**Approval of Loan Prediction Model with Machine Learning**

# Introduction

#### Problem Statement:

This dataset includes details of applicants who have applied for loan. The dataset includes details like credit history, loan amount, their income, dependents etc.

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.

Machine Learning is a boon in helping the Loan industry with this problem.

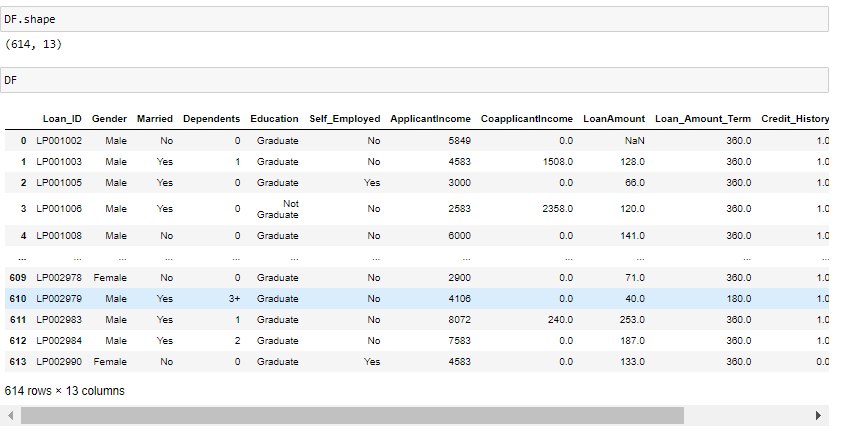
**Executive Summary:**

In this project, a dataset was provided with the details of the loan along with the customer details, as well as the personal details of the person who has applied for loan.

The Dataset was first cleaned, the various feature columns were analysed, then with feature engineering and based on strength of correlation and ANOVA f-score values, the feature columns were selected that would best predict the Target variable, to train and test machine learning models.

The loan dataset was worked with to build a predictive model that best predicts if the loan will be approved or rejected. Several models were trained and fitted with a part of the dataset and then tested with a different part of the dataset. The model that performed the best with the best confusion matrix performance, f1 score, ROC-AUC score and cross validation performance was then selected and tuned further with hyper parameter tuning techniques.

**About the Dataset:**

****

The given dataset consists of 13 columns and 614 rows.

**The Independent Feature columns are:**

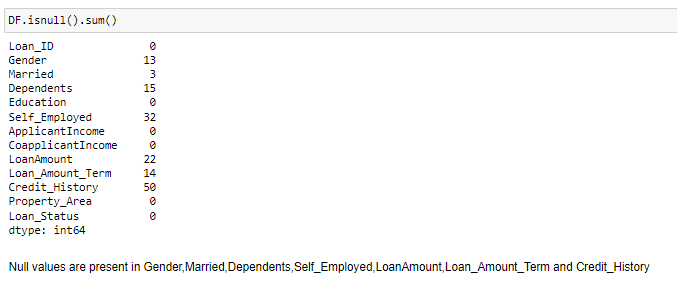
* Loan\_ID
* Gender
* Married
* Dependents
* Education
* Self\_Employed
* ApplicantIncome
* CoapplicantIncome
* Loan\_Amount
* Loan\_Amount\_Term
* Credit History
* Property\_Area

**The Target Variable to predict is given in the column:**

* Loan\_Status **:** Whether loan was approved or not as Yes or No

**Data Cleaning:**

Upon inspecting all the columns in the dataframe, it is observed that various columns have NaN values.



Gender – 13 Null values

Married – 3 Null values

Dependents – 15 Null values

Self\_Employed – 32 Null Values

LoanAmount – 22 Null Values

Loan\_Amount\_Term – 14 Null Values

Credit\_History – 50 Null Values

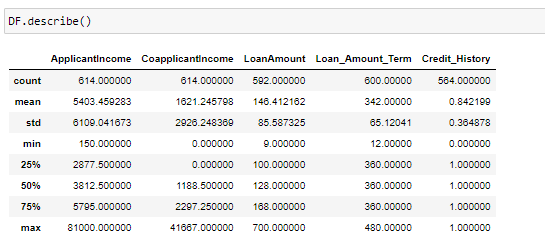
**Imputing Values to NaN values in all the columns with NaN (Null) Values**

The most Frequently occurring value in each of the above columns was imputed to the NaN values of the respective columns of the Frequently occurring values.

