hw4_4630

Akhil Havaldar

11/4/2022

Question 5

```
Data <- read.table("students2.txt", header=TRUE)

Data <- Data[,-1]

# Factors
Data$Gender <-factor(Data$Gender)
Data$Smoke <- factor(Data$Smoke)
Data$Marijuan <- factor(Data$Marijuan)
Data$DrivDrnk <- factor(Data$DrivDrnk)
```

a)

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

b)

```
result <- lm(GPA~., data=train)
summary(result)
```

```
##
## Call:
## lm(formula = GPA ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.26408 -0.28708  0.01616  0.35840  0.91389
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.234030   0.123314  26.226   <2e-16 ***
```

```
## Gendermale  -0.223777   0.092459  -2.420   0.0171 *
## SmokeYes    -0.295272   0.116809  -2.528   0.0129 *
## MarijuanYes  0.068902   0.115101   0.599   0.5507
## DrivDrnkYes  0.024632   0.104499   0.236   0.8141
## PartyNum    -0.002102   0.013203  -0.159   0.8738
## DaysBeer    -0.014080   0.011730  -1.200   0.2326
## StudyHrs     0.007215   0.004835   1.492   0.1385
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4776 on 110 degrees of freedom
## Multiple R-squared:  0.1787, Adjusted R-squared:  0.1265
## F-statistic:  3.42 on 7 and 110 DF,  p-value: 0.002404
```

```
y.test <- test[,"GPA"]
yhat.ols <- predict(result, newdata = test)
mse.ols <- mean((y.test - yhat.ols)^2)
mse.ols
```

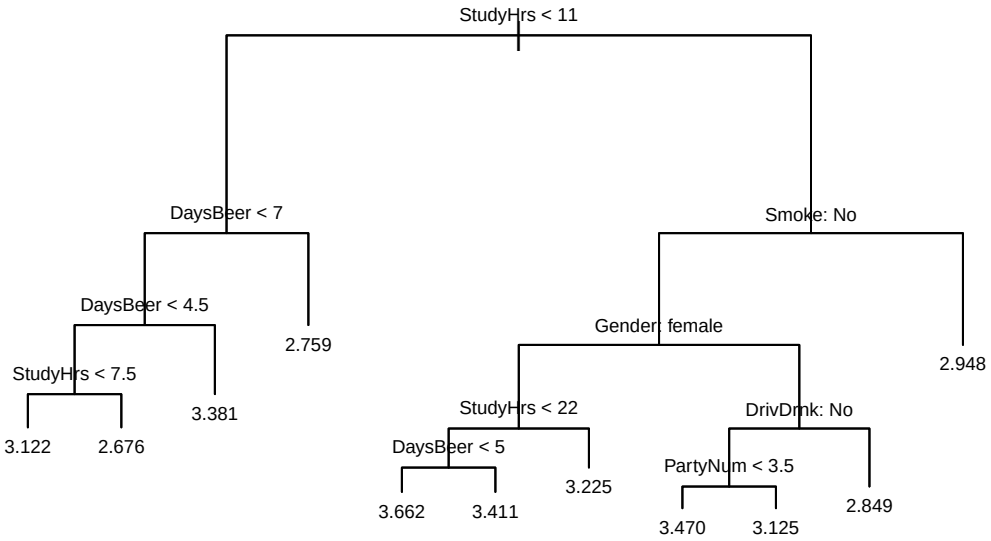
```
## [1] 0.1962592
```

c)

```
tree.result<-tree::tree(GPA~., data=train)
summary(tree.result)
```

```
##
## Regression tree:
## tree::tree(formula = GPA ~ ., data = train)
## Variables actually used in tree construction:
## [1] "StudyHrs" "DaysBeer" "Smoke"    "Gender"   "DrivDrnk" "PartyNum"
## Number of terminal nodes:  11
## Residual mean deviance:  0.1833 = 19.61 / 107
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -1.05500 -0.26630  0.05151  0.00000  0.26920  0.87110
```

```
plot(tree.result)
text(tree.result, cex=0.6, pretty=0)
```



