What's New in SAS® Enterprise Miner™ 5.2

Wayne Thompson, SAS Institute Inc., Cary, NC
David Duling, SAS Institute Inc., Cary, NC

ABSTRACT

SAS Enterprise Miner™ 5.2 for SAS 9.1.3 provides many new enhancements to help both business analysts and statisticians carry out the data mining process more efficiently and with greater control and flexibility. A major focus of this release is to deliver new interactive statistical and visualization tools. The tool set has been expanded to include the new SOM/Kohonen, Decisions, and Replacement nodes. Major improvements have been made to nearly every other node. System administration has been enhanced through the use of the SAS Analytics Platform, which provides both thin-client distribution and server management functionality. Grid processing is now supported to manage the workload created by a large group of data miners. Customers will find many reasons to upgrade to SAS Enterprise Miner 5.2.

INTRODUCTION

SAS Enterprise Miner 5.2 is the SAS solution for data mining, providing unparalleled model development and deployment opportunities. Delivered as a distributed client-server system, Enterprise Miner is well suited for joint workgroup collaborations and large data mining applications. Enterprise Miner's process flow diagram eliminates the need for manual coding and reduces the model development time for both business analysts and statisticians (see Figure 1). The system is customizable and extensible; users can integrate their code and build new nodes for redistribution. This paper provides an overview of the major enhancements of the latest release, Enterprise Miner 5.2 for SAS 9.1.3, delivered in November of 2005.

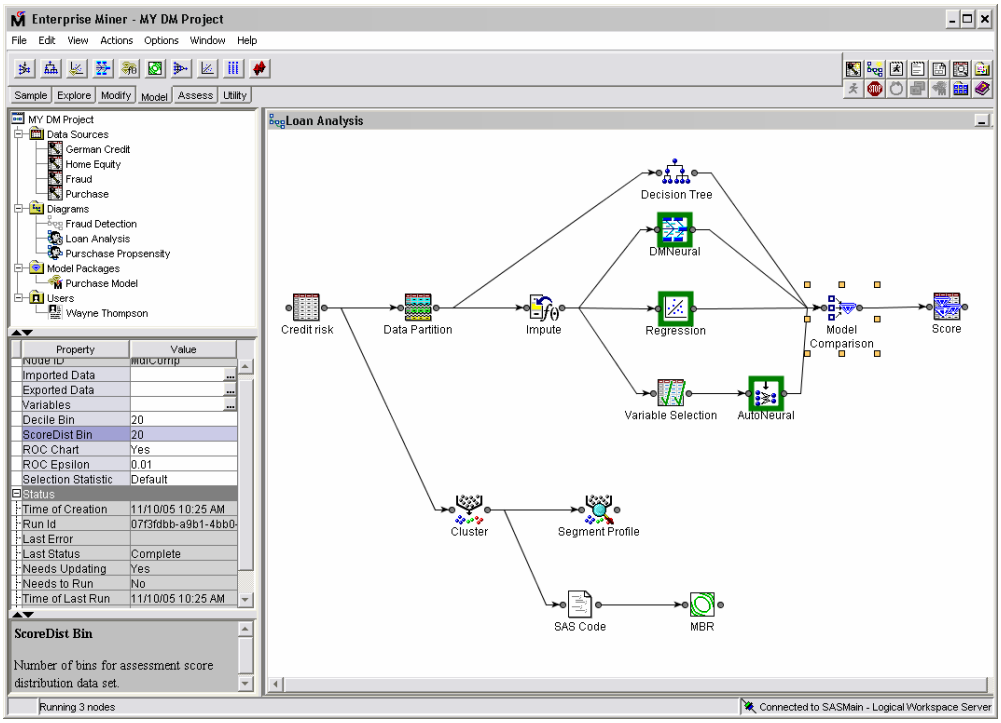

Figure 1. SAS Enterprise Miner 5.2 Graphical User Interface. Projects are persisted on the analytical server enabling data miners to collaborate on the analyses. The process flow diagram is a self-documenting template that can be easily updated or applied to new problems and shared with other analysts.

DATA VISUALIZATION AND MODIFICATIONS

Data exploration and preparation are important data mining tasks used to reveal systematic patterns and derive new features to ultimately help the analyst better understand, analyze, and model the data. SAS Enterprise Miner 5.2 delivers new interactive statistical and visualization tools to help the data miner better search for trends and anomalies and prepare the data in a manner more useful for model development.

EXPLORATORY GRAPHS

The graphics libraries in the SAS Enterprise Miner 5.2 client have been significantly enhanced with improved performance and many new plot types, including two-dimensional and three-dimensional graphics. Tables and graphs can be independently arranged. Interactive graphs are dynamically linked so that selecting data points in one plot updates the displays of corresponding plots and tables.

