Project Report on

'Segmentation and Predictive Analysis of Bank Marketing Data'

Undertaken at:

Sopra Steria, Sector - 135, Noida, India

22nd June, 2017 - 31st August 2017

Submitted by:

Sugandha Gupta

B.Tech. Electronics and Communication Engineering

Amity School of Engineering and Technology, New Delhi

Under the Guidance of:

Archana Singh

Senior Software Engineer

Sopra Steria Group

Noida

ACKNOWLEDGEMENT

The success and final outcome of this project required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along the duration of my project.

With deepest gratitude, I thank Sopra Steria for accepting me for their externship opportunity and for giving me a chance to work on a Machine Learning project.

With deepest gratitude, I thank my mentor Ms. Archana Singh for giving me this great opportunity to learn and work on a project that greatly interests me. Without her guidance and supervision, I could have never been able to complete the project.

I would also like to thank Ms. Dipinti Phutela for her efforts to match me to the right mentor so that I could work on something that complies with my interest.

Sugandha Gupta

CERTIFICATE

About Sopra Steria Group

Sopra Steria, a European leader in digital transformation, provides one of the most comprehensive portfolios of end-to-end service offerings on the market: consulting, systems integration, software development, infrastructure management and business process services. Sopra Steria is trusted by leading private and public-sector organisations to deliver successful transformation programmes that address their most complex and critical business challenges. Combining high quality and performance services, added value and innovation, Sopra Steria enables its clients to make the best use of digital technology. With over 40,000 employees in more than 20 countries, Sopra Steria had revenue of €3.7 billion in 2016.

Sopra Steria Group SA was established in September 2014 upon the merger of Sopra Group SA and Groupe Steria SCA. India is an Integral part of Sopra Steria's global business strategy. Sopra Steria has a strong local presence in India with more than 5,000 people working across 4 delivery centers: Noida, Bangalore, Chennai and Pune.

The India operation serves more than 70 customers, including 8 of the Group's top 10 clients. Sopra Steria India's proven offshore capabilities ensure superior value to its customers in terms of cost advantage, quality, speedy delivery and enhanced flexibility. In India, they offer a full spectrum of business solutions across a wide range of Industries to help solve customer's complex industry challenges.

CONTENTS

ABSTRACT

The project, 'Segmentation and Predictive Analysis of the Bank Marketing Data' is based on Machine Learning. The dataset for this project was downloaded from UCI Machine Learning Repository, where it was available as an open source dataset for free.

The dataset used in the project has 45211 observations, and 17 variables for each row. The task was to be able to create a model that predicts how many people will subscribe for the term deposit, using a number of variables.

Software Used

For this project, the RStudio software has been used for two applications:

- First, a dashboard based on the Bank Marketing dataset was made using the ShinyApp library, developed on RStudio which is based on R language.
- Second, the Machine Learning model was developed on RStudio, which is based on the R language as well.

INTRODUCTION

For the campaigning needs, organizations rely mainly on either mass campaign, or direct campaign. Mass campaign is focused on a large group of people of possibly different age groups, and even those people for whom the campaign is not even relevant. In contrast, the direct campaign focuses on specific potential clients and is more effective.

For this project, we work on a direct marketing campaign by a bank aimed to increase subscriptions to their term deposits. The aim of this project is to build a model to predict whether a particular client will subscribe to term deposit or not. If classifier has very high accuracy, it can help the bank manager to filter clients and use available resources more efficiently to achieve the campaign goal. Also, identifying the influential factors for customers' decision is also important so that the bank can establish efficient and precise campaigning strategy. Proper strategy would reduce cost and improve long term relations with the clients.

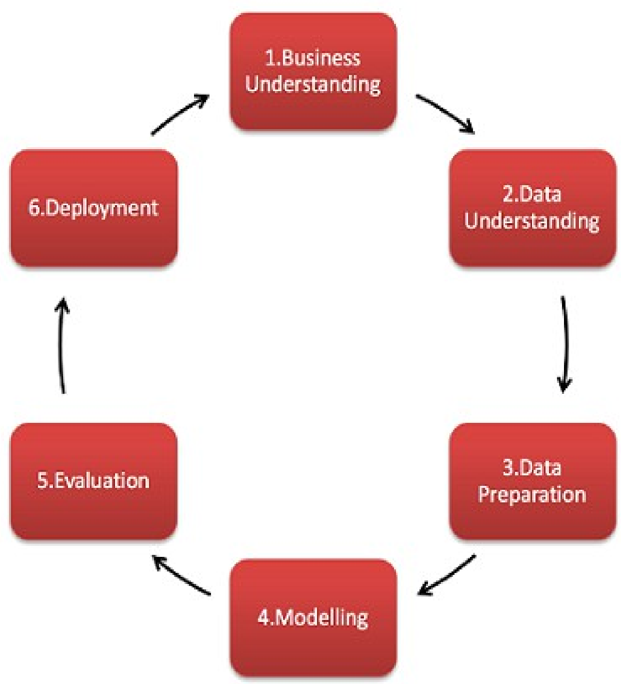
CRISP DM Methodology

CRISP-DM stands for cross-industry process for data mining. The CRISP-DM methodology provides a structured approach to planning a data mining project. It is a robust and well-proven methodology. It is the golden thread than runs through almost every client engagement. The CRISP-DM model is shown below.

This model is an idealized sequence of events. In practice many of the tasks can be performed in a different order and it will often be necessary to backtrack to previous tasks and repeat certain actions. The model does not try to capture all possible routes through the data mining process.

The different phases of this methodology are:

1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment



STAGE ONE – BUSINESS UNDERSTANDING

The steps involved in this stage are:
Determine the desired outputs of the project
Assess the current situation
Determine data mining goals
Produce project plan

STAGE TWO – DATA UNDERSTANDING

The steps involved in this stage are:
Describe data
Explore data
Verify data quality
Data quality report

STAGE THREE – DATA PREPARATION

The steps involved in this stage are:
Select your data
Clean your data
Construct required data
Integrate data

STAGE FOUR – MODELLING

The steps involved in this stage are:
Select modeling technique
Generate test design
Build model
Assess model

STAGE FIVE – EVALUATION

The steps involved in this stage are:
Evaluate your results
Review process

STAGE SIX – DEPLOYMENT

The steps involved in this stage are:
Plan deployment
Plan monitoring and maintenance
Produce final report
Review project

Shiny App Dashboard

Shiny is an open source R package that provides an elegant and powerful web framework for building web applications using R. Shiny helps you turn your analyses into interactive web applications without requiring HTML, CSS, or JavaScript knowledge.

