**Bank MARKETING Campaign ANALYTICS –**

**IDENTIFY TERM DEPOSIT SUBSCRIPTION PATTERNS IN BANKS**



**CAPSTONE PROJECT INTERIM REPORT**

**POST GRADUATE PROGRAM IN DATA SCIENCE AND ENGINEERING**

BY

**GROUP 4**

***Anindita Chatterjee***

***Deepshisha Singh***

***Joydeep Bose***

***Shubham Das***

***Sushovan Patra***

**EXTERNAL GUIDE**

**Ms ANJANA AGRAWAL**

**ACKNOWLEDGEMEMTS**

There are few people who deserve specific mention in this journey of ours, who have been a source of great inspiration to us all apart from our families.

We would like to express our sincere gratitude to our advisor, Ms Anjana Agrawal for guiding us throughout the journey. Ms Anjana Agrawal was always approachable and her constant encouragement and support helped us sail through this Capstone project journey very smoothly.

**ABSTRACT**

**Keywords:** direct marketing, term deposits, bank marketing

Term deposits are facing challenges from both economic pressure and marketing competition. There are a number of valuable studies concerning bank and deposit marketing. These studies mentioned the significance of customers and customer segmentation in bank and deposit marketing. However, problems like obsolescence of data, inadequate maps, lack of data and specific methods encounter in practical application of deposit market segmentation.

This research adopts data mining techniques of Logistics Regression and performs comprehensive exploratory data analysis to predict customers’ term deposit subscription behaviors and understand customers’ features to improve the effectiveness and accuracy of influential bank marketing.

**TABLE OF CONTENTS**

|  |
| --- |
| **acknowledgements** |
| **abstract** |
| **list of tables and figures** |
| **table of contents** |
| **CHAPTER 1: INTRODUCTIOn** |
| 1.1 Overview |
| 1.2 Types of marketing campaigns- mass campaigns and direct marketing |
| 1.3 Influential marketing |
| 1.4 Project outline |
| **CHAPTER 2: INDUSTRY RESEARCH AND LITERATURE REVIEW** |
| **CHAPTER 3: dataset & domain analysis** |
| 3.1 Overview |
| 3.2 Dataset |
| 3.3 Data Dictionary: Variable Categorization |
| 3.4 Data Description |
| 3.5 Data Exploration: Univariate Analysis & Transformation |
| 3.6 Data Exploration Analysis : Categorical vs Numerical |
| 3.7 Data Exploration Analysis : Categorical vs Categorical |
| 3.8 Data Exploration Analysis : Numerical vs Numerical |
| 3.9 Conclusion of Data Analysis |
| **CHAPTER 4: research methodology** |
| 4.1 Overview |
| 4.2 Statistical Test : T-Test, Z-Test for proportions, Chi-Square |
| **CHAPTER 5: model building** |
| **CHAPTER 6: references** |

**CHAPTER 1**

**Introduction**

* 1. **Overview**

Banking industry is an important sector of social economy. Bank sectors provide various products and services for clients. Deposits constitute one of the most traditional and fundamental operations of banks and meanwhile, deposits are a primary source of bank financing. There are many types of deposit accounts and some major types, including checking accounts, savings accounts, term deposit accounts and money market deposit accounts. This study will especially focus on term deposit accounts, because term deposit accounts provide bank sectors with the most stable sources of credit and profit. However, the global financial crisis in 2008 raised people’s distrusts on banks and the suspiciousness result in deposits shrank. In addition, due to the rapid development of capital market, the emergence of a large amount of financial intermediation and financial instruments provides more investment channels and opportunities for residents. Both economic pressure and marketing competition drive bank sectors to improve the effectiveness of marketing campaigns.

* 1. **Types of Marketing Campaigns: Mass Campaigns & Direct Marketing**

There are two typical marketing campaigns for companies to promote services and/or products, including mass campaigns and direct marketing. Mass campaigns aim at general indiscriminate public and direct marketing campaigns are implemented with the target of a specific group. According to a study implemented by, positive responses to mass campaigns are less than 1%; conversely, direct marketing campaigns are more effective. As a result, this research will mainly concentrate on direct marketing campaigns of term deposit accounts. Nevertheless, direct marketing might cause negative attitudes toward banks due to the intrusion of privacy. Therefore, pinpointing the target customer groups is the most important marketing strategy when adopting direct marketing campaigns. Direct marketing is one of the most effective marketing methods with an aim to maximize the customer’s lifetime value. Many cost-sensitive learning methods which identify valuable customers to maximize expected profit have been proposed.

This study will predict customers’ term deposit subscription behaviors and understand customers’ features to improve the effectiveness and accuracy of bank marketing.

**1.3 Influential Marketing**

Direct marketing campaigns promote products to potential customers by contacting them via a direct channel of communication, such as telephone or mail. The traditional practice for such a campaign typically involves the following steps. First, collect historical data from a previous campaign. Each customer record is associated with a number of individual characteristics (e.g. age, income), and a response variable indicating whether a customer responded after receiving the direct promotion. Second, perform data mining on the historical data to construct a model, with the goal of estimating the probability that a customer will respond to the promotion. Third, deploy the model to rank all potential customers in the current campaign by their estimated probability of responding.

The effectiveness of a campaign is determined by the response rate of contacted customers. This objective, which targets the most likely responders, has been widely adopted by both academia and industries An implicit assumption is that all purchases are generated by a direct contact.

Example 1. John has recently got married and the young couple has a joint account at bank X. Planning to invest for long-term, John who is very risk-averse has decided that he will invest only in term deposits of the bank as they come with higher rate of interest. The decision was made because John learnt about the interest rates on the term deposit during his visit to the bank. Applying traditional direct marketing strategies, bank X discovered that young newlyweds are more likely to respond to the direct promotion on the bank’s term deposit program. Hence, the bank sent John a brochure about its term deposit program. Though it is true that John will respond positively, he would have done so even without the promotion. In other words, contacting John added no new value to the campaign.

The example reveals two interesting points. First, certain customers base their purchase decisions on factors other than a direct marketing campaign, e.g. John. We call such customers “voluntary buyers.” Customers may become voluntary buyers due to word-of-mouth or viral marketing. The traditional strategy of maximizing the response rate aims to avoid non-buyers. When both strategies are in place, the campaign will focus on the customers who will buy if only if they are contacted. We believe that this focus is the right objective of a direct marketing campaign

**1.4 Project Outline**

Chapter 2 presents the Industry Research and Literature Review of the data mining algorithms used for direct and influencial marketing. Chapter 3 mentions the Datasets and Domain Analysis. Chapter 4 performs a statistical hypothesis testing of combination of variables. Chapter 5 mentions the model building and the results are presented. Chapter 6 concludes this research and presents the future scope of work.

**Chapter 2**

**INDUSTRY RESEARCH & Literature review**

Large amounts of data are nowadays available to companies about their customers. This data can be used to establish and maintain direct relationship with the customers in order to target them individually for specific products and services offer from the company. Large databases of customers and market data are maintained for this purpose. The customers to be targeted are selected from the database given different types of information such as demographic information and information on the customers’ personal characteristics like profession, age and sex and purchase history. Usually, the selected customers are contacted directly through one or some of the following means: personal contact, mail, e-mail, telephone and short message service (sms) to promote the new product or service. This type of marketing is called direct marketing. Among other companies, a growing number of banks and insurance companies are adopting direct marketing as their main strategy for interacting with their customers. Partly due to the growing interest in direct marketing, it has become an important application for data mining. In direct marketing, data mining has been used extensively to identify potential customers for new products.

**Data Survey Process**

Several key websites which host data like Kaggle and UCI were studied. Datasets on retail sector were studied and discussed with the the project advisor. Finally, a bank marketing dataset was selected as several researchers had focussed on direct marketing techniques for bank marketing, however only a handful of researchers had focussed on influential marketing for bank term deposits. We have performed analysis for influential marketing on this dataset.

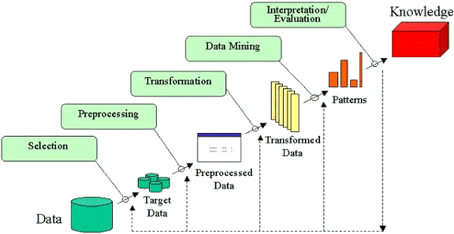
**Data Mining**

Data Mining or Knowledge Discovery in Databases can be defined as an activity that extracts some new nontrivial information contained in large databases. The goal is to discover hidden patterns, unexpected trends or other subtle relationships in the data using a combination of techniques from machine learning, statistics and database technologies. This new discipline today finds application in a wide and diverse range of business, scientific and engineering scenarios.

The overall knowledge discovery process was outlined by as an interactive and iterative process involving more or less the following steps: understanding the application domain, selecting the data, data cleaning and preprocessing, data integration, data reduction and transformation, selecting data mining algorithms, data mining, interpretation of the results and using the discovered knowledge. According to research, data mining tasks can be generally classified into two categories: descriptive and predictive. The former characterizes the general properties of the data in the database. The latter performs inference on the current data in order to make predictions.

Knowledge discovery in databases is the process of identifying valid, novel, potentially useful, and ultimately understandable patterns/models in data. Data mining is a step in the knowledge discovery process consisting of particular data mining algorithms that under some acceptable computational efficiency limitations, finds patterns or models in data.

