Analysis of a Portuguese Bank Marketing Dataset

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

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 Analysis

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

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

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

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

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

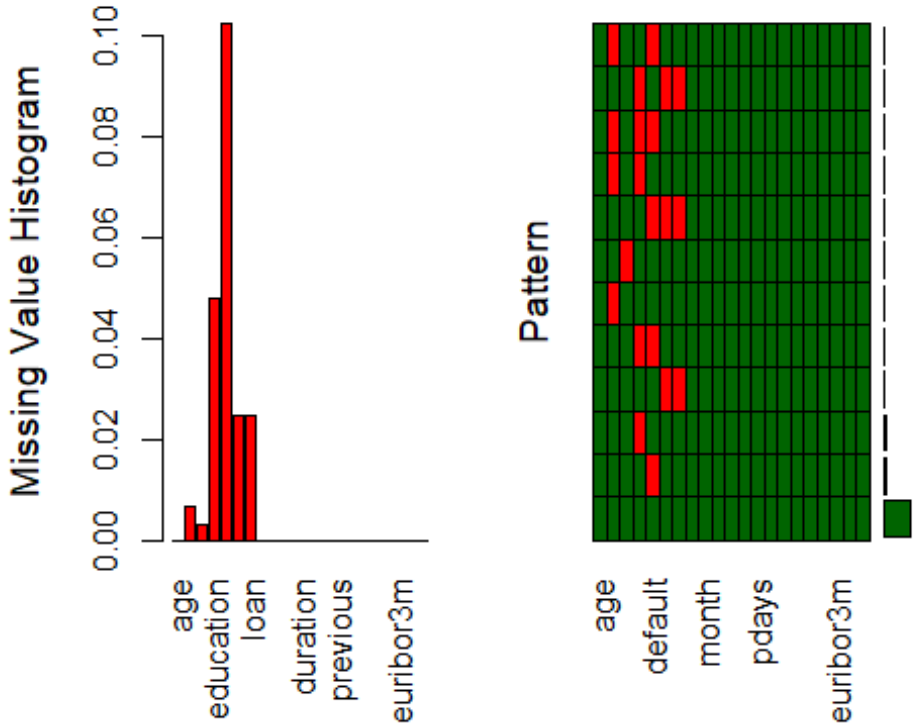
```
## VIM is ready to use.
##  Since version 4.0.0 the GUI is in its own package VIMGUI.
##
##           Please use the package to use the new (and old) GUI.
```

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

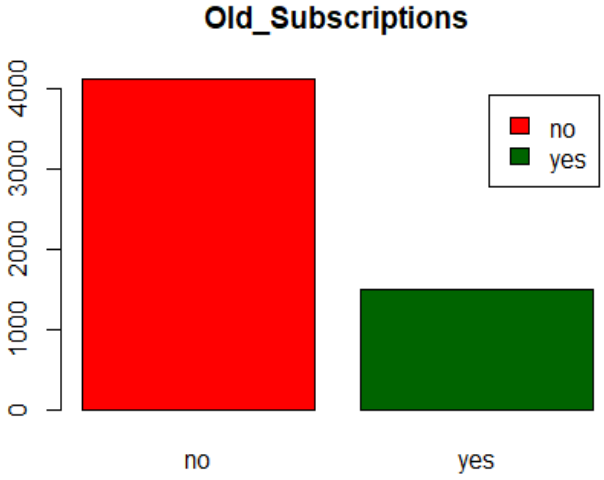
```
##
## Attaching package: 'VIM'
```

```
## 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")
```



```r
#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
```

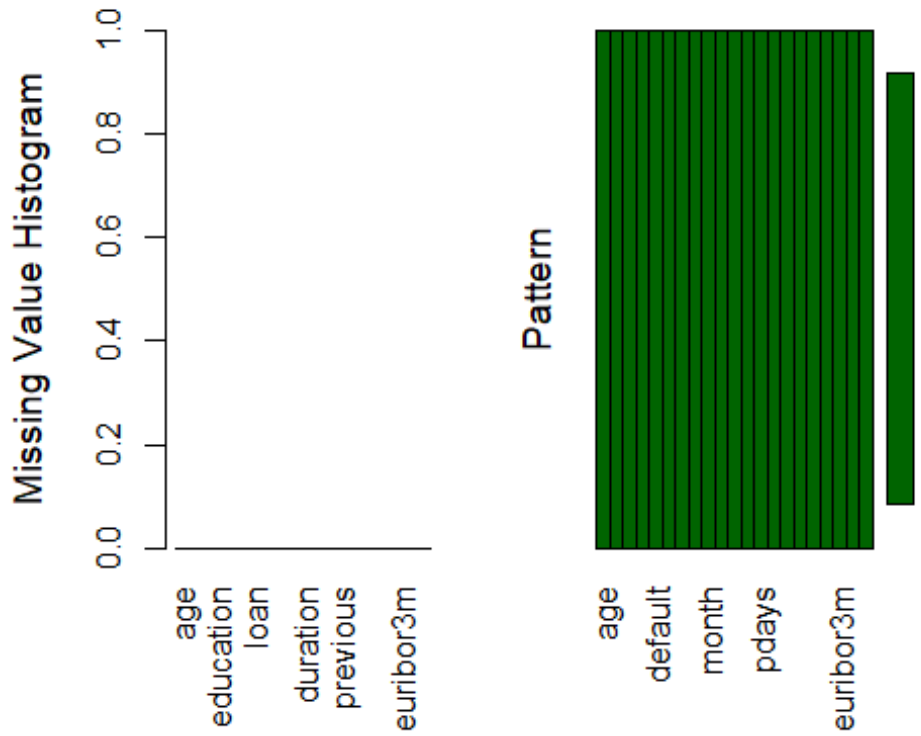
```
##
## Attaching package: 'mice'
```

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

```r
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,]
```

