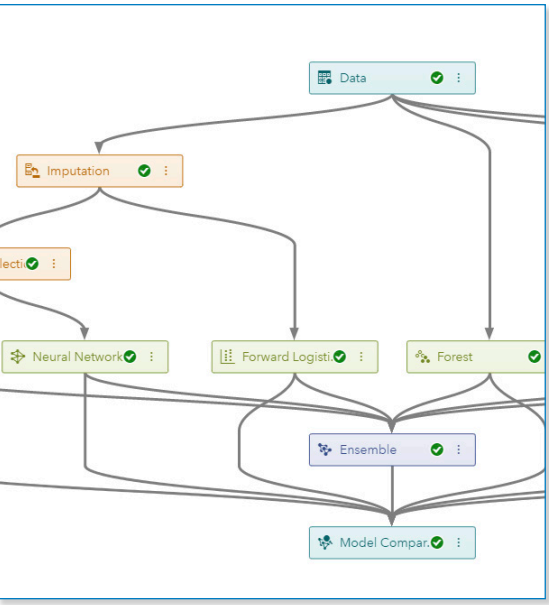
SAS® Visual Data Mining and Machine Learning

Everything you need to solve the most complex analytical problems within a single, integrated and collaborative solution



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

It provides a comprehensive, visual interface for accomplishing all steps related to the analytical life cycle. In addition to innovative machine learning and deep learning techniques for analyzing structured and unstructured data, it integrates all other tasks in your analytical processes. From data preparation and exploration to model development and deployment, multiple personas work in the same, integrated environment. Scalable and elastic processing provide flexibility and speed for faster answers to complex questions.

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

SAS Visual Data Mining and Machine Learning is the first solution that combines the most advanced analytics, data prep, visualization, model assessment and model deployment in a single environment. It also supports programming from popular open source languages. This consistent, collaborative environment produces repeatable results, helping improve organizational processes and uncover new opportunities for growth.

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

It is designed for anyone who needs to analyze large, complex data and build predictive models. This includes data scientists, statisticians, data miners, business analysts, citizen data scientists, data engineers and researchers.

Data collections 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 can help organizations solve critical issues.

SAS Visual Data Mining and Machine Learning addresses all of the steps necessary to turn raw data into insights – using an integrated and visual pipeline interface. A variety of analytical professionals can access and prepare data, engineer features, perform exploratory analysis, build and compare machine learning models, and create score code for implementing predictive models, faster than ever before.

Benefits

- **Boost the productivity of your analytical teams.** With support for the entire machine learning pipeline, this solution enables a variety of users to build and expand upon sophisticated models to get highly accurate results – all in a single, highly collaborative environment.
- **Reduce latency between data and deployment.** Interactive visual and programming interfaces dramatically shorten the time it takes to prepare data, build models and deploy them into production. High-speed processing delivers rapid results.
- **Explore multiple approaches to find optimal solutions – with confidence.** Superior performance from distributed processing and the feature-rich building blocks for machine learning pipelines let numerous users quickly explore and compare multiple approaches. Automated tuning tests different scenarios to find the best-performing model. Reproducibility in every stage of the analytical life cycle delivers answers and insights everyone can trust.
- **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.
- **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 machine learning models.
- **Empower users with language options.** Through the use of APIs, Python, R, Java, Lua and Scala 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.

Overview

SAS Visual Data Mining and Machine Learning offers an exciting, end-to-end visual environment that covers all aspects of machine learning and deep learning – from data access and data wrangling to sophisticated model building and deployment. In-memory, distributed processing handles large data and complex modeling, providing faster answers and efficient use of resources.

Flexible and approachable visual environment for analytics

SAS Drive is a full function, extensible content management application for SAS Viya. It gives users a straightforward way to create, manage and share content, and administer content permissions. The highly collaborative workspace enables users to easily see all work going on in a particular project. Content includes things like SAS Visual Analytics reports, SAS Data Management projects, SAS Studio code and more.

Another feature, the Exchange, organizes your favorite settings and lets you collaborate with others in one place. You can find a recommended node template or create your own template for a streamlined workflow for your team.

