Paper SAS311-2014

What's New in SAS® Enterprise Miner™ 13.1

Jared Dean and Jonathan Wexler, SAS Institute Inc.

ABSTRACT

Over the last year, the SAS® Enterprise Miner™ development team has made numerous and wide-ranging enhancements and improvements. New utility nodes that save data, integrate better with open-source software, and register models make your routine tasks easier. The area of time series data mining has three new nodes. There are also new models for Bayesian network classifiers, generalized linear models (GLMs), support vector machines (SVMs), and more.

INTRODUCTION

Big data is one of the most popular buzzwords in the field of analytics today. Although big data has become mainstream over the last few years, definitions vary widely between industries and practitioners. SAS considers that an organization has big data whenever its ability to handle, store, and analyze data exceeds its current capacity. Even if your company does not have big data problems today, you will surely face them in the future.

Big data connotes problems that can't be solved by traditional software, techniques, or infrastructure. Modernizing your approach is paramount to successful problem-solving. Rather than being reactive by buying more hardware or hiring more staff, you can benefit from using analytical software that adapts to the ever-changing landscape of big data to become more efficient and collaborative.

SAS Enterprise Miner development's response to this big data explosion centers on the following:

- Machine Learning: SAS Enterprise Miner has provided algorithmic support for machine learning for years by implementing techniques such as neural networks, clustering, and ensemble learning through decision trees and gradient boosting. Data miners and data scientists are looking for supervised learning techniques (such as random forest, support vector machines, and Bayesian networks) or unsupervised learning techniques (such as *k*-means clustering).
- Scalability: You want analytics that scale with your business problem. SAS Enterprise Miner high-performance nodes enable you to take advantage of the threaded processing power of your existing SAS environment. Threaded processing splits the processing among your cores, allowing for faster processing and increased productivity. These high-performance nodes can also run in distributed mode, using the memory available in your distributed environments. You can run these nodes in high-performance mode on your Hadoop clusters, such as Cloudera or Hortonworks, or on dedicated hardware from Teradata, Pivotal, and Oracle.
- Productivity: During a typical analytical process, there are many handoffs, from data preparation, to transformations, to modeling, and to deployment. In addition, this process requires strong collaboration and significant automation. You might also have other technologies that are not from SAS (such as R) nested within your analytical process. SAS Enterprise Miner now supports R within a data mining flow. You can use Enterprise Miner to read in, modify, and transform your data, and then use R to build your model. You can then assess the viability of the model against existing SAS models. You are still required to write the R code, but you can use the true power of SAS and especially of SAS Enterprise Miner to manipulate your data and take advantage of the advanced modeling techniques, automation, and most importantly, the true enterprise-grade support that SAS provides.

This paper reviews SAS Enterprise Miner 13.1, which focuses on these three themes; it provides 10 new nodes, three new procedures, and algorithmic and technological enhancements. SAS Enterprise Miner 13.1 was released in December 2013 and requires the first maintenance release of Base SAS® software

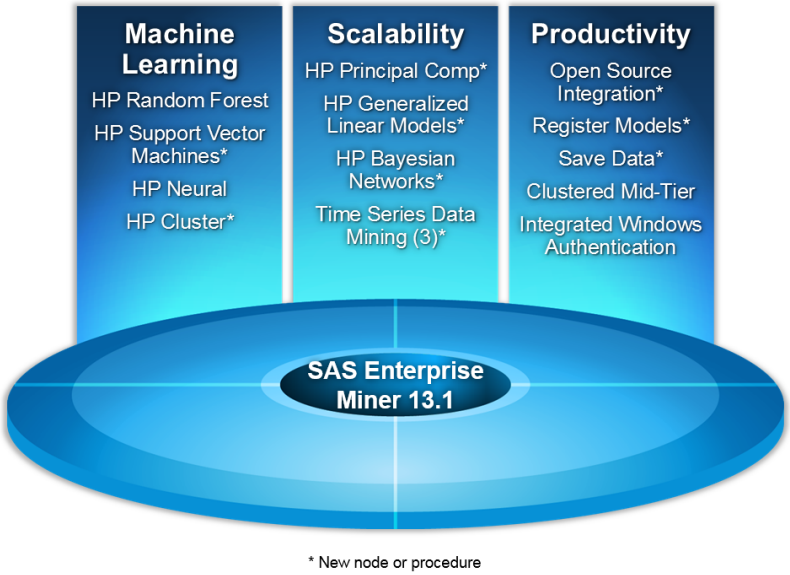

Figure 1. SAS Enterprise Miner 13.1 Themes and Features

NEW SAS ENTERPRISE MINER HIGH-PERFORMANCE PROCEDURES

PROC HPBNET

The HPBNET procedure is a high-performance procedure that learns a Bayesian network structure from an input data set. A Bayesian network is a directed acyclic graphical model in which nodes represent random variables and the links between nodes represent conditional dependency of the random variables. Because the Bayesian network provides conditional independence structure and a conditional probability table at each node, the model has been used successfully as a predictive model in supervised data mining. Supported structures include naive Bayes, tree augmented network (TAN), Bayesian network-augmented naïve (BAN), parent-child Bayesian network, and Markov blanket.

PROC HPSVM

The support vector machine (SVM) algorithm is popular in the data mining area of classification. The HPSVM procedure executes the SVM algorithm in multiple threads and, depending on licensing, uses multiple threads on multiple machines (distributed mode). PROC HPSVM provides two optimization techniques: the interior-point method and the active-set method.

PROC HPCLUS

The HPCLUS procedure performs clustering, a common step in data exploration, by using the *k*-means clustering technique. PROC HPCLUS uses least squares estimation to compute the cluster centroids. PROC HPCLUS computations have been written to take advantage of both parallel and distributed computing environments.

For some of the unsupervised clustering algorithms, such as *k*-means, one of the main questions is “How many clusters are in the data?” or said another way “What is the right number of clusters to segment my data?” Unsupervised clustering algorithms use several objective metrics, such as the cubic clustering criteria (CCC) and the gap statistic (Tibshirani, Walther, and Hastie 2001). CCC uses a simulated reference distribution to evaluate the right number of clusters. SAS has recently filed a patent that extends on the CCC method by generating many reference distributions, which are aligned with principal components. This invention is a new method, the aligned box criterion (ABC), for determining the true number of clusters in a set of data. ABC has the advantage of being more straightforward in diagnostic interpretation.

NEW HIGH-PERFORMANCE DATA MINING NODES

HP PRINCIPAL COMPONENTS NODE

The HP Principal Components node calculates eigenvalues and eigenvectors from the covariance matrix or the correlation matrix of input variables. Principal components are calculated from the eigenvectors and can be used as inputs for successor modeling nodes in the process flow diagram. Because interpreting principal components is often problematic or impossible, it is much safer to view them simply as a mathematical transformation of the set of original variables.

A principal components analysis is useful for data exploration and data dimension reduction. It is usually an intermediate step in the data mining process. Principal components are uncorrelated linear combinations of the original input variables; they depend on the covariance matrix or the correlation matrix of the original input variables. Principal components are usually treated as the new set of input variables for successor modeling nodes.