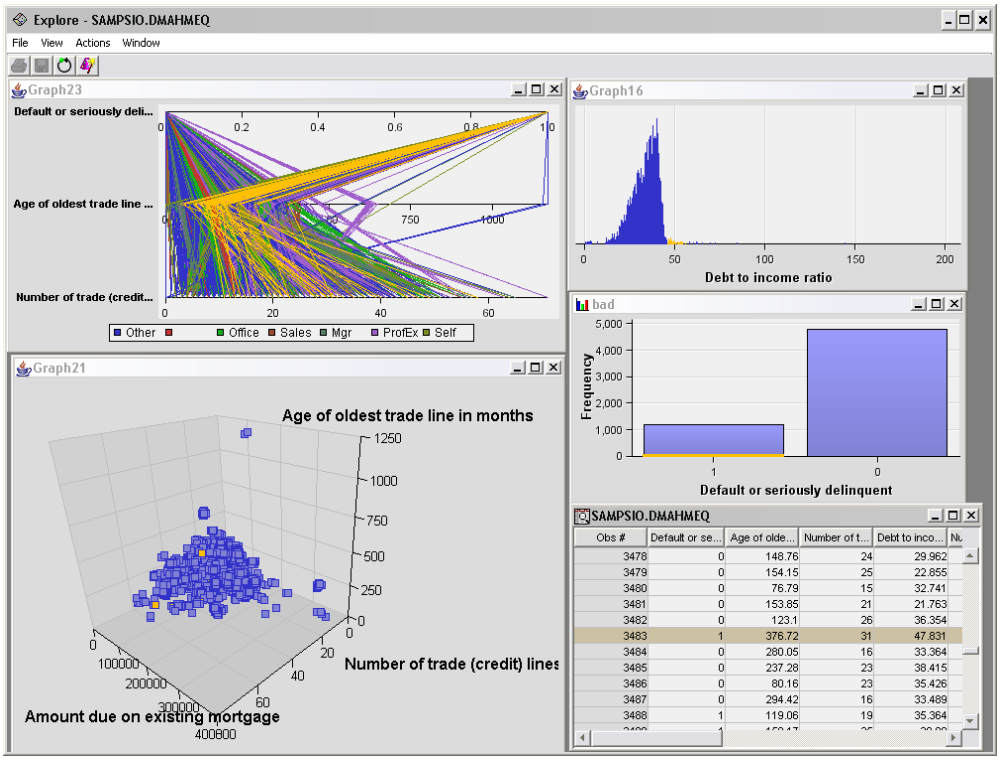

Figure 2. Explore your data interactively with parallel axis, density, 3-D rotating scatter, and other plots.

NEW NODES IN SAS ENTERPRISE MINER 5.2

Data preparation is known to be an important task for improving both the fit and reliability of a mining model. Variable transformations can be used to stabilize variances, remove nonlinearity, improve additivity, and correct non-normality in variables. Filtering extreme values from the training data can stabilize model parameter estimates. Replacing miscoded values and imputing missing values are other common data preparation acitvities.

Data preparation is also known to be one of the most time-consuming tasks, often cited as taking 50 to 75% of the total project effort. SAS Enterprise Miner 5.2 delivers several new interactive data preparation capabilities to help the data miner create and modify variables with more control and flexibility.

The **Tranform Variables node** now includes a formula expression builder to create customized variables from the input variables (Figure 3). The expression builder provides a host of operands and functions for defining the transformation, or you can type the SAS code. Distribution plots of the orignal variable and the new transformation can be viewed. Formulas are tested by executing against a sample and testing for errors. The user can easily modify the variable transformation definition at any time. The transformation logic is included the Enterprise Miner score code.