Kev Features

Interactive programming in a web-based development environment

- Visual interface for the entire analytical life cycle process.
- Drag-and-drop interactive interface requires no coding, though coding is an option.
- Supports automated code creation at each node in the pipeline.
- Best practice templates (basic, intermediate or advanced) help users get started quickly with machine learning tasks.
- Interpretability reports.
- Explore data from within Model Studio and launch directly into SAS Visual Analytics.
- View data within each node in Model Studio.
- Run SAS® Enterprise Miner ™ 14.3 batch code within Model Studio.
- Provides a collaborative environment for easy sharing of data, code snippets and best practices between different personas.

Highly scalable, distributed 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.
- Automated I/O disk spillover for improved memory management.

Model development with modern machine learning algorithms

- Decision forests:
  - Automated ensemble of decision trees to predict a single target.
  - Automated distribution of independent training runs.
  - Supports intelligent autotuning of model parameters.
  - Automated generation of SAS code for production scoring.

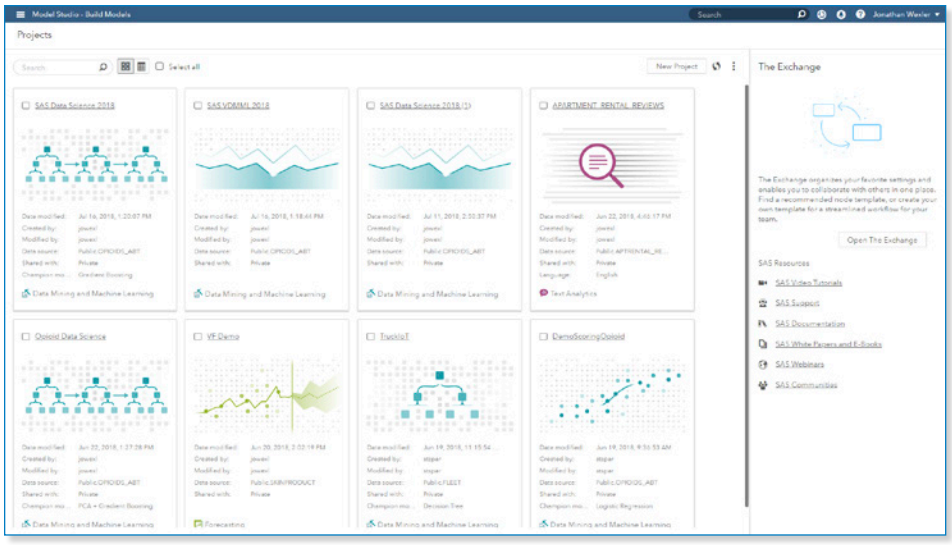


Figure 1: SAS Drive provides a collaborative workspace for users to create, share and manage projects and content.

The visual interface (Model Studio) provides an integrated environment for the most common machine learning steps: data wrangling, feature engineering, data exploration, model building and deployment. This highly collaborative environment is ideal for building, expanding and sharing models.

Multiple users can currently analyze any amount of structured and unstructured data with the Model Studio. Each project (goal) is defined by visual pipelines that break the analytics life cycle into a series of steps presented in a logical sequence. Pipeline branching can execute asynchronously. Within the pipeline, interactive tasks provide an easy way to apply sophisticated algorithms to large and complex data. These interactions also generate SAS code that can be save for later automation of tasks. In addition, code snippets and best practice templates are easily shared.

To enhance collaborative understanding, users are provided with business-friendly annotations within each node that describe what methods are being run, such as information about the methods, results and interpretation. Standard interpretability reports are also provided in all modeling nodes, including LIME, ICE, PD plots, etc.

Key Features (continued)

- Gradient boosting:
  - Automated iterative search for optimal partition of the data in relation to selected label variable.
  - Automated resampling of input data several times with adjusted weights based on residuals.
  - Automated generation of weighted average for final supervised model.
  - Supports binary, nominal and interval labels.
  - Ability to customize tree training with variety of options for numbers of trees to grow, splitting criteria to apply, depth of subtrees and compute resources.
  - Automated stopping criteria based on validation data scoring to avoid overfitting.
  - Automated generation of SAS code for production scoring.
- Neural networks:
  - Automated intelligent tuning of parameter set to identify optimal model.
  - Supports modeling of count data.
  - Intelligent defaults for most neural network parameters.
  - Ability to customize neural networks architecture and weights.
  - Techniques include deep forward neural network (DNN), convolutional neural networks (CNNs), recurrent neural networks (RNNs) and autoencoders.
  - Ability to use an arbitrary number of hidden layers to support deep learning.
  - Automatic standardization of input and target variables.
  - Automatic selection and use of a validation data subset.
  - Automatic out-of-bag validation for early stopping to avoid overfitting.
  - Supports intelligent autotuning of model parameters.
  - Automated generation of SAS code for production scoring.
- Support vector machines:
  - Models binary target labels.
  - Supports linear and polynomial kernels for model training.
  - Ability to include continuous and categorical in/out features.
  - Automated scaling of input features.
  - Ability to apply the interior-point method and the active-set method.
  - Supports data partition for model validation.

