SAS® FORUM
UNITED KINGDOM 2015

Deep Dive into High Performance Machine Learning Procedures

Tuba Islam, Analytics CoE, SAS UK

sas THE POWER TO KNOW.

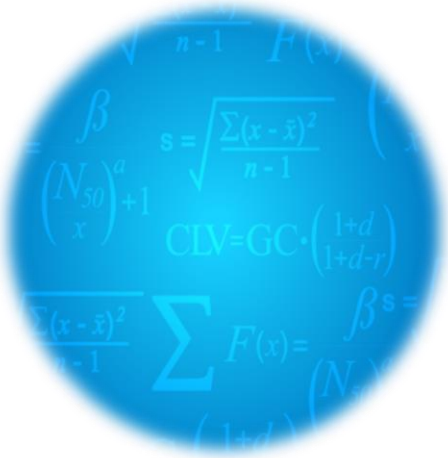
WHAT IS MACHINE LEARNING?

- **Wikipedia:** Machine learning, a branch of **artificial intelligence**, concerns the construction and study of systems that can **learn** from data.

- **SAS:** Machine learning is a branch of artificial intelligence that **automates** the building of systems that learn from data, identify patterns, and predict future results – with **minimal human intervention.** It shares many approaches with statistical modeling, data mining, information retrieval and other related fields.

MACHINE LEARNING

TWO FACETS



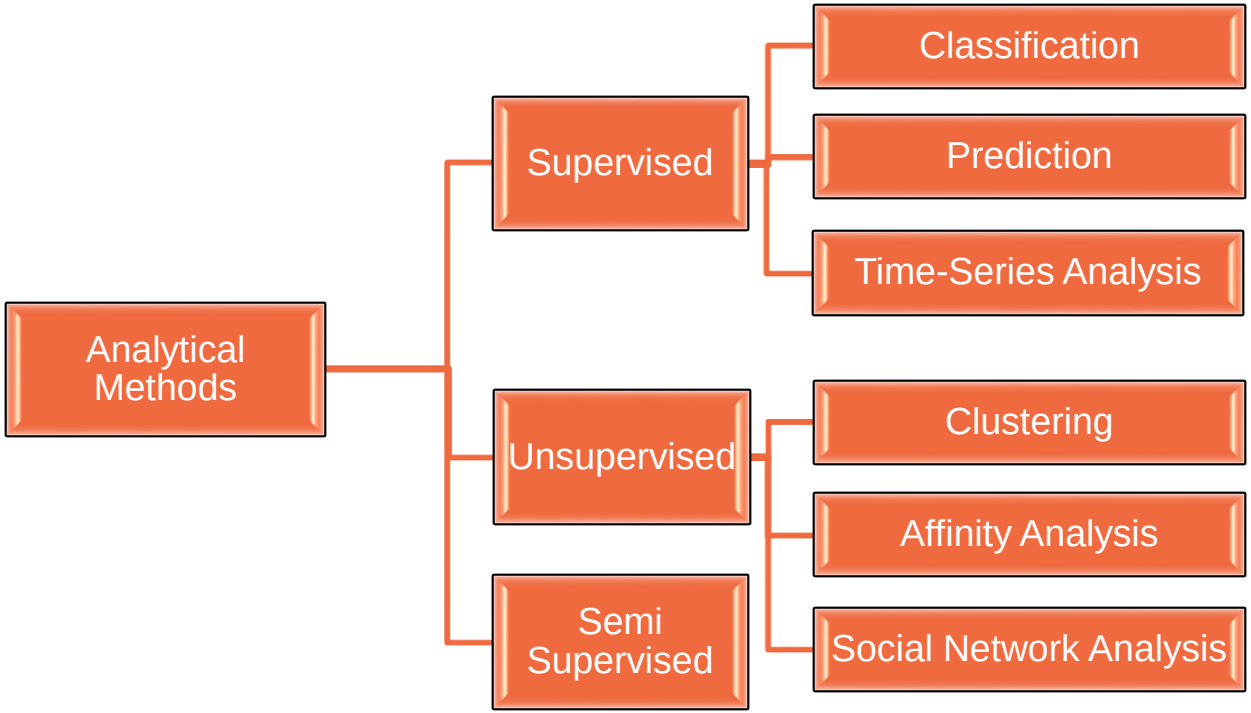
Algorithms

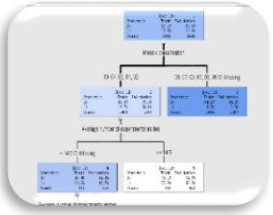
Step-by-step set of operations to be performed to solve a business problem.

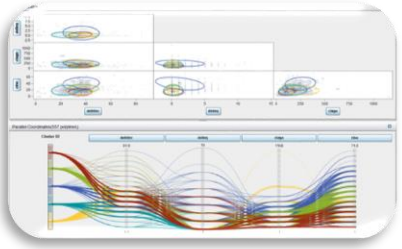
Automation (Scalability)

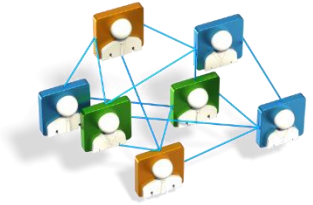
MACHINE LEARNING

ALGORITHM TAXONOMY









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
- Kernel density estimation
- Principal components analysis
- Singular value decomposition
- Gaussian mixture models
- Sequential covering rule building
- Model Ensembles

MACHINE LEARNING

TWO FACETS

Algorithms

Step-by-step set of operations to be performed to solve a business problem.

Automation (Scalability)

MACHINE LEARNING AT SCALE

WHAT DOES “AT SCALE” MEAN?

