Paper 147-2011

Creating a SAS® Model Factory Using In-Database Analytics

John Spooner, SAS Institute, United Kingdom

ABSTRACT

With the ever-increasing number of analytical models required to make fact-based decisions, as well as increasing audit compliance regulations, it is more important than ever that these models be created, monitored, retuned, and deployed as quickly and automatically as possible. This paper, using a case study from a major UK financial organization, shows how organizations can build a model factory efficiently using the latest SAS technology that uses the power of in-database processing. It focuses on how to integrate SAS® Enterprise Miner™, SAS® Model Manager, and SAS® Data Integration Studio, while leveraging the power of highly performant database applicances such as Teradata and Netezza.

INTRODUCTION

Today, organizations are experiencing an analytics revolution. The use of predictive analytics that enable business outcomes to be predicted accurately is increasing at an exponential rate. The output of analytical models is driving many operational processes for a wide range of organizations. Analytical models are now fundamental to the success of a business.

- It is critical that a bank can accurately predict whether a credit card transaction is fraudulent or valid within a nanosecond of the swipe. Get it wrong and the bank is at risk of losing a large amount of money from a fraudulent transaction, or annoying a high-value customer who will look into moving its business elsewhere.
- It is essential that a company can accurately predict when a customer is about to move to a competitor, understand what is driving that behaviour, and have the ability to make an offer that stops the customer from leaving. Get it wrong and the company is at risk of losing highly profitable customers to its competitors, or wasting money offering incentives to customers who are not at risk. This is particularly relevant for Communications Service Providers. They struggle to differentiate themselves by brand, coverage and customer service, and their customers will switch providers readily, based on who can offer the best deal right now.

Organizations have hours, minutes, or seconds to make critical business decisions. Analytical models are at the heart of these decisions, so it is critical that these analytical models are created using robust and industrial-strength processes. Data that is in the correct structure to build a model needs to be rapidly created, accessed, and used. Models need to be rapidly built and tested. These models need to be deployed into a production environment with minimal delay. The output from these production models must be generated quickly. These models must be constantly monitored to ensure that the output is as accurate as possible. Underperforming models need to be quickly replaced by more up-to-date models.

Organizations are taking months, sometimes years, to move through this end-to-end process. This results in less than optimal decisions, and organizations are either losing money or missing the opportunity to maximise their revenues.

This paper explains how an organization can create a more robust end-to-end process by using SAS technology to create a model factory. With a model factory, hundreds of models with a consistent framework can be quickly generated and deployed.

THE MODEL FACTORY IN ACTION

A major UK financial organization identified that its cycle time from model initiation to model deployment was unsatisfactory for the 21st century. The process was manual, error-prone, and resource-intensive. The process had little or no monitoring to identify model degradation.

In conjunction with SAS and the database vendor Teradata, the organization built a flexible and efficient platform for a model factory to seamlessly integrate SAS tools for data management, model development, and model deployment using in-database technology. The platform harnesses the power of the Teradata environment for data preparation and model scoring and uses the power of SAS analytics to build the models. Over 55 million records can be scored within Teradata many times during the day. This could not have been accomplished with the older process. The three months of lag time that it usually took for a model to be promoted to a production environment was dramatically reduced to days. There was a 40% reduction in data preparation. Analysts are 50% more productive.

THE MODELING FRAMEWORK

The modeling framework consists of six steps. All organizations must complete these six steps for predictive analytics to be effective in their processes. Figure 1 shows the six steps of the modeling framework.

