**Analysis of a Bank Marketing Campaign dataset**

A bank is a financial institution that takes deposits and gives out loans. Banks use depositors’ money to give loans. The amount of interest the banks collect on the loans is greater than the amount of interest they pay to customers with savings account or term deposit accounts.

*“A term deposit is a type of deposit account held at a financial institution where the money is locked up for a set period”.* The term deposit is an inexpensive source of funding for a bank. The higher subscription to term loans increases the lending capacity of the bank. Hence the term deposit is crucial for banks.

Banks leverage the information of their existing customers to target and cross-sell other services, to increase the overall engagement with the client. Banks often involve in various promotional activities like advertisements, inhouse branding, direct marketing, events, etc.

The Bank in this case is intending to promote the term deposit service to its existing customers through a telemarketing campaign.

**Business problem**: “To increase the number of subscriptions for a term deposit”.

The bank business problem can be solved with the help of strategies mentioned below:

* Improving marketing strategies to attract more customers to open new term deposits.
* Expand the reach to the target audience.
* Understand the influence of different factors on customer decision making.

The study consists of two main dimensions. First is to build a prediction model, which is appropriate to predict whether a client has subscribed to a term deposit or not, with the given telemarketing campaign data of the bank. Initially, exploratory data analysis has been performed to check for any existing patterns in the data and understand the right questions that can be asked by looking at the data. The study also applies and compares the Decision tree, Adaboost, Neural Network, Logistic Regression, and Random Forest classifiers. The second objective of this study is to enhance campaign effectiveness by determining the key features that influence the decision regarding the subscription to a term deposit by customers.

**The stakeholders** in this case are

**Customer:** The current user of bank services, to whom we are cross-selling the term deposit.

**Employees:** Bank employees who are facilitating the process of communicating and providing term deposits.

**Shareholders and Investors:** A shareholder owns a part of a bank through shares or stocks and has an interest in the performance of the company.

**Government:** Represented by its authorities and companies where it has a participating interest, is an important partner and a customer of the bank.

**Business Goal**: “To increase the subscription of term deposit accounts by enhancing the marketing campaigns to reach the existing customer and potential customers”.

**Data mining Goal**: “Identify the factors influencing the customer decision making for the subscription of term deposit, given the customer details hold by bank and details from previous marketing campaign”.

**Data Description:**

The dataset comprises of nominal (categorical and binary) and numerical attributes. Age, Balance, Day, Duration, Campaign, Pdays, and Previous makes up the numerical attributes. The categorical attributes are Job, Marital, Education, Contact, Month, and Poutcome. Default, Housing Loan, and Output dichotomous attributes. The attribute ‘Poutcome’ indicates the outcome of the previous marketing campaign, for instance, the campaign being unknown, other, failure, or successful. The target label is “deposit”.

The dataset comprises of 11162 rows (profiles) and 17 columns(attributes).

|  |  |  |
| --- | --- | --- |
| **S/N** | **Attribute Name** | **Attribute Type** |
| 1 | Age | Numeric |
| 2 | Job | Categorical |
| 3 | Marital status | Categorical |
| 4 | Education | Categorical |
| 5 | Default | Categorical (Binary) |
| 6 | Balance | Numeric |
| 7 | Housing | Categorical (Binary) |
| 8 | Loan | Categorical (Binary) |
| 9 | Contact | Categorical |
| 10 | Day | Numeric |
| 11 | Month | Categorical |
| 12 | Duration | Numeric |
| 13 | Campaign | Numeric |
| 14 | Pdays | Numeric |
| 15 | Previous | Numeric |
| 16 | Poutcome | Categorical |
| 17 | Deposit | Categorical (Binary) |

**Data understanding and preparation:**

Each row of the dataset represents the profile of an existing customer, while the column represents the demographics of each customer and information related to the current marketing campaign. There were not many missing values in the dataset.

**Exploratory data analysis:**

The target variable in the dataset was the attribute deposit. The goal of the analysis was to predict if a customer will open a term deposit account with the bank using the attributes in the dataset. In addition to this, the analysis was performed to understand the impact of different attributes on the outcome to improve the marketing campaigns.

Exploratory data analysis was performed to understand the essence of the data and decide the right questions to be asked,. EDA helped in taking important decisions about what attributes should be considered for analysis and what can be left out. It also helped us understand the correlation between some of the independent variables and the target variable.

Power BI and Tableau had been used to perform the exploratory data analysis. Box plots, bar charts, scatter plots, heat maps, ribbon charts and bubble charts have been used for the same.

