## EVALUATION PROJECT REPORT

## PROJECT NAME: LOAN PREDICTION

***Problem Description:-***

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

***Independent Variables:-***

1. *Loan ID - This refer to the unique identifier of the applicant's affirmed purchases*
2. *Gender - This refers to either of the two main categories (male and female) into which applicants are divided on the basis of their reproductive functions*
3. *Married - This refers to applicant being in a state of matrimony*
4. *Dependents - This refers to persons who depends on the applicants for survival*
5. *Education - This refers to number of years in which applicant received systematic instruction, especially at a school or university*
6. *Self-employed - This refers to applicant working for oneself as a freelancer or the owner of a business rather than for an employer*
7. *Applicant Income - This refers to disposable income available for the applicant's use under State law.*
8. *CoapplicantIncome - This refers to disposable income available for the people that participate in the loan application process alongside the main applicant use under State law.*
9. *Loan Amount - This refers to the amount of money an applicant owe at any given time.*
10. *Loan Amount Term - This refers to the duration in which the loan is availed to the applicant*
11. *Credit History - This refers to a record of applicant's ability to repay debts and demonstrated responsibility in repaying them.*
12. *Property Area - This refers to the total area within the boundaries of the property as set out in Schedule.*

***Dependent Variable (Target Variable):-***

*13. Loan Status - This refers to whether applicant is eligible to be availed the Loan requested.*

*You have to build a model that can predict whether the loan of the applicant will be approved (Loan status) or not on the basis of the details provided in the dataset.*

## Data Analysis

*Below are the steps for Exploratory Data Analysis*

***Variable Analysis;-***

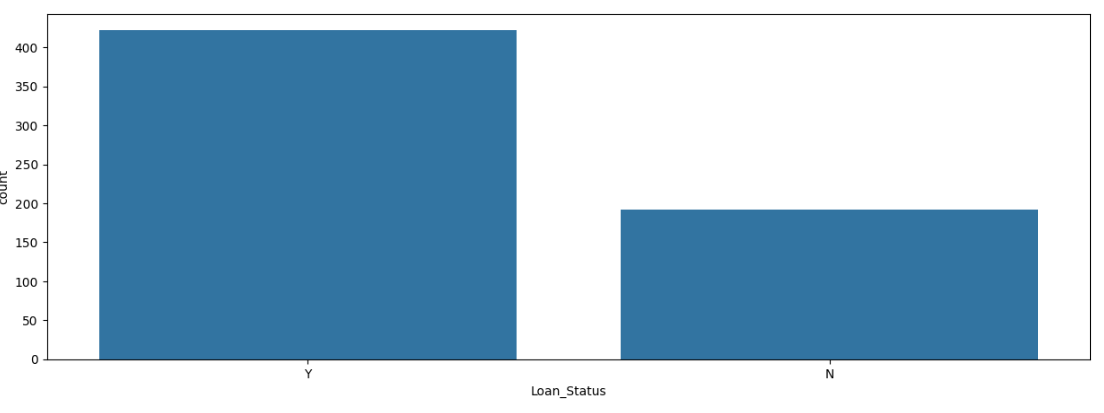
***1. Target Variable – Loan Status***

* ***Loan Status***

*Y 422*

*N 192*

*Name: count, dtype: int64*

**

***Predictor Variables:-***

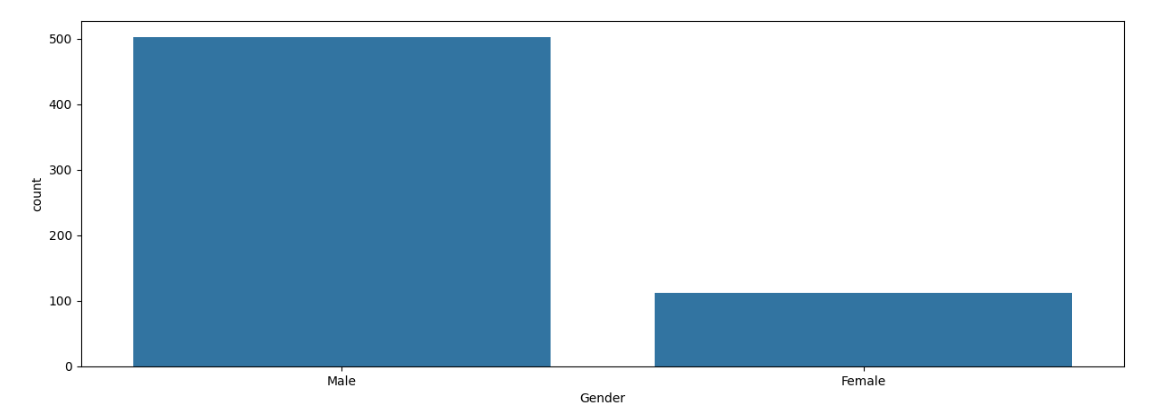
*There are 3 types of Independent Variables: Categorical, Ordinal & Numerical. Here the categorical variables are*

