Analysis of a Portugese bank Marketing Dataset

install.packages("ISwR") install.packages("VIM") install.packages("mice")
install.packages("caret") install.packages("ROCR") install.packages("randomForest")

original dataset in bankA

working dataset in bankB (unknown <- NA; age <- age_group)

age_group added in bankC (22nd Column: age_group)

```
#########################                              #########################
          #####           Bank marketing DATASET                #####
#########################                              #########################


library(ISwR)

# Load Data in Data Frame
bankA <- as.data.frame(read.csv("C:/Users/arup.roy/Documents/bank-marketing-
master/bankAdd.csv", sep= ";",header = T))

# Display the variables and first 10 records
str(bankA)

## '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 ...

head(bankA,10)

##    age         job marital           education default housing loan
## 1   56   housemaid married            basic.4y      no      no   no
## 2   57    services married         high.school unknown      no   no
## 3   37    services married         high.school      no     yes   no
## 4   40      admin. married            basic.6y      no      no   no
## 5   56    services married         high.school      no      no  yes
## 6   45    services married            basic.9y unknown      no   no
## 7   59      admin. married professional.course      no      no   no
## 8   41 blue-collar married             unknown unknown      no   no
## 9   24  technician  single professional.course      no     yes   no
## 10  25    services  single         high.school      no     yes   no
##      contact month day_of_week duration campaign pdays previous
## 1  telephone   may         mon      261        1   999        0
## 2  telephone   may         mon      149        1   999        0
## 3  telephone   may         mon      226        1   999        0
## 4  telephone   may         mon      151        1   999        0
## 5  telephone   may         mon      307        1   999        0
## 6  telephone   may         mon      198        1   999        0
## 7  telephone   may         mon      139        1   999        0
## 8  telephone   may         mon      217        1   999        0
## 9  telephone   may         mon      380        1   999        0
## 10 telephone   may         mon       50        1   999        0
##       poutcome emp.var.rate cons.price.idx cons.conf.idx euribor3m
## 1  nonexistent          1.1         93.994         -36.4     4.857
## 2  nonexistent          1.1         93.994         -36.4     4.857
## 3  nonexistent          1.1         93.994         -36.4     4.857
## 4  nonexistent          1.1         93.994         -36.4     4.857
## 5  nonexistent          1.1         93.994         -36.4     4.857
## 6  nonexistent          1.1         93.994         -36.4     4.857
## 7  nonexistent          1.1         93.994         -36.4     4.857
## 8  nonexistent          1.1         93.994         -36.4     4.857
## 9  nonexistent          1.1         93.994         -36.4     4.857
## 10 nonexistent          1.1         93.994         -36.4     4.857
```

```
##    nr.employed  y
## 1         5191 no
## 2         5191 no
## 3         5191 no
## 4         5191 no
## 5         5191 no
## 6         5191 no
## 7         5191 no
## 8         5191 no
## 9         5191 no
## 10        5191 no
```

```
# Replace all 'unknown' values with NA
bankB<-bankA
bankB[bankB=="unknown"]<-NA

summary(bankB$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   32.00   38.00   40.02   47.00   98.00
```

```
#Min 17 #Max 98 #Mean 40 #Median 38


# Dividing the People into Different Age Groups
for(i in 1 : nrow(bankB)){
  if (bankB$age[i] <= 19){bankB$age_group[i] = 'Teenagers'}
  else if (bankB$age[i] >= 20 & bankB$age[i] <= 29){bankB$age_group[i] =
'Twenties'}
  else if (bankB$age[i] >= 30 & bankB$age[i] <= 39){bankB$age_group[i] =
'Thirties'}
  else if (bankB$age[i] >= 40 & bankB$age[i] <= 49){bankB$age_group[i] =
'Forties'}
  else if (bankB$age[i] >= 50 & bankB$age[i] <= 59){bankB$age_group[i] =
'Fifties'}
  else if (bankB$age[i] >= 60 & bankB$age[i] <= 69){bankB$age_group[i] =
'Sixties'}
  else if (bankB$age[i] >= 70 ){bankB$age_group[i] = 'Seniors'}
}

# saving the data before replacing age_group with age
bankC<-bankB

bankB$age<-bankB$age_group
bankB<-bankB[1:21]
bankB$age<-as.factor(bankB$age)


# Separating New Customers from the Old ones
```

```r
oldCust <- subset(bankB, bankB$poutcome != "nonexistent")
summary(oldCust)
```

```
##         age                job           marital
##  Fifties  : 793   admin.     :1519   divorced: 631
##  Forties  :1149   blue-collar:1005   married :3107
##  Seniors  : 202   technician : 829   single  :1869
##  Sixties  : 244   services   : 518   unknown :   0
##  Teenagers:  34   management : 426   NA's    :  18
##  Thirties :2250   (Other)    :1291
##  Twenties : 953   NA's       :  37
##                education       default        housing          loan
##  university.degree  :1775   no     :5049   no     :2366   no     :4634
##  high.school        :1413   unknown:   0   unknown:   0   unknown:   0
##  basic.9y           : 741   yes    :   1   yes    :3120   yes    : 852
##  professional.course: 686   NA's   : 575   NA's   : 139   NA's   : 139
##  basic.4y           : 480
##  (Other)            : 260
##  NA's               : 270
##       contact          month      day_of_week    duration
##  cellular :5222   may    :2009   fri:1124    Min.   :   1.0
##  telephone: 403   nov    :1004   mon:1150    1st Qu.: 115.0
##                   apr    : 758   thu:1181    Median : 199.0
##                   aug    : 459   tue:1096    Mean   : 265.9
##                   jun    : 315   wed:1074    3rd Qu.: 328.0
##                   oct    : 304               Max.   :3509.0
##                   (Other): 776
##     campaign          pdays          previous            poutcome
##  Min.   : 1.000   Min.   :  0.0   Min.   :1.000   failure    :4252
##  1st Qu.: 1.000   1st Qu.: 13.0   1st Qu.:1.000   nonexistent:   0
##  Median : 1.000   Median :999.0   Median :1.000   success    :1373
##  Mean   : 1.957   Mean   :731.6   Mean   :1.266
##  3rd Qu.: 2.000   3rd Qu.:999.0   3rd Qu.:1.000
##  Max.   :16.000   Max.   :999.0   Max.   :7.000
##
##   emp.var.rate    cons.price.idx  cons.conf.idx      euribor3m
##  Min.   :-3.400   Min.   :92.20   Min.   :-50.80   Min.   :0.634
##  1st Qu.:-1.800   1st Qu.:92.89   1st Qu.:-46.20   1st Qu.:0.878
##  Median :-1.800   Median :92.89   Median :-42.00   Median :1.266
##  Mean   :-1.784   Mean   :93.13   Mean   :-41.66   Mean   :1.491
##  3rd Qu.:-1.700   3rd Qu.:93.20   3rd Qu.:-38.30   3rd Qu.:1.365
##  Max.   :-0.100   Max.   :94.77   Max.   :-26.90   Max.   :4.968
##
##   nr.employed     y
##  Min.   :4964   no :4126
##  1st Qu.:5018   yes:1499
##  Median :5099
##  Mean   :5077
##  3rd Qu.:5099
```

```
##  Max.   :5196
##
```

```
#05625 Old Customers
```

```
newCust <- subset(bankB, bankB$poutcome == "nonexistent")
summary(newCust)
```

```
##         age                 job           marital
##  Fifties  : 6069   admin.     :8903   divorced: 3981
##  Forties  : 9377   blue-collar:8249   married :21821
##  Seniors  :  267   technician :5914   single  : 9699
##  Sixties  :  480   services   :3451   unknown :    0
##  Teenagers:   41   management :2498   NA's    :   62
##  Thirties :14688   (Other)    :6255
##  Twenties : 4641   NA's       : 293
##                education        default         housing
##  university.degree  :10393   no     :27539   no     :16256
##  high.school        : 8102   unknown:    0   unknown:    0
##  basic.9y           : 5304   yes    :    2   yes    :18456
##  professional.course: 4557   NA's   : 8022   NA's   :  851
##  basic.4y           : 3696
##  (Other)            : 2050
##  NA's               : 1461
##       loan            contact          month       day_of_week
##  no     :29316   cellular :20922   may    :11760   fri:6703
##  unknown:    0   telephone:14641   jul    : 6946   mon:7364
##  yes    : 5396                     aug    : 5719   thu:7442
##  NA's   :  851                     jun    : 5003   tue:6994
##                                    nov    : 3097   wed:7060
##                                    apr    : 1874
##                                    (Other): 1164
##     duration         campaign          pdays        previous
##  Min.   :   0.0   Min.   : 1.000   Min.   :999   Min.   :0
##  1st Qu.: 100.0   1st Qu.: 1.000   1st Qu.:999   1st Qu.:0
##  Median : 177.0   Median : 2.000   Median :999   Median :0
##  Mean   : 257.1   Mean   : 2.664   Mean   :999   Mean   :0
##  3rd Qu.: 318.0   3rd Qu.: 3.000   3rd Qu.:999   3rd Qu.:0
##  Max.   :4918.0   Max.   :56.000   Max.   :999   Max.   :0
##
##         poutcome      emp.var.rate     cons.price.idx  cons.conf.idx
##  failure    :    0   Min.   :-3.4000   Min.   :92.20   Min.   :-50.80
##  nonexistent:35563   1st Qu.:-0.1000   1st Qu.:93.20   1st Qu.:-42.70
##  success    :    0   Median : 1.1000   Median :93.92   Median :-41.80
##                      Mean   : 0.3771   Mean   :93.65   Mean   :-40.32
##                      3rd Qu.: 1.4000   3rd Qu.:93.99   3rd Qu.:-36.40
##                      Max.   : 1.4000   Max.   :94.77   Max.   :-26.90
##
##    euribor3m      nr.employed     y
##  Min.   :0.634   Min.   :4964   no :32422
```

```
##  1st Qu.:4.021   1st Qu.:5191   yes: 3141
##  Median :4.859   Median :5196
##  Mean   :3.958   Mean   :5181
##  3rd Qu.:4.962   3rd Qu.:5228
##  Max.   :5.045   Max.   :5228
##
```

```
#35563 New Customers
```

