Bank Marketing for Term Deposits

Author: Anush Harish (21250164)

Supervisor: Dr Galatia Cleanthous

Department of Mathematics and Statistics

National University of Ireland, Maynooth

A thesis submitted in fulfillment of the requirements for the degree of MSc in Data Science and Analytics

2021-2022

Declaration

I hereby declare that the work described in this dissertation is, except where otherwise stated, entirely my own work.

Anush Harish (21250164)
August 2022

Acknowledgement

I gathered knowledge and ingenious ideas from many sources and brainstormed with many people to write this thesis. Author acknowledges all those directly or indirectly involved with this project for the support they provided during project execution.

It is with great gratitude that the author acknowledges his supervisor and mentor, Dr. Galatia Cleanthous, for providing him with the opportunity to work under her guidance and for all the guidance she provided throughout the project. She has a lot of knowledge and wisdom. This project would not have been a success without her critical analytical skills and invaluable input. Throughout the course, she motivated the author and helped him develop new approaches to the issues.

His heartfelt gratitude goes out to his friend and faculty members for the endless brainstorming sessions that helped the author refine the process.

Abstract

A bank's primary source of revenue is term deposits. An investment held with a bank for a short period of time is called a term deposit. A fixed rate of interest is paid on money invested for a fixed period of time. In order to offer term deposits, the bank uses a variety of marketing techniques, including email marketing, advertisements, telephonic marketing, and digital marketing. Telemarketing campaigns prove to be one of the most effective methods of reaching individuals. The telemarketing campaigns dataset is from UCI machine learning repository issued by a Portuguese banking institution.

Financial institutions need to identify which groups of customers are most likely to take advantage of term deposits. Similarly, this study considered the typical case of a bank direct marketing campaign dataset with two main objectives. For the first step, a variety of models are compared for accuracy, confusion matrix, and area under receiver operating characteristic curves (ROC) to determine which model is most suitable to fix the problem. Identifying the key characteristics of customers likely to subscribe to term deposits was the second objective of the study. With a prediction accuracy of 90.24%, Random Forest Classifier is the most prolific classifier in the study. Also, euribor 3m rate, number of employees, and job were the main key characteristics.

Keywords : Machine Learning, Statistical learning, Classification, Logistic Regression, Naive bayes, Random forest, Confusion matrix, Accuracy, Area under the ROC curve.

Table of Contents

List of Figures

List of Tables

Section 1

Introduction

Term deposits are a type of savings product where customers deposit money with a bank for a certain length of time. In other words, a term deposit is a loan that customers give to the bank for a specific period of time. As soon as the specified term is over, the bank returns your funds with the interest that has accrued based on the period and the amount placed. The consumer may not be interested in term deposits at all, thus targeting him/her is a waste of time and money. By classifying customers into high and low potentials, a predictive model can save expenses and target customers who are likely to subscribe. Telemarketing data has been provided by a Portuguese banking institution. As part of the marketing, customers were approached via phone to offer term deposit subscriptions.

The data is available on the UCI Machine Learning repository. There are 41188 rows and 21 attributes in the data set. Term deposits are either subscribed to or not subscribed to by the customer. Approximately 88% of the customers in the data set did not subscribe to a term deposit. A team of principal investigators is responsible for collecting the data, including S. Moro, P. Cortez and P. Rita (UCI Machine Learning Repository).

In order to understand which attributes are related to the outcome, exploratory analysis and feature engineering are conducted, as well as Chi-square tests of independence for categorical attributes. The numerical attributes are tested using the T-test (Guyon and Elisseeff). Missing values are dropped or transformed if they are related to the outcome. Once data is preprocessed different classification algorithms are developed, compared, and selected as the best predictive model. In order to enable the model to perform as accurately and efficiently as possible, different machine learning methods were used, including Logistic Regression, Naive Bayes, and Random Forest.

Attribute Information

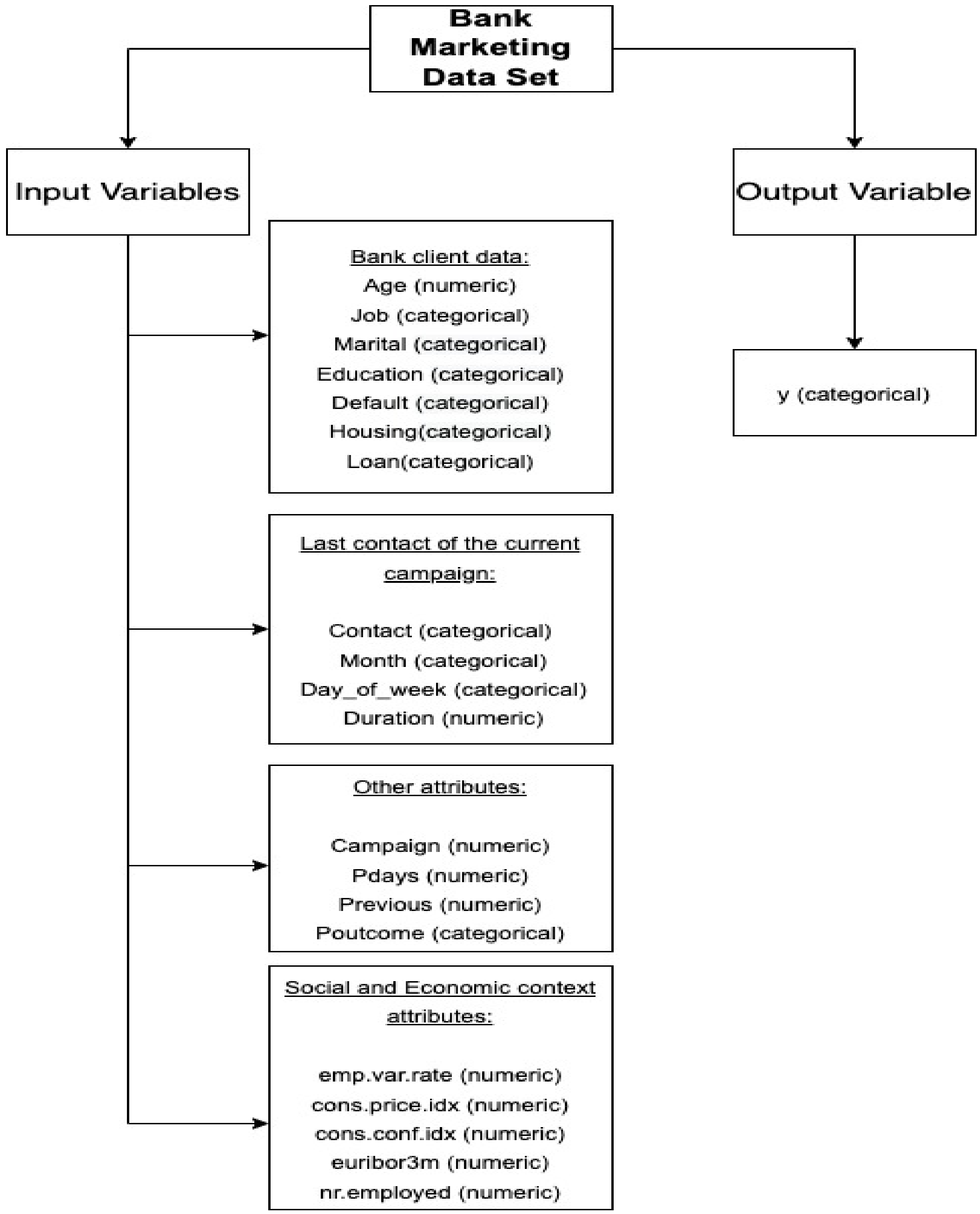

Fig 1.1 Structure of Dataset

Input variables

Bank Client Data:

1. Age : The age of the customers.
2. Job : Type of job - admin, blue-collar, entrepreneur, housemaid, management, retired, self employed, services, student, technician, unemployed, unknown.
3. Marital : Relationship status - married, divorced or widowed, single.
4. Education : Education level of the customer - basic.4y, basic.6y, basic.9y, high.school, illiterate, professional.course, university.degree, unknown.
5. Default: Does the customer have a credit default? - yes, no, unknown.
6. Housing: Does the customer have a mortgage loan? - no, yes, unknown.
7. Loan: Does the customer have a personal loan? - no, yes, unknown.

Last contact of the current campaign:

1. Contact: type of communication - cellular, telephone.
2. Month: Month of last contact - jan, feb, mar, apr, may, jun, jul, aug, sep, oct, nov, dec.
3. Day_of_week: Last contact day - mon, tue, wed, thu, fri, sat, sun.
4. Duration: The duration of the last contact, in seconds.

Other attributes:

1. Customer: The number of contacts made during this campaign.
2. Pdays: The number of days since the customer was last contacted by a previous campaign.
3. Previous: The number of contacts made before this campaign.
4. Poutcome: Results of previous marketing campaign - failure, nonexistent, success.

Social and economic context attributes:

1. Emp.var.rate: variation in employment - quarterly indicator.
2. Cons.price.idx: Index of consumer prices - monthly indicator.
3. Cons.conf.idx: Index of consumer confidence - monthly indicator.
4. Euribor3m: Three-month Euribor rate - daily indicator.
5. Nr.employed: employee count - quarterly indicator.

Output variable:

1. Y: Has the customer subscribed to a term deposit? - Yes, No. (UCI Machine Learning Repository)

Section 2

Literature review

According to (Moro et al.) due to the global financial crisis, banks' access to credit is restricted and they are instead concentrating on collecting money from their customers. Therefore, gathering such data and offering services in accordance with it may be really helpful to achieve effective marketing campaigns. To examine and enhance an institution's marketing capabilities, factors including behavior, psychology, mindset, and motivation must be taken into account (Raorane and R.V.Kulkarni).

In (Suebsing and Vajiramedhin) has many firms analyze the data from their prior customers before providing their services to new customers in order to make decisions that would prevent campaign failure. Predicting customer bank data can aid in the discovery of hidden trends and aid in the success of marketing initiatives. The selection of variables and features has become a major research topic in areas of application where datasets with tens of thousands or even millions of variables are accessible. In variable selection, three objectives are pursued: improving predictor performance, making predictors faster and more affordable, and exploring the underlying processes that produced the data (Guyon and Elisseeff).

Managing and maintaining customer data may help in identifying patterns and trends that can be used to develop new thoughts for attracting new customers. The capacity of machine learning to extract meaningful patterns from data is improved, and the use of data mining techniques in the banking industry is rapidly expanding (Ajay et al.). Machine learning algorithms may be used to do classifications, which can be used to classify the data into various classes (Radhakrishnan B et al.).

Using the CRISP-DM approach, (Moro et al.) analyzed a Portuguese bank's direct marketing dataset using data mining. The goal of their study was to develop a predictive model for enhancing direct marketing efforts by minimizing phone calls.

(Apampa) found that the balanced dataset with 17 attributes produced more accurate outcomes than the original unbalanced dataset. Based on the AUC value, Decision Trees outperformed Naive Bayes and Logistic Regression in the model.

A machine learning algorithm can be used to find different patterns in data, according to (Bishop). This analysis aims to identify the customers with the highest probability of applying for a long-term deposit with the bank. In order to determine whether a consumer is interested in placing a term deposit, banks can utilize various machine learning approaches. The R programming language has been used to implement three machine learning techniques. Due to its flexibility, Logistic Regression can be applied to arbitrary data sets. The class of the test data set can be predicted easily and quickly using Naive Bayes. The Random Forest algorithm is more accurate at predicting outcomes than decision trees.

The aim in (Parlar and Acaravci) was to define the relevant features for increasing the effectiveness of Bank Telemarketing in introducing term deposits to customers. In this case, Chi-square and Information Gain were used. The precision and recall measures were used in what seemed like a mini-evaluation. Concluded that classification performance was improved by reducing the number of attributes.

When compared using Accuracy, all of the classification methods listed above produced superior results. Different authors independently reached the results after applying unique characteristics to optimize the classification algorithms for the bank telemarketing dataset. The classification error, sensitivity, specificity, and accuracy were used to evaluate performance. However, accuracy was the metric that all authors used the most frequently. This study focuses on the algorithms with more accuracy and AUC value. The implementation section includes descriptions of each approach.

