

**BANK MARKETING**

**PREDICTION**

**Post Graduate Program in Data Science Engineering**

Location: **Bangalore** Batch:**PGPDSE Nov21**

**Submitted by**

Kilani Teja

Vimal Kanth

Mohammed Alzaman Nafisuddin Siddiqui

Rani Priya S

Swapnil Pattanshetty

**Mentored by**

Anjana Agarwal

**ACKNOWLEDGEMENT**

Any endeavor in a specific field requires the guidance and support of many people for successful completion. The sense of achievement on completing anything remains incomplete if the people who were instrumental in its execution are not properly acknowledged. We would like to take this opportunity to verbalize our deepest sense of indebtedness to our project mentor, Anjana Agarwal, who was a constant pillar of support and continually provided us with valuable insights to improve upon our project and make it a success. Further, we would like to thank our parents for encouraging us and providing us a platform wherein we got an opportunity to design our own project.

**DECLARATION**

We hereby declare, that this submission is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Table of Contents

[ACKNOWLEDGEMENT](#page2) [2](#page2)

[DECLARATION](#page3) [3](#page3)

[INTRODUCTION](#page6) [6](#page6)

[Dataset Information](#page6) 7

[Variable Categorization with Description](#page8) [8](#page8)

[Numerical](#page8) [8](#page8)

[Categorical](#page9) [9](#page9)

Problem Statement [10](#page10)

[Target Variable](#page10) [10](#page10)

[Methodology to be Followed](#page11) 10

[DATA PRE-PROCESSING](#page12) 11

[Datatype Verification](#page12) [12](#page12)

[Missing Value Treatment](#page14) 13

[Check for Outliers](#page15) 13

[EXPLORATORY DATA ANALYSIS](#page16) 14

[Univariate Analysis](#page16) 14

[Bi-Variate Analysis](#page26) 20

[Correlation Matrix](#page32) 25

[Statistical Tests](#page40) 26

[BASE MODEL](#page43) 28

[Decision Treel](#page43) 29

[MODEL BUILDING AND METHODS](#page45)

[Encoding of Categorical Variables and Numerical Values Treatment and Feature Engineering](#page45)

[Model Building](#page46)

[Model Understanding](#page47),[Precision and Recall Trade-Off](#page50),features importance,Roc curve,hyperparameter tuning,Accuracy Score.

[COMPARISON AND IMPLICATIONS](#page52)

[Comparison to Benchmark](#page53)

[Model Explanation](#page55)

[Inferences and Recommendations](#page55)

Limitation 42

Deployment....................................................................................................................................45

Conclusion.....................................................................................................................................49

**INTRODUCTION**

A Current practices:

The BANKING industry is an important sector of the 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 distrust of banks and the speciousness resulted in deposits shrank. In addition, due to the rapid development of the capital market, the emergence of a

large amount of financial inter mediation 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.

Background:

Bank direct marketing:

There are two main approaches for enterprises to promote products and/or services: through mass campaigns, targeting the general indiscriminate public, or directed marketing, targeting a specific set of contacts (Ling and Li 1998). Nowadays, in a globally competitive world, positive responses to mass campaigns are typically very low, less than 1%, according to the same study. Alternatively, directed marketing focus on targets that assumable will be keener to that specific product/service, making this kind of campaign more attractive due to its efficiency (Ou et al. 2003). Nevertheless, directed marketing has some drawbacks, for instance, it may trigger a negative attitude towards banks due to the intrusion of privacy ( Luding 2003). It should be stressed that due to internal competition and the current financial crisis, there are huge pressures for European banks to increase a financial assets. To solve this issue, one adopted

strategy is to offer attractive long-term deposit applications with good interest rates, in particular by using directed marketing campaigns. Also, the same drivers are pressing for a reduction in costs and time. Thus, there is a need for an improvement in efficiency: lesser contacts should be done, but an approximate number of successes (clients subscribing to the deposit) should be kept.

A sector that plays a very significant part in the Commercial and Economic backdrop of any country is the banking sector. Data Mining techniques can play a key role in providing different methods to analyze data and find useful patterns and extract knowledge in this sector (Vajiramedhin and Suebsing; 2014). Data mining helps in the extraction of useful information from the data (Turban et al.; 2011). According to (Venkatesh and Jacob; 2016), machine learning has more capability to gather information from the data, which results in the more frequent use of data mining methods in the banking sector. Due to the large amount

of data gathered in banks, data warehouses are required to store this data. Analyzing and identifying patterns from such data can be useful for Banks to identify trends and acquire knowledge from these data. With the acquired knowledge from these data, organizations can more clearly understand their customers and improve the services they provide. Such an understanding of data can help organizations gain success and improve the decision support system. As stated by (Raorane and Kulkarni; 2011), customers behavior must be understood by any organization to improve its business.

(Moro et al.; 2013) says, Analyzing the bank’s information and understanding the regular patterns can help banks to give better administer to their clients. From the Bank Telemarketing information, different examples can be dissected, and learning can be extricated to give better consumer loyalty and to make significant strides towards Mining valuable data from the information. As stated by (Keller and Kotler ; 2015), to enhance any business, advertising efforts assume an essential part in drawing in the clients to the administrations given by the associations.

**Dataset Information**

The dataset consists of 20 variables. Out of these variables 19 are independent variables and 1 is a target variable. The variables are a mixture of both numerical and categorical type. We divided the data into 4 groups as follows:

Bank client data:

|  |  |  |
| --- | --- | --- |
| VARIABLE | DATATYPE | DESCRIPTION |
| Age | Int | Client age |
| Job | Categorical | Type of job (categorical:"admin.","blue-collar",  "entrepreneur","management","retired","self-employed","services","housemaid"  "student","technician","unemployed","unknown") |
| Marital | Categorical | Marital status  (categorical:"Divorced","Married","Single","Unknown";  Note: "Divorced" Means divorced or widowed) |
| Education | Categorical | Education(categorical:"basic.4y""basic.6y","basic.9y", "high.school","illiterate","professional.course",  "university.degree","unknown") |
| Default | Categorical | Has credit in default?(categorical:"No","Yes","Unknown") |
| Housing | Categorical | Housing: Has housing loan?  (categorical:"No","Yes","Unknown") |
| Loan | Categorical | Has personal loan?(categorical:"No","Yes","Unknown") |

Related with the last contact of the current campaign:

|  |  |  |
| --- | --- | --- |
| Contact | Categorical | Contact communication type  (categorical:"Cellular","Telephone") |
| Month | Categorical | Contact month of year  (categorical: "Jan", "Feb", "Mar", ..., "Nov", "Dec") |
| Day\_of\_week | Categorical | Last contact day of the week (categorical:  "Mon","Tue","Wed","Thu","Fri") |
| Duration | Numeric | Last contact duration, in seconds |

Other attributes:

|  |  |  |
| --- | --- | --- |
| Campaign | Numeric | Number of contacts performed during this campaign  and for this client |
| Pdays | Numeric | Number of days that passed by after the client was  last contacted from a previous campaign (numeric;  999 means client was not previously contacted) |
| Previous | Numeric | Previous Numeric  Number of contacts performed before this campaign  and for this client (numeric) |
| Poutcome | Categorical | Outcome of the previous marketing campaign  (categorical: "Failure","Nonexistent","Success") |

Social and economic context attributes:

|  |  |  |
| --- | --- | --- |
| Emp.var.rate | Numeric | Employment variation rate - quarterly indicator |
| Cons.price.idx | Numeric | Consumer price index - monthly indicator |
| Cons.conf.idx | Numeric | Consumer confidence index - monthly indicator |
| Euribor3m | Numeric | Euribor 3 month rate - daily indicator |
| Nr.employed | Numeric | Number of employees - quarterly indicator |

Output variable (desired target):

|  |  |  |
| --- | --- | --- |
| Y | Binary | Has the client subscribed a term deposit  ?(binary:"Yes","No") |

**Problem Statement**

1. What is the main marketing campaign factor that can increase the customer’s decision to subscribe to a term deposit?

2. How accurate can we be in predicting the customer's decision to subscribe to a term deposit?

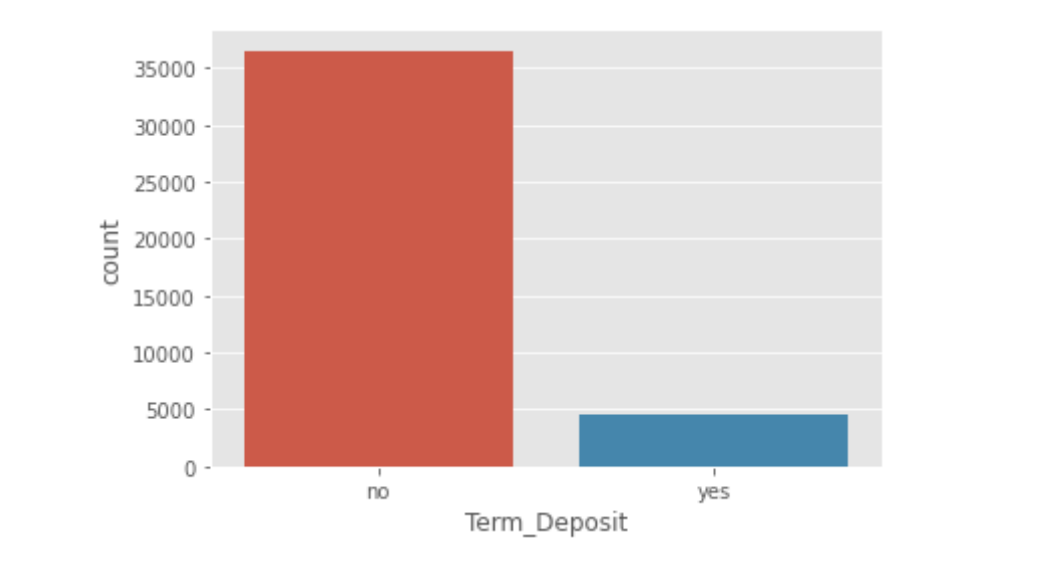
3. Business interpretation of the different models using Visualization

4. Business evaluation to convince that our model predicts the best.

Target Variable

The target variable of the above dataset is Y. We have to predict whether a customer is going to

subscribe to a term deposit? .



From the above plot we can observe that our dataset is highly imbalanced. Majority of the

data points belong to no class. Ratio of No class to yes class is 8:1.**We observe that there**

**is there is presence of moderate amount of class imbalance**.

**Methodology to be Followed**

CRISP-DM which stands for Cross Industry Standard Process for Data Mining is a methodology created to help shape data mining projects. It describes the different phases/tasks involved in the project and provides an overview of data mining life cycle.

1. **Business Understanding** - It focuses on determining the business requirements/objectives and understanding what outcome to achieve. Also determine the business units being affected. Convert this business problem into a data mining problem and carve out an initial plan.
   * Determine the business objectives: Understand what is needed to be accomplished for the customer.
   * Assess situation: Determine resources availability, project requirements, assess risks and contingencies, and conduct a cost-benefit analysis.
   * Determine data mining goals: Convert business problem to a data mining problem and recognize the data mining problem type such as classification, regression or clustering, etc.
   * Produce a project plan: Devise a step-to-step plan for executing the project.
2. **Data understanding -** This phase starts with collecting the data and then examining the data for its surface properties like data format, number of records, etc. The next step is to better understand the data by understanding each attribute and perform basic statistics on them. Understand the relationship between different attributes. Determine the quality of data by checking the missing values, outliers, duplicates, etc.
   * Collect initial data: Acquire the data and load it into the analysis tool to be used.
   * Describe data: Examine the data and document its surface properties like data format, number of records, or field identities. Understand the meaning of each attribute and attribute value in business terms. For each attribute, compute basic statistics so as to get a higher-level understanding.
   * Explore data: Find insights from the data. Query it, visualize it, and identify relationships among the data.
   * Verify data quality: Identify special values, missing attributes and null data. Determine how clean/dirty is the data.
3. **Data preparation -** This stage, which is often referred to as data wrangling, has the objective to develop the final data set for EDA and modelling. Covers all activities to construct the final dataset from the initial raw data. Some of the tasks include table, record and attribute selection as well as transformation and cleaning of data for modelling tools.
   * Select data: Determine which attributes/features will be used and document reasons for inclusion/exclusion.
   * Clean data: Correct, impute and remove the improper data.
   * Extract data: Derive new attributes from the existing ones
   * Integrate data: Create features by combining data from multiple sources.
   * Format data: Re-format data as necessary. For example, convert string values to numeric values so as to perform mathematical operations.
4. **Modelling -** In this stage we build and assess different models built using various techniques from the training dataset.
   * Select modelling technique: Determine the algorithms to be used to model the data based on the business requirement.
   * Generate test design: In order to build and test the model, we need to divide the dataset into training and testing data set. In this step we divide the data into train and test data set.
   * Build model: Based on the modelling technique selected, build the model on the input data set.
   * Assess model: Compare the results of different models based on confusion matrix. The outcome of this step frequently leads to model tuning iterations until the best model is found.
5. **Evaluation -** Evaluate the models and review the steps executed to construct the model to be certain it properly achieves the business objectives.
   * Evaluate results: Understand the data mining results and check how impactful they are in achieving the data mining goal. Select appropriate model based on confusion matrix.
   * Review process: Review the work accomplished and make sure that nothing was overlooked and all steps were properly executed. Summarize the findings and correct anything if needed.
   * Determine next steps: Based on the previous three tasks, determine whether to proceed to deployment, iterate further, or initiate new projects.

