Data Analytics Research Project

Title - Analyzing Portuguese Bank’s Campaign

Rishika Reddy Aleti

Abstract

This research aims to investigate the banking records for the Portuguese bank’s campaign. The main goal of the project is to predict if a customer would sign up for a term deposit or not. As the number of customer establishments increase across the banking sectors in Portuguese, it is important to ensure customer commitment for the banks. Since it is a highly populated country with vast number of populations that is increasing every year, the banks should be informed of potential risks they might be exposed to. This study uses data analytics, R programming, Python, visualizations, statistical analysis, and SQL to uncover insights on how the bank’s facilities have been performing with respect to their customer’s key attributes such as age, job, marital status, balance, loans, education, housing, subscriptions etc. The results show that most of the campaigns conducted have a high-risk factor especially for the people with married marital status and balance co-related to term deposit subscription. Therefore, Bank’s should be informed of these key factors that influence the annual term deposit subscription status and be cautious about maintaining a record of customer’s factors. In this research report, we expand on the important findings from the analysis, highlight the study's results, summarize the important findings by providing with suitable visualizations. The project is expected to demonstrate the understanding of the selected dataset through exploration and analysis of the dataset. A key requirement of the project is to identify the data types and appropriate selection and use of statistical and visualization tools associated with those types.

Using the selected dataset, the demonstration of analyses and interpretations using R, Python, and SQL to produce appropriate statistical summaries and visualizations that answer the research questions and support conclusions about the meaning and values derived from the dataset are derived. The data ingestion and exploration are carried out using R, Python, and SQL to determine the final term deposit subscription and success rate of the phone call campaign.

Introduction

The selected dataset is of a Portuguese banking institution’s direct marketing campaign via phone calls. The main objective is to predict whether the customers will sign up for a term deposit. The data is related with direct marketing campaigns of a Portuguese banking institution. This dataset is obtained from uci.edu which is a machine learning repository made open for public usage. The dataset characteristics are multivariate, and the attribute characteristics are real. The primary associated task is to classify the dataset consisting of 45211 number of instances and 17 attributes. This dataset was collected in the year 2012. My interest in the domain is to analyze these variety of 17 attributes that are listed in the dataset.

As this is a banking domain category related dataset it will have all the NOIR datatypes to perform qualitative and quantitative analysis required for the project.

In the profundity of analyzing the selected dataset led to these research questions. The research questions are as follows: -
1. Are the campaigns sufficient to attract all different types of customers?
2. Does marital status of an individual customer cause any change in the balance of the account?
3. Is the account's balance directly proportional to the marital status and subscription of the term deposit?

The potential benefits of answering research questions are numerous. The importance of studying the dataset is to provide information about the analysis of the customers, balance based on marital status and in-depth insights about the campaigns that were conducted to attract different types of customers. By knowing whether the campaigns were sufficient or were they not sufficient to achieve the goal we can get insights and help the bank in deciding on the campaign patterns. The final decision of how the campaigns should be carried out in future can be made by the analysis of the present campaigns. The marital status of an individual customer may cause changes in the account's main balance. Performing analysis on marital status provides us with information that divorced people are broke, which results in a low balance, and could be one of the reasons those customers do not choose to subscribe to the term deposit. Performing analysis on balance of the main account provides insights into the term deposits subscriptions. It gives us significant information related to the annual number of subscriptions taken in the bank and helps in decision making in the subscription plan by the bank. Any changes that need to be made in the plan will be made and implemented in the coming years to attract more customers.

Literature Review

In the one project report paper with the title Banking Dataset – Marketing Targets, the researchers have used the same dataset to predict if the term deposits were achieved or not. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed by the customer or not. Their provenance collection methodology is online collection and phone conversation tracking. Their project is a little similar but not completely similar as they are just finding out the binary answer for the term deposits of the customer whereas, we are even experimenting on the factors that influenced the subscription of term deposits. This paper helped with guidance on how to find the annual number of subscriptions and predict if the customers will convert to term deposits or not. The collection methods are also different in my project. This report confirms and differs from the current project.