Old Customer DATASET

```
#########################                          #########################
           #####         Old Customer DATASET          #####
#########################                          #########################
```

```
# Missing value Frequencies
library(VIM)
```

```
## Loading required package: colorspace
```

```
## Loading required package: grid
```

```
## Loading required package: data.table
```

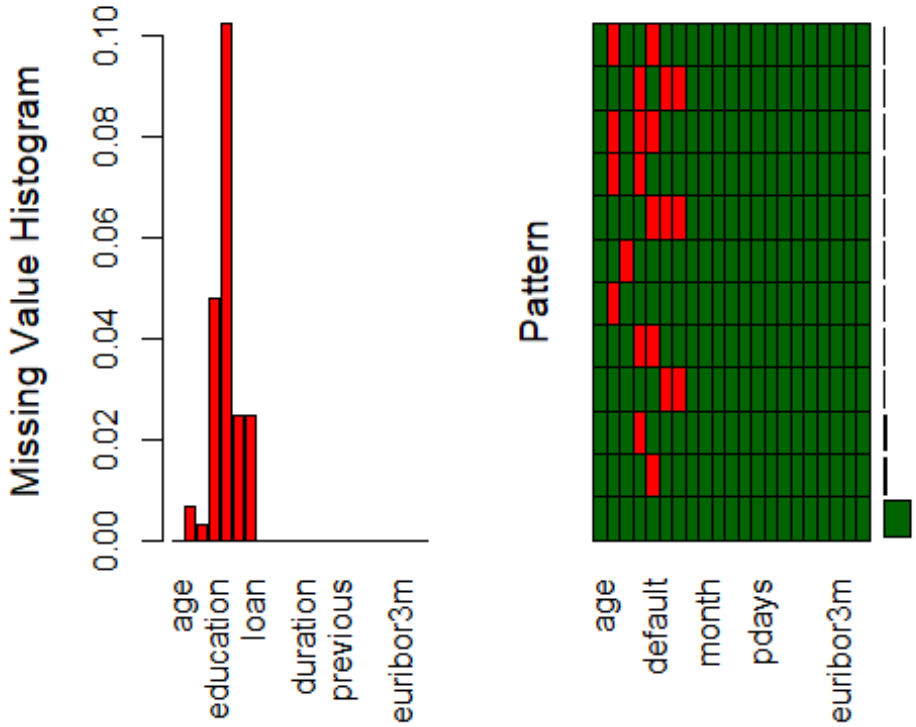
```
## VIM is ready to use.
##  Since version 4.0.0 the GUI is in its own package VIMGUI.
##
##           Please use the package to use the new (and old) GUI.
```

```
## Suggestions and bug-reports can be submitted at:
https://github.com/alexkowa/VIM/issues
```