- The tree has 11 terminal nodes.

d)

```
yhat <- predict(tree.result, newdata=test)
mse <- mean((y.test-yhat)^2)
mse
```

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

- Recursive binary test MSE is 0.3057565

e)

```
set.seed(1)
cv.gpa <- tree::cv.tree(tree.result, K=10)
cv.gpa
```

```
## $size
##  [1] 11 10  9  8  7  6  5  4  3  2  1
##
## $dev
```

```
##  [1] 35.07023 35.08824 35.28836 34.77547 33.74239 32.72005 32.87535 31.99070
##  [9] 31.48925 30.45872 31.27675
##
## $k
##  [1]      -Inf 0.3246136 0.3646147 0.4972900 0.5846633 0.8686436 1.0336225
##  [8] 1.2512398 1.3761691 1.6723170 2.9601458
##
## $method
## [1] "deviance"
##
## attr(,"class")
## [1] "prune"         "tree.sequence"
```

```
trees.num <- cv.gpa$size[which.min(cv.gpa$dev)]
trees.num
```

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

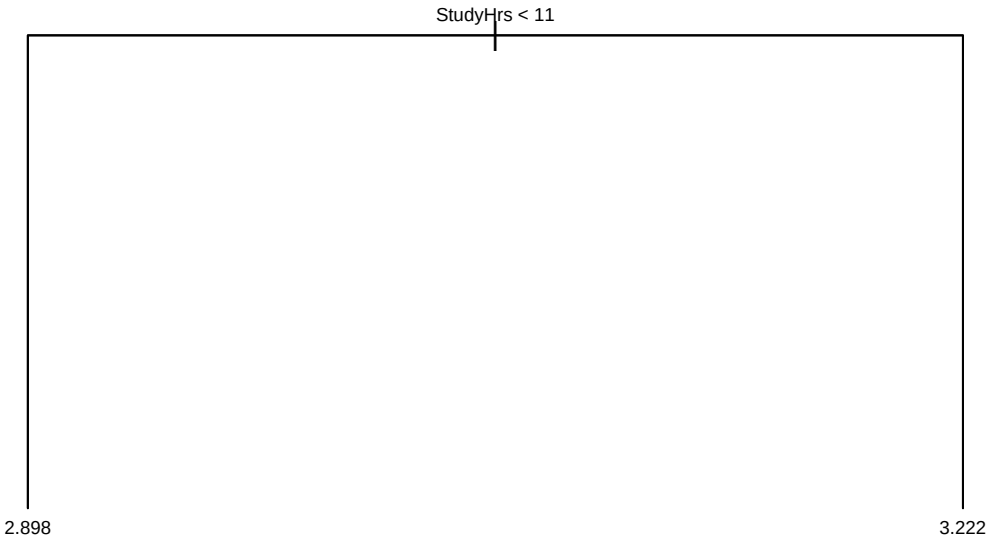
```
prune<-tree::prune.tree(tree.result, best=trees.num)
prune
```

```
## node), split, n, deviance, yval
##       * denotes terminal node
##
## 1) root 118 30.55 3.096
##   2) StudyHrs < 11 46 10.59 2.898 *
##   3) StudyHrs > 11 72 16.99 3.222 *
```

- With pruning, 2 terminal nodes gives us the smallest deviance.

f)

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



- StudyHrs is the most important predictor of GPA. Students who study less than 11 hours a week are predicted to have a GPA of 2.898. Students who study more than 11 hours a week are predicted to have a GPA of 3.222.

g)

```
yhat.prune <- predict(prune, newdata=test)
mse.prune <- mean((y.test-yhat.prune)^2)
mse.prune
```

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

- Pruned test MSE is 0.2170533.

h)

```
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.1.3
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
set.seed(2)

bag.gpa<-randomForest(GPA~., data=train, mtry=7, importance=TRUE)
bag.gpa
```

```
##
## Call:
##  randomForest(formula = GPA ~ ., data = train, mtry = 7, importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 7
##
##           Mean of squared residuals: 0.2395268
##                     % Var explained: 7.47
```

