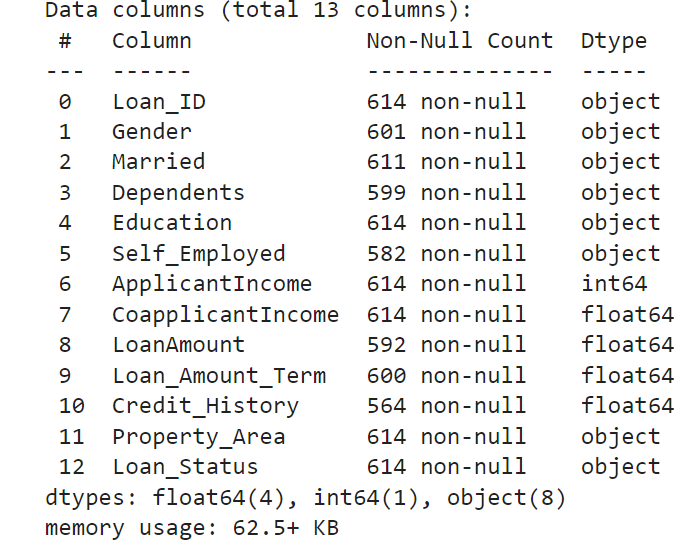
**INTRODUCTION**

**Problem Statement: -To build a model that can predict whether the loan of the applicant will be approved or not on the basis of the details provided in the dataset.**

A Prediction Model uses data mining, statistics and probability to forecast an outcome. Every model has some variables known as predictors that are likely to influence future results. The data that was collected from various resources then a statistical model is made. It can use a simple linear equation or a sophisticated neural network mapped using a complex software. As more data becomes available the model becomes more refined and the error decreases meaning then it’ll be able to predict with the least risk and consuming as less time as it can. The Prediction Model helps the banks by minimizing the risk associated with the loan approval system and helps the applicant by decreasing the time taken in the process. The main objective of the Project is to compare the Loan Prediction Models made implemented using various algorithms and choose the best one out of them that can shorten the loan approval time and decrease the risk associated with it. It is done by predicting if the loan can be given to that person on the basis of various parameters like credit history, Applicant income, Co-applicant income, age, married, gender, etc. The prediction model not only helps the applicant but also helps the bank by minimizing the risk and reducing the number of defaulters.

The parameters used in dataset are: -

Figure: -1

****

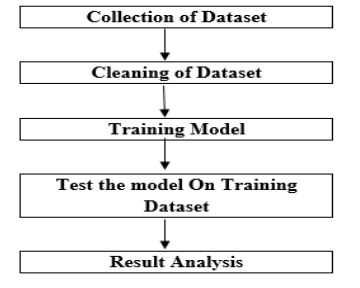
In the present circumstances, a loan needs to be approved manually by a representative of the bank which means that person will be responsible for whether the person is eligible for the loan or not and also calculating the risk associated with it. As it is done by a human it is a time-consuming process and is susceptible to errors.

If the loan is not repaid, then it accounts as a loss to the bank and banks earn most of their profits by the interest paid to them. If the banks lose too much money, then it will result in a banking crisis. These banking crisis affects the economy of the country. So it is very important that the loan should be approved with the least amount of error in risk calculation while taking up as the least time possible. So a loan prediction model is required that can predict quickly whether the loan can be passed or not with the least amount of risk possible.

The paper will be comparing different prediction models and deduce their limitations as well as advantages. The authors have used the same data for all the models which will give a clearer view on their performance and lead to a better comparison of the same. On the basis of the results, a modified prediction model will be created to ensure maximum accuracy and performance.

