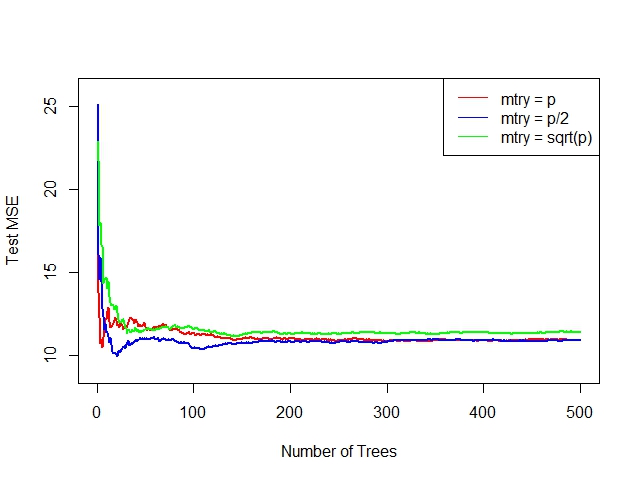
**HW 6 Chapter 8**

**Luyao Zhang (NetID: lzhang94)**

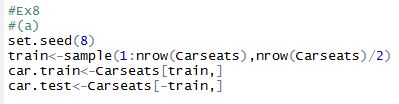
**Ex7**

The plot is displayed below. It seems that the test MSE drops dramatically when the number of trees goes up from 1. After 100 trees, the test MSE for all 3 situations becomes stable. Test MSE for the square root of the number of predictors is slightly higher than that of half or all of the predictors.



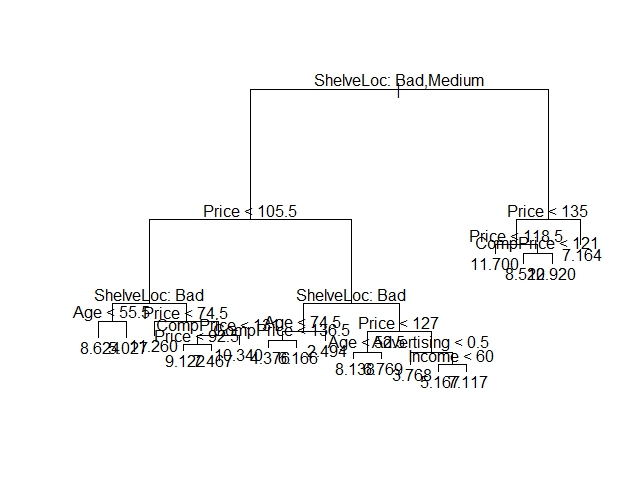
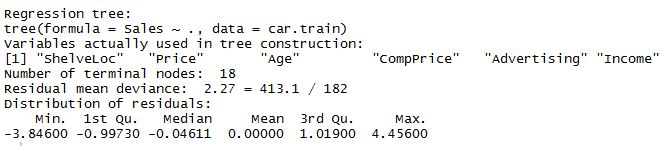
**Ex8**