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

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

```r
table(oldCust_test$y)  #no 1240 #yes 472
```

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

```r
######################   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
## -4.9070  -0.4892  -0.2377   0.3293   3.1358
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -7.031e+02  2.057e+02  -3.417 0.000633 ***
## ageForties                   -2.439e-01  1.802e-01  -1.353 0.176013
## ageSeniors                    7.941e-01  3.488e-01   2.277 0.022804 *
## ageSixties                    6.452e-01  2.762e-01   2.336 0.019475 *
## ageTeenagers                 -4.118e-01  6.176e-01  -0.667 0.504947
## ageThirties                  -2.141e-01  1.650e-01  -1.297 0.194508
## ageTwenties                  -1.694e-01  2.067e-01  -0.820 0.412290
## jobblue-collar               -9.971e-02  2.094e-01  -0.476 0.633897
## jobentrepreneur              -4.667e-01  3.970e-01  -1.176 0.239706
## jobhousemaid                 -5.371e-01  3.964e-01  -1.355 0.175503
## jobmanagement                 9.174e-02  2.011e-01   0.456 0.648234
## jobretired                   -4.359e-01  2.930e-01  -1.488 0.136838
## jobself-employed              4.973e-02  2.833e-01   0.176 0.860638
## jobservices                  -1.187e-01  2.231e-01  -0.532 0.594568
## jobstudent                    2.280e-01  2.350e-01   0.970 0.331963
## jobtechnician                 3.067e-01  1.744e-01   1.758 0.078674 .
## jobunemployed                 3.727e-01  2.978e-01   1.252 0.210712
## maritalmarried                3.946e-02  1.713e-01   0.230 0.817783
## maritalsingle                -4.918e-02  2.006e-01  -0.245 0.806333
## educationbasic.6y            -2.336e-01  3.231e-01  -0.723 0.469697
```

```
## educationbasic.9y             -1.510e-01  2.456e-01  -0.615 0.538541
## educationhigh.school          -1.625e-01  2.316e-01  -0.702 0.482987
## educationilliterate            1.683e+00  2.377e+00   0.708 0.479008
## educationprofessional.course  -5.515e-02  2.481e-01  -0.222 0.824077
## educationuniversity.degree    -1.238e-02  2.305e-01  -0.054 0.957174
## defaultyes                    -1.030e+01  2.943e+02  -0.035 0.972092
## housingyes                    -6.698e-02  9.946e-02  -0.673 0.500699
## loanyes                       -2.509e-01  1.442e-01  -1.740 0.081929 .
## contacttelephone              -4.811e-01  2.003e-01  -2.402 0.016323 *
## monthaug                       1.376e+00  3.834e-01   3.589 0.000332 ***
## monthdec                       1.425e+00  7.211e-01   1.977 0.048094 *
## monthjul                       6.105e-01  3.436e-01   1.777 0.075609 .
## monthjun                       1.749e-01  3.039e-01   0.575 0.565013
## monthmar                       2.586e+00  5.424e-01   4.767 1.87e-06 ***
## monthmay                       1.356e-01  2.014e-01   0.674 0.500612
## monthnov                       1.376e+00  7.105e-01   1.936 0.052811 .
## monthoct                       2.111e+00  8.474e-01   2.491 0.012741 *
## monthsep                       2.457e+00  9.205e-01   2.669 0.007613 **
## day_of_weekmon                -3.195e-01  1.626e-01  -1.965 0.049448 *
## day_of_weekthu                 1.740e-01  1.553e-01   1.121 0.262408
## day_of_weektue                 2.097e-01  1.605e-01   1.306 0.191472
## day_of_weekwed                 2.720e-01  1.622e-01   1.677 0.093478 .
## duration                       4.064e-03  2.297e-04  17.690  < 2e-16 ***
## campaign                      -9.469e-02  3.997e-02  -2.369 0.017841 *
## pdays                         -8.631e-04  2.706e-04  -3.190 0.001423 **
## previous                      -6.440e-02  7.569e-02  -0.851 0.394892
## poutcomesuccess                9.879e-01  2.653e-01   3.724 0.000196 ***
## emp.var.rate                  -1.849e+00  3.769e-01  -4.906 9.29e-07 ***
## cons.price.idx                 4.709e+00  1.172e+00   4.017 5.89e-05 ***
## cons.conf.idx                  1.310e-01  3.011e-02   4.352 1.35e-05 ***
## euribor3m                     -2.075e+00  9.337e-01  -2.222 0.026269 *
## nr.employed                    5.259e-02  1.964e-02   2.678 0.007397 **
## ---
## 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: 2669.3  on 3861  degrees of freedom
## AIC: 2773.3
##
## 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) 0.844042056074766"
```

```
#Accuracy = 84.40%
```