Features

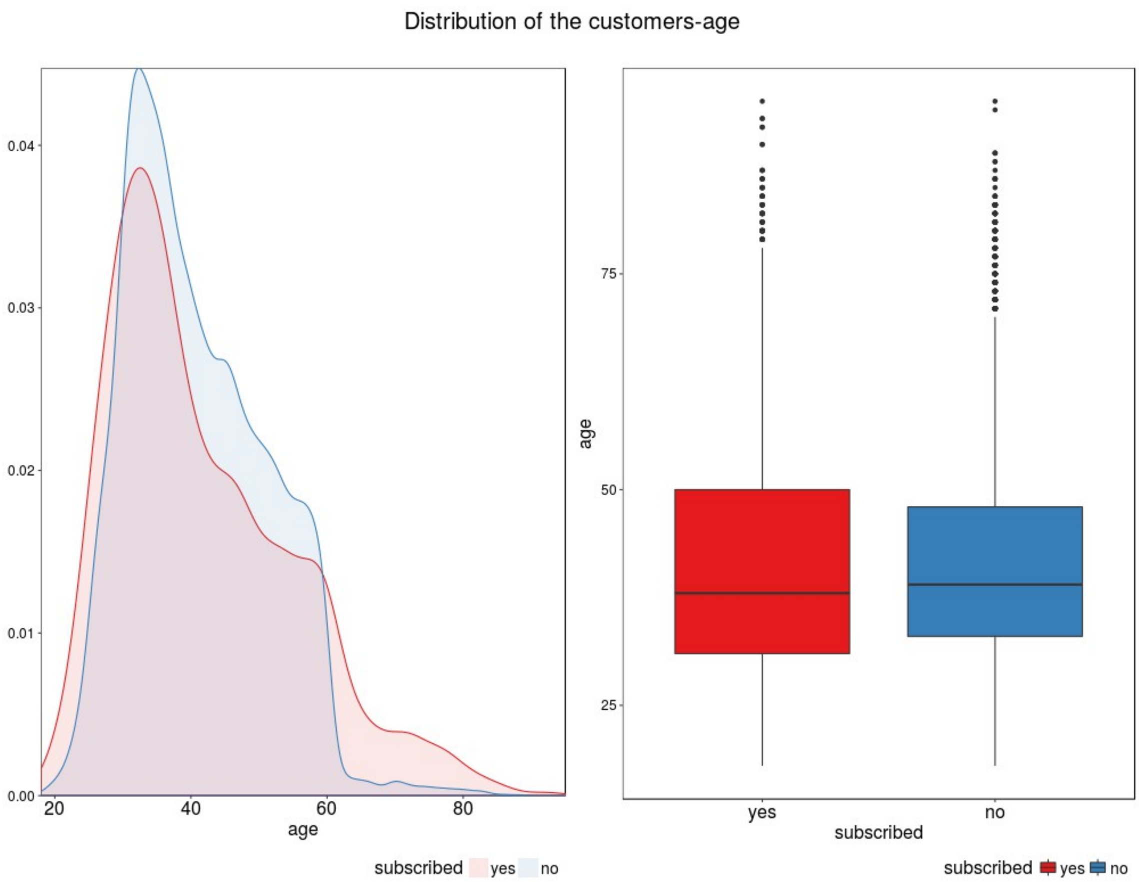
1. Build useful web applications with only a few lines of code
2. Shiny applications are automatically “live” in the same way that spreadsheets are live. Outputs change instantly as users modify inputs, without requiring a reload of the browser.
3. Shiny user interfaces can be built entirely using R, or can be written directly in HTML, CSS, and JavaScript for more flexibility.
4. Works in any R environment
5. Pre-built output widgets for displaying plots, tables, and printed output of R objects.
6. Fast bidirectional communication between the web browser and R.
7. Easy to develop and distribute.

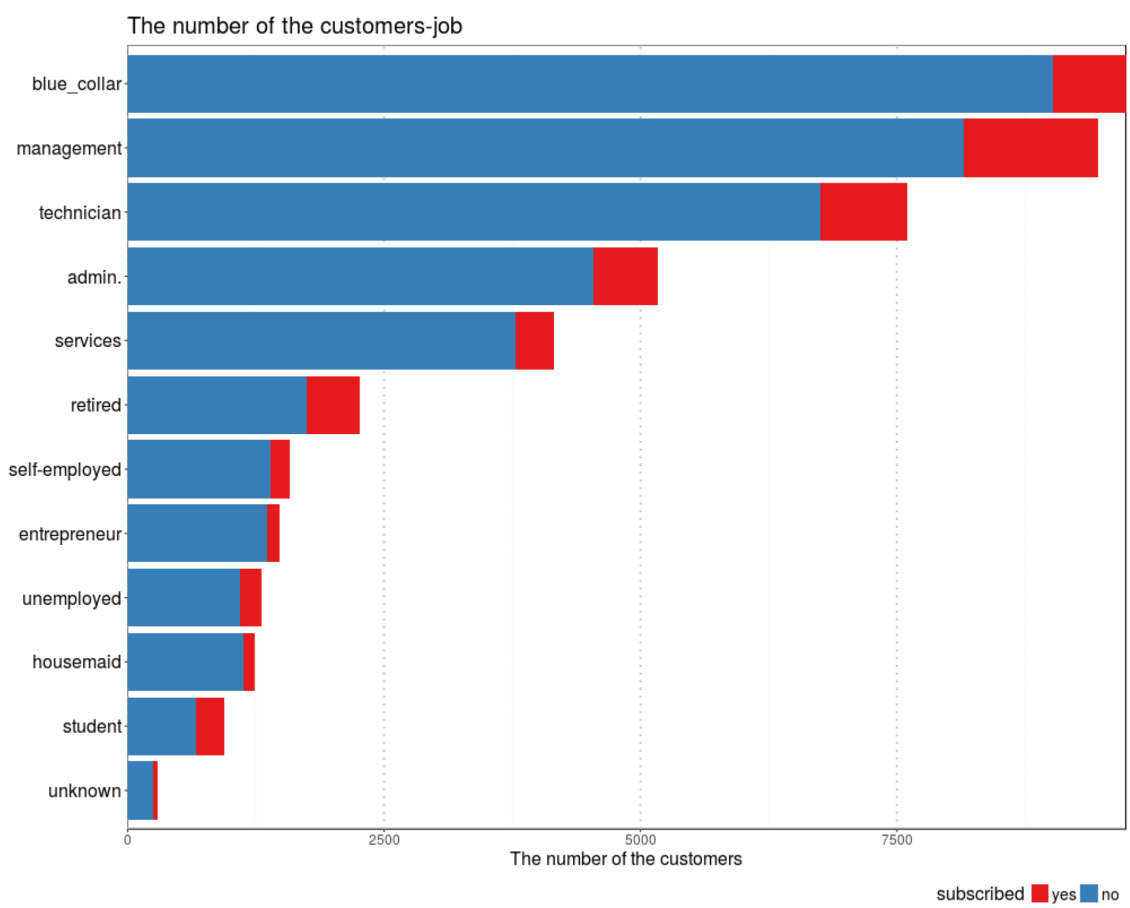
Installation

Shiny is available on CRAN, and can be installed from the console by entering the command:

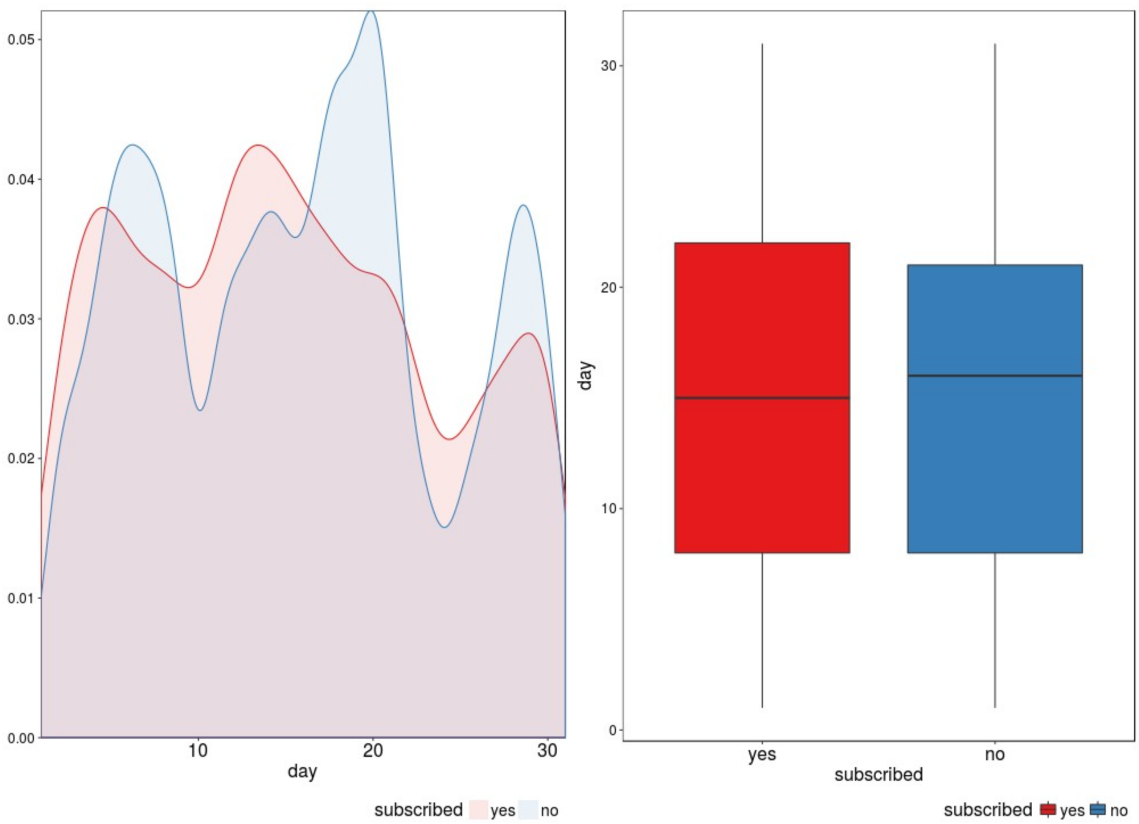
install.packages("shiny")

