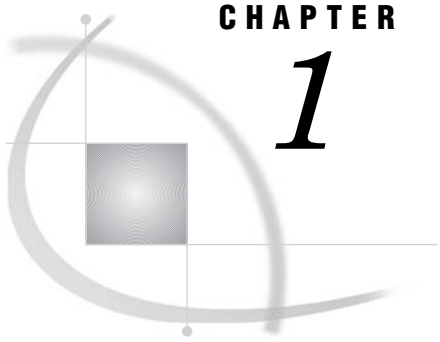


CHAPTER

1

Introduction to SAS Enterprise Miner 5.2 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 GUI 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 5.2 contains a collection of sophisticated analysis tools that have a common user-friendly interface that you can use to create and compare multiple models. Statistical tools include clustering, self-organizing maps / Kohonen, variable selection, trees, linear and logistic regression, and neural networking. Data preparation tools include outlier detection, variable transformations, 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.

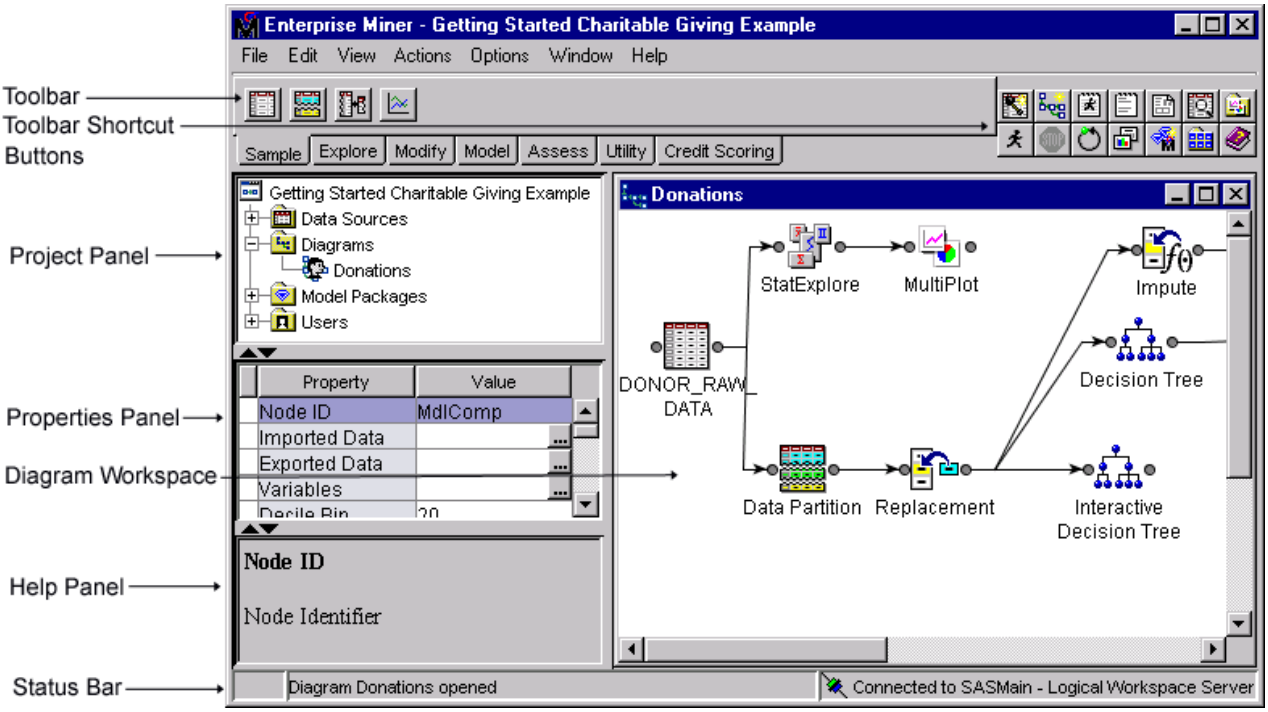
Layout of the Enterprise Miner Window

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



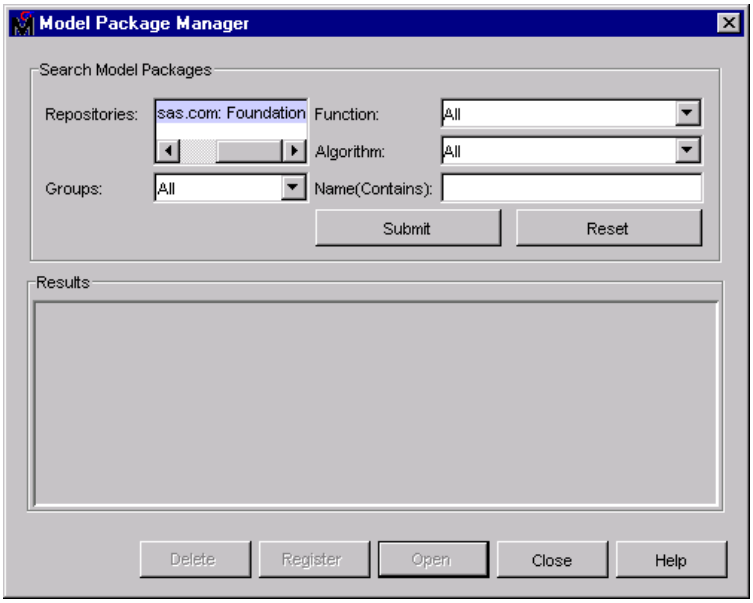
The Enterprise Miner window contains the following interface components:

- Toolbar and Toolbar shortcut buttons — The Enterprise Miner Toolbar is a graphic set of node icons that are organized by SEMMA categories. To the right side of 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 or tool 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, model packages, and users.
- 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.
- Help Panel — The 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.
  - 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.
  - 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.
  - Print Diagram — prints the contents of the window that is open in the Diagram Workspace.
  - Exit — ends the Enterprise Miner session and closes the window.
- Edit
  - Cut — deletes the selected item and copies it to the clipboard.

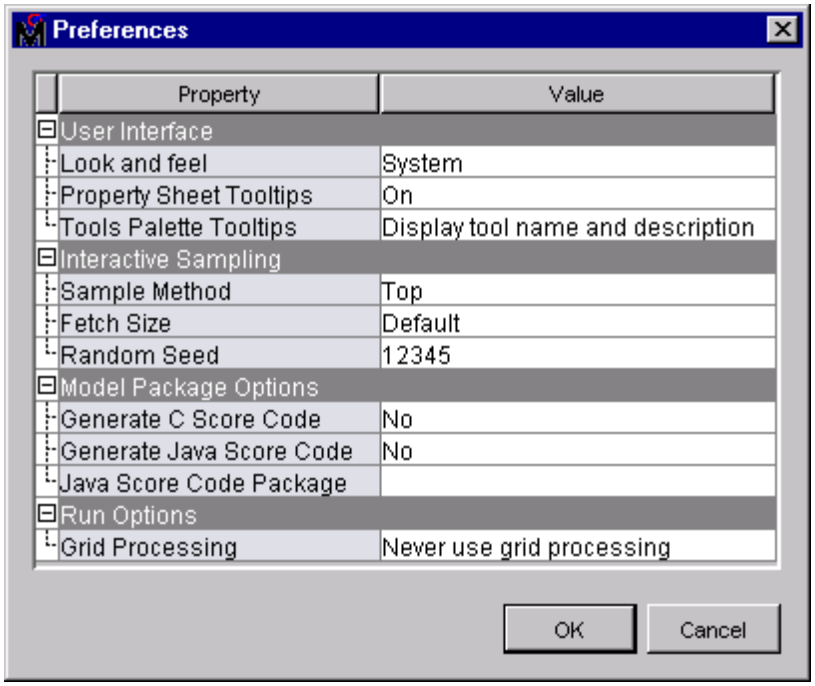
- Copy — copies the selected node to the clipboard.
- Paste — pastes a copied object from the clipboard.
- Delete — deletes the selected diagram, data source, or node.
- Rename — renames the selected diagram, data source, or node.
- Duplicate — creates a copy of the selected data source.
- Select All — 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.

- View
  - Property Sheet
    - Basic — displays the basic properties in the Properties Panel.
    - Advanced — displays the basic and advanced properties in the Properties Panel.
    - Hide — removes the Properties Panel and the Help Panel from the user interface.
  - 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.
  - Graphs — opens the Graphs window. Graphs that you create with SAS code in the Program Editor are displayed in this window.
  - Table — opens a table from the libraries that you have defined. You select a table from the Select a SAS Table 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.
  - 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 tools palette.
    - 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.
    - 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. By default, this option is selected.
    - Generate Java Score Code — creates Java score code when you create a report. By default, this option is selected. If you select **`Generate Java`**

Score Code, then 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 Nodes** — creates an orderly 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.

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.

All of the Enterprise Miner nodes are listed in a set of folders that are located on the **Tools** tab of the Enterprise Miner Project Navigator. The nodes are listed under the folder that corresponds to their data mining functions.

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, give each node's primary purpose, and supply examples and illustrations.

Sample Nodes

Node Name	Description
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.
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 user defined partitions to create partitioned data sets. See Chapter 3.