Paper 113-2010

Rapid Predictive Modeling for Customer Intelligence

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

ABSTRACT

Business analysts often need to develop campaigns based on predictive models to select which customers to target. The models often need to be developed in a short amount of time without always relying on a statistician. For these cases, SAS has developed the SAS® Rapid Predictive Modeler for use by customers of SAS® Enterprise Miner™. SAS Rapid Predictive Modeler runs as a customized task in either SAS® Enterprise Guide® or the SAS® Add-In for Microsoft Office. It automatically treats the data to handle outliers, missing values, rare target events, skewed data, collinearity, variable selection, and model selection. The results are presented in business terms as a simple-to-understand scorecard. The final model can be registered to the SAS® Model Manager and deployed to either a SAS® server or through a SAS® Scoring Accelerator to a relational database. The presentation covers both the implementation and results from an example analysis.

INTRODUCTION

The SAS Rapid Predictive Modeler has been created to ease the process of creating efficient, accurate, and robust data mining models. It requires minimal user input and produces reports that are suitable for business presentations. SAS Rapid Predictive Modeler models have been designed for common classification and prediction scenarios such as customer acquisition, up-selling, cross-selling, retention, churn, return on investment, and many others. In fact, SAS Rapid Predictive Modeler is suitable for any scenario in which the data are well-defined and a data mining model needs to be created quickly and accurately.

SAS Rapid Predictive Modeler models are easy to produce. You must supply a data set in which every row contains a set of independent predictor variables (known as inputs) and at least one dependent target variable. There are no other required selections. The SAS Rapid Predictive Modeler decides whether variables are continuous or categorical and which variables should be included in the model, and it uses a broad class of classical and modern modeling techniques to ensure the model is accurate and robust.

Three modeling methodologies are used: basic, intermediate, and advanced. The basic model develops a good baseline model quickly. The intermediate model starts with the basic model, adds a more complex set of transformations, and tries additional modeling algorithms. The intermediate model often results in a more accurate model, but it takes longer to run. The advanced model extends the intermediate model to include neural networks and ensemble models. All models generate a scorecard format report that can be used to make decisions in terms of which customers should be targeted.

SAS Rapid Predictive Modeler reports are easy to consume. You receive a listing of the significant terms in the model and common business graphics such as lift charts. You can choose additional details such as statistical goodness-of-fit measures and crosstabulations. The output is presented directly to you and also saved to a PDF or RTF file for inclusion in custom reports.

SAS Rapid Predictive Modeler models are easy to deploy into scoring systems. The models are saved as Base SAS code that can be executed on any Base SAS installation. You can register the models to the SAS® Metadata Server for direct use in other products such as SAS Enterprise Guide, SAS® Data Integration Studio, and SAS Model Manager. These products can automate the execution of score code and the deployment to other systems. SAS Rapid Predictive Modeler score code is also fully compatible with the SAS Scoring Accelerators for Teradata, Aster, Netezza, and DB2.

SAS Rapid Predictive Modeler models are compatible with SAS Enterprise Miner. In fact, the models are built using the functions of SAS Enterprise Miner. Models can be opened as projects and diagrams inside SAS Enterprise Miner for investigations into diagnostics, behaviors, and alternatives. SAS Rapid Predictive Modeler and Enterprise Miner users can collaborate on projects to produce the best model for each target. SAS Rapid Predictive Modeler models can also be executed later using the batch processing facility of SAS Enterprise Miner.

SAS Rapid Predictive Modeler builds the following models for standard data mining classification and regression problems:

- Classification models predict the value of a discrete variable such as True or False; High, Medium, or Low; Purchase or Decline; and Churn or Continue.
- Regression models predict the value of a real number variable such as Revenue, Sales, or Success Rate.

Creating models of last year's sales receipts combined with demographics, credit ratings, and segmentation strategies is called model training. You can save these models as SAS programs that can be used to predict values on new input data in the absence of target data. These score code programs can be used in any Base SAS environment. You can apply the score code to new data to make business decisions such as how much capacity to build or which customers to select for special offers. This process is named model scoring.

