Homework 07 Spring 2019

Daniel Smith

April 14, 2019

Homework 07 Spring 2019 - DUE April 17, 2019

```
library(knitr)
hook_output = knit_hooks$get('output')
knit_hooks$set(output = function(x, options) {
  # this hook is used only when the linewidth option is not NULL
  if (!is.null(n <- options$linewidth)) {
    x = knitr:::split_lines(x)
    # any lines wider than n should be wrapped
    if (any(nchar(x) > n)) x = strwrap(x, width = n)
    x = paste(x, collapse = '\n')
  }
  hook_output(x, options)
})
```

```
 #load everything needed
library(NHANES)
library(dplyr)
library(gmodels)
library(ROCR)
library(rpart)
library(partykit)
library(tidyverse)
library(RColorBrewer)
library(reshape)
library(plot3D)
library(parallel)
library(randomForestSRC)
library(ggRandomForests)
library(class)
library(mosaic)
library(mice)
```

Course Material to Review

Recall the NHANES dataset that we used in Lesson 12 on March 27, 2019, https://htmlpreview.github.io/?https://github.com/vhertzb/ml_supervised/blob/master/ML_supervised.html. And more on supervised learning on April 10, 2019, https://htmlpreview.github.io/?https://github.com/vhertzb/more-supervised-learning/blob/master/More_Supervised_Learning.html.

Also review the logistic regression examples in Homework 6 assignment, see https://htmlpreview.github.io/?https://github.com/melindahiggins2000/N741_Homework06_regression/blob/master/homework6.html.

Assignment

In the NHANES dataset there is a discrete variable called Depressed indicating whether each participant had “None”, “Several”, “Majority” or “AlmostAll” days in a month where the pariticpant felt down, depressed or hopeless. You are going to build a set of classifiers for this dependent variable. You may use any (set of) independent variable(s) you like except for the variable callsed DaysMentHlthBad (self-reported days that the participant’s mental health was not good out of 30 days).

Run this R code to get started and create 2 groups that either were depressed “None” versus more than “None” - the new variable is depressedYes.

```
# add depressedYes to NHANES dataset
NHANES <- NHANES %>%
  mutate(depressedYes = Depressed != "None")

# check recoding that "Several" and "Most"
# are coded as TRUE for depressedYes
# and "None" are coded FALSE for depressedYes
NHANES %>%
  select(Depressed, depressedYes) %>%
  with(table(Depressed, depressedYes))
```

```
##           depressedYes
## Depressed FALSE TRUE
##   None      5246    0
##   Several      0 1009
##   Most         0  418
```

PROBLEM 1: Run 4 classifier models for depressedYes:

- logistic regression A)Build the Classifier

```
#summarize the data set
summary(NHANES)
```