Section 3

Exploratory Analysis and Feature Engineering

An instrumental part of any Data Science project is Feature Engineering and EDA (Exploratory Data Analytics). Machine learning algorithms require features that have some specific characteristics in order to work effectively. These techniques improve the performance of our simple models. Initially, the data is in a raw format. These features must first be extracted from data before they can be used in algorithms. It is called Feature Engineering. Having relevant features in hand reduces the complexity of algorithms. The results are accurate, even if a non-ideal algorithm is used to solve the problem. Feature Engineering has two primary objectives. The input data should be prepared so that it is compatible with the constraints of the machine learning algorithm.

A machine learning model performs better when it is enhanced. There are many techniques used in Feature engineering, such as imputation, binning, outlier handling, log transformation, and extracting data.

Crosstable is a package based on a single function, crosstable(), that computes descriptive statistics on datasets quickly and conveniently. For categorical attributes, Chi-square tests of independence are used to determine which attributes are related to the outcome. In this test, two categorical variables are tested to see if they are independent or not. Changing the value of one categorical variable does not change its probability distribution if two categorical variables are independent. In R this test is performed using the chisq.test(data) function. Numeric attributes are tested with a T-test. Missing values associated with outcomes are dropped or transformed. T-tests are used to compare the means of two groups in statistics. A hypothesis test is used to determine whether an intervention affects the population of interest, or whether two groups differ. The t.test(data) function is used in R to perform this test.

Bank Client Data

1. Age : Who exactly was contacted as part of this marketing campaign?

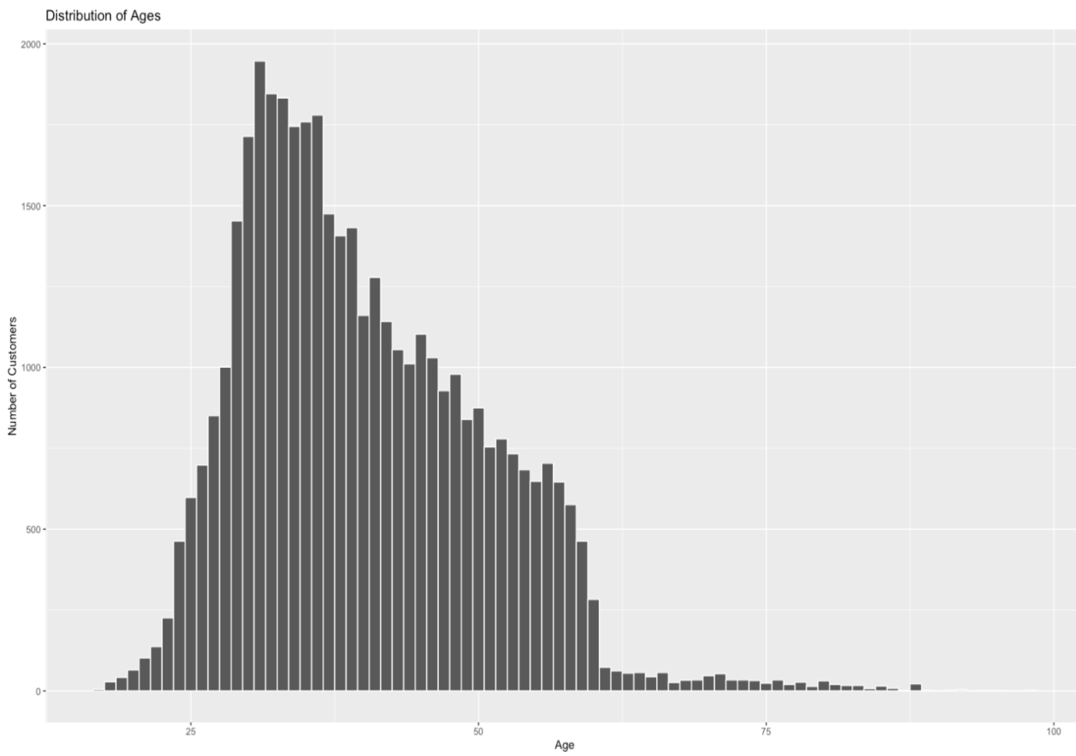

Fig 3.1 Distribution of Age

The distribution of age is from 17 to 98. 50% of customers are between the ages of 32 and 47. With an average of 40 and median of 38. Customers aged below 30 are categorized as Young_aged, ages from 30 to 60 are categorized as Middle aged, and customers aged above 60 are categorized as Old_aged (see Fig 3.1). So, variable age is converted from numerical to categorical.

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  41188


              | Bank_Data$y
Bank_Data$age |        no |       yes | Row Total |
--------------|-----------|-----------|-----------|
  Middle_aged |     31305 |      3304 |     34609 |
              |    11.522 |    90.756 |           |
              |     0.905 |     0.095 |     0.840 |
              |     0.857 |     0.712 |           |
              |     0.760 |     0.080 |           |
--------------|-----------|-----------|-----------|
     Old_aged |       496 |       414 |       910 |
              |   120.154 |   946.422 |           |
              |     0.545 |     0.455 |     0.022 |
              |     0.014 |     0.089 |           |
              |     0.012 |     0.010 |           |
--------------|-----------|-----------|-----------|
   Young_aged |      4747 |       922 |      5669 |
              |    15.962 |   125.729 |           |
              |     0.837 |     0.163 |     0.138 |
              |     0.130 |     0.199 |           |
              |     0.115 |     0.022 |           |
--------------|-----------|-----------|-----------|
 Column Total |     36548 |      4640 |     41188 |
              |     0.887 |     0.113 |           |
--------------|-----------|-----------|-----------|


        Pearson's Chi-squared test

data:  Bank_Data$age and Bank_Data$y
X-squared = 1310.5, df = 2, p-value < 2.2e-16
```

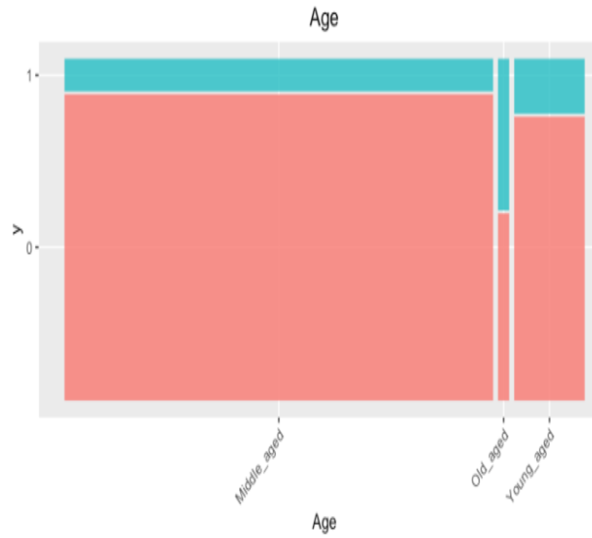

Fig 3.2 Proportions of each levels in Age

Since p-value is less than 0.05. It can be concluded that variable age is significant with the response variable i.e., y. The percentage of customers over 60 who subscribe to a term deposit is nearly 45.5%, which is higher than the percentage of younger individuals is 16.3% and the percentage of customers aged 30 to 60 is 9.5% (see Fig 3.2).

2. Job : What types of jobs does the customer pool represent?

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  41188


               | Bank_Data$y
Bank_Data$job  |        no |       yes | Row Total |
---------------|-----------|-----------|-----------|
        admin. |      9070 |      1352 |     10422 |
               |     3.423 |    26.961 |           |
               |     0.870 |     0.130 |     0.253 |
               |     0.248 |     0.291 |           |
               |     0.220 |     0.033 |           |
---------------|-----------|-----------|-----------|
   blue-collar |      8616 |       638 |      9254 |
               |    19.926 |   156.951 |           |
               |     0.931 |     0.069 |     0.225 |
               |     0.236 |     0.138 |           |
               |     0.209 |     0.015 |           |
---------------|-----------|-----------|-----------|
  entrepreneur |      1332 |       124 |      1456 |
               |     1.240 |     9.767 |           |
               |     0.915 |     0.085 |     0.035 |
               |     0.036 |     0.027 |           |
               |     0.032 |     0.003 |           |
---------------|-----------|-----------|-----------|
     housemaid |       954 |       106 |      1060 |
               |     0.191 |     1.507 |           |
               |     0.900 |     0.100 |     0.026 |
               |     0.026 |     0.023 |           |
               |     0.023 |     0.003 |           |
---------------|-----------|-----------|-----------|
    management |      2596 |       328 |      2924 |
               |     0.001 |     0.006 |           |
               |     0.888 |     0.112 |     0.071 |
               |     0.071 |     0.071 |           |
               |     0.063 |     0.008 |           |
---------------|-----------|-----------|-----------|
       retired |      1286 |       434 |      1720 |
               |    37.814 |   297.849 |           |
               |     0.748 |     0.252 |     0.042 |
               |     0.035 |     0.094 |           |
               |     0.031 |     0.011 |           |
---------------|-----------|-----------|-----------|
 self-employed |      1272 |       149 |      1421 |
               |     0.097 |     0.767 |           |
               |     0.895 |     0.105 |     0.035 |
               |     0.035 |     0.032 |           |
               |     0.031 |     0.004 |           |
---------------|-----------|-----------|-----------|
      services |      3646 |       323 |      3969 |
               |     4.375 |    34.458 |           |
               |     0.919 |     0.081 |     0.096 |
               |     0.100 |     0.070 |           |
               |     0.089 |     0.008 |           |
---------------|-----------|-----------|-----------|
       student |       600 |       275 |       875 |
               |    40.090 |   315.775 |           |
               |     0.686 |     0.314 |     0.021 |
               |     0.016 |     0.059 |           |
               |     0.015 |     0.007 |           |
---------------|-----------|-----------|-----------|
    technician |      6013 |       730 |      6743 |
               |     0.147 |     1.156 |           |
               |     0.892 |     0.108 |     0.164 |
               |     0.165 |     0.157 |           |
               |     0.146 |     0.018 |           |
---------------|-----------|-----------|-----------|
    unemployed |       870 |       144 |      1014 |
               |     0.985 |     7.758 |           |
               |     0.858 |     0.142 |     0.025 |
               |     0.024 |     0.031 |           |
               |     0.021 |     0.003 |           |
---------------|-----------|-----------|-----------|
       unknown |       293 |        37 |       330 |
               |     0.000 |     0.001 |           |
               |     0.888 |     0.112 |     0.008 |
               |     0.008 |     0.008 |           |
               |     0.007 |     0.001 |           |
---------------|-----------|-----------|-----------|
  Column Total |     36548 |      4640 |     41188 |
               |     0.887 |     0.113 |           |
---------------|-----------|-----------|-----------|


        Pearson's Chi-squared test

data:  Bank_Data$job and Bank_Data$y
X-squared = 961.24, df = 11, p-value < 2.2e-16
```

The "unknown" level in the data with a proportion of 0.008 so it should be removed because it doesn't provide any significant information. Rows with this value in the "job" column will be eliminated. Job variable is significant with the response variable because its p-value is less than 0.05.

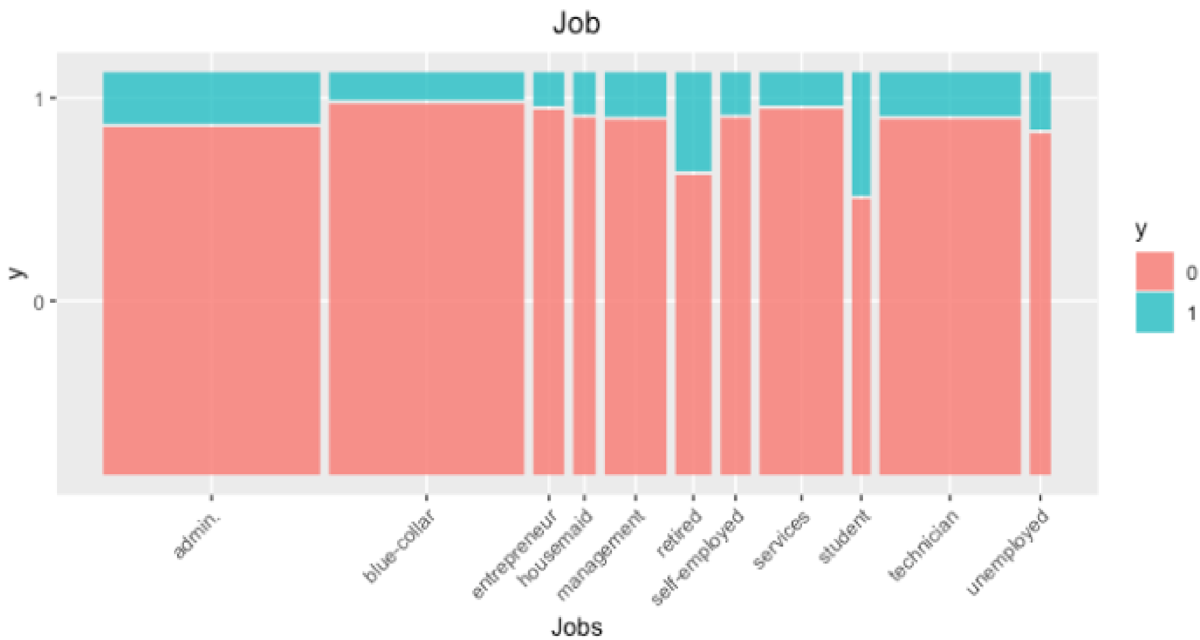

Fig 3.3 Proportions of each levels in Job

Students are the group that shows the greatest frequency of subscriptions to term deposits, with 31.4%. Term deposit subscriptions are highest among retired customers with 25.2% and unemployed with 14.2% (see Fig 3.3).

