# Now we can model the data  
#loadrequiredlibrariesforpreprocessing(TRUE)  
setwd('/Users/dsing001/LV/R')  
library(ggplot2)  
library(discretization)

## Warning: package 'discretization' was built under R version 3.1.2

library(randomForest)

## Warning: package 'randomForest' was built under R version 3.1.2

## randomForest 4.6-10  
## Type rfNews() to see new features/changes/bug fixes.

library(ROCR)

## Warning: package 'ROCR' was built under R version 3.1.2

## Loading required package: gplots

## Warning: package 'gplots' was built under R version 3.1.2

## KernSmooth 2.23 loaded  
## Copyright M. P. Wand 1997-2009  
##   
## Attaching package: 'gplots'  
##   
## The following object is masked from 'package:stats':  
##   
## lowess

library(LV)  
library(unbalanced)

## Warning: package 'unbalanced' was built under R version 3.1.2

## Loading required package: FNN

## Warning: package 'FNN' was built under R version 3.1.2

## Loading required package: RANN

library(StatMatch)

## Loading required package: proxy  
##   
## Attaching package: 'proxy'  
##   
## The following objects are masked from 'package:stats':  
##   
## as.dist, dist  
##   
## Loading required package: clue  
## Loading required package: survey  
## Loading required package: grid  
##   
## Attaching package: 'survey'  
##   
## The following object is masked from 'package:graphics':  
##   
## dotchart

clscol='Class'  
  
# read the data which we saved as part of part 2 aka feature selection  
train\_data <- read.csv('fs\_train\_data.csv')  
test\_data <- read.csv('fs\_test\_data.csv')  
valid\_data <- read.csv('fs\_validdata.csv')  
  
train\_data[,clscol] <- as.factor(as.numeric(train\_data[,clscol] ))  
test\_data[,clscol] <- as.factor(as.numeric(test\_data[,clscol] ))  
valid\_data[,clscol] <- as.factor(as.numeric(valid\_data[,clscol] ))  
  
head(train\_data)

## Loan.Amount Loan.Term Employment.Length Home.Ownership Annual.Income  
## 1 1200 36 Months 10.0 OWN 54000  
## 2 10800 36 Months 6.0 OWN 32000  
## 3 4000 36 Months 0.5 MORTGAGE 148000  
## 4 3200 36 Months 5.0 OWN 54000  
## 5 10625 36 Months 10.0 MORTGAGE 34000  
## 6 2800 60 Months 0.5 OWN 24000  
## Loan.Purpose Address.State Debt.To.Income.Ratio  
## 1 Debt Consolidation AL 5.47  
## 2 Debt Consolidation AL 11.63  
## 3 0 AL 16.98  
## 4 Debt Consolidation AL 6.51  
## 5 Debt Consolidation 22.09  
## 6 Debt Consolidation AZ 4.00  
## No..Delinquencies.In.Last.2.Years Earliest.Credit.Line.Opened  
## 1 Eight 31068.22  
## 2 Eight 35411.32  
## 3 Eight 34095.41  
## 4 Eight 38872.24  
## 5 Eight 33085.06  
## 6 Eight 36865.50  
## FICO.Credit.Score No..Inquiries.In.Last.6.Months  
## 1 675 0  
## 2 680 1  
## 3 725 0  
## 4 740 0  
## 5 675 2  
## 6 705 0  
## Months.Since.Last.Delinquency No..Of.Credit.Lines  
## 1 64 5  
## 2 58 14  
## 3 0 11  
## 4 0 7  
## 5 57 5  
## 6 0 7  
## No..Adverse.Public.Records Total.Credit.Balance Use.Of.Credit.Line  
## 1 Four 2584 40.4  
## 2 Four 3511 25.6  
## 3 Four 50000 51.0  
## 4 Four 3198 51.1  
## 5 One 13846 71.0  
## 6 One 2183 19.5  
## Total.Number.Of.Credit.Lines Loan.Application.Description  
## 1 31 748  
## 2 40 403  
## 3 19 363  
## 4 11 141  
## 5 16 62  
## 6 23 302  
## No..Of.Public.Record.Bankruptcies Class  
## 1 None 0  
## 2 None 0  
## 3 None 0  
## 4 None 0  
## 5 One 0  
## 6 One 0

frml1 <- as.formula('Class ~ .')  
  
