## Module 3 Assignment 2 - Classification with Logistic Regression

#install.packages("ROCR")  
#install.packages("e1071")

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.1.0 v purrr 0.3.0  
## v tibble 2.0.1 v dplyr 0.7.8  
## v tidyr 0.8.2 v stringr 1.3.1  
## v readr 1.3.1 v forcats 0.3.0

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(MASS)

##   
## Attaching package: 'MASS'

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

library(caret)

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

library(ROCR)

## Loading required package: gplots

##   
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':  
##   
## lowess

library(GGally)

##   
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':  
##   
## nasa

library(e1071)

parole <- read\_csv("parole.csv")

## Parsed with column specification:  
## cols(  
## male = col\_double(),  
## race = col\_double(),  
## age = col\_double(),  
## state = col\_double(),  
## time.served = col\_double(),  
## max.sentence = col\_double(),  
## multiple.offenses = col\_double(),  
## crime = col\_double(),  
## violator = col\_double()  
## )

View(parole)

str(parole)

## Classes 'spec\_tbl\_df', 'tbl\_df', 'tbl' and 'data.frame': 675 obs. of 9 variables:  
## $ male : num 1 0 1 1 1 1 1 0 0 1 ...  
## $ race : num 1 1 2 1 2 2 1 1 1 2 ...  
## $ age : num 33.2 39.7 29.5 22.4 21.6 46.7 31 24.6 32.6 29.1 ...  
## $ state : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ time.served : num 5.5 5.4 5.6 5.7 5.4 6 6 4.8 4.5 4.7 ...  
## $ max.sentence : num 18 12 12 18 12 18 18 12 13 12 ...  
## $ multiple.offenses: num 0 0 0 0 0 0 0 0 0 0 ...  
## $ crime : num 4 3 3 1 1 4 3 1 3 2 ...  
## $ violator : num 0 0 0 0 0 0 0 0 0 0 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. male = col\_double(),  
## .. race = col\_double(),  
## .. age = col\_double(),  
## .. state = col\_double(),  
## .. time.served = col\_double(),  
## .. max.sentence = col\_double(),  
## .. multiple.offenses = col\_double(),  
## .. crime = col\_double(),  
## .. violator = col\_double()  
## .. )

summary(parole)

## male race age state   
## Min. :0.0000 Min. :1.000 Min. :18.40 Min. :1.000   
## 1st Qu.:1.0000 1st Qu.:1.000 1st Qu.:25.35 1st Qu.:2.000   
## Median :1.0000 Median :1.000 Median :33.70 Median :3.000   
## Mean :0.8074 Mean :1.424 Mean :34.51 Mean :2.887   
## 3rd Qu.:1.0000 3rd Qu.:2.000 3rd Qu.:42.55 3rd Qu.:4.000   
## Max. :1.0000 Max. :2.000 Max. :67.00 Max. :4.000   
## time.served max.sentence multiple.offenses crime   
## Min. :0.000 Min. : 1.00 Min. :0.0000 Min. :1.000   
## 1st Qu.:3.250 1st Qu.:12.00 1st Qu.:0.0000 1st Qu.:1.000   
## Median :4.400 Median :12.00 Median :1.0000 Median :2.000   
## Mean :4.198 Mean :13.06 Mean :0.5363 Mean :2.059   
## 3rd Qu.:5.200 3rd Qu.:15.00 3rd Qu.:1.0000 3rd Qu.:3.000   
## Max. :6.000 Max. :18.00 Max. :1.0000 Max. :4.000   
## violator   
## Min. :0.0000   
## 1st Qu.:0.0000   
## Median :0.0000   
## Mean :0.1156   
## 3rd Qu.:0.0000   
## Max. :1.0000

parole = parole %>% mutate(male = as\_factor(as.character(male))) %>%  
mutate(male = fct\_recode(male,  
"female" = "0",  
"male" = "1"))%>%  
   
 mutate(age = as.numeric(as.character(age)))%>%  
   
mutate(time.served = as.numeric(as.character(time.served)))%>%   
   
 mutate(max.sentence = as\_factor(as.character(max.sentence)))%>%   
   
 mutate(race = as\_factor(as.character(race)))%>%  
   
mutate(race = fct\_recode(race,  
"white" = "1",  
"otherwise" = "2"))%>%   
   
 mutate(multiple.offenses = as\_factor(as.character(multiple.offenses)))%>%  
   
mutate(multiple.offenses = fct\_recode(multiple.offenses,  
"otherwise" = "0",  
"multipleoffenses" = "1"))%>%  
   
 mutate(state = as\_factor(as.character(state))) %>%  
   
mutate(state = fct\_recode(state,  
"anyotherstate" = "1",  
"Kentucky" = "2",  
"Louisiana" = "3",  
"Virginia" = "4"))%>%  
   
 mutate(crime = as\_factor(as.character(crime))) %>%  
   
   
mutate(crime = fct\_recode(crime,  
"othercrime" = "1",  
"larceny" = "2",  
"drug-relatedcrime" = "3",  
"driving-relatedcrime" = "4"))%>%   
   
 mutate(violator = as\_factor(as.character(violator)))%>%  
   
mutate(violator = fct\_recode(violator,  
"violatedparole" = "1",  
"noparoleviolation" = "0"))

str(parole)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 675 obs. of 9 variables:  
## $ male : Factor w/ 2 levels "male","female": 1 2 1 1 1 1 1 2 2 1 ...  
## $ race : Factor w/ 2 levels "white","otherwise": 1 1 2 1 2 2 1 1 1 2 ...  
## $ age : num 33.2 39.7 29.5 22.4 21.6 46.7 31 24.6 32.6 29.1 ...  
## $ state : Factor w/ 4 levels "anyotherstate",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ time.served : num 5.5 5.4 5.6 5.7 5.4 6 6 4.8 4.5 4.7 ...  
## $ max.sentence : Factor w/ 17 levels "18","12","13",..: 1 2 2 1 2 1 1 2 3 2 ...  
## $ multiple.offenses: Factor w/ 2 levels "otherwise","multipleoffenses": 1 1 1 1 1 1 1 1 1 1 ...  
## $ crime : Factor w/ 4 levels "driving-relatedcrime",..: 1 2 2 3 3 1 2 3 2 4 ...  
## $ violator : Factor w/ 2 levels "noparoleviolation",..: 1 1 1 1 1 1 1 1 1 1 ...