3. Marital : How are the customers' marital situations?

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  40858


                 | Bank_Data$y
Bank_Data$marital |        no |       yes | Row Total |
-----------------|-----------|-----------|-----------|
        divorced |      4126 |       473 |      4599 |
                 |     0.499 |     3.929 |           |
                 |     0.897 |     0.103 |     0.113 |
                 |     0.114 |     0.103 |           |
                 |     0.101 |     0.012 |           |
-----------------|-----------|-----------|-----------|
         married |     22178 |      2516 |     24694 |
                 |     3.229 |    25.431 |           |
                 |     0.898 |     0.102 |     0.604 |
                 |     0.612 |     0.547 |           |
                 |     0.543 |     0.062 |           |
-----------------|-----------|-----------|-----------|
          single |      9889 |      1605 |     11494 |
                 |     9.429 |    74.264 |           |
                 |     0.860 |     0.140 |     0.281 |
                 |     0.273 |     0.349 |           |
                 |     0.242 |     0.039 |           |
-----------------|-----------|-----------|-----------|
         unknown |        62 |         9 |        71 |
                 |     0.016 |     0.125 |           |
                 |     0.873 |     0.127 |     0.002 |
                 |     0.002 |     0.002 |           |
                 |     0.002 |     0.000 |           |
-----------------|-----------|-----------|-----------|
    Column Total |     36255 |      4603 |     40858 |
                 |     0.887 |     0.113 |           |
-----------------|-----------|-----------|-----------|


        Pearson's Chi-squared test

data:  Bank_Data$marital and Bank_Data$y
X-squared = 116.92, df = 3, p-value < 2.2e-16
```

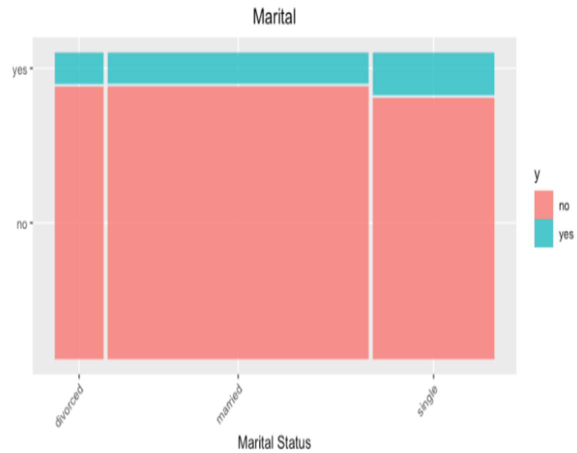

Fig 3.4 Proportions of each levels in Marital Status

The "unknown" level in the data with a proportion of 0.002. So, it should be removed because it doesn't provide any significant information. Rows with this value in the "marital" column will be eliminated. Marital variable is significant with the response variable because its p-value is less than 0.05.Single's show a high subscription rate of 14%. Both married and divorced have almost the same subscription rate of 10% (see fig 3.4).

4. Education : What is the educational qualification of customers?

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  40787


                    | Bank_Data$y
Bank_Data$education |        no |       yes | Row Total |
--------------------|-----------|-----------|-----------|
           basic.4y |      3695 |       423 |      4118 |
                    |     0.456 |     3.594 |           |
                    |     0.897 |     0.103 |     0.101 |
                    |     0.102 |     0.092 |           |
                    |     0.091 |     0.010 |           |
--------------------|-----------|-----------|-----------|
           basic.6y |      2077 |       187 |      2264 |
                    |     2.302 |    18.135 |           |
                    |     0.917 |     0.083 |     0.056 |
                    |     0.057 |     0.041 |           |
                    |     0.051 |     0.005 |           |
--------------------|-----------|-----------|-----------|
           basic.9y |      5536 |       470 |      6006 |
                    |     8.000 |    63.023 |           |
                    |     0.922 |     0.078 |     0.147 |
                    |     0.153 |     0.102 |           |
                    |     0.136 |     0.012 |           |
--------------------|-----------|-----------|-----------|
        high.school |      8436 |      1028 |      9464 |
                    |     0.172 |     1.352 |           |
                    |     0.891 |     0.109 |     0.232 |
                    |     0.233 |     0.224 |           |
                    |     0.207 |     0.025 |           |
--------------------|-----------|-----------|-----------|
         illiterate |        14 |         4 |        18 |
                    |     0.244 |     1.919 |           |
                    |     0.778 |     0.222 |     0.000 |
                    |     0.000 |     0.001 |           |
                    |     0.000 |     0.000 |           |
--------------------|-----------|-----------|-----------|
professional.course |      4631 |       594 |      5225 |
                    |     0.006 |     0.051 |           |
                    |     0.886 |     0.114 |     0.128 |
                    |     0.128 |     0.129 |           |
                    |     0.114 |     0.015 |           |
--------------------|-----------|-----------|-----------|
```

university.degree	10442	1654	12096
	7.921	62.403	
	0.863	0.137	0.297
	0.289	0.360	
	0.256	0.041	
unknown	1362	234	1596
	2.077	16.364	
	0.853	0.147	0.039
	0.038	0.051	
	0.033	0.006	
Column Total	36193	4594	40787
	0.887	0.113	

The “unknown” level contribution for subscription is high. So, the unknown level was changed to university.degree because it was the highest contribution of 29.7%. And illiterate level in the data with a proportion of 0.000. So, it should be removed because it doesn't provide any significant information. Rows with illiterate value in the "education" column will be eliminated. An education variable is significant with the response variable because its p-value is less than 0.05.

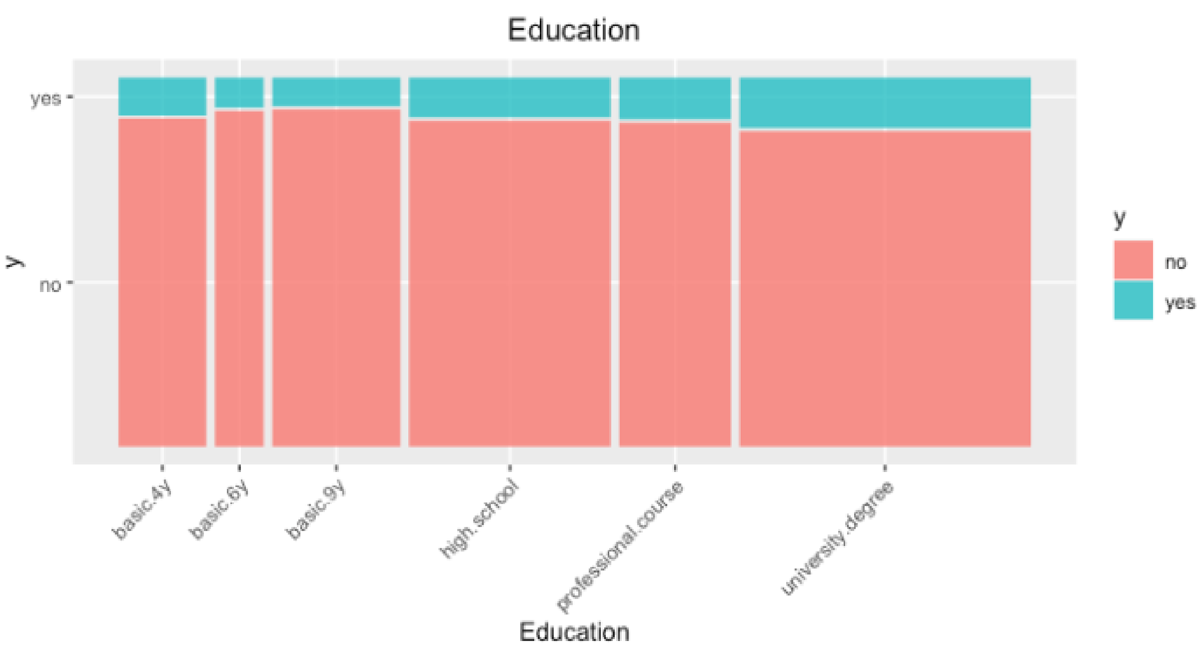

Fig 3.5 Proportions of each levels in Education

From Fig 3.5 it is observed that as the number of years in education increases the term deposit subscription also increases.

5. Default : Is the customer's credit in default?

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  40769


                  | Bank_Data$y
Bank_Data$default |        no |       yes | Row Total |
------------------|-----------|-----------|-----------|
               no |     28182 |      4155 |     32337 |
                  |     9.218 |    72.659 |           |
                  |     0.872 |     0.128 |     0.793 |
                  |     0.779 |     0.905 |           |
                  |     0.691 |     0.102 |           |
------------------|-----------|-----------|-----------|
          unknown |      7994 |       435 |      8429 |
                  |    35.318 |   278.381 |           |
                  |     0.948 |     0.052 |     0.207 |
                  |     0.221 |     0.095 |           |
                  |     0.196 |     0.011 |           |
------------------|-----------|-----------|-----------|
              yes |         3 |         0 |         3 |
                  |     0.043 |     0.338 |           |
                  |     1.000 |     0.000 |     0.000 |
                  |     0.000 |     0.000 |           |
                  |     0.000 |     0.000 |           |
------------------|-----------|-----------|-----------|
     Column Total |     36179 |      4590 |     40769 |
                  |     0.887 |     0.113 |           |
------------------|-----------|-----------|-----------|


        Pearson's Chi-squared test

data:  Bank_Data$default and Bank_Data$y
X-squared = 395.96, df = 2, p-value < 2.2e-16
```

Even though the default variable is significant to the response variable it is removed from the dataset because it has only 3 observations for yes and no as 79.1%, unknowns as 20.9%.

6. Housing : Is the customer in possession of a mortgage?

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  40769


                  | Bank_Data$y
Bank_Data$housing |        no |       yes | Row Total |
------------------|-----------|-----------|-----------|
               no |     16416 |      2003 |     18419 |
                  |     0.306 |     2.411 |           |
                  |     0.891 |     0.109 |     0.452 |
                  |     0.454 |     0.436 |           |
                  |     0.403 |     0.049 |           |
------------------|-----------|-----------|-----------|
          unknown |       877 |       107 |       984 |
                  |     0.016 |     0.129 |           |
                  |     0.891 |     0.109 |     0.024 |
                  |     0.024 |     0.023 |           |
                  |     0.022 |     0.003 |           |
------------------|-----------|-----------|-----------|
              yes |     18886 |      2480 |     21366 |
                  |     0.293 |     2.307 |           |
                  |     0.884 |     0.116 |     0.524 |
                  |     0.522 |     0.540 |           |
                  |     0.463 |     0.061 |           |
------------------|-----------|-----------|-----------|
     Column Total |     36179 |      4590 |     40769 |
                  |     0.887 |     0.113 |           |
------------------|-----------|-----------|-----------|


        Pearson's Chi-squared test

data:  Bank_Data$housing and Bank_Data$y
X-squared = 5.4627, df = 2, p-value = 0.06513
```

Since p-value is greater than 0.05, the housing variable is not significant for the response variable y. So, housing is removed from the dataset.

