Paper 114-2009

The Next Generation: SAS Enterprise Miner™ 6.1

Wayne Thompson and David Duling, SAS Institute, Cary, NC

ABSTRACT

SAS Enterprise Miner 6.1 delivers several new product enhancements to empower both business analysts and seasoned data miners to work more efficiently and produce improved results. A file import node is included for easy access to a broad range of input source types. Extended summary statistics for input variables are also generated to aid the analysts in defining variable roles and identify upfront data errors and trends during the data source definition process. The algorithmic suite has been extended to include linear and nonlinear Support Vector Machine learning along with powerful LARS and LASSO variable selection. Interactive decision-tree users will be able to model multiple targets for multi-objective segmentation strategy building and predictive modeling. The Reporter node includes new graphics for delivering an analysis-ready report journal to business constituents. Data miners can also share models using the new Model Viewer application. A major emphasis has been placed on extended model deployment capabilities to include market-basket scoring, native scoring in Teradata, and optimized scoring code for delivering faster answers.

INTRODUCTION

SAS Enterprise Miner is the premier data mining solution, delivering an abundance of descriptive and predictive modeling tools along with extensive model deployment alternatives. SAS Enterprise Miner 6.1 released in the first quarter of 2009 continues the SAS dedication to data mining productivity and deployment in which customer feedback helps drive a leading research and development staff. SAS Enterprise Miner 6.1 includes a vast range of new features including integration with the SAS 9.2 System, extended data preparation, enhanced modeling capabilities, improved reporting, and new scoring alternatives.

This paper provides an overview of the SAS Enterprise Miner 6.1new features and uses an example of mining data about charitable donations to illustrate some of these features.

OVERVIEW OF SAS ENTEPRISE MINER 6.1

SOFTWARE REQUIREMENTS

SAS Enterprise Miner 6.1 requires SAS 9.2 Platform. The SAS 9.2 system is an improved platform for managing and deploying analytical and business intelligence applications for both single-user applications and multi-user enterprises. SAS Enterprise Miner 6.1 contains changes related to the SAS 9.2 system that improve SAS Enterprise Miner installation, security, and administration.

SOFTWARE MIGRATION

SAS Enterprise Miner 5.3 does not operate with SAS 9.2. If you have existing SAS Enterprise Miner 5.3 project information stored in your SAS Metadata Server, the project information is converted from SAS 9.1.3 format to SAS 9.2 format during the SAS 9.2 / Enterprise Miner 6.1 installation. If you have existing SAS Enterprise Miner 5.3 project data folders that are stored on SAS Workspace Servers, the project data folders do not require conversion for use with SAS 9.2 and SAS Enterprise Miner 6.1. All SAS Enterprise Miner 5.3 project data folders, files, tables, views, and catalogs that are stored on SAS Workspace Servers are compatible for use with SAS 9.2 / Enterprise Miner 6.1.

In SAS Enterprise Miner 6.1 you can open existing SAS Enterprise Miner 5.3 projects without any manual conversion process. SAS Enterprise Miner 6.1 projects cannot be converted for use with SAS Enterprise Miner 5.3.

If you want to upgrade SAS Enterprise Miner 4.3 project data for use with SAS Enterprise Miner 6.1 you can use the Enterprise Miner project conversion macro. The project conversion macro upgrades SAS Enterprise Miner 4.3 project structures to SAS Enterprise Miner 5.3 project structures. SAS Enterprise Miner 6.1 opens Enterprise Miner 5.3 project structures the user creates by the SAS Enterprise Miner Project conversion macro.

PROJECTS

SAS Enterprise Miner 6.1 project information is now stored and managed in the SAS Metadata Folders. SAS Enterprise Miner 6.1 users create projects in a specific folder location. The default location for new SAS Enterprise Miner 6.1 projects is My Folder. The My Folder location is unique for every user and is a private location. When you create a SAS Enterprise Miner 6.1 project, you can accept the default project location or specify a different folder of your own.

For example, either individually or a member of a group, you might store mining projects in a common folder where the projects can be shared. You open projects by using a standard Open Project window that displays the SAS metadata folders tree structure by default (Figure 1).

