the form of textual comments from survey research and customer communications—information known as "unstructured data"—and log files from Web servers. These data increase both the accuracy and depth of the insights uncovered through your data mining efforts. Survey data can add valuable information about opinions and preferences—giving you the "why" behind actions and behaviors.

Data mining and predictive analytics
Data mining uncovers patterns in data using predictive techniques. Predictive analytics combines these advanced analytic techniques with decision optimization, which uses your analytical results to determine which actions will drive the best outcomes. These recommended actions, along with supporting information, are then delivered to the people and systems that can take action.



How is data mining different from OLAP and reporting?

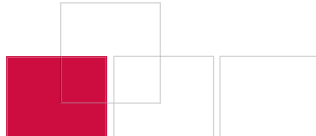
Reporting and online analytical processing (OLAP) are important tools for understanding what happened in the past. Data mining is a process for understanding what will happen in the future. Data mining uses predictive modeling, including statistics and machine-learning techniques such as neural networks, to predict what will happen. For example, queries and reports tell you the total sales for last month. OLAP goes a layer deeper to tell you sales by product for last month. Data mining, however, tells you who is likely to buy your products *next* month. And for the best business results, incorporate these insights into your marketing campaign strategy to determine, for example, how to make personalized offers that have the best likelihood of leading to sales.

How is data mining different from statistics?

Data mining doesn't replace statistics. In fact, statistics are a good complement to data mining. Traditional statistical techniques, such as regression, are used alongside data mining technologies, such as neural networks. Statistics are also used to validate data mining results.

Why use data mining?

When you have a reliable guide to the future of your business, you have the power to make the right decisions today. Data mining empowers you to manage and change the future of your



business by understanding the past and present, and delivering accurate predictions. For example, data mining tells you which prospects are likely to become profitable customers and which are most likely to respond to your offer. With this view of the future, you increase your return on investment (ROI) by making your offer to only those prospects likely to respond and become valuable customers. Your decisions are based on sound business insight, not on instinct or gut reactions. And those decisions deliver consistent results that keep you ahead of the competition.

What business problems does data mining solve?
You can use data mining to solve almost any business problem that involves data, including:

- Increasing revenues from customers
- Understanding customer segments and preferences
- Identifying profitable customers and acquiring new ones
- Improving cross-selling and up-selling
- Retaining customers and increasing loyalty
- Increasing ROI and reducing marketing campaign costs
- Detecting fraud, waste, and abuse
- Determining credit risks
- Increasing Web site profitability
- Increasing retail store traffic and optimizing layouts for increased sales
- Monitoring business performance

How does the data mining process work?
SPSS data mining products and services ensure timely, reliable results by supporting the CRoss-Industry Standard Process for Data Mining (CRISP-DM).* Created by industry experts, CRISP-DM provides step-by-step guidelines, tasks, and objectives for every stage of the data mining process. CRISP-DM is the industry-standard process for data mining projects.

There are six phases in CRISP-DM:

- Business understanding: Achieve a clear understanding of your business challenges
- Data understanding: Determine what data are available to mine for answers
- Data preparation: Prepare the data in the appropriate format to answer your business questions
- Modeling: Design data models to meet your requirements
- Evaluation: Test your results against the goals of your project
- Deployment: Make the results of the project available to decision makers

To learn more about CRISP-DM, visit **www.crisp-dm.org**.

** CRISP-DM is the exclusive property of the partners of the CRISP-DM consortium: NCR Systems Engineering Copenhagen (USA and Denmark), DaimlerChrysler AG (Germany), SPSS Inc. (USA), and OHRA Verzekeringen en Bank Groep B.V (The Netherlands). © 1999, 2000, 2001, 2002*

Data mining tips

Setting up for success

Follow CRISP-DM

Using CRISP-DM to guide your data mining projects helps to ensure a successful business outcome. It is critical to follow a proven methodology—complex data mining technologies and large volumes of available data can overwhelm a project that is not firmly grounded by the business problem you want to solve.