7. Loan : Is the customer in possession of a personal loan?

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  40769


               | Bank_Data$y
Bank_Data$loan |        no |       yes | Row Total |
---------------|-----------|-----------|-----------|
            no |     29799 |      3806 |     33605 |
               |     0.017 |     0.135 |           |
               |     0.887 |     0.113 |     0.824 |
               |     0.824 |     0.829 |           |
               |     0.731 |     0.093 |           |
---------------|-----------|-----------|-----------|
       unknown |       877 |       107 |       984 |
               |     0.016 |     0.129 |           |
               |     0.891 |     0.109 |     0.024 |
               |     0.024 |     0.023 |           |
               |     0.022 |     0.003 |           |
---------------|-----------|-----------|-----------|
           yes |      5503 |       677 |      6180 |
               |     0.064 |     0.507 |           |
               |     0.890 |     0.110 |     0.152 |
               |     0.152 |     0.147 |           |
               |     0.135 |     0.017 |           |
---------------|-----------|-----------|-----------|
  Column Total |     36179 |      4590 |     40769 |
               |     0.887 |     0.113 |           |
---------------|-----------|-----------|-----------|
```

```
        Pearson's Chi-squared test

data:  Bank_Data$loan and Bank_Data$y
X-squared = 0.86841, df = 2, p-value = 0.6478
```

Since p-value is greater than 0.05, the loan variable is not significant for the response variable y. So, the loan variable is removed from the dataset.

Last contact of the current campaign

1. Contact : How did the bank get in touch with the customer?

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  40769


                 | Bank_Data$y
Bank_Data$contact |        no |       yes | Row Total |
------------------|-----------|-----------|-----------|
         cellular |     22098 |      3815 |     25913 |
                  |    35.034 |   276.145 |           |
                  |     0.853 |     0.147 |     0.636 |
                  |     0.611 |     0.831 |           |
                  |     0.542 |     0.094 |           |
------------------|-----------|-----------|-----------|
        telephone |     14081 |       775 |     14856 |
                  |    61.110 |   481.674 |           |
                  |     0.948 |     0.052 |     0.364 |
                  |     0.389 |     0.169 |           |
                  |     0.345 |     0.019 |           |
------------------|-----------|-----------|-----------|
     Column Total |     36179 |      4590 |     40769 |
                  |     0.887 |     0.113 |           |
------------------|-----------|-----------|-----------|


        Pearson's Chi-squared test with Yates' continuity correction

data:  Bank_Data$contact and Bank_Data$y
X-squared = 853.01, df = 1, p-value < 2.2e-16
```

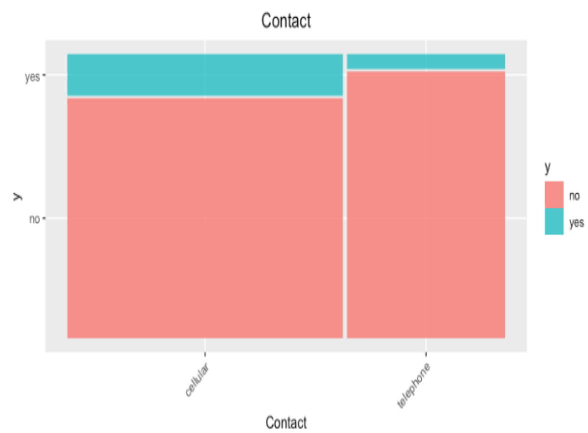

Fig 3.6 Proportions of each levels in Contact

Since p-value is less than 0.05. It can be concluded that variable contact is significant with the response variable i.e., y. The percentage of cellular responders who subscribed to term deposits was nearly 14.7%, and the percentage of telephone responders was 5.2% (see Fig 3.6).

2. Month : In which month customers are contacted?

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  40769


                | Bank_Data$y
Bank_Data$month |        no |       yes | Row Total |
----------------|-----------|-----------|-----------|
            mar |       267 |       274 |       541 |
                |    94.582 |   745.506 |           |
                |     0.494 |     0.506 |     0.013 |
                |     0.007 |     0.060 |           |
                |     0.007 |     0.007 |           |
----------------|-----------|-----------|-----------|
            apr |      2082 |       536 |      2618 |
                |    25.052 |   197.463 |           |
                |     0.795 |     0.205 |     0.064 |
                |     0.058 |     0.117 |           |
                |     0.051 |     0.013 |           |
----------------|-----------|-----------|-----------|
            may |     12734 |       882 |     13616 |
                |    35.070 |   276.428 |           |
                |     0.935 |     0.065 |     0.334 |
                |     0.352 |     0.192 |           |
                |     0.312 |     0.022 |           |
----------------|-----------|-----------|-----------|
            jun |      4697 |       548 |      5245 |
                |     0.388 |     3.060 |           |
                |     0.896 |     0.104 |     0.129 |
                |     0.130 |     0.119 |           |
                |     0.115 |     0.013 |           |
----------------|-----------|-----------|-----------|
            jul |      6471 |       642 |      7113 |
                |     3.996 |    31.498 |           |
                |     0.910 |     0.090 |     0.174 |
                |     0.179 |     0.140 |           |
                |     0.159 |     0.016 |           |
----------------|-----------|-----------|-----------|
            aug |      5459 |       644 |      6103 |
                |     0.343 |     2.705 |           |
                |     0.894 |     0.106 |     0.150 |
                |     0.151 |     0.140 |           |
                |     0.134 |     0.016 |           |
----------------|-----------|-----------|-----------|
            sep |       309 |       253 |       562 |
                |    72.176 |   568.904 |           |
                |     0.550 |     0.450 |     0.014 |
                |     0.009 |     0.055 |           |
                |     0.008 |     0.006 |           |
----------------|-----------|-----------|-----------|
            oct |       396 |       311 |       707 |
                |    85.347 |   672.717 |           |
                |     0.560 |     0.440 |     0.017 |
                |     0.011 |     0.068 |           |
                |     0.010 |     0.008 |           |
----------------|-----------|-----------|-----------|
            nov |      3672 |       412 |      4084 |
                |     0.630 |     4.969 |           |
                |     0.899 |     0.101 |     0.100 |
                |     0.101 |     0.090 |           |
                |     0.090 |     0.010 |           |
----------------|-----------|-----------|-----------|
            dec |        92 |        88 |       180 |
                |    28.722 |   226.395 |           |
                |     0.511 |     0.489 |     0.004 |
                |     0.003 |     0.019 |           |
                |     0.002 |     0.002 |           |
----------------|-----------|-----------|-----------|
   Column Total |     36179 |      4590 |     40769 |
                |     0.887 |     0.113 |           |
----------------|-----------|-----------|-----------|


        Pearson's Chi-squared test

data:  Bank_Data$month and Bank_Data$y
X-squared = 3076, df = 9, p-value < 2.2e-16
```

Since p-value is less than 0.05, it can be concluded that variable month is significant with the response variable i.e., y.

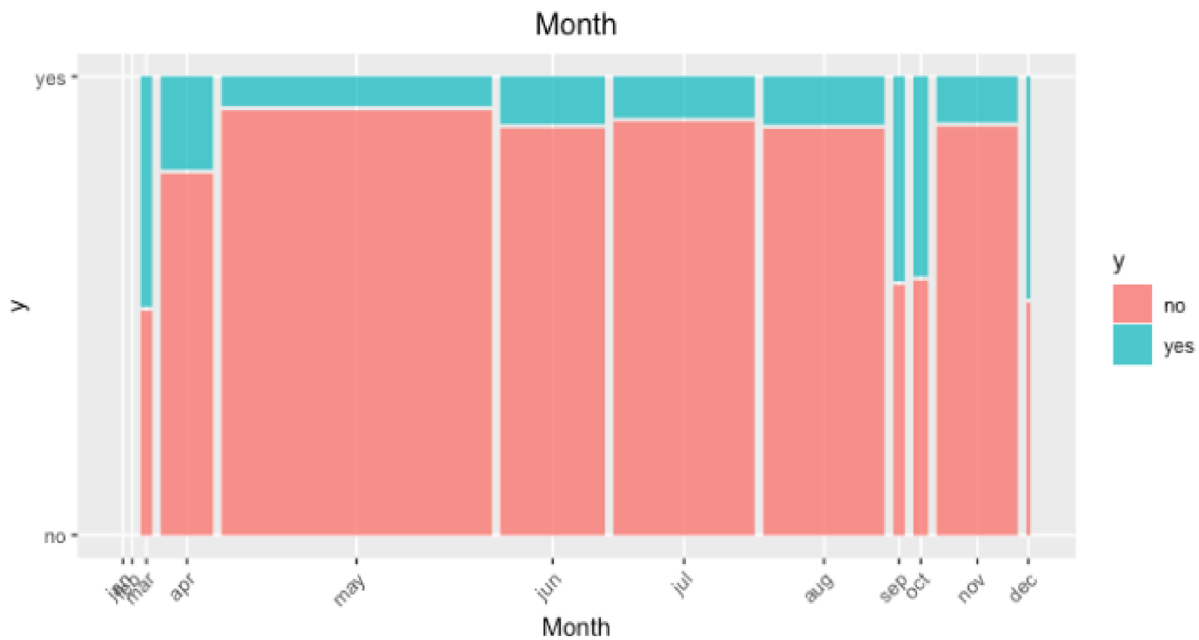

Fig 3.7 Proportions of each levels in Month

From Fig 3.7 it can be observed that there was no communication during January and February. The results are very strong for months with very low contact frequency, such as March, September, October, and December, with 44% to 51% of subscribers.

3. Day of the week : On what day of the week are customers contacted?

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  40769


                     | Bank_Data$y
Bank_Data$day_of_week |        no |       yes | Row Total |
---------------------|-----------|-----------|-----------|
                 mon |      7578 |       841 |      8419 |
                     |     1.528 |    12.047 |           |
                     |     0.900 |     0.100 |     0.207 |
                     |     0.209 |     0.183 |           |
                     |     0.186 |     0.021 |           |
---------------------|-----------|-----------|-----------|
                 tue |      7056 |       945 |      8001 |
                     |     0.275 |     2.169 |           |
                     |     0.882 |     0.118 |     0.196 |
                     |     0.195 |     0.206 |           |
                     |     0.173 |     0.023 |           |
---------------------|-----------|-----------|-----------|
                 wed |      7116 |       934 |      8050 |
                     |     0.107 |     0.846 |           |
                     |     0.884 |     0.116 |     0.197 |
                     |     0.197 |     0.203 |           |
                     |     0.175 |     0.023 |           |
---------------------|-----------|-----------|-----------|
                 thu |      7493 |      1031 |      8524 |
                     |     0.672 |     5.300 |           |
                     |     0.879 |     0.121 |     0.209 |
                     |     0.207 |     0.225 |           |
                     |     0.184 |     0.025 |           |
---------------------|-----------|-----------|-----------|
                 fri |      6936 |       839 |      7775 |
                     |     0.192 |     1.510 |           |
                     |     0.892 |     0.108 |     0.191 |
                     |     0.192 |     0.183 |           |
                     |     0.170 |     0.021 |           |
---------------------|-----------|-----------|-----------|
        Column Total |     36179 |      4590 |     40769 |
                     |     0.887 |     0.113 |           |
---------------------|-----------|-----------|-----------|
```

```
        Pearson's Chi-squared test

data:  Bank_Data$day_of_week and Bank_Data$y
X-squared = 24.646, df = 4, p-value = 5.925e-05
```

Since p-value is less than 0.05. It can be concluded that variable day of the week is significant with the response variable i.e., y.