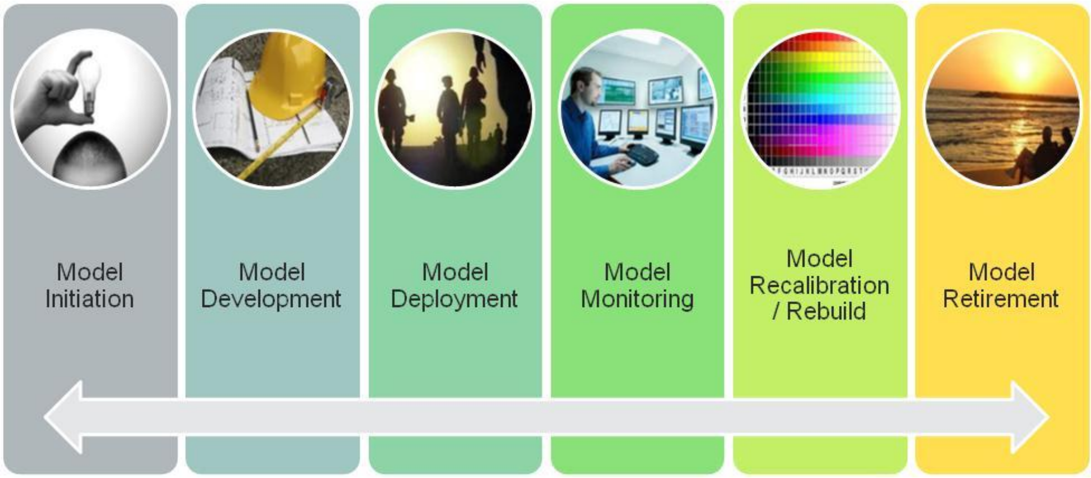


Figure 1: The Modeling Framework

Model Initiation

The need and scope of a model are specified by different business areas, such as marketing, fraud, or credit risk.

Model Development

The source data for modeling is extracted from a wide range of sources, including an enterprise data warehouse (EDW). These sources might be sporadically stored across the organization, and they might use different code, tools, and transformations. In this step, the data is combined to create an analytical data mart. The mart is used as the basis to build the model. A skilled analyst builds the model using statistical and data mining packages.

Model Deployment

Once a model is built, it is deployed into an environment where it can be executed. If the development and deployment coding languages are different, code must be translated manually from the development language (for example, SAS) to the deployment language (for example, COBOL or SQL).

A model is tested to make sure that it meets certain coding standards, generates the correct output that the analyst expects, and does not cause any technical issues when deployed in the production environment.

Following preproduction testing, formal approval is required by a release manager to promote the model to the production environment. The preproduction code is migrated to the production environment, and the model is run in real time or in a batch process.

Model Monitoring

The predictive performance of a model is monitored to ensure that the model is up-to-date and performing well.

Model Recalibration or Rebuild

Over time, the model degrades to a point where a new model needs to replace it. This need could be based on a change in environment, or simply because things change over time. This new model can be a recalibration, in which the new model uses the same characteristics as the existing model, but the weightings of the characteristics are updated. Or, the model can be rebuilt to include new and existing characteristics.

Model Retirement

A model that is no longer required by the business is retired.

CHALLENGES OF IMPLEMENTING THE MODELING FRAMEWORK

As organizations work through the framework, there are three main challenges that they encounter

Inconsistent Data in the Wrong Format

Typically, the EDW does not contain all of the data in the right format that is required for model development. Analysts tend to create their own analytical data marts. They create additional summarizations of the data in the EDW. And, they extract data from other systems. All of these data collection methods lead to inconsistent data, with different versions of data used to build the model. As a result, the data in the model development environment is different from the data in the model deployment environment. Some data is kept on file servers, some is on mapped drives, and, in extreme cases, some data is on local hard drives. This causes major issues with data governance. Data on local hard drives results in processing on the desktop, which can mean poor performance.

Cross-Functional Involvement

The steps of the framework require multiple areas of the business. For example, an analyst is required to build the model. A run-and-operate team in IT deploys and runs the model. A database administrator centrally manages the EDW. As each business area hands off the results to another area, there is a delay. Each area has to check and validate what has happened to the model and data. In many areas, it is common that a model is recoded. There might be many iterations of the model before the best model is ready. Because of so much manual involvement, incorrectly coded models can be deployed into a production environment.

