**Predicting Customer Response to Telemarketing Campaigns**

Finance Domain

Classification Technique

Harini G, 2048034

2MDS, Christ (Deemed to be University)

Nowadays, marketing spending in the banking industry is massive, meaning that it is essential for banks to optimize marketing strategies and improve effectiveness. Understanding customers' needs lead to more effective marketing plans, innovative product designs, and greater customer satisfaction.

**Project Description**

Increase the effectiveness of the bank's telemarketing campaign. This project will enable the bank to develop a more granular understanding of its customer base, predict customers' response to its telemarketing campaign and establish a target customer profile for future marketing plans. By analyzing customer features, such as demographics and transaction history, the bank will predict customer saving behaviors and identify which type of customers is more likely to make term deposits. The bank can then focus its marketing efforts on those customers. This will allow the bank to secure deposits more effectively and increase customer satisfaction by reducing unwanted advertisements for specific customers.

**Dataset**

The dataset contains 17 columns and 11162 observations. It has numeric and categorical variables. There are no missing and duplicated values present in the dataset. The dataset is very much clean and structured.

The outcome variable deposit is usually skewed.

Attribute Information:

1. Age
2. Type of job
3. Marital status(Categorical: ‘divorced’, ‘married’, ‘single’)
4. Education
5. Default: has defaulted on credit? (Binary: ‘yes’ or ‘no’)
6. Balance: has balance loan?
7. Housing: has housing loan? (Binary’: ‘yes’ or ‘no’)
8. Loan: has a personal loan? (Binary: 'yes' or 'no')
9. Contact: Cellular, Phone?
10. Month: Last contact month
11. Day: Last contact day
12. Duration: Last contact duration
13. Campaign: number of contacts performed during this campaign and for this client(numeric, includes the last contact)
14. Pdays: number of days that passed by after the client was last contacted from a previous campaign
15. Previous: number of contacts performed before this campaign and for this client(numeric)
16. Poutcome: outcomes of the previous marketing campaign(categorical: 'failure,' 'nonexistent, 'success')
17. Deposit: has the client subscribed to a term deposit? (binary: 'yes' or 'no')

## Data Exploration

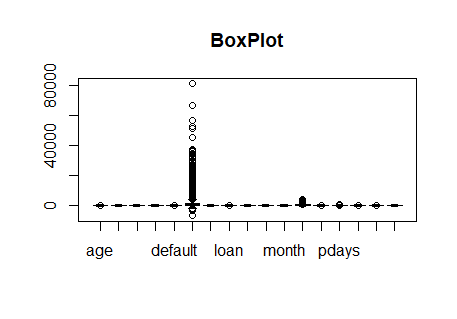
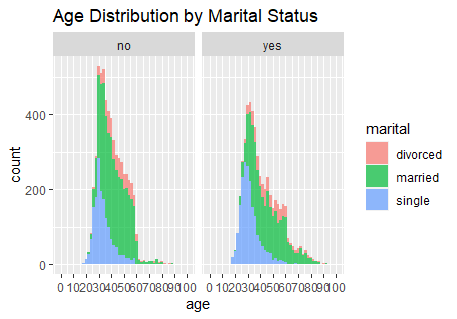


Fig1: Box plot of Numerical data

Box Plot for numerical data, as we can see, there are more outliers in the default column.



### Fig2: Age Distribution vs. Marital Status That Subscribes Term Deposit

The bulk of clients are married or divorced. The sharp drop of clients above age 60 with marital status 'divorced' and 'married.' \*Single clients drop in numbers above age 40.

