Due on November 30, 2018 (No Quiz)

Submit wordprocessed report with font size 11. No screenshots. Use set.seed(1) as frequent as needed.

1. (From Homework 5) The `hsb` dataframe from library `faraway` was collected as a subset of the High School and Beyond study conducted by the National Education Longitudinal Studies program of the National Center for Education Statistics. The response `prog` is multinomial with three levels.

   a) Fit a bagged categorical tree to predict the program of choice `prog` with two predictors `math` and `science`. Make a 2-in-1 color plot showing the observed and predicted responses with different colors. Find the overall error rate.

2. (Categorical trees) Consider the `Caravan` data set. Use `sapply(Caravan,table)` to see that some predictors are highly unbalanced (most rows belong to a few levels). Then remove variables `PVRAAUT` and `AVRAAUT`. Create a training set consisting of the first 1,000 observations, and a test set consisting of the remaining observations.

   a) Fit a random forest model to the training set with `Purchase` as the response and the other variables as predictors. Name the predictor that appears to be the most important?

   b) Fit a boosting model to the training set with `Purchase` as the response and the other variables as predictors. Use 1000 trees, and a shrinkage value of 0.01. Name the predictor that appears to be the most important?