HP CLUSTER NODE

The HP Cluster node performs observation clustering, which can be used to segment databases. Clustering places objects into groups (clusters) that are suggested by the data. The objects in each cluster tend to be similar to each other in some sense, and objects in different clusters tend to be dissimilar. If obvious clusters or groupings can be developed prior to the analysis, then the clustering analysis can be performed by simply sorting the data.

HP GLM NODE

The HP GLM node uses the HPGENSELECT procedure to fit a generalized linear model in a threaded and distributed computing environment. A wide range of target distributions and link functions are available. For interval targets, standard distributions in the exponential family are available; for binary targets, negative binomial and zero-inflated negative binomial are available.

NEW DATA MINING NODES

OPEN SOURCE INTEGRATION NODE

The Open Source Integration node enables you to write code in the R language and integrate that code with SAS Enterprise Miner. The Open Source Integration node seamlessly transfers both data and metadata from SAS data sets to R data frames and then, depending on the R package, can return model information for further integration with SAS Enterprise Miner. In addition to enabling you to train and score supervised and unsupervised R models, the Open Source Integration node enables you to transform and explore your data.

SAVE DATA NODE

The Save Data node makes it simple to save any of your data sources (training, validation, test, score, or transaction) along with temporary variables that are created during the process flow. You can save the data to any file type that is available in a location that you define and is outside the project folder structures.

MODEL REGISTER NODE

The Model Register node enables you to register segmentation, classification, or prediction models as metadata in SAS® Metadata Server. Models whose metadata are registered can be input into SAS® Model Manager, and they can be used to score data in either SAS® Enterprise Guide® or SAS Enterprise Miner. Information about input variables, output variables, target variables, target levels, mining function, training data, and SAS score code is registered to the metadata.

Before the Register Model node was available, several steps in SAS Enterprise Miner were necessary to register a segmentation, classification, or prediction model into metadata. The Model Registration node consolidates those steps and provides a model registration mechanism that can run in SAS Enterprise Miner batch mode.

TIME SERIES DIMENSION REDUCTION NODE

The Times Series Dimension Reduction node extracts features from each time series and reduces the dimensions of time. It provides dimension reduction techniques that include discrete wavelet and Fourier transformations, singular value decomposition (SVD), and line segment methods.

TIME SERIES CORRELATION NODE

The Time Series Correlation node enables you to perform correlation and cross-correlation analyses. It calculates numerous autocorrelation and cross-correlation statistics on time series data. This functionality is most useful in clustering tasks that involve time series and time covariate selection among multiple time series.

TIME SERIES DECOMPOSITION NODE

The Time Series Decomposition enables seasonal decomposition of time series. These seasonal decomposition components can be used in subsequent nodes in your Enterprise Miner diagram.

USING NEW FUNCTIONALITY TO BUILD A PREDICTIVE MODEL

Figure 2 illustrates how you can use the new SAS Enterprise Miner functionality to build predictive models and analyze your data. This example uses a data set of claims data from the auto insurance industry. The data consist of 9.8 million records and 50 variables, and they take up about 2.2GB of disk space. The original target was an interval level to measure the severity of a claim (how large it would be). The target was modified to have a binary level: claim or no claim. Because most drivers file no claim during an insurance period, the target event of a claim being filed is rare, appearing in about 0.01% of records (about 1,000 claims).

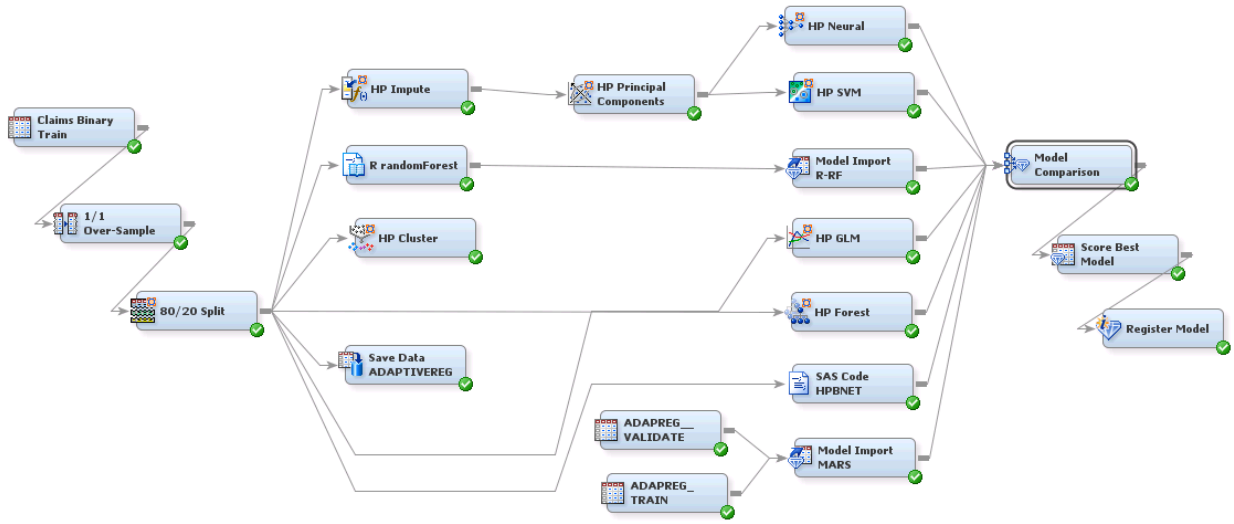

Figure 2. Using New Functionality to Analyze Claims Data

From left to right in the process flow diagram shown in Figure 2, the first node is an Input data Source node, which has been named “Claims Binary Train.” The next node is a Sample node (named “1/1 Over-Sample”), which oversamples the data to get an equal number of claim and non-claim observations. When you fit models for rare events, oversampling is a very effective technique for increasing the signal-to-noise ratio in the data. The next node is an HP Data Partition node (named “80/20 Split”), which partitions the data, assigning 80% of the data to the training partition and 20% of the data to the validation partition. The Input Data Source, Sample, and HP Data Partition nodes are unchanged for this release, but the ability to connect nodes from the HPDM toolbar to nodes that are located on other toolbars is new in SAS Enterprise Miner 13.1. This ability provides far greater flexibility in using all the tools that Enterprise Miner provides. For more information about restrictions on connecting HPDM nodes, see *SAS Enterprise Miner High-Performance Data Mining Node Reference*.

The new Save Data node is connected to the HP Data Partition node (“80/20 Split”). The Save Data node enables you to save any or all of the available data partitions and the new variables that have been created in the flow. This example saves the partitioned data so that they can be modeled by using the HPBNET procedure (which does not have a modeling node in SAS Enterprise Miner), thus ensuring comparability because the partitions in the models are identical. If you had imputed data, created a transformation, or performed any other data manipulation, those variables would also be included in the output from the Save Data node.