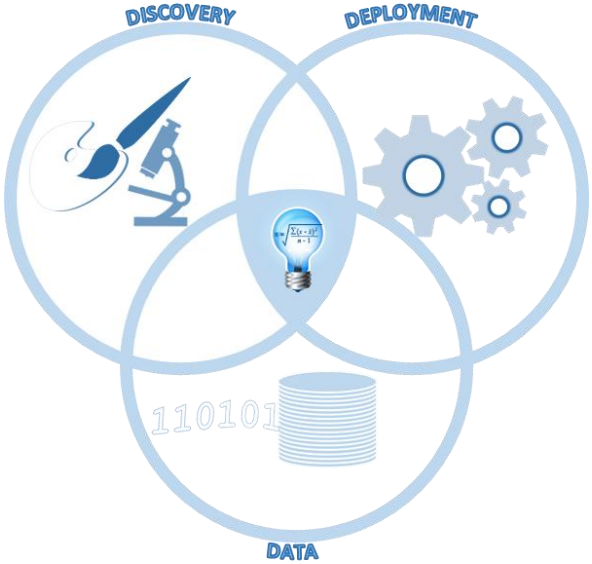
Craft Beer

Beer Production at Scale

SAS MACHINE LEARNING

*Go from **Machine Learning at Scale** to Making Decisions at Scale*

- More Predictive Algorithms
- More Iterations
- More Granular and Hierarchical Segments
- Automated Model Development
- Enabling Non-Technical Users



- Use ALL the Data
- Include More Attributes

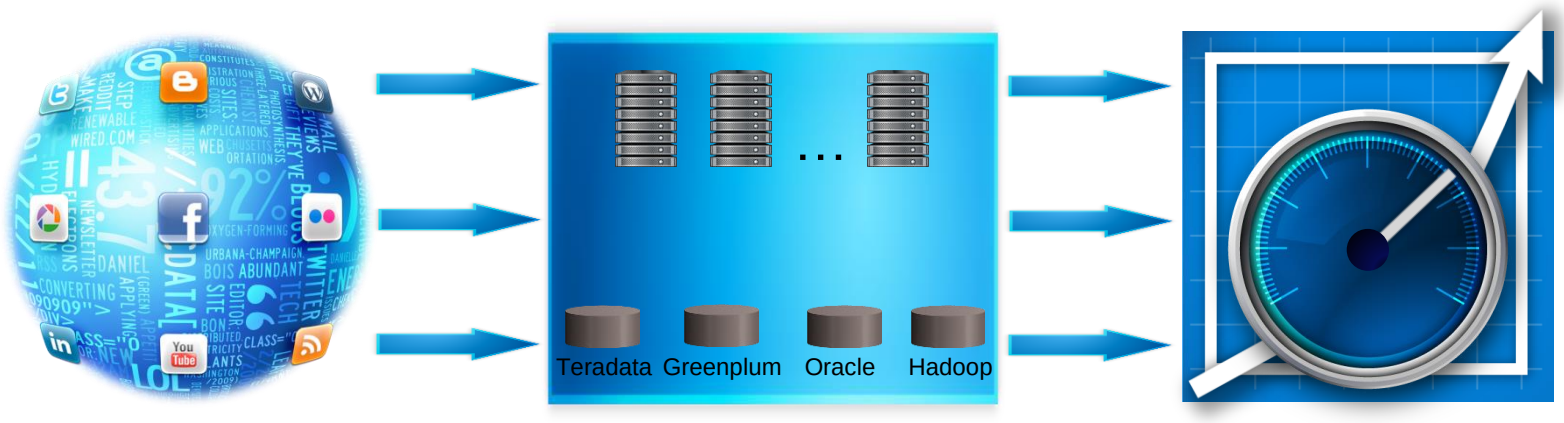
- Automated Implementation and Retraining
- Move Insights Closer to the Decision Maker
- Real-Time
- Scalable High Performance Architectures
- Execute scoring in-database

SAS® High-Performance Analytics

All of your data

Model extensively, iteratively, frequently

Better decisions at the right time



The SAS High-Performance Analytics solution provides in-memory set of offerings for quickly developing analytical models

How does it work?

SAS® High-Performance Analytics

- Automatically splits a single procedure into multiple threads that can be **run in parallel**
- Runs tasks in parallel across a **distributed environment** to provide massively parallel processing (MPP)
- Maximizes performance by operating on **data already loaded into memory**

SAS® HIGH-PERFORMANCE ANALYTICS

THE ROLE OF HPA IN THE ANALYTICS JOURNEY

- Prepare (“All data”, lots of variables, new events, unstructured data…)
- Explore (fast, interactive, anomaly detection…)
- Model (no. of iterations, complex models, retraining…)
- Deploy (operationalize, real-time…)

SAS® HIGH-PERFORMANCE ANALYTICS

BUSINESS BENEFITS

Efficiency

- Build, test and validate more models faster resulting in reduced cycle time to identify the best model
- Automate a greater deal of the total modeling flow in order to free up analytical resources to do more value-added work

Insight

- Using more data than previously possible, companies can now uncover unknown relationships and patterns to gain competitive advantage
- New data becomes available since sampling is no longer needed and you can use all the data combined with the fact that you are able to leverage data from multiple sources including both structured and unstructured information

Innovation

- Having improved efficiency and thereby freed up valuable resources in combination with new valuable data sources enable companies to conduct more experimentation and more iterations
- This will lead to the identification of new customer segments, potential new products and new service offerings

SAS® HIGH-PERFORMANCE ANALYTICS

SAS HIGH-PERFORMANCE PROCEDURES

High-Performance Statistics

- HPLOGISTIC
- HPREG
- HPLMIXED
- HPNLMOD
- HPSPLIT
- HPGENSELECT
- HPFMM
- HPCANDISC
- HPPRINCOMP
- HPQUANTSELECT
- HPPLS

