# vii.Exercise 8.4.8

Vidhi Shah, Sahil Shah, Kaarthik Sundaramoorthy

7/11/2020

In the lab, a classification tree was applied to the **Carseats** data set after converting **Sales** into a qualitative response variable. Now we will seek to predict **Sales** using regression trees and related approaches, treating the response as a quantitative variable.

```
library(ISLR)
library(tree)
library(randomForest)
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
attach(Carseats)
```

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

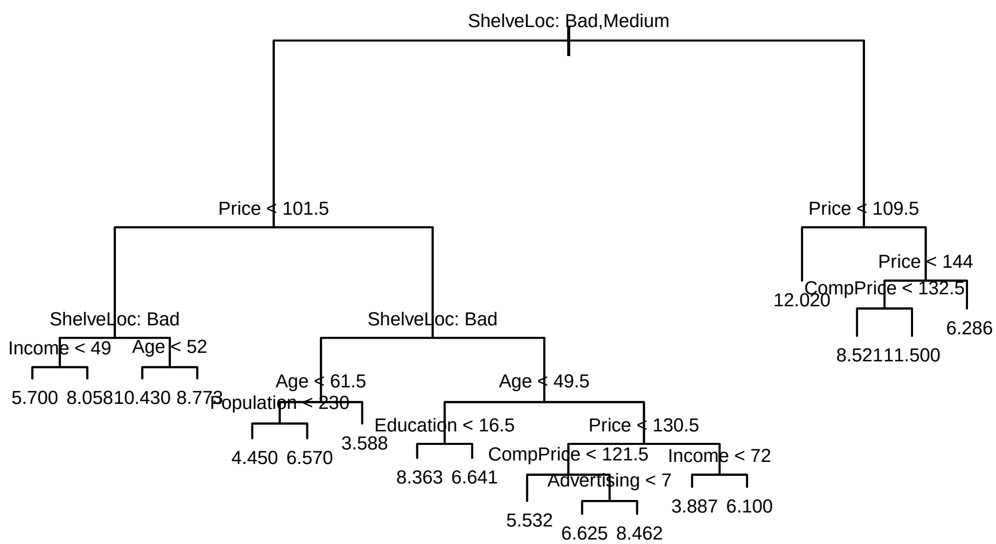
```
set.seed(123)
train <- sample(1:nrow(Carseats), nrow(Carseats) / 2)
Carseats.train <- Carseats[train, ]
Carseats.test <- Carseats[-train, ]
```

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

```
tree.carseats <- tree(Sales ~ ., data = Carseats, subset = train )
summary(tree.carseats)
```

```
##
## Regression tree:
## tree(formula = Sales ~ ., data = Carseats, subset = train)
## Variables actually used in tree construction:
## [1] "ShelveLoc"   "Price"       "Income"      "Age"         "Population"
## [6] "Education"   "CompPrice"   "Advertising"
## Number of terminal nodes:  18
## Residual mean deviance:  2.132 = 388.1 / 182
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -4.08000 -0.92870  0.06244  0.00000  0.87020  3.71700
```

```r
plot(tree.carseats)
text(tree.carseats,pretty=0,cex=0.6)
```




```r
tree.pred=predict(tree.carseats, newdata = Carseats.test)
mean((tree.pred-Carseats.test$Sales)^2)
```

```
## [1] 4.395357
```

The tree defines that the attribute shelf location is the primary factor and which results in lower sales. Following to that the price is the next important factor involving the split with both branches below the root. The test MSE here is about 4.3953574.

