ALY 6140 – Analytics Systems Technology

Professor: Rudhramoorthi Daya

Northeastern University, Boston

Logo

Description automatically generated

Prediction of telemarketing success in banks

Capstone Project Proposal

Submitted by – Santhi Kiran Chavali (Group -17)

**Introduction:**

Out of all the industries in that machine learning is transforming, the financial services sector is leading the way. Due in large part to digitization, banks are producing and storing an expanding amount of data. These data may be utilized by banks to create predictive analytics and use the knowledge to grow their businesses. The applications range widely, beginning with stock forecasting, marketing efficacy, term loan default risk, fraud detection in credit card transactions, and customer service utilizing natural language processing. One area where machine learning has made a significant difference is in the success of marketing campaigns; it has enabled marketing teams to better understand what is and is not effective. The big marketing budget, which is one of the key expenses on the profit and loss statement, is reduced as a result of their ability to focus on particular engagement mediums that are effective.

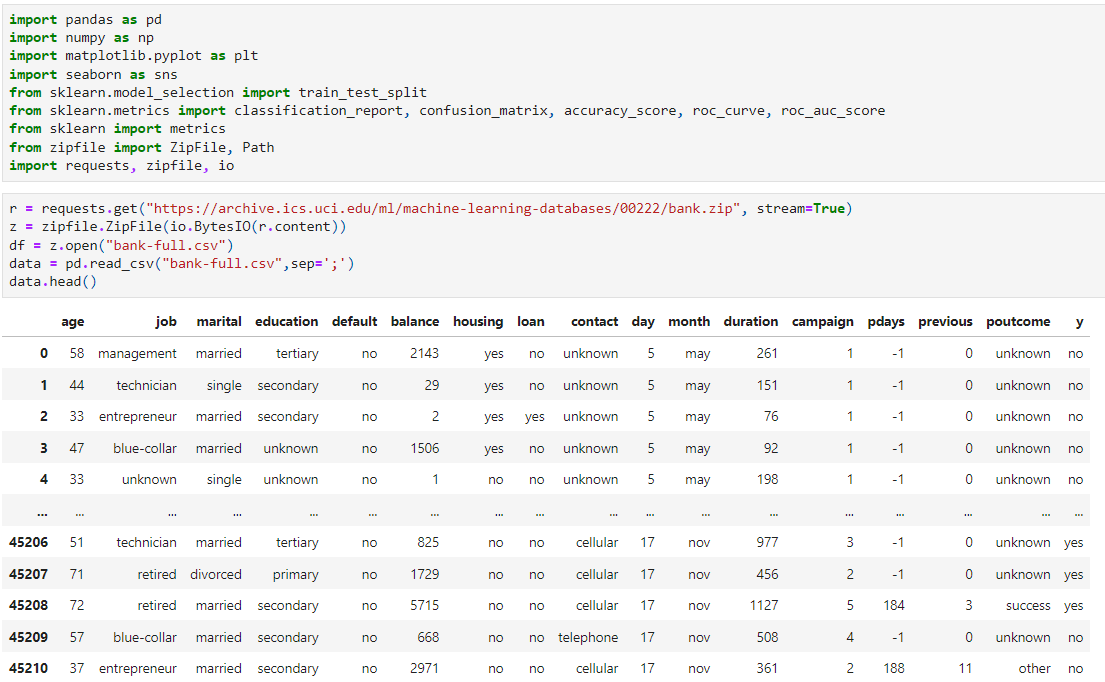
For the final project, I have selected the Bank marketing dataset from the UC Irvine Machine Learning Repository website (https://archive-beta.ics.uci.edu/dataset/222/bank+marketing). The dataset contains the details of the telemarketing campaigns conducted by a Portuguese banking institution. The main goal is to predict whether a customer is willing to make a term deposit while contacted through the telemarketing campaign. This project aims to put the problem into practice, try to increase model accuracy, and fit machine learning models like logistic regression, Decision tree models, and other techniques for feature engineering data into use, which should help produce better results.