High-Performance Data Mining

- HPREDUCE
- HPNEURAL
- HPFOREST
- HP4SCORE
- HPDECIDE
- HPCLUS
- HPSVM
- HPBNET

High-Performance Text Mining

- HPTMINE
- HPTMSCORE

High-Performance Optimization

- OPTLSO
- Select features in
  - OPTMILP
  - OPTLP
  - OPTMODEL

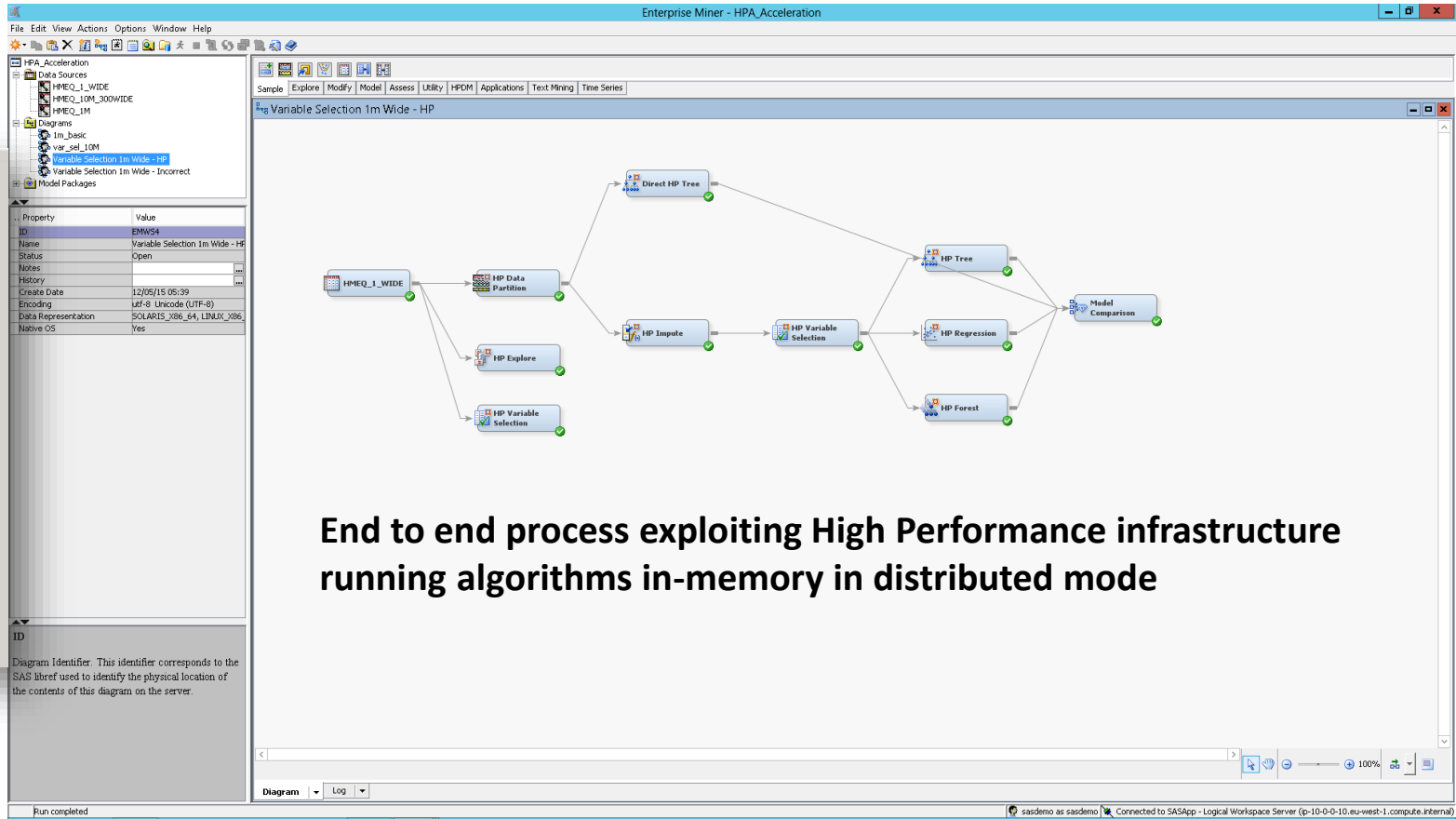
High-Performance Econometrics

- HPCOUNTREG
- HPSEVERITY
- HPQLIM
- HPPANEL
- HPCOPULA
- HPCDM

SAS® HIGH-PERFORMANCE DATA MINING

HIGH PERFORMANCE PROCEDURES IN SAS ENTERPRISE MINER

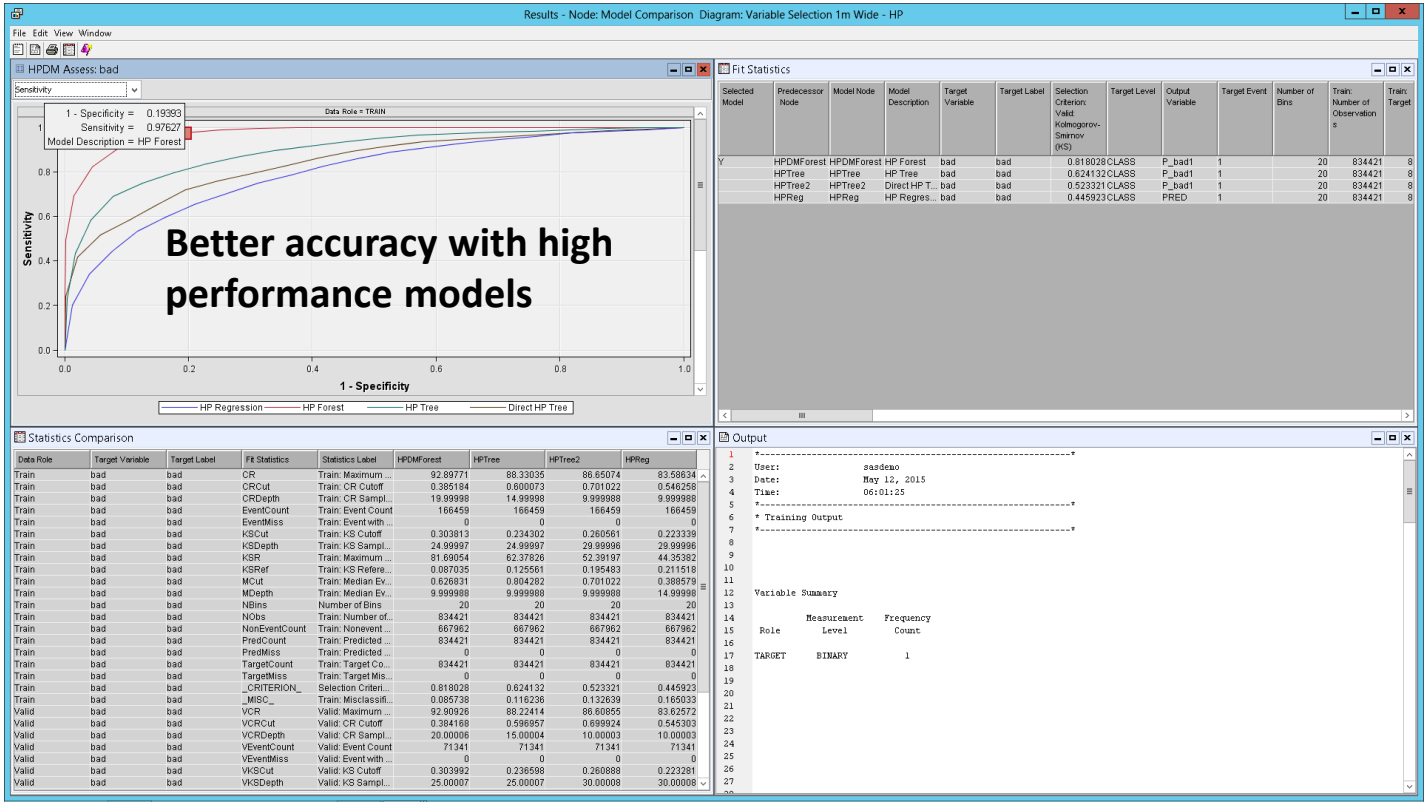
- HP Cluster
- HP Data Partition
- HP Explore
- HP Forest
- HP GLM
- HP Impute
- HP Neural
- HP Principal Components
- HP Regression
- HP SVM
- HP Text Miner
- HP Transform
- HP Tree
- HP Variable Selection



End to end process exploiting High Performance infrastructure running algorithms in-memory in distributed mode

