CHAPTER

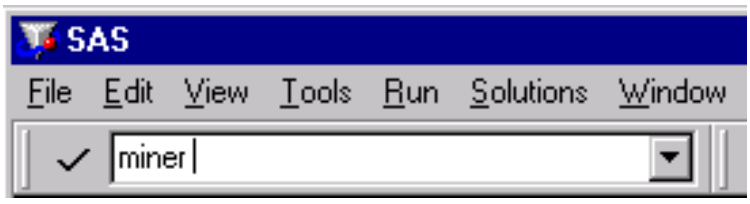
1

Introduction to SAS Enterprise Miner

Starting Enterprise Miner 1
Setting Up the Initial Project and Diagram 2
Identifying the Interface Components 3
Data Mining and SEMMA 4
Definition of Data Mining 4
Overview of the Data 4
Predictive and Descriptive Techniques 5
Overview of SEMMA 5
Overview of the Nodes 6
Sample Nodes 6
Explore Nodes 7
Modify Nodes 9
Model Nodes 11
Assess Nodes 13
Scoring Nodes 14
Utility Nodes 14
Some General Usage Rules for Nodes 15
Accessing SAS Data through SAS Libraries 16

Starting Enterprise Miner

To start Enterprise Miner, start SAS and then type **`miner`** on the SAS command bar. Submit the command by pressing the Return key or by clicking the check mark icon next to the command bar.



Alternatively, select from the main menu

Solutions ▶ Analysis ▶ Enterprise Miner

For more information, see *Getting Started with SAS Enterprise Miner*.

Setting Up the Initial Project and Diagram

Enterprise Miner organizes data analyses into projects and diagrams. Each project may have several process flow diagrams, and each diagram may contain several analyses. Typically each diagram contains an analysis of one data set.

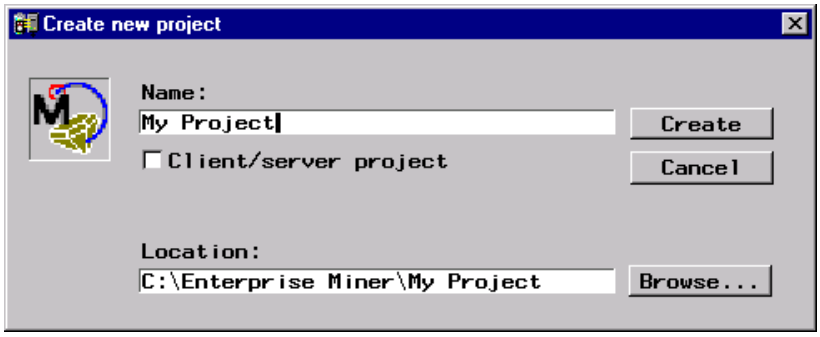
Follow these steps to create a project.

1 From the SAS menu bar, select

File ▶ New ▶ Project

2 Type a name for the project, such as My Project.

3 Select the **`Client/server project`** check box if necessary.



Note: You must have the access to a server that runs the same version of Enterprise Miner. For information about building a client/server project, see *Getting Started with SAS Enterprise Miner* or the online Help. △

4 Modify the location of the project folder by either typing a different location in the **`Location`** field or by clicking Browse.