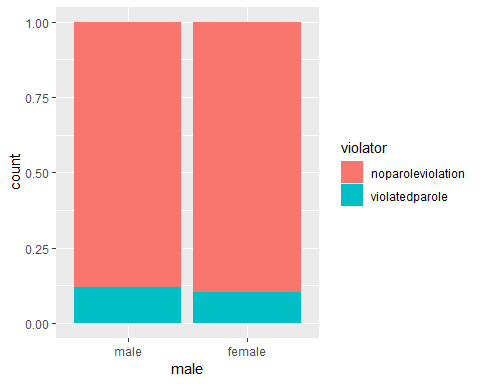
summary(parole)

## male race age state   
## male :545 white :389 Min. :18.40 anyotherstate:143   
## female:130 otherwise:286 1st Qu.:25.35 Kentucky :120   
## Median :33.70 Louisiana : 82   
## Mean :34.51 Virginia :330   
## 3rd Qu.:42.55   
## Max. :67.00   
##   
## time.served max.sentence multiple.offenses  
## Min. :0.000 12 :274 otherwise :313   
## 1st Qu.:3.250 18 : 78 multipleoffenses:362   
## Median :4.400 13 : 77   
## Mean :4.198 14 : 66   
## 3rd Qu.:5.200 15 : 60   
## Max. :6.000 16 : 36   
## (Other): 84   
## crime violator   
## driving-relatedcrime:101 noparoleviolation:597   
## drug-relatedcrime :153 violatedparole : 78   
## othercrime :315   
## larceny :106   
##   
##   
##

set.seed(12345)  
train.rows = createDataPartition(y = parole$violator, p=0.7, list = FALSE) #70% in training  
train = parole[train.rows,]   
test = parole[-train.rows,]

Male

ggplot(train,aes(x=male, fill = violator)) + geom\_bar(position = "fill")

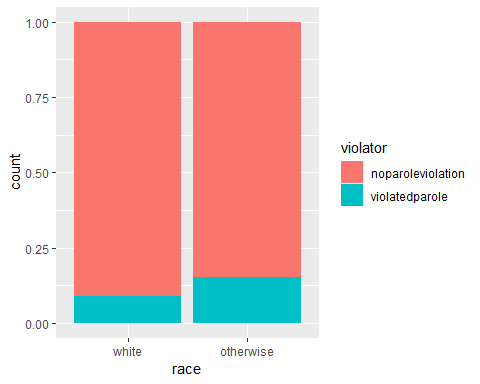


t1 = table(parole$violator, parole$male)  
prop.table(t1, margin = 2)

##   
## male female  
## noparoleviolation 0.8825688 0.8923077  
## violatedparole 0.1174312 0.1076923

Race

ggplot(train,aes(x=race, fill = violator)) + geom\_bar(position = "fill")

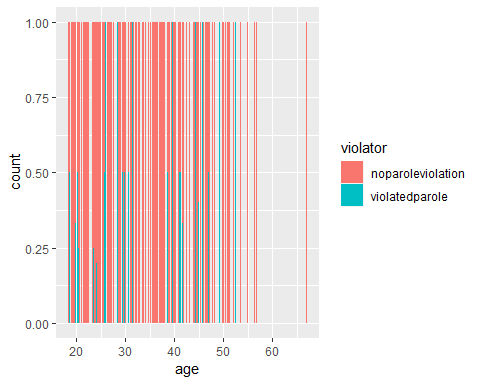


t2 = table(parole$violator, parole$race)  
prop.table(t2, margin = 2)

##   
## white otherwise  
## noparoleviolation 0.90488432 0.85664336  
## violatedparole 0.09511568 0.14335664

Age

ggplot(train,aes(x=age, fill = violator)) + geom\_bar(position = "fill")

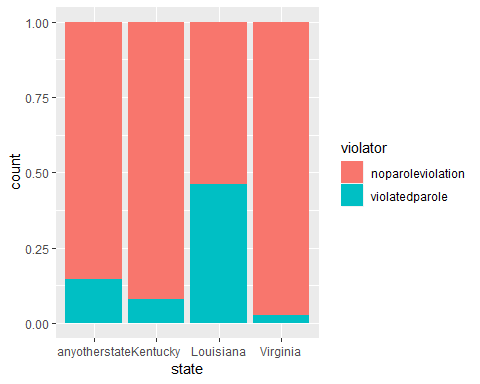


t3 = table(parole$violator, parole$age)  
prop.table(t3, margin = 2)

