***Predictive Modeling of Credit Risk Faced by a P2P lending platform***

**Problem Statement:** The Irish bank is a P2P lending platform, which provides funds to potential borrowers and earns a profit on the interest, which depends on the risk they take (the borrowers credit score). The platform provides loan to its loyal customers. The platform wants to understand whether the loan they have provided to their customers will be recovered or not.

Leveraging a dataset inspired by Lending Club but modified for educational purposes, this study aims to analyze the credit risk associated with loans provided by the bank and predict which loan shall go bad.

**Motivation:** The study aims at solving problem for the bank where it can provide loans to its customers safely i.e. with minimum chances of default. The problem can be solved by creating predictive models using Machine learning models which depend on inputs provided by the customer during the KYC. This will help the platform to differentiate between a good customer.

A study conducted by –

Clive Jude Ronald

Hitansh Lakkad

Pranav Nigam

Priya Singh

Source:[**https://www.kaggle.com/datasets/mrferozi/loan-data-for-dummy-bank**](https://www.kaggle.com/datasets/mrferozi/loan-data-for-dummy-bank)

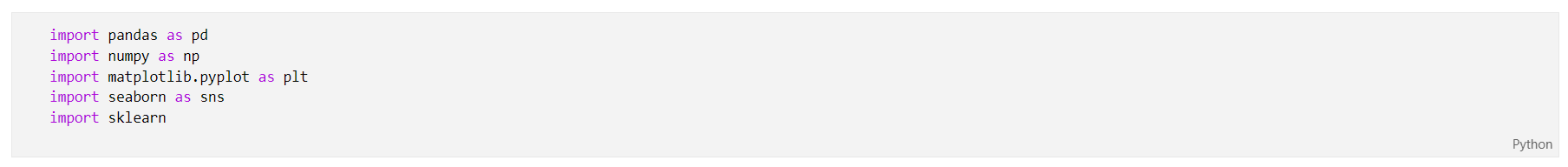
# ***Process Followed***

# ***Data Description***

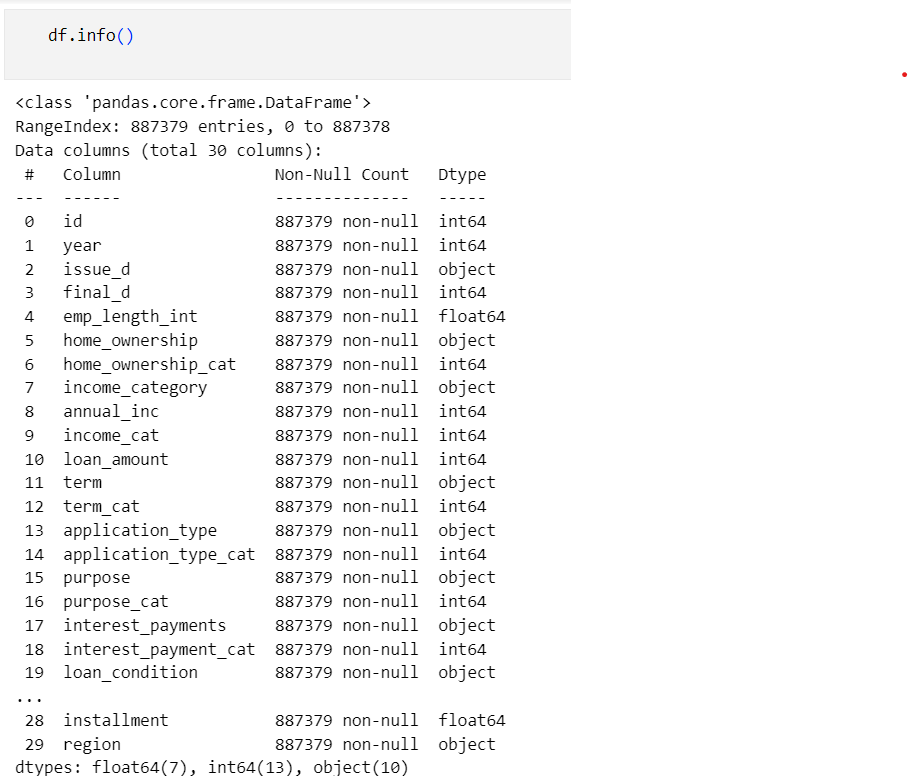
|  |  |
| --- | --- |
| Variable | Description |
| id | Customer ID |
| year | Year of availing the loan |
| issue\_d | Issue Date |
| final\_d | Date of Last payment |
| emp\_length\_int | Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years. |
| home\_ownership | The home ownership status provided by the borrower during registration. Our values are: rent, own, mortgage, other. |
| income\_category | Categorized Income (Low, Medium, High) |
| annual\_inc | The self-reported annual income provided by the borrower during registration. |
| loan\_amount | 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. |
| term | The number of payments on the loan. Values are in months and can be either 36 or 60. |
| application\_type | Indicates whether the loan is an individual application or a joint application with two co-borrowers |
| purpose | A category provided by the borrower for the loan request. |
| interest\_payments | Interest payments made by the customers (Actual Amounts) |
| loan\_condition | **Condition of the Loan [TARGET] (Good Loan = 0 , Bad Loan = 1)** |
| interest\_rate | Interest Rate on the loan |
| grade | LC assigned loan grade |
| 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. |
| total\_pymnt | Payments received to date for total amount funded |
| total\_rec\_prncp | Principal received to date |
| recoveries | post charge off gross recovery |
| installment | The monthly payment owed by the borrower if the loan originates. |
| region | region of Loan being executed |

# EDA and Preprocessing

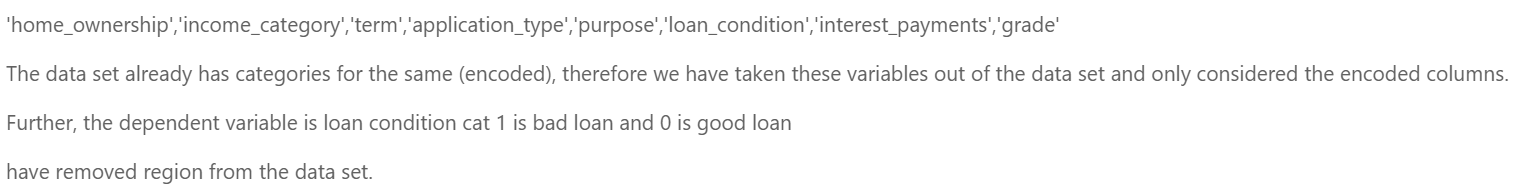
## Libraries Imported:



## About the data set:



There are 29 variables and 887379 observations. There are 10 object variables,13 int variables,7 float variables.

It is to be noted that the dataset itself has provided coding for categorical variables which has increased the number of columns. The number of columns / variables is similar to the one mentioned in Data Description.

### **Checking Null values and Duplicates**

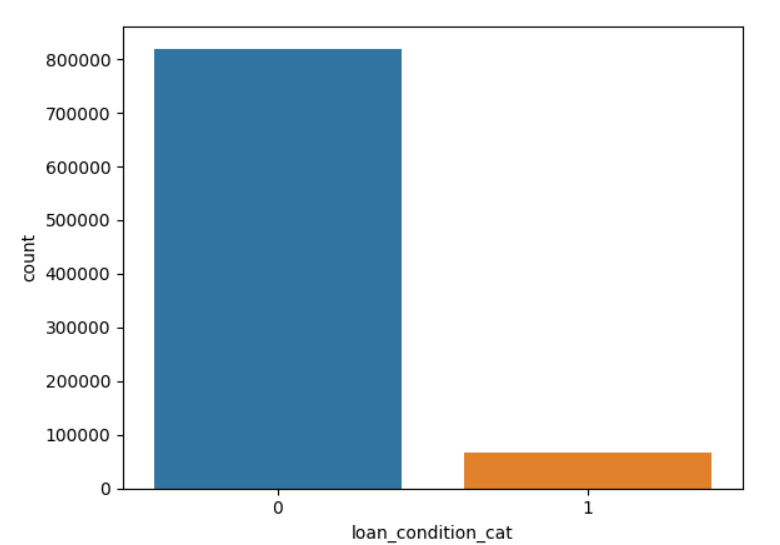
* The data did not have any null values.
* The data set also did not have any duplicates.

