Northeastern University

Khoury College of Computer Sciences,
Northeastern University,
Boston, MA
Fall 2022

Bank Marketing Prediction for Future Campaigns

Group 7:

Arya Dhorajiya

Rijul Saini

Sumit Hawal

Spandan Maaheshwari

1. <u>Summary:</u>

1.1. Overview

A Portuguese Bank wants to market one of its products, Term deposit. These are generally short-term with maturities ranging anywhere from a month to a few years. Term deposits are an extremely safe investment and are therefore very appealing to conservative, low-risk investors. Instead of mass marketing, the bank has opted to be more proactive in identifying potential buyers and contacting the customer directly.

1.2. Goals

Often, more than one contact to the same client was required, in order to assess if the product (bank term deposit) would subscribe or not. Our goal for this project is to perform post campaign analytics to identify the potential subscribers of the term deposit product for future campaigns. The dataset we have used has 16 attributes and an output variable which tells us whether a client subscribed for the term deposit. We intend to initially use exploratory data analysis to visualize the relationships between different input variables and the output, and from these insights select some variables to run linear and non-linear models on. Finally, using accuracy, specificity and sensitivity to select the best performing model.

1.3. Data Description

The bank marketing dataset was taken from the UCI Machine Learning Repository. Dataset includes 45211 instances with 17 variables of categorical, binary or numeric type. The attributes of the dataset are as follows:

1. *Age:* Age of client (numeric)
2. *Job:* Type of job (categorical)
3. *Marital:* Marital status (categorical)
4. *Education:* Level of education (categorical)
5. *Default:* Has credit in default? (binary)
6. *Balance:* Average yearly balance, in Euros (numeric)
7. *Housing:* Has a housing loan? (binary)
8. *Loan:* Has a personal loan? (binary)
9. *Contact:* Contact communication type (categorical)
10. *Day:* Last contact day of the month (numeric)
11. *Month:* Last contact month of year (categorical)
12. *Duration:* Last contact duration, in seconds (numeric)
13. *Campaign:* Number of contacts performed during this campaign and for this client, includes last contact (numeric)
14. *Pdays:* Number of days that passed by after the client was last contacted from a previous campaign (numeric)
15. *Previous:* Number of contacts performed before this campaign and for this client (numeric)
16. *Poutcome:* Outcome of the previous marketing campaign (categorical)
17. *Y:* Has the client subscribed to a term deposit? (binary)

2. Methods:

2.1. Programming Languages: R, Python

2.2. Libraries used: ggplot2, dplyr, tidyr, readr, ggthemes, tidyverse, gridExtra, numpy, pandas, sklearn, matplotlib, seaborn, XGBoost, yellowbrick, math

2.3. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is the process of finding important relationships in our dataset. These relationships can be helpful in determining what predictors can be useful in predicting our output variable successfully. In case of continuous variables, we look for a function that can explain the output variable. In case of categorical variables, we look for if the output is affected by different categories of the input categorical variable.

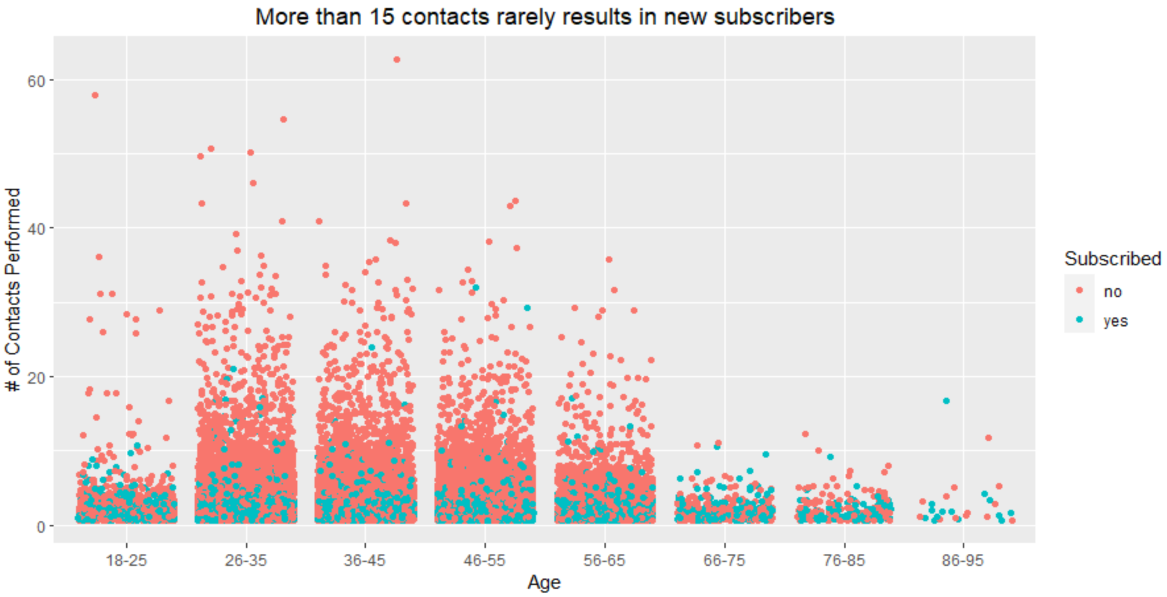


Fig 1. More contacts do not result in more subscriptions

Clearly there is room for improvement in the bank's outreach as most clients who were contacted more than 15 times did not subscribe to term deposit, as seen in *Fig 1*. Upon further inspection of the ones who subscribed in *Fig 2*, we see that although younger clients tend to subscribe more, fewer calls are needed for older clients to get subscribed.

We have identified a need here that shows that most of the bank's contacts are being wasted and not bearing any new customers. Narrowing down the target demographic using our project will result in more subscriptions and fewer wasted resources.