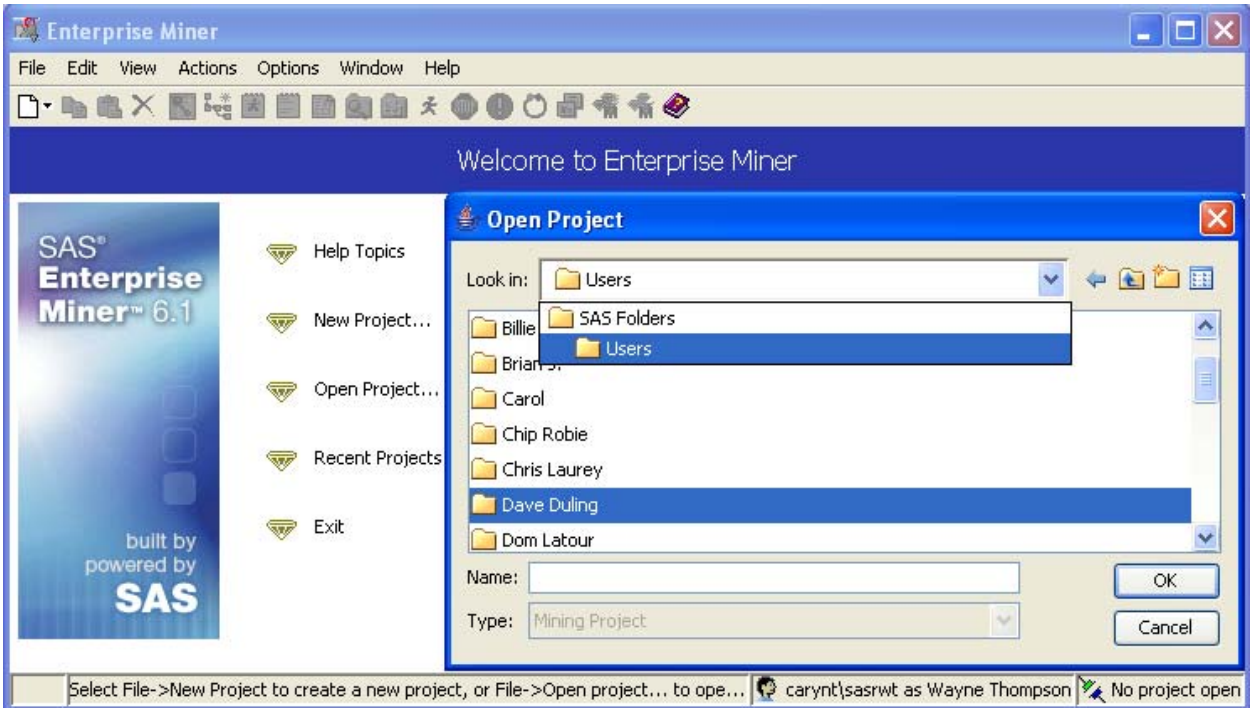


Figure 1. SAS Enterprise Miner Open Projects Window

When the SAS Metadata Server is upgraded from SAS 9.1.3 to SAS 9.2, existing SAS Enterprise Miner 5.3 project information that was stored in SAS Metadata Server is migrated to the shared data folder. SAS administrators can view SAS Enteprise Miner 6.1 project information via the SAS Management Console.

MODELS

SAS Enterprise Miner 6.1 models are stored and managed in the SAS Metadata Folders. You register models to a specific folder location. You can now open or import models by using a standard Open window that displays the SAS metadata folders tree structure by default.

When a SAS Metadata Server is upgraded from SAS 9.1.3 to SAS 9.2, existing SAS Enterprise Miner 5.3 models that were stored in SAS Metadata Server are migrated to the shared data folder. SAS Administrators can view SAS Enteprise Miner 6.1 model information via the SAS Management Console.

USABILITY

The user interface for SAS Enterprise Miner 6.1 been updated to include quick search code editors, faster dynamic sample generation for creating interactive plots, a totally new interactive decision tree component , and extended model import capabilities.

CODING IMPROVEMENTS

The SAS code editors and text viewers have been enhanced with a quick text search toolbar that highlights and navigates between selected text search results. This is a great aid when searching for text in SAS code, the SAS log, and SAS output listings. You can launch Quick Text Search from the SAS Enterprise Miner 6.1 main menu, or by using as a keyboard shortcut.

The Project Start Code Editor window has been modified to include the SAS log. Convenient access to the SAS log helps when you need to debug or modify Enterprise Miner project start code. The Project End Code Editor window has been eliminated.

INTERACTIVE GRAPHIC SAMPLES

Previous versions of SAS Enterprise Miner provided interactive exploratory graphics based on a sample of values in variable list tables. In Enterprise Miner 6.1, the sample table the software uses to generate interactive graphics has been improved to include only the attribute columns that the user selects, plus any additional Target, ID, Frequency, or Cost variables. This reduces the number of columns required to perform interactive graphic sampling and increases the number of rows of data that are available for graphics.

In addition, you can perform variable table list sampling for interactive graphics using a sampling algorithm that is stratified by categorical target variables. This change improves the representation of the sample in the presence of skewed data.

MODEL IMPORT AND EXPORT

In SAS Enterprise Miner 6.1 you can register models directly to the SAS metadata folders tree structure. This provides you with more control over the security, access privileges, and organization of models.

