ST. CLAIR COLLEGE OF APPLIED ARTS AND TECHNOLOGY

Ace Acumen Campus, Mississauga

Final Project Report

Bank marketing campaign prediction by using Machine Learning Algorithm

Submitted to :- Mrs. Savita Sherawat

Submitted by :- Aafiya Vahora (0785323)

Period:- May 2022 – August 2022

Table of content:-

Abstract:-

The information pertains to a Portuguese banking institution's direct marketing campaigns (phone calls). The classification purpose is to determine whether the client will sign a term deposit agreement (variable y).

Data Set Specifications

The information pertains to a Portuguese banking institution's direct marketing campaigns. Phone calls were used in the marketing activities. In order to determine whether the product (bank term deposit) would be subscribed ('yes') or not ('no'), many contacts to the same client were frequently required.

Keywords: Classification, Logistic regression, Backward Elimination, Random Forest and Artificial neural network, One hot encoding, K nearest neighbor, K-fold.

Research question: This is a binary classification problem, as it might have guessed by now. Our two classes are "yes" for customers who have signed up for a term deposit and "no" for customers who have not signed up.

Objective: The information pertains to a Portuguese banking institution's direct marketing campaigns(phone calls). The classification purpose is to determine whether or not the client will sign up for a term deposit (variable y).

Tools: Python, Machine learning algorithm.

GitHub Source:-

https://github.com/Aafiya1997/Predictive-Analysis-on-Bank-marketing

Attribute information

Bank client data:

- Age (numeric)
- Job: type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
- Marital: marital status (categorical: 'divorced', 'married', 'single', 'unknown' ; note: 'divorced' means divorced or widowed)
- Education (categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
- Default: has credit in default? (categorical: 'no', 'yes', 'unknown')
- Housing: has housing loan? (categorical: 'no', 'yes', 'unknown')
- Loan: has personal loan? (categorical: 'no', 'yes', 'unknown')

Other Attributes:

- Campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- Pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
- Previous: number of contacts performed before this campaign and for this client (numeric)
- Poutcome: outcome of the previous marketing campaign (categorical: 'failure','non existent','success')

Social and economic context attributes

- balance- quarterly indicator (numeric)
- Contact: Which method is used to contact the customer (numeric)
- day: consumer confidence index(at which date the customer is contacted) (numeric)
- month: in which month the customer is contacted (numeric)
- duration: quarterly indicator (numeric)

Output variable (desired target):

y - has the client subscribed a term deposit? (binary: 'yes', 'no')

Literature Review:-

Introduction:-

Marketing is a part of every small and big business. Every organization's strategy is their marketing skills. Marketing means actions taken by company to promote their products. It depends on which way businesses describes, offer, sale their products to the customers. How they interact with consumers that can make them buy their company's product. How fare and advantageous their deal is for their customers to have trust on them. A good marketing skill defines company's image.

Marketing strategy is a long-term plan for achieving a company's goals by understanding the needs of customers and creating a distinct and sustainable competitive advantage.

The 4 P's of Marketing: -

The 4 P's of marketing are place, price, product and promotion

Product: - It says about the product that what kind of product it is. What kind of features it has? What kind of service it provides? It can be anything either a thing or an online application. Product means that can be anything that used by others

Price: - It describes about price of the product. How much a customer going pay for it. Is it cost-effective or not and how are they going to offer a discount to customers, how much discount can apply on product? How they can provide monthly subscription of their product.

Place: -

It says about where a customer looking for your products. In a shop or online. A right place of your product to sell. Where can a customer find it easily. It matters a lot. It is more convenient for a customer to order online product by wasting time to visit a particular store.

Promotion: - It says that in what way we describe our products. How we going to describes each feature of the products to attract more customers. It says about promoting a product in a best way.

As a typical strategy to promote business development, marketing activities can generally be divided into mass marketing and direct marketing. Mass marketing is the use of newspapers, radio, television and other media to promote the general public, while direct marketing is through mobile phones, fixed phones, Email, etc. directly contact customers to promote products or provide customers with discounts. In today's highly competitive market environment, mass marketing is

no longer an effective and reliable method, and marketing is moving from traditional mass marketing moving to direct marketing (Elsalamony, 2014).

Banks usually have a large number of databases consisting of customer information and transaction information. This can not only provide banks with accurate and timely business and management information, but also perform functional query, analysis, and decision advice on information, and provide detailed information support for marketing activities (An, 2007).

Scope of the proposed work: -

The main purposed of this bank marketing campaign (by phone calls) was to predict if the customers will subscribe a long-term deposit 'Y' (yes or no). The goal of the study is to develop a predictive model capable of improving or increasing the efficiency of directed campaign for long term deposit subscriptions by reducing the number of contacts to do; that is a reduction in the number of customers to be contacted by phone. By using a classification model, we will predict this result. The classification model includes Decision tree, Random Forest, Logistic regression, Support Vector Machine (SVM), KNN, Naïve beyesian and many more. Here in this dataset, we will apply some classifiers to get the result.

Logistic Regression: -

Logistic regression is an example of supervised learning. It is used to calculate or predict the probability of a binary (yes/no) event occurring. An example of logistic regression could be applying machine learning to determine if a person is likely to be infected with COVID-19 or not.

LR and DT have the advantage of fitting models that tend to be easily understood by humans, while also providing good predictions in classification tasks. (S. Moro, P. Cortez and P. Rita, June 2014)

Decision trees: -

Decision tree builds classification models in the form of a hierarchical structure. Decision tree is developed through step-by-step incremental process of breaking down the dataset into smaller and smaller. At final process it generates a tree with decision nodes and leaf nodes. A decision node has two or more branches. Leaf node represents a classification or decision. The root node in a tree which corresponds to the best predictor from given datasets. (Sandhya N. dhage ,March 2018)

K Nearest Neighbors: -

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition based on their nearest neighbors and it should odd number. (Sandhya N. dhage ,March 2018)

Support Vector Machine (SVM): -

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. (Sandhya N. dhage ,March 2018)

Random Forest: -

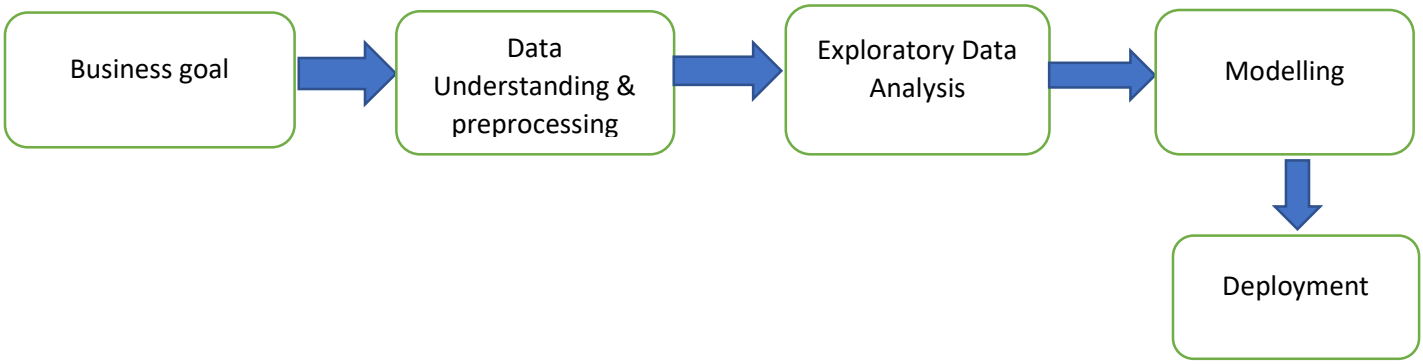
Random forest offers a natural way to rate the importance of variables since different variables being left out of the trees fitted in our forests are permuted (Aslett, L. J. M., Esperan¸ca, P. M. and Holmes, C. C. , 2015).

Random forest model in machine learning results in a better classification of the factors affecting. The potential for tuning interaction terms of the random forest like logistic regression offers more optimal prediction. In addition, they offer more optimal credit assessment out of confusion matrix and invaluable metrics in assessing the scientific relationship of data to predictive value through the variable importance metrics (Sharma, 2012).

Neural Network:-

Deep convolutional neural network, one of the promising deep neural networks, can handle the local relationship between their nodes which can make this model powerful in the area of image and speech recognition. (Kim, K. H., Lee, C. S., Jo, S. M., & Cho, S. B. ,2015)

Methodology:-



Business Goal:- To identify purpose of the business. The main goal to start this business.

Data understanding & Preprocessing:- It is very important to have detail data knowledge or to understand our data very well before pursuing further analysis. Data Preprocessing is like data preparation that converts raw data into understandable format. Importing necessary libraries to read csv and perform next tasks like checking data types and cleaning part.

Exploratory data Analysis:- It is about visualization and detection of outliers. Visualization of each field and comparison comes here.

Modelling:- Applying necessary Machine Learning models.

Deployment:- Deployment is the process of presenting a predictive model.

