# 

# Abstract

All business including the business of banking requires top line growth in terms of volumes of business to increase the bottom line of profit growth. Loan Default is the possibility of a loss resulting from a borrower's failure to repay a loan or meet contractual obligations. Loan Default also describes the risk that a bond issuer may fail to make payment when requested or that an insurance company will be unable to pay a claim. Thus, Loan default is always the threat to any financial institution and should be predicted in advance based on various features of the applicant.

This study aims at applying machine learning models, including Decision tree, Logistic regression, Random forest and Ada Boost classifier to classify applicants with and without loan default from a group of predicting variables, and evaluate their performance, based on the loan issued by XYZ Co-operation through 2007-2015

# CHAPTER 1: INTRODUCTION

## Loan Default Analysis

The objective of Loan Default Analysis project is to put ourselves in the shoes of a loan issuer and manage credit risk by using the past data and deciding whom to give the loan to in the future. Model has to analyzing the XYZ Co-operation Data to analyze and detect defaulters by building models on data from June 2007 to May 2015 and testing it on data from June 2015 to December 2015. Based on the accuracy of each model we can determine how good a model is for predicting that a person applying for loan will default or not.

## Need of the Study

In Loan Default Analysis, The purpose of the project is to predict whether a borrower will default or not, so that investors can avoid those borrowers using manual investing feature provided by lending club. This, however, does not necessarily lead to highest return on investment because by completely avoiding potential defaults, one also avoid riskier loans that may lead to higher return on investment even though they default at some point in the future. In order to maximize return on investment, one needs to optimize return on investment instead. In this project, we work on the simpler problem that is to predict loan defaults.

0 represents – Not Defaulter

1 represents – Defaulter

## Business or Enterprise under study

**Name of the Organisation:** XYZ Co-operation Bank

XYZ Corporation Lending Data is used under the study. Data of Loans issued by XYZ Co-operation through the year 2007-2015 is used for analysis. The data contains the indicator of default, payment information, credit history and many more variables.

**Database Description:**

**Rows:** 855696**Columns:** 73

|  |  |  |
| --- | --- | --- |
| **No** | **LoanStatNew** | **Description** |
| 1 | Id | A unique assigned ID for the loan listing. |
| 2 | member\_id | A unique Id for the borrower member. |
| 3 | loan\_amnt | The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value. |
| 4 | funded\_amnt | The total amount committed to that loan at that point in time. |
| 5 | funded\_amnt\_inv | The total amount committed by investors for that loan at that point in time. |
| 6 | Term | The number of payments on the loan. Values are in months and can be either 36 or 60. |
| 7 | int\_rate | Interest Rate on the loan |
| 8 | Instalment | The monthly payment owed by the borrower if the loan originates. |
| 9 | Grade | XYZ corp. assigned loan grade |
| 10 | sub\_grade | XYZ assigned assigned loan subgrade |
| 11 | emp\_title | The job title supplied by the Borrower when applying for the loan. |
| 12 | emp\_length | Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years. |
| 13 | home\_ownership | The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER. |
| 14 | annual\_inc | The self-reported annual income provided by the borrower during registration. |
| 15 | verification\_status | Was the income source verified |
| 16 | issue\_d | The month which the loan was funded |
| 17 | pymnt\_plan | Indicates if a payment plan has been put in place for the loan |
| 18 | Desc | Loan description provided by the borrower |
| 19 | Purpose | A category provided by the borrower for the loan request. |
| 20 | Title | The loan title provided by the borrower |
| 21 | zip\_code | The first 3 numbers of the zip code provided by the borrower in the loan application. |
| 22 | addr\_state | The state provided by the borrower in the loan application |
| 23 | Dti | A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested loan, divided by the borrower’s self-reported monthly income. |
| 24 | delinq\_2yrs | The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years |
| 25 | earliest\_cr\_line | The month the borrower's earliest reported credit line was opened |
| 26 | inq\_last\_6mths | The number of inquiries in past 6 months (excluding auto and mortgage inquiries) |
| 27 | mths\_since\_last\_delinq | The number of months since the borrower's last delinquency. |
| 28 | mths\_since\_last\_record | The number of months since the last public record. |
| 29 | open\_acc | The number of open credit lines in the borrower's credit file. |
| 30 | pub\_rec | Number of derogatory public records |
| 31 | revol\_bal | Total credit revolving balance |
| 32 | revol\_util | Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit. |
| 33 | total\_acc | The total number of credit lines currently in the borrower's credit file |
| 34 | initial\_list\_status | The initial listing status of the loan. Possible values are – W, F |
| 35 | out\_prncp | Remaining outstanding principal for total amount funded |
| 36 | out\_prncp\_inv | Remaining outstanding principal for portion of total amount funded by investors |
| 37 | total\_pymnt | Payments received to date for total amount funded |
| 38 | total\_pymnt\_inv | Payments received to date for portion of total amount funded by investors |
| 39 | total\_rec\_prncp | Principal received to date |
| 40 | total\_rec\_int | Interest received to date |
| 41 | total\_rec\_late\_fee | Late fees received to date |
| 42 | Recoveries | post charge off gross recovery |
| 43 | collection\_recovery\_fee | post charge off collection fee |
| 44 | last\_pymnt\_d | Last month payment was received |
| 45 | last\_pymnt\_amnt | Last total payment amount received |
| 46 | next\_pymnt\_d | Next scheduled payment date |
| 47 | last\_credit\_pull\_d | The most recent month XYZ corp. pulled credit for this loan |
| 48 | collections\_12\_mths\_ex\_med | Number of collections in 12 months excluding medical collections |
| 49 | mths\_since\_last\_major\_derog | Months since most recent 90-day or worse rating |
| 50 | policy\_code | publicly available policy\_code=1 new products not publicly available policy\_code=2 |
| 51 | application\_type | Indicates whether the loan is an individual application or a joint application with two co-borrowers |
| 52 | annual\_inc\_joint | The combined self-reported annual income provided by the co-borrowers during registration |
| 53 | dti\_joint | A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested loan, divided by the co-borrowers' combined self-reported monthly income |
| 54 | verified\_status\_joint | Indicates if the co-borrowers' joint income was verified by XYZ corp., not verified, or if the income source was verified |
| 55 | acc\_now\_delinq | The number of accounts on which the borrower is now delinquent. |
| 56 | tot\_coll\_amt | Total collection amounts ever owed |
| 57 | tot\_cur\_bal | Total current balance of all accounts |
| 58 | open\_acc\_6m | Number of open trades in last 6 months |
| 59 | open\_il\_6m | Number of currently active installment trades |
| 60 | open\_il\_12m | Number of installment accounts opened in past 12 months |
| 61 | open\_il\_24m | Number of installment accounts opened in past 24 months |
| 62 | mths\_since\_rcnt\_il | Months since most recent installment accounts opened |
| 63 | total\_bal\_il | Total current balance of all installment accounts |
| 64 | il\_util | Ratio of total current balance to high credit/credit limit on all install acct |
| 65 | open\_rv\_12m | Number of revolving trades opened in past 12 months |
| 66 | open\_rv\_24m | Number of revolving trades opened in past 24 months |
| 67 | max\_bal\_bc | Maximum current balance owed on all revolving accounts |
| 68 | all\_util | Balance to credit limit on all trades |
| 69 | total\_rev\_hi\_lim | Total revolving high credit/credit limit |
| 70 | inq\_fi | Number of personal finance inquiries |
| 71 | total\_cu\_tl | Number of finance trades |
| 72 | inq\_last\_12m | Number of credit inquiries in past 12 months |
| 73 | Default\_ind | Current status of the loan |

## Techniques Used

**LOGISTIC REGRESSION**

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

**DECISION TREE**

Decision Tree Classifier is a simple and widely used classification technique. It applies a straight forward idea to solve the classification problem. Decision Tree Classifier poses a series of carefully crafted questions about the attributes of the test record. Each time it receive an answer, a follow-up question is asked until a conclusion about the class label of the record is reached. The decision tree classifiers organized a series of test questions and conditions in a tree structure.

**RANDOM FOREST CLASSIFIER**

Random forests or random decision forests are an [ensemble learning](https://en.wikipedia.org/wiki/Ensemble_learning)  method for classification, regression and other tasks that operate by constructing a multitude of [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning) at training time and outputting the class that is the [mode](https://en.wikipedia.org/wiki/Mode_(statistics)) of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their [training set](https://en.wikipedia.org/wiki/Test_set).

**ADAPTIVE BOOST CLASSIFIER (ADABOOST)**

Boosting is an ensemble technique that attempts to create a strong classifier from a number of weak classifiers. [AdaBoost](https://en.wikipedia.org/wiki/AdaBoost) was the first really successful boosting algorithm developed for binary classification. It is the best starting point for understanding boosting.

