## **Description:**

1. Prediction of Loan Eligibility for Dream Housing Finance company is a Hackathon project on Datahack. ([link](https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/))
2. This project is implemented using Gradient Boosting Classifier.

## **Problem**

Company wants to automate the loan eligibility process based on customer detail provided while filling online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have provided a dataset to identify the customers segments that are eligible for loan amount so that they can specifically target these customers.

## **Data Source:**

[Datahack](datahack.analyticsvidhya.com)

Download Data Sets:

* [Training Data Set](https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/download/train-file)
* [Testing Data Set](https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/download/test-file)

Data Dictionary:

* Training Data:

Train file CSV containing the customers for whom loan eligibility is known as 'Loan\_Status'.

|  |  |
| --- | --- |
| Variable | Description |
| Loan\_ID | Unique Loan ID |
| Gender | Male/ Female |
| Married | Applicant married (Y/N) |
| Dependents | Number of dependents |
| Education | Applicant Education (Graduate/ Under Graduate) |
| Self\_Employed | Self employed (Y/N) |
| ApplicantIncome | Applicant income |
| CoapplicantIncome | Coapplicant income |
| LoanAmount | Loan amount in thousands |
| Loan\_Amount\_Term | Term of loan in months |
| Credit\_History | credit history meets guidelines |
| Property\_Area | Urban/ Semi Urban/ Rural |
| Loan\_Status | (Target) Loan approved (Y/N) |

* Testing Data:

Test file**:** CSVcontaining the customer information for whom loan eligibility is to be predicted.

|  |  |
| --- | --- |
| Variable | Description |
| Loan\_ID | Unique Loan ID |
| Gender | Male/ Female |
| Married | Applicant married (Y/N) |
| Dependents | Number of dependents |
| Education | Applicant Education (Graduate/ Under Graduate) |
| Self\_Employed | Self employed (Y/N) |
| ApplicantIncome | Applicant income |
| CoapplicantIncome | Coapplicant income |
| LoanAmount | Loan amount in thousands |
| Loan\_Amount\_Term | Term of loan in months |
| Credit\_History | credit history meets guidelines |
| Property\_Area | Urban/ Semi Urban/ Rural |

* Final Result Format:

The final submission file format should be in following manner. The Loan\_ID column contains unique loan IDs and the Loan\_Status column contain predicted result of loan status.

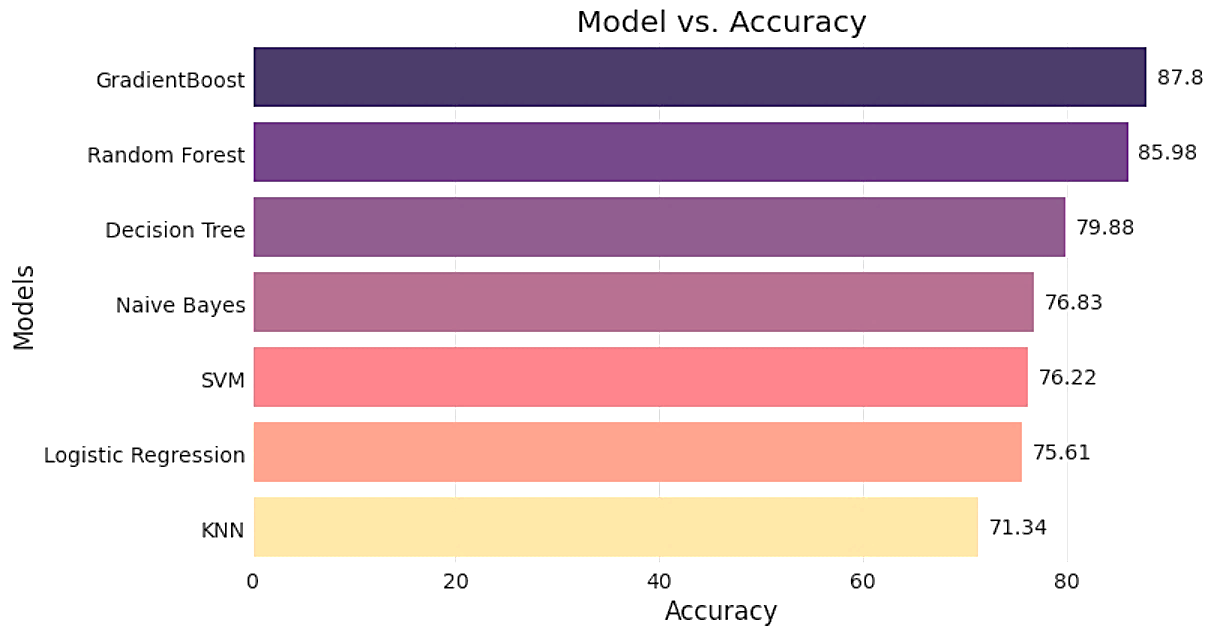
The final submission file should be in CSV format