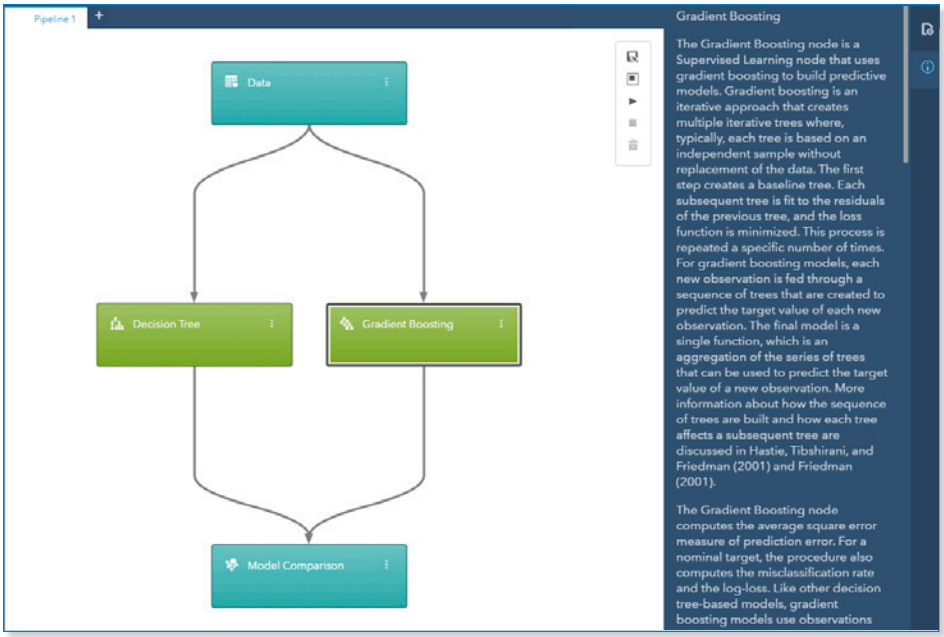


Figure 2: Annotations within each node provide information to aid collaboration.

Highly scalable, in-memory analytical processing

This solution provides a secure, multiuser environment for concurrent access to data in memory. Data and analytical workloads operations are distributed across nodes, in parallel, and are multithreaded on each node for very fast speed.

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.

You get dramatically reduced runtimes for large data and analytical processing, reduced network traffic and can take full advantage of modern, multicore architectures to find solutions much faster.

Innovative and robust statistical, data mining and machine learning techniques

SAS Visual Data Mining and Machine Learning delivers an incredibly broad set of modern statistical, machine learning, deep learning and text analytics algorithms within a single environment.

Analytical capabilities include clustering, different flavors of regression, decision forests, gradient boosting models, support vector machines, natural language processing, topic detection and more. These powerful methods drive the identification of new patterns, trends and relationships between data attributes in structured and unstructured data. The solution also provides matrix factorization for building customized recommendation systems.

With its ability to process high velocity and high-volume data sets, SAS Visual Data Mining and Machine Learning is uniquely suited for deep learning techniques. Deep learning algorithms include deep neural

Key Features (continued)

- Supports cross-validation for penalty selection.
- Automated generation of SAS code for production scoring.

- Factorization machines:
  - Supports the development of recommender systems based on sparse matrices of user IDs and item ratings.
  - Ability to apply full pairwise-interaction tensor factorization.
  - Includes additional categorical and numerical input features for more accurate models.
  - Supercharge models with timestamps, demographic data and context information.
  - Supports warm restart (update models with new transactions without full retraining).
  - Automated generation of SAS score code for production scoring.