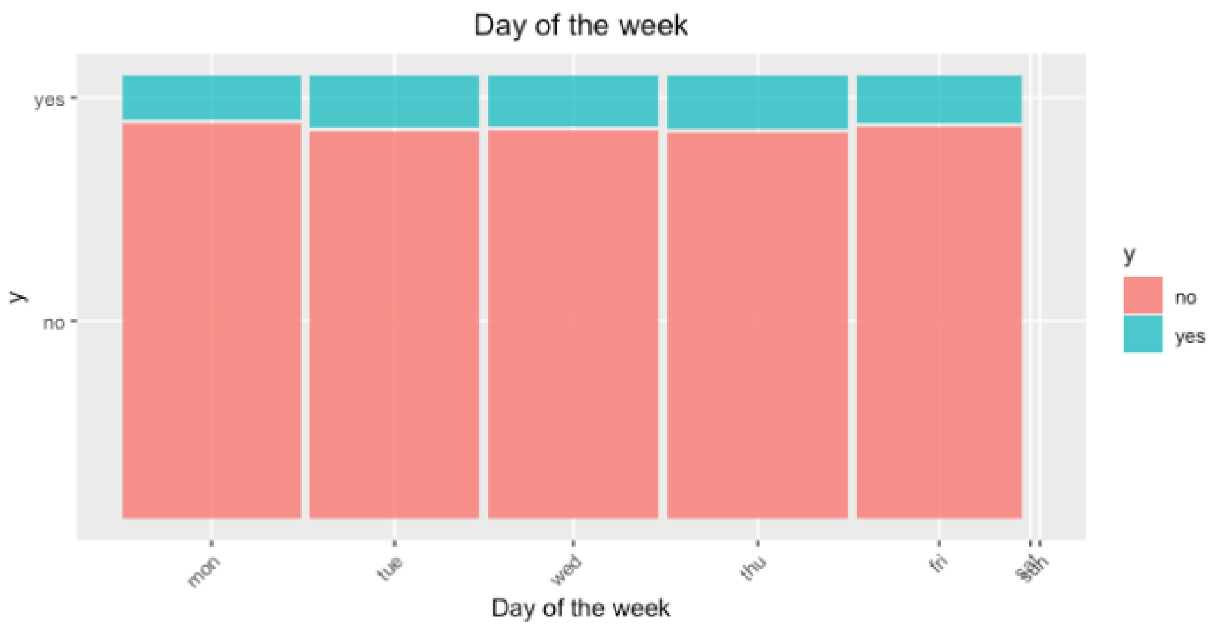

Fig 3.8 Proportions of each levels in Day of the week

From Fig 3.8 it is observed that weekend days are not used for making calls. Results tend to be better on Thursdays.

4. Duration: Before a call is made, the duration is not known. Y is also known after the call ends.

As a result, it is discarded.

Other Attributes

1. Campaign : How many times did the bank contact the customer throughout the campaign?

A numerical campaign was converted to a categorical campaign. During a single marketing campaign, calling the same person more than 6-7 times seems excessive (see Fig 3.9 and Fig 3.10).

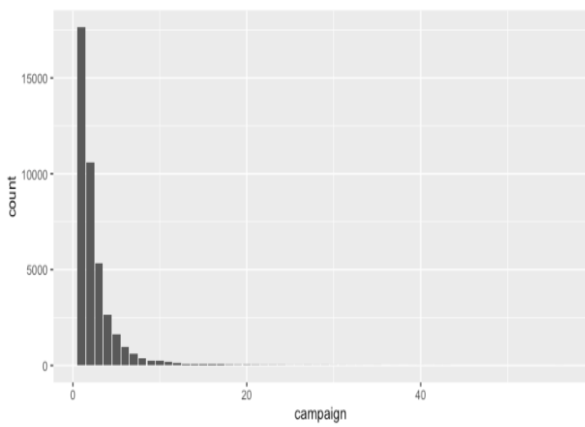

Fig 3.9 Distribution of Campaign

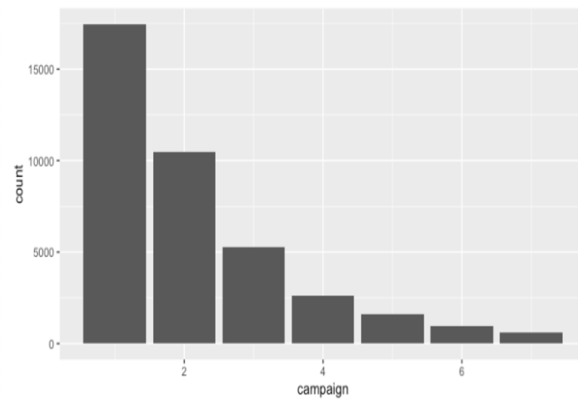

Fig 3.10 Distribution of Campaign after reduction

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  39019


                  | Bank_Data$y
Bank_Data$campaign |        no |       yes | Row Total |
-------------------|-----------|-----------|-----------|
                 1 |     15176 |      2268 |     17444 |
                   |     3.964 |    30.257 |           |
                   |     0.870 |     0.130 |     0.447 |
                   |     0.440 |     0.502 |           |
                   |     0.389 |     0.058 |           |
-------------------|-----------|-----------|-----------|
                 2 |      9280 |      1205 |     10485 |
                   |     0.010 |     0.076 |           |
                   |     0.885 |     0.115 |     0.269 |
                   |     0.269 |     0.267 |           |
                   |     0.238 |     0.031 |           |
-------------------|-----------|-----------|-----------|
                 3 |      4722 |       569 |      5291 |
                   |     0.412 |     3.146 |           |
                   |     0.892 |     0.108 |     0.136 |
                   |     0.137 |     0.126 |           |
                   |     0.121 |     0.015 |           |
-------------------|-----------|-----------|-----------|
                 4 |      2380 |       246 |      2626 |
                   |     1.459 |    11.134 |           |
                   |     0.906 |     0.094 |     0.067 |
                   |     0.069 |     0.054 |           |
                   |     0.061 |     0.006 |           |
-------------------|-----------|-----------|-----------|
                 5 |      1466 |       120 |      1586 |
                   |     2.896 |    22.102 |           |
                   |     0.924 |     0.076 |     0.041 |
                   |     0.042 |     0.027 |           |
                   |     0.038 |     0.003 |           |
-------------------|-----------|-----------|-----------|
                 6 |       892 |        74 |       966 |
                   |     1.682 |    12.838 |           |
                   |     0.923 |     0.077 |     0.025 |
                   |     0.026 |     0.016 |           |
                   |     0.023 |     0.002 |           |
-------------------|-----------|-----------|-----------|


-------------------|-----------|-----------|-----------|
                 7 |       583 |        38 |       621 |
                   |     2.098 |    16.010 |           |
                   |     0.939 |     0.061 |     0.016 |
                   |     0.017 |     0.008 |           |
                   |     0.015 |     0.001 |           |
-------------------|-----------|-----------|-----------|
      Column Total |     34499 |      4520 |     39019 |
                   |     0.884 |     0.116 |           |
-------------------|-----------|-----------|-----------|


        Pearson's Chi-squared test

data:  Bank_Data$campaign and Bank_Data$y
X-squared = 108.09, df = 6, p-value < 2.2e-16
```

Since p-value is less than 0.05. It can be concluded that variable campaign is significant with the response variable i.e.,y.

2. Pdays : How many days have passed since the consumer was contacted in a prior campaign?

Most of the values have 999. So, pdays are converted from numerical to categorical. If not contacted in pdays then 0(NO) else 1(YES). New column is added called cat_pdays with 0's and 1's and the existing column i.e., pdays is discarded.

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  39019


                   | Bank_Data$y
Bank_Data$cat_pdays |        no |       yes | Row Total |
-------------------|-----------|-----------|-----------|
                No |     33974 |      3564 |     37538 |
                   |    18.540 |   141.509 |           |
                   |     0.905 |     0.095 |     0.962 |
                   |     0.985 |     0.788 |           |
                   |     0.871 |     0.091 |           |
-------------------|-----------|-----------|-----------|
               Yes |       525 |       956 |      1481 |
                   |   469.930 |  3586.753 |           |
                   |     0.354 |     0.646 |     0.038 |
                   |     0.015 |     0.212 |           |
                   |     0.013 |     0.025 |           |
-------------------|-----------|-----------|-----------|
      Column Total |     34499 |      4520 |     39019 |
                   |     0.884 |     0.116 |           |
-------------------|-----------|-----------|-----------|


        Pearson's Chi-squared test with Yates' continuity correction

data:  Bank_Data$cat_pdays and Bank_Data$y
X-squared = 4211.4, df = 1, p-value < 2.2e-16
```

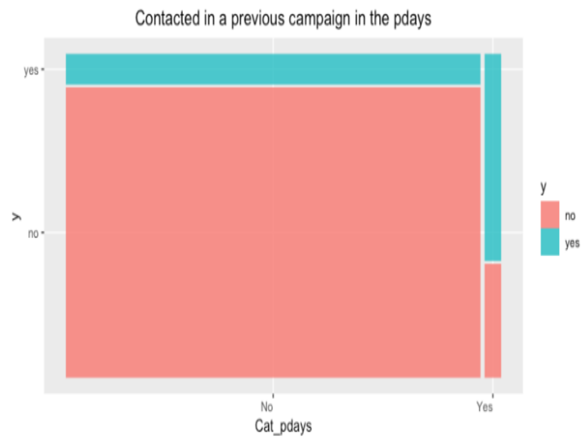

Fig 3.11 Proportions of each levels in Cat_pdays

Since p-value is less than 0.05. It can be concluded that variable cat_pdays is significant with the response variable i.e., y. Recontacting a customer after a prior campaign appears to significantly boost the likelihood of subscribing which can seen in Fig 3.11.

3. Previous : How many contacts were made prior to this campaign and for each customer?

Converted from numerical to categorical with 3 levels. Because in this attribute some levels show way not enough observations (see Fig 3.12 and Fig 3.13).

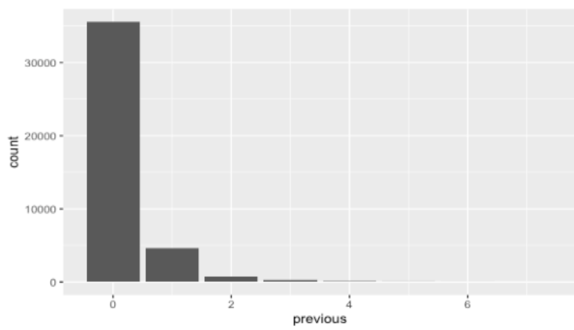

Fig 3.12 Distribution of Previous

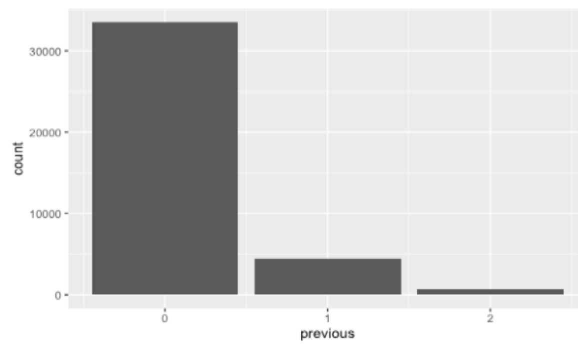

Fig 3.13 Distribution of Previous after reduction

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  38711


                   | Bank_Data$y
Bank_Data$previous |        no |       yes | Row Total |
-------------------|-----------|-----------|-----------|
                 0 |     30464 |      3043 |     33507 |
                   |    17.073 |   135.245 |           |
                   |     0.909 |     0.091 |     0.866 |
                   |     0.886 |     0.701 |           |
                   |     0.787 |     0.079 |           |
-------------------|-----------|-----------|-----------|
                 1 |      3516 |       955 |      4471 |
                   |    51.888 |   411.038 |           |
                   |     0.786 |     0.214 |     0.115 |
                   |     0.102 |     0.220 |           |
                   |     0.091 |     0.025 |           |
-------------------|-----------|-----------|-----------|
                 2 |       392 |       341 |       733 |
                   |   102.941 |   815.463 |           |
                   |     0.535 |     0.465 |     0.019 |
                   |     0.011 |     0.079 |           |
                   |     0.010 |     0.009 |           |
-------------------|-----------|-----------|-----------|
      Column Total |     34372 |      4339 |     38711 |
                   |     0.888 |     0.112 |           |
-------------------|-----------|-----------|-----------|


        Pearson's Chi-squared test

data:  Bank_Data$previous and Bank_Data$y
X-squared = 1533.6, df = 2, p-value < 2.2e-16
```

Since p-value is less than 0.05. It can be concluded that variable previous is significant with the response variable i.e., y.

4. Poutcome : Outcome of previously contacted customer?

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  38711


                   | Bank_Data$y
Bank_Data$poutcome |        no |       yes | Row Total |
-------------------|-----------|-----------|-----------|
           failure |      3496 |       552 |      4048 |
                   |     2.687 |    21.284 |           |
                   |     0.864 |     0.136 |     0.105 |
                   |     0.102 |     0.127 |           |
                   |     0.090 |     0.014 |           |
-------------------|-----------|-----------|-----------|
       nonexistent |     30464 |      3043 |     33507 |
                   |    17.073 |   135.245 |           |
                   |     0.909 |     0.091 |     0.866 |
                   |     0.886 |     0.701 |           |
                   |     0.787 |     0.079 |           |
-------------------|-----------|-----------|-----------|
           success |       412 |       744 |      1156 |
                   |   367.801 |  2913.588 |           |
                   |     0.356 |     0.644 |     0.030 |
                   |     0.012 |     0.171 |           |
                   |     0.011 |     0.019 |           |
-------------------|-----------|-----------|-----------|
      Column Total |     34372 |      4339 |     38711 |
                   |     0.888 |     0.112 |           |
-------------------|-----------|-----------|-----------|


        Pearson's Chi-squared test

data:  Bank_Data$poutcome and Bank_Data$y
X-squared = 3457.7, df = 2, p-value < 2.2e-16
```

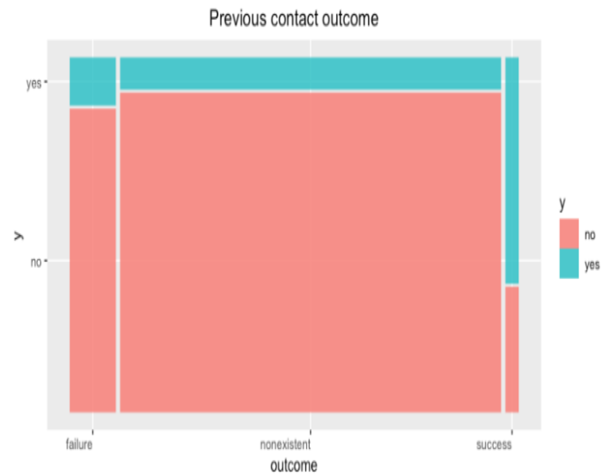

Fig 3.14 Proportion of each levels in Poutcome

Since p-value is less than 0.05. It can be concluded that variable Poutcome is significant with the response variable i.e., y. Almost 64.4% of customers who previously subscribed to a term deposit have agreed to do so again. Therefore, it is important to recontact customers (see Fig 3.14).

