Decision Tree and Random Forest

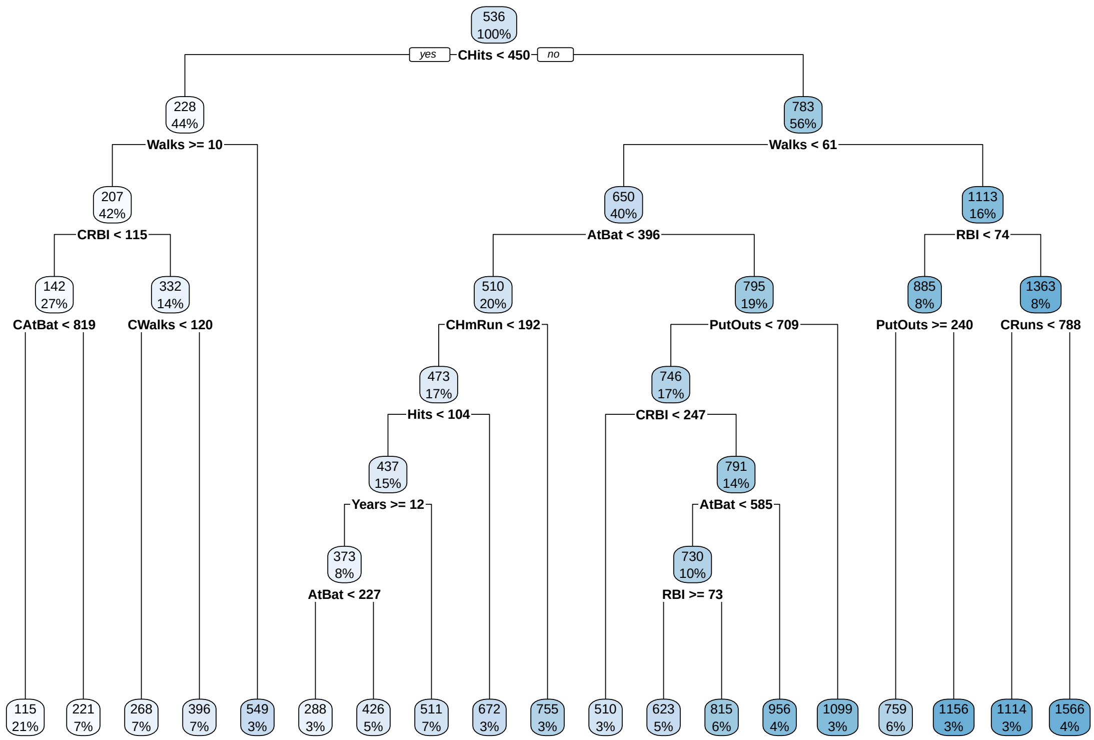
Aylin Mumcular

14th May 2018

```
res.rp <- rpart(Salary~.,d,cp=0.001)
res.rp
```

```
## n=263 (59 observations deleted due to missingness)
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##   1) root 263 53319110.00  535.9259
##     2) CHits< 450 117  5931094.00  227.8547
##       4) Walks>=10 110  1754378.00  207.4470
##         8) CRBI< 114.5 72   284426.40  141.6343
##          16) CAtBat< 818.5 54    76731.04  115.0926 *
##          17) CAtBat>=818.5 18    55531.94  221.2593 *
##         9) CRBI>=114.5 38   567215.00  332.1447
##          18) CWalks< 120 19   119848.20  268.2368 *
##          19) CWalks>=120 19   292166.40  396.0526 *
##       5) Walks< 10 7  3410996.00  548.5476 *
##     3) CHits>=450 146 27385210.00  782.8048
##       6) Walks< 61 104  9469906.00  649.6232
##        12) AtBat< 395.5 53  2859476.00  510.0157
##          24) CHmRun< 192 46  1613745.00  472.7718
##            48) Hits< 103.5 39  1192870.00  436.9872
##              96) Years>=11.5 21   354508.20  373.4127
##               192) AtBat< 226.5 8    26150.00  287.5000 *
##               193) AtBat>=226.5 13   232973.10  426.2821 *
##              97) Years< 11.5 18   654463.60  511.1574 *
##            49) Hits>=103.5 7    92692.86  