- Bayesian networks:
  - Learns different Bayesian network structures, including naive, tree-augmented naive (TAN), Bayesian network-augmented naive (BAN), parent-child Bayesian networks and Markov blanket.
  - Performs efficient variable selection through independence tests.
  - Selects the best model automatically from specified parameters.
  - Generates SAS code or an analytics store to score data.
  - Loads data from multiple nodes and performs computations in parallel.

- Dirichlet Gaussian mixture models (GMM):
  - Can execute clustering in parallel and is highly multithreaded.
  - Performs soft clustering, which provides not only the predicted cluster score but also the probability distribution over the clusters for each observation.
  - Learns the best number of clusters during the clustering process, which is supported by the Dirichlet process.
  - Uses a parallel variational Bayes (VB) method as the model inference method. This method approximates the (intractable) posterior distribution and then iteratively updates the model parameters until it reaches convergence.

- Semisupervised learning algorithm:
  - Highly distributed and multithreaded.

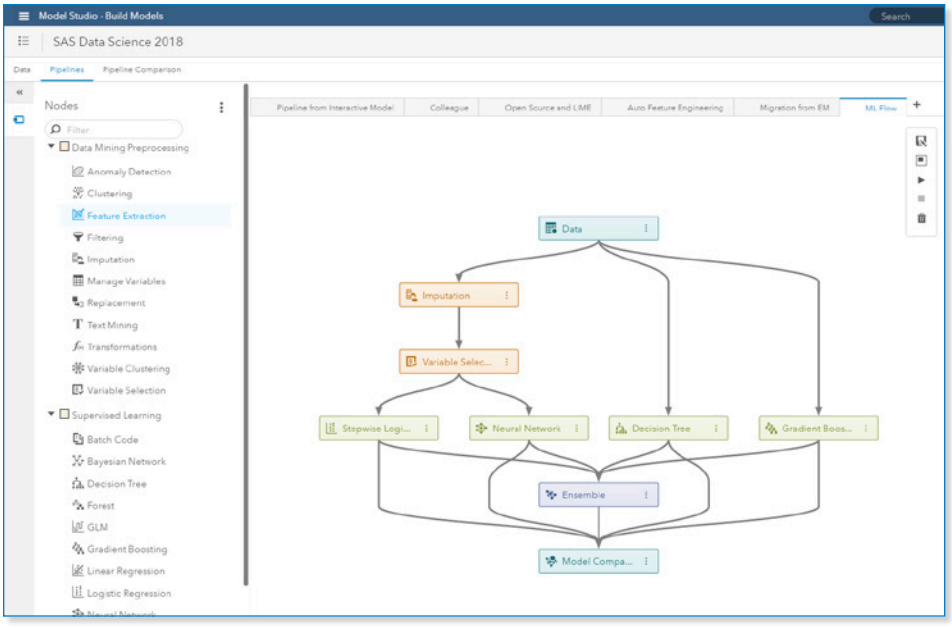


Figure 3: Visual pipelines break the analytics life cycle into a series of steps presented in a logical sequence.

networks, convolution neural networks for image classification and recurrent neural networks for improved text analysis. Users can customize these networks and support different types of layers such as convolution and pooling. They can also use networks built with Keras and Caffe, for example, and use as a 'warm start' within SAS.

Complex machine learning algorithms, such as decision trees, neural networks, support vector machines, gradient boosting and decision forests, can be automatically tuned for optimal performance, saving time and resources.

In addition, you can run SAS Enterprise Miner batch code within Model Studio. This enables you to easily compare SAS Enterprise Miner models against others, including open source.

Embedded support for Python and R languages

Users can embed open source code within an analysis and call open source algorithms seamlessly within a Model Studio flow. This facilitates collaboration across all personas within a department because users can program in the language of their choice The Open Source Code node in Model Studio is agnostic to Python or R software versions; any version can be used as the code is passed.

Key Features (continued)

- Returns the predicted labels for both the unlabeled data table and the labeled data table.

- T-distributed stochastic neighbor embedding (t-SNE):
  - Highly distributed and multithreaded.
  - Returns low-dimensional embeddings that are based on a parallel implementation of the t-SNE algorithm.