Also connected to the HP Data Partition node (“80/20 Split”) is the HP Impute node, which you can use to impute missing values. (This node is not changed for this release). You need this node to remove missing values and keep all the observations usable in the principal components analysis that is performed in the HP Principal Components node.

The HP Principal Components Node is new; it uses the HPRINCOMP procedure, which is new in SAS/STAT® 13.1. This node computes the principal components of the imputed claims data. It enables you to create either a fixed number of principal components or a maximum number of principal components that is based on cumulative proportional variance or incremental improvement. This analysis uses the default of 20 principal components. The decision to use 20 principal components was made after reviewing the eigenvalue plot shown in Figure 3. Because this plot is nearly flat at 20 principal components, most of the available information has been captured.

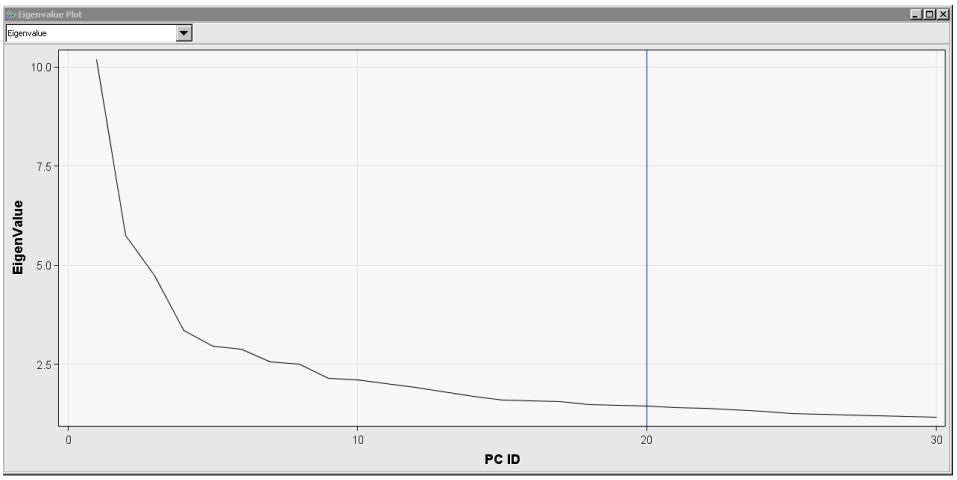

Figure 3. Eigenvalue Plot of Imputed Claims Data

Following the HP Principal Components node are the HP Neural node and the new HP SVM node. Although the HP Neural node is not new, it includes a number of enhancements. One enhancement is that you can define your own architecture in the Architecture property. This property enables you to specify up to 10 hidden layers for which you specify activation functions and the number of neurons. You can use this increased flexibility to achieve autoencoding and deep learning. For more information, see Hall (2014).

The HP SVM node provides a number of linear and nonlinear kernel options that use either interior point or active set optimization methods and support vector machines (SVMs) to model data. The HP SVM node replaces the SVM node, which was experimental for several releases. The HP SVM node is based on the new HPSVM procedure and provides excellent scalability and performance for large data. Like all SAS High-Performance Analytics procedures and high-performance data mining nodes, PROC HPSVM, and in turn the HPSVM node, use threaded algorithms and take advantage of modern multicore hardware to fit models more quickly than previously possible.

The next node connected to the HP Data Partition node is the new Open Source Integration node (named “R randomForest”). As the name implies, this node integrates open-source software with the SAS Enterprise Miner data mining platform. Currently, the only supported open-source language is R, but other languages will be supported in future releases. Integration with R happens in one of three ways, as designated by the value of the Output Mode property (PMML, Merge, or Name):

- PMML: Use of the predictive modeling markup language (PMML) enables you to seamlessly run an R script from SAS Enterprise Miner, and then use the R package to fit the model and convert it to a PMML file. The PMML file is then converted to DATA step code by using the PSCORE procedure. This process has restrictions that include the following:
  - The R package must conform such that the R PMML package can be used.
  - The model type that the R package creates must be supported by PROC PSCORE. Currently, PROC PSCORE supports the following PMML model types: linear models, multinomial log-linear models, generalized linear models, decision trees, neural networks, and *k*-means clustering.

  If these conditions are satisfied, then a model that is created in R can be converted to a DATA step and then used in SAS Enterprise Miner and a number of other SAS products, including SAS Model Manager and SAS® Scoring Accelerator.
- Merge: For R packages that do not meet all the criteria for PMML, you can still perform analyses, but you need to use the Model Import node to map the predicted variables to the modeling variables.
- None: This mode enables you to manage your entire analysis in an Open Source Integration node.

A main advantage of using the Open Source Integration node over working with R directly outside of SAS Enterprise Miner is that you can move your data between SAS and the open-source software. The partitioning, variable transformation, imputation, variable selection, binning, and so on that are part of nearly every data mining project do not need to be replicated in the open-source software (which is often very difficult, if not impossible).

The following R program fits a random forest model to the imputed claims data. Notice the macro variables that have been defined to simplify the references to variables and ensure a fair comparison of models. For complete details, see the documentation about the Open Source Integration node in *SAS Enterprise Miner: Reference Help*.

```
# Load library
library(randomForest)

# Train model
&EMR_MODEL <- randomForest(&EMR_CLASS_TARGET ~ &EMR_CLASS_INPUT + &EMR_NUM_INPUT,
ntree= 50, mtry= 5, data= &EMR_IMPORT_DATA, importance= TRUE)

# Results
&EMR_MODEL

# Plot available in Train Graphs
png("EMR_forestMsePlot.png")
plot(&EMR_MODEL, main= 'randomForest MSE Plot')
dev.off()

# Variable importance
round(importance(&EMR_MODEL),2)

# Export data to flow
&EMR_EXPORT_TRAIN <- predict(&EMR_MODEL, &EMR_IMPORT_DATA, type="prob")
&EMR_EXPORT_VALIDATE <- predict(&EMR_MODEL, &EMR_IMPORT_VALIDATE, type="prob")
```

In addition to using the Open Source Integration node to pass data objects between SAS and open-source software, you can also use it to display graphics that are created by R in the results window by following the guidelines that are described in the node documentation. Figure 4 is the graphical output from the PLOT statement in the preceding R program. It shows how increasing the number of trees in the forest affects the mean squared error (MSE) from the random forest model.

