Course Project

Topic 1: Analyzing blood pressure using ANOVA

Topic 2: Predicting the Client's deposit subscription using Bank's Marketing Data with Logistic Regression

Group members:

Anisha Ganeshkumar

Dali Zhou

Topic 1

I. Abstract

A) Introduction

Blood pressure usually refers to the pressure in large arteries of the systemic circulation. It is a vital sign of the health of human's heart. A too low blood pressure is called hypotension while a too high blood pressure is called hypertension. Both of them can cause heart diseases such as thrombus and chest pain. Several factors including age, gender and serum cholesterol may have influence on blood pressure.

B) Objective Statement

We are analyzing the factors that influence resting blood pressure in this study. The hypothesis is that people in different age and gender may have different resting blood pressure. Resting blood pressure is also related to heart rate and serum cholesterol. We use the data and the tool of ANOVA and we want to find out how do these factors influence resting blood pressure.

C) Data description

The database is from a research done by UCI. It contains more than 300 observations and about 10 variables. (*https://www.kaggle.com/ronitf/heart-disease-uci/data*)

The variables used in this study are: age, gender, serum cholesterol and maximum heart rate.

II. Data Description

A) Age

The database contains people aged from 29 to 77. The distribution of age and the relationship with blood pressure are in Figure 1.

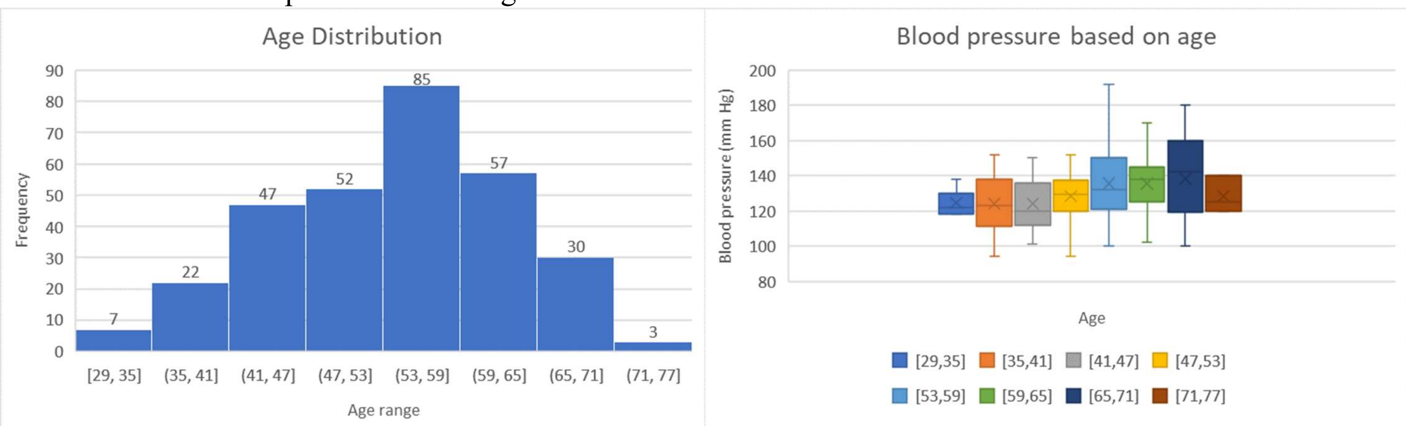


Figure 1. Relationship between age and blood pressure

B) Gender

The gender distribution and the relationship with blood pressure are in figure 2.

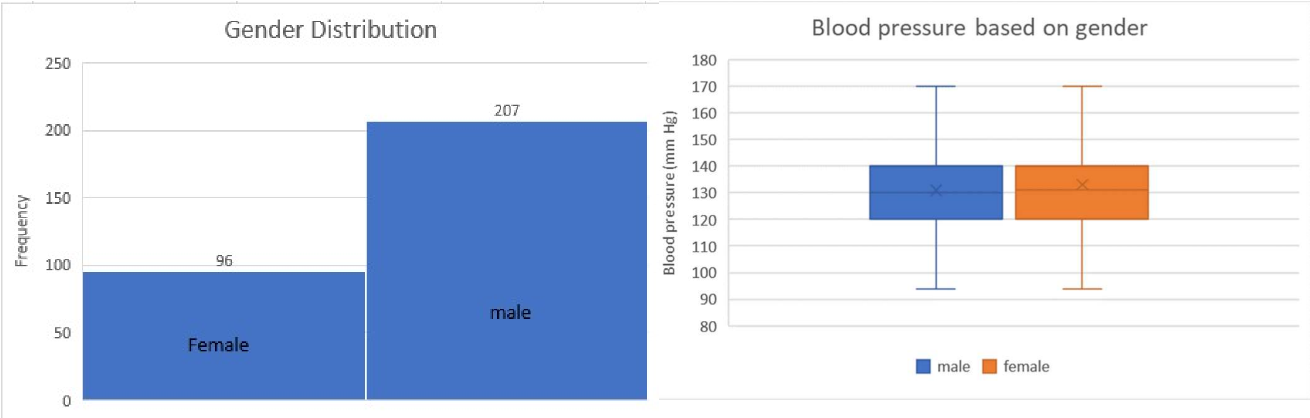


Figure 2. Relationship between gender and blood pressure

C) Serum cholesterol

The serum cholesterol is measured in mg/dl. We divide the data in 6 groups according to different levels. The results are in figure 3.

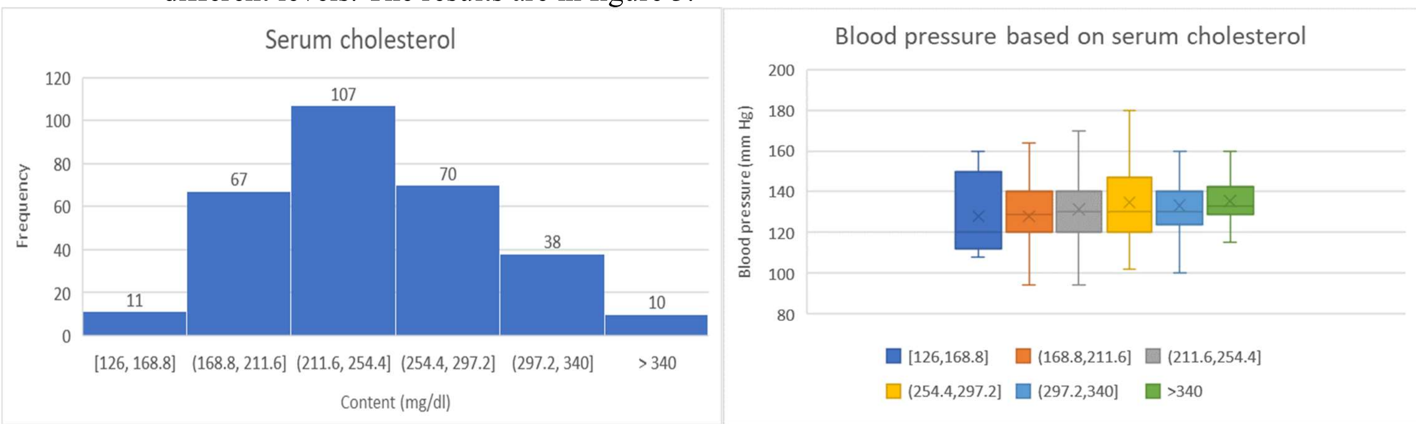


Figure 3. Relationship between serum cholesterol and blood pressure

D) Maximum heart rate

The maximum heart rate is measured in times/minute. We divide the data in 5 groups according to different levels. The results are in figure 4.

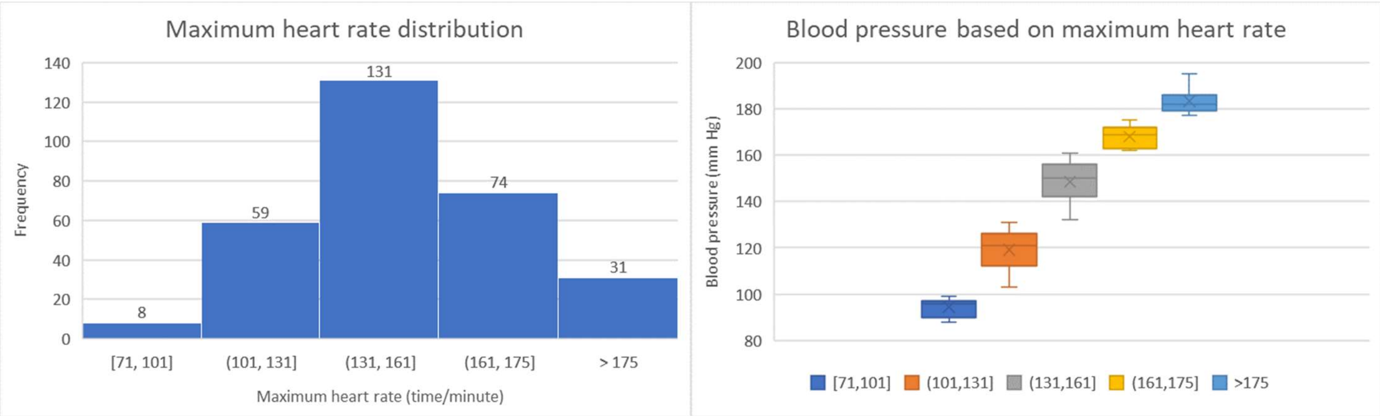


Figure 4. Relationship between maximum heart rate and blood pressure

III. Data Analysis

A) Age

Test of the effect on age.
H_0: The resting blood pressure does not differ from different age groups.
H_1: The resting blood pressure differs from different age groups.

