HW4_part1

Will Scheib

11/5/2021

5

```
library(tree)
library(randomForest)
library(tidyverse)
students <- read.table("data/students.txt", header=T)
students$Gender <- factor(students$Gender)
students$Smoke <- factor(students$Smoke)
students$Marijuan <- factor(students$Marijuan)
students$DrivDrnk <- factor(students$DrivDrnk)
students <- students %>% select(-Student)
```

a

```
set.seed(2013)
sample.data<-sample.int(nrow(students), floor(.50*nrow(students)), replace = F)
train<-students[sample.data, ]
test<-students[-sample.data, ]
```

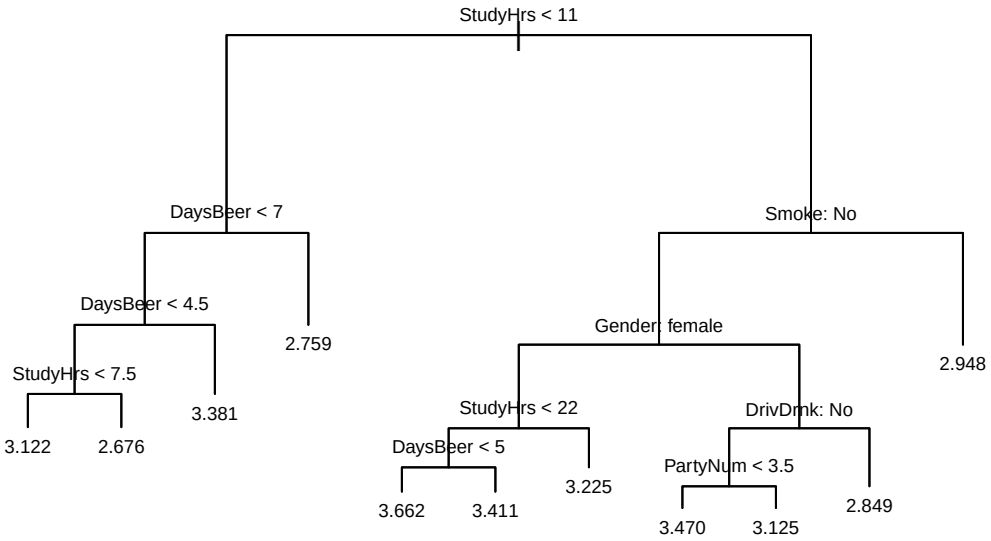
b

```
ols <- lm(GPA~., data=train)
ols.pred <- predict(ols, test)
ols.mse <- mean((ols.pred-test$GPA)^2)
ols.mse
```

```
## [1] 0.1962592
```

c

```
tree.class.train<-tree::tree(GPA~., data=train)
plot(tree.class.train)
text(tree.class.train, cex=0.6, pretty=0)
```




```
summary(tree.class.train)$size
```

```
## [1] 11
```

d

```
tree.pred <- predict(tree.class.train, test)
tree.mse <- mean((tree.pred-test$GPA)^2)
tree.mse
```

```
## [1] 0.3057565
```