In another project report of Data Mining for Bank Direct Using Data Mining for Bank Direct Marketing paper, the researchers have demonstrated on how the CRISP-DM approach was used to implement a Data Mining project. Real-world data was gathered from a Portuguese bank deposit subscription marketing effort. The commercial goal is to develop a model that can predict whether a contact will succeed, i.e., whether a client will subscribe to a deposit. Such a model can improve campaign efficiency by identifying the key factors that influence success, assisting in the better utilization of available resources (e.g., human labor, phone calls, and time), and the selection of a high-quality and affordable pool of potential buyers. This paper is very similar to my project as they have even described about the key factors that participate in influencing the final count of term deposit subscriptions. This paper has researched on things that will be a future of the current project as they have included attributes like the selection of high-quality and potential buyers that are not being focused on in the current project. This project has given me more knowledge and opened many options in which the work in this project can be done. This report duplicates and gives additional information when compared to the project.

In the third project report paper is of Decision Support Systems - A data-driven approach to predict the success of bank telemarketing Project. In this project the researchers have assessed a real problem with bank telemarketing for the sale of long-term deposit using newly proposed social and economic factors, a data driven method has been developed by them. The model features an engineering that prioritized term deposits, resulting in a model with 22 features that was fine-tuned. Using a realistic rolling-window method, they have compared four data mining models. In the end researchers have selected the half that is better classified, as it can target 79 percent of the customers. This project is almost duplicating the current project as they are experimenting on the factors, and they are making a model depending on these project as in the current project also we are trying to find out the influencing factors that affect the rate of annual term deposits of the bank.

Strategy, Methods, and Tools

This dataset was obtained as a CSV file, which made it easier to load into various tools and platforms of R and Python. The dataset consisted of 45,212 rows and 17 columns ordered by date from May 2008 to November 2010. Having a large dataset with multiple fields allowed this project to retrieve an abundance of information regarding the Portuguese Bank. It contained many attributes such as age, job, education, housing, marital status etc. that help to determine the result of the project.

The first step in the project was cleaning of the dataset that is done primarily in Python. The null values are removed, and the duplicate values were checked if any existed. After, finding out that there are no null values and no duplicate values the project is taken forward to analysis stage. Fortunately, no string values had to be adjusted based on their spellings of the same variable. After checking and cleaning the data the statistical analysis could set in motion. R programming language was used to create multiple visualizations and analyze the data. The Tables and visualizations have been demonstrated by a good design that comprised of captions, titles, variable and axis labels and readability. Co-relations between the attributes were recognized using R. Additionally, a univariate analysis of each type of NOIR data has been shown and an appropriate multivariate analysis is also shown by the help of scatterplot, correlation using the regression models. As the dataset has very many data items, all the items were not analyzed thoroughly, the most promising items related to the research questions were chosen by making sure that it includes all the above analysis types. In the end of the project SQL was used to show a schema for the selected dataset. The defining of a table, loading of the dataset, and execution of a few simple queries against the table to demonstrate the basic understanding of SQL for data summaries was performed.

Results

1. **Visualizations and interpretations made by using Python programming language**.