****

**After imputing the values it is observed that no more null values remain in the dataset.**

**Exploratory Data Analysis**

****

As there is huge difference between 75% and max in ApplicantIncome, CoapplicantIncome and LoanAmount, outliers are present.

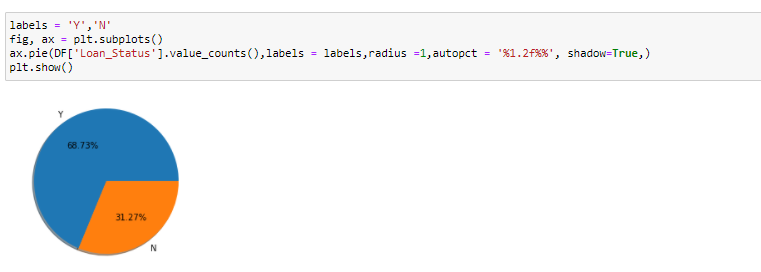
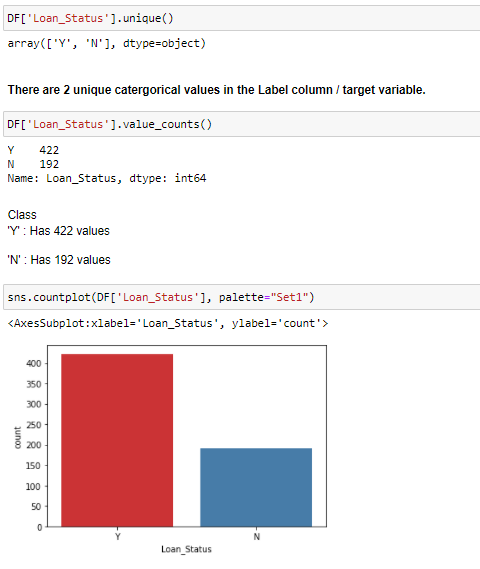
and skewness is present in ApplicantIncome and CoapplicantIncome as Standard deviation is higher than mean

### This is a Classification Problem since the Target variable / Label column ("Loan\_Status") has Categorical type of Data.

**Univariate Analysis**

**Analyzing the Target Class**

#### There are 2 unique categorical values in the Label column / target variable, viz. ‘Y’ and ‘N’.



**Class**

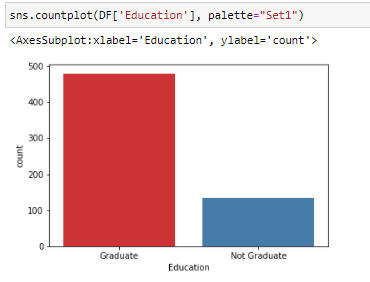
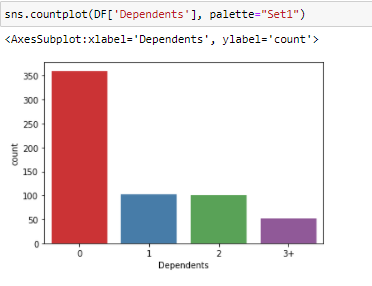
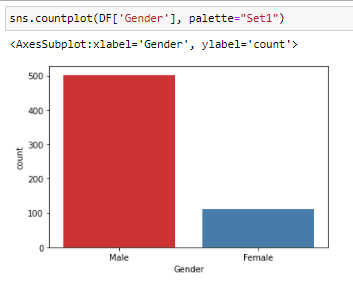
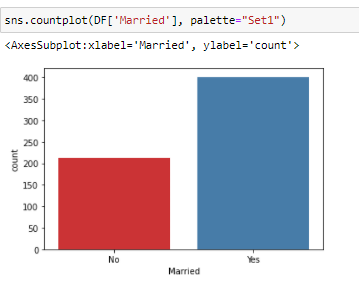
**'Y' : Has 68.73% of total values**

