Finding the Solution to Data Mining:

A Map of the Features and Components of SAS® Enterprise Miner™ Software

John Brocklebank, SAS Institute Inc., Cary, NC
Mark Brown, SAS Institute Inc., Cary, NC

Abstract

Enterprise Miner™ software is the data mining solution from SAS Institute Inc. This paper discusses three main features of Enterprise Miner—the graphical user interface (GUI), the SEMMA methodology, and client/server enablement—and maps the components of the solution to those features.

Introduction

Data mining is a process; not just a series of statistical analyses. Simply applying disparate software tools to a data mining project can take one only so far. Instead, what is needed to plan, implement, and successfully refine a data mining project is an integrated software solution—one that encompasses all steps of the process beginning with the sampling of data, through sophisticated data analyses and modeling, to the dissemination of the resulting business-critical information. In addition, the ideal solution should be intuitive and flexible enough that users with different degrees of statistical expertise can understand and use it.

To accomplish all this, the data mining solution must provide

- advanced, yet easy-to-use, statistical analyses and reporting techniques
- a guiding, yet flexible, methodology
- client/server enablement.

SAS® Enterprise Miner™ software is that solution. It synthesizes the world-renowned statistical analysis and reporting system of SAS Institute with an easy-to-use GUI that can be understood and used by business analysts as well as quantitative experts.

The components of the GUI can be used to implement a data mining methodology developed by the Institute. However, the methodology does not dictate the steps to be taken in projects. Instead, the methodology enables users to move data mining projects from raw data to information by going step-by-step as needed from sampling of data, through exploration and modification, to modeling, assessment and scoring of new data, and then to the dissemination of the results.

The GUI also contains components that help administrators set up and maintain the client/server deployment of Enterprise Miner software.

In addition, as a software solution from SAS Institute, Enterprise Miner fully integrates with the rest of the SAS® System including the award-winning SAS/Warehouse Administrator™ software, the SAS solution for online analytical process (OLAP), and SAS/IntrNet™ software, which enables applications deployment via intranets and the World Wide Web.

The Graphical User Interface

Enterprise Miner employs a single, graphical user interface (GUI) to give users all the functionality needed to uncover valuable information hidden in their volumes of data. With one point-and-click interface, users can perform the entire data mining process from inputting data from diverse sources, through preparing the data for modeling and accessing the value of the models, to scoring models for use in making business decisions.

The GUI is designed with two groups of users in mind: business analysts, who may have minimal statistical expertise, can quickly and easily navigate through the data mining process; and quantitative experts, who may want to go explore the details, can access and fine tune the underlying analytical processes. The GUI uses familiar desktop objects such as tool bars, menus, windows, and dialog pages to equip both groups with a full range of data mining tools.

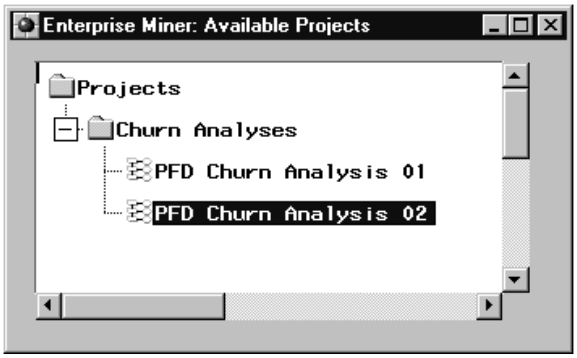
The main components of the GUI are the

- Available Projects window
- Enterprise Miner Workspace window
- Node Types window
- Message window.

Available Projects Window

The Available Projects window opens when you start Enterprise Miner.

Figure 1: Projects Displayed in the Available Projects Window

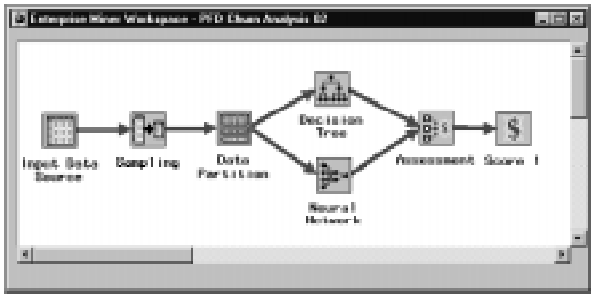


The window displays any existing projects in a familiar hierarchical form. Along with the tool bar and menus, the Available Projects window enables you to create and manage data mining projects.

Enterprise Miner Workspace Window

When you open an existing project or create a new one by using the Available Projects window, the Enterprise Miner Workspace window is displayed. Using the Enterprise Miner Workspace window, the Node Types window, the tool bar, and menus, you can build, edit, and run process flow diagrams (PFDs).

Figure 2: A PFD Displayed in the Enterprise Miner Workspace Window



With these easy-to-use tools, you can map out your entire data mining project, launch individual functions, and modify PFDs simply by pointing and clicking.

Node Types Window

When you open a project, the Node Types window appears along with the Enterprise Miner Workspace window.

Figure 3: Node Types Window

The Node Types window functions as a palette, which displays the data mining nodes that are available for constructing PFDs. When you place the cursor over a node icon, the name of the node appears in a pop-up field. Using the mouse, you can drag and drop nodes from the Node Types window onto the Enterprise Miner Workspace window and connect the nodes in the desired process flow.

Utility Nodes

In addition to nodes that perform specific data mining steps, such as sampling or data partitioning, Enterprise Miner includes the following utility nodes:

- The **SAS Code node** enables you to submit SAS software programming statements.
- The **Control Point node** enables you to establish a control point in process flow diagrams. A Control Point node can be used to reduce the number of connections that are made.
- The **Sub-diagram node** enables you to group a portion of a PFD into sub-units, which are themselves diagrams. For complex PFDs, you may want to create sub-diagrams to display various levels of detail.
- The **Data Mining Database node** enables you to create a data mining database (DMDB) for batch processing. For non-batch processing, DMDBs are automatically created as they are needed.

Common Features Among Nodes

The nodes of Enterprise Miner software have a uniform look and feel. For example, the tabbed dialog pages enable you to quickly access the appropriate options; the common node functionality enables you to learn the usage of the nodes quickly; and the Results Browser, which is available in many of the nodes, enables you to view the results of running the process flow diagrams.

Message Window

The Message window displays messages generated by the creation or execution of a PFD. It is hidden by default, and it can be toggled on and off using the View pull-down menu.

The SEMMA Methodology

One of the keys to the effectiveness of Enterprise Miner is the fact that the GUI makes it easy for business analysts as well as quantitative experts to organize data mining projects into a logical framework. The visible representation of this framework is a PFD. It graphically illustrates the steps taken to complete an individual data mining project.

In addition, a larger, more general, framework for staging data mining projects exists. This larger framework for data mining is the SEMMA methodology as defined by SAS Institute. SEMMA is simply an acronym for “Sample, Explore, Modify, Model, and Assess.” However, this logical superstructure provides users with a comprehensive method in which individual data mining projects can be developed and maintained. Not all data mining projects will need to follow each step of the SEMMA methodology—the tools in Enterprise Miner software give users the freedom to deviate from the process to meet their needs—but the methodology does give users a scientific, structured way of conceptualizing, creating, and evaluating data mining projects.

You can use the nodes of Enterprise Miner to follow the steps of the SEMMA methodology; the methodology is a convenient and productive superstructure in which to view the nodes logically and graphically. In addition, part of the strength of Enterprise Miner software comes from the fact that the relationship between the nodes and the methodology is flexible.

The relationship is flexible in that you can build PFDs to fit particular data mining requirements. You are not constrained by the GUI or the SEMMA methodology. For example, in many data mining projects, you may want to repeat parts of SEMMA by exploring data and plotting data at several points in the process. For other projects, you may want to fit models, assess those models, re-fit new data to the models, and then re-assess.

Sampling Nodes

In data-intensive applications such as data mining, using a sample of data as input rather than an entire database can greatly reduce the amount of time required for processing. If you can ensure the sample data are sufficiently representative of the whole, patterns that appear in the entire database also will be present in the sample. Although Enterprise Miner does not require the use of sample data, using the GUI and the sampling nodes, you can complete sophisticated data mining projects with a minimum amount of time and effort.

Warehouse Data

For obtaining samples of data from data warehouses and other types of data stores, Enterprise Miner provides a unique advantage—complete connectivity to the Institute’s award-winning SAS/Warehouse Administrator™ software.

