ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS
DEPARTMENT OF MANAGEMENT SCIENCE AND TECHNOLOGY
MSC IN BUSINESS ANALYTICS

ASSIGNMENT IN THE COURSE
«STATISTICS FOR BUSINESS ANALYTICS II»

PROJECT 2

ERASMIA KORNELATOU

ATHENS, 2020

Introduction

This data relates to telemarketing phone calls to sell long-term deposits. Within a campaign, the agents make phone calls to a list of clients to sell the product (outbound) or, if meanwhile the client calls the contact-center for any other reason, he is asked to subscribe the product (inbound). Thus, the result is a binary unsuccessful or successful contact.

This study considers real data collected from one of the retail bank, from May 2008 to June 2010, in a total of 39883 phone contacts. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

The purpose of our project is to be able to predict a successful contact (the client subscribes to the product). The rest variables are potential candidates for examining the variable under study. We are going to employ 3 methods and compare them.

About the Data:

Input variables:
bank client data:
1 - age (numeric)
2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
4 - education (categorical: basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
5 - default: has credit in default? (categorical: 'no','yes','unknown')
6 - housing: has housing loan? (categorical: 'no','yes','unknown')
7 - loan: has personal loan? (categorical: 'no','yes','unknown')

```
# related with the last contact of the current campaign:
8 - contact: contact communication type (categorical:
'cellular','telephone')
9 - month: last contact month of year (categorical:
'jan', 'feb', 'mar', ..., 'nov', 'dec')
10 - day_of_week: last contact day of the week
(categorical: 'mon','tue','wed','thu','fri')
11 - duration: last contact duration, in seconds
(numeric).


# other attributes:
12 - campaign: number of contacts performed during this
campaign and for this client (numeric, includes last
contact)
13 - pdays: number of days that passed by after the
client was last contacted from a previous campaign
(numeric; 999 means client was not previously contacted)
14 - previous: number of contacts performed before this
campaign and for this client (numeric)
15 - poutcome: outcome of the previous marketing campaign
(categorical: 'failure','nonexistent','success')
# social and economic context attributes
16 - emp.var.rate: employment variation rate - quarterly
indicator (numeric)
17 - cons.price.idx: consumer price index - monthly
indicator (numeric)
18 - cons.conf.idx: consumer confidence index - monthly
indicator (numeric)
19 - euribor3m: euribor 3 month rate - daily indicator
(numeric)
20 - nr.employed: number of employees - quarterly
indicator (numeric)

Output variable (desired target):
21 - SUBSCRIBED - has the client subscribed a term
deposit? (binary: 'yes','no')
```

Descriptive analysis and exploratory data analysis

In the first place we have a look at the structure of the variables.

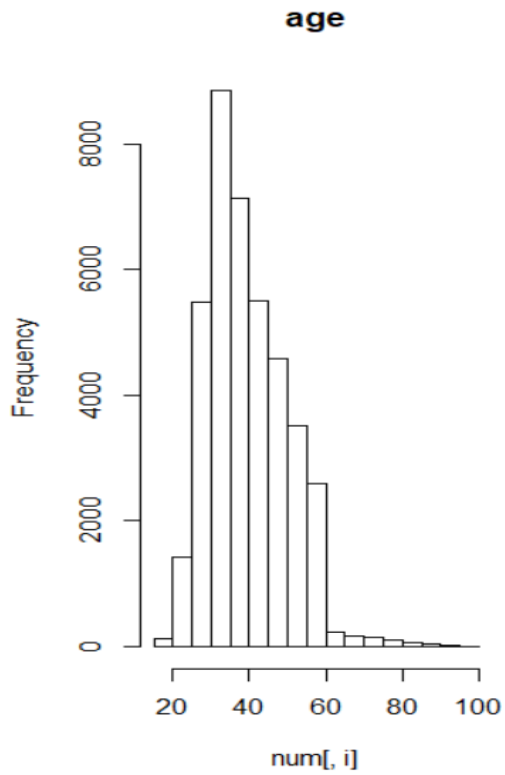
```
> #having a look at the structure of the data
> str(calls)
'data.frame':   39883 obs. of  21 variables:
 $ age            : int  56 57 37 40 56 45 59 41 24 25 ...
 $ job            : Factor w/ 12 levels "admin.","blue-collar",..: 4 8 8 1 8 8 1$
 $ marital        : Factor w/ 4 levels "divorced","married",..: 2 2 2 2 2 2 2 2 $
 $ education      : Factor w/ 8 levels "basic.4y","basic.6y",..: 1 4 4 2 4 3 6 8$
 $ 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$
 $ month          : Factor w/ 10 levels "apr","aug","dec",..: 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$
 $ 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 -$
 $ euribor3m      : num  4.86 4.86 4.86 4.86 4.86 ...
 $ nr.employed    : num  5191 5191 5191 5191 5191 ...
 $ SUBSCRIBED     : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```

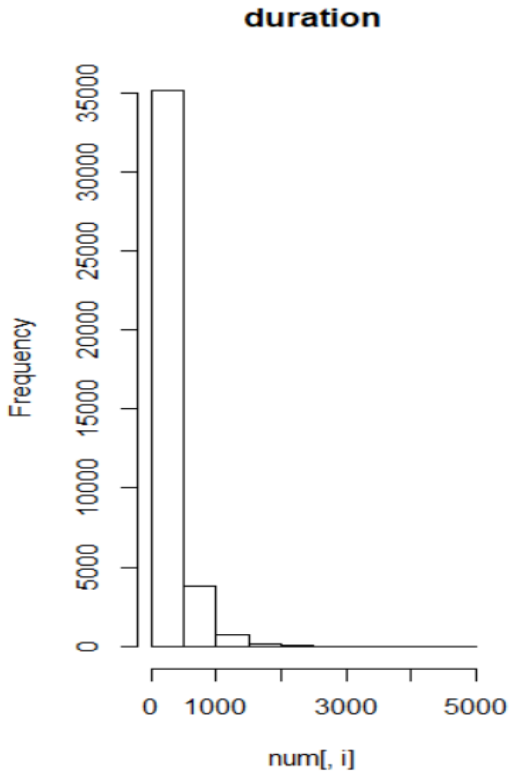
As we have already red in the decription, “job”, “marital”, “education”, “default”, “housing”, “loan”, “contact”, “month”, “day_of_week”, “poutcome” and “SUBSCRIBED” variables must be categorical variables whereas “age”, “duration”, “campaign”, “pdays”, “previous”, “emp.var.rate”, “cons.price.idx”, “cons.conf.idx”, “euribor3m”, “nr.employed” should be numeric

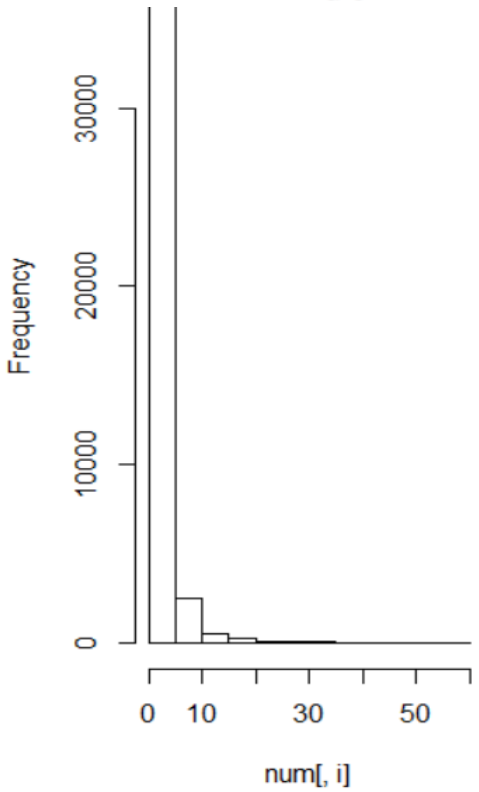
```
Classes 'tbl_df', 'tbl' and 'data.frame':       39883 obs. of  21 variables:
 $ age            : num  56 57 37 40 56 45 59 41 24 25 ...
 $ job            : Factor w/ 12 levels "admin.","blue-collar",..: 4 8 8 1 8 8 1 2 10 8 ...
 $ marital        : Factor w/ 4 levels "divorced","married",..: 2 2 2 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       : num  261 149 226 151 307 198 139 217 380 50 ...
 $ campaign       : num  1 1 1 1 1 1 1 1 1 1 ...
 $ pdays          : num  999 999 999 999 999 999 999 999 999 999 ...
 $ previous       : num  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 ...
 $ SUBSCRIBED     : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```