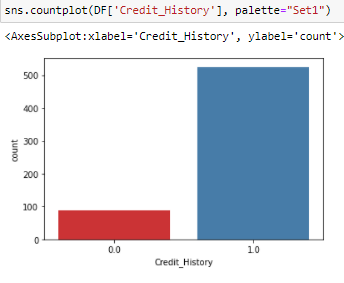
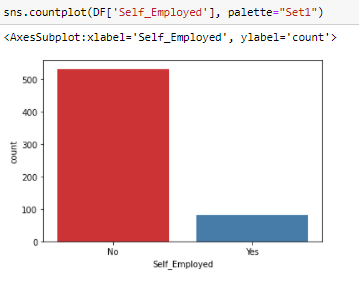
**'N' : Has 31.27% of total values**

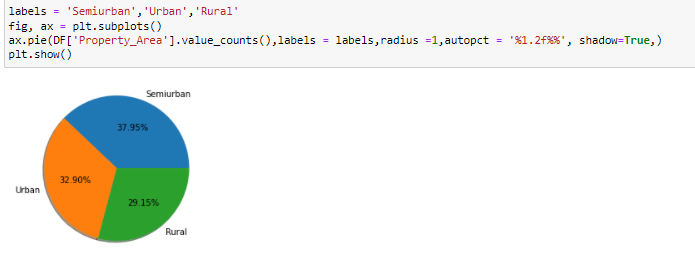
#### **Therefore, the Classes are imbalanced.**

**Upon analyzing the rest of the Feature Columns, following observations are made:**

* There are more Male Applicants than Female Applicants.
* There are more applicants who are married than those who are not.
* Most applicants have no dependents to support.
* Most applicants are Graduates.
* Most applicants are not Self Employed.
* Most applicants have a credit history of '1.0'.
* 37.95% applicants are from Semiurban, while 32.90% of applicants are from Urban and 29.15% are from Rural areas.

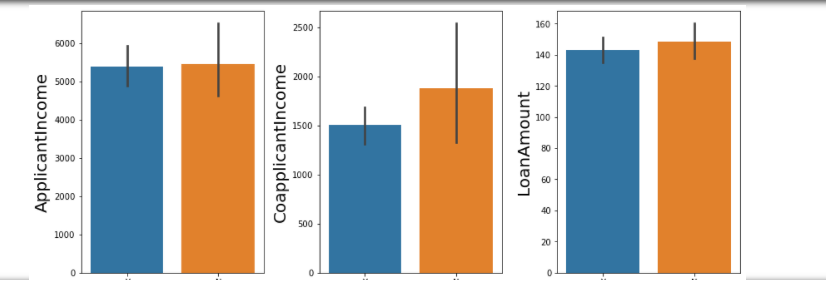


****

****

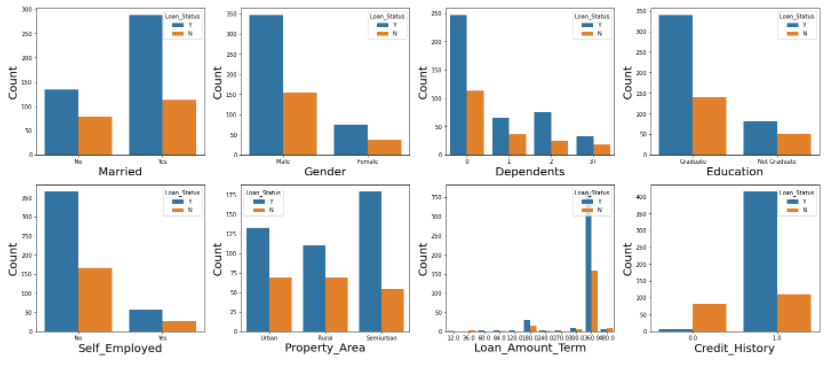
**Bivariate Analysis**

### Interpreting Relationship between Dependent Variable and Independent Variables



**Following observations can be made from above graphs:**

* Applicant income doesnt seem to contribute significantly to loan approval.
* Loan rejection is high for higher Coapplicant Income.
* LoanAmount doesn't seem to have a strong correlation with Loan Status.

