Data Mining from A to Z: Better Insights, New Opportunities

WHITE PAPER

Table of Contents

Content for this paper was provided by Tapan Patel and Wayne Thompson.

Patel is a Global Marketing Manager at SAS with more than 12 years of experience in enterprise software. He leads global product marketing initiatives for predictive analytics and data mining as well as in-database analytics, in-memory analytics and high-performance analytics. Before coming to SAS, Patel worked as a Senior Product Manager at HAHT Commerce Inc. He has an MBA from North Carolina State University, with a concentration in information technology management.

Thompson is the Manager of SAS Predictive Analytics Product Management. He is a renowned presenter, teacher, practitioner and innovator in the field of predictive analytics technology. Over the course of his 20-year tenure at SAS, he has been credited with bringing to market landmark SAS Analytics technologies (SAS Enterprise Miner, SAS Text Miner, Credit Scoring for SAS Enterprise Miner, SAS Model Manager, SAS Rapid Predictive Modeler, SAS Scoring Accelerator for Teradata and SAS Analytics Accelerator for Teradata). Thompson received his PhD and MS degrees from the University of Tennessee.

Introduction

Most organizations are awash in data – too much of it. And as many have learned, the ability to make effective, fact-based decisions is not dependent on the amount of data you have. Success is based on your ability to discover more meaningful and predictive insights from all the data you capture.

That's where predictive analytics and data mining come into play. Data mining looks for hidden patterns in your data that can be used to predict future behavior. Businesses, scientists and governments have used this approach for years to transform data into proactive insights. The same approach applies to business issues across virtually any industry.

Forward-thinking organizations use data mining and predictive analytics to help them detect fraud, minimize risk, anticipate resource demands, increase response rates for marketing campaigns, curb customer attrition and identify adverse drug effects during clinical trials. Because of its potential to yield predictive insights from masses of diverse data points, data mining is essential for improving performance and creating competitive advantage for all types of organizations.

Predictive analytics and data mining can help you to:

- Rapidly discover new, useful and relevant insights from your data.
- Make better decisions and act faster.
- Monitor analyses and results to verify their continued relevance and accuracy.
- Effectively manage a growing portfolio of predictive modeling assets.

The five steps of the SAS data mining process are at the heart of this approach, as defined by SEMMA. SEMMA refers to an approach in which you:

- **Sample** the data by creating a target data set large enough to contain the significant information.
- **Explore** the data by searching for anticipated relationships, unanticipated trends and anomalies – to gain deeper understanding and ideas.
- **Modify** the data by creating, selecting and transforming the variables to focus your model selection process.
- **Model** the data by using analytical tools to search for a combination of data that reliably predicts a desired outcome.
- **Assess** the data and models by evaluating the usefulness and reliability of the findings from the data mining process.

Forward-thinking organizations use data mining and predictive analytics to help them detect fraud, minimize risk, anticipate resource demands, increase response rates for marketing campaigns, curb customer attrition and identify adverse drug effects during clinical trials.

SAS provides several important components that play a part in this process. SAS Enterprise Miner is a comprehensive workbench for building predictive and descriptive models, following the SEMMA process. SAS Rapid Predictive Modeler enables business analysts to generate models without in-depth statistical knowledge. SAS Model Manager provides a structured framework for managing, validating, deploying and monitoring analytical models.

How Do Predictive Analytics and Data Mining Work?

Data mining is one of the core sets of technologies that helps organizations use predictive analytics to anticipate future outcomes, find new opportunities and improve business performance. Predictive analytics is an iterative process in which you:

- Select, explore and model large amounts of data.
- Identify meaningful patterns, trends and relationships among key variables.
- Deploy models.
- Assess the merits of various courses of action.

Data mining can be applied to a variety of customer issues in any industry – from segmentation and targeting to fraud detection and credit risk scoring.

Businesses, scientists and governments have used data mining for many years to transform data into predictive insights. It can be applied to a variety of customer issues in any industry – from segmentation and targeting to fraud detection and credit risk scoring, to identifying adverse drug effects during clinical trials.

"A lot of organizations use data mining techniques to segment customers by behavior, demographics or attitudes – to understand what products or services each segment would want or need in the future," said Patel. "Once you have properly identified the segments, you can create response models to predict which customers are likely to respond. You can further complement the customer acquisition model with a credit scoring model to find out which of those customers are good credit risks and worth the investment to acquire or retain."

Data mining outperforms rules-based systems for detecting fraud, even as fraudsters become more sophisticated in their tactics. "Models can be built to cross-reference data from a variety of sources, correlating nonobvious variables with known fraudulent traits to identify new patterns of fraud," Patel said. For its potential to yield predictive insights from masses of diverse data points, data mining has proven to be an invaluable component of any analytics initiative.

