**Project Report on**

**Bank Marketing (Long term deposit subscription)**

**Submitted by**

**Group No. 4 [Batch: Sep 2019,** **Location: Chennai]**

**Group Members**

1. **Balaaje V S - WFPBXFVJ7A**
2. **Govindha Raju - ZWTCM7P0RN**
3. **Hari krishna B - G55O0WR054**
4. **Rajasekaran R - B6CJWAGHCM**
5. **Santhosh G S – F1MTKDTJGM**
6. **Vignesh B – F4CU6UC8TX**

**Research Supervisor**

**Subramanian P V**

**Great Lakes Institute of Management**



Contents

[1. Introduction 3](#_Toc26427168)

[2. Problem Statement, Scope and Objective 5](#_Toc26427169)

[3. Data Source and Description 8](#_Toc26427170)

[4. Data Preprocessing 12](#_Toc26427171)

[5. Exploratory Data Analysis 16](#_Toc26427172)

[6. Modelling Approach 33](#_Toc26427173)

[7. Actionable insights and recommendations to the stakeholder 33](#_Toc26427174)

[8. References and Bibliography 44](#_Toc26427175)

[9. Appendix 45](#_Toc26427176)

# Introduction

## **Abstract:**

This is the classic marketing bank dataset uploaded originally in the UCI Machine Learning Repository. The dataset gives you information about a marketing campaign of a financial institution in which you will have to analyze in order to find ways to look for future strategies in order to improve future marketing campaigns for the bank. The classification goal is to predict if the client will subscribe a term deposit (variable y).

Find the best strategies to improve for the next marketing campaign. How can the financial institution have a greater effectiveness for future marketing campaigns? What the best algorithms to predict a term deposit and improve the next campaign efficiency? In order to answer this, we must analyze the last marketing campaign the bank performed and identify the patterns that will help us find conclusions in order to develop future strategies.

## **Business interpretation:**

A bank usually invests the customer’s term deposit deposits into riskier financial assets which can earn the better return than what they pay to their customer. The customer, on the other hand, is assured a risk-free return on his/her deposit. There is a stiff competition among the financial institutions/banks in increasing the customer base in their retail banking segment. Along with offering innovative products to the public, a huge amount of money is spent on marketing their products. The term deposit is very important among the diverse range of products and services offered by banks in retail banking segment.

With advancement in data science and machine learning and availability of data, most banks are adapting to a data-driven decision. This dataset here consists of direct marketing by contacting the clients and assessing the success rate of sales made.

In this project, we apply machine learning algorithms to build a predictive model of the data set to provide a necessary suggestion for marketing campaign team. The goal is to predict whether a client will subscribe a term deposit (variable y) with the help of a given set of dependent variables. This is a real dataset collected from a Portuguese bank that used its own contact-center to do direct marketing campaigns to motivate and attract the clients for their term deposit scheme to enhance the business.

# Problem Statement, Scope and Objective

## **Problem Statement:**

A Portuguese bank is rolling out term deposit for its customers. They have in the past connected to their customer base through phone calls. Results for these previous campaigns were recorded and have been provided to the current campaign manager to use the same in making this campaign more effective.

**Challenges that the manager faces are following:**

* Customers have recently started to complain that bank’s marketing staff bothers them with irrelevant product calls and this should immediately stop
* There is no prior framework for her decide and choose which customer to call and which one to leave alone

## **Project outcome:**

The manager has decided to use past data to automate this decision, instead of manually choosing through each customer. Previous campaign data which has been made available; contains customer characteristics, campaign characteristics, previous campaign information as well as whether customer ended up subscribing to the product as a result of that campaign or not.

Using this the manager plans to develop a statistical model which given this information predicts whether customer in question will subscribe to the product or not. A successful model which is able to do this, will make her campaign efficiently targeted and less bothering to uninterested customers.

## **Objective:**

To build a machine learning predictive model and predict which customers should be targeted for rolling out term deposits by bank. Our study will adopt data mining techniques to predict customers’ term deposit subscription behaviours and understand customers’ features to improve the effectiveness and accuracy of bank marketing. In order to achieve this objective, we break the whole approach into following questions.

I. How to predict whether a bank client will subscribe to a term deposit or not?

II. Which determinants would indicate a client is ready to subscribe to a term deposit through direct marketing?

III. How to segment term deposit market?

IV. Are there any common features of clients who have subscribed to a term deposit?

**Evaluation Criterion**: Accuracy, Recall and AUC Score

## **Literature review:**

A few studies have been conducted both in India and abroad over a period of time regarding the marketing strategies applicable in the banking sector. Followings few of the studies are reviewed hereunder as they would facilitate a clear backing for carrying out the present study.

* Mehta (2010) in his article” Personal Selling-A Strategy for promoting Bank Marketing “reported that there is lack of Marketing Communication in Indian Banks. He suggested for adopting banks suitable marketing promotion strategies for better business. He emphasized that on adoption of personal selling as a strategy for marketing promotion in Banks the banking business can improve considerably.
* Moro, Cortez and Laureano used the rminer Package to test three classification models (Decision Trees, Naïve Bayes and Support Vector Machines) and compare their performance through Receiver Operating Characteristic curve (ROC) and Lift curve analysis. Moro, Cortez and Rita also tested four data mining models, including logistic regression, decision trees (DT), neural network (NN) and support vector machine. After evaluating area of the receiver operating characteristic curve (AUC) and area of the LIFT cumulative curve (ALIFT), neural network presented the best performance. Nachev combined cross-validation and multiple runs to partition the data set into train and test sets . He also explored the impact of performance caused by different neural network designs.
* Wang, Song and Fang mentioned that the banking industry lacks scientific marketing management and they came up with the idea that carrying out market segmentation of deposit marketing and selecting the marketing target is the scientific way of marketing management.

# Data Source and Description

## **Data:**

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (or not) subscribed.

## **Data Dictionary:**

There are two datasets:

1. bank-full.csv with all examples, ordered by date (from May 2008 to November 2010).
2. bank.csv with 10% of the examples (4521), randomly selected from bank-full.csv.
3. The smallest dataset is provided to test more computationally demanding machine learning algorithms (e.g. SVM).
4. The classification goal is to predict if the client will subscribe a term deposit (variable y).
5. Number of Instances: 41188 for bank-full.csv
6. Number of Attributes: 21
7. Attribute information:

**# bank client data:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.No | Column Name | Description | Type | Values |
| 1 | age | age of the person | Numeric | 40,42…… |
| 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 | default | has credit in default? | Categorical | "no","yes","unknown" |
| 5 | housing | has housing loan? | Categorical | "no","yes","unknown" |
| 6 | loan | has personal loan? | Categorical | "no","yes","unknown" |
| 7 | education | Type of education | Categorical | "basic.4y","basic.6y","basic.9y","high.school","illiterate","professional.course","university.degree","unknown" |

**# related with the last contact of the current campaign:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.No | Column Name | Description | Type | Values |
| 8 | contact | contact communication type | Categorical | "cellular","telephone" |
| 9 | month | last contact month of year | Categorical | "jan", "feb", "mar", ..., "nov", "dec" |
| 10 | day\_of\_week | last contact day of the week | Categorical | "mon","tue","wed","thu","fri" |
| 11 | duration | last contact duration, in seconds | Numeric | 149,261…….. |