CHURN CLASSIFICATION EXAMPLE

SAS Rapid Predictive Modeler is best presented through an example analysis. Churn refers to the tendency of a subscriber to switch providers; it is one of the most common problems faced globally in the telecommunication industry. Reasons why a customer might churn include a competitive stimulation, unhappiness with service after the sale, dissatisfaction with quality of services, a move to another location, or disconnection by the provider due to account delinquency. The objective of this analysis is to build a classification model quickly and easily to measure the propensity for an active customer to churn. This enables service agents to take proactive steps to retain targeted profitable customers before churn occurs.

This churn analysis is presented through these steps:

- importing and summarizing the data
- developing models
- · reviewing the model results
- saving the model to a SAS Enterprise Miner project
- registering and scoring the model

IMPORTING AND SUMMARIZING THE DATA

Building a model requires data that represent historical events. You need input data that represent characteristics that can be used for prediction and target data that represent the event or value that you want to predict. Often, the input data are derived from one time period and the target data are derived from a later time period.

To make sure that your model will be effective in production, you should have a large number of observations stored as rows of data. For example, many retail customer models use input data with tens of thousands of observations. The data used for SAS Rapid Predictive Modeler should be organized into rows that represent observations and columns that represent values. One of the columns should represent a target (independent) variable. Consider the following example table:

Name	Current Bill Amount	Account Plan Type	Account Age	Complaint Code	Churn
Dominique	79	Sliver	28	Call quality	Υ
Susan	120	Gold	14		N
Brian	55	Silver	72	Billing problem	N

Table 1. Sample Model Training Table

Name is an ID column and is not used by SAS Rapid Predictive Modeler; *Current Bill Amount, Account Plan Type, Account Age*, and *Complaint Code* are input columns and are used. *Complaint Code* has a missing value for the second customer, which is automatically handled through binning or missing value replacement. You do not need to select input columns. SAS Rapid Predictive Modeler automatically detects which columns should be used as input columns. *Churn* is a target column and is used. You must select one target column to use, and all other columns are treated automatically. You can also select columns that will be excluded from the analysis.

You can select a frequency column that represents the relative weight that should be assigned to that row; for example, in some data sets a single row can represent more than one observation.

The data used for scoring should have all input columns. The target column is optional. When the model is used for making predictions on new data, the target column is missing. When the model is used for monitoring effectiveness, the target column is present. Scoring data usually contain the ID column.

Churn propensity scores are derived from analysis of the historical behavior of customers who churn. Figure 1 shows a partial listing of the sample training data which includes 4,708 customers along with columns that represent the customer id, the churn flag (1=churner, 0=non-churner), and thirteen candidate predictors that cover factors such as phone usage patterns, account billing and status information, technical support complaints, and age of the equipment. You can import data sources from a variety of formats including SAS tables, relational databases, and Microsoft Office Excel spreadsheets.

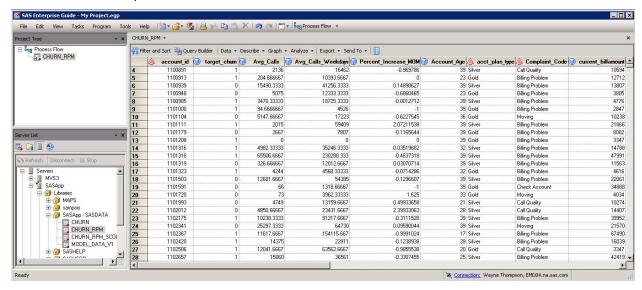


Figure 1. Churn Modeling Table

SAS Rapid Predictive Modeler is included in SAS Enterprise Guide and SAS Add-In for Microsoft Office not only because these are common interfaces for business analysts who are not experienced SAS programmers but also because these products include a complementary set of data preparation, summarization, exploration and reporting tasks to support the SAS Rapid Predictive Modeler analysis. Before you build a SAS Rapid Predictive Modeler model, you might want to use tasks such as *Characterize Data* and *Scatter Plot* to summarize and explore your data to understand trends and anomalies. Often in data mining the target variable contains a rare event; for example, if only 5% of your customers churn for a new offer, you want to make sure that you have a significant number of these customers in your data set. You might want to oversample your data to make sure you select all customers who churned and an equal number of customers who did not churn. Oversampling makes it easier for the model to find a stable solution. Figure 2 displays the results of the *One-Way Frequencies* task, which indicate that the proportion of churners to non-churners is about equal. Because the data have been oversampled, you want to specify prior probabilities in the task as described in the next section of this analysis.

