



## SAS® Enterprise Miner™ for Desktop 6.1

A fast, powerful data mining workbench delivered to your desktop

### What does SAS® Enterprise Miner™ for Desktop do?

SAS Enterprise Miner for Desktop is a complete data mining workbench that runs entirely within the confines of a Windows PC. It uses a wide variety of descriptive and predictive techniques to give you the insight to make profitable decisions.

### Why is SAS® Enterprise Miner™ for Desktop important?

Business leaders and researchers from organizations of all sizes are turning to predictive analytics to gain an unbeatable advantage in today's dynamic marketplace. SAS Enterprise Miner for Desktop delivers a comprehensive set of data mining capabilities to address a wide variety of business problems.

### For whom is SAS® Enterprise Miner™ for Desktop designed?

It is designed for quantitative analysts in small to medium-sized organizations, or those who work independently in departments faced with solving critical business issues or complex research problems.

To gain an edge in today's competitive market, powerful advanced analytic solutions are required to extract knowledge from vast stores of data and act on it. More and more organizations are turning to predictive analytics and data mining software to uncover patterns in data and discover hidden relationships.

Predictive analytics go beyond reporting on what *has* happened to discovering *why* it has happened and what is likely to happen next. Such insights empower decision makers to design innovative strategies not considered before, providing the potential for huge payoffs before your competitors can act. Data mining success stories have shown how to:

- Identify the most profitable customer relationship management (CRM) strategies.
- Set profitable credit or insurance rates by understanding how and why before making adjustments, unlike the traditional "black-box" approach.
- Better detect and deter fraudulent transactions and identify areas of potential high risk.
- Measure how maintenance schedules and operational processes affect manufacturing processes.
- Engage in pharmacovigilance, the process of evaluating and improving the safety of marketed medicines.

SAS Enterprise Miner for Desktop delivers the power of data mining right to your desktop PC. Highly evolved data classification, analysis and interpretation methods make it easy for business analysts and researchers to collaborate on projects essential to gaining a competitive advantage.

### Key benefits

- **An easy-to-use GUI helps both business analysts and statisticians build better models, faster.** An interactive process flow diagram environment eliminates the need for manual coding and dramatically shortens model development time. SAS' exclusive SEMMA (sample, explore, model, modify and assess) data mining approach combines a structured process with the logical organization of data mining tools so both experienced statisticians and less-seasoned business analysts can develop better predictive analytical models and start using results faster.
- **A rich set of data mining tools are delivered to your desktop, enabling you to spot trends and recognize business opportunities.** SAS Enterprise Miner for Desktop provides advanced predictive and descriptive modeling tools as well as numerous assessment features for comparing results from different modeling techniques. Decision trees, neural networks, clustering and associations, combined with advanced regression and statistical routines, deliver models with improved accuracy. Interactive options offer users highly flexible visualization tools for discovering patterns as well as controls for modifying the displays or resulting graphs. The result is improved performance and a higher ROI on all of your data assets.
- **Quick and painless installation that you can do yourself.** Software media delivered via electronic software download allows PC users to get up and running fast. You can start mining your data right away and make decisions that improve your business.



---

## Product overview

---

With SAS Enterprise Miner for Desktop, it is now possible for individual users to begin to realize the benefits they can reap from an interactive data mining workbench on their PC.

Data mining projects are set up and managed within a visual workspace. Users build their own process flow diagrams, add analysis nodes, compare models and generate SAS score code. Diagrams can be saved as XML files for reuse. Interactive statistical and visualization tools help data miners spot trends and anomalies quickly, so they can focus their energies on developing better models rather than being bogged down with formatting output reports or documenting previous attempts that resulted in the current model.

The SAS Code node enables users to integrate SAS DATA step processing and procedures into a SAS Enterprise Miner for Desktop process flow diagram.

---

## An organized and logical GUI for data mining success

---

SAS Enterprise Miner for Desktop provides a flexible framework for conducting all phases of data mining using the SEMMA approach. Using drag-and-drop tools, process flow diagrams are created, updated and easily modified for the next analytical study. These saved visual diagrams can be referred to later when communicating the various analytical investigations conducted during the study.