**Figure 1: Knowledge Discovery Process**



Diagram

Description automatically generated**Figure 2: Data Mining Process**

**2.11 Research Questions**

This study will adopt data mining techniques to predict customers’ term deposit subscription behaviors and understand customers’ features to improve the effectiveness and accuracy of influential bank marketing. In order to achieve this objective, the following questions will be addressed-

* How to predict whether a bank client will subscribe to a term deposit or not?
* Which determinants would indicate a client is ready to subscribe to a term deposit through influencing marketing?
* Are there any common features of clients who have subscribed to a term deposit?

**CHAPTER 3**

**DATASET & DOMAIN ANALYSIS**

* 1. **Overview**

In this chapter, we present the results of the exploratory data analysis carried out for identifying the determinants of term deposit subscription behavior that may be applicable in this specific context of this project.

**3.2 Dataset**

The objective of the exploratory field study was to identify determinants of influential variables that impact term deposit marketing for banks. 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 assess if the product (bank term deposit) would be (or not) subscribed.

**3.3 Data Dictionary: Variable Categorization**

The influential marketing dataset used in this study was provided by Moro et al. [Moro et al., 2011], also available in [Bache, & Lichman, 2013].

It consists of 45,211 samples, each having 17 attributes, one of which, y is the target label. There are no missing values.

The attributes are both categorical and numeric and can be grouped as:

* demographical (age, education, job, marital status);
* bank information (balance, prior defaults, loans);
* direct marketing campaign information (contact type, duration, days since last contact, outcome of the prior campaign for that client, etc.)

**Table

Description automatically generated**

**Table 1: Data Description**

**3.4 Data Description**

**Table

Description automatically generated**

The output variable y shows that the dataset is unbalanced. Indeed, the successful samples corresponding to the class 'yes' are 5,289, which is 11.7% of all samples; all other samples belong to the 'no' class, which is 88.3% of the dataset.

**3.5 Data Exploration: Univariate Analysis & Transformation**

A picture containing text, receipt

Description automatically generated**JOB:**

Job is of categorical nominal data type. Here we have taken the frequency of each class occurrences and then plot it with different jobs. Here by looking at the data we can see people with admin jobs were targeted most. Although they have higher number of "yes" amount still the ration between "yes" and "no" is very low. Best ratio can be seen in people who are retired and self employed. If we consider the age data distribution it was clear that people over 60 tend to say yes more and here it has been restated.

**Chart, bar chart

Description automatically generatedTable

Description automatically generated**

Chart, box and whisker chart

Description automatically generated**AGE:**

Age is of metric discrete data type. For the first part of the descriptive analysis wee processed the data of the age column and found followings, \* Minimum Age : 17 \* Maximum Age : 98 \* Mean Age : 40 \* Standard deviation of the age : 10.4 In this section age data plot with number of "yes" and "no" responses. Here it can be clearly seen that target of the telemarketers of the bank were people from early thirties late thirties.A lthough the number of people who said yes are high in these region, the number of people who said no is also high. But when it comes to people who are above 60 has same number of "yes" and "no"s.

Mean age group of clients not subscribed 40.8 yrs

Mean age group of clients subscribed 41.7 yrs.

Median age group of clients not subscribed 39.0 yrs.

Median age group of clients subscribed 38.0 yrs.

Range of age 18 to 95 yrs

Skewness of the age column 0.68.

Chart, bar chart

Description automatically generatedAge Data is clustered around the range of 20 to 60 – hence it has few outliers as well. Log Transformation can be used to reduce the skewness of the Age Data.

The Age Data is grouped into 4 clusters:

* Cluster 1: 25 yrs to 45 yrs
* Cluster 2: 45 yrs to 65 yrs
* Cluster 3: Less than 25 yrs
* **Chart, histogram

  Description automatically generated**Cluster 4: Greater than 65 yrs

**Text

Description automatically generatedChart, box and whisker chart

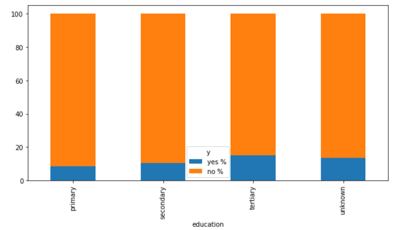
Description automatically generatedText, table

Description automatically generatedChart, bar chart

Description automatically generatedMARITAL:**

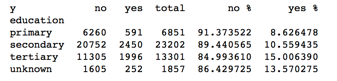
Marital status is of categorical nominal data type. Here we have taken the frequency of each class occurrences and then plot it with different relationship types Out of the people who were contacted, most of them are married people and it’s obvious because bank has contacted more people in their mid thirties. Although they have the higher "yes" percentage still the ratio between "yes" and "no" is very low. But single people and divorced people have higher ratio compared to married people

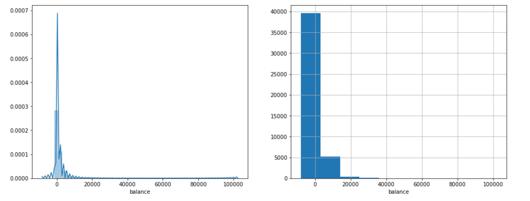
Age and Marital status have high correlation amongst themselves.

**EDUCATION:**

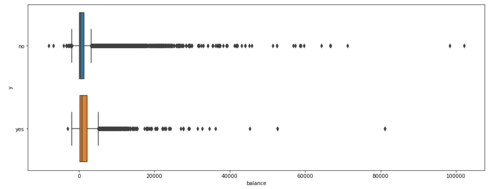
Education is of categorical ordinal data type. Here we have taken the frequency of each class occurrences and then plot it with different education levels. Fig.. Out of the people who were contacted, most of them are people who has a university level education

**Text

Description automatically generated**

**BALANCE:**

**Chart, histogram

Description automatically generated**

****Numerous customers have zero balance. The count is 3222 customers who have 0 balance in their account. These data need to be looked into during feature transformation. The Balance curve is highly scewed – may require cube root transformation later.

**TARGET:**

**Text

Description automatically generated with medium confidenceHOUSING:**

**Chart, bar chart

Description automatically generated**

**Chart, bar chart

Description automatically generatedGraphical user interface, text, application, email

Description automatically generated**

**Text, letter

Description automatically generatedLOAN:**

**Graphical user interface, text, application, email

Description automatically generated**

efault, Housing and Loan are of categorical nominal data type. Here we have taken the frequency of each class occurrences and then plot it with the corresponding feature values. Fig. 5. Loan Chart . . When considering the default data we can see there are lot more unknown data. People with no default credit has said yes and people with credit haven’t respond at all. According to the data people with default credit haven’t been taken in this campaign. When considering the Housing loan data the people with an already housing loan have said yes more. Here the selection of people to be called must had a favour for people with loans. When considering the Personal Loan Data the people with an already personal have said yes more.

**Text

Description automatically generatedChart, bar chart

Description automatically generatedText, table

Description automatically generatedLAST CAMPAIGN DATA:**

**Graphical user interface, text

Description automatically generated**

**Chart, histogram

Description automatically generatedChart, bar chart

Description automatically generatedDAY:**

**Text, table

Description automatically generatedMONTH:**

Month This is the last contact month of year. This data type is categorical nominal. Here the categories are months in theiv year. Here we have plot a graph with months and number of "yes" and "nos"s. Fig. 8. Month Chart . Here there are no any call in the first quarter of the year. But then from march to may there is a sudden spike in the number of calls. Although the number of "yes" in may is higher the ratio between the "no" is lesser. But in September and October no of "yes" and "no" are going together.

It is clearly seen that 71% of the contacts were made from May to August.

**Chart, bar chart

Description automatically generated**

**DURATION:**

**A picture containing text, receipt

Description automatically generatedChart, bar chart

Description automatically generated**Duration has been clustered into 5 sub-sets as shown.

**Chart, line chart

Description automatically generatedCAMPAIGN:**

It is seen that 87% of subscribers comes from first 3 attempts only and the no. of subscribers decreases from so on such that .55% of subscribers comes from more than 10 attempts.

If we look at the success rate it is more than 10% for first 10 attempts.

**PDAYS:**

**Chart, histogram

Description automatically generated**

For PDays less than 0, the customer was never contacted earlier and is a new customer.

Chart, histogram

Description automatically generated

If the old client is too old (>200 days), the chances of subscribing are very less.

**Chart

Description automatically generatedChart, waterfall chart

Description automatically generated**

It is visible that there is high collinearity with pdays since -1 in pday means new client which is 36954 and hence number of contacts=0 is also 36954 as before 3384 of the approached 33570 approached folks turned out to be subscribing.

The above figure clearly shows higher change of folks to subscribe in the first 3 attempts to contact.

**POUTCOME:**

**Chart, bar chart

Description automatically generatedTable

Description automatically generated**

**A picture containing background pattern

Description automatically generated**

It is clearly visible that there is high correlation with previous, pdays.

**3.6 Data Exploration: Analysis- Categorical vs Numerical**

On performing Categorical vs Numerical analysis- it was found that cellular and mobile phone connects have higher conversion rate. Customers without personal loans are more inclined towards term deposits. Similarly, customers without home loans are highly inclined to take term deposits.

In the below section, all possible combinations of the categorical vs numerical data has been visualized – to draw the above inferences.

