

MineSet™ 3.0 Enterprise Edition Tutorial for Windows®

Document Number 007-4006-001

CONTRIBUTORS

Written by Helen Vanderberg, Pam Sogard, and Sandra Motroni

Illustrated by Dany Galgani

Production by Linda Rae Sande

Engineering contributions by Barry Becker, Amit Bleiweiss, Jeff Brainerd,
Cliff Brunk, Eben Haber, Ara Jerahian, Eser Kandogan, Andy Kar, Ed Karrels,
Alex Kozlov, Brian Lovrin, Alan Norton, Peter Rathman, Gerald Rousselle,
Mario Schkolnick, Dan Sommerfield, Peter Welch, and Brett Zane-Ulman,

© 1999 Silicon Graphics, Inc.— All Rights Reserved

The contents of this document may not be copied or duplicated in any form, in whole or in part, without the prior written permission of Silicon Graphics, Inc.

LIMITED AND RESTRICTED RIGHTS LEGEND

Use, duplication, or disclosure by the Government is subject to restrictions as set forth in the Rights in Data clause at FAR 52.227-14 and/or in similar or successor clauses in the FAR, or in the DOD, DOE or NASA FAR Supplements. Unpublished rights reserved under the Copyright Laws of the United States. Contractor/manufacturer is Silicon Graphics, Inc., 1600 Amphitheatre Pkwy., Mountain View, CA 94043-1351.

Silicon Graphics is a registered trademark, and SGI, MineSet and the Silicon Graphics logo are trademarks, of Silicon Graphics, Inc. Oracle is a registered trademark of Oracle Corporation. Windows and Windows NT are registered trademarks, and MicroSoft SQL Server is a trademark of MicroSoft Corporation.

The Tree Visualizer is patented under United States Patents No. 5,528,735; 5,555,354 5,671,381; and 5,861,885. The Splat Visualizer is patented under United States Patent No. 5,861,891. Patent pending for the 2D slider in the Map Visualizer, Scatter Visualizer and Splat Visualizer. Patent pending for the Evidence Visualizer, Decision Table and Splatviz animation.

Contents

	About This Tutorial	v
	Audience for This Tutorial	v
	Prerequisites for This Tutorial	v
	Structure of This Tutorial	vi
	Typographical Conventions	vi
1.	Data Mining Fundamentals	1
	About Data Mining	1
	Terms Used in This Tutorial	2
	Data Mining Methods	2
	Analytical Data Mining Algorithms	3
	Supervised Modeling	4
	Unsupervised Modeling	5
	Visual Data Mining	6
	MineSet Tools for Data Mining Tasks	7
2.	Data Mining Process	9
	Identifying the Data	9
	Preparing the Data	10
	Transforming the Data	11
	Building a Model	12
	Evaluating a Model	12
	Deploying a Model	12
	Applying the Process to a Specific Database	12

- 3. **Churn Tutorial** 13
 - About the Raw Data 13
 - Starting MineSet 14
 - Viewing the Records 15
 - Building an Evidence Classifier 18
 - Viewing Probabilities With Splat Visualizer 21
 - Visualizing Geographic Distributions 25
 - Creating a Decision Tree Classifier 30

- 4. **Further Explorations** 33
 - Exploring Data Clusters 33
 - Relating the Columns and Axes in the Model 36
 - Finding Important Columns in the Clustered Model 38
 - Mapping to Scatter Visualizer 39
 - Invoking a Decision Table 42
 - Targeting Customers Using a Model 44
 - Creating a Training Sample 45
 - Applying a Model 46
 - Reducing Misclassification Costs 51
 - Displaying a Confusion Matrix 51
 - Defining a Loss Matrix 54
 - Viewing a Return on Investment Curve 55
 - Further Exploration of MineSet 57

- A. **Navigating in the MineSet Visualizers** 59
 - Navigating in the Tree Visualizers 59
 - Navigating in Non-Tree Visualizers 61

About This Tutorial

MineSet 3.0 Enterprise Edition Tutorial for Windows introduces MineSet in a Windows environment. MineSet is an integrated suite of data mining and visualization tools, and provides a swift survey of the concepts and processes of data mining. This tutorial describes a few basic tasks to help you use MineSet immediately. Once you are familiar with the interface, refer to the *MineSet 3.0 Enterprise Edition User's Guide for Windows* for a full description of other MineSet features. The user's guide is delivered online as part of the MineSet product. See also <http://mineset.sgi.com> for more information.

Audience for This Tutorial

This tutorial is for end users. No experience in programming is required, nor is any previous knowledge of statistics (although it is helpful). However, a basic knowledge of Windows is assumed.

Prerequisites for This Tutorial

To work with this tutorial, MineSet should be installed on your system, or you should have access to such a system. The examples depend on it. Instructions for installing MineSet are available in the *MineSet 3.0 Enterprise Edition User's Guide for Windows* and on the MineSet Web page <http://mineset.sgi.com>, where MineSet itself can be downloaded for evaluation purposes.

For this tutorial you do not need access to a database. The data needed is included in the MineSet distribution.

Structure of This Tutorial

[Chapter 1, “Data Mining Fundamentals,”](#) introduces the concept of data mining and explains how it can be used to solve problems. Common data mining tasks are aligned with the various MineSet tools; the details of each tool are covered in later chapters.

[Chapter 2, “Data Mining Process,”](#) describes the tasks involved in the process of data mining. A case study of data mining using MineSet is provided.

[Chapter 3, “Churn Tutorial,”](#) provides a detailed tutorial for the process of data mining using MineSet. It begins from the initial screen and steps screen by screen through the MineSet tools, using *churn*, a dataset provided with the MineSet distribution.

[Chapter 4, “Further Explorations,”](#) continues the exploration of MineSet with more complex variations of data mining.

[Appendix A, “Navigating in the MineSet Visualizers,”](#) explains the various ways to move around in and manipulate the visualizer windows.

Typographical Conventions

This tutorial uses several font conventions:

italics Italics are used for commands, filenames, variables, and user interface button names.

Courier Courier is used for examples of system output and for the contents of files.

Courier bold Courier bold is used for commands and other text that you type literally.

Data Mining Fundamentals

This chapter surveys data mining methods, model building and assessment, and the role of MineSet in connection with these topics:

- [“About Data Mining” on page 1](#)
- [“Terms Used in This Tutorial” on page 2](#)
- [“Data Mining Methods” on page 2](#)
- [“Analytical Data Mining Algorithms” on page 3](#)
- [“Visual Data Mining” on page 6](#)
- [“MineSet Tools for Data Mining Tasks” on page 7](#)

About Data Mining

The purpose of data mining is to discover patterns in data so that this knowledge can be applied to problem solving. Analytical data mining integrated with powerful visualizations presents new pathways to knowledge discovery. The data mining system can automatically find and show you new patterns that can lead to fresh insight. Examples of this might be determining correlations among attributes, discriminating among subsets of the data with differing characteristics, and inferring probabilities of future events from historical data.

In ordinary database queries or online analytic processing (OLAP), you must specify directly any relationships between data elements. Data mining can discover relationships that may be unknown or unseen by you.

Data to be analyzed, or mined, is often initially retrieved when a business or scientific process is performed, such as acquiring data from customer billing, pharmaceutical testing, or point-of-sale transactions. The amount of data retrieved may be so large as to preclude analysis by means other than data mining. Such data, once properly transformed, is often stored in a data warehouse. See [“Preparing the Data” on page 10](#) for further details.

Terms Used in This Tutorial

The data files that MineSet uses can be thought of as large tables. The rows are the individual records, and the columns are the *attributes* for each record. The *label* in classification tasks is a specific value of the attribute chosen for classification. For instance, in the example file used throughout this tutorial (*churn*), the task is to classify the records into customers who have quit the company (churned) and those who have not. The label attribute (column) is “churned,” and the possible label values are “yes” and “no.”

A *discrete* label is one that can have only a limited number of values, such as gender, salary ranges (for instance, less than \$40,000; \$40,000 to \$80,000; and over \$80,000), and age ranges (for instance, under 21, 21 to 35, 36 to 50, and over 50). A continuous label can have any value in a large range, for example, yearly salary, yearly sales, and miles per gallon.

Data Mining Methods

Data mining combines hypothesis testing and data-driven discovery. In hypothesis testing, the investigator tests an idea against a body of data to confirm or reject its validity. In some cases, the data itself may drive discovery. In discovery, the investigator draws conclusions from the data, allowing the data itself to suggest conclusions. Often data mining problems are resolved by using a blend of both methods. For example, conclusions may give rise to new hypotheses that can be tested and confirmed or rejected. Data mining is where statistics and machine learning converge.

The MineSet suite of tools lets you analyze, mine, and graphically display data so that you can visualize, explore, and understand your data. You can organize and examine your data in different ways. The mining tools automatically find patterns and build models that can be viewed using the visualization tools. When you apply the visualization tools directly to the data, you gain a deeper, intuitive understanding of your data, often discovering hidden patterns and important trends.

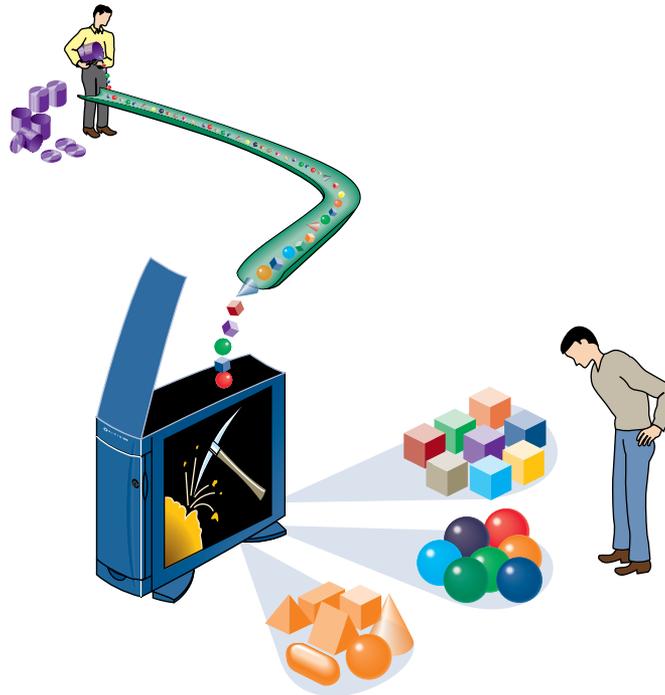


Figure 1-1 Analytical Data Mining Discovers Patterns in Data

The results of a typical analytical data mining operation in MineSet include both a model describing the data and a visualization of the model. This visualization is a 3D interface that lets you manipulate objects interactively, as well as perform animations. The visualization allows you to understand the model, and survey complex data patterns to gain invaluable insight when making decisions. MineSet is an integrated system in which the analytical algorithms can generate the visualization, and you can select visualization elements for further mining.

Analytical Data Mining Algorithms

Analytical data mining algorithms automatically build models from the data. Two families of modeling algorithms are commonly used—supervised and unsupervised. Predictive modeling tasks, where the goal is to predict the value of one column based on the values of other columns, are called *supervised* tasks. These tasks are similar to the supervision of a teacher who gives you the correct answer for the question, to teach you.

The goal in descriptive modeling is to discover patterns and segments of the data. These are *unsupervised* tasks. There is no notion of a correct answer nor an obvious agreed-upon measure of performance. Unsupervised tasks provide insight into the data as a whole by showing patterns and segments that behave similarly.

Supervised Modeling

In supervised modeling, there is a special attribute called the “label” that you intend to predict. By encoding the relation between the label and the other attributes, the model can make predictions about new, unlabeled data. In addition, by visualizing the model itself, you can gain insight into the relationship between labels and other attributes. For example, if a customer has left your company (typically called attrition or churn), you can build a model that not only predicts which customers are likely to churn, but also help you understand the reasons and patterns that lead to this behavior.

The two most common supervised modeling tasks are called classification and regression. If the label is discrete (that is, containing a fixed set of values), the task is called classification; if the label is a continuous value (that is, can take a value in a continuous range—for example, income, or stock price), the task is called regression.

Classification

Classification is the task of assigning a discrete label value to an unlabeled record (see [“Terms Used in This Tutorial” on page 2](#) for definitions of these terms). In doing so, records are divided into predefined groups. For example, a simple classification might group customer billing records into two specific classes: those who pay their bills within 60 days, and those who take longer than 60 days to pay. Other data classification tasks might divide customers by gender or income. Classifiers can also predict the probability that the label will take on a specific value for a particular record. For example, MineSet can compute the probability that a certain customer will pay his or her bill within 60 days, given the values of other attributes in the customer’s record.

A classifier is a model that predicts one attribute of a set of data when given other attributes. MineSet can induce (build) a classifier automatically from a training set (a subset of the dataset.) When a classifier is induced, MineSet also generates a visualization of the model that can help you understand how the classifier operates, thus providing valuable insight. Once a classifier is generated, it can be used to classify or predict class probabilities for unlabeled records (that is, for records that are missing the label attribute). This concept is explained further in [Chapter 3](#).

MineSet has inducers for four classification models: Decision Trees, Option Trees, Evidence (Simple Bayes), and Decision Table Classifiers. Each model can be viewed using the 3D Visualizer.

Regression

Regression is a supervised modeling task similar to classification, except that the label is not discrete. For example, predicting salary or the price of a stock is a regression, whereas predicting whether the salary is in a given range or whether a stock will go up or down is a classification task.

Assessing the Accuracy of Models

Predictive models are rarely perfect, therefore estimating their accuracy is an important part of the data mining process. The tool used to measure accuracy depends on the model type. Classifiers are usually evaluated according to their error rate. The most common such measure is misclassification, or proportion of misclassified records. When assessing the accuracy of a model, it is important to test it on data that was not used in building the model. MineSet provides a number of methods for evaluating errors. See [Chapter 4, “Further Explorations,”](#) for details.

Unsupervised Modeling

In unsupervised modeling, the aim is to discover rules and segments of the data that behave similarly (clusters). Unsupervised modeling is a descriptive task, not a predictive task. The models cannot be used directly to make predictions, hence it is not necessary to set aside part of the data as a test set with which to validate the classifier. MineSet provides two methods for unsupervised modeling—associations and clustering.

Associations

To generate associations, the task is to determine rules of implication between data attributes A and B, such that A implies B. Associations are often used to find affinity groupings that discover what items are usually purchased with others. The classic affinity grouping is market basket analysis, predicting the frequency with which certain items are purchased at the same time. For example, discovering that baby food implies a higher probability that a customer will buy low-tar cigarettes rather than regular cigarettes might help stores arrange their shelves differently.