**Implementation**

Figure:-2

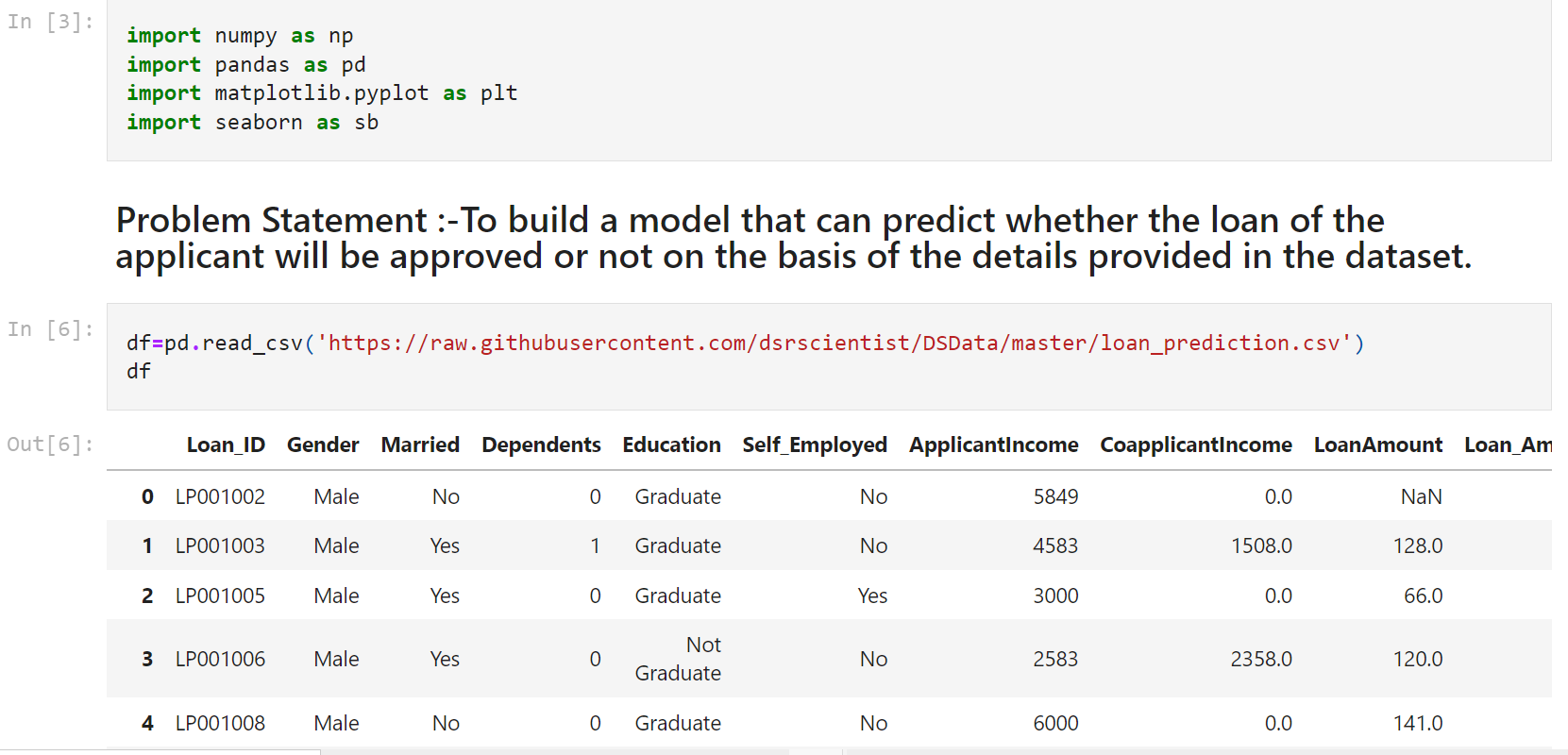
****

**PROPOSED MODEL**

* This system predict whether the loan is approve or reject .
* This System refers the following things or ways.
* Data Collection Data Pre-processing (Data Cleaning)
* Model Selection
* Model Evaluation
* Classification Result (output)

