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

Table of Contents

INTRODUCTION

Industry Review

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 [1]. There are many types of deposit accounts and some major types, including checking accounts, savings accounts, term deposit accounts, and money market deposit accounts [2]. 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 suspiciousness resulted in deposits shrank [3]. In addition, due to the rapid development of the 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.

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.

Literature Survey - Publications, Application, past and Undergoing research:

Research Question

How can Predictive Analysis Using Multiple Machine Learning Techniques support the decision-making process in the banking sector to improvise the model to predict whether a customer will apply for a long-term deposit in a bank and in improving the business of the bank?

Literature Review

According to (Suebsing and Vajiramedhin; 2013), Many organizations before offering the services to their customers, analyze the data from the previous customers and takes decisions to avoid any failuoffor the campaigns. Predicting such bank data of the customers can help in finding the hidden patterns help in the success of such marketing campaigns. According to (Moro et al.; 2013), due to the worldwide budgetary crisis, the credits for banks are limited, and the focus for banks is to accumulate funds from their customers. So, accumulating such data and providing services according to that can be of immense help to gaining a successful marketing campaign. Issues such as Behavior, Psychology, Mindset, and Motivation need to be considered to analyze and improve the marketing ability of the organizations (Raorane and Kulkarni; 2011). Managing and Maintaining the data of the customers can help in getting patterns and trends to analyze to generate new strategies to attract new customers (Fayyad et al.; 1996). Machine Learning has a better ability to capture meaningful patterns from the data, also applications of the data mining methods in the sector of banks are increasing enormously (Venkatesh and Jacob; 2016). Classifications can be performed using machine learning algorithms, which can be used to segregate the data into distinct categories (Radhakrishnan et al.; 2013).

There are several valuable studies concerning bank and deposit marketing. Different recommendations are put forward from different marketing aspects based on qualitative methods or quantitative analysis. Data mining techniques have been widely applied in bank marketing as well. We came up with the idea that the association rules can be applied to cross-selling of bank products and customer risk control [6]. However, many studies just compare the performance of different classification algorithms in predicting the success rate of bank marketing campaigns. For example, Moro, Cortez, and Laureano used the miner Package and R Tool to test three classification models (Decision Trees, Naïve Bayes, and Support Vector Machines) and compare their performance through Receiver Operating Characteristic curve (ROC) and Lift curve analysis. Similarly, Moro, Cortez, and Rita tested four data mining models, including logistic regression, decision trees (DT), neural network (NN), and support vector machine. After evaluating the area of the receiver operating characteristic curve (AUC) and the area of the LIFT cumulative curve (ALIFT), the neural network presented the best performance. Nachev combined cross-validation and multiple runs to partition the data set into train and test sets. He also explored the impact of performance caused by different neural network designs.

All research above focuses on predicting customers' behaviors resulting from bank marketing. To avoid marketing campaigns being annoying rather than attractive, the right promotional messages should be delivered to the right customer groups. As early as 1974, Robert put forward the idea of the use of census data in bank marketing [9]. He mentioned that census data can be applied to location analysis and marketing segmentation. Wang, Song, and Fang mentioned that the banking industry lacks scientific marketing management and banks generally adopt some traditional marketing methods, including relationship marketing (use employees' relationships to find deposit clients), self-interest marketing (obtain deposits by satisfying clients' interests, such as gifts), passive marketing (attract customers to increase the deposit by offering warm and thoughtful counter service) and simple service marketing (attract deposits by meeting the low-level requirements of customers, such as providing door-to-door services) [10]. They came up with the idea that carrying out market segmentation of deposit marketing and selecting the marketing target is the scientific way of marketing management. However, problems like obsolescence of data, inadequate maps, lack of data, and specific methods are encountered in the practical application of deposit market segmentation.

Dataset Information

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

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

Data Dictionary

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	Number of contacts performed before this campaign and for this client (numeric)
Poutcome	Numeric	Outcome of the previous marketing campaign (categorical: "Failure","Nonexistent","Success")

Social and economic context attributes

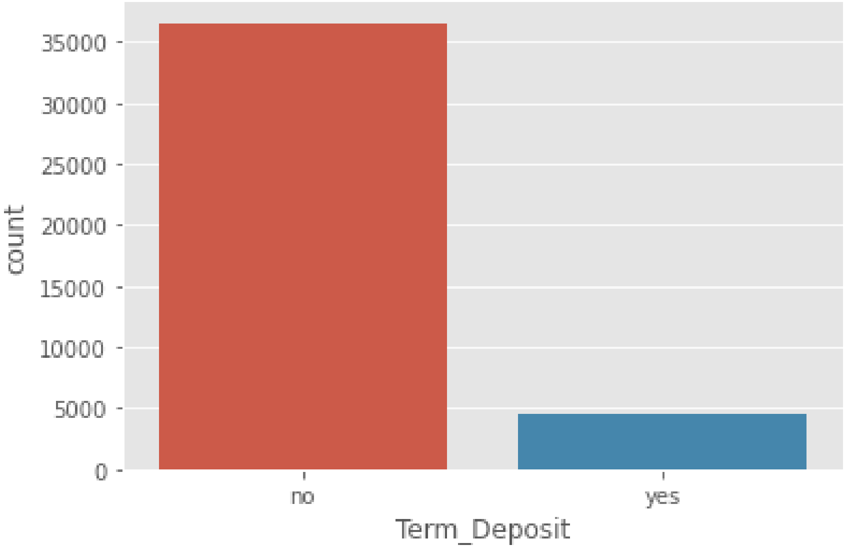
VARIABLE	DATATYPE	DESCRIPTION
Emp.var.rate	Numeric	Employment variation rate - quarterly indicator
Cons.price.idx	Numeric	Consumer price index - monthly indicator
Cons.conf.idx	Numeric	Consumer confidence index - monthly indicator
Euribor3m	Numeric	Euribor 3 month rate - daily indicator
Nr.employed	Numeric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

Output variable (desired target):

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

Target Variable

The target variable of the above dataset is Term_Deposit. We have to predict whether a customer subscribes for a term deposit or not.



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 moderateamount of class imbalance**.

DATA PRE-PROCESSING

Data pre-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 pre-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 pre-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 data consists of 119390 rows and 32 columns. Out of these we have 13 categorical columns and the rest as numerical.

Variable Categorization with Description

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.

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

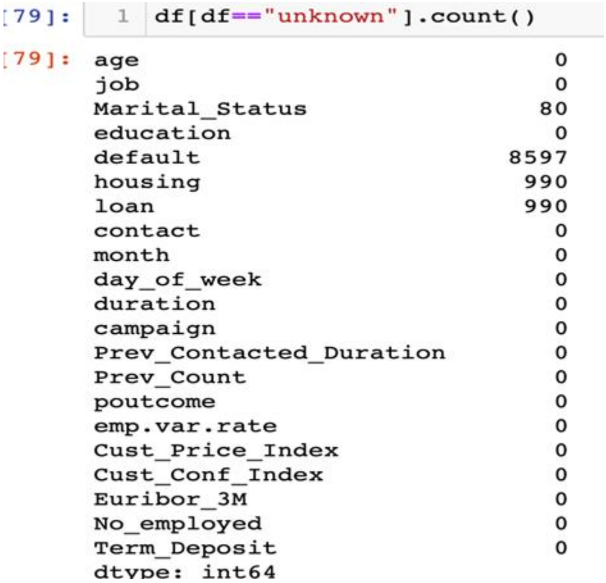
Poutcome	object
Emp.var.rate	float64
Cons.price.idx	float64
Cons.conf.idx	float64
Euribor3m	float64
Nr.employed	float64
Term_Deposit	object

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.



```
[79]:  df[df=="unknown"].count()

[79]:  age                           0
       job                           0
       Marital_Status               80
       education                     0
       default                    8597
       housing                     990
       loan                        990
       contact                       0
       month                         0
       day_of_week                   0
       duration                      0
       campaign                      0
       Prev_Contacted_Duration       0
       Prev_Count                    0
       poutcome                      0
       emp.var.rate                  0
       Cust_Price_Index              0
       Cust_Conf_Index               0
       Euribor_3M                    0
       No_employed                   0
       Term_Deposit                  0
       dtype: int64
```

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

Check for Outliers

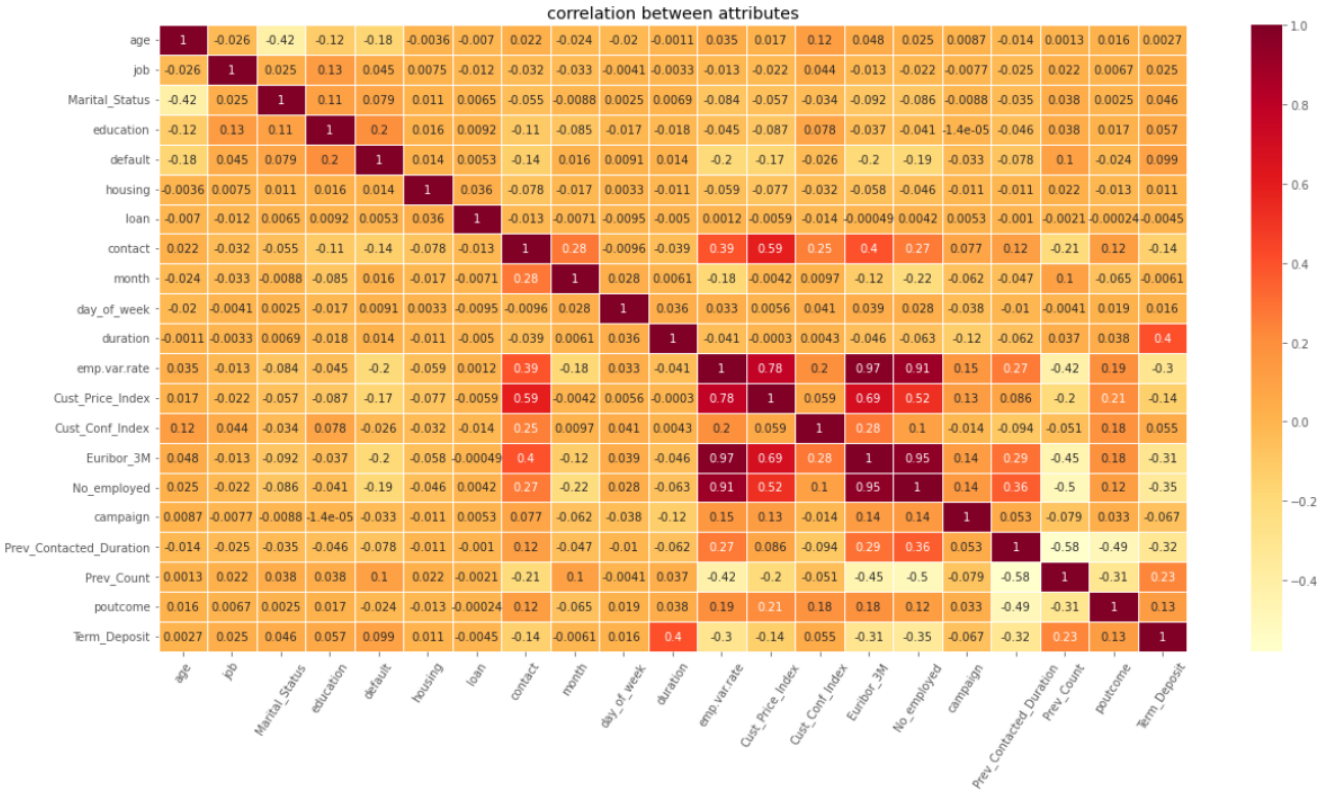
Data has outliers present in each of the numerical columns. For making the base model, we do not perform any outlier treatment and retain all the rows present in the data.

