ANALYSIS OF POTENTIAL CUSTOMERS FOR TERM DEPOSIT IN PORTUGESE BANKING INSTITUTION

(CAPSTONE PROJECT - GROUP 3)

Bona fide record of work done by

Gowtham Raj M
Harsha Vardhan S
Kavin R
Santhosh M
Vignesh V

Mentored by

Ms Anjana Agrawal

POST GRADUATE PROGRAM IN
DATA SCIENCE AND ENGINEERING

September 2020

Great Lakes Institute of Management
Great Learning Institute

CHENNAI – 600 096

Table of Contents

ACKNOWLEDGEMENT

A project work is a job of great enormity and it cannot be accomplished by an individual without the help of experts. This work is a result of selfless help of many individuals. Our heartfelt thanks to our guide Ms. AnjanaAgarwal for her inspiring guidance and support given to us throughout the course of this work. We would like to express our gratitude to Mr.Mahesh Anand, Mr.YL Prasad, Mr.Regi Mathew in helping us understand the concepts of Machine Learning, Mr.Chandrasekhar who lead us through the path of Statistics, Mr.Aniruddha Kalabhande who taught us the ways of Exploring and infering Data, Mr.Raghuraman helping us presnt the inferences in an appealing way through Tableau, Mr.Sachin and Ms.Vaidehi Talikar for setting our foundation strong in Python, Ms.Meena helping us perceive the art of extracting required data from the database. Our special thanks to the members of the committee, all the faculty and supporting staff of the Great Learnings for assisting us in this project. We are much indebted to all those who are contributing directly or indirectly to make this project a success.

ABSTRACT

Marketing is the channel through which a product is exposed to a potential customer who can buy or subscribe. Over the time marketing is used to persuade the customer to buy the product through multiple approaches. One type of such marketing is Direct Marketing. Direct marketing is the process through which the company gets in direct contact to the customers, either via face to face meeting or through a phone call to get them introduced to the product and persuade them to buy it. Of these two types of direct marketing techniques the financial and banking institutions prefers the phone call method as the customer base is high and also the cost involved in this type is relatively less.

The major attributes which helps in optimizing the cost of marketing via phone call is the amount of time spent with the potential customers. But it is very less to impossible to find a potential customer via phone calls and most of the time the salespeople tend to spend more time with the customers who might not interested in the product. This will add up to the operational cost and will have a direct impact in the profit. This is where the data driven approach can be used to predict the potential customer, so that most of the time can be concentrated on these customers.

SYNOPSIS

The data taken for our project is one such dataset that belongs to direct marketing through phone calls where the main objective is to classify the potential customers who can subscribe to the term deposit. This belongs to a Portuguese bank dated from May 2008 to November 2010. The main purpose of this project is to create a predictive machine learning model, using various supervised classification models and identifying the best model that can classify the potential customers from the dataset who might be able to subscribe to the term deposit based on the given features. This will in turn help to find the features which have higher impact on the outcome.

Objectives:

1. To propose a predictive model that can assist in finding the potential customer who can subscribe the term deposit with the bank
2. To analyze based on the customer profile, previous marketing campaigns and the socio-economic status of the country.

CHAPTER – 1

Introduction

1.1 Description of the Data Set

The features of the dataset can be broadly classified into three categories:

1. Basic Information about the customers:

    - Age : Age of the customer contacted
    - Job : The job nature of the customer contacted
    - Education : Major Qualification of the customer
    - Default : Status of any default connected with the customer in the past
    - Housing : Whether the customer has a Housing Loan
    - Loan : Whether the customer has a Personal Loan

2. Features related to the Marketing Campaigns:

    - Contact : Type of the device used by the customer while contacting (Cellular / Telephone)
    - Month : Month on which the Marketing call is made
    - Day of Week : In which day the customer was contacted
    - Duration : Last contacted duration in seconds
    - Campaign : Number of contacts performed to the particular customer during this campaign
    - Pdays : Number of Days passed by after the customer has been contacted for a previous campaign
    - Previous : Number of contacts performed before this campaign, for this customer

3. Socio – Economic context:

    - Emp.var.rate : Employment Variation Rate – Quarterly Indicator
    - Cons.price.idx : Consumer Price Index – Monthly Indicator
    - Cons.conf.idx : Consumer Confidence Index – Monthly Indicator
    - Euribor3m : Euribor 3 month rate – Daily Indicator
    - Nr.Employed : Number of Employees – Quarterly Indicator

Above mentioned features are classified into Categorical and Continuous as follows:

Categorical : Job, Education, Default, Housing, Loan, Contact, Month, Day of week, Pdays, Previous, Emp.var.rate, Nr.Employed

Continuous : Age, Duration, Campaign, Pdays, Cons.Price.Idx, Cons.Conf.Idx, Euribor3m

1.2 Project Justification:

As mentioned in the synopsis, the direct marketing through call is an important aspect in terms of selling their products to the new customers, which is eventually new customer acquisition. The major part of the profit acquired by the bank is cross selling their products to the existing customer. In all these marketing activities direct marketing plays the vital role.

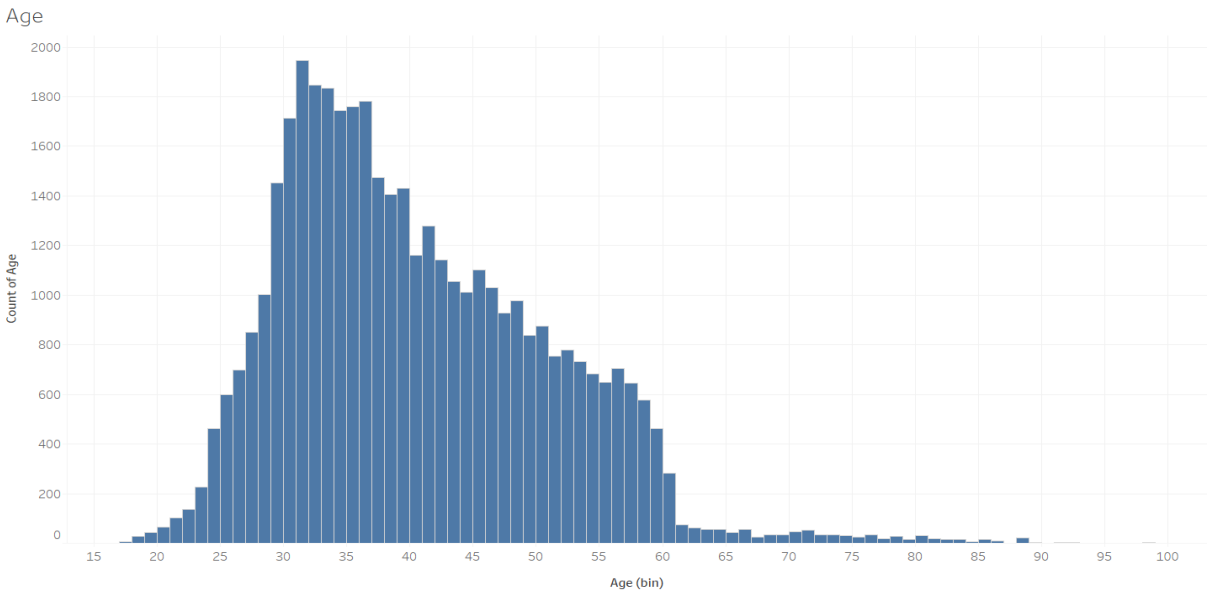
The major unit economics involved in the direct marketing through call is the time spent by the sales person with the potential customers. If the sales person spends more time with customer without any potential, then it will have a direct impact on the conversion metrics of the particular marketing campaign. Apart from this, channelizing the time spent by the sales person on the potential customer will boost the morale of the sales person as this is one of the most stressful jobs out there in the market and the sales person are prone to stress easily, within few calls.

The outcome of the project is to classify the customers who have potential to subscribe for the term deposit of the bank. With this outcome the sales person can be directed to give extra attention and effort towards the customers who are more potential. And the basic nature of direct marketing through call is that the possibility of customer getting converted is very less in the first attempt and the sales person has to persuade the customer for a certain period time. This persuasion time should optimal, as the time duration is less, the chances of customer getting converted will be less and on the other hand if the time duration extends than the optimal duration, it will result in the customer getting irritated and eventually result in losing the customer.

CHAPTER - 2
Exploratory Data Analysis

2.1 Univariate Analysis:

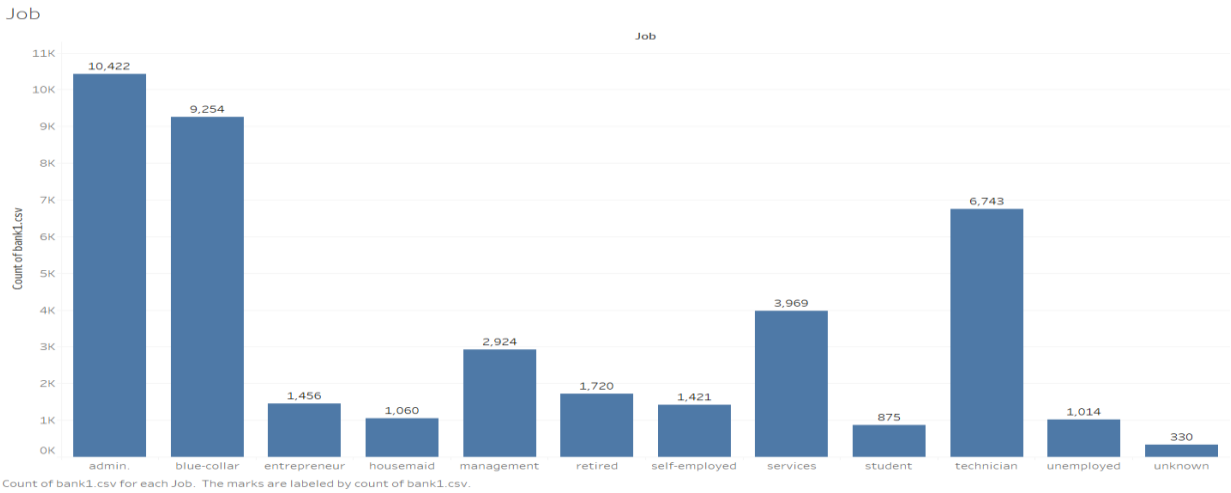
Age:



The trend of count of Age for Age (bin).

The above plot is the histogram plot of the age feature. From that we can see that the distribution is not normal. The majority of the customers considered for this marketing campaign is between the age criteria of 24 and 60.

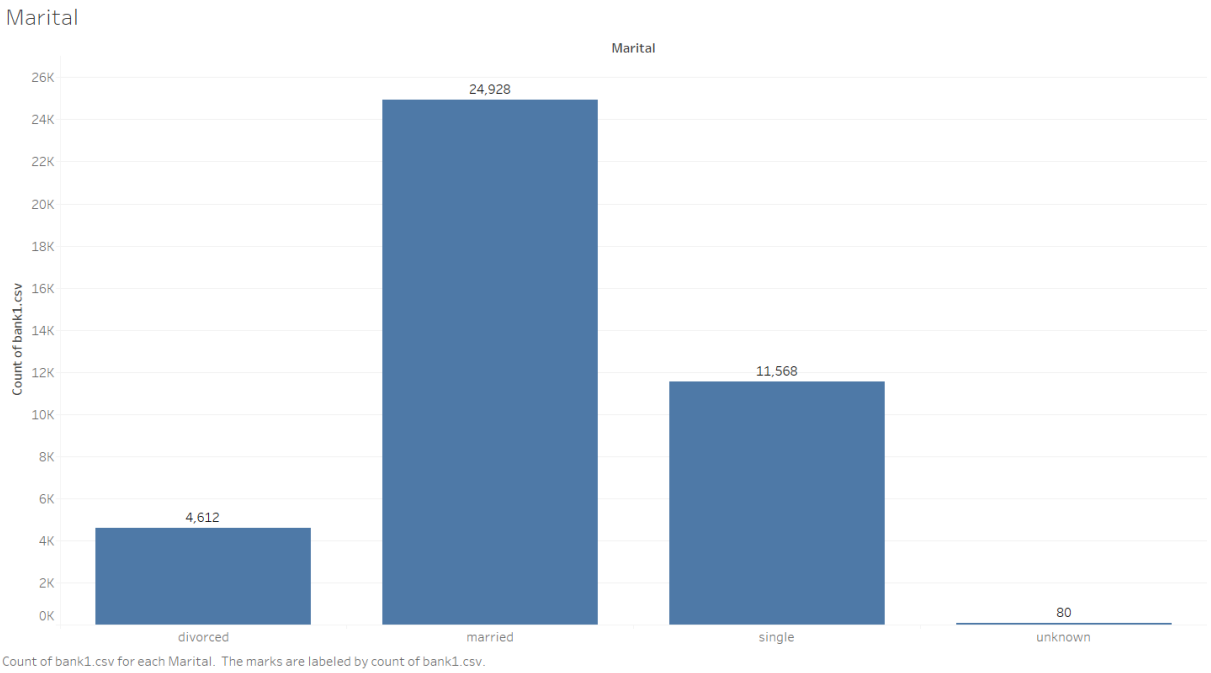
Job:



Count of bank1.csv for each Job. The marks are labeled by count of bank1.csv.

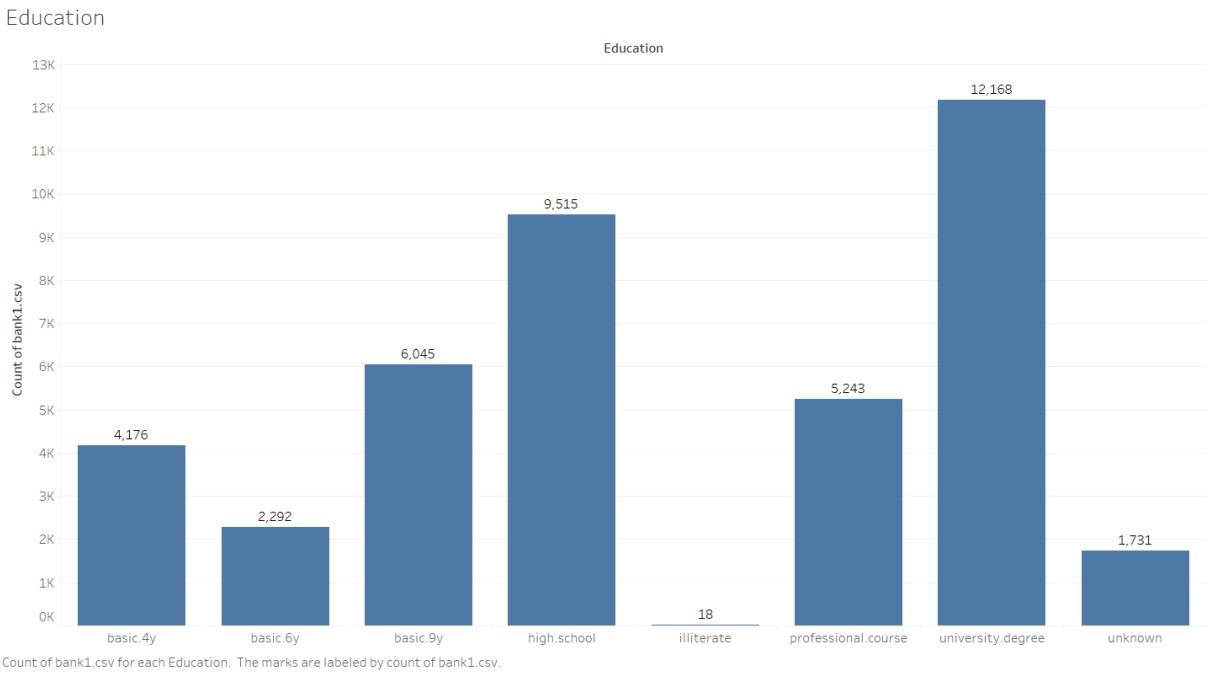
The above is the bar plot of the Job feature. Admin, Blue-Collar and Technician are the top three job types of the customers involved in this campaign.

Marital:



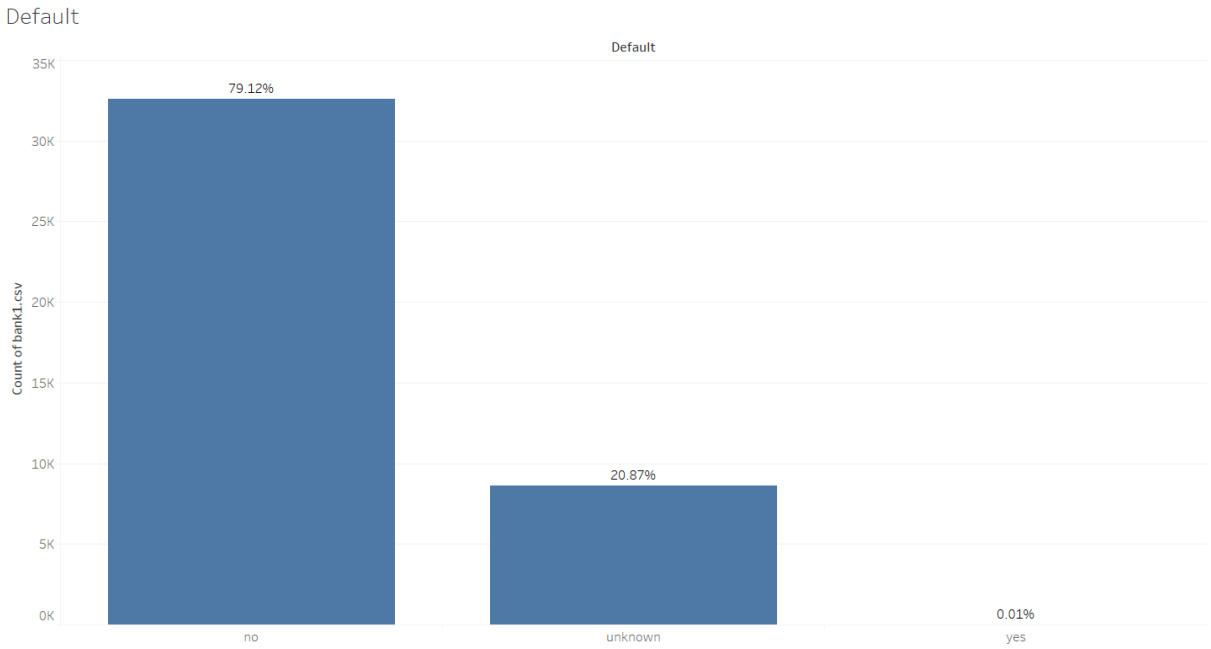
From the above bar plot we can infer that the majority of the customers were married, followed by single and divorced.