You can import a registered model into an existing data mining process flow diagram by using the Model Import node. The score code of the imported model is applied to the data in the process flow diagram, generating new model assessment statistics.

The Model Repository window has been removed from SAS Enterprise Miner 6.1. The former flat list of registered models has been replaced by a hierarchical view of models in the SAS metadata folders. The Model Import node provides a list of available models. You can select File --> Open Model from the main menu to open a file utility window to browse the SAS Metadata Folders tree structure and choose a model for inspection. You can also use the Model Import tool to navigate the SAS metadata folders tree structure and choose a model for addition to the process flow diagram.

INTERACTIVE TREE

The Tree Desktop Application has also been replaced by an entirely new interactive decision tree component which requires no separate install or documentation. SAS Enterprise Miner 6.1 enables users to invoke the software using Java Webstart. Examples of the new interactive tree component are provided in the Extended Modeling Features section.

DATA MANAGEMENT AND SUMMARIZATION

Preparing representative input data sources is one of the most important and often tedious tasks of data mining. The process of defining and summarizing data sources for analysis in Enterprise Miner 6.1 has been improved tremendously.

FILE IMPORT NODE

The new File Import node provides a quick and easy-to-use tool for non programmers to import external data files into their SAS Enteprise Miner process flows. The File Import node is located on the Sample tab of the SAS Enterprise Miner tool bar. The node supports importing dBase files, Stata, Microsoft Excel .XLS files, SAS .JMP files, Paradox .DB files, SPSS .SAV files, Lotus .WK1, .WK3, and .WK4 files, in addition to tab-delimited .TXT files, comma-delimited .CSV files, and user-defined delimited .DLM files. The data must be located on the SAS Enterprise Miner client machine or in a network location accessible to the SAS Workspace server (Figure 2).

The File Import node provides properties for controlling the maximum size of file to be imported along with other useful options, such as the number of rows to skip, the delimiter that separates columns, and whether or not to use the Data Source Advance Advisor for allocating variable roles. You can also preview before importing them to verify that the data structure is consistent and correct with your expectation.

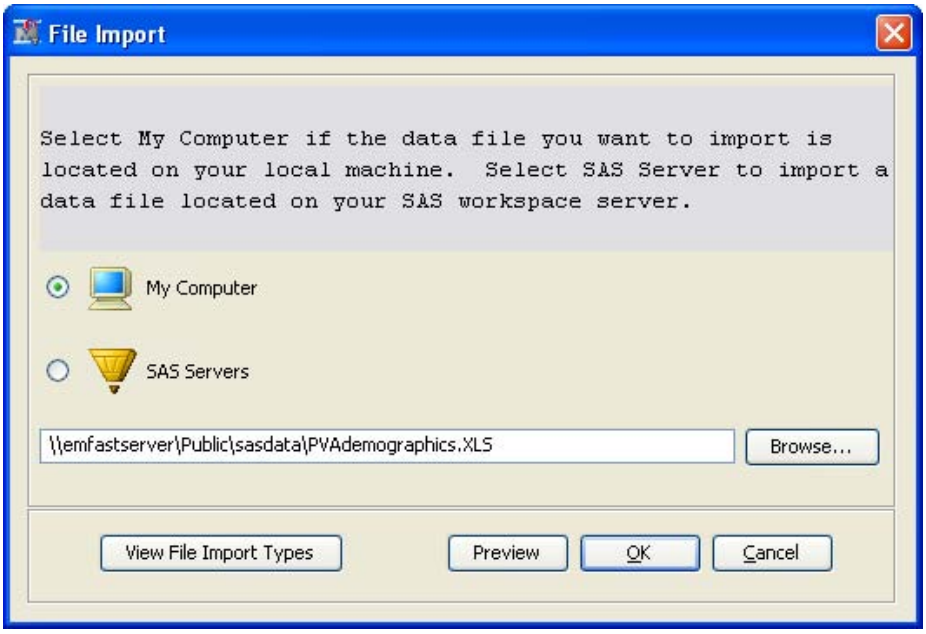
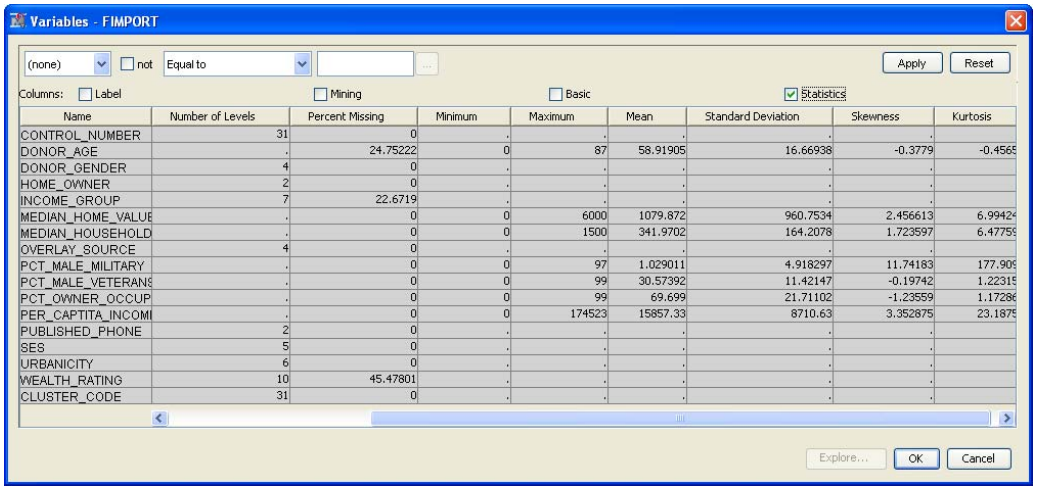


