The Power to Know.

SAS® Enterprise Miner 5.2

Unearthing valuable insight—profitable data mining results with less time and effort

What does SAS Enterprise Miner do?

SAS Enterprise Miner streamlines the entire data mining process from data access to model assessment. It supports all necessary tasks within a single, integrated solution while providing the flexibility for efficient collaborations.

Why is SAS Enterprise Miner important?

SAS provides the most powerful, complete data mining solution on the market with unparalleled model development and deployment opportunities. Delivered as a distributed client-server system, it is especially well suited for data mining in large organizations.

For whom is SAS Enterprise Miner designed?

SAS Enterprise Miner is designed for data miners, marketing analysts, database marketers, risk analysts, fraud investigators, business managers, engineers and scientists who play strategic roles in identifying and solving critical business or research issues.

Turning increasing amounts of raw data into useful information remains a challenge for most organizations because the answers that identify key opportunities often lie buried in mountains of data. Which customers will purchase what products and when? Which customers are leaving and what can be done to retain them? How should rates be set to ensure profitability? How are maintenance schedules and operational influences affecting a component's time-to-failure?

To gain an edge in today's competitive market, powerful advanced analytic solutions are required to extract knowledge from vast stores of data. The emerging field of data mining incorporates the process of selecting, exploring and modeling. Discovering previously unknown patterns can deliver actionable strategies for decision makers across your enterprise. For those who choose to implement data mining, the payoffs can be huge.

Unfortunately, data mining can be unwieldy and inefficient. Some analytical tools are limited to a particular set of algorithms, and many do not integrate with software from other niche vendors. Typically, data is scattered across various computer hardware platforms and must be gathered and transformed for the analysis.

The data preparation step is compounded further as the complexities of business problems increase. As a result, quantitative experts spend considerable time accessing and manipulating disparate data sources before beginning to apply their expertise to building the models required to solve business problems.

The demand for actionable analytical information is growing in every industry, putting increased pressure on data miners to produce more and better models in less time. Today's organizations require enterprisewide collaboration on data mining projects and call for powerful, multipurpose solutions that can be tailored to meet different needs.

SAS Enterprise Miner provides an optimized architecture for mining large quantities of data to provide data miners with more time to create highly accurate predictive and descriptive models. Results of the data mining process can be shared easily throughout an organization to deliver actionable analytical information and incorporate models into business processes.

SAS®9

Key benefits

- **A broad set of tools supports the complete data mining process.** Regardless of your data mining needs, SAS provides flexible software that supports all steps necessary to address the complex problems at hand in a single, integrated solution. Going from raw data to accurate, business-driven data mining models becomes a seamless process, enabling the statistical modeling group, business managers and the IT department to collaborate more efficiently.
- **An easy-to-use GUI helps both business analysts and statisticians build more models, faster.** SAS Enterprise Miner's process flow diagram environment eliminates the need for manual coding and dramatically shortens model development time for both business analysts and statisticians. The process flow diagrams also serve as self-documenting templates that can be updated easily or applied to new problems without starting over from scratch. Users can tailor their experience with SAS Enterprise Miner 5.2 via the flexible, interactive display settings. Diagrams can be shared easily with other analysts throughout the enterprise.
- **Makes it easier to surface reliable business information.** SAS Enterprise Miner offers numerous assessment features for comparing results from different modeling techniques. Both statistical and business users share a single, easy-to-interpret view. Model results are shared quickly across the enterprise with the unique model repository system, which links metadata and model management capabilities together in an integrated framework.
- **Provides model deployment and scoring capabilities with unprecedented ease.** Scoring is the process of applying a model to new data and is the end result of many data mining endeavors. SAS Enterprise Miner automates the tedious scoring process and supplies complete scoring code for all stages of model development in SAS, C, Java and PMML. The scoring code can be deployed in a variety of real-time or batch environments within SAS, on the Web or directly in relational databases. The result is faster implementation of data mining results.

Product overview

SAS Enterprise Miner 5.2 is delivered as a modern, distributed client-server system. To streamline the data mining process, this software is designed to work seamlessly with SAS' data integration, advanced analytics and business intelligence capabilities. It also provides a proven model deployment architecture.

