573 Final Paper RMD HT

2022-11-08

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(gbm)
```

```
## Loaded gbm 2.1.8.1
```

```
library(ggplot2)
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(glmnet)
```

```
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
##
## Loaded glmnet 4.1-4
```

```r
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
##
## Attaching package: 'pROC'
##
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```r
library(class)
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```r
library(randomForest)
```

```
## randomForest 4.7-1.1
## Type rfNews() to see new features/changes/bug fixes.
##
## Attaching package: 'randomForest'
##
## The following object is masked from 'package:dplyr':
##
##     combine
##
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```r
library(tree)
library(e1071)
```

Read and clean data

```r
set.seed(12345)
data <- read.csv("Drug_Consumption.csv")
head(data)
```

```
##   ID   Age Gender                         Education Country Ethnicity   Nscore
## 1  2 25-34      M                  Doctorate degree      UK     White -0.67825
## 2  3 35-44      M Professional certificate/ diploma      UK     White -0.46725
## 3  4 18-24      F                    Masters degree      UK     White -0.14882
## 4  5 35-44      F                  Doctorate degree      UK     White  0.73545
## 5  6    65      F             Left school at 18 years  Canada     White -0.67825
```

```
## 6  7 45-54      M                    Masters degree     USA     White -0.46725
##     Escore   Oscore   AScore   Cscore Impulsive       SS Alcohol Amphet Amyl
## 1  1.93886  1.43533  0.76096 -0.14277  -0.71126 -0.21575     CL5    CL2  CL2
## 2  0.80523 -0.84732 -1.62090 -1.01450  -1.37983  0.40148     CL6    CL0  CL0
## 3 -0.80615 -0.01928  0.59042  0.58489  -1.37983 -1.18084     CL4    CL0  CL0
## 4 -1.63340 -0.45174 -0.30172  1.30612  -0.21712 -0.21575     CL4    CL1  CL1
## 5 -0.30033 -1.55521  2.03972  1.63088  -1.37983 -1.54858     CL2    CL0  CL0
## 6 -1.09207 -0.45174 -0.30172  0.93949  -0.21712  0.07987     CL6    CL0  CL0
##   Benzos Caff Cannabis Choc Coke Crack Ecstasy Heroin Ketamine Legalh LSD Meth
## 1    CL0  CL6      CL4  CL6  CL3   CL0     CL4    CL0      CL2    CL0 CL2  CL3
## 2    CL0  CL6      CL3  CL4  CL0   CL0     CL0    CL0      CL0    CL0 CL0  CL0
## 3    CL3  CL5      CL2  CL4  CL2   CL0     CL0    CL0      CL2    CL0 CL0  CL0
## 4    CL0  CL6      CL3  CL6  CL0   CL0     CL1    CL0      CL0    CL1 CL0  CL0
## 5    CL0  CL6      CL0  CL4  CL0   CL0     CL0    CL0      CL0    CL0 CL0  CL0
## 6    CL0  CL6      CL1  CL5  CL0   CL0     CL0    CL0      CL0    CL0 CL0  CL0
##   Mushrooms Nicotine Semer VSA
## 1       CL0      CL4   CL0 CL0
## 2       CL1      CL0   CL0 CL0
## 3       CL0      CL2   CL0 CL0
## 4       CL2      CL2   CL0 CL0
## 5       CL0      CL6   CL0 CL0
## 6       CL0      CL6   CL0 CL0
```

```
table(data$Age)
```

```
##
## 18-24 25-34 35-44 45-54 55-64    65
##   643   481   355   294    93    18
```

```
table(data$Gender)
```

```
##
##   F   M
## 941 943
```

```
table(data$Education)
```

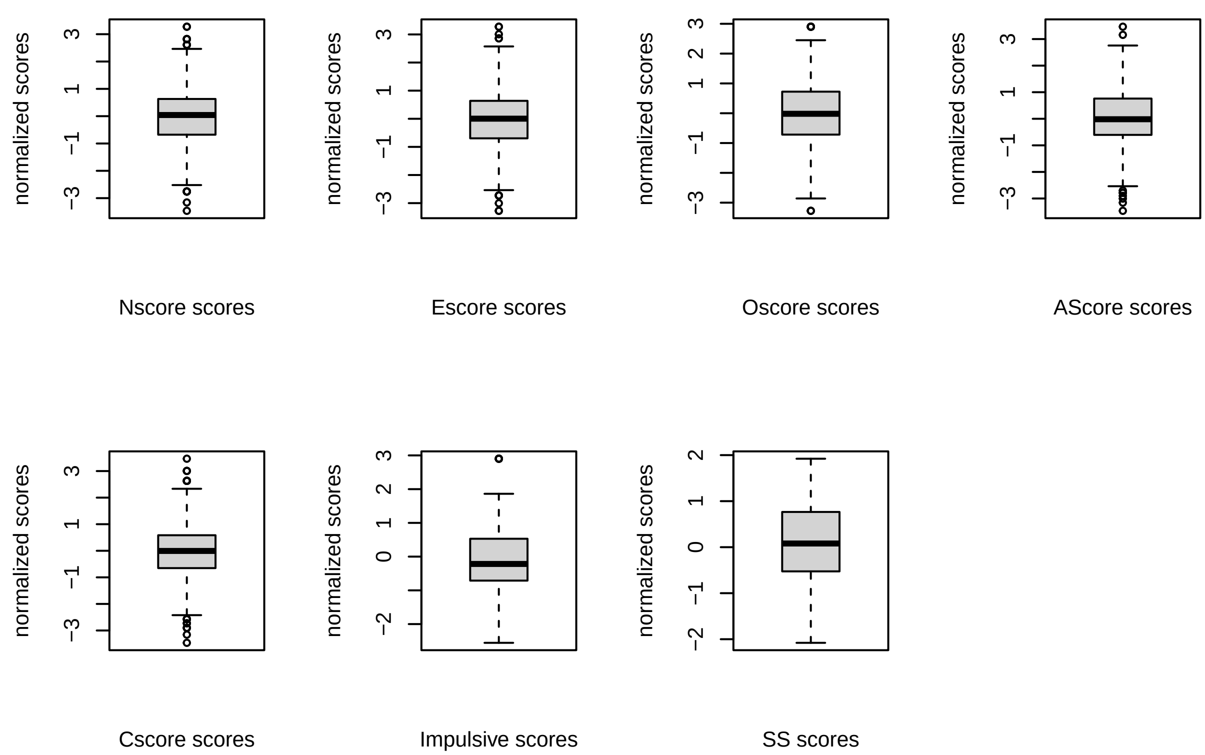
```
##
##                                    Doctorate degree
##                                                  89
##                              Left school at 16 years
##                                                  99
##                              Left school at 17 years
##                                                  30
##                              Left school at 18 years
##                                                 100
##                          Left school before 16 years
##                                                  28
##                                      Masters degree
##                                                 283
##                   Professional certificate/ diploma
##                                                 269
```

```
## Some college or university, no certificate or degree
##                                                   506
##                                     University degree
##                                                   480
```

```
table(data$Country)
```

```
##
##            Australia             Canada        New Zealand              Other
##                   54                 87                  5                118
## Republic of Ireland                 UK                USA
##                   20               1043                557
```

```
par(mfrow = c(2,4))
for(i in 7:13){
  boxplot(data[,i], xlab = paste(colnames(data)[i], "scores"), ylab = "normalized scores")
}
```



Read and clean data

```
set.seed(12345)
data <- read.csv("Drug_Consumption.csv")

# Remove the over-claimers using the control drug "Semer"
data <- subset(data, data$Semer == "CL0")
```

```r
for(i in 14:ncol(data)){
  data[,i] <- as.numeric(data[, i] == "CL4" | data[, i] == "CL5" | data[, i] == "CL6")
}

# Drop 65+
data <- data %>% mutate(dummy=1) %>%
spread(key=Age,value=dummy,fill=0)

# Drop Doctorate
data <- data %>% mutate(dummy=1) %>%
spread(key=Education,value=dummy,fill=0)

# Drop other
data <- data %>% mutate(dummy=1) %>%
spread(key=Country,value=dummy,fill=0)

# Drop other
data <- data %>% mutate(dummy=1) %>%
spread(key=Ethnicity,value=dummy,fill=0)

# Drop 'F' variable and rename to gender
data <- data %>% mutate(dummy=1) %>%
spread(key=Gender,value=dummy, fill=0)

# Drop variables that we aren't using.
drop <- c("ID", "65+","Doctorate degree","Other","F","Amphet","Amyl","Benzos","Choc","Crack", "Coke","E
data <- data[,!(names(data) %in% drop)]

names(data)[names(data) == "M"] <- "Gender"

# Split into test and train data
test.i <- sample(1:nrow(data), .3*nrow(data))
test.data <- data[test.i,]
train.data <- data[-test.i,]
```

Generate Tables for Data