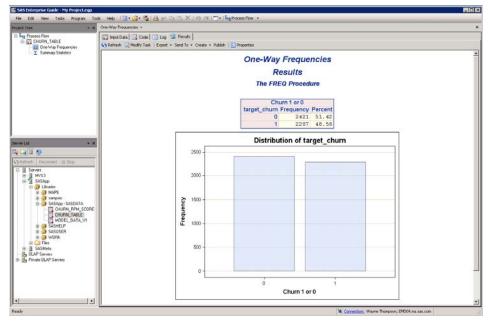


Figure 2. One-Way Frequencies for the Target Churn

BUILDING A CHURN CLASSIFICATION MODEL

SAS Rapid Predictive Modeler can be executed from within SAS Enterprise Guide or the SAS Add-In for Microsoft Office. The task is identical in either product and requires a SAS Enterprise Miner license. Figure 3 shows that the SAS Rapid Predictive Modeler task is organized along with the model scoring task within the Data Mining group. Many business analysts prefer to work directly in Microsoft Office Excel. One benefit of using SAS Enterprise Guide is that it provides project-based storage for your analyses along with a process flow diagram to manage the steps in your analysis.

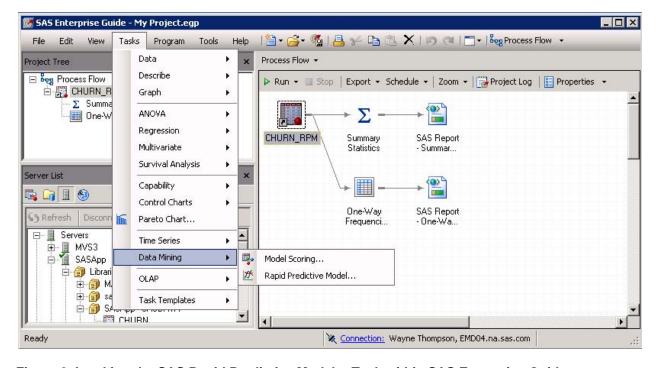


Figure 3. Invoking the SAS Rapid Predictive Modeler Task within SAS Enterprise Guide

SAS Rapid Predictive Modeler prompts you for the input data source if an active data set is not already defined. Figure 4 shows the Data panel of the SAS Rapid Predictive Modeler task. You assign modeling roles by dragging them from the **Inputs Variables** list to the appropriate **Modeling Roles** list. You are only required to specify the dependent variable (target) from your table that represents the value that you want to predict. The SAS Rapid Predictive Modeler automatically decides whether variables are continuous or categorical. The SAS Rapid Predictive Modeler automatically sets all variables as input variables (predictors), which is very convenient when you have a large number of candidate inputs. You can select variables that you want to exclude. You can also select a frequency variable that represents a weight that you want to apply to the rows of data. You can select ID variables that are to be excluded from the analysis. It is often helpful to identify these variables for reporting and score selection.

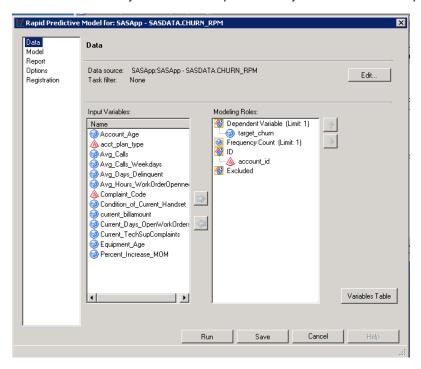


Figure 4. SAS Rapid Predictive Modeler Data Panel

You specify the type of model you want to build in the Model panel. SAS Enterprise Miner 6.1 modeling functions are executed when you run the SAS Rapid Predictive Modeler to generate each model. The basic model samples the data only in the case where you have a rare target event and then partitions the data using the target as a stratification variable. It then applies a one-level decision-tree-based variable selection step. The selected input variables are then binned with respect to their relationship with the target and are passed to a forward stepwise regression model.