**Chart-1:** **Months of Marketing Activity**

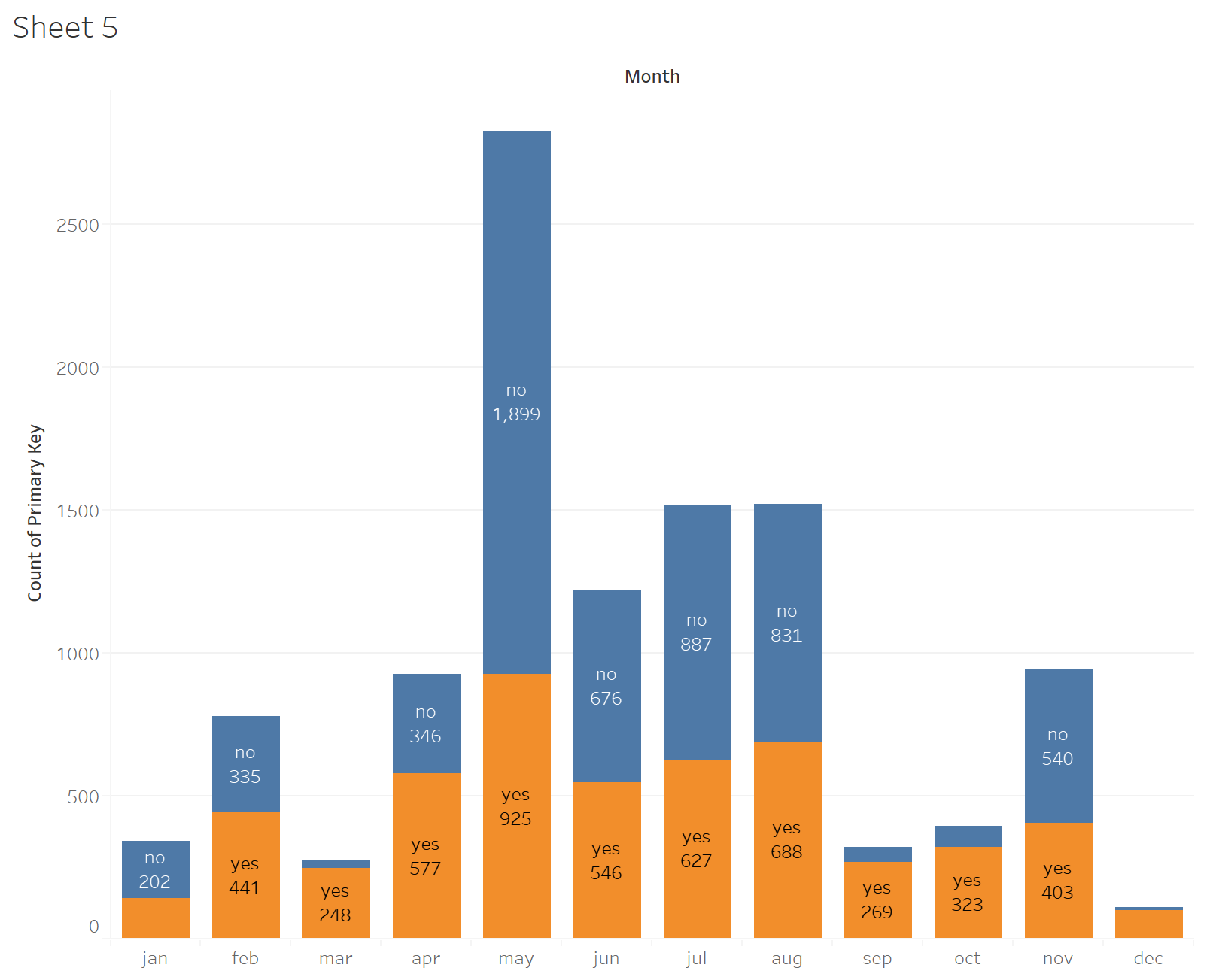


Fig1: Campaign calls and responses by month

**It can be inferred from the chart that** marketing activity was highest in the month of **May.** However, this was also the month when the highest number of potential clients rejected term deposits offers. As seen from the chart, May saw the lowest effective rate totaling to -34.49%. As the chart suggests, it will be wise for the bank to focus the marketing campaigns during the months of **March, September, October, and December.** The marketing activity in the month of December seems very low and the bank should find a way to find ways to reach their customers during December. The reason for this low marketing activity could be the holiday season.

**Chart-2: Campaign calls**

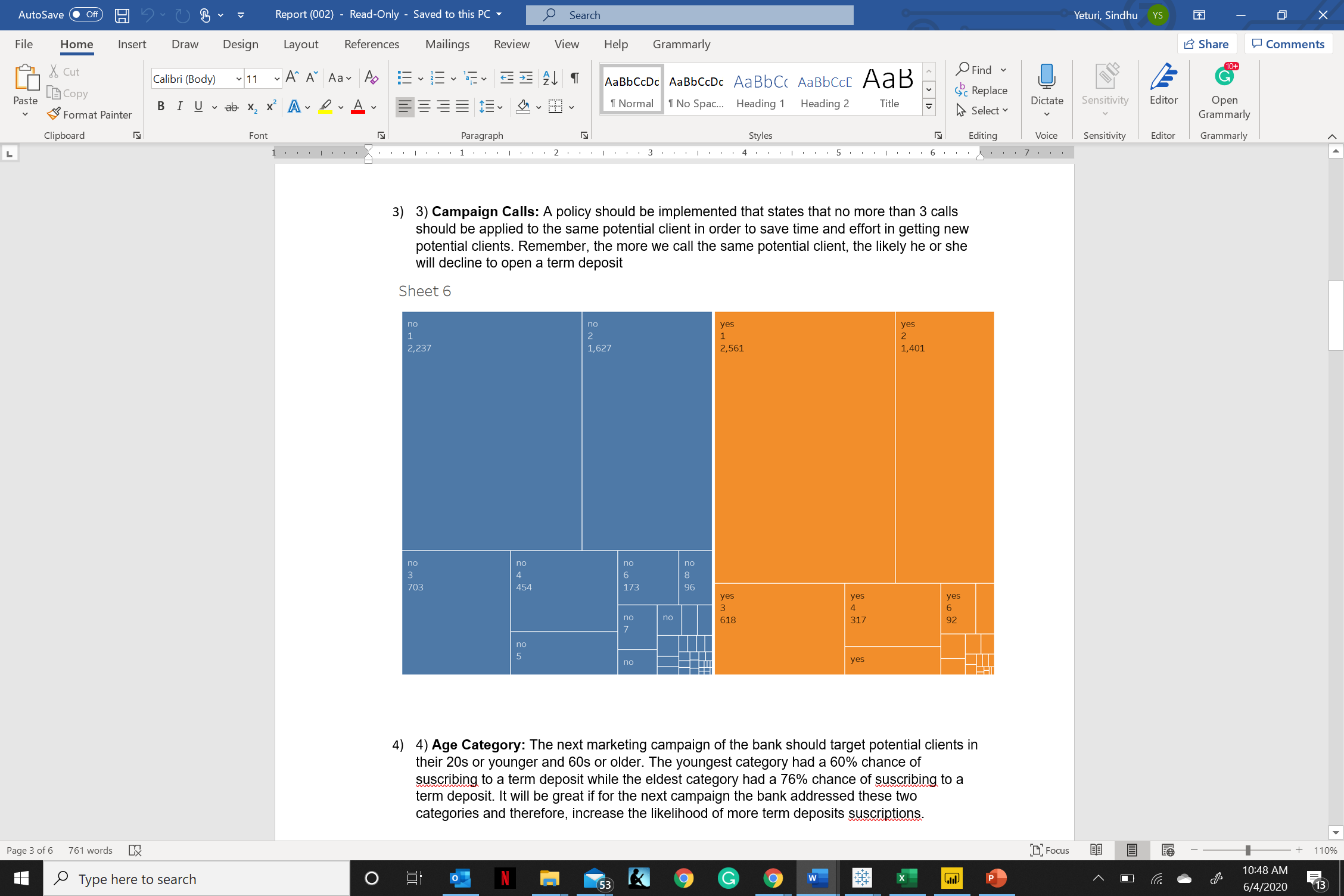


Fig2: Number of calls and conversions

From the chart above, it can be seen that the response rate was good when the number of marketing calls was 1. As the number of calls increased, the response rate reduced. Therefore it is ideal not to make more than 3 calls to the same potential client and also save time and effort which can be used in reaching new potential clients. The likeliness of declining to open a term deposit increases with the number of calls made to the same customer.

