How does SAS® Support Machine Learning?

Getting the most out of your data

How does SAS support Machine Learning?

Agenda

- What is Machine Learning?
- Terminology and key characteristics
- How you can use machine learning in SAS
- Examples in Enterprise Miner
- Examples in SAS Viya

Machine Learning

What is Machine Learning?

Definition

- Automatic
- Adaptive

*Using iterative processes, machine learning builds models that **automatically adapt** with little or no human intervention.*

SAS Analytics in Action

Machine Learning

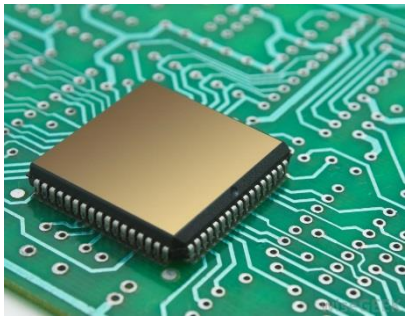
Automate
- Provide automation to the model building process by minimizing human intervention

Customize
- Build powerful models using SAS's state-of-the-art algorithms in conjunction with open source tools

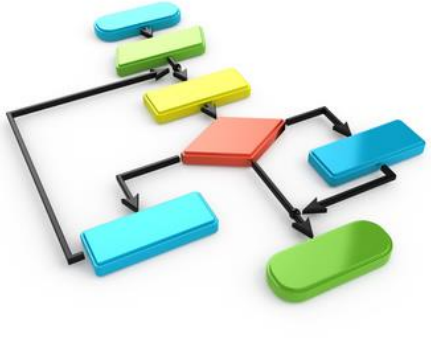
Speed
- Fast response time for sophisticated analytics applied to data of any size or complexity

Machine learning : Why is it so important now?

Data

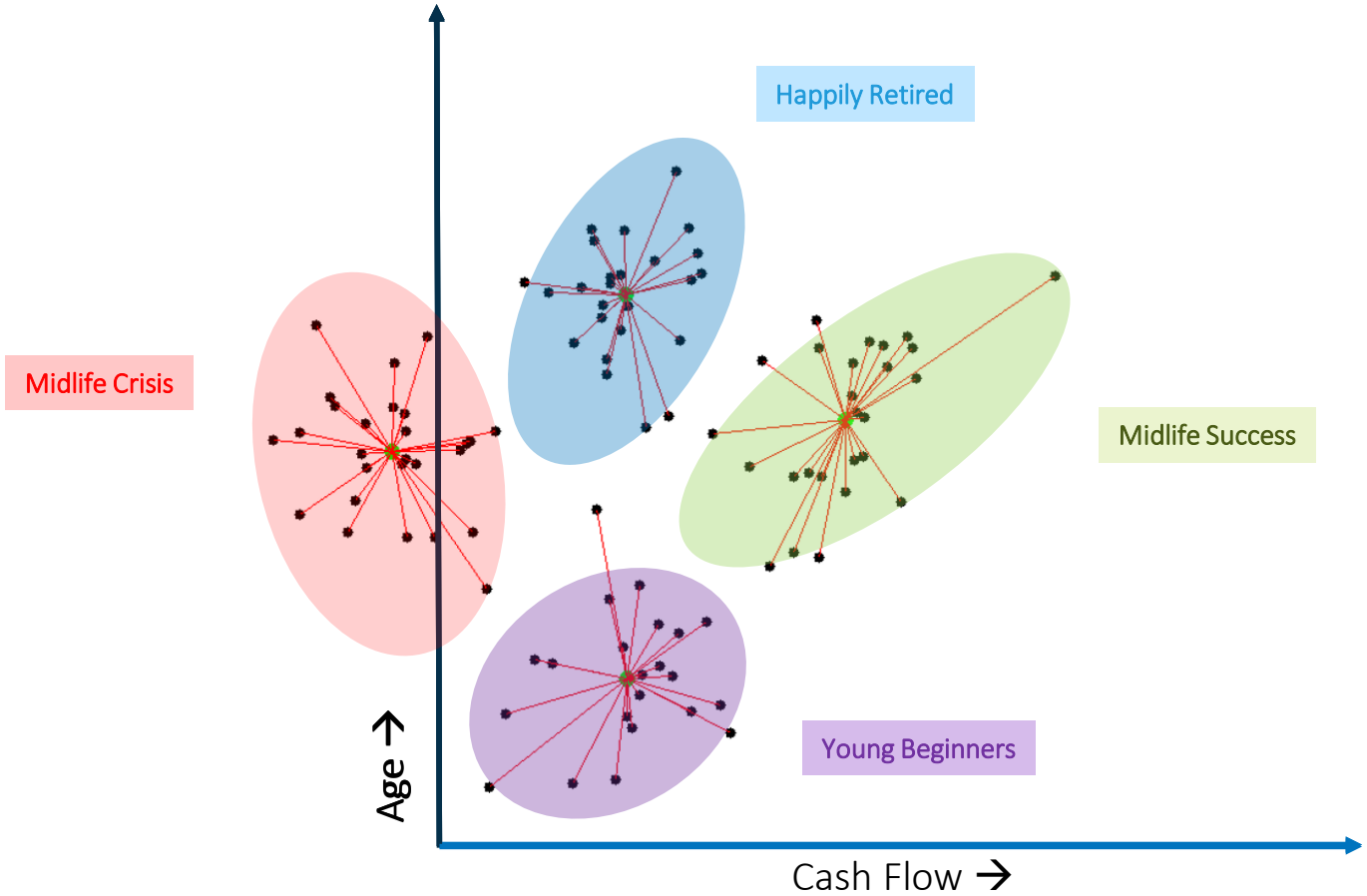


Computing Power



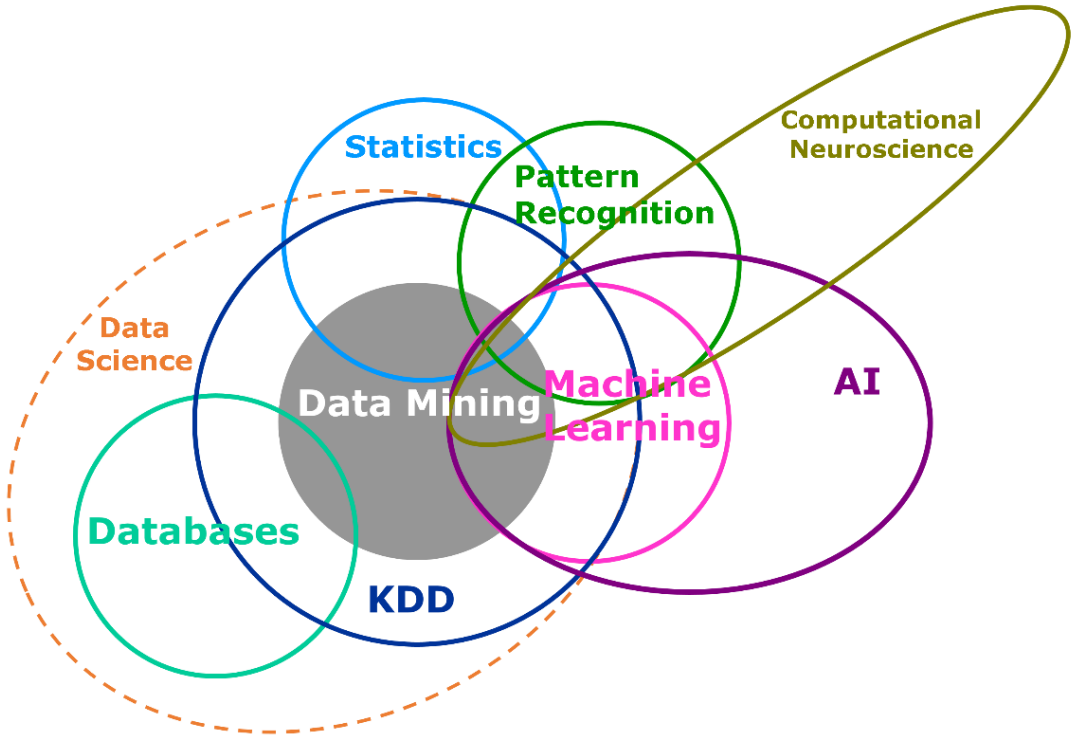
Algorithms

The iterative nature of Machine learning



What is Machine learning?

Machine learning and inferential statistics



There is a lot of overlap!

Machine Learning

- Results-driven
- Black box analytics

Inferential Statistics

- Inferential
- White box analytics

Terminology

Terminology

What are Machine Learning's terminologies?

Data Mining - a superset of many different methods to extract insights from data.

Machine Learning - use computers to probe the data for structure, even if we do not have a theory of what that structure looks like.

Deep learning - combines advances in computing power and special types of neural networks to learn complicated patterns in large amounts of data.

Data Mining

What are Machine Learning's terminologies?

Data Mining - a superset of many different methods to extract insights from data.

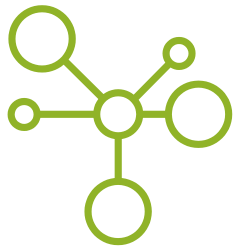
- may involve traditional statistical methods and machine learning.
- applies methods from many different areas to identify previously unknown patterns from data.
- can include statistical algorithms, machine learning, text analytics, time series analysis and other areas of analytics.
- also includes the study and practice of data storage and data manipulation.

Machine Learning

What are Machine Learning's terminologies?

Machine Learning - use computers to probe the data for structure, even if we do not have a theory of what that structure looks like.

The test for a machine learning model is a validation error on new data, not a theoretical test that proves a null hypothesis.



Deep Learning

What are Machine Learning's terminologies?

Deep learning - combines advances in computing power and special types of neural networks to learn complicated patterns in large amounts of data.

- Deep learning techniques are currently state-of-the-art for identifying objects in images and words in sounds.
- Researchers are now looking to apply these successes in pattern recognition to more complex tasks
- Automatic language translation, medical diagnoses and numerous other important social and business problems.

Terminology

Machine learning terms versus inferential statistics terms

What are all these archaic, outmoded and confusing terms?

What are all these new fangled and confusing terms?

- Feature
- Input
- Target
- Object

- Variable
- Independent Variable
- Dependent Variable
- Observation

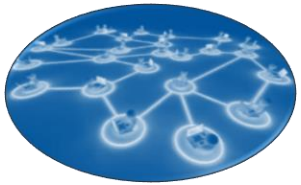
Machine Learning Usage

Practical Applications

Credit Risk

Object recognition

Self-driving Cars



Cyber Security

Sentiment Analysis in Text

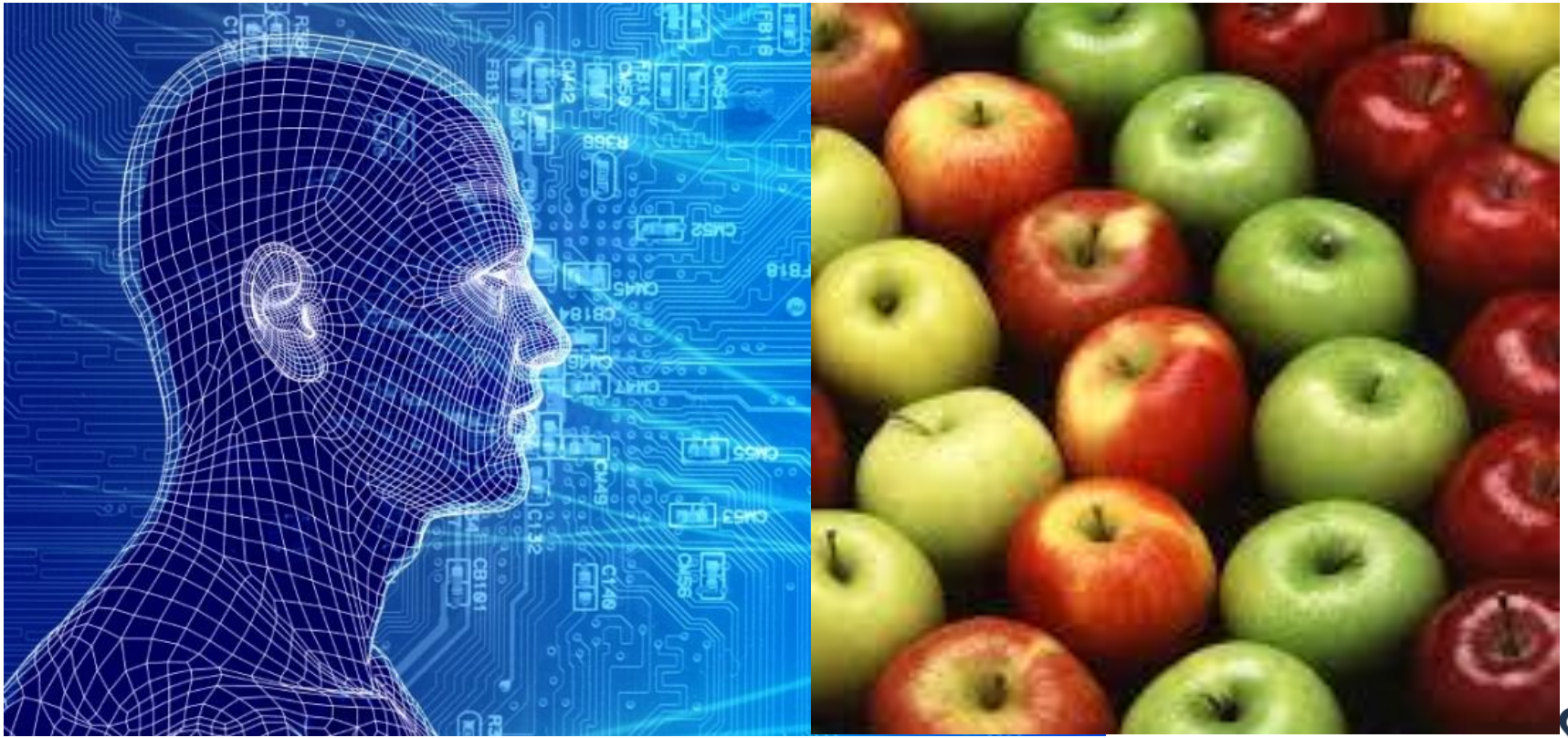
Fraud Detection

How Does Machine Learning Work?