```r
head(data)
```

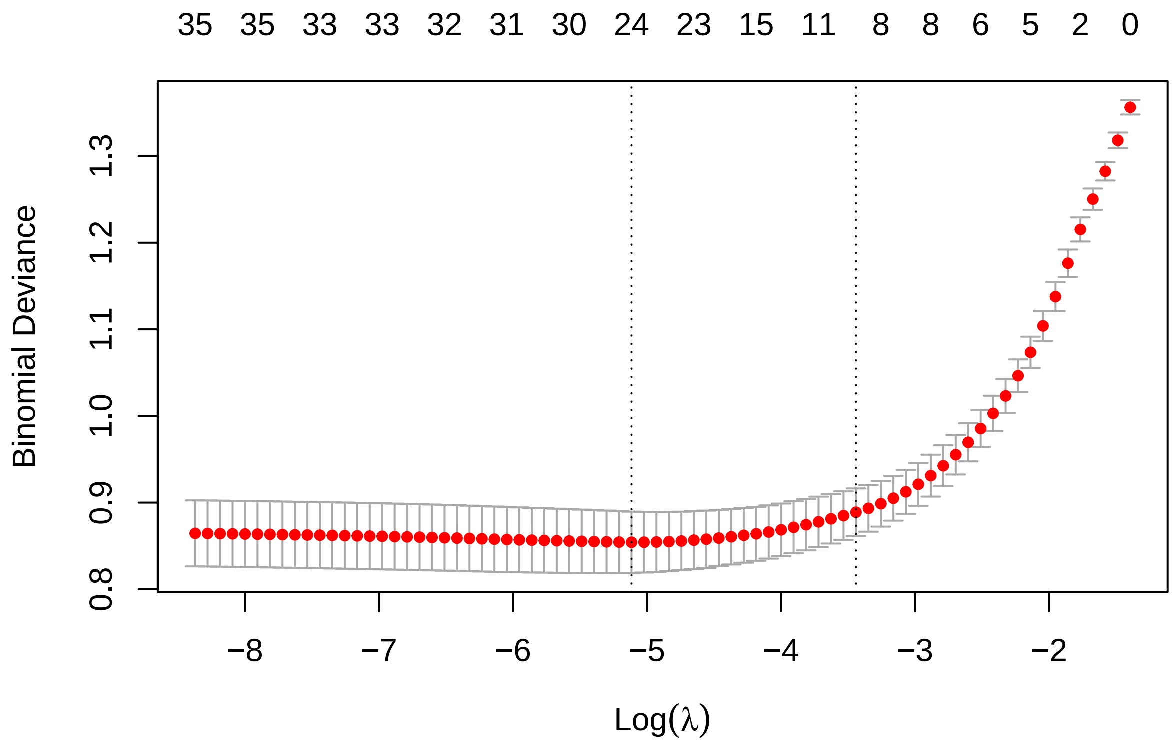
```
##     Nscore   Escore   Oscore   AScore   Cscore Impulsive       SS Alcohol Caff
## 1 -0.67825  1.93886  1.43533  0.76096 -0.14277  -0.71126 -0.21575       1    1
## 2 -0.46725  0.80523 -0.84732 -1.62090 -1.01450  -1.37983  0.40148       1    1
## 3 -0.14882 -0.80615 -0.01928  0.59042  0.58489  -1.37983 -1.18084       1    1
## 4  0.73545 -1.63340 -0.45174 -0.30172  1.30612  -0.21712 -0.21575       1    1
## 5 -0.67825 -0.30033 -1.55521  2.03972  1.63088  -1.37983 -1.54858       0    1
## 6 -0.46725 -1.09207 -0.45174 -0.30172  0.93949  -0.21712  0.07987       1    1
##   Cannabis Nicotine 18-24 25-34 35-44 45-54 55-64 65 Left school at 16 years
## 1        1        1     0     1     0     0     0  0                       0
## 2        0        0     0     0     1     0     0  0                       0
## 3        0        0     1     0     0     0     0  0                       0
## 4        0        0     0     0     1     0     0  0                       0
## 5        0        1     0     0     0     0     0  1                       0
## 6        0        1     0     0     0     1     0  0                       0
```

```
##   Left school at 17 years Left school at 18 years Left school before 16 years
## 1                       0                       0                           0
## 2                       0                       0                           0
## 3                       0                       0                           0
## 4                       0                       0                           0
## 5                       0                       1                           0
## 6                       0                       0                           0
##   Masters degree Professional certificate/ diploma
## 1              0                                 0
## 2              0                                 1
## 3              1                                 0
## 4              0                                 0
## 5              0                                 0
## 6              1                                 0
##   Some college or university, no certificate or degree University degree
## 1                                                    0                 0
## 2                                                    0                 0
## 3                                                    0                 0
## 4                                                    0                 0
## 5                                                    0                 0
## 6                                                    0                 0
##   Australia Canada New Zealand Republic of Ireland UK USA Asian Black
## 1         0      0           0                   0  1   0     0     0
## 2         0      0           0                   0  1   0     0     0
## 3         0      0           0                   0  1   0     0     0
## 4         0      0           0                   0  1   0     0     0
## 5         0      1           0                   0  0   0     0     0
## 6         0      0           0                   0  0   1     0     0
##   Mixed-Black/Asian Mixed-White/Asian Mixed-White/Black White Gender
## 1                 0                 0                 0     1      1
## 2                 0                 0                 0     1      1
## 3                 0                 0                 0     1      0
## 4                 0                 0                 0     1      0
## 5                 0                 0                 0     1      0
## 6                 0                 0                 0     1      1
```

```r
par(mfrow = c(1,4))
Alc_table <- table(data$Alcohol)
Caff_table<- table(data$Caff)
Cann_table <- table(data$Cannabis)
Nic_table <- table(data$Nicotine)
```

LASSO Exploration

```r
set.seed(123)
#Setting up matrices for lasso
x <- model.matrix(Cannabis~., data = data)[, -1]
y <- data$Cannabis
x.test <- as.matrix(test.data[,-10])
y.test <- test.data$Cannabis

#CV for Optimal Lambda
cv.out <- cv.glmnet(x, y, alpha = 1, family = 'binomial')
plot(cv.out)
```

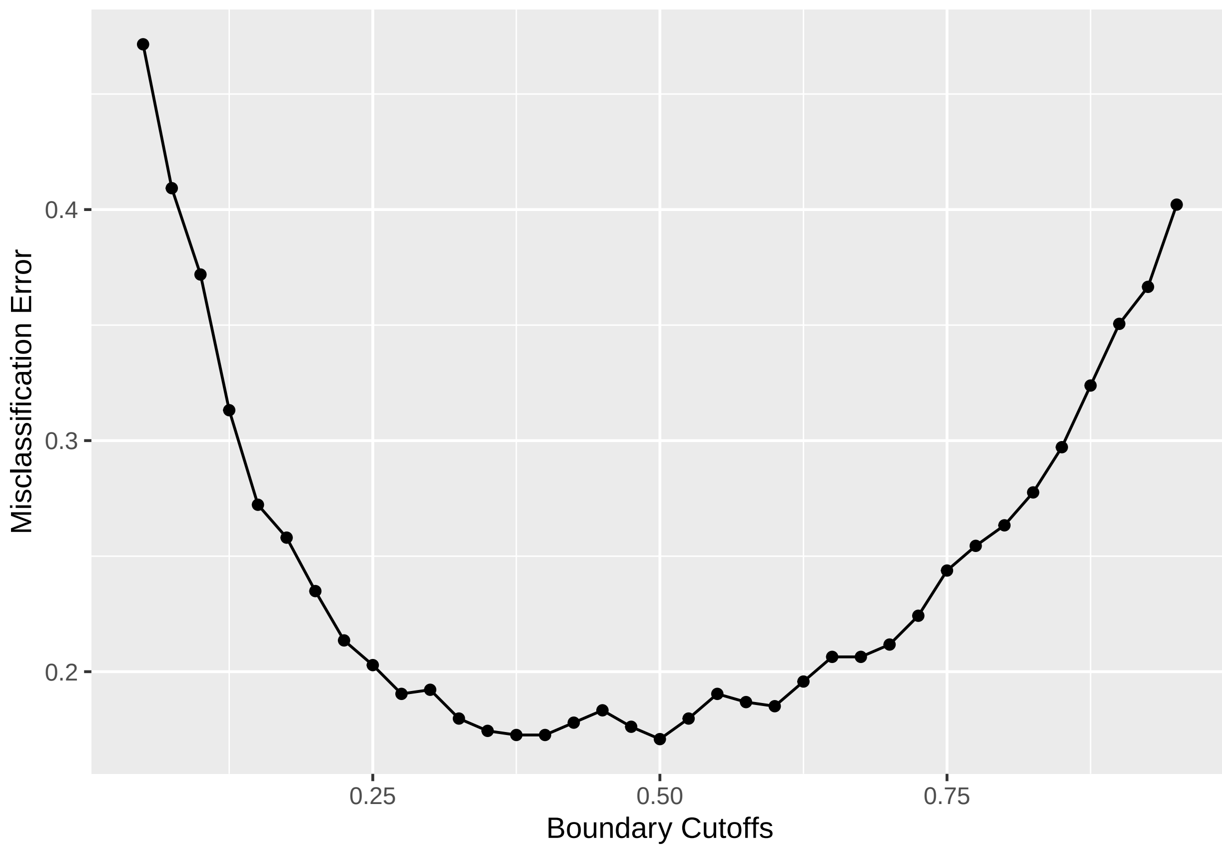