Figure 2. File Import Node Window.

You can redefine variable roles and control what column metadata about the variables is displayed in the variables table display. Instead of displaying enormous tables that have many variable attribute columns, you can configure the variables table by selecting only those attributes that are important to their work.

The new descriptive statistics are especially helpful for identifying trends and making decisions about how to treat variables downstream in the analysis. In Figure 3 DONOR_AGE, INCOME_GROUP, and WEALTH_RATING have a high percentage of missing values. Many of these interval variables also have large standard deviation and extreme skewness statistics suggesting that you may want to try transforming the data prior to fitting a regression model.



Name	Number of Levels	Percent Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
CONTROL_NUMBER	31	0	.	.	.	.	.	
DONOR_AGE	.	24.75222	0	87	58.91905	16.66938	-0.3779	-0.4565
DONOR_GENDER	4	0	.	.	.	.	.	
HOME_OWNER	2	0	.	.	.	.	.	
INCOME_GROUP	7	22.6719	.	.	.	.	.	
MEDIAN_HOME_VALUE	.	0	0	6000	1079.872	960.7534	2.456613	6.99424
MEDIAN_HOUSEHOLD	.	0	0	1500	341.9702	164.2078	1.723597	6.47759
OVERLAY_SOURCE	4	0	.	.	.	.	.	
PCT_MALE_MILITARY	.	0	0	97	1.029011	4.918297	11.74183	177.909
PCT_MALE_VETERANS	.	0	0	99	30.57392	11.42147	-0.19742	1.22315
PCT_OWNER_OCCUP	.	0	0	99	69.699	21.71102	-1.23559	1.17286
PER_CAPTITA_INCOM	.	0	0	174523	15857.33	8710.63	3.352875	23.1875
PUBLISHED_PHONE	2	0	.	.	.	.	.	
SES	5	0	.	.	.	.	.	
URBANICITY	6	0	.	.	.	.	.	
WEALTH_RATING	10	45.47801	.	.	.	.	.	
CLUSTER_CODE	31	0	.	.	.	.	.	

Figure 3. Variables Table Display provides control over what metadata including new descriptive statistics is displayed in columns.

You can use the imported data source with other SAS Enteprise Miner nodes to develop an analysis. For example, you can merge the demographics with the aggregated transactional customer data and partition the resulting table into training, validation, and test data sources.

DEFINE DATA SOURCES FROM THE EXPLORER WINDOW

The SAS Explorer window provides convenient access to allocated SAS tables along with SAS Enterprise project data (Figure 4). You can use the Explorer to quickly locate and view table listings or to develop a plot using one of the many interactive graph components. You can now define input data sources to be used in your analysis simply by dragging and dropping a table onto the diagram workspace The Input Data Source wizard automatically opens, guiding you through the process of defining metadata information about the data source, such as variable roles, measurement levels, and data source type. You can specify whether or not to add the data source as an entry to the Data Source folder of the project. The SAS Explorer has also been enhanced to view and edit (when appropriate) catalog entries of the types SOURCE, LOG, OUTPUT, and XML.