For this project, a shiny app was developed and uploaded on the online server of the Shiny App website. A few screenshots of the dashboard are shown:

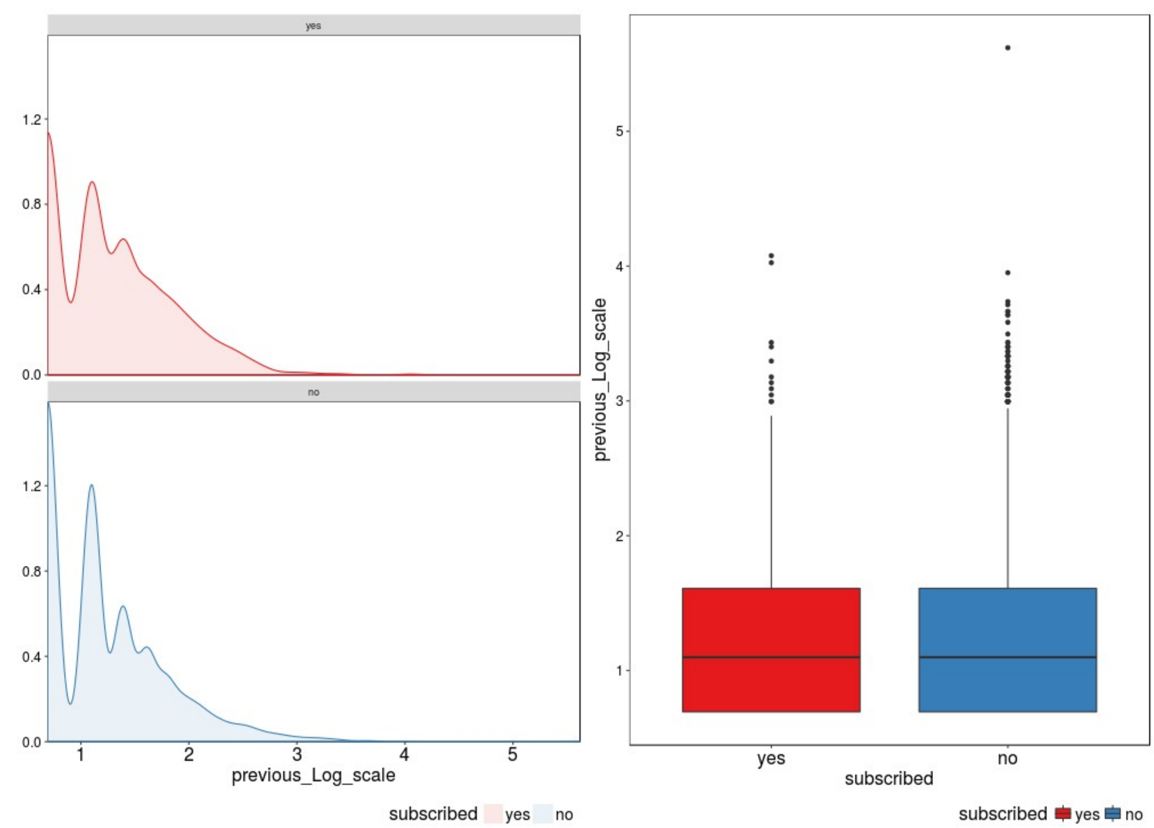

Distribution of the customers-age
0.04
0.03
0.02
0.01
0.00
20
40
60
80
age
subscribed yes no
75
50
25
age
yes
no
subscribed
subscribed yes no

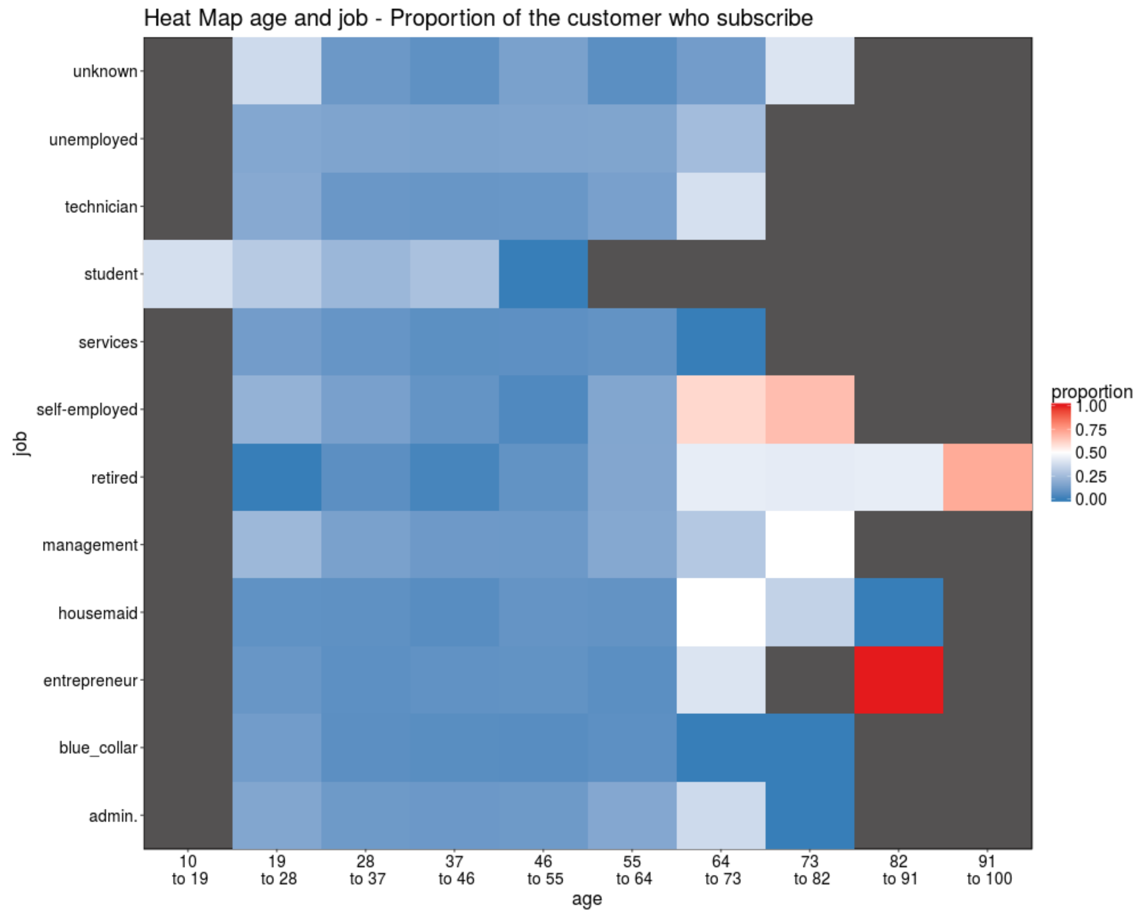

The number of the customers-job
blue_collar
management
technician
admin.
services
retired
self-employed
entrepreneur
unemployed
housemaid
student
unknown
0
2500
5000
7500
The number of the customers
subscribed yes no

