Loan Prediction Problem

NOTE: There are 3 files in the submission folder-

a) **NIRA_Assignment_Ananlysis.ipynb** which contains the Analysis and Model building parts in detail. My Notebook is self Explanatory, It will give a proper insight in a step by step fashion where anyone could understand the logical steps.
b) **NIRA_Assignment_function.ipynb** which contains the use of Logistic Regression Model and all the steps are given in a single function call.
c) **NIRA_Assignment_Report.pdf** which contains the Report in pdf format as asked.

---

1. Understanding the Problem Statement

Here, we are building an automated model which will reduce the time taken by the bank to manually validate the customer eligibility for loan based on customer information, and will also ensure the Bank gets the money back.

2. Generating Hypothesis

Now, after carefully studying the problem and looking at the variable/column features in the dataset, we can come up with some of our own assumptions/hypotheses. It is basically thinking of the features that might affect the output, though, our assumption may or may not be always correct in reality. Let's talk about some of them below:

- Loan Amount : Higher the Loan Amount the lesser is the chance to repay back and vice-versa.

- Salary : Higher the salary, higher the chances of the loan repayment and hence Loan approved chances should be high.
- Loan_Amount_Term : For a lower amount with lower term, chances to approve the loan should be higher.
- Credit_History : If the previous history is good means that higher the chance of loan approval.

3. Exploratory Data Analysis

a. VARIABLE IDENTIFICATION

Here, we will try to identify the **Input** data and the **Target Variable**(Output Data) and their **data-types**.

There are 614 rows and 13 columns in the dataset.

- Categorical Variables:Gender, Married, Dependents, Education, Self_Employed, Property_Area, Loan_Status, Credit_History.
- Continuous Variables: { ApplicantIncome (int64 format - integer variable)}, {CoapplicantIncome, LoanAmount, Loan_Amount_Term (float64 format - Numerical values having decimal values involved)}

Target Variable : Loan_Status.

As, we can clearly see that this is a Classification Problem where we have to predict whether the loan would be approved or not on the basis of customer information. Since. there are two classes as Y/N this is a **Binary Classification Problem**.

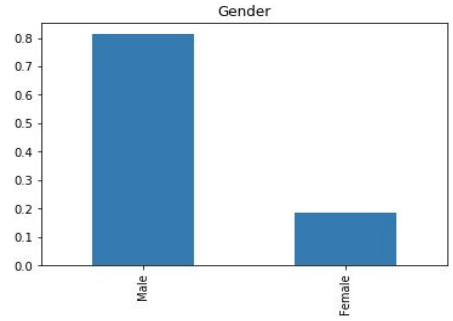
b. UNIVARIATE ANALYSIS

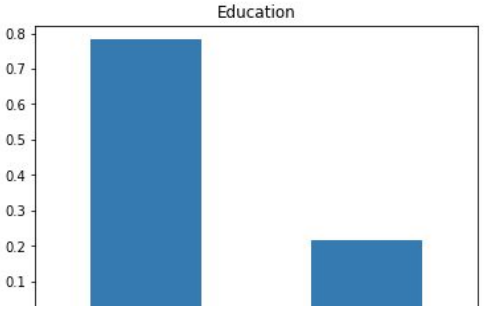
Plotting Bar Graph for Categorical variables and Histogram for Continuous Variables distribution observations.

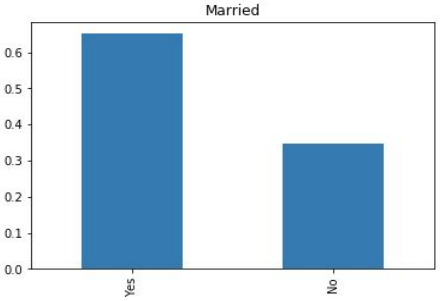
- 80% observations for Gender were Male in the form.
- 65% of the people who applied for Loan were Married.
- Roughly 58% have "0" dependents, 18% have "1" dependents, 18% have "2" dependents, 6% have "3+" dependents in the dataset.

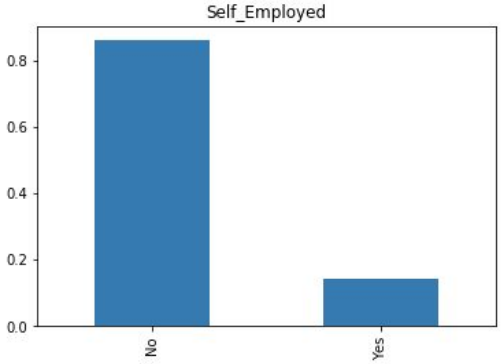
- Around 78% of the people are Graduates and 22% are non-graduates.
- Around 15% of people are self-employed.
- Majority of people (~38%) have their property area in semi urban area, followed by nearly 33% in Urban areas and 29% in Rural areas.
- Roughly 84% people fall in the "1" credit_history category which might mean they have a positive credit_history (which is denoted here by 1).
- Applicant Income, CoApplicantIncome showed a right skewed distribution and the profusion of outliers. It can be because of income differences of individuals when compared on Education basis, as Educated People have higher income.
- LoanAmount and LoanAmountTerm, showed profusion of outliers.