Data Preprocessing:-

- The very first I imported necessary libraries in Python like, Pandas to read csv file in pyth on.
- Then I Imported Numpy for Mathematical calculations in Python and Warning To ignore warnings in python.
- Seaborn and matplotlib for Visualization. Then I create a dataframe named df to store csv file in it.
- Then by using shape command I checked how many rows and columns of my dataset.
- After that by using dtypes I checked the data types of each attributes.
- Then I used describe command to get a statistical summary of a dataset.
- Hence, I changed all the object attributes to category by using astype command.

Missing Values:-

by using isnull function I checked the null values in my dataset and I found that I'm not having a ny null values in my dataset

Visualize and detect outliers for numeric attributes:-

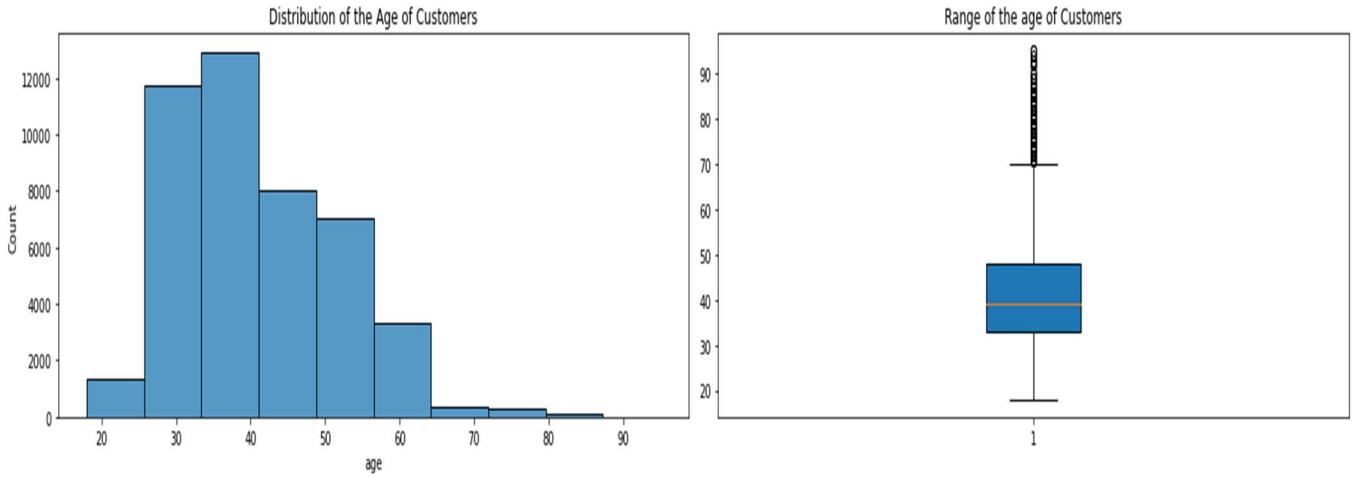


Figure 1 outliers for Age attributes

Here, I've visualized outliers for Age attribute. The first one is a histogram with the title of distri bution of the Age of customers. I take 10 bins here. We can clearly see the histogram is a right s kewed. On the right side I've made a boxplot named as Range of the Age of Customers for the ou tliers, and here we can clearly see the outliers. After that I find the values of Q1, Q3, IQR, upper bound and lower bound to drop the outliers. Mention below is the visualization of a removed ou tliers: -

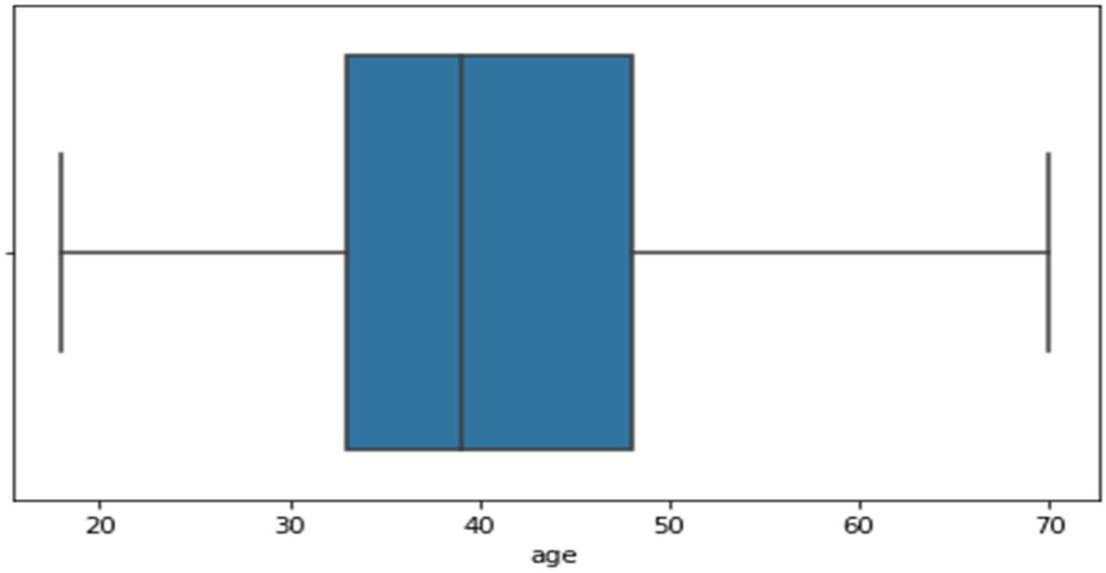


Figure 2 detected outliers of Age attribute

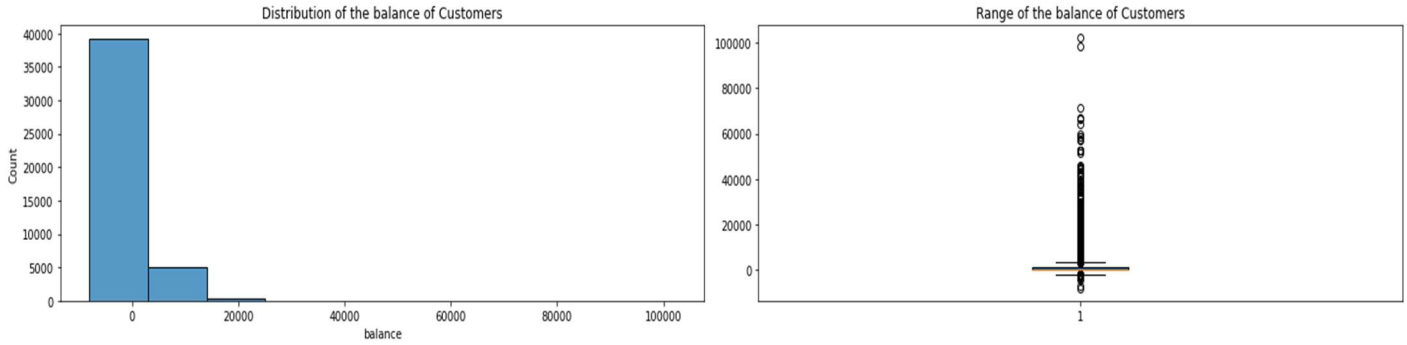


Figure 3 For balance attribute

It shows the outliers of Balance attribute, histogram with title of distribution of the balance of cu stomers with the outliers of range of the balance of customers. We have many outliers in balance attribute, again I remove the outliers by implementing statistical operations, and here's the visual ization of the balance attribute decreased outliers