##   
## 18.4 18.5 18.7 18.8 19  
## noparoleviolation 1.0000000 1.0000000 0.5000000 1.0000000 1.0000000  
## violatedparole 0.0000000 0.0000000 0.5000000 0.0000000 0.0000000  
##   
## 19.1 19.2 19.3 19.4 19.5  
## noparoleviolation 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000  
## violatedparole 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  
##   
## 19.6 19.7 19.9 20 20.2  
## noparoleviolation 1.0000000 1.0000000 0.6666667 1.0000000 0.7500000  
## violatedparole 0.0000000 0.0000000 0.3333333 0.0000000 0.2500000  
##   
## 20.3 20.4 20.5 20.6 20.7  
## noparoleviolation 0.5000000 1.0000000 0.7500000 0.6666667 1.0000000  
## violatedparole 0.5000000 0.0000000 0.2500000 0.3333333 0.0000000  
##   
## 20.8 20.9 21 21.1 21.2  
## noparoleviolation 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000  
## violatedparole 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  
##   
## 21.3 21.4 21.5 21.6 21.7  
## noparoleviolation 1.0000000 1.0000000 1.0000000 0.5000000 0.5000000  
## violatedparole 0.0000000 0.0000000 0.0000000 0.5000000 0.5000000  
##   
## 21.8 21.9 22 22.1 22.2  
## noparoleviolation 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000  
## violatedparole 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  
##   
## 22.3 22.4 22.5 22.6 22.8  
## noparoleviolation 1.0000000 0.6666667 1.0000000 1.0000000 0.6666667  
## violatedparole 0.0000000 0.3333333 0.0000000 0.0000000 0.3333333  
##   
## 22.9 23 23.1 23.2 23.3  
## noparoleviolation 1.0000000 1.0000000 1.0000000 1.0000000 0.8333333  
## violatedparole 0.0000000 0.0000000 0.0000000 0.0000000 0.1666667  
##   
## 23.4 23.6 23.7 23.8 24  
## noparoleviolation 1.0000000 0.8000000 0.6666667 1.0000000 1.0000000  
## violatedparole 0.0000000 0.2000000 0.3333333 0.0000000 0.0000000  
##   
## 24.2 24.3 24.4 24.5 24.6  
## noparoleviolation 0.8333333 1.0000000 1.0000000 1.0000000 1.0000000  
## violatedparole 0.1666667 0.0000000 0.0000000 0.0000000 0.0000000  
##   
## 24.7 24.8 24.9 25 25.1  
## noparoleviolation 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000  
## violatedparole 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  
##   
## 25.2 25.3 25.4 25.5 25.6  
## noparoleviolation 1.0000000 0.7500000 1.0000000 1.0000000 0.8571429  
## violatedparole 0.0000000 0.2500000 0.0000000 0.0000000 0.1428571  
##   
## 25.7 25.8 25.9 26 26.3  
## noparoleviolation 1.0000000 0.5000000 1.0000000 0.0000000 1.0000000  
## violatedparole 0.0000000 0.5000000 0.0000000 1.0000000 0.0000000  
##   
## 26.4 26.5 26.6 26.8 26.9  
## noparoleviolation 1.0000000 1.0000000 1.0000000 0.6666667 1.0000000  
## violatedparole 0.0000000 0.0000000 0.0000000 0.3333333 0.0000000  
##   
## 27 27.1 27.2 27.3 27.4  
## noparoleviolation 1.0000000 1.0000000 1.0000000 1.0000000 0.0000000  
## violatedparole 0.0000000 0.0000000 0.0000000 0.0000000 1.0000000  
##   
## 27.5 27.6 27.7 27.8 27.9  
## noparoleviolation 0.5000000 1.0000000 1.0000000 1.0000000 1.0000000  
## violatedparole 0.5000000 0.0000000 0.0000000 0.0000000 0.0000000  
##   
## 28 28.1 28.2 28.3 28.4  
## noparoleviolation 1.0000000 0.6666667 1.0000000 1.0000000 0.6666667  
## violatedparole 0.0000000 0.3333333 0.0000000 0.0000000 0.3333333  
##   
## 28.5 28.7 28.8 28.9 29  
## noparoleviolation 0.5000000 1.0000000 0.8000000 0.6666667 1.0000000  
## violatedparole 0.5000000 0.0000000 0.2000000 0.3333333 0.0000000  
##   
## 29.1 29.2 29.5 29.6 29.7  
## noparoleviolation 1.0000000 1.0000000 0.6666667 1.0000000 1.0000000  
## violatedparole 0.0000000 0.0000000 0.3333333 0.0000000 0.0000000  
##   
## 29.9 30 30.1 30.2 30.3  
## noparoleviolation 0.2500000 1.0000000 1.0000000 1.0000000 1.0000000  
## violatedparole 0.7500000 0.0000000 0.0000000 0.0000000 0.0000000  
##   
## 30.4 30.7 30.8 31 31.1  
## noparoleviolation 1.0000000 0.5000000 0.6666667 0.8333333 1.0000000  
## violatedparole 0.0000000 0.5000000 0.3333333 0.1666667 0.0000000  
##   
## 31.2 31.3 31.4 31.5 31.6  
## noparoleviolation 1.0000000 1.0000000 0.5000000 0.5000000 1.0000000  
## violatedparole 0.0000000 0.0000000 0.5000000 0.5000000 0.0000000  
##   
## 31.7 31.8 32 32.1 32.2  
## noparoleviolation 1.0000000 1.0000000 1.0000000 1.0000000 0.6666667  
## violatedparole 0.0000000 0.0000000 0.0000000 0.0000000 0.3333333  
##   
## 32.3 32.4 32.5 32.6 32.7  
## noparoleviolation 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000  
## violatedparole 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  
##   
## 32.8 32.9 33 33.2 33.3  
## noparoleviolation 0.7500000 1.0000000 1.0000000 1.0000000 0.0000000  
## violatedparole 0.2500000 0.0000000 0.0000000 0.0000000 1.0000000  
##   
## 33.4 33.5 33.6 33.7 33.8  
## noparoleviolation 1.0000000 0.5000000 1.0000000 0.5000000 1.0000000  
## violatedparole 0.0000000 0.5000000 0.0000000 0.5000000 0.0000000  
##   
## 33.9 34 34.1 34.2 34.3  
## noparoleviolation 1.0000000 1.0000000 0.0000000 0.6666667 1.0000000  
## violatedparole 0.0000000 0.0000000 1.0000000 0.3333333 0.0000000  
##   
## 34.4 34.5 34.6 34.7 34.8  
## noparoleviolation 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000  
## violatedparole 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  
##   
## 34.9 35 35.1 35.2 35.3  
## noparoleviolation 0.6666667 1.0000000 1.0000000 1.0000000 1.0000000  
## violatedparole 0.3333333 0.0000000 0.0000000 0.0000000 0.0000000  
##   
## 35.4 35.5 35.6 35.8 35.9  
## noparoleviolation 