Contributing author: Matthew David

Decision Tree and Forest Models

Summary: These are two classification models. These models help identify what group a data point belongs to. Decision Tree and Forest models can help predict classification of categorical or continuous variables.

STEP 1: Create sample

In any classification problem you will need to set an estimation sample and a validation sample of your data. This helps us compare different classification models to see which better fit the data.

Useful Alteryx tool: Create Sample

STEP 2: Model Settings

Select a target variable and predictor variables, you can include as many predictor variables as you would like because the model will only use variables that work best. Specify the number of records needed to allow for a split, the smaller the number the more splits you will get. In the Forest Model you can choose the number of trees to use.

Useful Alteryx tool: Forest, Decision Tree

STEP 3: Interpreting the Report

Root Node Error in the Decision Tree model is the percentage of how many of the data points went to the incorrect terminal node (predicted incorrectly) when all of the data points are validated against themselves within the entire training set (the Estimation dataset). The Pruning Plot lists out the levels in the decision tree with their related error terms with cross-validation samples.

The Variable Importance Plot is a bar graph that's length indicates the importance of the predictor variables. The Confusion Matrix is a matrix (or table) that lists out all of the possible prediction results when we validate our model against itself.

The Out of the Bag Error Rate for the Forest Model explains how well the model performed with the cross-validation set in the estimation data. Similar to R-squared. The Percentage Error for Different Number of Trees graph helps us see what the correct number of trees is to use, so we can avoid over computing in the future. What we are looking for where does the graph flatline?

Useful Alteryx tools: Forest, Decision Tree

STEP 4: Model Comparison

Use the fit and error measures, Accuracy which represents the overall accuracy, the number of correct predictions of all classes divided by total sample number. The F1 score is calculated the following way, precision * recall / (precision + recall) You can read more about precision and recall. There will also be a confusion matrix in this reprot to show how the models compared to the validation set. This confusion matrix is one of the best methods to review the accuracy and precision of your model as well as to understand any model bias in classifying your data points.

Useful Alteryx tool: Model Comparison

STEP 5: Score Data

Apply the model by attaching a score tool to the data set you are trying to classify and the model object.

Useful Alteryx tool: Score