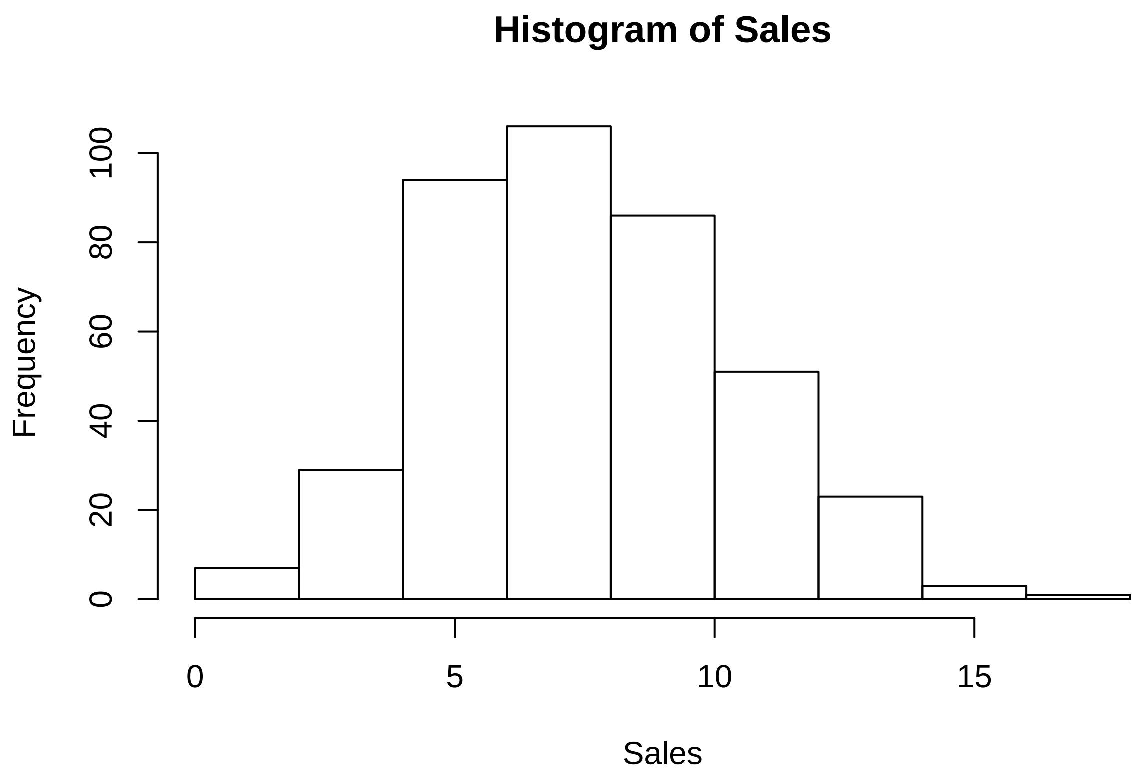
Tree-Based Methods

Contents

Introduction

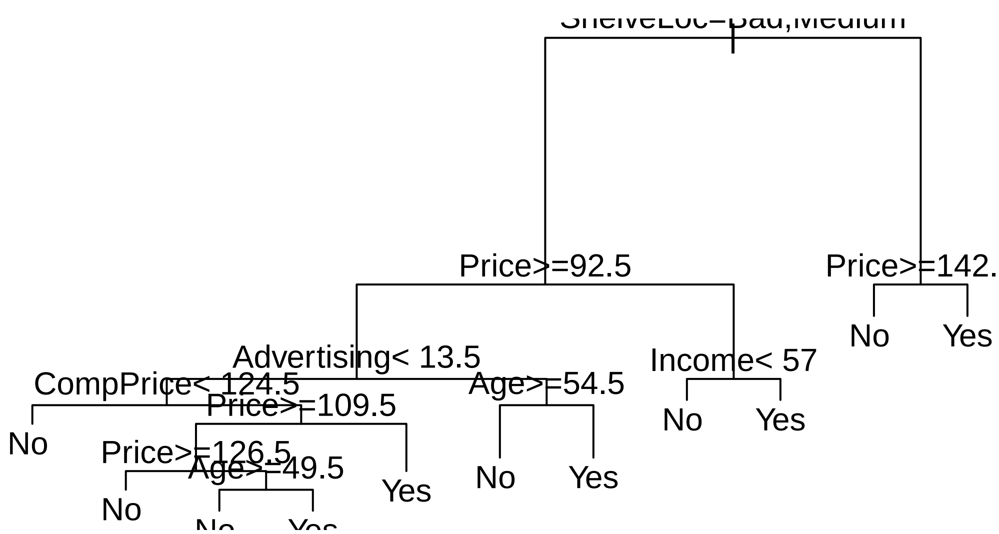
We will have a look at the `Carseats` data using the `tree`, `rpart`, and `party` packages in R, as in the lab in the book. We create a binary response variable `High` (for high sales), and we include it in the same dataframe.

```
library(ISLR)
attach(Carseats)

hist(Sales)
```



Now we fit a tree to these data, and summarize and plot it. Notice that we have to *exclude* `Sales` from the right-hand side of the formula, because the response is derived from it.

```
High <- ifelse(Sales <= 8, "No", "Yes")
Carseats <- data.frame(Carseats, High)

library(rpart)
tree.carseats <- rpart(High~.-Sales, data = Carseats)

plot(tree.carseats)
text(tree.carseats, pretty=1)
```




