CHAPTER

1

Introduction to SAS Enterprise Miner 5.3 Software

Data Mining Overview

SAS defines *data mining* as the process of uncovering hidden patterns in large amounts of data. Many industries use data mining to address business problems and opportunities such as fraud detection, risk and affinity analyses, database marketing, householding, customer churn, bankruptcy prediction, and portfolio analysis.The SAS data mining process is summarized in the acronym SEMMA, which stands for sampling, exploring, modifying, modeling, and assessing data.

- *Sample* the data by creating one or more data tables. The sample should be large enough to contain the significant information, yet small enough to process.
- *Explore* the data by searching for anticipated relationships, unanticipated trends, and anomalies in order to gain understanding and ideas.
- *Modify* the data by creating, selecting, and transforming the variables to focus the model selection process.
- *Model* the data by using the analytical tools to search for a combination of the data that reliably predicts a desired outcome.
- *Assess* the data by evaluating the usefulness and reliability of the findings from the data mining process.

You might not include all of these steps in your analysis, and it might be necessary to repeat one or more of the steps several times before you are satisfied with the results.

After you have completed the assessment phase of the SEMMA process, you apply the scoring formula from one or more champion models to new data that might or might not contain the target. The goal of most data mining tasks is to apply models that are constructed using training and validation data in order to make accurate predictions about observations of new, raw data.

The SEMMA data mining process is driven by a process flow diagram, which you can modify and save. The Graphical User Interface is designed in such a way that the business analyst who has little statistical expertise can navigate through the data mining methodology, while the quantitative expert can go "behind the scenes" to fine-tune the analytical process.

SAS Enterprise Miner 5.3 contains a collection of sophisticated analysis tools that have a common user-friendly interface that you can use to create and compare multiple models. Analytical tools include clustering, association and sequence discovery, market basket analysis, path analysis, self-organizing maps / Kohonen, variable selection, decision trees and gradient boosting, linear and logistic regression, two stage modeling, partial least squares, support vector machines, and neural networking. Data preparation tools include outlier detection, variable transformations, variable clustering, interactive binning, principal components, rule building and induction, data imputation, random sampling, and the partitioning of data sets (into train, test, and validate data sets). Advanced visualization tools enable you to quickly and easily examine large amounts of data in multidimensional histograms and to graphically compare modeling results.

Enterprise Miner is designed for PCs or servers that are running under Windows XP, UNIX, Linux, or subsequent releases of those operating environments. The figures and screen captures that are presented in this document were taken on a PC that was running under Windows XP.

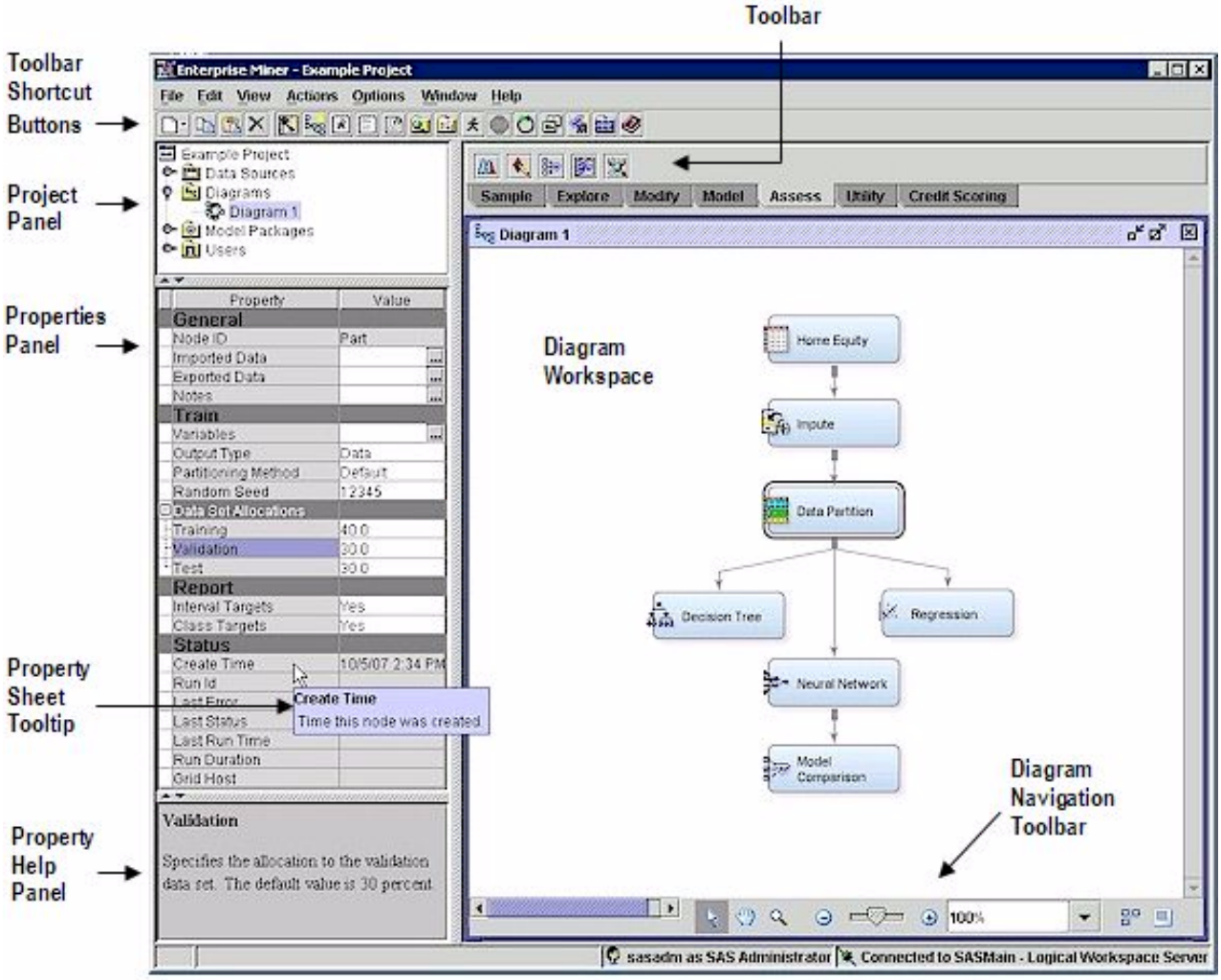
Layout of the Enterprise Miner Window

About the Graphical Interface

You use the Enterprise Miner graphical interface to build a process flow diagram that controls your data mining project.

