Paper 155-2008

Cool New Features in SAS® Enterprise Miner™ 5.3

David Duling, Wayne Thompson, Sascha Schubert
SAS Institute, Inc.

ABSTRACT

SAS released Enterprise Miner 5.3 in late 2007 with a veritable plethora of cool new features for data miners everywhere. Nearly every module of the software has been updated. New interactive data preparation tools make it easier to manipulate data and construct a sample for mining. For data exploration, Enterprise Miner now supports hierarchical market baskets to isolate interesting rules at different product category levels, multivariate graphical data exploration that persists a user's interactive selections, a new scalable variable clustering node for dimension reduction, and more interactive user control over feature selection. Variable creation has been enhanced with a new interactive binning tool, an interactive rule building tool, and new transformation options. There are three new core predictive modeling techniques in Gradient Boosting, Support Vector Machines, and Partial Least Squares, along with a tool to make it easier to import models previously produced with SAS/STAT® code. For model assessment, a new Cutoff node examines posterior probability distributions where users can enter cutoff values, and a new Reporter tool uses SAS ODS to produce reports spanning the entire analysis for printing and editing. The user interface is revised with more navigation controls, smarter property sheets, better graphics, and improved code editors. Users should see significant productivity gains from the software, and have even more fun data mining.

INTRODUCTION

SAS Enterprise Miner has been an industry-leading tool in the data mining field for nearly 10 years. This might lead you to believe that data mining products are in maintenance mode; however, that is most definitely not the case. On the contrary, the field of data mining is rapidly evolving to include new transactional and Web-based data sources; new applications such as social network analysis, rate making, and time series classification; and new modeling algorithms to detect global and local features. The latest release of Enterprise Miner contains a host of new productivity, statistical, interactive, and graphical tools designed to improve the productivity of the SAS data miner. This paper will focus on the new features in Enterprise Miner 5.3 with analytical examples.

MIGRATION

Before we can start data mining, we have to consider platforms and migration. Enterprise Miner 5.3 runs on SAS 9.1.3 Service Pack 4. Installation requires updates to the SAS Foundation, the SAS Analytics Platform, and the SAS Enterprise Miner client. Those users who need to preserve their Enterprise Miner 4.3 projects will find a new project conversion utility that moves all Enterprise Miner 4.3 diagrams into an Enterprise Miner 5.3 project. This function preserves the diagram structure, many of the node properties, and many of the tools results such as log and output listings, source and score code, and results tables needed for producing gains charts. The Enterprise Miner 4.3 result sets are visible inside the Enterprise Miner 5.3 Node Results window so that users can then run the diagrams in Enterprise Miner 5.3 and compare output. This will satisfy users' needs to archive and retrieve their Enterprise Miner 4.3 results from within Enterprise Miner 5.3. Users of Enterprise Miner 5.2 will not need to perform any migration action because these projects are directly usable in Enterprise Miner 5.3.

NEW FEATURES

The Enterprise Miner 5.3 documentation and product literature provide a detailed list of new and enhanced features. That list is too lengthy to discuss in detail in this paper. Instead, we will focus on a few key features that will affect users in the areas of usability, graphical exploration, feature selection, variable binning, group processing and model building, and post processing.

Usability

The first thing an Enterprise Miner 5.3 user will notice is a revised user interface that incorporates common design principles established for SAS software. The main interface element is the PFD (process flow diagram) that now includes navigation tools for easily moving around the workspace. Data mining is often an exploratory exercise where several options and alternatives are attempted before a final model is created, and we have found that users often have upward of fifty nodes in their diagrams. All interactive graphics sport more easily usable controls and cleaner, more technical renderings, along with several new multivariate graphs. Administrators will appreciate the ability to use SSL to secure authentications.