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

```
## Call:
## rpart(formula = High ~ . - Sales, data = Carseats)
##   n= 400
##
##           CP nsplit rel error    xerror       xstd
## 1 0.28658537      0 1.0000000 1.0000000 0.05997967
## 2 0.10975610      1 0.7134146 0.7134146 0.05547692
## 3 0.04573171      2 0.6036585 0.6646341 0.05429826
## 4 0.03658537      4 0.5121951 0.6707317 0.05445296
## 5 0.02743902      5 0.4756098 0.6402439 0.05365767
## 6 0.02439024      7 0.4207317 0.6768293 0.05460552
## 7 0.01219512      8 0.3963415 0.6402439 0.05365767
## 8 0.01000000     10 0.3719512 0.6402439 0.05365767
##
## Variable importance
##       Price   ShelveLoc         Age Advertising   CompPrice      Income
##          34          25          11          11           9           5
##  Population   Education
##           3           1
##
## Node number 1: 400 observations,    complexity param=0.2865854
##   predicted class=No   expected loss=0.41  P(node) =1
##     class counts:   236   164
##    probabilities: 0.590 0.410
```

```
##   left son=2 (315 obs) right son=3 (85 obs)
##   Primary splits:
##       ShelveLoc   splits as  LRL,       improve=28.991900, (0 missing)
##       Price       < 92.5  to the right, improve=19.463880, (0 missing)
##       Advertising < 6.5   to the left,  improve=17.277980, (0 missing)
##       Age         < 61.5  to the right, improve= 9.264442, (0 missing)
##       Income      < 60.5  to the left,  improve= 7.249032, (0 missing)
##
## Node number 2: 315 observations,    complexity param=0.1097561
##   predicted class=No   expected loss=0.3111111  P(node) =0.7875
##     class counts:   217    98
##    probabilities: 0.689 0.311
##   left son=4 (269 obs) right son=5 (46 obs)
##   Primary splits:
##       Price       < 92.5  to the right, improve=15.930580, (0 missing)
##       Advertising < 7.5   to the left,  improve=11.432570, (0 missing)
##       ShelveLoc   splits as  L-R,       improve= 7.543912, (0 missing)
##       Age         < 50.5  to the right, improve= 6.369905, (0 missing)
##       Income      < 60.5  to the left,  improve= 5.984509, (0 missing)
##   Surrogate splits:
##       CompPrice < 95.5  to the right, agree=0.873, adj=0.13, (0 split)
##
## Node number 3: 85 observations,    complexity param=0.03658537
##   predicted class=Yes  expected loss=0.2235294  P(node) =0.2125
##     class counts:    19    66
##    probabilities: 0.224 0.776
##   left son=6 (12 obs) right son=7 (73 obs)
##   Primary splits:
##       Price       < 142.5 to the right, improve=7.745608, (0 missing)
##       US          splits as  LR,        improve=5.112440, (0 missing)
##       Income      < 35    to the left,  improve=4.529433, (0 missing)
##       Advertising < 6     to the left,  improve=3.739996, (0 missing)
##       Education   < 15.5  to the left,  improve=2.565856, (0 missing)
##   Surrogate splits:
##       CompPrice < 154.5 to the right, agree=0.882, adj=0.167, (0 split)
##
## Node number 4: 269 observations,    complexity param=0.04573171
##   predicted class=No   expected loss=0.2453532  P(node) =0.6725
##     class counts:   203    66
##    probabilities: 0.755 0.245
##   left son=8 (224 obs) right son=9 (45 obs)
##   Primary splits:
##       Advertising < 13.5  to the left,  improve=10.400090, (0 missing)
##       Age         < 49.5  to the right, improve= 8.083998, (0 missing)
##       ShelveLoc   splits as  L-R,       improve= 7.023150, (0 missing)
##       CompPrice   < 124.5 to the left,  improve= 6.749986, (0 missing)
##       Price       < 126.5 to the right, improve= 5.646063, (0 missing)
##
## Node number 5: 46 observations,    complexity param=0.02439024
##   predicted class=Yes  expected loss=0.3043478  P(node) =0.115
##     class counts:    14    32
##    probabilities: 0.304 0.696
##   left son=10 (10 obs) right son=11 (36 obs)
##   Primary splits:
```

```
##       Income      < 57    to the left,  improve=4.000483, (0 missing)
##       ShelveLoc   splits as  L-R,       improve=3.189762, (0 missing)
##       Advertising < 9.5   to the left,  improve=1.388592, (0 missing)
##       Price       < 80.5  to the right, improve=1.388592, (0 missing)
##       Age         < 64.5  to the right, improve=1.172885, (0 missing)
##
## Node number 6: 12 observations
##   predicted class=No   expected loss=0.25  P(node) =0.03
##     class counts:     9     3
##    probabilities: 0.750 0.250
##
## Node number 7: 73 observations
##   predicted class=Yes  expected loss=0.1369863  P(node) =0.1825
##     class counts:    10    63
##    probabilities: 0.137 0.863
##
## Node number 8: 224 observations,    complexity param=0.02743902
##   predicted class=No   expected loss=0.1830357  P(node) =0.56
##     class counts:   183    41
##    probabilities: 0.817 0.183
##   left son=16 (96 obs) right son=17 (128 obs)
##   Primary splits:
##       CompPrice   < 124.5 to the left,  improve=4.881696, (0 missing)
##       Age         < 49.5  to the right, improve=3.960418, (0 missing)
##       ShelveLoc   splits as  L-R,       improve=3.654633, (0 missing)
##       Price       < 126.5 to the right, improve=3.234428, (0 missing)
##       Advertising < 6.5   to the left,  improve=2.371276, (0 missing)
##   Surrogate splits:
##       Price      < 115.5 to the left,  agree=0.741, adj=0.396, (0 split)
##       Age        < 50.5  to the right, agree=0.634, adj=0.146, (0 split)
##       Population < 405   to the right, agree=0.629, adj=0.135, (0 split)
##       Education  < 11.5  to the left,  agree=0.585, adj=0.031, (0 split)
##       Income     < 22.5  to the left,  agree=0.580, adj=0.021, (0 split)
##
## Node number 9: 45 observations,    complexity param=0.04573171
##   predicted class=Yes  expected loss=0.4444444  P(node) =0.1125
##     class counts:    20    25
##    probabilities: 0.444 0.556
##   left son=18 (20 obs) right son=19 (25 obs)
##   Primary splits:
##       Age       < 54.5  to the right, improve=6.722222, (0 missing)
##       CompPrice < 121.5 to the left,  improve=4.629630, (0 missing)
##       ShelveLoc splits as  L-R,       improve=3.250794, (0 missing)
##       Income    < 99.5  to the left,  improve=3.050794, (0 missing)
##       Price     < 127   to the right, improve=2.933429, (0 missing)
##   Surrogate splits:
##       Population  < 363.5 to the left,  agree=0.667, adj=0.25, (0 split)
##       Income      < 39    to the left,  agree=0.644, adj=0.20, (0 split)
##       Advertising < 17.5  to the left,  agree=0.644, adj=0.20, (0 split)
##       CompPrice   < 106.5 to the left,  agree=0.622, adj=0.15, (0 split)
##       Price       < 135.5 to the right, agree=0.622, adj=0.15, (0 split)
##
## Node number 10: 10 observations
##   predicted class=No   expected loss=0.3  P(node) =0.025
```