## EDA on Categorical Variables – Bivariate Analysis

The Bivariate analysis on categorical variables consists of the relationship between the categorical variables and the loan being good or bad. It is to be understood that as mentioned above, loan\_condition\_cat is the target / dependent variable we are more concerned with identification of the bad loans.

But first we have to identify the occurrences of bad loans, which is provided as follows:

|  |  |  |
| --- | --- | --- |
| Loan Type | Actual | Percentage |
| Good Loan | 819950 | 92.4013% |
| Bad Loan | 67429 | 7.5987% |



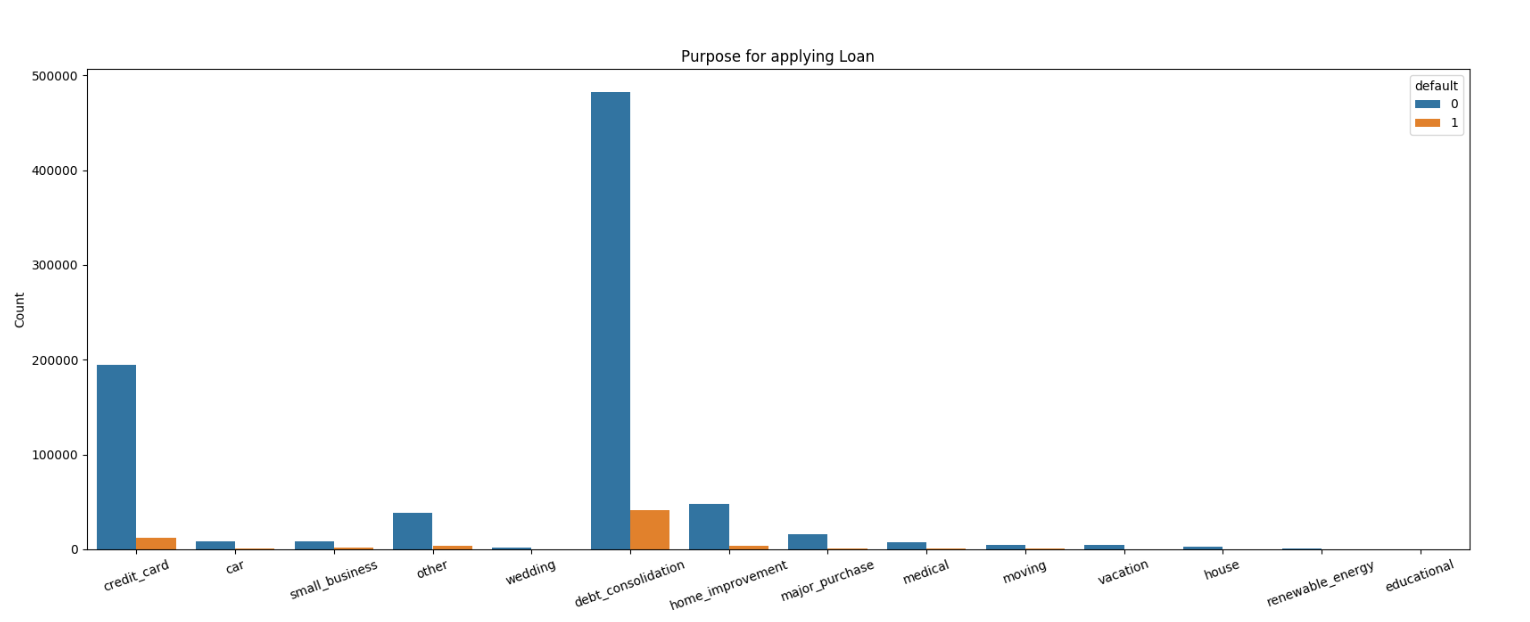
From the information above we can understand that the set is an imbalanced one where the number of good loans is quite high as compared to bad loans. Some variables have also been removed from the data set as follows:



This has been done to remove variables which may not have directly affected the target variable. Please note that the decision is based on duplication of categorical variables which have been encoded in the dataset and other variables like customer ID, Year of Issue, Region, etc.

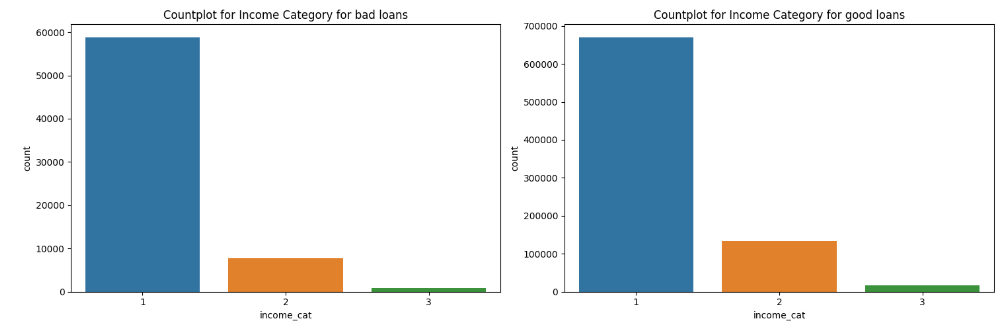
With this, let’s proceed with the Bivariate analysis.

### **Purpose for Applying Loan**



The above image suggests that most amount of loans taken is for debt consolidation where people have defaulted the most. This is then followed by loan taken for credit cards, where we have the 2nd highest count of defaults. This suggests that it is risky to provide loans to people who want to clear off other loans.

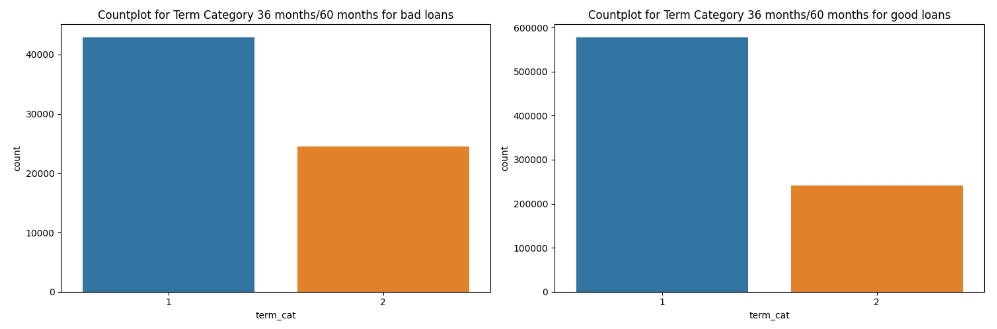
### **Income Category**



|  |  |
| --- | --- |
| Income | Encoding |
| High | 1 |
| Medium | 2 |
| Low | 3 |

The income category suggests the income for the people who have defaulted on loans. The above diagram suggests that most of the defaults have happened for people with higher income, these maybe the people who have taken loans to pay off their other loans or they may be people who are less financially educated.