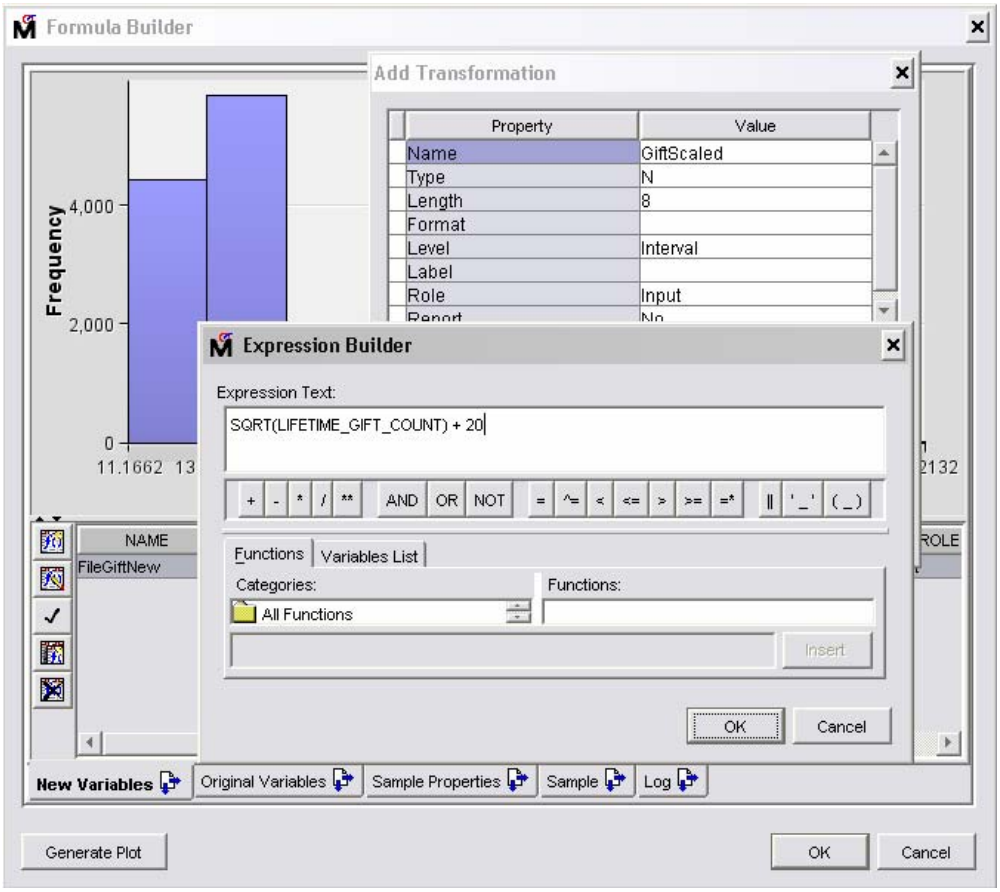

Figure 3. Develop customized transformations using the interactive Transform Variables node Expression Builder.

The **Replacement node** is a new tool for manipulating class variable values and includes both automated and manual replacement options. You can use the node to interactively specify replacement values for class variable levels, for instance to correct data miscoding, to combine levels or reduce dimensionality, or to regroup values created by binning. The training data is scanned on the SAS server to generate a summary table containing all class input and target levels. The summary table is displayed in the client where the user enters replacement values.

In Figure 4, known and unknown values of input variables have been mapped to new values for the JOB and REASON variables. The node also has options for automatically replacing unknown values in the scoring code.

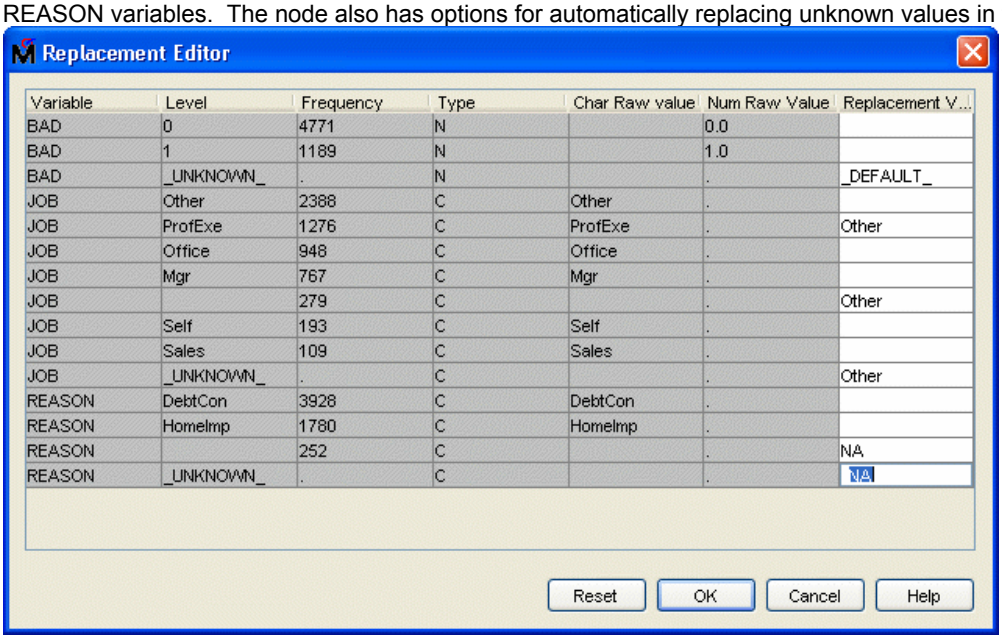

Figure 4. Map new values manually in the Replacement node.

The **Filter node** has been enhanced to support interactive user selection of filter values from both categorical and interval inputs (Figure 5). A new algorithm has been implemented to identify regions of outliers in the data set. Rather than creating a naïve histogram style distribution, the new code recursively builds and merges percentiles to create a more valid representation of the data. The user can then manually enter filter ranges or select them using a slider bar for interval variables or selecting **discrete** to exclude for categorical variables. Optionally, the score code will create a new variable that identifies observations selected for removal but not apply the **where** clause; thus, the node can be used to identify outliers and selected regions for subsequent processing.

