Week 7 Assignment (60 Points)

1. This assignment is on classification tree (20 Points)
   1. Load the bostonhousing.csv into RStudio.
   2. Explore data.
      1. How many rows of data do we have?
      2. How many attributes?
   3. Apply classification trees to the *MEDV\_CAT* (for classification tree, target attribute should be categorical and numeric format and this attribute it is).
   4. Compute the a-priori (prior probability of each class) probabilities in the data file.
   5. Partition the data, training (70%) and test (30%). Use 2021 as the seed.
   6. Build the tree model. Ignore minisplit, use minibucket = 10 and cp = 0
   7. Create the tree graph. Interpret the root and the level 1 nodes
   8. Evaluating the tree by creating confusion matrix of the training and test datasets.
   9. Rebuild the tree on training dataset. Ignore minibucket, use minisplit = 10 and cp = 0.
   10. Create the new tree graph
   11. Plot the *CP* graph of the new tree model. What value of the *CP* gives the best tree model?
   12. Prune the tree with the best CP value.
   13. Create the graph of the pruned tree.
   14. Apply the pruned tree model on the test dataset. Evaluate the confusion matrix of result.
2. Use the UniversalBank.csv to build a logistic regression model. Include the following in this task. (20 Points)
   1. Use the following R code to convert the education attribute into categorical format. This attribute is already categorical but has numeric format.

factor(<datasetname>$<attributename>, levels = c(current level), labels = c(category names))

* 1. Partition dataset into training 60% and valid 40%. Use 2023 for seed
  2. Build the model on training set.
  3. View the output summary
  4. What attributes are non-significant.
  5. What is legit? What is the logistic regression probability function? (You may abbreviate the predictors’ name.)
  6. Rebuild the model with only with significant attributes.
  7. Apply the model on the validation set.
  8. Build the validation dataset confusion matrix with 50% cutoff.
  9. Plot lift and decile-wise charts
  10. Export a csv file which has the valid dataset as well as predicted values.

1. Do all sections of problem 10.3 in the page 290 of your textbook except part g.

(20 Points)

Part g is a bonus. Use 50% cutoff probability (5 points)

My notes on these LR algorithms as well as textbook chapter 10 have all info you need to do this assignment

**Submit all R codes, all requested information, all requested plot and graphs, and all confusion matrices with your evaluations.**