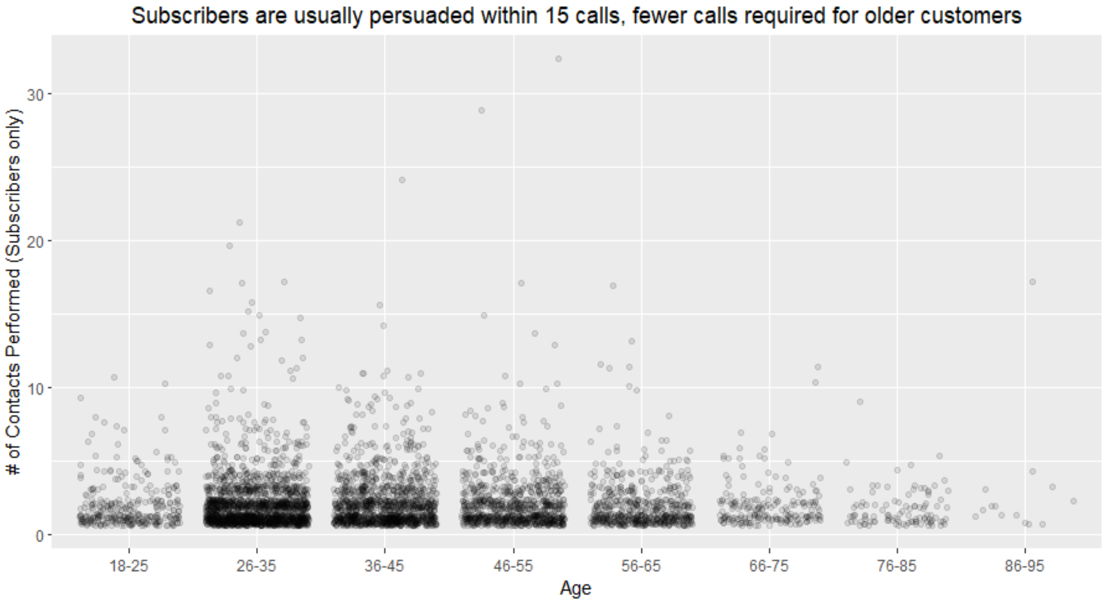


Fig 2. Younger clients tend to subscribe more

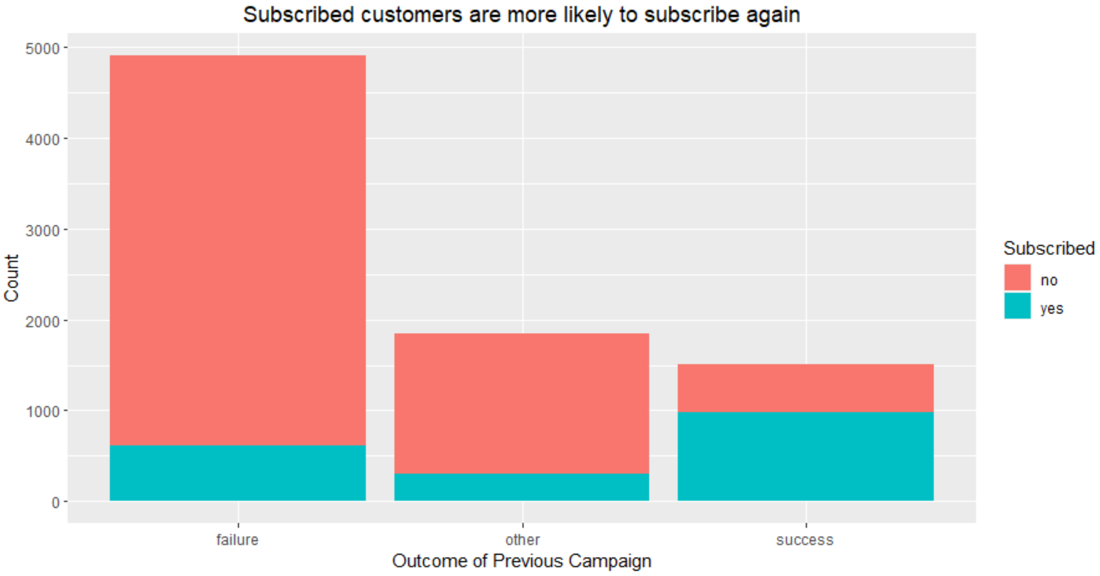


Fig 3. Outcome of previous campaign is important in predicting future outcome

Fig 3 and *Fig 4* show the trends of poutcome (outcome of previous campaign) and balance (average yearly balance) with subscriptions. We can see in *Fig 3* that customers who subscribed in previous campaign are more likely to resubscribe again. *Fig 4* shows that the younger clientele with average yearly balance of about 30,000 Euros or less are more likely to subscribe, while the trends changes with increase in age of client. With age, the average yearly balance of clients who subscribe becomes more than those who don't subscribe. This shows that older retired people with more disposable income at hand are likely to subscribe, while younger earning clients who want a safe investment with low risk and who don't have a lot of disposable income are willing to subscribe.