Analytical data preparation

- Feature engineering best practice pipeline includes best transformations.
- Distributed data management routines provided via a visual front end.
- Large-scale data exploration and summarization.
- Cardinality profiling:
  - Large-scale data profiling of input data sources.
  - Intelligent recommendation for variable measurement and role.
- Sampling: Supports random and stratified sampling, oversampling for rare events and indicator variables for sampled records.

Data exploration, feature engineering and dimension reduction

- T-distributed stochastic neighbor embedding (t-SNE).
- Feature binning.
- High-performance imputation of missing values in features with user-specified values, mean, pseudo median and random value of nonmissing values.
- Feature dimension reduction.
- Large-scale principal components analysis (PCA), including moving windows and robust PCA.
- Unsupervised learning with cluster analysis and mixed variable clustering.

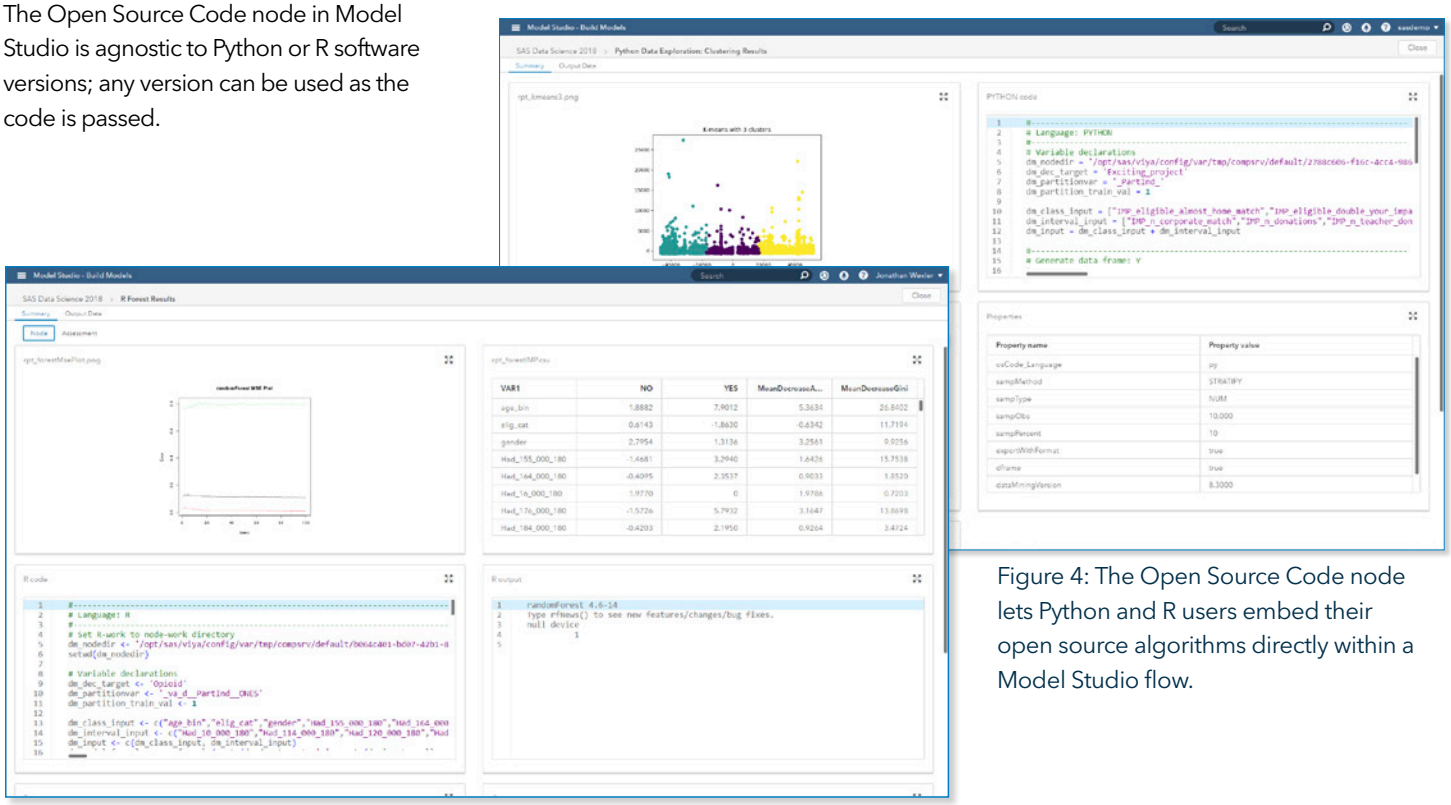


Figure 4: The Open Source Code node lets Python and R users embed their open source algorithms directly within a Model Studio flow.