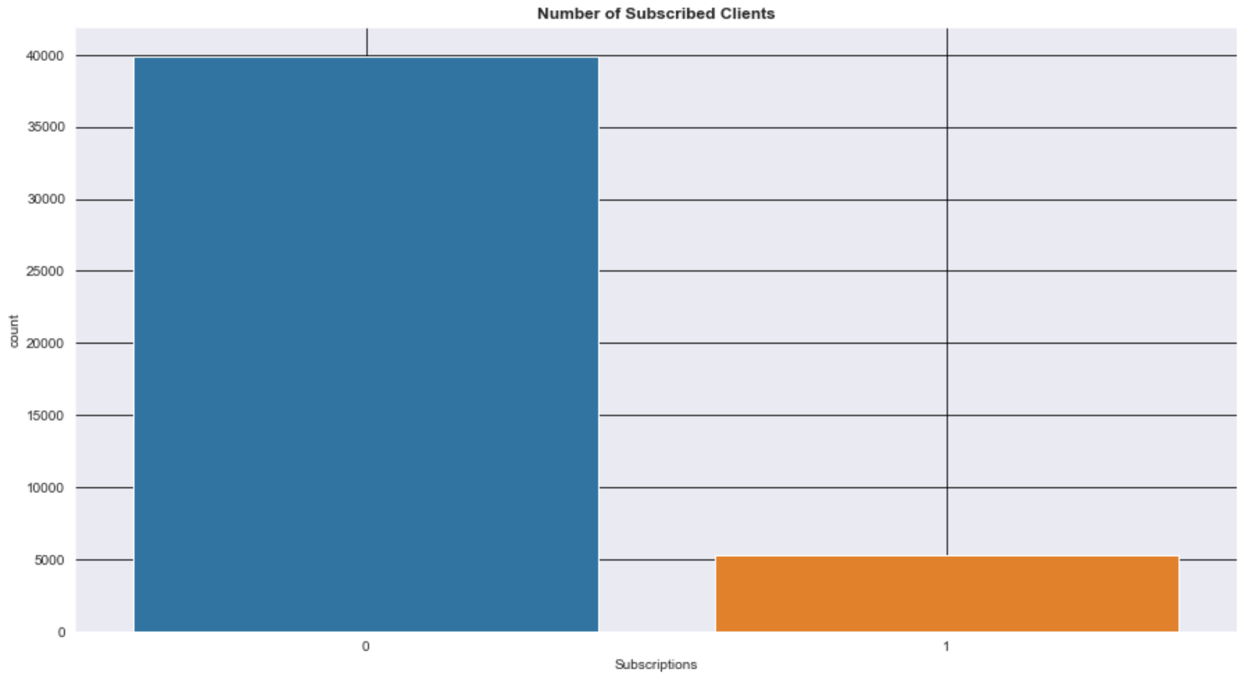


Fig 1 – Number of Subscribed Clients(blue is unsubscribed and orange is subscribed)