**CROSS VALIDATION TECHNIQUE -K-FOLD CROSS VALIDATION**

Cross-validation is a statistical method used to estimate the skill of machine learning models .It is commonly used in applied machine learning to compare and select a model for a given predictive modelling problem because it is easy to understand, easy to implement, and results in skill estimates that generally have a lower bias than other methods.

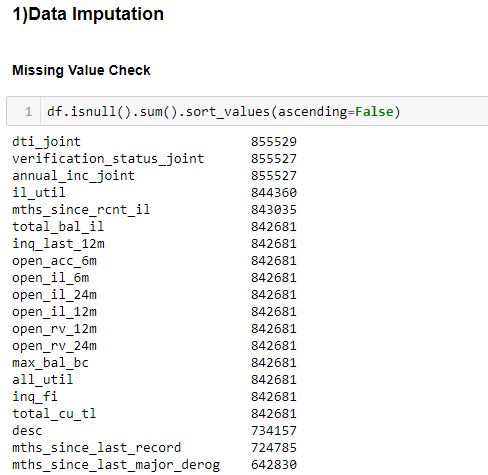
**SMOTE –** **SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE**

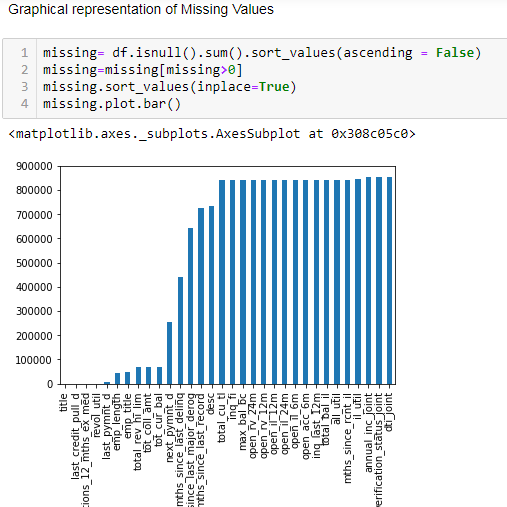
Classification using class-imbalanced data is biased in favor of the majority class. The bias is even larger for high-dimensional data, where the number of variables greatly exceeds the number of samples. The problem can be attenuated by under sampling or oversampling, which produce class-balanced data. Generally under sampling is helpful, while random oversampling is not. The Synthetic Minority Over-sampling TEchnique (SMOTE) is an oversampling approach that creates synthetic minority class samples. It potentially performs better than simple oversampling and it is widely used.

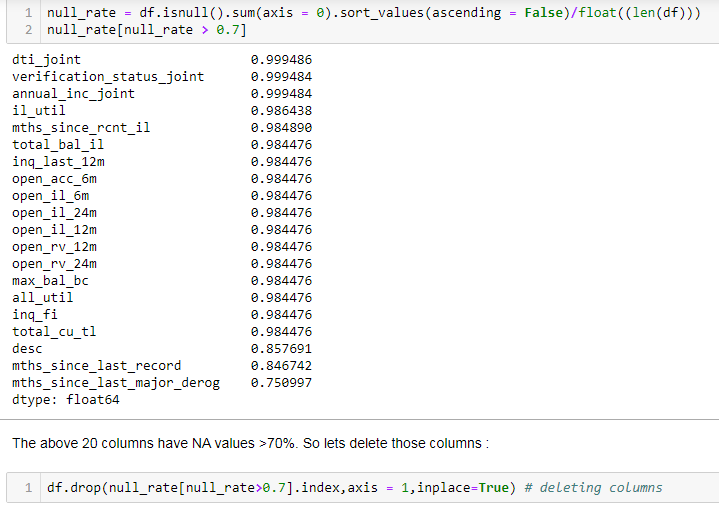
# CHAPTER 2: DATA PREPARATION AND UNDERSTANDING

One of the first steps we engaged in was to outline the sequence of steps that we will be following for our project. Each of these steps are elaborated below

1. **Missing Value Analysis and Deleting columns with NA >70% :**

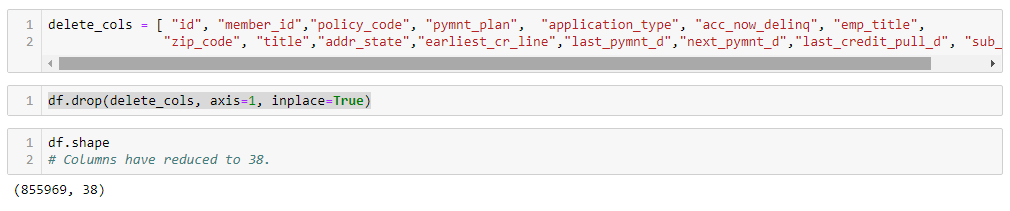
****

****

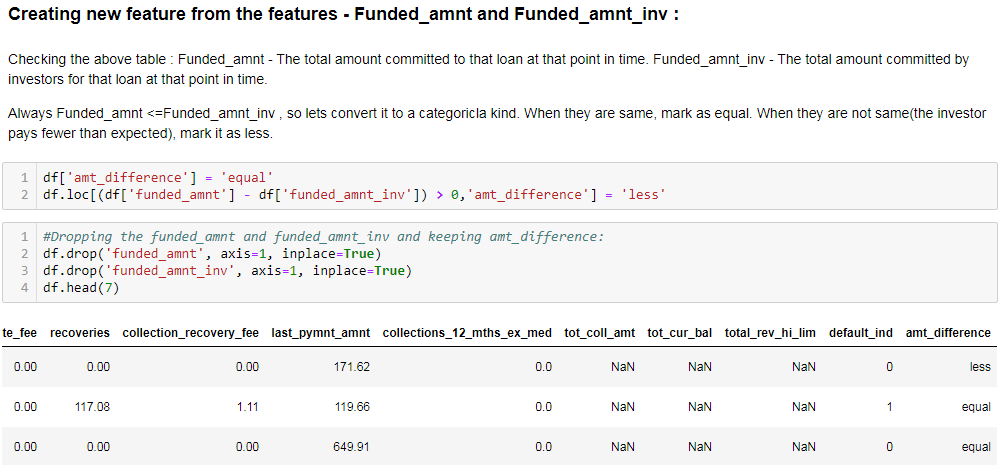
****

1. **Deleting irrelevant Features :**

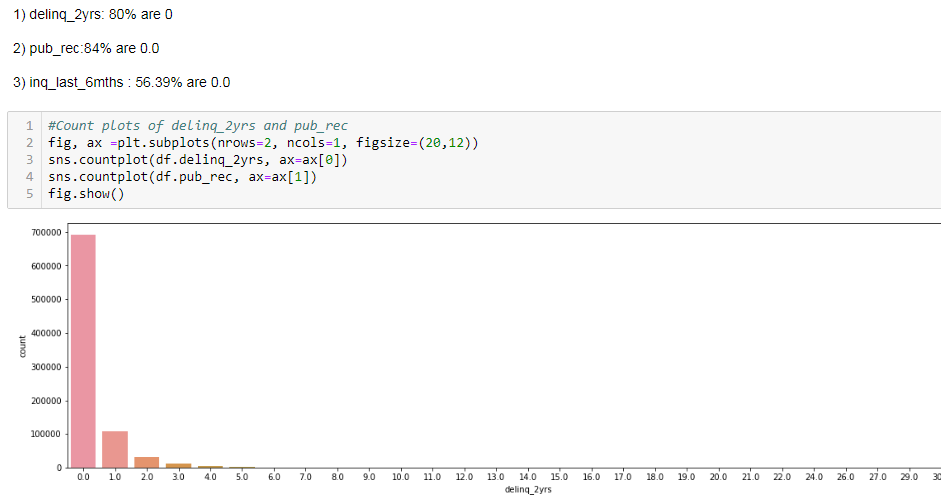
Id and member\_id are 100% unique .Policy\_code has just 1 value throughout the dataset. Payment\_plan, application\_type and acc\_now\_delinq are highly unbalanced (more than 90% are single value). Also columns - desc, emp\_title, zip\_code, title, addr\_state, earliest\_cr\_line, last\_pymnt\_d, next\_pymnt\_d, last\_credit\_pull\_d and sub\_grade are too descriptive. Considering these points the columns are deleted as these will not help in analysis

