BUAN/MIS 6356.001

Assignment 2 (50 points)

Assignment Date: 9/20/2024

Assignment Due Date: 10/03/2024

The data files required for the assignment are available in the 'Data' folder on the course elearning page. Download the zip file in the folder to access the data. The folder also has several other data files that are required for this assignment.

Submit all your answers in a single document. For the questions that require you to create a visualization or code using R, include the code snap and plots in the document. A separate .r file with a working R code that generates all the plots should be submitted as well.

1. **Predicting Boston Housing Prices:** The file BostonHousing.csv contains information collected by the US Bureau of the Census concerning housing in the area of Boston, Massachusetts. The dataset includes information on 506 census housing tracts in the Boston area. The goal is to predict the median house price in new tracts based on information such as crime rate, pollution, and number of rooms. The dataset contains 13 predictors, and the response is the median house price (MEDV). The following terms describe each of the predictors and the response.

   CRIM - Per capita crime rate by town

   ZN - Proportion of residential land zoned for lots over 25,000 ft2

   INDUS - Proportion of nonretail business acres per town

   CHAS - Charles River dummy variable (= 1 if tract bounds river; = 0 otherwise)

   NOX - Nitric oxide concentration (parts per 10 million)

   RM - Average number of rooms per dwelling

   AGE - Proportion of owner-occupied units built prior to 1940

   DIS - Weighted distances to five Boston employment centers

   RAD - Index of accessibility to radial highways

   TAX - Full-value property-tax rate per $10,000

   PTRATIO - Pupil/teacher ratio by town

   LSTAT - Percentage lower status of the population

MEDV - Median value of owner-occupied homes in $1000s

BUAN/MIS 6356.502

a. Why should the data be partitioned into training and validation sets? What will the training set be used for? What will the validation set be used for? (5 points)
b. Fit a multiple linear regression model to the median house price (MEDV) as a function of CRIM, CHAS, and RM. Write the equation for predicting the median house price from the predictors in the model. (5 points)
c. Using the estimated regression model, what median house price is predicted for a tract in the Boston area that does not bind the Charles River, has a crime rate of 0.1, and where the average number of rooms per house is 6? What is the prediction error? (5 points)
d. Reduce the number of predictors: (35 points)
    i. Which predictors are likely to be measuring the same thing among the 13 predictors? Discuss the relationships among INDUS, NOX, and TAX. (10 points)
    ii. Compute the correlation table for the 12 numerical predictors and search for highly correlated pairs. These have potential redundancy and can cause multi-collinearity.

Choose which ones to remove based on this table. (10 points) iii. Use stepwise regression with the three options (backward, forward, both) to reduce the remaining predictors as follows: Run stepwise on the training set. Choose the top model from each stepwise run. Then use each of these models separately to predict the validation set. Compare RMSE, MAPE, and mean error, as well as lift charts. Finally, describe the best model. (15 points) ** Refer to the textbook chapter 6 for further reference!

ALL THE BEST!