Social - Economical context attributes

The five continuous variables are indicators of social and economic conditions. They are Variation in employment rate, Consumer price index, Consumer confidence index, Euribor 3 months rate, Number of employees. The correlation between these variables is calculated and plotted as shown in Fig 3.15.

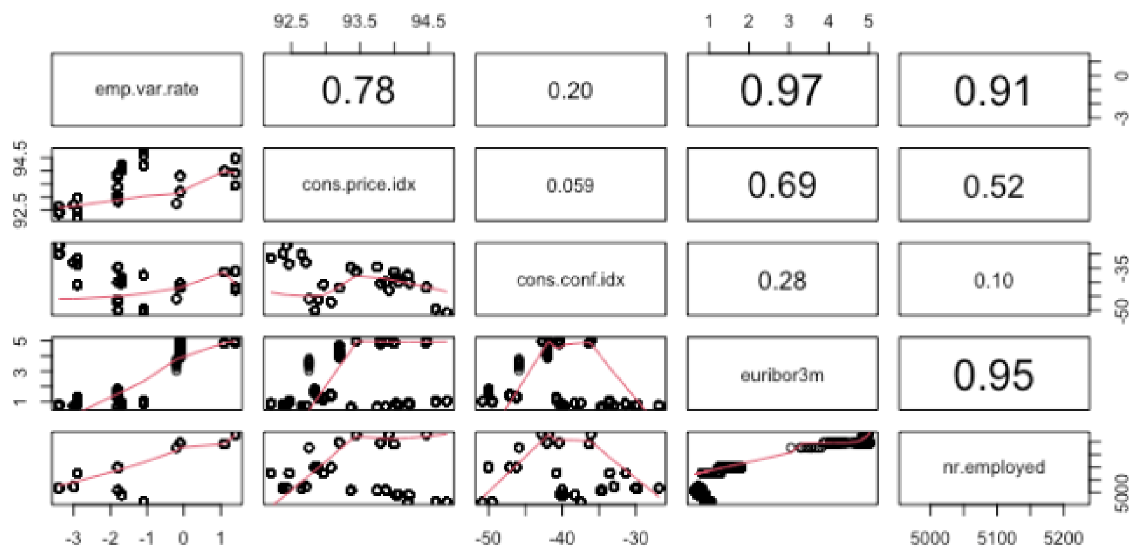

Fig 3.15 Correlation Plot of Social-Economical context attributes

More than 0.90 correlation coefficients were found in three pairs, which is far too high. Emp.var.rate is not significant. In order to soften the correlations between those five variables (see Fig 3.15), this variable is discarded. While two variables, euribor 3m and nr.employed, still show a strong correlation of 95%, these variables are retained. Due to the fact that the number of employees is not related to the euribor 3 months rate, this is most likely a misleading association.

```
        Welch Two Sample t-test

data:  Bank_Data$cons.price.idx by Bank_Data$y
t = 24.914, df = 5129.9, p-value < 2.2e-16
alternative hypothesis: true difference in means between group no and group yes is not equal to 0
95 percent confidence interval:
 0.2432471 0.2847972
sample estimates:
 mean in group no mean in group yes
         93.59021          93.32619


        Welch Two Sample t-test

data:  Bank_Data$cons.conf.idx by Bank_Data$y
t = -8.1332, df = 4910.9, p-value = 5.25e-16
alternative hypothesis: true difference in means between group no and group yes is not equal to 0
95 percent confidence interval:
 -0.9782358 -0.5982372
sample estimates:
 mean in group no mean in group yes
        -40.61108         -39.82284


        Welch Two Sample t-test

data:  Bank_Data$euribor3m by Bank_Data$y
t = 59.063, df = 5361.4, p-value < 2.2e-16
alternative hypothesis: true difference in means between group no and group yes is not equal to 0
95 percent confidence interval:
 1.591484 1.700760
sample estimates:
 mean in group no mean in group yes
         3.787124          2.141002


        Welch Two Sample t-test

data:  Bank_Data$nr.employed by Bank_Data$y
t = 57.46, df = 4965.1, p-value < 2.2e-16
alternative hypothesis: true difference in means between group no and group yes is not equal to 0
95 percent confidence interval:
 75.03565 80.33669
sample estimates:
 mean in group no mean in group yes
         5175.396          5097.710
```

From the Welch two sample t-test it can be seen that all other variables in social-economical context are significant with the response variable.

From the Fig 3.16,

- There is a similar difference in the average consumer price index between subscribers and non-subscribers: 93.4055 for subscribers and 93.5345 for non-subscribers.

- It is not apparent that the consumer confidence index differs significantly between subscribers and non-subscribers: -39.55 for non-subscribers and -41.15 for subscribers.
- Euribor 3 month subscribers have a lower median and are more variable than non-subscribers.
- There is a significant difference between the number of bank employees by customer group. Among non-subscribed customers, the median number of employees 5196 is higher than the median number of subscribers 5099.

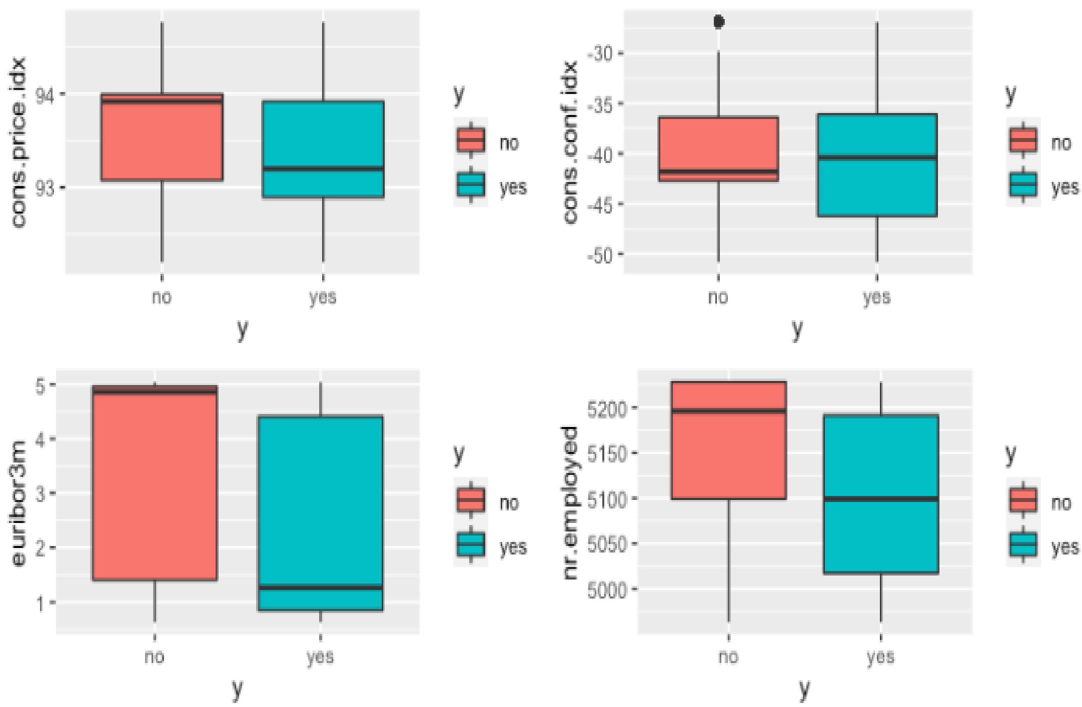

Fig 3.16 Boxplot of Social-Economical context Attributes

Summary of Data Pre-processing

After Preprocessing the data set contains 38711 rows and 15 Predictors and 1 response (see Fig 3.17).

Accepted	Rejected
Age	Default
Job	Housing
Marital	Loan
Education	Duration
Contact	emp.var.rate
Month	
Day_of_week	
Campaign	
Pdays	
Previous	
Poutcome	
cons.price.idx	
cons.conf.idx	
euribor3m	
nr.employed	

Fig 3.17 Proposed dataset

Rejected variables : Five variables are rejected. A lack of variability in default, a lack of significance in housing, a lack of significance in loans, a meaninglessness in duration, and a lack of significance in variation in employment rate.

