Bank Marketing

Which customers to target?

Jyothirmayee Nagireddy - October 2020

Contents

1. The Problem

Everywhere in the world, most people trust their money with already existing financial institutions like the banks for savings, financial transactions, or just storing their money. Banks are considered to be a very safe option for their money.

A term deposit is a cash investment held at a financial institution. Your money is invested for an agreed rate of interest over a fixed amount of time, or term. The cons associated with this method are quite less, depending on the financial institution you are dealing with. Typically, the money can only be withdrawn at the end of the period - or earlier with a penalty attached.

Instead of earning the usual rate of interest on their balance, many would like a higher rate of interest. Among other things, an easy way to achieve this is by subscribing to a term deposit. Given the option, few customers subscribe for a term deposit. Hence the goal of this project is to predict whether a customer will subscribe to a term deposit or not.

Based on the data, the clients are classified into two categories. ‘Yes’ for clients who subscribed for a term deposit and ‘no’ for clients who did not subscribe for a term deposit. Usually, the number of clients who subscribe to a term deposit would be very small. The challenges for this problem arises due to the skewed nature of the data. This project is an imbalanced classification problem with an imbalance ratio of 10%.

Problem: Improve the marketing campaign of a Portuguese bank by analyzing client’s data and previous marketing campaign data and predict which customers to target.

2. The Client

The data was provided by the Portuguese bank to better target its customers such that more customers subscribe to a term deposit. The idea of solving and predicting which customers to target does not end with the banking industry. It can be used in various other industries that deal with increasing their income by addressing the right customer.

3. The Data

The data used in this project was acquired from the UCI Repository.

The data is related to the direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe to a term deposit (variable y).

The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, to assess if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

The data has 45,211 instances and has both categorical data as well as numerical data.

Dataset: http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#

3.1 The Attributes

There are 16 feature attributes and 1 target attribute (y). The feature attributes are of three classes:

- Features related to information regarding bank clients.
- Features related to the last contact of the current campaign.
- Features related to the previous campaign.

Bank client data:

1. **age** (numeric)
2. **job** : type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur","student", "blue-collar","self-employed","retired","technician","services")
3. **marital** : marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)
4. **education** (categorical: "unknown","secondary","primary","tertiary")
5. **default**: has credit in default? (binary: "yes","no")
6. **balance**: average yearly balance, in euros (numeric)
7. **housing**: has a housing loan? (binary: "yes","no")

8. **loan**: has personal loan? (binary: "yes","no")

Related with the last contact of the current campaign:

9. **contact**: contact communication type (categorical: "unknown","telephone","cellular")
10. **day**: last contact day of the month (numeric)
11. **month**: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
12. **duration**: last contact duration, in seconds (numeric)

    Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

Other attributes:

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 (numeric, -1 means the client was not previously contacted)
15. **previous**: number of contacts performed before this campaign and for this client (numeric)
16. **poutcome**: outcome of the previous marketing campaign (categorical: "unknown", "other", "failure", "success")

Output variable (desired target):

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

3.2 Data Wrangling

Overview: This section describes the various data cleaning and data wrangling methods applied to the Bank Marketing Data.

Data reading: The data was loaded into the Jupyter Notebook as a Pandas DataFrame with a delimiter ';'.

Various steps in data wrangling:

- **.head()** function is used to get a clear picture of what the data looks like.
- **.describe()** - To compute various statistics on the DataFrame columns.

  It computes count, mean, standard deviation, minimum, 25th percentile, 50th percentile, 75th percentile, and maximum for the numerical data columns.
- **.info()** - gives the basic information of each of the columns. It returns the column name along with the count of the non-null rows of the data along with the data type of the column.
- **Missing values check: isnull()** function of the panda library was used to check if there are any missing values in the DataFrame. A logical function, that checks if there are any missing values was passed for each row inside the DataFrame. The output was applied to each column by specifying (axis =1). The count of isnull() function returned no missing values.
- **Knowing the categorical variables- .unique()** was used to display the various values of the categorical columns. It was computed for 10 categorical columns.
- **Knowing the numerical variables- boxplots** were plotted for the numerical columns. The boxplots revealed **outliers** for the numerical data. But the outliers are not removed for now for further analysis.

Data Cleaning: The data is clean and it does not require any pivoting or melting of the DataFrame. It does not require any further modifications.

3.3 Data Storytelling

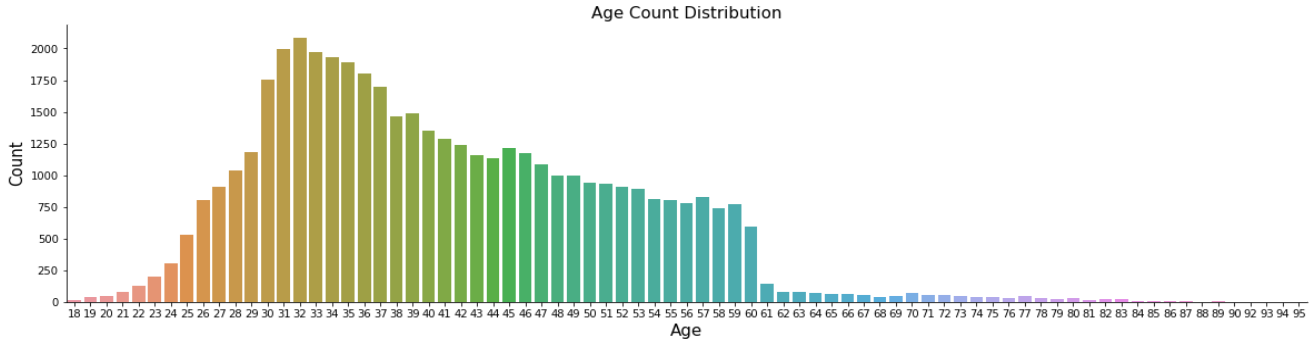
This section of the project deals with data storytelling. Different features are analyzed to see the general trends in the data. For the continuous or numerical features, histograms and bar plots are used to understand the data, whereas for categorical features countplot was used. In addition to this, categorical columns are further analyzed to understand the category of each feature that effectively produced the most 'yes' in the target variable y.

3.3.1 Analyzing various attributes

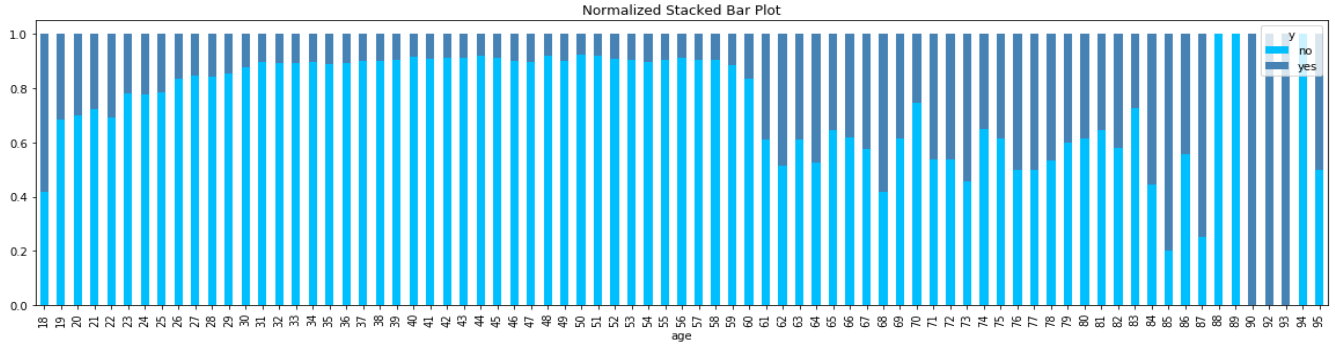
Bank Clients Information -

AGE - age of clients.