**Problem Statement:**

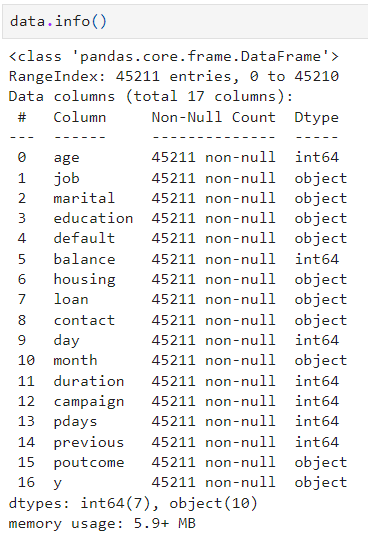
The main aim of the project is to use the available data to predict whether the customer going to subscribe to the term deposit or not.

**Exploratory Data Analysis:**

1. First I have imported all the necessary libraries like pandas, numpy, matpotlib, etc. The dataset is imported directly from the website to jupyter notebook using requests.get() function. The required file is in Zip format so further extracted the zip file which consists of three datasets and stored the required dataset csv file to jupyter notebook using zipfile library. Used pd.read\_csv() function in pandas to read the csv file.

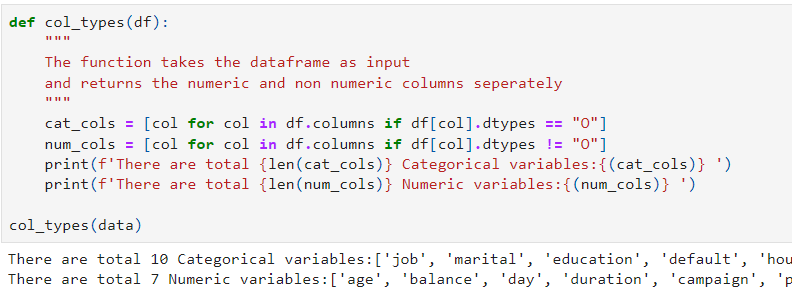


1. To get the overall information about the dataset I have used the info() function.



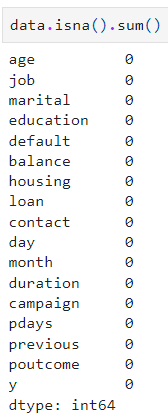
From the above image, we can see that the dataset consists of 45211 records/rows with 17 columns/variables. Out of the 17 variables, we have 7 numeric variables(age, balance, day, duration, campaign, pdays, previous) and 10 non-numeric variables(job, marital, education, default, housing, loan, contact, month, poutcome, y).

1. A self-defined function col\_types() is created which takes the dataset as input and gives the numeric and non-numeric variables separately.

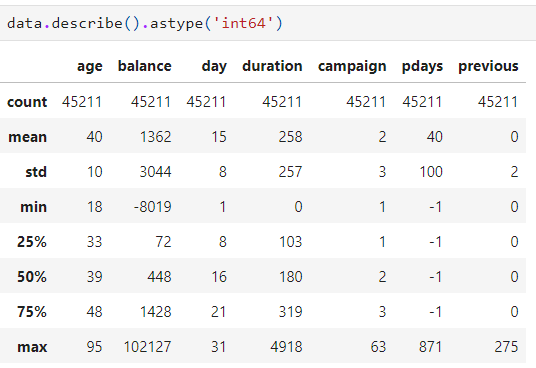


From the above image we can see that, using the function we can get the numeric and non-numeric columns separately.

1. To check the missing values/na values isna().sum() is used and found that the dataset does not contain any missing values.