Distribution of the customers-day

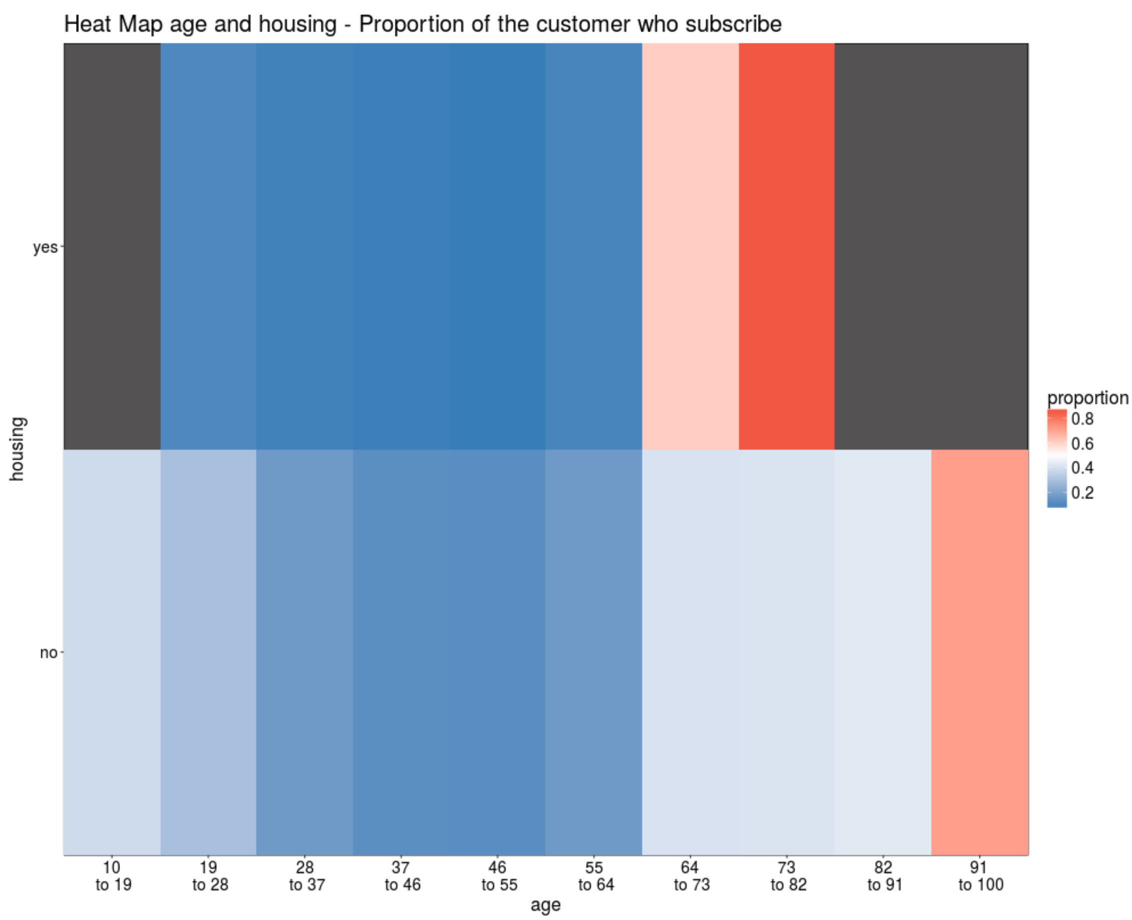


Distribution of the customers-previous
36954 rows are deleted by Log scaling

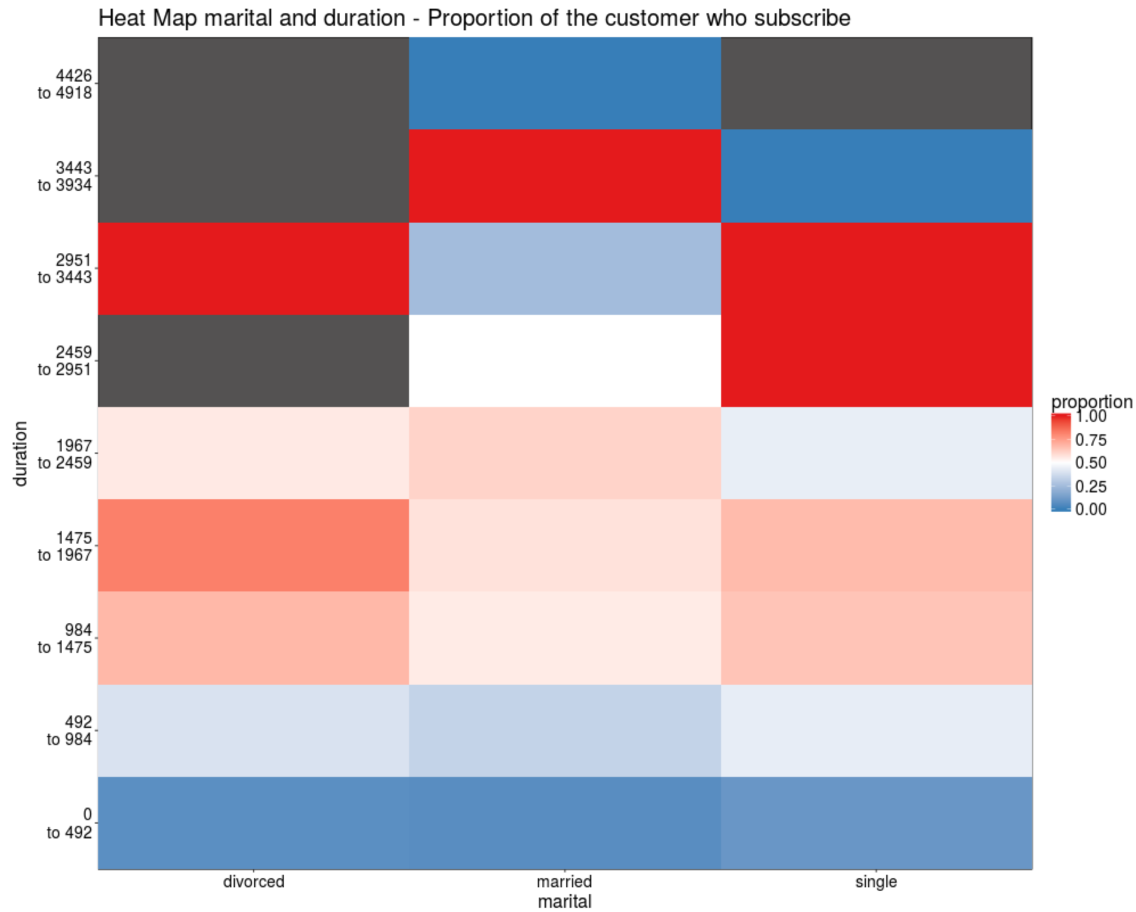


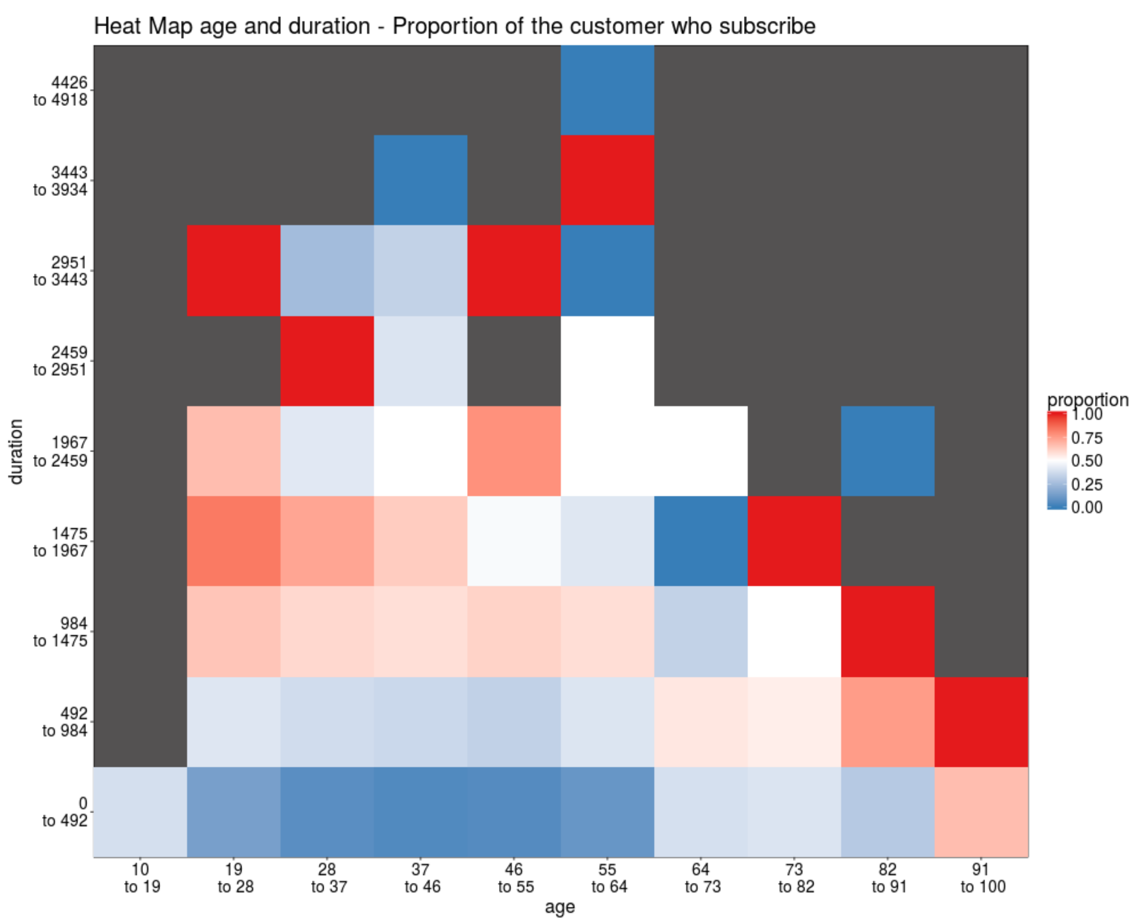


Heat Map age and job - Proportion of the customer who subscribe
unknown
unemployed
technician
student
services
self-employed
retired
management
housemaid
entrepreneur
blue_collar
admin.
job
10 to 19
19 to 28
28 to 37
37 to 46
46 to 55
55 to 64
64 to 73
73 to 82
82 to 91
91 to 100
age
proportion
1.00
0.75
0.50
0.25
0.00



Heat Map age and housing - Proportion of the customer who subscribe
yes
no
housing
10 to 19
19 to 28
28 to 37
37 to 46
46 to 55
55 to 64
64 to 73
73 to 82
82 to 91
91 to 100
age
proportion
0.8
0.6
0.4
0.2


Heat Map marital and duration - Proportion of the customer who subscribe
4426 to 4918
3443 to 3934
2951 to 3443
2459 to 2951
1967 to 2459
