Chapter 8 Tree Based Methods - Problems 8

kartheek raj mulasa

1/16/2020

8 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.

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

Data pulling from ISLR library and data snapshot.

```
require(MASS)
```

```
## Loading required package: MASS
```

```
require(tree)
```

```
## Loading required package: tree
```

```
## Warning: package 'tree' was built under R version 3.6.2
```

```
require(ISLR)
```

```
## Loading required package: ISLR
```

```
require(randomForest)
```

```
## Loading required package: randomForest
```

```
## Warning: package 'randomForest' was built under R version 3.6.2
```

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

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

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


carseats<-data.frame(Carseats)
head(carseats)
```

```
##   Sales CompPrice Income Advertising Population Price ShelveLoc Age Education
## 1  9.50       138     73          11        276   120       Bad  42        17
## 2 11.22       111     48          16        260    83      Good  65        10
## 3 10.06       113     35          10        269    80    Medium  59        12
## 4  7.40       117    100           4        466    97    Medium  55        14
## 5  4.15       141     64           3        340   128       Bad  38        13
## 6 10.81       124    113          13        501    72       Bad  78        16
##   Urban  US
## 1   Yes Yes
## 2   Yes Yes
## 3   Yes Yes
```

```
## 4   Yes Yes
## 5   Yes  No
## 6    No Yes
```

```
str(carseats)
```

```
## 'data.frame':    400 obs. of  11 variables:
##  $ Sales      : num  9.5 11.22 10.06 7.4 4.15 ...
##  $ CompPrice  : num  138 111 113 117 141 124 115 136 132 132 ...
##  $ Income     : num  73 48 35 100 64 113 105 81 110 113 ...
##  $ Advertising: num  11 16 10 4 3 13 0 15 0 0 ...
##  $ Population : num  276 260 269 466 340 501 45 425 108 131 ...
##  $ Price      : num  120 83 80 97 128 72 108 120 124 124 ...
##  $ ShelveLoc  : Factor w/ 3 levels "Bad","Good","Medium": 1 2 3 3 1 1 3 2 3 3 ...
##  $ Age        : num  42 65 59 55 38 78 71 67 76 76 ...
##  $ Education  : num  17 10 12 14 13 16 15 10 10 17 ...
##  $ Urban      : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 2 2 1 1 ...
##  $ US         : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 2 1 2 1 2 ...
```