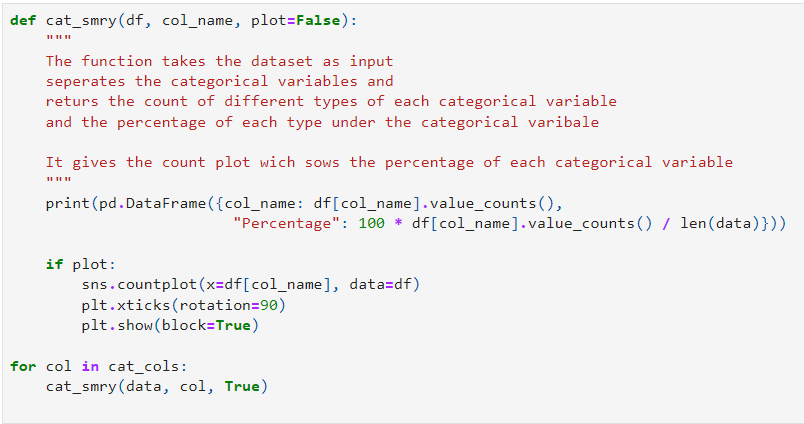
1. Descriptive statistics of numeric variables are found using the describe() function.



It is observed that:

* The age of the customers in the dataset ranges from 18 to 95 years and 40 being the average age.
* The balance maintained by the customers in their bank accounts in the dataset is between -8019 to 102127 euros.
* The duration of calls made during the campaigns range from 0 seconds to 4918 seconds and the average call is for 258 seconds
* Pdays are the number of days since a previous campaign's final interaction with the client. It is between -1 to 871 days.
* Previous column shows how many contacts were made for this client and before that particular campaign which ranges from 0 to 275.

1. To get the overall picture of categorical variables, I have created a function cat\_smry() which gives the detailed structure along with plots.



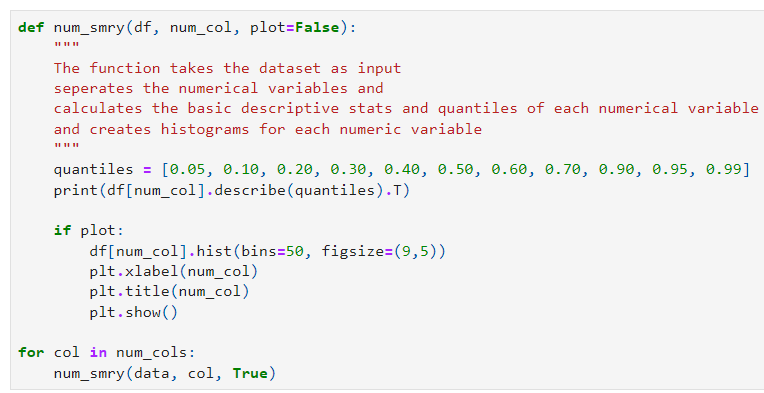
Plots of all Categorical variables:

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

From the above plots, it is observed that:

* Blue collar, management, and technician job holders are targeted mostly during the campaigns conducted and students are given the least preference.
* Calls were made mostly to married customers and divorced customers got fewer calls.
* Customers with secondary education were concentrated more when compared to tertiary and primary education.
* Customers with no default status, having a personal loan, and a housing loan are given priority to make calls. And, most of the calls were made to cellular phone holders.
* Maximum calls were made during the May month and December being the least.
* Out of the calls made during the campaigns only 5289 customers that is 11.70 percent are willing to subscribe for a term deposit.

1. To get the overall picture of numerical variables, I have created a function num\_smry() which gives the detailed structure along with plots.



Plots for the numerical variables:

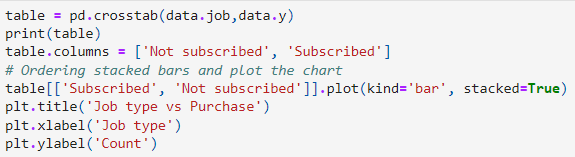
|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

It is observed from the above histograms that targeted age group is between 30 to 40 with an average balance of 1362 euros and more than that and the average duration of the call was 258 seconds. More campaigns were conducted during the earlier stage and gradually decreased in later years. And, the average number of calls made previously to particular customers is 0.5.