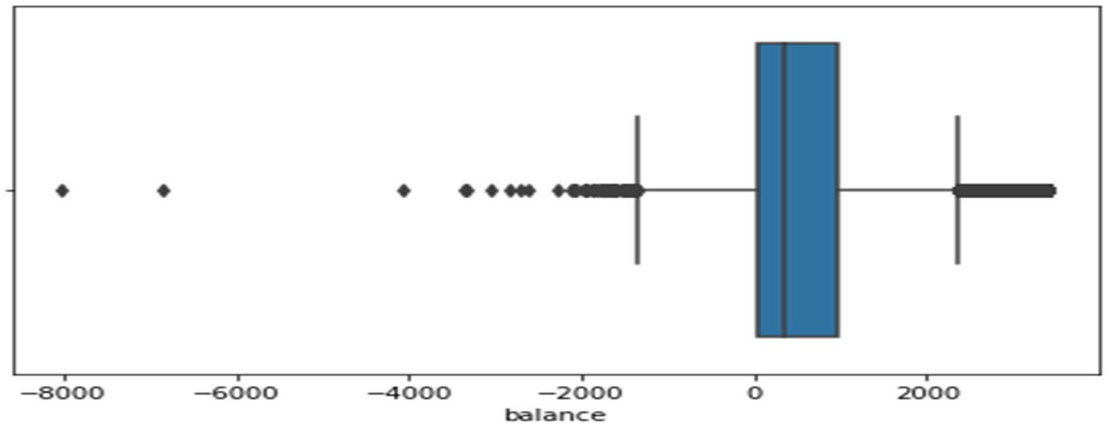


Figure 4 removed outliers in Balance attribute

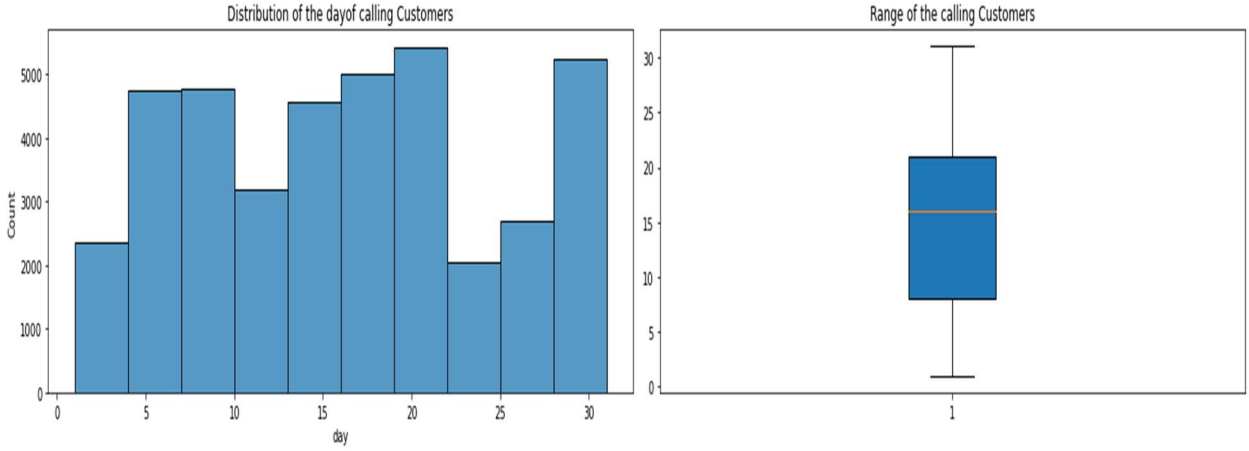


Figure 5 Outliers of day attribute

Here, it is the outlier of day attribute, in which I've taken 10 bins . We can clearly see that we do n't have any outliers in day attribute, so here I am not removing any outliers from this attribute, I' m keeping this Outliers as it is.

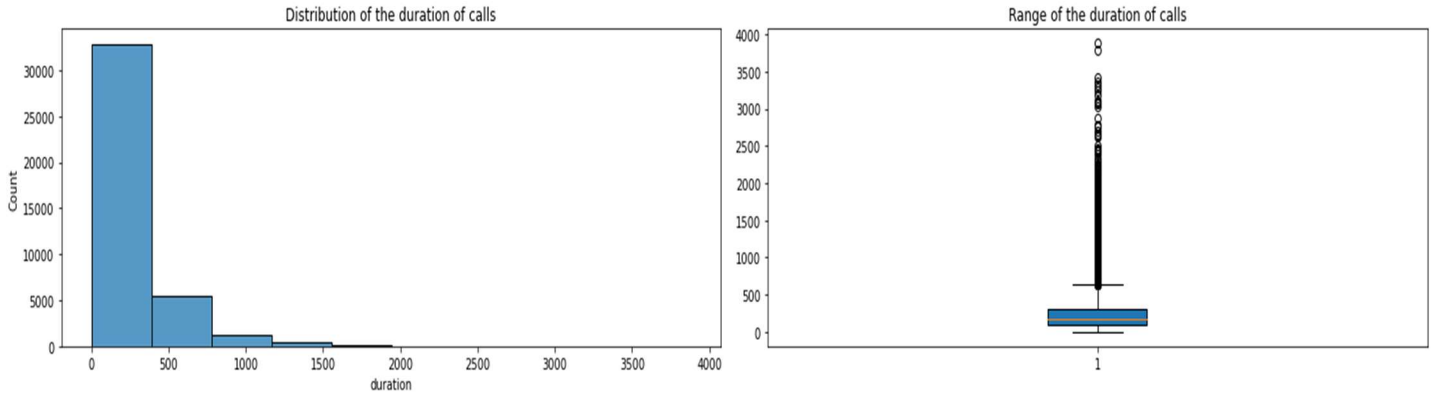


Figure 5 Distribution of call duration

It is the outliers of Duration attribute. The first one is the histogram named as Distribution of the duration of calls, as shown in figure it is a right skewed and besides this is a box plot of a duratio n attribute. It clearly shows that it has many outliers, so again I find the values of Q1, Q3, IQR, u pper bound and lower bound to replace this outliers, mention below is the figure for removed outl iers in our boxplot.

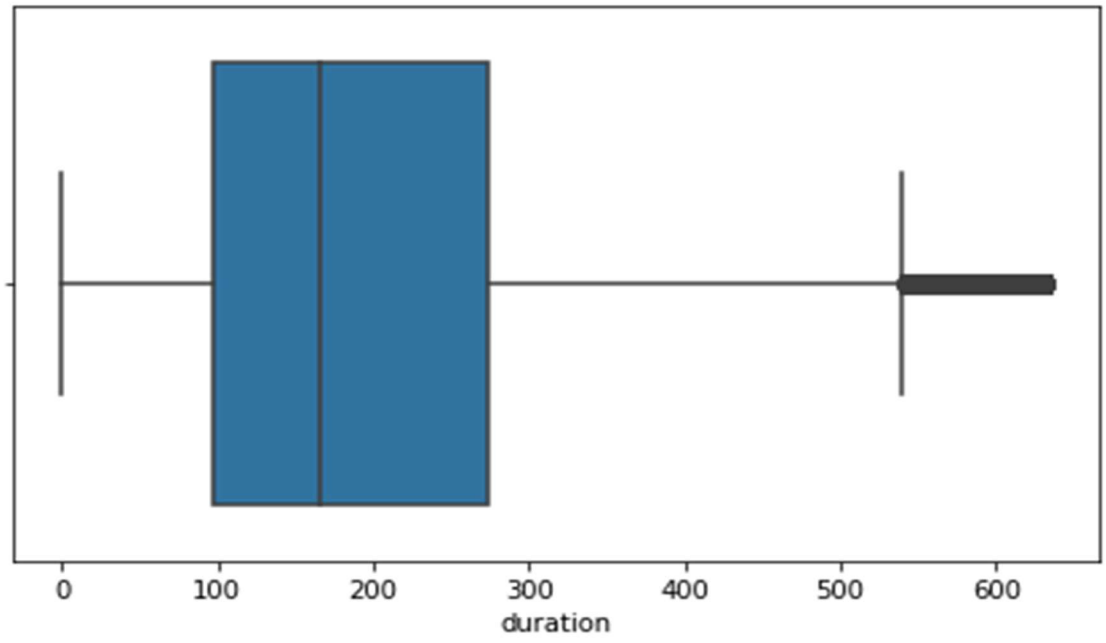


Figure 6 replaced outliers for call duration

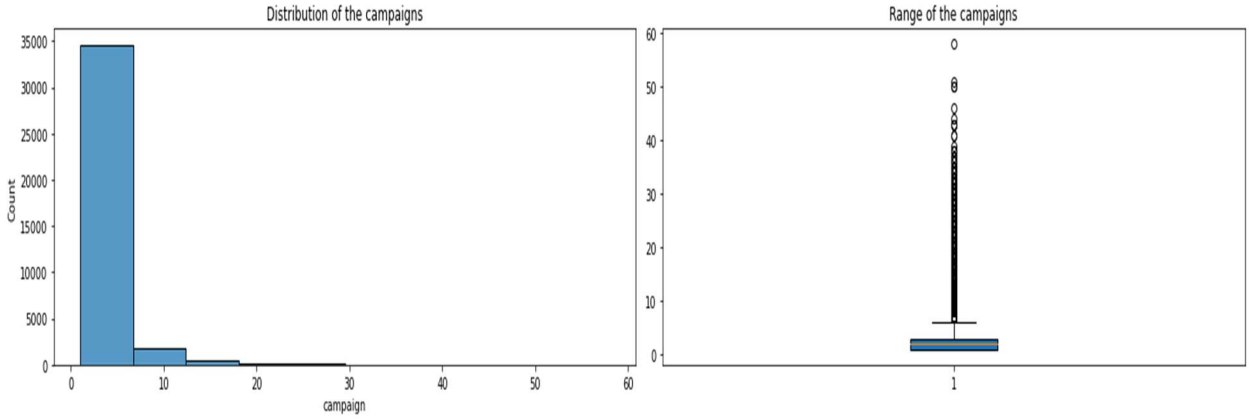


Figure 7 For previous campaign

It is the histogram and box plot for campaign attribute. In the box plot we can clearly say in the r ange of 0 to 1 we have many outliers in campaign attribute. By using statistical calculation to re move or replace outliers, we can detect outliers by finding values of Q1, Q3, IQR upper bound a nd lower bound to remove outliers. Below are the outliers after detection of it.

