Lab7

Sri Seshadri

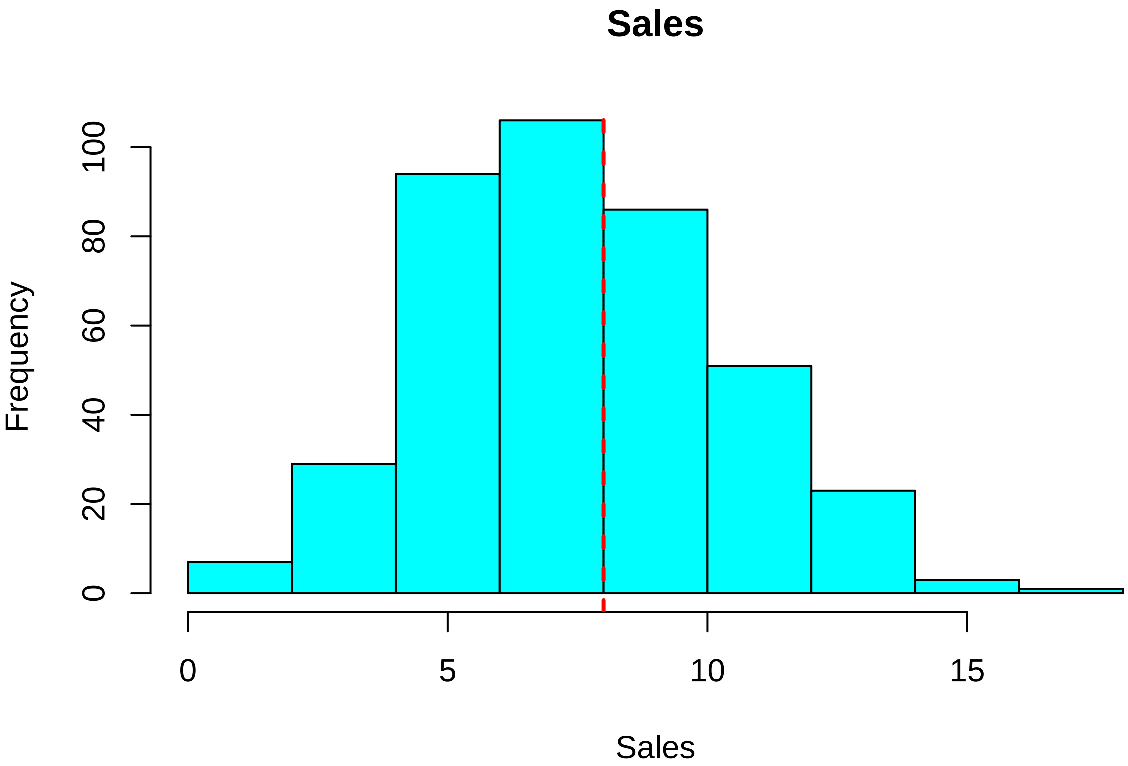
3/4/2018

8.3 Lab: Decision Trees Page - 323

8.3.1 Fitting classification Trees

We'll use Carseats data set. An additional variable "High" is created as a logical vector. High being "Yes", when sales is > 8.

```
data("Carseats")
#skimr::skim(Carseats)
hist(Carseats$Sales,xlab = "Sales", main = "Sales", col = c("#00FFFF"))
abline(v=8,col= "red", lty = 2, lwd = 2)
```



