



SAS® FORUM
UNITED KINGDOM 2015

Deep Dive into High Performance Machine Learning Procedures

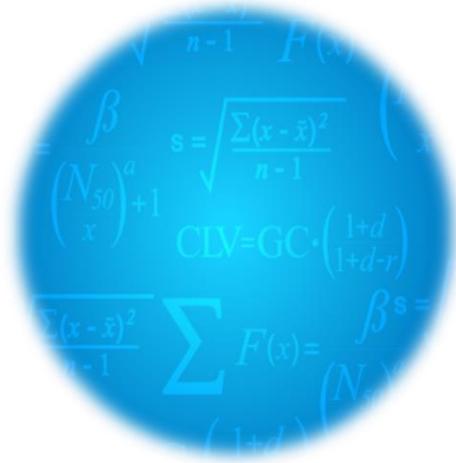
Tuba Islam, Analytics CoE, SAS UK

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.

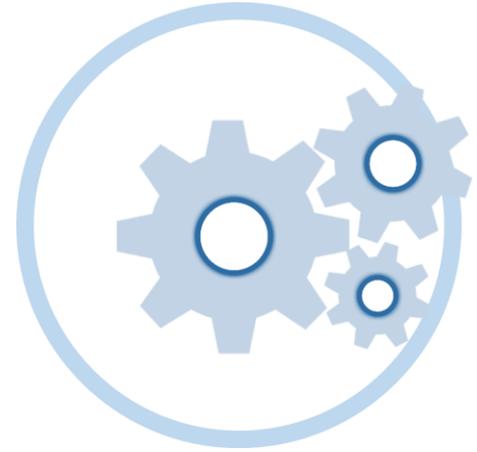
MACHINE LEARNING

TWO FACETS



Algorithms

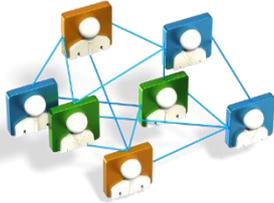
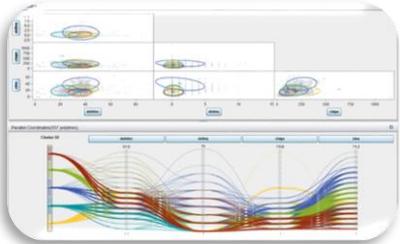
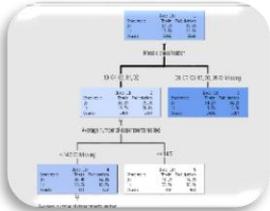
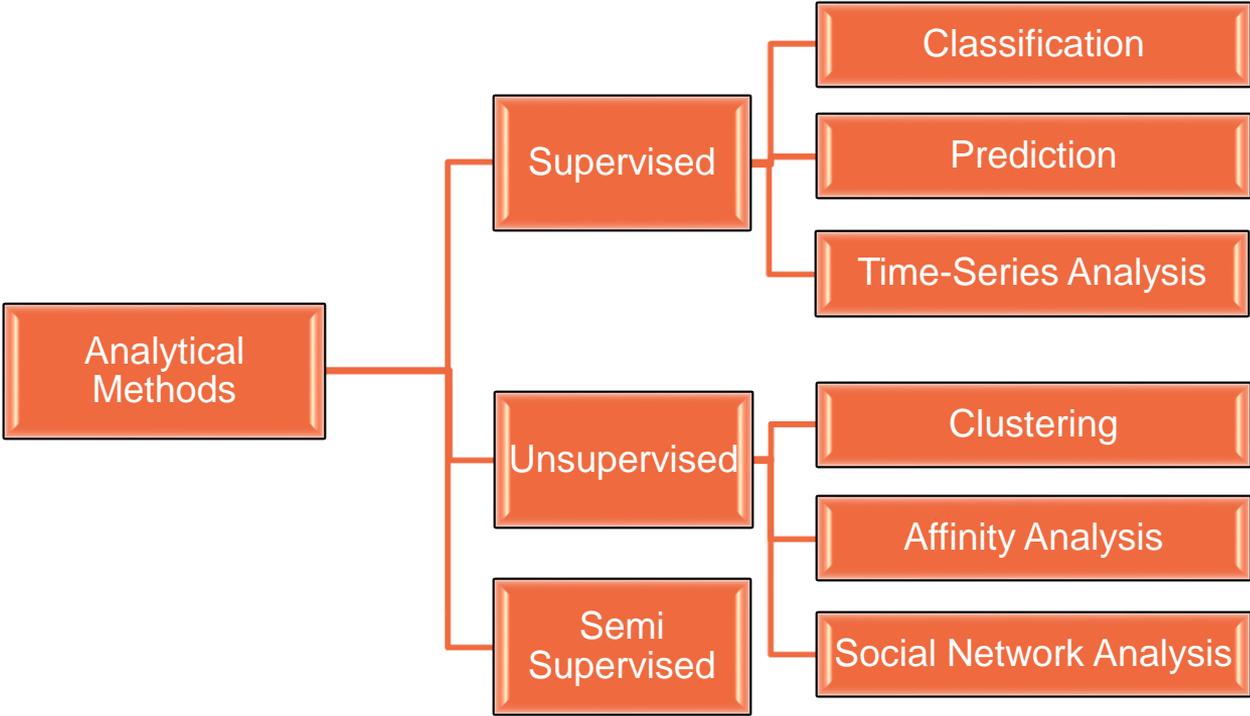
Step-by-step set of operations to be performed to solve a business problem.