Application	What Is Predicted?	Resulting Business Decision
Profiling and segmentation	Customer behaviors and needs by segment.	How to better target product/ service offers.
Cross-sell and up-sell	What customers are likely to buy.	Which product/service to recommend.
Acquisition and retention	Customer preferences and purchase patterns.	How to grow and maintain valuable customers.
Campaign management	The success of customer communications.	How to direct the right offer to the right person at the right time.
Profitability and lifetime value	Drivers of future value (margin and retention).	Which customers to invest in and how to best appeal to them.

Figure 1: Common applications for data mining across industries.

Application	What Is Predicted?	Resulting Business Decision
Credit scoring (banking)	Creditworthiness of new and existing sets of customers.	How to assess and control risk within existing (or new) consumer portfolios.
Market basket analysis (retail)	Products that are likely to be purchased together.	How to increase sales with cross-sell/up-sell, loyalty programs, promotions.
Asset maintenance (utilities, manufacturing, oil and gas)	The real drivers of asset or equipment failure.	How to minimize operational disruptions and maintenance costs.
Health and condition management (health insurance)	Patients at risk of chronic, treatable/preventable illness.	How to reduce health care costs and satisfy patients.
Fraud management (government, insurance, banks)	Unknown fraud cases and future risks.	How to decrease fraud losses and lower false positives.
Drug discovery (life sciences)	Compounds that have desirable effects.	How to bring drugs to the marketplace quickly and effectively.

Figure 2: Industry-specific data mining applications.

Predictive analytics and data mining activities can help you to:

- Discover patterns, trends and relationships represented in data.
- Develop models to better understand and describe characteristics and activities based on these patterns.
- Use those insights to help evaluate future options and make fact-based decisions.
- Deploy scores and results for timely, appropriate actions.
- Manage the life cycle of models and monitor their performance to avoid decay.

The Data Mining Process

Let's consider the steps of the entire SAS data mining process (SEMMA) in more detail.

Sample the Data

To sample the data, create one or more data tables that represent the target data sets. Since data mining can only uncover patterns already present in the data, the sample should be large enough to contain the significant information, yet small enough to process. Target data is generally divided into two sets, the training set and the test set. The training set is used to train the data mining algorithm(s), while the test set is used to verify the accuracy of any patterns found.

Explore the Data

To explore the data, search for anticipated relationships, unanticipated trends and anomalies in order to gain understanding and ideas. You can use several types of techniques for this preliminary data exploration. For example, clustering discovers groups or structures in the data that are similar, beyond the structures known in the data. Classification generalizes a known structure to apply to new data, such as classifying a customer as a good or poor credit risk. Regression works to find a function that models the data with the least error. Association-rule learning searches for relationships among variables, such as products frequently bought together, known as market basket analysis.

Modify the Data

To modify the data, you should create, select and transform the variables to focus the model selection process. Based on your discoveries in the exploration phase, you may need to manipulate your data to introduce new variables, fill in missing values, or look for outliers so you can reduce the number of variables to only the most significant ones.

Model the Data

Model the data by using analytical techniques to search for a combination of the data that reliably predicts a desired outcome. Depending on the data and the issue at hand, you may use a variety of modern techniques to solve your problem. For example, you may use neural networks, random forests, incremental response or time series data mining, as well as industry-specific techniques such as credit scoring in banking or rate making in insurance.

Assess the Data and Models

Assess the data and models by evaluating the usefulness and reliability of the findings from the data mining process. Not all patterns found by the data mining algorithms will be valid. The algorithms might find patterns in the training data set that are not present in the general data set. This is called overfitting. To address this concern, patterns are validated against a test set of data. The patterns learned on the training data will be applied to the test set, and the resulting output is compared to the desired (or known) output.

For example, a data mining algorithm that had been trained to distinguish fraudulent credit card transactions from legitimate ones would then be applied to the test set of transactions on which it had not been trained. The accuracy of the patterns can then be measured from how many credit card transactions are correctly classified. If the learned patterns do not meet desired standards, modifications are made to the preprocessing and data mining until the result is satisfactory and the learned patterns can be applied to operational systems.

You might not include all of these steps in your analysis, and it might be necessary to repeat one or more of the steps several times before you are satisfied with the results.

“A common misconception is to think of SEMMA as a data mining methodology,” said Patel. “It should be considered as a logical model development process – and as such, it can fit in any iterative data mining methodologies or analytical life cycle you have adopted.”