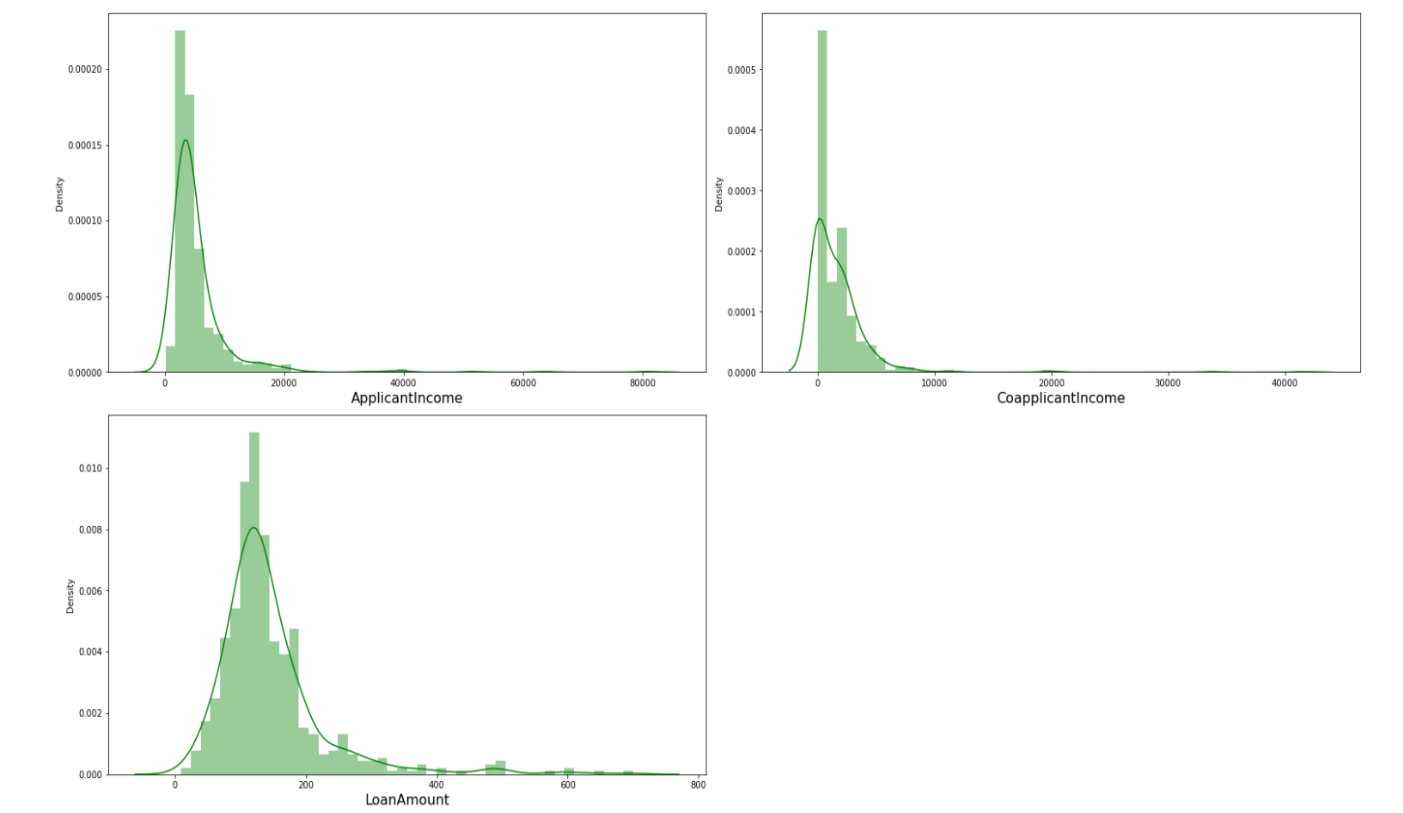
**Data collecting and importing**

* The Data set was imported in csv format and a simple function is applied to which is pd**. read\_csv (“ ”) for** read the csv data and use for further analysis.
* Figure:-3



**Distribution of attribute**

Figure:-4



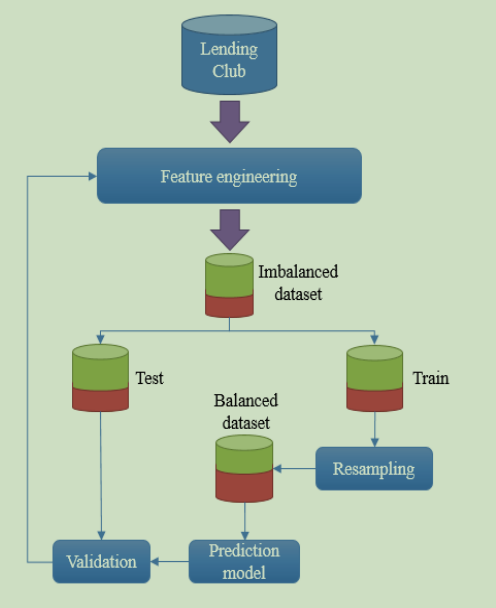
**It shows the skewness of the columns .**

**Skewness is only explore in continous columns not in categorical columns.**

**Research Methodology**

To help lenders evaluate the creditworthiness of borrowers in social lending platforms, we developed a decision support system that

includes a novel prediction model to reduce the risk of loan defaults.

****

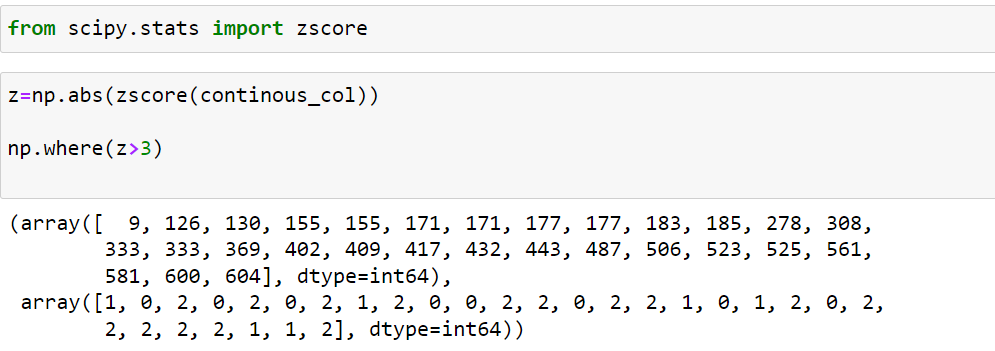
**Cleaning outliers**

Its main purpose is to enhance data reliability by cleaning data and selecting the subset of data features with the most discriminatory power. In credit risk prediction, ignoring irrelevant features can

increase classification accuracy and decrease the computational costs associated with running several machine learning models. Feature selection also reduces the dimensionality of the data, which helps mitigate the risk of overfitting. Our model comprises four important steps: data cleaning, leaky data removal, data transformation and correlation analysis, and deriving new attributes. The data is cleaned by first removing missing and

null values from dataset , then all outliers are removed according to the acceptable range defined in.