**# other attributes:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.No | Column Name | Description | Type | Values |
| 12 | campaign | number of contacts performed during this campaign and for this client | Numeric | 1,2…. |
| 13 | pdays | number of days that passed by after the client was last contacted from a previous campaign | Numeric | 1,2….. 999,999- means client was not previously contacted |
| 14 | previous | number of contacts performed before this campaign and for this client | Numeric | 0,1,2.. |
| 15 | poutcome | outcome of the previous marketing campaign | Categorical | "failure","nonexistent","success" |

# **social and economic context attributes**:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.No | Column Name | Description | Type | Values |
| 16 | emp.var.rate | employment variation rate - quarterly indicator | Numeric | 1.1,1.5…. |
| 17 | cons.price.idx | consumer price index - monthly indicator | Numeric | 93.99,93.84…. |
| 18 | cons.conf.idx | consumer confidence index - monthly indicator | Numeric | -36.4,-35.5… |
| 19 | euribor3m | The Euro Interbank Offered Rate is a daily reference rate, published by the European Money Markets Institute- 3 month rate | Numeric | 4.857,5.025…. |
| 20 | nr.employed | number of employees - quarterly indicator | Numeric | 5191,5250… |

**#Output variable (desired target):**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.No | Column Name | Description | Type | Values |
| 21 | y | has the client subscribed a term deposit? | Categorical | "yes","no" |

8. **Missing Attribute Values**: None

## **Variable categorization**

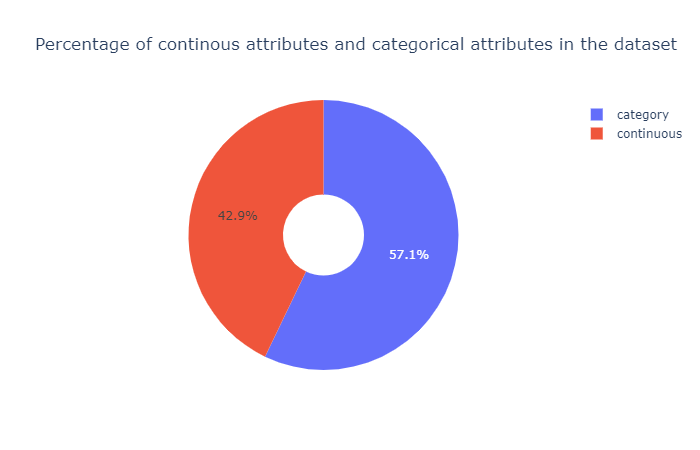


Figure no 1 – Percentage of continuous attributes and categorical attributes in the dataset.

# Data Preprocessing

## **Outliers:**

**Isolation Forest:**

Isolation Forest is an unsupervised learning algorithm that belongs to the ensemble decision trees family. This approach is different from all previous methods. All the previous ones were trying to find the normal region of the data then identifies anything outside of this defined region to be an outlier or anomalous.

This method works differently. It explicitly isolates anomalies instead of profiling and constructing normal points and regions by assigning a score to each data point. It takes advantage of the fact that anomalies are the minority data points and that they have attribute-values that are very different from those of normal instances. This algorithm works great with very high dimensional datasets and it proved to be a very effective way of detecting anomalies.

Outliers were found using isolation forest,

|  |  |  |
| --- | --- | --- |
| S.No | Column name | Outlier percentage |
| 1 | age | 1.14 |
| 2 | duration | 7.19 |
| 3 | campaign | 5.84 |
| 4 | pdays | 3.68 |
| 5 | emp.var.rate | 0.00 |
| 6 | cons.price.idx | 0.00 |
| 7 | cons.conf.idx | 1.09 |
| 8 | euribor3m | 0.00 |
| 9 | nr.employed | 0.00 |

* Outliers with outlier percentage greater than 5 were removed and imputed.
* Various imputation techniques were tested and finally fixed with median.
* bfill produced very less outlier percentage and followed close to normality than other methods (mean, median, ffill, mode).
* Even though bfill produced less outlier produced it produced less precision compared to median imputation.
* Hence, we are going with median imputation.

Before outlier treatment,

A screenshot of a cell phone

Description automatically generated

Figure no 4.1 – Percentage of outliers in each continuous column.

A screenshot of a cell phone

Description automatically generated

Figure no 4.2 – Percentage of long-term subscription.

Duration and campaign column before Treatment:

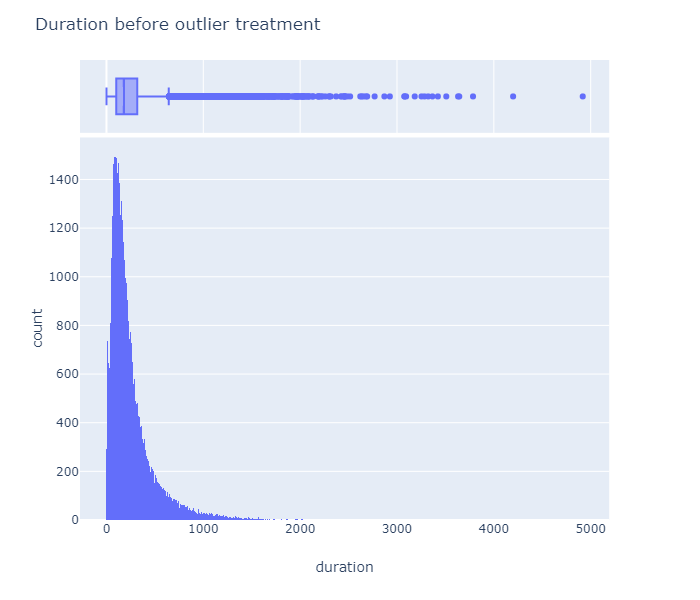
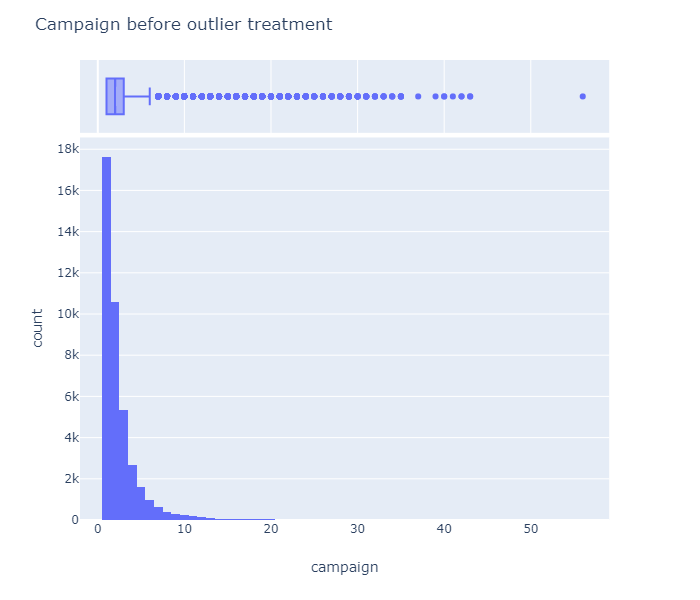


Figure no 4.2.1 – Campaign before outlier treatment. Figure no 4.2.2 – Duration before outlier treatment

After outlier treatment,

A screenshot of a cell phone

Description automatically generated

Figure no 4.3 – Percentage of outliers in each continuous column.

A close up of a logo

Description automatically generated

Figure no 4.4 – Percentage of long-term subscription.