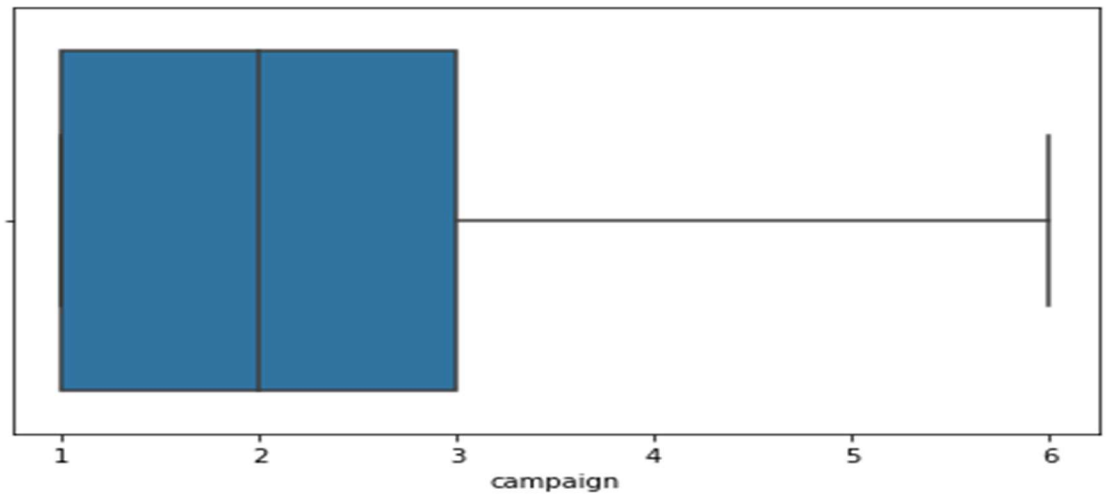


Figure 8 Removed one

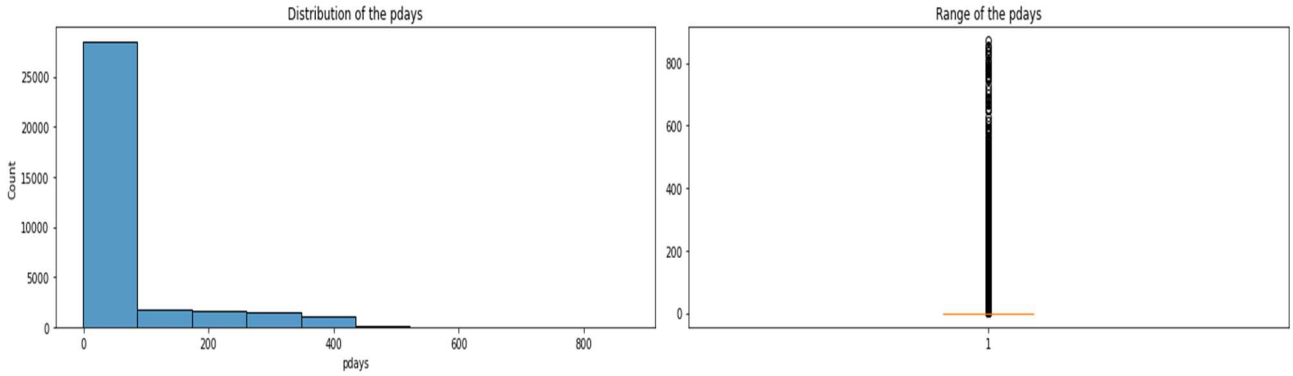


Figure 9 Visualization for Pdays Outliers

Here the outliers of Pdays attribute. It shows about how much times have left after last call to the customers. It differentiates time between last campaign and current one. By using statistical tools, we removed the outliers. The mention bellowed is removed outliers from Pdays attribute.

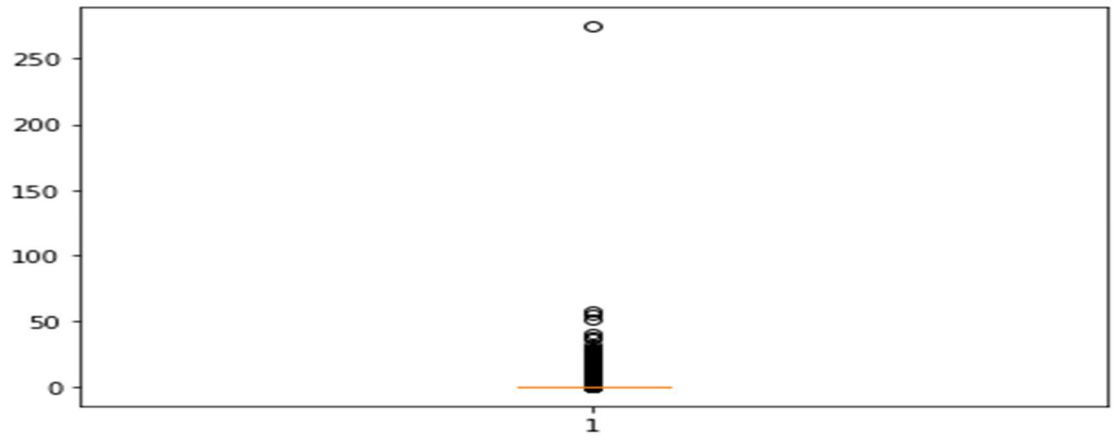


Figure 10 Removed outliers in Pdays

Correlation heatmap: -

Then I check the correlation of a numeric attribute then created heatmap of it. Here I've used spearman correlation method. We can say that there is no correlation between variables in my dataset, as the values are close to 1 then it is correlated, and here I'm not having any value close to 1, so there is no correlation.

Figure 11 Heatmap of a dataset

Visualization of Categorical Attribute: -

For Default attribute:-

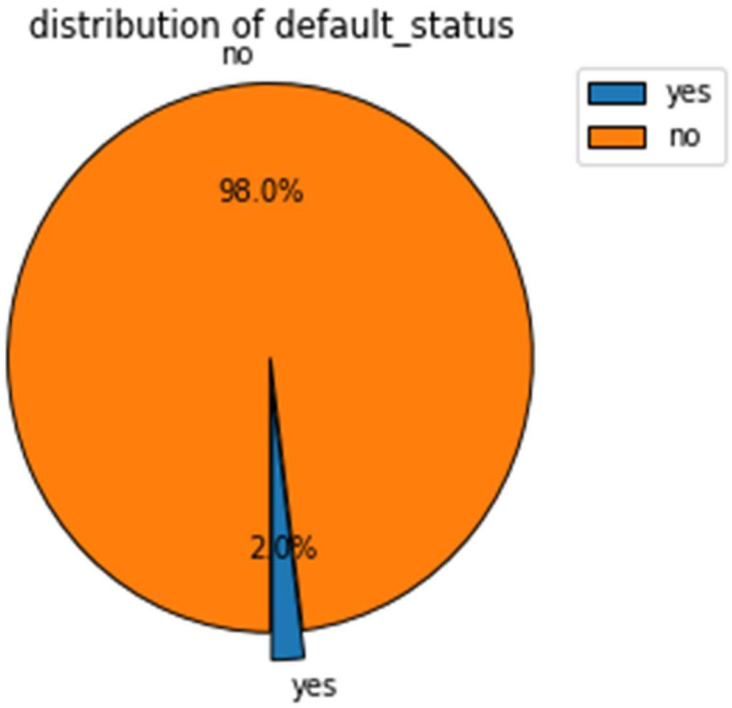


Figure 12 Default status visualization

First, I've checked the unique values of the default attribute. It has a two category in a form of yes and no. The default attribute stands for, is a customer having their bank status as a default or not. I've created pie chart for this attribute with the title of distribution of a default status it says that 2% of customers are having their credit in default, while 98% of the customers do not have their credit in default.

For Marital attribute: -

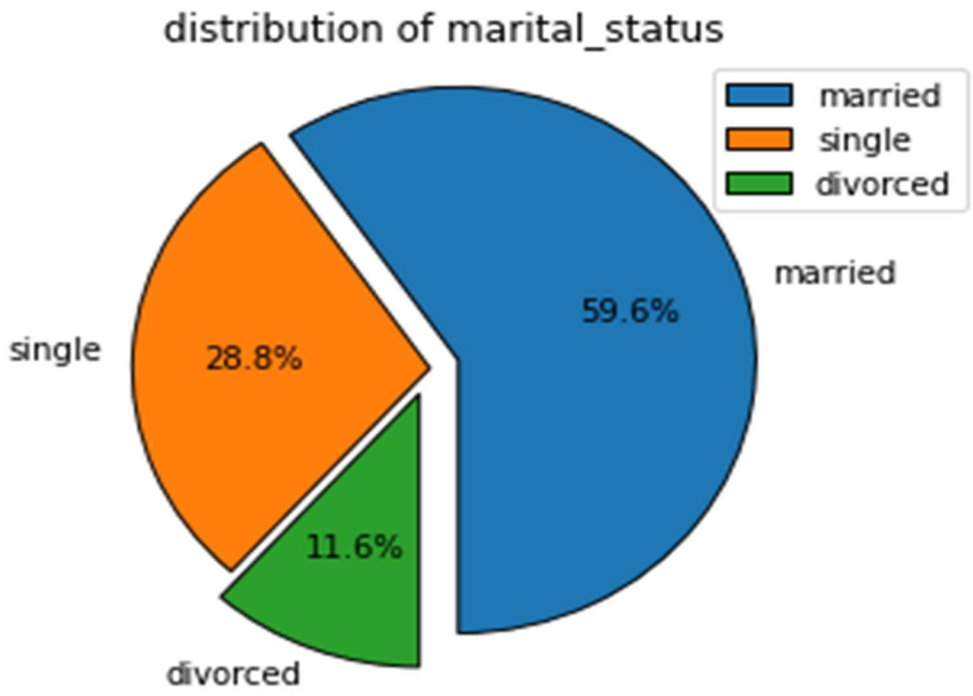


Figure 13 marital status visualization

Here, the marital attribute in my dataset provides information about customers marital status. It has a three-category single, married and divorced. This pie chart shows that the Portuguese bank is having 59.6% customers are married marked as a blue color, 28.8% are single and 11.6% are divorced customers.