Begin with the end in mind

To be able to show ROI at the end of the project, know before you start how you will evaluate the results (i.e., Which business measures should you use? How are they calculated or derived?).

For example, do you want to find 70 percent of churners in 20 percent of your subscribers? Would you know how to translate this information into expected revenue improvement, based on sound assumptions about the cost and response of your retention programs? Or, would you know how much you would save if you identified ten additional cases of fraud?

Set expectations

Make sure project stakeholders know that data mining is not a silver bullet that magically solves business problems. Data mining is a business process aided by computer support. As with any business problem, stakeholders need to find a solvable problem and work on the solution.

If you plan to segment customers for your marketing department, let department members know the type of information they are likely to receive as a result of your project (i.e., "We're using product information and demographic data, so we expect to provide segments based on age, income, etc., that will show the product mix favored by these customers").

Limit the scope of your initial project

Start with realistic objectives and schedules. When you achieve success, move on to more complex projects.

For example, rather than attempting to immediately improve customer acquisition, cross-selling, up-selling, and retention in every region, focus on a smaller, more realistic goal.

Identify a steering committee

A data mining project is a group effort. Data mining requires business users who understand the issues and the data, as well as people who understand analysis. Those who own the data will need to provide access, as well.

For example, you may need a data mining analyst, a database analyst, and a marketing manager. These titles may fall into different functional areas with goals that do not align well with the goals of the project. It's important to find ways for these roles to work together.

Avoid the data dump

Always set up the business problem, define the project goals, and get support. If you simply begin analyzing a pile of data with no project structure, you will get lost in the data and waste time.

Don't let the volume of data drive your project. Focus on the business goal. You may not use all of your data—only some may be relevant for the project. You may even discover that your data are not sufficient to resolve your business problem. A large volume of data is no guarantee that you have the *right* data. For example, recent information usually offers more accurate predictions for customer behavior than volumes of historical data.

Following the phases of CRISP-DM

This section outlines tips excerpted from the data mining guide, "*CRISP-DM™ 1.0*." This detailed tool expands on the information presented here and includes a user guide. It can be downloaded at **www.crisp-dm.org**.

Business understanding

Know "who, what, when, where, why, and how" from a business perspective

Develop a thorough understanding of the project parameters: the current business situation, the primary business objective of the project, the criteria for success, and who will determine the success of the project.

Create a deployment strategy

Think about how you want to use the results of the data mining project. For example:

- Will the results be used by specialists who don't need translations of the results?
- Will the results be used by a wide range of employees who need differing levels of translations?
- Will the results be deployed via a particular medium (online, paper, etc.) that requires a certain format?

Develop a maintenance strategy

How will you manage data once the initial project is completed? If the project is part of an ongoing strategy, will you:

- Analyze new data periodically?
- Analyze new data in real time?

Assess the situation and inventory resources

Make sure to go over every aspect of the project in advance to ensure you have what you need for success:

- Personnel (project sponsor, business, and technical experts)
- Data sources (access to warehouse or operational data)
- Computing resources (hardware, platforms)
- Software (data mining and other relevant software)

What are the project requirements?

List all of the requirements of the project:

- Schedule for completion
- Comprehensibility and quality of results
- Security
- Legal restrictions on data access

What assumptions are being made about the project?

List and clarify all of the assumptions you have made about:

- Data quality (accuracy, availability)
- External factors (economic issues, competition, technical advances)
- Internal factors (the business problem)
- Models (Is it necessary to understand, describe, or explain the models to senior management?)

Under what constraints will the project operate?

Check and develop solutions for the following:

- General constraints (legal issues, budget, timing, resources)
- Access rights to data sources (restrictions, necessary passwords)
- Technical accessibility of data (operating systems, data management system, file or database format)
- Accessibility of relevant knowledge

Does everyone speak the same language?

Make sure that everyone involved understands the terms and concepts that will be used throughout the project.

Facilitate interdepartmental understanding by creating a glossary of the business and technical terms that are specific or relevant to the project.

Translate business objectives into data mining tasks

Determine which data mining tasks you must complete in order to achieve your business objective. Define the data mining tasks using technical terms.

