Suraj Rimal

Lab Assignment: Chapter 8

8a.

library(ISLR)

set.seed(1)

train = sample(1:nrow(Carseats), nrow(Carseats) / 2)

Car.train = Carseats[train,]

Car.test = Carseats[-train,]

8b.

library (tree)

reg.tree = tree(Sales~.,data = Carseats, subset=train)

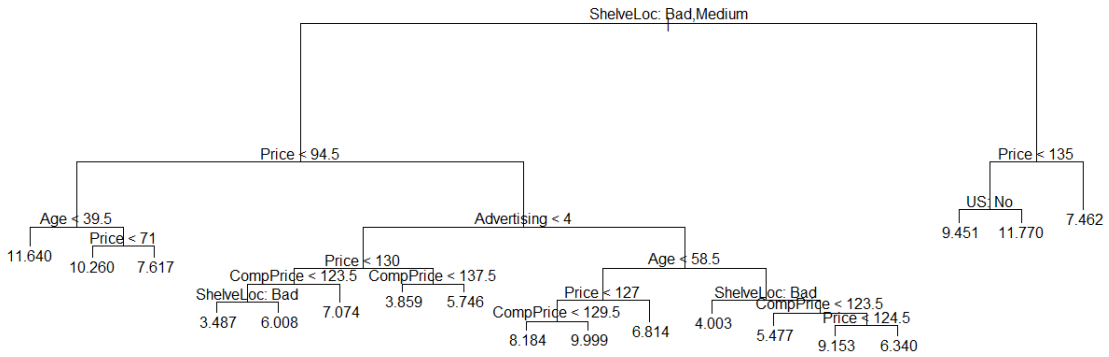
summary(reg.tree)

```
> reg.tree = tree(Sales~.,data = Carseats, subset=train)
> summary(reg.tree)

Regression tree:
tree(formula = Sales ~ ., data = Carseats, subset = train)
Variables actually used in tree construction:
[1] "ShelveLoc"   "Price"       "Age"         "Advertising" "CompPrice"   "US"
Number of terminal nodes:  18
Residual mean deviance:  2.167 = 394.3 / 182
Distribution of residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-3.88200 -0.88200 -0.08712  0.00000  0.89590  4.09900
```

plot(reg.tree)

text(reg.tree ,pretty =0)



yh = predict(reg.tree,newdata = Car.test)

mean((yh - Car.test$Sales)^2)

```
> mean((yh - Car.test$Sales)^2)
[1] 4.922039
```

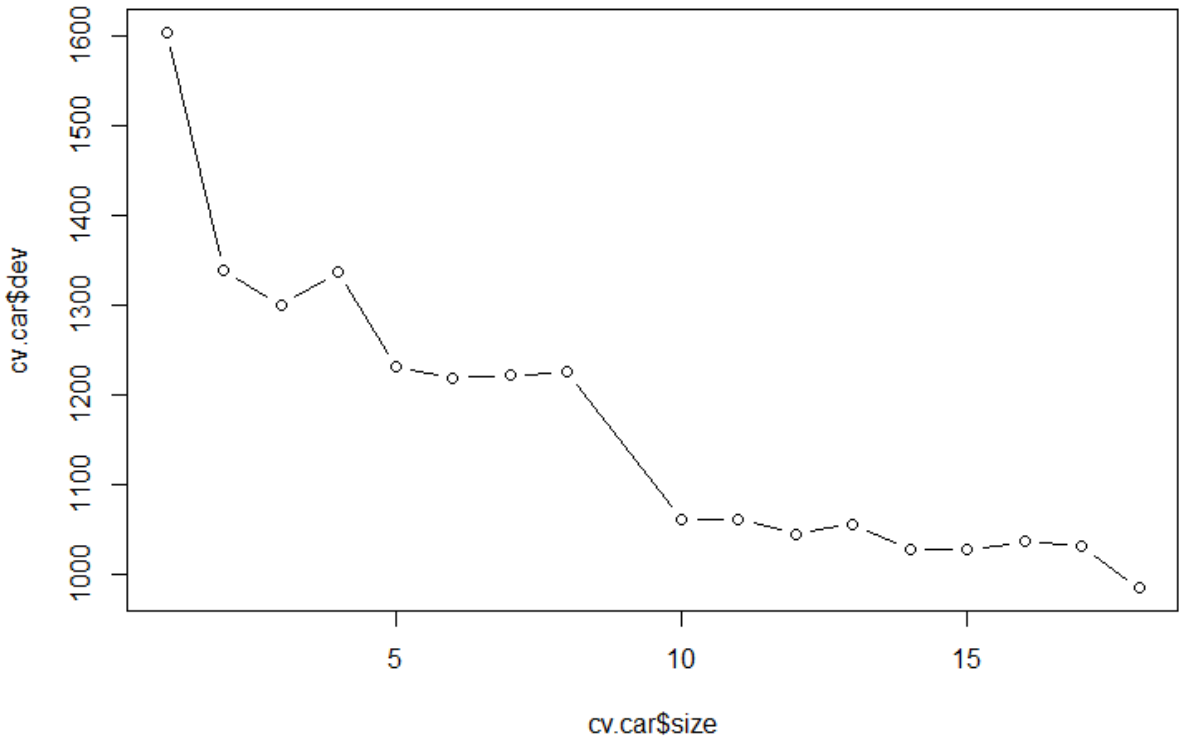
The test MSE is about 4.92.