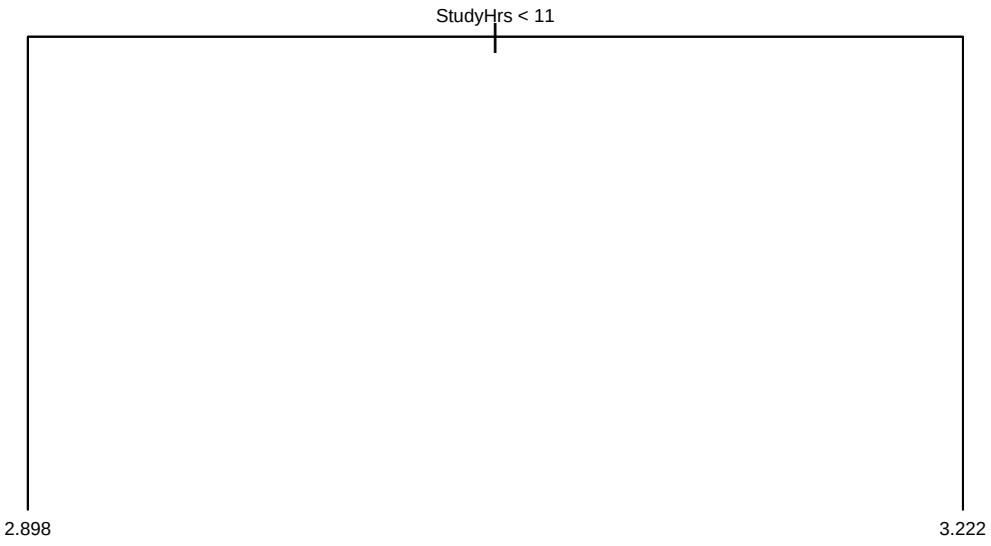
e

```
set.seed(1)
cv.class<-tree::cv.tree(tree.class.train, K=10, FUN=prune.tree)
trees.num.class<-cv.class$size[which.min(cv.class$dev)]
trees.num.class
```

```
## [1] 2
```

```
prune.class<-tree::prune.tree(tree.class.train, best=trees.num.class)

plot(prune.class)
text(prune.class, cex=0.6, pretty=0)
```



The predicted GPA for students who study fewer than 11 hours studying per week is 2.898, and for students who study 11 hours per week or more is 3.222.

g

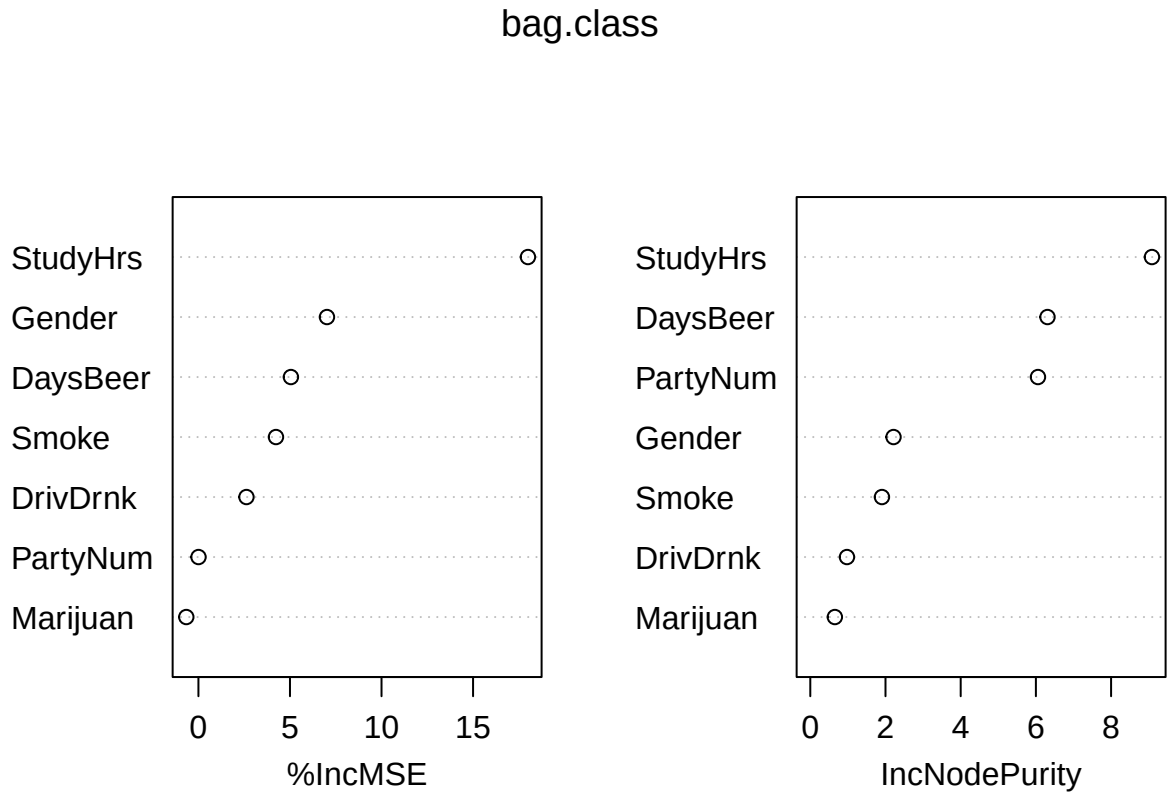
```
pruned.pred <- predict(prune.class, test)
pruned.mse <- mean((pruned.pred-test$GPA)^2)
pruned.mse
```

```
## [1] 0.2170533
```