672.1429 *
##          25) CHmRun>=192 7   762619.10  754.7619 *
##        13) AtBat>=395.5 51  4503956.00  794.7054
##          26) PutOuts< 709 44  2358329.00  746.3631
##            52) CRBI< 246.5 7   292530.40  509.6429 *
##            53) CRBI>=246.5 37  1599333.00  791.1480
##             106) AtBat< 585 27   975014.60  729.9065
##               212) RBI>=72.5 12   353972.80  623.4702 *
##               213) RBI< 72.5 15   376342.30  815.0555 *
##             107) AtBat>=585 10   249641.40  956.5000 *
##          27) PutOuts>=709 7  1396458.00 1098.5710 *
##       7) Walks>=61 42 11502830.00 1112.5880
##        14) RBI< 73.5 22  3148182.00  885.2651
##          28) PutOuts>=239.5 15   656292.30  758.8889 *
##          29) PutOuts< 239.5 7  1738973.00 1156.0710 *
##        15) RBI>=73.5 20  5967231.00 1362.6430
##          30) CRuns< 788 9   581309.70 1114.4440 *
##          31) CRuns>=788 11  4377879.00 1565.7150 *
```

```
rpart.plot(res.rp)
```




```r
#Recursive partitioning

predict(res.rp,newdata=d[17,]) #First grow a very large tree, then prune it
```

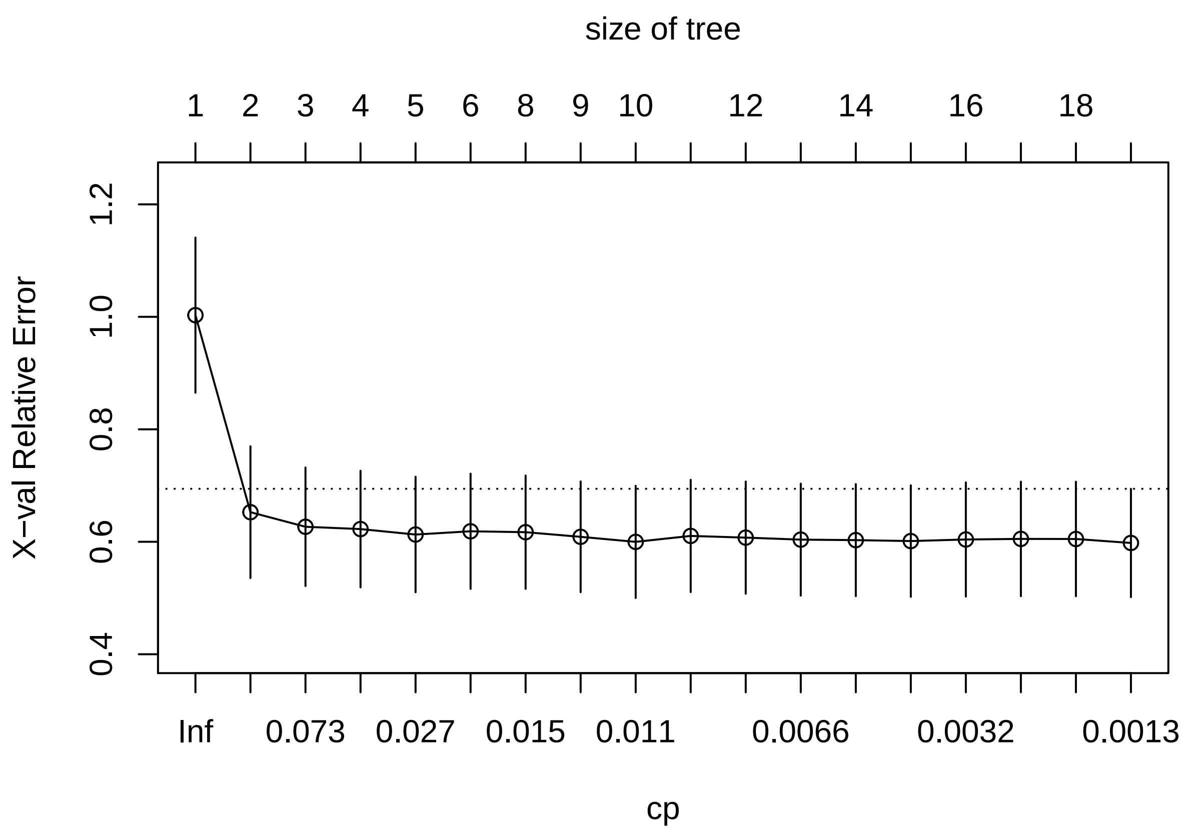
```
## -Buddy Bell
##    1565.715
```

```r
printcp(res.rp)
```

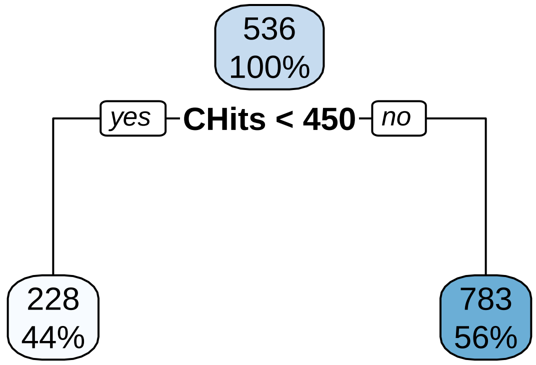
```