The interactive interface guides users as they:

- Apply statistical and visualization techniques to see and become familiar with the data quality and trends.

- Explore and transform the data to identify a candidate set of predictor variables.
- Create models with those variables to predict outcomes. Novice data miners can build initial models quickly with default settings, while more experienced users can tweak settings to specify unique parameters to enhance their models.
- Combine modeling techniques for additional accuracy.
- Compare models and try multiple approaches and options. Easy-to-interpret displays help users communicate why a particular model is the best predictor.
- Validate the accuracy of decision models with new data before deploying results into the operational, day-to-day business environment.
- Apply the champion model against new data using automatically generated, complete score code.

The interactive, easy-to-use drag-and-drop process flow diagram approach shortens the model development time for both experienced statisticians and business domain experts.

The process flow diagrams also serve as self-documenting templates that can be updated later or applied to new problems without starting over from scratch.

---

## Data preparation, summarization and exploration tools provide quality results suited to individual problems

---

Preparing data is the most time-consuming aspect of data mining endeavors. SAS Enterprise Miner for Desktop combines powerful data mining capabilities with data exploration

and data preparation features, making it easy to read in data from files other than SAS as a fully integrated part of the data mining process.

Extensive, descriptive summarization features and advanced visualization tools enable users to examine large amounts of data in dynamically linked, multidimensional plots that support interactive exploration tasks.

Critical preprocessing tasks include merging files, appending data, sampling, choosing appropriate methods for handling incomplete entries and missing values, binning variables, clustering observations, dropping variables and filtering outliers. Bad data is bad business, and only by starting with quality inputs (careful cleansing of data) can you expect to get quality results.

---

## An integrated suite of unmatched modeling techniques

---

SAS Enterprise Miner for Desktop provides sophisticated analytical depth for the PC user with a suite of advanced predictive and descriptive modeling algorithms, including decision trees, gradient boosting, neural networks, clustering, linear and logistic regression, associations and more. You can take previously created SAS/STAT® software models and incorporate them into the SAS Enterprise Miner for Desktop environment for even further fine-tuning and integrated model comparisons.

---

## Model comparisons, reporting and management

---

SAS Enterprise Miner for Desktop offers numerous assessment tools for comparing results from different modeling techniques. Results are presented in both statistical and business terms

within a single, easy-to-interpret framework. Models generated from different modeling algorithms can be consistently evaluated across a highly visual, interactive user interface.

