# Module 3

## Christina Hall

### Logistic Regression (Classification)

tidyverse.quiet=TRUE  
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.5.2

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

## v ggplot2 3.1.0 v purrr 0.2.5  
## 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

## Warning: package 'ggplot2' was built under R version 3.5.2

## Warning: package 'tibble' was built under R version 3.5.2

## Warning: package 'tidyr' was built under R version 3.5.2

## Warning: package 'readr' was built under R version 3.5.2

## Warning: package 'purrr' was built under R version 3.5.2

## Warning: package 'dplyr' was built under R version 3.5.2

## Warning: package 'stringr' was built under R version 3.5.2

## Warning: package 'forcats' was built under R version 3.5.2

## -- 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)

## Warning: package 'caret' was built under R version 3.5.2

## Loading required package: lattice

##   
## Attaching package: 'caret'

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

library(ROCR)

## Warning: package 'ROCR' was built under R version 3.5.2

## Loading required package: gplots

## Warning: package 'gplots' was built under R version 3.5.2

##   
## Attaching package: 'gplots'

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

library(readr)  
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)

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, "Kentucky"="2", "Louisiana"="3", "Virginia"="4", "AnyOtherState"="1")) %>%  
 mutate(crime = as\_factor(as.character(crime))) %>%  
 mutate(crime = fct\_recode(crime, "larceny"="2", "drug-related"="3", "driving-related"="4", "AnyOtherCrime"="1")) %>%  
 mutate(multiple.offenses = as\_factor(as.character(multiple.offenses))) %>%  
 mutate(multiple.offenses = fct\_recode(multiple.offenses, "Multiple"="1", "otherwise"="0")) %>%  
 mutate(violator = as\_factor(as.character(violator))) %>%  
 mutate(violator = fct\_recode(violator, "violated"="1", "completed"="0"))

## Warning: package 'bindrcpp' was built under R version 3.5.2

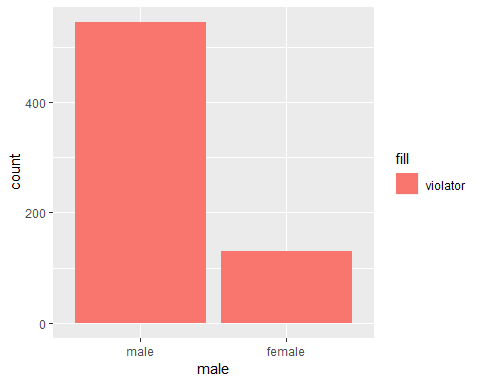
# Task 1

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

# Task 2

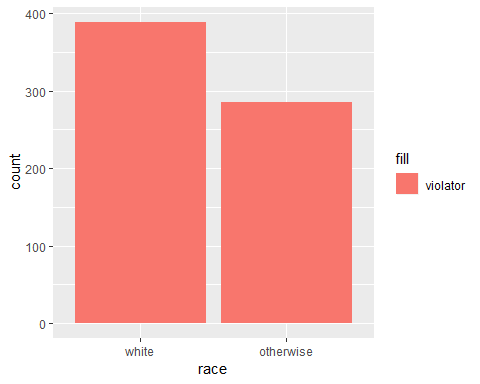
Male

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

 This data supports the notion that more males violated parole than females.

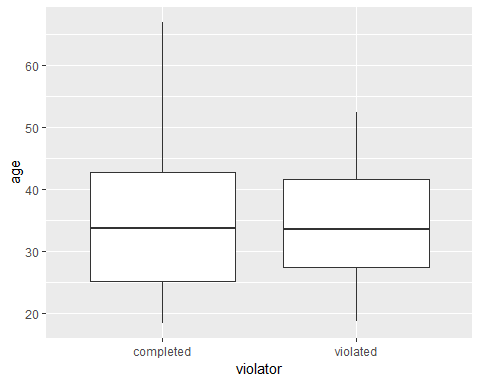
race

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

 This data supports that the white race violates parole more than combined other races.