Figure 1.1 shows the components of the Enterprise Miner window.

Figure 1.1 The Enterprise Miner Window



The Enterprise Miner window contains the following interface components:

- Toolbar and Toolbar shortcut buttons — The Enterprise Miner Toolbar is a graphic set of node icons that are organized by SEMMA categories. Above the Toolbar is a collection of Toolbar shortcut buttons that are commonly used to build process flow diagrams in the Diagram Workspace. Move the mouse pointer over any node, or shortcut button to see the text name. Drag a node into the Diagram Workspace to use it. The Toolbar icon remains in place and the node in the Diagram Workspace is ready to be connected and configured for use in your process flow diagram. Click on a shortcut button to use it.
- Project Panel — Use the Project Panel to manage and view data sources, diagrams, model packages, and project users.
- Properties Panel — Use the Properties Panel to view and edit the settings of data sources, diagrams, nodes, and model packages.
- Diagram Workspace — Use the Diagram Workspace to build, edit, run, and save process flow diagrams. This is where you graphically build, order, sequence and connect the nodes that you use to mine your data and generate reports.
- Property Help Panel — The Property Help Panel displays a short description of the property that you select in the Properties Panel. Extended help can be found

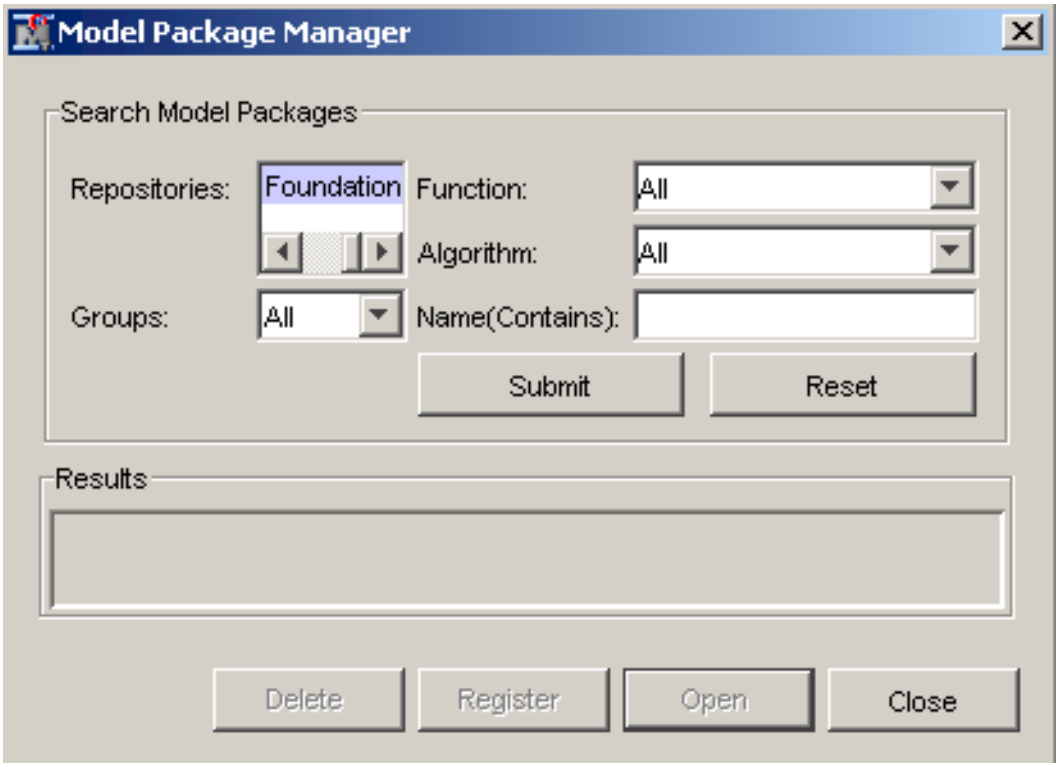
in the Help Topics selection from the Help main menu or from the Help button on many windows.

- Status Bar — The Status Bar is a single pane at the bottom of the window that indicates the execution status of a SAS Enterprise Miner task.

Enterprise Miner Menus

Here is a summary of the Enterprise Miner menus:

- File
  - New
    - Project — creates a new project.
    - Diagram — creates a new diagram.
    - Data Source — creates a new data source using the Data Source wizard.
    - Library — creates a new SAS library.
  - Open Project — opens an existing project. You can also create a new project from the Open Project window.
  - Recent Projects — lists the projects on which you were most recently working. You can open recent projects using this menu item.
  - Open Model Package — opens a model package SAS Package (SPK) file that you have previously created.
  - Explore Model Packages — opens the Model Package Manager window, in which you can view and compare model packages.

