Bank\_problem.R

Yash Vora

Wed Feb 27 15:57:07 2019

#Importing Libraries  
library(tidyverse)

## Loading tidyverse: ggplot2  
## Loading tidyverse: tibble  
## Loading tidyverse: tidyr  
## Loading tidyverse: readr  
## Loading tidyverse: purrr  
## Loading tidyverse: dplyr

## Warning: package 'purrr' was built under R version 3.4.4

## Conflicts with tidy packages ----------------------------------------------

## filter(): dplyr, stats  
## lag(): dplyr, stats

library(rockchalk)

## Warning: package 'rockchalk' was built under R version 3.4.4

##   
## Attaching package: 'rockchalk'

## The following object is masked from 'package:dplyr':  
##   
## summarize

library(caret)

## Warning: package 'caret' was built under R version 3.4.4

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

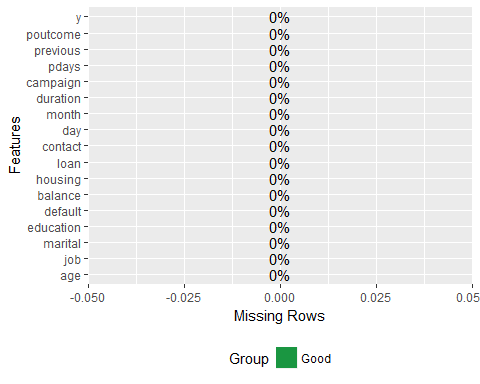
#Importing Dataset  
training\_set <- read.csv('bank-full.csv', sep = ';')  
test\_set <- read.csv('bank.csv', sep = ';')  
dataset <- rbind(training\_set, test\_set)  
  
#Data Preprocessing  
summary(dataset)

## age job marital education   
## Min. :18.00 blue-collar:10678 divorced: 5735 primary : 7529   
## 1st Qu.:33.00 management :10427 married :30011 secondary:25508   
## Median :39.00 technician : 8365 single :13986 tertiary :14651   
## Mean :40.96 admin. : 5649 unknown : 2044   
## 3rd Qu.:48.00 services : 4571   
## Max. :95.00 retired : 2494   
## (Other) : 7548   
## default balance housing loan contact   
## no :48841 Min. : -8019 no :22043 no :41797 cellular :32181   
## yes: 891 1st Qu.: 72 yes:27689 yes: 7935 telephone: 3207   
## Median : 448 unknown :14344   
## Mean : 1368   
## 3rd Qu.: 1431   
## Max. :102127   
##   
## day month duration campaign   
## Min. : 1.00 may :15164 Min. : 0.0 Min. : 1.000   
## 1st Qu.: 8.00 jul : 7601 1st Qu.: 103.0 1st Qu.: 1.000   
## Median :16.00 aug : 6880 Median : 180.0 Median : 2.000   
## Mean :15.82 jun : 5872 Mean : 258.7 Mean : 2.767   
## 3rd Qu.:21.00 nov : 4359 3rd Qu.: 320.0 3rd Qu.: 3.000   
## Max. :31.00 apr : 3225 Max. :4918.0 Max. :63.000   
## (Other): 6631   
## pdays previous poutcome y   
## Min. : -1.00 Min. : 0.0000 failure: 5391 no :43922   
## 1st Qu.: -1.00 1st Qu.: 0.0000 other : 2037 yes: 5810   
## Median : -1.00 Median : 0.0000 success: 1640   
## Mean : 40.16 Mean : 0.5769 unknown:40664   
## 3rd Qu.: -1.00 3rd Qu.: 0.0000   
## Max. :871.00 Max. :275.0000   
##

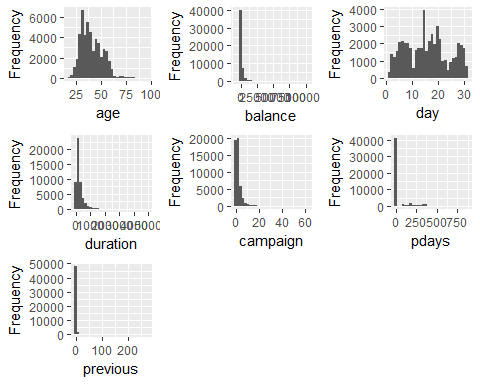
# Quick Data exploration (Missing values)  
library(DataExplorer)