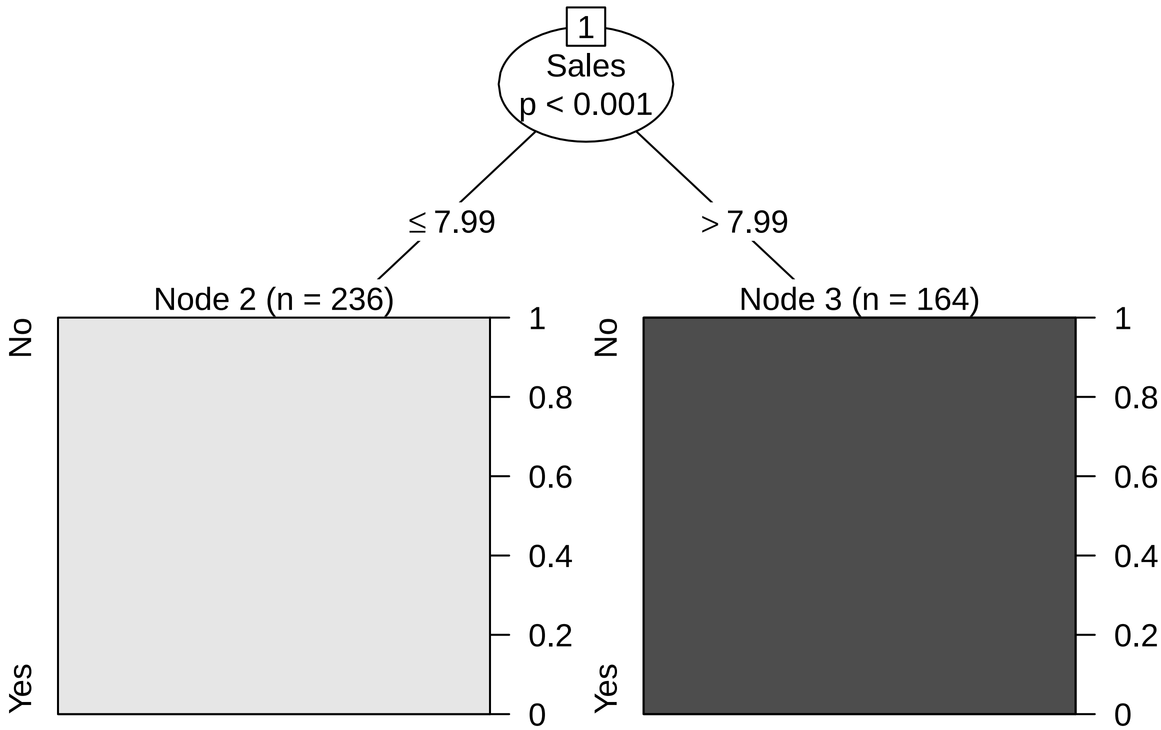
```
##     class counts:     7     3
##    probabilities: 0.700 0.300
##
## Node number 11: 36 observations
##   predicted class=Yes  expected loss=0.1944444  P(node) =0.09
##     class counts:     7    29
##    probabilities: 0.194 0.806
##
## Node number 16: 96 observations
##   predicted class=No   expected loss=0.0625  P(node) =0.24
##     class counts:    90     6
##    probabilities: 0.938 0.062
##
## Node number 17: 128 observations,    complexity param=0.02743902
##   predicted class=No   expected loss=0.2734375  P(node) =0.32
##     class counts:    93    35
##    probabilities: 0.727 0.273
##   left son=34 (107 obs) right son=35 (21 obs)
##   Primary splits:
##       Price     < 109.5 to the right, improve=9.764582, (0 missing)
##       ShelveLoc splits as  L-R,       improve=6.320022, (0 missing)
##       Age       < 49.5  to the right, improve=2.575061, (0 missing)
##       Income    < 108.5 to the right, improve=1.799546, (0 missing)
##       CompPrice < 143.5 to the left,  improve=1.741982, (0 missing)
##
## Node number 18: 20 observations
##   predicted class=No   expected loss=0.25  P(node) =0.05
##     class counts:    15     5
##    probabilities: 0.750 0.250
##
## Node number 19: 25 observations
##   predicted class=Yes  expected loss=0.2  P(node) =0.0625
##     class counts:     5    20
##    probabilities: 0.200 0.800
##
## Node number 34: 107 observations,    complexity param=0.01219512
##   predicted class=No   expected loss=0.1869159  P(node) =0.2675
##     class counts:    87    20
##    probabilities: 0.813 0.187
##   left son=68 (65 obs) right son=69 (42 obs)
##   Primary splits:
##       Price     < 126.5 to the right, improve=2.9643900, (0 missing)
##       CompPrice < 147.5 to the left,  improve=2.2337090, (0 missing)
##       ShelveLoc splits as  L-R,       improve=2.2125310, (0 missing)
##       Age       < 49.5  to the right, improve=2.1458210, (0 missing)
##       Income    < 60.5  to the left,  improve=0.8025853, (0 missing)
##   Surrogate splits:
##       CompPrice   < 129.5 to the right, agree=0.664, adj=0.143, (0 split)
##       Advertising < 3.5   to the right, agree=0.664, adj=0.143, (0 split)
##       Population  < 53.5  to the right, agree=0.645, adj=0.095, (0 split)
##       Age         < 77.5  to the left,  agree=0.636, adj=0.071, (0 split)
##       US          splits as  RL,        agree=0.626, adj=0.048, (0 split)
##
## Node number 35: 21 observations
```

```
##   predicted class=Yes  expected loss=0.2857143  P(node) =0.0525
##     class counts:     6    15
##    probabilities: 0.286 0.714
##
## Node number 68: 65 observations
##   predicted class=No   expected loss=0.09230769  P(node) =0.1625
##     class counts:    59     6
##    probabilities: 0.908 0.092
##
## Node number 69: 42 observations,    complexity param=0.01219512
##   predicted class=No   expected loss=0.3333333  P(node) =0.105
##     class counts:    28    14
##    probabilities: 0.667 0.333
##   left son=138 (22 obs) right son=139 (20 obs)
##   Primary splits:
##       Age         < 49.5  to the right, improve=5.4303030, (0 missing)
##       CompPrice   < 137.5 to the left,  improve=2.1000000, (0 missing)
##       Advertising < 5.5   to the left,  improve=1.8666670, (0 missing)
##       ShelveLoc   splits as  L-R,       improve=1.4291670, (0 missing)
##       Population  < 382   to the right, improve=0.8578431, (0 missing)
##   Surrogate splits:
##       Income      < 46.5  to the left,  agree=0.595, adj=0.15, (0 split)
##       Education   < 12.5  to the left,  agree=0.595, adj=0.15, (0 split)
##       CompPrice   < 131.5 to the right, agree=0.571, adj=0.10, (0 split)
##       Advertising < 5.5   to the left,  agree=0.571, adj=0.10, (0 split)
##       Population  < 221.5 to the left,  agree=0.571, adj=0.10, (0 split)
##
## Node number 138: 22 observations
##   predicted class=No   expected loss=0.09090909  P(node) =0.055
##     class counts:    20     2
##    probabilities: 0.909 0.091
##
## Node number 139: 20 observations
##   predicted class=Yes  expected loss=0.4  P(node) =0.05
##     class counts:     8    12
##    probabilities: 0.400 0.600
```

```r
library(party)
```

```
## Loading required package: grid

## Loading required package: mvtnorm

## Loading required package: modeltools

## Loading required package: stats4

## Loading required package: strucchange

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich
```

```r
tree_carseats <- ctree(High~.-Sales, data = Carseats)

plot(tree_carseats)
```




```r
summary(tree_carseats)
```

```
##     Length      Class       Mode
##          1 BinaryTree         S4
```