Exploratory Data Analysis

EDA is a way of interpreting, summarising and visualisation the information from dataset. It could help us to find out the patterns and relationships that may not be understood or visible. It is the one of the important things in data science life cycle.
Let’s have a look at some of the Common Analysis we do.

Relationship Between Variables

The Best Way To check the relationship Between Variables is to check the correlation between them
Here is a heatmap of Correlation Matrix to understand the relation between variables



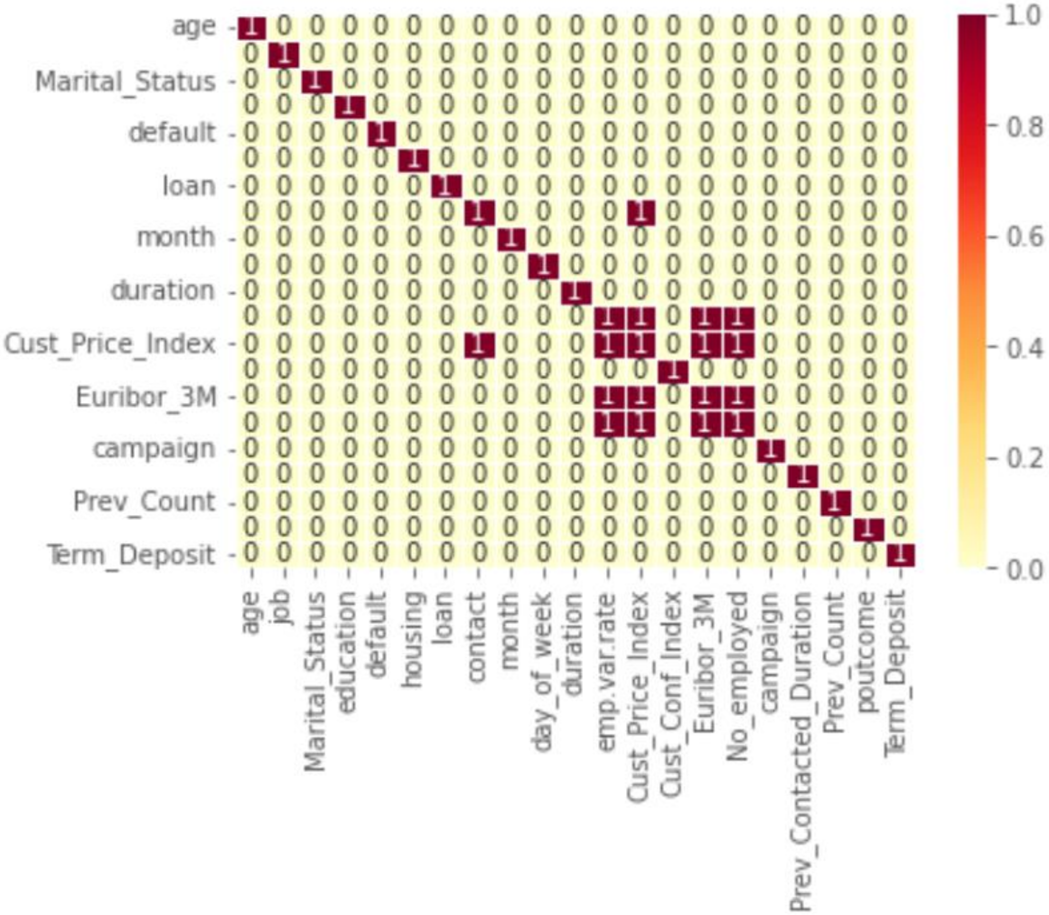
Let’s see with Respect to our Target Variable how does different variables correlate

No_employed	-0.354678
Prev_Contacted_Duration	-0.320945
Euribor_3M	-0.307771
emp.var.rate	-0.298334
contact	-0.144773
Cust_Price_Index	-0.136211
campaign	-0.066532
month	-0.006065
loan	-0.004466
age	-0.002676
housing	-0.011085
day_of_week	-0.015967
job	-0.025482

Marital_Status	-0.045849
Cust_Conf_Index	-0.054878
education	-0.057268
default	-0.099199
poutcome	-0.129789
Prev_Count	-0.230181
duration	-0.404378

From the above matrix of Correlation of variables with respect to our target variable, We can infer that column No of employed and duration are the most correlated to the target variable followed by Prev_Contacted_Duration and Prev_count, whereas age, loan, housing are some of the least correlated variable with respect to our target variable.

Now Let's check If the variables correlation with each other are greater than 0.5 to consider it as a strong relation between each other.



From the above Heatmap We can infer that Customer_Price_Index, Euribor_3M and campaign variables have very high correlation almost equal to 1 so there is a chance of them being multicollinear with each other and only one of the variable is required for model building

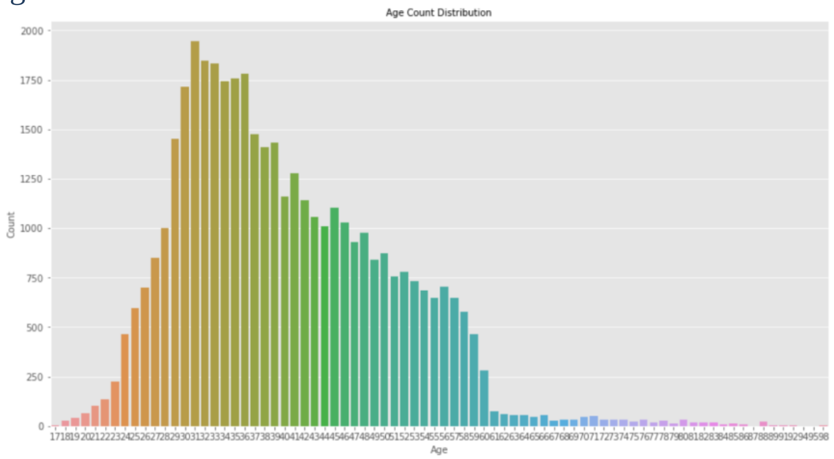
To Check for Multicollinearity

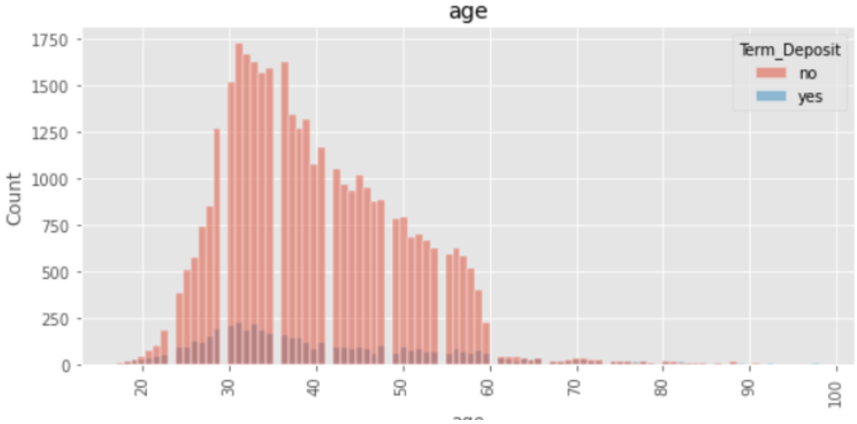
Multicollinearity is the occurrence of high intercorrelations among two or more independent variables in a multiple regression model. Multicollinearity can lead to skewed or misleading results when a researcher or analyst attempts to determine how well each independent variable can be used most effectively to predict or understand the dependent variable in a statistical model.

In general, multicollinearity can lead to wider confidence Intervals that produce less reliable probabilities in terms of the effect of independent variables in a model. We can check for multicollinearity using Variation Inflation Factor. In our Data There seems to be no to little multicollinearity between variables so we continue with our EDA without removal of any attribute.

Distribution Of variables.

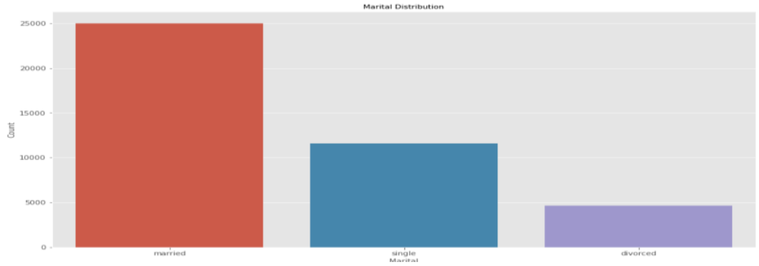
Age:



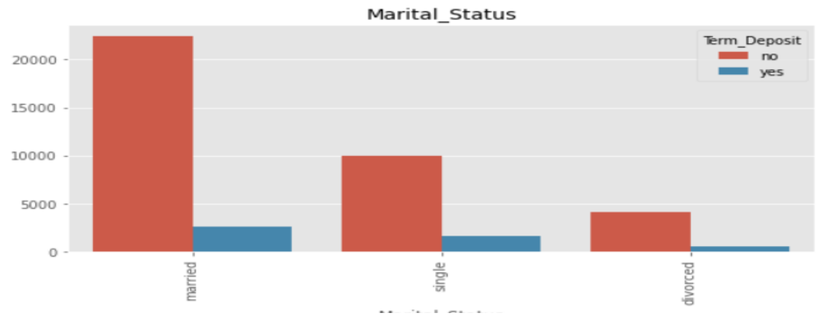


This is Univariate Analysis of the age attribute, As we can see the age can be divided into 3 sections of young medium and old, the old people above 60 are least contacted and people with medium age around 25 to 35 are maximum and hence take the maximum term deposit but old people who are contacted have taken the term deposit with highest conversion ratio as u can see from the bivariate analysis with our target variable.

