Assignment #6: Tree-Based Methods

SANDEEP DASARI **UIN- 829002252**

Problem 1

In the lab, a classification tree was applied to the Carseats data set after converting Sales into a binary response variable. This question will seek to predict Sales using regression trees and related approaches, treating the response as a quantitative variable (that is, without the conversion).

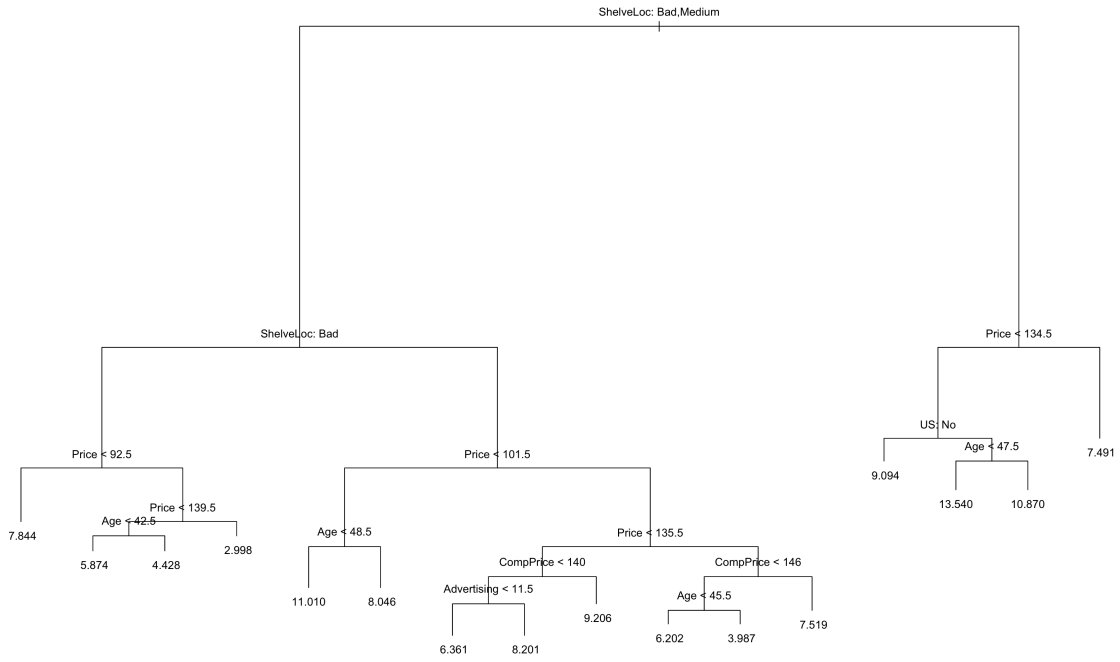
(a) Split the data set into a training set and a test set.

```
install.packages("tree")
install.packages("randomForest")
> library(ISLR)
>   library(tree)
>   library(randomForest)
randomForest 4.6-14
Type rfNews() to see new features/changes/bug fixes.
>   set.seed(3)
>   train = sample(1:nrow(Carseats), nrow(Carseats) / 2)
>   train_car = Carseats[train, ]
>   test_car = Carseats[-train,]
```

(b) Fit a regression tree to the training set. Plot the tree, and interpret the results. Then compute the test MSE.

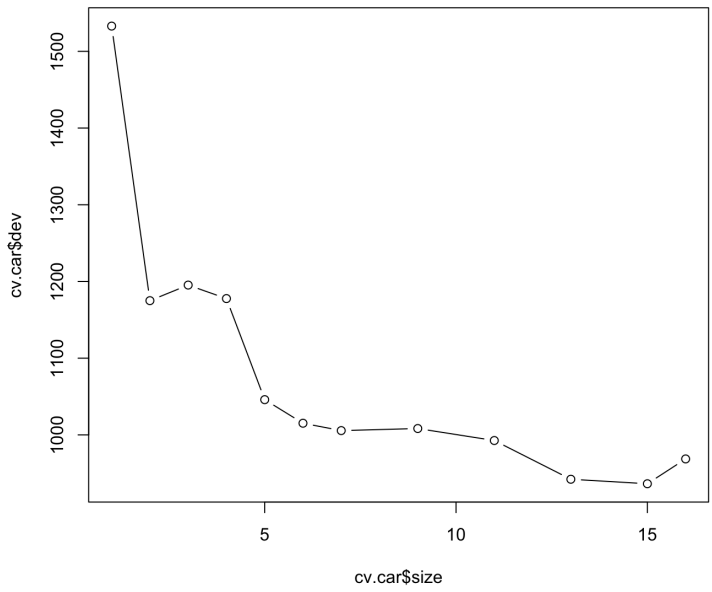
```
> reg_tree = tree(Sales~.,data = train_car)
>   summary(reg_tree)
Regression tree:
tree(formula = Sales ~ ., data = train_car)
Variables actually used in tree construction:
[1] "ShelveLoc"   "Price"       "Age"         "CompPrice"   "Advertising" "US"
Number of terminal nodes:  16
Residual mean deviance:  2.134 = 392.6 / 184
Distribution of residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-4.37400 -0.90790 -0.05181  0.00000  0.92840  3.82600
>   plot(reg_tree)
>   text(reg_tree,pretty=0)
>   y_pred = predict(reg_tree, newdata = test_car)
```

```
> mean((y_pred - test_car$Sales)^2)
[1] 4.784151
```

