MSCI 446 Assignment 3 - Question 2

K. Morris, H. Gomaa, M. Harper

29/03/2021

```
library('tidyverse')
library('caret')
library('ggplot2')
library('gridExtra')
library('AmesHousing')
library('plotly')
library('ISLR')
library('glmnet')
library('leaps')

library('tree')
library('rpart')
library('rattle')
library('randomForest')
library('e1071')
```

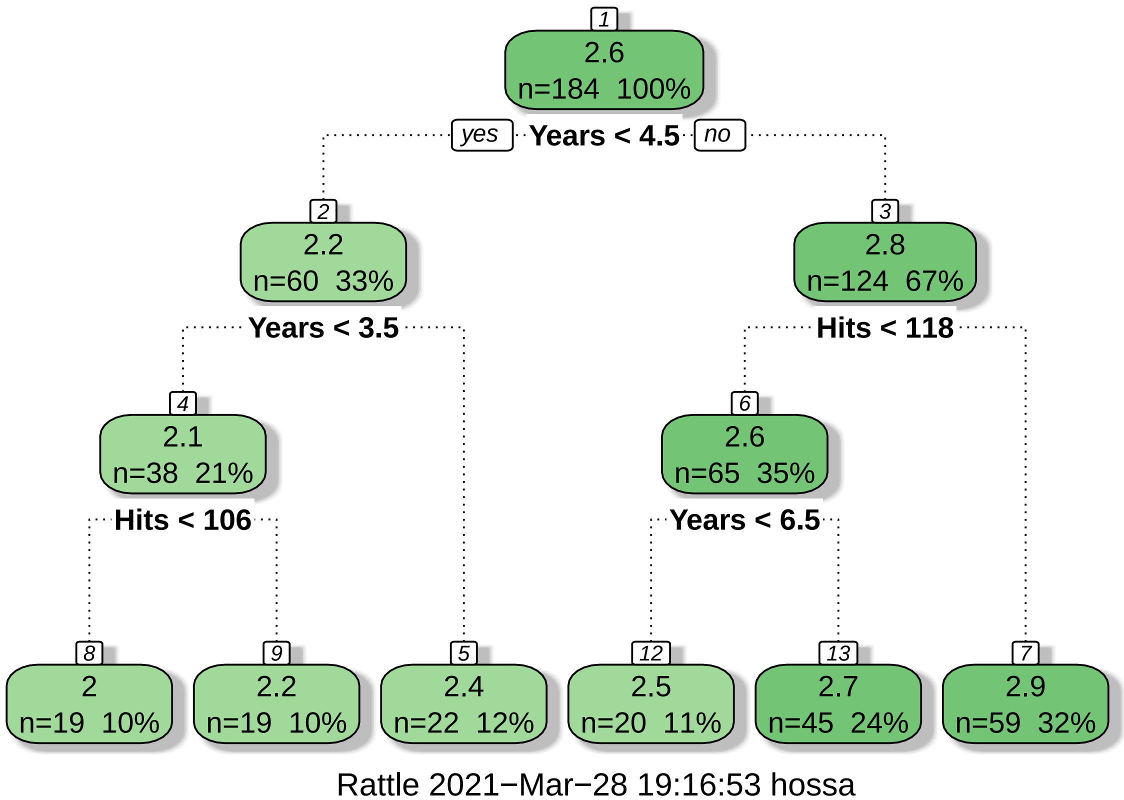
```
theme_set(theme_minimal())
```

2.1 Preprocess data:

```
datHeart <- read.csv('Heart.csv')[,-1]
datHeart <- datHeart[complete.cases(datHeart),]
datHeart$AHD <- as.factor(datHeart$AHD)

# clean up NA data
datHitters = Hitters
datHitters$Salary = log(datHitters$Salary, 10)
datHitters <- datHitters[complete.cases(datHitters),]

set.seed(112)
hitters.train_inds <- sample(1:nrow(datHitters), size = floor(0.7*nrow(datHitters)))
hitters.train <- datHitters[hitters.train_inds,]
hitters.test  <- datHitters[-hitters.train_inds,]

# set seed again for next split
#set.seed(112)
heart.train_inds <- sample(1:nrow(datHeart), size = floor(0.7*nrow(datHeart)))
heart.train <- datHeart[heart.train_inds,]
heart.test  <- datHeart[-heart.train_inds,]
```

