Chapter 1: Introduction

1.1 Getting Started

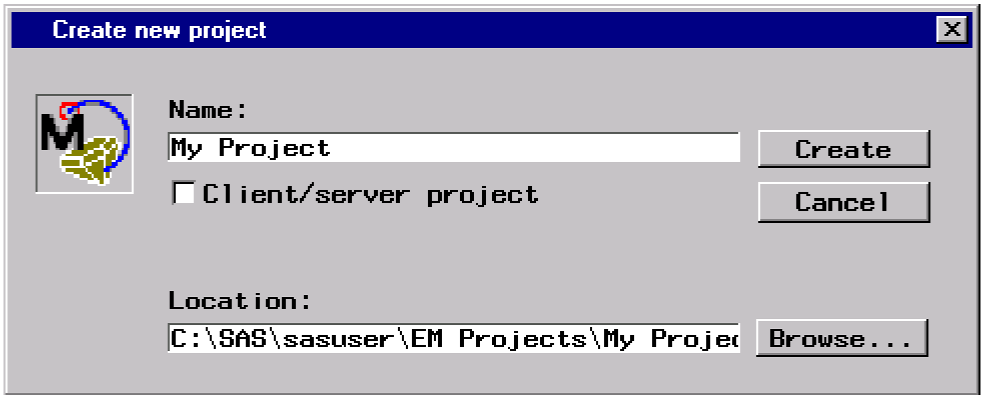
Opening Enterprise Miner

To start Enterprise Miner, double-click on the Enterprise Miner icon on your desktop. If no icon is available and you are running on Windows, use the Start menu and select **Start → Programs → Enterprise Miner → Enterprise Miner *X.x*,** where *X.x* represents the version number of Enterprise Miner that you have installed on your machine.

Setting Up the Initial Project and Diagram

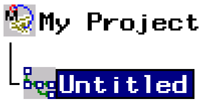
Enterprise Miner organizes analyses into projects and diagrams. Each project may have several diagrams, and each diagram may contain several analyses. Typically each diagram contains an analysis of one data set.

1. Select **File → New → Project**.
2. Type in the name of the project (for example, `My Project`).
3. Check the box for **Client/server project** if necessary.
   Note: You must have the access to a server that runs the same version of Enterprise Miner. See *Getting Started with the Enterprise Miner* or the online help if you want to build a client/server project.



Name:
My Project
□ Client/server project
Location:
C:\SAS\sasuser\EM Projects\My Projec

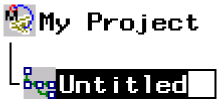
4. Modify the location of the project folder if you want by selecting **Browse**.
5. Select **Create**. The project opens with an initial untitled diagram.



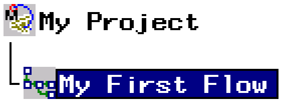
My Project
Untitled

6. Click on the diagram title and type in a new title (for example, `My First Flow`).

After Selecting Name

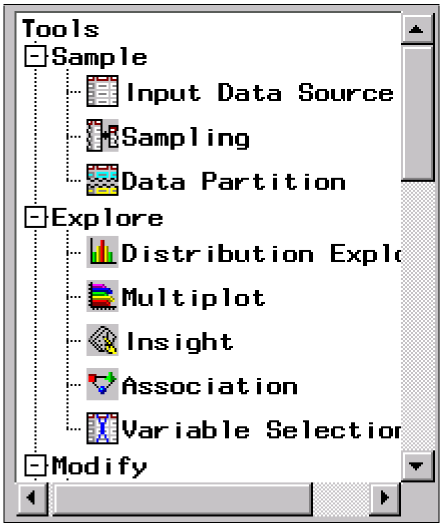


Final Appearance



Identifying the Workspace Components

7. Observe that the project window opens with the Diagrams tab activated. Select the **Tools** tab that is located to the right of the Diagrams tab in the lower-left portion of the project window. This tab enables you to see all of the tools (or nodes) that are available in Enterprise Miner.



Many of the commonly used tools are shown on the toolbar at the top of the window. You can add additional tools to this toolbar by dragging them from the window above onto the toolbar. In addition, you can rearrange the tools on the toolbar by dragging each tool to a new location on the bar.

8. Select the **Reports** tab that is located to the right of the Tools tab. This tab displays any reports that have been generated for this project. This is a new project, so no reports are currently available.

9. Return to the Tools tab.

1.2 Data Mining Using SEMMA

Definition of Data Mining

There are many techniques that can be grouped under the name *Data Mining*. SAS Institute defines data mining as "advanced methods for exploring and modeling relationships in large amounts of data."

Overview of the Data

The 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. A typical data set has many thousand observations. An observation may represent an individual customer, a specific transaction, or a certain household. The data set also contains specific information (or variables) about each observation such as demographic information, sales history, or financial information. How this information is used depends on the research question of interest.

When talking about types of data, it is important to consider the *measurement level* of each variable. You can generally classify each variable into one of the following categories:

- *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 (taking the average does not make sense, such as *average gender*), then it is categorical. Categorical data can be split up 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 (i.e. gender).
  - *nominal* - a variable that has more than two levels, but the ordering of the levels has no implied order (i.e., pie flavors - cherry, apple, and peach).
  - *ordinal* - a variable that has more than two levels, and the ordering of the levels has an implied order (i.e., drink size - small, regular, large). Note: Ordinal variables may be treated as nominal variables (i.e., if you are not interested in the ordering of the levels), but nominal variables cannot be treated as ordinal variables since there is no implied ordering by definition.

