**BU.510.650 Homework #5**

**Data Analytics Page 1 of 2**

**Spring 2021 The Johns Hopkins University**

**Arnab Bisi Carey Business School**

**Homework #5**

Due: 03/07/21, 11:59pm

**Attention:** Homeworks #1 to #6 are all individual assignments but you are always free to ask the instructor/teaching assistant questions as needed. Copy your e-mail to both of us.

For all homework assignments prepare two files. The first is an MSWord file for your answers including any necessary figures. Name this file based on your name as in “*LastName.FirstName.HWNumber.docx”.* This file should be both concise and complete. We should not have to open your **R** file to find your answers. On the other hand, this file should only contain what is necessary to answer the question.

The second file is your **R** Script. Make this file name *“LastName.FirstName.HWNumber.R*”. We will use this file if we need more information about your answers. Both files are to be submitted via Blackboard.

1. For the Smarket data set from “ISLR” library, using the 2005 data as the test data and the remaining as the training data, what is your recommended value of K in the K-Nearest Neighbors (KNN) approach and why? (2 points)

1. Using Carseats data set from “ISLR” library and installing “tree” library, run the R codes provided in the file “Carseats-DecisionTree-Example.R” to demonstrate decision tree method and submit your output here. (2 points)
2. In this question, you will use the K-Nearest Neighbors (KNN) algorithm to predict whether a passenger will survive or not.

To begin your work on this question, first read the data from the file "TitanicforKNN.csv" to a data frame named Titanic.

\*\*Note: Please review the data before proceeding. You will notice that I already converted all the categorical variables (Gender, Fare, Class) into 0-1 columns. I did so, because KNN does not work well with non-numeric variables.\*\*

Next, split the data into training data and test data, using random selection. Include half of the records in the training data and the rest in the test data. (\*\*Remember to include set.seed(1) before the random selection in your code, so we all end up making the same split.\*\*)

1. Run the KNN algorithm to predict the response variable Survived for each passenger in the test data. Do this for K = 1, 3, and 5. According to these predictions for K = 1, 3, and 5, what is the proportion of passengers in the test data that will survive? (2 points)

R Hints: To run the function knn(), recall that you need four inputs: (i) a matrix that contains the values of predictors in the training data, (ii) a matrix that contains the values of predictors in the test data, (iii) a vector containing the values of the response (Survived) in the training data, (iv) a value for K. To obtain (i), remove the Survived column from the training data. To obtain (ii), remove the Survived column from the test data. To obtain (iii), create a vector that stores the values of Survived column in the training data. See the Smarket-KNN-Example.R in Session 5 for a related example.

1. For each K, compute the accuracy of predictions for the test data. Which K works best in this case?

(1 point)

1. In this question, you will estimate a decision tree for the AutoLoss data. The data file for this question, AutoLoss-DT.csv, is slightly different from the previous data file. In particular, instead of the actual loss amount for each vehicle, it has a column called HighLoss, which indicates whether the loss is high (“Yes”) or low (“No”) for each vehicle. Our goal is to create a decision tree that predicts whether the loss for a vehicle will be high or low.

To begin your work on this question, run the following two lines of code: The first one replaces ?s with NA while reading the data from the .csv file, and the second one removes all the observations with any NA.

AutoLoss <- read.csv("AutoLoss-DT.csv", na.strings = "?")

AutoLoss <- na.omit(AutoLoss)

**\*\*Please include set.seed(15) once at the beginning of your code, so we all get the same results.\*\***

1. Fit a decision tree to the entire data, with HighLoss as the response and all other variables as predictors. Plot the tree (including the names of predictors in the plot) and answer the following questions: Which predictors are used at the nodes of the tree? How many terminal nodes (leaves) does the tree have? (1 point)
2. Determine the best tree size, using cross-validation and pruning. (See how we accomplished this in TASK 7 of Carseats example.) Plot the tree you obtained (including the names of predictors in the plot). (1 point)
3. Use the best tree to answer the following question (you do not need to use R for this): Suppose my car fits the description shown below. Will this car incur a high loss or not?

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| FuelType | Aspiration | NumDoors | BodyStyle | DriveWheels | Length | Width | Height |
| gas | std | four | wagon | 4wd | 177 | 71 | 61 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Weight | EngineSize | Horsepower | PeakRPM | Citympg | Price |
| 3527 | 122 | 241 | 5000 | 22 | 25000 |

(1 point)