**Chart

Description automatically generatedDAY-POUTCOME**

**Chart, waterfall chart

Description automatically generatedDAY-MONTH**

**Chart, waterfall chart

Description automatically generatedDAY-CONTACT**

**Chart

Description automatically generated**

**DAY-EDUCATION**

**Chart, bar chart

Description automatically generatedDAY MARITAL**

**Chart

Description automatically generatedDAY-JOB**

**Chart, box and whisker chart

Description automatically generatedDAY-LOAN**

**Chart, box and whisker chart

Description automatically generatedDAY-HOUSING**

**Chart, waterfall chart

Description automatically generatedDAY-DEFAULT**

**Chart

Description automatically generated with medium confidenceBALANCE-JOB**

**BALANCE-POUTCOME**

**Chart, histogram

Description automatically generated**

**BALANCE-CONTACTChart, histogram

Description automatically generated**

**BALANCE-EDUCATIONChart, histogram

Description automatically generated**

**BALANCE-MARITALChart, histogram

Description automatically generated**

**Chart, histogram

Description automatically generatedLOAN-BALANCE**

**Chart, histogram

Description automatically generatedHOUSING-BALANCE**

**DEFAULT-BALANCEChart, histogram

Description automatically generated**

**Chart, bar chart

Description automatically generatedChart, histogram

Description automatically generatedAGE-POUTCOME AGE\_CONTACT**

**AGE-EDUCATIONChart, histogram

Description automatically generated**

**JOB-AGE**

**Graphical user interface, chart, histogram

Description automatically generatedChart

Description automatically generated JOB-AGE MARITAL-AGE**

**AGE-LOANChart, histogram

Description automatically generated**

**AGE-HOUSINGChart, histogram

Description automatically generated**

**AGE-DEFAULTChart, histogram

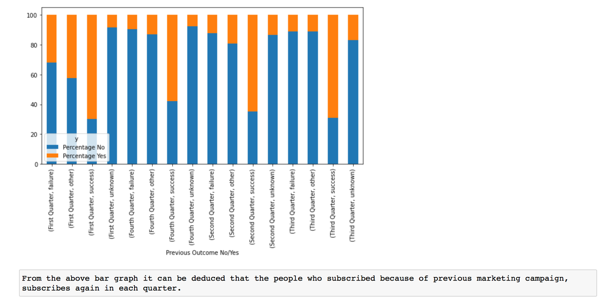
Description automatically generated**

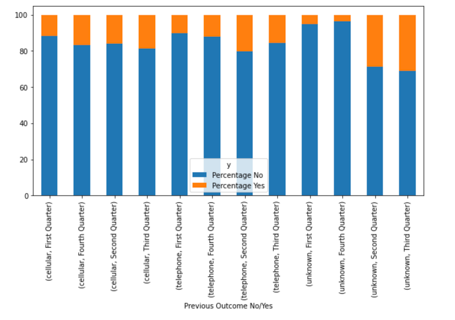
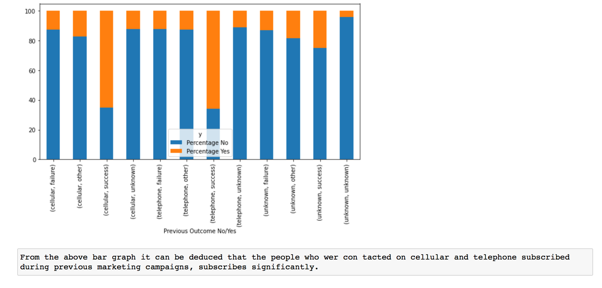
**3.7 Data Exploration: Analysis- Categorical vs Categorical**

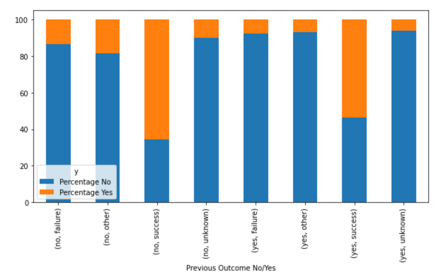
On performing Categorical vs Categorical analysis- it was found that folks who tend to subscribe in each month- are regular repeat customers. Cellular and mobile connects have higher conversion rate for term deposits. Folks who default in credit card payment do not subscribe to term deposits. An interesting study was done on marital status- although the risk appetite (Attitude to Take Risk & the Capacity to Take Risk) change post a marriage- nevertheless, the folks are indifferent when it came to re-subscription to term deposits post marriage. So, there is no significant change in the wealth planning of these indivuals post marriage- they may diversity into other funds, but this data does not suffice to indicate that. It can be safely assumed that marriage is not a significant factor in influencing term deposit subscription/ non-subscription.

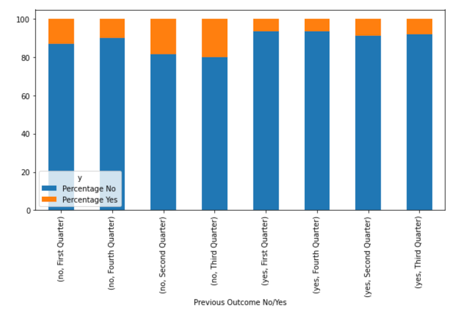
In the below section, all possible combinations of the categorical vs categorical data has been visualized – to draw the above inferences.

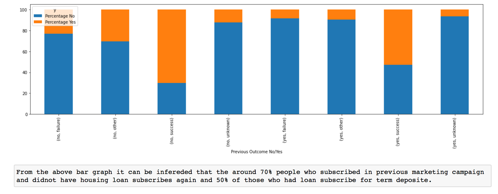
**MONTH-POUTCOME**

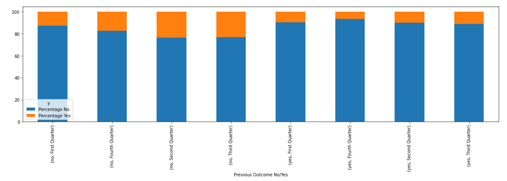
****

**CONTACT-POUTCOME**

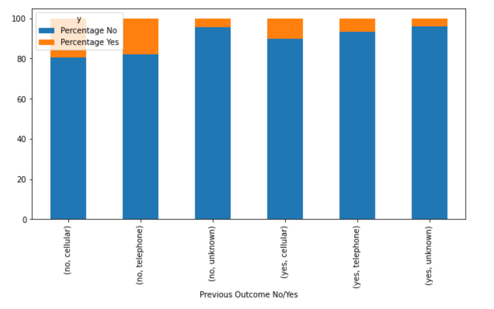
**CONTACT-MONTH**

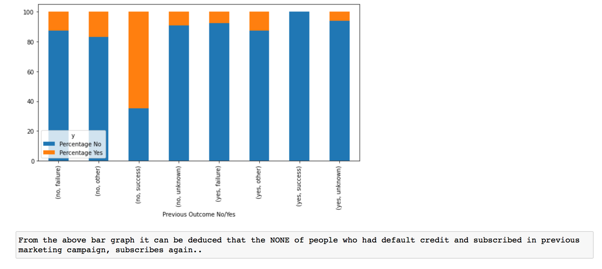
**LOAN-POUTCOME**

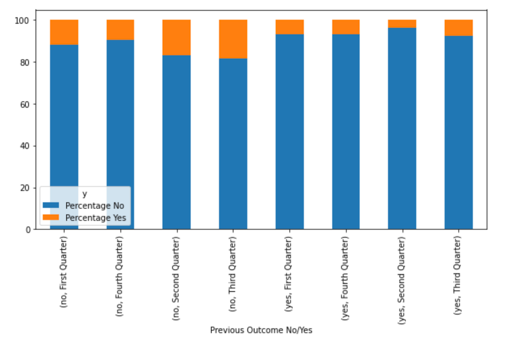
**LOAN-CONTACT**

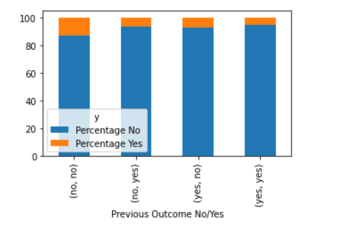
**HOUSING-POUTCOME**

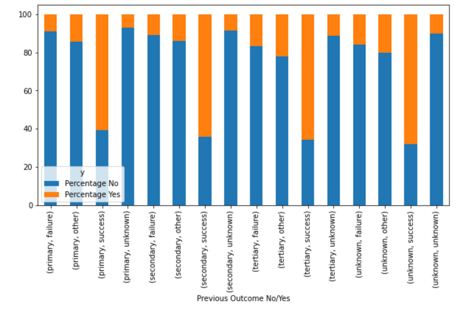
**HOUSING-MONTH**

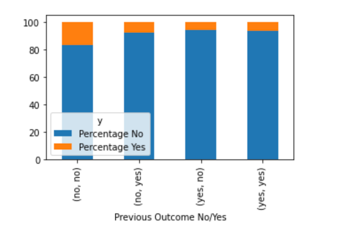
****

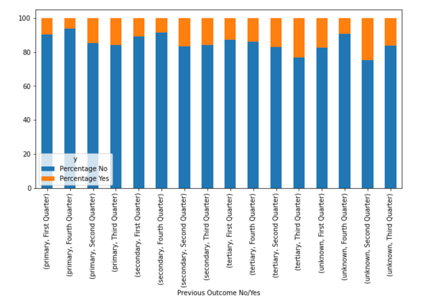
**DEFAULT-POUTCOME**

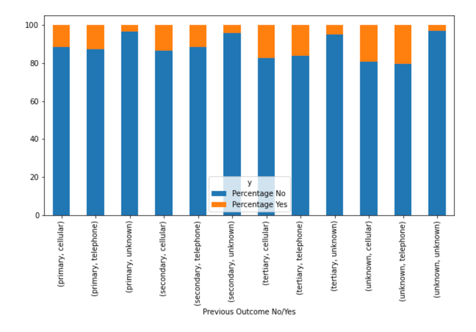
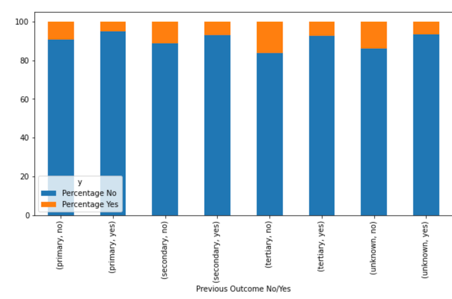
**DEFAULT-MONTH MARITAL-MONTH**

