Appendix 3 - Customer’s decision to open a term deposit account

## Part 2 - Naive Bayes

### Step 1 : Collecting Data

Same as of Part 1

### Step 2 : Exploring, preprocessing and cleaning the data

Primary setup

knitr::opts\_knit$set(root.dir = '/Users/sobil/Documents/MSC/Sem 1/Data Mining & Machine Learning/Project/Bank/')  
remove(list = ls())  
set.seed(1)

loading all the libraries required

library(ggplot2)  
library(ggthemes)  
library(caret)

## Loading required package: lattice

library(e1071)

Same as part 1

source("bank\_import\_primaryExplore.R")

## 'data.frame': 41188 obs. of 21 variables:  
## $ age : int 56 57 37 40 56 45 59 41 24 25 ...  
## $ job : Factor w/ 12 levels "admin.","blue-collar",..: 4 8 8 1 8 8 1 2 10 8 ...  
## $ marital : Factor w/ 4 levels "divorced","married",..: 2 2 2 2 2 2 2 2 3 3 ...  
## $ education : Factor w/ 8 levels "basic.4y","basic.6y",..: 1 4 4 2 4 3 6 8 6 4 ...  
## $ default : Factor w/ 3 levels "no","unknown",..: 1 2 1 1 1 2 1 2 1 1 ...  
## $ housing : Factor w/ 3 levels "no","unknown",..: 1 1 3 1 1 1 1 1 3 3 ...  
## $ loan : Factor w/ 3 levels "no","unknown",..: 1 1 1 1 3 1 1 1 1 1 ...  
## $ contact : Factor w/ 2 levels "cellular","telephone": 2 2 2 2 2 2 2 2 2 2 ...  
## $ month : Factor w/ 10 levels "apr","aug","dec",..: 7 7 7 7 7 7 7 7 7 7 ...  
## $ day\_of\_week : Factor w/ 5 levels "fri","mon","thu",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ duration : int 261 149 226 151 307 198 139 217 380 50 ...  
## $ campaign : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ pdays : int 999 999 999 999 999 999 999 999 999 999 ...  
## $ previous : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ poutcome : Factor w/ 3 levels "failure","nonexistent",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ emp.var.rate : num 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 ...  
## $ cons.price.idx: num 94 94 94 94 94 ...  
## $ cons.conf.idx : num -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 ...  
## $ euribor3m : num 4.86 4.86 4.86 4.86 4.86 ...  
## $ nr.employed : num 5191 5191 5191 5191 5191 ...  
## $ y : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...  
##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 41188   
##   
##   
## | no | yes |   
## |-----------|-----------|  
## | 36548 | 4640 |   
## | 0.887 | 0.113 |   
## |-----------|-----------|  
##   
##   
##   
##

### Step 3 - Data transformation & preparation

#### 1) creating a new df to convert the numeric columns (continous) to factors by binning

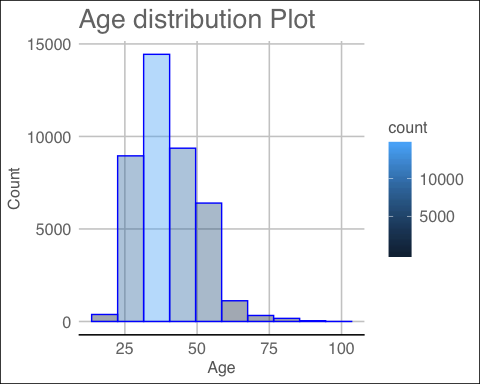
bank\_n <- bank

checking histograms and creating factors

theme\_set(theme\_gdocs())

age

pl.age <- ggplot(bank\_n, aes(x=age)) + geom\_histogram(bins = 10, color = 'blue', aes(fill=..count..), alpha = 0.4) + xlab('Age') + ylab('Count') + ggtitle('Age distribution Plot')  
pl.age