Different Technologies Being Used

Data in an EDW that is being used by many different analytical teams results in multiple copies of large data tables. All of these copies lead to duplicate storage and large data transfers across networks. Analysts might use different analytical software to build the models. When it is time to promote the model to a production environment, analysts have to translate the model into a common standard. Models degrade over time, so they must be constantly monitored and managed. Model monitoring in the production environment is not consistent because different Microsoft Excel spreadsheets can be used to monitor the model. Models degrade over time, so they must be constantly monitored and managed.

THE MODEL FACTORY SOLUTION USING IN-DATABASE ANALYTICS

Organizations need to reduce the impact and challenges of implementing the modeling framework. The time that it takes to complete the steps of the framework needs to be reduced dramatically. The solution to this is a model factory using in-database analytics. Figure 2 shows how the model factory is constructed.

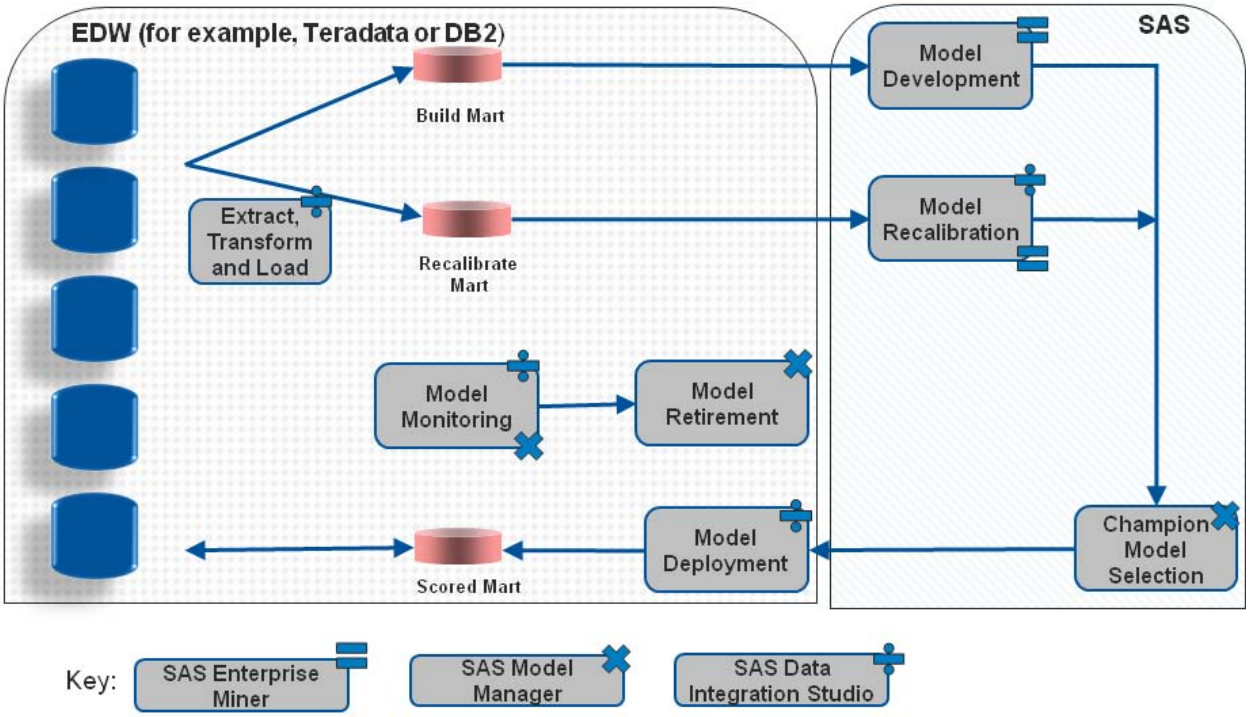


Figure 2: The Model Factory in Production

The model factory uses SAS in-database processing technology to take advantage of the massively parallel processing (MPP) architecture of the database or data warehouse for scalability and better performance. Moving data management and analytics to where the data resides is beneficial in terms of speed. This move reduces other unnecessary data moves, and promotes better data governance. The SAS technology stack that is needed to power the model factory is SAS Enterprise Miner, SAS Data Integration Studio, SAS Model Manager, SAS/ACCESS® software for the EDW, and SAS® Scoring Accelerator for the EDW. More information about these products can be found at http://www.sas.com/software/.