**for removal of outliers**



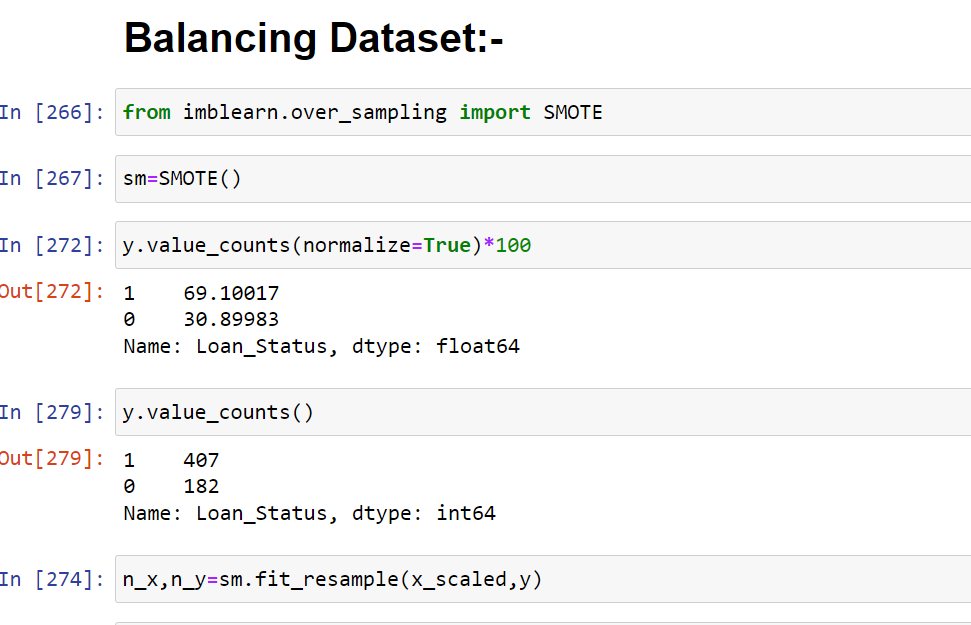
**Imbalanced Learning Approaches**

This study employs resampling approach to deal with the imbalance problem. Demonstrates three categories of resampling approach including undersampling, over-sampling, and hybrid methods. We

address the state of the art algorithms in each of these categories. The under-sampling approach includes random under-sampling (RUS), and instance hardness threshold(IHT) algorithms. For over-sampling approach, random over-sampling (ROS), synthetic

minority over-sampling technique (SMOTE), and adaptive synthetic sampling (ADASYN) are studied. Finally, SMOTE + Tomek links (SMOTE-TOMEK) and SMOTE+ edited nearest neighbour (SMOTE-ENN) are two prominent hybrid approaches that considered

by this research.

****

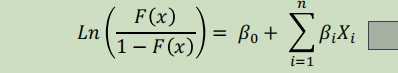
**Classification Models**

To the best of our knowledge, Logistic regression, Decision tree, Gradient boosting, and Random forest have demonstrated the best performance in the area of classification. Therefore, these three algorithms were selected for loans evaluation in this research.

**Logistic Regression**

Logistic regression is a standard industry algorithm that is commonly used in practice because of its simplicity and balanced error distribution . It is a binary classification technique that generates one of two variables as its result, e.g., good or bad borrowers. The

logistic regression formula is shown in Eq.



Where F(x) refers to the probability prediction, *β0* is the constant coefficient, and *βi* is the coefficient for the feature *xi*, which is calculated using maximum likelihood. Therefore, for a set of features, *xi i=1,…,n*, the logistic regression algorithm predicts the

probability that a sample belongs to a specific class

**Linear Discriminate Analysis**

Linear discriminate analysis is a statistical algorithm that determines the relationship between a target variable and a set of independent variables . Many studies into credit scoring have used linear discriminate analysis because it tends to achieve better performance than other classifiers when linear patterns are involved

**Random Forest**

Random forest algorithms are based on ensemble trees. This method, which can be seen as an enhanced bagging technique, is a powerful way to construct a forest of random decision trees. A random forest algorithm can also build multiple decision trees that have been trained on bootstrap samples from the training data. Rather than considering all available features, the algorithm randomly chooses a subset of attributes when building the trees or splitting the nodes. Once all the trees have been generated, the most

popular class is decided with a voting function

**Validation**

The dataset was divided into a train set and a test set at a ratio of 80:20. Only the training set was balanced, through resampling, then validated with the still imbalanced test set.

**Feature Description**

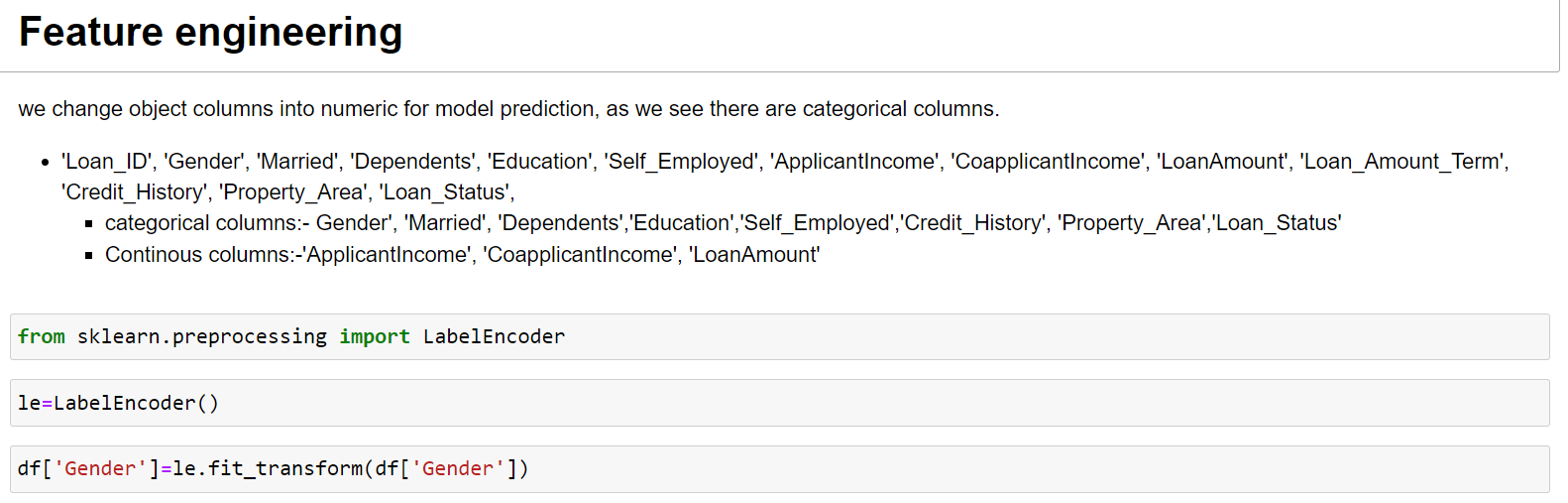
Loan\_ID, Gender, Married, Dependents, Education,Self\_Employed,