Designed around an organized and logical GUI for data mining success

With SAS Enterprise Miner, you get an easy-to-use process flow diagram approach that eliminates the need for manual coding and supports collaborative model development. With the proven, self-guiding (SEMMA) data mining process, both experienced statisticians and less seasoned business analysts can develop more and better predictive analytical models. SEMMA provides a flexible framework for conducting the core tasks of data mining, encompassing five primary steps – sampling, exploration, modification, modeling and assessment. Driven by process flow diagrams that can be modified, saved and shared, SEMMA makes it easy to apply exploratory statistical and visualization techniques, select and transform the most significant variables, create models with those variables to predict outcomes, validate accuracy and deploy these rich insights as decision models for the operational day-to-day business environment.

High-performance grid-enabled workbench

The innovative Java client/SAS server architecture provides unprecedented flexibility for configuring an efficient installation that scales from a single-user system to very large enterprise solutions. Powerful servers may be dedicated to computing, while end users move from office to home to remote sites without losing access to mining projects or services. Many process intensive server tasks such as data sorting, summarization, variable selection and regression modeling are multithreaded, and processes can be run in parallel for distribution across a grid of servers or scheduled for batch processing.

An integrated suite of unmatched modeling techniques

SAS Enterprise Miner provides superior analytical depth with an unmatched suite of advanced predictive and descriptive modeling algorithms, including decision trees, neural networks, memory-based reasoning, clustering, linear and logistic regression, associations, time series and more. Critical preprocessing tasks such as merging data files, addressing missing values, clustering, dropping variables and filtering for outliers are all handled within SAS Enterprise Miner.

Sophisticated set of data preparation, summarization and exploration tools

Preparing data for mining usually is the most time-consuming aspect of data mining endeavors — but not with SAS Enterprise Miner. SAS Enterprise Miner 5.2 includes several tools that make data preparation a fully integrated and efficient part of the data mining process by providing interactive capabilities to explore and transform data for optimal model training. Extensive descriptive summarization features as well as advanced visualization tools enable users to examine quickly and easily large amounts of data in dynamically-linked multidimensional plots that support interactive tasks. The outcome? Quality data mining results that are uniquely and optimally suited to individual business problems.

Business-based model comparisons, reporting and management

Assessment features for comparing models in terms of lift curves and overall profitability allow analysts to easily share and discuss essential results with business users. Models generated from different modeling algorithms can be consistently evaluated across a highly visual user interface. Business domain experts and statisticians alike can compare data mining from a common framework. Model results packages that contain all relevant information of a data mining process flow provide easy model reporting and management. These model result packages are centrally managed through the SAS Metadata Server and can be viewed and queried by data miners, business managers and data managers via a Web-based model repository viewer – the industry's only Web-based system for effectively managing and distributing large model portfolios throughout the organization.

Model scoring and deployment across the enterprise with unprecedented ease

Model deployment is the final and most important phase in which the ROI from the entire mining process is realized. This can be a tedious process and often entails manually writing or converting scoring code, which can delay model implementation and introduce potentially costly mistakes. Scoring code needs to mirror the entire process that led to the final model, including every data preprocessing step. Often, organizations have different environments for model building and model deployment and scoring code must be provided in different languages. SAS Enterprise Miner automatically generates the score code for the entire process flow and supplies complete scoring code in SAS, C, Java and PMML. Production data can be scored within SAS Enterprise Miner or on any other machine, and the scoring code can be deployed in batch, real-time on the Web or directly in relational databases.

Open, extensible design for ultimate flexibility

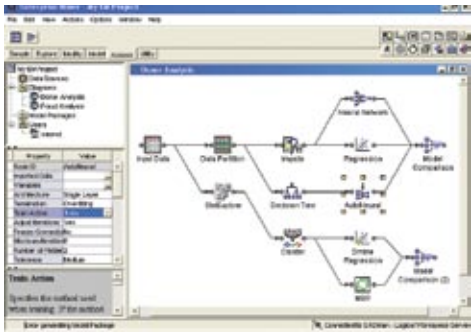
The customizable modeling environment of SAS Enterprise Miner provides the ability to add tools and include personalized SAS code. A Java API is available to embed data mining algorithms into operational business production systems in a fast and easy manner. Default selection lists can be extended with custom developed tools written with SAS code or XML logic, which opens the entire world of SAS to data miners.

