Lachlan Sinclair

2/24/2020

Week 7 Assignment 2

Reflections

a. 10 fold-cross validation for this dataset suggests that the CART model performs better than the logistic regression model. For both models results I am considering the average rate of correctly assigning the class and the variance of the test set to be more important than the training set. Also, as it wasn't specified, I went with the sample variance since the two subsets are only partial samples of the overall population. The CART model not only had significantly better probability of predicting correctly but also had a lower variance. This is also the case for the average and variance of the training dataset as well. Both models produce relatively low variances for the training set (with the CART models still being lower), however the variance of the test set for the LR model was twice the CART models test set variance. This was also the case for the training set but the magnitude was quite smaller. The CART model's average prediction rate for the test set was about 6% greater than that of the LR models, it is clear that the 10-fold-cross validation suggests the CART model outperforms the LR model for this dataset.

b. A K-fold-cross validation allows for multiple models, that are trained in the same way, to be developed over a single dataset. More specifically it is randomly partitioning the dataset into distinct folds, the same model is then trained over specific combinations of these folds in that create unique training/testing sets. The focus of the K-fold-cross validation process is to validate a specific model and give the modeler some insight into its potential viability once it is generalized outside of the provided data. A random forest is a method of generating different tree models randomly. Random forest are focused on the way the model is being generated and not a method of partitioning data. Random forests generate random trees by allowing trees to select from a limited number of randomly selected attributes to split on, it repeatedly selects random sets of attributes to choose from until the tree is complete, it then repeats this process again for multiple trees. Random forests ensure a multitude of trees are explored, allowing for the best performing one the be identified. This is because the optimal split order that will be selected by whatever information gain selecting algorithm is being used may not result in the optimal model. So, K-fold-cross validation is used to provide valuable statistical information about a model using a given dataset and the random forest is used to find the optimal tree model for a given set of data.