**Chart-3: Correlation between term deposit accounts and education**

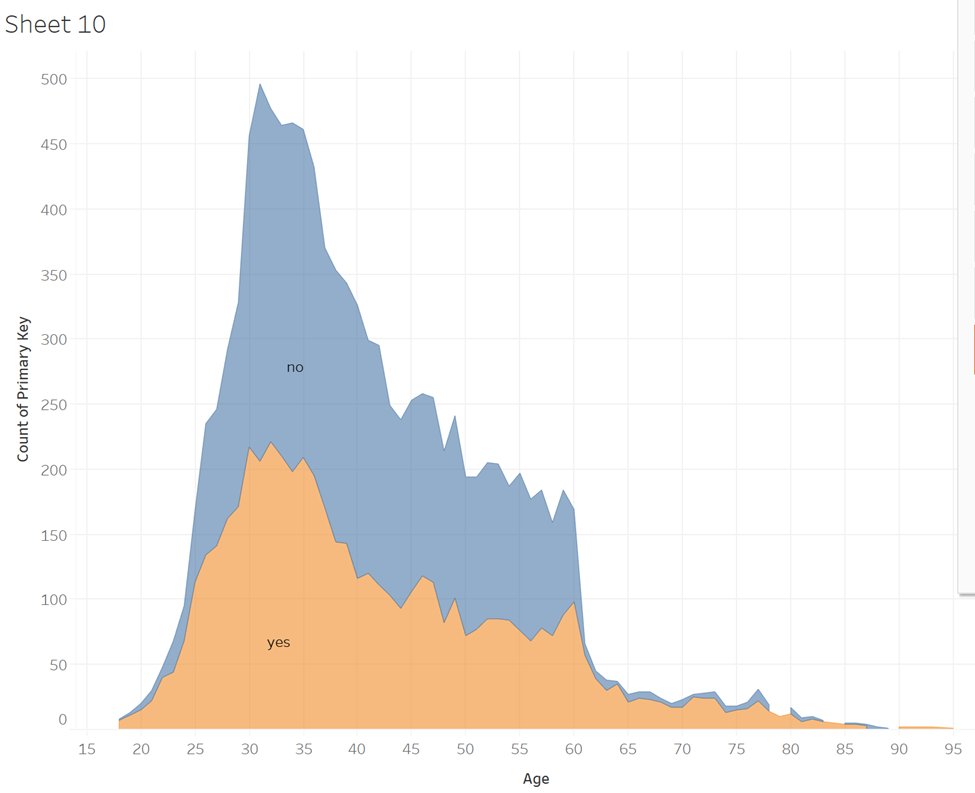


Fig3: Term deposit subscription and age

The graph above shows that the response rate to the term deposits is higher from the individuals belonging to the 25 to 45 years age group. The highest number of people that said yes to a term deposit account belonged to the same age group. The next marketing campaign of the bank should target potential clients in the age group between 25 to 45. The marketing campaigns can be customized to these two age groups to increase the subscription rate to the term deposits.

**Chart-4: Correlation between education and average account balance**

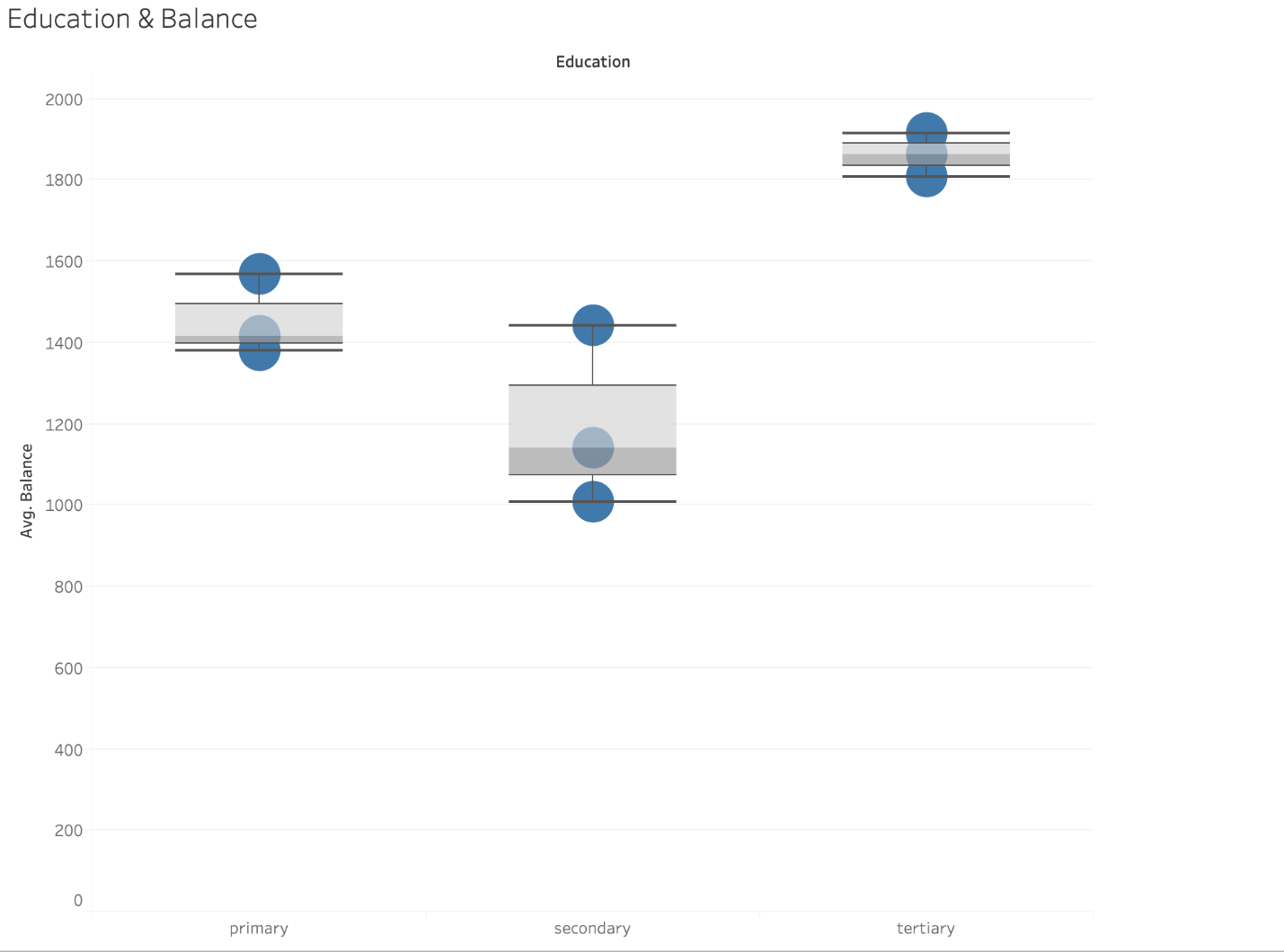


Fig4: Education, marital status and average balance

Primary Secondary Tertiary

The above chart represents the correlation between the average account balances, marital status, and education level. It is evident from the graph that people with a tertiary level of education had the highest account balances. This could be because the management level employees mostly consisted of people with a tertiary level of education and they seem to earn higher salaries. The dots in the box plot represent the marital status of individuals. The topmost dot represents the divorced individuals, the middle dot represents the married individuals and bottom-most represents the single people. Married people seem to have a higher account balance compared to single and divorced people.