An integrated, complete view of all your enterprise data

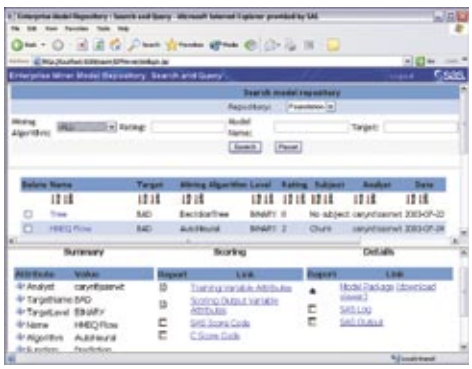
Data mining is most effective when it is part of an integrated information delivery strategy. SAS Enterprise Miner is seamlessly integrated across the SAS Enterprise Intelligence Platform, which provides an end-to-end framework for creating and sharing enterprise intelligence. In addition to creating actionable intelligence from structured data, SAS Text Miner can be easily added to incorporate unstructured textual content into predictive modeling analysis. Together, these technologies offer a synergistic solution that encompass a full spectrum of data analysis and knowledge discovery issues facing the dynamic enterprise of today.

Modern, distributable data mining system suited for large enterprises

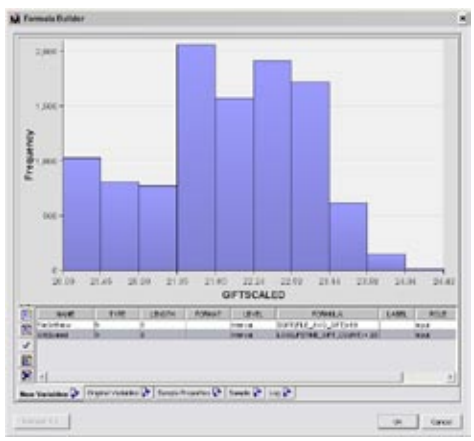
SAS Enterprise Miner is deployable via a thin-client Web portal for easy distribution to multiple users with minimal maintenance of the clients. Alternatively, the complete system can be configured on a stand-alone PC. And, while many data mining applications run only on one or two platforms, SAS Enterprise Miner supports Windows servers and different UNIX platforms, making it the software of choice for organizations with large-scale data mining projects.



Build more models faster with SAS Enterprise Miner's easy-to-use GUI.



Manage models with the Web-based repository viewer.



Easily define variable transformations.

SAS Enterprise Miner 5.2 Key Features

Multiple interfaces

- Easy to use GUI for building process flow diagrams:
  - Build more and better models faster.
  - Web deliverable.
  - Access to SAS programming environment.
  - XML diagram exchange.
  - Reuse diagrams as templates for other projects or users.
- Batch processing:
  - Encapsulates all features of the GUI.
  - SAS macro based.
- Experimental Java API.
- Web-based model repository:
  - Manage large model portfolios.
  - Query models by algorithm, rating, target, etc.
  - Distribute results such as lift charts, tree diagrams and score code to business and data managers.

Scalable processing

- Server-based processing– asynchronous model training. Stop processing cleanly.
- Grid computing:
  - Distribute mining process across a cluster.
  - Schedule training and scoring tasks.
  - Load balancing and resource allocation.
- Parallel processing– run multiple tools and diagrams concurrently.
- Multithreaded predictive algorithms.
- All storage located on servers.

Accessing data

- Access to more than 50 different file structures.
- Integrated with SAS ETL Studio through SAS Metadata Server:
  - Use SAS ETL Studio to define training tables for mining in SAS Enterprise Miner.
  - Use SAS ETL Studio to retrieve and deploy SAS Enterprise Miner scoring code.

Sampling

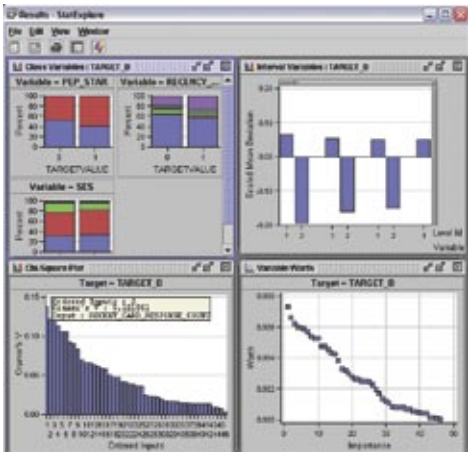
- Simple random.
- Stratified.
- Weighted.
- Cluster.
- Systematic.
- First *N*.
- Rare event sampling.

Data partitioning

- Create training, validation and test data sets.
- Ensure good generalization of your models through use of holdout data.
- Default stratification by the class target.
- Balanced partitioning by any class variable.

Filtering outliers

- Apply various distributional thresholds to eliminate extreme interval values.
- Combine class values with fewer than n occurrences.
- Interactively filter class and numeric values.



SAS Enterprise Miner provides flexible data visualizations.

Develop customized interactive plots with the graphics wizard.