Distinguish apple from orange

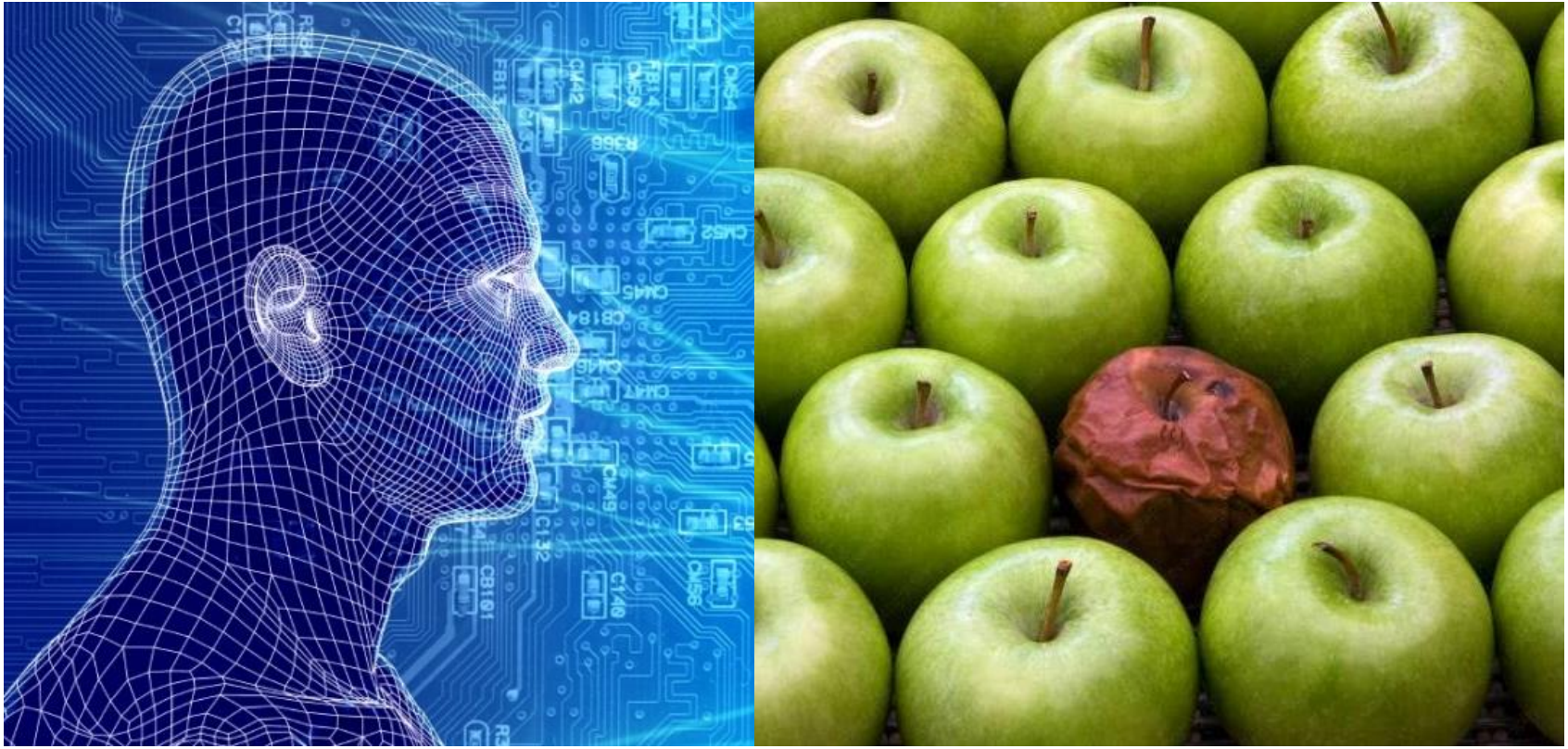
How Does Machine Learning Work?

Distinguish Granny Smith apple from Fuji apple



How Does Machine Learning Work?

Finding the rotten apple

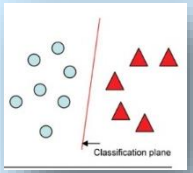


How Does Machine Learning Work?

Types of Learning

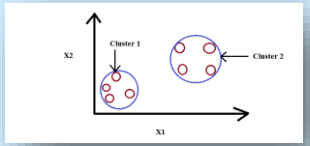
SUPERVISED LEARNING

- Trained on labeled data (Target variable)
- Classification, Prediction
- Algorithms: Logistic Regression, Gradient Boosting etc.



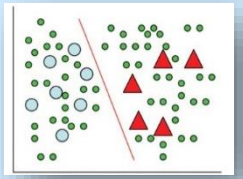
UNSUPERVISED LEARNING

- Unlabeled data (no Target variable)
- Clustering, Feature Extraction
- Algorithms: K-means clustering, PCA, etc.



SEMI-SUPERVISED LEARNING

- Combination of labeled and unlabeled data
- Classification, Regression, Prediction
- Algorithms: Autoencoders, TSVM etc.



*In semi-supervised learning, supervised prediction and classification algorithms are often combined with clustering.

How Does Machine Learning Work?

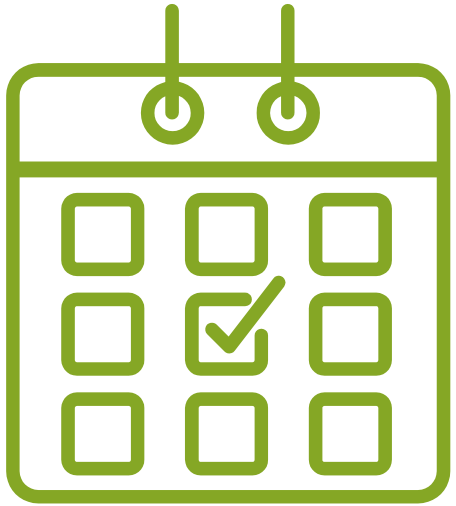
Machine Learning Types and Examples

Supervised	Classification models	Predict class membership • 0 = no, 1 = yes: Whether a person responded or not • Low, Medium, High: A customer's likeliness to respond
	Prediction models	Predict a number • $217.56: Total profit, expense, cost for a customer • 37: Number of months before a customer churns
Unsupervised	Clustering	Grouping together similar people, things, events • Transactions that are likely to be fraudulent • Customers that are likely to have similar behaviors
	Association	Affinity, or how frequently things occur together; sometimes in what order • Customers who purchase A also purchase B

Key Characteristics

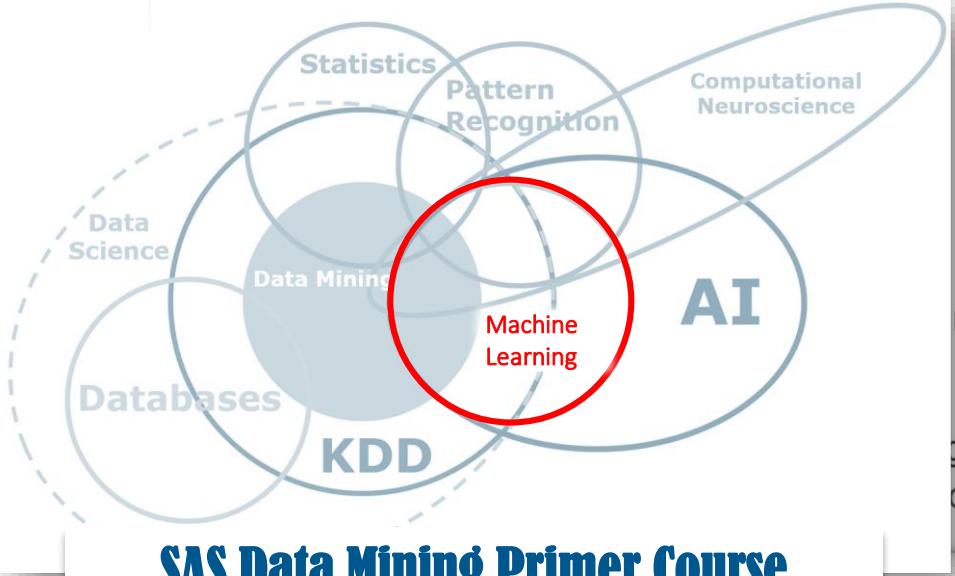
What's required to create valuable machine learning systems?

1. Data preparation capabilities
2. Algorithms – basic and advanced
3. Automation and iterative processes
4. Scalability
5. Workforce



PROC DISCRIM (K-nearest-neighbor discriminant analysis) – James Goodnight, SAS founder and CEO, 1979

- 1950's:
  - Samuel's checker program



SAS Data Mining Primer Course
SAS Institute, 1998

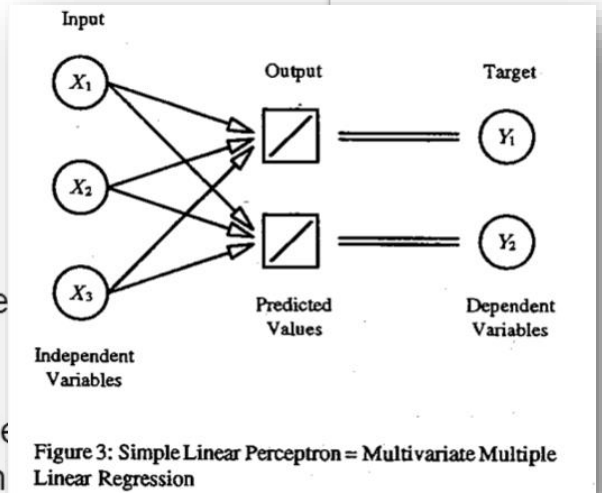


Figure 3: Simple Linear Perceptron = Multivariate Multiple Linear Regression

Neural Networks and Statistical Models,
SAS Institute, 1994

How Does Machine Learning Work?

Example Algorithms in SAS (STAT, EM, VS, VDMML)

SUPERVISED LEARNING

- Tree-based models (Decision Trees, Random Forests, Gradient Boosting)
- Regressions
- Neural Networks
- Support Vector Machines
- Recommender Systems

UNSUPERVISED LEARNING

- Clustering
- Principal Component Analysis
- Text Topic Detection

SEMI-SUPERVISED LEARNING

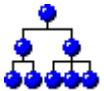
- Clustering
- Factorization Machines

How Does Machine Learning Work?

Not New for SAS

Machine Learning has been available in both SAS/STAT and Enterprise Miner for decades

Neural Networks



Decision Trees

Random Forests

Gradient Boosting

Text Analytics

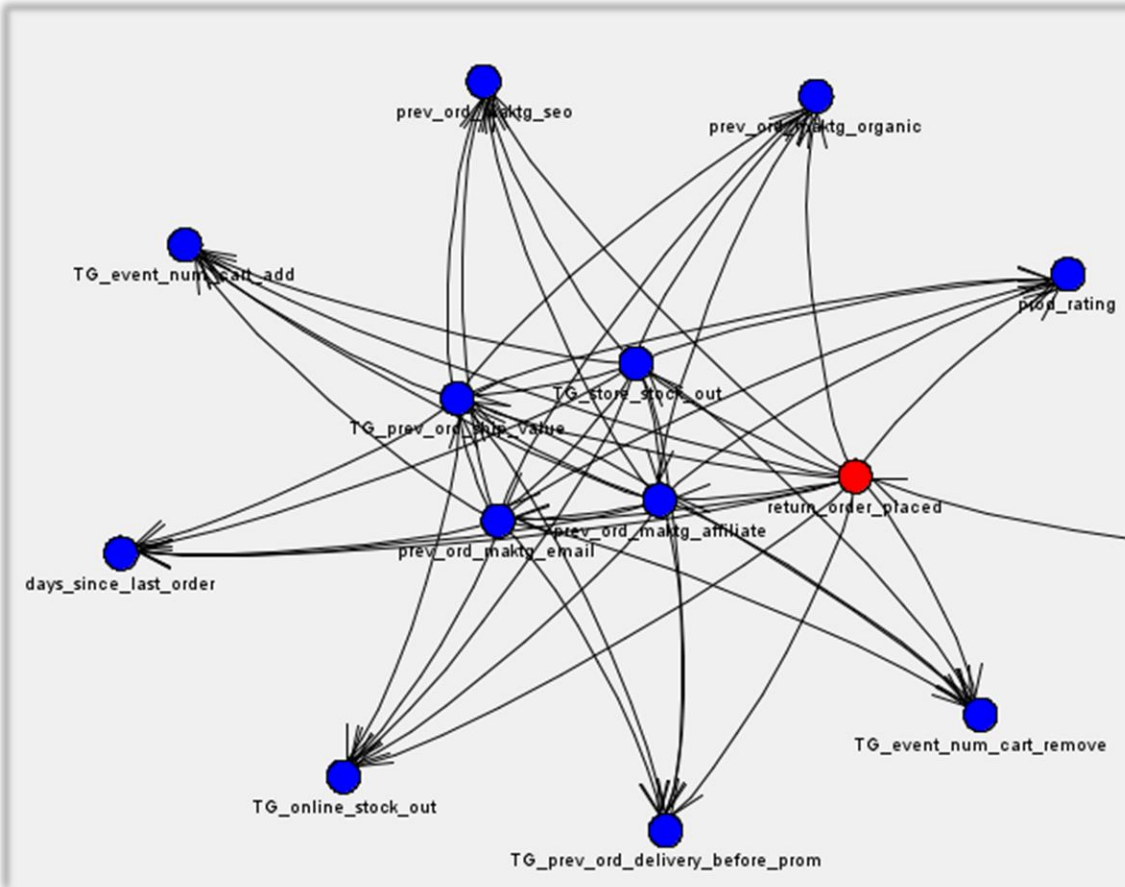
SAS Machine Learning

ALGORITHMS

- Neural networks
- Decision trees
- Random forests
- Associations and sequence discovery
- Gradient boosting and bagging
- Support vector machines
- Nearest-neighbor mapping
- k-means clustering
- Self-organizing maps
- Local search optimization techniques such as Genetic algorithms
- Regression
- Expectation maximization
- Multivariate adaptive regression splines
- Bayesian networks
- Factorization Machines
- Kernel density estimation
- Principal components analysis
- Singular value decomposition
- Gaussian mixture models
- Sequential covering rule building
- Model Ensembles
- And More…….

