Predictive Modeling Using SAS Enterprise Miner 14 Exam

During the testing of these objectives; you will be expected to perform common tasks, such as:

- Create a new project in Enterprise Miner
- Open an existing project in Enterprise Miner
- Add diagrams to projects in Enterprise Miner
- Create libraries within Enterprise Miner
- Add nodes to diagrams in Enterprise Miner
- Copy nodes within Enterprise Miner
- Connect nodes to create process flows in Enterprise Miner
- Change interactive sampling methods for data exploration
- Work with the Help functionality within Enterprise Miner

Data Sources - 20-25%

Create data sources from SAS tables in Enterprise Miner

- Use the Basic Metadata Advisor
- Use the Advanced Metadata Advisor
- Customize the Advanced Metadata Advisor
- Set Role and Level meta data for data source variables
- Set the Role of the table (raw, scoring, transactional, etc)

Explore and assess data sources

- Create and interpret plots, including Histograms, Pie charts, Scatter plot, Time series, Box plot
- Identify distributions
- Find outlying observations
- Find number (or percent) of missing observations
- Find levels of nominal variables
- Explore associations between variables using plots by highlighting and selecting data
- Compare balanced and actual response rates when oversampling has been performed
- Explore data with the STAT EXPLORER node.
- Explore input variable sample statistics
- Browse data set observations (cases)

Modify source data

- Replace zero values with missing indicators using the REPLACEMENT node
- Use the TRANSFORMATION node to be able to correct problems with input data sources, such as variable distribution or outliers.
Use the IMPUTE node to impute missing values and create missing value indicators
Reduce the levels of a categorical variable
Use the FILTER node to remove cases

Prepare data to be submitted to a predictive model

Select a portion of a data set using the SAMPLE node
Partition data with the PARTITION Node
Use the VARIABLE SELECTION node to identify important variables to be included in a predictive model.
Use the PARTIAL LEAST SQUARES node to identify important variables to be included in a predictive model.
Use a DECISION TREE or REGRESSION nodes to identify important variables to be included in a predictive model.

Building Predictive Models - 35-40%

Describe key predictive modeling terms and concepts

Data partitioning: training, validation, test data sets
Observations (cases), independent (input) variables, dependent (target) variables
Measurement scales: Interval, ordinal, nominal (categorical), binary variables
Prediction types: decisions, rankings, estimates
Dimensionality, redundancy, irrelevancy
Decision trees, neural networks, regression models
Model optimization, overfitting, underfitting, model selection
Describe ensemble models

Build predictive models using decision trees

Explain how decision trees identify split points
Build decision trees in interactive mode
Change splitting rules
Explain how missing values can be handled by decision trees
Assess probability using a decision tree
Prune decision trees
Adjust properties of the DECISION TREE node, including: subtree method, Number of Branches, Leaf Size, Significance Level, Surrogate Rules, Bonferroni Adjustment
Interpret results of the decision tree node, including: trees, leaf statistics, treemaps, score rankings overlay, fit statistics, output, variable importance, subtree assessment plots
Explore model output (exported) data sets

Build predictive models using regression

Explain the relationship between target variable and regression technique
Explain linear regression
Explain logistic regression (Logit link function, maximum likelihood)
Explain the impact of missing values on regression models
Select inputs for regression models using forward, backward, stepwise selection techniques
Adjust thresholds for including variables in a model
Interpret a logistic regression model using log odds
Interpret the results of a REGRESSION node (Output, Fit Statistics, Score Ranking Overlay charts)
Use fit statistics and iteration plots to select the optimum regression model for different decision types
Add polynomial regression terms to regression models.
Determine when to add polynomial terms to linear regression models.

Build predictive models using neural networks

- Theory of neural networks (Hidden units, Tanh function, bias vs intercept, variable standardization)
- Build a neural network model
- Use regression models to select inputs for a neural network
- Explain how neural networks optimize their model (stopped training)
- Recognize overfit neural network models.
- Interpret the results of a NEURAL NETWORK node, including: Output, Fit Statistics, Iteration Plots, and Score Rankings Overlay charts

Predictive Model Assessment and Implementation - 25-30%

Use the correct fit statistic for different prediction types

- Misclassification
- Average Square Error
- Profit/Loss
- Other standard model fit statistics

Use decision processing to adjust for oversampling (separate sampling)

- Explain reasons for oversampling data
- Adjust prior probabilities

Use profit/loss information to assess model performance

- Build a profit/loss matrix
- Add a profit/loss matrix to a predictive model
- Determine an appropriate value to use for expected profit/loss for primary outcome
- Optimize models based on expected profit/loss

Compare models with the MODEL COMPARISON node

- Model assessment statistics
- ROC Chart
- Score Rankings Chart, including (cumulative) % response chart, (cumulative) Lift chart, gains chart.
• Total expected profit
• Effect of oversampling

Score data sets within Enterprise Miner

• Configure a data set to be scored in Enterprise Miner
• Use the SCORE node to score new data
• Save scored data to an external location with the SAVE DATA node
• Export SAS score code

Pattern Analysis - 10-15% (new content)

Identify clusters of similar data with the CLUSTER and SEGMENT PROFILE nodes

• Select variables to use to define the clusters
• Standardize variable scales
• Explore clusters with results output and plots
• Compare distribution of variables within clusters

Perform association and sequence analysis (market basket analysis)

• Explain association concepts (Support, confidence, expected confidence, lift, difference between association and sequence rules)
• Create a data set for association analysis
• Interpret the results and graphs of the ASSOCIATION node.