# 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 <- as.formula('Class ~ .')  
  
md\_prms <- train\_and\_predict\_random\_forest\_and\_ret\_auc(frml1,train\_data,valid\_data)  
auc <- md\_prms$auc  
tst\_with\_prob <- md\_prms$tst\_with\_prob  
#  
AUC <- auc$AUC  
GC <- (2\*AUC) - 1  
KS <- auc$KS  
KSRealized <- auc$KSRealized  
AUC

## [1] 0.7213728

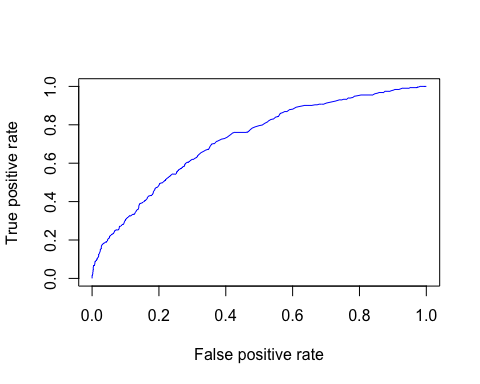
GC

## [1] 0.4427457

KS

## [1] 0.3418823

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] 0.198

#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 0.47 1

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

## [1] 98 22

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

## [1] 500 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 0.082 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 0.21 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  
## 85 10000 36 Months 5 MORTGAGE 69000  
## 1490 6400 36 Months 5 MORTGAGE 45600  
## 1480 2800 36 Months 4 MORTGAGE 88000  
## 1648 15000 36 Months 3 OWN 80000  
## Loan.Purpose Address.State Debt.To.Income.Ratio  
## 1423 Debt Consolidation AZ 11.14  
## 85 Debt Consolidation AZ 7.32  
## 1490 Debt Consolidation AL 13.39  
## 1480 Debt Consolidation AZ 13.66  
## 1648 Debt Consolidation AZ 11.84  
## No..Delinquencies.In.Last.2.Years Earliest.Credit.Line.Opened  
## 1423 Eight 37859.32  
## 85 Eight 37165.23  
## 1490 Eight 37249.38  
## 1480 Eight 37006.25  
## 1648 Eight 35898.36  
## FICO.Credit.Score No..Inquiries.In.Last.6.Months  
## 1423 670 0  
## 85 685 2  
## 1490 715 0  
## 1480 660 1  
## 1648 680 1  
## Months.Since.Last.Delinquency No..Adverse.Public.Records  
## 1423 0 Four  
## 85 0 Four  
## 1490 0 Four  
## 1480 56 Four  
## 1648 0 Four  
## Total.Credit.Balance Use.Of.Credit.Line Total.Number.Of.Credit.Lines  
## 1423 11260 54.9 26  
## 85 9859 63.3 18  
## 1490 7486 53.5 22  
## 1480 6050 79.6 17  
## 1648 7848 46.8 24  
## Loan.Application.Description No..Of.Public.Record.Bankruptcies Class  
## 1423 146 None 1  
## 85 322 None 0  
## 1490 405 None 1  
## 1480 159 None 1  
## 1648 275 None 1  
## predprob predclass  
## 1423 0.082 0  
## 85 0.086 0  
## 1490 0.088 0  
## 1480 0.186 0  
## 1648 0.084 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  
## 1174 8400 36 Months 7 MORTGAGE 60000  
## 851 9000 36 Months 1 MORTGAGE 92000  
## 1301 7400 36 Months 4 MORTGAGE 94800  
## Loan.Purpose Address.State Debt.To.Income.Ratio  
## 1 Small Business AL 19.67  
## 630 Small Business AL 6.14  
## 1174 Medical AL 19.04  
## 851 Small Business FL 15.85  
## 1301 Small Business AL 10.22  
## No..Delinquencies.In.Last.2.Years Earliest.Credit.Line.Opened  
## 1 Eight 35359.40  
## 630 Eight 36227.22  
## 1174 Eight 36585.12  
## 851 Eight 35413.51  
## 1301 Eight 37277.06  
## FICO.Credit.Score No..Inquiries.In.Last.6.Months  
## 1 735 1  
## 630 745 1  
## 1174 665 1  
## 851 730 0  
## 1301 755 0  
## Months.Since.Last.Delinquency No..Adverse.Public.Records  
## 1 0 Four  
## 630 0 Four  
## 1174 0 Four  
## 851 0 Four  
## 1301 0 Four  
## Total.Credit.Balance Use.Of.Credit.Line Total.Number.Of.Credit.Lines  
## 1 22738 45.9 22  
## 630 9153 36.5 26  
## 1174 11320 52.7 20  
## 851 27458 63.4 32  
## 1301 0 0.0 15  
## Loan.Application.Description No..Of.Public.Record.Bankruptcies Class  
## 1 1000 None 0  
## 630 121 None 0  
## 1174 814 None 0  
## 851 1000 None 0  
## 1301 723 None 0  
## predprob predclass  
## 1 0.210 1  
## 630 0.250 1  
## 1174 0.204 1  
## 851 0.250 1  
## 1301 0.208 1

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

##   
## 0 1  
## 0 917 98  
## 1 500 215

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

## [1] 0.686901

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

## [1] 0.6543353

#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\_random\_forest\_and\_ret\_auc(frml1,train\_data,test\_data)  
auc <- md\_prms$auc  
  
tst\_with\_prob <- md\_prms$tst\_with\_prob  
#  
AUC <- auc$AUC  
GC <- (2\*AUC) - 1  
KS <- auc$KS  
KSRealized <- auc$KSRealized  
AUC