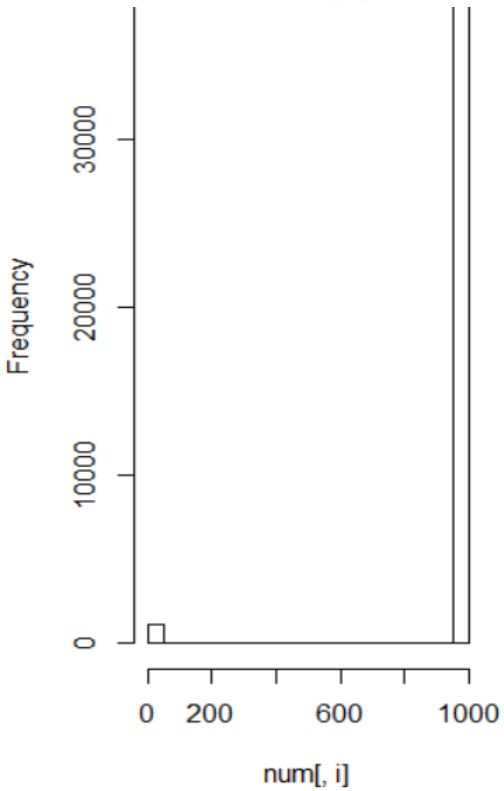
variables. After converting them to the right types, the types of the variables are as above:

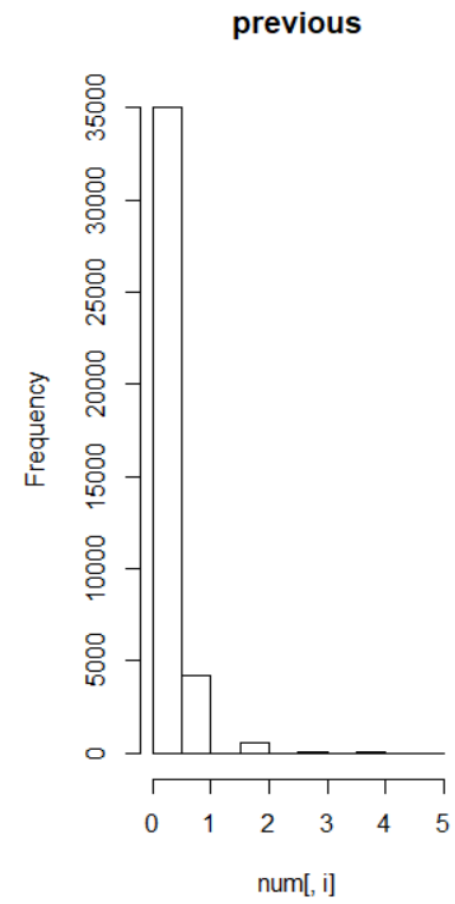
Analysis for numerical variables

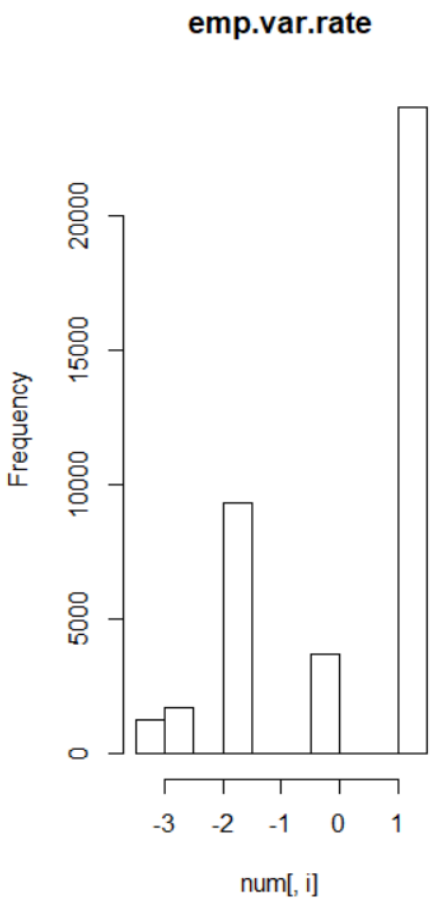


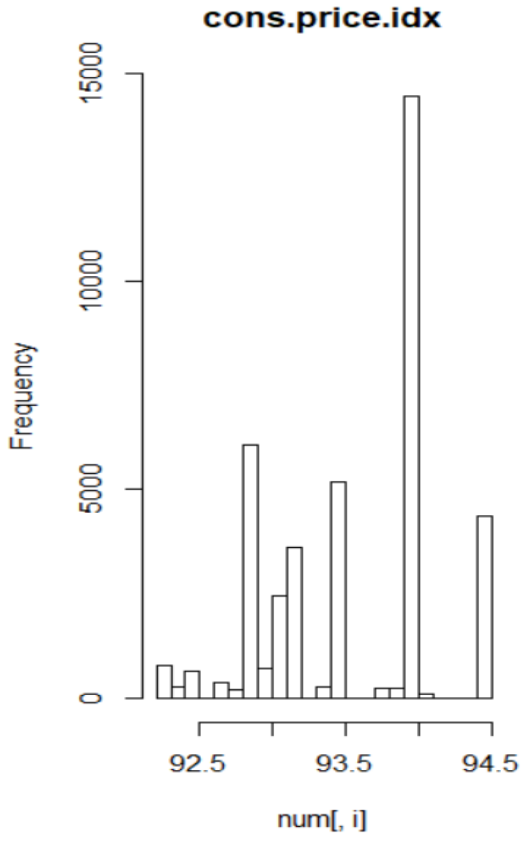


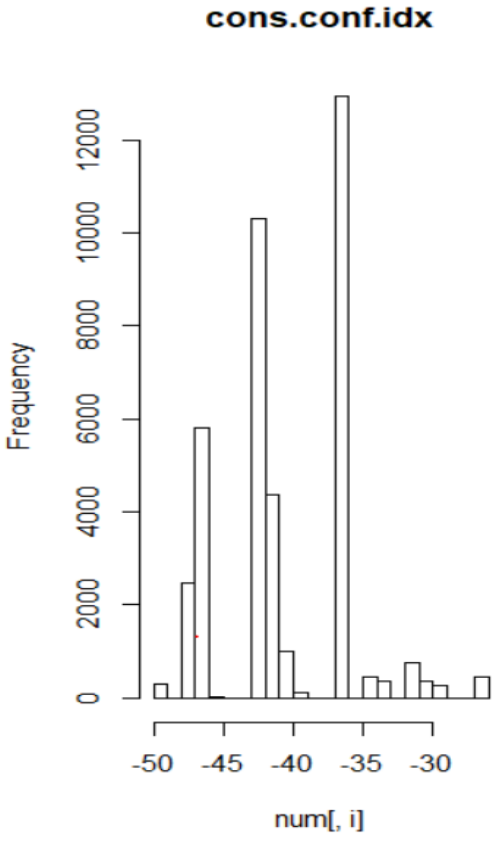


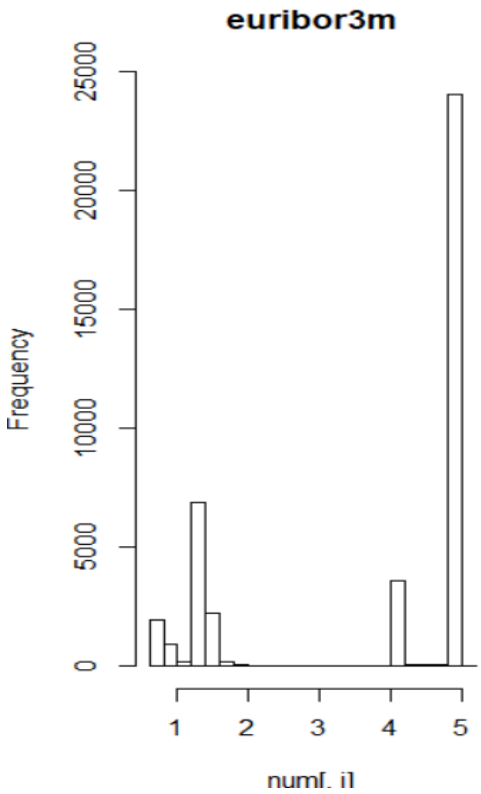


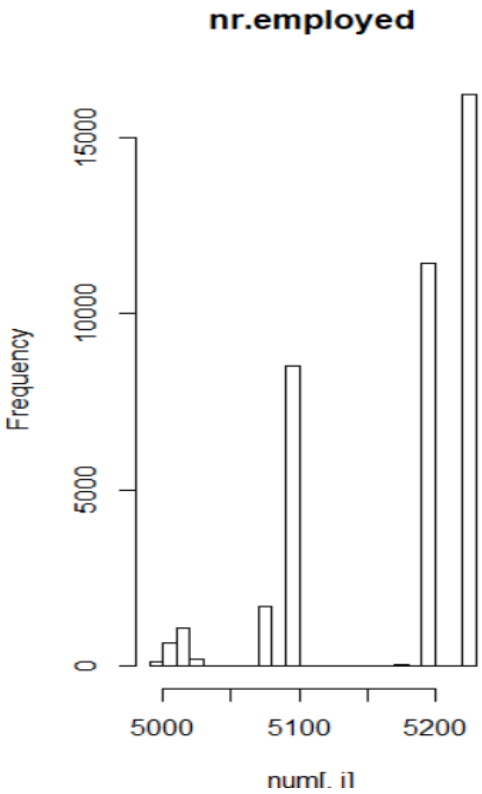

previous
Frequency
0
5000
10000
15000
20000
25000
30000
35000
0
1
2
3
4
5
num[, i]

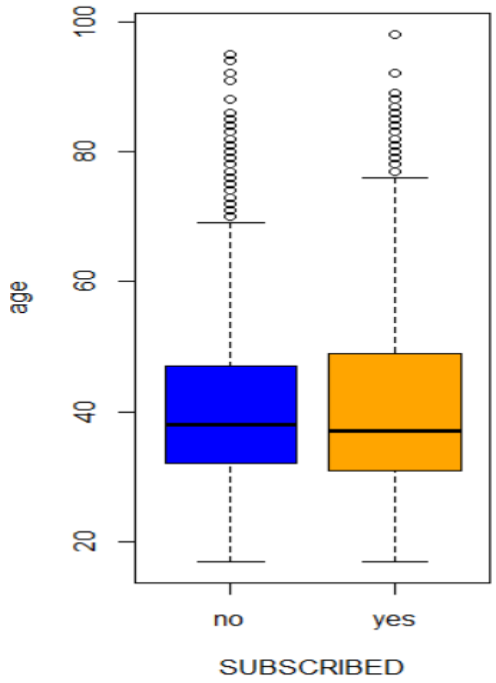

emp.var.rate
Frequency
0
5000
10000
15000
20000
-3
-2
-1
0
1
num[, i]

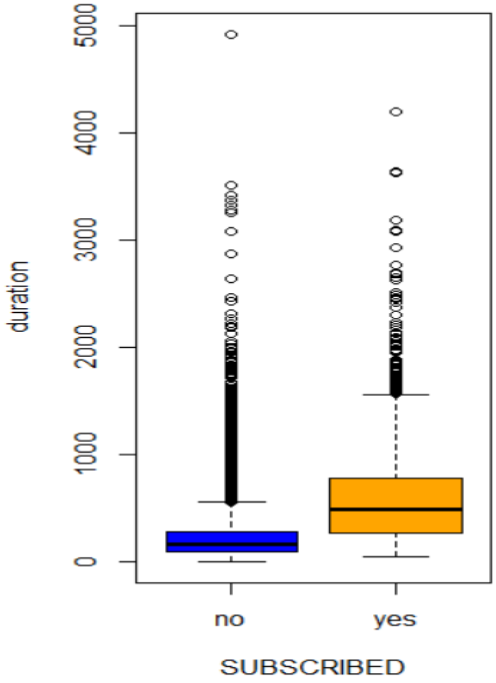

cons.price.idx
Frequency
0
5000
10000
15000
92.5
93.5
94.5
num[, i]

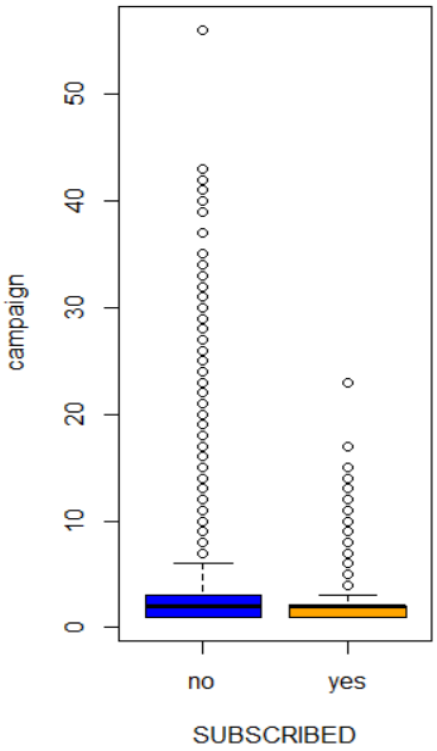

cons.conf.idx
Frequency
0
2000
4000
6000
8000
10000
12000
-50
-45
-40
-35
-30
num[, i]

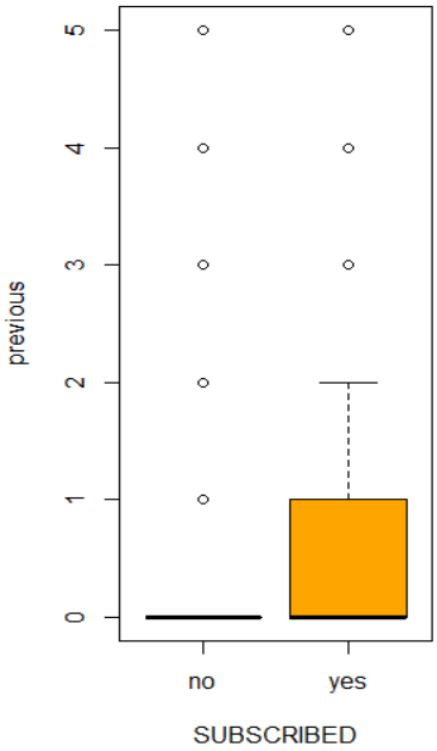

euribor3m
Frequency
0
5000
10000
15000
20000
25000
1
2
3
4
5
num[, i]


nr.employed
Frequency
0
5000
10000
15000
5000
5100
5200
num[, i]