Figure 1 - Better Usability in the Primary User Interface

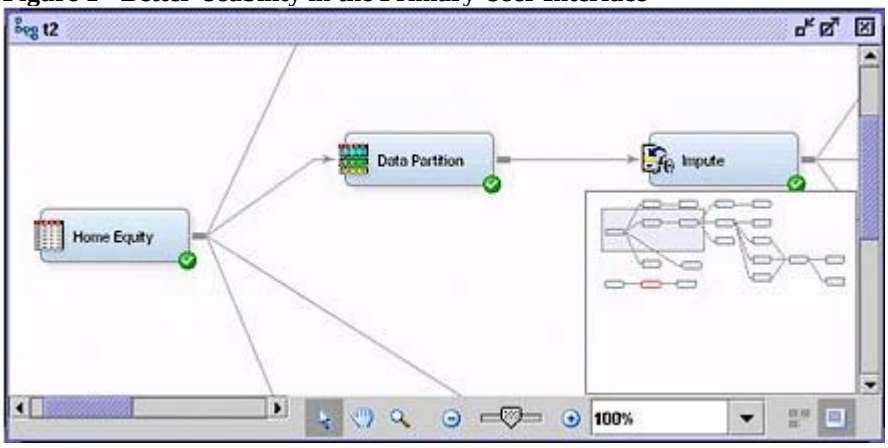


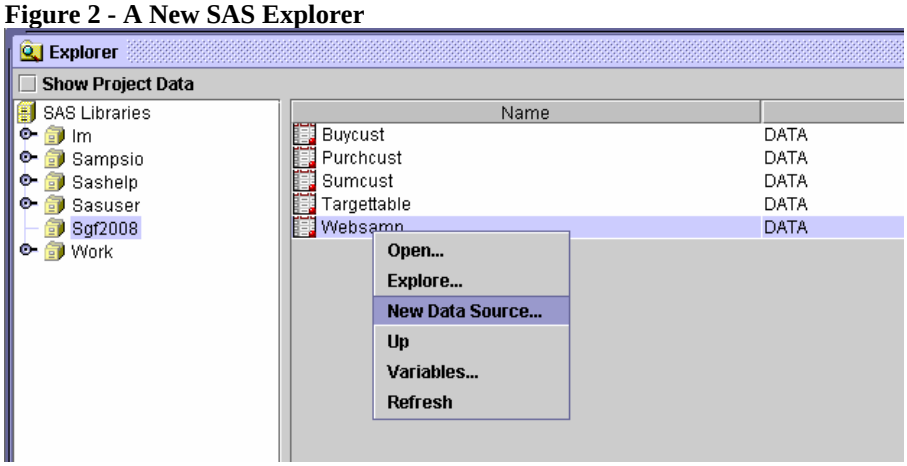
Diagram nodes have properties that control their behavior (for example, the size of a sample or the complexity of a neural network). These properties have been significantly reorganized into three categories in the property sheet:

- Training properties control the process of building a model. These properties have the greatest affect on run times.
- Scoring properties control the generation of score code, metadata, and exported data sets.
- Reporting properties control the generation of printed and graphical output.

This change gives the user the most direct control over the data mining process and will greatly improve efficiency. For example, changing the value of a scoring or reporting property will not force a retraining of a model, potentially saving long periods of run time.

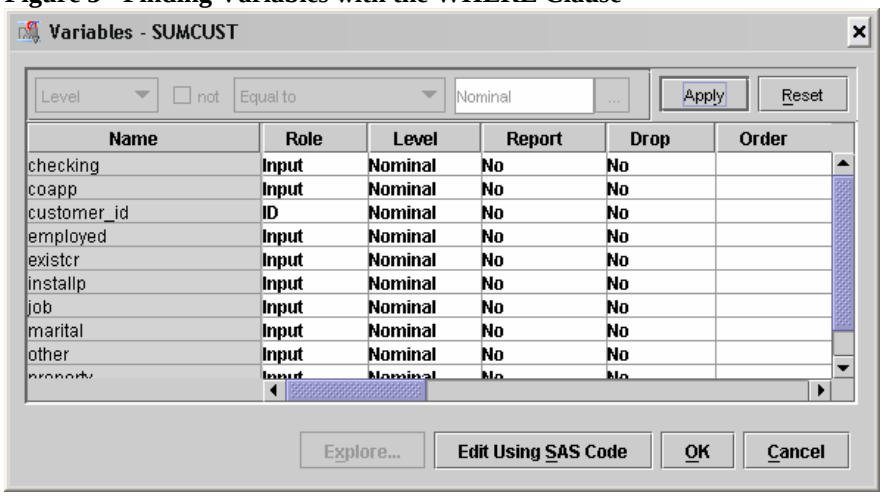
Users will find improved data access actions. A new library creation wizard will lead users through the process of creating SAS library names, an alternative to typing SAS code libref statements. Through this, users can also access Base SAS data and Microsoft Excel files. Once libraries have been created, table access is consolidated in the SAS Explorer window with actions for browsing large data, graphically exploring data, and a wizard for creating project data sources with metadata.

Figure 2 - A New SAS Explorer



Data mining tables often feature a large number of columns. To help users find and manipulate variables, Enterprise Miner 5.3 provides a new variables table with a WHERE clause that can be applied to any attribute in the metadata. For tables with thousands of columns, this feature will be a tremendous help.