Duration and campaign column after outlier treatment:

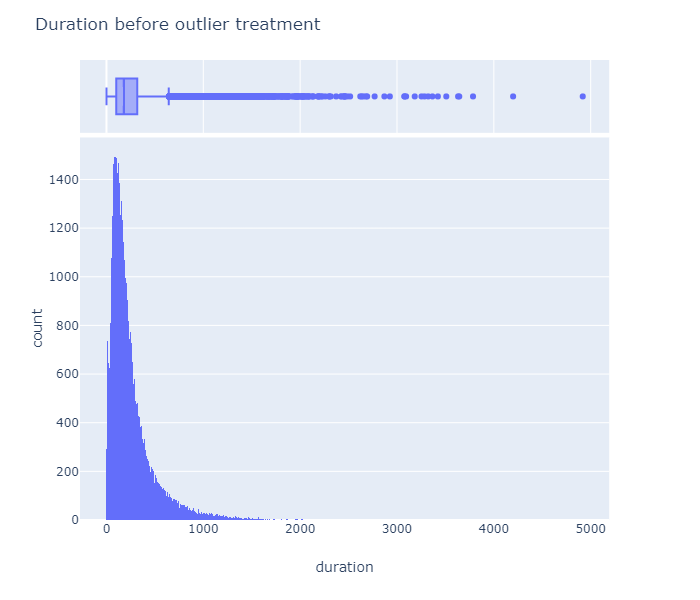
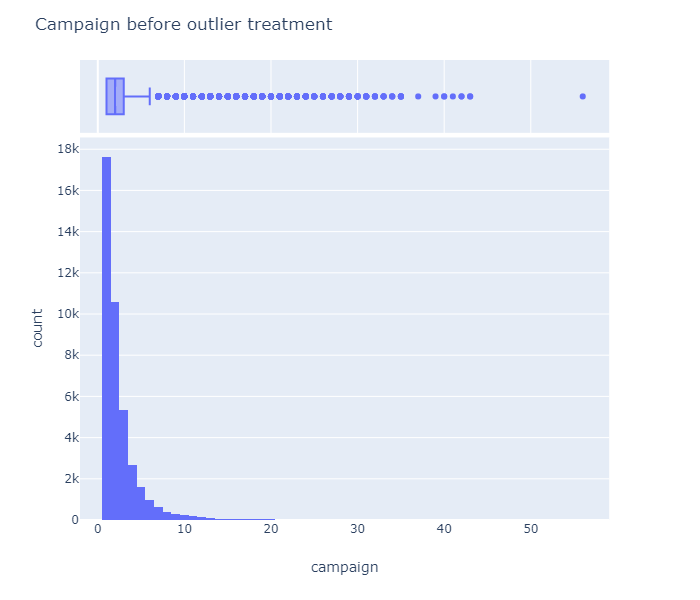


Figure no 4.4.1–Campaign After outlier treatment. Figure no 4.4.2 – Duration After outlier treatment

# Exploratory Data Analysis

A close up of a logo

Description automatically generated

Figure no 5.1 – Information on Term deposit Subscription

**Inference:**

From the plot we can see that only, 11.3% of people are subscribed to the term deposits and this implies that our data is imbalanced. We can try to balance the data using balancing techniques like SMOTE and over sampling techniques which can help in improving the accuracy.

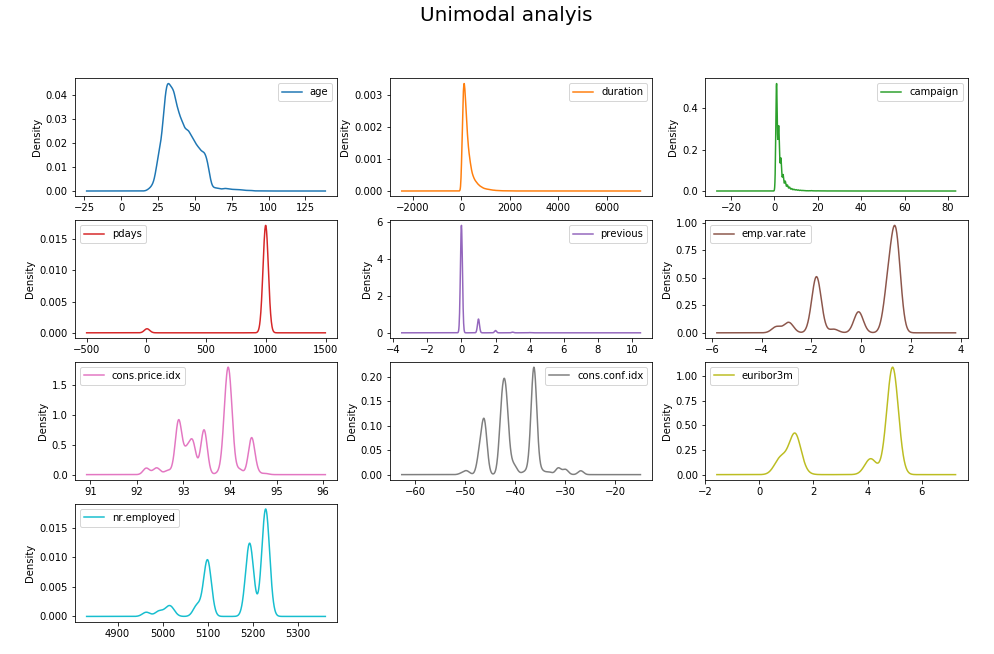


Figure no 5.2 – Unimodal Analysis with KDE plot

**Inference:**

Most of the continuous columns are skewed and varies very much in terms of scale. Hence scaling and outlier removal is necessary for this dataset.

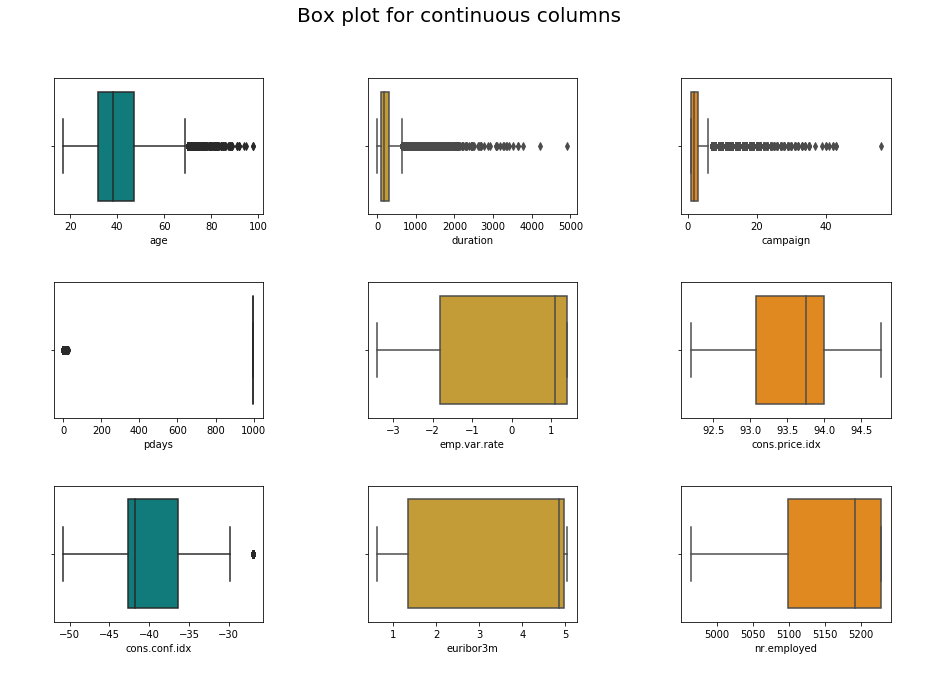


Figure no 5.3 – Box plot for continuous columns

**Inference:**

The box plot shows that there are outliers present in a couple of columns. These should be treated before proceeding to model building.

