

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 is 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 these tools and 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 don't even need to know SAS programming and have little statistical expertise in the development of your EM 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 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 to the EM diagram graphical workspace. The SAS EM diagram workspace environment looks similar to the desktop in Windows 95, 98, XP, and Vista. EM is very easy to use and can save a tremendous amount of time having to program in SAS. 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 does a wide variety of statistics from descriptive and univariate statistics, numerous types of charts and plots, goodness-of-fit modeling assessment statistics, decision tree analysis, principal component analysis, cluster analysis, association analysis, link analysis, and traditional regression modeling with automatically generated graphs that can be directed to the SAS output window.

Opening Enterprise Miner

- The most common way of opening Enterprise Miner is by selecting **Solutions > Analysis > Enterprise Miner** from the Enterprise Miner main menu.
- Enter or type *miner* at the command line that is located at the upper left hand corner underneath the command bar.
- Create an Enterprise Miner icon onto your Microsoft Windows desktop, and then double click on the Enterprise Miner icon.

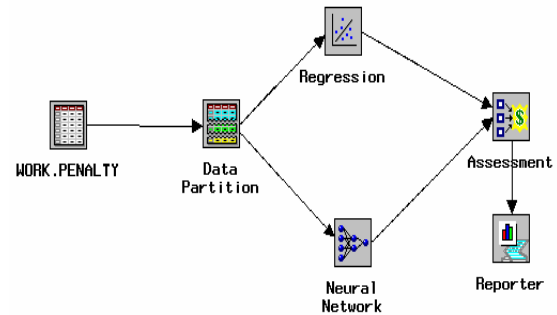
Running the Diagram

There is one of two ways to run an Enterprise Miner SEMMA design.

- Run each node separately after connecting each node in succession to the EM process flow diagram with the exception being the initial **Input Data Source** node.
- The standard procedure in compiling the Enterprise Miner diagram is to run the entire SEMMA EM diagram all at once by constructing the entire Enterprise Miner process flow diagram and selecting the last node connected then right-click the mouse and select **Run** from the pop-up menu items. Therefore in the following diagram that is displayed in the next page, the **Reporter** node is the last node to the process flow diagram to execute.

Basic Steps in Designing the Workflow Diagram to Compare Neural Network and Regression Prediction Estimates.

1. Drag and drop the **Input Data Source** node onto the diagram workspace.
 2. Double click to open the **Input Data Source** node.
 3. Select the **Data** tab and press the **Select...** button to browse the available library references to select the permanent or temporary SAS data set to create the metadata sample.
 4. Select the **Variables** tab to set the variables roles of target, input and id variables to the model.
 5. Drag and drop the **Data Partition** node to the right of the **Input Data Source** node onto the diagram workspace and connect the **Input Data Source** node to the **Data Partition** node
 6. Double click the mouse to open the **Data Partition** node.
 7. Select the **Partition** tab to specify the percentage of data allocated to the training and validation data sets. Set the percentages option with a 50% allocation to both the training and validation data sets.
 8. Drag and drop the **Regression** node to the right of the **Data Partition** node onto the diagram workspace.
 9. Connect the **Data Partition** node to the **Regression** node. Double click the **Regression** node to open the node and select the **Output** tab, then select the **Process or Score** check box to create an output scored data set with the predicted values, then close the **Regression** node.
- Note:** Repeat steps 8-10 with the **Neural Network** node.
10. Drag and drop the **Assessment** node to the right of the **Regression** node and the **Neural Network** node in the diagram.
 11. Connect both the **Regression** node and the **Neural Network** node to the **Assessment** node.
 12. Drag and drop the **Reporter** node onto the diagram workspace.
 13. To run the process flow, select the Reporter node and right-click the mouse to select **Compile** from the shortcut pop-up menu items.
 14. After running process flow diagram, open the Reporter node to view the results.