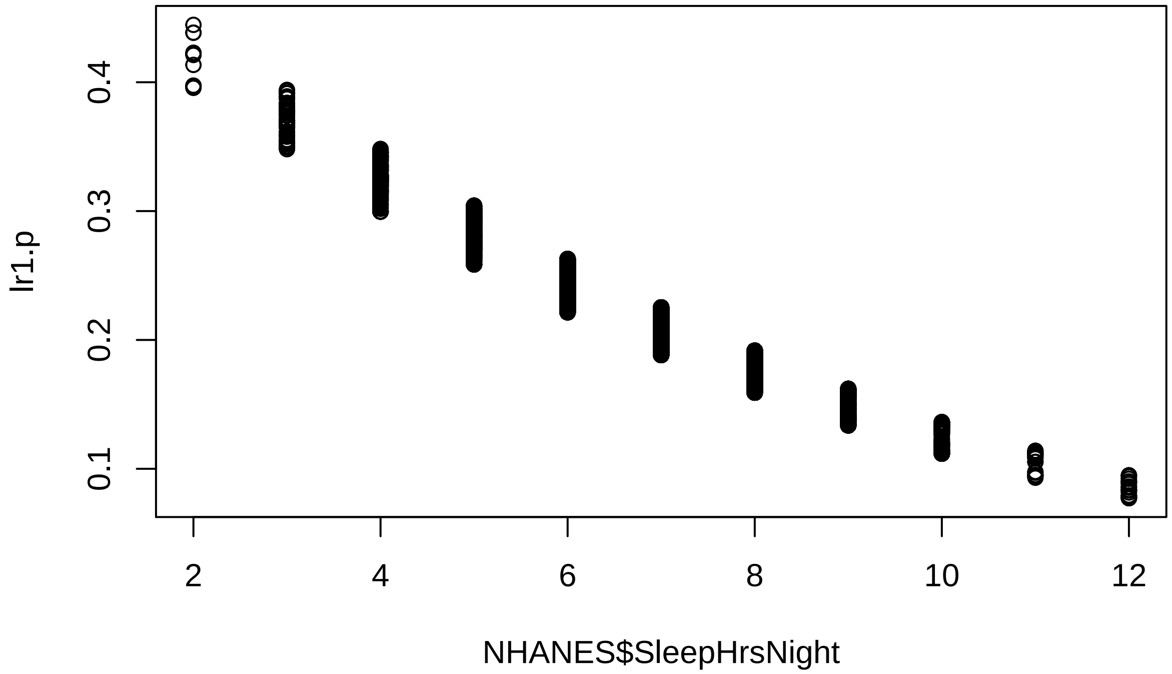
```
##        ID           SurveyYr       Gender          Age
##  Min.   :51624   2009_10:5000   female:5020   Min.   : 0.00
##  1st Qu.:56904   2011_12:5000   male  :4980   1st Qu.:17.00
##  Median :62160                                Median :36.00
##  Mean   :61945                                Mean   :36.74
##  3rd Qu.:67039                                3rd Qu.:54.00
##  Max.   :71915                                Max.   :80.00
##
##    AgeDecade      AgeMonths          Race1           Race3
##   40-49 :1398   Min.   :  0.0   Black   :1197   Asian   : 288
##   0-9   :1391   1st Qu.:199.0   Hispanic: 610   Black   : 589
##   10-19 :1374   Median :418.0   Mexican :1015   Hispanic: 350
##   20-29 :1356   Mean   :420.1   White   :6372   Mexican : 480
##   30-39 :1338   3rd Qu.:624.0   Other   : 806   White   :3135
##  (Other):2810   Max.   :959.0                   Other   : 158
##  NA's   : 333   NA's   :5038                    NA's    :5000
##            Education          MaritalStatus         HHIncome
##  8th Grade     : 451   Divorced    : 707   more 99999 :2220
##  9 - 11th Grade: 888   LivePartner : 560   75000-99999:1084
##  High School   :1517   Married     :3945   25000-34999: 958
##  Some College  :2267   NeverMarried:1380   35000-44999: 863
##  College Grad  :2098   Separated   : 183   45000-54999: 784
##  NA's          :2779   Widowed     : 456   (Other)    :3280
##                        NA's        :2769   NA's       : 811
##   HHIncomeMid        Poverty        HomeRooms       HomeOwn
##  Min.   :  2500   Min.   :0.000   Min.   : 1.000   Own  :6425
##  1st Qu.: 30000   1st Qu.:1.240   1st Qu.: 5.000   Rent :3287
##  Median : 50000   Median :2.700   Median : 6.000   Other: 225
##  Mean   : 57206   Mean   :2.802   Mean   : 6.249   NA's :  63
##  3rd Qu.: 87500   3rd Qu.:4.710   3rd Qu.: 8.000
##  Max.   :100000   Max.   :5.000   Max.   :13.000
##  NA's   :811      NA's   :726     NA's   :69
##          Work          Weight           Length          HeadCirc
##  Looking   : 311   Min.   :  2.80   Min.   : 47.10   Min.   :34.20
##  NotWorking:2847   1st Qu.: 56.10   1st Qu.: 75.70   1st Qu.:39.58
##  Working   :4613   Median : 72.70   Median : 87.00   Median :41.45
##  NA's      :2229   Mean   : 70.98   Mean   : 85.02   Mean   :41.18
##                    3rd Qu.: 88.90   3rd Qu.: 96.10   3rd Qu.:42.92
##                    Max.   :230.70   Max.   :112.20   Max.   :45.40
##                    NA's   :78       NA's   :9457     NA's   :9912
##      Height           BMI           BMICatUnder20yrs         BMI_WHO
##  Min.   : 83.6   Min.   :12.88   UnderWeight:  55    12.0_18.5   :1277
##  1st Qu.:156.8   1st Qu.:21.58   NormWeight : 805    18.5_to_24.9:2911
##  Median :166.0   Median :25.98   OverWeight : 193    25.0_to_29.9:2664
##  Mean   :161.9   Mean   :26.66   Obese      : 221    30.0_plus   :2751
##  3rd Qu.:174.5   3rd Qu.:30.89   NA's       :8726    NA's        : 397
##  Max.   :200.4   Max.   :81.25
##  NA's   :353     NA's   :366
##      Pulse           BPSysAve        BPDiaAve          BPSys1
##  Min.   : 40.00   Min.   : 76.0   Min.   :  0.00   Min.   : 72.0
##  1st Qu.: 64.00   1st Qu.:106.0   1st Qu.: 61.00   1st Qu.:106.0
##  Median : 72.00   Median :116.0   Median : 69.00   Median :116.0
##  Mean   : 73.56   Mean   :118.2   Mean   : 67.48   Mean   :119.1
##  3rd Qu.: 82.00   3rd Qu.:127.0   3rd Qu.: 76.00   3rd Qu.:128.0
##  Max.   :136.00   Max.   :226.0   Max.   :116.00   Max.   :232.0
##  NA's   :1437     NA's   :1449    NA's   :1449     NA's   :1763
##      BPDia1           BPSys2          BPDia2           BPSys3
##  Min.   :  0.00   Min.   : 76.0   Min.   :  0.00   Min.   : 76.0
##  1st Qu.: 62.00   1st Qu.:106.0   1st Qu.: 60.00   1st Qu.:106.0
##  Median : 70.00   Median :116.0   Median : 68.00   Median :116.0
##  Mean   : 68.28   Mean   :118.5   Mean   : 67.66   Mean   :117.9
##  3rd Qu.: 76.00   3rd Qu.:128.0   3rd Qu.: 76.00   3rd Qu.:126.0
##  Max.   :118.00   Max.   :226.0   Max.   :118.00   Max.   :226.0
##  NA's   :1763     NA's   :1647    NA's   :1647     NA's   :1635
##      BPDia3      Testosterone       DirectChol      TotChol
##  Min.   :  0.0   Min.   :   0.25   Min.   :0.390   Min.   : 1.530
##  1st Qu.: 60.0   1st Qu.:  17.70   1st Qu.:1.090   1st Qu.: 4.110
##  Median : 68.0   Median :  43.82   Median :1.290   Median : 4.780
##  Mean   : 67.3   Mean   : 197.90   Mean   :1.365   Mean   : 4.879
##  3rd Qu.: 76.0   3rd Qu.: 362.41   3rd Qu.:1.580   3rd Qu.: 5.530
##  Max.   :116.0   Max.   :1795.60   Max.   :4.030   Max.   :13.650
##  NA's   :1635    NA's   :5874      NA's   :1526    NA's   :1526
##    UrineVol1       UrineFlow1        UrineVol2       UrineFlow2
##  Min.   :  0.0   Min.   : 0.0000   Min.   :  0.0   Min.   : 0.000
##  1st Qu.: 50.0   1st Qu.: 0.4030   1st Qu.: 52.0   1st Qu.: 0.475
##  Median : 94.0   Median : 0.6990   Median : 95.0   Median : 0.760
##  Mean   :118.5   Mean   : 0.9793   Mean   :119.7   Mean   : 1.149
##  3rd Qu.:164.0   3rd Qu.: 1.2210   3rd Qu.:171.8   3rd Qu.: 1.513
##  Max.   :510.0   Max.   :17.1670   Max.   :409.0   Max.   :13.692
##  NA's   :987     NA's   :1603      NA's   :8522    NA's   :8524
##  Diabetes     DiabetesAge        HealthGen     DaysPhysHlthBad
##  No  :9098   Min.   : 1.00   Excellent: 878   Min.   : 0.000
##  Yes : 760   1st Qu.:40.00   Vgood    :2508   1st Qu.: 0.000
##  NA's: 142   Median :50.00   Good     :2956   Median : 0.000
##              Mean   :48.42   Fair     :1010   Mean   : 3.335
##              3rd Qu.:58.00   Poor     : 187   3rd Qu.: 3.000
##              Max.   :80.00   NA's     :2461   Max.   :30.000
##              NA's   :9371                     NA's   :2468
##  DaysMentHlthBad  LittleInterest   Depressed     nPregnancies
##  Min.   : 0.000   None   :5103   None   :5246   Min.   : 1.000
##  1st Qu.: 0.000   Several:1130   Several:1009   1st Qu.: 2.000
##  Median : 0.000   Most   : 434   Most   : 418   Median : 3.000
##  Mean   : 4.127   NA's   :3333   NA's   :3327   Mean   : 3.027
##  3rd Qu.: 4.000                                  3rd Qu.: 4.000
##  Max.   :30.000                                  Max.   :32.000
##  NA's   :2466                                    NA's   :7396
##     nBabies         Age1stBaby    SleepHrsNight    SleepTrouble
##  Min.   : 0.000   Min.   :14.00   Min.   : 2.000   No  :5799
##  1st Qu.: 2.000   1st Qu.:19.00   1st Qu.: 6.000   Yes :1973
##  Median : 2.000   Median :22.00   Median : 7.000   NA's:2228
##  Mean   : 2.457   Mean   :22.65   Mean   : 6.928
##  3rd Qu.: 3.000   3rd Qu.:26.00   3rd Qu.: 8.000
##  Max.   :12.000   Max.   :39.00   Max.   :12.000
##  NA's   :7584     NA's   :8116    NA's   :2245
##  PhysActive  PhysActiveDays       TVHrsDay        CompHrsDay
##  No  :3677   Min.   :1.000   2_hr     :1275   0_to_1_hr:1409
##  Yes :4649   1st Qu.:2.000   1_hr     : 884   0_hrs    :1073
##  NA's:1674   Median :3.000   3_hr     : 836   1_hr     :1030
##              Mean   :3.744   0_to_1_hr: 638   2_hr     : 589
##              3rd Qu.:5.000   More_4_hr: 615   3_hr     : 347
##              Max.   :7.000   (Other)  : 611   (Other)  : 415
##              NA's   :5337    NA's     :5141   NA's     :5137
##  TVHrsDayChild   CompHrsDayChild Alcohol12PlusYr   AlcoholDay
##  Min.   :0.000   Min.   :0.000   No  :1368       Min.   : 1.000
##  1st Qu.:1.000   1st Qu.:0.000   Yes :5212       1st Qu.: 1.000
##  Median :2.000   Median :1.000   NA's:3420       Median : 2.000
##  Mean   :1.939   Mean   :2.198                   Mean   : 2.914
##  3rd Qu.:3.000   3rd Qu.:6.000                   3rd Qu.: 3.000
##  Max.   :6.000   Max.   :6.000                   Max.   :82.000
##  NA's   :9347    NA's   :9347                    NA's   :5086
##   AlcoholYear    SmokeNow    Smoke100         Smoke100n       SmokeAge
##  Min.   :  0.0   No  :1745   No  :4024   Non-Smoker:4024   Min.   : 6.00
##  1st Qu.:  3.0   Yes :1466   Yes :3211   Smoker    :3211   1st Qu.:15.00
##  Median : 24.0   NA's:6789   NA's:2765   NA's      :2765   Median :17.00
##  Mean   : 75.1                                             Mean   :17.83
##  3rd Qu.:104.0                                             3rd Qu.:19.00
##  Max.   :364.0                                             Max.   :72.00  
##  NA's   :4078                                              NA's   :6920   
##  Marijuana   AgeFirstMarij   RegularMarij  AgeRegMarij    HardDrugs  
##  No  :2049   Min.   : 1.00   No  :3575    Min.   : 5.00   No  :4700  
##  Yes :2892   1st Qu.:15.00   Yes :1366    1st Qu.:15.00   Yes :1065  
##  NA's:5059   Median :16.00   NA's:5059    Median :17.00   NA's:4235  
##              Mean   :17.02                Mean   :17.69              
##              3rd Qu.:19.00                3rd Qu.:19.00              
##              Max.   :48.00                Max.   :52.00              
##              NA's   :7109                 NA's   :8634               
##  SexEver         SexAge      SexNumPartnLife   SexNumPartYear 
##  No  : 223   Min.   : 9.00   Min.   :   0.00   Min.   : 0.000  
##  Yes :5544   1st Qu.:15.00   1st Qu.:   2.00   1st Qu.: 1.000  
##  NA's:4233   Median :17.00   Median :   5.00   Median : 1.000  
##              Mean   :17.43   Mean   :  15.09   Mean   : 1.342  
##              3rd Qu.:19.00   3rd Qu.:  12.00   3rd Qu.: 1.000  
##              Max.   :50.00   Max.   :2000.00   Max.   :69.000  
##              NA's   :4460    NA's   :4275      NA's   :5072    
##  SameSex          SexOrientation  PregnantNow   depressedYes   
##  No  :5353   Bisexual    : 119   Yes    :  72   Mode :logical  
##  Yes : 415   Heterosexual:4638   No     :1573   FALSE:5246     
##  NA's:4232   Homosexual  :  85   Unknown:  51   TRUE :1427     
##              NA's        :5158   NA's   :8304   NA's :3327     
##                                                                
##                                                                
## 
```

```
#Split the data into a training and test dataset (90/10 split) based on a fix seed
set.seed(123456)
lr1 <- glm(depressedYes ~ SleepHrsNight + Age, data=NHANES, family=binomial)
summary
```