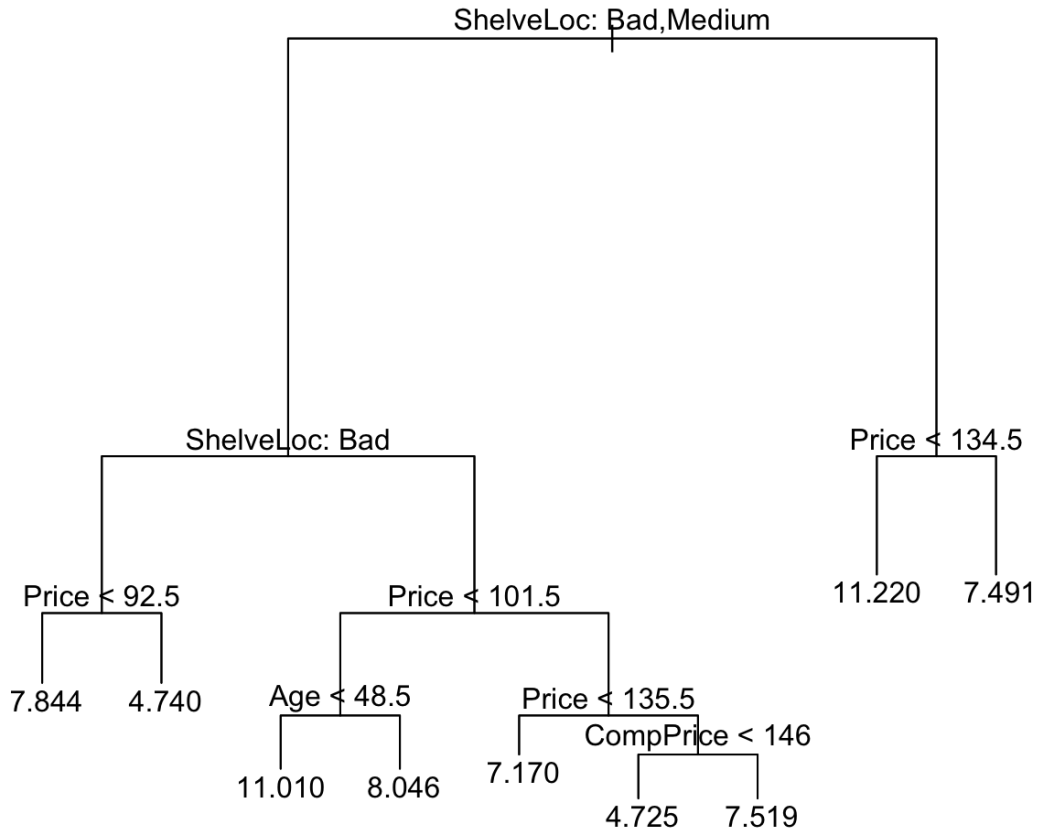


(c) Prune the tree obtained in (b). Use cross validation to determine the optimal level of tree complexity. Plot the pruned tree and interpret the results. Compute the test MSE of the pruned tree. Does pruning improve the test error?

```
> pruned = prune.tree(reg_tree, best = 8)
> set.seed(3)
> cv.car = cv.tree(reg_tree)
> plot(cv.car$size, cv.car$dev, type = "b")
```



```
>   plot(pruned)
>   text(pruned,pretty=0)
```




```
>   y_pred = predict(pruned, newdata= test_car)
>   mean((y_pred - test_car$Sales)^2)
[1] 5.075903
```