```
lambda.opt <- cv.out$lambda.min
lambda.opt # 0.006588544
```

```
## [1] 0.006003236
```

```
# Lasso
lasso <- glmnet(x, y, alpha = 1, lambda = lambda.opt, family = "binomial")

#Lasso Regression
lasso.pred <- predict(lasso, s = lambda.opt, newx = x.test, type = "response")

# Assign a class to predictions based on boundary optimization found by this
# code.

cutoffs <- seq(.05, .95, by = .025)
preds <- rep(0,length(lasso.pred))
error.lasso <- rep(0,length(lasso.pred))
lasso.test.err <- rep(NA, length(cutoffs))

for(i in 1:length(cutoffs)){
  preds <- ifelse(lasso.pred < cutoffs[i], 0, 1)

  for(e in 1:length(preds)){
    error.lasso[e] <- (preds[e] == y.test[e])
  }
```

```
  lasso.test.err[i] = (length(error.lasso)-sum(error.lasso))/length(error.lasso)
}

df <- data.frame(cutoffs, lasso.test.err)
ggplot(data = df, aes(x = cutoffs, y = lasso.test.err)) +
  geom_point() +
  geom_line() +
  xlab("Boundary Cutoffs") +
  ylab("Misclassification Error")
```




```
min(lasso.test.err) # 0.1725979
```

```
## [1] 0.1708185
```

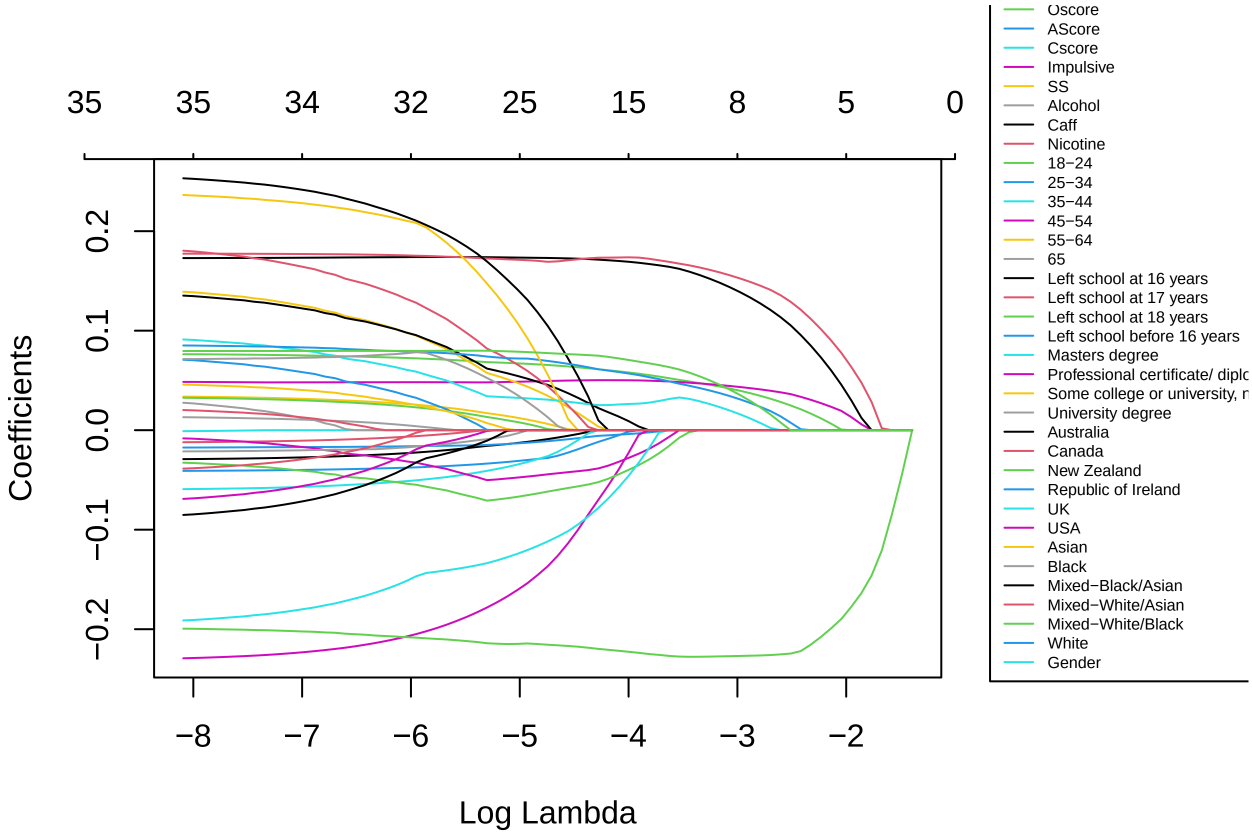
```
cutoffs[which.min(lasso.test.err)] # 0.5
```

```
## [1] 0.5
```

```
# This process verified that 0.5 is the optimal cutoff to minimize test error
# using this lasso regression. We reached a test error rate of 0.1725979 or
# a success rate of 82.74%

#Predictor Coefficients after Lasso
coef(lasso)
```

```
## 38 x 1 sparse Matrix of class "dgCMatrix"
##                                                                      s0
## (Intercept)                                                 -1.03621469
## Nscore                                                      -0.08503095
## Escore                                                       .
## Oscore                                                       0.45119455
## AScore                                                       .
## Cscore                                                      -0.08470724
## Impulsive                                                    .
## SS                                                           0.33743838
## Alcohol                                                      0.12387857
## Caff                                                         .
## Nicotine                                                     1.07268340
## `18-24`                                                      0.91697043
## `25-34`                                                      .
## `35-44`                                                      .
## `45-54`                                                     -0.23307344
## `55-64`                                                     -0.13107620
## `65`                                                        -1.73335963
## `Left school at 16 years`                                    0.36636759
## `Left school at 17 years`                                    .
## `Left school at 18 years`                                    0.36093765
## `Left school before 16 years`                                0.39234000
## `Masters degree`                                            -0.42727518
## `Professional certificate/ diploma`                          .
## `Some college or university, no certificate or degree`       0.16860336
## `University degree`                                         -0.31771326
## Australia                                                    0.04905410
## Canada                                                       .
## `New Zealand`                                                0.92819062
## `Republic of Ireland`                                        .
## UK                                                          -1.14315070
## USA                                                          0.40420861
## Asian                                                       -1.15089309
## Black                                                        .
## `Mixed-Black/Asian`                                          0.67707959
## `Mixed-White/Asian`                                          0.37697095
## `Mixed-White/Black`                                          .
## White                                                        .
## Gender                                                       0.49309850
```

```
# Make a new data set removing the variables considered insignificant by the
# lasso regression.
data.lasso <-  subset(data, select = -c(Escore, AScore, Impulsive, Caff, `35-44`, `Left school at 17 yea

