****

**PGPDSE FT Capstone Project – Interim Report**

**Project Group Info:**

|  |  |
| --- | --- |
| Batch details | April 2022 |
| Team members | 5 |
| Domain of Project | Finance (Risk Management) |
| Proposed project title | Peer-To-Peer Credit Risk Analysis |
| Group Number | Group-4 |
| Team Leader | Mr. Gaurav Aher |
| Mentor Name | Ms. Anjana Agrawal |

Date:

G.Y. Aher

Signature of the Mentor Signature of the Team Leader

***Industry Review***

**Current practices:**

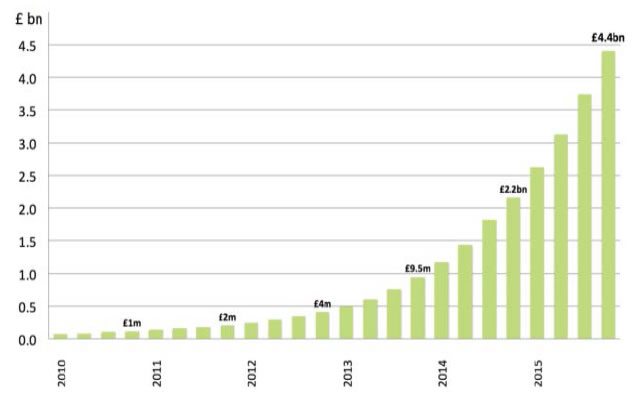
**Peer-To-Peer Lending (P2P Lending)**

P2P lending is a form of crowd-funding used to raise loans which are paid back with interest. It can be defined as the use of an online platform that matches lenders with borrowers in order to provide unsecured loans. The borrower can either be an individual or a legal person requiring a loan. The interest rate may be set by the platform or by mutual agreement between the borrower and the lender. Fees are paid to the platform by both the lender as well as the borrower. The borrowers pay an origination fee (either a flat rate fee or as a percentage of the loan amount raised) according to their risk category. The lenders, depending on the terms of the platform, have to pay an administration fee and an additional fee if they choose to use any additional service (e.g. legal advice etc.), which the platform may provide. The platform provides the service of collecting loan repayments and doing preliminary assessment on the borrower’s creditworthiness. The fees go towards the cost of these services as well as the general business costs. The platforms do the credit scoring and make a profit from arrangement fees and not from the spread between lending and deposit rates as is the case with normal financial intermediation.

While crowd funding - equity, debt based and fund based- would fall under the purview of capital markets regulator (SEBI), P2P lending would fall within the domain of the Bank.

**Market Size**

According to data released by P2PFA, the cumulative lending through P2P platforms globally, at the end of Q4 of 2015, has reached 4.4 billion GBP1. Lending through P2P has grown dramatically from 2.2 million GBP in 2012 to 4.4 billion GBP in 2015.



**Regulatory Practices (US)**

There are two levels of regulation, Federal regulation through the Securities and Exchange Commission (SEC) and State level, where platforms have to apply on a state-by-state basis. One level below the federal requirements is state regulation. Some states outright ban the practice of P2P lending (e.g. Texas). Other states place limits on the type of investors using the platforms to lend (e.g. California). In addition, if a platform wishes to operate across multiple state boundaries, it must apply to each state separately.

**Background Research**

Lending Club is a financial services company headquartered in San Francisco, California. It was the first peer-to-peer lender to register its offerings as securities with the Securities and Exchange Commission (SEC), and to offer loan trading on a secondary market. At its height, Lending Club was the world's largest peer-to-peer lending platform. The company reported that $15.98 billion in loans had been originated through its platform up to December 31, 2015.

**History**

LendingClub was initially launched on Facebook as one of Facebook's first applications. After receiving $10.26 million in a Series A funding round in August 2007, from venture capital investors Norwest Venture Partners and Canaan Partners, LendingClub was developed into a full-scale peer-to-peer lending company.

**Recognition**

In 2011 and 2012 the company was named to as one of the AlwaysOn Global 250LendingClub is the winner of the World Economic Forum 2012 Technology Pioneer Award. It has been recognized by Forbes as one of America’s 20 most promising companies in 2011and 2012,and by Fast Company as one of the ten most innovative financial companies in the world. It was named one of the Disruptor 50 by CNBC in May 2013 and 2014, as a disruptive innovator in next generation financial services. In 2014, LendingClub was recognized by Inc. as one of the 500 Fastest Growing Private Companies in America at #248. Renaud Laplanche, the company’s founder and CEO, also received The Economist Innovation Award in 2014 for the consumer products category

**Literature Survey - Publications, Application, past and undergoing research**

1. Application of Machine Learning Algorithms to Default Prediction using Naïve Bayes and Decision Tree, Random Forest Classifiers and various Boosting techniques
2. It is a peer-reviewed journal Author by (Mogi Jordan Christ 1, Rahmanto Nikolaus Permana Tri2, Wiranto Chandra3, Tuga Mauritsius4) &  
   published by **Binus University.(**[**link**](https://www.researchgate.net/publication/337047465_Lending_Club_Default_Prediction_using_Naive_Bayes_and_Decision_Tree)**)**

**4.Problem understanding**

The reason behind describing customer credit risk analytics in the preceding paragraphs is because, the goal of our next Machine Learning project is to develop a model that can predict customers who are most likely to default. With the help of various visualization libraries that are at our disposal, we will be able to figure out possible parameters that govern a customer’s decision to default.

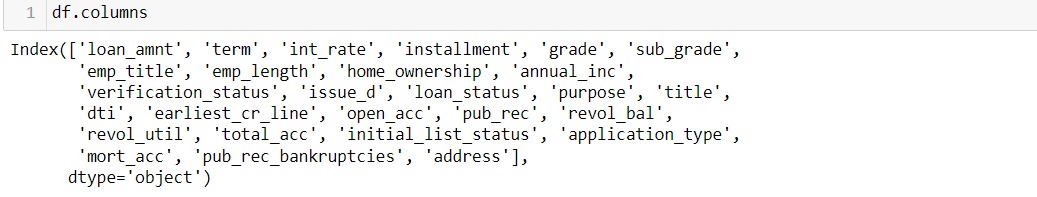
**Dataset and Domain**

**Data Dictionary**

The data set was downloaded from Kaggle namely **lending\_club\_loan\_two.csv**  
The lending\_club.csv gives information about the loans borrowed by borrower’s from 2007-2015  
and also information about the defaulters.

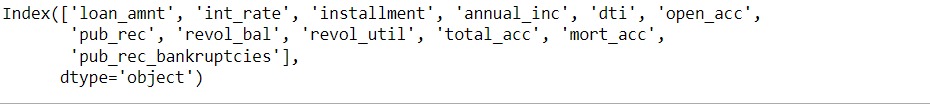
**It has 27 attributes and 396030 rows**.

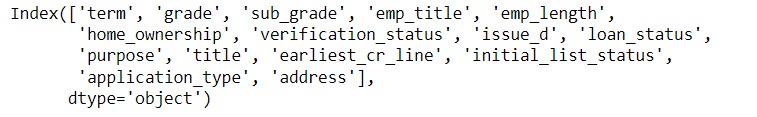
The attribute/feature/column names are given below.

****

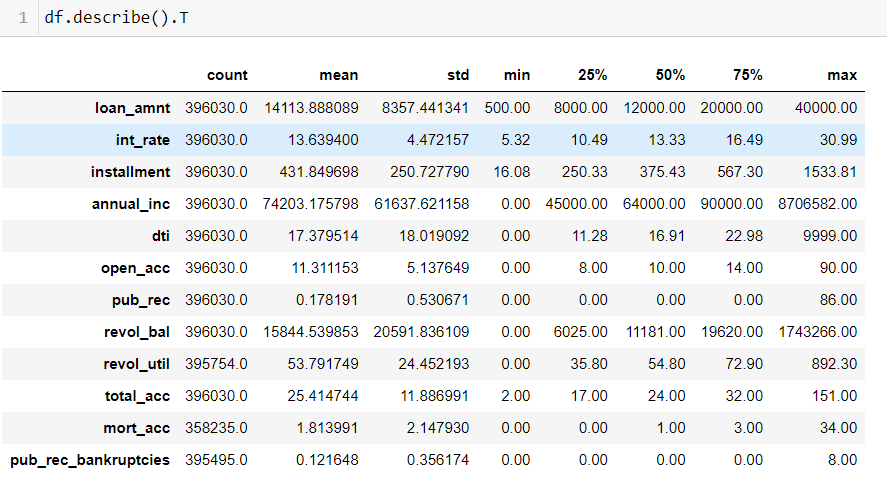
|  |  |  |
| --- | --- | --- |
| **Features** | **Describtions** | **Data Type** |
| **loan\_amnt** | The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.(between 500 $ to 40000 $) | **float64** |
| **term** | The number of payments on the loan. Values are in months and can be either 36 or 60. | **object** |
| **int\_rate** | Interest Rate on the loan(between 5.32 % to 30.99 %) | **float64** |
| **installment** | The monthly payment owed by the borrower if the loan originates. | **float64** |
| **grade** | LC assigned loan grade | **object** |
| **sub\_grade** | LC assigned loan subgrade | **object** |
| **emp\_title** | The job title supplied by the Borrower when applying for the loan. | **object** |
| **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. | **object** |
| **home\_ownership** | The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are RENT, OWN, MORTGAGE, OTHER | **object** |
| **annual\_inc** | The self-reported annual income provided by the borrower during registration. | **float64** |
| **verification\_status** | Indicates if income was verified by LC, not verified, or if the income source was verified | **object** |
| **issue\_d** | The month which the loan was funded | **object** |
| **loan\_status** | Current status of the loan | **object** |
| **purpose** | A category provided by the borrower for the loan request. | **object** |
| **title** | The loan title provided by the borrower | **object** |
| **dti** | A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income. | **float64** |
| **earliest\_cr\_line** | The month the borrower's earliest reported credit line was opened | **object** |
| **open\_acc** | The number of open credit lines in the borrower's credit file. | **float64** |
| **pub\_rec** | Number of derogatory public records | **float64** |
| **revol\_bal** | Total credit revolving balance | **float64** |
| **revol\_util** | The total number of credit lines currently in the borrower's credit file | **float64** |
| **total\_acc** | The total number of credit lines currently in the borrower's credit file | **float64** |
| **initial\_list\_status** | The initial listing status of the loan. Possible values are – W, F | **object** |
| **application\_type** | Indicates whether the loan is an individual application or a joint application with two co-borrowers | **object** |
| **mort\_acc** | Number of mortgage accounts. | **float64** |
| **pub\_rec\_bankruptcies** | Number of public record bankruptcies | **float64** |
| **address** | Borrowers address | **object** |

**The numerical Variable –**

****

**The categorical columns are:-**

**Data Pre-processing**

****

**Interpretation:**

1. Count of all features are not same so we can say that there is missing values in the Dataset.

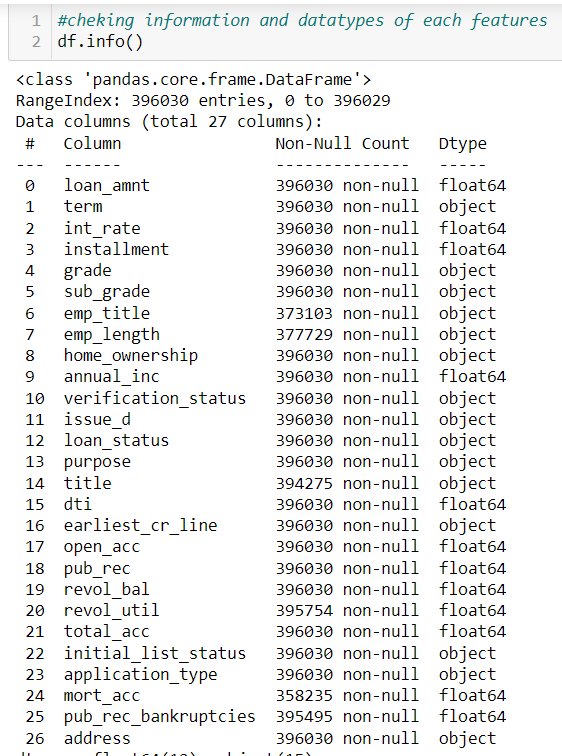
2. Minimum values of Dataset is positive so we can say that Dataset has no negative Datapoint.