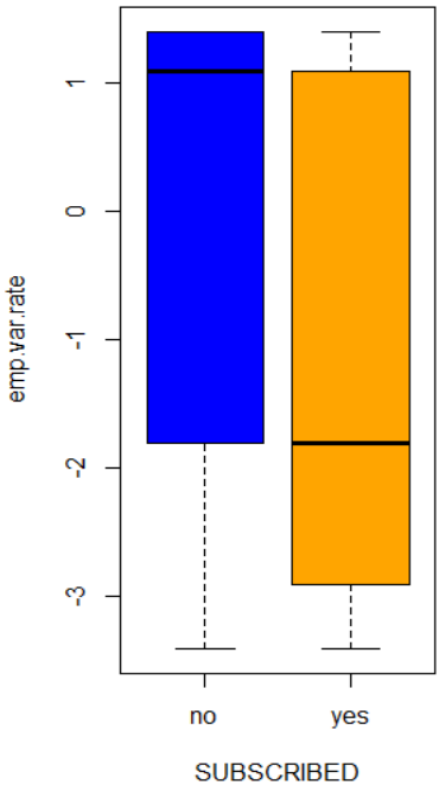

age
20
40
60
80
100
no
yes
SUBSCRIBED

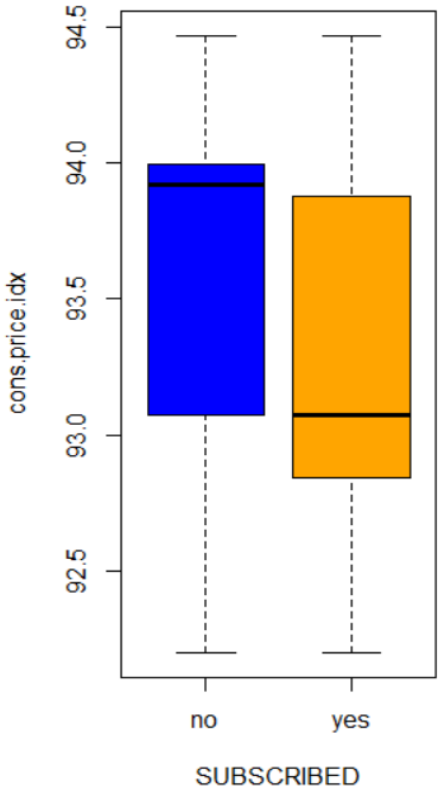

duration
0
1000
2000
3000
4000
5000
no
yes
SUBSCRIBED

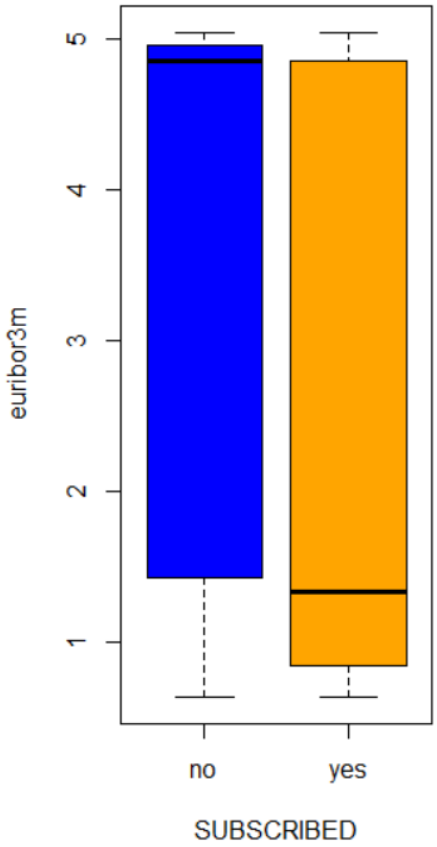


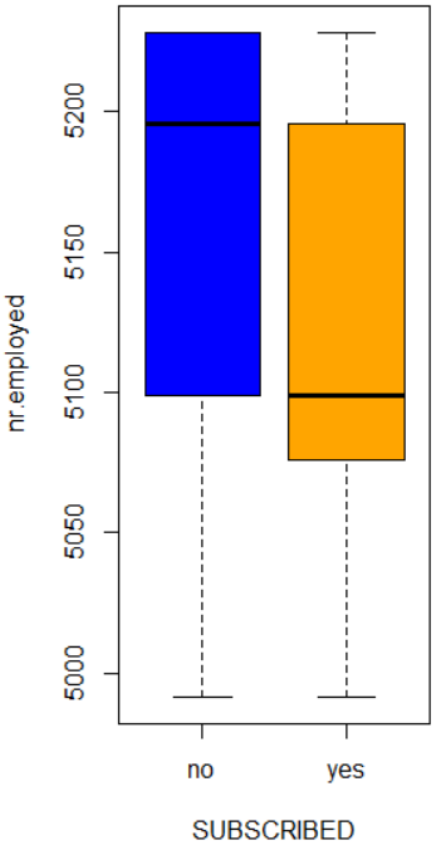


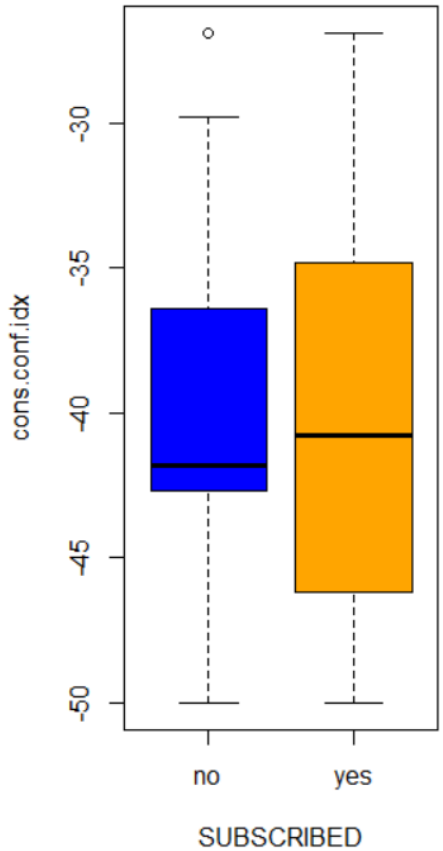
Data Preprocessing

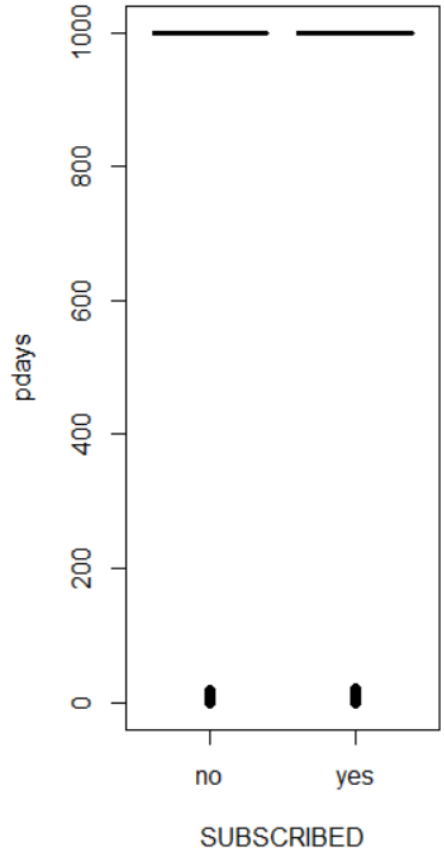




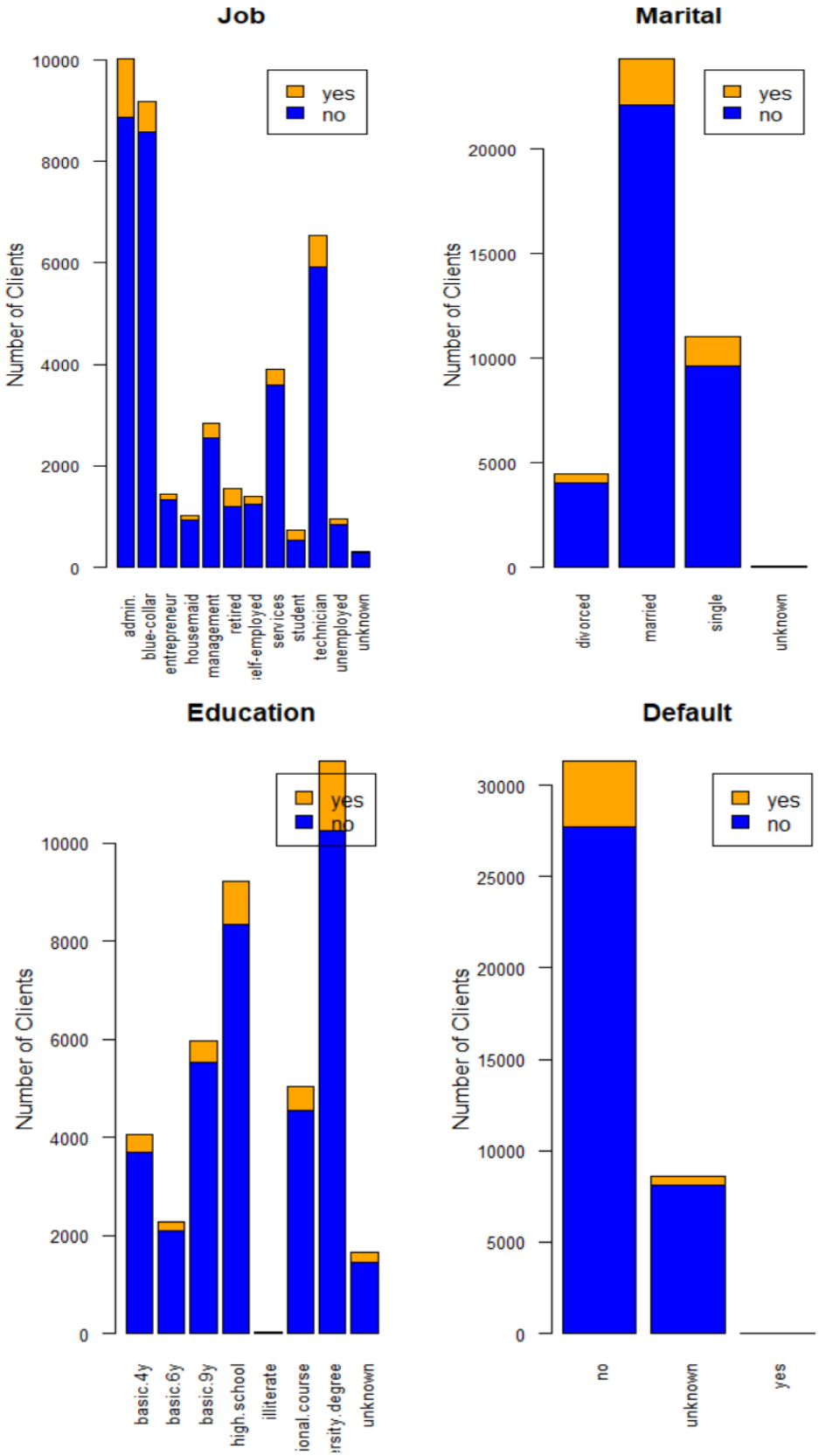

euribor3m
1
2
3
4
5
no
yes
SUBSCRIBED

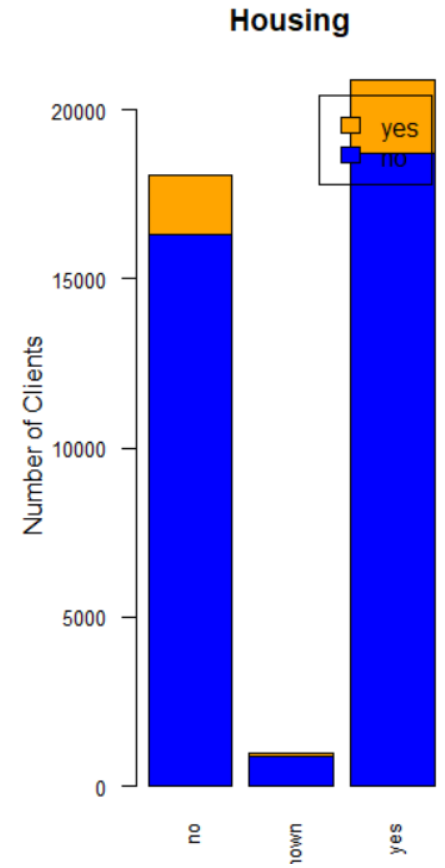

nr.employed
5000
5050
5100
5150
5200
no
yes
SUBSCRIBED

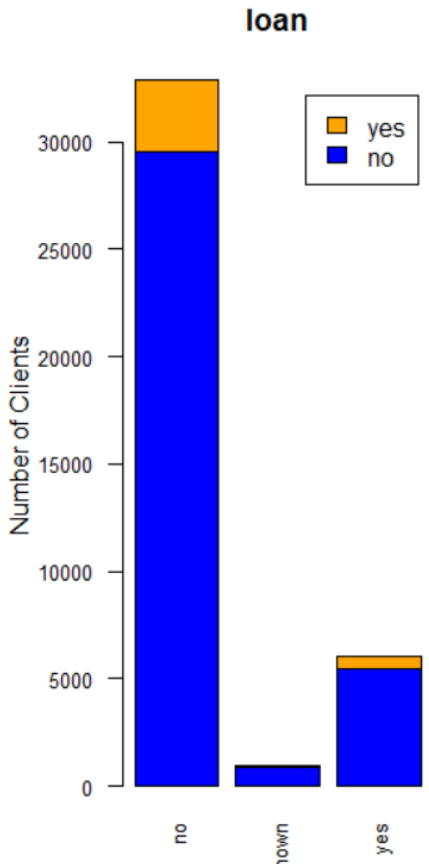

cons.conf.idx
-50
-45
-40
-35
-30
no
yes
SUBSCRIBED

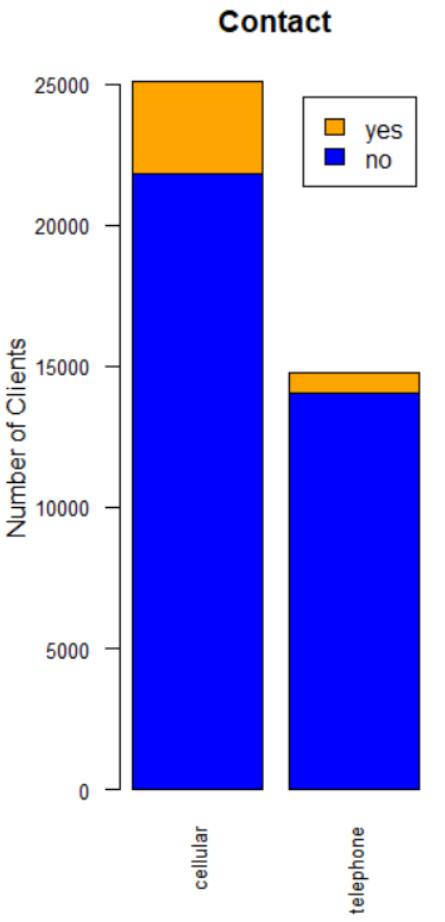

pdays
0
200
400
600
800
1000
no
yes
SUBSCRIBED

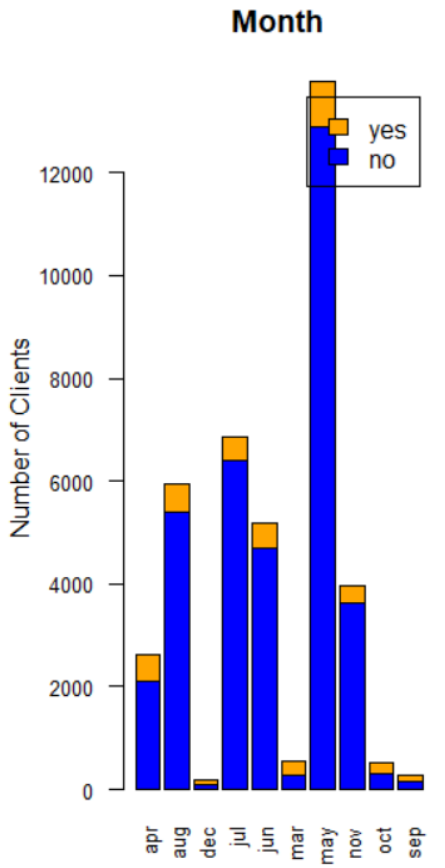
Analysis for factorial variables

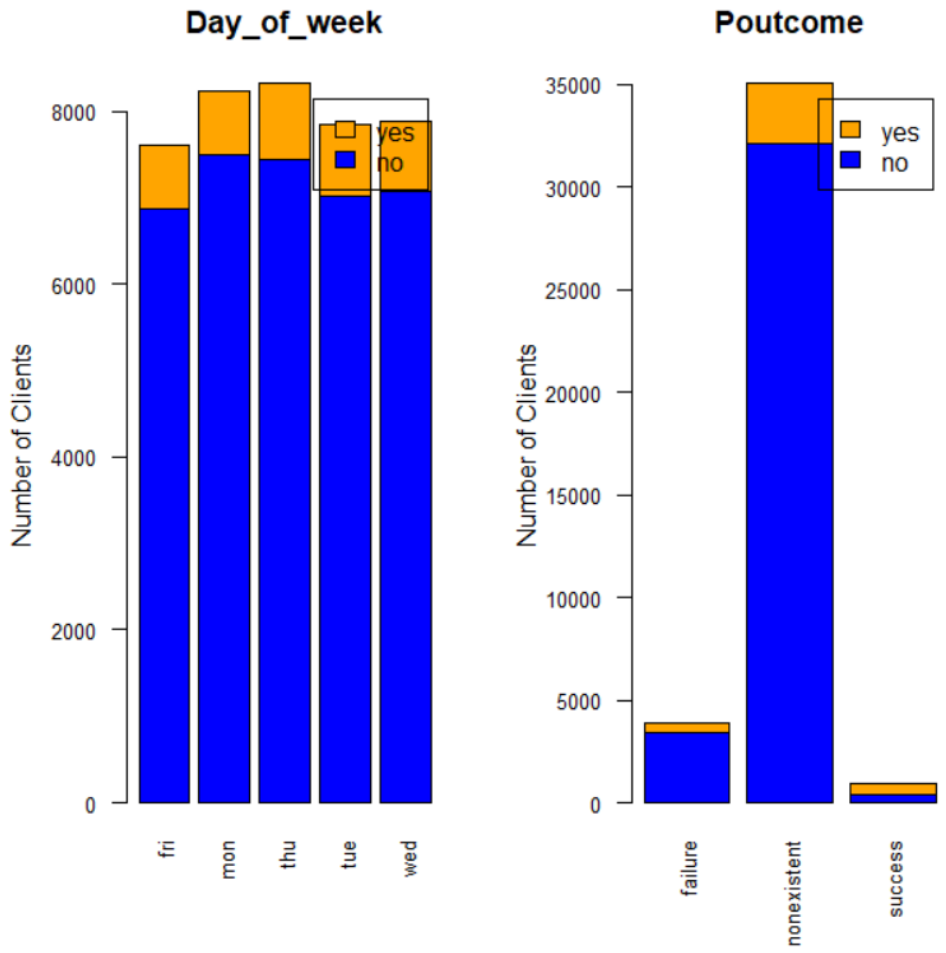



Housing
Number of Clients
20000
15000
10000
5000
0
yes
no
no
yes


loan
Number of Clients
30000
25000
20000
15000
10000
5000
0
yes
no
no
yes


Contact
Number of Clients
25000
20000
15000
10000
5000
0
yes
no
cellular
telephone


Month
