library(tidyverse)  
library(MASS)  
library(caret)  
library(ROCR)

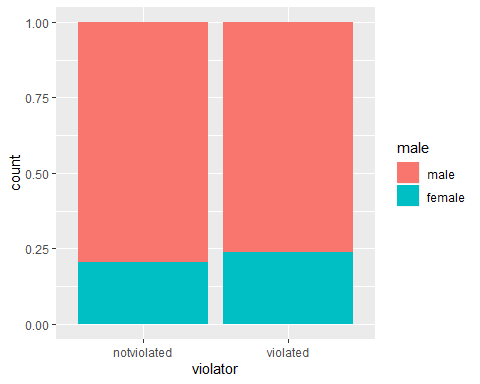
parole = read\_csv("parole.csv")

## 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()  
## )

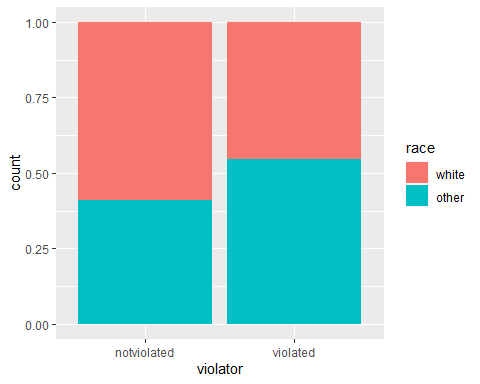
parole = parole %>%   
 mutate(male = as\_factor(as.character(male))) %>%  
 mutate(male = fct\_recode(male,  
 "male" = "1",  
 "female" = "0")) %>%  
 mutate(race = as\_factor(as.character(race))) %>%  
 mutate(race = fct\_recode(race,  
 "white" = "1",  
 "other" = "2")) %>%  
 mutate(state = as\_factor(as.character(state))) %>%  
 mutate(state = fct\_recode(state,  
 "Kentucky" = "2",  
 "Louisiana" = "3",  
 "Virginia" = "4",  
 "other" = "1")) %>%  
 mutate(crime = as\_factor(as.character(crime))) %>%  
 mutate(crime = fct\_recode(crime,  
 "larceny" = "2",  
 "drug" = "3",  
 "driving" = "4",  
 "other" = "1")) %>%  
 mutate(multiple.offenses = as\_factor(as.character(multiple.offenses))) %>%  
 mutate(multiple.offenses = fct\_recode(multiple.offenses,  
 "multiple" = "1",  
 "notmultiple" = "0")) %>%  
 mutate(violator = as\_factor(as.character(violator))) %>%  
 mutate(violator = fct\_recode(violator,  
 "violated" = "1",  
 "notviolated" = "0"))

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

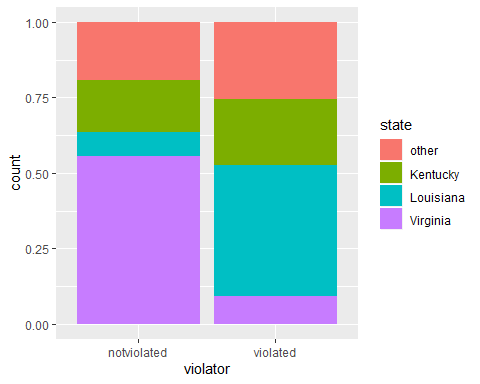
ggplot(train, aes(violator, fill = male)) + geom\_bar(position="fill")



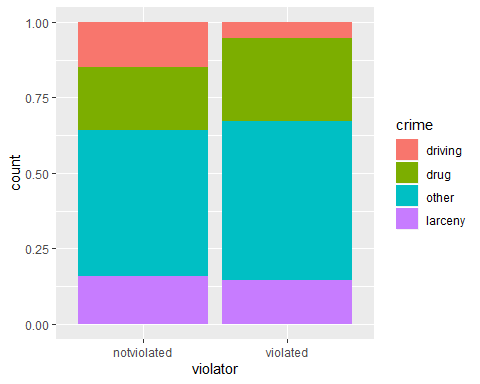
ggplot(train, aes(violator, fill = race)) + geom\_bar(position="fill")



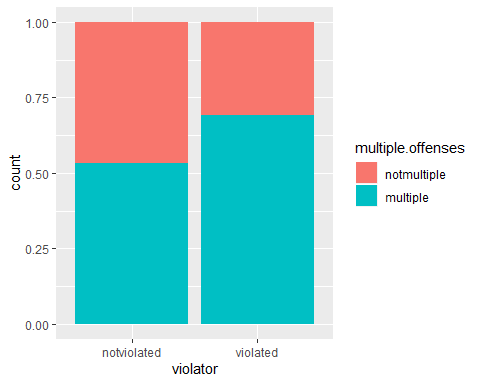
ggplot(train, aes(violator, fill = state)) + geom\_bar(position="fill")