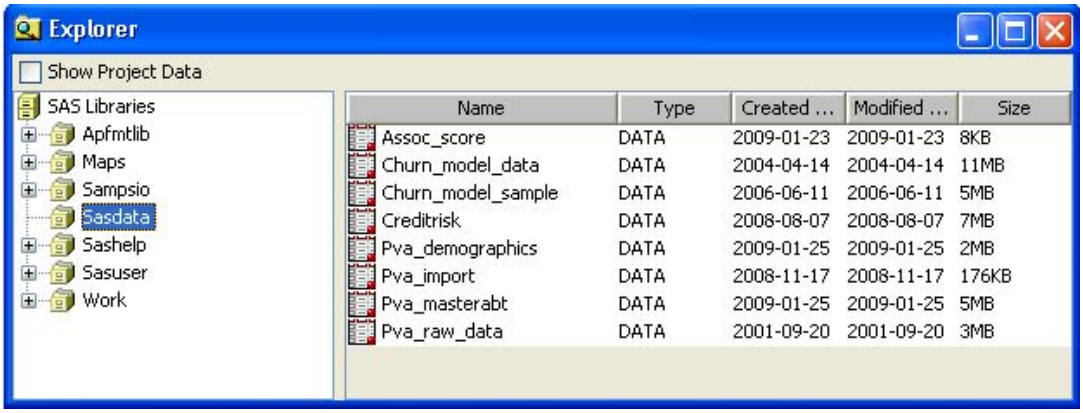


Figure 4. SAS Enterprise Miner 6.1 Explorer.

After you have defined the master analytical base table containing primarily aggregated customer donation transactions and other core overlay data, you can easily merge it with the demographic data using the Merge node (Figure 5).

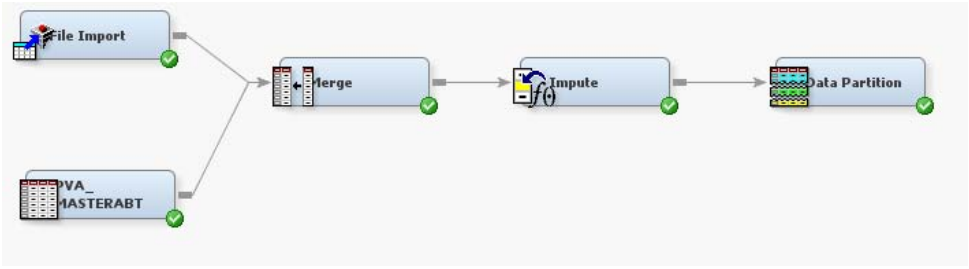


Figure 5. Integrate external data sources using the new File Import node and other SAS Enterprise Miner data preparation nodes.

The Append node has also been updated for Enterprise Miner 6.1 to combine training, validation, and test data sets into one training data set to calculate summary statistics on. This feature is very useful when the user wants to compare an imported model developed on a full training table versus an Enterprise Miner model developed on partitioned data.

EXTENDED MODELING FEATURES

LAR AND LASSO

Data miners often have training tables that contain several hundred and even thousands of potential predictors. Feature selection is an important task in data mining both to help ultimately develop parsimonious models that are not overly contaminated with collinear effects and also to generalize reasonably well when the model is applied to new data. The new LARS node of SAS Enterprise Miner 6.1 implements model fitting and variable selection for interval targets by using the SAS/STATGLMSELECT procedure. Methods include not only extensions to standard general linear modeling selection methods (forward, backward, and stepwise) but also the newer least absolute shrinkage and

selection operator (LASSO) and least angular regression (LAR) methods of Efron, Hastie, Johnston, and Tibshirani. (2004), respectively. The LAR method starts with no effects in the model and adds effects. At the first step the variable chosen to enter the model is the one most correlated with the target. The absolute value of this coefficient grows until a second effect becomes as correlated with the current residual as the effect in the model. This process continues until all variables enter the model or a selection stopping criterion is met. LASSO deletes parameters based on a version of ordinary least squares where the sum of the absolute regression coefficients is constrained. Additional information about the LAR and LASSO methods is provided in Cohen (2006).

The LARS node supports selection criteria, such as, Akaike's information criterion, Sawa's Bayesian information criterion, and Mallow C(p) statistics. You can also incorporate validation data or use *k*-fold cross validation for model evaluation. As do other SAS Enterprise Miner modeling nodes, the LARS node generates score. A variety of diagnostic plots are also provided to evaluate the selection process, as shown in Figure 6. One useful plot is the coefficient paths plot, which displays the change in the coefficients at the different steps as variables enter the model. The vertical line corresponds to the optimal model based on the user-defined model selection criterion, which in this case is the averaged square error for the validation data. The Iteration Plot shows the selected step for the optimal model along with the option for you to choose what selection statistic to display on the vertical axis.