Number of Clients
12000
10000
8000
6000
4000
2000
0
yes
no
apr
aug
dec
jul
jun
mar
may
nov
oct
sep



From the "Job" barplot, we can see that the customers who have a job of admin have the highest rate of subscribing a term deposit, but they are also the highest when it comes to not subscribing. This is simply because we have more customers working as admin than any other profession.

Based on the "Marital" barplot, the majority of the customers are married.

As for the "housing" plot, we can see that the majority of the customers have a housing loan.

For most of the customers, the previous marketing campaign outcome does not exists. It means that most of the customers are new customers who have not been contacted earlier. Also one thing to note here that, for the customers who had a successful outcome from the previous campaign, majority of those customers did subscribe for a term deposit. As it has the class distribution of 2.2% for positive class, and 1.2%

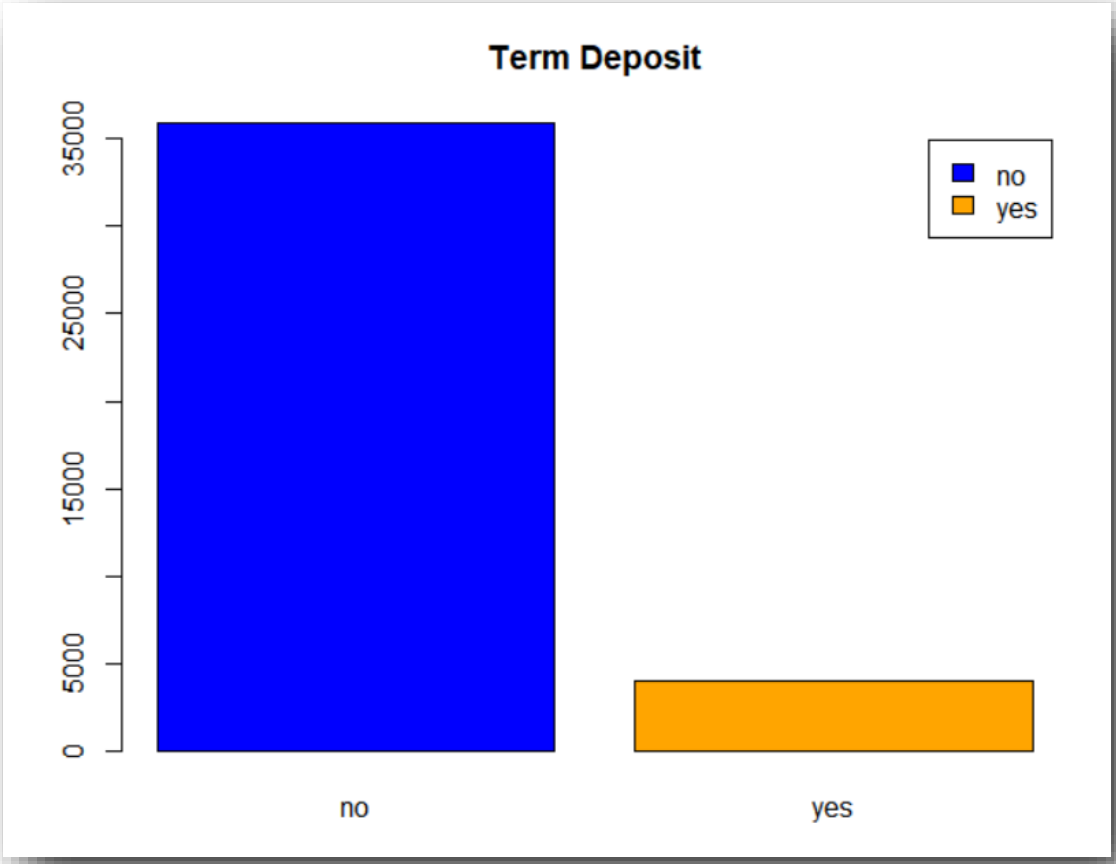