For example, the business goal, "Increase catalog sales to existing customers," might translate into the data mining goal, "Predict how many widgets customers will buy, given their purchases over the previous three years, relevant demographic information, and the price of the item."

Determine data mining success criteria

Using technical terms, describe which criteria must be met in order to consider the project a success.

For example, the model must display a specific level of predictive accuracy or the propensity-to-purchase profile must have a specific degree of lift.

Produce a project plan

Create a plan that outlines the steps you will take to achieve your data mining goals and meet your business objective. Assess which tools and techniques are available to enable you to complete your project.

Data understanding

Make sure the data are available

Gather all of the data you will need for your project. If your data will come from more than one source, make sure your data mining tool can integrate the data.

Data collected from online activity can improve the quality and accuracy of your models. Use a Web mining tool to add a deeper level of insight to your data mining project.

Survey data can add critical attitudinal insights to your models. A combination of behavioral and attitudinal data is best for comprehensive insight.

Up to 80 percent of your data may be hidden in text documents. Use a text mining tool to efficiently search these sources for valuable information.

Try some exploratory data mining

Help data warehouse builders set priorities by analyzing small amounts of data from multiple sources and communicating any discoveries.

Do your data cover relevant attributes?

Ensure success by choosing data that best represent the behavior or situation you want to analyze.

Describe existing data

Get a clear picture of your data by creating a report that describes data formats, the number of records and fields, field identities, and other relevant features.

Check data quality

To prevent future problems, assess the quality of your data and make a plan for addressing any problems that are detected.

- Do the attribute names and the values they contain fit together?
- Are any attributes missing? Are there any blank fields?
- Check for multiple spellings of values to eliminate repetition
- Look for data that deviate from the norm and determine the causes

Review any attributes that give answers that conflict with common sense (i.e., pregnant males).

Exclude any data that are not relevant (i.e., if you're checking on home loan behavior, eliminate customers who have never owned a home, etc.).

Generate a data quality report

Check for duplicate data, potential data errors (i.e., customers shown to have churned before they even became customers), and mandatory database fields that may contain invalid information.

Data preparation

Select your data

Decide what data to use for analysis and list the reasons for your decisions. This involves:

- Performing significance and correlation tests to determine which fields to include
- Selecting data subsets
- Using sampling techniques to review small chunks of data for appropriateness

Decide whether certain attributes are more important than others and weight them accordingly.

For more accurate models, be sure to include non-traditional types of data, such as survey data, key concepts from customer communications, and data about online activity. Combining multiple types of data gives you a more complete picture of your customers and your organization.

Address data quality problems

To ensure reliable results, take the time now to fix any data quality problems. Activities may include:

- Determining how to deal with data noise
- Addressing special values and their meaning. For example, a special value can be a default value used when a survey question is not answered or when data is shortened for space considerations (“2004” becomes “04”).

Some fields may be irrelevant to your goals and don’t need to be cleaned. Track actions taken or not taken for those fields, as you may decide to use them later in the process.

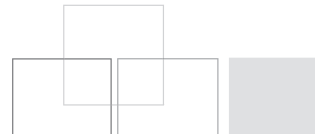
Choose a flexible data construction tool

Make sure the data mining tool you choose is capable of constructing the data according to project needs. Your tool should also allow you to add new fields as needed. Remember that data mining is a discovery-driven process—it’s impossible to know in advance where the data will take you.

Determine whether to create derived attributes

You may wish to create derived attributes for the following reasons:

- Due to your experience with the situation at hand, you know that a particular attribute is important to the data even though it doesn’t currently exist
- The modeling algorithm only handles certain data types, therefore, important information won’t be included unless it is recreated
- Modeling results reveal that relevant facts are not represented



Before you add derived attributes, determine whether and how they will help the modeling process.

Consolidate information by merging data

When you join new tables to consolidate information, you may also want to generate new fields and aggregate values.

Make sure that your data mining tool can combine different types of data—such as survey, text, and Web data—from multiple sources without costly, time-consuming customization.