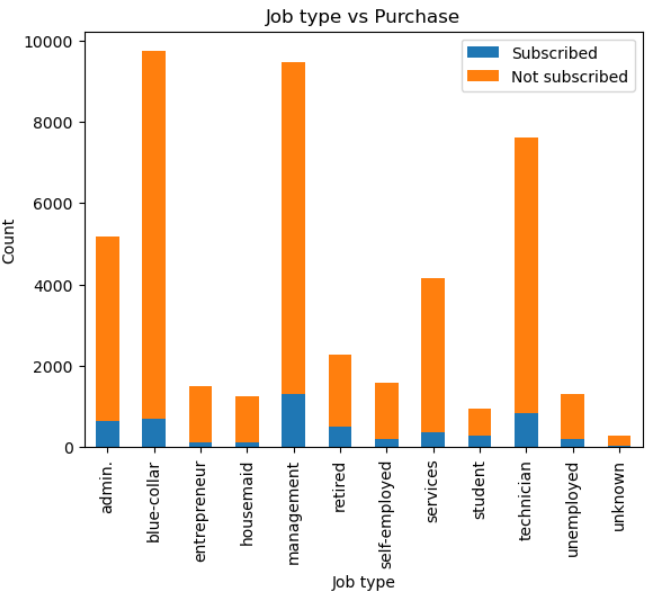
**Analysis:**

For the analysis part I have focused on three variables Job, education, and month and compared with whether customers are willing to subscribe for the term deposit. I have used pd.crosstab() to perform this task.

**Job type Vs Purchase:**

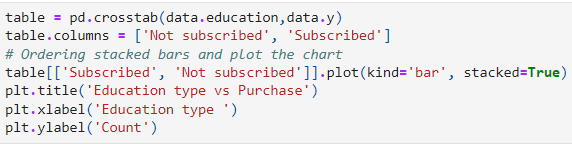
****

Result:

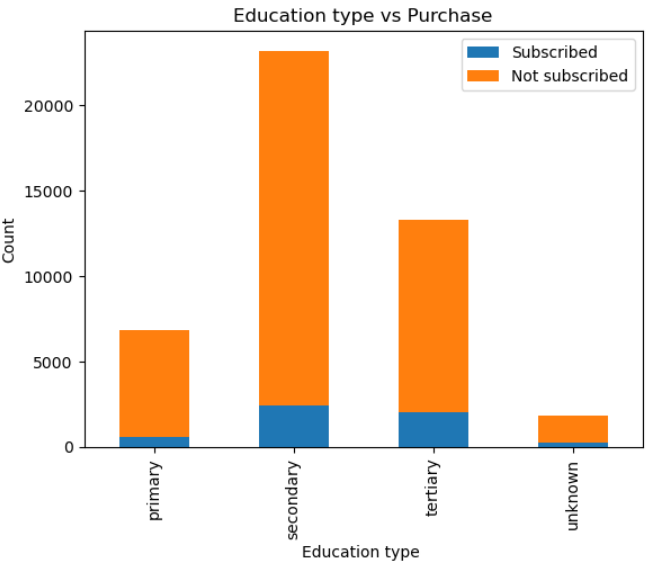


From the above output obtained it can be observed that customers with management job type are subscribed more for the term deposit followed by technicians and blue-collar job holders and entrepreneurs being the least after unknown job holders