for negative class. From this, we can make an assumption, that this feature may hold some value in predicting the target variable. specially poutcome_success category.

From the “age” boxplot we know that for both the customers that subscibed or didn’t subscribe a term deposit, has a median age of around 40. And the boxplot for both the classes overlap quite a lot, which means that age isn’t necessarily a good indicator for which customer will subscribe and which customer will not.

The month and day_of_week seems not to have an important role in our prediction .

Imbalanced Dataset

After calculating the number of “yes” and “no” in the variable “SUBSCRIBED” we come to the conclusion that our dataset is imbalanced as there are:



Yes: 3987 rows
No: 35896 rows

Finding only the accuracy of our model is not a very good measure of evaluating our model as it can better predict the "No" values than "Yes" values which is our purpose.

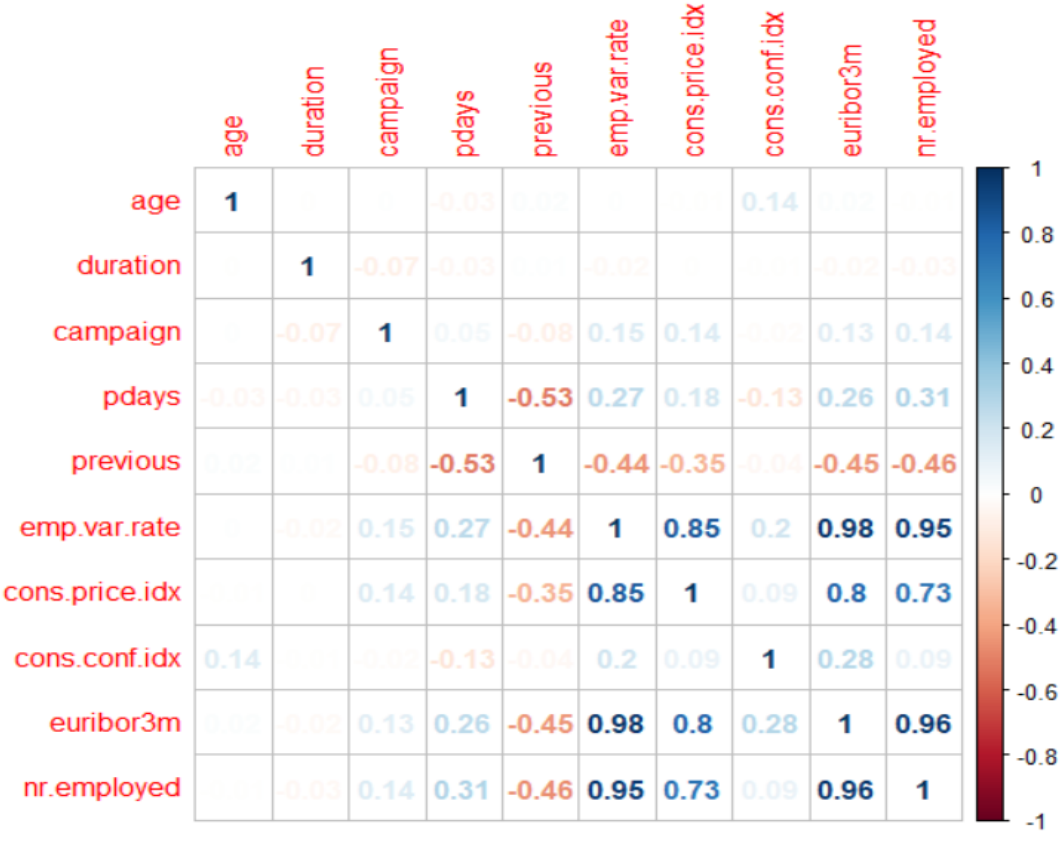
We should also check if there are any duplicate lines. Before removing duplicates the lines where 39883 and then there were 39871, so the duplicate lines were 12.