**Chart-5: Correlation between marital status and deposit accounts**

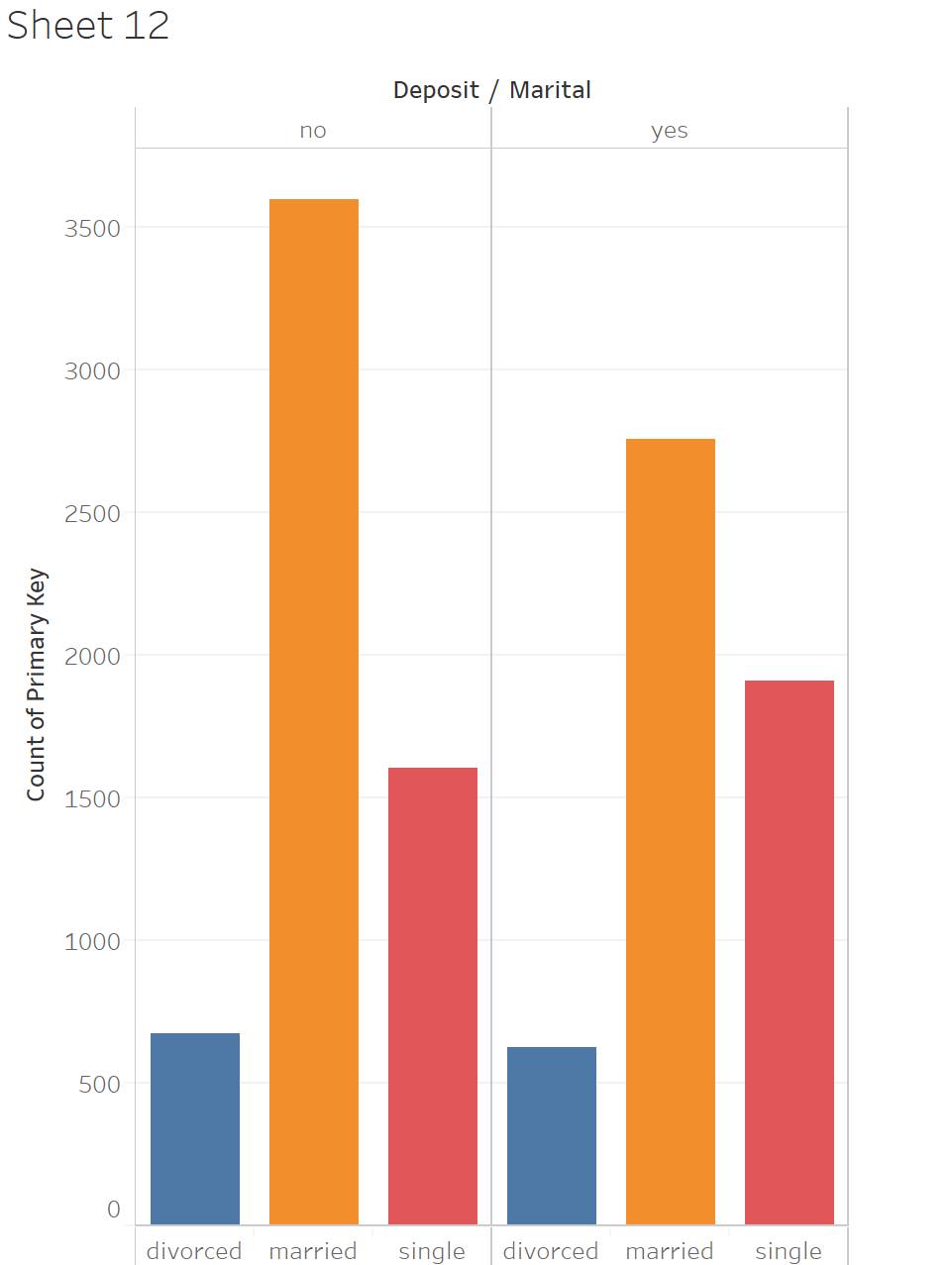


Fig5: Deposit accounts by marital status

The above bar chart shows that married people had the highest response rate. They also had the most number of deposit accounts. As seen in fig4, married people had the highest average account balance and this graph supports that information by showing that they also had the highest deposit accounts.