**DATA PRE-PROCESSING**

Data pee-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data pee-processing task.

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data per-processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

The dataset consists of 21 variables. Out of these variables 20 are independent variables and 1 is a target variable. The variables are a mixture of both numerical and categorical type.

**Datatype Verification**

We first check the data types of each of the columns of the data.

|  |  |
| --- | --- |
| Attribute | DataType |
| Age | int64 |
| Job | object |
| Marital | object |
| Education | object |
| Default | object |
| Housing | object |
| Loan | object |
| Contact | object |
| Month | object |
| Day\_of\_week | object |
| Duration | int64 |
| Campaign | int64 |
| Pdays | int64 |
| Previous | int64 |
| Poutcome | object |
| Emp.var.rate | float64 |
| Cons.price.idx | float64 |
| Cons.conf.idx | float64 |
| Euribor3m | float64 |
| Nr.employed | float64 |
| Term\_Deposit | object |

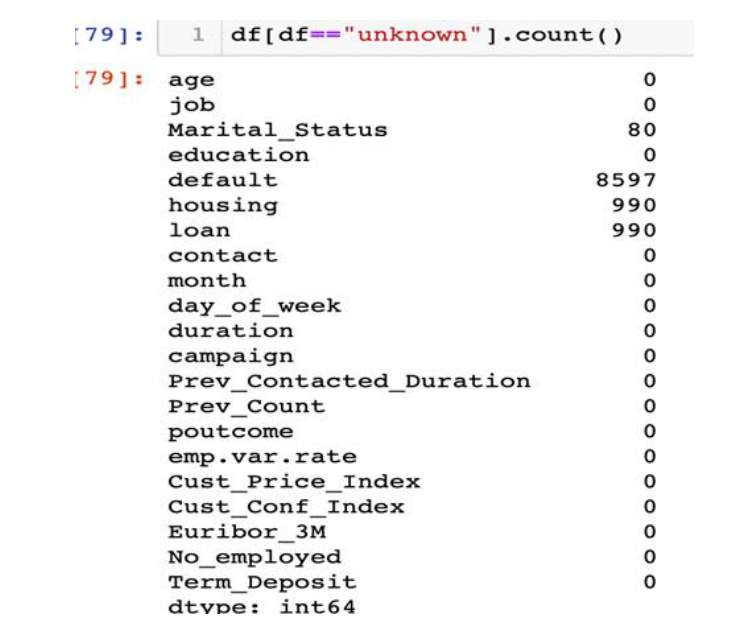
Missing Value Treatment

The next step of data pre-processing is to handle missing data in the datasets. If our dataset contains some missing data, then it may create a huge problem for our machine learning model. Hence it is necessary to handle missing values present in the dataset.

We have have unknown value in the data set according to they have below table we have taken care of them like mode imputation techniques:

|  |  |
| --- | --- |
| Attribute | Null Value Percentage |
| housing | 0.0240361 |
| Marital\_status | 0.208725 |
| Default | 0.000024 |
| loan | 0.0240361 |

For this research work, the bank marketing dataset was downloaded from the Data world1. It comprises 41188 rows and 21 columns.



We have done missing value treatment using median imputation,for categorical we have performed mode imputation

**Check for Outliers :**

We have numeric column which are discrete in nature and for rest of numeric we have done outlier treatment so we have done capping of the oultier with boundary limit.

Age, duration, campaign, Prev\_contacted\_duration,Prev\_count and cust\_conf\_index has outliers, Lets treat these outliers

Age : There are only 422 people above the age of 70 which is very less compared to the total population and these can be considered old all together so we can cap them at 70 itself or say 98th percentile

Duration : The attribute duration has high outliers stating that with some people the duration of conversation has crossed beyond 1000 seconds but majority of the duration lies between 0 to 600 capping the high outliers to a more reasonable number with majority so that the the high duration data does not affect while building the model there are still some outlier but these lie in around 80th percentile so we can probably allow that

Prev Contacted : 999 just refers to the people who have not contacted not at all, it does not have a significance so number of days passed doesn't matter, so we can just change the number to 28 .We just removed extreme values, even though there are outliers its not extreme outliers

Campaign : Campaign has ignorable instances once it crosses 10 so we can cap it at around 10 since most of the people have been contacted only once or twice while some have been contacted more but instances where people were contacted more than 10 are very less about 800 out of 40k which is very less.

**[EXPLORATORY DATA ANALYSIS](#page16)**

**Univariate Analysis with Box Plot and Histogram distribution of all the columns were done and as part of Bivariate Analysis, pair plots of all the features were made and the following observations were made:**

**Term Deposit: (**our target variable) This is a categorical column. It is highly imbalanced, so we use Smote Technique to balance the data.

**Age:** age distribution is Right-Skewed. (here we got mean value is 40,median value is 38,mode value is 31).mean is greater than median and mode so it is right-skewed.

**duration**: The attribute duration has high outliers stating that with some people the duration of conversation has crossed beyond 1000 seconds but the majority of the duration lies between 0 to 600 capping the high outliers to a more reasonable number with the majority so that the high duration data does not affect while building the model there is still some outlier but these lie in around 80th percentile so we can probably allow that.

**Campaign:** Campaign has ignorable instances once it crosses 10 so we can cap it at around 10 since most of the people have been contacted only once or twice while some have been contacted more but instance, where people were contacted more than 10,10 is very less about 800 out of 40k which is very less.

**Prev\_Contacted\_Duration:** Here we use boxplot.999 just refers to the people who have not contacted not at all, it does not have significance so the number of days passed doesn't matter, so we can just change the number to 28.

**prev\_count:** This has only 4 significant values remaining and has very less instances to call them significant since this makes the values discrete in nature and hence can be treated as categorical.

**Jobs:** jobs are a categorical column here we use a bar plot in that people with admin jobs have been contacted more by the bank. People with unknown jobs are very few. As we can see people with admin jobs subscribed the most.

**Marital status:** Marital status is a categorical column here we use a bar plot in that we can see that married people subscribe to term deposit more compared to others while divorced people are the least.

**emp.var.rate:** Here we use histplot .No clear relationship with term-deposit.

**Cust\_Conf\_Index:** Here we use countplot. No clear relationship with term-deposit.

**Cust\_Price\_index:** Here we use countplot . It says about account balance of the client.

**Euribor\_3M**: Here we use distplot .loans maturity of some period of time.No clear relationship with term-deposit.

**No\_employed:** Here we use countplot of number of employees.

**Poutcome:** outcome of the previous marketing campaign(failure,non existent,success)non existent gives more outcome for subscription of term deposit.

**Education:** Here we use countplot. (university degree, High school, illiterate, professional course and so on)the university degree have been contacted more and they have subscribed the most.

**Loan:** In loan attribute who don’t have loan they have taken more subscription in term deposit.

**Month**: In month variable most of the subscription happened during month of May but during September,October,march with respect to the highest conversation ratio almost contacted person taken subscription.

**Contact:** The people who are contact by cellular phones they are getting more subscription.

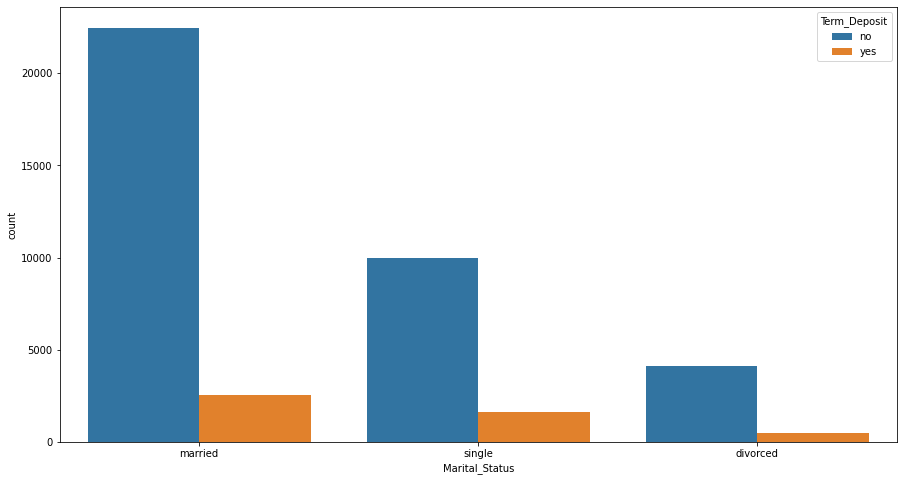
**Default**: Has credit in default(no,yes,unknown) but here most of the members is no.

**Housing:** Has housing loan(no,yes,unknown) but here most of the members has housing loan.

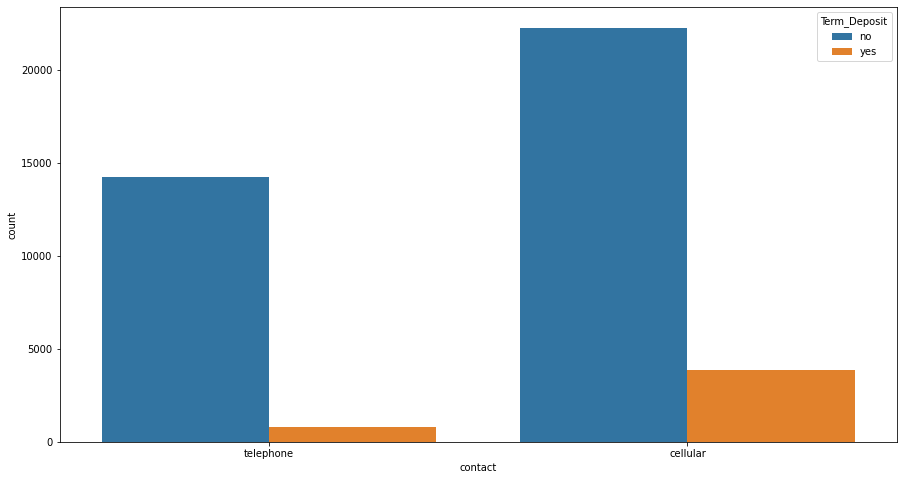
**day\_of\_week:** number of days passed after the last contact.

**Following are few observations which can be seen how the term-deposit are related to various features:**

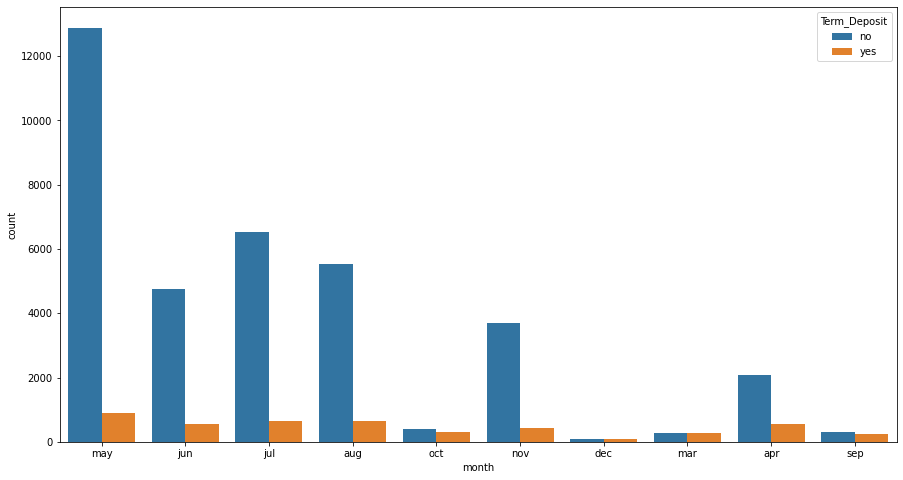
1.In marital status there is a clear that most of the married people are interested to subscribe the term deposit compared to others while divorced people are the least.

