# Now we can model the data  
  
  
#loadrequiredlibrariesforpreprocessing(TRUE)  
setwd('/Users/dsing001/myr/R\_generic\_code/myrpreprocess/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(myrpreprocess)  
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

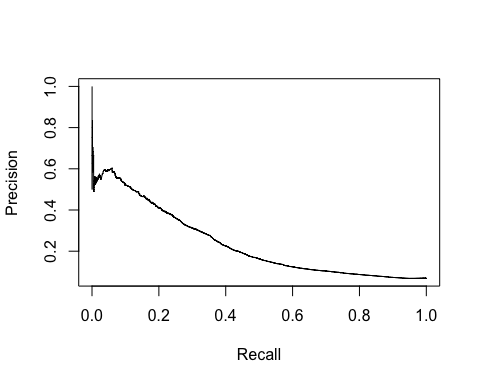
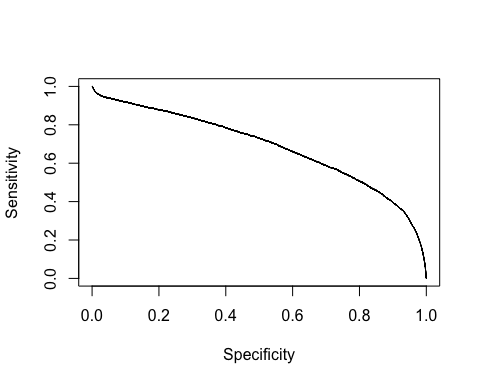
## Loading required package: RANN

clscol='SeriousDlqin2yrs'  
  
# 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')  
  
  
# Now remove the test and valid data  
  
frml1 <- 'SeriousDlqin2yrs ~ .'  
colnames(train\_data)

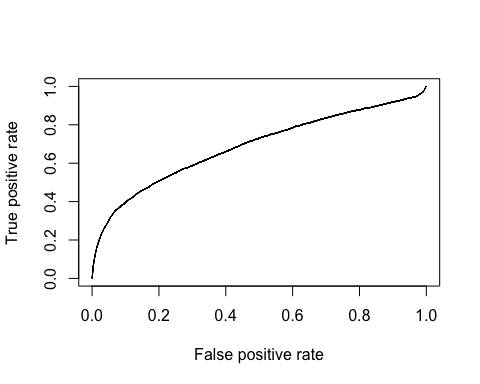
## [1] "X"   
## [2] "RevolvingUtilizationOfUnsecuredLines"  
## [3] "age"   
## [4] "NumberOfTime30.59DaysPastDueNotWorse"  
## [5] "DebtRatio"   
## [6] "MonthlyIncome"   
## [7] "NumberOfOpenCreditLinesAndLoans"   
## [8] "NumberOfTimes90DaysLate"   
## [9] "NumberRealEstateLoansOrLines"   
## [10] "NumberOfTime60.89DaysPastDueNotWorse"  
## [11] "NumberOfDependents"   
## [12] "SeriousDlqin2yrs"

# Let us model a normal logistic regression. Check the performance on valid\_dt and if needed calibrate parameters.   
# Once model is ready try finally on test\_data. Also, do not change anything after that even that is bad. If test performance is bad. Regenerate the train, test and valid and try samething over and again.  
train\_and\_predict\_log\_reg\_and\_ret\_auc(frml1,train\_data,valid\_data,predict\_type='response')

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## [1] 0.6872396



## $auc  
## [1] 0.6872396

```