Machine Learning Assignment 2
June(Liqiong) Sun# 75956961

1.1 Explain, in two or three brief sentences, what bagging is and why it is useful.

Bagging is a method that generates B data set, from randomly sampling n observations with replacement from a data set with n observations B times. It is useful because averaging a set of observations reduces variance.

1.2 Consider the two-predictor case (say X_1 and X_2). Which of the following partitions of the predictor space correspond to a tree -- Figure 1 or Figure 2? If the letters represent the decisions in each case, write out the decision tree for this partition.

Figure 1

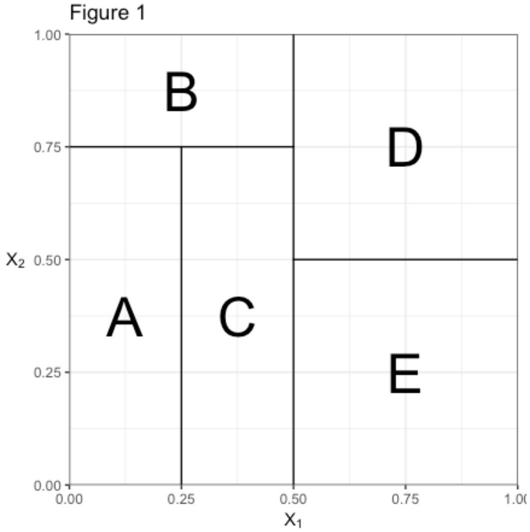
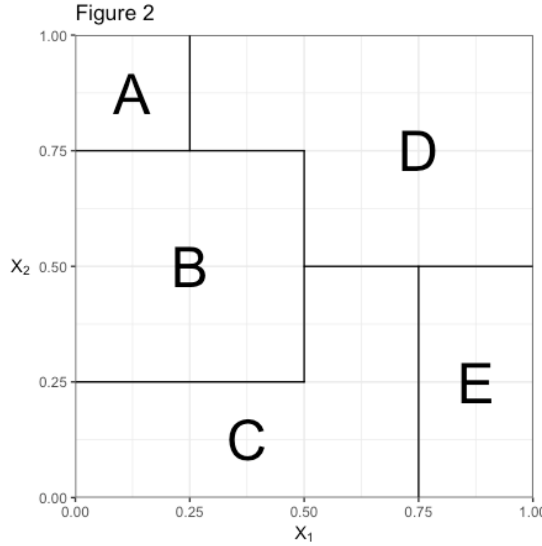
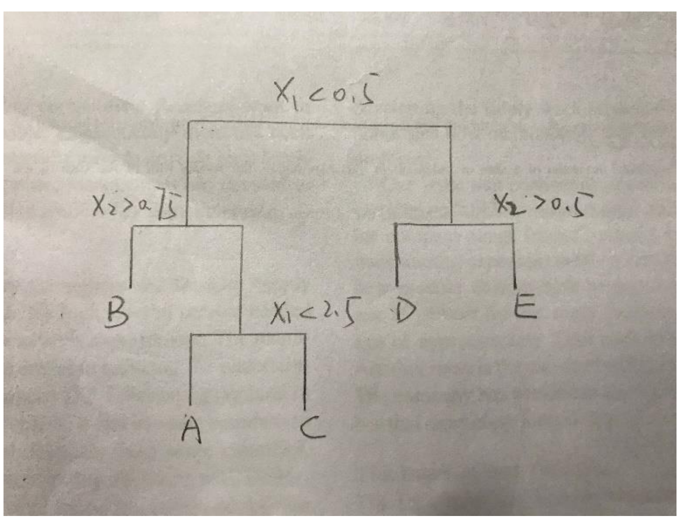


Figure 2



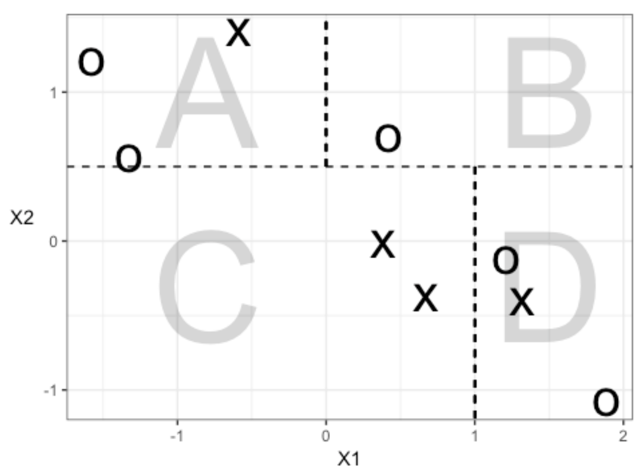
In figure 1, the partitions of the predictor space correspond to a tree. The partition of the decision tree in figure 1 is drawn as below, and for each level the left branch means "yes" and the right branch means "no".