****

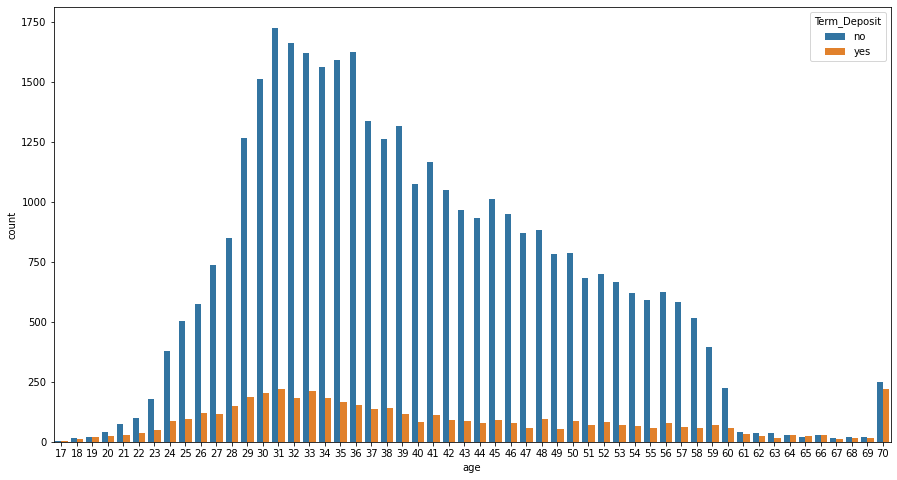
2. In contact there is a clear that people who are contacted by cellular phones are getting more subscriptions.



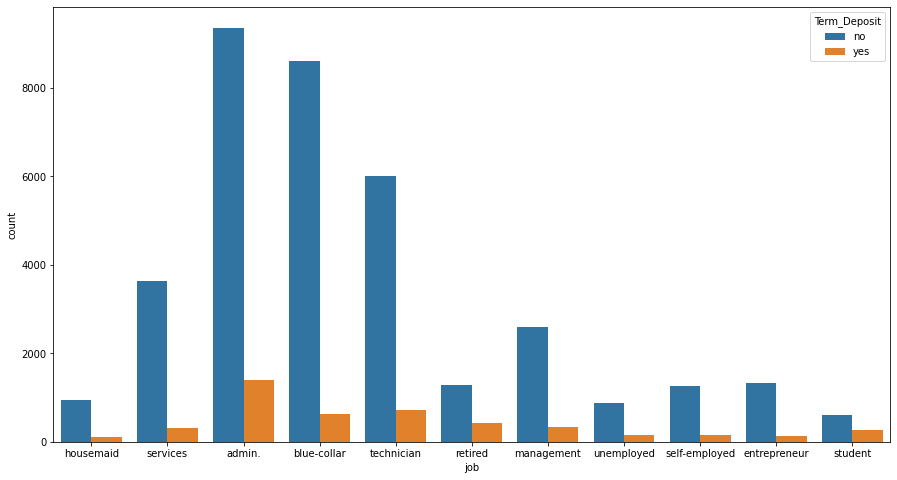
3. In month variable there is a clear that most of the subscriptions happened during month of May but during September, October, March with respect to the highest conversation ratio almost contacted people has taken subscription.



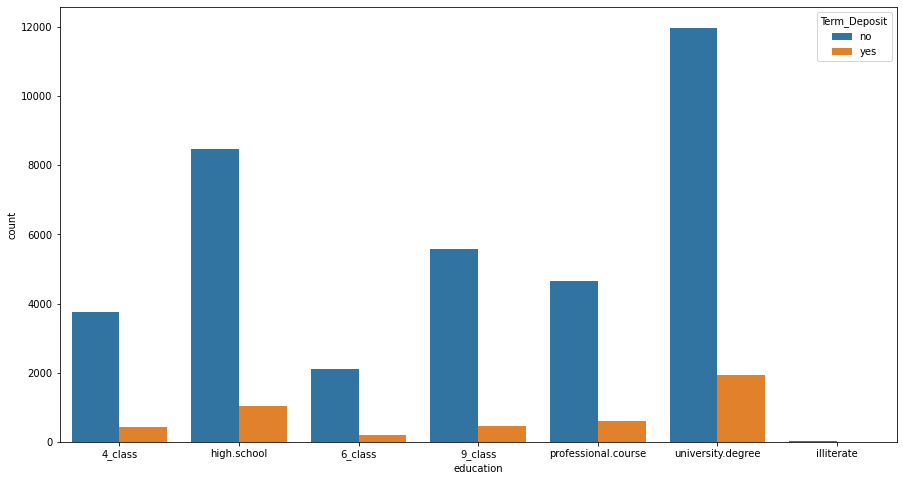
4. In age there is a clear that medium age are contacted more and they were taken maximum term deposit. Though old people are less contacted they were taken subscription with the highest conversation ratio.



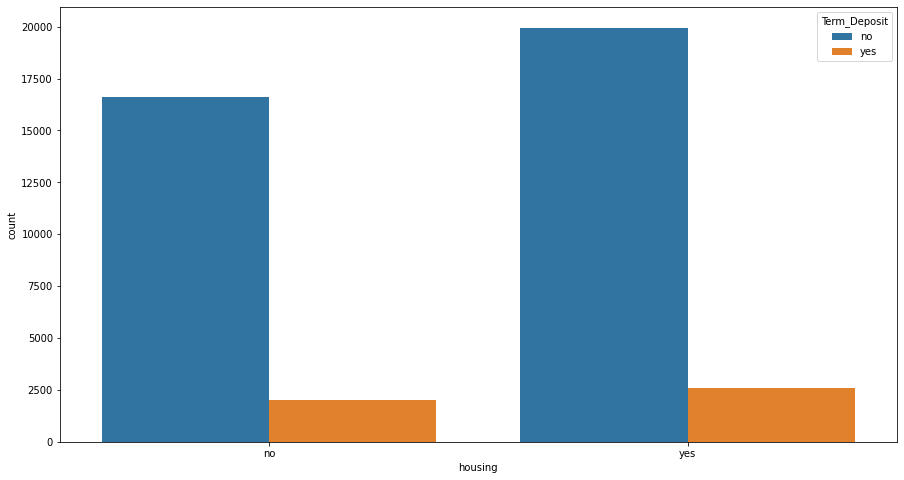
5. In jobs most of the members of admin, technician and blue color were taken subscription. Though students are least contacted they were taken term-deposit with highest conversation ratio.



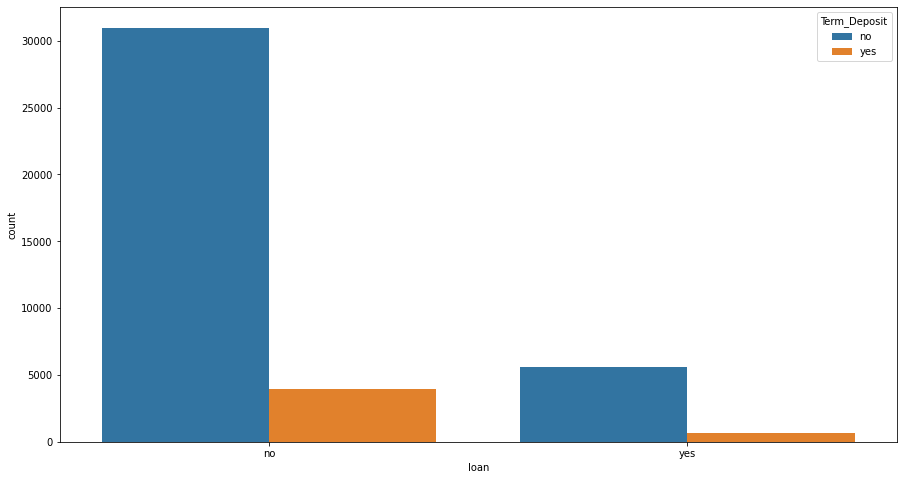
6. In education there is a clear that university degree and high school students were taken subscription.



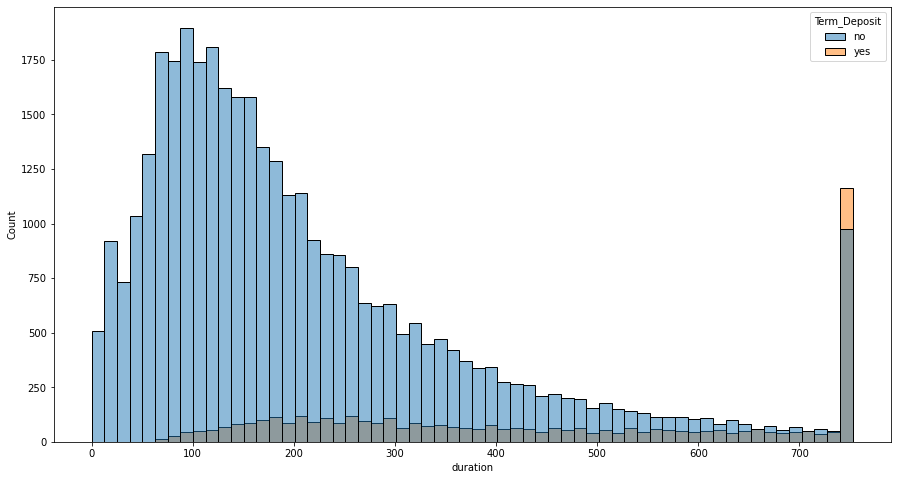
7. In housing most of the members have house loan yet they were taken for subscription.



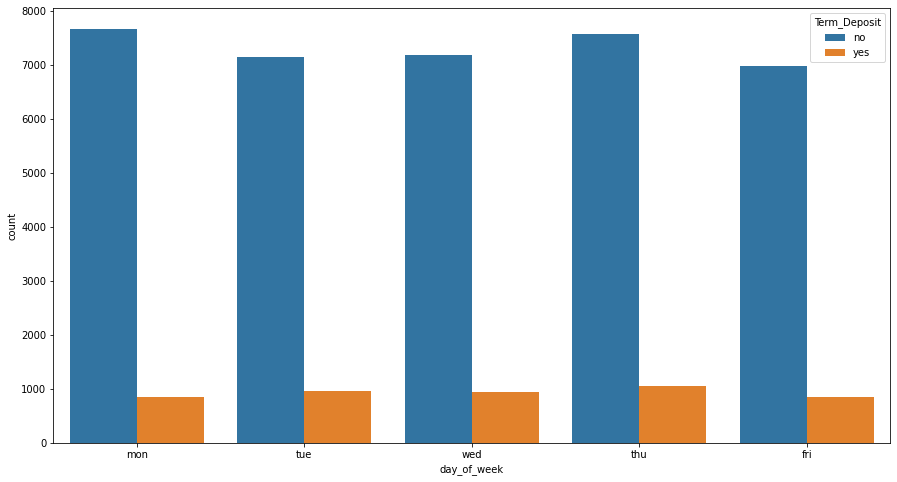
8. In loan variable most of the members doesn’t have loan.



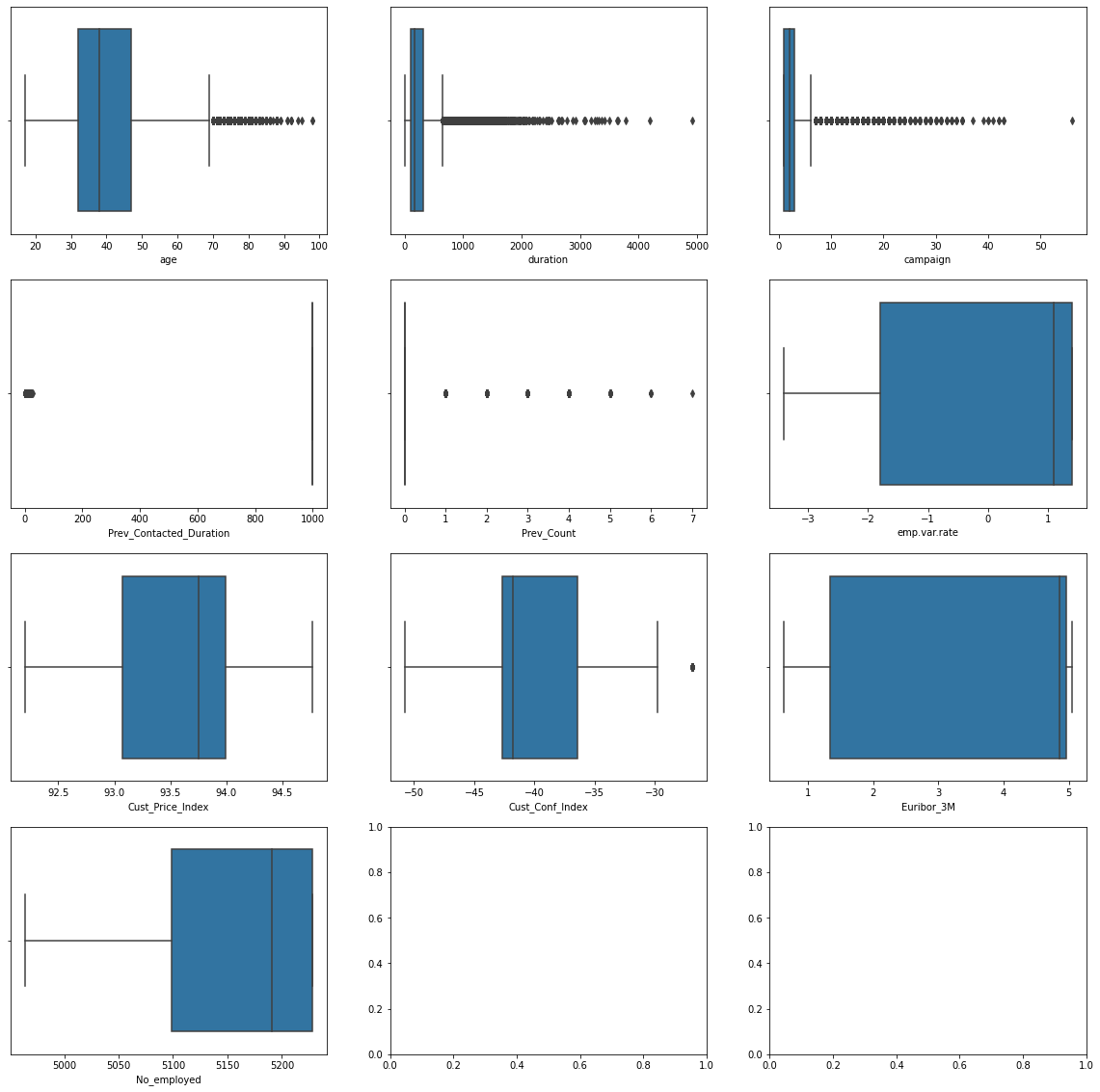
9. In duration variable, last call duration people were taken subscription with a high conversation ratio.