SUMMARY						
Group	n	sum	ave	var		
[29,35]	7	872	124.5714	54.28571		
[35,41]	22	2728	124	206.5714		
[41,47]	47	5828	124	180.087		
[47,53]	52	6669	128.25	174.3873		
[53,59]	85	11517	135.4941	421.3244		
[59,65]	57	7732	135.6491	224.1961		
[65,71]	30	4151	138.3667	494.5851		
[71,77]	3	385	128.3333	108.3333		
ANOVA						
	SS	df	MS	F	P-value	F crit
SSA	8543.782	7	1220.54	4.26877	0.000164	2.040681
SSE	84347.33	295	285.9231			
SST	92891.11	302				

P-value is less than the level of significance (0.05), so we reject H_0.
We have enough evidence to conclude that the effect of age on resting blood pressure is significant.

B) Gender

Test of the effect on gender
H_0: The resting blood pressure does not differ from different gender.

H_1: The resting blood pressure differs from different gender.

SUMMARY

Group	n	sum	ave	var
male	207	27106	130.9469	277.4972
female	96	12776	133.0833	372.9193

ANOVA

	SS	df	MS	F	P-value	F crit
SSA	299.3601	1	299.3601	0.973169	0.324683	3.872538
SSE	92591.75	301	307.6138			
SST	92891.11	302				

P-value is larger than the level of significance (0.05), so we do not reject H_0.
We have enough evidence to conclude that the effect of gender on resting blood pressure is insignificant.

C) Serum cholesterol

Test of the effect on serum cholesterol.
H_0: The resting blood pressure does not differ from different serum cholesterol content groups.
H_1: The resting blood pressure differs from different serum cholesterol content groups.

SUMMARY

Group	n	sum	ave	var
[126,168.8]	11	1408	128	349.6
168.8,211.6	67	8572	127.9403	222.5115
211.6,254.4	107	14046	131.271	292.7277
254.4,297.2	70	9439	134.8429	404.1344
(297.2,340]	38	5062	133.2105	329.9004
>340	10	1355	135.5	172.2778

ANOVA

	SS	df	MS	F	P-value	F crit
SSA	2038.12	5	407.6241	1.33253	0.25029	2.244392
SSE	90852.99	297	305.9023			
SST	92891.11	302				

P-value is larger than the level of significance (0.05), so we do not reject H_0.
We have enough evidence to conclude that the effect of serum cholesterol content on resting blood pressure is insignificant.

D) Maximum heart rate

Test of the effect on maximum heart rate.
H_0: The resting blood pressure does not differ from different maximum heart rate groups.
H_1: The resting blood pressure differs from different maximum heart rate groups.

SUMMARY						
Group	n	sum	ave	var		
[71,101]	7	661	94.42857	15.61905		
(101,131]	59	7031	119.1695	69.55698		
(131,161]	131	19459	148.542	69.38861		
(161,175]	74	12438	168.0811	19.3358		
>175	31	5683	183.3226	36.29247		
ANOVA						
	SS	df	MS	F	P-value	F crit
SSA	136588.6	4	34147.14	648.0807	2.5E-145	2.402043
SSE	15648.83	297	52.68965			
SST	152237.4	301				

P-value is less than the level of significance (0.05), so we reject H_0.
We have enough evidence to conclude that the effect of maximum heart rate on resting blood pressure is significant.

IV. Discussion

According to the results above, age and maximum heart rate have influence on resting blood pressure. Elderly people tend to have higher blood pressure since the average blood pressure in larger age groups is higher. People with higher heart rate are easier to develop hypertension. The F-value is much more than the critical F-value, which indicates that maximum heart rate influences greatly on resting blood pressure.

The analysis shows that gender does not influence resting blood pressure. Men and women do not have a significantly difference on blood pressure. Many people may think that high content of serum cholesterol causes hypertension. However, according to the analysis of blood pressure and serum cholesterol content, the P-value is about 0.25. This result does not support many people’s hypothesis. In fact, cholesterol is an essential component of animal cell membranes. High cholesterol may be a consequence of obesity and may incur hyperlipidemia. Evidence showing that high cholesterol may lead to hypertension is limited.

Topic 2

I. Abstract

A) Introduction

In this part of the project, we will be using the Bank Marketing Dataset to build a Logistic regression model. The data consists of direct marketing campaigns of a Portuguese banking institution. The data consists of a total of 45211 records and 17 variables. It has both Numeric and Categorical variables like age, balance, Job, marital status, loan information etc. of the Client's, as well as the promotion details like duration, month, day, outcome etc. (Data source: *https://archive.ics.uci.edu/ml/datasets/Bank+Marketing*)

B) Objective

The objective of this part of the project is to predict if a Client would subscribe to the term deposit subscription as the result of the marketing campaigns. The model used to predict this information is Logistic regression model.

C) Procedure

The Software tool used to perform the analysis and build the model is R in RStudio. R Packages such as data. Table, ggplot2, CA Tools, caret and information Value were used to perform exploratory analysis, build the logistic regression model and to evaluate the results. Since this is a binary classification model, logistic regression seemed to be a good fit.

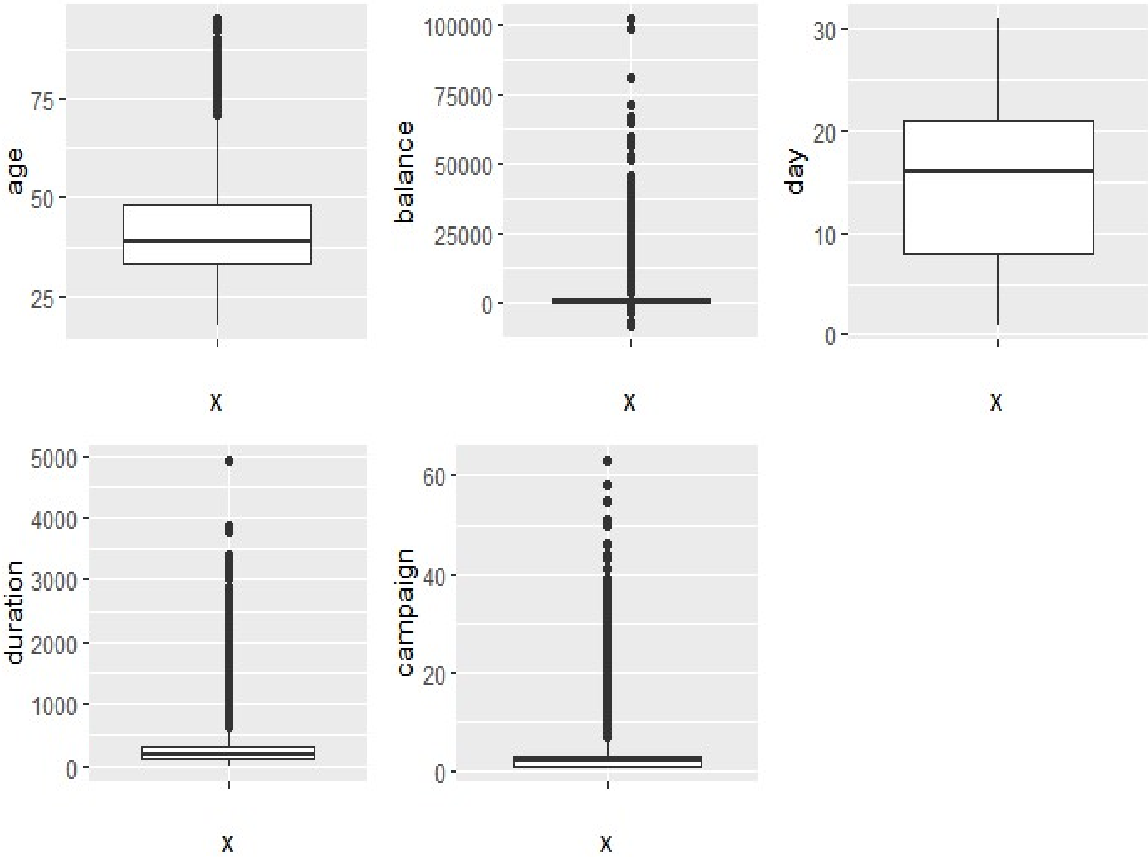
II. Analysis

The data did not have any duplicates or missing values and the summary of the data can be given as:

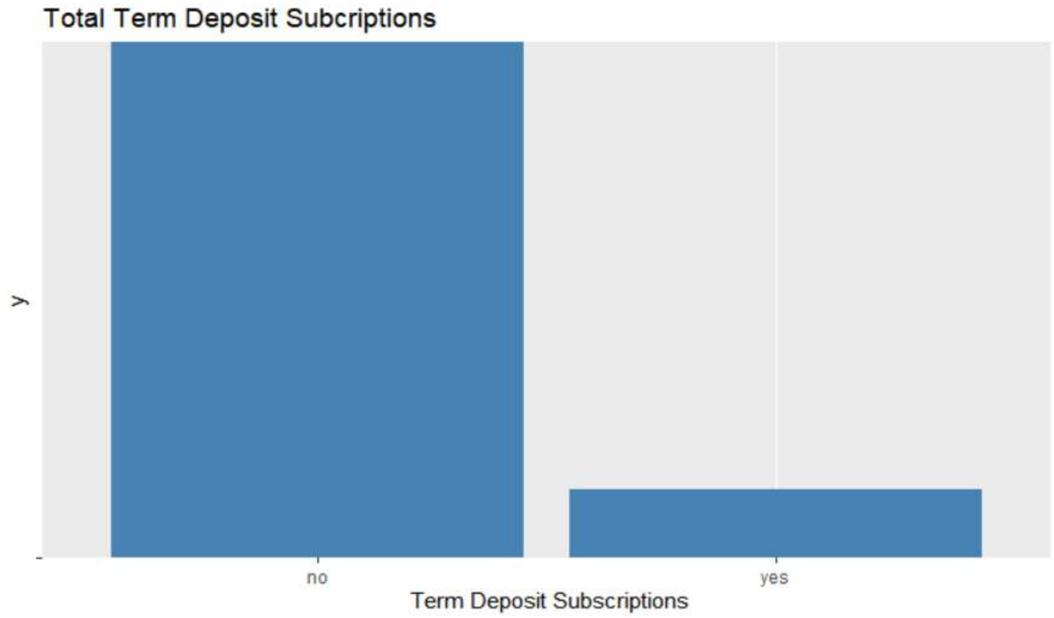
```
      age                 job           marital          education      default    
 Min.   :18.00   blue-collar:9732   divorced: 5207   primary  : 6851   no :44396  
 1st Qu.:33.00   management :9458   married :27214   secondary:23202   yes:  815  
 Median :39.00   technician :7597   single  :12790   tertiary :13301              
 Mean   :40.94   admin.     :5171                    unknown  : 1857              
 3rd Qu.:48.00   services   :4154                                                 
 Max.   :95.00   retired    :2264                                                 
                 (Other)    :6835                                                 
    balance       housing      loan            contact           day            month      
 Min.   : -8019   no :20081   no :37967   cellular :29285   Min.   : 1.00   may    :13766  
 1st Qu.:    72   yes:25130   yes: 7244   telephone: 2906   1st Qu.: 8.00   jul    : 6895  
 Median :   448                           unknown  :13020   Median :16.00   aug    : 6247  
 Mean   :  1362                                             Mean   :15.81   jun    : 5341  
 3rd Qu.:  1428                                             3rd Qu.:21.00   nov    : 3970  
 Max.   :102127                                             Max.   :31.00   apr    : 2932  
                                                                            (Other): 6060  
    duration         campaign          pdays          previous           poutcome    
 Min.   :   0.0   Min.   : 1.000   Min.   : -1.0   Min.   :  0.0000   failure: 4901  
 1st Qu.: 103.0   1st Qu.: 1.000   1st Qu.: -1.0   1st Qu.:  0.0000   other  : 1840  
 Median : 180.0   Median : 2.000   Median : -1.0   Median :  0.0000   success: 1511  
 Mean   : 258.2   Mean   : 2.764   Mean   : 40.2   Mean   :  0.5803   unknown:36959  
 3rd Qu.: 319.0   3rd Qu.: 3.000   3rd Qu.: -1.0   3rd Qu.:  0.0000                  
 Max.   :4918.0   Max.   :63.000   Max.   :871.0   Max.   :275.0000                  
                                                                                     
   y        
 no :39922  
 yes: 5289  
```

```
Classes ‘data.table’ and 'data.frame':  45211 obs. of  17 variables:
 $ age      : int  58 44 33 47 33 35 28 42 58 43 ...
 $ job      : Factor w/ 12 levels "admin.","blue-collar",..: 5 10 3 2 12 5 5 3 6 10 ...
 $ marital  : Factor w/ 3 levels "divorced","married",..: 2 3 2 2 3 2 3 1 2 3 ...
 $ education: Factor w/ 4 levels "primary","secondary",..: 3 2 2 4 4 3 3 3 1 2 ...
 $ default  : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 2 1 1 ...
 $ balance  : int  2143 29 2 1506 1 231 447 2 121 593 ...
 $ housing  : Factor w/ 2 levels "no","yes": 2 2 2 2 1 2 2 2 2 2 ...
 $ loan     : Factor w/ 2 levels "no","yes": 1 1 2 1 1 1 2 1 1 1 ...
 $ contact  : Factor w/ 3 levels "cellular","telephone",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ day      : int  5 5 5 5 5 5 5 5 5 5 ...
 $ month    : Factor w/ 12 levels "apr","aug","dec",..: 9 9 9 9 9 9 9 9 9 9 ...
 $ duration : int  261 151 76 92 198 139 217 380 50 55 ...
 $ campaign : int  1 1 1 1 1 1 1 1 1 1 ...
 $ pdays    : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
 $ previous : int  0 0 0 0 0 0 0 0 0 0 ...
 $ poutcome : Factor w/ 4 levels "failure","other",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ y        : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```

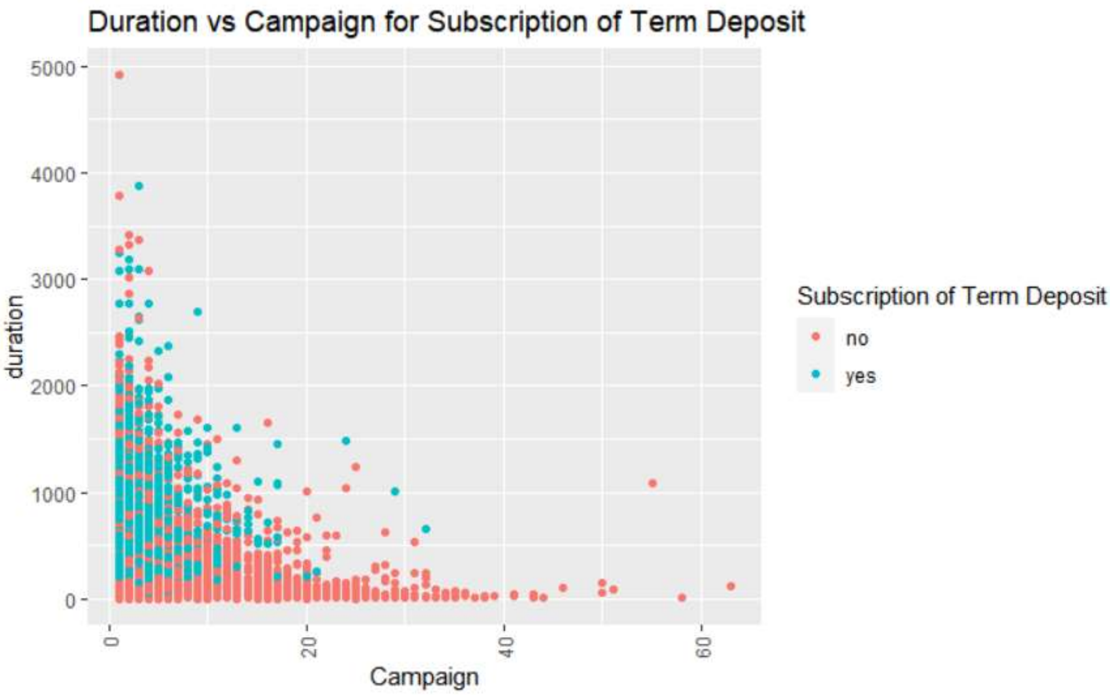
To find if the Predictor variables had outliers that might affect the Target variable, boxplots are considered. The variables age, balance, duration and campaign had a significant amount of outliers.



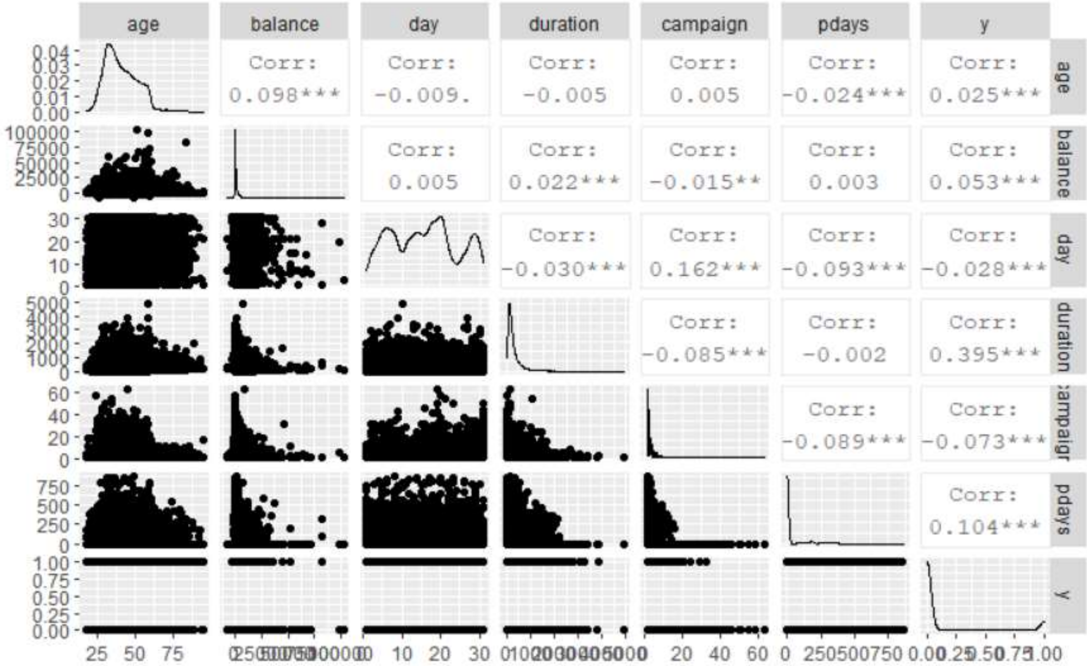
Also the Target variable is skewed towards NO (0).



Most of the Subscriptions occur with less number of Total Campaigns and with average to higher duration. Successful Campaigns occur before first 10 calls and they decrease to much lower rate after that. Duration of the call is also similar in the first 10 contacts.



The Target Variable is then changed to binary variable and the Correlation between the different variables are analysed. The correlation between the numerical values are observed and clearly there isn’t much direct correlation between the variables



III. Logistic regression model

The data is split into Training set and test set. 75% of the data is split into Train set and 25% as test set. Then the Numerical predictors with outliers were standardized to compensate the skewness and to reduce the effect of outliers. Further the regression model is built.

