STAT2450 Assignment #7

James MacPhee

B00768516

#(a)  
library(ISLR)  
set.seed(12)  
index=sample(1:nrow(OJ),800,replace=F)  
OJtrain=OJ[index,]  
OJtest=OJ[-index,]

#(b)  
library(tree)  
OJtree=tree(Purchase~.,data=OJtrain)  
summary=summary(OJtree)  
summary

##   
## Classification tree:  
## tree(formula = Purchase ~ ., data = OJtrain)  
## Variables actually used in tree construction:  
## [1] "LoyalCH" "PriceDiff"  
## Number of terminal nodes: 7   
## Residual mean deviance: 0.7513 = 595.8 / 793   
## Misclassification error rate: 0.1688 = 135 / 800

trainRate=summary$misclass[1]/summary$misclass[2]  
paste("Error rate: ",trainRate)

## [1] "Error rate: 0.16875"

paste("# of terminal nodes: ",summary$size)

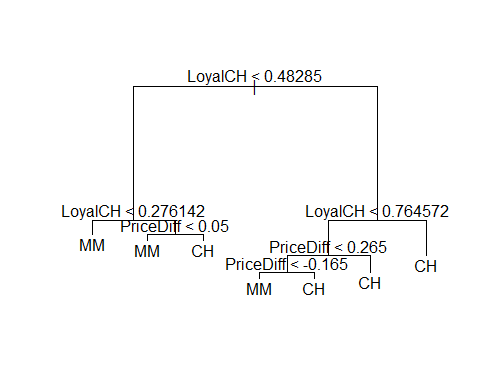
## [1] "# of terminal nodes: 7"

#(c)  
OJtree

## node), split, n, deviance, yval, (yprob)  
## \* denotes terminal node  
##   
## 1) root 800 1063.00 CH ( 0.62000 0.38000 )   
## 2) LoyalCH < 0.48285 290 315.90 MM ( 0.23448 0.76552 )   
## 4) LoyalCH < 0.276142 152 102.30 MM ( 0.10526 0.89474 ) \*  
## 5) LoyalCH > 0.276142 138 182.80 MM ( 0.37681 0.62319 )   
## 10) PriceDiff < 0.05 63 66.74 MM ( 0.22222 0.77778 ) \*  
## 11) PriceDiff > 0.05 75 104.00 CH ( 0.50667 0.49333 ) \*  
## 3) LoyalCH > 0.48285 510 449.80 CH ( 0.83922 0.16078 )   
## 6) LoyalCH < 0.764572 246 300.90 CH ( 0.69919 0.30081 )   
## 12) PriceDiff < 0.265 154 210.30 CH ( 0.57143 0.42857 )   
## 24) PriceDiff < -0.165 36 44.32 MM ( 0.30556 0.69444 ) \*  
## 25) PriceDiff > -0.165 118 152.40 CH ( 0.65254 0.34746 ) \*  
## 13) PriceDiff > 0.265 92 54.36 CH ( 0.91304 0.08696 ) \*  
## 7) LoyalCH > 0.764572 264 71.70 CH ( 0.96970 0.03030 ) \*

*#Choosing node 8) the tree will predict that Minute Maid will be choosen if Citrus Hill cost less than 0.13 more than Minute Maid and that customer loyalty in less than 0.469289*

#(d)  
plot(OJtree)  
text(OJtree)



*#From the tree we can see that when there is high customer loyalty Citrus Hill will be choosen unless the week of purchase is very high (late), and when customer loyalty is low Minute Maid will be choosen as long as the it is cheaper or at least not too much more expensive than Citrus Hill.*

#(e)  
OJpred=predict(OJtree,OJtest,type='class')  
table=table(OJpred,OJtest[,'Purchase'])  
table

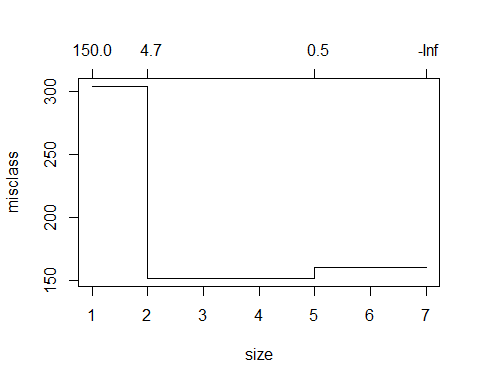
##   
## OJpred CH MM  
## CH 143 31  
## MM 14 82

testRate=(sum(table)-sum(diag(table)))/sum(table)  
testRate

## [1] 0.1666667

#(f)  
OJcv=cv.tree(OJtree, FUN=prune.misclass)

#(g)  
plot(OJcv)



#(h)  
bestsize=OJcv$size[OJcv$dev==min(OJcv$dev)]

#(i)  
prunedTree=prune.misclass(OJtree, best=bestsize)

#(j)  
prunedTrainPred=predict(prunedTree,OJtrain,type='class')  
table=table(prunedTrainPred,OJtrain[,'Purchase'])  
prunedTrainRate=(sum(table)-sum(diag(table)))/sum(table)  
paste("Unpruned training error rate is: ",trainRate)

## [1] "Unpruned training error rate is: 0.16875"

paste("Pruned training error rate is slightly higher at: ",prunedTrainRate)

## [1] "Pruned training error rate is slightly higher at: 0.1875"

#(k)  
prunedTestPred=predict(prunedTree,OJtest,type='class')  
table=table(prunedTestPred,OJtest[,'Purchase'])  
prunedTestRate=(sum(table)-sum(diag(table)))/sum(table)  
paste("Unpruned testing error rate is: ",testRate)

## [1] "Unpruned testing error rate is: 0. 166666666666667"

paste("Pruned testing error rate is slightly higher at: ",prunedTestRate)

## [1] "Pruned testing error rate is slightly higher at: 0.2 "

I had to change the results slightly after knitting to Word document as it seemed that the seed I was using didn’t carry through so the data change and therefore some of the results