3. The difference Between mean and median of each variable is more, so we can say that data is not normally distributed.

4. The difference Between min and max of each variable is more, so we can say that Some of the features also contains potential outliers.

**Check Data type**

**Check the data type of each variable. If the data type is not as per the data definition, change the data type.**

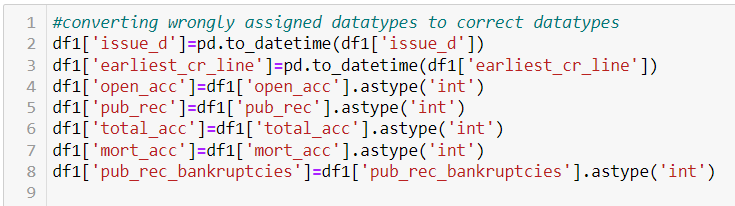
****

**Inference:**

Dataset has 27 features in which 12 are numerical`(float64)` type and 15 are categorical`(object)` type features and also

issue\_d , open\_acc , pub\_rec , total\_acc , mort\_acc , pub\_rec\_bankruptcies datatypes are wrongly specified.

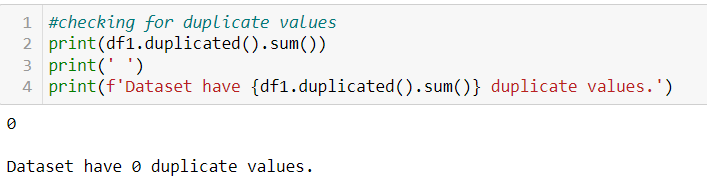
**Data type conversions**

****

**Interpretation:** Now all the variables are similar Data types**.**

**Data cleaning**

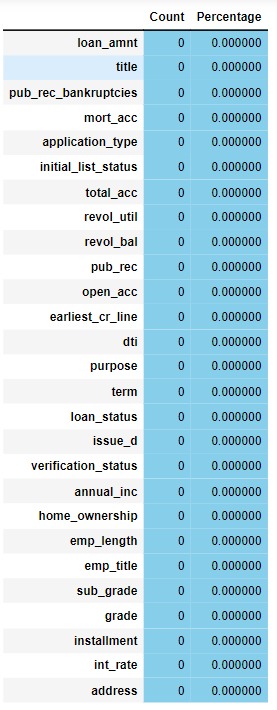
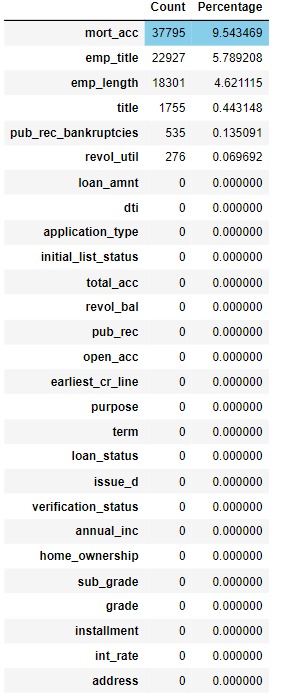
**Checking duplicate values**

****

**Missing Values Treatment:**

Missing values plays a prominent role in the dataset. Generally, we can drop the columns or rows depending the percentage of missing values. We can also replace the missing values with optimum values. In order to perform such operations, we will first look into the overall missing values in each column using the below python code.

**BEFORE IMMPUTING NULL VALUES** **AFTER TREATING NULL VALUES**

  
we found that for most of the `NaN` Values of 'title' in the 'application type' column is `individual(94.87%)`, so in our further checks we consider application type `Individual` and check the corresponding values against individual in the title column and found it to be 'Debt consolidation' .Thus, we impute the null values of the title column with 'Debt consolidation'

**Project -Statement-**

Default analysis using various machine Learning technique’s

**Complexity involved /Project Outcome/** **Commercial/ Academic / Social value -**

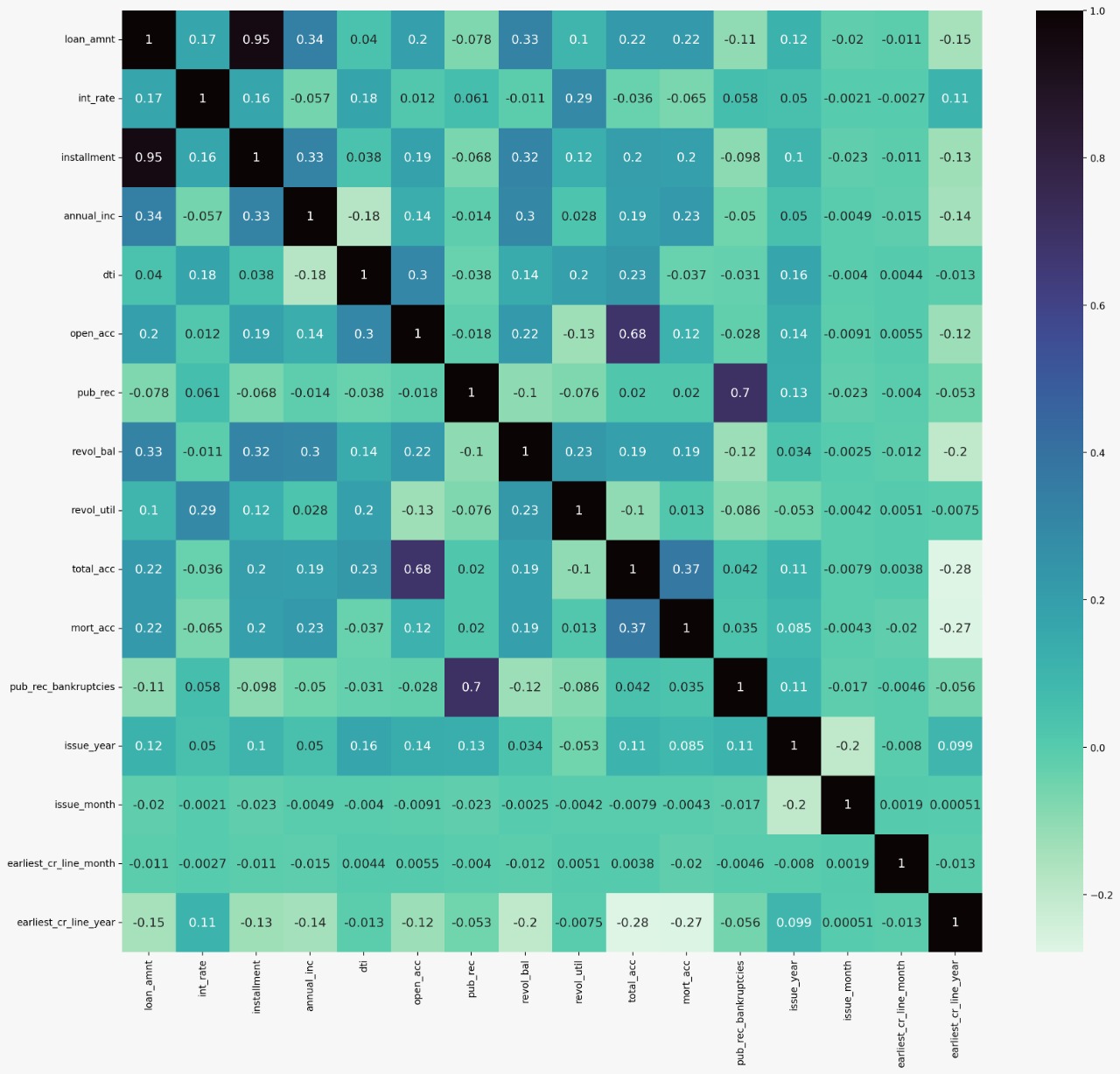
Our project scope is to run the exploratory data analysis to find the business insights from our loan data, and to build a learning model using data mining techniques / machine leaning algorithms that will use the historic loan data to learn and helps to identify loans/borrowers which are likely to default.

As per the recent studies, 3-4% of the total loans defaults every year. This is a huge risk for the investors who is funding the loans. Investors require more comprehensive assessment of these borrowers than what is presented by Lending Club to make a smart business decision. Data mining techniques and Machine Learning model/analysis could help predicting the loan default likelihood which may allow investors to avoid loan defaults thus limiting the risk of their investments.

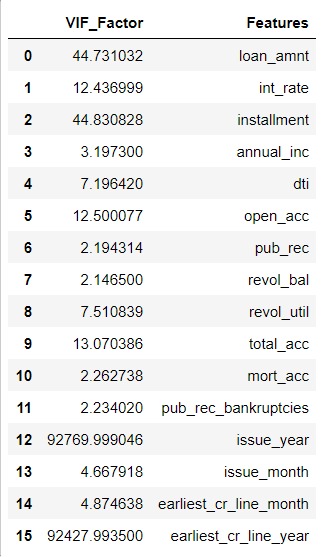
* **Data Exploration (EDA)**

**Relationship between variables**

**Dealing with High Correlation:**

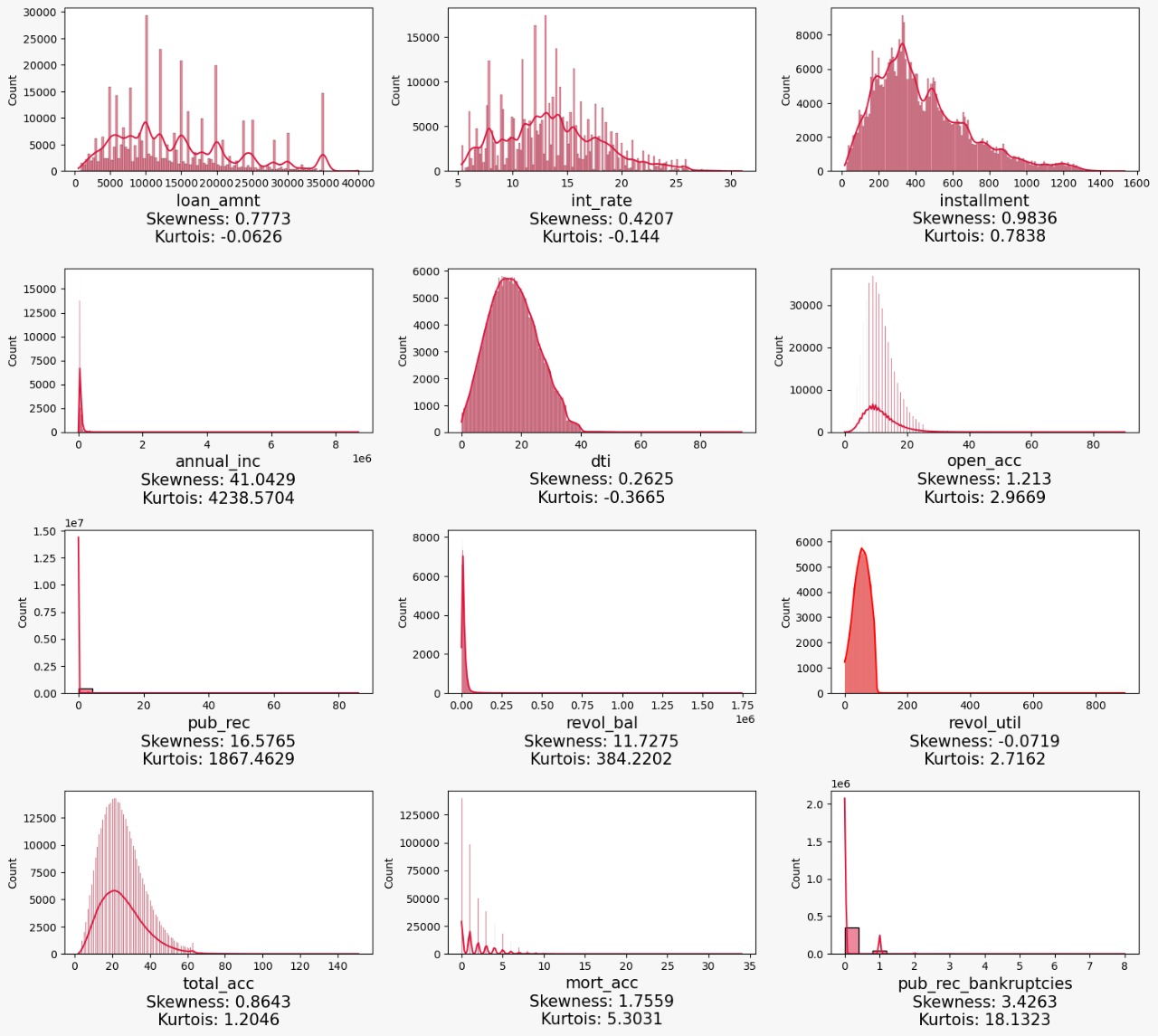
****

**Treating Multi-** **Collinearity by using The Variance Inflation Factor (VIF) The value of VIF equal to 1 indicates that no features are correlated. We calculate VIF of the numerical independent variables.**

