Quiz # 3. AMS 580

Name:________________________________SBU ID:_________________________________

Dear all, you have the entire lecture time to do this quiz. It is open book however you must do the work entirely on your own. Please turn on your video and mute your audio. For those of you who are not in the classroom, please connect to zoom with both your computer and your cellphone. Please turn in your quiz by 9:50am.

For each part, please submit your answer in two documents: 1. The Rmd file, and 2. The output file (word or pdf) – to the SBU Brightspace (BS). Therefore, if you have tried both parts, you should submit 4 files. If you have difficulty submitting to BS, you may then email the solutions (with the subject of "Quiz 3, AMS 580", and in one email with the two attachments), to your TA Ian at: weihao.wang@stonybrook.edu

Part I. CART with the GreatUnknown Data – Classification Task

The **GreatUnknown.csv** data contain 12 predictors and one binary response variable y (= 0 or 1), which is the true class label**.**

For this dataset, **sensitivity** is defined as a case labeled 1 being classified to label 1, while **specificity** is defined as a case labeled 0 being classified to label 0.

1. For the entire dataset, please perform the data cleaning as instructed before; namely, delete observations with missing value(s). Please report how many cases (namely, data points) are left after this step. Then please use the random seed 456 to divide the cleaned data into 75% training and 25% testing.

2. Please first build a fully grown tree using the training data, and draw the tree plot using *rattle*. Next please use this tree to predict class label in the testing data. Please compute the Confusion matrix and report the sensitivity, specificity, and the overall accuracy for the testing data.

3. To make the tree more robust, we shall prune the fully grown tree using the training data with 10-fold cross-validation. Please (1) show the complexity plot, (2) report the best CP value, and (3) draw the pruned tree using *rattle*.

4. Please use this optimal pruned tree to predict the class label in the testing data. Please compute the Confusion matrix and report the sensitivity, specificity and the overall accuracy for the testing data.

5. Please fit a logistic regression model using the training data and use this model to predict whether each check is forged or not in the testing data. Please compute the Confusion matrix and report the sensitivity, specificity, and the overall accuracy for the testing data.

6. Please output the test data along with the prediction results from (1) the Full Tree, (2) the Optimal Pruned Tree, and (3) the logistic regression, in an excel file. Please generate the majority vote using these three prediction methods. Please compute the Confusion matrix and report the sensitivity, specificity and the overall accuracy for the testing data based on the majority vote.

Part II. Decision Tree with the Tennis Data – Extra Credit Problem

For the **Tennis.xlsx** data, **p**lease write an Rmd program to decide on the root node in the decision tree using the **Gini index**.

This resembles the example in your lecture notes where we used the **Shannon entropy** to determine the root node for the Tennis data.