The intermediate model is an extension of the basic. Multiple transformations are provided to several variable selection techniques. A decision tree, regression model, and a logistic regression (which contains the node variable from a decision tree from a predecessor node) are used as modeling techniques. The node variable exported from a decision tree is one way to represent interactions. The advanced model extends the intermediate methodology to include a neural network model, advanced regression analysis, and ensemble models.

The basic model runs faster than the intermediate model but might create a less accurate model. The same is often true as you progress from using the intermediate to the advanced model.

Figure 5 shows the Decisions and Priors window, where you set optional information about categorical target variables. SAS Rapid Predictive Modeler automatically makes a pass only through the data for the target variable to set the event level based on descending order and also to calculate the data proportion for each target level. The model comparison statistics are dependent on the target event level which for this analysis is correctly set at 1.

Before you develop predictive models, it is important to specify the correct priors to properly adjust model predictions regardless of what the proportions are in the training data. If no prior probabilities are used, the estimated posterior probability for the churn=1 event class would be too high. In this example the training data have been oversampled to include 49% churners (1's) and 51% non-churners (0's). However, it is known that the population historically contains about 4% churners and 96% non-churners, so the priors are adjusted accordingly as shown in Figure 5.

You use the decision options to bias the model construction to a particular event. For example, if it is more important to identify churners as churners than non-churners as non-churners, then enter a higher weight in the (1, 1) cell of the decision table. Select inverse priors to bias the model to identify both true positive and true negative predictions at the risk of incorrectly estimating false positive and false negative predictions.

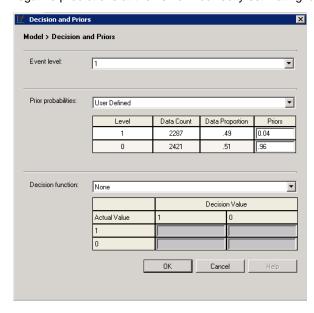


Figure 5. Decisions and Priors Model Options

The SAS Rapid Predictive Modeler automatically generates a standard report including lift, receiver-operator characteristic (ROC) charts, and a scorecard. You select additional reports in the Report panel; these include model summarization, variable rankings, crosstabulations, classification matrix, and detailed fit statistics. Reports are reviewed in the next section.

Once you are satisfied with your model settings, click **Run** to execute the task. Figure 6 shows the SAS Rapid Predictive Modeler task executing within a SAS Enterprise Guide process flow. You can modify the task to select a different candidate set of input variables, modeling method, or reports items (or any combination) and rerun it. You can also add more than one SAS Rapid Predictive Modeler task to your SAS Enterprise Guide process flow.

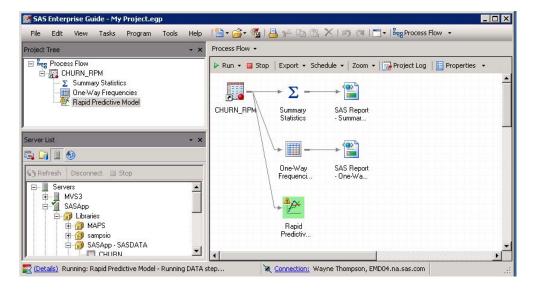


Figure 6. SAS Rapid Predictive Modeler Task Running within a SAS Enterprise Guide Process Flow

REVIEWING THE MODEL RESULTS

Business analysts and statisticians often spend a significant amount of time compiling model results for review with their coworkers. The SAS Rapid Predictive Modeler includes a concise set of core reports for reviewing what data source and variables were used for modeling, a ranking of the important predictor variables, several fit statistics for evaluating the accuracy of the model, and a model scorecard. The SAS Rapid Predictive Modeler automatically selects the best fitting model and displays the report. For brevity, only a subset of the reports is presented here.