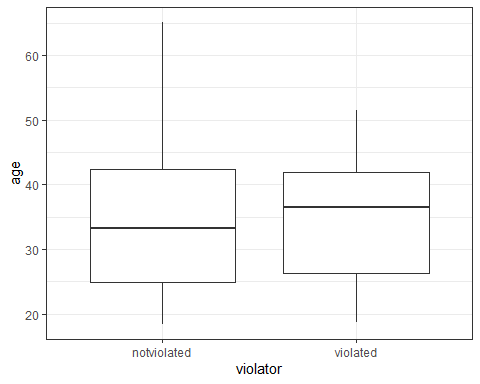
ggplot(train, aes(violator, fill = crime)) + geom\_bar(position="fill")



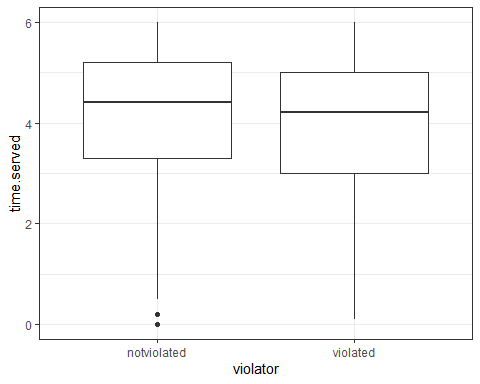
ggplot(train, aes(violator, fill = multiple.offenses)) + geom\_bar(position="fill")



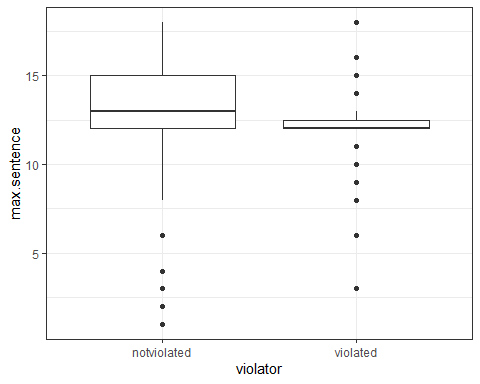
ggplot(train, aes(violator, age)) + geom\_boxplot() + theme\_bw()



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



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



The variables of the training set that appear to be most predictive of the response variable “violator” are “multiple.offenses”, “state”, and “race”. This is due in large part to the distributions we see between “notviolated” and “violated” within “violator” for each of these variables; there seems to be enough variance to make predictions as to whether or not someone will or will not violate their parole based on these variables. For the other variables, the distributions among “notviolated” and “violated” are relatively close, so we can assume that they might not be the best indicator of “violator”.

The most predictive variable of “violator” appears to be “multiple.offenses”. This makes sense logically as someone incarcerated with multiple offenses will probably have more strict parole terms, making it easier to violate their parole.

mod1 = glm(violator ~ multiple.offenses, train, family = "binomial")  
summary(mod1)

##   
## Call:  
## glm(formula = violator ~ multiple.offenses, family = "binomial",   
## data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.5610 -0.5610 -0.4089 -0.4089 2.2465   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.4398 0.2529 -9.648 <2e-16 \*\*\*  
## multiple.offensesmultiple 0.6702 0.3078 2.177 0.0295 \*   
## ---  
## 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: 335.02 on 471 degrees of freedom  
## AIC: 339.02  
##   
## Number of Fisher Scoring iterations: 5

In this model, we see that the “base” level of “multiple.offenses” that is being comapared to “violator” is “nonmultipleoffenses”, meaning the person does not have multiple offenses. This level is significant to “violator”. However, “multiple.offensesmultiple”, meaning the person does have multiple offenses, is not very signficant to violator. This is the opposite of what I thought, so I am going to run another model for one of the other variables I thought was significant. This model also has an AIC of 339.02.

mod2 = glm(violator ~ state, train, family = "binomial")  
summary(mod2)

##   
## 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

In this model, the base is “other” and it appears to be significant. The rest of the states are significant too, except for Kentucky. This makes sense since, when looking at the bar graph above, the distribution for Kentucky among “nonviolated” and “violated” has very little variance. This model also has an AIC of 283.18, which is less than the previous model (less is better).

mod3 = glm(violator ~., train, family = "binomial")  
summary(mod3) #manual model

##   
## 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   
## raceother 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 1.711032 0.396463 4.316 1.59e-05 \*\*\*  
## crimedrug 0.516479 0.739095 0.699 0.48468   
## crimeother 0.727043 0.690775 1.053 0.29257   
## crimelarceny 1.119953 0.797552 1.404 0.16025   
## ---  
## 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