Countplot from seaborn gives a very good explanation of how the age of the clients is distributed from the minimum age of 18 to the maximum age of 95. Apart from countplot, histogram and boxplot were plotted.



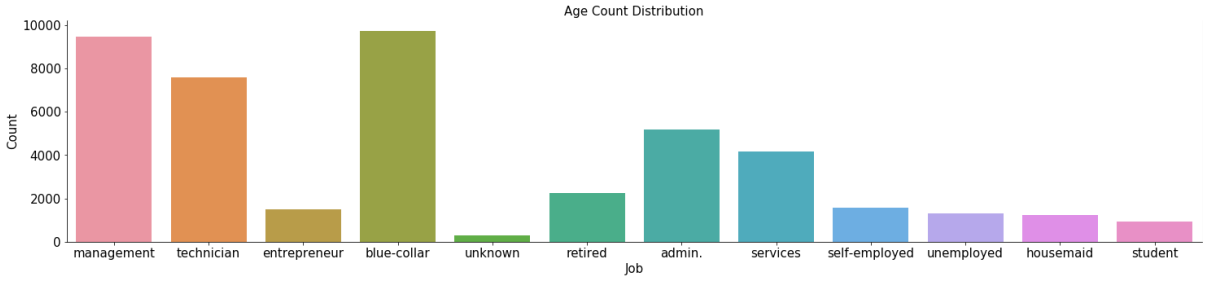
Even though the minimum and maximum age are largely at the end of the graphs, most of the age groups are centered in the middle. The mean age is about 40yrs and the standard deviation is 10 years. Most of the clients are middle-aged.

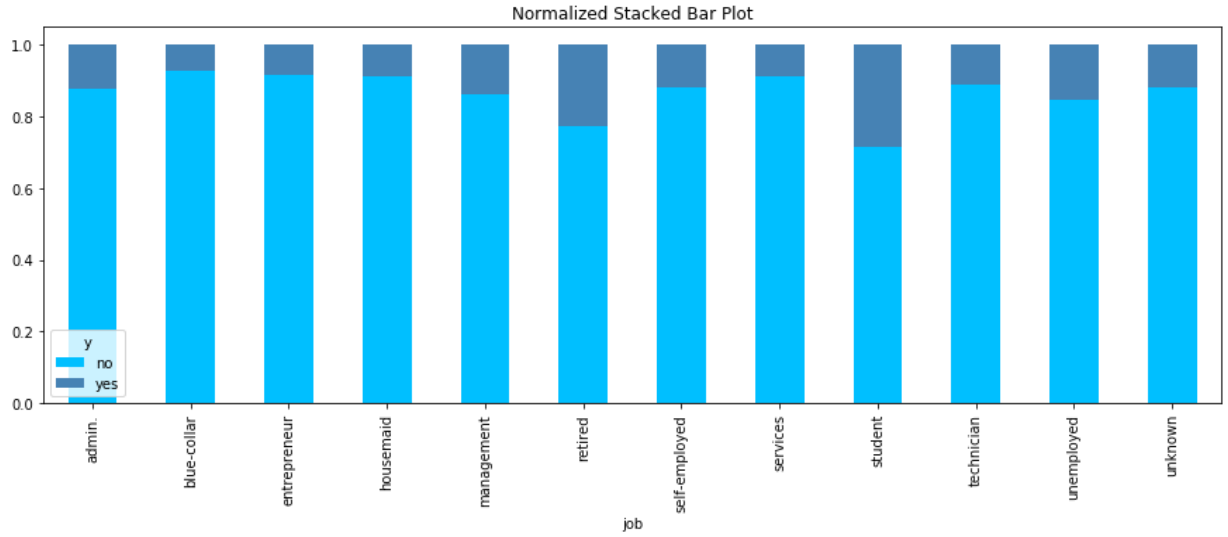


It can also be seen that customers older than 60 years have a higher tendency to opt for a term deposit. Ages below 22 and above 60 are very important and they should be targeted by the marketing campaign more than the other age groups.

JOB

The job categories - management, technician, and blue-collar have a high count of jobs compared to all other job categories.

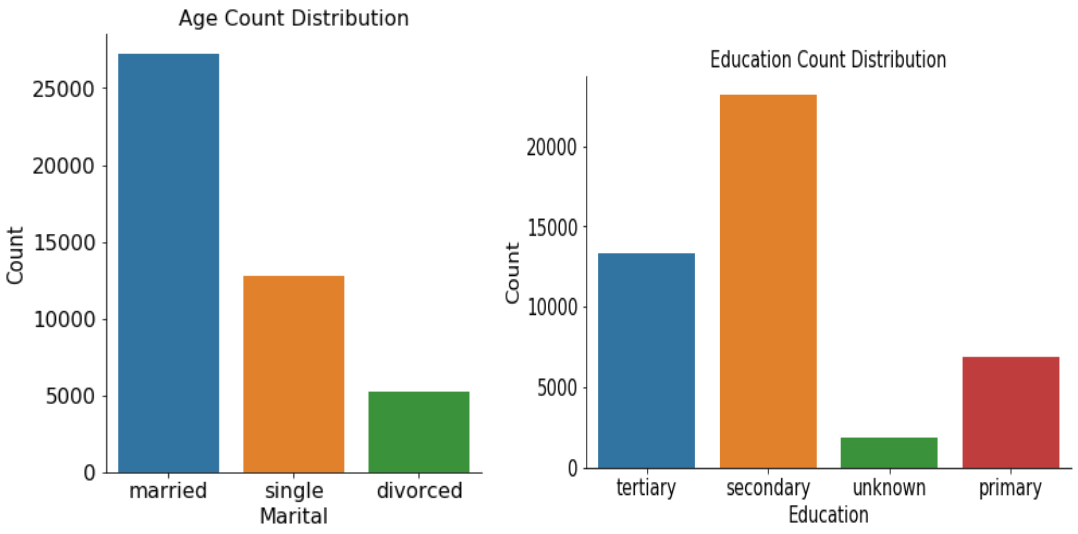




The proportion of bank clients who are getting low or no income (student, housemaid, unemployed, self-employed, unknown, entrepreneur) is very less compared to that of clients who are regularly paid. Maybe they are less prone to opening a bank account.

Students and retired clients are more prone to subscribing to a term deposit.

MARITAL, EDUCATION



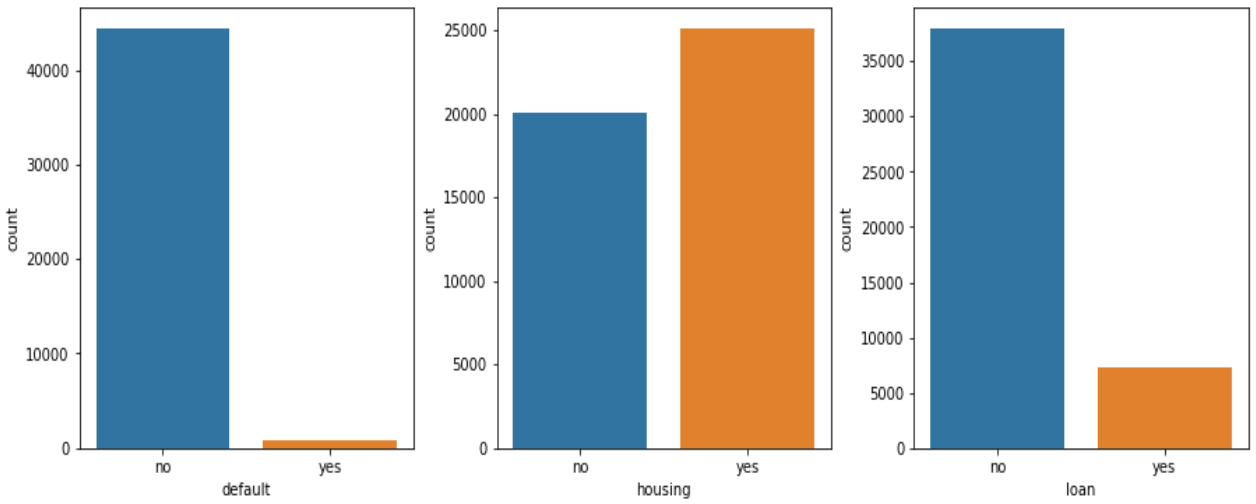
Married clients having a bank account is higher than single and divorced. Since the data for this feature is not that skewed, the data may represent the general population.