#oversample the rare class and undersample the class with creditworthy the data and then use that for modelling.  
train\_smt <- ubSMOTE(train\_data[,-ncol(train\_data)], train\_data$Class, perc.over = 200, k = 5, perc.under = 175, verbose = TRUE)

## Time SMOTE: 13.14 ; perc.over 200 ; perc.under 175 ; k 5

table(train\_smt$Y)

##   
## 0 1   
## 9884 8472

train\_data <- train\_smt$X  
train\_data$Class <- train\_smt$Y  
md\_prms <- train\_and\_predict\_log\_reg\_and\_ret\_auc(frml1,train\_data,valid\_data,predict\_type='response')  
auc <- md\_prms$auc  
mdl <- md\_prms$model  
tst\_with\_prob <- md\_prms$tst\_with\_prob  
#  
AUC <- auc$AUC  
GC <- (2\*AUC) - 1  
KS <- auc$KS  
KSRealized <- auc$KSRealized  
AUC

## [1] 0.6916989

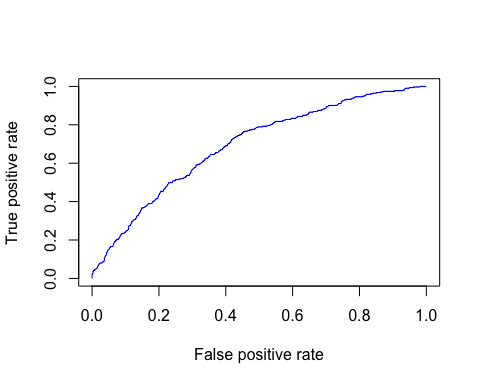
GC

## [1] 0.3833979

KS

## [1] 0.3119198

rocperf <- auc$rocperf  
plot(rocperf,col='blue',xlim = c(0,1), ylim = c(0,1))



#Plot the roc curve  
# 1- specificity  
fpr <- as.vector(attr(rocperf,'x.values')[[1]])  
tpr <- as.vector(attr(rocperf,'y.values')[[1]])  
pr <- data.frame(cbind(fpr,tpr))  
colnames(pr) <- c('fpr','tpr')  
  
  
  
cutoffvalues <- as.vector(attr(rocperf,'alpha.values')[[1]])  
cutoffvalue <- cutoffvalues[KSRealized]  
#cutoffvalue <- 0.5  
cutoffvalue

## [1] 0.256966

#generate the confusion matrix. To get the cutoff I will use the fact that we have KS score. Thus, where we have that value occuring we will have the best accuracy/recall/precision  
tst\_with\_prob$predclass <- ifelse(tst\_with\_prob$predprob>cutoffvalue,1,0)  
#missclassification each one example say where class 1 was predicted as class 0 and vice versa  
tst\_with\_prob[1422,]

## Loan.Amount Loan.Term Employment.Length Home.Ownership Annual.Income  
## 1422 14000 60 Months 0 MORTGAGE 50000  
## Loan.Purpose Address.State Debt.To.Income.Ratio  
## 1422 Medical FL 7.08  
## No..Delinquencies.In.Last.2.Years Earliest.Credit.Line.Opened  
## 1422 Eight 37096.22  
## FICO.Credit.Score No..Inquiries.In.Last.6.Months  
## 1422 660 3  
## Months.Since.Last.Delinquency No..Of.Credit.Lines  
## 1422 30 17  
## No..Adverse.Public.Records Total.Credit.Balance Use.Of.Credit.Line  
## 1422 Four 5132 41.4  
## Total.Number.Of.Credit.Lines Loan.Application.Description  
## 1422 35 204  
## No..Of.Public.Record.Bankruptcies Class predprob predclass  
## 1422 None 1 0.899709 1

tst\_with\_prob\_1 <- tst\_with\_prob[tst\_with\_prob$Class==1 & tst\_with\_prob$predclass==0,]  
dim(tst\_with\_prob\_1)

## [1] 75 23

tst\_with\_prob\_0 <- tst\_with\_prob[tst\_with\_prob$Class==0 & tst\_with\_prob$predclass==1,]  
dim(tst\_with\_prob\_0)

## [1] 640 23

tst\_with\_prob\_1 <- tst\_with\_prob\_1[1,]  
tst\_with\_prob\_0 <- tst\_with\_prob\_0[1,]  
tst\_with\_prob\_1