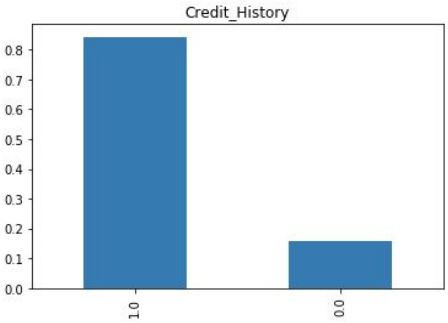
- Target Variable (Loan_Status) : Roughly 69% observations fall in "Y"

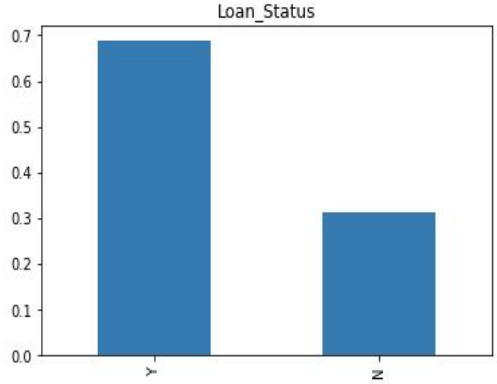




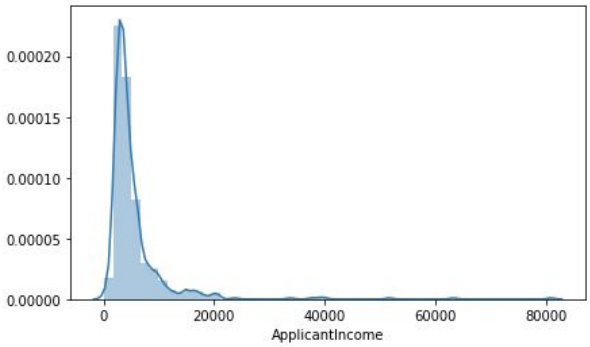


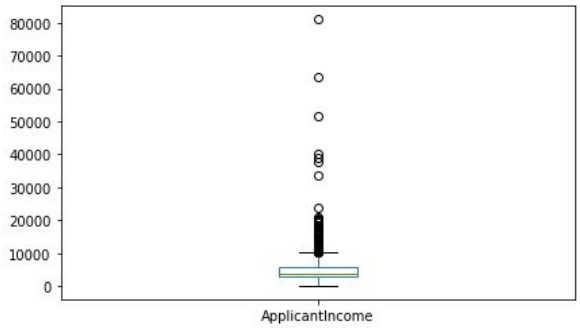


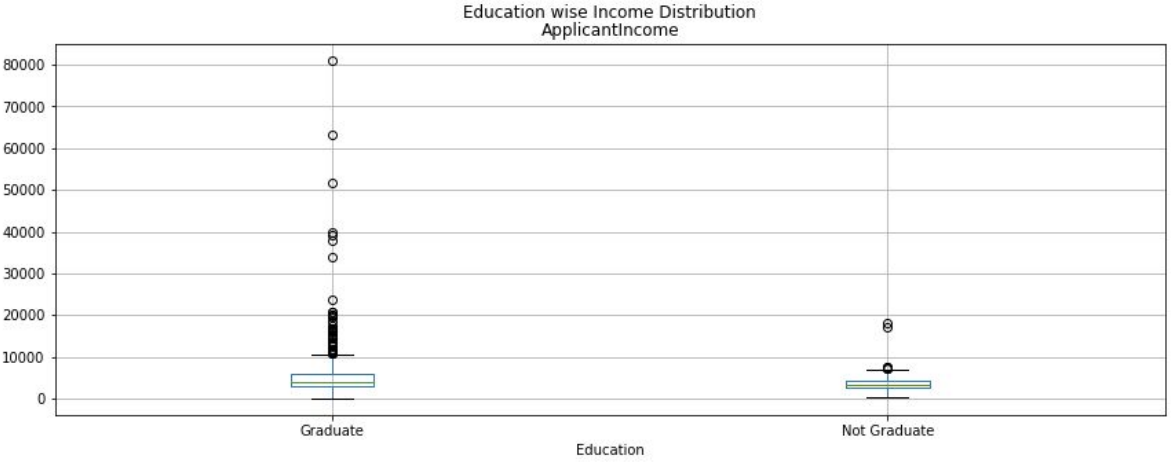


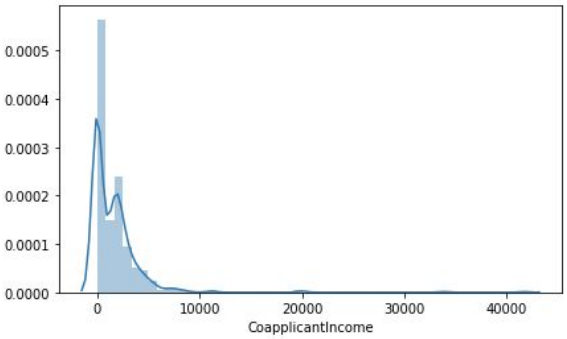


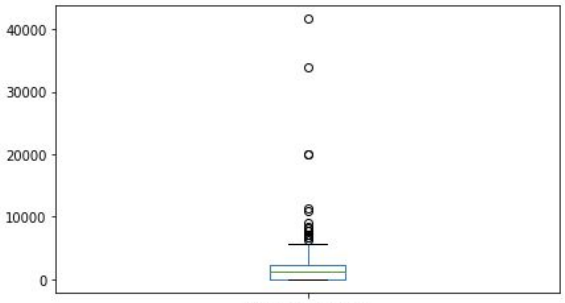
seg.

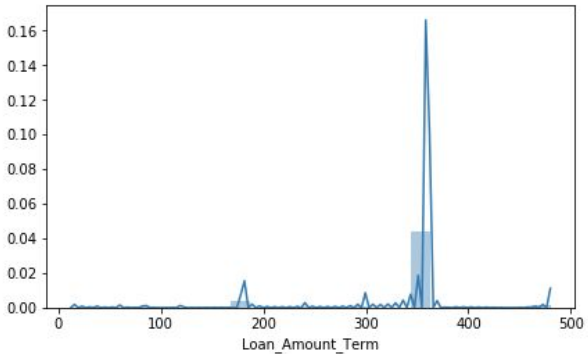

0.00020
0.00015
0.00010
0.00005
0.00000
0
20000