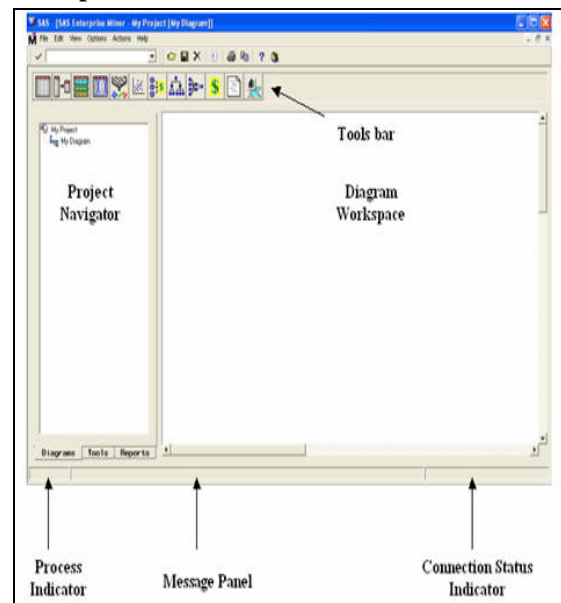


Layout of the EM workflow diagram to compare the Regression and Neural Network models

The Working Environment of Enterprise Miner

The Enterprise Miner layout is designed for ease of use. The working environment is similar to the Windows desktop. However, the complexity of Enterprise Miner SEMMA designing is setting the correct configuration settings and constructing the appropriate process flow diagram.

Enterprise Miner Window



By default, the last project that was opened will be displayed when opening Enterprise Miner. The Enterprise Miner window and working environment is composed of three major components as follows:

The Enterprise Miner Tools Bar

The Enterprise Miner tools bar is a graphical set of node icons to build a SEMMA process flow diagrams in the diagram workspace. The same tools bar node icons are also located in the **Project Navigator Tools** tab. The **Tools Bar** is designed to give you easy access in dragging the more popular nodes onto the diagram workspace.

The Project Navigator

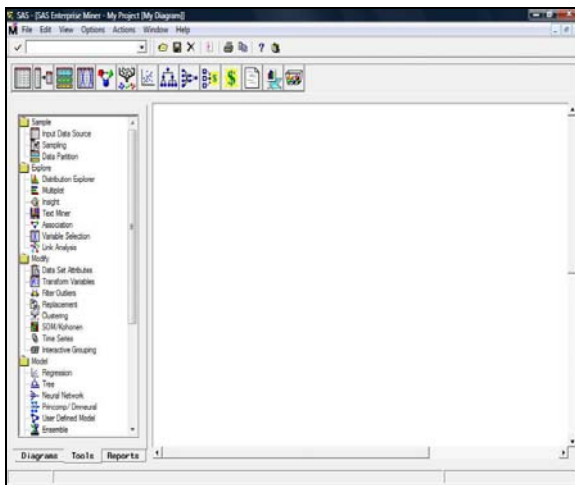
The Project Navigator manages the projects, diagrams, and reports. The window display looks similar to the Windows Explorer style interface in which the displayed Enterprise Miner project and diagrams look similar to system folders. The Project Navigator is listed to the left of the Diagram Workspace. The Project Navigator window is composed of three separate tabs, the **Diagrams** tab, **Tools** tab, and **Reports** tab.

Diagrams tab: The diagram tab displays the current EM project that is opened and the list of the various diagrams associated with the currently opened project. The diagram workspace can be displayed by simply selecting the corresponding diagram from the **Diagram** window.