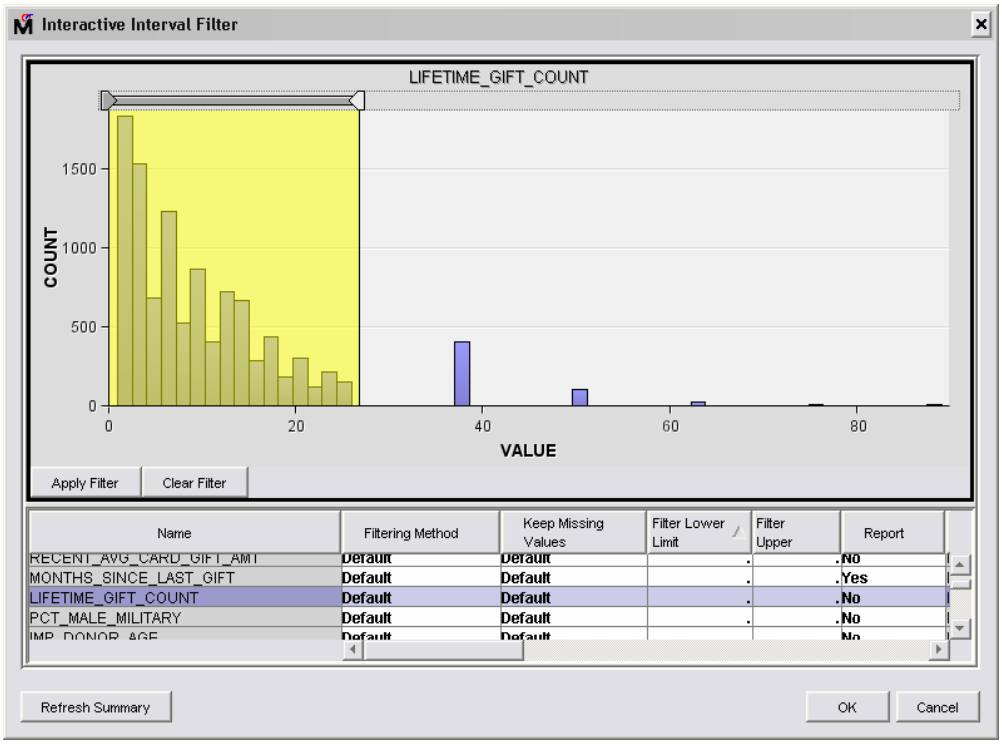

Figure 5. Filter extreme values interactively with the Filter node. The shaded region defines the variable range to keep.

The **SOM node** provides both data visualization and clustering capabilities. Self-organizing maps (SOMs) are a data visualization technique used to reduce the dimensions of data through the use of self-organizing neural networks. This process of reducing the dimensionality of vectors is essentially a data compression technique known as vector quantization. In addition, the Kohonen technique creates a network that stores information in such a way that any topological relationships within the training set are maintained. Figure 6 shows the distribution of the variable DONOR_AGE across the grid of SOM dimensions. The user can select any variable for display, even variables not used in building the SOM model.

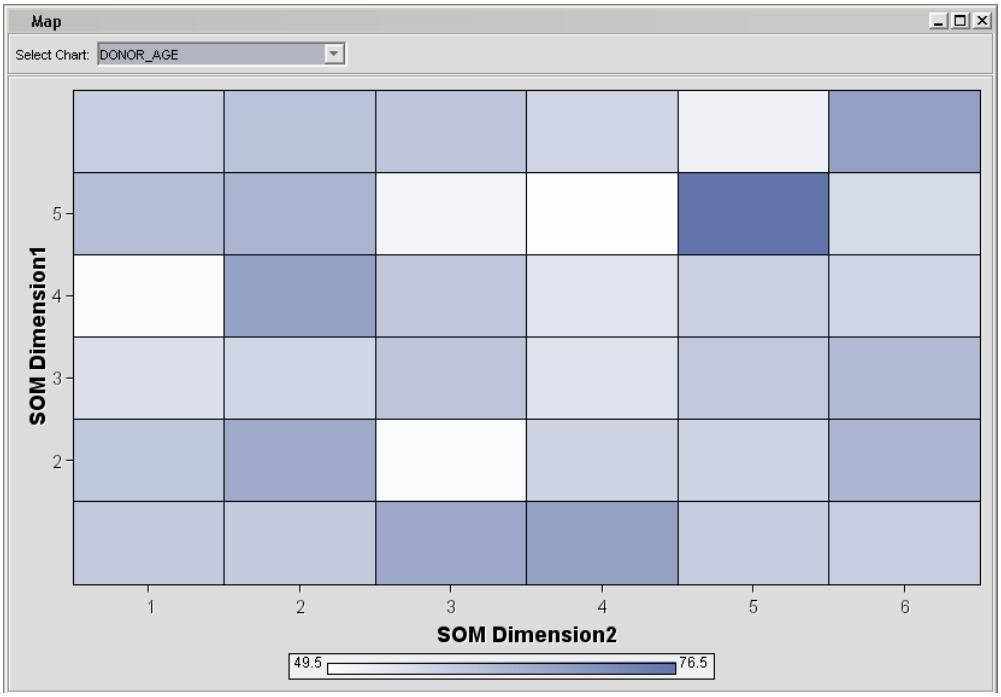

Figure 6. Cluster or reduce the dimensionality of the data using the SOM/Kohonen node.

The **Decision Node** can make a new decision for each case in training and scoring data sets based on numerical consequences specified via a decision matrix and cost variables or cost constants (Figure 7). This is useful for selecting models that maximize profit or minimize loss rather than relying on traditional statistical measures, such as misclassification rates or Bayesian Information Criterion. Users can define a decision matrix specifying profit, loss, or revenue, and define prior probabilities, which are often necessary when the sample proportions of the target classes in the historical training set differ considerably from the proportions in the operational data to be scored either through deliberate sampling bias or inherent variation. The decision function will multiply the decision values by the event probabilities to both select the optimal decision and also to estimate the profit or loss of the consequence. This facilitates what-if analysis where a user can change the profit or loss and cost parameters to estimate their effect on decision distributions. The node can also be used to apply the decision function to models created outside of Enterprise Miner and to create comparable fit statistics, lift charts, and distribution charts.