40000
60000
80000
ApplicantIncome

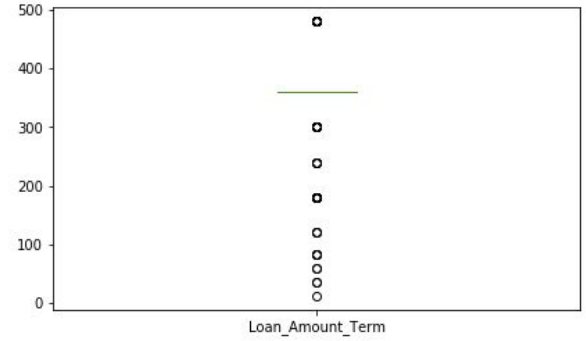

80000
70000
60000
50000
40000
30000
20000
10000
0
ApplicantIncome

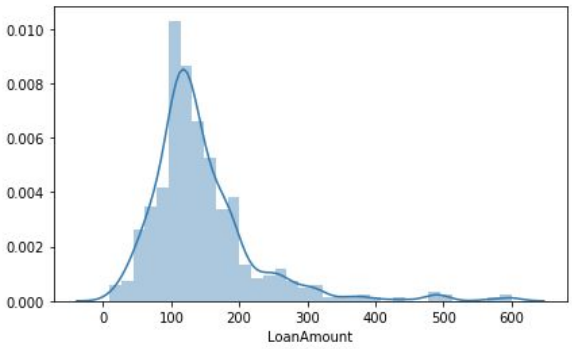

Education wise Income Distribution
ApplicantIncome
80000
70000
60000
50000
40000
30000
20000
10000
0
Graduate
Not Graduate
Education

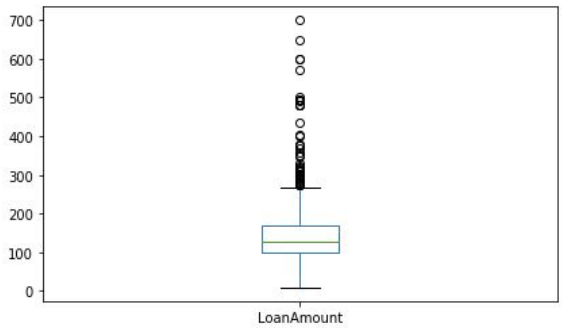

0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
0
10000
20000
30000
40000
CoapplicantIncome