Automation (Scalability)

MACHINE LEARNING

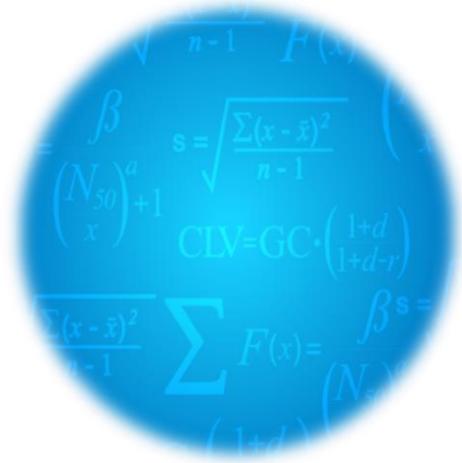
ALGORITHM TAXONOMY



- 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

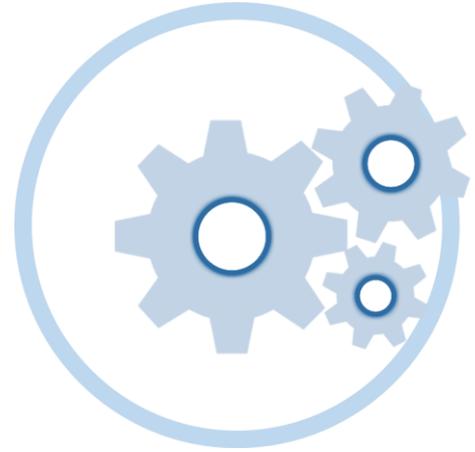
MACHINE LEARNING

TWO FACETS



Algorithms

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



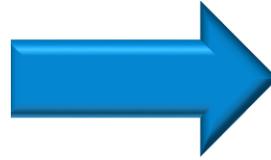
Automation (Scalability)

MACHINE LEARNING AT SCALE

WHAT DOES “AT SCALE” MEAN?