**Chart-6:**

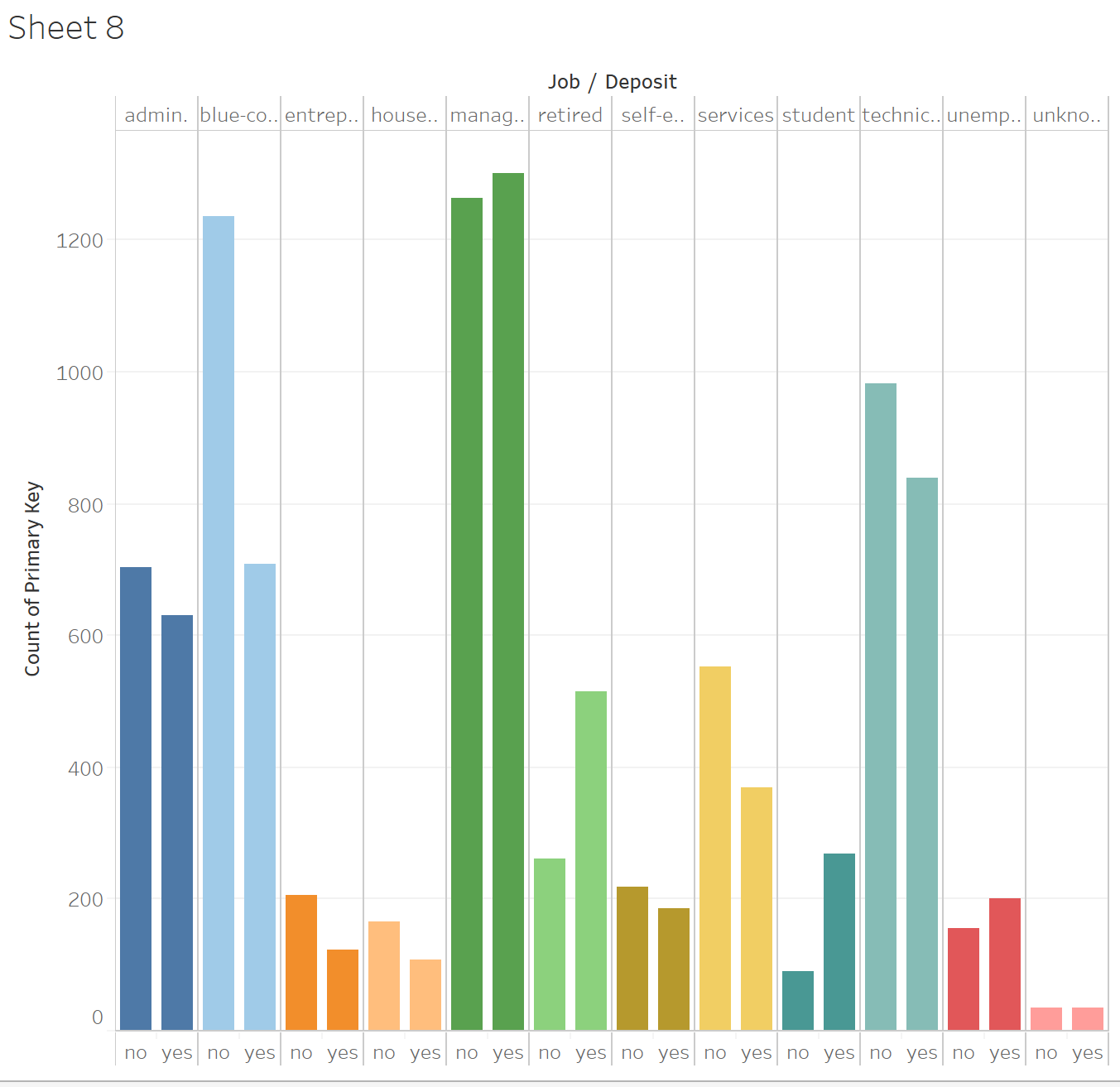


Fig6: Deposit accounts by job

The above chart shows the correlation between the job category and the number of deposit accounts by the category. From this graph, it can be inferred that the management level employees had the highest number of deposit accounts on average as compared to the other job categories. The management level employees also seem to have the highest response rate.

**Results of Exploratory Data Analysis:**

* People in the age group between 30-33 had the highest account balances followed by self-employed and the management level employees.
* People aged between 25-55 years had the highest positive response rate to the marketing calls made regarding the term deposit accounts.
* The conversion rate to term deposits was highest when the number of campaign calls were between 1 and decreased as the number of calls increased.
* It was interesting to see that the balances in the month of December and November were highest followed by the month of March.
* Most numbers of customers that had a loan account were married.
* Married people had a higher number of term deposit accounts in comparison with divorced and single people.
* People working at the management level had the highest average account balances.
* People with tertiary education were higher in number in the management roles and had more loan accounts in comparison with other job categories.
* People at the management level held the highest number of deposit accounts.

**Constraints to Exploratory data analysis:**

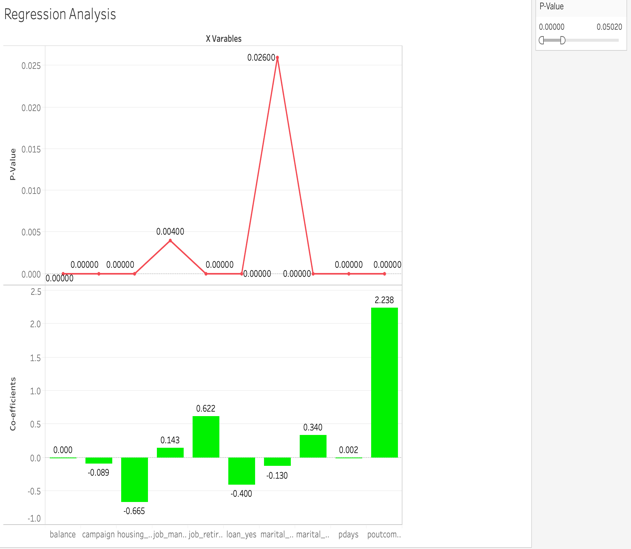
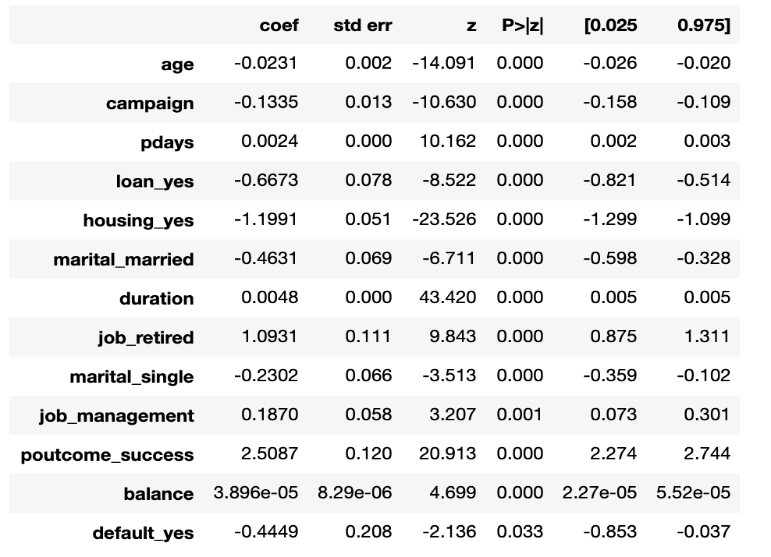
* Retired people, self-employed, and management-level employees can be targeted with customized offers for the term deposits.
* The right time to pitch the campaigns could be between September and March.

**Modeling and analysis**

**Logistic Regression**

This analysis helped to determine whether a customer will open a term deposit account with the bank based on different predictor variables. Y (categorical variable)– Whether a customer will belong to this category – “Yes’ for opening term deposit or not. X -variables (age, campaign, pdays, loan, housing loan, marital, duration, job, poutcome, default etc. It is used for classification.

The categorical variables amongst the x variables are loan, housing loan, poutcome, job, marital. We have introduced dummy variable to convert it into numerical value for the regression analysis with the help of Python code. The result of the analysis are as follows:



1. Almost all the x-variables are statistically significant (p<0.05)
2. Helped in identifying the coefficients which are important and reliable for determining the Y variable for this project from the list of many predictor variables.
3. The important variable for analysis will be loans, housing loans, job, balance, duration, and balance of the individual.