****

**MARITAL-POUTCOME**

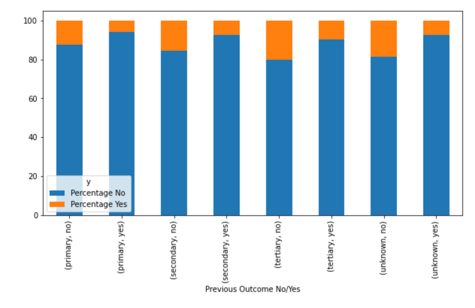
****

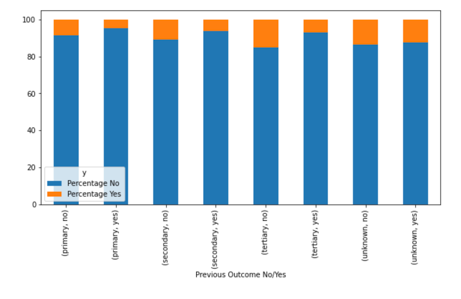
**EDUCATION-DEFAULT**

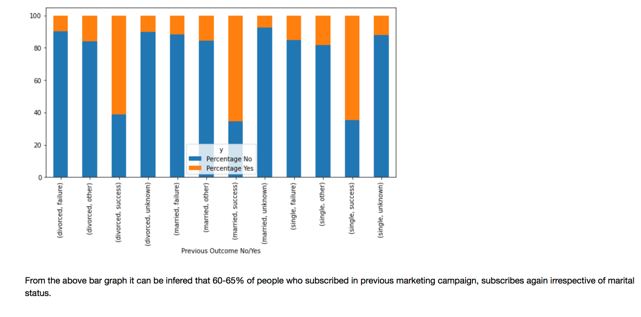
**EDUCATION-HOUSING**

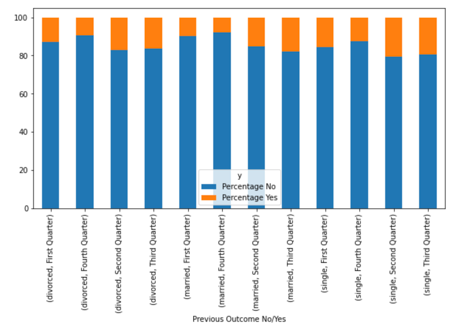
**EDUCATION-LOAN**

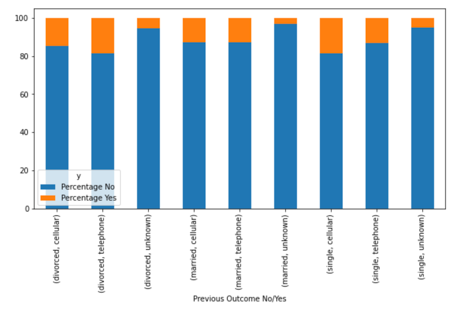
**EDUCATION-CONTACT**

**EDUCATION-MONTH**

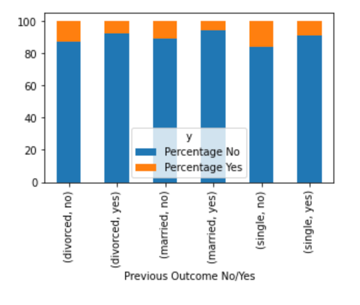
**EDUCATION-POUTCOME**

**DEFAULT-HOUSING**

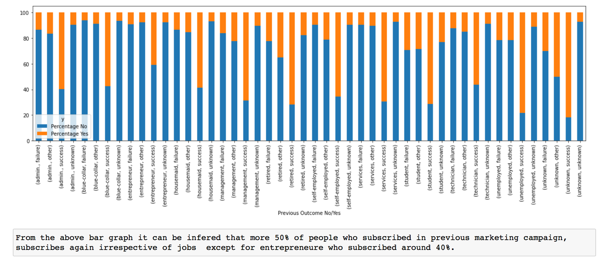
**DEFAULT-LOAN**

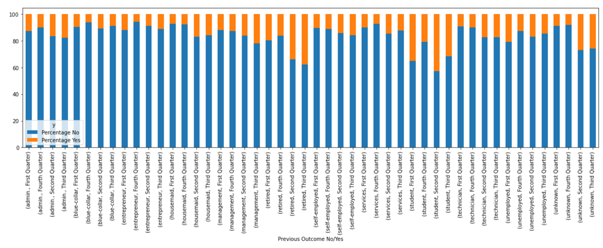
**MARITAL-CONTACT**

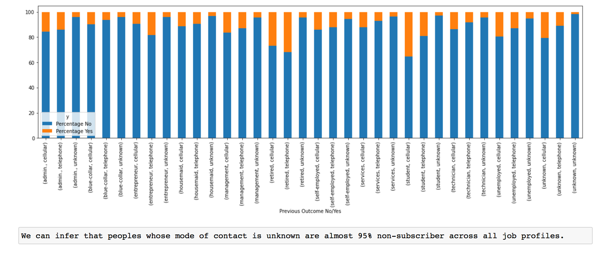
**MARITAL-LOAN**

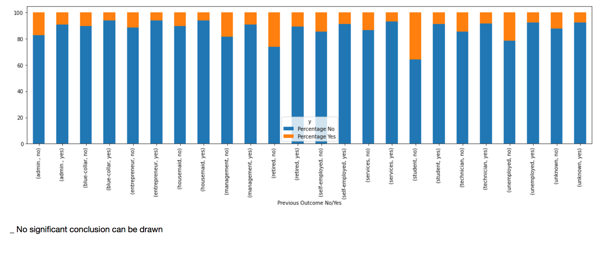
****

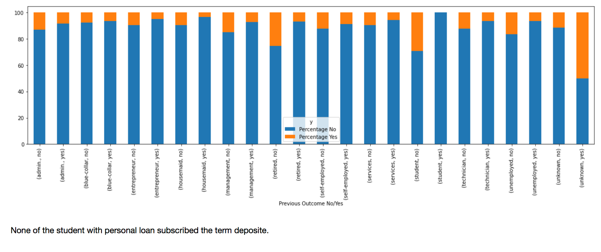
**JOB-POUTCOME**

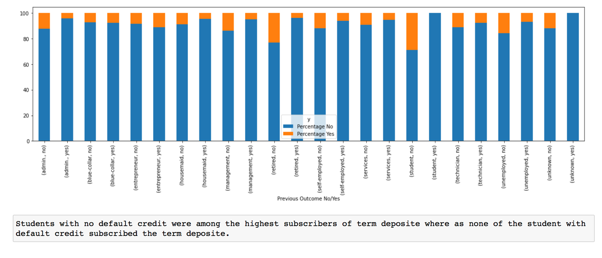
****

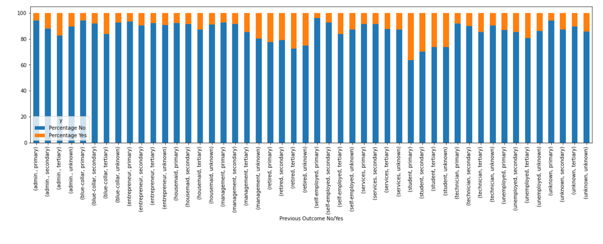
**JOB-MONTH**

**JOB-CONTACT**

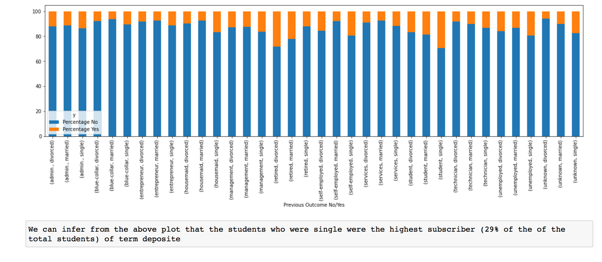
**JOB-HOUSING**

**JOB-LOAN**

**JOB-DEFAULT**

**JOB-EDUCATION**

**JOB-MARITAL**

****

**3.8 Data Exploration: Analysis- Numerical vs Numerical**

In the below section, all possible combinations of the numerical vs numerical data has been visualized – to draw the above inferences.

**Chart

Description automatically generatedPDAYS-DURATION**

**Chart, surface chart

Description automatically generated**

**DAY-DURATION**

**DURATION-AGEChart, bar chart

Description automatically generated**

**Chart, box and whisker chart

Description automatically generated**

**BALANCE-PDAYS**

**BALANCE-DAYChart, histogram

Description automatically generated**

**BALANCE-DURATIONChart, histogram

Description automatically generated**

**BALANCE-AGEA picture containing text, map, boat

Description automatically generated**

**3.9 Conclusion of Data Analysis**

To conclude this chapter, term deposits of banks are facing the challenges from both economic pressure and marketing competition. A study needs to conducted using data mining techniques to predict customers’ term deposit subscription behaviors and understand customers’ features to improve the effectiveness and accuracy of bank marketing.