A screenshot of a cell phone

Description automatically generated

Figure no 5.4 – Contact details on Term deposit Subscription

**Observation:**

* This bar plot consists of two contact details as cellular and telephone. While the blue shaded region shows who opened term deposit and orange shaded region shows who not opened term deposit. X-axis represents contact category whereas Y-axis represents subscribers count.
* From this bar plot, we can visualize that the customers not subscribed is high in both the categories- cellular and telephone. 3,853 customers contacted through cellular have subscribed which gives 14.7% subscription rate while customers contacted through telephone has only 5.2% subscription rate.

**Inference:**

* From this bar graph, we can say that customers contacted through cellular has higher subscription rate than those contacted through telephone.

A screenshot of a cell phone

Description automatically generated

Figure no 5.5 – Default Status on Term deposit Subscription

**Observation:**

* This bar plot consists of three default status details as yes, no and unknown. While the blue shaded region shows who opened term deposit and orange shaded region shows who not opened term deposit. X-axis represents default status whereas Y-axis represents subscribers count.
* From this bar plot, we can visualize that the customers not subscribed is high in all the categories - no, unknown and yes. The category ‘no’ has the highest number of subscribers with 4197, while ‘unknown’ has 443 and ‘yes’ has none.
* We can clearly see that when customers with no or unknown default are more inclined to subscribe to term deposits. Those who have default never tends to subscribe to term deposits.

**Inference:**

From the bar plot, we can say that lesser the default, better are the chances of the customer getting subscribed to the term deposits. The people who don’t have any default have more probability of opening the term deposit since they pay the premiums on time.

A screenshot of a cell phone

Description automatically generated

Figure no 5.6 – Education details on Term deposit Subscription

**Observation:**

* This bar plot consists of different education details like university degree, high school, basic 9 years, professional course, basic 4 years, basic 6 years, unknown, illiterate. While the blue shaded region shows who opened term deposit and orange shaded region shows who not opened term deposit. X-axis represents default status whereas Y-axis represents subscribers count.
* From this bar plot, we can visualize that the customers not subscribed is high in all the categories. The highest subscription is from category university degree with 1670 subscribers and high school with 1,031 subscribers, while others have comparatively a smaller number of subscribers. basic 9 years has 595, professional course has 473, basic 4 years has 428 subscribers, basic 6 years has 251, unknown has 188 and illiterate has a meagre 4 subscribers.

**Inference:**

From this bar plot, we can infer that education is one of the key factors to decide whether the customer will be able to subscribe to term deposits. More the university degree, more is the probability of customer opening the term deposit.

A screenshot of a social media post

Description automatically generated

Figure no 5.7 – Job details on Term deposit Subscription

**Observation:**

* This bar plot consists of job details such as admin, blue-collar, technician, services, management, retired, entrepreneur, self-employment, housemaid, unemployed, student and unknown. While the blue shaded region shows who opened term deposit and orange shaded region shows who not opened term deposit. X-axis represents default status whereas Y-axis represents subscribers count.
* From this bar plot, we can visualize we can visualize that the customers not subscribed is high in all the categories. The highest subscription is from categories admin with 1,352 subscribers, blue-collar with 730 subscribers and technician with 638 subscribers, while others have comparatively a smaller number of subscribers.
* Even though retired and entrepreneur have only 323 and 275 subscribers, they have a very high percentage of people in their category compared to others.

**Inference:**

* + From the bar plot, we can infer that retired and entrepreneur customers have higher probability of opening the term deposit. Also, the probability of housemaid /unemployed/student subscribing to long term deposits is very low.

A screenshot of a cell phone

Description automatically generated

Figure no 5.8 – Martial Status details on Term deposit Subscription

**Observation:**

* This bar plot consists marital status of customers as married, single, divorced, unknown. While the blue shaded region shows who opened term deposit and orange shaded region shows who not opened term deposit. X-axis represents default status whereas Y-axis represents subscribers count.
* From this bar plot, we can visualize we can visualize that the customers not subscribed is high in all the categories. The subscription is from categories married with 2,532, single with 1,620, divorced with 476 and unknown with 12 subscribers.

**Inference:**

* From the bar plot, we can infer that married peoples has higher number of customers opening the term deposit. But single, divorced and unknown category have higher percentage of customers opening term deposits.

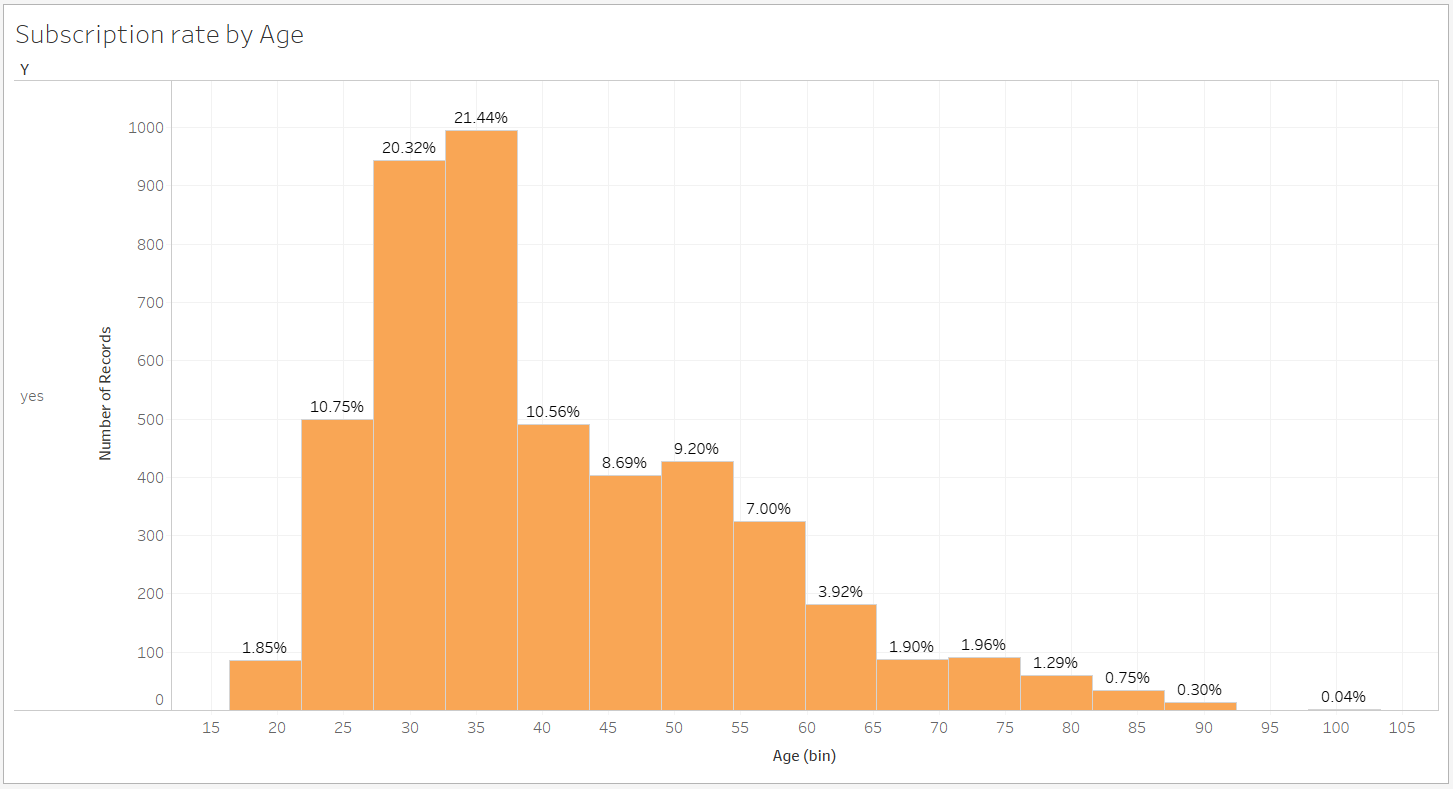


Figure no 5.9 – Subscription rate by Age

**Observation:**

* This bar plot shows the age distribution of customers who have subscribed to term deposits. X-axis represents age of the subscribers whereas Y-axis represents number of records.