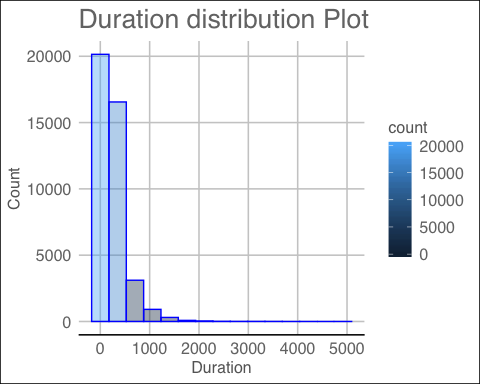


bank\_n$age <- cut(bank\_n$age,  
 breaks = c(1, 10, 20, 30, 40, 50, 60, 70, 80, max(bank\_n$age)),  
 labels = c("1-10", "10-20", "20-30", "30-40", "40-50", "50-60", "60-70", "70-80", "80+"))  
table(bank\_n$age)

##   
## 1-10 10-20 20-30 30-40 40-50 50-60 60-70 70-80 80+   
## 0 140 7243 16385 10240 6270 488 303 119

duration

pl.duration <- ggplot(bank\_n, aes(x=duration)) + geom\_histogram(bins = 15, color = 'blue', aes(fill=..count..), alpha = 0.4) + xlab('Duration') + ylab('Count') + ggtitle('Duration distribution Plot') +stat\_bin(breaks=c(-0.004, seq(0.001,1.0, by=0.005)))  
pl.duration



summary(bank\_n$duration)

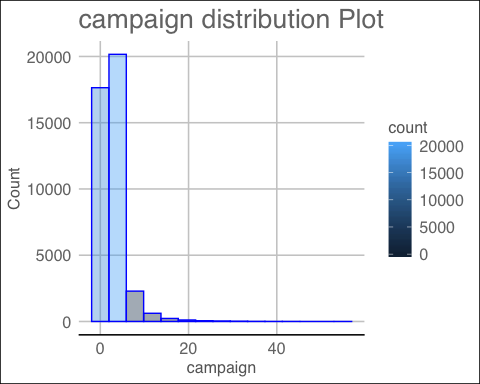
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 102.0 180.0 258.3 319.0 4918.0

bank\_n$duration <- cut(bank\_n$duration,  
 breaks = c(0, 10, 30, 70, 110, 150, 190, 250, 350, 600, 900, 1200, max(bank\_n$duration)),  
 labels = c("1-10","30-70", "30-70", "70-110", "110-150", "150-190", "190-250", "250-350", "350-600", "600-900", "900-1200", "1200+"))  
table(bank\_n$age)

##   
## 1-10 10-20 20-30 30-40 40-50 50-60 60-70 70-80 80+   
## 0 140 7243 16385 10240 6270 488 303 119

campaign

pl.campaign<- ggplot(bank\_n, aes(x=campaign)) + geom\_histogram(bins = 15, color = 'blue', aes(fill=..count..), alpha = 0.4) + xlab('campaign') + ylab('Count') + ggtitle('campaign distribution Plot')  
pl.campaign



summary(bank\_n$campaign)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 1.000 2.000 2.568 3.000 56.000

bank\_n$campaign <- cut(bank\_n$campaign,  
 breaks = c(1, 2, 3, 5, 10, 15, 20, max(bank\_n$campaign)),  
 labels = c("1-2","2-3", "3-5", "5-10", "10-15", "15-20", "20+"))  
table(bank\_n$campaign)

##   
## 1-2 2-3 3-5 5-10 10-15 15-20 20+   
## 10570 5341 4250 2516 514 198 157

pdays

summary(bank\_n$pdays)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 999.0 999.0 962.5 999.0 999.0

table(bank\_n$pdays)

##   
## 0 1 2 3 4 5 6 7 8 9 10 11 12   
## 15 26 61 439 118 46 412 60 18 64 52 28 58   
## 13 14 15 16 17 18 19 20 21 22 25 26 27   
## 36 20 24 11 8 7 3 1 2 3 1 1 1   
## 999   
## 39673

bank\_n$pdays <- cut(bank\_n$pdays,  
 breaks = c(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 28, max(bank\_n$pdays)),  
 labels = c("0","1", "2", "3", "4", "5", "6", "7", "8", "9", "10-15", "15-20", "20-28", "999"))  
summary(bank\_n$pdays)

## 0 1 2 3 4 5 6 7 8 9 10-15 15-20 20-28   
## 26 61 439 118 46 412 60 18 64 52 166 30 8   
## 999 NA's   
## 39673 15

previous

summary(bank\_n$previous)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 0.000 0.000 0.173 0.000 7.000

table(bank\_n$previous)