Integrated data preparation, exploration and feature engineering

To overcome time-consuming analytical data preparation activities, the drag-and-drop interface enables data engineers to quickly build and run transformations, augment data and join data within the integrated visual pipeline of activities. All actions are performed in memory to maintain a consistent data structure. 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.

Integrated text analytics

SAS Visual Data Mining and Machine Learning includes integrated text analytics for users who want to incorporate features derived from free-form text into a predictive model for text parsing and topic discovery, automatic Boolean-rule generation for categorical target variables and scoring data for text topics.

Designed with big data in mind, you can examine extremely large collections of text documents. Explore all of your textual data, not just a subset, to gain new insights about unknown themes and connections. Combining structured data with text data uncovers previously undetected relationships and adds even more predictive power to analytical models.

Key Features (continued)

Integrated text analytics

- Supports 32 native languages out of the box: English, Arabic, Chinese, Croatian, Czech, Danish, Dutch, Farsi, Finnish, French, German, Greek, Hebrew, Hindi, Hungarian, Indonesian, Italian, Japanese, Korean, Norwegian, Polish, Portuguese, Romanian, Russian, Slovak, Slovenian, Spanish, Swedish, Tagalog, Turkish, Thai and Vietnamese.
- Automated parsing, tokenization, part-of-speech tagging and lemmatization.
- Predefined concepts extract common entities such as names, dates, currency values, measurements, people, places and more.
- Automated feature extraction with machine-generated topics (singular value decomposition and latent Dirichlet allocation).
- Supports machine learning and rules-based approaches within a single project.
- Automatic rule generation with the BoolRule.
- Classify documents more accurately with deep learning (recurrent neural networks).

Model assessment

- Automatically calculates supervised learning model performance statistics.
- Produces output statistics for interval and categorical targets.
- Creates lift table for interval and categorical target.
- Creates ROC table for categorical target.

Model scoring

- Automatically generates SAS DATA step code for model scoring.
- Applies scoring logic to training, holdout data and new data.

SAS® Viya® in-memory engine

- CAS (SAS Cloud Analytic Services) performs processing in memory and distributes processing across nodes in a cluster.
- User requests (expressed in a procedural language) are translated into actions with the parameters needed to process in a distributed environment. The result set and messages are passed back to the procedure for further action by the user.

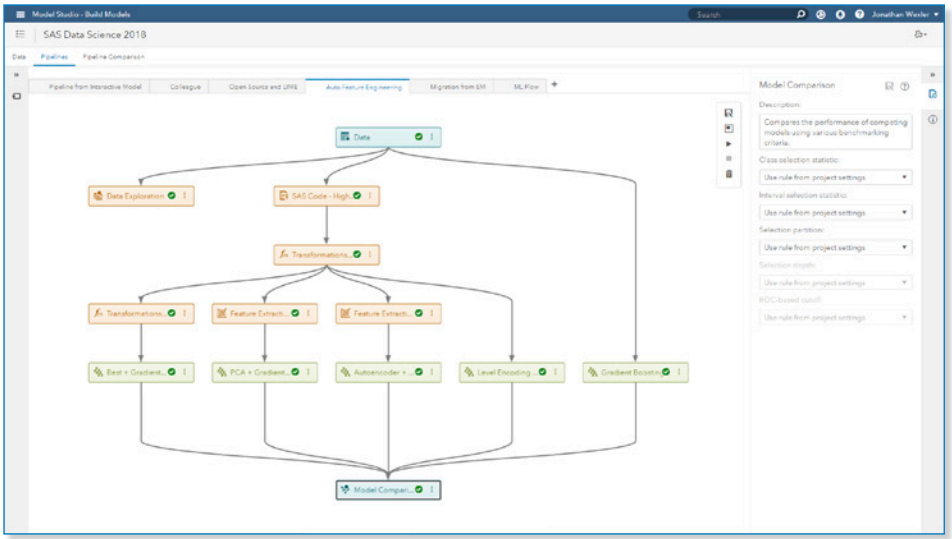


Figure 5: A best practice feature engineering template helps identify the best set of predictors, transformations and extracted features to use in machine learning models.