##
## Regression tree:
## rpart(formula = Salary ~ ., data = d, cp = 0.001)
##
## Variables actually used in tree construction:
##  [1] AtBat   CAtBat  CHits   CHmRun  CRBI    CRuns   CWalks  Hits
##  [9] PutOuts RBI     Walks   Years
##
## Root node error: 53319113/263 = 202734
##
## n=263 (59 observations deleted due to missingness)
##
##           CP nsplit rel error  xerror     xstd
## 1  0.3751526      0   1.00000 1.00296 0.138011
## 2  0.1202660      1   0.62485 0.65261 0.117220
## 3  0.0447760      2   0.50458 0.62667 0.105404
## 4  0.0395069      3   0.45981 0.62259 0.103688
## 5  0.0189058      4   0.42030 0.61289 0.102734
```

```
## 6  0.0156460      5   0.40139 0.61865 0.102423
## 7  0.0141210      7   0.37010 0.61703 0.100734
## 8  0.0140507      8   0.35598 0.60881 0.098552
## 9  0.0090608      9   0.34193 0.59981 0.099742
## 10 0.0087486     10   0.33287 0.61039 0.099877
## 11 0.0070271     11   0.32412 0.60739 0.099788
## 12 0.0061551     12   0.31709 0.60381 0.099624
## 13 0.0045893     13   0.31094 0.60293 0.099682
## 14 0.0034490     14   0.30635 0.60124 0.099191
## 15 0.0029108     15   0.30290 0.60407 0.101550
## 16 0.0028538     16   0.29999 0.60509 0.101883
## 17 0.0017889     17   0.29713 0.60494 0.101853
## 18 0.0010000     18   0.29535 0.59787 0.096345
```

```r
plotcp(res.rp) #Anything below the line is statistically indifferent. I make a choice in favor of a sim