Education:



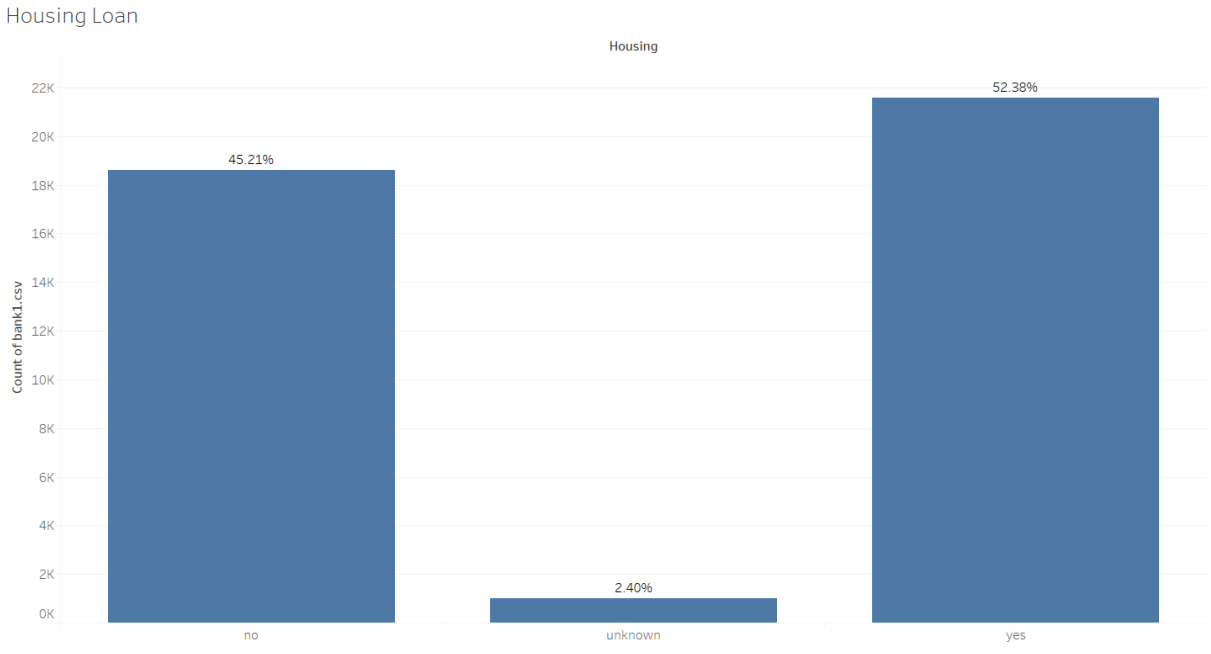
From the above plot we can infer that almost all of the customer involved in this marketing campaign has been acquired basic education as the count of illiterate is very low compared to the other categories.

Default:



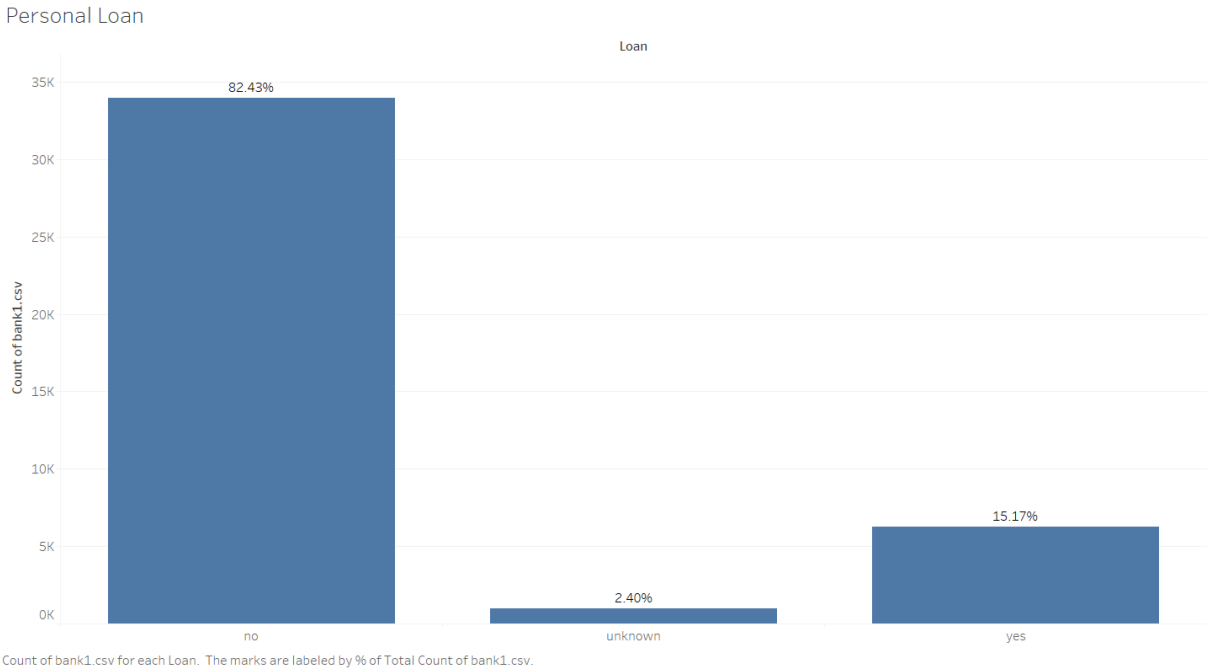
From the above plot we can infer that about 79% of the customers does not have any default history associated with them.

Housing:



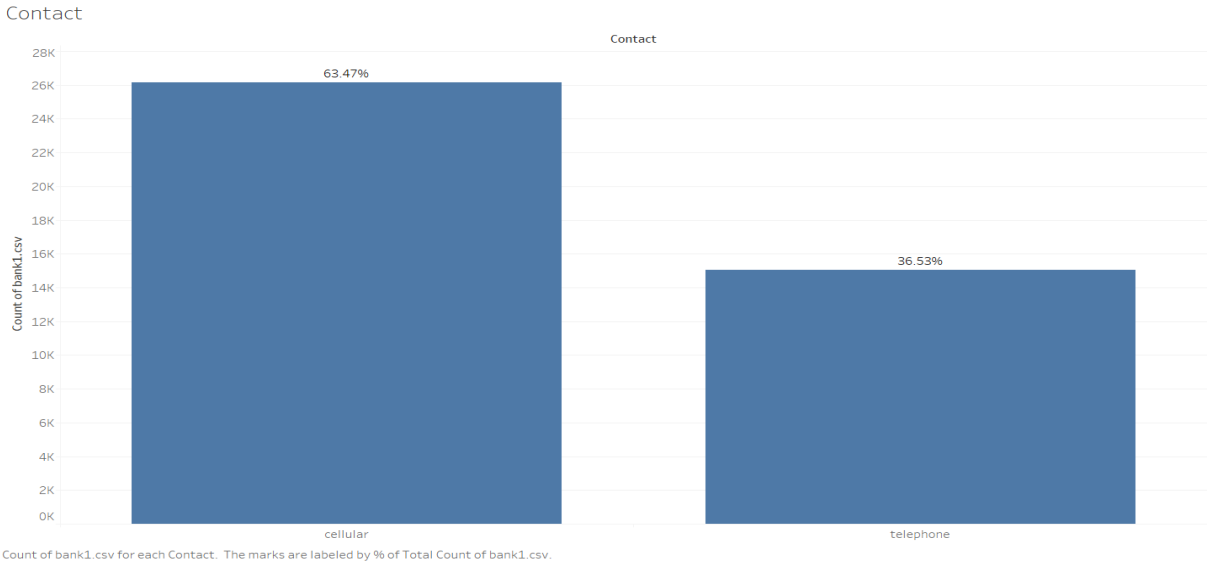
The percentage of customers who have availed a housing loan already is just above the 50% mark and around 45% of customers don't have housing loan history. About 2.5% of the customers housing loan history are unknown.

Loan:



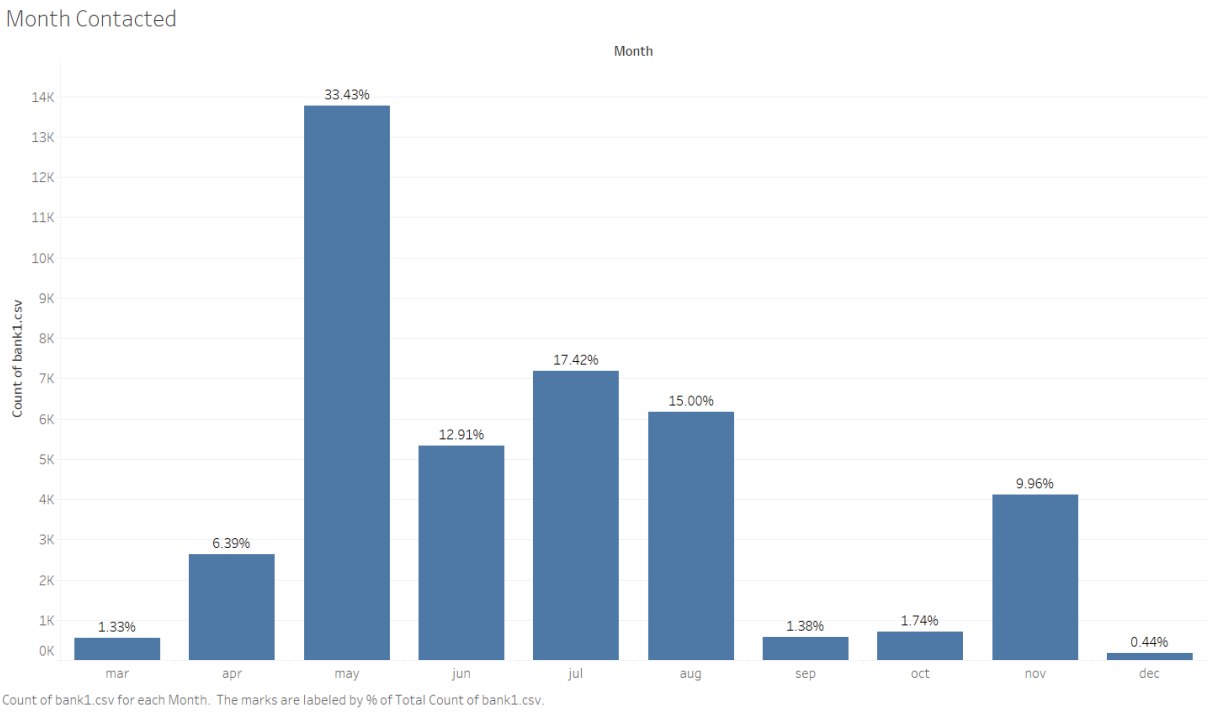
The number of customers who doesn't have any personal loan is just above 82%. Therefore, majority of the customers in this marketing campaign doesn't have a personal loan.

Contact:



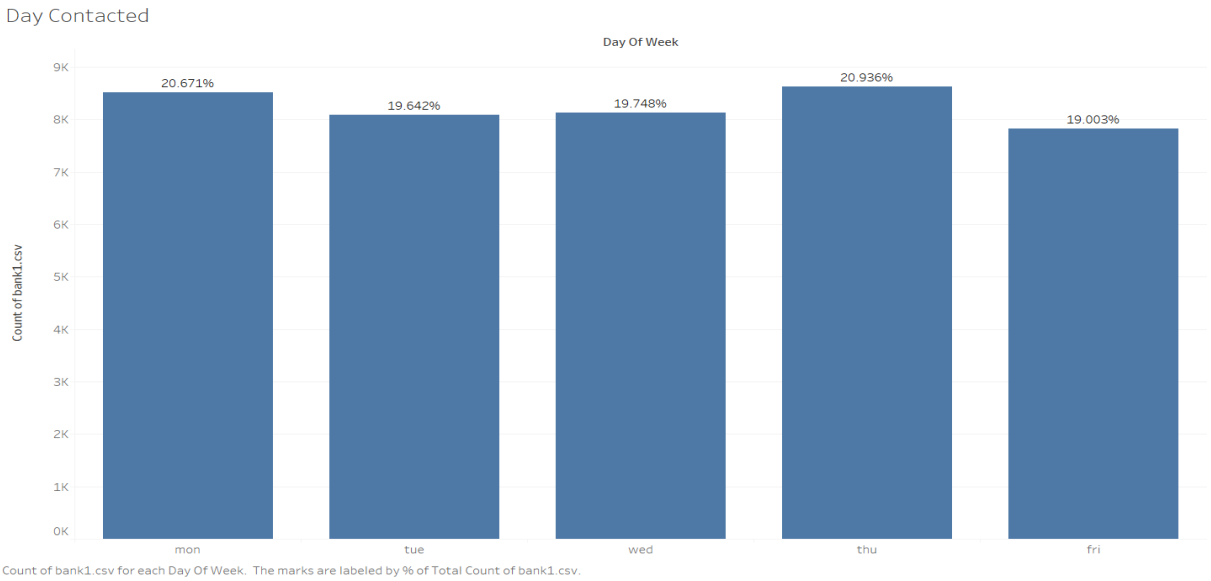
About 63% of the customer who have been contacted have been using Cellular phones for this campaign. The remaining have used Telephone.

Month:



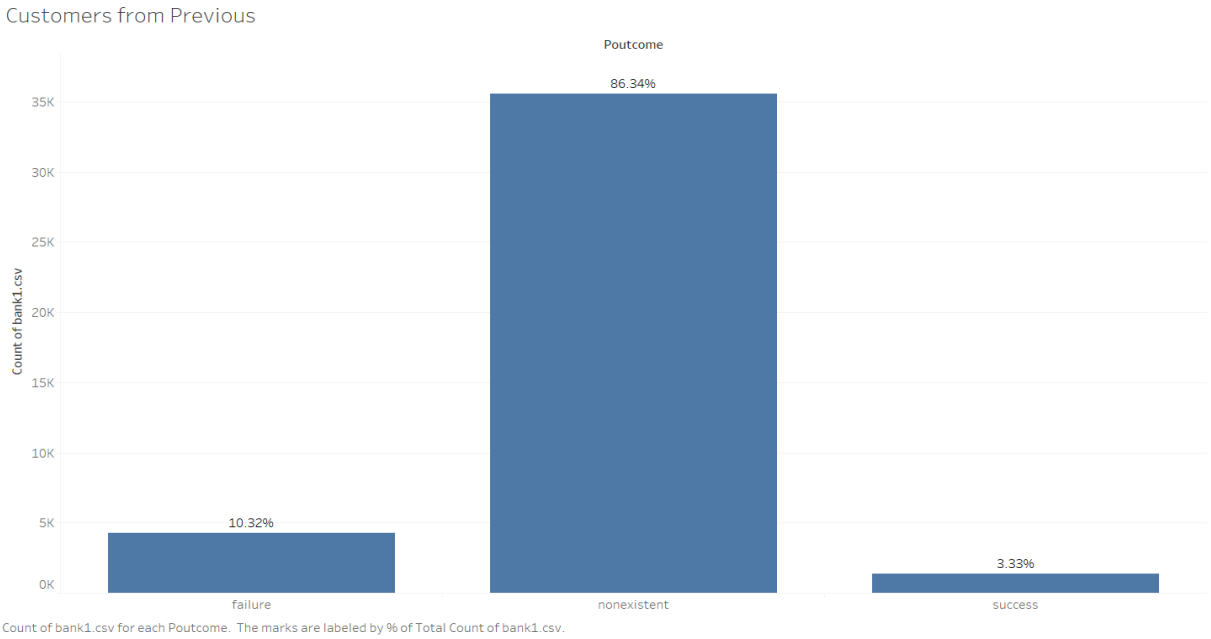
Almost 33% of the total customers for this marketing campaign have been contacted in the month of May, followed by July which has 17% and then by August which accounts to 15%. December has very lowest number of customers being contacted, whose percentage stands at 0.44% of the total customers.

Day of the week:



The customers were not contacted during the weekends. And during the weekdays the percentage of the total customers being contacted was almost equal among all days.

Poutcome:

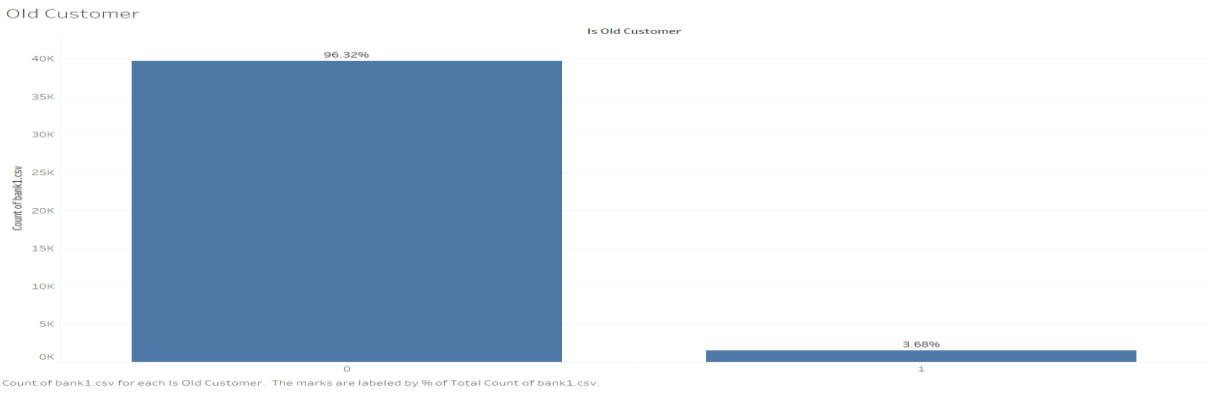


Around 86% of the total customer contacted for this campaign has been new customers who haven't existed in the previous campaign. About 10.32% customers were the previous customers who were not successful with the previous campaign and about 3.33% were the customers who were successful with the previous campaign.

Pdays / is_old_customer:

The feature Pdays consists of data in which if the customer is denoted as '999' the customer has not been contacted previously and then if the customer has been contacted previously, then the number of days passed since the customer has been contacted.

For better understanding of this feature we have converted this into another feature called 'is_old_customer' where if the data is '0' then they are new customer and if the data denoted '1' then they are old customer.