For a detailed summary of the tree, print it:

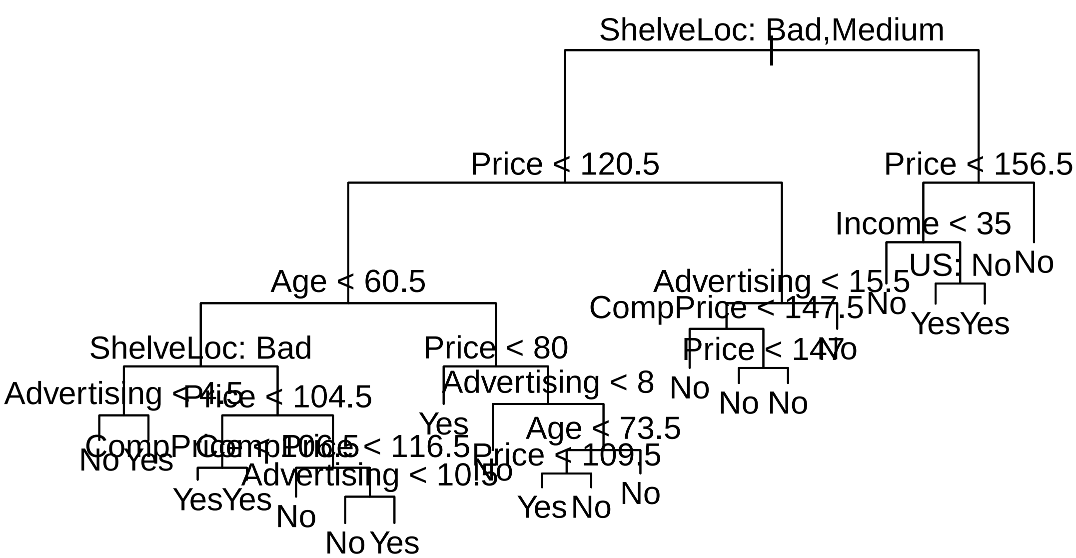
```r
tree.carseats
```

```
## n= 400
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##   1) root 400 164 No (0.59000000 0.41000000)
##     2) ShelveLoc=Bad,Medium 315  98 No (0.68888889 0.31111111)
##       4) Price>=92.5 269  66 No (0.75464684 0.24535316)
##         8) Advertising< 13.5 224  41 No (0.81696429 0.18303571)
##          16) CompPrice< 124.5 96   6 No (0.93750000 0.06250000) *
##          17) CompPrice>=124.5 128  35 No (0.72656250 0.27343750)
##            34) Price>=109.5 107  20 No (0.81308411 0.18691589)
##              68) Price>=126.5 65   6 No (0.90769231 0.09230769) *
##              69) Price< 126.5 42  14 No (0.66666667 0.33333333)
##               138) Age>=49.5 22   2 No (0.90909091 0.09090909) *
```

```
##                139) Age< 49.5 20   8 Yes (0.40000000 0.60000000) *
##             35) Price< 109.5 21   6 Yes (0.28571429 0.71428571) *
##          9) Advertising>=13.5 45  20 Yes (0.44444444 0.55555556)
##           18) Age>=54.5 20   5 No (0.75000000 0.25000000) *
##           19) Age< 54.5 25   5 Yes (0.20000000 0.80000000) *
##        5) Price< 92.5 46  14 Yes (0.30434783 0.69565217)
##         10) Income< 57 10   3 No (0.70000000 0.30000000) *
##         11) Income>=57 36   7 Yes (0.19444444 0.80555556) *
##      3) ShelveLoc=Good 85  19 Yes (0.22352941 0.77647059)
##        6) Price>=142.5 12   3 No (0.75000000 0.25000000) *
##        7) Price< 142.5 73  10 Yes (0.13698630 0.86301370) *
```