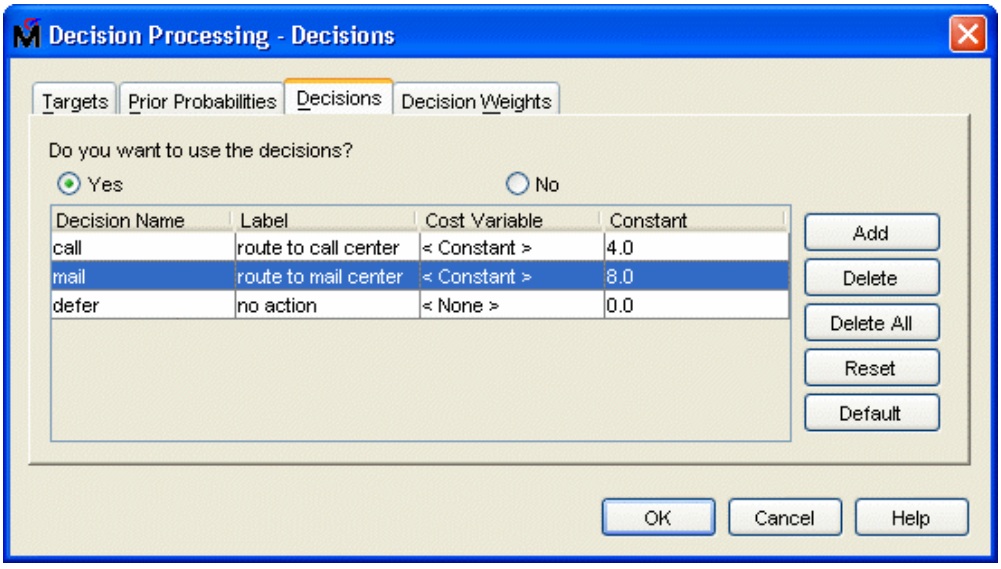

Figure 7. Configure decision processing using the Decision mode.

CHANGES TO EXISTING NODES

The **Decision Tree node** provides a tree growth iteration plot that displays the value of a model assessment measure on the vertical axis for different subtrees on the horizontal axis (Figure 8). A reference line is superimposed on the plot indicating the subtree selected as the final model. The iteration plot is very helpful in understanding how large a tree is needed for sufficient fit and whether large trees overfit the training data.

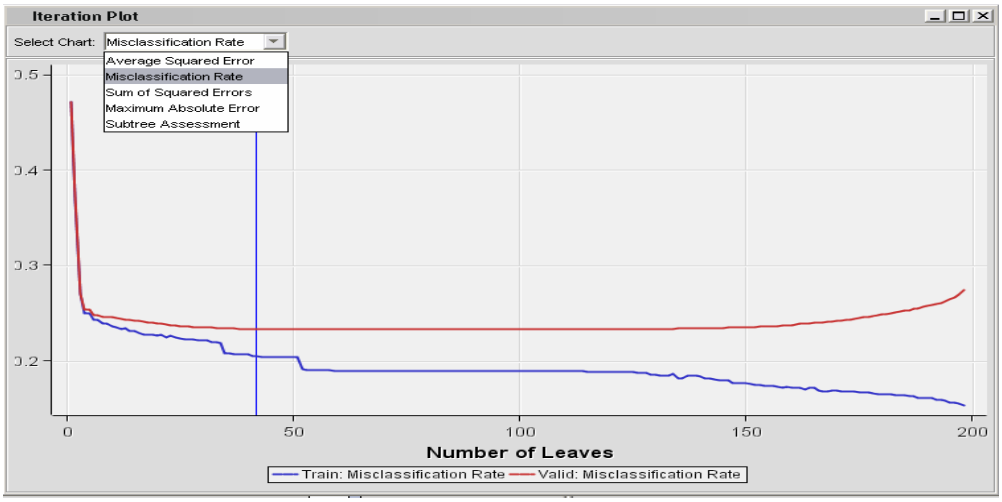

Figure 8. Evaluate different-sized decision trees using the Decision Tree Iteration Plot.

The **Enterprise Miner Tree Desktop Application** is used for interactive training or viewing decision trees created by the Decision Tree node (Figure 9). Recent enhancements include the ability to export all graphics to a JPG file, the ability to collapse and expand nodes in a tree display, and the choice of displaying variables labels or names. Users also gain greater flexibility in selecting splitting candidates when manually building the tree.