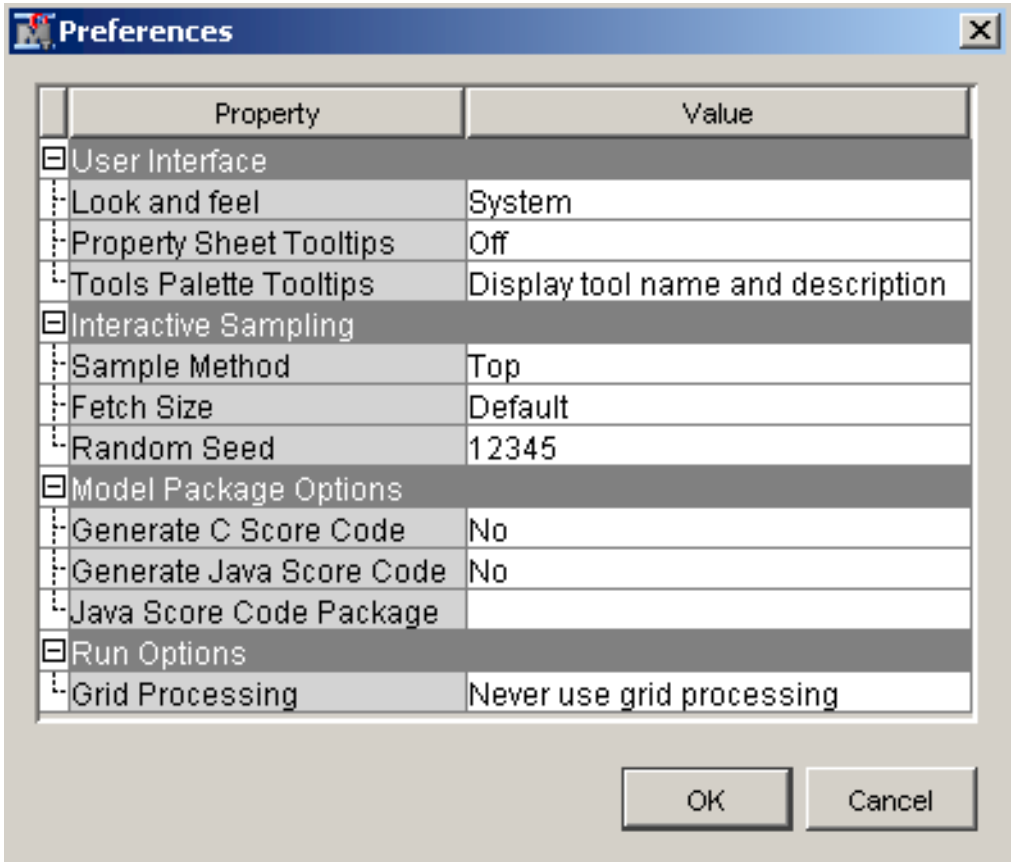


  - Open Diagram — opens the diagram that you select in the Project Panel.
  - Close Diagram — closes the open diagram that you select in the Project Panel.
  - Close this Project — closes the current project.

- Delete this Project — deletes the current project.
- Import Diagram from XML — imports a diagram that has been defined by an XML file.
- Save Diagram As — saves a diagram as an image (BMP or GIF) or as an XML file. You must have an open diagram and that diagram must be selected in the Project Panel. Otherwise, this menu item appears as Save As and is dimmed and unavailable.
- Print Diagram — prints the contents of the window that is open in the Diagram Workspace. You must have an open diagram and that diagram must be selected in the Project Panel. Otherwise, this menu item is dimmed and unavailable.
- Print Preview — displays a preview of the Diagram Workspace that can be printed. You must have an open diagram and that diagram must be selected in the Project Panel. Otherwise, this menu item is dimmed and unavailable.
- Exit — ends the Enterprise Miner session and closes the window.

- Edit
  - Cut — deletes the selected item and copies it to the clipboard.
  - Copy — copies the selected node to the clipboard.
  - Paste — pastes a copied object from the clipboard.
  - Delete — deletes the selected diagram, data source, or node.
  - Rename — renames the selected diagram, data source, or node.
  - Duplicate — creates a copy of the selected data source.
  - Select All — selects all of the nodes in the open diagram, selects all texts in the Program Editor, Log, or Output windows.
  - Clear All — clears text from the Program Editor, Log, or Output windows.
  - Find/Replace — opens the Find/Replace window so that you can search for and replace text in the Program Editor, Log, and Results windows.
  - Go To Line — opens the Go To Line window. Enter the line number on which you want to enter or view text.
  - Layout
    - Horizontally — creates an orderly horizontal arrangement of the layout of nodes that you have placed in the Diagram Workspace.
    - Vertically — creates an orderly vertical arrangement of the layout of nodes that you have placed in the Diagram Workspace.
  - Zoom — increases or decreases the size of the process flow diagram within the diagram window.
  - Copy Diagram to Clipboard — copies the Diagram Workspace to the clipboard.
- View
  - Program Editor — opens a SAS Program Editor window in which you can enter SAS code.
  - Log — opens a SAS Log window.
  - Output — opens a SAS Output window.
  - Explorer — opens a window that displays the SAS libraries (and their contents) to which Enterprise Miner has access.
  - Graphs — opens the Graphs window. Graphs that you create with SAS code in the Program Editor are displayed in this window.
  - Refresh Project — updates the project tree to incorporate any changes that were made to the project from outside the Enterprise Miner user interface.

- Actions
  - Add Node — adds a node that you have selected to the Diagram Workspace.
  - Select Nodes — opens the Select Nodes window.
  - Connect nodes — opens the Connect Nodes window. You must select a node in the Diagram Workspace to make this menu item available. You can connect the node that you select to any nodes that have been placed in your Diagram Workspace.
  - Disconnect Nodes — opens the Disconnect Nodes window. You must select a node in the Diagram Workspace to make this menu item available. You can disconnect the selected node from a predecessor node or a successor node.
  - Update — updates the selected node to incorporate any changes that you have made.
  - Run — runs the selected node and any predecessor nodes in the process flow that have not been executed, or submits any code that you type in the Program Editor window.
  - Stop Run — interrupts a currently running process flow.
  - View Results — opens the Results window for the selected node.
  - Create Model Package — generates a mining model package.
  - Export Path as SAS Program — saves the path that you select as a SAS program. In the window that opens, you can specify the location to which you want to save the file. You also specify whether you want the code to run the path or create a model package.
- Options
  - Preferences — opens the Preferences window. Use the following options to change the user interface:



- Look and Feel — you can select **`Cross Platform`**, which uses a standard appearance scheme that is the same on all platforms, or **`System`** which uses the appearance scheme that you have chosen for your platform.
- Property Sheet Tooltips — controls whether tooltips are displayed on various property sheets appearing throughout the user interface.
- Tools Palette Tooltips — controls how much tooltip information you want displayed for the tool icons in the Toolbar.
- Sample Methods — generates a sample that will be used for graphical displays. You can specify either **`Top`** or **`Random`**.
- Fetch Size — specifies the number of observations to download for graphical displays. You can choose either Default or Max.
- Random Seed — specifies the value you want to use to randomly sample observations from your input data.
- Generate C Score Code — creates C score code when you create a report. The default is No.
- Generate Java Score Code — creates Java score code when you create a report. The default is No. If you select Yes for **`Generate Java Score Code`**, you must enter a filename for the score code package in the Java Score Code Package box.
- Java Score Code Package — identifies the filename of the Java Score Code package.
- Grid Processing — enables you to use grid processing when you are running data mining flows on grid-enabled servers.

- Window
  - Tile — displays windows in the Diagram Workspace so that all windows are visible at the same time.
  - Cascade — displays windows in the Diagram Workspace so that windows overlap.
- Help
  - Contents — opens the Enterprise Miner Help window, which enables you to view all the Enterprise Miner Reference Help.
  - Component Properties — opens a table that displays the component properties of each tool.
  - Generate Sample Data Sources — creates sample data sources that you can access from the Data Sources folder.
  - Configuration — displays the current system configuration of your Enterprise Miner session.
  - About — displays information about the version of Enterprise Miner that you are using.

Diagram Workspace Pop-up Menus

You can use the Diagram Workspace pop-up menus to perform many tasks. To open the pop-up menu, right-click in an open area of the Diagram Workspace. (Note that you can also perform many of these tasks by using the pull-down menus.) The pop-up menu contains the following items:

- **Add node** — accesses the Add Node window.
- **Paste** — pastes a node from the clipboard to the Diagram Workspace.
- **Select All** — selects all nodes in the process flow diagram.
- **Select Nodes** — opens a window that displays all the nodes that are on your diagram. You can select as many as you want.
- **Layout** — creates an orderly horizontally or vertically aligned arrangement of the nodes in the Diagram Workspace.
- **Zoom** — increases or decreases the size of the process flow diagram within the diagram window by the amount that you choose.
- **Copy Diagram to Clipboard** — copies the Diagram Workspace to the clipboard.

Organization and Uses of Enterprise Miner Nodes

About Nodes

The nodes of Enterprise Miner are organized according to the Sample, Explore, Modify, Model, and Assess (SEMMA) data mining methodology. In addition, there are also Credit Scoring and Utility node tools. You use the Credit Scoring node tools to score your data models and to create freestanding code. You use the Utility node tools to submit SAS programming statements, and to define control points in the process flow diagram.

Note: The **Credit Scoring** tab does not appear in all installed versions of Enterprise Miner. △

Remember that in a data mining project, it can be an advantage to repeat parts of the data mining process. For example, you might want to explore and plot the data at several intervals throughout your project. It might be advantageous to fit models, assess the models, and then refit the models and then assess them again.

The following tables list the nodes and give each node's primary purpose.

Sample Nodes

Node Name	Description
Append	Use the Append node to append data sets that are exported by two different paths in a single process flow diagram. The Append node can also append train, validation, and test data sets into a new training data set.
Data Partition	Use the Data Partition node to partition data sets into training, test, and validation data sets. The training data set is used for preliminary model fitting. The validation data set is used to monitor and tune the model weights during estimation and is also used for model assessment. The test data set is an additional hold-out data set that you can use for model assessment. This node uses simple random sampling, stratified random sampling, or clustered sampling to create partitioned data sets. See Chapter 3.
Filter	Use the Filter node to create and apply filters to your training data set and optionally, to the validation and test data sets. You can use filters to exclude certain observations, such as extreme outliers and errant data that you do not want to include in your mining analysis. Filtering extreme values from the training data tends to produce better models because the parameter estimates are more stable. By default, the Filter node ignores target and rejected variables.
Input Data Source	Use the Input Data Source node to access SAS data sets and other types of data. This node introduces a predefined Enterprise Miner Data Source and metadata into a Diagram Workspace for processing. You can view metadata information about your data in the Input Data Source node, such as initial values for measurement levels and model roles of each variable. Summary statistics are displayed for interval and class variables. See Chapter 3.
Merge	Use the Merge node to merge observations from two or more data sets into a single observation in a new data set.

Node Name	Description
Sample	Use the Sample node to take random, stratified random samples, and to take cluster samples of data sets. Sampling is recommended for extremely large databases because it can significantly decrease model training time. If the random sample sufficiently represents the source data set, then data relationships that Enterprise Miner finds in the sample can be extrapolated upon the complete source data set. The Sample node writes the sampled observations to an output data set and saves the seed values that are used to generate the random numbers for the samples so that you can replicate the samples.
Time Series	Use the Time Series node to convert transactional data to time series data to perform seasonal and trend analysis. This node enables you to understand trends and seasonal variations in the transaction data that you collect from your customers and suppliers over the time, by converting transactional data into time series data. Transactional data is time-stamped data that is collected over time at no particular frequency. By contrast, time series data is time-stamped data that is collected over time at a specific frequency. The size of transaction data can be very large, which makes traditional data mining tasks difficult. By condensing the information into a time series, you can discover trends and seasonal variations in customer and supplier habits that might not be visible in transactional data.