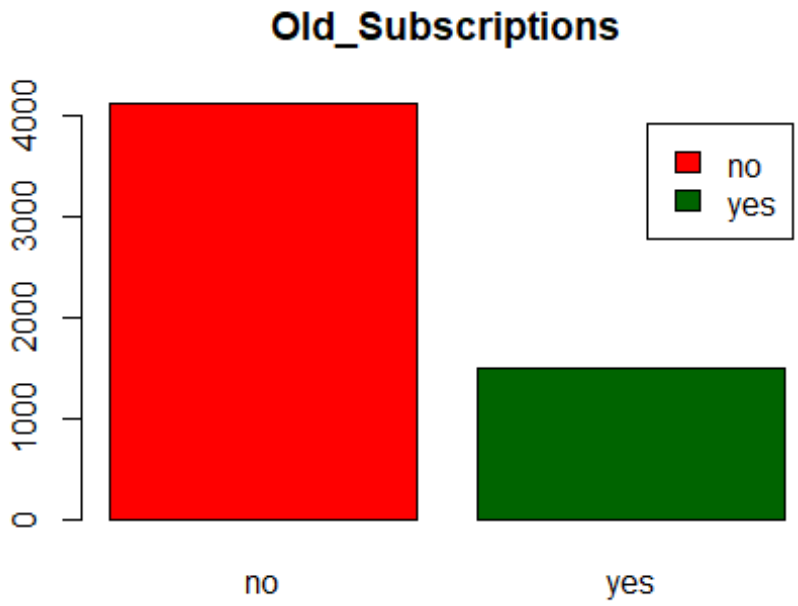
```
##
## Attaching package: 'VIM'
```

```
## The following object is masked from 'package:datasets':
##
##     sleep
```

```
aggrPlot <- aggr(oldCust, col=c('darkgreen','red'), ylab=c("Missing Value
Histogram","Pattern"))
```

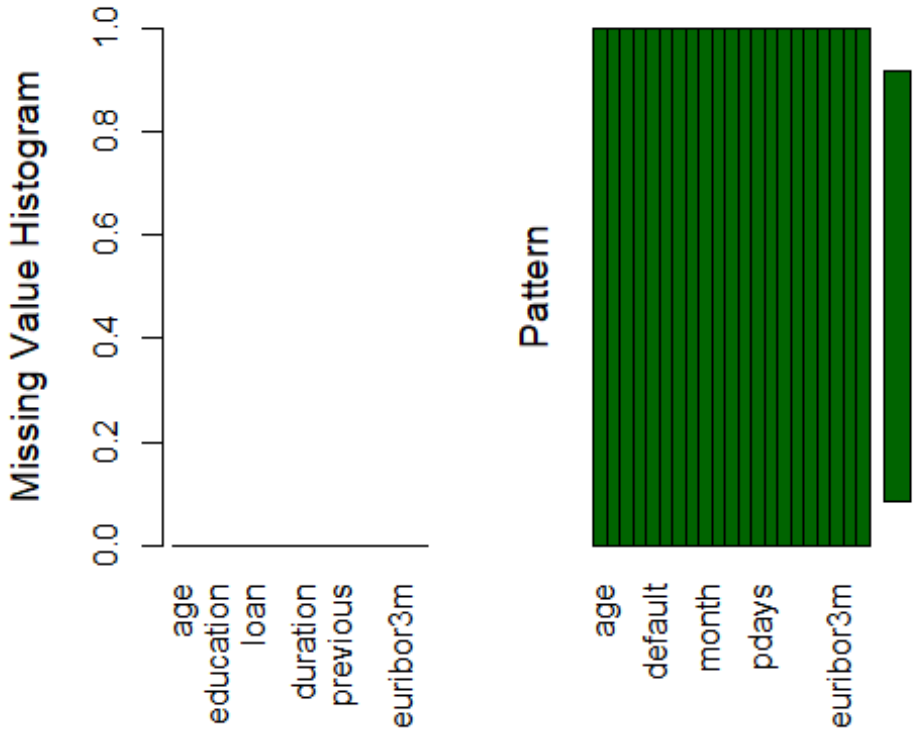



```
#default 0.1022 #education 0.0480 #housing 0.0247 #Loan 0.0247 #job 0.0065
#marital 0.0032

#Subscription Count
oldCount <- table(oldCust$y)
barplot(oldCount,col=c("red","darkgreen"),legend = rownames(oldCount), main =
"Old_Subscriptions")
```




```
#no 4126 #yes 1499


# Impute Missing Values and Check
library(mice)

## Loading required package: lattice

## Registered S3 methods overwritten by 'lme4':
##   method                           from
##   cooks.distance.influence.merMod car
##   influence.merMod                 car
##   dfbeta.influence.merMod          car
##   dfbetas.influence.merMod         car

##
## Attaching package: 'mice'

## The following objects are masked from 'package:base':
##
##     cbind, rbind

oldCust2 <- mice(oldCust)

##
##  iter imp variable
##   1   1  job  marital  education  default  housing  loan
##   1   2  job  marital  education  default  housing  loan
```

```
##   1   3  job  marital  education  default  housing  loan
##   1   4  job  marital  education  default  housing  loan
##   1   5  job  marital  education  default  housing  loan
##   2   1  job  marital  education  default  housing  loan
##   2   2  job  marital  education  default  housing  loan
##   2   3  job  marital  education  default  housing  loan
##   2   4  job  marital  education  default  housing  loan
##   2   5  job  marital  education  default  housing  loan
##   3   1  job  marital  education  default  housing  loan
##   3   2  job  marital  education  default  housing  loan
##   3   3  job  marital  education  default  housing  loan
##   3   4  job  marital  education  default  housing  loan
##   3   5  job  marital  education  default  housing  loan
##   4   1  job  marital  education  default  housing  loan
##   4   2  job  marital  education  default  housing  loan
##   4   3  job  marital  education  default  housing  loan
##   4   4  job  marital  education  default  housing  loan
##   4   5  job  marital  education  default  housing  loan
##   5   1  job  marital  education  default  housing  loan
##   5   2  job  marital  education  default  housing  loan
##   5   3  job  marital  education  default  housing  loan
##   5   4  job  marital  education  default  housing  loan
##   5   5  job  marital  education  default  housing  loan

## Warning: Number of logged events: 150

oldCust_com <- complete(oldCust2)

aggrPlot <- aggr(oldCust_com, col=c('darkgreen','red'), ylab=c("Missing Value
Histogram","Pattern"))
```




```
#none


#Split data into Train and Test subsets
library(caret)

## Loading required package: ggplot2

set.seed(101)

oldCust_com$y<-ifelse(oldCust_com$y =='no', 0,1)
oldCust_com$y<-as.factor(oldCust_com$y)

ids <- sample(seq(1, 2), size = nrow(oldCust_com), replace = TRUE, prob =
c(.7, .3))

oldCust_train <- oldCust_com[ids==1,]
oldCust_test  <- oldCust_com[ids==2,]


table(oldCust_train$y)

##
##    0    1
## 2886 1027
```

```r
#no 2886 #yes 1027
table(oldCust_test$y)
```

```
##
##    0    1
## 1240  472
```

```r
#no 1240 #yes 472


#########################   Logistic Model (oldCust)
#########################


oldCust_logit <- glm(y ~., family=binomial(link='logit'), data =
oldCust_train)
summary(oldCust_logit)
```