MODEL DEVELOPMENT

Data Preparation

The EDW provides the primary source of data for analysis. SAS Data Integration Studio is used to create extract, load, and transform (ELT) routines that produce the analytical build marts. During the extraction phase, a data source or system is accessed, and only the data that is needed is extracted. The data is staged in the EDW using high-speed loaders to ensure that the data is loaded quickly. The data is transformed into a structure that is fit for model building, and raw data is summarized to create derived fields. ELT routines are performed using standard SAS Data Integration Studio transforms, and SQL is automatically generated that performs the field derivation directly inside the EDW.

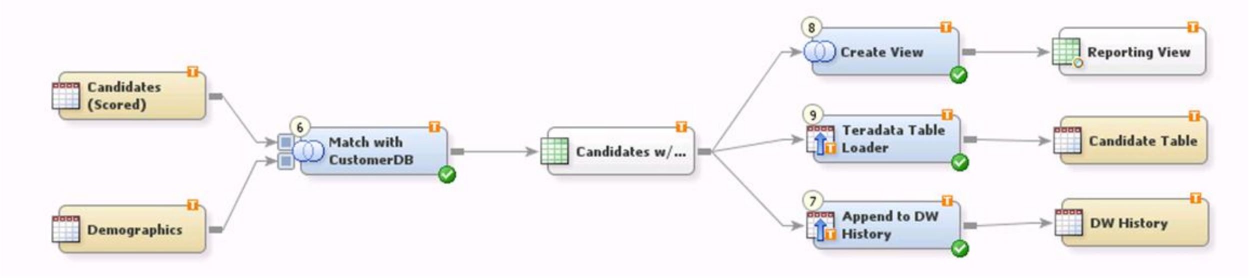


Figure 3: An Example Process Flow from SAS Data Integration Studio

Model Build

Model building is performed by analysts in SAS Enterprise Miner. Analysts can build their own processes using prebuilt nodes aligned to the SAS SEMMA (Sample, Explore, Modify, Model, and Assess) methodology. Alternatively, the SAS Rapid Predictive Modeler, a component of SAS Enterprise Miner, can automatically guide an analyst through a behind-the-scenes workflow of data preparation and data mining tasks. The modeler provides basic, intermediate, and advanced templates that automatically treat the data to handle outliers, missing values, rare target events, skewed data, correlated variables, variable selection, and model building and selection.

