**Interim Project Report on**

**“Prediction of Term Deposit Subscription based on Bank Marketing”**

**Submitted by**

**Group No. 4 [Batch: Sep 2019,** **Location: Chennai]**

**Group Members**

1. **Balaaje V S - WFPBXFVJ7A**
2. **Govindha Raju - ZWTCM7P0RN**
3. **Hari Krishna B - G55O0WR054**
4. **Rajasekaran R - B6CJWAGHCM**
5. **Santhosh G S - F1MTKDTJGM**
6. **Vignesh B - F4CU6UC8TX**

**Research Supervisor**

**Subramanian P V**

**Great Lakes Institute of Management**



Contents

[1. Introduction 3](#_Toc26427168)

[2. Problem Statement, Scope and Objective 3](#_Toc26427169)

[3. Data Source and Description 3](#_Toc26427170)

[4. Data Preprocessing 3](#_Toc26427171)

[5. Exploratory Data Analysis 3](#_Toc26427172)

[6. Modelling Approach 3](#_Toc26427173)

[7. Actionable insights and recommendations to the stakeholder 4](#_Toc26427174)

[8. References and Bibliography 4](#_Toc26427175)

[9. Appendix 4](#_Toc26427176)

# Introduction

<<Relevant details about the project, industry, company and a need of the present study.>>

**Abstract:**

This is the classic marketing bank dataset uploaded originally in the UCI Machine Learning Repository. The dataset gives you information about a marketing campaign of a financial institution in which you will have to analyze in order to find ways to look for future strategies in order to improve future marketing campaigns for the bank.

The classification goal is to predict if the client will subscribe a term deposit (variable y).

Find the best strategies to improve for the next marketing campaign. How can the financial institution have a greater effectiveness for future marketing campaigns? What the best algorithms to predict a term deposit and improve the next campaign efficiency? In order to answer this, we must analyze the last marketing campaign the bank performed and identify the patterns that will help us find conclusions in order to develop future strategies.

**Business interpretation:**

A bank usually invests the customer’s long-term deposits into riskier financial assets which can earn the better return than what they pay to their customer. The customer, on the other hand, is assured a risk-free return on his/her deposit. There is a stiff competition among the financial institutions/banks in increasing the customer base in their retail banking segment. Along with offering innovative products to the public, a huge amount of money is spent on marketing their products. The term deposit is very important among the diverse range of products and services offered by banks in retail banking segment.

With advancement in data science and machine learning and availability of data, most banks are adapting to a data-driven decision. This dataset here consists of direct marketing by contacting the clients and assessing the success rate of sales made.

In this project, we apply machine learning algorithms to build a predictive model of the data set to provide a necessary suggestion for marketing campaign team. The goal is to predict whether a client will subscribe a term deposit (variable y) with the help of a given set of dependent variables. This is a real dataset collected from a Portuguese bank that used its own contact-center to do direct marketing campaigns to motivate and attract the clients for their term deposit scheme to enhance the business

# Problem Statement, Scope and Objective

<<Defining the core problem statement and listing down the objectives of the project. Ideally here you will state the overall objective and break it down into the steps that you are going to follow to achieve the objective.>

A Portuguese bank is rolling out term deposit for its customers. They have in the past connected to their customer base through phone calls. Results for these previous campaigns were recorded and have been provided to the current campaign manager to use the same in making this campaign more effective.

Challenges that the manager faces are following:

* Customers have recently started to complain that bank’s marketing staff bothers them with irrelevant product calls and this should immediately stop
* There is no prior framework for her decide and choose which customer to call and which one to leave alone

The manager has decided to use past data to automate this decision, instead of manually choosing through each customer. Previous campaign data which has been made available; contains customer characteristics, campaign characteristics, previous campaign information as well as whether customer ended up subscribing to the product as a result of that campaign or not.

Using this the manager plans to develop a statistical model which given this information predicts whether customer in question will subscribe to the product or not. A successful model which can do this, will make the marketing campaign efficiently targeted and less bothering to uninterested customers.

## **Objective:**

To Build a machine learning predictive model and predict which customers should be targeted for rolling out term deposits by bank. Our study will adopt data mining techniques to predict customers’ term deposit subscription behaviors and understand customers’ features to improve the effectiveness and accuracy of bank marketing. In order to achieve this objective, we break the whole approach into following questions.

I. How to predict whether a bank client will subscribe to a term deposit or not?

II. Which determinants would indicate a client is ready to subscribe to a term deposit through direct marketing?

III. How to segment term deposit market?

IV. Are there any common features of clients who have subscribed to a term deposit?

Evaluation Criterion: AUC score and Accuracy

**Literature Survey:**

A few studies have been conducted both in India and abroad over a period of time regarding the marketing strategies applicable in the banking sector. Followings few of the studies are reviewed hereunder as they would facilitate a clear backing for carrying out the present study.