Does your data mining tool require a specific order?

At this stage, you may need to sort your dataset if your data mining tool requires that your records be in a particular order.

Should the data be balanced?

Determine whether your modeling technique requires balanced data.

For example, direct mail campaigns often return response information skewed toward "no response." Some techniques make no response predictions and, as a result, have a high degree of accuracy. To predict positive responses accurately, however, some techniques may require you to have roughly equal numbers of positive and negative responses.

Modeling

Selecting modeling techniques

To match your data to the right technique, check which assumptions each technique makes about data format and quality. In some cases, only one technique may be appropriate for your situation. Be sure to consider:

- Which techniques are appropriate for your problem
- Whether there are any political requirements (management expectations, understandability)
- Whether there are any constraints (unusual data characteristics, staff expertise, timing issues)

To ensure that you have the right technique for each model and situation, choose a data mining tool that offers a wide range of techniques and modeling options.

Test before you build

Before you create your model, test the quality and validity of the techniques you plan to use. Create a test design that incorporates a training test, a test set, and a validation set. Then build the model on the training set and assess its effectiveness with the test dataset.

Build your model

To create a model, run your modeling tool on the dataset you have prepared. Describe the result and assess its expected accuracy, effectiveness, and potential shortcomings.

Create a detailed model report that lists the rules produced, the parameter settings used, the model's behavior and interpretation, and any conclusions about patterns revealed in the data.

Use only attributes that will be available to the model in the right state at the time of deployment
For example, if you want to create a model that predicts customer attrition after one year, use customer data from the one-year point to accurately reflect customer behavior at that point in time.

Using induction to produce a rule
Rules are essentially parameters within which the data must fall in order to be considered. They are usually in an "if/then" format. Induction enables you to automatically choose which rules are most effective for obtaining specific results. For example, use induction to create a set of rules for qualifying loan prospects:

- If employed for more than two years, then credit risk is good
- If older than 30, then credit risk is good
- If declared bankruptcy at any time, then credit risk is bad

Test after you build

Make sure your model delivers results that will help you achieve your data mining goal.

Use lift and gains tables to test a model's predictive ability.

Try several models to get the right fit

To improve model performance, try adding or removing fields or experimenting with the available options. Also, since each technique works slightly differently, try a variety (such as clustering and association) to find all of the relevant patterns.

Statistical models are good for:
Initial analysis—Statistical analysis is useful in the early stages of a data mining project to gain an overview of the structure of the data. Developing a concise description of the characteristics of the data can help project members develop hypotheses and plan further analysis.

Propensity models are good for:
Predicting customer behavior—Discover who is most likely to purchase, most likely to churn, most likely to default on loans, and much more. Use this information to determine which customers and prospects offer the best long-term profitability.

Clustering is good for:
Finding natural groupings of cases that have the same characteristics—Detect fraud by using clustering to group similar cases of unusual credit card transactions.

Association rules are good for:
Basket analysis—Discover which items are most likely to be purchased together. Use this information to improve cross-selling through catalog and store layout, recommendation engines, phone and direct mail offers, and more.

Evaluation

Evaluate your data mining results

Determine whether and how well the results delivered by a given model will help you achieve your business goals. Is there any business reason why the model is deficient?

If the time and resources are available, try testing the model or models on test applications within the real application.

Review the data mining process for outstanding tasks

When you have confirmed the quality and effectiveness of your results, review your work to date to determine whether you have missed any important steps or information.

- Was each stage of the data mining process necessary in retrospect?
- Was each stage executed as well as possible?