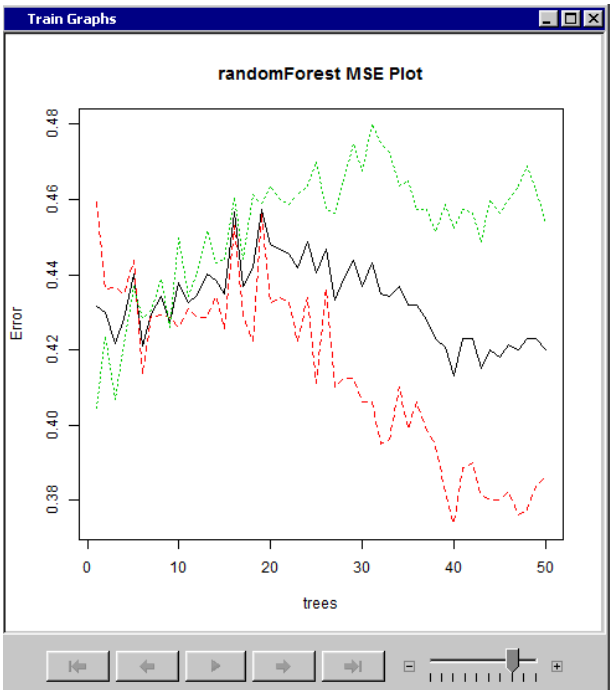

Figure 4. Plot of MSE from Using R Graphing Facility within Open-Source Integration Node

This example could not use the PMML facility to convert the random forest model from the R package into SAS DATA step code, so it uses the Merge output mode instead. The Merge output mode requires the use of the Model Import node, which has been included in SAS Enterprise Miner for many releases. The Model Import node enables you to map the prediction variables from a model that is built outside of SAS Enterprise Miner modeling nodes (the predict function in this case) to the proper event level for assessment and comparison in the Model Compare node. You can see this mapping component in Figure 7.

The next node connected to the HP Data Partition node ("80/20 Split") is the HP Cluster node. This new node uses the HPCLUS procedure to perform *k*-means clustering. You can use these clusters as segments for building many models by using the Start Group and End Group nodes, or you can use them as an additional nominal variable. Figure 5 shows the number and relative size of the clusters that are determined by the HP Cluster node.

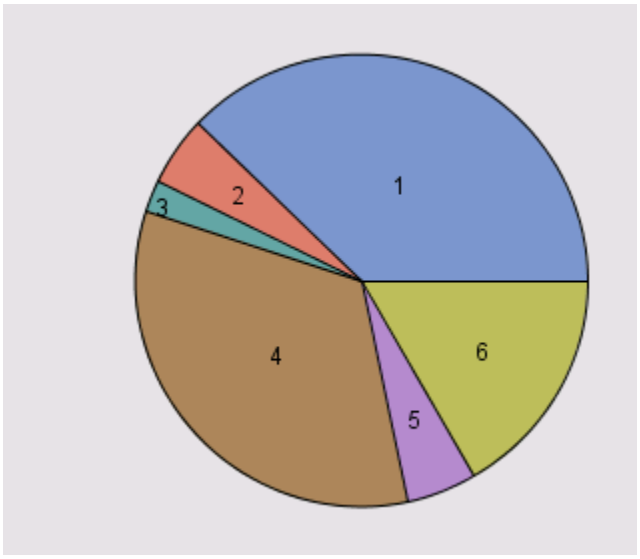

Figure 5. Size of Clusters Selected by the HP Cluster Node

The next node connected to the HP Data Partition node ("80/20 Split") is the HP GLM node. This node is based on the HPGENSELECT procedure; it fits generalized linear models by using a variety of probability distributions and link

functions for the target variable. One of the supported probability distributions is the Tweedie distribution, which is essential for modeling interval targets that have a large number of observations whose value is 0, which is usually true in insurance data where most policyholders do not make a claim in a specific time period. The HP GLM node replaces the Ratemaking node as the means of fitting GLM-like models. The HP GLM node has several advantages over the Ratemaking node, including the ability to treat interval variables as continuous. Figure 6 shows that several terms in the model have very large effects.

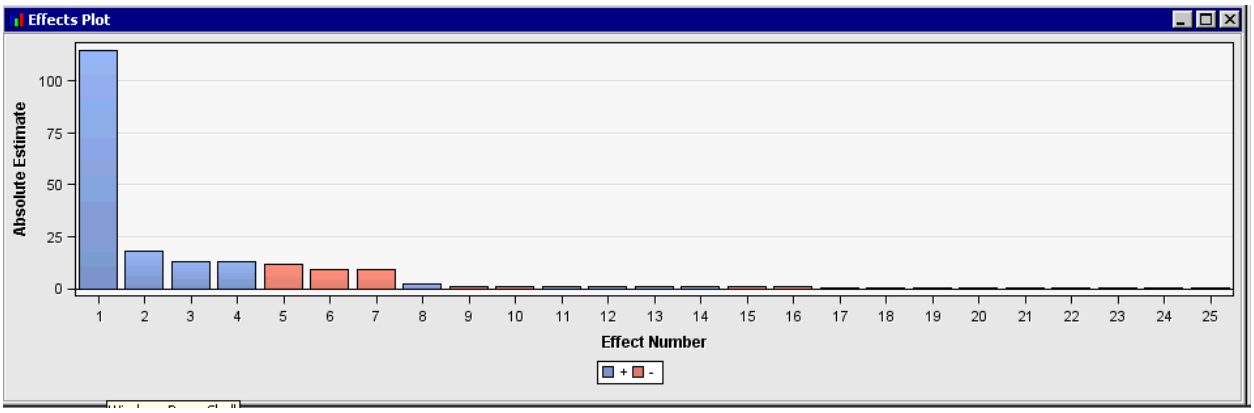

Figure 6. Effects Plot from the HP GLM Node

The next node connected to the HP Data Partition node is the Save Data node. This node enables you to export the various partitions of your data to an external library or file for analysis or presentation outside SAS Enterprise Miner. This example uses the ADAPTIVEREG procedure (new in SAS/STAT 13.1) to fit a model and compare its performance to other methods. PROC ADAPTIVEREG fits multivariate adaptive regression splines (Friedman 1991). SAS Enterprise Miner does not provide a node specifically for PROC ADAPTIVEREG, but it uses the SCORE statement in PROC ADAPTIVEREG to assess the predictive power of the model relative to the other methods in the diagram. The following statements use the two partitions that were exported by the Save Data node to fit the model and to score the training and validation data sets:

```
*Library where data was saved;
libname j "\\d78970\public";

*Macro variables;
%let Interval_var=Model_Year NVVar1 NVVar2 NVVar3 NVVar4 Var1 Var2
Var3 Var4 Var5 Var6 Var7 Var8 Vehicle blind_model_risk blind_submodel_risk;
%let Nominal_var= Calendar_Year Cat1 Cat10 Cat11 Cat2 Cat3 Cat4 Cat5 Cat6
Cat7 Cat8 IMP_Cat12 IMP_OrdCat NVCat Cat9 claim_in_2005;

*Procedure call to fit model for binary target and score data sets;
proc adaptivereg data=j.EM_SAVE_TRAIN valdata=j.EM_SAVE_VALIDATE;
   class claim_flag &nominal_var;
   model claim_flag(descending) = &interval_var &nominal_var /dist=binomial;
   score data=j.EM_SAVE_TRAIN out=j.Adapreg_Train PREDICTED(ilink);
   score data=j.EM_SAVE_VALIDATE out=j.Adapreg_Validate;
run;
```

After the model has fit both the training and validation data sets (Adapreg_Train and Adapreg_Validate, respectively), those data sets need to be created as data sources in the SAS Enterprise Miner project and assigned the correct data source role. Both of these new data sources are then added to the diagram and connected to a Model Import node.