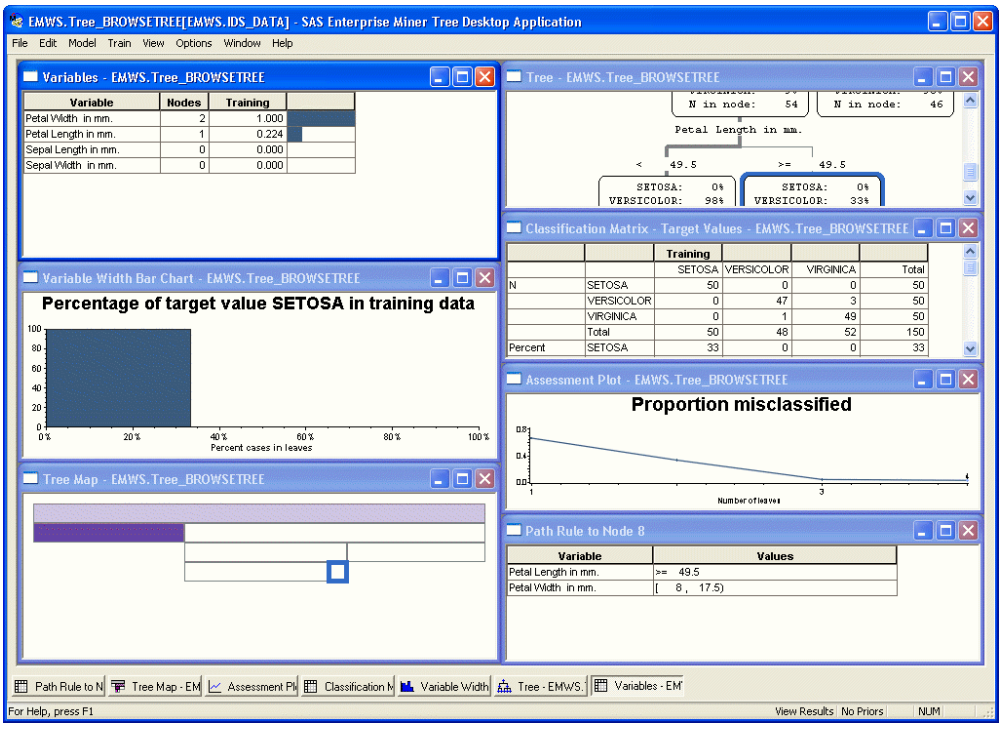

Figure 9. Build decision trees interactively with the Tree Desktop Application.

The **Regression node** now includes an iteration plot with a reference line for the selected step in the stepwise regression along with a combo box for choosing the display statistic. The new Estimate Selection Plot displays parameter estimates across the model selection steps (Figure 10). You can use the plot to characterize the size and stability of each parameter across iterations of the stepwise regression.

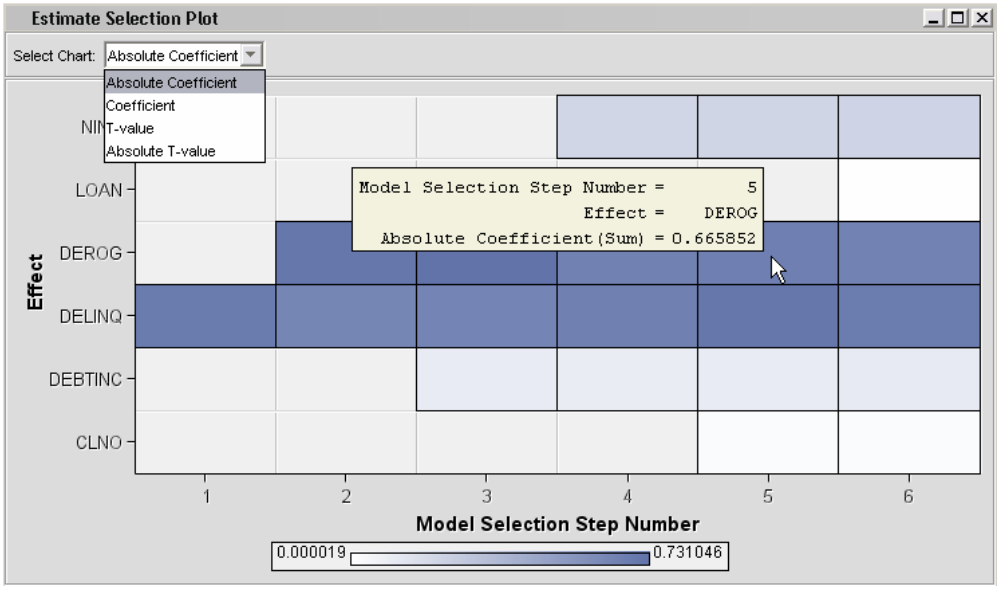

Figure 10. Characterize the overall size and consistency of the parameter estimates in a stepwise regression with the Estimate Selection Plot.

All modeling nodes benefit from improved assessment lift charts, classification charts, and distribution charts that have been updated with new controls for easily changing the assessment statistic displayed. You can more quickly switch from captured response to cumulative captured response, for instance, without finding the corresponding variable in the output table.

Changes have been made to the statistics and plots displayed in the **Model Comparison node** and also to the predictive modeling nodes to improve the consistency between reports. The Score Rankings and Score Distribution Plots now provide a selection control for conveniently changing the displayed statistics (Figure 11). The Model Comparison node now includes the Kolmogrov-Smirnov statistic in the results table. A baseline has been added to the Receiver Operating Characteristic (ROC) chart that is produced for binary targets.