Determine next steps

Now is the time to determine whether the project is successful enough to move ahead to deployment. If not, take any further steps necessary to achieve satisfactory results. Keep in mind:

- The deployment potential of each result
- How the process could be improved
- Whether the resources exist for additional steps or repetitions of previous steps

Deployment

Create a deployment plan

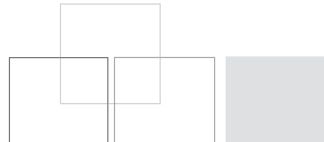
Take the project results and decide how best to use them to address your business issue:

- Summarize deployable models or software results
- Develop and evaluate alternative deployment plans
- Confirm how the results will be distributed to recipients
- Determine how to monitor the use of the results and measure the benefits
- Identify possible problems and pitfalls of deployment

Monitor and maintain your plan

Ensure the best use of your data mining results by creating a results maintenance plan that addresses:

- What could change in the future that would affect the use of the results
- How to monitor accurate use of the results
- When, if necessary, to discontinue deployment or use of the results



Create a final report

Depending on your deployment plan, the report may be either a project summary or a final presentation of the data mining results. To create your final report, first:

- Identify which reports are needed (slides, management summary, etc.)
- Analyze how well the data mining goals were met
- Identify report recipients
- Outline the structure and content of the report
- Select which discoveries to include

Execute your deployment plan

Put your data mining results to optimal use by distributing them according to the deployment plan. Even the most brilliant discovery will not generate ROI if it isn't used to improve your business.

Review the project

This is your opportunity to assess what went right, what didn't, major accomplishments, and any necessary improvements. For a complete review, try the following:

- Interview all significant project members about their experiences
- Interview any end users of your data mining results about their experiences
- Document and analyze the specific data mining steps that you took
- Create any recommendations for future projects

Selecting a data mining tool

The tips in this section are excerpted from the CRISP-DM document, "Performing a data mining tool evaluation."

Look for a tool with a proven record of solving the business problems your project addresses

Choose a tool that you know to be useful in solving problems within your industry and that has a successful track record with the types of applications you're planning.

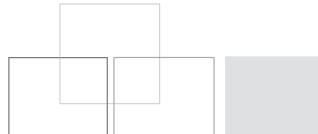
Select a tool that bridges business understanding and the technical aspect of data mining

Make sure that the steps used by the tool match the business needs of data mining:

- Does the tool present data mining concepts clearly?
- Does the tool integrate with project management software or other tools you may be using? Will you have to create applications to bridge the gap if it doesn't?

Make sure your tool works with your existing data sources and format

You will save time and money, and maximize your chances for reliable results, by choosing a tool that can pull and combine data from multiple sources and formats. This is particularly important if discoveries later in the data mining process lead you to add data from a new source.



A data mining tool that enables you to combine behavioral and attitudinal data, in the form of both structured and unstructured data, will deliver more accurate results and provide greater flexibility in terms of the types of data mining projects you're able to undertake.

Look for interactive exploration and visualization capabilities

Make it easy to explore and understand the data by choosing a tool that provides interactive visualization techniques. These techniques allow you to quickly gain insights by making changes within graphs and creating new graphs based on different dimensions of the data.

Choose a tool with efficient, comprehensible data preparation steps

Save time and resources by choosing a data mining tool that prepares data efficiently (from initial stages to model building) and that presents data preparation steps in an easy-to-understand way. This enables project members with varying levels of expertise to obtain effective results.

Make sure that your tool can automatically extract data

Avoid writing time-consuming manual queries by choosing a tool that extracts data automatically for the various data preparation steps.

Can the tool build effective models in a reasonable time?

Look for a tool that enables analysts to quickly find the most effective models. The tool should support efficient building and testing of multiple models.

Choose a tool with a wide range of techniques

To ensure the best results, make sure your tool offers a wide range of techniques or algorithms for visualization, classification, clustering, association, and regression. For example, you might discover that one technique works better than another for specific data. Flexibility will enable you to try a number of techniques to get accurate, effective results. The tool should also be able to combine techniques in situations where that would produce the best results.

Can the tool use the data and equipment you already have?

Choose a data mining tool that can use your data where it exists today, regardless of whether it is in databases or files, and that is compatible with your existing analysis and visualization tools. You don't want to waste time and resources building another database because you are unable to analyze the data you already have.

Choose a tool that delivers consistent, high-quality results

Get accurate results with a variety of data with an adaptable tool that performs well in a variety of situations, rather than one designed for a specific type of data or situation. Your tool should be able to manage any data that you may need to effectively address your business problem.