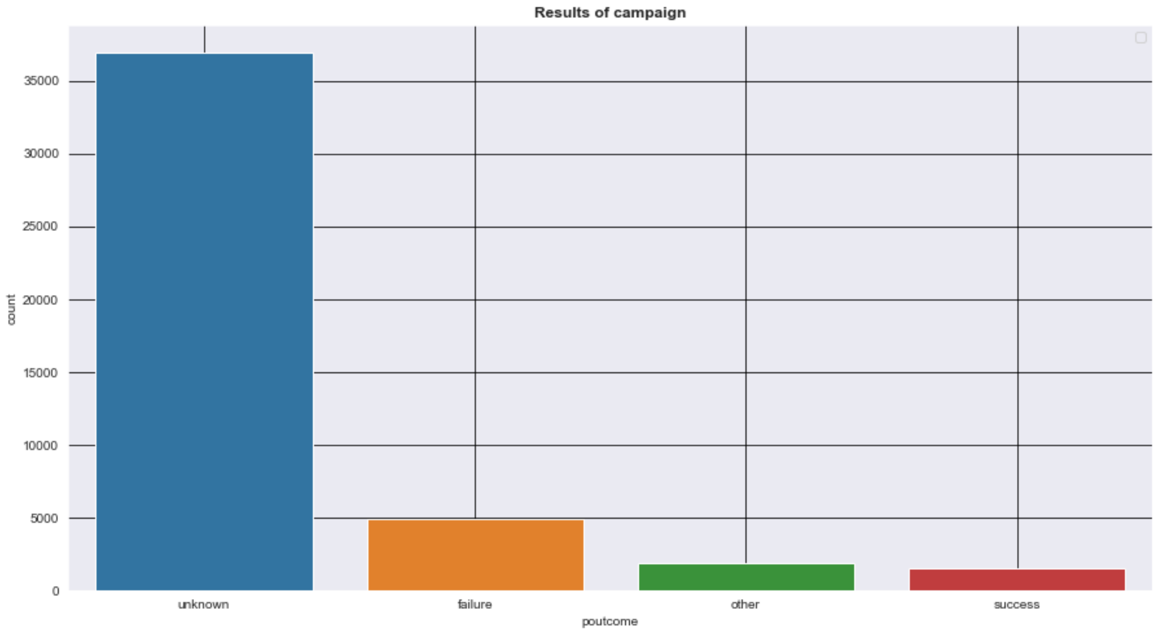


Fig 2 – Results of campaign

This plot gives the viewer a in depth information about the results of the campaign with grid lines making it easy to identify the insights of dataset by providing the answer to first research question about determining whether the campaigns are sufficient or not.

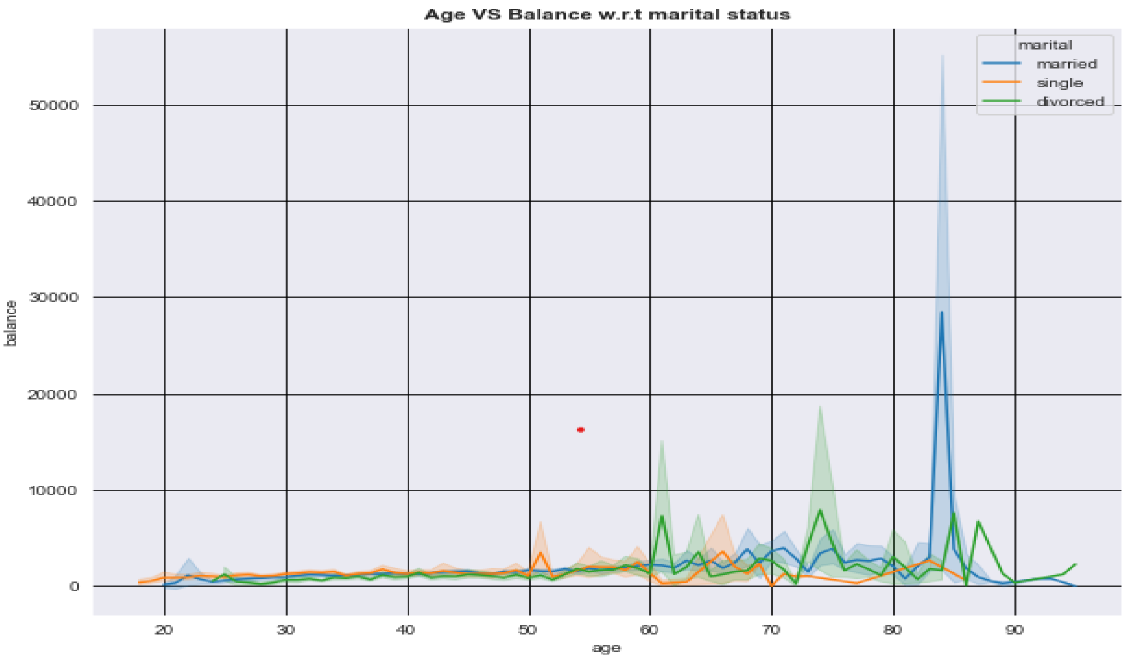


Fig 3 – Affected on balance with respect to marital status.

This plot helps the viewer to identify and answer the second research question about how the balance is effected by the customer's marital status and age. It is observed that married people tend to have more balance when compared to the others.