ApplicantIncome, CoapplicantIncome, LoanAmount,

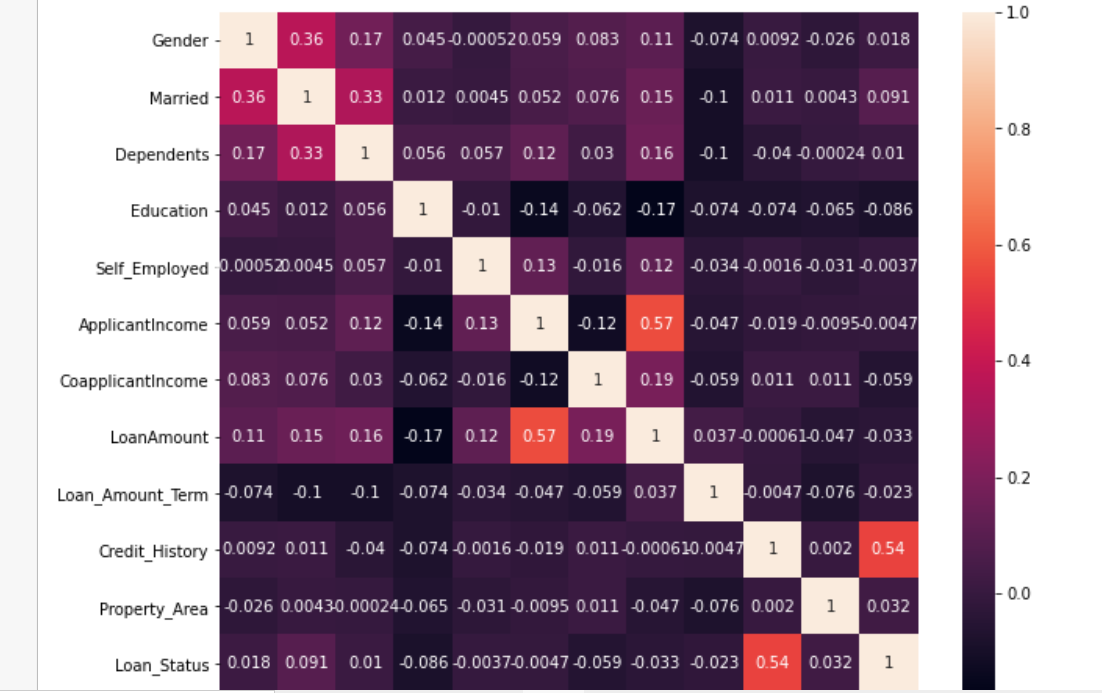
Loan\_Amount\_Term, Credit\_History, Property\_Area, Loan\_Status

**Feature engineering** :-

Feature engineering is done by using label encoder. It transform all the categorical columns which is in object type into numeric type for model training and prediction

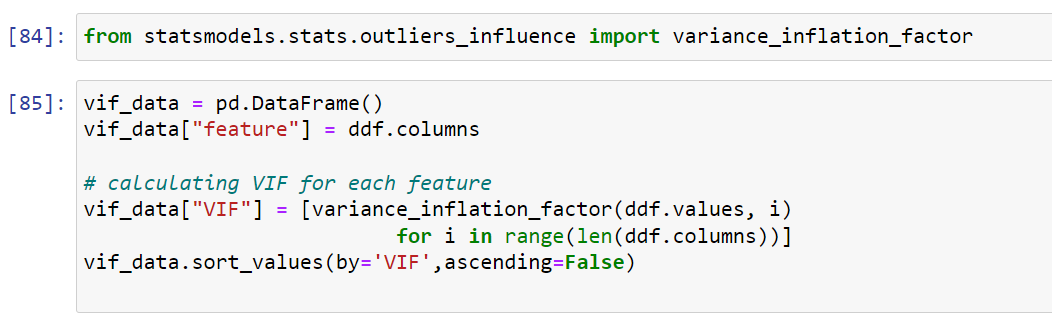


**Correlation: - threshold limit is >= 0.91**

****

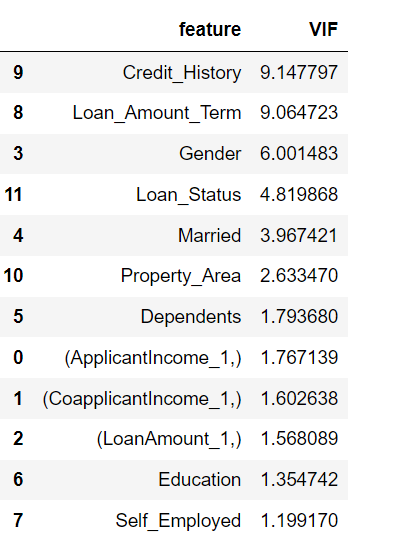
This is correlation chart shows the features columns are how much correlated. Whether they are strong correlated or weak, positive or negative. It also indicates multicollinearity between the feature