### Scoring with unprecedented ease

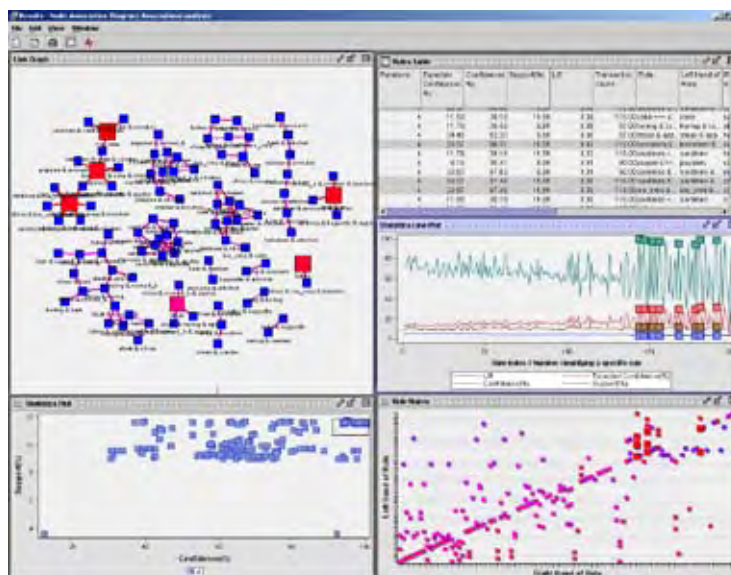
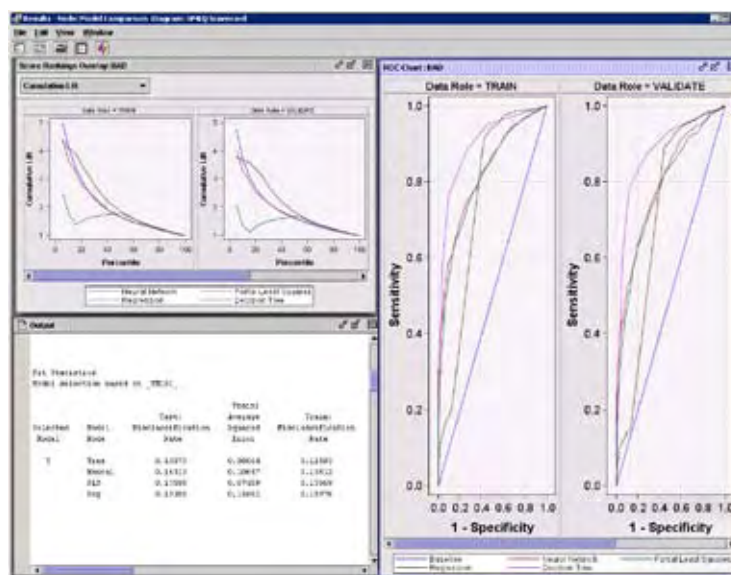
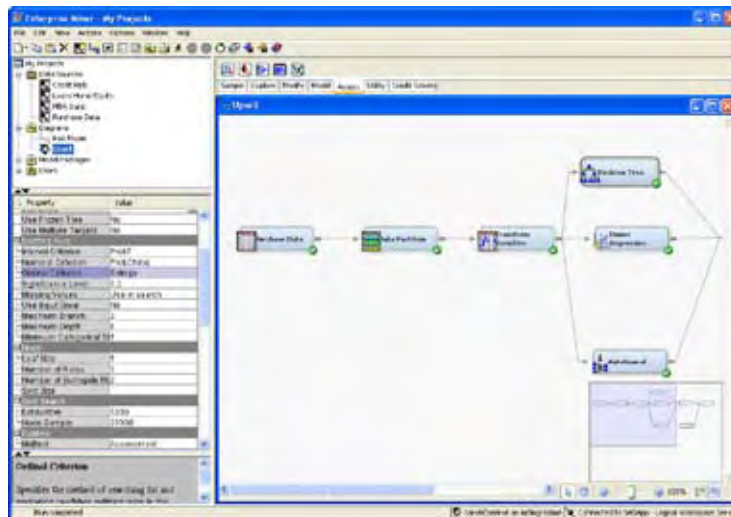
The final and most important phase in data mining projects occurs when new data is “scored” — when the model scoring code is applied to new data, a predicted outcome is produced and appropriate decisions are identified for action.

Once the data mining models are developed, SAS Enterprise Miner for Desktop allows you to export SAS score code for rapid deployment into your operational environment with a single click. Manual conversion of scoring code not only causes delays for model implementation, it also can introduce potentially costly mistakes. Unless the entire process that led to the final model is mirrored in the score code (including all data preprocessing steps), the real-world application will miss the mark. SAS Enterprise Miner for Desktop automatically generates score code for the entire process flow and supplies this scoring code in SAS.

**Top screen (Figure 1):** With SAS Enterprise Miner for Desktop’s GUI, you can quickly develop process flow diagrams that are self-documenting and can be updated easily or applied to new problems.

**Center screen (Figure 2):** Evaluate multiple models together in one easy-to-interpret framework using the Model Comparison node.

**Bottom screen (Figure 3):** View market basket profiles. Interactively subset the rules based on lift, confidence, support, chain length, etc.



## Key Features

### Multiple interfaces

- Easy-to-use GUI for building process flow diagrams:
  - Build more and better models faster.
  - Access the SAS programming environment.
  - Provides XML diagram exchange.
  - Reuse diagrams as templates for other projects or users.
- Batch processing:
  - Encapsulates all features of the GUI.
  - SAS macro-based.
  - Embed training and scoring processes into customized applications.

