

# Data Mining Using SAS Enterprise Miner

Randall Matignon, Piedmont, CA

## An Overview of SAS Enterprise Miner

The following article is in regards to Enterprise Miner v.4.3 that is available in SAS v9.1.3. Enterprise Miner an awesome product that SAS first introduced in version 8. It consists of a variety of analytical tools to support data mining analysis. Data mining is an analytical tool that is used to solving critical business decisions by analyzing large amounts of data in order to discover relationships and unknown patterns in the data. The Enterprise Miner data mining SEMMA methodology is specifically designed to handling enormous data sets in preparation to subsequent data analysis. In SAS Enterprise Miner, the SEMMA acronym stands for Sampling, Exploring, Modifying, Modeling, and Assessing large amounts of data.

The reason that SAS Enterprise Miner has been given this acronym is that usually the first step in data mining is to sample the data in order to acquire a representative sample of the data. The next step is to usually explore the distribution or the range of values of each variable to the selected data set. This might be followed by modifying the data set by replacing missing values or transforming the data in order to achieve normality in the data since many of the various analytical tools depend on the variables having a normal distribution. The reason is because many of the nodes in Enterprise Miner calculate the square distances between the variables that are selected to the analysis. The next step might be to model the data. In other words, there might be interest in predicting certain variables in the data. The final steps might be to determine which models are best by assessing the accuracy between the different models that have been created.

## The Ease of Use to Enterprise Miner

SAS Enterprise Miner is a powerful new module introduced in version 8. But, more importantly SAS Enterprise Miner is very easy application to learn and very easy to use. SAS Enterprise Miner is visual programming with a GUI interface. The power of the SAS Enterprise Miner product is that you do not even need to know SAS programming and have very little statistical expertise in the development of your Enterprise Miner project since it is as simple as selecting icons or nodes from the EM tool palette or menu bar and dragging the icons onto the EM diagram workspace or desktop. Yet, an expert statistician can adjust and fine-tune the default settings and run the SEMMA process flow diagram to their own personal specifications. The nodes are then connected to one another in a graphical diagram workspace. SAS Enterprise Miner is visual programming with SAS icons within a graphical EM diagram workspace. It is as simple as dragging and dropping icons onto the EM diagram graphical workspace. The SAS Enterprise Miner diagram workspace environment looks similar to the desktop in Windows 95, 98, XP, and Vista. Enterprise Miner is very easy to use and can save a tremendous amount of time having to program in SAS. However, SAS Enterprise Miner has a powerful **SAS Code** node that brings in the capability of SAS programming into the SEMMA data mining process through the use of a SAS data step in accessing a wide range of the powerful SAS procedures into the SAS Enterprise Miner process flow diagram. Enterprise Miner produces a wide variety of statistics from descriptive, univariate, and goodness-of-fit statistics, numerous types of charts and plots, traditional regression modeling, decision tree analysis, principal component analysis, cluster analysis, association analysis, link analysis, along with automatically generated graphs that can be directed to the SAS output window.

## The Purpose of the Enterprise Miner Nodes

Data Mining is a sequential process of Sampling, Exploring, Modifying, Modeling, and Assessing large amounts of data to discover trends, relationships, and unknown patterns in the data. SAS Enterprise Miner is designed for SEMMA data mining. SEMMA stands for the following.

**S**ample – Identify the analysis data set with the data that is large enough to make significant findings, yet small enough to compile the code in a reasonable amount of time. The nodes create the analysis data set, randomly sample the source data set, or partition the source data set into a training, validation, and test data set.

**E**xplore – Explore the data sets to view the data set to observe for unexpected trends, relationships, patterns, or unusual observations while at the same time getting familiar with the data. The nodes plot the data, generate a wide variety of analysis, identify important variables, or perform association analysis.

**M**odify – Prepares the data for analysis. The nodes can create additional variables or transform existing variables for analysis by modifying or transforming the way in which the variables are used in the analysis, filter the data, replace missing values, condense and collapse the data in preparation to time series modeling, or perform cluster analysis.

**M**odel – Fits the statistical model. The nodes predicts the target variable against the input variables by using either least-squares or logistic regression, decision tree, neural network, dmneural network, user-defined, ensemble, nearest neighbor, or two-stage modeling.

**A**ssess – Compare the accuracy between the statistical models. The nodes compare the performance of the various classification models by viewing the competing probability estimates from the lift charts, ROC charts, and threshold charts. For predictive modeling designs, the performance of each model and the modeling assumptions can be verified from the prediction plots and diagnosis charts.

**Note:** Although, the **Utility** nodes are not a part of the SEMMA acronym, the nodes will allow you to perform group processing, create a data mining data set to view various descriptive statistics from the entire data set, and organize

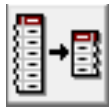
the process flow more efficiently by reducing the number of connections or condensing the process flow into smaller more manageable subdiagrams.

## The Enterprise Miner Nodes

### Sample Nodes



The purpose of the **Input Data Source** node is to read in a SAS data set or import and export other types of data through the SAS import Wizard. The **Input Data Source** node reads the data source and creates a data set called a *metadata sample* that automatically defines the variable attributes for later processing within the process flow. In the metadata sample, each variable is automatically assigned a level of measurement and a variable role assignment to the analysis. For example, categorical variables with more than two class levels and less than ten class levels are automatically assigned a nominal measurement level with an input variable role. By default, a metadata sample takes a random sample of 2,000 observations from the source data set. If the data set is smaller than 2,000 observations, then the entire data set is used to create the data mining data set. From the metadata sample, the node displays various summary statistics for both interval-valued and categorical-valued variables. For the interval-valued variables, it's important that the variables share the same range of values, otherwise various transformations such as standardizing, are recommended. The reason is because a large majority of the data mining analysis designs apply the squared distance between the data points. The node has the option of editing the target profile for categorical-valued target variables in order to assign prior probabilities to the categorical response levels that truly represent the appropriate level of responses in addition to predetermined profit and cost amounts for each target-specified decision consequences in order to maximum expected profit or minimize expected loss from the following statistical models.