**Checking VIF**

****

**Importing VIF and the values**

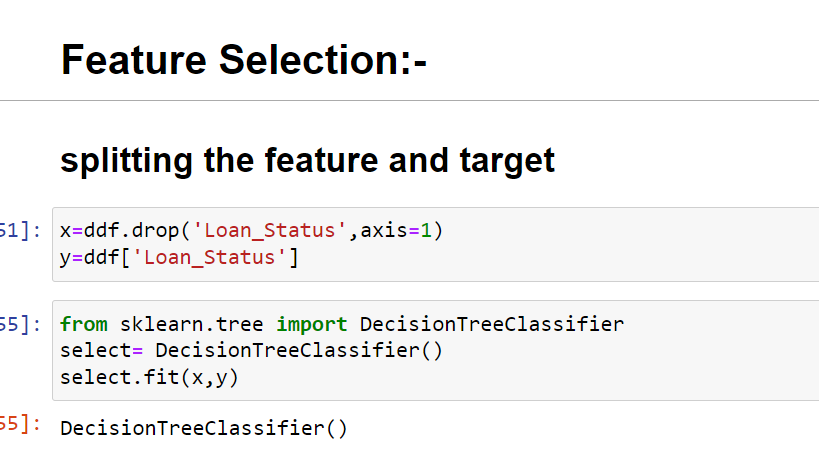
The **variance inflation factor** (VIF) quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing collinearity/multicollinearity. Higher values signify that it is difficult to impossible to assess accurately the contribution of predictors to a model.

****

I am not removing any feature here all the VIF values of features is less than 10. And for my view it not indicating multicollinearity.

**So, I am keep the all feature and use for prediction.**

**Feature selection**

****

Splitting the feature and target columns and then by using the Decision tree classifier extract the features values.

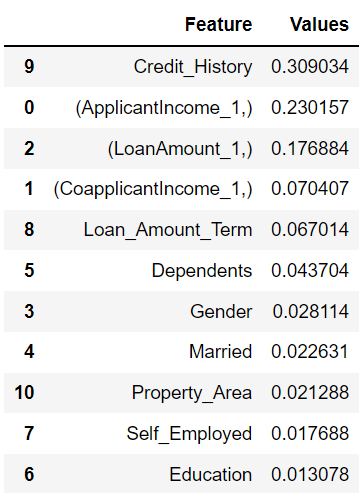
**== select. feature\_importances**\_

**[0.23015726,0.07040705,0.17688404,0.02811426,**

**0.02263133,0.0437044,0.01307788,0.01768774,**

**0.06701418,0.30903385,0.02128802]**

These are features values in ascending order





0.23015726**+** 0.07040705**+** 0.17688404**+** 0.02811426**+** 0.02263133**+**0.0437044 **+** 0.01307788**+** 0.01768774**+** 0.06701418**+** 0.30903385**+**0.02128802

Out[260]:1.00000001

It gives the value 1.00000001 which is complete 100%

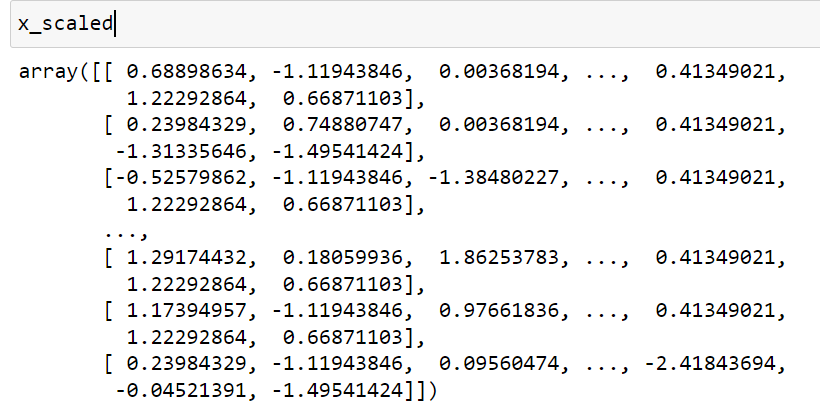
* so, we may drop one columns which is very close to 0.
* but, I have not dropped any columns, selecting all the columns.

# **SCALING DATA**

from sklearn.preprocessing import StandardScaler

scale= StandardScaler()

x\_scaled=scale.fit\_transform(ddf)



**why is data scaled?**

The purpose of scaling the data cause of different types of features

Some of them are categorical and some of them are continuous. And all of them are measures with different parameter to make them equal or we say that to bring all the feature columns into single scale we do the scaling.

I am using standard scaler. Standard scaler removes the mean and scales each feature/variable to unit variance. This operation is performed feature-wise in an independent way.

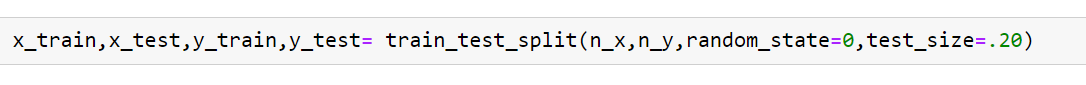
**After balancing the data, the next step is model building and training. Before that we find out the best random state**

**Model building and training :-**

****

Using logistic regression, we get the best random state and accuracy is seen here is very good.