### Accessing and managing data

- Access to more than 50 file structures.
- File Import node for easy access to Microsoft Excel, comma-delimited files, SAS, JMP® and other common file formats.
- Enhanced Explorer window to quickly locate and view table listings or to develop a plot using interactive graph components.
- SAS Library Explorer and Library Assignment wizard.
- Drop variables node.
- Merge data node.
- Append node.
- Filter outliers:
  - Apply various distributional thresholds to eliminate extreme interval values.
  - Combine class values with fewer than  $n$  occurrences.
  - Interactively filter class and numeric values.
- Metadata node for modifying columns metadata such as role, measurement level and order.

### 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.

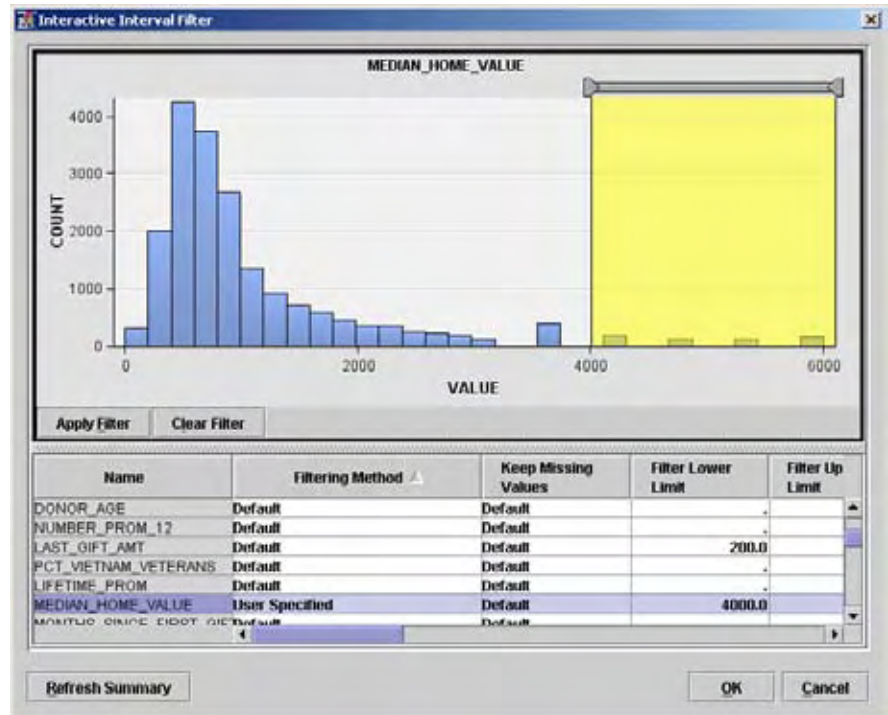


Figure 4: Filter extreme values interactively with the Filter node. The shaded region defines the variable range to keep.

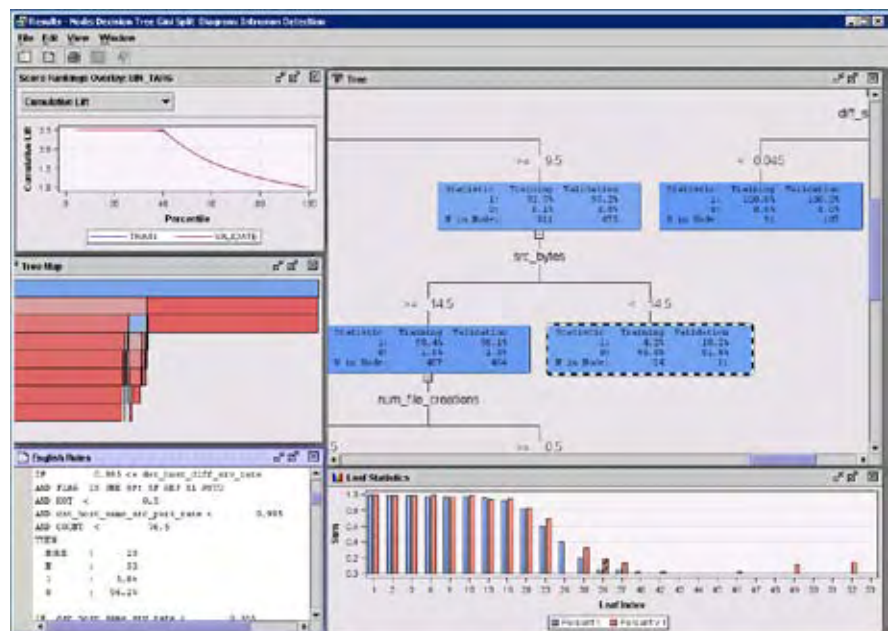


Figure 5: Develop decision trees interactively or in batch. Numerous assessment plots to help gauge overall tree stability are included.