# Test and training sets for lasso
test.lasso <- data.lasso[test.i,]
train.lasso <- data.lasso[-test.i,]

#Lasso Plot
par(mar=c(5, 4, 4, 8), xpd=TRUE)
lasso.plot <- glmnet(x, y, alpha = 1)
plot(lasso.plot, "lambda", col = 1:36)
legend("topright", inset=c(-0.6, -.4), lwd = 1, col= 1:37, legend = colnames(data[,-10]), cex = 0.5)
```

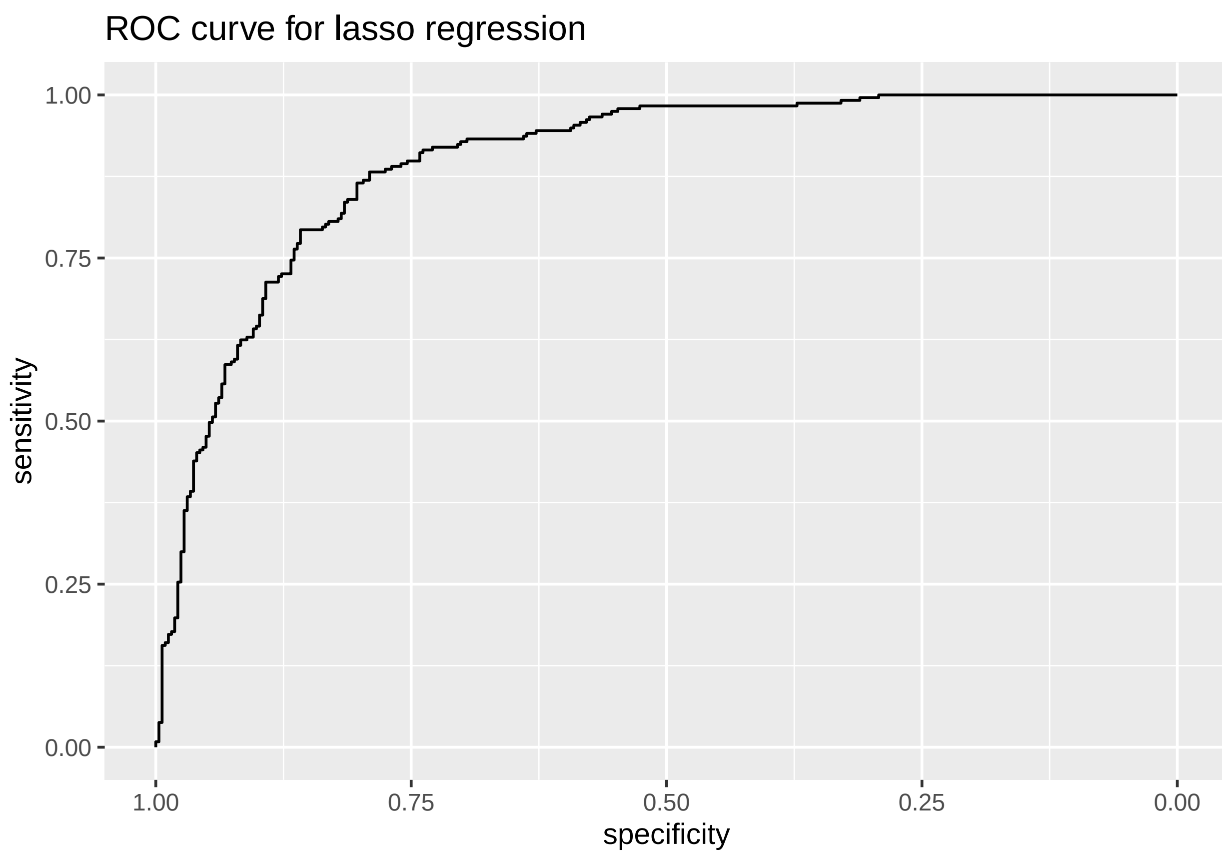



```
lasso.pred <- as.numeric(lasso.pred)
ROC.score.lasso <- roc(test.data$Cannabis, lasso.pred)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

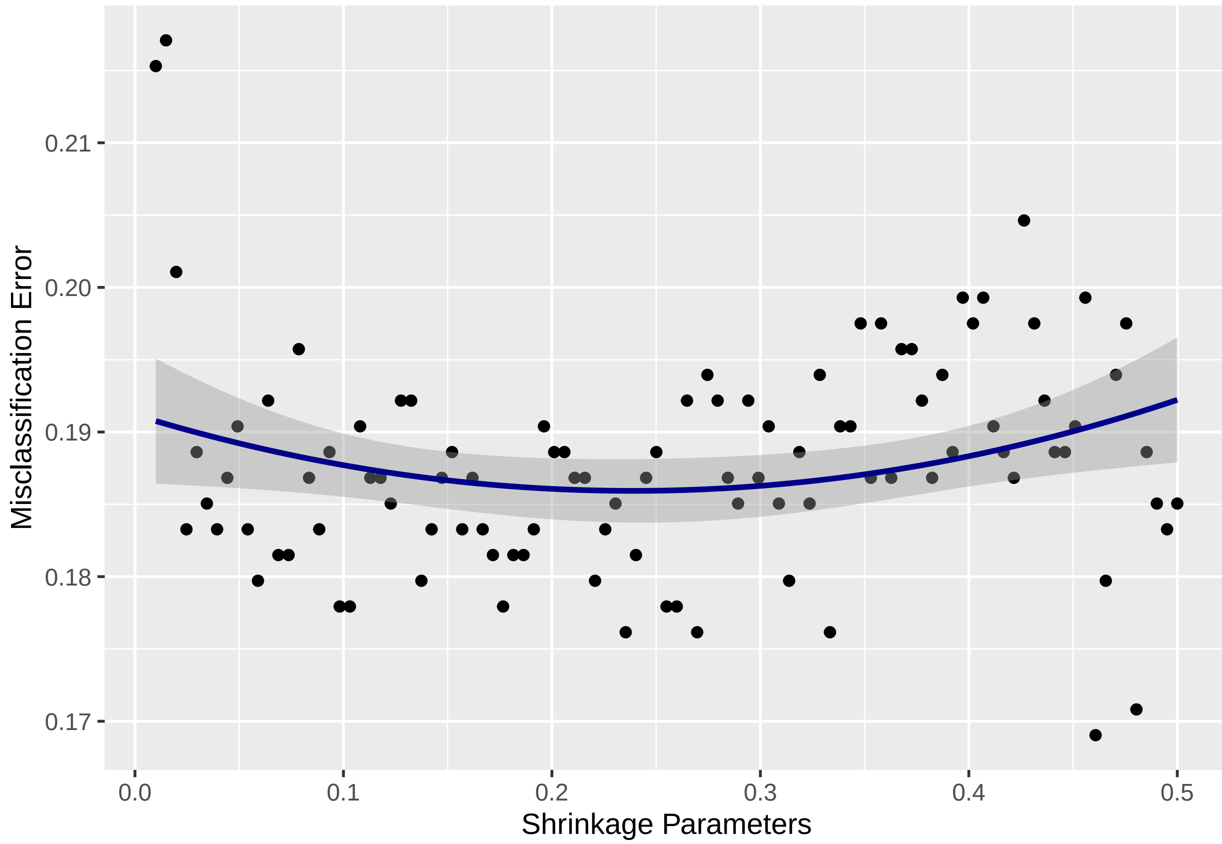
```
ggroc(ROC.score.lasso, legacy.axes = FALSE) +
  ggtitle("ROC curve for lasso regression")
```



Boosting - Finding the optimal shrinkage parameter

```
# Cycle through the shrinkage parameters to find the ideal value based on
# test MSE. Plot test error along different shrinkage values to find the
# ideal value.

# We will find the optimal cutoff for the prediction boundary using the
# optimal shrinkage coefficient found through through this process. We will use
# 0.5 as the cutoff for this process and we will optimize the decision
# boundary based on the optimal shrinkage value to compensate for the unequal
# distribution of class 0 (not used Cannabis within the last month) and 1 (used
# Cannabis within the last month) in the data set. We also aim to reduce
# test error by optimizing the decision boundary.
set.seed(12345)
shrinkage <- seq(from = 0.01, to = .5, by = .0049)
boost.test.err <- rep(0, length(shrinkage))
error <- rep(0, nrow(test.data))

for(i in 1:length(shrinkage)){
  boost <- gbm(Cannabis ~ ., data = train.data,
               distribution = 'bernoulli',
               n.trees = 200,
               shrinkage = shrinkage[i])

  pred.boost <- predict(boost,
                        n.trees=100,
```

```
                        newdata = test.data,
                        type = 'response')

  pclass.boost <- rep(NA, length(pred.boost))

  for(n in 1:length(pred.boost)){
    if(pred.boost[n] < 0.5){
      pclass.boost[n] = 0
    }else{
      pclass.boost[n] = 1
    }
  }

  for(e in 1:length(pclass.boost)){
    error[e] <- ((pclass.boost[e]) == test.data$Cannabis[e])
  }

  boost.test.err[i] = (length(error)-sum(error))/length(error)
}

df <- data.frame(shrinkage, boost.test.err)
ggplot(data = df, aes(x = shrinkage, y = boost.test.err)) +
  geom_point() +
  stat_smooth(method = "glm", formula = y ~ x + I(x^2), size = 1, col = "dark blue") +
  xlab("Shrinkage Parameters") +
  ylab("Misclassification Error")
```




```
shrinkage[which.min(boost.test.err)] # Use .23
```

```
## [1] 0.4608
```

```
min(boost.test.err) # 0.186
```