h

```
set.seed(2)
bag.class<-randomForest::randomForest(GPA~., data=train, mtry=7, importance=TRUE)

randomForest::varImpPlot(bag.class)
```




```
bagging.pred <- predict(bag.class, test)
bagging.mse <- mean((bagging.pred-test$GPA)^2)
bagging.mse
```

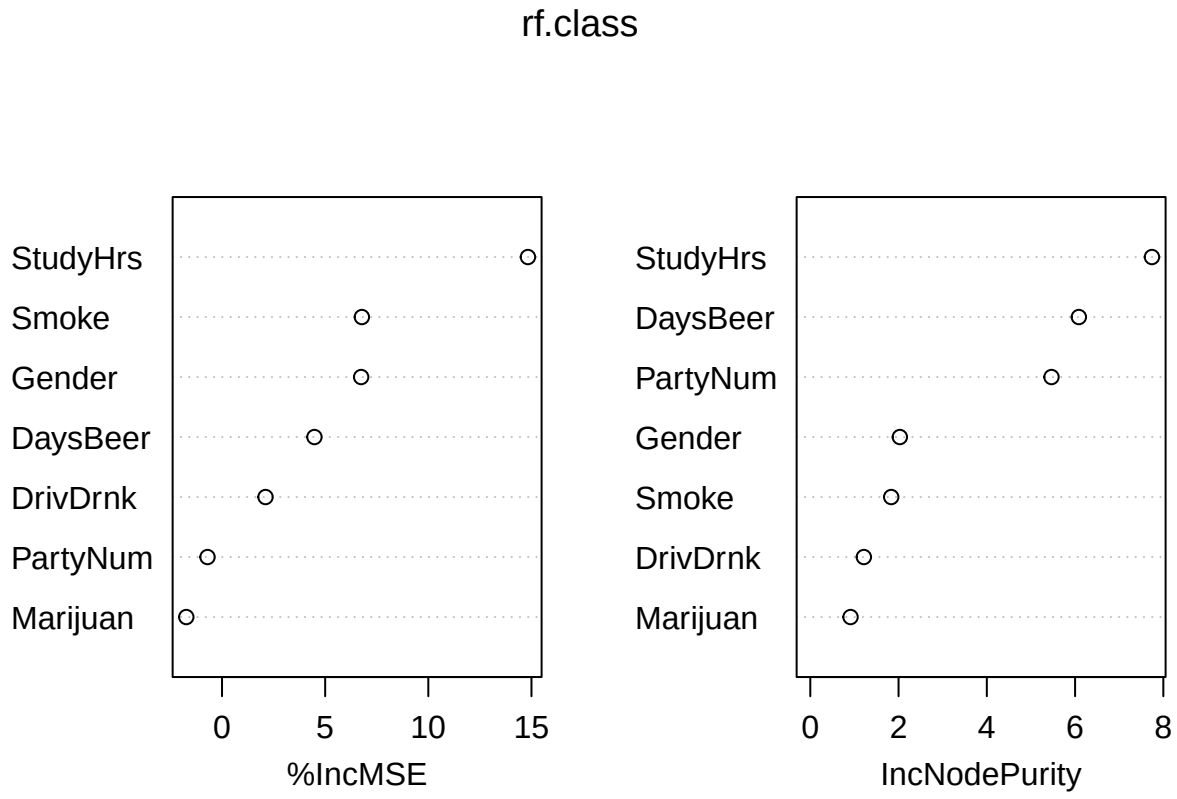
```
## [1] 0.2671621
```

StudyHrs is easily the most important predictor of GPA.