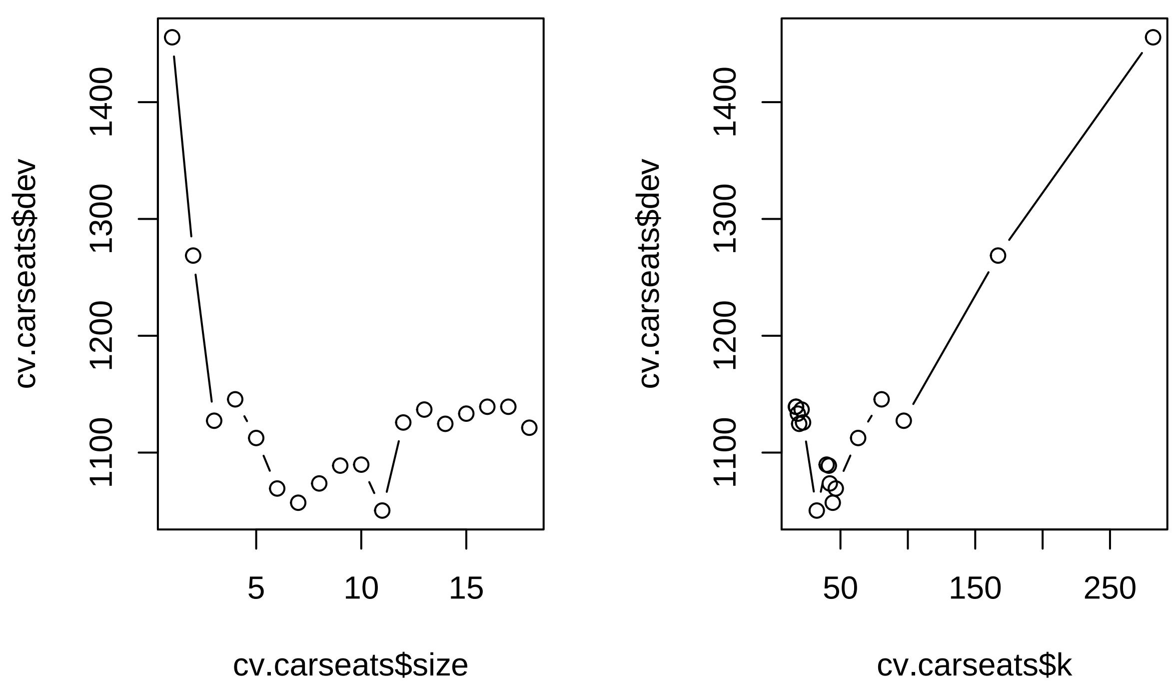
(c) Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test MSE?

```r
set.seed(2)
cv.carseats <- cv.tree(tree.carseats)
cv.carseats
```

```
## $size
##  [1] 18 17 16 15 14 13 12 11 10  9  8  7  6  5  4  3  2  1
##
## $dev
```

```
##  [1] 1121.271 1139.239 1139.231 1133.335 1124.619 1136.810 1125.763 1050.377
##  [9] 1089.685 1088.844 1073.570 1057.068 1069.233 1112.483 1145.583 1127.277
## [17] 1268.653 1455.501
##
## $k
##  [1]      -Inf  16.93456  17.10094  18.36513  19.36805  21.17568  22.24728
##  [8]  32.41359  39.73324  41.38229  42.06332  44.28093  46.53155  63.12600
## [15]  80.51350  96.97503 166.93502 281.96185
##
## $method
## [1] "deviance"
##
## attr(,"class")
## [1] "prune"         "tree.sequence"
```

```r
par(mfrow = c(1, 2))
plot(cv.carseats$size, cv.carseats$dev, type = "b")
plot(cv.carseats$k, cv.carseats$dev, type = "b")
```

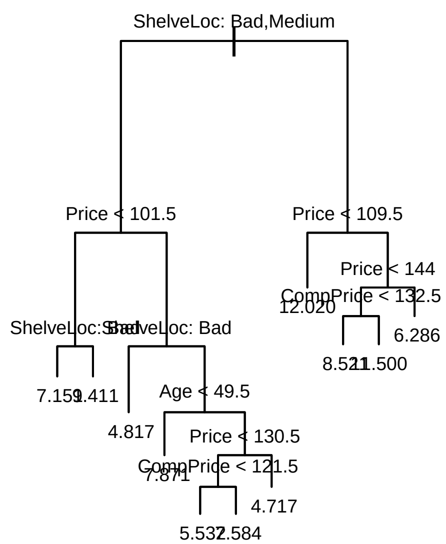



```r
best.size <- cv.carseats$size[which.min(cv.carseats$dev)]
best.size
```

```
## [1] 11
```

```r
prune.carseats <- prune.tree(tree.carseats, best = best.size)
plot(prune.carseats)
text(prune.carseats, pretty = 0,cex=0.6)
yhat <- predict(prune.carseats, newdata = Carseats.test)
mean((yhat - Carseats.test$Sales)^2)
```

```
## [1] 4.646409
```




The best size here is 11 . Pruning the tree in this case scenario increases the test MSE to 4.6464094.