Lineage viewer

A lineage viewer visually displays the relationships between decisions, models, data and decisions. Relationships can be created to define the lineage between various objects, and the relationships are displayed as arcs between the models.

The SAS lineage viewer surfaces both business and technical metadata, enabling users to trace lineages from source to report, search content and add content to the metadata repository.

Model assessment and scoring

Now it's easy to test different modeling approaches in a single run, and compare results of multiple machine learning algorithms with standardized tests to automatically identify champion models.

Then to deliver real value, you can quickly operationalize analytics in all kinds of environments (distributed, traditional) with automatically generated SAS score code.

With just one click, models can be registered and published or APIs can be created.

Accessible and cloud-ready

Whether it's Python, R, Java, Lua or Scala, modelers and data scientists can access SAS capabilities from their preferred coding environment. And with SAS Viya REST APIs, they can add the power of SAS to other applications.

Additionally, SAS Visual Data Mining and Machine Learning can be deployed where it makes the most sense for your organization – whether that's on-site, in a private cloud via technologies such as Cloud Foundry, or in public clouds (like Amazon Web Services and Microsoft Azure).

You can also access this software via the predeployed and preconfigured managed software-as-a-service offerings provided by SAS.

Key Features (continued)

- Data is managed in blocks and can be loaded in memory and on demand.
- If tables exceed memory capacity, the server caches the blocks on disk. Data and intermediate results are held in memory as long as required, across jobs and user boundaries.
- Includes highly efficient node-to-node communication. An algorithm determines the optimal number of nodes for a given job.
- Communication layer supports fault tolerance and lets you remove or add nodes from a server while it is running. All components can be replicated for high availability.
- Support for legacy SAS code and direct interoperability with SAS 9.4M5 clients.
- Supports multitenancy deployment, allowing for a shared software stack to support isolated tenants in a secure manner.

SAS® procedures (PROCs) and CAS actions

- A programming interface (SAS Studio) allows IT or developers to access a CAS server, load and save data directly from a CAS server, and support local and remote processing on a CAS server.
- Python, Java, R, Lua and Scala programmers or IT staff can access data and perform basic data manipulation against a CAS server, or execute CAS actions using PROC CAS.
- Integrate and add the power of SAS to other applications using REST APIs.

Deployment options

- On-site deployments:
  - Single-machine server to support the needs of small to midsize organizations.
  - Distributed server to meet growing data, increasing workloads and scalability requirements.
- Cloud deployments:
  - Enterprise hosting; private or public cloud (e.g., BYOL in Amazon) infrastructure; SAS managed software as a service (SaaS); and Cloud Foundry platform as a service (PaaS) to support multiple cloud providers.

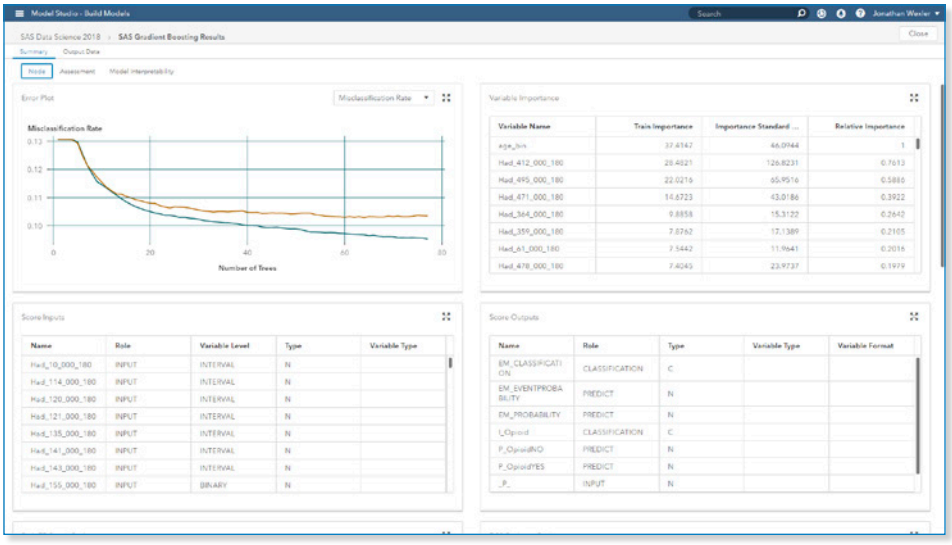


Figure 6: Model assessment features compare results from multiple algorithms to automatically identify champion models.