Clustering

Clustering algorithms segment the data into groups of records, or clusters, that have similar characteristics. For instance, a health-insurance company may discover that these characteristics define a segment: 20-to-45 years old, technical worker, fewer than two children, television science-fiction fan, and a disposable income of \$10,000 to \$20,000 per year.

The segment can then be targeted more effectively with a health insurance package well-suited for these people, by using television ads in new science-fiction episodes.

Visual Data Mining

An analytical data mining algorithm can be complemented with data visualization techniques taking advantage of the human brain's amazing pattern recognition capability. The following MineSet visualizers are available:

- Map Visualizer—Data is displayed on a map, commonly a geographical map.
- Scatter Visualizer—Data points are shown in one, two, or three dimensions. Additional attributes can be mapped to color, size, and shape. Finally, two additional attributes may be mapped to sliders, allowing animation and fly-throughs, for a total of eight dimensions. The column importance operation in MineSet can help you identify the important dimensions to map for a given task.
- Splat Visualizer—Similar to Scatter Visualizer, with the distinction that data density is shown by opacity of color, which appears as a blurred translucent cloud. The result approximates the effect of rendering each data point individually.
- Tree Visualizer—Data is mapped to nodes in order to see the hierarchical breakdowns of the data.

MineSet Tools for Data Mining Tasks

If you have data mining problems requiring classification, regression, and clustering, you will find these MineSet tools useful:

- **Decision Tree Inducer and Classifier**—Induces a classifier resulting in a decision tree visualization.
- **Option Tree Inducer and Classifier**—Induces a classifier similar to a decision tree inducer and classifier. However, it builds alternative options and averages them during classification, usually leading to improved accuracy.
- **Evidence Inducer and Classifier**—Creates its own classifier and produces a visualization to display evidence based on the data provided.
- **Decision Table Inducer and Classifier**—Creates a hierarchical visualization displaying pairs of dimensions at every level. You can drill up and drill down quickly, while maintaining context.
- **Clustering Algorithm**—Groups data according to similarity of characteristics, then displays it as a series of box plots and histograms, similar to the Statistics Visualizer. The clustering algorithm displays results using the Cluster Visualizer by default, but other visual tools may be used as an alternative.
- **Regression Tree**—Induces a regressor that predicts a real value, that is, results with gradations of value rather than specific predetermined limits.
- **Column Importance**—Determines the importance of specific columns in discriminating one label value from another. Used to observe the varying effects of changing variables, or to suggest columns to map to the axes of the Scatter and Splat Visualizers.

MineSet contains additional tools to aid the knowledge discovery process:

- **Statistics Visualizer**—Data is displayed in the form of box plots and histograms, one per column. Continuous columns are shown as box plots, discrete columns are shown as histograms.
- **Histogram Visualizer**—Data is displayed in the form of histograms. Continuous columns are binned (broken down into ranges).
- **Record Viewer**—The original data is displayed as a spreadsheet.

The next chapter, [Chapter 2](#), describes a typical data mining process and how the tools are used.

Data Mining Process

This chapter introduces specific tasks involved in the knowledge discovery process. The knowledge discovery process is iterative (represented in [Figure 2-1](#)) where the you go back to earlier stages once you discover new patterns and improve your understanding of the data.

The common steps of this process are:

1. Identify the source of the data—see [“Identifying the Data” on page 9](#).
2. Prepare the data—see [“Preparing the Data” on page 10](#).
3. Build a model—see [“Building a Model” on page 12](#).
4. Evaluate the model—see [“Evaluating a Model” on page 12](#).
5. Deploy the model—see [“Deploying a Model” on page 12](#).

Identifying the Data

The task of identifying the data begins by deciding what data is needed to solve a problem. For example, predictability about customer behavior is often a necessary goal recast in terms of a problem. In defining the problem, the investigator must identify the data needed to solve that problem and explore other possible sources of data.

Data may be in a difficult location or in an obscure form. Sometimes there are several initial databases that may be incompatible with each other. Further, if data is scanty or incomplete, more data may be needed. The form in which new data is to be collected depends on the form of existing data. MineSet supports native interfaces to several commercial databases (Oracle, Informix, SQL), ODBC interface, as well as reading data from different file formats (tab-separated flat file, MineSet binary file, Excel, SPSS, Mutable, etc.)



Figure 2-1 Data Mining Process

Preparing the Data

Data may need modification before loading into MineSet (a step often called cleaning.) Specifically, the following problems are common:

- Data may be in a format incompatible with MineSet representation (for example, binary, encoded, or EBCDIC strings from old mainframe computers).
- Data may be misspelled or erroneous, or have incomplete, or erroneous values.
- Field descriptions may be unclear or confusing, or may mean different things depending on the source. For example, order date may mean the date that the order was sent, postmarked, received, or keyed in.
- Data may be out of date; for example, customers may have moved, changed households, or changed spending patterns.

Even clean data may need to be transformed before it is suitable for mining and visualization.

Transforming the Data

Transformations can greatly improve model performance. If you were analyzing telephone company data, for instance, you may find that long distance rate (sales divided by total minutes used) is a better predictor of customer behavior than either element given separately. Data transformations are at the heart of developing a sound model. As you progress, you may even go back and transform the data differently. You can transform the data by:

- Adding columns, usually by applying a mathematical formula to existing data to create a new field.
- Removing columns that are not pertinent, are redundant, or contain obvious, uninteresting predictors.
- Filtering visualizations. For example, you may want to see only the strongest rules or the most profitable customer segments.
- Binning data—breaking up a continuous range of data into discrete segments (for instance, [1 - 10], [11 - 20], and so on).
- Aggregating data—grouping records together, and finding the sum, maximum, minimum, or average values.
- Sampling the data to get a random subset of the data (by percentage or count).
- Applying a classifier that you have previously created, to label new records with a class label, or to estimate the probability of a given label value.

In MineSet, most of these transformations take place using the Data Transformation pane in Tool Manager.

Building a Model

At the core of the knowledge discovery process is model building, automatically done by analytical data mining algorithms. This is clarified in [Chapter 3](#).

Evaluating a Model

Evaluating the accuracy of a model refines your understanding of that model and its usefulness. This allows you to improve the model by filtering the data, eliminating columns, creating new columns, and so on.

MineSet implements four model assessment methods: error estimation, confusion matrix, lift curve, and ROI (return-on-investment) curve.

Deploying a Model

A model can be deployed by applying it to new data. New data can give rise to further questions, which may require further refinements.

In the telecommunications example in [Chapter 3](#), a model can be created to determine which customers are likely to leave their phone carrier. Customer records can then be evaluated through the model to identify the specific customers most likely to leave. These customers can be given incentives to stay.

Applying the Process to a Specific Database

The next two chapters step you through the knowledge discovery process on the churn dataset—a prepared dataset of telecommunication customers. As you work through the examples, think of the process presented here and how your operations progress forward and loop back as shown in [Figure 2-1](#).

Churn Tutorial

This chapter steps you through a possible knowledge discovery process using the *churn* dataset provided with MineSet. It is assumed that MineSet is installed on the system you use, together with all the sample files. Each step is explained in detail. Unless otherwise noted, each step builds on the step before. These steps are:

- [“Starting MineSet” on page 14](#)
- [“Viewing the Records” on page 15](#)
- [“Building an Evidence Classifier” on page 18](#)
- [“Viewing Probabilities With Splat Visualizer” on page 21](#)
- [“Visualizing Geographic Distributions” on page 25](#)
- [“Creating a Decision Tree Classifier” on page 30](#)

About the Raw Data

The *churn* dataset deals with telecommunications customers—people who use the phone regularly. Customers have a choice of carriers, or companies providing them with telephone service. When these customers change carriers they are said to “churn,” which results in a loss of revenue for the previous carrier. A telecommunications company is likely to have a database of call records containing call information (source, destination, date, duration), a billing database, a customer database, and a customer service database. Relevant information about the customer appears in all these databases. This information, when combined, yields a set of customer signatures. The churn dataset provided with MineSet is such a set; the step of identifying the data and creating customer signatures into records has already been done. This dataset, which is used in the rest of the chapter, contains one record per customer.

Starting MineSet

1. Start MineSet by choosing Start > Programs > MineSet 3.0 Enterprise Edition > MineSet, or double-click the MineSet icon on your desktop.
2. If the “Log in to Server” dialog box (Figure 3-1) does not come up by default, choose File > Connect to Server. In the resulting dialog box, click “This machine as current user,” if you wish to use your current system as both client and server. If you wish to use another system as a server, type in the server name, your login name, and your password (if any).



Figure 3-1 Tool Manager Login Window

3. Click *OK*. If you have previously logged in to MineSet, you may be presented with a restored session. For the purposes of this tutorial, you must open a new file (step 4).
4. In the Tool Manager window, choose File > Open New Data File. If the resulting dialog box does not show the data directory, go to the directory where MineSet was installed, in which the default is MineSet 3.0 > data.
5. Select *churn.schema*. A series of entries appears in the right-hand Preview Columns pane as shown in Figure 3-2.
6. Click *Open*.

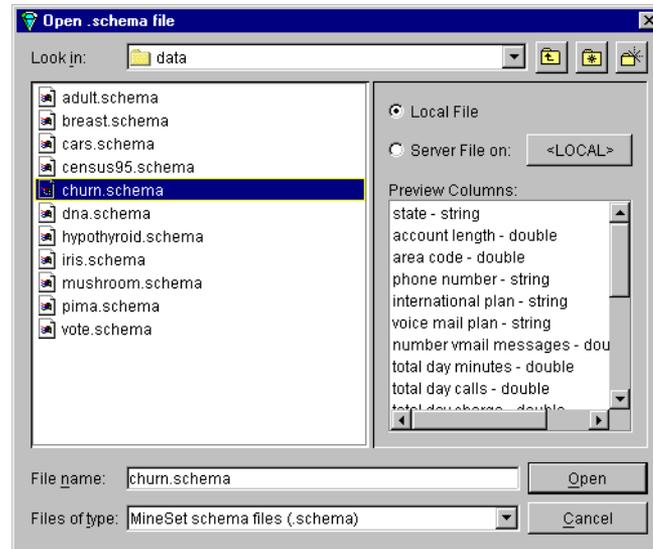


Figure 3-2 Open New Data File Window

This gives you access to a dataset of telecommunications customers. The next time you run MineSet, you will be automatically returned to this position (or wherever you were when you last exited MineSet), and any option selections you made will be saved.

Viewing the Records

You can see the records in spreadsheet form, after bringing up MineSet Tool Manager, by following these steps:

1. In the upper row of tabs in the Data Destinations pane of the Tool Manager, click the Viz Tools tab; then in the lower row of tabs, click *Records* to access the Record Viewer tab.

The *churn* dataset used in the rest of the chapter contains one record per customer. The Data Transformations pane on the left side of Tool Manager lists columns with their type: state (string), account length (double), and so forth. Columns are defined as double or float if they are numbers, or string if they are made up of characters.

2. Click *Invoke Tool* at the lower right.

The data appears as a spreadsheet. The columns and their meanings are shown in [Table 3-1](#).

Table 3-1 Details of Columns Shown for churn Dataset by MineSet Record Viewer

Column name	Value
state	Two-letter abbreviation for the customer's U.S. state of residence
account length	Numerical value indicating the number of months the customer has been with the long-distance carrier
area code	Three-digit telephone company designations
phone number	Three+four-digit telephone company designations
international plan	Special pricing package for international calls, expressed as a yes/no value
voice mail plan	Special pricing package for customers with voice mail provided by the carrier, expressed as a yes/no value
number of voice mail messages	Average number of voice mail messages per day
total day minutes total eve minutes total night minutes total intl minutes	Average number of minutes charged at the carrier's day, evening, night, or international rate
total day calls total eve calls total night calls total intl calls	Average number of calls made during the carrier's day, evening, night, or international hours
total day charge total eve charge total night charge total intl charge	Average amount charged at the carrier's day, evening, night, or international rate
number customer service calls	Number of calls this customer made to carrier customer support in the last six months
churned	Whether this customer changed long-distance carriers in the last six months, expressed as a yes/no value

3. Close the Record Viewer window. You should see the Tool Manager window again, still using the *churn* data source.
4. In the Data Destination pane of the Tool Manager window, the Viz Tools tab should still be displayed; in the lower row of tabs, click the Statistics tab.
5. Click *Invoke Tool*.

The Statistics Visualizer display appears consisting of a number of histograms and box plots. The histograms show the distribution of values for discrete variables, and the box plots show summary statistics for continuous variables.

Each box plot (on the right in [Figure 3-3](#)) shows statistics about data from a single column, including the minimum, maximum, mean (in red), median, and two out of four quartiles (25th and 75th percentiles). These values are marked as lines, and the standard deviation (in red) is shown after the +/- sign.

The mean is the number found by adding the data in a column, then dividing by the number of records. The median is the middle number when numbers in a given column are arranged in order of size. The standard deviation is a measure of the dispersion of the data in a column.

The histograms consist of specific discrete values: state names or yes/no values. Scroll down to find the churned histogram in the display (see [Figure 3-3](#), left). It shows that 707 customers out of 5,000 have left the carrier. The churned column is important throughout this tutorial.

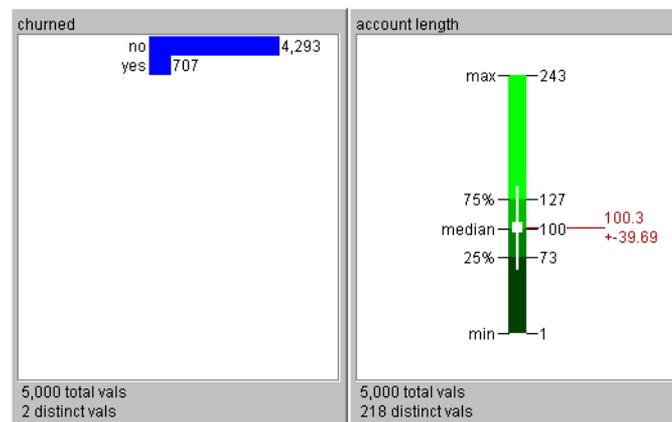


Figure 3-3 Representative Histogram and Box Plot Produced by Statistics Visualizer

6. Close the Statistics Visualizer window and return to the Tool Manager window.

Building an Evidence Classifier

You are now ready to perform analytical data mining. Verify that MineSet is connected to the appropriate server, and that the data source is *churn.schema*. If you exited MineSet between sessions, the history file automatically returns to where you left off.