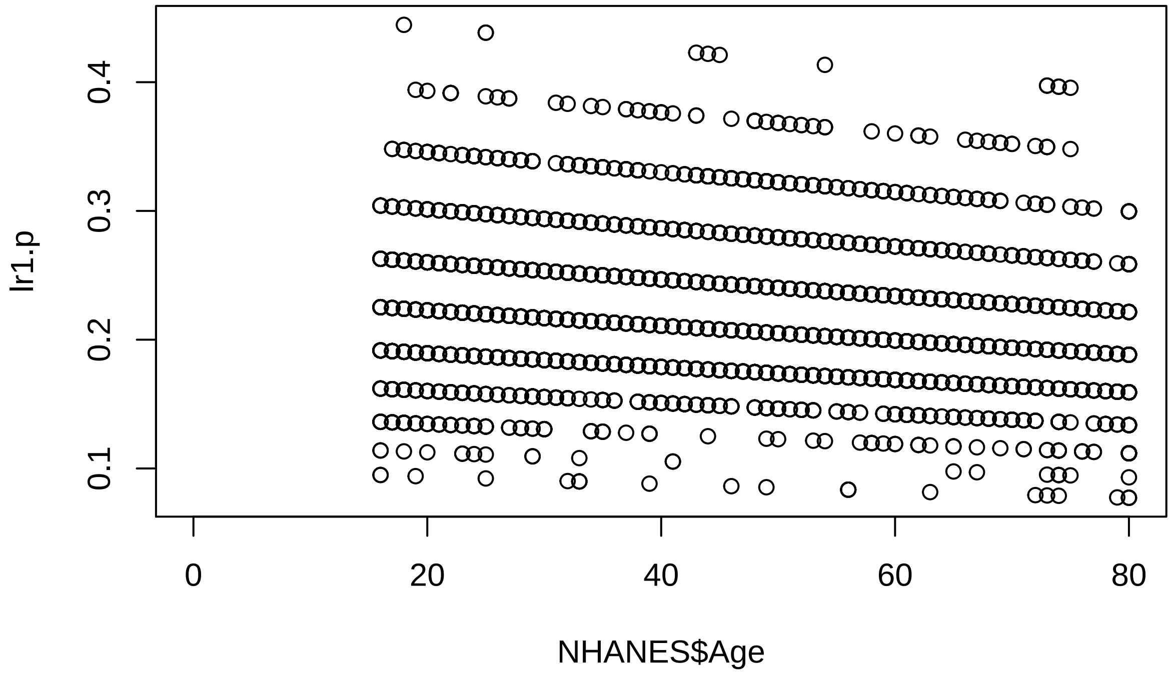
```
## standardGeneric for "summary" defined from package "base"
## 
## function (object, ...) 
## standardGeneric("summary")
## <environment: 0x7fbb1f790d38>
## Methods may be defined for arguments: object
## Use  showMethods("summary")  for currently available ones.
```

B)Report Effectiveness on NHANES Dataset AND C) Appropriate Visualizations Not a very good model. Sensitivity Is not good, Sepcificity is Acceptable.

```
#How did LR1 do in prediction?
lr1.p <- predict(lr1, newdata=NHANES, type = "response")
#plot for continuous predictor SleepHrsNight
plot(NHANES$SleepHrsNight, lr1.p) #plot tells us we need a probability of outcome around 0.25.
```



```r
plot(NHANES$Age, lr1.p)
```



```r
#Confusion Matrix
CrossTable(NHANES$depressedYes, lr1.p > 0.25)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  6660
##
##
##                       | lr1.p > 0.25
## NHANES$depressedYes |     FALSE |      TRUE | Row Total |
## --------------------|-----------|-----------|-----------|
##               FALSE |      4353 |       885 |      5238 |
##                     |     4.377 |    18.081 |           |
##                     |     0.831 |     0.169 |     0.786 |
##                     |     0.812 |     0.682 |           |
##                     |     0.654 |     0.133 |           |
## ---------------------|-----------|-----------|-----------|
##                 TRUE |      1009 |       413 |      1422 |
##                      |    16.122 |    66.601 |           |
##                      |     0.710 |     0.290 |     0.214 |
##                      |     0.188 |     0.318 |           |
##                      |     0.152 |     0.062 |           |
## ---------------------|-----------|-----------|-----------|
##         Column Total |      5362 |      1298 |      6660 |
##                      |     0.805 |     0.195 |           |
## ---------------------|-----------|-----------|-----------|
##
##
```

```
#OR we can get TPR and FPR with a 0.25 probability
#confusion matrix
t1 <- table(lr1.p > 0.25, NHANES$depressedYes)
t1 #gives same results as the CrossTable() funciton above.
```

```
##
##         FALSE TRUE
##   FALSE  4353 1009
##   TRUE    885  413
```

```
#calculate sensitivity
tpr <- t1[2,2]/(t1[2,2]+t1[1,2])
tpr  #not very good; only 19.5%
```

```
## [1] 0.290436
```

```
#calculate specificity
tnr <- t1[1,1]/(t1[1,1]+t1[2,1])
tnr #Pretty good at 89.7%
```

```
## [1] 0.8310424
```

```
#Look at Area under the curve
lr1.pr <- prediction(lr1.p, NHANES$depressedYes)
lr1.prf <- performance(lr1.pr, measure = "tpr", x.measure= "fpr") #I can't get this to run properly and
plot(lr1.prf)
abline(a=0, b=1, col="red")
#AUC
auc <- performance(lr1.pr, measure = "auc")
auc <- auc@y.values[[1]]
auc #auc "not enough distinct predicitons to compute area under the ROC curve" However, I have a feelin
```

D) Interpret

```
#Get ORs from lr1
exp(coef(lr1))
```

```
##   (Intercept) SleepHrsNight           Age
##     1.2820980     0.8155201     0.9964828
```

For everyone one hour increase in the number of sleep per night, the odds of being classified as depressed decreases by 0.1 when controlling for age. Controlling for number of hours of sleep/night, age pracitcally has no effect on the odds of being classified as depressed.