Education is grouped into 4 categories- primary, secondary, tertiary, and unknown. Among these 4 categories, secondary and tertiary categories are on the higher side.

The higher count of secondary and tertiary education can be because of being able to better provide for themselves economically and hence the need for a bank account.

Default, Housing & Loan

Whether the client has credit in default, whether the client has a housing loan, and whether the client has a personal loan.

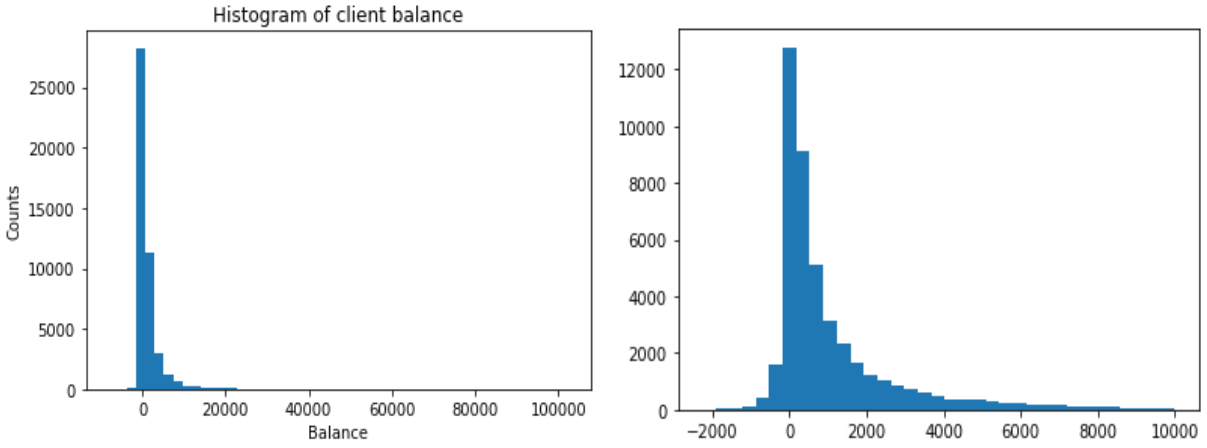


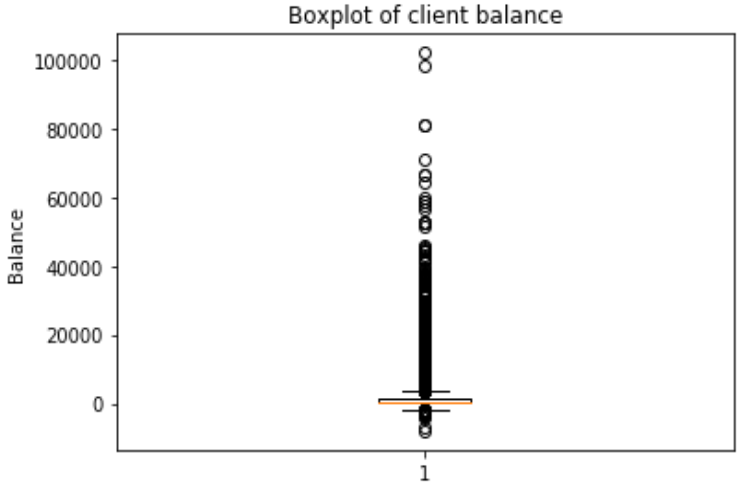
It can be understood from the bar plots that most of the clients do not have any credit in default and do not have a personal loan.

The clients having housing loans are nearly equal to the clients who do not have housing loans.

Balance

Average yearly balance, in euros. This is one of the numerical columns of the data set.





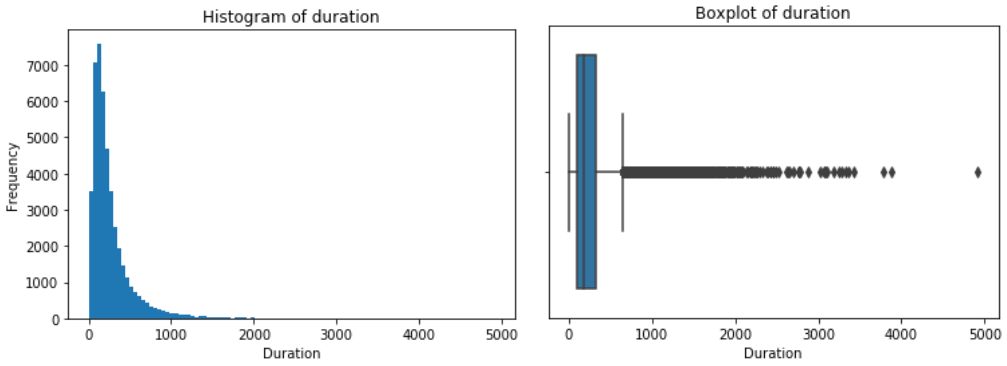
Balance ranges from a minimum of -8019 to a maximum of 102127. But 25th and 75th percentiles range between 72 to 1428 euros. It can be seen from the histogram that many clients have a negative balance. There are a lot of outliers, but they are kept for further analysis.

Clients who have negative balances were further analyzed. The value_counts() method on the term deposit ‘y’ for clients who had negative balances was performed. Out of 7280 clients, 502 clients subscribed for a term deposit. Hence, even clients who had negative balances are statistically important.

Attributes related to the last contact of the current campaign :

Duration

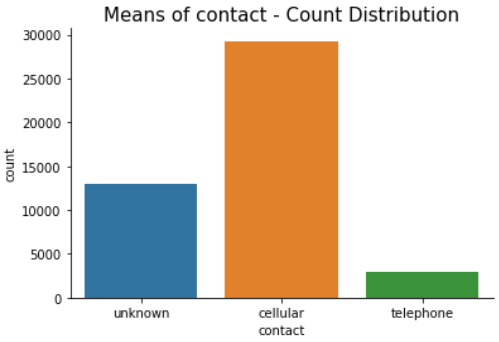
Call duration made by the telephone executive regarding the term deposit.



A histogram and boxplot were plotted to see the distribution of the call duration. It is understood that most of the call duration is grouped around the mean duration at around 258s.

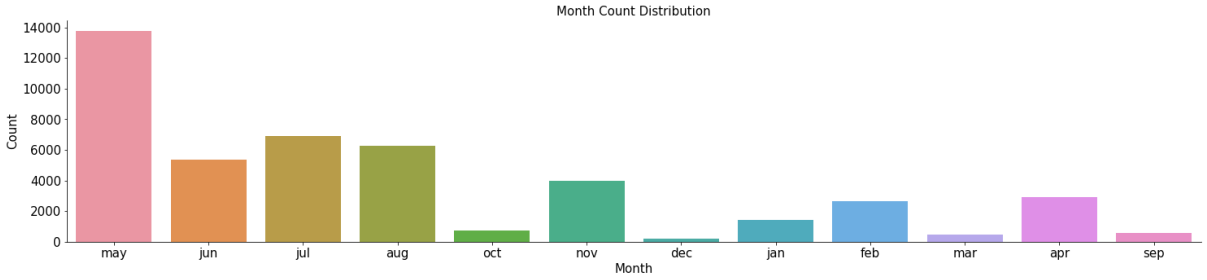
Most of the data is skewed towards the left. Data between the minimum and 75th percentile lies between 0 and 319 seconds. Whereas the maximum duration is 4918 seconds. The data below the 75th percentile for the duration of the call gives a clearer picture. The call duration is usually about 5 minutes. But occasionally it got a little higher. And the maximum value of 4918 s could be an outlier.