* ***Gender***

*Male 502*

*Female 112*

*Name: count, dtype: int64*

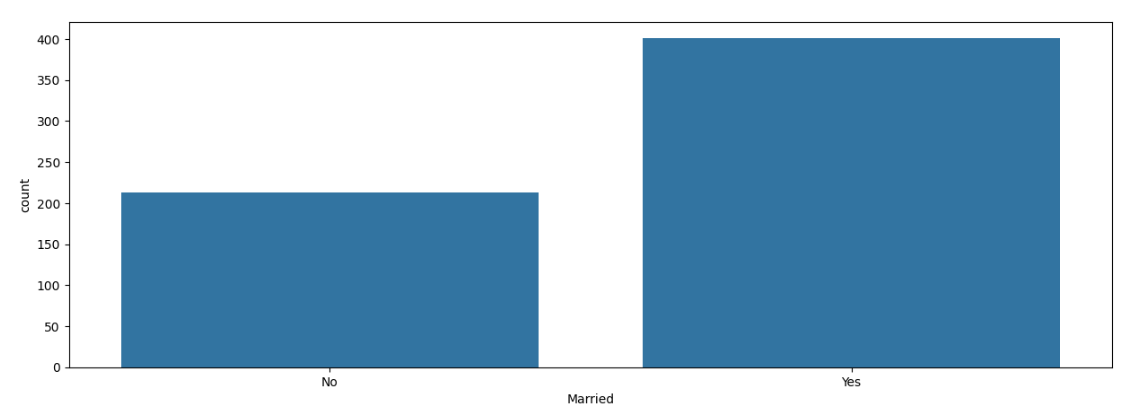
**

* ***Married***

*Yes 401*

*No 213*

*Name: count, dtype: int64*

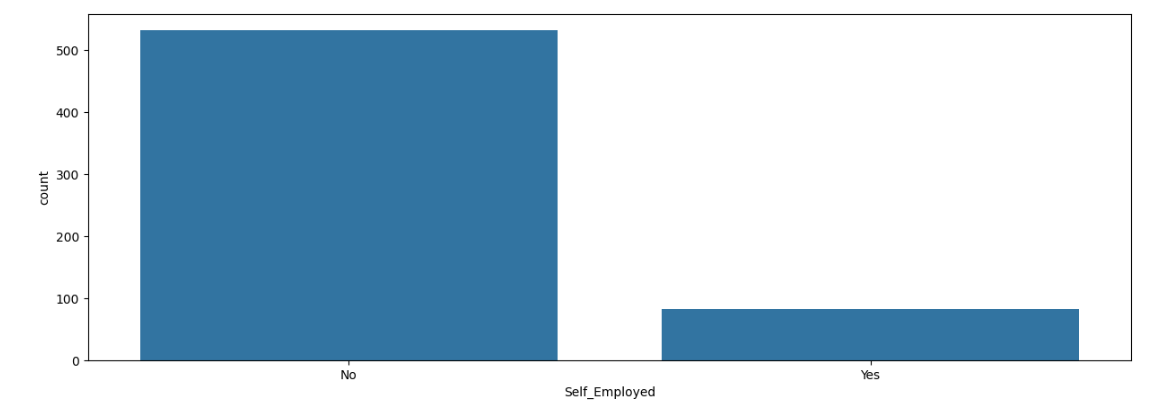
**

* ***Self Employed***

*No 532*

*Yes 82*

*Name: count, dtype: int64*

**

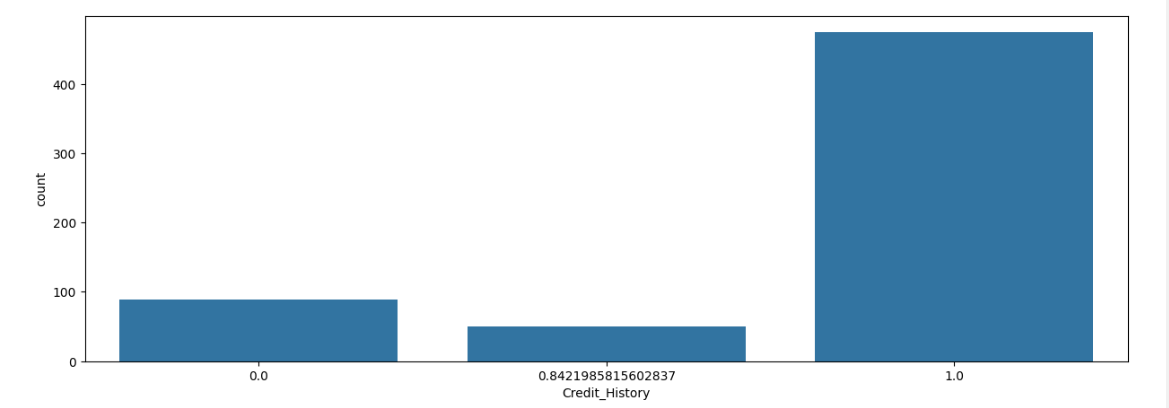
* ***Credit History***

*1.000000 475*

*0.000000 89*

*0.842199 50*

*Name: count, dtype: int64*

**

***Ordinal Features:-***

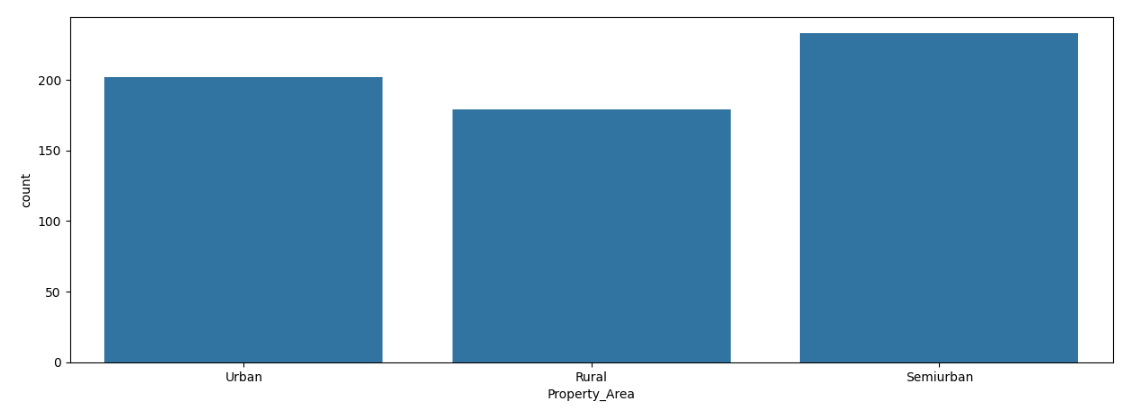
* ***Property Area***

*Semi urban 233*

*Urban 202*

*Rural 179*

*Name: count, dtype: int64*

**

* ***Number of Dependents***

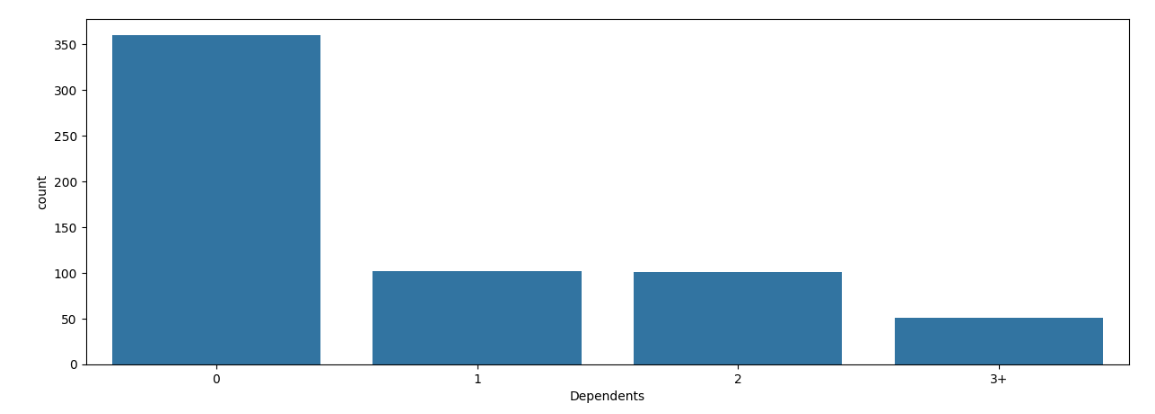
*0 360*

*1 102*

*2 101*

*3+ 51*

*Name: count, dtype: int64*

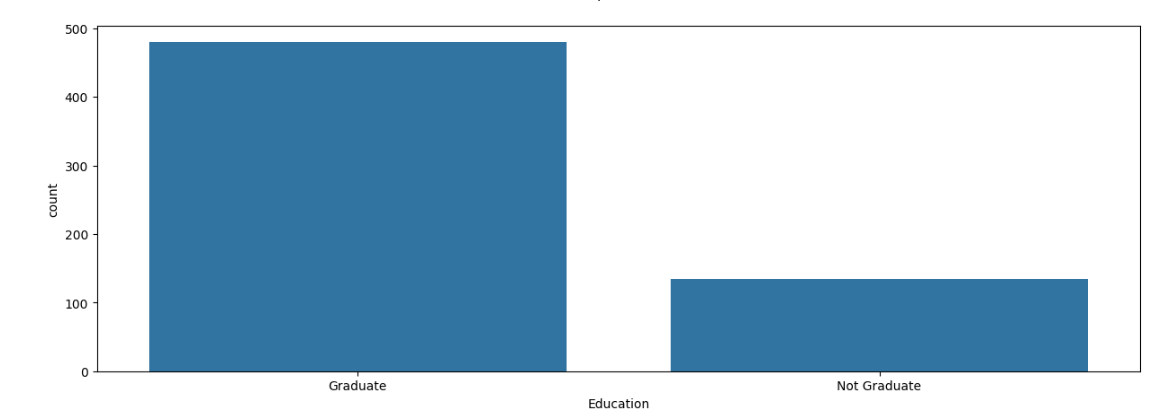