Few generic inferences from this chapter: The contact duration has significant impact on the success rate of telemarketing. Secondly, the number of contacts performed during the marketing campaign should be controlled. It is better to control the number of contacts less than 3 times; otherwise, too frequent contacts may cause aversion. Thirdly, it is better to call customers’ telephone numbers (such as mobile or cellular number), however rather than their cellular, it may be worthwhile to study the impact of privacy concerns for telemarking for new customers. Fourthly, bank sectors can launch targeted marketing campaigns to attract specific customers in accordance with the results of clustering, such as children’s growth deposit scheme, retirement term deposit scheme, student term deposit scheme, housing loan deposit scheme, etc.

In the next chapter, the base model is generated by the application of various algorithms.

**CHAPTER 4**

**RESEARCH METHODOLOGY**

**4.1 Overview**

This chapter mentions the statistical analysis studied during the course of this project. Several research hypothesis were framed and tests of statistical significance performed at 95% confidence interval.

**4.2 Statistical Tests- t-tests, ztest for proportions, Chi-squares**

**Hypothesis 1**:

H0: Mean age of Subscribers equals mean age of non-subscribers / Age is Insignificant Feature.

Ha: Mean age of Subscribers is not Equal to mean age of non-subscribers / Age is Significant Feature.

mean age of Subscribers= 41.670069956513515

mean age of Non-Subscribers= 40.83898602274435

Ttest\_indResult(statistic=5.350255423036203, pvalue=8.825643691922395e-08)

Since pval<0.05 and tstat>1.96 considering 95% confidence, we reject the Null Hypothesis. Hence Ha Holds True and Age has a significant effect on Target.

**Hypothesis 2**:

H0: Mean Call Duration of Subscribers equals mean Call Duration of non-subscribers/ Call Duration is Insignificant Feature

Ha: Mean Call Duration of Subscribers not Equal to mean Call Duration of non-subscribers / Call Duration is a significant feature.

mean call duration of Subscribers= 8.954909560723484

mean call duration of NonSubscribers= 3.6863801078770093

Ttest\_indResult(statistic=91.28943612670857, pvalue=0.0)

Since pval<0.05 and tstat>1.96 considering 95% confidence, we reject the Null Hypothesis. Hence Ha Holds True and Call Duration has a significant effect on Target.

**Hypothesis 3**:

H0: Mean Balance of Subscribers equals mean Call Duration of non-subscribers/ Mean Balance is Insignificant Feature

Ha: Mean Balance of Subscribers not Equal to mean Call Duration of non-subscribers / Mean Balance is a significant feature.

mean balance of Subscribers= 1804.2679145396105

mean balance of NonSubscribers= 1303.7149691899203

Ttest\_indResult(statistic=11.25043445878562, pvalue=2.5211136918751468e-29)

Since pval<0.05 and tstat>1.96 considering 95% confidence, we reject the Null Hypothesis. Hence Ha Holds True and Balance has a significant effect on Target.

**Hypothesis 4**:

H0: Mean Positive Balance of Subscribers equals mean Positive Balance of non-subscribers/ Having Positive Balance is Insignificant Feature

Ha: Mean Positive Balance of Subscribers not Equal to mean Positive Balance of non-subscribers / Mean Positive Balance is a significant feature.

mean balance of Subscribers having balance>0 = 2006.7635262168374

mean balance of NonSubscribers having balance>0= 1604.5039524499155

no of positive Balance Subscribers= 4787

no of positive Balance NonSubscribers= 33144

Ttest\_indResult(statistic=8.036095910859034, pvalue=9.540855117273107e-16)

Since pval<0.05 and tstat>1.96 considering 95% confidence, we reject the Null Hypothesis. Hence Ha Holds True and Positive Balance has a significant effect on Target.

**Hypothesis 5**:

H0: Mean Zero Balance of Subscribers equals mean Zero Balance of non-subscribers/ Having Zero Balance is Insignificant Feature

Ha: Mean Zero Balance of Subscribers not Equal to mean Zero Balance of non-subscribers / Mean Zero Balance is a significant feature.

proportions\_ztest([4997,292],[(36700+4997),(3222+292)]

(6.508420013090475, 7.594532905140344e-11)

Since pval<0.05 and tstat>1.96 considering 95% confidence, we reject the Null Hypothesis. Hence Ha Holds True and having Zero Balance has a significant effect on Target.

**Hypothesis 6**: Job effect to Target

H0: Job Type is Insignificant feature to Target.

Ha: Job Type is significant feature to Target.

stats.chi2\_contingency(pd.crosstab(df['job'],df['y']))

tstat: 775.380810637894 pvalues: 4.040337361025556e-160

Since pval<0.05 and tstat>1.96 considering 95% confidence, we reject the Null Hypothesis. Hence Ha Holds True and Job Type has a significant effect on Target.

observed-Expected

y no yes Expected\_no Expected\_yes

job

admin. 4540 631 4566.071576 604.928424

blue-collar 9024 708 8593.503882 1138.496118

entrepreneur 1364 123 1313.043596 173.956404

housemaid 1131 109 1094.938842 145.061158

management 8157 1301 8351.557718 1106.442282

retired 1748 516 1999.146402 264.853598

self-employed 1392 187 1394.280994 184.719006

services 3785 369 3668.045122 485.954878

student 923 303 1082.576630 143.423370

technician 6757 840 6708.266440 888.733560

unemployed 1101 202 1150.568800 152.431200

**Hypothesis 7**: Marital effect to Target

H0: Marital Type is Insignificant feature to Target.

Ha: Marital Type is significant feature to Target.

stats.chi2\_contingency(pd.crosstab(df['marital'],df['y']))

tstat: 196.49594565603957 pvalues: 2.1450999986791792e-43

Since pval<0.05 and tstat>1.96 considering 95% confidence, we reject the Null Hypothesis. Hence Ha Holds True and Job Type has a significant effect on Target.

observed-Expected

y no yes Expected\_no Expected\_yes

marital

divorced 4585 622 4597.860123 609.139877

married 24459 2755 24030.375528 3183.624472

single 10878 1912 11293.764349 1496.235651

**Hypothesis 8**: Education effect to Target

H0: Education Type is Insignificant feature to Target.

Ha: Education Type is significant feature to Target.

stats.chi2\_contingency(pd.crosstab(df['education'],df['y']))

tstat: 204.35576231814383 pvalues: 4.2141504857990294e-45

Since pval<0.05 and tstat>1.96 considering 95% confidence, we reject the Null Hypothesis. Hence Ha Holds True and Education Type has a significant effect on Target.

observed-Expected

y no yes Expected\_no Expected\_yes

education

primary 7865 843 7689.296322 1018.703678

secondary 20752 2450 20487.718564 2714.281436

tertiary 11305 1996 11744.985114 1556.014886

**Hypothesis 9**: Credit Card Default to Target

H0: Default (having credit card by default Y/N) is Insignificant feature to Target.

Ha: Default (having credit card by default Y/N) is significant feature to Target.

proportions\_ztest([5237,52],[(39159+5237),(763+52)

(4.766917467205618, 1.8706589343075603e-06)

Since pval<0.05 and tstat>1.96 considering 95% confidence, we reject the Null Hypothesis. Hence Ha Holds True and Default (having credit card by default Y/N) has a significant effect on Target.

**Hypothesis 10**: Housing Loan to Target

H0: Having housing loan (Y/N) is Insignificant feature to Target.

Ha: Having housing loan (Y/N) is significant feature to Target.

proportions\_ztest([3354,1935],[(16727+3354),(23195+1935)

(29.592122568437706, 1.887098648854402e-192)

Since pval<0.05 and tstat>1.96 considering 95% confidence, we reject the Null Hypothesis. Hence Ha Holds True and Having housing loan (Y/N) has a significant effect on Target.

**Hypothesis 11**: Personal Loan to Target

H0: Personal loan (Y/N) is Insignificant feature to Target.

Ha: Personal loan (Y/N) is significant feature to Target.

proportions\_ztest([4805,484],[(33162 +4805),(6760 +484)

(14.498101054828755, 1.2454800673760236e-47)

Since pval<0.05 and tstat>1.96 considering 95% confidence, we reject the Null Hypothesis. Hence Ha Holds True and Personal loan (Y/N) has a significant effect on Target.

**Hypothesis 12**: Mode of Contact to Target

H0: Mode of Contact is Insignificant feature to Target.

Ha: Mode of Contact is significant feature to Target.

stats.chi2\_contingency(pd.crosstab(df['contact'],df['y']))

tstat: 1035.714225356292 pvalues: 1.251738325340638e-225

Since pval<0.05 and tstat>1.96 considering 95% confidence, we reject the Null Hypothesis. Hence Ha Holds True and Personal loan (Y/N) has a significant effect on Target.

observed-Expected

y no yes Expected\_no Expected\_yes

contact

cellular 24916 4369 25859.099998 3425.900002

telephone 2516 390 2566.042158 339.957842

unknown 12490 530 11496.857844 1523.142156

**Hypothesis 13**: Month to Target

H0: Monthly quarters is Insignificant feature to Target.

Ha: Monthly quarters is significant feature to Target.

stats.chi2\_contingency(pd.crosstab(df['quarters'],df['y']))

tstat: 440.8126749659026 pvalues: 3.1900158096027513e-95

Since pval<0.05 and tstat>1.96 considering 95% confidence, we reject the Null Hypothesis. Hence Ha Holds True and Monthly quarters have a significant effect on Target.

observed-Expected

y no yes Expected\_no Expected\_yes

quarters

First Quarter 12137 1584 12115.851496 1605.148504

Fourth Quarter 19991 2048 19460.771892 2578.228108

Second Quarter 4096 826 4346.200792 575.799208

Third Quarter 3698 831 3999.175820 529.824180

**Hypothesis 14**: poutcome to Target

H0: Previous marketing campaign is Insignificant feature to Target.