Accepted variables : In order to interpret character variables, they must be transformed into factor variables (see Fig 3.18). Finally, there are 15 predictors and 1 response variable without missing values.

```
'data.frame':   38711 obs. of  16 variables:
 $ age           : Factor w/ 3 levels "Middle_aged",..: 1 1 1 1 1 1 1 1 3 3 ...
 $ job           : Factor w/ 11 levels "admin.","blue-collar",..: 4 8 8 1 8 8 1 2 10 8 ...
 $ marital       : Factor w/ 3 levels "divorced","married",..: 2 2 2 2 2 2 2 2 3 3 ...
 $ education     : Factor w/ 6 levels "basic.4y","basic.6y",..: 1 4 4 2 4 3 5 6 5 4 ...
 $ contact       : Factor w/ 2 levels "cellular","telephone": 2 2 2 2 2 2 2 2 2 2 ...
 $ month         : Factor w/ 12 levels "jan","feb","mar",..: 5 5 5 5 5 5 5 5 5 5 ...
 $ day_of_week   : Factor w/ 7 levels "mon","tue","wed",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ campaign      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ cat_pdays     : Factor w/ 2 levels "No","Yes": 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/ 3 levels "failure","nonexistent",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ cons.price.idx: num  94 94 94 94 94 ...
 $ cons.conf.idx : num  -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 ...
 $ euribor3m     : num  4.86 4.86 4.86 4.86 4.86 ...
 $ nr.employed   : num  5191 5191 5191 5191 5191 ...
 $ y             : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

Fig 3.18 Levels and Data types of each variable in Proposed Dataset

Section 4

Model Implementation

Categorizing data into classes is the process of classification. It can be performed on both structured and unstructured data. In the process, the first step involves predicting the class of provided data points. A class can also be called a target, a label, or a category. As part of classification predictive modeling, input variables are transformed into discrete output variables. It is important to identify the category or class to which the new data belong. The binary classification process is based on categorical variables being predicted using only two categories.

Splitting a data set into training and testing sets is known as data splicing. The data is split 80:20, so that 80% of the data is used for training and 20% for testing the model; this was done using the random samples and permutation function sample() in R. Customer subscriptions to term deposits are output (y). The implementation is based on several machine learning algorithms. In order to get the greatest accuracy and the maximum contributions, Logistic Regression, Naive Bayes, and Random Forest algorithms are applied. A popular classification method is logistic regression. When the target variable is binary, this method is used. The output y in our model is either yes or no. Both continuous and categorical predictors can be used. The Naive Bayes algorithm is a probabilistic approach to solving classification problems based on the Bayes Theorem. An independent predictor variable is the basis for a Machine Learning model. Algorithms such as the Random Forest algorithm are used to perform supervised classification and regression. In this algorithm, several trees are randomly planted in a forest.

- Linear algorithms: Logistic Regression.
- Nonlinear algorithms: Naive Bayes.
- Bagging algorithms: Random Forest.

Logistic Regression

Logistic regression is fundamental to machine learning and is used most often to classify data. The basic approach to Logistic Regression is quite similar to Linear Regression. Using logistic regression, an independent variable or predictor X is used to predict a categorical dependent variable Y. (K.Domijan)

Logistic Regression for Binary response :

Random component:

$$Y \sim Bernoulli(\pi)$$

$$E[Y] = \pi$$

Systematic component:

The linear combination of explanatory variables used in the model.

$$\eta = \beta 0 + \beta 1X1+\beta 2X2+.....+\beta nXn$$

Link function: logit link

$$g(\pi) = \log(\pi/1-\pi)=\eta$$

Note that since we are interested in the parameter:

$$\log(\pi/1-\pi)=\eta$$
$$\pi/1-\pi= e^{\eta}$$
$$\pi=e^{\eta}/1+e^{\eta}$$

This is called the logistic function. This function might look like this:

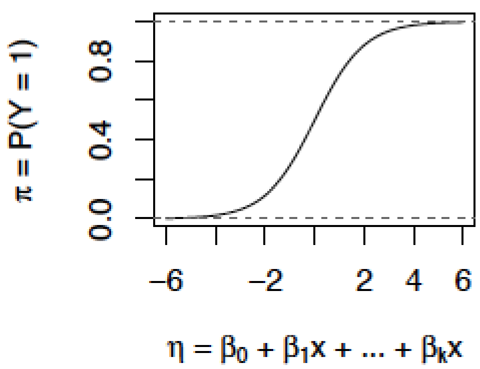

Fig 4.1 Logistic Regression

Where, $0<e^{\eta}/1+e^{\eta}<1$

This gives us the logistic regression function:

$$\log(\pi/1-\pi)= \beta 0 + \beta 1X1+\beta 2X2+.....+\beta nXn$$

The glm() method in R is used to create a logistic regression model. Logistic regression is a technique of model known as a Generalized Linear Model (GLM), which can be built using the glm() function.

A glm() function has the following syntax:

glm (formula, data, family)

Where,

Formula: The relationship between independent and dependent variables can be represented by the formula.

Data: Formula applied to a collection of data.
Family: The type of regression model is specified in this field. In order to analyze bank data, binary logistic regression is applied.
With the glm() function, the maximum likelihood method is used to compute the model. Using this method, the coefficients (β0, β1) are determined so that the predicted probabilities are as close as possible to the true probabilities. Essentially, the maximum likelihood estimator will find values (β0, β1) that result in probabilities closest to 0 or 1 for a binary classification.

Logistic Regression Model 1 : Fitted for the entire cleaned dataset.

```
Call:
glm(formula = y ~ ., family = binomial(link = "logit"), data = Bank_Data_Train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2267  -0.4013  -0.3305  -0.2630   2.7998

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  94.966805  25.064273   3.789 0.000151 ***
ageOld_aged                   0.259477   0.122581   2.117 0.034278 *
ageYoung_aged                 0.108043   0.061209   1.765 0.077540 .
jobblue-collar               -0.145967   0.076660  -1.904 0.056900 .
jobentrepreneur              -0.068497   0.122815  -0.558 0.577031
jobhousemaid                 -0.064745   0.146494  -0.442 0.658515
jobmanagement                -0.030585   0.085726  -0.357 0.721260
jobretired                    0.162942   0.115412   1.412 0.158001
jobself-employed             -0.091821   0.117706  -0.780 0.435340
jobservices                  -0.116780   0.084183  -1.387 0.165375
jobstudent                    0.174516   0.113381   1.539 0.123756
jobtechnician                 0.011671   0.070109   0.166 0.867790
jobunemployed                -0.048816   0.126746  -0.385 0.700129
maritalmarried                0.019554   0.068132   0.287 0.774105
maritalsingle                 0.060650   0.076075   0.797 0.425317
educationbasic.6y             0.090414   0.118228   0.765 0.444425
educationbasic.9y            -0.077195   0.093320  -0.827 0.408120
educationhigh.school          0.068466   0.090344   0.758 0.448551
educationprofessional.course  0.056362   0.100087   0.563 0.573347
educationuniversity.degree    0.098985   0.088076   1.124 0.261072
contacttelephone             -0.533973   0.069354  -7.699 1.37e-14 ***
monthapr                     -0.734472   0.132382  -5.548 2.89e-08 ***
monthmay                     -1.435425   0.121282 -11.835  < 2e-16 ***
monthjun                     -0.433547   0.146536  -2.959 0.003090 **
monthjul                     -0.617212   0.135079  -4.569 4.89e-06 ***
monthaug                     -0.985622   0.135398  -7.279 3.35e-13 ***
monthsep                     -1.407262   0.166207  -8.467  < 2e-16 ***
monthoct                     -1.017031   0.155838  -6.526 6.75e-11 ***
monthnov                     -1.206819   0.131679  -9.165  < 2e-16 ***
monthdec                     -0.596799   0.220285  -2.709 0.006744 **
day_of_weektue                0.306913   0.064851   4.733 2.22e-06 ***
day_of_weekwed                0.359219   0.065077   5.520 3.39e-08 ***
day_of_weekthu                0.300492   0.063661   4.720 2.36e-06 ***
day_of_weekfri                0.275504   0.065874   4.182 2.89e-05 ***
campaign                     -0.027312   0.015419  -1.771 0.076514 .
cat_pdaysYes                  1.079148   0.281069   3.839 0.000123 ***
previous                     -0.064955   0.121868  -0.533 0.594035
poutcomenonexistent           0.425950   0.150339   2.833 0.004608 **
poutcomesuccess               0.789054   0.280200   2.816 0.004862 **
cons.price.idx               -0.233017   0.138357  -1.684 0.092149 .
cons.conf.idx                 0.008615   0.008036   1.072 0.283672
euribor3m                     0.204881   0.134649   1.522 0.128110
nr.employed                  -0.014611   0.002580  -5.663 1.48e-08 ***
---
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 21895  on 30967  degrees of freedom
Residual deviance: 17471  on 30925  degrees of freedom
AIC: 17557

Number of Fisher Scoring iterations: 6
```

Regression models can be simplified by eliminating insignificant terms. It is easier to work with a model if the number of terms is reduced. A model with insignificant terms can reduce the precision of the predictors if they are left in the model. So, model 2 is developed where the model 1 is reduced to significant predictors.

Logistic Regression Model 2 : Reduced model.

```
Call:
glm(formula = y ~ age + contact + month + day_of_week + cat_pdays +
    poutcome + nr.employed, family = binomial(link = "logit"),
    data = Bank_Data_Train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1780  -0.3964  -0.3381  -0.2461   2.7821

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         55.9295534  1.7010369  32.880  < 2e-16 ***
ageOld_aged          0.3807619  0.0922969   4.125 3.70e-05 ***
ageYoung_aged        0.1650518  0.0533939   3.091  0.00199 **
contacttelephone    -0.4963491  0.0574531  -8.639  < 2e-16 ***
monthapr            -0.8756932  0.1198878  -7.304 2.79e-13 ***
monthmay            -1.5413479  0.1144890 -13.463  < 2e-16 ***
monthjun            -0.5452852  0.1240738  -4.395 1.11e-05 ***
monthjul            -0.6686275  0.1233858  -5.419 5.99e-08 ***
monthaug            -0.8158432  0.1221457  -6.679 2.40e-11 ***
monthsep            -1.1998422  0.1530669  -7.839 4.55e-15 ***
monthoct            -0.8026928  0.1434966  -5.594 2.22e-08 ***
monthnov            -1.0964711  0.1269559  -8.637  < 2e-16 ***
monthdec            -0.4015104  0.2142048  -1.874  0.06087 .
day_of_weektue       0.3288295  0.0645135   5.097 3.45e-07 ***
day_of_weekwed       0.3670810  0.0648740   5.658 1.53e-08 ***
day_of_weekthu       0.3047630  0.0632923   4.815 1.47e-06 ***
day_of_weekfri       0.2795046  0.0656667   4.256 2.08e-05 ***
cat_pdaysYes         1.0697597  0.2617830   4.086 4.38e-05 ***
poutcomenonexistent  0.5115671  0.0635522   8.050 8.31e-16 ***
poutcomesuccess      0.8210450  0.2658216   3.089  0.00201 **
nr.employed         -0.0112218  0.0003372 -33.282  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 21895  on 30967  degrees of freedom
Residual deviance: 17526  on 30947  degrees of freedom
AIC: 17568

Number of Fisher Scoring iterations: 6
```

For comparing two models Model 1 and Model 2 ANOVA test is performed. In R ANOVA test is performed using anova(model 1, model 2) function.

```
Analysis of Deviance Table

Model 1: y ~ age + job + marital + education + contact + month + day_of_week +
    campaign + cat_pdays + previous + poutcome + cons.price.idx +
    cons.conf.idx + euribor3m + nr.employed
Model 2: y ~ age + contact + month + day_of_week + cat_pdays + poutcome +
    nr.employed
  Resid. Df Resid. Dev  Df Deviance  Pr(>Chi)
1     30925      17471
2     30947      17526 -22   -55.44 0.0001028 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

AIC is less for model 1. And also in anova test we reject the null hypo and conclude Model 1 is better.

Performance evaluation of the Logistic Regression Model 1 is based on confusion matrix, accuracy and AUC from the ROC curve (see Fig 4.2).

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6809  648
         1  106  180

               Accuracy : 0.902
                 95% CI : (0.8958, 0.9091)
    No Information Rate : 0.8931
    P-Value [Acc > NIR] : 0.003123

                  Kappa : 0.2838

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

            Sensitivity : 0.21739
            Specificity : 0.98467
         Pos Pred Value : 0.62937
         Neg Pred Value : 0.91310
              Precision : 0.62937
                 Recall : 0.21739
                     F1 : 0.32316
             Prevalence : 0.10694
         Detection Rate : 0.02325
   Detection Prevalence : 0.03694
      Balanced Accuracy : 0.60103

       'Positive' Class : 1
```

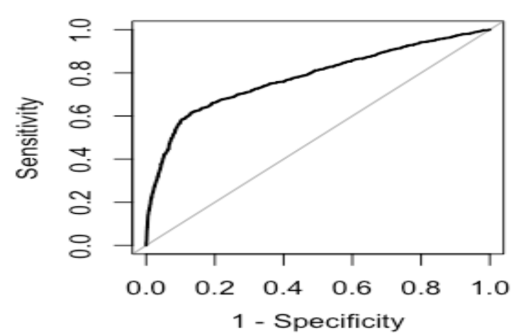

Fig 4.2 ROC - Logistic Regression

Area under the curve is 0.7787

In R, confusionMatrix() function is used to calculate the confusion matrix, predicted accuracy, confidence interval, sensitivity, specificity, and other metrics. Here, the predicted accuracy of the logistic regression model is 90.2% and the misclassification rate is 0.09737828. The area under the curve is 0.7787 which is shown in Fig 4.2.

Naive Bayes

Naive Bayes algorithm is based on the Bayes Theorem and uses a probabilistic approach to solve classification problems (Keshari).

Bayes Theorem

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Where,

P(A|B) – A probability of event A occurring if event B has already occurred Posterior Probability

P(B|A) – A probability of event B occurring if event A has already occurred Likelihood

P(A) – the probability of event A - Prior Probability of the proposition

P(B) – the probability of event B - Prior Probability of evidence

In the Naive Bayes algorithm, each predictor variable is considered independent of any other variable. 'Naive' is the name given to it.

Fitting a Naive Bayes model in which predictors are believed to be independent within each class label is done using naive_bayes() in R. Performance evaluation of the Naive Bayes Model is based on confusion matrix, accuracy and AUC from the ROC curve (see Fig 4.3).

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6386  431
         1  529  397

               Accuracy : 0.876
                 95% CI : (0.8685, 0.8833)
    No Information Rate : 0.8931
    P-Value [Acc > NIR] : 0.999999

                  Kappa : 0.383

 Mcnemar's Test P-Value : 0.001744

            Sensitivity : 0.47947
            Specificity : 0.92350
         Pos Pred Value : 0.42873
         Neg Pred Value : 0.93678
              Precision : 0.42873
                 Recall : 0.47947
                     F1 : 0.45268
             Prevalence : 0.10694
         Detection Rate : 0.05127
   Detection Prevalence : 0.11959
      Balanced Accuracy : 0.70148

       'Positive' Class : 1
```

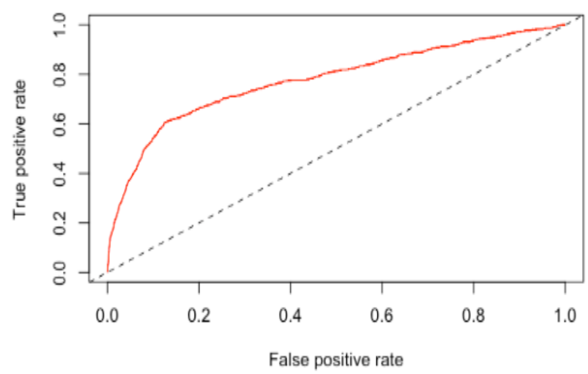

Fig 4.3 ROC - Naive Bayes

Area under the curve is 0.7781189

Using the function confusionMatrix() to calculate the confusion matrix, predicted accuracy, confidence interval, sensitivity, specificity, and other metrics. Here, the predicted accuracy of the naive bayes model is 87.6% and the misclassification rate is 0.123983. The area under the curve is 0.7781189 which is shown in Fig 4.3.