The purpose of the **Sampling** node is to perform various sampling techniques to the input data set. Sampling is recommended for extremely large data sets to reduce both the memory resources and processing time to data mining. The node performs random, systematic, stratified, sequential, and cluster sampling. From the node, you have the option to specify the desired sample size by entering either the appropriate number of records or the percentage of allocation. The node enables you to define the method for stratified sampling. In stratified sampling, observations are randomly selected within each non-overlapping group or strata that are created. The type of stratified sampling that can be performed from the node is either selecting stratified samples by the same proportion of observations within each strata, an equal number of observations in each strata, creating the stratified groups by the proportion of observations and the standard deviation of a specified variable within each group or a user-defined stratified sample in which each stratified group is created by various class levels of the categorical-valued variable. For cluster sampling, the node is designed for you to specify a random sample of clusters where the clusters are usually of unequal sizes, and then specifying the number of cluster to the sample based on all the selected clusters of either every *n*th cluster or the first *n* clusters. You may also specify a certain percentage of clusters based on all of the clusters that are created. The random seed number determines the sampling. Therefore, using an identical random seed number to select the sample from the same SAS data set will create an identical random sample of the data set. However, the exception is when the random seed is set to zero, then the random seed number is set to the computer's clock at run time. An output data set is created from the sample selected that is passed on through the process flow diagram.



The purpose of the **Data Partition** node is to partition or split the metadata sample into a training, validation, and test data set. The purpose of splitting the original source data set into separate data sets is to prevent overfitting and achieve good generalization to the statistical modeling design. Overfitting is when the model generates an exceptional fit to the data. However, fitting the same model to an entirely different random sample of the same data set will result in a poor fit to the data. Generalization is analogous to interpolation or exploration in generating unbiased and accurate estimates by fitting the model to data that is entirely different data that was used in fitting the statistical model. The node will allow you to select either simple random sample, stratified random sample, or user-defined sample to create the partitioned data sets. The random seed number determines the random sampling that follows a uniform distribution between zero and one along with a counter number that is created for each data set in order to regulate the correct number of records that are allocated into the partitioned data sets. The node will allow you to perform user-defined sampling where the class levels of the categorical-valued variable determines the partitioning of the data. User-defined sampling is advantageous in time series modeling where the data must be retained in chronological order over time.

### Explore Nodes



The purpose of the **Distribution Explorer** node is to view multidimensional histograms to graphically view the multitude of variables in the analysis. Observing the distribution or the range of values of each variable is usually the first step to data mining. Although, the node is designed to view the distribution of each variable separately, however, the node has the added capability of viewing the distribution of up to three separate variables at the same time. In other words, the node displays up to a 3-D frequency bar chart based on either the frequency percentage, mean, or sum. The node will allow you to select the axes variables to the multi-dimensional bar chart. From the tab, the node will allow you to display a frequency bar chart of each variable. For categorical-valued variables, the bar chart has the option of specifying the number of bins or bars that will be

displayed in the multi-dimensional bar chart. For interval-valued variables, the node will allow you to set the range of values that will be displayed within the bar chart. Descriptive statistics are generated for the interval-valued variables by each categorical-valued variable. Otherwise, if the selected axes variables are all categorical variables, then frequency tables will be generated.



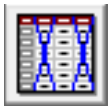
The purpose of the **Multiplot** node is a visualization tool to graphically view the numerous variables in the analysis through a built-in slide show. The node creates various bar charts, stacked bar charts, and scatter plots. The graphs are designed to observe trends, patterns, and extreme values between the variables in the active training data set. The **Multiplot** node gives you the added flexibility to add or remove various charts and graphs from the slide show. The slide show will allow you to browse the distribution between the variables by scrolling back and forth through the multitude of charts and graphs that are automatically created.



The purpose of the **Insight** node is to browse the corresponding data set to perform a wide assortment of analysis. The node opens the SAS/INSIGHT session. The node creates various graphs and statistics. Initially, when the node is opened, the node displays a table listing that is similar to a SAS Table View environment. The node can generate various charts and graphs such as histogram, box plots, probability plots, line plots, scatter plots, contour plots, and rotating plots. The node generates numerous univariate statistics, trimmed mean statistics, and robust measure of scale statistics. In addition, the node performs a wide range of analysis from regression modeling, logistic regression modeling, multivariate analysis, and principal component analysis. The node is capable of transforming and creating new variables in the data set. However, it is advised by SAS not to load extremely large data set into SAS/INSIGHT. Therefore, the node has an option of taking a random sample of the entire data set or a subset random sample of the data.