In the scoring process, a new variable, PRED, is created in both the training and validation data sets. That variable does not match the SAS Enterprise Miner naming convention for prediction variables, so it must be mapped to the correct level for correct assessment. Figure 7 shows the mapping component. Because PRED represents the probability of the event occurring, PRED is assigned to P_CLAIM_FLAG1.

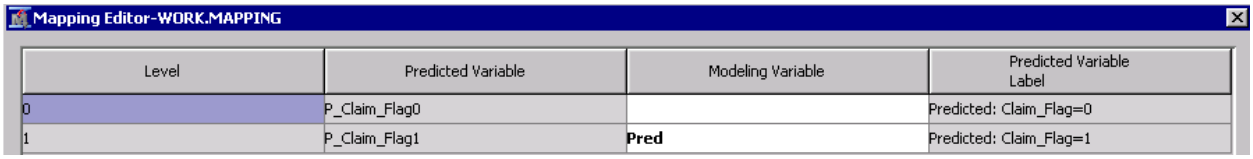

Figure 7. Model Import Node Dialog Box for Mapping Predicted Values

The next node is the HP Forest Node. This node is not new, but it benefits from improved performance of the HPFOREST procedure for more scalable model fitting and additional variable selection methods.

The final node that is connected to the HP Data Partition node is a SAS Code node. Although SAS Enterprise Miner has not yet implemented a modeling node that uses the HPBNET procedure, you can enter the following PROC HPBNET statements in the SAS Code node to take advantage of the powerful modeling technique that PROC HPBNET provides:

```
*PROC HPBNET code;
proc HPBNET data=&EM_IMPORT_DATA;
    target %EM_TARGET;
    input %EM_INTERVAL_INPUT / level=interval;
    input %EM_BINARY_INPUT %EM_NOMINAL_INPUT /level=nominal;
    code file="&EM_FILE_EMPUBLISHSCORECODE";
run;
```

The CODE statement, the SAS Enterprise Miner macro variables, and setting the Tool Type property to the value Model enables you to take full advantage of PROC HPBNET.

Now that all seven paths are described, you can use the Model Comparison node to assess which model is best. The assessment of many models to determine the best one for a specific application is very common and is a best practice in model building. The Ensemble node enables you to incorporate another best practice: to combine the different models to create a better prediction. (This node is not new in SAS Enterprise Miner 13.1.) For more information about this node, see Maldonado (2014).

The flexibility of SAS Enterprise Miner is shown in this example's evaluation and comparison of four models from new or enhanced SAS Enterprise Miner modeling nodes: an R Random Forest model that is built in an Open Source Integration node, the ADAPTIVEREG procedure in the Save Data and Model Import nodes, and a naïve Bayes model that is fit by the HPBNET procedure in a SAS Code node. Figure 8 shows the comparison of the cumulative lift from all of these models. The HPSVM model has the best cumulative lift at the second decile, in part because of the balanced nature of the oversample.

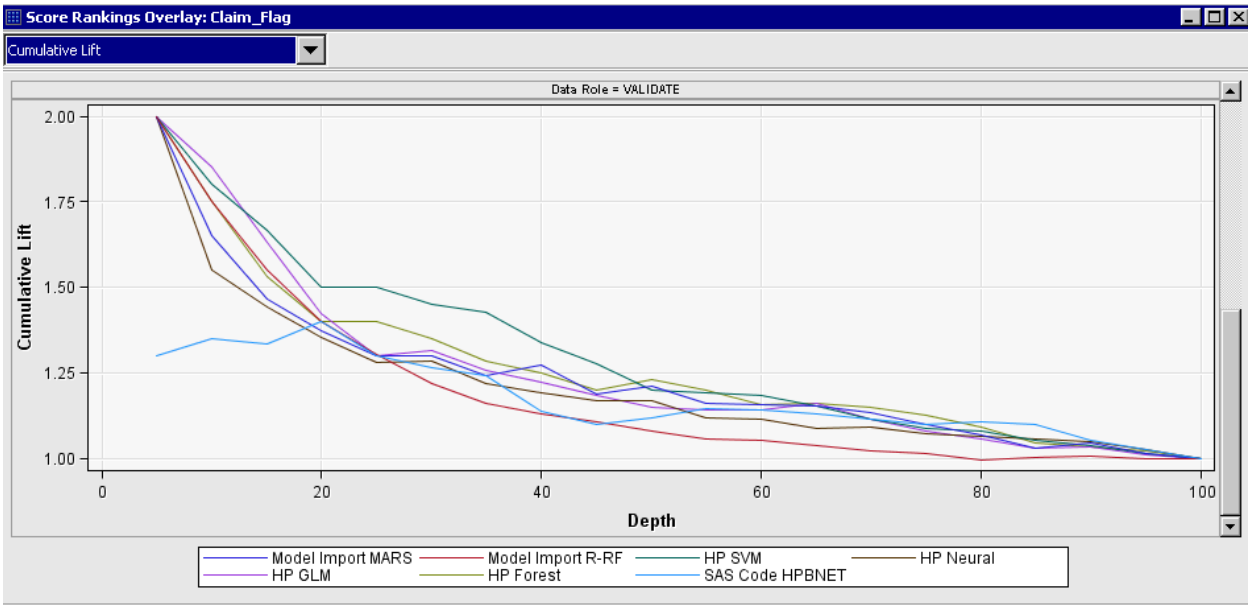

Figure 8. Comparison of Validation Partition of Seven Models

After you run the model comparison and determine the best model, you face the challenge of deploying this model into your production systems. You must first run a Score node on the selected model to generate the necessary inputs and outputs.

The final node in this diagram is the Model Registration node. This node streamlines the functionality of registering a model that is built using SAS Enterprise Miner to metadata for deployment to the SAS Scoring Accelerator or for management within SAS Model Manager. Before the Model Registration node was implemented, you had to create a model package and register that model package to metadata through a number of steps. In addition, the registration of a model was always a manual step that had to be performed after a diagram had completed. With the new Model Registration node, you can automate the process for registering a model as part of the flow. So you can register models even if diagrams are run as a batch operation. Because the best model from this example's Model Comparison node was the HP SVM node (as shown in Figure 8), the HP SVM node model is registered to the SAS Metadata Server. If the best model changes as a result of different tuning parameters of the models or new data, then the new best model is automatically registered to metadata. You can use the Model Import node to compare models that are created outside SAS Enterprise Miner, but you can register metadata only for models that contain SAS DATA step score code.

TIME SERIES DATA MINING EXAMPLE THAT USES NEW FUNCTIONALITY

The Nike+ FuelBand is a device that you wear on your wrist and that contains a number of sensors to track your activity. Seeing the status of your activity motivates you to be more active, which helps you achieve your fitness goals more easily. The Nike+ FuelBand collects data on steps, calories, and NikeFuel (a measure of exertion) as a time series in which a minute is the lowest interval.

This example shows how you can use the new features of time series data mining nodes to analyze the data that the Nike+ FuelBand collected on an individual subject. For more information, see Dean (2014). Figure 9 shows the process flow diagram.