1. In the Data Transformations pane of the Tool Manager, delete the following columns. Click a column, then press and hold the Ctrl key to gather the rest. Then click the *Remove Column* button.

phone number
total day minutes
total day calls
total eve minutes
total eve calls
total night minutes
total night calls
total intl minutes
total intl calls

The phone number column has no predictive value. The total minutes and total calls columns correlate with the total charge columns and add little extra information. Removing these columns shortens processing time for inducing the model and produces a more succinct visualization.

2. In the upper row of tabs in the Data Destinations pane of the Tool Manager window, click the Mining Tools tab.
3. In the lower row of tabs click the Classify tab, and make selections from these pulldown menus:

Mode: Classifier & Error

Inducer: Evidence

Discrete Label: churned

You are about to induce an evidence classifier to help characterize the customers who are likely to churn. The default mode, Classifier & Error, uses a holdout method on the data, inducing the classifier from two-thirds of the data and leaving the remainder as a test set to estimate the error rate.

4. Click *Go*

As the inducer reads the data you may be warned that the column “state” will be removed because it has too many unique values. If this happens, go to the Tool Manager File > Preferences menu and change the default maximum attribute values to 100.

The Status window on the bottom of Tool Manager shows progress and summary information about the induction process, including the estimated error rate of 11.40% plus or minus .78%. When the induction step is done, the Evidence Visualizer is automatically invoked, showing the model visually (Figure 3-4).

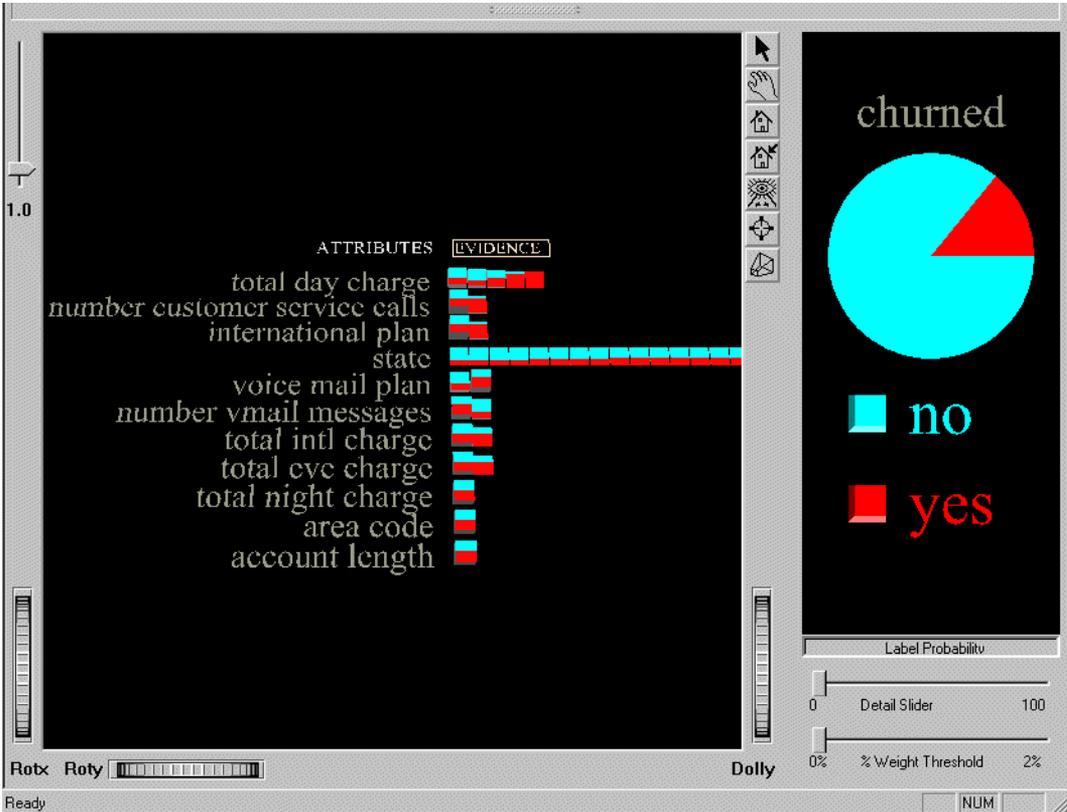


Figure 3-4 Evidence Visualizer Window

The Evidence Visualizer sorts the columns by discriminating power for the label “churned,” starting from the top. To adjust the view quickly, use the Dolly thumbwheel. For this tutorial both the square cake charts in the left pane and the pie charts in the right pane are termed “charts.”

The Label Probability pane (on the right side in [Figure 3-4](#)) shows a pie chart representing *prior probability*. Prior probability means the probability of a random record having a churn value of “yes” (red wedge) or “no” (blue wedge), without taking into account any of the attribute values. Mathematically, this is the number of records with the class label, divided by the total number of records.

In the Evidence pane (on the left in [Figure 3-4](#)), the columns in the dataset are represented by charts, each one with a value or range of values for that attribute. Switch the cursor mode from grasp (hand) to pick (arrow), and click on the box titled Evidence. The cake charts show *conditional probability*, that is, the probability that a customer with the particular attribute value represented by the pie chart (for instance a “yes” value in the “voice mail plan” column) have a churn value “yes.”

Click on a chart to update the right pane to show the expected probability according to the model.

Navigate using the thumbwheels at the screen’s border, or use the mouse buttons and Ctrl key in various combinations. See [Appendix A, “Navigating in the MineSet Visualizers,”](#) for more details about navigation controls.

You can see the factors that affect churning, because the slice representing churn increases from left to right on the first and second rows in [Figure 3-4](#), so a serious problem is evident. Customers that use the company’s service the most also churn at a higher rate. The company is losing its most valuable customers.

To find out about a class label (for example, a churn value of “yes”), select a value in the Label Probability pane on the right. Click on the button by the label “yes” in the right pane, and the evidence is shown as bars. Pointing to bars will show you the estimated probabilities.

The discriminating attributes shown here also can be used to choose axes for a scatterplot visualization. The state attribute, shown relatively high on the list of attributes, hints at a possible geographical relationship.

Evidence models use and show attributes independently; however in many datasets attributes are not independent, and a set of attributes, considered in combination, is better for determining the label. The error estimation in the status window shows that the classifier expects about an 12% error rate. Later we generate a decision tree that is more accurate. Close the Evidence Visualizer before moving on to work with the Splat Visualizer.

Viewing Probabilities With Splat Visualizer

The Splat Visualizer requires that the column mapped to color must have a numerical value. The churned column is a string that must be converted to a number (p_churned, indicating the probability of churning), before mapping it in Splat Visualizer:

1. In the upper row of tabs in the Data Destinations pane, click the Viz Tools tab; then, in the lower row of tabs, click the Splat tab to access the Splat Visualizer.
2. In the Data Transformations pane, click *Add Column*.

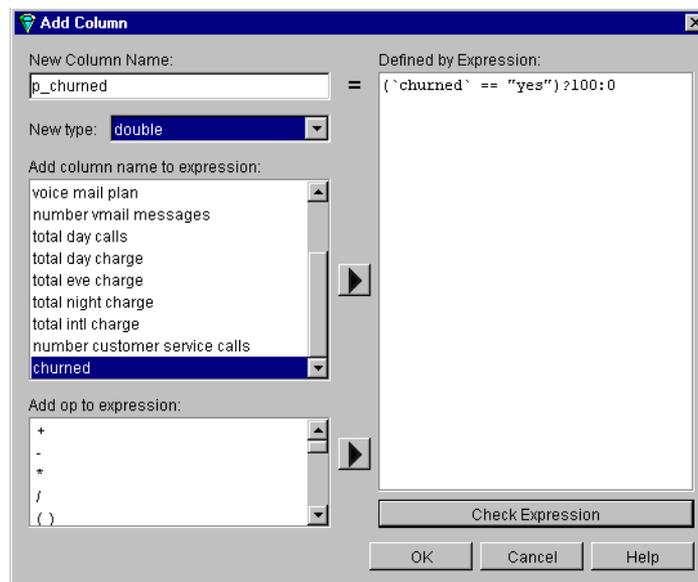


Figure 3-5 Adding a New Column

3. In the Add Column dialog box ([Figure 3-5](#)), in the New Column Name text field, enter the new name, `p_churned`. The intention is to make a column of numbers, based on the churned column.

In the Defined by Expression text field, create the expression

`(`churned` == "yes") ? 100 : 0`. You can create this expression from the two scrolling lists on the left: “Add column name to expression” and “Add op to expression,” or you can type it in directly. This expression translates to “if the value in the churned column is “yes,” give `p_churned` a value of 100, otherwise give it a value of 0.” The purpose of this is to translate a string (yes or no) into a numerical value. Verify that the “New type” text field is set to double.

Click *Check Expression* to ensure there are no syntax errors. Click *OK* to dismiss the dialog box and *OK* to add the column.

4. In the Data Transformations pane, under the Splat tab, map columns to the visual elements by selecting a column in the pulldown menu next to each element. For this tutorial, from the pulldown menu:

for Axis 1 choose “total day charge”

for Axis 2 choose “number customer service calls”

for Axis 3 choose “international plan”

for Color choose “`p_churned`,” so the result is similar to [Figure 3-6](#).

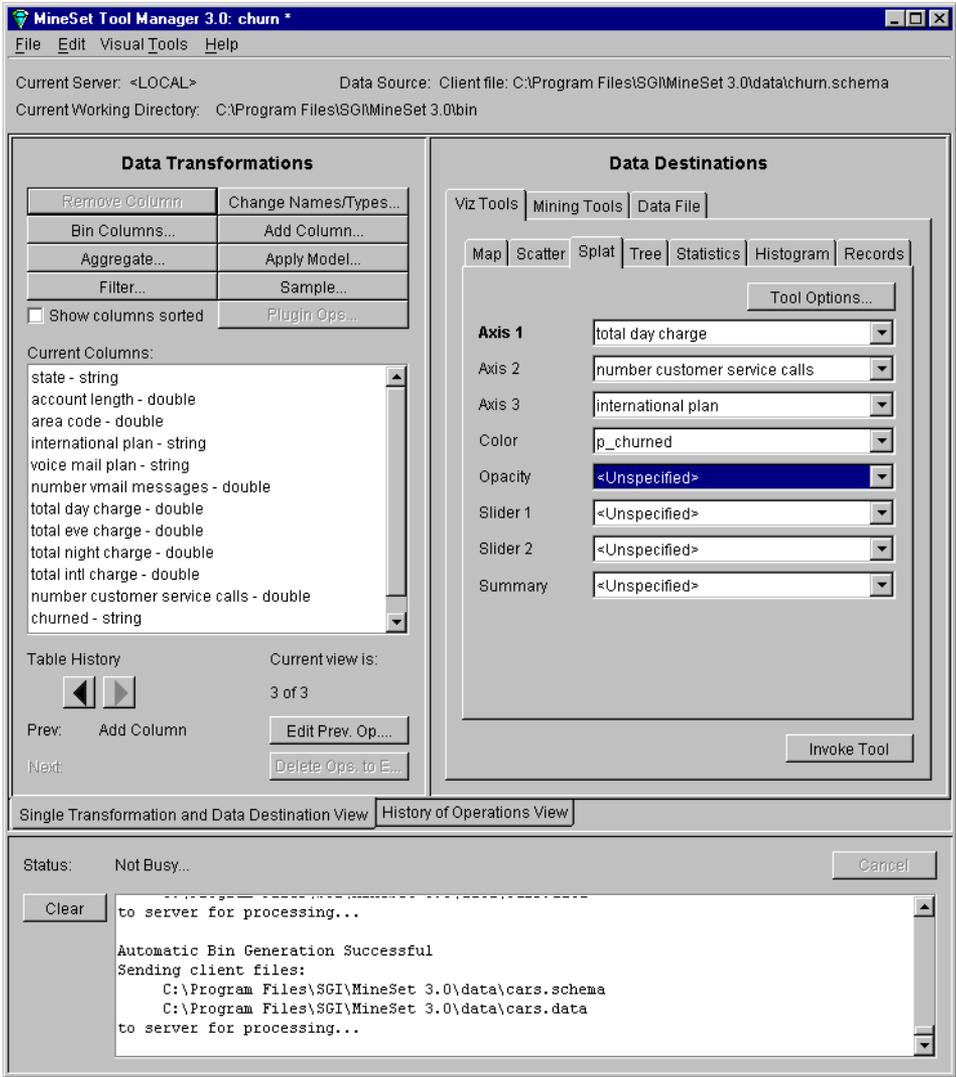


Figure 3-6 Mapping Columns to Elements for Splat Visualizer

5. Click *Invoke Tool*.

The data is plotted on the Splat Visualizer window, shown in [Figure 3-7](#). The slider bar in the upper left varies the color density. See [Appendix A, “Navigating in the MineSet Visualizers,”](#) for help on window manipulation. To navigate in the scene and examine different areas, click and hold both left and right mouse buttons, and move the cursor across the scene. Splat Visualizer allows you to analyze complex data by viewing the varying behavior in several dimensions.

You can save the current state of the Tool Manager including special options by choosing `File > Save Current Session As`, and specifying `churn1.mineset`.

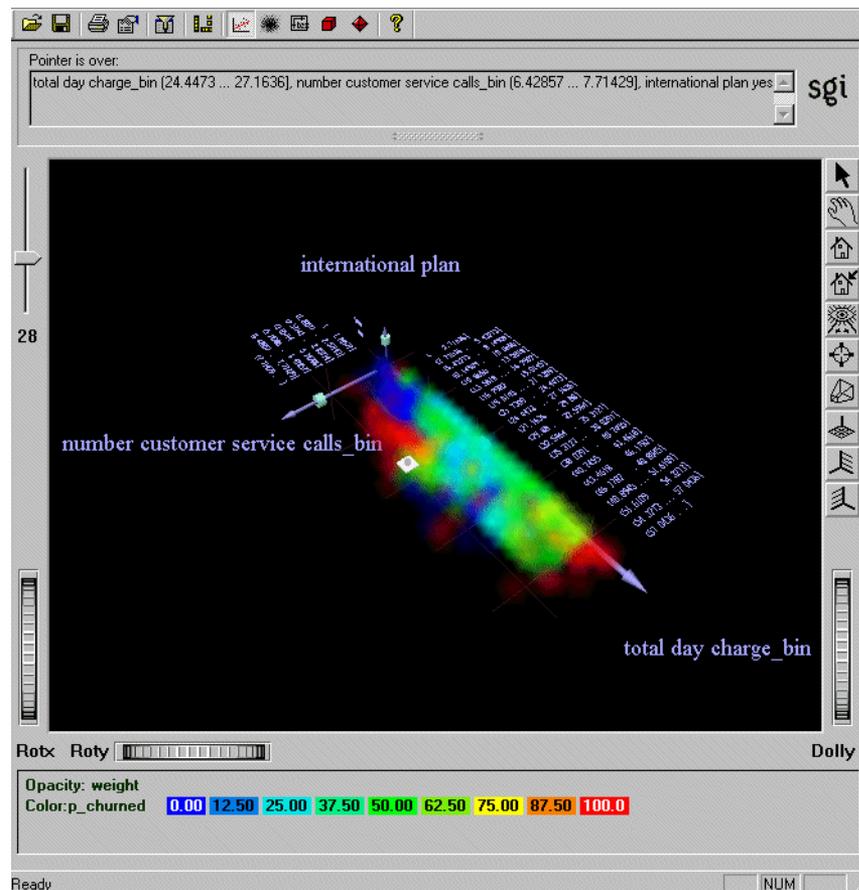


Figure 3-7 Splat Visualizer Window

In the visualization shown in [Figure 3-7](#), the highest probability of churn occurs in two places: in the yellow to red areas when total day charge is high, shown in the bottom of this figure; and when the total day charge is low and customer service calls are high (near the upper left of this figure). Low-paying customers who make many customer service calls leave. These are customers you may not want to keep, because they cost you money and bring little reward. The high-paying customers at the bottom of the figure are a better target.