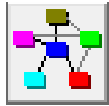


The purpose of the **Association** node is to identify associations or relationships of certain items or events that occur together or in a particular sequence. The node is designed to either perform association or sequential discovery analysis. This type of analysis is often called *market basket analysis*. As an example of market basket analysis is “if a customer buys product A, what is the probability that the customer will also purchase product B?”. In other words, the node is designed to determine the association or relationships between one item or event that might be attributed to another item or event in the data. Measuring the strength of the associations or interactions between the items of the if-then rule are determined by the three evaluation criterion statistics, that is, the support probability, confidence probability, and lift value. The goal to the analysis is to determine the association of items that occur most frequently or frequently occur together. The association is based on an id variable that identifies the various customers and the target variable that identifies the various items in the data. The difference between association analysis and sequence analysis is that sequence discovery takes into account the order in which the items or event occur. In sequence discovery, a sequence variable or a timestamp variable must exist in the data to identify the order of occurrence that the items or events occur. Sequence analysis is used to identify the sequence of events that are likely to occur during an interval of time. For example, there might be an interest in determining customers buying item A before item B based on the difference in time in which the items are purchased that is identified by a variable in the data with a sequence variable role. The node will allow you to reduce processing time by setting the minimum rate of occurrence and the number of possible combination of items that occur together that defines the if-then rule between the items. In sequence discovery, the node will allow you to determine the minimum number of sequences and the time interval to the sequence. From the results, the node will create separate bar charts of the three evaluation criterion statistics along with a grid plot to graphically view the corresponding evaluation criterion statistics between the items in the left and right hand side of the if-then association rule. The grid plot will allow you to graphically view the strength and weakness of the associated items that are created. And finally, the node will allow you to interactively create your own association rules that will allow you to view the three evaluation criterion statistics.



The purpose of the **Variable Selection** node is to select important input variables in the model that best predicts the target variable from a combination of potential input variables. Selecting the right combination of input variables is one of the most important steps to a well-designed statistical modeling design. From the node, the variable selection routine is a three step process. In the first step, the node initially performs correlation analysis based on the simple linear relationship between each input variable and the target variable. However, the node will allow you to set the criterion value to zero that will prevent the correlation analysis procedure from being performed. In the second step, the node performs forward stepwise regression to whose input variables that are not rejected from the previous correlation analysis. Again, the node will allow you to set the modeling selection criterion value. For binary-valued target variables to predict, there is an additional third step that is performed. The additional step is to fit the regression model to the fitted values that is the only input variable to the fitted model. For interval-valued targets, the node selects the most significant input variables based on the r-squared criterion. For categorical-valued targets, the node applies a chi-square criterion. For categorical input variables, the node produces a series of 2x2 frequency tables to select the best input variables to the classification model based on the highest chi-square statistic. However, the node has the added flexibility in performing the variable selection routine by applying the multiple linear regression model when fitting the binary-valued target variable. The node will allow you to remove input variables from the model that have a large amount of missing values, categorical input variables with a large number of unique values, or remove redundant input variables. Redundant variables are two separate variables that essentially have the same meaning or the same range of values. In addition, the node will

allow you to remove the input variables from the model, altogether. From the results, the node will display a table listing of the input variables that were selected from the variable selection routine. From the table listing, you may remove certain input variables from the model. The node displays bar charts from the first two steps in the variable selection routine. In other words, the node displays a bar chart of the individual r-squared statistic from the correlation analysis that is performed in the first step. The following tab will display a bar chart of the sequential increase in the r-square statistic from the forward stepwise regression procedure with each input variable entered into the model one at a time that is performed in the second step.



The purpose of the **Link Analysis** node is to visually display the relationship between the variables in order to define the characteristics between the variables. In addition, the node is capable of creating separate cluster groupings. The node generates various link analysis graphs to display the relationship between the variables. The graph consists of various nodes and links. The nodes represent the variables and the links represent the connections between the nodes. The nodes within the link graph are analogous to icons that are displayed on your desktop. The nodes are connected to each other to represent the relationship between one another. By default, the link graph displays a circle of nodes with numerous lines or links that are connected to a certain number of nodes. The nodes that are connected by a line will indicate the corresponding nodes that are related to each other. By default, the size of the node is determined by the relative frequency count and the thickness of the lines is determined by the marginal frequency count between the two separate nodes that are connected to each other. For interval-valued variables, three separate nodes are created by binning its values of equal length. For categorical-valued variables, a node is created for each class level. At times, the link graph can potentially amount to numerous links that might inhibit your ability in visualizing of the various links. Therefore, the node is designed so that you may only view the links that originate from the selected nodes. The node will also allow you to add or remove each node from the graph. The node calculates two separate centrality measurements. The first centrality measures the importance of the node based on the number of connections it has. The second centrality measures the number of nodes connected to it.

### Modify Nodes



The purpose of the **Data Set Attributes** node is to change the attributes to the metadata sample such as the data set name, description, and role of the data mining data set within the process flow. The node is designed to export the data mining data set into many different file formats. In addition, the node will allow you to assign the variable role, variable label, and measurement level. From the node, you may delete variables from the analysis. For categorical variables, the node will allow you to set the ordering levels that is important in the classification modeling designs. For categorical-valued target variables or interval-valued target variable stratified into separate intervals, the node will allow you to set the target profiles. One of the purposes of the node within the process flow is to assign the appropriate variable roles to the data set in which the data is read from the **SAS Code** node rather than the **Input Data Source** node.



The purpose of the **Transform Variables** node is to transform the interval-valued variables in the active training data set. The node will allow you to transform the variables from the numerous built-in functions in SAS. The node performs log, square root, inverse, squared, exponential, and standardized transformations. In addition, the node is capable of transforming the interval-valued variables from various robust statistics that are not seriously affected by outliers or extreme values in the data with comparison to the previous traditional maximum-likelihood estimates. The node has various options to transform the interval-valued variables into categorical variables by binning its values into buckets or quartiles of equal size. The node has various power transformations that are designed to either maximizes normality to the variable or maximize the correlation between the target variable. The node is designed such that if the transformation that is applied results in undefined values due to an illegal transformation that is performed to some of its values, then the node will automatically set the values to the variable to conform to the transformation that is applied to the variable. For example, applying a log transformation to a variable within the node will result in the node automatically adding the appropriate constant value to the variable that will prevent the node from performing an illegal transformation.