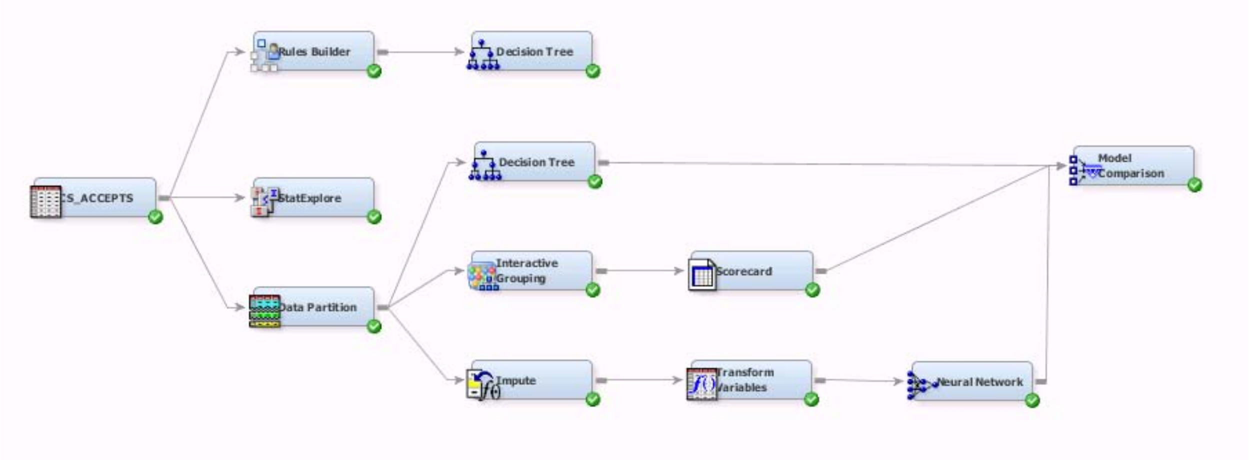


Figure 4: An Example Process Flow from SAS Enterprise Miner

Champion Model Selection

When model development is complete, analysts register a model package centrally in SAS metadata. A model package contains all of the SAS scoring code that is needed when the model is deployed. It contains optimized scoring code that includes only the necessary transformations from the scoring code, eliminating any transformations that are redundant.

The model package contains information about the name of the algorithm that was used, the name of the data mining analyst, and the time of creation. This is useful information for data governance.
This model package can be shared with other SAS applications, such as SAS Model Manager. SAS Model Manager

enables organizations to standardize the process of creating, managing, deploying, and monitoring analytical models. It provides a central and secure repository for storing the analytical models.

A model package is associated with a project. This enables the model deployment team to see the model that has been built by the analyst and all of the associated output. A review of the model ensures that the right steps have been taken and a suitable and robust model is released into the production environment.

MODEL DEPLOYMENT

Model Deployment

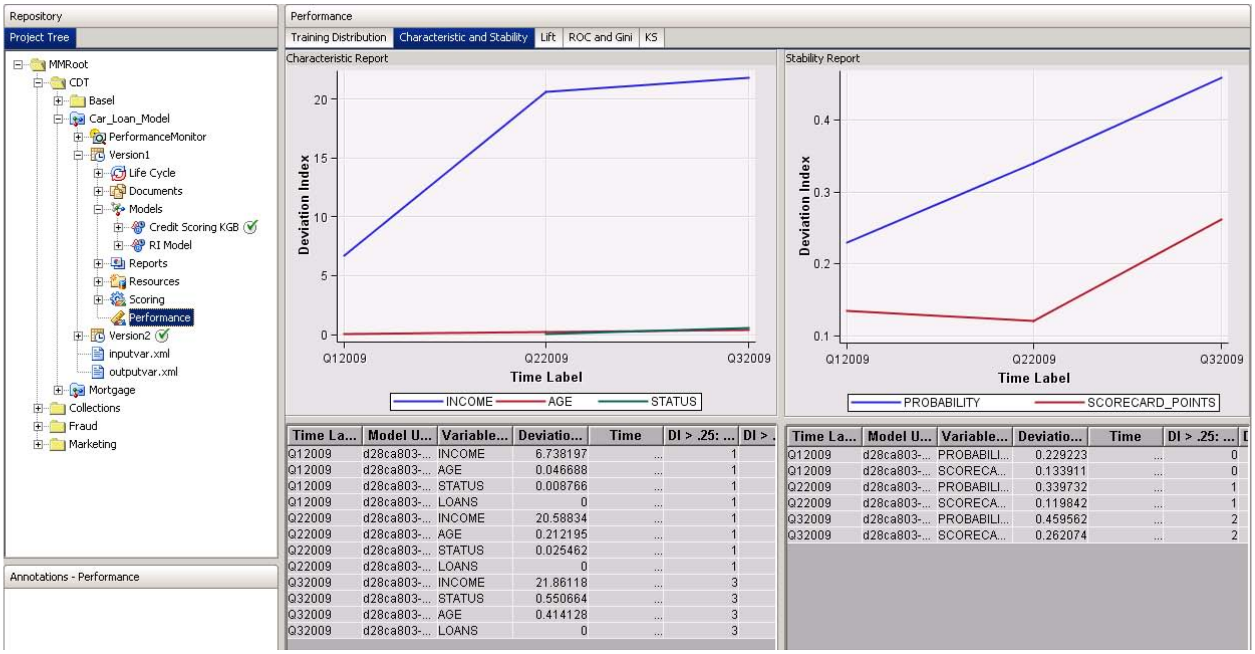
Once a model has been reviewed, signed off, and declared ready for production, it is considered champion status in SAS Model Manager. The model is converted into a vendor-defined function (VDF) in SAS Model Manager, and it is placed in the EDW. This is performed with a click of a button in SAS Model Manager, and it is powered by the SAS Scoring Accelerator for the database that the model will reside in. A VDF is a database-specific function that contains all of the scoring logic that is required for a model to run. By placing the VDF in the EDW, the common security, auditing, and administration offered by the EDW can be honored and leveraged.