##   
## 0 1 2 3 4 5 6 7   
## 35563 4561 754 216 70 18 5 1

bank\_n$previous <- as.factor(bank\_n$previous)  
table(bank\_n$previous)

##   
## 0 1 2 3 4 5 6 7   
## 35563 4561 754 216 70 18 5 1

emp.var.rate

summary(bank\_n$emp.var.rate)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -3.40000 -1.80000 1.10000 0.08189 1.40000 1.40000

table(bank\_n$emp.var.rate)

##   
## -3.4 -3 -2.9 -1.8 -1.7 -1.1 -0.2 -0.1 1.1 1.4   
## 1071 172 1663 9184 773 635 10 3683 7763 16234

bank\_n$emp.var.rate <- as.factor(bank\_n$emp.var.rate)  
summary(bank\_n$emp.var.rate)

## -3.4 -3 -2.9 -1.8 -1.7 -1.1 -0.2 -0.1 1.1 1.4   
## 1071 172 1663 9184 773 635 10 3683 7763 16234

cons.price.idx

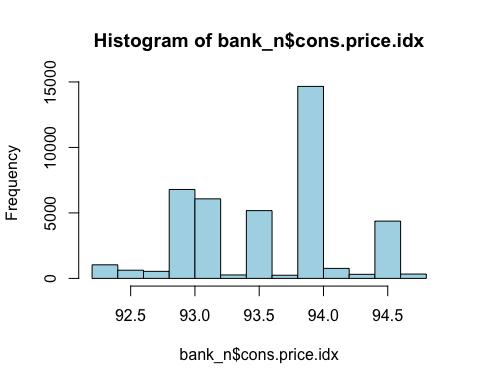
summary(bank\_n$cons.price.idx)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 92.20 93.08 93.75 93.58 93.99 94.77

table(bank\_n$cons.price.idx)

##   
## 92.201 92.379 92.431 92.469 92.649 92.713 92.756 92.843 92.893 92.963 93.075   
## 770 267 447 178 357 172 10 282 5794 715 2458   
## 93.2 93.369 93.444 93.749 93.798 93.876 93.918 93.994 94.027 94.055 94.199   
## 3616 264 5175 174 67 212 6685 7763 233 229 303   
## 94.215 94.465 94.601 94.767   
## 311 4374 204 128

hist(bank\_n$cons.price.idx, col = "lightblue")



bank\_n$cons.price.idx <- cut(bank\_n$cons.price.idx,  
 breaks = c(92,93, 93.5, 94, max(bank\_n$cons.price.idx)),  
 labels = c("92-93", "93-93.5", "93.5-94", "94+"))  
summary(bank\_n$cons.price.idx)

## 92-93 93-93.5 93.5-94 94+   
## 8992 11513 14901 5782

cons.conf.idx

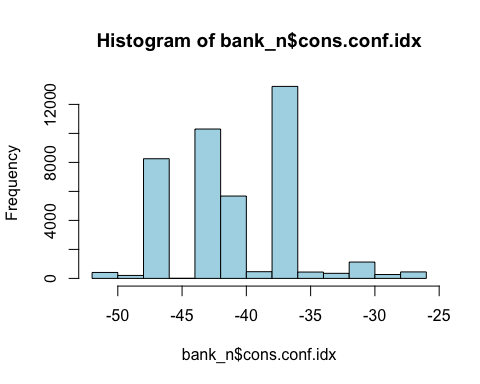
summary(bank\_n$cons.conf.idx)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -50.8 -42.7 -41.8 -40.5 -36.4 -26.9

table(bank\_n$cons.conf.idx)

##   
## -50.8 -50 -49.5 -47.1 -46.2 -45.9 -42.7 -42 -41.8 -40.8 -40.4 -40.3 -40   
## 128 282 204 2458 5794 10 6685 3616 4374 715 67 311 212   
## -39.8 -38.3 -37.5 -36.4 -36.1 -34.8 -34.6 -33.6 -33 -31.4 -30.1 -29.8 -26.9   
## 229 233 303 7763 5175 264 174 178 172 770 357 267 447

hist(bank\_n$cons.conf.idx, col = "lightblue")



bank\_n$cons.conf.idx <- cut(bank\_n$cons.conf.idx,  
 breaks = c(-51, -45, -40, -35, max(bank\_n$cons.conf.idx)),  
 labels = c("(-50)-(-45)", "(-45)-(-40)", "(-40)-(-35)", "(-35)+"))  
summary(bank\_n$cons.conf.idx)