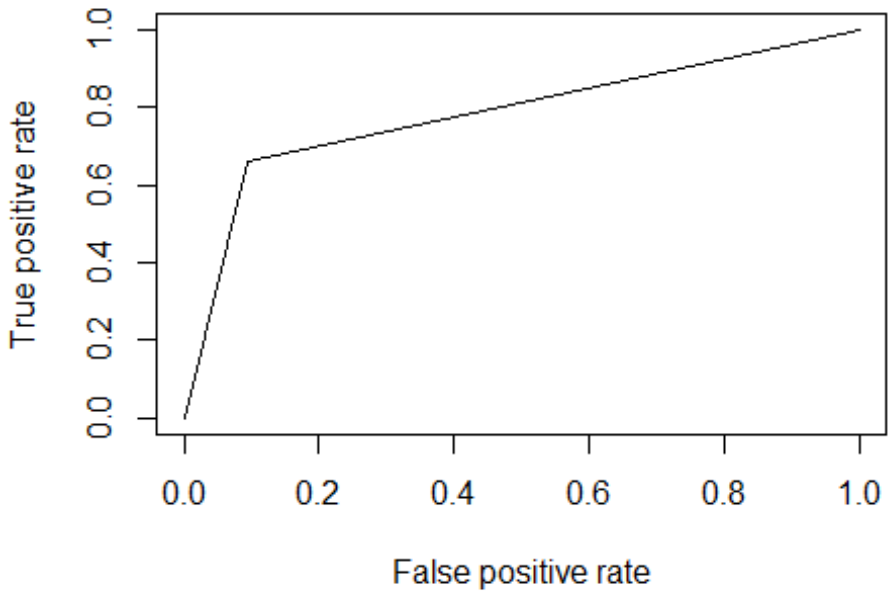
```
##
## Call:
## glm(formula = y ~ ., family = binomial(link = "logit"), data =
oldCust_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.9038  -0.4871  -0.2374   0.3311   3.1466
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -6.920e+02  2.059e+02  -3.361 0.000776 ***
## ageForties                   -2.540e-01  1.800e-01  -1.411 0.158234
## ageSeniors                    9.177e-01  3.518e-01   2.609 0.009083 **
## ageSixties                    7.057e-01  2.793e-01   2.527 0.011508 *
## ageTeenagers                 -3.820e-01  6.108e-01  -0.625 0.531767
## ageThirties                  -2.299e-01  1.646e-01  -1.397 0.162527
## ageTwenties                  -1.822e-01  2.066e-01  -0.882 0.377699
## jobblue-collar               -1.138e-02  2.091e-01  -0.054 0.956595
## jobentrepreneur              -4.709e-01  3.954e-01  -1.191 0.233613
## jobhousemaid                 -4.068e-01  4.004e-01  -1.016 0.309739
## jobmanagement                 5.645e-02  2.007e-01   0.281 0.778501
## jobretired                   -5.063e-01  2.954e-01  -1.714 0.086566 .
## jobself-employed             -7.301e-03  2.886e-01  -0.025 0.979819
## jobservices                  -9.161e-02  2.228e-01  -0.411 0.681005
## jobstudent                    2.855e-01  2.343e-01   1.219 0.222907
## jobtechnician                 2.697e-01  1.746e-01   1.545 0.122408
## jobunemployed                 4.726e-01  2.971e-01   1.591 0.111636
## maritalmarried                2.613e-02  1.714e-01   0.152 0.878796
## maritalsingle                -7.039e-02  2.009e-01  -0.350 0.726051
## educationbasic.6y            -1.920e-01  3.270e-01  -0.587 0.557153
## educationbasic.9y            -1.006e-01  2.461e-01  -0.409 0.682635
```

```
## educationhigh.school          -4.730e-02  2.326e-01  -0.203 0.838908
## educationilliterate            1.805e+00  2.457e+00   0.734 0.462744
## educationprofessional.course   5.069e-02  2.491e-01   0.203 0.838781
## educationuniversity.degree     1.654e-01  2.309e-01   0.716 0.473730
## defaultyes                    -1.098e+01  2.715e+02  -0.040 0.967748
## housingyes                    -9.632e-02  9.959e-02  -0.967 0.333464
## loanyes                       -1.964e-01  1.433e-01  -1.371 0.170496
## contacttelephone              -4.850e-01  2.008e-01  -2.415 0.015739 *
## monthaug                       1.358e+00  3.834e-01   3.541 0.000398 ***
## monthdec                       1.363e+00  7.216e-01   1.889 0.058850 .
## monthjul                       6.002e-01  3.436e-01   1.747 0.080693 .
## monthjun                       1.722e-01  3.043e-01   0.566 0.571369
## monthmar                       2.566e+00  5.419e-01   4.735 2.19e-06 ***
## monthmay                       1.363e-01  2.014e-01   0.677 0.498515
## monthnov                       1.342e+00  7.114e-01   1.886 0.059266 .
## monthoct                       2.068e+00  8.481e-01   2.438 0.014763 *
## monthsep                       2.415e+00  9.208e-01   2.623 0.008715 **
## day_of_weekmon                -3.174e-01  1.627e-01  -1.951 0.051055 .
## day_of_weekthu                 1.801e-01  1.554e-01   1.159 0.246371
## day_of_weektue                 2.114e-01  1.607e-01   1.315 0.188400
## day_of_weekwed                 2.692e-01  1.620e-01   1.662 0.096567 .
## duration                       4.061e-03  2.296e-04  17.691  < 2e-16 ***
## campaign                      -9.683e-02  3.997e-02  -2.423 0.015397 *
## pdays                         -8.595e-04  2.709e-04  -3.173 0.001507 **
## previous                      -6.735e-02  7.573e-02  -0.889 0.373771
## poutcomesuccess                9.915e-01  2.655e-01   3.734 0.000188 ***
## emp.var.rate                  -1.841e+00  3.769e-01  -4.884 1.04e-06 ***
## cons.price.idx                 4.645e+00  1.173e+00   3.960 7.48e-05 ***
## cons.conf.idx                  1.284e-01  3.015e-02   4.259 2.06e-05 ***
## euribor3m                     -2.023e+00  9.350e-01  -2.164 0.030502 *
## nr.employed                    5.153e-02  1.965e-02   2.622 0.008740 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4504.7  on 3912  degrees of freedom
## Residual deviance: 2668.6  on 3861  degrees of freedom
## AIC: 2772.6
##
## Number of Fisher Scoring iterations: 12
```

```
oldCust_logitResult <- predict(oldCust_logit, newdata=oldCust_test,
type='response')
oldCust_logitResult <- ifelse(oldCust_logitResult >= 0.5,1,0)
oldCust_logitError  <- mean(oldCust_logitResult != oldCust_test$y)

print(paste('Accuracy for Logistic Model (oldCust)',1-oldCust_logitError))
```

