# 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$No..Of.Credit.Lines <- NULL  
test\_data$No..Of.Credit.Lines <- NULL  
valid\_data$No..Of.Credit.Lines <- NULL  
  
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] ))  
  
#Now add the interaction terms  
# from decision tree  
  
frml1 <- 'Class ~ . + Loan.Purpose:FICO.Credit.Score + Loan.Purpose:Use.Of.Credit.Line + Loan.Term:FICO.Credit.Score + Loan.Term:Annual.Income + Loan.Term:Use.Of.Credit.Line + Annual.Income:FICO.Credit.Score + Loan.Term:Loan.Purpose + FICO.Credit.Score:No..Inquiries.In.Last.6.Months + No..Inquiries.In.Last.6.Months:Use.Of.Credit.Line + FICO.Credit.Score:Total.Number.Of.Credit.Lines + Loan.Term:No..Inquiries.In.Last.6.Months'  
  
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] 0.7133642

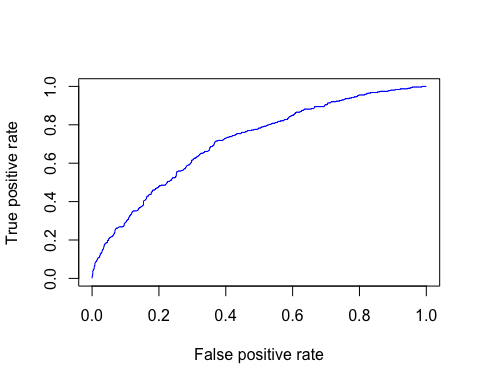
GC

## [1] 0.4267284

KS

## [1] 0.3458596

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] 0.1718012

#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..Adverse.Public.Records  
## 1422 30 Four  
## Total.Credit.Balance Use.Of.Credit.Line Total.Number.Of.Credit.Lines  
## 1422 5132 41.4 35  
## Loan.Application.Description No..Of.Public.Record.Bankruptcies Class  
## 1422 204 None 1  
## predprob predclass  
## 1422 0.5520149 1

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

## [1] 90 22

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

## [1] 524 22

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..Adverse.Public.Records  
## 1423 0 Four  
## Total.Credit.Balance Use.Of.Credit.Line Total.Number.Of.Credit.Lines  
## 1423 11260 54.9 26  
## Loan.Application.Description No..Of.Public.Record.Bankruptcies Class  
## 1423 146 None 1  
## predprob predclass  
## 1423 0.1531905 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..Adverse.Public.Records  
## 1 0 Four  
## Total.Credit.Balance Use.Of.Credit.Line Total.Number.Of.Credit.Lines  
## 1 22738 45.9 22  
## Loan.Application.Description No..Of.Public.Record.Bankruptcies Class  
## 1 1000 None 0  
## predprob predclass  
## 1 0.2900987 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  
## 1702 15000 36 Months 6 MORTGAGE 98000  
## 1418 5400 36 Months 6 MORTGAGE 30000  
## 1490 6400 36 Months 5 MORTGAGE 45600  
## 607 13725 36 Months 5 MORTGAGE 80000  
## Loan.Purpose Address.State Debt.To.Income.Ratio  
## 1423 Debt Consolidation AZ 11.14  
## 1702 Debt Consolidation AZ 11.41  
## 1418 Debt Consolidation AZ 13.10  
## 1490 Debt Consolidation AL 13.39  
## 607 Debt Consolidation AZ 20.68  
## No..Delinquencies.In.Last.2.Years Earliest.Credit.Line.Opened  
## 1423 Eight 37859.32  
## 1702 Eight 35684.06  
## 1418 Eight 39056.53  
## 1490 Eight 37249.38  
## 607 Eight 36140.38  
## FICO.Credit.Score No..Inquiries.In.Last.6.Months  
## 1423 670 0  
## 1702 675 0  
## 1418 685 0  
## 1490 715 0  
## 607 695 0  
## Months.Since.Last.Delinquency No..Adverse.Public.Records  
## 1423 0 Four  
## 1702 0 Four  
## 1418 0 Four  
## 1490 0 Four  
## 607 0 Four  
## Total.Credit.Balance Use.Of.Credit.Line Total.Number.Of.Credit.Lines  
## 1423 11260 54.9 26  
## 1702 4414 81.7 14  
## 1418 7589 70.9 16  
## 1490 7486 53.5 22  
## 607 27629 61.9 28  
## Loan.Application.Description No..Of.Public.Record.Bankruptcies Class  
## 1423 146 None 1  
## 1702 84 None 1  
## 1418 0 None 1  
## 1490 405 None 1  
## 607 172 None 0  
## predprob predclass  
## 1423 0.15319048 0  
## 1702 0.17180120 0  
## 1418 0.19962647 1  
## 1490 0.08091781 0  
## 607 0.14387758 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  
## 1174 8400 36 Months 7 MORTGAGE 60000  
## 243 15000 36 Months 1 MORTGAGE 85000  
## 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  
## 1174 Medical AL 19.04  
## 243 Small Business AZ 11.92  
## No..Delinquencies.In.Last.2.Years Earliest.Credit.Line.Opened  
## 1 Eight 35359.40  
## 630 Eight 36227.22  
## 851 Eight 35413.51  
## 1174 Eight 36585.12  
## 243 Eight 34425.37  
## FICO.Credit.Score No..Inquiries.In.Last.6.Months  
## 1 735 1  
## 630 745 1  
## 851 730 0  
## 1174 665 1  
## 243 720 0  
## Months.Since.Last.Delinquency No..Adverse.Public.Records  
## 1 0 Four  
## 630 0 Four  
## 851 0 Four  
## 1174 0 Four  
## 243 0 Four  
## Total.Credit.Balance Use.Of.Credit.Line Total.Number.Of.Credit.Lines  
## 1 22738 45.9 22  
## 630 9153 36.5 26  
## 851 27458 63.4 32  
## 1174 11320 52.7 20  
## 243 27662 73.4 41  
## Loan.Application.Description No..Of.Public.Record.Bankruptcies Class  
## 1 1000 None 0  
## 630 121 None 0  
## 851 1000 None 0  
## 1174 814 None 0  
## 243 1000 None 0  
## predprob predclass  
## 1 0.2900987 1  
## 630 0.2662893 1  
## 851 0.2620036 1  
## 1174 0.2031360 1  
## 243 0.2665185 1

