• SAS Webinar – 10th May 2016 at 10:00 AM BST
• Enterprise Miner: Data Exploration and Visualisation
• The session looks at:
  - Data Visualisation and Sampling
  - Variable Selection
  - Missing Value Imputation
  - Outlier Detection
ROLE OF SAS ENTERPRISE MINER
THE ANALYTICS LIFECYCLE

PREDICTIVE ANALYTICS AND DATA MINING

IDENTIFY / FORMULATE PROBLEM
DATA PREPARATION
DATA EXPLORATION
TRANSFORM & SELECT
BUILD MODEL
VALIDATE MODEL
DEPLOY MODEL
EVALUATE / MONITOR RESULTS

Domain Expertise
Decision Making
Process and ROI Evaluation

Model Validation
Model Deployment
Model Monitoring
Data Preparation

Data Exploration
Data Visualization
Report Creation

Exploratory Analysis
Descriptive Segmentation
Predictive Modeling

Copyright © 2012, SAS Institute Inc. All rights reserved.
THE ANALYTICS LIFECYCLE

PREDICTIVE ANALYTICS AND DATA MINING

Domain Expertise
Decision Making
Process and ROI Evaluation

Data Preparation
Data Exploration
Data Visualization
Report Creation

Model Validation
Model Deployment
Model Monitoring
Data Preparation

Exploratory Analysis
Descriptive Segmentation
Predictive Modeling
• Modern, collaborative, easy-to-use data mining workbench
• Sophisticated set of data preparation and exploration tools
• Modern suite of modeling techniques and methods
• Interactive model comparison, testing and validation
• Automated scoring process delivers faster results
• Open, extensible design for ultimate flexibility
SAS® ENTERPRISE MINER™
MODEL DEVELOPMENT PROCESS

Sample
- Input Data
- File Import
- Sample
- Data Partition
- Merge
- Filter
- Append
- Association
- Cluster
- Variable Selection
- Market Basket
- StatExplore
- Variable Clustering
- MultiPlot
- Path Analysis

Explore
- DMDB
- SOM/Kohonen
- Graph Explore

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

Model
- Decision Tree
- AutoNeural
- Data Mining Regression
- DMNeural
- Ensemble
- Gradient Boosting
- TwoStage
- LARS
- MBR

Assess
- Neural Network
- Model Comparison
- SVM
- Score
- Partial Least Squares
- Segment Profile
- Decisions
- Cutoff
SAS® ENTERPRISE MINER™
MODEL DEVELOPMENT PROCESS

Utility
- Metadata
- SAS Code
- Start Groups
- End Groups
- Control Point
- Reporter
- Score Code Export
- Ext Demo

Apps.
- Open Source Integration
- Register Model
- Incremental Response
- Survival

Time Series
- TS Correlation
- TS Data Preparation
- TS Decomposition
- TS Dimension Reduction
- TS Exponential Smoothing
- TS Similarity

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

Credit Scoring
- Interactive Grouping
- Reject Inference
- Scorecard
SAS® ENTERPRISE MINER™

SEMMA IN ACTION – REPEATABLE PROCESS
DATA VISUALISATION AND SAMPLING
“Quickly find related patterns within a set of data via interactive pictures.”
DATA VISUALISATION AND SAMPLING

SAMPLE AND EXPLORE

- Data selection
  - Required & excluded fields
  - Sample balancing
  - Data partitioning

- Data evaluation
  - Statistical measures
  - Visualization
  - Identifying outliers
  - Analytical segmentation
  - Variable creation & selection
• Sampling
  • **Sample** Node:
    • Stratified / Simple Random Sampling
    • Used for over/under sampling input data
  • **Data Partition** Node:
    • Random sampling into Training, Validation and Test sets
    • Prevent model over fitting
  • **Filter** Node:
    • Select time period of interest
    • Filter based on pre-defined flag
• Segmentation
  • **Cluster** Node
    • Unsupervised, k-means clustering algorithm
    • Data driven
    • Output tree based descriptions
Variable Selection

- **Variable Selection** Node
  - Relationship of independent variables to dependent target
  - R-Square and Chi-square selection criteria

- **Variable Clustering** Node
  - Identify correlations and covariance's between input variables
  - Select Best variable from cluster or Cluster Component

- **Interactive Grouping** Node
  - Computes Weights of Evidence
  - GINI and Information Values for variable selection
**MISSING VALUE IMPUTATION**

**IMPUTE**

- **Missing Value Imputation**
  - **Impute** Node
    - Complete case required for models such as Regression
    - Multiple imputation techniques e.g. Tree, Distribution, Mean, Mode

<table>
<thead>
<tr>
<th>Class (categorical) variables</th>
<th>Interval (numeric) variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input/Target</td>
<td>Input/Target</td>
</tr>
<tr>
<td>Count</td>
<td>Mean</td>
</tr>
<tr>
<td>Default Constant Value</td>
<td>Median</td>
</tr>
<tr>
<td>Distribution</td>
<td>Midrange</td>
</tr>
<tr>
<td>Tree (only for inputs)</td>
<td>Distribution</td>
</tr>
<tr>
<td>Tree Surrogate (only for inputs)</td>
<td>Tree (only for inputs)</td>
</tr>
<tr>
<td>Mid-Minimum Spacing</td>
<td>Tree Surrogate (only for inputs)</td>
</tr>
<tr>
<td>Tukey’s Biweight</td>
<td>Mid-Minimum Spacing</td>
</tr>
<tr>
<td>Huber</td>
<td>Tukey’s Biweight</td>
</tr>
<tr>
<td>Andrew’s Wave</td>
<td>Huber</td>
</tr>
<tr>
<td>Default Constant</td>
<td>Andrew’s Wave</td>
</tr>
<tr>
<td></td>
<td>Default Constant</td>
</tr>
</tbody>
</table>
• **Outlier Detection**
  - **Filter** Node
    - Automated and Interactive filtering
    - Identify and exclude extreme outliers
  - **Replacement** Node
    - Generates score code to process unknown levels when scoring
    - Interactively specify replacement values for class and interval levels

Replace any value $> a$ with the mean.
SUMMARY

• Comprehensive data mining toolset
  • Variety of visualisation and sampling methodologies
  • Number of approaches to data and dimension reduction
  • Importance of enhancing data prior to model development
• Garbage in = Garbage out (GIGO)