1475 to 1967
984 to 1475
492 to 984
0 to 492
duration
divorced
married
single
marital
proportion
1.00
0.75
0.50
0.25
0.00


Heat Map age and duration - Proportion of the customer who subscribe
4426 to 4918
3443 to 3934
2951 to 3443
2459 to 2951
1967 to 2459
1475 to 1967
984 to 1475
492 to 984
0 to 492
duration
10 to 19
19 to 28
28 to 37
37 to 46
46 to 55
55 to 64
64 to 73
73 to 82
82 to 91
91 to 100
age
proportion
1.00
0.75
0.50
0.25
0.00

Machine Learning Model Development

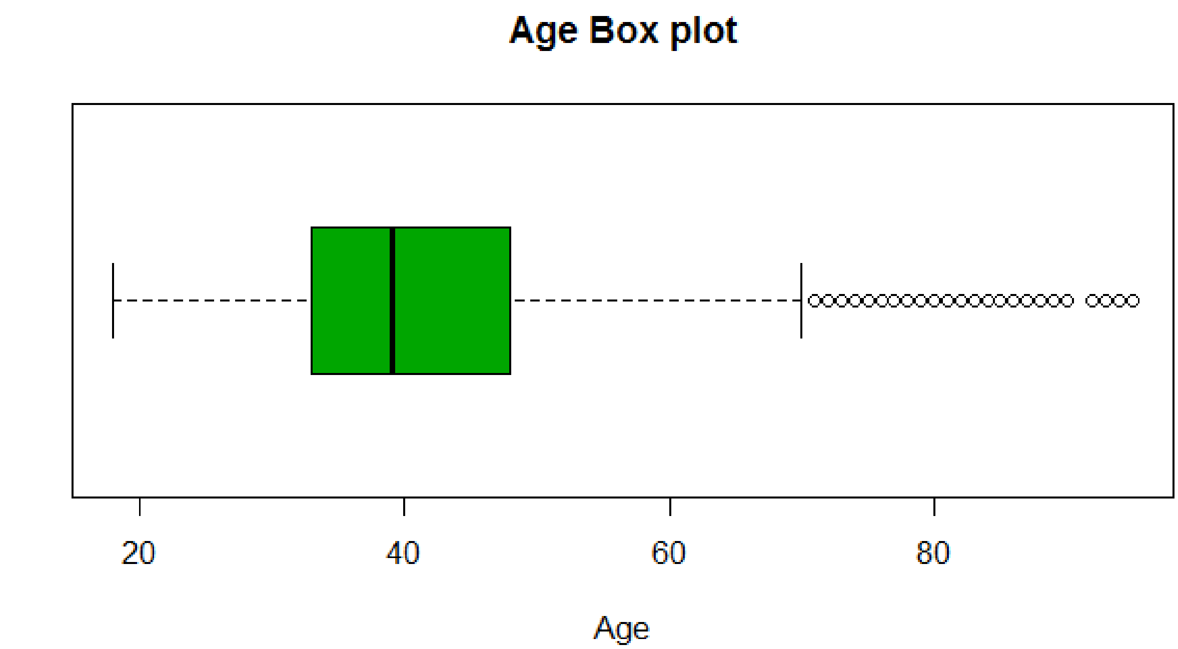
The dataset was loaded in the Rstudio software. The dataset contained 45211 observations, each containing 17 variables.
The structure of the dataset is as follows:

```
'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 ...
```

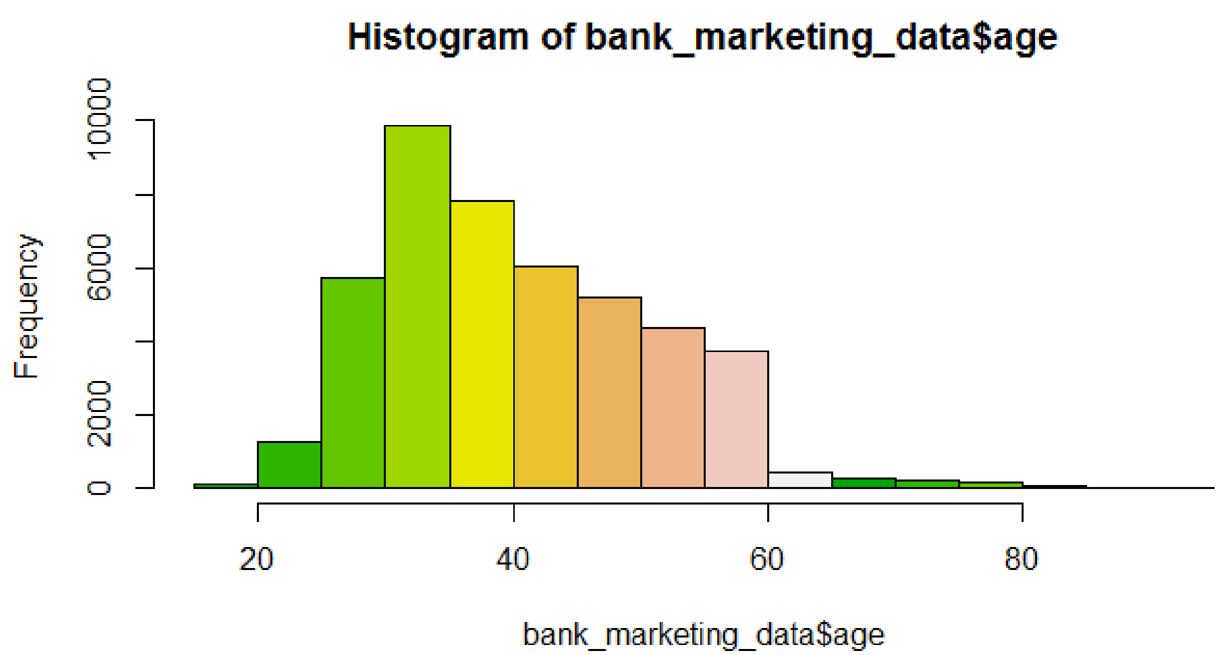
Summary of the data is as follows:

```
      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  
```