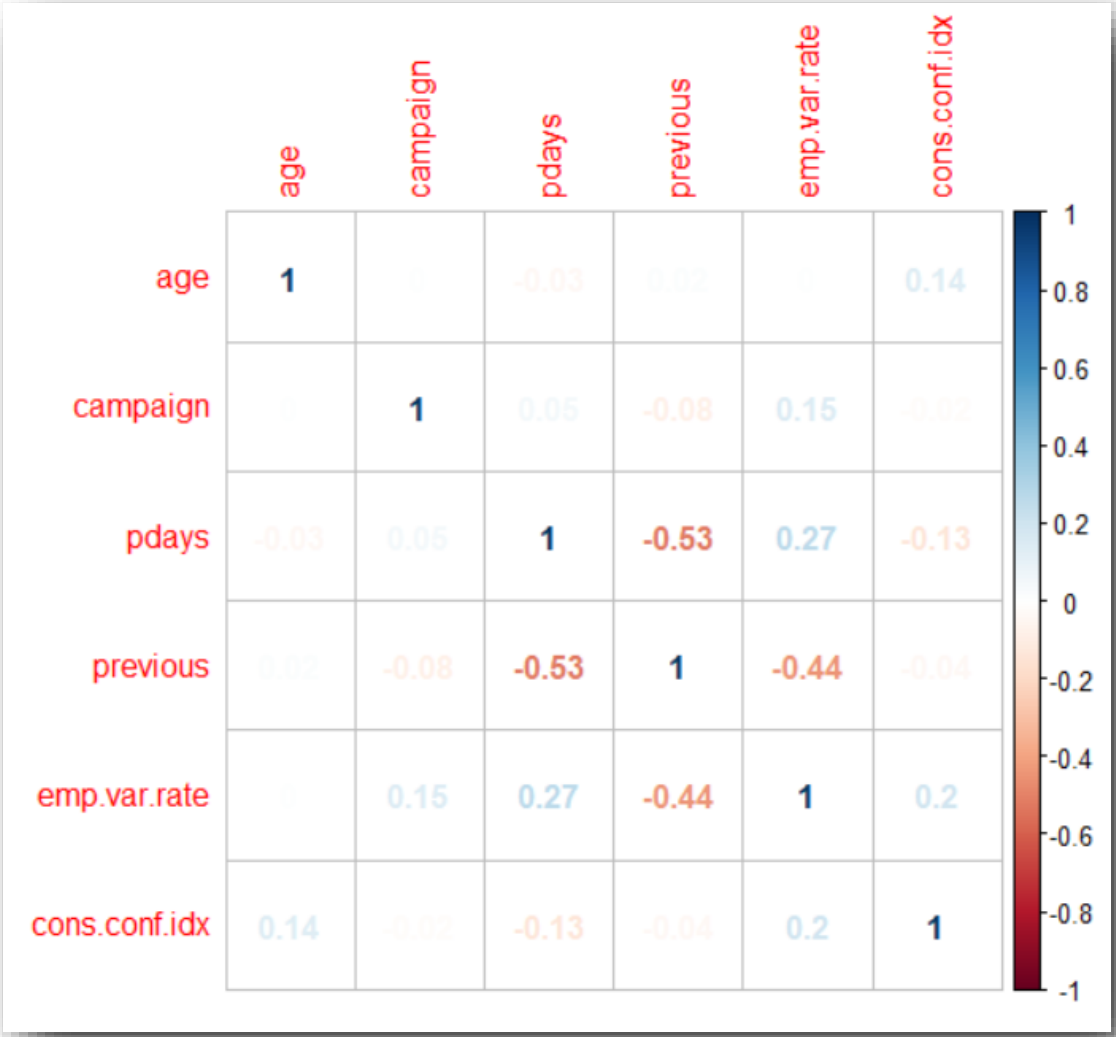
Our next step is to standardize all the numeric variables in order to increase the predictive power of our impending model.

Vizualization of bivariate assosiations

When predictors are correlated with other predictors in the data, then there is an issue of collinearity. Using highly correlated predictors in some modeling techniques can result in highly unstable models and produce poor predictive performance

At first glance, we notice that there is a multicollinearity issue in our data. Specifically, there is high correlation between "emp.var.rate" ,"cons.price.idx", "cons.conf.idx" and "euribor3m" and "nr.employed" variables. From these variables, we are going to keep only "emp.var.rate" .After removing these variables we have faced the collinearity issue.





We will also remove "duration" variable as it highly affects the output target . If duration=0 then y='no' and for this reason it should be removed as we would like to have a realistic predictive model.

Train

We are going to split our data into 70% for the train dataset and 30% for the test set.

Predictive Models

Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Confusion Matrix

```
Prediction    no   yes
       no  10520   935
       yes   245   260

               Accuracy : 0.9013
                 95% CI : (0.8959, 0.9066)
    No Information Rate : 0.9001
    P-Value [Acc > NIR] : 0.3304

                  Kappa : 0.2621

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9772
            Specificity : 0.2176
         Pos Pred Value : 0.9184
         Neg Pred Value : 0.5149
             Prevalence : 0.9001
         Detection Rate : 0.8796
   Detection Prevalence : 0.9578
      Balanced Accuracy : 0.5974

       'Positive' Class : no
```

Based on our results, Random Forest , gives :
Accuracy: 0.904
Sensitivity: 0.9772
Specificity: 0.2176
Balanced Accuracy: 0.5974

K-NN Model

In pattern recognition, the *k*-nearest neighbors algorithm (*k*-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the *k* closest training examples in the feature space. In *k-NN classification*, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (*k* is a positive integer, typically small). If *k* = 1, then the object is simply assigned to the class of that single nearest neighbor.

```
Confusion Matrix and Statistics

          Reference
Prediction    no   yes
       no  10577   989
       yes   188   206

               Accuracy : 0.9016
                 95% CI : (0.8961, 0.9069)
    No Information Rate : 0.9001
    P-Value [Acc > NIR] : 0.2978

                  Kappa : 0.2207

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9825
            Specificity : 0.1724
         Pos Pred Value : 0.9145
         Neg Pred Value : 0.5228
             Prevalence : 0.9001
         Detection Rate : 0.8844
   Detection Prevalence : 0.9671
      Balanced Accuracy : 0.5775
```

Based on our results, KNN Model , gives :
Accuracy: 0.9016
Sensitivity: 0.9825
Specificity: 0.1724

CART TREE

Decision tree learning is one of the predictive modeling approaches used in statistics, data mining and machine learning. It uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees.

```
Confusion Matrix and Statistics

          Reference
Prediction    no   yes
       no  10703  1041
       yes    62   154

               Accuracy : 0.9078
                 95% CI : (0.9024, 0.9129)
    No Information Rate : 0.9001
    P-Value [Acc > NIR] : 0.002412

                  Kappa : 0.1936

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9942
            Specificity : 0.1289
         Pos Pred Value : 0.9114
         Neg Pred Value : 0.7130
             Prevalence : 0.9001
         Detection Rate : 0.8949
   Detection Prevalence : 0.9819
      Balanced Accuracy : 0.5616
```