The purpose of the **Filter Outliers** node is to identify and remove observations from the active training data set based on outliers or missing values in the data. From the node, you have the option of eliminating rare values from the process flow diagram and keeping missing values in the analysis. For categorical valued variables, the node removes observations that do not occur within a certain number of times in each category. For interval-valued variables, the node will allow you to remove observations from the data that fall outside various ranges. In other words, the node will allow you to remove the values of the interval-valued variable by various interval settings such as the standard deviation from the mean, median absolute deviance, modal center, and extreme percentiles. The node gives you the added flexibility of removing observations from each variable in the active training data set. For each categorical-valued input variable, the node will allow you to remove observations from the training data set by removing certain class levels. For each interval-valued variable, the node will allow you to remove observations from the training data set by specifying certain intervals or range of values.



The purpose of the **Replacement** node is to impute or fill-in values that are missing. The node is designed so that you may impute missing values or redefine values based on predetermined intervals or class levels in the active training data set. In other words, the node will allow you to trim non-missing values by replacing values that might be incorrectly coded from the active training data set. By default, Enterprise Miner is designed to replace the missing values, and then replace these same imputed missing values by a specified interval. The node will allow you to globally impute missing values to each variable in the active training data set. However, the node will allow you to override the global settings and impute missing variable for each variable separately. For example, you might want to impute missing values in some variables by their own mean and other variables by their own median. By default, for interval-valued variables, its missing values are replaced by their own mean. For categorical-valued variables, its missing values are replaced by their most frequent category. However, the node consists of a wide assortment of imputation methods to replace missing values to the interval-valued variables such as the median, midrange, distribution-based replacement, tree imputation, tree imputation with surrogates, and various robust estimators. The node is also designed to allow you to replace extreme values or missing values with a constant value or redefine values between a specified range of values.



The purpose of the **Clustering** node is to perform cluster analysis. Clustering is the process of dividing the data set into mutually exclusive or non-overlapping groups with the data points within each group as close as possible to one another and different groups that are separated as far apart from one another. Once the clusters are created, the next step is to identify the difference between the clusters and observe the characteristics of each cluster that is created. As an example, there might be interest in clustering the various baseball hitters in the game based on numerous hitting statistics, and then determining the various hitting attributes between the separate clusters that are created. In other words, one cluster might be composed of mainly home run and rbi hitters with a different cluster consisting of base stealers and triples hitters, and so on. The node performs both hierarchical and partitive clustering techniques. For the hierarchical techniques, the node performs the average, centroid, or Wald's methods. For partitive clustering technique, the node performs the traditional k-means clustering technique. The node will allow you to standardize the input variables to the analysis before creating the cluster groupings since the squared distance between each observation and the cluster mean is applied to create the clustering assignments. There are various options settings that are available within the node in defining the clustering assignments such as specify the number of clusters or the minimum and maximum number of clusters, cluster size, clustering criterion statistic, and the convergence criterion that determines the process of replacement of the cluster seeds. The clustering seeds are the initial data points that create the temporary clusters. The cluster seeds are usually the first few observations with no missing observations among the input variables. From the results, the node will display a pie chart of the frequency counts between the clusters and the normalized mean plot to profile each cluster by identifying the standardized input variables that are significantly different from the overall mean. This will indicate the input variables which best characterize the corresponding cluster. The node will also create bubble plots that will allow you to graphically view the size and distance between the cluster groupings. The node displays table listings of the various clustering statistics. The table listings will allow you to observe the stability and reliability in the clustering assignments. Instability in the clusters will be indicated by a small number of observations and a small squared distance created within each cluster. In addition, the node will create a scored data set with a segment identifier variable that can be used in the following statistical modeling designs.



The purpose of the **SOM/Kohonen** node is to perform cluster analysis through network training. The three Kohonen network training unsupervised learning techniques that are available within the **SOM / Kohonen** node are either Kohonen vector quantization (VQ), Kohonen self-organizing maps (SOMs), or batch SOMs with Nadaraya-Watson or local-linear smoothing. Kohonen VQ technique is a clustering method as opposed to the SOMs techniques that are primarily dimension-reduction methods. Kohonen SOMs are similar to a neural network design where each input unit is connected to each output unit. However, the difference is that the output layer consists of numerous units that are arranged into a squared grid that is similar to a checkerboard where each square represents a cluster. The size of the two-dimensional squared grid is determined by the number of rows and columns. From the node, you may specify the number of rows and columns to the SOM design. As each observation is fed into Kohonen SOM training for cluster membership, the output units compete with one another based on the input unit assigned to the numerous possible output layer units where the rules of the game are "winner takes all". The winning output unit is the unit with the smallest squared distance between it and the input unit. The reward is that the winning unit, weight, or cluster seed is adjusted to strengthen the connection between it and the input layer units by superimposing the multi-dimensional squared grid onto the multi-dimensional input space. In SOM training, a neighborhood concept is incorporated to preserve the topological mapping by adjusting the neighboring units closer to the winning unit within the output layer. The purpose of the adjustment is to increase the likelihood of the input units assigned to it and the surrounding neighboring units. The node will allow you to specify the amount of adjustment to the neighboring units. Since Kohonen VQ training is similar to k-means clustering, the node enables you to specify the number of clusters. From the results, the node will graphically display the frequency counts for each cluster from the SOM/Kohonen map and an input means plot that is similar to the **Clustering** node. The plot will allow you to determine which input variables contributes best to each cluster that is created. The node will generate various clustering statistics to observe the stability and reliability of the clustering assignments such as the number of observations assigned to each cluster and the squared distance from the cluster mean and the furthest data point within each cluster.