OUTLIER DETECTION AND TREATMENT



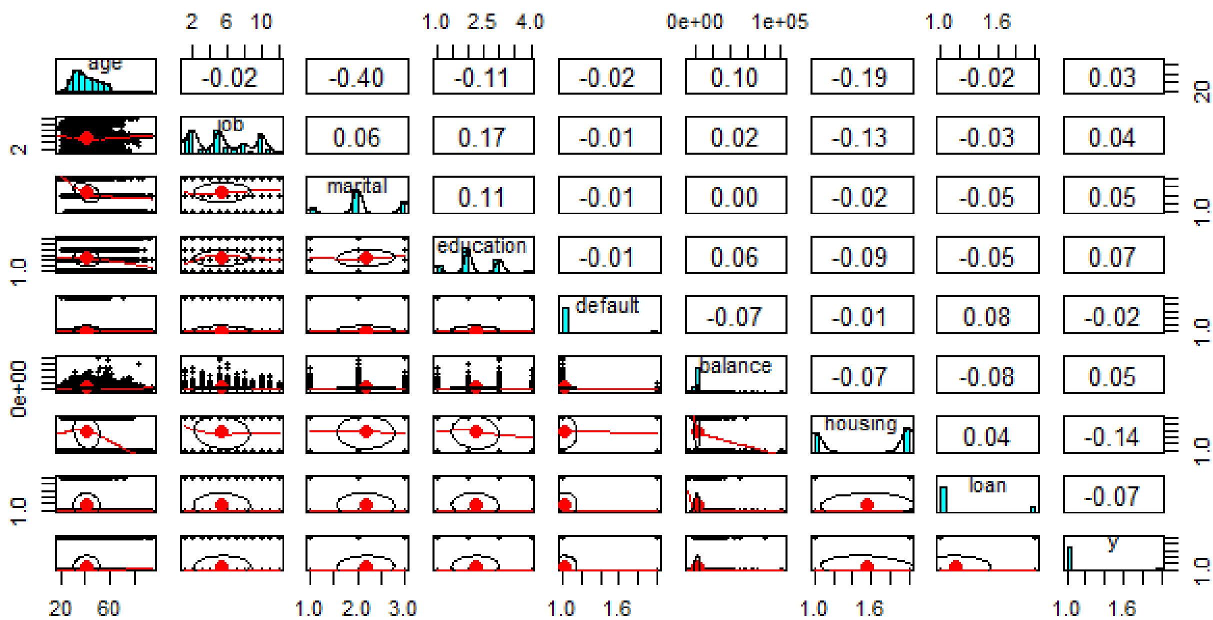
Next, we obtain a boxplot of the values of age. We can see the median and the 1st and 3rd quartiles. Outliers are also seen as the circles outside the higher side of range.Next, we obtain a histogram to confirm the presence of outliers.

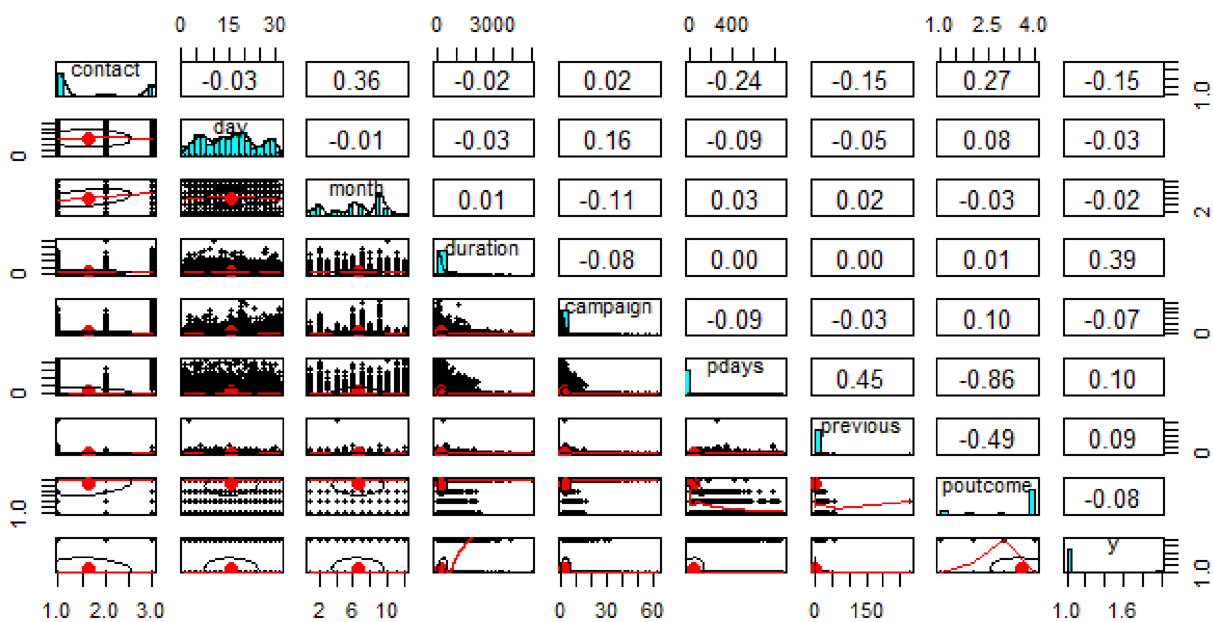


It is confirmed that there are no outliers in the age column.

CORRELATION ANALYSIS

What we saw in the box plot can be emphasized by correlation plot, It can tell if predictor is a good predictor or not a good predictor. This analysis can help us decide if we can drop some columns/predictors depending upon its correlation with the outcome variable. We obtain the correlation plot by using psych library's pairs.panels function.