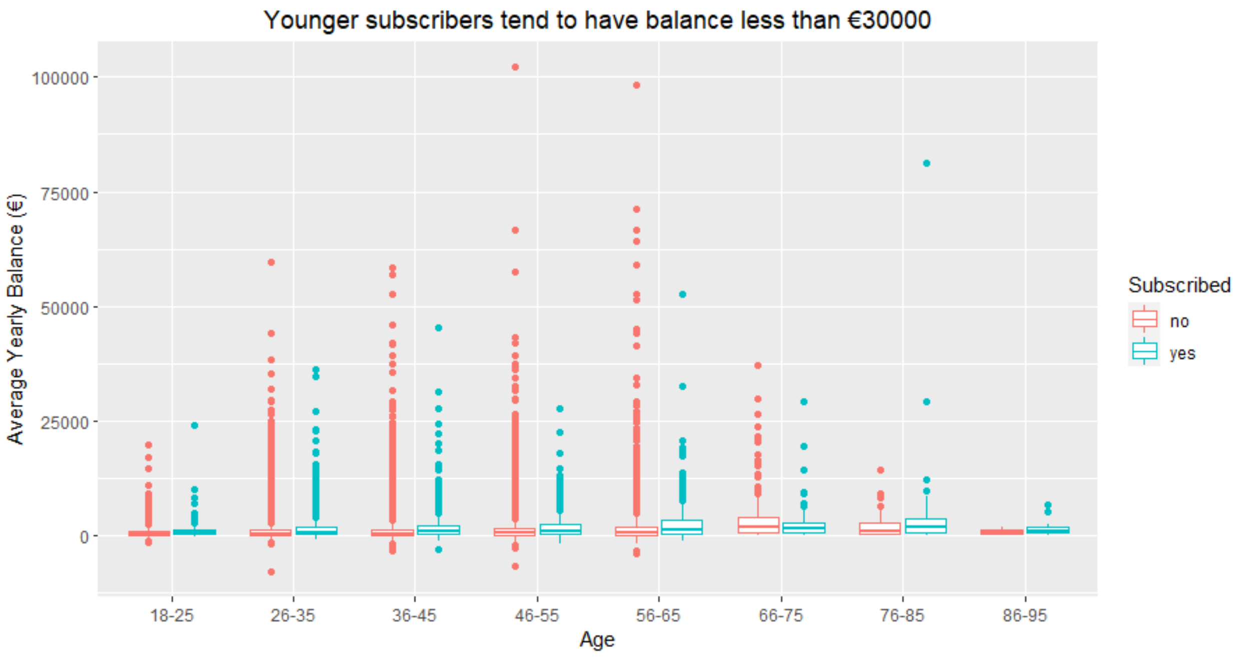


Fig 4. Trends in average yearly balance change with age of client

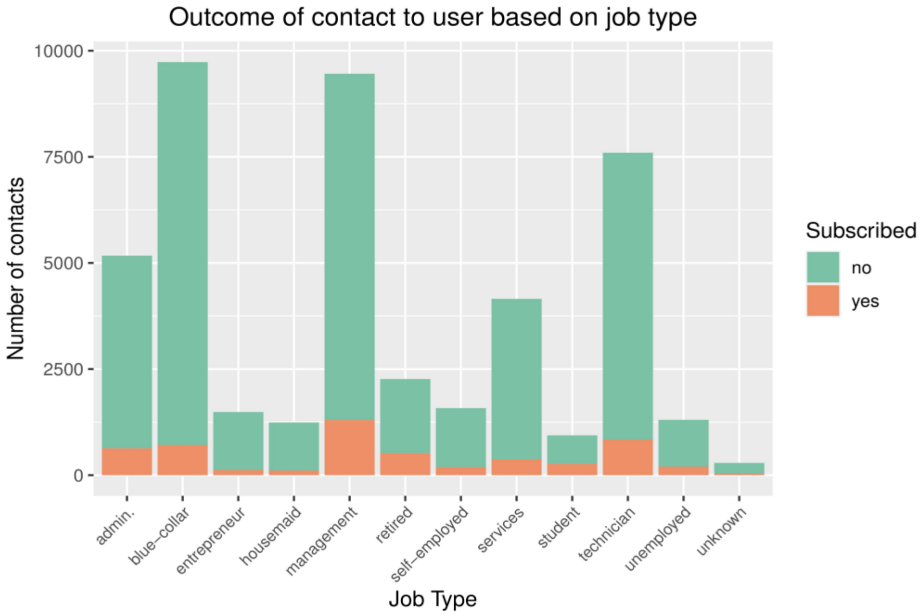


Fig 5. Outcome of contact to user based on job type

In the above stacked bar chart, we can see that students, retired and people with management type job say ‘yes’ to term subscription proportionally more than other categories of jobs. So here, the banks can focus more on targeting these people with such backgrounds to improve their success rate. While at the same time, targeting less those of categories like blue-collar, admin, services to reduce their rejection rate.

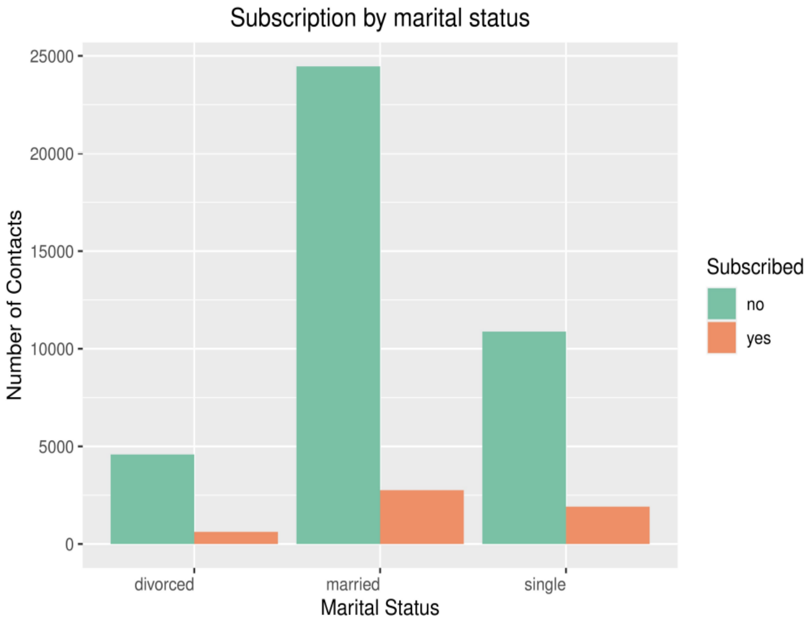


Fig 6. Subscription by Marital Status

In *Fig 6*, we can see the ‘married’ and ‘divorced’ people tend to subscribe less to term deposits as compared to ‘singles’. Thus, the banks can focus on finding as to why these people tend to reject their term deposit subscriptions.

Fig 7 is a stacked bar graph with proportions of the number of people who subscribe and did not subscribe to the term plan for every month of the year. We can see that in the months September, October, March, April had more subscription to term deposit than rejection. So, increasing the amount of contacts during these months could improve the rate of people saying ‘yes’ to term deposits.