From the above plot we can infer that about 96% of the customers who were involved in this campaign were not involved in the previous campaign and they have not been contacted by the bank for the previous campaign.

Price.Idx.Range:

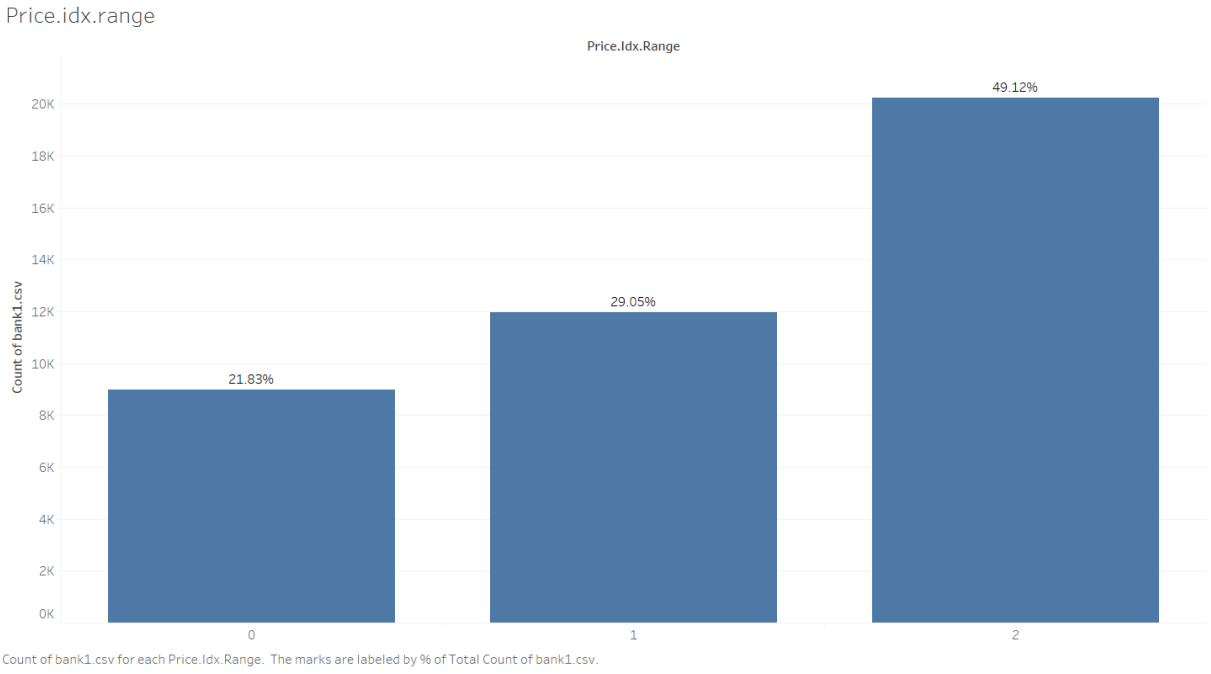
The feature Cons.Price.Idx has been converted into Price.Idx.Range to obtain a clear inference from the feature.

The feature Price.Idx.Range has been derived as follows:

If the Cons.Price.Range is between 92.198 and 93.056, then the Price.Idx.Range is 0,

If the Cons.Price.Range is between 93.056and 93.912, then the Price.Idx.Range is 1,

If the Cons.Price.Range is greater than 93.056, then the Price.Idx.Range is 2.



About 49% of the customers contacted falls in the 2 category. The 0 and 1 category takes 21.83% and 29.05%, respectively.

Conf.Idx.Range:

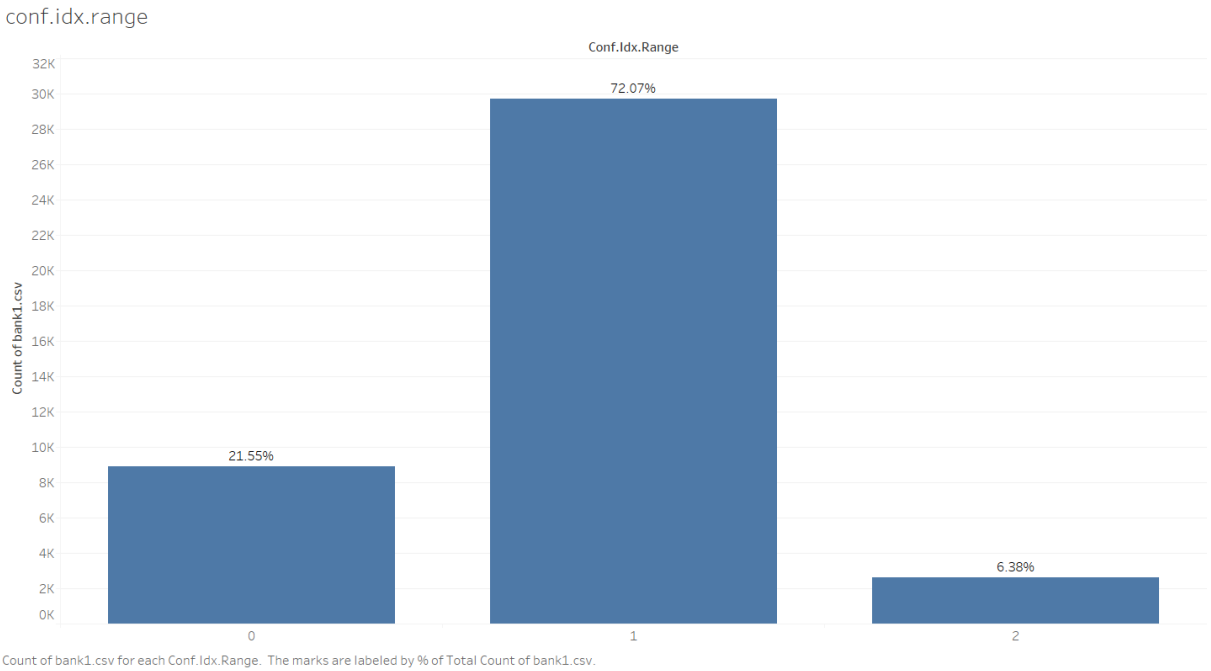
The feature Cons.Conf.Idx has been converted into Conf.Idx.Range to obtain a clear inference from the feature.

The feature Conf.Idx.Range has been derived as follows:

If the Cons.Conf.Range is between -50.824 and -42.833, then the Conf.Idx.Range is 0,

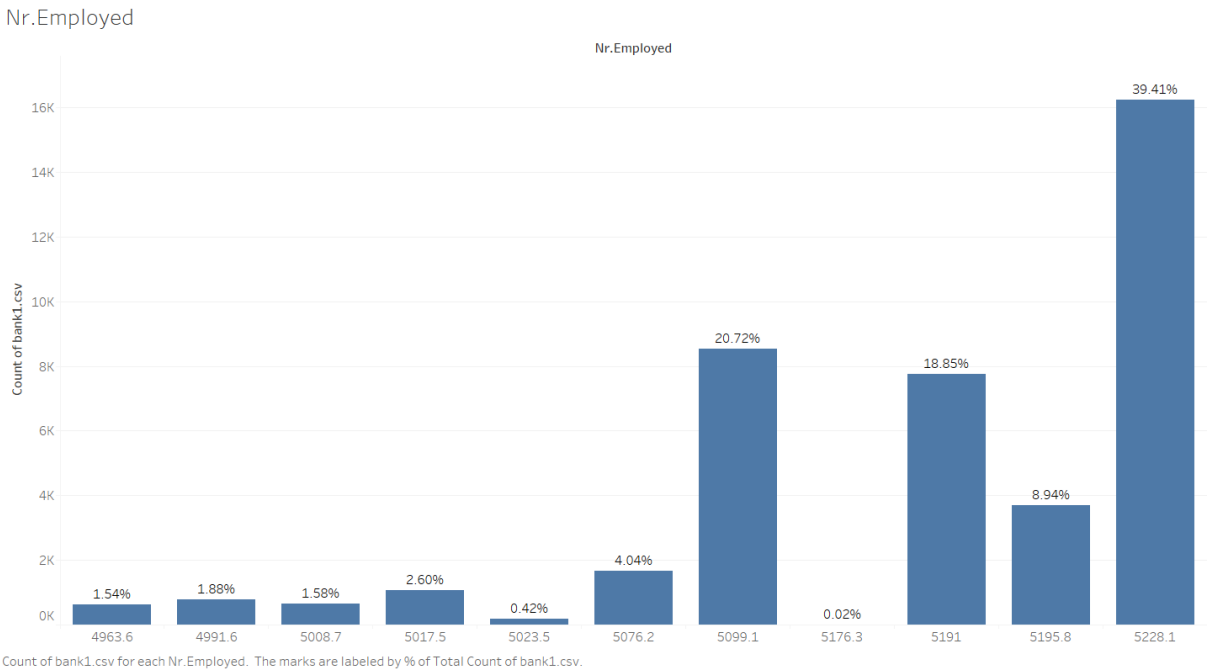
If the Cons.Price.Range is between -42.833 and -34.867, then the Conf.Idx.Range is 1,

If the Cons.Price.Range is greater than -34.867, then the Conf.Idx.Range is 2.



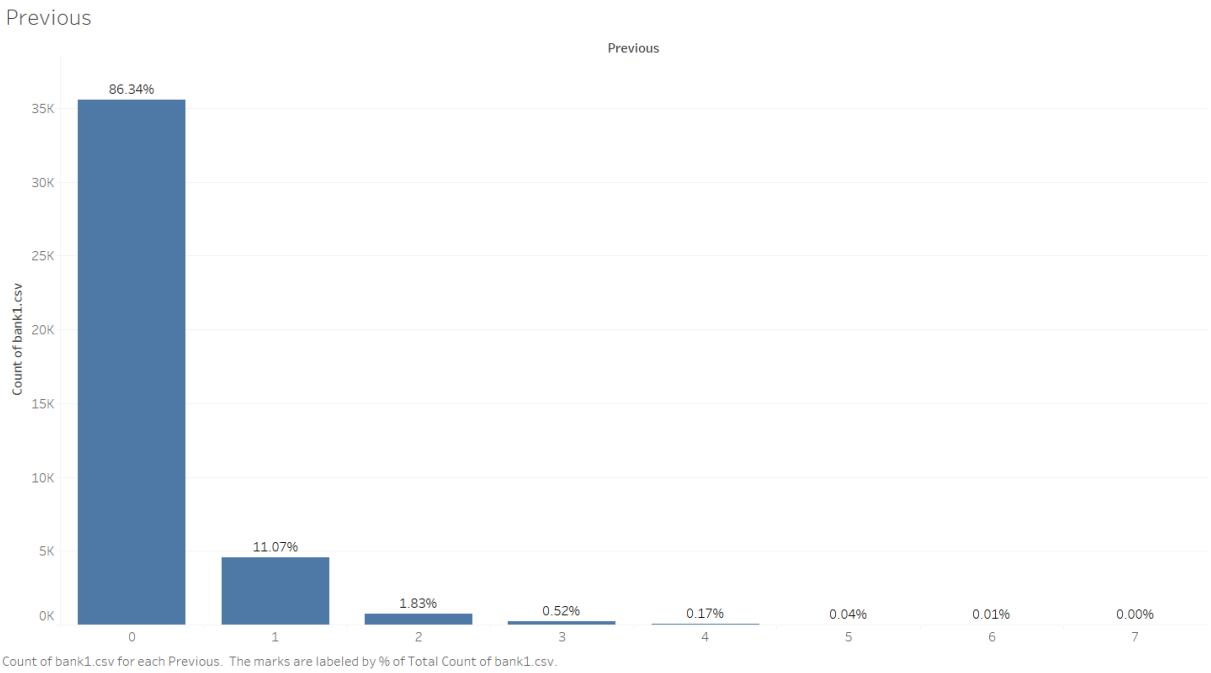
About 72% of the customers contacted falls in the 1 category. The 0 and 2 category takes 21.55% and 6.38%, respectively.

Nr.Employed:



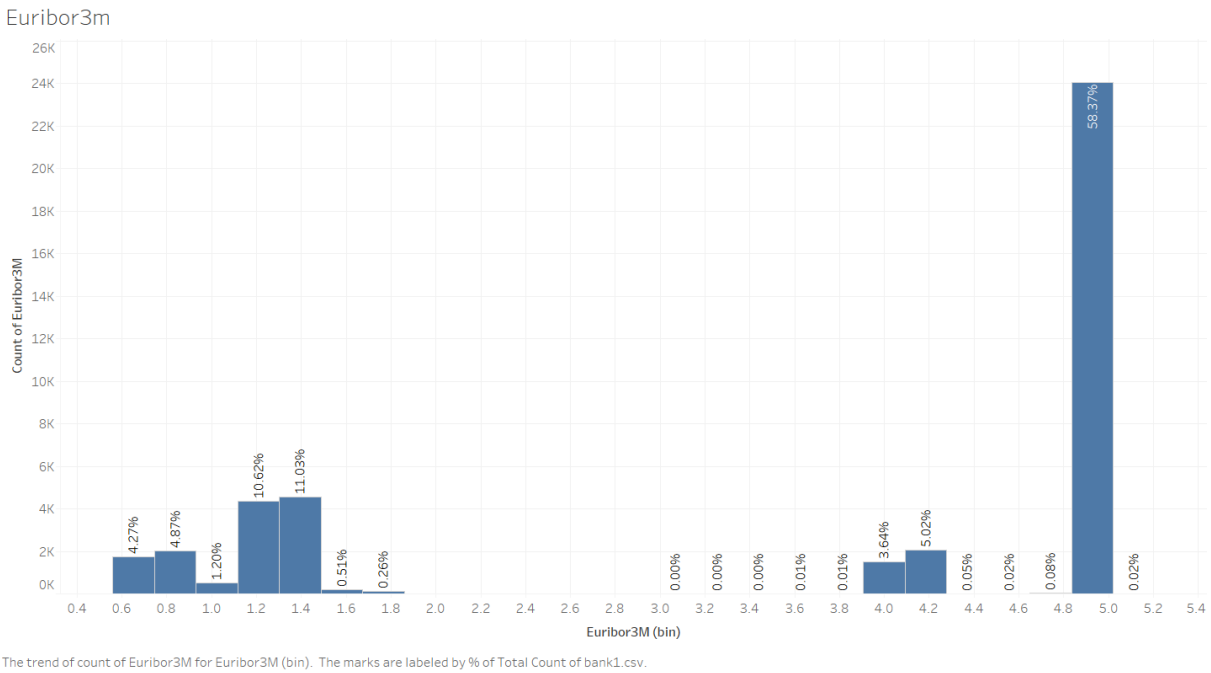
The number of customers contacted is higher towards the higher Nr.Employed and about 39% of the customers are contacted when the Nr.Employed is at 5228.1

Previous:



About 86% of the customers were not even contacted single time before this campaign. They were the new customers for the bank.

Euribor3m:



Around 58% of the customers were contacted when the Euribor percentage stood at 5.

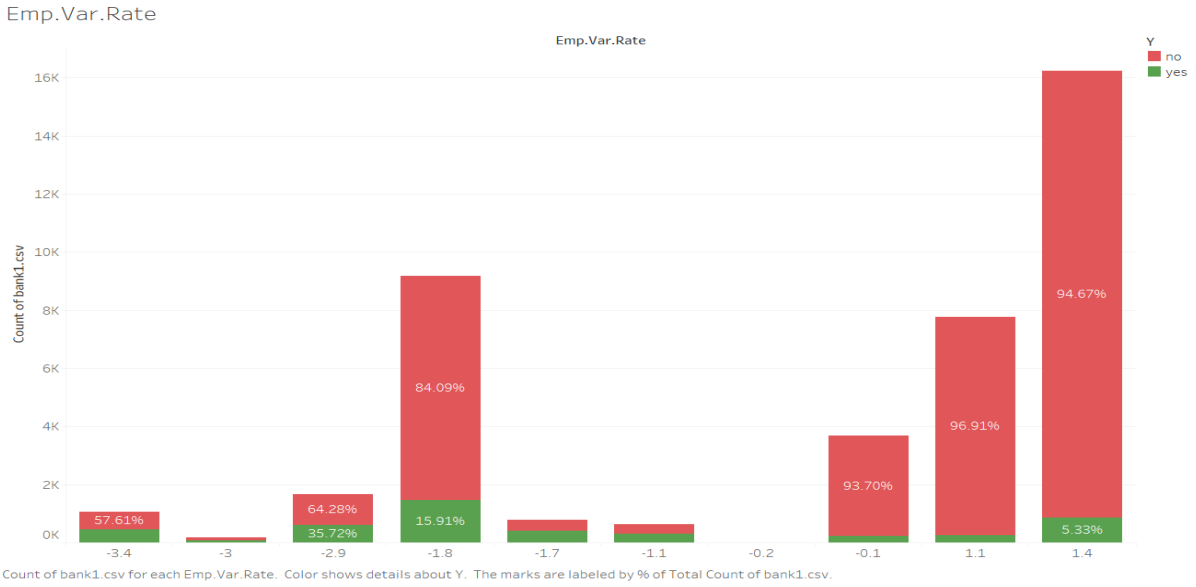
Based on the Univariate analysis, some of the features have been converted into Categorical. Below is the revised list of Categorical and Continuous features,

Categorical : Job, Education, Default, Housing, Loan, Contact, Month, Day of week, Previous, Pdays, Emp.var.rate, Nr.Employed, is_old_customer, Price.Idx.range, Conf.Idx.range

Continuous : Age, Duration, Campaign, Euribor3m

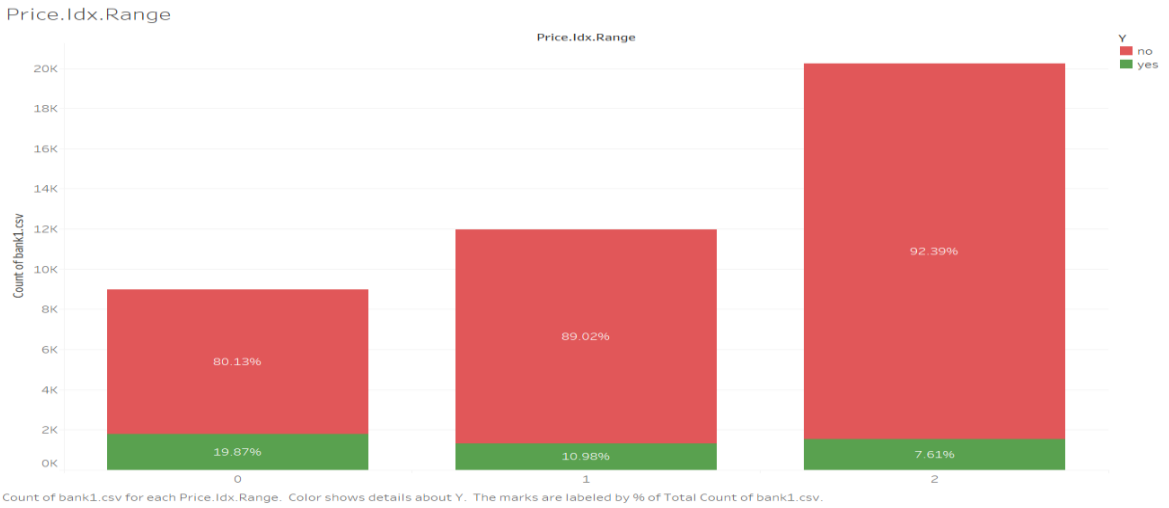
2.2 Bivariate / Multivariate Analysis:

Emp.Var.Rate vs Y:



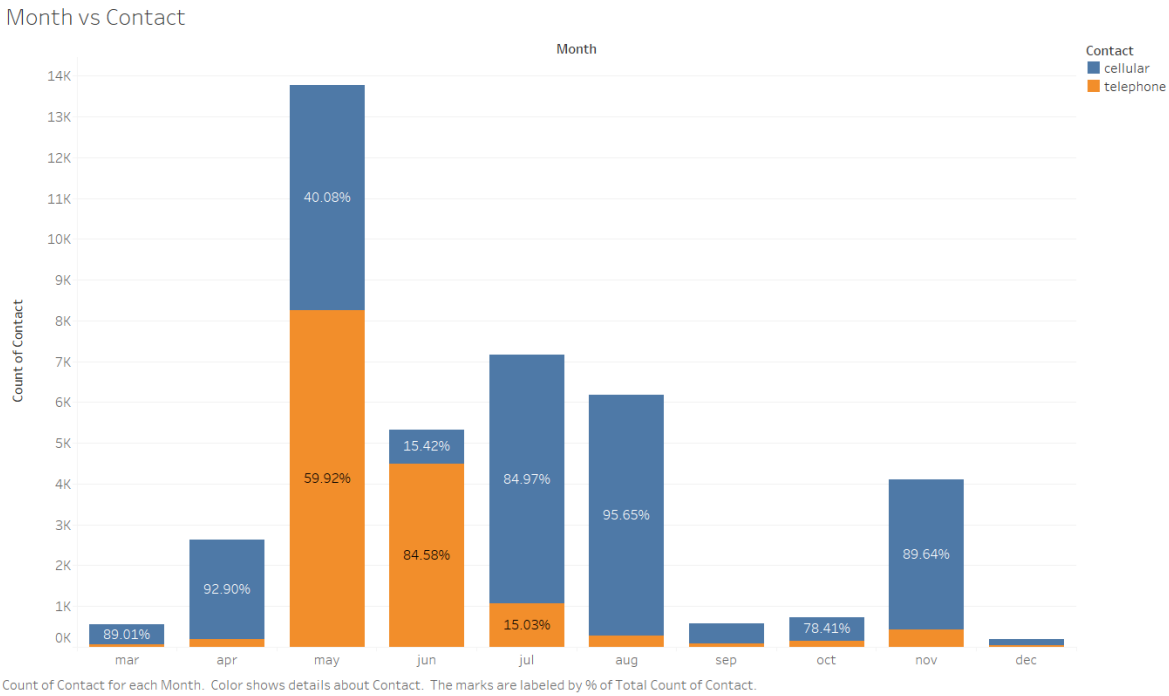
- The above plot is the comparison of the emp.var.rate feature against the target feature y. From the plot we can infer that majority of the customers has been contacted when the emp.var.rate is at 1.4.
- In the plot the red bar denotes the number customer who said no and the green bar denotes the number of customer who said yes. We can see that the ratio of customer, who has said yes, tends to be higher at the low emp.var.rate when we compare the same with that of the higher emp.var.rate.

Price.Idx.Range vs Y :

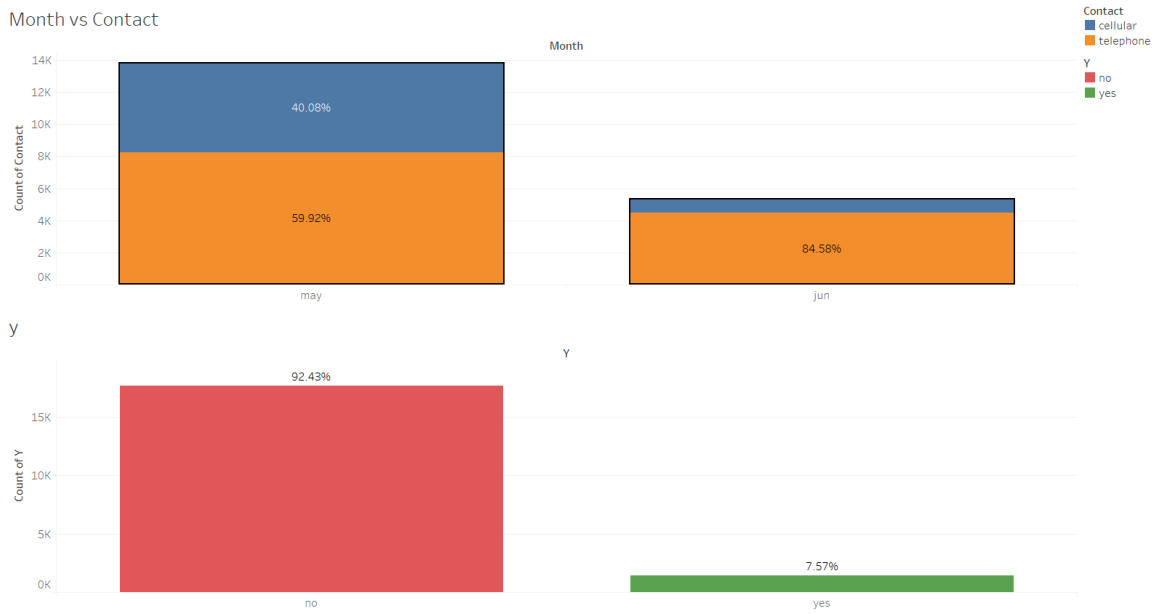


As we see in the above plot as the Price.Idx.Range increases the percentage of customer who say yes tends to decrease.

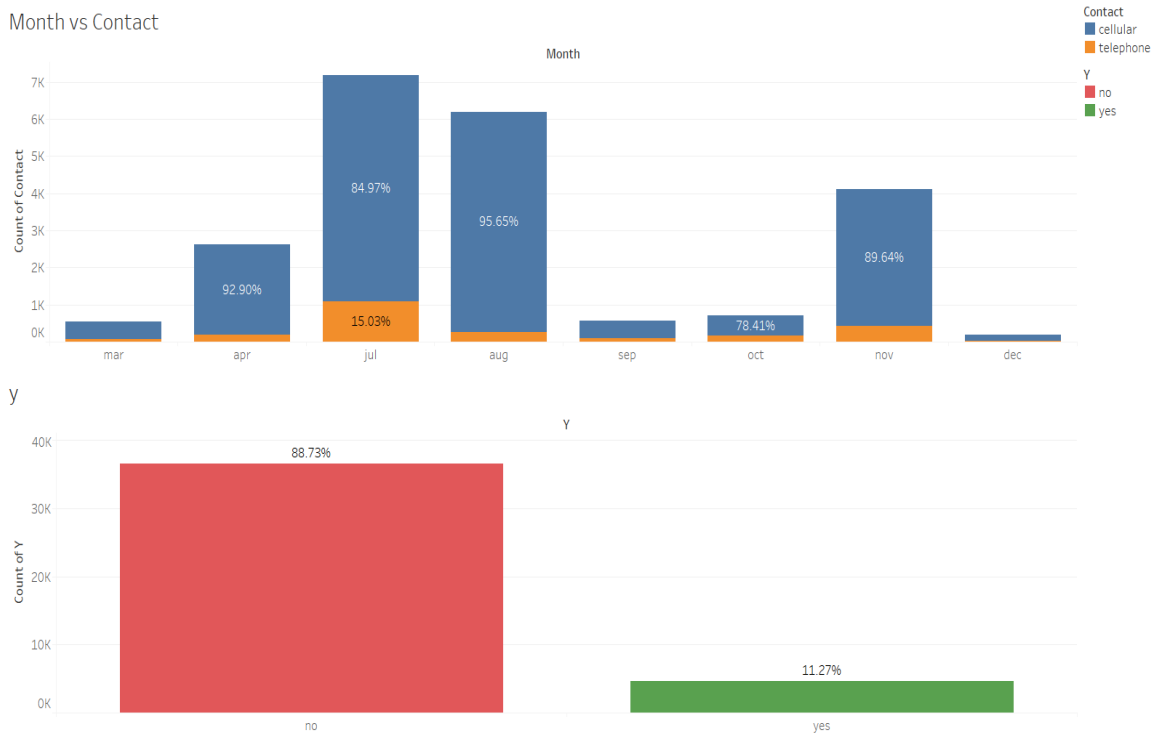
Analysis of Contact Feature:



From the above plot we can infer that contrary to the trend, the month of May and June has Telephone as the higher percentage of the medium of communication used by the customer. All other month have Cellular as the higher percentage.

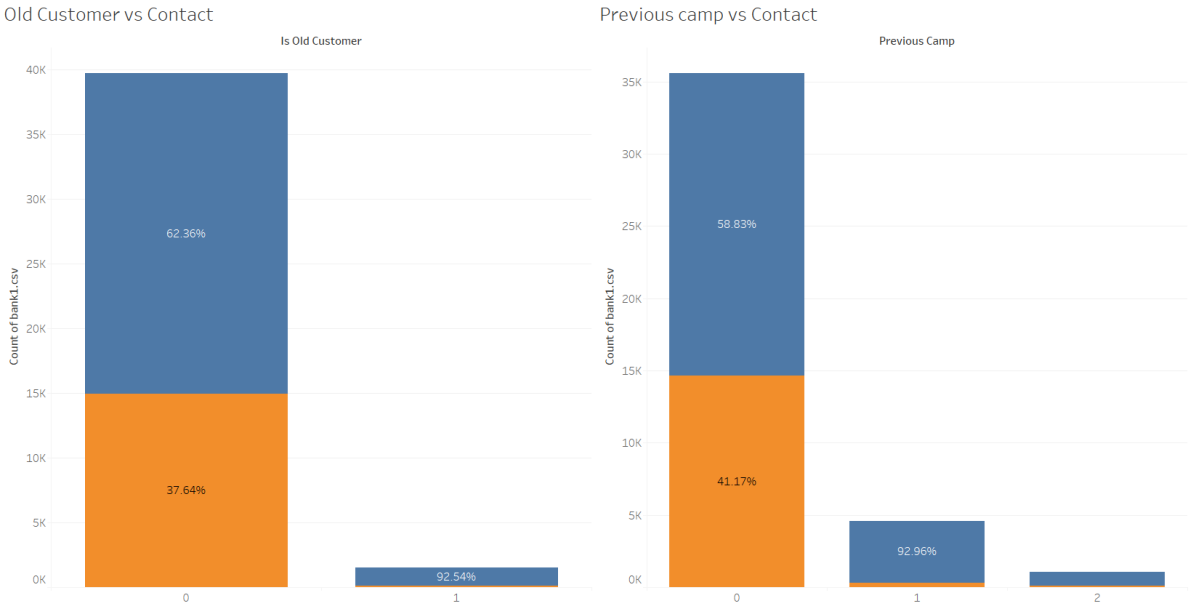


Therefore when we take the May and June alone where the percentage of customers with Telephone being contacted the percentage of customers saying yes stands at 7.57%



Whereas when we consider all other months where the percentage of customers with the Cellular is contacted is high, the percentage of Customers saying yes stands at 11.27% which is almost 4% increase compared to the above.

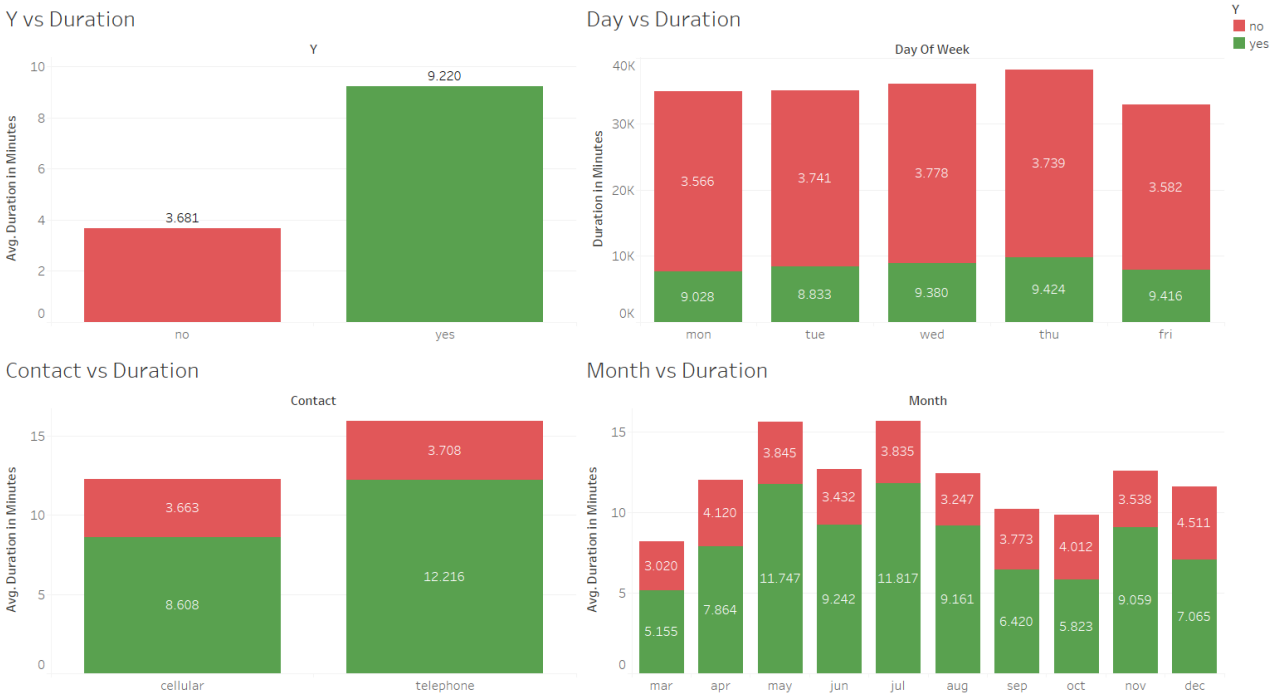
Therefore it is suggested that if the number of customers with Cellular is contacted in the months of May and June, we can improve the conversion percentage in those months.



Further from the above plot we can infer that as the customer with Cellular phones tends to stick with the campaign or the company as the percentage of customers with telephone drops drastically. And also, if we see the old customers almost 92% of the old customer are the one with the Cellular phone.

Therefore it is suggested that for improving the conversion percentage and also to have the customer to be in touch with the bank for the longer period of time it is recommended to concentrate on the customers who are with the Cellular phones.

Analysis of Duration feature:



The above plot is the comparison of the Duration feature with various other features.

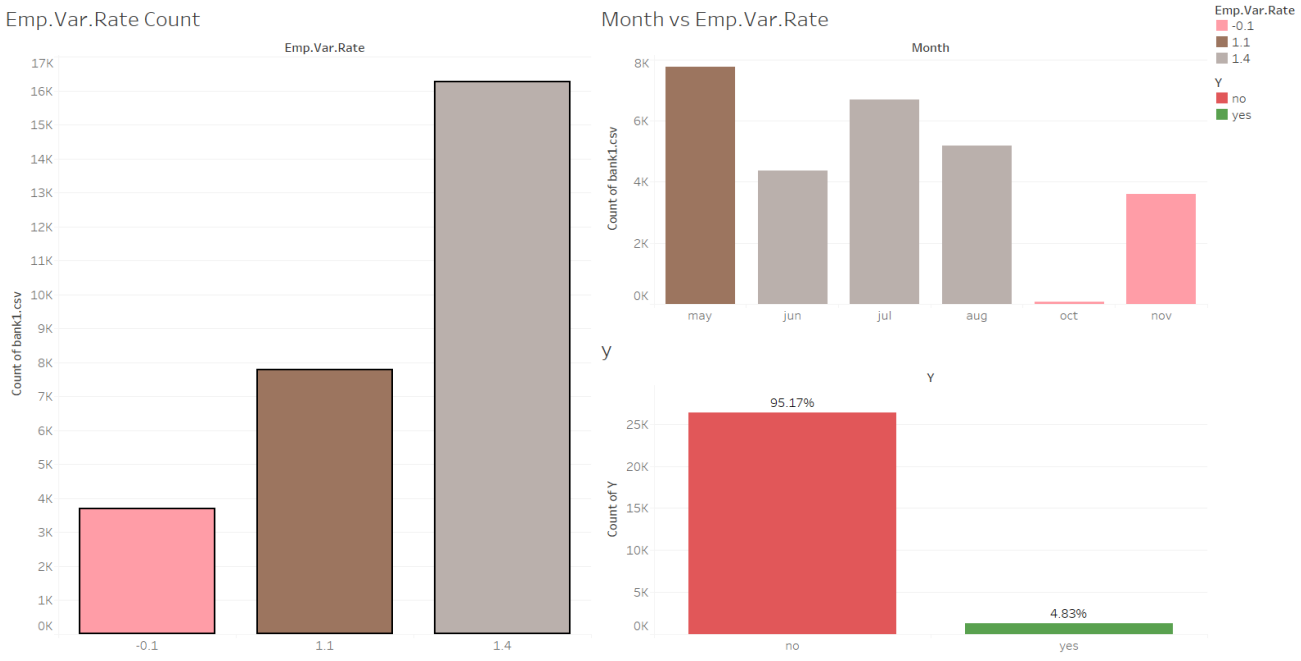
First plot in the top left is the plot between the target variable and the average duration in minutes. We can see that the average duration of the customer who says no was at 3.6 minutes. Whereas, the average duration for the customer who says yes, was at, 9.2 minutes. Therefore the average duration of engagement is higher for the customer who says yes compared to the customer who say no.