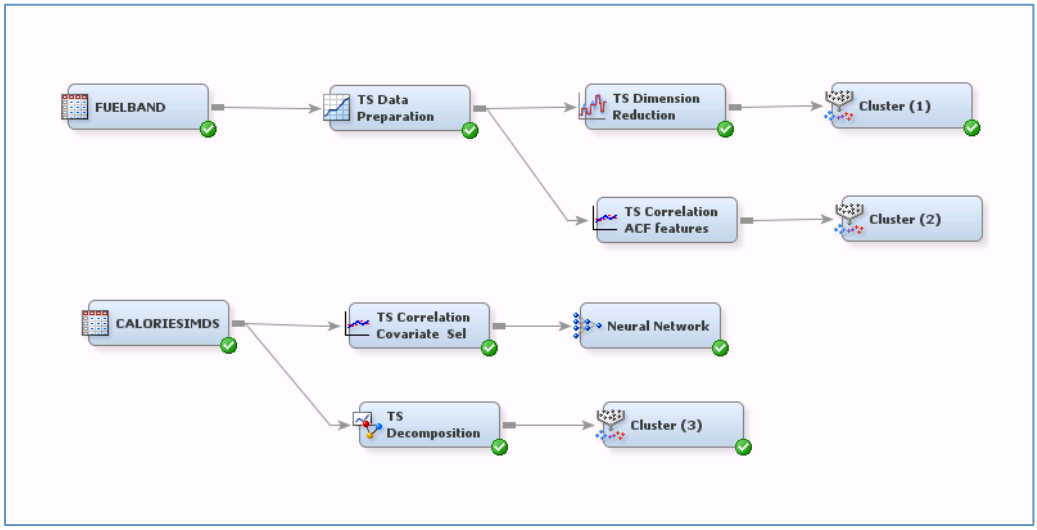

Figure 9. Process Flow Diagram for New Times Series Nodes

In the Input Data Source node (titled FUELBAND), the variable MONTH is set to the role of CROSSID, the variable TS_CALORIES is set to the role of INPUT, and DAY is assigned the role of TIMEID. In the TS Data Preparation node, setting the aggregation option to Average summarizes the input data into a daily averaged calories series for each month. The missing option is also set to Average because the number of days differs in each month. These settings modify the times series to make them all the same length so that the final data have 12 time series, from January to December, and each time series has 31 time points, one for each day of the month.

The TS Dimension Reduction node extracts features from each time series. The default setting, discrete wavelet transformation (DWT), is used and the number of dimensions is set to six. This dimension reduction reduces each series to about 20% of the original dimension (6/31). For the DWT, the node uses a simple Haar transform, which consists of wavelet scale and detail coefficients. In this example, four scale coefficients (dimensions 1, 2, 4, and 5) and two detail coefficients (dimensions 3 and 6) are stored and exported for further analysis, such as time series clustering. The six extracted features for each series are plotted, and the clustering result uses the extracted features instead of the original time series.

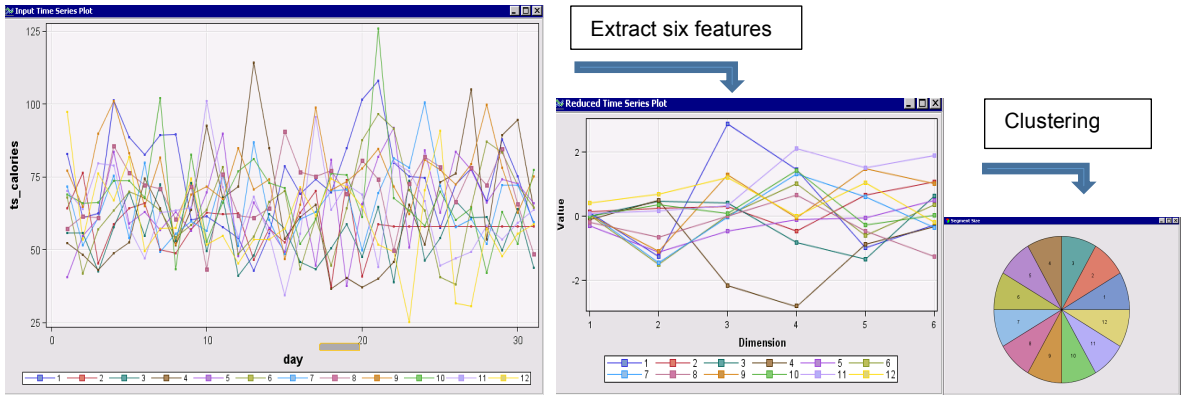

Figure 10. Process of Time Series Dimension Reduction to Clustering

The Cluster node discovers an interesting result: the time series are sufficiently distinct such that each is assigned to its own cluster. Over the 15 months during which the data were gathered, each month was quite different from other months. Some months had regular exercise, and some months had no exercise. Activity reduced dramatically as the subject began writing a book on data mining and related topics, but it increased again when the subject suffered writer's block.

Similarly, you can use TS Correlation node to extract autocorrelation coefficients for clustering input variables because the autocorrelation function (ACF) is usually used to determine the autoregressive time series model. In addition to extracting autocorrelation, the TS Correlation node provides a time covariate selection function that uses cross-correlation analysis between target series and input series. The second flow from Figure 9 uses the CALORIESIMDS data to perform time covariate selection and time series seasonal decomposition. The

CALORIESIMDS data are collected from the same FuelBand, but the first two months of data are used to make a target series that is an hourly sequence averaged over two months. In other words, the target series is the average hourly pattern in a day from the first two months of data. The remaining data are divided into daily series that have 24 points, one for each hour. The correlation criterion of 0.85 in the TS Correlation node selects seven input time series that meet the criteria. Any modeling node, depending on the application, can be connected for further analysis. This example uses the Neural node. This combination of TS Correlation followed by a modeling node provides a useful alternative variable selection method when you have time-varying covariates in your modeling input data.

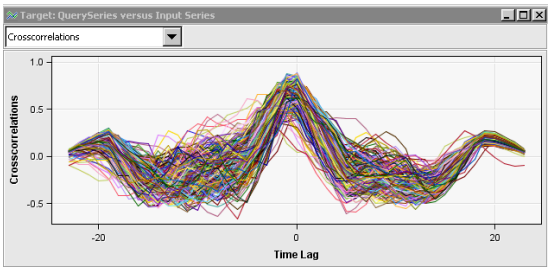

Figure 11. Time Series Cross-Correlation between Reference Series and Input Series

The TS Decomposition node decomposes the original series into seasonal, trend, cyclic, and irregular components. It also provides some combination of components such as trend-cycle. You can understand the time series behavior and underlying structure better if you look at the decomposed series instead of the original series. For example, a FuelBand user might be interested only in a positive trend in the series. The decomposed components can also be used for time series clustering. In Figure 11, the trend-cycle components are exported as Cluster node inputs because this example did not want to consider seasonal and irregular components in its time series clustering.