Explore Nodes

Node Name	Description
Association	Use the Association node to identify association relationships within the data. For example, if a customer buys a loaf of bread, how likely is the customer to also buy a gallon of milk? You use the Association node to perform sequence discovery if a time-stamped variable (a sequence variable) is present in the data set. Binary sequences are constructed automatically, but you can use the Event Chain Handler to construct longer sequences that are based on the patterns that the algorithm discovered.
Cluster	Use the Cluster node to segment your data so that you can identify data observations that are similar in some way. When displayed in a plot, observations that are similar tend to be in the same cluster, and observations that are different tend to be in different clusters. The cluster identifier for each observation can be passed to other nodes for use as an input, ID, or target variable. This identifier can also be passed as a group variable that enables you to automatically construct separate models for each group.
DMDB	The DMDB node creates a data mining database that provides summary statistics and factor-level information for class and interval variables in the imported data set. In Enterprise Miner 4.3, the DMDB database optimized the performance of the Variable Selection, Tree, Neural Network, and Regression nodes. It did so by reducing the number of passes through the data that the analytical engine needed to make when running a process flow diagram. Improvements to the Enterprise Miner 5.3 software have eliminated the need to use the DMDB node to optimize the performance of nodes, but the DMDB database can still provide quick summary statistics for class and interval variables at a given point in a process flow diagram.
Graph Explore	The Graph Explore node is an advanced visualization tool that enables you to explore large volumes of data graphically to uncover patterns and trends and to reveal extreme values in the database. You can analyze univariate distributions, investigate multivariate distributions, create scatter and box plots, constellation and 3D charts, and so on. If the Graph Explore node follows a node that exports a data set in the process flow, it can use either a sample or the entire data set as input. The resulting plot is fully interactive: you can rotate a chart to different angles and move it anywhere on the screen to obtain different perspectives on the data. 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. You may also want to use the node downstream in the process flow to perform tasks, such as creating a chart of the predicted values from a model developed with one of the modeling nodes.

Node Name	Description
Market Basket	The Market Basket node performs association rule mining over transaction data in conjunction with item taxonomy. Transaction data contain sales transaction records with details about items bought by customers. Market basket analysis uses the information from the transaction data to give you insight about which products tend to be purchased together. This information can be used to change store layouts, to determine which products to put on sale, or to determine when to issue coupons or some other profitable course of action. The market basket analysis is not limited to the retail marketing domain. The analysis framework can be abstracted to other areas such as word co-occurrence relationships in text documents. The Market Basket node is not included with SAS Enterprise Miner for the Desktop.
MultiPlot	Use the MultiPlot node to explore larger volumes of data graphically. The MultiPlot node automatically creates bar charts and scatter plots for the input and target variables without requiring you to make several menu or window item selections. The code that is created by this node can be used to create graphs in a batch environment. See Chapter 3.
Path Analysis	Use the Path Analysis node to analyze Web log data and to determine the paths that visitors take as they navigate through a Web site. You can also use the node to perform sequence analysis.
SOM/Kohonen	Use the SOM/Kohonen node to perform unsupervised learning by using Kohonen vector quantization (VQ), Kohonen self-organizing maps (SOMs), or batch SOMs with Nadaraya-Watson or local-linear smoothing. Kohonen VQ is a clustering method, whereas SOMs are primarily dimension-reduction methods.
StatExplore	Use the StatExplore node to examine variable distributions and statistics in your data sets. You can use the StatExplore node to compute standard univariate distribution statistics, to compute standard bivariate statistics by class target and class segment, and to compute correlation statistics for interval variables by interval input and target. You can also combine the StatExplore node with other Enterprise Miner tools to perform data mining tasks such as using the StatExplore node with the Metadata node to reject variables, using the StatExplore node with the Transform Variables node to suggest transformations, or even using the StatExplore node with the Regression node to create interactions terms. See Chapter 3.

Node Name	Description
Variable Clustering	Variable clustering is a useful tool for data reduction, such as choosing the best variables or cluster components for analysis. Variable clustering removes collinearity, decreases variable redundancy, and helps to reveal the underlying structure of the input variables in a data set. When properly used as a variable-reduction tool, the Variable Clustering node can replace a large set of variables with the set of cluster components with little loss of information.
Variable Selection	Use the Variable Selection node to evaluate the importance of input variables in predicting or classifying the target variable. To preselect the important inputs, the Variable Selection node uses either an R-Square or a Chi-Square selection (tree-based) criterion. You can use the R-Square criterion to remove variables in hierarchies, remove variables that have large percentages of missing values, and remove class variables that are based on the number of unique values. The variables that are not related to the target are set to a status of rejected. Although rejected variables are passed to subsequent nodes in the process flow diagram, these variables are not used as model inputs by a more detailed modeling node, such as the Neural Network and Decision Tree nodes. You can reassign the status of the input model variables to rejected in the Variable Selection node. See Chapter 5.

Modify Nodes

Node Name	Description
Drop	Use the Drop node to drop certain variables from your scored Enterprise Miner data sets. You can drop variables that have roles of Assess, Classification, Frequency, Hidden, Input, Predict, Rejected, Residual, Target, and Other from your scored data sets.
Impute	Use the Impute node to impute (fill in) values for observations that have missing values. You can replace missing values for interval variables with the mean, median, midrange, mid-minimum spacing, distribution-based replacement. Alternatively, you can use a replacement M-estimator such as Tukey's biweight, Hubers, or Andrew's Wave. You can also estimate the replacement values for each interval input by using a tree-based imputation method. Missing values for class variables can be replaced with the most frequently occurring value, distribution-based replacement, tree-based imputation, or a constant. See Chapter 5.

