**Assignment: Tree Based Methods, Regression Trees**

BUAD 5082 – Spring 2019

1. **Objectives**

The purpose of this assignment is to provide you with an opportunity to investigate some of the ideas regarding tree based methods and how to create regression trees.

1. **What You Will Need**

* Access to a Windows computer with R, and to the following files, which can be downloaded from the Class Schedule page of the course web site:

1. **What You Will Hand In**

Nothing, it is not expected for you to submit this assignment.

1. **Due Date**

February 27th, 2019: just before midnight. After this date we will post the solutions to the problem set.

1. **Note on Collaboration**

This is not an official assignment. You may work with others solve this problem set.

1. **Preliminaries:**

To get set up for the assignment, follow these steps:

1. As the first statement in your script file, enter rm(list=ls())
2. Each question and part in the assignment should begin with the following three comment lines where n is the question number and p is the part letter:

###########################

####QUESTION n PART p####

###########################

1. Do not create console output other than what is asked of you explicitly.
2. We suggest that you read the entire assignment before starting – there are often notes and suggestions at the end of assignment documents.
3. **Assignment Tasks:**

**Question 1: Predicting MPG: Tree Based Methods using Regression Trees**

**In the following steps you will use the Auto dataset to construct a tree regression model to predict mpg using most of the predictors.**

1. rm(list=ls()) and set the random seed to 5072.
2. Create a data frame for the Auto dataset. Be sure to omit any NA values.
3. Omit non-numeric predictors.
4. Split the data into training, validation, and test sets using the sample() function as usual. The training set should be approximately 60% of the total number of rows. The validate and test sets should be approximately 20% of the total number of rows each.
5. Perform tree-based linear regression on the training data in order to predict mpg using all the variables except for the name variable.

Hint: use the rpart function for this regression

1. Print the total number of terminal nodes found from the Part E regression
2. Plot the regression tree in the graphics window. Be sure to include a Title.
3. Print the name of the variables in order in which they are used to split the regression tree.
4. For the test set, use the model just created to predict mpg for each data point. Use head to display only the first six predictions.

**Question 2: Predicting Hitters’ Salaries: Tree Based Methods using Regression Trees**

**In the following steps you will use the Hitters dataset to construct and prune a regression tree.**

1. rm(list=ls()) and set the random seed to 5072.
2. Omit any NA values from the Hitters dataset.
3. Take the log of the dependent variable for this problem which is salary.
4. Omit any non-numeric variables. Prepare the Hitters dataset to be split.
5. Now repeat Steps D to H in Question 1 on this revised Hitters dataset.
6. For the training set, use the model just created to predict salaries for each data point. Use head to display only the first six predictions.
7. Calculate the test MSE.
8. Evaluate the predictive power of this model using the validate dataset.
9. Prune the regression tree by minimizing the cross-validated error. Display this regression tree in the graphics window.
10. Print the number of terminal nodes in the pruned tree.

**Question 3: To Gain More Understanding about Regression Trees**

1. Describe the parts of the regression tree.
2. Define recursive binary splitting. Be sure to mention greedy algorithms.
3. Explain the advantages of why one would choose to use a regression tree to explain data.
4. Why should one prune a regression tree?
5. I want to build a regression tree but I don’t want it to be too big for whatever reason. What are some stopping criteria we can implement to cause the regression tree to be smaller?