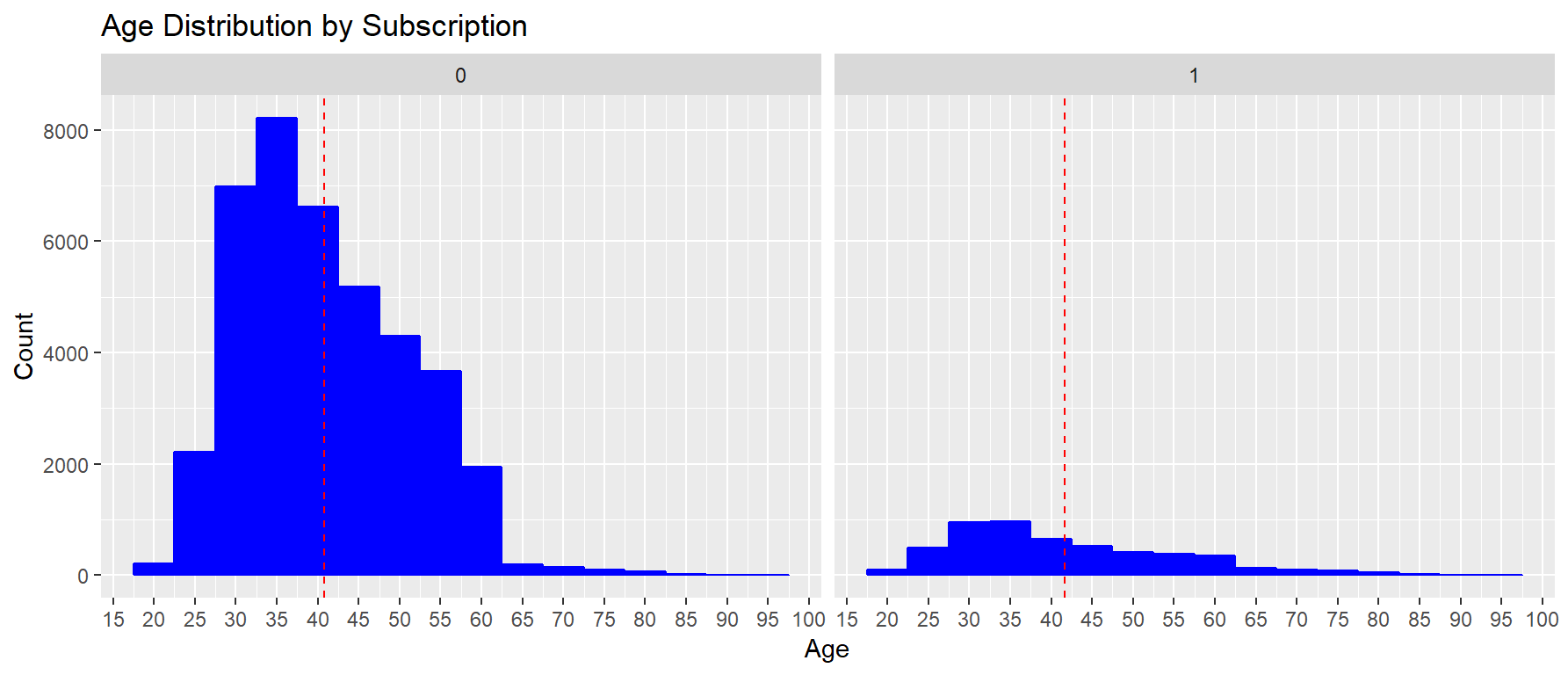


Fig3: Age Distribution by Subscription

Most clients that subscribe are between the age of 25 to 45. The mean age for all clients is above 40 years of age.

