**Santander Customer Transaction Prediction**

**Pravin Chaudhary**

**18-11-2019**

1. **Introduction**
   1. **Problem Statement**
   2. **Data**
2. **Methodology**
   1. **Pre – Processing**
   2. **Outlier detection & removal**
   3. **Correlation**
   4. **Data Distribution**
   5. **Feature Scaling**
   6. **Feature selection**
3. **Modelling**
   1. **Model Selection**
   2. **Model Evaluation metrices**
   3. **Logistic Regression**
   4. **Random Forest**
   5. **SMOTE**
   6. **LightGBM**
4. **Conclusion** 
   1. **Model Evaluation – Confusion Metrix, AUC Score**
   2. **Model Selection**

**Chapter 1. Introduction**

**1.1 Problem Statement :**

At ​Santander​, mission is to help people and businesses prosper. We are always looking for ways to help our customers understand their financial health and identify which products and services might help them achieve their monetary goals. Our data science team is continually challenging our machine learning algorithms, working with the global data science community to make sure we can more accurately identify new ways to solve our most common challenge, binary classification problems such as: is a customer satisfied? Will a customer buy this product? Can a customer pay this loan?

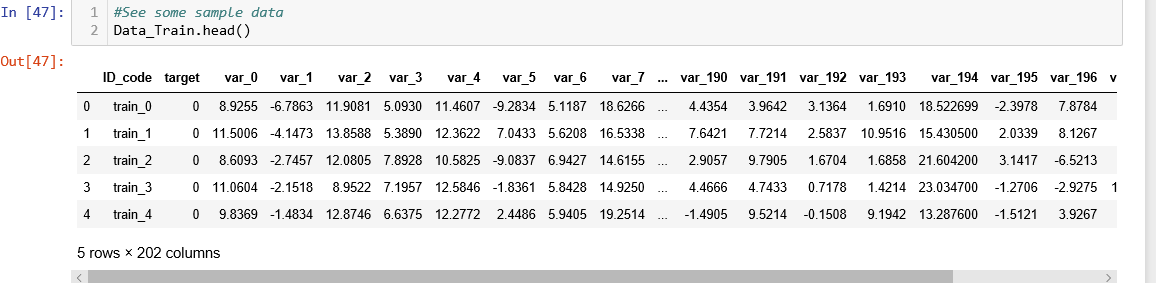
In this challenge, we need to identify which customers will make a specific transaction in the future, irrespective of the amount of money transacted.

**1.2 Data :**

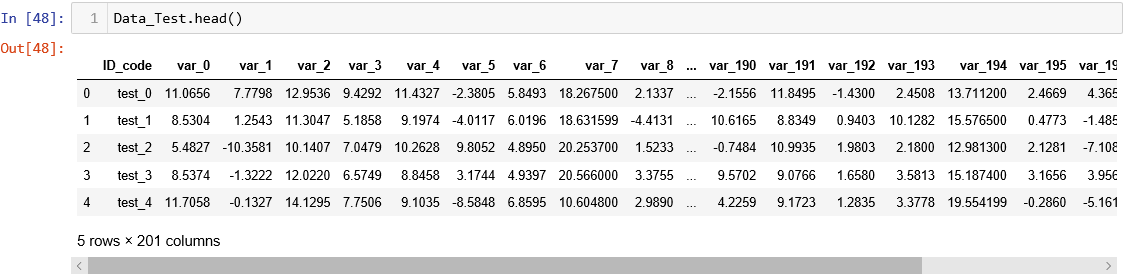
Our Aim is to build a classification model which will predict which customer will make a transaction in future.

Below is the sample of the data :

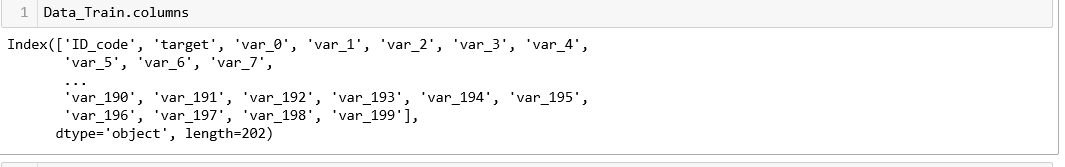
**Train Data Set :**



**Test Data Set :**



**Features :**



So we are having around 201 Variables based on which we need to predict if a customer will make a transaction or not.