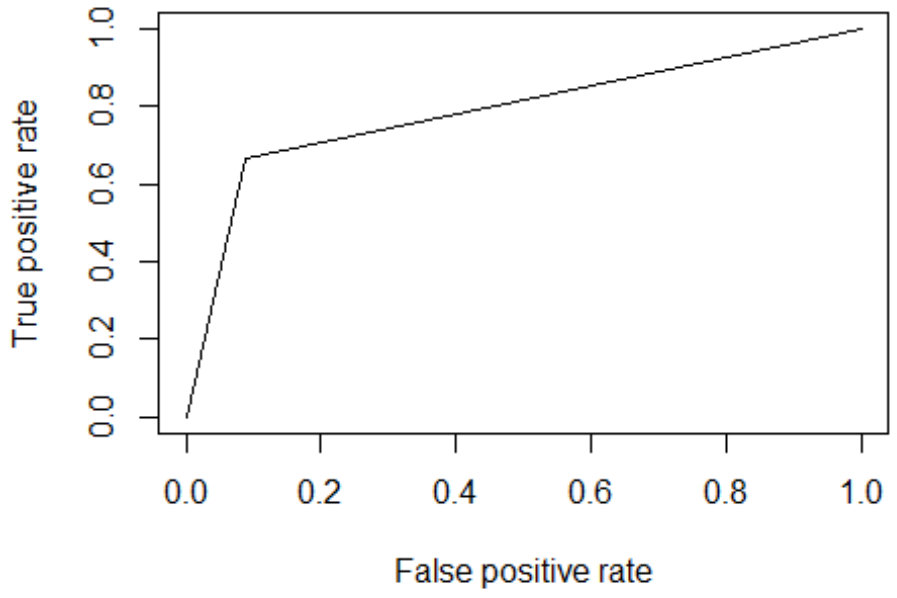
```
library(ROCR)
```

```
## Loading required package: gplots
```

```
##
## Attaching package: 'gplots'
```

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

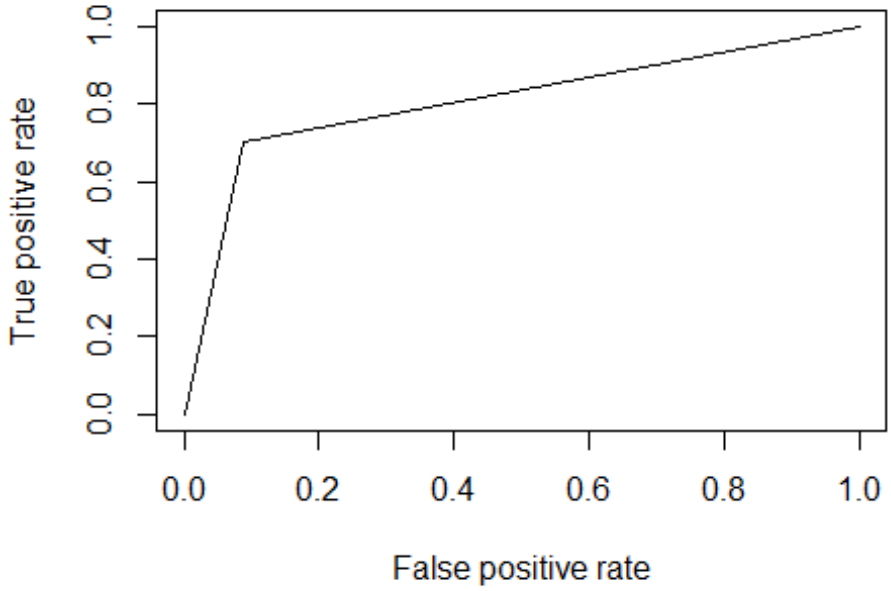
```
oldCust_logitPred <- prediction(oldCust_logitResult, oldCust_test$y)
oldCust_logitPerf <- performance(oldCust_logitPred, measure = "tpr",
x.measure = "fpr")
plot(oldCust_logitPerf)
```




```
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))
```