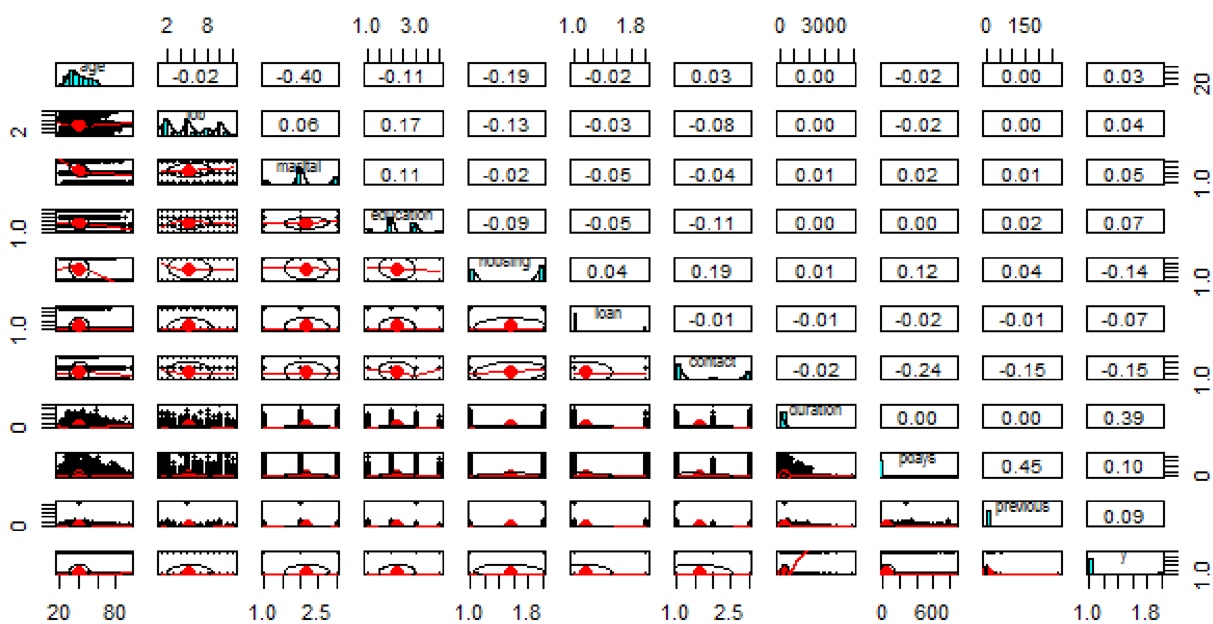




Based on the correlation values above, we create a subset to reduce the feature space. This is the structure of data after subset creation.

```
'data.frame':   45211 obs. of  11 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 ...
 $ 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 ...
 $ duration : int  261 151 76 92 198 139 217 380 50 55 ...
 $ pdays    : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
 $ previous : int  0 0 0 0 0 0 0 0 0 0 ...
 $ y        : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```

The correlation plot after making subset:

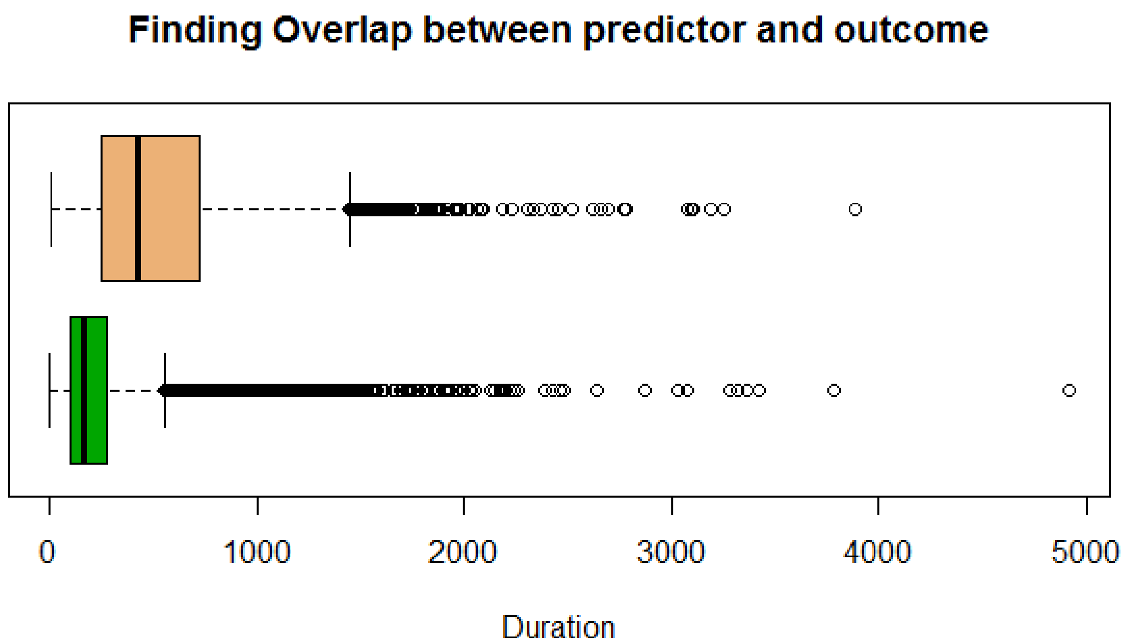


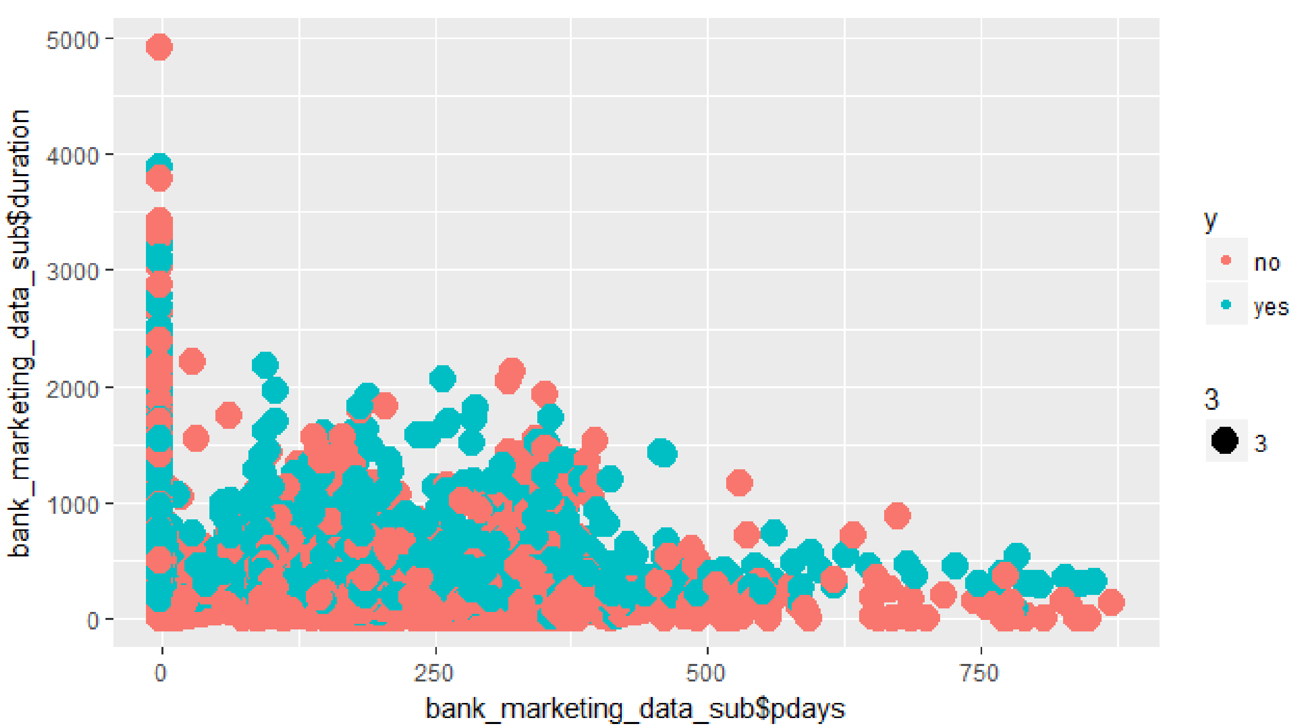
Next, we perform data transformation. We convert the categorical values into numerical values.