What are the tool's deployment capabilities?

It is critical to choose a tool capable of integrating your results into operational applications now and in the future. Also consider:

- Whether integration will be cost effective or whether it will require additional time and money
- How easily the tool can update data mining results and what additional investments, if any, are required

Assess the potential costs of ownership associated with the tool

Analyze the potential ROI for each tool:

- What will be the cost of ownership over the product's lifetime, including any additional software or services required by the tool? When can you expect a positive ROI?
- How long will it take to implement your data mining tool? Is it designed for technical experts or can it accommodate users of varying expertise? What training cost are involved now and in the future?
- Is the tool customizable for your particular users and business needs? Can you save common processes and automate tasks?

About SPSS Inc.

SPSS Inc. (NASDAQ: SPSS) is the world's leading provider of predictive analytics software and solutions. The company's predictive analytics technology improves business processes by giving organizations consistent control over decisions made every day. By incorporating predictive analytics into their daily operations, organizations become Predictive Enterprises—able to direct and automate decisions to meet business goals and achieve measurable competitive advantage.

More than 250,000 public sector, academic, and commercial customers, including more than 95 percent of the FORTUNE® 1000, rely on SPSS technology to help increase revenue, reduce costs, and detect and prevent fraud. Founded in 1968, SPSS is headquartered in Chicago, Illinois.

What makes SPSS unique?

For more than 35 years, SPSS has been the clear leader in analytics technology. Here are some of the reasons that customers have selected SPSS software to drive their decision making:

- **A complete, 360° view**—SPSS software enables you to develop in-depth understanding by using all of your information together, both traditional structured data and unstructured data, for a 360° view of your organization

- **Easy integration with operational systems**—SPSS predictive analytics technologies and products are designed to work well both independently and with other SPSS technologies or external systems
- **Open, standards-based architecture**—SPSS software follows industry standards such as OLE DB for data access, XMLA for data/format sharing, PMML for predictive model sharing, SSL for Internet security management, and LDAP/Active Directory Services for authentication and authorization, just to name a few
- **A faster return on your software investment**—According to a recent study by Nucleus Research, an independent analyst firm, 94 percent of SPSS customers achieve a positive return on investment within an average payback period of just 10.7 months
- **A lower total cost of ownership**—SPSS technology is designed to work with your existing technology infrastructure and staff resources, and SPSS keeps both your short- and long-term costs of ownership low by providing open technology and flexible licensing options

Learn more about what makes SPSS unique at **www.spss.com/corpinfo/spss_edge.htm.**

SPSS Inc. products

With SPSS Inc. products, you build a flexible analytics system that enables you to both meet your needs today and achieve tomorrow's goals.

Data mining

AnswerTree®—AnswerTree reveals segments and predicts how groups will respond, using scalable decision trees. You can easily see the groups that matter, because AnswerTree displays models visually.

Clementine®—Clementine's interactive data mining process incorporates your valuable expertise at every step to create powerful predictive models that address your specific business issues. The Clementine family of data mining products includes:

- **Clementine Application Templates (CATs)**—CATs provide pre-built streams of common data mining applications that you can apply directly to your data or use as the foundation for customized streams
- **Text Mining for Clementine**—Extract key concepts, sentiments, and relationships from unstructured data, and convert them to structured format for predictive modeling with Clementine
- **Web Mining for Clementine**—Easily transform raw Web data into analysis-ready business events within Clementine's intuitive visual workflow interface

- **SPSS Predictive Enterprise Services™**—Centralize and organize models and modeling processes
- **Cleo™**—Deploy Web applications that enable decision makers to perform real-time scoring with Clementine predictive models

Predictive analytics

Predictive Analytic Applications—Deliver on-demand and real-time recommendations to systems and decisions makers through a combination of advanced analytics and decision optimization. SPSS offers the following solutions:

- **PredictiveCallCenter™**—Turn inbound customer calls into sales opportunities
- **PredictiveClaims™**—Increase customer satisfaction and reduce insurance claim fraud
- **PredictiveMarketing™**—Generate more profit from outbound marketing campaigns
- **Predictive Text Analytics™**—Improve the depth and accuracy of business insights by incorporating unstructured textual data into your analyses
- **Predictive Web Analytics™**—Improve online marketing campaigns and site effectiveness
- **PredictiveWebSite™**—Turn Web visits into sales opportunities

Statistical analysis

SPSS Base—Once you understand your data, you need to prepare them for analysis. SPSS is a modular, tightly integrated, full-featured product line for the analytical process—from planning to data collection, data access and management, analysis, reporting, and deployment—and a critical component of the data mining process. Add the complementary products below to increase your analysis capabilities:

- **SPSS Advanced Models™**—Improve the accuracy of your analyses and provide more dependable conclusions with procedures designed to fit the inherent characteristics of your data
- **SPSS Regression Models™**—Apply more sophisticated models for greater accuracy in market research, medical research, financial risk assessment, and many other areas
- **SPSS Text Analysis for Surveys™**—Categorize text responses to open-ended survey questions so you can integrate them with your quantitative survey data. SPSS Text Analysis for Surveys extracts key concepts from text for further analysis in SPSS or Microsoft® Excel®.
- **SmartViewer® Web Server**—Get the analysis you need to make more informed decisions with convenient, Web-based delivery of SPSS reports

The SPSS family of products includes a full range of add-on modules and stand-alone products. For a complete list, go to **www.spss.com/spss/family.cfm**.

Survey and market research

Dimensions™—Conduct both large-scale, multi-mode research projects and smaller, one-of-a kind surveys with this open, scalable, and customizable survey research platform. The Dimensions platform includes products for every step of the survey process, from creating survey scripts to collecting and analyzing data and reporting the results.

Training and Services

SPSS Training—SPSS offers a full suite of data mining courses, as well as product-specific training. Most courses are available at an SPSS facility or at your company site.

SPSS Worldwide Services—Let our experienced consultants help you determine which problems to address and how best to solve them.

SPSS products are available for Windows, UNIX®, and other platforms.

Glossary

Association: The process of discovering which events occur together or are related. For example, use association techniques to determine which products are often purchased together. Contrast with **sequence detection**, which can be used to discover the order in which the products were purchased.

Attitudinal data: Data that relate to or are expressive of personal attitudes or opinions. Attitudinal data is often gathered through survey research, such as responses to open-ended survey questions, and analysis of textual communications, such as customer e-mails.

Attribute: A property or characteristic of an entity; also known as a **variable** or **field**.

Balanced data: In cases where you have two or more categories of data to analyze, each category should have an equal amount of data to simplify the modeling process.

Behavioral data: Data that relate to or reflect behavior or actions. Behavioral data is the data type used most extensively in data mining.

Churn: Churn describes the process of customer attrition and is a primary source of aggravation for many industries, particularly telecommunications and financial services.

Classification: The process of identifying the group to which an object belongs by examining characteristics of the object. In classification, the groups are defined by an external criterion (contrast with **clustering**).

Clustering: The process of grouping records based on similarity. For example, an insurance company might use clustering to group customers according to income, age, type of policy purchased, or prior claims history. Clustering divides a dataset so that records with similar content are in the same group, and groups are as different as possible from each other (contrast with **classification**).

CRoss-Industry Standard Process for Data Mining (CRISP-DM): CRISP-DM provides a structure for data mining projects, as well as guidance on potential problems and their solutions. CRISP-DM is comprised of six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

Cross-selling: The practice of offering and selling additional products or services to existing customers.

Data mining: The process of analyzing data to discover hidden patterns and relationships that can help you manage and improve your business.

Data warehouse: The database in which data is collected and stored for analysis.

Decision trees: Graphical, tree-like displays that clearly show segments, patterns, and hierarchies in data.

Deployment: The distribution and use of results obtained from data mining.

Derived attributes: Derived attributes are new attributes that are constructed from one or more existing attributes in the same record.