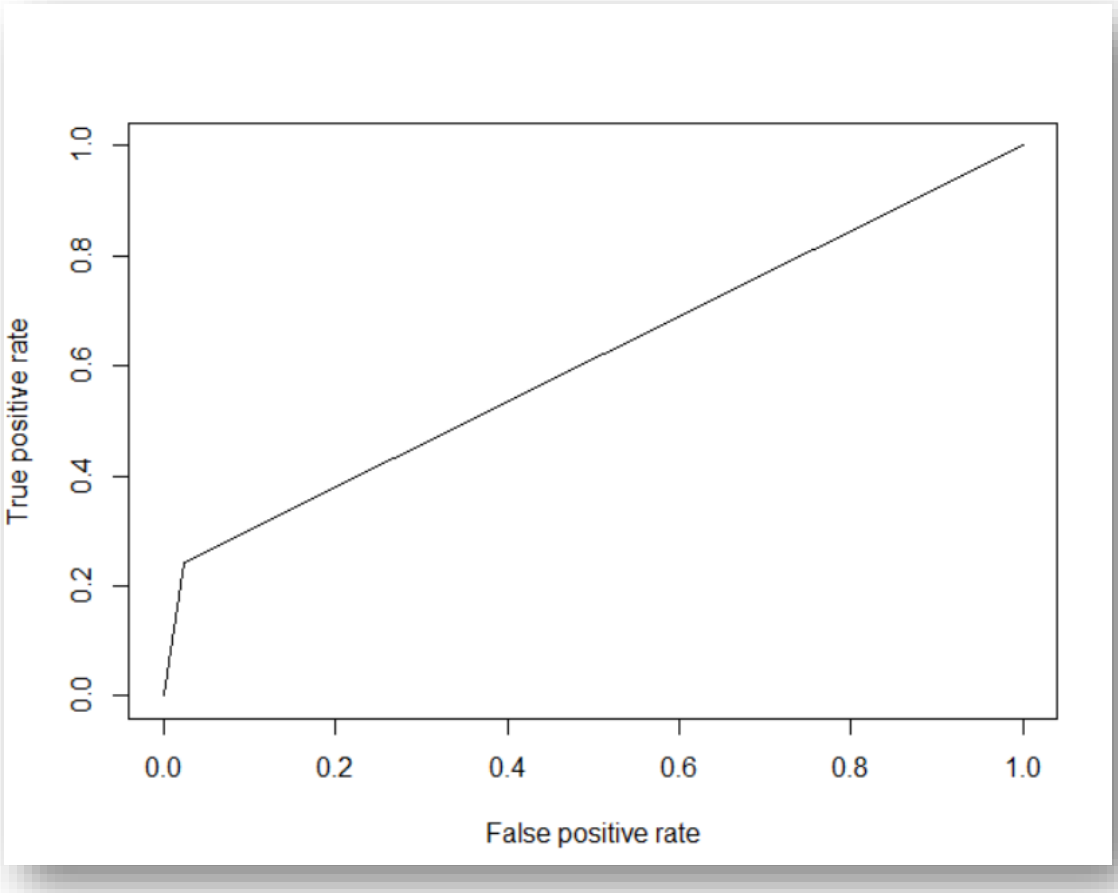
Based on our results, Cart Tree , gives :
Accuracy: 0.9078
Sensitivity: 0.9942
Specificity: 0.1289

Sensitivity is the accuracy rate for only the positive class, in this case customers who purchased the term deposit while specificity is the accuracy rate for the negative class - customers who did not purchase a term deposit. A perfect model that completely separates the two classes would have 100% sensitivity and specificity.

Roc Curve - AUC

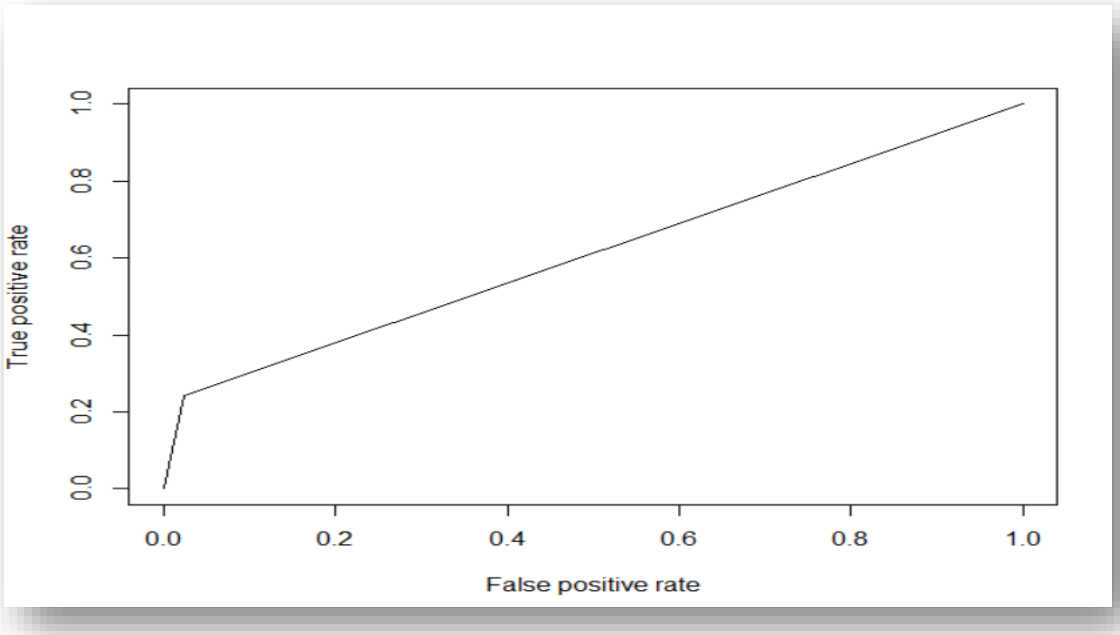
As our dataset is imbalanced , the best measure to characterize models is ROC Curve since it is a function of sensitivity and specificity and it is insensitive to disparities in the class proportions. Now we are going to split our dataset into train and test for 10 times for each model and find the average AUC value resulting from Roc Curve.

Random Forest

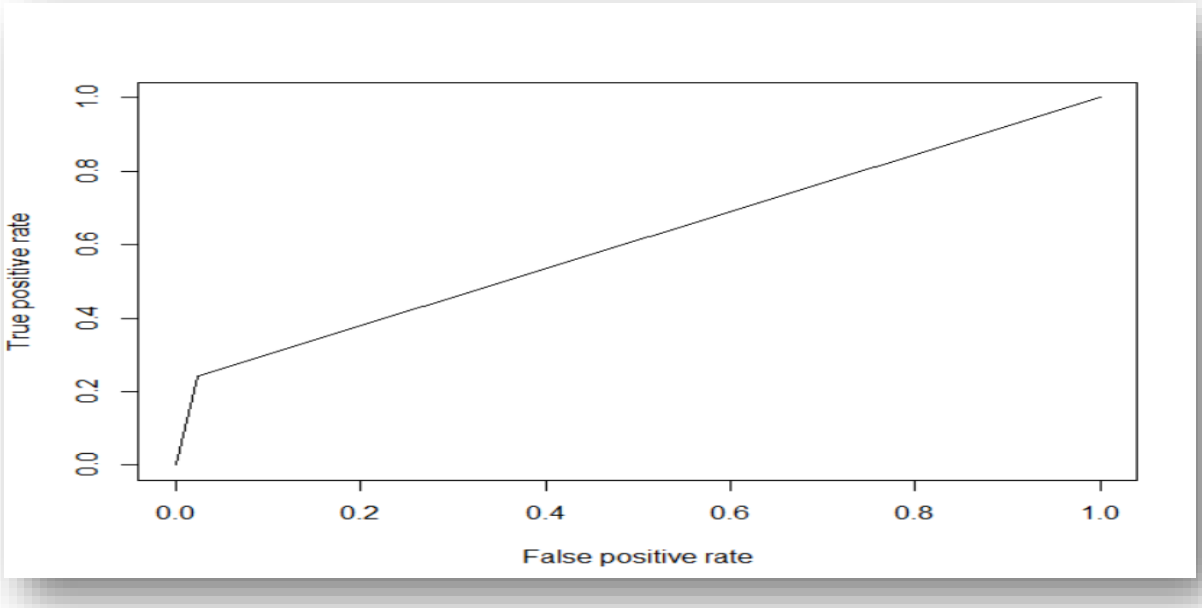


The average AUC is equal to 60.43 %

K-NN Model



The average AUC is equal to 59.08 %Cart Tree



The average AUC is equal to 59.71 %.

Conclusion

According to the AUC measure, the best model for predicting long-term deposits is Random Forest.