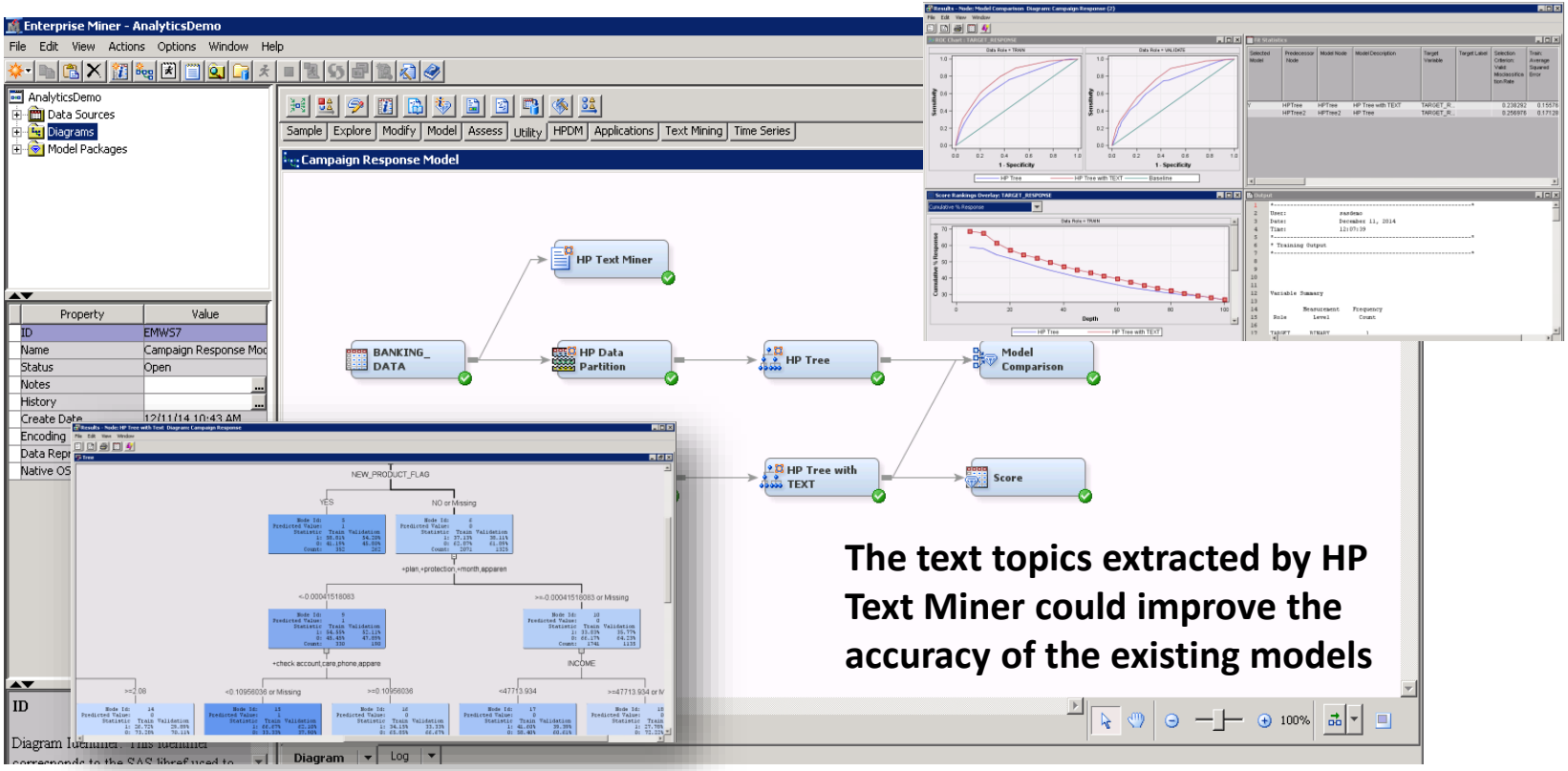
SAS® HIGH-PERFORMANCE DATA MINING

HIGH PERFORMANCE MODEL ACCURACY



SAS® HIGH-PERFORMANCE DATA MINING

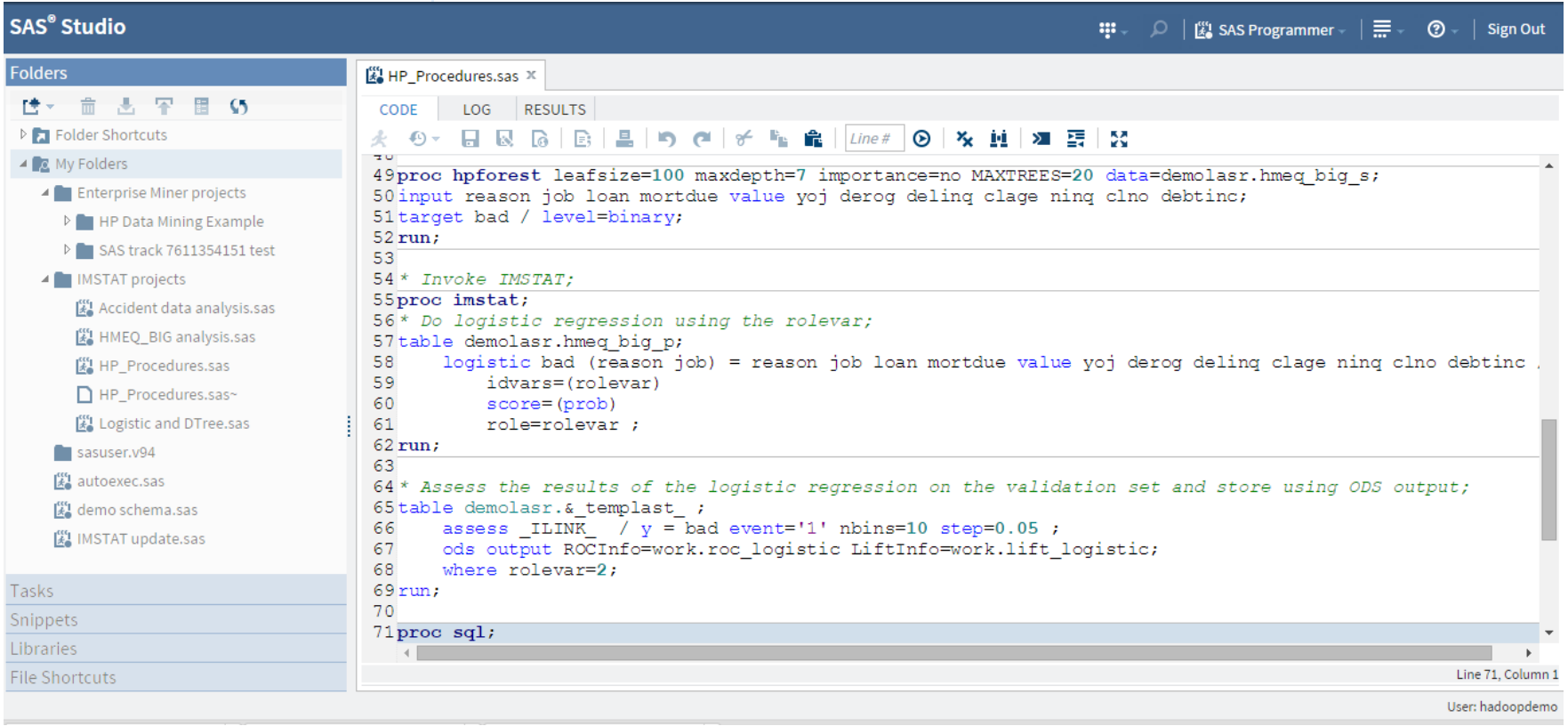
INSIGHT FROM UNSTRUCTURED DATA



The text topics extracted by HP Text Miner could improve the accuracy of the existing models

SAS® STUDIO

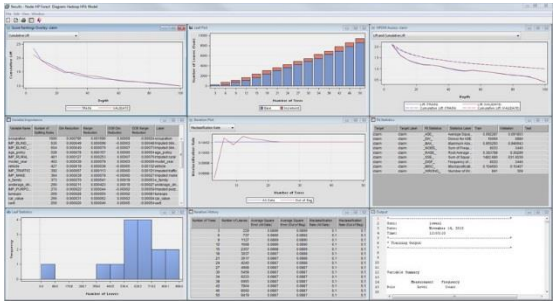
HIGH PERFORMANCE PROCEDURES



SAS® HPA PROCEDURES

PROC HPFOREST

- The HPFOREST procedure creates an ensemble of hundreds of decision trees to predict a single target of either interval or nominal measurement level.
- The decision trees differ from each other in two ways:
  - First, the training data for a tree is a sample, without replacement, from the original training data of the forest.
  - Second, the input variables considered for splitting a node are randomly selected from all available inputs.
- The HPFOREST procedure searches for rules that maximize the measure of worth that is associated with the splitting criterion. For binary, nominal, and interval targets, the worth of a split is the reduction in node impurity.