````
classifier.lm = glm(formula = y ~ .,
                    family = binomial,
                    data = training_set)
```
```{r}
pred_lm = predict(classifier.lm, type='response', newdata=test_set[,-17])
```
````

```
Call:
glm(formula = y ~ ., family = binomial, data = training_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.9042  -0.3760  -0.2552  -0.1505   3.4118

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.4961316  0.1647564  -9.081  < 2e-16 ***
age                -0.0064149  0.0270229  -0.237 0.812357
```
```

```
jobblue-collar     -0.2672889  0.0830689  -3.218 0.001292 **
jobentrepreneur    -0.3051980  0.1418816  -2.151 0.031470 *
jobhousemaid       -0.5146324  0.1596849  -3.223 0.001269 **
jobmanagement      -0.1684416  0.0844593  -1.994 0.046113 *
jobretired          0.2221087  0.1120026   1.983 0.047360 *
jobself-employed   -0.3136744  0.1294951  -2.422 0.015423 *
jobservices        -0.2191598  0.0972702  -2.253 0.024253 *
jobstudent          0.3726750  0.1271339   2.931 0.003375 **
jobtechnician      -0.1741671  0.0796773  -2.186 0.028822 *
jobunemployed      -0.1583676  0.1291398  -1.226 0.220076
jobunknown         -0.1824658  0.2620261  -0.696 0.486200
maritalmarried     -0.1298653  0.0685918  -1.893 0.058317 .
maritalsingle       0.0713095  0.0784880   0.909 0.363593
educationsecondary  0.2130456  0.0744832   2.860 0.004232 **
educationtertiary   0.3946393  0.0866480   4.555 5.25e-06 ***
educationunknown    0.2572531  0.1193561   2.155 0.031135 *
defaultyes          0.0235249  0.1885961   0.125 0.900732
balance             0.0463666  0.0179409   2.584 0.009754 **
housingyes         -0.6885606  0.0508506 -13.541  < 2e-16 ***
loanyes            -0.4310351  0.0686728  -6.277 3.46e-10 ***
contacttelephone   -0.1364545  0.0858760  -1.589 0.112067
contactunknown     -1.6140543  0.0850402 -18.980  < 2e-16 ***
day                 0.0800815  0.0240176   3.334 0.000855 ***
monthaug           -0.6741007  0.0900908  -7.482 7.29e-14 ***
monthdec            0.6700482  0.2035746   3.291 0.000997 ***
monthfeb           -0.1727000  0.1033309  -1.671 0.094656 .
monthjan           -1.1250409  0.1358690  -8.280  < 2e-16 ***
monthjul           -0.8141951  0.0889740  -9.151  < 2e-16 ***
monthjun            0.4040684  0.1084617   3.725 0.000195 ***
monthmar            1.6820029  0.1374889  12.234  < 2e-16 ***
monthmay           -0.3929589  0.0831272  -4.727 2.28e-06 ***
monthnov           -0.8685786  0.0973779  -8.920  < 2e-16 ***
monthoct            0.9108856  0.1255391   7.256 3.99e-13 ***
monthsep            0.8480700  0.1387113   6.114 9.72e-10 ***
duration            1.0698324  0.0190733  56.091  < 2e-16 ***
campaign           -0.2809236  0.0365892  -7.678 1.62e-14 ***
pdays              -0.0002158  0.0003484  -0.619 0.535605
previous            0.0081202  0.0064007   1.269 0.204567
poutcomeother       0.2528714  0.1014263   2.493 0.012661 *
poutcomesuccess     2.2485136  0.0942386  23.860  < 2e-16 ***
poutcomeunknown    -0.1814308  0.1061418  -1.709 0.087391 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 24474  on 33908  degrees of freedom
Residual deviance: 16205  on 33866  degrees of freedom
AIC: 16291

Number of Fisher Scoring iterations: 6
```

With the cutoff threshold value for the Binary classifier as 0.39, the Confusion Matrix of the model is:

	0 <int>	1 <int>
0	9646	740
1	334	582

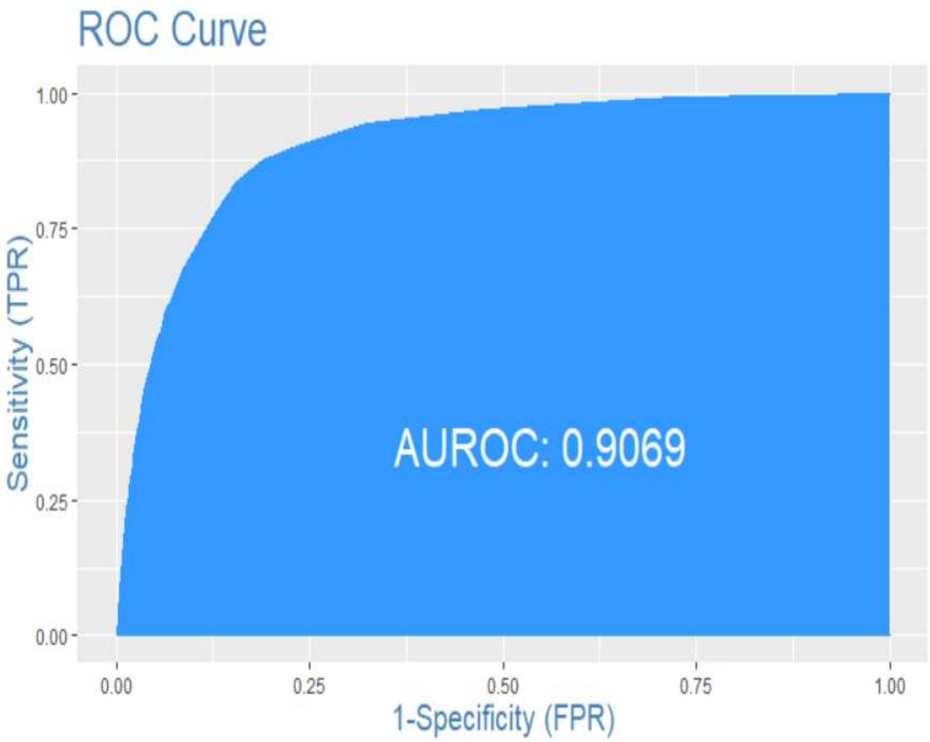
The column values are actuals, while the row values are predicted values.

True Positive: 9646

False Negative: 740

True Negative: 582

False Positive: 334



ROC Curve traces the percentage of accurately predicted True positives. The curve is rising steeply. As the Cutoff Score decreases, the sensitivity (True Positive Rate) is increasing faster than the False Negative Rate. The Greater the area below the curve, the greater the predictability of the model. AUROC = 0.9069 (closer to 1) shows that the model is successful.

The Misclassification error of the model is 0.095 which shows the mismatch of predicted vs actual values. The lesser the error, the better.

Concordance is the percentage of pairs, whose scores of actual positive's are greater than the scores of actual negative's. For a perfect model, this will be 100%. So, the higher the concordance, the better is the quality of model. Concordance of our model is 0.9078

Sensitivity (True Positive Rate) = 0.44

Specificity (1-False Positive Rate) = 0.966

The overall accuracy of the Logistic Regression model is 90%

IV. Conclusion

With the above used data we can successfully predict if a Client would subscribe for a Term deposit using the promotion data. A better prediction accuracy would have been possible if the Target variable was not left skewed to this extent. This data can also be used for predicting the duration of the Marketing call using Linear regression.

REFERENCE:

[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

APPENDIX: R code

Importing and summarizing the data

```
bank <- read.csv("~/NEU_COURSES/Stastical_methods_in_Eng/Project/bank-full.csv", s
ep=";")

summary(bank)

##       age                 job           marital          education
##  Min.   :18.00   blue-collar:9732   divorced: 5207   primary  : 6851
##  1st Qu.:33.00   management :9458   married :27214   secondary:23202
##  Median :39.00   technician :7597   single  :12790   tertiary :13301
##  Mean   :40.94   admin.     :5171                    unknown  : 1857
##  3rd Qu.:48.00   services   :4154
##  Max.   :95.00   retired    :2264
##                  (Other)    :6835
##  default        balance       housing      loan           contact
##  no :44396   Min.   : -8019   no :20081   no :37967   cellular :29285
##  yes:  815   1st Qu.:    72   yes:25130   yes: 7244   telephone: 2906
##              Median :   448                           unknown  :13020
##              Mean   :  1362
##              3rd Qu.:  1428
##              Max.   :102127
##
##       day            month          duration         campaign
##  Min.   : 1.00   may    :13766   Min.   :   0.0   Min.   : 1.000
##  1st Qu.: 8.00   jul    : 6895   1st Qu.: 103.0   1st Qu.: 1.000
##  Median :16.00   aug    : 6247   Median : 180.0   Median : 2.000
##  Mean   :15.81   jun    : 5341   Mean   : 258.2   Mean   : 2.764
##  3rd Qu.:21.00   nov    : 3970   3rd Qu.: 319.0   3rd Qu.: 3.000
##  Max.   :31.00   apr    : 2932   Max.   :4918.0   Max.   :63.000
##                  (Other): 6060
##      pdays          previous          poutcome       y
##  Min.   : -1.0   Min.   :  0.0000   failure: 4901   no :39922
##  1st Qu.: -1.0   1st Qu.:  0.0000   other  : 1840   yes: 5289
##  Median : -1.0   Median :  0.0000   success: 1511
##  Mean   : 40.2   Mean   :  0.5803   unknown:36959
##  3rd Qu.: -1.0   3rd Qu.:  0.0000
##  Max.   :871.0   Max.   :275.0000
##
```

Checking for duplicates and missing variables

```
library(data.table)
bank <- as.data.table(bank)
bank[duplicated(bank)]

## Empty data.table (0 rows and 17 cols): age,job,marital,education,default,balanc
e...
```

```r
sum(!complete.cases(bank))
```

```
## [1] 0
```

```r
sapply(bank, function(x) sum(is.na(x)))
```

```
##       age       job   marital education   default   balance   housing 
##         0         0         0         0         0         0         0 
##      loan   contact       day     month  duration  campaign     pdays 
##         0         0         0         0         0         0         0 
##  previous  poutcome         y 
##         0         0         0
```

The data doesn't have any duplicates or missing values.