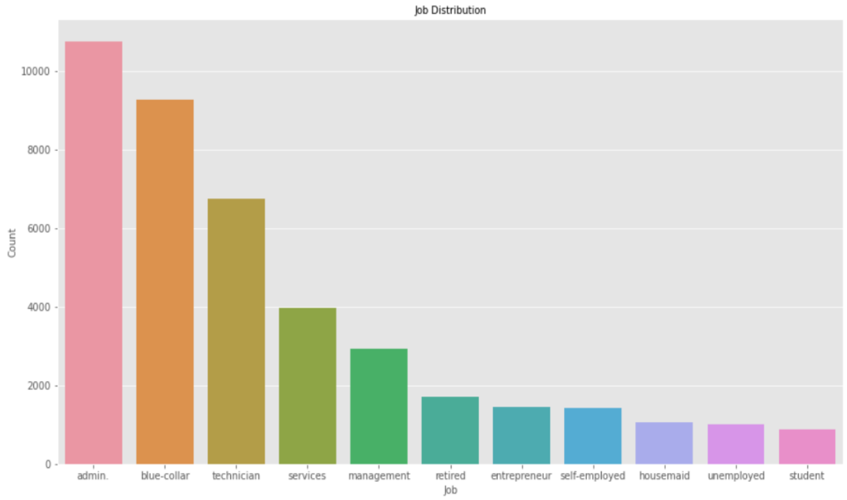
Marital Status:



From these Two plots of Marital Status We can see that married people subscribe to term deposit more compared to other while divorced people are the least

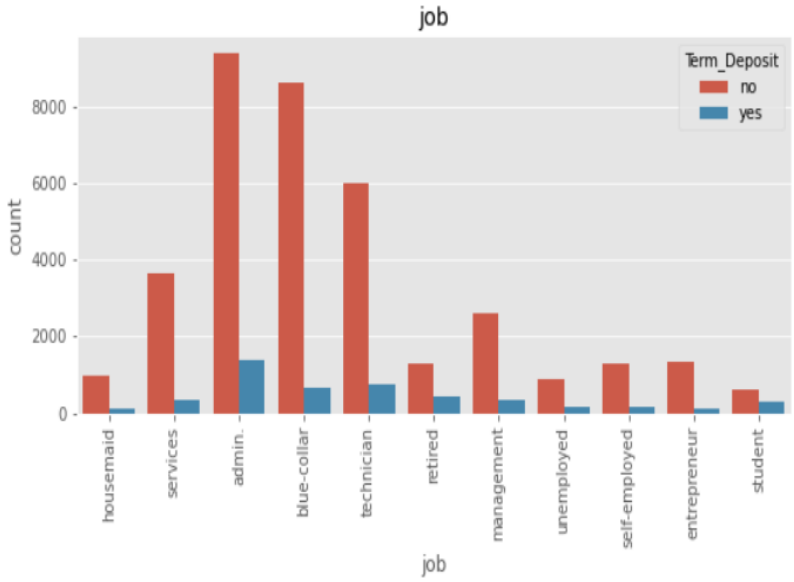


Job:

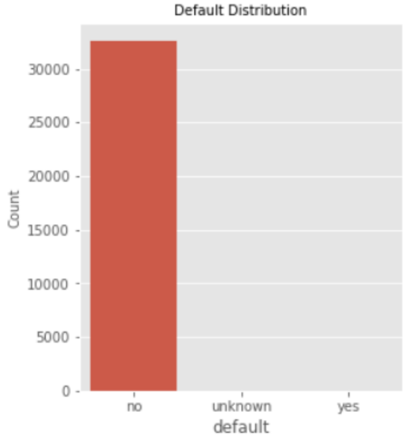


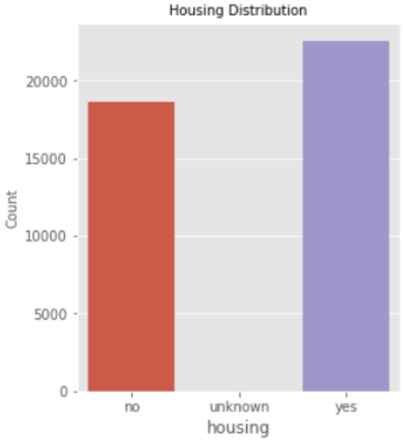
From the above plot we can observe that people with admin jobs have been contacted more by the bank. People with unknown jobs are very few. Let's check people with which jobs have subscribed for the deposits.

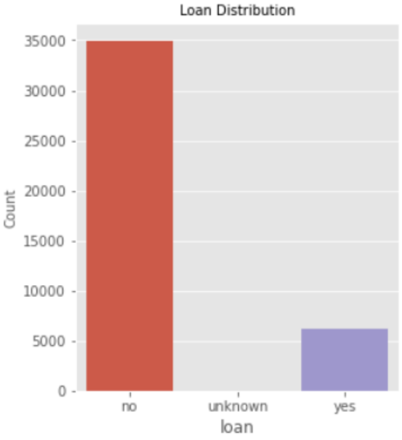
As we can see people with admin job have subscribed the most.

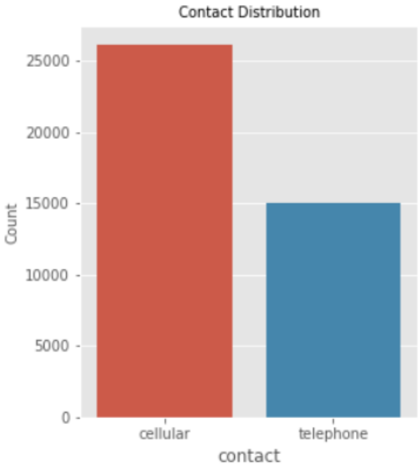


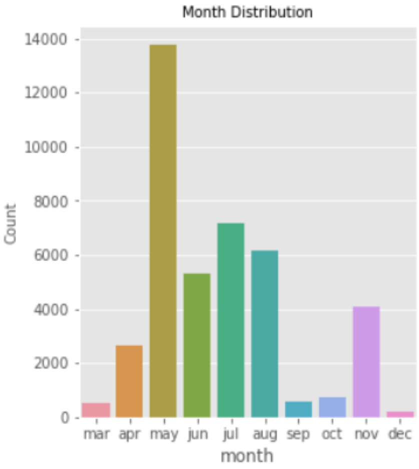
Now let’s check some Univariate analysis distribution of variables.

