**PROJECT REPORT**

**XYZ Corporation Lending Data Project**

*Submitted towards the partial fulfillment of the criteria for award of Genpact Data Science Prodegree by Imarticus*

*Submitted By:*

*Anchal Sriwastava*

*Babita Pant*

*Course and Batch: DSP- 01 Feb2020*



# Abstract

**Keywords**

*Disclaimer: \*Data shared by the customer is confidential and sensitive information. It should not be used for any purposes apart from capstone project submission for DSP. The Name and demographic details of the enterprise is kept confidential as per their owners’ request and binding.*

# Acknowledgements

We are using this opportunity to express my gratitude to everyone who supported us throughout the course of this group project. I am thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the project work. We are sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project.

Further, We are fortunate to have such a great mentor who readily shared his immense knowledge in data analytics and guide us in a manner that the outcome resulted in enhancing our data skills.

We wish to thank, all the faculties, as this project utilized knowledge gained from every course that formed the DSP program.

I certify that the work done by us for conceptualizing and completing this project is original and authentic.

Date: December 15, 2020 Anchal Sriwastava

Place: Noida, UP Babita Pant

# Certificate of Completion

I hereby certify that the project titled **“XYZ Corporation Lending Data Project”** was undertaken and completed by Anchal Sriwastava & Babita Pant from the batch of DSP - 01 (Feb 2020)

Date: December 15, 2020

Place – Noida, UP

Table of Contents

[Abstract 2](#_Toc510332824)

[Acknowledgements 2](#_Toc510332825)

[Certificate of Completion 3](#_Toc510332826)

[CHAPTER 1: INTRODUCTION 6](#_Toc510332827)

[1.1 Title & Objective of the study 6](#_Toc510332828)

[1.2 Need of the Study 6](#_Toc510332829)

[1.3 Business or Enterprise under study 6](#_Toc510332830)

[1.4 Business Model of Enterprise 6](#_Toc510332831)

[1.4 Data Sources 6](#_Toc510332832)

[1.5 Tools & Techniques 6](#_Toc510332833)

[1.6 Infrastructure Challenges 6](#_Toc510332834)

[CHAPTER 2: DATA PREPARATION AND UNDERSTANDING 6](#_Toc510332835)

[2.1 Phase I – Data Extraction and Cleaning: 6](#_Toc510332836)

[2.2 Phase II - Feature Engineering 6](#_Toc510332837)

[2.3 Data Dictionary: 7](#_Toc510332838)

[2.4 Exploratory Data Analysis: 7](#_Toc510332839)

[CHAPTER 3: FITTING MODELS TO DATA 7](#_Toc510332840)

[4.1 LINEAR REGRESSION MODEL 7](#_Toc510332841)

[4.1.1 First Linear Regression Model 7](#_Toc510332842)

[4.1.2 Second Linear Regression model 7](#_Toc510332843)

[4.1.3 Third Linear Regression model 7](#_Toc510332844)

[4.2 RANDOM FOREST 8](#_Toc510332845)

[4.2.1 INFRASTRUCTURE CHALLENGES 8](#_Toc510332846)

[CHAPTER 5: KEY FINDINGS 8](#_Toc510332847)

[CHAPTER 6: RECOMMENDATIONS AND CONCLUSION 9](#_Toc510332848)

[CHAPTER 7: REFERENCES 9](#_Toc510332849)

# 

# CHAPTER 1: INTRODUCTION

## Title & Objective of the study

‘XYZ Corporation Lending Data Project’ is the project we are working upon which falls under the BFSI domain (Banking Financial services and Insurance sector). The text files contain complete loan data for all loans issued by XYZ Corp. through 2007-2015. The primary purpose of working on this project is to predict the probability of default, whether the customer will default the loan or not by using the past data. That means, given a set of new predictor variables, we need to predict the target variable as 1 -> Defaulter or 0 -> Non Defaulter.

## Need of the Study

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

## 1.3 Business or Enterprise under study

XYZ Corporation Lending Data is under the study. Data of Loans issued by XYZ Corp. through 2007-2015 is used for analysis. The data contains the indicator of default, payment information, credit history, etc.

## 1.4 Business Model of Enterprise

Selecting the relevant variables from the dataset and arranging their values in order of importance to create models to predict the probability of default of an individual in the future by performing different types of algorithms on the data.

## 1.5 Data Sources

XYZ Corp Lending Data- Data contains the information about the status of the loan defaulter.

|  |  |
| --- | --- |
| addr\_state | The state provided by the borrower in the loan application |
| annual\_inc | The self-reported annual income provided by the borrower during registration. |
| annual\_inc\_joint | The combined self-reported annual income provided by the co-borrowers during registration |
| application\_type | Indicates whether the loan is an individual application or a joint application with two co-borrowers |
| collection\_recovery\_fee | post charge off collection fee |
| collections\_12\_mths\_ex\_med | Number of collections in 12 months excluding medical collections |
| delinq\_2yrs | The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years |
| desc | Loan description provided by the borrower |
| 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. |
| 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 |
| earliest\_cr\_line | The month the borrower's earliest reported credit line was opened |
| 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. |
| emp\_title | The job title supplied by the Borrower when applying for the loan. |
| funded\_amnt | The total amount committed to that loan at that point in time. |
| funded\_amnt\_inv | The total amount committed by investors for that loan at that point in time. |
| grade | XYZ corp. assigned loan grade |
| home\_ownership | The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER. |
| id | A unique assigned ID for the loan listing. |
| initial\_list\_status | The initial listing status of the loan. Possible values are – W, F |
| inq\_last\_6mths | The number of inquiries in past 6 months (excluding auto and mortgage inquiries) |
| installment | The monthly payment owed by the borrower if the loan originates. |
| int\_rate | Interest Rate on the loan |
| issue\_d | The month which the loan was funded |
| last\_credit\_pull\_d | The most recent month XYZ corp. pulled credit for this loan |
| last\_pymnt\_amnt | Last total payment amount received |
| last\_pymnt\_d | Last month payment was received |
| 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. |
| loan\_status | Current status of the loan |
| member\_id | A unique Id for the borrower member. |
| mths\_since\_last\_delinq | The number of months since the borrower's last delinquency. |
| mths\_since\_last\_major\_derog | Months since most recent 90-day or worse rating |
| mths\_since\_last\_record | The number of months since the last public record. |
| next\_pymnt\_d | Next scheduled payment date |
| open\_acc | The number of open credit lines in the borrower's credit file. |
| out\_prncp | Remaining outstanding principal for total amount funded |
| out\_prncp\_inv | Remaining outstanding principal for portion of total amount funded by investors |
| policy\_code | publicly available policy\_code=1 new products not publicly available policy\_code=2 |
| pub\_rec | Number of derogatory public records |
| purpose | A category provided by the borrower for the loan request. |
| pymnt\_plan | Indicates if a payment plan has been put in place for the loan |
| recoveries | post charge off gross recovery |
| revol\_bal | Total credit revolving balance |
| revol\_util | Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit. |
| sub\_grade | XYZ assigned assigned loan subgrade |
| term | The number of payments on the loan. Values are in months and can be either 36 or 60. |
| title | The loan title provided by the borrower |
| total\_acc | The total number of credit lines currently in the borrower's credit file |
| total\_pymnt | Payments received to date for total amount funded |
| total\_pymnt\_inv | Payments received to date for portion of total amount funded by investors |
| total\_rec\_int | Interest received to date |
| total\_rec\_late\_fee | Late fees received to date |
| total\_rec\_prncp | Principal received to date |
| verified\_status\_joint | Indicates if the co-borrowers' joint income was verified by XYZ corp., not verified, or if the income source was verified |
| zip\_code | The first 3 numbers of the zip code provided by the borrower in the loan application. |
| open\_acc\_6m | Number of open trades in last 6 months |
| open\_il\_6m | Number of currently active installment trades |
| open\_il\_12m | Number of installment accounts opened in past 12 months |
| open\_il\_24m | Number of installment accounts opened in past 24 months |
| mths\_since\_rcnt\_il | Months since most recent installment accounts opened |
| total\_bal\_il | Total current balance of all installment accounts |
| il\_util | Ratio of total current balance to high credit/credit limit on all install acct |
| open\_rv\_12m | Number of revolving trades opened in past 12 months |
| open\_rv\_24m | Number of revolving trades opened in past 24 months |
| max\_bal\_bc | Maximum current balance owed on all revolving accounts |
| all\_util | Balance to credit limit on all trades |
| total\_rev\_hi\_lim | Total revolving high credit/credit limit |
| inq\_fi | Number of personal finance inquiries |
| total\_cu\_tl | Number of finance trades |
| inq\_last\_12m | Number of credit inquiries in past 12 months |
| acc\_now\_delinq | The number of accounts on which the borrower is now delinquent. |
| tot\_coll\_amt | Total collection amounts ever owed |
| tot\_cur\_bal | Total current balance of all accounts |
| verification\_status | Was the income source verified |