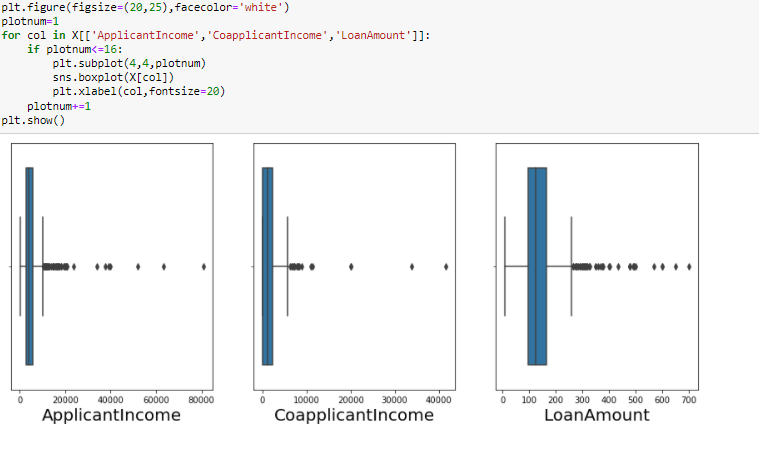


**Following observations can be made from above graphs:**

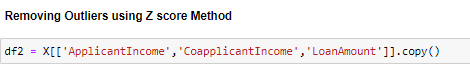
* More Married applicants have loan approval.
* Male applicants have higher loan approval than Female applicants. This may also be due to the fact that there are more male applicants than female applicants.
* Applicants with 0 dependents have highest loan approval.
* Graduates have higher loan approval.
* Applicants who are not self-employed have higher loan approval.
* Applicants from semi urban areas have highest loan approval.
* Highest loan approval is for 360.
* Applicants with Credit History of 1.0 has highest loan approval.

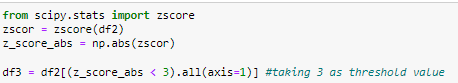
Column ‘Loan\_ID’ is not required for the prediction of values of the target column, hence, it can be dropped.

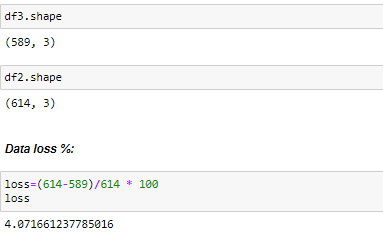
**Before Proceeding with finding the correlations of the columns, the outliers need to be checked and removed.**



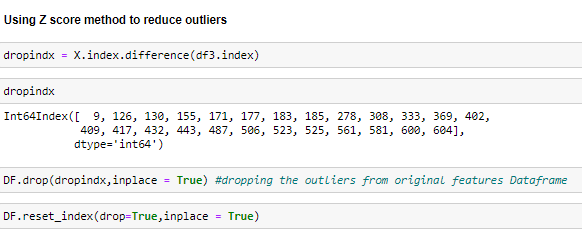
**With the help of z-score method, outliers will be removed.**