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

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 1.73482 0.39421 4.401 1.08e-05 \*\*\*  
## raceother 1.09382 0.38974 2.807 0.00501 \*\*   
## ---  
## 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

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 \*\*\*  
## raceother 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 1.73482 0.39421 4.401 1.08e-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: 246.98 on 467 degrees of freedom  
## AIC: 258.98  
##   
## Number of Fisher Scoring iterations: 6

The forward stepwise and backward stepwise models both had lower AIC values than my manual model (a value of 258.98 compared to 268.09). Additionally, the forward stepwise and backward stepwise models both came to the same conclusion for significant variables - race, state, and multiple.offenses. I’ll call the backward stepwise my best model. In this model, stateVirginia and multiple.offensesmultiple are the most significant levels of the state and multiple.offenses variables respectively, while raceother is slightly less significant but not insignificant. This model seems to make sense/be intuitive based on these variables and the relatively low AIC.

mod4 = glm(violator ~ state + multiple.offenses + race, train, family = "binomial")  
summary(mod4)

##   
## 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 1.73482 0.39421 4.401 1.08e-05 \*\*\*  
## raceother 1.09382 0.38974 2.807 0.00501 \*\*   
## ---  
## 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

This model is similar to the forward and backward stepwise models; it has the same AIC value and significant levels of variables. The most significant levels of variables are stateVirginia and multiple.offensesmultiple, while raceother is slightly less significant

Parolee1 = data.frame(state = "Louisiana", multiple.offenses = "multiple", race = "white")  
predict(mod4, Parolee1, type = "response")

## 1   
## 0.3379961

Parolee2 = data.frame(state = "Kentucky", multiple.offenses = "notmultiple", race = "other")  
predict(mod4, Parolee2, type = "response")

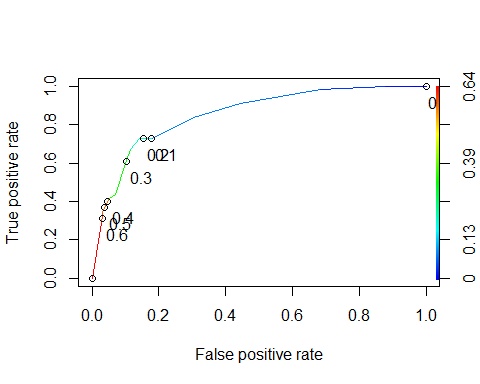
## 1   
## 0.2069629

Predicted probability of parole violation for Parolee1 and Parolee2 is 0.338 and 0.207 respectively.

predictions = predict(mod4, 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))



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

Sensitivity and specificity are 0.7272727 and 0.8588517 respectively.

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

##   
## FALSE TRUE  
## notviolated 359 59  
## violated 15 40

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

## [1] 0.8435518

Accuracy is 0.8435518. The implications of incorrectly classifying a parolee as someone who would violate or would not violate parole are that a person may be denied parole when they were unlikely to violate it in the first place, or that someone may be granted parole when they were likely to violate it. If someone who is granted parole and is likely to violate it does not violate it, then great. But someone granted parole, that is likely to violate, and does violate presents a greater risk to society depending on the crime and hurts the reputation of the court/entity that did grant their parole.

t2 = table(train$violator, predictions > 0.3)  
t2

##   
## FALSE TRUE  
## notviolated 376 42  
## violated 22 33

(t2[1,1]+t2[2,2])/nrow(train)

## [1] 0.8646934

t3 = table(train$violator, predictions > 0.4)  
t3

##   
## FALSE TRUE  
## notviolated 405 13  
## violated 36 19

(t3[1,1]+t3[2,2])/nrow(train)

## [1] 0.8964059

t4 = table(train$violator, predictions > 0.5)  
t4

##   
## FALSE TRUE  
## notviolated 405 13  
## violated 36 19

(t4[1,1]+t4[2,2])/nrow(train)

## [1] 0.8964059

t5 = table(train$violator, predictions > 0.6)  
t5

##   
## FALSE TRUE  
## notviolated 406 12  
## violated 39 16

(t5[1,1]+t5[2,2])/nrow(train)

## [1] 0.8921776

Accuracy starts to drop off after increasing cutoff past 0.5. Increased accuracy from about 84.3% to about 89.6%.

mod5 = glm(violator ~ state + multiple.offenses + race, test, family = "binomial")  
  
predictions2 = predict(mod5, type = "response")  
  
t5 = table(test$violator, predictions2 > 0.5)  
t5

##   
## FALSE TRUE  
## notviolated 175 4  
## violated 13 10

(t5[1,1]+t5[2,2])/nrow(test)

## [1] 0.9158416

After creating a model for the testing set using the same variables, we get an accuracy of about 91.6%.