The purpose of the **Time Series** node is to prep the data to perform time series modeling by condensing the data into chronological order of equally-spaced time intervals. Time series modeling is designed to predict the seasonal variability of the target variable based on its own past values over time. Therefore, the node requires both a target variable and a time identifier variable. The node will allow you to define the interval type by defining the length of each time interval such as year, quarter, month, week, day, etc. In addition, the node has the added option of entering a fixed length in time by specifying the start and end time. The node will allow you to specify the way in which the data is accumulated within each time interval by its sum, mean, median, minimum, maximum, and so on. In time series analysis, the observations must be in chronological order, therefore, the missing values might need to be estimated or imputed. The node has various imputation option settings in replacing missing data points in each time interval such as the mean, median, minimum, maximum, constant values to name a few. From the results, the node will allow you to view a trend plot and seasonal plot in order to view stationarity in the data over time. The trend plot will allow you to view both the variability and trend in the data at each data point over time. The seasonality plot will allow you to view the seasonal variability and trend in which the plot displays the accumulated data points over time.

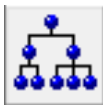


The purpose of the **Interactive Grouping** node is to automatically create group variables for the data mining analysis. The node will allow you to transform interval-valued variables into categorically-valued variables or redefine each categorically-valued variable into entirely different groups based on a binary-valued target variable. The various groups that are created within the node are determined by certain criterion statistics called the weight of evidence statistic and the information value statistic based on a binary-valued target variable. The WOE statistic measures the relative risk of the input grouping variable, that is, the logarithm difference of the response rate that is the difference between the proportion of the target nonevent and target event. The WOE statistic is very similar to the log odds-ratio statistic in logistic regression modeling. The information value statistic calculates the weighted difference between the proportion of the target nonevent and target event. The node has the added flexibility of allowing you to interactively group each input variable in the active training data set one at a time by viewing various frequency bar charts, line plots, and table listings of the grouping criterion statistics across each group that has been created. A scored data set is created with the corresponding grouping results that can be used as inputs in the subsequent modeling nodes. From the results, the node will display a table listing of the grouping performance of each input variable in the active training data set based on the various grouping criterion statistics.

### Model Nodes

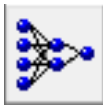


The purpose of the **Regression** node is to perform either linear and logistic regression modeling in the process flow. However, the added capability of the node is that it will allow you to apply the least-squares model in predicting the binary-valued target variable. The reason for fitting the least-squares model is to avoid computational difficulties and long lengthy runs due to the iterative maximum likelihood algorithm that is applied to calculate the parameter estimates and standard errors in the logistic regression model which may take many passes of the data to reach stable parameter estimates. From the node, polynomial models can be constructed by including higher-order modeling terms into the regression model. In addition, interaction terms may be added to the statistical model. For logistic regression modeling, various link functions may be specified. From the node, stepwise regression may be performed. The purpose of stepwise regression is to select the best set of input variables to the model by fitting the model to the validation data set. Furthermore, the node will give you the added flexibility of specifying the order in which the variables are entered into the regression model during the modeling selection procedure. From the node, you may specify the target profile that is similar to the following modeling nodes. For categorical-valued targets or stratified interval-valued targets, various numerical target-specified consequences can be predetermined in creating various business modeling scenarios that are designed to maximize expected profit or minimize expected loss from the validation data set. For categorical-valued targets, prior probabilities can be specified in order to increase the classification performance of the modeling design. From the results, the node will display bar charts of both the parameter estimates and the corresponding t-test statistics by fitting the active training data set. The standard table listing of various modeling assessment statistics will be displayed to view the stability in the model and overfitting in the data. This can be determined by observing a wide discrepancy between the modeling assessment statistics between the partitioned data sets. Scatter plots will be generated to view the functional relationship between the target variable and input variables in the model. In addition, the residuals may be plotted to validate the various statistical assumptions of the model. The node will also display the PROC DMREG procedure output listing and the internal score code that can be applied to a new set of data and input values. Similar to the other modeling nodes, the node generates the scored data set with the fitted values and residuals along with all the other variables in the data set.



The purpose of the **Tree** node is to perform decision tree analysis. Decision tree analysis is designed to perform either classification or predictive modeling. Decision tree modeling performs a series of if-then decision rules that forms a series of partitions that gradually divides the target values into smaller and smaller homogenous groups based on the input values. The *if* condition to the rule corresponds to the tree path and the *then* condition to the rule corresponds to the leaf node. The decision tree has a starting point at the top called the root and ends at the bottom called the leaves. In other words, an iterative branching process splits the target values until the data is partitioned at the leaf nodes. The leaves of the tree contain the final fitted values or estimated probabilities. For interval-valued targets, the fitted values are the target means at each leaf. For categorical-valued targets, the estimated probabilities are the target proportions at each leaf. The first step to the iterative split search procedure is to determine the best split for each input variable. The second step is to choose the