6. Close the Splat Visualizer and return to the Tool Manager Window.

Visualizing Geographic Distributions

As shown in [Figure 3-4 on page 19](#), the Evidence model indicated that state was a good discriminating attribute. If your model does not show this, you may have failed to change the maximum values as described in Step 4 of “[Building an Evidence Classifier](#)” on [page 18](#). This section builds on previous computations to display data geographically to illustrate how churn varies by state.

You have already added the column (p_churned) from existing columns in the dataset. You can now transform the data into a smaller dataset that contains the average churn per state. Such a transformation is called aggregation.

1. In the Data Transformations pane of the Tool Manager window, click *Aggregate*.

In the Aggregate dialog box move p_churned into the left column (highlight it and click the left arrow). Highlight p_churned in the new window, check Average and Count on, and ensure Sum, Min, and Max are unchecked. Leave state in the central column and move all the rest to the right column. (Hold down the Ctrl key to gather multiple columns.) Make sure your screen looks like [Figure 3-8](#). Click *OK* to apply your choices.

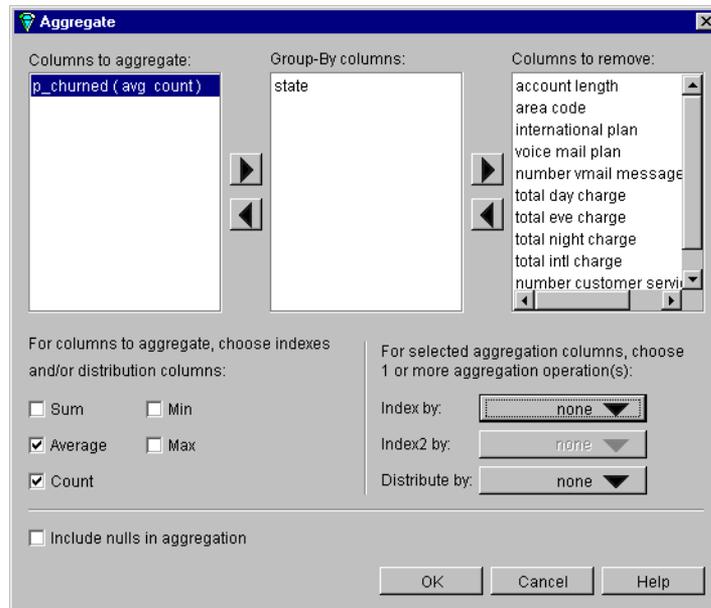


Figure 3-8 Aggregate Dialog Box

2. In the upper row of tabs in the Data Destinations pane, click the Viz Tools tab, then the Records tab. Click *Invoke Tool* to see a record for each state, with the average churning customers and the total number of customers who churned for that state.
3. Close the Record Viewer window and return to the Tool Manager window. You will now link this data to a map of the United States.
4. From the lower row of tabs in the Tool Manager Data Destination pane, click the Map Visualizer tab. Then click the *Tool Options* button. The Map Visualizer Options dialog box appears (Figure 3-9).
5. Click the button to the right of the Entities File text field. From there, navigate to the directory where MineSet was installed and choose *config > mapviz > gfx_files>usa.state.hierarchy*.

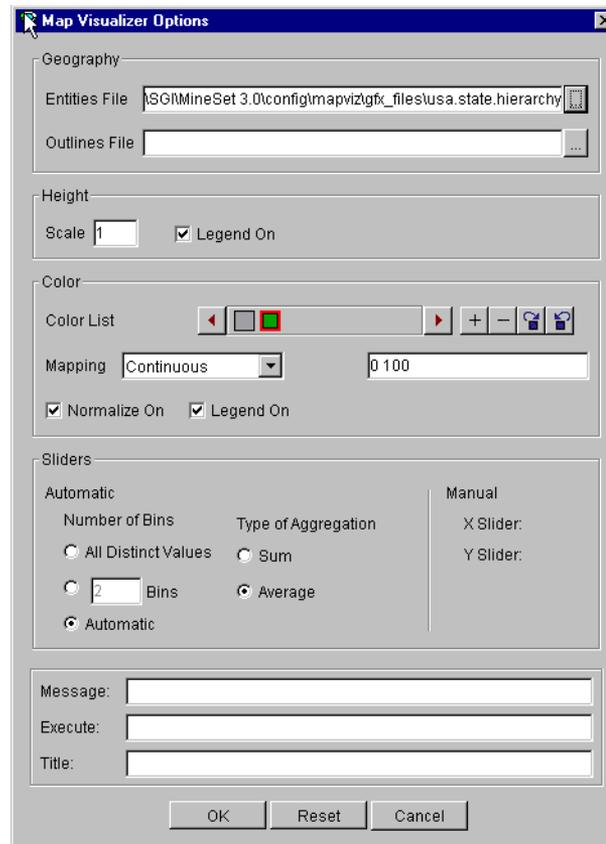


Figure 3-9 Map Visualizer Options Dialog Box

6. Click *Open* to retrieve that file, and *OK* to dismiss the Map Visualizer Options panel.

The next step is to link the visual elements to the columns.

7. Map the current columns in the Data Transformations pane to elements in the Data Destinations pane by doing the following (see [Figure 3-10](#)). From the pulldown menus next to the visual elements:

Entity-Bars choose state

Height-Bars choose count_p_churned

Color-Bars choose avg_p_churned.

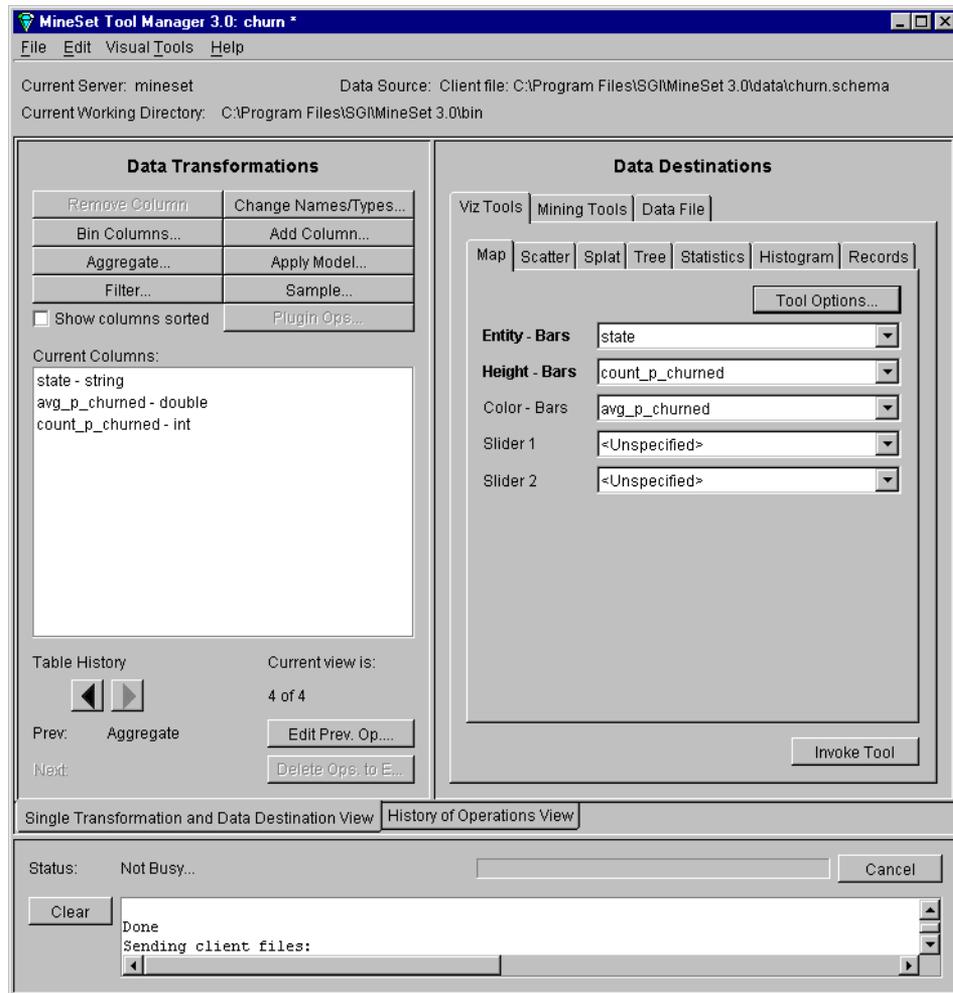


Figure 3-10 Mapping Columns to Visual Entities for Map Visualizer

8. Click *Invoke Tool* to view the map distribution of churned customers according to state (Figure 3-11).

The visualization shows the distribution of churned customers across the United States. For each state, the color indicates the probability of churn and the height indicates the number of customers in that state. For example, in [Figure 3-11](#), Maine is chosen, showing an average churn rate of 18.4466%, but based on the churn count of 103. In other words, the average is based on only 103 customers. West Virginia shows the greatest height, with a probability of churn based on 158 customers. States showing the clearest, brightest colors calculate an average churn rate over 21% (Texas, Montana, Washington, California, and New Jersey). This visualization indicates that there is no obvious relationship between churn and geography, although different states do have different churn rates.

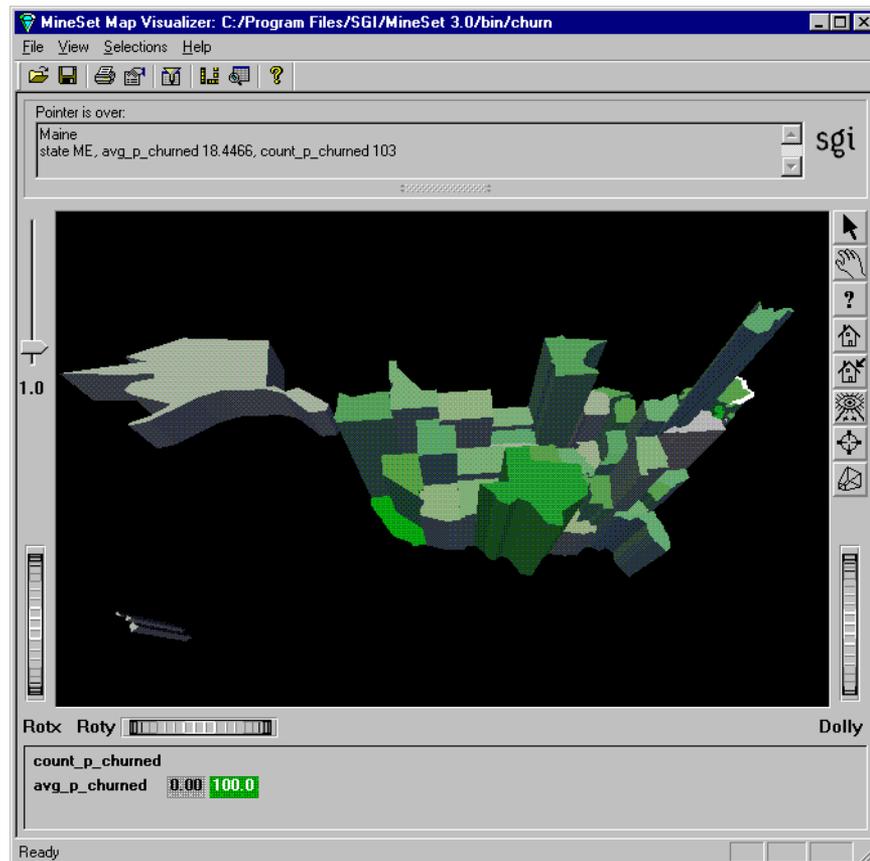


Figure 3-11 Map Visualizer Window With Average Churn Distribution

Close Map Visualizer using File > Exit. The next example explores the Decision Tree classifier, using the same dataset to produce a different visualization.

Creating a Decision Tree Classifier

Unlike the Evidence classifier, the Decision Tree classifier can show attribute interactions, that is, combinations of attribute values that affect the classification. For this section, begin using the history mode. Follow these steps to build a decision tree classifier and visualize it.

1. Switch to history mode by clicking on the History of Operations View tab at the bottom of the Tool Manager Data Destinations panel.
2. Delete the Add Column and Aggregate operations from the history by right clicking on them and selecting delete.
3. Switch back to Single Transformation and Data Destination View. You should now be at “Current view is: 2 of 2”
4. In the upper row of tabs in the Data Destinations panel, click the Mining Tools tab.
5. In the lower row of tabs, click the Classify tab, and make selections the following selections from the pulldown menus:

Mode: Classifier & Error

Inducer: Decision Tree

Discrete Label: churned

6. Click *Go*

MineSet classifies and creates the Decision Tree model as shown in [Figure 3-12](#). The estimated error rate is significantly improved (6.36% +/-0.60%) over the Evidence Visualizer in [Figure 3-4](#), confirming the earlier hypothesis that interactions between attributes are significant. In [Figure 3-12](#), every node in the decision tree has two bars, one for each label value. Pointing to a bar shows the record count and percentage for that label value. The base of every node indicates the number of records that reach it, and a color, and the estimated error rate for the subtree (see legend on bottom of the visualization).

In this example the root of the decision tree is total day charge, indicating that this is the single most important factor—how much money the customers spent on daytime calls, with a dividing threshold of 44.96.

