SAS® FORUM
UNITED KINGDOM 2015

Advanced Modelling Techniques in SAS Enterprise Miner

Dr Iain Brown, Senior Analytics Specialist Consultant,
SAS UK & Ireland

THE POWER TO KNOW.

Agenda

- **SAS Presents – Thursday 11th June 2015 – 15:45**
- **Advanced Modelling Techniques in SAS Enterprise Miner**
- *The session looks at:*
  - *Supervised and Unsupervised Modelling*
  - *Classification and Prediction Techniques*
  - *Tree Based Learners*

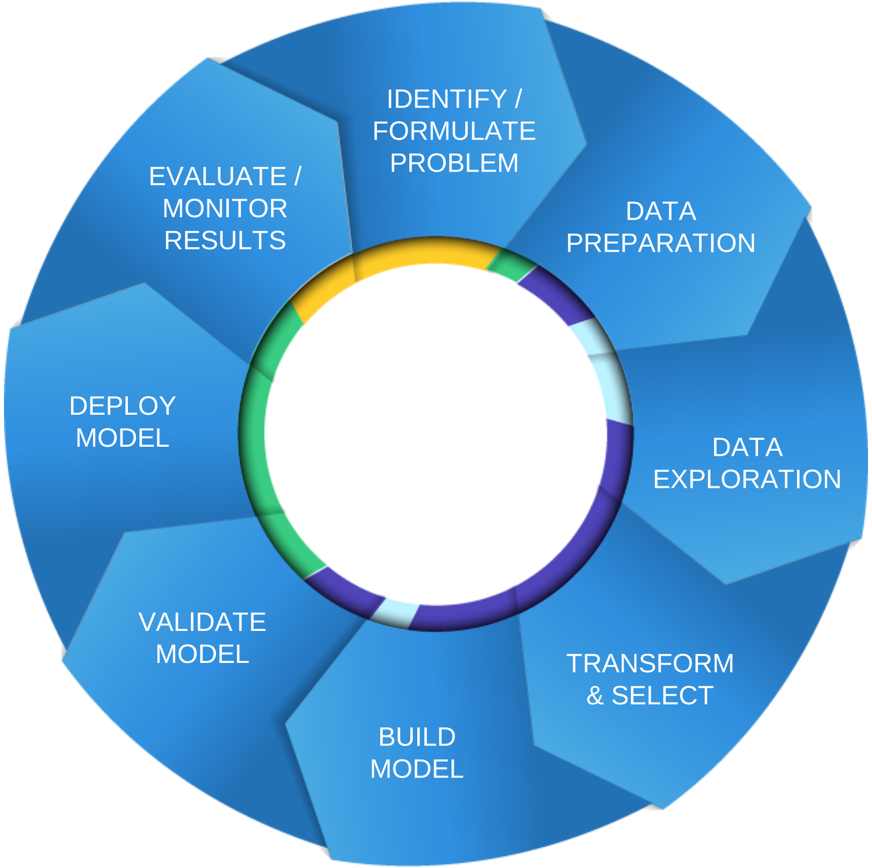
The Analytics Lifecycle

BUSINESS MANAGER

Domain Expert
Makes Decisions
Evaluates Processes and ROI

IT SYSTEMS / MANAGEMENT

Model Validation
Model Deployment
Model Monitoring
Data Preparation



BUSINESS ANALYST

Data Exploration
Data Visualization
Report Creation

DATA MINER / STATISTICIAN

Exploratory Analysis
Descriptive Segmentation
Predictive Modeling

SAS FORUM UNITED KINGDOM 2015

sas THE POWER TO KNOW.

