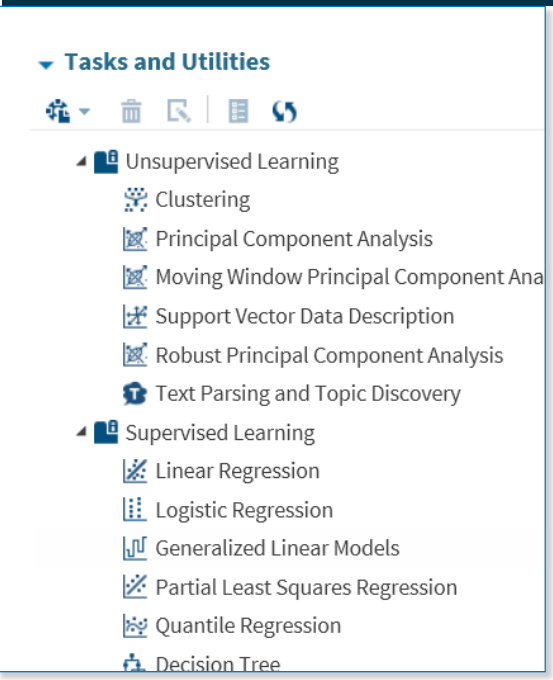
SAS® Visual Data Mining and Machine Learning

Everything needed to solve your most complex problems within a single, integrated in-memory environment



What does SAS® Visual Data Mining and Machine Learning do?

SAS Visual Data Mining and Machine Learning combines data wrangling, data exploration, visualization, feature engineering, and modern statistical, data mining and machine-learning techniques all in a single, integrated in-memory processing environment. This provides faster, more accurate answers to complex business problems, increased deployment flexibility, and reliable, secure analytics management and governance for agile IT.

Why is SAS® Visual Data Mining and Machine Learning important?

It enables data scientists and others to solve previously unfeasible business problems by removing barriers created by data sizes, data diversity, limited analytical depth and computational bottlenecks. Dramatic performance gains and innovative algorithms mean greater productivity and faster, more creative answers to your most complex problems.

For whom is SAS® Visual Data Mining and Machine Learning designed?

It is designed for those who want to use powerful and customizable in-memory algorithms in interactive and programming interfaces to analyze large, complicated data and uncover new insights faster. This includes business analysts, data scientists, experienced statisticians, data miners, engineers, researchers and scientists.

Data volumes continue to grow. Highly skilled data scientists and analytical professionals are in short supply. Organizations struggle to find timely answers to increasingly complex problems. Whether it's analyzing every transaction to identify emerging fraud patterns, analyzing growing amounts of social media chatter to improve customer experience or producing an accurate and fast recommendation system to predict next-best offers, sophisticated machine-learning software gives organizations a way to solve their most important issues.

SAS Visual Data Mining and Machine Learning addresses all of the steps necessary to turn data into new insights. Data scientists can access and prepare data, engineer features, perform exploratory analysis, build and compare machine-learning models, and create score code for implementing predictive models, more quickly than ever before.

Benefits

- **Solve complex analytical problems faster.** This solution runs on SAS® Viya™, the latest addition to the SAS Platform, delivering predictive modeling and machine-learning capabilities at breakthrough speeds. In-memory data persistence eliminates the need to load data multiple times during iterative analyses. Analytical model processing time is measured in seconds or minutes rather than hours so you can find solutions to difficult problems faster than ever.
- **Boost the productivity of your data scientists.** With support for the entire machine-learning pipeline, this solution enables data scientists to get highly accurate results quicker – all in a single environment.
- **Explore multiple approaches to find optimal solutions.** Superior performance from distributed processing and the feature-rich building blocks for machine-learning pipelines let you quickly explore and compare multiple approaches. With automated tuning, you can test different scenarios in an integrated environment to find the best performing model and provide answers with high levels of confidence.
- **Empower users with language options.** Python, R, Java and Lua programmers can experience the power of this solution without having to learn how to program in SAS. Give them access to trusted and tested SAS machine-learning algorithms they can use from other languages.
- **Use interactive interfaces for common machine-learning tasks.** Intuitive interfaces are part of the web-based programming environment and allow for easy configuration of common machine-learning tasks. The associated SAS code is automatically generated for later batch runs and automation. Users can share data sources and code snippets for improved collaboration.
- **Quickly deploy your predictive models with automatically generated SAS score code.** Shorten the time to value even more with easy-to-implement score code that is automatically generated in multiple programming languages for all your machine-learning models.