Input Data Source Node

The Input Data Source node enables you to access and query SAS data sets and other types of data structures that will be used for data mining projects. You can use more than one Input Data Source node in a single project to define multiple sources of input.

The node includes a set of dialog pages and other windows that enable you to specify the details about the data source and the variables in the data source.

Dialog Pages

The interface to the Input Data Source node is a dialog box that includes the following pages:

Data

When you open an Input Data Source node, a Data dialog page is displayed in which you can enter the name of the input data set. After entering the name, you press the return key to begin pre-processing the data source. Pre-processing includes the automatic generation of meta data, such as model roles and summary statistics.

By default, the Input Data Source node uses a sample of 2000 observations to determine the meta data. Samples are stored on the local client when you run a remote project. They are used for several tasks, such

as viewing a variable's distribution in a histogram, determining variable hierarchies in the Variable Selection node, and as input to the Insight node.

You can control the size of the meta data sample as well as the purpose of the input data source. By default, the purpose of the sample is for use as raw data. Alternatively, you can define the purpose of the data source to be training, validating, testing, or scoring.

The meta data sample is not intended for use beyond the Input Data Source node. To obtain a sample to be used for modeling or other data mining analysis, you should use the Sampling node, which is specifically designed to give you many options for sampling data.

SQL Query Window

You can also use the SQL Query Window to define a data source. By selecting the Query button, you open the SQL Query window, which enables you to build a SAS View in a point-and-click environment.

After the SQL view is completed, Enterprise Miner pre-processes the view for faster processing in subsequent mining activities.

When data pre-processing is completed, the Data dialog page re-opens and displays meta data including the name of the view, its description, its role or purpose in the project, and measurements of size. This automatic assignment of variable roles and measurement levels by Enterprise Miner saves you from the tedious task of defining this information for each variable.

Variables

The Variables dialog page contains a data table, which enables you to redefine model roles, measurement levels, formats, and labels for the variables in the input data. These settings are determined initially by Enterprise Miner during preprocessing of the data source. The settings are based on the meta data sample. Changing the role of a model or the measurement level for a variable is as easy as placing the cursor in the desired cell and single clicking the right mouse button. This action opens a pop-up menu from which you can make changes.

Other features of the Variables dialog page enable you to

- specify formats and labels for the variables by typing in the values directly
- sort in ascending order or subset (if relevant) that column by its values
- display a histogram showing the distribution of any variable.

Interval Variables

The Interval Variables dialog page displays a sortable table of summary statistics for all interval variables from the input data set including minimum and maximum values, the mean, standard deviation, percentage of observations with missing values, and the skewness and kurtosis.

Class Variables

The Class Variables dialog page displays a sortable table of information about all non-interval level variables from the input data set including the name of the variable, the number of unique values, the order in which the values are sorted in the data set, and the names of other variables from the input data set on which the variable depends. The Class Variables dialog page is where you set the target level for class variables.

Sampling Node

The Sampling node enables you to extract a sample of your input data source. Sampling is recommended for extremely large data bases, because it can tremendously decrease model fitting time. The Sampling node performs simple random sampling, *n*th-observation sampling, stratified sampling, or first-*n* sampling of an input data set. For any type of sampling, you can specify either a number of observations or a percentage of the population to select for the sample. The Sampling node writes the sampled observations to an output data set. The Sampling Node saves the seed values used to generate the random numbers for the samples so that you may replicate the samples.

The Sampling node must be preceded by a node that exports at least one raw data table, which is usually an Input Data Source node.

To partition the sample into training, validation, and test data sets, follow the Sampling node with a Data Partition node. Exploratory and modeling nodes also can follow a sampling node.

After you run the Sampling node, you can use the Actions pull-down menu to browse the results. The Sampling Results window is a tabbed dialog interface that provides easy access to a table view of the sample data set, the Output window, the SAS Log, and the Notes window.

Dialog Pages

The interface to the Sampling node is a dialog box that includes the following pages:

General

In the General dialog page, you specify the sampling methods, the sampling size, and the random seed.

Sampling Methods – The Sampling node supports simple random sampling (the default), sampling every *n*th observation, stratified sampling, and sampling the first *n* observations.

- **Every *n*th Observation** – With *n*th observation sampling (also called systematic sampling), the random sample node computes the percentage of the population that is required for the sample, or uses the percentage specified in the General tab.
- **Stratified Sampling** – In stratified sampling, you specify categorical variables from the input data set to form strata (or subsets) of the total population. Within each stratum, all observations have an equal probability of being selected for the sample. Across all strata, however, the observations in the input data set generally do not have equal probabilities of being selected for the sample. You perform stratified sampling to preserve the strata proportions of the population within the sample. This may improve the classification precision of fitted models.
- **First *n* Observations** – With first *n* sampling, the Sampling node selects the first *n* observations from the input data set for the sample. You can specify either a percentage or an absolute number of observations to sample in the General tab.
- **Cluster** – Cluster specifies that the sample data is to be based on a cluster variable. For values of the cluster variable that are selected by the Sampling node, all records associated with the selected cluster are included in the sample. Selecting cluster sampling on the General dialog page enables (ungrays) the Cluster dialog page, which is where you specify the cluster variable, the cluster sampling method, and the number of clusters.

Sample Size – You can specify sample size as a percentage of the total population, or as an absolute number of observations to be sampled.

Random Seed – The Sampling node displays the seed value used in the random number function for each sample. The default seed value is a random number generated from a call to the system clock. You may type in a new seed directly, or select the Generate New Seed button to have the Sampling node generate a new seed automatically. The Sampling node saves the seed value used for each sample so that you may replicate samples exactly.

Stratification Variables

The Stratification dialog page contains two sub-pages that enable you to control the variables and options used for stratification.

Variables – The Variables sub-page contains a data table that lists the variables that are appropriate for use as stratification variables. Stratification variables must be categorical (binary, ordinal, or nominal). Continuous variables have too many levels with too few observations per level to form useful strata. For example, a variable with two levels of gender or a variable with five ordinal categories of income are appropriate stratification variables. A variable with interval-level measurements of income is not an appropriate stratification variable.

Options – In the Options sub-page you may specify the stratification criteria, a deviation variable, and the minimum stratum size for each stratum.

Data Partition Node

Most data mining projects have large volumes of data that can be sampled with the Sampling node. After sampling, the data can be partitioned with the Data Partition node before you begin constructing data models. The Data Partition node enables you to partition the input data source or sample into data sets for the following purposes:

- **Training** – used to fit initial models.

- **Validation** – used by default for model assessment. A validation data set also is used for fine tuning the model. The Decision Tree and Neural Network nodes have the capacity of over-fitting training data sets. To prevent these nodes from over fitting, validation data sets are automatically used to retreat to a simpler fit than the fit based on the training data alone. Validation data sets also can be used by the Regression node as stepwise selection criterion.
- **Testing –** an additional data set type that can be used for model assessment.

Partitioning provides a mutually exclusive data set(s) for cross validation and model assessment. A mutually exclusive data set does not share common observations with another data set. Partitioning the input data also helps to speed preliminary model development.

Simple or stratified random sampling is performed to partition the input observations into training, validation, and test data sets. You specify the proportion of sampled observations to write to each partitioned data set. The Data Partition node saves the seed values used to generate the random numbers for the samples so that you can replicate the data sets.

Dialog Pages

The interface to the Data Partition node is a dialog box that includes the following pages:

Partition

In the Partition page, you can specify the

- sampling method used to write observations to the partition data sets
- random seed used to generate the random sample
- percentage of observations to allocate to the training, validation and test data sets.

Sampling Methods

The Partition node supports simple random sample, which is the default, and stratified sampling.

- **Simple Random Sampling** – In simple random sampling, every observation in the data set has the same probability of being written to the sample. For example, if you specify that 40 percent of the population should be selected for the training data set, then each observation in the input data set has a 40 percent chance to be selected.
- **Stratified Sampling** – In stratified sampling, you specify variables from the input data set to form strata (or subsets) of the total population. Within each stratum, all observations have an equal probability of being selected for the sample. Across all strata, however, the observations in the input data set generally do not have equal probabilities of being selected for the sample. You perform stratified sampling to preserve the strata proportions of the population within the sample. This may improve the classification precision of fitted models.

Exploration and Modification Nodes

Data mining is a dynamic, iterative process through which you can gain insights at various stages of a project. One perspective on a problem can lead to another and to the need for further modification and exploration. Enterprise Miner gives you numerous tools and techniques to help you explore and modify your data including:

- **Graphical Displays** from simple graphs to multidimensional bar charts,
- **Outlier Filters** to stabilize parameter estimates,
- **Transformations** that enable you to normalize or linearize the data and to stabilize the variance
- **Advanced Visualization Techniques** including OLAP, which enable you to interact with multidimensional graphical displays to evaluate issues such as data structure and the applicability of variables.