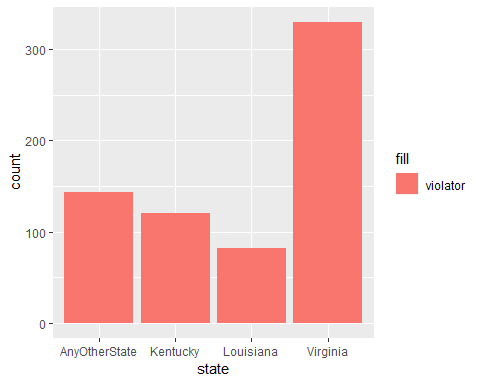
Age

ggplot(parole, aes(x=violator, y= age)) + geom\_boxplot()

 Age does not seem to be a good predictor of whether or not a parolee will violate parole.

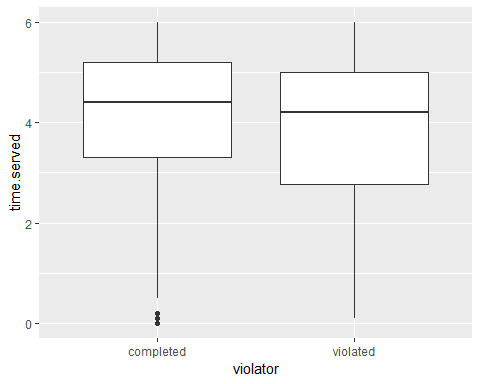
state

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

 Based on the data the state of Virginia parolees violate parole more than prolees in Louisiana.

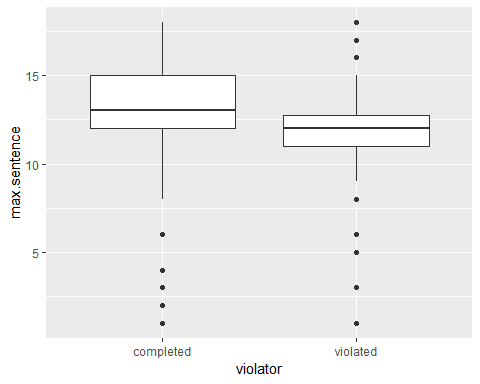
time.served

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

 Based on the data it is unlikely that time.served has a large influence on a parolee to violate parole.

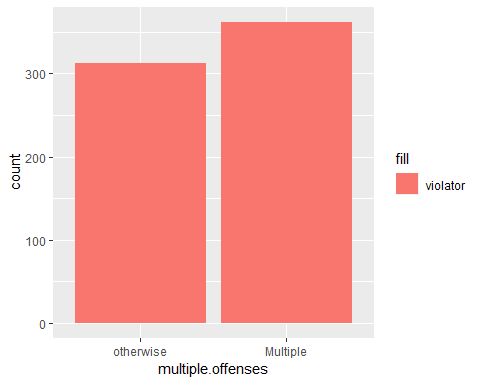
max.sentence

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

 Based on the data max.sentence has a small affect on a parolee violating parole.

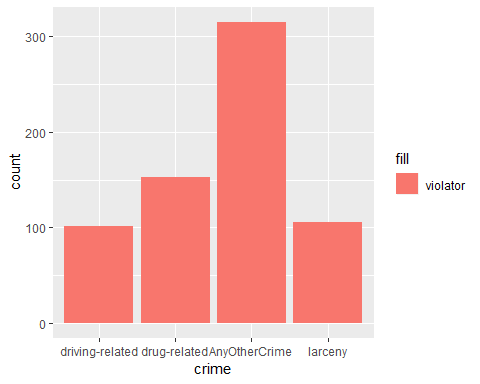
multiple.offenses

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

 Based on the data multiple offenses is a small indicator of whether or not parole will be violated.

crime

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

 Based on the data any crimes other than driving-related, drug-related, and larceny have a higher rate of violators.

State seems to be the most predictive of “violator.”

# Task 3

mod1 = glm(violator ~ state , parole, family = "binomial")  
summary(mod1)

##   
## Call:  
## glm(formula = violator ~ state, family = "binomial", data = parole)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.0955 -0.4981 -0.2071 -0.2071 2.7760   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.8165 0.2411 -7.534 4.92e-14 \*\*\*  
## stateKentucky -0.2079 0.3728 -0.558 0.577   
## stateLouisiana 1.6207 0.3277 4.946 7.58e-07 \*\*\*  
## stateVirginia -2.0153 0.4517 -4.461 8.15e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 483.27 on 674 degrees of freedom  
## Residual deviance: 382.89 on 671 degrees of freedom  
## AIC: 390.89  
##   
## Number of Fisher Scoring iterations: 6

The P values are below 1, but p value losses meaning as the data sets get larger. The coefficients make sence. The AIC is 390.89, which is low and show that the model is a good model.