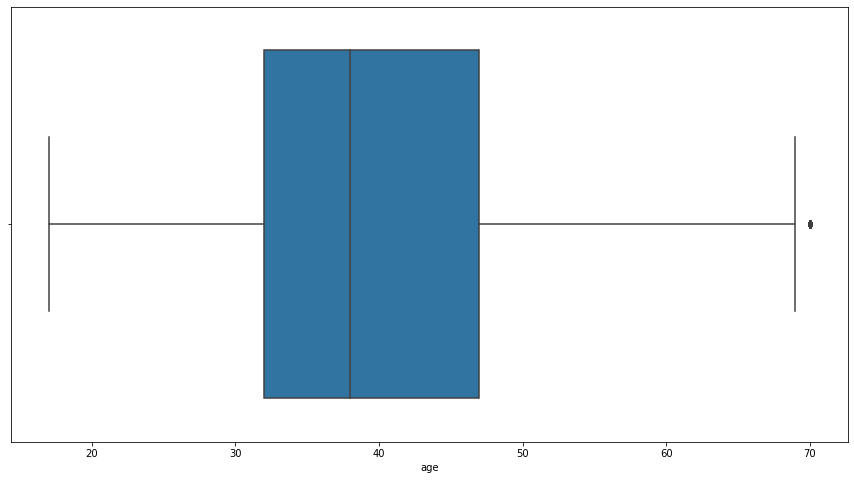
10. In day\_of\_week variable almost every day is equal that people were taken subscription.



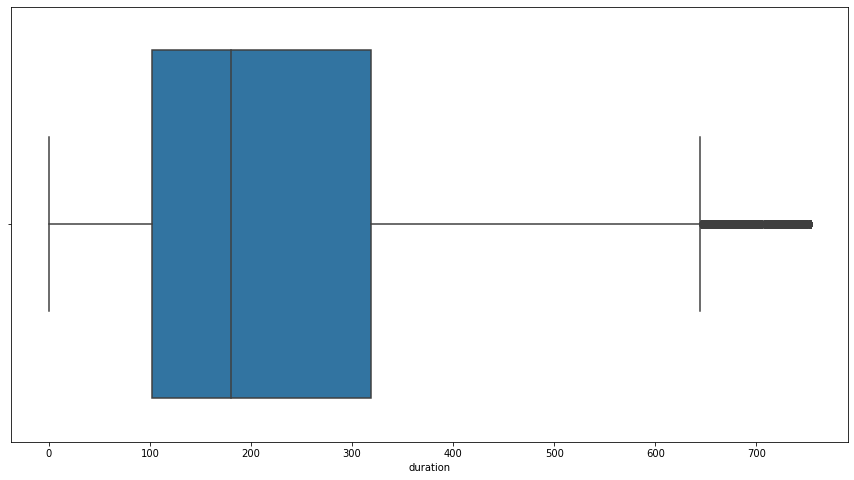
Univariate Analysis with Box Plot. Here we can see that some of the variables has outliers i.e, Age, duration, campaign, Prev\_contacted\_duration, Prev\_count and cust\_conf\_index has outliers, Later we treat these outliers with IQR method.



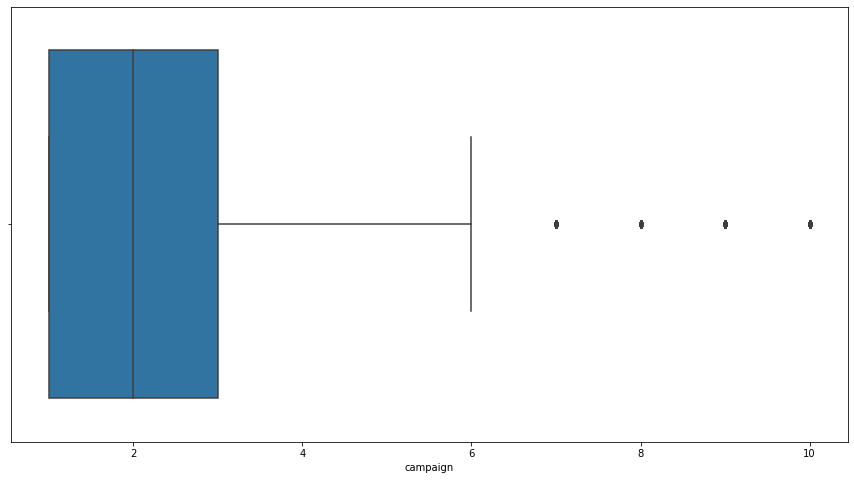
1. Age column after treating outlier:



1. The attribute duration has high outliers stating that with some people the duration of conversation has crossed beyond 1000 seconds but majority of the duration lies between 0 to 600 capping the high outliers to a more reasonable number with majority so that the high duration data does not affect while building the model there are still some outlier but these lie in around 80th percentile so we can probably allow that.



1. Campaign has ignorable instances once it crosses 10 so we can cap it at around 10 since most of the people have been contacted only once or twice while some have been contacted more but instances where people were contacted more than 10 are very less about 800 out of 40k which is very less.



### Some Univariate Analysis with barplot.

### default, housing, loan

### 

### contact, month, days

### 

### Poutcome

### 

### 

### Distribution of Euribor\_3M by using distplot.

### 

### Emp\_var\_rate

### 

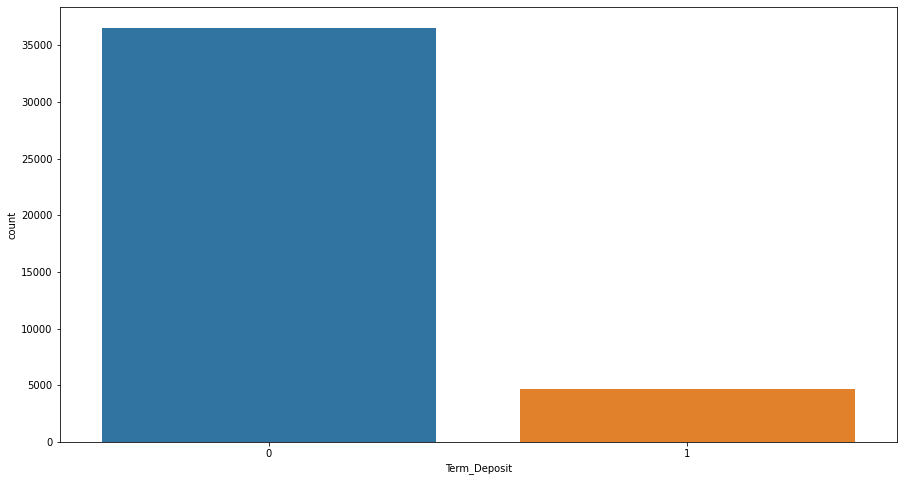
### Cust\_price\_index

### 

### Cust\_conf\_index

### 

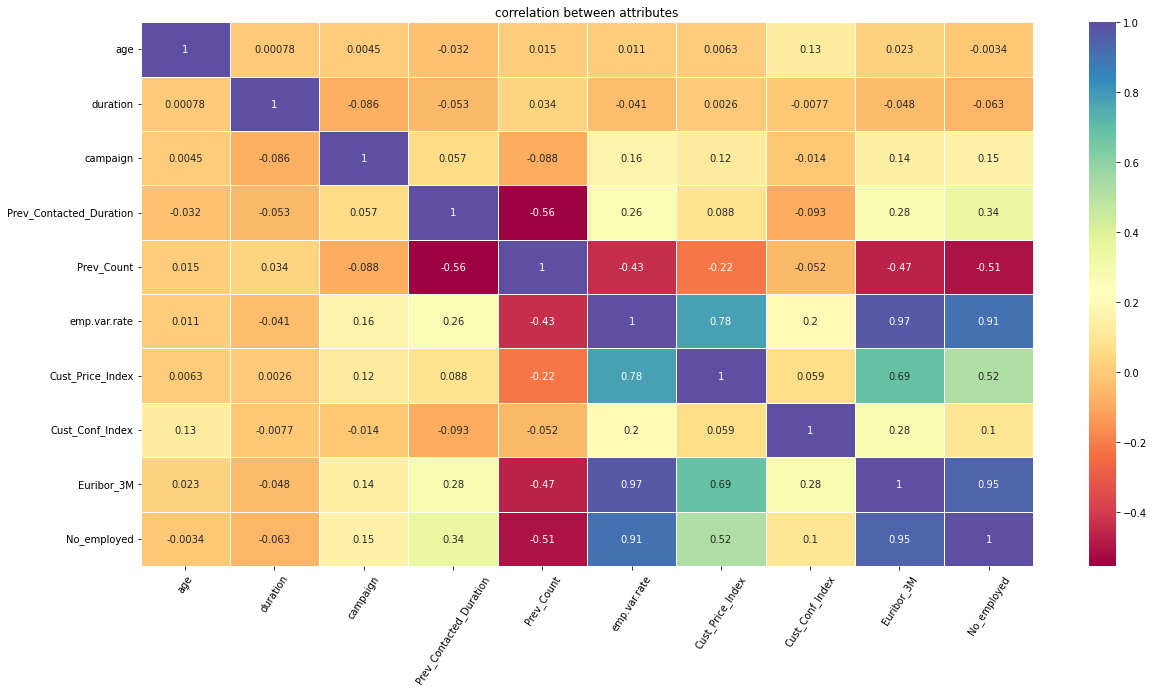
Term-deposit: (Target variable) Here we can see that target variable is highly imbalanced with a ratio of 88:11



So here we use Smote Technique to balance the data:



Checking for correlation between the features, using heat maps for a better understanding.

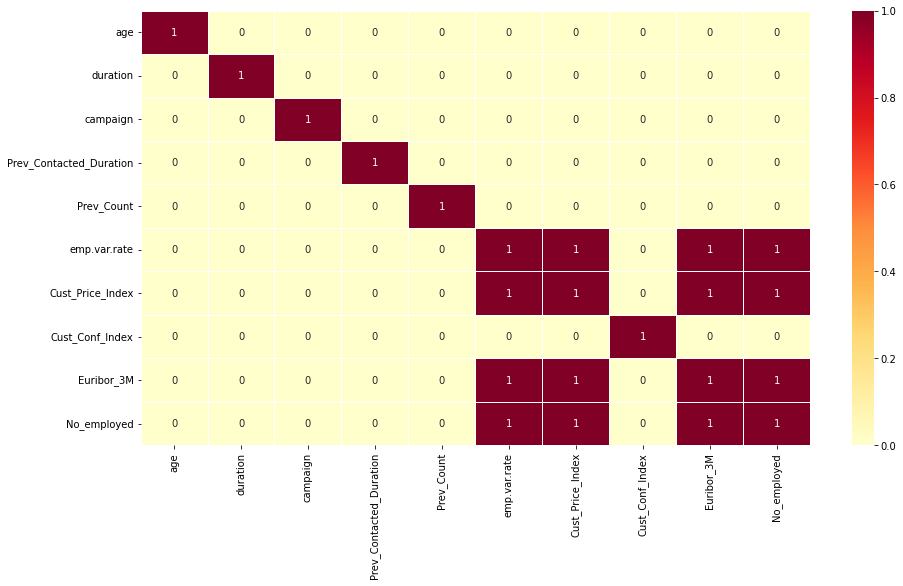


Based on Correlation matrix we can say that some variables are so highly related that only one of them would suffice for model building 'Positive high correlation between:

'emp.var.rate' and 'no\_employed'

'emp.var.rate' and 'Euribor\_3m'

'Euribor\_3m' and 'nr.no\_employed'

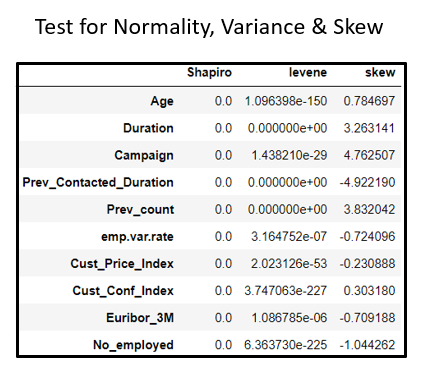


So we drop No\_employed, emp.var.rate but we keep Euribor\_3M because it gives more information.