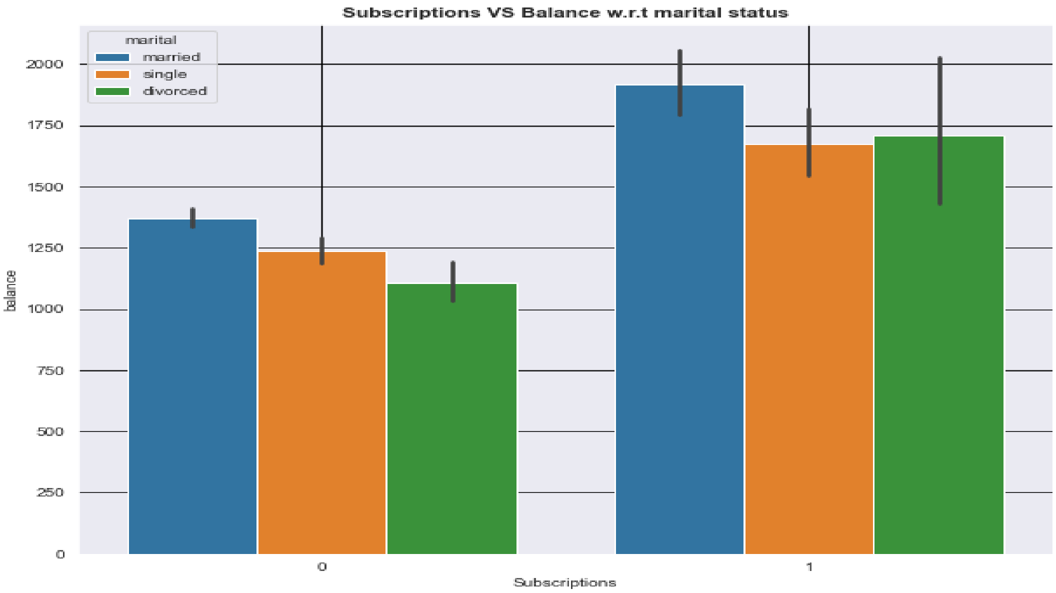


Fig 4 – Analysis of balance and marital status on subscriptions.

This plot helps to answer the third research question of the project i.e., the analysis of final term deposit subscription by observing the influence of marital status and balance as the primary attributes.

2. **Visualizations and interpretation made by using R programming language**.

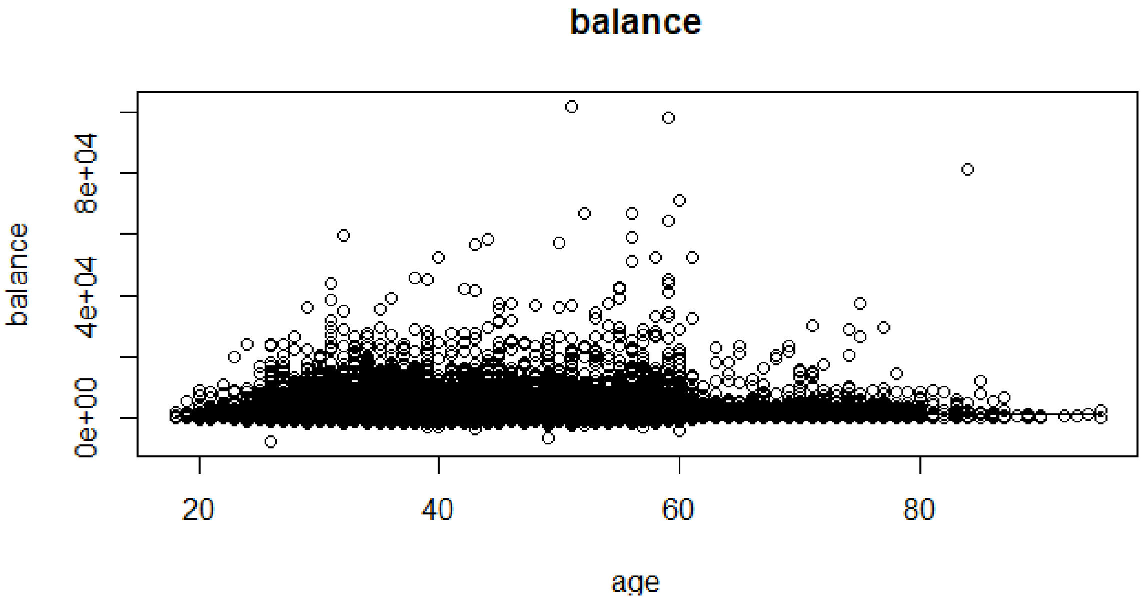


Fig 5- Balance and Age using R.