Mehta (2010) [2] in his article” Personal Selling-A Strategy for promoting Bank Marketing “reported that there is lack of Marketing Communication in Indian Banks. He suggested for adopting banks suitable marketing promotion strategies for better business. He emphasized that on adoption of personal selling as a strategy for marketing promotion in Banks the banking business can improve considerably.

Moro, Cortez and Laureano [3] used the rminer Package to test three classification models (Decision Trees, Naïve Bayes and Support Vector Machines) and compare their performance through Receiver Operating Characteristic curve (ROC) and Lift curve analysis. Moro, Cortez and Rita [4] also tested four data mining models, including logistic regression, decision trees (DT), neural network (NN) and support vector machine. After evaluating area of the receiver operating characteristic curve (AUC) and area of the LIFT cumulative curve (ALIFT), neural network presented the best performance. Nachev combined cross-validation and multiple runs to partition the data set into train and test sets [5]. He also explored the impact of performance caused by different neural network designs.

Wang, Song and Fang [6] mentioned that the banking industry lacks scientific marketing management and they came up with the idea that carrying out market segmentation of deposit marketing and selecting the marketing target is the scientific way of marketing management.

# Data Source and Description

<<Identifying and providing information about the data to be used for the analysis. Students should mention the source of data, contents and time period to be used. Provide detail description of data including data dictionary, amount of missing data and all other relevant information in detail. For long list of variables and their descriptions, use appendix >>

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (or not) subscribed.

There are two datasets:

1. bank-full.csv with all examples, ordered by date (from May 2008 to November 2010).
2. bank.csv with 10% of the examples (4521), randomly selected from bank-full.csv.
3. The smallest dataset is provided to test more computationally demanding machine learning algorithms (e.g. SVM).
4. The classification goal is to predict if the client will subscribe a term deposit (variable y).
5. Number of Instances: 41188 for bank-full.csv
6. Number of Attributes: 21
7. Attribute information:

# bank client data:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.No | Column Name | Description | Type | Values |
| 1 | age | age of the person | Numeric | 40,42…… |
| 2 | job | type of job | Categorical | "admin.","unknown","unemployed","management","housemaid","entrepreneur","student","blue-collar","self-employed","retired","technician","services" |
| 3 | marital | marital status | Categorical | "married","divorced","single";  **note**: "divorced" means divorced or widowed |
| 4 | default | has credit in default? | Categorical | "no","yes","unknown" |
| 5 | housing | has housing loan? | Categorical | "no","yes","unknown" |
| 6 | loan | has personal loan? | Categorical | "no","yes","unknown" |
| 7 | education | Type of education | Categorical | "basic.4y","basic.6y","basic.9y","high.school","illiterate","professional.course","university.degree","unknown" |

# related with the last contact of the current campaign:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.No | Column Name | Description | Type | Values |
| 8 | contact | contact communication type | Categorical | "cellular","telephone" |
| 9 | month | last contact month of year | Categorical | "jan", "feb", "mar", ..., "nov", "dec" |
| 10 | day\_of\_week | last contact day of the week | Categorical | "mon","tue","wed","thu","fri" |
| 11 | duration | last contact duration, in seconds | Numeric | 149,261…….. |

# other attributes:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.No | Column Name | Description | Type | Values |
| 12 | campaign | number of contacts performed during this campaign and for this client | Numeric | 1,2…. |
| 13 | pdays | number of days that passed by after the client was last contacted from a previous campaign | Numeric | 1,2….. 999,999- means client was not previously contacted |
| 14 | previous | number of contacts performed before this campaign and for this client | Numeric | 0,1,2.. |
| 15 | poutcome | outcome of the previous marketing campaign | Categorical | "failure","nonexistent","success" |