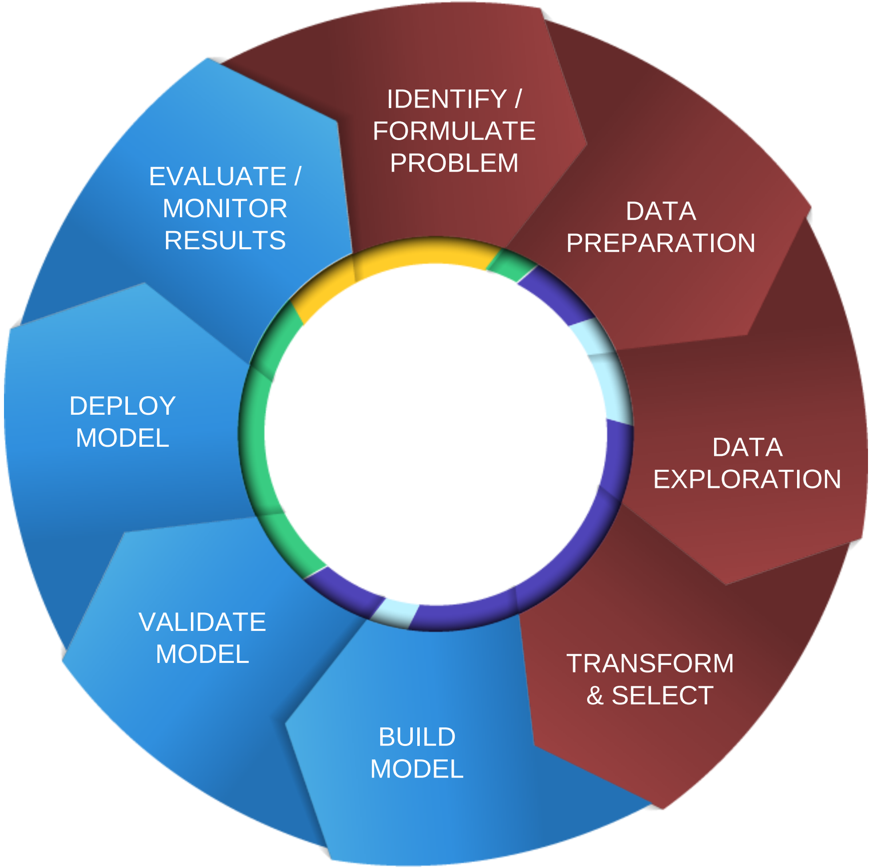
The Analytics Lifecycle

BUSINESS MANAGER

Domain Expert
Makes Decisions
Evaluates Processes and ROI

IT SYSTEMS / MANAGEMENT

Model Validation
Model Deployment
Model Monitoring
Data Preparation



BUSINESS ANALYST

Data Exploration
Data Visualization
Report Creation

DATA MINER / STATISTICIAN

Exploratory Analysis
Descriptive Segmentation