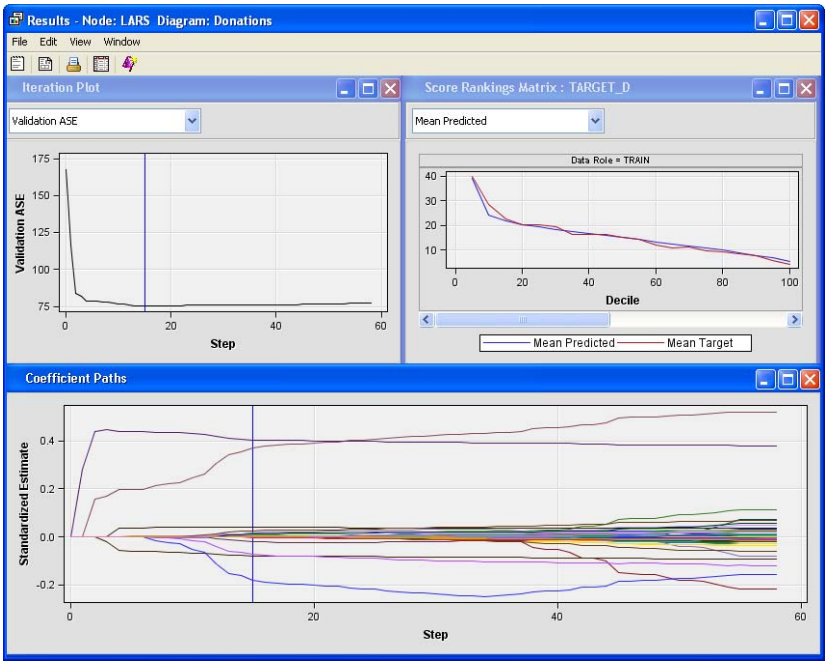


Figure 6. Example LARS Node Results

You can use the LARS node as a competitor variable reduction method to the Variable Clustering, Variable Selection, and Decision Tree nodes. Customers also asked for the ability to control how selected variable inputs are combined from two or more predecessor nodes into a successor node. The Merge node of SAS Enterprise Miner 6.1 has been updated to enable you to define rules that specify how to combine input variables from multiple predecessor nodes. The ANY rule sets the input to the rejected role if the variable is rejected in any predecessor node. The MAJORITY rule rejects an input if the input is rejected in the majority of the predecessor nodes. The ALL rule rejects an input variable only if it is rejected in all predecessor nodes. In the example process flow in Figure 7, the union of the input variables from the LARS and the Variable Clustering node are passed to the Regression node. This example passes only the selected inputs from the LARS node to another Regression node for integrated model comparison.

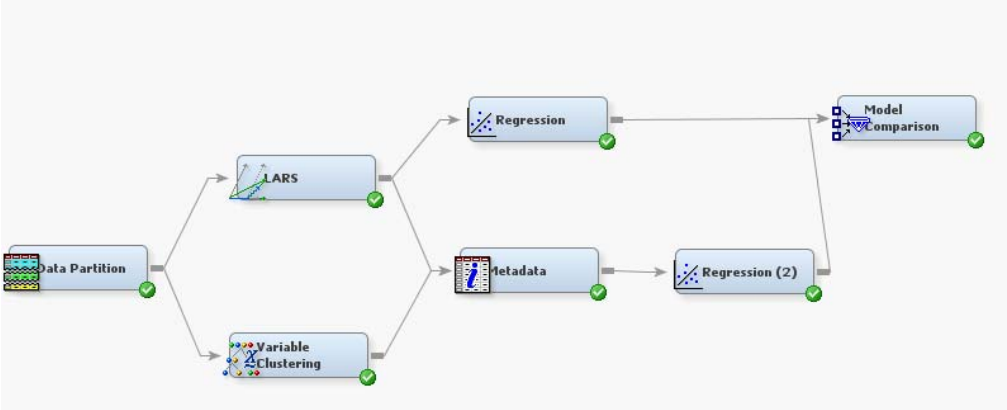


Figure 7. The Metadata node supports defining rules to enable users handle how they want to combine candidate variable inputs.

DEVELOPING SEGMENTATION STRATEGIES INTERACTIVELY BY USING SWITCH TARGETS

A switch targets feature has been added to SAS Enterprise Miner 6.1 so you can select a new dependent variable in a tree leaf and make new splits based on the new target. This is a powerful analytical feature for designing decision trees for segmentation strategies. The donations data contains two target variables: the likelihood that a customer will make a donation (TARGET_B with values of 0 or 1) and the dollar donation amount given that the customer responds (TARGET_D). Rather than develop a more classical two-stage predictive model, you may want to develop rule based segmentation strategies that identify dense pockets of high-dollar donors. Midstream during construction of the tree you can switch targets and spit on the second target as shown in Figure 8.