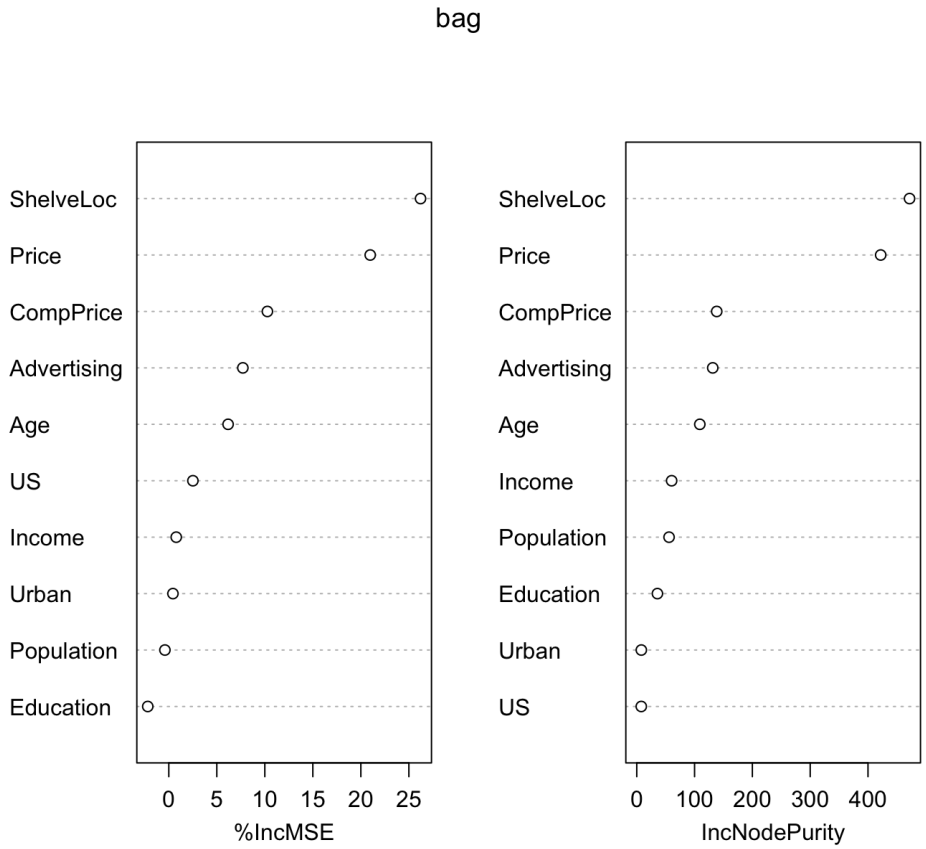
- No, pruning increases the MSE from 4.78 to 5.07. The tree complexity is 14.

(d) Use the bagging approach to analyze the data. What test MSE do you obtain? Determine which variables are most important.

```
> set.seed(1)
>   bag = randomForest(Sales~.,data = train_car, mtry = 10, ntree=100, importance = TRUE)
>   yhat_bag = predict(bag , newdata = test_car)
>   mean((yhat_bag - test_car$Sales)^2)
[1] 2.823931
```

> To determine the important variables

```
>   importance(bag)
              %IncMSE IncNodePurity
CompPrice   10.2664420    138.331489
Income       0.7746575     60.346494
Advertising  7.7110681    131.426841
Population  -0.4034994     55.866688
Price       20.9735768    421.939709
ShelveLoc   26.2150513    471.880483
Age          6.1745874    109.087540
Education   -2.1910308     35.733702
Urban        0.4313970      7.901116
US           2.5098922      7.802182
>   varImpPlot(bag)
```



- Important variables are ShelveLoc, Price, CompPrice, Age, Advertising, US in reducing importance.

(e) Use random forests to analyze the data. What test MSE do you obtain? Determine which variables are most important.

```
> set.seed(1)
>   random_forest = randomForest(Sales~.,data = train_car, mtry = 3, ntree = 100, importance = TRUE)

>   yhat_rf = predict(random_forest, newdata = test_car)
>   mean((yhat_rf - test_car$Sales)^2)
[1] 3.259455
```

- The MSE reduces from original tree but is not better than bagging (2.82).

```
>   importance(random_forest)
              %IncMSE IncNodePurity
CompPrice    4.6970336     131.48730
Income       2.8787020     111.53743
Advertising  5.8485187     158.56082
Population   0.5580396      86.96019
Price       13.0253273     340.86780
ShelveLoc   15.1947496     342.30045
Age          4.1282300     135.31990
Education   -1.9557642      56.71704
Urban       -1.2468731      14.78818
US           2.7246318      31.80294

>   varImpPlot(random_forest)
```

- Important variables are Price, ShelveLoc, CompPrice, Age, Advertising, US and Income.