Associations Node

The Associations node enables you to perform either association or sequence discovery. Association discovery enables you to identify items that occur together in a given event or record. The technique is also known as *market basket analysis*. Rules of form are discovered based on frequency counts of the number of times items occur alone and in combination in a database. The rules can be expressed in statements such as "if item A is part of an event, then item B is also part of the event X percent of the time."

Such rules should not be interpreted as a direct causation, but as an association between two or more items. However, identifying credible associations can help the business technologist make business

decisions such as when to distribute coupons, when to put a product on sale, or how to lay out items in a store.

Dialog Pages

The interface to the Associations node is a dialog box that includes the following pages:

General

The General dialog page enables you to set the following rule-forming parameters:

- **Minimum Transaction Frequency** – a measure of support that indicates the percentage of times the items occur together in the data base.
- **Maximum Number of Items in an Association** – determines the maximum size of the item set to be considered. For example, the default of 4 items indicates that you wish to examine up to 4-way associations.
- **Minimum Confidence** – specifies the minimum confidence level to generate a rule. The default level is 10 percent.

Sequences

The Sequences page enables you to set parameters that are used to set sequence rules. For example, you may want to set a minimum support for sequence to filter out items from the sequence analysis that occur too infrequently in the data base.

Advanced

The Advanced page enables you to set the following rule-forming parameters:

- **Calculate Maximum Number of Associations Using 2**_n_** – This option determines the maximum number of associations that are computed.
- **Customize Sort** – By default, if the Association node creates 100,000 rules or less, then it sorts the support values in descending (highest to lowest) order within each relation in the Associations results table. If there is more than 100,000 rules, then the sort routine is not executed; the node runs faster but the support values are listed in ascending order within each relation when you view the results in the browser.

Sequence Discovery

The Association node also enables you to perform sequence discovery. Sequence discovery goes a step further than association discovery by taking into account the ordering of the relationships (time sequence) among items. For example, sequence discovery rules may indicate relationships such as, “Of those customers who currently hold an equity index fund in their portfolio, 15 percent of them will open an international fund in the next year,” or, “Of those customers who purchase a new computer, 25 percent of them will purchase a laser printer in the next month.”

3-Dimensional Histogram

You can view a grid plot of the left side and right side items simply by selecting the "Graph" item from the View pull-down menu. The support for each rule determines the size of the square in the plot. The confidence level establishes the color of the square. A confidence density legend is annotated in the bottom left of the graph.

Subsetting

You can subset the table to see a more specific set of rules. For example, you can subset only those rules having a lift greater than one.

Bar Chart Node

The Bar Chart node is an advanced visualization tool that enables you to explore large volumes of data graphically. You can use the node to uncover patterns and trends, and to reveal extreme values in the data base. You can generate multi-dimensional histograms for discrete or continuous variables. The node is fully interactive—you can rotate a chart to different angles and move it anywhere on the screen. You can also probe the data by positioning the cursor over a particular bar within the chart. A text window displays the values that correspond to that bar.

Interactive Tools

The Bar Chart node includes easy-to-use input fields, buttons, and sliders that you use to interact with the graphical display. Pull-down menus and a Histograms Toolbar provide additional functionality to explore the data.

Clustering Node

The Clustering node performs observation clustering, which can be used to segment databases. Clustering

places objects into groups or clusters suggested by the data. The objects in each cluster tend to be similar to each other in some sense, and objects in different clusters tend to be dissimilar. If obvious clusters or groupings could be developed prior to the analysis, then the clustering analysis could be performed by simply sorting the data.

The clustering methods perform disjoint cluster analysis on the basis of Euclidean distances computed from one or more quantitative variables and seeds that are generated and updated by the algorithm. You can specify the clustering criterion that is used to measure the distance between data observations and seeds. The observations are divided into clusters such that every observation belongs to at most one cluster.

After clustering is performed, the characteristics of the clusters can be examined graphically using the results browser. Often of interest is the consistency of clusters across variables. The three-dimensional charts and plots enable you to graphically compare the clusters.

Lastly, the cluster identifier for each observation can be passed to other nodes for use as an input, id, group, or target variable. For example, you could form clusters based on different age groups you want to target. Then you could build predictive models for each age group by passing the cluster variable as a group variable to a modeling node.

Dialog Pages

The interface to the Clustering node is a dialog box that includes the following pages:

Clusters

In the Clusters dialog page, you specify options for the maximum number of clusters. The terms *segment* and *profile segment* both refer to a cluster of observations; therefore, for practical purposes the terms *segment* and *cluster* are equivalent.

Segment Identifier – The segment identifier consists of the variable name, its label, and its role in the model.
Maximum Number of Clusters – The default maximum number of clusters is 10. You may want to perform cluster analysis using different values for the maximum number of clusters. A preliminary cluster analysis may identify outlying observations. Severe outlier problems can distort the remaining clusters.

Seeds

The Seeds dialog page consists of three sub-pages.

General – The General sub-page of the Seeds dialog page enables you to specify the clustering criterion. The default clustering criterion is Least Squares (OLS), and clusters are constructed so that the sum of the squared distances of observations to the cluster means is minimized. Other criteria you can specify are as follows:

- Mean Absolute Deviation (Median)
- Modified Ekblom-Newton
- Root-Mean-Square Difference
- Least Squares (fast)
- Least Squares
- Newton
- Midrange.

Initial – The Initial sub-page of the Seeds dialog page specifies how the cluster seeds are to be updated or replaced.

Final – The Final sub-page of the Seeds dialog page controls the stopping criterion for generating cluster seeds.

Missing Values

In the Missing Values dialog page, you specify how data observations containing some missing values are to be handled. Observations that have missing values cannot be used as cluster seeds. You can choose to handle missing values by excluding the observations containing missing values, or by selecting one of the available methods for imputing values for the missing values.

The imputation methods include

- Seed of Nearest Cluster
- Mean of Nearest Cluster
- Conditional Mean
- Multiple Mean
- Multiple Stochastic.

When performing imputations, you can further specify that unequal variances be accounted for, by using the options for Unequal Variance Adjustment.

Output

The Output dialog page consists of the Clustered Data sub-pages.

- **Print Page –** You use the Print sub-page of the Output dialog page to specify the output to be produced. The default output is the “Cluster Statistics.” Optionally, you can specify “No Output,” “Distance Between Cluster Mean,” and **“**Cluster Listing.”
- **Statistics Data Sets –** The Statistics Data Sets sub-page of the Output dialog page lists data sets that contain the cluster statistics and the seed statistics. The cluster statistic data set contains statistics about each cluster.
- **Clustered Data** – The Clustered Data sub-page of the Output dialog page lists the data libraries and output data sets for training, validation, testing, and scoring.

Viewing the Results

After you run the Clustering node, you can view the results by using the Results Browser.

- **Partition Page –** The Partition dialog page provides a graphical representation of key characteristics of the clusters. A three-dimensional pie chart displays the count (slice width), the variability (height), and the central tendency (color). The labels correspond to the segment number. The toolbox enables you to interact with the pie chart.

- **Cluster Distances Page –** The Cluster Distances dialog page provides a graphical representation of the size of each cluster and the relationship among clusters.

- **Cluster Profiles Page –** The Cluster Profiles dialog page provides a graphical representation of the input variables for each cluster.

- **Statistics Page –** The Statistics dialog page displays information about each cluster in a tabular format. To sort the table, single click on the appropriate column heading. To reverse the sort order, single click again on the column heading.

- **Output Page –** The Output dialog page displays the output from running the underlying SAS software programming statements.

Data Replacement Node

Data sources can contain records that have missing values for one or more variables, which can be the result of any number of situations such as data entry errors, incomplete customer responses, or transaction system and measurement failures.

By default, if an observation contains a missing value, then Enterprise Miner will not use that observation for modeling by the Variable Selection, Neural Network, or Regression nodes. As a remedy, you could discard incomplete observations, but that may mean throwing away useful information from the variables that have non-missing values. Discarding incomplete observations also may bias the sample, because observations that have missing values may have other characteristics in common as well. As an alternative to discarding observations, you could use the Decision Tree node to define a decision alternative or surrogate rules to group missing values into a special category. You also could use the Clustering node to impute (fill in) missing values in the Missing Values dialog tab.