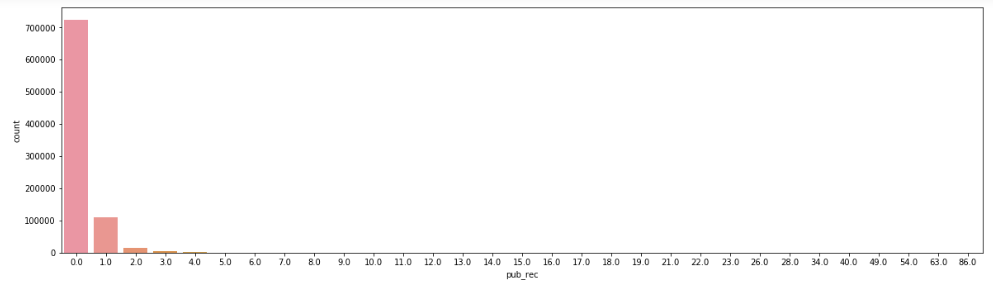
****

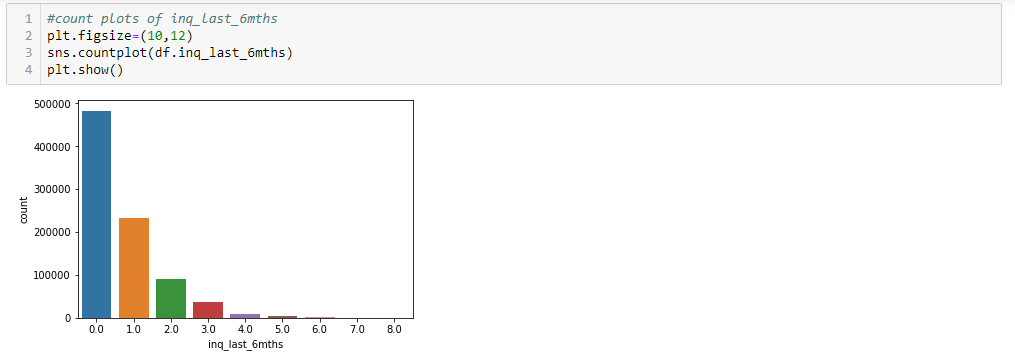
1. **Creating new Feature :**



1. **Converting highly biased columns to categorical features :**

****

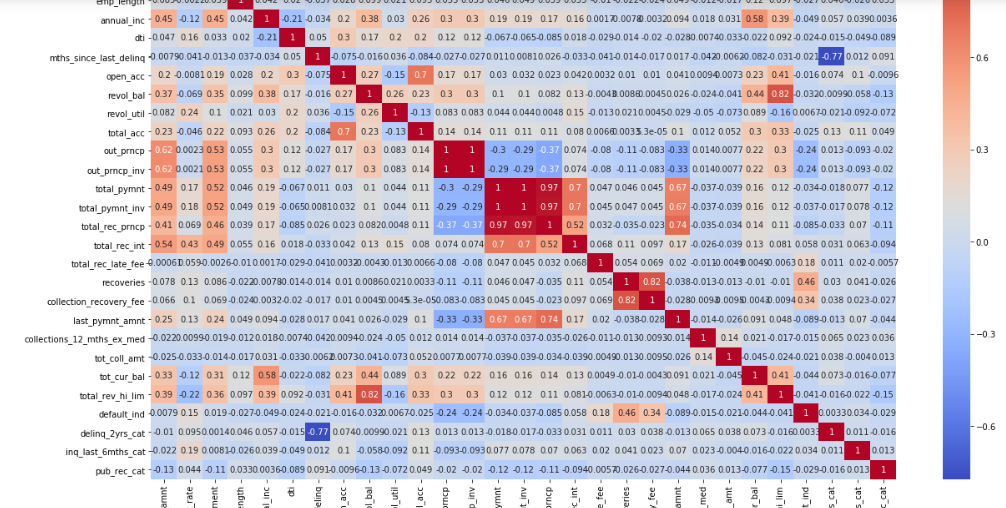
****

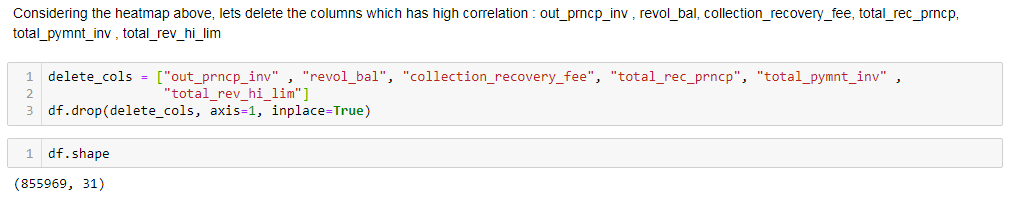
****

****

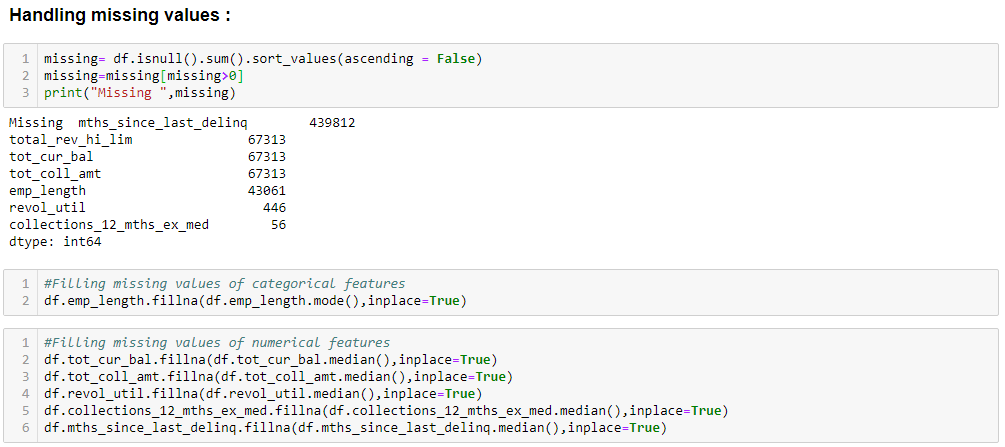
1. **Checking Multi Collinearity and deleting highly correlated:**