Model Execution

Model execution is controlled centrally using SAS Data Integration Studio jobs. These jobs control which data tables are used as scoring marts, the model that is used to score the mart, and the creation of a file that contains the scores. The scoring mart is created using in-database processing, and it resides in the database. The model execution job is scheduled to run at a specific time interval, or when it is initiated by a trigger. The model is executed from within a generated transformation in SAS Data Integration Studio, and the model is run directly in the database.

MODEL MONITORING

Once a model is in a production environment, and is being executed at regular intervals, the champion model is centrally monitored because its predictive performance will degrade over time. SAS Model Manager enables you to view a variety of monitoring reports that have been created by scheduled jobs in SAS Data Integration Studio. The champion model can be retired when its performance degradation hits a certain threshold, and it can be replaced with a new model that has been recalibrated or rebuilt.



Time La...	Model U...	Variable...	Deviatio...	Time	DI > .25: ...	DI > .
Q12009	d28ca803-...	INCOME	6.738197	...	1	
Q12009	d28ca803-...	AGE	0.046688	...	1	
Q12009	d28ca803-...	STATUS	0.008766	...	1	
Q12009	d28ca803-...	LOANS	0	...	1	
Q22009	d28ca803-...	INCOME	20.58834	...	1	
Q22009	d28ca803-...	AGE	0.212195	...	1	
Q22009	d28ca803-...	STATUS	0.025462	...	1	
Q22009	d28ca803-...	LOANS	0	...	1	
Q32009	d28ca803-...	INCOME	21.86118	...	3	
Q32009	d28ca803-...	STATUS	0.550664	...	3	
Q32009	d28ca803-...	AGE	0.414128	...	3	
Q32009	d28ca803-...	LOANS	0	...	3	

Time La...	Model U...	Variable...	Deviatio...	Time	DI > .25: ...
Q12009	d28ca803-...	PROBABILI...	0.229223	...	0
Q12009	d28ca803-...	SCORECA...	0.133911	...	0
Q22009	d28ca803-...	PROBABILI...	0.339732	...	1
Q22009	d28ca803-...	SCORECA...	0.119842	...	1
Q32009	d28ca803-...	PROBABILI...	0.459562	...	2
Q32009	d28ca803-...	SCORECA...	0.262074	...	2

Figure 5: Model Monitoring Reports in SAS Model Manager

AUTOMATIC MODEL RECALIBRATION OR REBUILD

SAS Enterprise Miner supports a concept known as batch processing. Batch processing is a SAS macro-based interface to the SAS Enterprise Miner client/server environment that operates without running the SAS Enterprise Miner client. Batch processing supports the building, running, and reporting of SAS Enterprise Miner process flow diagrams.

SAS Enterprise Miner batch processing code can be created directly from the Enterprise Miner process flow that was used to create the model. The batch processing code is accessed through a custom-generated transformation in SAS Data Integration Studio, and is scheduled to run at regular time intervals as part of a SAS Data Integration Studio job.