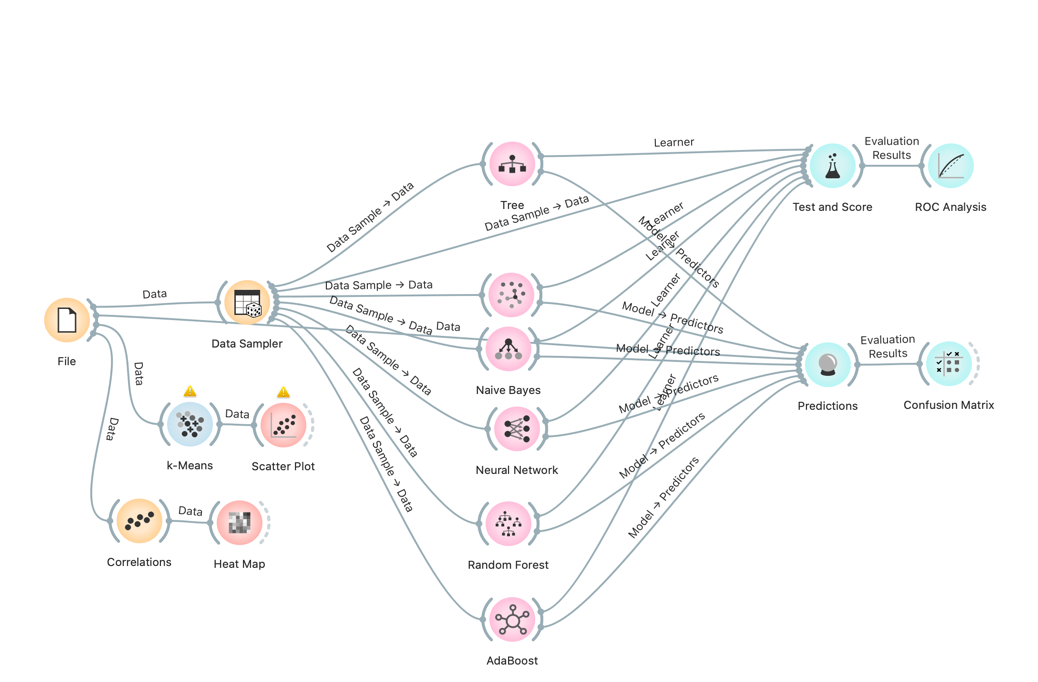
* If a person has a loan, the probability of opening a term deposit decreases by 0.667unit.
* If a person is retired the probability of opening a term deposit increases by 1.093 units
* If the job of a person ids management the probability of opening a term deposit increases by 0.187 units.
* If the duration of the campaign increases by 1 unit the probability of the opening a tern deposit increases by 0.048 units.

**Classification Models**

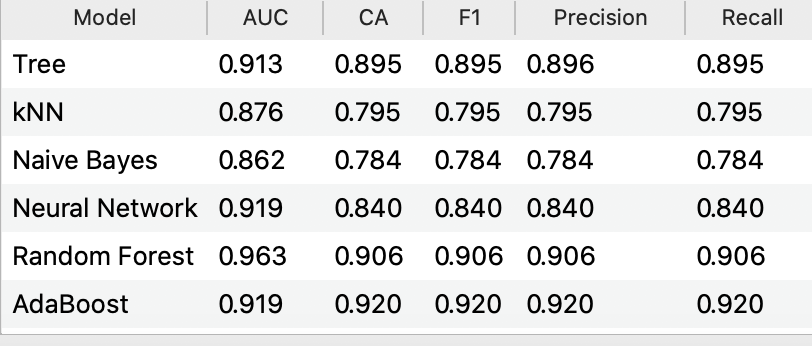
Other Classification models used to determine how well various classification models can help in predicting the success/failure of the next marketing campaign and if it will result in opening a term deposit account or not.

Orange is used as the data mining tool. The data set was imported in the orange canvas with the help of a File widget. The target variable(label) was set as a deposit (Yes/No). The data was then connected to the data sampler widget where we split the dataset into training and test data set in (70:30) ratio. To avoid over-fitting of data, cross-validation was set to 10 folds to ensure that the features which have the greatest influence on our label(deposit) are equally distributed. This algorithm will iterate through all the training and test sets and the main purpose of this is to grab the overall pattern of the data.

The Orange data mining canvas screenshot:



The correlation widget helped to plot the correlation between various numeric variables in the data. The result of the analysis states that there is a positive correlation between campaign, day, age, balance, and duration.

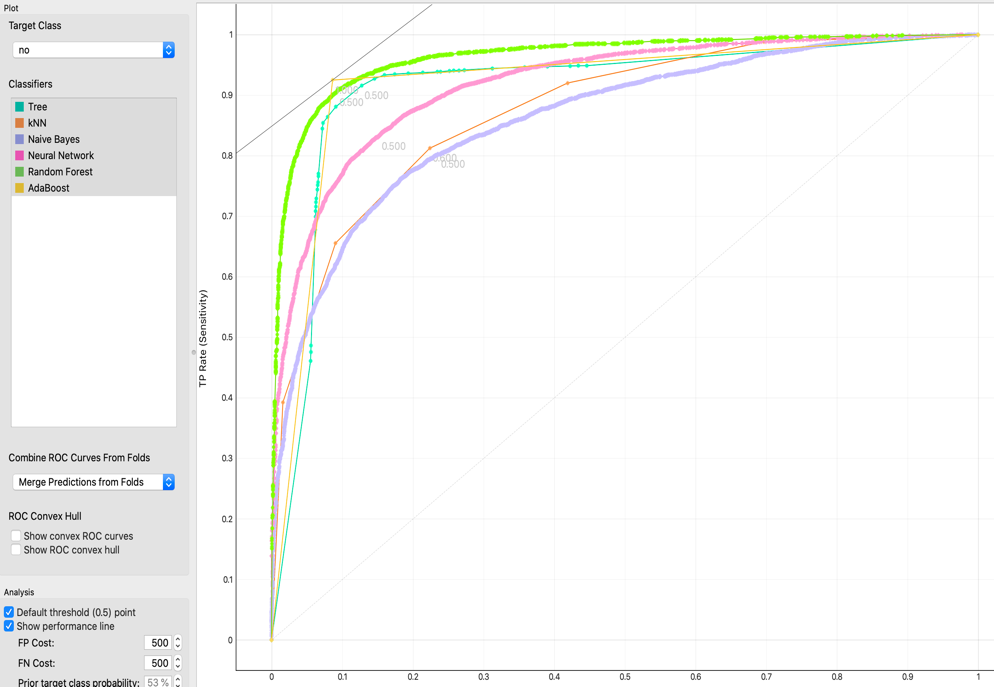


Prediction algorithm used to predict the classification accuracy of opening a term deposit for the test data.

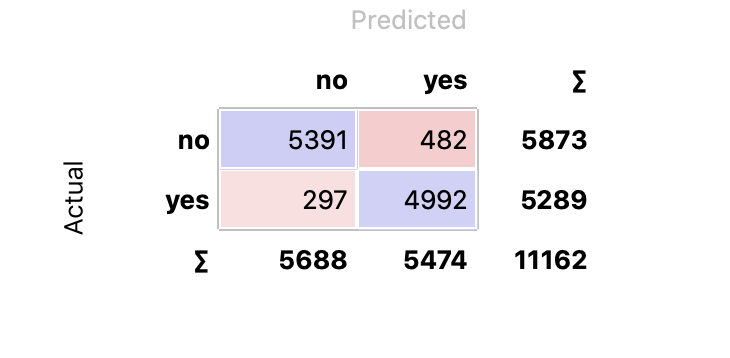
The Models Used are - **Decision Tree, Random Forest, Naïve Bayes, Neural network, Random Forest, AdaBoost**

The classification accuracy is highest for **AdaBoost** **(92) %** followed by **Random Forest (90.6)%**.