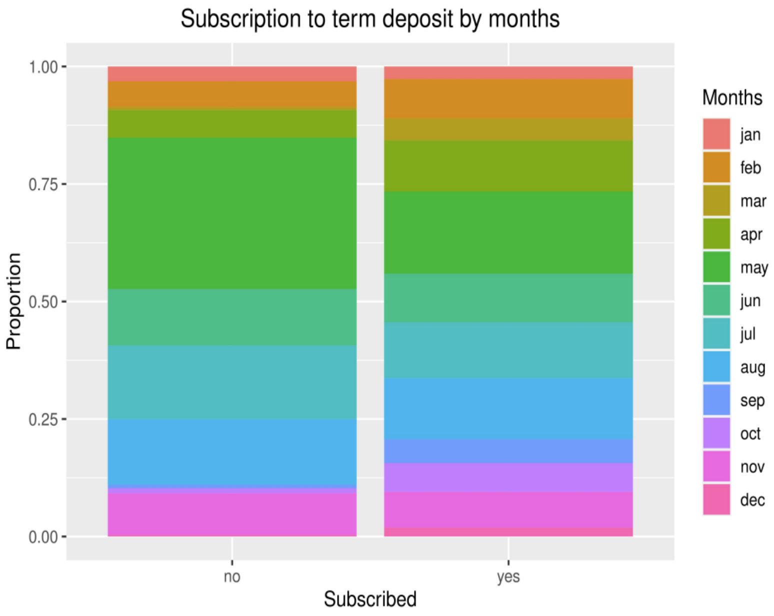


Fig 7. Subscription to term deposit by months

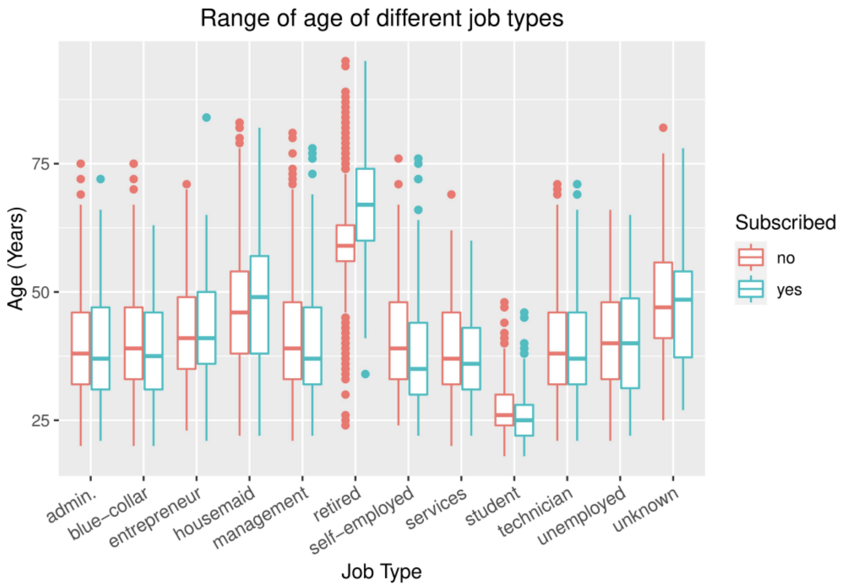


Fig 8, shows boxplots of age of people in different job types and whether they subscribe to term deposit or not. Here, we can see that each job category approximately equally rejects('no') and accepts('yes') to the term deposit of the bank except the ones in management and self-employed. The management job-type tend to say yes more to term deposit then reject it, while those who are 'self-employed' tend to reject the term deposit more than accept it.

Fig 8. Range of age of different job types

Fig 9 is a density plot showing duration of calls by whether they say 'yes'/'no', to term deposit. We can see that people who subscribe to term deposits tend to stay on call for a longer period of time. This can be due to the fact that they may be interested to learn more about the deposit and that is why they could stay on call for a longer period of time. While the people who didn't want to subscribe would quickly end the call.

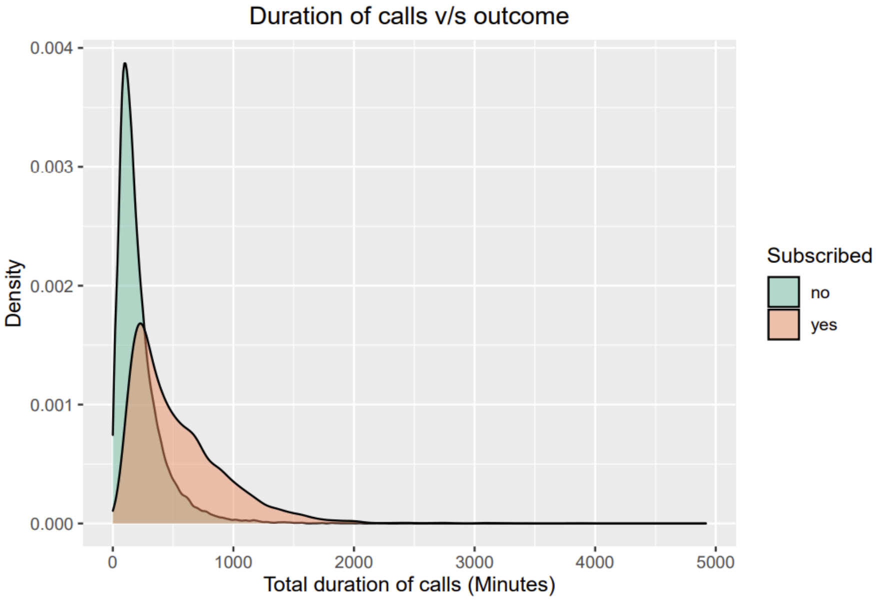
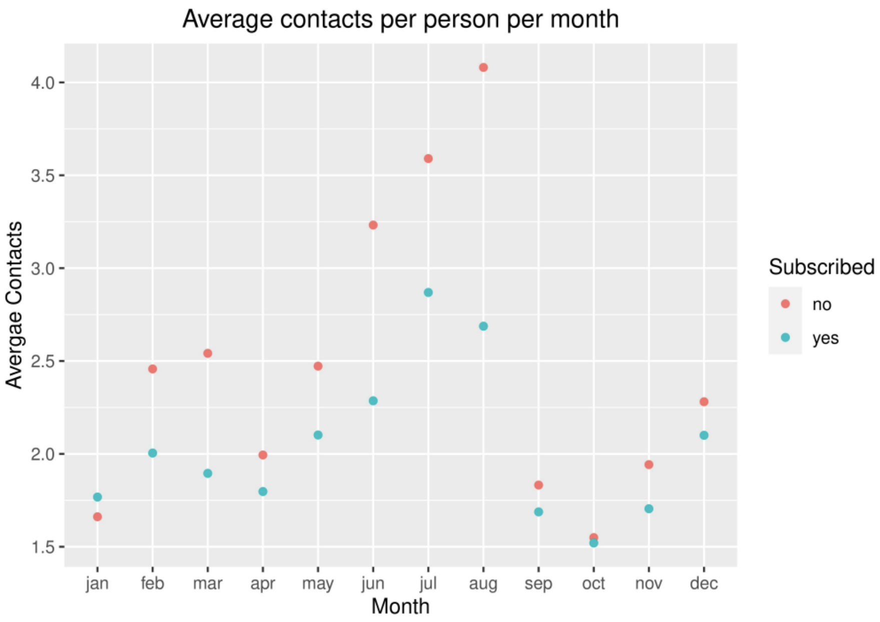


Fig 9. Duration of calls v/s outcome



The scatterplot shown by *Fig 10*, on average how many times a person was contacted and if they decide to subscribe to term deposit or not. We can see that people who subscribe to term deposits usually are contacted less than 2.5 times. Therefore, the bank can optimize the number of contacts they make to convince a person to subscribe to their term deposit. As they know if that person wants to subscribe, they will usually do before contacting them the 3rd time.

Fig 10. Average contacts per person per month

Fig 11 shows the scatter plot of average duration of calls by months and the subscription rate of the calls. We can do the same thing as we saw in Fig **5**, that on average if people subscribe, they tend to stay on call for a longer time. Here the average duration of people who subscribed is greater than 400 minutes.