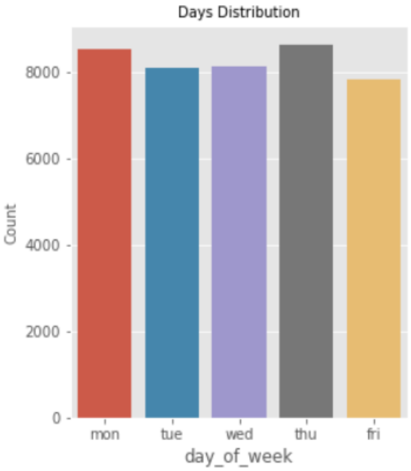


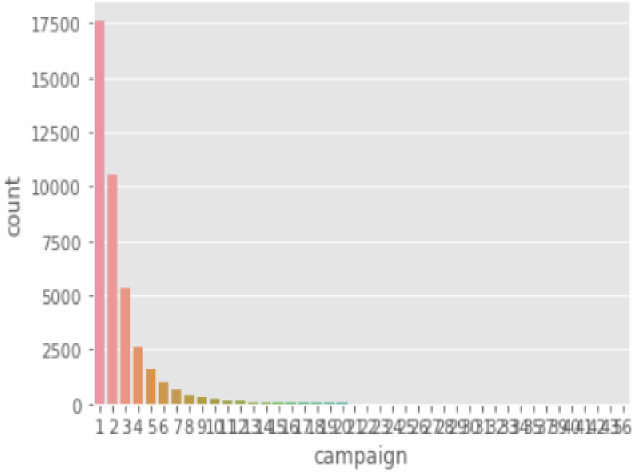


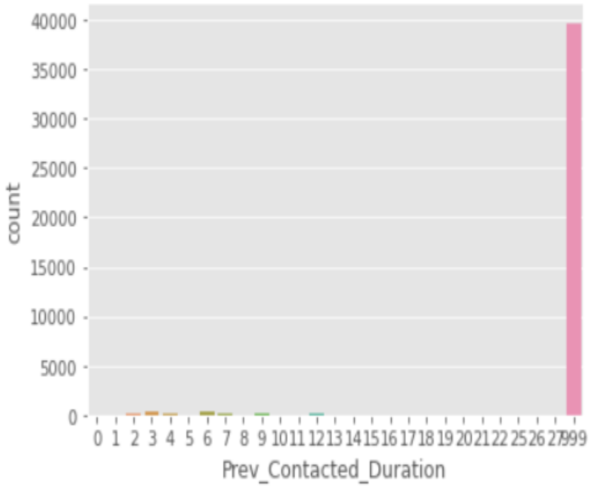




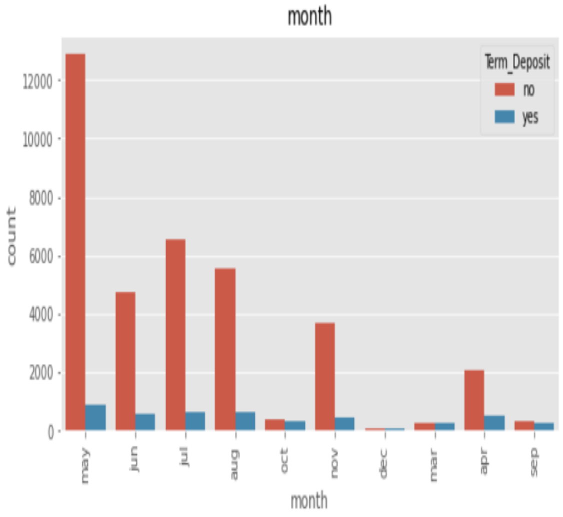


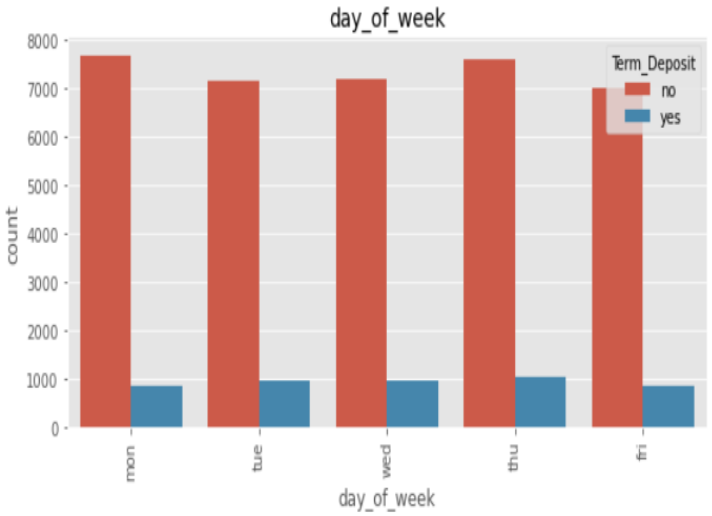


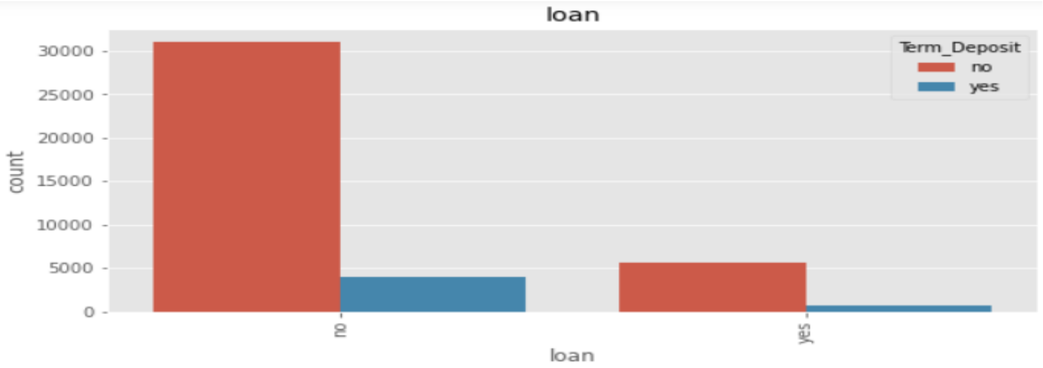


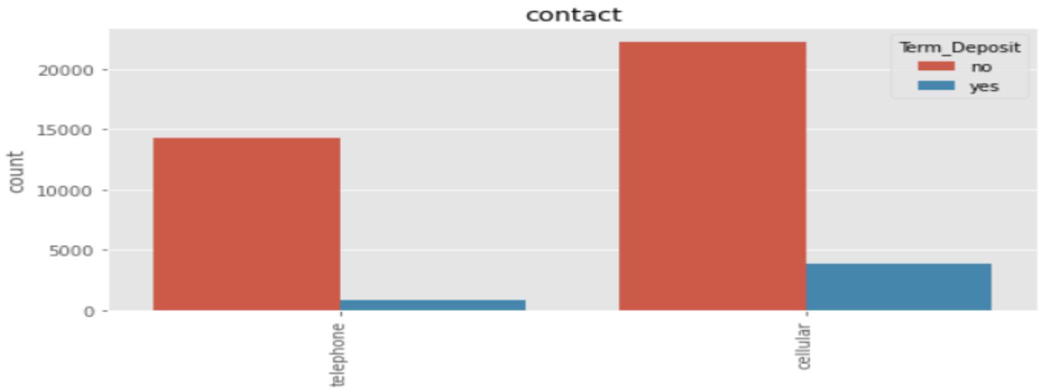


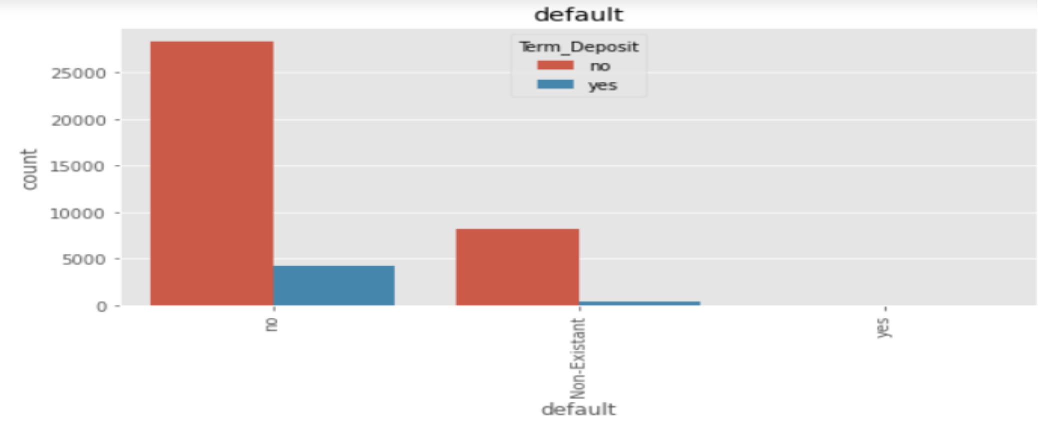
Now let’s check Bivariate analysis distribution of the remaining variables.

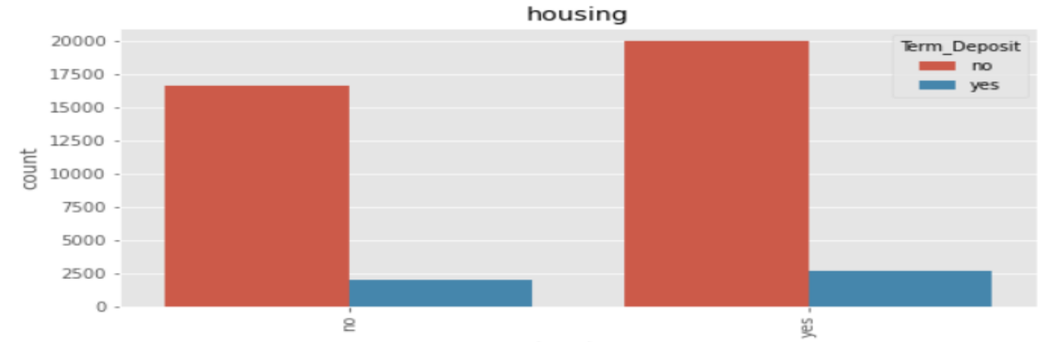

month
count
0
2000
4000
6000
8000
10000
12000
Term_Deposit
no
yes
may
jun
jul
aug
oct
nov
dec
mar
apr
sep
month

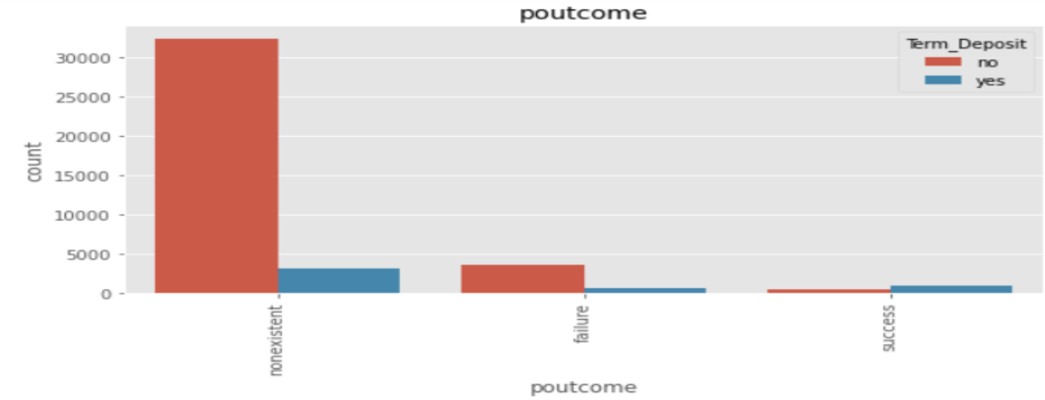

day_of_week
count
0
1000
2000
3000
4000
5000
6000
7000
8000
Term_Deposit
no
yes
mon
tue
wed
thu
fri
day_of_week


loan
count
0
5000
10000
15000
20000
25000
30000
Term_Deposit
no
yes
no
yes
loan

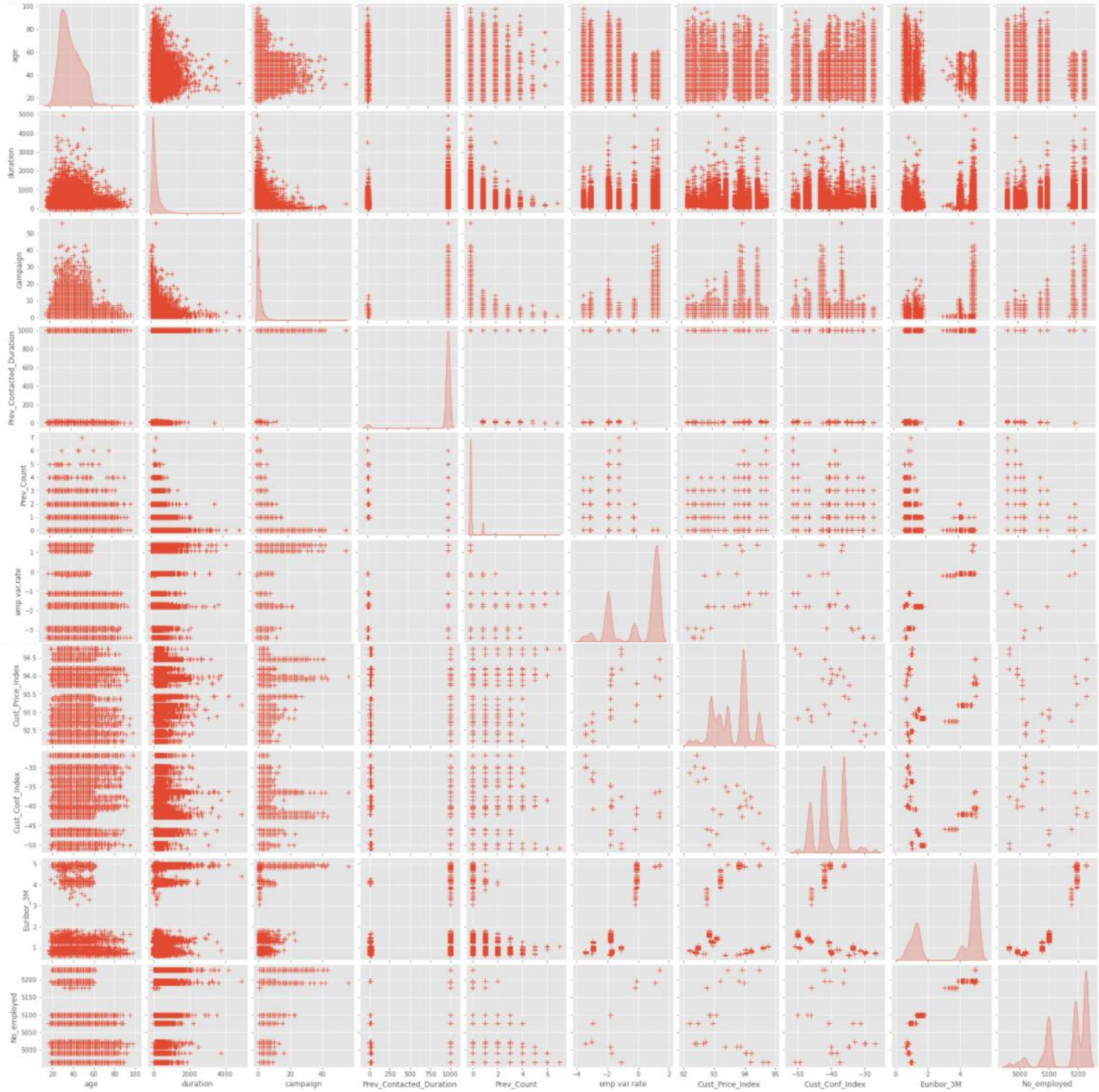

contact
count
0
5000
10000
15000
20000
Term_Deposit
no
yes
telephone
cellular


default
Term_Deposit
no
yes
count
25000
20000
15000
10000
5000
0
no
Non-Existant
yes
default


housing
Term_Deposit
no
yes
count
20000
17500
15000
12500
10000
7500
5000
2500
0
no
yes


poutcome
Term_Deposit
no
yes
count
30000
25000
20000
15000
10000
5000
0
nonexistent
failure
success
poutcome