The second plot in the top right compares the duration feature with that of the day of the week in which the customer being contacted. There is not much variation between the days as the average time duration for the customer who says no remains between 3.5 to 3.7 minutes and the average duration of customer who says yes remains between 8.8 to 9.4 minutes

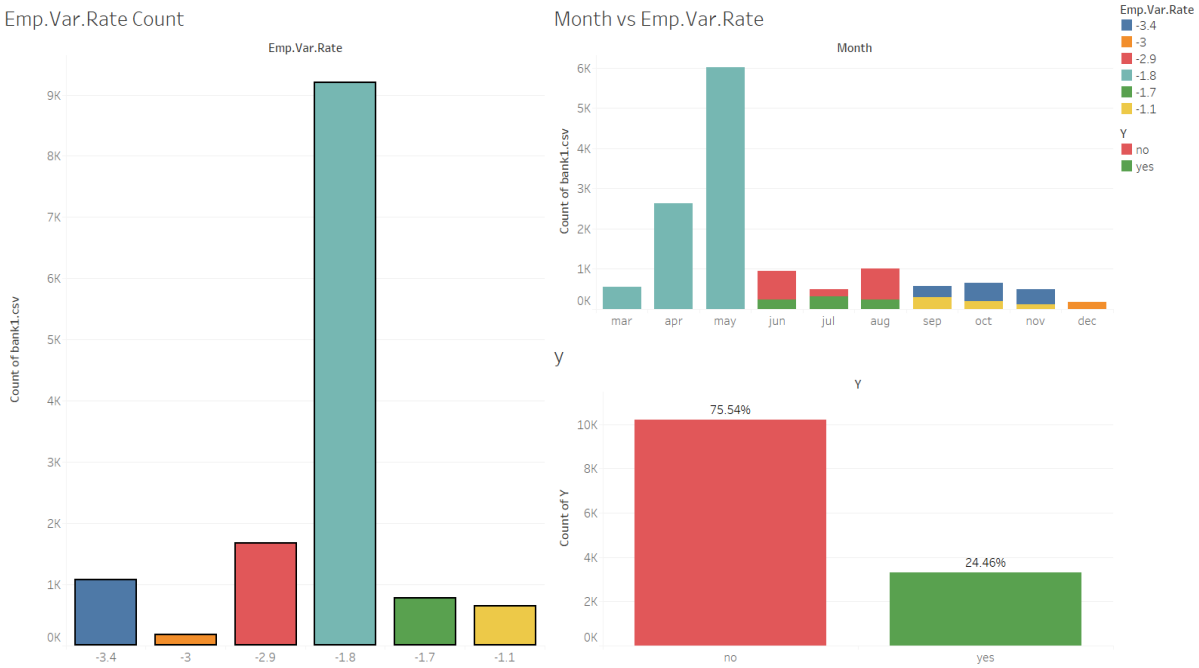
The third plot in the bottom left compares the contact feature with that of the average duration. The average duration a customer who says yes for the customer having cellular stands at 8.6 minutes compared to that of 12.2 minutes for the customer having telephone. Therefore the average time duration is approximately 4 minutes lesser for the customer who have cellular phone. On the contrary, the average time duration for the customer who says no, for both the customer having cellular and telephone is almost same.

The fourth plot compares the month feature with that of the duration. The average time duration for the customer who said yes is highest for the month of May and July, where it stands at 11.7 minutes and 11.8 minutes, respectively and the lowest average time is seen in the month of March, where it is at 5.1 minutes. On the contrary, the range of average time duration of the customers, who says no, across all months tends to remain same which is between 3 to 4.5 minutes.

Analysis of Emp.Var.Rate:

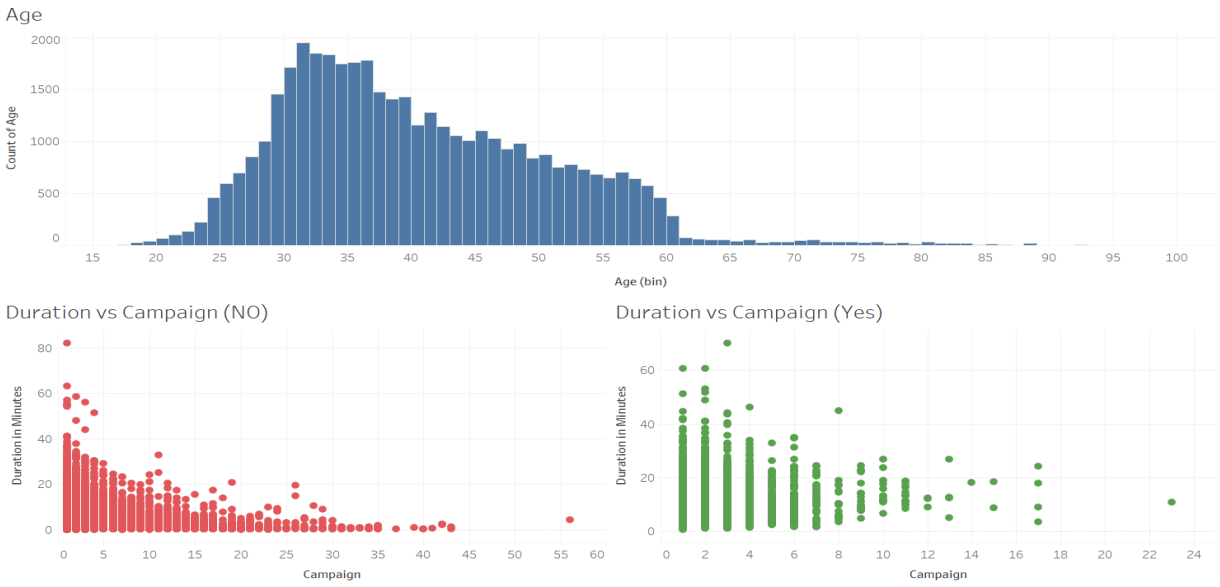


The above plot has the higher end of the emp.var.rate which consists of -0.1,1.1 and 1.4. These emp.var.rate prevail during the months of May, June, July, August and November. The number in October is very negligible. As we during this period the percentage of customers saying yes stands at 4.83%. Let us compare this with the lower range.



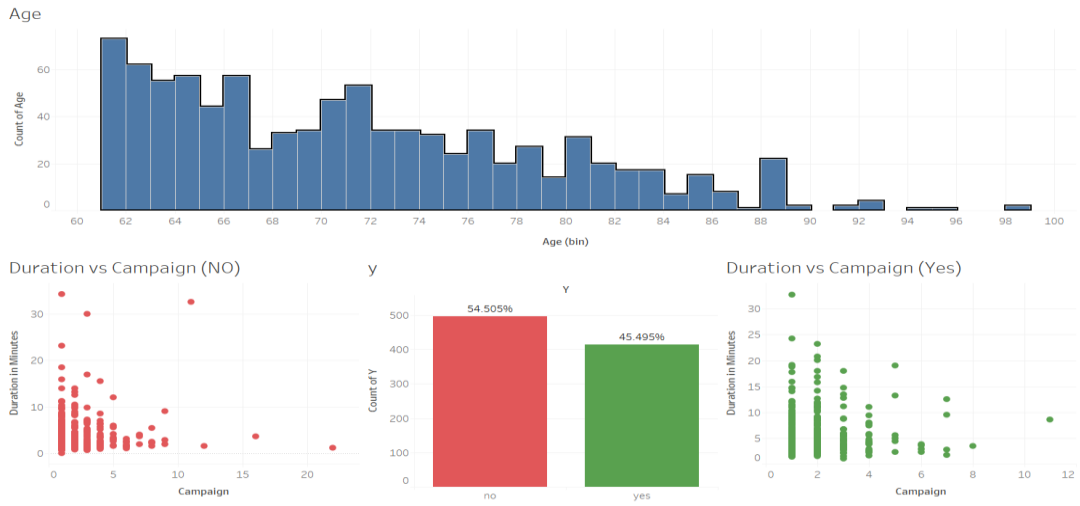
The emp.var.rate at the lower ranges are -3.4, -3, -2.9, -1.8, -1.7, -1.1. These rates were present in almost all the months. The percentage of customers saying yes in these emp.var.rate stands at 25.46%. Almost 20% difference in the percentage of customers saying yes between the lower range and upper range. Therefore it is inferred that the percentage of customers saying yes to the marketing campaign tend to be higher when the emp.var.rate is lower.

Analysis on Age, Duration & Campaign:



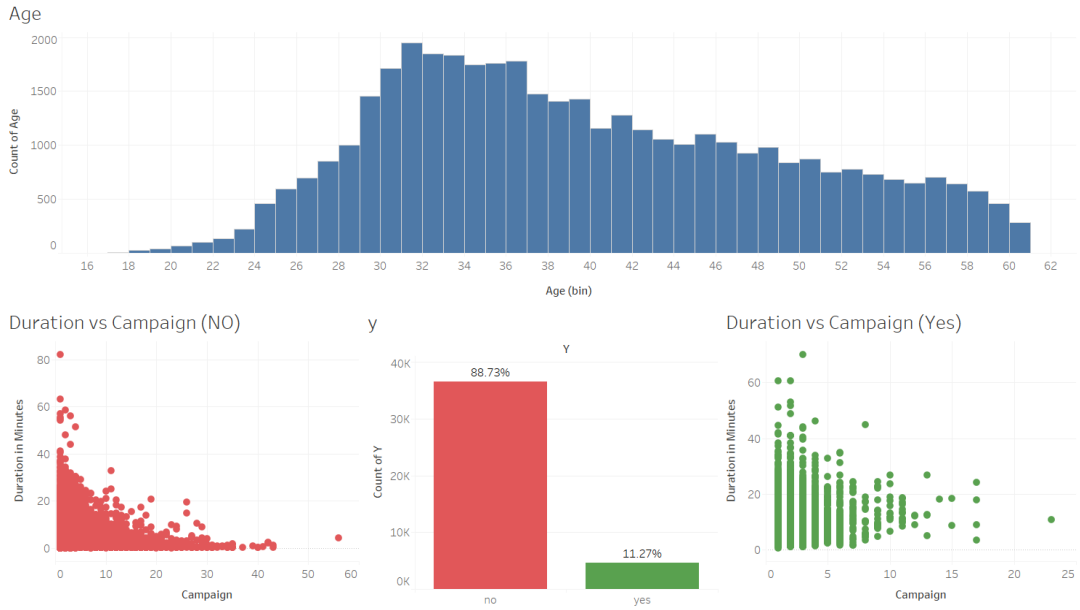
In general across all the age, the number of calls done to a customer who tend to say no, during this marketing campaign tend to extend till 45, with an outlier at around 56. When we compare it with the customer saying yes, the number of calls reduces to 17, with an outlier at 23. There is a considerable amount of difference in the number of calls made to the customer saying yes and no.

We have already discussed the time duration previously based on the average time duration. Here the number of minutes engaged to a customer saying no, extends till 60 minutes, with a outlier around 80 minutes. Whereas the time duration for the customer saying yes extends till 40, with some scarce data points above that.



When we take the age range of above 60, the percentage of customer saying yes is around 45%. And the number calls that has to be made for customer saying yes is around 7 with two outliers, whereas, for the customer saying no, the number of calls extends till 10 and there are four outliers outside of that.

If we take the duration for this age group, the customer saying yes tends to have the duration around 20 minutes and if we compare the same with the customer saying no it is around 14 minutes, with some outliers.



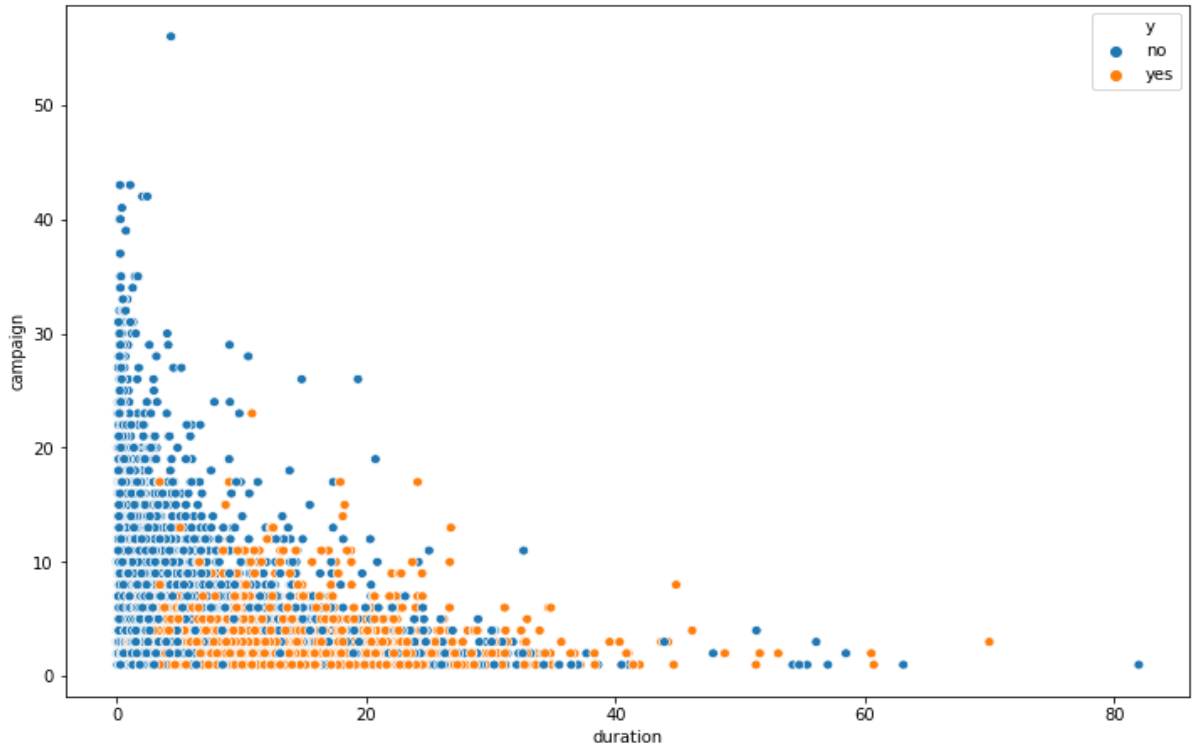
If we consider the same features for age group less than 60, the percentage customers saying yes drops to 11.27%.

If we have a look at the campaign, the maximum number of calls that has been made for customer saying yes stands at 17, whereas the maximum number of calls for the customer saying no spreads up to 40.

And when we compare the duration, the pattern exists for both the customer saying yes and no are similar to that of the age group above 60.

Therefore as we see there is a great increase in the number of customer saying yes, in the age group above 60, when compared with the age group below 60. But the number of calls made to the age group above 60 is very much less. And also the customer tends to say yes if the number of calls made to the customer is kept minimal. We recommend capping it around 20, and still if the customer says no, it is advisable to move on to the next customer, rather calling the same customer as the conversion percentage tends to drop above this.

Duration vs Campaign:



From the plot we can infer that the cluster of yes starts to appear around 7 minutes and before that it is full of no cluster. As we see the campaign, above 10, the cluster density of yes gets decreasing. Therefore we can infer that the minimum duration to get yes is around 7 minutes and the maximum calls that can be capped for this campaign should be around 10. Above 10 calls the customer tends to say no, therefore we can concentrate those calls on other customers. And also we can cap the duration around 40 minutes as the density becomes very thin after the 40 minute mark.

2.3 Null Values:

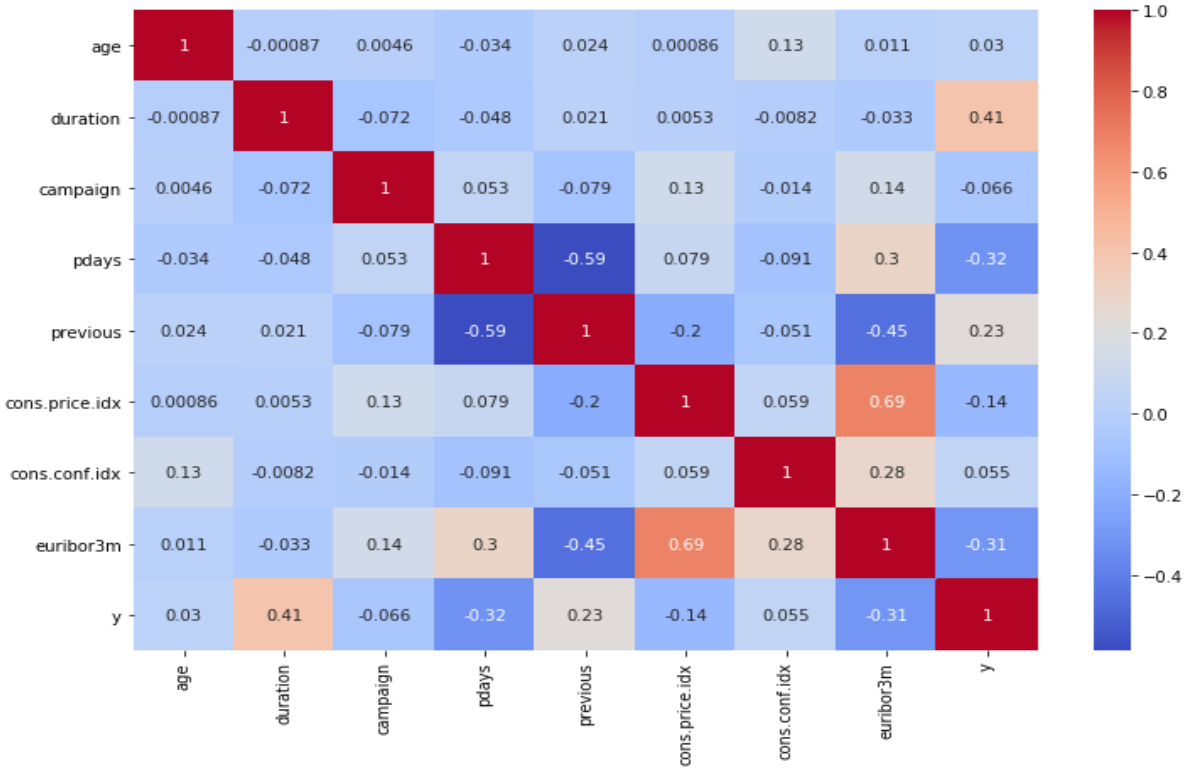
The null values are present in the following features:

Feature	No. Null Values
Job	330
Marital	80
Education	1731
Housing	990
Loan	990

We have used the KNN Imputer to the dataset to impute the null values in the above features.