**Chapter 2: Methodology**

**2.1 Pre – Processing**

Before we feed the data to the model. We need to do some pre-processing on the data. It includes cleansing the data, finding the relationship of input features against the output label and also finding the relationship of the input features with each other

Also visualizing the data to find meaningful features and their distrubutions.

**2.2 Outlier detection and removal**



From above what we understand is that our target class is imbalance.

That is the customers making the transaction are way less than the customers who are not making a transaction.

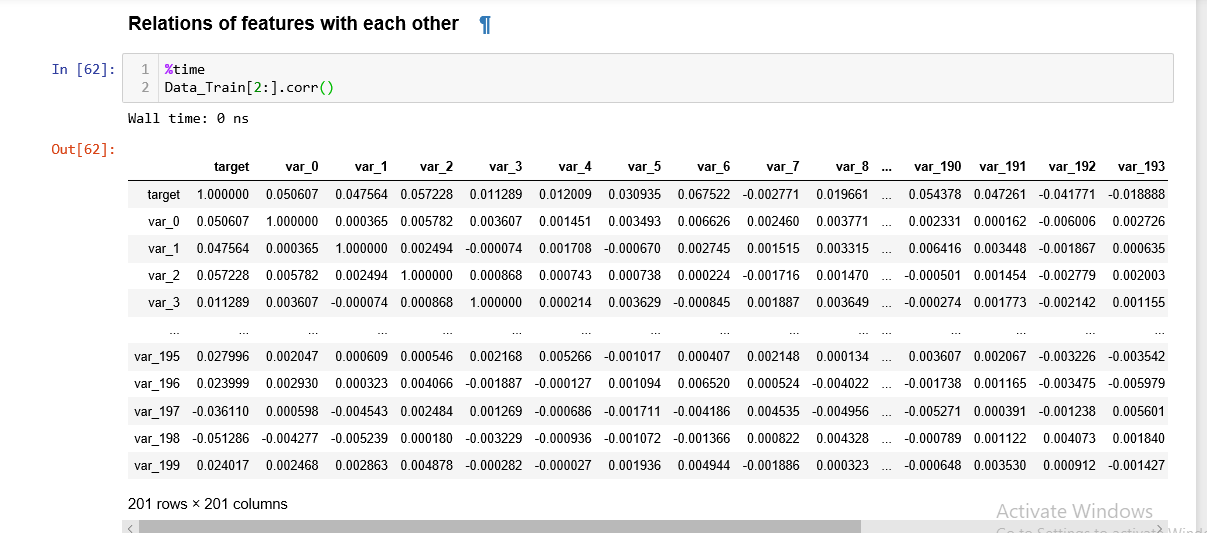
So we will not do any Outlier removal. The reason is we might lose some useful information of our minority class

**2.3 Correlation :**

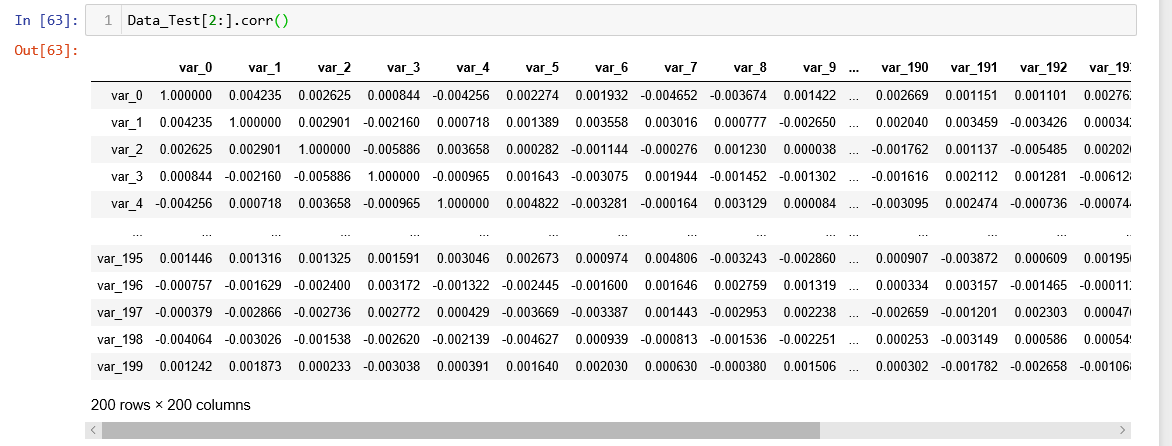
This method is used to find the how strong or weak is the relationship between the variable is. Which inturn help us to eliminate variables which are not required and only select those which are useful.