## Loan.Amount Loan.Term Employment.Length Home.Ownership Annual.Income  
## 1423 10000 36 Months 6 MORTGAGE 75000  
## Loan.Purpose Address.State Debt.To.Income.Ratio  
## 1423 Debt Consolidation AZ 11.14  
## No..Delinquencies.In.Last.2.Years Earliest.Credit.Line.Opened  
## 1423 Eight 37859.32  
## FICO.Credit.Score No..Inquiries.In.Last.6.Months  
## 1423 670 0  
## Months.Since.Last.Delinquency No..Of.Credit.Lines  
## 1423 0 17  
## No..Adverse.Public.Records Total.Credit.Balance Use.Of.Credit.Line  
## 1423 Four 11260 54.9  
## Total.Number.Of.Credit.Lines Loan.Application.Description  
## 1423 26 146  
## No..Of.Public.Record.Bankruptcies Class predprob predclass  
## 1423 None 1 0.2481038 0

tst\_with\_prob\_0

## Loan.Amount Loan.Term Employment.Length Home.Ownership Annual.Income  
## 1 23500 36 Months 5 MORTGAGE 46000  
## Loan.Purpose Address.State Debt.To.Income.Ratio  
## 1 Small Business AL 19.67  
## No..Delinquencies.In.Last.2.Years Earliest.Credit.Line.Opened  
## 1 Eight 35359.4  
## FICO.Credit.Score No..Inquiries.In.Last.6.Months  
## 1 735 1  
## Months.Since.Last.Delinquency No..Of.Credit.Lines  
## 1 0 9  
## No..Adverse.Public.Records Total.Credit.Balance Use.Of.Credit.Line  
## 1 Four 22738 45.9  
## Total.Number.Of.Credit.Lines Loan.Application.Description  
## 1 22 1000  
## No..Of.Public.Record.Bankruptcies Class predprob predclass  
## 1 None 0 0.4034823 1

#find all the data which is similar and show that  
k=5  
simitms\_1 <- order(gower.dist(tst\_with\_prob\_1,tst\_with\_prob))[seq(1:k)]  
#show 5 nearst point to the misclassified example  
tst\_with\_prob[simitms\_1,]

## Loan.Amount Loan.Term Employment.Length Home.Ownership Annual.Income  
## 1423 10000 36 Months 6 MORTGAGE 75000  
## 1523 8400 36 Months 5 MORTGAGE 94000  
## 1702 15000 36 Months 6 MORTGAGE 98000  
## 1418 5400 36 Months 6 MORTGAGE 30000  
## 1490 6400 36 Months 5 MORTGAGE 45600  
## Loan.Purpose Address.State Debt.To.Income.Ratio  
## 1423 Debt Consolidation AZ 11.14  
## 1523 Debt Consolidation AZ 22.54  
## 1702 Debt Consolidation AZ 11.41  
## 1418 Debt Consolidation AZ 13.10  
## 1490 Debt Consolidation AL 13.39  
## No..Delinquencies.In.Last.2.Years Earliest.Credit.Line.Opened  
## 1423 Eight 37859.32  
## 1523 Eight 36003.26  
## 1702 Eight 35684.06  
## 1418 Eight 39056.53  
## 1490 Eight 37249.38  
## FICO.Credit.Score No..Inquiries.In.Last.6.Months  
## 1423 670 0  
## 1523 680 1  
## 1702 675 0  
## 1418 685 0  
## 1490 715 0  
## Months.Since.Last.Delinquency No..Of.Credit.Lines  
## 1423 0 17  
## 1523 0 14  
## 1702 0 10  
## 1418 0 9  
## 1490 0 9  
## No..Adverse.Public.Records Total.Credit.Balance Use.Of.Credit.Line  
## 1423 Four 11260 54.9  
## 1523 Four 50000 81.5  
## 1702 Four 4414 81.7  
## 1418 Four 7589 70.9  
## 1490 Four 7486 53.5  
## Total.Number.Of.Credit.Lines Loan.Application.Description  
## 1423 26 146  
## 1523 30 73  
## 1702 14 84  
## 1418 16 0  
## 1490 22 405  
## No..Of.Public.Record.Bankruptcies Class predprob predclass  
## 1423 None 1 0.2481038 0  
## 1523 None 1 0.2551118 0  
## 1702 None 1 0.2810061 1  
## 1418 None 1 0.3228883 1  
## 1490 None 1 0.1479780 0

simitms\_0 <- order(gower.dist(tst\_with\_prob\_0,tst\_with\_prob))[seq(1:k)]  
#show 5 nearst point to the misclassified example  
tst\_with\_prob[simitms\_0,]