**Education type Vs Purchase:**

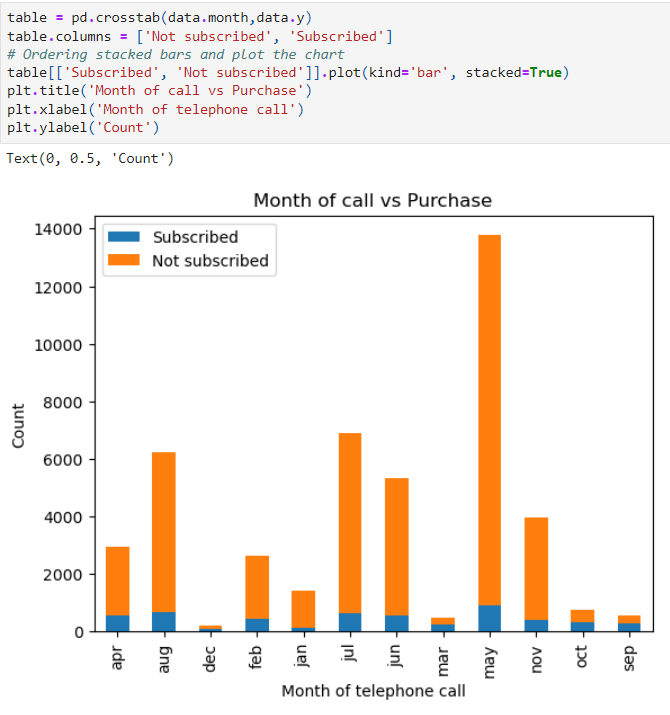
****

**Result:**

****

From the above output it can be observed that customers with secondary education received more calls and the subscription is also more followed by tertiary and unknown education being the least for the subscription.

**Month Vs Purchase:**

****

From the above result it is observed that maximum number of calls made in the months on may august and July and the maximum subscriptions received are more in month of may and followed by august and July.

The above results are the counts of subscriptions but when compared to the calls made and the subscriptions received the results might vary.

**Predictive models:**

For this part of the assignment, I have chosen the logistic regression model and the random forest regression. A Logistic regression model is a type of statistical model that is frequently used for categorization and predictive analytics. Based on a given dataset of independent variables, logistic regression calculates the likelihood that a particular event will occur, such as voting or not voting. Random Forest is an ensemble technique capable of handling both regression and classification tasks with the aid of numerous decision trees and a process known as Bootstrap and Aggregation, also referred to as bagging, The fundamental idea behind this is to mix numerous decision trees to determine the final output rather than depending just on one decision tree.

To perform the regression models, dataset is prepared in the following steps:

1. Even the duration column highly impacts the output. Yet, the duration is unknown prior to making a call. Additionally, the call's outcome is obviously apparent when it ends. Thus, the goal is to create a prediction model, duration is eliminated from the dataset.



1. A new variable ‘target’ is created to convert categorical column y to numerical and y column is removed from the dataset





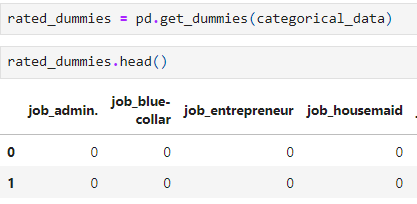
1. New sets of numerical data and categorical data is obtained to further modify the data



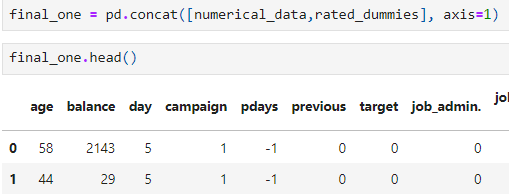




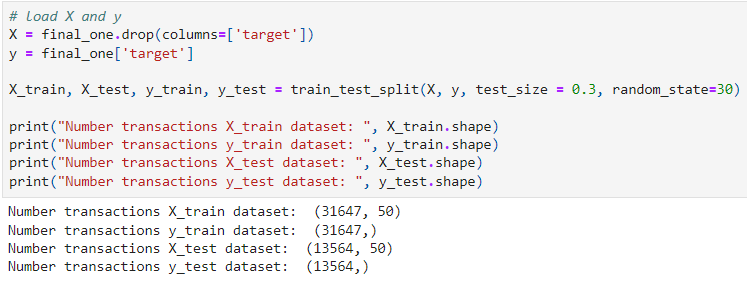
1. Created dummy variables for the all the categorical variables



1. Using the concat() function combined the numerical data and categorical data with dummy variables to get the final dataset.



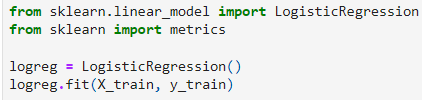
1. The dataset is split into a training dataset and a testing dataset. For this, I have used train\_test\_split() function.