2.2 Decision Trees for Regresssion

```r
# 2.2.1
hitters.tree <- rpart(Salary ~ Hits + Years, data = hitters.train)

# 2.2.2
fancyRpartPlot(hitters.tree)
```




```r
#2.2.3 and 2.2.4
hitters.preds <- predict(hitters.tree, newdata = hitters.test)
hitters.errs <- hitters.test$Salary - hitters.preds
hitters.sq.errs <- hitters.errs^2
hitters.SSE <- sum(hitters.sq.errs)
hitters.SSE
```

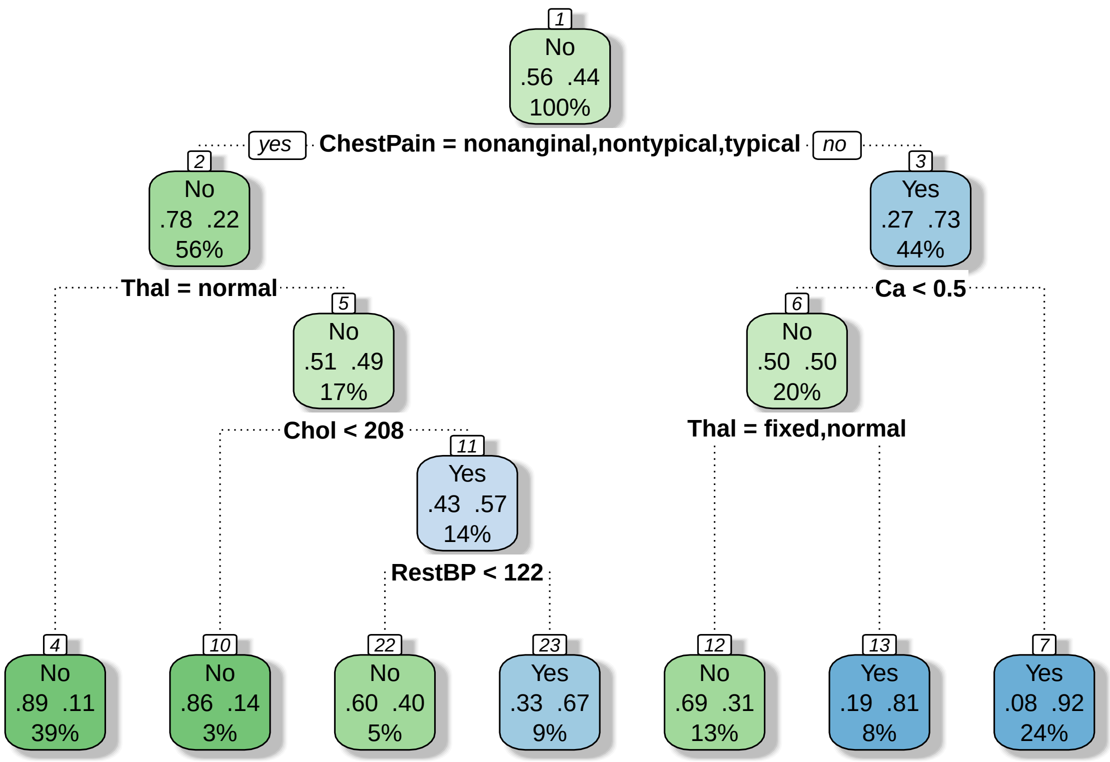
```
## [1] 6.216793
```

2.2.3: Following the above tree, the logarithmic salary value for a player with 6 seasons with 125 hits is 2.9. Converting to thousands of dollars: $10^{2.9} = 794.328$. Therefore, salary is expected to be $794,328.

