SAS® Data Mining Roadmap

David Duling
Development Director
R&D

Wayne Thompson PhD
Product Manager
Product Management

Copyright © 2010 SAS Institute Inc. All rights reserved.
Topics

- Current Releases
  - Enterprise Miner 6.2 SAS 9.2
  - SAS Analytics Accelerator 1.3 for Teradata
- Upcoming Releases
  - Enterprise Miner 7.1 SAS 9.3
  - SAS Analytics Accelerator 2.1 for Teradata
- Future Development Projects
  - Enterprise Miner & JMP Pro Integration
  - SAS High Performance Data Mining
  - Tools
  - Applications
- Summary
Enterprise Miner 6.2 SAS 9.2 – August 2010

- SAS Rapid Predictive Modeler
  - Automated data mining for the masses
  - EG and AMO 4.3 task that executes prebuilt Enterprise Miner best practice models
- Early Adopters: Dow Chemical & Deutsche Telekom
- Conference Papers
  - *Automated Modeling at ING: Enhancement Data Set.* ING.
  - *Creating a Model Factory Using In-Database Analytics.* SAS.
  - *Propensity to Buy Factory.* Cisco.
  - *Developing and Managing Analytical Models.* Nedbank.
Enterprise Miner 6.2 SAS 9.2 – continued

- Interactive Decision Tree
  - Validation data
  - Node summaries and statistics
  - All EM customers

- Others features
  - Validvarname=any
  - New statistical transformations

- Credit Scoring for EM 6.2
  - Improved interactive grouping
  - Intercept parameter for scorecard
SAS Analytics Accelerator 1.3 for Teradata

- Procedures and Macros tuned for in-database performance
- Move critical data intensive parts of the algorithm to the database.
- Create SAS brand functions that run inside the database
- Essential tools coverage
  - Exploratory data analysis
  - Basic multivariate statistics
  - Sampling
  - Linear and Logistic Regression
  - Model Scoring
Time Series and Survival Data Mining Are the Most Desired Predictive Modeling Enhancements

Question: What other predictive modeling methods would you like SAS Enterprise Miner to support?

- Time Series Data Mining 71%
- Survival Data Mining 61%
- Bayesian Belief Networks 44%
- Multi Adaptive Regression Analysis 33%
- Generalized Additive Models 32%
- Differential Response Analysis 32%
- Genetic Algorithm 31%
- Random Forests 29%
- Support Vector Machine 28%
- Link Analysis 27%
- Other 8%

n=238 (Respondents who view predictive modeling enhancements as somewhat to very important.)
Enterprise Miner 7.1 SAS 9.3 - July 2011
Enterprise Miner 7.1 for the Desktop
SAS Data Mining for the Classroom

- Updated configuration and deployment
- New tools for application mining
- New and extended algorithms
- Rapid automated modeling
- Other features
- SAS Credit Scoring for Enterprise Miner 7.1 extensions
EM 7.1 Configuration and Deployment

- Standardized configuration for SAS EM, Forecasting, Reporting, OLAP, etc…
  - SAS Analytics Platform is OUT
  - SAS Web Infrastructure Platform is IN
  - Standard Web Server Deployments
  - Standard Web Services Implementation

- Single user DM workstation configuration for academia users and individual contributors
  - Standard Windows Application
  - No extra configuration
  - No extra security
EM 7.1 Projects and Models

- No change to EM Projects data
  - No porting or migration required

- No change to EM Registered models
  - No porting or migration required

- Project Migration code provided for users that also change platforms.
  - For example, Windows XP to Windows 7
  - For example, Windows to Unix
EM 7.1 User Interface

- No major changes to learn
- Many usability *challenges* have been improved
- Focus on speed and large data rows and columns
**EM 7.1 - Survival Mining Node**

- Discrete time to event regression with additive logistic regression.
- The time effect is modeled with cubic splines to allow for flexible shapes of the hazard functions.
- Next the proportional hazard function is fitted with covariates.
- Data Prep to define time interval and time training range
- Scoring with Mean Residual life calculations
EM 7.1 - Insurance Ratemaking Node

- Determination of what premiums to charge for insurance.
- Models types include
  - Claim frequency
  - Severity
  - Pure Premium
- Focused on scalability
  - Bins the predictor variables
  - Analyzes the data in-memory
  - Approximate estimation of complex underlying distributions
- Handles the common over dispersed data condition where most of the claim values are zero.
- Computes relativities which are used by actuaries to derive a rating plan
Credit Scoring for EM 7.1 - Scorecard Enhancements

- Constrained optimized variable binning for automating credit scorecard development
  - Min and Max number of bins
  - Min and Max bin size
  - Min bin response rates
  - Monotonic response curve
- Reverse Scorecard - increasing points with event equal to good
- Adverse Characteristic Reporting – Reason or turn down codes
EM 7.1 SVM node (experimental)

- Estimation Methods
  - Full Dense Quadratic Programming
  - Decomposed Quadratic Programming
  - Lagrangian SVM
  - Least Squares SVM
- Grid and Optimal Search
- Regularization Parameter tuning
- Kernels
  - Polynomial
  - RBF
  - Sigmoid
- Cross validation
- Sampling
- Procedural based scoring

![SVM node interface](image)
SAS Rapid Predictive Modeler

- Great customer feedback
- Scalability improvements
- Automatic data scoring in EG process
- Flexible project naming options
EM 7.1 Decision Trees - Interactive Subtree Selection
EM 7.1 Performance