5 Click Create. The project opens with an initial untitled diagram.

6 Select the diagram title and type a new name, such as My First Flow.

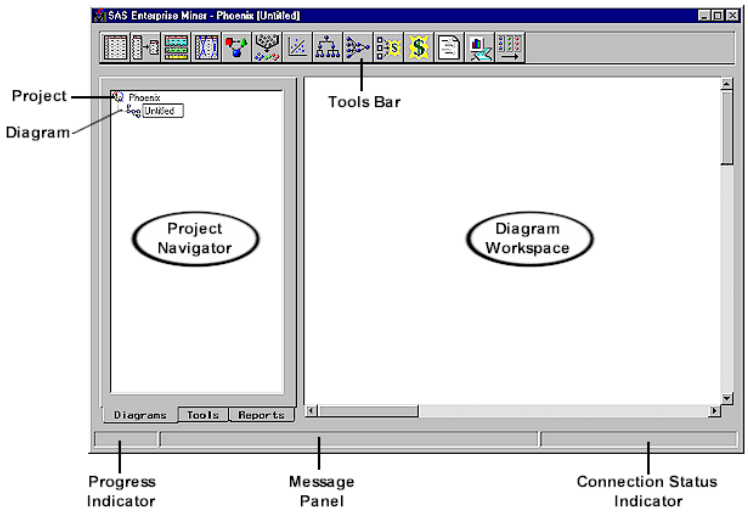
After selecting diagram title

Final Appearance

Identifying the Interface Components

The SAS Enterprise Miner window contains the following interface components:

The SAS Enterprise Miner v.4 Window



- Project Navigator — enables you to manage projects and diagrams, add tools to the Diagram Workspace, and view HTML reports that are created by the Reporter node. Note that when a tool is added to the Diagram Workspace, the tool is referred to as a node. The Project Navigator has three tabs:
  - Diagrams tab — lists the current project and the diagrams within the project. By default, the project window opens with the Diagrams tab activated.
  - Tools tab — contains the Enterprise Miner tools palette. This tab enables you to see all of the tools (or nodes) that are available in Enterprise Miner. The tools are grouped according to the SEMMA data-mining methodology.

    Many of the commonly used tools are shown on the Tools Bar at the top of the window. You can add additional tools to the Tools Bar by dragging them from the Tools tab onto the Tools Bar. In addition, you can rearrange the tools on the Tools Bar by dragging each tool to a new location on the Tools Bar.
  - Reports tab — displays the HTML reports that are generated by using the Reporter node.
- Diagram Workspace — enables you to build, edit, run, and save process flow diagrams.
- Tools Bar — contains a customizable subset of Enterprise Miner tools that are commonly used to build process flow diagrams in the Diagram Workspace. You can add or delete tools from the Tools Bar.
- Progress Indicator — displays a progress indicator bar that indicates the execution status of an Enterprise Miner task.
- Message Panel — displays messages about the execution of an Enterprise Miner task.
- Connection Status Indicator — displays the remote host name and indicates whether the connection is active for a client/server project.

Data Mining and SEMMA

Definition of Data Mining

This document defines *data mining* as advanced methods for exploring and modeling relationships in large amounts of data.

Overview of the Data

Your data often comes from several different sources, and combining information from these different sources may present quite a challenge. The need for better and quicker access to information has generated a great deal of interest in building data warehouses that are able to quickly assemble and deliver the needed information in usable form. To download documentation that discusses the Enterprise Miner add-ins to SAS/Warehouse Administrator, go to the SAS Customer Support Center Web site (**`http://support.sas.com`**). From Software Downloads, select **`Product and Solution Updates`**. From the Demos and Downloads page, select **`SAS/Warehouse Administrator Software`**, and download the version that you want.

A typical data set has many thousand observations. An observation may represent an entity such as an individual customer, a specific transaction, or a certain household. Variables in the data set contain specific information such as demographic information, sales history, or financial information for each observation. How this information is used depends on the research question of interest.

When talking about types of data, consider the measurement level of each variable. You can generally classify each variable as one of the following:

- interval — a variable for which the mean (or average) makes sense, such as average income or average temperature.
- categorical — a variable consisting of a set of levels, such as gender (male or female) or drink size (small, regular, large). In general, if the variable is not continuous (that is, if taking the average does not make sense, such as average gender), then it is categorical. Categorical data can be grouped in several ways. For the purposes of Enterprise Miner, consider these subgroupings of categorical variables:
  - unary — a variable that has the same value for every observation in the data set.
  - binary — a variable that has only two possible levels. Gender is an example.
  - nominal — a variable that has more than two levels, but the values of the levels have no implied order. Pie flavors such as cherry, apple, and peach are examples.
  - ordinal — a variable that has more than two levels, and the values of the levels have an implied order. Drink sizes such as small, regular, and large are examples.