Segment your data using clustering or self-organizing maps.

Transformations

- Simple: log, square root, inverse, square, exponential, standardized.
- Binning: bucketed, quantile, optimal binning for relationship to target.
- Best power: maximize normality, maximize correlation with target, equalize spread with target levels.
- Interactions editor: define polynomial and nth degree interaction effects.
- Interactively define transformations:
  - Define customized transformations using the expression builder.
  - Compare the distribution of the new variable with the original variable.

Data replacement

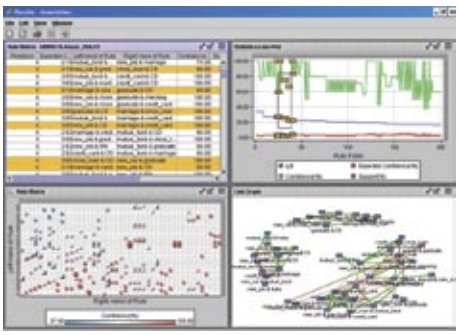
- Measures of centrality.
- Distribution-based.
- Tree imputation with surrogates.
- Mid-medium spacing.
- Robust M-estimators.
- Default constant.
- Replacement Editor:
  - Specify new values for class variables.
  - Assign replacement values for unknown values.

Descriptive statistics

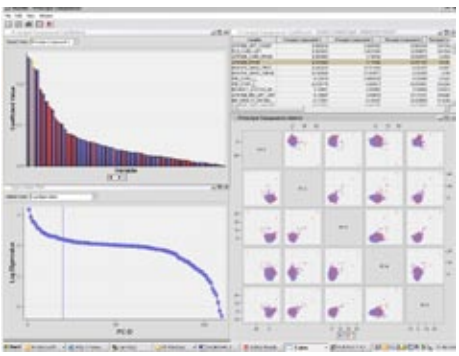
- Univariate statistics and plots:
  - Interval variables – *n*, mean, median, min, max, standard deviation, scaled deviation and percent missing.
  - Class variables – number of categories, counts, mode, percent mode, percent missing.
  - Distribution plots.
  - Statistics breakdown for each level of the class target.
- Bivariate statistics and plots:
  - Ordered Pearson and Spearman correlation plot.
  - Ordered chi-square plot with option for binning continuous inputs into *n* bins.
  - Coefficient of variation plot.
- Variable selection by logworth.
- Other interactive plots:
  - Variable worth plot ranking inputs based on their worth with the target.
  - Class variable distributions across the target and/or the segment variable.
- Scaled mean deviation plots.

Graphs/visualization

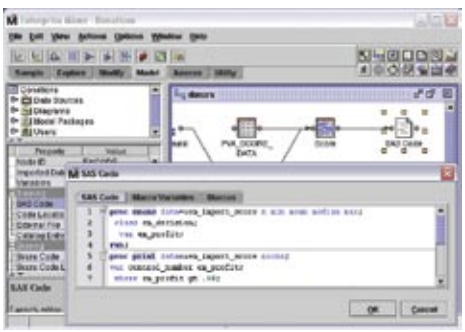
- Batch and interactive plots: scatter plots, scatter plot matrix plots, lattice plots, 3D charts, density plots, histograms, multidimensional plots, pie charts and areabar charts.
- Segment profile plots:
  - Interactively profile segments of data created by clustering and modeling tools.
  - Easily identify variables that determine the profiles and the differences between groups.
- Easy-to-use graphics wizard:
  - Titles and footnotes.
  - Apply a WHERE clause.
  - Choose from several color schemes.
  - Easily rescale axes.
  - Surface the underlying data from standard SAS Enterprise Miner results to develop customized graphics.
- Plots and tables are interactively linked supporting tasks such as brushing and banding.
- Copy and paste data and plots easily into other applications or save as BMP files.



View market basket profiles.



Perform preliminary variable selection.



Integrate customized SAS code.

Clustering and self-organizing maps

- Clustering:
  - User defined or automatically chooses the best *k* clusters.
  - Several strategies for encoding class variables into the analysis.
  - Handles missing values.
  - Variable segment profile plots showing the distribution of the inputs and other factors within each cluster.
  - Decision tree profile using the inputs to predict cluster membership.
  - PMML score code.
- Self-organizing maps:
  - Batch SOMs with Nadaraya-Watson or local-linear smoothing.
  - Kohonen networks.
  - Overlay the distribution of other variables onto the map.
  - Handles missing values.