****

**Features involved in multicollinearity are loan\_amt , int\_rate, installment, open\_acc, total\_acc, issue\_year, earliest\_cr\_line\_year.  
We can treat this features by using PCA transform or we can drop it based upon model need.**

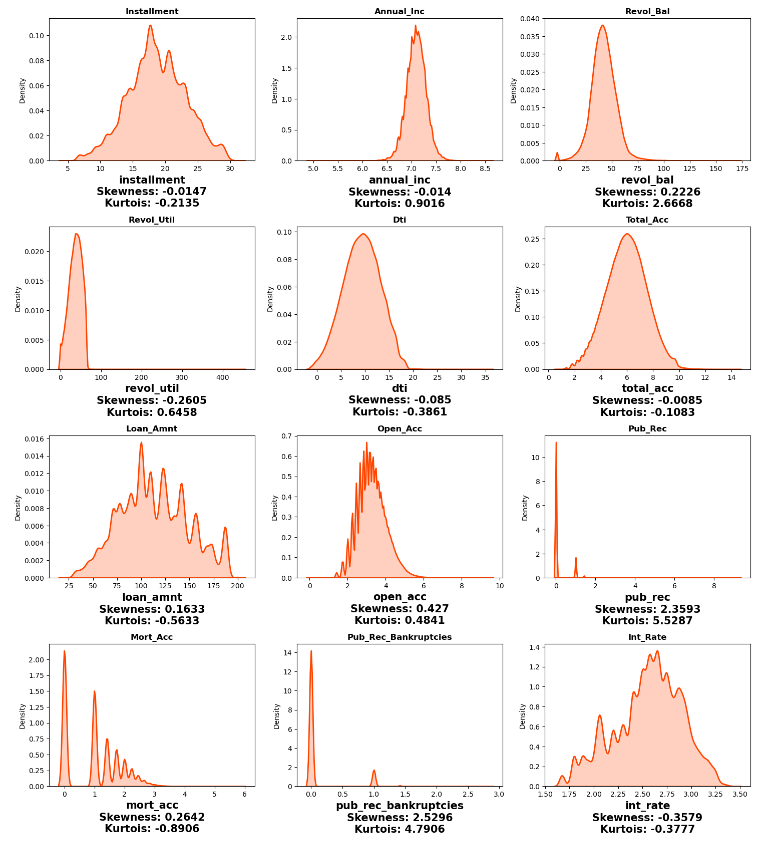
**Distribution of Variables**

****

**Here by visualizing distribution plot we can see that the Features in plotted in `light red color` are `positively skewed` and Features plotted in dark red colour are Negatively Skewed.**

**To reduce the impact of skewness we can use various transformation technique’s here we are using box cox transformation, Square root transformation and log transformation.**

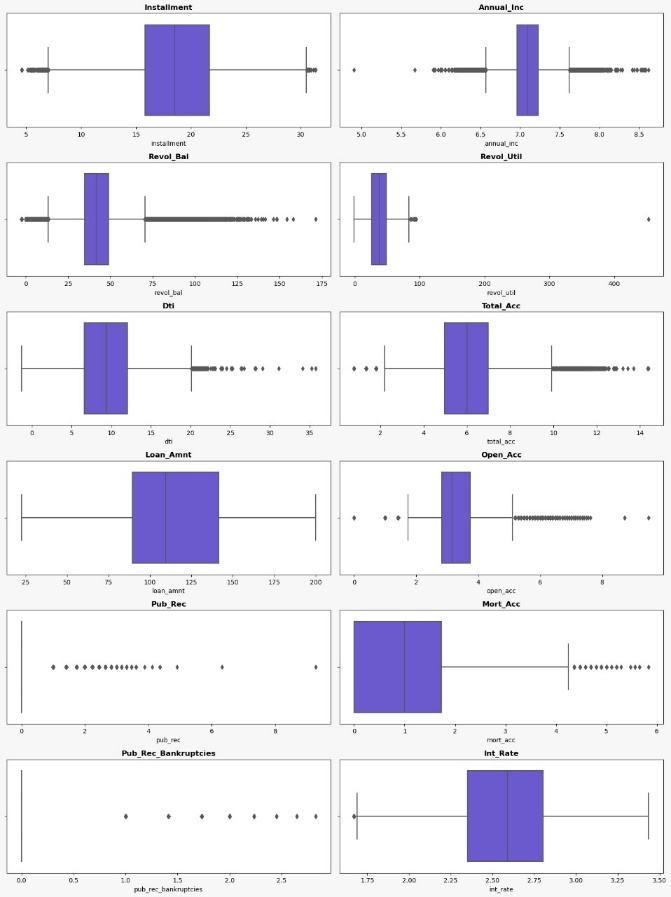
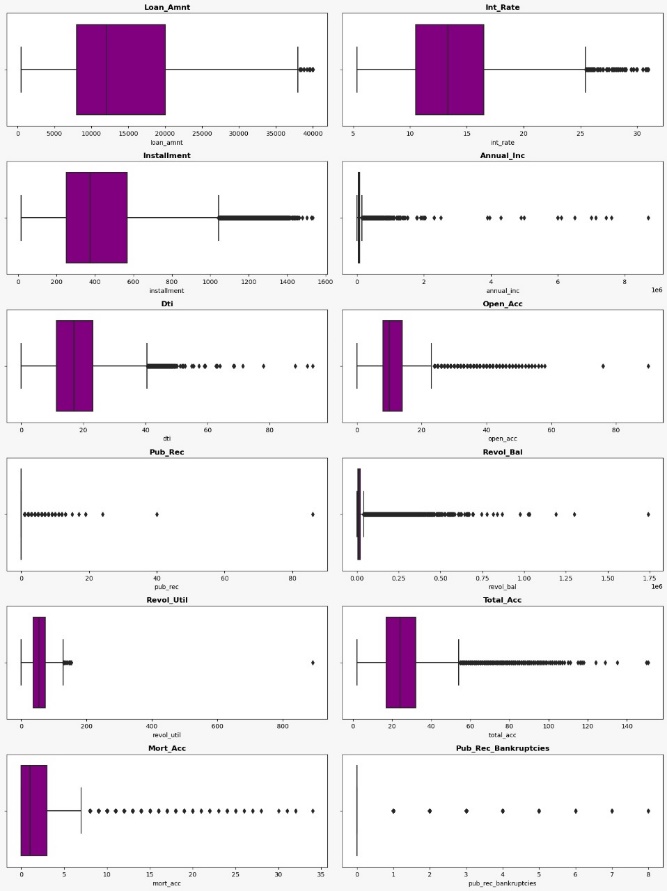
**Skewness after Transformation**

**  
  
  
  
Inference:** Here by visualizing distribution plot we can observe that there is a reduction of skewness after Transformation.

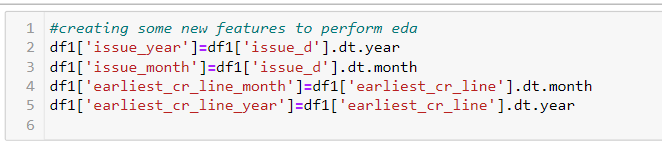
**Dealing with Extreme values**

• Multiple Extreme values detected in numerical features using box plot   
• Same was taken care of by using Box-cox transformation, Log transformation &  
 Square-root transformation.   
• Observed using box plot, not all Extreme values are removed.   
• But we managed to get them reduced to significant extent.