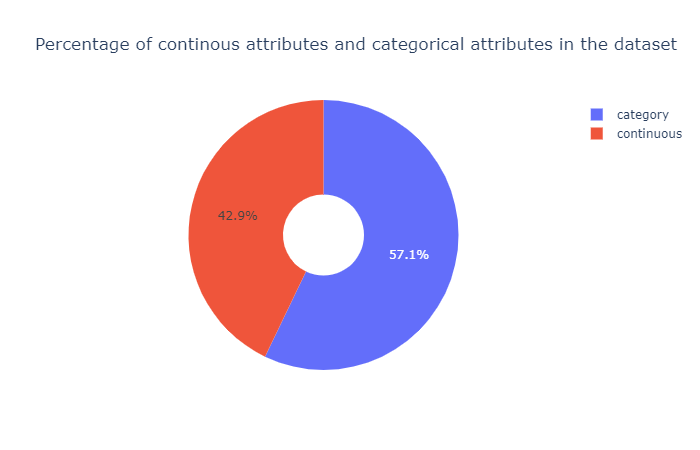
# social and economic context attributes:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.No | Column Name | Description | Type | Values |
| 16 | emp.var.rate | employment variation rate - quarterly indicator | Numeric | 1.1,1.5…. |
| 17 | cons.price.idx | consumer price index - monthly indicator | Numeric | 93.99,93.84…. |
| 18 | cons.conf.idx | consumer confidence index - monthly indicator | Numeric | -36.4,-35.5… |
| 19 | euribor3m | The Euro Interbank Offered Rate is a daily reference rate, published by the European Money Markets Institute- 3 month rate | Numeric | 4.857,5.025…. |
| 20 | nr.employed | number of employees - quarterly indicator | Numeric | 5191,5250… |

#Output variable (desired target):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.No | Column Name | Description | Type | Values |
| 21 | y | has the client subscribed a term deposit? | Categorical | "yes","no" |

8. Missing Attribute Values: None



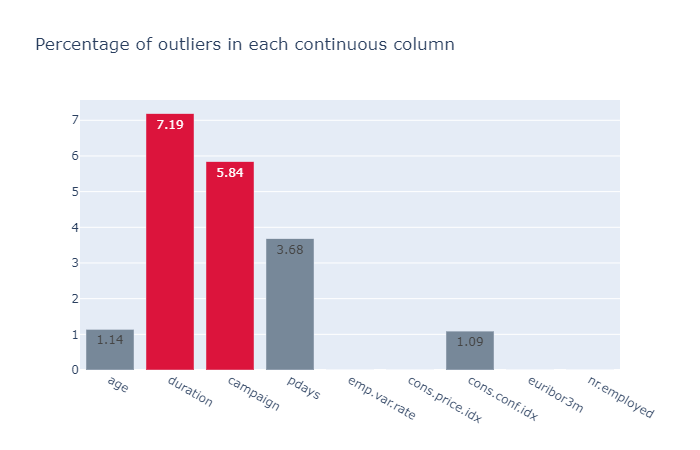
# Data Preprocessing

## **Outliers:**

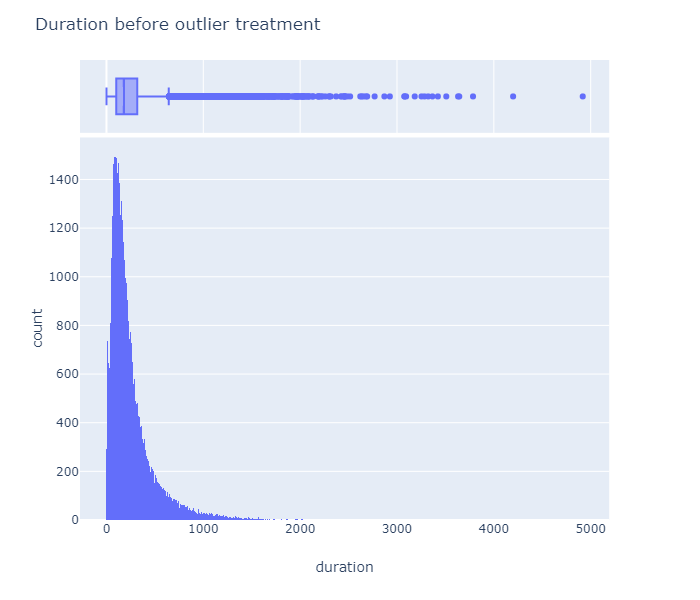
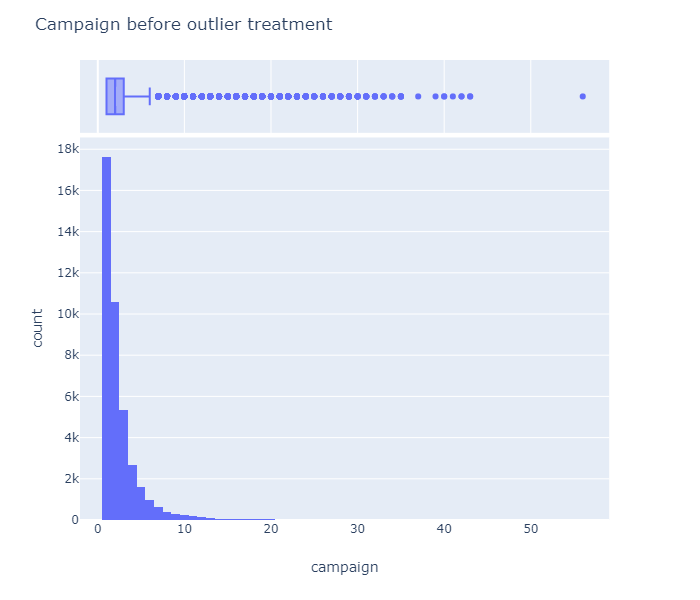
* Outliers were found using IQR method .if value<fence\_low or value> fence\_high then those values were removed.