best split among a multitude of possible splits from many different input variables. The best splits are selected when the target values are divided into unique groups where each target group falls into one and only one node. The **Tree** node is designed to determine the best splits based on the  $p$ -values from the chi-square test statistic by the class levels of the categorical-valued target variable and the  $p$ -values from the F-test statistic by the range of values of the interval-valued target variable. However, the node will allow you to specify other assessment statistics in determining the best splits. The node provides various stopping rules. The purpose of the stopping rules are designed to avoid overfitting and improve stability in the tree. The various stopping criteria that can be selected from the node are setting the number of records required for a split search, setting the number of records allowed to a node, and reducing the depth of the tree. One of the advantages of the node is that it performs an internal variable selection routine to reduce the number of inputs variable and, therefore, the number of splits to search for. The node performs a variety of the more popular decision tree algorithms such as CART, CHAID, and C4.5. The node will allow you to interactively construct your own decision tree. The advantage of building your own decision tree is that you may prune the tree and define your own splits in order to increase the accuracy and stability of the tree and prevent overfitting in the model. In order to avoid overfitting in the tree, the training data set is used to build the tree and the validation data set is used to measure the accuracy of the tree model in which the best tree is selected. For interval-valued targets, the default goodness-of-fit statistic is the average squared error. For categorical-valued targets, the default assessment statistic is the proportion of the target class levels incorrectly classified. One of the advantages to the node is the way in which the tree model handles missing values. When there are missing values in the primary splitting input variable, then the node uses the surrogate input variable. The surrogate input variable is an entirely different input variable that will generate a similar split as the primary input variable. From the results, the node will allow you to update the tree and the corresponding scored data set based on the number of leaves that are selected. The node will display a line plot, table listing, and tree ring all in one window or tab. The line plot displays the goodness-of-fit statistic across the number of leaves to select the best decision tree. The line plot will display the number of leaves that are created from the training data set with a white vertical line indicating to you when the final tree was selected from the validation data set. In addition, the node will display a table listing of the assessment statistic between the partitioned data sets in order to determine stability in the decision tree. From the line plot or table listing, the node will allow you to perform tree pruning by selecting a smaller number of leaves. The tree ring can be used as a navigational tool that will allow you to quickly scan through the various assessment statistics at each split that was performed within the tree from the training data set. A subsequent table listing will allow you to view the variables added and removed from the decision tree model.

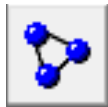


The purpose of the **Neural Network** node is to perform neural network modeling. Neural network modeling is essentially non-linear modeling within the process flow diagram. The default neural network architecture is the multilayer perceptron MLP network with one hidden layer consisting of three hidden units. Generally each input is connected to the first layer, each hidden layer is fully connected to the next hidden layer, and the hidden layer is fully connected to the output. The node is designed to perform various neural network architectures such as MLP and RBF designs. The node will allow you to select from a wide variety of activation functions, combination functions, error functions, and optimization techniques. In addition, you may perform preliminary runs, add multiple hidden units or hidden layers, weight decay regularization, direct connections, and apply various standardization methods to the input variables, target variables, and the bias or scale parameters. The node is also designed to perform interactive training which is sequentially constructing a neural network model by reconfiguring and refitting the neural network model any number of times all within the same window environment. From the results, the node generates a table view and graphical view of the weight estimates. In addition, an optimization line plot is displayed that plots the modeling assessment statistic or goodness-of-fit statistic at each iteration of the iterative gradient search with a vertical white line indicating the iteration in which the final weight estimates were determined based on the smallest average error or misclassification error from the validation data set. The node will allow you to perform updates to the weight estimates and corresponding modeling assessment statistics from the optimization plot.



The purpose of the **Princomp/Dmneural** node is to perform either principal component analysis or dmneural network modeling. Principal components analysis is designed to explain the variability in the data as opposed to dmneural network modeling that is designed to explain the variability in the target variable. The purpose of principal components is both input variable reduction and interpretation in the linear combination of the input variables or eigenvectors in explaining the variability in the data. In principal components, the linear combination of input variables that accounts for the highest amount of variation is called the first principle component. This will be followed by the second principal component that explains the next most variability in the data. The first step is to determine the smallest number of the principal components that accounts for a sufficient amount of variability. The next step is hopefully identifying the characteristic grouping in the linear combination of input variables and eigenvectors. This can be achieved by observing the larger eigenvectors. For principal component analysis, the node will allow you to specify either the covariance or correlation matrix to calculate the principal components in explaining the variability in the data. In addition, the node will allow you to specify the maximum number of principle components to output. In both modeling designs, the node will allow you to standardize the input variables since the modeling designs are based on the variability in the input variables in the model. Dmneural network training is an additive nonlinear model. The reason is because the node automatically performs three separate stages to the nonlinear modeling design. In the first step, the model predicts the target values. In the subsequent models, the residual values are then used as the target variable to the model in order to

retrieve additional information from the residual values to predict the variability in the target variable. In dmneural network modeling, the fitted values are based on a multidimensional frequency table that consists of the frequency counts of the selected principal components at a specified number of discrete grid points in order to accelerate network training. An activation function is applied, like neural network modeling, to the linear combination of input variables and eigenvalues, i.e. principal components. The node will allow you to specify the goodness-of-fit statistic, number of principal components, maximum number of stages (up to 10 stages), number of iterations, and minimum r-square statistic. From the principal component results, the node displays various bar charts and line plots that display the amount of variability explained by the model across the number of principal components. The plots will allow you to select the number of principal components to retain to the analysis. The node displays a table listing of the eigenvalues and eigenvectors. The table listing will allow you to view the proportion of the variability explained by each principal component. In addition, the table listing will allow you to characterize between the input variables that are explained by each principal component. A scored data set can be created based on the number of principal components that have been selected from the node. The scored data set consists of the input variables along with the principle components that have been selected. For dmneural modeling, the node will display scatter plots or bar charts of the fitted values to view the accuracy of the network model. The node will display a table listing at each stage of the eight separate activation functions that are automatically applied to the additive nonlinear model. A subsequent table listing will be displayed that lists the best activation functions with the smallest modeling assessment statistic at each stage of the nonlinear modeling design.