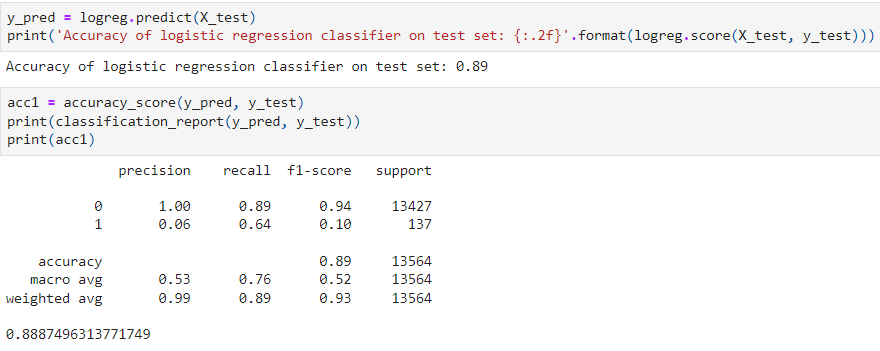


To perform the regression model we have to split the dataset into two sets. One is the training set in which we perform the regression model and check the output and the other is the test set used to check the performance of the model. In our case, I have split the data in 70:30 size where training is made on the 70% of the data and testing is made on 30% of the data.

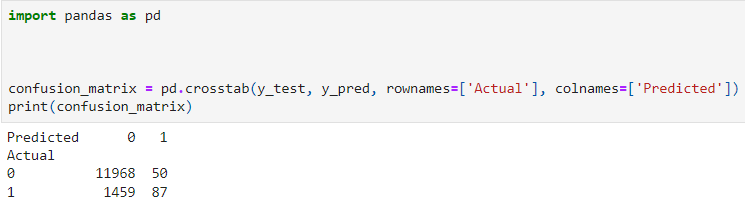
**Logistic regression model:**

I have imported the logistic regression library from sklearn.linear\_model and metrics from sklearn to perform logistic regression