Figure 3 - Finding Variables with the WHERE Clause



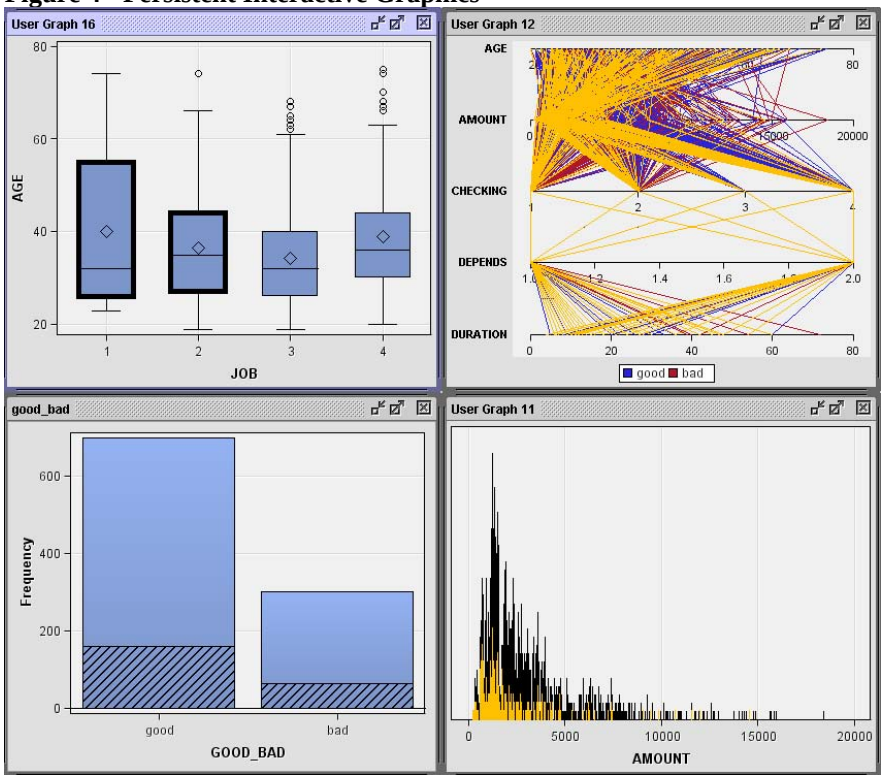
Graphical Exploration

As computer memory becomes more abundant and video technology improves, data mining applications increasingly rely on graphical exploration of data. Enterprise Miner 5.3 adds a new Graph Explore node for managing interactive plot creation. When the node is run, the training action will extract a data sample sized appropriately for downloading and display on the client. In the Results window, the sample is displayed and the user has full capability to create any number of these plots:

Scatter	Lattice Containers
Line	Parallel Axis
Histogram – 1D and 2D	Constellation
Density – 1D and 2D	3D Surface, Bar, Scatter
Box	2D Contour
Graphical Tables	Bar and Pie
Vector	Band
Scatter Matrix	Needle

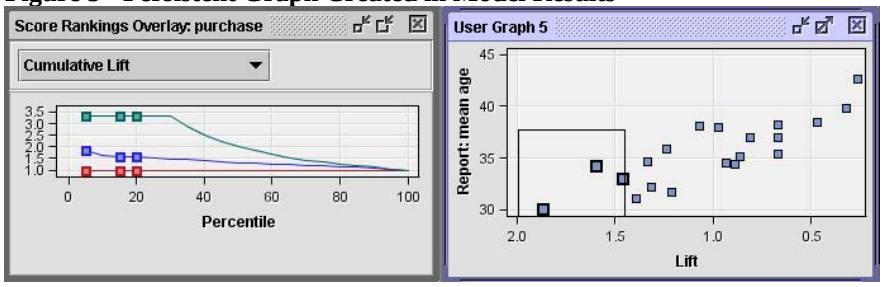
All plots are linked through the common data model so selections in one plot are visible in another. However, the really cool new feature is that all plots that are created by the user are persisted with the node and can be reopened at any time. This ability to persist interactively created graphs will greatly aid graphical ad hoc data exploration by saving the time to re-create graphs for continuing analysis or display for colleagues and managers. In Figure 4, the bar plot for the target variable GOOD_BAD was automatically created by the node, and the three other plots were created by the user. The region for the variable JOB with value “1” has been selected in the box plot and the corresponding regions highlighted in the three other plots.

Figure 4 - Persistent Interactive Graphics



Even better, the Results windows for all nodes have this behavior. For example, the results of any classification modeling tool will include a table that contains values for lift, gain, cumulative profit, and so on. Enterprise Miner also has a Report attribute for variables that is used in creating summary statistics. In this case, both the AGE and DURATION variables are enabled with the Report attribute and therefore have been summarized at every percentile. The user can now make a plot of average AGE by Lift and see a decreasing relationship: younger customers are predicted to provide more lift. No coding is required, and the plots are saved with the diagram and project.

Figure 5 - Persistent Graph Created in Model Results



Interactive Programming

One of the joys of being a SAS user is writing SAS code. Even though Enterprise Miner does present a nice user interface, writing code is still a great way to build and extend your analysis. Many customers use Enterprise Miner as a project organizer for large amounts of SAS code. Enterprise Miner 5.3 users will find a vastly improved SAS code node interface. You can enter training, scoring, and reporting code directly, run immediately for development and debugging, and then run the entire analysis path without leaving the Code Editor window. In this example, we use the following code and actions:

Training code	`proc means data=&em_import_data; run ;`
Scoring code	`logamt= log(amount) ;`

Reporting code	`proc means data=&em_export_train ;`
	`var amount logamt ;`
	`run ;`

1. First we enter the training code and press F3 or the Run toolbar icon to submit the code. The macro variable &EM_IMPORT_DATA points to the training data that is imported into the node. Examining the log and output shows a strong left skew for the variable AMOUNT, which is not unexpected.