**

* ***Education***

*Graduate 480*

*Not Graduate 134*

*Name: count, dtype: int64*

**

***Numerical Features:-***

* *The Applicant’s Income*
* *The Co – Applicant’s Income*

## EDA Concluding Remarks

* *80% of loan applicants are male in the training dataset.*
* *Nearly 70% are married.*
* *Nearly 85-90% loan applicants are self-employed.*
* *The loan has been approved for more than 65% of applicants.*
* *Almost 58% of the applicants have no dependents.*
* *Highest number of applicants are from Semi urban areas, followed by urban areas.*
* *Around 80% of the applicants are Graduate.*

## pre – processing pipeline

***A. Data Cleaning***

***a. Finding the Data type of variables***

Loan\_ID object

Gender object

Married object

Dependents object

Education object

Self\_Employed object

ApplicantIncome int64

CoapplicantIncome float64

LoanAmount float64

Loan\_Amount\_Term float64

Credit\_History float64

Property\_Area object

Loan\_Status object

dtype: object

***b. Handling Missing / Null values***

Loan\_ID 0

Gender 13

Married 3

Dependents 15

Education 0

Self\_Employed 32

ApplicantIncome 0

CoapplicantIncome 0

LoanAmount 22

Loan\_Amount\_Term 14

Credit\_History 50

Property\_Area 0

Loan\_Status 0

dtype: int64

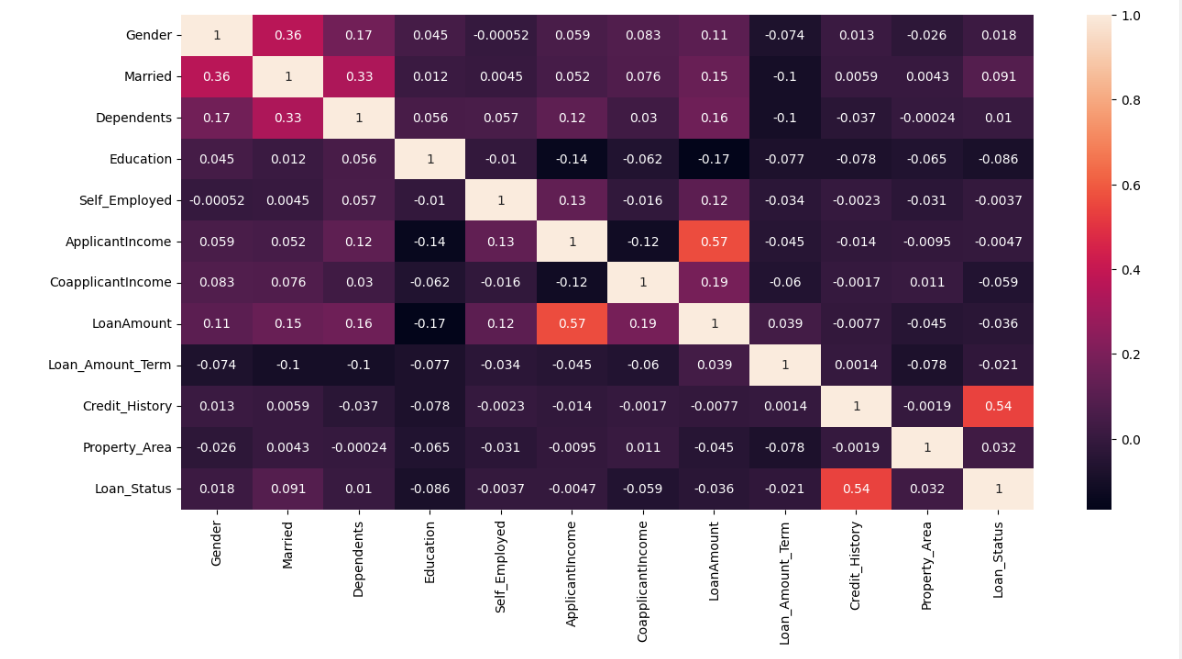
*Now we know that the missing / null values are replaced by mean or mode of the remaining values depending on the type of data whether float or object type. For float type it is mean, and for object type it is mode. Thus handled the missing values.*