|  |  |  |
| --- | --- | --- |
| s.no | column name | outlier percentage |
| 1 | age | 1.14 |
| 2 | duration | 7.19 |
| 3 | campaign | 5.84 |
| 4 | pdays | 3.68 |
| 5 | emp.var.rate | 0.00 |
| 6 | cons.price.idx | 0.00 |
| 7 | cons.conf.idx | 1.09 |
| 8 | euribor3m | 0.00 |
| 9 | nr.employed | 0.00 |

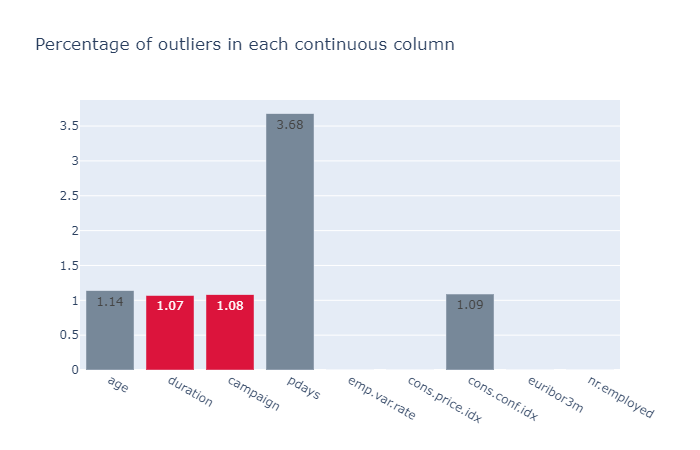
* Outliers with outlier percentage greater than 5 were removed and imputed.
* Various imputation techniques were tested and finally fixed with bfill.
* bfill produced very less outlier percentage and followed close to normality than other methods (mean, median, ffill, mode).



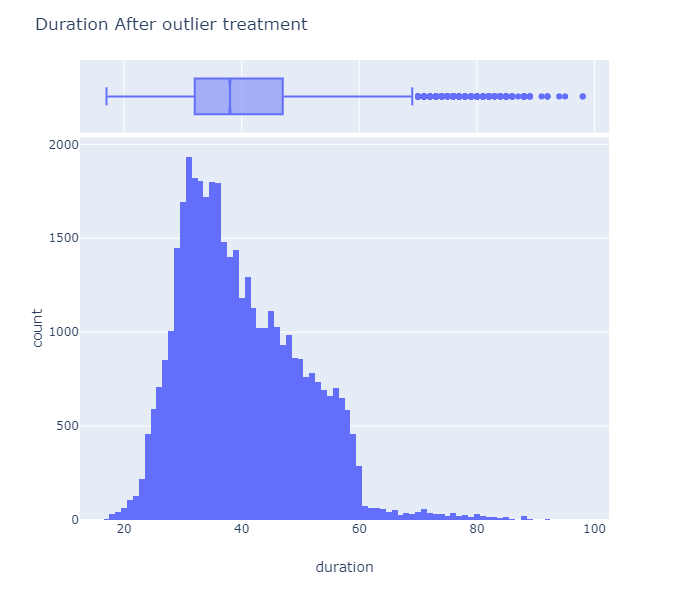
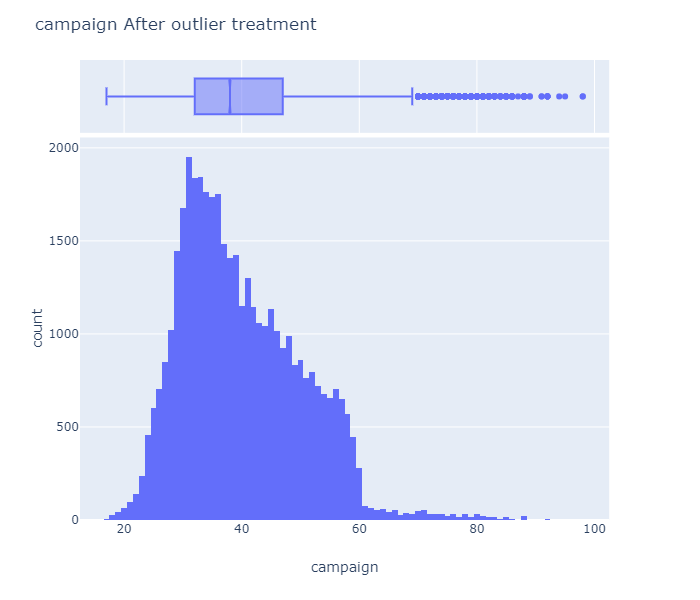
Duration and campaign column before Treatment:



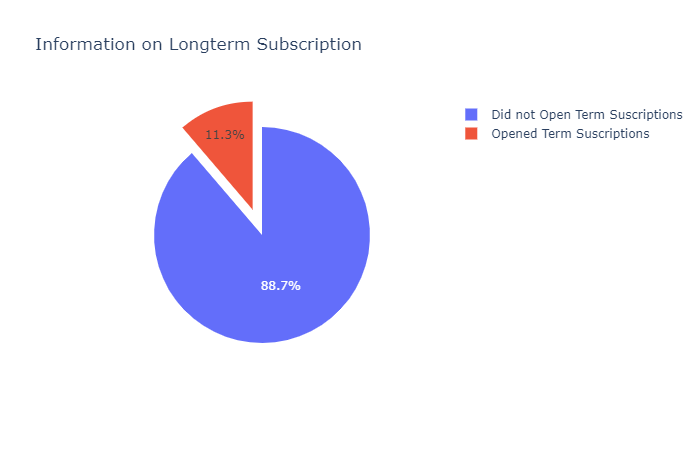
After Outlier Treatment:



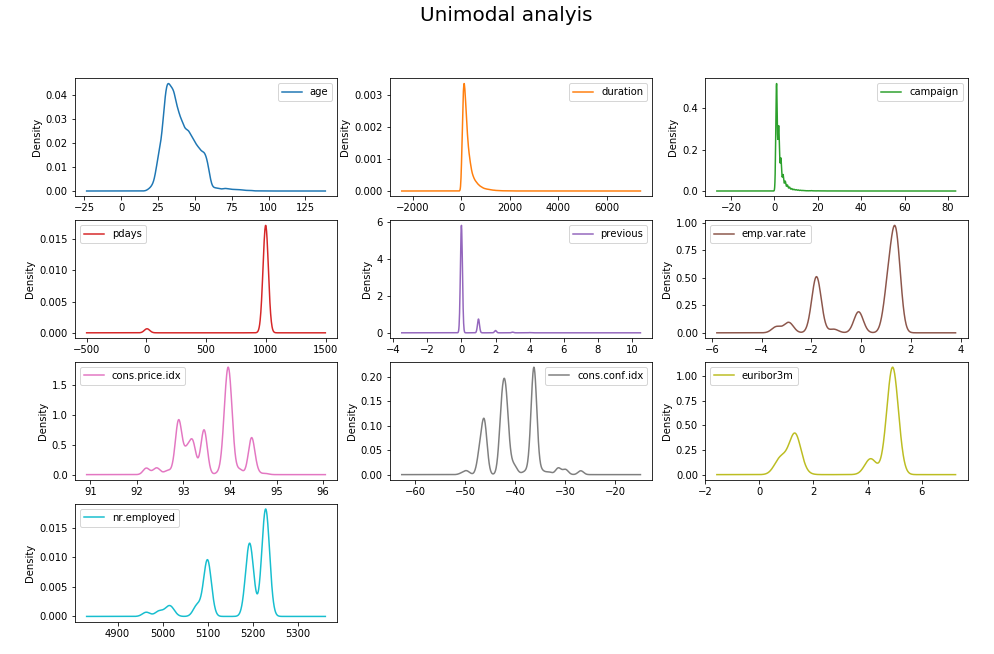
Duration and campaign column After Treatment:



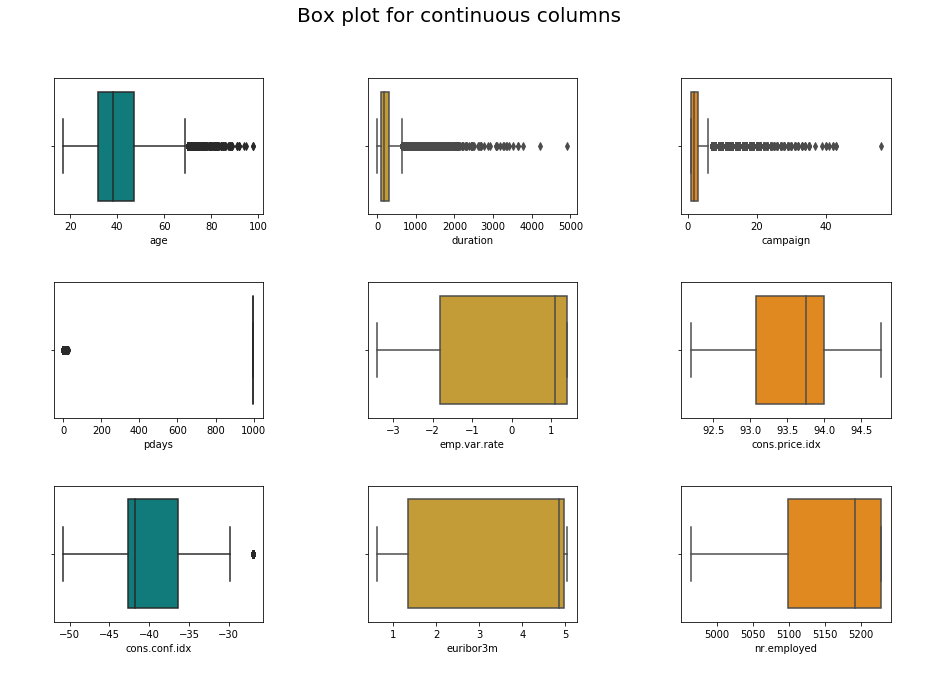
# Exploratory Data Analysis



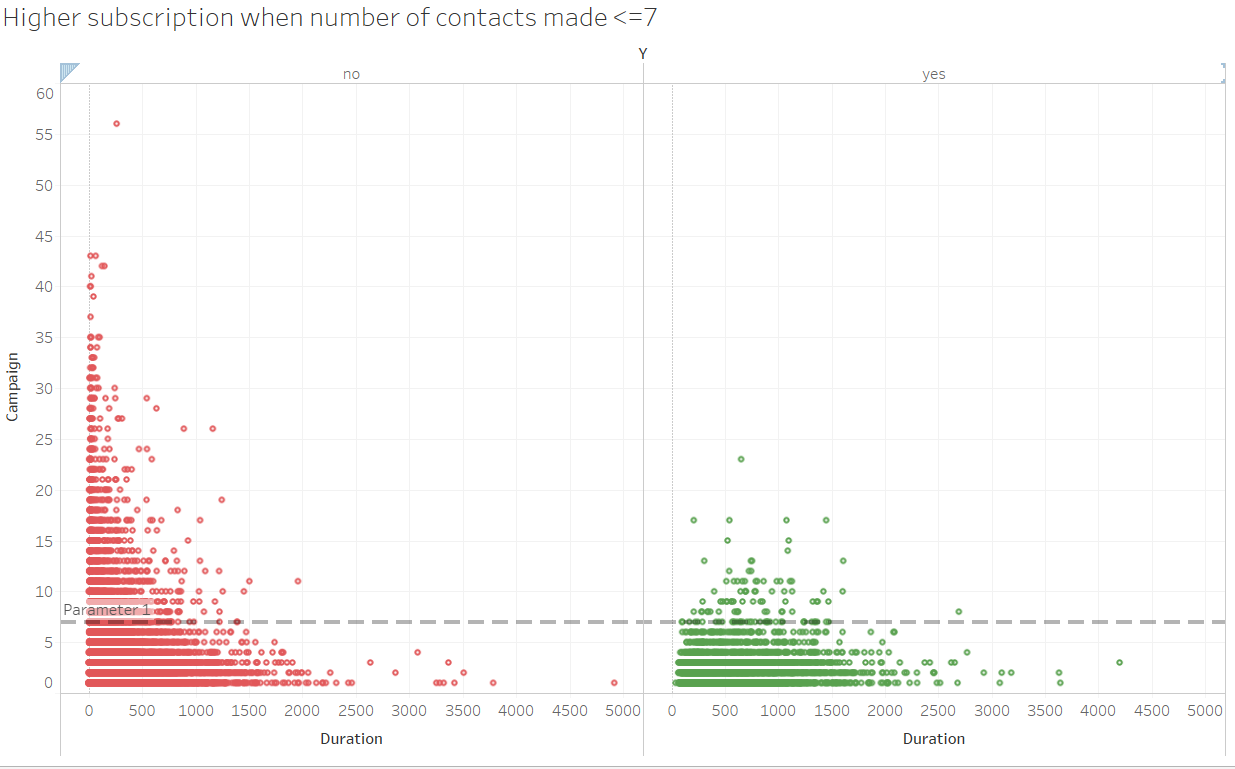
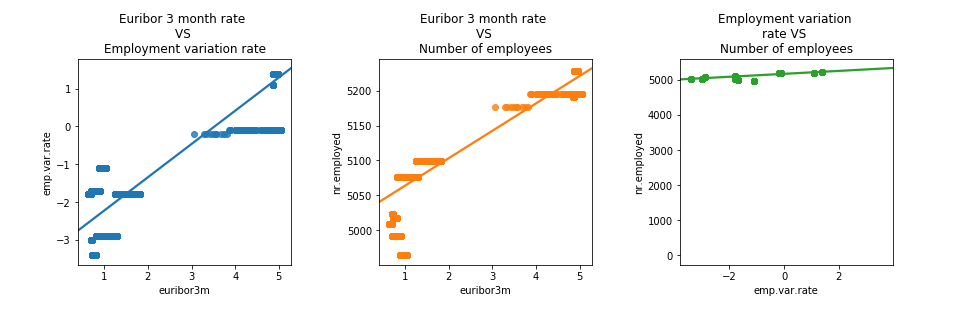
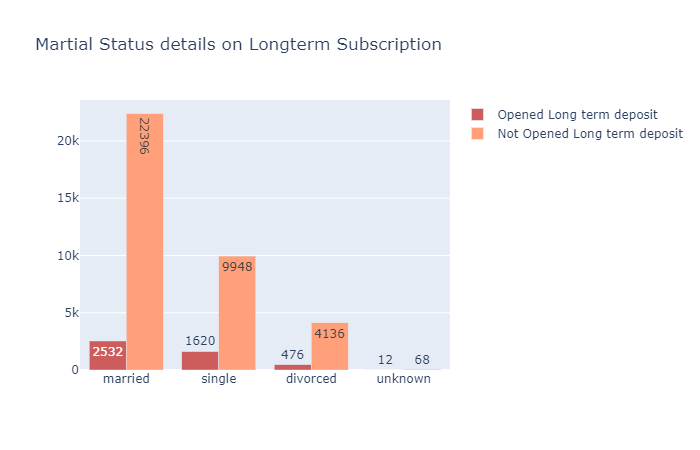
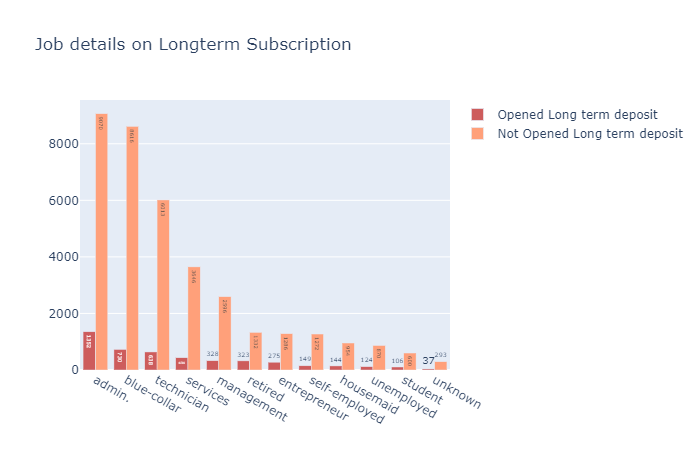
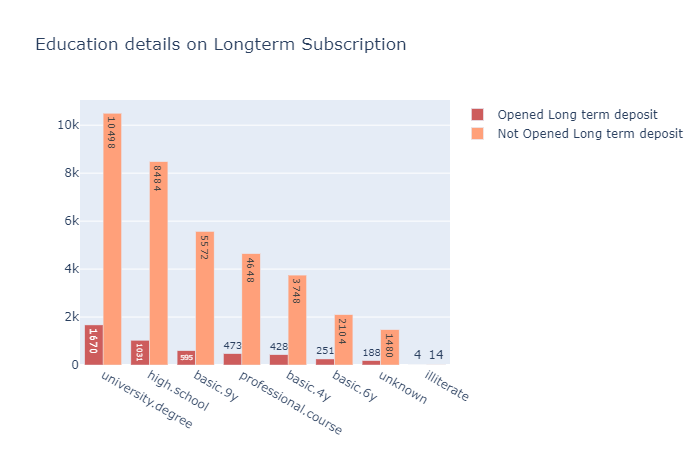
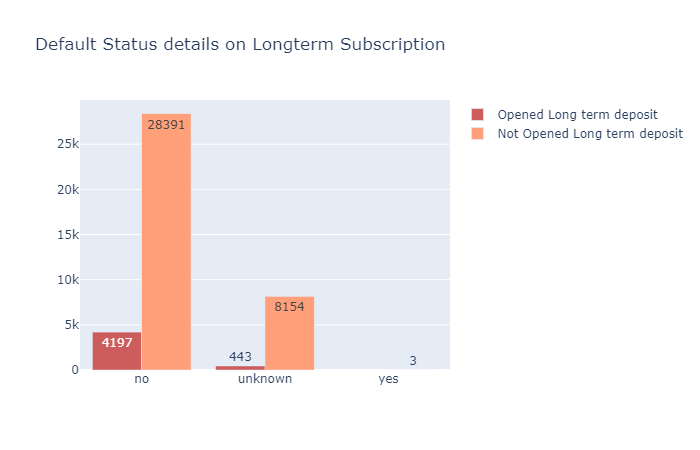