Another alternative is to use the Data Replacement node to replace missing values for interval and class variables. The Data Replacement node provides the following interval replacement statistics:

- mean
- median
- midrange.

Using the Data Replacement node, you can impute missing values for class variables with the variables mode or leave the value as missing. You can customize the default replacement statistics by specifying your own replacement values for missing and non-missing data. Missing values for the training, validation, test, and score data sets are replaced using replacement statistics that are calculated from the training predecessor data set or from the meta data sample file of the Input Data Source node.

If you created more than one predecessor data set that has the same model role, then the Data Replacement node automatically chooses one of the data sources. If a valid predecessor data set exists, then you can assign a different data source to a role.

Default Method

You set the default interval and class imputation statistics in the Default Method dialog page. Interval variables contain continuous values, such as AGE and INCOME. Class variables have discrete levels, such as DEPT or ITEM. You also specify whether to calculate the imputation statistic based on the sample or the entire training data set.

The default imputation statistic is used to impute missing values for all variables of the same type (interval or class). You can assign different imputation statistics to different variables, and specify your own imputation criteria in the Interval Variables or Class Variables tabs.

The Default Method dialog page includes radio buttons that enable you to set the default interval imputation statistic as one of the following:

- **Mean** – the arithmetic average, which is calculated as the sum of all values divided by the number of observations. The mean is the most common measure of a variable's central tendency; it is an unbiased estimate of the population mean.
- **Median** – the 50th percentile, which is either the middle value or the arithmetic mean of the two middle values for a set of numbers arranged in ascending order.
- **Midrange** – the average of the range, where the range is the difference between the maximum and minimum values.
- **None** – do not replace the missing values.

The mean may be preferable for data replacement when the variable values have, in general, a normal distribution. The mean and median are equal for a normal distribution. The median is less sensitive to extreme values than is the mean or midrange. Therefore, the median is preferable when you want to replace missing values for variables that have skewed distributions. The median is also useful for ordinal data. The midrange is a rough measure of central tendency that is easy to calculate.

By selecting a radio button, you can specify to replace missing values for class variables with values from one of two variables. The replacement variables are

- **Most Frequent** – replaces missing class values with the variable's mode, which is the value that occurs with the greatest frequency. For example, if MEDIUM is the most common value for SIZE, then all missing values for SIZE are replaced with a value of MEDIUM.
- **None** – missing values are left as missing.

You also can specify whether the data source used to calculate the interval and class imputation statistics is the meta data sample (created when you run the Input Data Source node), or the entire training data set. The meta data sample is used by other nodes to perform many tasks, such as setting variable roles and calculating summary statistics, viewing a variable's distribution in a histogram, and determining variable hierarchies. You may want to use the meta data sample file for data replacement if the entire training data set is very large.

Weights

The Weights dialog page enables you to specify a weight variable that contains relative weights for each observation in the data source. The observation weights are used to calculate a weighted mean. Weight variables are not used to calculate the median or midrange.

Interval Variables

The Interval Variables dialog page enables you to customize the default interval method that you set in the Default Method tab. The Interval Variables tab lists the name, model role, status, imputation statistic, replacement values for non-missing data, format, and label.

The Internal Variables dialog page displays the default imputation statistic. You can assign different imputation statistics to different variables or specify your own numeric replacement values.

You also can use one of these methods to replace non-missing values that are less than (or greater than) a particular value with a new value. These data replacement methods enable you to replace extreme values on either side of the variable's distribution with a more centrally located data value.

Before non-missing values are replaced with a new value, missing values are imputed with the method displayed in the Imputation Method column.

Class Variables

The Class Variables dialog page enables you to customize the default class imputation statistic that is set to either "Most Frequent" or "None" in the Default Method tab. The Class Variables dialog page lists the

name, model role, status, imputation statistic, replacement values, measurement level, type, format, and label. You can assign different imputation statistics to different variables or specify your own replacement value.

Managing Data Replacement Graphs

The tool box at the top of the Enterprise Miner GUI enables you to manage the bar charts and histograms that are displayed in the Select Value window including functions to do the following:

- print the graph
- paste it to the clipboard
- select points on the graph
- display a text box that lists the frequency and variable value for a bar
- move the graph
- zoom in and out.

The File pull-down menu also contains items that are useful for managing the graph including functions for saving the graph as a **bmp**, **gif**, **tif**, or **ps** file; printing the graph to your default printer, and **e-mailing** the graphical image to others.

Data Set Attributes Node

The Data Set Attributes node enables you to modify data set attributes, such as data set names, descriptions, and roles. You also can use this node to modify the meta data sample that is associated with a data set. For example, you could generate a data set in the SAS Code node and then modify its meta data sample with the Data Set Attributes node.

Filter Outliers Node

The Filter Outliers node enables you to apply a filter to your data to exclude observations, such as outliers or other observations that you do not want to include in further data mining analysis. Filtering extreme values from the data tends to produce better models because the parameter estimates are more stable.

Automatic Filter Options Window

The Automatic Filter Options window enables you to

- eliminate rare values for classification variables with fewer than 25 unique values
- eliminate extreme values for interval variables.

Classification Variable with Fewer than 25 Unique Values

This automatic filter option enables you to eliminate rare values of classification variables, such as REGION, that have less than or equal to *n* unique levels. You eliminate rare class values because of the lack of precision in their parameter estimates.

Eliminate Extreme Values for Interval Variables

This automatic filter option enables you to eliminate extreme values for interval variables, such as SALARY, using one of the following methods:

- **Standard Deviations from the Mean –** eliminates values that are more than *n* standard deviations from the mean.
- **Extreme Percentiles –** eliminates values that are in the top and bottom *p*th percentile. For measures arranged in order of magnitude, the bottom *p*th percentile is the value such that *p* percent of the interval measurements are less than or equal to that value.
- **Modal Centroid –** eliminates values more than *n* spacings from the modal center.
- **Median Absolute Deviations (MAD) –** eliminates values that are more than *n* deviations from the median.

Apply Filter Window

The Apply Filter window enables you to examine and adjust the automatic filtering options that apply to class variables and interval variables.

Class Variables

The Class Variables dialog page is a data table that lists the names and labels of all the classification variables in the input data set. This page also lists the minimum frequency cutoff and excluded values. Pop-up menus are provided that enable you to adjust the minimum frequency cutoff value, view histograms of the density of each level of a class variable, and adjust settings interactively.

Interval Variables

The Interval Variables dialog page is a data table that lists the name, label, and range of each interval variable in the input data set. The range for each interval variable is determined by the method you selected in the Automatic Filter Options window.

Observations outside of the range are excluded from the output data set. Pop-up menus and tool boxes enable you to make adjustments to the range of any internal variable and display the results graphically.

Group Processing Node

The Group Processing node enables you to define group variables, such as GENDER, to obtain separate analyses for each level of the grouping variable(s). If you defined more than one target variable in the Input Data Source node, then a separate analysis is also done for each target. You can have one Group Processing node per process flow diagram. By default, group processing occurs automatically when you run a process flow diagram that contains a Group Processing node.

Dialog Pages

The interface to the Group Processing node is a dialog box that includes the following pages:

General

The General dialog page enables you to define whether or not you want to perform group processing when the process flow diagram is run. By default, the node loops through each level of the group variable(s) when you run the process flow diagram. When training preliminary models, you can improve runtime performance by selecting an option to suppress the automatic looping. The General dialog page also displays the number of targets used, number of groups, and total number of loops.

Group Variables

The Group Variables dialog page contains a data table that enables you to specify various characteristics of variables used in group processing such as:

- Whether group variables are used as input or for grouping.
- The levels contained in each group variable. By default, group processing is performed on all levels of the group variable. If a group variable has several levels, you may want to perform group processing on only a few levels.
- The sort sequence of the variables.

Target Variables

The Target Variables dialog page contains a table that enables you to define the target variables that you want to use for group processing.

By default, all variables that you defined as targets in the Input Data Source node are used as targets during group processing (each target is analyzed separately when the process flow diagram is run). Processing only the desired targets reduces group processing time.

Insight Node

The Insight node enables you to interactively explore and analyze your data through multiple graphs and analyses that are linked across multiple windows. For example, you can analyze univariate distributions, investigate multivariate distributions, create scatter and box plots, display mosaic charts, and examine correlations. In addition, you can fit explanatory models using analysis of variance, regression, and the generalized linear model.