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000  
## violatedparole 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  
##   
## 36 36.1 36.2 36.3 36.4  
## noparoleviolation 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000  
## violatedparole 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  
##   
## 36.5 36.6 36.7 36.8 37  
## noparoleviolation 0.8000000 1.0000000 1.0000000 1.0000000 1.0000000  
## violatedparole 0.2000000 0.0000000 0.0000000 0.0000000 0.0000000  
##   
## 37.2 37.3 37.4 37.5 37.6  
## noparoleviolation 0.6666667 0.5000000 0.5000000 1.0000000 1.0000000  
## violatedparole 0.3333333 0.5000000 0.5000000 0.0000000 0.0000000  
##   
## 37.8 38 38.1 38.2 38.3  
## noparoleviolation 1.0000000 1.0000000 1.0000000 1.0000000 0.5000000  
## violatedparole 0.0000000 0.0000000 0.0000000 0.0000000 0.5000000  
##   
## 38.4 38.5 38.6 38.7 38.8  
## noparoleviolation 1.0000000 1.0000000 1.0000000 0.5000000 1.0000000  
## violatedparole 0.0000000 0.0000000 0.0000000 0.5000000 0.0000000  
##   
## 38.9 39 39.1 39.2 39.4  
## noparoleviolation 1.0000000 1.0000000 1.0000000 0.7500000 1.0000000  
## violatedparole 0.0000000 0.0000000 0.0000000 0.2500000 0.0000000  
##   
## 39.5 39.6 39.7 39.8 39.9  
## noparoleviolation 1.0000000 1.0000000 0.5000000 0.6666667 1.0000000  
## violatedparole 0.0000000 0.0000000 0.5000000 0.3333333 0.0000000  
##   
## 40 40.1 40.3 40.4 40.6  
## noparoleviolation 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000  
## violatedparole 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  
##   
## 40.8 40.9 41 41.1 41.2  
## noparoleviolation 1.0000000 1.0000000 1.0000000 0.6666667 1.0000000  
## violatedparole 0.0000000 0.0000000 0.0000000 0.3333333 0.0000000  
##   
## 41.3 41.4 41.6 41.7 41.9  
## noparoleviolation 0.6000000 0.5000000 1.0000000 0.6666667 1.0000000  
## violatedparole 0.4000000 0.5000000 0.0000000 0.3333333 0.0000000  
##   
## 42 42.1 42.3 42.4 42.5  
## noparoleviolation 1.0000000 0.0000000 1.0000000 1.0000000 1.0000000  
## violatedparole 0.0000000 1.0000000 0.0000000 0.0000000 0.0000000  
##   
## 42.6 42.8 43 43.1 43.2  
## noparoleviolation 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000  
## violatedparole 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  
##   
## 43.3 43.4 43.5 43.6 43.7  
## noparoleviolation 1.0000000 1.0000000 1.0000000 0.5000000 1.0000000  
## violatedparole 0.0000000 0.0000000 0.0000000 0.5000000 0.0000000  
##   
## 43.8 44 44.1 44.2 44.3  
## noparoleviolation 1.0000000 1.0000000 0.7500000 1.0000000 1.0000000  
## violatedparole 0.0000000 0.0000000 0.2500000 0.0000000 0.0000000  
##   
## 44.4 44.5 44.6 44.7 44.8  
## noparoleviolation 0.0000000 1.0000000 1.0000000 0.5000000 1.0000000  
## violatedparole 1.0000000 0.0000000 0.0000000 0.5000000 0.0000000  
##   
## 44.9 45 45.1 45.4 45.5  
## noparoleviolation 0.6666667 0.6000000 1.0000000 1.0000000 1.0000000  
## violatedparole 0.3333333 0.4000000 0.0000000 0.0000000 0.0000000  
##   
## 45.6 45.8 45.9 46 46.1  
## noparoleviolation 1.0000000 0.0000000 1.0000000 1.0000000 1.0000000  
## violatedparole 0.0000000 1.0000000 0.0000000 0.0000000 0.0000000  
##   
## 46.2 46.3 46.4 46.5 46.6  
## noparoleviolation 1.0000000 1.0000000 1.0000000 0.0000000 1.0000000  
## violatedparole 0.0000000 0.0000000 0.0000000 1.0000000 0.0000000  
##   
## 46.7 46.8 46.9 47 47.1  
## noparoleviolation 1.0000000 1.0000000 1.0000000 0.5000000 1.0000000  
## violatedparole 0.0000000 0.0000000 0.0000000 0.5000000 0.0000000  
##   
## 47.2 47.3 47.5 47.7 47.8  
## noparoleviolation 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000  
## violatedparole 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  
##   
## 48 48.2 48.4 48.5 48.7  
## noparoleviolation 1.0000000 0.6666667 0.0000000 0.6666667 1.0000000  
## violatedparole 0.0000000 0.3333333 1.0000000 0.3333333 0.0000000  
##   
## 48.8 48.9 49 49.3 49.9  
## noparoleviolation 0.5000000 1.0000000 1.0000000 0.0000000 1.0000000  
## violatedparole 0.5000000 0.0000000 0.0000000 1.0000000 0.0000000  
##   
## 50.1 50.2 50.5 50.6 50.9  
## noparoleviolation 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000  
## violatedparole 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  
##   
## 51 51.1 51.2 51.3 51.4  
## noparoleviolation 1.0000000 1.0000000 1.0000000 1.0000000 0.0000000  
## violatedparole 0.0000000 0.0000000 0.0000000 0.0000000 1.0000000  
##   
## 51.7 51.8 52.1 52.5 52.6  
## noparoleviolation 1.0000000 1.0000000 1.0000000 0.0000000 1.0000000  
## violatedparole 0.0000000 0.0000000 0.0000000 1.0000000 0.0000000  
##   
## 53 53.5 53.8 53.9 54.1  
## noparoleviolation 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000  
## violatedparole 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  
##   
## 54.4 54.5 54.8 54.9 55  
## noparoleviolation 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000  
## violatedparole 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  
##   
## 55.7 56.4 56.5 56.8 57.5  
## noparoleviolation 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000  
## violatedparole 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  
##   
## 58.5 59.4 61.4 61.6 63.4  
## noparoleviolation 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000  
## violatedparole 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  
##   
## 65.1 67  
## noparoleviolation 1.0000000 1.0000000  
## violatedparole 0.0000000 0.0000000