## 1.6 Tools & Techniques

Tools: Google Colab

Techniques: Logistic Regression, Decision Tree Classifier, Random Forest Classifier, XG Boost Classifier & AdaBoost Classifier.

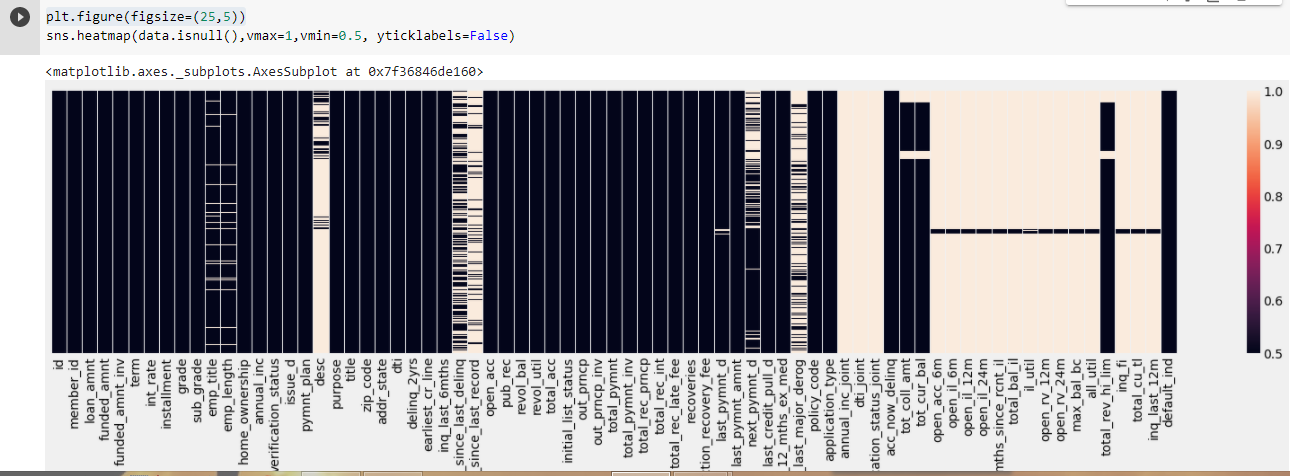
# CHAPTER 2: DATA PREPARATION AND UNDERSTANDING

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

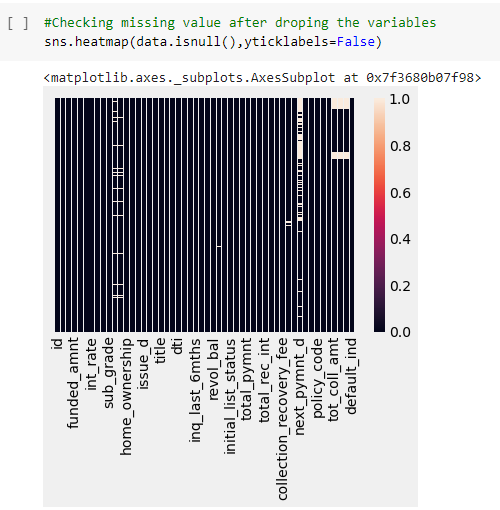
## 2.1 Phase I – Data Extraction and Cleaning:

* **Missing Value Analysis and Treatment**

After printing the shape of the data, we gain that the dataset consists of 855969 observations and 73 variables. The initial step was to check the missing values in each variable and for a better view, plot a heatmap of the dataset for visualizing the missing values as shown below:



It is evident from the above heatmap that our dataset contains a lot of missing values and we cannot use feature that has so many missing values. Above heatmap shows the intensity of values that are missing in every columns. All the light colored columns represents the amount of missing values present in that specific column. Firstly, setting a threshold of 50%, i.e. dropping the columns which have more than or equal to 50% missing values. We are then left with 52 variables. Then visualizing the missing values in each column after dropping the variables, we get the following heatmap:



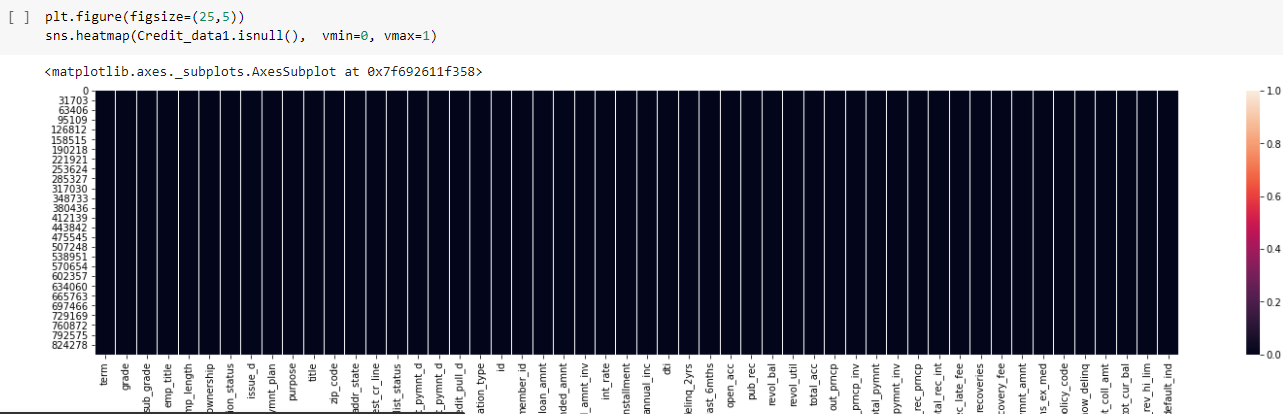
By comparing the above two heatmaps, it is clearly seen that the amount of missing values have been reduced drastically. Also the dataset does not consists any duplicate records.