- Default stratification by the class target.
- Balanced partitioning by any class variable.
- Output SAS tables or views.

### **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  $n$ th degree interaction effects.
- Interactively define transformations:
  - Define customized transformations using the Expression Builder or SAS Code editor.
  - Compare the distribution of the new variable with the original variable.
- Predefine global transformation code for reuse.

### **Interactive variable binning**

- Quantile or bucket.
- Gini variable selection.
- Handle missing values as separate group.
- Fine and coarse classing detail.
- Profile bins by target.
- Modify groups interactively.
- Save binning definitions.

### **Rules Builder node**

- Create ad hoc, data-driven rules and policies.
- Interactively define the value of the outcome variable and paths to the outcome.

### **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.
  - Interactively cap extreme interval values to a replacement threshold.

### **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, matrix plots, box plots, constellation plots, contour plots, needle plots, lattice plots, 3-D charts, density plots, histograms, multidimensional plots, pie charts and area bar 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 Explorer wizard and Graph Explore node:
  - Create titles and footnotes.
  - Apply a WHERE clause.
  - Choose from several color schemes.
  - Easily rescale axes.
  - Surface underlying data from standard SAS Enterprise Miner for Desktop results to develop customized graphics.
- Plots and tables are interactively linked, supporting tasks such as brushing and banding.
- Data and plots can be easily copied and pasted into other applications or saved as BMP files.
- Interactive graphs are automatically saved in the Results window of the node.

### **Clustering and self-organizing maps**

- Clustering:
  - User defined or automatically chooses the best clusters.
  - Several strategies for encoding class variables into the analysis.
  - Handles missing values.
  - Variable segment profile plots show the distribution of the inputs and other factors within each cluster.
  - Decision tree profile uses the inputs to predict cluster membership.
- 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 rules based on lift, confidence, support, chain length, etc.
- Seamless integration of rules with other inputs for enriched predictive modeling.
- Hierarchical associations:
  - Derive rules at multiple levels.
  - Specify parent and child mappings for the dimensional input table.

### **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.
- LARS (Least Angle Regression) variable selection:
  - AIC, SBC, Mallows C(p), cross validation and other selection criterion.
  - Plots include: parameter estimates, coefficient paths, iteration plot, score rankings and more.
  - Generalizes to support LASSO (least absolute shrinkage and selection operator).
  - Supports class inputs.
  - Score code generation.
- 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.
- Variable clustering:
  - Divide variables into disjoint or hierarchical clusters.
  - Eigenvalue or Principal Components learning.
  - Includes class variable support.
  - Dendrogram tree of the clusters.
  - Selected variables table with cluster and correlation statistics.
  - Cluster network and R-Square plot.
  - Interactive user override of selected variables.
- 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.