***c. Dummy Variables for categorical variables:-***

*Now before model building, we need to encode the object data type variables in 0 to 1 so as to work on the data. Also, we need to drop the columns we don’t need, like here the loan ID is unique thus dropped it off.*

***d. Correlation between Quantitative Variables:-***

*Correlation is a statistical term describing the degree to which two variables move in coordination with one another. If the two variables move in the same direction, then those variables are said to have a positive correlation. If they move in opposite directions, then they have a negative correlation. Here we see the below graph.*

**

Loan\_Status 1.000000

Credit\_History 0.540483

Married 0.091478

Property\_Area 0.032112

Gender 0.017987

Loan\_ID 0.011773

Dependents 0.010118

Self\_Employed -0.003700

ApplicantIncome -0.004710

Loan\_Amount\_Term -0.020974

LoanAmount -0.036416

CoapplicantIncome -0.059187

Education -0.085884

Name: Loan\_Status, dtype: float64

*We see that the factors affecting the loan status are Credit History, Married or not followed by property area and gender. But the most important positively affecting factor is Credit History as Credit History is important in banks to give any person loan. A credit history is the record of how a person has managed his or her credit in the past, including total debt load, number of credit lines, and timeliness of payment. Lenders look at a potential customer’s credit history to decide whether or not on gender as seen in previous scenarios that males are less likely to save and return loan amount than women. Also, the negative factors affecting the loan application status is the education level and applicant income.*

## Building machine learning models

*For model building, we first need to separate the dependent and independent variables x and y. Here the factors are the independent variables and Loan\_Status is the dependent variable as discussed thus dropping off Loan\_Status and proceeding.*

***a. Checking the skewness:-***

*Skewness is a measure of the symmetry of a distribution. The highest point of a distribution is its mode. The mode marks the response value on the x-axis that occurs with the highest probability. A distribution is skewed if the tail on one side of the mode is fatter or longer than on the other: it is asymmetrical.*

*Is an asymmetrical distribution a negative skew indicates that the tail on the left side in longer than on the right side (left-skewed), conversely a positive skew indicates the tail on the right side is longer than on the left (right-skewed).Asymmetric distributions occur when extreme values lead to a distortion of the normal distribution.*

*In the data we get below skewness:*

CoapplicantIncome 7.491531

ApplicantIncome 6.539513

LoanAmount 2.726601

Self\_Employed 2.159796

Education 1.367622

Dependents 1.015551

Property\_Area -0.066196

Married -0.644850

Gender -1.648795

Credit\_History -1.963600

Loan\_Amount\_Term -2.389680

dtype: float64

*We see that the data is skewed and to bring the skewneww in the range of (-0.5,0.5) using power transform*