```r
str(bank)
```

```
## Classes 'data.table' and 'data.frame':   45211 obs. of  17 variables:
##  $ age      : int  58 44 33 47 33 35 28 42 58 43 ...
##  $ job      : Factor w/ 12 levels "admin.","blue-collar",..: 5 10 3 2 12 5 5 3
6 10 ...
##  $ marital  : Factor w/ 3 levels "divorced","married",..: 2 3 2 2 3 2 3 1 2 3
 ...
##  $ education: Factor w/ 4 levels "primary","secondary",..: 3 2 2 4 4 3 3 3 1 2
 ...
##  $ default  : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 2 1 1 ...
##  $ balance  : int  2143 29 2 1506 1 231 447 2 121 593 ...
##  $ housing  : Factor w/ 2 levels "no","yes": 2 2 2 2 1 2 2 2 2 2 ...
##  $ loan     : Factor w/ 2 levels "no","yes": 1 1 2 1 1 1 2 1 1 1 ...
##  $ contact  : Factor w/ 3 levels "cellular","telephone",..: 3 3 3 3 3 3 3 3 3 3
 ...
##  $ day      : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ month    : Factor w/ 12 levels "apr","aug","dec",..: 9 9 9 9 9 9 9 9 9 9 ...
##  $ duration : int  261 151 76 92 198 139 217 380 50 55 ...
##  $ campaign : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ pdays    : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
##  $ previous : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ poutcome : Factor w/ 4 levels "failure","other",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ y        : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

The Target variable is skewed towards 0(NO)

Boxplots analyzing Outliers

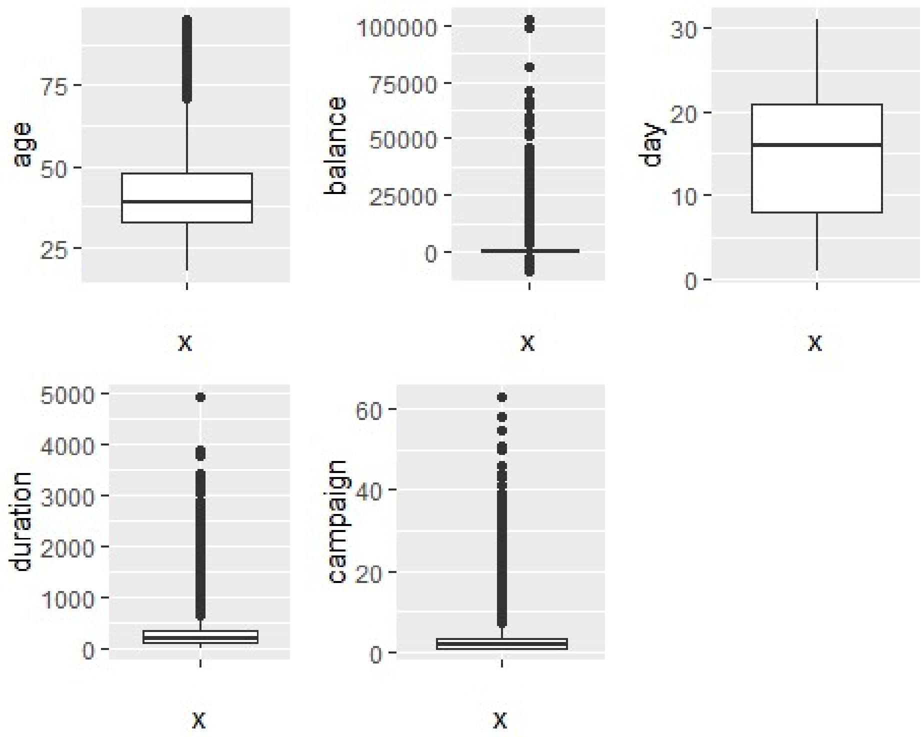
```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```r
library(pdp)
```

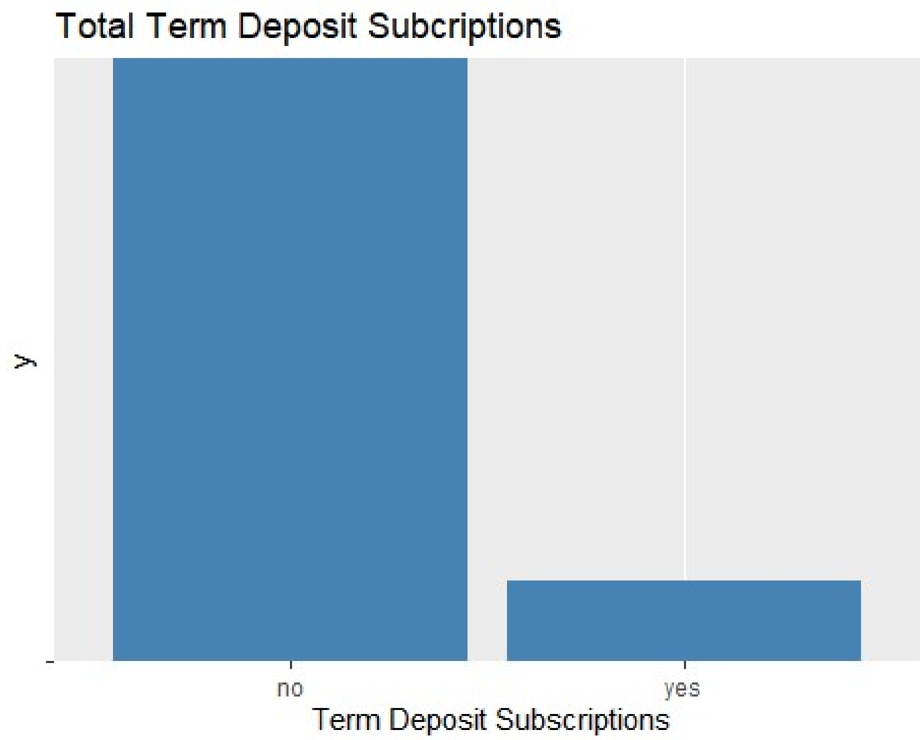
```
## Warning: package 'pdp' was built under R version 3.6.3
```

```
p1 <- ggplot(bank, aes(x='', y=age)) +
  geom_boxplot()
p2 <- ggplot(bank, aes(x='', y=balance)) +
  geom_boxplot()
p3 <- ggplot(bank, aes(x='', y=day)) +
  geom_boxplot()
p4 <- ggplot(bank, aes(x='', y=duration)) +
  geom_boxplot()
p5 <- ggplot(bank, aes(x='', y=campaign)) +
  geom_boxplot()
grid.arrange(p1,p2,p3,p4,p5,ncol =3,nrow=2)
```



Some Visualizations are performed for better understanding of the data:

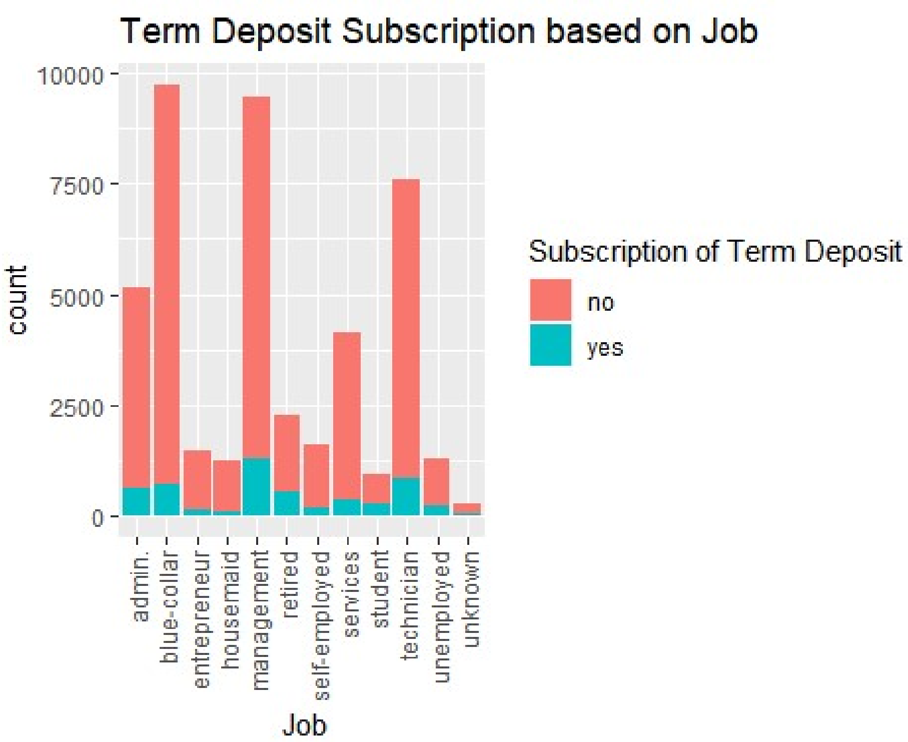
```
barp1 <-ggplot(data=bank, aes(x=y, y='')) +
  geom_bar(stat="identity", fill="steelblue")+ggtitle("Total Term Deposit Subcript
ions") +
         xlab(" Term Deposit Subscriptions")
barp1
```