SAS ENTERPRISE MINER INFRASTRUCTURE

INTEGRATED WINDOWS AUTHENTICATION

SAS has supported Integrated Windows Authentication (IWA) since version 9.1.3., and now SAS Enterprise Miner 13.1 supports IWA. You see this feature in the logon dialog box in a client/server topology. For more information about setting up a network domain that is compatible for IWA, see *SAS® Enterprise Miner™ 13.1: Administration and Configuration* and *SAS® 9.4 Intelligence Platform: Security Administration Guide*. To enable IWA, select its check box, as shown in Figure 12. When IWA is selected, you are no longer prompted for a user name and password; instead you will use the credentials that were supplied to the Windows operating system.

Figure 12. SAS Enterprise Miner Log On Screen, Highlighting IWA

MIDDLE-TIER CLUSTERING

Another new infrastructure feature in SAS Enterprise Miner 13.1 is clustering the middle tier. Figure 13 shows an example topology of a clustered middle tier. The advantages of clustering the middle tier are load balancing of users and the automatic failover that is enabled by having multiple middle-tier machines. The middle-tier clustering feature is often used heavily in large enterprise deployments where service level agreements (SLAs) exist or in multinational organizations where users are actively using the system at all times of day. For more information about how to configure a clustered middle tier, see *SAS Enterprise Miner 13.1: Administration and Configuration.*

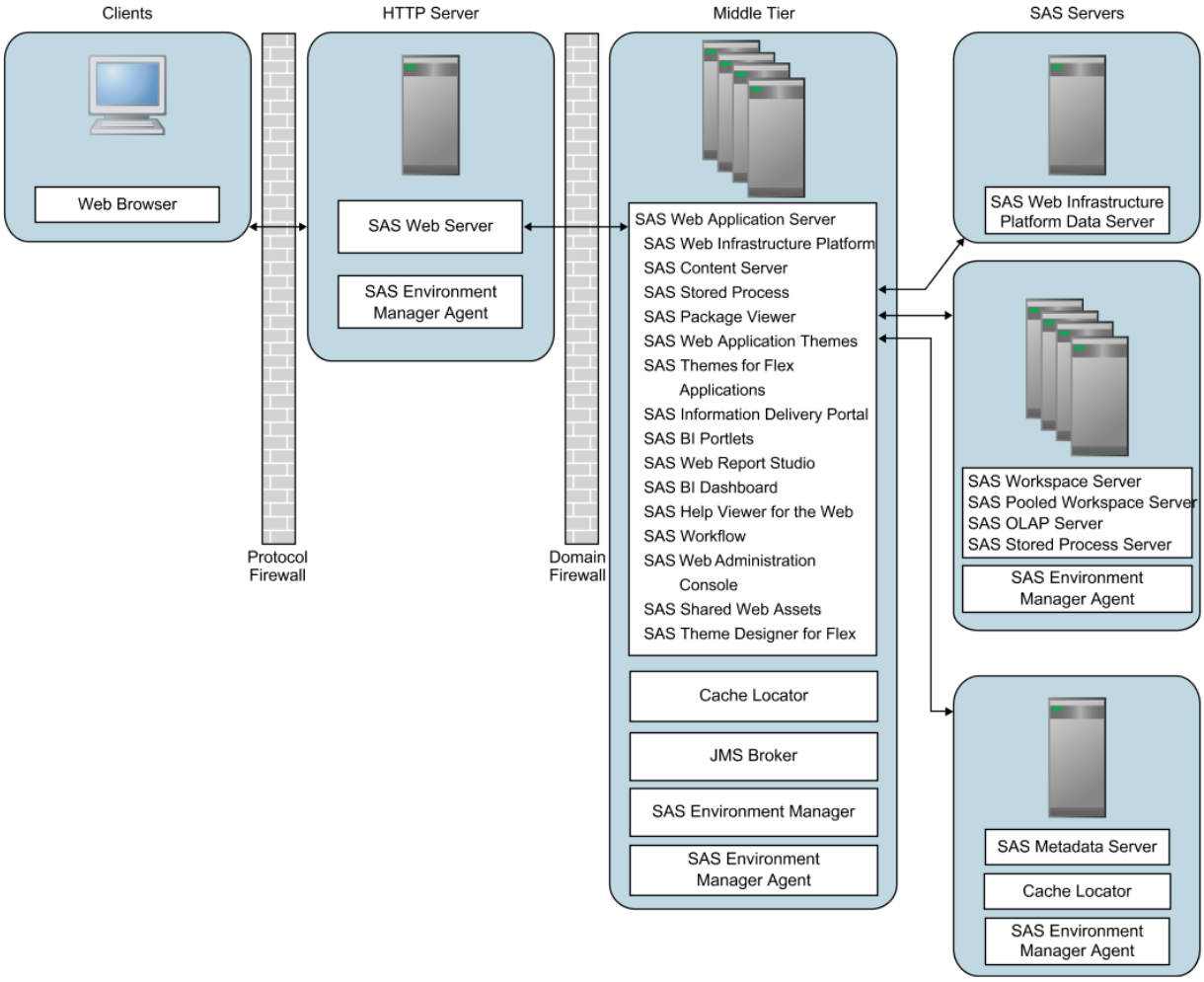

Figure 13. Topology of Clustered Middle-Tier Deployment

Figure 13 illustrates a relatively complex topology of SAS Enterprise Miner. The user communicates through the Enterprise Miner client that was either installed or downloaded through the Java Web Start delivery mechanism. Then that client communicates with a SAS Web Server. The SAS Web Server communicates with one or many middle-tier machines, which provide backup in failover and balance the workload of multiple users. Then the middle-tier machines pass along requests to the SAS Workspace Server or the SAS Metadata Server to respond to users' requests to run diagrams, open projects, or interact with SAS Enterprise Miner in other ways. Like any enterprise system, the clustered middle tier is largely ignored if things are running well. SAS Enterprise Miner users never need to wonder which middle-tier server their diagram is running on, because the particular sever is irrelevant to their use of Enterprise Miner.

LOCKDOWN

SAS Enterprise Miner also supports the platform option of Lockdown, which was released in SAS 9.4 in July 2013. Lockdown restricts a user's ability to access files by filtering requests through a "white-list." This feature is particularly useful when access needs to be controlled between groups or organizational entities either for business or regulatory requirements.

For more information about the lockdown feature, see *SAS Enterprise Miner: Administration and Configuration*. A recommended best practice is that if you use the lockdown feature, you also restrict the location where SAS Enterprise Miner projects can be created. You can specify this location in SAS® Management Console, as shown in Figure 14. That project creation location must also be included in the "white-list" for the lockdown option.