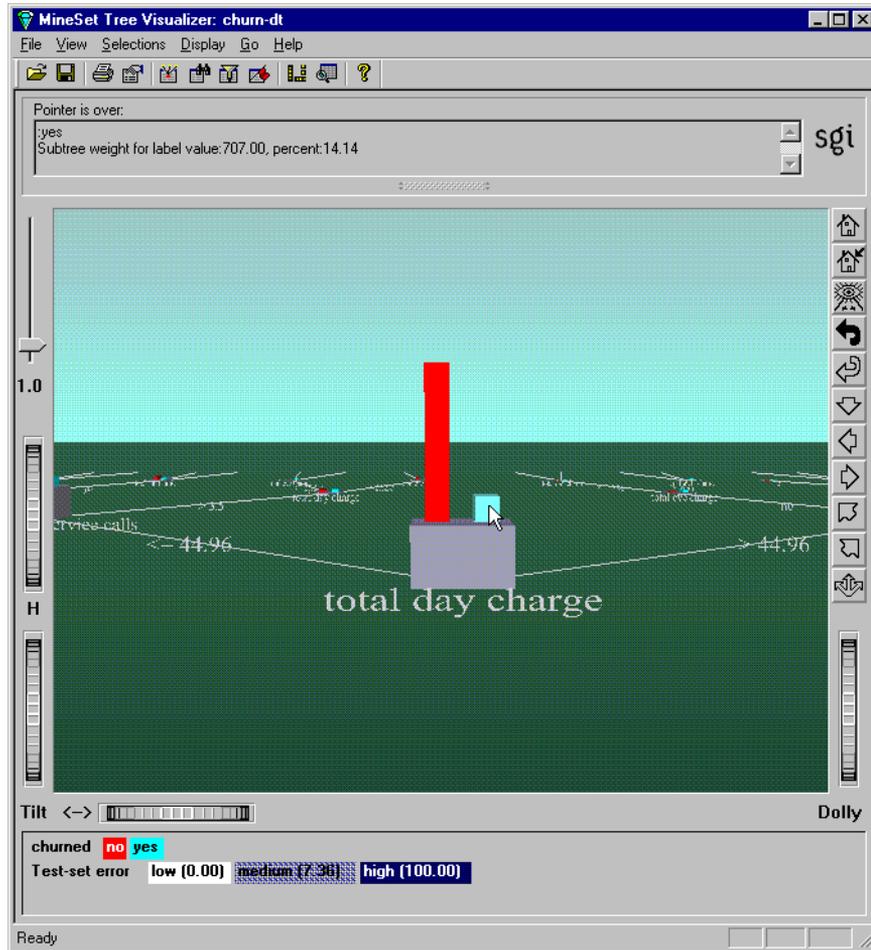


Figure 3-12 Tree Visualizer Window

You can fly through the Tree Visualizer landscape using the Dolly wheel or mouse button combinations. (See [Appendix A, “Navigating in the MineSet Visualizers,”](#) for more navigation information). By selecting the red bar (churned: yes) at the root, the right you can see that 14.14% of the customers churn. Click the right line (total day charge > \$44.96) to move to the child node, which contains customers who spend the most on daytime calls. Of these customers, 59.31% churn. Again click either branch line to move to the next node, which shows those customers who spend the most on daytime calls and may or may not have a voice mail plan. Of these customers, those with voice mail churn at the much lower rate of 9.33%. Perhaps offering voice mail to customers can help reduce churning.

It is important to understand that this tree was automatically induced from the data. The attributes chosen for nodes and the thresholds are determined by the process of induction.

To drill-through and see the original data, select a node base or a bar and choose Selections > Show Original Data. A Record Viewer shows the records matching the node you selected.

If you would like to explore MineSet further, and discover more about applying a classifier, continue to the next chapter, [Chapter 4, “Further Explorations.”](#)

Further Explorations

This chapter continues to explore the MineSet tools. It assumes that you have worked through Chapter 3, “Churn Tutorial,” and prepares you to use other aspects of MineSet:

- “Exploring Data Clusters” on page 33
- “Invoking a Decision Table” on page 42
- “Targeting Customers Using a Model” on page 44
- “Reducing Misclassification Costs” on page 51
- “Further Exploration of MineSet” on page 57

Exploring Data Clusters

When confronted with an unfamiliar dataset, you can discover interesting attributes or characteristics using the clustering algorithm. This non-predictive algorithm segments records into clusters that are similar in several ways. For this example, return to the Tool Manager window, and begin a new history by reopening the *churn.schema* file.

1. In the upper row of tabs in the Data Destinations pane, click the Mining Tools tab.
2. In the lower row of tabs, click the Cluster tab, and make the following selections:

Method: Single k-Means

Number of Clusters: 3

- In the Data Transformations pane of Tool Manager, select and remove the following columns from the Current Columns pane by highlighting them and then clicking *Remove Column* (see [Figure 4-1](#)):

state
 account length
 area code
 phone number
 international plan (because it correlates to total intl charge)
 voice mail plan (because it correlates to number vmail messages)
 Multiple selections can be made using the Ctrl key.

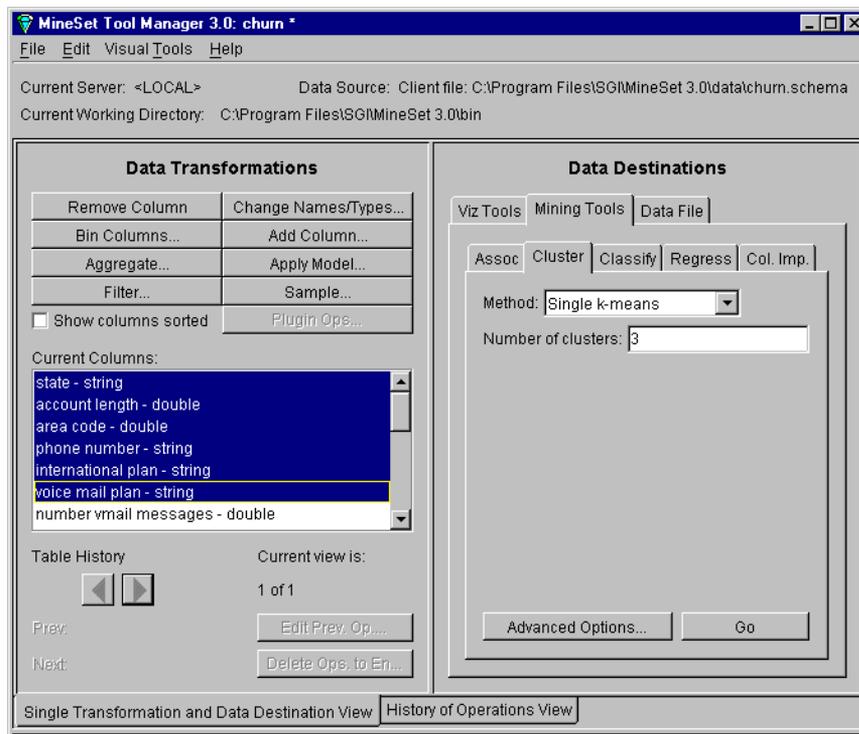


Figure 4-1 Removing Columns to Prepare for Clustering

The columns that are removed are those that are not likely to influence the clustering productively. The column “churned” is retained to help explain results. You can experiment with which columns to remove as you explore the dataset further.

4. Click *Advanced Options* to set the weight of the attributes.

By default, the weight of each column is set to one (1), which means each column is given equal importance. Set the churned column to 0 for this example, to see if this attribute is generated spontaneously as the dataset is clustered. Click *Set* then *OK*.

5. Click *Go* on the right side of the Tool Manager window.

The Status window on the bottom of Tool Manager shows the progress of the clustering operation, as the algorithm selects significant characteristics by which to group the records. The model is saved as *churn.cluster*.

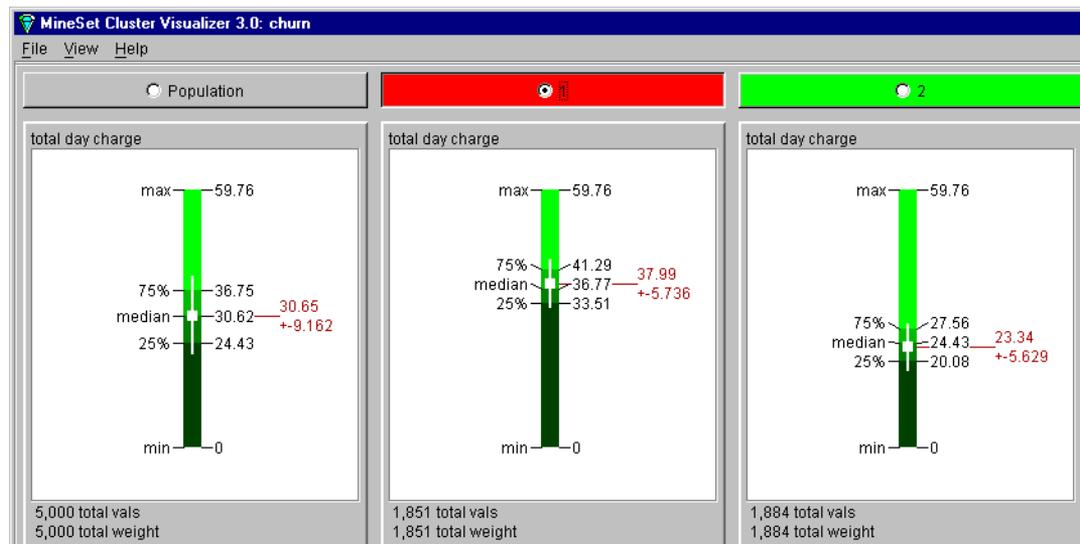


Figure 4-2 Box Plots Produced by Cluster Visualizer

Figure 4-2 shows a partial view of the result of Cluster Visualizer. Adjust the window to see the full width of all the clusters. The columns are sorted by their power in discriminating how one cluster differs from another. Clearly the number of voice mail messages, total day minutes, and total day charge are the most important columns. Color and means are quite different between clusters at the top of columns, yet as you scroll down the display, the differences become minimal.

6. At the top of the display, click the circle next to the cluster number. This changes the attribute ordering, so that attributes important in discriminating this cluster from the others change order.
7. Choose File > Exit in the Cluster Visualizer window to close the window and return to the Tool Manager window.

Relating the Columns and Axes in the Model

With Cluster Visualizer you can look at independent attributes in a dataset, examine the most prominent, and see how each differs. However, to see how attributes relate to each other between clusters, Scatter Visualizer provides a clearer view. To apply the clustered model to Scatter Visualizer, you need to determine which columns should be mapped to the various axes.

1. In the Data Transformations pane of Tool Manager, click *Apply Model* and select *churn.cluster* from the list of available models. Click *OK*.

Although Cluster Visualizer indicated three columns as the most important, each cluster's order of importance was independent, with no indication of interactions between attributes. At this point, the Column Importance tool is useful.

2. In the upper row of tabs in the Data Destinations pane, click the Data File tab, then click the Server checkbox. In the text field, type the filename `churn-crop`, and click *Create File*. This saves the abbreviated version of the churn dataset that is used later in this tutorial.

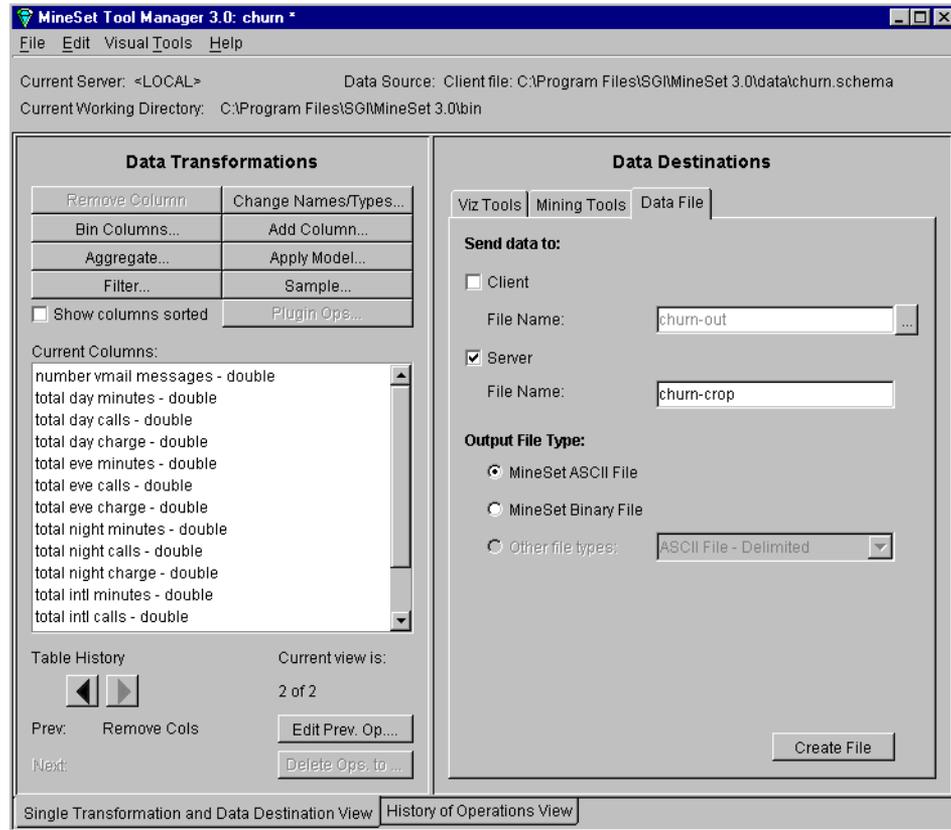


Figure 4-3 Saving Data File to Server

Finding Important Columns in the Clustered Model

1. In the Data Destinations pane of Tool Manager, click the Mining Tools tab, then click the Col. Imp. tab (Column Importance). By default the tool selects the top three columns in terms of importance. The discrete label is Cluster.

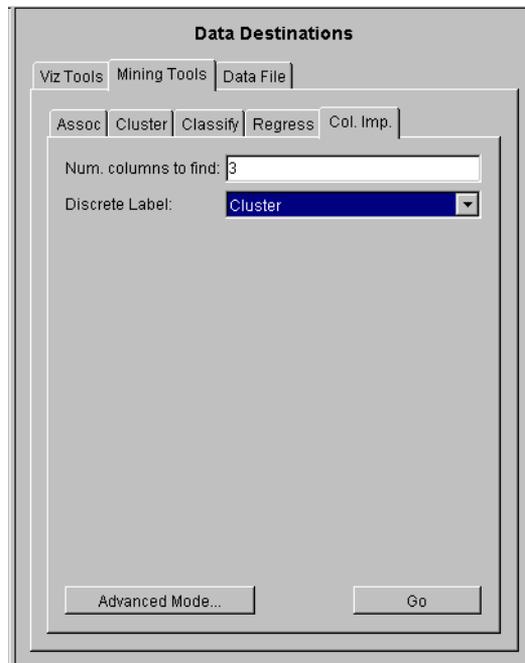


Figure 4-4 Column Importance Selections for Cluster

2. Click *Go*.

The displayed panel that shows:

1. number vmail messages
2. total day minutes
3. total eve minutes

The status window shows that time spent on the phone during the day is a factor, with all other columns showing a correlation. The next step is to map these columns to axes in Scatter Visualizer.