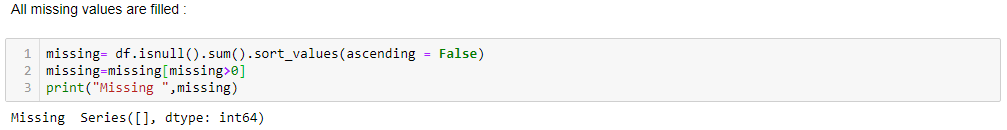




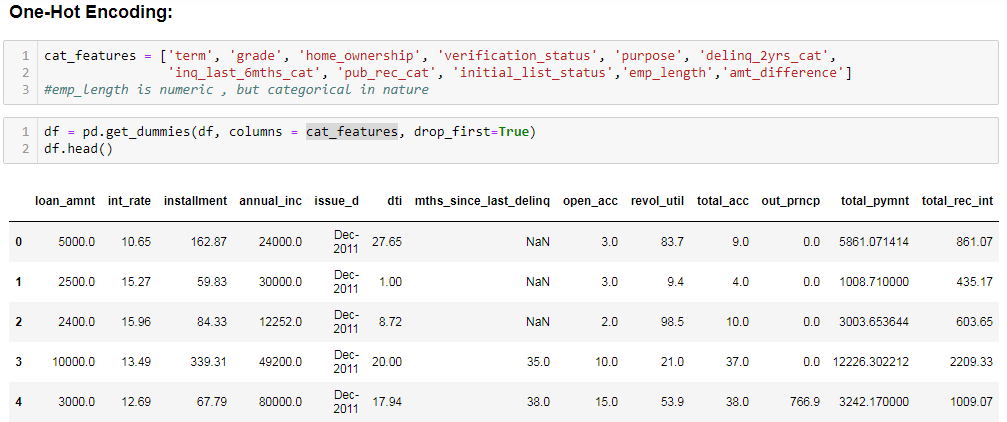
****

1. **Handling remaining missing values:**

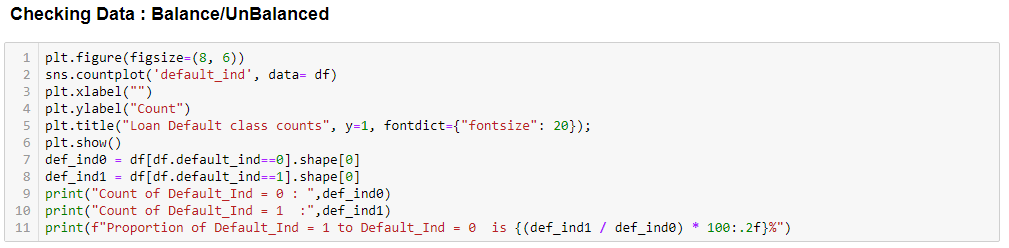


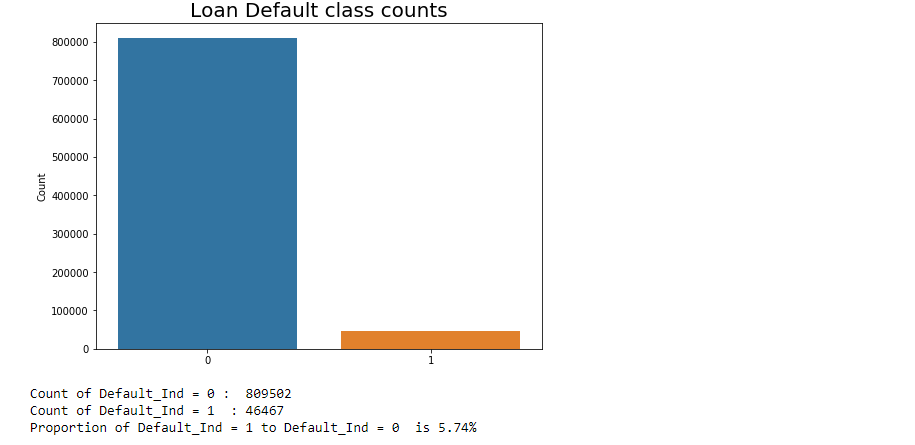
****

1. **One-Hot Encoding –Converting Categorical Variables to Numeric:**



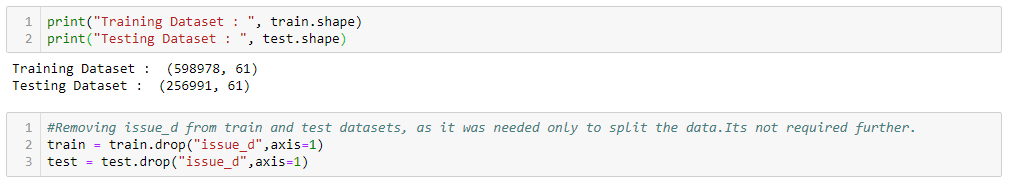
1. **Checking whether the Data is Balanced or not:**

****

****

1. **Splitting the Data into Testing and Training Data using issue\_d Column** :