Node Name	Description
Interactive Binning	The Interactive Binning node is an interactive grouping tool that you use to model nonlinear functions of multiple modes of continuous distributions. The interactive tool computes initial bins by quantiles; then you can interactively split and combine the initial bins.You use the Interactive Binning node to create bins or buckets or classes of all input variables. You can create bins in order to reduce the number of unique levels as well as attempt to improve the predictive power of each input. The Interactive Binning node enables you to select strong characteristics based on the Gini statistic and to group the selected characteristics based on business considerations. The node is helpful in shaping the data to represent risk ranking trends rather than modeling quirks, which might lead to overfitting.
Principal Components	Use the Principal Components node to perform a principal components analysis for data interpretation and dimension reduction. The node generates principal components that are uncorrelated linear combinations of the original input variables and that depend on the covariance matrix or correlation matrix of the input variables. In data mining, principal components are usually used as the new set of input variables for subsequent analysis by modeling nodes.
Replacement	Use the Replacement node to impute (fill in) values for observations that have missing values and to replace specified non-missing values for class variables in data sets. You can replace missing values for interval variables with the mean, median, midrange, or mid-minimum spacing, or with a distribution-based replacement. Alternatively, you can use a replacement M-estimator such as Tukey's biweight, Huber's, or Andrew's Wave. You can also estimate the replacement values for each interval input by using a tree-based imputation method. Missing values for class variables can be replaced with the most frequently occurring value, distribution-based replacement, tree-based imputation, or a constant. See Chapters 3, 4, and 5.
Rules Builder	The Rules Builder node accesses the Rules Builder window so you can create ad hoc sets of rules with user-definable outcomes. You can interactively define the values of the outcome variable and the paths to the outcome. This is useful in ad hoc rule creation such as applying logic for posterior probabilities and scorecard values. Any Input Data Source data set can be used as an input to the Rules Builder node. Rules are defined using charts and histograms based on a sample of the data.
Transform Variables	Use the Transform Variables node 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 nonlinearity, improve additivity, and correct nonnormality in variables. In Enterprise Miner, the Transform Variables node also enables you to transform class variables and to create interaction variables. See Chapter 5.

Model Nodes

Node Name	Description
AutoNeural	Use the AutoNeural node to automatically configure a neural network. It conducts limited searches for a better network configuration. See Chapters 5 and 6.
Decision Tree	Use the Decision Tree node to fit decision tree models to your data. The implementation includes features that are found in a variety of popular decision tree algorithms such as CHAID, CART, and C4.5. The node supports both automatic and interactive training. When you run the Decision Tree node in automatic mode, it automatically ranks the input variables, based on the strength of their contribution to the tree. This ranking can be used to select variables for use in subsequent modeling. You can override any automatic step with the option to define a splitting rule and prune explicit tools or subtrees. Interactive training enables you to explore and evaluate a large set of trees as you develop them. See Chapters 4 and 6.
Dmine Regression	Use the Dmine Regression node to compute a forward stepwise least-squares regression model. In each step, an independent variable is selected that contributes maximally to the model R-square value.
DMNeural	Use DMNeural node to fit an additive nonlinear model. The additive nonlinear model uses bucketed principal components as inputs to predict a binary or an interval target variable.
Ensemble	Use the Ensemble node to create new models by combining the posterior probabilities (for class targets) or the predicted values (for interval targets) from multiple predecessor models.
Gradient Boosting	Gradient boosting is a boosting approach that creates a series of simple decision trees that together form a single predictive model. Each tree in the series is fit to the residual of the prediction from the earlier trees in the series. Each time the data is used to grow a tree, the accuracy of the tree is computed. The successive samples are adjusted to accommodate previously computed inaccuracies. Because each successive sample is weighted according to the classification accuracy of previous models, this approach is sometimes called stochastic gradient boosting. Boosting is defined for binary, nominal, and interval targets.
MBR (Memory-Based Reasoning)	Use the MBR (Memory-Based Reasoning) node to identify similar cases and to apply information that is obtained from these cases to a new record. The MBR node uses *k*-nearest neighbor algorithms to categorize or predict observations.
Model Import	Use the Model Import node to import and assess a model that was not created by one of the Enterprise Miner modeling nodes. You can then use the Model Comparison node to compare the user-defined model with one or more models that you developed with an Enterprise Miner modeling node. This process is called integrated assessment.

Node Name	Description
Neural Network	Use the Neural Network node to construct, train, and validate multilayer feedforward neural networks. By default, the Neural Network node automatically constructs a multilayer feedforward network that has one hidden layer consisting of three neurons. In general, each input is fully connected to the first hidden layer, each hidden layer is fully connected to the next hidden layer, and the last hidden layer is fully connected to the output. The Neural Network node supports many variations of this general form. See Chapters 5 and 6.
Partial Least Squares	The Partial Least Squares node is a tool for modeling continuous and binary targets that are based on SAS/STAT PROC PLS. Partial least squares regression produces factor scores that are linear combinations of the original predictor variables. As a result, no correlation exists between the factor score variables that are used in the predictive regression model. Consider a data set that has a matrix of response variables Y and a matrix with a large number of predictor variables X. Some of the predictor variables are highly correlated. A regression model that uses factor extraction for the data computes the factor score matrix T=XW, where W is the weight matrix. Next, the model considers the linear regression model Y=TQ+E, where Q is a matrix of regression coefficients for the factor score matrix T, and where E is the noise term. After computing the regression coefficients, the regression model becomes equivalent to Y=XB+E, where B=WQ, which can be used as a predictive regression model.
Regression	Use the Regression node to fit both linear and logistic regression models to your data. You can use continuous, ordinal, and binary target variables. You can use both continuous and discrete variables as inputs. The node supports the stepwise, forward, and backward selection methods. A point-and-click term editor enables you to customize your model by specifying interaction terms and the ordering of the model terms. See Chapters 5 and 6.
Rule Induction	Use the Rule Induction node to improve the classification of rare events in your modeling data. The Rule Induction node creates a Rule Induction model that uses split techniques to remove the largest pure split node from the data. Rule Induction also creates binary models for each level of a target variable and ranks the levels from the most rare event to the most common. After all levels of the target variable are modeled, the score code is combined into a SAS DATA step.
Support Vector Machines (Experimental)	Support Vector Machines are used for classification. They use a hyperplane to separate points mapped on a higher dimensional space. The data points used to build this hyperplane are called support vectors.
TwoStage	Use the TwoStage node to compute a two-stage model for predicting a class and an interval target variables at the same time. The interval target variable is usually a value that is associated with a level of the class target.