Overview

SAS Visual Data Mining and Machine Learning uses the latest statistical, text analysis and machine-learning algorithms to accelerate structured and unstructured data exploration and model development. It embraces the major programming languages for analytics, including SAS, and unifies previously disconnected and time-consuming tasks. This comprehensive, enterprise-grade solution is designed specifically with data scientists in mind.

Interactive, web-based visual and programming interfaces

SAS Visual Data Mining and Machine Learning provides a web-based interface for the most common machine-learning steps – from data prep and feature engineering to model building, assessment and scoring. You choose whether to code projects using SAS or other programming languages such as Python, R, Java or Lua. Each task in the machine-learning program generates SAS code behind the scenes for later batch runs, editing and automation. You can also create advanced machine-learning algorithms using a visual drag-and-drop interface without ever having to code.

Highly scalable, in-memory analytical processing

This solution takes advantage of the SAS Viya engine, which is optimized for multi-pass analytical computations. It provides secure, concurrent access to data in memory so many users can collaborate to explore the same raw data and build models simultaneously. Data and analytical workload operations are automatically distributed across the cores of a single server or the nodes of a massive compute cluster, taking advantage of parallel processing for blazingly fast speeds. All data, tables and objects are held in memory as long as required, allowing for efficient processing. With built-in fault tolerance and memory management, advanced workflows can be applied to data, ensuring that processes always finish.

Key Features

Interactive, web-based visual and programming interfaces

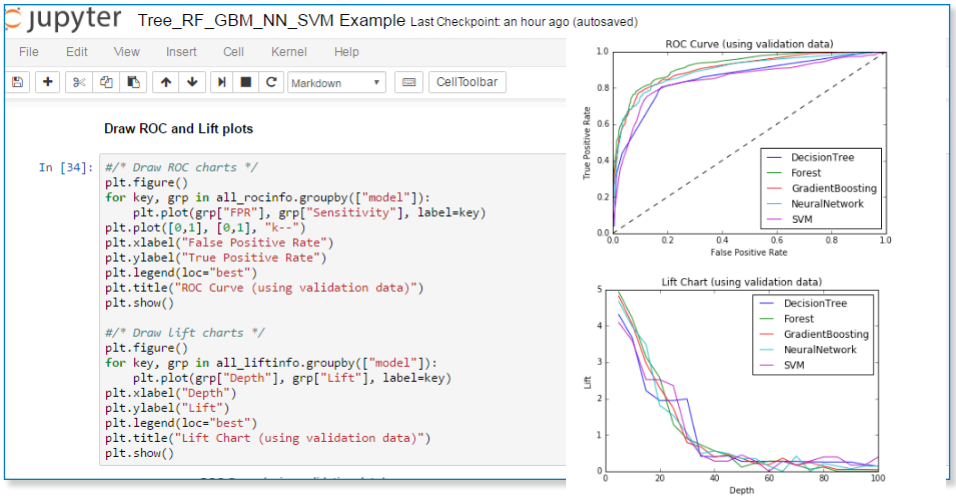
- Low maintenance, web-based interfaces for programming and point-and-click workflows.
- Generates SAS code for getting started quickly and automating machine-learning tasks.
- Collaborative environment enables easy sharing of data, code snippets and best practices.

Highly scalable, in-memory analytical processing

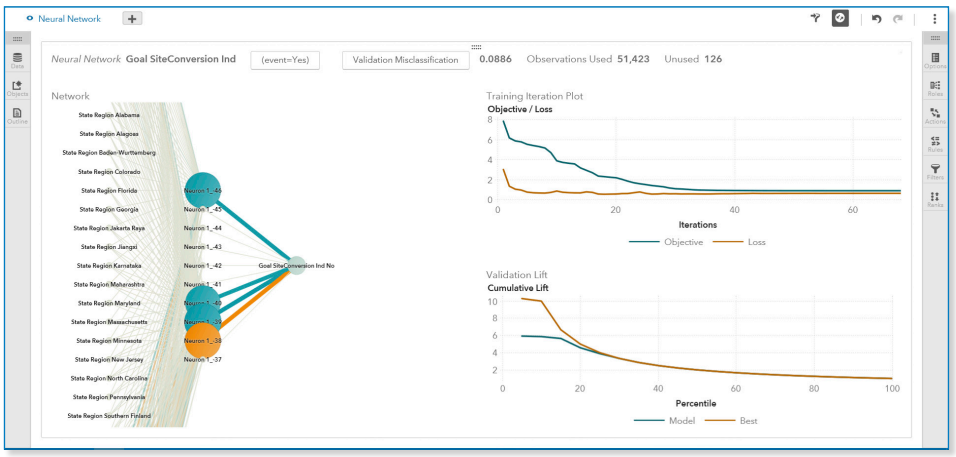
- Distributed, in-memory processing of complex analytical calculations on large data sets provides low-latency answers.
- Analytical tasks are chained together as a single in-memory job without having to reload the data or write out intermediate results to disks.
- Concurrent access to the same data in memory by many users improves efficiency.
- Data and intermediate results are held in memory as long as required, reducing latency.
- Built-in workload management ensures efficient use of compute resources.
- Built-in failover management guarantees submitted jobs always finish.