**Correlation of input features with each other :**

**Train Data :**



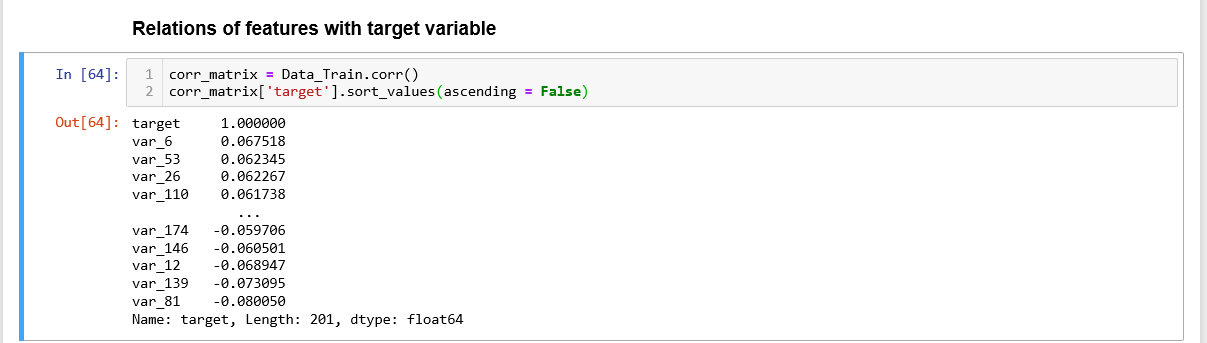
**Test Data :**



If we see the above :

1. The variables are very less correlated to each other means they are independent of each other.

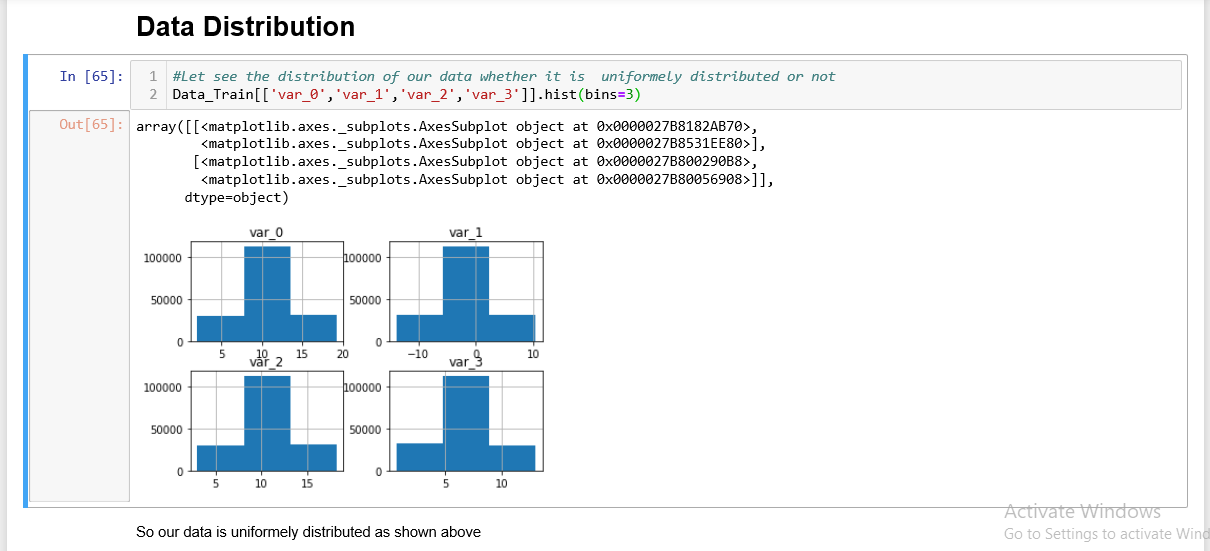
**Correlation of input features with target variable :**



If we see above what we find that features are not strongly correlated with the target variable.

**2.4 Data Distribution :**

Lets see whether our data is uniformly distributed or not.



So from above fig, we can see that our data is uniformly distributed.

Hence we will do the standardization of our data in feature scaling.

* 1. **Feature Scaling :**

The purpose of Feature Scaling is to standardize the features of our dataset into a specific range. So that when we feed this data to our model it performs well.