- decision tree

A) Build the Classifier

```
#Use Logisitc Regression Model from Above
summary(NHANES)
```

```
##        ID           SurveyYr       Gender          Age
##  Min.   :51624   2009_10:5000   female:5020   Min.   : 0.00
##  1st Qu.:56904   2011_12:5000   male  :4980   1st Qu.:17.00
##  Median :62160                                Median :36.00
##  Mean   :61945                                Mean   :36.74
##  3rd Qu.:67039                                3rd Qu.:54.00
##  Max.   :71915                                Max.   :80.00
##
##   AgeDecade      AgeMonths          Race1           Race3
##   40-49 :1398   Min.   :  0.0   Black   :1197   Asian   : 288
##   0-9   :1391   1st Qu.:199.0   Hispanic: 610   Black   : 589
##   10-19 :1374   Median :418.0   Mexican :1015   Hispanic: 350
##   20-29 :1356   Mean   :420.1   White   :6372   Mexican : 480
##   30-39 :1338   3rd Qu.:624.0   Other   : 806   White   :3135
##  (Other):2810   Max.   :959.0                   Other   : 158
##  NA's   : 333   NA's   :5038                    NA's    :5000
##           Education         MaritalStatus         HHIncome
##  8th Grade     : 451   Divorced     : 707   more 99999 :2220
##  9 - 11th Grade: 888   LivePartner : 560   75000-99999:1084
##  High School   :1517   Married     :3945   25000-34999: 958
##  Some College  :2267   NeverMarried:1380   35000-44999: 863
##  College Grad  :2098   Separated   : 183   45000-54999: 784
##  NA's          :2779   Widowed     : 456   (Other)    :3280
##                        NA's        :2769   NA's       : 811
##   HHIncomeMid        Poverty        HomeRooms       HomeOwn
##  Min.   :  2500   Min.   :0.000   Min.   : 1.000   Own  :6425
##  1st Qu.: 30000   1st Qu.:1.240   1st Qu.: 5.000   Rent :3287
##  Median : 50000   Median :2.700   Median : 6.000   Other: 225
##  Mean   : 57206   Mean   :2.802   Mean   : 6.249   NA's :  63
##  3rd Qu.: 87500   3rd Qu.:4.710   3rd Qu.: 8.000
##  Max.   :100000   Max.   :5.000   Max.   :13.000
##  NA's   :811      NA's   :726     NA's   :69
##         Work           Weight            Length          HeadCirc
##  Looking   : 311   Min.   :  2.80   Min.   : 47.10   Min.   :34.20
##  NotWorking:2847   1st Qu.: 56.10   1st Qu.: 75.70   1st Qu.:39.58
##  Working   :4613   Median : 72.70   Median : 87.00   Median :41.45
##  NA's      :2229   Mean   : 70.98   Mean   : 85.02   Mean   :41.18
##                    3rd Qu.: 88.90   3rd Qu.: 96.10   3rd Qu.:42.92
##                    Max.   :230.70   Max.   :112.20   Max.   :45.40
##                    NA's   :78       NA's   :9457     NA's   :9912
##      Height           BMI          BMICatUnder20yrs         BMI_WHO
##  Min.   : 83.6   Min.   :12.88   UnderWeight:  55    12.0_18.5   :1277
##  1st Qu.:156.8   1st Qu.:21.58   NormWeight : 805    18.5_to_24.9:2911
##  Median :166.0   Median :25.98   OverWeight : 193    25.0_to_29.9:2664
##  Mean   :161.9   Mean   :26.66   Obese      : 221    30.0_plus   :2751
##  3rd Qu.:174.5   3rd Qu.:30.89   NA's       :8726    NA's        : 397
##  Max.   :200.4   Max.   :81.25
##  NA's   :353     NA's   :366
##      Pulse           BPSysAve        BPDiaAve          BPSys1
##  Min.   : 40.00   Min.   : 76.0   Min.   :  0.00   Min.   : 72.0
##  1st Qu.: 64.00   1st Qu.:106.0   1st Qu.: 61.00   1st Qu.:106.0
##  Median : 72.00   Median :116.0   Median : 69.00   Median :116.0
##  Mean   : 73.56   Mean   :118.2   Mean   : 67.48   Mean   :119.1
##  3rd Qu.: 82.00   3rd Qu.:127.0   3rd Qu.: 76.00   3rd Qu.:128.0
##  Max.   :136.00   Max.   :226.0   Max.   :116.00   Max.   :232.0
##  NA's   :1437     NA's   :1449    NA's   :1449     NA's   :1763
##      BPDia1           BPSys2          BPDia2           BPSys3
##  Min.   :  0.00   Min.   : 76.0   Min.   :  0.00   Min.   : 76.0
##  1st Qu.: 62.00   1st Qu.:106.0   1st Qu.: 60.00   1st Qu.:106.0
##  Median : 70.00   Median :116.0   Median : 68.00   Median :116.0
##  Mean   : 68.28   Mean   :118.5   Mean   : 67.66   Mean   :117.9
##  3rd Qu.: 76.00   3rd Qu.:128.0   3rd Qu.: 76.00   3rd Qu.:126.0
##  Max.   :118.00   Max.   :226.0   Max.   :118.00   Max.   :226.0
##  NA's   :1763     NA's   :1647    NA's   :1647     NA's   :1635
##      BPDia3       Testosterone       DirectChol       TotChol
##  Min.   :  0.0   Min.   :   0.25   Min.   :0.390   Min.   : 1.530
##  1st Qu.: 60.0   1st Qu.:  17.70   1st Qu.:1.090   1st Qu.: 4.110
##  Median : 68.0   Median :  43.82   Median :1.290   Median : 4.780
##  Mean   : 67.3   Mean   : 197.90   Mean   :1.365   Mean   : 4.879
##  3rd Qu.: 76.0   3rd Qu.: 362.41   3rd Qu.:1.580   3rd Qu.: 5.530
##  Max.   :116.0   Max.   :1795.60   Max.   :4.030   Max.   :13.650
##  NA's   :1635    NA's   :5874      NA's   :1526    NA's   :1526
##    UrineVol1       UrineFlow1       UrineVol2       UrineFlow2
##  Min.   :  0.0   Min.   : 0.0000   Min.   :  0.0   Min.   : 0.000
##  1st Qu.: 50.0   1st Qu.: 0.4030   1st Qu.: 52.0   1st Qu.: 0.475
##  Median : 94.0   Median : 0.6990   Median : 95.0   Median : 0.760
##  Mean   :118.5   Mean   : 0.9793   Mean   :119.7   Mean   : 1.149
##  3rd Qu.:164.0   3rd Qu.: 1.2210   3rd Qu.:171.8   3rd Qu.: 1.513
##  Max.   :510.0   Max.   :17.1670   Max.   :409.0   Max.   :13.692
##  NA's   :987     NA's   :1603      NA's   :8522    NA's   :8524
##  Diabetes     DiabetesAge        HealthGen    DaysPhysHlthBad
##  No  :9098   Min.   : 1.00   Excellent: 878   Min.   : 0.000
##  Yes : 760   1st Qu.:40.00   Vgood    :2508   1st Qu.: 0.000
##  NA's: 142   Median :50.00   Good     :2956   Median : 0.000
##              Mean   :48.42   Fair     :1010   Mean   : 3.335
##              3rd Qu.:58.00   Poor     : 187   3rd Qu.: 3.000
##              Max.   :80.00   NA's     :2461   Max.   :30.000
##              NA's   :9371                     NA's   :2468
##  DaysMentHlthBad  LittleInterest   Depressed     nPregnancies
##  Min.   : 0.000   None   :5103   None   :5246   Min.   : 1.000
##  1st Qu.: 0.000   Several:1130   Several:1009   1st Qu.: 2.000
##  Median : 0.000   Most   : 434   Most   : 418   Median : 3.000
##  Mean   : 4.127   NA's   :3333   NA's   :3327   Mean   : 3.027
##  3rd Qu.: 4.000                                 3rd Qu.: 4.000
##  Max.   :30.000                                 Max.   :32.000
##  NA's   :2466                                   NA's   :7396
##     nBabies         Age1stBaby    SleepHrsNight    SleepTrouble
##  Min.   : 0.000   Min.   :14.00   Min.   : 2.000   No  :5799
##  1st Qu.: 2.000   1st Qu.:19.00   1st Qu.: 6.000   Yes :1973
##  Median : 2.000   Median :22.00   Median : 7.000   NA's:2228
##  Mean   : 2.457   Mean   :22.65   Mean   : 6.928
##  3rd Qu.: 3.000   3rd Qu.:26.00   3rd Qu.: 8.000
##  Max.   :12.000   Max.   :39.00   Max.   :12.000
##  NA's   :7584     NA's   :8116    NA's   :2245
##  PhysActive  PhysActiveDays       TVHrsDay        CompHrsDay
##  No  :3677   Min.   :1.000   2_hr     :1275   0_to_1_hr:1409
##  Yes :4649   1st Qu.:2.000   1_hr     : 884   0_hrs    :1073
##  NA's:1674   Median :3.000   3_hr     : 836   1_hr     :1030
##              Mean   :3.744   0_to_1_hr: 638   2_hr     : 589
##              3rd Qu.:5.000   More_4_hr: 615   3_hr     : 347
##              Max.   :7.000   (Other)  : 611   (Other)  : 415
##              NA's   :5337    NA's     :5141   NA's     :5137
##  TVHrsDayChild   CompHrsDayChild Alcohol12PlusYr   AlcoholDay
##  Min.   :0.000   Min.   :0.000   No  :1368       Min.   : 1.000
##  1st Qu.:1.000   1st Qu.:0.000   Yes :5212       1st Qu.: 1.000
##  Median :2.000   Median :1.000   NA's:3420       Median : 2.000
##  Mean   :1.939   Mean   :2.198                   Mean   : 2.914
##  3rd Qu.:3.000   3rd Qu.:6.000                   3rd Qu.: 3.000
##  Max.   :6.000   Max.   :6.000                   Max.   :82.000
##  NA's   :9347    NA's   :9347                    NA's   :5086
##   AlcoholYear    SmokeNow    Smoke100         Smoke100n       SmokeAge
##  Min.   :  0.0   No  :1745   No  :4024   Non-Smoker:4024   Min.   : 6.00
##  1st Qu.:  3.0   Yes :1466   Yes :3211   Smoker    :3211   1st Qu.:15.00
##  Median : 24.0   NA's:6789   NA's:2765   NA's      :2765   Median :17.00
##  Mean   : 75.1                                             Mean   :17.83
##  3rd Qu.:104.0                                             3rd Qu.:19.00
##  Max.   :364.0                                             Max.   :72.00
##  NA's   :4078                                              NA's   :6920
##  Marijuana   AgeFirstMarij   RegularMarij  AgeRegMarij    HardDrugs
##  No  :2049   Min.   : 1.00   No  :3575    Min.   : 5.00   No  :4700
##  Yes :2892   1st Qu.:15.00   Yes :1366    1st Qu.:15.00   Yes :1065
##  NA's:5059   Median :16.00   NA's:5059    Median :17.00   NA's:4235
##              Mean   :17.02                Mean   :17.69
##              3rd Qu.:19.00                3rd Qu.:19.00
##              Max.   :48.00                Max.   :52.00
##              NA's   :7109                 NA's   :8634
##  SexEver         SexAge      SexNumPartnLife   SexNumPartYear
##  No  : 223   Min.   : 9.00   Min.   :   0.00   Min.   : 0.000
##  Yes :5544   1st Qu.:15.00   1st Qu.:   2.00   1st Qu.: 1.000
##  NA's:4233   Median :17.00   Median :   5.00   Median : 1.000
##              Mean   :17.43   Mean   :  15.09   Mean   : 1.342
##              3rd Qu.:19.00   3rd Qu.:  12.00   3rd Qu.: 1.000
##              Max.   :50.00   Max.   :2000.00   Max.   :69.000
##              NA's   :4460    NA's   :4275      NA's   :5072
##  SameSex          SexOrientation  PregnantNow   depressedYes
##  No  :5353   Bisexual    : 119   Yes    :  72   Mode :logical
##  Yes : 415   Heterosexual:4638   No     :1573   FALSE:5246
##  NA's:4232   Homosexual  :  85   Unknown:  51   TRUE :1427
##              NA's        :5158   NA's   :8304   NA's :3327
##
```

```
##
##
```