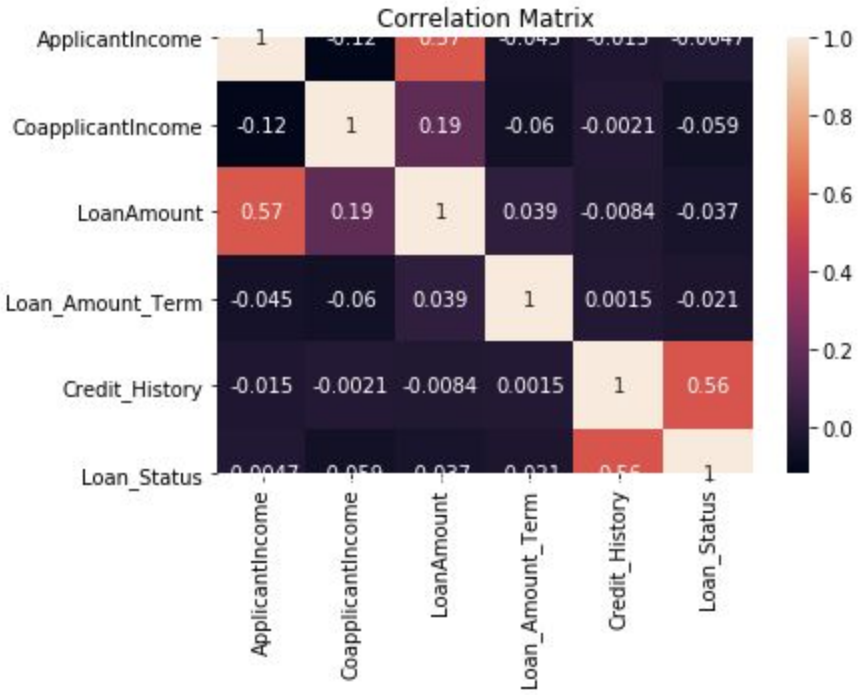

40000
30000
20000
10000
0
CoapplicantIncome


0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
100
200
300
400
500
Loan_Amount_Term