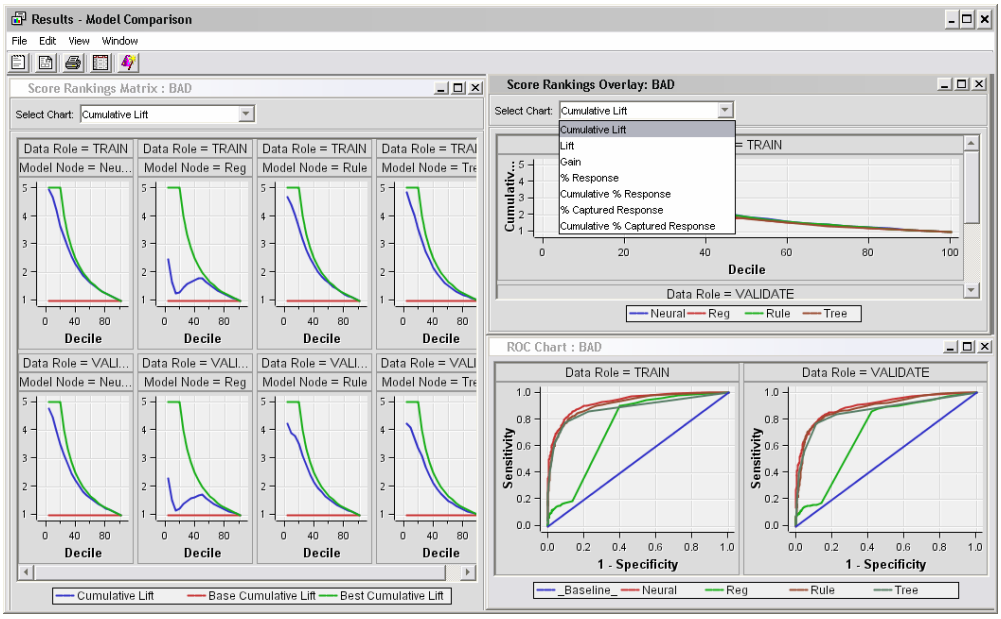

Figure 11. Evaluate multiple models together in one easy-to-interpret framework using the Model Comparison node.

The **Principal Components node,** used for both data exploration and dimension reduction, now provides a matrix plot that is useful in looking for patterns separated by the dimensions. The plots are color coded by the target event, if present (Figure 12). You select the the number of subplots that appear. The node now includes an interactive principal components selector to choose how many components should be exported for subsequent analysis.

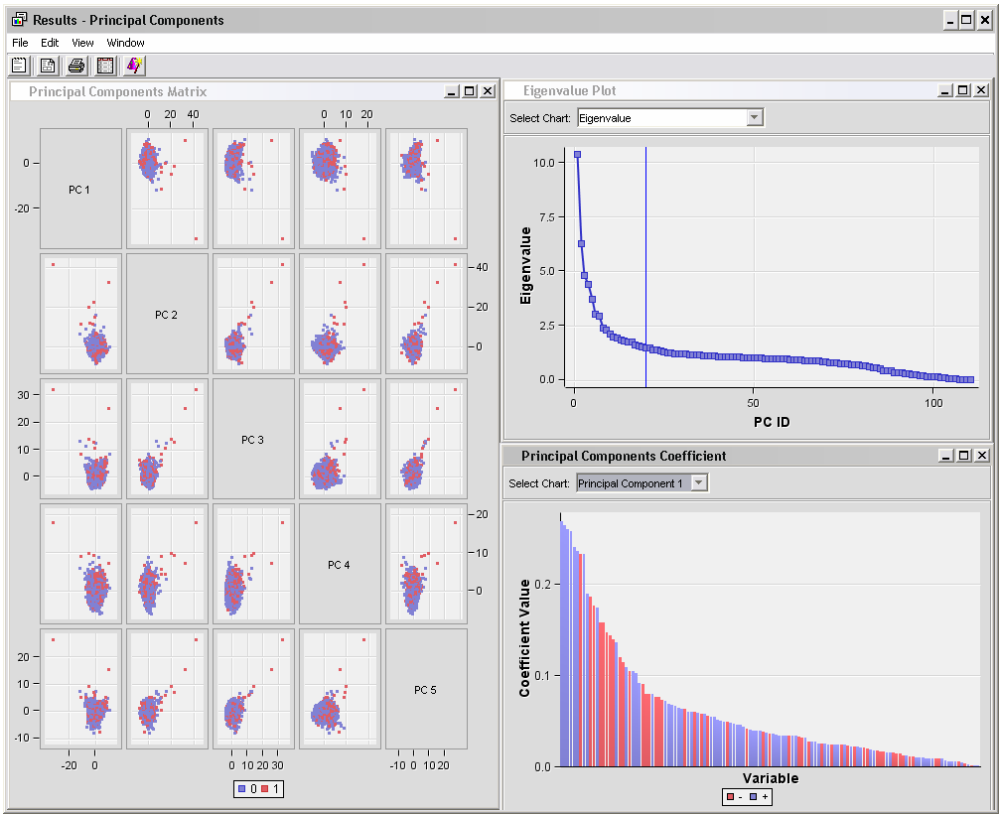

Figure 12. Visualize the output of principal components.

The **Path node** is used to search transactional data such as Web logs for frequent sequential patterns. New graphical enhancements have been added to help better explore the navigation habits of visitors. Plots include a funnel count plot depicting the drop-off in the number of visitors along a particular path of interest and a sequence rules matrix view.