Mapping to Scatter Visualizer

1. In the upper row of tabs in the Data Destinations pane, click the Viz Tools tab; then from the lower row of tabs click the Scatter tab to access the Scatter Visualizer.

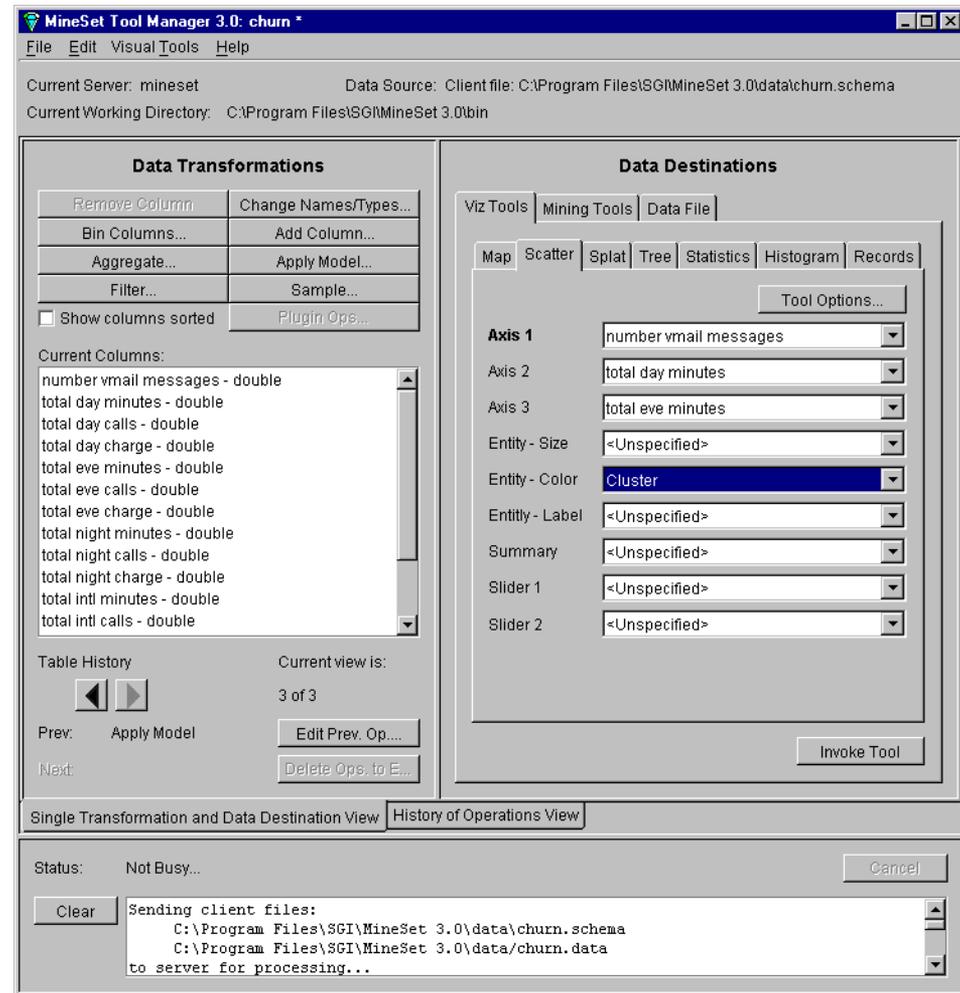


Figure 4-5 Mapping Columns to Axes for Scatter Visualizer

2. In the Data Destination pane, map these elements to the following columns using the pulldown menus (see [Figure 4-5](#)):

Axis 1 choose number vmail messages

Axis 2 choose total day minutes

Axis 3 choose total eve minutes

Entity-color choose Cluster (created when you applied the model)

3. Click *Invoke Tool*.

The Scatter Visualizer window in [Figure 4-6](#) shows the clusters clearly differentiated in color. The blue scatter cubes represent cluster 2, and the flat pancake shape is split evenly between red and green—clusters 1 and 3. This pancake indicates very low numbers of voice mail messages. Clearly, total day minutes and total evening minutes are interdependent. If you click on an interesting visual point, the supporting data is displayed. Dismiss the Scatter Visualizer window and return to Tool Manager for the next step.

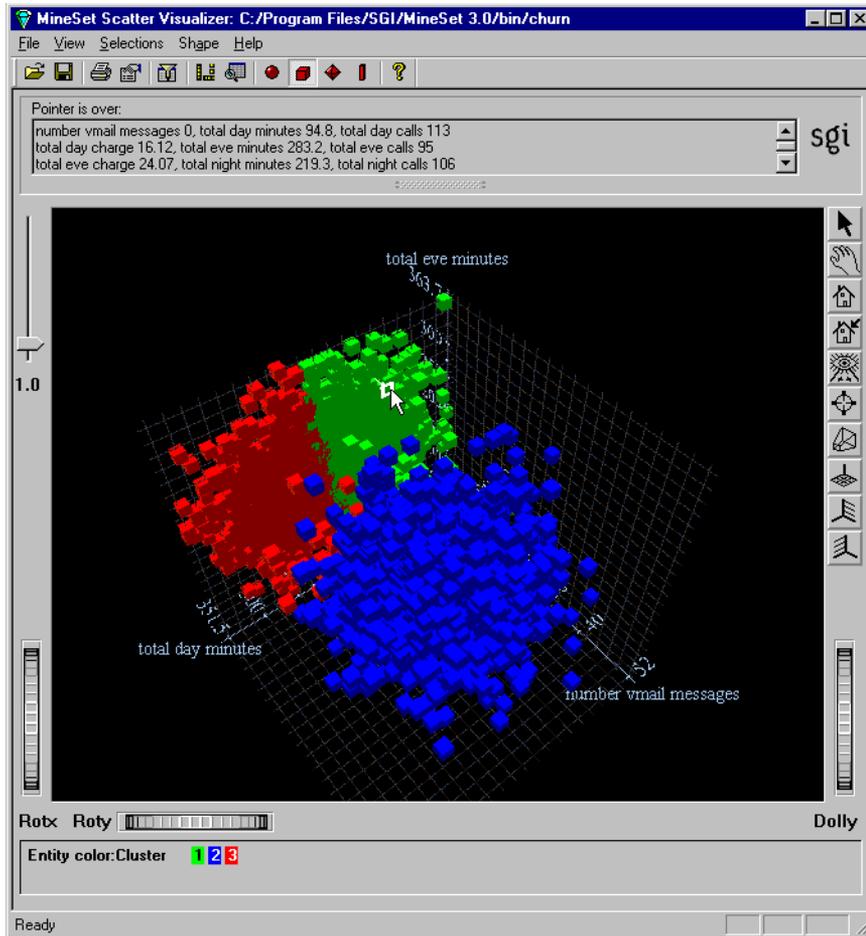


Figure 4-6 Scatter Visualization Plotted From Clustering

Invoking a Decision Table

For this example, instead of using Scatter Visualizer to see your clustered data, you can visualize the same data as a Decision Table.

1. In the Tool Manager window, choose File > Open New Data File.
2. In the Open .schema file window, click the *Server File* button, and select *churn-crop.schema*. This is the file saved earlier. If you exited MineSet between sessions, you are automatically returned to where you left off. This can be set from the Tool Manager File > Preferences menu.
3. Click on the file and click *Open*.
4. In the upper row of tabs in the Data Destinations pane, click the Mining Tools tab.
5. In the lower row of tabs, click the Classify tab, and make the following selections from the pulldown menus:

Mode: Classifier & Error

Inducer: Decision Table

Discrete Label: churned

Make sure you have the correct discrete label. You are about to induce a decision table and allow the algorithm to suggest which columns are most important to map to the X and Y axes.

6. Verify the *Suggest* checkbox is checked, then click *Go*.

You can see columns being mapped to axes as the tool determines appropriate mappings. The Status window on the bottom of Tool Manager shows progress and summary information about the induction process, including the classification error rate. When the induction step is done, the Decision Table Visualizer is automatically invoked, showing the model visually. Manipulate the Dolly thumbwheel to enhance the display, or work with the mouse buttons for navigation— see [Appendix A, “Navigating in the MineSet Visualizers.”](#)

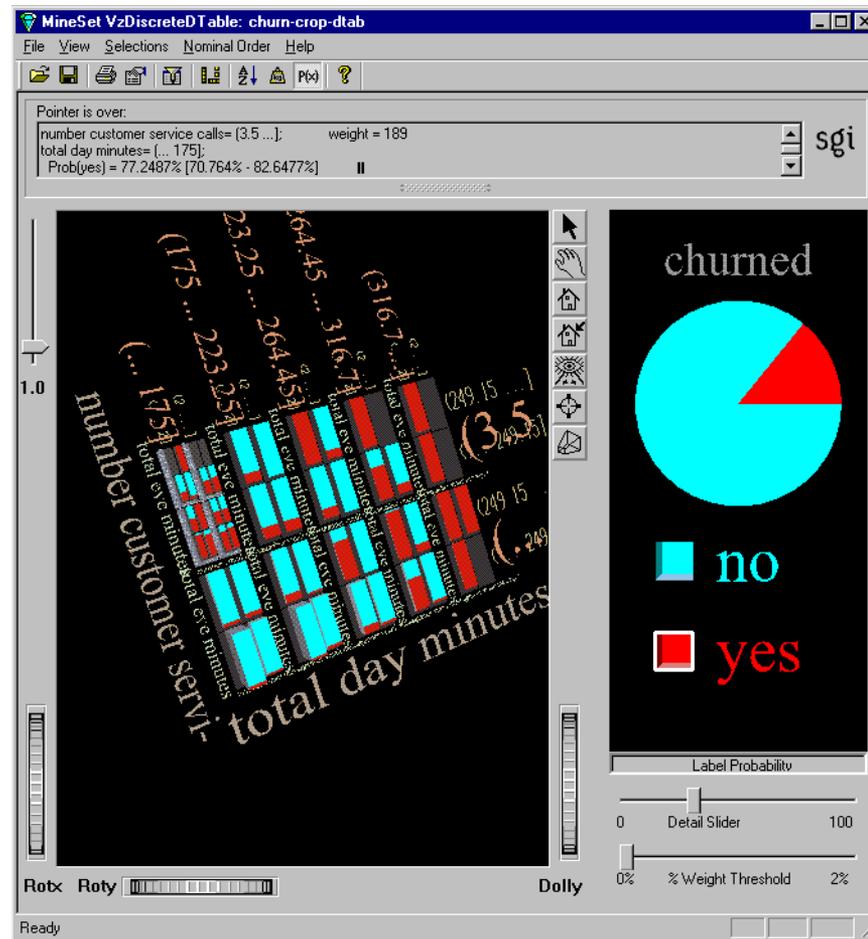


Figure 4-7 Decision Table Showing Clustered Churn Results

Figure 4-7 shows a round pie chart in the Label Probability pane, like Evidence Visualizer, indicating the overall percentage of churn. In the left pane, the data is shown as cake charts, that tell you, within this subset of the data, how much churn exists. Clearly, customers with high total day minutes always churn.

The Decision Table shows you data at different levels of detail, taking only a few columns into consideration at first, adding more detail as you examine further. Change the cursor mode from grasp to pick, and pass the cursor across the scene to display data above the window. Notice the bar that falls out of the expected pattern—total day minutes less than 175, and customer service calls over 3.5. Drill up and down using the mouse buttons, see [Appendix A, “Navigating in the MineSet Visualizers.”](#) Dismiss the display when you are finished examining the Decision Table, and return to Tool Manager.

Targeting Customers Using a Model

Previously, you created models to predict which customers are likely to churn. Now that you have such a model, you may want to target customers who are likely to churn *before* they churn. The lift curve helps accomplish this goal.

A lift curve is a plot in which the X axis shows the number of records from 0 to 100% and the Y axis shows the number of records corresponding to customers who have a given label value (`Churn=yes` in this case). Two curves are shown on the graph in [Figure 4-10](#). The lower curve or line (red) shows the number of customers expected to churn given a random ordering of the records. The upper curve (white) shows the percentage of customers who churn when placed in order according to the classifier's score (probability estimate) for each record. Records representing customers that the model identifies as most likely to churn appear first; those less likely to churn appear last. The advantage that the model ordering provides can be seen by the difference between the model's curve and the random curve.

In building this lift curve, a selected model is applied to the test set. In the example below, a specified segment of the dataset is used for training. Then the induced model is run on the remainder of the dataset. Although lift curves can be generated easily by selecting Lift Curve from the Advanced Options for classifiers, in this tutorial a more complex scenario is shown, one that involves sampling and application of a model to a dataset.

Creating a Training Sample

For this example, return to the Tool Manager base window, and begin a new history by using File > Open New Data File and returning to the local file *churn.schema*.

1. In the Data Transformations pane, click *Sample*. In the Sampling dialog box type 40 for the percentage of sampling, and click *OK*.

This choice simply samples a random 40% of the total dataset, from which the classifier is induced.

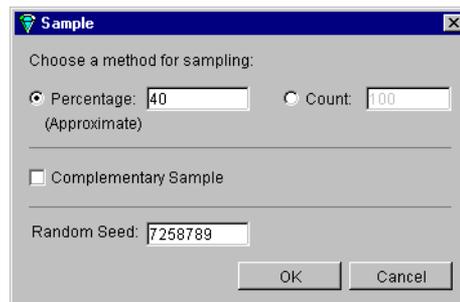


Figure 4-8 Selecting a Sampling for Testing

2. In the upper row of tabs in the Data Destinations pane, click the Mining Tools tab; then from the lower row of tabs click Classify and make the following selections from the pulldown menus:

Mode: Classifier only

Inducer: Decision Tree

Discrete Label: churned

You are inducing a decision tree classifier based on the random 40% sampling, and choosing “Classifier only” because this is the training set. The test set is the remainder of the dataset (excluding the 40% sampled records).

3. Click *Go*

The resulting decision tree demonstrates the model, which is required in the next stage. The root weight is substantially diminished, because the size of the sample is less than the complete dataset, and no color appears at the base of each node, indicating that no error estimation is available.