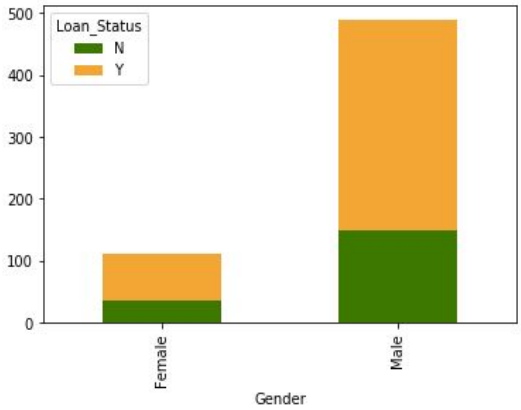

500
400
300
200
100
0
Loan_Amount_Term

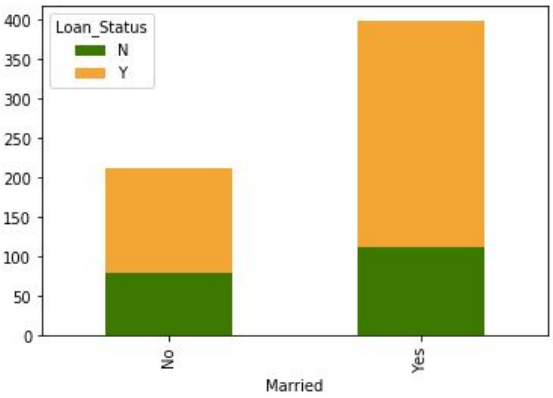


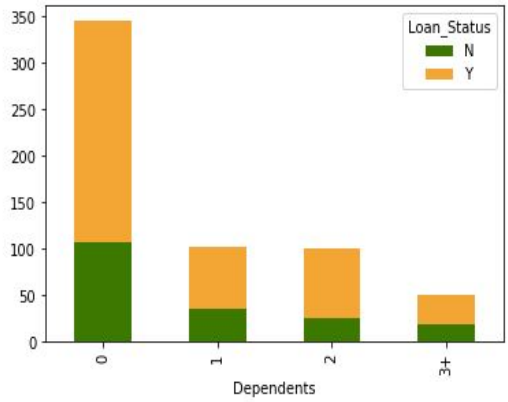


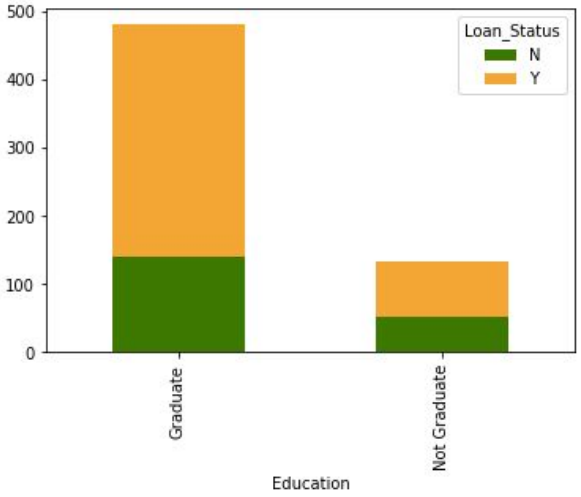


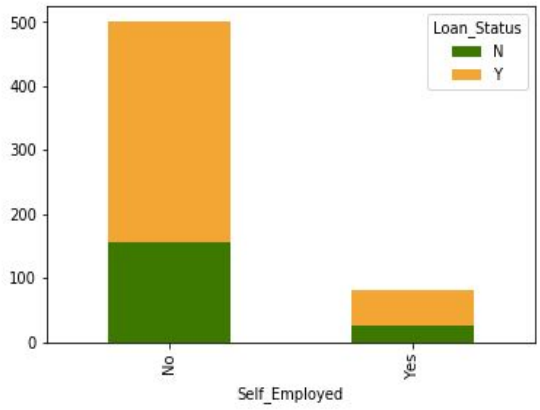
c. BIVARIATE ANALYSIS

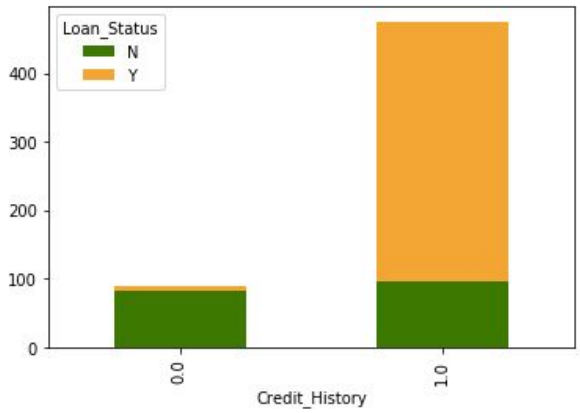

Loan_Status
N
Y
500
400
300
200
100
0
Female
Male
Gender

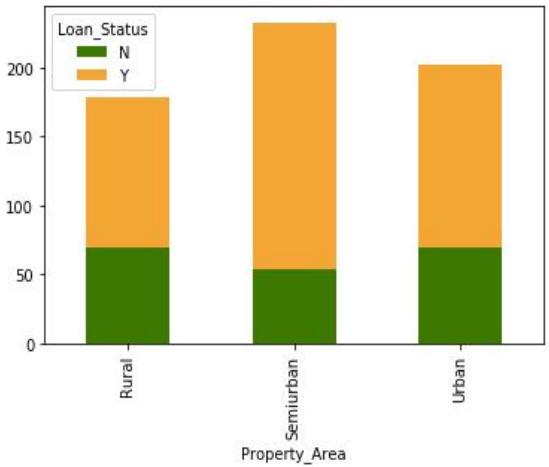

Loan_Status
N
Y
400
350
300
250
200
150
100
50
0
No
Yes
Married


Loan_Status
N
Y
350
300
250
200
150
100
50
0
0
1
2
3+
Dependents


Loan_Status
N
Y
500
400
300
200
100
0
Graduate
Not Graduate
Education


Loan_Status
N
Y
500
400
300
200
100
0
No
Yes
Self_Employed


Loan_Status
N
Y
400
300
200
100
0
0.0
1.0
Credit_History


Loan_Status
N
Y
200
150
100
50
0
Rural
Semiurban
Urban
Property_Area

I tried to identify the relation if any between the target variable and input variables one by one.

- Males have high probability of Laon Approval.
- Married people have higher probability of Loan Approval
- People without any Dependent have higher chances of Loan Approval.
- Graduates have higher chances of Loan Approval.
- Self_Employed people have lower chance of Loan Approval.
- If the credit history is good, then chances of Loan Approval are higher.
- People demanding Loan for house in in SemiUrban area have higher chance of Loan Approval.
- High Correlation between:(ApplicantIncome - LoanAmount) and (Credit_History - Loan_Status).

4. Data Preprocessing Methods

e. MISSING VALUES TREATMENT

I was able to find Missing Values. Though the percentage of missing values in the dataset was low.

Credit_History: 8.143322%

Self_employed: 5.211726%

Loan_Amount: 3.583062%

Dependents: 2.442997%

Loan_Amount_Term: 2.280130%

Gender: 2.117264%

Married: 0.488599

- Gender, Married, Dependents, Self_Employed, Credit_History being categorical were filled by Mode Values in place of Missing values.
- Regarding the Loan_Amount_Term variable we can clearly see in the plot that "360" value occurs mostly and the distribution is also not normal. So, we fill it with "360" mode value.

- As for the Loan_Amount variable the plot clearly shows that it is not a normally distributed curve. Also, it has profusion of outliers, so Mean is not a good way to fill missing values because it is highly affected by presence of outliers and hence we will use Median to fill missing values.