    Note: Ordinal variables may be treated as nominal variables, if you are not interested in the ordering of the levels. However, nominal variables cannot be treated as ordinal variables since there is no implied ordering by definition. △

  Missing values are not included in the counts.

To obtain a meaningful analysis, you must construct an appropriate data set and specify the correct measurement level for each of the variables.

Predictive and Descriptive Techniques

Predictive modeling techniques enable you to identify whether a set of input variables is useful in predicting some outcome variable. For example, a financial institution may try to determine if knowledge of an applicant's income and credit history (input variables) helps to predict whether the client is likely to default on a loan (outcome variable).

To distinguish the input variables from the outcome variables, set the model role for each variable in the data set. Identify outcome variables by using the target model role, and identify input variables by using the input model role. Examples of model roles include cost, freq, ID, and input. If you want to exclude some of the variables from the analysis, identify these variables by using the rejected model role. Specify a variable as an ID variable by using the ID model role.

Predictive modeling techniques require one or more outcome variables of interest. Each technique attempts to predict the outcome as well as possible according to some criteria such as maximizing accuracy or maximizing profit. This document shows you how to use several predictive modeling techniques through Enterprise Miner including regression models, decision trees, and neural networks. Each of these techniques enables you to predict a binary, nominal, ordinal, or continuous outcome variable from any combination of input variables.

Descriptive techniques enable you to identify underlying patterns in a data set. These techniques do not have a specific outcome variable of interest. This document explores how to use Enterprise Miner to perform the following descriptive analyses:

- Cluster analysis: This analysis attempts to find natural groupings of observations in the data, based on a set of input variables. After grouping the observations into clusters, you can use the input variables to try to characterize each group. When the clusters have been identified and interpreted, you can decide whether to treat each cluster independently.
- Association analysis: This analysis identifies groupings of products or services that tend to be purchased at the same time or at different times by the same customer. The analysis answers questions such as
  - What proportion of the people who purchased eggs and milk also purchased bread?
  - What proportion of the people who have a car loan with some financial institution later obtain a home mortgage from the same institution?

Overview of SEMMA

Enterprise Miner nodes are arranged into the following categories according the SAS process for data mining: SEMMA.

- Sample — identify input data sets (identify input data; sample from a larger data set; partition data set into training, validation, and test data sets).
- Explore — explore data sets statistically and graphically (plot the data, obtain descriptive statistics, identify important variables, perform association analysis).
- Modify — prepare the data for analysis (create additional variables or transform existing variables for analysis, identify outliers, replace missing values, modify the way in which variables are used for the analysis, perform cluster analysis, analyze data with self-organizing maps (known as SOMs) or Kohonen networks).
- Model — fit a predictive model (model a target variable by using a regression model, a decision tree, a neural network, or a user-defined model).

- Assess — compare competing predictive models (build charts that plot the percentage of respondents, percentage of respondents captured, lift, and profit).

The Score and Score Converter nodes form another group, Score, and are designed to capture score code for the models and to translate the SAS DATA step score code into the C and Java programming languages. The SAS DATA step score code can be saved as a SAS program outside Enterprise Miner. The SAS program can then be run on any platform that runs base SAS. Thus, you can perform the actual scoring on almost any type of platform. Code that is based on the C or Java languages can be integrated into standalone C or Java programs that operate outside SAS.

Additional tools are available under the Utility nodes group.