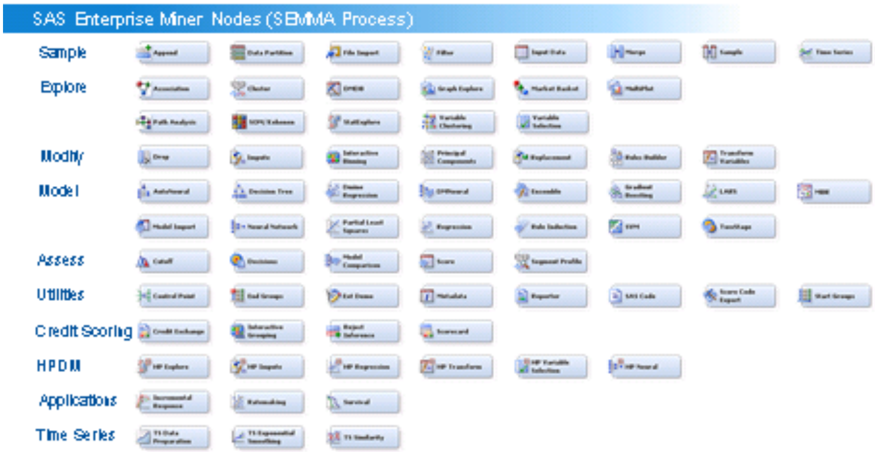

Figure 1: Data mining is an iterative process of learning from the data and applying new insights to continually improve models and processes.

Using SAS® Enterprise Miner™ for Predictive Analytics and Data Mining

SAS Enterprise Miner is a comprehensive, graphical workbench for data mining that follows the SEMMA process. The platform provides capabilities for each step of the process to identify the most significant variables, develop models using the latest algorithms, validate the accuracy and fitness of the model(s), and generate a scored data set with predictive values that can be deployed into your operational applications.

A quality data sample is critical to data mining success. SAS Enterprise Miner provides powerful data preparation tools that address data problems, such as missing values and outliers, and help you develop segmentation rules. Interactive data exploration rules enable users to create dynamic, linked plots to identify relationships within the data. Data modeling takes advantage of a suite of predictive and descriptive modeling algorithms, such as decision trees, neural networks, clustering, linear regression and logistic regression.

Complete, optimized scoring code is delivered in SAS, C, Java and PMML for scoring data in SAS as well as in other environments – or it is delivered as an in-database function for scoring inside industry-leading databases such as Teradata, IBM, Oracle, Pivotal, Aster Data, etc.

The SEMMA data mining process is driven by a process flow diagram that you can modify and save. The drag-and-drop graphical user interface enables the business analyst who has little statistical expertise to navigate through the data mining methodology, while the quantitative expert can go behind the scenes to fine-tune the analytical process.

SAS® Enterprise Miner™

- *Sample:* Create training and test sample data sets with high predictive value.
- *Explore:* Interactively explore relationships and anomalies in the data.
- *Modify:* Create, transform and select the most appropriate variables for analysis.
- *Model:* Apply a range of modeling techniques to identify patterns in the data.
- *Assess:* Validate the usefulness and reliability of findings from the data mining process.

Using SAS® Rapid Predictive Modeler to Make Analytics Mainstream for Business Analysts

As analytics becomes more prevalent across the organization, there is a growing need for business analysts and subject-matter experts to have more self-sufficiency with predictive analysis, said Patel. “Analysts shouldn’t have to rely on statisticians and modelers every time they need to develop new insights from the data, because the turnaround often needs to be fast, and lots of models need to be developed.

“On the other hand, these users do not necessarily understand all the aspects of statistics and data mining, nor do they have the time to develop the data mining process flow diagrams or set up options to drop, modify and add variables. Modelers and statisticians will always play a key role in analytic data preparation and model development, but there is a need to make their work more readily available in a self-service type of application.

“With SAS Rapid Predictive Modeler, business analysts can build models without in-depth statistical knowledge,” said Patel. “The platform automatically steps through a workflow of analytical tasks and draws on SAS Enterprise Miner behind the scenes to come up with a better fitting model to yield better results. Business analysts can find on-demand insights and act on them quickly and effectively in a self-sufficient manner.”

> **“With SAS Rapid Predictive Modeler, business analysts can build models without in-depth statistical knowledge. The platform automatically steps through a workflow of analytical tasks and draws on SAS Enterprise Miner behind the scenes to come up with a better fitting model to yield better results.”**
>
> **Tapan Patel**
> *Global Marketing Manager at SAS*

When business analysts can generate models in a few simple steps, it opens up the door for more people to contribute toward solving problems and seizing opportunities. Business analysts just select data and choose inputs for the outcome they desire. The software automatically processes the input variables and selects the best predictive model. Easy-to-interpret charts deliver dynamic results, helping business analysts determine which predictions will be most valuable.