Predictive Modeling

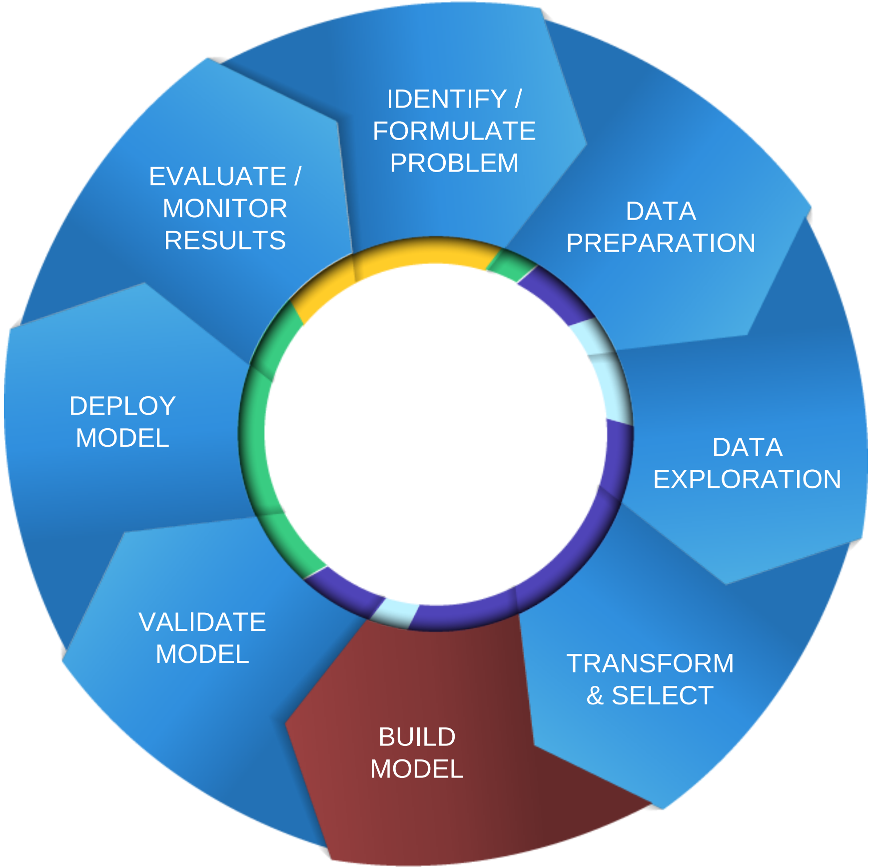
The Analytics Lifecycle

BUSINESS MANAGER

Domain Expert
Makes Decisions
Evaluates Processes and ROI

IT SYSTEMS / MANAGEMENT

Model Validation
Model Deployment
Model Monitoring
Data Preparation



BUSINESS ANALYST

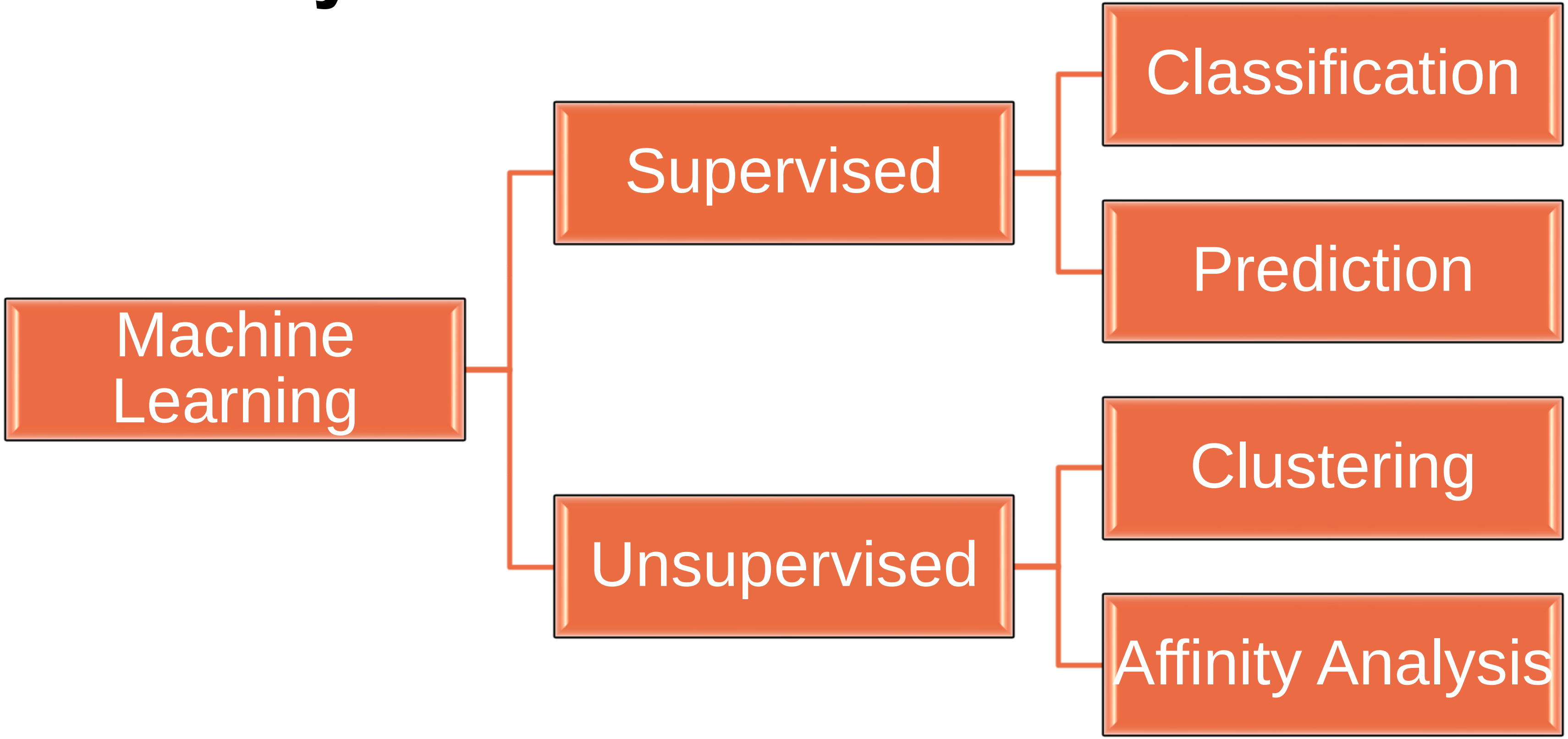
Data Exploration
Data Visualization
Report Creation