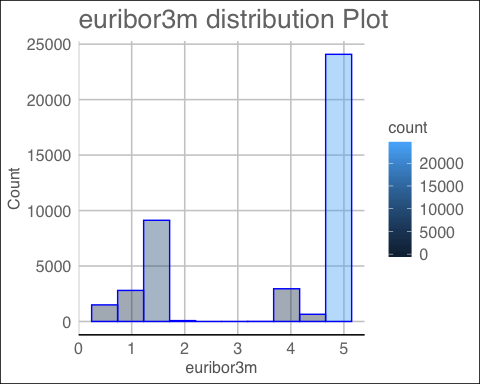
## (-50)-(-45) (-45)-(-40) (-40)-(-35) (-35)+   
## 8876 15980 13703 2629

euribor3m

summary(bank\_n$euribor3m)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.634 1.344 4.857 3.621 4.961 5.045

pl.euribor3m <- ggplot(bank\_n, aes(x=euribor3m)) + geom\_histogram(bins = 10, color = 'blue', aes(fill=..count..), alpha = 0.4) + xlab('euribor3m') + ylab('Count') + ggtitle('euribor3m distribution Plot')  
pl.euribor3m



bank\_n$euribor3m <- cut(bank\_n$euribor3m,  
 breaks = c(0,1, 2, 4, max(bank\_n$euribor3m)),  
 labels = c("0-1", "1-2", "2-4", "4+"))  
summary(bank\_n$euribor3m)

## 0-1 1-2 2-4 4+   
## 3908 9590 14 27676

nr.employed

summary(bank\_n$nr.employed)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 4964 5099 5191 5167 5228 5228

table(bank\_n$nr.employed)

##   
## 4963.6 4991.6 5008.7 5017.5 5023.5 5076.2 5099.1 5176.3 5191 5195.8 5228.1   
## 635 773 650 1071 172 1663 8534 10 7763 3683 16234

bank\_n$nr.employed <- as.factor(bank\_n$nr.employed)  
summary(bank\_n$nr.employed)

## 4963.6 4991.6 5008.7 5017.5 5023.5 5076.2 5099.1 5176.3 5191 5195.8 5228.1   
## 635 773 650 1071 172 1663 8534 10 7763 3683 16234

summary of df

summary(bank\_n)

## age job marital   
## 30-40 :16385 admin. :10422 divorced: 4612   
## 40-50 :10240 blue-collar: 9254 married :24928   
## 20-30 : 7243 technician : 6743 single :11568   
## 50-60 : 6270 services : 3969 unknown : 80   
## 60-70 : 488 management : 2924   
## 70-80 : 303 retired : 1720   
## (Other): 259 (Other) : 6156   
## education default housing loan   
## university.degree :12168 no :32588 no :18622 no :33950   
## high.school : 9515 unknown: 8597 unknown: 990 unknown: 990   
## basic.9y : 6045 yes : 3 yes :21576 yes : 6248   
## professional.course: 5243   
## basic.4y : 4176   
## basic.6y : 2292   
## (Other) : 1749   
## contact month day\_of\_week duration campaign   
## cellular :26144 may :13769 fri:7827 70-110 : 5952 1-2 :10570   
## telephone:15044 jul : 7174 mon:8514 110-150: 5521 2-3 : 5341   
## aug : 6178 thu:8623 350-600: 5452 3-5 : 4250   
## jun : 5318 tue:8090 250-350: 5325 5-10 : 2516   
## nov : 4101 wed:8134 30-70 : 5198 10-15 : 514   
## apr : 2632 (Other):13736 (Other): 355   
## (Other): 2016 NA's : 4 NA's :17642   
## pdays previous poutcome emp.var.rate   
## 999 :39673 0 :35563 failure : 4252 1.4 :16234   
## 2 : 439 1 : 4561 nonexistent:35563 -1.8 : 9184   
## 5 : 412 2 : 754 success : 1373 1.1 : 7763   
## 10-15 : 166 3 : 216 -0.1 : 3683   
## 3 : 118 4 : 70 -2.9 : 1663   
## (Other): 365 5 : 18 -3.4 : 1071   
## NA's : 15 (Other): 6 (Other): 1590   
## cons.price.idx cons.conf.idx euribor3m nr.employed y   
## 92-93 : 8992 (-50)-(-45): 8876 0-1: 3908 5228.1 :16234 Yes: 4640   
## 93-93.5:11513 (-45)-(-40):15980 1-2: 9590 5099.1 : 8534 No :36548   
## 93.5-94:14901 (-40)-(-35):13703 2-4: 14 5191 : 7763   
## 94+ : 5782 (-35)+ : 2629 4+ :27676 5195.8 : 3683   
## 5076.2 : 1663   
## 5017.5 : 1071   
## (Other): 2240