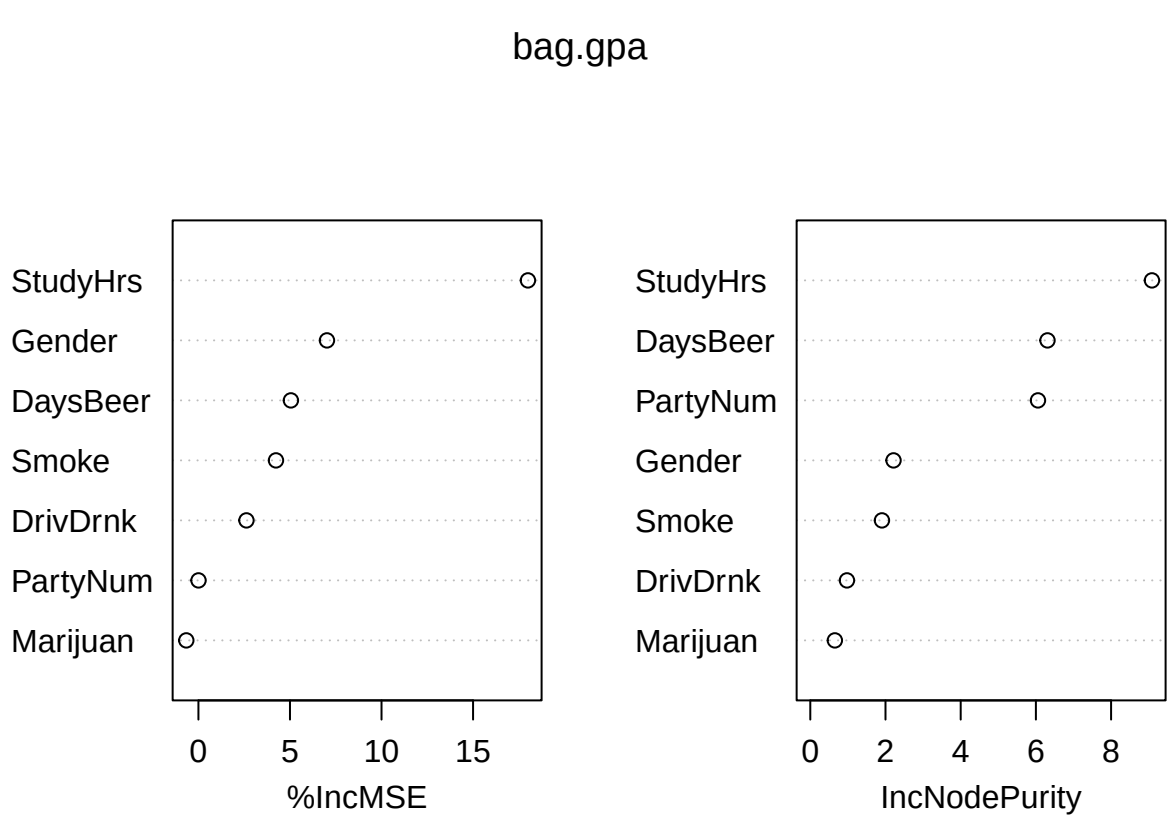
```
# test mse
yhat.bag <- predict(bag.gpa, newdata=test)
mse.bag <- mean((y.test-yhat.bag)^2)
mse.bag
```

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

```
# importance
randomForest::importance(bag.gpa)
```

```
##               %IncMSE IncNodePurity
## Gender    7.021042504     2.2128855
## Smoke     4.235681594     1.9051378
## Marijuan -0.661512657     0.6529114
## DrivDrnk  2.623311201     0.9747505
## PartyNum  0.005320738     6.0548493
## DaysBeer  5.050952109     6.3082402
## StudyHrs 17.996637110     9.0859584
```

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



- Test MSE with bagging is 0.2671621. StudyHrs is by far the most important variable as seen from the plot.

i)