Craft Beer

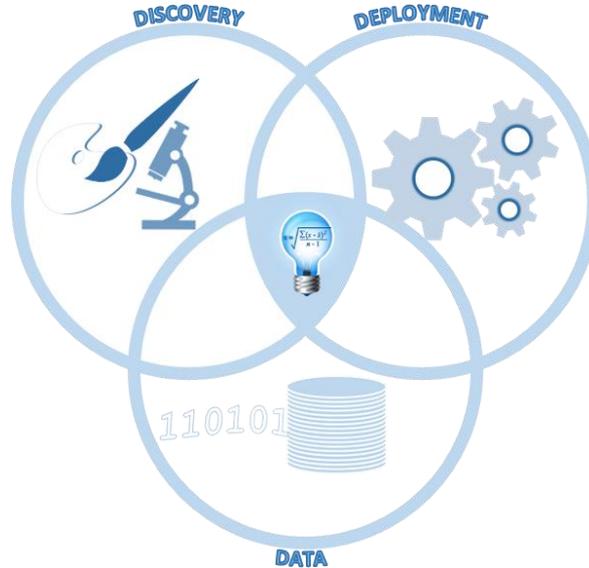


Beer Production at Scale

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



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

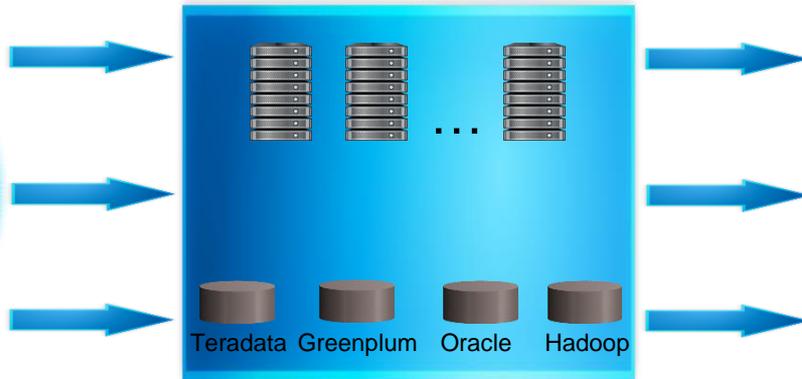
- Use ALL the Data
- Include More Attributes

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

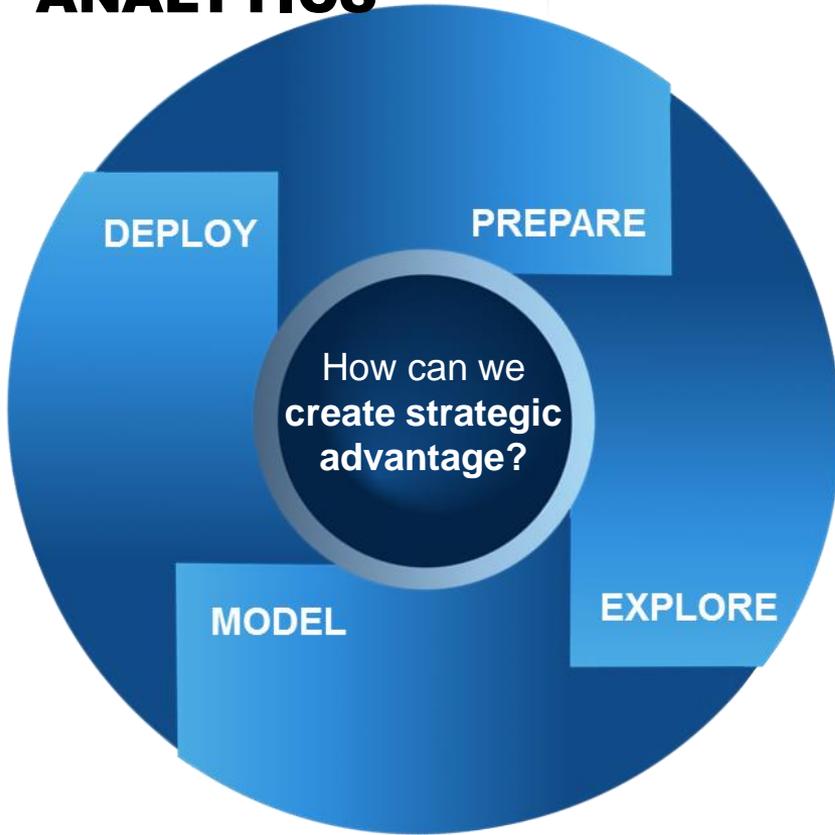
How does it work?