## [1] 0.7163036

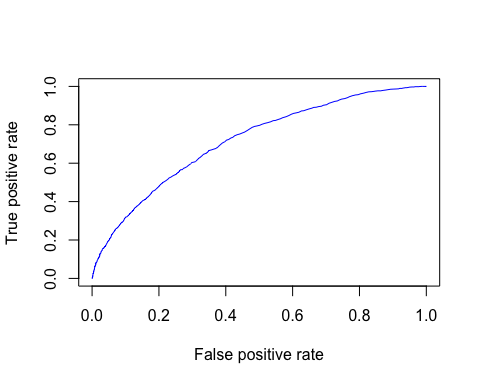
GC

## [1] 0.4326072

KS

## [1] 0.3175701

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] 0.184

#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 0.358 1

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

## [1] 387 22

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

## [1] 2411 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 0.114 0

tst\_with\_prob\_0

## Loan.Amount Loan.Term Employment.Length Home.Ownership Annual.Income  
## 3 12500 36 Months 7 MORTGAGE 47000  
## Loan.Purpose Address.State Debt.To.Income.Ratio  
## 3 Home Improvement AZ 18.33  
## No..Delinquencies.In.Last.2.Years Earliest.Credit.Line.Opened  
## 3 Four 33383.18  
## FICO.Credit.Score No..Inquiries.In.Last.6.Months  
## 3 710 2  
## Months.Since.Last.Delinquency No..Adverse.Public.Records  
## 3 6 Four  
## Total.Credit.Balance Use.Of.Credit.Line Total.Number.Of.Credit.Lines  
## 3 6991 25.9 29  
## Loan.Application.Description No..Of.Public.Record.Bankruptcies Class  
## 3 333 None 0  
## predprob predclass  
## 3 0.262 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  
## 6965 6000 36 Months 6 OWN 48000  
## 6826 7000 36 Months 10 OWN 60000  
## Loan.Purpose Address.State Debt.To.Income.Ratio  
## 6073 0 AZ 11.68  
## 7220 0 AZ 8.88  
## 6876 0 AZ 15.21  
## 6965 0 AZ 25.85  
## 6826 0 AZ 20.24  
## No..Delinquencies.In.Last.2.Years Earliest.Credit.Line.Opened  
## 6073 Eight 35520.40  
## 7220 Eight 37358.15  
## 6876 Eight 36390.35  
## 6965 Eight 37495.35  
## 6826 Eight 37606.17  
## FICO.Credit.Score No..Inquiries.In.Last.6.Months  
## 6073 755 0  
## 7220 725 0  
## 6876 730 0  
## 6965 725 0  
## 6826 695 1  
## Months.Since.Last.Delinquency No..Adverse.Public.Records  
## 6073 0 Four  
## 7220 0 Four  
## 6876 0 Four  
## 6965 0 Four  
## 6826 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  
## 6965 4286 27.8 16  
## 6826 6057 35.8 21  
## Loan.Application.Description No..Of.Public.Record.Bankruptcies Class  
## 6073 0 None 1  
## 7220 0 None 1  
## 6876 0 None 1  
## 6965 0 None 1  
## 6826 58 None 1  
## predprob predclass  
## 6073 0.114 0  
## 7220 0.140 0  
## 6876 0.102 0  
## 6965 0.094 0  
## 6826 0.182 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  
## 3 12500 36 Months 7 MORTGAGE 47000  
## 5877 10000 36 Months 4 MORTGAGE 50000  
## 1640 6000 60 Months 4 MORTGAGE 42000  
## 812 3000 36 Months 2 MORTGAGE 53000  
## 366 8000 36 Months 0 MORTGAGE 28776  
## Loan.Purpose Address.State Debt.To.Income.Ratio  
## 3 Home Improvement AZ 18.33  
## 5877 Home Improvement AZ 17.48  
## 1640 Home Improvement AZ 20.89  
## 812 Home Improvement AZ 19.83  
## 366 Home Improvement AZ 9.47  
## No..Delinquencies.In.Last.2.Years Earliest.Credit.Line.Opened  
## 3 Four 33383.18  
## 5877 Four 33127.52  
## 1640 Four 34492.40  
## 812 Four 32169.54  
## 366 Four 30000.00  
## FICO.Credit.Score No..Inquiries.In.Last.6.Months  
## 3 710 2  
## 5877 700 2  
## 1640 710 2  
## 812 690 0  
## 366 730 3  
## Months.Since.Last.Delinquency No..Adverse.Public.Records  
## 3 6 Four  
## 5877 19 Four  
## 1640 7 Four  
## 812 5 Four  
## 366 19 Four  
## Total.Credit.Balance Use.Of.Credit.Line Total.Number.Of.Credit.Lines  
## 3 6991 25.9 29  
## 5877 2517 8.7 18  
## 1640 3659 37.7 43  
## 812 20168 55.3 33  
## 366 3270 16.6 14  
## Loan.Application.Description No..Of.Public.Record.Bankruptcies Class  
## 3 333 None 0  
## 5877 672 None 0  
## 1640 196 None 0  
## 812 597 None 0  
## 366 126 None 0  
## predprob predclass  
## 3 0.262 1  
## 5877 0.250 1  
## 1640 0.316 1  
## 812 0.204 1  
## 366 0.218 1

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

##   
## 0 1  
## 0 3661 387  
## 1 2411 957

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

## [1] 0.7120536

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

## [1] 0.6227077