SAS Enterprise Miner

Algorithms – basic and advanced



SAS Enterprise Miner

- Linear & Logistic Regression
- Decision Trees
- Random Forest
- Gradient Boosting
- Support Vector Machines
- Neural Networks
- Clustering
- Bayesian Networks
- Principal Components
- Open Source Models

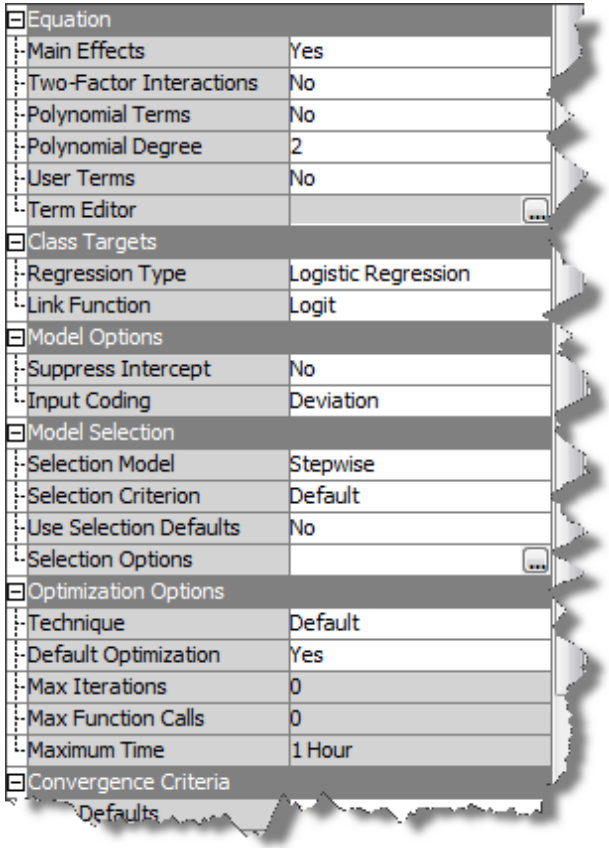
Machine Learning

Supervised Algorithms

Regression

- Linear
- Logistic
- Many Options

- Computes a forward stepwise least-squares regression
- Optionally computes all 2-way interactions of classification variables
- Optionally uses AOV16 variables to identify non-linear relationships between interval variables and the target variable.
- Optionally uses group variables to reduce the number of levels of classification variables.



Machine Learning Supervised Algorithms

Decision Tree

- Classify observations based on the values of nominal, binary, or ordinal targets
- Predict outcomes for interval targets
- Easy to interpret
- Interactive Trees available

Train	
Variables	
Interactive	
Use Frozen Tree	No
Use Multiple Targets	No
Precision	4
Splitting Rule	
Interval Criterion	ProbF
Nominal Criterion	ProbChisq
Ordinal Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Split Precision	4
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Decision
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	No

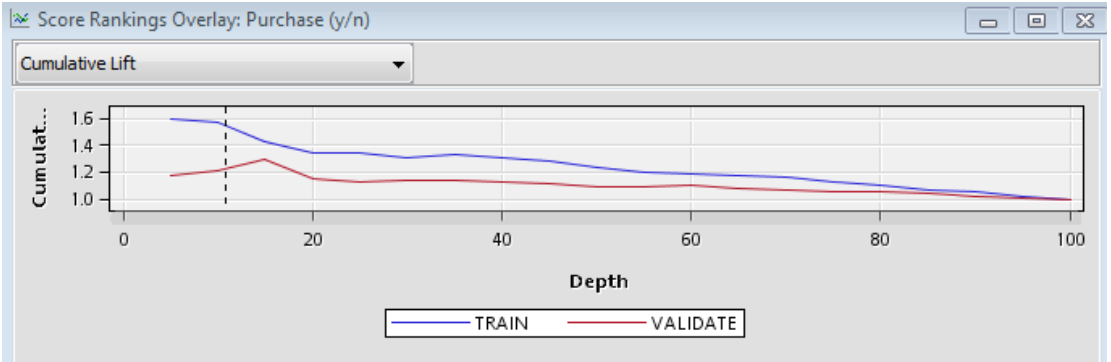
Machine Learning Supervised Algorithms

Gradient Boosting

- Sequential ensemble of many trees
- Extremely good predictions
- Very effective at variable selection

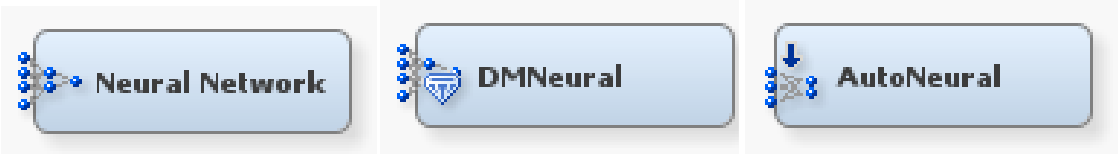
Variable Importance

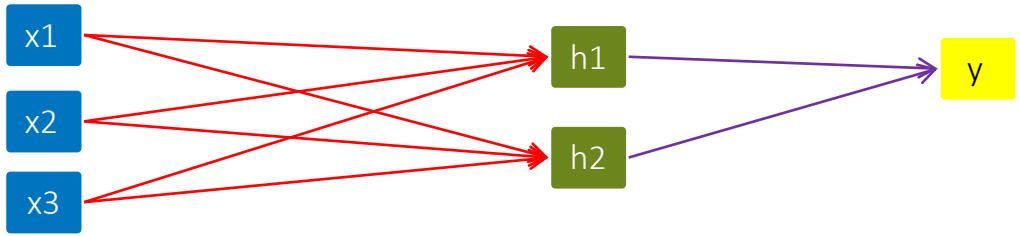
Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
RECENCY	Recency	4	1	0.656234	0.656234
STATECOD	State Code	4	0.944464	0	0
LEISURE	Leisure Prod.	1	0.537544	0	0
DOMESTIC	Domestic Prod.	1	0.491743	1	2.033584
APPAREL	Apparel Purch.	1	0.330797	0.518451	1.567281
AGE	Age	0	0	0	.
COATS	Coats Purch.	0	0	0	



Machine Learning Supervised Algorithms

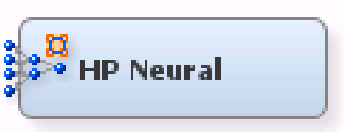
Neural Networks



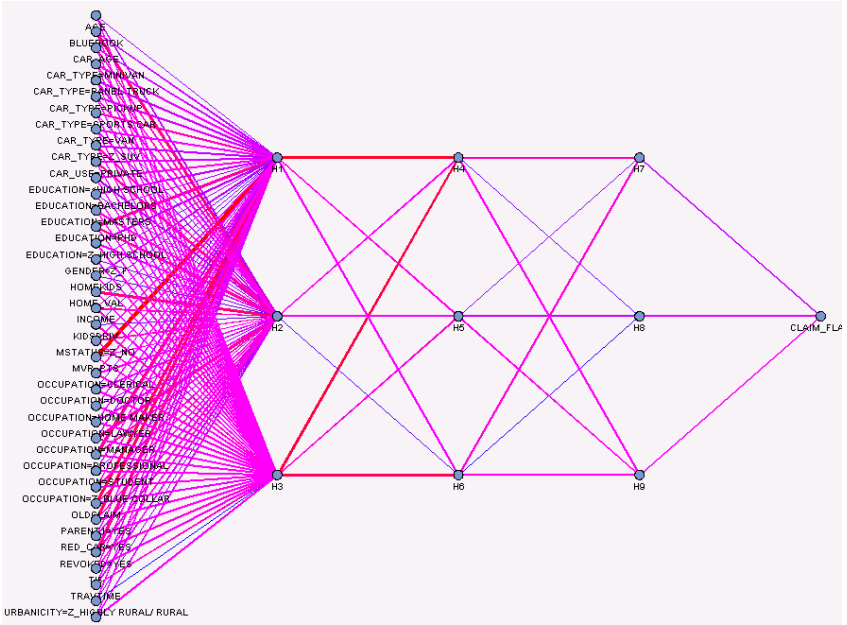


- Non-linear relationship between inputs and output
- Prediction more important than ease of explaining model
- Requires a lot of training data

Neural Network

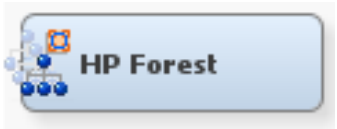


- Provides a User-Defined Architecture, which gives users more control over the construction of the neural network.
- Users can specify the number of hidden layers, the number of hidden neurons, and associated activation functions for each layer.
- Users can configure Input and Target Standardizations, Target Error, and Activation Functions.

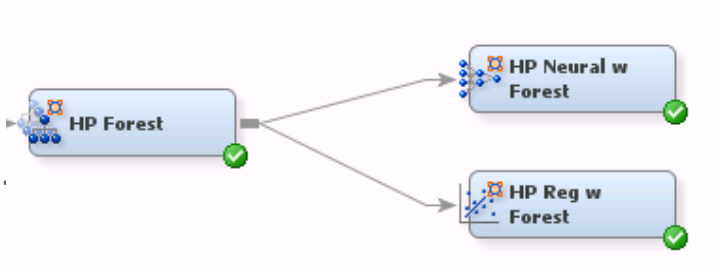


Machine Learning Supervised Algorithms

Random Forest

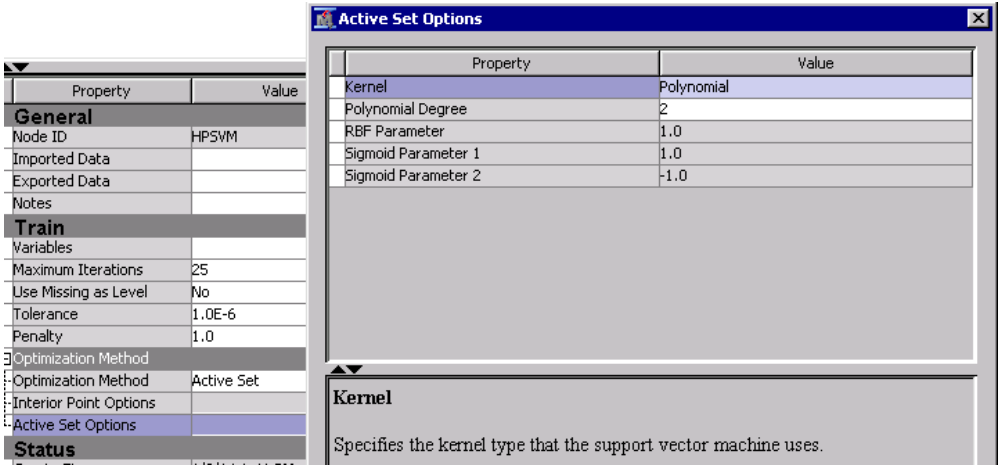


- Creates a predictive model called a forest.
- Consists of several decision trees that differ from each other in 2 ways.
  1. Training data for the tree is a sample w/o replacement.
  2. The input variables that are considered for splitting a node are randomly selected from all available inputs.
- Also allows users to optionally perform variable selection based on either OOB Average Error for interval targets or OOB marginal reduction for class targets.



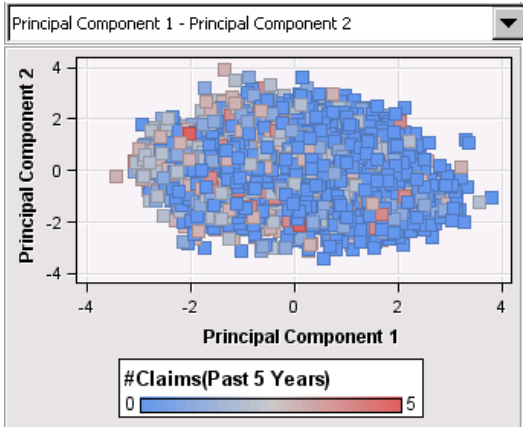
Support Vector Machines

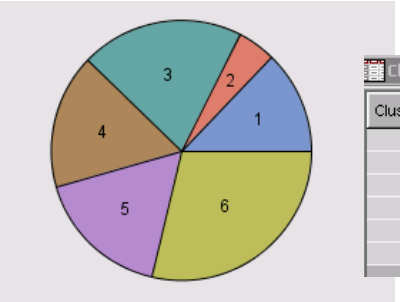
- Enables the creation of linear and nonlinear support vector machine models
- Constructs separating hyperplanes that maximize the margin between two classes
- Enables use of a variety of kernels: linear, polynomial, radial basis function, and sigmoid function. The node also provides interior point and active set optimization methods.