Two sets of statistics are reported: training and validation. The SAS Rapid Predictive Modeler process divides the input data into training data and validation data. This partitioning is carefully controlled to make sure that the division results in two data sets that accurately represent the population. The training data are used to compute the parameters for each model, resulting in the training fit statistics. The validation data are then scored with each model, resulting in the validation fit statistics are used to compare models and detect overfitting. If the training statistics are significiantly better than the validation statistics, then you would suspect overfitting, which is when the model is trained to detect random signals in the data. Models with the best validation statistics are generally preferred unless you have some reason to believe that either the model was biased by the validation data or that the validation data are not representative as in the case of a rapidly changing population. In this case, you continue with a model selected based on validation data statistics.

Figure 7 shows the model fit statistics for the selected model. Sample statistics include the misclassification rate, root average squared error, Kolmogorov-Smirnov statistic, Gini coefficient, and lift at various depths of file. Fit statistics are included for both the training and valdiation data. The SAS Rapid Predictive Modeler also outputs a model fit comparison report when the advanced model is selected. In this analysis, the SAS Enterprise Miner Autoneural algorithm was selected as the best model based on maximizing cumulative lift for the validation data. The Reg2 model is a forward logistic regression from the basic modeling methodology that resulted in a baseline model by selecting no terms from your input variables. This can happen when there is a weak signal or a signal that depends on combinations of multiple terms.

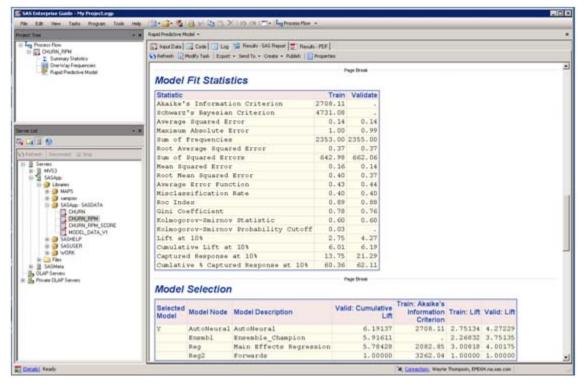


Figure 7. Goodness of Model Fit and Comparison Statistics

The most common report feature for a class target variable model is the Model Gains Chart. For this chart the customer cases are sorted from left to right by the individuals who are most likely to have churned as predicted by the selected model. The cumulative captured response is a measure of how many class target events are identified in each percentile. Figure 8 shows that about 41% of the events have been identified in the first 5% of cases as ranked by the predicted values. At the twentieth percentile, just over three-fourths of the target event cases have been identified. Lift is a measure of the ratio of target events identified by the model to target events found by random selection. Most business users know how to use a gains chart. This plot is available only for models of class target variables.

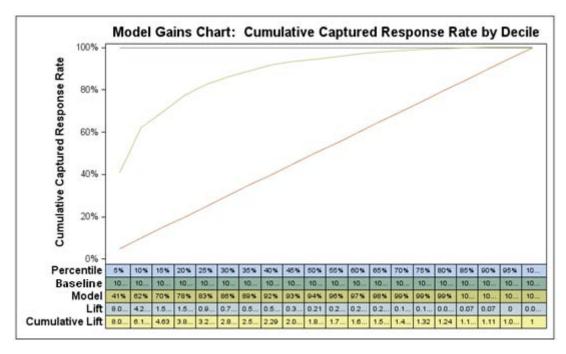


Figure 8. Cumulative Model Gains Chart

The Receiver-Operator Characteristic (ROC) plot is adapted from the field of engineering. This plot shows the maximum predictive power for a model for the entire sample rather than for a single decile. The data are plotted as sensitivity versus (one minus) specificity. The separation between the model curve and the diagonal, representing a random selection model, is termed the Komolgorov-Smirinov (KS) value. A higher KS value represents a more powerful model. Figure 9 shows the ROC plot for the model.