```
summary(carseats)
```

```
##      Sales          CompPrice       Income        Advertising
##  Min.   : 0.000   Min.   : 77   Min.   : 21.00   Min.   : 0.000
##  1st Qu.: 5.390   1st Qu.:115   1st Qu.: 42.75   1st Qu.: 0.000
##  Median : 7.490   Median :125   Median : 69.00   Median : 5.000
##  Mean   : 7.496   Mean   :125   Mean   : 68.66   Mean   : 6.635
##  3rd Qu.: 9.320   3rd Qu.:135   3rd Qu.: 91.00   3rd Qu.:12.000
##  Max.   :16.270   Max.   :175   Max.   :120.00   Max.   :29.000
##    Population        Price        ShelveLoc        Age          Education
##  Min.   : 10.0   Min.   : 24.0   Bad   : 96   Min.   :25.00   Min.   :10.0
##  1st Qu.:139.0   1st Qu.:100.0   Good  : 85   1st Qu.:39.75   1st Qu.:12.0
##  Median :272.0   Median :117.0   Medium:219   Median :54.50   Median :14.0
##  Mean   :264.8   Mean   :115.8                Mean   :53.32   Mean   :13.9
##  3rd Qu.:398.5   3rd Qu.:131.0                3rd Qu.:66.00   3rd Qu.:16.0
##  Max.   :509.0   Max.   :191.0                Max.   :80.00   Max.   :18.0
##  Urban       US
##  No :118   No :142
##  Yes:282   Yes:258
##
##
##
##
```

Test and train data splits

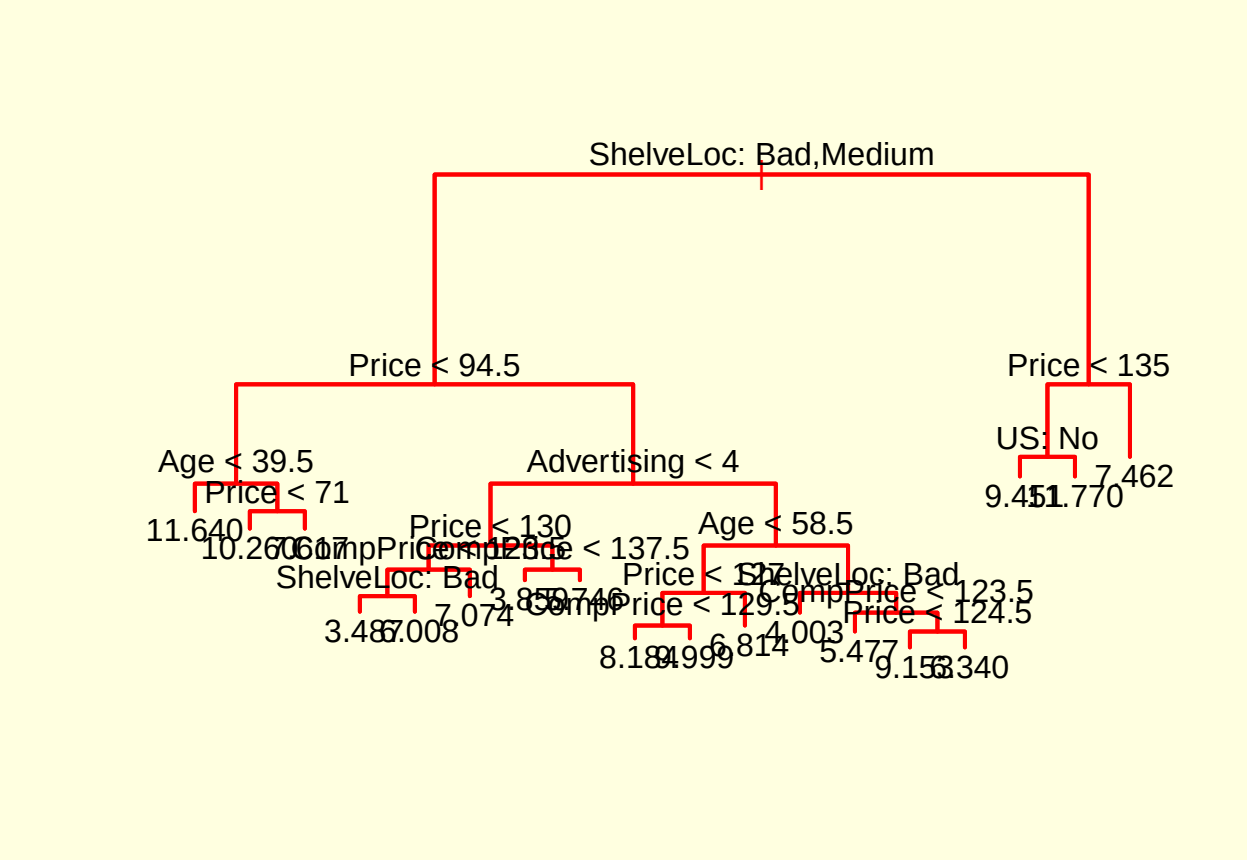
```
set.seed(1)
sd<-sample(1:nrow(carseats), round(nrow(carseats)/2))
train<-carseats[sd,]
test<-carseats[-sd,]
```

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

Answer

Building a random forest regression model to train.

```
ctree<-tree(Sales~.,data = train)
par(mfrow=c(1,1),bg="lightyellow")
plot(ctree,col="red",lwd=2)
text(ctree, pretty=0)
```