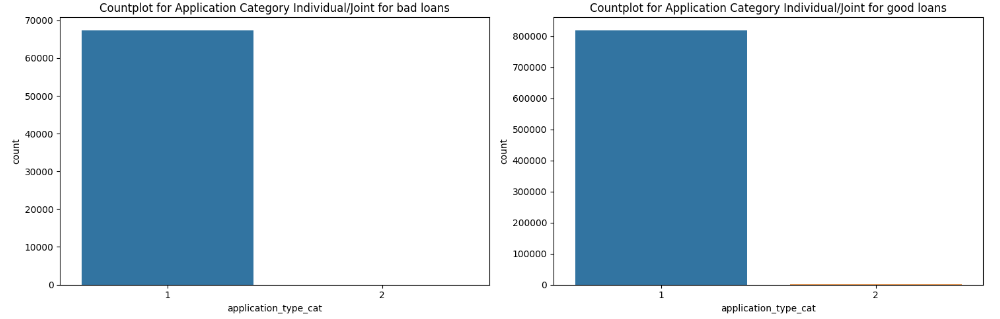
### **Term Category**



|  |  |
| --- | --- |
| Term | Encoding |
| 36 months | 1 |
| 60 months | 2 |

The above image suggests that most of the defaults have happened on shorter duration loans which range up to 36 months. This may again be attributed to the fact that most of the loans were taken to clear off other debts, which generally are taken for a shorter term. Hence, loans taken from shorter terms are more likely to be defaulted.

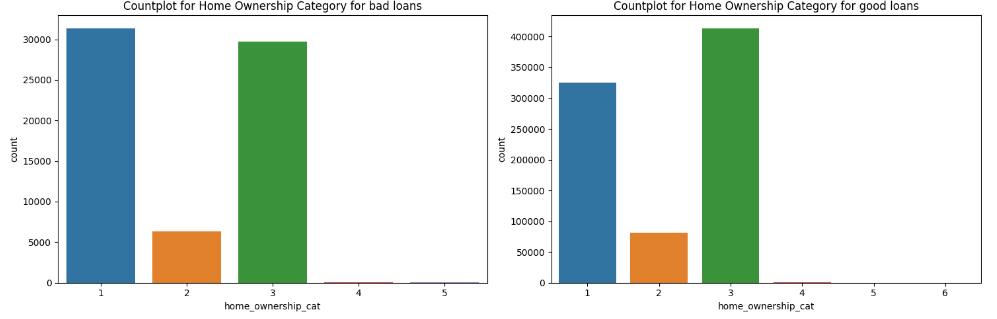
### **Application Category**



|  |  |
| --- | --- |
| Application Type | Encoding |
| Individual | 1 |
| Joint | 2 |

The above diagram suggests that most of the defaulters are individuals availing the loan.

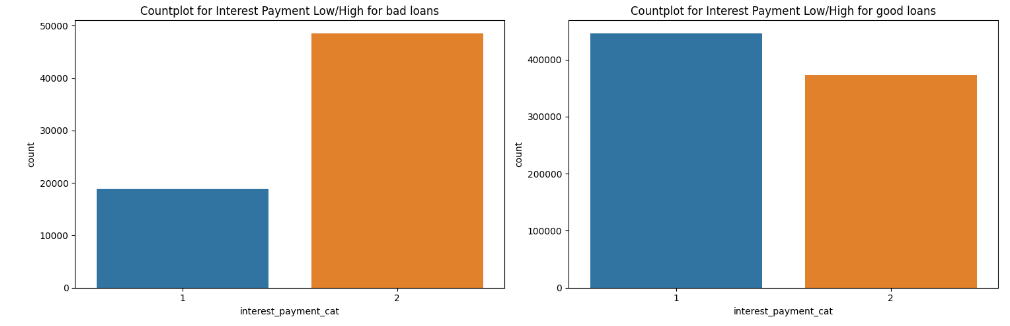
### **Home Ownership Type**



|  |  |
| --- | --- |
| Home Ownership | Encoding |
| Rent | 1 |
| Own | 2 |
| Mortgage | 3 |
| Other | 4 |
| None | 5 |
| Any | 6 |

The maximum number of defaulters are people who live in a rented abode. This is followed by people who have mortgaged properties. Both these factors lead us to believe that most of the defaulters are people who do not own houses.

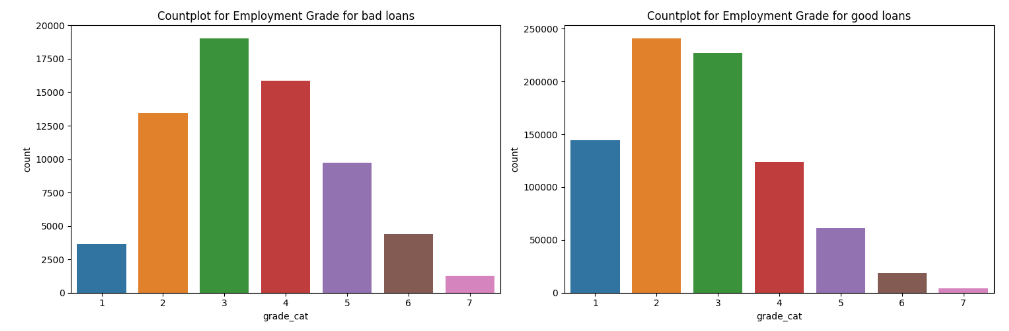
### **Interest Payment**



|  |  |
| --- | --- |
| Interest payment | Encoding |
| Low | 1 |
| High | 2 |

The above image sows that people who pay higher interest rates are more likely to default as compared to people who pay less interest.

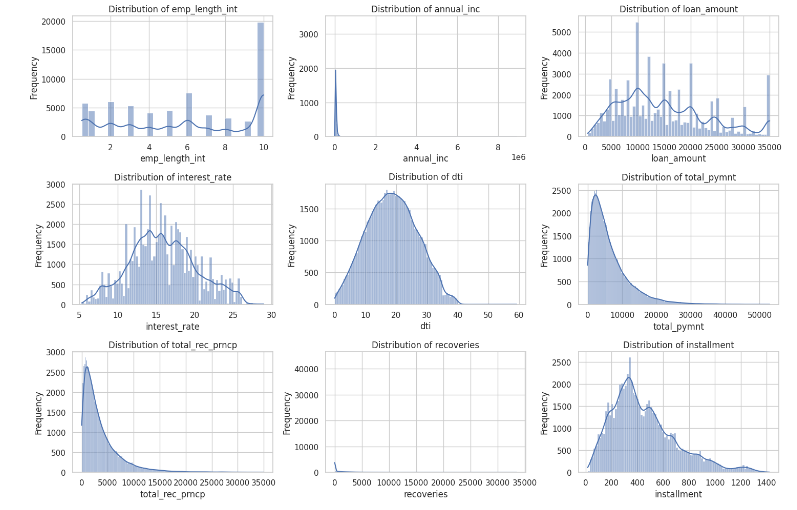
### Employment Grade



|  |  |
| --- | --- |
| Employment Grade | Encoding |
| A | 1 |
| B | 2 |
| C | 3 |
| D | 4 |
| E | 5 |
| F | 6 |
| G | 7 |