This SAS Data Integration Studio job accesses the latest version of the analytical data mart. It can either recalibrate or rebuild a model, depending on the process. A model package is created and automatically registered in SAS Model Manager. Users can view the model in SAS Model Manager, and decide whether the model should replace the existing champion model.

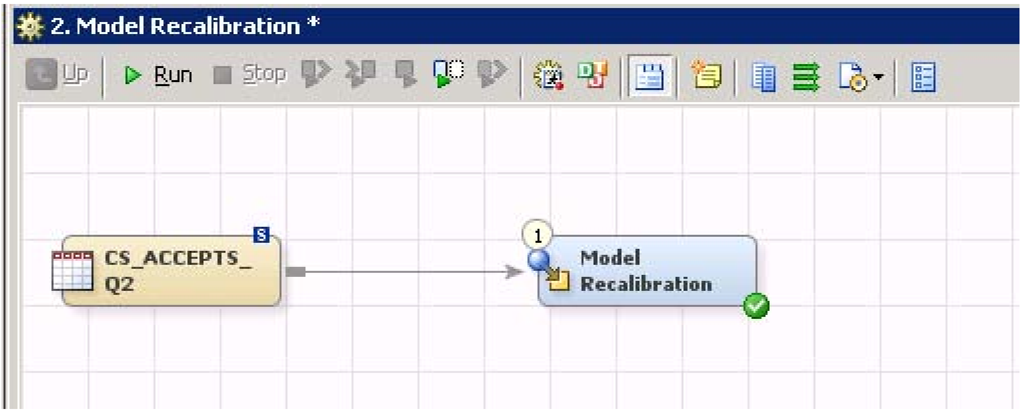


Figure 6: A Model Recalibration Job in SAS Data Integration Studio

CONCLUSION

By implementing a SAS model factory that uses in-database processing, organizations gain the following benefits:

- **Reduced development time**. A model factory uses integrated SAS components to reduce the modeling lifecycle by eliminating redundant steps.
- **Reduced deployment time**. The process of converting scoring code into logic that is placed directly in the EDW happens automatically. This eliminates the timely and error-prone manual process of translating the model.
-
- **Faster scoring processes**. Because the model is scored directly in the database, the model execution job uses the scalability and processing speed offered in the database. This reduces scoring times from hours and days to minutes and seconds.
- **Active monitoring and management of models**. The model factory allows standard monitoring reports to be created and reviewed.
- **Reduced risk**. By using consistent processes and technologies for model development and deployment, any risks involved in the modeling process can be reduced.

The model factory provides a robust framework for any modeling environment. It supports a wide range of modeling applications, such as churn modeling, risk scorecard development, and cross-sell up-sell modeling.

ACKNOWLEDGMENTS

The author would like to thank Neil Griffin, SAS UK Professional Services, for all of his work implementing the model factory at the major UK financial organisation. The author would also like to thank all of the analytics team at a major UK communications organisation who first worked on the requirement of automating the rebuilding and recalibrating process.

REFERENCES

SAS Institute Inc. 2007. SAS Institute white paper. “Best Practices for Managing Predictive Models in a Production Environment.” http://www.sas.com/resources/whitepaper/wp_3515.pdf.

Schubert, Sascha. 2008. “Tailoring the Use of SAS® Enterprise Miner™.” *Proceedings of the SAS Global Forum 2008 Conference*. Cary, NC: SAS Institute Inc. Available at http://www2.sas.com/proceedings/forum2008/145-2008.pdf.

Stander, Jeff. 2010. “SAS® Data Integration Studio: Tips and Techniques for Implementing ELT.” Proceedings of the SAS Global Forum 2010 Conference. Cary, NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings10/116-2010.pdf.

CONTACT INFORMATION

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

John Spooner
SAS Software
Wittington House, Henley Road
Marlow, SL7 3HA, United Kingdom
John.Spooner@suk.sas.com