Contact, Month, Day



Communication type - unknown, telephone, and cellular.

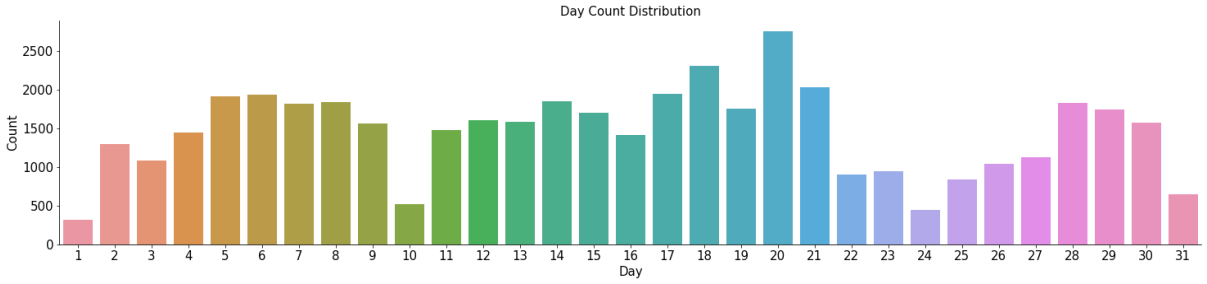
Cellular contact is the highest of the three categories. The higher proportion of cellular contact shows that it is the usual method of contact. And most people have and use cellular telephones rather than a usual (landline) telephone.



- last contact month of the year.

May month has the highest data points. Jun, July, August have the next highest data points. The rest of the months do not have many data points.

Many clients are contacted during these months - May to August. Summer seems to be the usual contact period regarding a term deposit. Maybe because the financial year starts in April and the banks may be under pressure to create a new stimulus. Winter might be a dormant period for contacting clients.



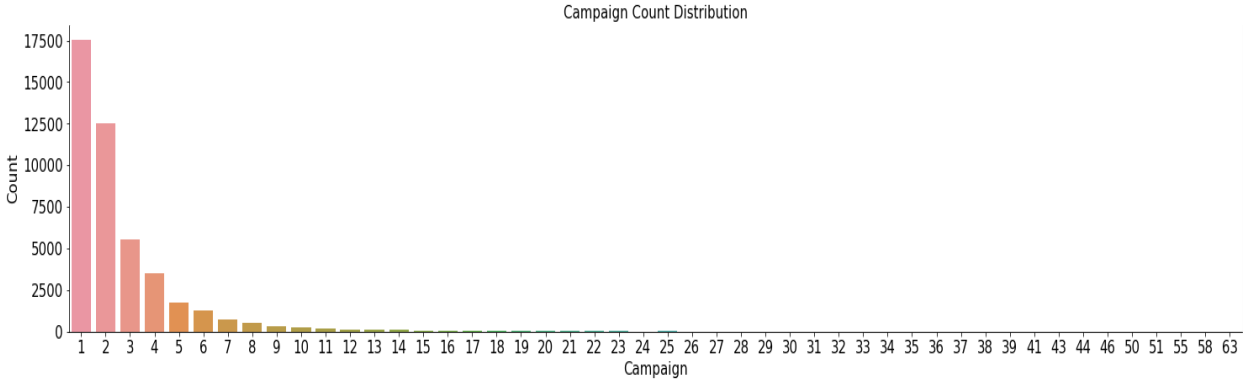
- last contact day of the month.

The last contact day of the month is nearly equally distributed. Among all the days of the month. This means that the clients are equally contacted all the time of the month.

OTHER ATTRIBUTES

Campaign

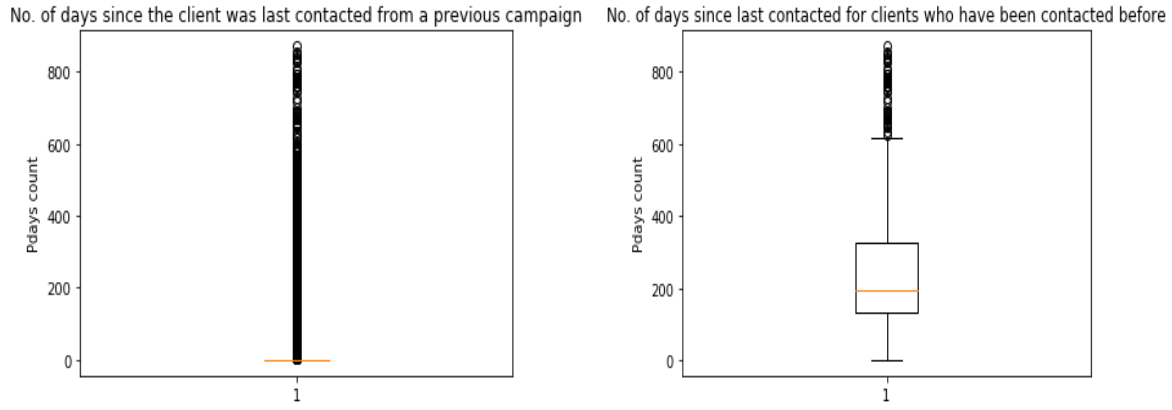
The number of contacts performed during this campaign and for this client.



75 percent of the clients were contacted once, twice, or thrice. Making clients who were contacted more than 6 times as outliers. About 3064 outliers are making the total percentage of the outliers to about 6.78%. But the outliers are not removed because the data among the outliers have a significant number of clients who have subscribed for a term deposit ‘y’.

Pdays

The number of days that passed by after the client was last contacted from a previous campaign.

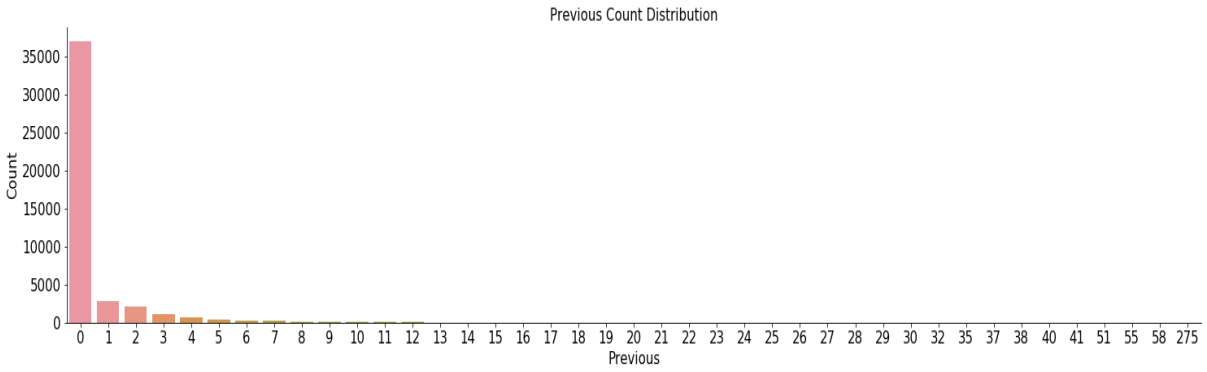


The first plot shows that nearly all the values are at point 0 (i.e., many clients were not connected before).