Input

If the Insight node follows a node that exports a data set in process flow, then the Insight node can use as input either the meta data sample file or the entire data set. The default is to use the meta data sample file as input. For representative random samples, patterns found in the meta data sample file should generalize to the entire data set. An option is provided that enables you to load the entire data set into the Insight node.

Transform Variables Node

The Transform Variables node enables you to create new variables that are transformations of existing variables in your data. Transformations are useful when you want to improve the fit of a model to the data. For example, transformations can be used to stabilize variances, remove non-linearity, and correct non-normality in variables. You can choose from the following types of transformations:

- log
- square root
- inverse
- square

- exponential
- standardize
- bucket
- quantile.

You also can create your own transformation or modify a transformation by defining a formula in the Computed Column window. The Computed Column window is a graphical interface that contains a column list box, a number pad, an operator pad, and a functions list box to build an expression.

Transform Variables Table

The interface to the Transform Variables node is a table editor in which each row represents an original variable or a transformed variable. The Transform Variables data table enables you to specify interactively the keep status, delete variables, transform variables, modify transformed variables, and change formats and labels. The interface includes columns for the following:

- **Name** - the name of the original or transformed variable.
- **Keep** - whether the variable is to be output to a subsequent node.
- **Mean** - the mean value.
- **Std Dev** - the standard deviation, which is a measure of dispersion about the mean.
- **Skew** - the skewness value, which is measure of the tendency for the distribution of values to be more spread out on one side than the other. Positive skewness indicates that values located to the right of the mean are more spread out than are values located to the left of the mean. Negative skewness indicates the opposite.
- **Kurtosis** - the kurtosis statistic, which is a measure of the shape of the distribution of values. Large values of kurtosis indicates the data contain some values that are very distant from the mean, as compared to most of the other values in the data set.
- **C.V.** - the coefficient of variation is a measure of spread relative to the mean. It is calculated as the standard deviation divided by the mean.
- **Formula** - the formula for the transformation.
- **Format** - the format for the variable.
- **Label** - the variable label.

Binning Transformations

Binning transformations enable you to collapse an interval variable, such as debt-to-income ratio, into an ordinal grouping variable. There two types of binning transformations: *quantile* and *bucket*.

A **quantile** is any of the four values that divide the values of a variable frequency distribution into four classes. You have the flexibility to divide the data values into *n* equally spaced classes. The quantile transformation is useful when you want to create uniform groups. For example, you may want to divide the variable values into 10 uniform groups (10th, 20th, 30th,.... percentiles).

Buckets are created by dividing the data values into *n* equally spaced intervals based on the difference between the minimum and maximum values. Unlike a quantile transformation, the number of observations in each bucket is typically unequal.

Creating New Columns

Creating new columns is quick and easy by selecting the Create Column item from the Actions pull-down menu or by selecting the Add Computed Column tool icon on the tool bar.

Building Equations

The Customize window contains a column list box, a number pad, an operator pad, and a functions list box to build an equation, which is displayed in the equation box at the bottom of the window and verified when you close the window.

Modifying a Variable's Definition

To change a transformed variable's definition, right click anywhere in the variable row of the transformed variable and select the Modify Definition pop-up menu. This action opens the Computed Column window, which enables you to modify the variable's definition.

Variable Selection Node

Many data mining databases have hundreds of potential model inputs (independent variables). The Variable Selection node can assist you in reducing the number of inputs by dropping those that are unrelated to the target. The Variable Selection node quickly identifies input variables that are useful for predicting the target variable(s) based on a linear models framework. Then, the remaining information-rich inputs can be passed to one of the modeling nodes, such as the Regression node, for more detailed evaluation.

The Variable Selection node facilitates ordinary least squares or logistic regression methods.

Parameters

You can use the Variable Selection node to pre-select variables using an R-square or Chi-square criterion method. The following parameters are available:

- **Remove Variables Unrelated to the Target –** This method provides a fast preliminary variable assessment and facilitates the rapid development of predictive models with large volumes of data. You can quickly identify input variables which are useful for predicting the target variable(s) based on a linear models framework.
- **Remove Variables in Hierarchies** – This variable selection method searches for input variables with a hierarchical relationship. For example, if there is a relationship between state and zip code then you may not want to use both of these inputs. The information may be redundant between the two variables. The Variable selection node finds these hierarchical relationships and gives you the option of keeping the input that has the most detail or the variable that has the least detail.
- **Remove Variables by Percentage of Missing –** This method removes variables with large percentages of missing values. By default, variables having more than 50 percent missing values will be removed. You can type another value in the Percent Missing field.
- **Remove Class Variables with Greater Than or Equal to *n* Values –** This method removes class variables if they have only a single value or if they have greater than or equal to *n* unique values. The default number of unique values for class variable removal is 80.

You select and deselect these methods interactively in the Variable Selection Parameter window. All four methods are used by default for variable selection, and they work together to determine which variables to eliminate. Class variables that have many levels, such as zip code, can be related to the target, but eliminating these variables tends to speed the processing time of the modeling nodes, often without a great loss of information.

Viewing the Results

When you run the Variable selection node, the results are generated and displayed in a results browser. The browser automatically opens when the node finishes running. The Results browser is a tabbed dialog composed of the following tabs:

- **Variables –** summarizes the decisions generated by the variable selection algorithms.
- **Log –** displays the log generated by the DMINE procedure.
- **Output –** displays the output from the DMINE procedure.
- **Code –** shows the code generated by the DMINE procedure.
- **R-square –** displays a horizontal bar chart of the simple R-square value for each model term.
- **Effects –** displays a horizontal bar chart showing the incremental increase in the model R-square value for selected inputs.

Modeling Nodes

The three main tools in Enterprise Miner software for performing statistical modeling are

- **Decision Trees** – for classification trees
- **Regression** – for linear and logistic regression
- **Neural Networks** – for nonlinear or linear modeling.

Decision Tree Node

The Decision Tree node enables you to create decision trees that either

- classify observations based on the values of nominal or binary targets,
- predict outcomes for interval targets, or
- predict the appropriate decision when you specify decision alternatives.

Decision trees produce a set of rules that can be used to generate predictions for a new data set. This information can then be used to drive business decisions. For example, in database marketing, decision trees can be used to develop customer profiles that can help target promotional mailings in order to generate a higher response rate.

The Decision Tree Node finds multi-way splits based on nominal, ordinal, and interval inputs. You choose the splitting criteria that you would like to use to create the tree. The available options represent a hybrid of the options from the CHAID (Chi-squared automatic interaction detection), CART (classification and

regression trees), and C4.5 algorithms. You also can set the options to simulate traditional CHAID, CART, or C4.5.

The Decision Tree Node supports both automatic and interactive training. When you run the node in automatic mode, it automatically ranks the input variables based on the strength of their contribution to the target. This ranking may be used to select variables for use in subsequent modeling. In addition, dummy variables that represent important "interactions" between variables can be automatically generated for use in subsequent modeling. You may override any automatic step with the option to define a splitting rule and prune explicit nodes or sub-trees. Interactive training enables you to explore and evaluate a large set of trees that you develop heuristically.

In addition, the Decision Tree Node enables you to

- **Use prior probabilities and frequencies to train data** in proportions that are different from those of the populations on which predictions are made. For example, if fraud occurs in one percent of transactions, then one tenth of the non-fraud data is often adequate to develop the model using prior probabilities that adjust the 10-to-1 ratio in the training data to the 100-to-1 ratio in the general population.

- **Base the criterion for evaluating a splitting rule on either a statistical significance test**, namely an F-test or a Chi-square test, or on the reduction in variance, entropy, or gini impurity measure. The F-test and Chi-square test accept a p-value input as a stopping rule. All criteria allow the creation of a sequence of sub-trees. You can use validation to select the best sub-tree.

- **Evaluate a tree (or sub-tree) by incorporating a profit matrix** that is associated with a particular decision alternative. In the special situation in which you predict the value of a categorical variable, the profit matrix implements misclassification costs. For example, the incorrect prediction of a transaction as fraudulent might cost less than the incorrect prediction of that transaction as non-fraudulent.

Dialog Pages

The interface to the Decision Tree node is a dialog box that includes the following pages:

General

In the General dialog page, you can specify the splitting criterion and values related to the size of the tree. For nominal or binary targets, you have a choice of three splitting criteria:

- **Chi-Square Test** – the Pearson Chi-Square measure of the target vs. the branch node, with a default significance level of 0.20.
- **Entropy Reduction** – the entropy measure of node impurity
- **Gini Reduction** – the Gini measure of node impurity.