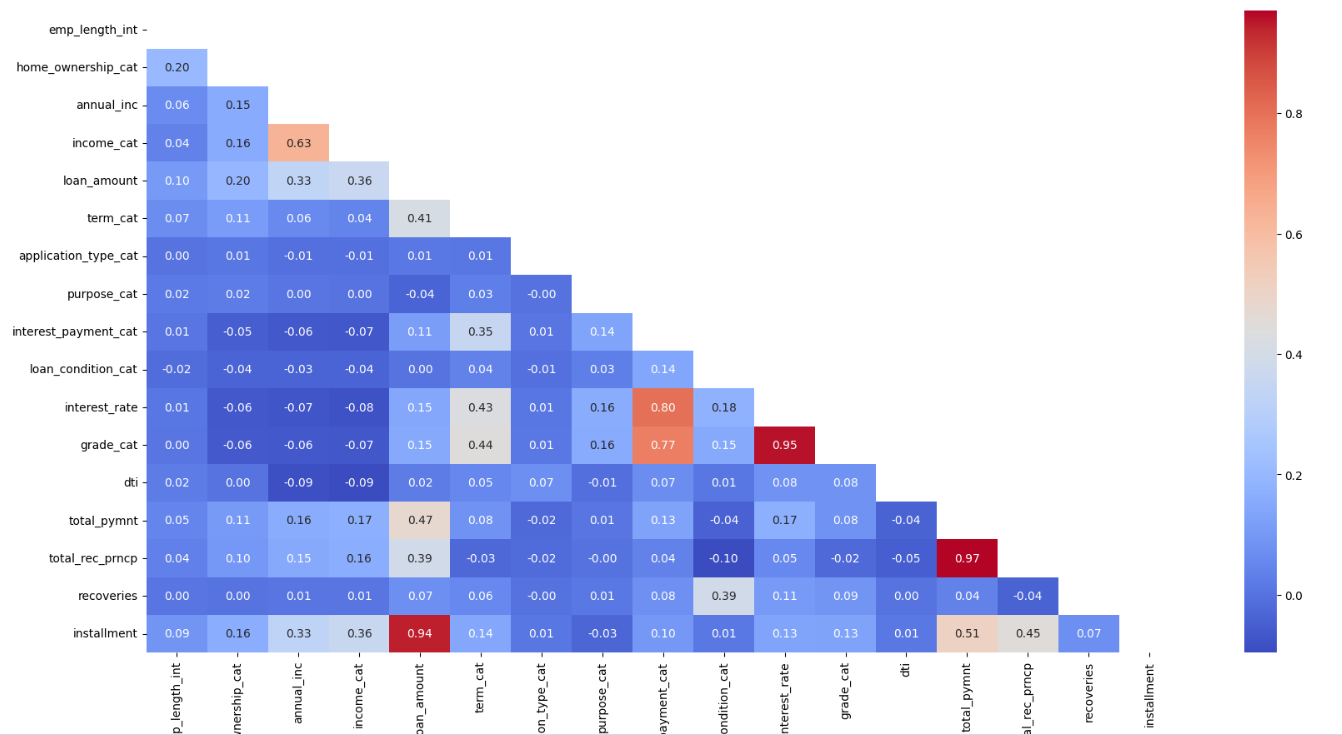
This suggests that most of the defaulters lie in grade 3 followed by grade 4 employment category, which helps us to conclude that out of the people who have defaulted, have an employment category C and below. This may also include business men or entrepreneurs who have taken loans to pay off other debts.

## Distribution of continuous variables:



|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Continuous Variable | Mean | SD | Min | 25th Perc | Median | 75th perc | Max |
| Employment Length | 6.05 | 3.5074 | 0.5 | 3 | 6.05 | 10 | 10 |
| Annual Income | 75027.59 | 64698.15 | 0 | 45000 | 65000 | 90000 | 9500000 |
| Loan Amount | 14755.26 | 8435.456 | 500 | 8000 | 13000 | 20000 | 35000 |
| Interest rate | 13.2467 | 4.381867 | 5.32 | 9.99 | 13.29 | 16.2 | 28.99 |
| Dti | 18.157 | 17.19063 | 0 | 11.91 | 17.65 | 23.95 | 9999 |
| Total Payment | 7558.83 | 7871.243 | 0 | 1914.59 | 4895 | 10616.8 | 57777.6 |
| Total Principal Amount Received | 5757.71 | 6625.441 | 0 | 1200.57 | 3215.32 | 8000 | 35000 |
| Recoveries | 45.9192 | 409.6939 | 0 | 0 | 0 | 0 | 33520.3 |
| Installment | 436.717 | 244.1866 | 15.67 | 260.705 | 382.55 | 572.6 | 1445.46 |

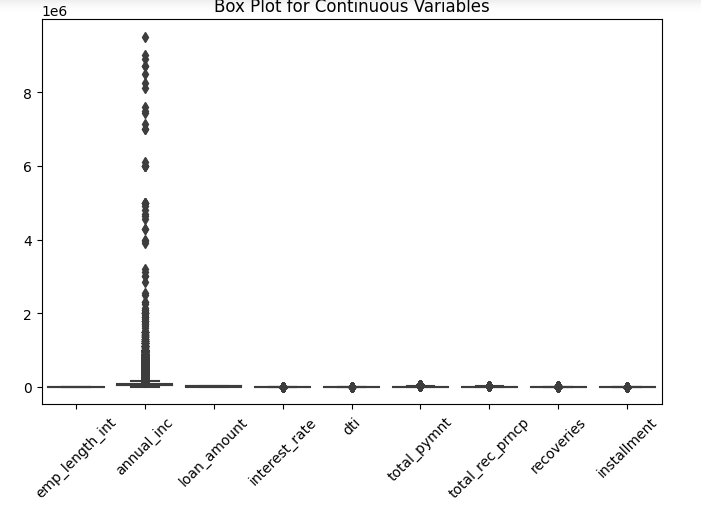
## Correlation Matrix

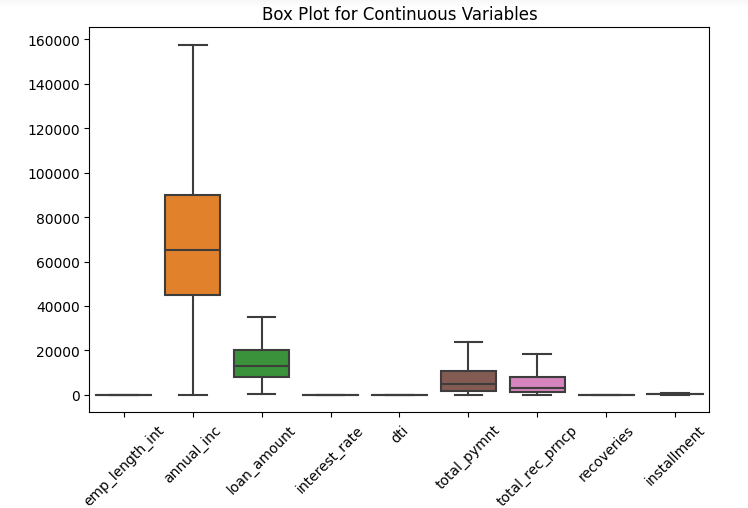


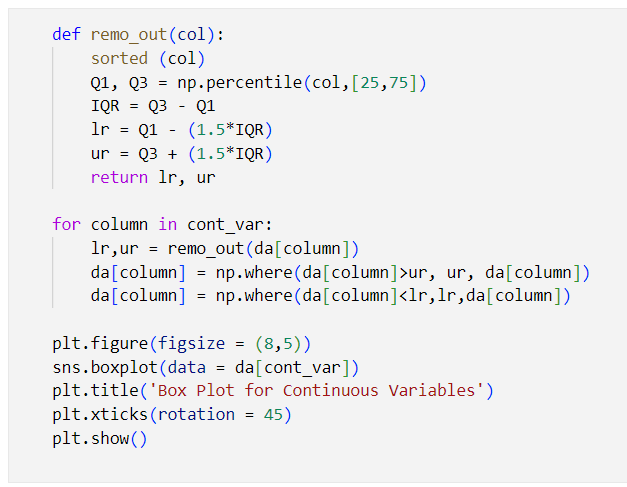
Hence, we can conclude that the target / dependent variable is not greatly affected by the independent variables.