8c.

set.seed(1)

cv.car = cv.tree(reg.tree)

plot(cv.car$size, cv.car$dev, type = "b")

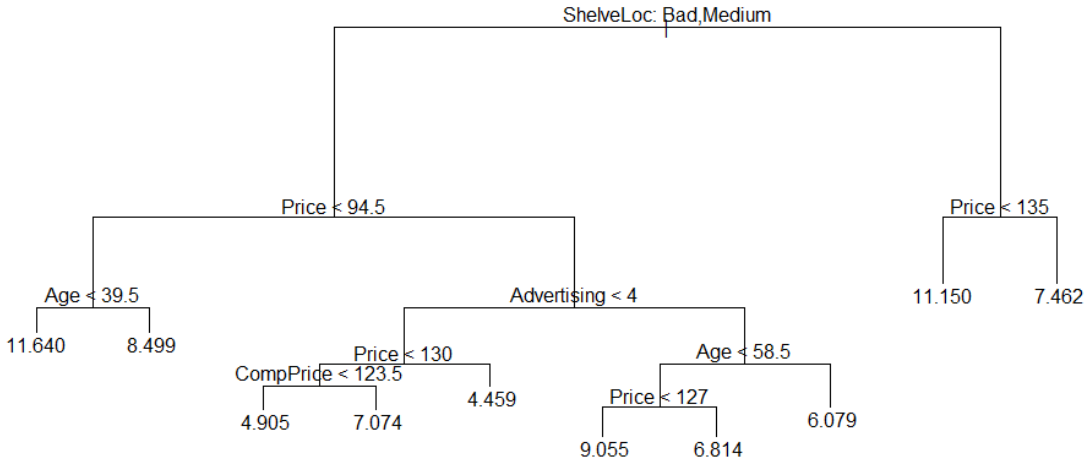


We now prune the tree to obtain the 10-node tree.

prune.car = prune.tree(reg.tree, best = 10)

plot(prune.car)

text(prune.car,pretty=0)



yh=predict(prune.car, newdata= Car.test)

mean((yh-Car.test$Sales)^2)

```
> yh=predict(prune.car, newdata= Car.test)
> mean((yh-Car.test$Sales)^2)
[1] 4.918134
```

We see that pruning the tree slightly decreases the Test MSE to 4.918.