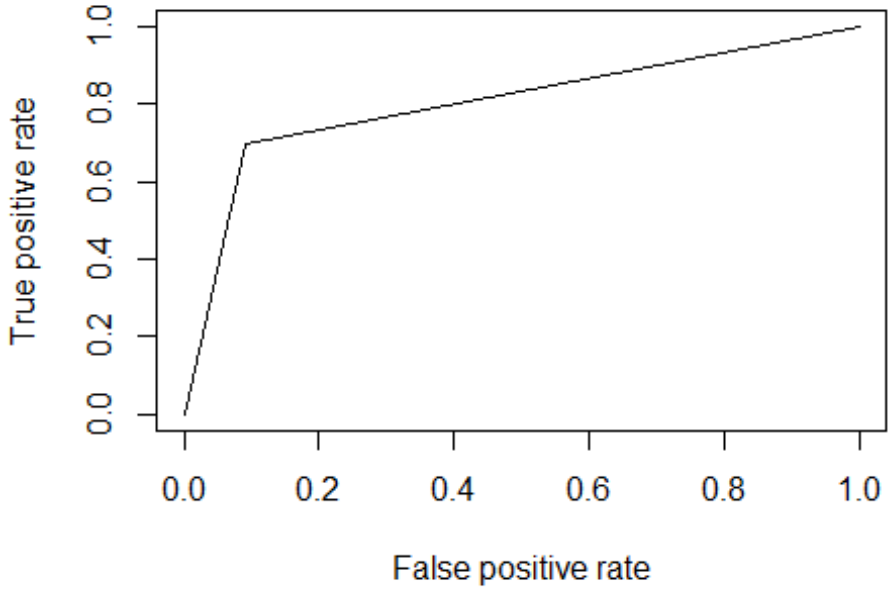
```
## [1] "Accuracy for Logistic Model (oldCust) 0.839369158878505"
```

```
#"Accuracy for Logistic Model (oldCust) 0.839369158878505"

library(ROCR)

## Loading required package: gplots

##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess

oldCust_logitPred <- prediction(oldCust_logitResult, oldCust_test$y)
oldCust_logitPerf <- performance(oldCust_logitPred, measure = "tpr",
x.measure = "fpr")
plot(oldCust_logitPerf)
```




```
oldCust_logitAUC <- performance(oldCust_logitPred, measure = "auc")
oldCust_logitAUC <- oldCust_logitAUC@y.values[[1]]

print(paste('Area under the Curve for Logistic Model
(oldCust)',oldCust_logitAUC))

## [1] "Area under the Curve for Logistic Model (oldCust) 0.784793603061782"

#"Area under the Curve for Logistic Model (oldCust) 0.784793603061782"
```

```
#########################   Random Forest Model (oldCust)
#########################


library(randomForest)

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':
##
##     margin

oldCust_rf<-randomForest(y ~.,data = oldCust_train, importance=TRUE,
ntree=1000)

oldCust_rfResult <- predict(oldCust_rf, oldCust_test)
oldCust_rfError  <- mean(oldCust_rfResult != oldCust_test$y)

print(paste('Accuracy for Random Forest Model (oldCust)',1-oldCust_rfError))

## [1] "Accuracy for Random Forest Model (oldCust) 0.851635514018692"

#"Accuracy for Random Forest Model (oldCust) 0.851635514018692"

library(ROCR)

oldCust_rfPred <- prediction(as.numeric(oldCust_rfResult),
as.numeric(oldCust_test$y))
oldCust_rfPerf <- performance(oldCust_rfPred, measure = "tpr", x.measure =
"fpr")
plot(oldCust_rfPerf)
```