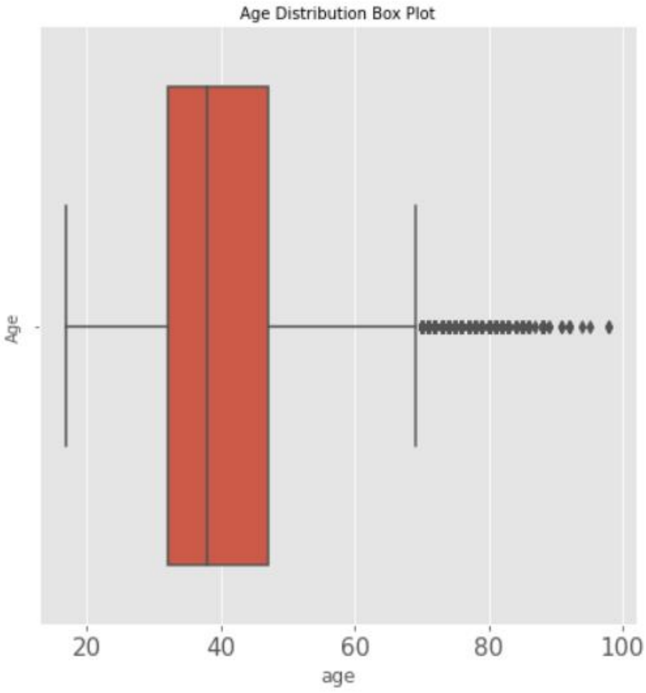
Multivariate Analysis:

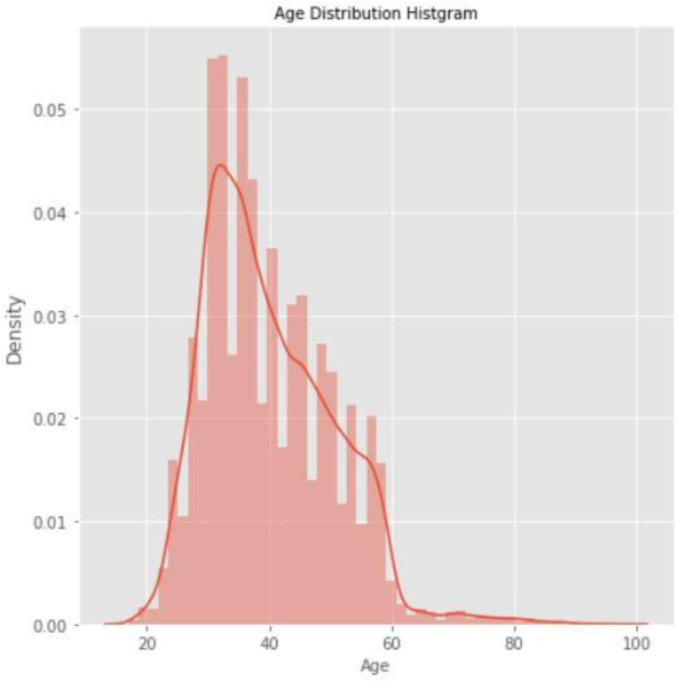


Detection of Outlier and it’s treatment

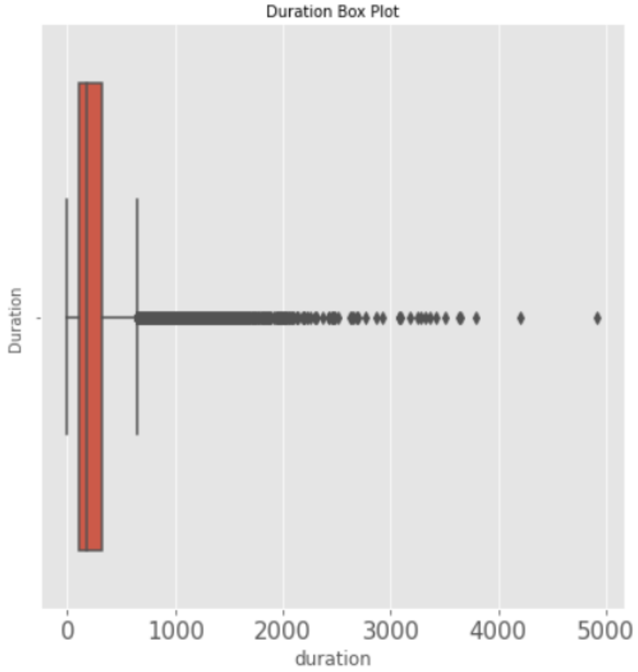
We have checked for outliers for all of the numerical variables using Box Plot but most of them didn’t had any outliers except 2 which are age and duration attributes, so to treat them we have tried multiple approaches such as binning in which we converted the numerical variables into categorical variables of 3-4 values by setting a limit eventhough this was working but we didn’t want the numerical variable be treated as categorical so, we kept them as it is, we could have also performed IQR treatment for it but we didn’t want to remove any rows and keep them as it is so we transformed age using log transformation and duration using square root transformation to get its skewness reduced and make it as normal as possible.

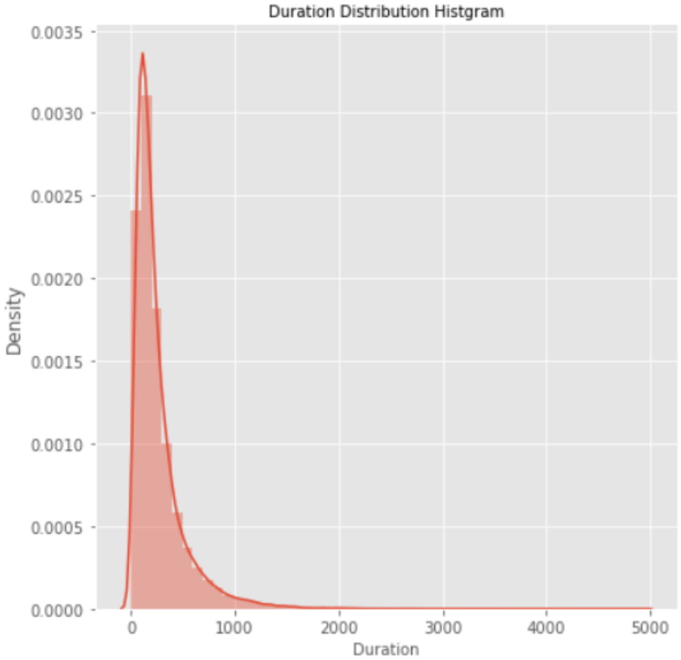
Age:





Duration:



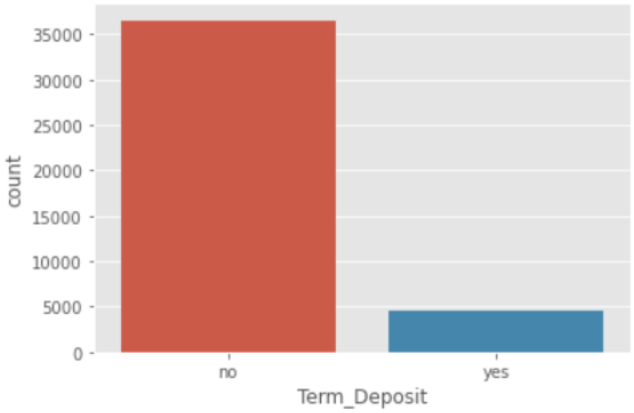


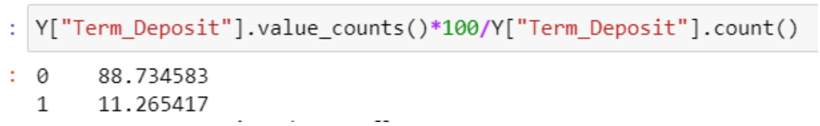
Handling the Imbalanced Data

Imbalance data refers to the categorical dataset in which class distribution is not uniform. This means the distribution of classes is unequal in the dataset. For example, you have a dataset with 1000 records and 2 classes(Yes/No). Out of 1000, only 50 records belong to class ‘Yes’ and the remaining 950 records to class ‘No’. The distribution of records among class ‘Yes’ and ‘No’ is unequal.

Undersampling, Oversampling and generating synthetic data.: These methods are often presented as great ways to balance the dataset before fitting a classifier on it. In a few words, these methods act on the dataset as follows:

- undersampling consists in sampling from the majority class in order to keep only a part of these points
- oversampling consists in replicating some points from the minority class in order to increase its cardinality
- generating synthetic data consists in creating new synthetic points from the minority class (see SMOTE method for example) to increase its cardinality

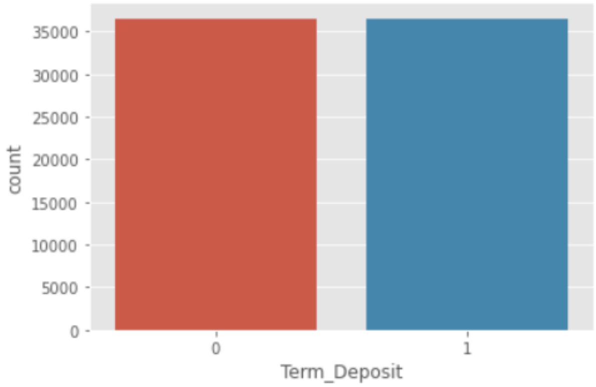


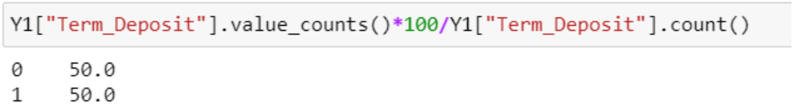


```
: Y["Term_Deposit"].value_counts()*100/Y["Term_Deposit"].count()

: 0    88.734583
  1    11.265417
```

To deal with this we use SMOTE technique to add more synthetic rows although we have built model on both imbalanced and balanced data

After SMOTE:





```
Y1["Term_Deposit"].value_counts()*100/Y1["Term_Deposit"].count()

0    50.0
1    50.0
```

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.

Feature Engineering

From the exploratory analysis carried out, the target variable was an imbalance, and the resampling technique which is a function in python was applied to the balance of the dataset for better analysis. Before the balancing of the data, 89% of the customers did not subscribe to a term deposit of the bank while the remaining 11% of customers manage to subscribe.

Encoding of categorical Features

To convert the categorical features to numerical values we use encoding. For this particular dataset we have used Label Encoder, which is part of Scikit-learn library. This approach involves converting each value in a column to a number. In this technique, each label is assigned a unique integer based on alphabetical ordering.