For Loan attribute: -

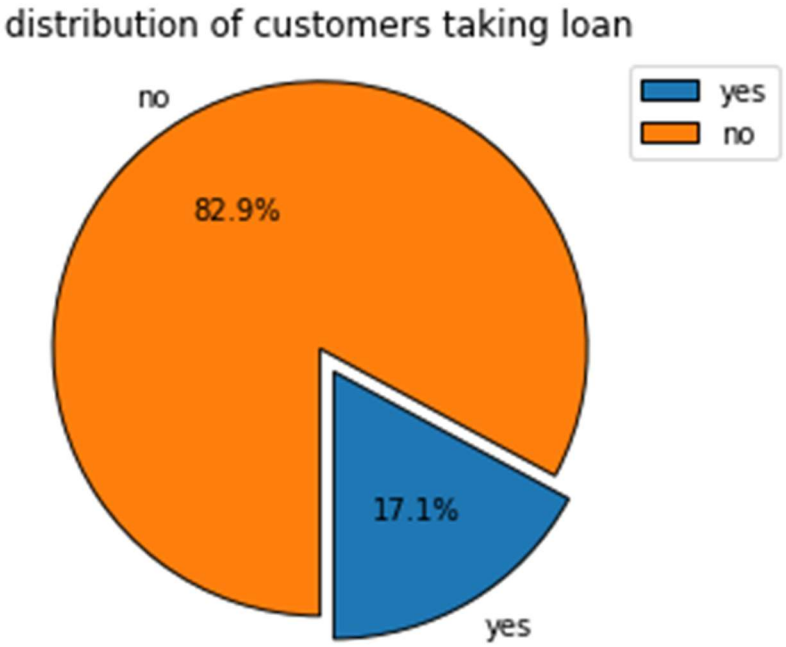


Figure 14 chart for loan attribute

The loan attribute tells us about the loan status of the customers. It shows that is the customers are having personal loan or not. It has two categories as a yes and no. In my dataset only 17.1% of clients are having personal loan while majority of people(82.9%) are not having any personal loan. I named this pie chart as a distribution of a customers taking personal loan.

For Housing attribute:-

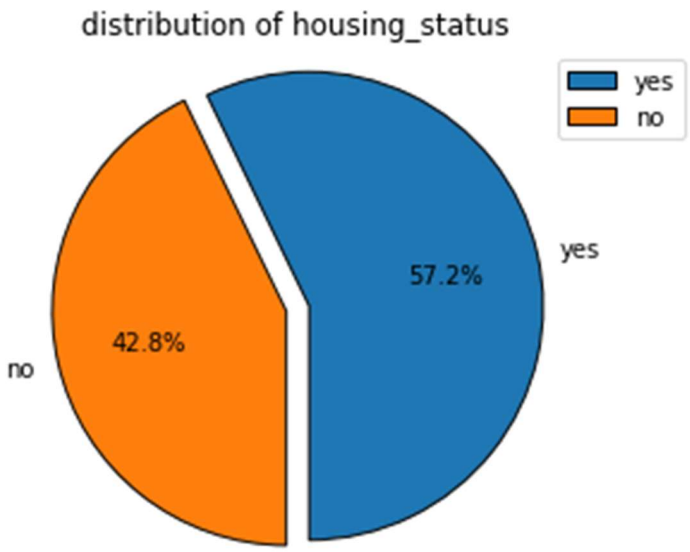


Figure 15 Housing status of customer

This pie chart titled as a distribution of housing status says about housing loan. It shows that 57.2% of customers are having housing loan which are labeled as a blue and rest which are not having any housing loan marked as a orange.

For Contact attribute: -

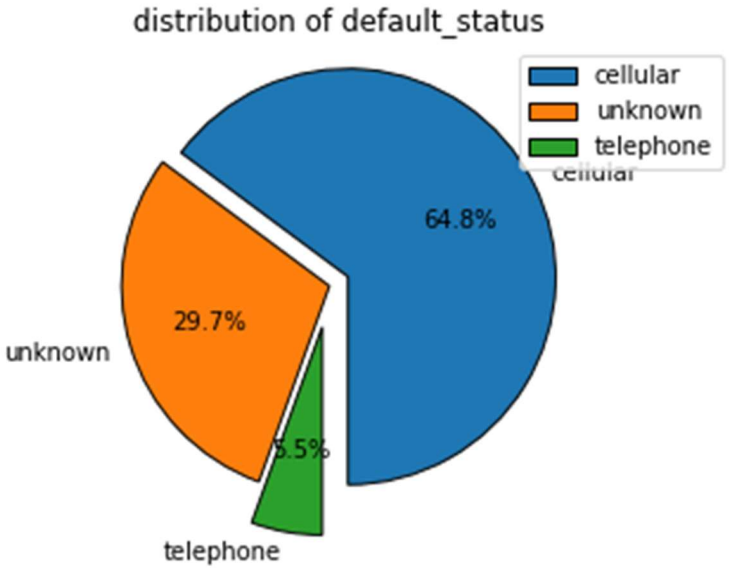


Figure 16 Contact method off campaign

The Contact attribute states that how a customer got contacted. In which way bank contacted customers in this campaign. First, we checked the unique values here, it has a three categorical attribute cellular, telephone and unknown. The pie chart shows that 64.8% of clients are contacted via cellular, 5.5% via telephone and rest method are unknown.

For Education attribute: -

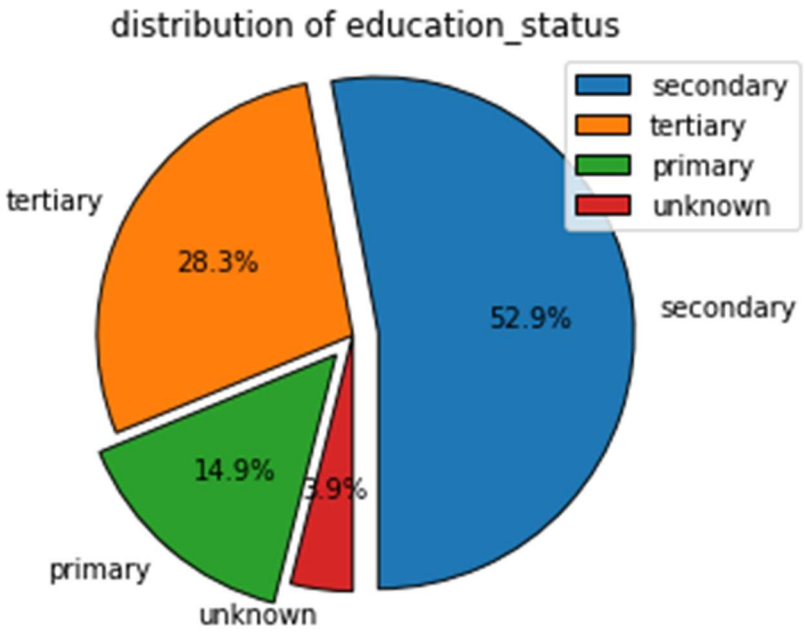


Figure 17 Education status of the customers

This attribute says about education status of the customers. It divides into primary, secondary, tertiary and unknows. 52.9% of customers are having secondary education and rest of them are having tertiary and primary, while some of them are not even mentioned so that information is marked as unknown.