DATA MINER / STATISTICIAN

Exploratory Analysis
Descriptive Segmentation
Predictive Modeling

Supervised and Unsupervised Modelling

Taxonomy


Machine Learning
Supervised
Unsupervised
Classification
Prediction
Clustering
Affinity Analysis

Learning Methods

Supervised:

- Discover patterns in the data that relate attributes to labels.
- Patterns are used to predict the values of the label in future data instances.

Unsupervised:

- The data have no label attribute.
- Goal is to explore the data to find some intrinsic structures in them.

Supervised Learning (Classification & Prediction)

Logistic Regression

Decision Trees, CART

Decision Trees, CHAID

Gradient Boosting

Random Forests

kth Nearest Neighbor

Neural Networks

Nonlinear SVMs

Bayesian Networks

Regression, least square

Generalized Linear Models

LASSO, LAR

Splines, MARS

Unsupervised Learning

K-means

Fuzzy K-means

Hierarchical Clustering

Vector Quantization

Multidimensional Scaling

Principal Components

Nonnegative Matrix Factorization

Assocations, Apriori

Classification and Prediction Techniques

Model Development Process

Sample
- Input Data
- File Import
- Sample
- Data Partition
- Merge
- Filter
- Append

Explore
- Association
- Cluster
- Variable Selection
- Market Basket
- StatExplore
- Variable Clustering
- MultiPlot
- Path Analysis
- DMDB
- SOM/Kohonen
- Graph Explore

Modify
- Transform Variables
- Impute
- Replacement
- Interactive Binning
- Rules Builder
- Drop
- Principal Components

Model
- Decision Tree
- AutoNeural
- Dmine Regression
- DMNeural
- Ensemble
- Gradient Boosting
- LARS
- MBR
- Neural Network
- SVM
- Partial Least Squares
- Regression
- Rule Induction
- TwoStage
- Model Import

HPDM
- HP Regression
- HP Forest
- HP SVM
- HP Tree
- HP GLM
- HP Neural

Regression

- Linear
- Logistic

Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Stepwise
Selection Criterion	Default
Use Selection Defaults	No
Selection Options	...
Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour
Convergence Criteria	
Defaults	

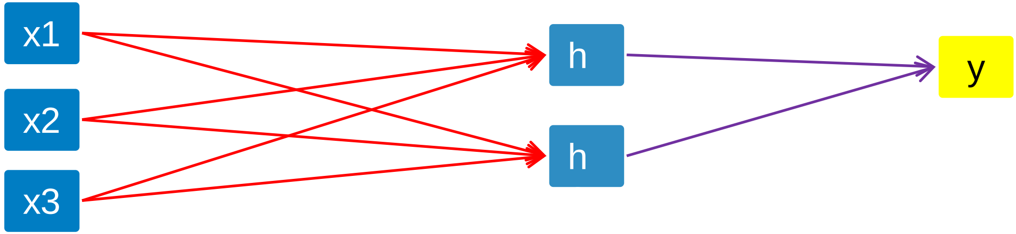
- Computes a forward stepwise least-squares regression
- Optionally computes all 2-way interactions of classification variables
- Optionally uses AOV16 variables to identify non-linear relationships between interval variables and the target variable.
- Optionally uses group variables to reduce the number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

Generalised Linear Models

- Uses the high-performance HPGENSELECT procedure to fit a generalized linear model in a threaded or distributed computing environment.
- Several response probability distributions and link functions are available.
- Provides model selection methods.

