# Natalie Chandler

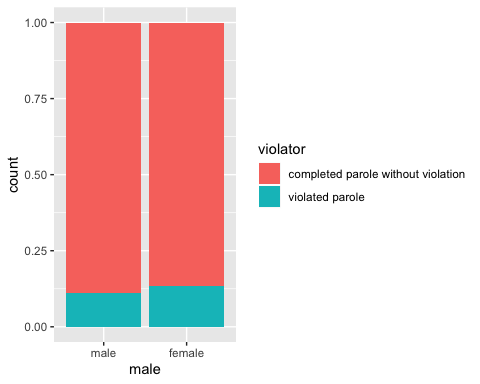
## Module 3 - Classification and Logistic Regression

library(tidyverse)  
library(MASS)  
library(caret)  
library(ROCR)  
  
parole = read\_csv("parole.csv")

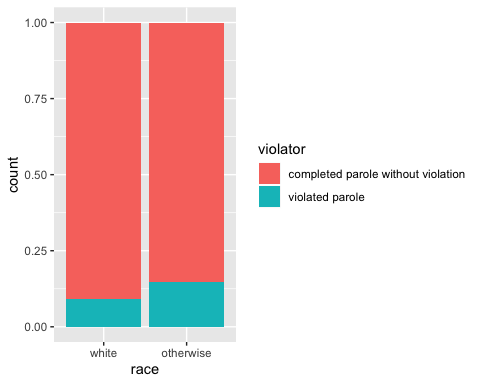
parole = parole %>% mutate(male = as\_factor(as.character(male))) %>%  
mutate(male = fct\_recode(male,  
"male" = "1",  
"female" = "0")) %>%  
  
mutate(race = as\_factor(as.character(race))) %>%  
mutate(race = fct\_recode(race,  
"white" = "1",  
"otherwise" = "2")) %>%  
   
mutate(state = as\_factor(as.character(state))) %>%  
mutate(state = fct\_recode(state,  
"any other state" = "1",  
"Kentucky" = "2",  
"Louisiana"="3",  
"Virginia"="4")) %>%  
   
mutate(crime = as\_factor(as.character(crime))) %>%  
mutate(crime = fct\_recode(crime,  
"any other crime" = "1",  
"larceny" = "2",  
"drug-related crime" = "3",  
"driving-related crime" = "4")) %>%  
  
mutate(multiple.offenses = as\_factor(as.character(multiple.offenses))) %>%  
mutate(multiple.offenses = fct\_recode(multiple.offenses,  
"multiple offenses" = "1",  
"otherwise" = "0")) %>%  
  
mutate(violator = as\_factor(as.character(violator))) %>%  
mutate(violator = fct\_recode(violator,  
"violated parole" = "1",  
"completed parole without violation" = "0"))

# Task 1  
set.seed(12345)  
train.rows = createDataPartition(y = parole$violator, p=0.7, list = FALSE)   
train = parole[train.rows,]   
test = parole[-train.rows,]

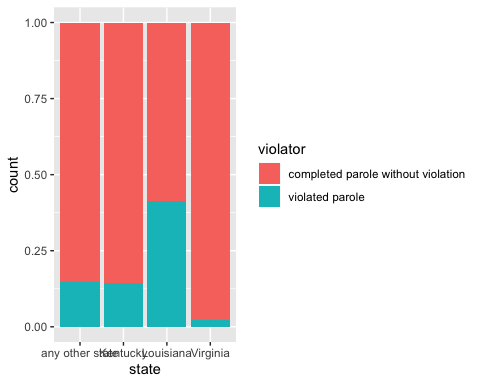
#Task 2  
ggplot(train, aes(x=male, fill=violator))+ geom\_bar(position = "fill")



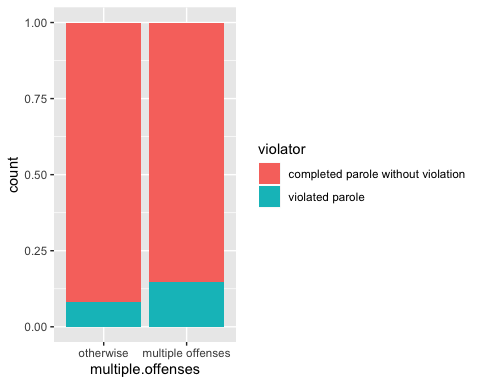
ggplot(train, aes(x=race, fill=violator))+ geom\_bar(position = "fill")