2.2.4: total sum squared error (SSE) for every hitter salary in the test set: SSE = 6.216793

2.3 Decision Trees for Classification

```
# 2.3.1
heart.tree <- rpart(AHD ~ ., data = heart.train)

# 2.3.2
fancyRpartPlot(heart.tree)
```



Rattle 2021−Mar−28 19:16:53 hossa

```
#2.3.3
preds.heart <- predict(heart.tree, newdata = heart.test, type='class')

first5hearts <- heart.test[1:5, 'AHD']
first5preds <- preds.heart[1:5]
first5acc <- mean(first5preds==first5hearts)*100

cat("2.3.3:\nFirst 5 heart accuracy =", first5acc, "%\n")
```

```
## 2.3.3:
## First 5 heart accuracy = 80 %
```

```
#2.3.4
cat("2.3.4: Confusion Matrix\n")
```

```
## 2.3.4: Confusion Matrix
```

```r
confusionMatrix(preds.heart, heart.test$AHD)$table
```

```
##           Reference
## Prediction No Yes
##        No  40   6
##        Yes  5  39
```

2.4 Bagging: Regression

```r
hitters.bagging <- randomForest(Salary~., data=hitters.train, mtry=ncol(hitters.train)-1)
log.preds <- predict(hitters.bagging, newdata = hitters.test)
dollar.preds <- 10^(log.preds)
first4 <- dollar.preds[1:4]
first4
```

```
##     -Andre Dawson -Andres Galarraga    -Andres Thomas     -Alex Trevino
##          817.1282          101.9939          101.9691          412.4586
```

```r
errs <- hitters.test$Salary - log.preds
errs.squared <- errs^2
SSE <- sum(errs.squared)
SSE
```

```
## [1] 4.744524
```

The SSE from the standard regression tree in **2.2** was calculated to be 6.217. The bagging method shows a significant improvement to an SSE of 4.745.

2.5 Bagging: Classification

```r
#2.5.1: Na values handled in 2.2

#2.5.2
heart.bagging <- randomForest(AHD~., data=heart.train, mtry=ncol(heart.train)-1)
preds <- predict(heart.bagging, newdata = heart.test, type='class')

#2.5.3
first4preds <- preds[1:4]
first5preds <- preds[1:5]
first4preds
```

```
##   1   4   6  10
##  No Yes  No Yes
## Levels: No Yes
```

```
#2.5.4
confusionMatrix(preds, heart.test$AHD)$table
```

```
##           Reference
## Prediction No Yes
##        No  39   7
##        Yes  6  38
```

```
first4hearts <- heart.test[1:4, 'AHD']
first4acc <- mean(first4preds==first4hearts)*100

cat("2.5.5:\nFirst 4 (bagged) heart accuracy =", first4acc, "%\n")
```

```
## 2.5.5:
## First 4 (bagged) heart accuracy = 75 %
```

```
first5hearts <- heart.test[1:5, 'AHD']
first4acc <- mean(first5preds==first5hearts)*100

cat("2.5.5:\nFirst 5 (bagged) heart accuracy =", first4acc, "%\n")
```

```
## 2.5.5:
## First 5 (bagged) heart accuracy = 60 %
```

The accuracy has appeared to drop 5% from before bagging. Also a smaller sample size used (4 vs 5), so not really a fair comparison.

Upon examining the first 5 results, a direct comparison can be made. Here, we observe that classification accuracy is 20% lower with bagging.

2.6 Random Forest: Regression

```
# Na values handled in 2.2

hitters.rf <- randomForest(Salary ~., data=hitters.train, mtry = (ncol(hitters.train)-1)^0.5)
preds.rf <- predict(hitters.rf, newdata = hitters.test)

first4preds.log <- preds.rf[1:4]
first4preds.dollars <- 10^(first4preds.log)
first4preds.dollars
```

```
##     -Andre Dawson -Andres Galarraga    -Andres Thomas     -Alex Trevino
##          831.6390          104.4547          102.4247          395.5203
```

```
errs.rf <- hitters.test$Salary - preds.rf
squared.errs.rf <- errs.rf^2
SSE.rf <- sum(squared.errs.rf)
SSE.rf
```

```
## [1] 4.388885
```

The random forest model above is essentially identical to the bagging one from before, however, the *mtry* argument in the randomForest function is set to the square root of the number of predictors, whereas *mtry* for the bagging model uses all the predictors at each step of tree growing. This essentially de-correlates the numerous trees that are grown, in theory making for a less biased prediction.

The SSE from the bagging and randomForest models were 4.745 and 4.389 respectively. It appears as though the random forest model has a slightly lower SSE for this particular model and R-seed.