**Inference:**

* From this bar plot, we can visualize that customers in the age group of 30-40 are most likely to open a term deposit.

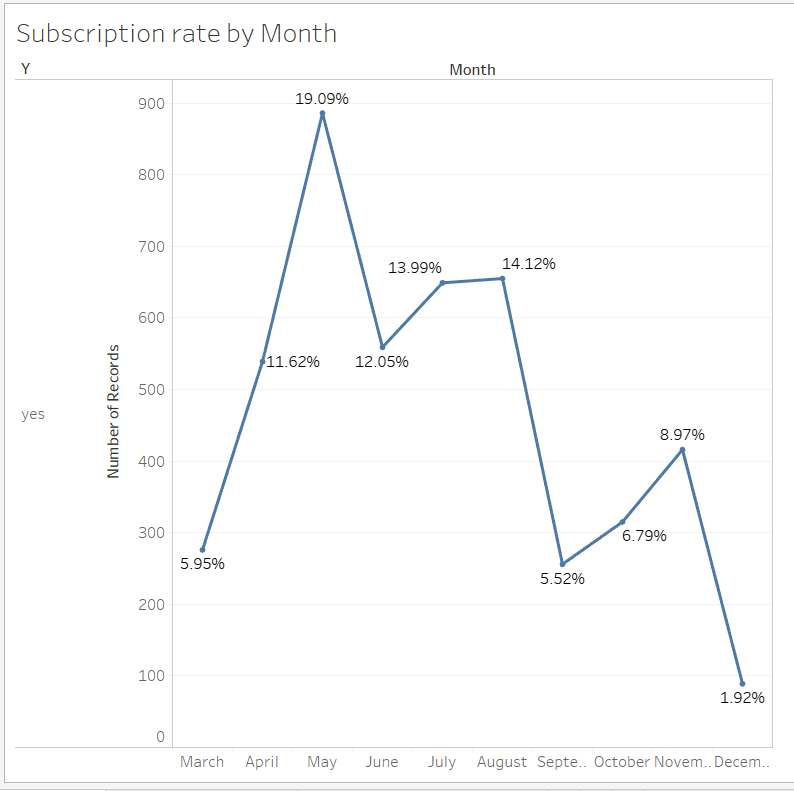


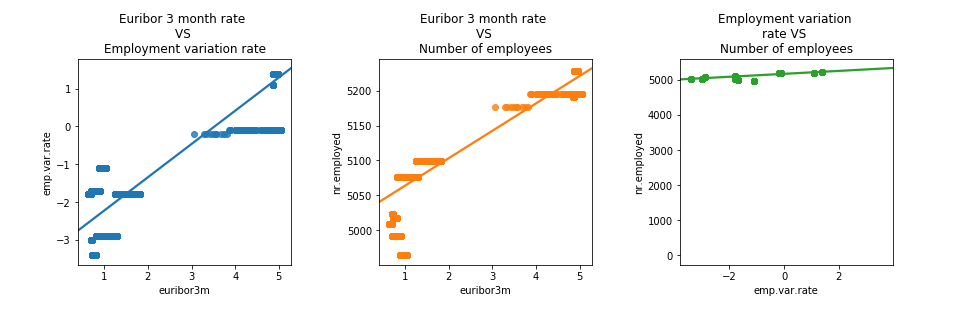
Figure no 5.10 – Subscription rate by Month

**Observation:**

* This line plot shows the months in which the customers are contacted who have subscribed to term deposits. X-axis represents month whereas Y-axis represents number of records.

**Inference:**

* From the Line graph, we can visualize that the month of May has the highest number of subscriptions.



**Figure no 5.9.1 – Regression plot between Figure no 5.9.1 – Regression plot between Figure no 5.9.1 – Regression plot between**

**Euribor 3-month rate vs Employment variation rate Euribor 3-month rate vs Number of employees Employment variation vs Number of employees**

**Inference:**

* + 5.9.1 regression plot shows there is a high correlation among Euribor 3-month rate and Employment variation with r = 0.97.
  + 5.9.2 regression plot shows there is a high correlation among Euribor 3-month rate and Number of employees with r = 0.95.
  + 5.9.3 regression plot shows there is a high correlation among Employment variation and Number of employees with r = 0.91.

**STATISTICAL SIGNIFICANCE OF ALL FEATURES WITH TARGET('Y)**

**1.AGE VS TARGET('Y'):**

* H0: Mean of customer subscribed = Mean of customer not subscribed
* HA: Mean of customer subscribed != Mean of customer not subscribed

By performing 2 sample t-test

p-value <= alpha(0.05): significant result, reject null hypothesis (H0), dependent.

**2.JOB VS TARGET('Y):** P-Proportion

* H0: P admin = P blue-collar = P technician = P services = P management = P retired = P entrepreneur = P self-employed = P housemaid = P unemployed = P student = P unknown
* HA: P admin != P blue-collar != P technician != P services != P management != P retired != P entrepreneur != P self-employed != P housemaid != P unemployed != P student != P unknown

By performing chi-square: p-value = 4.189

p-value <= alpha(0.05): significant result, reject null hypothesis (H0), dependent.

**3.MARITAL VS TARGET('Y')**

* H0: Proportion of Married = Proportion of Single = Proportion of Divorced
* HA: Proportion of Married != Proportion of Single != Proportion of Divorced

By performing chi-square: p-value = 2.068

p-value <= alpha(0.05): significant result, reject null hypothesis (H0), dependent.

**4.EDUCATION VS TARGET('Y')**

By performing chi-square: p-value =

p-value <= alpha(0.05): significant result, reject null hypothesis (H0), dependent.

**5.DEFAULT VS TARGET('Y')**

* H0 : Proportion of Yes = Proportion of No = Proportion of Unknown
* HA : Proportion of Yes != Proportion of No != Proportion of Unknown

By performing chi-square: p-value =5.161

p-value <= alpha(0.05): significant result, reject null hypothesis (H0), dependent.

**6.HOUSING VS TARGET('Y)**

* H0 : Proportion of Yes = Proportion of No = Proportion of Unknown
* HA : Proportion of Yes != Proportion of No != Proportion of Unknown

By performing chi-square: p-value = 0.058

p-value > alpha(0.05): not significant result, fail to reject null hypothesis (H0), independent.

**7.LOAN VS TARGET('Y')**

* H0 : Proportion of Yes = Proportion of No = Proportion of Unknown
* HA : Proportion of Yes != Proportion of No != Proportion of Unknown

By performing chi-square: p-value = 0.578

p-value > alpha(0.05): not significant result, fail to reject null hypothesis (H0), independent.