After that we have segregated the categorical and numerical data for missing value treatment.

The missing values in categorical variable have been treated with **None** and the numerical variables have been treated with **Median** by using for loop.

After that we have concatenated the categorical and numerical features.

After the complete treatment of the missing values, it is evident from the below heatmap that the dataset is now clean and ready for EDA.



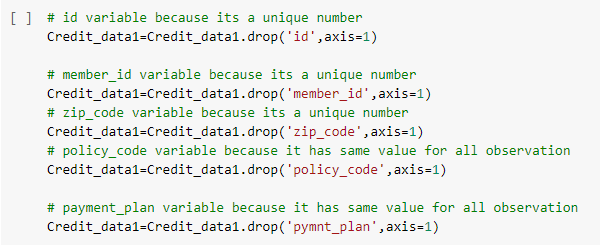
* **Handling Outliers**

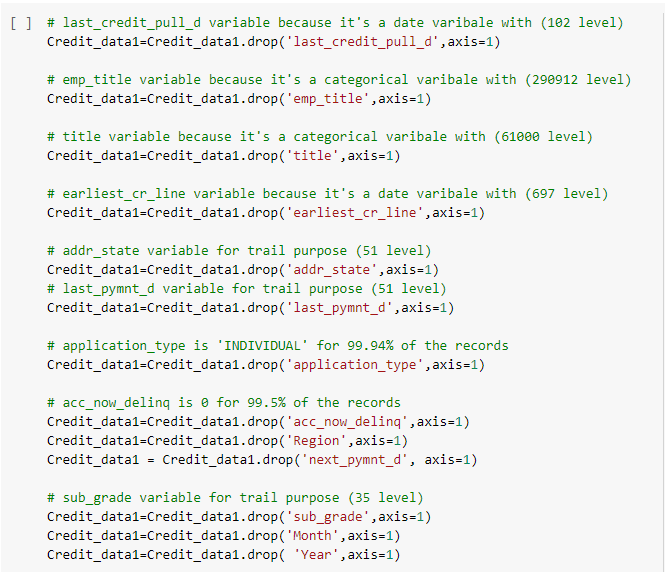
Outlier Treatment was not done because of the following reasons:

* Presence of Clusters in the outliers.
* Less number of outliers as compared to the huge number of observations whose effect will be negligible.
* Lack of Domain knowledge.

## 2.2 Phase II - Feature Engineering

After the missing value treatment we are dropping few more Irrelevant Columns





Now we have total of 855969 observations and 37 variables in our data

After imputing the missing data for categorical variable with none and for numerical variable with median, we split the dataset into Train and Test.





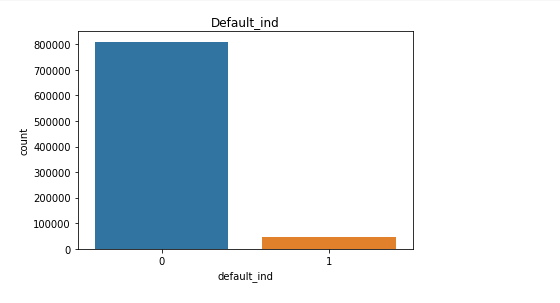






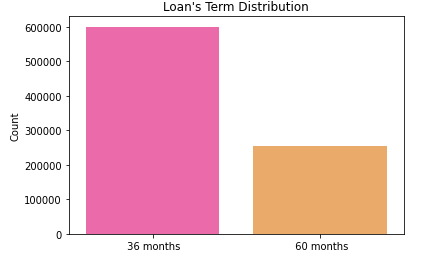
**2.3 Exploratory Data Analysis:** EDA is the process of performing initial investigations on data to discover patterns, to test hypothesis and to check assumptions with the help of descriptive statistics and graphical representations. The response variable in this data is ‘default\_ind’ which indicates that the customer will Default (‘1’) or Non-Default (‘2’)

* **Plot showing the count of the Default customers and Non-default customers in ‘default\_ind’ variable.**

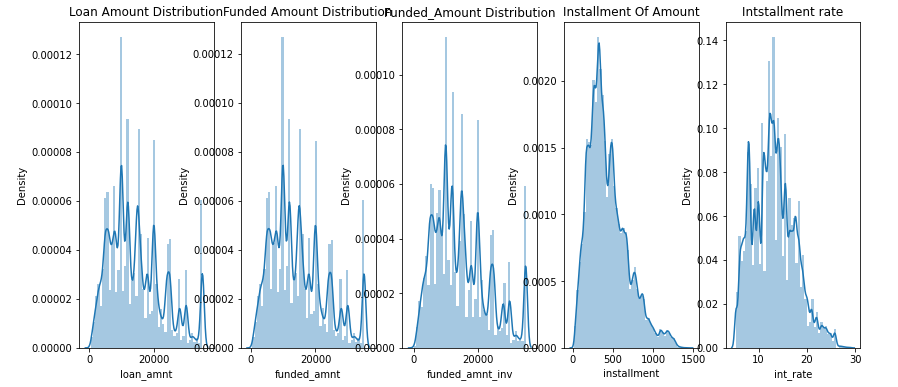


Non-Default Customer: 94.57 % of the dataset. Default Customer: 5.43 % of the dataset. From the above graph, we gain that the dataset is highly unbalanced.