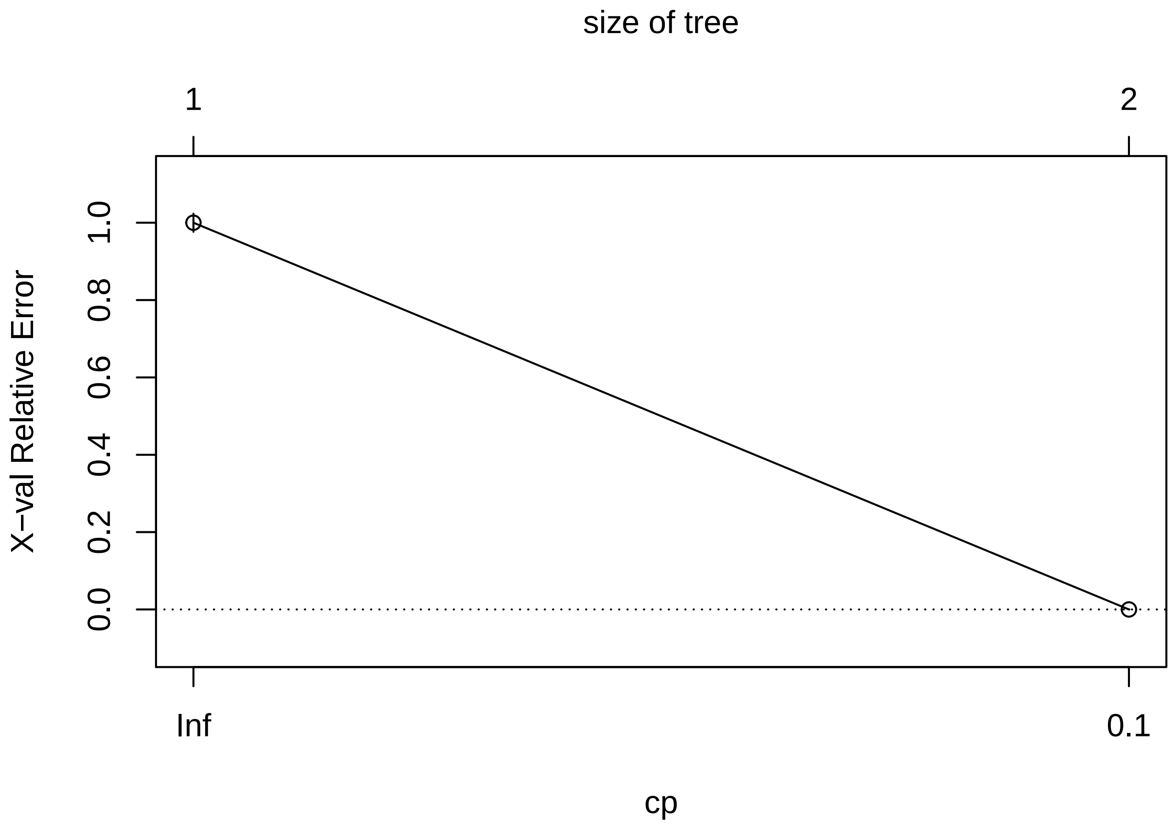
```
#grow tree
fitd <- rpart(depressedYes~., method="class", data = NHANES) #decided to include all possible predictor
class(fitd)
```

```
## [1] "rpart"
```

```
#display results
printcp(fitd)
```

```
##
## Classification tree:
## rpart(formula = depressedYes ~ ., data = NHANES, method = "class")
##
## Variables actually used in tree construction:
## [1] Depressed
##
## Root node error: 1427/6673 = 0.21385
##
## n=6673 (3327 observations deleted due to missingness)
##
##     CP nsplit rel error xerror     xstd
## 1 1.00      0         1      1 0.023472
## 2 0.01      1         0      0 0.000000
```

```
#Visualize Cross-Validation Restuls
plotcp(fitd)
```



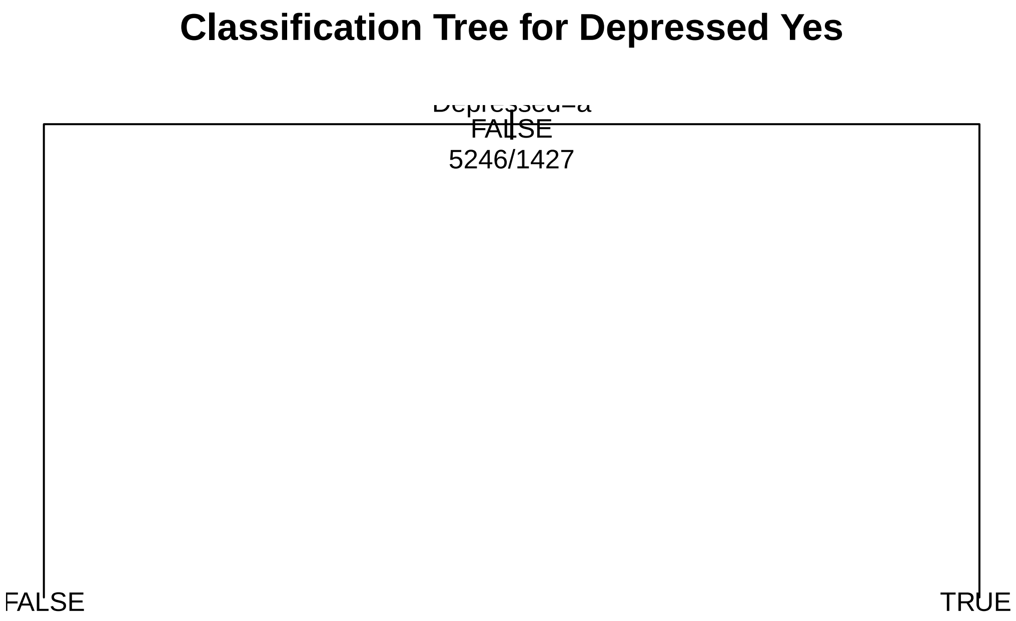
```
#Summary of Splits
summary(fitd)
```

```
## Call:
## rpart(formula = depressedYes ~ ., data = NHANES, method = "class")
##   n=6673 (3327 observations deleted due to missingness)
##
##     CP nsplit rel error xerror       xstd
## 1 1.00      0         1      1 0.02347154
## 2 0.01      1         0      0 0.00000000
##
## Variable importance
##       Depressed  LittleInterest DaysMentHlthBad       HealthGen
##              68              15              15               1
##
## Node number 1: 6673 observations,    complexity param=1
##   predicted class=FALSE  expected loss=0.2138468  P(node) =1
##     class counts:  5246  1427
##    probabilities: 0.786 0.214
##   left son=2 (5246 obs) right son=3 (1427 obs)
##   Primary splits:
##       Depressed       splits as  LRR,          improve=2243.68100, (0 missing)
##       LittleInterest  splits as  LRR,          improve= 625.24820, (8 missing)
##       DaysMentHlthBad < 2.5     to the left,  improve= 540.91780, (3 missing)
##       HealthGen       splits as  LLLRR,        improve=  86.13322, (0 missing)
##       SleepTrouble    splits as  LR,           improve=  80.23511, (0 missing)
##   Surrogate splits:
##       LittleInterest  splits as  LRR,         agree=0.834, adj=0.226, (0 split)
##       DaysMentHlthBad < 9.5     to the left,  agree=0.834, adj=0.221, (0 split)
##       HealthGen       splits as  LLLLR,       agree=0.789, adj=0.014, (0 split)
##       ID              < 51638.5 to the right, agree=0.787, adj=0.003, (0 split)
##
## Node number 2: 5246 observations
##   predicted class=FALSE  expected loss=0  P(node) =0.7861532
##     class counts:  5246     0
##    probabilities: 1.000 0.000
##
## Node number 3: 1427 observations
##   predicted class=TRUE   expected loss=0  P(node) =0.2138468
##     class counts:     0  1427
##    probabilities: 0.000 1.000
```

B) Report Effectiveness Was not effective on the NHANES data set. When including all variables, the only varibale included in the decision tree was "Depressed" which was used to make the "depressedYes" variable.

C) Visualization

```r
# Plot the tree
plot(fitd, uniform = TRUE, main = "Classification Tree for Depressed Yes")
text(fitd, use.n = TRUE, all = TRUE, cex = 0.8)
```



D) INterpret Results… There isn't much that is meaningful from this tree since the variable "depressed" was used to create "depressedYes"… Did I do something wrong?