The second figure is a boxplot of the clients who have been contacted before. There are about 36954 clients who have not been contacted by a previous campaign (they are given as pdays = 0). Pdays = 0 data is removed and plotted with a boxplot to get a general sense of the data. Apart from Pdays = 0, the rest of the data is normally distributed.

Previous

The number of contacts performed before this campaign and for this client.

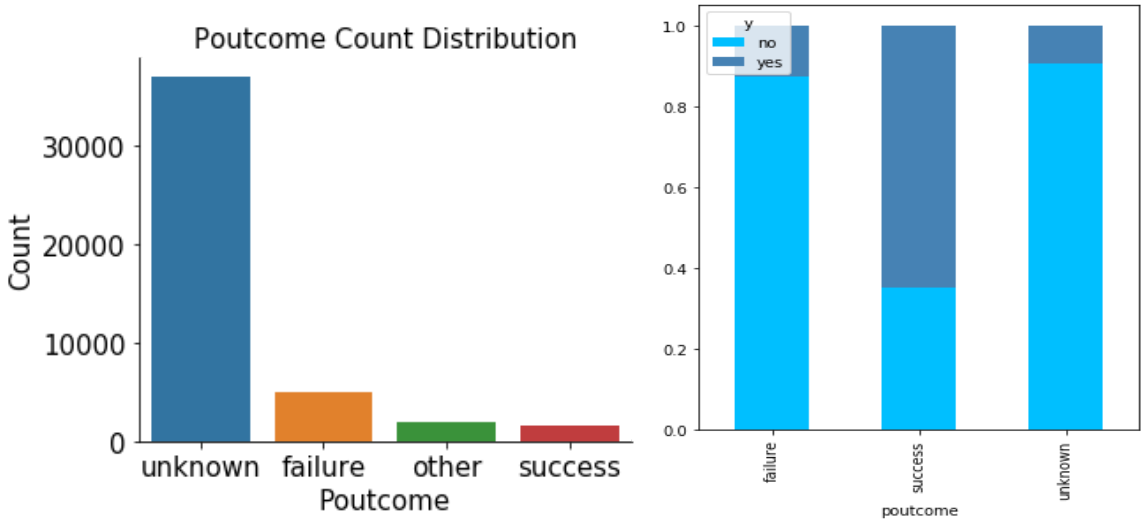


Like the previous attribute, this attribute also has about 36,954 clients who have not been contacted by a previous campaign (denoted by previous = 0). Apart from that, the rest of the data is normally distributed.

Poutcome

The outcome of the previous marketing campaign.

Poutcome is grouped into 4 categories- unknown, other, failure, and success.



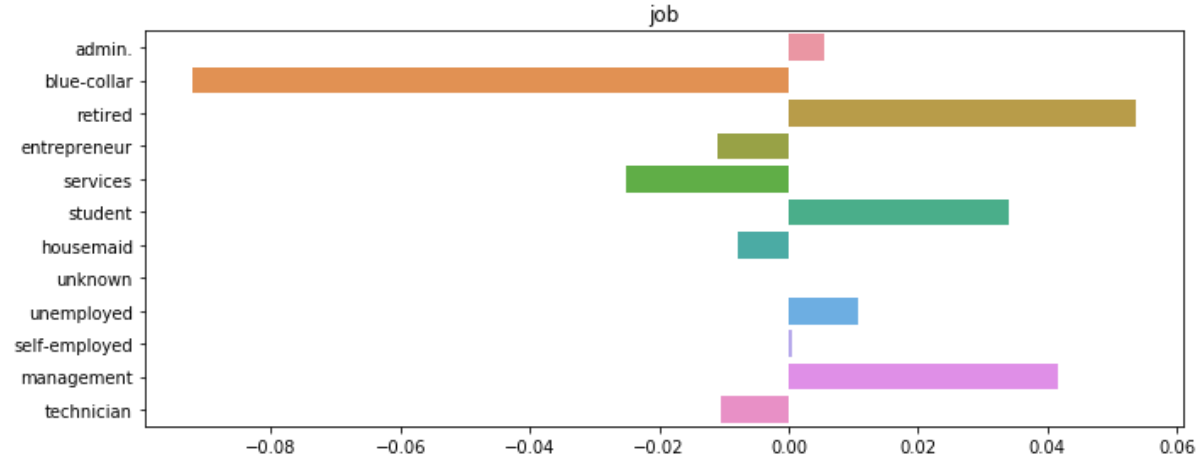
Unknown and other categories are pretty much the same and hence they are grouped and it has the highest proportion of data points. It is understood from this that the data regarding the previous outcome is not known.

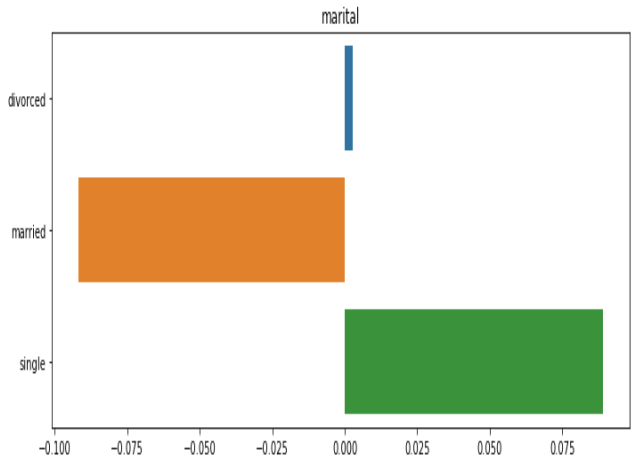
There is a higher chance that a customer will subscribe for a term deposit if he/she has already done that in the previous campaign.

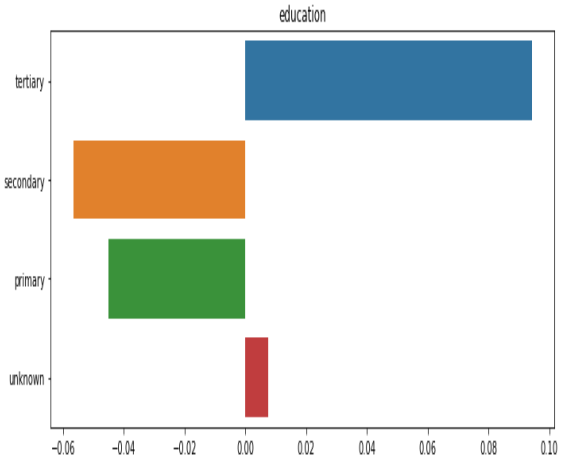
3.3.2 Categorical Columns

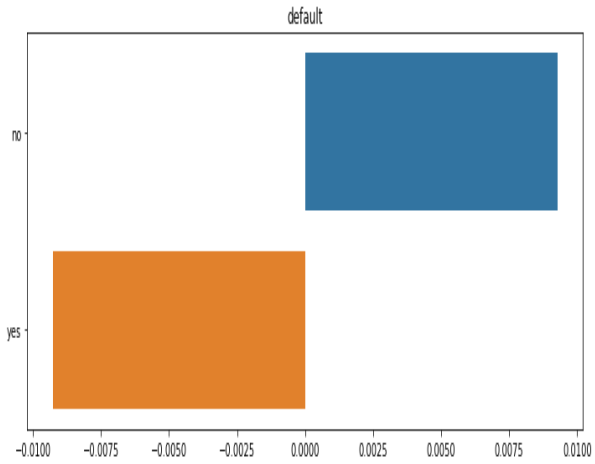
Categorical columns are further treated to analyze every class of each feature. This is done to understand what attributes of the clients or attributes relating to the campaign have a higher chance of clients subscribing to a term deposit.