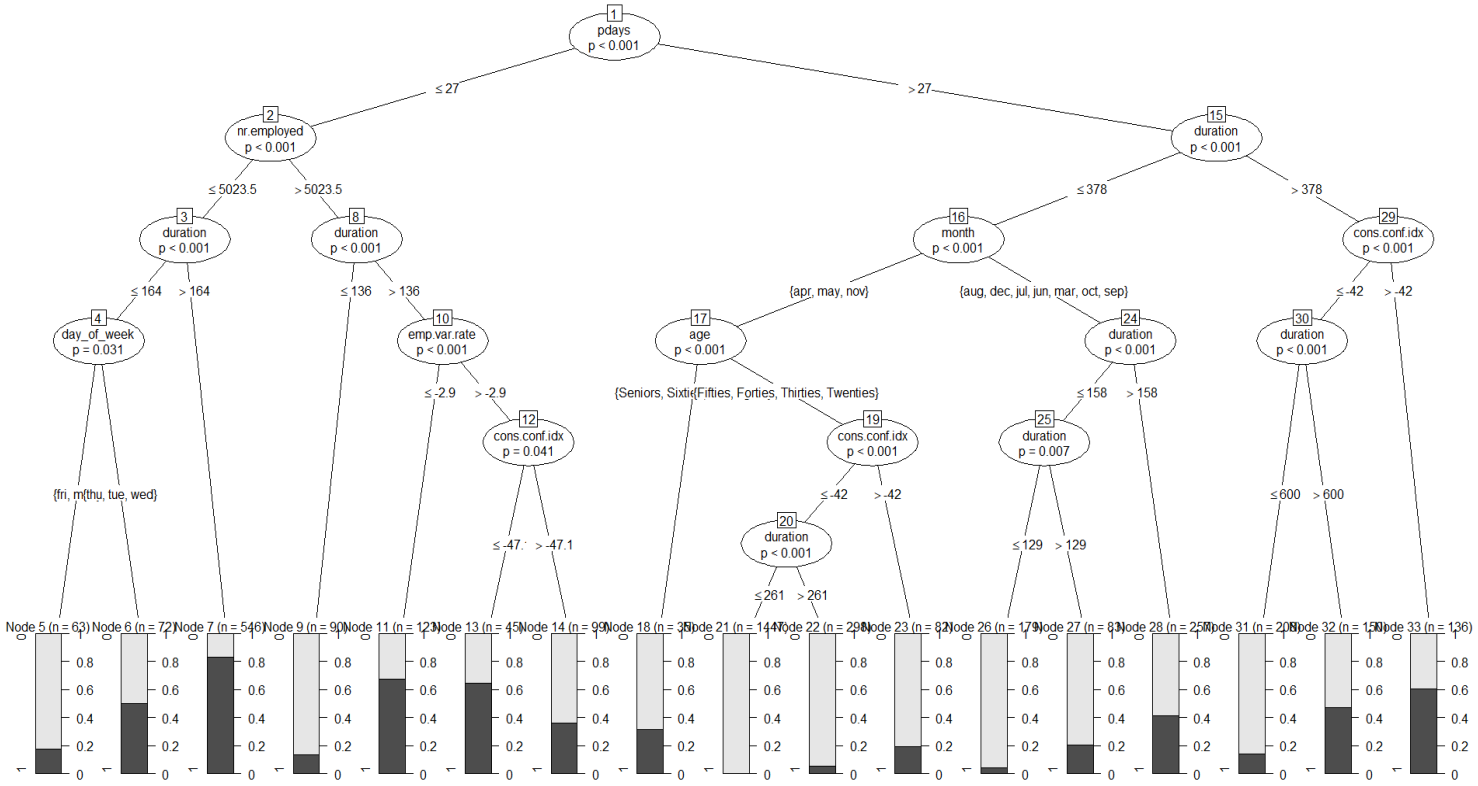
```
## [1] "Area under the Curve for Logistic Model (oldCust) 0.789331601968289"
```

```r
#Area under Curve = 78.93%

####################   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) 0.854556074766355"

#Accuracy = 85.46%

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)
```

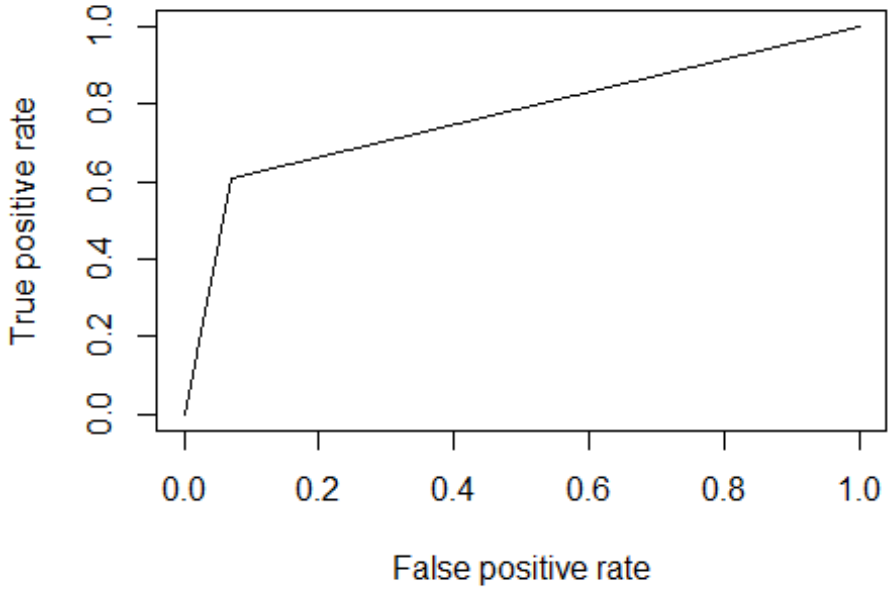



```
oldCust_rfAUC <- performance(oldCust_rfPred, measure = "auc")
oldCust_rfAUC <- oldCust_rfAUC@y.values[[1]]

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

## [1] "Area under the Curve for Random Forest Model (oldCust)
0.807743302351011"

#Area under Curve = 80.77%

##########################   Tree Model (oldCust)   ##########################


library(party)

## Loading required package: mvtnorm

## Loading required package: modeltools

## Loading required package: stats4

## Loading required package: strucchange

## Loading required package: zoo

##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

oldCust_tree<-ctree(y ~.,data = oldCust_train)
plot(oldCust_tree)
```




```
oldCust_treeResult <- predict(oldCust_tree, oldCust_test)
oldCust_treeError  <- mean(oldCust_treeResult != oldCust_test$y)

print(paste('Accuracy for Tree Model (oldCust)',1-oldCust_treeError))

## [1] "Accuracy for Tree Model (oldCust) 0.842289719626168"

#Accuracy = 84.23%

library(ROCR)

oldCust_treePred <- prediction(as.numeric(oldCust_treeResult),
as.numeric(oldCust_test$y))
oldCust_treePerf <- performance(oldCust_treePred, measure = "tpr", x.measure
= "fpr")
plot(oldCust_treePerf)
```