Random Forest

Random Forest is a Bagging algorithm. Bagging is a way of improving the performance of a tree, by calculating multiple trees from different samples and averaging the results. (Hurley)The steps are:

Repeat B times

1. Sample n observations with replacement from the data.
2. Fit a tree $\hat{f}_k$ the kth sample
3. At each step, in deciding on the optimal split, use only a random selection of the m available predictors.

Typically m =√p, where p is number of predictors

The reasoning behind this concept is that, if there is a very strong predictor in the data, it will be the first pick for all bagged trees. Other predictors may be neglected. As a result, the majority of the bagged trees will appear the same, and their projections will be the same. Averaging identical predictions will not help to reduce variance. By limiting the predictor choice at a node to just a subset, the trees are not correlated i.e. the predictions are less correlated. Averaging these predictions reduces variance even more.

The randomForest() function in R implements Breiman and Cutler's original Fortran code for Breiman's random forest algorithm for classification and regression. The Random Forest Model is evaluated based on confusion matrix, accuracy, and AUC (see Fig 4.4).

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6757  598
         1  158  230

               Accuracy : 0.9024
                 95% CI : (0.8955, 0.9089)
    No Information Rate : 0.8931
    P-Value [Acc > NIR] : 0.003918

                  Kappa : 0.3328

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

            Sensitivity : 0.27778
            Specificity : 0.97715
         Pos Pred Value : 0.59278
         Neg Pred Value : 0.91869
              Precision : 0.59278
                 Recall : 0.27778
                     F1 : 0.37829
             Prevalence : 0.10694
         Detection Rate : 0.02970
   Detection Prevalence : 0.05011
      Balanced Accuracy : 0.62746

       'Positive' Class : 1
```

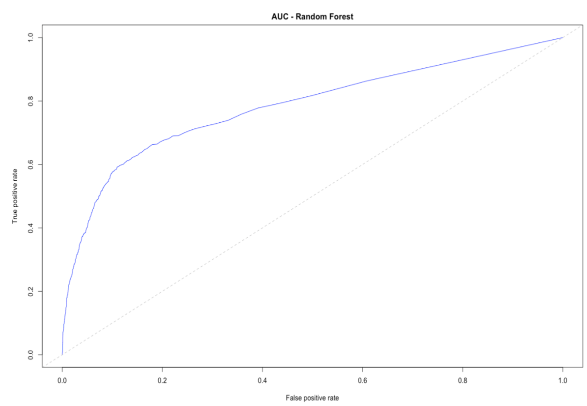

Fig 4.4 ROC - Random Forest

Area under the curve is 0.7829

To calculate the confusion matrix, predicted accuracy, confidence interval, sensitivity, specificity, and other metrics confusionMatix() function is used in R. Here, the predicted accuracy of the random forest model is 90.24% and the misclassification rate is 0.09763657. The area under the curve is 0.7829 which is shown in Fig 4.4.

Section 5

Model Evaluation

When the evaluation process is completed, the most appropriate model is chosen to predict which Customer will apply for a Term Deposit in the Bank. A model's performance is measured in terms of its predictive accuracy and Area under the ROC curve, which is the basis for evaluating its performance. An R model is used to test the accuracy of each model and to visualize the results. The evaluation and decision-making process is conducted using a more accurate model. AUC - ROC curves can be used to measure performance for a wide range of threshold levels for classification issues. The AUC measures the degree of separability, whereas the ROC measures probability. This reflects how good a model is at discriminating between classes. As the AUC increases, the model is more likely to predict 0 classes as 0 and 1 classes as 1(Sarang Narkhede). In the same way, higher the AUC, better the model distinguishes between customers who sign for term deposits and those who do not.

Models	Accuracy	AUC
Logistic Regression	0.9020	0.7787
Naive Bayes	0.8760	0.7781
Random Forest	0.9024	0.7829

Table 5.1 Model Selection

From the results obtained from different model implementations, the Random Forest algorithm gave the highest Accuracy of 90.24 percent, which was followed by Logistic Regression with 90.20 percent of accuracy and so on. Whereas, AUC of random forest has the highest value of 0.7829, followed by logistic regression with 0.7787.

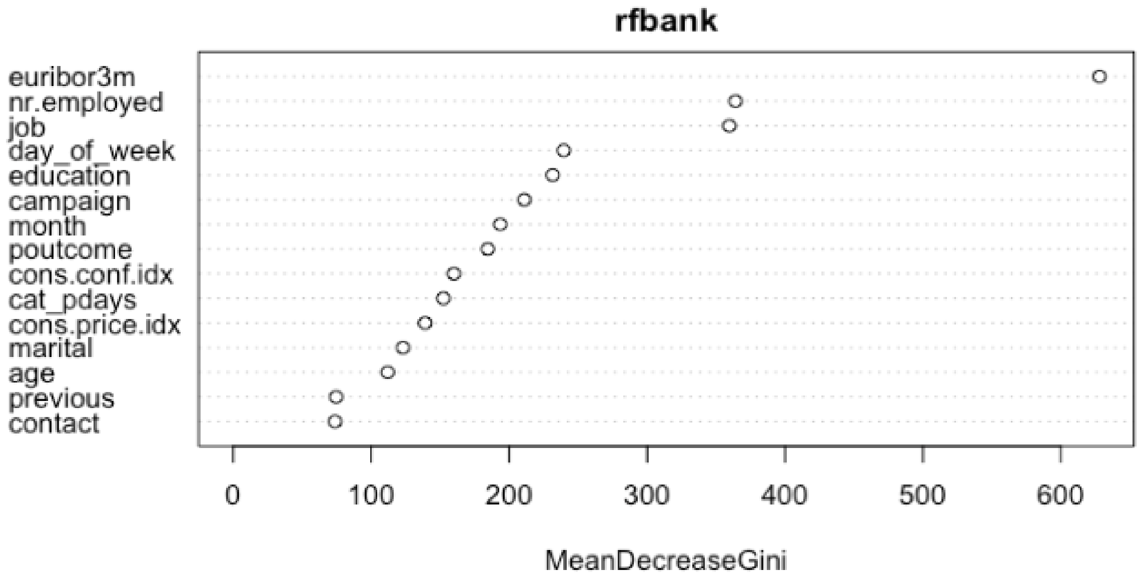

Fig 5.1 Variable of Importance from Random Forest Algorithm

In the R varImpPlot() function is used to plot the variable of importance plot from the random forest algorithm, it has been determined that the Euribor 3m rate contributed the highest percentage to the bank dataset, followed by the number of employees and jobs.

Section 6

Conclusion

The main study goal was to examine customer behavior when initiating a term deposit in a bank using certain basic criteria. Three distinct machine learning approaches were used to evaluate the study (Logistic Regression, Naive Bayes, Random Forest). Random Forest had the highest performance, with the maximum predicted accuracy of 90.24 percent. Also, euribor 3m rate, nr.employed, and job are the three most important variables in the forecast. Using this algorithm, banks will be able to generate more revenue while saving expenses by contacting customers who are more likely to sign up for a term deposit. In the future, the model may be enhanced by comparing it to a larger dataset. Also, different models like boosting algorithms and ensemble models must be used to get more accuracy.

Reference

Abhijit Raorane and R.V.Kulkarni. "Data Mining Techniques: A source for consumer behavior analysis." *Cambridge Dictionary | English Dictionary, Translations & Thesaurus*, 2016, https://arxiv.org/pdf/1109.1202.pdf.

Ajay, Ajay Venkatesh, Shomona Gracia Jacob. "Prediction of Credit-Card Defaulters: A Comparative Study on Performance of Classifiers." *International Journal of Computer Applications*, 2016, https://www.ijcaonline.org/archives/volume145/number7/ajay-2016-ijca-910702.pdf.

Anirut Suebsing and Chakarin Vajiramedhin. "Accuracy Rate of Predictive Models in Credit Screening." *Hikari Ltd*, 2013, http://www.m-hikari.com/ams/ams-2013/ams-109-112-2013/suebsingAMS109-112-2013.pdf.

Catherine Hurley. *Statistical machine learning.* Notes, 2022.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Edited by Springer Science+Business Media, Springer, 2007. *Pattern Recognition and Machine Learning*, https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf.

Gianluca Bontempi. "Statistical foundations of Machine learning the handbook." *ResearchGate*, 2022,https://www.researchgate.net/publication/242692234_Statistical_foundations_of_machine_learning_the_handbook.

Isabelle Guyon and Andre Elisseeff. "An Introduction to Variable and Feature Selection 1 Introduction." *Journal of Machine Learning Research*, 2003, https://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf.

Katarina Domijan. *Generalized linear model*. Notes, 2022.

Kislay Keshari. “Naive Bayes Tutorial | Naive Bayes Classifier in Python.” *Edureka*, 28 July 2020, https://www.edureka.co/blog/naive-bayes-tutorial/.

Nestor Pereira. “Using Machine Learning Classification Methods to Detect the Presence of Heart Disease.” *Arrow@TU Dublin*, 11 December 2019, https://arrow.tudublin.ie/cgi/viewcontent.cgi?article=1226&context=scschcomdis.

Olatunji Apampa. “Evaluation of Classification and Ensemble Algorithms for Bank Customer Marketing Response Prediction.” *CSUSB ScholarWorks*, 2016, https://scholarworks.lib.csusb.edu/cgi/viewcontent.cgi?article=1296&context=jitim.

Radhakrishnan B, Shineraj G, Anver Muhammed K.M. “Application of Data Mining In Marketing.” *Cambridge Dictionary | English Dictionary, Translations & Thesaurus*, 2013, https://arxiv.org/pdf/1310.8462.pdf.

Sarang Narkhede, “Understanding AUC-ROC curve”, 2018, https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5#:~:text=the%20multiclass%20model%3F-,What%20is%20the%20AUC%20%2D%20ROC%20Curve%3F,capable%20of%20distinguishing%20between%20classes.

Sergio Moro, Paulo Cortez, Paulo Rita. “A data-driven approach to predict the success of bank telemarketing.” *Cambridge Dictionary | English Dictionary, Translations & Thesaurus*, 2014,https://reader.elsevier.com/reader/sd/pii/S016792361400061X?token=75BFC1D43AA6191939C211D4B9FB4696EF20901F035D11BBD66352A3F1F19926218C5363229437BA1D98EC6728655E63&originRegion=eu-west-1&originCreation=20220728174559.

Sergio Moro, Paulo Cortez, Raul Laureano. “A data mining approach for bank telemarketing using the rminer package and r tool.” *ResearchGate*, 2013, https://www.researchgate.net/publication/256464440_A_data_mining_approach_for_bank_telemarketing_using_the_rminer_package_and_r_tool.

Tuba Parlar and Songul Kakilli Acaravci. “Using Data Mining Techniques for Detecting the Important Features of the Bank Direct Marketing Data | International Journal of Economics and Financial Issues.” *EconJournals.com*, 14 April 2017, https://www.econjournals.com/index.php/ijefi/article/view/4580.

UCI Machine Learning Repository. “Bank Marketing Data Set.” *UCI Machine Learning Repository*, 14 February 2012, https://archive.ics.uci.edu/ml/datasets/bank+marketing.