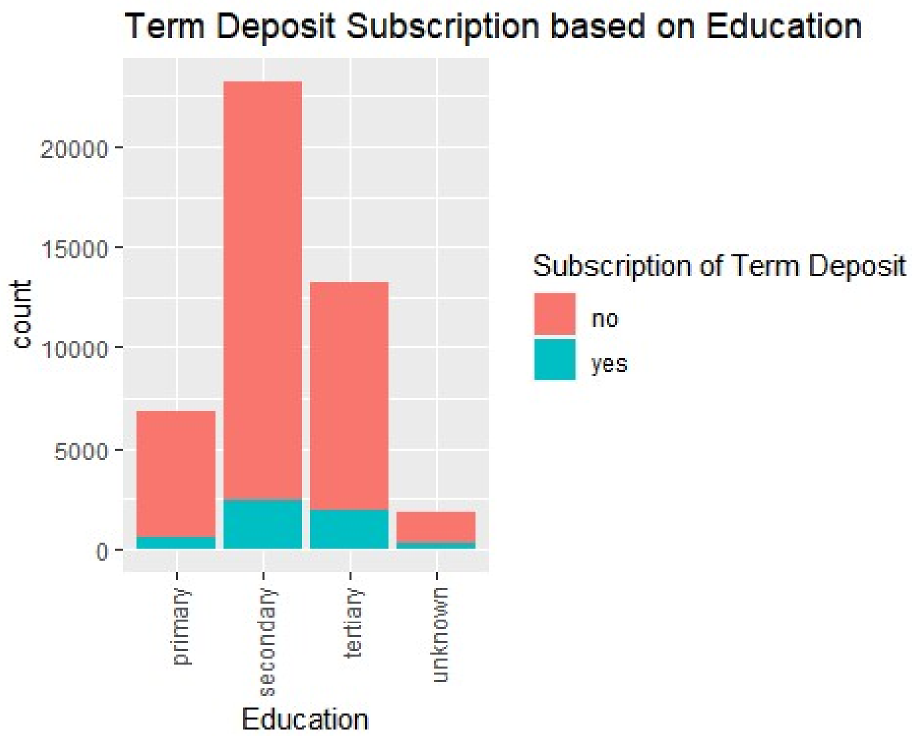
The Target variable is clearly skewed towards 0(NO)

Barplots:

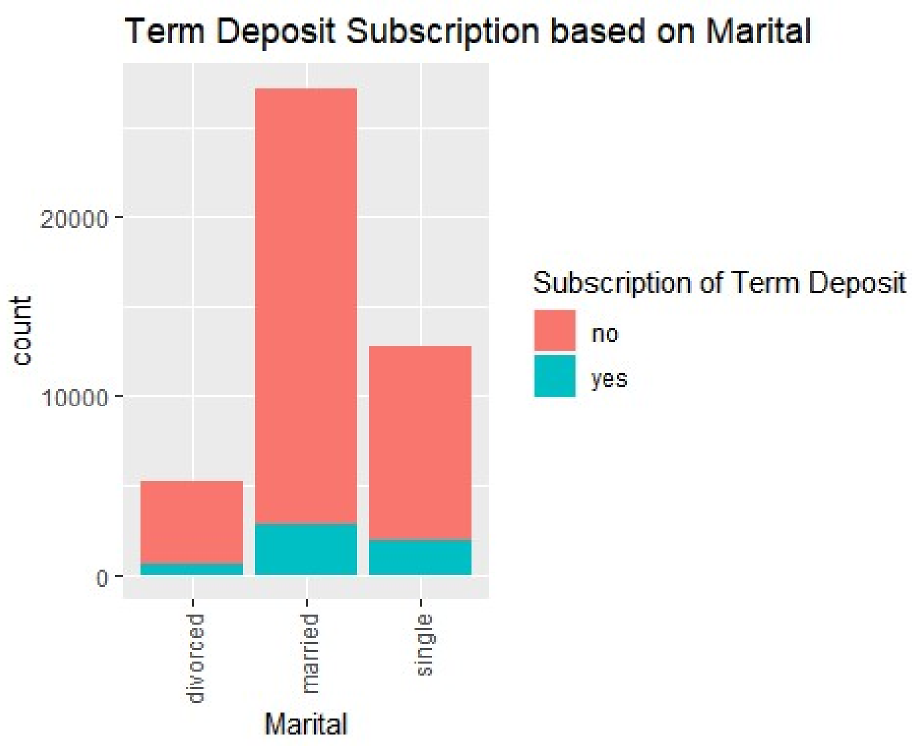
```
barp2 <- ggplot(data = bank, aes(x=job, fill=y)) +
         geom_bar() +
         ggtitle("Term Deposit Subscription based on Job") +
         xlab(" Job") + guides(fill=guide_legend(title="Subscription of Term Depos
it")) + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
barp2
```




```
barp3 <- ggplot(data = bank, aes(x=education, fill=y)) + geom_bar() +
         ggtitle("Term Deposit Subscription based on Education") +
         xlab("Education") + guides(fill=guide_legend(title="Subscription of Term
Deposit")) + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
barp3
```

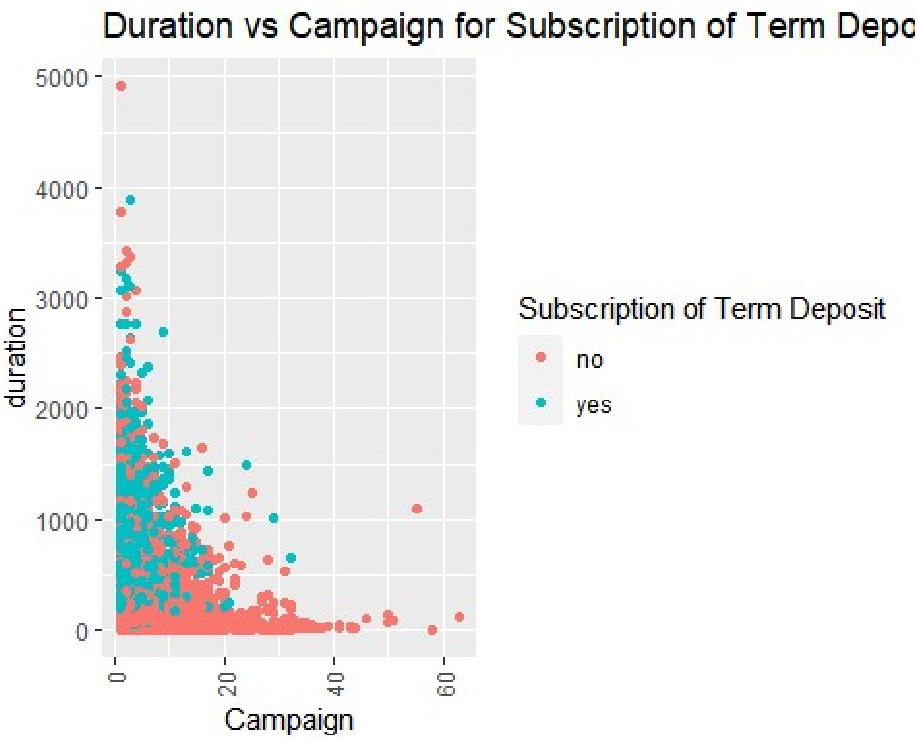



```
barp4 <- ggplot(data = bank, aes(x=marital, fill=y)) + geom_bar() +
         ggtitle("Term Deposit Subscription based on Marital") +
         xlab("Marital") + guides(fill=guide_legend(title="Subscription of Term De
posit")) + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
barp4
```



Scatterplots:

```
Scatterp1 <- ggplot(data = bank, aes(x=campaign,y=duration, color=y)) + geom_point
() +
         ggtitle("Duration vs Campaign for Subscription of Term Deposit") +
         xlab("Campaign") + guides(color=guide_legend(title="Subscription of Term
Deposit")) + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
Scatterp1
```



Changing Target variable to binary

```
bank$y = ifelse(bank$y=='yes',1,0)
str(bank)
```

```
## Classes 'data.table' and 'data.frame':   45211 obs. of  17 variables:
##  $ age      : int  58 44 33 47 33 35 28 42 58 43 ...
##  $ job      : Factor w/ 12 levels "admin.","blue-collar",..: 5 10 3 2 12 5 5 3
6 10 ...
##  $ marital  : Factor w/ 3 levels "divorced","married",..: 2 3 2 2 3 2 3 1 2 3
 ...
##  $ education: Factor w/ 4 levels "primary","secondary",..: 3 2 2 4 4 3 3 3 1 2
 ...
##  $ default  : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 2 1 1 ...
##  $ balance  : int  2143 29 2 1506 1 231 447 2 121 593 ...
##  $ housing  : Factor w/ 2 levels "no","yes": 2 2 2 2 1 2 2 2 2 2 ...
##  $ loan     : Factor w/ 2 levels "no","yes": 1 1 2 1 1 1 2 1 1 1 ...
##  $ contact  : Factor w/ 3 levels "cellular","telephone",..: 3 3 3 3 3 3 3 3 3 3
 ...
##  $ day      : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ month    : Factor w/ 12 levels "apr","aug","dec",..: 9 9 9 9 9 9 9 9 9 9 ...
##  $ duration : int  261 151 76 92 198 139 217 380 50 55 ...
##  $ campaign : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ pdays    : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
##  $ previous : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ poutcome : Factor w/ 4 levels "failure","other",..: 4 4 4 4 4 4 4 4 4 4 ...
```

```
##  $ y         : num  0 0 0 0 0 0 0 0 0 0 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```r
summary(bank)
```