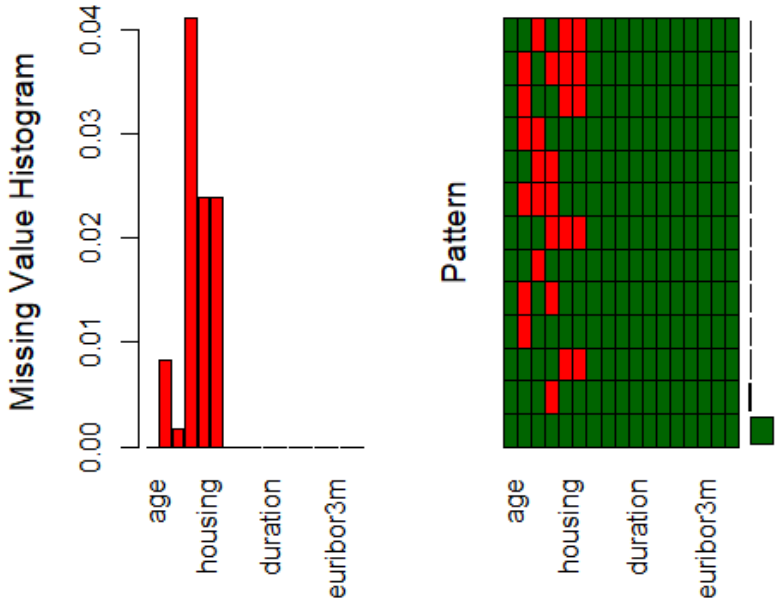



```
oldCust_treeAUC <- performance(oldCust_treePred, measure = "auc")
oldCust_treeAUC <- oldCust_treeAUC@y.values[[1]]
print(paste('Area under the Curve for Tree Model (oldCust)',oldCust_treeAUC))

## [1] "Area under the Curve for Tree Model (oldCust) 0.770407326407873"

#Area under Curve = 77.04%
```

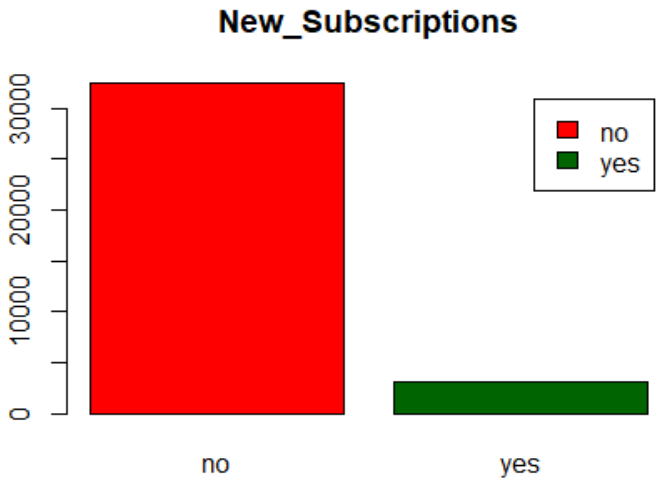
New Customer DATASET Analysis

```
#########################                              #########################
          #####          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 and
# their default credit with the bank

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

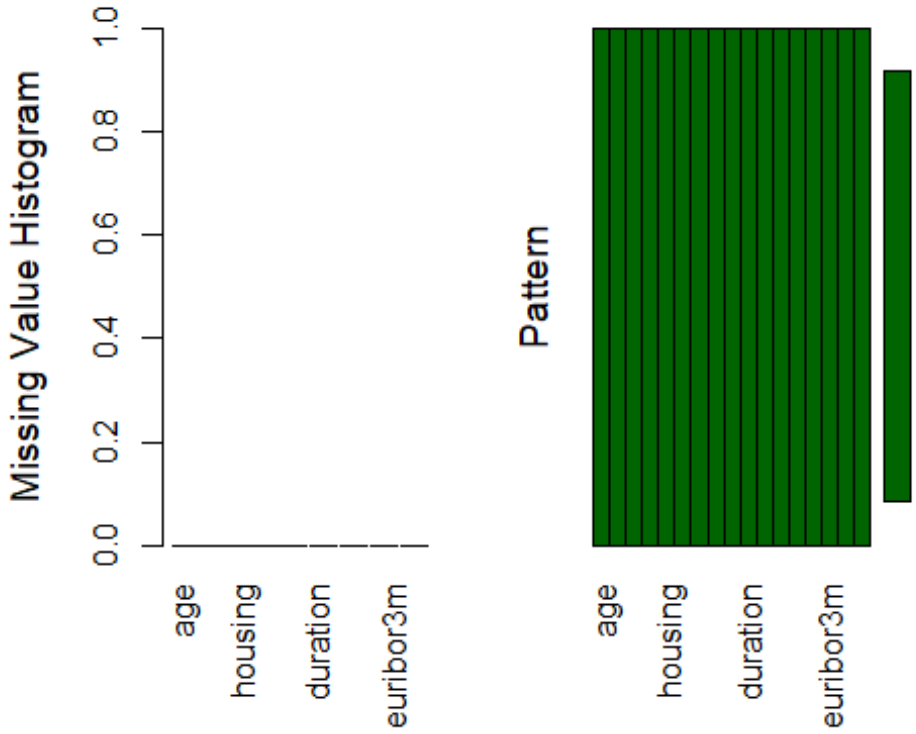
```r
# Missing value Frequencies
library(VIM)