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

The screenshot displays the SAS Enterprise Miner interface with a workflow diagram titled "Variable Selection 1m Wide - HP". The workflow starts with the "HMEQ_1_WIDE" data source, which branches into three parallel paths: "HP Data Partition", "HP Explore", and "HP Variable Selection". The "HP Data Partition" path leads to "Direct HP Tree", which then feeds into "HP Tree". The "HP Explore" path leads to "HP Impute", which then feeds into "HP Variable Selection". The "HP Variable Selection" path also leads to "HP Variable Selection". From "HP Variable Selection", the workflow branches into "HP Regression" and "HP Forest". Both "HP Regression" and "HP Forest" feed into the final "Model Comparison" step. The interface includes a menu bar (File, Edit, View, Actions, Options, Window, Help), a toolbar, and a task pane on the left showing a project tree and a property window for the selected node.

Property	Value
ID	EMW54
Name	Variable Selection 1m Wide - HP
Status	Open
Notes	
History	
Create Date	12/05/15 05:39
Encoding	utf-8 Unicode (UTF-8)
Data Representation	SOLARIS_x86_04, LINUX_x86_04
Native OS	Yes

Diagram Identifier. This identifier corresponds to the SAS href used to identify the physical location of the contents of this diagram on the server.

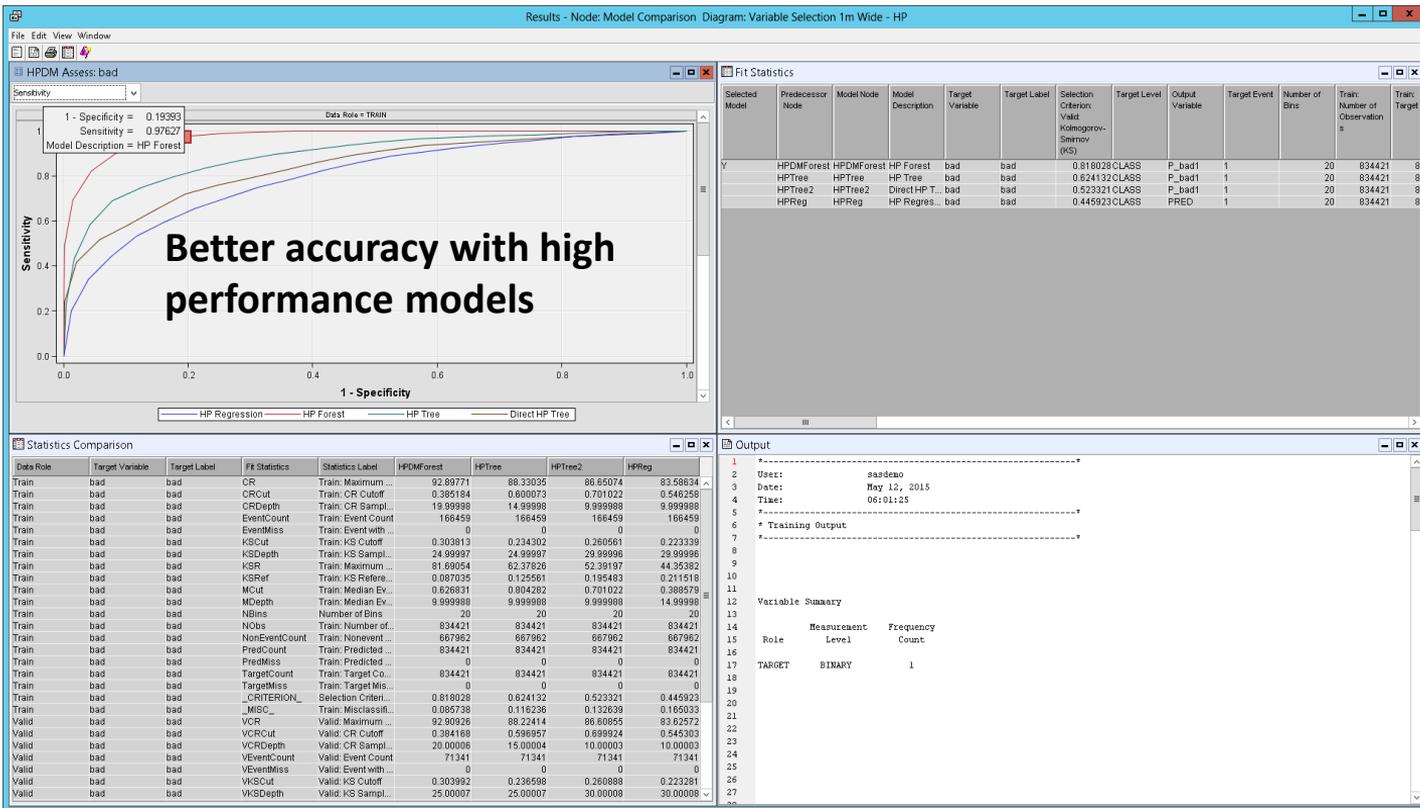
Run completed

sasdemo as sasdemo Connected to SASApp - Logical Workspace Server (p-10-0-0-10.eu-west-1.compute.intern...