We perform label encoding for numerical features such as campaign, Prev_Count, Poutcome and Prev_contacted_duration as these features contain discrete fixed values.

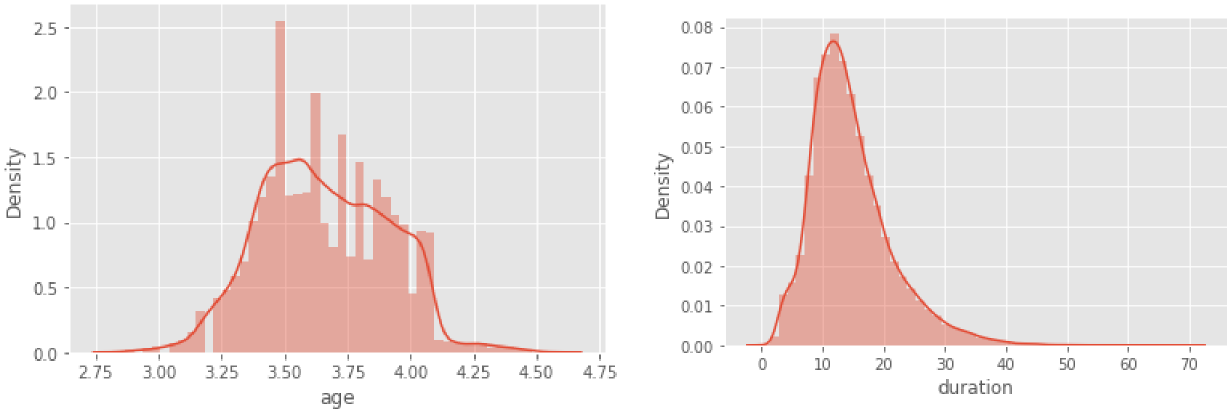
Feature Scaling of Numerical Features

Scaling of the data comes under the set of steps of data pre-processing when we are performing machine learning algorithms in the data set. In the machine learning algorithms if the values of the features are closer to each other there are chances for the algorithm to get trained well and faster instead of the data set where the data points or features values have high differences with each other will take more time to understand the data and the accuracy will be lower.

Like normalization, standardization is also required in some forms of machine learning when the input data points are scaled in different scales. Standardization can be a common scale for these data points. The basic concept behind the standardization function is to make data points centred about the mean of all the data points presented in a feature with a unit standard deviation. This means the mean of the data point will be zero and the standard deviation will be 1. This technique also tries to scale the data point between zero to one but in it, we don't use max or minimum. Here we are working with the mean and the standard deviation.

Feature Transformation

Reason of doing log transformation is that, most of our dataset are right skewed. In data analysis transformation is the replacement of a variable by a function of that variable. The logarithm, x to log base 10 of x, or x to log base e of x (ln x), or x to log base 2 of x, is a strong transformation with a major effect on distribution shape. It is commonly used for reducing right skewness and is often appropriate for measured variables. It can not be applied to zero or negative values. One unit on a logarithmic scale means a multiplication by the base of logarithms being used.



Data distribution after Log Transformation

MODEL DEPLOYMENT

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 model.

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 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:

- Logistic Regression
- KNN classifier
- Random Forest Classifier

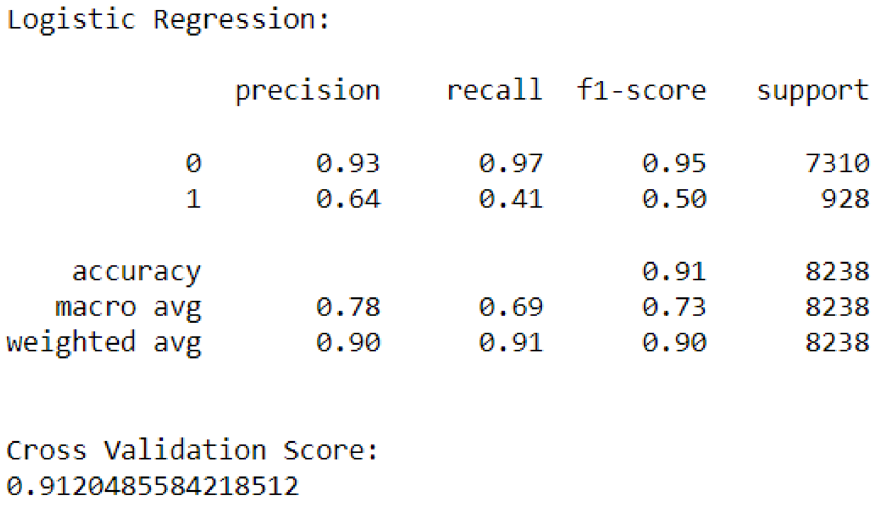
LOGISTIC REGRESSION

Logistic regression is a traditional machine model that fits a linear decision boundary between the positive and negative samples. Logistic regression uses a line (Sigmoid function) in the form of an "S" to predict if t dependent variable is true or false based on the independent variables.
One advantage of logistic regression is the model is interpretable, we know how features are important for predicting positive or negative. Take note that the modeling is sensitive to the scaling of the features.
we will use logistic regression to build a model. To evaluate the accuracy of these logistic regression models, we will analyze AUC, predicted accuracy, and weighted accuracy. AUC measures the area under the ROC Curve; thus, predicting true positives more accurately in the model will maximize it.

Logistic regression without SMOTE

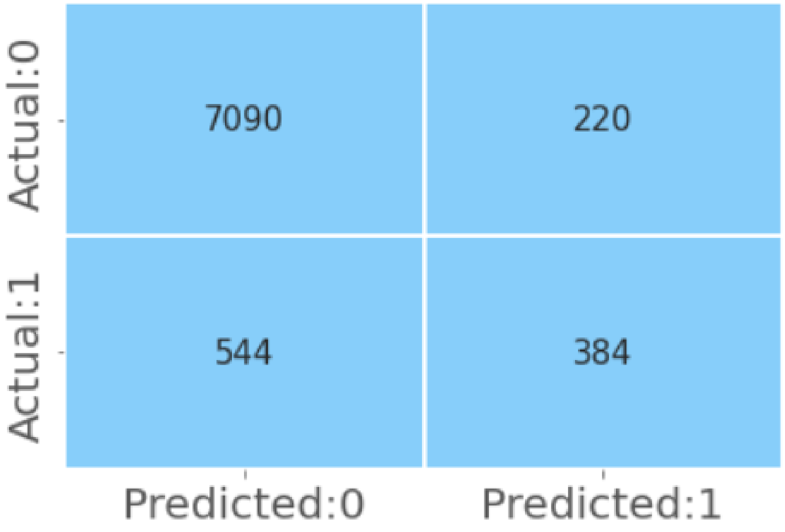
On applying logistic regression algorithm, we get the following observations:

CLASSIFICATION REPORT:

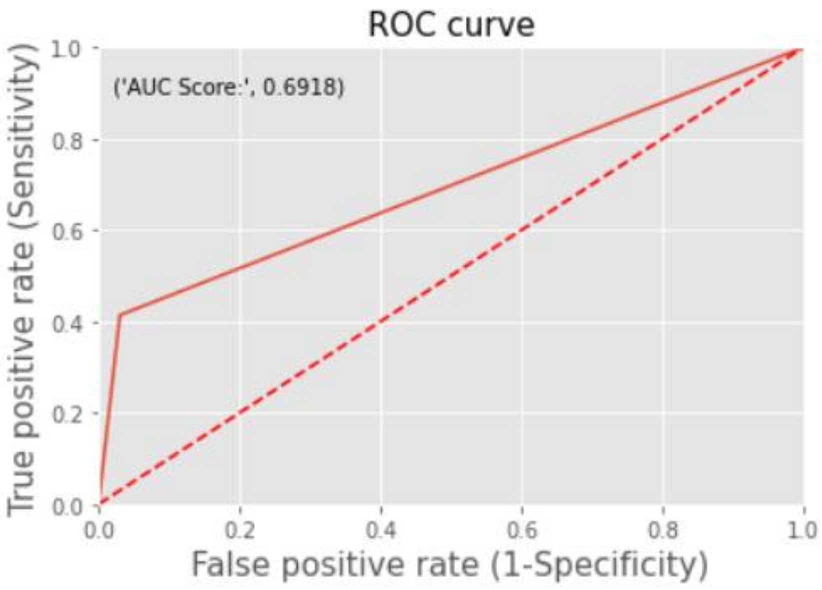


```
Logistic Regression:

              precision    recall  f1-score   support

           0       0.93      0.97      0.95      7310
           1       0.64      0.41      0.50       928

    accuracy                           0.91      8238
   macro avg       0.78      0.69      0.73      8238
weighted avg       0.90      0.91      0.90      8238


Cross Validation Score:
0.9120485584218512
```

CONFUSION MATRIX

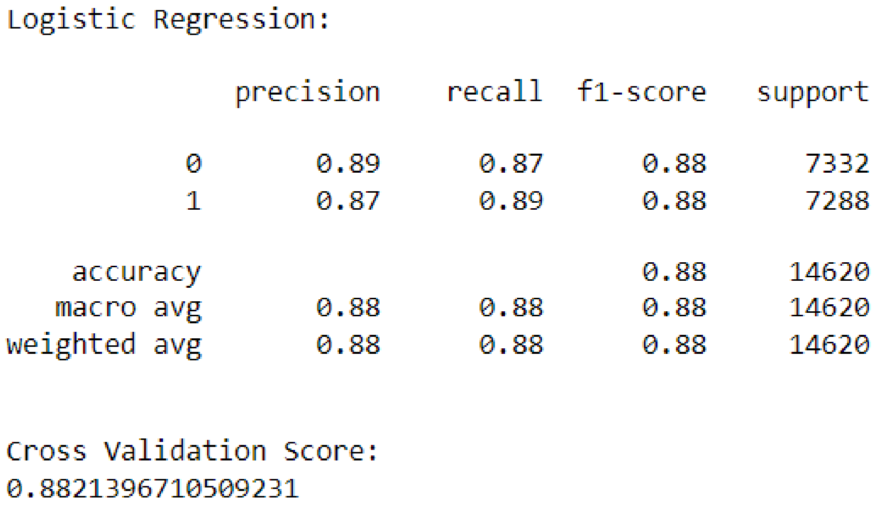


ROC CURVE



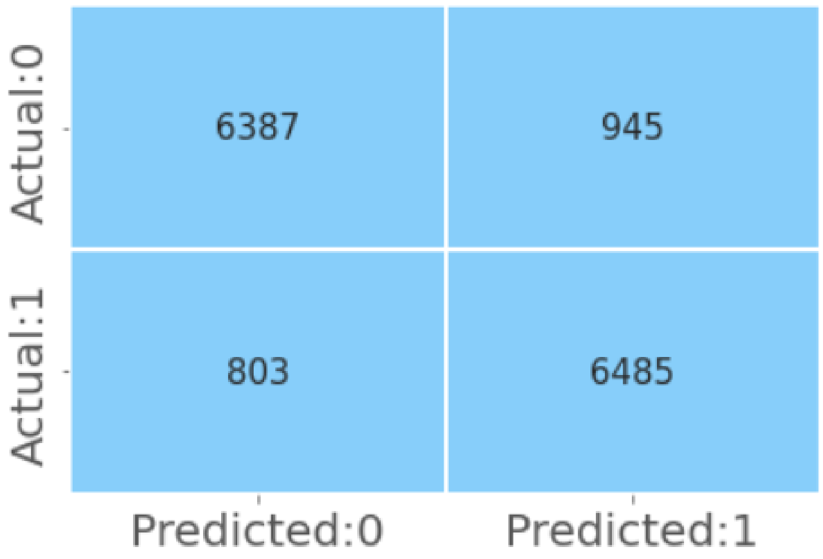
LOGISTIC REGRESSION – BASE MODEL WITH SMOTE