Field: A space for an individual piece of data or information. Also known as a **variable** or **attribute**. For example, one data field may contain a customer's first name. The next data field may contain the customer's last name.

Gains tables: Gains tables measure model effectiveness by showing the difference between results obtained by the model and results obtained without using the model.

Lift charts: Enable users to measure model effectiveness by showing the ratio between results obtained using the model and results obtained without using the model. The farther the lift line from the baseline, the more effective the model.

Machine-learning techniques: A set of methods that enables a computer to learn a specific task—such as decision making, estimation, classification, or prediction—without manual programming.

Model: A set of representative rules, behaviors, or characteristics against which data are analyzed to find similarities. Descriptive models are used to analyze past events. Predictive models are used to discover what will happen in the future. With predictive models, data miners can explore alternative scenarios to determine which actions will produce the desired future outcome.

Neural network: A model for predicting or classifying cases using a complex mathematical scheme that simulates an abstract version of brain cells. A neural network is trained by presenting it with a large number of observed cases, one at a time, and allowing it to update itself repeatedly until it learns the task.

Noise: The difference between a model and its predictions. Sometimes data is referred to as noisy when it contains errors, such as many missing or incorrect values or when there are extraneous columns.**

Online analytical processing (OLAP): OLAP enables users to analyze many layers of current and historical data. Though OLAP can tell you what is happening and what happened previously with your data, it can't tell you what will happen in the future.

Pivot tables: Interactive tables that enable users to get different views of information by easily repositioning rows, columns, and layers of data.

Predictive analytics: Predictive analytics is a combination of advanced analytic techniques and decision optimization. Predictive analytics uses historical information to make predictions about future behavior, and then delivers recommended actions to the people and systems that can use them.

Predictive modeling: The process of creating models to predict future activity, behavior, or characteristics. For example, a predictive model may show which customers are most likely to churn in the future, based on the characteristics and actions of previous churners.

Query: A request sent to a database for information based on specified characteristics or properties.

Record: A record is a set of related data stored together. Also known as a row (in spreadsheets) or a case (in statistics).

Regression: The process of discovering and predicting relationships between two or more variables.

Reporting: The process of deploying or distributing the results of data analysis in a format that is comprehensible to the recipient.

Return on investment (ROI): ROI is the value that is returned or obtained from various investments in technology, infrastructure, etc.

Rule induction technique: The process of automatically deriving decision-making rules for predicting or classifying future cases from example cases.

Sequence detection: The process of discovering the order of events in data. For example, use sequence detection to discover the order in which customers purchase certain products. Contrast with **association**, which reveals which products are purchased together.

Statistics: The mathematics of the collection, organization, and interpretation of numerical data.***

Structured data: Data in traditional numerical format, such as transactional data.

Test set: A dataset independent of the **training set**, used to fine-tune the estimates of the model parameters.**

Text mining: The process of analyzing textual information—such as documents, e-mails, and call center transcripts—to extract relevant concepts.

Training set: A dataset used to estimate or train a model.**

Unstructured data: Data in a text format or other non-numerical format. Combining unstructured and structured data in your data mining projects can help you produce more accurate, valuable results.

Up-selling: The practice of offering and selling more profitable products or services to existing customers than those they currently own or use.

Variable: Any measured characteristic or **attribute** that differs for different subjects.

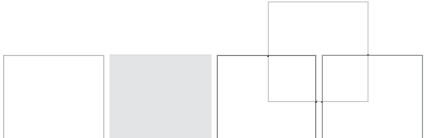
Web cubes: An online, multi-layered display used to examine the relationships between symbolic data fields.

Web mining: The process of analyzing data from online activities—including pay-per-click advertising and other marketing campaigns—to discover relevant patterns and important behavioral insights.

*** From "Two Crows: data mining glossary" (www.twocrows.com/glossary.htm)*

**** From Webster's II New College Dictionary, © 1999, 1995 by Houghton Mifflin Company*

SPSS has a worldwide network of distributors. To locate the SPSS office nearest you, go to **www.spss.com**.



ISBN 1-56827-282-0 Printed in the U.S.A.

$8.95