For Job attribute:-

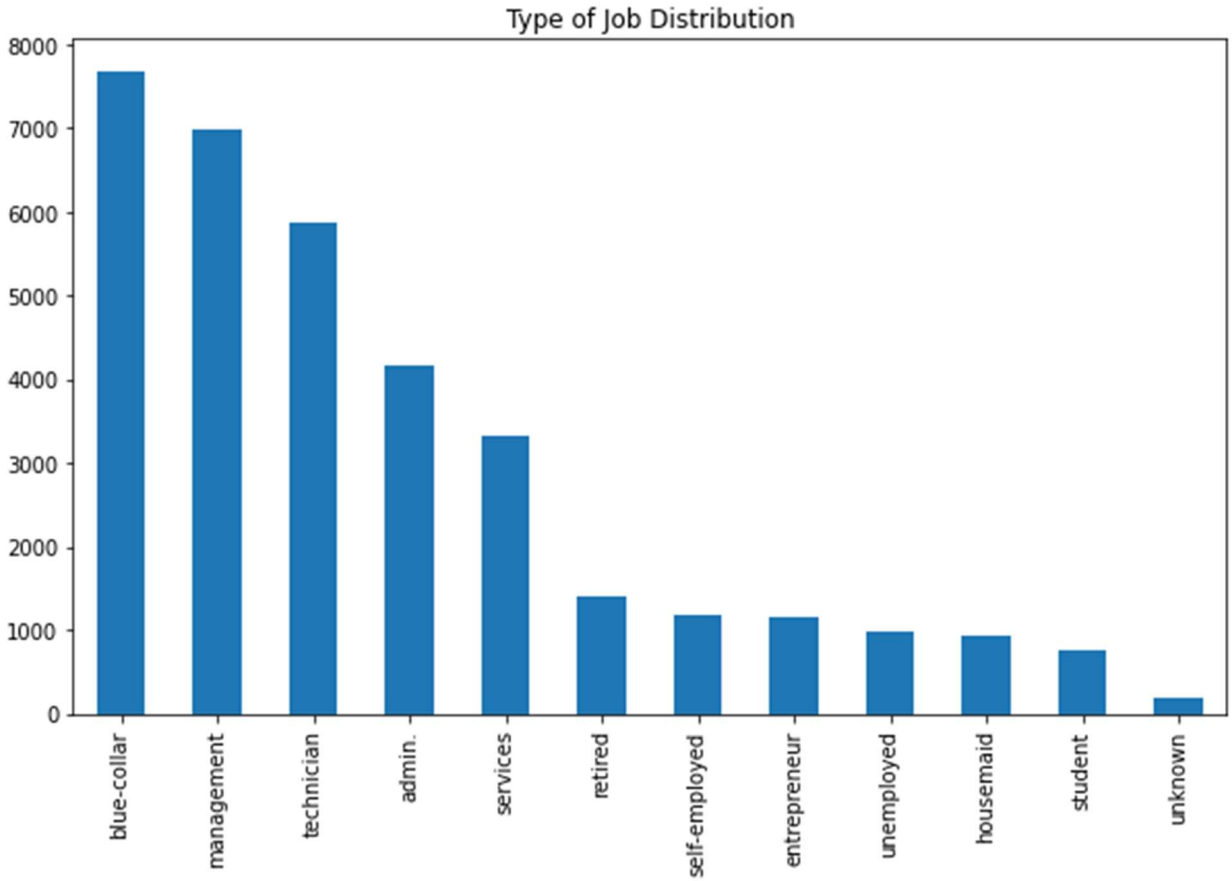


Figure 18 Employment status of the customers

This is the bar chart for job attribute. It has a information about employment status of clients. It shows that most of the clients of bank are having blue collar job, some are retired, some are unemployed while some of them are student and many more categories are there

For Poutcome attribute: -

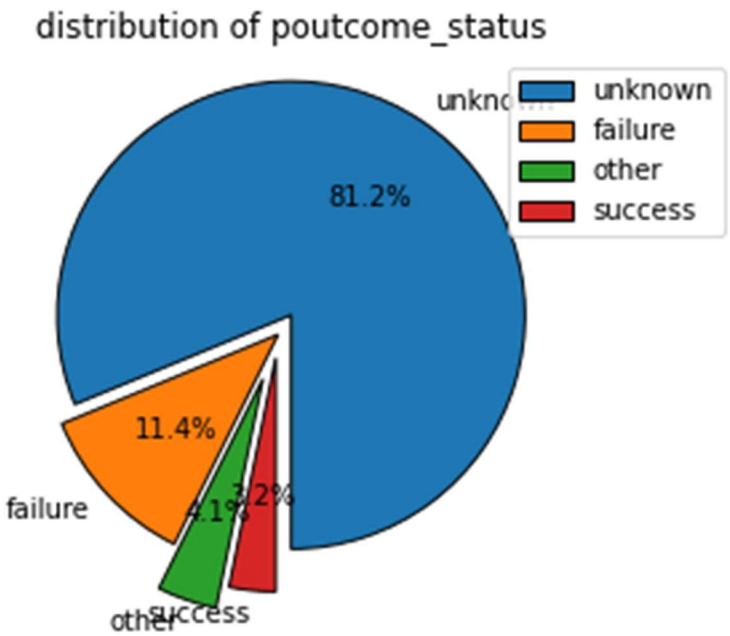


Figure 19 Previous outcome of campaign

This poutcome attribute gives us information about previous campaign results details. It's let us know about how was the previous campaign. It has four categorical attributes as a failure, success, others and unknown. This pie chart shows that majority result of previous outcome was unknown, while it has only only 4.1% success rate and 11.4% failure rate, which is higher than success rate.

For y attribute:-

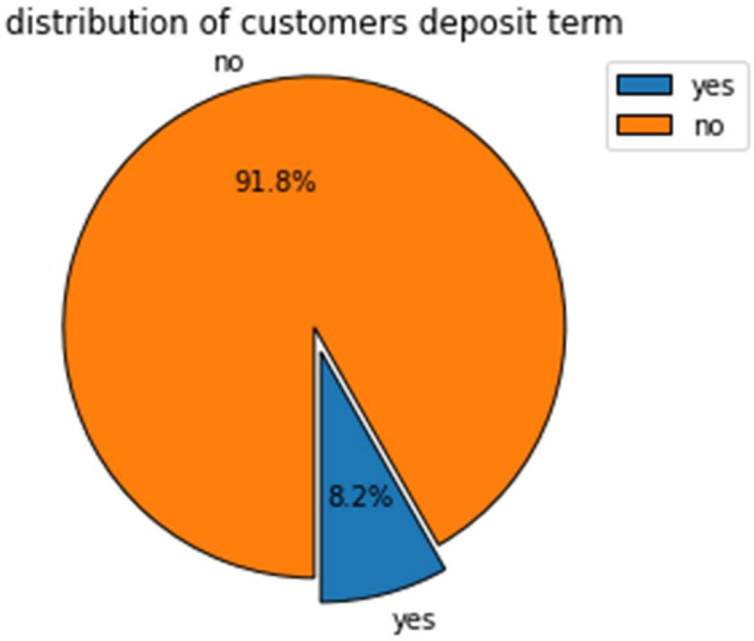


Figure 20 deposit term subscribtion

This is my target variable. It has information about customer deposit term. It shows that 91.8% of clients are not subscribing the bank deposit, however only 8.2% of the customers subscribing the bank deposit term.

Comparison of the target variable with rest of the variable: -

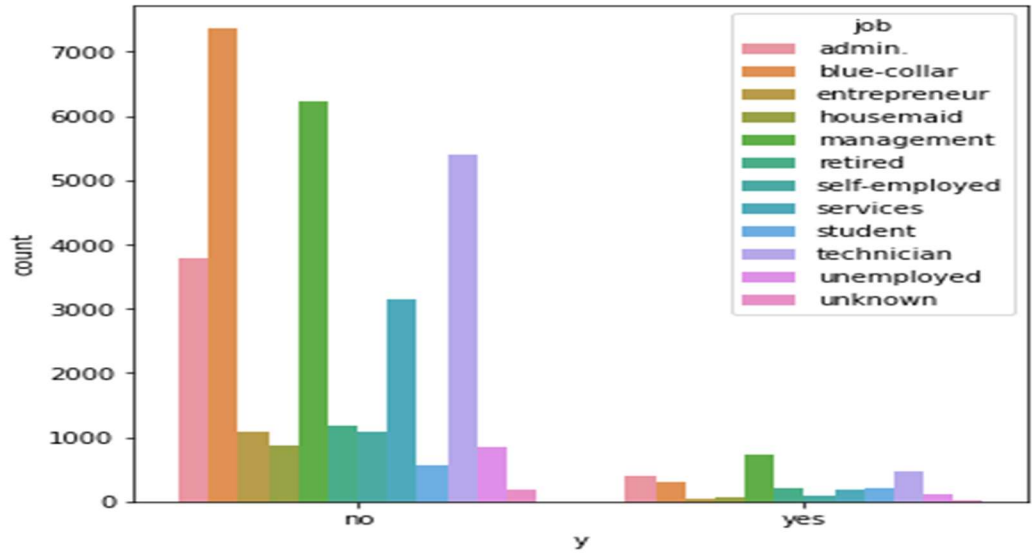


Figure 21 comparison of job attribute with job

Here, I've compared my target variable with job attribute, it shows comparison that how many of clients subscribed the term from different employment status.