#### 2) removing NAs after transformation

bank\_n <- na.omit(bank\_n)  
summary(bank\_n)

## age job marital education   
## 30-40 :9376 admin. :6012 divorced: 2692 university.degree :6896   
## 40-50 :5885 blue-collar:5323 married :14252 high.school :5436   
## 20-30 :4122 technician :3857 single : 6551 basic.9y :3488   
## 50-60 :3649 services :2315 unknown : 44 professional.course:3046   
## 60-70 : 243 management :1642 basic.4y :2333   
## 70-80 : 131 retired : 932 basic.6y :1339   
## (Other): 133 (Other) :3458 (Other) :1001   
## default housing loan contact   
## no :18422 no :10754 no :19306 cellular :14380   
## unknown: 5116 unknown: 575 unknown: 575 telephone: 9159   
## yes : 1 yes :12210 yes : 3658   
##   
##   
##   
##   
## month day\_of\_week duration campaign pdays   
## may :8001 fri:4811 70-110 :3451 1-2 :10568 999 :22824   
## jul :4500 mon:5302 30-70 :3218 2-3 : 5338 5 : 209   
## aug :3579 thu:4593 110-150:3066 3-5 : 4249 2 : 205   
## jun :3275 tue:4379 350-600:3057 5-10 : 2515 10-15 : 80   
## nov :1981 wed:4454 250-350:2933 10-15: 514 3 : 61   
## apr :1306 190-250:2870 15-20: 198 8 : 29   
## (Other): 897 (Other):4944 20+ : 157 (Other): 131   
## previous poutcome emp.var.rate cons.price.idx  
## 0 :20768 failure : 2125 1.4 :10263 92-93 :4730   
## 1 : 2258 nonexistent:20768 -1.8 : 4978 93-93.5:6228   
## 2 : 362 success : 646 1.1 : 4622 93.5-94:9056   
## 3 : 107 -0.1 : 1743 94+ :3525   
## 4 : 34 -2.9 : 740   
## 5 : 7 -3.4 : 443   
## (Other): 3 (Other): 750   
## cons.conf.idx euribor3m nr.employed y   
## (-50)-(-45):4819 0-1: 1779 5228.1 :10263 Yes: 2338   
## (-45)-(-40):9433 1-2: 5132 5099.1 : 4658 No :21201   
## (-40)-(-35):8096 2-4: 0 5191 : 4622   
## (-35)+ :1191 4+ :16628 5195.8 : 1743   
## 5076.2 : 740   
## 5017.5 : 443   
## (Other): 1070

#### 3) creating training and testing dataset from exisitng sample

set.seed(1)  
indx <- caret::createDataPartition(bank\_n$y, p = 0.8, list = FALSE)

creating lables for test and training data sets

bank\_train\_labels <- bank\_n[indx,21]  
bank\_test\_labels <- bank\_n[- indx,21]

removing dependent variable

bank\_n <- bank\_n[,-21]  
summary(bank\_n)