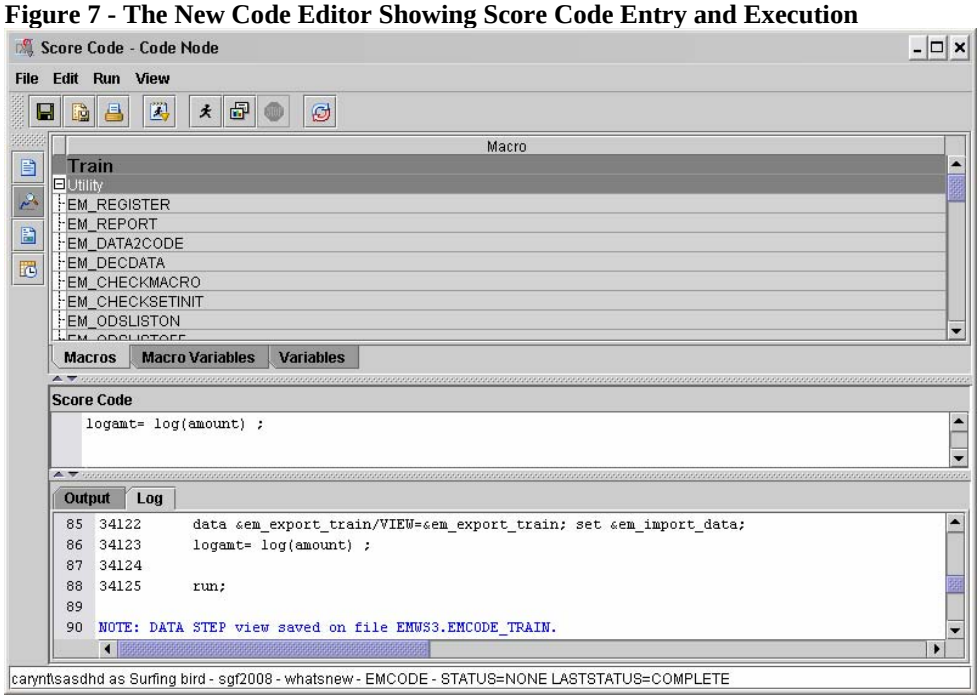
Figure 6 - Output from Running Training Code

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
checking	1000	2.5770000	1.2576377	1.0000000	4.0000000
duration	1000	20.9030000	12.0588145	4.0000000	72.0000000
history	1000	2.5450000	1.0831196	0	4.0000000
amount	1000	3271.26	2822.74	250.0000000	18424.00
savings	1000	2.1050000	1.5800226	1.0000000	5.0000000
employed	1000	3.3840000	1.2083063	1.0000000	5.0000000
installp	1000	2.9730000	1.1187147	1.0000000	4.0000000
marital	1000	2.6820000	0.7080801	1.0000000	4.0000000
coapp	1000	1.1450000	0.4777062	1.0000000	3.0000000
resident	1000	2.8450000	1.1037179	1.0000000	4.0000000
property	1000	2.3580000	1.0502090	1.0000000	4.0000000
age	1000	35.5460000	11.3754686	19.0000000	75.0000000
other	1000	2.6750000	0.7056011	1.0000000	3.0000000
housing	1000	1.9290000	0.5312638	1.0000000	3.0000000
existcr	1000	1.4070000	0.5776545	1.0000000	4.0000000
job	1000	2.9040000	0.6536140	1.0000000	4.0000000
depends	1000	1.1550000	0.3620858	1.0000000	2.0000000
telephon	1000	1.4040000	0.4909430	1.0000000	2.0000000
foreign	1000	1.0370000	0.1888562	1.0000000	2.0000000

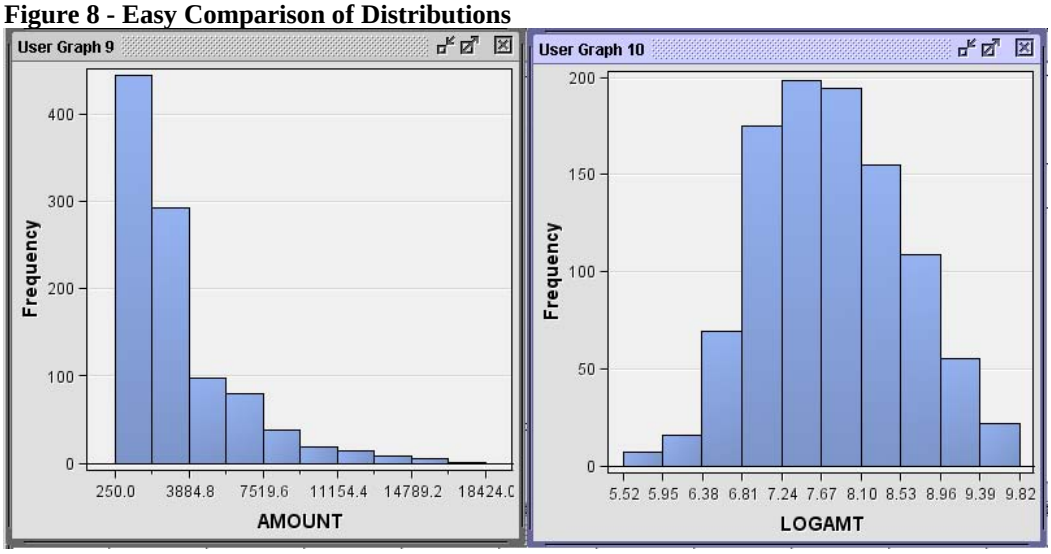
2. Thus we enter the scoring code for creating a new variable named LOGAMT and press F3 again. The result is shown in Figure 7. Enterprise Miner has wrapped our score code inside DATA, SET, and RUN lines, and then executed the entire DATA step for each of the Train, Validate, and Test data partitions. Enterprise Miner will also add our scoring code to the total score code for the diagram, along with all the score code that is generated by the Enterprise Miner nodes.
3. Finally, we enter the report code and press F3 again to produce an output listing that compares the two variables.