Analytical data preparation

- Quickly interpret complex relationships or key variables that influence modeling outcomes within large data sets.
- Filter observations and understand a variable's level of influence on the overall model lift.
- Use the most advanced techniques to detect rare events, outliers and/or influence points to help you determine, capture or remove them from downstream analysis (e.g., models).



With the Python interface, you can use Python code to call SAS Analytics.



Fit complex nonlinear relationships using neural networks.

Powerful data manipulation and management

Take advantage of powerful data manipulation and management capabilities within the distributed, in-memory environment to prepare data for analytics. Access data, join tables, subset and filter data, and create the final table for machine-learning projects.

Data exploration, feature engineering and dimension reduction

Explore your data with descriptive statistics and powerful graphical programming. Discover data issues and fix them with advanced analytical techniques. Quickly identify potential predictors, reduce the dimensions of large data sets and easily create new features from your original data.

Modern statistical, data mining and machine-learning techniques

Apply powerful unsupervised and supervised learning algorithms, such as clustering, principal component analysis, linear and nonlinear regression, logistic regression, decision trees, random forests, gradient boosting, neural networks and support vector machines to your structured and unstructured data, and quickly identify the champion model.

With matrix factorization, you can build customized recommendation systems. And more generalized and powerful than matrix factorization, tensor factorizations create low-dimensional representations of multiway data in a tensor, as opposed to two-way data in a matrix. Tensor factorization allows you to customize recommendations for n-way combinations instead of just two-way combinations.

Open ecosystem drives greater analytics value

Empower your data scientists with SAS Analytics that are easily available from a variety of programming languages. Whether it's Python, R, Java or Lua, modelers and data scientists can access the power of SAS from their preferred coding environment. Take advantage of SAS analytical capabilities without having to develop SAS coding skills.

Key Features (continued)

- Explore data using bar charts, histograms, box plots, heat maps, bubble plots, geographic maps and more.
- Derive predictive outputs or segmentations that can be used directly in other modeling or visualization tasks.

Data exploration, feature engineering and dimension reduction

- Large-scale binning of continuous features.
- High-performance imputation of missing values in features with user-specified values, mean, pseudo-median and random value of nonmissing values.
- Large-scale dimension reduction for continuous and categorical features.
- Large-scale principal components analysis (PCA), including moving windows and robust PCA.
- Unsupervised learning with cluster analysis and mixed variable clustering.

Model development with statistical algorithms

- Linear regression models.
- Logistic regression models.
- Generalized linear models.
- Nonlinear regression models.
- Quantile regression models.
- Predictive partial least squares.
- Decision trees.

Model development with modern machine-learning algorithms

- Random forests for binary, nominal and interval labels:
  - Automated ensemble of decision trees predict a single target.
  - Automated distribution of independent training runs.
  - Automated intelligent tuning of parameter set to identify optimal model.
- Gradient boosting for binary, nominal and interval labels:
  - Automated iterative search for optimal partitioning of data in relation to selected label variable.
  - Automated generation of weighted averages for final supervised model.
  - Automated stopping criteria based on validated data scoring to avoid overfitting.
- Neural networks for binary, nominal and interval labels:
  - Intelligent defaults for most neural network parameters.
  - Customizable neural networks architecture and weights.
  - Ability to use an arbitrary number of hidden layers to support deep learning.
  - Automatic out-of-bag validation for early stopping to avoid overfitting.
  - Automated intelligent tuning of parameter set to identify optimal model.
  - Surface autoencoders for enhanced dimension reduction.
- Support vector machines for binary labels:
  - Model training with linear and polynomial kernels.
  - Ability to apply the interior-point method and the active-set method.
  - Support for data partitioning for model validation.
  - Support for cross-validation for penalty selection.
  - Automated intelligent tuning of parameter set to identify optimal model.
- Factorization machines:
  - Develop recommender systems based on sparse matrices of user IDs and item ratings.
  - Apply full pairwise-interaction tensor factorization.
  - Supercharge models with time stamps, demographic data and context information.
  - Supports warm restart so you can update models with new transactions without having to fully retrain.
  - Automatically invoke tensor factorization for recommendations involving more than two nominal variables.