- Improved speed for Large data sets.
- Procedures optimized for utility file usage.
- SAS code optimized for efficient file usage.
- Group processing optimizations.
- 10 to 50% improvement is expected for many common predictive modeling cases.
- KDD2009 run time now under 27 minutes on my desktop, well under 10 minutes on cheap Windows servers.
  - 15k columns!!
  - 50k rows
  - RPM model flows.
SAS Data Mining Scoring

- New DMSCORE product component added to BASE SAS
- Adds PROC MBSCORE to SAS BASE installs.
- DMRAN function and DMNORM macros also moved to SAS BASE.
- Standardizes EM scoring on BASE SAS engine
  - Previously users could only score EM data step score code but not models that require EM procedures
  - Additional procedures will be added post EM 7.1, e.g. PROC SVMSCORE
- PMML version upgraded to 4.0.
SAS Analytics Accelerator 2.1 for Teradata

1. Optional sampling – Useful for rare target events or client side discovery

2. *Equal height variable binning with WOE (new SAS 9.3)*

3. *Variable selection using the DMINE procedure (new SAS 9.3)*

4. Linear and Logistic Regression modeling

5. Scoring via the Scoring Accelerator for Teradata

Double byte character support
New Application: Time Series Data Mining

The process of discovering and extracting meaningful patterns and relationships from time series data

- Integration of transactional data into a time series to support enriched analyses:
  - Trend analysis and seasonal decomposition
  - Cluster the time series
  - Similarity analysis
  - Incorporating the time series as additional predictors
  - Dimension reduction for subsequent forecasting

- Early adopters:
  - Dow Chemical – dimension reduction and new product forecasting
  - Swiss bank – fraud detection
Data Preparation
  - Provides a tool of aggregation, differencing, summarization, etc.

Exponential Smoothing
  - Fits ESM to interval variables

Similarity
  Computes several similarity measures among time series.
  Similarity measures automatically used for clustering time series.
Topics

- **Current Releases**
  - Enterprise Miner 6.2 SAS 9.2
  - SAS Analytics Accelerator 1.3 for Teradata

- **Upcoming Releases**
  - Enterprise Miner 7.1 SAS 9.3
  - SAS Analytics Accelerator 2.1 for Teradata

- **Future Development Projects**
  - SAS High Performance Data Mining
  - Enterprise Miner & JMP Pro Integration
  - Tools & Algorithms
  - Applications

- **Summary**
SAS High Performance Analytics

- Large Appliance Model
  - Multiple blades, Multi-core / blade, Multi-GB RAM / Blade

- Data Management
  - Standard Relational Data Base storage and tools
  - Fast Parallel Data Read and Write

- In-Memory Analytics
  - Load once, Access many times
  - Eliminates excessive disk I/O
  - Alongside the Data Base model

- Parallel Algorithms
  - Fit the model to the parallel data storage
  - Maximize parallel processing, Minimize inter-process communication
SAS Enterprise Miner and JMP Pro Integration

- JMP Pro is the premier in-memory statistical discovery tool.
- SAS Enterprise Miner is the market leading distributive and collaborative data mining workbench.
- Leveraging the best of both will suit the needs of data mining professionals and workgroups.

Vision - A single unified user experience that supports discovery through production predictive analytics.
SAS Predictive Analytics Core Spectrum

Product Strategy:

- Unified look & feel
- Logical progression with no loss in functionality

SAS Enterprise Miner

- Multi user collaboration
- Bigger and complex data
- Server side execution
- C, Java, PMML, in-database scoring
- Model registration to SAS Metadata

SASEnterprise Miner for Desktop

- Repeatable process flow
- Self-documenting analysis
- Additional methods, including application specific
- Integrated model comparison
- Complete, optimized SAS scoring code
- Automated data mining
- Add-ons
  - Text Mining
  - Credit Scoring

JMP Pro

- Visual discovery & modeling
- Single-user starting point
- In-memory computations
- Fully integrated SAS client
- Extensible with JSL, SAS, R, Excel, Add-ins
Revised Enterprise Miner Pricing & Packaging

- JMP Pro client license(s) added at *no* extra cost to:
  - Enterprise Miner 6.2 SAS 9.2
  - Enterprise Miner 6.2 for Desktop SAS 9.2
  - SAS Data Mining for Classroom (Teaching & Research)

- Separate install for JMP Pro 9

- Existing EM 6.2 customers can request JMP Pro through SAS contracts

- Use of JMP Pro in EM bundle is restricted to the scope of SAS Data Mining

- Customers can always license JMP Pro separately from SAS EM and use JMP Pro entirely separately from EM

- Number of JMP Pro clients is tied to the number of EM personal clients
JMP Pro and SAS Enterprise Miner Roadmap

- SAS EM & JMP Pro Extension Tool Kit: September 2011
- Rate Making Platform: Q3 of 2012
- SAS EM and SAS EM Desktop JMP client with PFD: Dec 2012+
Tool & Algorithms

- Data Preparation tools
- Automated Learning
- Interactive Modeling
  - Neural Networks
  - Link Analysis
- Survival DM extensions
  - Time Varying Covariates
  - Automate cubic spline knot formation
- Time Series Data Mining extensions
  - Dimension reduction using SVD, Wavelet, and Fourier transformation etc.
  - Compute lagged cross-correlation among time series
- R integration
Net Lift Modeling

- Isolate the treatment effect in promotional campaigns.
- Difference Method based on DMREG regression procedure.
- Built in two stage modeling.
Summary

- Rich extensible suite of predictive and descriptive modeling methods
- Focus on applications
- Committed to making data mining more accessible and relevant.
- Ability to analyze complex massive data sources
- Customer driven roadmap