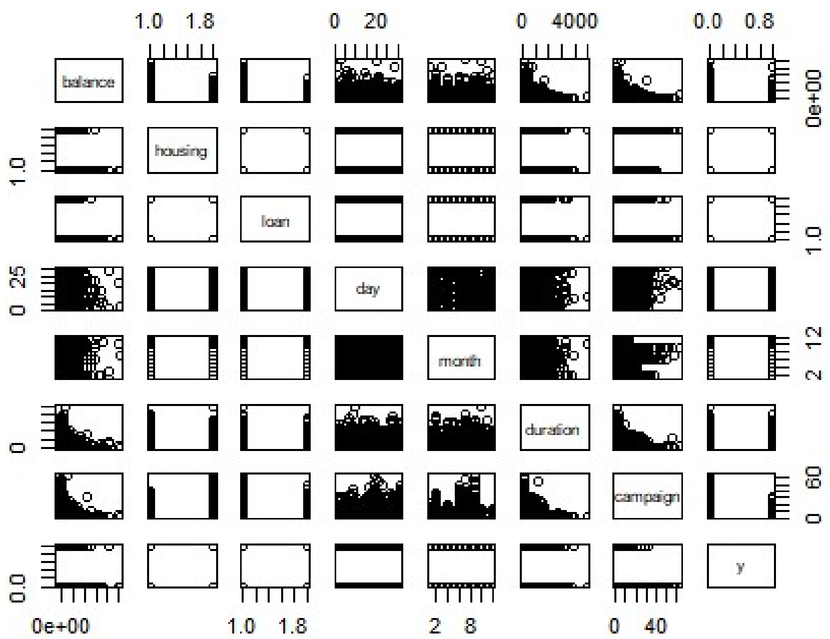
```
##       age                 job           marital          education    
##  Min.   :18.00   blue-collar:9732   divorced: 5207   primary  : 6851  
##  1st Qu.:33.00   management :9458   married :27214   secondary:23202  
##  Median :39.00   technician :7597   single  :12790   tertiary :13301  
##  Mean   :40.94   admin.     :5171                    unknown  : 1857  
##  3rd Qu.:48.00   services   :4154                                     
##  Max.   :95.00   retired    :2264                                     
##                  (Other)    :6835                                     
##  default        balance       housing      loan            contact     
##  no :44396   Min.   : -8019   no :20081   no :37967   cellular :29285  
##  yes:  815   1st Qu.:    72   yes:25130   yes: 7244   telephone: 2906  
##              Median :   448                           unknown  :13020  
##              Mean   :  1362                                            
##              3rd Qu.:  1428                                            
##              Max.   :102127                                            
##                                                                        
##       day            month          duration         campaign     
##  Min.   : 1.00   may    :13766   Min.   :   0.0   Min.   : 1.000  
##  1st Qu.: 8.00   jul    : 6895   1st Qu.: 103.0   1st Qu.: 1.000  
##  Median :16.00   aug    : 6247   Median : 180.0   Median : 2.000  
##  Mean   :15.81   jun    : 5341   Mean   : 258.2   Mean   : 2.764  
##  3rd Qu.:21.00   nov    : 3970   3rd Qu.: 319.0   3rd Qu.: 3.000  
##  Max.   :31.00   apr    : 2932   Max.   :4918.0   Max.   :63.000  
##                  (Other): 6060                                    
##      pdays          previous            poutcome           y        
##  Min.   : -1.0   Min.   :  0.0000   failure: 4901   Min.   :0.000  
##  1st Qu.: -1.0   1st Qu.:  0.0000   other  : 1840   1st Qu.:0.000  
##  Median : -1.0   Median :  0.0000   success: 1511   Median :0.000  
##  Mean   : 40.2   Mean   :  0.5803   unknown:36959   Mean   :0.117  
##  3rd Qu.: -1.0   3rd Qu.:  0.0000                   3rd Qu.:0.000  
##  Max.   :871.0   Max.   :275.0000                   Max.   :1.000  
## 
```

```r
prop.table(table(bank$y))
```

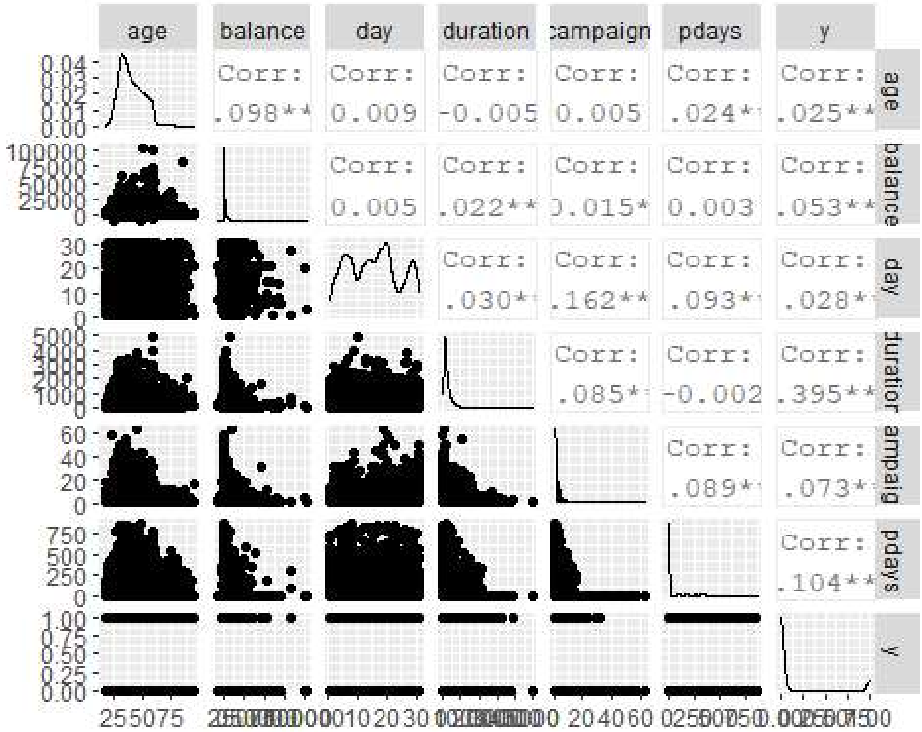
```
## 
##         0         1 
## 0.8830152 0.1169848
```

Correlation matrix:

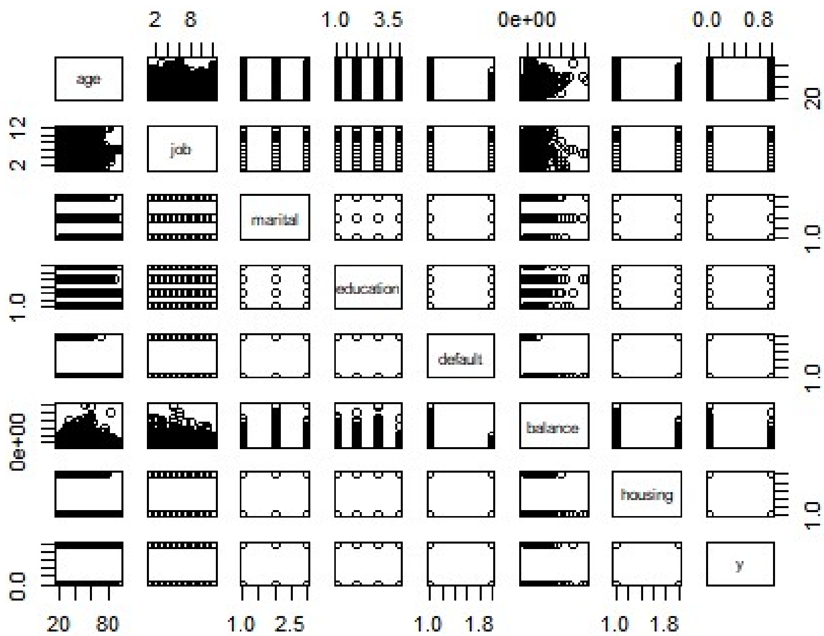
```r
bank.select <- bank[,c(6,7,8,10,11,12,13,17)]
pairs(bank.select)
```