2.4 Multi-Collinearity:

Let us check the collinearity in the dataset using a correlation plot:



Following table shows the features with high collineariy:

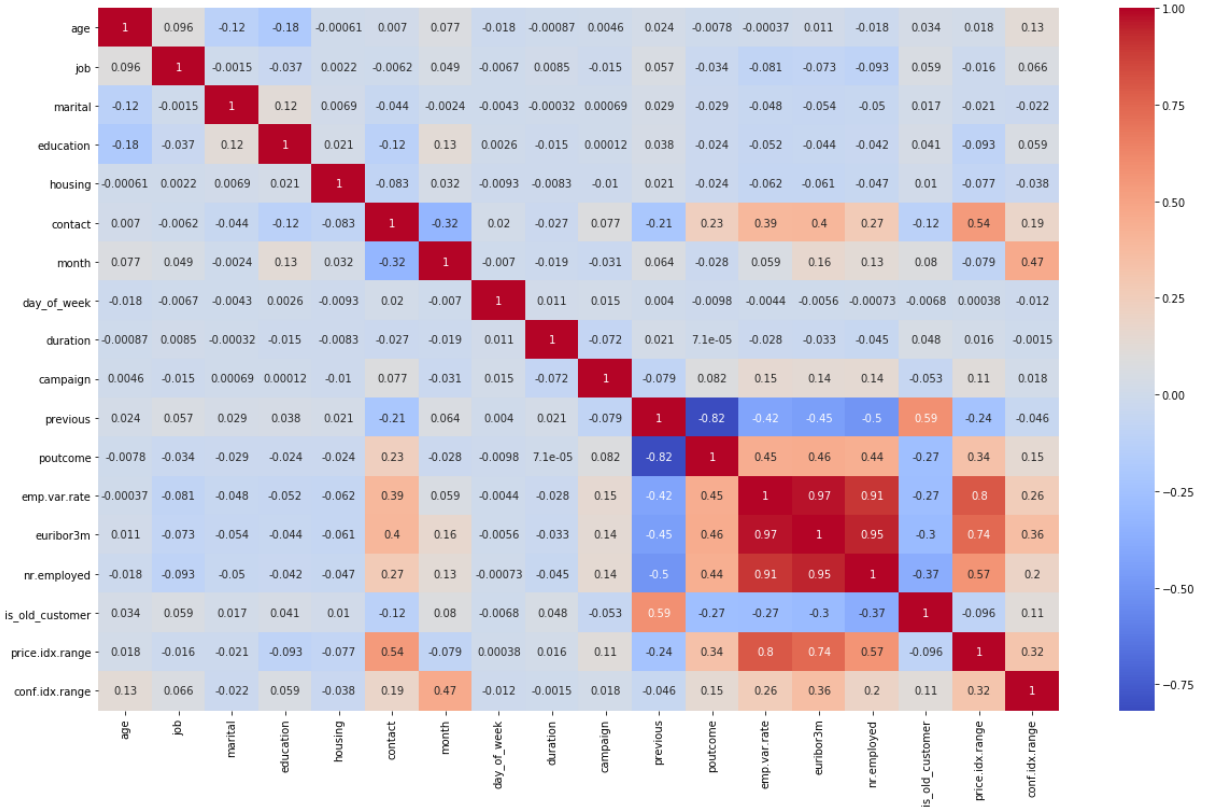
Feature 1	Feature 2	Collineartiy
cons.price.idx	euribor3m	0.69
previous	pdays	-0.59
previous	euribor3m	-0.45
pdays	euribor3m	0.3
cons.conf.idx	euribor3m	0.28
cons.price.idx	previous	-0.2

- Euribor3m feature is collinear with 4 other features
- Previous feature is collinear with 3 other features
- Conf.price.idx is collinear with 2 other features

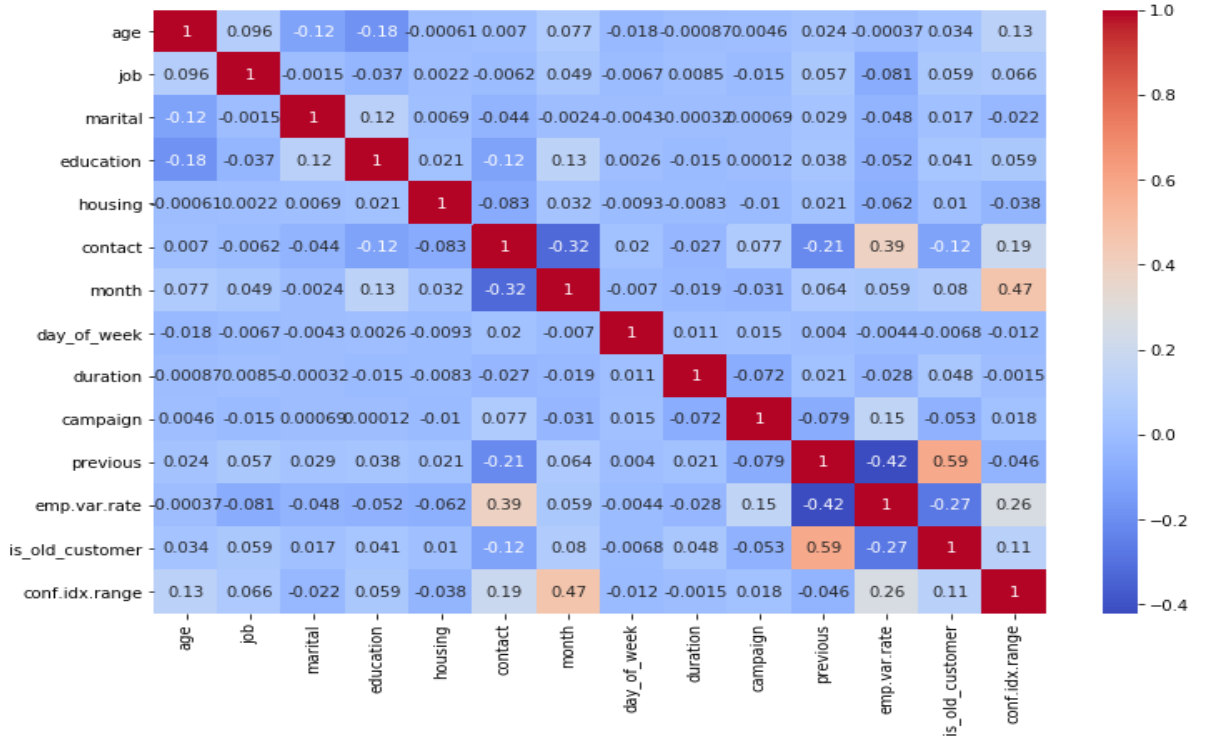
Further we have used the Variation Inflation Factor (VIF) from the statsmodels.stats.outliers_influence to select the features which reduces the collinearity within the features. It has returned the following features which results in the reduced collinearity between the features:

- age
- job
- marital
- education
- housing
- contact
- month
- day_of_week
- duration
- campaign
- previous
- emp.var.rate
- is_old_customer
- conf.idx.range

Correlation after Imputation:



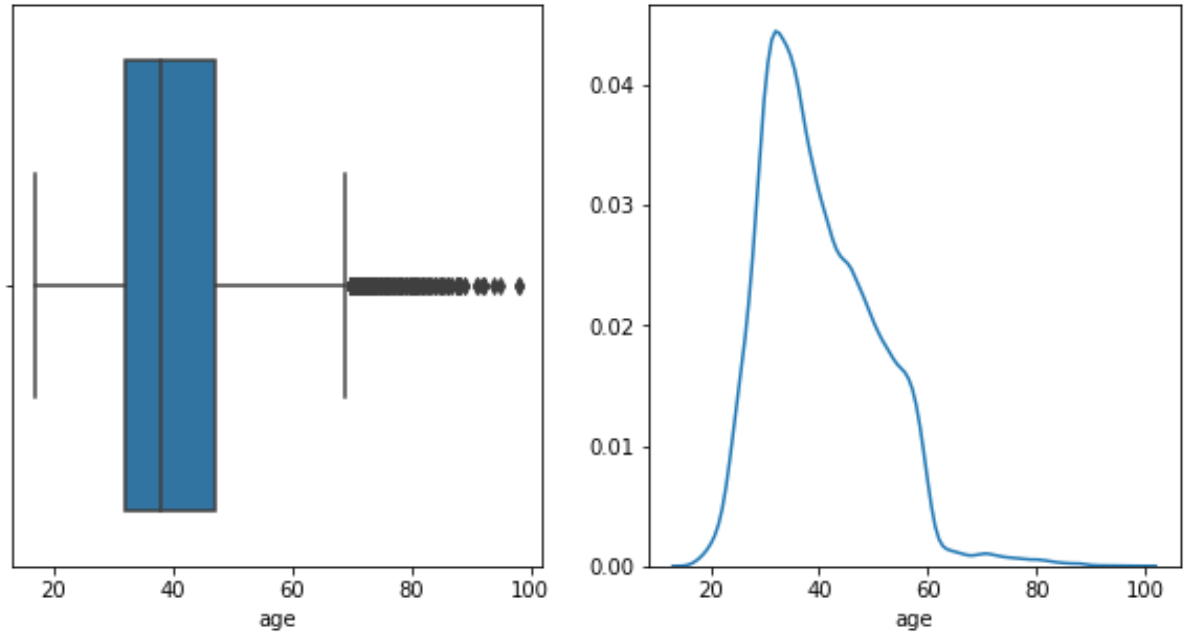
Correlation plot with the features selected using VIF:



As we compare both the plots we can see that the collinearity between the features has been greatly reduced. Though we have to check how the model built with these features perform with the respect to other models. The same comparison has been reported in the later part of the report.

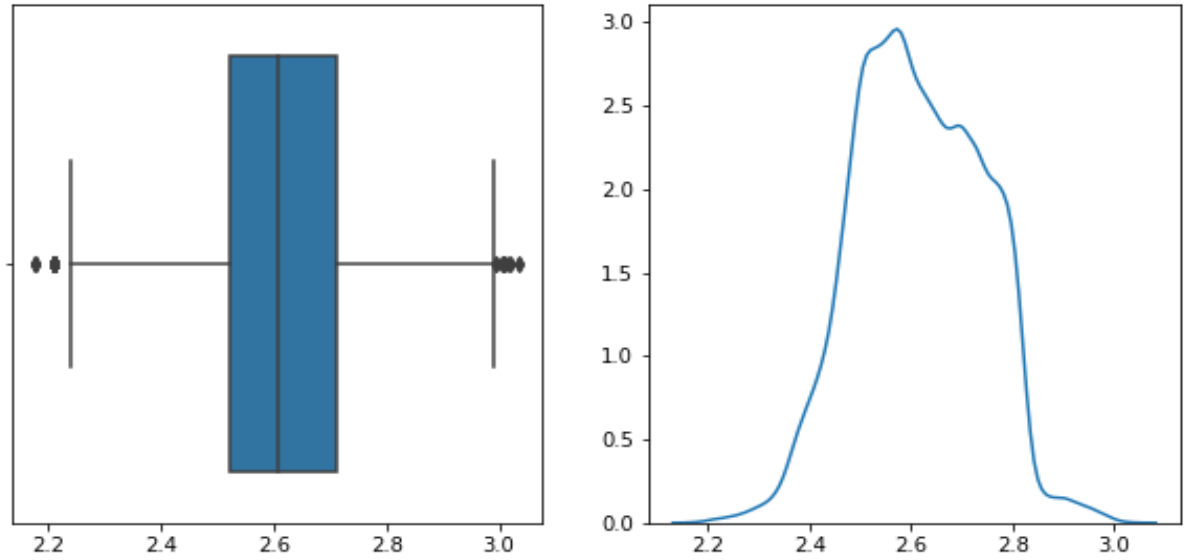
2.5 Distribution and Outliers:

Age:



The age feature has the outliers on the higher side and as a result we can see that the distribution is right skewed.

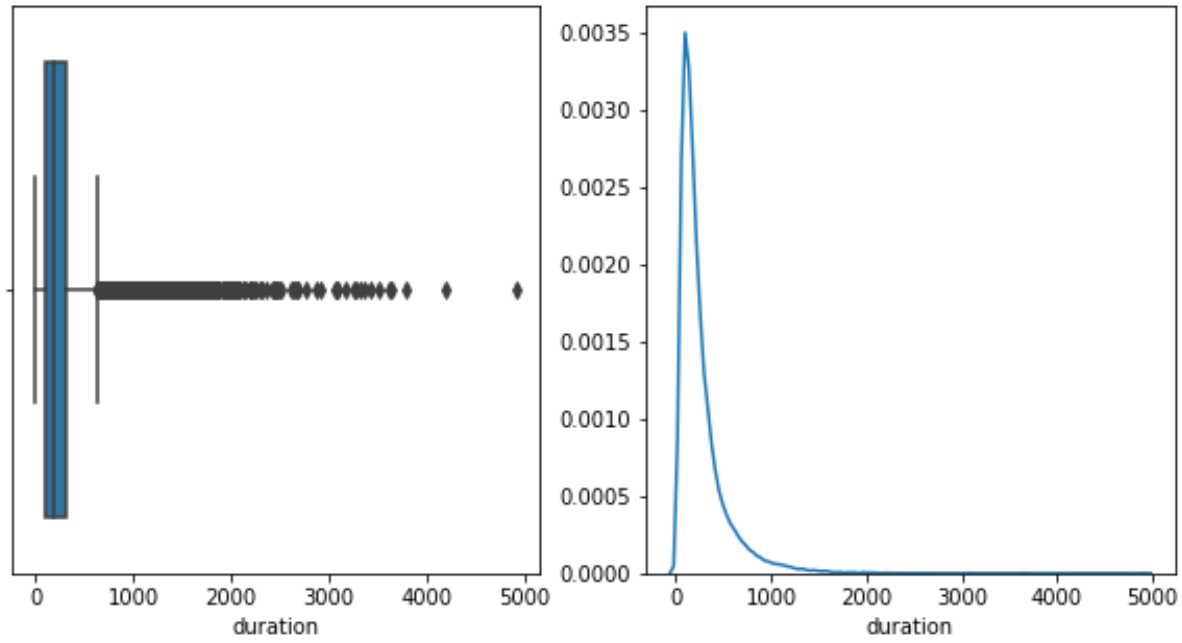
To reduce the outliers and also to get a normal distribution, we have tried a box-cox transformation and the resulting plot is shown below:



As a result the outliers on the higher side have reduced and two new outliers have emerged on the lower side. The right skew has also reduced and the plot has become somewhat normal.

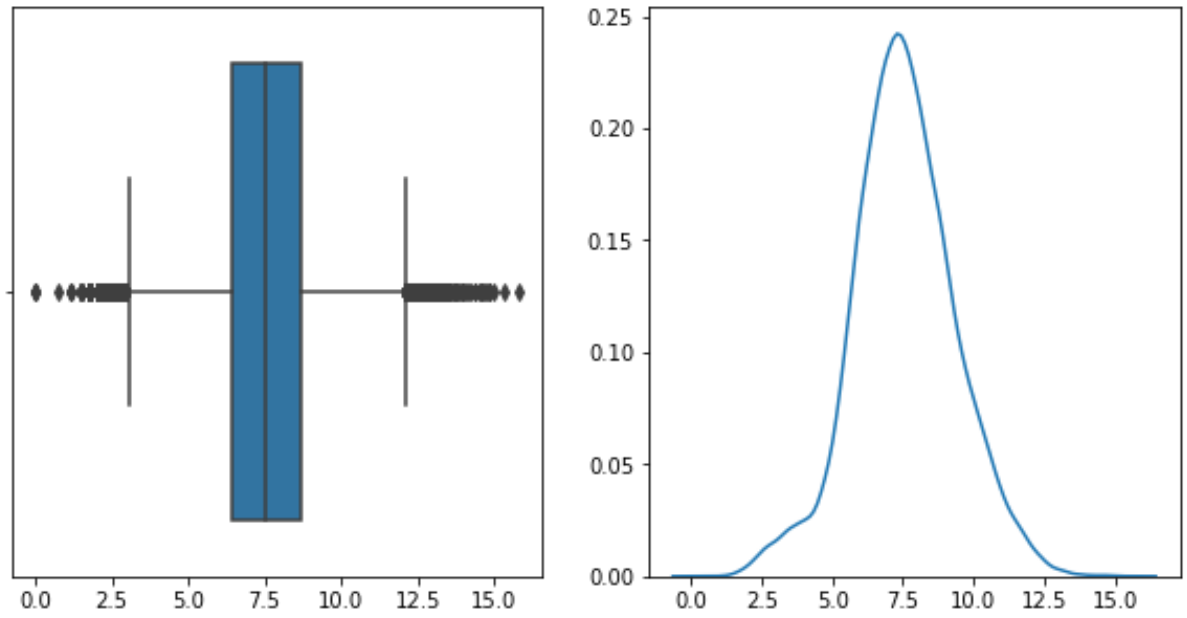
Lambda value for this transformation is -0.19444047844673226

Duration:



The duration has lot of outliers towards the higher end and as a result the distribution plot is highly right skewed.