```
summary(ctree)
```

```
##
## Regression tree:
## tree(formula = Sales ~ ., data = train)
## Variables actually used in tree construction:
## [1] "ShelveLoc"   "Price"       "Age"         "Advertising" "CompPrice"
## [6] "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
```

Predicting the results and MSE on test data.

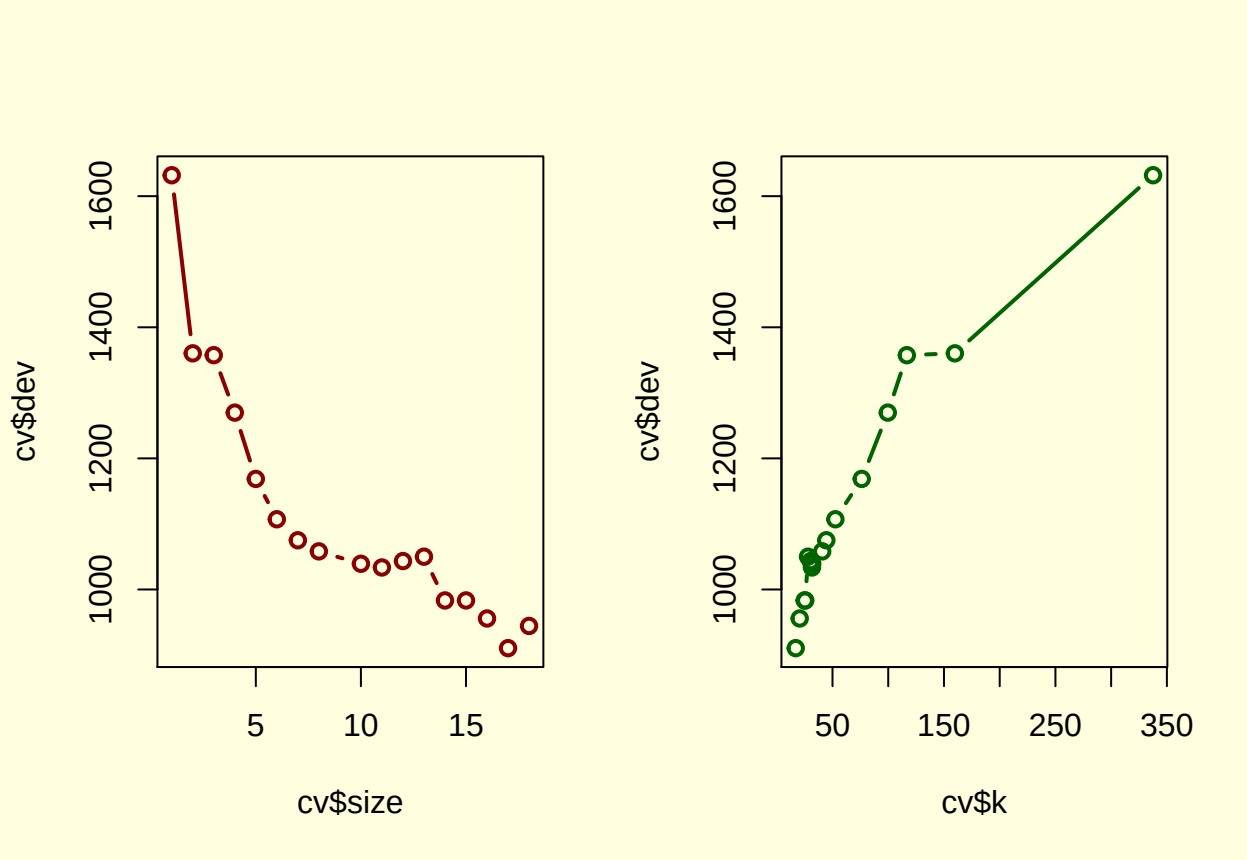
```
 TESTMSE<-function(x) {
  h=x
  predictvalues<-predict(h,newdata=test)
  mean((predictvalues-test$Sales)^2)
}
TESTMSE(ctree)
```