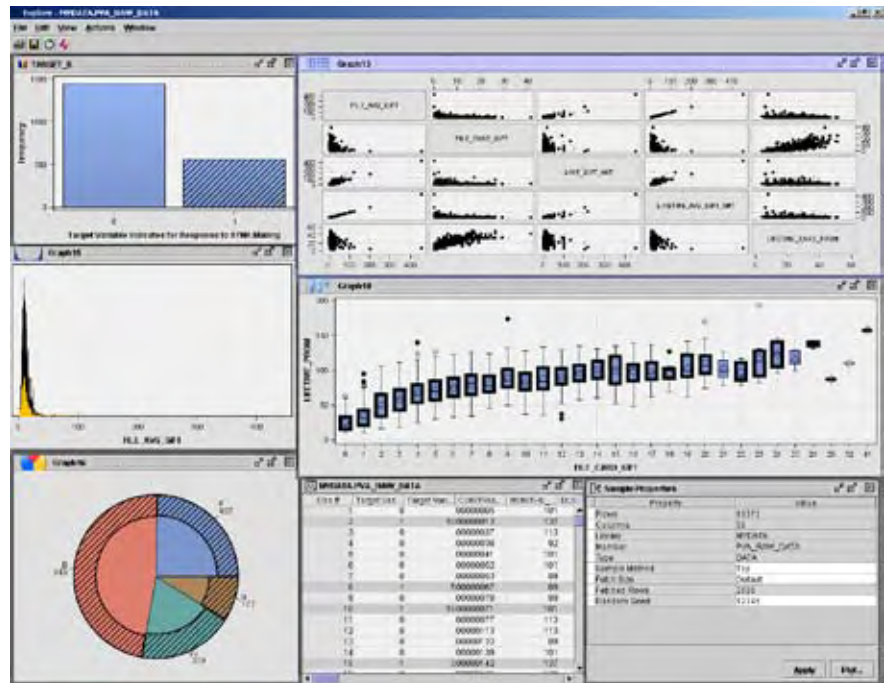


Figure 6: Explore your data interactively with parallel axis, density, 3-D rotating scatter plots and other plots. Interactive graphs are automatically saved within the results of any node.

- Manage time metrics with descriptive data.

#### SAS Code node

- Write SAS code for easy-to-complex data preparation and transformation tasks.
- Incorporate procedures from other SAS products.
- Develop custom models.
- Create SAS Enterprise Miner for Desktop extension nodes.
- Augment score code logic.
- Easy-to-use program development interface:
  - Macro variables to reference data sources, variables, etc.
  - Interactive code editor and submit.
  - Separately manage training, scoring and reporting code.
  - SAS Output and SAS LOG.
- Create graphics.

#### Consistent modeling features

- Select models based on either the training, validation (default) or test data using several criteria 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.
- Decisions node for setting target event and defining priors and profit/loss matrices.

#### 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.
- Dmine Regression node:
  - Fast forward stepwise least squares regression.
  - Optional variable binning to detect nonlinear relationships.
  - Optional class variable reduction.
- Include interaction terms.

### Decision trees

- Methodologies:
  - CHAID, classification and regression trees, bagging and boosting, gradient boosting.
  - Tree selection based on profit or lift objectives and prune accordingly.
  - K-fold cross validation.
- Splitting criterion: Prob Chi-square test, Prob F-test, Gini, Entropy, variance reduction.
- Switch targets for designing multi-objective segmentation strategies.
- Automatically output leaf IDs as inputs for modeling and group processing.
- Displays English rules.
- Calculates variable importance for preliminary variable selection and model interpretation.
- 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 multiway splits.
  - Split on any candidate variable.
  - Copy split.
  - More than 13 tables and plots are dynamically linked to better evaluate the tree performance.
  - Easy to print tree diagrams on a single page or across multiple pages.
- Based on the fast ARBORETUM procedure.

### Neural networks

- Neural Network node:
  - Flexible network architectures with combination and activation functions.
  - 10 training techniques.
  - Preliminary optimization.
  - Automatic standardization of inputs.
  - Supports direction connections.