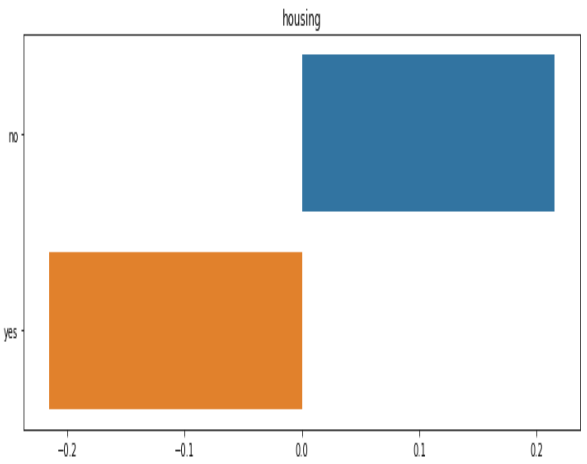
Each class per feature is normalized and the differences are plotted between positive and negative frequencies. Positive values imply this category favors clients that will subscribe and the negative values category that favor not buying the product.

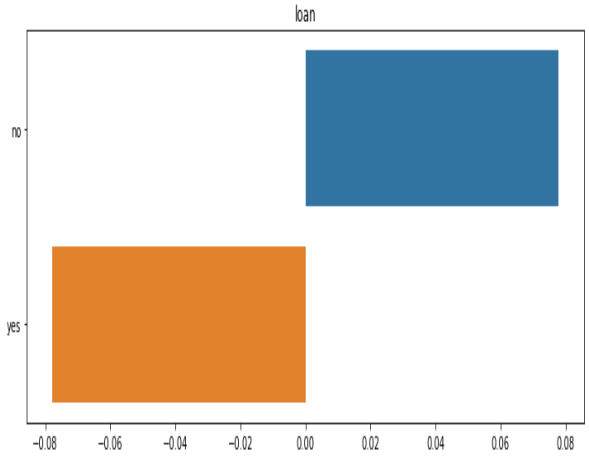


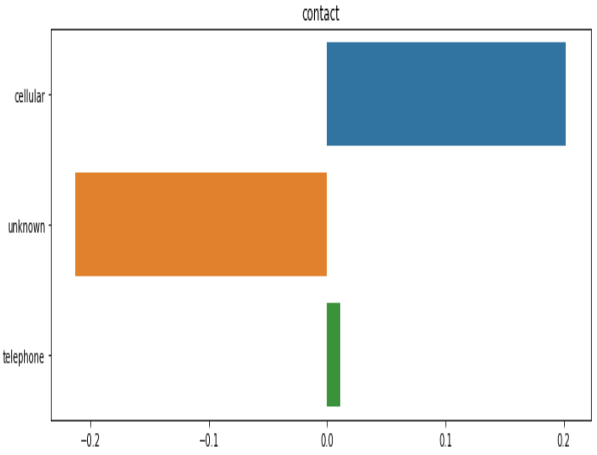


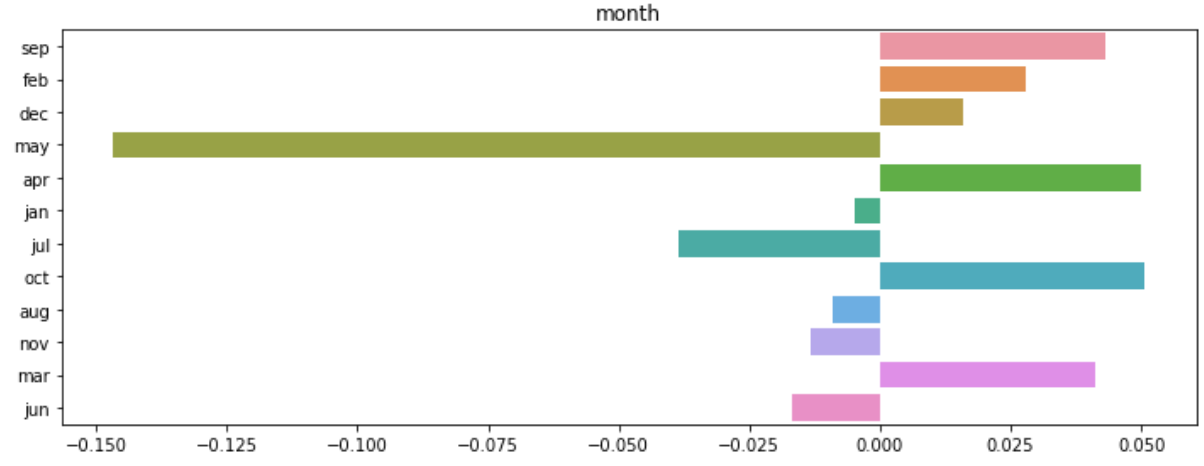


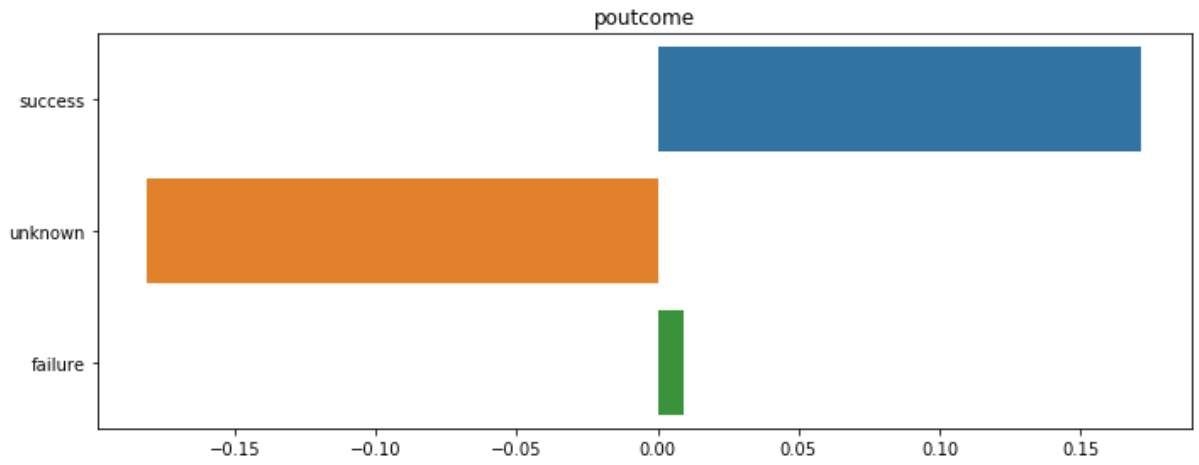

default
no
yes
−0.0100
−0.0075
−0.0050
−0.0025
0.0000
0.0025
0.0050
0.0075
0.0100


housing
no
yes
−0.2
−0.1
0.0
0.1
0.2


loan
no
yes
−0.08
−0.06
−0.04
−0.02
0.00
0.02
0.04
0.06
0.08


contact
cellular
unknown
telephone
−0.2
−0.1
0.0
0.1
0.2


month
sep
feb
dec
may
apr
jan
jul
oct
aug
nov
mar
jun
−0.150
−0.125
−0.100
−0.075
−0.050
−0.025
0.000
0.025
0.050



There are many take-away for the marketing team from the above plots:

- Retired clients, students, and management professionals are to be more targeted and blue-collar professionals less.
- Target single clients compared to married clients
- Target clients with higher education
- Target clients who do not have any credits in default, no housing loans, and no personal loan.
- Cellular contact is better.
- Do not contact in May month. It may be because of the summer vacations that families generally take during this month.
- Target the clients for whom the outcome of the previous campaign was successful.

3.3.2 Conclusion from data storytelling

Client information attributes such as age, job, education, marital status gives a general representation of people having a bank account. Balance in the bank account gives an interesting story. There are various clients with a balance 0 or less than zero and they still subscribe to a term deposit. The marketing campaign must target retired clients and also clients with higher education first.