For ordinal or interval targets, you have a choice of two splitting criteria:

- **F Test** – with a default significance level of 0.20.
- **Variance Reduction.**

Advanced

In the Advanced page, you can select from the following sub-tree options:

- Best Assessment Value
- Distinct Distributions in Leaves
- Most Leaves
- At Most Indicated Number of Leaves.

If you have selected either the Chi-Square Test or F Test in the Options page, you also can specify a method to adjust the p-values, either

- **KASS** – multiplies the p-value by a factor that depends on the number of branches and number of distinct values of the inputs
- **DEPTH** – adjusts the final p-value for a partition to simultaneously accept all previous partitions used to create the current subset being partitioned.

Priors

In the Priors dialog page, you can select one of the following prior probabilities options to implement prior class probabilities for nominal targets:

- **Proportional to the Data** – implements prior class probabilities for the target based on the distribution of the target values in the training data set. For example, if 20 percent of the observations have a target value of 1 and 80 percent have a target value of 0, then the prior probabilities for the target would be .2 and .8.

- **Equal probability** – applies equal probabilities for the target, .5 and .5 respectively for the target variable.
- **Explicit** – enables you to apply explicit prior probabilities to the target values.

When you select the Explicit option, a table opens that enables you to enter explicit prior probabilities for each target value.

Assessment

In the Assessment dialog page, you can define how to perform overall assessment of the target and specify decision alternatives and threshold values.

Output

The Output page consists of two sub-pages: the **Data Sub-page**, which enables you to score the model and lists output data set details; and the **Variables Sub-page**, which enables you to select output variables to be used by subsequent modeling nodes.

Tree Diagram Pop-up Menu

If you right click on the background of the Tree Diagram window, a pop-up menu opens that enables you to specify tree customizations, save the tree, and print the tree.

Input

The Decision Tree node requires one target variable and at least one input variable. The target variable can be nominal, binary, or interval. The input variables can be nominal, binary, ordinal, or interval. The bonus, frequency, and weight variables must be interval. The target, input, bonus, frequency, and weight variables are exclusive. Optionally, you can specify bonus variables, a frequency variable, and a weight variable.

Viewing the Results

Output of the Decision Tree node includes the following:

- **Summary Table** – provides summary statistics for the currently selected tree. For nominal target variables, the Summary Table presents *n x m* tables for the training data and the validation data.
- **Tree Ring Navigator** – presents a graphical display of possible data segments from which to form a tree. The Tree Ring Navigator also enables you to view specific segments in the Tree Diagram. You can use the tool box at the top of the application to control the Tree Ring. Tool tips are displayed when you place your cursor over a tool icon.
- **Assessment Table** – provides a measure of how well the tree describes the data. For a nominal target, the default measure is the proportion of observations correctly classified. For an interval target, the default measure is the average sum of squared differences of an observation from its predicted value. The table displays the assessment for several candidate partitions of the data. If a validation data set is used, the assessment based on the validation data will be more reliable than that based on the training data.
- **Assessment Graph** – plots the assessment values from the Assessment Table.

Tree Diagram

The Tree diagram displays node (segment) statistics, the names of variables used to split the data into nodes, and the variable values for several levels of nodes in the tree.

Output Data Sets

The Decision Tree node includes an Output page that is part of the Tree Browser, which enables you to specify a scoring output data set, and to select the variables you want to output for subsequent modeling.

Neural Network Node

An artificial neural network is a computer application that attempts to mimic the neurophysiology of the human brain in the sense that the network learns to find patterns in data from a representative data sample. More specifically, it is a class of flexible nonlinear regression models, discriminant models, and data reduction models, which are interconnected in a nonlinear dynamic system. By detecting complex non-linear relationships in data, neural networks can help you make predictions about real-world problems.

An important feature of the Neural Network node is its built-in intelligence about neural network architecture. The node surfaces this intelligence to the user by making functions available or unavailable in the GUI according to what is mathematically compatible within a neural network. Unavailable functions are grayed out, which simplifies the building process for the user and ensures that all available functions are compatible with neural network architecture.

The following neural network architectures are available in Enterprise Miner:

- generalized linear model (GLIM)
- multi-layer perceptron (MLP), which is often the best architecture for prediction problems
- radial basis function (RBF), which is often best for clustering problems
- equal-width RBF
- normalized RBF
- normalized equal-width RBF.

Dialog Pages

The interface to the Neural Network node is a dialog box that includes the following pages:

Initialization

In the Initialization dialog page, you can accomplish the following tasks:

- **generate a random seed**, by selecting "Generate New Seed." The random seed affects the starting point for training the network. If the starting point is close to the final settings, then the training time can be dramatically reduced. Conversely, if the starting point is not close to the final settings, then training time tends to increase. You may want to first accept the default random seed setting, and then in later runs, specify other random seeds.
- **select a distribution**, by clicking the down arrow and selecting from the resulting menu. Choices are uniform, normal, and cauchy.
- **select a scale.**
- **select a location.**
- **select initial estimates.**

Preliminary Optimization

In the Preliminary Optimization dialog page, you can specify

- the number of preliminary runs
- the training technique
- the maximum iterations
- whether model defaults are allowed
- the maximum CPU time.

Training

In the Training dialog page, you can:

- specify the training technique
- specify the maximum iterations
- allow model defaults to be provided
- specify the maximum CPU time
- request a plot of error history
- specify whether to always retrain the network.

Output

In the Output dialog page, you can view properties of output data sets. By clicking Properties, you can view the administrative details about the data set and view the data set in a table. These data sets can also be viewed in the Data Sets tab of the Results Browser. Output includes the following:

- **estimates data sets** – for preliminary optimization and training
- **output data sets** – for training, validation, testing, and scoring
- **fit statistics data sets** – for training, validation, and testing.

Advanced

In the Advanced dialog page, you can specify the

- objective function
- maximum number of function calls
- default layer size
- convergence criteria.

Regression Node

The Regression node enables you to fit both linear and logistic regression models to a predecessor data set in an Enterprise Miner process flow. Linear regression attempts to predict the value of a continuous target as a linear function of one or more independent inputs. Logistic regression attempts to predict the probability that a binary or ordinal target will acquire the event of interest as a function of one or more independent inputs.

The node includes a point-and click "Interaction Builder" to assist you in creating higher-order modeling terms. The Regression node, like the Decision Tree and Neural Network nodes, also provides you with a directory table facility, called the Model Manager, in which you can store and access models on demand. The node supports forward, backward, and stepwise selection methods. Data sets that have a role of score are automatically scored when you train the model.

In addition, Enterprise Miner enables you to build regression models in a batch environment.

Data sets used as input to the Regression Node can include cross-sectional data, which are data collected across multiple customers, products, geographic regions, and so on, but typically, not across multiple time periods.

Dialog Pages

The interface to the Regression node is a dialog box that includes the following pages:

Variables

The Variables dialog page contains a data table, which enables you to specify the status for the variables in the input data set, sort the variables, and an Interaction Builder, which enables you to add interaction terms to the model.

For example, if the effect of one input on the target depends on the level of one or more inputs, you may want to use the Interaction Builder to add interaction terms to your model. An interaction term, or multiplicative effect, is a product of existing explanatory inputs. For example, the interaction of SALARY and DEPT is SALARY*DEPT.

Another example involves a polynomial model, which includes powers of existing explanatory inputs. In such a situation, you may want to use the Variables dialog page to include polynomial terms if you suspect a nonlinear relationship exists between the input(s) and the target.

Model Options

The Model Options dialog page provides details about, and enables you to specify options for, the target variable and the regression process. The Model Options dialog page includes sub-pages for the Target Definition and the type of Regression.

- The **Target Definition** sub-page lists the name and measurement level of the target variable.
- The **Regression** sub-page enables you to specify whether the regression type is linear or logistic and what type of link functions to use. For binary or ordinal targets, the default regression type is logistic. For interval targets, the default regression type is linear. For a linear regression, the identity link function is used. For a logistic regression, you can select either logit, cloglog (complementary log-log), or probit as the link function.

The Model Options dialog page also enables you to specify the input coding as either deviation or GLM as well as suppress or not suppress the intercept.

Selection Method

The Selection Method dialog page enables you to specify details about model selection. You can choose from the following selection methods:

- **backward** – begins with all inputs in the model and then systematically removes inputs that are not related to the target.
- **forward** – begins with no inputs in the model and then systematically adds inputs that are related to the target.
- **stepwise** – systematically adds and deletes inputs from the model. Stepwise selection is similar to forward selection except that stepwise may remove an input once it has entered the model and replace it with another input.
- **none** – all inputs are used to fit the model.