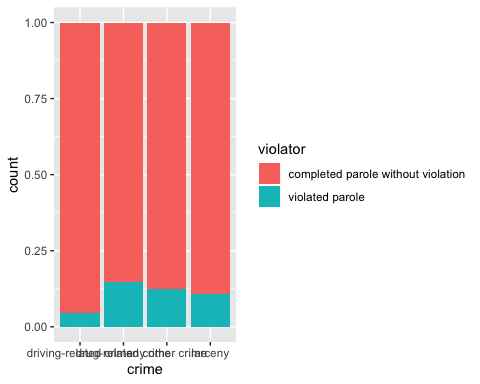
ggplot(train, aes(x=state, fill=violator))+ geom\_bar(position = "fill")



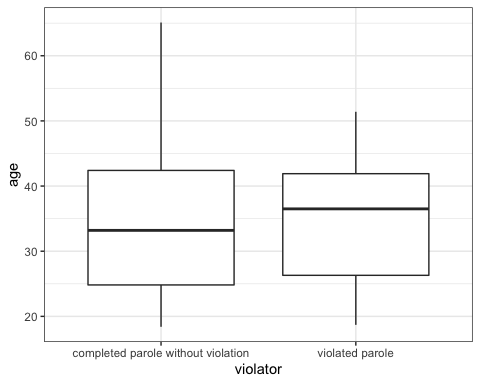
ggplot(train, aes(x=multiple.offenses, fill=violator))+ geom\_bar(position = "fill")



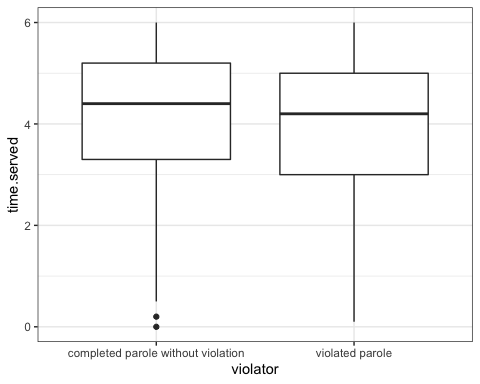
ggplot(train, aes(x=crime, fill=violator))+ geom\_bar(position = "fill")



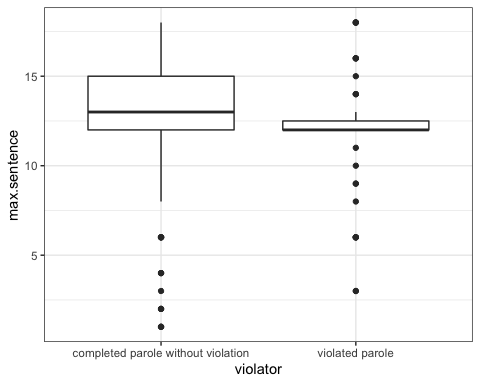
ggplot(train, aes(x=violator, y=age))+ geom\_boxplot()+theme\_bw()



ggplot(train, aes(x=violator, y=time.served))+ geom\_boxplot()+theme\_bw()



ggplot(train, aes(x=violator, y=max.sentence))+ geom\_boxplot()+theme\_bw()



**The bar plots above indicate that race, state, number of offenses, and type of crime may contribute to violation of parole. State showed greatest variation and perhaps is a signficant link to parole violations. Interestingly, there did not seem to be a significant difference in parole violation in men to women. Bar plots were used as a means of visually comparing these groups, as they are categorical. The boxplots consider age, time served, and maximum sentence. Age does not seem to be a predictor of parole violation, as there is little difference in distribution of parole violator and non-violator age. The results are similar for time served.**

#Task 3  
LogReg1= glm(violator~state, train, family = "binomial")  
summary(LogReg1)

##   
## Call:  
## glm(formula = violator ~ state, family = "binomial", data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.0335 -0.5589 -0.2065 -0.2065 2.7780   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.75539 0.28944 -6.065 1.32e-09 \*\*\*  
## stateKentucky -0.02238 0.42567 -0.053 0.958067   
## stateLouisiana 1.40709 0.39351 3.576 0.000349 \*\*\*  
## stateVirginia -2.08191 0.53672 -3.879 0.000105 \*\*\*  
## ---  
## 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: 275.18 on 469 degrees of freedom  
## AIC: 283.18  
##   
## Number of Fisher Scoring iterations: 6