You can see in the status field that the classifier is automatically saved under the name *churn-dt.class*. The next step is to use this classifier on the remainder of the churn dataset.

Applying a Model

Dismiss the Decision Tree window and return to the Tool Manager window. Because you have used the first 40% of the dataset to build the model, you have the remaining 60% to use as a test set.

1. In the Data Transformations pane click *Edit Prev. Op.* You are presented with the Sampling Dialog box again.
2. In the Sampling Dialog box, enter 40 in the Percentage text field again, but this time click the Complementary Sample box to indicate you want the other part of the sample.
3. Click *OK*.
4. Click the *Apply Model* button in the Data Transformations pane.
5. From the list of available models choose *churn-dt.class*. This is the Decision Tree model built on the churn dataset.
6. Click the Test Model tab in the lower part of the pane; turn on *Show lift curve*, and set the *ROI/Lift label* pulldown menu to yes.

Having built a classifier based on the random sample, you now apply it to the remainder of the churn dataset.

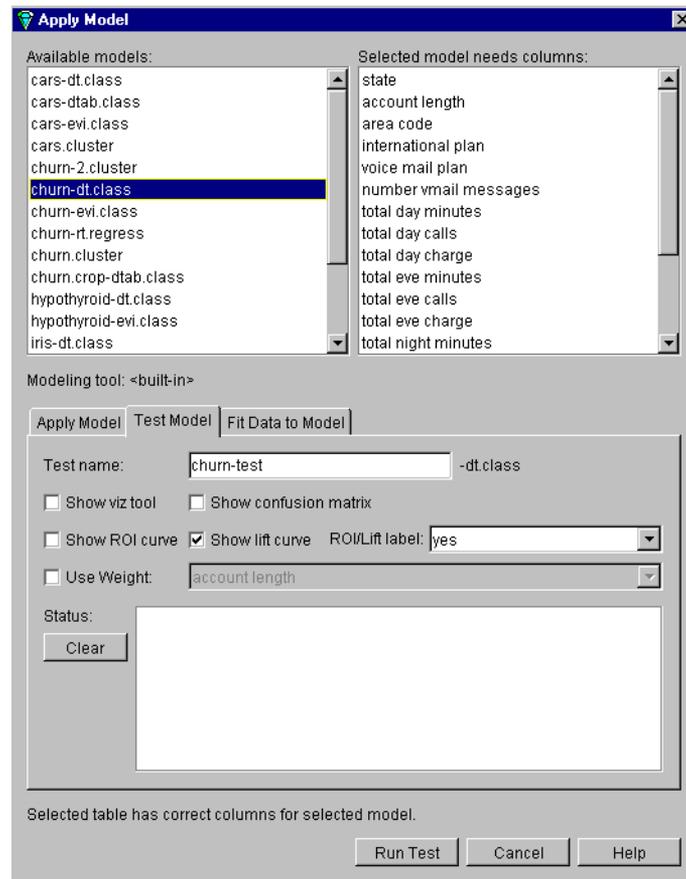


Figure 4-9 Preparing to Test Classifier on Full Dataset

7. Click *Run Test*. The process takes some time. The resulting lift curve is shown in [Figure 4-10](#), with the details of any selected point shown in the upper banner.

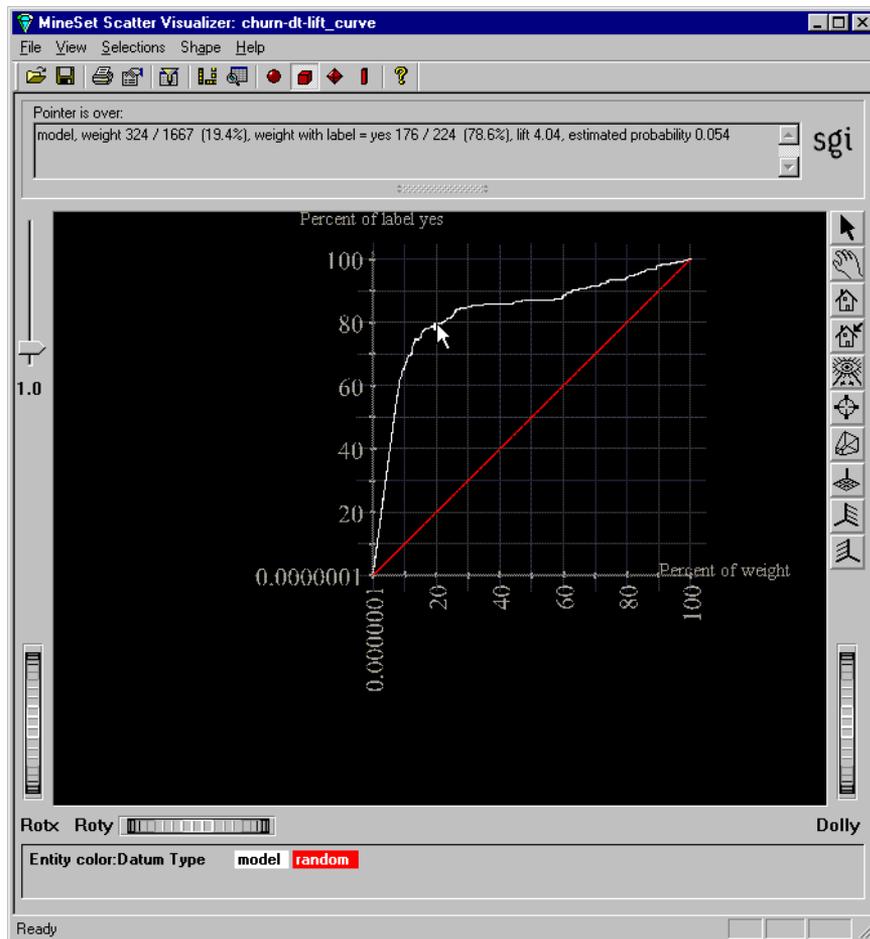


Figure 4-10 Lift Curve

Move the pointer along the white (model) line, clicking at various points to see the lift and percentage of customers with `churn=yes`. Look for the knee of the curve, in this example where the estimated probability of the classifier is 0.054.

This is the point at which the return on investment of sending incentives to customers that may churn diminishes rapidly. The next step is to apply the classifier to the full dataset.

8. Return to the Apply Model dialog box; click the Apply Model tab, select *churn-dt.class* and make these selections:

Estimated probability values for label yes

New column name: p_churned (You must type this in.)

When you click *Estimated probability values for label*, yes is chosen to match the corresponding selection in the Test Model step. This process adds a new column representing the likelihood that certain people will churn (*p_churned*.) Click *OK*.

9. On the Data Transformations pane of Tool Manager click *Filter*, in the “Defined by Expression” text field create the expression `p_churned > 0.054`. Check expression before clicking *OK*.

This is the estimated probability figure retrieved from Step 8 shown in [Figure 4-10](#). The intention is to select only those customers with the greatest likelihood of churning. In a real-life situation, this step would be executed against unlabeled data to predict which of the existing customers are likely to churn.

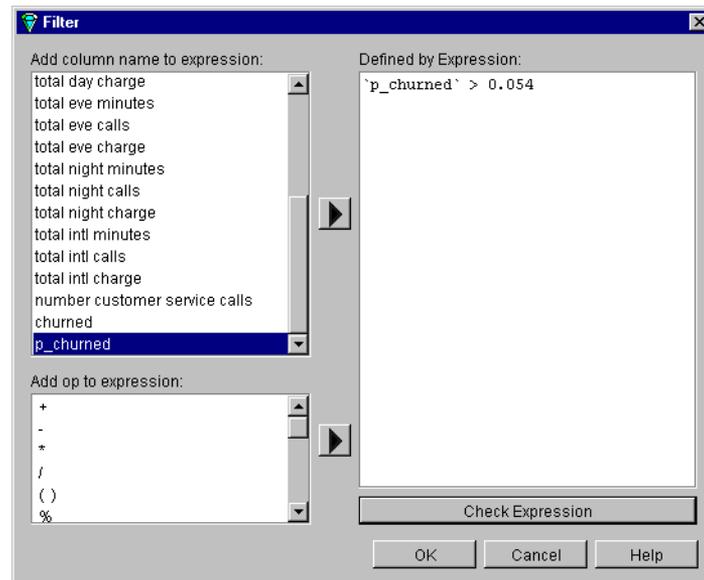
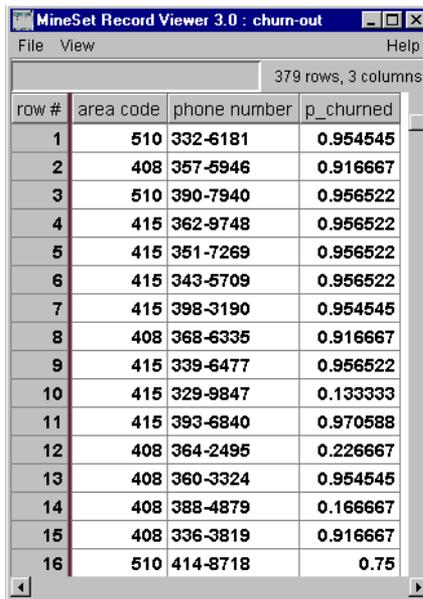


Figure 4-11 Filtering for the Probability of Churn

Finally, you can view the results in Record Viewer, eliminating unnecessary columns for easier reference, as detailed next.

10. In the Data Destinations pane of Tool Manager, click the Viz Tools tab, then click the Records tab. In the Data Transformations pane, select all columns except area code, phone number, and p_churned, then click the *Remove Column* button. Select multiple columns by pressing the Shift key for a range, or the Ctrl key for specific selections.
11. Click Invoke Tool.

The result is a useful phone list, shown in [Figure 4-12](#), of those customers who have the greatest likelihood of churning based on the model.



The screenshot shows a window titled "MineSet Record Viewer 3.0 : churn-out". The window has a menu bar with "File", "View", and "Help". Below the menu bar, it indicates "379 rows, 3 columns". The main area contains a table with the following data:

row #	area code	phone number	p_churned
1	510	332-6181	0.954545
2	408	357-5946	0.916667
3	510	390-7940	0.956522
4	415	362-9748	0.956522
5	415	351-7269	0.956522
6	415	343-5709	0.956522
7	415	398-3190	0.954545
8	408	368-6335	0.916667
9	415	339-6477	0.956522
10	415	329-9847	0.133333
11	415	393-6840	0.970588
12	408	364-2495	0.226667
13	408	360-3324	0.954545
14	408	388-4879	0.166667
15	408	336-3819	0.916667
16	510	414-8718	0.75

Figure 4-12 Record Viewer Results

In Record Viewer, for every record there is a number estimating the probability that the customer will churn. Filtering has retained those customers with the highest numbers. That provides the list of only those potential churn customers to whom you should send incentives (for example, solicit by phone, send mail, and so forth). Dismiss the Record Viewer, and continue exploring the churn dataset.

Reducing Misclassification Costs

You can reduce the cost of making mistakes in building the model using three important tools in MineSet: confusion matrix to give a detailed picture of errors and incorrect predictions, loss matrix to take into account that some mistakes are worse than others, and return-on-investment curve to show when investing more time or money is fruitless.

Displaying a Confusion Matrix

Return to the Tool Manager window, and reopen *churn.schema*.

1. In the Data Destinations pane, click the Mining Tools tab; then click the Classify tab and make the following selections from the pulldown menus:

Mode: Classifier & Error

Inducer: Decision Tree

Discrete Label: churned

2. Click *Advanced options* and the Classifier options pane shown in [Figure 4-13](#) appears.

In the process a message may appear that the attribute “phone_number” is being removed because it had more than 100 distinct values. Click OK and proceed.

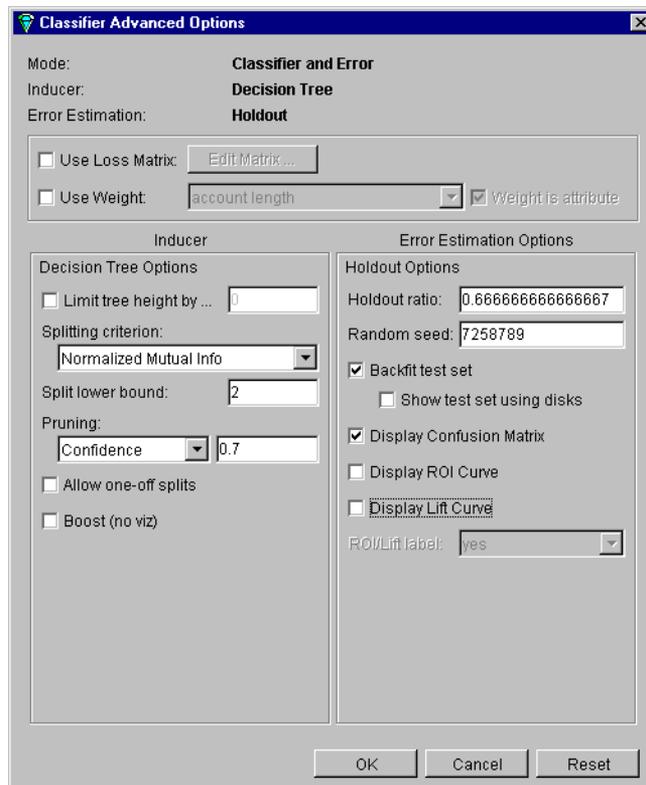


Figure 4-13 Classifier Advanced Options Panel

3. Turn on *Display Confusion Matrix* and *Backfit test set*. Make sure *Display Lift Curve* and *Display ROI Curve* are turned off, then click *OK*.
4. In the Classify pane of the Tool Manager Data Destination pane click *Go*.

The Confusion Matrix displays where the classifier makes mistakes in classifying. Dismiss the Tree Visualizer and examine the Confusion Matrix. From this, you can construct a Loss Matrix based on what you now know about the data, to make some errors less tolerable than others.