## Warning: package 'DataExplorer' was built under R version 3.4.4

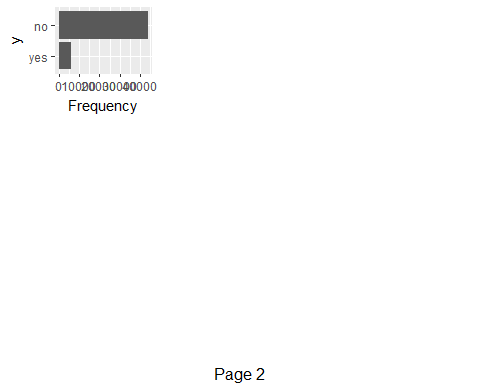
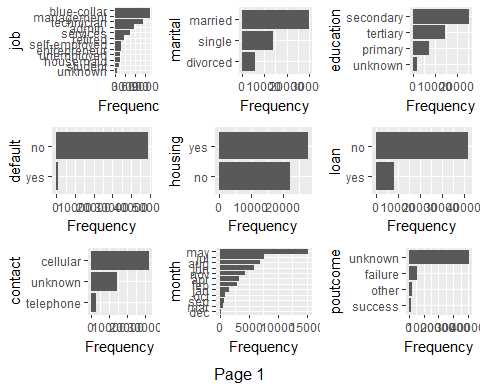
#Plot missing values  
plot\_missing(dataset)



#Continuous variables  
plot\_histogram(dataset)



#Categorical Variables  
plot\_bar(dataset)



#Encoding Categorical Data  
  
#Combining levels with similar proportions to avoid redundancy  
round(prop.table(table(dataset$job,dataset$y),1)\*100,1)

##   
## no yes  
## admin. 87.8 12.2  
## blue-collar 92.7 7.3  
## entrepreneur 91.7 8.3  
## housemaid 90.9 9.1  
## management 86.3 13.7  
## retired 77.1 22.9  
## self-employed 88.3 11.7  
## services 91.1 8.9  
## student 71.8 28.2  
## technician 89.0 11.0  
## unemployed 85.0 15.0  
## unknown 87.4 12.6

dataset$job = combineLevels(dataset$job, levs = c("self-employed","unknown","technician"), newLabel = "job\_1")

## The original levels admin. blue-collar entrepreneur housemaid management retired self-employed services student technician unemployed unknown   
## have been replaced by admin. blue-collar entrepreneur housemaid management retired services student unemployed job\_1

dataset$job = combineLevels(dataset$job, levs = c("services","housemaid","entrepreneur", "blue-collar"),   
 newLabel = "job\_2")

## The original levels admin. blue-collar entrepreneur housemaid management retired services student unemployed job\_1   
## have been replaced by admin. management retired student unemployed job\_1 job\_2

dataset$job = combineLevels(dataset$job, levs = c("management", "admin."), newLabel = "job\_3")

## The original levels admin. management retired student unemployed job\_1 job\_2   
## have been replaced by retired student unemployed job\_1 job\_2 job\_3

dataset$job = combineLevels(dataset$job, levs = c("student"), newLabel = "job\_4")

## The original levels retired student unemployed job\_1 job\_2 job\_3   
## have been replaced by retired unemployed job\_1 job\_2 job\_3 job\_4

dataset$job = combineLevels(dataset$job, levs = c("retired"), newLabel = "job\_5")

## The original levels retired unemployed job\_1 job\_2 job\_3 job\_4   
## have been replaced by unemployed job\_1 job\_2 job\_3 job\_4 job\_5

dataset$job = combineLevels(dataset$job, levs = c("unemployed"), newLabel = "job\_6")

## The original levels unemployed job\_1 job\_2 job\_3 job\_4 job\_5   
## have been replaced by job\_1 job\_2 job\_3 job\_4 job\_5 job\_6

dataset$job = factor(dataset$job,  
 levels = c("job\_1", "job\_2", "job\_3", "job\_4", "job\_5", "job\_6"),  
 labels = c(1, 2,3,4,5,6))  
  
table(dataset$marital)

##   
## divorced married single   
## 5735 30011 13986

dataset$marital = factor(dataset$marital,  
 levels = c("divorced", "married", "single"),  
 labels = c(1, 2, 3))  
  
table(dataset$education)

##   
## primary secondary tertiary unknown   
## 7529 25508 14651 2044

dataset$education = factor(dataset$education,  
 levels = c("primary", "secondary", "tertiary", "unknown"),  
 labels = c(1,2,3,4))  
  
dataset$default = factor(dataset$default,  
 levels = c("no", "yes"),  
 labels = c(0,1))  
  
dataset$housing = factor(dataset$housing,  
 levels = c("no", "yes"),  
 labels = c(0,1))  
  
dataset$loan = factor(dataset$loan,  
 levels = c("no", "yes"),  
 labels = c(0,1))  
  
table(dataset$contact)