Figure 7 - The New Code Editor Showing Score Code Entry and Execution



At this point, the user has run a basic exploration very easily, transformed a variable, saved score code, and viewed the results. We can now follow this code with another Graph Explore node to visually compare the distributions of AMOUNT and LOGAMT.

Figure 8 - Easy Comparison of Distributions

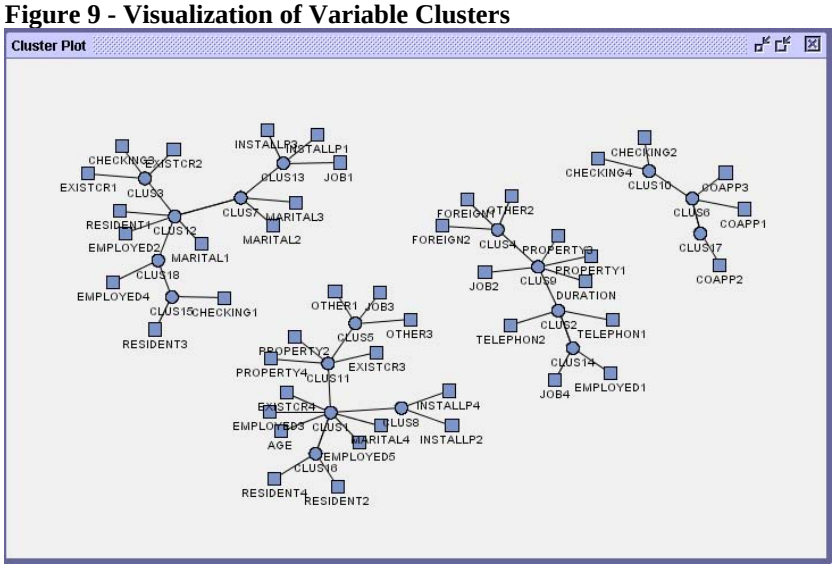


Variable Clustering

The explosion of features that are stored in databases makes the problem of feature selection increasingly more important. Many data miners need to create models that will run in production environments for up to two years or more. These models must be robust to unexpected changes in population distributions; that is, small changes in relative characteristic values should not create large changes in relative score values. Variable collinearity can often be the cause of these unexpected changes. One technique that is used by statisticians to address collinearity is variable clustering. SAS/STAT includes the VARCLUS procedure, which groups variables together into clusters by their

relative values. Enterprise Miner now uses PROC VARCLUS in a two-stage algorithm so that users can run variable clustering on very large data sets. Several diagnostic statistics and visualizations are generated by the new node. In the following cluster plot, squares represent the individual terms and circles represent the cluster centers. The names of the variables that have been grouped together show obvious similarity relationships. An analyst can interactively choose to select individual terms from each cluster, or select the latent cluster components for export, or both. An interactive control makes it easy for the user to select which terms from each cluster should be used in further model building. In this example, we will select terms rather than cluster components and follow the variable clustering node with a Logistic Regression function to build a robust model with a minimal number of largely uncorrelated terms.