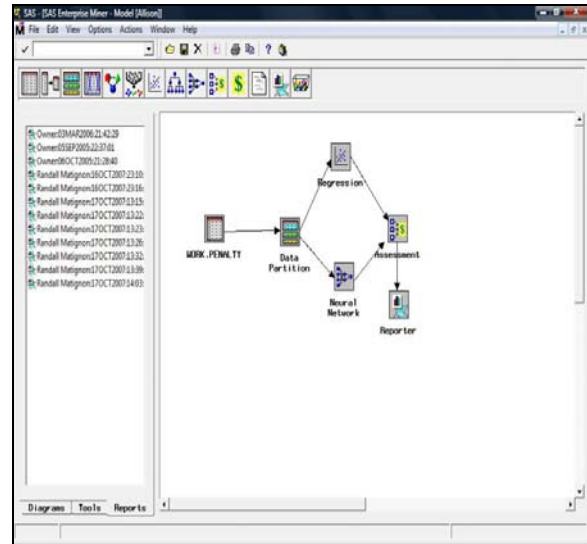
The EM project and the associated diagrams are listed in a hierarchical fashion underneath the Enterprise Miner project that is listed at the top. The **Project Navigator** will allow you to create, open, rename and delete the currently listed project, diagrams and reports. The window will also display the status of the project and diagrams.

Tools tab: The tools tab displays all of the Enterprise Miner tool nodes that are available. It is organized in a hierarchical fashion according to the SEMMA (Sample, Explore, Modify, Model and Access) and Utility nodes listing. The tools nodes can be dragged and dropped onto the diagram workspace to become part of the process flow diagram.

Reports tab: The reports tab contains the HTML report entries created from the **Reporter** node. The reports tab displays all the reports that are associated with the currently opened EM project.



The Enterprise Miner Tools tab to select the various nodes to place onto the Diagram Workspace.



The Enterprise Miner Reports tab to view the various HTML reports that have been created.

The Diagram Workspace

The diagram workspace is a desktop environment to construct the Enterprise Miner SEMMA process flow diagrams. The workspace is used to build, run, and save the currently opened Enterprise Miner diagram.

- The **Progress Indicator** is a progress bar indicating to you of the execution status of the Enterprise Miner task.
- The **Message Panel** displays the task being executed and the associated user messages.
- The **Connection Status Indicator** is active during the client-server projects. The indicator displays the remote host name and the connection status.

User Preference Options in Enterprise Miner

The purpose of the user preference options within the **User Preferences** window is to specify the start-up configuration settings to the Enterprise Miner environment. In addition, these options gives you the flexibility to redirect the compiled Enterprise Miner results to either the SAS system output window or the SAS system log window, set-up server profile settings for the client/server Enterprise Miner project and customizing the layout of the HTML report created from the **Reporter** node. From the **Enterprise Miner** window, select **Options > User preferences** from the Enterprise Miner main menu and the **User Preferences** window will appear.

The following are the various tabs that are available from the **User preference** window.

Session tab: The purpose of the **Session** tab is to allow you to select the various option settings to the default working environment in Enterprise Miner when you first open the EM project. The tab will allow you to specify various option settings such as redirecting the procedure log and output listings to the SAS log and output window. In addition, the tab will allow you to prevent various goodness-of-fit statistics and various performance charts from being generated from the classification models within the **Assessment** node.

Projects tab: The purpose of the **Projects** tab is to specify the default directory when you create a new Enterprise Miner project, autosave settings to automatically save the opened Enterprise Miner diagram a specified amount of time, and network configuration settings in automatically connecting to a client/server network.

Server Profiles tab: The purpose of the **Server Profiles** tab is to create, modify or delete the network server profiles used in client/server connections. The server profile contains information for the network administrator to set-up configuration settings that Enterprise Miner needs to establish remote connections to the network server.

Reports tab: The purpose of the **Reports** tab is to allow you to customize the layout of the HTML listing that is created from the **Reporter** node. The tab will give you the flexibility of organizing the way in which the hyperlinks and the corresponding results are displayed in the **Reporter** node. In other words, the tab will give you the capability in specifying the order in which the various reporting topics are listed in the HTML report by selecting the **Create Report** option from the **Reporter** node. A window will display two separate list boxes that are side-by-side. The **Available report items** list box that is positioned to the left of the window will display all the possible topics that the **Reporter** node can display. The **Display in reports** list box that is displayed to the right of the window displays the selected topics that will be displayed in the HTML report (in that specific order).