****

****

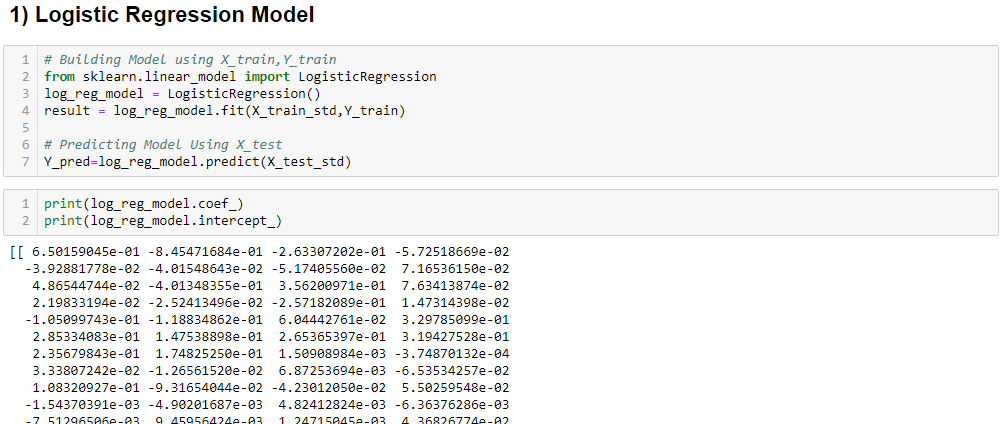
1. **Standardizing the data :**

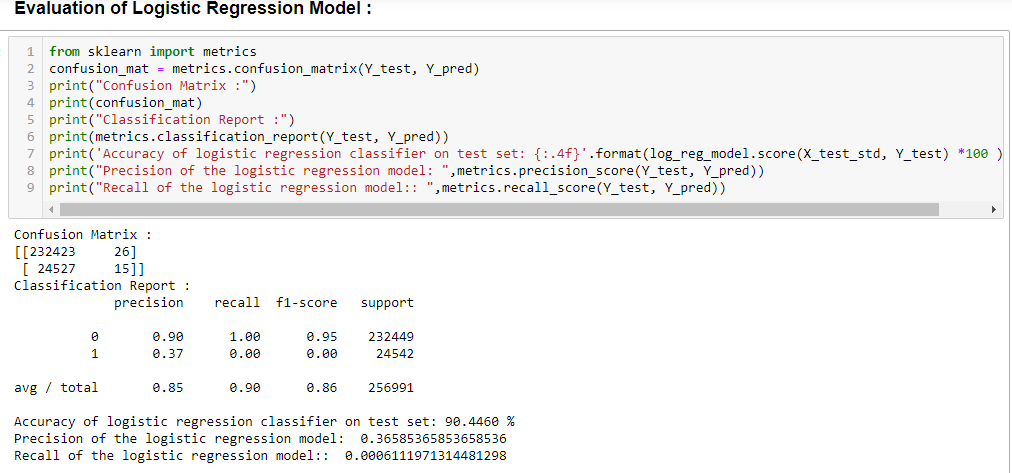
****

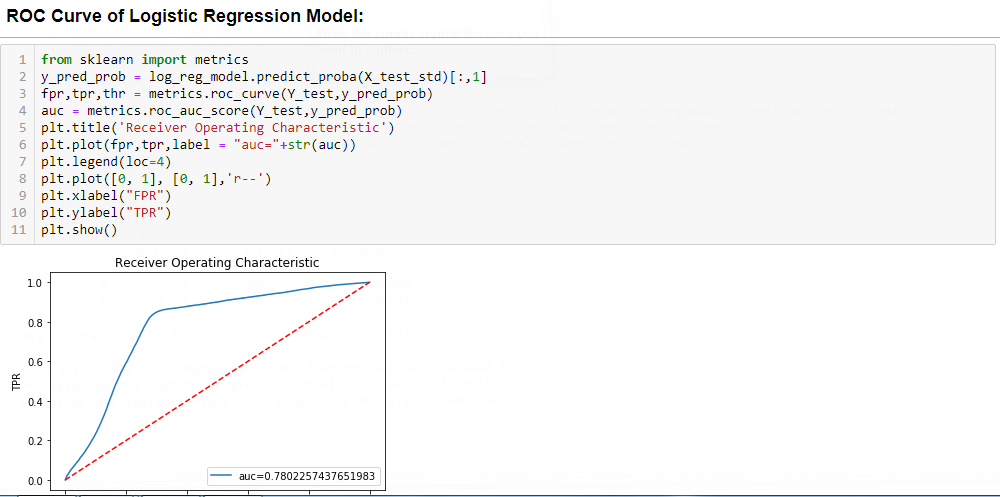
**CHAPTER 3: FITTING MODELS TO DATA**

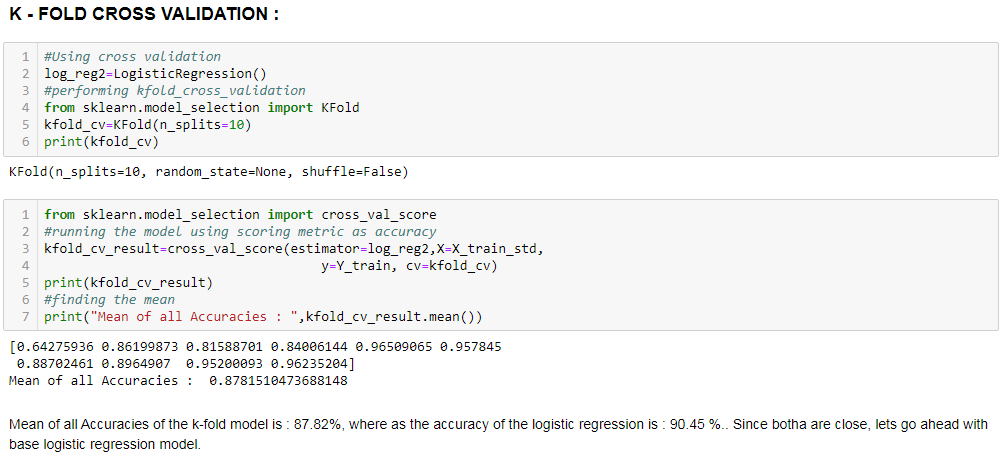
**3.1) Using UnBalanced Data:**

**3.1.1) Logistic Regression Model**

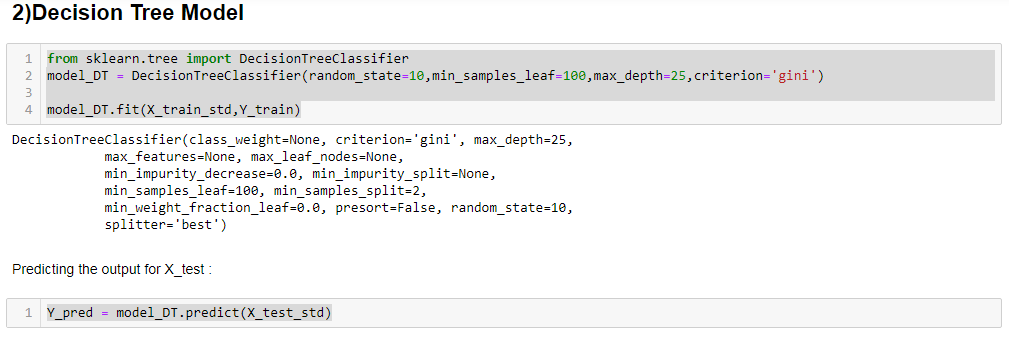
****

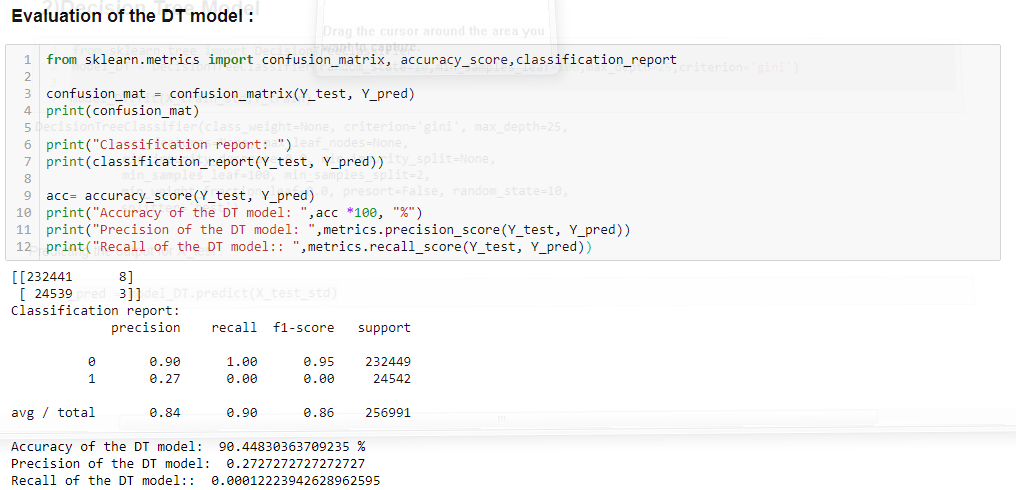
****

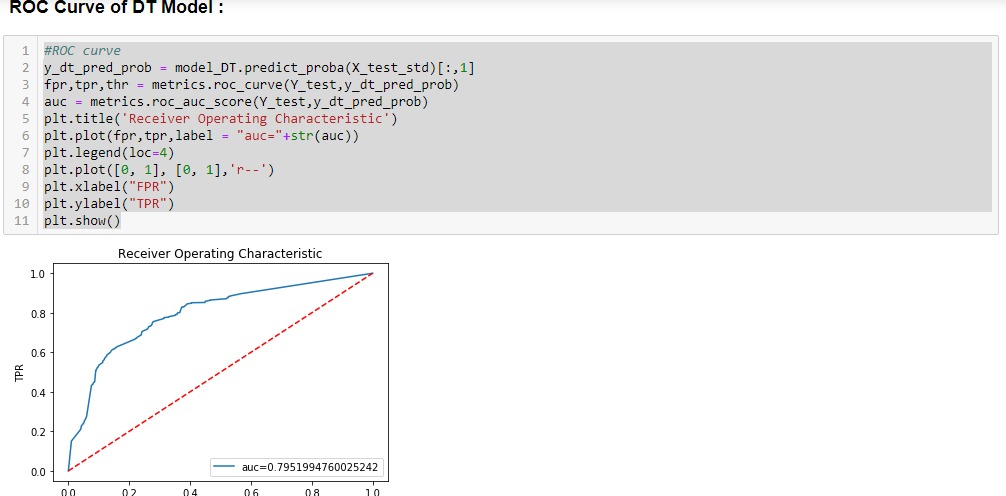
****

****

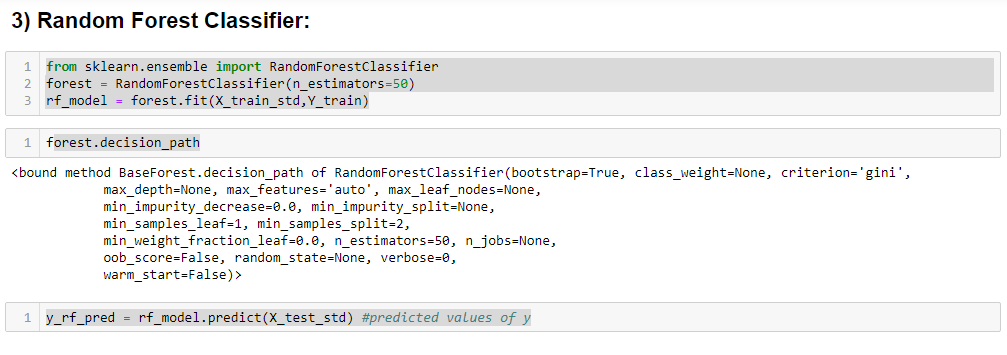
**3.1.2) Decision Tree Model**

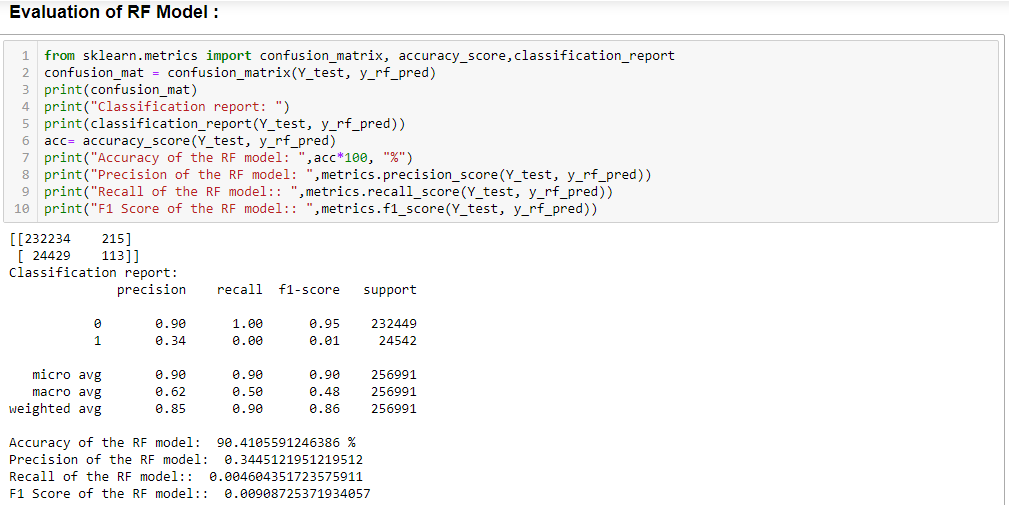
****

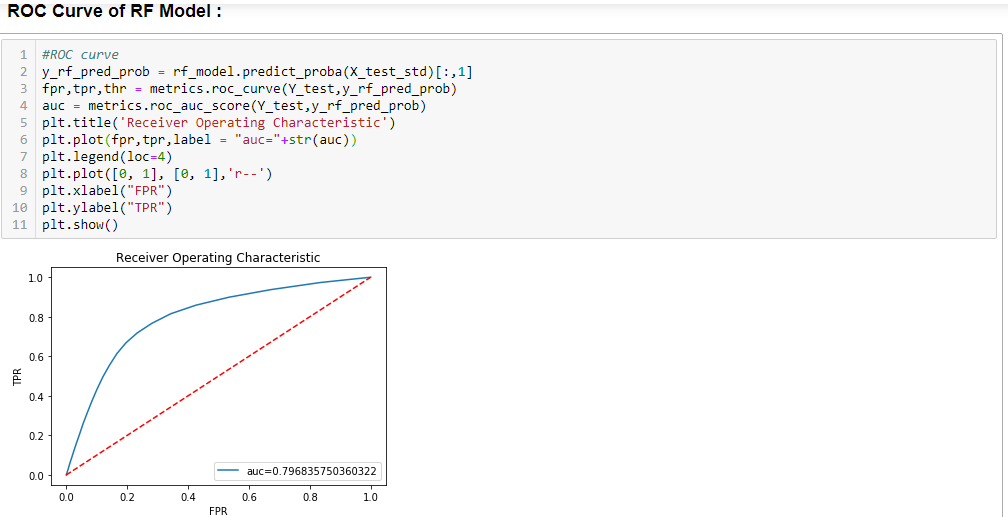
****

****

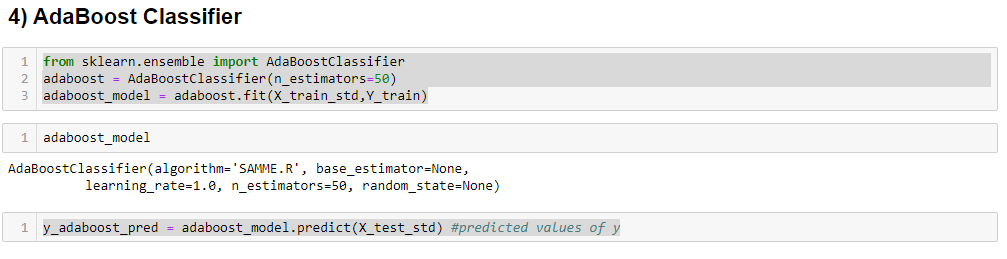
**3.1.3) Random Forest Model**

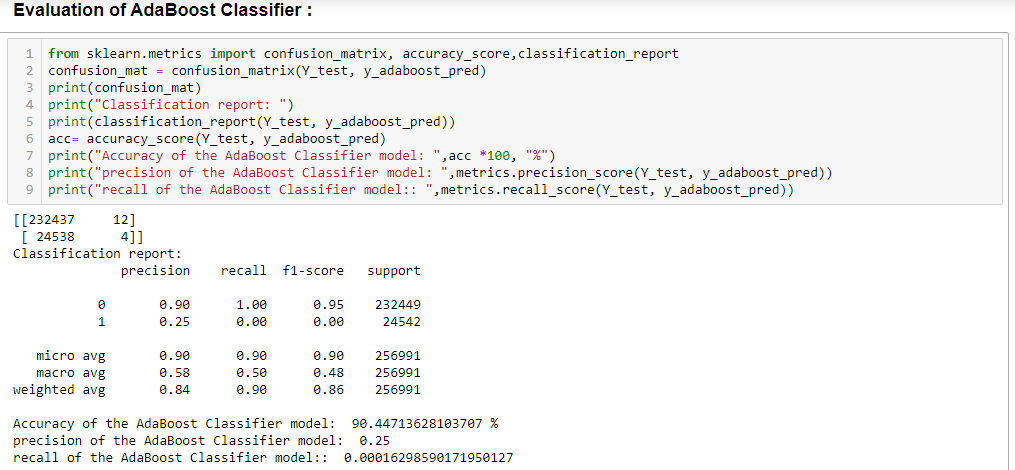
****

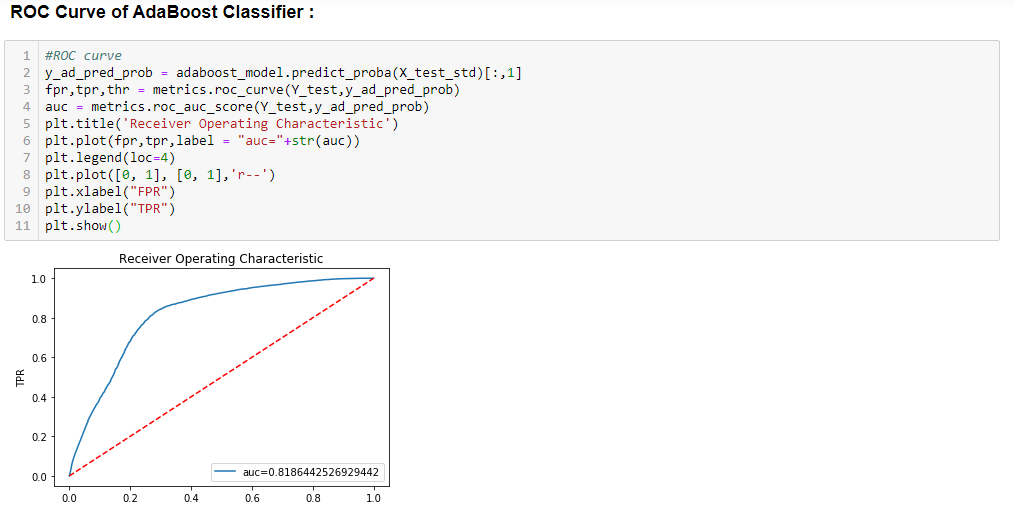
****

****

**3.1.4) AdaBoost Classifier Model**

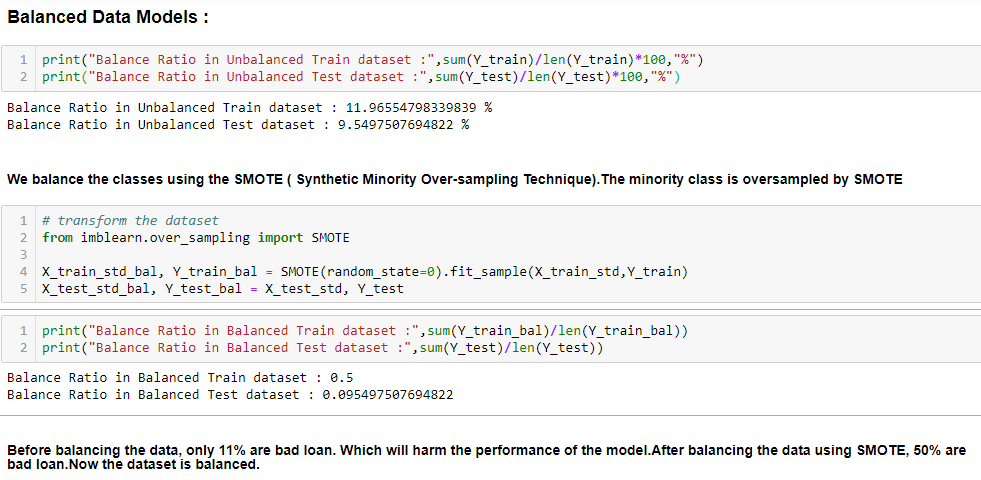
****

****

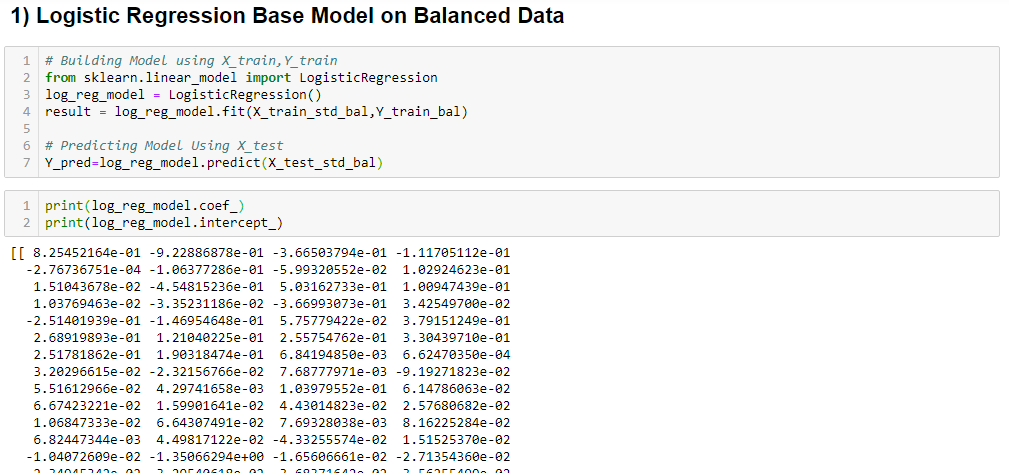