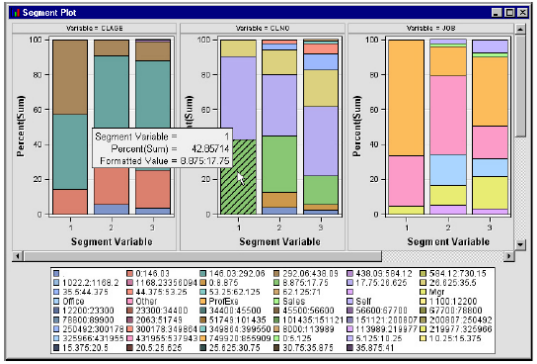


```
proc hpforest data=diabetes ;
   input NumPregnancies
         plasmaGlucose
         diastolicBloodPr
         tricepsSkinfold
         hrSerumInsulin
         BodyMassIndex
         DiabetesPedigreeFn
         Age ;
   target diabetes;
   save file="model";
run;
```

SAS® HPA PROCEDURES

PROC HPCLUS

- The HPCLUS procedure uses the k-means algorithm for clustering numeric interval input variables and the k-modes algorithm for nominal input variables. Uses only numeric interval or only nominal variables, it does not perform clustering for mixed levels of inputs.
- Provides a new technique called the aligned box criterion (ABC) for estimating the number of clusters in the data set.
- Enables you to use parallel execution in a distributed computing environment, while you can still run in single-machine mode on the server.




```
proc hpclus data=sashelp.iris maxclusters=3 outiter outstat=hpclusOutstat1;
   score out=hpclusOut1;
   input SepalLength SepalWidth PetalLength PetalWidth;
   id SepalLength SepalWidth PetalLength PetalWidth Species;
run;
```

SAS® HPA PROCEDURES

PROC HPBNET (HP BAYESIAN NETWORK)

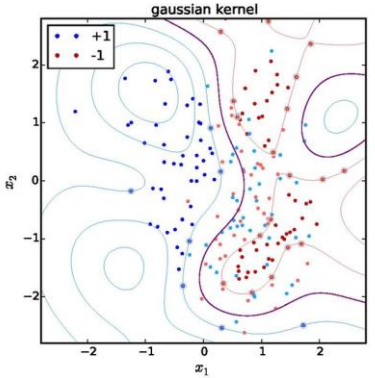
- The HPBNET procedure analyses different types of Bayesian network structures, including naive, tree augmented naive (TAN), Bayesian network-augmented naive (BAN), parent-child Bayesian network, and Markov blanket.
- Performs efficient variable selection through independence tests, and it automatically selects the best model from the specified parameters.

```
proc hpbnet data=sampsio.dmagecr numbin=10 alpha=0.05
          structure=Naive TAN PC MB varselect=0 1 bestmodel;
    target Good_Bad;
    input  Checking History Purpose Savings Employed Installp
           Resident Property Other Housing Existcr Job Depend
           Foreign/level=NOM;
    input Age Amount Duration/level=INT;
    partition FRACTION (VALIDATE=0.3);
    output network=network fit=fit validinfo=validinfo;
run;
```

SAS® HPA PROCEDURES

PROC HPSVM

- The HPSVM procedure supports both continuous and categorical inputs
- Supports classification of a binary target
- Supports the interior-point method and the active-set method
- Supports cross-validation for penalty selection
- Supports scoring of models

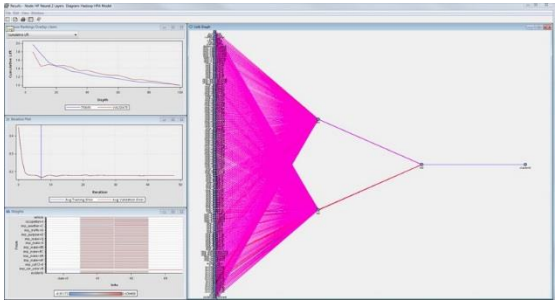



```
proc hpsvm data=sampsio.dmagecr;
    input checking history purpose savings employed marital coapp
          property other job housing telephon foreign/level=nominal;
    input duration amount installp resident existcr depends age/level=interval;
    target good_bad;
run;
```