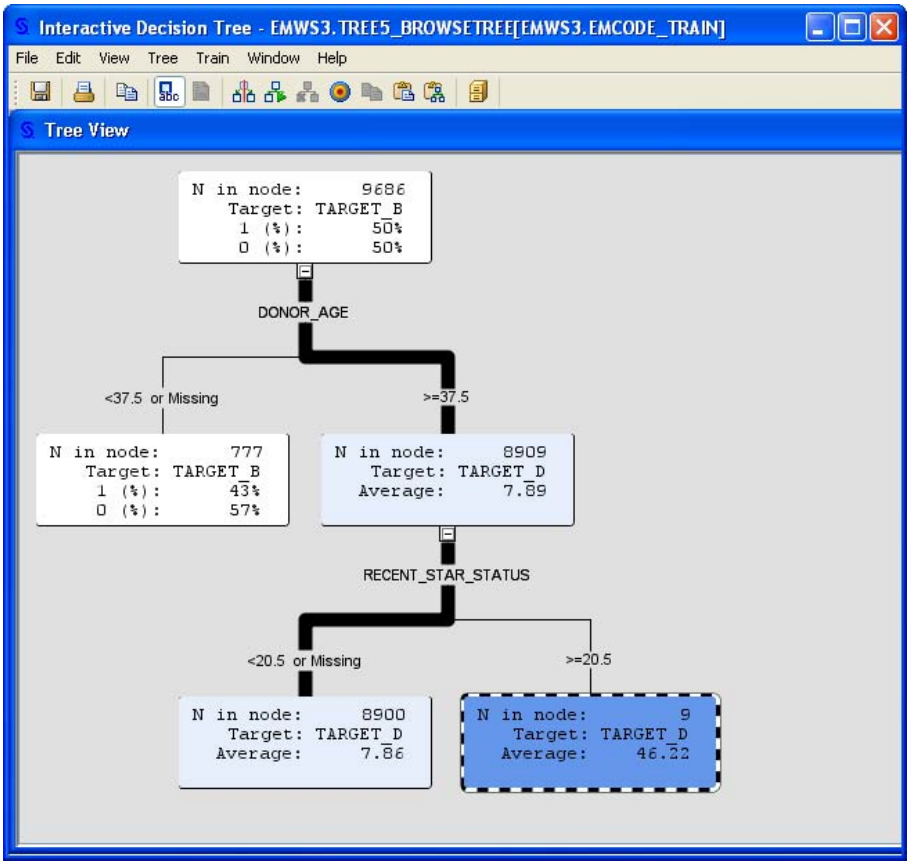


Figure 8. Splitting on multiple targets

Essentially the switch target feature provides a very convenient way to evaluate performance based on alternative targets. The segments can be used to define powerful strategies that are easy to understand by business managers. Score code is created for all targets but model assessment is done only for the primary target. The switch target feature complements well the copy-and-paste descendents feature added to SAS Enterprise Miner 5.3.

SUPPORT VECTOR MACHINE LEARNING

The new Support Vector Machine (SVM) node performs binary classification and regression estimation. Support Vector Machines map an n-dimensional input space into a high-dimensional feature space. In this high-dimensional feature space a linear classifier is constructed. In SAS Enterprise Miner 6.1, the SVM node is enhanced by adding Gaussian, sine, and polynomial kernel functions. These nonlinear kernel functions fit a wider variety of models.

A new data mining procedure, SVMSCORE, applies the nonlinear kernel functions to new data. The procedure processes the support vector rows of data that are required in order to score these functions. The SVM node also features improved optimization if you need to include source code.

REPORTER NODE

The Reporter node generates an analysis-ready report which along with SAS Enterprise Miner model packages provides a concise summary of the analysis for archiving and results sharing. The report is generated in PDF and RTF format and contains all information about the variables, functions, parameters, and the graphs displayed in the Node Results windows. In SAS Enterprise Miner 6.1, the Reporter node provides new SAS ODS (Output Delivery System) functions. The new functions create document graphs, process flow diagrams, and analytical plots that match the graphics that are displayed in the SAS Enterprise Miner user interface.

The SAS Enterprise Miner 6.1 Reporter node also provides new Decision Tree results plots for use in PDF and RTF documents (Figure 9). In Reporter node output, the properties list for each node tool indicates the property settings that have been changed from their default values. The Reporter results window now contains a standard external file viewer that you can use to view the document that was produced.