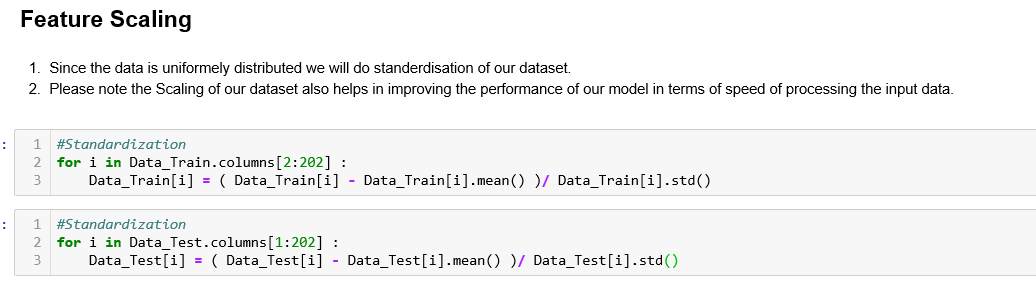
Now there are two ways for Feature Scaling :

1. Normalisation
2. Standardization

In case if our data is not normally distributed then we go with Normalization

But here since it is normally distributed, We will do the standardization of our data.

**Logic for Standardization :**



**4.4 Feature Selection :**

Feature selection plays an important role in building our model.There might be lot of features in our dataset but not all features are important. We need to only select those features which plays an important role in predicting our output variable.

There are different methods to find the correlation of input features with the output variable. One of them is pearson correlation.

It is really difficult to select the some specific features out of all bcz :

1. The Names of the features are not meaningful which inturn does not helps us in understanding their meaning.

2. Also the correlation does not show if a variable is strongly related with the output variable.

**Chapter 3. Modelling**

**3.1 Model Selection:**

Our output variable is having binary values that is 1’s and 0’s. So We will go with the Classification model to predict our outcome.

**3.2 Model Evaluation metrices :**

The Model Evaluation is an importand aspect of data science project because based on this we decide whether to select a model or not.

We have choose two metrices over here :

1. Confusion metrix
2. AUC Score, precision, recall, f1-score

First we will use a simple classification model then we will use more complex ones.

**3.3 Logistic Regression :**

Logistic regression is a classification algorithm used to assign observation to a discrete set of classes.



**3.4 Random Forest :**

Random Forest consist of one or more decision trees.

There are two types :

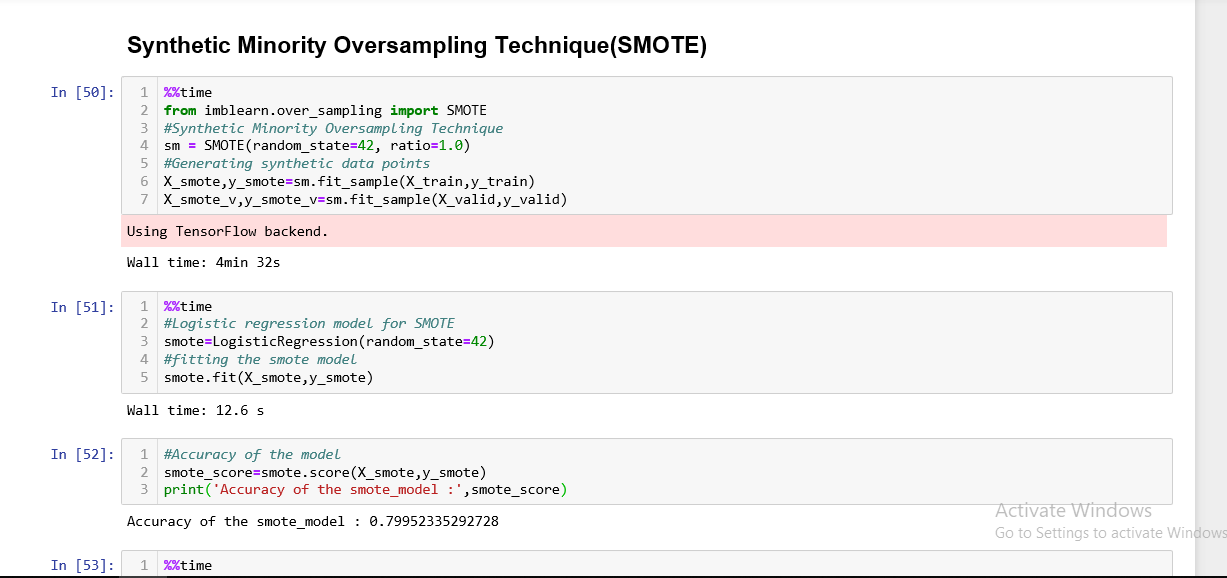
1. Random Forest regression
2. Random Forest Classifier

Now here since our output variable is a class having binary value 0 and 1. We will choose Random Forest Classifier.