Ha: Previous marketing campaign is significant feature to Target.

proportions\_ztest([3384,1905],[(33570+3384),(6352+1905)])

(-35.56502186559602, 4.868264925584431e-277)

Since pval<0.05 and tstat>1.96 considering 95% confidence, we reject the Null Hypothesis. Hence Ha Holds True and previously marketing campaign have a significant effect on Target.

observed-Expected

y no yes Expected\_no Expected\_yes

poutcome

failure 4283 618 3770.273949 1130.726051

other 1533 307 1415.487465 424.512535

success 533 978 1162.392152 348.607848

unknown 3 2 3.846433 1.153567

**Hypothesis 15**: PreviousDays to Target

H0: Number of days to last contact is Insignificant feature to Target.

Ha: Number of days to last contact is significant feature to Target.

proportions\_ztest([3384,1905],[(33570+3384),(6352+1905)])

(-35.56502186559602, 4.868264925584431e-277)

Since pval<0.05 and tstat>1.96 considering 95% confidence, we reject the Null Hypothesis. Hence Ha Holds True and Number of days to last contact has a significant effect on Target.

**CHAPTER 5**

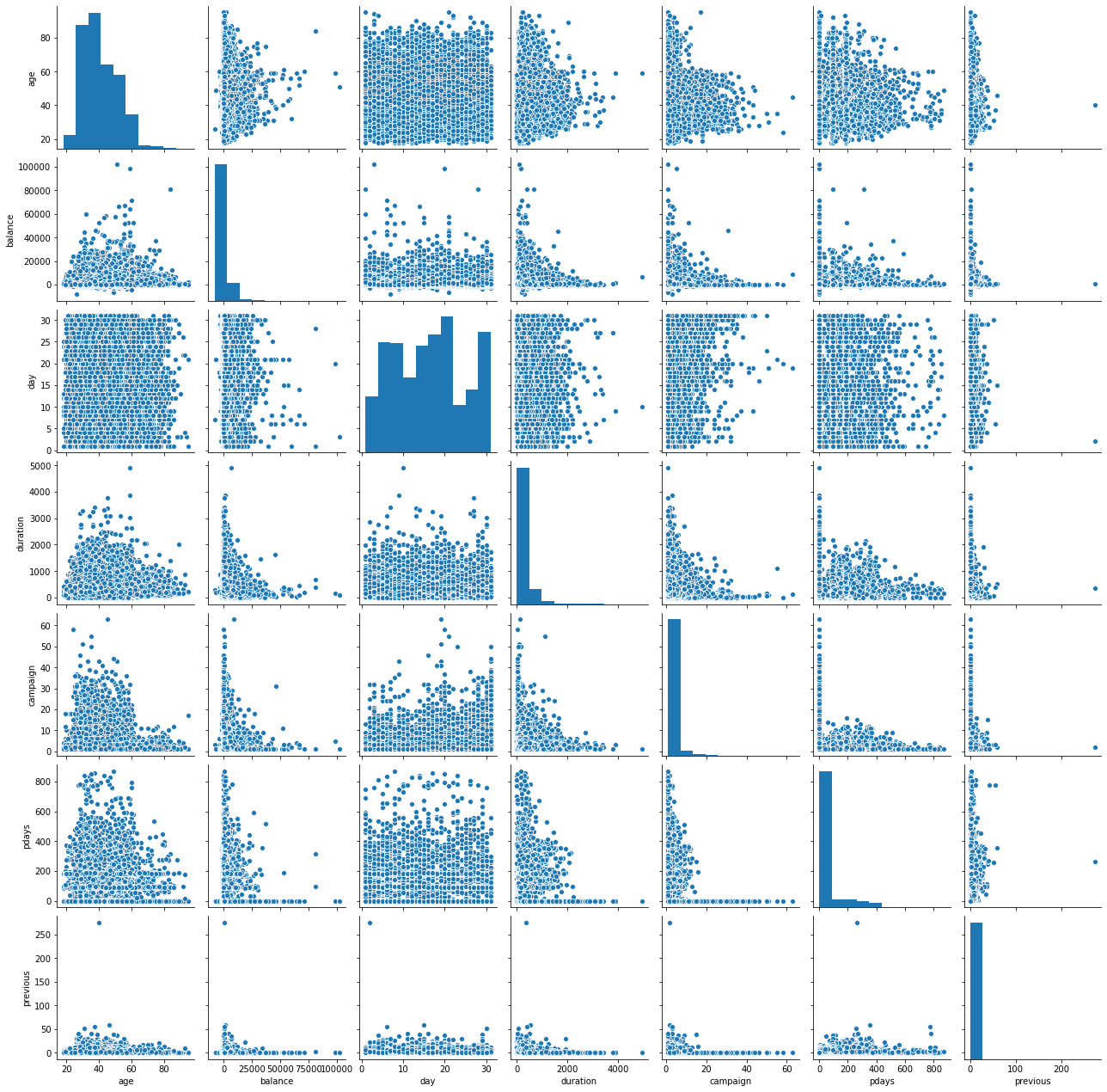
**MODEL BUILDING**

**5.1 Overview**

This chapter mentions the Logistics Regression model used to study this data.

**5.2 Reducing Skewness of the variables**

* *Reducing Skewness in Age using log transformation:*
* *Reducing Skewness of balance using cubic transformation:*
* *Reducing Skewness of duration using cubic transformation:*

****

**Preprocessing** There are NO missing values .The data set contains a feature ‘duration’ which is the last contact duration, in seconds. Since the value is not known before a call is performed, and since the output value (yes / no) is known once the call is ended, this feature is removed for the predictive modelling. The labels (non-numerical) are encoded to numerical labels The data set consists of features with different data types; numeric and categorical. The numerical features were standardize by removing the mean and by scaling to unit variance (StandardScaler) There was one feature that can be considered as having Categorical Ordinal data type; education. This feature was converted to numerical representation using OrdinalEncoder (this results in single column of integers per feature. Integers are 0 to (no of categories - 1). All the other categorical features were nominal. So they were encoded to numeric using OneHotEncoder. The features cannot be converted to numeric values since these values do not have an order. Therefore, here a binary column is created for each nominal value. These nominal and categorical preprocessing is added to a pipeline, where the classifier is added as the next step.

**Target of Model :** High sensitivity and low false negative ratio (high negative pred-value) as to reduce the cost of misclassifying a potential subscriber as non-subscriber.

As Our Base Model, the Classification Algorithm considered is Logistic Regression, which is statistical model that in its basic form uses a logistic function to predict our binary dependent variable(whether a client will subscribe to term deposit{‘yes’ or ‘no’})

Based on Cross-validation scores, other algorithm considered are:

**❏ Decision Tree Classifier: This is a tree based model where the data are split according to the given parameters. For this problem, Gini impurity is used as the criteria to measure quality of a split, best split as the strategy to choose the split, maximum depth as 21 with 45 leaf nodes at most , minimum number of samples required to split as 4 and Complexity parameter used for Minimal Cost-Complexity Pruning as 0.002. (ccp-alpha)**

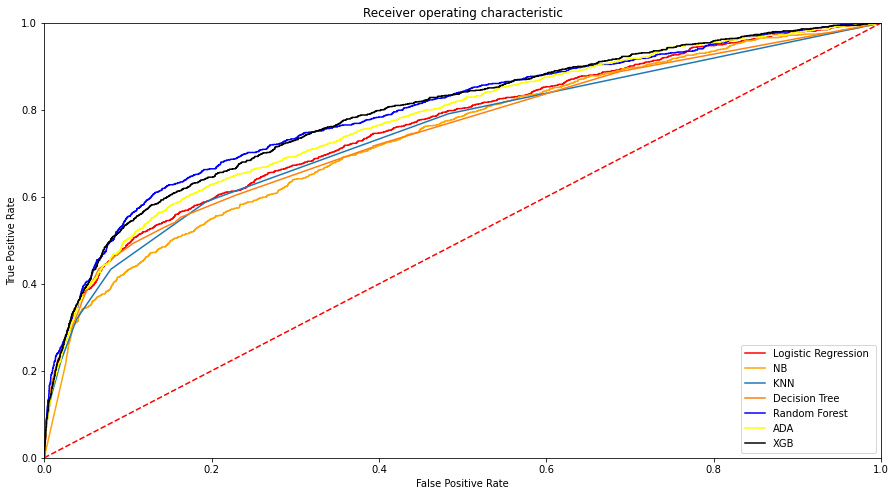
**❏ K Nearest Neighbors Classifier: This is a neighbor based method. This works directly on learned samples instead of creating rules. This method assigns the class of the majority of its k neighbors as the class of the test instance. For this problem, the best k value was 10.**

**❏ Naïve Bayes Classifier: This is a probabilistic classifier based on Bayes theorem of probability. Here, Gaussian is used as the probability distribution.**

**❏ XG Boost: is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. For this Problem the hyper parameters considered after tuning are n\_estimators =700, learning\_rate =0.01 ,gamma=.5, max\_depth=7, colsample\_bytree=0.6**

**❏ AdaBoost : classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult 7cases. For this problem the hyper parameters considered post tuning, n estimators 250 and Learning\_rate=1.**

**❏ random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. For this problem, Entropy as impurity is used as the criteria to measure quality of a split, with max features to be SQRT along with 250 estimators. maximum depth as 45 and minimum number of samples required to split as 8**



|  |  |
| --- | --- |
| **Algorothm** | **Crossval Scores Before Over Sampling** |
| **Logistic Regression** | **0.8925292452599043** |
| **KNN Classifier** | **0.8855894104376232** |
| **Decision Tree Classifier** | **0.8928886629611823** |
| **Random Forest Classifier** | **0.8931375108885063** |
| **ADA Boost Classifier** | **0.7790282636040845** |
| **XGBoost Classifier** | **0.8905108863992796** |
| **Naïve Bayes** | **0.8543464295083194** |

**Performance Evaluation :** Since this is an imbalanced data set, confusion matrix, precision, recall and F1 score are used along with the accuracy score, since accuracy score alone can be misleading. In this table, each cell for Precision, Recall and F1 score gives the scores for label ‘no’ and ‘yes’ respectively. F1 Score is used to select the best classifier since these scores are for an imbalanced data set and weighted average of F1 Score is used since it accounts this imbalance in its calculation. According to these results, accuracy score is proportional to the F1 Score weighted average. This is because the class imbalance was taken into consideration.