- random forest

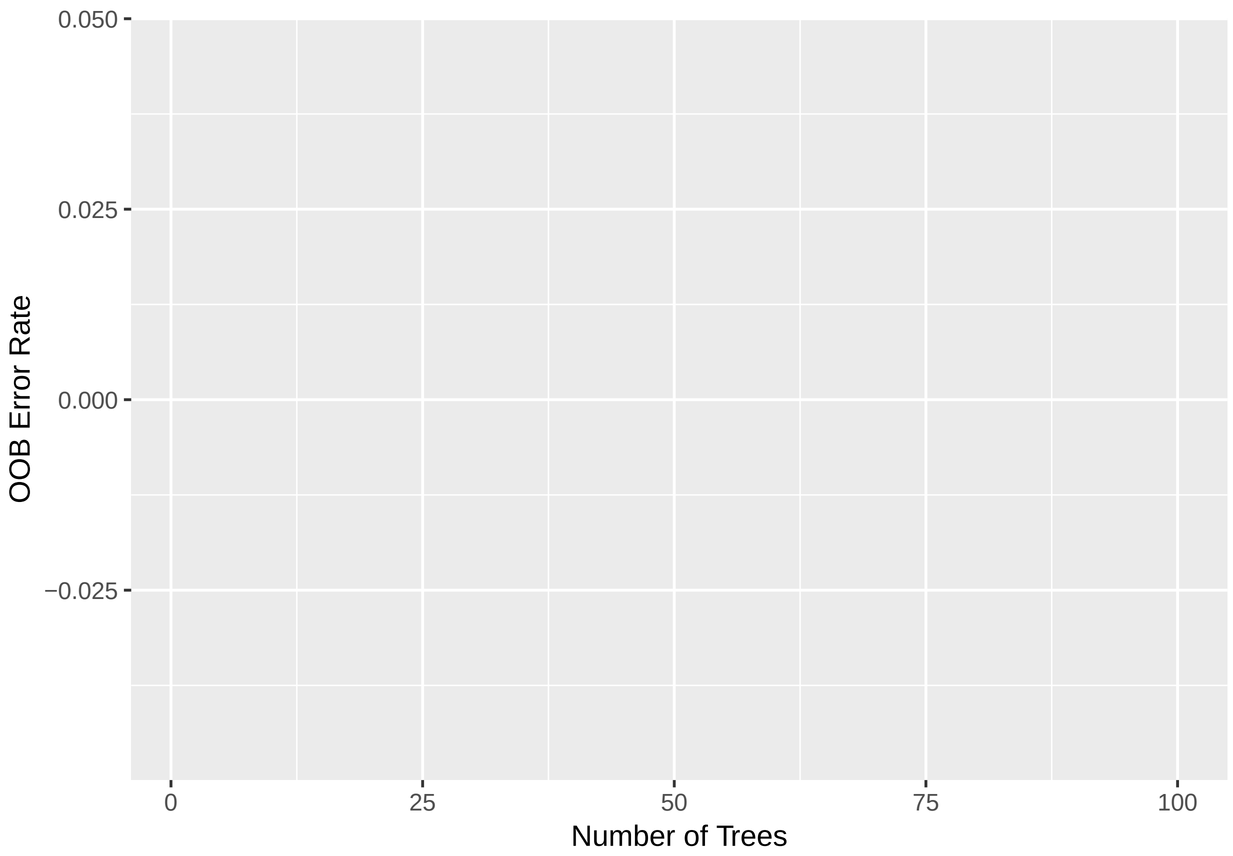
A) Build Classifier

```
NHANES.df <- as.data.frame(NHANES)
set.seed(456789)
# Random Forest for the ozone dataset
fitallrf <- rfsrc(depressedYes~., data=NHANES.df, ntree = 100, tree.err=TRUE, na.action = c("na.impute")
# view the results
fitallrf
```

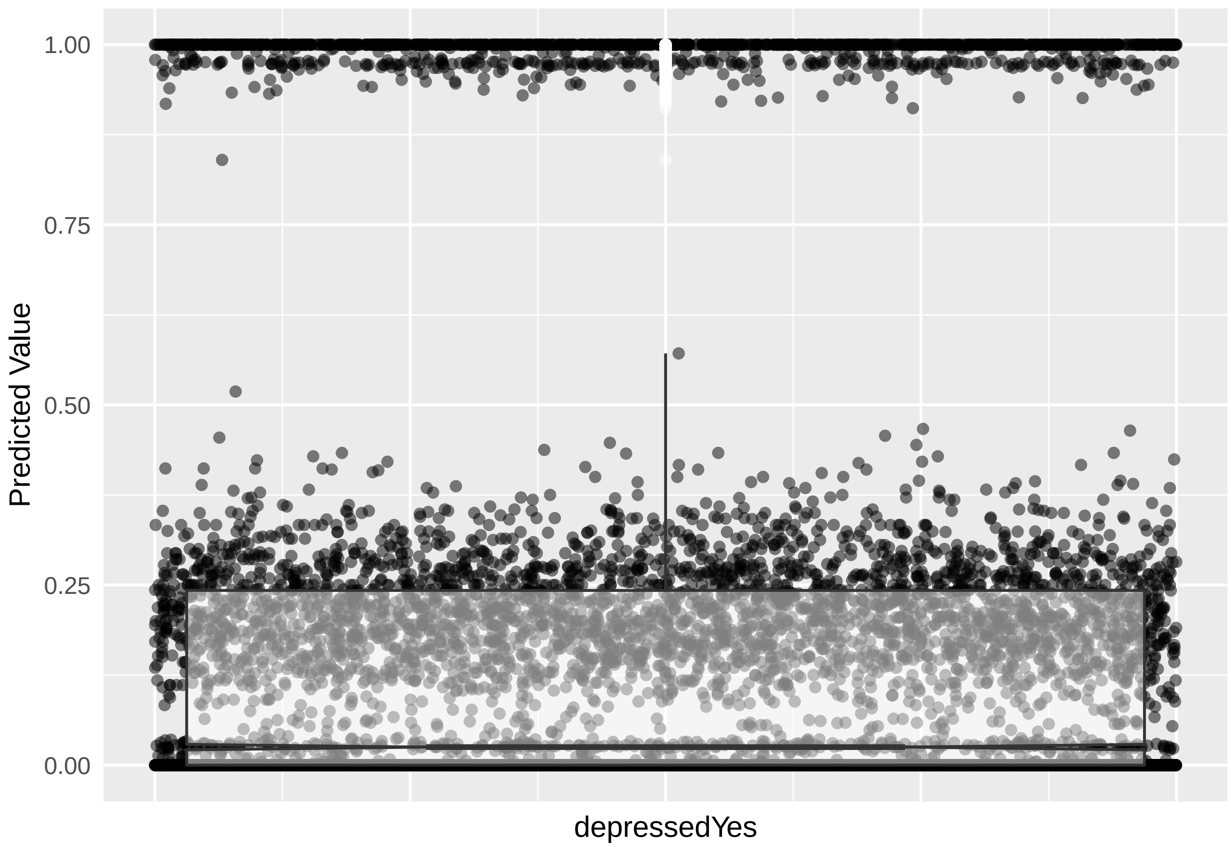
```
##                          Sample size: 10000
##                     Was data imputed: yes
##                      Number of trees: 100
##            Forest terminal node size: 5
##        Average no. of terminal nodes: 23.15
## No. of variables tried at each split: 26
##               Total no. of variables: 76
##        Resampling used to grow trees: swr
##     Resample size used to grow trees: 10000
##                             Analysis: RF-R
##                               Family: regr
##                       Splitting rule: mse *random*
##        Number of random split points: 10
##                 % variance explained: 99.95
##                           Error rate: 0
```

B) Report its effectiveness on the NHANES dataset and C) Make an approriate visualization of the model (I think this is answered here?)