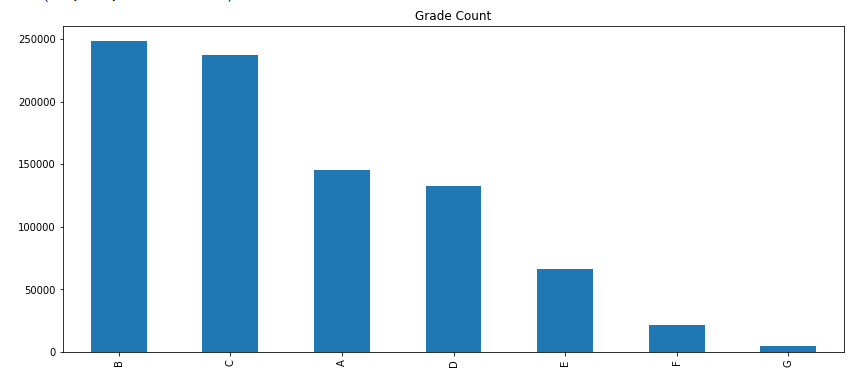
* **Plot showing the distribution of ‘term’ variable.**



• **Plot showing the distribution of loan amount, funded amount, Installments distribution and Interest rates distribution.**

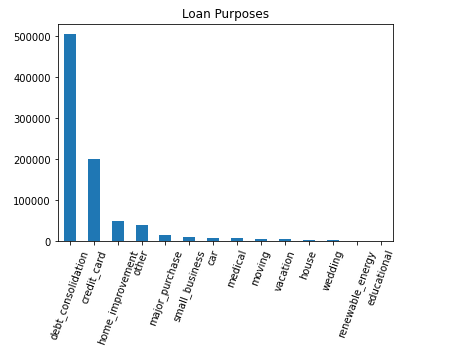


• **Plot showing the Grade count.**



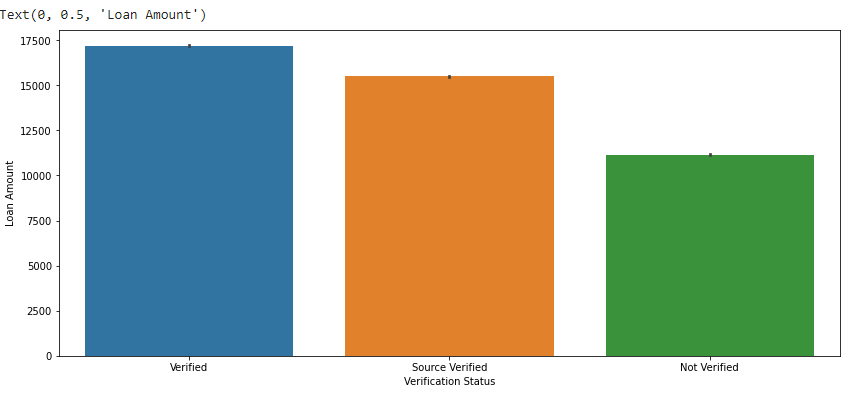
It appears that B and C are the dominant grades.

• **Plot showing the purpose for which the loan was taken by every individual.**

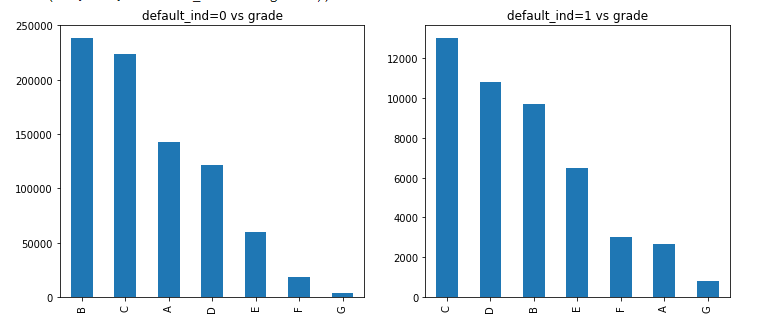


From the above figure, it is observed that huge loans are taken for debt consolidation.

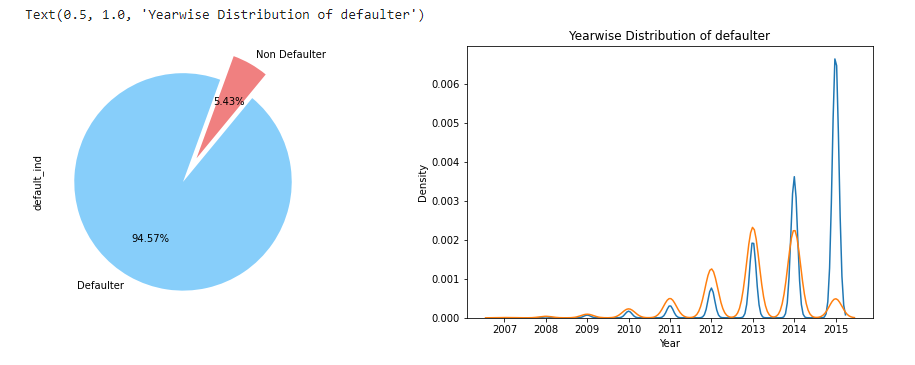
• **Plot showing Loan amount by verification status.**



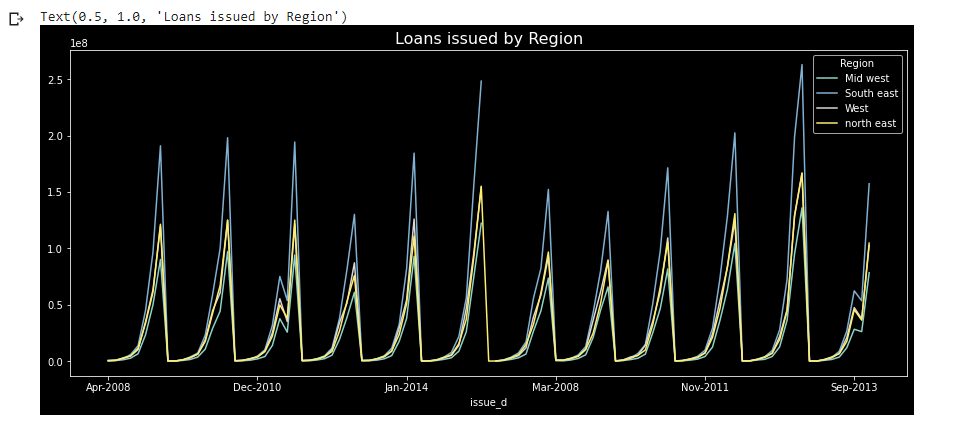
• **Plot showing Defaulter and non-defaulter grade wise.**



• **Plot showing Defaulter and non-defaulter year wise.**



• **Plot showing lone issued by region.**



## 2.4 Encoding:

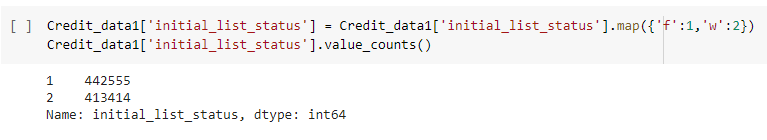
LABEL ENCODING The SciKit Learn library in Python consists of two encoders which are used to convert categorical data or text data into numbers which will help our model to understand. The two encoders are Label Encoder and One Hot Encoder. By importing the LabelEncoder class from the sklearn library, a categorical data or text data can be converted to numbers, fit and transform the respective categorical variable data and then replace the existing text data with the new encoded data. Now when the data has been encoded into numbers, the model might get confused into thinking that a column has data with some kind of order or hierarchy. Therefore, to overcome this One Hot Encoder is used.