The purpose of the **User Defined Model** node is to allow you to include the fitted values from some other external modeling procedure into the Enterprise Miner process flow diagram. In other words, the node will allow you to generate assessment statistics from other modeling procedures that are not a part of Enterprise Miner such as PROC NLIN, PROC ARIMA, PROC GENMOD, and so on. In addition, you may create a scored data set that will allow you to compare the fitted values to the other modeling designs within Enterprise Miner. The node will allow you to specify prior probabilities to each class level of the categorical target variable to predict. The purpose of the prior probabilities is that they are used to adjust the estimated probabilities to increase the accuracy of the classification model. In addition, the node will allow you to specify the target profile to determine the best profit or loss of the user-defined modeling design.



The purpose of the **Ensemble** node is to combine, stratify, or resample the fitted models. The purpose of the ensemble modeling designs is to achieve more accurate models. The reason is because the node has the capability of averaging the prediction estimates from various models or averaging the prediction estimates based on successive fits from the same predictive model where the analysis data set is randomly sampled any number of times. This will result in more stable models that can significantly improve generalization. One restriction is that the estimates must be generated from the same training data set with the target variable in each separate model that must be compatible with the same level of measurement. The node will also allow you to fit stratified models by generating separate estimates for each group or segment from the training data set. In addition, the node will allow you to specify a couple resampling techniques called bagging and boosting. Bagging is essentially bootstrap sampling. Bootstrap sampling randomly samples the same data set with replacement by fitting the same model numerous times. The bagged estimates are calculated by refitting the predictive model to generate the fitted values for each bootstrap sample of equal size, and then dividing by the number of bootstrap samples. Boosting is a classification modeling design that refits the model numerous times. In this iterative classification modeling design, weight estimates are applied to adjust the estimated probabilities for each observation that has incorrectly classified the target class levels from the previous fit. Stratified modeling and both the resampling techniques work in conjunction with the **Group Processing** node.



The purpose of the **Memory-Based Reasoning** node is to perform nearest neighbor modeling. Nearest neighbor modeling is a nonparametric modeling technique. The reason is because the model does not assume any distributional relationship or functional form between the input variables and the target variable you want to predict. The only parameter estimate that this type of smoothing predictive model technique needs is the number of neighbors  $k$ . By default, the node sets the smoothing constant to 16 neighbors. The number of nearest neighbors  $k$  is usually determined by trial-and-error that depends on the distribution of the data and the number of variables in the nonparametric model. For categorical-valued targets, the fitted values are determined by the number of occurrences of each target group falling within the hypersphere surrounding the probe  $x$ , and then dividing the number of occurrences by the number of nearest neighbors. The number of neighbors  $k$  determines the degree of smoothing to the decision boundaries between the target groups. For interval-valued targets, the node calculates each fitted value within a predetermined window of  $k$  target values. The distance function, usually the smallest sum-of-squares, is applied to locate the nearest neighbors and the combination function is used to combine the target values from the nearest neighbors, and then averaged to generate the fitted values. For interval-valued target variables, this type of model fitting is analogous to moving average time series modeling. Since it's critical to reduce the number of input variables in order to condense the multidimensional window of data points, therefore, it's recommended that principal components are the only input variables in the nonparametric model. Hence, the sorted values from the first principle component scores determine the fitted values of the  $k$ -nearest neighbor estimates. The reason for applying the set of principal components is to provide the model with the best linear combination of the input variables to explain the variability in the data. In addition, the principal components are independent of each other that will resolve algebraic problems in the nonparametric model. The node will provide you with various methods

to determine the most appropriate neighbors to calculate the prediction estimates. The node also has a weight dimension option that enables you to smooth-out the jagged nonlinear fit. The weights are designed to adjust the distance function by favoring some input variables over others in the model.



The purpose of the **Two Stage Model** node is to predict an interval-valued target variable based on the estimated probabilities of the target event from the categorical target variable. In other words, two separate models are fit in succession. In the first stage, the class model is fit to predict the categorical target variable. In the second stage, the value model is fit to predict the interval-valued target variable with the estimated probabilities of the target event that are used as one of the input variables in the predictive second-stage model. The node requires two separate target variables to fit in the two-stage model. The reason for this type of modeling design is because at times the interval-valued target might be associated with the class levels of the categorical-valued target variable. For example, two-stage modeling might be useful in predicting the total sales of an item based on an identifier of this same item that was purchased. By default, decision tree modeling is applied in the first stage by fitting the categorical target variable to the classification model. In the second stage, least-squares regression modeling is applied to the interval-valued target variable to predict. However, the node performs a wide variety of modeling techniques to both stages of the two-stage modeling design such as decision-tree modeling, regression modeling, MLP and RBF neural network modeling, and GLIM modeling. The node will allow you to include either the estimated probabilities or the classification identifier of the target event from the first-stage model as one of the input variables to the second-stage model. The node gives you the added capability to adjust the estimated probabilities of the target event from the first-stage model through the use of the bias adjustment option. In addition, there is a filtering option that is designed to remove all the observations from the second-stage model in which the first-stage classification model has incorrectly identified the target event. From the results, the node displays the classification table to evaluate the classification performance of the first-stage model and the standard assessment statistics to evaluate the predictive performance of the second-stage modeling design. The standard summary statistics are listed by fitting the interval-valued target variable by each combination of the actual and predicted levels of the categorical target variable in the first-stage model.