State

ggplot(train,aes(x=state, fill = violator)) + geom\_bar(position = "fill")

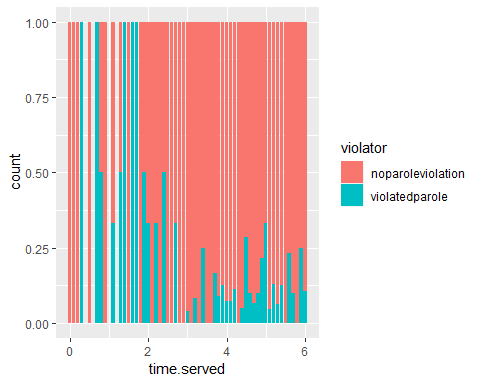


t4 = table(parole$violator, parole$state)  
prop.table(t4, margin = 2)

##   
## anyotherstate Kentucky Louisiana Virginia  
## noparoleviolation 0.86013986 0.88333333 0.54878049 0.97878788  
## violatedparole 0.13986014 0.11666667 0.45121951 0.02121212

Time Served

ggplot(train,aes(x=time.served , fill = violator)) + geom\_bar(position = "fill")

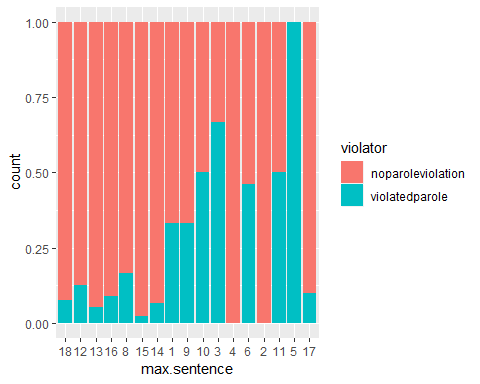


t5 = table(parole$violator, parole$time.served)  
prop.table(t5, margin = 2)

##   
## 0 0.1 0.2 0.3 0.5  
## noparoleviolation 1.00000000 0.50000000 1.00000000 0.00000000 1.00000000  
## violatedparole 0.00000000 0.50000000 0.00000000 1.00000000 0.00000000  
##   
## 0.7 0.8 0.9 1.1 1.2  
## noparoleviolation 0.50000000 0.33333333 1.00000000 0.75000000 1.00000000  
## violatedparole 0.50000000 0.66666667 0.00000000 0.25000000 0.00000000  
##   
## 1.3 1.4 1.5 1.6 1.7  
## noparoleviolation 0.50000000 0.00000000 1.00000000 0.00000000 0.33333333  
## violatedparole 0.50000000 1.00000000 0.00000000 1.00000000 0.66666667  
##   
## 1.8 1.9 2 2.1 2.2  
## noparoleviolation 1.00000000 0.50000000 0.66666667 1.00000000 0.66666667  
## violatedparole 0.00000000 0.50000000 0.33333333 0.00000000 0.33333333  
##   
## 2.3 2.4 2.5 2.6 2.7  
## noparoleviolation 1.00000000 0.66666667 1.00000000 1.00000000 0.66666667  
## violatedparole 0.00000000 0.33333333 0.00000000 0.00000000 0.33333333  
##   
## 2.8 2.9 3 3.1 3.2  
## noparoleviolation 1.00000000 1.00000000 0.96721311 1.00000000 0.89473684  
## violatedparole 0.00000000 0.00000000 0.03278689 0.00000000 0.10526316  
##   
## 3.3 3.4 3.5 3.6 3.7  
## noparoleviolation 1.00000000 0.75000000 1.00000000 1.00000000 0.82352941  
## violatedparole 0.00000000 0.25000000 0.00000000 0.00000000 0.17647059  
##   
## 3.8 3.9 4 4.1 4.2  
## noparoleviolation 0.93333333 0.83333333 0.84210526 0.94736842 0.85714286  
## violatedparole 0.06666667 0.16666667 0.15789474 0.05263158 0.14285714  
##   
## 4.3 4.4 4.5 4.6 4.7  
## noparoleviolation 1.00000000 0.89655172 0.85185185 0.93333333 0.95454545  
## violatedparole 0.00000000 0.10344828 0.14814815 0.06666667 0.04545455  
##   
## 4.8 4.9 5 5.1 5.2  
## noparoleviolation 0.93750000 0.68181818 0.71428571 0.96666667 0.87500000  
## violatedparole 0.06250000 0.31818182 0.28571429 0.03333333 0.12500000  
##   
## 5.3 5.4 5.5 5.6 5.7  
## noparoleviolation 0.92307692 0.92307692 1.00000000 0.81250000 0.91666667  
## violatedparole 0.07692308 0.07692308 0.00000000 0.18750000 0.08333333  
##   
## 5.8 5.9 6  
## noparoleviolation 0.92307692 0.88888889 0.84615385  
## violatedparole 0.07692308 0.11111111 0.15384615