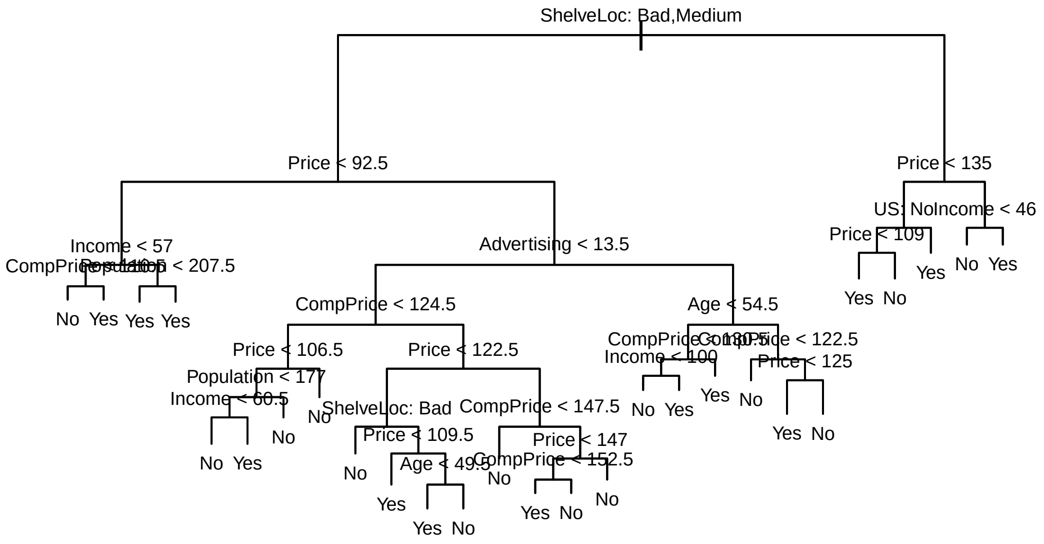
```
Carseats <- Carseats %>% dplyr::mutate(High = as.factor(if_else(Sales <= 8, "No", "Yes")))
```

page 324:

We'll use a tree() to fit a classiifcation tree to predict if the "High" variable is "Yes" or "No"

```
set.seed(10)
tree.carseats <- tree(High ~ . , data = Carseats[,-1])
```

```r
plot(tree.carseats)
text(tree.carseats,pretty=0,cex = .6)
```



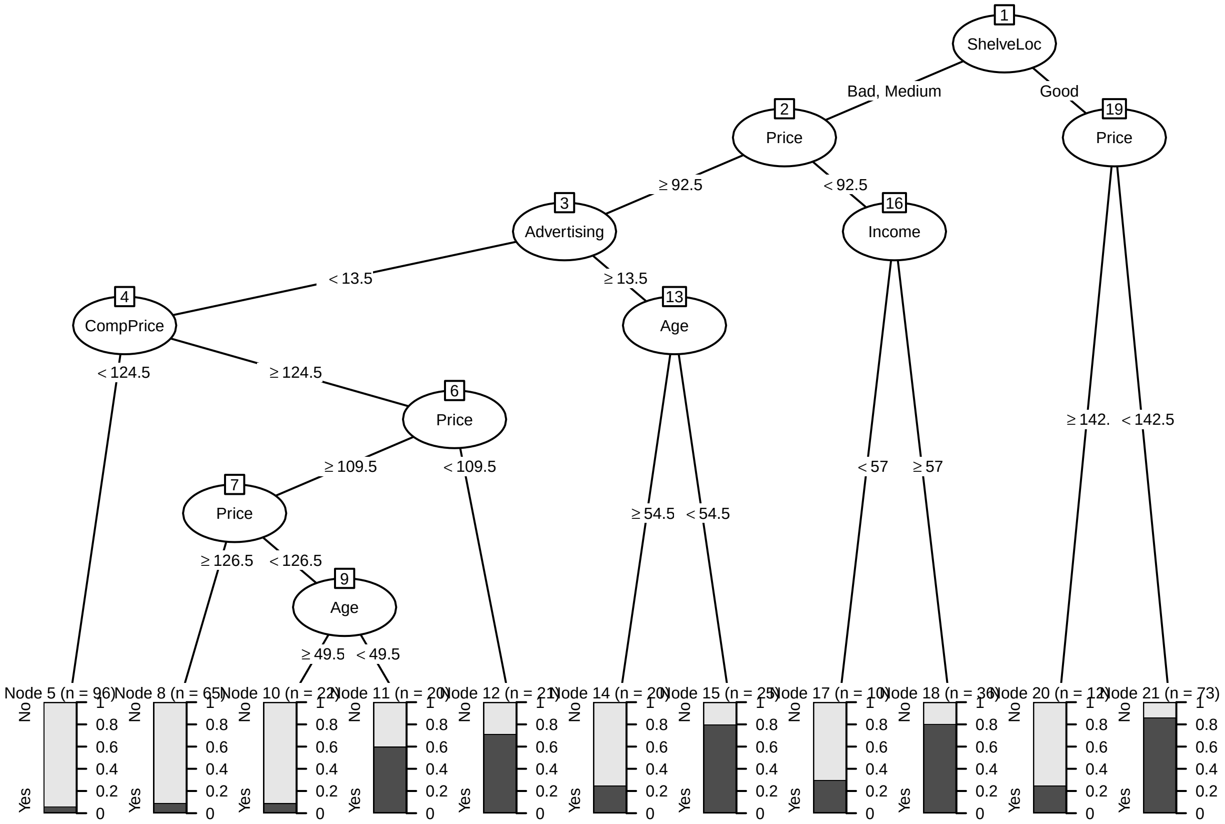
```r
summary(tree.carseats)
```

```
##
## Classification tree:
## tree(formula = High ~ ., data = Carseats[, -1])
## Variables actually used in tree construction:
## [1] "ShelveLoc"   "Price"       "Income"      "CompPrice"   "Population"
## [6] "Advertising" "Age"         "US"
## Number of terminal nodes:  27
## Residual mean deviance:  0.4575 = 170.7 / 373
## Misclassification error rate: 0.09 = 36 / 400
```