Lets create a training and test set (250,150) split of the 400 observations, grow the tree on the training set, and evaluate its performance on the test set.

```r
library(tree)
set.seed(1011)
train <- sample(1:nrow(Carseats), 250)
tree.carseats <- tree(High ~ .-Sales, Carseats, subset = train)

plot(tree.carseats);text(tree.carseats, pretty=0)
```




```r
tree.pred <- predict(tree.carseats, Carseats[-train,], type = "class")
with(Carseats[-train,], table(tree.pred,High))
```

```
##           High
## tree.pred No Yes
##       No  72  27
```

```
##       Yes 18  33
```

```r
(72+33)/150
```

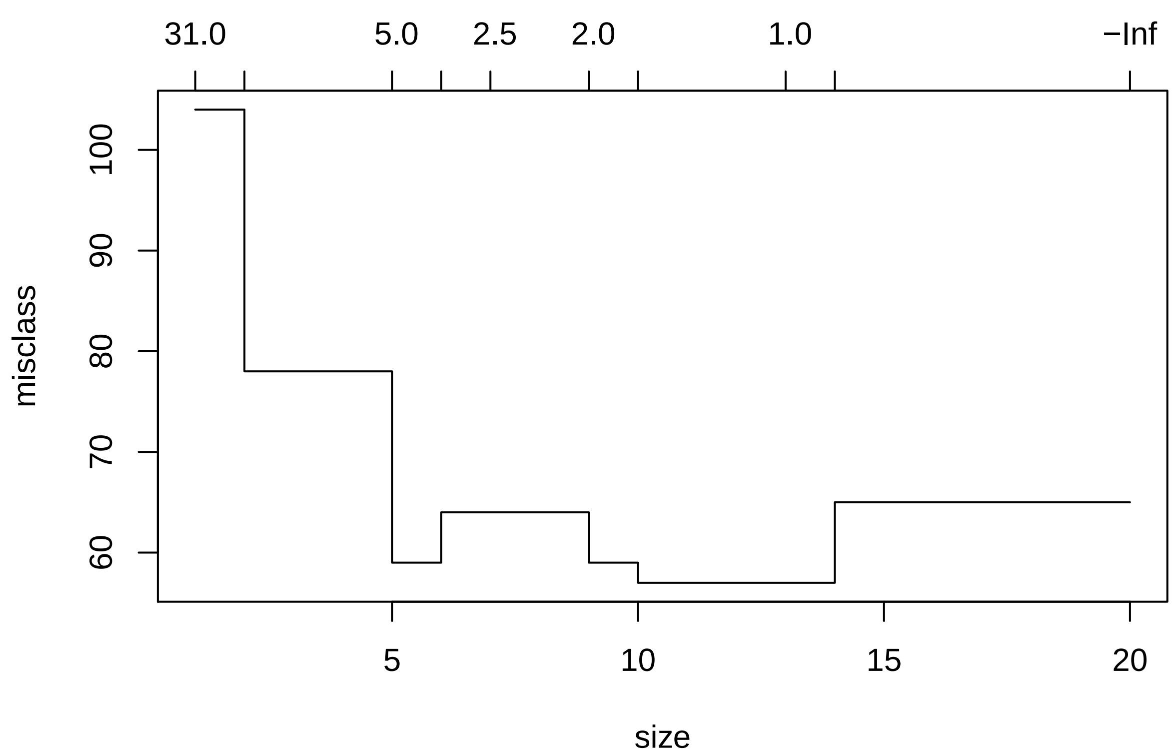
```
## [1] 0.7
```

This tree was grown to full depth, and might be too variable. We now use CV to prune it.

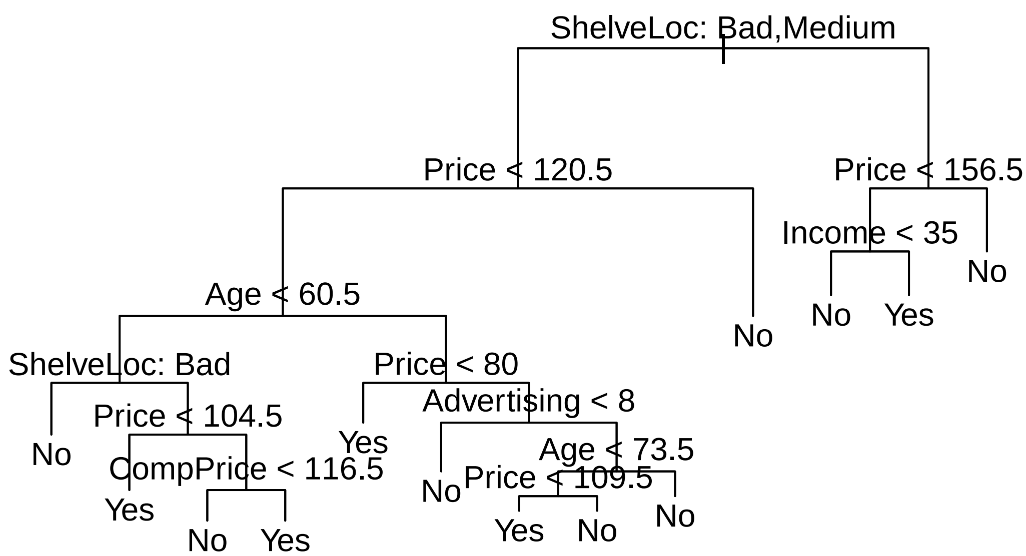
```r
library(tree)
cv.carseats <- cv.tree(tree.carseats, FUN = prune.misclass)
cv.carseats
```

```
## $size
##  [1] 20 14 13 10  9  7  6  5  2  1
##
## $dev
##  [1]  65  65  57  57  59  64  64  59  78 104
##
## $k
##  [1]      -Inf  0.000000  1.000000  1.333333  2.000000  2.500000  4.000000
##  [8]  5.000000  9.000000 31.000000
##
## $method
## [1] "misclass"
##
## attr(,"class")
## [1] "prune"         "tree.sequence"
```

```r
plot(cv.carseats)
```



```r
prune.carseats <- prune.misclass(tree.carseats, best=13)
plot(prune.carseats);text(prune.carseats, pretty=0)
```



Now lets evaluate this pruned tree on the test data.

```r
tree.pred <- predict(prune.carseats, Carseats[-train,], type="class")
with(Carseats[-train, ], table(tree.pred, High))
```

```
##          High
## tree.pred No Yes
##       No  72  28
##       Yes 18  32
```

```r
(72+32)/150
```

```
## [1] 0.6933333
```

It has done about the same as our original tree. So pruning did not hurt us wrt misclassification errors, and gave us a simpler tree.