**MANUAL LABEL ENCODING**

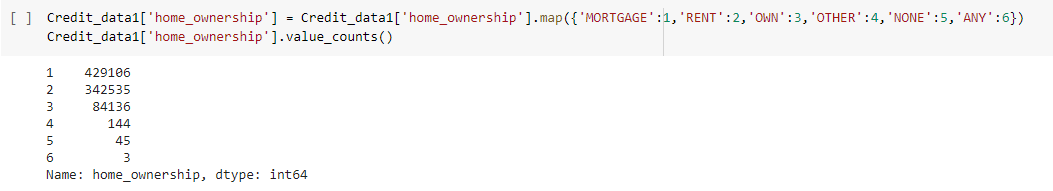
* Employee length in years has 11 levels. The possible values we can assign is from 0 to 10 with 0 indicating less than one year and 10 indicating experience of ten or more years.



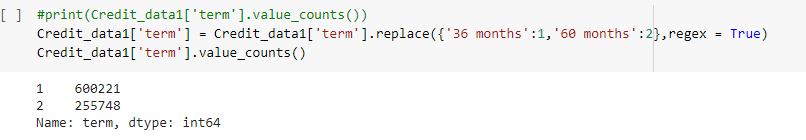
* Initial list status which indicates whether the loan is an individual application or a joint application with two co-borrowers. Replacing f and w with 1 and 2 respectively.



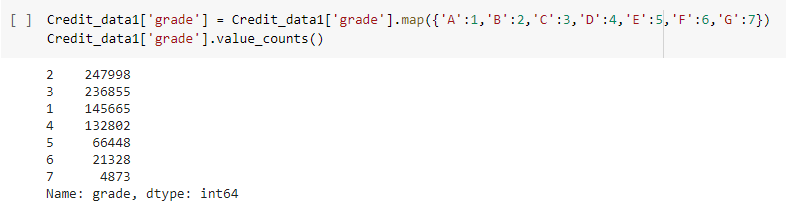
* Home ownership which has 6 levels such as ‘Mortgage’,’ Rent’, ‘Own’, ‘Other’, ‘None’ and ‘Any’ have been label encoded as well.



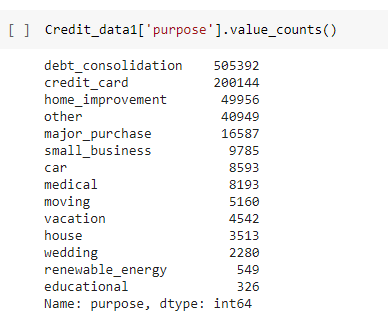
* Similarly the term which consists of 2 levels (36months and 60 months) are label encoded with 1 and 2 respectively.



* 7 levels of Grades which was assigned by XYZ Corp also needed label encoding as well as the purpose variable with 14 levels provided by the borrower for the loan request. Grade:

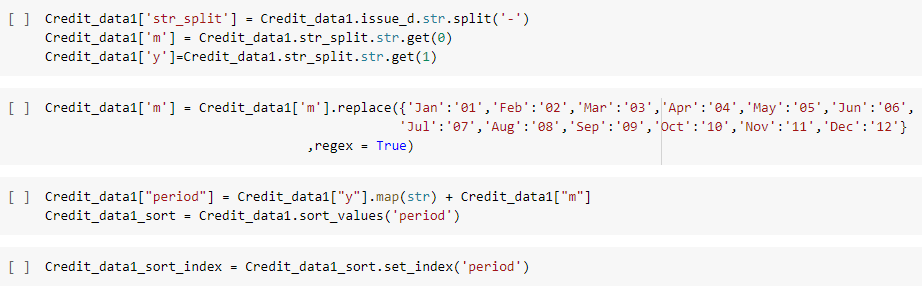


* Purpose:

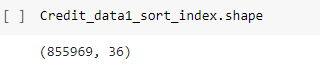


# CHAPTER 3: FITTING MODELS TO DATA

3.1 Data Partition: The data is divided based on the ‘issue\_d’ variable from which the records from June-2007 to May-2015 will go into Training data while the records from June-2015 to Dec-2015 will fall in the Testing data. So to treat the date column i.e. ‘issue\_d’, Split the column into two different columns and replace the values as per the requirement. Then with the help of map function join the split columns and merge them one with a different name (‘period’). Followed by sorting the ‘period’ column and making it an index for slicing according to the requirement.



Followed by dropping the irrelevant columns such as ‘issue\_d’, ‘str\_split’, ‘m’ and ‘y’, we are left with 36 variables.



**Slicing the data into train and test**



**Creating the x\_train, y\_train, x\_test and y\_test dataframes:**



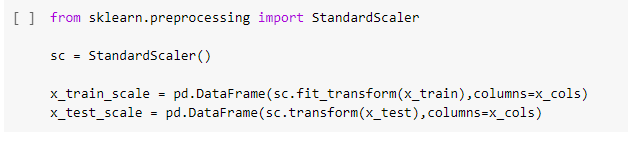






**3.2 Feature Scaling**:

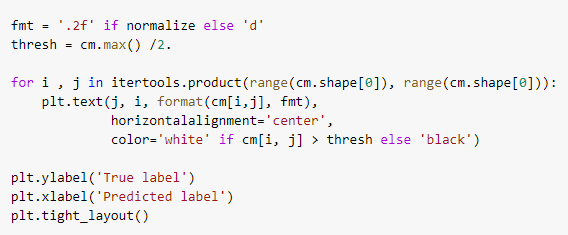
Feature scaling involves rescaling the features so as to limit the range of variables so that they can be compared on common grounds. Using the sklearn library and importing the StandardScaler class, we can use feature scaling.



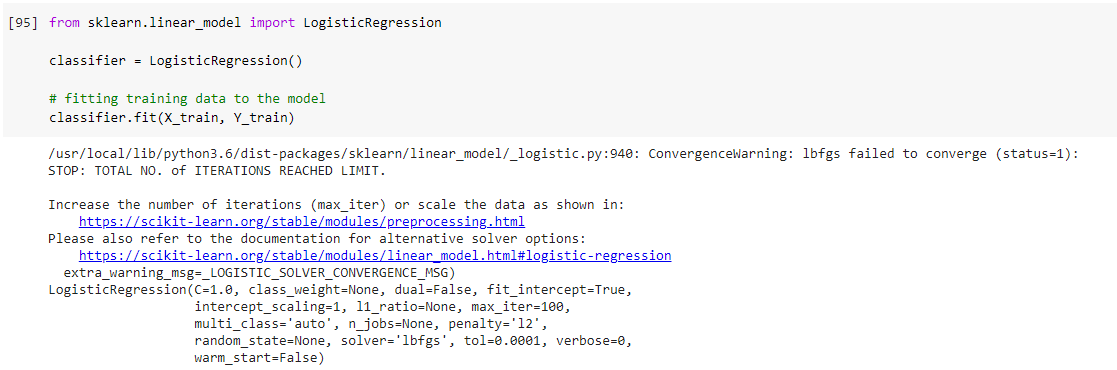
**4.1 MODEL BUILDING**

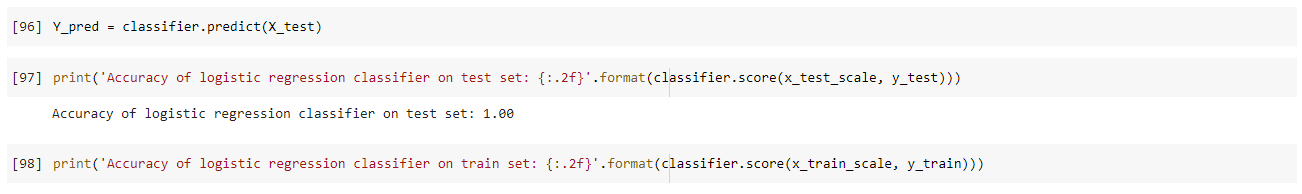
Firstly, we created a custom function for **Confusion Matrix** for better understanding and organized look.