# Task 4

allmod = glm(violator ~., parole, family = "binomial")   
summary(allmod)

##   
## Call:  
## glm(formula = violator ~ ., family = "binomial", data = parole)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6633 -0.4123 -0.2574 -0.1589 2.8738   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.182313 1.041740 -3.055 0.00225 \*\*   
## malefemale -0.270624 0.370506 -0.730 0.46513   
## raceotherwise 0.757252 0.324581 2.333 0.01965 \*   
## age 0.006554 0.013724 0.478 0.63297   
## stateKentucky 0.208399 0.417528 0.499 0.61769   
## stateLouisiana 0.893812 0.447042 1.999 0.04557 \*   
## stateVirginia -3.280842 0.526952 -6.226 4.78e-10 \*\*\*  
## time.served -0.076548 0.099531 -0.769 0.44184   
## max.sentence 0.053293 0.043824 1.216 0.22396   
## multiple.offensesMultiple 1.531547 0.325794 4.701 2.59e-06 \*\*\*  
## crimedrug-related -0.123182 0.542514 -0.227 0.82038   
## crimeAnyOtherCrime 0.157812 0.484705 0.326 0.74474   
## crimelarceny 0.494705 0.584656 0.846 0.39747   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 483.27 on 674 degrees of freedom  
## Residual deviance: 348.68 on 662 degrees of freedom  
## AIC: 374.68  
##   
## Number of Fisher Scoring iterations: 6

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

##   
## Call:  
## glm(formula = violator ~ 1, family = "binomial", data = parole)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.4956 -0.4956 -0.4956 -0.4956 2.0775   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.0352 0.1204 -16.9 <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: 483.27 on 674 degrees of freedom  
## Residual deviance: 483.27 on 674 degrees of freedom  
## AIC: 485.27  
##   
## Number of Fisher Scoring iterations: 4

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

## Start: AIC=485.27  
## violator ~ 1  
##   
## Df Deviance AIC  
## + state 3 382.89 390.89  
## + max.sentence 1 465.68 469.68  
## + multiple.offenses 1 475.81 479.81  
## + time.served 1 477.05 481.05  
## + race 1 479.56 483.56  
## <none> 483.27 485.27  
## + male 1 483.17 487.17  
## + age 1 483.25 487.25  
## + crime 3 480.48 488.48  
##   
## Step: AIC=390.89  
## violator ~ state  
##   
## Df Deviance AIC  
## + multiple.offenses 1 358.69 368.69  
## + race 1 376.71 386.71  
## <none> 382.89 390.89  
## + time.served 1 381.65 391.65  
## + max.sentence 1 381.93 391.93  
## + male 1 382.16 392.16  
## + age 1 382.87 392.87  
## + crime 3 380.87 394.87  
##   
## Step: AIC=368.69  
## violator ~ state + multiple.offenses  
##   
## Df Deviance AIC  
## + race 1 353.26 365.26  
## <none> 358.69 368.69  
## + max.sentence 1 356.73 368.73  
## + time.served 1 358.02 370.02  
## + male 1 358.04 370.04  
## + age 1 358.64 370.64  
## + crime 3 357.47 373.47  
##   
## Step: AIC=365.26  
## violator ~ state + multiple.offenses + race  
##   
## Df Deviance AIC  
## <none> 353.26 365.26  
## + max.sentence 1 351.62 365.62  
## + time.served 1 352.43 366.43  
## + male 1 352.71 366.71  
## + age 1 353.20 367.20  
## + crime 3 351.81 369.81

summary(forwardmod)

##   
## Call:  
## glm(formula = violator ~ state + multiple.offenses + race, family = "binomial",   
## data = parole)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.4012 -0.4051 -0.2604 -0.1801 2.8739   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.50359 0.30055 -8.330 < 2e-16 \*\*\*  
## stateKentucky 0.04449 0.39449 0.113 0.9102   
## stateLouisiana 0.75016 0.39147 1.916 0.0553 .   
## stateVirginia -3.12945 0.51147 -6.119 9.44e-10 \*\*\*  
## multiple.offensesMultiple 1.51964 0.32027 4.745 2.09e-06 \*\*\*  
## raceotherwise 0.74594 0.31828 2.344 0.0191 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 483.27 on 674 degrees of freedom  
## Residual deviance: 353.26 on 669 degrees of freedom  
## AIC: 365.26  
##   
## Number of Fisher Scoring iterations: 6

The quality of the model is good. The AIC is 365.26, which is lower than the module build with just state variable. The state, multiple.offenses, and male variables are significant. The model is intuitive based on the visual data collected in task 2.