The plot obtained after the box-cox transformation is shown below:

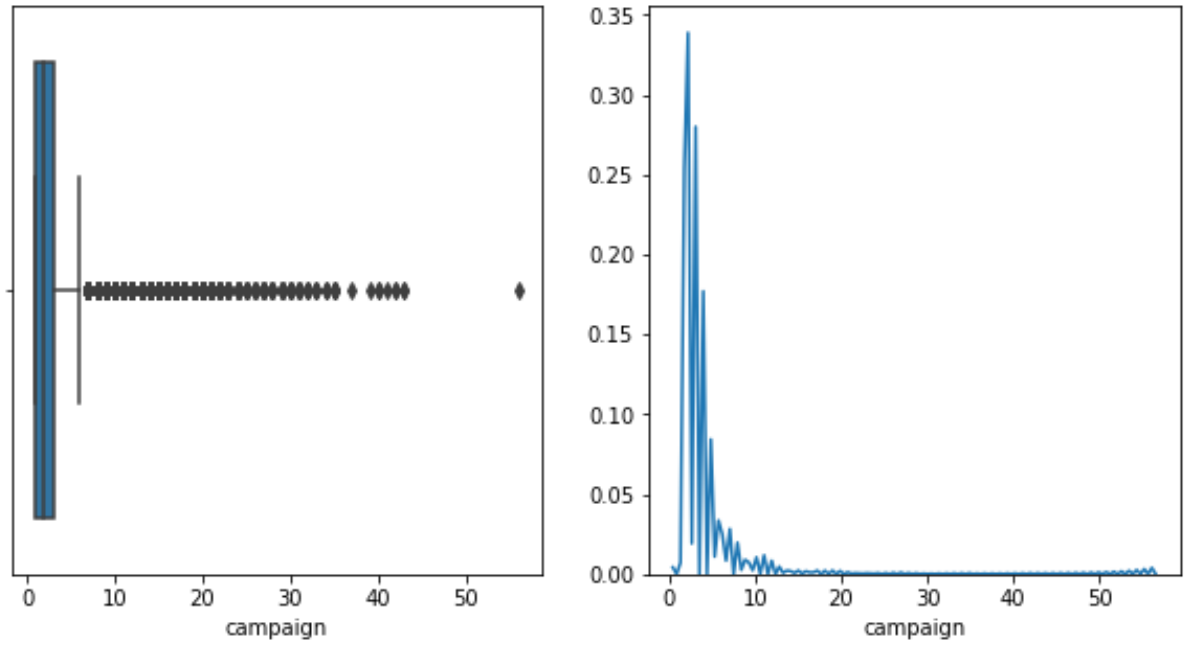


This also follows the same path as we have seen in the age feature. The outliers on the higher end have been reduced considerably but it has resulted in the emergence of few outliers on the lower end.

A far as the distribution plot, the skewness of the plot has been reduced and it has become almost a normally distributed curve.

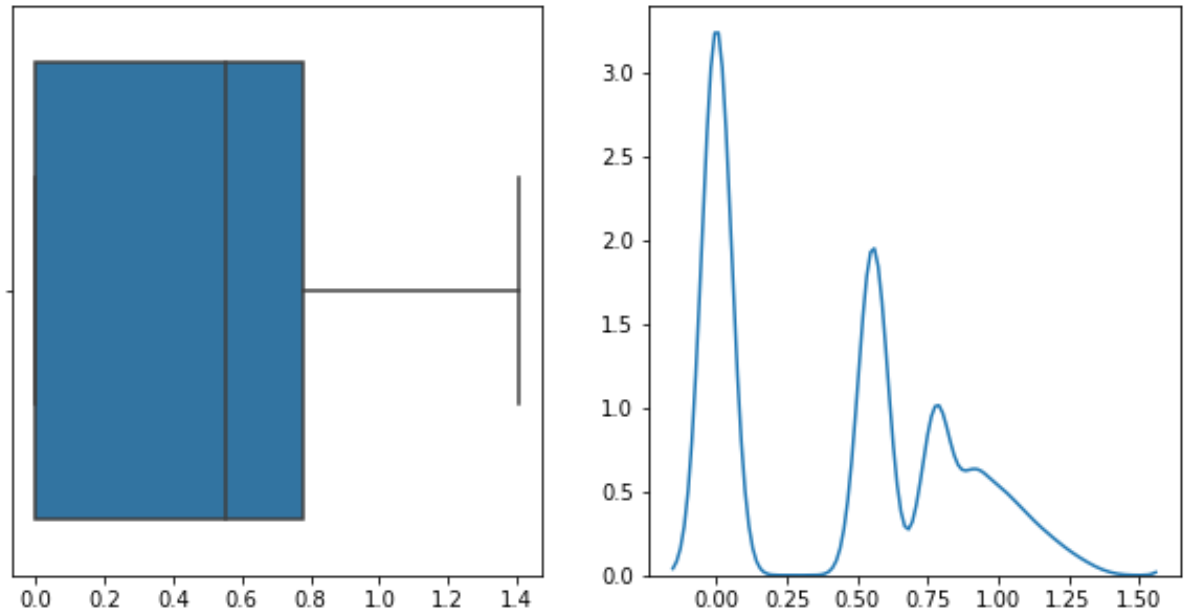
Lambda value for this transformation is 0.13343230837941755

Campaign:



In the campaign feature as well there has been lot of outliers on the higher end. In the distribution plot, the plot is right skewed and also there have been multiple peaks in the distribution.

After boxcox transformation, the plot resembles as follows:



The outliers have been reduced to zero and the number of peaks in the distribution has been reduced to three.

Lambda value for this transformation is -0.6627357510367072

2.6 Statistical Test:

Hypothesis Testing:

Two Sample T – test for Continuous Variables:

H0 : There is no significant relationship between default and y / means are equal

H1 : There is significant relationship between default and y / means are not equal

Chi squared test for Independence for Categorical Variables:

H0 : The two variables are independent

H1 : The two variables are associated

Level of Significance:

The level of significance for both the Two sample T – test and Chi Squared Test = 0.05

Declaration of Significance of a feature:

- If the obtained p – value for each feature undergoing the respective statistical test is less than 0.05, then we reject the null hypothesis and therefore the feature is declared significant and can be used to predict the target variable.
- If the obtained p – value for each feature undergoing the respective statistical test is greater than 0.05, then we fail to reject the null hypothesis and therefore the feature is declared non-significant and cannot be used to predict the target variable

The statistical test has been conducted before imputing the null values and after imputing the null values. Both the p – values has been mentioned for comparison in the following table,

Feature	Statistical Test Used	Before Imputation		After Imputation	
		p value	Result	p value	Result
Age	Two sample T-test	6.80213641846347E-10	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.	6.80213641846347E-10	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.
Duration	Two sample T-test	0.0	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.	0.0	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.
campaign	Two sample T-test	2.00777999061757E-41	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.	2.00777999061757E-41	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.
euribor3m	Two sample T-test	0.0	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.	0.0	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.
Job	Chi Squared test for Independence	4.18976328756386E-199	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.	3.33792135073991E-201	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.
Marital	Chi Squared test for Independence	2.06801464844221E-26	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.	3.44698056008601E-27	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.
education	Chi Squared test for Independence	3.30518901440250E-38	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.	2.79973828098257E-35	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.
Housing	Chi Squared test for Independence	0.05829447669453	The p-value is **greater** than 0.05, so we **accept** the null hypothesis.	0.02074085297508	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.
Loan	Chi Squared test for Independence	0.57867528704418	The p-value is **greater** than 0.05, so we **accept** the null hypothesis.	0.37356361930075	The p-value is **greater** than 0.05, so we **accept** the null hypothesis.
Contact	Chi Squared test for Independence	1.52598565231299E-189	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.	1.52598565231299E-189	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.
Month	Chi Squared test for Independence	0.0	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.	0.0	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.
day_of_week	Chi Squared test for Independence	2.95848200527853E-05	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.	2.95848200527853E-05	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.

previous	Chi Squared test for Independence	0.0	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.	0.0	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.
poutcome	Chi Squared test for Independence	0.0	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.	0.0	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.
emp.var.rate	Chi Squared test for Independence	0.0	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.	0.0	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.
nr.employed	Chi Squared test for Independence	0.0	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.	0.0	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.
is_old_customer	Chi Squared test for Independence	0.0	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.	0.0	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.
price.idx.range	Chi Squared test for Independence	1.82569873552716 E-204	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.	1.82569873552716 E-204	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.
conf.idx.range	Chi Squared test for Independence	0.0	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.	0.0	The p-value is **lesser** than 0.05, so we **reject** the null hypothesis.

From the above table we can infer that the housing and loan features fails the statistical test before imputation. After the imputation is done the housing features passes the test, but the loan feature still fails. Therefore it will not be useful in predicting the target variable.

Further, the default feature is not considered for the statistical test. The feature has 99.999% of the category No and only 0.001% of the category Yes. Therefore the feature is almost redundant and cannot be used in predicting the target variable.

2.7 Class Imbalance:

The dataset consists of binary class, namely Yes and No. Yes denotes that the customer has subscribed to the term deposit of the bank and No represent that the customer doesn't subscribe to the term deposit.

In this dataset the Class No is of 36,548, which is of 89% and Class Yes is of 4640, which is of 11%. There is an imbalance in the dataset, but this imbalance is the real-world scenario as the number of customers saying Yes, tends to be very much less than the one's saying No. Therefore it is better to train the model with this imbalance so that the model can learn this.

However, we will try the methods available to counter this class imbalance, such as SMOTE to understand how the model performance and report the same in the later part of the report.

2.8 Feature Engineering:

Three features has been modified to get the better understanding and inference from those features:

pdays :

The feature Pdays consists of data in which if the customer is denoted as '999' the customer has not been contacted previously and then if the customer has been contacted previously, then the number of days passed since the customer has been contacted.

For better understanding of this feature we have converted this into another feature called 'is_old_customer' where if the data is '0' then they are new customer and if the data denoted '1' then they are old customer.

price.idx.range :

The feature Cons.Price.Idx has been converted into Price.Idx.Range to obtain a clear inference from the feature.

The feature Price.Idx.Range has been derived as follows:

If the Cons.Price.Range is between 92.198 and 93.056, then the Price.Idx.Range is 0,

If the Cons.Price.Range is between 93.056and 93.912, then the Price.Idx.Range is 1,

If the Cons.Price.Range is greater than 93.056, then the Price.Idx.Range is 2.

conf.idx.range :

The feature Cons.Conf.Idx has been converted into Conf.Idx.Range to obtain a clear inference from the feature.

The feature Conf.Idx.Range has been derived as follows:

If the Cons.Conf.Range is between -50.824 and -42.833, then the Conf.Idx.Range is 0,

If the Cons.Price.Range is between -42.833 and -34.867, then the Conf.Idx.Range is 1,

If the Cons.Price.Range is greater than -34.867, then the Conf.Idx.Range is 2.

2.9 Scaling:

To make it much easier for the model to learn the information from the data, we scale the data using the StandardScaler object from the Sklearn library.

The StandardScaler is first fit with the train data. This helps the StandardScaler object to learn the statistical data of the train data. Then we transform the train data with the StandardScaler object. Then

with the same object the test data is also transformed, so that the test data is transformed with the same statistics of the train data.

2.10 Feature Selection:

The feature selection is used to select the statistically important features to predict the target variable. This usually helps in increasing the performance of the model by eliminating the less statistically important features. But it solely depends upon the nature of the dataset.

Here we have used the inbuilt Recursive Feature Elimination (RFE) technique from the sklearn.feature_selection library to select the statistically important features.

While using the Logistic Regression model the RFE returned 11 features as the most important features. Those features are,

- education
- contact
- month
- duration
- poutcome
- emp.var.rate
- euribor3m
- nr.employed
- is_old_customer
- price.idx.range
- conf.idx.range

While using the Decision Tree model the RFE returned 16 features as the most important features. Those features are,

- age
- job
- Marital
- Education
- Housing
- Contact
- month
- day_of_week
- duration
- campaign
- previous
- poutcome
- emp.var.rate
- euribor3m
- nr.employed
- is_old_customer

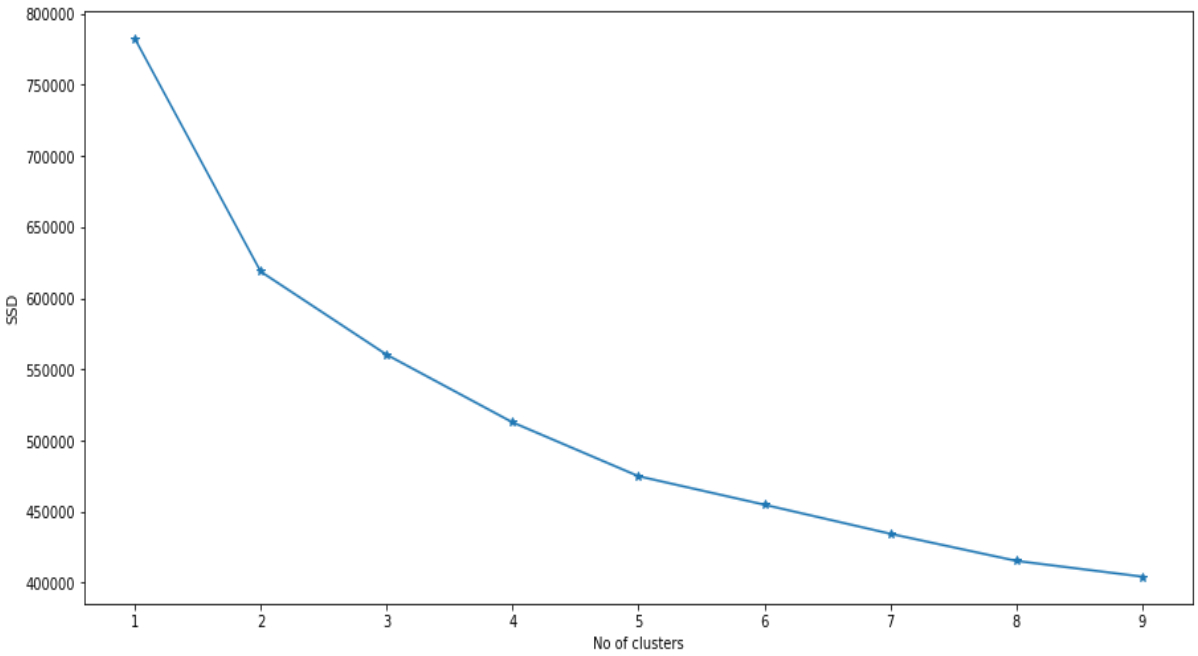
While using the Random Forest model the RFE returned 4 features as the most important features. Those features are,

- age
- duration
- euribor3m
- nr.employed

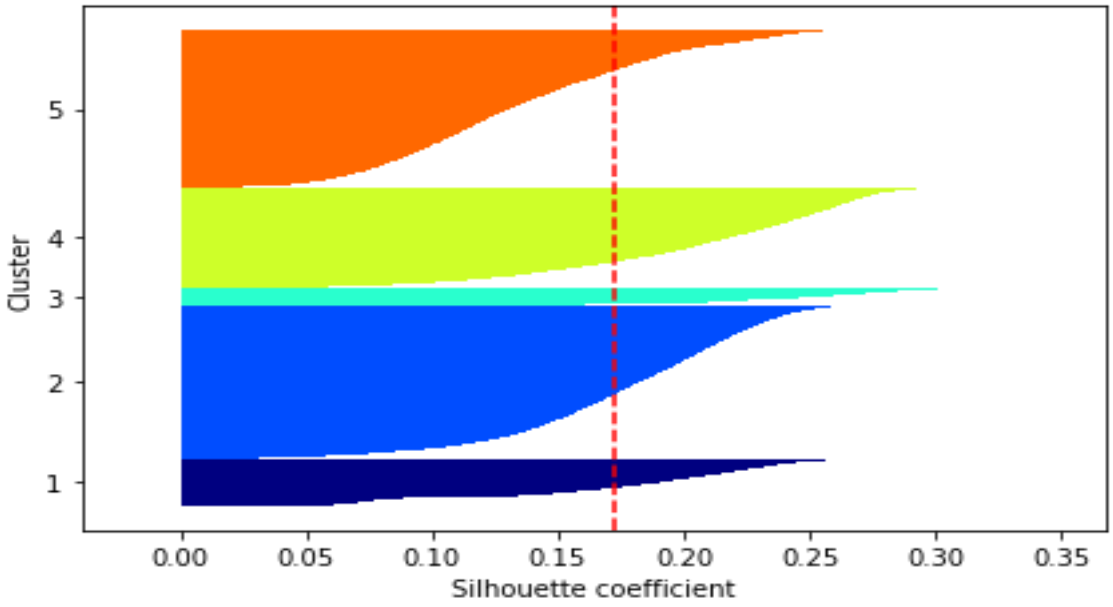
2.11 Unsupervised Learning:

Apart from building the Classification model to classify the target variable in the dataset, we also tried to find if there are any hidden clusters which help in better understanding and classification of the dataset. Therefore we removed the target variable and tried to implement the unsupervised learning to understand if the dataset exhibits different number of clusters.

We have applied KMeans clustering to the dataset and from the obtained result we have plotted the elbow plot to get the optimum number of clusters.



From the plot we can see that the first elbow happens at cluster number 2 and that has been already present in the dataset. Next to that we have identified that the second elbow happens at cluster number 5. Therefore we have applied the same to the dataset to classify the data into 5 clusters.



From the plot we can see that there is no negative classification among the clusters. The number of data in each class is as follows:

Cluster	Value Counts
3	13672
0	13284
1	8607
4	4110
2	1515

We have done a cross verification to see how the target variable is distributed among the clusters

Cluster	Y	
	Yes	No
2	63.82%	36.17%
1	21.25%	78.74%
4	12.94%	87.05%
3	6.10%	93.89%
0	3.59%	96.40%

Of all the clusters the cluster 2 has highest conversion rate as we see that about 63.82% of customer tends to subscribe the marketing campaign and in turn the term deposit of the bank.

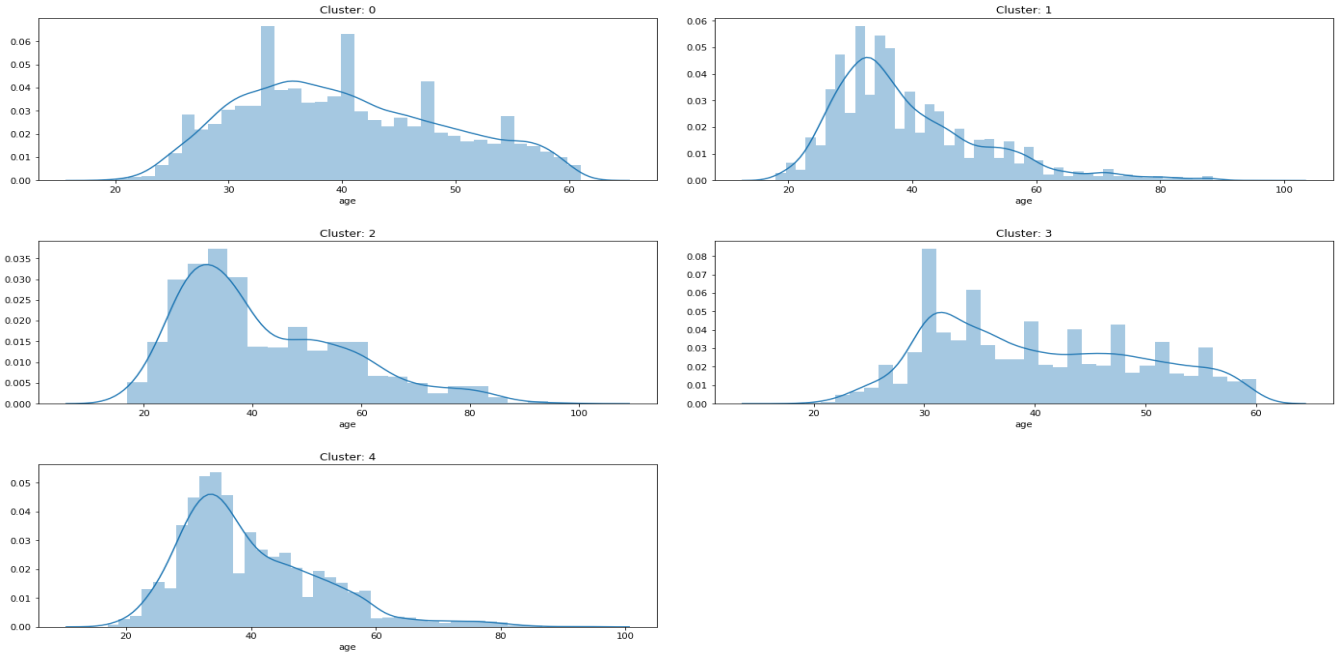
It is followed by cluster 1, 4, 3 and 0, respectively. Cluster 3 and 0 have very poor conversion rates as their percentage of customer saying yes turns out to be 6.10% and 3.59% respectively.