(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.

```r
set.seed(1)
bag.carseats = randomForest(Sales~.,data=Carseats.train,mtry = 10, importance = TRUE)
yhat.bag <- predict(bag.carseats, newdata = Carseats.test)
mean((yhat.bag - Carseats.test$Sales)^2)
```

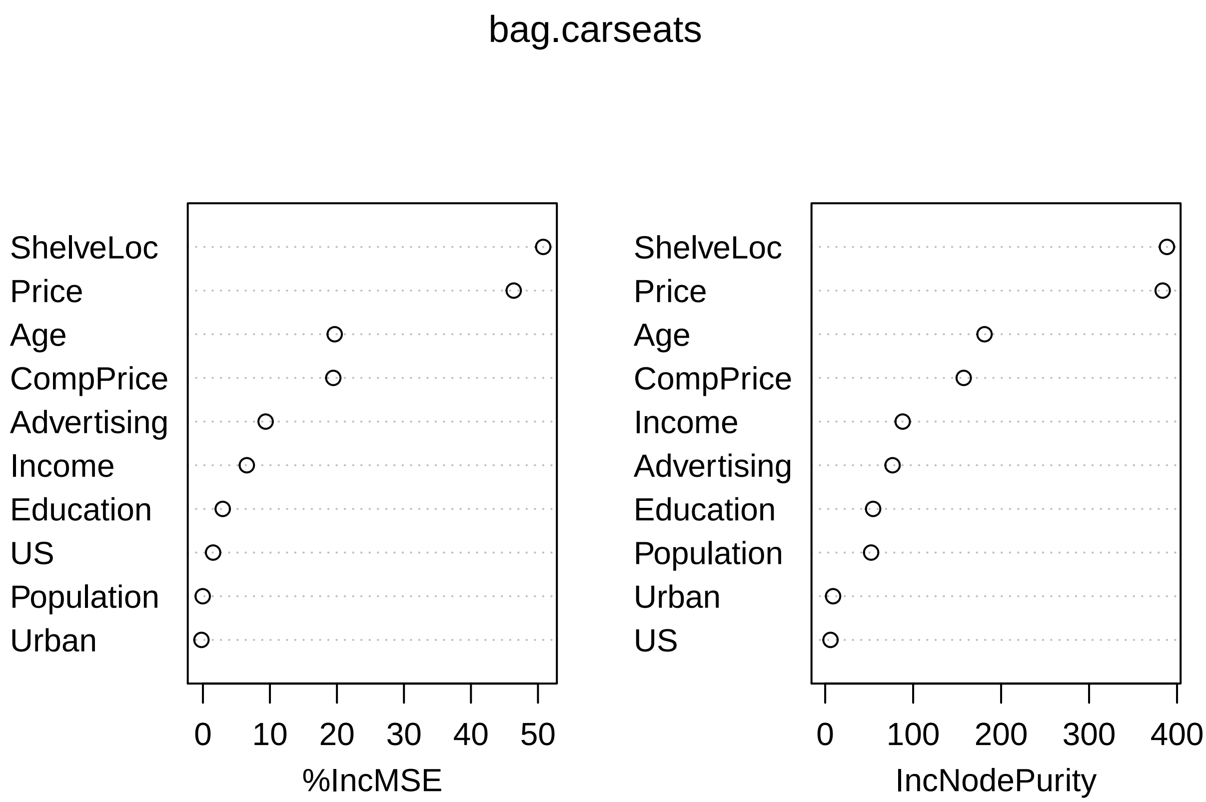
```
## [1] 2.719623
```

```r
importance(bag.carseats)
```

```
##                  %IncMSE IncNodePurity
```

```
## CompPrice   19.45546048    157.493242
## Income       6.54808190     88.084674
## Advertising  9.35661044     76.584090
## Population  -0.03332117     52.232916
## Price       46.38347866    383.642775
## ShelveLoc   50.78194122    388.510326
## Age         19.66414864    181.155199
## Education    2.95721774     54.469563
## Urban       -0.22123191      8.959947
## US           1.51594678      6.052909
```

```r
varImpPlot(bag.carseats)
```




The bagging method improves the MSE to 2.7196233. From the graph we can also see that the **Price**, **ShelveLoc**, and **Age** are the important predictors with respect to the **Sale**.