## Loan.Amount Loan.Term Employment.Length Home.Ownership Annual.Income  
## 1 23500 36 Months 5 MORTGAGE 46000  
## 630 22250 36 Months 10 MORTGAGE 42000  
## 851 9000 36 Months 1 MORTGAGE 92000  
## 170 14250 36 Months 3 MORTGAGE 87950  
## 1174 8400 36 Months 7 MORTGAGE 60000  
## Loan.Purpose Address.State Debt.To.Income.Ratio  
## 1 Small Business AL 19.67  
## 630 Small Business AL 6.14  
## 851 Small Business FL 15.85  
## 170 Small Business AL 12.20  
## 1174 Medical AL 19.04  
## No..Delinquencies.In.Last.2.Years Earliest.Credit.Line.Opened  
## 1 Eight 35359.40  
## 630 Eight 36227.22  
## 851 Eight 35413.51  
## 170 Eight 35041.05  
## 1174 Eight 36585.12  
## FICO.Credit.Score No..Inquiries.In.Last.6.Months  
## 1 735 1  
## 630 745 1  
## 851 730 0  
## 170 750 3  
## 1174 665 1  
## Months.Since.Last.Delinquency No..Of.Credit.Lines  
## 1 0 9  
## 630 0 10  
## 851 0 10  
## 170 0 9  
## 1174 0 13  
## No..Adverse.Public.Records Total.Credit.Balance Use.Of.Credit.Line  
## 1 Four 22738 45.9  
## 630 Four 9153 36.5  
## 851 Four 27458 63.4  
## 170 Four 39 0.3  
## 1174 Four 11320 52.7  
## Total.Number.Of.Credit.Lines Loan.Application.Description  
## 1 22 1000  
## 630 26 121  
## 851 32 1000  
## 170 25 350  
## 1174 20 814  
## No..Of.Public.Record.Bankruptcies Class predprob predclass  
## 1 None 0 0.4034823 1  
## 630 None 0 0.4565923 1  
## 851 None 0 0.3851382 1  
## 170 None 0 0.3024482 1  
## 1174 None 0 0.3322827 1

#calculate accurayce and recall  
ab <- table(tst\_with\_prob$predclass,tst\_with\_prob$Class)  
ab

##   
## 0 1  
## 0 777 75  
## 1 640 238

recall <- ab[2,2]/(ab[1,2] + ab[2,2])  
recall

## [1] 0.7603834

acc <- (ab[2,2] + ab[1,1])/(sum(ab))  
acc

## [1] 0.5867052

#part below this should be run only when you are fine with your model on validation data and you should not cheat by running below an calibrating your model :)  
#on test data accuracy  
train\_data <- rbind(train\_data,valid\_data)  
md\_prms <- train\_and\_predict\_log\_reg\_and\_ret\_auc(frml1,train\_data,test\_data,predict\_type='response')  
auc <- md\_prms$auc  
mdl <- md\_prms$model  
tst\_with\_prob <- md\_prms$tst\_with\_prob  
#  
AUC <- auc$AUC  
GC <- (2\*AUC) - 1  
KS <- auc$KS  
KSRealized <- auc$KSRealized  
AUC

## [1] 0.7022514

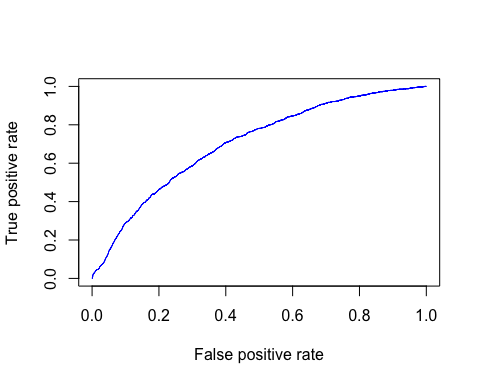
GC

## [1] 0.4045029

KS

## [1] 0.3093562

rocperf <- auc$rocperf  
plot(rocperf,col='blue',xlim = c(0,1), ylim = c(0,1))



#Plot the roc curve  
# 1- specificity  
fpr <- as.vector(attr(rocperf,'x.values')[[1]])  
tpr <- as.vector(attr(rocperf,'y.values')[[1]])  
pr <- data.frame(cbind(fpr,tpr))  
colnames(pr) <- c('fpr','tpr')  
  
  
  
cutoffvalues <- as.vector(attr(rocperf,'alpha.values')[[1]])  
cutoffvalue <- cutoffvalues[KSRealized]  
#cutoffvalue <- 0.5  
cutoffvalue