##   
## cellular telephone unknown   
## 32181 3207 14344

dataset$contact = factor(dataset$contact,  
 levels = c("cellular", "telephone", "unknown"),  
 labels = c(1,2,3))   
  
table(dataset$month)

##   
## apr aug dec feb jan jul jun mar may nov oct sep   
## 3225 6880 234 2871 1551 7601 5872 526 15164 4359 818 631

#Combining levels with similar count to avoid redundancy  
round(prop.table(table(dataset$month,dataset$y),1)\*100,1)

##   
## no yes  
## apr 80.4 19.6  
## aug 88.9 11.1  
## dec 53.4 46.6  
## feb 83.3 16.7  
## jan 89.8 10.2  
## jul 90.9 9.1  
## jun 89.8 10.2  
## mar 48.9 51.1  
## may 93.3 6.7  
## nov 89.9 10.1  
## oct 56.0 44.0  
## sep 54.7 45.3

dataset$month = combineLevels(dataset$month, levs = c("dec", "sep", "oct"), newLabel = "month1")

## The original levels apr aug dec feb jan jul jun mar may nov oct sep   
## have been replaced by apr aug feb jan jul jun mar may nov month1

dataset$month = combineLevels(dataset$month, levs = c("aug","jan", "jun", "jul", "nov"), newLabel = "month2")

## The original levels apr aug feb jan jul jun mar may nov month1   
## have been replaced by apr feb mar may month1 month2

dataset$month = combineLevels(dataset$month, levs = c("mar"), newLabel = "month3")

## The original levels apr feb mar may month1 month2   
## have been replaced by apr feb may month1 month2 month3

dataset$month = combineLevels(dataset$month, levs = c("apr"), newLabel = "month4")

## The original levels apr feb may month1 month2 month3   
## have been replaced by feb may month1 month2 month3 month4

dataset$month = combineLevels(dataset$month, levs = c("feb"), newLabel = "month5")

## The original levels feb may month1 month2 month3 month4   
## have been replaced by may month1 month2 month3 month4 month5

dataset$month = combineLevels(dataset$month, levs = c("may"), newLabel = "month6")

## The original levels may month1 month2 month3 month4 month5   
## have been replaced by month1 month2 month3 month4 month5 month6

dataset$month = factor(dataset$month,  
 levels = c("month1", "month2", "month3", "month4", "month5", "month6"),  
 labels = c(1,2,3,4,5,6))  
  
  
dataset$poutcome = factor(dataset$poutcome,  
 levels = c("failure", "other", "success", "unknown"),  
 labels= c(1,2,3,4))  
  
dataset$y = factor(dataset$y,  
 levels = c("no", "yes"),  
 labels = c(0,1))  
  
# Duration is not counted to create a more realistic model  
dataset = dataset[-12]  
  
#Splitting dataset  
library(caTools)  
set.seed(123)  
split = sample.split(dataset$y, SplitRatio = 0.80)  
training\_set = subset(dataset, split == TRUE)  
test\_set = subset(dataset, split == FALSE)  
  
# Fitting Random Forest Classification to the Training set  
library(randomForest)

## randomForest 4.6-12

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

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':  
##   
## combine

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

set.seed(123)  
classifier = randomForest(x = training\_set[-16],  
 y = training\_set$y,  
 ntree = 100)  
  
# Predicting the Test set results  
y\_pred = predict(classifier, newdata = test\_set[-16])  
  
confusionMatrix(y\_pred, test\_set$y)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 8664 866  
## 1 120 296  
##   
## Accuracy : 0.9009   
## 95% CI : (0.8948, 0.9067)  
## No Information Rate : 0.8832   
## P-Value [Acc > NIR] : 1.056e-08   
##   
## Kappa : 0.3341   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.9863   
## Specificity : 0.2547   
## Pos Pred Value : 0.9091   
## Neg Pred Value : 0.7115   
## Prevalence : 0.8832   
## Detection Rate : 0.8711   
## Detection Prevalence : 0.9582   
## Balanced Accuracy : 0.6205   
##   
## 'Positive' Class : 0   
##

varImp(classifier)

## Overall  
## age 851.39329  
## job 309.68680  
## marital 169.78330  
## education 219.45916  
## default 18.12773  
## balance 987.63118  
## housing 151.50122  
## loan 70.73018  
## contact 146.60927  
## day 739.08880  
## month 597.57470  
## campaign 341.34502  
## pdays 447.83779  
## previous 222.05373  
## poutcome 538.18621

#Therefore, a supervised model has been created to predict term deposit product with an accuracy of 91%.