(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.

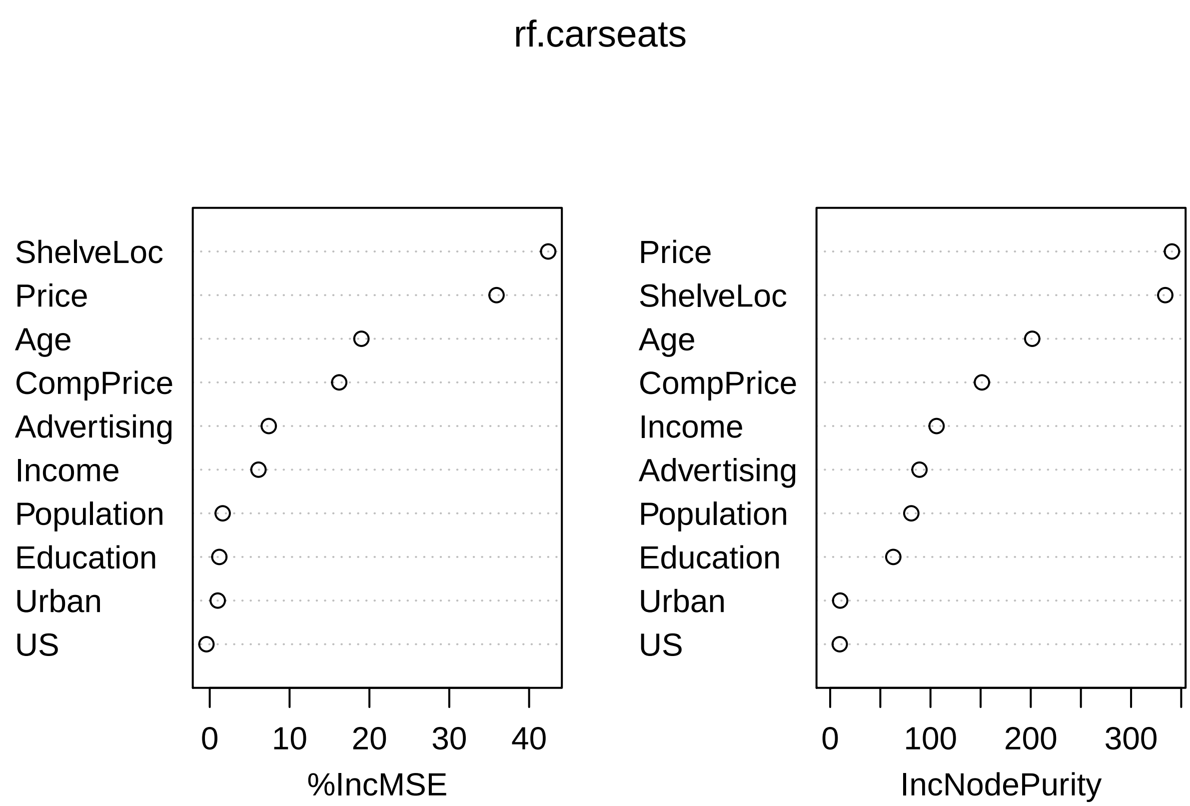
```r
rf.carseats <- randomForest(Sales ~ ., data = Carseats.train, mtry = 5, importance = TRUE)
yhat.rf <- predict(rf.carseats, newdata = Carseats.test)
mean((yhat.rf - Carseats.test$Sales)^2)
```

```
## [1] 3.029787
```

```r
importance(rf.carseats)
```

```
##                %IncMSE IncNodePurity
## CompPrice   16.2153774    151.303116
## Income       6.1058841    106.073448
## Advertising  7.3929755     89.039402
## Population   1.6217460     80.898770
## Price       35.9193599    340.758704
## ShelveLoc   42.3911696    334.120925
## Age         18.9986143    201.454684
## Education    1.2019161     62.932906
## Urban        1.0025134      9.970047
## US          -0.4123717      9.542462
```

```r
varImpPlot(rf.carseats)
```




The random forest increases the MSE to 3.0297873. The change in the $m$ of the test MSE is seen between **2.7** to **3.2**. From the above plot we see that the variable importance factors for prediction are same as in the bagging approach i.e **Price**, **ShelveLoc**, and **Age**.