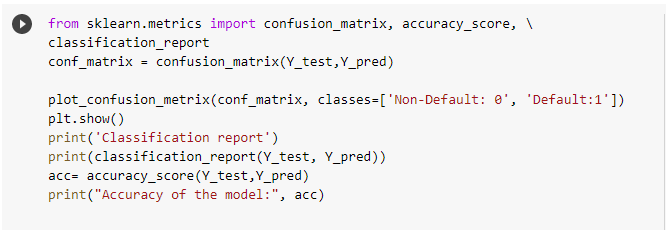




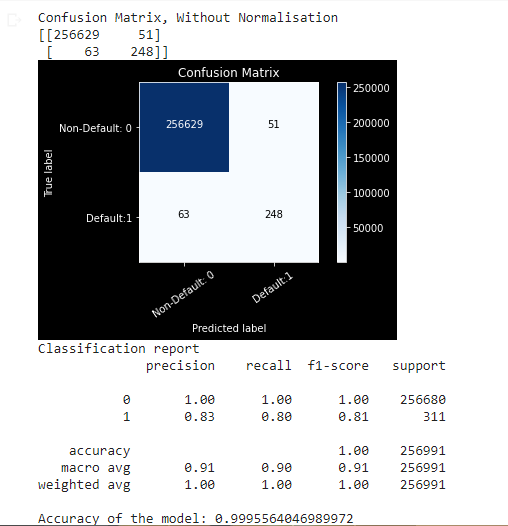
**4.1.1 Logistic Regression Model**







From the sklearn library using the ‘LogisticRegression’ class, we created a logistic regression model and following results were interpreted:



Referring to the above confusion matrix, we can clearly see that the **Type I** error is **51** while the **Type II** error is **63**.

Since the data is unbalanced, we would not focus on the accuracy of the model but instead tune the model for less Type I and Type II errors.

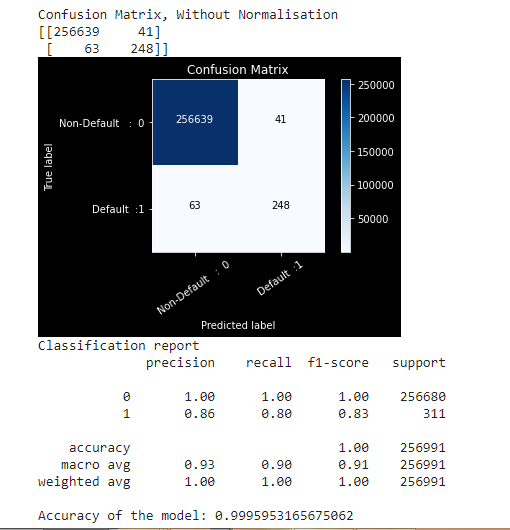
**TUNING THE MODEL**

Adjusting the threshold level of the probabilities to 0.60:

After tuning



**After tuning the model, we get the following results:**



Now the **Type I** error has decreased to **41** after tuning while the Type II error is still the same.

**USING CROSS VALIDATION:**

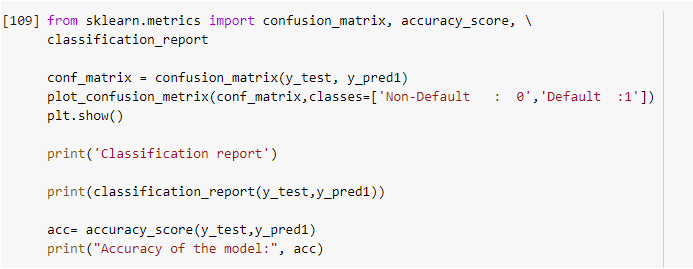


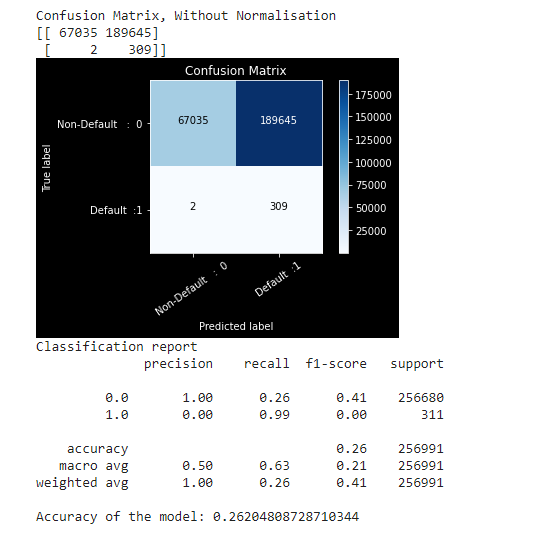


**4.1.2 Decision Tree Classification**

Training the model on the train set and then predicting on the test set using ‘**Entropy**’ for splitter selection and using the ‘DecisionTreeClassifier’ class.