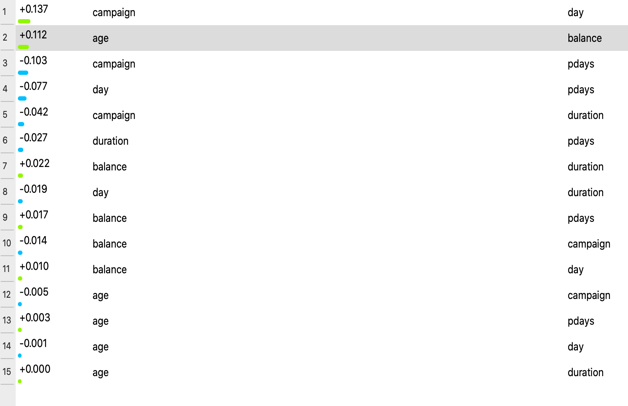
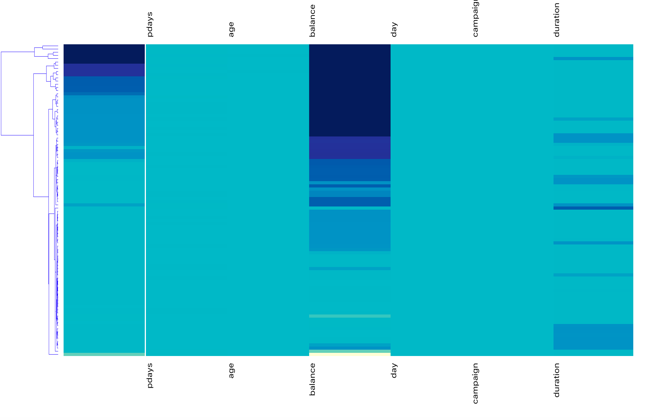
The **ROC curve** tells us how well our classifier is classifying between term deposit subscriptions (True Positives) and non-term deposit subscriptions. The closer is the line to our top-left corner the better is our model separating both classes. Random Forest has the maximum ROC score of **96.3%.**



The Confusion matrix output for Random Forest to verify how the model performed in classifying the customers that are likely to subscribe to a term deposit.

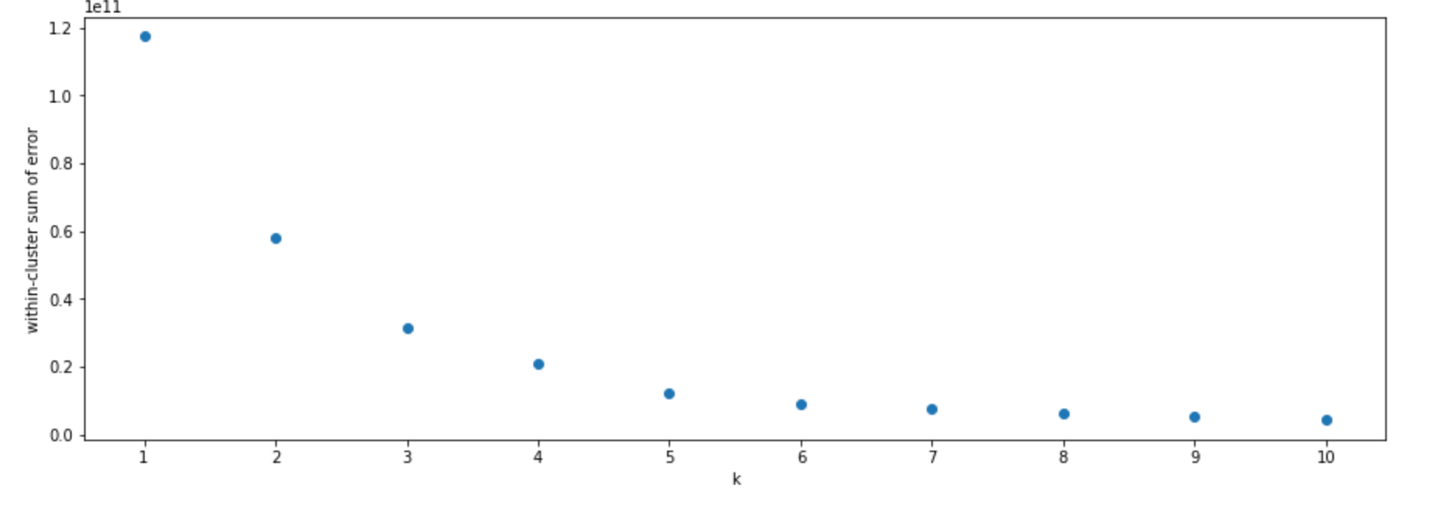


**Heat Map** – Correlation heat map for numerical variables. The values are represented by color stating the higher a certain value, the darker the represented color.

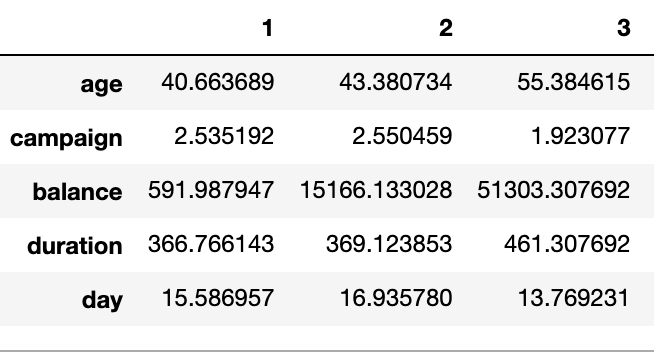
 

**K- Means Clustering** – It is used to group similar instances together and create segments. We did not have any survey data so it was a bit difficult to find out the clusters.