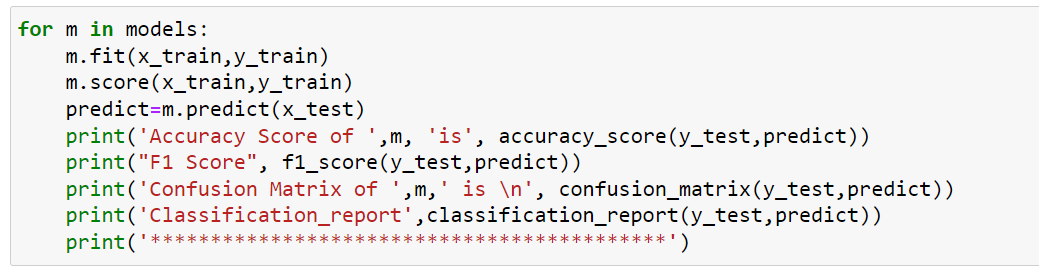
**Splitting the test and train data**

****

**This is use for prediction**

**Model Building: -**

****

****

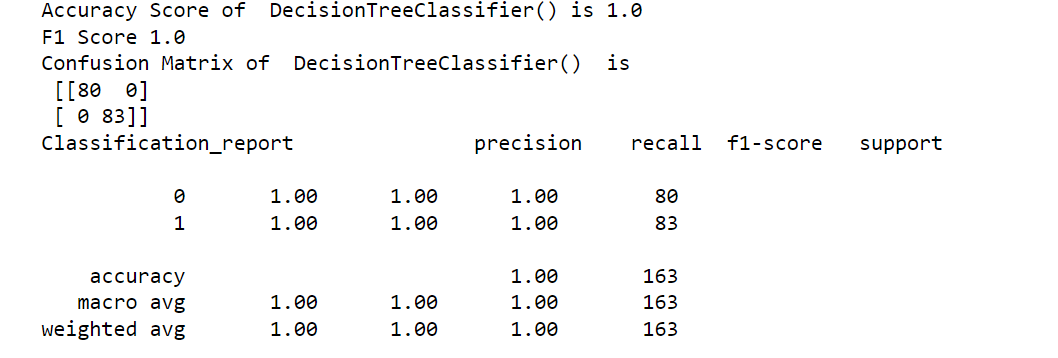
Imported all the different types of Machine learning algorithm.

In all these models who shows best result after cross validation. we save that for further prediction