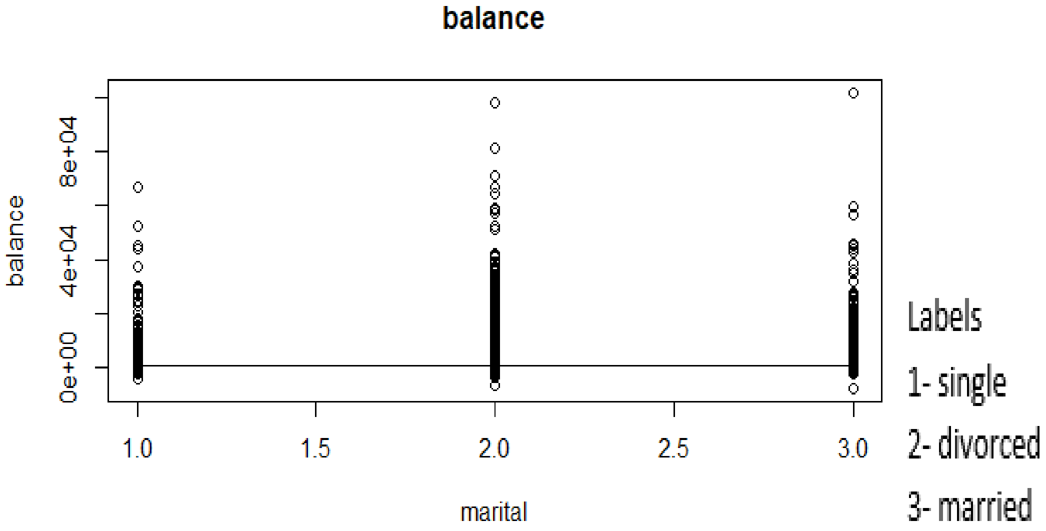
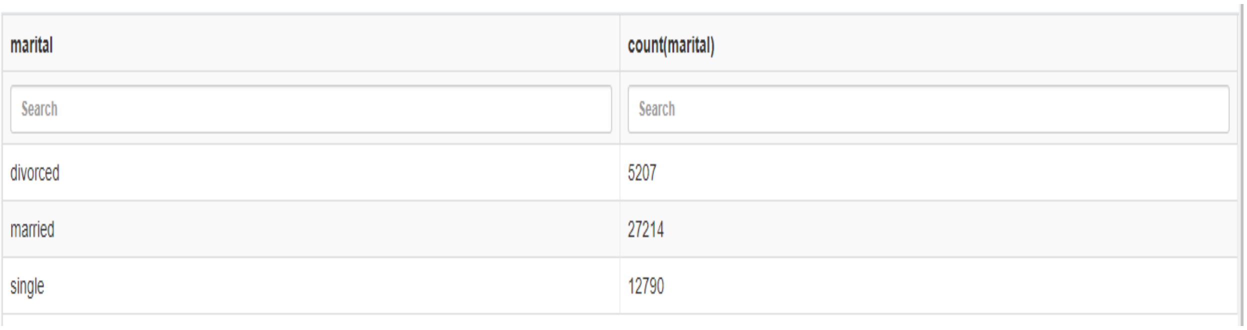


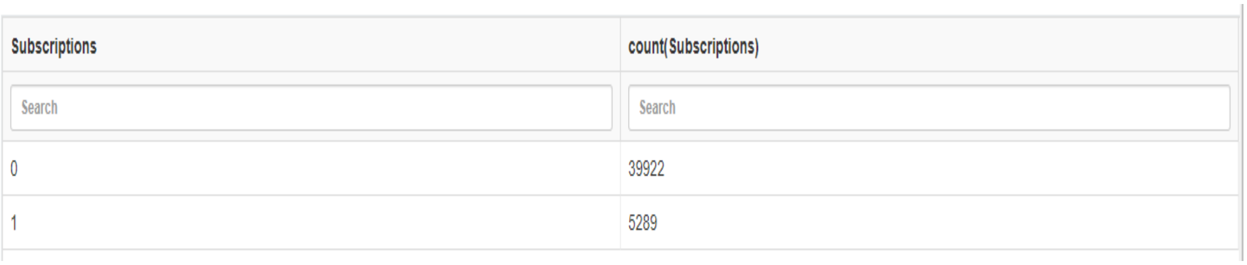
Fig 6- Balance and marital status using R.