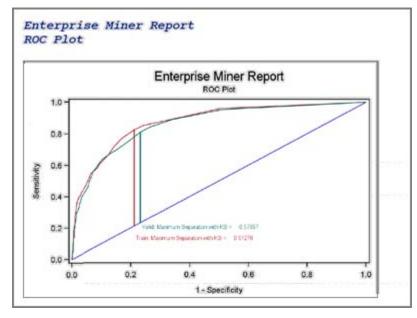


Figure 9. Receiver Operating Characteristic Plot

The SAS Rapid Predictive Modeler always outputs a scorecard to interpret the model's characteristics for business purposes. Each interval variable is binned into distinct ranges of values. Each variable is ranked by importance in the model and scaled to a maximum of 1,000 points. Each distinct value of each variable then receives a portion of the

scaled point total. Scorecards provide a quick view into the behavior of the model. Figure 10 displays the churn scorecard developed using the SAS Rapid Predictive Modeler task within the SAS Add-In for Microsoft Office. Customers with account ages less than or equal to 21.5 months who have lower average call durations measured in seconds during the weekdays and who also tend to have both smaller percent usage increases from month to month and older equipment are more likely churners. Involuntary churn occurs when the customer account is closed by the provider. Customers with high average numbers of days in delinquency are naturally more at more risk of being canceled. Customers who are dissatisfied with call quality or have billing problems are assigned a much higher score than customers who are just calling to check their account status. Moving and the other attributes for complaint code also have fairly high scores.

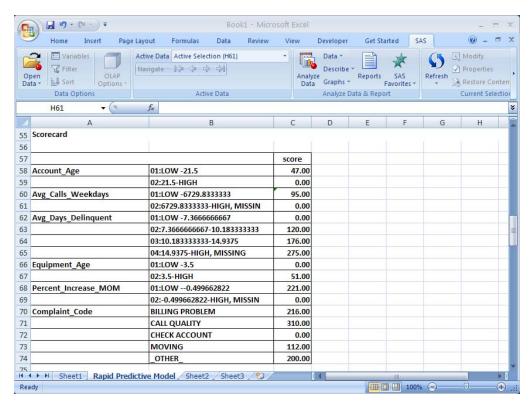


Figure 10. SAS Rapid Predictive Modeler Scorecard Shown in SAS Add-In for Microsoft Office

SAVING THE ANALYSIS TO A SAS ENTERPRISE MINER PROJECT

You can optionally save your SAS Rapid Predictive Modeler analysis to a SAS Enterprise Miner project via the Options panel of the SAS Rapid Predictive Modeler task. This provides tremendous flexibility to share and customize the model using additional SAS Enterprise Miner tools. The ability to extend SAS Rapid Predictive Modeler models within SAS Enterprise Miner supports more of a white-box-versus-black-box automated modeling delivery and also fosters collaboration between business analysts and experienced data miners. The model can be registered from SAS Enterprise Miner to the SAS Metadata Repository for importing into other SAS Enterprise Miner process flows to support integrated model comparison, or it can be deployed to other SAS applications.

Several functions in SAS Enterpise Miner were modified to enable the automated analysis of the SAS Rapid Predictive Modeler; these functions include the Input Data Node, the Transform Node, the Score Node, the Metadata Node, and the Reporter node. These changes also directly benefit SAS Enterprise Miner users.

The basic, intermediate, and advanced model processes were developed based on numerius tests with customer data and simulations of various known modeling scenarios. The basic model process developed includes the following steps implemented as SAS Enterprise Miner process flow diagrams:

- 1. data summarization and detection of ID, class, and continuous variables
- 2. imputation of missing values by using SAS Enterprise Miner techniques

- statistical sampling to produce an efficient and representative data set that accounts for rare values and skewed distributions
- 4. partitioning rows into training and validation data
- 5. transformations that reveal nonlinearity and contain the effects of outliers
- 6. variable selection methods
- 7. reliable modeling techniques that produce interpretable functions
- 8. model selection based on statistics that promote generalization

You can open SAS Rapid Predictive Modeler projects inside SAS Enterprise Miner to examine these details and try out modifications and alternatives.

REGISTERING AND SCORING THE MODEL

After the model has been computed, you can also register the model to the SAS Metadata Repository from within the SAS Rapid Predictive Modeler task. A registered model can be used by your colleagues in their scoring processes in SAS Enterprise Guide or SAS Add-In for Microsoft Office, or imported into their SAS Enterprise Miner sessions for comparison, review, and further development.