**3.5 SMOTE :**

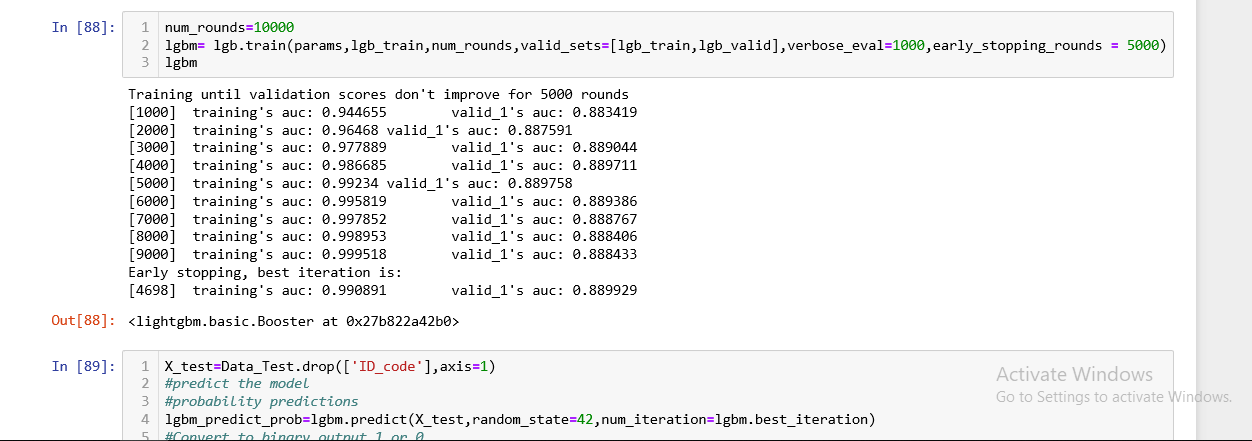
SMOTE uses a nearest neighbor’s algorithm to generate new and synthetic data to use for training the model. In order to balance imbalanced data we are going to use SMOTE sampling method.



**3.6 LightGBM :**

LightGBM is a gradient boosting framework that uses tree based learning algorithms.