****

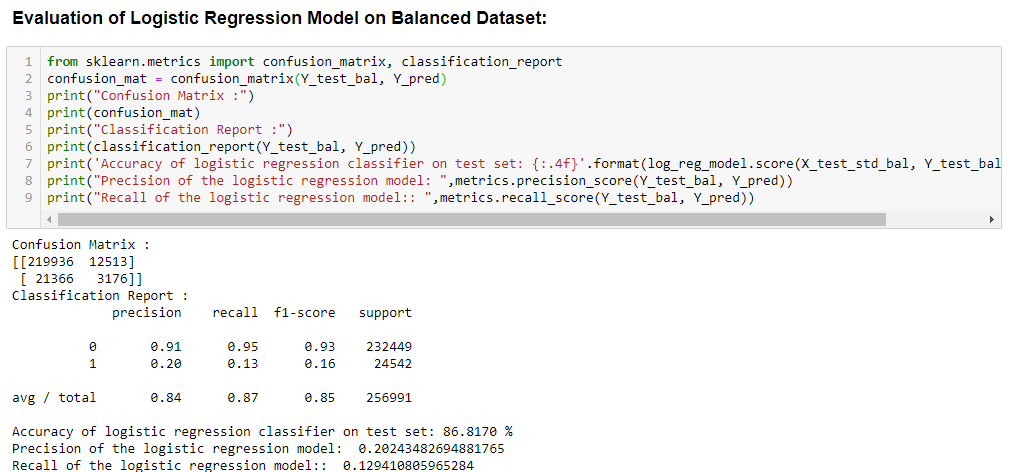
**3.2) Using Balanced Data:**

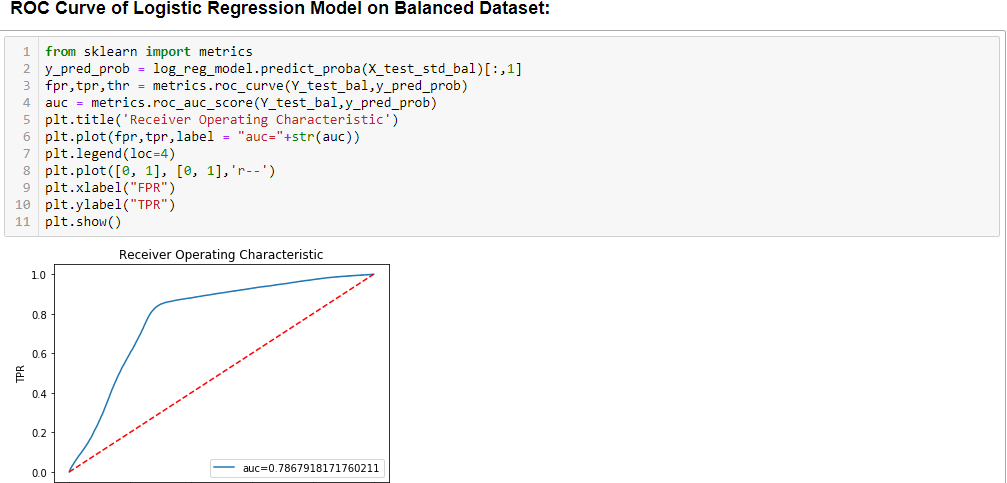
**Class Imbalance Solved using SMOTE:**

****

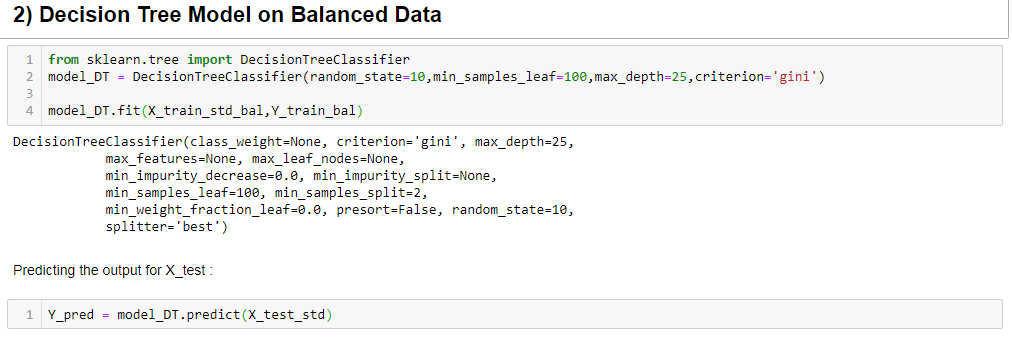
**3.2.1) Logistic Regression Model**

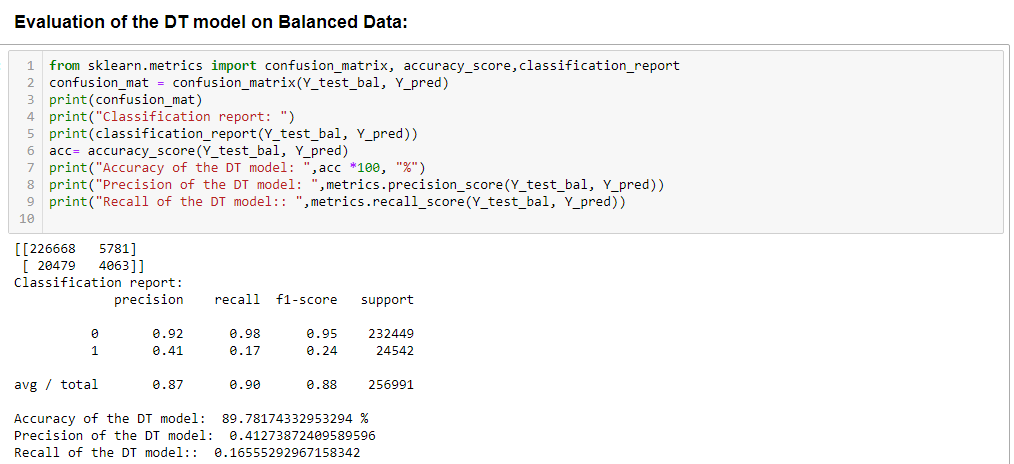
****

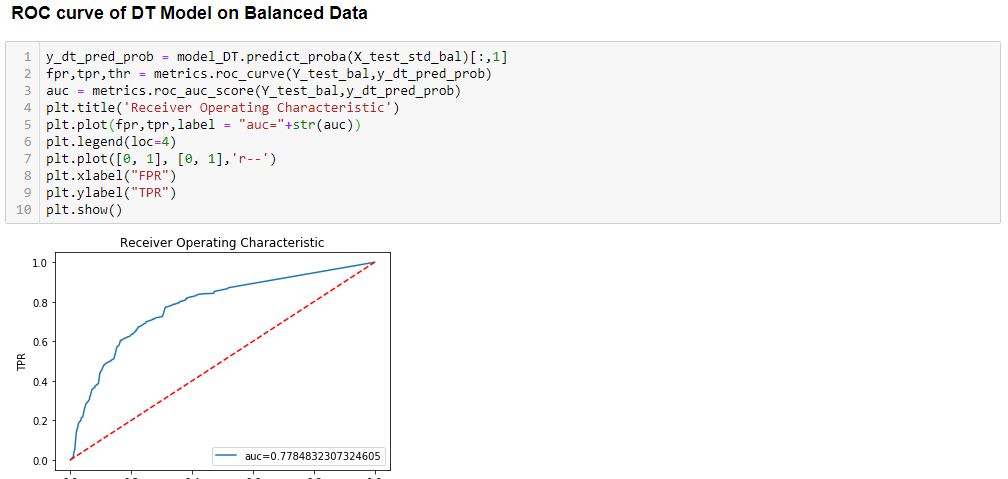
****

****

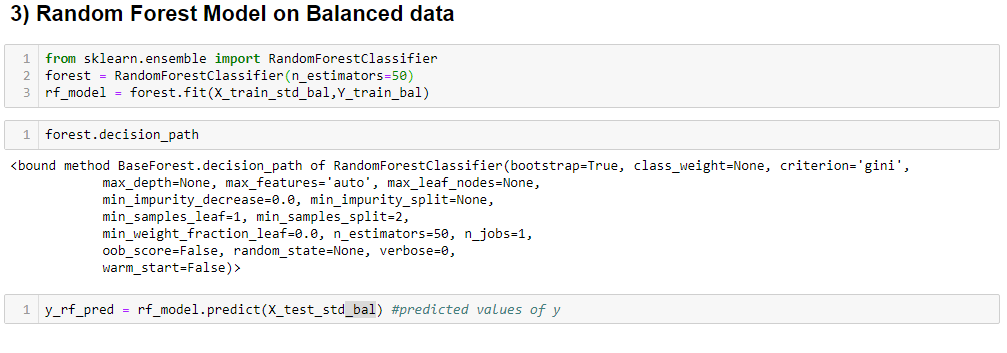
**3.2.2) Decision Tree Model**

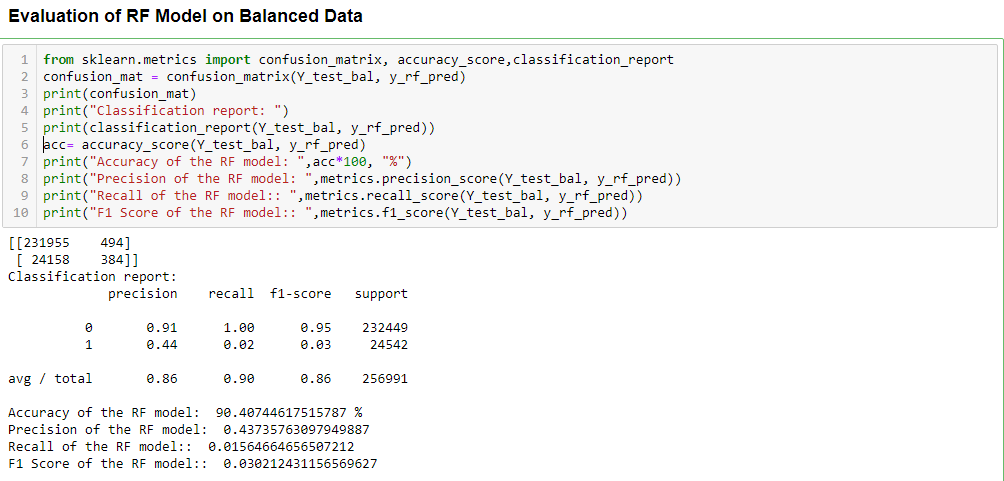
****

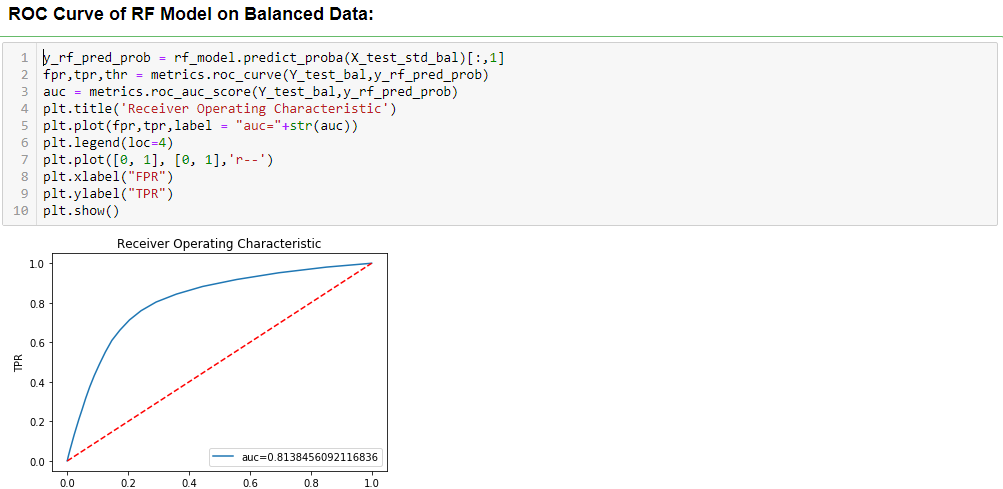
****

****

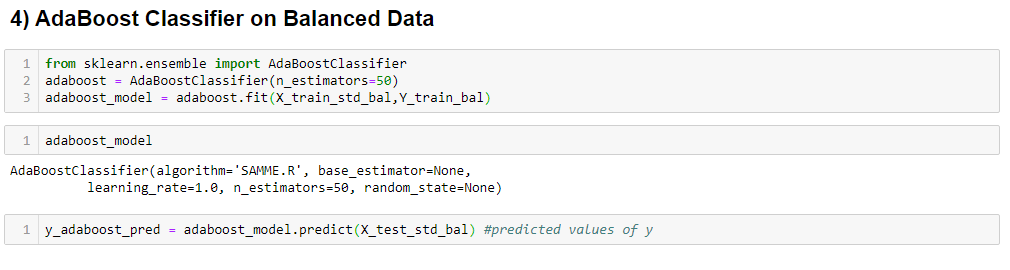
**3.2.3) Random Forest Model**

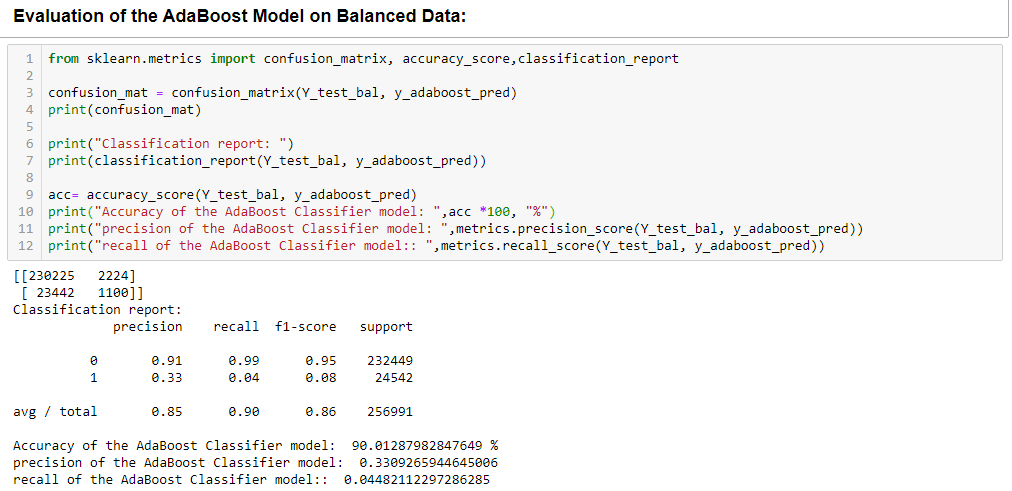
****

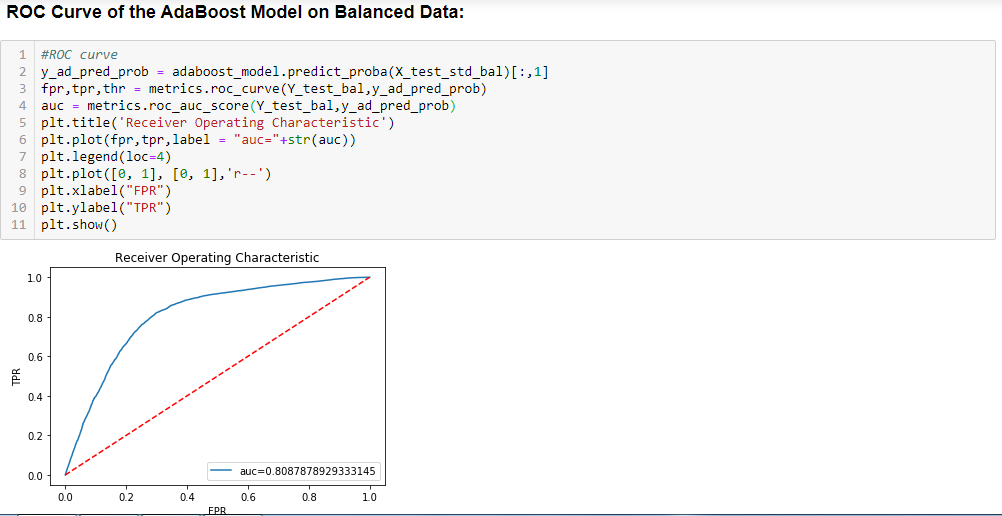
****

****

**3.2.4) AdaBoost Classifier Model**

****

****

****

CHAPTER 4: MODEL COMPARISON

Significant Variables identified in logistic model are also used in other models as well.

Below tables provide a snapshot of the various models which the business can choose from based on the pros and cons of each model.