Principal Components and Cluster Nodes

- Perform principal component analysis for data dimension reduction, a frequent intermediate step in the data mining process
- Perform *k*-means clustering analysis in threaded and distributed computing environments, using numeric interval variables as inputs





Cluster Centroids

Cluster ID	AGE	BLUEBOOK	CAR_AGE	HOMEKIDS	H
1	47.7153	20525.15	11.38766	0.503249	
2	44.57949	16567.28	8.746701	0.692308	
3	46.33524	16717.82	9.018731	0.580176	
4	45.41172	14163.27	7.104052	0.719665	
5	43.00772	12547.33	6.02052	0.943684	
6	43.23257	14626.02	7.922706	0.797468	

New in SAS® Enterprise Miner™ 13.1

Machine Learning Supervised and Unsupervised Algorithms

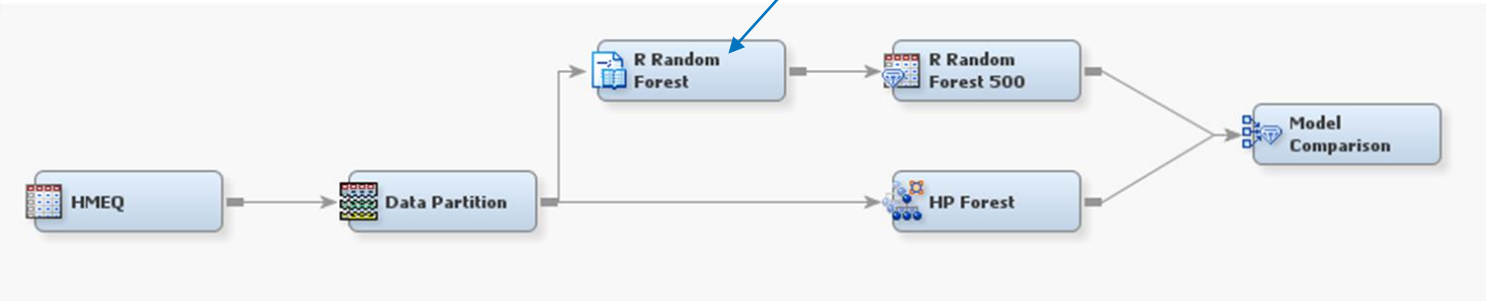
Open Source

- Integrate R code inside Enterprise Miner

Notable Features

- Model Comparison
- Ensemble
- Score Code Generation

```
library(randomForest)

&EMR_MODEL <- randomForest(&EMR_CLASS_TARGET ~  &EMR_CLASS_INPUT + &EMR_NUM_INPUT, ntree= 500, mtry= 5, data= &EMR_IMPORT_DATA, importance= TRUE)

&EMR_EXPORT_TRAIN <- predict(&EMR_MODEL, &EMR_IMPORT_DATA, type="prob")


&EMR_EXPORT_VALIDATE <- predict(&EMR_MODEL, &EMR_IMPORT_VALIDATE, type="prob")
&EMR_EXPORT_TEST <- predict(&EMR_MODEL, &EMR_IMPORT_TEST, type="prob")

&EMR_EXPORT_TRAIN[1:10,]

png("EMR_forestMsePlot.png")
plot(&EMR_MODEL, main= 'randomForest MSE Plot')
dev.off()

write.table(round(importance(&EMR_MODEL),2), file= "EMR_forestImportance.csv", sep= ",", row.names= TRUE, col.names= TRUE)

print(&EMR_MODEL)
  round(importance(&EMR_MODEL),2)
```

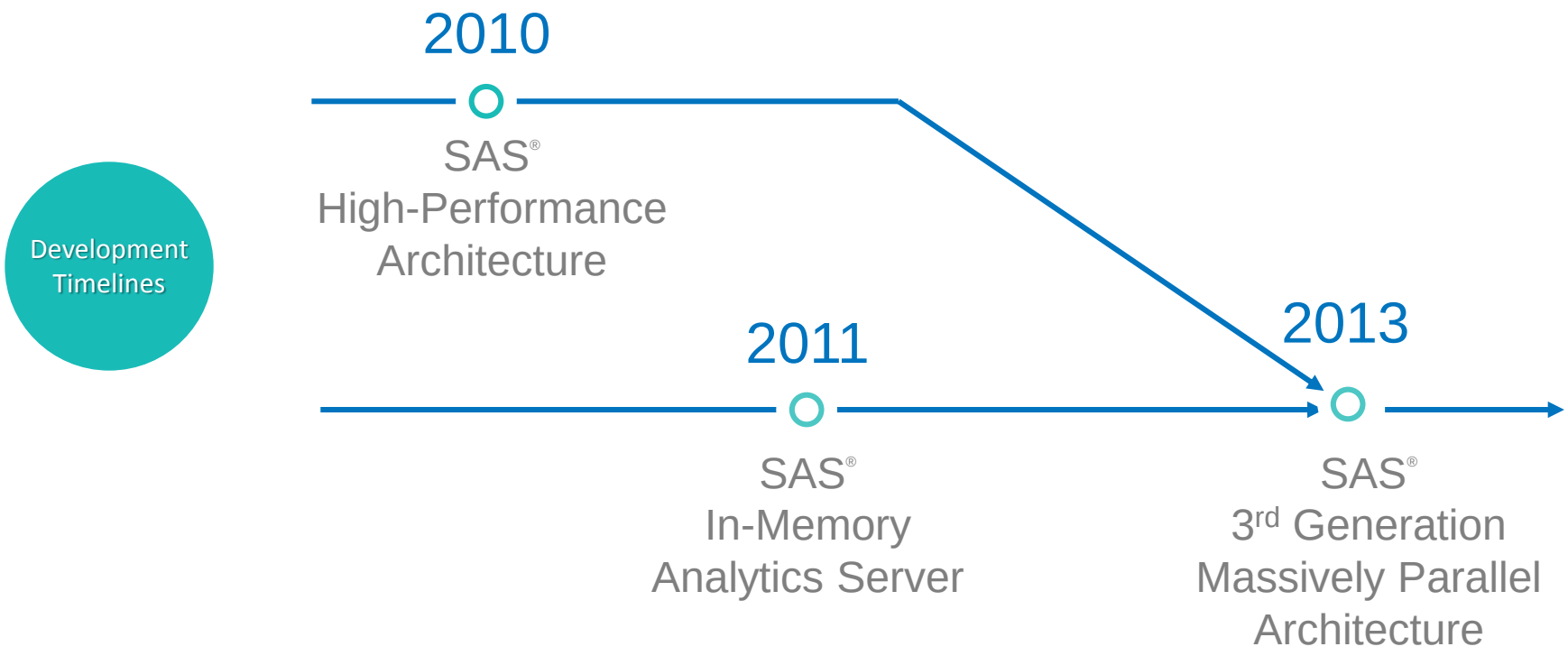


SAS Enterprise Miner

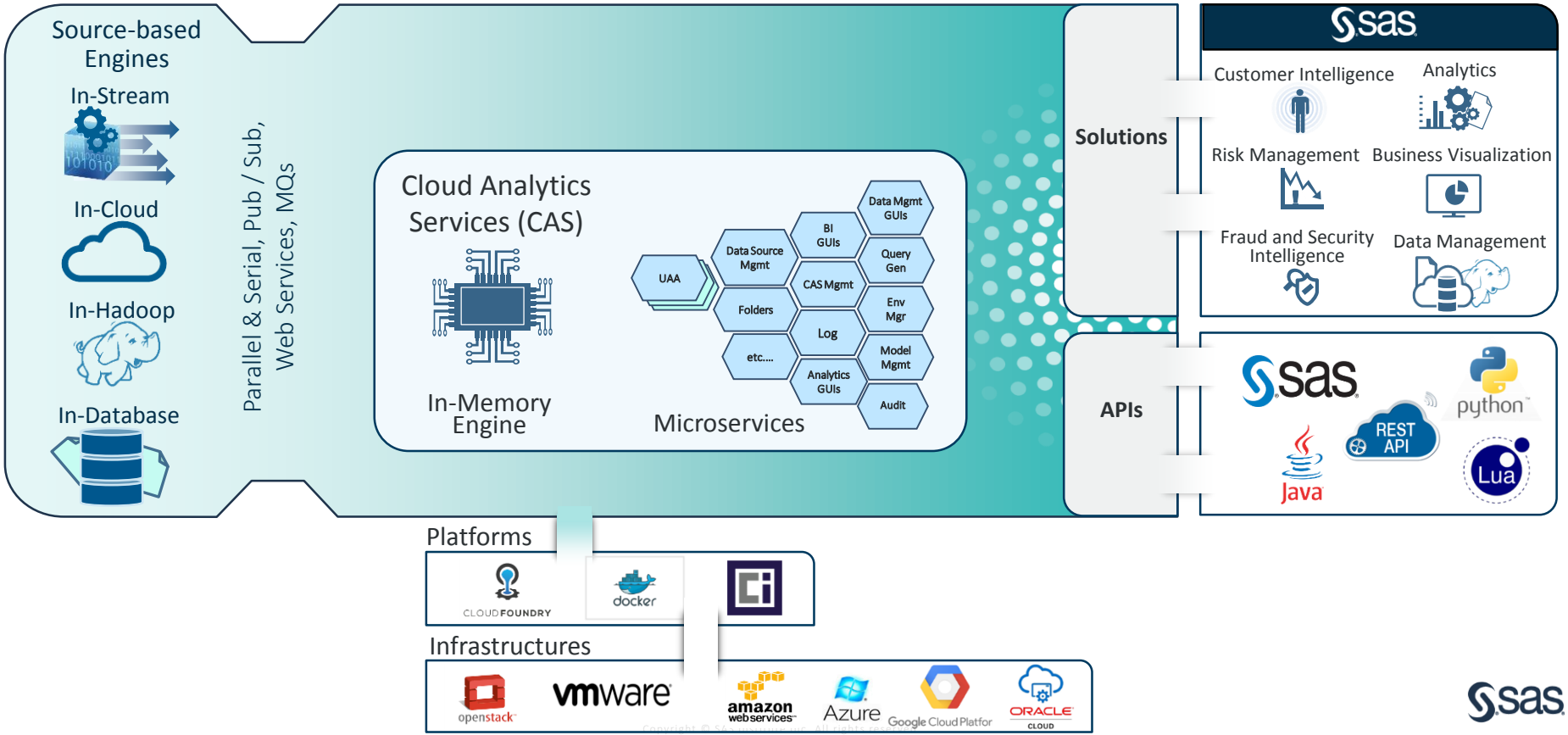
Machine Learning Demo

SAS Viya

Unifying Our Architecture



Platform Architecture



Introduction: Software Overview

SAS Viya Increases Openness of SAS Platform

Visual Interfaces

Programming Interfaces

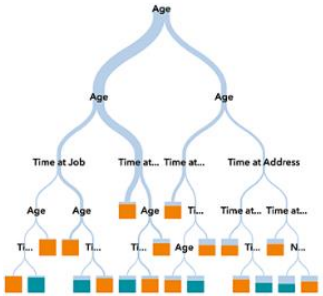
python™ sas Java

Future Plans:

R

API Interfaces

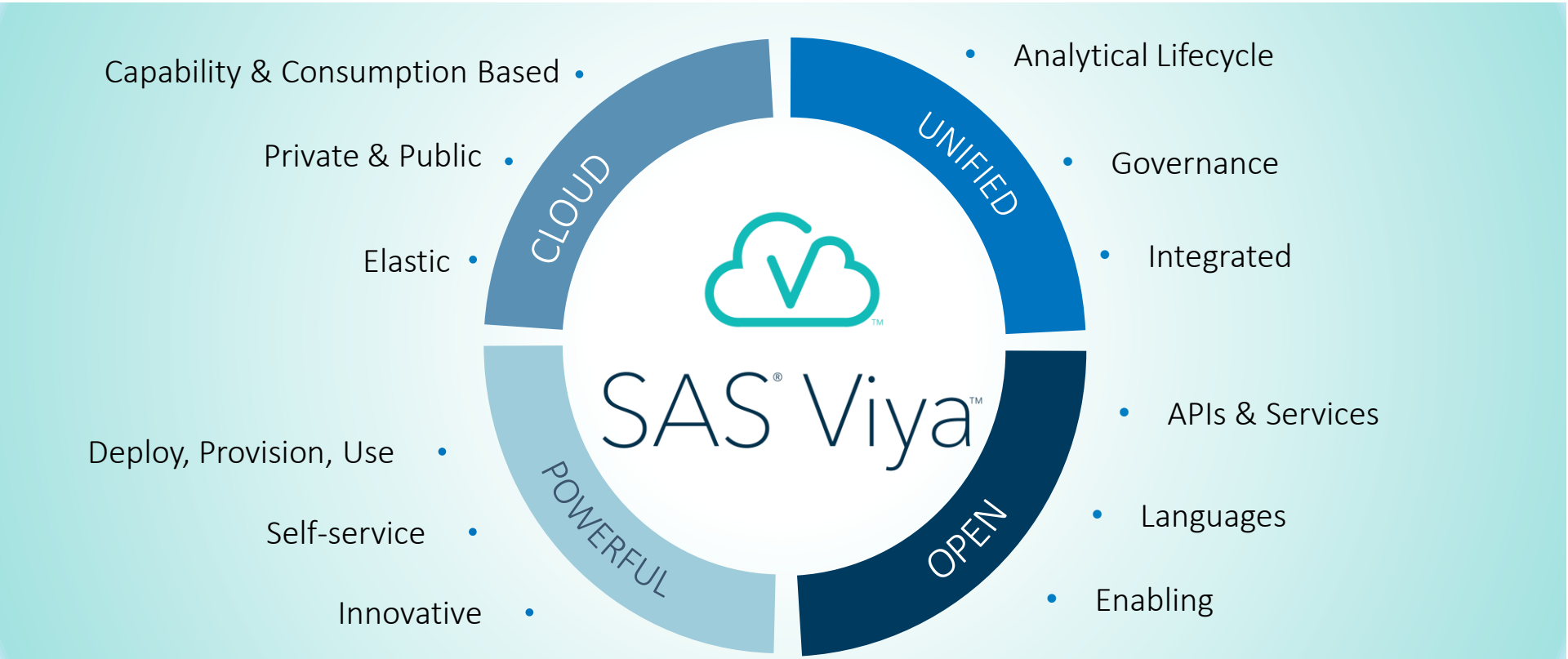
REST API



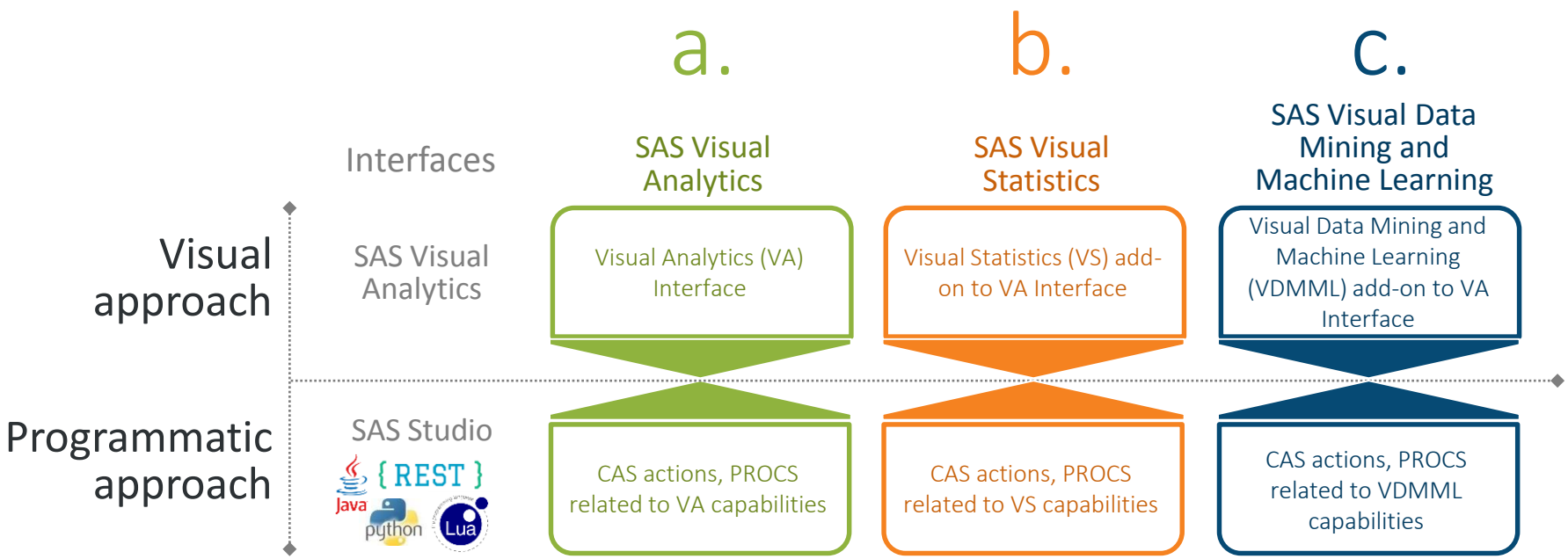
SAS® Viya™

sas

SAS® viya™ new analytic architecture

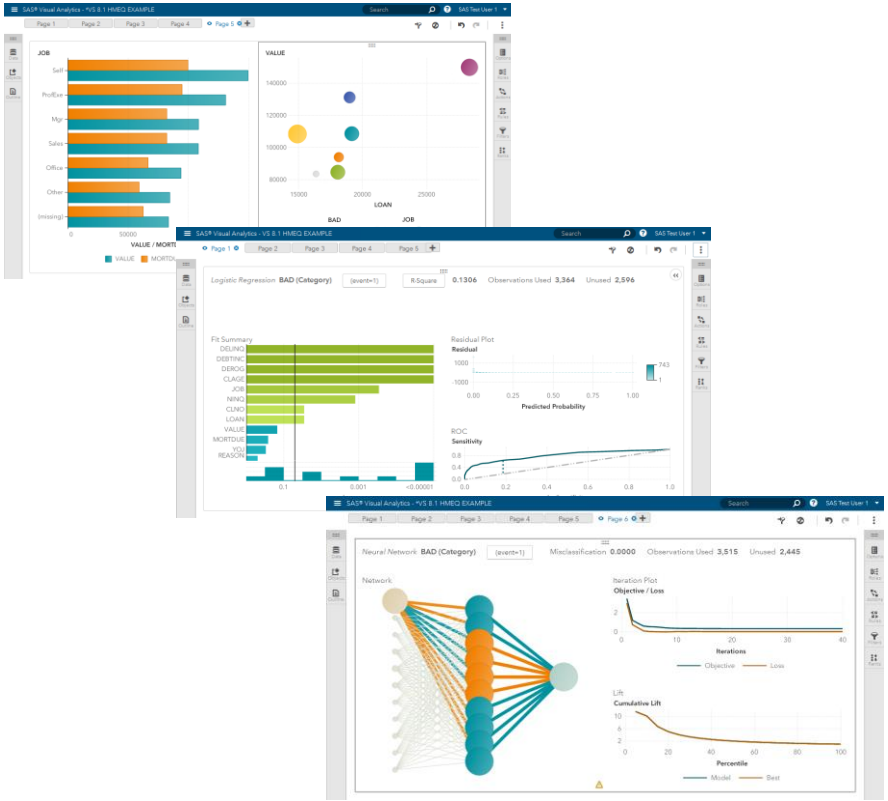


SAS® Viya™ Based Product offerings



Introduction: Software Overview

Comparing VA, VS, and VDMML



SAS Visual Analytics 8.1

- Explore data and discovery relationships
- Examine distributions and summary statistics
- Post-model analysis and reporting

SAS Visual Statistics 8.1

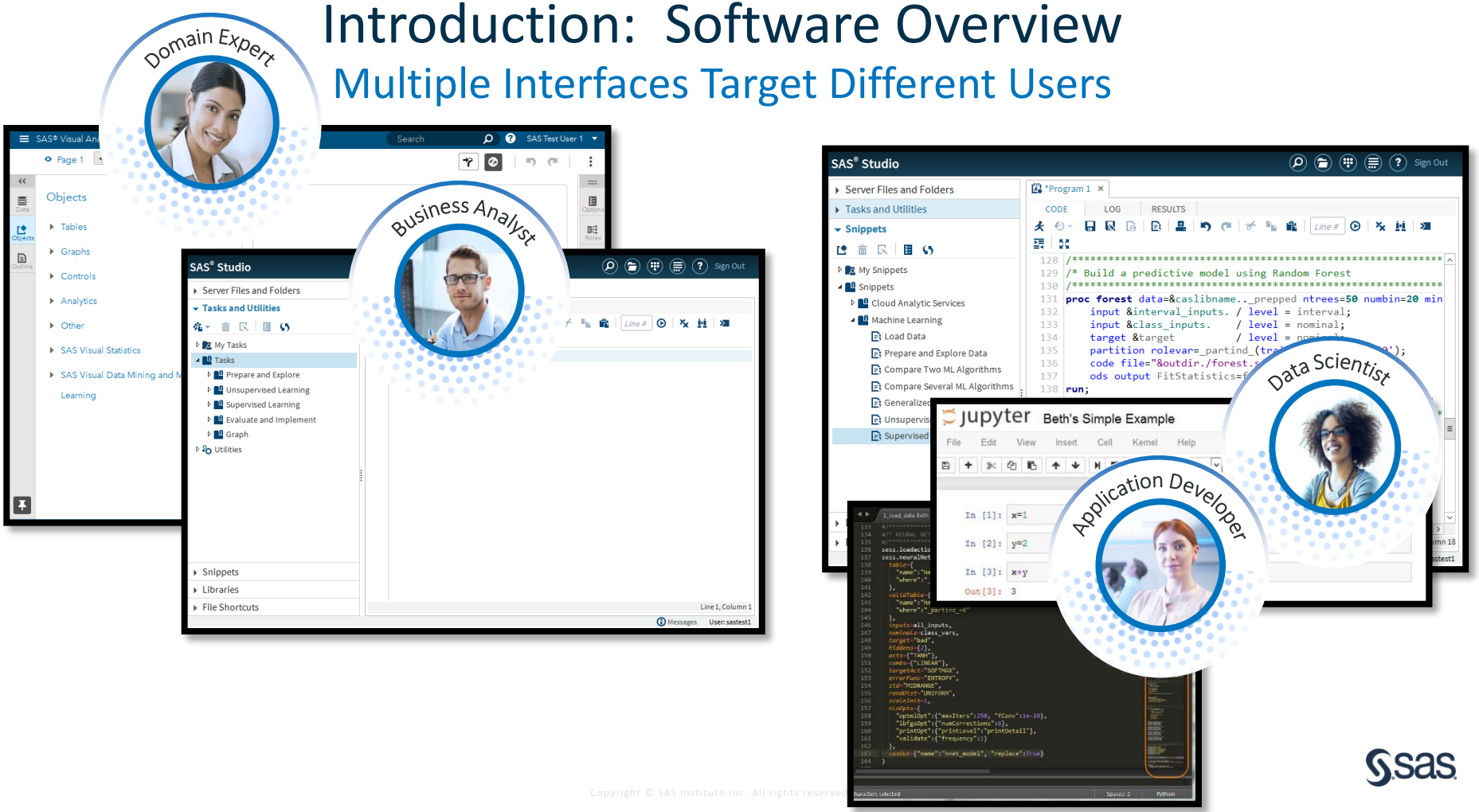
- Build unsupervised and supervised learning models
- Interactively refine candidate models
- Compare models and generate score code

SAS VDMML 8.1

- Five additional machine learning models

Introduction: Software Overview

Multiple Interfaces Target Different Users

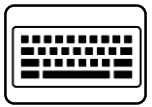


Introduction: Software Overview

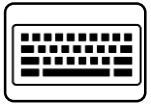
Building a New Model

- There are many ways to build a model in SAS VS/VDMML:

1. In the reporting interface you can:
   a. Build a model from a Visual Analytics graph
   b. Build a model from scratch using the reporting interface



2. In the SAS Studio interface you can:
   a. Build a model using SAS Studio Tasks
   b. Bring in SAS Studio Snippets
   c. Type in code from scratch