8d.

library(randomForest)

set.seed(1)

bag.car = randomForest(Sales~.,data=Car.train,mtry = 10, importance = TRUE)

yh.bag = predict(bag.car,newdata=Car.test)

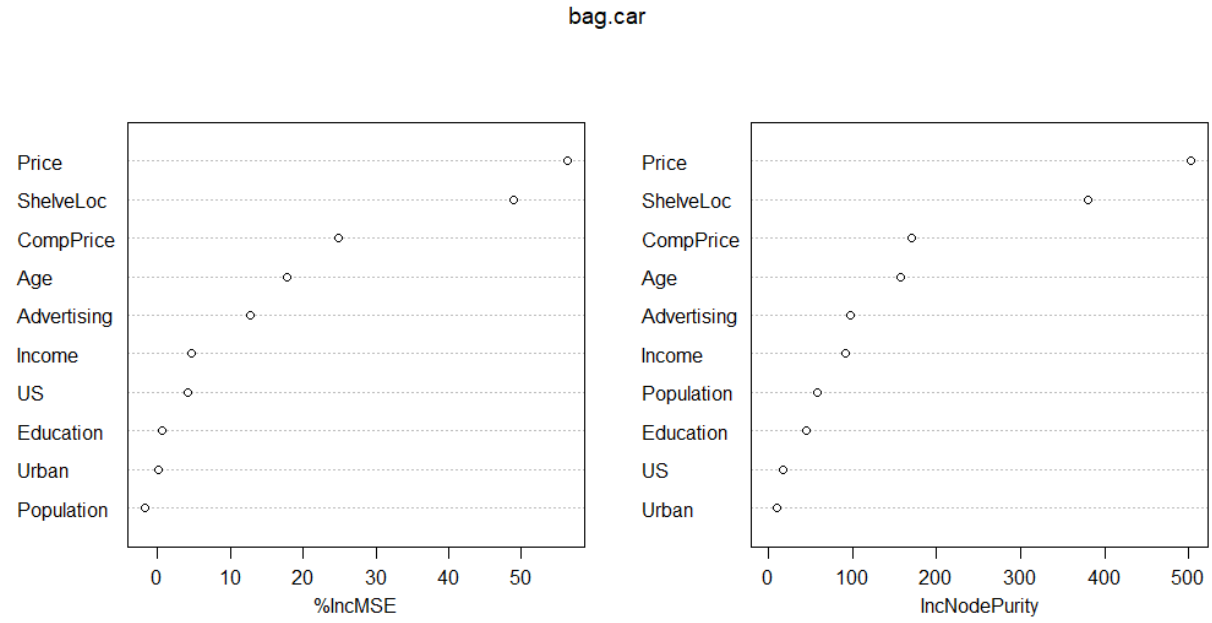
mean((yh.bag-Car.test$Sales)^2)

```
> mean((yh.bag-Car.test$Sales)^2)
[1] 2.605253
```

importance(bag.car)

```
> importance(bag.car)
               %IncMSE IncNodePurity
CompPrice   24.8888481    170.182937
Income       4.7121131     91.264880
Advertising 12.7692401     97.164338
Population  -1.8074075     58.244596
Price       56.3326252    502.903407
ShelveLoc   48.8886689    380.032715
Age         17.7275460    157.846774
Education    0.5962186     44.598731
Urban        0.1728373      9.822082
US           4.2172102     18.073863
> |
```

varImpPlot(bag.car)



The price that the company charges for car seats at each location, as well as the quality of the shelving location for the car seats at each location, are the most significant variables. The bagging approach regression tree's test MSE is 2.60, which is very less than that of a single tree that has been pruned optimally.

8e.

set.seed(1)

rf.car = randomForest(Sales~.,data=Car.train,mtry = 3, importance = TRUE)

yh.rf = predict(rf.car,newdata=Car.test)

mean((yh.rf-Car.test$Sales)^2)

```
> yh.rf = predict(rf.car,newdata=Car.test)
> mean((yh.rf-Car.test$Sales)^2)
[1] 2.960559
```

The MSE of the test set is 2.96, when m = 3.