Max Sentence

ggplot(train,aes(x=max.sentence, fill = violator)) + geom\_bar(position = "fill")



t6 = table(parole$violator, parole$max.sentence)  
prop.table(t6, margin = 2)

##   
## 18 12 13 16 8  
## noparoleviolation 0.89743590 0.86496350 0.96103896 0.88888889 0.85000000  
## violatedparole 0.10256410 0.13503650 0.03896104 0.11111111 0.15000000  
##   
## 15 14 1 9 10  
## noparoleviolation 0.98333333 0.95454545 0.75000000 0.66666667 0.66666667  
## violatedparole 0.01666667 0.04545455 0.25000000 0.33333333 0.33333333  
##   
## 3 4 6 2 11  
## noparoleviolation 0.33333333 1.00000000 0.56250000 1.00000000 0.60000000  
## violatedparole 0.66666667 0.00000000 0.43750000 0.00000000 0.40000000  
##   
## 5 17  
## noparoleviolation 0.00000000 0.93333333  
## violatedparole 1.00000000 0.06666667

Multiple Offenses

ggplot(train,aes(x=multiple.offenses, fill = violator)) + geom\_bar(position = "fill")

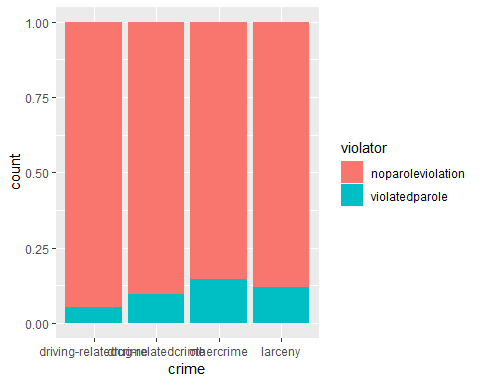


t7 = table(parole$violator, parole$multiple.offenses)  
prop.table(t7, margin = 2)

##   
## otherwise multipleoffenses  
## noparoleviolation 0.9201278 0.8535912  
## violatedparole 0.0798722 0.1464088

Crime

ggplot(train,aes(x=crime, fill = violator)) + geom\_bar(position = "fill")



t8 = table(parole$violator, parole$crime)  
prop.table(t8, margin = 2)

##   
## driving-relatedcrime drug-relatedcrime othercrime  
## noparoleviolation 0.93069307 0.87581699 0.87619048  
## violatedparole 0.06930693 0.12418301 0.12380952  
##   
## larceny  
## noparoleviolation 0.87735849  
## violatedparole 0.12264151

model1 = glm(violator ~ state, train, family = "binomial")  
summary(model1)

##   
## Call:  
## glm(formula = violator ~ state, family = "binomial", data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.1106 -0.4084 -0.2255 -0.2255 2.7147   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.7677 0.2892 -6.113 9.79e-10 \*\*\*  
## stateKentucky -0.6747 0.5146 -1.311 0.189803   
## stateLouisiana 1.6086 0.3841 4.188 2.81e-05 \*\*\*  
## stateVirginia -1.8916 0.5046 -3.749 0.000178 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 340.04 on 472 degrees of freedom  
## Residual deviance: 264.58 on 469 degrees of freedom  
## AIC: 272.58  
##   
## Number of Fisher Scoring iterations: 6

We see the same information that we did above in the graph; parolees are less likely to break parole in every other state than Louisiana. We see that the negative coefficients in the “Estimate Std” column point to the higher liklihood being in Louisiana.

Since this is our first logistic regression model we will compare our AIC of 272.58 to the others.

allmod = glm( violator ~ state + male + race + age + time.served + max.sentence + multiple.offenses + crime, train, family = "binomial")  
summary(allmod)

##   
## Call:  
## glm(formula = violator ~ state + male + race + age + time.served +   
## max.sentence + multiple.offenses + crime, family = "binomial",   
## data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.0554 -0.3590 -0.2200 -0.1101 3.0389   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -4.267e+00 1.185e+00 -3.599 0.000319  
## stateKentucky -3.758e-01 5.954e-01 -0.631 0.527990  
## stateLouisiana 9.567e-01 5.846e-01 1.637 0.101731  
## stateVirginia -4.088e+00 8.448e-01 -4.839 1.3e-06  
## malefemale -4.436e-01 5.078e-01 -0.874 0.382326  
## raceotherwise 1.130e+00 4.389e-01 2.574 0.010051  
## age 3.701e-02 1.791e-02 2.066 0.038815  
## time.served -3.217e-02 1.318e-01 -0.244 0.807120  
## max.sentence12 4.847e-03 6.799e-01 0.007 0.994312  
## max.sentence13 7.518e-01 1.041e+00 0.722 0.470168  
## max.sentence16 4.965e-01 1.157e+00 0.429 0.667940  
## max.sentence8 -6.542e-01 1.137e+00 -0.575 0.565043  
## max.sentence15 -4.543e-02 1.263e+00 -0.036 0.971313  
## max.sentence14 1.219e+00 1.072e+00 1.137 0.255545  
## max.sentence1 -9.938e-01 1.739e+00 -0.572 0.567624  
## max.sentence9 -7.802e-01 1.256e+00 -0.621 0.534537  
## max.sentence10 -1.141e+00 1.711e+00 -0.667 0.505009  
## max.sentence3 -2.062e-01 1.507e+00 -0.137 0.891158  
## max.sentence4 -1.602e+01 1.133e+03 -0.014 0.988718  
## max.sentence6 -8.758e-01 9.766e-01 -0.897 0.369835  
## max.sentence2 -1.634e+01 1.693e+03 -0.010 0.992298  
## max.sentence11 3.415e-01 1.307e+00 0.261 0.793876  
## max.sentence5 1.653e+01 2.400e+03 0.007 0.994505  
## max.sentence17 2.243e+00 1.431e+00 1.568 0.116993  
## multiple.offensesmultipleoffenses 1.708e+00 4.514e-01 3.784 0.000155  
## crimedrug-relatedcrime 8.259e-02 7.526e-01 0.110 0.912613  
## crimeothercrime 7.582e-01 6.622e-01 1.145 0.252229  
## crimelarceny 8.066e-01 7.822e-01 1.031 0.302461  
##   
## (Intercept) \*\*\*  
## stateKentucky   
## stateLouisiana   
## stateVirginia \*\*\*  
## malefemale   
## raceotherwise \*   
## age \*   
## time.served   
## max.sentence12   
## max.sentence13   
## max.sentence16   
## max.sentence8   
## max.sentence15   
## max.sentence14   
## max.sentence1   
## max.sentence9   
## max.sentence10   
## max.sentence3   
## max.sentence4   
## max.sentence6   
## max.sentence2   
## max.sentence11   
## max.sentence5   
## max.sentence17   
## multiple.offensesmultipleoffenses \*\*\*  
## crimedrug-relatedcrime   
## crimeothercrime   
## crimelarceny   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 340.04 on 472 degrees of freedom  
## Residual deviance: 220.41 on 445 degrees of freedom  
## AIC: 276.41  
##   
## Number of Fisher Scoring iterations: 15

