ENTERPRISE MINER: ANALYTICAL MODEL DEVELOPMENT
Enterprise Miner: Analytical Model Development

- The session looks at:
  - Supervised and Unsupervised Modelling
  - Classification and Prediction Techniques
  - Tree Based Learners
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

BUSINESS MANAGER

Domain Expert
Makes Decisions
Evaluates Processes and ROI

BUSINESS ANALYST

Data Exploration
Data Visualization
Report Creation

IT SYSTEMS / MANAGEMENT

Model Validation
Model Deployment
Model Monitoring
Data Preparation

DATA MINER / STATISTICIAN

Exploratory Analysis
Descriptive Segmentation
Predictive Modeling

Copyright © 2012, SAS Institute Inc. All rights reserved.
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 EXPERT
Makes Decisions
Evaluates Processes and ROI

BUSINESS MANAGER

DATA EXPLORATION
Data Exploration
Data Visualization
Report Creation

BUSINESS ANALYST

IT SYSTEMS / MANAGEMENT
Model Validation
Model Deployment
Model Monitoring
Data Preparation

DATA MINER / STATISTICIAN
Exploratory Analysis
Descriptive Segmentation
Predictive Modeling

BUSINESS MANAGER
Domain Expert
Makes Decisions
Evaluates Processes and ROI

DATA EXPLORATION

DATA ANALYST

IT SYSTEMS / MANAGEMENT
Model Validation
Model Deployment
Model Monitoring
Data Preparation

DATA MINER / STATISTICIAN
Exploratory Analysis
Descriptive Segmentation
Predictive Modeling

DATA EXPLORATION

DATA ANALYST

IT SYSTEMS / MANAGEMENT
Model Validation
Model Deployment
Model Monitoring
Data Preparation

DATA MINER / STATISTICIAN
Exploratory Analysis
Descriptive Segmentation
Predictive Modeling
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

BUSINESS MANAGER
Domain Expert
Makes Decisions
Evaluates Processes and ROI

BUSINESS ANALYST
Data Exploration
Data Visualization
Report Creation

IT SYSTEMS / MANAGEMENT
Model Validation
Model Deployment
Model Monitoring
Data Preparation

DATA MINER / STATISTICIAN
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
- Path Analysis

Explore
- Association
- Cluster
- Variable Selection
- Market Basket
- StatExplore
- Variable Clustering
- MultiPlot
- Graph Explore

Modify
- DMDB
- SOM/Kohonen
- Transform Variables
- Impute
- Replacement
- Interactive Binning
- Rules Builder
- Drop
- Principal Components

Model
- Decision Tree
- AutoNeural
- Random Regression
- DMNeural
- Ensemble
- Gradient Boosting
- LARS
- MBR

Assess
- Neural Network
- Model Comparison
- SVM
- Partial Least Squares
- Regression
- Rule Induction
- TwoStage
- Model Import
SAS® ENTERPRISE MINER™
SEMMA IN ACTION – REPEATABLE PROCESS
SUPERVISED AND UNSUPERVISED MODELLING
SUPERVISED AND UNSUPERVISED MODELLING

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 AND UNSUPERVISED MODELLING

SUPERVISED LEARNING (CLASSIFICATION & PREDICTION)

<table>
<thead>
<tr>
<th>Supervised Learning</th>
<th>Neural Networks</th>
<th>Regression, least square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decision Trees, CART</td>
<td>Nonlinear SVMs</td>
<td>Generalized Linear Models</td>
</tr>
<tr>
<td>Decision Trees, CHAID</td>
<td></td>
<td>LASSO, LAR</td>
</tr>
<tr>
<td>Gradient Boosting</td>
<td></td>
<td>Splines, MARS</td>
</tr>
<tr>
<td>Random Forests</td>
<td></td>
<td></td>
</tr>
<tr>
<td>kth Nearest Neighbor</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Supervised and Unsupervised Modelling

**Unsupervised Learning**

- **K-means**
- **Fuzzy K-means**
- **Hierarchical Clustering**
- **Vector Quantization**

- **Multidimensional Scaling**
- **Principal Components**
- **Nonnegative Matrix Factorization**
- **Associations, Apriori**
Regression

- Linear
- Logistic

- 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.
Generalized 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.
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

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

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
TREE BASED LEARNERS
• 3 key tree based learning algorithms:

1. Decision Trees
2. Gradient Boosting
3. Random Forests
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
- CART, CHAID, C4.5 approximate
GRADIENT BOOSTING
Gradient Boosting

- 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.
RANDOM FORESTS
What is a Random Forest?

At each node:
- choose some small subset of variables at random
- find a variable (and a value for that variable) which optimizes the split
• 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
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
QUESTIONS AND ANSWERS