Attributes related to the last contact of the current campaign tell us that most clients are contacted in summer and it could be any day of the month with a usual contact method being cellular and the duration mostly under 5 minutes. But the clients contacted in May have the highest no's in the target variable.

And some attributes talk about the previous campaigns. It is better to target clients who have already been contacted before.

4. Inferential statistics

The section on inferential statistics is looking at the statistical significance of observations made and thoughts had during the EDA. This is an essential step to understanding whether or not the differences between the customers who subscribed to term deposit and the customers who did not, are factual or just happened by chance.

The focus for us lies in three categories:

- Hypothesis testing for continuous features.
- Analysis of discrete features.
- Collinearity and Heatmap between features.

4.1 Hypothesis Testing

Hypothesis testing was done for continuous variables to check if the differences to the target variable ‘y’ caused by the independent features are statistically significant or not. A T-test is used for hypothesis testing since the data does not represent the whole population.

Three variables that are analyzed are:

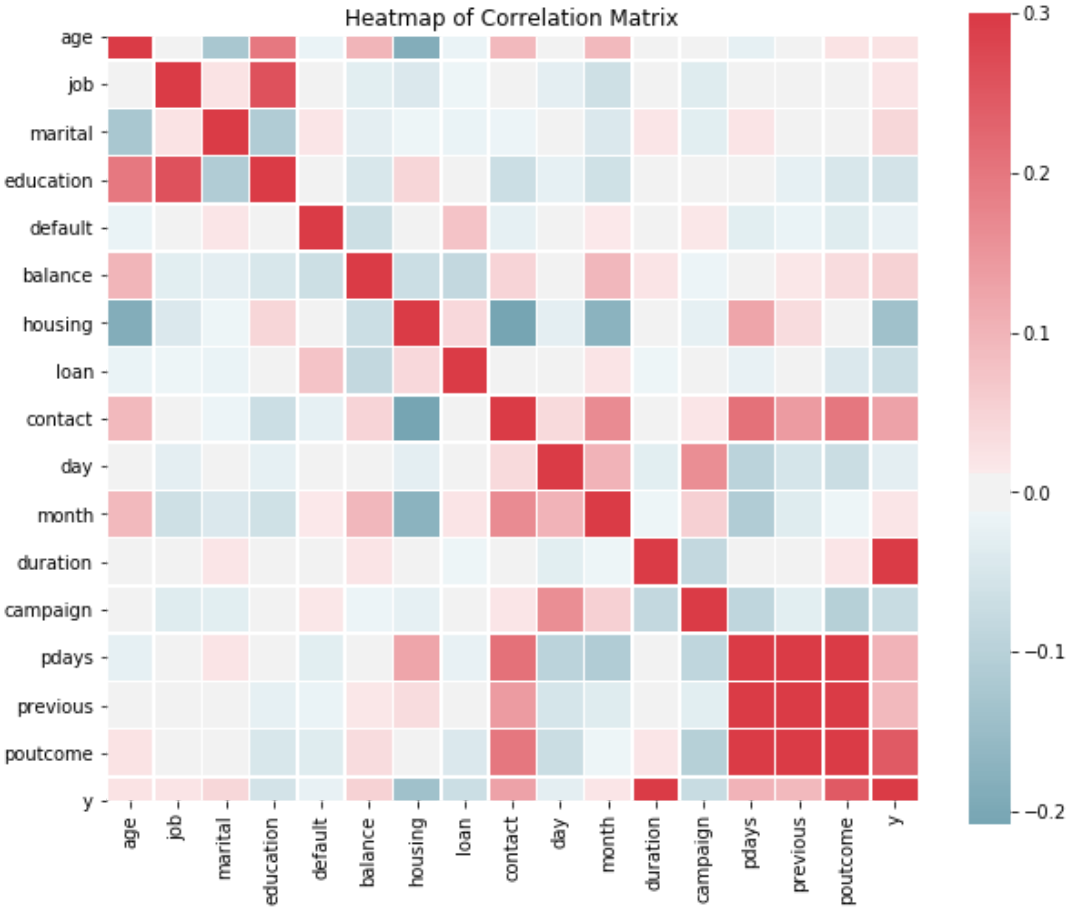
- Age - age of the customer
- Balance - balance in the customer’s account
- Duration - call duration of the latest contact.

A hypothesis to determine if the differences within the independent features that lead to customers subscribing to a term deposit or not are statistically significant. The p-value for all three tests is less than 10^-5.

One more hypothesis test was done on the balance in the customer’s account. The Null hypothesis was to determine whether the customers with zero or negative balance who subscribed to term deposits are similar to customers with a positive balance. The Null hypothesis was rejected with a p-value of 10^-300.

Having a positive balance leads to higher customers subscribing to a term deposit.

4.2 Collinearity and Heatmap between features



The target feature ‘y’ has a positive correlation with the continuous variables age, balance, and duration(duration is strongly correlated).

The values of discrete variables are converted manually from string to numerical quantities.

A strong correlation can also be seen with the ‘poutcome’ feature as well. But the data has a very high number of unknown values.

4.3 Conclusion - inferential statistics

The data proves that the variation in the features that caused different categories in the target variable (yes/no) did not occur by chance/ are statistically significant. This allows us to reject the Null Hypothesis.

5. Machine Learning

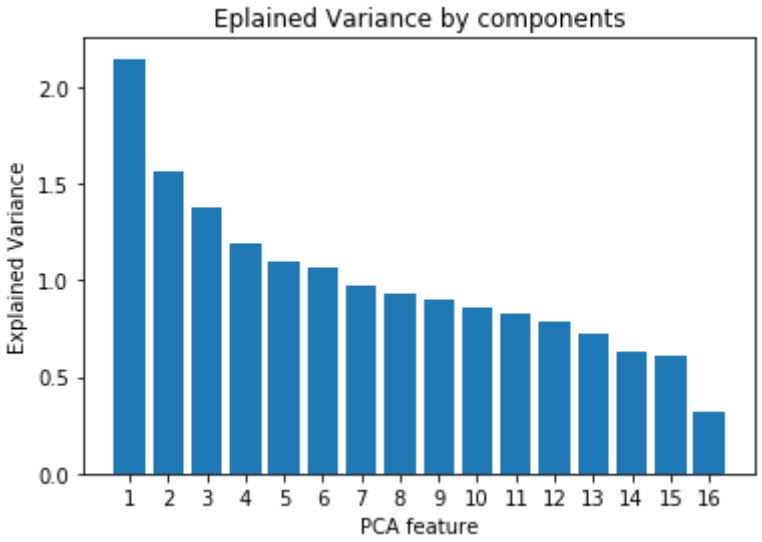
This section deals with building a predictive model for the client data for predicting whether or not the client subscribes for a term deposit.