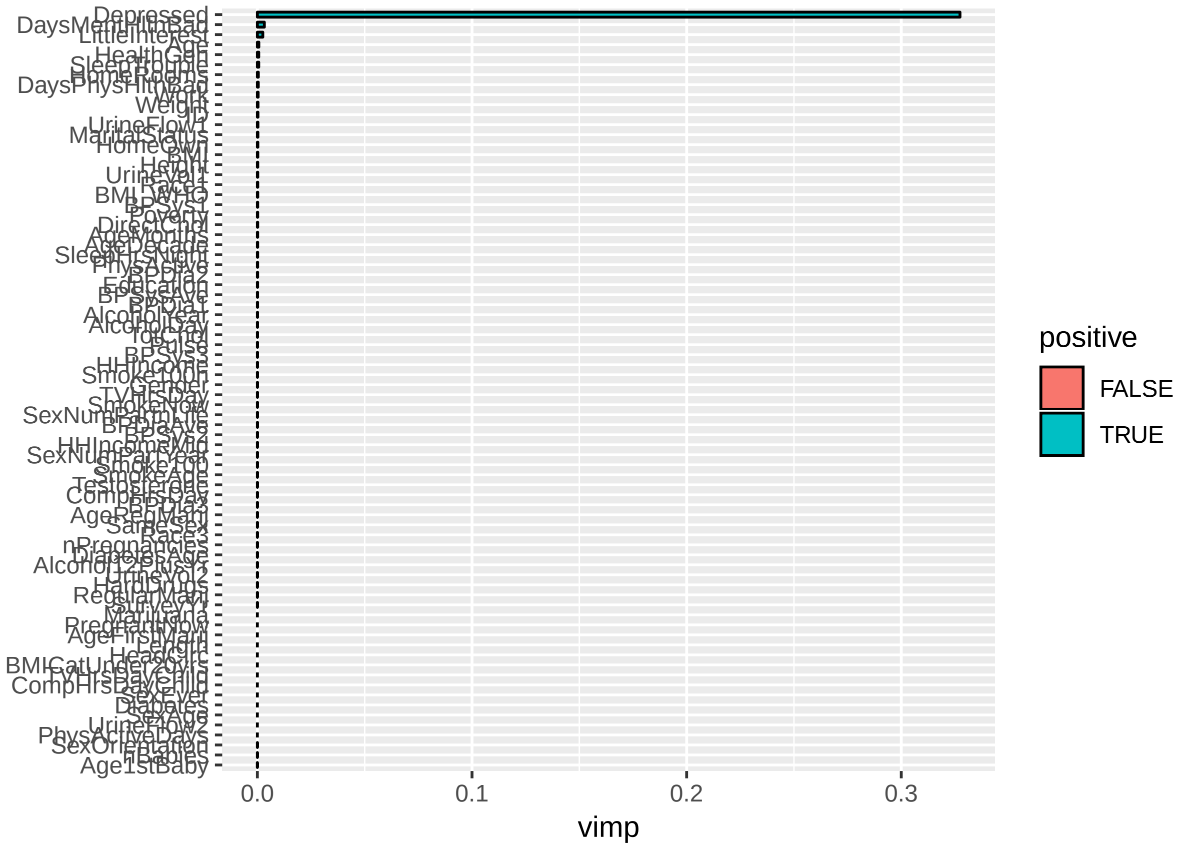
```
# Plot the OOB errors against the growth of the forest
gg_e <- gg_error(fitallrf) #only one tree reported an error value and it was tree '100' and the rate wa
plot(gg_e)
```



```
# Plot the predicted depressedYes values
plot(gg_rfsrc(fitallrf), alpha = 0.5)
```



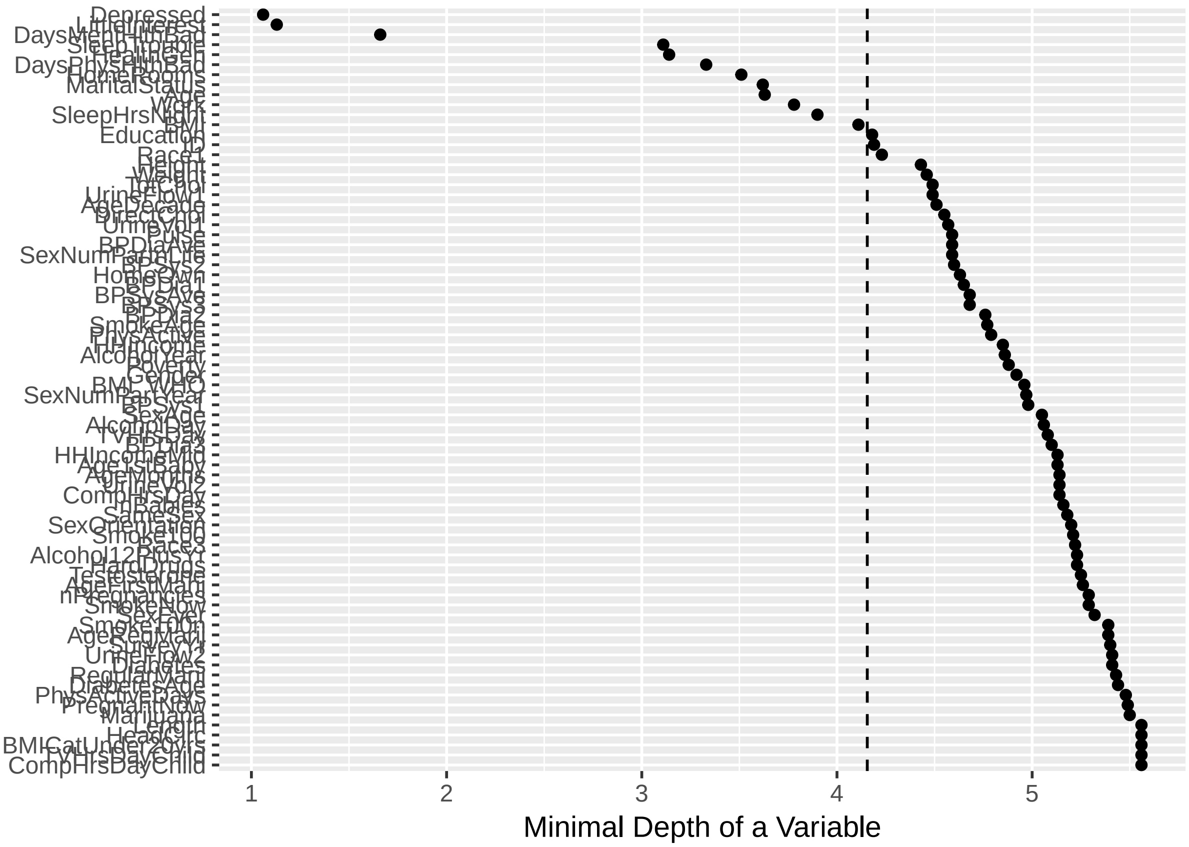
```r
#Plot VIMP rankings of independent variables
plot(gg_vimp(fitallrf))
```



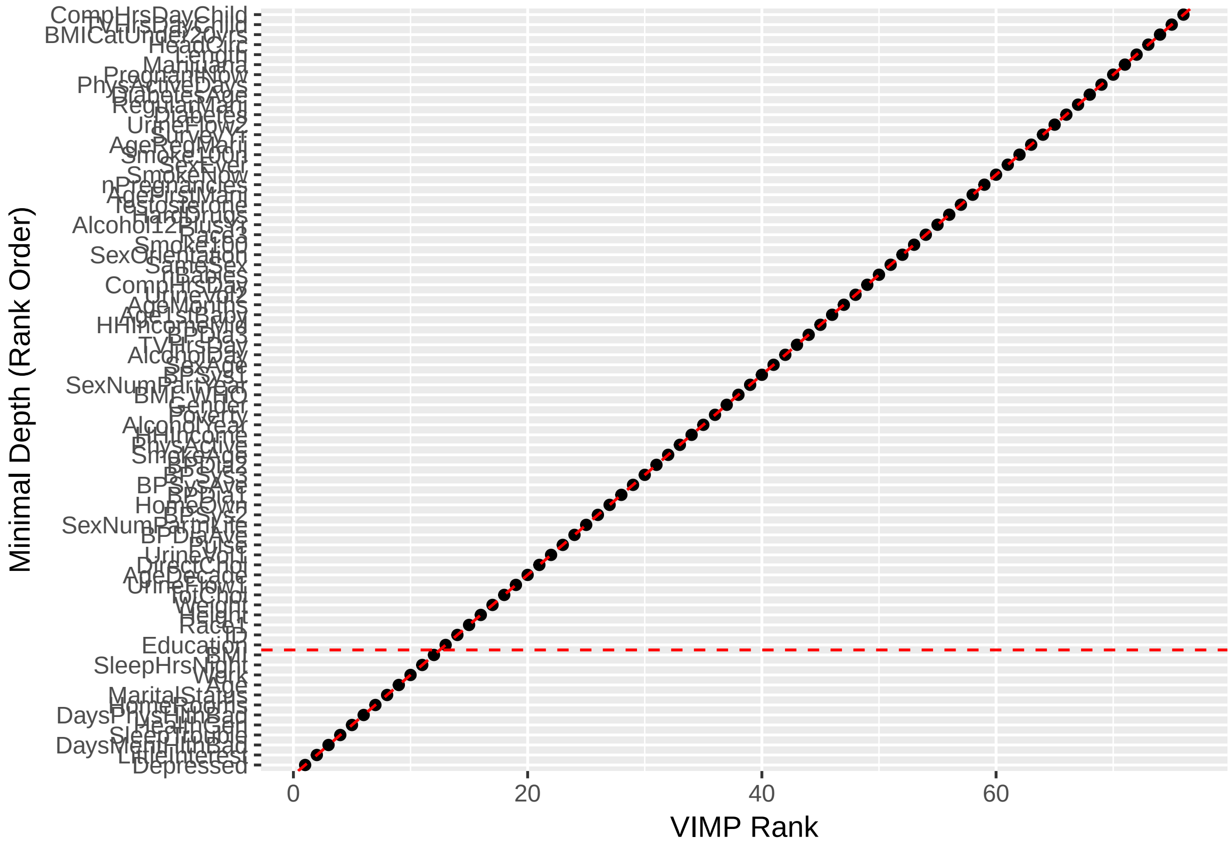
```r
#minimal depth
varsel_depressedYes <- var.select(fitallrf)
```

```
## minimal depth variable selection ...
##
##
## -----------------------------------------------------------
## family              : regr
## var. selection      : Minimal Depth
## conservativeness    : medium
## x-weighting used?   : TRUE
## dimension           : 76
## sample size         : 10000
## ntree               : 100
## nsplit              : 10
## mtry                : 26
## nodesize            : 5
## refitted forest     : FALSE
## model size          : 12
## depth threshold     : 4.1553
## PE (true OOB)       : 1e-04
##
##
## Top variables:
##                 depth vimp
## Depressed        1.06   NA
## LittleInterest   1.13   NA
## DaysMentHlthBad  1.66   NA
## SleepTrouble     3.11   NA
## HealthGen        3.14   NA
## DaysPhysHlthBad  3.33   NA
## HomeRooms        3.51   NA
## MaritalStatus    3.62   NA
## Age              3.63   NA
## Work             3.78   NA
## SleepHrsNight    3.90   NA
## BMI              4.11   NA
## -----------------------------------------------------------
```