Note: These modeling nodes use a directory table facility, called the Model Manager, in which you can store and access models on demand. The modeling nodes also enable you to modify the target profile or profiles for a target variable. △

Assess Nodes

Node Name	Description
Cutoff	The Cutoff node provides tabular and graphical information to assist users in determining an appropriate probability cutoff point for decision making with binary target models. The establishment of a cutoff decision point entails the risk of generating false positives and false negatives, but an appropriate use of the Cutoff node can help minimize those risks. You will typically run the node at least twice. In the first run, you obtain all the plots and tables. In subsequent runs, you can change the values of the Cutoff Method and Cutoff User Input properties, customizing the plots, until an optimal cutoff value is obtained.
Decisions	Use the Decisions node to define target profiles for a target that produces optimal decisions. The decisions are made using a user-specified decision matrix and output from a subsequent modeling procedure.
Model Comparison	Use the Model Comparison node to use a common framework for comparing models and predictions from any of the modeling tools (such as Regression, Decision Tree, and Neural Network tools). The comparison is based on the expected and actual profits or losses that would result from implementing the model. The node produces the following charts that help to describe the usefulness of the model: lift, profit, return on investment, receiver operating curves, diagnostic charts, and threshold-based charts. See Chapter 6.
Segment Profile	Use the Segment Profile node to assess and explore segmented data sets. Segmented data is created from data BY-values, clustering, or applied business rules. The Segment Profile node facilitates data exploration to identify factors that differentiate individual segments from the population, and to compare the distribution of key factors between individual segments and the population. The Segment Profile node outputs a Profile plot of variable distributions across segments and population, a Segment Size pie chart, a Variable Worth plot that ranks factor importance within each segment, and summary statistics for the segmentation results. The Segment Profile node does not generate score code or modify metadata.
Score	Use the Score node to manage, edit, export, and execute scoring code that is generated from a trained model. Scoring is the generation of predicted values for a data set that might not contain a target variable. 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. See Chapter 6.

Utility Nodes

Node Name	Description
Control Point	Use the Control Point node to establish a control point to reduce the number of connections that are made in process flow diagrams. For example, suppose three Input Data nodes are to be connected to three modeling nodes. If no Control Point node is used, then nine connections are required to connect all of the Input Data nodes to all of the modeling nodes. However, if a Control Point node is used, only six connections are required.
End Groups	The End Groups node is used only in conjunction with the Start Groups node. The End Groups node acts as a boundary marker that defines the end of group processing operations in a process flow diagram. Group processing operations are performed on the portion of the process flow diagram that exists between the Start Groups node and the End Groups node. If the group processing function that is specified in the Start Groups node is stratified, bagging, or boosting, the End Groups node functions as a model node and presents the final aggregated model. Enterprise Miner tools that follow the End Groups node continue data mining processes normally.
Start Groups	The Start Groups node is useful when your data can be segmented or grouped, and you want to process the grouped data in different ways. The Start Groups node uses BY-group processing as a method to process observations from one or more data sources that are grouped or ordered by values of one or more common variables. BY variables identify the variable or variables by which the data source is indexed, and BY statements process data and order output according to the BY-group values. You can use the Enterprise Miner Start Groups node to perform these tasks: □ define group variables such as GENDER or JOB, in order to obtain separate analyses for each level of a group variable □ analyze more than one target variable in the same process flow □ specify index looping, or how many times the flow that follows the node should loop □ resample the data set and use unweighted sampling to create bagging models □ resample the training data set and use reweighted sampling to create boosting models
Metadata	Use the Metadata node to modify the columns metadata information at some point in your process flow diagram. You can modify attributes such as roles, measurement levels, and order.

Node Name	Description
Reporter	The Reporter node uses SAS Output Delivery System (ODS) capability to create a single PDF or RTF file that contains information about the open process flow diagram. The PDF or RTF documents can be viewed and saved directly and are included in Enterprise Miner report package files. The report contains a header that shows the Enterprise Miner settings, process flow diagram, and detailed information for each node. Based on the Nodes property setting, each node that is included in the open process flow diagram has a header, property settings, and a variable summary. Moreover, the report also includes results such as variable selection, model diagnostic tables, and plots from the Results browser. Score code, log, and output listing are not included in the report. Those items are found in the Enterprise Miner package folder.
SAS Code	Use the SAS Code node to incorporate new or existing SAS code into process flows that you develop using Enterprise Miner. The SAS Code node extends the functionality of Enterprise Miner by making other SAS procedures available in your data mining analysis. You can also write a SAS DATA step to create customized scoring code, to conditionally process data, and to concatenate or to merge existing data sets. See Chapter 6.

Usage Rules for Nodes

Here are some general rules that govern the placement of nodes in a process flow diagram:

- The Input Data Source node cannot be preceded by any other nodes.
- All nodes except the Input Data Source and SAS Code nodes must be preceded by a node that exports a data set.
- The SAS Code node can be defined in any stage of the process flow diagram. It does not require an input data set that is defined in the Input Data Source node.
- The Model Comparison node must be preceded by one or more modeling nodes.
- The Score node must be preceded by a node that produces score code. For example, the modeling nodes produce score code.
- The Ensemble node must be preceded by a modeling node.
- The Replacement node must follow a node that exports a data set, such as a Data Source, Sample, or Data Partition node.

Overview of the SAS Enterprise Miner 5.3 Getting Started Example

This book uses an extended example that is intended to familiarize you with the many features of Enterprise Miner. Several key components of the Enterprise Miner process flow diagram are covered.

In this step-by-step example you learn to do basic tasks in Enterprise Miner: you create a project and build a process flow diagram. In your diagram you perform tasks

such as accessing data, preparing the data, building multiple predictive models, comparing the models, selecting the best model, and applying the chosen model to new data (known as scoring data). You also perform tasks such as filtering data, exploring data, and transforming variables. The example is designed to be used in conjunction with Enterprise Miner software.