### **Statistical Significance of Variables**

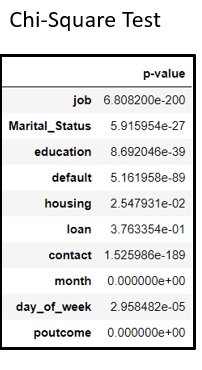
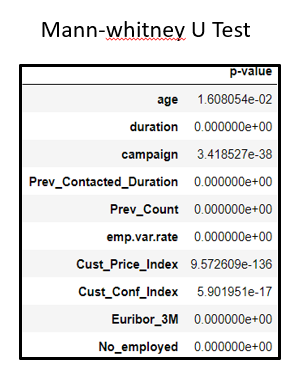
We perform statistical hypothesis testing to understand the significance of the independent variable in predicting the dependent variable(target variable). Before we proceed to the hypothesis testing, we need to check the if the data follows the following assumptions:

* Data has normal distribution
* Data has equal variance

If these two assumptions are satisfied, then we can perform parametric test on the numerical data. For categorical variables, since there is no variance and normality as they are based on proportions of their subclasses, we proceed to perform non-parametric test.

In this dataset, we perform the Chi Square Test to assess the significance of the categorical variable in predicting the Term\_Deposit (dependent variable). Here we are using the p-value method to test the significance of an independent variable. We can notice that the p-value of job, Marital\_Status,education,housing,contact,month,day\_of\_week,poutcome are lesser than the significance level (0.05). Thus we reject the null hypothesis, concluding that these variables are significant in predicting the outcome of Term\_Deposit. Therefore Term\_deposit is dependent over these variable. Whereas loan and default variables are insignificant for the analysis

For the Numerical data we checked if it passes the Shapiro and Levene Test. As the p-value of all variables are less than the significance level we reject the null Hypothesis.Thus we can conclude that numeric variables do not have a normal distribution and equal variance. Since these variables do not satisfy the assumptions of being normal and having equal variance, We proceed with non-parametric test, Mannwhitney-U test. From result obtained we observe that all numerical variables have p-value less than significance level, we reject the null hypothesis and conclude that all the numerical variables significantly impact the prediction of Term\_deposit.



Statistical tests conducted to assess the significance of features

**MODEL BUILDING**

**Classification Predictive Modeling**

In this Bank Marketing Dataset, we need to do a predictive classification modelling, we chose classification modelling since there are 2 class labels, subscribed to a term deposit or not. There are many different types of classification algorithms for modelling classification predictive modelling so in this we try out 3 different types of models.

Classification predictive modelling algorithms are evaluated based on their results. Classification accuracy is a popular metric used to evaluate the performance of a model based on the predicted class labels. Classification accuracy can’t be trusted alone, so other metrices are used to ensure the accuracy of model.

The various models built, must be evaluated based on certain model performance measures to identify the most robust models. The choice of the right model performance measures is highly critical since the dataset is a highly imbalanced dataset and the conversion rate. Model accuracy alone may not be enough to evaluate a model. Hence the following model performance measures have been used to evaluate the models, based on the confusion matrix built for the predictions on the training and test dataset:

Accuracy: Accuracy is the number of correct predictions made by the model by the total number of records. The best accuracy is 100% indicating that all the predictions are correct. Sensitivity or recall: Sensitivity (Recall or True positive rate) is calculated as the number of correct positive predictions divided by the total number of positives. It is also called recall or true positive rate (TPR).

Specificity: Specificity (true negative rate) is calculated as the number of correct negative predictions divided by the total number of negatives.

Precision: Precision (Positive predictive value) is calculated as the number of correct positive predictions divided by the total number of positive predictions.

F1-Score: F1 score is an overall measure of a model's accuracy that combines precision and recall A good F1 score means that you have low false positives and low false negatives, so you're correctly identifying real threats and you are not disturbed by false alarms. An F1 score is considered perfect when it's 1, while the model is a total failure when it's 0.

The target variable that the client subscribed to the term deposit is a binary classification task which has one class i.e. –Subscribed and another class is not Subscribed to the term deposit. The class subscribed is assigned the class label 1 and the class not subscribed is assigned with the class label 0.

The algorithms that are used for classification are:

• Decision Tree

• Random Forest Classifier

• XG Boosts

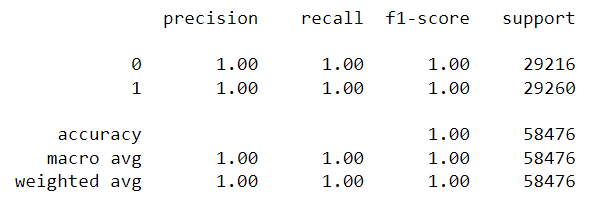
**Decision Tree**

One of the popular machine learning models is tree-based methods. The simplest tree-based method is known as a decision tree. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules gotten from training data. In Decision Trees, for predicting a class label for a record we start from the root of the tree. One advantage of tree-based methods is that they have no assumptions about the structure of the data and are able to pick up non-linear effects if given sufficient tree depth.

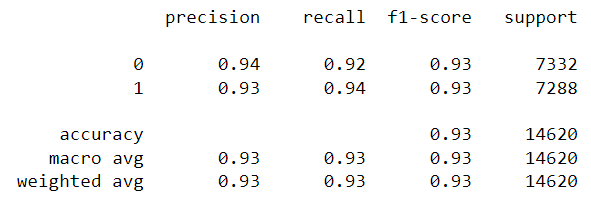
## **Base Model without Hyperparameters**

Creating base model with the help of Decision tree to check the accuracy, precision, recall values to evaluate the model of the present data.