Overview of the Nodes

Sample Nodes

The Input Data Source node reads data sources and defines their attributes for later processing by Enterprise Miner. This node can perform various tasks:

- access SAS data sets and data marts. Data marts can be defined by using the SAS Data Warehouse Administrator, and they can be set up for Enterprise Miner by using the Enterprise Miner Warehouse Add-ins.
- automatically create a metadata sample for each variable when you import a data set with the Input Data Source node. By default, Enterprise Miner obtains the metadata sample by taking a random sample of 2,000 observations from the data set that is identified in the Input Data Source. Optionally, you can request larger samples. If the data is smaller than 2,000 observations, the entire data set is used
- use the metadata sample to set initial values for the measurement level and the model role for each variable. You can change these values if you are not satisfied with the automatic selections that are made by the node.
- display summary statistics for interval and class variables.
- define target profiles for each target in the input data set.

Note: This document uses the term *data sets* instead of *data tables*. △

The Sampling node enables you to perform random sampling, stratified random sampling, and cluster sampling. Sampling is recommended for extremely large databases because it can significantly decrease model-training time. If the sample is sufficiently representative, relationships that are found in the sample can be expected to

generalize to the complete data set. The Sampling 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. You can replicate the samples by using the same seed value.

The Data Partition node enables you 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 data set that you can use for model assessment. This node uses simple random sampling, stratified random sampling, or a user-defined partition to create training, test, or validation data sets. Specify a user-defined partition if you have determined which observations should be assigned to the training, validation, or test data sets. This assignment is identified by a categorical variable that is in the raw data set.

Explore Nodes

The Distribution Explorer node enables you to explore large volumes of data in multidimensional histograms. You can view the distribution of up to three variables at a time with this node. When the variable is binary, nominal, or ordinal, you can select specific values to exclude from the chart. To exclude extreme values for interval variables, you can set a range cutoff. The node also generates simple descriptive statistics for the interval variables.

The Multiplot node enables you to explore large volumes of data graphically. Unlike the Insight or Distribution Explorer nodes, the Multiplot node automatically creates bar charts and scatter plots for the input and target variables without making several menu or window item selections. The code that is created by this node can be used to create graphs in a batch environment, whereas the Insight and Distribution Explorer nodes must be run interactively.

Insight

The Insight node enables you to open a SAS/INSIGHT session. SAS/INSIGHT software is an interactive tool for data exploration and analysis. With it, you explore samples of data through graphs and analyses that are linked across multiple windows. You can analyze univariate distributions, investigate multivariate distributions, and fit explanatory models by using generalized linear models.

Association

The Association node enables you to identify association relationships within the data. For example, if a customer buys a loaf of bread, how likely is the customer to buy a gallon of milk as well? The node also enables you to perform sequence discovery if a time-stamp variable (a sequence variable) is present in the data set.

Variable Selection

The Variable Selection node enables you to evaluate the importance of input variables in predicting or classifying the target variable. To select the important inputs, the node uses either an R-square or a Chi-square selection (tree-based) criterion. The R-square criterion enables you to remove variables that have large percentages of missing values, remove class variables that are based on the number of unique values, and remove variables in hierarchies. Variables can be hierarchical because of levels of generalization (Zipcode generalizes to State, which generalizes to Region) or because of formulation (variable A and variable B may have interaction A*B). 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 more detailed modeling nodes, such as the Neural Network and Tree nodes. Certain variables of interest may be rejected by a variable selection technique, but you can force these variables into the model by reassigning the input model role to these variables in any modeling node.

Link Analysis

The Link Analysis node enables you to tranform data from different sources into a data model that can be graphed. The data model supports simple statistical measures,

presents a simple interactive graph for basic analytical exploration, and generates cluster scores from raw data. The scores can be used for data reduction and segmentation.

Modify Nodes

The Data Set Attributes node enables you to modify data set attributes, such as data set names, descriptions, and roles. You can also use this node to modify the metadata sample that is associated with a data set and to specify target profiles for a target. An example of a useful Data Set Attributes application is to generate a data set in the SAS Code node and then modify its metadata sample with this node.