TO LEARN MORE »

To learn more about SAS Visual Data Mining and Machine Learning, download white papers, view screenshots and see other related material, please visit sas.com/vdmml.

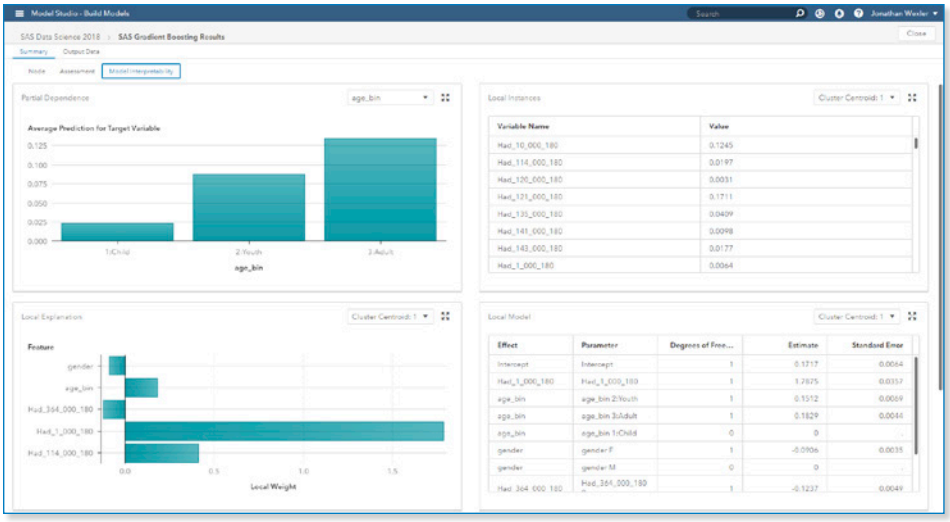


Figure 7: Standard interpretability reports are provided in all SAS Visual Data Mining and Machine Learning modeling nodes, including LIME, ICE, PD plots, etc.

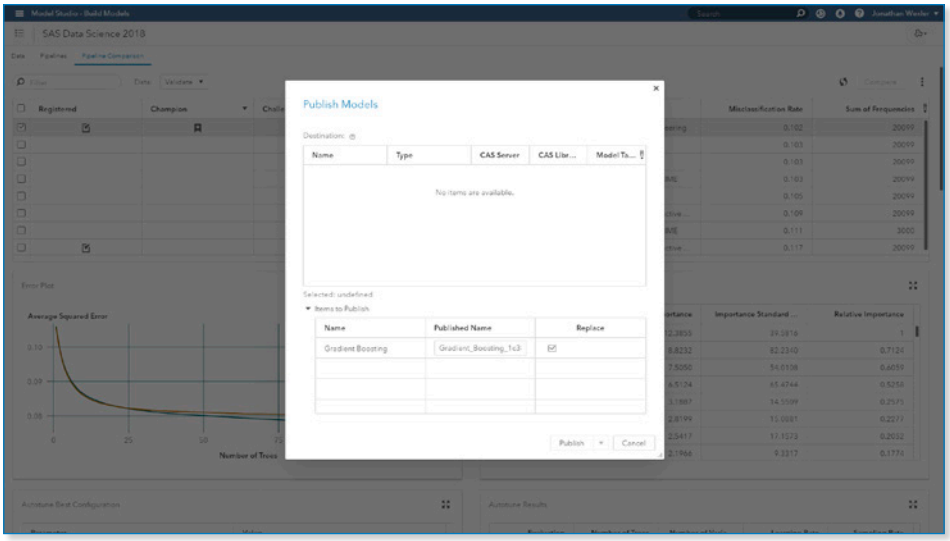


Figure 8: With just a single click, models can be registered and published or APIs can be created.