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




```r
#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")
```



```r
#no 32422 #yes 3141


# Impute Missing Values and Check
library(mice)

newCust2 <- mice(newCust)

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

## Warning: Number of logged events: 125

newCust_com <- complete(newCust2)

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




```
#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)

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) #no 22745 #yes 2191

##
##     0     1
## 22745  2191

table(newCust_test$y)  #no 9677  #yes 950

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

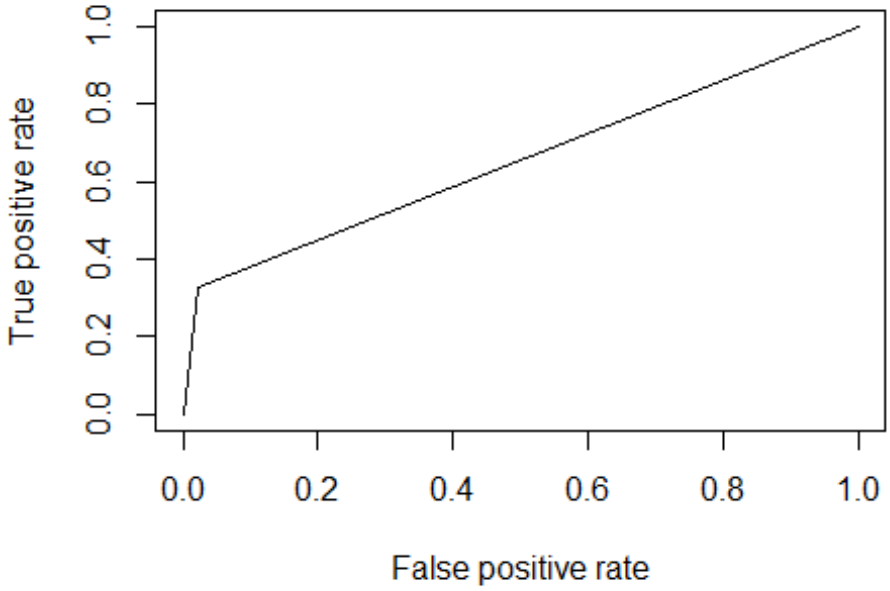
```r
newCust_trains<- newCust_train
# saving train data before making the model

######################   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
## -6.0989  -0.2696  -0.1771  -0.1306   3.4308
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -2.760e+02  5.716e+01  -4.828 1.38e-06 ***
## ageForties                   -1.746e-01  9.814e-02  -1.779 0.075282 .
## ageSeniors                    1.348e-01  2.298e-01   0.587 0.557492
## ageSixties                    1.639e-01  1.880e-01   0.871 0.383496
## ageTeenagers                  6.788e-01  4.888e-01   1.389 0.164912
## ageThirties                  -3.170e-02  9.195e-02  -0.345 0.730281
## ageTwenties                   8.324e-02  1.161e-01   0.717 0.473289
## jobblue-collar               -2.844e-01  1.107e-01  -2.568 0.010219 *
## jobentrepreneur              -3.195e-02  1.631e-01  -0.196 0.844710
## jobhousemaid                  1.241e-02  1.960e-01   0.063 0.949540
## jobmanagement                -1.413e-01  1.181e-01  -1.197 0.231371
## jobretired                    1.190e-01  1.692e-01   0.703 0.482102
## jobself-employed             -1.683e-01  1.597e-01  -1.054 0.291912
## jobservices                  -1.981e-01  1.200e-01  -1.651 0.098658 .
## jobstudent                    1.395e-01  1.647e-01   0.847 0.396976
## jobtechnician                -1.387e-01  1.003e-01  -1.383 0.166553
## jobunemployed                -2.978e-01  1.900e-01  -1.567 0.117052
## maritalmarried               -7.268e-02  9.514e-02  -0.764 0.444887
## maritalsingle                 7.056e-02  1.079e-01   0.654 0.513199
## educationbasic.6y             2.449e-01  1.583e-01   1.547 0.121820
## educationbasic.9y             1.214e-01  1.279e-01   0.948 0.342879
## educationhigh.school          1.772e-01  1.269e-01   1.396 0.162565
## educationilliterate           1.328e+00  7.162e-01   1.854 0.063747 .
## educationprofessional.course  2.921e-01  1.401e-01   2.085 0.037032 *
## educationuniversity.degree    3.515e-01  1.269e-01   2.770 0.005605 **
## housingyes                    2.909e-02  5.702e-02   0.510 0.609966
## loanyes                      -2.397e-02  7.898e-02  -0.304 0.761481
## contacttelephone             -6.053e-01  1.069e-01  -5.660 1.51e-08 ***
## monthaug                      1.205e+00  1.867e-01   6.457 1.07e-10 ***
```

```
## monthdec                      6.254e-01  3.246e-01   1.927 0.053996 .
## monthjul                      7.643e-02  1.327e-01   0.576 0.564742
## monthjun                     -8.804e-01  1.920e-01  -4.586 4.52e-06 ***
## monthmar                      2.102e+00  2.105e-01   9.984  < 2e-16 ***
## monthmay                     -6.148e-01  1.205e-01  -5.103 3.34e-07 ***
## monthnov                     -5.631e-01  1.678e-01  -3.357 0.000789 ***
## monthoct                      2.960e-01  2.213e-01   1.337 0.181172
## monthsep                      4.101e-01  2.712e-01   1.512 0.130472
## day_of_weekmon               -4.346e-02  9.176e-02  -0.474 0.635745
## day_of_weekthu                7.998e-02  8.939e-02   0.895 0.370957
## day_of_weektue                1.234e-01  9.182e-02   1.344 0.178881
## day_of_weekwed                1.782e-01  9.158e-02   1.946 0.051646 .
## duration                      4.820e-03  9.738e-05  49.496  < 2e-16 ***
## campaign                     -2.664e-02  1.433e-02  -1.858 0.063138 .
## emp.var.rate                 -2.118e+00  2.170e-01  -9.760  < 2e-16 ***
## cons.price.idx                2.573e+00  3.812e-01   6.749 1.49e-11 ***
## cons.conf.idx                 2.088e-03  1.204e-02   0.173 0.862334
## euribor3m                     5.473e-01  1.816e-01   3.014 0.002582 **
## nr.employed                   5.691e-03  4.557e-03   1.249 0.211688
## ---
## 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:  9072.7  on 24888  degrees of freedom
## AIC: 9168.7
##
## Number of Fisher Scoring iterations: 6
```

```
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))
```