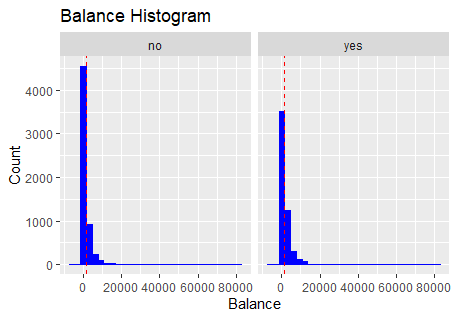
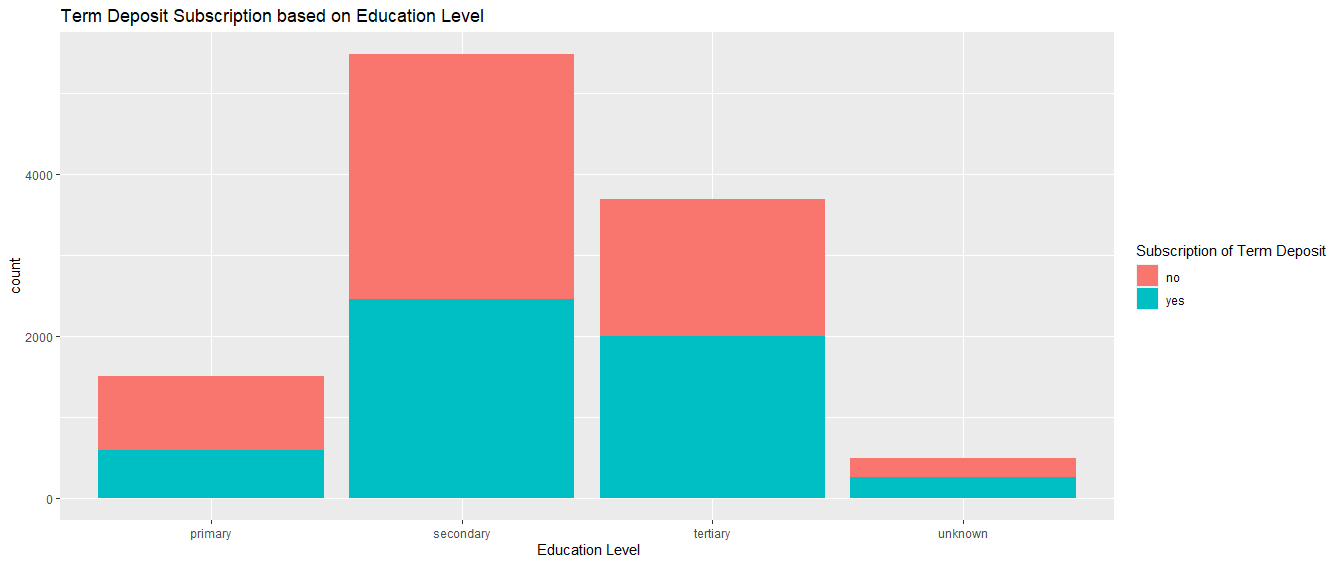


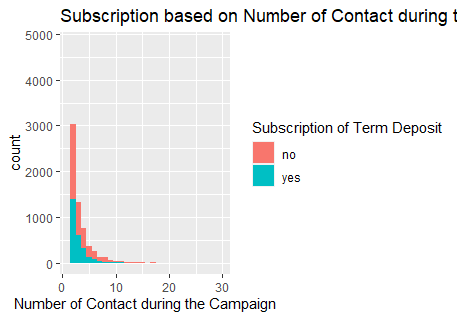
Fig4: Balance Histogram

Clients that subscribe to term deposits have lower loan balances.



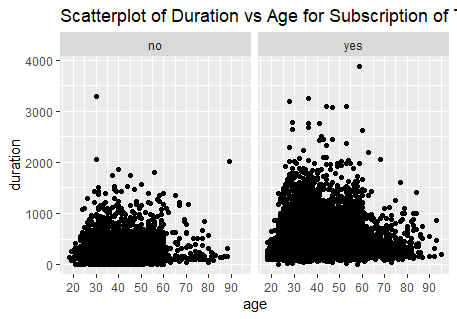
### Fig5: Education vs. Subscription

Having higher education is seen to contribute to higher subscriptions of term deposits. Most clients who subscribe are from 'secondary' and 'tertiary' education levels. Tertiary educated clients have a higher rate of subscription (15%) from total clients called.