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

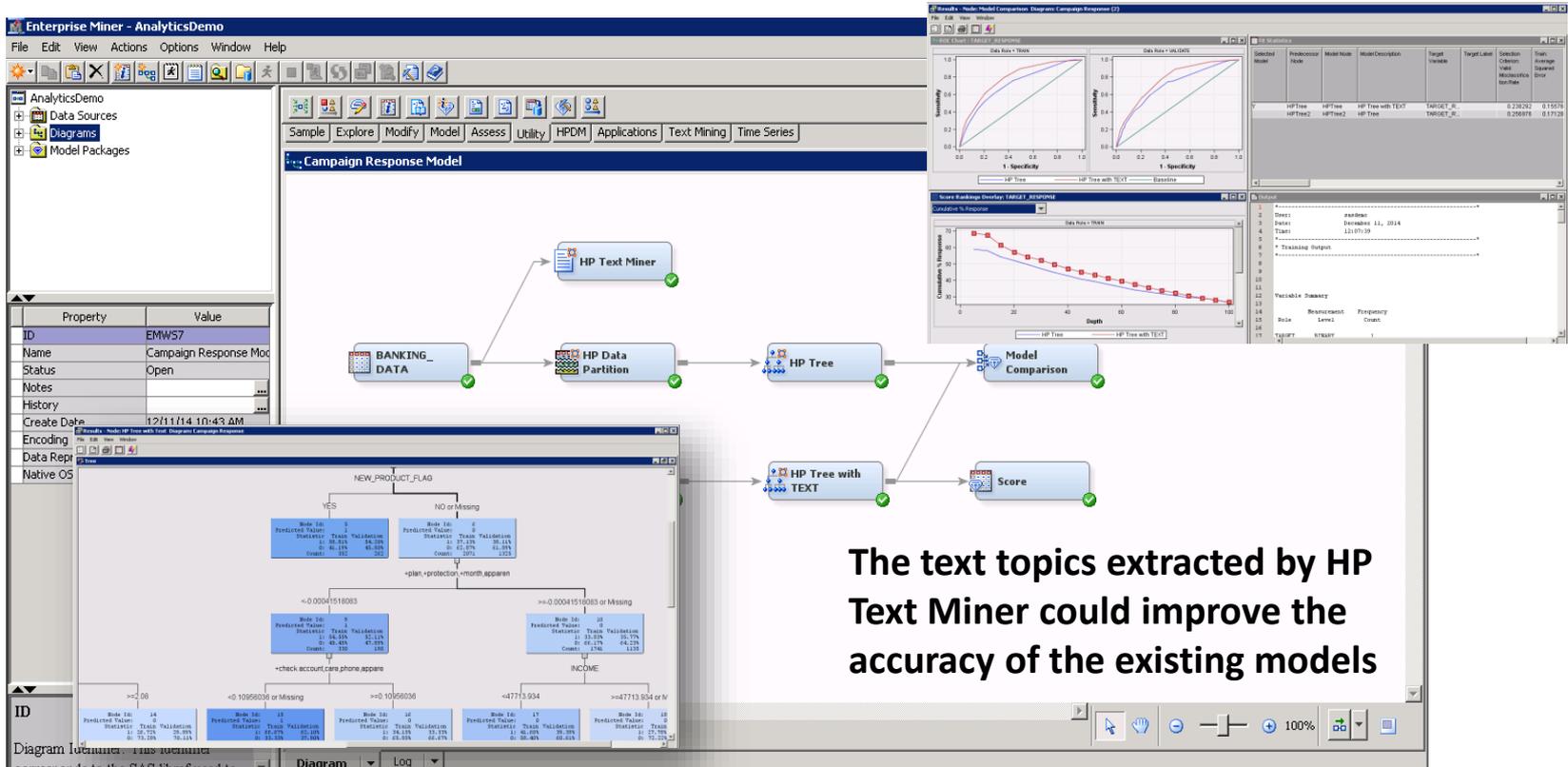
SAS® HIGH-PERFORMANCE DATA MINING

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

The screenshot displays the SAS Studio environment. On the left is a 'Folders' pane with a tree view of project folders including 'Enterprise Miner projects', 'IMSTAT projects', and 'sasuser.v94'. The main workspace shows a SAS program named 'HP_Procedures.sas' with the following code:

```
49 proc hpforest leafsize=100 maxdepth=7 importance=no MAXTREES=20 data=demolasr.hmeq_big_s;
50 input reason job loan mortdue value yoj derog delinq clage ning clno debtinc;
51 target bad / level=binary;
52 run;
53
54 * Invoke IMSTAT;
55 proc imstat;
56 * Do logistic regression using the rolevar;
57 table demolasr.hmeq_big_p;
58   logistic bad (reason job) = reason job loan mortdue value yoj derog delinq clage ning clno debtinc
59     idvars=(rolevar)