**The logistic regression does not compare probability, but log odds. The coefficient signs conclude that residing in the state of Louisiana is a significant contributor to parole violation, while Kentucky, Virginia, and other states were less likely contributors. While computing AIC values of other variables using the logistic model, the state variable seemed have the lowest AIC. This perhaps indicates state as one of the most significant contributors to the model. Moreover, Kentucky was not statistically relevant according to p-values. In contrast, Virginia showed p-value significance, yet a more negative link to parole violation.**

# Task 4  
allmod = glm(violator ~.,train, family = "binomial")   
summary(allmod)

##   
## Call:  
## glm(formula = violator ~ ., family = "binomial", data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6055 -0.3932 -0.2643 -0.1384 2.9470   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -3.750397 1.318165 -2.845 0.00444  
## malefemale 0.137577 0.411340 0.334 0.73803  
## raceotherwise 1.143719 0.403890 2.832 0.00463  
## age 0.005279 0.016910 0.312 0.75490  
## stateKentucky 0.124282 0.492370 0.252 0.80072  
## stateLouisiana 0.217202 0.556154 0.391 0.69614  
## stateVirginia -3.801561 0.666733 -5.702 1.19e-08  
## time.served -0.109344 0.118901 -0.920 0.35777  
## max.sentence 0.065956 0.054593 1.208 0.22700  
## multiple.offensesmultiple offenses 1.711032 0.396463 4.316 1.59e-05  
## crimedrug-related crime 0.516479 0.739095 0.699 0.48468  
## crimeany other crime 0.727043 0.690775 1.053 0.29257  
## crimelarceny 1.119953 0.797552 1.404 0.16025  
##   
## (Intercept) \*\*   
## malefemale   
## raceotherwise \*\*   
## age   
## stateKentucky   
## stateLouisiana   
## stateVirginia \*\*\*  
## time.served   
## max.sentence   
## multiple.offensesmultiple offenses \*\*\*  
## crimedrug-related crime   
## crimeany other crime   
## 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: 242.09 on 460 degrees of freedom  
## AIC: 268.09  
##   
## Number of Fisher Scoring iterations: 6

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=268.09  
## violator ~ male + race + age + state + time.served + max.sentence +   
## multiple.offenses + crime  
##   
## Df Deviance AIC  
## - crime 3 244.47 264.47  
## - age 1 242.18 266.18  
## - male 1 242.20 266.20  
## - time.served 1 242.93 266.93  
## - max.sentence 1 243.57 267.57  
## <none> 242.09 268.09  
## - race 1 250.24 274.24  
## - multiple.offenses 1 261.96 285.96  
## - state 3 316.24 336.24  
##   
## Step: AIC=264.47  
## violator ~ male + race + age + state + time.served + max.sentence +   
## multiple.offenses  
##   
## Df Deviance AIC  
## - age 1 244.48 262.48  
## - male 1 244.85 262.85  
## - time.served 1 245.04 263.04  
## - max.sentence 1 246.00 264.00  
## <none> 244.47 264.47  
## - race 1 252.62 270.62  
## - multiple.offenses 1 265.46 283.46  
## - state 3 321.69 335.69  
##   
## Step: AIC=262.48  
## violator ~ male + race + state + time.served + max.sentence +   
## multiple.offenses  
##   
## Df Deviance AIC  
## - male 1 244.86 260.86  
## - time.served 1 245.04 261.04  
## - max.sentence 1 246.01 262.01  
## <none> 244.48 262.48  
## - race 1 252.65 268.65  
## - multiple.offenses 1 265.52 281.52  
## - state 3 322.14 334.14  
##   
## Step: AIC=260.86  
## violator ~ race + state + time.served + max.sentence + multiple.offenses  
##   
## Df Deviance AIC  
## - time.served 1 245.31 259.31  
## - max.sentence 1 246.33 260.33  
## <none> 244.86 260.86  
## - race 1 252.80 266.80  
## - multiple.offenses 1 265.93 279.93  
## - state 3 322.54 332.54  
##   
## Step: AIC=259.31  
## violator ~ race + state + max.sentence + multiple.offenses  
##   
## Df Deviance AIC  
## - max.sentence 1 246.98 258.98  
## <none> 245.31 259.31  
## - race 1 253.11 265.11  
## - multiple.offenses 1 266.89 278.89  
## - state 3 323.88 331.88  
##   
## Step: AIC=258.98  
## violator ~ race + state + multiple.offenses  
##   
## Df Deviance AIC  
## <none> 246.98 258.98  
## - race 1 254.96 264.96  
## - multiple.offenses 1 267.66 277.66  
## - state 3 332.93 338.93

summary(backmod)