set.seed(1)

rf.car = randomForest(Sales~.,data=Car.train,mtry = 5, importance = TRUE)

yh.rf = predict(rf.car,newdata=Car.test)

mean((yh.rf-Car.test$Sales)^2)

```
> mean((yh.rf-Car.test$Sales)^2)
[1] 2.714168
```

slightly less when m = 5

set.seed(1)

rf.car = randomForest(Sales~.,data=Car.train,mtry = 7, importance = TRUE)

yh.rf = predict(rf.car,newdata=Car.test)

mean((yh.rf-Car.test$Sales)^2)

```
> mean((yh.rf-Car.test$Sales)^2)
[1] 2.678559
```

set.seed(1)

rf.car = randomForest(Sales~.,data=Car.train,mtry = 9, importance = TRUE)

yh.rf = predict(rf.car,newdata=Car.test)

mean((yh.rf-Car.test$Sales)^2)

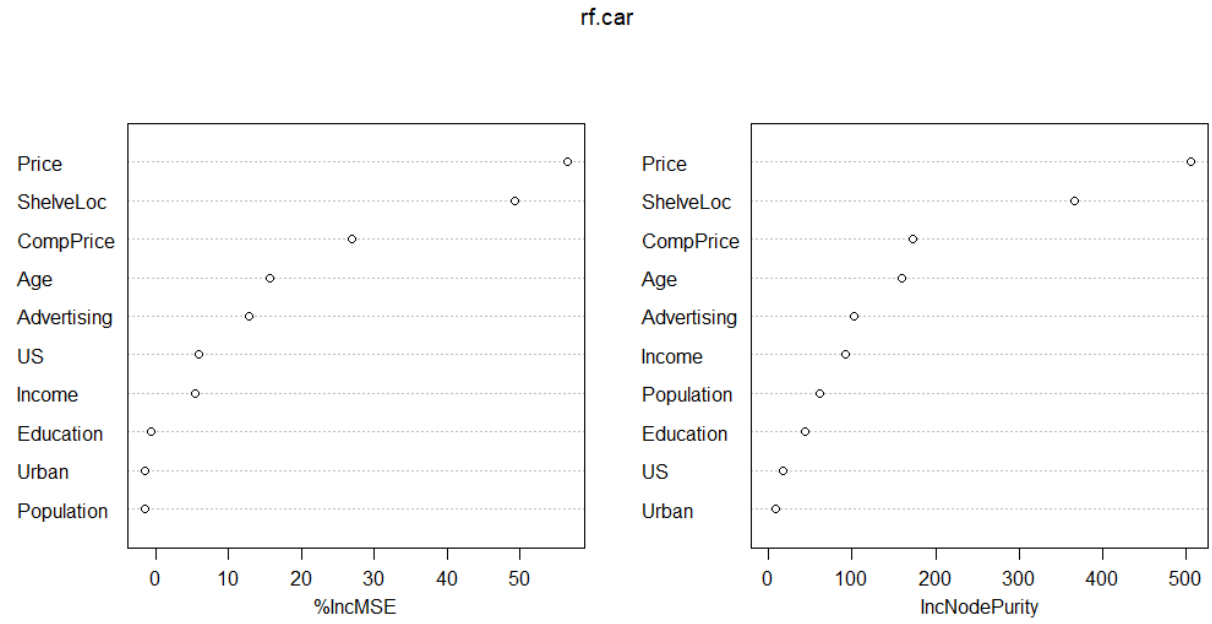
```
> mean((yh.rf-Car.test$Sales)^2)
[1] 2.590855
```

In this particular case, as m grows closer to 10 (the total number of predictor variables), the MSE decreases.

importance(rf.car)

```
> importance(rf.car)
               %IncMSE IncNodePurity
CompPrice   27.0180857     172.83514
Income       5.4318733      92.50209
Advertising 12.7898803     102.32473
Population  -1.6067153      61.43735
Price       56.5767031     506.02790
ShelveLoc   49.2910725     366.85186
Age         15.7203983     159.45251
Education   -0.7086443      43.74559
Urban       -1.5245078       8.46724
US           5.8813465      18.04286
```

varImpPlot(rf.car)



Here again, the price that the company charges for car seats at each location, as well as the quality of the shelving location for the car seats at each location, are the most significant variables.