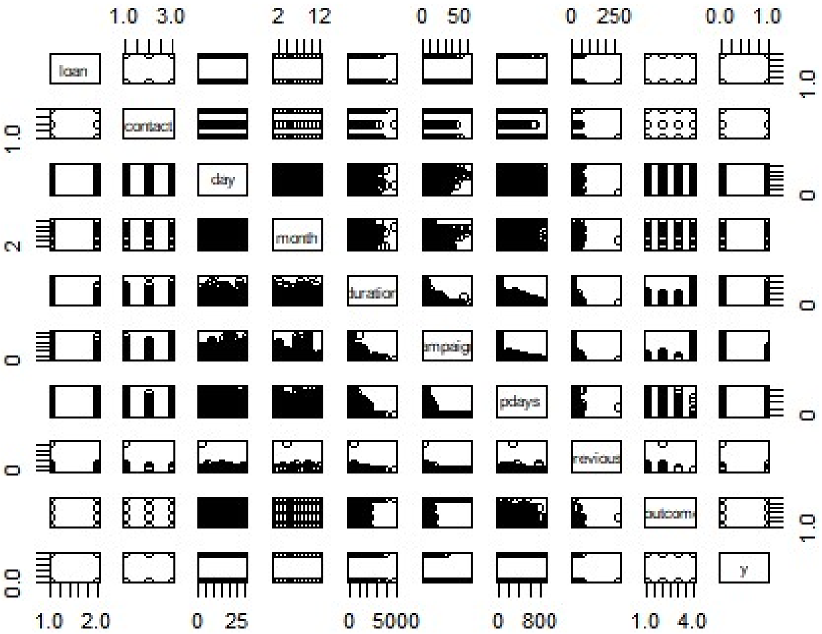

```
library(GGally)

## Warning: package 'GGally' was built under R version 3.6.3

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

bank1 <- bank[,c(1,6,10,12,13,14,17)]
ggpairs(bank1)
```



```r
bank2 <- bank[,c(1,2,3,4,5,6,7,17)]
pairs(bank2)
```



```r
bank3 <- bank[,c(8,9,10,11,12,13,14,15,16,17)]
pairs(bank3)
```



Splitting Training and testing data:

```r
library(caTools)

## Warning: package 'caTools' was built under R version 3.6.3

set.seed(123)
split = sample.split(bank$y,SplitRatio = 0.75)
training_set = subset(bank, split == TRUE)
test_set = subset(bank, split == FALSE)
```

Scaling Numeric Variables

```r
training_set[,c(1)] <- scale(training_set[,c(1)])
training_set[,c(6)] <- scale(training_set[,c(6)])
training_set[,c(10)] <- scale(training_set[,c(10)])
training_set[,c(12)] <- scale(training_set[,c(12)])
training_set[,c(13)] <- scale(training_set[,c(13)])
test_set[,c(1)] <- scale(test_set[,c(1)])
test_set[,c(6)] <- scale(test_set[,c(6)])
test_set[,c(10)] <- scale(test_set[,c(10)])
test_set[,c(12)] <- scale(test_set[,c(12)])
test_set[,c(13)] <- scale(test_set[,c(13)])
```

Building a Logistic Regression model:

```r
classifier.lm = glm(formula = y ~ .,
                 family = binomial,
                 data = training_set)

pred_lm = predict(classifier.lm, type='response', newdata=test_set[,-17])

predicted_y <- data.frame(y = test_set$y, pred = NA)
predicted_y$pred <- pred_lm
```

Confusion matrix: Finding the Optimum Cutoff

```r
library(InformationValue)

## Warning: package 'InformationValue' was built under R version 3.6.3

optCutOff <- optimalCutoff(test_set$y, pred_lm)[1]
optCutOff

## [1] 0.3999999

Results <- confusionMatrix(test_set$y, pred_lm, threshold = optCutOff)
Results

##      0   1
## 0 9646 740
## 1  334 582
```

Summary of the Regression Model

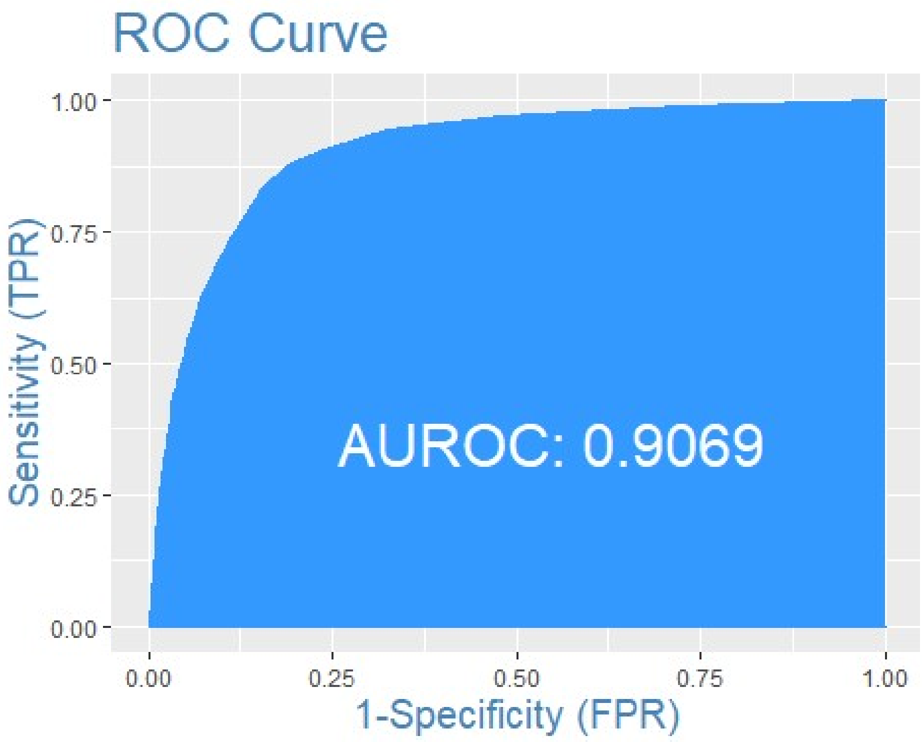
```r
summary(classifier.lm)

##
## Call:
## glm(formula = y ~ ., family = binomial, data = training_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.9042  -0.3760  -0.2552  -0.1505   3.4118
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -1.4961316  0.1647564  -9.081  < 2e-16 ***
## age                -0.0064149  0.0270229  -0.237 0.812357
## jobblue-collar     -0.2672889  0.0830689  -3.218 0.001292 **
## jobentrepreneur    -0.3051980  0.1418816  -2.151 0.031470 *
## jobhousemaid       -0.5146324  0.1596849  -3.223 0.001269 **
## jobmanagement      -0.1684416  0.0844593  -1.994 0.046113 *
## jobretired          0.2221087  0.1120026   1.983 0.047360 *
## jobself-employed   -0.3136744  0.1294951  -2.422 0.015423 *
## jobservices        -0.2191598  0.0972702  -2.253 0.024253 *
## jobstudent          0.3726750  0.1271339   2.931 0.003375 **
## jobtechnician      -0.1741671  0.0796773  -2.186 0.028822 *
```

```
## jobunemployed      -0.1583676  0.1291398  -1.226 0.220076
## jobunknown         -0.1824658  0.2620261  -0.696 0.486200
## maritalmarried     -0.1298653  0.0685918  -1.893 0.058317 .
## maritalsingle       0.0713095  0.0784880   0.909 0.363593
## educationsecondary  0.2130456  0.0744832   2.860 0.004232 **
## educationtertiary   0.3946393  0.0866480   4.555 5.25e-06 ***
## educationunknown    0.2572531  0.1193561   2.155 0.031135 *
## defaultyes          0.0235249  0.1885961   0.125 0.900732
## balance             0.0463666  0.0179409   2.584 0.009754 **
## housingyes         -0.6885606  0.0508506 -13.541  < 2e-16 ***
## loanyes            -0.4310351  0.0686728  -6.277 3.46e-10 ***
## contacttelephone   -0.1364545  0.0858760  -1.589 0.112067
## contactunknown     -1.6140543  0.0850402 -18.980  < 2e-16 ***
## day                 0.0800815  0.0240176   3.334 0.000855 ***
## monthaug           -0.6741007  0.0900908  -7.482 7.29e-14 ***
## monthdec            0.6700482  0.2035746   3.291 0.000997 ***
## monthfeb           -0.1727000  0.1033309  -1.671 0.094656 .
## monthjan           -1.1250409  0.1358690  -8.280  < 2e-16 ***
## monthjul           -0.8141951  0.0889740  -9.151  < 2e-16 ***
## monthjun            0.4040684  0.1084617   3.725 0.000195 ***
## monthmar            1.6820029  0.1374889  12.234  < 2e-16 ***
## monthmay           -0.3929589  0.0831272  -4.727 2.28e-06 ***
## monthnov           -0.8685786  0.0973779  -8.920  < 2e-16 ***
## monthoct            0.9108856  0.1255391   7.256 3.99e-13 ***
## monthsep            0.8480700  0.1387113   6.114 9.72e-10 ***
## duration            1.0698324  0.0190733  56.091  < 2e-16 ***
## campaign           -0.2809236  0.0365892  -7.678 1.62e-14 ***
## pdays              -0.0002158  0.0003484  -0.619 0.535605
## previous            0.0081202  0.0064007   1.269 0.204567
## poutcomeother       0.2528714  0.1014263   2.493 0.012661 *
## poutcomesuccess     2.2485136  0.0942386  23.860  < 2e-16 ***
## poutcomeunknown    -0.1814308  0.1061418  -1.709 0.087391 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24474  on 33908  degrees of freedom
## Residual deviance: 16205  on 33866  degrees of freedom
## AIC: 16291
##
## Number of Fisher Scoring iterations: 6
```

```r
misClassError(test_set$y, pred_lm, threshold = optCutOff)
```

```
## [1] 0.095
```

```r
plotROC(test_set$y, pred_lm)
```




```r
Concordance(test_set$y, pred_lm)

## $Concordance
## [1] 0.9078979
##
## $Discordance
## [1] 0.09210213
##
## $Tied
## [1] 2.775558e-17
##
## $Pairs
## [1] 13193560

sensitivity(test_set$y, pred_lm, threshold = optCutOff)

## [1] 0.4402421

specificity(test_set$y, pred_lm, threshold = optCutOff)

## [1] 0.9665331

accuracy = (Results['1','1']+Results['0','0'])/(Results['0','1'] + Results['1','0
'] + Results['1','1'] + Results['0','0'])
accuracy

## [1] 0.9049726
```