60     score=(prob)
61     role=rolevar ;
62 run;
63
64 * Assess the results of the logistic regression on the validation set and store using ODS output;
65 table demolasr.&_templast_ ;
66   assess _ILINK_ / y = bad event='1' nbins=10 step=0.05 ;
67   ods output ROCInfo=work.roc_logistic LiftInfo=work.lift_logistic;
68   where rolevar=2;
69 run;
70
71 proc sql;
```

The status bar at the bottom right indicates 'Line 71, Column 1' and 'User: hadoopdemo'.



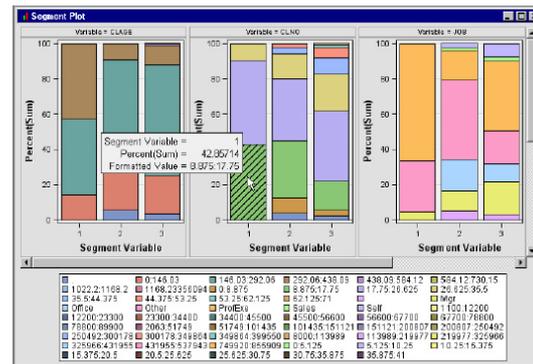
- The HPFOREST procedure creates an ensemble of hundreds of decision trees to predict a single target of either interval or nominal measurement level.
- The decision trees differ from each other in two ways:
 - First, the training data for a tree is a sample, without replacement, from the original training 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;
```



- 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;
```