## age job marital education   
## 30-40 :9376 admin. :6012 divorced: 2692 university.degree :6896   
## 40-50 :5885 blue-collar:5323 married :14252 high.school :5436   
## 20-30 :4122 technician :3857 single : 6551 basic.9y :3488   
## 50-60 :3649 services :2315 unknown : 44 professional.course:3046   
## 60-70 : 243 management :1642 basic.4y :2333   
## 70-80 : 131 retired : 932 basic.6y :1339   
## (Other): 133 (Other) :3458 (Other) :1001   
## default housing loan contact   
## no :18422 no :10754 no :19306 cellular :14380   
## unknown: 5116 unknown: 575 unknown: 575 telephone: 9159   
## yes : 1 yes :12210 yes : 3658   
##   
##   
##   
##   
## month day\_of\_week duration campaign pdays   
## may :8001 fri:4811 70-110 :3451 1-2 :10568 999 :22824   
## jul :4500 mon:5302 30-70 :3218 2-3 : 5338 5 : 209   
## aug :3579 thu:4593 110-150:3066 3-5 : 4249 2 : 205   
## jun :3275 tue:4379 350-600:3057 5-10 : 2515 10-15 : 80   
## nov :1981 wed:4454 250-350:2933 10-15: 514 3 : 61   
## apr :1306 190-250:2870 15-20: 198 8 : 29   
## (Other): 897 (Other):4944 20+ : 157 (Other): 131   
## previous poutcome emp.var.rate cons.price.idx  
## 0 :20768 failure : 2125 1.4 :10263 92-93 :4730   
## 1 : 2258 nonexistent:20768 -1.8 : 4978 93-93.5:6228   
## 2 : 362 success : 646 1.1 : 4622 93.5-94:9056   
## 3 : 107 -0.1 : 1743 94+ :3525   
## 4 : 34 -2.9 : 740   
## 5 : 7 -3.4 : 443   
## (Other): 3 (Other): 750   
## cons.conf.idx euribor3m nr.employed   
## (-50)-(-45):4819 0-1: 1779 5228.1 :10263   
## (-45)-(-40):9433 1-2: 5132 5099.1 : 4658   
## (-40)-(-35):8096 2-4: 0 5191 : 4622   
## (-35)+ :1191 4+ :16628 5195.8 : 1743   
## 5076.2 : 740   
## 5017.5 : 443   
## (Other): 1070

creating training and testing data

bank\_train <- bank\_n[indx,]  
bank\_test <- bank\_n[- indx,]

### Step 4 : Train the model

bank\_classifier <- naiveBayes(bank\_train, bank\_train\_labels)

### Step 5 : Evaluate the model

bank\_test\_pred <- predict(bank\_classifier, bank\_test)  
caret::confusionMatrix(bank\_test\_pred, bank\_test\_labels)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Yes No  
## Yes 277 371  
## No 190 3869  
##   
## Accuracy : 0.8808   
## 95% CI : (0.8712, 0.8899)  
## No Information Rate : 0.9008   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : 0.4313   
##   
## Mcnemar's Test P-Value : 2.97e-14   
##   
## Sensitivity : 0.59315   
## Specificity : 0.91250   
## Pos Pred Value : 0.42747   
## Neg Pred Value : 0.95319   
## Prevalence : 0.09921   
## Detection Rate : 0.05885   
## Detection Prevalence : 0.13767   
## Balanced Accuracy : 0.75282   
##   
## 'Positive' Class : Yes   
##

# Step 6 : Improving the model

bank\_classifier2 <- naiveBayes(bank\_train, bank\_train\_labels, laplace = 1)  
bank\_test\_pred2 <- predict(bank\_classifier2, bank\_test)  
caret::confusionMatrix(bank\_test\_pred2, bank\_test\_labels)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Yes No  
## Yes 277 367  
## No 190 3873  
##   
## Accuracy : 0.8817   
## 95% CI : (0.8721, 0.8908)  
## No Information Rate : 0.9008   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : 0.4335   
##   
## Mcnemar's Test P-Value : 8.827e-14   
##   
## Sensitivity : 0.59315   
## Specificity : 0.91344   
## Pos Pred Value : 0.43012   
## Neg Pred Value : 0.95324   
## Prevalence : 0.09921   
## Detection Rate : 0.05885   
## Detection Prevalence : 0.13682   
## Balanced Accuracy : 0.75330   
##   
## 'Positive' Class : Yes   
##

confusionMatrix(bank\_test\_pred2, bank\_test\_labels, mode="prec\_recall")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Yes No  
## Yes 277 367  
## No 190 3873  
##   
## Accuracy : 0.8817   
## 95% CI : (0.8721, 0.8908)  
## No Information Rate : 0.9008   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : 0.4335   
##   
## Mcnemar's Test P-Value : 8.827e-14   
##   
## Precision : 0.43012   
## Recall : 0.59315   
## F1 : 0.49865   
## Prevalence : 0.09921   
## Detection Rate : 0.05885   
## Detection Prevalence : 0.13682   
## Balanced Accuracy : 0.75330   
##   
## 'Positive' Class : Yes   
##