And with SAS Viya REST APIs, you can add the power of SAS Analytics to your other applications.

Automated model tuning

Discover optimal model configurations with automated, intelligent hyperparameter tuning. Based on your selection of which hyperparameters to explore and within what ranges, the integrated autotuning process uses any of a number of available search strategies to train and assess models in parallel. Eighteen different model assessment metrics can be used as your tuning objective, evaluated using either a validation partition or the built-in cross-validation mechanism. (See an example on Page 1.)

Integrated text analytics

Designed with big data in mind, you can examine large collections of text documents to gain new insights about unknown themes and connections using powerful text preprocessing, natural language processing, topic detection and more. The integrated text analytics capabilities also enable data scientists to use the insights hidden in unstructured data for improved supervised learning.

Model assessment and scoring

Test different modeling approaches in a single run and compare results of multiple supervised learning algorithms with standardized tests to quickly identify champion models. When the champion model has been identified, operationalize analytics in distributed and traditional environments with automatically generated SAS score code.

TO LEARN MORE »

To learn more about SAS Visual Data Mining and Machine Learning, view screenshots and see other related materials, please visit sas.com/dmml.

Key Features (continued)

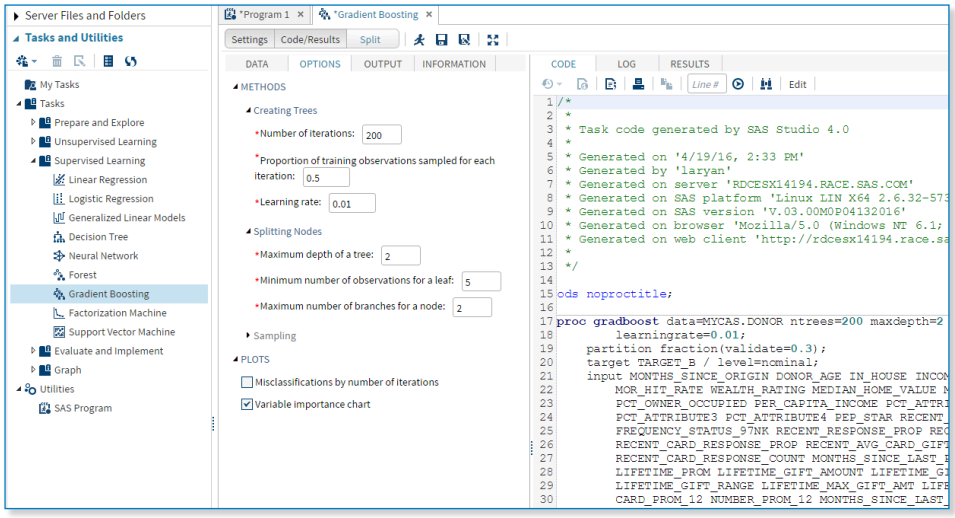
- Automated intelligent tuning of parameter set to identify optimal model.
- Network analytics and community detection:
  - Augment data mining and machine-learning approaches with graph theory and network analysis algorithms.
  - Apply pairwise interaction between entities of interest.
  - Assumptionless approach improves detection of the many ways a network may arise.
  - Model the weights of the network links based on the strength of interaction frequency.

Integrated text analytics

- Supports 27 native languages out of the box (English, Arabic, Croatian, Czech, Danish, Dutch, Finnish, French, German, Greek, Hebrew, Italian, Japanese, Korean, Norwegian, Polish, Portuguese, Russian, Spanish, Slovak, Slovenian, Swedish, Turkish, Chinese, Vietnamese, Indonesian, Thai).
- Automatically identifies term part of speech (more than 15 are system defined).
- Extracts standard entities such as location, time, date and address from predefined options.
- Detects noun groups and multiterm lists, and creates single terms for processing.
- Finds term variants automatically with synonym detection.
- Uses default start and stop lists to manage terms for parsing and downstream processing.
- Machine-learned topics represent the term-by-document, matrix-generated text processing as a structured numeric representation of the document collection.
- Extract Boolean rules from large-scale transactional data.
- Automatically generate a set of Boolean rules by analyzing a text corpus.

Model assessment and scoring

- Supervised learning model performance statistics are automatically calculated for a selected model with binary, nominal or interval label.
- Creates lift table for interval and categorical target, and ROC table for categorical target.
- Automated generation of SAS DATA step code for model scoring.
- Apply scoring logic to training, holdout data and new data.
- Score new textual data.



Selecting model options in the middle panel generates code that can be edited and shared, in this case for a gradient boosting model.

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

THE POWER TO KNOW.