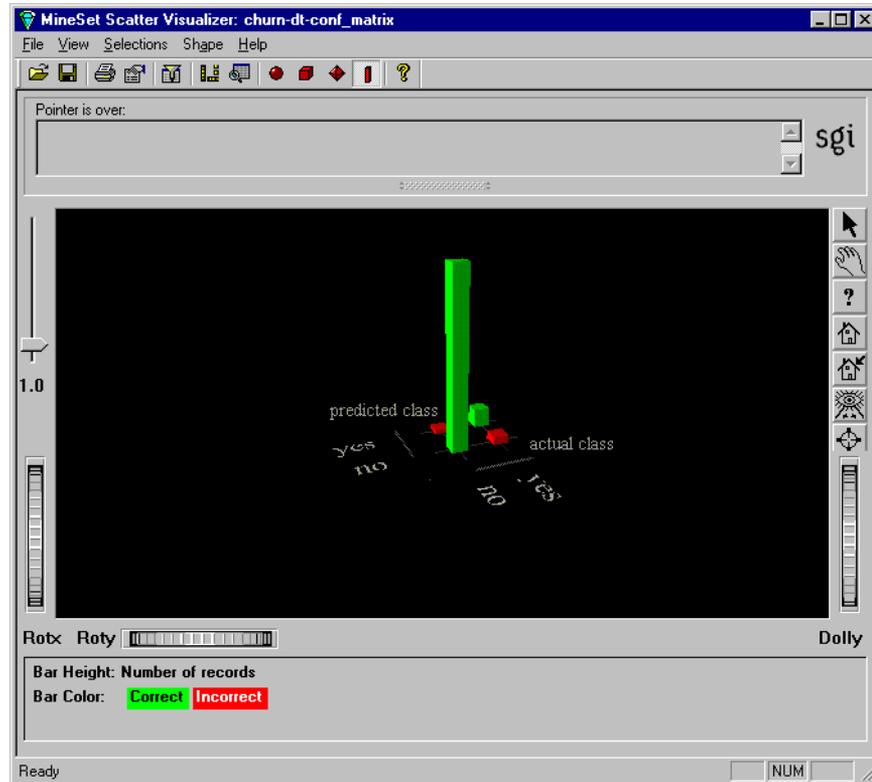


Figure 4-14 Confusion Matrix Showing Correct and Incorrect Classifications

In the window shown in [Figure 4-14](#), the two green bars (one tall and one short) represent correct classifications. The two red bars represent misclassifications. Substantial misclassification occurs in the category represented by the larger of the two red bars — “predicted class: no, actual class: yes.” These customers were predicted not to churn, but actually did so, a costly mistake even at 4.7% (to see the percentages, choose Selections > Show values). You can try to reduce that error by constructing a Loss Matrix based on what you now know about the data, and weight the errors represented by this particular red bar more heavily.

Defining a Loss Matrix

The purpose of constructing a Loss Matrix is to control which errors the classifier will favor and which it will avoid.

1. Dismiss the Confusion Matrix display with File > Exit and return to the Tool Manager window.
2. Click *Advanced options* to return to the Classifier options pane.
3. In the Classifier Advanced Options dialog box, turn on *Use Loss Matrix*.
4. Click *Edit Matrix* to weight the cost of making errors. A Loss Matrix pane similar to that shown in [Figure 4-15](#) appears.

		Predicted Values		
		?	no	yes
Actual Values	no	10	0	3
	yes	10	10	-10

Figure 4-15 Loss Matrix Showing Weighting

5. Set the following values across the rows of the Loss Matrix, reading from left to right:

Actual Values: no: 10—0—3

Actual Values: yes: 10—10—(-10)

The value in the column under the question mark should be somewhat high, to prevent the classifier from predicting “unknown.”

Using these values, if you predict a customer will not churn, and you are correct, you neither win nor lose (represented by zero). If you predict a customer will not churn (and therefore fail to send them any incentives), and they do churn, you incur a loss of 10 (represented by positive 10, since the numbers represent loss). If you incorrectly predict a customer will churn, and they do not, you lose three, representing the cost of sending a mailing unnecessarily. If your mailing program works, and you retain some of the customers who would have churned, you gain 10 (represented by minus 10). Your next step is to investigate the return on your investment.

Viewing a Return on Investment Curve

The Return on Investment curve lets you see the cost of making certain kinds of errors, and indicates to you the point at which it is no longer fruitful to continue taking action

1. Make sure *Backfit test set*, *Display Confusion Matrix*, and *Use Loss Matrix* are turned on in the Classifier Advanced options pane.
2. Ensure *Display ROI Curve* also is checked on.
3. Ensure *ROI/Lift label* is set to yes and click *OK*.
4. Click *Go* in the Classify pane of the Tool Manager window.

Three display windows appear; the Decision Tree and Confusion Matrix, and the ROI Curve. The Confusion Matrix shows the classifier is more conservative in making churn=no predictions, thus reducing false negatives. The errors on one side have been increased, but those on the other have been decreased. Dismiss both the Confusion Matrix and the Decision Tree display, and examine the ROI curve window.

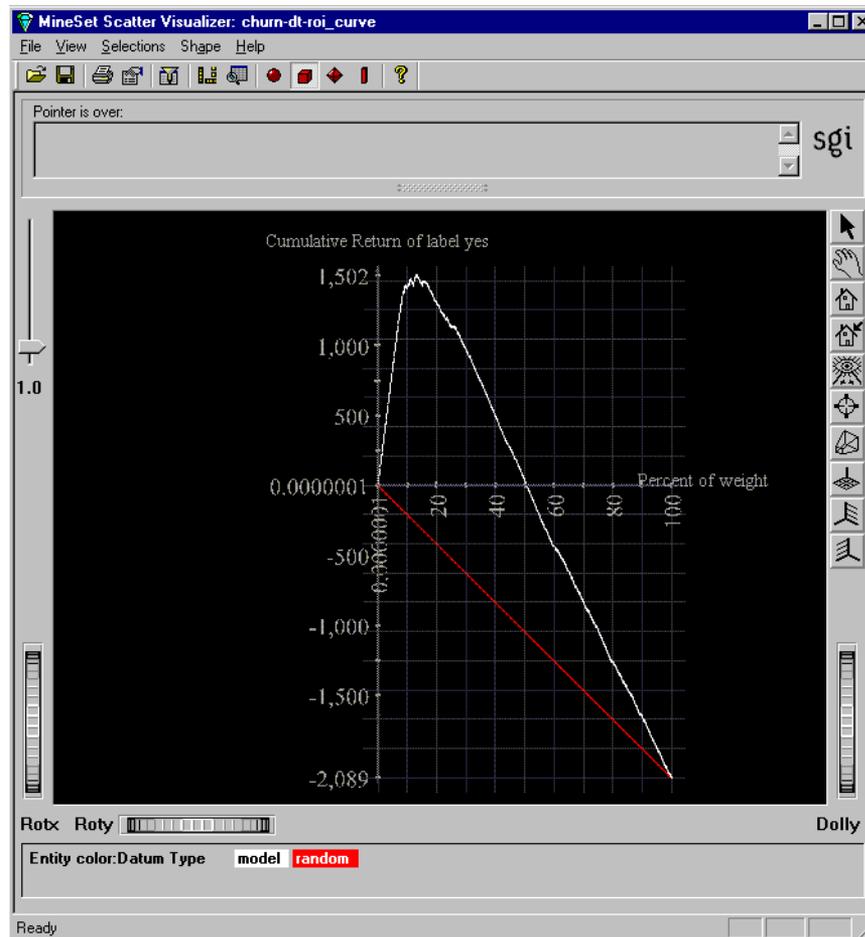


Figure 4-16 Return on Investment Curve

The ROI curve shown in [Figure 4-16](#) bears a marked resemblance to a lift curve. The horizontal line across the middle represents zero profit and loss. The red line represents the expected performance if you were to take a random sample of the population and send them mail. You expect a loss if you mail to everyone, because of the cost of mailing. However, there is a point of optimum return on investment, represented by the knee of the curve, at 1488 or 12.6% of the population.

Further Exploration of MineSet

See the *MineSet 3.0 Enterprise Edition User's Guide*, the *MineSet 3.0 Enterprise Edition Reference Guide*, and the *MineSet 3.0 Enterprise Edition Interface Guide* for descriptions of these tools and what the analytical data mining algorithms can show. The manual is online and can be launched by selecting Help > MineSet User's Guide.

This tutorial has only been a brief introduction to the MineSet tool suite. Other aspects covered in the *MineSet User's Guide* include:

- Scatter Visualizer.
- Tree Visualizer for visualizing hierarchies.
- Option Tree Inducer and Classifier.
- Association Rules Generator and Visualizer.
- Regression, to allow you to predict a continuous value instead of discrete.
- Transformations, including binning, distribution, and indexing of arrays.
- Record weighting, which allows assigning different weights to different records, because some records are more important than others (for example, highly profitable customers).
- Learning Curve, which can help you determine whether sampling can be done on your dataset to speed up the knowledge discovery process, without losing much of the accuracy of the induced classifiers.
- Many tool options, including color manipulation, message boxes.
- Animation sliders for visual tools.
- Batch processing. The program `mineset_batch` can be used to execute operations non-interactively. This is useful if a job needs to run regularly (for example, once a night).
- Error estimation using advanced techniques such as cross-validation.

Also described in the *MineSet User's Guide* are the technical details of file and data manipulation.

Note: Data mining algorithms find correlations that may not be causal. A well-known discovery is the strong correlation between shoe size and reading ability: the larger one's shoe size, the better the reading ability. This correlation, while true, is not causal; both shoe size and reading ability improve with age (as children get older, their shoe size and ability to read both increase.) You are cautioned against attributing causality to discovered correlations. Wearing larger shoes is unlikely to increase your reading ability.

Navigating in the MineSet Visualizers

Navigating in the Tree Visualizers

The Tree Visualizer display is best thought of as though you are viewing the scene through a camera. To change the view, you change the position of the camera (the viewpoint). This section consists of two tables that serve as a quick reference for the Tree, Decision Tree, Option Tree, and Regression Tree Visualizer controls. [Table A-1](#) describes the navigation buttons.

Table A-1 Navigation Icons in the Tree Visualizers

icon	Action
	Returns the chart to the size and position designated as the home view. By default this is the size and position of the chart when the visualizer is first invoked. You can change the home position by using the next icon.
	Sets a new home view for the chart. Use this to save a certain view or position.
	Moves the chart to a position where it is centered and all of it is visible in the window.
	Undoes the previous move (like the Back button on a Web browser).
	Redoes a move that has been undone (like the Forward button on a Web browser).
	Moves one node closer to root of the tree.
	Moves one node or bar to the left.
	Moves one node or bar to the right.
	Moves one node down the tree on the left path.
	Moves one node down the tree on the right path.
	Pops up a menu of possible paths from the current node.

[Table A-2](#) lists several manipulations you can perform on the scene in the tree visualizer. Most of these manipulations can be done either with one of the controls on the visualizer window or with a mouse action.

Table A-2 Manipulating the Tree Visualizer Scene

Action	Slider or Wheel	Mouse Equivalent
Fly over surface of the scene	N/A	Hold down the left and right mouse buttons (or middle mouse button) and move the mouse.
Raise or lower bar heights to emphasize differences	Height slider (upper left)	N/A
Move viewpoint up and down	H wheel	Hold down the right mouse button, and move the mouse up and down.
Move the viewpoint from side to side	Side to side wheel (<-->)	Hold down the left and right mouse buttons (or middle mouse button) and move the mouse side to side.
Move the viewpoint backwards and forwards	Dolly wheel	Hold down the left and right mouse buttons (or middle mouse button) and move the mouse up and down.
Change the up and down tilt of the camera	Tilt wheel	N/A
Move forward in the direction you are pointing	N/A	Hold down Alt key and left and right mouse buttons (or middle mouse button) and move mouse. While moving forward, the viewpoint also moves down, based on the current tilt. Similarly, while moving backward, the viewpoint moves up, based on the tilt.
Select a child of a node	N/A	Hold down the Ctrl key and click the right mouse button on parent node, then click on the child to move there (or use the branching navigation icon).

Navigating in Non-Tree Visualizers

This section consists of two tables that serve as a quick reference for the Evidence, Decision Table, Map, Scatter, and Splat Visualizer navigation controls. [Table A-3](#) describes the navigation buttons.

Table A-3 Navigation Buttons in Non-Tree Visualizers

Button	Name	Action
	Pick	Changes the program to pick mode (an arrow). In pick mode, you can highlight (brush over) or select (click) elements of the chart.
	Grasp	Changes the program to grasp mode (a hand). In grasp mode, you can move the chart around in the window: <ul style="list-style-type: none"> — To move chart in window, hold down right mouse button and move mouse. — To rotate chart, hold down left mouse button and move mouse. — To dolly the chart in and out, hold down left and right mouse buttons (or use middle mouse button) and move mouse.
	Home	Returns the chart to the size and position designated as the home view. By default this is the size and position of the chart when the visualizer is first invoked. You can change the home position by using the set home icon.
	Set home	Sets a new home view for the chart. Use this when you want to save a certain view or position.
	View All	Moves the chart to a position where it is centered and all of it is visible in the window.
	Zoom	Moves the point you select to the middle of the pane and zooms to it. When the mouse cursor becomes a targeting sight, move it to the spot you want to see more clearly, then click the left mouse button.
	3D	Toggles the 3D perspective.
	Top View	Changes the chart to a top view (Scatter and Splat Visualizers only).
	Front View	Changes the chart to a front view (Scatter and Splat Visualizers only).
	Side View	Changes the chart to a side view (Scatter and Splat Visualizers only).

[Table A-4](#) describes the adjustment sliders and wheels in the non-tree visualizers.

Table A-4 Manipulating Non-Tree Visualizer Scene

Action	Slider or Wheel	Mouse or Keyboard Action
Toggle between Select and Grasp mode	N/A	Press the Esc key or navigation buttons.
Move scene	N/A	Click and hold the right mouse button. Move the cursor in the direction you want to move the chart.
Raise or lower cake, pie, or bar heights to emphasize differences	Height slider (upper left)	N/A
Rotate scene around X axis	Rotx wheel	Click and hold the left mouse button. Move the cursor in the direction you want to rotate the chart.
Rotate scene around Y axis	Roty wheel	Click and hold the left mouse button. Move cursor in the direction you want to rotate the chart.
Zoom scene in and out	Dolly wheel	Click and hold the left and right mouse buttons (or middle mouse button). Move the mouse down to zoom in and up to zoom out.
Filter out less important attributes	Detail slider (Evidence and Decision Table Visualizers only)	N/A
Filter out attribute values with record weights less than a specified percentage of the total weight of records in the dataset, up to 2%	% Weight Threshold slider (Evidence and Decision Table Visualizers only)	N/A

Table A-4 (continued) Manipulating Non-Tree Visualizer Scene

Action	Slider or Wheel	Mouse or Keyboard Action
Drill down through levels of detail (Decision Table and Map Visualizers only)	N/A	Put the mouse arrow over a specific chart (or the background for all charts) and click the right mouse button.
Drill up through levels of detail (Decision Table and Map Visualizers only)	N/A	Put the mouse arrow over a specific chart (or the background for all charts) and Ctrl-click the right mouse button (or click the middle mouse button).