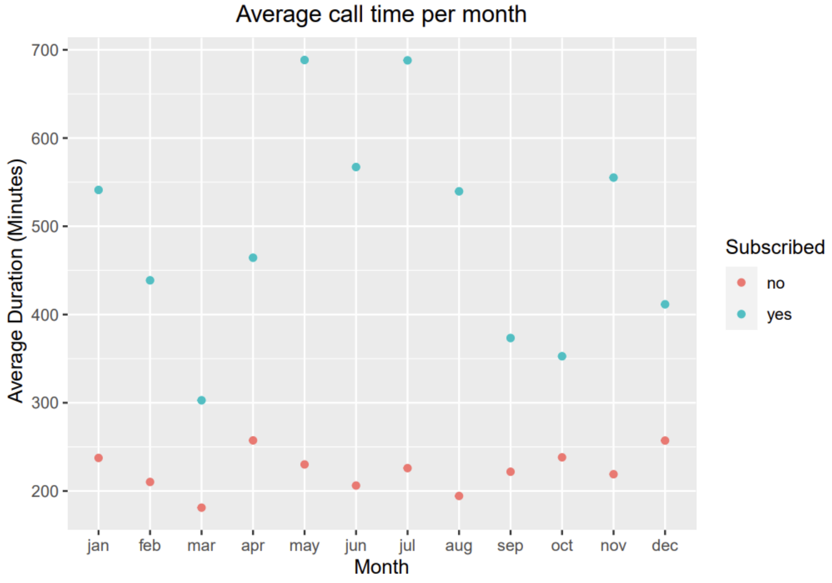


Fig 11. Average call time per month

2.4. Data Preprocessing:

Any type of processing done on raw data to get it ready for another data processing operation is referred to as data preprocessing, which is a part of data preparation. It is a crucial first step in the data analysis process. Inferences can now be made against our variables to learn how the predictors will affect the output variable.

Our dataset is already tidy, but to construct a good predictive model on the given data, some data pre-processing steps need to be performed as we identified a class imbalance in our dataset through our Exploratory Data Analysis.

We performed the data pre-processing steps as mentioned below :

- **Removed Duplicate Data**: As the given dataset didn't have any NA values, we further explored and saw that it had 12 rows, which were duplicates of other data points, so we dropped those to reduce the bais of one category in data.

- **Downsampled the data**: To downsample data from an 88-12% split to a 50-50% split, we reduced the number of samples in the larger class to match the number of samples in the smaller class. This can be done by randomly selecting a subset of the larger class. The resulting dataset will have an equal number of samples in each class, with a 50-50% split.

- **Converted label into binary classes**: The label required numeric encoding to fit a model, so we used label encoding to convert the class 'yes' and 'no' into 0 and 1. This encoding will help produce a better model for future inference of results.

- **Dropped 'default' and 'contact' columns:** The 'default' and 'contact' columns had to be dropped because 'default' contained the data if the customers have unpaid debt or not, which didn't give any inference while performing Exploratory Data Analysis, being similar to 'contact', which was classified into Cellular and Telephone, didn't help in model prediction.

- **One Hot Encoded the Categorical Variables:** We had to one hot encode categorical variables, in order to create a new binary column for each unique category in the variable. The resulting columns will have a value of 1 for the rows that belong to the corresponding category, and 0 for all other rows. This allows the categorical data to be used in Classification machine learning algorithms that expect numeric input.

- **Scaling the Quantitative Variables using Min-Max Scalar**: Data Scaling was performed for quantitative variables using a min-max scaler, to transform each variable to have a minimum value of 0 and a maximum value of 1. It is done by subtracting the minimum value from each value in the variable and then dividing by the range of the variable (i.e., the difference between the maximum and minimum values). This ensures that all variables are on the same scale, which is very useful for Classification machine learning algorithms.

- **Utilizing Cross-Validation on XGBoost and Random Forest models**: Cross-validation is a technique that can be used to evaluate the performance of XGBoost and random forest models. It involves splitting the training data into multiple folds, training the model on each fold, and then evaluating the model on the remaining fold. This process is repeated multiple times, and the results are averaged to obtain an estimate of the model's performance on unseen data. Cross-validation helps us to prevent overfitting by ensuring that the model has not memorized the training data, and it can also provide a more accurate estimate of the model's generalization error.

2.5. Data Modelling:

As our dataset is complex, we are gonna use a combination of Linear and non-Linear models to identify complex relations and get inferences, which is not possible through a simple Linear model.

A. Logistic Regression :

Logistic regression is a type of regression analysis that is used to model the probability of a binary outcome, such as the likelihood that a customer will churn or that a patient will have a certain disease. In logistic regression, the dependent variable (i.e. the variable that we are trying to predict) is binary, while the independent variables (i.e. the variables that we use to predict the outcome) can be continuous or categorical.

The equation for logistic regression is as follows:

$$\hat{y} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}}$$

In this equation, yˆ is the predicted probability that the binary outcome will occur, x1, x2, ..., xn are the independent variables, and β0, β1, ..., βn are the coefficients that are learned during the model training process. The coefficients represent the effect that each independent variable has on the predicted outcome. For example, a positive coefficient for an independent variable would indicate that the variable has a positive effect on the outcome, while a negative coefficient would indicate a negative effect.

The loss function used by logistic regression is the cross-entropy loss, which is the negative log-likelihood of the true labels given the predicted probabilities. This loss function is commonly used for classification tasks because it measures the difference between the predicted probabilities and the true labels in a way that is differentiable and can be optimized using gradient descent.

Some advantages of logistic regression include its simplicity, interpretability, and efficiency. It is easy to train and requires little hyperparameter tuning, which makes it a good choice for many classification tasks. The coefficients of the model can also provide insights into the relationship between the independent variables and the predicted outcome, which can be useful for understanding the data and making decisions based on the model's predictions. Overall, logistic regression is a popular and widely used method for classification tasks.

B. Random Forest:

Random forest is an ensemble learning method that is used for both regression and classification tasks. It is a type of decision tree algorithm that constructs a large number of decision trees during training and then combines the predictions of each tree to make a final prediction.

The equation for a random forest model is as follows:

$$\hat{y} = \frac{1}{T} \sum_{i=1}^{T} \hat{y}_i$$

In this equation, yˆ is the final predicted value, T is the number of decision trees in the forest, and y^i is the predicted value for the i^{th} decision tree. To make a prediction, the random forest model first constructs T decision trees using a random subset of the training data. Each tree is trained to make predictions based on a different subset of the features in the data. Then, the model combines the predictions of each tree by taking the average (for regression tasks) or by taking the majority vote (for classification tasks).

A decision tree is a type of machine learning algorithm that is used for both regression and classification tasks. It works by dividing the input data into regions, called leaf nodes, that have similar values for the target variable. The tree is constructed by making a series of decision splits, based on the values of the input features, that divide the data into increasingly pure leaf nodes. The final prediction for a given example is the average or majority vote of the training examples in the leaf node that the example falls into.

For classification tasks, Loss function used by random forest for regression tasks is cross-entropy loss, which is the negative log- likelihood of the true labels given the predicted probabilities. The loss function is used to evaluate the model's performance and to guide the training process by minimizing the loss.

Some advantages of using random forest include its ability to capture complex non-linear relationships, its ability to handle a large number of features, and its robustness to overfitting. Random forest is an ensemble method, which means that it constructs a large number of decision trees during training and then combines their predictions to make a final prediction. This allows the model to capture complex non-linear relationships in the data, which can improve the model's accuracy and generalization ability. Additionally, random forest can handle a large number of features, which is often the case in real-world applications where high-dimensional data is common. Finally, a random forest is less prone to overfitting than a single decision tree, because the randomness in the training process helps to prevent the trees from becoming too similar.