Market basket analysis

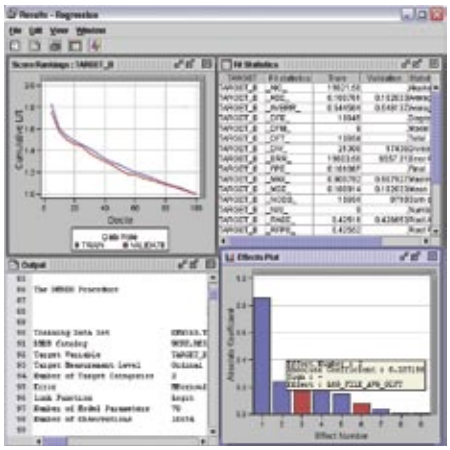
- Associations and sequence discovery:
  - Grid plot of the rules ordered by confidence.
  - Statistics line plot of the lift, confidence, expected confidence, and support for the rules.
  - Statistics histogram of the frequency counts for given ranges of support and confidence.
  - Expected confidence versus confidence scatter plot.
  - Rules description table.
  - Network plot of the rules.
- Interactively subset the rules based on lift, confidence, support, chain length, etc.
- Seamless integration of the rules with other inputs for enriched predictive modeling.
- Output rules easily for clustering customers by their purchase behavior.
- PMML score code.

Web path analysis

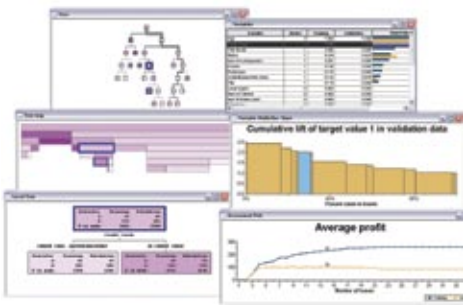
- Scalable and efficient mining of the most frequently navigated paths from clickstream data.
- Mine frequent consecutive subsequences from any type of sequence data.

Dimension reduction

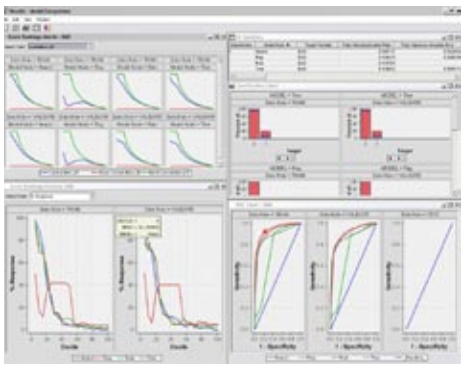
- Variable selection:
  - Remove variables unrelated to target based on a chi-square or R2 selection criterion.
  - Remove variables in hierarchies.
  - Remove variables with many missing values.
  - Reduce class variables with large number of levels.
  - Bin continuous inputs to identify nonlinear relationships.
  - Detect interactions.
- Principal components:
  - Calculate Eigenvalues and Eigenvectors from correlation and covariance matrices.
  - Plots include: principal components coefficients, principal components matrix, Eigenvalue, Log Eigenvalue, Cumulative Proportional Eigenvalue.
  - Interactively choose the number of components to be retained.
  - Mine the selected principal components using predictive modeling techniques.
- Time series mining:
  - Reduce transactional data into a times series using several accumulation methods and transformations.
  - Analysis methods include seasonal, trend, time domain, seasonal decomposition.
  - Mine the reduced time series using clustering and predictive modeling techniques.
- Manage time metrics with descriptive data.



Develop regression models.



Create interactive decision trees.



Evaluate multiple models.

SAS Code node

- Write SAS code for easy to complex data preparation and transformation tasks.
- Incorporate procedures from other SAS products.
- Import external models.
- Develop custom models and SAS Enterprise Miner nodes.
- Includes macro variables to easily reference data sources, variables, etc.
- Augment score code logic.

Consistent modeling features

- Select models based on either the training, validation (default) or test data using several criterion such as: profit or loss, AIC, SBC, average square error, misclassification rate, ROC, Gini, KS (Kolmogorov-Smirnov).
- Incorporate prior probabilities into the model development process.
- Supports binary, nominal, ordinal and interval inputs and targets.
- Easy access to score code and all partitioned data sources.
- Display multiple results in one window to help better evaluate model performance.

Regression

- Linear and logistic.
- Stepwise, forward and backward selection.
- Equation terms builder: polynomials, general interactions, effect hierarchy support.
- Cross validation.
- Effect hierarchy rules.
- Optimization techniques include: Conjugate Gradient, Double Dogleg, Newton-Raphson with Line Search or Ridging, Quasi-Newton, Trust Region.
- PMML score code.