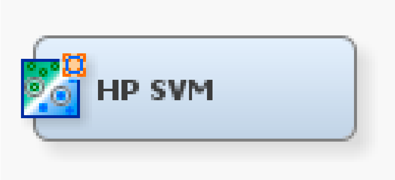
Property	Value
General	
Node ID	HPGLM
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Set Reference Level	Default
Reference Level	...
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
Suppress Intercept	No
Use Missing as Level	No
Modeling	
Interval Target Probabilit	Poisson
Interval Target Link Func	Log
Binary Target Link Functi	Logit
Optimization Options	...
Convergence Options	...
ZI Model Options	...
Tweedie Model Options	...
Model Selection	
Selection Method	Forward
Stop Criterion	DEFAULT (SL)
Selection Options	...

Neural Networks

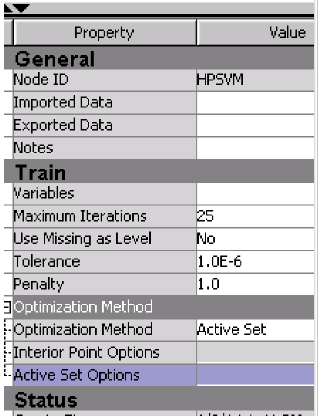


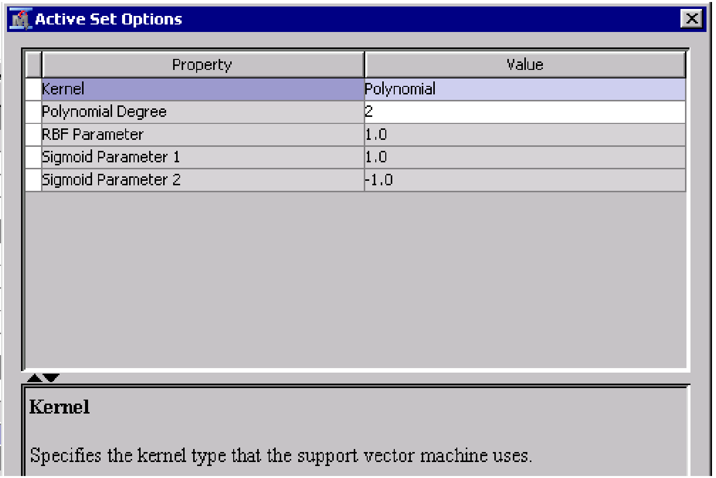
- Non-linear relationship between inputs and output
- Prediction more important than ease of explaining model
- Requires a lot of training data

Support Vector Machines

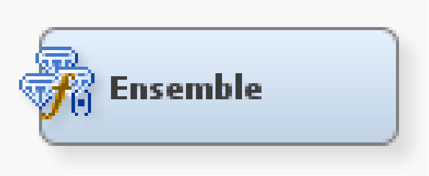


- Enables the creation of linear and non-linear support vector machine models.
- Constructs separating hyperplanes that maximize the margin between two classes.
- Enables the use a variety of kernels: linear, polynomial, radial basis function, and sigmoid function. The node also provides Interior point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