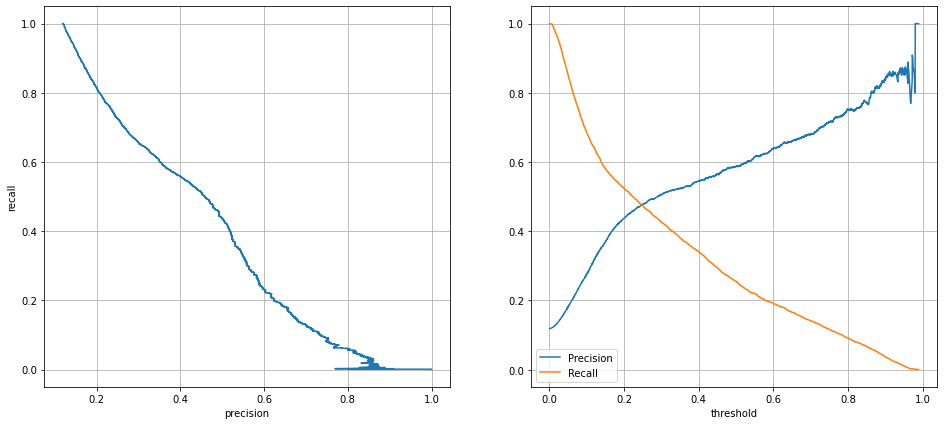
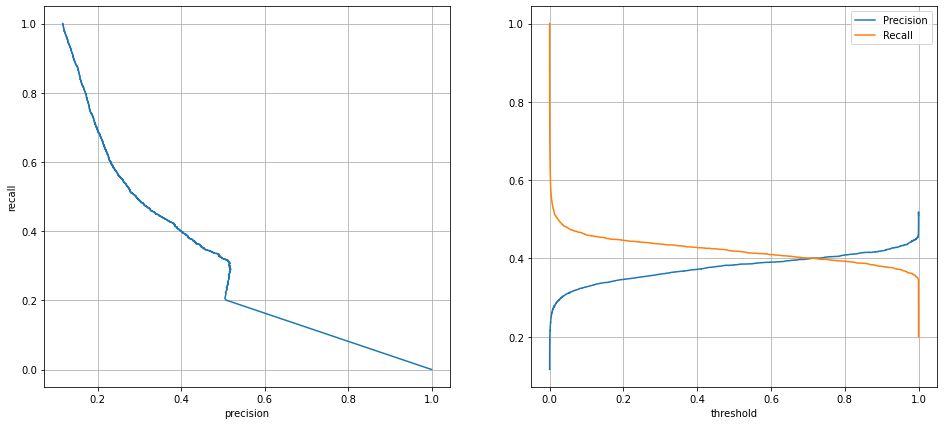
According to the above results, the highest F1 Score weighted average is scored by Naïve Bayes. The 2nd and 3rd best classifiers are Random Forest Classifier and XG Boost Classifier. Based on these results, considering f1 score , Roc Score and accuracy score we considered Naïve Bayes to be the algorithm used for this Model .

In Further Steps Since we are Not satisfied with the prediction of the Model , Over sampling is Going to be used for generating best prediction Model.

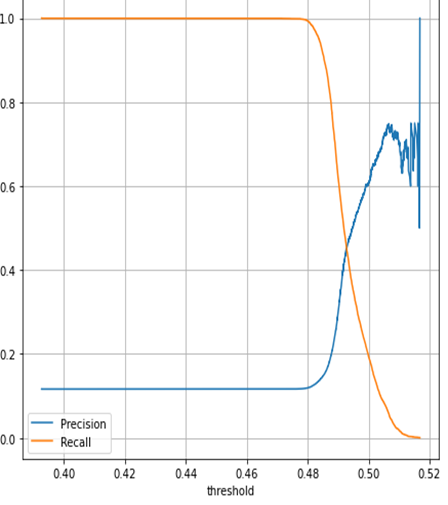
|  |  |  |  |
| --- | --- | --- | --- |
| **Classifier** | **Precision** | **Recall** | **Avg F1 Score** |
| **Logistic Regression** | 0.90  0.68 | 0.99  0.18 | 0.61 |
| **KNN Classifier** | 0.90  0.59 | 0.98  0.21 | 0.58 |
| **Decision Tree Classifier** | 0.91  0.31 | 0.90  0.33 | 0.62 |
| **Random Forest Classifier** | 0.90  0.66 | 0.98  0.23 | 0.64 |
| **ADA Boost Classifier** | 0.90  0.63 | 0.98  0.20 | 0.62 |
| **XGBoost Classifier** | 0.90  0.62 | 0.98  0.24 | 0.63 |
| **Naïve Bayes** | 0.92  0.39 | 0.91  0.42 | 0.66 |

|  |  |  |
| --- | --- | --- |
| **Classifier** | **Roc Auc Score** | **Accuracy score** |
| **Logistic Regression** | 0.7599 | 0.8906 |
| **KNN Classifier** | N/A |  |
| **Decision Tree Classifier** | 0.7450 | 0.8357 |
| **Random Forest Classifier** | 0.7927 | 0.8970 |
| **ADA Boost Classifier** | 0.7750 | 0.8891 |
| **XGBoost Classifier** | 0.7980 | 0.8909 |
| **Naïve Bayes** | 0.7357 | 0.8520 |

Precision Recall VS Threshold curve

****🡨NB.

XGB🡪

RF****

**Random Oversampling: Randomly duplicate examples in the minority class. Random oversampling involves randomly selecting examples from the minority class, with replacement, and adding them to the training dataset. They are referred to as “*naive resampling*” methods because they assume nothing about the data and no heuristics are used. This makes them simple to implement and fast to execute, which is desirable for very large and complex datasets.**

**Over Fitting Of Decision tree and Random Forest Algorithm :**

**❏ Decision Tree Classifier: For this oversampling problem, Gini impurity is used as the criteria to measure quality of a split, best split as the strategy to choose the split, minimum number of samples required to split as 2 and it is found that or Decision Tree model is Prone To Over-Fit.**

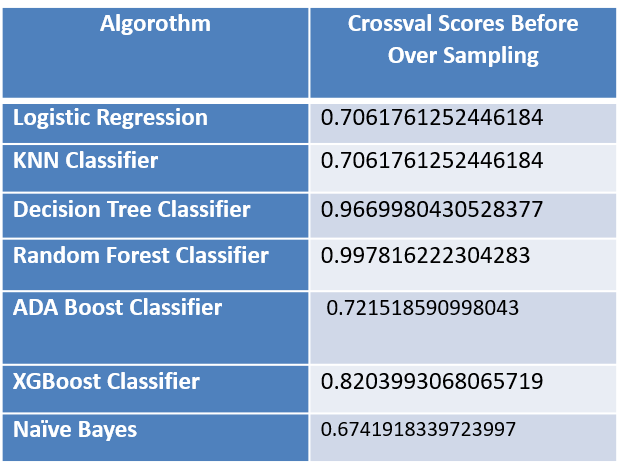
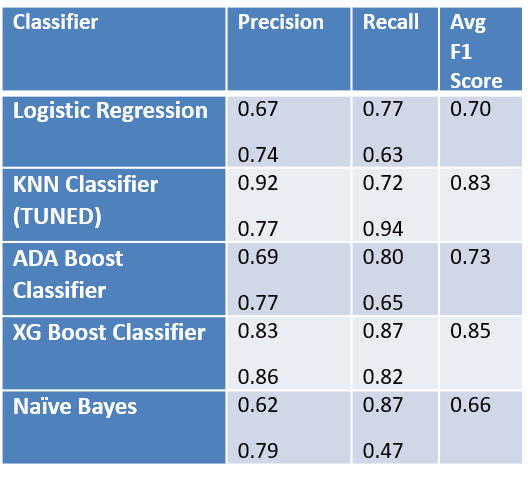
**❏ random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. For this problem, Gini as impurity is used as the criteria to measure quality of a split, with 100estimators. maximum depth as none and minimum number of samples required to split as 2. and it is found that or Decision Tree model is Prone To Over-Fit.**

**NOTE: Decision Tree and Random Forest Classifier shows very high precision and recall scores subjective to overfitting the data hence we can say that Decision tree and random forest Classifier DoesNot applies to this Model**