Overall, random forest is a powerful and widely used method for both regression and classification tasks.

C. XGBoost

XGBoost (eXtreme Gradient Boosting) is a gradient boosting algorithm that is used for both regression and classification tasks. It is an implementation of gradient boosting that is specifically designed to be highly efficient and scalable.

The equation for an XGBoost model is as follows:

$$\hat{y} = \sum_{i=1}^{T} f_i(x)$$

In this equation, yˆ is the predicted value, T is the number of boosting rounds, fi is the ith boosting round, and x is the input data. XGBoost works by iteratively fitting weak learners (e.g., decision trees) to the residuals of the previous boosting round. The residuals are the differences between the true values and the predicted values of the previous round. This process continues for T rounds, and the final prediction is the sum of the predictions from each round.

The loss function used by XGBoost is the sum of the squared residuals, which is commonly used for regression tasks. For classification tasks, XGBoost uses a variant of the logistic loss function, which is the negative log-likelihood of the true labels given the predicted probabilities. The loss function is used to evaluate the model’s performance and to guide the training process by minimizing the loss.

Some advantages of using XGBoost include its efficiency, scalability, and predictive power. XGBoost is designed to be highly efficient, both in terms of training time and memory usage, which makes it well suited for large-scale problems and real-time applications. It is also highly scalable, which means that it can handle very large datasets and a large number of features. In terms of predictive power, XGBoost has been shown to perform well on a variety of regression and classification tasks, and is often used in competitions and real-world applications where high prediction accuracy is important. Additionally, XGBoost provides a number of features for tuning and regularization, which can help to prevent overfitting and improve the model's generalization ability.

3. Results:

The following table represent the integrated results of all the models for the subscription of term deposit:

Testing				
Model	**Accuracy**	**Sensitivity**	**Specificity**	**Balanced**
Logistic Regression	0.8030	0.7788	0.8266	0.8027
Random Forest (CV)	0.8586	0.8557	0.8615	0.8586
eXtreme Gradient Boosted Trees	0.8670	0.8703	0.8628	0.8666

Table 1. Classification results comparison between all models

The relevant indicators are as follows:

- **Classification Rate/Accuracy**: proportion of correctly predicting the subscription of the term deposit
- **Specificity**: how many customers who will not actually subscribe term deposit are successfully identified
- **Sensitivity**: how many customers who subscribe term deposit are successfully identified
- **Balanced accuracy**: average of sensitivity and specificity to account for fair prediction whether a customer is a potential subscriber or not.

The AUC value is generally from 0.5 to 1, 0.5 means completely random classification, 1 means perfect classification. The larger the value, the better the classifier performance. As shown in *Fig 12* below the ROC curves of the XGBoost model in the test set have been calculated to obtain AUC values of 0.92, and the model classification effect is satisfactory.

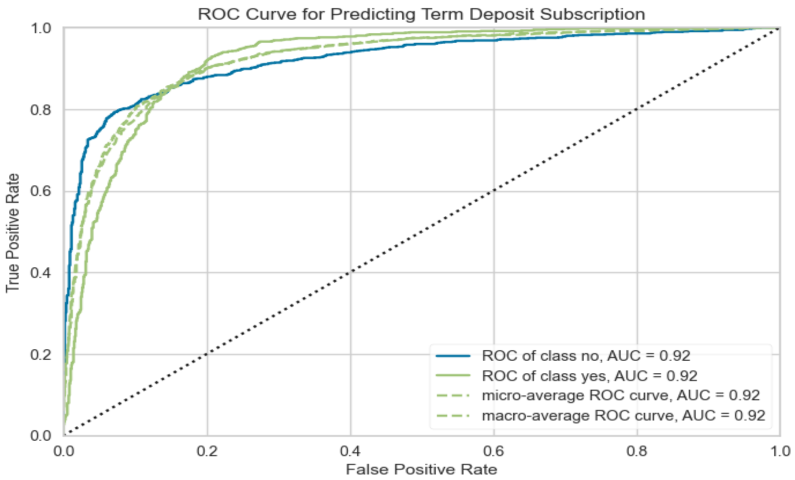


Fig 12. ROC curve of XGBoost classifier

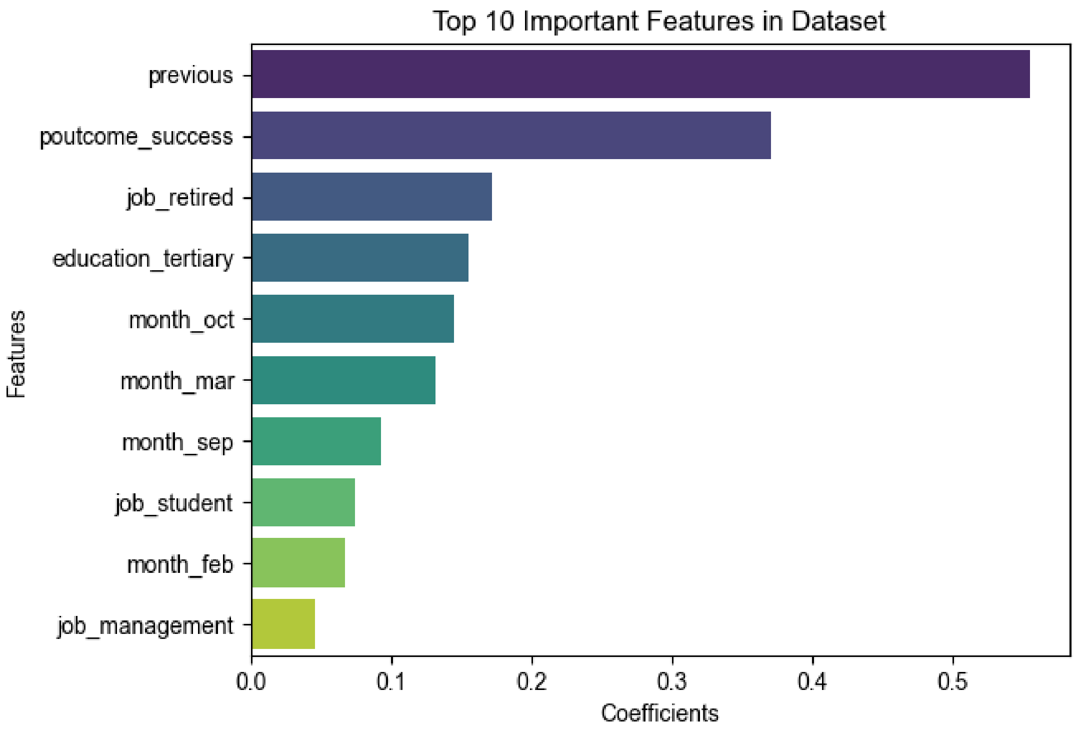
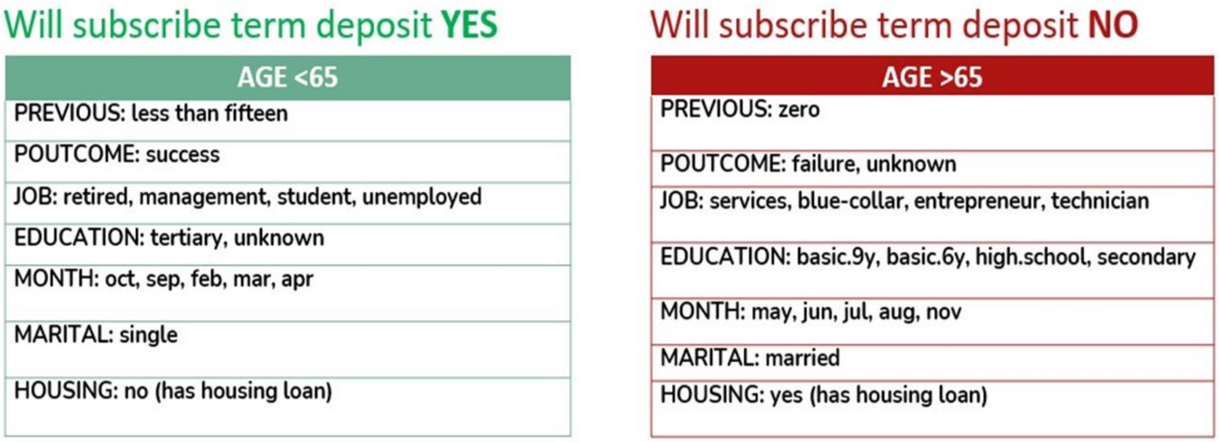