#### Access Nodes



The purpose of the **Assessment** node is to compare the prediction estimates from several different models by viewing various comparison plots and modeling assessment statistics. For interval-valued target variables, the node will allow you to view various frequency bar charts and scatter plots of the target values or the residuals to the model in order to view the accuracy of the prediction model and validate the various statistical assumptions. For categorical-valued target variables, the node will allow you to select each classification model separate or as a group. The node is designed to compare the classification performance from several different models by displaying various lift charts or performance charts that are line plots that plot the predicted probabilities across the ordered percentile estimates from the validation data set. In addition, the node displays response threshold bar charts by fitting the binary-valued target variable. The response threshold bar charts display the rate of accuracy across the range of the threshold probabilities. The threshold probability is a cut-off probability that is used in assigning an observation into one of the two target class levels based on the estimated probability. By selecting each classification model separately, the node will display the threshold-based bar charts. Threshold-based bar charts display the frequency bar chart of the classification matrix. The classification matrix is a two-way frequency table between the actual and predicted class levels of the categorical target variable. The node will allow you to adjust the threshold probability in order to determine the accuracy of the model in classifying each target class level from the specified threshold probability. ROC charts are generated that are designed to display the predictive power of the classification model at each level of the binary-valued target variable. The one requirement of the node is that it must proceed anyone of the modeling nodes that must be connect to the node.



The purpose of the **Score** node is to view, edit, save, delete, combine, or execute the score code program. The score code is the internal Enterprise Miner scoring formulas that are generated from the corresponding node by executing the node. In addition, the score code will list the data step processing that was involved to generate the results. The score code can be used within the SAS editor to generate new prediction estimates by providing a new set of values from the input variables. One of the purposes of the node is that you may score the incoming data set from the most desirable modeling node that is part of the process flow diagram.



The purpose of the **Reporter** node is to assemble and organize the results from the various nodes within the process flow diagram into an HTML report to be displayed by your local Web browser. The **Reporter** node condenses the statistical results into a well-organized HTML layout for presentational purposes. From the main menu options, you may customize the layout design of the HTML listing. The HTML listing consists of various table listings, graphs, and hyperlinks. The hyperlinks will navigate you to the results that are generated from the various nodes.

The following are the remaining utility nodes available in the SEMMA Enterprise Miner process flow diagram.

### Utility Nodes



The purpose of the **Group Processing** node is to perform a separate analysis by each class level of the grouping variable. The node will also allow you to analyze multiple targets separately or analyze the same data set repeatedly by performing bagging, i.e. bootstrap sampling, or boosting resampling to the subsequent analysis node in the process flow. The node has the added flexibility of controlling the number of times the subsequent nodes will loop in the process flow diagram that are connected to the node. The node is often used in conjunction with the **Ensemble** node. For instance, the **Group Processing** node can be used to instruct Enterprise Miner to create the fitted values from each stratified model that can then be combined from the subsequent **Ensemble** node. In addition, the bagging or boosting resampling techniques cannot be performed from the **Ensemble** node unless the resampling techniques are selected within the **Group Processing** node.



The purpose of the **Data Mining Database** node is to create a data mining database. The node will provide you with the only mechanism in the process flow to browse the statistics or view the distribution of the variables based on all the observations from the input data set. The DMDB data mining data set is designed to optimize the performance of the various nodes in the process flow. The DMDB procedure is used to create the DMDB data mining data set. The procedure compiles and computes the metadata information of the input data set by the variable roles. The metadata stores both the data set information as well as the statistical information for the variables associated with the assigned roles. For categorical variables, the catalog of the metadata information contains information such as the class level values, class level frequencies, number of missing values for each class level, its ordering information, and the target profile information. For interval-valued variables, the metadata consists of the range of values, number of missing values, moments of the variables, and the target profile information that are then used in many of the other Enterprise Miner nodes that require a DMDB data set. From the results, table listings will display various descriptive statistics that are based on all the observations from the input data set for the interval-valued variables. For categorical variables, the node will display the number of class levels and frequency bar charts that are based on all the observations from the input data set.



The purpose of the **SAS Code** node is to give you the ability to incorporate new or existing SAS programming code into the process flow diagram. The node is one of the most powerful Enterprise Miner node to the SEMMA process diagram. The reason is because it gives the you the ability to access many of the powerful SAS procedures into the process flow. In addition, the node will allow you to perform data step programming within the process flow to manipulate the various data mining data sets that are a part of the process flow. One of the tabs will display a hierarchical listing of the numerous macro variables that are available within the node and the corresponding SAS session. The numerous macro variables are in reference to the data mining libraries, data set roles, scored data sets, variables or variable roles, measurement levels, and exported data sets. A separate tab will allow you to write SAS programming code. However, the node will allow you to access the standard SAS editor to write the corresponding SAS programming code.



The purpose of the **Control Point** node is to establish a control point in the process flow diagram. The **Control Point** node is used to reduce the number of connections that are made in the process flow diagram in order to keep the appearance of the various nodes that are connected to one another within the diagram easier to interpret. As an example, connecting multiple data sets to each modeling node can be reduced from the node by connecting each **Input Data Source** node for each respective data set to the **Control Point** node that is then connected to each of the modeling nodes.



The purpose of the **Subdiagram** node is to create a portion of the process flow diagram by subdividing the process flow diagrams into separate diagrams. The advantage of subdividing the process flow diagram is to subdivide the numerous nodes and connections into smaller more manageable diagrams that are then reconnected to one another. Subdividing the nodes into separate subdiagrams will give you the added capability to copy or export the subdiagrams into separate Enterprise Miner projects and diagrams.

### CONCLUSION

Enterprise Miner v4.3 is a powerful product that is available within the SAS software. I hope after reading this article that Enterprise Miner v4.3 will become very easy SAS analytical tool for you to use in order to incorporate in your SAS analysis tools.

### CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Randall Matignon  
Piedmont, CA 94611  
Phone: 510-547-4282  
E-mail: [statrat594@aol.com](mailto:statrat594@aol.com)  
Web: <http://www.sasenterpriseminer.com>

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. © indicates USA registration.

Other brand and product names are trademarks of their respective companies.