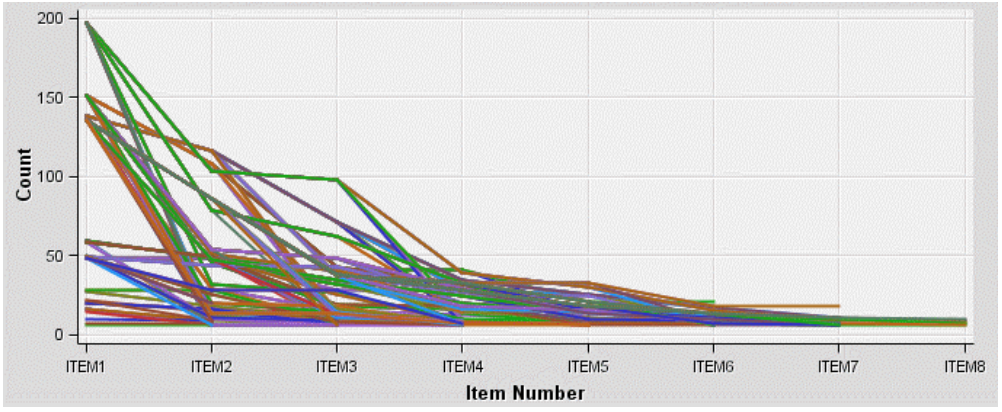

Figure 13. Use the funnel count plot to show the decay in Web session lengths.

The program editor of the **SAS Code node** has Macro Variables and Macros tabs that contain lists of macro variables and macros that are used in SAS training code. You can use your mouse to drag items from the Macro Variables and Macros lists and drop them in the SAS Code editor to enhance and simplify SAS code creation.

You now have better control of the sample makeup when using the **Sample node**. The Level Based value has been added as an option to the Criterion property in the stratified sample properties. If Level Based is selected, then the sample is based on the proportion captured and sample proportion of a specific level.

ADMINISTRATION AND CONFIGURATION

The following changes and enhancements improve the administration and configuration of SAS Enterprise Miner.

SAS ANALYTICS PLATFORM

The SAS Analytics Platform provides a common client/server architecture and implementation for a family of products that includes SAS Enterprise Miner, SAS Forecast Studio, and SAS Inventory Policy Studio. One instance of the Analytics Platform middle-tier server can serve all three applications. It is easier for SAS administrators to install and configure multiple SAS analytical applications.

In three-tier environments, administrators can monitor the status of the Analytics Platform server through both Web-based and client-based tools. Users can download and install the Enterprise Miner client directly from the Analytics Platform server. Administrators who configure multi-user environments must manually configure and start the Analytics Platform middle-tier services.

The SAS Analytics Platform cannot be licensed separately. The Analytics Platform installation is triggered by the installation of any of the existing SAS Analytics Platform products.

SAS MANAGEMENT CONSOLE PLUG-IN FOR ENTERPRISE MINER ADMINISTRATION

Some of the functionality that was in the configuration wizard in SAS Enterprise Miner 5.1 has been moved to a SAS Management Console plug-in. The SAS Management Console plug-in gives administrators better control over Enterprise Miner configurations. The following information is now stored in the SAS Management Console plug-in:

- server start-up code
- default project path and security
- maximum concurrent nodes per process execution
- model group definitions
- WebDAV location is now stored in the SAS Metadata Repository and configured by SAS Management Console. As a result, users never need to know about the WebDAV location that is used by the Model Manager functions. WebDAV is Enterprise Miner's Web-based Distributed Authoring and Versioning. It is a set of extensions to the HTTP protocol that enable you to collaboratively edit and manage files on remote Web servers.

GRID PROCESSING

Grid processing is available for enterprises that need to perform computing over multiple logical or physical systems. The execution of the process flow diagram in SAS Enterprise Miner is sent to a load balancing manager that distributes the jobs to a grid of systems. This is expected to benefit users who run multiple, large-process flow diagrams, or users who manage a large multi-user environment.

CONCLUSION

SAS is proud to present Enterprise Miner 5.2 to users at this year's SUGI conference. A significant amount of work has been devoted to meeting and exceeding users' demands. Progress has been made in data exploration, selection, and transformation; in new and enhanced reports for predictive models; in ease of use; and in system administration and multi-server grid processing. All these issues are expected to significantly enhance the user experience and lead to better predictive models for data mining.

REFERENCES

Administrator Guide for SAS Analytics Platform (2005).
http://support.sas.com/documentation/onlinedoc/apcore/index.html

What's New in SAS® Enterprise Miner™ 5.2 (2005). http://support.sas.com/software/91x/emgui52whatsnew900.htm

ACKNOWLEDGMENTS

Appreciation is extended to the entire SAS Enterprise Miner family (development, testing, technical support, education, publications, marketing, and field strategy/support) for helping bring this product to market. We are also very grateful to our customer base for all of the great feedback and support of our product. Thank you.

CONTACT INFORMATION

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

David Duling
Development Director
SAS Institute Inc.
SAS Campus Dr., S6102
Cary, NC 27513
Work Phone: (919) 531-5267
Email: david.duling@sas.com

Wayne Thompson
Product Manager
SAS Institute Inc.
SAS Campus Dr., S6100
Cary. NC 27513
Work Phone: (919) 531-6485
Email: wayne.thompson@sas.com