Comparison between Marital and target variable: -

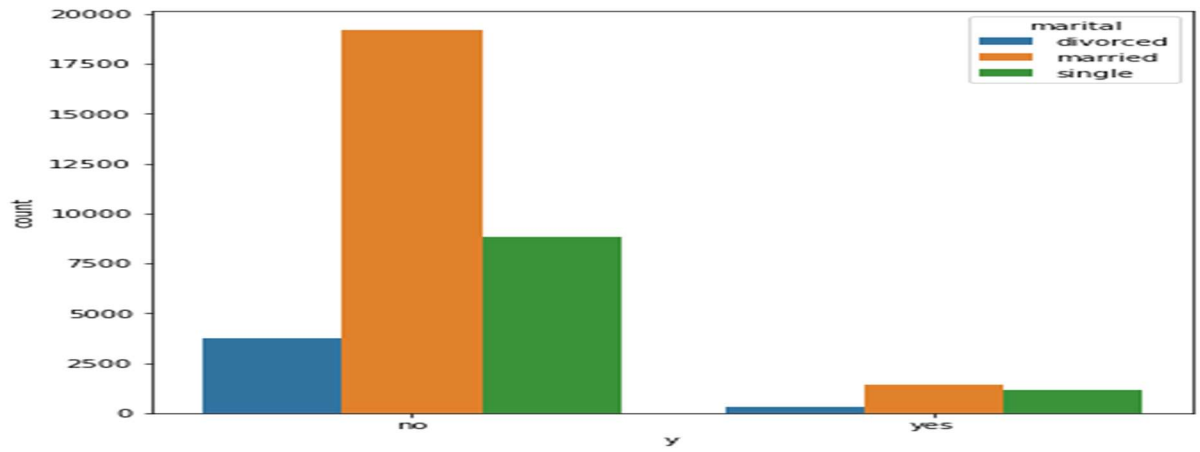


Figure 22 Marital status in terms of target variable

Here, it shows that how many of married, single and divorced has subscribed the deposit term.

As we can see that I've majority of marital status vote as no.

Comparison with Education attribute:-

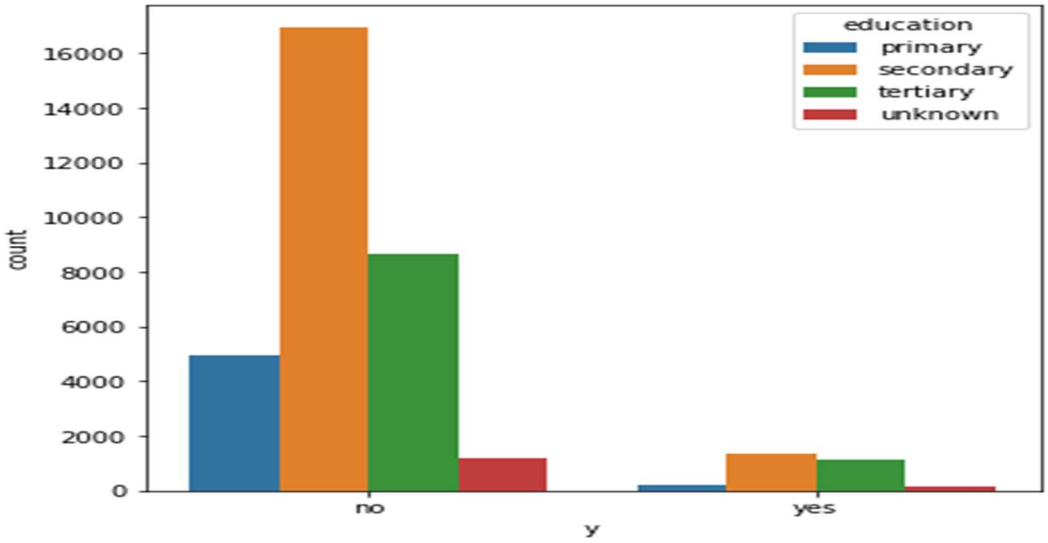
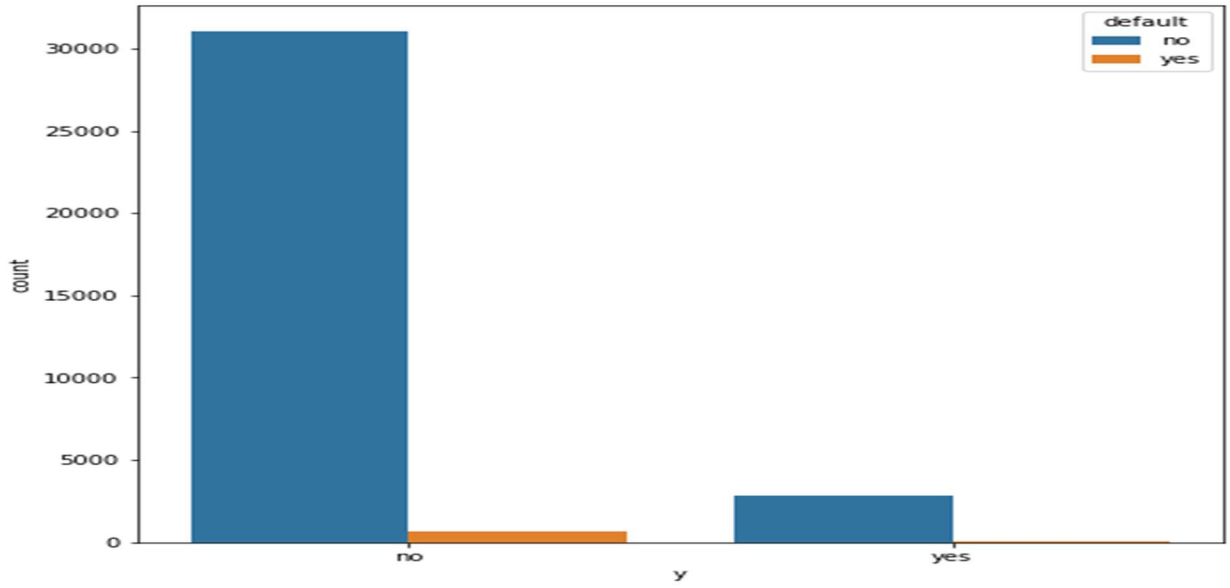


Figure 23 target variable according to education status

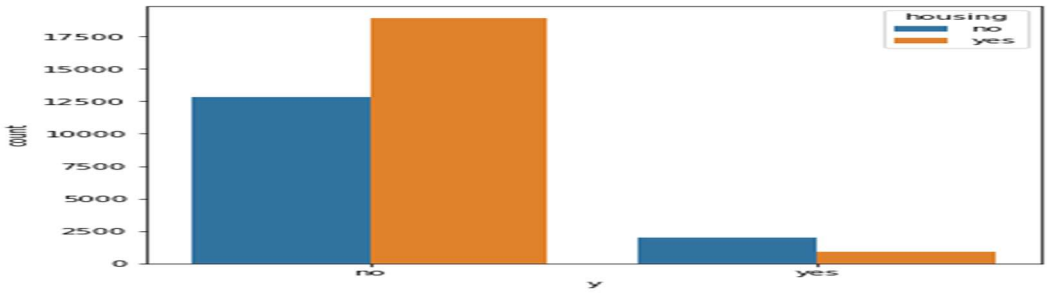
Here, it differentiate between education status of the customers who have subscribed the deposit term and who has not.

Comparison with default attribute:-



Here it shows that most of the customers are not having their credit in default and this are the same who haven't subscribed the deposit term.

Comparison with Housing attribute:-



Here it compares that customers who are having housing loan and also has subscribed the bank deposit term. It compares both yes and no with housing loan

Comparison with loan attribute:-

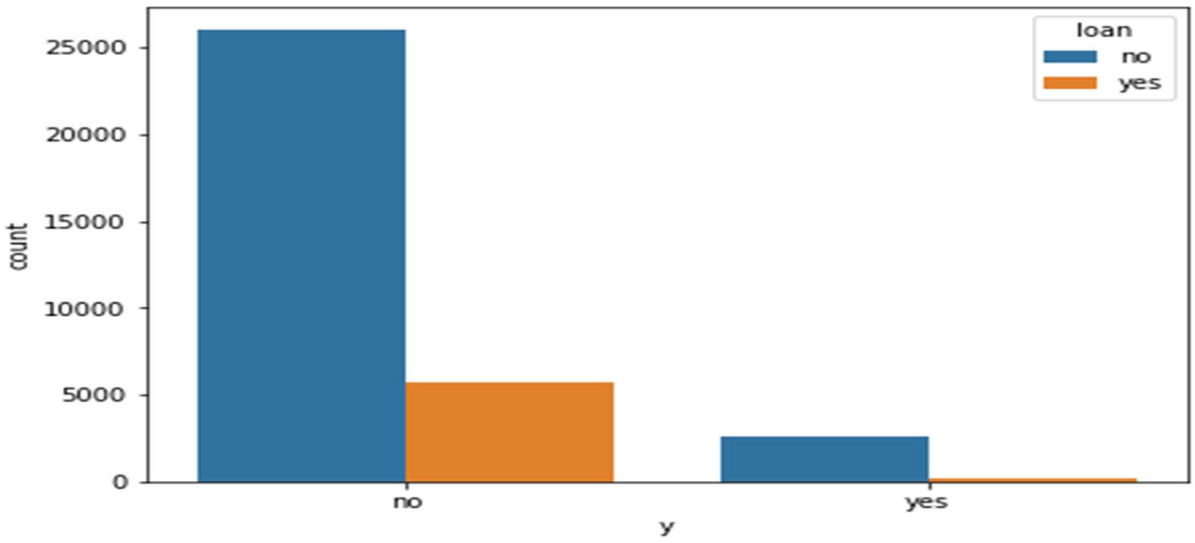


Figure 24 loan status as per deposit term

Here it depicts that customers with personal loan does subscribed the deposit term yes or no