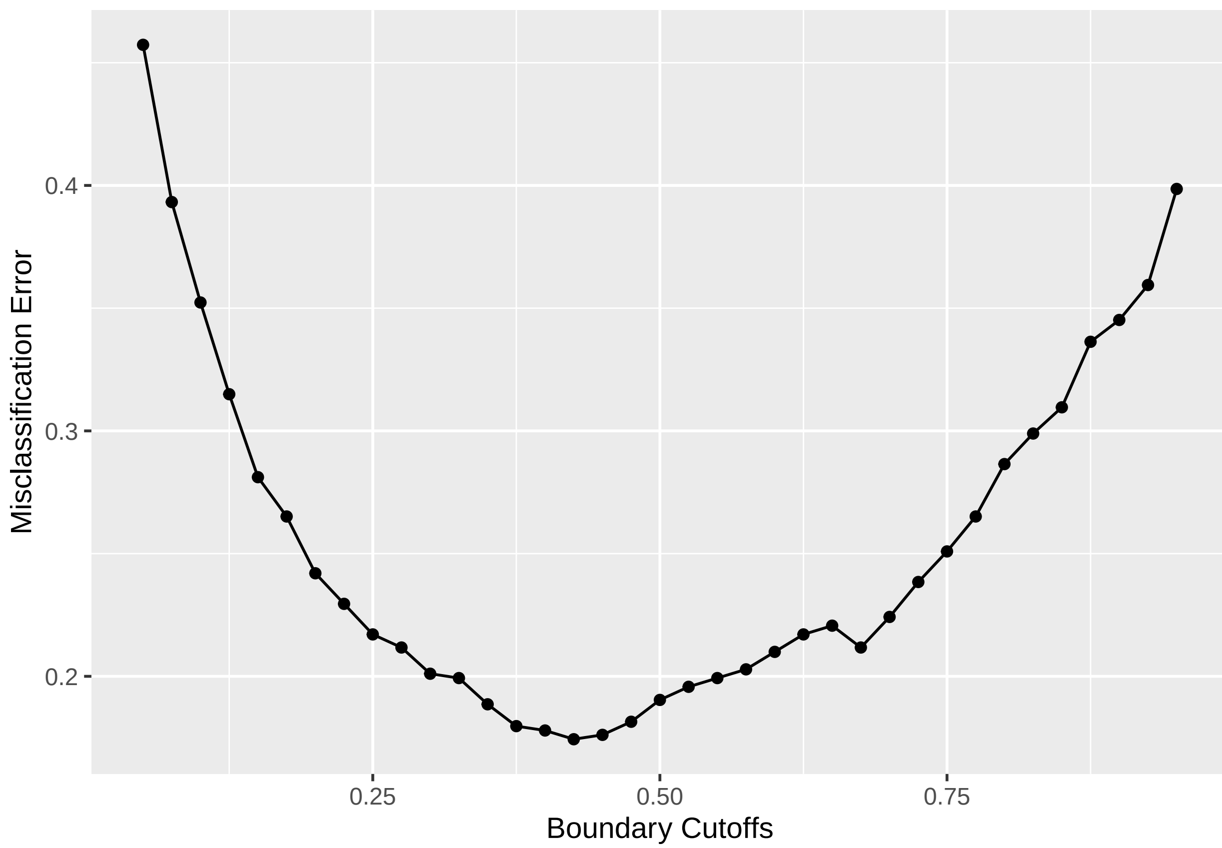
```
## [1] 0.1690391
```

```
# From this chart, we see that shrinkage coefficients between .01 and .5 are
# the ideal values. I will not use the shrinkage value with the lowest test
# error (0.4804) because it appears to be an outlier. I will stick within the
# ideal range and use the shrinkage value of 0.186 as it had a low test error
# and it is the approximate bottom of the regression line of test errors.
```

Boosting - Finding the optimal decision boundary

```
# Pick the ideal boundary cutoff using the ideal shrinkage value
# Plot the test MSE along different cutoff values of class 0/1
cutoffs <- seq(.05, .95, by = .025)
set.seed(12345)
boost.test.err <- rep(0, length(cutoffs))
error <- rep(0, nrow(test.data))

boost.2 <- gbm(Cannabis ~ ., data = train.data,
```

```r
             distribution = 'bernoulli',
             n.trees = 200,
             shrinkage = .23)

pred.boost.2 <- predict(boost.2,
                      n.trees=100,
                      newdata = test.data,
                      type = 'response')

pclass.boost.2 <- rep(NA, length(pred.boost.2))

for(i in 1:length(cutoffs)){
  pclass.boost.2 <- ifelse(pred.boost.2 < cutoffs[i], 0, 1)

  for(e in 1:length(pclass.boost.2)){
    error[e] <- (pclass.boost.2[e] == test.data$Cannabis[e])
  }
  boost.test.err[i] = (length(error)-sum(error))/length(error)

}

df <- data.frame(cutoffs, boost.test.err)
ggplot(data = df, aes(x = cutoffs, y = boost.test.err)) +
  geom_point() +
  geom_line() +
  xlab("Boundary Cutoffs") +
  ylab("Misclassification Error")
```




```
cutoffs[which.min(boost.test.err)] # 0.425
```

```
## [1] 0.425
```

```
min(boost.test.err) # 0.1743772
```

```
## [1] 0.1743772
```

Boosting - Combine ideal shrinkage coefficient and ideal cutoff value

```
set.seed(12345)
error <- rep(0, nrow(test.data))

boost <- gbm(Cannabis ~ ., data = train.data,
             distribution = 'bernoulli',
             n.trees = 500,
             shrinkage = 0.23)

pred.boost <- predict(boost,
                      n.trees=100,
                      newdata = test.data,
                      type = 'response')

pclass.boost <- ifelse(pred.boost < .425, 0, 1)
```

```
for(e in 1:length(pclass.boost)){
  error[e] <- (pclass.boost[e] == test.data$Cannabis[e])
}

boost.test.err = (length(error)-sum(error))/length(error)
boost.test.err # 0.1743772
```

```
## [1] 0.1743772
```

```
boost.success.rate <- 1 - boost.test.err
boost.success.rate # 0.8256228
```

```
## [1] 0.8256228
```

```
# This code runs the model using the optimized shrinkage parameter and boundary
# cutoff. We reached an error rate of 17.43%, or a success rate of 82.56%.
```

Logistic Regression

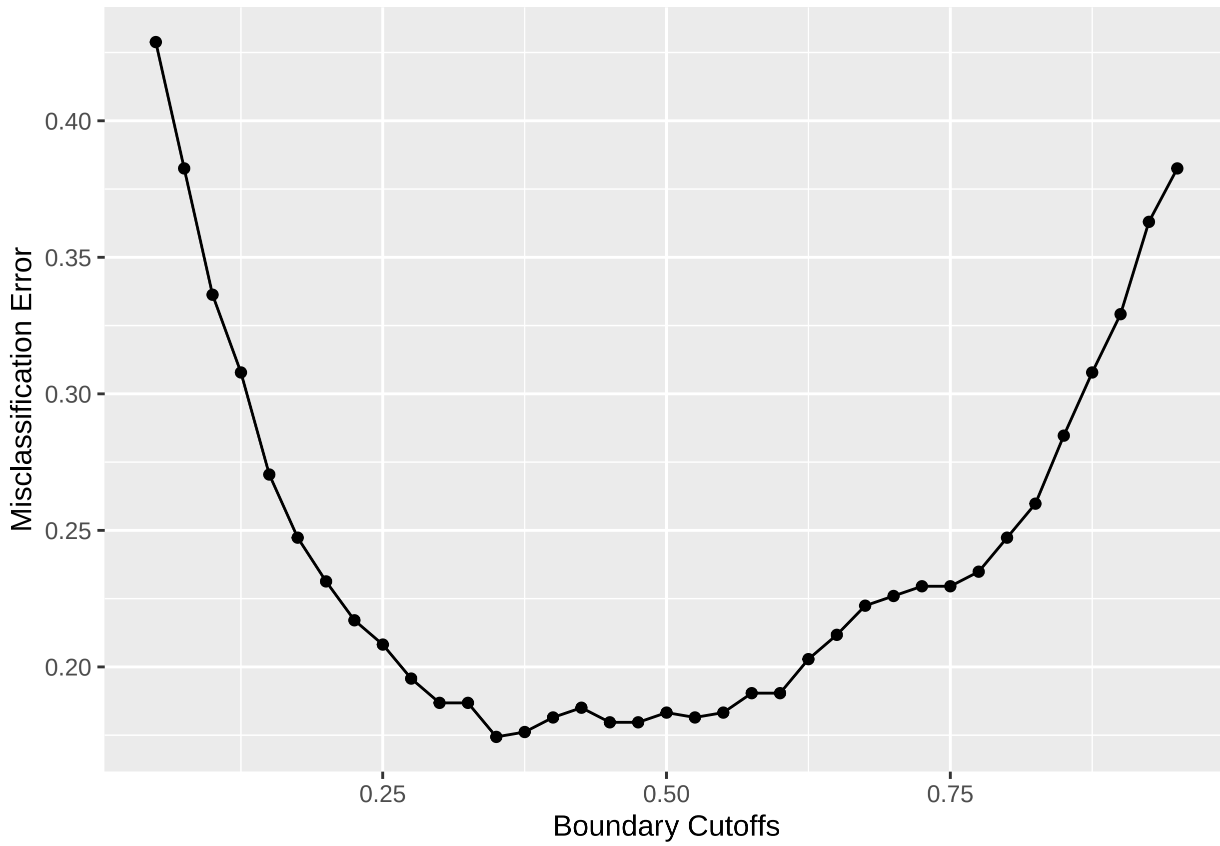
```
# In this code, we  use logistic regression to generate a binary prediction
# model to predict if an individual has used Cannabis within the last month.

# We will cycle through decision boundaries from 5% to 95% and calculate test
# error at each cutoff. This will be used to find the error-minimizing decision
# boundary of our model.

set.seed(12345)
log.fit <- glm(Cannabis ~ ., data = train.data, family = "binomial")
```