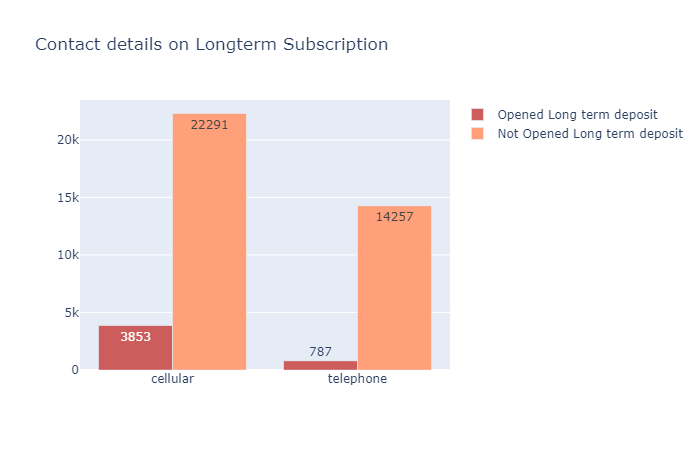
**Inference:** from the plot we can see that only 11.3% percentage of people subscribed to the long term subscription plan and this further implies our data is imbalanced for prediction .hence, we need to balance the data using balancing techniques like SMOTE and over sampling techniques.



