FE-582 – Assignment 4
Spring 2021

Problem 1
This problem use the OJ data set (OJ.csv).

a) Create a training set containing a random sample of 800 observations, and a test set containing the remaining observations.
b) Fit a tree to the training data, with Purchase as the response and the other variables as predictors. Use the summary() function to produce summary statistics about the tree, and describe the results obtained. What is the training error rate? How many terminal nodes does the tree have?
c) Type in the name of the tree object in order to get a detailed text output. Pick one of the terminal nodes, and interpret the information displayed.
d) Create a plot of the tree, and interpret the results.
e) Predict the response on the test data, and produce a confusion matrix comparing the test labels to the predicted test labels. What is the test error rate?
f) Apply the cv.tree() function to the training set in order to determine the optimal tree size.
g) Produce a plot with tree size on the x-axis and cross-validated classification error rate on the y-axis.

Problem 2
This problem use the Caravan data set (CARAVAN.csv).

a) Create a training set consisting of the first 1,000 observations, and a test set consisting of the remaining observations.
b) Fit a boosting model to the training set with Purchase as the response and the other variables as predictors. Use 1,000 trees, and a shrinkage value of 0.01. Which predictors appear to be the most important?
c) Use the boosting model to predict the response on the test data. Predict that a person will make a purchase if the estimated probability of purchase is greater than 20 %. Form a confusion matrix. What fraction of the people predicted to make a purchase do in fact make one? How does this compare with the results obtained from applying KNN or logistic regression to this data set?