1. The code for splitting the data into a training set and a test set is as below:

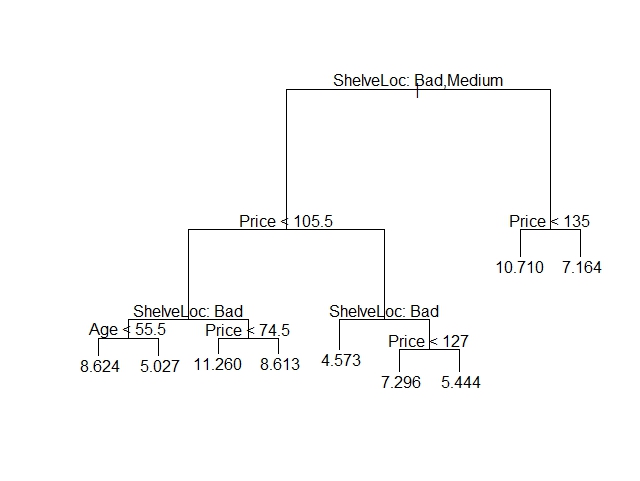


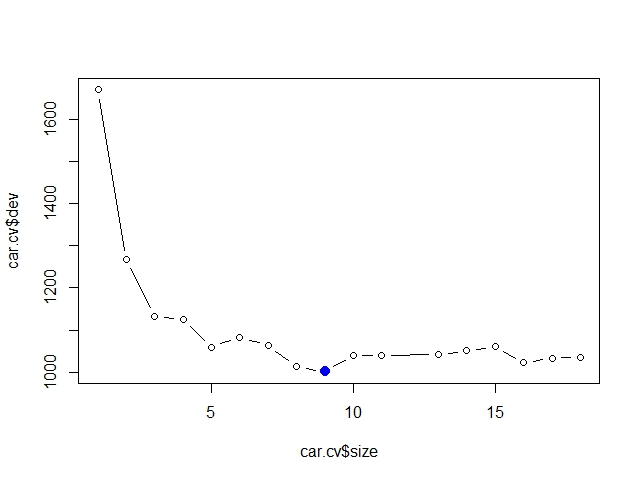
1. Below are the summary of the plot of the model. The test MSE is 4.50.

C:\Users\lzhang94\AppData\Local\Microsoft\Windows\INetCache\Content.Word\8_b_2_MSE.JPG

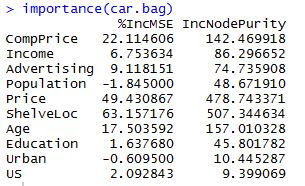


1. Based on the plot below, CV supported the tree of size 9. Using this result for pruning the tree, and plot can be found below, and the test MSE is 4.95, which is larger than that obtained in (b).

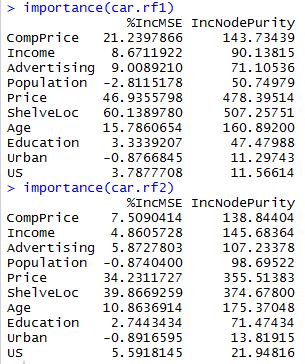
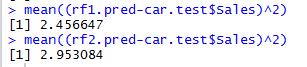




1. Using the bagging approach, we obtained test MSE equal to 2.41. According the importance analysis, “Price” and “ShelveLoc” are the most important variables.

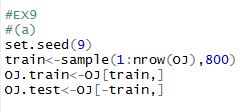


1. Using random forests, the test MSE is equal to 2.46 when m = p = 9, and the test MSE is equal to 2.95 when m = square root of p = 3. Also, both values of m indicate that “Price” and “ShelveLoc” are the most important variables.

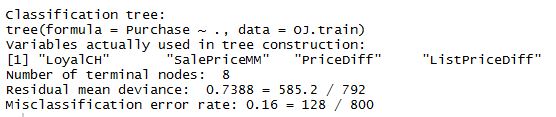


**Ex9**

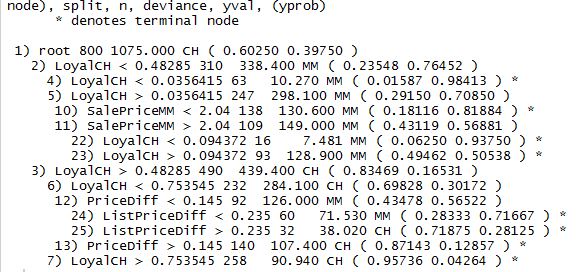
1. The code for splitting the data into a training and a test set is as below:



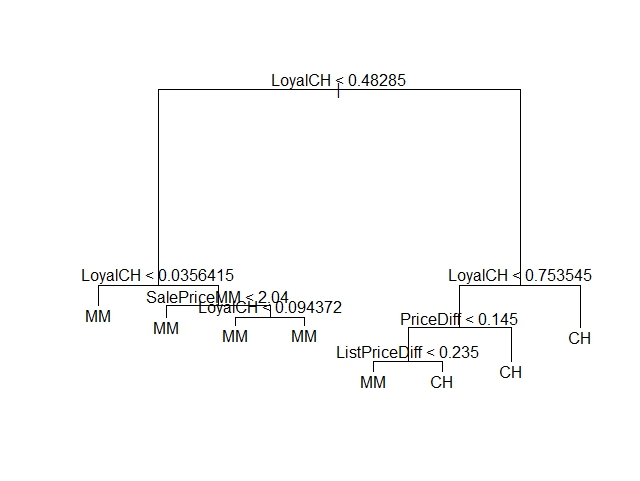
1. The summary of the model is as below. Apparently, the fitted tree has 8 terminal nodes, and the training error rate is equal to 0.16.



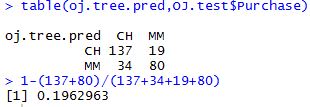
1. We picked #7, which is a terminal node. The split criterion is LoyalCh > 0.754, the sample size is 258, with the deviance equal to 90.94, and an overall prediction for the branch of CH. More than 95% of the points in this node have the value of CH, and the remaining 4% have the value of MM.



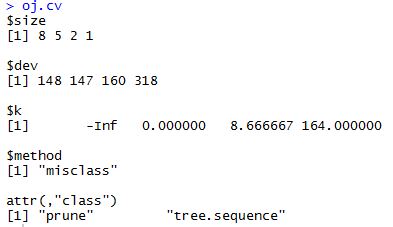
1. The plot of the tree is as below. According to the plot, it seems that “LoyalCH” is the most important predictor for “Purchase”, because the first branch of tree categorizes customers based on “LoyalCH”, and the top three nodes all have “LoyalCH”.



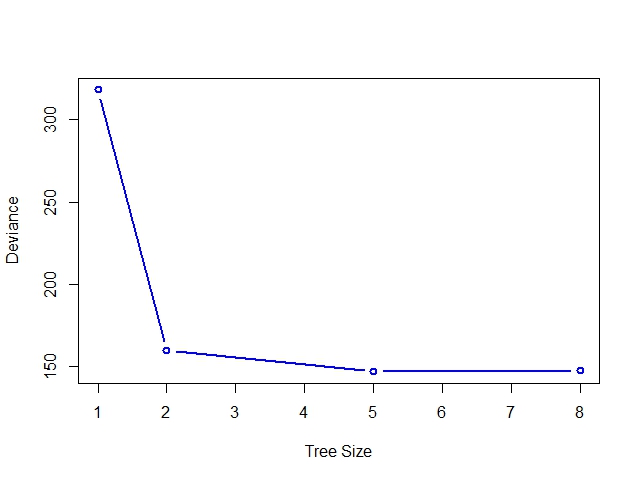
1. The table is as below. The test error rate is 0.20.



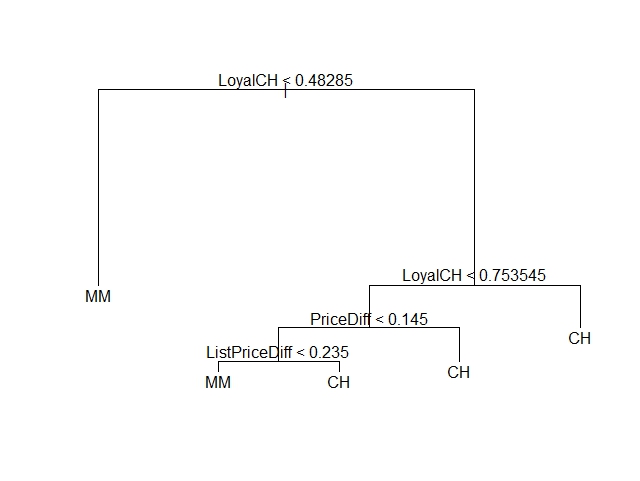
1. The optimal size of the tree is 5.