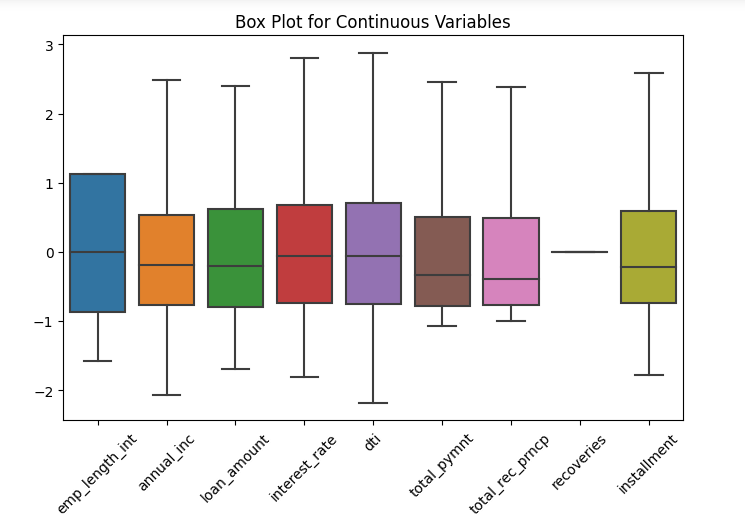
# ***Logistic Regression***

The very first step that we took was create a logistic regression model to classify the good loans and the bad loans. To this end, we first normalized the continuous variables in the following manner:



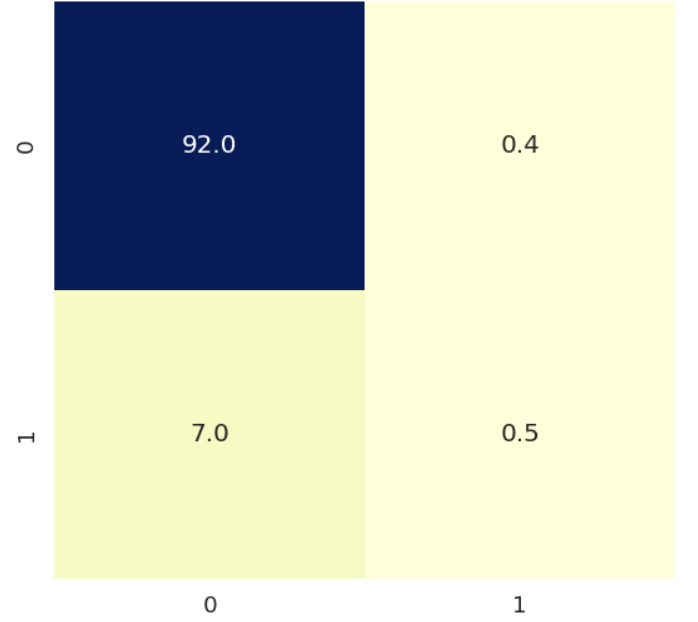


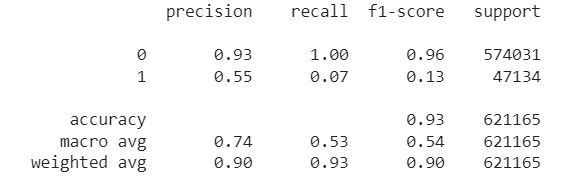




The next step was splitting the data into train and test data set in the ratio 70:30. Thereafter the code for Logistic Regression was run and the following results were observed:

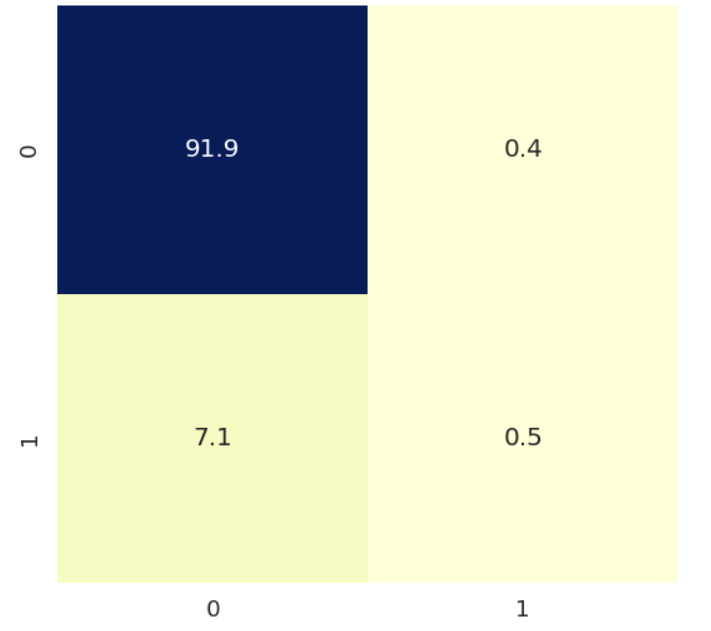
## Train Data Set

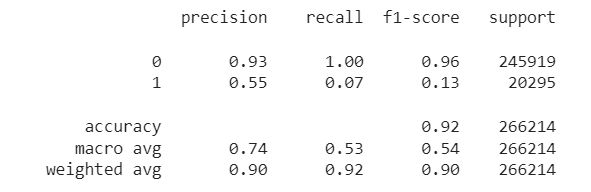




The train data set suggests that the model is quite effective in identifying the good loans as the F1 score is quite high, however, the concerned category, i.e. the bad loans (denoted by 1),

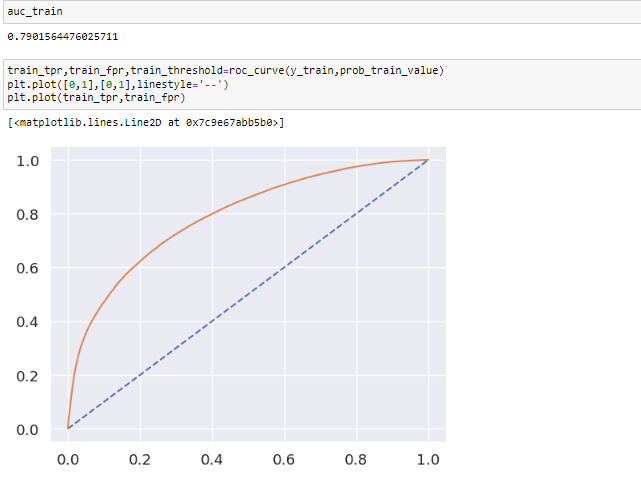
## Test Data Set





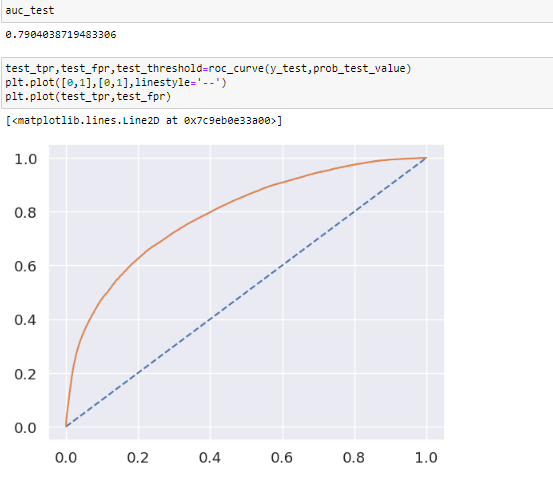
The same observation can be made regarding the test data set as well since our concerned category is not being identified correctly.

## AUC Score – Train



The AUC score suggests that the classification is not optimum since the AUC score is only 79%.

## AUC Score – Test



Along similar lines, the AUC test score for the test data set also shows that the model is not optimum.

Observations:

* The Logistic regression model is a good fit one.
* The model needs to be optimized to identify bad loans.
* The cause of this bad classification maybe because of imbalance in the data set.
* The recall for the bad loans need to be increased.