**Inference:** all the continuous columns are skewed and not scaled. Hence scaling and outlier removal is necessary for this dataset



**Inference:** Box plot shows there are outliers present in the dataset and it should be treated before proceeding to model building.



# Modeling Approach

<< Interim reports are not expected to include final results but for effective feedback you must provide some interim results. For example, if you are working to predict a binary response variable, it is not enough to state that you will be using logistic regression and CART. You should include logistic regression model output and CART output. Before using any model, test the model assumptions and prove that all model assumptions are satisfied. For all supervised learning models, try K-fold cross validation on data or at least split the data into training and test data; use the training data to build the model and test the model using the test data. It is strongly encouraged that you try to provide some insights into the problem and mention the next steps.>>

# Actionable insights and recommendations to the stakeholder

<<Based on your EDA and the models employed on the data, get actionable insights that the stakeholder can act upon. You may also mention the key findings of the project.>>

# References and Bibliography

<< Details about any reference books, articles, web resources etc. that are to be used for the project.>>

[1] UCI Machine Learning Repository (2014). Bank Marketing Data Set  
[Online]. Available from:  
[http://archive.ics.uci.edu/ml/datasets/Bank+Marketing](https://slack-redir.net/link?url=http%3A%2F%2Farchive.ics.uci.edu%2Fml%2Fdatasets%2FBank%2BMarketing) (Accessed: 4  
September 2017).

[2] Mehta, S. (2010) “Personal Selling –A Strategy for promoting Bank Marketing” State Bank of India Monthly Review.

[3] Moro, S., Cortez, P. & Laureano, R. (2013) A data mining approach for bank telemarketing using the rminer package and r tool [Online].

Available from:

<https://www.researchgate.net/publication/256464440_A_data_mining_approach_for_bank_telemarketing_using_the_rminer_package_and_r_tool>

[4] Moro, S., Cortez, P. & Rita, P. (2014) A Data-Driven Approach to Predict the Success of Bank Telemarketing [Online]. Decision Support Systems, Elsevier, 62:22-31, June 2014

Available from:

https://pdfs.semanticscholar.org/4a27/709545cfa225d8983fb4df8061fb205b9116.pdf (Accessed: 14 September 2017).

[5] Nachev, A. (2015) Application of data mining techniques for directmarketing [Online].

Available from:

http://www.foibg.com/ibs\_isc/ibs-30/ibs-30-p09.pdf (Accessed: 14 September 2017).

[6] Wang, B.Z., Song, J.L., & Fang, C. (2002) ‘Opinions on Deposit Marketing of Commercial Banks’, Financial Theory and Practice, 2002(9), August, pp.32-33.

# Appendix

<<All raw codes and raw outputs must be placed in Appendix.

In the appendix, you may include data dictionary and a view of the data for at least five observations.>>

**For detail instructions see Interim Report Guidelines**