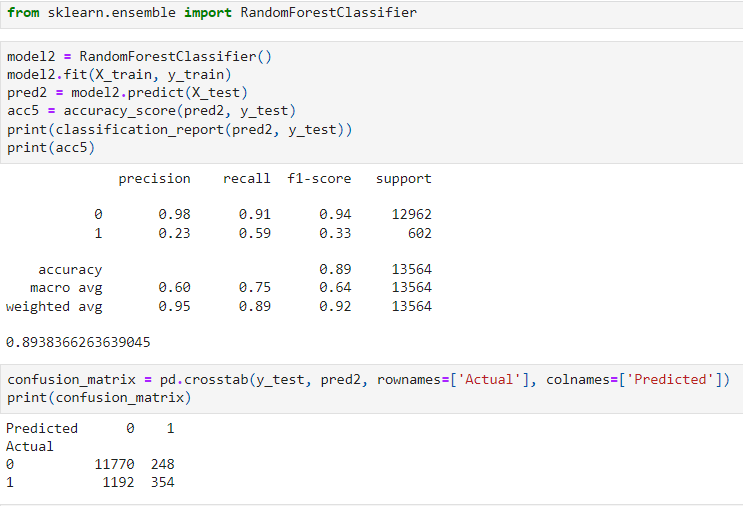


Created confusion matrix to make the final conclusion

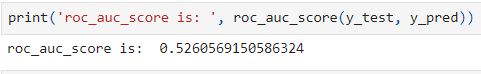


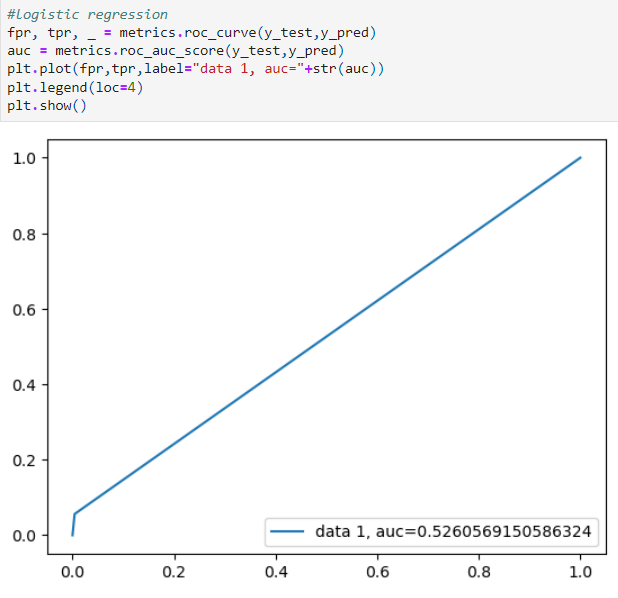
**Random forest:**

To perform the random forest classification I have used RandomforestClassifier library from sklearn.ensemble.

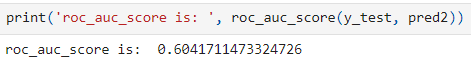


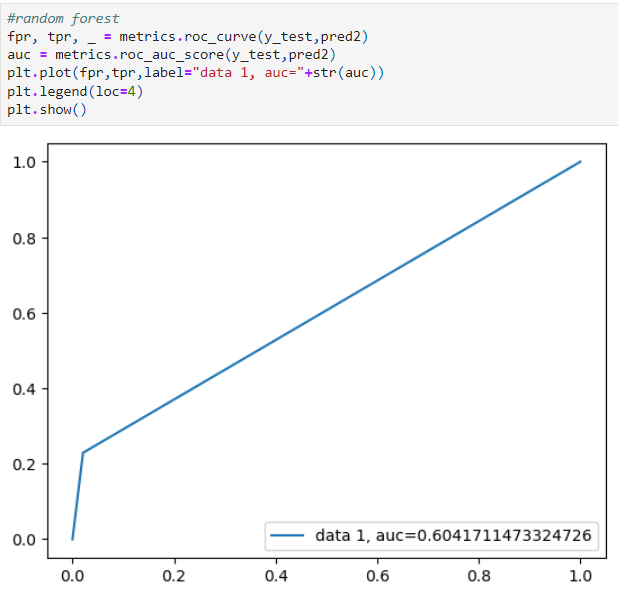
ROC and AUC Score for Logistic regression and Plot:





ROC and AUC Score for Random forest and plot:





**Conclusion:**

Classification models are used to draw some conclusions for the given input data. Accuracy and ROC curves are mostly used when the data in the dataset is balanced. But, when we have imbalanced data precision and recall could be the better measures to consider.

For the logistic regression performed above we have got an accuracy of 89% but it is better to consider the recall measure which is 64% percent. For this bank dataset, the subscription percentage was 11.70. Using the logistic regression model the percentage can be increased up to 64%. Hence it is a good model to consider.

For the random forest regression performed above we have got an accuracy of 89% but it is better to consider the recall measure which is 59% percent. For this bank dataset, the subscription percentage was 11.70. Using the random forest classification model the percentage can be increased up to 59% which makes it a decent model to consider.

When comparing both the Logistic regression model and the random forest model. Logistic regression is a better option in this case. The recall measure in logistic regression(64%) is more than the random forest(59%). Hence in my opinion logistic regression is better than the random forest.

**References:**

1. UC Irvine Machine Learning Repository. (2012). Uci.edu. <https://archive-beta.ics.uci.edu/dataset/222/bank+marketing>
2. Dutta, A. (2019, June 14). Random Forest Regression in Python - GeeksforGeeks. GeeksforGeeks. <https://www.geeksforgeeks.org/random-forest-regression-in-python/>
3. What is Logistic regression? | IBM. (n.d.). Www.ibm.com. Retrieved February 13, 2023, from <https://www.ibm.com/topics/logistic-regression#:~:text=Resources->
4. Scikit-learn. (2018). 3.2.4.3.1. sklearn.ensemble.RandomForestClassifier — scikit-learn 0.20.3 documentation. Scikit-Learn.org. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>
5. sklearn.linear\_model.LogisticRegression — scikit-learn 0.21.2 documentation. (2014). Scikit-Learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.LogisticRegression.html