**Before Treating Extreme value After Treating Extreme values**

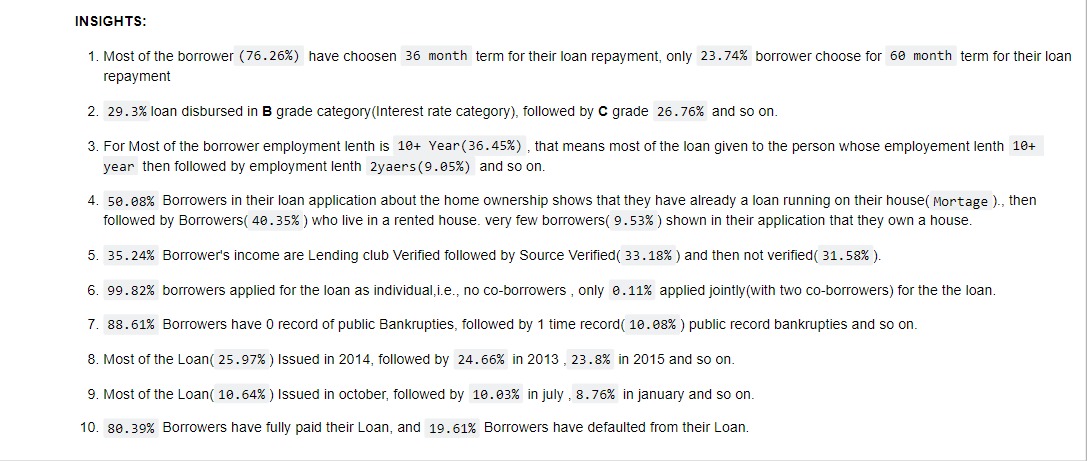
** **

**Feature Engineering**

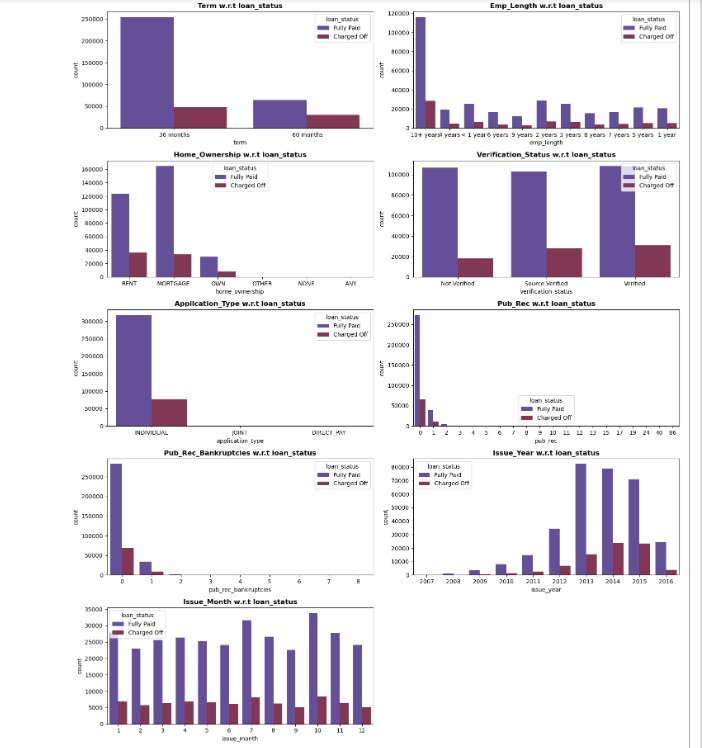
****

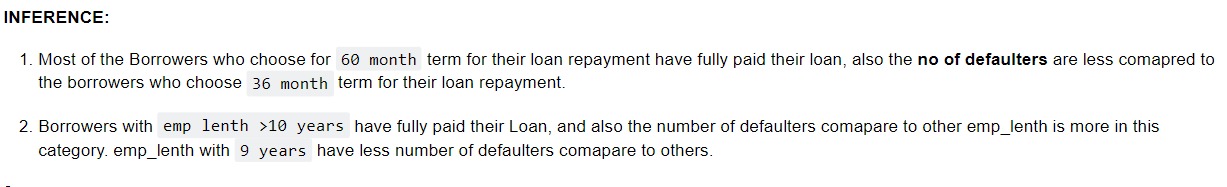
**Univariant Analysis**

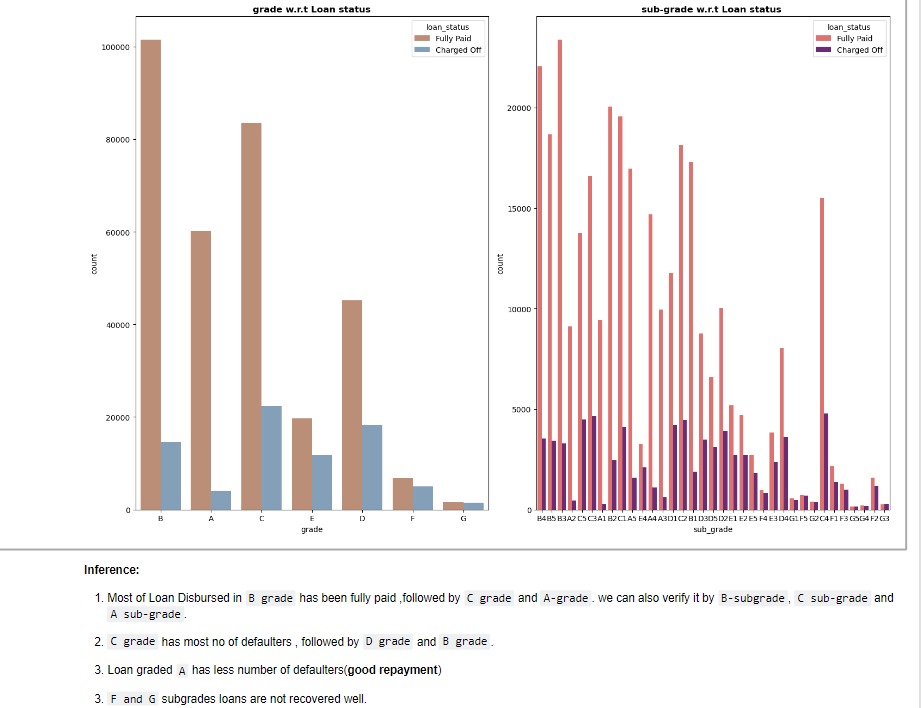
****

****

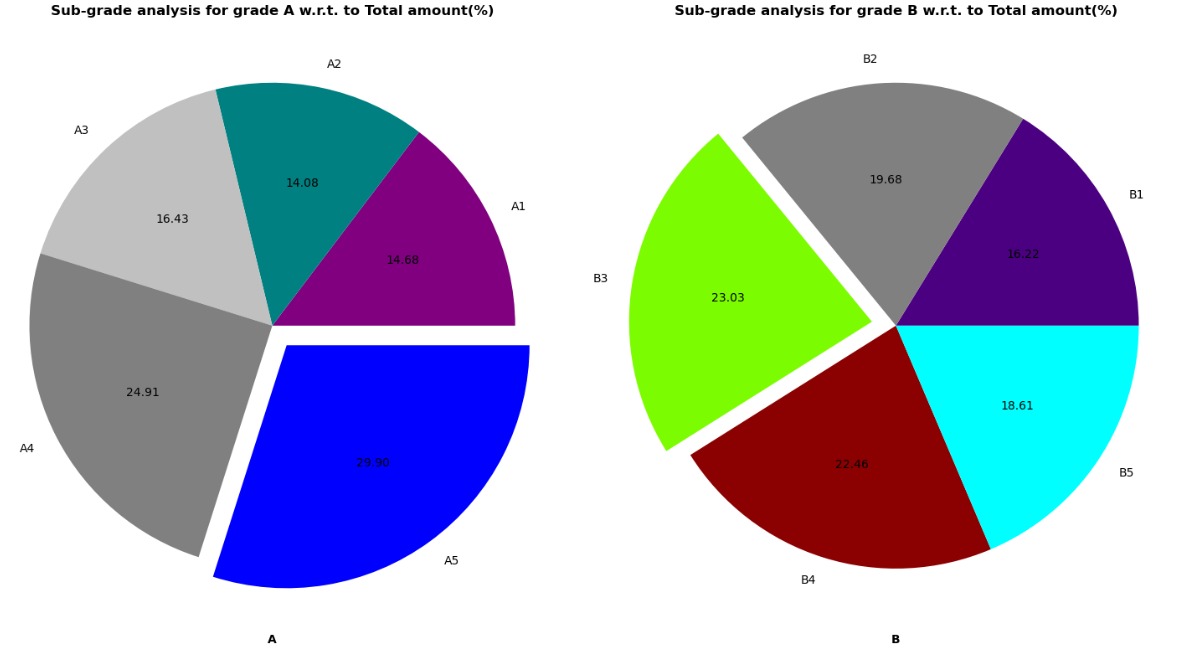
**Bivariate Analysis**

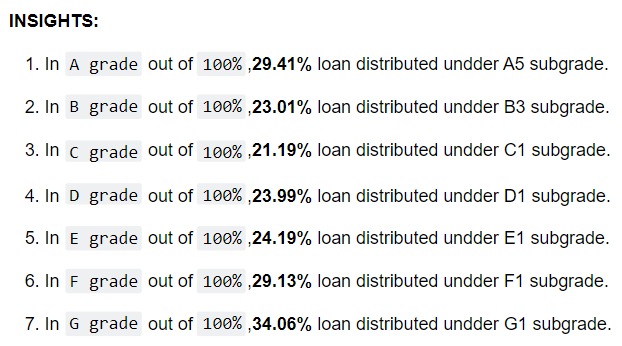
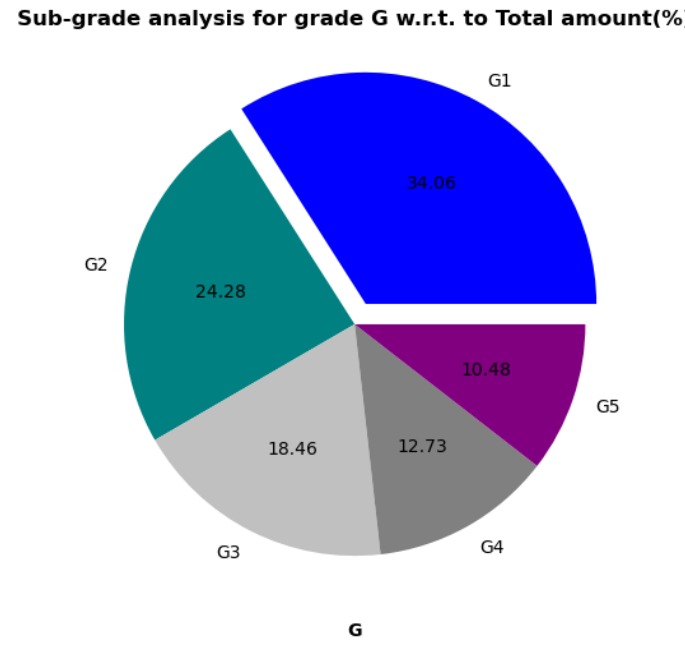
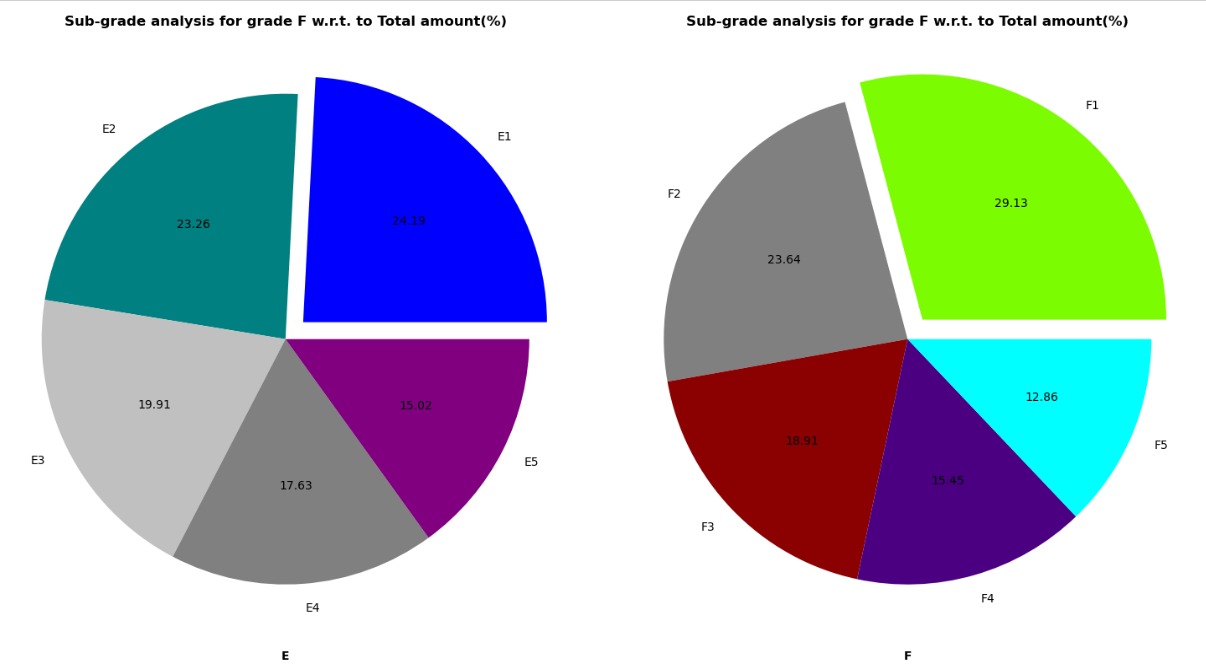
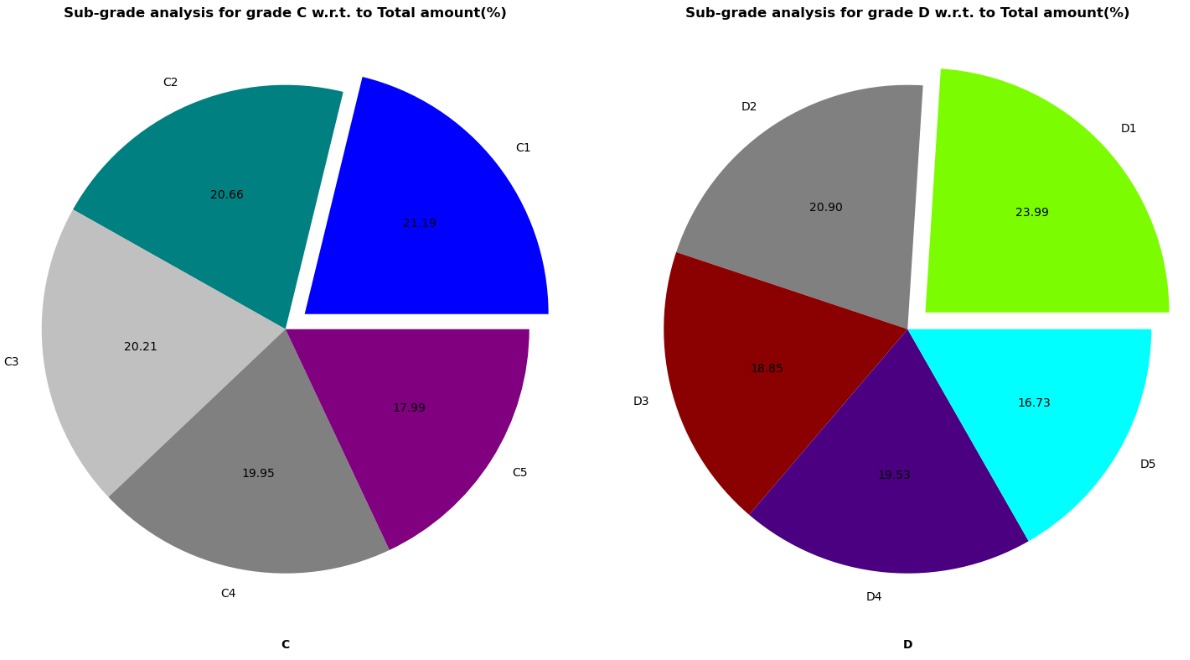
****