The Transform Variables node enables you to transform variables; for example, you can transform variables by taking the square root of a variable, by taking the natural logarithm, maximizing the correlation with the target, or normalizing a variable. Additionally, the node supports user-defined formulas for transformations and enables you to group interval-valued variables into buckets or quantiles. This node also automatically places interval variables into buckets by using a decision tree-based algorithm. Transforming variables to similar scale and variability may improve the fit of models and, subsequently, the classification and prediction precision of fitted models.

The Filter Outliers node enables you to identify and remove outliers from data sets. Checking for outliers is recommended, as outliers may greatly affect modeling results and, subsequently, the classification and prediction precision of fitted models.

Replacement

The Replacement node enables you 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, or distribution-based replacement, or 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.

Clustering

The Clustering node enables you to segment your data; that is, it enables you to identify data observations that are similar in some way. 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. It can also be passed as a group variable that enables you to automatically construct separate models for each group.

SOM/Kohonen

The SOM/Kohonen node generates self-organizing maps, Kohonen networks, and vector quantization networks. Essentially the node performs unsupervised learning in which it attempts to learn the structure of the data. As with the Clustering node, after the network maps have been created, the characteristics can be examined graphically by using the Results browser. The node provides the analysis results in the form of an interactive map that illustrates the characteristics of the clusters. Furthermore, it provides a report that indicates the importance of each variable.

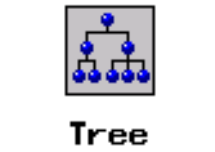
Time Series

The Time Series node enables you to convert transactional data to time series data. It also performs seasonal and trend analysis on time-stamped transactional data.

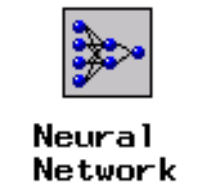
The Interactive Grouping node enables you to interactively group variable values into classes. Statistical and plotted information can be interactively rearranged as you explore various variable groupings. The Interactive Grouping node requires a binary target variable.

Model Nodes

The Regression node enables you to fit both linear and logistic regression models to your data. You can use both continuous and discrete variables as inputs. The node supports the stepwise, forward, and backward-selection methods. A point-and-click interaction builder enables you to create higher-order modeling terms.



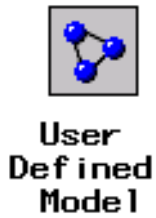
The Tree node enables you to perform multiway splitting of your database, based on nominal, ordinal, and continuous variables. This is the SAS implementation of decision trees, which represents a hybrid of the best of CHAID, CART, and C4.5 algorithms. The node supports both automatic and interactive training. When you run the Tree node in automatic mode, it automatically ranks the input variables by the strength of their contribution to the tree. This ranking can be used to select variables for use in subsequent modeling. In addition, dummy variables can be generated for use in subsequent modeling. Using interactive training, you can override any automatic step by defining a splitting rule or by pruning a node or subtree.



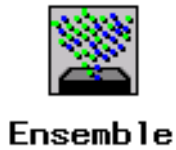
The Neural Network node enables you to construct, train, and validate multilayer feed-forward neural networks. By default, the Neural Network node automatically constructs a multilayer feed-forward network that has one hidden layer consisting of three neurons. In general, each input is fully connected 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.

The Princomp/Dmneural node enables you to fit an additive nonlinear model that uses the bucketed principal components as inputs to predict a binary or an interval target variable. The node also performs a principal components analysis and passes the principal components to the successor nodes.



The User Defined Model node enables you to generate assessment statistics by using predicted values from a model that you built with the SAS Code node (for example, a logistic model that uses the SAS/STAT LOGISTIC procedure) or the Variable Selection node. You can also generate assessment statistics for models that are built by a third-party software product when you create a SAS data set that contains the predicted values from the model. The predicted values can also be saved to a SAS data set and then imported into the process flow with the Input Data Source node.



The Ensemble node creates a new model by averaging the posterior probabilities (for class targets) or the predicted values (for interval targets) from multiple models. The new model is then used to score new data. One common approach is to resample the training data and fit a separate model for each sample. The Ensemble node then integrates the component models to form a potentially stronger solution. Another common approach is to use multiple modeling methods, such as a neural network and a decision tree, to obtain separate models from the same training data set.