****

****

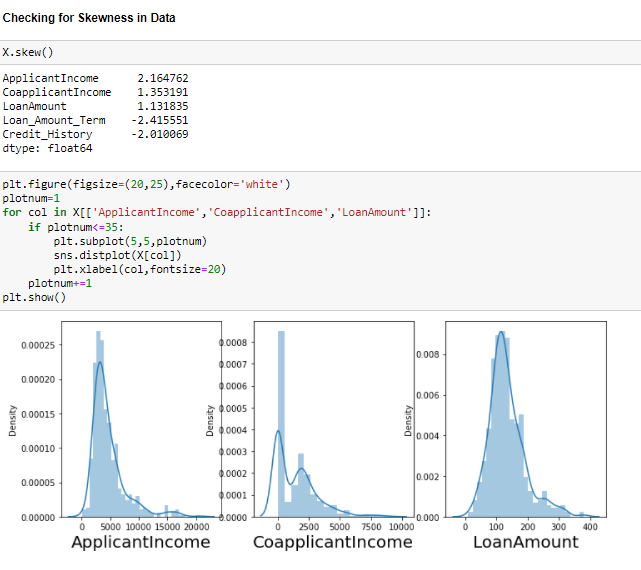
****

**As, the loss of data is only ~4.071%, we can proceed with it.**

****

**So, outliers are removed**

**And now, we will proceed to check the skewness and its removal.**

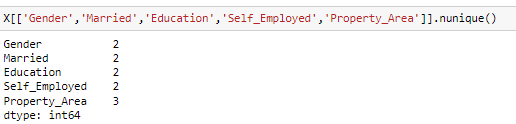
****

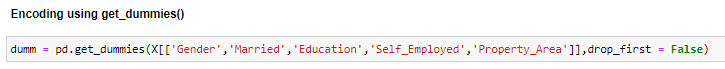
**Considerable skewness exist in the data and we need to remove it to move further.**

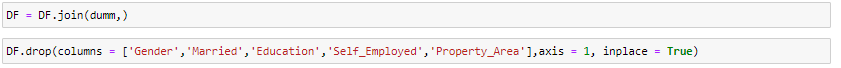
****

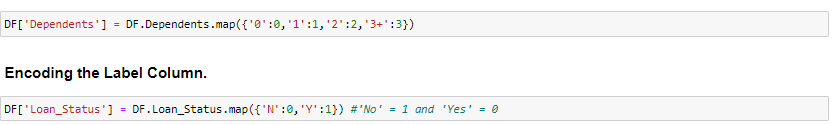
**Lot of skewness has been removed using this method and now, we will proceed further.**