## [1] 0.2756175

#generate the confusion matrix. To get the cutoff I will use the fact that we have KS score. Thus, where we have that value occuring we will have the best accuracy/recall/precision  
tst\_with\_prob$predclass <- ifelse(tst\_with\_prob$predprob>cutoffvalue,1,0)  
#missclassification each one example say where class 1 was predicted as class 0 and vice versa  
tst\_with\_prob[1422,]

## Loan.Amount Loan.Term Employment.Length Home.Ownership Annual.Income  
## 1422 3600 36 Months 0.5 OWN 50000  
## Loan.Purpose Address.State Debt.To.Income.Ratio  
## 1422 Debt Consolidation AZ 23.21  
## No..Delinquencies.In.Last.2.Years Earliest.Credit.Line.Opened  
## 1422 Eight 38319.35  
## FICO.Credit.Score No..Inquiries.In.Last.6.Months  
## 1422 685 3  
## Months.Since.Last.Delinquency No..Of.Credit.Lines  
## 1422 33 16  
## No..Adverse.Public.Records Total.Credit.Balance Use.Of.Credit.Line  
## 1422 Four 8792 43.5  
## Total.Number.Of.Credit.Lines Loan.Application.Description  
## 1422 22 0  
## No..Of.Public.Record.Bankruptcies Class predprob predclass  
## 1422 None 0 0.343899 1

tst\_with\_prob\_1 <- tst\_with\_prob[tst\_with\_prob$Class==1 & tst\_with\_prob$predclass==0,]  
dim(tst\_with\_prob\_1)

## [1] 398 23

tst\_with\_prob\_0 <- tst\_with\_prob[tst\_with\_prob$Class==0 & tst\_with\_prob$predclass==1,]  
dim(tst\_with\_prob\_0)

## [1] 2400 23

tst\_with\_prob\_1 <- tst\_with\_prob\_1[1,]  
tst\_with\_prob\_0 <- tst\_with\_prob\_0[1,]  
tst\_with\_prob\_1

## Loan.Amount Loan.Term Employment.Length Home.Ownership Annual.Income  
## 6073 9000 36 Months 8 OWN 30000  
## Loan.Purpose Address.State Debt.To.Income.Ratio  
## 6073 0 AZ 11.68  
## No..Delinquencies.In.Last.2.Years Earliest.Credit.Line.Opened  
## 6073 Eight 35520.4  
## FICO.Credit.Score No..Inquiries.In.Last.6.Months  
## 6073 755 0  
## Months.Since.Last.Delinquency No..Of.Credit.Lines  
## 6073 0 6  
## No..Adverse.Public.Records Total.Credit.Balance Use.Of.Credit.Line  
## 6073 Four 9358 48.7  
## Total.Number.Of.Credit.Lines Loan.Application.Description  
## 6073 18 0  
## No..Of.Public.Record.Bankruptcies Class predprob predclass  
## 6073 None 1 0.1392759 0

tst\_with\_prob\_0

## Loan.Amount Loan.Term Employment.Length Home.Ownership Annual.Income  
## 1 6000 36 Months 6 OWN 60000  
## Loan.Purpose Address.State Debt.To.Income.Ratio  
## 1 Debt Consolidation AL 2.2  
## No..Delinquencies.In.Last.2.Years Earliest.Credit.Line.Opened  
## 1 Eight 36657.24  
## FICO.Credit.Score No..Inquiries.In.Last.6.Months  
## 1 695 2  
## Months.Since.Last.Delinquency No..Of.Credit.Lines  
## 1 57 3  
## No..Adverse.Public.Records Total.Credit.Balance Use.Of.Credit.Line  
## 1 One 3719 39.1  
## Total.Number.Of.Credit.Lines Loan.Application.Description  
## 1 16 0  
## No..Of.Public.Record.Bankruptcies Class predprob predclass  
## 1 None 0 0.5067637 1

#find all the data which is similar and show that  
k=5  
simitms\_1 <- order(gower.dist(tst\_with\_prob\_1,tst\_with\_prob))[seq(1:k)]  
#show 5 nearst point to the misclassified example  
tst\_with\_prob[simitms\_1,]