```
## [1] 4.922039
```

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

Answer

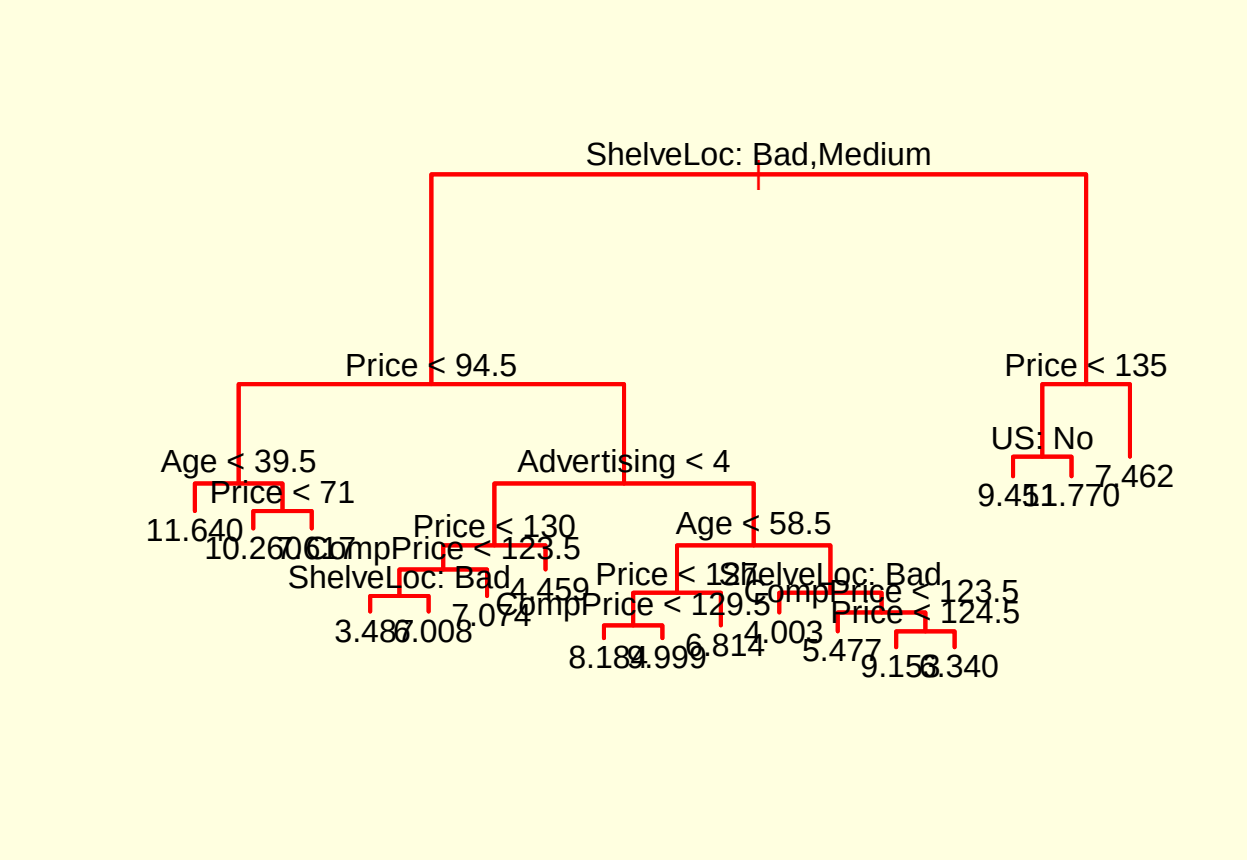
```
set.seed(5)
cv<-cv.tree(ctree)
par(mfrow=c(1,2),bg="lightyellow")
plot(cv$size,cv$dev,type="b",col="darkred",lwd=2)
plot(cv$k,cv$dev,type="b",col="darkgreen",lwd=2)
```



Insight3 : Lowest deviance at the size of 17 is achieved by above plots

Apply this size and prune to tree

```
set.seed(1)
pcv<-prune.tree(ctree,best = 17)
par(mfrow=c(1,1),bg="lightyellow")
plot(pcv,col="red",lwd=2)
text(pcv,pretty=0)
```




```
TESTMSE(pcv)
```

```
## [1] 4.827162
```

Insight4: MSE slightly reduced after prune the model from 4.9 to 4.8.

(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.

```
bc<-randomForest(Sales~.,data=train,mtry=10,importance=T)
TESTMSE(bc)
```

```
## [1] 2.605253
```

```
bc$importance
```

```
##                %IncMSE IncNodePurity
## CompPrice    1.258348871    170.182937
## Income       0.129497944     91.264880
## Advertising  0.416330655     97.164338
## Population  -0.041240944     58.244596
## Price        5.007734520    502.903407
## ShelveLoc    3.351130076    380.032715
## Age          0.752727319    157.846774
## Education    0.012742260     44.598731
## Urban        0.001441772      9.822082
## US           0.054097611     18.073863
```

Insight 4: MSE lowered from 4.8 to 2.6 by apply bagging.Price,Shelveloc,Compprice and age in that order are top importance variables to predict the Sales.

(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.