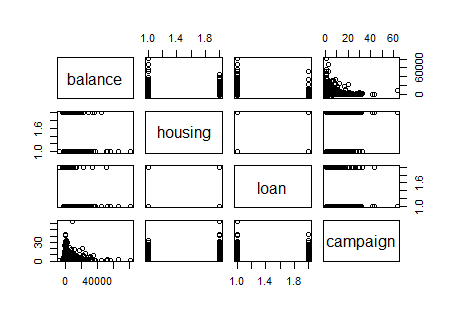
### Fig6: Subscription-based on Number of Contacts during Campaign

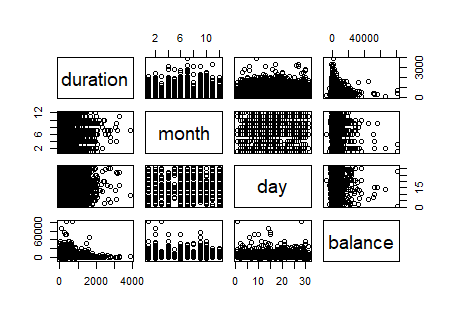
It can be observed from the bar chart that there will be no subscription beyond seven contacts during the campaign. A future campaign could improve resource utilization by setting limits to contacts during a campaign. Future campaigns can focus on the first three contacts as they will have a higher subscription rate.



### Fig7: Scatterplot of Duration by Age

Fewer clients after the age of 60. The duration of the call looks similar.

Fig8: Scatterplot Matrix

Fig9: Scatterplot Matrix

Due to a large number of attributes (17 total), eight were chosen for correlation. No clear correlation pattern can be observed as most features are categorical.

**Preprocessing**

In the dataset, few variables are containing categorical variables. They are converted into factors. Namely,

1. Deposit: 0 – No, 1 – Yes
2. Job
3. Marital: married – 0, single – 1, divorced – 2
4. Education
5. Default - 0 – No, 1 – Yes
6. Housing - 0 – No, 1 – Yes
7. Loan - 0 – No, 1 – Yes
8. Month

**Processing**

The data needs to be split into training and testing. I am using the sample. Split () the data is divided into training 70% and testing 30%, i.e., 7813 observations in training and 3349 in testing.

**Post-processing**

The output variable of the data is categorical. So, implementing classification algorithms and standardizing the data.

**Model Implementation:**

1. Random Forest: Random forest algorithm is a supervised classification and regression algorithm. This algorithm randomly creates a forest with several trees. The more trees in the forest, the more robust the forest looks. Similarly, in the random forest classifier, the higher the number of trees in the forest, the greater is the accuracy of the results.

Random forest builds multiple decision trees (called the forest) and

glues them together to get a more accurate and stable prediction.

The forest it builds is a collection of Decision Trees trained with the

bagging method.

The accuracy of the model by setting the number of trees to 500 is 0.8615(86.15%).

Sensitivity : 0.9061

Specificity : 0.8207

Pos Pred Value : 0.8198

Neg Pred Value : 0.9066

Balanced Accuracy : 0.8634

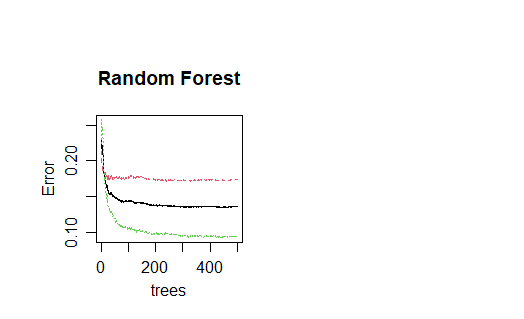
****

Fig 10: Random Forest Graph

The Redline represents the MCR of class not deposited. The green line represents the MCR of class deposited, and the black line represents overall MCR or OOB error.