1.3 In random forests, how do we ensure that the trees being averaged are sufficiently different from each other as to make the averaging effective?

We choose a random sample of m predictors as split candidates from the full set of p predictors. The split is forced to consider only one of those m predictors. By doing so, we are literally decorrelating the trees and thereby making the average of the resulting trees less variable.

1.4 Consider the following partition of the (two-predictor) predictor space. This time, the letters are there for you to refer to label the four regions. The response is categorical, and can either be "x" or "o".



(a) For this regression tree, what predictions would be made for each of the regions A through D? (b) What is the classification error for this decision?

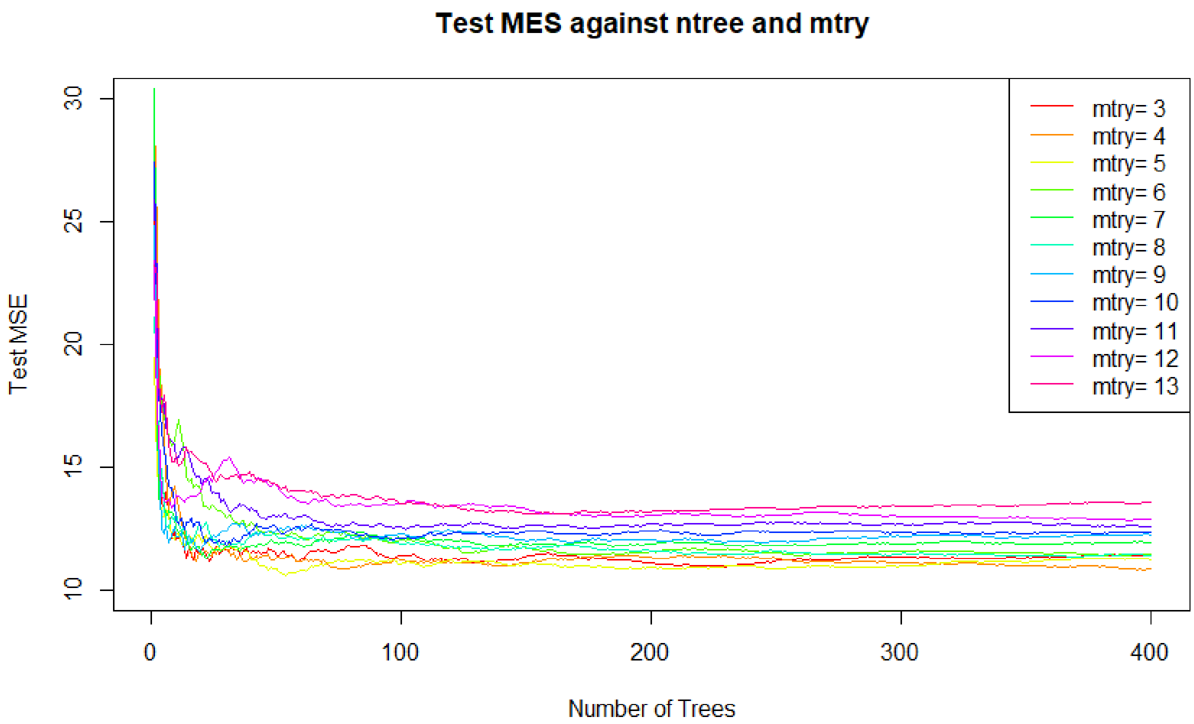
(a) According to the most commonly occurring class in each region, I predict O for region A, B and D, and X for region C.

(b) The classification error for this decision is 2/9.

1.5 Why might we not want a decision tree to split until the training data are perfectly classified? What aspect of a decision tree can we control to prevent creating a decision tree that large?

Because it will end up with overfitting problem, and not generalized enough for prediction for new data. We can set a bottom rule to stop the tree to split further, such as the minimum number of data fall in a class. As long as the number fall in a class reaches the bottom line, the tree cannot split further.