**Now, we need to encode the categorical columns using get\_dummies()**

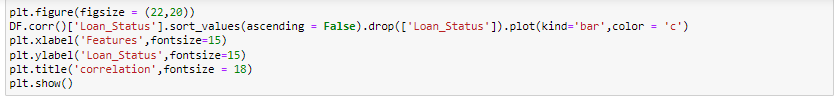
****

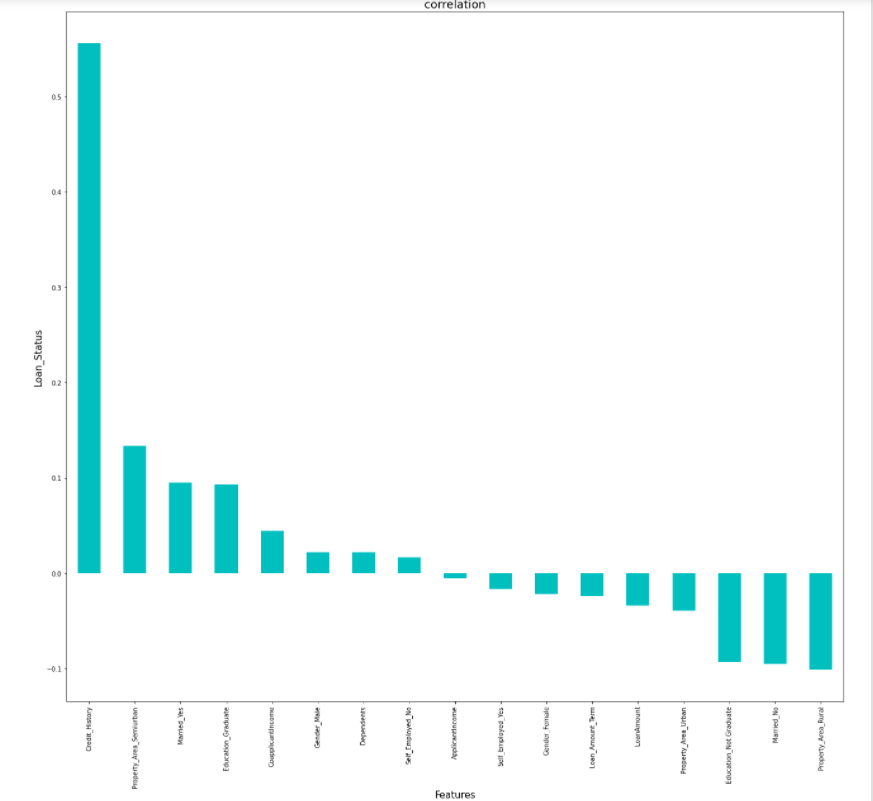
****

****

****

**Finding the correlations**

****



Credit\_History, Property\_Area\_Semiurban,Married\_Yes,Education\_Graduate have the highest positive correlation with Loan\_Status, while Propert\_Area\_Rural,Married\_No,Education\_Not Graduate have the highest negative correlation with Loan\_Status.

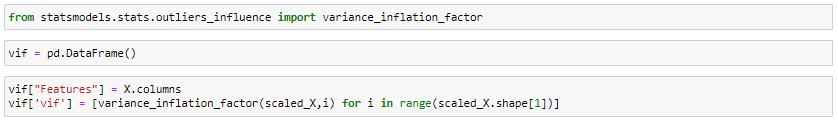
### 

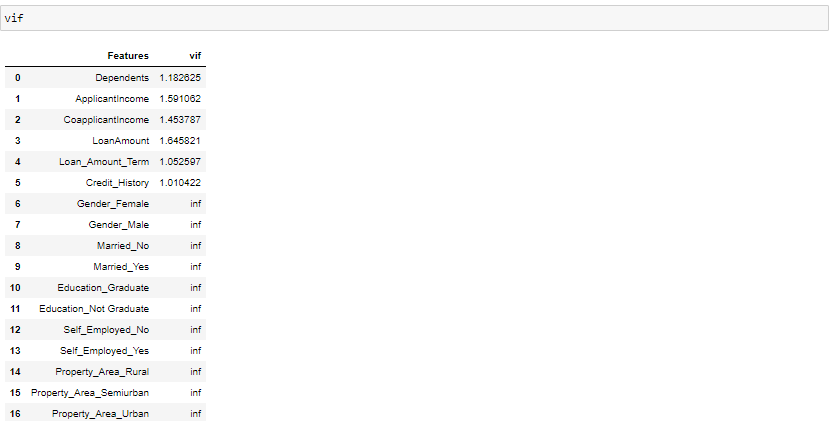
### Standardization needs to be done with the help of Standard Scaler

### 

### Checking for Multicollinearity using Variance Inflation Factor

Variance inflation factor measures how much the variance of an independent variable is influenced / inflated, by its interaction/correlation with other independent variables.