Random Forests and Boosting

These methods use trees as building blocks to build more complex models. Here we will use the Boston housing data to explore random forests and boosting. These data are in the `MASS` package. It gives housing values and other statistics in each of 506 suburbs of Boston based on a 1970 census.

Random Forests

Random forests build lots of bushy trees, and then average them to reduce the variance.

```
library(randomForest)
```

```
## randomForest 4.6-12
## Type rfNews() to see new features/changes/bug fixes.
```

```
library(MASS)

?Boston
```

```
## starting httpd help server ...
##  done
```

```
dim(Boston)
```

```
## [1] 506  14
```

```
set.seed(101)
train <- sample(1:nrow(Boston), 300)
```

Lets fit a random forest and see how well it performs. We will use the response `medv`, the median housing value (in $1K dollars)

```
rf.boston <- randomForest(medv ~ ., data = Boston, subset = train)
rf.boston
```

```
##
## Call:
##  randomForest(formula = medv ~ ., data = Boston, subset = train)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##           Mean of squared residuals: 12.34243
##                     % Var explained: 85.09
```

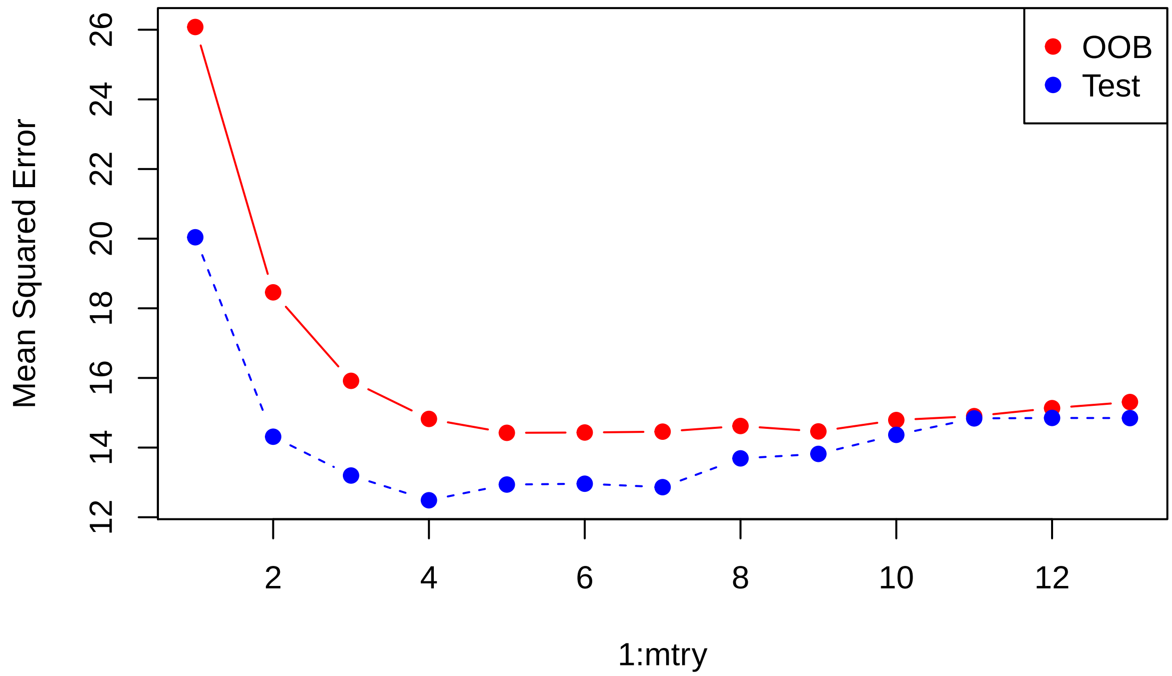
The MSR and % variance explained are based on OOB or *out-of-bag* estimates, a very clever device in random forests to get honest error estimates. The model reports that `mtry=4`, which is the number of variables randomly chosen at each split. Since $p = 13$ here, we could try all 13 possible values of `mtry`. We will do so, record the results, and make a plot.

```
oob.err <- double(13)
test.err <- double(13)

for(mtry in 1:13){
                   fit <- randomForest(medv ~., data=Boston, subset=train, mtry=mtry, ntree=400)
         oob.err[mtry] <- fit$mse[400]
                  pred <- predict(fit, Boston[-train,])
        test.err[mtry] <- with(Boston[-train,], mean((medv-pred)^2))
```

```
        cat(mtry," ")
}
```

```
## 1  2  3  4  5  6  7  8  9  10  11  12  13
```

```
matplot(1:mtry, cbind(test.err,oob.err), pch=19, col=c("red","blue"), type="b", ylab="Mean Squared Erro
legend("topright", legend=c("OOB","Test"), pch=19, col=c("red","blue"))
```