If you choose the Forward, Backward, or Stepwise selection method, then you can specify the selection criteria as either AIC (Akaike's Information Criterion), SBC (Schwarz's Bayesian Criterion), Validate, Cross Validate, or None.

Advanced

You set the optimization method, iteration controls, and convergence criteria in the Advanced dialog page. Nonlinear optimization methods include the following:

- gradient
- double dogleg
- Newton-Raphson with line search
- Newton-Raphson with ridging
- quasi-Newton
- trust-region.

Viewing the Results

The interface to the Regression results is a dialog box that includes the following dialog pages:

Estimates – displays a bar chart of the standardized or non-standardized parameter estimates from the regression analysis. A standardized parameter estimate is obtained by standardizing all the inputs to zero mean and unit variance prior to running the regression.

Statistics – lists fit statistics, in alphabetical order, for the training data, validation data, and test data analyzed with the regression model.

Output – lists the standard SAS output from linear or logistic regression analysis, depending on what type of regression analysis you specified in the Regression node. For linear regression, the standard output lists the following information about the model:

- R-square
- adjusted R-square
- AIC (Akaike's Information Criterion)
- SBC (Schwarz's Bayesian Criterion)
- BIC (Bayesian Information Criterion)
- C_P (Mallows' C_P statistic).

For logistic regression, the standard output lists the following information about the target and input variables:

- **Response Profile** – For each level of the target, it lists the ordered value of the response variable, and the count or frequency.
- **Class Level Information** – For each class input variable, it lists the values of the design matrix.

Properties – lists the following information about the model:

- name you specified for the model settings
- description you specified for the model
- date that you created the model
- last date that the model was modified
- type of regression (linear, or logistic)
- name of the target variable.

User-Defined Model Node

The User-Defined Model node enables you to generate assessment statistics using predicted values from a model built with the SAS Code node (such as a logistic model using the SAS/STAT™ LOGISTIC procedure) or from the Variable Selection node. Also, the predicted values can be saved to a data set and then imported into the process flow with the Input Data Source node.

Model Manager

The Regression node, the Decision Tree node, and the Neural Network node includes a directory table facility, called the Model Manger, in which you can store and access models on demand.

Dialog Pages

The interface to the Model Manager is a dialog box that includes the following pages:

Models

The Model Manager opens with the Models dialog page, which lists the trained models. For each model, information about when and how the model was created is listed along with fit statistics generated from the training, validation, and/or test data sets.

Profit Matrix

You use the Profit Matrix dialog page to define a table of expected revenues and costs for each decision alternative for each level of the target variable.

Assessment Options

In the Assessment Options dialog page, you set the partitioned data set that is used for model assessment and for determining if an exact model is created or not.

Assessment Reports

You use the Assessment Reports dialog page to select the assessment charts you want to create for a model.

Assessment Nodes

Assessment provides a common framework to compare models and predictions from any analytical tool in Enterprise Miner. The common criteria for all modeling and predictive tools are the expected and actual profits for a project that uses the model results. These are the criteria that enable you to make cross-model comparisons and assessments, independent of all other factors such as sample size or the type of modeling tool used.

Assessment Node

The Assessment node provides a common framework to compare models and predictions from the Decision Tree, Neural Network, and Regression nodes. The common criteria for all modeling and predictive tools are the expected and actual profits obtained from model results. These are the criteria that enable the user to make cross-model comparisons and

assessments, independent of all other factors such as sample size and modeling node.

Assessment statistics are automatically computed when you train a model with a modeling node. You can compare models with either the Assessment node or the Model Manager of a modeling node. The Assessment node and the Model Manager provide the same assessment reports.

An advantage of the Assessment node is that it enables you to compare models created by multiple modeling nodes. The Model Manager is restricted to comparing models trained by the respective modeling node. An advantage of the Model Manager is that it enables you to re-define the cost function (Profit Matrix) for a model. The Assessment node uses the Profit Matrix defined in the Model Manager as input. Essentially, the Assessment node serves as a browser. Therefore, you cannot re-define the Profit Matrix in the Assessment node.

Initial assessment report options are defined in the Enterprise Miner Administrator. You can re-define these options in the Model Manager of the respective modeling node. Assessment options defined in the Administrator or the Model Manager cannot be changed in the Assessment node.

Input

The Assessment node requires the following two inputs:

- **scored data set** – consists of a set of posterior probabilities for each level of a binary-level, nominal-level or ordinal-level target variable. The Regression, Neural Network, and Decision Tree nodes automatically produce a scored data set as output. If the target is interval-level, then the scored data set does not contain posterior probabilities. Instead, it contains the predicted values for the target. You produce a scored data set when you train a model with a modeling node.

- **cost function** – is a table of expected revenues and expected costs for each decision alternative for each level of the target variable. Also known as the **profit matrix**. An optional parameter in the cost function is a value for unit cost. The costs of the decision alternatives may be the same for each level of the target, but they may differ, depending on the actions required by the business decision. The expected profit depends on the level of the target variable. The quality of the assessment depends upon how accurately the users can estimate the cost function. Cost functions should be regularly examined and updated as necessary. You incorporate a cost function into the model assessment by defining a profit matrix in the Model Manager.

Expected Profits

The Assessment node combines the posterior probabilities from the scored data set with the cost function to produce expected profits.

Standard Charts for Assessment

The Assessment node uses the expected profits and actual profits to produce standard charts and tables that describe the usefulness of the model that was used to create the scored data set.

You view the results of the Assessment node by selecting one or more assessment charts. The Tools pull-down menu enables you to specify which charts to create and to browse the data sources. The data sources can be training, validation, or test data sets. Assessment charts include the following:

- lift charts (or *gains charts*)
- profit charts
- return on investment (ROI) charts
- diagnostic classification charts
- statistical receiver operating characteristic (ROC) charts
- business ROC charts
- top-bottom charts (or *top 10 marginal impact variables charts*)
- mosaic charts
- MDDB charts
- threshold-based charts
- interactive profit/loss assessment charts.

Lift Charts

In a lift chart (also known as a *gains chart*) all observations from the scored data set are sorted from highest expected profit to lowest expected profit. Then the observations are grouped into cumulative deciles. If a profit matrix was not defined in the Model Manager, then a default profit matrix is used, which has the expected profit equal to the posterior probability.

Lift charts show the percent captured positive response, percent positive response, or the lift value on the vertical axis. These statistics can be displayed

as either cumulative or non-cumulative values for each decile.

An index value, called a *lift index*, scaled from -100 to 100, represents this area of gain in the lift chart. Useful models have index values closer to 100, while weaker models have index values closer to zero. In rare situations, the index may have negative values.

Profit Chart

In a profit chart, the cumulative or non-cumulative profits within each decile of expected profits is computed. For a useful predictive model, the chart will reach its maximum fairly quickly. For a model that has low predictive power, the chart will rise slowly and not reach its maximum until a high cumulative decile.

Return on Investment Chart

The return on investment (ROI) chart displays the cumulative or non-cumulative ROI for each decile of observations in the scored data set. The return on investment is the ratio of actual profits to costs, expressed as a percentage.

Diagnostic Classification Charts

Diagnostic classification charts provide information on how well the scored data set predicts the actual data. The type of chart you can produce depends on if the target is a non-interval or interval target.

Classification Charts for Non-Interval Targets

Classification charts display the agreement between the predicted and actual target variable values for non-interval-level target variables.

Classification Plots for Interval-Level Targets

For interval-level target variables, the Assessment node plots the actual values against the values predicted by the model.

The Assessment node also plots the residuals (actual values - predicted values) for the target against the predicted target values.

Receiver Operating Characteristic (ROC) Charts

ROC charts display the sensitivity (true positive / total actual positive) and specificity (true positive / total actual negative) of a classifier for a range of cutoffs. ROC charts require a binary target.

Statistical ROC Charts

Each point on the curve represents a cutoff probability. Points closer to the upper-right corner correspond to low cutoff probabilities. Points in the lower left correspond to higher cutoff probabilities. The extreme points (1,1) and (0,0) represent no-data rules where all cases are classified into class 1 or class 0, respectively.

Business ROC Charts

Business ROC charts display the prediction accuracy of the target across a range of decision threshold values. An advantage of a business ROC chart over a statistical ROC chart, is that you have an indication of model performance across a range of threshold levels.

Top-Bottom Charts

Top-bottom charts (also know as *top 10 marginal impact charts*) compare observations in the scored data set that are ranked the highest in expected profits with those that are ranked the lowest in expected profits to show which input variables are important in making that distinction.