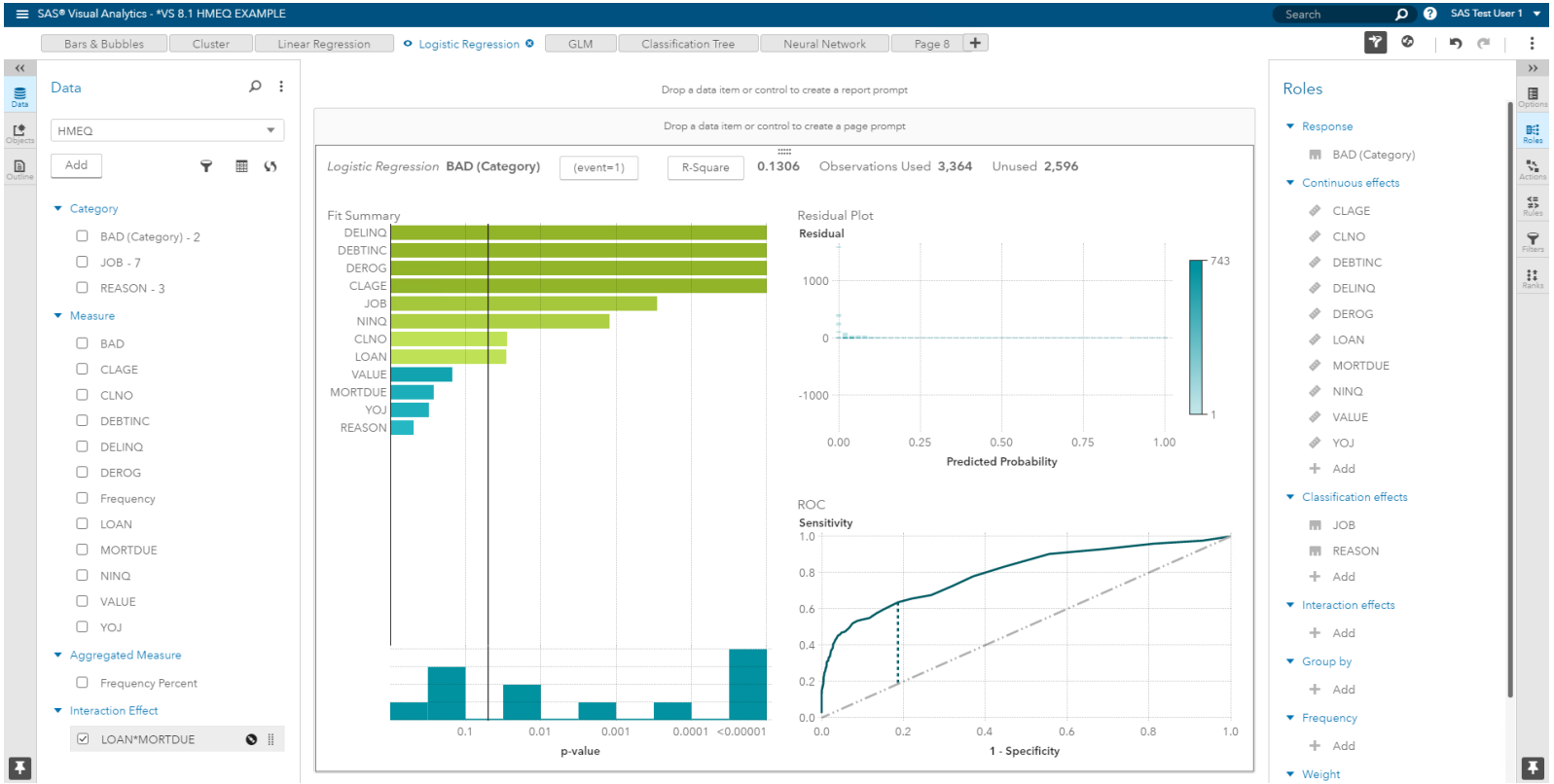


3. You can also program in Python, Java or Lua to run CAS actions

4. You can use Enterprise Miner to build models and run them in CAS from SAS 9

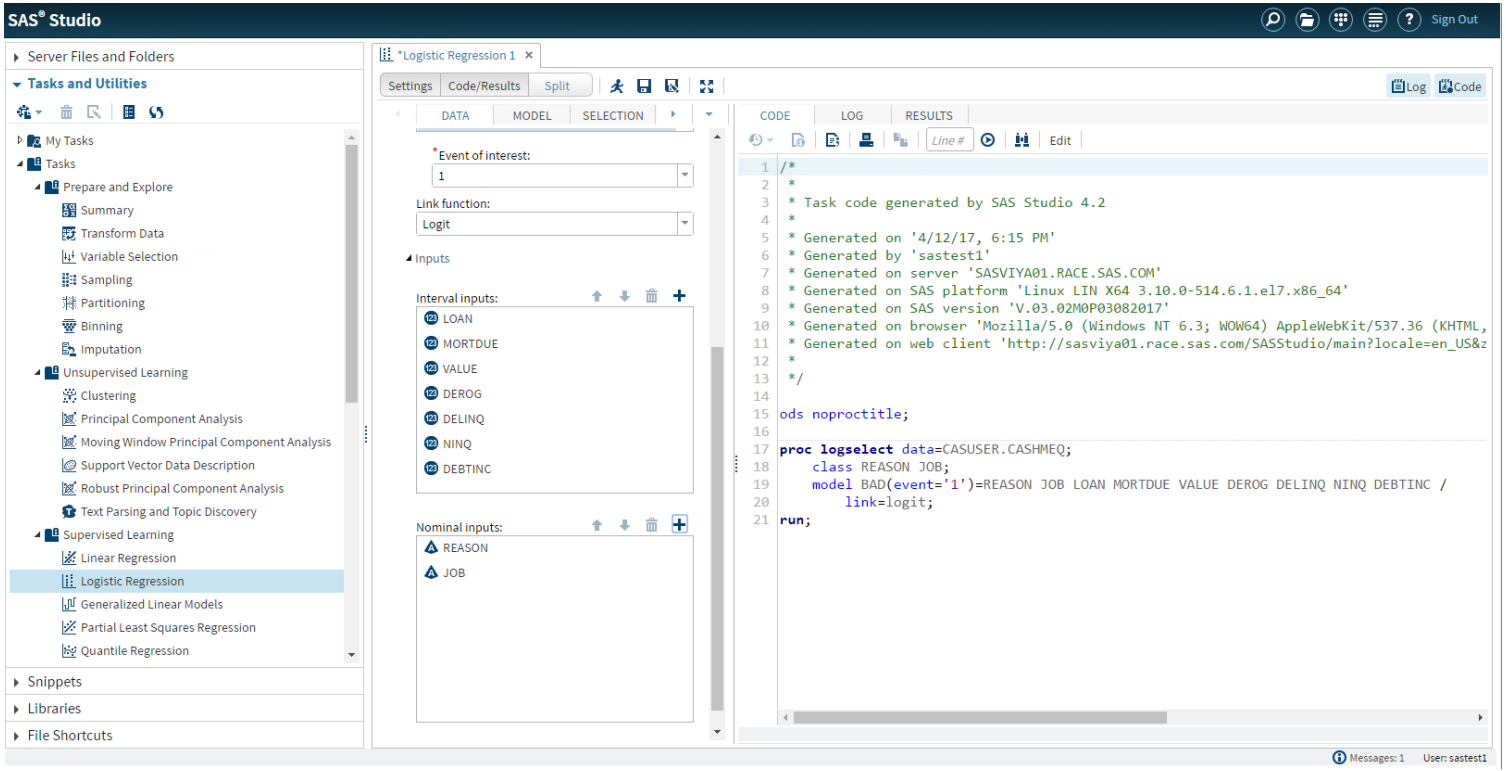
Interfaces

Building a Model from Scratch in the Reporting Interface



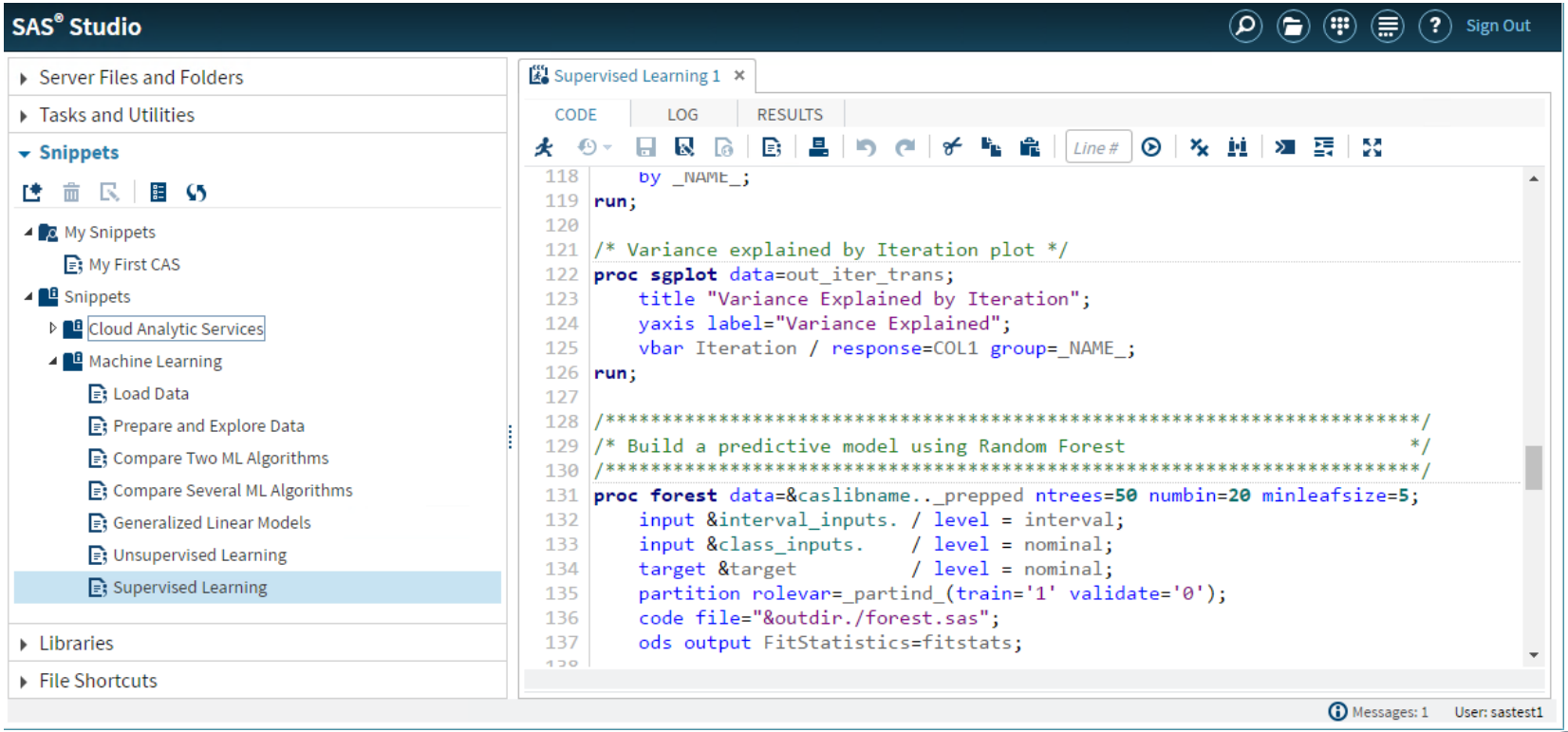
Interfaces

Building a Model Using SAS Studio Tasks



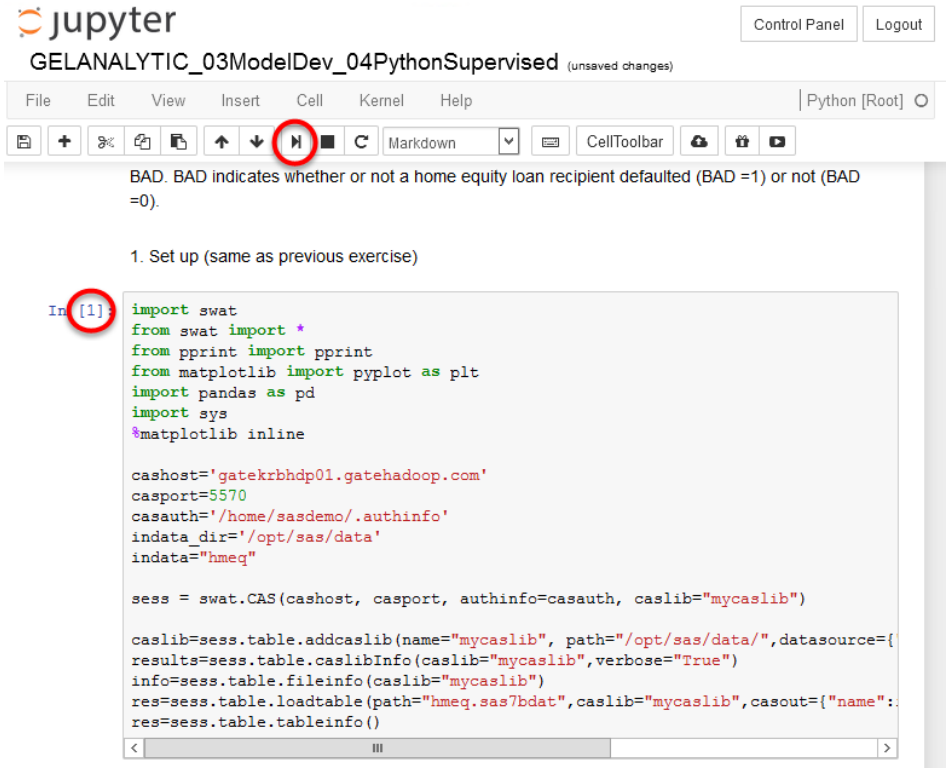
Interfaces

Building a Model Using SAS Studio Snippets



Interfaces

Building a Model Using Python



Visual Statistics 8.1

VDMML 8.1

Intro: Software Overview

Summary of Procedures

VA GUI Object

SAS Studio Task / Code

1 Data Wrangling

CARDINALITY	Variable cardinality analysis
BINNING	Variable binning
VARIMPUTE	Missing value imputation
VARREDUCE	Variable selection
PARTITION	Sampling and partitioning
MDSUMMARY	Basic Descriptive Statistics
FEDSQL	Run SQL query in CAS
TEXTMINE	Text mining
TMSCORE	Scoring of new text docs

2 Unsupervised Learning

KCLUS	*k*-means & *k*-modes clustering
PCA	Principal component analysis
SVDD	Support Vector Data Description
RPCA	Robust PCA
MWPCA	Moving Windows PCA