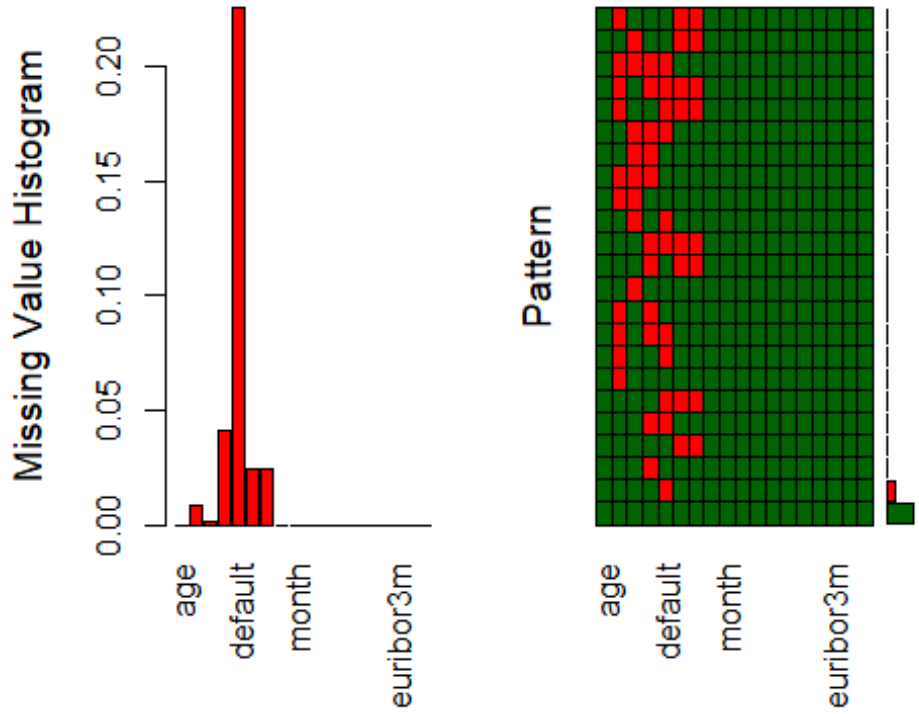



```
oldCust_rfAUC <- performance(oldCust_rfPred, measure = "auc")
oldCust_rfAUC <- oldCust_rfAUC@y.values[[1]]

print(paste('Area under the Curve for Random Forest Model
(oldCust)',oldCust_rfAUC))

## [1] "Area under the Curve for Random Forest Model (oldCust)
0.803758884636413"

#"Area under the Curve for Random Forest Model (oldCust) 0.803758884636413"
```

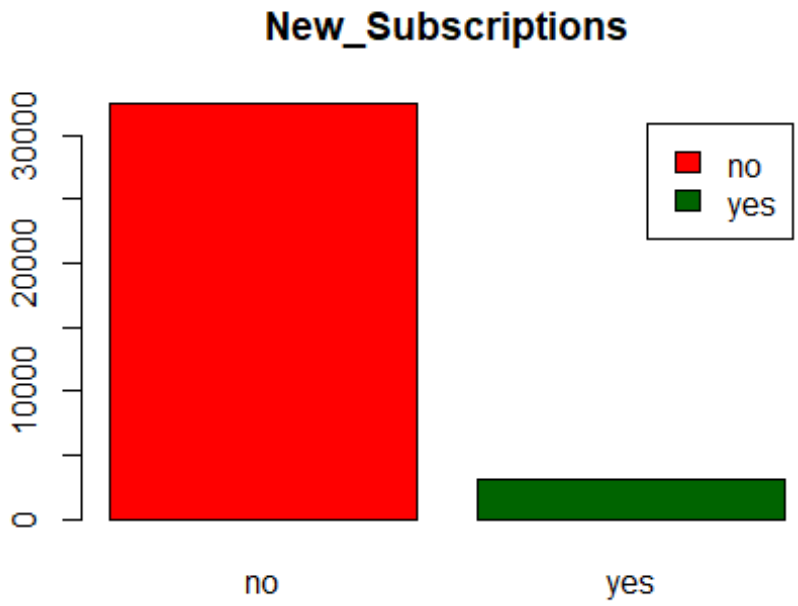
New Customer DATASET

```
#########################                          #########################
           #####         New Customer DATASET          #####
#########################                          #########################




# Since they are the new customers, it does not make sense to know
# their outcome from the previous campaign,
# number of previous contacts,
# amount of day passed from their last contact

newCust$poutcome <-NULL
```

```
newCust$previous <-NULL
newCust$pdays    <-NULL


# Missing value Frequencies
library(VIM)

aggrPlot <- aggr(newCust, col=c('darkgreen','red'), ylab=c("Missing Value
Histogram","Pattern"))
```



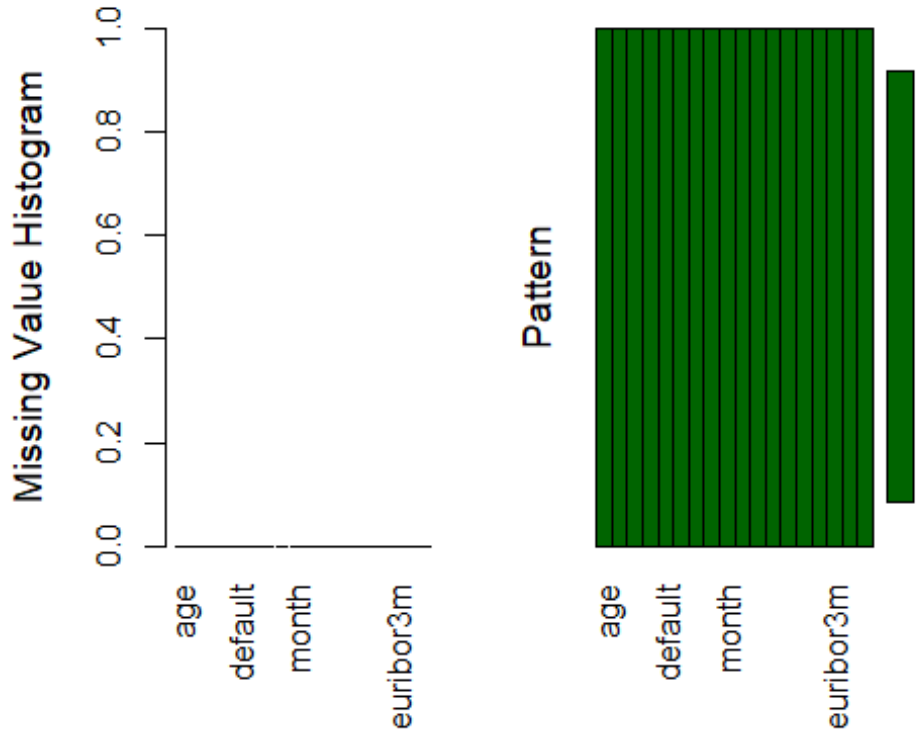

```
#education 0.0411 #housing 0.0239 #loan 0.0239 #job 0.0082 #marital 0.0017

#Subscription Count
newCount <- table(newCust$y)
barplot(newCount,col=c("red","darkgreen"),legend = rownames(newCount), main =
"New_Subscriptions")
```




```
#no 32422 #yes 3141


# Impute Missing Values and Check
library(mice)

newCust2 <- mice(newCust)

##
##  iter imp variable
##   1   1  job  marital  education  default  housing  loan
##   1   2  job  marital  education  default  housing  loan
##   1   3  job  marital  education  default  housing  loan
##   1   4  job  marital  education  default  housing  loan
##   1   5  job  marital  education  default  housing  loan
##   2   1  job  marital  education  default  housing  loan
##   2   2  job  marital  education  default  housing  loan
##   2   3  job  marital  education  default  housing  loan
##   2   4  job  marital  education  default  housing  loan
##   2   5  job  marital  education  default  housing  loan
##   3   1  job  marital  education  default  housing  loan
##   3   2  job  marital  education  default  housing  loan
##   3   3  job  marital  education  default  housing  loan
##   3   4  job  marital  education  default  housing  loan
##   3   5  job  marital  education  default  housing  loan
##   4   1  job  marital  education  default  housing  loan
```

```
##    4    2  job  marital  education  default  housing  loan
##    4    3  job  marital  education  default  housing  loan
##    4    4  job  marital  education  default  housing  loan
##    4    5  job  marital  education  default  housing  loan
##    5    1  job  marital  education  default  housing  loan
##    5    2  job  marital  education  default  housing  loan
##    5    3  job  marital  education  default  housing  loan
##    5    4  job  marital  education  default  housing  loan
##    5    5  job  marital  education  default  housing  loan

## Warning: Number of logged events: 150
```

```r
newCust_com <- complete(newCust2)

aggrPlot <- aggr(newCust_com, col=c('darkgreen','red'), ylab=c("Missing Value Histogram","Pattern"))
```




```r
#none


#Split data into Train and Test subsets
library(caret)
set.seed(102)

newCust_com$y<-ifelse(newCust_com$y =='no', 0,1)
newCust_com$y<-as.factor(newCust_com$y)
```

```r
id <- sample(seq(1, 2), size = nrow(newCust_com), replace = TRUE, prob =
c(.7, .3))

newCust_train <- newCust_com[id==1,]
newCust_test  <- newCust_com[id==2,]

table(newCust_train$y)
```

```
##
##     0     1
## 22745  2191
```

```r
#no 22745 #yes 2191
table(newCust_test$y)
```

```
##
##    0    1
## 9677  950
```

```r
#no 9677 #yes 950


#########################   Logistic Model (newCust)
#########################


newCust_logit <- glm(y ~., family=binomial(link='logit'), data =
newCust_train)
summary(newCust_logit)
```