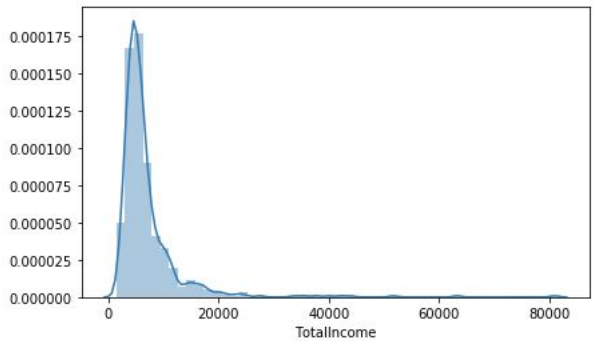
	Total	Percent
Credit_History	50	8.143322
Self_Employed	32	5.211726
LoanAmount	22	3.583062
Dependents	15	2.442997
Loan_Amount_Term	14	2.280130
Gender	13	2.117264
Married	3	0.488599
Loan_Status	0	0.000000
Property_Area	0	0.000000
CoapplicantIncome	0	0.000000
ApplicantIncome	0	0.000000
Education	0	0.000000
Loan_ID	0	0.000000

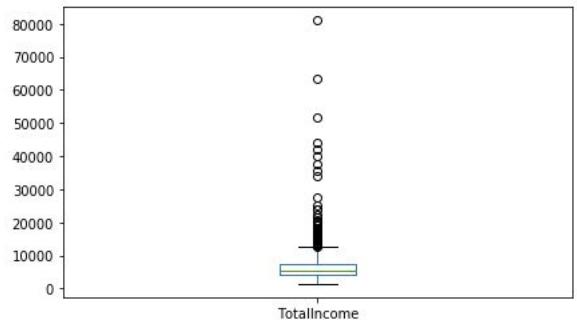
f. OUTLIER TREATMENT

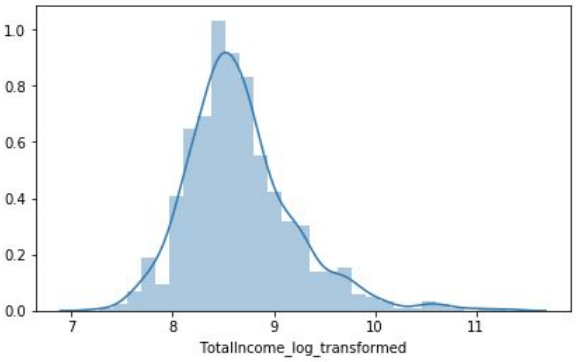
Treated Outliers by log transformations as the distribution was Right Skewed, e.g. for LoanAmount. It does not affect smaller values to that degree but reduces extreme values and makes the curve more like a Normal Distribution Curve.

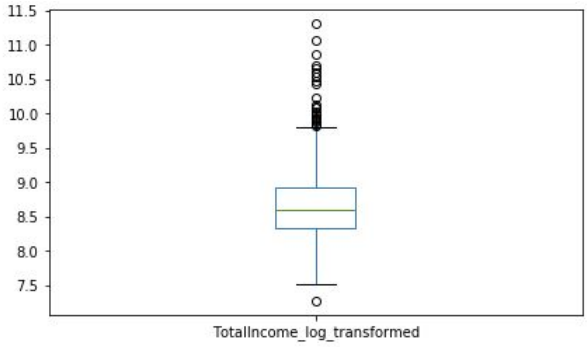
g. FEATURE ENGINEERING

- Feature1 - Looking at the ApplicantIncome, and CoapplicantIncome columns. We can introduce a new feature from these two and make a new feature called TotalIncome. General assumption can be higher the TotalIncome, higher the Loan Approval chances.i.e. TotalIncome = ApplicantIncome + CoApplicantIncome







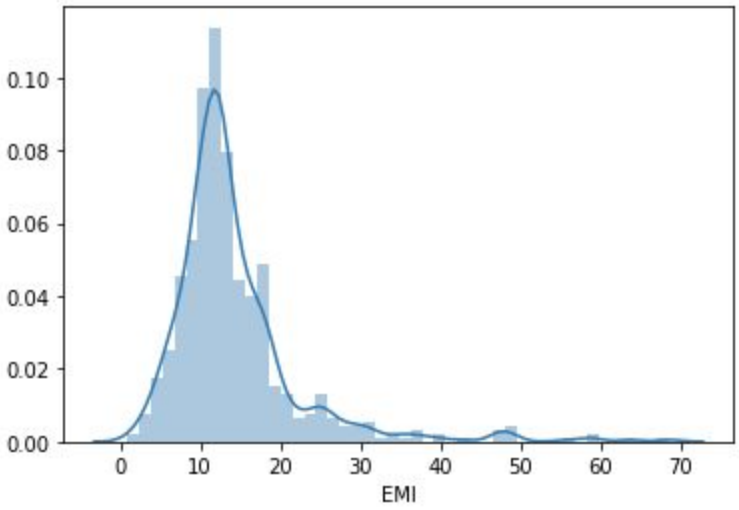


- Feature2 - EMI i.e. calculated as -

  A = P*r*(1+r)^N;

  B = (1+r)^(N-1);

  EMI = A/B;



P=Loan Amount, N Loan Amount Term, r = rate of interest (taking 9% here)

- Feature3 - Balanace_Income (after paying EMI) i.e. TotalIncome - EMI

h. Encoding the Data

Converted the Categorical Variables into Numerical forms so as to be able to feed to the Model.

i. Splitting the Dataset for Training and Testing.