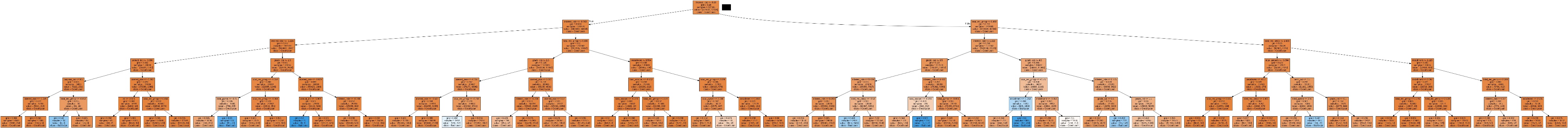
# Decision Tree

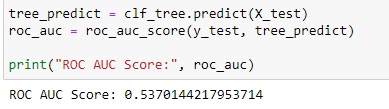
In the next step, we created a decision tree based on the standardized set to understand whether the algorithm would help us in categorizing the good loans and the bad loans more effectively.

To this end, a grid search algorithm was run to find the proper hyperparameters:



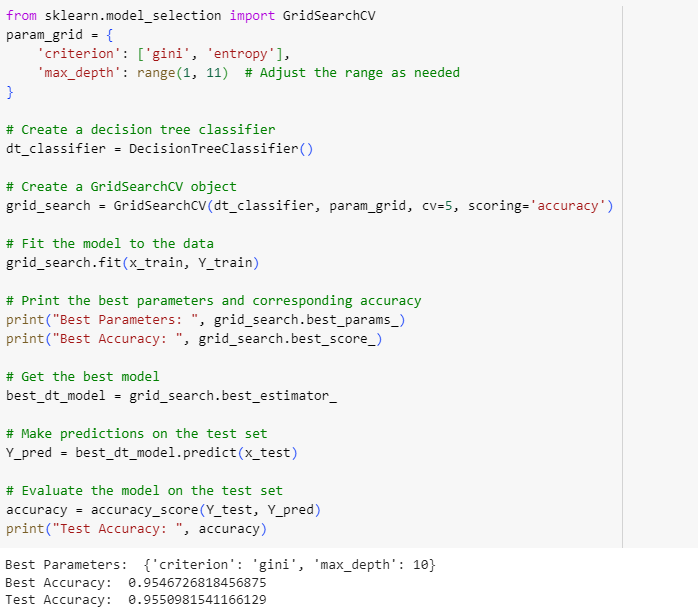
Hence the optimum splitting criteria is: Gini and the max depth is 10, which gives the best accuracy in train and test data. However, the interpretability of the model will go for a toss in case of a model with 10 layers. Therefore, a tree with a depth of 6 was created:





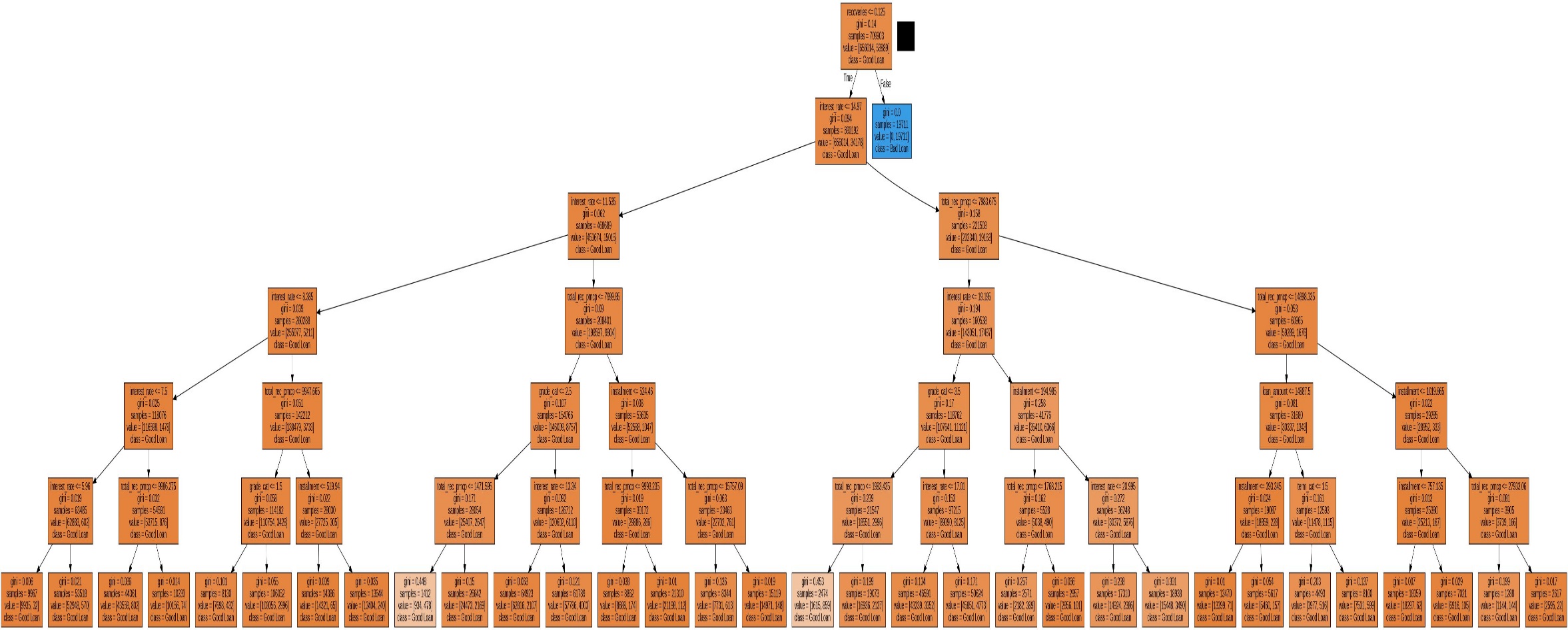
Alternatively, since decision tree algorithm is not prone to scale of the variables, we have used the original data for analysis as well.

We ran the same grid search algorithm, and there was an increase in the accuracy in the following manner.

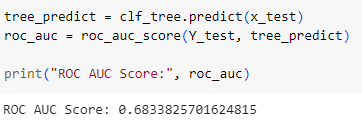


However, due to the interpretability issue, we decided to go with decision tree with 6 layers again.

The results are as follows:

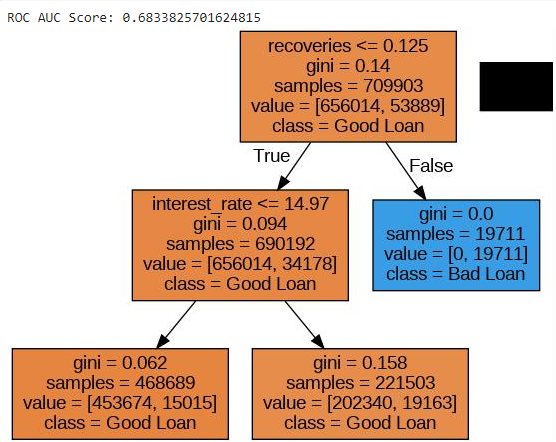


Here we can see that the decision tree has split the tree into two classes where the blue box represents the bad loans and the orange boxes represent the good loan. However, the ROC AUC score is:



The ROC AUC suggests that the model created is not a good fit to the test data.

Given the decision tree correctly split the train data into 2 classes we ran the algorithm once again with 2 layers.