The model can also be imported into SAS Model Manager, both for management with all of your other modeling assets and also for monitoring model degradation. SAS Data Integration Studio users can import models to create managed and scheduled scoring processes. Figure 11 shows applying a registered SAS Rapid Predictive Modeler model to score new data with the Model Scoring task. SAS Rapid Predictive Modeler models can also be published using the SAS Scoring Accelerator for Teradata, DB2, or Netezza.

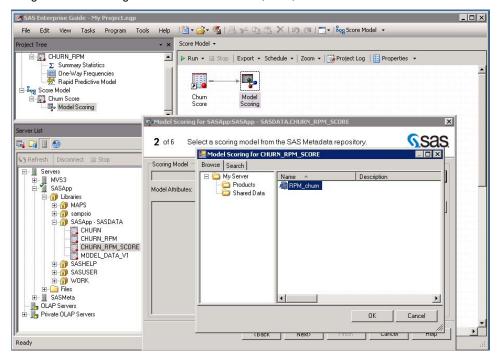


Figure 11. Model Registration from SAS Enterprise Guide

CONCLUSION

Organizations today are increasing their use of predictive analytics to more accurately predict their business outcomes, to improve business performance, and to increase profitability. SAS Rapid Predictive Modeler provides business analysts with self-sufficient access to an automated easy-to-use predictive modeling task to support a variety of applications such as cross-selling, up-selling, retention, and acquisition. SAS Rapid Predictive Modeler also delivers proven best-practices modeling methodologies for routine modeling building tasks, freeing up time for the experienced statistician to focus on more complex data mining problems. Although the SAS Rapid Predictive Modeler is very automated and easy to use, it is not a black box. The analysis can be saved to a SAS Enterprise Miner project for support further customizing. The SAS Rapid Predictive Modeler also generates key reporting elements necessary for reviewing the model results with decision makers. SAS Rapid Predictive Modeler models can also be registered to SAS metadata, enabling full integration with the SAS Business Analytics Framework. Models can be scored using the scoring transformation of SAS Data Integration Studio or the model scoring task of SAS Enterprise Guide and SAS Add-In for Microsoft Office. The SAS Rapid Predictive Modeler score code is also fully compatible with the SAS Scoring Accelerators to support in-database scoring. Models can be registered to SAS Model Manager for ongoing maintenance and monitoring of the model. The authors anticipate that many new SAS users will leverage the power of SAS data mining through this easy-to-use modeling tool.

REFERENCES

- SAS Institute Inc. 2009. "SAS Add-in for Microsoft Office." http://www.sas.com/resources/factsheet/sas-ms-office-addin-factsheet.pdf
- SAS Institute Inc. 2009. "SAS Enterprise Guide 4.2 Fact Sheet." 2009. http://www.sas.com/resources/factsheet/sas-enterprise-guide-factsheet.pdf
- SAS Institute Inc. 2009. SAS Enterprise Guide 4.2 Reference Help. Cary, NC: SAS Institute Inc.
- SAS Institute Inc. 2009. "SAS Enterprise Miner 6.1 Fact Sheet." http://www.sas.com/technologies/analytics/datamining/miner/factsheet.pdf
- SAS Institute Inc. 2009. SAS Enterprise Miner 6.1 Reference Help. Cary, NC: SAS Institute Inc.
- Wielenga, Doug. 2007. "Identifying and Overcoming Common Data Mining Mistakes." Proceedings of the SAS Global Forum 2007 Conference. Cary, NC: SAS Institute Inc.

ACKNOWLEDGMENTS

The authors thank Billie Anderson, Michael Burke, Susan Haller, Kevin Hodge, Brian Johnson, Jagruti Kanjia, Dominique Latour, Bob Lucas, Renee Sember, Carol Thompson and Doug Wielenga for helping design, develop and test SAS RPM.

CONTACT INFORMATION

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

David Duling Wayne Thompson

S6102 S6100

SAS Institute Inc.
SAS Campus Drive
SAS Campus Drive
SAS Campus Drive

Cary, North Carolina, 277513 Cary, North Carolina, 277513

<u>David.Duling@SAS.com</u> Wayne.Thompson@sas.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.