The Model is splitted into x_train, y_train, x_test, and y_test.

5. Model Selection

Since the problem is a binary classification problem, I used Classification Algorithms.

Mostly we can make use of Logistic Regression, SVM in linear problems, and Decision Tree, Random Forest, etc in complex non-linear problems.

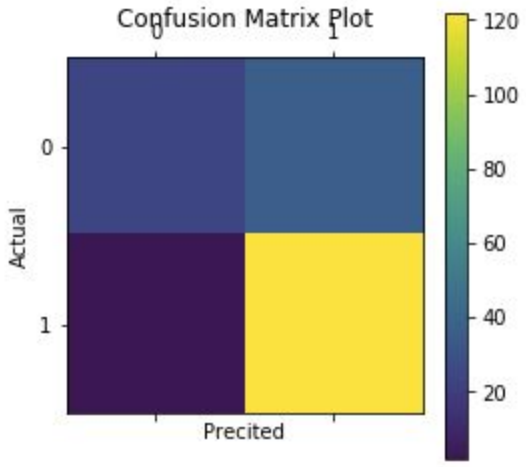
Further, after Training the Models, I tried to tune the model using Grid Search and cross validation.

Here, in our case for the time being Logistic Regression performed best and gave maximum area for AUC_ROC curve.

6. Model Evaluation.

We have various methods to assess the fit of Model. Some of the methods are :

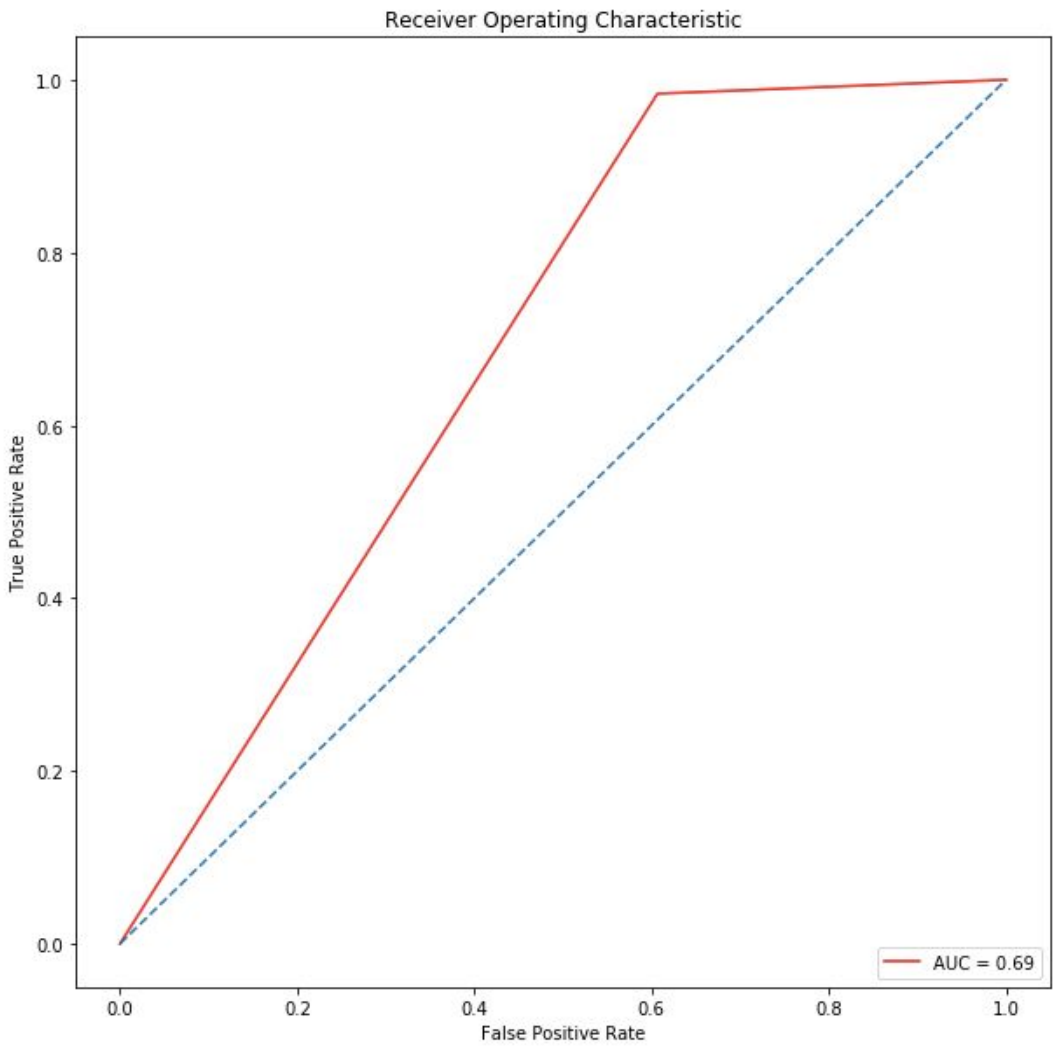
1. Confusion Matrix
2. AUC ROC Curve : Best Model considered is supposed to have maximum area under the Curve.