**Using UnBalanced Data:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sr.no | Model Name | AUC | Precision | Recall | Accuracy |
| 1 | Logistic Regression | 78.02% | 0.3658 | 0.0006 | 90.45% |
| 2 | Decision Tree | 79. 52% | 0.2727 | 0.0001 | 90.46% |
| 3 | Random Forest | 80.20% | 0.3770 | 0.0048 | 90.42 % |
| 4 | Ada Boost Classifier | 81.86 % | 0.2500 | 0.0002 | 90.44 % |

**Using Balanced Data:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sr.no | Model Name | AUC | Precision | Recall | Accuracy |
| 1 | Logistic Regression | 78.68% | 0.2024 | 0.1294 | 86.82% |
| 2 | Decision Tree | 77.85% | 0.4127 | 0.1656 | 89.78% |
| 3 | Random Forest | 81.38% | 0.4373 | 0.0156 | 90.47 % |
| 4 | Ada Boost Classifier | 80.88% | 0.3309 | 0.0448 | 90.01 % |

CHAPTER 5: CONCLUSION

.This project focused on building different machine learning models and evaluating the performance of logistic regression, random forest, decision tree and AdaBoost Classifier with unbalanced train data and balanced train data.

We found that after oversampling the minority class by Synthetic Minority Over-Sampling Technique (SMOTE) in the training set, the precision and recall score improves for every model. Considering all the cases,random forest model that works better than logistic regression.The best model selected out of all models that have been tested is Random Forest model with an accuracy of 90.47% with unbalanced data and 90.42% with balanced data.

Loan Default Analysis is a very crucial part of the banking sector and it plays an important role in the growth of the bank’s profit. Using analyzing techniques one can predict or analyze that a person applying for loan will repay the loan or not. So, multiples algorithms have been implemented to analyze a defaulter. To overcome the crisis of loss of revenue, Loan Default analysis will help banks to grow and earn profit.