Self\_Employed 2.159796

Education 1.367622

Dependents 0.441404

Loan\_Amount\_Term 0.389827

LoanAmount 0.024376

ApplicantIncome -0.092946

CoapplicantIncome -0.145646

Property\_Area -0.158267

Married -0.644850

Credit\_History -1.599912

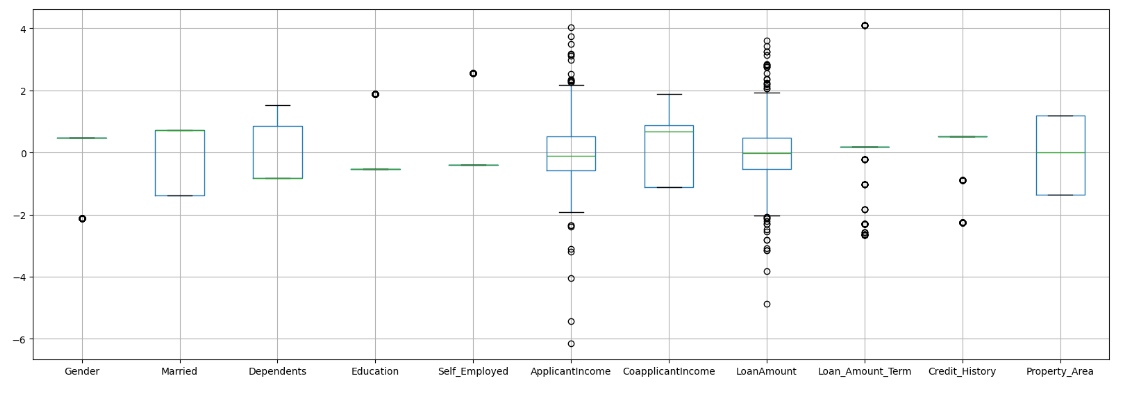
Gender -1.648795

dtype: float64

*Now we see that the data is not skewed, and we have removed the skewness. Thus, we can now proceed.*

***b. Handling Outliers:-***

*Now we need to check whether outliers are present in the data or not. For that we need to check if the z-value / z-score of all factors is exceeding the range (-3, 3). As we want the data to be normally distributed in the range of (-3, 3) as the data be in the 99% domain. Thus this will be the best data to work with.*

**

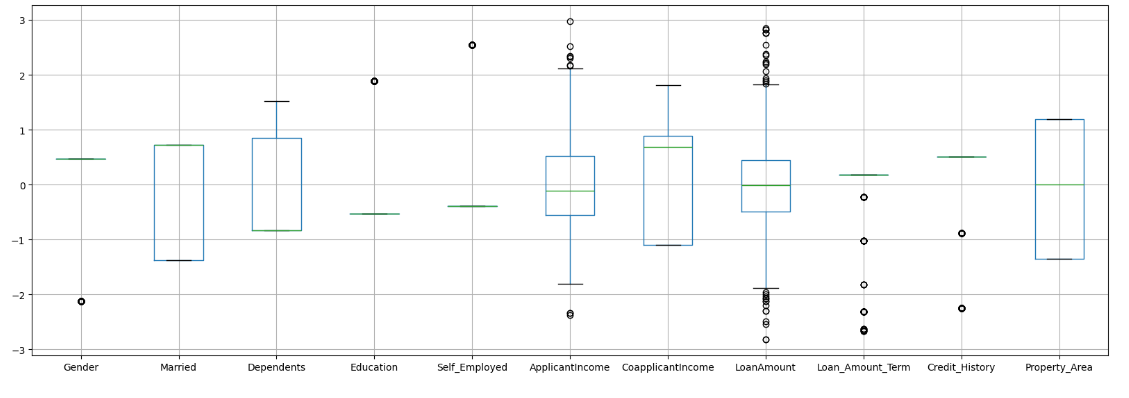
*Here we see there are a few outliers, thus removing them and bringing the data in the range of (-3, 3)*

**from** scipy.stats **import** zscore

z **=** np**.**abs (zscore (new\_df))

new\_df **=** new\_df [(z**<**3)**.**all (axis**=**1)]

*Using the above code, we were successfully able to remove the outliers and we get the below range of data.*

**

*As we see there are no outliers, thus can proceed with modelling*

***c. Regression Model:-***

*To predict the Loan Application Status we need to do Logistic Regression modelling as the value of Loan\_status is 0 or 1 thus binary mapping depending whether the loan application of a person gets approved or not*

*Running the regression, we get best accuracy as 0.8974358974358975 on Random state 370.*

***d. Train\_test\_split:-***

*The data we use is usually split into training data and test data. The training set*

*Contains a known output and the model learns on this data in order to be generalize to other data later on. We have the test dataset (or subset) in order to test our*

*Model’s prediction on this subset*

*Using the random state, we divide the training and testing data in the ratio of 0.8 and0.2 meaning 20% of the data is the testing data and training data is 80%*

*Thus,*

*Train data = 465 records*

*Test data = 117 records*

*As total records are 595 as we removed the outliers.*

***Now running the Regression Models, below are the accuracy and result we get***

***1. Logistic Regression:-***

*Logistic Regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic Regression (or logit regression) is estimating the*