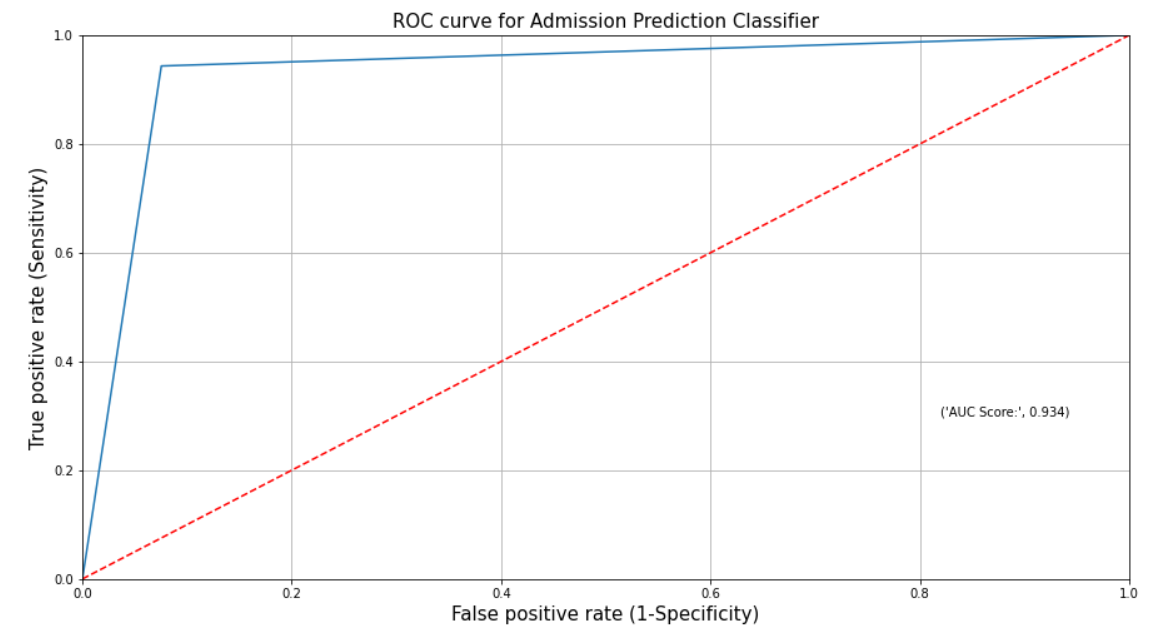
Classification report on train set



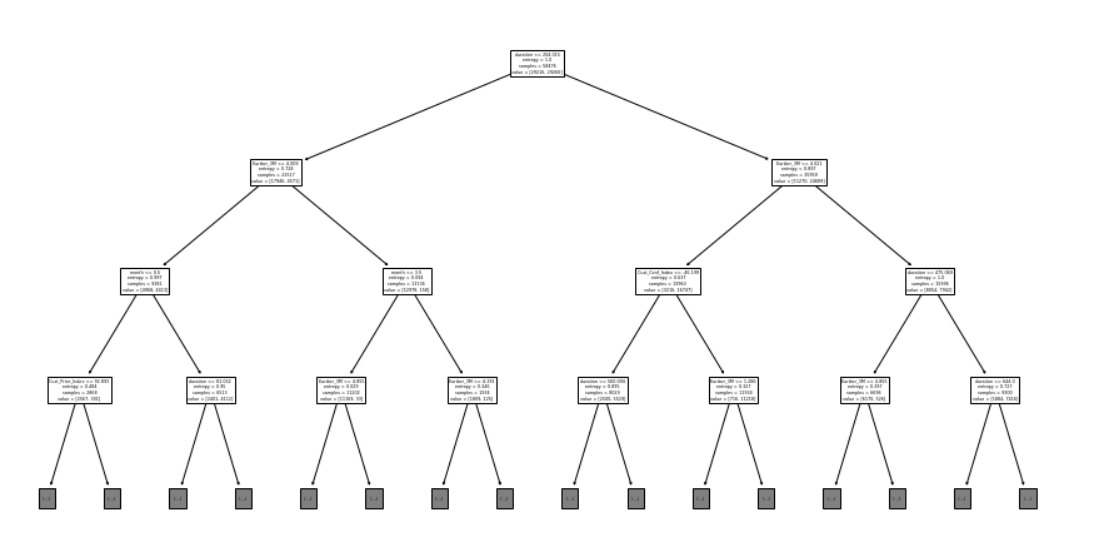
Classification report on test set



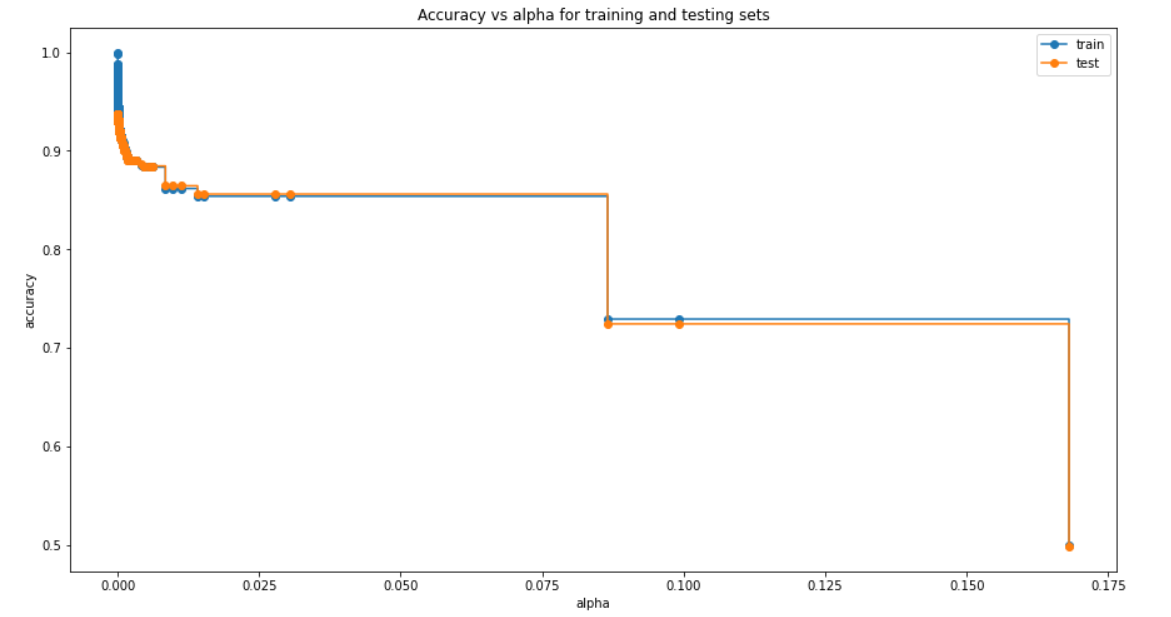
ROC Curve



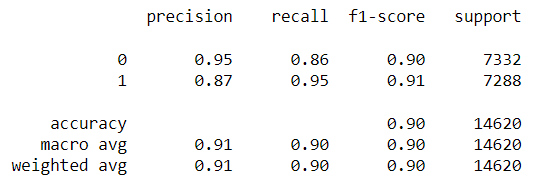
Decision Tree plot



## **Decision tree after post pruning**

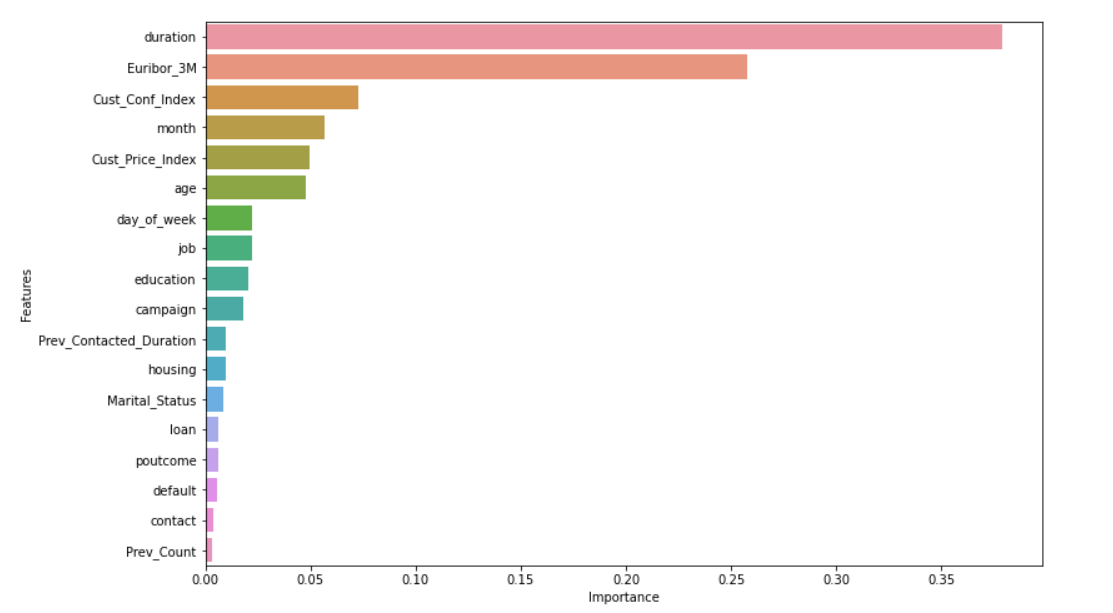


Classification report

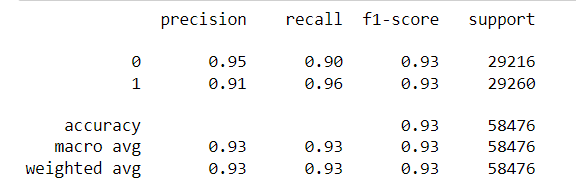


## **Model after selecting important features and hyperparameter tuning**

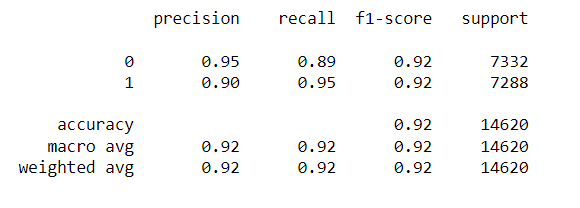
## **Important Features**



Classification report on train set



Classification report on test set



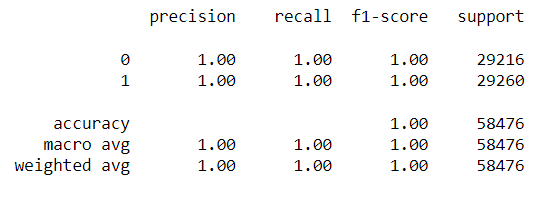
Decision Tree Accuracy



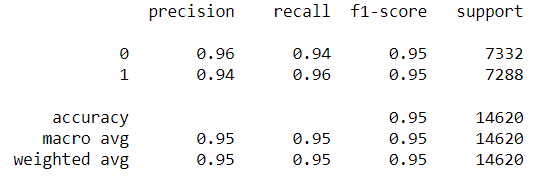
**RANDOM FOREST CLASSIFIER**

Random forest belongs to supervised learning method algorithm used for classification that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or means prediction of the individual trees. The decision tree is a tree structure (which can be a binary tree or a non-binary tree). Each of its non-leaf nodes corresponds to a test of a feature, each branch representing the output of the feature attribute over a range of values, and each leaf node storing a category. The decision tree starts with a root node, tests the corresponding feature attributes in the category to be classified, and output branches are selected according to their values until the leaf node is reached, finally the category stored by the leaf node is regards as the decision result. A random forest is a collection of decision trees in which each decision tree is unrelated.

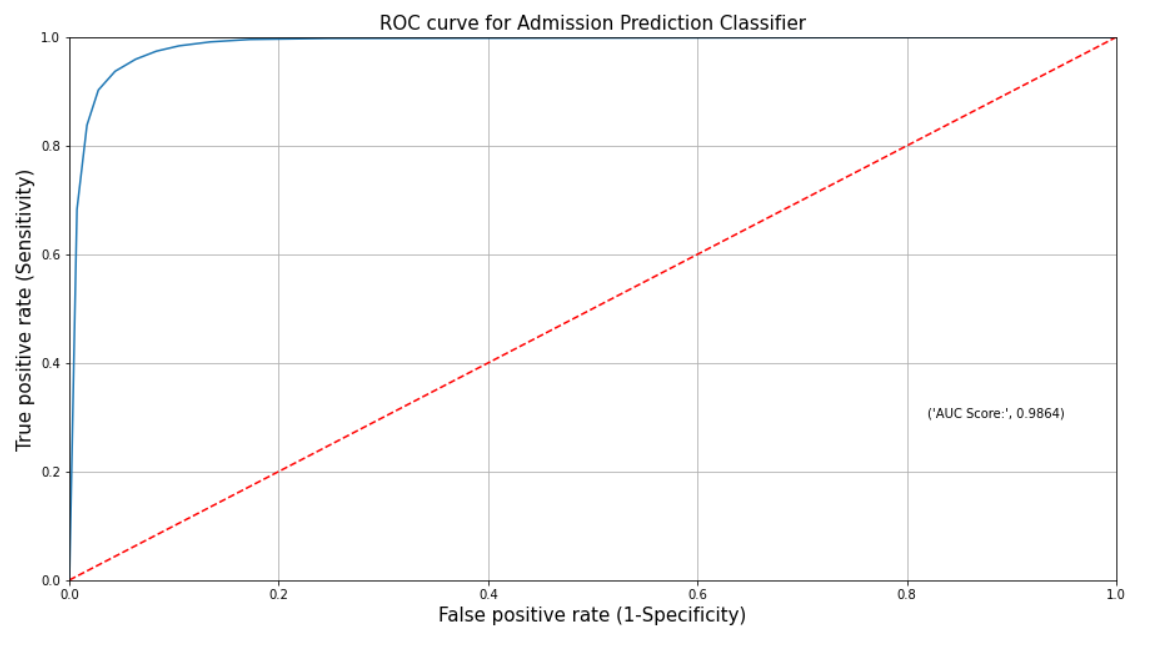
Classification report on train set



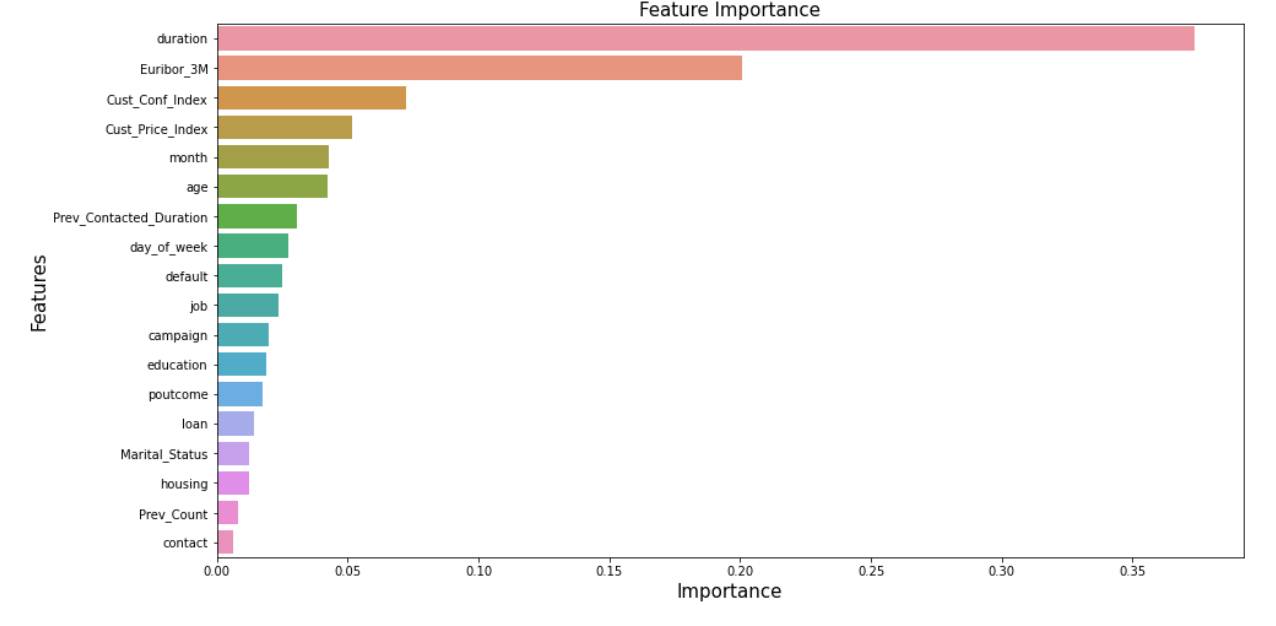
Classification report on test set



ROC Curve



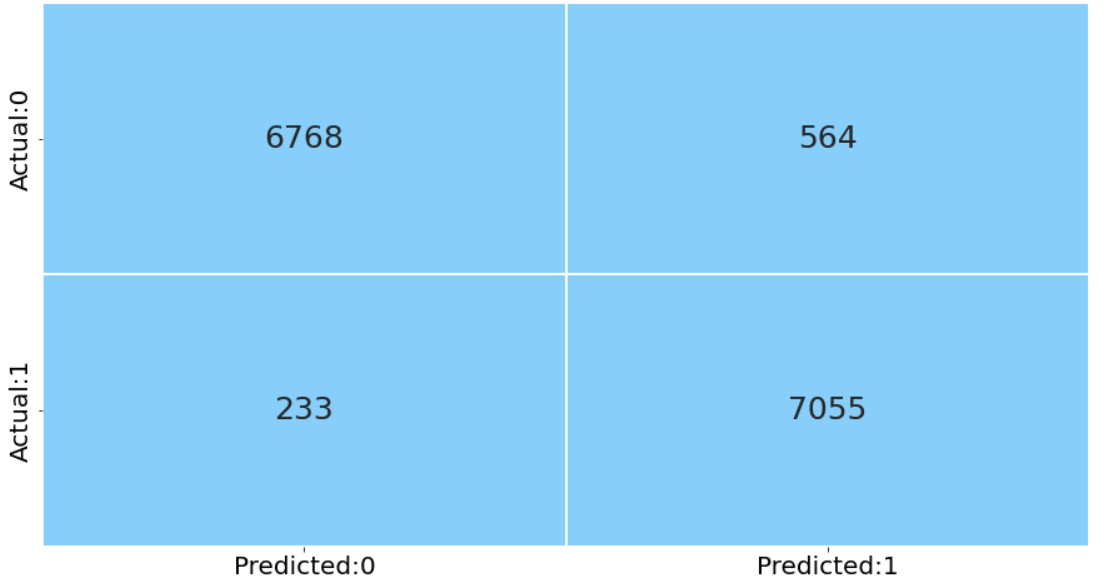
## **Important Features**



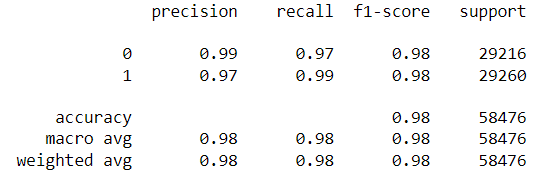
**XG Boost**

XGBoost is a scalable and highly accurate implementation of gradient boosting that pushes the limits of computing power for boosted tree algorithms, being built largely for energizing machine learning model performance and computational speed. With XGBoost, trees are built in parallel, instead of sequentially like GBDT. It follows a level-wise strategy, scanning across gradient values and using these partial sums to evaluate the quality of splits at every possible split in the training set