For nominal-level or ordinal-level input variables in the top-bottom (TB) chart, the chart displays the total frequency for each level and the percentage of that total in the top and bottom categories. For interval-level input variables, the TB chart displays the total frequency and the average value in the top and bottom categories.

The Assessment node can produce the following top-bottom charts:

- **TB10** – compares the top 10 percent of the data to the bottom 10 percent of the data
- **TB25** – compares the top 25 percent of the data to the bottom 25 percent of the data
- **TB50** – compares the top 50 percent of the data to the bottom 50 percent of the data.

Mosaic Charts

Mosaic charts provide a graphical representation of the discriminative power of the input variables listed in the top-bottom charts.

MDDB Charts

MDDB charts display the top ten input variables, as determined by the modeling tool.

Threshold-based Charts

Threshold-based charts enable you to display the agreement between the predicted and actual target values across a range of threshold levels (the cutoff that is used to classify an observation based on the event level posterior probabilities).

Interactive Profit/Loss Assessment Charts

You can also create an interactive profit/loss assessment chart that enables you to interactively access how a profit/loss matrix impacts the total return over a range of threshold levels.

Score Node

The Score node enables you to manage, edit, export, and execute scoring code that is generated from trained models. Scoring is the generation of predicted values for a new data set that may not contain a target. Scoring a new data set is the end result of most data mining problems. For example,

- A marketing analyst may want to score a data base to create a mailing list of customers most likely to make a purchase.
- A financial analyst may want to score credit union members to identify probable fraudulent transactions.

The Score node generates and manages scoring formulas in the form of a single SAS data step, which can be used in most SAS environments even without the presence of Enterprise Miner software.

Any node that modifies the observations of the input variables or creates scoring formula generates components of score code. The following nodes generate components of scoring code:

- **Transformation Node –** creates new variables from data set variables
- **Data Replacement Node –** imputes missing values
- **Clustering Node –** creates a new segment ID column and imputes missing values
- **Group Processing Node –** subsets the data using IF statements
- **Regression Node –** creates predicted values
- **Neural Network Node –** creates predicted values
- **Decision Tree Node –** creates predicted values
- **SAS Code Node –** avenue for customized scoring code.

Dialog Pages

The interface to the Score node is a dialog box that includes the following pages:

Score Code Page

The Score Code dialog page provides access to the score code management functions. Management functions include **current imports**, which lists the scoring code currently imported from node predecessors, and **accumulated runs**, which lists scoring code exported by the nodes predecessors during the most recent path run (training action).

The Score Code dialog page also includes a pop-up menu that provides access to the saved management functions, which enable you to display the score code in the viewer and save currently imported or accumulated run entries to a file.

Saving Code

If you decide that a model provides good scoring code, you may want to save it for future scoring. Saved code can be edited, deleted, exported, or submitted using the pop-up menu.

Run Action

The Run Action dialog page contains the following sub-pages:

- **General Sub-page** – enables you to select the action that will be performed when the Score node is run within a diagram path (default)
- **Merge Options Sub-page** – enables you to select variables that you want to keep in the merged data set.

Client/Server Enablement

The GUI includes dialog pages and other graphical tools such as radio buttons and menus that make establishing connections between servers and clients fast, efficient, and easy to understand. The client/server functionality of Enterprise Miner software provides advantages because the solution

- distributes data-intensive processing to the most appropriate machine
- minimizes network traffic by processing the data on the source machine

- minimizes data redundancy by maintaining one central data source
- distributes server profiles to multiple clients
- can regulate access to data sources
- can toggle between remote and local processing.

With Enterprise Miner, you can access diverse data sources from database management systems on servers to use with your local SAS Enterprise Miner session.

Administrator and Users

Due to the complexities that can be encountered in configuring client/server connections, Enterprise Miner relies on a person to act as an administrator, someone who will perform all system setup functions. The administrator role is established when you first invoke the Administration Dialog Box and enter a password.

A user is the beneficiary of server profiles that are defined and distributed by the administrator. A user can establish a remote connection, but does not have the authority to define new server profiles.

Administration Window

The interface to the client/server Administration functions is a tabbed dialog box as shown in Figure 4.

Figure 4: Administration Window

Dialog Pages

The interface to the Administration window is a dialog box that includes the following pages:

General

In the General dialog page of the Administration window, you set the privilege mode as either administrator or user. An administrator can define and modify server profiles; a user cannot.

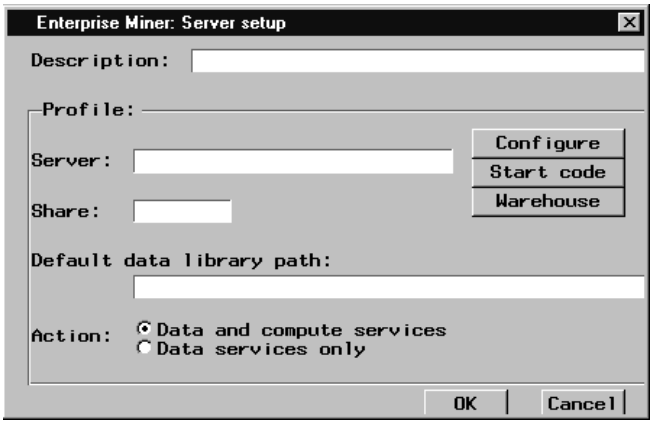
Servers

The Server dialog page enables you to define server and query profiles. A server profile contains all the configuration information necessary to establish a connection on a remote server. You must define a server profile before you can establish a remote connection or define a query profile.

The Server Setup Window

To add a server profile, begin by simply selecting “Add” from the Administration window. This action opens the Server Setup window, which is shown in Figure 5.

Figure 5: Server Setup Window



The Server Setup window enables you to enter the server profile including information such as a description, the network address of the server in either name or number format, and the default data library pathname. By selecting a setting of the radio button at the bottom of the Server Setup window, you quickly can specify whether you want processing to take place on the remote server or download data samples from the server to be processed locally.

The Server Setup window also includes options that enable you to configure aspects of the remote session such as start code, which will run when the server is initialized.

Defining Query Profiles

The Server dialog page also provides an option for adding an SQL query profile. Query profiles are used by the Input Data Source node to automatically load your query profile preferences.

Modifying and Removing a Server Profile

A Modify button and a Remove button are also a part of the Server dialog page, which provide convenient ways for the administrator to change or delete server profiles as needed.

Projects

The Projects dialog page of the Administration window enables you to define the project libraries that will appear in the Available Projects window. Adding, modifying, and removing project libraries are all completed using simple data entry fields, pop-up menus, and radio buttons.

Assessment

The Assessment dialog page enables you to define global assessment chart options for model assessment in subsequent diagrams. Assessment charts help you describe the usefulness of the model that was used to create a scored data set. For example, you may want to delimit globally the display of assessment charts to one type such as lift charts, or you may want to enable the display of mosaic charts, which, by default, are not displayed.

Options

By default, results of the log and output are sent to the respective nodes. The Options dialog page enables you to redirect log and output from the node to the SAS System log and output windows of SAS Display Manager.

Defaults

The Defaults dialog page enables you to customize the default node options.

Conclusion

Enterprise Miner software provides all the functionality one needs to plan, implement, and successfully refine data mining projects. The software fully integrates all steps of the data mining process beginning with the sampling of data, through sophisticated data analyses and modeling, to the dissemination of the resulting information. The functionality of Enterprise Miner is surfaced through an intuitive and flexible GUI, which enables users, who may have different degrees of statistical expertise, to mine volumes of data for valuable information.

References and Further Reading

SAS Institute Inc. (1995), SAS Institute White Paper: *Building a SAS® Data Warehouse*, Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1996), SAS Institute White Paper: *OLAP Tools and Techniques within the SAS® System*, Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1996), SAS Institute White Paper: *The SAS® System and Web Integration*, Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1997), SAS Institute White Paper: *Business Intelligence Systems and Data Mining*, Cary, NC: SAS Institute Inc.

Acknowledgments

Contributors to *Finding the Solution to Data Mining: A Map of the Features and Components of SAS® Enterprise Miner ™ Software* include the following SAS Institute employees: Brent L. Cohen, James D. Seabolt, R. Wayne Thompson, and John S. Williams.

Contact

Mark Brown
Program Manager, Data Mining
SAS Institute Inc.
SAS Campus Drive
Cary, NC, 27513

Voice: 919-677-8000, ext. 7165
E-mail: sasjub@unx.sas.com