## Loan.Amount Loan.Term Employment.Length Home.Ownership Annual.Income  
## 6073 9000 36 Months 8 OWN 30000  
## 7220 10000 36 Months 7 OWN 30000  
## 6243 19000 36 Months 10 OWN 54000  
## 6876 15000 36 Months 7 OWN 95600  
## 6965 6000 36 Months 6 OWN 48000  
## Loan.Purpose Address.State Debt.To.Income.Ratio  
## 6073 0 AZ 11.68  
## 7220 0 AZ 8.88  
## 6243 0 AZ 7.00  
## 6876 0 AZ 15.21  
## 6965 0 AZ 25.85  
## No..Delinquencies.In.Last.2.Years Earliest.Credit.Line.Opened  
## 6073 Eight 35520.40  
## 7220 Eight 37358.15  
## 6243 Eight 35497.36  
## 6876 Eight 36390.35  
## 6965 Eight 37495.35  
## FICO.Credit.Score No..Inquiries.In.Last.6.Months  
## 6073 755 0  
## 7220 725 0  
## 6243 750 0  
## 6876 730 0  
## 6965 725 0  
## Months.Since.Last.Delinquency No..Of.Credit.Lines  
## 6073 0 6  
## 7220 0 12  
## 6243 0 3  
## 6876 0 12  
## 6965 0 11  
## No..Adverse.Public.Records Total.Credit.Balance Use.Of.Credit.Line  
## 6073 Four 9358 48.7  
## 7220 Four 13689 46.2  
## 6243 Four 2606 57.9  
## 6876 Four 9723 40.7  
## 6965 Four 4286 27.8  
## Total.Number.Of.Credit.Lines Loan.Application.Description  
## 6073 18 0  
## 7220 18 0  
## 6243 4 1  
## 6876 25 0  
## 6965 16 0  
## No..Of.Public.Record.Bankruptcies Class predprob predclass  
## 6073 None 1 0.13927586 0  
## 7220 None 1 0.17858365 0  
## 6243 None 1 0.15094940 0  
## 6876 None 1 0.09220341 0  
## 6965 None 1 0.11910256 0

simitms\_0 <- order(gower.dist(tst\_with\_prob\_0,tst\_with\_prob))[seq(1:k)]  
#show 5 nearst point to the misclassified example  
tst\_with\_prob[simitms\_0,]

## Loan.Amount Loan.Term Employment.Length Home.Ownership Annual.Income  
## 1 6000 36 Months 6 OWN 60000  
## 1427 11000 36 Months 2 OWN 52000  
## 4413 3600 36 Months 8 OWN 54996  
## 1386 10000 36 Months 3 OWN 28600  
## 3292 2400 36 Months 10 OWN 50000  
## Loan.Purpose Address.State Debt.To.Income.Ratio  
## 1 Debt Consolidation AL 2.20  
## 1427 Debt Consolidation AL 9.39  
## 4413 Debt Consolidation AZ 1.24  
## 1386 Medical AL 5.37  
## 3292 Debt Consolidation AL 1.13  
## No..Delinquencies.In.Last.2.Years Earliest.Credit.Line.Opened  
## 1 Eight 36657.24  
## 1427 Eight 36800.15  
## 4413 Eight 33600.15  
## 1386 Eight 36768.45  
## 3292 Eight 36916.17  
## FICO.Credit.Score No..Inquiries.In.Last.6.Months  
## 1 695 2  
## 1427 685 0  
## 4413 720 2  
## 1386 685 2  
## 3292 680 2  
## Months.Since.Last.Delinquency No..Of.Credit.Lines  
## 1 57 3  
## 1427 0 10  
## 4413 37 7  
## 1386 44 2  
## 3292 27 4  
## No..Adverse.Public.Records Total.Credit.Balance Use.Of.Credit.Line  
## 1 One 3719 39.1  
## 1427 One 12678 43.6  
## 4413 One 1919 19.8  
## 1386 One 3245 54.1  
## 3292 One 1429 40.8  
## Total.Number.Of.Credit.Lines Loan.Application.Description  
## 1 16 0  
## 1427 14 187  
## 4413 29 19  
## 1386 6 31  
## 3292 21 196  
## No..Of.Public.Record.Bankruptcies Class predprob predclass  
## 1 None 0 0.5067637 1  
## 1427 None 0 0.5504471 1  
## 4413 None 0 0.5207685 1  
## 1386 None 0 0.7982077 1  
## 3292 One 0 0.8627278 1

#calculate accurayce and recall  
ab <- table(tst\_with\_prob$predclass,tst\_with\_prob$Class)  
ab

##   
## 0 1  
## 0 3672 398  
## 1 2400 946

recall <- ab[2,2]/(ab[1,2] + ab[2,2])  
recall

## [1] 0.703869

acc <- (ab[2,2] + ab[1,1])/(sum(ab))  
acc

## [1] 0.6227077