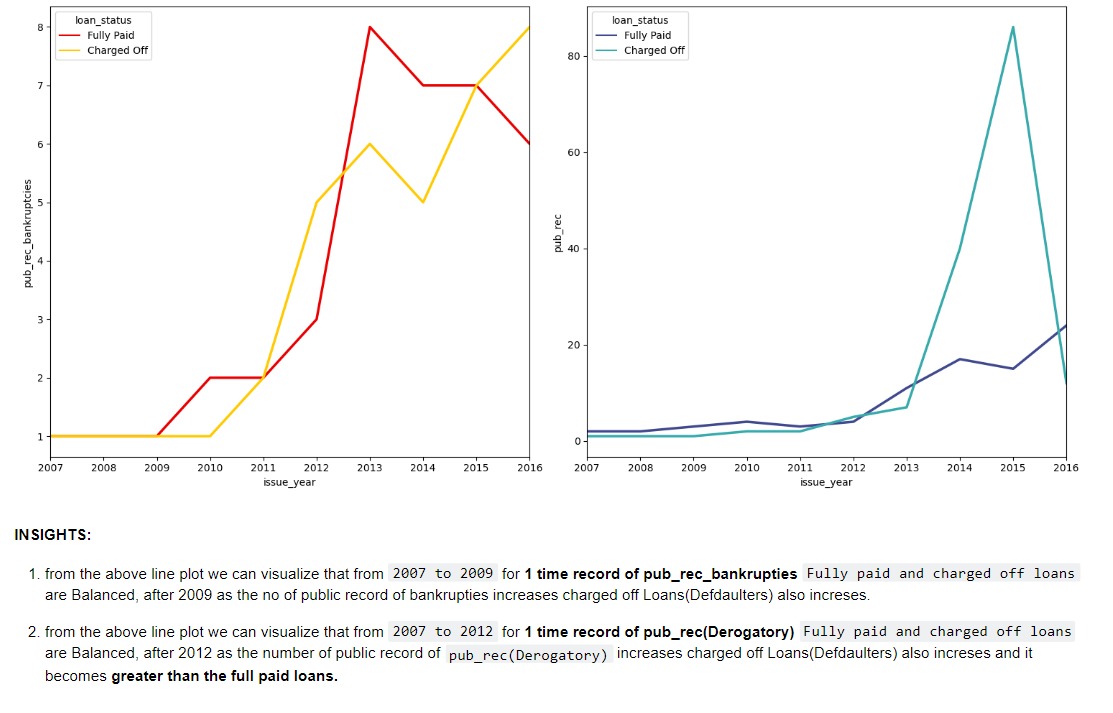
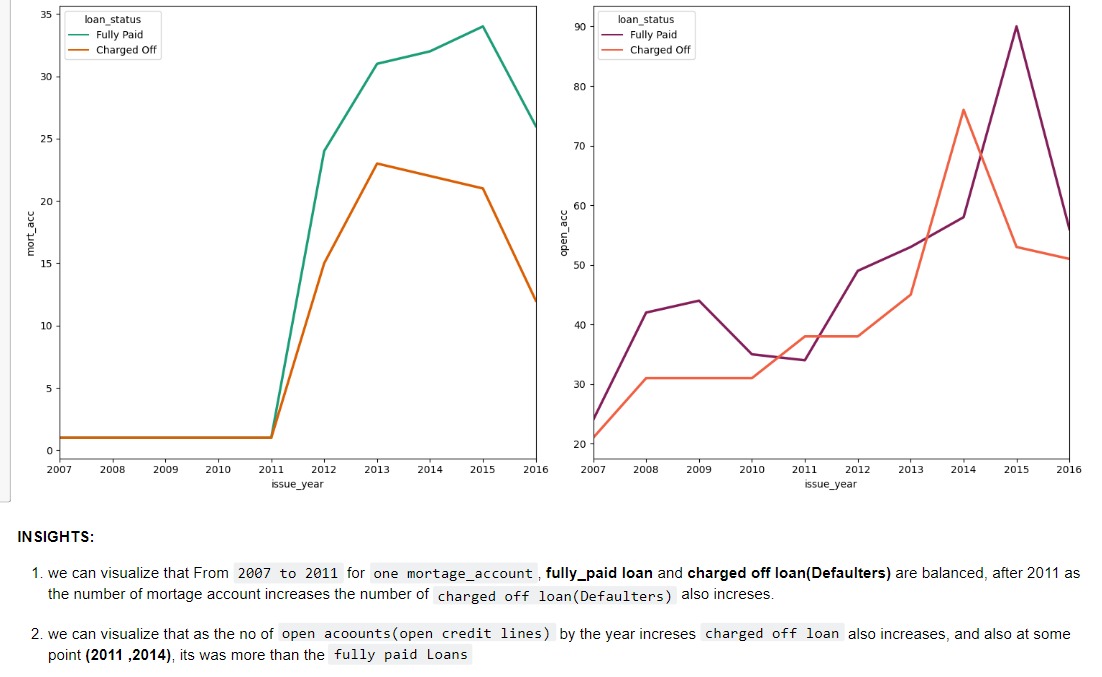
****

****

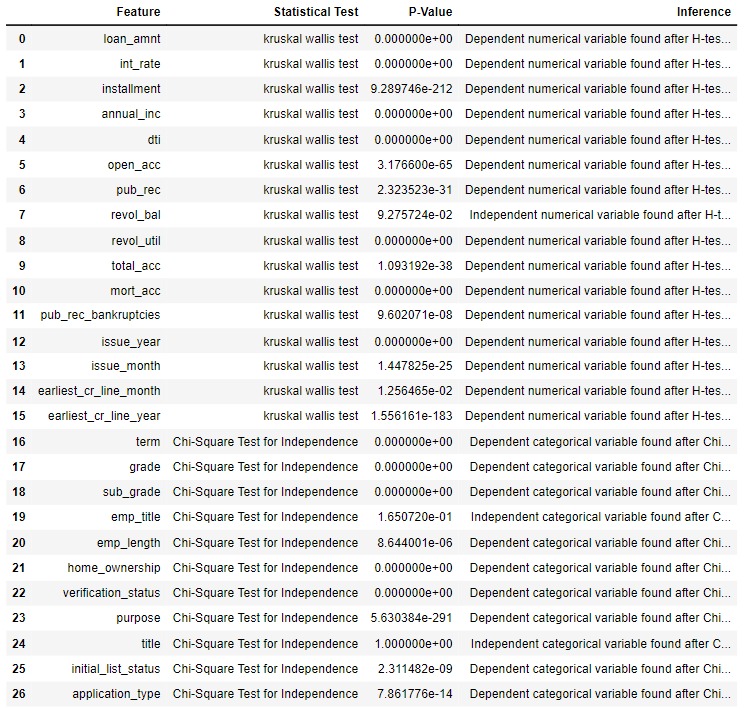
**Multi-Variate Analysis**

****

****

****

**Statistical significance of variables:**

****

We have used Chi-Square Test for Independence to test whether the categorical variables are independent or not.

**𝐻0 : The variables are independent**

**𝐻1: The variables are not independent (i.e. variables are dependent)**

We have used **Jarque bera** test to check the normality of data  
**𝐻0 : The data is normally distributed.**

**𝐻1: The data is not normally distributed.**

**We found that data is not normal therefore we use kruskal wallis test to check its dependence on the target variable**

The null and alternative hypothesis for Kruskal Wallis test is given as:

**𝐻0 : The data samples are with equal median (independent).**

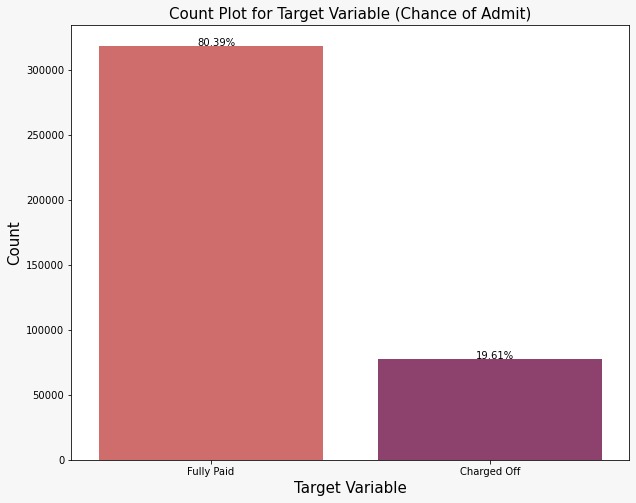
**𝐻1 : The data samples do not have equal median(i.e. variables are dependent).**

**Transformation**

We have used the Box-cox transformation, Log transformation & Square-root transformation technique. to reduce the outliers and make the data more normally distributed



**Class Imbalance and its Treatment:**

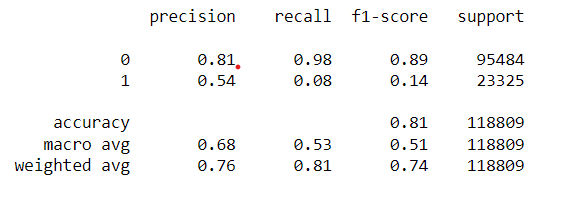
**Here we can see that our target variable is imbalance, and to treat that we are going to use various oversampling techniques like smote.**

**Feature selection:** Feature selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in. Having irrelevant features in your data can decrease the accuracy of the models and make your model learn based on irrelevant feature.

**PCA:-** Here we checked for multicollinearity, and found that there is lots of features have multicollinearity among them. To dealt with multicollinearity we have to go with PCA (Principal Component Analysis). PCA is used when we have features which are multicollinearity effect and we are not sure that which feature we should drop. It is a type of transformation where it combines the existing predictors in a way only to keep the most informative. PCA cuts the number of interdependent variables to a smaller set of uncorrelated components. Instead of using highly correlated variable we will use components in the model that have eigenvalue greater than 1.

**Future Scope :** - As from problem statement and Dataset it is cleared that it is an Unsupervised Learning, where we are not going to predict that either flight is going to be delayed or not. Or we are not going to explore out that what is the factor or feature that are responsible for delay. This dataset gives information about historical data, means it only talks about the previous events. So here we are collecting the insight of inferences from past data that what was the reason for delay, is any specific Airline which are getting more delayed, if it is delayed then is there specific day or months.