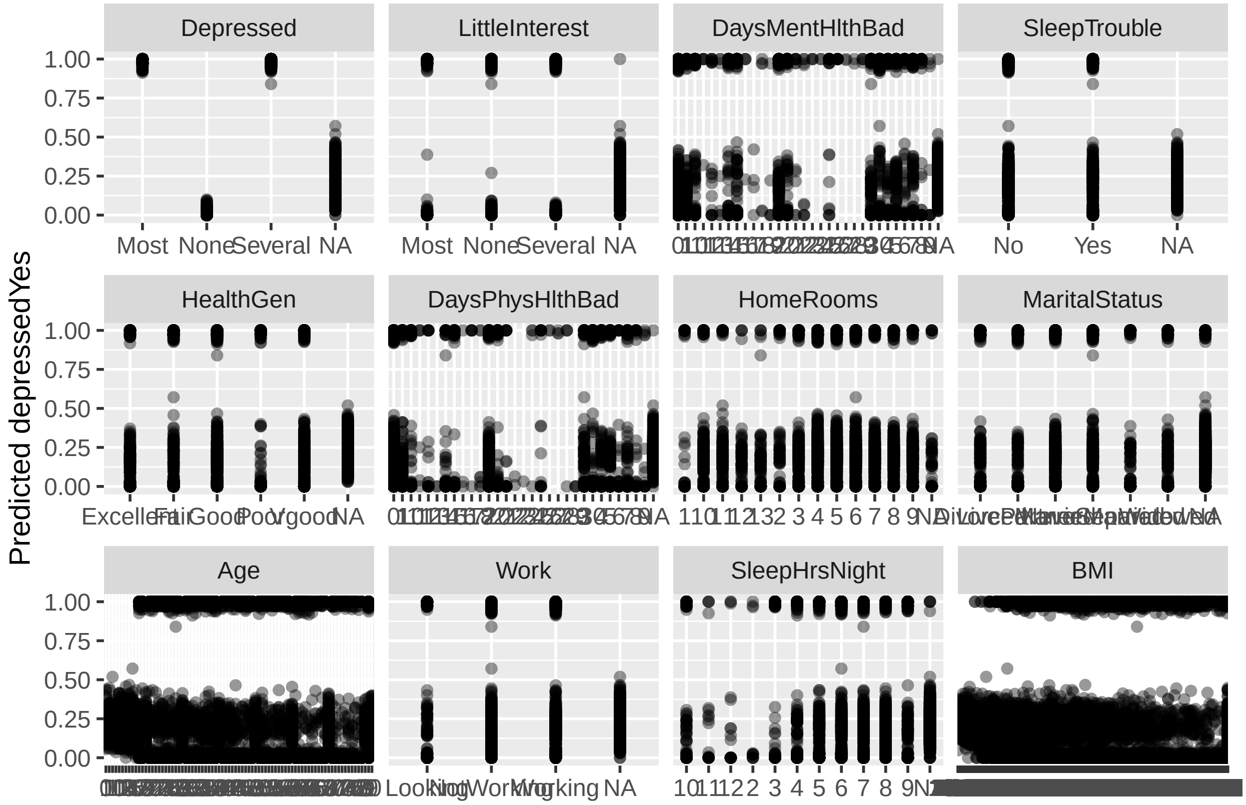
```
# Save the gg_minimal_depth object for later use
gg_md <- gg_minimal_depth(varsel_depressedYes)
# Plot the object
plot(gg_md)
```



```
# Plot minimal depth v VIMP
gg_mdVIMP <- gg_minimal_vimp(gg_md)
plot(gg_mdVIMP) #Honestly, I don't know why I have two lines... BUT if we go off the diagonal line, the
```



```r
gg_v <- gg_variable(fitallrf)
# Use the top ranked minimal depth variables only, plotted in minimal depth rank order
xvar <- gg_md$topvars
# Plot the variable list in a single panel plot
plot(gg_v, xvar = xvar, panel = TRUE, alpha = 0.4) +
  labs(y="Predicted depressedYes", x="")
```



D) Interpretation According to the minimal depth, the top 3 important variables in the prediciton of depressedYes are: Depressed, LittleInterest, and DaysMentHlthBad. This makes sense since depressedYes is a derivative of Depressed and DaysMentHlthBad is known to correlate with depressedYes. Clinically, LittleInterest ebing important in the prediction of depressedYes since having little interest in things you previously enjoyed is part of the clinical diagnosis of depression.

- k-nearest neighbor A)Build the Classifier

```
#Create a dataset from NHANES
NHANES2 <- NHANES %>%
  dplyr::select(Age, Gender, Diabetes, SleepHrsNight, BMI, HHIncome, PhysActive, depressedYes) %>%
  na.omit()
glimpse(NHANES2)
```

```
## Observations: 6,110
## Variables: 8
## $ Age           <int> 34, 34, 34, 49, 45, 45, 45, 66, 58, 54, 50, 33, ...
## $ Gender        <fct> male, male, male, female, female, female, female...
## $ Diabetes      <fct> No, No, No, No, No, No, No, No, No, No, No, No, ...
## $ SleepHrsNight <int> 4, 4, 4, 8, 8, 8, 8, 7, 5, 4, 7, 6, 6, 7, 7, 6, ...
## $ BMI           <dbl> 32.22, 32.22, 32.22, 30.57, 27.24, 27.24, 27.24,...
## $ HHIncome      <fct> 25000-34999, 25000-34999, 25000-34999, 35000-449...
## $ PhysActive    <fct> No, No, No, No, Yes, Yes, Yes, Yes, Yes, Yes, Ye...
## $ depressedYes  <lgl> TRUE, TRUE, TRUE, TRUE, FALSE, FALSE, FALSE, FAL...
```

```r
#Convert everything to numeric
NHANES2$Gender <- as.numeric(NHANES2$Gender)
NHANES2$Diabetes <- as.numeric(NHANES2$Diabetes)
NHANES2$HHIncome <- as.numeric(NHANES2$HHIncome)
NHANES2$PhysActive <- as.numeric(NHANES2$PhysActive)
NHANES2$depressedYes <- as.numeric(NHANES2$depressedYes)
```

B) Report effectiveness 100% prediciton using k of 1, 3, 5, 20, 50

```r
# Apply knn procedure to predict Diabetes
# Let's try different values of k to see how that affects performance
knn.1 <- knn(train = NHANES2, test = NHANES2, cl = NHANES2$depressedYes, k = 1)
knn.3 <- knn(train = NHANES2, test = NHANES2, cl = NHANES2$depressedYes, k = 3)
knn.5 <- knn(train = NHANES2, test = NHANES2, cl = NHANES2$depressedYes, k = 5)
knn.20 <- knn(train = NHANES2, test = NHANES2, cl = NHANES2$depressedYes, k =20)
knn.50 <- knn(train = NHANES2, test = NHANES2, cl = NHANES2$depressedYes, k =50)
#knn.1 amount correctly predicted
100*sum(NHANES2$depressedYes == knn.1)/length(knn.1)
```

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

```r
#knn.3 correct prediciton
100*sum(NHANES2$depressedYes == knn.1)/length(knn.3)
```

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

```r
#knn.5 correct
100*sum(NHANES2$depressedYes == knn.1)/length(knn.5)
```

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

```r
#knn.20 correct prediciton
100*sum(NHANES2$depressedYes == knn.1)/length(knn.20)
```

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

```r
#perfect prediction for all values of K... Let's try knn.50?
100*sum(NHANES2$depressedYes == knn.1)/length(knn.50)#still 100
```

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

C) Appropriate Visualization

D) Interpret the Results. What have you learned about people who self-report being depressed?

For each model do the following:

(A) Build the classifier.
(B) Report its effectiveness on the NHANES dataset.

(C) Make an appropriate visualization of this model.
(D) Interpret the results. What have you learned about people who self-report being depressed?

PROBLEM 2: Repeat problem 1 except now use the quantitative variable called `DaysMentHlthBad` as your outcome variable. Run 3 models:

- multiple linear regression,
- regression tree, and
- random forest.

And answer parts A, B, C, and D again for each model.

NOTE: `depressedYes` and `DaysMentHlthBad` are correlated but were 2 separate questions and are not perfectly aligned. The amount of missing data `NA's` are different between the 2 variables. To learn more about the variables in the dataset, run `help(NHANES, package = "NHANES")`.