2. Describe the results obtained in Questions 7.



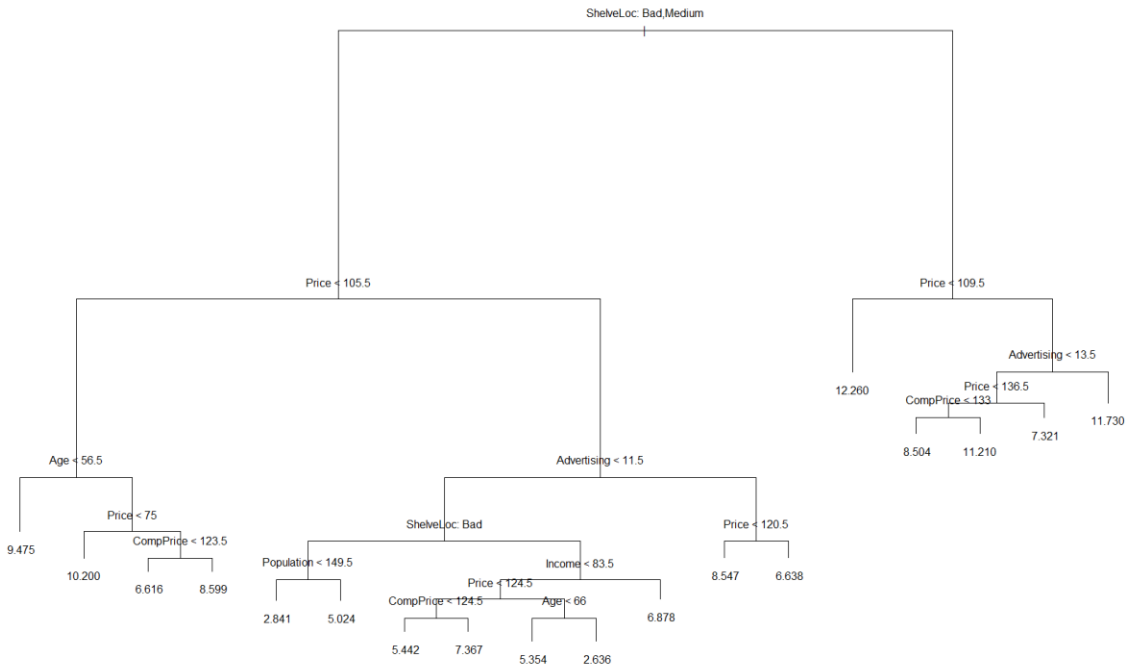
With the increase of number of trees, test set MES decreases at first and then stabilized after certain number of trees. Test set MES has the highest stabilized value when mtry equals to the total number of predictors available in the dataset. When mtry equals to 5, the test set MSE get its minimum value.

3. Question 8

(a) Split the data set into a training set and a test set.

I split 70% of the data set into the training set and set the rest to the test set.

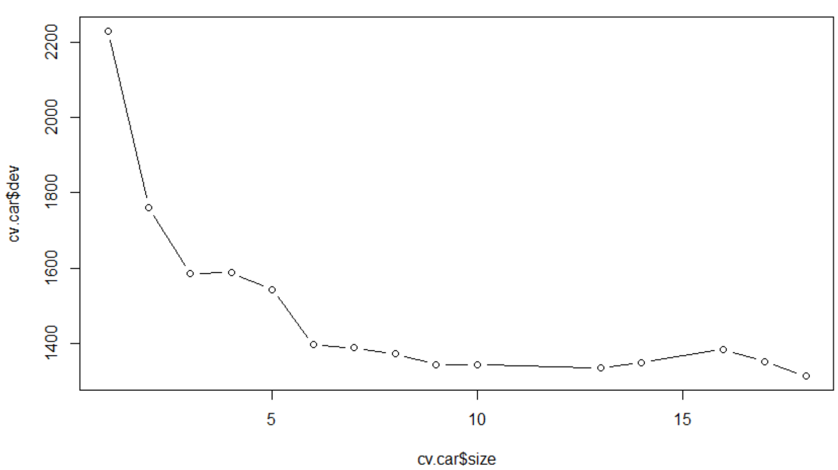
(b) Fit a regression tree to the training set. Plot the tree, and interpret the results. What test MSE do you obtain?



The tree shown above indicates that high quality shelve location and lower price correspond to higher sales. Bad and medium quality shelve location and lower price yields higher sales as well. For cars with bad quality shelve location, the sales are not good.