```




```r
res.pruned <- prune(res.rp,cp=.2)
rpart.plot(res.pruned)
```




```
predict(res.pruned,newdata=d[17,])
```

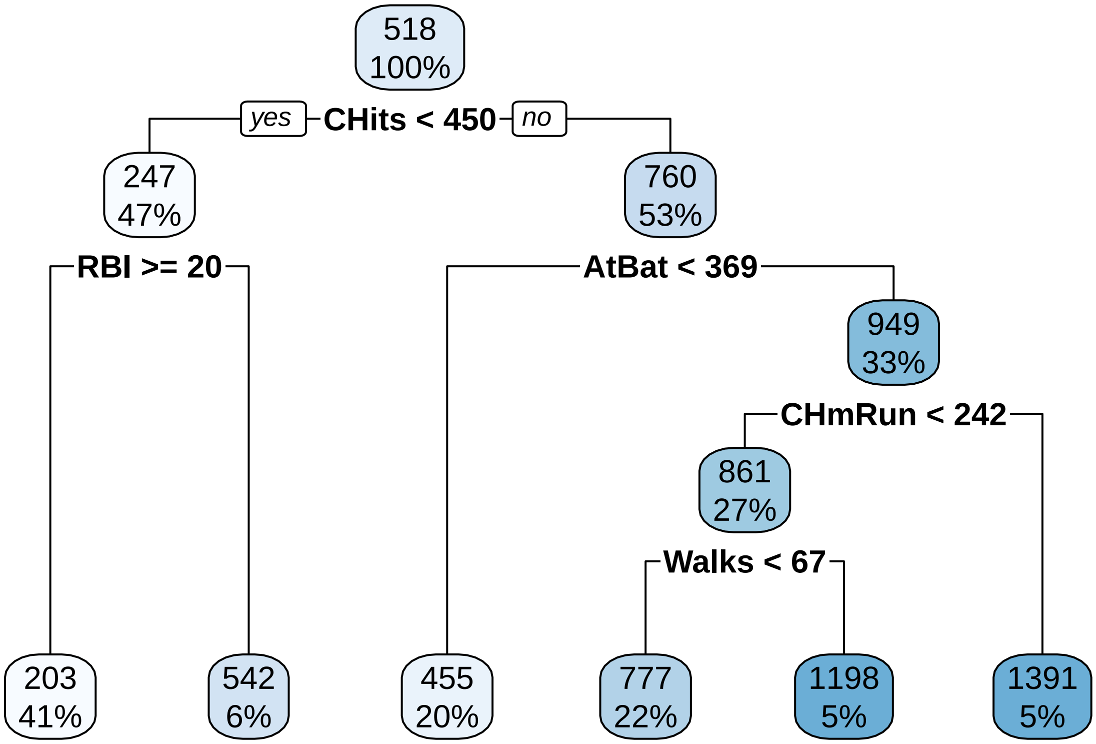
```
## -Buddy Bell
##    782.8048
```

```
#Use MSE to check goodness of fit. This is a nonparametric model, I can just use predictions. Pick the
```

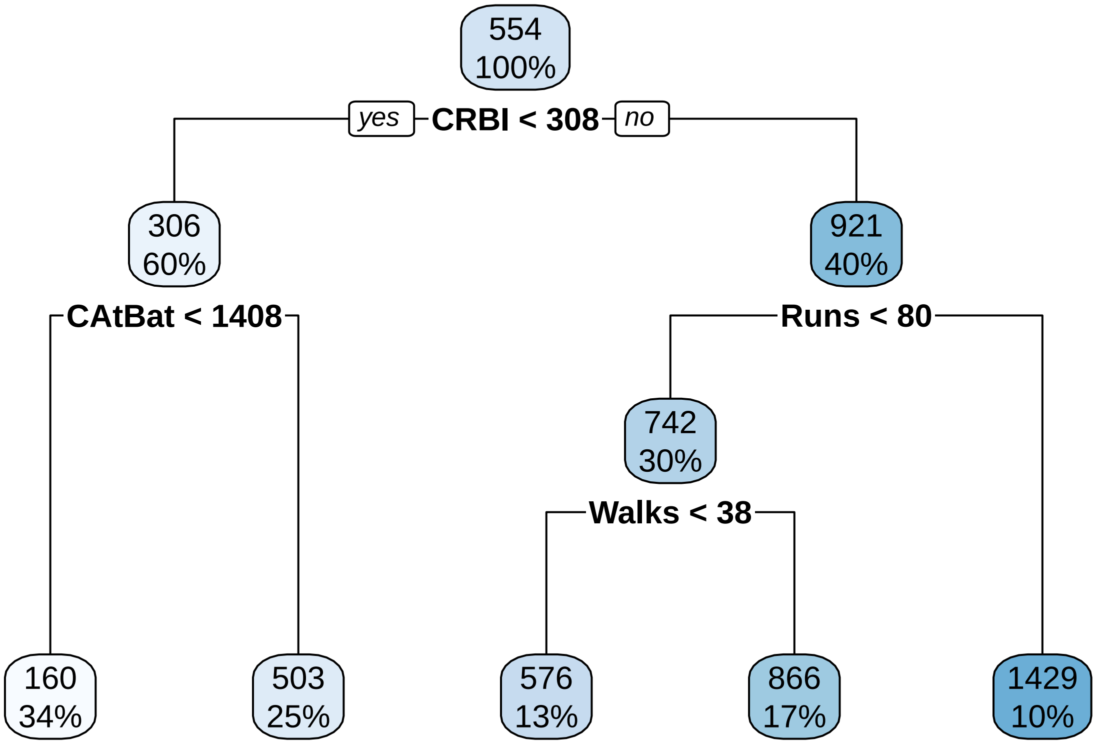
```
set.seed(1)

id <- sample(c(FALSE,TRUE),nrow(d),rep=TRUE)

table(id)
```

```
## id
## FALSE  TRUE
##   162   160
```

```
d1 <- d[id,]
d2 <- d[!id,]
res.rp1 <- rpart(Salary~.,d1,cp=0.02)
res.rp2 <- rpart(Salary~.,d2,cp=0.02)

rpart.plot(res.rp1)
```




```
rpart.plot(res.rp2)
```




```r
set.seed(3)
s <- sample(1:40,rep=TRUE)
table(s) #T1 and T2 are not independent
```

```
## s
##  2  3  4  5  6  8  9 10 11 12 15 16 18 19 20 22 23 25 29 31 33 36 37 39 40
##  1  1  1  2  1  3  2  3  1  2  1  1  1  1  1  1  1  1  2  1  1  1  3  1  6
```

Bagging

```r
res.rf <- randomForest(Salary~.,na.omit(d),mtry=ncol(d)-1)

res.rf
```

```
##
## Call:
##  randomForest(formula = Salary ~ ., data = na.omit(d), mtry = ncol(d) -      1)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 19
##
##           Mean of squared residuals: 79865.04
##                     % Var explained: 60.61
```

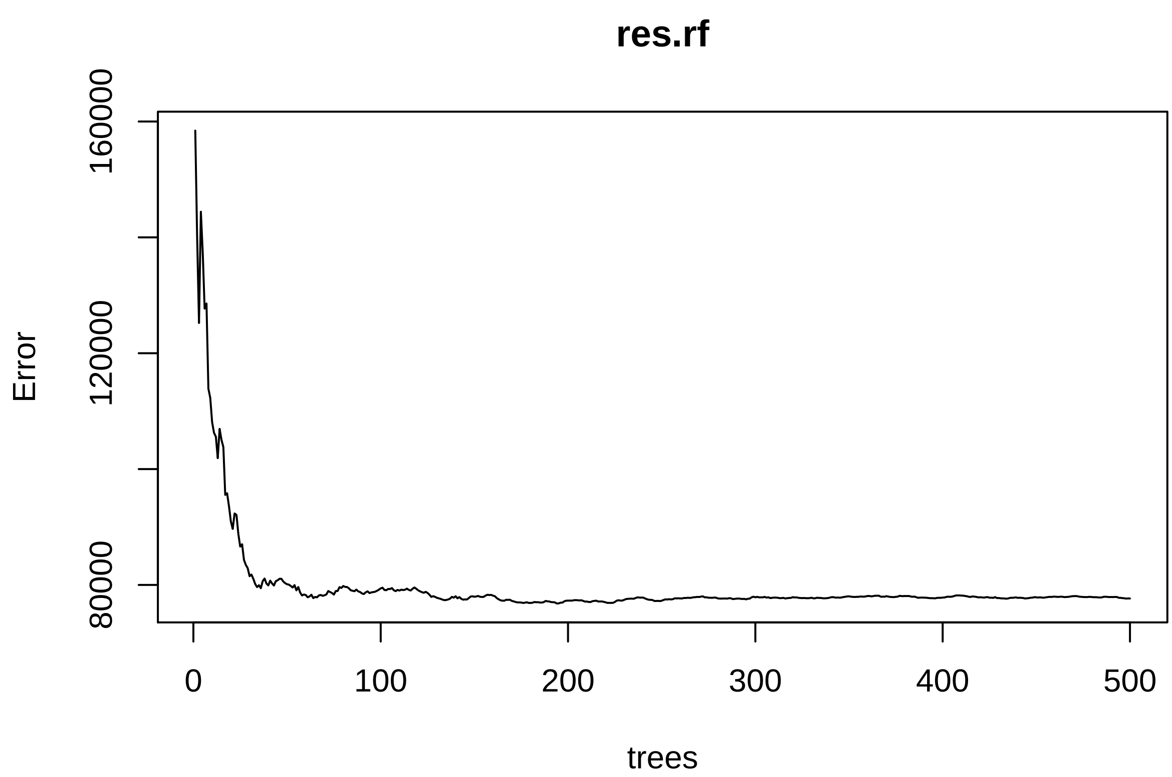
RandomForest

```r