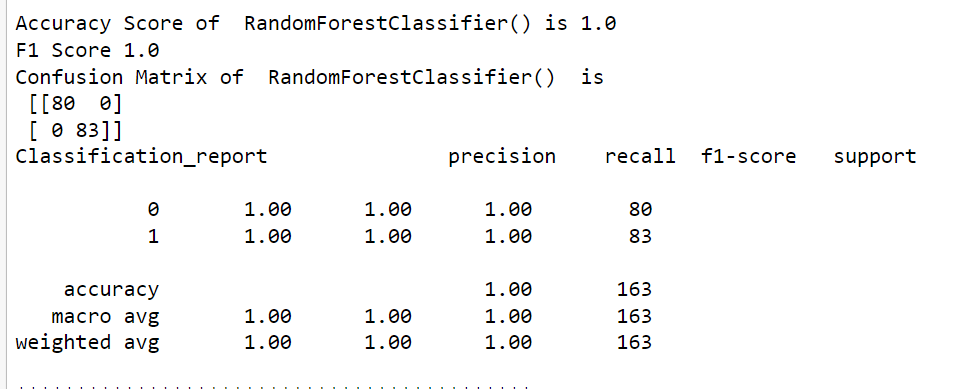
**These are some models results: -**

Showing some of 3 important models

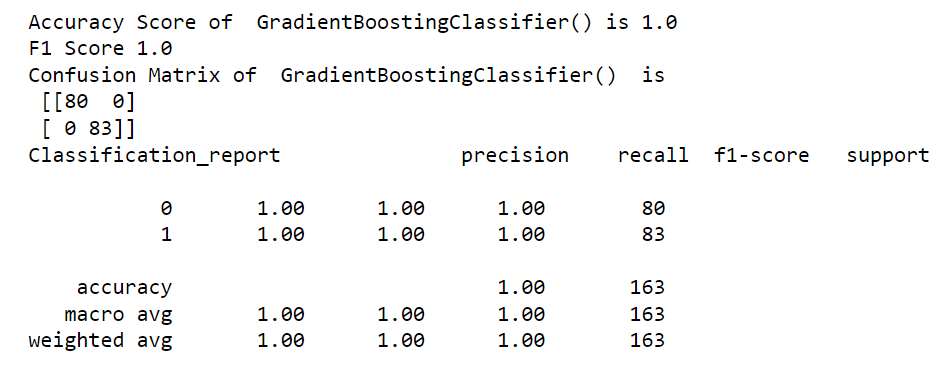
**1.**

****

**2.**

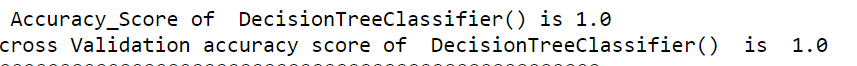
****

**3.**

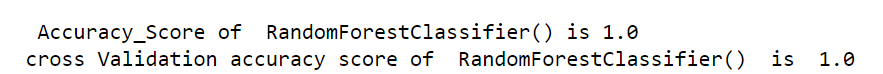
****

**Cross validation: -**

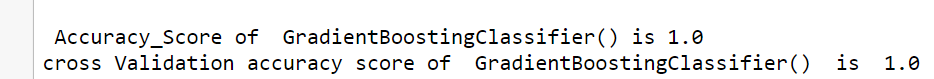
**1.**



**2.**



**3.**

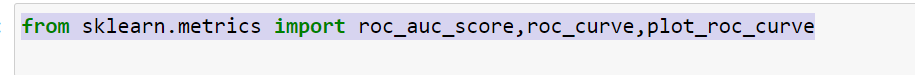


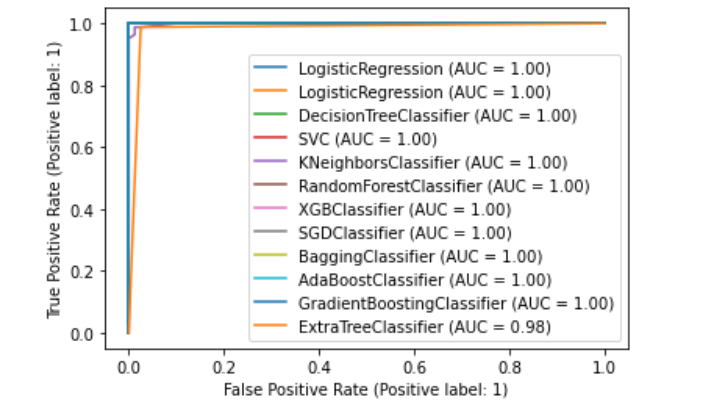
**These all three cross validation and accuracy score are same there is no change, no problem of over fitted model so, we can choose any one of them for prediction.**

# **Hyperparameter Tuning: -**

We don’t need any hyperparameter tuning as we see it already showing its best accuracy, and not only one or two model most of model is showing its best accuracy score and also after cross validation score its shows model is not overfitted so, we can save the models.

# **Plot roc curve: -**





All models except one model Extra tree classifier shows (AUC=0.98)

AUC shows area under the curve ,higher the AUC higher the accuracy.

## **Here, we see all the model overlapped each other, and all the model has AUC=1**

### **so, we save any of the models, they are showing almost 100% accuracy**

### **The model is saved for further prediction**

**Conclusions and Future Works**

Identifying the risk score for a potential borrower is crucial for the healthy functioning of social lending markets, where class imbalance problems are prevalent. However, few studies into social lending

platforms have considered the characteristics of imbalanced data. Moreover, the efficiency of resampling techniques in evaluating P2P loans is a controversial issue. To calculate the creditworthiness of borrowers in P2P lending platforms, we used the most recent data

published by the Lending Club. Appropriate features were selected through comprehensive feature. Engineering process, and we introduced a non-standard financial feature to increase the reliability of the computed risk scores. Additionally, given the Lending Club dataset contains imbalanced classes, we also compared different resampling methods to determine the best overall technique. Accordingly, the state-ofthe-art classifiers – random forest, logistic regression and linear discriminate analysis – were combined with

different resampling techniques and tested on Lending Club’s data. Our experiments show that random forest and random under-sampling may be an efficient combination of classifier and resampling strategy to compute risk scores for loan applicants in social lending markets.

P2P lenders can take advantage of the credit risk prediction modelling discussed in this study to make smarter decisions when evaluating loan applications. Moreover, lenders might apply the attributes

identified in this study to compute the creditworthiness of borrowers. Identifying default borrowers in advance can prevent financial loss. Furthermore, more accurate assessments of the probability of default may also help when developing strategies to compensate for risk, such as increased interest rates.

**References**

1. Guo, Y., Zhou, W., Luo, C., Liu, C., and Xiong, H.: ‘Instance-based credit risk assessment for investment decisions in P2P lending’, European Journal of Operational Research, 2016, 249, (2), pp. 417-426

2. Malekipirbazari, M., and Aksakalli, V.: ‘Risk assessment in social lending via random forests’, Expert Systems with Applications, 2015, 42, (10), pp. 4621- 4631

3. Emekter, R., Tu, Y., Jirasakuldech, B., and Lu, M.: ‘Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending’, Applied Economics, 2015, 47, (1), pp. 54-70

4. Xia, Y., Liu, C., and Liu, N.: ‘Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending’, Electronic Commerce Research and Applications, 2017, 24, pp. 30-49

5. Lin, W.-C., Tsai, C.-F., Hu, Y.-H., and Jhang, J.-S.: ‘Clustering-based undersampling in class-imbalanced data’, Information Sciences,2017, 409, pp. 17-26

6. Ala'raj, M., and Abbod, M.F.: ‘Classifiers consensus system approach for credit scoring’, Knowledge-Based Systems, 2016, 104, pp. 89-105

7. Byanjankar, A., Heikkilä, M., and Mezei, J.: ‘Predicting credit risk in peer-to-peer lending: A neural network approach’, in Editor (Ed.)^(Eds.): ‘Book Predicting credit risk in peer-to-peer lending: A neural network approach’ (IEEE, 2015, edn.), pp. 719-725

8. Siami, M., Gholamian, M.R., Basiri, J., and Fathian, M.: ‘An Application of Locally Linear Model Tree Algorithm for Predictive Accuracy of Credit Scoring’, in Editor (Ed.)^(Eds.): ‘Book An Application of Locally Linear Model Tree Algorithm for Predictive Accuracy

of Credit Scoring’ (Springer, 2011, edn.), pp. 133-142

9. Siami, M., Gholamian, M.R., and Basiri, J.: ‘An