Figure 9 - Visualization of Variable Clusters

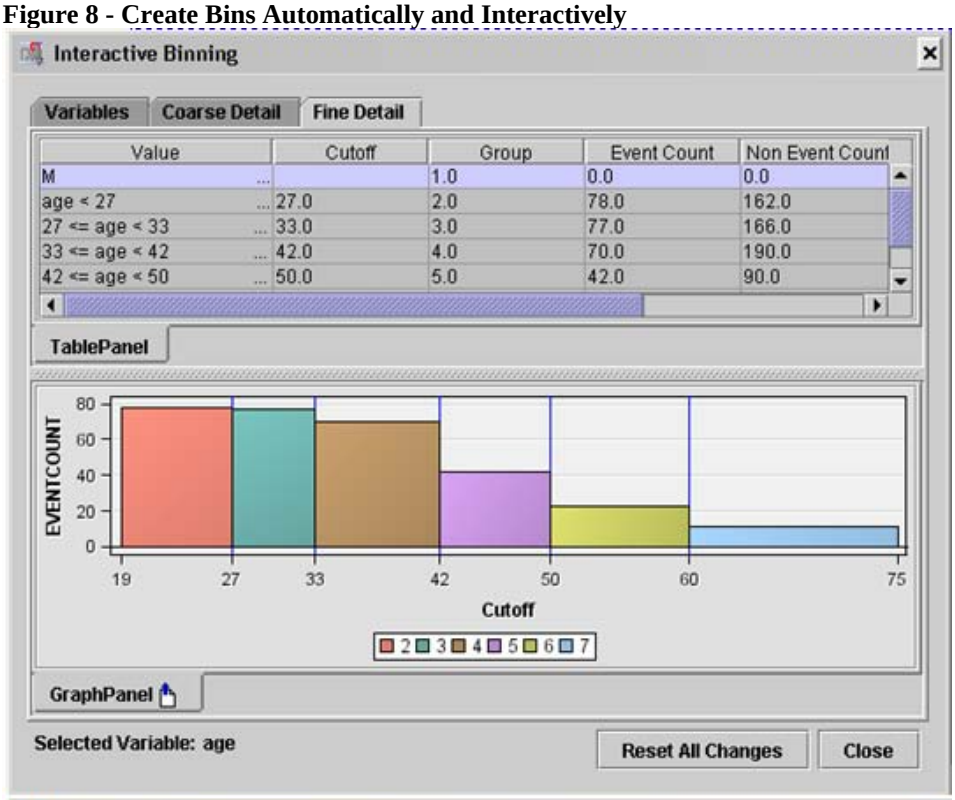


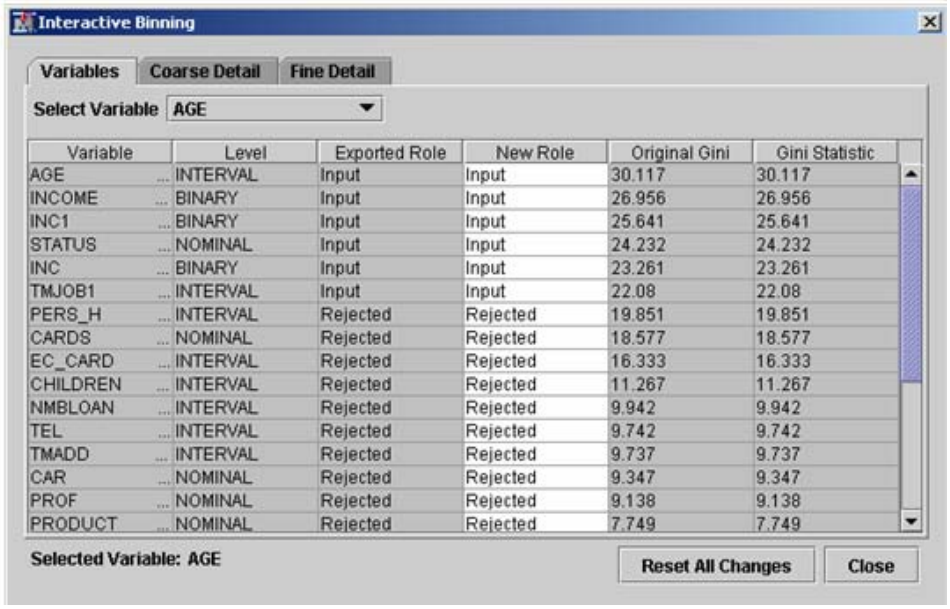
Variable Binning

Many data miners need to discretize continuous data or group together categorical values for either building additive models such as scorecards, or to better model nonlinear relationships in their data. Enterprise Miner 5.3 now contains a special node and interactive interface just for binning data. When the node runs, the train action will create an automatic set of bins based on quantiles that contain equal numbers of observations or buckets that span equal ranges. The node also implements variable selection based on the Gini criteria. In our example, eight variables were retained and thirteen were rejected.

Users can then enter an interactive mode and change the bins to match their needs. In the following plot, the quantile algorithm binned all values above 60; however, the data miner needed to create new bins at values 50 and 60. This is often done to conform to a business policy or industry regulation as well as the intuition of the data miner.