**8.CONTACT VS TARGET('Y')**

* H0: Proportion of Telephone = Proportion of Cellular
* HA: Proportion of Telephone != Proportion of Cellular

By performing chi-square: p-value = 1.525

p-value <= alpha(0.05): significant result, reject null hypothesis (H0), dependent.

**9.MONTH VS TARGET('Y')**

P=Proportion

* H0:P may = P july = P august = P june = P november = P april = P october = P september = P march = P December
* HA:P may != P july != P august != P june != P november != P april != P october != P september != P march != P December

By performing chi-square: p-value = 0.0

p-value <= alpha(0.05): significant result, reject null hypothesis (H0), dependent.

**10.DAY OF WEEK VS TARGET('Y')**

* H0: Proportion of Monday = Proportion of Tuesday = Proportion of Wednesday = Proportion of Thursday = Proportion of Friday
* HA: Proportion of Monday != Proportion of Tuesday != Proportion of Wednesday != Proportion of Thursday != Proportion of Friday

By performing chi-square: p-value = 2.958

p-value <= alpha(0.05): significant result, reject null hypothesis (H0), dependent.

**11.DURATION VS TARGET('Y')**

* H0 : Mean of customer subscribed Duration = Mean of customer not subscribed Duration
* HA : Mean of customer subscribed Duration != Mean of customer not subscribed Duration

By performing 2 sample t-test

p-value <= alpha(0.05): significant result, reject null hypothesis (H0), dependent.

**12.CAMPAIGN VS TARGET('Y')**

* H0 : Mean of customer subscribed Campaign = Mean of customer not subscribed Campaign
* HA : Mean of customer subscribed Campaign != Mean of customer not subscribed Campaign

By performing 2 sample t-test

p-value <= alpha(0.05): significant result, reject null hypothesis (H0), dependent.

**13.PREVIOUS VS TARGET('Y)**

* H0: Proportion of 0 = Proportion of 1 = Proportion of 2 = Proportion of 3 = Proportion of 4 = Proportion of 5 = Proportion of 6 = Proportion of 7
* HA: Proportion of 0 != Proportion of 1 != Proportion of 2 != Proportion of 3 != Proportion of 4 != Proportion of 5 != Proportion of 6 != Proportion of 7

By performing chi-square: p-value = 0.0

p-value <= alpha(0.05): significant result, reject null hypothesis (H0), dependent.

**14.P OUTCOME VS TARGET('Y')**

* H0 : Proportion of Nonexistant = Proportion of Failure = Proportion of Success
* HA : Proportion of Nonexistant != Proportion of Failure != Proportion of Success

By performing chi-square: p-value = 0.0

p-value <= alpha(0.05): significant result, reject null hypothesis (H0), dependent.

**15.EMP.VAR.RATE VS TARGET('Y')**

* H0 : Mean of customer subscribed emp.var.rate = Mean of customer not subscribed emp.var.rate
* HA : Mean of customer subscribed emp.var.rate != Mean of customer not subscribed emp.var.rate

By performing 2 sample t-test

p-value <= alpha(0.05): significant result, reject null hypothesis (H0), dependent.

**16.CONS.PRICE.IDX VS TARGET('Y')**

* H0 : Mean of customer subscribed cons.price.idx = Mean of customer not subscribed cons.price.idx
* HA : Mean of customer subscribed cons.price.idx != Mean of customer not subscribed cons.price.idx

By performing 2 sample t-test

p-value <= alpha(0.05): significant result, reject null hypothesis (H0**)**, dependent.

**17.CONS.CONF.IDX VS TARGET('Y')**

* H0 : Mean of customer subscribed cons.conf.idx = Mean of customer not subscribed cons.conf.idx
* HA : Mean of customer subscribed cons.conf.idx != Mean of customer not subscribed cons.conf.idx

By performing 2 sample t-test

p-value <= alpha(0.05): significant result, reject null hypothesis (H0), dependent.

**18.EURIBOR3M VS TARGET('Y')**

* H0 : Mean of customer subscribed euribor3m = Mean of customer not subscribed euribor3m
* HA : Mean of customer subscribed euribor3m != Mean of customer not subscribed euribor3m

By performing 2 sample t-test

p-value <= alpha(0.05): significant result, reject null hypothesis (H0), dependent.

**19.NR.EMPLOYED VS TARGET ('Y')**

* H0 : Mean of customer subscribed nr.employed = Mean of customer not subscribed nr.employed
* HA : Mean of customer subscribed nr.employed != Mean of customer not subscribed nr.employed

By performing 2 sample t-test

p-value <= alpha(0.05): significant result, reject null hypothesis (H0), dependent.

# Modeling Approach

## **Model Prerequisites done :**

* All the data were scaled using Standard scaler
* We used two methods for modelling approach:
  + Point Estimation: using Train-Test Split-(Train data -70%,Test data-30%)
  + Using Cross -validation Technique: we used K-fold cross validation with 5 splits
* Balancing of dataset

## **Before Balancing:**

A close up of a logo

Description automatically generated

Figure no 6.1 – Target column before balancing

## **1.1Train-Test Split result summarization Before balancing:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Name | Accuracy\_Train | Accuracy\_Test | precision | Recall | Roc\_AUC | F1-score |
| 1. | LOGISTIC REGRESSION | 0.93 | 0.93 | 0.53 | 0.21 | 0.60 | 0.30 |
| 2. | KNN | 0.94 | 0.93 | 0.49 | 0.18 | 0.58 | 0.26 |
| 3. | NAVIES BAYES | 0.09 | 0.08 | 0.07 | 1.00 | 0.51 | 0.13 |
| 4. | DECISION TREE ENTROPY | 1.00 | 0.92 | 0.40 | 0.43 | 0.69 | 0.42 |
| 5. | DECISION TREE GINI | 1.00 | 0.92 | 0.40 | 0.45 | 0.70 | 0.43 |
| 6. | RANDOM FOREST | 1.00 | 0.94 | 0.60 | 0.31 | 0.65 | 0.41 |
| 7. | BAGGING CLASSIFIER | 0.99 | 0.93. | 0.51 | 0.34 | 0.66 | 0.41 |
| 8. | ADABOOST CLASSIFIER | 0.93 | 0.93 | 0.51 | 0.27 | 0.62 | 0.35 |
| 9. | GRADIENTBOOST CLASSIFIER | 0.94 | 0.94 | 0.59 | 0.34 | 0.66 | 0.43 |

## **1.2. K-fold result summarization Before Balancing:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Name | BIAS ERROR | VARIANCE ERROR |
| 1. | LOGISTIC REGRESSION | 0.079730 | 0.000003 |
| 2. | KNN | 0.213391 | 0.000047 |
| 3. | NAVIES BAYES | 0.314031 | 0.000049 |
| 4. | DECISION TREE ENTROPY | 0.310452 | 0.000059 |
| 5. | DECISION TREE GINI | 0.373558 | 0.004189 |
| 6. | RANDOM FOREST | 0.061935 | 0.000017 |
| 7. | BAGGING CLASSIFIER | 0.102905 | 0.000054 |
| 8. | ADABOOST CLASSIFIER | 0.064416 | 0.000004 |
| 9. | GRADIENTBOOST CLASSIFIER | 0.055136 | 0.000013 |