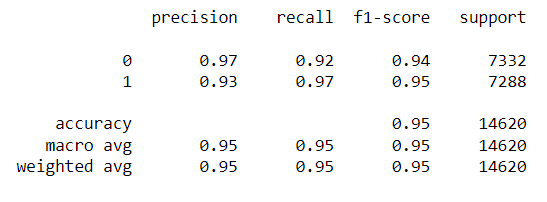
Confusion Matrix



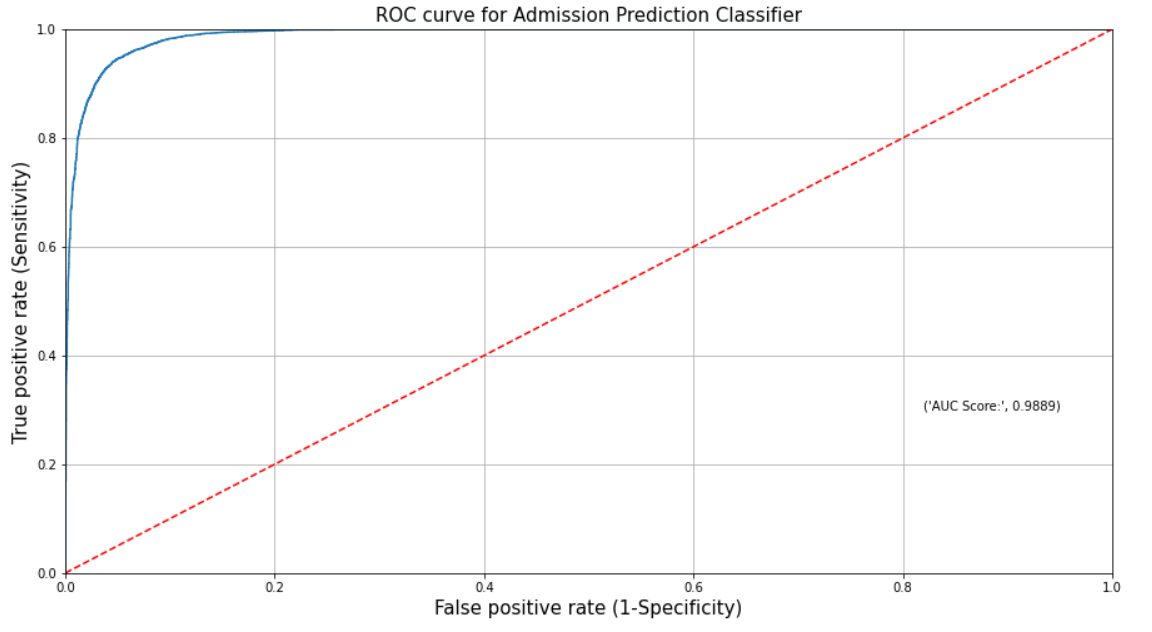
Classification report on train set



Classification report on test set



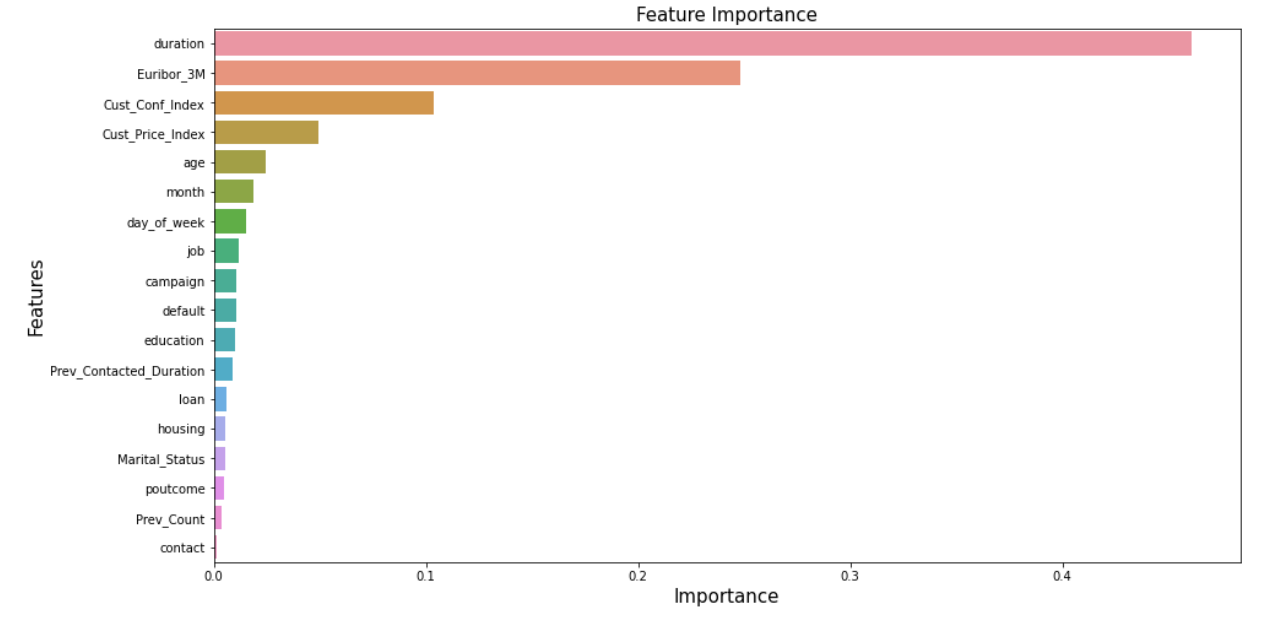
ROC Curve



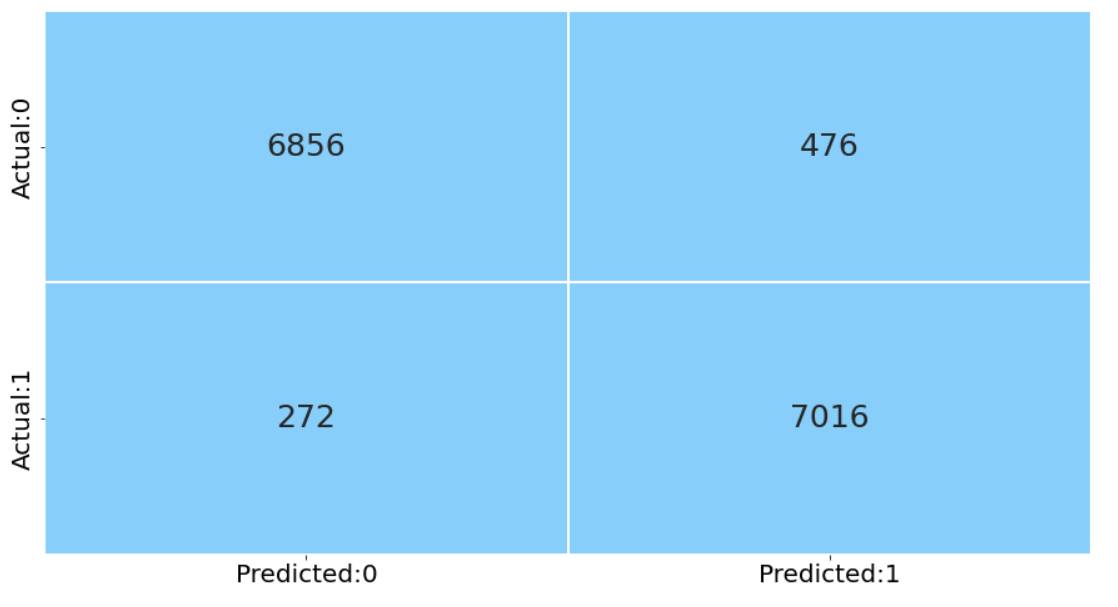
Accuracy Score :



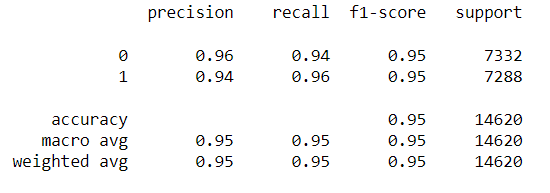
## **Model after selecting important features and hyperparameter tuning.**



Confusion matrix



Classification Report



Accuracy Score:



Here the accuracy is found to be 94% and this would be a good fit for the model.

**Overall Model Results**

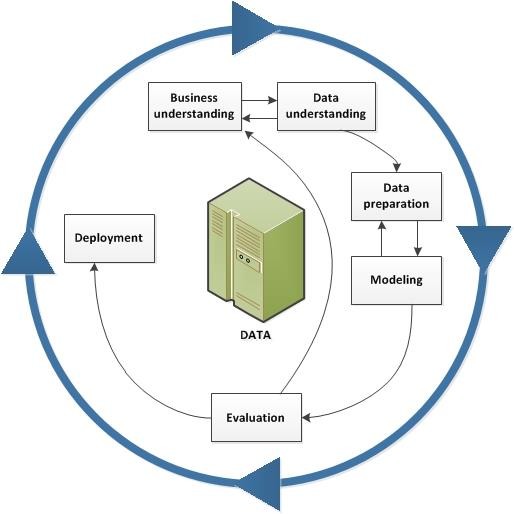
|  |  |
| --- | --- |
| **Model** | **Accuracy Score** |
| Decision Tree | 0.92 |
| Random Forest | 0.95 |
| XG Boost | 0.95 |

When compared with different types of models and its corresponding accuracy scores it has been found that the XG Boost seems to have the highest accuracy score.

We are considering XG boost to be our final model ,this model is better fit compared with other models and also has better performance

**Research Methodology**

The methodology to be used for this research follows the cross-industry standard process for data mining (CRISP-DM).it provides a systematic approach to planning a data mining project. This approach is reliable and well-proven due to its step by step process and its general applicability (Gregory, 2018). This CRISP-DM includes five phases which are hierarchical and will be implemented during a data mining project (Rudiger and Jochen; 2000). These are shown in the diagram below;



**Figure CRISP-DM Approach**

**Business Understanding**

The first phase of the approach chosen for this data mining project is market segment awareness. For the main objective of this project to be achieved, an appropriate dataset needs to be put into consideration. The historical dataset used for this project was derived from the data.world. Containing 41188 rows with 21 columns.

Data Preparation

The data for this analysis will be designed finally after having understood the data. The data is expected to be used for their potential modeling for this analysis. this analysis will start by gathering data from the data.world. The data will be obtained by downloading the bank marketing dataset which will be followed by the selected data exploration. The exploration will help to provide a clearer understanding of the data characteristics, size, and structure. Exploration will also assist in choosing the variables to be used while keeping in mind the study's question and purpose, Better analysis of the data for this research will expose data quality problems and observations. It will eventually help in defining the type of data mining technique. It will ultimately help to decide what type of data mining technique to use for the research to better achieve a reasonable outcome

**Data Modelling**

The data will Present its independent variables and target variable after the data preparation has been completed to a certain level which will be further divided by a certain percentage for validation into test and training data. A certain percentage will be allocated to the training, while its percentage will be allocated to the test though smaller. The training is assigned higher so that the classifier can learn from the training larger part of the Dataset. This breaking will make it possible to apply the chosen modeling technique. This research will be limited to three chosen modeling techniques from other modeling approaches used for a data mining problem. Decision Tree, Random Forest, XGBoost are the models to be used for this research work.

Model Evaluation

These model(s) shall be evaluated for every analytical function. The models to be implemented for this analysis will be tested, as this will shape as necessary an important part of the research process. The models that will be used will be reviewed to determine their accuracy, and based on their accuracy, this will be done using the confusion matrix.

**Deployment**

This is the last stage of the study. The information gained from the results obtained will be reported and provided for use after thorough modeling with its assessment. This can also imply that the models used will be contrasted and therefore the one considered to have performed better will be suggested to the finance sector, especially the financial service providers to serve as means of improving their services. This will allow them to use this research to identify the main problems and focus on ways to develop innovative approaches to fix the existing problems, thereby providing customers with a better financial service that could reduce grievances and disputes.