**Logistic Regression (Base Model)**

****

* **where 1 represents charged off and 0 represents fully paid.**
* **Accuracy score for the base model is 81 percent**

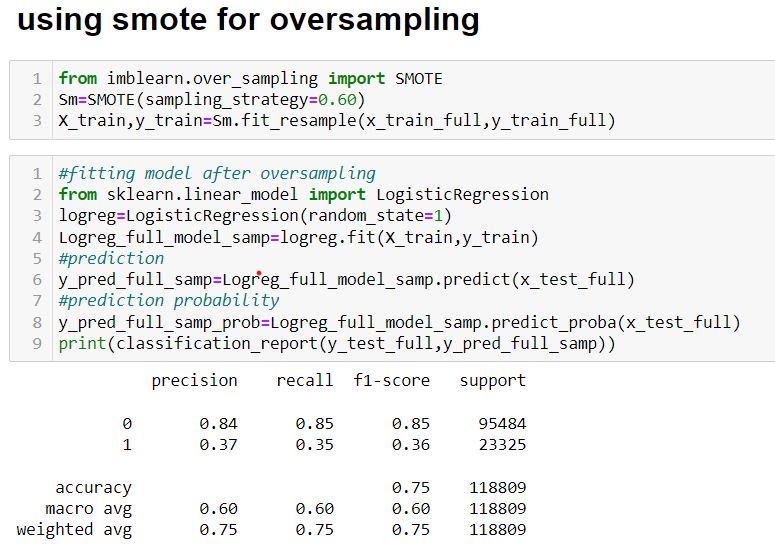
**Inference**: Here by looking at classification report, the precision for predicting 1 i.e defaulted is 0.54 , which is low so based on our business problem we have to increase both the precision and the Recall score, i.e we Have to reduce FN.

**Assumptions.**

linearity in the logit for continuous variables  
absence of multicollinearity  
lack of strongly influential outliers

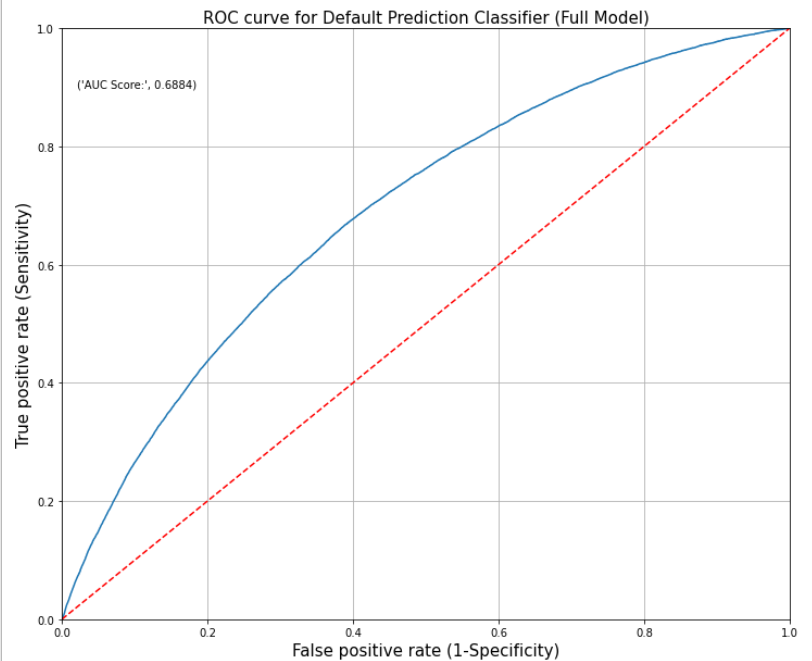
**Oversampling:**

As Our Target class is imbalance so, we are using oversampling technique reduce the class imbalance.

****

**Confusion matrix ROC-Curve:**

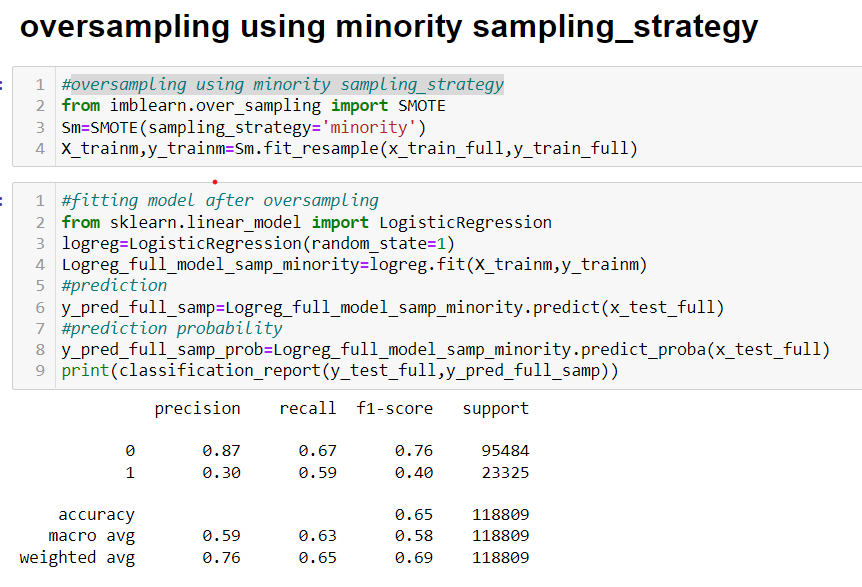
**Chart, treemap chart

Description automatically generated**

**Inference:**

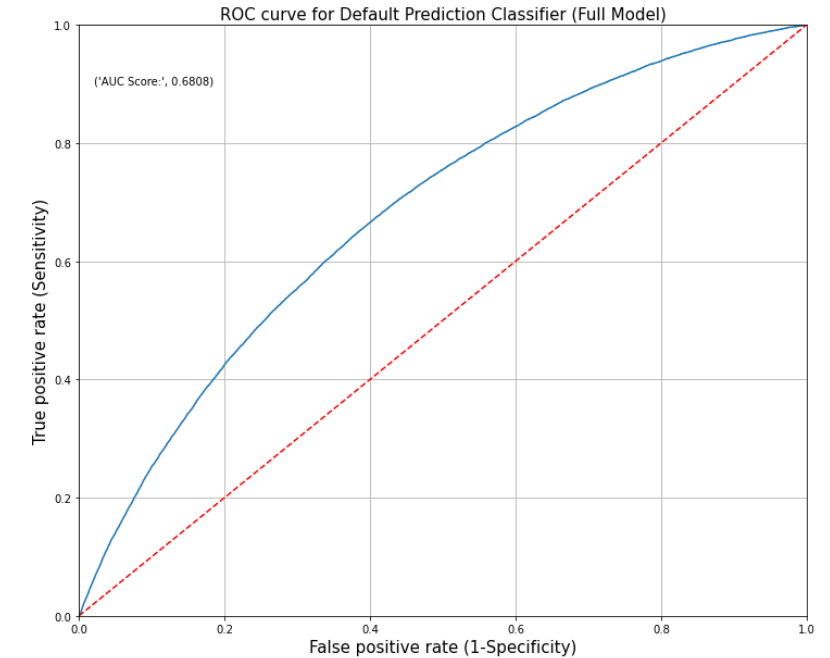
Our main goal is to reduce the FN as per our business requirement, and after oversampling it get reduced but not at a good extent.

* The red dotted line represents the ROC curve of a pure random classifier; a good classifier stays as far away from that line as possible (towards top-left corner)
* From the above plot, we can see that our classifier (logistic regression) is away from the dotted line; with the AUC score 0.6884

****

**Confusion matrix: ROC-Curve:**

**Treemap chart

Description automatically generated**

**Inference:**

Here we can see that FN has been reduced at a certain level when we used smote oversampling using minority strategy.

**Score Card for Logistic Regression:**

**Table

Description automatically generated with low confidence**

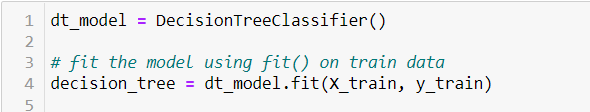
**INTERPRETATION:**

 if we compare all these models then the Logreg\_full\_model\_samp\_minority can be considered as a good model because all the model parameters like accuracy, Kappa score,'f1-score is high for this model and also precision, and, recall are more balanced for this model. but overall model is not doing great in predicting defaulters.

**Conclusion:**

* Logreg\_full\_model\_samp\_minority model performance is good, but it can be improved by passing different values and obtain a cut-off based on the passed values (Youden's Index, Cost-based Method).
  + Hyperparameter tuning can lead to more optimized model.
  + We have to implement various classiﬁcation machine learning algorithms and take feedback for them. This will take us to the best suited model

**Decision Tree Model**

****

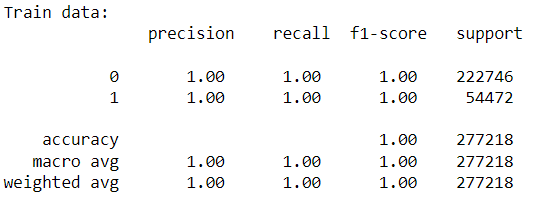
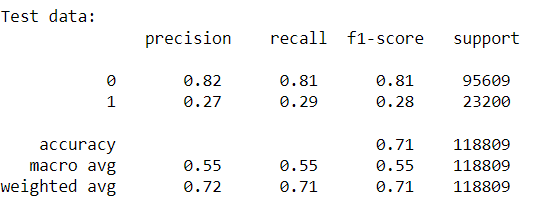
**Model performance:**

**1.Confusion matrix**

Chart, treemap chart

Description automatically generated

**Report:**

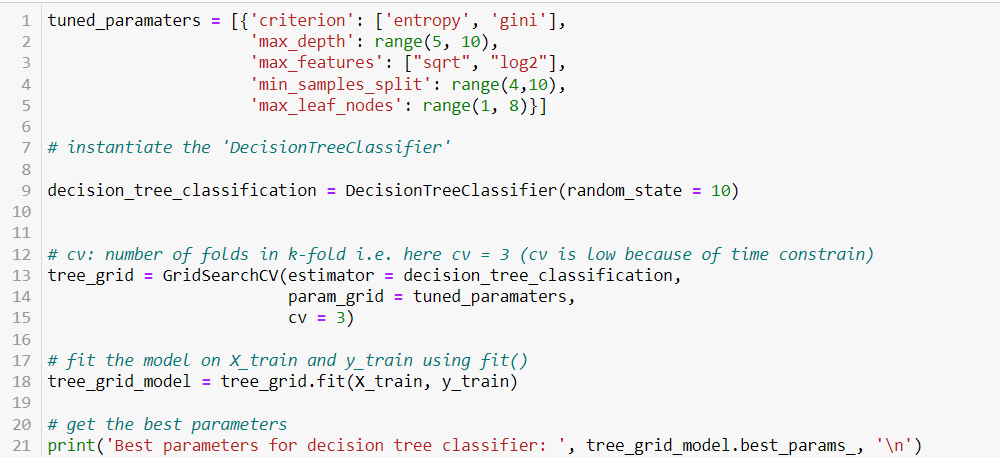
**** ****

**Inference: -**

* From The above model, our train accuracy is 1 and test accuracy is 0.71, result is Overfitting
* As the model is over fitted, our false Negative and false Positive is inaccurate
* In the next step we tuned the Hyperparameters and rebuild the model.

building a model using the best parameters

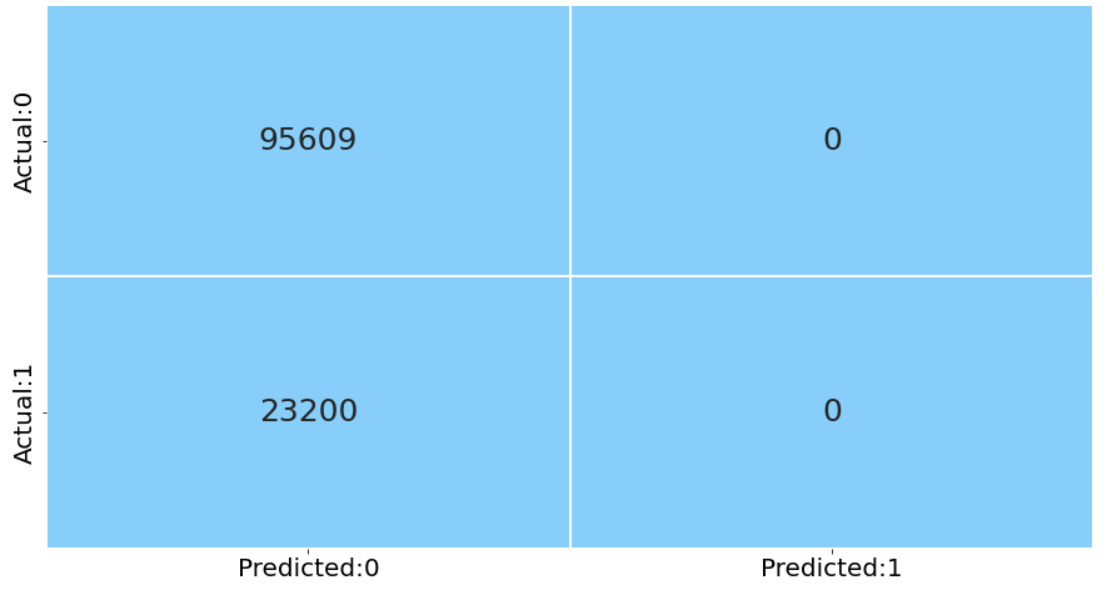
(Grid Search CV)