3 Supervised Learning

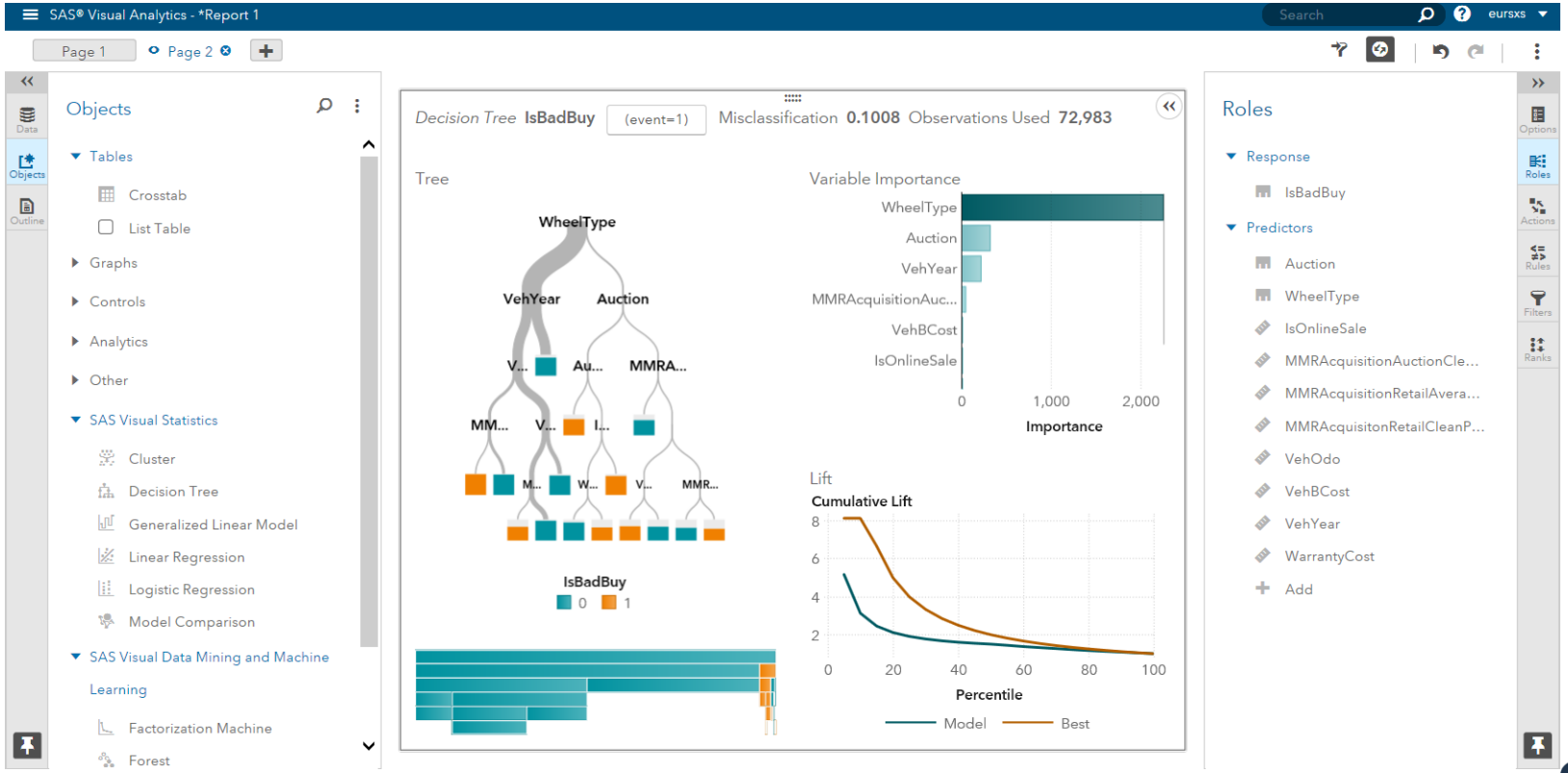
REGSELECT	Ordinary least squares models	Regression
LOGSELECT	Logistic regression models	Regression
GENSELECT	Generalized linear models	Regression
NLMOD	Nonlinear regression models	Regression
PLSMOD	Partial least square models	Regression
QTRSELECT	Quantile regression models	Regression
TREESPLIT	Decision tree models	
FOREST	Random forest models	Tree Based
GRADBOOST	Gradient boosting models	Tree Based
NNET	Neural network models	
SVMACHINE	Support vector machine models	
FACTMAC*	Factorization machine models	

4 Model Assessment

ASSESS	Assess supervised models

SAS Visual Statistics & Machine Learning

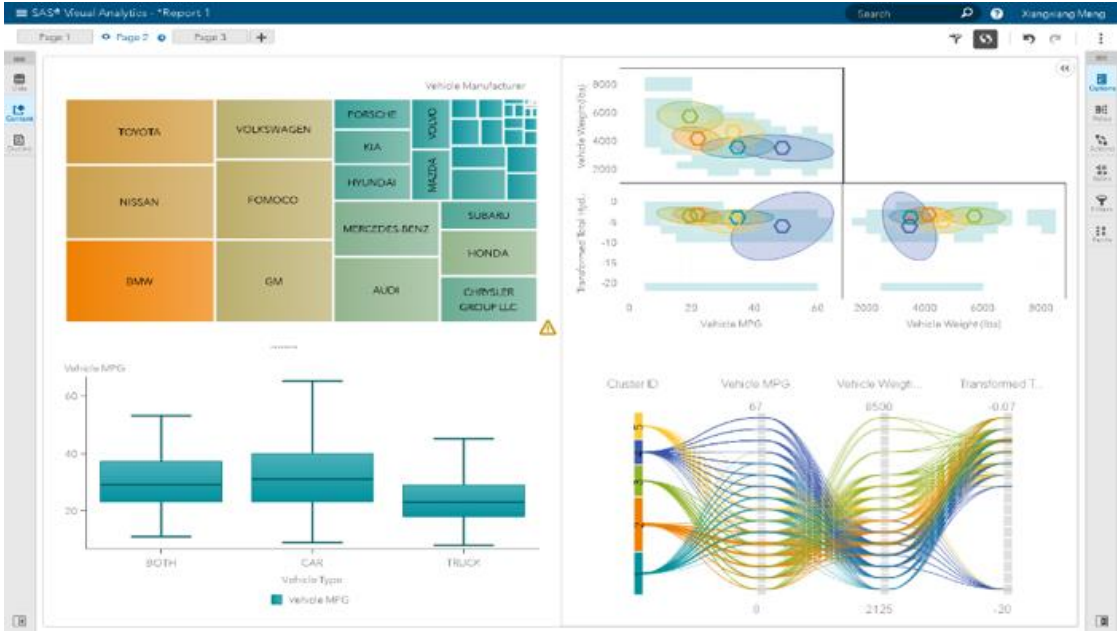
Visual Interface



Introduction: Software Overview

Key Features of SAS® Visual Statistics 8.1

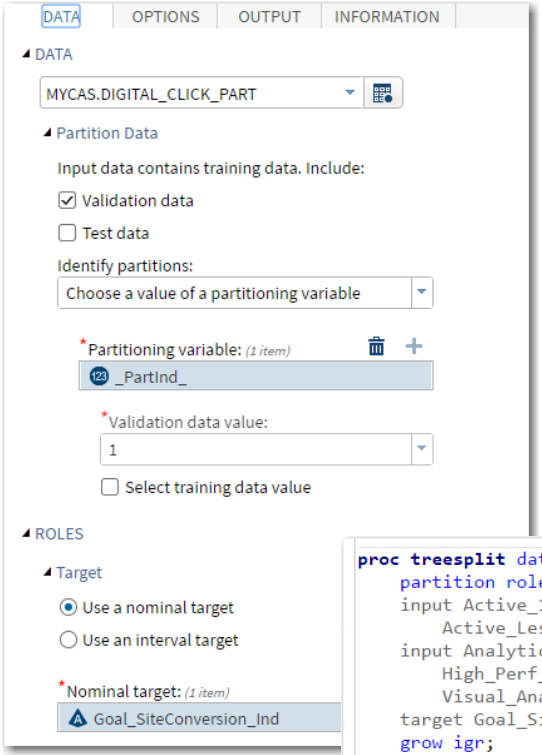
- Modeling Techniques
- (Reporting Interface)
  - Clustering (k-means)
  - Linear Regression
  - Logistic Regression
  - GLM Regression
  - Decision Trees
- Common Features
  - Training-validation partitioning
  - Variable Importance / Profile
  - Model Assessment
  - Model comparison
  - Derivation of predictive outputs
  - Ability to export model statistics into Excel
  - Score Code

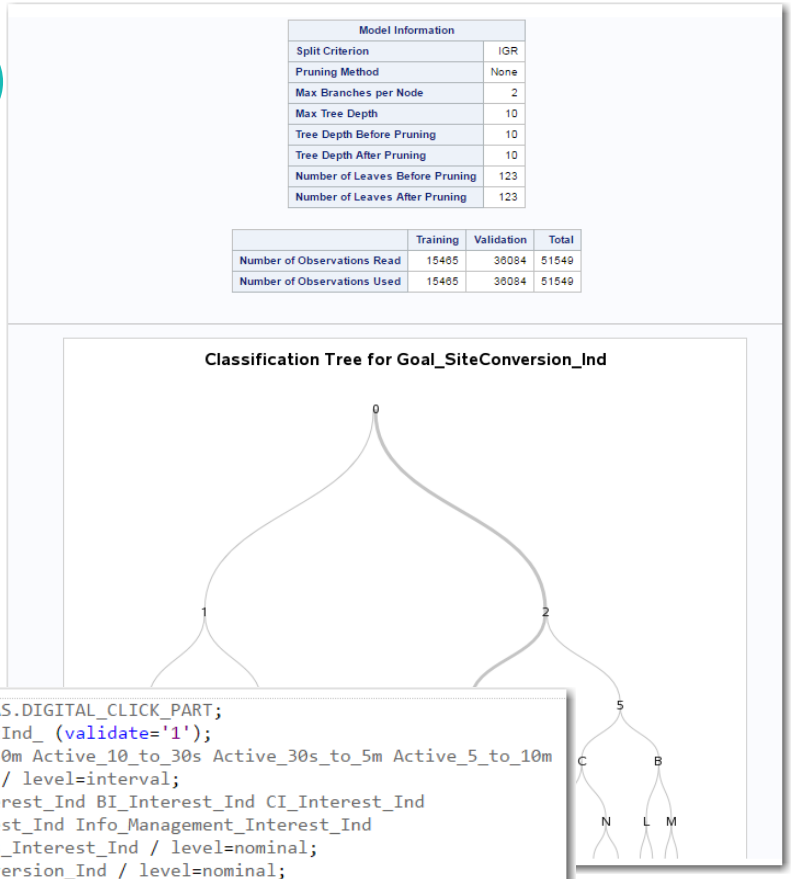


SAS® Visual Statistics 8.1

Decision Tree (PROC TREESPLIT)

- Task included.
- Allows for autotuning.
- Does not allow for reusing predictors.
- Ability to decide splitting criterion.
- Save variable importance table.
- Create scored table and generate score code.





Model Information	
Split Criterion	IGR
Pruning Method	None
Max Branches per Node	2
Max Tree Depth	10
Tree Depth Before Pruning	10
Tree Depth After Pruning	10
Number of Leaves Before Pruning	123
Number of Leaves After Pruning	123

	Training	Validation	Total
Number of Observations Read	15465	36084	51549
Number of Observations Used	15465	36084	51549


```
proc treesplit data=MYCAS.DIGITAL_CLICK_PART;
    partition role=_PartInd_ (validate='1');
    input Active_10_to_30m Active_10_to_30s Active_30s_to_5m Active_5_to_10m
        Active_Less_10s / level=interval;
    input Analytics_Interest_Ind BI_Interest_Ind CI_Interest_Ind
        High_Perf_Interest_Ind Info_Management_Interest_Ind
        Visual_Analytics_Interest_Ind / level=nominal;
    target Goal_SiteConversion_Ind / level=nominal;
    grow igr;
    prune none;
run;
```

Link to documentation

SAS® Visual Data Mining and Machine Learning 8.1

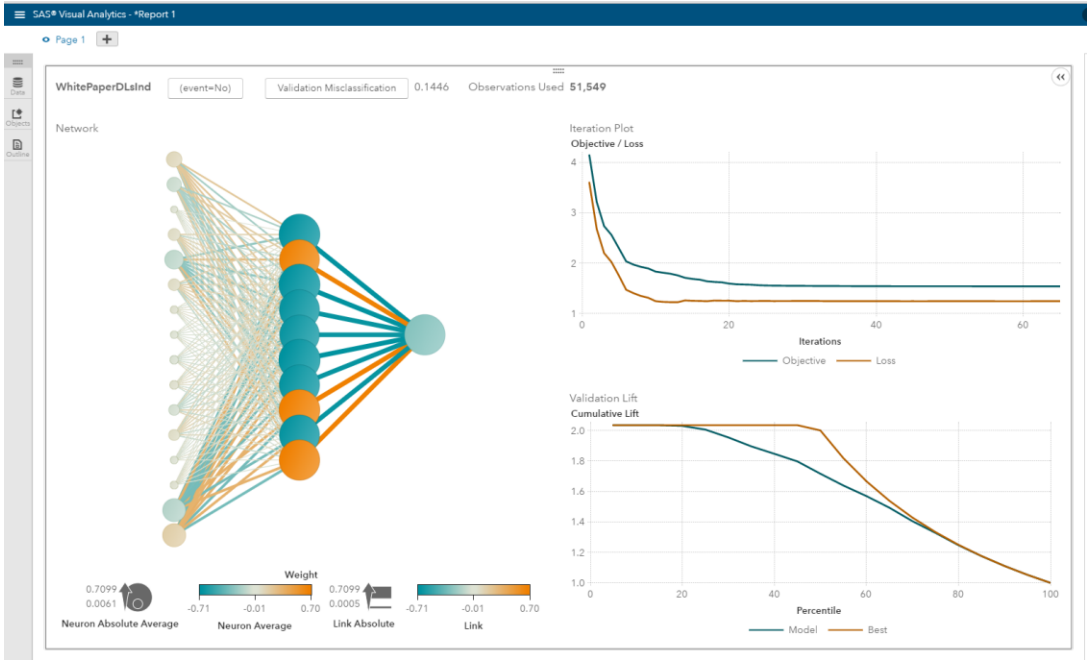
Visual Interface

Machine Learning Techniques

- Forest
- Factorization Machine
- Gradient Boosting
- Neural Network
- Support Vector Machine