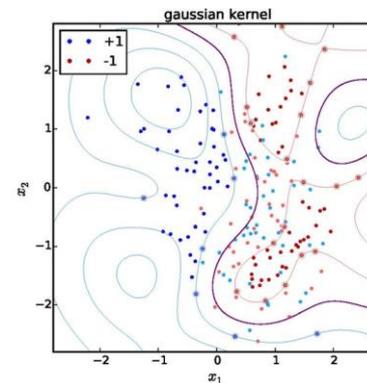


- The HPBNET procedure analyses different types of Bayesian network structures, including naive, tree augmented naive (TAN), Bayesian network-augmented naive (BAN), parent-child Bayesian network, and Markov blanket.
- Performs efficient variable selection through independence tests, and it automatically 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;
```



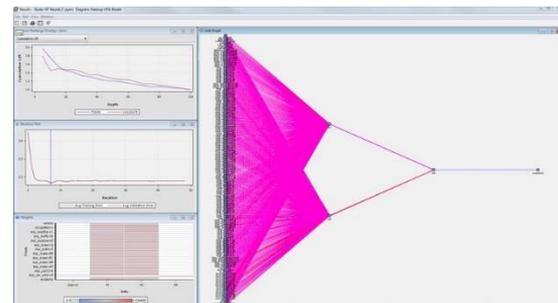
- The HPSVM procedure supports both continuous and categorical inputs
- Supports classification of a binary target
- Supports the interior-point method and the active-set method
- Supports cross-validation for penalty selection
- Supports scoring of models



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



- 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® Decision Manager

File Help

My Tasks

Data

Business Rules

Models

Projects

Portfolios

Workflows

Models: Projects (3 o)

Name

- Fraud
- HMEQ_Model
 - Application Scoring
 - Behavioral Scoring
 - EM_HMEQ
- Marketing

Publish Models

Publish destination: Hadoop

Publish method: SAS Embedded Process

Select one or more models to publish, and specify a publish name for each model.

Select	Model Name	Role	Version	Model Type	Publish Name	Date Published
<input checked="" type="checkbox"/>	InitiaL_Test_18_4...	Champion	1.0	Classification	EM_HMEQ	Apr 21, 2015 12:01 PM

Replace scoring files that have the same publish name

Specify an identifier to add to the database target table for each model:

EM_HMEQ

Validate scoring results

Train table: [Browse](#)

Hadoop Settings

Server and port number:

Directory path:

MapReduce server and port number:

User ID: Password:

[More Options...](#)

View