What does rpart do?

Why is there a difference between misclassification between rpart and tree? Also notice the number of terminal nodes (of course the rpart's terminal node collapes the Yes and No into a stacked bar… but still look at the difference)

```r
set.seed(10)
rpartTree <- rpart::rpart(High ~ . , data = Carseats[,-1])
plot(partykit::as.party(rpartTree),gp = gpar(fontsize = 6))
```



```r
#summary(rpartTree)
rpartTreePred <- predict(rpartTree,newdata = Carseats[,-1])
rpartTreePred <- as.factor(if_else(rpartTreePred[,2] > 0.5,"Yes","No"))
caret::confusionMatrix(rpartTreePred,Carseats$High, positive = "Yes")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  200  25
##        Yes  36 139
##
##                Accuracy : 0.8475
##                  95% CI : (0.8085, 0.8813)
##     No Information Rate : 0.59
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.688
##  Mcnemar's Test P-Value : 0.2004
##
##             Sensitivity : 0.8476
##             Specificity : 0.8475
##          Pos Pred Value : 0.7943
##          Neg Pred Value : 0.8889
##              Prevalence : 0.4100
##          Detection Rate : 0.3475
##    Detection Prevalence : 0.4375
##       Balanced Accuracy : 0.8475
```

```
##
##        'Positive' Class : Yes
##
```

Let's predict outcome of tree() with predict()

The misclassifications are different between tree() and rpart(). What makes the difference?

```
tree.carseats.Pred <- predict(tree.carseats,newdata = Carseats[,-1])
tree.carseats.Pred <- as.factor(if_else(tree.carseats.Pred[,2] > 0.5,"Yes","No"))
caret::confusionMatrix(tree.carseats.Pred,Carseats$High, positive = "Yes")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  213  13
##        Yes  23 151
##
##                Accuracy : 0.91
##                  95% CI : (0.8776, 0.9362)
##     No Information Rate : 0.59
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.8157
##  Mcnemar's Test P-Value : 0.1336
##
##             Sensitivity : 0.9207
##             Specificity : 0.9025
##          Pos Pred Value : 0.8678
##          Neg Pred Value : 0.9425
##              Prevalence : 0.4100
##          Detection Rate : 0.3775
##    Detection Prevalence : 0.4350
##       Balanced Accuracy : 0.9116
##
##        'Positive' Class : Yes
##
```