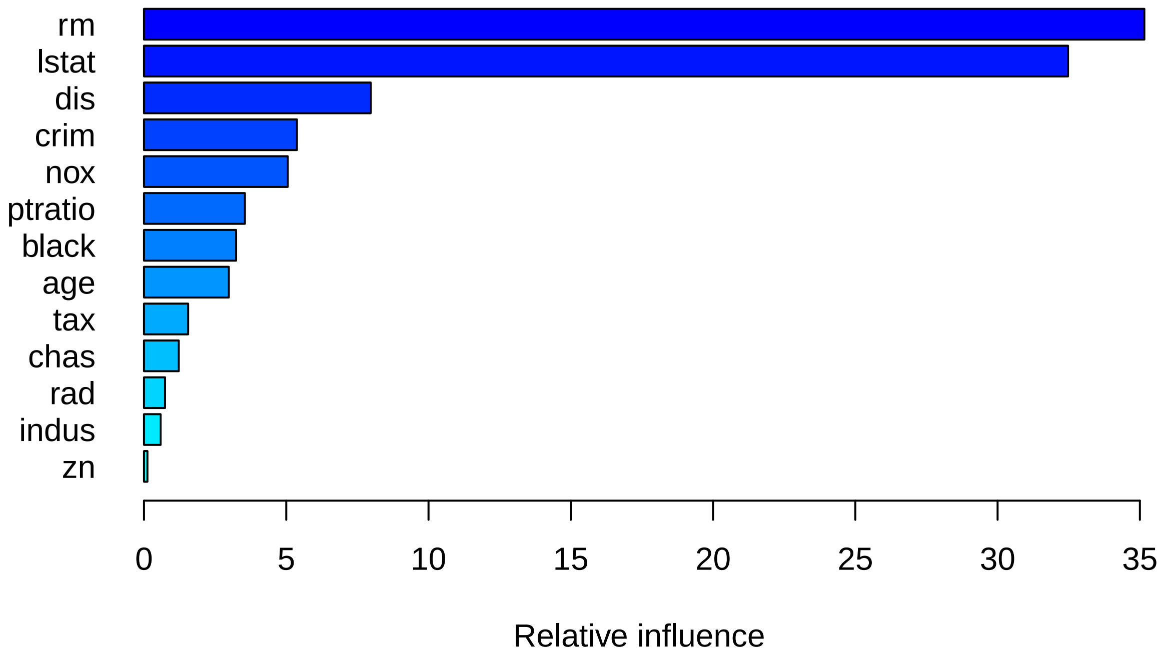


Not too difficult! Although the test-error curve drops below the OOB curve, these are estimates based on data, and so have their own standard errors (which are typically quite large). Notice that the points at the end with `mtry=13` correspond to bagging.

Boosting

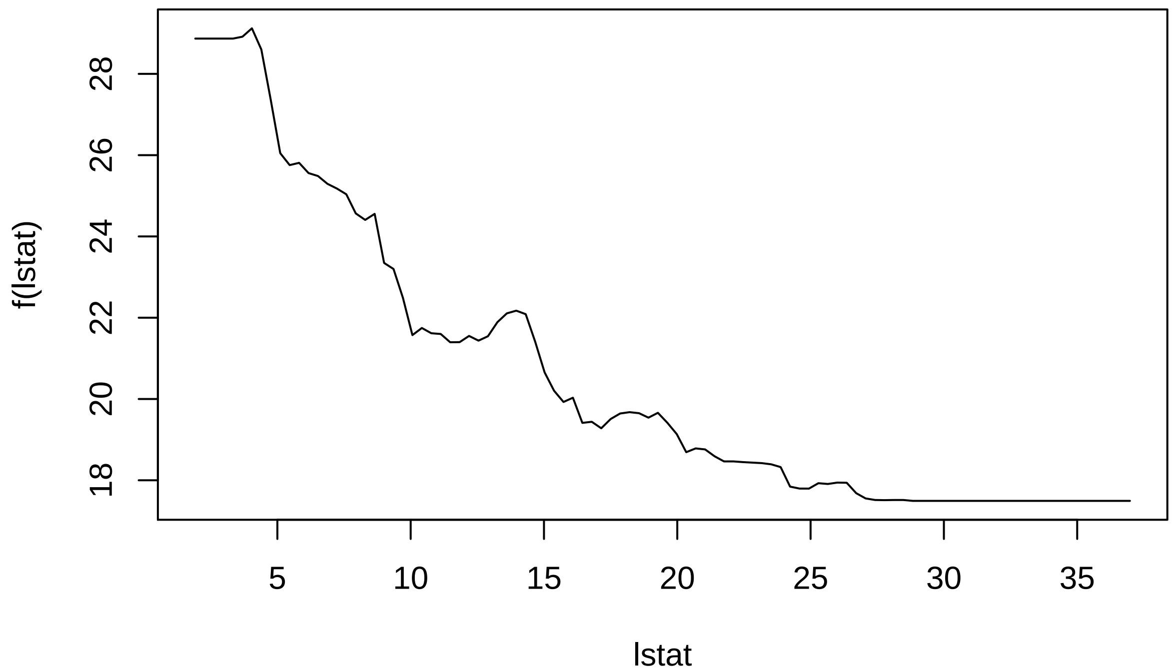
Boosting builds lots of smaller trees. Unlike random forests, each new tree in boosting tries to patch up the deficiencies of the current ensemble.

```
library(gbm3)
boost.boston <- gbm(medv ~ ., data = Boston[train, ], distribution = "gaussian",
    n.trees = 10000, shrinkage = 0.01, interaction.depth = 4)

summary(boost.boston)
```