Decision trees

- General methodology:
  - CHAID, classification and regression trees, C 4.5.
  - Tree selection based on profit or lift objectives and prune accordingly.
- Splitting criterion: Prob Chi-square test, Prob F-test, Gini, Entropy, variance reduction.
- Automatically output leaf IDs as inputs for subsequent modeling.
- Displays English rules.
- Calculates variable importance for preliminary variable selection.
- Unique consolidated tree map representation of the tree diagram.
- Interactive tree desktop application:
  - Interactive growing/pruning of trees; expand/collapse tree nodes.
  - Define customized split points including binary or multi-way splits.
  - Split on any candidate variable.
  - More than 13 tables and plots are dynamically linked to better evaluate the tree performance.
  - Easy to print the tree diagram on a single page or across multiple pages.
- Based on the fast underlying ARBORETUM procedure.

Neural networks

- Neural Network node:
  - Flexible network architectures with extensive combination and activation functions.
  - 10 training techniques.
  - Preliminary optimization.
  - Automatic standardization of inputs.
  - Supports direction connections.
- Autoneural Neural node:
  - Automated multilayer perceptron building to search for the optimal configuration.
  - Type and activation function selected from four different types of architectures.

SAS® Enterprise Miner Technical Requirements

Client environment

- AIX; HP/UX-IPF; Solaris; Linux for Intel (x86 32-bit): Red Hat Linux 8.0, RHAS 2.1, RHEL 3.0, SuSE SLES 8, SLES 9; Windows (x86-32): Windows XP

Server environment

- AIX (64-bit), Release 5.1+
- HP/UX (64-bit), Release 11i+
- HP/UX Itanium (64-bit), Release 11i+
- Linux for Intel (x86 32-bit): Red Hat Linux 8.0, RHAS 2.1, RHEL 3.0, SuSE SLES 8
- Linux for Itanium (64-bit): Red Hat RHEL 3.0
- Solaris (64-bit) 8, 9, 10 on SPARC
- Tru64 UNIX (64-bit), Version 5.1A or 5.1B
- Windows (x86-32): Windows NT 4 Server, Windows 2000 Server, Windows Server 2003
- Windows (64-bit on Itanium): Windows Server 2003

Enterprise Model Repository Viewer (optional Web tier configuration)

SAS includes a reference implementation of Apache Tomcat. Sites can optionally choose to license another Web server or WebDAV component directly from the vendor.

Required software

Base SAS and SAS/STAT

- PMML score code.
- DM Neural node:
  - Model building with dimension reduction and function selection.
  - Fast training; linear and nonlinear estimation.

Rule induction

- Recursive predictive modeling technique.
- Especially useful for modeling rare events.

Two-stage modeling

- Sequential and concurrent modeling for both the class and interval target.
- Choose a decision tree, regression, or neural network model for each stage.
- Control how the class prediction is applied to the interval prediction.
- Accurately estimate customer value.

Memory-based reasoning

- *k*-nearest neighbor technique to categorize or predict observations.
- Patented Reduced Dimensionality Tree and Scan.

Model ensembles

- Combine model predictions to form a potentially stronger solution.
- Methods include: Averaging, Voting, Maximum.

Model comparison

- Compare multiple models in a single framework for all holdout data sources.
- Automatically selects the best model based on the user defined model criterion.
- Extensive fit and diagnostics statistics.
- Lift charts; ROC curves.
- Profit and loss charts with decision selection; Confusion (classification) matrix.
- Class probability score distribution plot; Score ranking matrix plots.
- Interval target score rankings and distributions.

Scoring

- Score node for interactive scoring in the GUI.
- Automated score code generation in SAS, C, Java and PMML.
- SAS, C and Java scoring code capture modeling, clustering, transformations and missing value imputation code.
- Deploy models in multiple environments.

Utility tools

- Drop variables node.
- Merge data node.
- Metadata node for modifying columns metadata such as role, measurement level and order.

World Headquarters
and SAS Americas
SAS Campus Drive
Cary, NC 27513 USA
Tel: (1) 919 677 8000
Fax: (1) 919 677 4444
U.S. & Canada sales:
(1) 800 727 0025

SAS International
PO Box 10 53 40
Neuenheimer Landsr. 28-30
D-69043 Heidelberg, Germany
Tel: (49) 6221 4160
Fax: (49) 6221 474850

www.sas.com

101369US_358921.1105