SAS® HPA PROCEDURES

PROC HPNEURAL

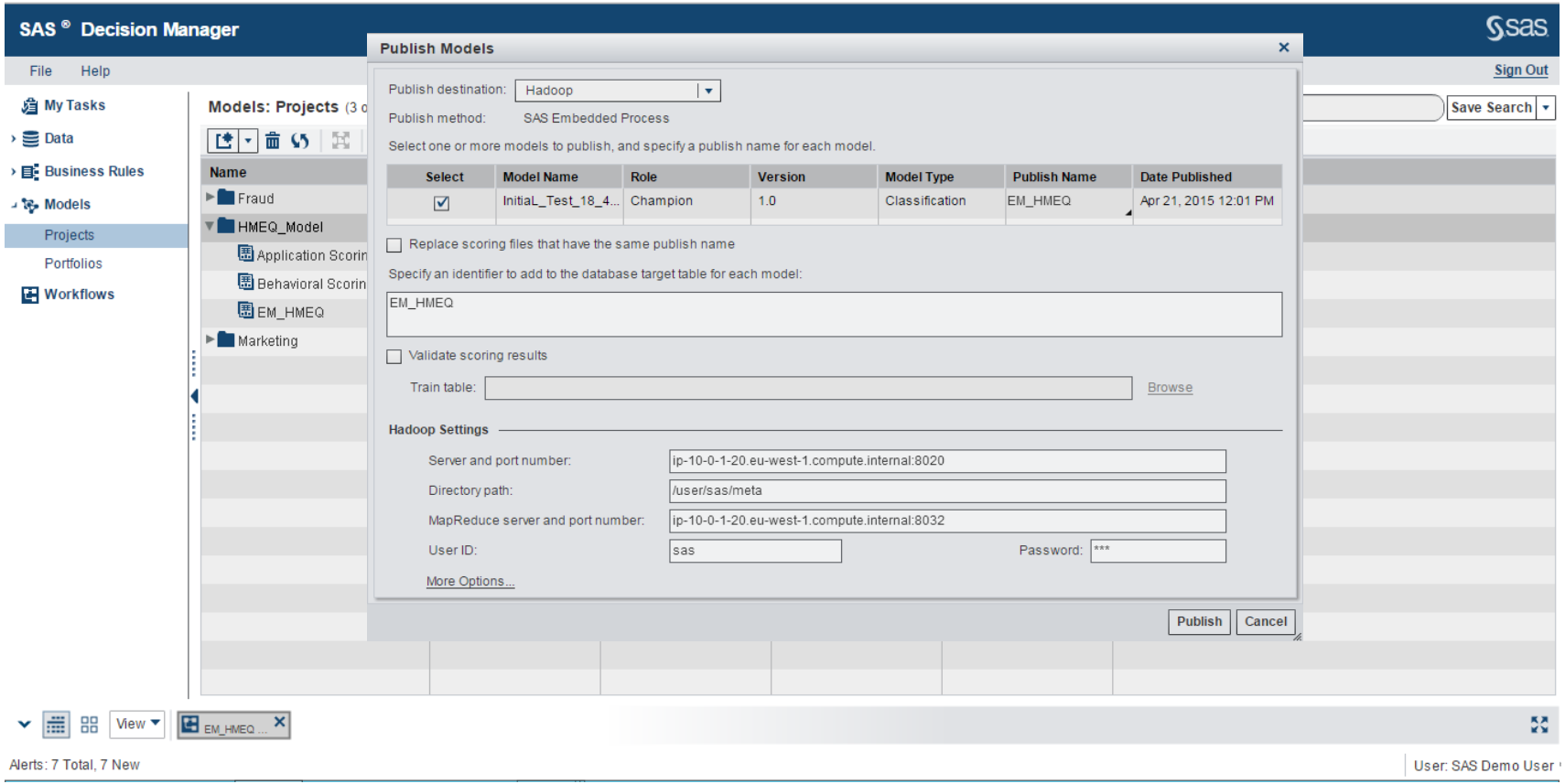
- The HPNEURAL procedure does not have many parameters that you must specify so that users with minimal experience with neural networks can obtain good results
- Uses the limited memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) optimization method with proprietary enhancements. LBFGS was chosen for the HPNEURAL procedure because of both its speed of training and its limited use of memory, which can be especially important for problems with large amounts of training data.



```
proc hpneural data=iris;
   input SepalLength SepalWidth PetalLength PetalWidth;
   target Species / level=nom;
   hidden 2;
   train outmodel=model_iris maxiter=1000;
   score out=scores_iris;
run;
```