|  |  |
| --- | --- |
| Variable | Description |
| Loan\_ID | Unique Loan ID |
| Loan\_Status | (Target) Loan approved (Y/N) |

The Project is divided into Two parts:

1. **Building Machine Learning Model.**
2. **Predicting the Outcomes of Test Dataset**.

## **Why Gradient Boosting Classifier?**

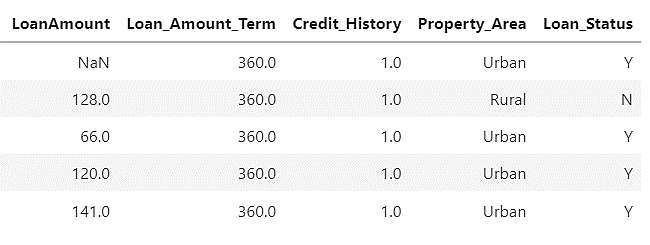
****

**As we see in above bar plot the Gradient Boosting Classifier has the highest accuracy rate hence, we choose Gradient Boosting Classifier.**

**Advantages -**

* It provides predictive scores which is better than other classifiers.
* It often provides predictive accuracy that cannot be trumped.
* Optimize on differentiable loss function.
* Provides several Hyper Parameter Tuning options that make the function fit very flexible.

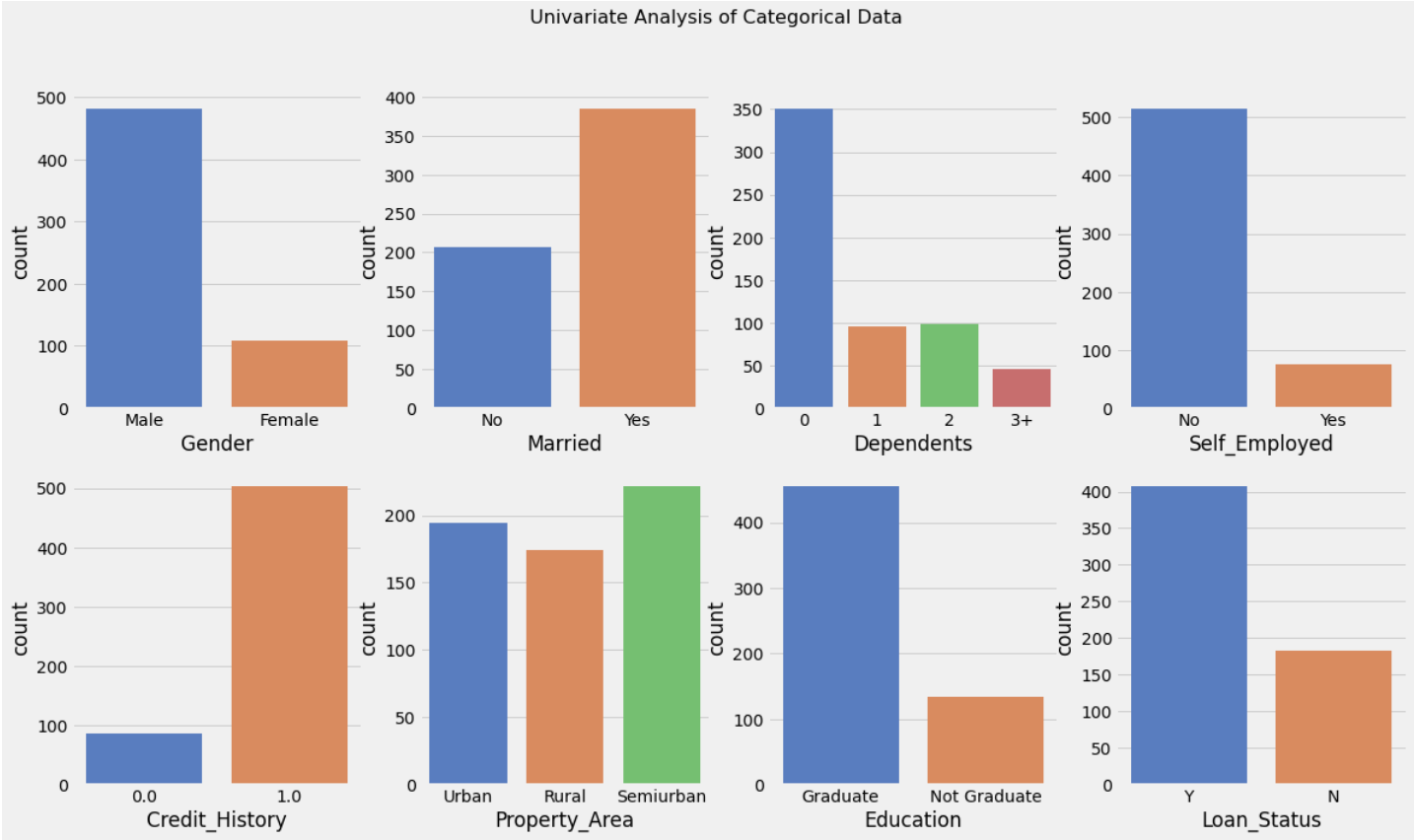
## **Input/Training Data**



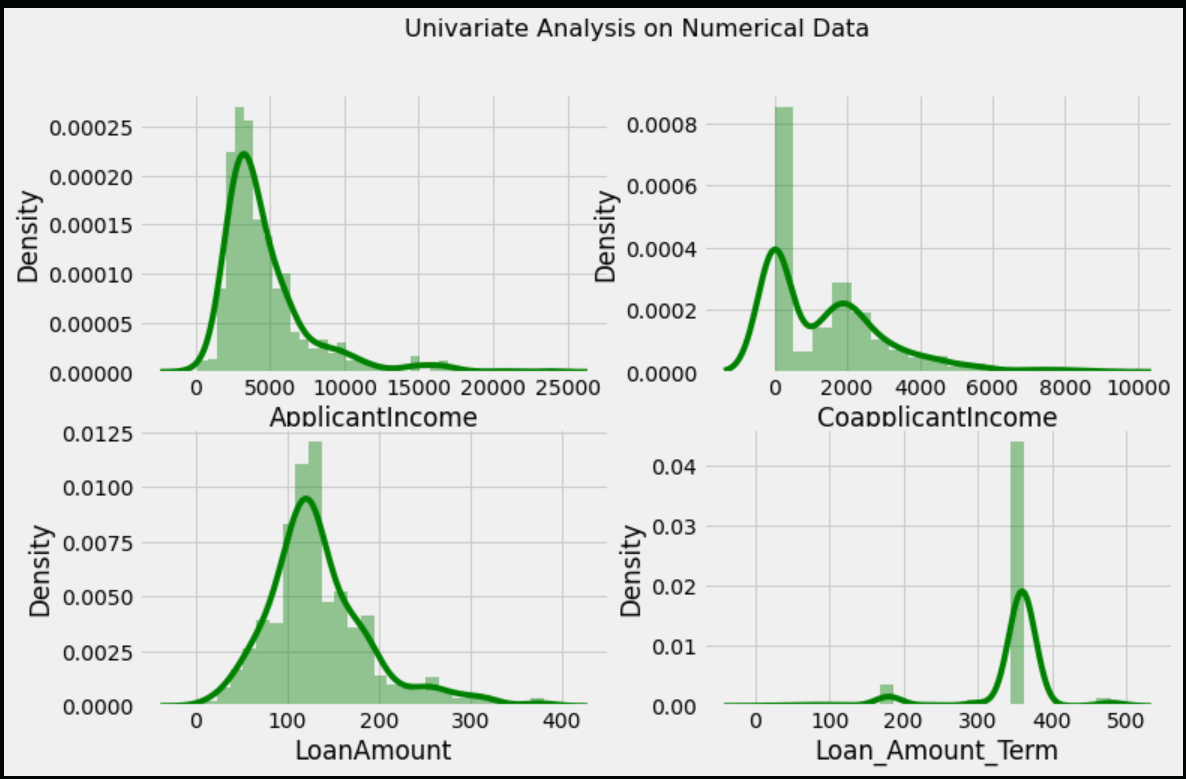
## **Exploratory Data Analysis**

### **Univariate Data Analysis**

1. **On Qualitative Features of Data**

****

1. **On Quantitative Features of Data**