i

```
set.seed(2)
rf.class<-randomForest::randomForest(GPA~., data=train, mtry=3, importance=TRUE)

randomForest::varImpPlot(rf.class)
```




```
rf.pred <- predict(rf.class, test)
rf.mse <- mean((rf.pred-test$GPA)^2)
rf.mse
```

```
## [1] 0.2389143
```

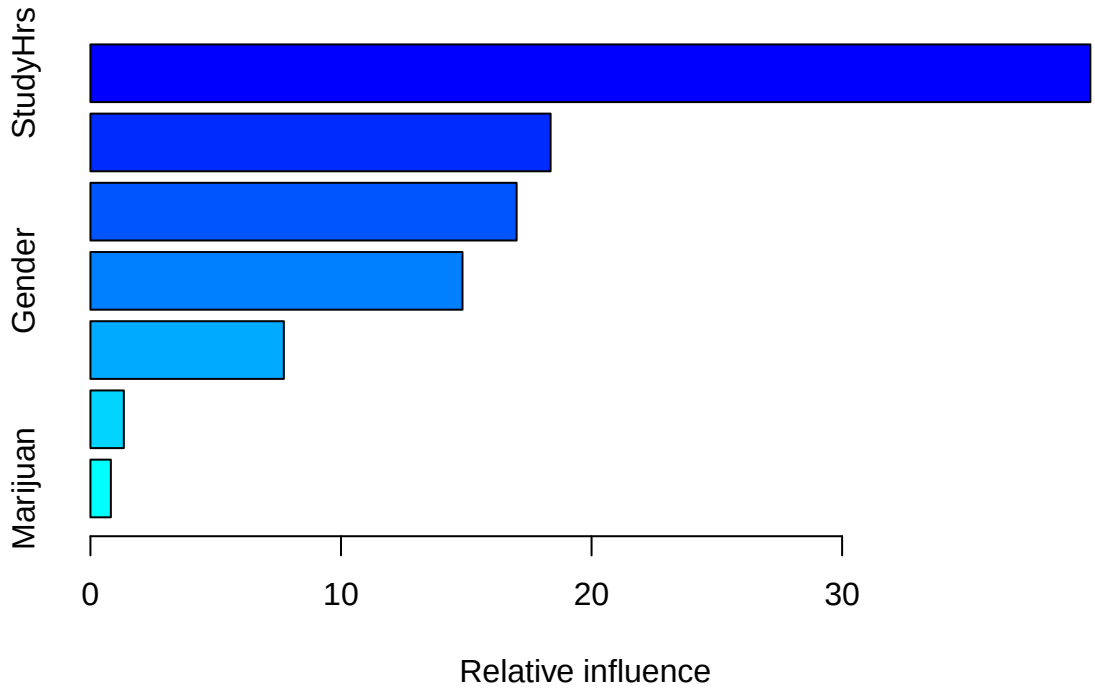
StudyHrs is easily the most important predictor of GPA.

j

```
set.seed(2)
boost.class<-gbm::gbm(GPA~., data=train, shrinkage=0.0001, n.trees=5000, interaction.depth=1)
```