*Parameters of a logistic model (a form of binary regression)*

***Logistic Regression***

**Accuracy 88.03418803418803**

[[22 8]

[ 6 81]]

precision recall f1-score support

0.0 0.79 0.73 0.76 30

1.0 0.91 0.93 0.92 87

accuracy 0.88 117

macro avg 0.85 0.83 0.84 117

weighted avg 0.88 0.88 0.88 117

***2. Decision Tree Classification***

**Accuracy 76.06837606837607**

[[20 10]

[18 69]]

precision recall f1-score support

0.0 0.53 0.67 0.59 30

1.0 0.87 0.79 0.83 87

accuracy 0.76 117

macro avg 0.70 0.73 0.71 117

weighted avg 0.78 0.76 0.77 117

***3. Random Forest Classification***

**Accuracy 88.88888888888889**

[[23 7]

[ 6 81]]

precision recall f1-score support

0.0 0.79 0.77 0.78 30

1.0 0.92 0.93 0.93 87

accuracy 0.89 117

macro avg 0.86 0.85 0.85 117

weighted avg 0.89 0.89 0.89 117

***4. Support vector classification***

**Accuracy 88.88888888888889**

[[21 9]

[ 4 83]]

precision recall f1-score support

0.0 0.84 0.70 0.76 30

1.0 0.90 0.95 0.93 87

accuracy 0.89 117

macro avg 0.87 0.83 0.85 117

weighted avg 0.89 0.89 0.89 117

*Here we see that the best classifier is Logistic Regression with 88% accuracy*

***e. Cross Validation:-***

*Cross validation is primarily used in applied machine learning to estimate the skill of a machine leaning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.*

*We got below CV scores.*

* *Cross Validation score of Logistics Regression Model is* ***0.7921161214264664***
* *Cross Validation Score of Random Forest Classifier is* ***0.7852785145888593***
* *Cross Validation Score of Decision Tree Classifier is* ***0.6976274683171235***
* *Cross Validation Score of Support Vector Classifier is* ***0.8024462127910403***

*We see that the best accuracy is given by Logistic Regression of 88% and cross validation score is 79% which may be because the data is slightly overfitted, also we*

*Have little data to test and train.*

***d. Hyper Parameter Testing:-***

*In machine learning, hyperparameter optimization or tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a*

*Parameter whose value is used to control the learning process. By contrast, the*

*Values of other parameters (typically node weights) are learned.*

*There are two types of search used for hyper parameters testing: Research and*

*Randomized search CV.*

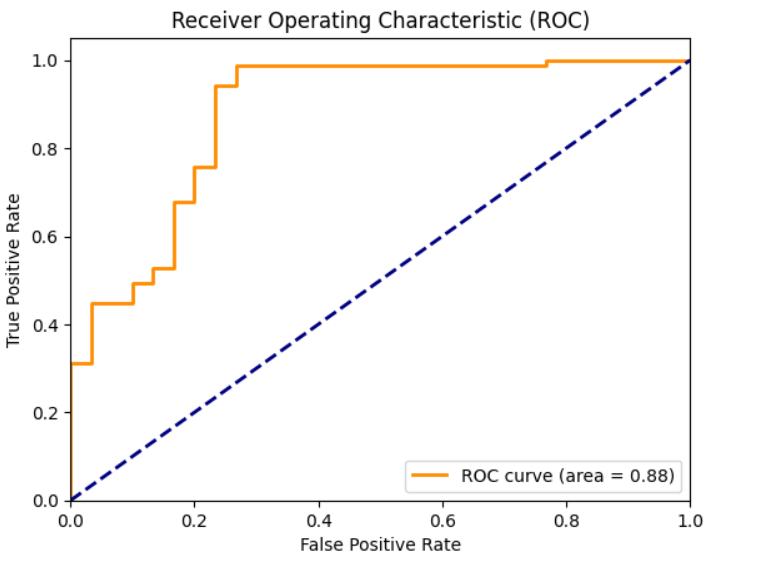
*With small dataset and lots of resources, Grid Search will produce accurate results.*

*However, with large data sets, the high dimensions will significantly slow down*