Figure 8 - Create Bins Automatically and Interactively



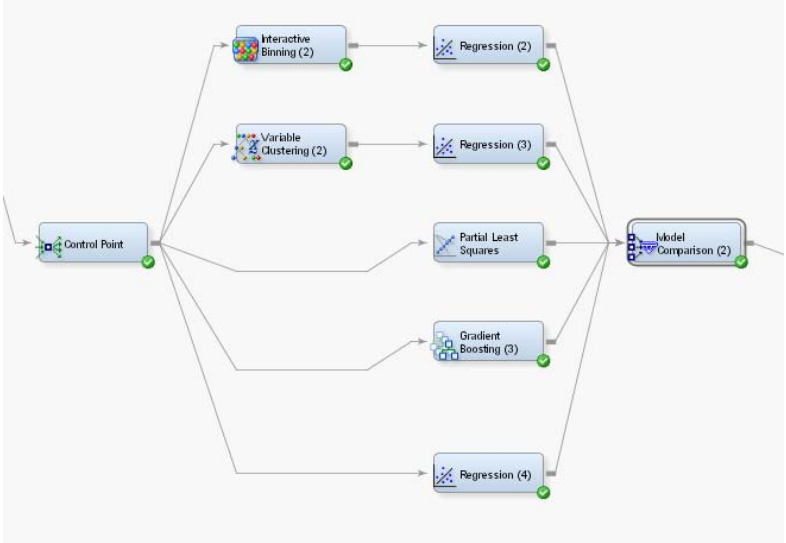


Modeling

Now that we have selected terms by clustering and binned our continuous variables, we are ready to build models. We can use the Logistic Regression tool to build reference models, but we also have two new production model tools we can try on this data. First, we have a new Partial Least Squares model node that implements the PLS procedure from SAS/STAT into the Enterprise Miner process. The PLS options have been selected for data mining, and score code is generated for deployment to production systems. The PLS model is particularly good at including latent effects in the model and is suited for both wide and tall data. Our second new modeling technique is found in the Gradient Boosting node and the

new GRADBOOST procedure in Enterprise Miner. This procedure implements the stochastic gradient boosting methodology from recent machine learning literature. The technique combines a series of decision trees that are trained on residuals defined in terms of the derivative of a loss function. Gradient Boosting models are thought to be adaptive to a wide variety of structures found in data and are sometimes claimed to be the best general purpose classifier. One extra benefit of this node is a new measure of variable importance based on the number of times a variable is used in splits in all the models. This measure is found in the results of the Gradient Boosting node.

Figure 9 - Comparing Multiple Models



We can now easily compare these models on a common statistic. Enterprise Miner 5.3 gives the user much greater control over automatic selection of models. In this case, we have selected the top 10% lift computed on the validation data.

Figure 10 - Flexible Model Selection in the Model Comparison Node

Model Selection	
Selection Table	Validation
Selection Statistic	Lift
Selection Depth	10

In the following model comparison, we can see that while the Gradient Boosting node shows the lowest validation misclassification rate, the combination of variable binning and logistic regression produces the best 10% lift on validation data. Remembering that variable binning incorporated an aggressive variable selection, we can infer that that method was well placed to avoid overfitting. Models selection order often changes with different selection statistics.

Figure 11 - Model Comparison and Selection

Fit Statistics
Model selection based on _VLIFT_

Selected Model	Model Node	Valid: Lift	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error	Valid: Misclassification Rate
Y	Reg2	1.42381	0.14701	0.2050	0.17910	0.27090
	PLS	1.13905	0.16211	0.2000	0.16614	0.27425
	Boost3	1.13810	0.14294	0.1975	0.17119	0.25418
	Reg4	1.13810	0.12604	0.1925	0.21061	0.29097
	Reg3	1.13333	0.12910	0.2050	0.21300	0.29097

Group Processing

For data with multiple natural BY groups, analytically computed BY groups, or multiple target variables, a single modeling process can be repeated many times. Automating this task is a key productivity enhancement in Enterprise Miner 5.3. Use the Start Groups and End Groups nodes to define a diagram segment that will be repeated many times.

Because our champion model was produced by output from the Variable Binning node, we can try to improve on this model by building clusters to find groups of observations that can be treated with the same model. Within each cluster, we will build Neural Network, Decision Tree, and Logistic Regression functions. The Model Comparison node will select the best function for each cluster. The End Group node will append the functions together inside IF statements for each cluster. The final result is a single model that has independently trained functions for each computed cluster. Using an independent model function for each cluster has the potential for tuning the predictions to local structure found in each cluster.