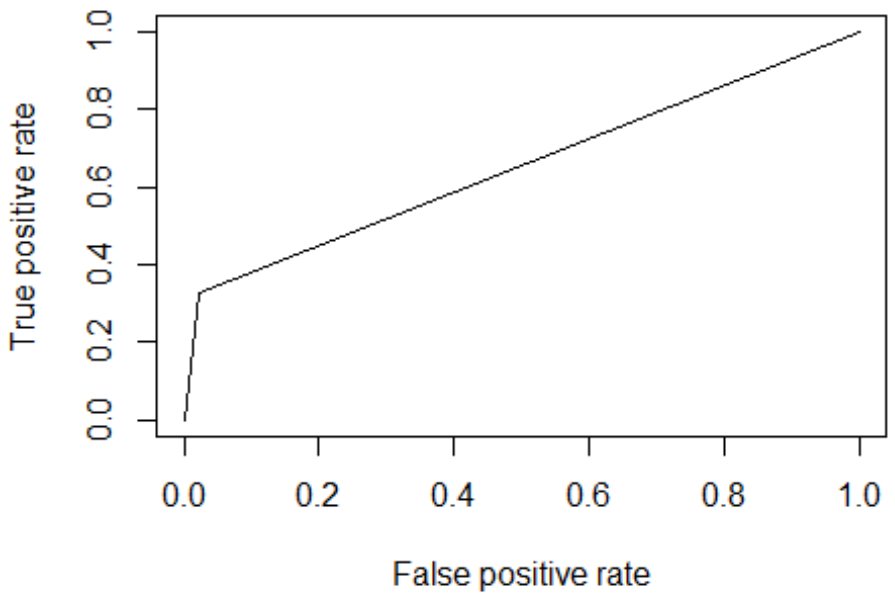
```
##
## Call:
## glm(formula = y ~ ., family = binomial(link = "logit"), data =
newCust_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.0972  -0.2700  -0.1773  -0.1307   3.4303
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -2.768e+02  5.716e+01  -4.842 1.29e-06 ***
## ageForties                  -1.712e-01  9.809e-02  -1.745 0.080908 .
## ageSeniors                   1.197e-01  2.300e-01   0.520 0.602804
## ageSixties                   1.697e-01  1.882e-01   0.902 0.367273
## ageTeenagers                 6.472e-01  4.895e-01   1.322 0.186093
## ageThirties                 -2.845e-02  9.198e-02  -0.309 0.757088
## ageTwenties                  8.585e-02  1.160e-01   0.740 0.459357
## jobblue-collar              -3.047e-01  1.106e-01  -2.755 0.005868 **
## jobentrepreneur             -5.556e-02  1.638e-01  -0.339 0.734540
```

```
## jobhousemaid                   -2.028e-02  1.961e-01  -0.103 0.917637
## jobmanagement                  -1.350e-01  1.179e-01  -1.146 0.251995
## jobretired                      1.215e-01  1.689e-01   0.719 0.471948
## jobself-employed               -1.737e-01  1.597e-01  -1.088 0.276788
## jobservices                    -2.124e-01  1.203e-01  -1.766 0.077390 .
## jobstudent                      1.172e-01  1.652e-01   0.710 0.477845
## jobtechnician                  -1.411e-01  1.005e-01  -1.405 0.160153
## jobunemployed                  -3.080e-01  1.901e-01  -1.620 0.105167
## maritalmarried                 -7.620e-02  9.515e-02  -0.801 0.423176
## maritalsingle                   7.386e-02  1.079e-01   0.684 0.493749
## educationbasic.6y               2.411e-01  1.581e-01   1.525 0.127209
## educationbasic.9y               1.285e-01  1.280e-01   1.004 0.315428
## educationhigh.school            1.509e-01  1.270e-01   1.188 0.234877
## educationilliterate             9.088e-01  7.976e-01   1.139 0.254535
## educationprofessional.course    2.514e-01  1.400e-01   1.796 0.072537 .
## educationuniversity.degree      3.183e-01  1.270e-01   2.506 0.012222 *
## defaultyes                     -6.301e+00  1.195e+02  -0.053 0.957937
## housingyes                      3.168e-02  5.701e-02   0.556 0.578358
## loanyes                        -4.040e-02  7.932e-02  -0.509 0.610483
## contacttelephone               -6.046e-01  1.069e-01  -5.655 1.56e-08 ***
## monthaug                        1.219e+00  1.866e-01   6.531 6.52e-11 ***
## monthdec                        6.379e-01  3.248e-01   1.964 0.049519 *
## monthjul                        8.298e-02  1.327e-01   0.625 0.531788
## monthjun                       -8.788e-01  1.920e-01  -4.578 4.69e-06 ***
## monthmar                        2.104e+00  2.106e-01   9.993  < 2e-16 ***
## monthmay                       -6.112e-01  1.205e-01  -5.072 3.95e-07 ***
## monthnov                       -5.562e-01  1.678e-01  -3.316 0.000914 ***
## monthoct                        2.999e-01  2.214e-01   1.355 0.175466
## monthsep                        4.209e-01  2.711e-01   1.552 0.120590
## day_of_weekmon                 -4.322e-02  9.178e-02  -0.471 0.637680
## day_of_weekthu                  8.027e-02  8.938e-02   0.898 0.369163
## day_of_weektue                  1.227e-01  9.183e-02   1.336 0.181533
## day_of_weekwed                  1.766e-01  9.158e-02   1.928 0.053870 .
## duration                        4.818e-03  9.737e-05  49.483  < 2e-16 ***
## campaign                       -2.638e-02  1.432e-02  -1.842 0.065462 .
## emp.var.rate                   -2.122e+00  2.170e-01  -9.781  < 2e-16 ***
## cons.price.idx                  2.578e+00  3.812e-01   6.763 1.35e-11 ***
## cons.conf.idx                   1.981e-03  1.203e-02   0.165 0.869287
## euribor3m                       5.476e-01  1.815e-01   3.017 0.002556 **
## nr.employed                     5.743e-03  4.557e-03   1.260 0.207539
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14840.4  on 24935  degrees of freedom
## Residual deviance:  9075.7  on 24887  degrees of freedom
## AIC: 9173.7
##
## Number of Fisher Scoring iterations: 9
```

```r
newCust_logitResult <- predict(newCust_logit, newdata=newCust_test,
type='response')
newCust_logitResult <- ifelse(newCust_logitResult >= 0.5,1,0)
newCust_logitError  <- mean(newCust_logitResult != newCust_test$y)

print(paste('Accuracy for Logistic Model (newCust)',1-newCust_logitError))
```

```
## [1] "Accuracy for Logistic Model (newCust) 0.920956055330761"
```