	precision	recall	f1-score	support
0	0.923077	0.393443	0.551724	61
1	0.767296	0.983871	0.862191	124
accuracy			0.789189	185
macro avg	0.845186	0.688657	0.706957	185
weighted avg	0.818661	0.789189	0.759821	185

ROC Area Under Curve is : 0.6886567953463776

ROC CURVE FOR LOGISTIC REGRESSION:



A skilful model will assign a higher probability to a randomly chosen real positive occurrence than a negative occurrence on average. This is known as a model having skill. Generally, skilful models are represented by curves that **bow up to the top left of the plot.**

A no-skill classifier is one that cannot discriminate between the classes and would predict a random class or a constant class in all cases. A model with no skill is represented at the point (0.5, 0.5). A model with no skill at each threshold is represented by a diagonal line from the bottom left of the plot to the top right and has an AUC of 0.5.(blue dotted line)

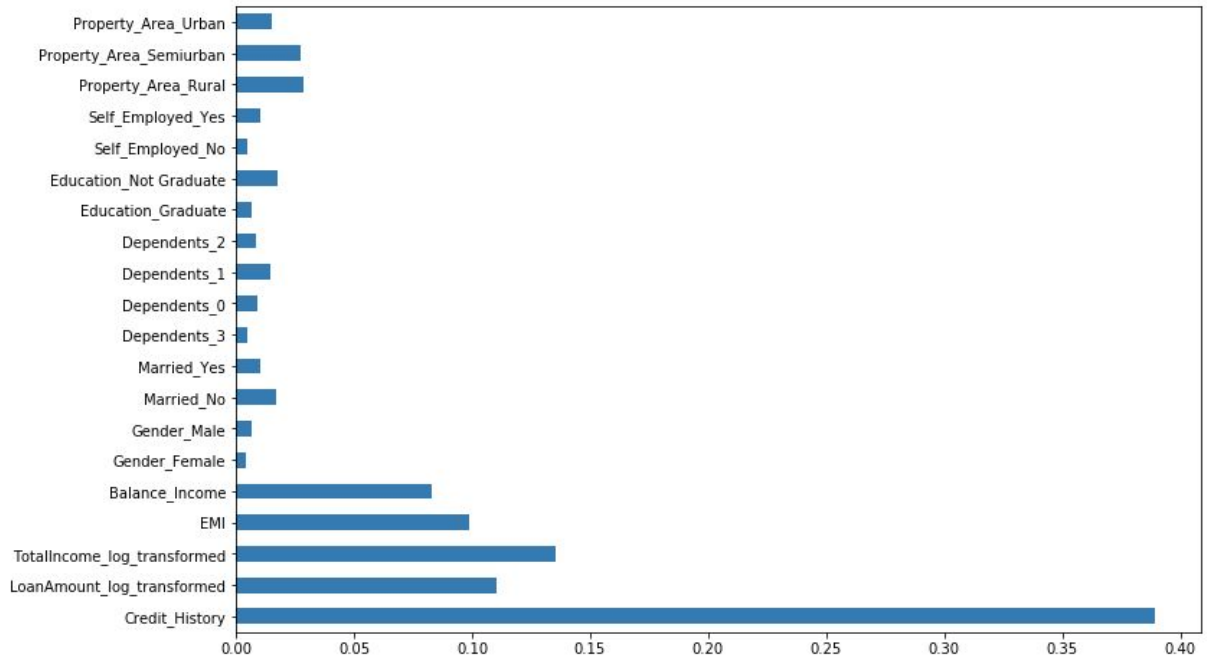
A model with perfect skill is represented at a point (0,1). A model with perfect skill is represented by a line that travels from the bottom left of the plot to the top left and then across the top to the top right.

Here, in the given Problem Statement, Logistic Regression Performed best with maximum auc_roc score of 0.69 and best accuracy of 81.82% on Grid Search CV. Though, the Predictions can be improved further.

7. Model Interpretation

Being a non-complex data and lesser order of complexity, Logistic Regression performed well on the DataSet. It comes out to be a skillful model.

Feature Importance : A common point on model interpretability is that with an increase in model complexity, model interpretability goes down at least as fast. Feature importance is a basic approach to interpret the model. It is a great way to see **what** the model is learning and whether our hypothesis is correct or not.



As, we can see that while training a Random Forest Model, we obtain the above feature importance chart. Hence, Feature Engineering helped us here definitely as they had high importance.

Further Scope of Improvement.

- In addition to the work done, further we can Tune the Model More.

- Gather More Dataset for better Learning or can perform Data Augmentation.
- Ensemble Models can be used and Fine Tuned.
- We can think of some additional features for feature engineering.

Report Ends Here.