##   
## Call:  
## glm(formula = violator ~ race + state + multiple.offenses, family = "binomial",   
## data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.3609 -0.4094 -0.2705 -0.1575 2.9653   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -2.51087 0.36354 -6.907 4.96e-12  
## raceotherwise 1.09382 0.38974 2.807 0.00501  
## stateKentucky 0.07372 0.46051 0.160 0.87282  
## stateLouisiana 0.10381 0.50018 0.208 0.83559  
## stateVirginia -3.60795 0.63788 -5.656 1.55e-08  
## multiple.offensesmultiple offenses 1.73482 0.39421 4.401 1.08e-05  
##   
## (Intercept) \*\*\*  
## raceotherwise \*\*   
## stateKentucky   
## stateLouisiana   
## stateVirginia \*\*\*  
## multiple.offensesmultiple offenses \*\*\*  
## ---  
## 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: 246.98 on 467 degrees of freedom  
## AIC: 258.98  
##   
## Number of Fisher Scoring iterations: 6

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

## Start: AIC=342.04  
## violator ~ 1  
##   
## Df Deviance AIC  
## + state 3 275.18 283.18  
## + max.sentence 1 331.01 335.01  
## + multiple.offenses 1 335.02 339.02  
## + race 1 336.51 340.51  
## + time.served 1 336.61 340.61  
## <none> 340.04 342.04  
## + crime 3 335.07 343.07  
## + male 1 339.72 343.72  
## + age 1 339.95 343.95  
##   
## Step: AIC=283.18  
## violator ~ state  
##   
## Df Deviance AIC  
## + multiple.offenses 1 254.96 264.96  
## + race 1 267.66 277.66  
## <none> 275.18 283.18  
## + max.sentence 1 274.27 284.27  
## + time.served 1 274.44 284.44  
## + age 1 275.11 285.11  
## + male 1 275.13 285.13  
## + crime 3 271.72 285.72  
##   
## Step: AIC=264.96  
## violator ~ state + multiple.offenses  
##   
## Df Deviance AIC  
## + race 1 246.98 258.98  
## <none> 254.96 264.96  
## + max.sentence 1 253.11 265.11  
## + time.served 1 254.47 266.47  
## + male 1 254.91 266.91  
## + age 1 254.94 266.94  
## + crime 3 252.75 268.75  
##   
## Step: AIC=258.98  
## violator ~ state + multiple.offenses + race  
##   
## Df Deviance AIC  
## <none> 246.98 258.98  
## + max.sentence 1 245.31 259.31  
## + time.served 1 246.33 260.33  
## + male 1 246.78 260.78  
## + age 1 246.98 260.98  
## + crime 3 244.78 262.79

summary(forwardmod)

##   
## Call:  
## glm(formula = violator ~ state + multiple.offenses + race, family = "binomial",   
## data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.3609 -0.4094 -0.2705 -0.1575 2.9653   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -2.51087 0.36354 -6.907 4.96e-12  
## stateKentucky 0.07372 0.46051 0.160 0.87282  
## stateLouisiana 0.10381 0.50018 0.208 0.83559  
## stateVirginia -3.60795 0.63788 -5.656 1.55e-08  
## multiple.offensesmultiple offenses 1.73482 0.39421 4.401 1.08e-05  
## raceotherwise 1.09382 0.38974 2.807 0.00501  
##   
## (Intercept) \*\*\*  
## stateKentucky   
## stateLouisiana   
## stateVirginia \*\*\*  
## multiple.offensesmultiple offenses \*\*\*  
## raceotherwise \*\*   
## ---  
## 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: 246.98 on 467 degrees of freedom  
## AIC: 258.98  
##   
## Number of Fisher Scoring iterations: 6

**The model showed an improvement with the addition of multiple offenses and race variables to the AIC value. Interestingly, the p-value indicates less statistical significance for state (in particular, Louisiana). In the first model, Louisiana showed significant statistical relevance. Perhaps this is linked to correlation of the variables (multicollinearity), as discussed in linear regression. Multiple offenses and race variables showed statistical significance according to p-values. Increases in number of offenses, as expected, is indicative of increases in parole violation. A less obvious link to parole violations was race. However, this was less significant. Overall, this model was intutive based on the data visualized in early questions. However, changes in significance in certain variables was not expected.**