```
set.seed(2)
mtrail<- function(x){
h=x
carrftree<-randomForest(Sales~.,data=train,mtry=h,importance=T)
TESTMSE(carrftree)

}

for (i in 1:10){
  mtrail(i)
  print(mtrail(i))

}
```

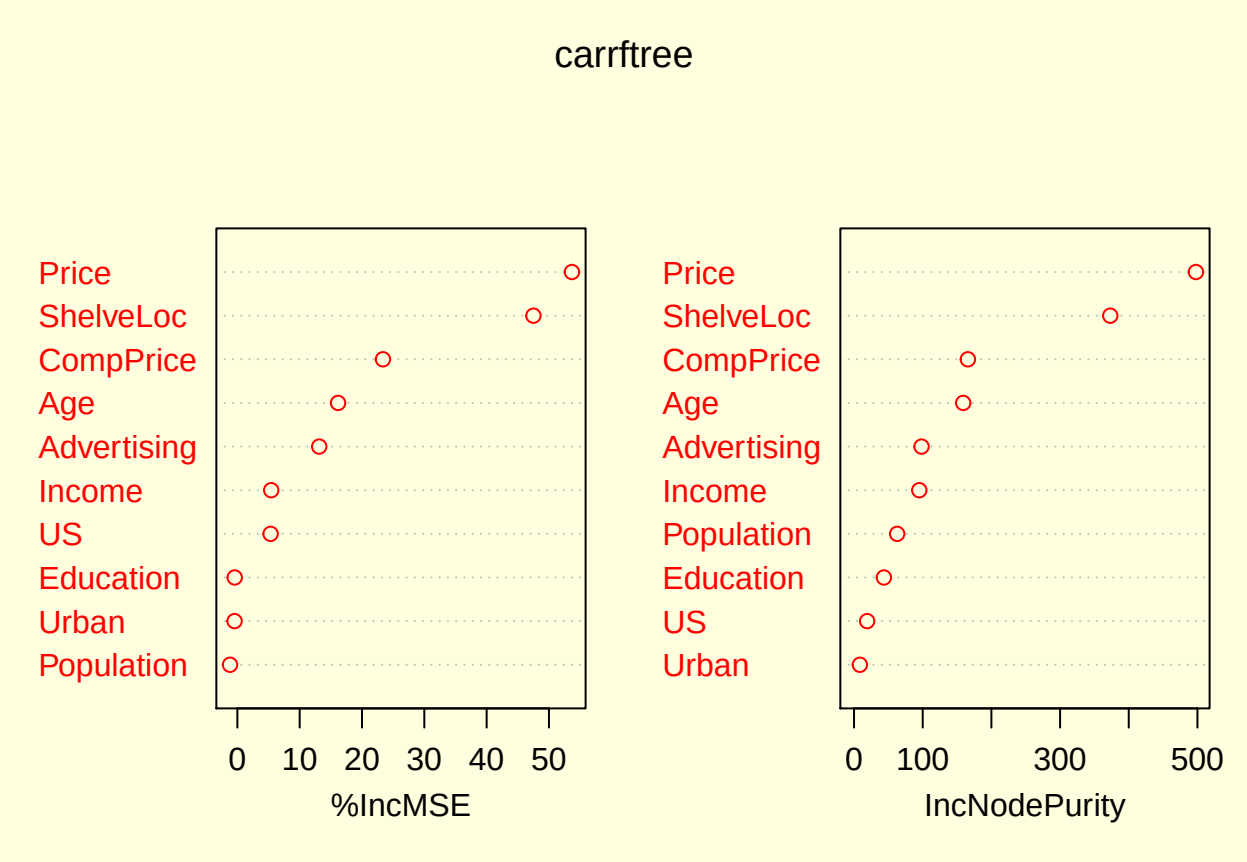
```
## [1] 4.886725
## [1] 3.501771
## [1] 3.029013
## [1] 2.760433
## [1] 2.723557
## [1] 2.668834
## [1] 2.657156
## [1] 2.65958
## [1] 2.614046
## [1] 2.631261
```

Insight 5: as m number increases Test MSE values decreases till m=9 the next m value plugged the Test MSE increases.Best fit in this instance 9 varaiables to predict the Sales.

```
carrftree<-randomForest(Sales~.,data=train,mtry=9,importance=T)
carrftree$importance
```

```
##                %IncMSE IncNodePurity
## CompPrice    1.350541540    165.752045
## Income       0.144647070     95.025145
## Advertising  0.428040522     98.298566
## Population  -0.028370465     62.832793
## Price        5.044630902    497.791260
## ShelveLoc    3.301238427    373.078529
## Age          0.681935767    158.980198
## Education   -0.008287301     43.465417
## Urban       -0.003716672      8.492259
## US           0.071805729     19.119600
```

```
par(mfrow=c(1,1),bg="lightyellow")
varImpPlot(carrftree,col="red",lwd=2)
```



Same as insight 4 i.e Price,Shelveloc,Compprice and age in that order are top importance variables to predict the Sales.