```
'data.frame':   45211 obs. of  13 variables:
 $ age        : Factor w/ 4 levels "(1,20]","(20,40]",..: 3 3 2 3 2 2 2 3 3 3 ...
 $ job        : Factor w/ 12 levels "admin.","blue-collar",..: 5 10 3 2 12 5 5 3 6
10 ...
 $ education  : Factor w/ 4 levels "primary","secondary",..: 3 2 2 4 4 3 3 3 1
2 ...
 $ 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 ...
```

```
$ duration   : int  261 151 76 92 198 139 217 380 50 55 ...
$ pdays      : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
$ previous   : int  0 0 0 0 0 0 0 0 0 0 ...
$ y          : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
$ is_divorced: num  0 0 0 0 0 0 0 1 0 0 ...
$ is_single  : num  0 1 0 0 1 0 1 0 0 1 ...
$ is_married : num  1 0 1 1 0 1 0 0 1 0 ...
```

Next, we obtain plots to determine the overlap between the predictors and output to check whether they can be major predictor.





Next, for training and testing, we can use CreateDataPartition method present in caret package to split in such a way that training and testing data will have same ratio of target variable.

Decision Tree Model

In R, decision tree algorithm can be implemented using *rpart package.* In addition, we'll use *caret package* for doing cross validation. Cross validation is a technique to build robust models which are not prone to overfitting.
Summary of the decision tree model:

```
Call:
rpart(formula = y ~ ., data = training)
  n= 31649

          CP nsplit rel error    xerror       xstd
1 0.02984067      0 1.0000000 1.0000000 0.01544198
2 0.01000000      2 0.9403187 0.9432892 0.01505397

Variable importance
duration
     100

Node number 1: 31649 observations,    complexity param=0.02984067
  predicted class=no   expected loss=0.1170021  P(node) =1
    class counts: 27946  3703
   probabilities: 0.883 0.117
  left son=2 (28215 obs) right son=3 (3434 obs)
  Primary splits:
      duration < 523.5 to the left,  improve=818.3451, (0 missing)
      pdays    < 8.5   to the left,  improve=178.2470, (0 missing)
      previous < 0.5   to the left,  improve=176.1903, (0 missing)
      age      splits as  RLLR,      improve=155.8784, (0 missing)
      contact  splits as  RRL,       improve=147.7386, (0 missing)

Node number 2: 28215 observations
  predicted class=no   expected loss=0.07733475  P(node) =0.8914974
    class counts: 26033  2182
   probabilities: 0.923 0.077

Node number 3: 3434 observations,    complexity param=0.02984067
  predicted class=no   expected loss=0.4429237  P(node) =0.1085026
    class counts:  1913  1521
   probabilities: 0.557 0.443
  left son=6 (2227 obs) right son=7 (1207 obs)
  Primary splits:
      duration   < 835.5 to the left,  improve=82.22512, (0 missing)
      contact    splits as  RRL,       improve=34.07002, (0 missing)
      is_married < 0.5   to the right, improve=16.23152, (0 missing)
      pdays      < 44.5  to the left,  improve=14.36751, (0 missing)
      previous   < 0.5   to the left,  improve=13.75921, (0 missing)
```

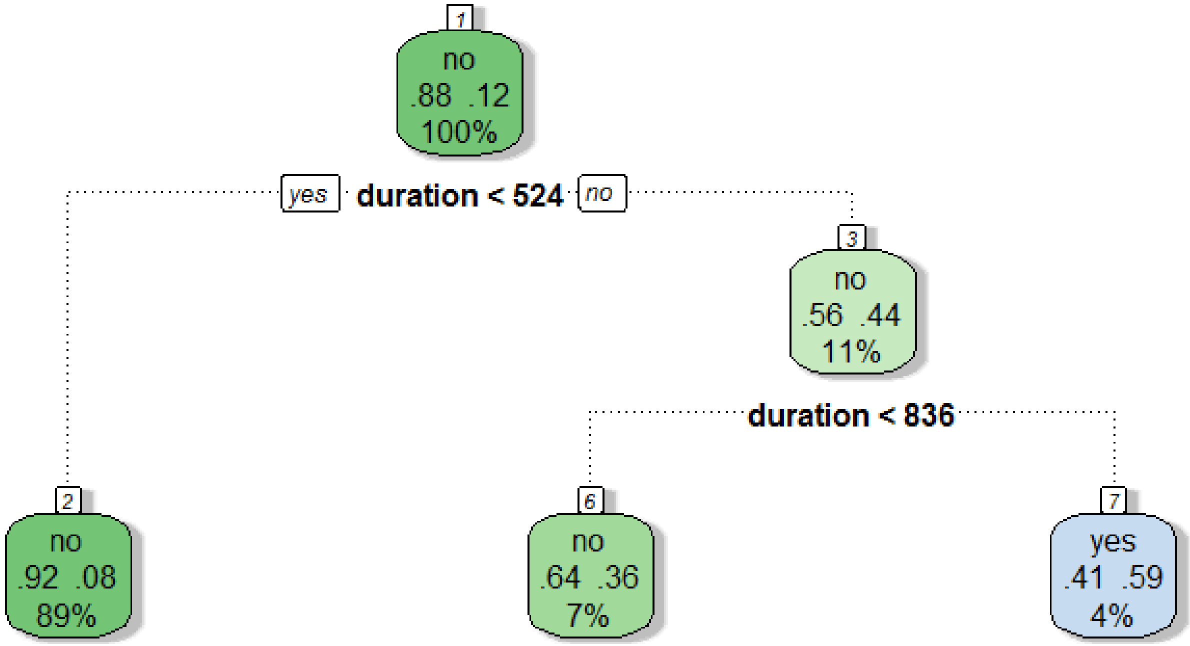
```
  Surrogate splits:
      previous < 17.5  to the left,  agree=0.649, adj=0.001, (0 split)

Node number 6: 2227 observations
  predicted class=no   expected loss=0.3623709  P(node) =0.07036557
    class counts:  1420   807
   probabilities: 0.638 0.362

Node number 7: 1207 observations
  predicted class=yes  expected loss=0.4084507  P(node) =0.03813707
    class counts:   493   714
   probabilities: 0.408 0.592
```

Testing the decision tree:
Predictions table:

```
predictions
   no   yes
13037   525
```

Confusion matrix:

```
predictions         no        yes
        no  0.86617018 0.09511871
        yes 0.01688542 0.02182569


Confusion Matrix and Statistics

          Reference
Prediction    no   yes
       no  11747  1290
       yes   229   296

               Accuracy : 0.888
                 95% CI : (0.8826, 0.8933)
    No Information Rate : 0.8831
    P-Value [Acc > NIR] : 0.03717

                  Kappa : 0.236
 Mcnemar's Test P-Value : < 2e-16

            Sensitivity : 0.9809
            Specificity : 0.1866
         Pos Pred Value : 0.9011
         Neg Pred Value : 0.5638
             Prevalence : 0.8831
         Detection Rate : 0.8662
   Detection Prevalence : 0.9613
      Balanced Accuracy : 0.5838

       'Positive' Class : no
```



Rattle 2017-Oct-25 11:49:19 gupta

Application of the Project

This project has been designed to predict the efficacy of the marketing campaign of a bank. It can be used in similar scenarios, i.e., for prediction of customer churning. The churn rate, also known as the rate of attrition, is the percentage of subscribers to a service who discontinue their subscriptions to that service within a given time period. For a company to expand its clientele, its growth rate, as measured by the number of new customers, must exceed its churn rate.

Hence, this model can be an excellent predictor of customer churn.

Conclusion

The following conclusions have been derived through the project.

1. “Duration” has positive effect on people saying “yes”. This is because the longer the conversations on the phone, the higher interest the customer will show to the term deposit. The Bank ought to focus on the potential clients who have significant call duration.
2. Although the duration of the call plays a major role in the people's decision of subscribing or not, the out-bound calls might create a negative attitude towards the bank due to the intrusion of privacy. Bank should decrease the outbound call rate and use inbound calls for cross-selling intelligently to increase the duration of the call.
3. Bank may target clients of job category of housemaid, services, technician etc as these set of people are averse to taking risks and look for safe deposit of their savings with fixed returns.
4. To improve their lead generation, banks may hire more people or develop analytic solution, as an alternative. This would improve the quality of conversation as agents would be spending more time with selective clients only.

References

1. A Complete Tutorial to learn Data Science in R from Scratch, Analytics Vidhya.

2. Analytics Edge MOOC by EdX.