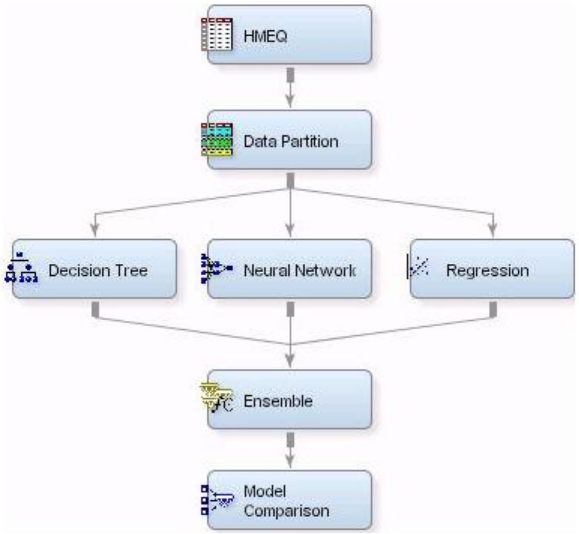


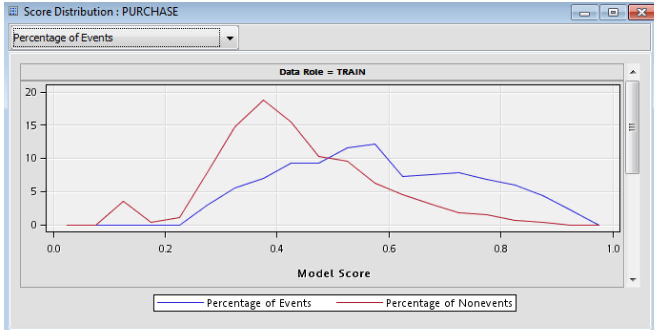


Ensemble

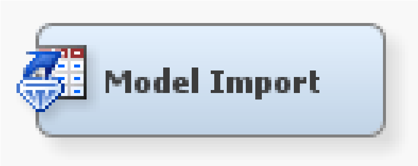


- Creates new models by combining the posterior probabilities (for class targets) or the predicted values (for interval targets) from multiple predecessor models.
- 3 Methods
  - Average
  - Maximum
  - Voting



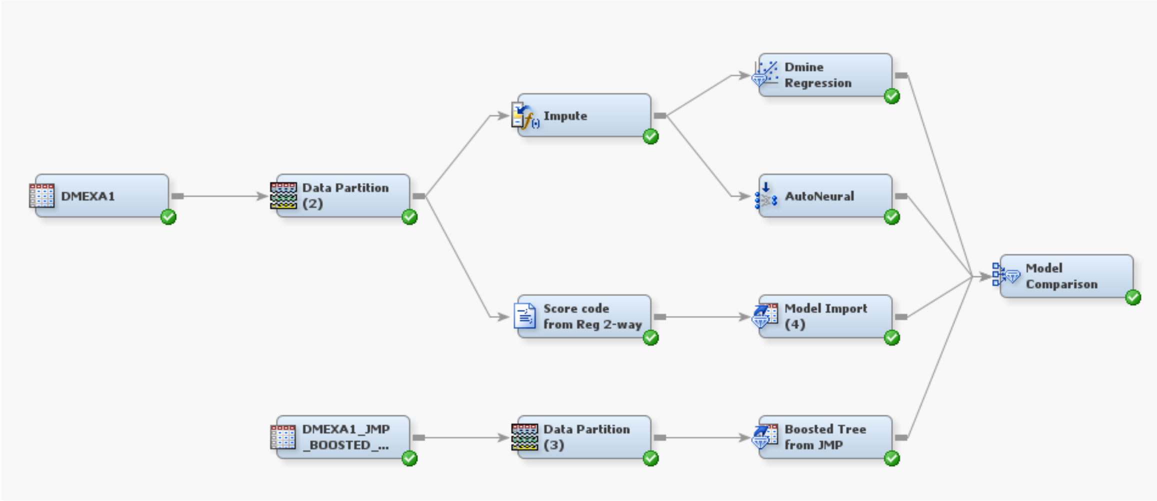


Model Import



- Reads all model details from Metadata Repository
- Applies models to new data and generates all fit statistics
- Compatible with model selection tools
- Useful for sharing models with other users
- Useful testing old models with updated data

- Importing already scored records/cases
- Importing registered SAS Model Package
- Importing SAS Score Code



Tree Based Learners

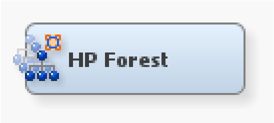
SAS EM Tree Algorithms

- 3 key tree based learning algorithms:

1. Decision Trees

2. Gradient Boosting

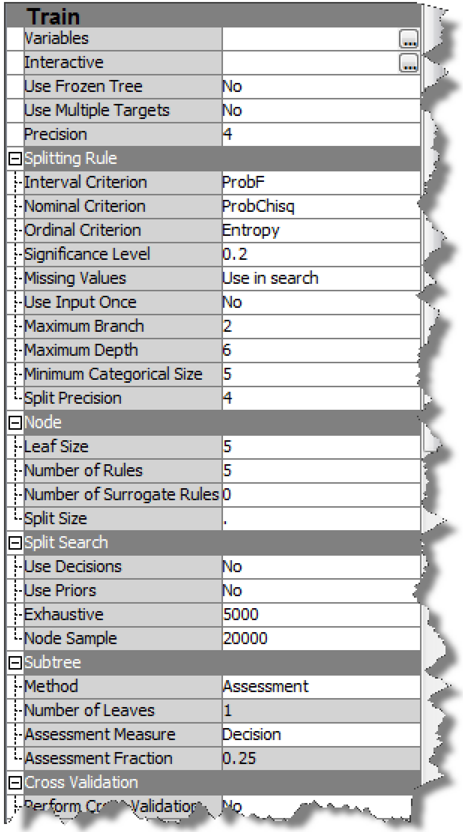
3. Random Forests



Decision Trees

Decision Trees

- Classify observations based on the values of nominal, binary, or ordinal targets
- Predict outcomes for interval targets
- Easy to interpret
- Interactive Trees available
- CART, CHAID, C4.5 approximate



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

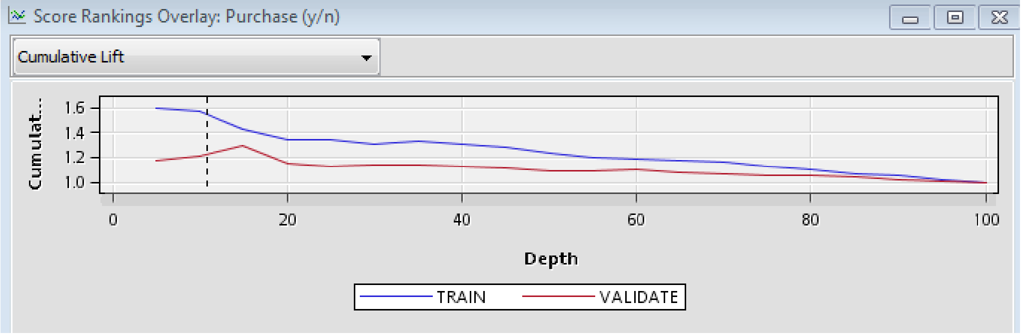
Gradient Boosting

Modelling Algorithms

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	

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

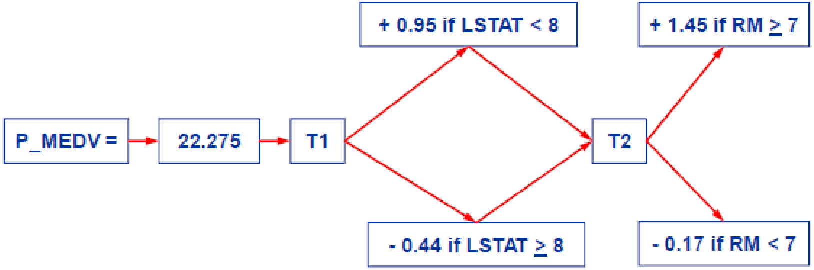


Gradient Boosting

- Approach that resamples the analysis data set several times to generate results that form a weighted average of the re-sampled data set.
- Tree boosting creates a series of decision trees which together form a single predictive model.
- A tree in the series is fit to the residual of the prediction from the earlier trees in the series.
- The residual is defined in terms of the derivative of a loss function.
- The successive samples are adjusted to accommodate previously computed inaccuracies.