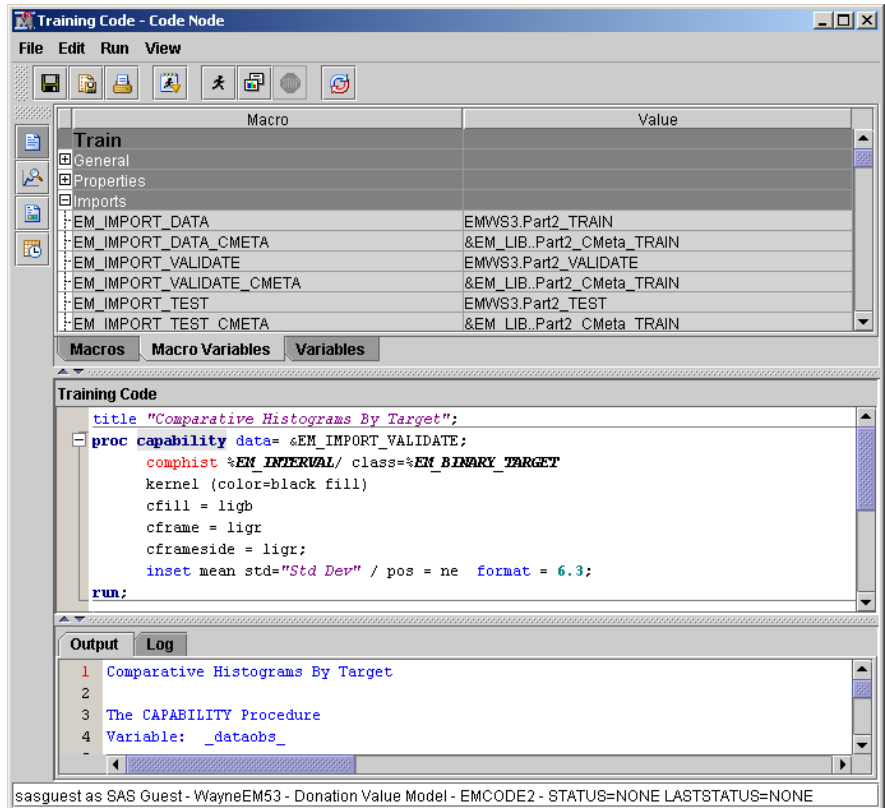


Figure 7: Integrate customized SAS code to create variable transformations, incorporate SAS procedures, develop new nodes, augment scoring logic, tailor reports and more.

- Autoneural Neural node:
  - Automated multilayer perceptron building searches for optimal configuration.
  - Type and activation function selected from four different types of architectures.
- DM Neural node:
  - Model building with dimension reduction and function selection.
  - Fast training; linear and nonlinear estimation.

### Partial Least Squares node

- Especially useful for extracting factors from a large number of potential correlated variables.
- Also performs principal components regression and reduced rank regression.
- User or automatic selection of the number of the factors.
- Choose from five cross-validation strategies.
- Supports variable selection.

### 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 and Maximum.

### Group processing with the Start and End Groups nodes

- Repeat processing over a segment of the process flow diagram.
- Use cases: stratified modeling, bagging and boosting, multiple targets, cross validation.

### Model Import node

- Register SAS Enterprise Miner for Desktop models for reuse in other diagrams and projects.
- Import and evaluate external models.

### Model evaluation

- Model Comparison node to compare multiple models in a single framework for all holdout data sources.
- Automatically selects the best model based on the user-defined model criterion.
- Supports user override.
- 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.
- Cutoff node to determine probability cutoff point(s) for binary targets.
- User override for default selection.
- Max KS Statistic.
- Min Misclassification Cost.
- Maximum Cumulative Profile.
- Max True Positive Rate.
- Max Event Precision from Training Prior.
- Event Precision Equal Recall.

### Reporter node

- Uses SAS Output Delivery System to create a PDF or RTF document of a process flow.
- Helps document the analysis process and facilitate results sharing.
- Document can be saved and is included in the SAS Enterprise Miner for Desktop Results Packages.
- Includes image of the process flow diagram.
- User-defined notes entry.

### Scoring

- Score node for interactive scoring in the SAS Enterprise Miner for Desktop GUI.
- Score node creates optimized score code by default, eliminating unused variables.
- Automated score code generation in SAS.
- SAS scoring code captures modeling, clustering, transformations and missing value imputation code.

## SAS® Enterprise Miner™ for Desktop 6.1 Technical Requirements

### Client environment

- Microsoft Windows (x86-32):  
Windows XP Professional, Windows Server 2003, Windows Vista\*
- Microsoft Windows on x64 (EM64T/AMD64): Windows XP Professional for x64, Windows Vista\* for x64, Windows Server 2003 for x64

\* NOTE: Windows Vista Editions that are supported include Enterprise, Business and Ultimate.

### Required software

- Base SAS and SAS/STAT®