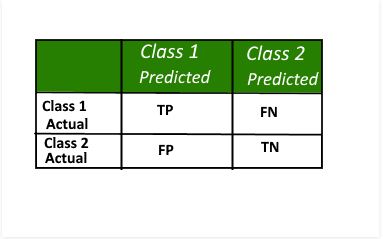


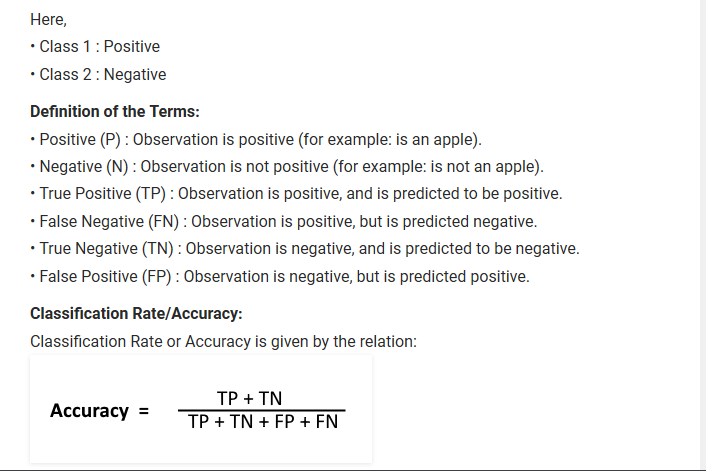


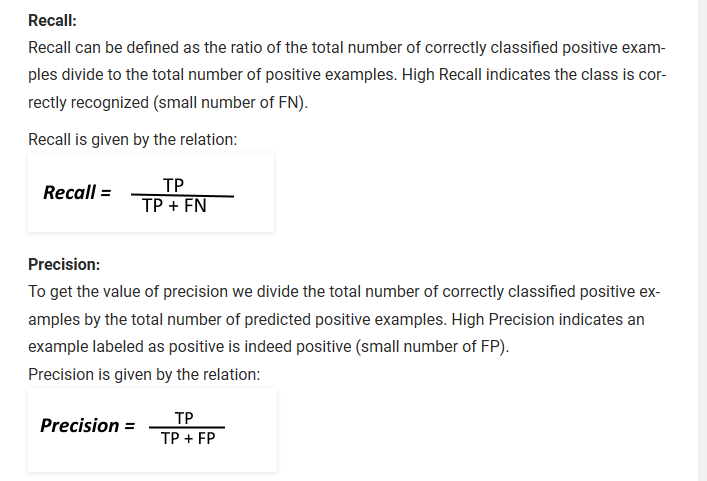
1. **Conclusion** 
   1. **Model Evaluation :**

Now we have used 4 models. So we need to select the best model out of the 4 based on their some accuracy parameters like confusion matrix, auc score, recall, precision.

Confusion Matrix: - It is a technique for summarizing the performance of a classification algorithm. The number of correct predictions and incorrect predictions are summarized with count values with respect to each class.







**Area under curve(AUC) Score :**

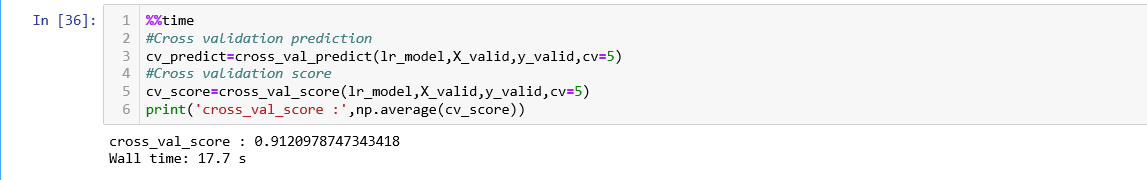
auc\_score :- It is a metric that computes the area under the Roc curve and also used metric for imbalanced data.

Roc curve is plotted true positive rate or Recall on y axis against false positive rate or specificity on x axis. The larger the area under the roc curve better the performance of the model.

**Logistic Regression :**

**Confusion matrix :**

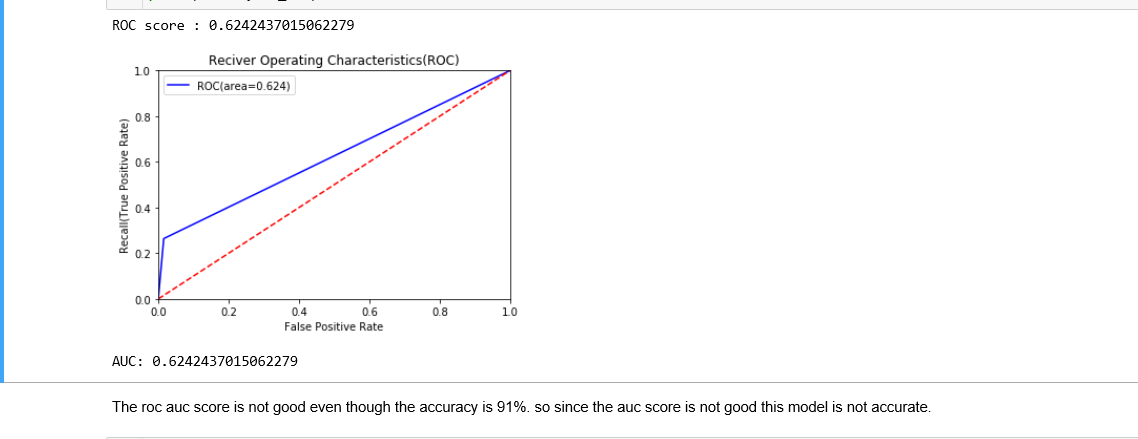
As per the confusion matrix, we got the accuracy of 0.91.

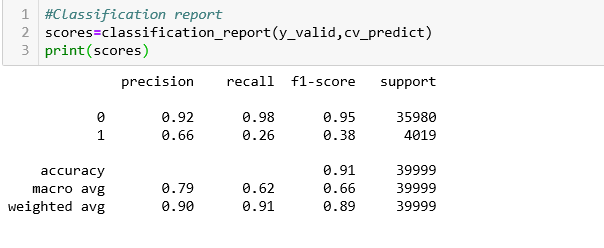


Accuracy of the model is not the best metric to use when evaluating the imbalanced datasets as it may be misleading. So, we are going to change the performance metric.

Hence we will go with the AUC score.