emptymod = glm(violator ~1, train, family = "binomial")  
summary(emptymod)

##   
## Call:  
## glm(formula = violator ~ 1, family = "binomial", data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.4972 -0.4972 -0.4972 -0.4972 2.0745   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.0281 0.1434 -14.14 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 340.04 on 472 degrees of freedom  
## Residual deviance: 340.04 on 472 degrees of freedom  
## AIC: 342.04  
##   
## Number of Fisher Scoring iterations: 4

backmod = stepAIC(allmod, direction = "backward", trace = TRUE)

## Start: AIC=276.41  
## violator ~ state + male + race + age + time.served + max.sentence +   
## multiple.offenses + crime  
##   
## Df Deviance AIC  
## - max.sentence 16 232.46 256.46  
## - crime 3 223.61 273.61  
## - time.served 1 220.47 274.47  
## - male 1 221.21 275.21  
## <none> 220.41 276.41  
## - age 1 224.70 278.70  
## - race 1 227.18 281.18  
## - multiple.offenses 1 235.72 289.72  
## - state 3 281.50 331.50  
##   
## Step: AIC=256.46  
## violator ~ state + male + race + age + time.served + multiple.offenses +   
## crime  
##   
## Df Deviance AIC  
## - time.served 1 232.50 254.50  
## - crime 3 237.35 255.35  
## - male 1 233.75 255.75  
## <none> 232.46 256.46  
## - age 1 235.78 257.78  
## - race 1 238.59 260.59  
## - multiple.offenses 1 249.02 271.02  
## - state 3 317.35 335.35  
##   
## Step: AIC=254.5  
## violator ~ state + male + race + age + multiple.offenses + crime  
##   
## Df Deviance AIC  
## - crime 3 237.41 253.41  
## - male 1 233.88 253.88  
## <none> 232.50 254.50  
## - age 1 235.78 255.78  
## - race 1 238.60 258.60  
## - multiple.offenses 1 249.70 269.70  
## - state 3 319.80 335.80  
##   
## Step: AIC=253.41  
## violator ~ state + male + race + age + multiple.offenses  
##   
## Df Deviance AIC  
## - male 1 238.31 252.31  
## <none> 237.41 253.41  
## - age 1 239.86 253.86  
## - race 1 243.87 257.87  
## - multiple.offenses 1 255.53 269.53  
## - state 3 325.68 335.68  
##   
## Step: AIC=252.31  
## violator ~ state + race + age + multiple.offenses  
##   
## Df Deviance AIC  
## <none> 238.31 252.31  
## - age 1 240.42 252.42  
## - race 1 245.01 257.01  
## - multiple.offenses 1 256.59 268.59  
## - state 3 326.11 334.11

summary(backmod)

##   
## Call:  
## glm(formula = violator ~ state + race + age + multiple.offenses,   
## family = "binomial", data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6006 -0.3634 -0.2710 -0.1533 2.8440   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.41616 0.71485 -4.779 1.76e-06 \*\*\*  
## stateKentucky -0.47239 0.54431 -0.868 0.3855   
## stateLouisiana 0.59117 0.48037 1.231 0.2184   
## stateVirginia -3.18805 0.60453 -5.274 1.34e-07 \*\*\*  
## raceotherwise 1.02399 0.39793 2.573 0.0101 \*   
## age 0.02383 0.01640 1.453 0.1462   
## multiple.offensesmultipleoffenses 1.64389 0.39957 4.114 3.89e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 340.04 on 472 degrees of freedom  
## Residual deviance: 238.31 on 466 degrees of freedom  
## AIC: 252.31  
##   
## Number of Fisher Scoring iterations: 6

forwardmod = stepAIC(emptymod, direction = "forward", scope=list(uppper=allmod, lower=emptymod), trace = TRUE)

## Start: AIC=342.04  
## violator ~ 1

summary(forwardmod)

##   
## Call:  
## glm(formula = violator ~ 1, family = "binomial", data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.4972 -0.4972 -0.4972 -0.4972 2.0745   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.0281 0.1434 -14.14 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 340.04 on 472 degrees of freedom  
## Residual deviance: 340.04 on 472 degrees of freedom  
## AIC: 342.04  
##   
## Number of Fisher Scoring iterations: 4

Based on all of the models created, the best model as far as fit goes when analyzing the AIC is the backward stepwise approach. Using the backward stepwise model we get an AIC of 252. This is better than the original model just based on “state” (AIC 272) and the forward stepwise model (AIC 342).