res.rf <- randomForest(Salary~.,na.omit(d))
res.rf
```

```
##
## Call:
##  randomForest(formula = Salary ~ ., data = na.omit(d))
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 6
##
##           Mean of squared residuals: 77656.8
##                     % Var explained: 61.7
```

```
#Choose the lowest MSE
```

```
plot(res.rf) #Error after so many trees. Error sharply decreases.
```




```
res.rf <- randomForest(Salary~.,na.omit(d),importance=TRUE) #Importance of a variable calculated. For e
res.rf
```

```
##
## Call:
##  randomForest(formula = Salary ~ ., data = na.omit(d), importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 6
##
##           Mean of squared residuals: 78050.06
##                     % Var explained: 61.5
```

```r
importance(res.rf,type=1) #first column
```

```
##              %IncMSE
## AtBat      8.3053494
## Hits       6.4552783
## HmRun      6.0570968
## Runs       6.6815317
## RBI        6.1764620
## Walks      5.2095354
## Years      7.0732857
## CAtBat    14.1930385
## CHits     13.2423782
## CHmRun     7.8666030
## CRuns     10.5996034
## CRBI      11.1924379
## CWalks     6.9982971
## League    -1.0493647
## Division   0.7546720
## PutOuts    4.4171769
## Assists    0.1772433
## Errors     1.7773874
## NewLeague  1.3284999
```

```r
varImpPlot(res.rf,type=1)
```