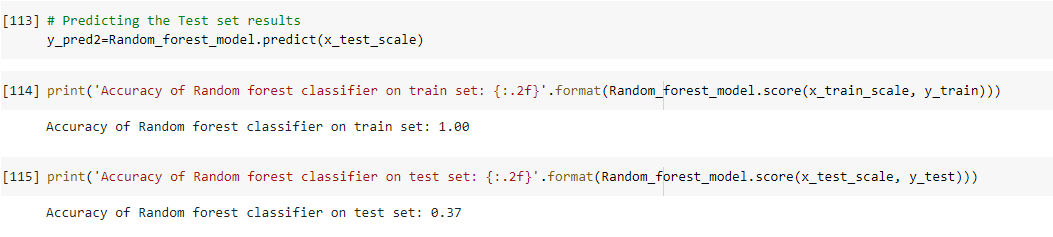


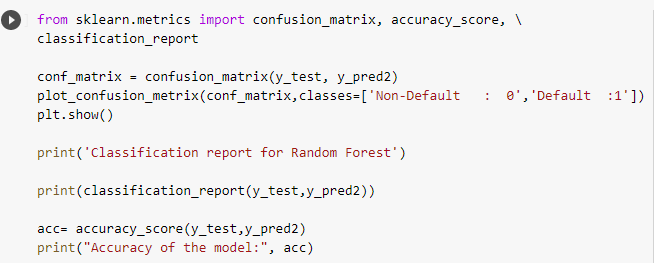


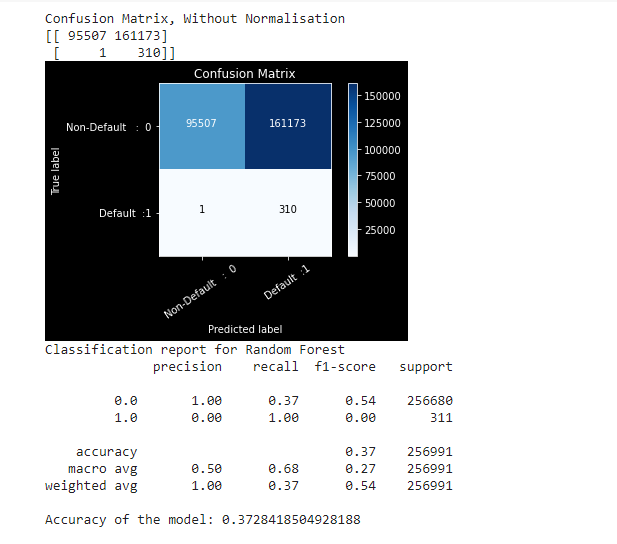
**In this model, the Type II error is low but the Type I error is extremely high which is not acceptable.**

**4.1.3 Random Forest**





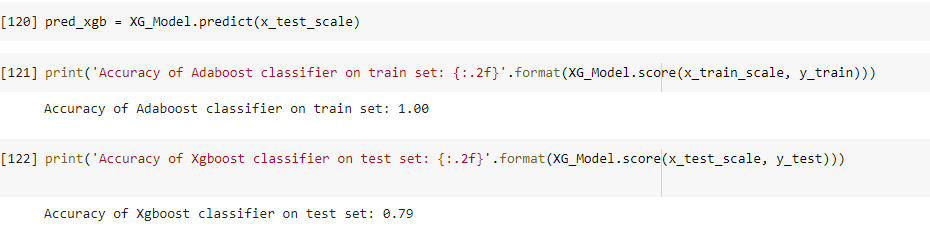


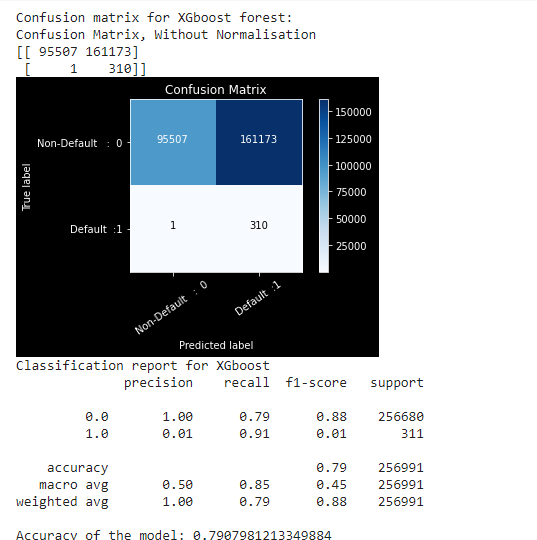


**In this model, the Type II error is low but the Type I error is extremely high which is not acceptable.**

**4.1.4 XgBoost Classifier**

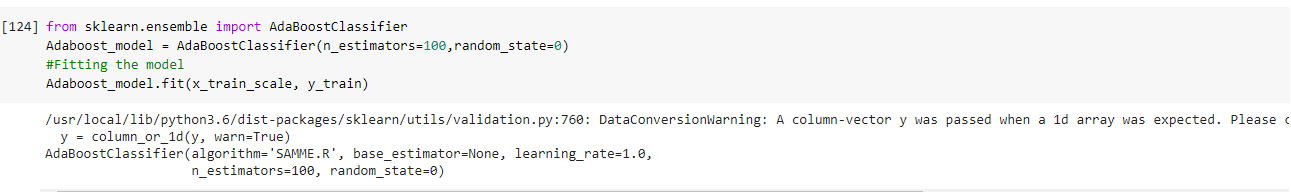


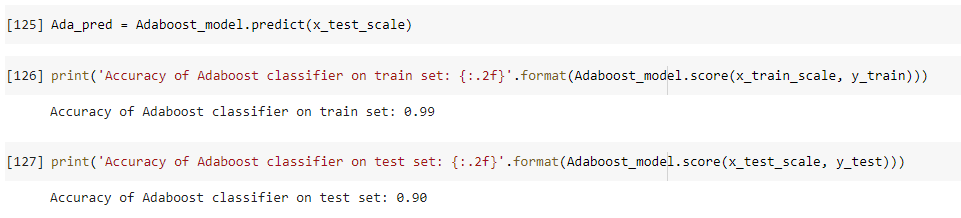


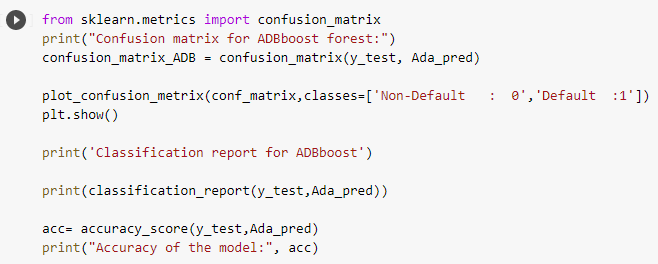


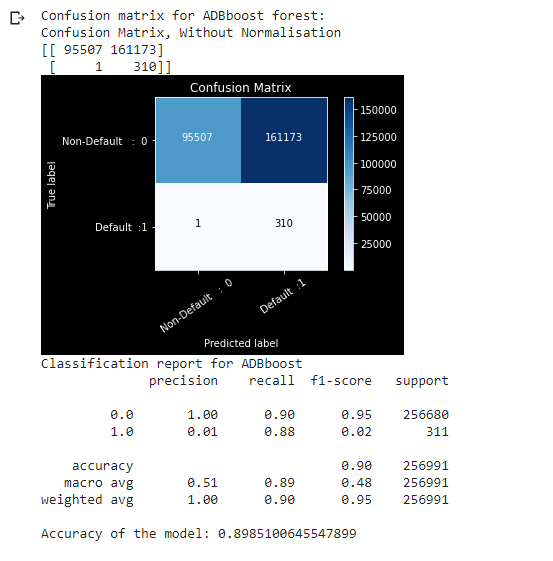
**In this model, the Type II error is low but the Type I error is high but the accuracy is 79%.**