# Task 5  
LogReg2= glm(violator~state+multiple.offenses+race, train, family = "binomial")  
summary(LogReg2)

##   
## Call:  
## glm(formula = violator ~ state + multiple.offenses + race, family = "binomial",   
## data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.3609 -0.4094 -0.2705 -0.1575 2.9653   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -2.51087 0.36354 -6.907 4.96e-12  
## stateKentucky 0.07372 0.46051 0.160 0.87282  
## stateLouisiana 0.10381 0.50018 0.208 0.83559  
## stateVirginia -3.60795 0.63788 -5.656 1.55e-08  
## multiple.offensesmultiple offenses 1.73482 0.39421 4.401 1.08e-05  
## raceotherwise 1.09382 0.38974 2.807 0.00501  
##   
## (Intercept) \*\*\*  
## stateKentucky   
## stateLouisiana   
## stateVirginia \*\*\*  
## multiple.offensesmultiple offenses \*\*\*  
## raceotherwise \*\*   
## ---  
## 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: 246.98 on 467 degrees of freedom  
## AIC: 258.98  
##   
## Number of Fisher Scoring iterations: 6

**The model above indicates that race and multiple offenses are significant indicators of expected parole violations. While state is included in this model, it shows that many variables are not statistically relevant. As stated above, the change in statistical significance in state variables may indicate multiculinearity.**

#Task 6  
  
Parolee1 = data.frame(state = "Louisiana", multiple.offenses = "multiple offenses", race = "white")  
predict(forwardmod, Parolee1, type="response")

## 1   
## 0.3379961

Parolee2 = data.frame(state = "Kentucky", multiple.offenses = "otherwise", race = "otherwise")  
predict(forwardmod, Parolee2, type="response")

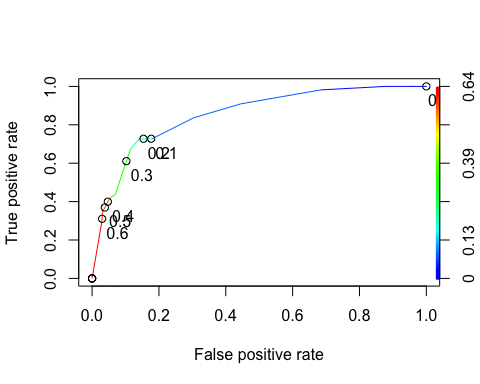
## 1   
## 0.2069629

**The probability that parolee 1 will violate parole is .34 and parolee 2 is .21.**

# Task 7  
predictions = predict(LogReg2, type="response")   
head(predictions)

## 1 2 3 4 5 6   
## 0.07509978 0.19512504 0.19512504 0.07509978 0.07509978 0.19512504

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))



as.numeric(performance(ROCRpred, "auc")@y.values)

## [1] 0.8524576

#Task 8  
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 0.7272727  
## specificity 0.8588517  
## cutoff 0.2069629

t1 = table(train$violator,predictions > 0.2069629)  
t1

##   
## FALSE TRUE  
## completed parole without violation 359 59  
## violated parole 15 40

(t1[1,1]+t1[2,2])/nrow(train)

## [1] 0.8435518

**The accuracy of the model on the training set is 0.8435518, with sensitivity of 0.7272727 and specificity of 0.8588517. In this case, we are sacrificing some accuracy in order to balance sensitivity and specificity. Therefore, parolees may be incorrectly classfied as parole violators or non-violators. Implications for this could be costly to the government and society. This includes both safety and negative monetary implications for government*.***

#Task 9  
t1 = table(train$violator,predictions > 0.3)  
t1

##   
## FALSE TRUE  
## completed parole without violation 376 42  
## violated parole 22 33

(t1[1,1]+t1[2,2])/nrow(train)

## [1] 0.8646934

t1 = table(train$violator,predictions > 0.4)  
t1

##   
## FALSE TRUE  
## completed parole without violation 405 13  
## violated parole 36 19

(t1[1,1]+t1[2,2])/nrow(train)

## [1] 0.8964059

**The probability threshold that best maximizes accuracy on the training set is 0.4.**

#Task 10  
LogReg3= glm(violator~state+multiple.offenses+race, test, family = "binomial")  
  
predictions = predict(LogReg3, type="response")   
head(predictions)

## 1 2 3 4 5 6   
## 0.07636941 0.07636941 0.08005369 0.07636941 0.23794683 0.08005369

t1 = table(test$violator,predictions > 0.4)  
t1

##   
## FALSE TRUE  
## completed parole without violation 175 4  
## violated parole 13 10

(t1[1,1]+t1[2,2])/nrow(test)

## [1] 0.9158416

**The accuracy of the of the test set with a threshold of 0.4 is 0.9158416.**