The Ensemble node integrates the component models from the complementary modeling methods to form the final model solution. The Ensemble node can also be used to combine the scoring code from stratified models. The modeling nodes generate different scoring formulas when they operate on a stratification variable (for example, a group variable such as Gender) that you define in a Group Processing node. The Ensemble node combines the scoring code into a single DATA step by logically dividing the data into blocks by using IF-THEN DO/END statements.

It is important to note that the ensemble model that is created from either approach can be more accurate than the individual models only if the individual models differ.

The Memory-Based Reasoning node is a modeling tool that uses a k-nearest neighbor algorithm to categorize or predict observations.

The Two Stage Model node computes a two-stage model to predict a class target and an interval target. The interval target variable is usually a value that is associated with a level of the class target variable.

Assess Nodes

The Assessment node provides a common framework for comparing models and predictions from any of the modeling nodes (Regression, Tree, Neural Network, and User Defined Model nodes). 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.

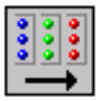
The Reporter node assembles the results from a process flow analysis into an HTML report that can be viewed with a Web browser. Each report contains header information, an image of the process flow diagram, and a separate report for each node in the flow including node settings and results. Reports are managed in the Reports tab of the Project Navigator.

Scoring Nodes

The Score node enables you to generate and manage predicted values from a trained model. Scoring formulas are created for both assessment and prediction. Enterprise Miner generates and manages scoring formulas in the form of SAS DATA step code, which can usually be used in SAS even without the presence of Enterprise Miner.

The Score Converter node provides scored data mining output in both the C and Java languages. The choices of language output enable you to use Enterprise Miner output in programs that operate outside SAS.

Utility Nodes



The Group Processing node enables you to perform an analysis for each level of a class variable such as Gender. You can also use this node to specify multiple targets or process the same data source repeatedly. When multiple targets are selected, Enterprise Miner analyzes each target separately.

The Data Mining Database node enables you to create a data mining database (DMDB) for batch processing. For nonbatch processing, Enterprise Miner automatically creates DMDBs as they are needed.

The SAS Code node enables you to incorporate new or existing SAS code into process flow diagrams. You can also use a SAS DATA step to create customized scoring code, to conditionally process data, and to concatenate or to merge existing data sets. The node provides a macro facility to dynamically reference data sets (used for training, validation, testing, or for scoring) and variables, such as input, target, and predict variables. After you run the SAS Code node, you can then export the results and the data sets for use by subsequent nodes in the diagram.

The Control Point node enables you to establish a control point to reduce the number of connections that are made in process flow diagrams. For example, suppose that you want to connect three Input Data Source nodes to three modeling nodes. If you omit the Control Point node, then you need nine connections to connect all of the Input Data Source nodes to all of the modeling nodes. However, if you use the Control Point node, you need only six connections.

The Subdiagram node enables you to group a portion of a process flow diagram into a subdiagram. For complex process flow diagrams, you may want to create subdiagrams to better design and control the process flow.

Some General Usage Rules for Nodes

These are some general rules that govern placing nodes in a process flow diagram:

- The Input Data Source cannot be preceded by any other node.
- The Sampling node must be preceded by a node that exports a data set.
- The Assessment node must be preceded by one or more model nodes.
- The Score node or Score Converter node must be preceded by a node that produces score code. Any node that modifies the data or builds models generates score code.
- The SAS Code node can be defined in any stage of the process flow diagram. It does not require an input data set to be defined in the Input Data Source node.