The screenshot displays the SAS HPA software interface. The main window shows a workflow diagram titled "Example_Decision_process". The workflow starts with an input table "SAS_HMEQ_1M" (SA_S_HMEQ_1M) which feeds into a "Segment creation rules..." task. The output of this task is an "OUTPUTS_test" table (based on sas input). This table is then processed by a "Split by Segment" task. The output of the split task is distributed into three parallel paths, each leading to a "Create Scoring Metadata from Hadoop" task (for segment 1, 2, and 3 respectively). Each of these tasks then feeds into a "Model Scoring in Hadoop" task. The interface includes a menu bar (File, Edit, View, Check Outs, Actions, Debug, Tools, Window, Help), a toolbar, and a left-hand navigation pane with categories like Control, Data, Data Quality, Hadoop, and High-Performance Analytics. A "Basic Properties" window is open at the bottom left, showing details for the "Business Rules" task. The bottom status bar shows "SASApp", "sasdemo as Unrestricted", and "ip-10-0-0-10 : 8561".

File Edit View Check Outs Actions Debug Tools Window Help

New Workspace Server

Example_Decision_process *

Transformations

Control

- Data
 - Append
 - Business Rules
 - Compare Tables
 - Data Transfer
 - Data Validation
 - Enterprise Decision Management
 - Key Effective Date
 - Lookup
 - Model Scoring
 - Rank
 - SCD Type 1 Loader
 - SCD Type 2 Loader
 - Sort
 - Splitter
 - Standardize
 - Surrogate Key Generator
 - Transpose
 - User Written Code
- Data Quality
- Hadoop
 - Create Scoring Metadata from H...
 - Hadoop Container
 - Hadoop File Reader
 - Hadoop File Writer
 - Hive
 - Map Reduce
 - Model Scoring in Hadoop
 - Pig
 - Transfer From Hadoop
 - Transfer To Hadoop
- High-Performance Analytics

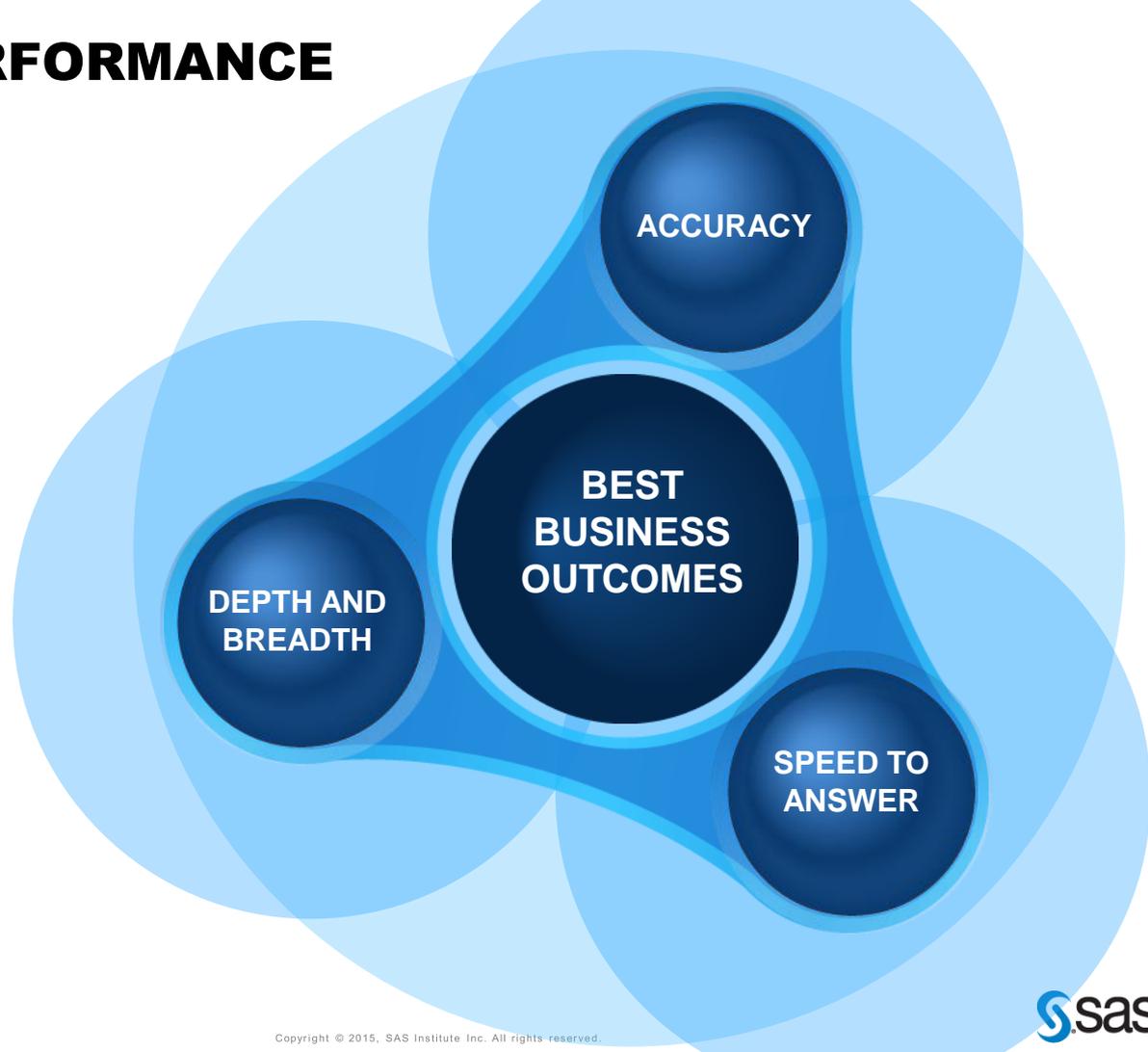
Basic Properties

Name	Value
Name	Business Rules
Description	

Diagram Code Log Output

SASApp sasdemo as Unrestricted ip-10-0-0-10 : 8561

SAS® HIGH-PERFORMANCE ANALYTICS





SAS® FORUM
UNITED KINGDOM 2015



THE
POWER
TO KNOW.

www.SAS.com