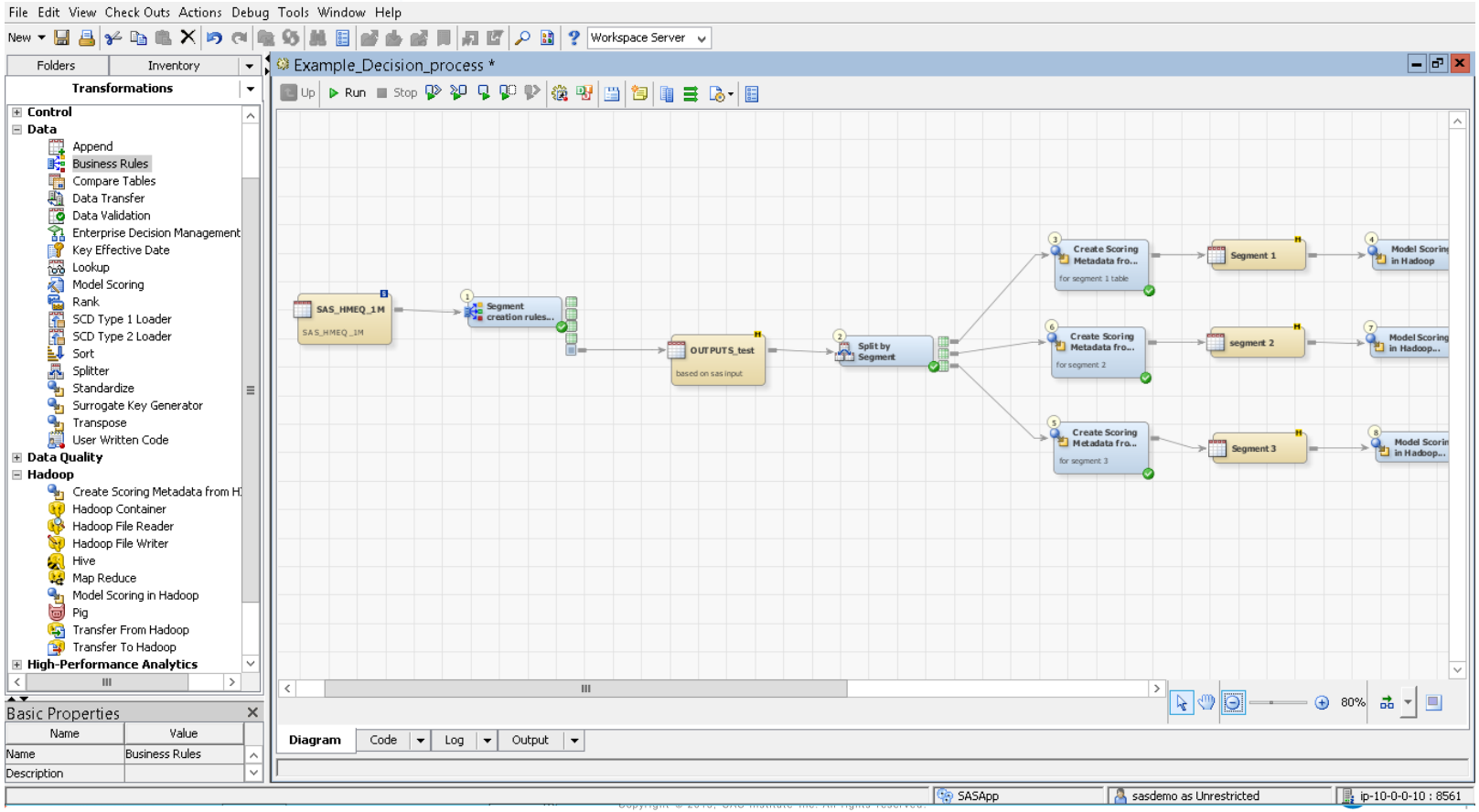
SAS® HPA AUTOMATION

PUBLISHING HIGH PERFORMANCE MODELS IN-DATABASE

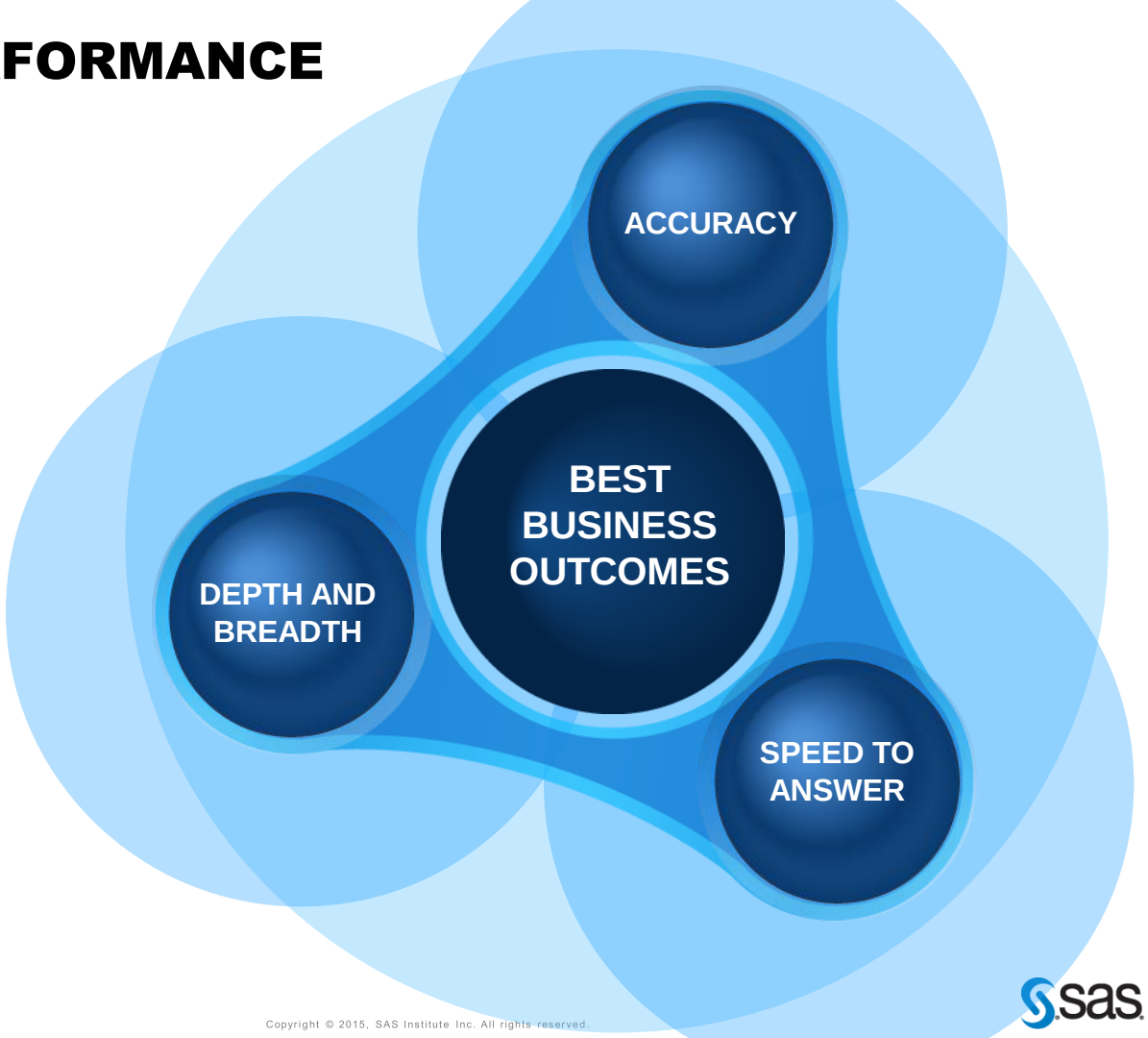


SAS® HPA AUTOMATION

EXECUTING MODELS IN PRODUCTION



SAS® HIGH-PERFORMANCE ANALYTICS



SAS® FORUM
UNITED KINGDOM 2015

www.SAS.com