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
cutoffs <- seq(.05, .95, by = .025)
probs <- predict(log.fit, test.data, type = "response")
preds <- rep(0, length(probs))
error.log <- rep(0,length(probs))
log.test.err <- rep(NA, length(cutoffs))

for(i in 1:length(cutoffs)){
  preds <- ifelse(probs < cutoffs[i], 0, 1)

  for(e in 1:length(preds)){
    error.log[e] <- (preds[e] == test.data$Cannabis[e])
  }

  log.test.err[i] = (length(error.log)-sum(error.log))/length(error.log)
}

df <- data.frame(cutoffs, log.test.err)
```

```r
ggplot(data = df, aes(x = cutoffs, y = log.test.err)) +
  geom_point() +
  geom_line() +
  xlab("Boundary Cutoffs") +
  ylab("Misclassification Error")
```




```r
min(log.test.err) # 0.1761566
```

```
## [1] 0.1743772
```

```r
cutoffs[which.min(log.test.err)] # 0.35
```

```
## [1] 0.35
```

```r
# =====================================================
# Identified ideal cutoff at 0.325 Rerun logistic regression using the ideal
# cutoff and calculate the confusion matrix to see the false positive rate,
# false negative rate, and model accuracy.

log.fit <- glm(Cannabis ~ ., data = train.data, family = "binomial")
```

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
summary(log.fit)
```

```
##
## Call:
## glm(formula = Cannabis ~ ., family = "binomial", data = train.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7728  -0.5792  -0.2688   0.5949   2.4302
##
## Coefficients:
##                                                      Estimate Std. Error
## (Intercept)                                        -6.797e+13  9.079e+13
## Nscore                                             -1.995e-01  9.307e-02
## Escore                                             -9.156e-02  9.697e-02
## Oscore                                              5.252e-01  9.564e-02
## AScore                                              1.701e-03  8.564e-02
## Cscore                                             -8.587e-02  9.667e-02
## Impulsive                                          -1.218e-02  1.087e-01
## SS                                                  3.742e-01  1.151e-01
## Alcohol                                             5.020e-01  2.080e-01
## Caff                                                2.393e-01  3.447e-01
## Nicotine                                            1.175e+00  1.601e-01
## `18-24`                                             6.797e+13  9.079e+13
## `25-34`                                             6.797e+13  9.079e+13
## `35-44`                                             6.797e+13  9.079e+13
## `45-54`                                             6.797e+13  9.079e+13
## `55-64`                                             6.797e+13  9.079e+13
## `65`                                               -4.436e+15  9.079e+13
## `Left school at 16 years`                           5.997e-01  5.131e-01
## `Left school at 17 years`                           1.333e+00  7.342e-01
## `Left school at 18 years`                           8.872e-01  5.047e-01
## `Left school before 16 years`                       1.585e+00  7.303e-01
## `Masters degree`                                   -1.500e-01  4.113e-01
## `Professional certificate/ diploma`                 5.638e-01  4.174e-01
## `Some college or university, no certificate or degree`  4.859e-01  4.087e-01
## `University degree`                                -1.335e-01  3.925e-01
## Australia                                           9.315e-01  5.236e-01
## Canada                                              3.252e-02  4.267e-01
## `New Zealand`                                       1.863e+00  1.414e+00
## `Republic of Ireland`                              -2.967e-01  6.786e-01
## UK                                                 -9.892e-01  2.979e-01
## USA                                                 6.479e-01  3.030e-01
## Asian                                              -1.483e+00  9.593e-01
## Black                                              -1.018e+00  8.492e-01
## `Mixed-Black/Asian`                                 2.555e+01  2.161e+05
## `Mixed-White/Asian`                                 5.387e-01  8.629e-01
## `Mixed-White/Black`                                -5.882e-01  9.162e-01
## White                                              -4.016e-01  4.088e-01
## Gender                                              6.471e-01  1.661e-01
##                                                    z value Pr(>|z|)
## (Intercept)                                         -0.749 0.454066
## Nscore                                              -2.144 0.032047 *
```

```
## Escore                                                      -0.944 0.345052
## Oscore                                                       5.492 3.98e-08 ***
## AScore                                                       0.020 0.984158
## Cscore                                                      -0.888 0.374375
## Impulsive                                                   -0.112 0.910797
## SS                                                           3.253 0.001143 **
## Alcohol                                                      2.414 0.015792 *
## Caff                                                         0.694 0.487549
## Nicotine                                                     7.337 2.19e-13 ***
## `18-24`                                                      0.749 0.454066
## `25-34`                                                      0.749 0.454066
## `35-44`                                                      0.749 0.454066
## `45-54`                                                      0.749 0.454066
## `55-64`                                                      0.749 0.454066
## `65`                                                       -48.856  < 2e-16 ***
## `Left school at 16 years`                                    1.169 0.242562
## `Left school at 17 years`                                    1.816 0.069438 .
## `Left school at 18 years`                                    1.758 0.078784 .
## `Left school before 16 years`                                2.171 0.029949 *
## `Masters degree`                                            -0.365 0.715286
## `Professional certificate/ diploma`                          1.351 0.176837
## `Some college or university, no certificate or degree`       1.189 0.234541
## `University degree`                                         -0.340 0.733867
## Australia                                                    1.779 0.075241 .
## Canada                                                       0.076 0.939248
## `New Zealand`                                                1.318 0.187582
## `Republic of Ireland`                                       -0.437 0.661984
## UK                                                          -3.321 0.000898 ***
## USA                                                          2.138 0.032505 *
## Asian                                                       -1.546 0.122194
## Black                                                       -1.199 0.230713
## `Mixed-Black/Asian`                                          0.000 0.999906
## `Mixed-White/Asian`                                          0.624 0.532427
## `Mixed-White/Black`                                         -0.642 0.520857
## White                                                       -0.982 0.325891
## Gender                                                       3.896 9.78e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1782.5  on 1313  degrees of freedom
## Residual deviance: 1069.9  on 1276  degrees of freedom
## AIC: 1145.9
##
## Number of Fisher Scoring iterations: 25
```

```r
probs <- predict(log.fit, test.data, type = "response")
preds <- rep(0, length(probs))
preds[probs > 0.35] = 1

preds <- as.factor(preds)
test.data$Cannabis <- as.factor(test.data$Cannabis)
confusionMatrix(test.data$Cannabis, preds) #  82.38%
```