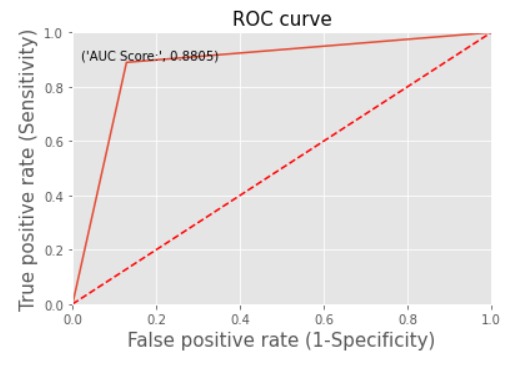
**Discussion**

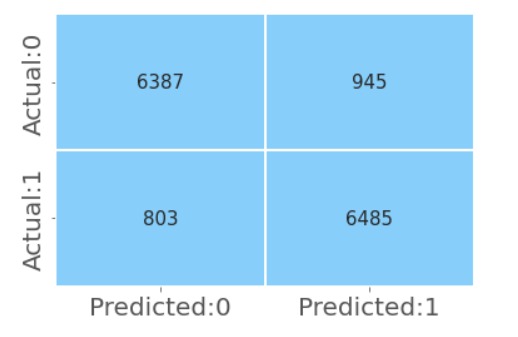
This thesis started as a guideline for researching by setting out a research question and its specified objectives. On this basis, a methodology of data mining was chosen and implemented with the analysis of three different algorithms to facilitate the specified objectives, address the question, and accomplish a meaningful analysis. This research adopted three classifiers approach to accommodate its specified goals based on performance conducted. on this basis, different outcomes or findings were generated using the three chosen algorithms adopted for this research: Decision Tree, Random Forest and XGBoost. the result generated from the analysis is to help the banks in identifying the main factor that can increase customer subscriptions to a term deposit. based on the result from the models carried out, XGBoost has the highest accuracy rate of 94% in the test data. This indicates that XGBoost will be a good model for predicting if customers will subscribe to a term deposit or not. however, the other two models also perform well but XGBoost performs exceptionally as the best model. Besides, two experiments were carried out in this study, the first experiment was done using all the variables present in the dataset after pre- processing, while the second experiment was performed using 10 important variables to improve the output of the model using lesser variables.

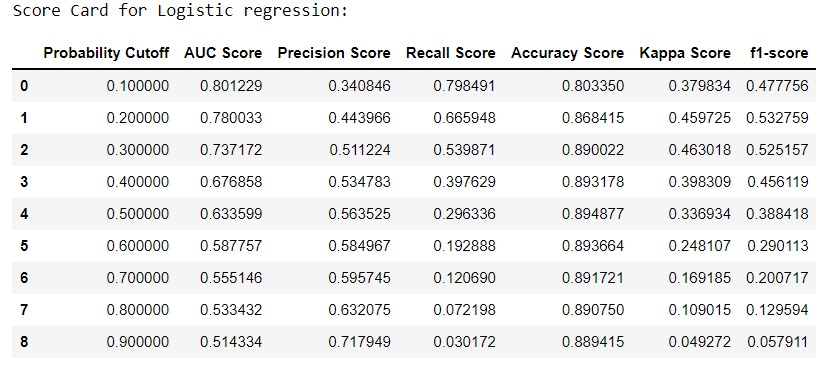
In experiment one, XGBoost had 93% for both Accuracy and AUC. Looking at the model that correctly classifies the number of subscribers i.e. Recall, KNN had 92%, Decision Tree and Random forest had recall of 92% and 90% respectively.

The second experiment was done to achieve the research objectives and research question as highlighted in sections by identifying the main factors that can influence the customer's decision in signing up for a term deposit in the bank using Feature Importance. The results have gotten shows that XGBoost has the highest Accuracy with 94% which is the same as that of test data and AUC score of (94) which has the same values as that of experiment one. Looking at the model that correctly classifies the number of subscribers i.e. Recall, XGBoost had 97%, Decision Tree and Random Forest had reduced scores of 89% and 91% respectively. for Precision, XGBoost had a score of 96%, Decision Tree had 95%.

**Comparison to benchmark:**







As we have taken logistic regression as our bench mark model with different Probabillity cutoff

**Implications**

From the processing, EDA and model building we have found out the following attributes and what are the changes to be done on how to approach the attributes differently with respect to target variable

1. Age: Most of the data we got was from the ages between 25-50 but the people who has subscribed to term deposit are the older people who are above 55 so the conclusion for this attribute is to target older generation people more
2. Jobs: The people who have been target the most are admin, blue-collar and technicians and these people have subscribed more as well but other occupations such as who are students eventhough they have been targeted less but the conversion rate for students is better than other attributes so need to target more people who are students
3. Marital Status: Marital status had less impact on target variable which we have noticed in the feature importance, so this attribute need not be considered in making important decisions on whom to target.
4. Education: People with university degree and highschool have been targeted the most but we can see that the people who subscribe to term deposit are mostly from university degree and people who are pursuing a professional course since they are more mature to understand the importance of term deposit.
5. Housing , Loan and Default: From these 3 attributes we can notice that people who are financially independent and less burdened by loan, or housings tend to subscribe to more term deposit, so we can target accordingly to more financially independent people
6. Month and Day of week: May is the month where the people have been targeted along with some other months June July, this indicates that people have been targeted more after the beginning of new fiscal year but results show that the conversion rate has been the most during the end of the year starting from September to December there is a high chance of subscribing to term deposit, The day of the week doesn’t matter much interms of subscriptions and they have been targeted equally.
7. Social and Economic Attributes: These attributes such as Customer Confidence index, Customer price index, Euribor\_3M are some of the most important attributes which are significant in deciding whether the person will subscribe to term deposit or not, so the people with higher numbers in these attributes should be targeted more and these attributes are significant in model building as well.
8. Duration: duration is the most important attribute of all the attributes, this tells us that people who have been engaged more with the customer executives are tend to subscribe more to the term deposit, so we can conclude that people should be more engaged one on one to influence them into subscribing to term deposit more.
9. Camoaign: The campaign attributes which are related to previous campaign success rate are not that important from the exisiting data plus these lack data points to inference more from them, so we can conclude that the success of previous outcome doesn’t decide the success of current campaign.

**Limitations**

* Lack of sample size for statistical measurement is one of the limitations of the study. Studies reveal that statistical measurement requires a large sample size. In this case, the researcher would need to analyze many banks, which would be challenging. Following the restrictions on the banking industry, the researcher may not identify the most appropriate banks to study and the number of institutions to study. Lack of large sample size is also problematic in terms of concluding. Research that focuses on a small sample size will not determine the actual phenomenon of the market. Therefore, it is important to focus on a large sample size.
* Also, this research is subject to limited data. Concerning the restrictions in the banking industry, the research will have limited access to the data. Many banks would not allow researchers to get information on some essential aspects of the business. The nature of the banking business does not allow them to offer access to various information. Also, some banks are not willing to share information about their marketing strategies due to fear of competition. As a result, this research may not identify the actual campaign strategies used by the selected banks.
* Time constraints are another limitation of the study. Like many organizations, banks work on a fixed schedule. The management of any bank ensures that time is utilized effectively. In this case, the researcher will write a letter to request for an appointment with various stakeholders. Due to the varying schedules of each bank, the research may not meet the guidelines. The limitation of time is also attributable to data collection and analysis.
* The customers social and economic attributes and the contact related information are influencing target more than personal information about the client, Marital Status, has he taken a loan or housing or other attributes except age are not that relevant in predicting the outcome of Term Deposit, In real world personal attributes are important so our data and model are lacking in accessing these attributes thoroughly.
* Some attributes such as poutcomes which is basically outcome of previous campaign, or default which has a lot of unknown values, these data lack outputs in majority, the data is either non-existent or unknown, these attributes lack too many data points to use them as prediction for our target variable, so more data should be added on such attributes.
* Previous Contacted Duration is number of days that passed by after the client was last contacted from a previous campaign but most of them haven’t been contacted at all, almost 37000 of our 44000 haven’t been contacted this tells us that the data we have is on new customers who haven’t been contacted previously in any campaign.

### **Deployment**

### **Folder Structure**

This is the folder structure we are following

* Deploy
  + app.py
  + new\_model.pkl
  + requirements.txt
  + Static
    - CSS
      * Style.css
    - Img
      * Image.jpg
  + Template
    - Index.html

## Requirements.txt

Contains all the package versions used. Has to be installed before trying to deploy it.

## App.py

We have used the pickle file of XGB. This is where we write the code for creating an API using Flask

## new\_model.pkl

## the final model on our local machine is stored along with other key model related variables in pickle file (.pkl)

## Style.css

This file contains the styling for all the elements we have used in the html form. And we have linked this to the HTML file.

## Image.jpg

This is the image file which we have used for the background of the webpage.

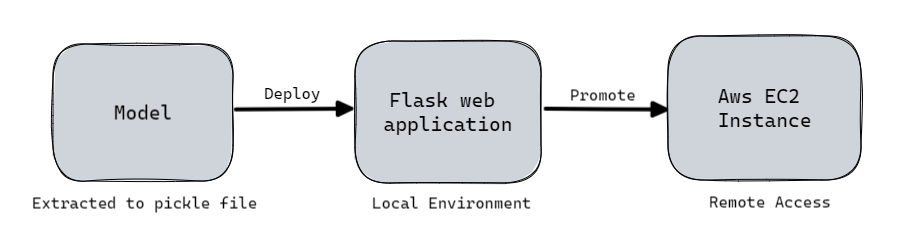
## Index.html

## This is used to build a simple web page, where we are providing the values of duration,Euribor\_3M,Cust\_Conf\_Index,Cust\_Price\_Index,age,month,day\_of\_week,campaign,job, and default for the customer being assessed. On clicking the predict button we are getting the prediction if the Customer subscribes or does not subscribe for term deposit

### **Introduction**

We have used Flask. Flask is a lightweight web framework and is used to manage HTTP requests/responses. It runs on WSGI server toolkit and Jinja2 template engine.

* Request sent via REST API
* Cleaned & preprocessed by Feature Extractor
* Trained model is used in giving the prices.
* Model can be deployed as a pickle file.
* Flask is used to manage HTTP requests/responses. It runs on WSGI server toolkit and Jinja2 template engine.



### **Process**

We have used the HTTP form to collect the user input. Information on Features such as duration,Euribor\_3M,Cust\_Conf\_Index,Cust\_Price\_Index,age,month,day\_of\_week,campaign,job, and default are crucial in predicting the subscription of term deposit.

On clicking the predict button we are getting the prediction if the Customer subscribes or does not subscribe for term deposit. The model used is XGBoost Classifier. The model has been saved into a pickle file and have been used in the ***app.py*** file. When the url for the deployed model is loaded flask receives a **GET** request and it renders the HTML file which has the form created in it.

Once the user fills out the HTML form, selects the model and clicks the Predict button, the flask will receive a **POST** request. It will help us extract the data from the form and we will preprocess it and run it through the model selected and shows the prediction.

### **Screenshot**

Link : [Term Deposit Subscription Predictor](http://ec2-13-232-112-243.ap-south-1.compute.amazonaws.com:8080/)



### **Conclusion**

We have used the model already built along with the flask and AWS to deploy the model. Users can remotely access the predictor on their local browser.

**Conclusion and Future work.**

Many banks use direct marketing strategies to enable customers to access adequate information about the products. Researchers suggest that the applicability of data mining techniques depends on the availability of customers’ information. Also, studies reveal that machine learning techniques determine customer response to bank products. The ability of the customers to subscribe to term deposits depends on the marketing campaign by the bank. In this research work, the resampling method was used in dealing with the problem of imbalanced data, and three machine learning algorithms (Decision Tree, Random Forest and XGBoost) were deployed to find out the main factor that influences customers decision to subscribe to a term deposit in the bank. Feature Importance and Hyperparameter Tuning was then used to get the most important factors that influence customer decisions and was then retrained to perform the second experiment. Both experiments had good accuracies ranging from 89% to 94% with XGBoost performing better with an improved recall score of 97% and AUC score of 94% compared to other algorithms used for this research. The correlation heatmap highlighted five factors that can influencing the customer's decision and ‘duration’ has the highest correlation coefficient with dependent variable (positive correlation) of 0.41. Which means that the longer the bank continue to advertise their product and service, the more customers can subscribe to a term deposit. Banks should focus on direct marketing techniques when applying statistical and mathematical approaches to determine customer response. This project can further be enhanced by using other techniques like univariate selection and logical imputation of missing values on the dataset to identify more factors that can influence customer's decision to subscribe to a term deposit in the bank.