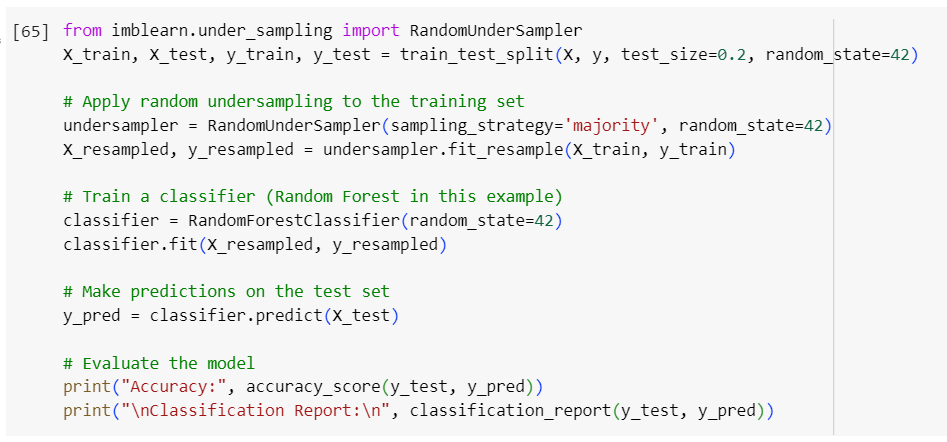


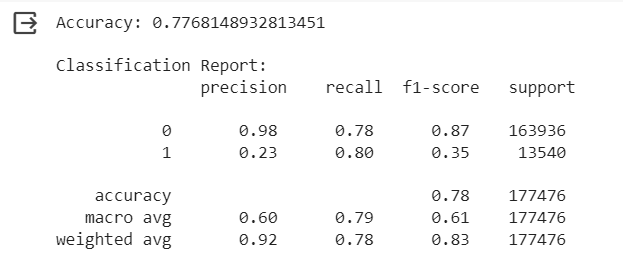
However, the ROC AUC score is still quiet, which prompted us to shift to other models.

# Over Sampling and Under Sampling

Considering that the set is an imbalanced one, we also tried to reduce the effect of the major class by using under sampling and over sampling.

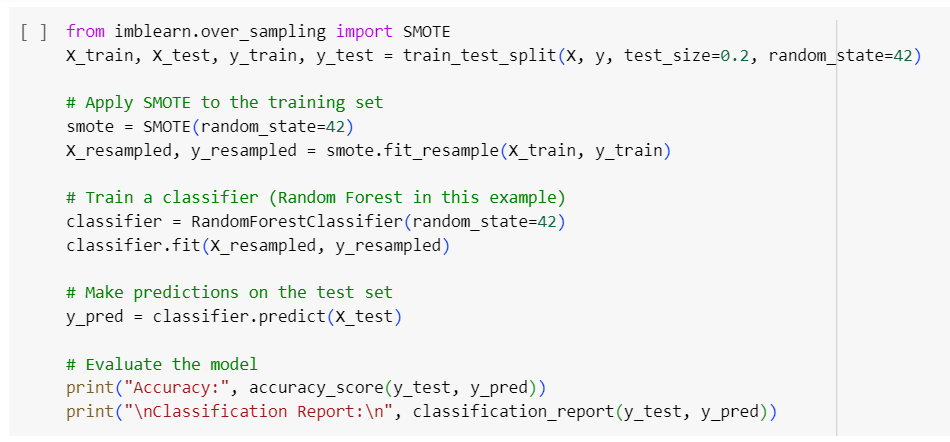
## Under Sampling

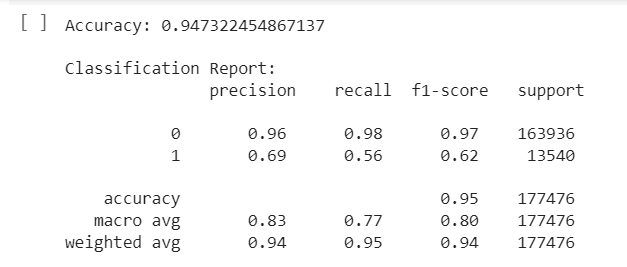




The under sampling has not provided to be fruitful since we have not been able to efficiently classify the bad loans. Even though the recall has increased as compared to Logistic Regression model, it is still not as efficient as is desired.

## Over Sampling using Smote





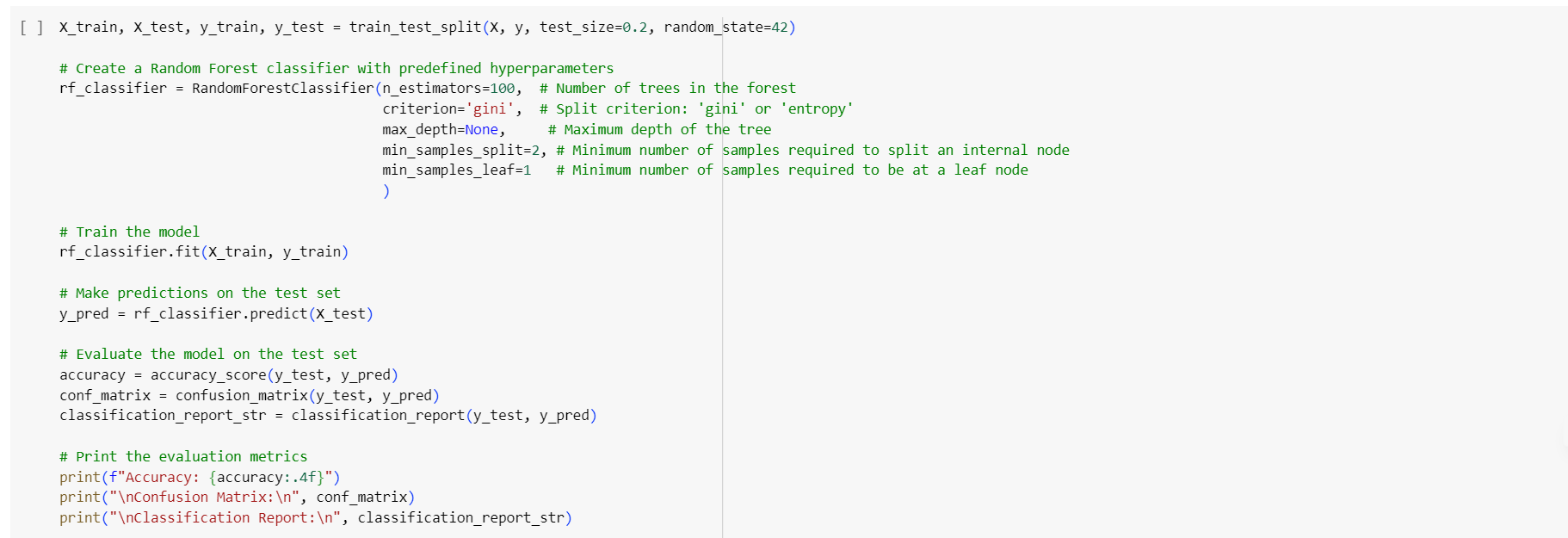
The over sampling has provided better results as compared to the Logistic Regression and Under Sampling models, where the classification of the bad loans is better than the above-mentioned models. Thus, it is a much better model. However, we can still improve the results, by using ensemble models.

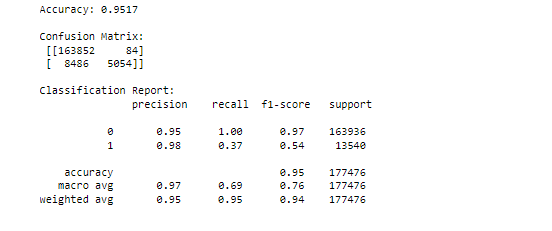
# Ensemble Models

Lastly, we tried increasing the accuracy of the model by using the ensemble models.