**4.1.5 AdaBoost Classifier**



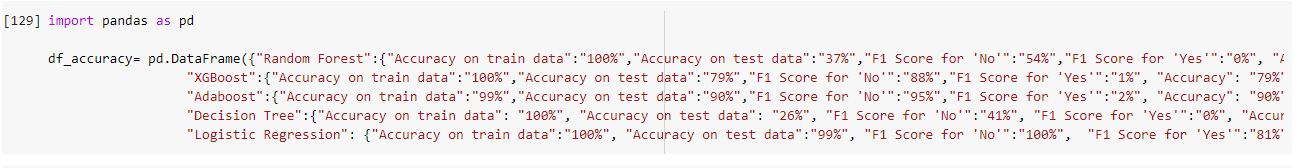


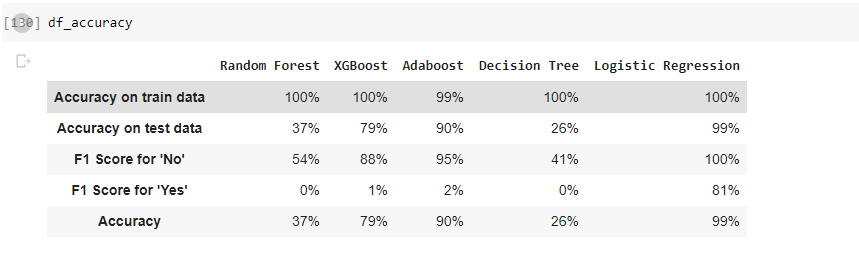




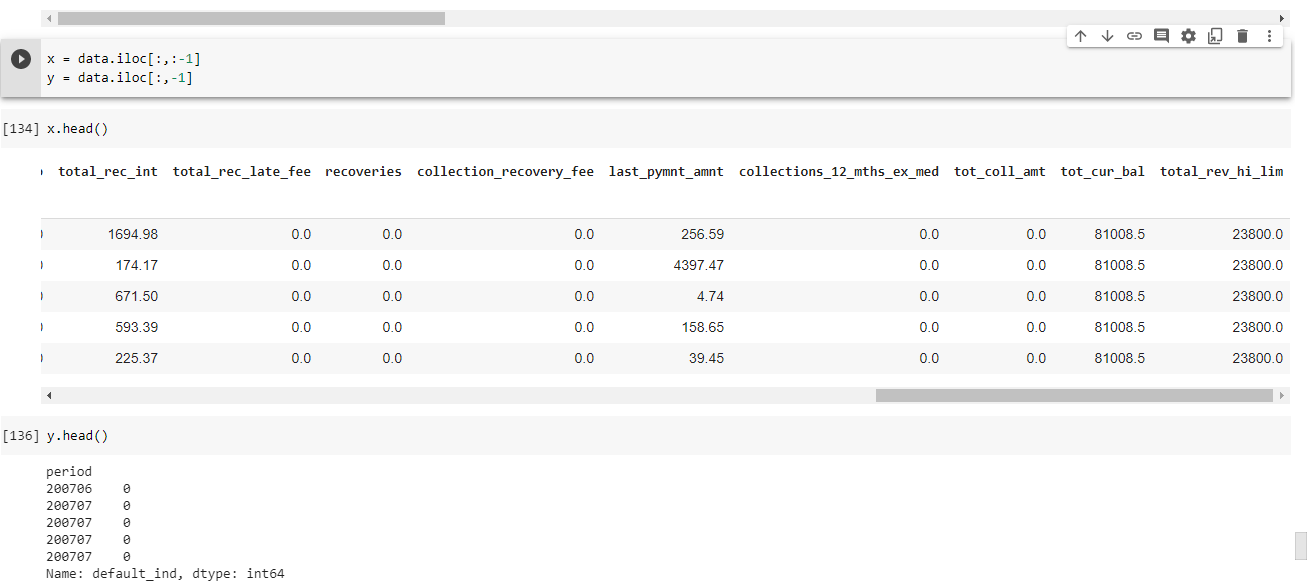
**In this model, the Type II error is 1 but the Type I error is high with the accuracy is 89%.**

CHAPTER 5: Final Model

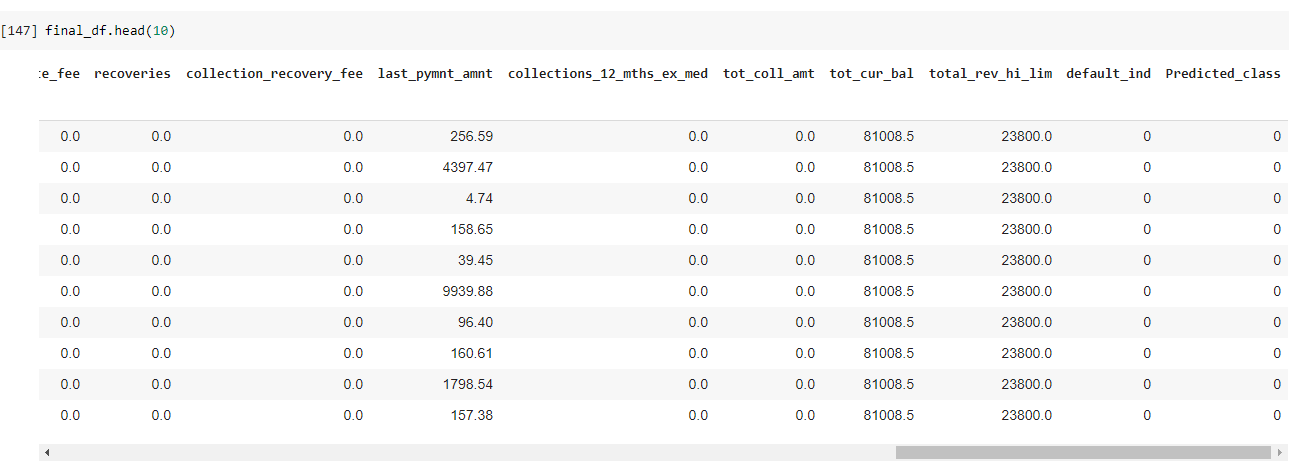




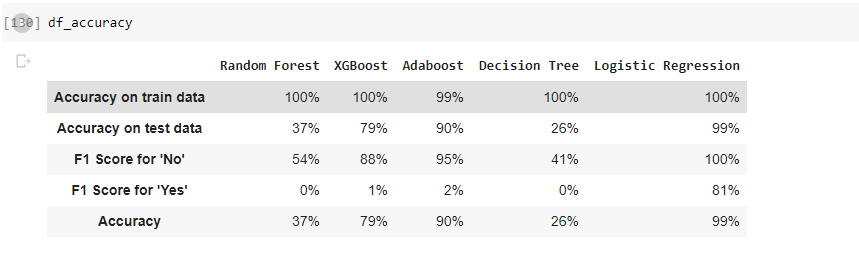
From above accuracy report and checking Type I and Type II error, we have observed that logistic regression model has the highest accuracy with 99% and lowest Type I and Type II error 51 and 63 respectively. Now we will perform prediction on the whole dataset which consists of around 8.55 lacs observations and then concatenated the predicted variable to the dataset for final submission to the client for comparing the actual and predicted values.







**CHAPTER 6: CONCLUSION**



From above accuracy report and checking Type I and Type II error, we have observed that logistic regression model has the highest accuracy with 99% and lowest Type I and Type II error 51 and 63 respectively.