By Mean Decreasing Gini, the essential variables in predicting the output variable are:

1. Duration
2. Month
3. Balance
4. Age
5. Day

By Mean Decreasing Accuracy, the Important variables in the predicting output variable are:

1. Duration
2. Month
3. Contact
4. Day
5. Poutcome

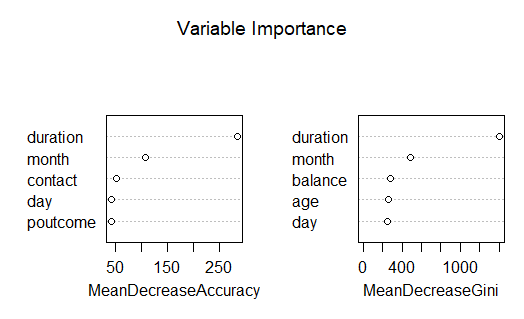
****

Fig 11: Important variables for predicting the output

Out of 3349 testing observations, 335 variables are missed classified.

1. SVM(Support Vector Machine): Support vector machines (SVM) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. It is mostly used in classification problems. In this algorithm, each data item is plotted as a point in n-dimensional space (where n is the number of features), with the value of each feature being the value of a particular coordinate. Then, classification is performed by finding the hyper-plane that best differentiates the two classes.

The model's accuracy by setting kernel as linear and type as C-classification is 0.8274(82.74%).

Out of 3349 testing observations, 578 variables are missed classified.

Sensitivity : 0.8399

Specificity : 0.8161

Pos Pred Value : 0.8045

Neg Pred Value : 0.8499

Balanced Accuracy : 0.8280

1. Logistic Regression: Logistic regression is a predictive modeling algorithm used when the Y variable is binary categorical. It can take only two values, like 1 or 0. The goal is to determine a mathematical equation that can predict the probability of event 1. It can be used to predict the Y when only the X’s are known.

Logistic regression is also called

binary classification problems. Logistic regression is a classic predictive modeling technique and remains a popular choice for modeling binary categorical variables.

The accuracy of the model by setting the family as binomial is 0.9037(90.37%).

Out of 3349 testing observations, 151 variables are missed classified.

Precision: 0.5913163

Recall: 0.5475431

FScore: 0.5685885

**Conclusions**

According to the analysis, a target customer profile can be established. The most responsive customers possess these features:

* Feature 1: age < 30 or age > 60
* Feature 2: students or retired people
* Feature 3: a balance of more than 5000 euros

By applying classification model were successfully built. The bank will predict a customer's response to its telemarketing campaign before calling this customer. In this way, the bank can allocate more marketing efforts to the clients who are classified as highly likely to accept term deposits and call less to those who are unlikely to make term deposits.

In addition, predicting duration before calling and adjusting the marketing plan benefit both the bank and its clients. On the one hand, it will increase the efficiency of the bank's telemarketing campaign, saving time and effort. On the other hand, it prevents some clients from receiving unwanted advertisements, raising customer satisfaction. With the aid of logistic and ridge regression models, the bank can enter a virtuous cycle of effective marketing, more investments, and happier customers.

**References**

1. <https://marathon.csee.usf.edu/~sarkar/IEEEformat.html#:~:text=22.5%20cm).,1%2F4%20inch%20(approx>.
2. <https://www.coep.org.in/page_assets/491/IEEE_Template_4.pdf>
3. <https://rpubs.com/mc_chiwaye/627326>
4. <https://rpubs.com/shienlong/wqd7004_RRookie>
5. <https://www.datacamp.com/community/tutorials/support-vector-machines-r>
6. <https://github.com/krishna7189/Rcodeeasy/blob/master/SUPPORT%20VECTOR%20MACHINE%20(SVM)%20-%20Detailed%20Example%20on%20Classification%20in%20R>
7. <https://www.geeksforgeeks.org/classifying-data-using-support-vector-machinessvms-in-r/>
8. <https://app.grammarly.com/ddocs/1134167267>