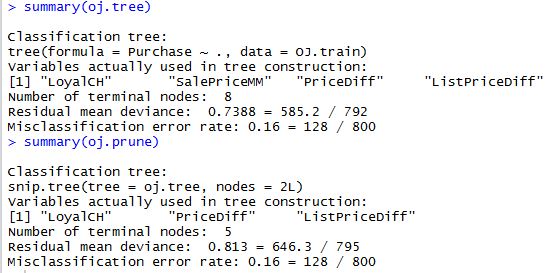
1. The plot is as below.



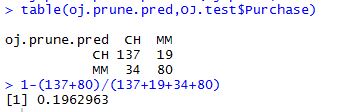
1. The optimal size of the tree should be 5.
2. The plot of the tree is as below.



1. According to the output below, the pruned tree and the original tree have the same misclassification error rate, which is 0.16.



1. The table and test error rate of the pruned tree is as below. The pruned tree has the same test error rate as the original tree, which is 0.20.

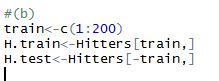


**Ex 10**

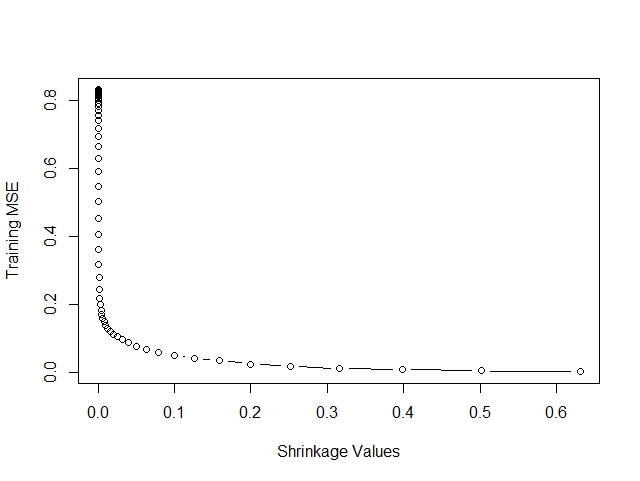
1. The code for removing some observations and log-transformation is as below:



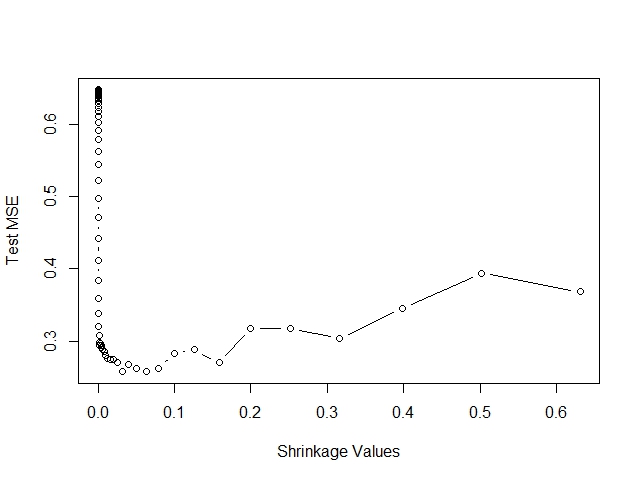
1. The code for creating a training set and a test set is as below:



1. The plot is as shown below:



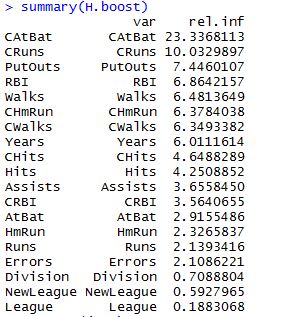
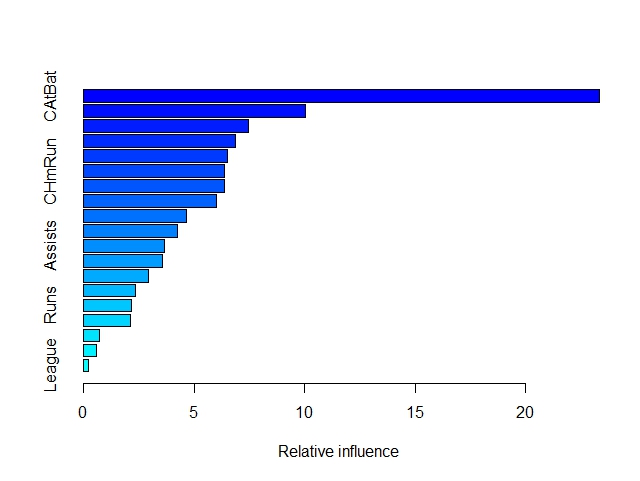
1. The plot is shown below. The minimum test MSE is 0.258, and the corresponding lambda is 0.063.



1. The test MSE resulting from linear regression is 0.492, and the test MSE resulting from the ridge regression is 0.457, and they are both larger than the test MSE obtained from boosting, which is equal to 0.258.



1. “CAtBat” seems to be the most important variable.

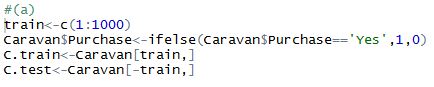


1. The test MSE obtained from using the bagging approach is 0.232, which is by far the smallest among all 4 approaches.

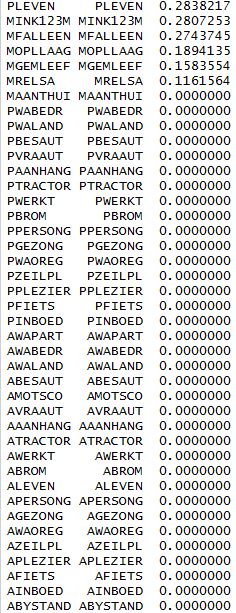
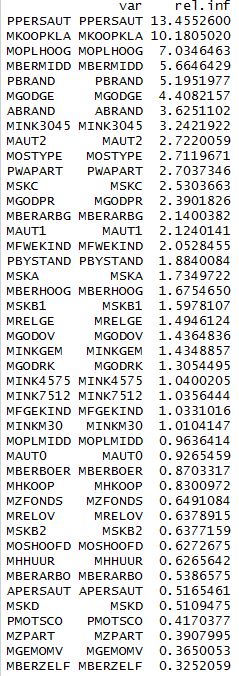
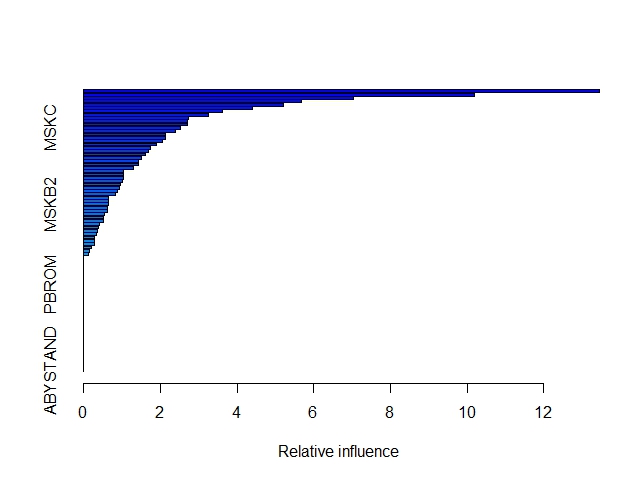


**Ex 11**

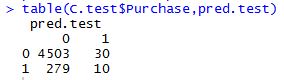
1. The code for creating the training set and the test set is as below:



1. According to the summary of the model shown below, “PPERSAUT” and “MKOOPKLA” are the most important variables.



1. The table resulting from the boosting approach is as below, and the fraction of people predicted to make a purchase do in fact make one is about 10 / (30+10) = 25%.



The table resulting from the logistic regression is shown below, and the fraction is 58 / (350+58) = 14.22%, which is smaller than that obtained by boosting.