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.

Sample – 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 samples the source data set or partitions the source data set into a training, validation and test data set.

Explore – 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.

Modify – 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, identify outliers, replace missing values or perform cluster analysis.

Model – Fits the statistical model. The nodes predict the target variable against the input variables by using either regression, decision tree, neural network, dmneural network, nearest neighbor, ensemble, two-stage or user-defined modeling.

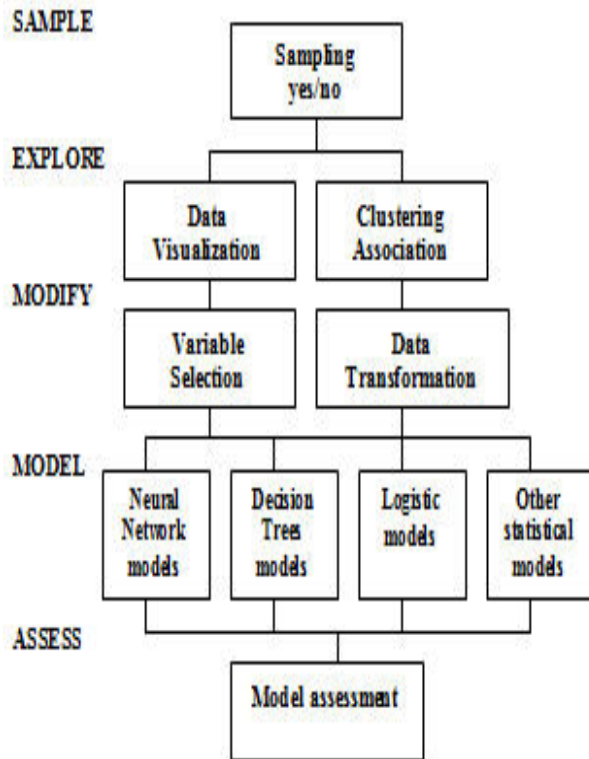
Assess – Compare the accuracy between the statistical models. The nodes compare the performance of the various classification models by viewing the competing probability estimates from the lift charts, ROC charts, and threshold charts. For predictive modeling designs, the performance of each model and 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 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.

Relationship between SEMMA and the Enterprise Miner Nodes

SEMMA	Enterprise Miner Nodes
1. Sample	Input Data Source, Data Partition, Sampling
2. Explore	Distribution Explorer, Multiplot, Insight, Association, Variable Selection, Link Analysis
3. Modify	Data Set Attributes, Transform Variables, Filter Outliers, Replacement, Clustering, SOM/Kohonen, Time Series
4. Model	Regression, Tree, Neural Network, Princomp/Dmneural, User Defined, Ensemble, Memory-Based Reasoning, Two Stage Model
5. Assess	Assessment, Score, Reporter

The Flowchart to the SEMMA Design



Basic Steps to Enterprise Miner SEMMA Predictive or Classification Modeling

- Read the source data set.
- Partition the outputted source data set into usually two data sets, that is, a training and validation data set, and even third data set called a test data set (assuming that there is a sufficient amount of data).
- Generate various models such as regression, neural networks, decision trees, ensemble modeling, two-stage modeling, or user-defined modeling.
- Assess the performance of various models by analyzing various modeling assessment statistics or comparison plots that can be created in order to determine the best predictive or classification model.
- Select the best predictive or classification model that predicts the target responses.
- Add custom nodes to produce desire results such as the SAS code node that will allow you to access numerous SAS procedures or perform data step manipulation.
- Create HTML reports for presentational purposes in which Enterprise Miner embeds the generated output into a HTML format layout for presentational purposes.

Conclusion

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

Contact Information

Randall Matignon
 Piedmont, CA 94611
 510-547-4282
 e-mail: statrat594@aol.com
 website: www.sasenterpriseminer.com