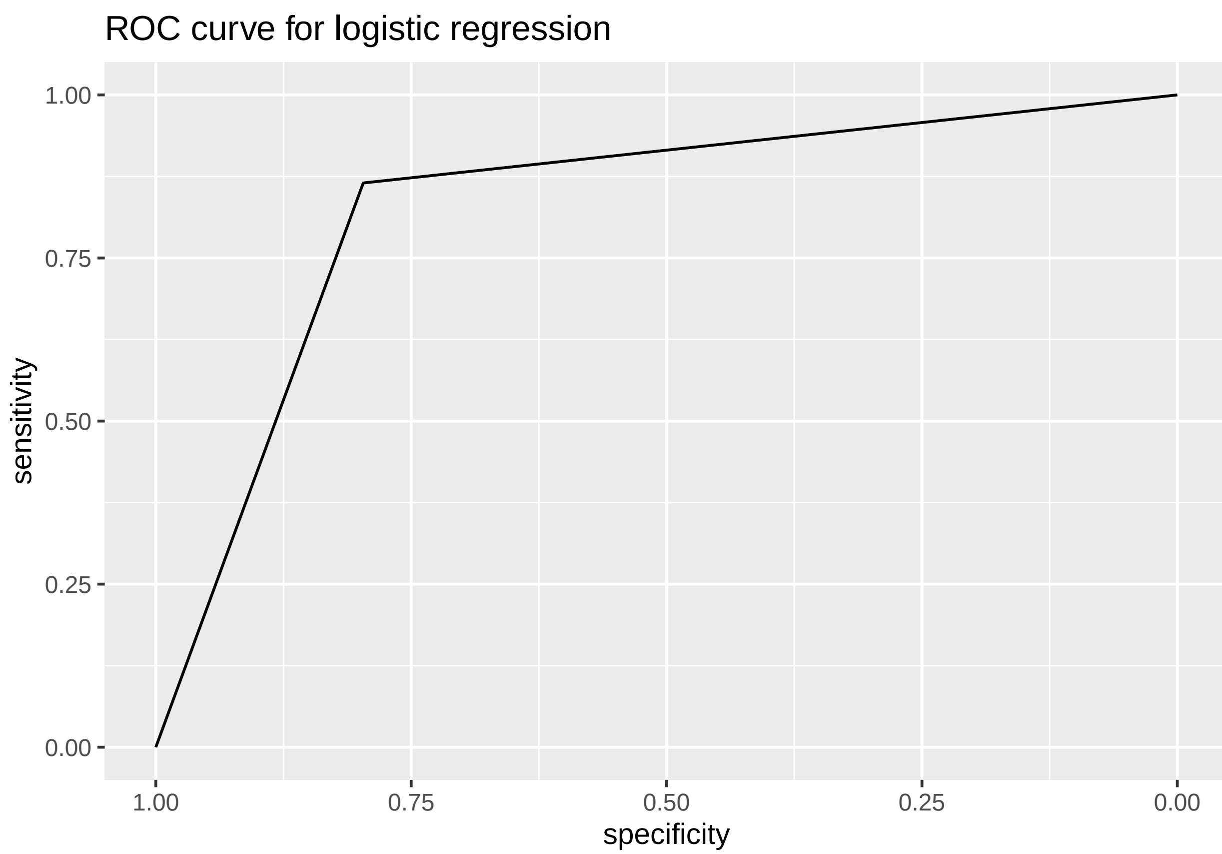
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 259  66
##          1  32 205
##
##                Accuracy : 0.8256
##                  95% CI : (0.7917, 0.8561)
##     No Information Rate : 0.5178
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6493
##
##  Mcnemar's Test P-Value : 0.0008576
##
##             Sensitivity : 0.8900
##             Specificity : 0.7565
##          Pos Pred Value : 0.7969
##          Neg Pred Value : 0.8650
##              Prevalence : 0.5178
##          Detection Rate : 0.4609
##    Detection Prevalence : 0.5783
##       Balanced Accuracy : 0.8232
##
##        'Positive' Class : 0
##
```

```r
# Accuracy of 82.38%
# FPR = 5.69395%
# FNR = 11.92171%

#ROC-curve using pROC library
test.data$Cannabis <- as.numeric(test.data$Cannabis)
preds <- as.numeric(preds)
ROC.score.log <- roc(test.data$Cannabis, preds)
```

```
## Setting levels: control = 1, case = 2
```

```
## Setting direction: controls < cases
```

```r
ggroc(ROC.score.log, legacy.axes = FALSE) +
  ggtitle("ROC curve for logistic regression")
```



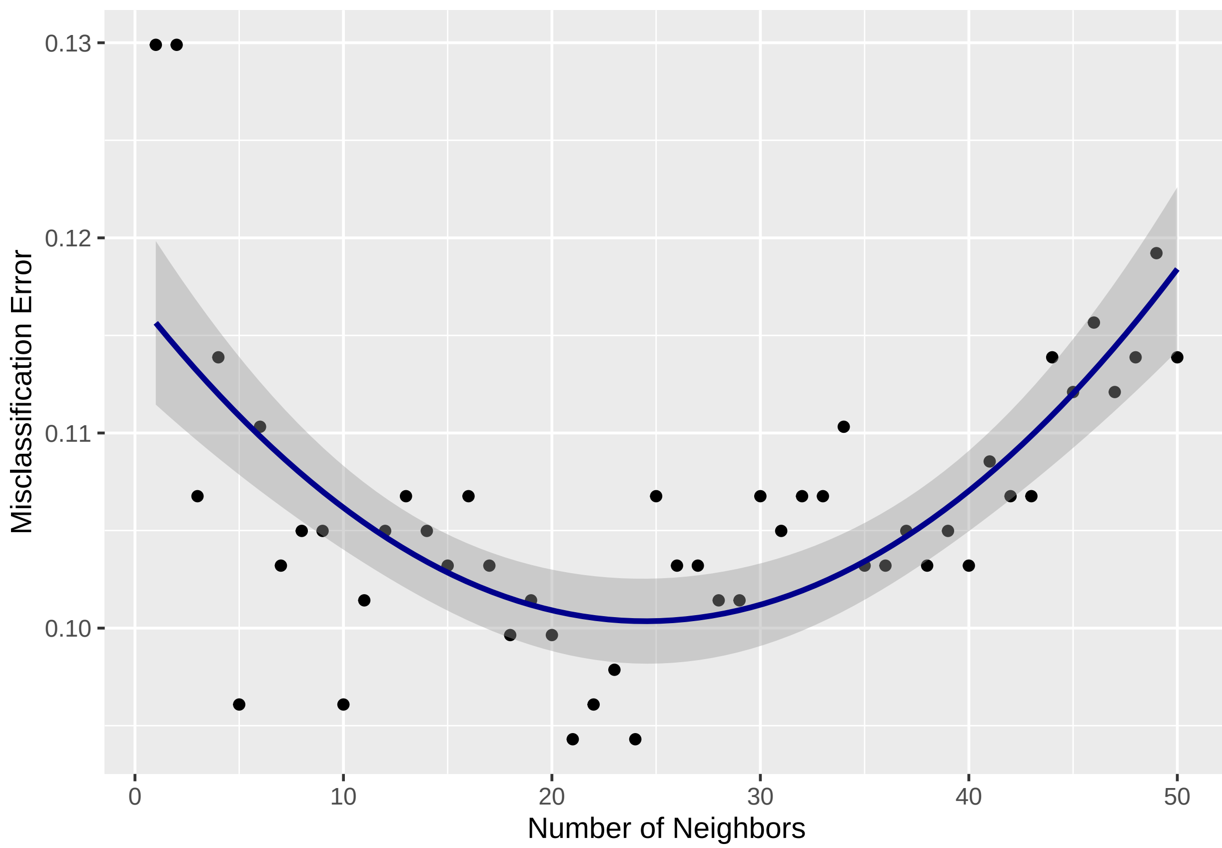
RECLEAN DATA

```
set.seed(12345)
data <- read.csv("Drug_Consumption.csv")

# Remove the over-claimers using the control drug "Semer"
data <- subset(data, data$Semer == "CL0")

for(i in 14:ncol(data)){
  data[,i] <- as.numeric(data[, i] == "CL4" | data[, i] == "CL5" | data[, i] == "CL6")
}

# Drop 65+
data <- data %>% mutate(dummy=1) %>%
spread(key=Age,value=dummy,fill=0)

# Drop Doctorate
data <- data %>% mutate(dummy=1) %>%
spread(key=Education,value=dummy,fill=0)

# Drop other
data <- data %>% mutate(dummy=1) %>%
spread(key=Country,value=dummy,fill=0)

# Drop other
data <- data %>% mutate(dummy=1) %>%
```

```r
spread(key=Ethnicity,value=dummy,fill=0)

# Drop 'F' variable and rename to gender
data <- data %>% mutate(dummy=1) %>%
spread(key=Gender,value=dummy, fill=0)

# Drop variables that we aren't using.
drop <- c("ID", "65+","Doctorate degree","Other","F","Amphet","Amyl","Benzos","Choc","Crack", "Coke","E
data <- data[,!(names(data) %in% drop)]

names(data)[names(data) == "M"] <- "Gender"

# Split into test and train data
test.i <- sample(1:nrow(data), .3*nrow(data))
test.data <- data[test.i,]
train.data <- data[-test.i,]
```

kNN

```r
set.seed(12345)
ks <- 1:50
knn.error <- rep(0, length(ks))

for(i in 1:length(ks)){
  pred.knn <- knn(train.data, test.data, train.data$Cannabis, k = ks[i])
  table.knn <- table(pred.knn, test.data$Cannabis)
  knn.error[i] <- (table.knn[1,2] + table.knn[2,1])/(table.knn[1,2] + table.knn[2,1] + table.knn[2,2] +
}

df.knn = data.frame(ks, knn.error)
ggplot(data = df.knn, aes(x = ks, y = knn.error)) +
  geom_point() +
  stat_smooth(method = "glm", formula = y ~ x + I(x^2), size = 1, col = "dark blue") +
  xlab("Number of Neighbors") +
  ylab("Misclassification Error")
```




```
which.min(knn.error) # k = 22 results in the minimum error
```

```
## [1] 21
```

```
min(knn.error) # 0.09252669, or a success rate of 90.74733%
```

```
## [1] 0.09430605
```

SVM

```
x_SVM <- train.data[,-10]
y_SVM <- train.data[,10]
SVM_data <- data.frame(x = x_SVM, y = as.factor(y_SVM))
SVM_model <- svm(y~., data = SVM_data, kernel = "linear", scale = FALSE, cost = 10)
SVM_model
```