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

Overview of the Methods

Predictive Modeling Techniques enable the analyst 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. If you wish 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 book shows you how to use several predictive modeling techniques by using Enterprise Miner including
Regression Models, *Decision Trees*, and *Neural Networks*. Each of these techniques enables you to predict a binary, nominal, or continuous outcome variable from any combination of input variables. Decision Trees and Neural Networks can also model ordinal targets.

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

- *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, the analyst can use the input variables to try to characterize each group. Once the clusters have been identified and interpreted, the analyst may decide 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 would answer 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?

Understanding SEMMA

The tools are arranged according the SAS process for Data Mining, SEMMA.

SEMMA stands for

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 set 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 percentage of respondents, percentage of respondents captured, lift, and profit).

The Score node is grouped with the tools under **A**ssess and is designed to capture scoring code for the models that have been fit. The scoring 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.

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:

1. It enables you to 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.
2. It automatically creates 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.
3. It uses 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.
4. It displays summary statistics for interval and class variables.
5. It enables you to define target profiles for each target in the input data set.

Note: For the purposes of this document, **data sets** and **data tables** are equivalent terms.

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 so that you may replicate the samples.

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 holdout 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, validation, or test data sets. Specifying a user-defined partition indicates that you have determined which observations should be assigned to the training, validation, or test data sets, and this assignment is identified by a categorical variable that is in the raw data set.

Explore Nodes

The Distribution Explorer node is a visualization tool that enables you quickly and easily 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 is another visualization tool that enables you to explore larger 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.

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.

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 also buy a gallon of milk? The node also enables you to perform sequence discovery if a time stamp variable (a sequence variable) is present in the data set.

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 a more detailed modeling node, 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 these variables the input model role in any modeling node.

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 provides a visual interface for grouping 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.

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.

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.

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.

Model Nodes

The Regression node enables you to fit both linear and logistic regression models to your data. You can use continuous, nominal, 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 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 System 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 may 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 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 once 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 ensemble 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 two 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 IF-THEN-DO/END blocks.

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 disagree with one another.

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 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 Reporter node assembles the results from a process flow analysis into an HTML report that can be viewed with your favorite 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.

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, DMDBs are automatically created as they are needed.

The SAS Code node enables you to incorporate new or existing SAS code into process flow diagrams. The ability to write SAS code enables you to include additional SAS System procedures into your data mining analysis. 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, the results and the data sets can then be exported 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 three Input Data Source 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 Source nodes to all of the modeling nodes. However, if a Control Point node is used, only six connections are required.

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 (PFD):

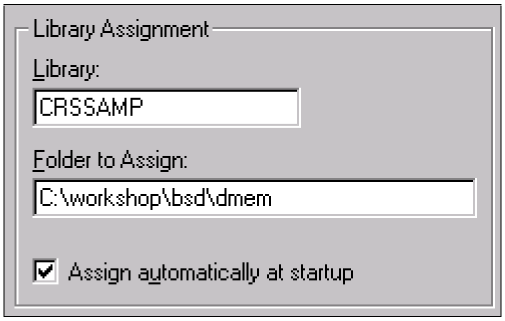
- The Input Data Source cannot be preceded by any other node.
- The Sampling node must be preceded by a node that exports a data set.
- The Assessment node must be preceded by one or more modeling nodes.
- The Score node must be preceded by a node that produces score code. Any node that modifies the 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.

1.3 Accessing Data in SAS Software

Using SAS Libraries

SAS uses libraries to organize files. These libraries point to folders where data and programs are stored. In Version 3 of Enterprise Miner, libraries must conform to the naming conventions that are used in Release 6.12. These conventions require the library name to have no more than eight alphanumeric characters, and the name cannot contain special characters such as asterisks (*) and ampersands (&). To create a new library or to view existing libraries, use the Globals menu and select **Access → Display libraries**.

You can see the files in a library by selecting the library name from the list of libraries in the upper-left portion of the dialog box. To create a new library, say CRSSAMP, select **New Library** and fill in the resulting dialog box with the desired library name and associated path. The following library identifies the folder whose path is `C:\workshop\bsd\dmem`.



Library Assignment

Library: CRSSAMP

Folder to Assign: C:\workshop\bsd\dmem

☑ Assign automatically at startup

Observe that the box for **Assign automatically at startup** is checked. This library will be reassigned every time that the SAS session starts. If you do not check this box, the library name is not automatically assigned when the SAS or Enterprise Miner session starts. As a result, the contents of the library will be unavailable for use by the SAS System or Enterprise Miner in future sessions unless you reassign the library name manually. Select **Assign** to finish assigning the library name.

There are several libraries that are automatically assigned when you open Enterprise Miner. One of these libraries (SAMPSIO) contains sample data sets that are used in Enterprise Miner reference material to illustrate important concepts. For the purposes of this document, assume that the data sets 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.