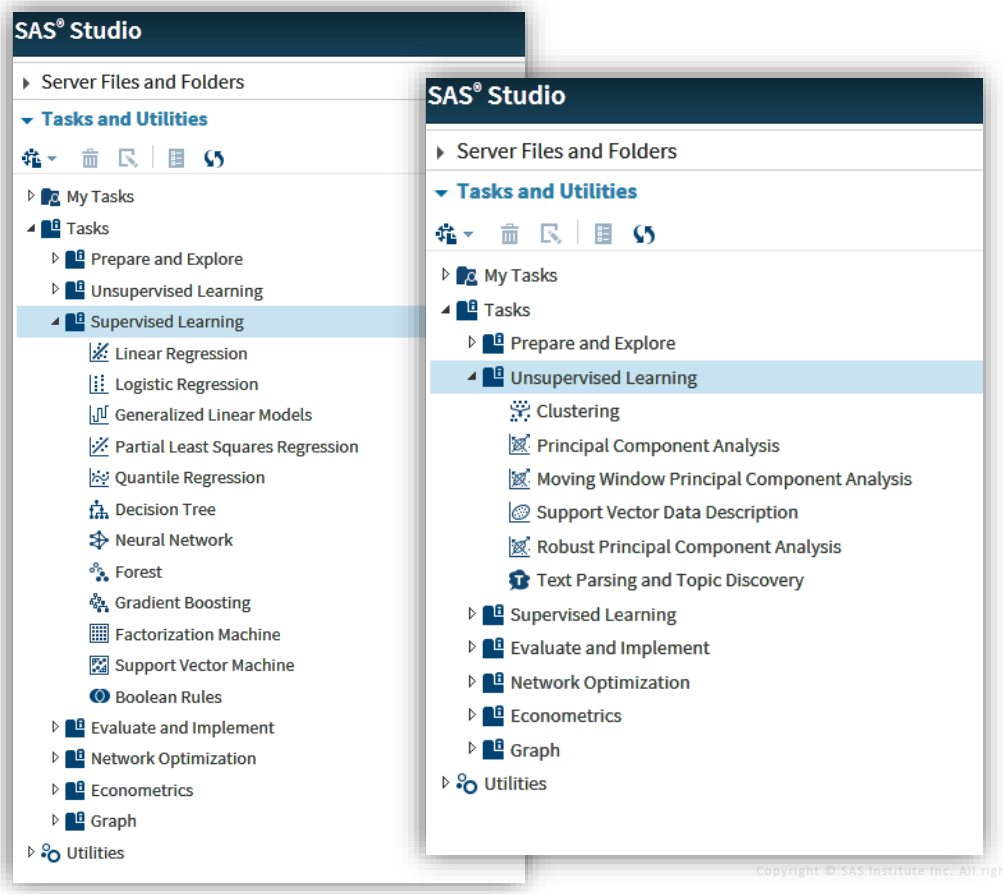
Common Features

- Training-validation*
- Auto-tuning
- Model Assessment
- Score Code or Astore Table
- Model comparison
- Ability to export model statistics into Excel



SAS Visual Data Mining and Machine Learning 8.1

Tasks in SAS Studio



- Unsupervised Learning
  - Moving Window PCA
  - Robust PCA
  - Support Vector Data Description
  - Text Parsing and Topic Discovery
- Supervised Learning
  - Partial Least Square Regression
  - Quantile Regression
  - Boolean Rules

Supervised Learning: VDMML

Common Features

- All five VDMML models have
  - Training and validation partitioning (except FACTMAC)
  - Autotuning
  - Score Code (Neural Network) OR Astore Table (Other four models)
  - Model Assessment
  - Model Comparison
  - Ability to Export Model Statistics into Excel

Automating

Methods

- VDMML offers:
  - Random search highly parallelizable)
  - Latin Hypercube (highly parallelizable)
  - LH + proprietary SAS/OR algorithm (sequential in nature)
- All 5 VDMML procs as well as Decision Trees can be autotuned via SAS Studio interface

SAS Viya

Machine Learning Demo

How does SAS support Machine Learning?

Agenda

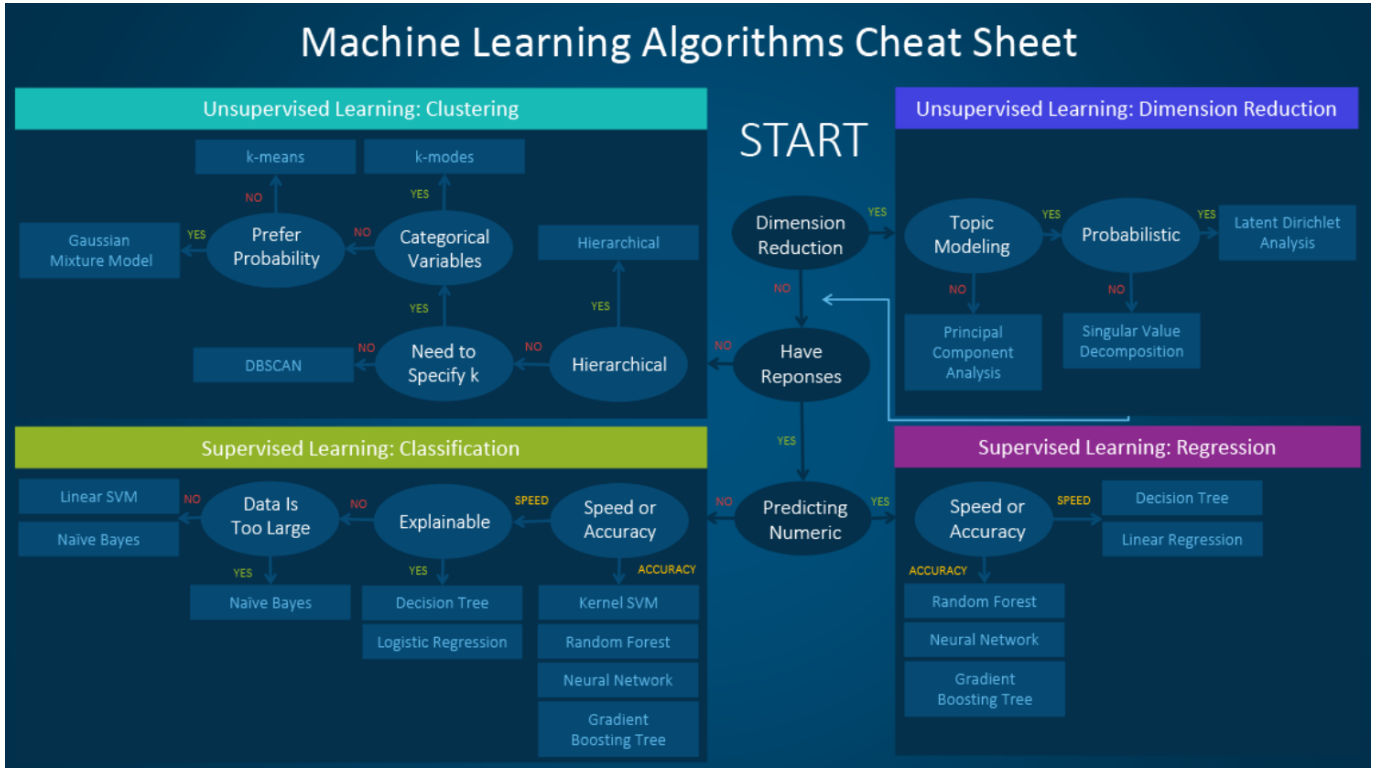
- What is Machine Learning?
- Terminology and key characteristics
- How you can use machine learning in SAS
- Examples in Enterprise Miner
- Examples in SAS Viya

Resources

Where I can I learn more?

Model Selection

Machine Learning Algorithms Cheat Sheet



Access Here

Recommended Resources

Hungry for more?

Video - Automated Machine Learning at Scale
http://www.sas.com/en_us/webinars/automated-machine-learning-scale.html

Machine learning - what it is and why it matters (reading)
http://www.sas.com/en_us/insights/analytics/machine-learning.html

Live web and classroom training - Big Data, Data Mining, and Machine Learning
Big Data course

Machine learning
Public Customer references

- DIRECT TV http://www.sas.com/en_us/customers/directv.html
- DSW http://www.sas.com/en_us/customers/dsw.html
- CNA http://www.sas.com/en_us/customers/cna.html
- Octo Telematics http://www.sas.com/en_us/news/press-releases/2015/february/insurance-analytics-octo.html
- Allianz http://www.sas.com/en_us/customers/allianz-fraud-management.html
- Blue Cross And Blue Shield of North Carolina http://www.sas.com/success/bcbsnc_readmission.html

SAS® Visual Data Mining and Machine Learning

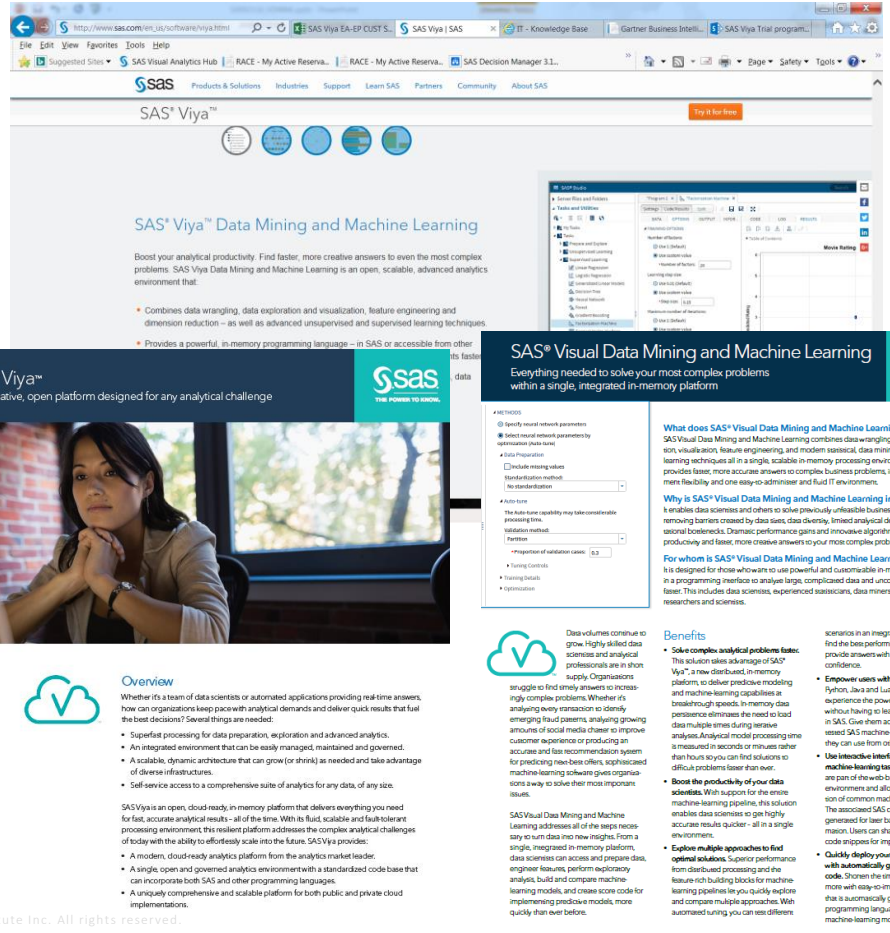
Key Resources

- SAS VDMML Product Web Page
- Factsheet
- SAS Viya Brochure
- Documentation
- VDMML Community
- Videos

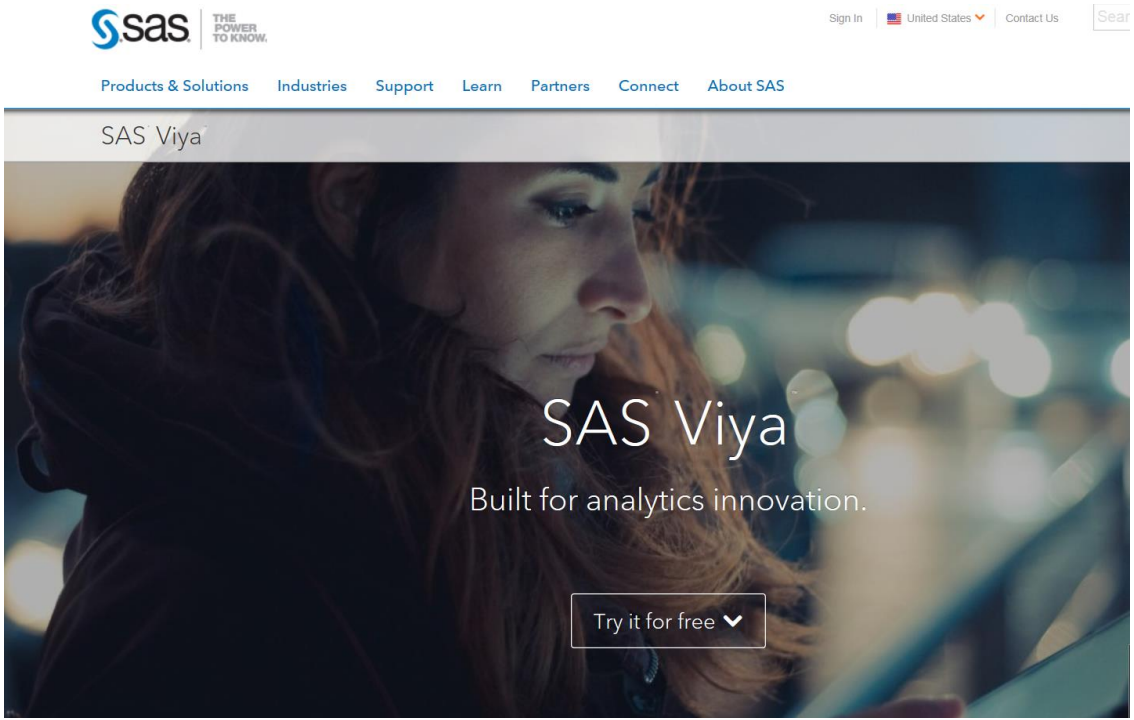
SAS Visual Data Mining and Machine Learning
https://youtu.be/X0AU4gDUc_Y

SAS Visual Data Mining and Machine Learning with Python
https://youtu.be/LXoikPWQJ3o

'How do I' videos
http://support.sas.com/training/tutorial/viya/index.html



VIYA Trial



https://www.sas.com/en_us/software/viya.html

Questions?

Thank you for your time and attention!

sas.com

CUSTOMER LOYALTY TEAM • Support You Can Count On

THE POWER TO KNOW.