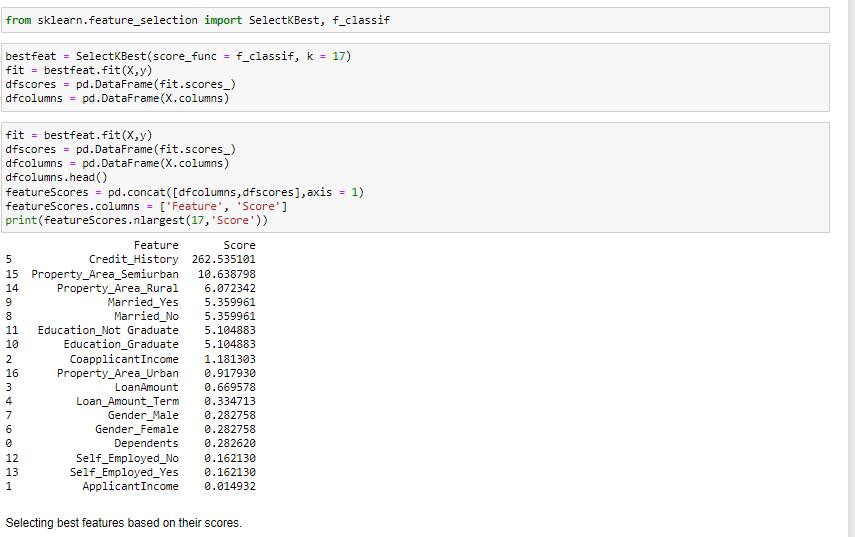




It is found that there is no multicollinearity

### Selecting Kbest Features

Based on the respective ANOVA f-score values, the feature columns are selected that would best predict the Target variable, to train and test machine learning models.





ApplicantIncome, Self\_Employed\_yes, Self\_Employed\_No, Dependents will be dropped as they are not the best features.

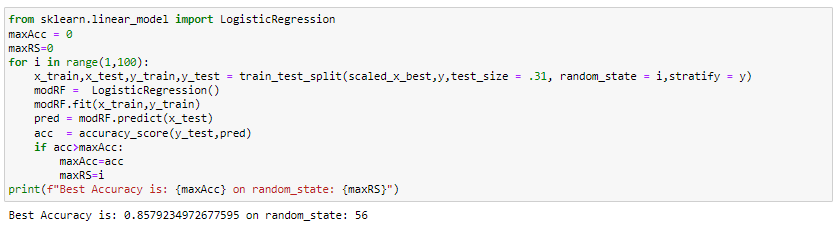
### Then, it will be standardize using Standard Scaler



**Classification Model Building**

****

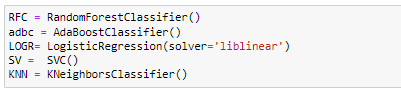
**Finding the Best Random State**



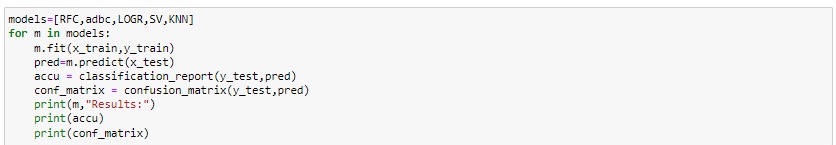
**Creating Train-Test split based on random state obtained above:**

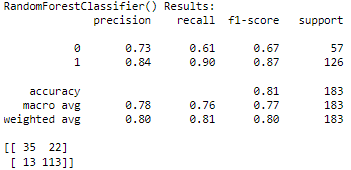
****

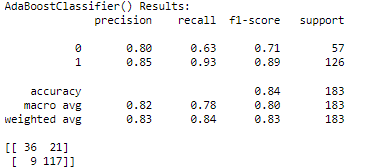
### Training the Models

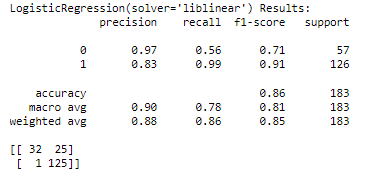
****

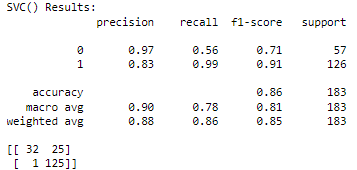
**We will use the for loop in order to get the results for all the ML models**

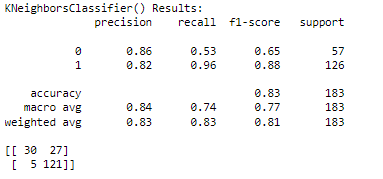
****

****

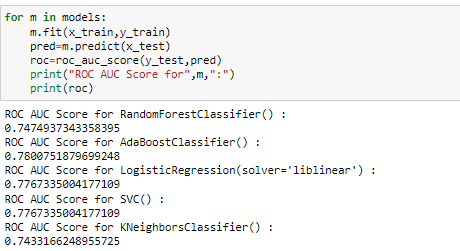
****

****

****

****

**ROC\_AUC\_Score**

****

**Model Cross Validation was done to check the over fitting**

****

****

**Analyzing Model Accuracies**

### Logistic Regression Model Accuracy

The trained Logistic Regression Model shows

F1 score of 0.86

Roc\_auc score of 0.7767

Cross validation score of 0.8149

Sensitivity(Recall for ‘ Loan Approval’ cases) is 0.56 and Specificity (recall of Loan Rejection cases) is 0.99

