Notes for Training

Ali Elsayed

Imputation:-
- replaces missing values in data sets
- include replacing a missing value in an input variable with the mean or mode of that variable's non-missing values

Manage Variables:-
- enables you to make modifications (such as changing the role of a variable or adding new transformations) to the data

Replacement:-
- enables you to replace outliers and unknown class levels with specified values

Text Mining:-
- enables you to process text data in a document collection
- **Text parsing** processes textual data into a term-by-document frequency matrix

Transformations:-
- enables you to alter your data by replacing an input variable with some function of that variable
- used to stabilize variances, remove nonlinearity, and correct non-normality

Variable Clustering:-
- divides numeric variables into disjoint clusters and chooses a variable that represents each cluster.
- removes collinearity, decreases redundancy, and helps reveal the underlying structure of the data set.

Variable Selection:-
- uses several unsupervised and supervised methods to determine which variables have the most impact on the model.

Linear regression:-
- Fits an ordinary least squares regression model for an interval target.
- Linear regression and GLM for interval target.

Logistic regression:-
- Fits a logistic regression model for a binary or nominal target.

SVM:-
- Fits a support vector machine via interior-point optimization for a binary or interval target.
- Modeling linear or linearly separable phenomena by using linear kernels or polynomial kernels up to degree three

Neural Network:-
- For Interval or Binary and Nominal
- Fits a fully-connected neural network model.

Notes for Training

Ali Elsayed

Gradient Boosting:-

- evolved from the application of boosting methods to regression trees
- builds a sequential series of decision trees.

Prediction Type:-

- Decisions
- Rankings
- Estimates

Some tables

Data Exploration Results:-

- **Important Inputs table:-**
  - bar chart and examine the relative importance of the ranked variables
- **Interval Variable Moments table:-**
  - This table displays the interval variables with their associated statistics, which include minimum, maximum, mean, standard deviation, skewness, kurtosis, relative variability, and the mean plus or minus two standard deviations
- **the Interval Variable Summaries scatter plot:-**
  - Observe that several variables have deviation from normality—that is, high kurtosis on the Y axis and high skewness on the X axis

Replacement Results:-

- **The Interval Variables table:-**
  - shows which variables now have a lower limit of 0

Text Mining Results:-

- **the Kept Terms and Dropped Terms tables:-**
  - These tables include terms used and ignored, respectively, during the text analysis
- **Topics table:-**
  - This table shows topics created by the Text Mining node.

Transformations Results:-

- **Transformed Variables Summary table:-**
  - This table displays information about the transformed variables, including how they were transformed, the corresponding input variable, the formula applied, the variable level, type, and variable label.

Variable Selection Results:-

- **Variable Selection table:-**
  - This table contains the output role for each variable. At the top of the table are the input variables selected by the node. These variables have a blank cell in the Reason column.

Notes for Training

Ali Elsayed

Decision Tree Results:-

- **Tree Diagram:-**
    - which presents the final tree structure for this particular model, such as the depth of the tree and all end leaves.
- **The Pruning Error plot:-**
    - shows the model's performance based on the misclassification rate because the target is binary
    - This plot shows how the average squared error changes for subtrees
- **The Variable Importance table:-**
    - shows the input variables that are most significant to the final model. The most important input variable has its relative importance as 1 and all others are measured based on the most important input.
- **The Output window:-**
    - shows the final decision tree model parameters, the Variable Importance table, and the pruning iterations.

Assessment tab…

- **the Cumulative Lift:-**
    - showing the model's performance ordered by the percentage of the population
- **the ROC curve:-**
    - which shows the model's performance considering the true positive rate and the false positive rate and F1
    - For a binary target only
    - ROC is very useful for deployment
- **The Fit Statistics output:-**
    - shows the model's performance based on several assessment measures, such as average squared error and misclassification rate for (binary data)
- **the Event Classification output:-**
    - which shows the confusion matrix at various cutoff values for each partition. The default view is based on percentages.

Notes for Training

Ali Elsayed

Neural Network Results:-

- **Network Diagram:-**
  - which presents the final neural network structure for this model, including the hidden layer and the hidden units
  - displays the input nodes, hidden nodes, connections, and output nodes of a neural network. Nodes are represented as circles, and links between the nodes are lines connecting two circles. The size of the circle represents the magnitude of the absolute value of that node, relative to the model, and the colour indicates whether that value is positive or negative. Similarly, the size of the line between two nodes indicates the strength of the link, and the colour indicates whether that value is positive or negative
- **The Iteration plot**
  - shows the model's performance based on the valid error throughout the training process when new iterations are added to achieve the final model

The Insights tab:-

- contains summary information in the form of a report for the project, the champion model, and any challenger models. For the purposes of the Insights tab, a champion model is the overall project champion model, and a challenger model is one that is a pipeline champion, but not the overall project champion
- At the top of the report is a summary of the project and a list of any project notes. Summary information about the project includes the target variable, the champion model, the event rate, and the number of pipelines in the project.
- **Most Common Variables Selected Across All Models table:-**
  - This plot summarizes common variables used in the project by displaying the number of pipeline champion models that the variables end up in. Only variables that appear in models used in the pipeline comparison are displayed.
- **The Assessment for All Models plot:-**
  - This plot summaries model performance for the champion model across each pipeline and the overall project champion.
- **The Most Important Variables for Champion Model plot:-**
  - This plot shows the most important variables, as determined by the relative importance calculated using the actual overall champion model.
- **The Cumulative Lift for Champion Model plot:-**
  - This plot displays the cumulative lift for the overall project champion model for both the training and validation partitions.