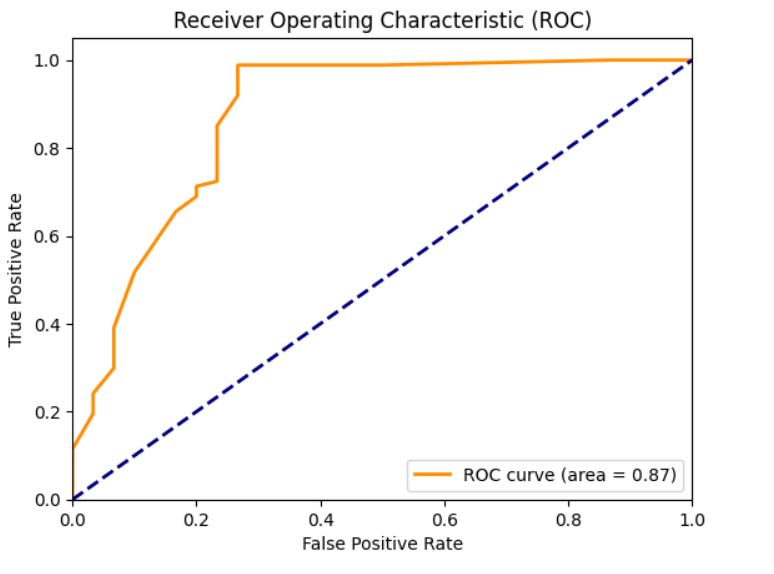
*Computation time and be very expensive. In this instance, it is advised to use*

*Grid Search CV as the data is too little thus better accuracy.*

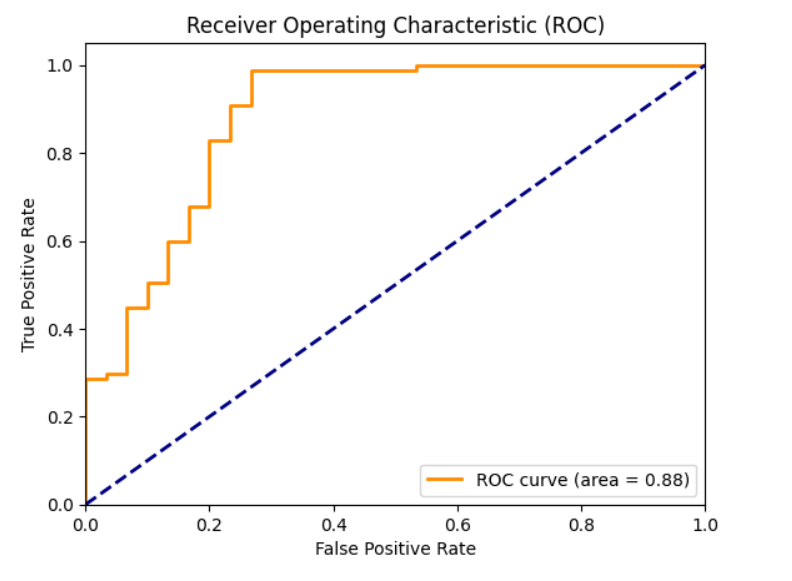
* *Used GridSearchCV for RandomForestClassifier – Accuracy is 92.3%*

**

* *Used GridSearchCV for DecisionTreeClassifier – Accuracy is 87.17%*

******

* *Used GridSearchCV for Logistic Regression – Accuracy is 92.3%*

******

*Model accuracy is 92.30% through Logistic Regression and using hyper parameter*

*testing also the accuracy is same. But for Random Forest Classifier and Decision*

*Tree Classifier the accuracy has increased and AUC accuracy is 87 % using the*

*Logistic Regression and 88% using the Random Forest Classifier and Decision Tree Classifier.*

## Concluding remarks

*We did Exploratory information Investigation on the highlights of this dataset and*

*saw how each include is distributed.*

*We dissected each variable to check in the event that information is cleaned and*

*ordinarily distributed. We cleaned the information and evacuated NA values. We*

*tried to find the correlation and based on the outcomes, we accepted whether or not*

*there’s a relation between the Loan Application Status and the other factors, we saw*

*that the Credit History is the most important factor positively related followed by the*

*Marital Status, which means the higher the credit history the higher chances of loan application to get rejected. And the Education is negatively related which means the higher the candidate’s education the lower chances of his/her loan application to get rejected. We used the logistic regression model to classify the loan application status of whether they will be approved or not or whether the loan will be given or not*

*depending on the various factors.*

*We got a very good model with a 92.30% accuracy which is very good. Just a*

*Disadvantage that the data available to train the model is too less and thus needs to be improved. Thus, this model will help the banks to identify the loan defaulters. But*

*to improve the model accuracy and the cross-validation score, we need to have*

*more data to train and test and then we can move to conclusion.*

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\*\*\*\*\*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*