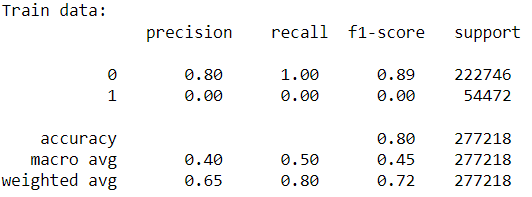
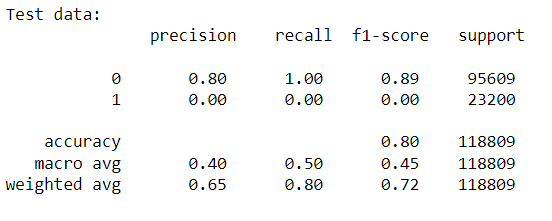


**Model Performance after Tunning:**

**Confusion matrix**

****

**Classification Report**

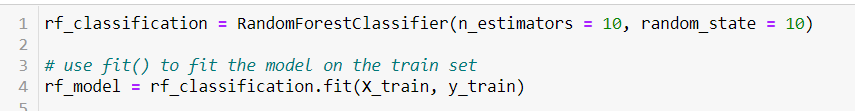
 

**Inference: -**

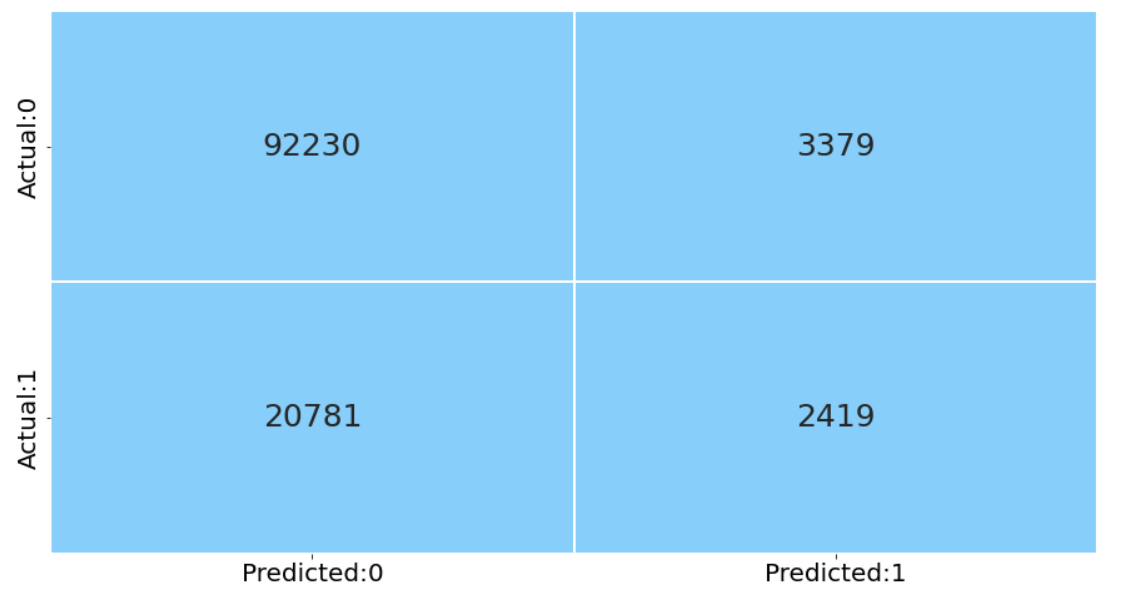
* The train and test Accuracy are comparable, which shows the reduction in overfitting.
* In this case the false negative and false positive values can be trusted and the FN value are quite High, but as our focus is on reduction of False negative values

Typically, Random Forest classifier is more accurate than a single decision tree, we rebuild the model using the same to reduce the FN and increase the accuracy

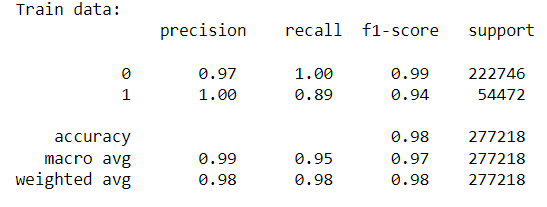
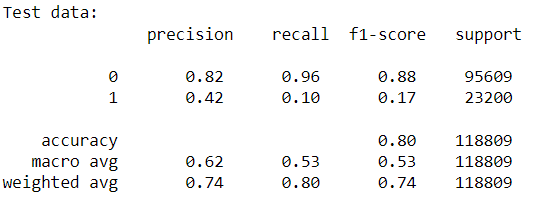
**Random Forest Classifier**

****

**Confusion matrix:**

****

**Classification Report:**

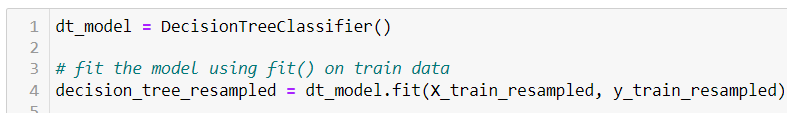
**** ****

**Inferences:**

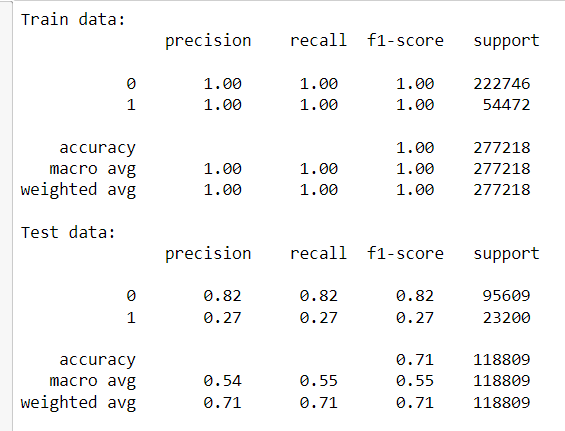
* From The above model, our train accuracy is 0.98 and test accuracy is 0.80, result is Overfitting
* As the model is over fitted, our false Negative and false Positive is inaccurate

**Building models on oversampled data**

**Decision Tree Model**

****

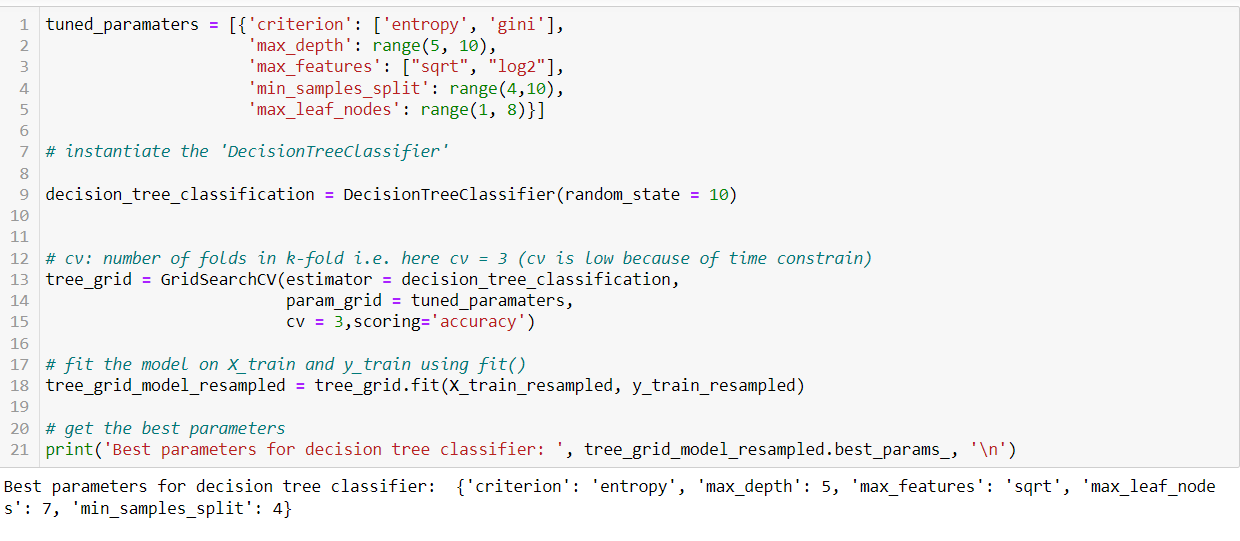
**Report;**

****

**Inferences:**

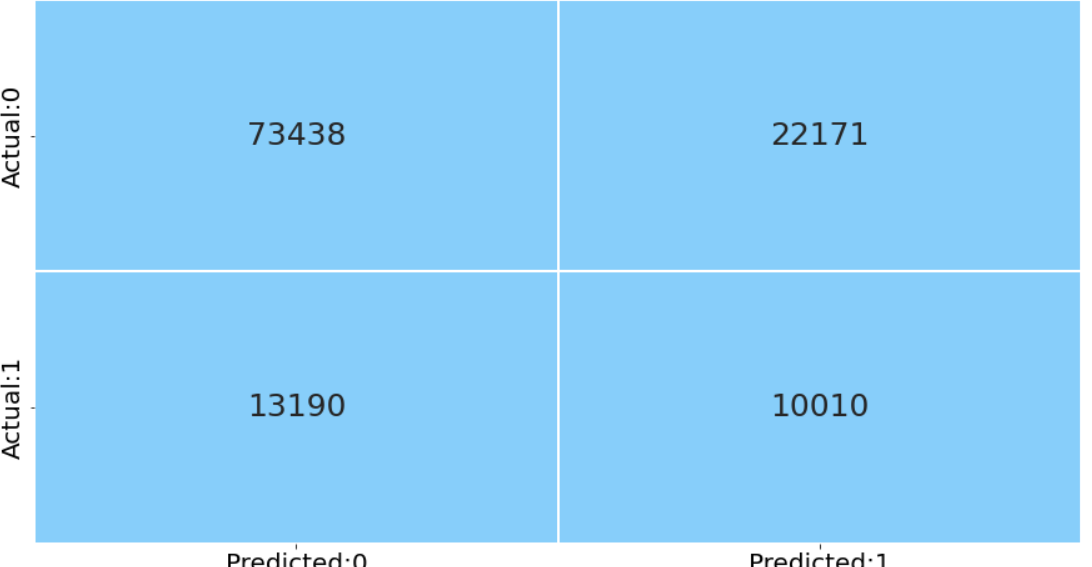
* From The above model, our train accuracy is 1 and test accuracy is 0.71, result is Overfitting
* As the model is over fitted, our false Negative and false Positive is inaccurate
* In the next step we tuned the Hyperparameters and rebuild the model.