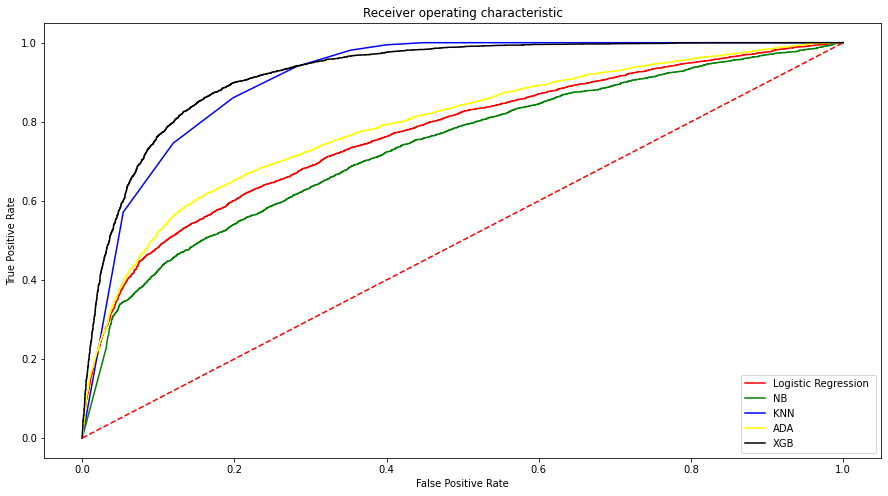
**❏ XG Boost classifier : a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework, For this case of over sampling Dataset the hyper parameters considered post tuning by randomized Search CV , n estimators 700 , Learning\_rate=0.1, gamma=1.5, max\_depth=5, subsample=0.8.**

**❏ AdaBoost : classifier is a meta-estimator that corrects weights of estimators to adjust co-eff. For this problem the hyper parameters considered post Randomized Search CV tuning, n estimators 750 and Learning\_rate=1.**

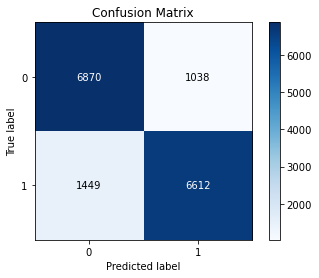
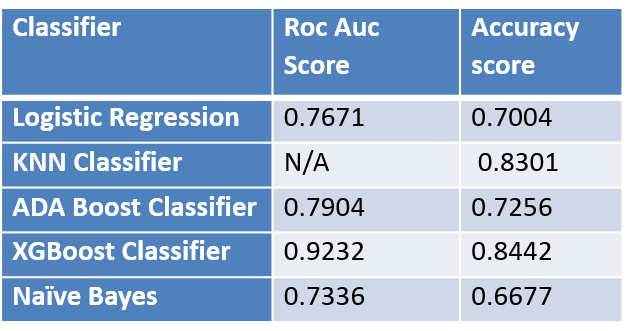
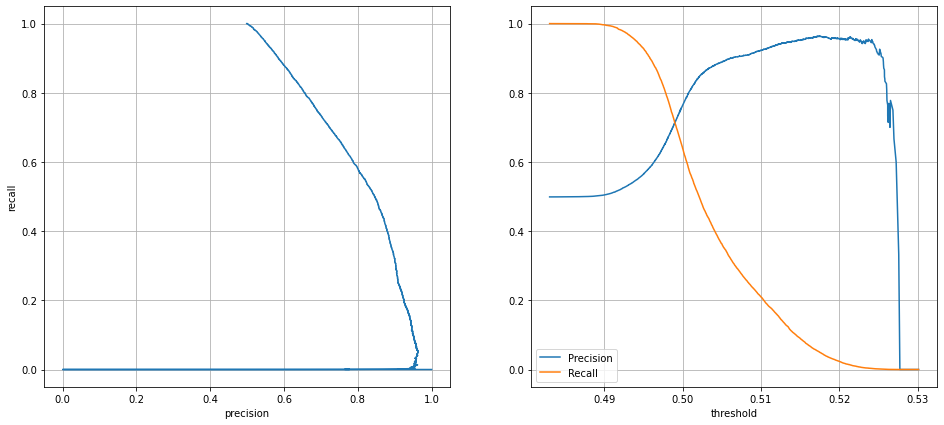
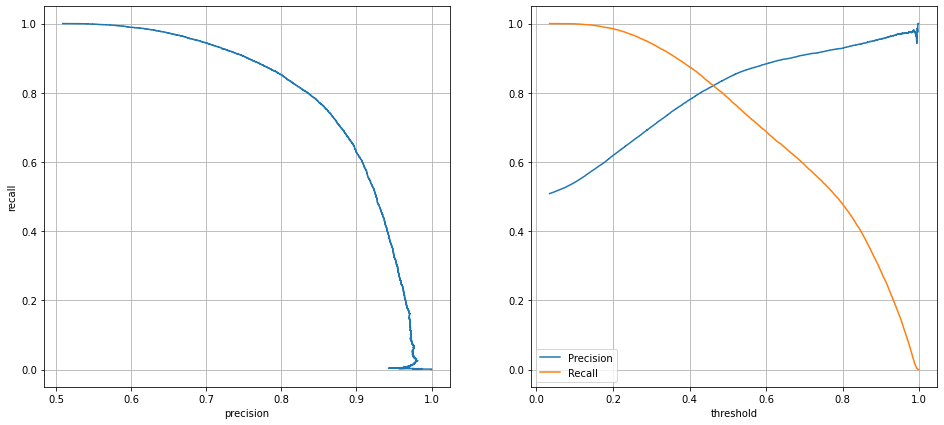
**❏ K Nearest Neighbors Classifier: This is a neighbor based method. For this problem, the best k value was 7. Model is prone to overfit on changing weights =‘Distance’ and algorithm =‘Brute’**

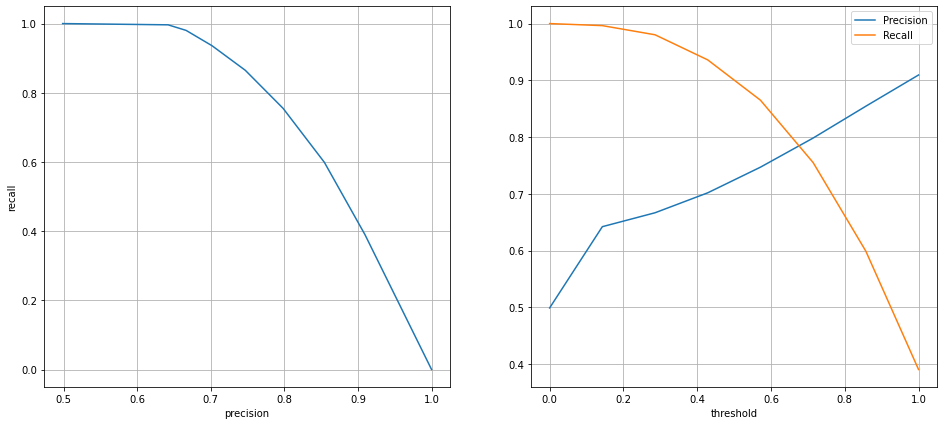
**Algorithms used in Bank Marketing**

**2.6 Logistic Regression**



**Performance Evaluation** : Since this was an**imbalanced** **data** **set**, and then after Oversampling y (yes/no) are in same proportion, confusion matrix, precision, recall and F1 score are used along with the accuracy score, to select the best model.

In This Case based on Precision, Recall, F1 score, Accuracy score and Roc Curve we can clearly identify that our first best Model is XG Boost Classifier while 2nd and 3rd best classifiers are KNN Classifier and ADA Boost Classifier.

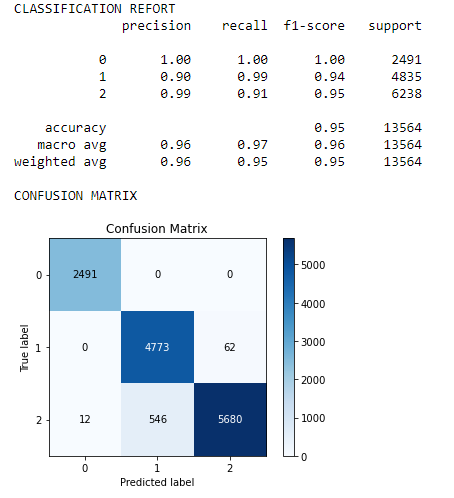
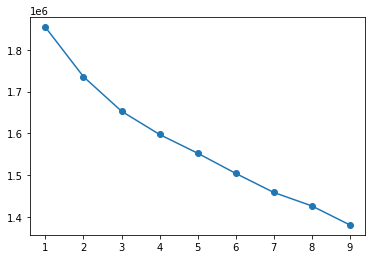


**Clustering:**

Unsupervised Learning Employed to access the accuracy of the data set and analyze the possible classes expected,

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. In this topic, we will learn what is K-means clustering algorithm, how the algorithm works, along with the Python implementation of k-means clustering. In our Problem statement obtained Elbow curve pointed out 3-4 clusters moving ahead with 3 cluster’s , then adding those labels as 0,1,2 and training on our best model for dataset without over sampling (Naïve Bayes )we reached the following classification matrix.

We can conclude that the outcome of bank Marketing is not enclosed to Yes/No , more diversity is present

****

**CHAPTER 6**

**REFERENCES**

|  |  |
| --- | --- |
| Original owner of data | Paulo Cortez (Univ. Minho) and Sérgio Moro (ISCTE-IUL) @ 2012 |
| Reference | 1. Bank Management -Marketing <https://www.tutorialspoint.com/bank_management/bank_management_marketing.htm> 2. <https://towardsdatascience.com/machine-learning-case-study-a-data-driven-approach-to-predict-the-success-of-bank-telemarketing-20e37d46c31c> |
| Link to web page | <https://archive.ics.uci.edu/ml/datasets/bank+marketing> |