Comparison with Contact attribute:-

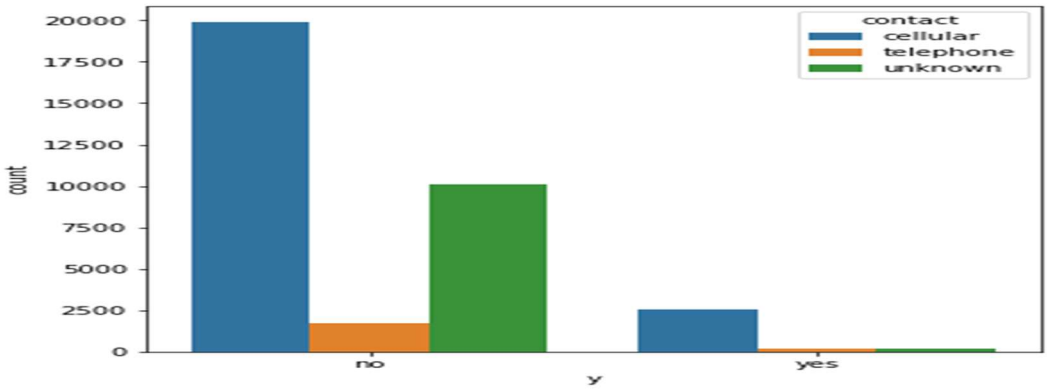


Figure 25 deposit term with methods of contact

It shows how a customer get contacted in this campaign. It differentiate how the yes one are contacted and how the no one are contacted

Comparison with month attribute:-

Between months and

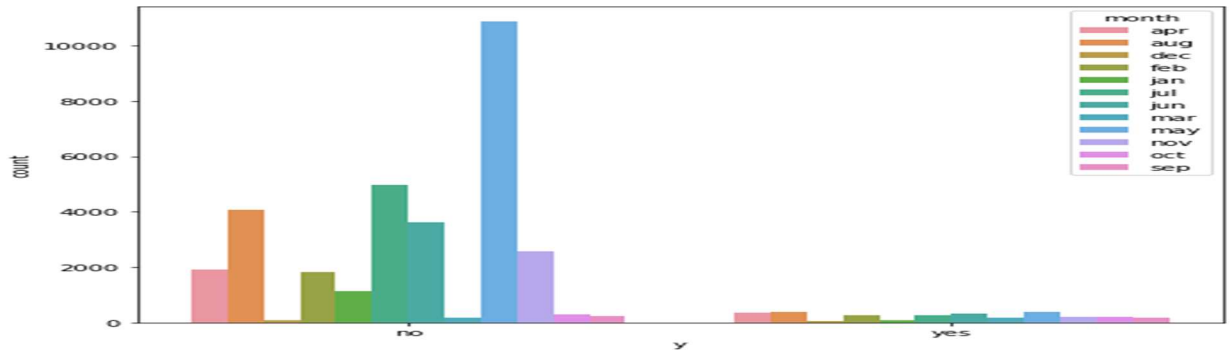


Figure 26 between job and month

It shows the bank deposit term according to a month of contacted customer.

Comparison with Poutcome variable: -

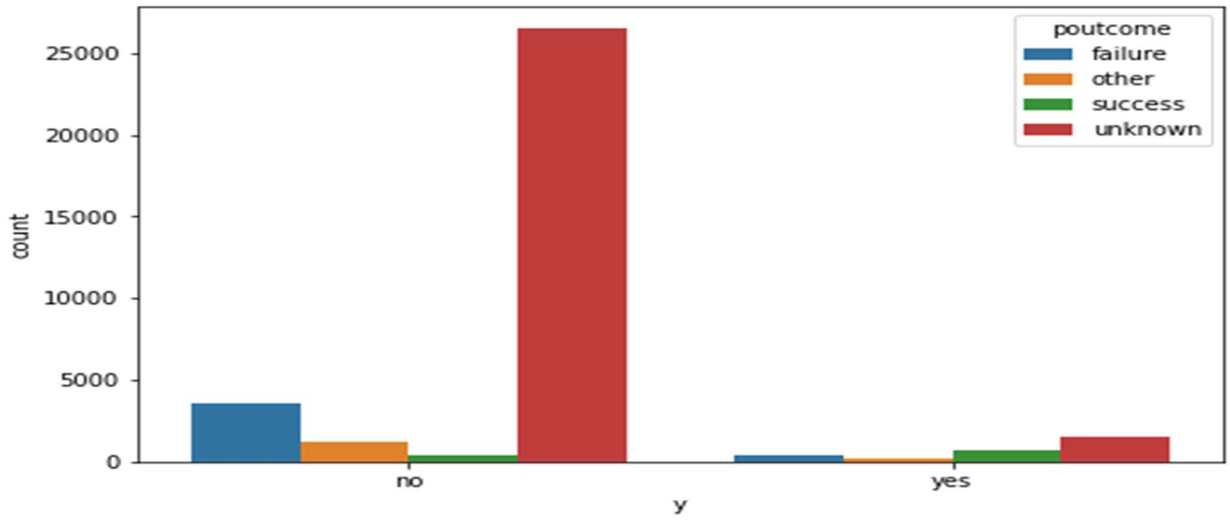


Figure 27 Previous outcome with current outcome

It compares previous outcome result that are failure , success, others and unknown with the bank deposit term.

For Normalization:-

We perform normalization to remove unnecessary and duplicated data from our dataset for better visualization. After visualization I've converted all attributes except target variable to numeric using one hot encoding method for normalization. In this method first I made a subset of the attributes, then I made X variable to store all the dummies in there, or to store converted numeric attributes there.

Then I import the Libra MinMaxScaler from sklearn to scale my dataset into Min-Max. In this I named a variable scaled_df to create one more data frame to fit and transform X. In other words I scaled X variable to Min and Max. Next I create a variable Y to store my target variable

Train-Test-Split:-

To split my dataset into train and test I imported train-test-split from sklearn.model selection.

Here I put test size as 0.20 to test 20% of my dataset and rest for training. Then I checked if my my target variable is balance or not by using .value_count() command and I got to know that my target is not balanced.

SMOTE:-

To balance imbalanced dataset I've used SMOTE here. First I imported smote from imblern.over sampling. Then I resample it with smote, and checked whether my dataset is still balanced or not, as a result it is balanced.

Classification models:-

After balancing my dataset I applied classification models to find accuracy and check which model suits my mode best.

Logistic regression:-

Logistic regression is used to predict the probability of categorical variable. First I import logistic regression from linear model and fit my LR model in a variable named model. Then I predict my model using .predict() command and find accuracy. I got an accuracy of 84.12% from this model.

K nearest neighbor:-

Knn is a classifier that gives decision based on its neighbors. I imported knn from sklearn.neighbors. Then I check the accuracy and it is 84.5%

Random Forest Classifier:-

Here, I imported Random forest classifier from sklearn.ensemble. I got an accuracy of 92.82%. Which is highest accuracy among all the models

Decision Tree:-

Decision Tree is a Supervised Machine Learning Algorithm that uses a set of rules to make decisions.

In this classifier I got an accuracy of 89.74%

K – Fold cross validation:-

In this method, we split the data-set into k number of subsets(known as folds) then we perform training on the all the subsets but leave one(k-1) subset for the evaluation of the trained model. In this method, we iterate k times with a different subset reserved for testing purpose each time.

Conclusion: -

As my dataset is about predictive analysis, like whether the customer will subscribed the term deposit yes or no, so here I applied all the classification models and find accuracy from each model. Here I can say that I got the best accuracy from the random forest model, so here I can say that random forest gives best accuracy and which is best suitable model for my dataset.

REFERENCES: -

1. Machine learning Algorithm Decision tree - Sandhya N. dhage, International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 4 Issue: 3
2. Machine Learning Algorithm K nearest neighbor - Sandhya N. dhage, International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 4 Issue: 3
3. Machine Learning Algorithm SVM - Sandhya N. dhage, International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 4 Issue: 3

4. Machine Learning Logistic Regression - Sandhya N. dhage, International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 4 Issue: 3

5. Machine Learning Naïve Beyesian - Sandhya N. dhage, International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 4 Issue: 3

6. 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

7. 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]

8. Elsalamony, H. A. (2014). Bank Direct Marketing Analysis of Data Mining techniques. International Journal of Computer Applications, 85, 12-22. https://doi.org/10.5120/14852-3218

9. An, H.-X. (2007). On Our Commercial Banks Marketing Strategy Questions. Journal of Central University of Finance & Economics, 4, 42-48.

10. Aslett, L. J. M., Esperan¸ca, P. M. and Holmes, C. C. , 2015, Bank Marketing Analysis of Random Forest Classifier.

11. Sharma 2012, the random forest model in machine learning results in a better classification.

12. Kim, K. H., Lee, C. S., Jo, S. M., & Cho, S. B. (2015). Predicting the Success of Bank Telemarketing Using Deep Convolutional Neural Network. In 2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR) (pp. 314-317). Piscataway: Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/SOCPAR.2015.7492828