The graph between within clusters sum of errors and k and with the help of elbow method we tried to determine the optimal number of K = 3



The three different clusters are as follows:



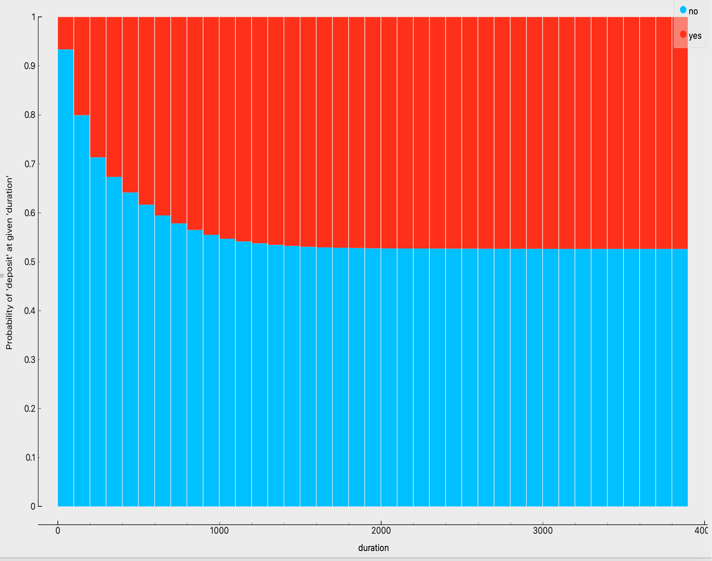
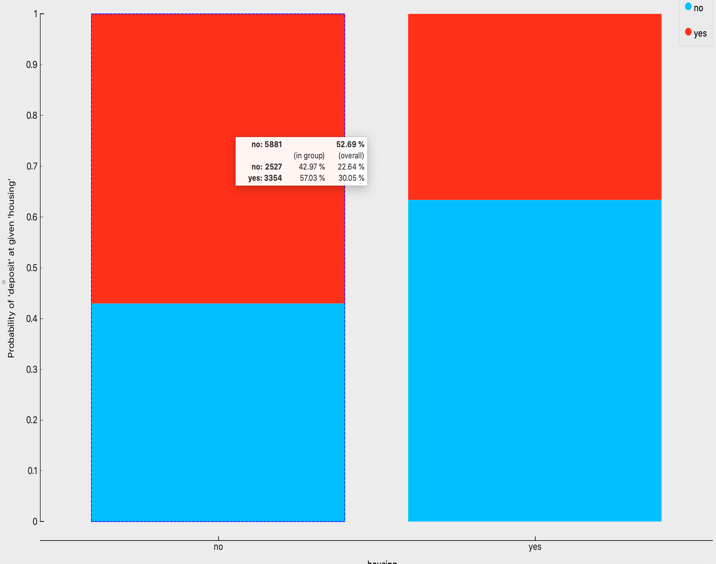
Cluster 1 represents the average age of 40 with the number of contacts performed during the campaign is 2 having a balance of 591and the campaign duration (call) was 366 seconds) on the 15th day.

Cluster 2 represents the average age of 43 with the number of contacts performed during the campaign is 2 having a balance of 15166 and the campaign duration (call) was 369 seconds) on the 16th day.

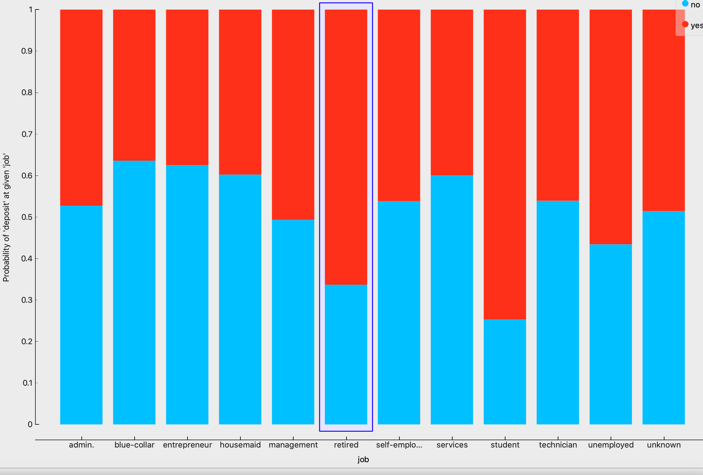
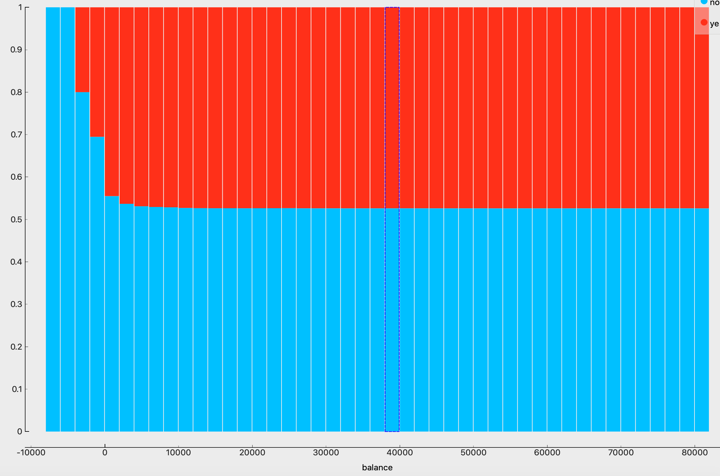
Cluster 3 represents the average age of 55 with the number of contacts performed during the campaign is 2 having a balance of 51303 and the campaign duration (call) was 461 seconds) on 14th day.

**Note: There was very less data for young group (18-25).**

**Data Visualization after clustering:**

** **

* The chart above represents that the probability of deposit is more in a cluster where there is no housing loan. This means that the potential client has financial compromises to pay back its house loan and thus, there is no cash for he or she to subscribe to a term deposit account.
* Target individuals with a higher duration above average, there is a highly likelihood that this target group would open a term deposit account.

Potential clients were retired who were the most likely to subscribe to a term deposit. Retired individuals tend to not spend big and are more likely to put their cash to work by lending it to the financial institution.

Customers with higher balance have higher probability to open a term deposit with the bank.

**Recommendations:**

Recommendations are based on predictive models and exploratory data analysis to improve the next marketing campaign.

1. Preferred month for marketing activities: The next marketing campaign should focus on months like March, September, October, and December, as these months shows the highest percentage of individuals subscribing to a term deposit.
2. A policy should be implemented that states that no more than 3 calls should be applied to the same potential client to save time and effort in getting new potential clients. As exploratory data analysis suggest, the more we call the same potential client, the likelihood of he or she will decline to open a term deposit
3. Occupation: Potential clients that were students or retired were the most likely to subscribe to a term deposit. Retired individuals tend to have more term deposits to gain some cash through interest payments. Students were the other group that used to subscribe to term deposits.

The job roles like admin, technician, blue-collar, management contribute to more than 60% of the total targeted individual. Hence with the assumption that the sample represents the demography of all client, the marketing campaign should focus on them with different

1. Balance**:** The next campaign should focus on potential clients with average and high balances to increase the likelihood of subscribing to a term deposit.
2. House Loan: The marketing campaign should focus on customers who don’t have any ongoing housing loan, as the probability of them subscribing to term deposit goes down drastically, as suggested by logistic regression results.
3. Target individuals with a higher duration (above 375): Target the target group that is above average in duration, there is a high likelihood that this target group would open a term deposit account.
4. Develop a Questionnaire: Since the duration of the call is the feature that most positively correlates with whether a potential client will open a term deposit or not, by providing an interesting questionnaire for potential clients during the calls the conversation length might increase