## Random Forest

We first tried the Random Forest with 100 estimators and Gini index is the splitting criteria. I was unfortunately unable to run the Grid Search algorithm due to lack of computing power. The following code was run:

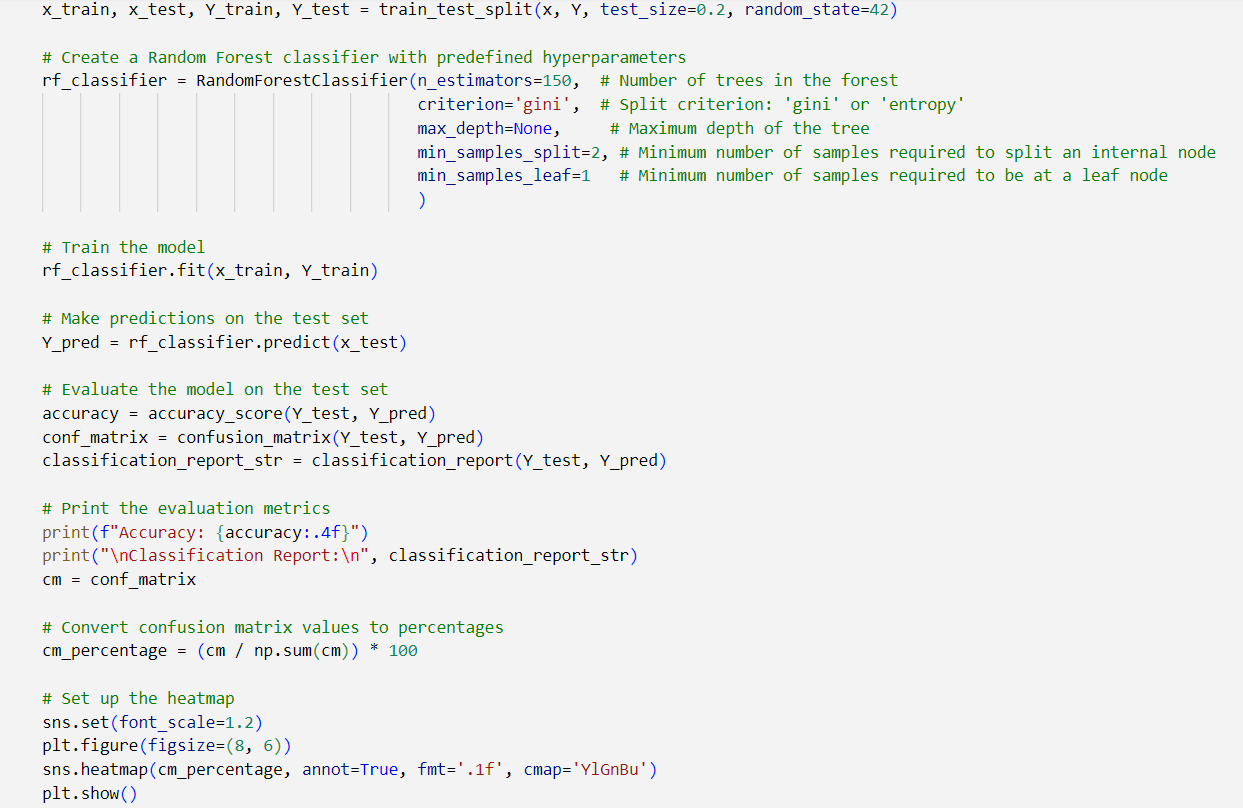


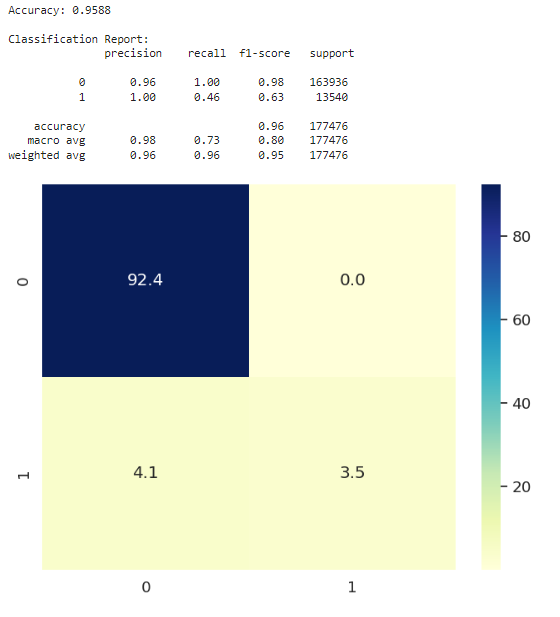


***It is to be noted that the ensemble models were run on the original data set without standardization of the continuous variables, since these models are quite robust to outliers in the data set.***

It can be observed that the random forest with 100 estimators does not quite improve the accuracy of the model. Although, it is a better model than the Logistic regression and Decision tree, but it is still not as efficient as the SMOTE model.

We therefore, ran the model again with 150 estimators:

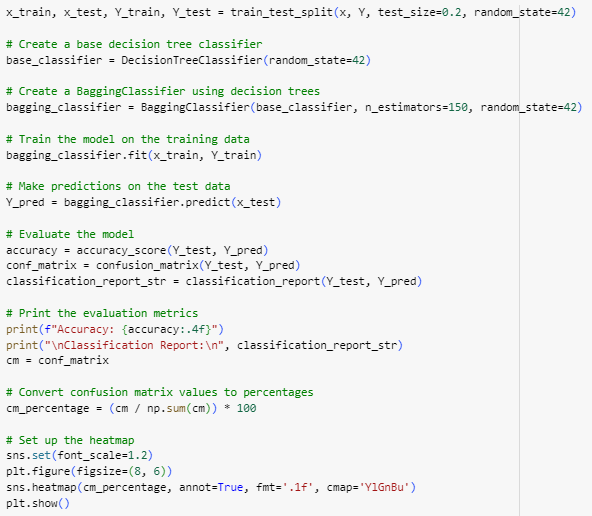


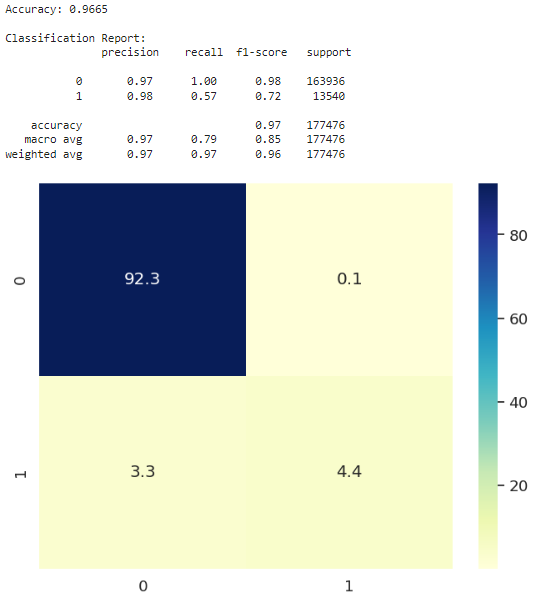


This model is a further enhancement on the Random Forest with 100 estimators, which is an improvement upon all the models above. However, the concerned class is still not getting identified to a desired extent (of 90% - F1 score) since the recall is quite low.

We did not carry on with the Random Forests because of lack of computing power and shifted to bagging.

## Bagging





It can be observed that the bagging model is the best model where we have used 150 estimators. It has the highest accuracy amongst all the models and has classified the concerned class most effectively.

It is more desirable that the F! be further increased by either using more estimators or by using bagging model. However, the computing power is limited.

# Conclusion and Suggestions

## Problems with the Data Set

* The data was too big for our computer to handle, which made it hard to create a decision tree, and when we tried to improve it.
* The computing resources were Limited and not powerful enough, which affected how well our model could predict things from the data.
* To fix this, we need to either get a more powerful computer or find different ways to work with such large amounts of data, making sure our predictions are still accurate for example use of Cloud Computing. Conclusion
* Implementing Random Forest with bagging has proven effective in improving the F1 score for distinguishing between good and bad loans.
* While boosting could offer additional enhancements, computing power constraints make bagging a pragmatic choice.
* Despite the potential benefits of iterative approaches, the current limitations call for a balanced consideration of efficiency and accuracy.
* In conclusion, Random Forest with bagging provides a commendable strategy for creditworthiness assessment, offering notable predictive improvements within our computational constraints.