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

##   
## 0 1  
## 0 893 90  
## 1 524 223

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

## [1] 0.7124601

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

## [1] 0.6450867

#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] 0.7152062

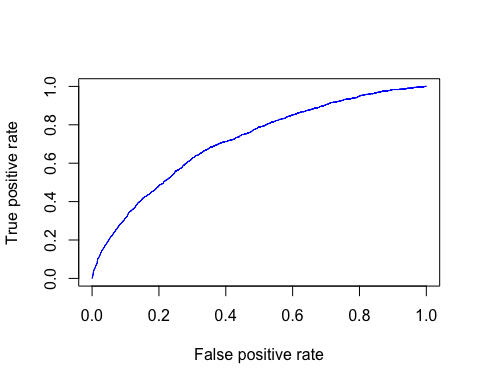
GC

## [1] 0.4304124

KS

## [1] 0.3307718

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] 0.1818469

#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..Adverse.Public.Records  
## 1422 33 Four  
## Total.Credit.Balance Use.Of.Credit.Line Total.Number.Of.Credit.Lines  
## 1422 8792 43.5 22  
## Loan.Application.Description No..Of.Public.Record.Bankruptcies Class  
## 1422 0 None 0  
## predprob predclass  
## 1422 0.2541424 1

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

## [1] 452 22

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

## [1] 2026 22

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..Adverse.Public.Records  
## 6073 0 Four  
## Total.Credit.Balance Use.Of.Credit.Line Total.Number.Of.Credit.Lines  
## 6073 9358 48.7 18  
## Loan.Application.Description No..Of.Public.Record.Bankruptcies Class  
## 6073 0 None 1  
## predprob predclass  
## 6073 0.0901913 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..Adverse.Public.Records  
## 1 57 One  
## Total.Credit.Balance Use.Of.Credit.Line Total.Number.Of.Credit.Lines  
## 1 3719 39.1 16  
## Loan.Application.Description No..Of.Public.Record.Bankruptcies Class  
## 1 0 None 0  
## predprob predclass  
## 1 0.2225536 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  
## 6876 15000 36 Months 7 OWN 95600  
## 6243 19000 36 Months 10 OWN 54000  
## 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  
## 6876 0 AZ 15.21  
## 6243 0 AZ 7.00  
## 6965 0 AZ 25.85  
## No..Delinquencies.In.Last.2.Years Earliest.Credit.Line.Opened  
## 6073 Eight 35520.40  
## 7220 Eight 37358.15  
## 6876 Eight 36390.35  
## 6243 Eight 35497.36  
## 6965 Eight 37495.35  
## FICO.Credit.Score No..Inquiries.In.Last.6.Months  
## 6073 755 0  
## 7220 725 0  
## 6876 730 0  
## 6243 750 0  
## 6965 725 0  
## Months.Since.Last.Delinquency No..Adverse.Public.Records  
## 6073 0 Four  
## 7220 0 Four  
## 6876 0 Four  
## 6243 0 Four  
## 6965 0 Four  
## Total.Credit.Balance Use.Of.Credit.Line Total.Number.Of.Credit.Lines  
## 6073 9358 48.7 18  
## 7220 13689 46.2 18  
## 6876 9723 40.7 25  
## 6243 2606 57.9 4  
## 6965 4286 27.8 16  
## Loan.Application.Description No..Of.Public.Record.Bankruptcies Class  
## 6073 0 None 1  
## 7220 0 None 1  
## 6876 0 None 1  
## 6243 1 None 1  
## 6965 0 None 1  
## predprob predclass  
## 6073 0.09019130 0  
## 7220 0.10618364 0  
## 6876 0.05914413 0  
## 6243 0.10486826 0  
## 6965 0.08049394 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  
## 1601 3000 36 Months 2 OWN 47208  
## 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  
## 1601 Debt Consolidation AL 3.05  
## 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  
## 1601 Eight 37541.11  
## FICO.Credit.Score No..Inquiries.In.Last.6.Months  
## 1 695 2  
## 1427 685 0  
## 4413 720 2  
## 1386 685 2  
## 1601 675 3  
## Months.Since.Last.Delinquency No..Adverse.Public.Records  
## 1 57 One  
## 1427 0 One  
## 4413 37 One  
## 1386 44 One  
## 1601 24 Four  
## Total.Credit.Balance Use.Of.Credit.Line Total.Number.Of.Credit.Lines  
## 1 3719 39.1 16  
## 1427 12678 43.6 14  
## 4413 1919 19.8 29  
## 1386 3245 54.1 6  
## 1601 4339 51.5 11  
## Loan.Application.Description No..Of.Public.Record.Bankruptcies Class  
## 1 0 None 0  
## 1427 187 None 0  
## 4413 19 None 0  
## 1386 31 None 0  
## 1601 47 None 0  
## predprob predclass  
## 1 0.2225536 1  
## 1427 0.2325213 1  
## 4413 0.2295266 1  
## 1386 0.4436585 1  
## 1601 0.2109709 1

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

##   
## 0 1  
## 0 4046 452  
## 1 2026 892

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

## [1] 0.6636905

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

## [1] 0.6658576