Precision for ‘ Loan Approval’ cases is 0.97 and Precision for Loan Rejection cases is 0.83

### Random Forest Classifier Model Accuracy

The trained Random Forest Classifier Model shows

F1 score of 0.84

Roc\_auc score of 0.7474

Cross validation score of 0.7690

Sensitivity(Recall for ‘ Loan Approval’ cases) is 0.61 and Specificity (recall of Loan Rejection cases) is 0.90

Precision for ‘ Loan Approval’ cases is 0.73 and Precision for Loan Rejection cases is 0.84

### AdaBoost Classifier Model Accuracy

The trained AdaBoost Classifier Model shows

F1 score of 0.84

Roc\_auc score of 0.7800

Cross validation score of 0.7877

Sensitivity(Recall for ‘ Loan Approval’ cases) is 0.63 and Specificity (Recall of Loan Rejection cases) is 0.93

Precision for ‘ Loan Approval’ cases is 0.80 and Precision for Loan Rejection cases is 0.85

### SV Classifier Model Accuracy

The trained SV Classifier Model shows

F1 score of 0.86

Roc\_auc score of 0.7767

Cross validation score of 0.8176

Sensitivity(Recall for ‘ Loan Approval’ cases) is 0.63 and Specificity (Recall of Loan Rejection cases) is 0.93

Precision for ‘ Loan Approval’ cases is 0.97 and Precision for Loan Rejection cases is 0.83

### K Nearest Neighbours Classifier Model Accuracy

The trained K Nearest Neighbours Classifier Model shows

F1 score of 0.83

Roc\_auc score of 0.7433

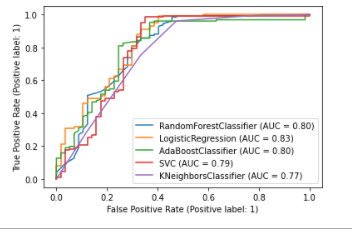
Cross validation score of 0.7708

Sensitivity(Recall for ‘ Loan Approval’ cases) is 0.53 and Specificity (Recall of Loan Rejection cases) is 0.96

Precision for ‘ Loan Approval’ cases is 0.86 and Precision for Loan Rejection cases is 0.82

### 

### ROC AUC curves



Logistic Regression has the largest area under their curve and scored 0.83 on AUC

Since , Sensitivity summarizes the true positive rate, ie. how many we got correct out of all the positive cases and Specificity summarizes our true negative rate, which is how many we got correct out of all the negative cases. The model that performs the best in those criteria will be chosen.

Both Logistic Regression and Support Vector Classifier have performed the best on Loan\_Prediction based on the fact that their Sensitivity, Specificity and Precision scores are the highest amongst all the model performances.

### Therefore, based on the above graph and roc\_auc\_scores, Logistic Regression is the best model for the dataset, with AUC = 0.83 and roc\_auc\_score = 0.7767

### Hyper Parameter Tuning

GridSearchCV was used for Hyper Parameter Tuning of the Logistic Regression model.

Based on the input parameter values and after fitting the train datasets,

The Logistic Regression Model was further tuned based on the parameter values yielded from GridsearchCV.

The Tuned Logistic Regression Model displayed an accuracy of 85.79%

### Concluding Remarks

In conclusion, Logistic Regression Model is able to correctly distinguish between Loan Approval and Loan Rejection with high accuracy.

The dataset had very limited data which is problematic as models show greater stability when the dataset is of a good size. However, a large set of feature columns enabled selecting a smaller feature size that provides the best accuracy for the model and obtaining results in optimal time.