CLASSIFICATION REPORT



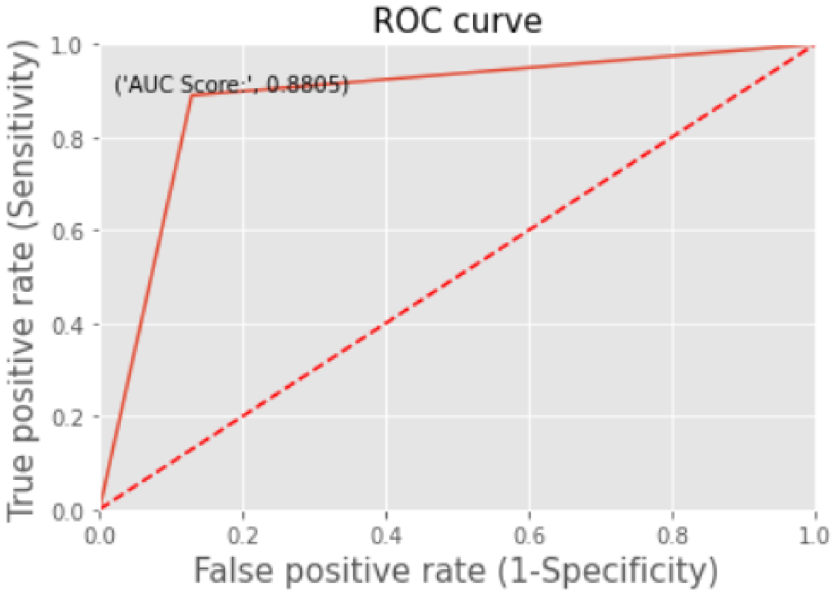
```
Logistic Regression:

              precision    recall  f1-score   support

           0       0.89      0.87      0.88      7332
           1       0.87      0.89      0.88      7288

    accuracy                           0.88     14620
   macro avg       0.88      0.88      0.88     14620
weighted avg       0.88      0.88      0.88     14620


Cross Validation Score:
0.8821396710509231
```

CONFUSION MATRIX



	Predicted:0	Predicted:1
Actual:0	6387	945
Actual:1	803	6485

ROC CURVE

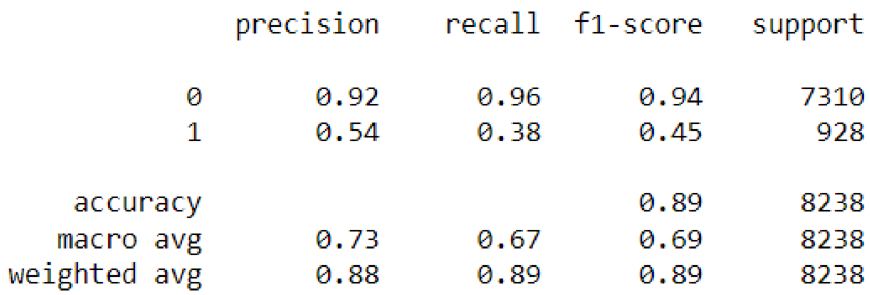


KNN CLASSIFIER

KNN algorithms use data and classify new data points based on similarity measures (e.g. distance function). Classification is done by a majority vote to its neighbors. The data is assigned to the class which has the nearest neighbors. As we increase the number of nearest neighbors, the value of k, accuracy might increase.

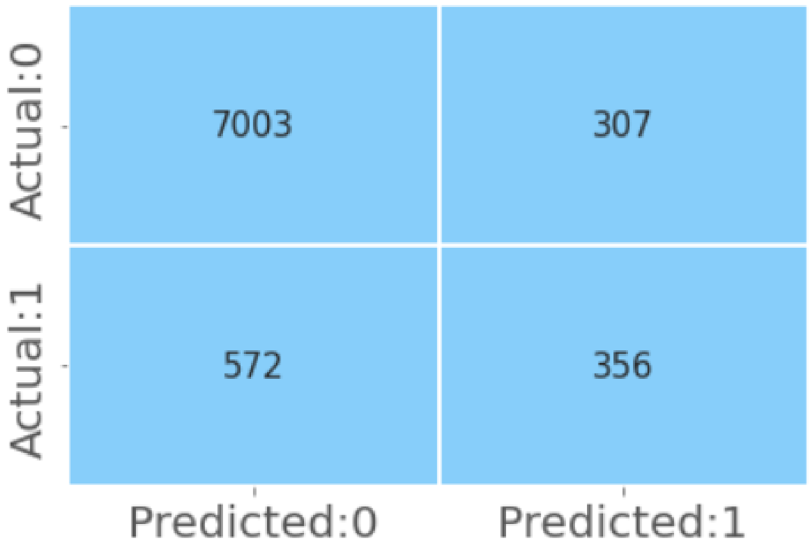
KNN – BASE MODEL WITHOUT SMOTE

CLASSIFICATION REPORT



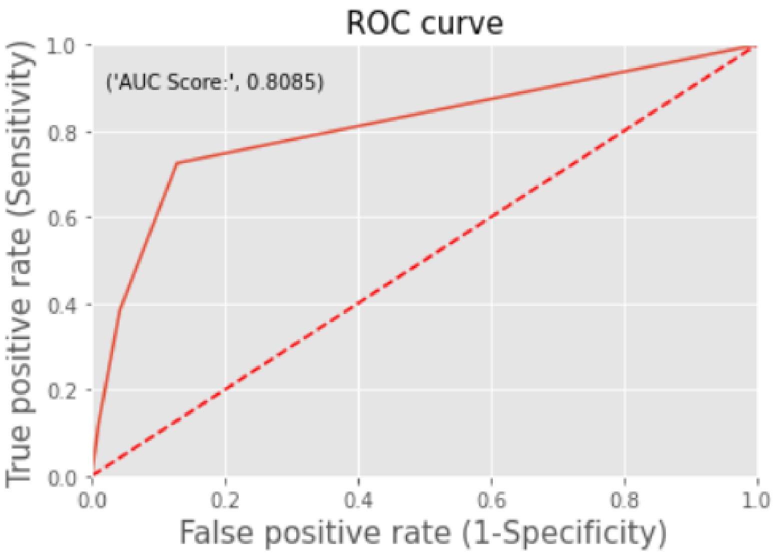
```
              precision    recall  f1-score   support

           0       0.92      0.96      0.94      7310
           1       0.54      0.38      0.45       928

    accuracy                           0.89      8238
   macro avg       0.73      0.67      0.69      8238
weighted avg       0.88      0.89      0.89      8238
```

CONFUSION MATRIX



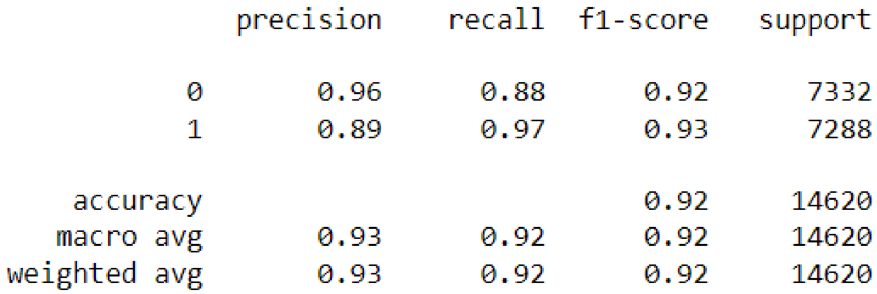
	Predicted:0	Predicted:1
Actual:0	7003	307
Actual:1	572	356

ROC CURVE



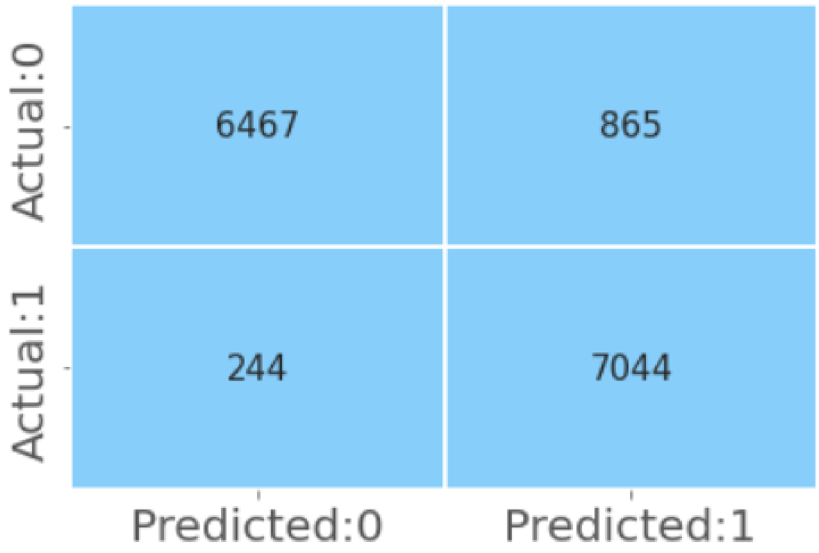
KNN – BASE MODEL WITH SMOTE

CLASSIFICATION REPORT



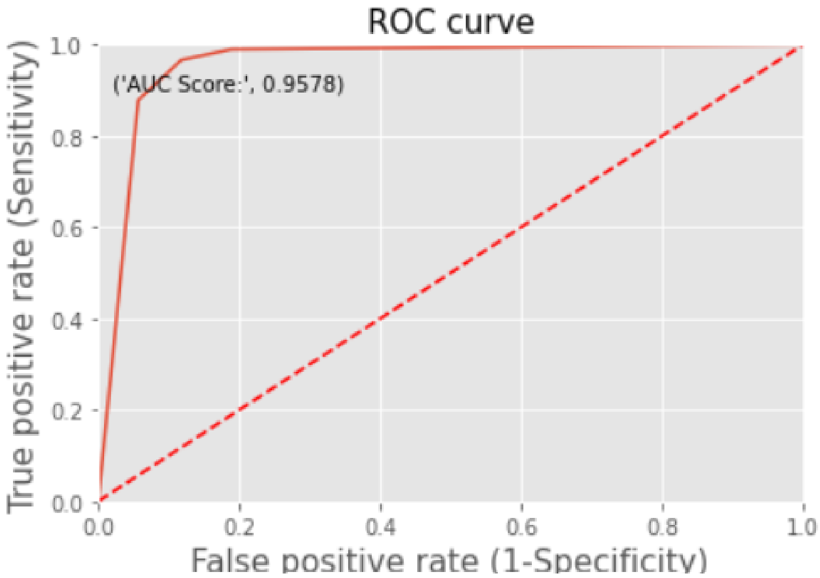
```
              precision    recall  f1-score   support

           0       0.96      0.88      0.92      7332
           1       0.89      0.97      0.93      7288

    accuracy                           0.92     14620
   macro avg       0.93      0.92      0.92     14620
weighted avg       0.93      0.92      0.92     14620
```