It focuses majorly on:

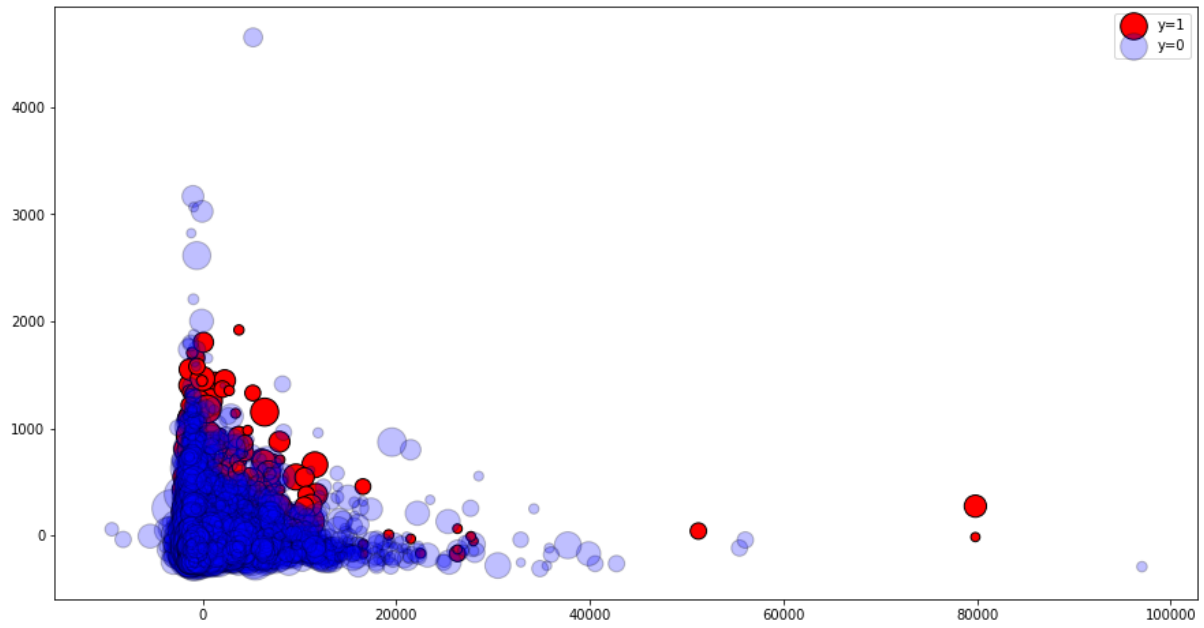
- Principal component analysis (PCA)
- Upsampling the skewed data
- Base model selection from various models trained.
- Test metric
- Hyper-parameter tuning
- Predicting new data

5.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is used to explain the variance-covariance structure of a set of variables through linear combinations. In this case, it is mainly done for two reasons. To understand the variance-covariance structure for the features and for visualization of the data that is multi-dimensional.



From the above barplot, the difference of explained variance is minimal after the number of features increases to 3. For that reason, feature reduction is done later after building the model during hyperparameter tuning.



PCA can also be used as a great visualization tool to get an idea about data we are about to predict. The scatter plot below shows how this data might look in a two-dimensional space.

5.2 Upsampling

As we can see from the visualization above that the data is highly skewed. For building a model for such skewed data, the minority class is upsampled by sampling from the existing data repetitively with repetition until the majority class and minority class has equal samples.

This is done using the package sklearn.utils.resample.

5.3 Base Model Selection from various models

The upsampled data is then used to build the following models:

- Linear regression
- Knn
- Support Vector Machine (SVM)
- Decision Tree
- Random Forest
- Extreme Gradient Boosting
- Gradient Boosting Classifier

The data is split into training and test data in the ratio of 4:1. The data is then scaled using the StandardScalar. Each model is trained using the training data and the model is then used to predict using the test data.

5.4 Test Metric

There are three test metrics computed on each of the models built.

- Accuracy score
- Recall
- Mathew's correlation coefficient (MCC)

Mathew's correlation coefficient uses all the four components of a confusion matrix. It returns a value between (-1, 1) with -1 being a poorly fitted model and 1 being the rightly fitted model.

The entire data is split into training and testing datasets in the ratio of 4:1. The following models are trained using the training dataset and various test metrics are computed using the test dataset. The model is selected based majorly on MCC.

The MCC scores obtained after running various models on test data is as follows:

Models	MCC
Random Forest Classifier	0.941119
Decision Tree Classifier	0.927532
K-Near Neighbors	0.863852
Support Vector Machine	0.732137
Gradient Boosting	0.722849
XGBoost	0.718817
Logistic Model	0.607161

5.5 The Model

The Random Forest classifier performs much better than all the other models.

Training Accuracy score : 100.0
Confusion matrix :
[31995 0]
[0 31880]
Recall - train : 1.0
MCC : 1.0

Testing Accuracy score : 97.0
Confusion matrix :
[7450 477]
[6 8036]
Recall - test : 0.9992
MCC : 0.9411

5.5.1 Hyper-parameter tuning

Hyper-Parameter is done on the base model selected to make sure that the model works well with the available data. Feature reduction is also done here.
From this step, it is understood that out of the 16 available features, it is best to use 6 features for this model.

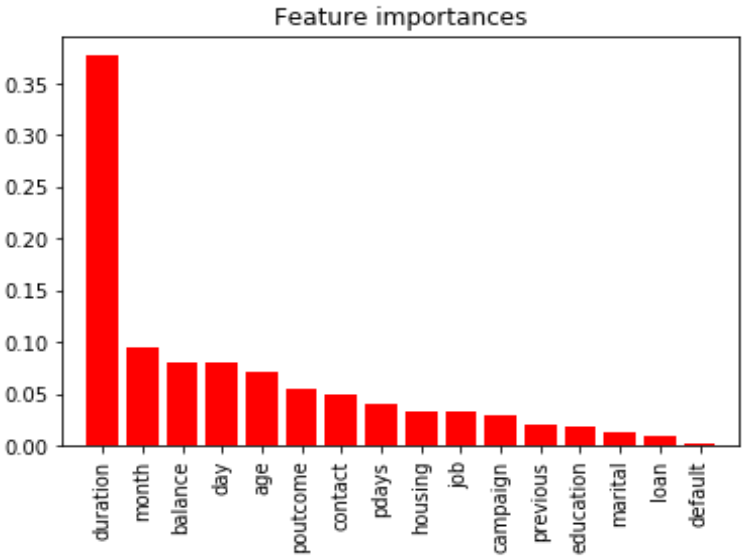
The test metric is computed again after hyperparameter tuning.

Training Accuracy score : 100.0
Confusion matrix :
[31995 0]
[0 31880]
Recall - train : 1.0
MCC : 1.0

Testing Accuracy score : 97.0
Confusion matrix :
[7534 393]
[11 8031]
Recall - test : 0.9985
MCC : 0.9574

5.5.2 Feature Importance

The features used in training the model have different importance associated. Few features are more important than the others.



The duration of the call made to the clients is very important in determining whether a client will subscribe to a term deposit or not. This is the most important feature of the entire data. This was discussed earlier in the data storytelling section of this project.

5.6 Predicting using unseen data

The new data that is imported is again treated into categorical columns as the test data was treated earlier. StandardScalar is also applied to the new data to make the new data similar to the training data.

The Random forest classifier performs very well in predicting new data. Mathew's correlation coefficient is 0.95, which is as good as it works on test data.

```
Accuracy score :  0.9891
Confusion matrix :
 [3952       48]
 [   1       520]
MCC :  0.9500
```

5.7 Conclusion from Machine Learning

The Random Forest classifier has the highest performance with Mathew's correlation coefficient of 0.957 achieved with test data. Hyperparameter tuning of the model increased the MCC from 0.9411 to 0.9574. The best number of features to be used (i.e., 6 features) is also determined using the hyperparameter tuning. This model also has very low false negatives compared to other models.

6. Conclusion

The data was acquired from the marketing campaign of a Portuguese bank by calling its customers to market their product (Term deposit).

The duration of the call played a very important role in determining whether a customer subscribed for a term deposit or not. If the customer was engaged in the call long enough to understand the merits of the term deposit, the customer has a higher chance of subscribing to a term deposit.

Most customers were contacted in May. Interestingly, Most clients who were contacted in May did not subscribe to a term deposit. This can be avoided in the future.

During the start of the campaign, it is better to target the customers with whom there was a success in the previous campaign. They are more likely to continue since they already know the merits and demerits.

Targeting old people is better because they would want a relatively safe investment and a steady income. It can also be seen from the exploratory data analysis that old people (people aged above 60) are more likely to subscribe to a term deposit compared to others.

The model has very low false negatives. It is a good sign since not many people will be missed by the marketing campaign.