## **2.After balancing using SMOTE:**

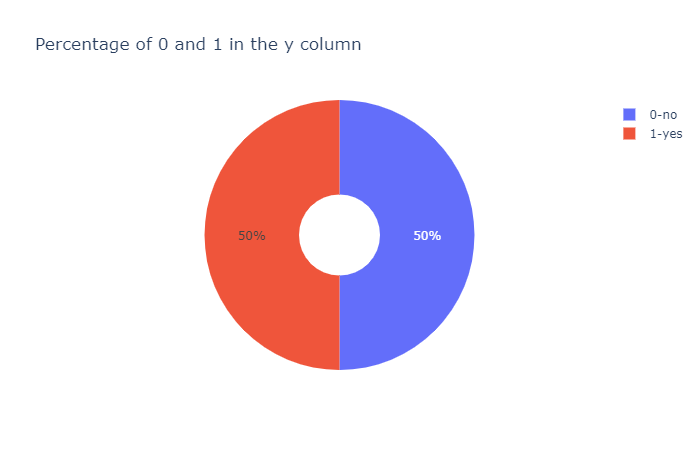


Figure no 6.2 – Target column using SMOTE balancing technique

## **2.1Train-Test Split result summarization After balancing:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Name | Accuracy Train | Accuracy Test | precision | Recall | ROC\_AUC | F1-score |
| 1. | LOGISTIC REGRESSION | 0.88 | 0.87 | 0.87 | 0.88 | 0.87 | 0.88 |
| 2. | KNN | 0.95 | 0.93 | 0.88 | 1.00 | 0.93 | 0.94 |
| 3. | NAVIES BAYES | 0.51 | 0.51 | 0.51 | 1.00 | 0.51 | 0.67 |
| 4. | DECISION TREE ENTROPY | 1.00 | 0.94 | 0.94 | 0.95 | 0.94 | 0.94 |
| 5. | DECISION TREE GINI | 1.00 | 0.94 | 0.94 | 0.94 | 0.94 | 0.94 |
| 6. | RANDOM FOREST | 1.00 | 0.97 | 0.96 | 0.98 | 0.97 | 0.97 |
| 7. | BAGGING CLASSIFIER | 1.00 | 0.95 | 0.95 | 96 | 0.95 | 0.96 |
| 8. | ADABOOST CLASSIFIER | 0.93 | 0.93 | 0.93 | 0.93 | 0.93 | 0.93 |
| 9. | GRADIENTBOOST CLASSIFIER | 0.95 | 0.95 | 0.93 | 0.96 | 0.95 | 0.95 |

## **2.2.K-fold result summarization After Balancing:**

|  |  |  |  |
| --- | --- | --- | --- |
| S.NO | Name | BIAS ERROR | VARIANCE ERROR |
| 1. | LOGISTIC REGRESSION | 0.062326 | 6.78e-06 |
| 2. | KNN | 0.025466 | 1.64e-07 |
| 3. | NAVIES BAYES | 0.058892 | 1.34e-05 |
| 4. | DECISION TREE ENTROPY | 0.058841 | 8.79e-06 |
| 5. | DECISION TREE GINI | 0.395095 | 1.45e-05 |
| 6. | RANDOM FOREST | 0.003133 | 1.56e-07 |
| 7. | BAGGING CLASSIFIER | 0.009636 | 8.53e-07 |
| 8. | ADABOOST CLASSIFIER | 0.019733 | 8.73e-07 |
| 9. | GRADIENTBOOST CLASSIFIER | 0.007946 | 5.08e-07 |

## **3.After balancing using Over Sampling Technique:**

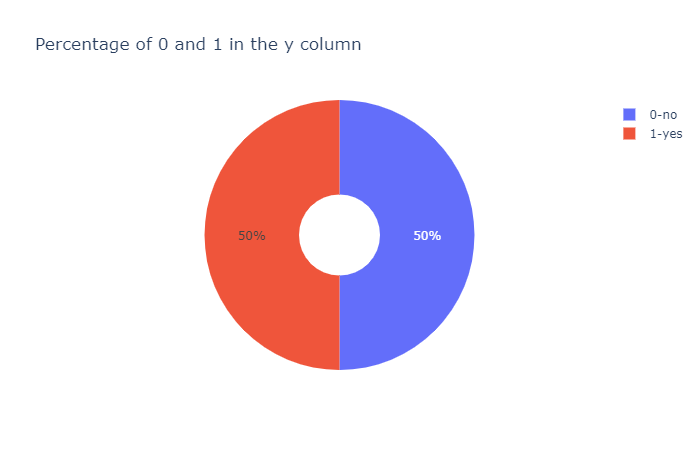


Figure no 6.3 – Target column using Over Sampling balancing technique

## **3.1Train-Test Split result summarization After balancing:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Name | Accuracy\_Train | Accuracy\_Test | precision | Recall | ROC\_AUC | F1-score |
| 1. | LOGISTIC REGRESSION | 0.88 | 0.87 | 0.87 | 0.88 | 0.87 | 0.88 |
| 2. | KNN | 0.95 | 0.93 | 0.88 | 1.00 | 0.93 | 0.94 |
| 3. | NAVIES BAYES | 0.51 | 0.51 | 0.51 | 1.00 | 0.51 | 0.67 |
| 4. | DECISION TREE ENTROPY | 1.00 | 0.94 | 0.94 | 0.95 | 0.94 | 0.94 |
| 5. | DECISION TREE GINI | 1.00 | 0.94 | 0.94 | 0.94 | 0.94 | 0.94 |
| 6. | RANDOM FOREST | 1.00 | 0.97 | 0.96 | 0.98 | 0.97 | 0.97 |
| 7. | BAGGING CLASSIFIER | 1.00 | 0.95 | 0.95 | 0.96 | 0.95 | 0.96 |
| 8. | ADABOOST CLASSIFIER | 0.93 | 0.93 | 0.93 | 0.83 | 0.93 | 0.83 |
| 9. | GRADIENTBOOST CLASSIFIER | 0.95 | 0.95 | 0.93 | 0.96 | 0.95 | 0.85 |

## **3.2.K-fold result summarization After Balancing:**

|  |  |  |  |
| --- | --- | --- | --- |
| S.NO | Name | BIAS ERROR | VARIANCE ERROR |
| 1. | LOGISTIC REGRESSION | 0.062326 | 6.7810e-06 |
| 2. | KNN | 0.025466 | 1.3433e-07 |
| 3. | NAVIES BAYES | 0.058892 | 1.3499e-05 |
| 4. | DECISION TREE ENTROPY | 0.058841 | 8.7990e-06 |
| 5. | DECISION TREE GINI | 0.395095 | 1.4556e-05 |
| 6. | RANDOM FOREST | 0.003133 | 1.5699e-07 |
| 7. | BAGGING CLASSIFIER | 0.009636 | 8.5332e-07 |
| 8. | ADABOOST CLASSIFIER | 0.019733 | 8.7505e-07 |
| 9. | GRADIENTBOOST CLASSIFIER | 0.007946 | 5.08125e-07 |

From the above results the best model is light gbm classifier because it has good generalization in terms of train and test accuracy. And the model also has low bias and low variance (good bias and variance trade off) and the recall score of the model got improved. When the model is built after target class data balancing by smote/oversampling, we get improved accuracy and recall score, but it comes at the cost of artificial values, so they are negated.

**Final Model Evaluation - Light GBM Algorithm:**

There are mainly two reasons that why we have used Light GBM Algorithm,

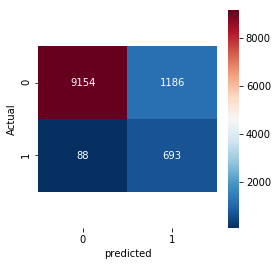
1. Balance the data – Our target column has two categories which are distributed unequally. So, we can balance the data by using parameter class\_weight in the algorithm.