```
##
## Call:
## svm(formula = y ~ ., data = SVM_data, kernel = "linear", cost = 10,
##     scale = FALSE)
##
##
## Parameters:
##    SVM-Type:  C-classification
```

```
##  SVM-Kernel:  linear
##        cost:  10
##
## Number of Support Vectors:  588
```

```
SVM_predict <- predict(SVM_model, data.frame(x = test.data[,-10], y = test.data[,10]))
# Ideal cost is 1.92875e-22
# minimum error is 0.192923

table.SVM <- table(SVM_predict, test.data$Cannabis)
table.SVM
```

```
##
## SVM_predict   0   1
##           0 276  51
##           1  49 186
```

```
(table.SVM[1,2] + table.SVM[2,1])/(table.SVM[1,2] + table.SVM[2,1] + table.SVM[1,1] + table.SVM[2,2])
```

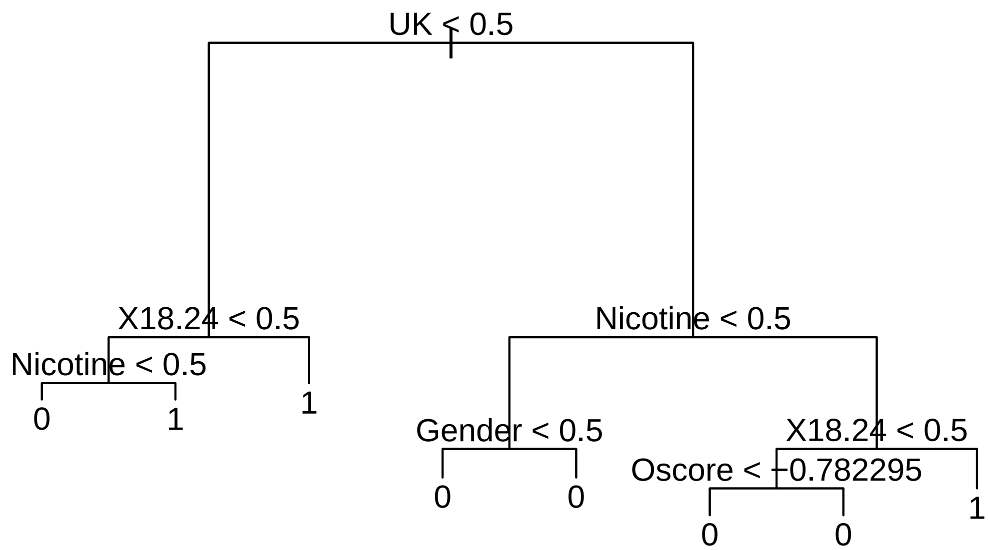
```
## [1] 0.1779359
```

Decision Trees

```
set.seed(12345)
tree_train <- data.frame(train.data)
tree_test <- data.frame(test.data)
treefit <- tree(as.factor(Cannabis)~. ,data = tree_train)
summary(treefit)
```

```
##
## Classification tree:
## tree(formula = as.factor(Cannabis) ~ ., data = tree_train)
## Variables actually used in tree construction:
## [1] "UK"       "X18.24"   "Nicotine" "Gender"   "Oscore"
## Number of terminal nodes:  8
## Residual mean deviance:  0.8736 = 1141 / 1306
## Misclassification error rate: 0.2032 = 267 / 1314
```

```
# variables used : UK, 18-24, Oscore, Nicotine, gender, and SS
plot(treefit)
text(treefit)
```




```
tree.predict <- predict(treefit, tree_test, type = "class")
tree.table <-table(tree.predict, tree_test$Cannabis)
tree.error <- (tree.table[1,2] + tree.table[2,1])/(tree.table[1,2] + tree.table[2,1] + tree.table[1,1] +
tree.error # 0.1992883
```

```
## [1] 0.1992883
```

Random Forest

```
set.seed(12345)
rF <- randomForest(as.factor(Cannabis)~., data = tree_train, importance = TRUE)
rf.predict <- predict(rF, tree_test)
rf.table <-table(rf.predict, tree_test$Cannabis) # .2009 error rate
rf.table
```

```
##
## rf.predict   0   1
##          0 272  47
##          1  53 190
```

```
rf.error <- (rf.table[1,2] + rf.table[2,1])/(rf.table[1,2] + rf.table[2,1] + rf.table[1,1] + rf.table[2
rf.error # 0.1814947
```

```
## [1] 0.1779359
```

LDA

```r
set.seed(12345)
lda.fit <- lda(as.factor(Cannabis)~., data = train.data)
```

```
## Warning in lda.default(x, grouping, ...): variables are collinear
```

```r
lda.fit
```

```
## Call:
## lda(as.factor(Cannabis) ~ ., data = train.data)
##
## Prior probabilities of groups:
##        0        1
## 0.585997 0.414003
##
## Group means:
##        Nscore        Escore     Oscore      AScore     Cscore  Impulsive
## 0 -0.05089368 -0.0002935974 -0.2980838  0.07746401  0.1988430 -0.2055659
## 1  0.02314386  0.0304835294  0.4436558 -0.12047708 -0.2295484  0.3013005
##           SS   Alcohol      Caff  Nicotine   ‘18-24‘   ‘25-34‘   ‘35-44‘
## 0 -0.3303218 0.8155844 0.9259740 0.2974026 0.1701299 0.2844156 0.2428571
## 1  0.4434787 0.8400735 0.9522059 0.6875000 0.5588235 0.2150735 0.1213235
##      ‘45-54‘    ‘55-64‘       ‘65‘ ‘Left school at 16 years‘
## 0 0.22597403 0.05844156 0.01818182                0.06363636
## 1 0.07904412 0.02573529 0.00000000                0.03492647
##   ‘Left school at 17 years‘ ‘Left school at 18 years‘
## 0                0.01038961                0.03246753
## 1                0.02389706                0.07536765
##   ‘Left school before 16 years‘ ‘Masters degree‘
## 0                    0.01168831       0.20259740
## 1                    0.01654412       0.08639706
##   ‘Professional certificate/ diploma‘
## 0                           0.1649351
## 1                           0.1286765
##   ‘Some college or university, no certificate or degree‘ ‘University degree‘
## 0                                              0.1441558           0.3090909
## 1                                              0.4283088           0.1764706
##    Australia     Canada ‘New Zealand‘ ‘Republic of Ireland‘        UK       USA
## 0 0.01558442 0.03896104   0.001298701           0.009090909 0.7597403 0.1298701
## 1 0.04411765 0.05147059   0.003676471           0.016544118 0.2591912 0.5275735
##         Asian      Black ‘Mixed-Black/Asian‘ ‘Mixed-White/Asian‘
## 0 0.023376623 0.02597403         0.000000000         0.009090909
## 1 0.003676471 0.01102941         0.005514706         0.016544118
##   ‘Mixed-White/Black‘     White    Gender
## 0         0.007792208 0.9142857 0.3805195
## 1         0.011029412 0.8897059 0.6801471
##
## Coefficients of linear discriminants:
##                                                                 LD1
## Nscore                                                 -0.128719708
## Escore                                                 -0.063894941
## Oscore                                                  0.309557993
```

```
## AScore                                                   -0.009474028
## Cscore                                                   -0.044445079
## Impulsive                                                -0.006703214
## SS                                                        0.205778013
## Alcohol                                                   0.273200620
## Caff                                                      0.099991319
## Nicotine                                                  0.747972175
## `18-24`                                                   0.528292204
## `25-34`                                                  -0.030754587
## `35-44`                                                  -0.171371758
## `45-54`                                                  -0.343321680
## `55-64`                                                  -0.197907859
## `65`                                                     -1.129097364
## `Left school at 16 years`                                 0.388558815
## `Left school at 17 years`                                 0.814517525
## `Left school at 18 years`                                 0.522752868
## `Left school before 16 years`                             0.944261957
## `Masters degree`                                         -0.092648764
## `Professional certificate/ diploma`                       0.347386663
## `Some college or university, no certificate or degree`  0.344774309
## `University degree`                                      -0.038493892
## Australia                                                 0.476439450
## Canada                                                   -0.080914690
## `New Zealand`                                             1.228977238
## `Republic of Ireland`                                    -0.174970800
## UK                                                       -0.756964941
## USA                                                       0.402876112
## Asian                                                    -0.649234781
## Black                                                    -0.407009859
## `Mixed-Black/Asian`                                       0.989634489
## `Mixed-White/Asian`                                       0.109803221
## `Mixed-White/Black`                                      -0.393579887
## White                                                    -0.282101194
## Gender                                                    0.380486085
```

```
lda.pred <- predict(lda.fit, test.data)$class
table.lda <- table(lda.pred, test.data$Cannabis)

lda.error <- (table.lda[1,2] + table.lda[2,1])/(table.lda[1,2] + table.lda[2,1] + table.lda[1,1] + table
lda.error # 0.1886121
```

```
## [1] 0.1886121
```