Figure 12 - Group Processing with Analytical Clusters

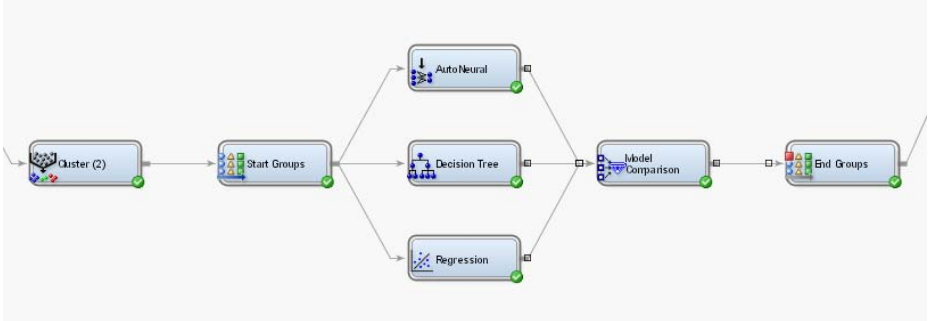


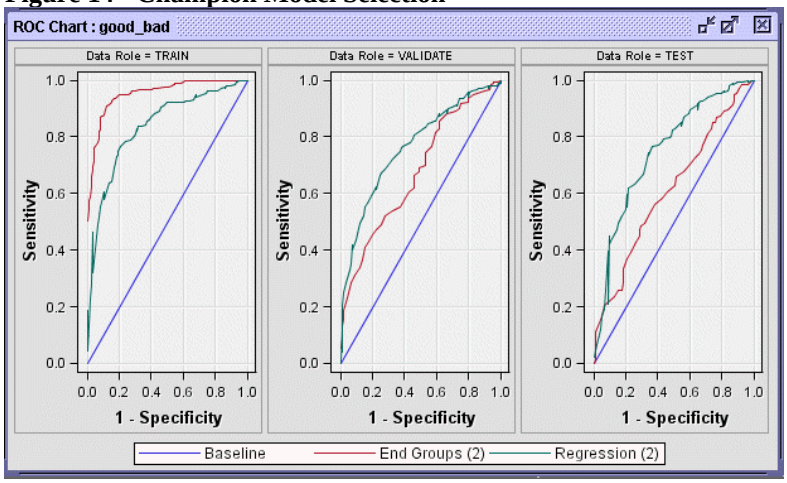
Figure 13 displays the model selection for each cluster. Notice that different model functions were selected in each cluster (Neural, Regression, Neural, and Tree), and also that cluster 3 shows a higher lift (1.736) than any other cluster. If we are looking for a model group to target, we might want to profile that group for differences with the rest of the sample.

Figure 13 - Model Selection for Each Cluster

Loop Number	Selected Model	Predecessor Node	Model Node	Valid: Lift
1	Y	AutoNeural	AutoNeural	1.305556
1		Reg5	Reg5	1.175
1		Tree2	Tree2	1.133124
2	Y	AutoNeural	AutoNeural	1.359375
2		Reg5	Reg5	1.273437
2		Tree2	Tree2	1
3	Y	Tree2	Tree2	1.736842
3		AutoNeural	AutoNeural	1.736842
3		Reg5	Reg5	0.315789
4	Y	Reg5	Reg5	1.069959
4		Tree2	Tree2	1
4		AutoNeural	AutoNeural	0.888889

Finally, we can compare our new stratified model with our previous champion by using a new Model Comparison node to select the grand champion. Once again, the minimalist model is selected based on 10% lift computed on validation data. While the stratified model did improve the training statistics, it did not generalize as well and lost the race on the validation data. This is clearly evident in the ROC chart shown in Figure 14. We can now proceed with confidence that the grand champion model does the best job of balancing model performance and robustness.

Figure 14 - Champion Model Selection



REPORTING

We need to report our findings to our colleagues, managers, auditors, and regulators. Enterprise Miner 5.3 adds a new Reporter node that uses SAS ODS to produce PDF and RTF documents that contain all information about the variables, functions, parameters, and even the graphs displayed in the Node Results windows. This includes critical information such as variable selection, model fit statistics, and lift charts. The user can choose to run the report for the entire diagram, the analysis path, or only the predecessor node. These reports are expected to make life much easier for the perpetually overworked data miner.

CONCLUSION

With the release of SAS Enterprise Miner 5.3, analysts have a great set of cool new features for the full range of data mining activities. This paper has presented only a subset of available new tools. For complete information about changes and enhancements in Enterprise Miner 5.3, visit the SAS Web site for the fact sheet at the address listed in the references.

REFERENCES

Friedman, Jerome H. 2002. "Stochastic Gradient Boosting." *Computational Statistics & Data Analysis* 38: 367–378.

Tobias, R. 1995. "An Introduction to Partial Least Squares Regression." *Proceedings of the Twentieth Annual SAS Users Group International Conference.* Cary, NC: SAS Institute Inc. 1250–257.

Installation and Configuration Instructions for SAS Enterprise Miner 5.3. 2007. SAS Institute, Inc. Cary, NC. http://support.sas.com/documentation/onlinedoc/miner/install53.pdf

What's New in SAS Enterprise Miner 5.3. 2007. SAS Institute, Inc. Cary, NC. http://support.sas.com/documentation/whatsnew/91x/emgui53whatsnew.htm

SAS Enterprise Miner 5.3 Fact Sheet. 2007. SAS Institute Inc., Cary, NC. http://www.sas.com/technologies/analytics/datamining/miner/factsheet.pdf

ACKNOWLEDGMENTS

The authors thank the entire Enterprise Miner Development staff for its dedication to developing the software and for its help in creating this paper.

CONTACT INFORMATION

David Duling
SAS Institute, Inc.
SAS Campus Drive
Cary, North Carolina, 277513
David.Duling@SAS.com

Wayne Thompson
SAS Institute, Inc.
SAS Campus Drive
Cary, North Carolina, 277513
Wayne.Thompson@sas.com

Sascha Schubert
SAS France
Domaine de Grégy
Grégy-sur-Yerres
77257 Brie Comte Robert Cedex
Sascha.Schubert@sas.com