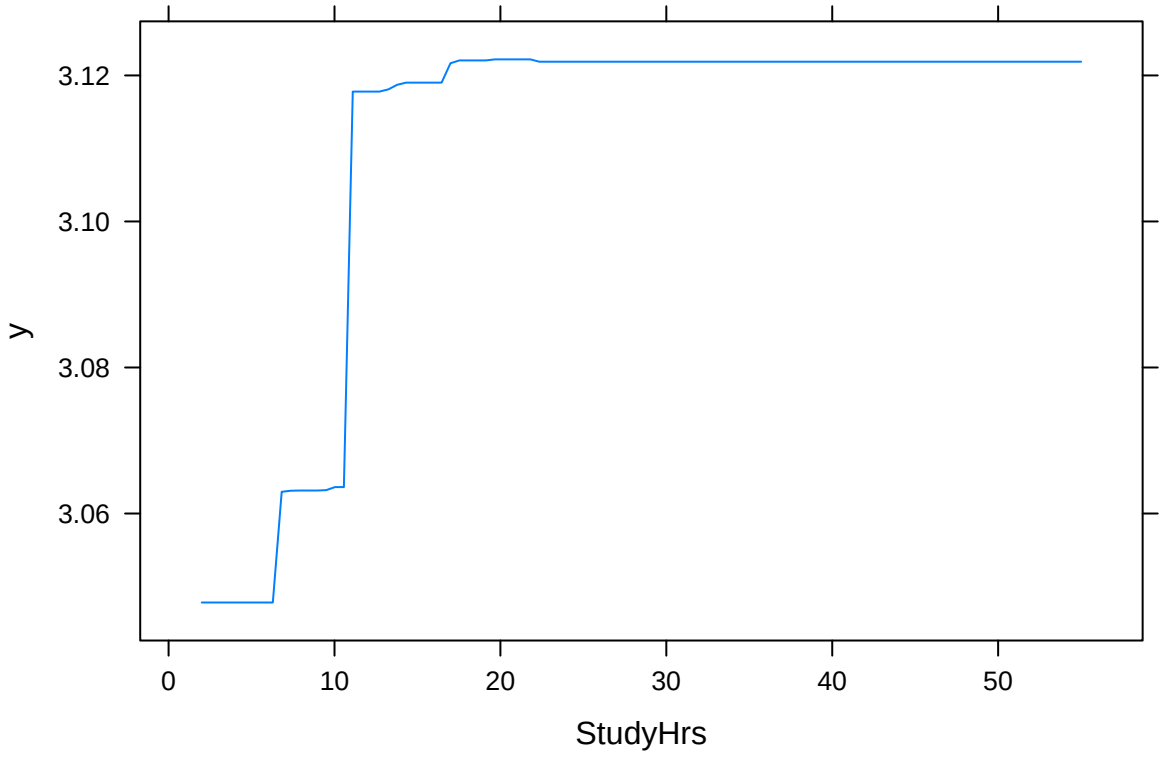
```
## Distribution not specified, assuming gaussian ...
```

```
summary(boost.class)
```




```
##                var   rel.inf
## StudyHrs StudyHrs 39.907643
## DaysBeer DaysBeer 18.365373
## Smoke       Smoke 17.009407
## Gender     Gender 14.849067
## PartyNum PartyNum  7.718825
## DrivDrnk DrivDrnk  1.333953
## Marijuan Marijuan  0.815733
```

```
plot(boost.class,i="StudyHrs")
```




```
boost.pred<-predict(boost.class, newdata=test, n.trees=5000, type = "response")
boost.mse <- mean((boost.pred-test$GPA)^2)
boost.mse
```

```
## [1] 0.202228
```

StudyHrs is easily the most important predictor of GPA.

k

```
c(
  ols.mse=ols.mse,
  tree.mse=tree.mse,
  pruned.mse=pruned.mse,
  bagging.mse=bagging.mse,
  rf.mse=rf.mse,
  boost.mse=boost.mse
)
```

```
##     ols.mse    tree.mse  pruned.mse bagging.mse      rf.mse   boost.mse
##   0.1962592   0.3057565   0.2170533   0.2671621   0.2389143   0.2022280
```

OLS had the lowest test MSE.

l

```
summary(ols)$coefficients
```

```
##                 Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)  3.234029644 0.123313763 26.2260235 3.838411e-49
## Gendermale  -0.223776957 0.092459050 -2.4202818 1.714618e-02
## SmokeYes    -0.295272306 0.116809214 -2.5278169 1.289747e-02
## MarijuanYes  0.068901538 0.115100719  0.5986195 5.506573e-01
## DrivDrnkYes  0.024631912 0.104498794  0.2357148 8.140923e-01
## PartyNum    -0.002101670 0.013203013 -0.1591811 8.738181e-01
## DaysBeer    -0.014080480 0.011730006 -1.2003814 2.325700e-01
## StudyHrs     0.007214838 0.004835235  1.4921381 1.385254e-01
```

StudyHrs was clearly the best predictor according to all of the tree models, but OLS gives a very different result. It doesn't see StudyHrs as the most important predictor at all.