



SAS® DESKTOP DATA MINING FOR MIDSIZE BUSINESS

A fast, powerful data mining workbench designed for small to midsize organizations

What does SAS® Desktop Data Mining for Midsize Business do?

SAS Desktop Data Mining for Midsize Business is a complete data mining workbench that runs entirely within the confines of a Windows PC. The interactive GUI makes it easy to discover patterns and relationships from a comprehensive ensemble of modeling algorithms.

Why is SAS® Desktop Data Mining for Midsize Business important?

Business leaders and researchers from small to midsize organizations are turning to predictive analytics to gain an unbeatable advantage in today's dynamic marketplace. SAS Desktop Data Mining for Midsize Business delivers a comprehensive set of features necessary for performing data mining.

For whom is SAS® Desktop Data Mining for Midsize Business designed?

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



**THE
POWER
TO KNOW®**

To gain an edge in today's competitive market, powerful analytic solutions are required to extract knowledge from vast stores of data. 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 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. Finding these patterns and hidden relationships before your competitors do is key. Data mining success stories have shown how to:

- Identify the most profitable Customer Relationship Management (CRM) strategies.
- Set profitable credit or promotions by understanding how and why before making adjustments.
- Better detect and deter fraudulent transactions and identify areas of potential high risk.
- Improve the effectiveness of all channels of targeted consumer marketing.
- Predict the success of collection strategies and manage account receivables.

SAS Desktop Data Mining for Midsize Business 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 discover patterns and relationships essential to gaining a competitive advantage.

Key benefits

- **An easy-to-use GUI helps both business analysts and statisticians build more 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 more and better predictive analytical models.
- **An affordable set of data mining tools are delivered to your desktop, enabling you to spot trends and recognize business opportunities.** This data mining workbench for the desktop PC 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.
- **Quick and painless installation that you can do yourself.** A DVD or CD install to the Windows environment allows PC users to get up and running fast. You can start mining your data right away.

Product overview

With SAS Desktop Data Mining for Midsize Business, it is now possible for smaller organizations to begin to realize the benefits that can be reaped from a process-based data mining solution.

Data mining projects are set up and managed within a visual workspace. Users lay out 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 users 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.

Sophisticated analysis capabilities are presented in an intuitive layout of icons (drag-and-drop nodes) that include all the common elements of a data mining process. Each node for data transformation, advanced descriptive and predictive analysis, model assessment and deployment has a uniform look and feel that makes it easy to learn and manipulate.

An organized and logical GUI for data mining success

SAS Desktop Mining for Midsize Business provides a flexible framework for conducting all phases of data mining using the SAS SEMMA approach. Using drag-and-drop icons, 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. Data quality assessments are conducted and automated suggestions are provided to correct for incomplete entries or data errors.
- Explore and transform the data to identify the most significant variables.
- Create models with those variables to predict outcomes. Novice users 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.

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 Desktop Data Mining for Midsize Business 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, choosing appropriate methods for handling incomplete entries and missing values, grouping, clustering and dropping variables, and filtering for 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 Desktop Data Mining for Midsize Business provides an affordable option for delivering analytical depth to the PC with a suite of advanced predictive and descriptive modeling algorithms, including decision trees, neural networks, clustering, linear and logistic regression, associations, and more.

Model comparisons, reporting and management

SAS Desktop Data Mining for Midsize Business 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 new data is put into the model, a predicted outcome is produced and appropriate decisions are identified for action. Once the data mining models are developed, SAS Desktop Data Mining for Midsize Business allows you to export 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 Desktop Data Mining for Midsize Business automatically generates score code for the entire process flow and supplies this scoring code in SAS.

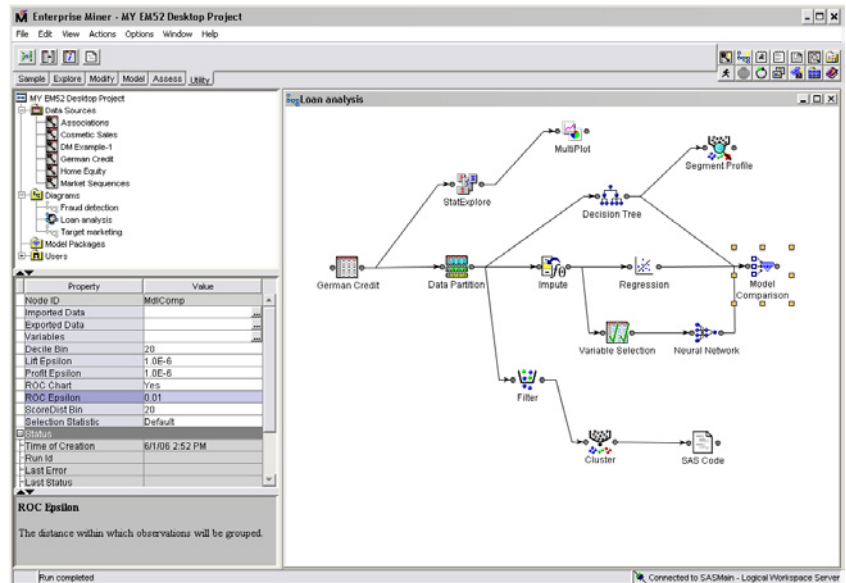
SAS® Desktop Data Mining for Midsize Business technical requirements