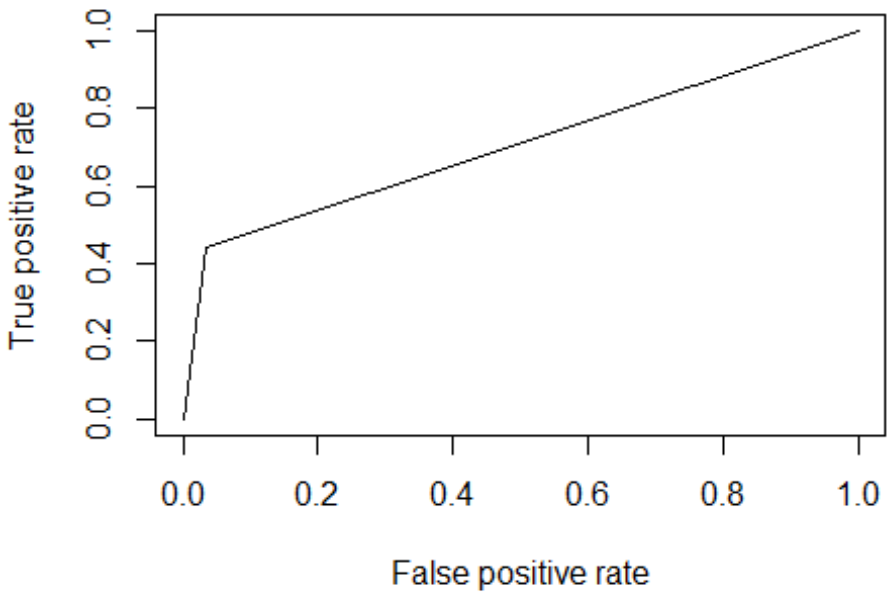
```
## [1] "Accuracy for Logistic Model (newCust) 0.920956055330761"
```

```
#Accuracy = 92.09%

library(ROCR)

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

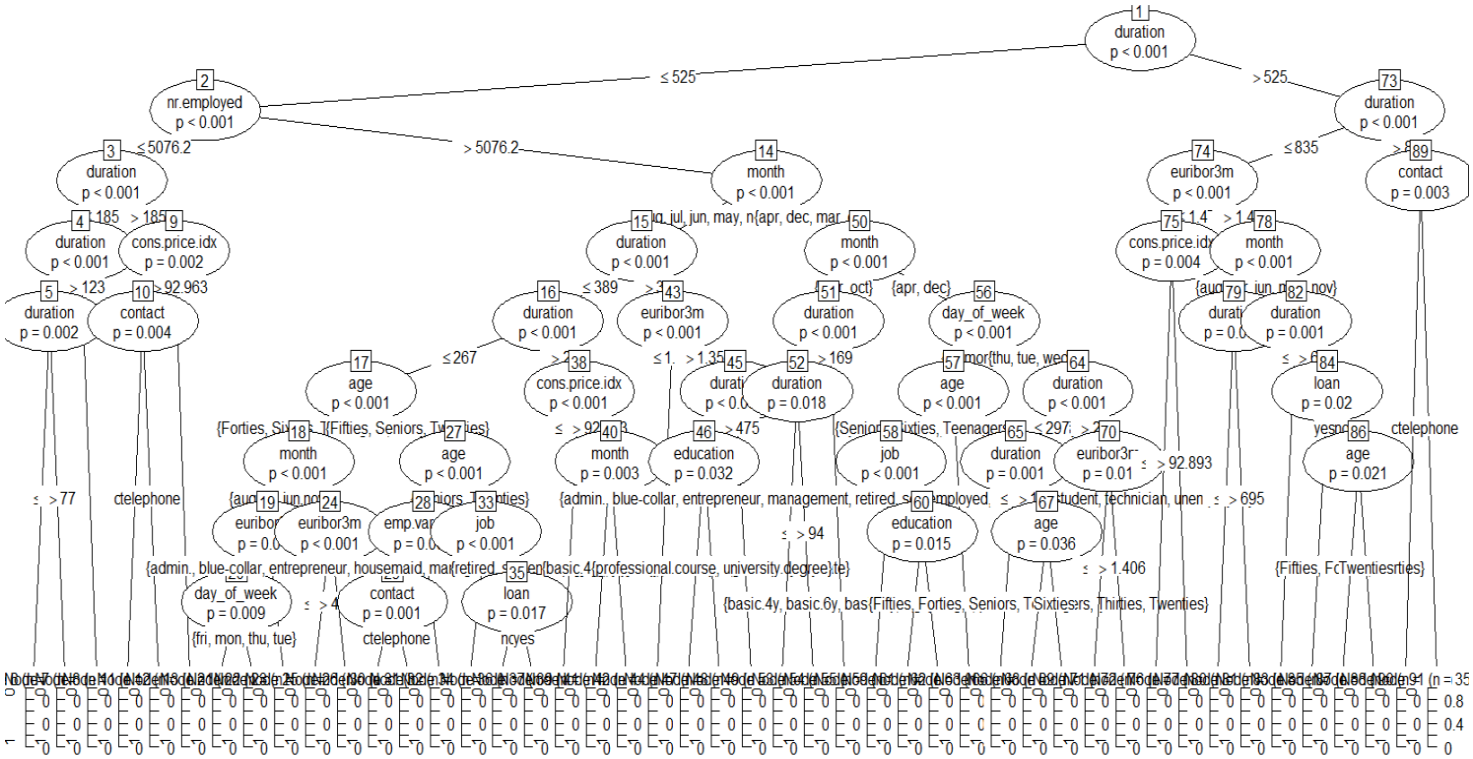