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

```
##
## Call:
##  randomForest(formula = GPA ~ ., data = train, mtry = 3, importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##           Mean of squared residuals: 0.2466555
##                     % Var explained: 4.71
```

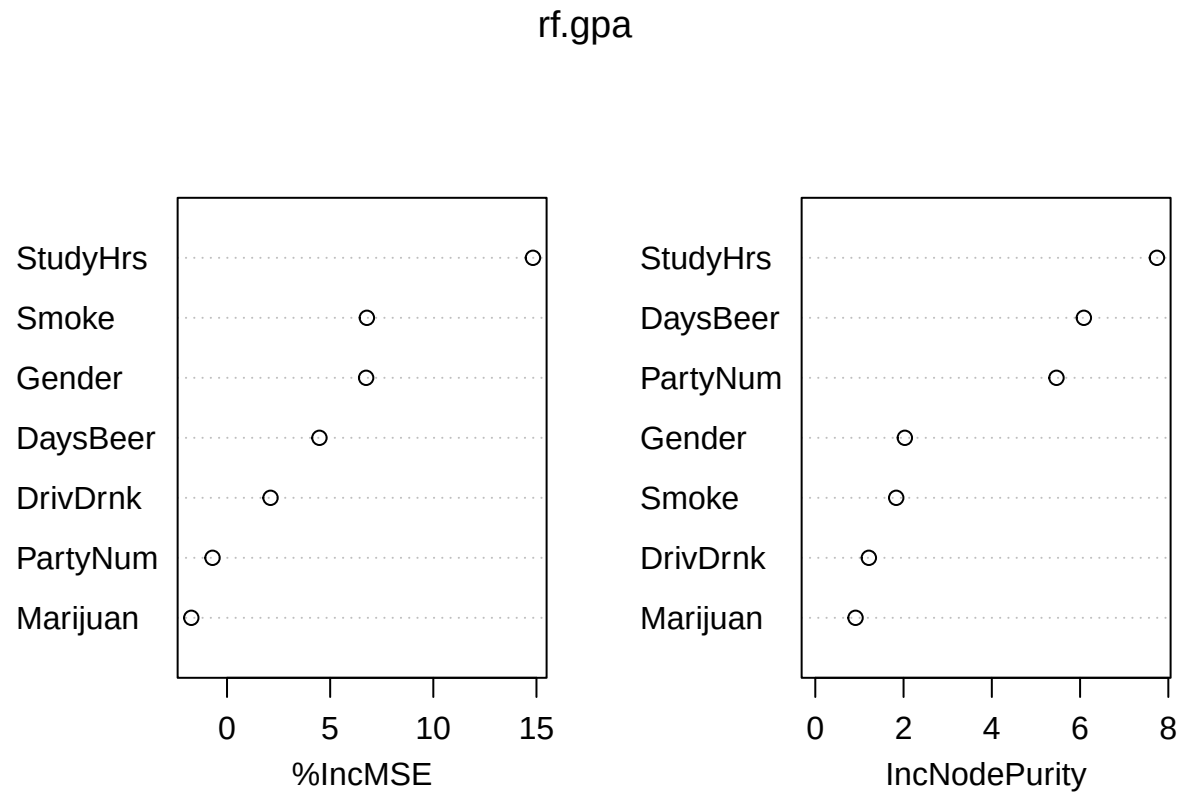
```
yhat.rf<-predict(rf.gpa, newdata=test)
mse.rf<-mean((y.test-yhat.rf)^2)
mse.rf
```

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

```
randomForest::importance(rf.gpa)
```

```
##             %IncMSE IncNodePurity
## Gender    6.7434457     2.0289152
## Smoke     6.7783763     1.8335276
## Marijuan -1.7311133     0.9118012
## DrivDrnk  2.1080169     1.2133161
## PartyNum -0.7022751     5.4652140
## DaysBeer  4.4793034     6.0856794
## StudyHrs 14.8282927     7.7414319
```

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



- The test MSE is 0.2389143 which is lower than the previous example with bagging. Studyhrs is still the most important variable as seen from the plot.