SAS Rapid Predictive Modeler is also useful for modelers and statisticians – who are in short supply and often juggling a high workload. With SAS Rapid Predictive Modeler, quantitative specialists can offload some of the simple, day-to-day model-building tasks and use the software themselves to quickly generate baseline models.

You can use SAS Rapid Predictive Modeler to:

- Quickly generate accurate predictive models that help you make better decisions.
- Make business analysts self-sufficient by providing a user-friendly interface.
- Deliver analytic results in a simple, consumable way.
- Support model improvement and customization using SAS Enterprise Miner.

Using SAS® Model Manager for Governance

“As more predictive models are put into production to address business issues, organizations are challenged to manage the growing model portfolio,” Patel said. “Models should be considered an important enterprise asset, and managed as such – but this is largely a manual process today.”

SAS Model Manager can help you to:

- Register and compare models.
- Promote champion and challenger models.
- Validate models.
- Deploy and put models into production quickly.
- Monitor and track model performance.

"Model decay is another serious challenge," said Patel. "Retaining poorly performing models can lead to inaccurate results and poor business decisions. And for many organizations, good management of analytical models is critical for regulatory compliance. If you're not able to justify why a champion model was chosen or how a particular score was calculated, the organization could face fines, losses or penalties."

SAS Model Manager provides a structured framework for registering, managing, validating and deploying analytical models for timely decisions. Models are stored in a unified model repository. The registration, use and modification of models is supported by a rich metadata structure and project templates, and is documented in audit trails. The platform supports collaborative model development and validation, as well as careful version control.

SAS Model Manager also helps in monitoring the performance of models, so you can address any degradation in a model's value, said Patel. "As the champion model goes through the test phase and production life cycle, an audit trail is created, and this helps in understanding whether the model is performing up to standards, or whether we should retire it or upgrade it before the next iteration of the model life cycle."

> "Models should be considered an important enterprise asset, and managed as such – but this is largely a manual process today."
>
> **Tapan Patel**
> *Global Marketing Manager at SAS*

A Tour of the Data Mining Process

Imagine using SAS data mining products to predict churn for a telecommunications provider. The process entails looking at various models previously created by data miners, testing them for validity, registering the models, selecting champion models (the ones that perform best for predicting and preventing churn) and then putting those models into production to score customer cases.

In this example, the desire is to classify customers based on historical observations of whether or not they're likely to churn. "The approach applies to other business problems as well, such as determining purchase propensity, identifying fraudulent claims and assessing credit risk," explained Wayne Thompson, Manager of SAS Predictive Analytics Product Management.

SAS® Enterprise Miner™ and SAS® Rapid Predictive Modeler

The data set can be a view into a relational database such as Teradata or IBM DB2, a SAS table or even an Excel spreadsheet. This sample data set shows dimensions or inputs that will be used to classify churn, such as complaints, equipment age and contract duration.

SAS® Enterprise Miner™:

- Explores data relationships and anomalies, enriches data, and selects variables.
- Develops predictive and descriptive models using a variety of algorithms.
- Assesses, tests and validates to ensure that the model generalizes well when applied to new data.
- Registers models for deployment against operational systems.

It only takes a few clicks to develop some exploratory plots, such as the distribution of churners relative to nonchurners. You might discover that some variables do not seem to be correlated with churn – but others, such as account age, are related.

You can get summary statistics such as minimum, maximum, mean and standard deviation just by clicking a button. Some of these variables have missing values, which would create problems for some modeling methods, such as regression and neural networks. You might need to consider data imputation to replace missing values.

"Imagine that you have dragged and dropped a modeling table onto the workspace," said Thompson. "The data partition node enables you to split the data into training, validation and test sets. We initially fit a model with the training set, then use the validation data set to help prevent overfitting the training data; and finally, we use the test data to make sure the model scores accurately when applied against data that was not used in the training process. This test step helps ensure that the model will generalize well when presented with new cases it has not seen before."

Drag-and-drop nodes create a decision tree that incorporates a variety of data modeling methods. In the example, data is partitioned into the three data sets. The decision tree will recursively partition the data to separate the churners from nonchurners based on a series of *if-then-else* rules.

"As we go through the decision tree, we see that other variables become important in identifying propensity to churn, such as customer complaints and payment history," said Thompson. "You get your SAS scoring from these models, which can be applied to a database to score new customer cases. It is very easy to do this."