Fig 13. XGBoost feature importance

The most important predictors depicted in *Fig 13* are **previous, poutcome, job, education and month**.



Will subscribe term deposit **YES**

AGE <65
PREVIOUS: less than fifteen
POUTCOME: success
JOB: retired, management, student, unemployed
EDUCATION: tertiary, unknown
MONTH: oct, sep, feb, mar, apr
MARITAL: single
HOUSING: no (has housing loan)

Will subscribe term deposit **NO**

AGE >65
PREVIOUS: zero
POUTCOME: failure, unknown
JOB: services, blue-collar, entrepreneur, technician
EDUCATION: basic.9y, basic.6y, high.school, secondary
MONTH: may, jun, jul, aug, nov
MARITAL: married
HOUSING: yes (has housing loan)

Table 2. Indicators for identification of potential term deposit subscribers

Additionally, using all our analysis, a table has been created for marketing companies which contains the best set of indicators for singling out potential term deposit subscribers. A snapshot of the tables created for banking institutions have been displayed above in *Table 2.*

4. Discussion:

The aim was to predict the set of factors/attributes that play a key role in determining the potential subscribers of the term deposit product for future campaigns. By using feature importance with gradient boosted trees, we were able to select the best set of predictors that play a key role in determining the customer as a potential subscriber.

After performing all the models, the gradient boosted trees turned out to be the champion model for the data with a sensitivity of **87%**, specificity of **86%**, and balanced accuracy of **86.7%**. The champion model provides a highly efficient implementation of the stochastic gradient boosting algorithm and access to a suite of model hyperparameters designed to effectively identify whether a customer will subscribe to term deposit or not.

This not only reduces the intrusiveness brought to many non-target customers, but also helps the bank to effectively improve marketing efficiency and reduce the cost of resources such as manpower, so that the bank can carry out more targeted marketing activities. The empirical research on the marketing data set of Portuguese banks show that the prediction model proposed in this report has a relatively satisfactory learning ability and generalization ability, which can provide applications for marketing companies and banking institutions to achieve precise marketing.

We can further examine how it influences our subscription rate using factors like employee variation rate, consumer price index, and consumer confidence index. In order to get more accurate findings from the test data, we could use additional techniques like neural networks and SVM for future advancements, which requires more balanced dataset with greater number of observations.

5. Statement of Contribution:

After weighing our options for dataset selection, we chose to use Bank Marketing data suggested by Sumit. Along with ideation, Sumit and Rijul worked on the EDA part, visualizing different variable relationships and helping to identify significant features for the modeling phase.

Following EDA, the provided dataset had a class imbalance, therefore Arya down-sampled the data to only include the categories with the bulk of data points, randomly resampled the data, and preprocessed the data to get it ready for data modeling. Spandan performed model selection, model evaluation, performance metric comparison, projected ROC curve to understand the performance of the Boost model, and generated a bar chart representing the Top 10 Important Features in the dataset after preprocessing the data.

6. References:

- Data Source: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing#
- Jianguo Che, Sai Zhao, Yongfan Li, Kai Li (2020). Bank Telemarketing Forecasting Model Based on t-SNE-SVM. *Journal of Service Science and Management, 2020, 13, 435-448*.

7. Appendix:

Some interesting relationship between different variables in the dataset:

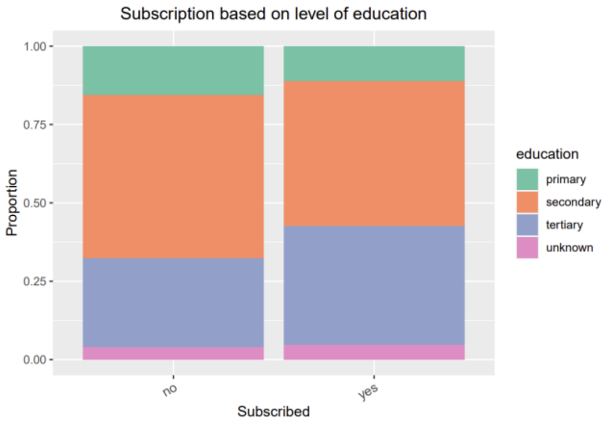


Fig 14. Subscription based on level of education

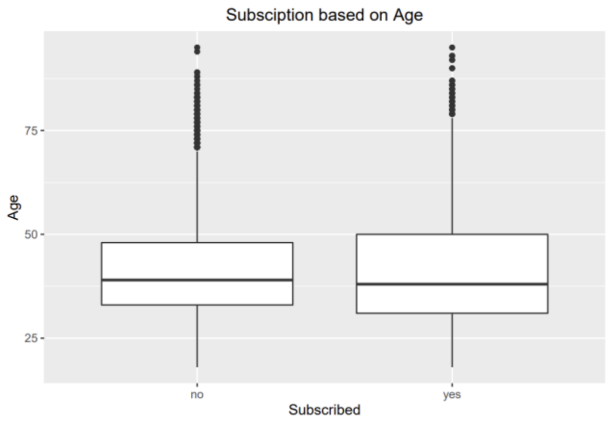


Fig 15. Subscription based on Age

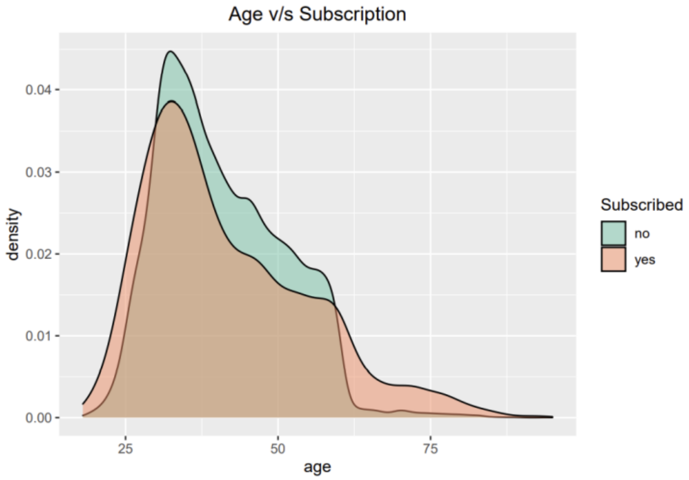


Fig 17. Subscription varying with Age

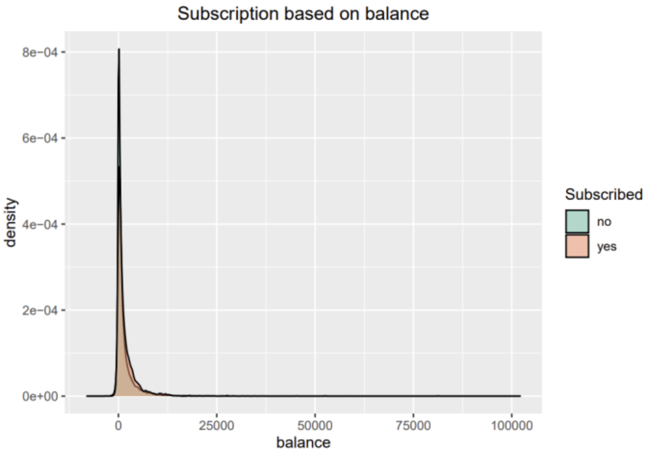


Fig 16. Subscription based on balance

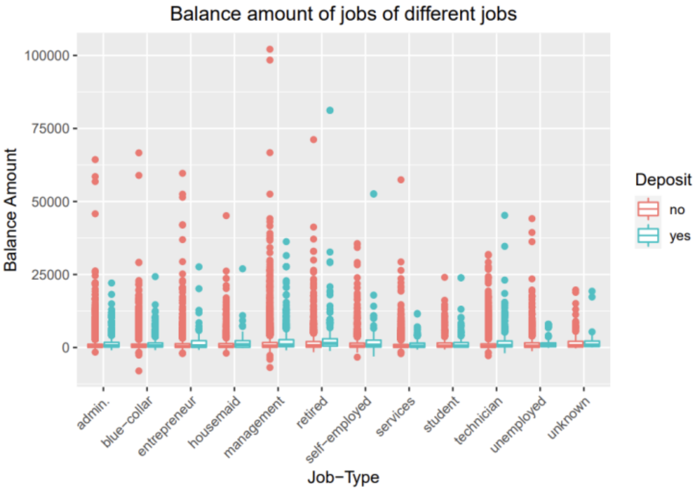


Fig 19. Balance amount of jobs of different jobs

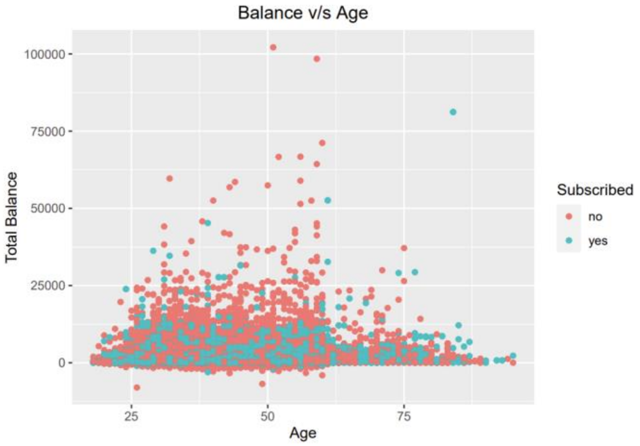


Fig 18. How bank balance differs with age

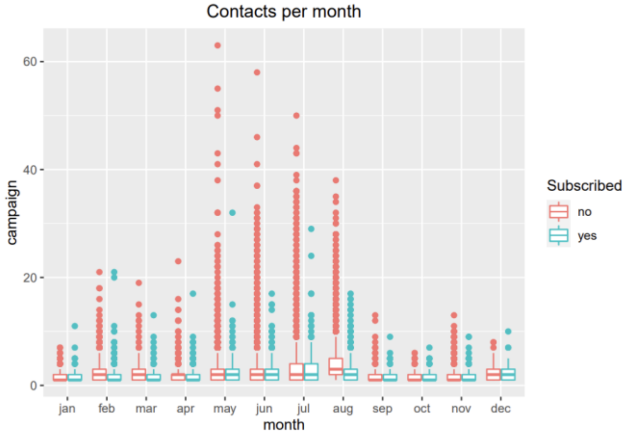


Fig 20. Number of people contacted per month

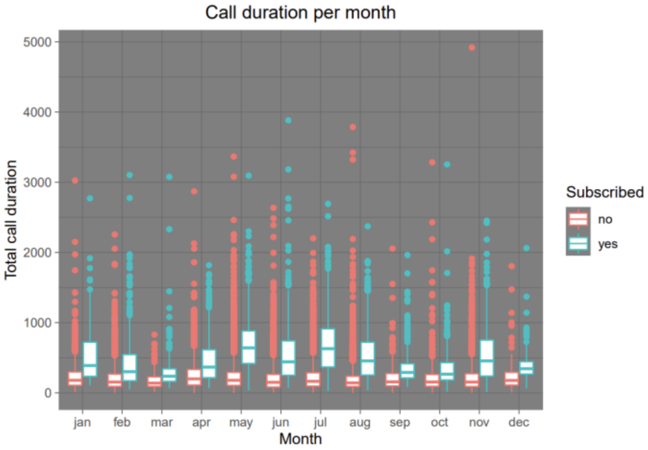


Fig 21. Duration of call monthly (total)