j)

```
library(gbm)
```

```
## Warning: package 'gbm' was built under R version 4.1.3
```

```
## Loaded gbm 2.1.8.1
```

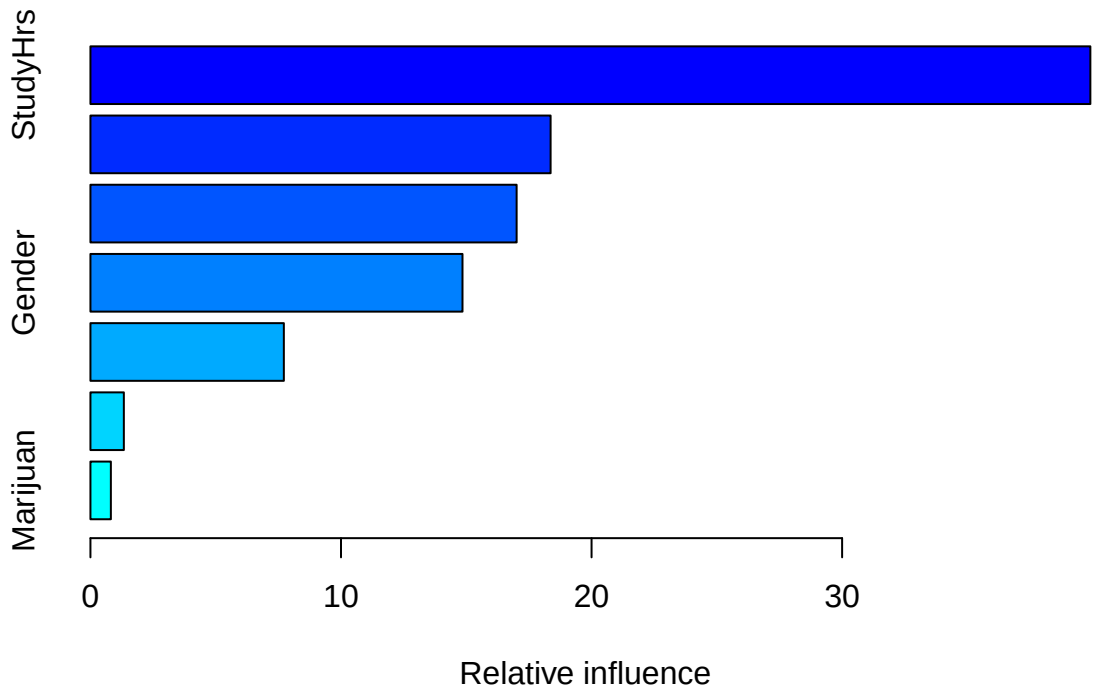
```
set.seed(2)
boost.gpa <- gbm(GPA~., data=train, distribution="gaussian", n.trees=5000,
                 interaction.depth=1, shrinkage=0.0001)

yhat.boost <- predict(boost.gpa, newdata=test, n.trees=5000, interaction.depth=1,
                      shrinkage=0.0001)

mse.boost<-mean((yhat.boost-y.test)^2)
mse.boost
```

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

```
summary(boost.gpa)
```




```
##                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
```

- Test MSE with boosting is 0.202228. This is lower than both previous examples with bagging and random forest. StudyHrs is still by far the most important predictor as seen from the plot.

k)

```
c(mse.ols, mse, mse.prune, mse.bag, mse.boost)
```

```
## [1] 0.1962592 0.3057565 0.2170533 0.2671621 0.2022280
```

- OLS had the lowest test MSE with 0.1962592. For tree based methods, boosting had the lowest overall test mse with 0.2022280.

l)

- StudyHrs was by far the most important predictor variable among all tree based methods. However in OLS, StudyHrs was shown to be insignificant.

Question 1

- a) flexibility decreases
- b)
- c) test mse decreases initially then starts to increase
- d) var decreases
- e) bias increases
- f) remains constant

Question 2

- a) 50.44
- b) 58.86
- c) 52.80

Question 3

- football : 6
- basketball : 9
- Basketball will be predicted