Gradient Boosting

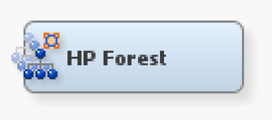
- A gradient boosting tree with an interval target (Median Home Value, MEDV) :



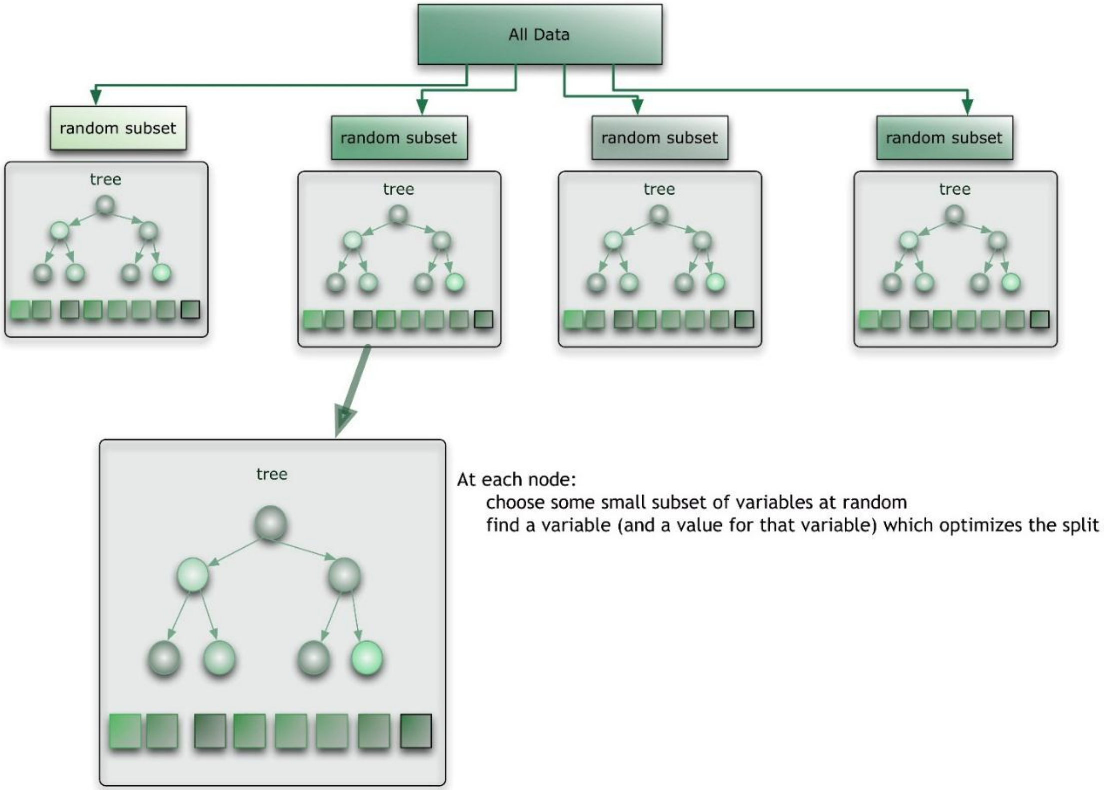
- Number of iterations, M=2; Maximum tree depth = 1
- Resulting model is combination of two decision trees (T1 and T2) each with 2 leaves.
- The value of 22.275 is the mean MEDV, while P_MEDV is the predicted value
- An observation with LSTAT = 6 and RM = 5 would have a P_MEDV value of 22.275 + .95 - .17 = 23.055

Random Forests

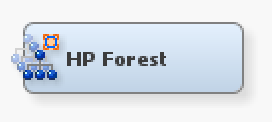
Random Forest Node



What is a Random Forest?



HPForest



- HP node provides increased processing speed
- Random Forest ensemble methodology
  - Samples without replacement
  - Random selection of variables for each tree
  - Uses measures of association to select variable
  - Creates a prediction that is aggregated across the value in the leaf of each tree

Tree Demonstration

Summary

Summary

- EM supports a variety of both supervised and unsupervised modelling algorithms

- Linear / Non-Linear modelling

- Benefits from Tree based learning algorithms include:
  - Interoperability
  - Model performance
  - Outliers/ Missing Values

SAS® FORUM
UNITED KINGDOM 2015

Questions and Answers

Iain.Brown@sas.com

www.SAS.com

SAS® THE POWER TO KNOW®