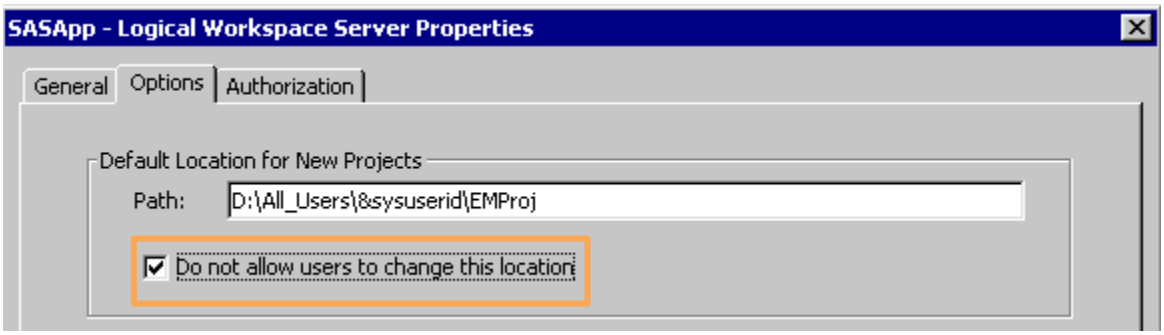

Figure 14. SAS Enterprise Miner Plug-in of SAS Management Console

FUTURE DIRECTION OF SAS ENTERPRISE MINER

RAPID DEPLOYMENT

A main goal of SAS Enterprise Miner development is to simplify the installation process in order to enable rapid distribution and installation. A step in that direction will be the release of a fully configured distribution using SAS vApp technology. The SAS vApp technology (vApp) is a group of products that can be deployed quickly in a preconfigured virtual environment and updated easily. The vApp for SAS Enterprise Miner enables you to deploy a standard three-tier single-machine topology (the most common deployment choice for departmental servers) and go from a clean machine to a running client/server instance of SAS Enterprise Miner in just a few minutes after the download completes. This will be a tremendous timesaver for IT staff in installation, migration, and upgrades of SAS Enterprise Miner.

BROWSER-BASED DATA MINING AND AUTOMATION

Along with development efforts that are described in this paper, recent work on creating a new browser-based client for creating predictive models has been progressing. This client was released to selected customers along with SAS Enterprise Miner 13.1, and their feedback is being incorporated as functionality is being added. A new web-based client that supports two main use cases is focused on productivity and simplification. One use case is rapid prototyping of predictive models. Many SAS customers find that they often need to develop many more models than time and staffing allow. The web-based client increases productivity through a new, simplified design so that you can evaluate more ideas in a shorter period of time and focus on the ideas that have the greatest interest for your organization. Increased productivity leads to competitive advantage for your organization. The other primary use case is to support a true, innovative model factory. Imagine having a system that enables you to build models for all the stock keeping units (SKUs) in your store. Then, when you need to track and retrain models, you will need only one environment to do so. Not only can you build more models, but you can also collaborate with your colleagues and share what you discover. You will hear much more about these use cases throughout 2014.

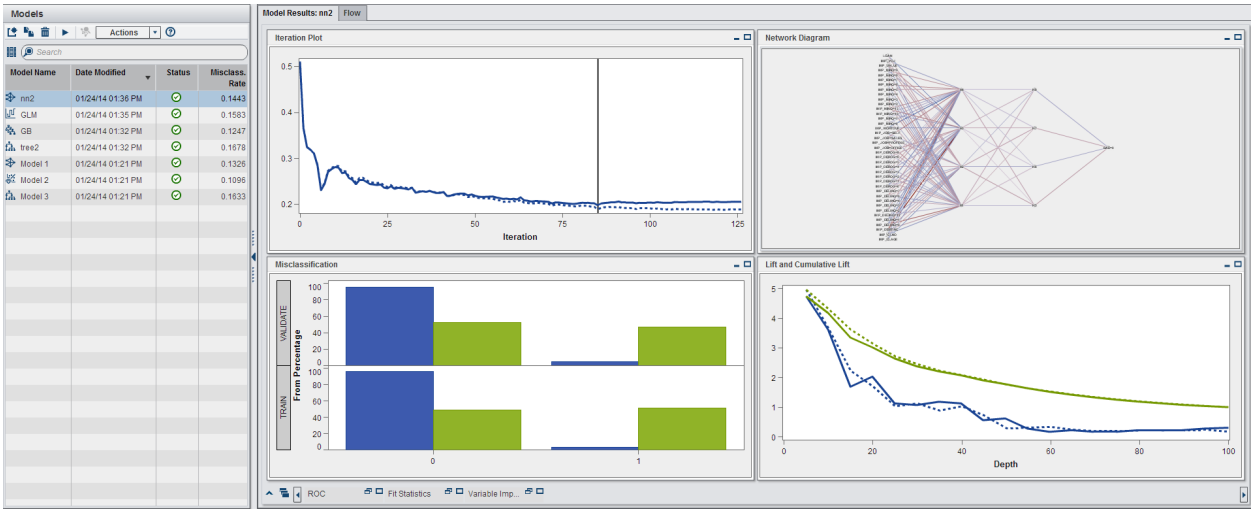

Figure 15. SAS Data Mining Web Client, Early Adopter Version

INNOVATION IN SAS ENTERPRISE MINER

SAS continues to release new high-performance machine learning algorithms that can execute in single-machine and distributed modes. You will see new algorithmic support for rules and associations, ensemble methods, and unsupervised learning. SAS continues to embrace open source as an extension of the SAS ecosystem. There will be close linkages between open-source packages and SAS Enterprise Miner, creating a complete analytic flow.

CONCLUSION

SAS Enterprise Miner 13.1 was a significant release. Support for big data analytics will continue and accelerate through the next several releases. Building models requires many steps, and each new release will make your process more streamlined, scalable, and, last but not least, collaborative. SAS wants you to build more models, use all your data, and intelligently deploy your SAS analytics.

REFERENCES

Dean, J. 2014. *Big Data, Data Mining, and Machine Learning: Creating Value for Business Leaders and Practitioners*. New York: John Wiley & Sons.

Friedman, J. H. 1991. “Multivariate Adaptive Regression Splines.” *Annals of Statistics* 19:1–141.

Hall, P., Dean, J., Kaynar, K. I., and Silva, J. “Overview of Machine Learning with SAS Enterprise Miner.” *Proceedings of the SAS Global Forum 2014 Conference.* Cary, NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings14/SAS313-2014.pdf.

Maldonado, M., Dean, J., Czika, W., and Haller, S. “Leveraging Ensemble Models in SAS Enterprise Miner.” *Proceedings of the SAS Global Forum 2014 Conference*. Cary, NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings14/SAS133-2014.pdf.

Tibshirani, R., Walther, G., and Hastie, T. 2001. “Estimating the Number of Clusters in a Data Set via the Gap Statistic.” *Journal of the Royal Statistical Society*, *Series B* 63:411–423.

ACKNOWLEDGMENTS

The authors would like to thank Susan Haller, Patrick Hall, Taiyeong Lee, Min Lu, and Dominique Latour for their contributions to the paper. They are also grateful to Anne Baxter and Ed Huddleston for their editorial contributions.

CONTACT INFORMATION

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

Jared Dean
SAS Institute Inc
Cary, NC 27519
Jared.Dean@SAS.com

Jonathan Wexler
SAS Institute Inc
Cary, NC 27519
Jonathan.Wexler@SAS.com