Test MSE is 5.288256.

(c) Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test MSE?



According to the plot above, I choose best size=9 as it has low deviance and relatively small tree size. Test MES after pruning the tree is 5.110397. Pruning decrease the test MSE from 5.288256 to 5.110397.

(d) Use the bagging approach in order to analyze this data. What test MSE do you obtain? Use the importance() function to determine which variables are most important.

```
> importance(car_bag)
              %IncMSE IncNodePurity
CompPrice   13.578245     192.17922
Income       7.163158     174.50515
Advertising 15.882820     216.15361
Population   2.864505     147.60184
Price       43.848974     532.21286
ShelveLoc   43.620496     453.25549
Age         14.160376     218.37852
Education    1.815317      89.24899
Urban       -2.406505      17.89143
US           3.945516      32.45247
```

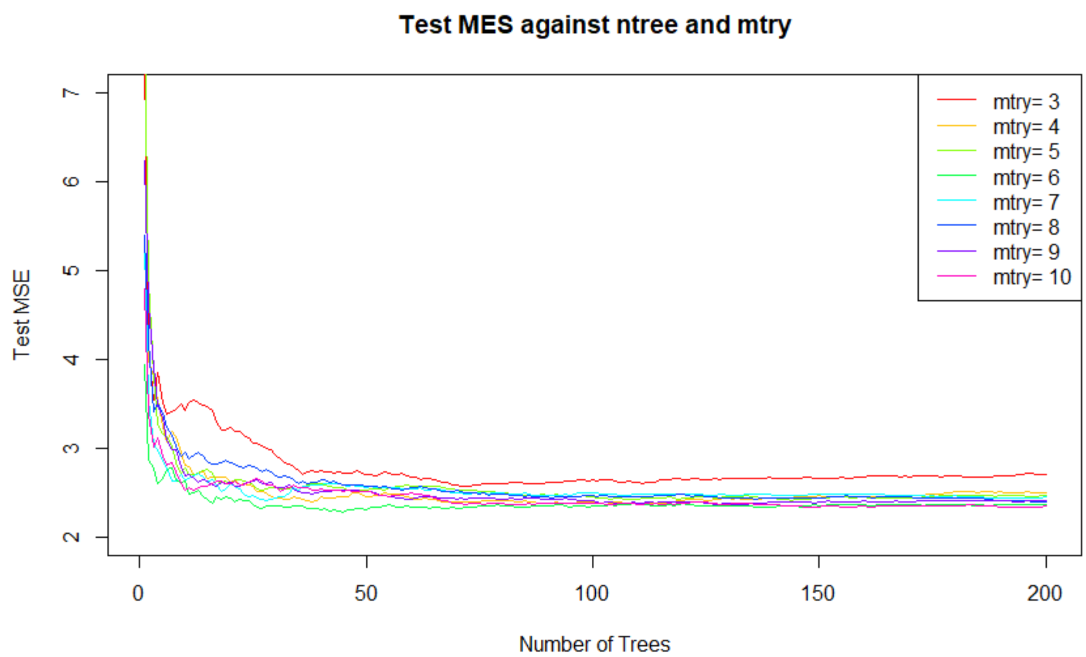
Test MSE is 2.74983 using bagging approach. Price and shelveloc are the most important variables as they have high %IncMSE and IncNodePurity.

(e) Use random forests to analyze this data. What test MSE do you obtain? Use the importance() function to determine which variables are most important. Describe the effect of m, the number of variables considered at each split, on the error rate obtained.

```
> importance(car_forest)
              %IncMSE IncNodePurity
CompPrice   21.003194     196.25087
Income       6.463616     130.94251
Advertising 18.795702     199.56999
Population   3.265927     115.52837
Price       60.636842     637.35623
ShelveLoc   61.583862     552.43691
Age         17.431946     206.59632
Education    3.183782      65.68864
Urban       -3.572691      11.04080
US           2.681270      19.62691
```

Test MSE is 2.384143 using random forest approach when mtry=6. Price and shelveloc are the most important variables as they have high %IncMSE and IncNodePurity.

Machine Learning Assignment 2
June(Liqiong) Sun# 75956961



From the plot above, it is obvious that when mtry equals to the 3, test MSE gets the maximum value. While when mtry equals to 6, test MSE gets the minimum value. When mtry use the other values in the legend, the value of stabilized test MSE fluctuated a little between 2 and 3.