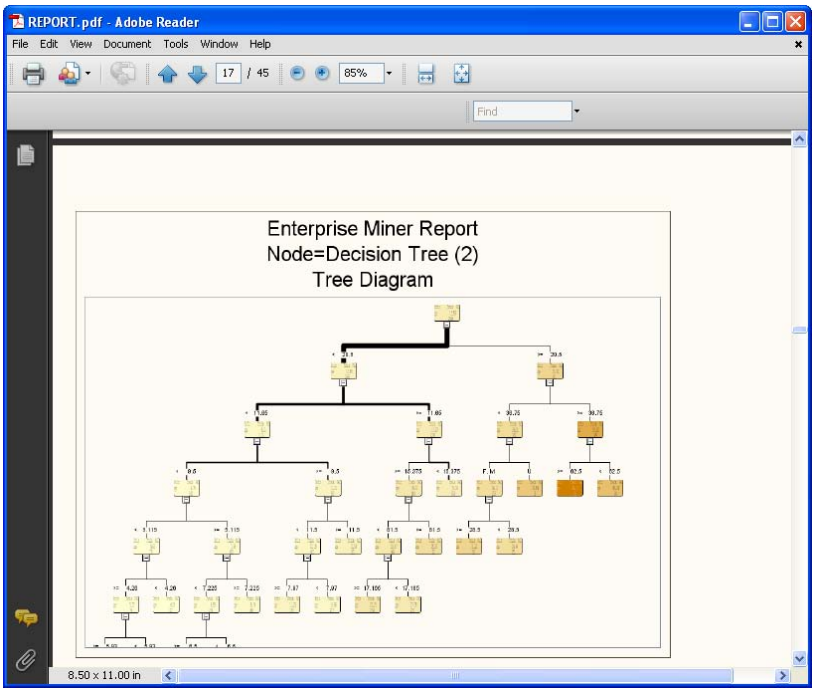


Figure 9. Decision Tree Diagram output from the Reporter node.

MODEL DEPLOYMENT

SCORE NODE

The Score node aggregates score code from the process flow diagram to create a single, deployable score code object. In SAS Enterprise Miner 6.1, the Score node scans and manipulates the SAS score code that the process flow diagram generates to eliminate intermediate code that produces terms that are not deployed in the final model function. The internally manipulated code is called optimized score code. The Score node now creates optimized score code by default. The Score node can also output the non-optimized score code for comparison.
For example, the Imputation node can add SAS code that creates many new variables, but a subsequent model selection step may keep only a few of the new terms. The optimized code eliminates unused terms that were created by the Imputation node.

The optimized code has a major positive impact on scoring and deployment processes. Fewer variables will need to be saved in the score input data sets in operational systems, which can save enterprises large amounts of resources and labor.

SCORING MODELS IN TERADATA

Many SAS customers store their operational in a Teradata Enterprise Data Warehouse (EDW). SAS Enterprise Miner generates SAS score code which coupled with SAS/Access Interface to Teradata can connect to a Teradata server to extract rows to SAS for scoring.

Customers requested the ability to score SAS Enteprise Miner models directly in the database to prevent extracting data and to leverage the shared nothing-architecture of Teradata. In the mid 2000s, SAS Enterprise Miner began supporting the Predictive Modeling Markup Language (PMML) for a core set of data mining functions. SAS Enterprise Miner PMML models can be deployed directly in Teradata using the Teradata Analytic Dataset Generator PMML scoring engine.

The SAS Scoring Accelerator for Teradata 1.4 is a new product for publishing SAS Enterprise Miner models into Teradata specific scoring functions for execution directly in Teradata. A primary advantage of scoring models using the SAS Scoring Accelerator for Teradata versus PMML is that a larger class of SAS Enterprise Miner data modification and modeling algorithms are supported. The SAS Enterprise Miner 6.1 Score Code Export node exports score files that are used as input to the publishing macro of the SAS Scoring Accelerator for Teradata (Figure 10). The Score Code Export node is an extension node delivered with the SAS Scoring Accelerator for Teradata media and can be added to your Enterprise Miner 5.3 or 6.1 installations.

Figure 10. SAS Enteprise Miner Process Flow which includes the Score Code Export node.

CONCLUSION

SAS is honored to present SAS Enterprise Miner 6.1 to its loyal and growing user base who are attending this year's SAS Global Forum conference. Many new customer enhancements have been added to this release which is expected to tremendously enhance user productivity and result in better predictive models. SAS is dedicated to delivering the most flexible and extensible data mining system to its user community. The authors look forward to seeing you at SAS Global Forum this year and encourage additional feedback and questions about the product. The future of SAS data mining is very promising. We look forward to working with you on the next chapter of SAS Enterprise Miner.

REFERENCES

- Cohen, R.A. 2006. "Introducing the GLMSELECT Procedure for Model Selection". Proceedings of the 31st SAS Users Group International Conference. Paper 207-31, Cary, NC: SAS Institute Inc.
- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression (with discussion)," Annals of Statistics, 32, 407–499.

- Mangasarian, O.L. and Musicant, D.R (2000), *Lagrangian Support Vector Machines*, Technical Report 00-06, Data Mining Institute, University of Wisconsin, Madison, Wisconsin. Also in *Journal of Machine Learning Research* 1, March 2001, 161-177.
- SAS Institute Inc. SAS® Scoring Accelerator 1.4 for Teradata: User's Guide. Cary, NC: SAS Institute Inc. 2008
- What's New in SAS® Enterprise Miner™ 6.1 (2009). See http://support.sas.com/documentation/onlinedoc/miner/index.html

ACKNOWLEDGMENTS

The authors express the upmost appreciation to the entire SAS data mining community for helping develop, test and deliver this new release on the SAS 9.2 platform. We also grateful to our customers, partners, and analysts for the excellent feedback that has helped shape many of these new features.

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