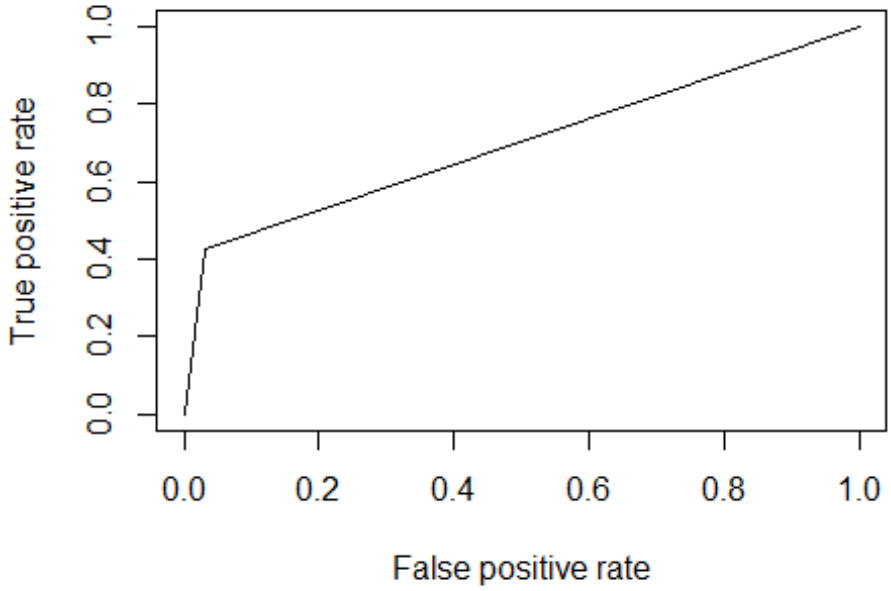
```r
#"Accuracy for Logistic Model (newCust) 0.920956055330761"

library(ROCR)

newCust_logitPred <- prediction(newCust_logitResult, newCust_test$y)
newCust_logitPerf <- performance(newCust_logitPred, measure = "tpr",
x.measure = "fpr")
plot(newCust_logitPerf)
```




```r
newCust_logitAUC <- performance(newCust_logitPred, measure = "auc")
newCust_logitAUC <- newCust_logitAUC@y.values[[1]]

print(paste('Area under the Curve for Logistic Model
(newCust)',newCust_logitAUC))
```

```
## [1] "Area under the Curve for Logistic Model (newCust) 0.653773407373969"
```

```r
#"Area under the Curve for Logistic Model (newCust) 0.653773407373969"


#########################   Random Forest Model (newCust)
#########################


library(randomForest)

newCust_rf<-randomForest(y ~.,data = newCust_train, importance=TRUE,
ntree=1000)

newCust_rfResult <- predict(newCust_rf, newCust_test)
newCust_rfError  <- mean(newCust_rfResult != newCust_test$y)

print(paste('Accuracy for Random Forest Model (newCust)',1-newCust_rfError))

## [1] "Accuracy for Random Forest Model (newCust) 0.91935635645055"

#"Accuracy for Random Forest Model (newCust) 0.91935635645055"

library(ROCR)

newCust_rfPred <- prediction(as.numeric(newCust_rfResult),
as.numeric(newCust_test$y))
newCust_rfPerf <- performance(newCust_rfPred, measure = "tpr", x.measure =
"fpr")
plot(newCust_rfPerf)
```




```
newCust_rfAUC <- performance(newCust_rfPred, measure = "auc")
newCust_rfAUC <- newCust_rfAUC@y.values[[1]]

print(paste('Area under the Curve for Random Forest Model
(newCust)',newCust_rfAUC))

## [1] "Area under the Curve for Random Forest Model (newCust)
0.696562549289417"

#"Area under the Curve for Random Forest Model (newCust) 0.696562549289417"
```