Further we have looked into the average time duration the customers being engaged in each cluster.

Cluster	Duration (in Minutes)
2	5.35
1	4.47
3	4.23
0	4.21
4	4.08

We had an inference in the supervised learning EDA part that the customer who tends to say yes has the highest average call duration when compared to others. The pattern in here also tends to be same as the clusters with highest conversion rate 2 and 1 tend to have highest average call duration.

Let's have a look at the distribution of age across all the clusters



As we see in the above distribution plot the Clusters with high conversion rate 2,1 and 4, respectively have include the age above 60 years, whereas the Clusters with low conversion rate 1 and 3 have their maximum age around 60 years. This supports our previous claim in the Supervised learning EDA that the conversion rate increases as we move above 80 years of age.

While we look into the allocation of old and new customers between the clusters:

Cluster	Old Customer	New Customer
0	0%	100%
1	0%	100%
2	100%	0%
3	0%	100%
4	0%	100%

The entire old customer has been clustered into one and they are present in Cluster 2 and all the other Clusters have the new customers. This also supports our previous inference in the Supervised Learning were if the customer already knows about the bank then the conversion rate tends to increase. This has also become true with this clustering.

The next important feature we have assessed is contact:

Cluster	Cellular	Telephone
3	97%	3%
2	93%	7%
4	92%	7%
1	11%	89%
0	0%	100%

The cluster 0 has 100% of customers who use telephones. And as we see the conversion in this cluster is very less among all. The most interesting cluster of all is cluster 3. It has almost 97% of the customers who use Cellular phones. But contrary to what we have seen in previous analysis, the conversion in this cluster is second lowest. Having only 11% of the Cellular Phone users in the cluster, Cluster 1 is the second best cluster in terms of conversion.

Next feature is month,

Cluster	April (%)	August (%)	December (%)	July (%)	June (%)	March (%)	May (%)	November (%)	October (%)	September (%)
0	0.00	1.11	0.00	7.45	32.93	0.00	58.44	0.00	0.08	0.00
1	21.77	6.32	1.03	3.03	7.31	4.26	46.44	2.46	4.05	3.31
2	7.59	15.58	3.04	7.79	10.17	5.94	16.37	12.54	10.36	10.63
3	0.00	36.78	0.07	41.66	0.00	0.00	0.00	21.10	0.39	0.00
4	15.64	5.43	0.92	2.68	3.92	2.17	42.85	19.81	3.58	3.02

The cluster 2 has the best conversion rate among all and the percentage of calls made has been spread across all the months. If we take the cluster with worst conversion rate, i.e., Cluster 0, more than half the percent of the calls is made in the month of May. The cluster 3 has the second worst conversion rate and it has 42% of calls made on July and 36% calls made on August. Therefore if the percentage of calls made in each month is gets evenly spread out like what we see in Cluster 2, the conversion might tend to increase.

The below table represents the maximum number of calls done per client for this marketing campaign alone:

Cluster	Campaign (Max)
2	13
4	16
1	23
3	43
0	56

As we have seen in the analysis part of the supervised learning, the customer who tends to get converted has got fewer numbers of calls in the particular campaign and as the number of calls gets increased, the customer didn't get converted. We have also suggested reducing the number of calls done to a particular customer during this campaign to 15 and the remaining calls can be concentrated towards the newer customers.

We can see the same kind of pattern here as well as the Cluster 2, with maximum conversion rate has the maximum number of calls made in the campaign standing at 13. If we compare this with the cluster with worst conversion rate, Cluster 0, the maximum number of calls made stands ate 56, followed by the cluster with second worst conversion rate Cluster 3 having maximum calls at 43. We can see a huge increase in the maximum number of calls made when we compare Cluster 2,4,1 and Cluster 3 and 0.

Cluster	Previous							
	0	1	2	3	4	5	6	7
0	100%	0%	0%	0%	0%	0%	0%	0%
1	100%	0%	0%	0%	0%	0%	0%	0%
2	0%	57%	27%	11%	4%	1%	0.26%	0.06%
3	100%	0%	0%	0%	0%	0%	0%	0%
4	0%	90%	8%	1%	0.29%	0.05%	0.02%	0%

The Cluster with high conversion rate, 2 has 0% of customers who are new and all of their customers have been contacted at least once before this campaign. The Cluster 4, which has third best conversion rate too have the same pattern as that of Cluster 2.

Cluster 1 has 100% new customers, despite of that it has second best conversion rate. Having the customers whose age are above 60 and having the second longest average call duration might have worked in favor of conversion in this cluster.

Cluster	emp.var.rate									
	-3.4	-3	-2.9	-1.8	-1.7	-1.1	-0.2	-0.1	1.1	1.4
0	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	58.43%	41.47%
1	6.90%	1.03%	12.87%	72.47%	3.78%	2.92%	0.00%	0.00%	0.00%	0.00%
2	15.97%	3.03%	15.31%	29.90%	18.21%	15.24%	0.00%	2.31%	0.00%	0.00%
3	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	21.49%	0.00%	78.43%
4	5.70%	0.90%	7.85%	60.65%	4.16%	3.69%	0.02%	16.98%	0.00%	0.00%

On comparing the emp.var.rate with the cluster, it too follows the inference obtained from the previous analysis done in supervised learning. With increase in emp.var.rate the conversion rate tends to decrease and the clusters with the worst conversion rate, 0 & 3, falls in the highest emp.var.rate.

Cluster	nr.employed										
	4963.6	4991.6	5008.7	5017.5	5023.5	5076.2	5099.1	5176.3	5191	5195.8	5228.1
0	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	58.43%	0.00%	41.47%
1	2.92%	3.78%	3.07%	6.90%	1.03%	12.87%	69.39%	0.00%	0.00%	0.00%	0.00%
2	15.24%	18.21%	14.38%	15.97%	3.03%	15.31%	15.51%	0.00%	0.00%	2.31%	0.00%
3	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.06%	0.00%	21.49%	78.43%
4	3.69%	4.16%	4.06%	5.71%	0.90%	7.85%	56.59%	0.02%	0.00%	16.98%	0.00%

Insights:

- It is more clear that the Cluster 2 is the one with the best conversion rate among all the clusters
- Cluster 1 and 4 are the second and third best clusters in terms of the conversion rate.
- Cluster 0 and 3 are the one with the worst conversion rates.
- All the cluster analysis follows the same pattern as that of the insights obtained from the supervised learning

Suggestions:

- The sales team can be divided based on the clusters
- The Cluster 2 has the more favorable features for conversion and hence the customers in the sales persons with basic call etiquette are sufficient to deal with these customers.
- The Cluster 1 and 4 have almost the same features of 2 but the sales person will have to put considerable amount of effort to convert these customers. The sales persons with better understanding of the bank and the products can be deployed to handle these clusters.
- The Cluster 0 and 3 have worst conversion rate of all. Further most of the customers in these clusters are tend to be new and will not have any prior idea of the bank and the products. Therefore the effort needed to pursue the customer to build trust and eventually convert them to subscribe to the campaign is very much high. Therefore the sales person with more experience within the bank has to be deployed in engaging with the customers of these clusters as they tend have more knowledge about the culture and product of the bank and they will find it easier to engage with these customers.

CHAPTER - 3

Model Building – Supervised Learning

3.1 Model Building:

The model building is done with the default parameters and then the hyper parameters of the model are tuned to refine the performance of the model to obtain the best possible result.

Classes in Target Variable:

0 - Not subscribed to the Term Loan (Majority Class)

1 - Subscribed to the Term Loan (Minority Class)

VIF Model: Variation Inflation Factor (VIF) from the statsmodels.stats.outliers_influence is used to select the features which reduce the collinearity within the features. Then using the selected features the model is built.

RFE Model: Recursive Feature Elimination (RFE) technique from the sklearn.feature_selection library is used to select the statistically important features. Using the selected features the model is built.

SMOTE Model: SMOTE from imblearn.over_sampling is used to improve the class imbalance in the target variable. The minority class is synthesized to represent certain percentage of the total dataset. With the obtained dataset the model is built. Here we have used the sampling strategy as 0.3, which in turn synthesizes the minority class to represent 30% of the total dataset.

Model Tuning - Threshold:

In usual cases the classification threshold is set at 0.5, where the model classifies the output less than 0.5 probabilities as 0, while more than 0.5 are classified as 1. In the cases of class imbalance, the model will find it hard to classify the minority class when the threshold for classification is set at 0.5. In such cases the threshold of classification is iterated to find if we can obtain better tradeoff between the majority and minority class to get better classification.

In our model building we have iterated over various thresholds and found out that most of the models returns better minority class classification at the threshold of 0.2. We have updated the result obtained using the threshold of 0.2 in each model section for comparison.

Model Parameters:

There are various parameters to measure the performance of the classification model. Accuracy, ROC AUC Score, Precision, Recall, Sensitivity, Specificity are widely used one.

The performance of the classification models which uses the datasets with class imbalance are measured using the recall of the minority class. If the model has better recall value of the minority class then that model is selected to be the best model. We have also build the model in line of

obtaining the better recall score of the minority class and the model with better recall of minority class will be selected as the best model.

3.2 Base Models:

We have done a basic model building and the performance of the models has been provided in the table below:

Model	Recall	
	Mean	Variance
Naive Bayes	0.609	0.00041
Decision Tree	0.527	0.0001
Gradient Boost	0.527	0.00048
Random Forest	0.514	0.0002
Bagging Classifier	0.475	0.00018
KNN	0.429	0.00053
Logistic	0.409	0.00135
Ada Boost	0.399	0.00103

Of all the models Naïve Bayes, Decision Tree, Gradient Boost and Random Forest comes in the top four based on the Recall value. And of the four Decision Tree and Random Forest have very less variance.

Above is the performance of the base models. Moving on, we will be doing the hyper parameter tuning for the applicable models and come to the conclusion of which model performs better in classifying the target variable.

Further the threshold of classifying the minority class is iterated, and all the models are compared based on specific model to obtain the best model.

3.3 Logistic Regression:

The Logistic Regression model is built in the following methods

- Base Model
- Feature Selected using the Variation Inflation Factor
- Feature Selected using the RFE
- SMOTE

Model Results:

Model	CV Score		Recall (Majority)		Recall (Minority)		ROC	
	Mean	Variance	Train	Test	Train	Test	Train	Test
Base	0.395	0.023	0.97	0.97	0.4	0.41	0.924	0.929
VIF	0.358	0.022	0.98	0.97	0.36	0.38	0.921	0.925
RFE	0.394	0.0217	0.97	0.97	0.4	0.41	0.924	0.929
SMOTE	0.645	0.017	0.94	0.94	0.65	0.64	0.93	0.929

Of all the models the model where we have SMOTE seems to perform better out of all. But though the SMOTE model is used for the data set with large imbalance where the minority class accounts to maximum of 3%. Therefore we are not considering SMOTE model for this dataset. Next to that the RFE model performs better where it outdoes the base model with less variance.

Threshold (set at 0.2):

Model	Recall (Majority)	Recall (Minority)
Base	0.88	0.81
VIF	0.82	0.89
RFE	0.832	0.892
SMOTE	0.88	0.817

3.4 Naïve Bayes:

Model Results:

Model	Cross Validation		Recall (Majority)		Recall (Minority)		ROC	
	Mean	Variance	Train	Test	Train	Test	Train	Test
Naïve Bayes	0.609	0.0004	0.87	0.86	0.61	0.62	0.85	0.852

Threshold (set at 0.2):

Model	Recall (Majority)	Recall (Minority)
Naïve Bayes	0.83	0.709

Hyper Parameter Tuning:

The hyper parameter present in each of the model is tuned using Randomized Search Class from sklearn.model_selection library. The data set is validated by splitting the dataset into ten equal parts using the KFold object form sklearn.model_selection library. Then the hyper parameters combination is tested via 50 iteration of Randomized Search and the resulting hyper parameter that gives best score is chosen to build the model. This method is used across all the model building to verify the robustness of the result obtained.

3.5 KNearest Neighbors:

Hyper parameters:

Knn:

- Weights : Distance
- n_neighbors : 2

Bagged Knn:

- n_estimators : 43

Model Results:

Model	Cross Validation		Recall (Majority)		Recall (Minority)		ROC	
	Mean	Variance	Train	Test	Train	Test	Train	Test
Knn	0.468	0.0006	1	0.94	1	0.47	1	0.786
Bagged Knn	0.457	0.0007	1	0.95	0.99	0.46	0.99	0.878

Threshold (set at 0.2):

Model	Recall (Majority)	Recall (Minority)
Knn	0.901	0.658
Bagged Knn	0.894	0.69

3.6 Decision Tree:

Hyper Parameters:

Decision Tree:

- max_depth : 13
- max_features : 11
- criterion : entropy

Bagged Decision Tree:

- n_estimators : 13

Gradient Boost Decision Tree:

- n_estimators : 194

Ada Boost Decision Tree:

- n_estimators : 139

Model Results:

Model	Cross Validation		Recall (Majority)		Recall (Minority)		ROC	
	Mean	Variance	Train	Test	Train	Test	Train	Test
Decision Tree	0.52	0.0007	0.98	0.95	0.76	0.54	0.985	0.831
Bagged Decision Tree	0.534	0.0002	1	0.96	0.98	0.53	0.99	0.92
Gradient Boosted Decision Tree	0.533	0.0008	0.97	0.97	0.55	0.54	0.955	0.951
Ada Boost Decision Tree	0.491	0.0002	0.97	0.97	0.42	0.44	0.943	0.943

Threshold (set at 0.2):

Model	Recall (Majority)	Recall (Minority)
Decision Tree	0.908	0.71
Bagged Decision Tree	0.878	0.863
Gradient Boosted Decision Tree	0.899	0.853
Ada Boost Decision Tree	0.972	0.442

3.7 Random Forest:

Hyper Parameters:

Random Forest:

- n_estimators : 200
- max_depth : 9
- max_features : 17
- criterion : entropy

Boosted Random Forest:

- n_estimators : 56

Model Results:

Model	Recall (Majority)		Recall (Minority)		ROC	
	Train	Test	Train	Test	Train	Test
Random Forest	0.98	0.96	0.64	0.56	0.972	0.951
Ada Boost Random Forest	1	0.96	1	0.56	1	0.947

Threshold (set at 0.2):

Model	Recall (Majority)	Recall (Minority)
Random Forest	0.892	0.887
Ada Boost Random Forest	0.879	0.893

3.8 Best Model:

Comparing all the model results, the following have given best results:

Model	Cross Validation		Recall (Majority)		Recall (Minority)		ROC	
	Mean	Variance	Train	Test	Train	Test	Train	Test
Random Forest	0.564	0.0008	0.98	0.96	0.64	0.56	0.972	0.951
Gradient Boosted Decision Tree	0.533	0.0008	0.97	0.97	0.55	0.54	0.955	0.951

Threshold (set at 0.2):

Model	Recall (Majority)	Recall (Minority)
Random Forest	0.892	0.887
Gradient Boosted Decision Tree	0.899	0.853

Therefore by comparing the two models the Random Forest has given better Recall for the minority class and also the Recall (Majority), Recall (Minority) tradeoff for the Random Forest model turns out to be better one of the two.

Hence, Random Forest is the best model for classifying this dataset.

CHAPTER - 4

Conclusion

We have applied Logistic Regression, Naïve Bayes, KNN, Decision Tree, Random Forest, Bagging Classifier, Ada Boost Classifier, and Gradient Boosting Classifier to the dataset and also tuned their hyper parameters to identify the best model that can classify the customers who might subscribe to the term deposit.

At the end we have finalized the Random Forest Classifier to classify as it gave better recall of the minority class, which are the customers who might subscribe to the term deposit. Further we have iterated over the threshold of classification to have better classification of minority class. In that too, the Random Forest comes out to be the best model with better Specificity and Sensitivity.

Further we have also done an unsupervised learning to find if there are any hidden clusters in the data and we have identified five clusters within the dataset.

Further based on the Exploratory Data Analysis we are suggesting the following changes to be considered:

- Concentration given to the customers above the age of 60 has to be increased as the conversion rate over the age of 60 tends to increase, but the number of calls done to those customers was very less.
- The Customers tends to subscribe to the term loan tend to engage for higher call duration.
- The conversion rate of the customer who has cellular phone tends to be on higher side compared to that of the customer with telephones. This might also help in contact the target customer directly via cellphone and also keep them engaged via SMS and other means as well.
- The number of calls made in the particular campaign as tends to be capped at around 15 as when the numbers of calls to a particular customer increase above this the conversion rate reduces drastically. Instead of engaging the same customer above this, we can concentrate these calls on new customers.
- The customer who have subscribed to the previous marketing campaign tends to subscribe to this campaign as well. Therefore the customer who have already subscribed for the previous marketing campaigns can be included in this campaign to improve the conversion.

Based on Unsupervised Learning:

- The Cluster 2, 1 and 4 have the best conversion rate. Therefore more number of sales persons can be allocated to the customers who belong to these clusters to improve the conversion rate.
- A separate set of sales persons can be allocated to the clusters 0 and 3, which has low conversion rate. These are the cluster which is full of new customers. As we engage further with these customers, they might acquire the features of the cluster 2, 1 and 4 and eventually the conversion will increase.