**AUC Score :**





If you see above,

The precision and recall is really good for class 0 whereas these are not good for class 1.

Note :

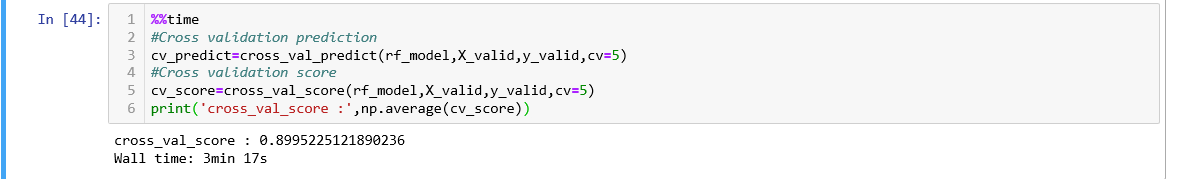
class 0 - customer will not make a transaction

class 1 - customer will make a transaction

So since the f1-score(harmonic mean of precision and recall) for customer who will be making a transaction is 0.38 which is not good and also the auc score is 0.62 which is not a good accuracy. Hence we are rejecting this model.

**Random Forest Classifier :**

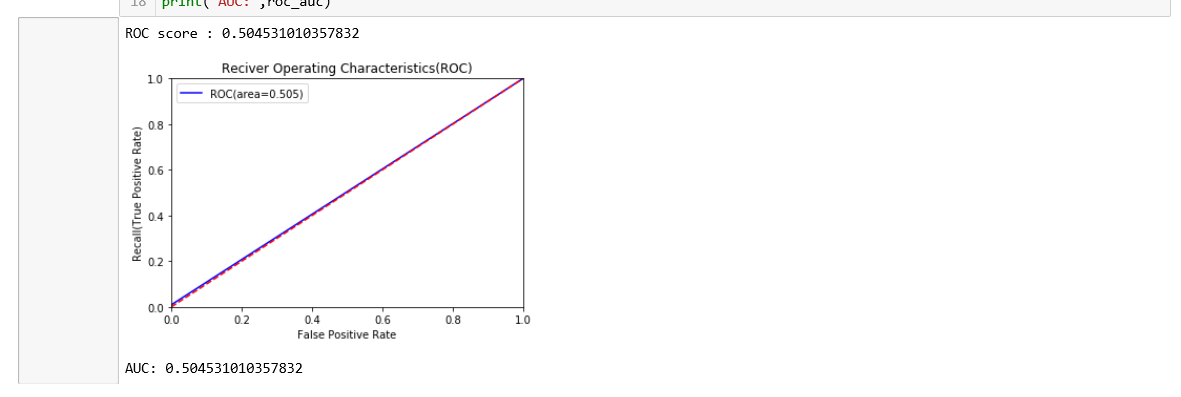
**Confusion matrix** : As per the confusion matrix, we got the accuracy of 0.89.

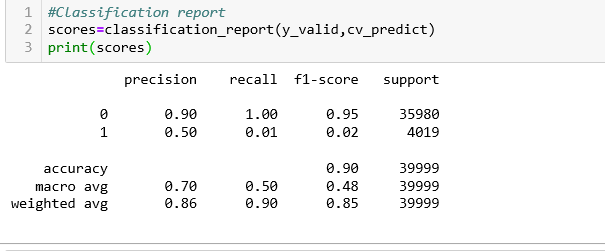


Accuracy of the model is not the best metric to use when evaluating the imbalanced datasets as it may be misleading. So, we are going to change the performance metric.

Hence we will go with the AUC score.

**AUC Score :**





If you see above,

The precision and recall is really good for class 0 whereas these are not good for class 1.

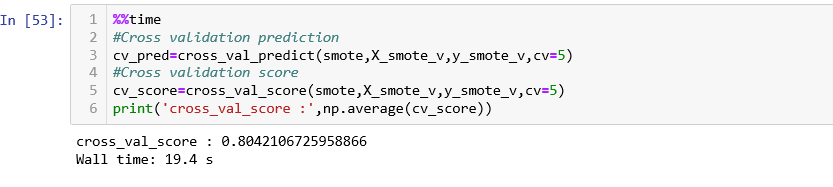
note : class 0 - customer will not make a transaction

class 1 - customer will make a transaction

So since the f1-score(harmonic mean of precision and recall) for customer who will be making a transaction is 0.02 which is not good and also auc score is 0.50 which is not a good accuracy. Hence we are rejecting this model.