It seems as if the most important variables where whether or not you have multiple offenses (in this case if you do), your race (in this case is you are a race other than white), and your state (in this case if you are in Virginia). This model seems relatively intuitive but there are limited variables that are significant.

model2 = glm(violator ~ state + multiple.offenses + race, train, family = "binomial")  
summary(model2)

##   
## Call:  
## glm(formula = violator ~ state + multiple.offenses + race, family = "binomial",   
## data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.4553 -0.3862 -0.2931 -0.1787 2.8791   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.5582 0.3709 -6.898 5.28e-12 \*\*\*  
## stateKentucky -0.4816 0.5417 -0.889 0.3740   
## stateLouisiana 0.5292 0.4769 1.110 0.2672   
## stateVirginia -3.2301 0.6028 -5.358 8.39e-08 \*\*\*  
## multiple.offensesmultipleoffenses 1.6596 0.3985 4.165 3.12e-05 \*\*\*  
## raceotherwise 1.0024 0.3966 2.528 0.0115 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 340.04 on 472 degrees of freedom  
## Residual deviance: 240.42 on 467 degrees of freedom  
## AIC: 252.42  
##   
## Number of Fisher Scoring iterations: 6

The quality of this model is good though slightly worse than the model created with the backward stepwise process.

It seems as if the most important variables where whether or not you have multiple offenses (in this case if you do), your race (in this case is you are a race other than white), and your state (in this case if you are in Virginia). This model seems relatively intuitive but there are limited variables that are significant. Looking at just these variables though the significance of the race went down.

Parolee1

newdata = data.frame (state = "Louisiana", race = "white", multiple.offenses = "multipleoffenses")  
predict(forwardmod, newdata, type = "response")

## 1   
## 0.1162791

Parolee2

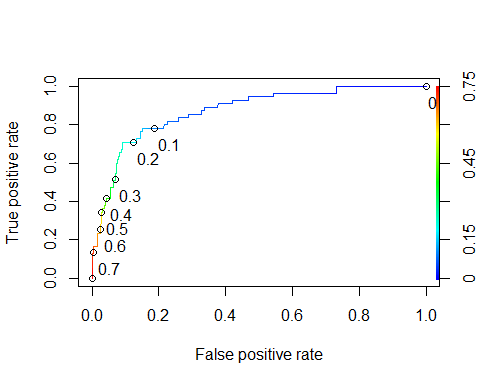
newdata = data.frame (state = "Kentucky", race = "otherwise", multiple.offenses = "otherwise")  
predict(forwardmod, newdata, type = "response")

## 1   
## 0.1162791

predictions = predict(backmod, type = "response")  
head(predictions)

## 1 2 3 4 5 6   
## 0.06753690 0.15586516 0.05302695 0.13267145 0.21763144 0.06430986

ROCRpred = prediction(predictions, train$violator)  
  
ROCRperf = performance(ROCRpred, "tpr", "fpr")  
plot(ROCRperf, colorize=TRUE, print.cutoffs.at=seq(0,1,by=0.1), text.adj=c(-0.2, 1.7))



opt.cut = function(perf, pred){  
 cut.ind = mapply(FUN=function(x, y, p){  
 d = (x - 0)^2 + (y-1)^2  
 ind = which(d == min(d))  
 c(sensitivity = y[[ind]], specificity = 1-x[[ind]],   
 cutoff = p[[ind]])  
 }, perf@x.values, perf@y.values, pred@cutoffs)  
}  
print(opt.cut(ROCRperf, ROCRpred))

## [,1]  
## sensitivity 0.7818182  
## specificity 0.8492823  
## cutoff 0.1452056

acct1 = table(train$violator,predictions > 0.1452056)  
acct1

##   
## FALSE TRUE  
## noparoleviolation 355 63  
## violatedparole 12 43

(acct1[1,1]+acct1[2,2])/nrow(train)

## [1] 0.8414376

Based on the cutoff of .1452 the sensitivity is .7818, the specificity is .8493, and the accuracy is .8414. The implications of incorrectly classifying a parolee could be mean that people could be arrested or subject to criminal complications.

acct1 = table(train$violator,predictions > 0.5)  
acct1

##   
## FALSE TRUE  
## noparoleviolation 406 12  
## violatedparole 37 18

(acct1[1,1]+acct1[2,2])/nrow(train)

## [1] 0.8964059

acct1 = table(train$violator,predictions > 0.6)  
acct1

##   
## FALSE TRUE  
## noparoleviolation 408 10  
## violatedparole 42 13

(acct1[1,1]+acct1[2,2])/nrow(train)

## [1] 0.8900634

acct1 = table(train$violator,predictions > 0.4)  
acct1

##   
## FALSE TRUE  
## noparoleviolation 400 18  
## violatedparole 33 22

(acct1[1,1]+acct1[2,2])/nrow(train)

## [1] 0.8921776

acct1 = table(train$violator,predictions > 0.5)  
acct1

##   
## FALSE TRUE  
## noparoleviolation 406 12  
## violatedparole 37 18

(acct1[1,1]+acct1[2,2])/nrow(train)

## [1] 0.8964059

#acct1 = table(test$violator,predictions > 0.5)  
#acct1  
#(acct1[1,1]+acct1[2,2])/nrow(test)

acct1 = table(train$violator,predictions > 0.5)  
acct1

##   
## FALSE TRUE  
## noparoleviolation 406 12  
## violatedparole 37 18

(acct1[1,1]+acct1[2,2])/nrow(test)

## [1] 2.09901

\*\*Hey Dr. Hill,

Could you walk me through how I am supposed to get number 10? I had quite a bit of trouble with this assignment (hence the late submission) but I have managed to get through all of them except this one.

I have tried to introduce the test data set into the table with the predictions but it tells me ‘all arguments must have the same length’. Not entirely what is causing this unless there is an N/A observation?

Then the second version gives me 2.099 which I am assuming is way wrong because it is saying this model would work 200% of the time?

Thank you in advance for any clarity.