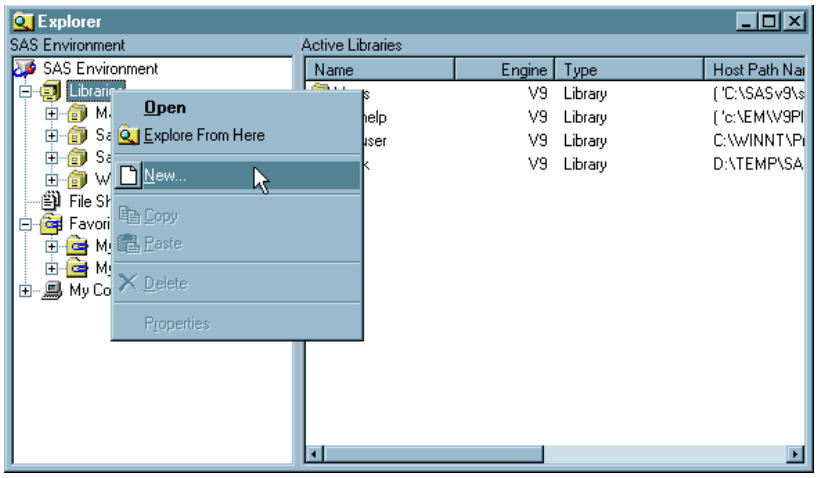
Accessing SAS Data through SAS Libraries

SAS uses libraries to organize files. These libraries point to folders where data and programs are stored. In Enterprise Miner, Release 4.2, libraries must conform to the SAS Version 8 naming conventions. These conventions require the library name to have no more than eight alphanumeric characters. The first character of the name must be a letter or an underscore (_). Subsequent characters can be characters, numeric digits, and underscores. The name cannot contain special characters such as asterisks (*) and ampersands (&). For more information, see Names in the SAS Language in SAS Help and Documentation.

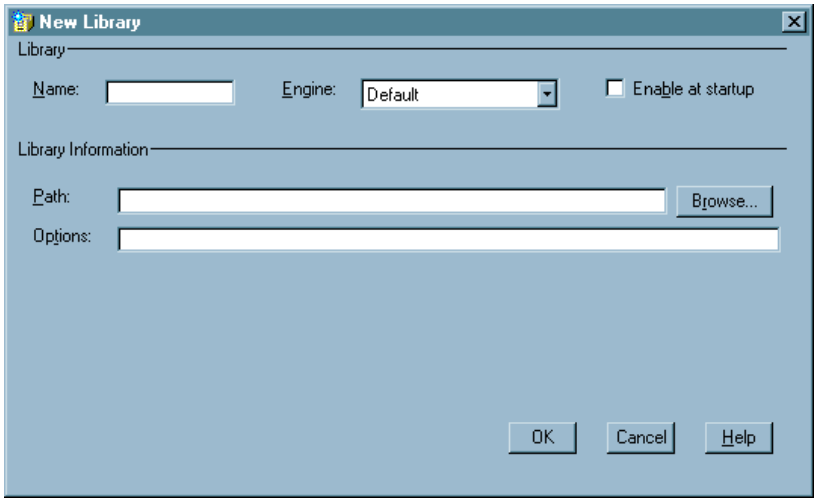
To create a new library or to view existing libraries, select from the main SAS menu

View ▶ Explorer

The following display shows an example of the Explorer window.



You can view the files in a library by selecting the library name from the list of libraries in the left panel of the Explorer window. To create a new library, right-click **Libraries** and select **New**. The New Library window opens.



Specify the library name, engine, associated path, and options.

In the display, the **Enable at Startup** check box is not selected. This library will not be reassigned every time that the SAS session starts. Select the check box if you want SAS to automatically assign the library each time SAS starts.

Several libraries are automatically assigned when you start Enterprise Miner. One of these libraries (SAMPSIO) contains sample data sets that are used in Enterprise Miner reference material. This document uses the data sets that are in SAMPSIO. Any data set in the library can then be referenced by the two-part name that is constructed by using the SAS library name and the SAS data set name. For example, the HMEQ data set in the SAMPSIO library is identified by the two-part name SAMPSIO.HMEQ.