2. Feature Engineering - There are two parameters in algorithm called reg\_alpha and reg\_beta for feature engineering.

We observe that mostly all algorithms having overfit and the recall score is not so high. By using this algorithm, we got better accuracy not much overfit and good recall score.

**Applying Light GBM, we get**

**CONFUSION MATRIX**



**EVALUATION METRICS:**

ACCURACY = 0.88

PRECISION = 0.36

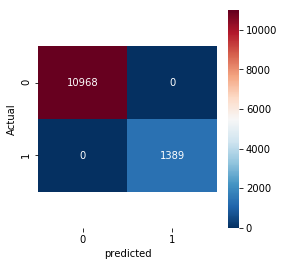
RECALL = 0.88

F1 SCORE = 0.52

ROC\_AUC = 0.88

**Removing redundant columns using VIF, we get**

**CONFUSION MATRIX**



**EVALUATION METRICS:**

ACCURACY = 1.0

PRECISION = 1.0

RECALL = 1.0

F1 SCORE = 1.0

ROC\_AUC = 1.0

# Actionable insights and recommendations to the stakeholder

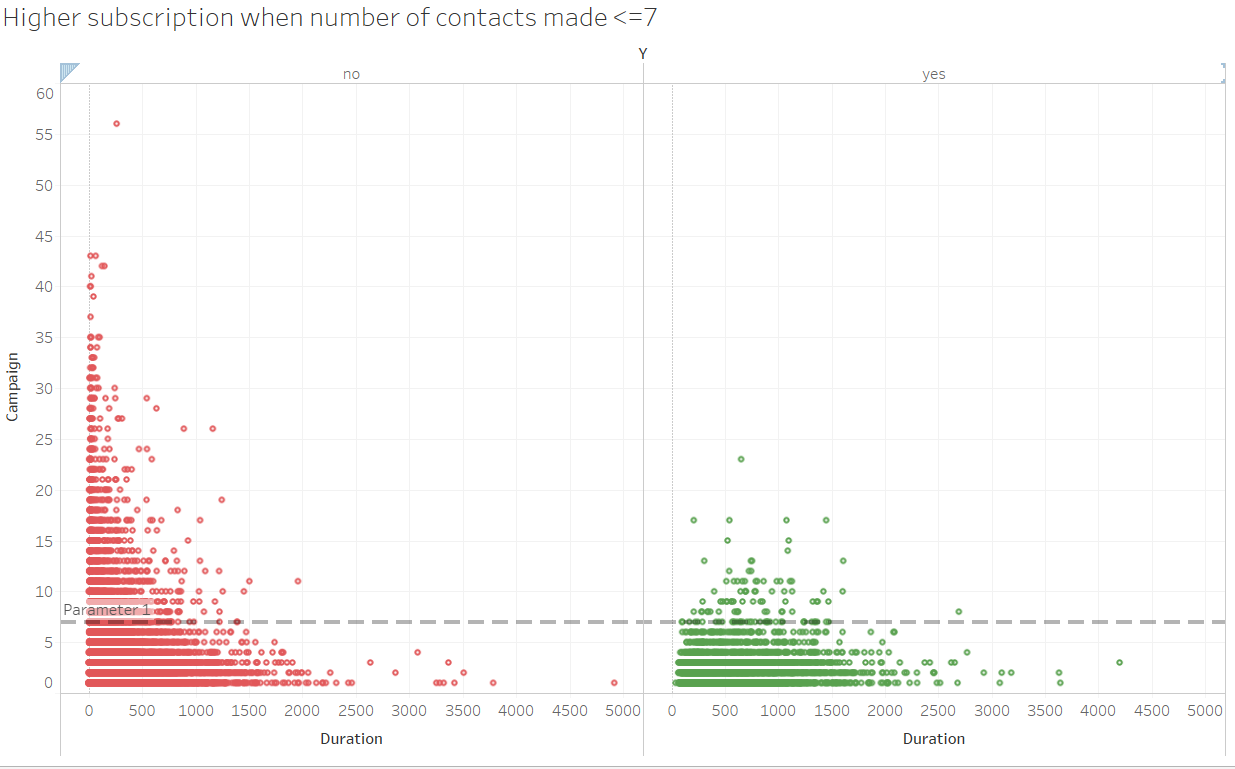


Figure no 7.1(A). Figure no 7.2(B)

Figure no 7.1 – Higher Subscription when number of contacts made <=7

* + These plots show two different categories (Yes and No) based-on number of times the customer has been contacted. X-axis represents duration of the call whereas Y-axis represents number of times the customer has been campaigned.
  + From the scatter plots, we can see that if the customer receives more than 7+ calls, the subscription rate is low. The rejection rate is higher than the success rate because a greater number of calls were made to contact the customer.
  + The number of calls is a key factor for the probability of opening the term deposit. So, we got to keep in mind not to contact a customer more than 7 times.

A screenshot of a social media post

Description automatically generated

* This bar plot shows various job roles of our customers. While the blue shaded region shows who opened term deposit and orange shaded region shows who not opened term deposit. X-axis represents default status whereas Y-axis represents subscribers count.
  + We must target retired and entrepreneur customers because they have higher percentage of customers opening the term deposit.

**Project outcome:**

* We have decided to use past data to automate this decision, instead of manually choosing through each customer. Previous campaign data shows whether the customer subscribed to term deposits as a result of that campaign or not.
* Using this, we developed various Machine Learning models to predict whether customer will subscribe to term deposits or not. Light BGM model can do this with great evaluation scores when compared to other models.
* When we implement this model to our new marketing campaign, it can efficiently target the interested customers and will be less bothering to uninterested customers.

**Business outcome:**

* Let’s assume we use 10 Tele-marketing executives to call current customers and market our term deposits. Each person has a salary of 12 Euro/hr. Each day, those 10 executives can talk approximately 40 hours.
* Previous marketing campaign has total duration of around 3,000 hours which takes around 75 to 80 days to complete. So, the maximum cost for the bank in the time period is 76,800 Euros.
* By using this model, we can target only the potential customers and can reduce the time spent on uninterested customers.

# References and Bibliography

* **Dataset Source:** <http://archive.ics.uci.edu/ml/datasets/Bank+Marketing>
* **Research Paper on this dataset:**
  + [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

**Link:**<https://www.researchgate.net/publication/256464440_A_data_mining_approach_for_bank_telemarketing_using_the_rminer_package_and_r_tool>

* **Research paper on Bank marketing Evaluation Criteria:**
  + Using AUC and Accuracy in Evaluating Learning Algorithms by Jin Huang Charles X. Ling Department of Computer Science The University of Western Ontario London, Ontario, Canada N6A 5B7.

**Link:**<https://home.cse.ust.hk/~qyang/Teaching/537/Papers/AUC-evaluation.pdf>

* **Books Referred:**
  + Applied Predictive modelling by Max Kuhn and kjell Johnson.
  + Machine Learning Mastery with python by Jason Brownlee.
* Plotly Python Open Source Graphing Library

<https://plot.ly/python/>

# Appendix

**Data Dictionary:**

* UCI-University of California, Irvine
* CSV - Comma Separated Values
* AUC - Area under the ROC Curve
* ROC - Receiver Operating Characteristic Curve
* SVM - Support Vector Machine
* IQR - Interquartile Range
* Fence low- q1-1.5\*iqr
* Fence High-q3+1.5\*iqr
* BFILL - Backward Fill Method
* SMOTE - Synthetic Minority Over-Sampling Technique
* CART - Classification and Regression Trees
* EDA - Exploratory Data Analysis