CONFUSION MATRIX



	Predicted:0	Predicted:1
Actual:0	6467	865
Actual:1	244	7044

ROC CURVE



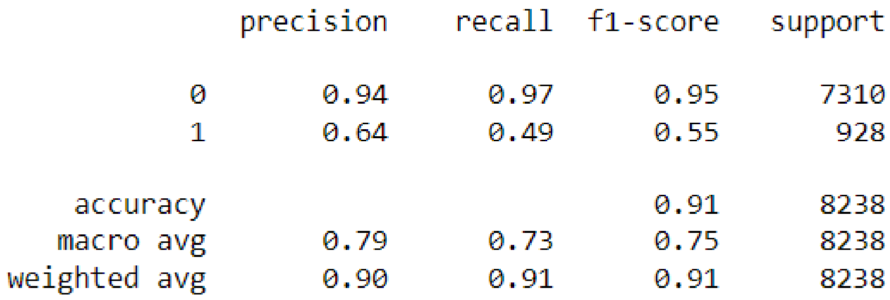
With smote there is significant change in confusion matrix in true positive and true negative rates.

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.

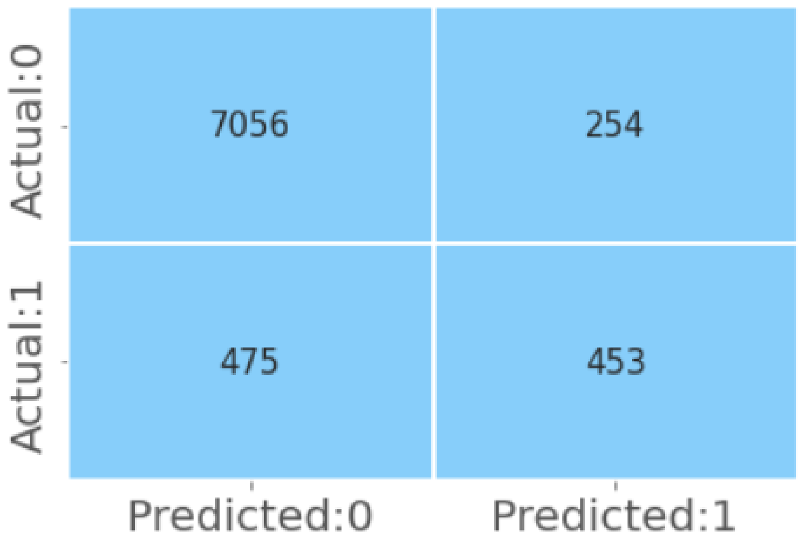
Random Forest – BASE MODEL WITHOUT SMOTE

CLASSIFICATION REPORT



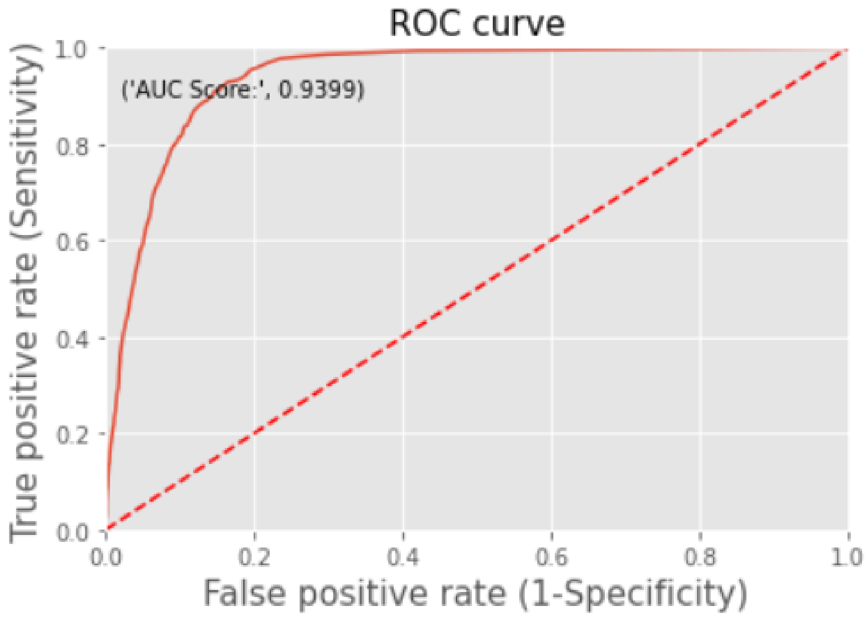
```
              precision    recall  f1-score   support

           0       0.94      0.97      0.95      7310
           1       0.64      0.49      0.55       928

    accuracy                           0.91      8238
   macro avg       0.79      0.73      0.75      8238
weighted avg       0.90      0.91      0.91      8238
```

CONFUSION MATRIX

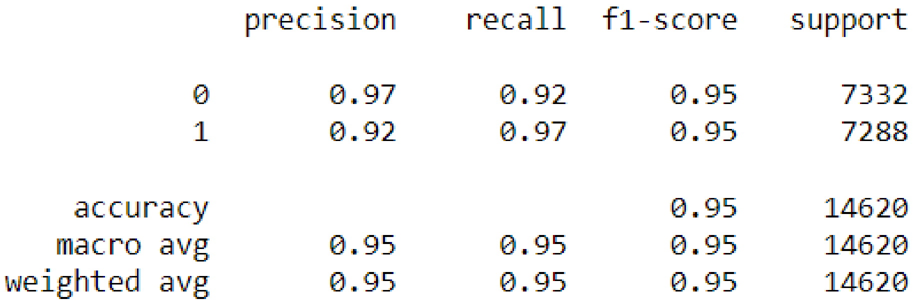


	Predicted:0	Predicted:1
Actual:0	7056	254
Actual:1	475	453

ROC CURVE

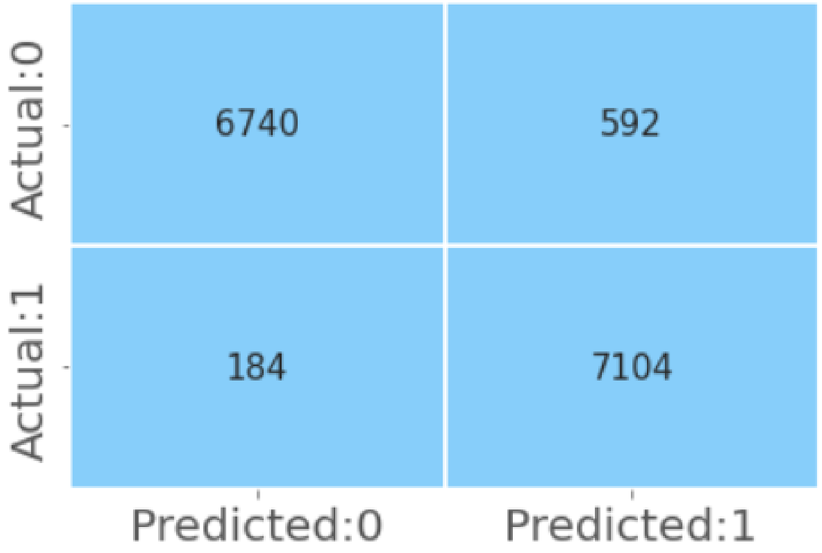


Random Forest – BASE MODEL WITH SMOTE
CLASSIFICATION REPORT



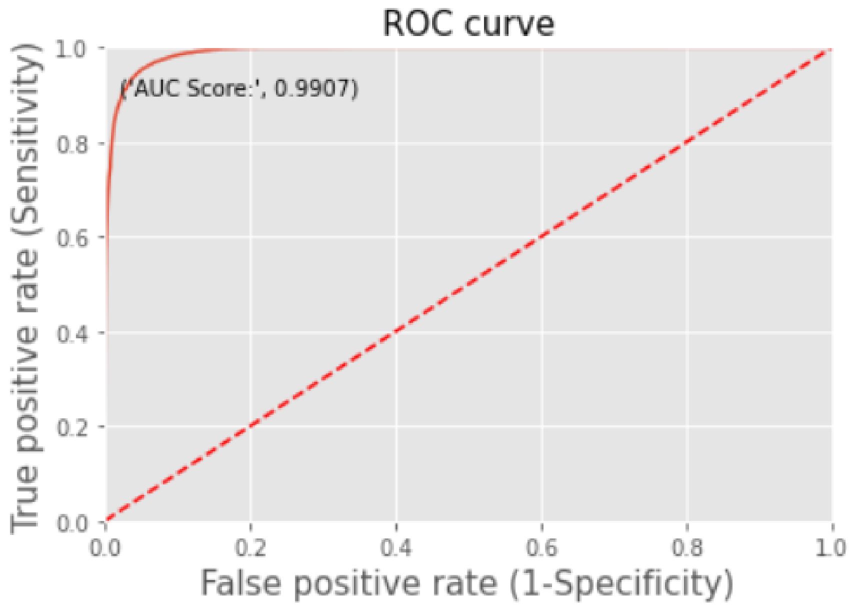
```
              precision    recall  f1-score   support

           0       0.97      0.92      0.95      7332
           1       0.92      0.97      0.95      7288

    accuracy                           0.95     14620
   macro avg       0.95      0.95      0.95     14620
weighted avg       0.95      0.95      0.95     14620
```

CONFUSION MATRIX



	Predicted:0	Predicted:1
Actual:0	6740	592
Actual:1	184	7104

ROC CURVE



With smote there is significant change in confusion matrix in true positive and true negative rates.

FUTURE WORK

Further after building the base model, we will proceed with building non-linear models followed by feature selection and hyperparameter tuning. Also, from the EDA we learnt that various categorical columns have high cardinality. Cardinality means presence there is presence of many unique values in a categorical column. Encoding such columns with One Hot Encoding leads to increase in the dimension of the dataset and higher computation. Hence, we need to explore methods to deal with such columns. Along with this according to our business problem we have to select an evaluation metric to compare the results of various models which we will build.