3. SQL Server query results.



marital	count(marital)
Search	Search
divorced	5207
married	27214
single	12790

Fig – Running a SQL query to count the marital status of the customers.



Subscriptions	count(Subscriptions)
Search	Search
0	39922
1	5289

Fig – Running a SQL query to get the count of subscriptions.

Limitations

During my study, I encountered a few restrictions. First, there was a dearth of prior research in this field. Few people have investigated banking data, particularly banking data from Europe. As previously stated in this study, I was only able to locate a few research papers that used Portuguese Banking Dataset. Nonetheless, these investigations were conducted in 2014, indicating that there have been no recent studies or advancements in this field. This makes it difficult to develop a theoretical foundation and comprehend the research topics.

Additionally, there were some limitations when analyzing the dataset for sentiment value. Since the dataset was so large, it was hard to load the violations text files into Python. Thus, I had to separate the files by specific attributes for it to load correctly.

Future Research

The IT revolution has increased the Indian banking sector substantially, and technology has improved our perceptions of banking, let alone our ability to participate in it. From "traditional banking" to "convenient banking," from long lines to opening a bank account with just a few clicks, the value created for customers and society has been enormous. Although banks have been rapidly adopting technology, this growth has been shown to be concentrated primarily in urban and metro areas. The actual benefits of information technology have yet to reach the rural populace. This is something that the banking industry must take advantage of in order to fulfill its full potential. One excellent technique to make financial services more accessible to rural customers is to make programs and software available in regional languages. Organizations all across the world are pursuing a variety of digital transformations. While the banking industry has already begun to adopt several of these technologies, particularly digital payments, and secure transactions, it is expected that cloud technology and advanced analytics will soon be adopted to expedite processes. The top three trends and predictions are taking a customer-centric approach and attempting to eliminate friction from the customer's journey, Application of artificial intelligence, big data, and cognitive computing to harness real-time data received by banks and the last one is the Application programming interfaces (APIs) being used to convert traditional banking platforms to open banking platforms (33%)

Conclusion

By the end of the project, we can answer all the research questions that were made in the beginning of the project. The campaigns are clearly not sufficient to attract all different types of customers as we can see from the derived visualizations that the success rate of campaigns is the least and the failure rate of campaigns is much greater than that of success rate.

The marital status of an individual customer causes change in the balance of the account as from the graphs we can identify that the customers with single and divorced seem to have less balance when compared to the one's whose marital status is married. The subscriptions also are likely to be greater for the married people when compared to divorced and single customers.
The account's balance is directly proportional to the marital status and subscription of the term deposit.

According to my study questions, customers aged 40 to 60 are willing to subscribe to a term deposit. The customers' balance is unaffected by their marital status because the association between them is low when compared to other variables. Because most of the customers' addresses are unknown, the bank may or may not have contacted them about the subscription, thus I concluded that the bank had neglected to contact the clients. Housing and loans clearly do not have a significant impact on the target variable.

Reference

S. Moro, P. C. (2014, June). Bank Marketing Data Set. Retrieved from https://archive.ics.uci.edu/ml/datasets/bank+marketing#

RATHI, P. (2020). Banking Dataset - Marketing Targets. kaggle. Retrieved from https://www.kaggle.com/datasets/prakharrathi25/banking-dataset-marketing-targets

S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimaraes, Portugal, October, 2011. EUROSIS. [bank.zip]. Retrieved from https://core.ac.uk/download/pdf/55616194.pdf

S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014 Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S016792361400061X