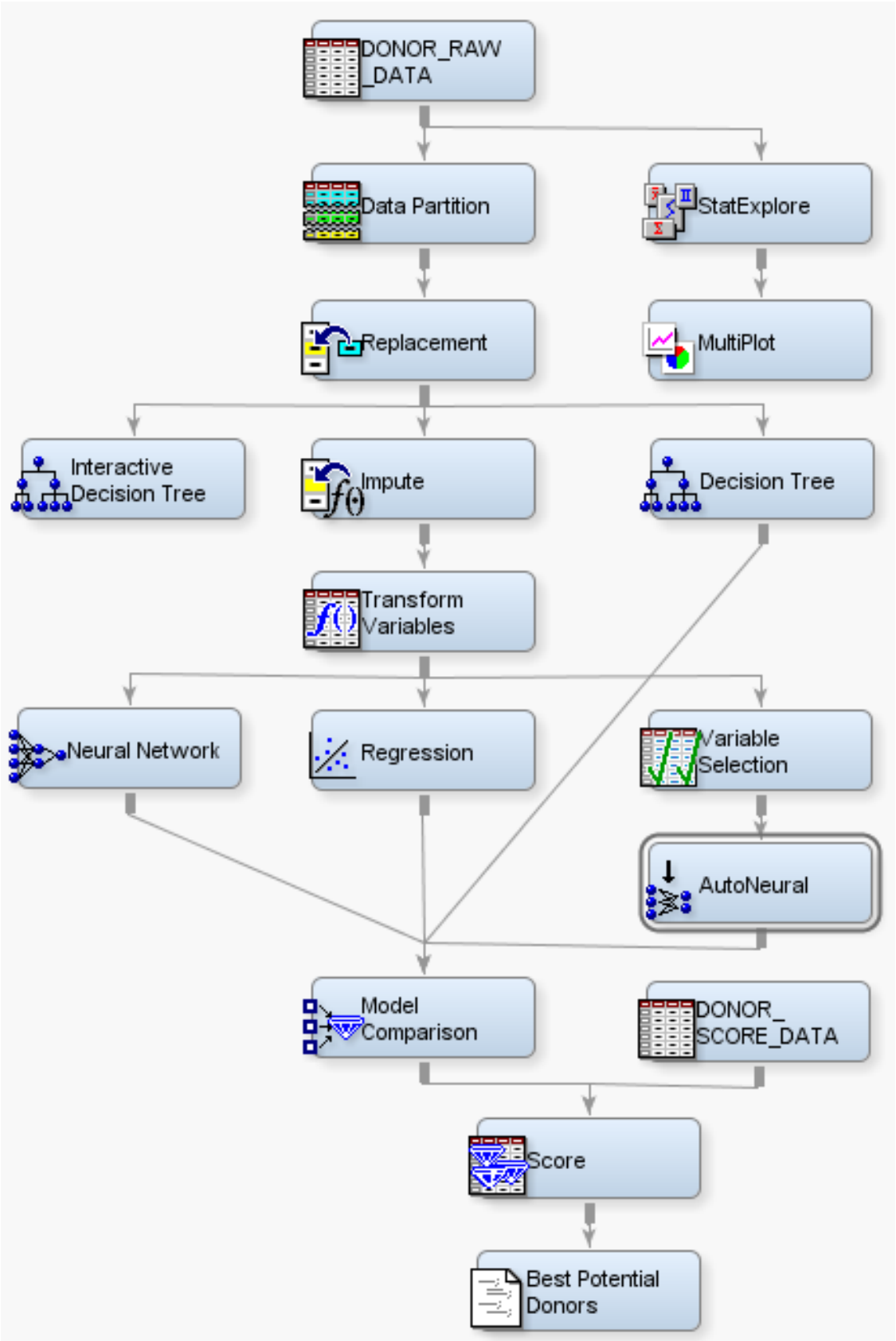
Example Problem Description

A national charitable organization seeks to better target its solicitations for donations. By only soliciting the most likely donors, less money will be spent on solicitation efforts and more money will be available for charitable concerns. Solicitations involve sending a small gift to an individual along with a request for a donation. Gifts include mailing labels and greeting cards.

The organization has more than 3.5 million individuals in its mailing database. These individuals have been classified by their response to previous solicitation efforts. Of particular interest is the class of individuals who are identified as lapsing donors. These individuals have made their most recent donation between 12 and 24 months ago. The organization has found that by predicting the response of this group, they can use the model to rank all 3.5 million individuals in their database. The campaign refers to a greeting card mailing sent in June of 1997. It is identified in the raw data as the 97NK campaign.

When the most appropriate model for maximizing solicitation profit by screening the most likely donors is determined, the scoring code will be used to create a new score data set that is named Donor.ScoreData. Scoring new data that does not contain the target is the end result of most data mining applications.

When you are finished with this example, your process flow diagram will resemble the one shown below.


DONOR_RAW_DATA
Data Partition
StatExplore
Replacement
MultiPlot
Interactive Decision Tree
Impute
Decision Tree
Transform Variables
Neural Network
Regression
Variable Selection
AutoNeural
Model Comparison
DONOR_SCORE_DATA
Score
Best Potential Donors

Here is a preview of topics and tasks in this example:

Chapter	Task
2	Create your project, define the data source, configure the metadata, define prior probabilities and profit matrix, and create an empty process flow diagram.
3	Define the input data, explore your data by generating descriptive statistics and creating exploratory plots. You will also partition the raw data and replace missing data.
4	Create a decision tree and interactive decision tree models.
5	Impute missing values and create variable transformations. You will also develop regression, neural network, and autoneural models. Finally, you will use the variable selection node.
6	Assess and compare the models. Also, you will score new data using the models.
7	Create model results packages, register your models, save and import the process flow diagram in XML.

Note: This example provides an introduction to using Enterprise Miner in order to familiarize you with the interface and the capabilities of the software. The example is not meant to provide a comprehensive analysis of the sample data. △

Software Requirements

In order to re-create this example, you must have access to SAS Enterprise Miner 5.3 software, either as client/server application, or as a complete client on your local machine.