**Synthetic Minority Oversampling Technique(SMOTE) :**

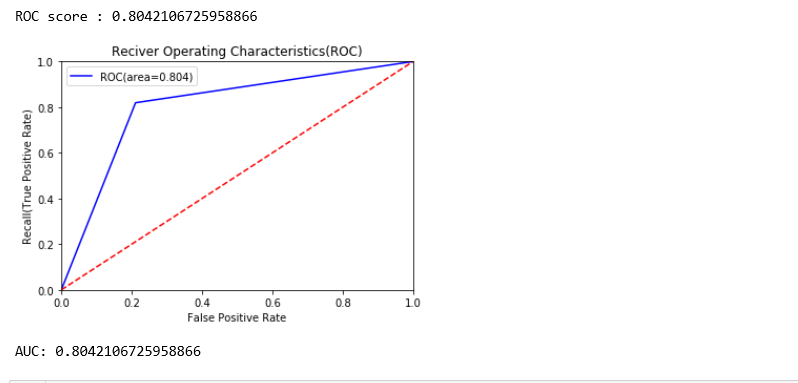
**Confusion matrix :** As per the confusion matrix, we got the accuracy of 0.80.

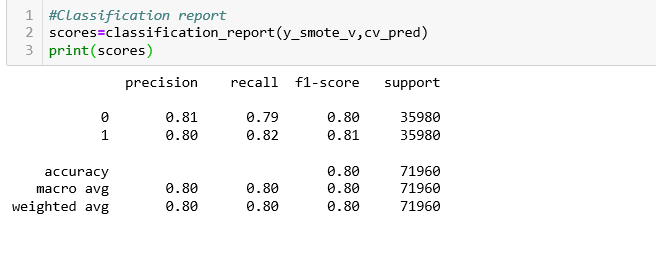


Accuracy of the model is not the best metric to use when evaluating the imbalanced datasets as it may be misleading. So, we are going to change the performance metric.

Hence we will go with the AUC score.

**AUC Score :**





If you see above,

The precision and recall is really good for both class 0 and class 1 which is a positive outcome as compared to the previous model Logistic Regression and Random forest

note : class 0 - customer will not make a transaction

class 1 - customer will make a transaction

So since the f1-score(harmonic mean of precision and recall) for customer who will be making a transaction is 0.81 which is good and also auc score is 0.80 which is a good accuracy.

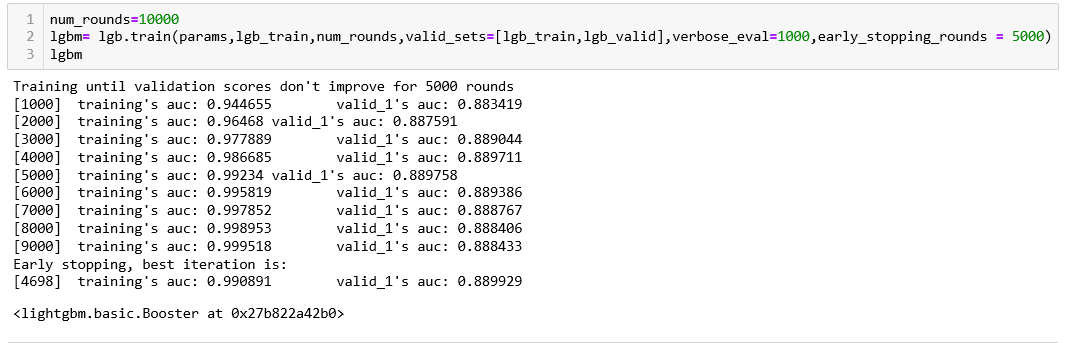
Still we will see other model to see if that has good accuracy that this or not

**LightGBM :**

**Confusion matrix :** As per the confusion matrix, we got the accuracy of 0.87.



**AUC Score :** auc score of LightGBM is 0.88.



* 1. **Model Selection :**

Now, Finally lets select the based model out of 4 models based on their AUC score.

We got below conclusion :

1. Logistic regression and Random Forest Classifier did not work well on our imbalanced dataset.
2. SMOTE and LightGBM worked well on imbalance dataset.
3. Now if we compare SMOTE and lightGBM, AUC score of ligthGBM is 0.89 whereas AUC score of SMOTE is 0.80
4. So lightGBM is having good accuracy than the SMOTE.

Hence we choose lightGBM Model as our final model for the Project to predict the whether a customer will make a transaction or not.