```
newCust_logitAUC <- performance(newCust_logitPred, measure = "auc")
newCust_logitAUC <- newCust_logitAUC@y.values[[1]]

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

## [1] "Area under the Curve for Logistic Model (newCust) 0.65424805425779"

#Area under Curve = 65.42%

####################   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 = 91.93%
```

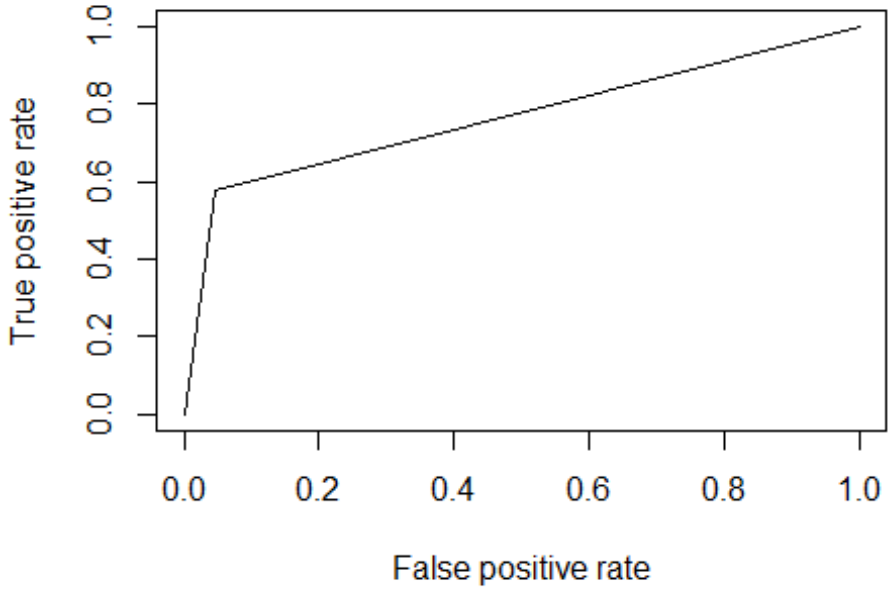
```
library(ROCR)

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)
0.703207605662912"

#Area under Curve = 70.32%

##########################   Tree Model (newCust)   ##########################


library(party)

newCust_tree<-ctree(y ~.,data = newCust_train)
plot(newCust_tree)
```




```
newCust_treeResult <- predict(newCust_tree, newCust_test)
newCust_treeError  <- mean(newCust_treeResult != newCust_test$y)

print(paste('Accuracy for Tree Model (newCust)',1-newCust_treeError))

## [1] "Accuracy for Tree Model (newCust) 0.921144255199021"

#Accuracy = 92.11%

library(ROCR)

newCust_treePred <- prediction(as.numeric(newCust_treeResult),
as.numeric(newCust_test$y))
newCust_treePerf <- performance(newCust_treePred, measure = "tpr", x.measure
= "fpr")
plot(newCust_treePerf)
```




```
newCust_treeAUC <- performance(newCust_treePred, measure = "auc")
newCust_treeAUC <- newCust_treeAUC@y.values[[1]]
print(paste('Area under the Curve for Tree Model (newCust)',newCust_treeAUC))

## [1] "Area under the Curve for Tree Model (newCust) 0.765418762883234"

#Area under Curve = 76.54%
```