This example was based on a data set with 35,000 observations and 106 variables – a small to midsize data set. You could potentially have thousands of input variables and millions of rows or observations. In that case, you might want to subset the number of variables before you fit another technique. Thompson pointed out that you can use the results from variable clustering to better explain the underlying relationships. Distilling to the most relevant variables helps reduce the processing burden for data sets with a high number of variables.

"The data mining can be multilayered," said Thompson. "For example, you can do a decision tree and then pass that to a neural network, combine models and average them to form a stronger solution. You trial these different algorithms in your bag of modeling tricks, and then use integrated model comparison to determine which ensemble model actually had the best classification for the business problem at hand."

"Data mining models that show good predictive value can be registered as champion models and reused by others," said Thompson. "A business analyst using SAS Rapid Predictive Modeler can execute a prebuilt model. Behind the scenes, SAS Enterprise Miner goes through the SEMMA process, automatically selects the best variables to use and shows how well the model is fitting in different files. An advanced user can delve into the project that created the model.

With SAS® Rapid Predictive Modeler:

- Business analysts can automatically develop a baseline predictive model to support self-service analytics.
- Data miners and statisticians can customize or improve upon the baseline model using SAS Enterprise Miner.

> **"A business analyst using SAS Rapid Predictive Modeler can execute a prebuilt model. Behind the scenes, SAS Enterprise Miner goes through the SEMMA process, automatically selects the best variables to use and shows how well the model is fitting in different files. An advanced user can delve into the project that created the model."**
>
> **Wayne Thompson**
> *Manager of SAS Predictive Analytics Product Management*

In this example, SAS Model Manager helped to:

- Import, test, validate and select a champion model.
- Monitor model performance for degradation.
- Use the model to score a target list of customers to identify those with high lifetime value and high propensity to churn.

A Guided Tour of SAS® Model Manager

SAS Model Manager provides a structured, auditable way to manage the business of using analytical models and applying the results to operational systems.

Models are managed in a hierarchical model repository. If you drill down, you can see who developed the project, the data sets that were involved, the model version and other key pieces of information, such as the variables associated with the project and the SAS modeling code.

A scoring officer or someone responsible for model deployment, for example, often has to field questions about the variables. He may be asked: "Why did you choose these variables? How does this flow work?" Now he can be empowered to answer those questions, without relying on the model development team.

Validation reports compare the cumulative lift of various models that were developed for the business problem at hand. Based on reports such as this and other types of investigation, you can select a champion – a model to put into production.

SAS Model Manager enables you to create a scoring task, which maintains a lot of metadata information, the SAS code, variables, results of scoring, tables used and the model developed. You can take that scoring function and push it out into a number of different channels for consumption by operational applications. Scoring can be performed right in the database, to a relational database or in other SAS tools.

After the model has been in production for a period of time, you'll want to understand how well it is performing. If cumulative lift has declined over time, it indicates that the model has degraded. You can then look at other statistical measures and at the models in the training set to see which input variables might be at play in the model degradation.

A performance-monitoring dashboard provides an at-a-glance view of which models are performing well. For example, good-performing models are shown in green, whereas those that are in decline are shown in yellow or red. You can drill down to see another layer of detail, such as which predictors have slipped.

You can see who the original modeler was, circle back with the developer, and perhaps determine that you need a new model. Then you can start this whole champion and challenge process again. All the functions of this performance monitoring and validation have been tracked, so it's easy to see who did what.

Closing Thoughts

Today, many organizations recognize the value of predictive analytics as an increasingly important source of competitive advantage.

SAS Enterprise Miner streamlines the data mining process to create highly accurate predictive and descriptive models based on analysis of vast amounts of data from across the enterprise. SAS Model Manager helps organize and track the tasks of model registration, validation, deployment and performance monitoring in a production environment. SAS Rapid Predictive Modeler empowers business analysts and subject-matter experts with easy-to-use capabilities for quickly generating their own predictive models, without always relying on quantitative specialists. Find out how these technologies help businesses all around the world achieve their goals.

For More Information

Read why SAS is considered a “powerhouse” in big data analytics:
sas.com/news/preleases/forrester-wave.html

Learn more about SAS solutions:
sas.com/datamining

About SAS

SAS is the leader in business analytics software and services, and the largest independent vendor in the business intelligence market. Through innovative solutions, SAS helps customers at more than 65,000 sites improve performance and deliver value by making better decisions faster. Since 1976 SAS has been giving customers around the world THE POWER TO KNOW®. For more information on SAS® Business Analytics software and services, visit **sas.com**.

SAS Institute Inc. World Headquarters +1 919 677 8000
To contact your local SAS office, please visit: **sas.com/offices**