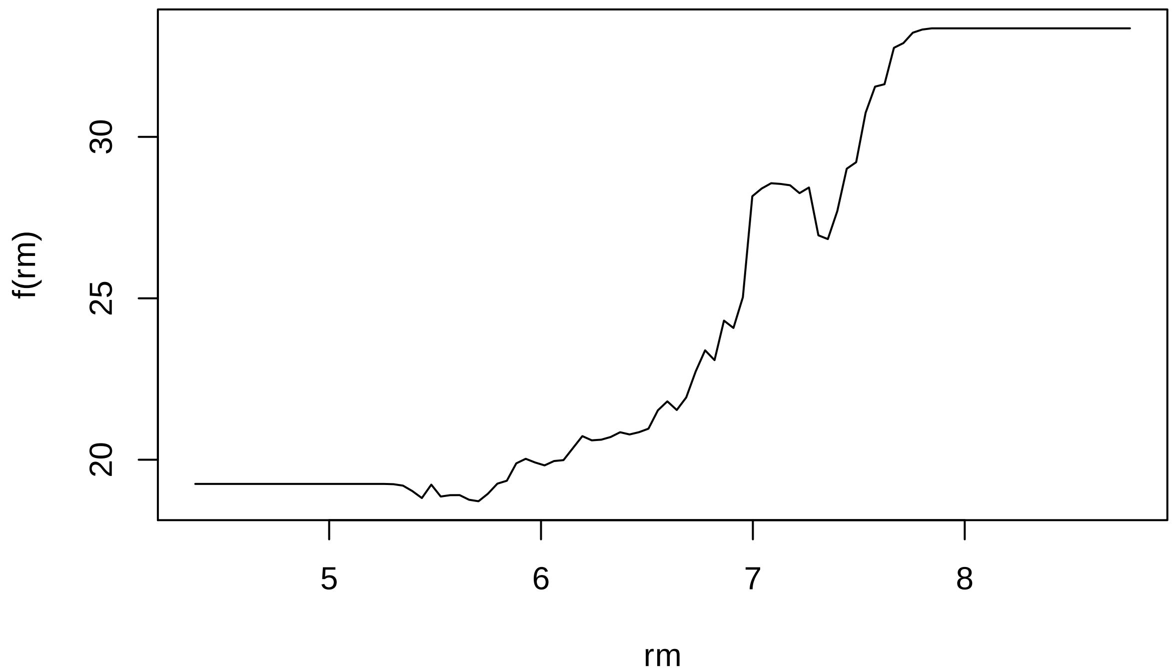



```
##              var    rel_inf
## rm            rm 35.1573249
## lstat      lstat 32.4749785
## dis          dis  7.9656899
## crim        crim  5.3735504
## nox          nox  5.0503321
## ptratio ptratio  3.5460553
## black      black  3.2374385
## age          age  2.9799571
## tax          tax  1.5512722
## chas        chas  1.2211187
## rad          rad  0.7399728
## indus      indus  0.5820085
## zn            zn  0.1203010
```

```
plot(boost.boston, var_index = "lstat")
```




```r
plot(boost.boston, var_index = "rm")
```

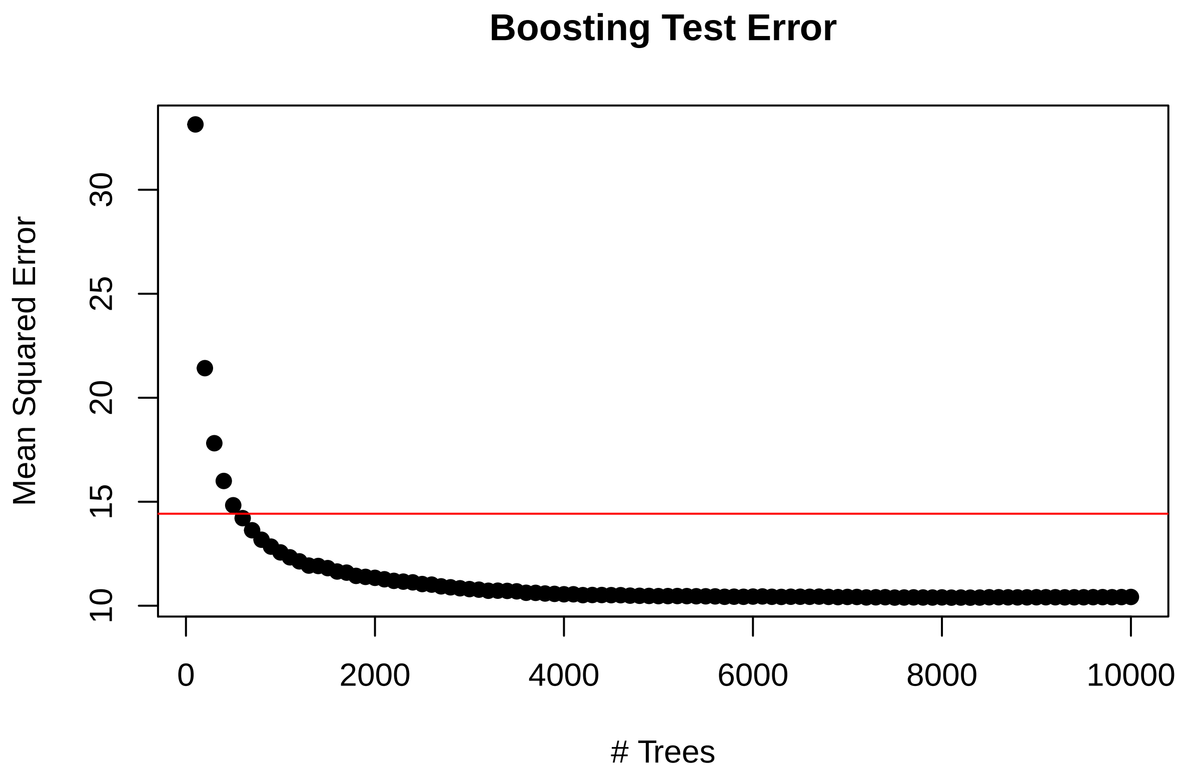


Lets make a prediction on the test set. With boosting, the number of trees is a tuning parameter, and if we have too many we can overfit. So we should use cross-validation to select the number of trees. We will leave this as an exercise. Instead, we will compute the test error as a function of the number of trees, and make a plot.

```
n.trees <- seq(from=100, to=10000, by=100)
predmat <- predict(boost.boston, newdata = Boston[-train, ], n.trees=n.trees)
dim(predmat)
```

```
## [1] 206 100
```

```
berr <- with(Boston[-train, ], apply((predmat-medv)^2,2,mean))
plot(n.trees, berr, pch=19, ylab="Mean Squared Error", xlab="# Trees", main="Boosting Test Error")
abline(h=min(test.err), col="red")
```


Boosting Test Error
Mean Squared Error
10
15
20
25
30
0
2000
4000
6000
8000
10000
Trees