Client environment

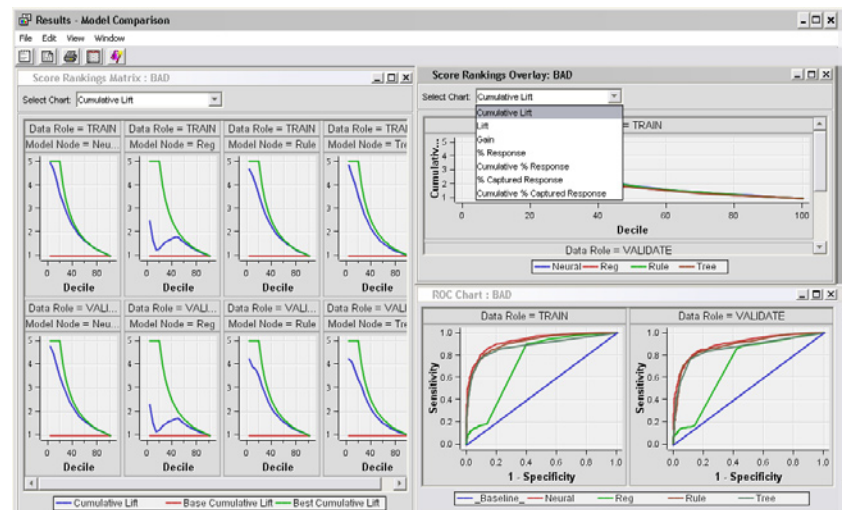
Windows (x86-32):

- Windows XP Professional
- Windows 2000 Professional
- Windows NT

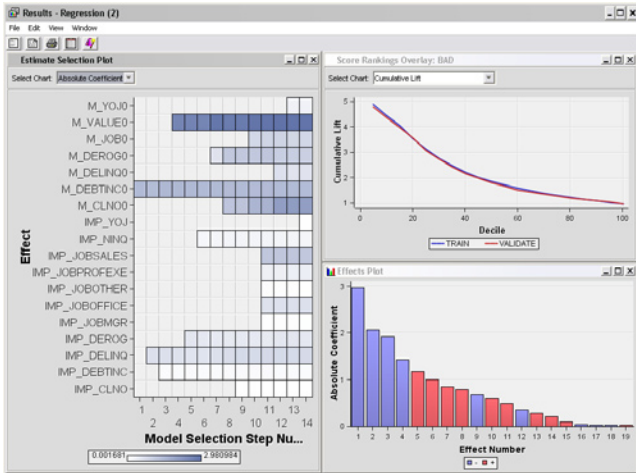
Please contact your SAS Reseller for further information on technical requirements.



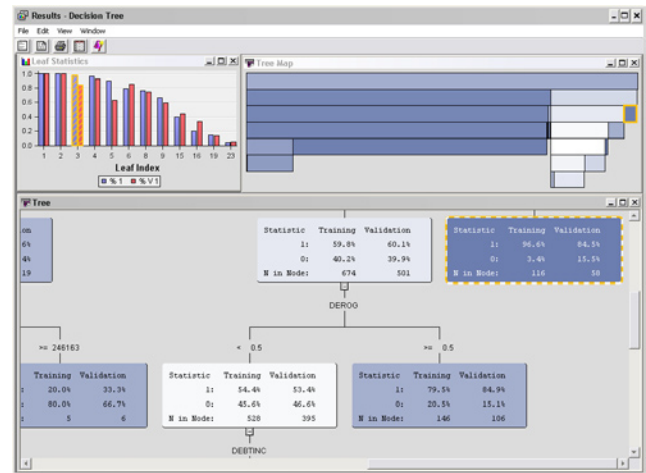
Business analysts as well as quantitative experts can build more models faster with the easy-to-use GUI.



Models generated from different modeling algorithms can be consistently evaluated within a highly visual user interface.



Develop regression models.



Create interactive decision trees.

Key features

Easy-to-use GUI for building process flow diagrams

- Build more and better models faster.
- Access to SAS programming environment.
- XML diagram exchange.
- Reuse diagrams as templates for other projects or users.
- Ability to capture institutional knowledge for repeatability and traceability.

Easy installation

- CD or DVD insert and install.

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.

Key features (continued)

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 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 wizard:
 - Titles and footnotes.
 - Apply a WHERE clause.
 - Choose from several color schemes.
 - Easily rescale axes.
- Surface the underlying data from standard SAS Desktop Data Mining for Midsize Business 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.

Key features (continued)

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.
- Self-organizing maps:
- Batch SOMs with Nadaraya-Watson or local-linear smoothing.
- Kohonen networks.
- Overlay the distribution of other variables onto the map.
- Handle 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.

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.

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 Desktop Data Mining for Midsize Business 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.

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.

Key features (continued)

Neural networks

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

Ensemble models

- 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.
- SAS scoring codes capture modeling, clustering, transformations and missing value imputation code.
- Deploy models in multiple environments requiring only Base SAS.

Utility tools

- Drop variables node.
- Merge data node.
- Metadata node for modifying columns metadata such as role, measurement level and order.
- Access to SAS Enterprise Guide® for additional data processing, and reporting.



THE
POWER
TO KNOW.

SAS Institute Inc. World Headquarters +1 919 677 8000

To contact your local SAS office, please visit: www.sas.com/offices

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. Copyright © 2007, SAS Institute Inc. All rights reserved. 000000_000000.0007