# Task 5

train= train %>% dplyr::select(c(violator, state, multiple.offenses, race))  
train1= glm(violator ~ state + multiple.offenses + race, train, family= "binomial")  
summary(train1)

##   
## 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.offensesMultiple 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

AIC is 252.42, which is smaller than the forward stepwise model. The train1 model is a good model. Two of the state variables show a negative coefficient to show they decrease the violator coefficient. Multiple.offenses-multiple and race-otherwise show positive coefficient to show they increase the violator coefficient.

# Task 6

newdata= data.frame(state = "Louisiana", multiple.offenses = "Multiple", race = "white")  
predict(train1, newdata, type="response")

## 1   
## 0.408682

Parolee 1: 40.87% chance of violating parole

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

## 1   
## 0.152755

Parolee 2: 15.28% chane of violating parole

# Task 7

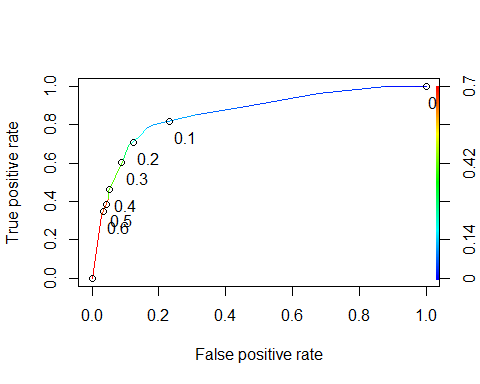
set.seed(12345)  
trainmod= train(violator~.,train, method= "glm")  
summary(trainmod)

##   
## Call:  
## NULL  
##   
## 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.offensesMultiple 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

predictions= predict( train1, type= "response")

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] 0.8586124

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 0.8373206  
## cutoff 0.1161882

# Task 8

t1 = table(train$violator,predictions > 0.1161882)  
t1

##   
## FALSE TRUE  
## completed 357 61  
## violated 14 41

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

## [1] 0.8414376

The accuracy, sensitivity, and dspecificity of the model is 84%. The implications of incorrectly classifying a parolee would cost extra money and risk to the community of the parolee.

# Task 9

t1 = table(train$violator,predictions > 0.5)  
t1

##   
## FALSE TRUE  
## completed 406 12  
## violated 37 18

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

## [1] 0.8964059

t1 = table(train$violator,predictions > 0.6)  
t1

##   
## FALSE TRUE  
## completed 406 12  
## violated 37 18

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

## [1] 0.8964059

The probability treshold that best describes maximizes the accuracy on the training set is 0.5.

# Task 10

test= test %>% dplyr::select(c(violator, state, multiple.offenses, race))  
test1= glm(violator ~ state + multiple.offenses + race, test, family= "binomial")  
summary(test1)

##   
## Call:  
## glm(formula = violator ~ state + multiple.offenses + race, family = "binomial",   
## data = test)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.3191 -0.5607 -0.1746 -0.1529 2.8950   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.4666 0.5282 -4.670 3.01e-06 \*\*\*  
## stateKentucky 0.6960 0.6275 1.109 0.26731   
## stateLouisiana 1.0817 0.7241 1.494 0.13522   
## stateVirginia -3.4208 1.1513 -2.971 0.00297 \*\*   
## multiple.offensesMultiple 1.4436 0.5854 2.466 0.01367 \*   
## raceotherwise 0.2683 0.5792 0.463 0.64318   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 143.22 on 201 degrees of freedom  
## Residual deviance: 107.85 on 196 degrees of freedom  
## AIC: 119.85  
##   
## Number of Fisher Scoring iterations: 7

testpredictions= predict( test1, type= "response")

t1 = table(test$violator,testpredictions > 0.5)  
t1

##   
## FALSE TRUE  
## completed 175 4  
## violated 16 7

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

## [1] 0.9009901

With the probablility threshold of .5, test model has a accuracy of 90%.