**Tuned the Hyperparameters using GridSearchCV (Decision tree)**

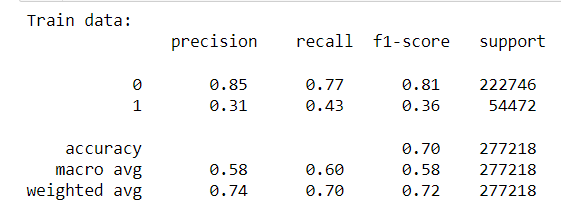
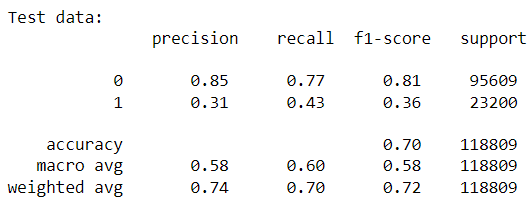
****

**Model performance after tuning;**

**Confusion matrix:**

****

**Classification report**

**** ****

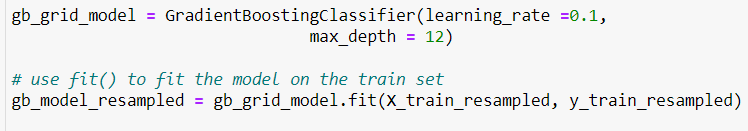
**Inference:**

The train and test Accuracy are comparable, which shows the reduction in overfitting.

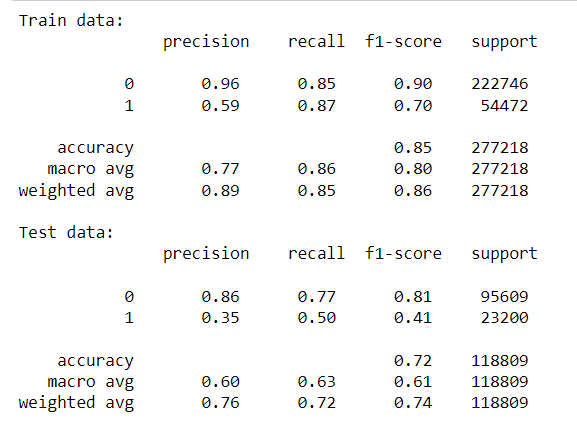
* In this case the false negative and false positive values can be trusted and the FN has reduced significantly.

**Boosting Models**

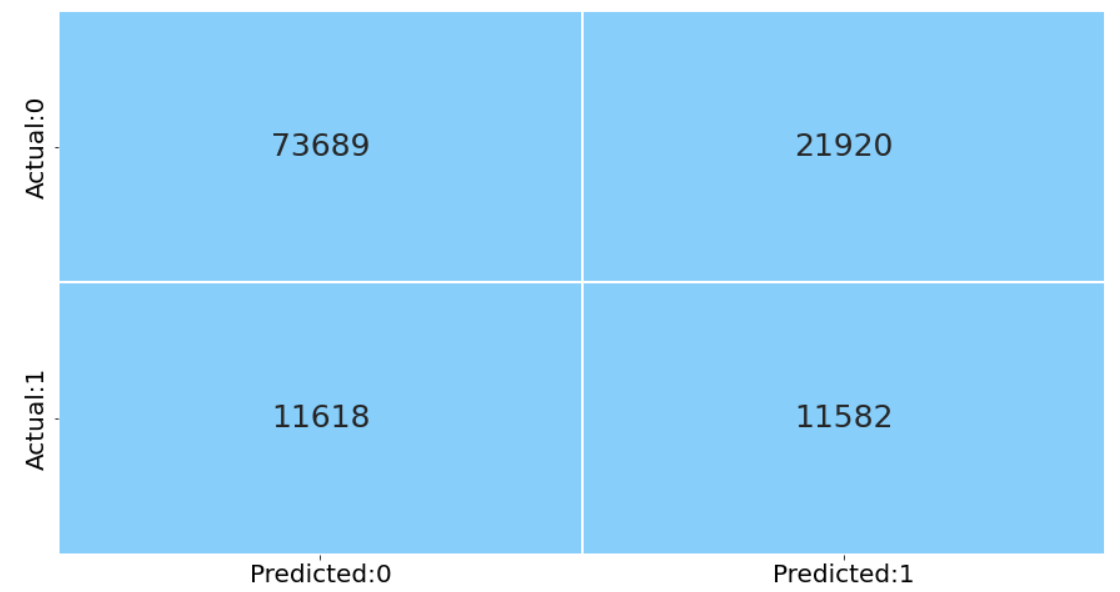
* **Gradiant Boost**

****

**Classification report:**

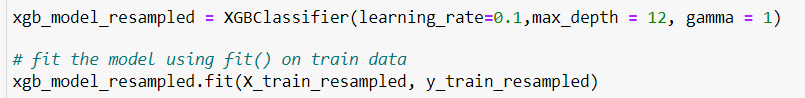
****

**Confusion Matrix:**

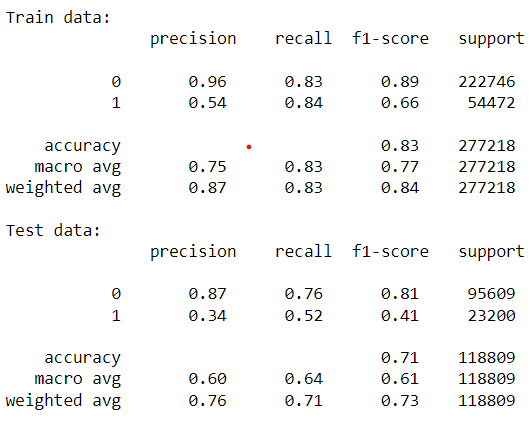
****

**Inference:**

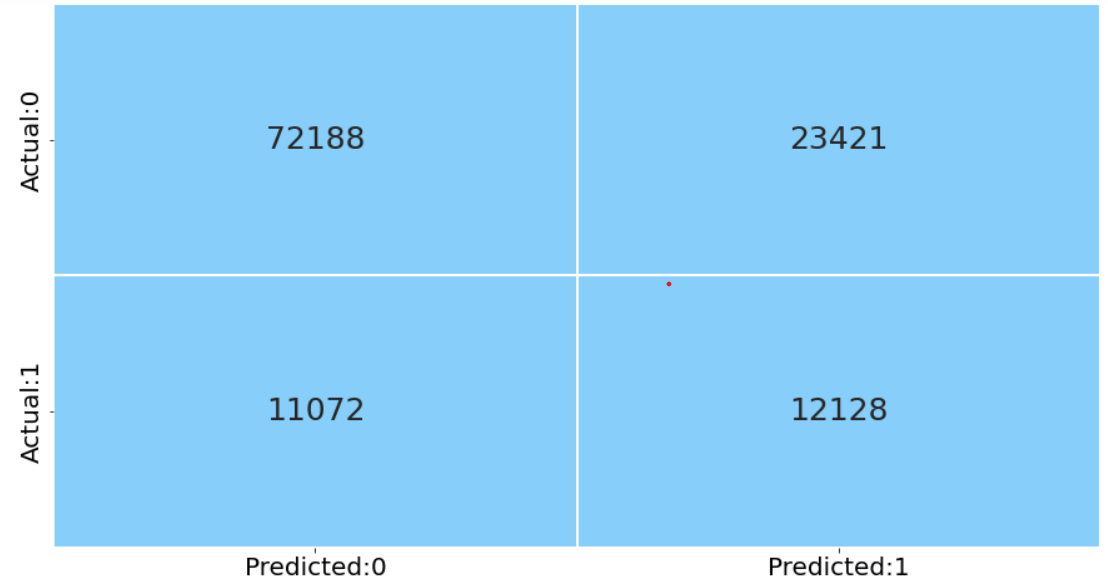
* According to the train-test report, train value shows 0.85 and test value shows 0.72 from this we can say the model is performing well but not that extent, it seems to be overfitted.
* FN has significantly reduced if we compare to the previous model.
* **XG Boost**

****

**Classification Report:**

****

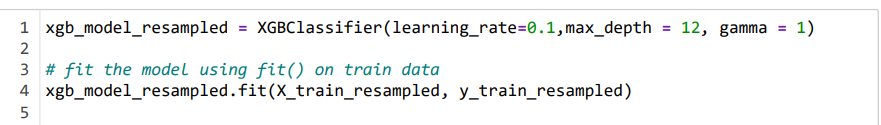
**Confusion matrix:**

****

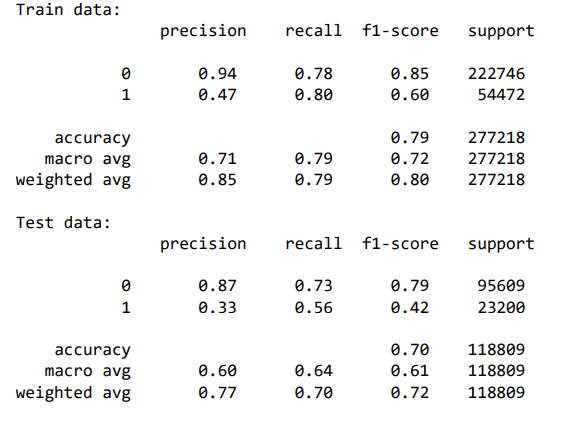
**Inference:**

* According to XGBoost model report, we can observe that model is overfitted.
* Compared to previous models, the FN has reduced.
* Lets build another XGBoost model using RFE

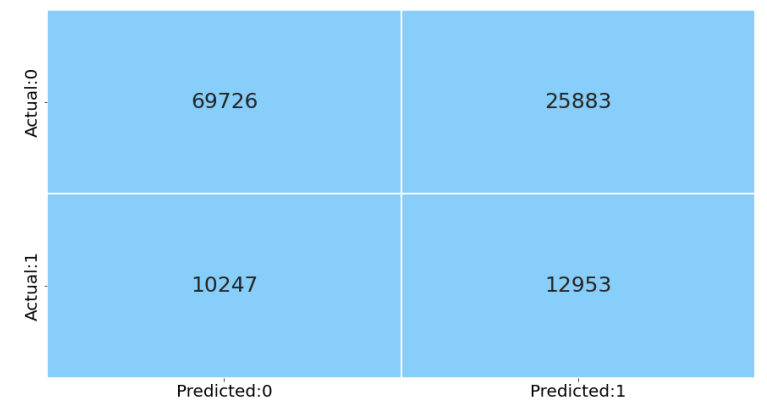
**XGBoost model using RFE**



**Classification Report:**

****

**Confusion Matrix:**

****

**Inference:**

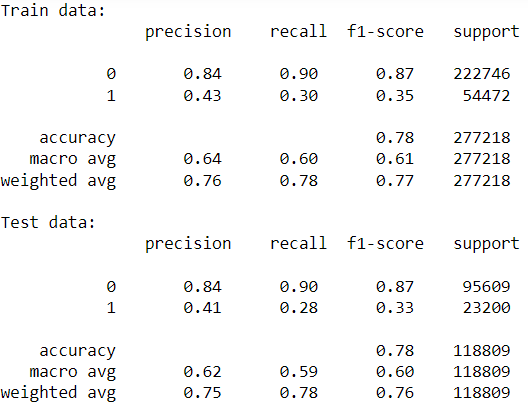
* According to XGBoost model report, we can observe that train-test values are good.
* Compared to previous models XGBoost model, the errors have been reduced.
* Comparatively the recall value has been boosted from 7% to 56%
* Compare to base Decision model, the overfitting is reduced and FN errors are reduced by

nearly 50%

**Stack Generalization:**

****

**Classification Report:**

****

**Inference:**

* According to stack Generalization report, we can observe that train test values are good
* Compare to base logistic model, the overfitting is reduced and FN errors are reduced by

But not as good as XGBoost model

**Business Insights:**

* XGBoost model is performing well in predicting defaulters based on the Historical Data. Company can deploy this model for the prediction to check which borrower can be possible defaulters in future and they will be able to reduce the number of defaulters and from that they will be able to reduce losses.
* Company can also work on some of the weak areas which we have analysed during EDA analysis like from 2007 to 2009 the fully paid and defaulters were balanced but as the number of public record bankruptcies increases the defaulters also increased.