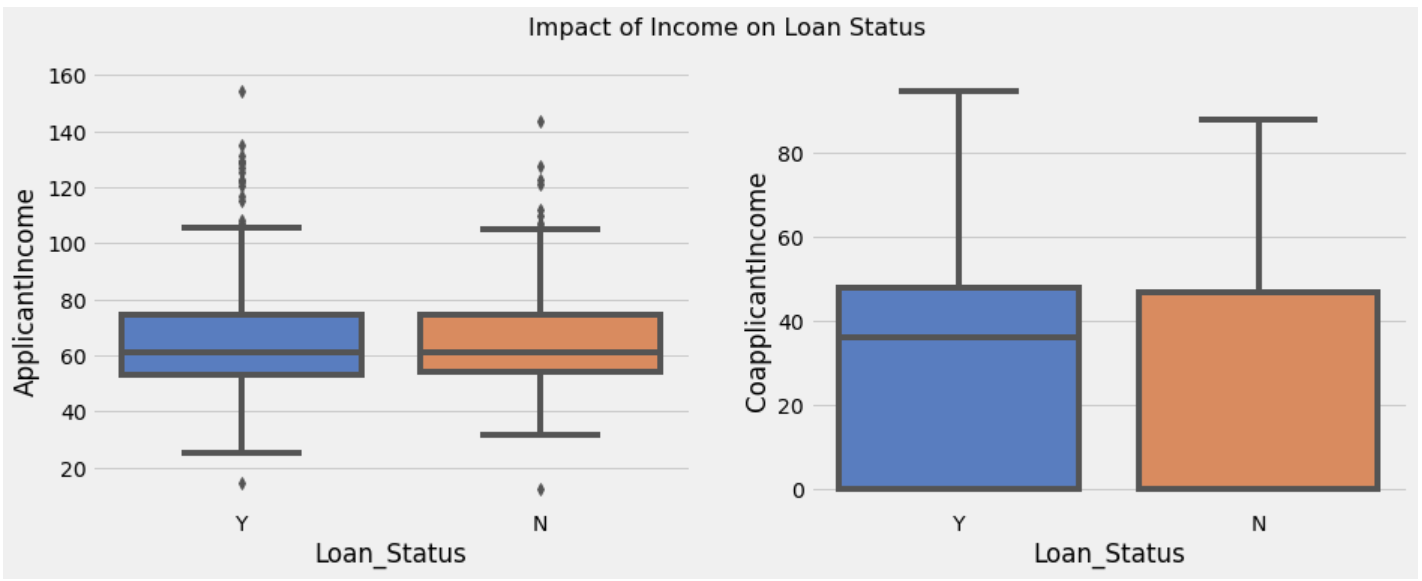
****

**As we see in above plot the data is highly skewed.**

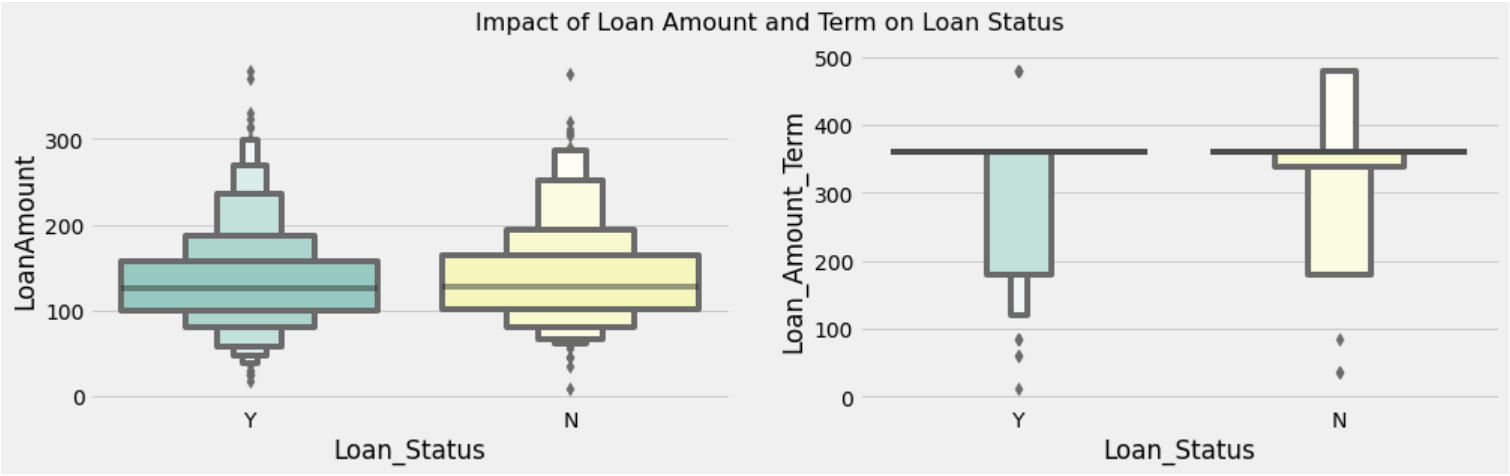
### **Bivariate Data Analysis**

**Bivariate Analysis on Quantitative Features**

Visualize the impact of Applicant Income and Co-applicant Income of the Loan Status

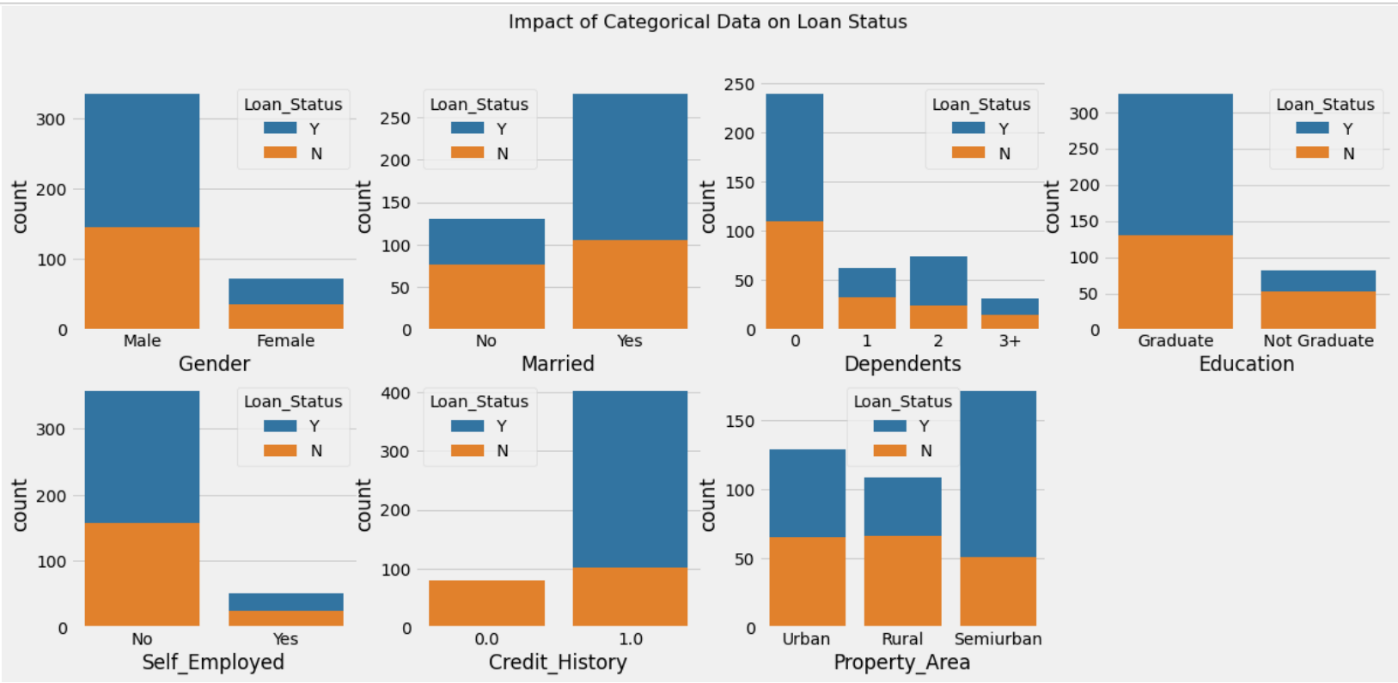
****

Visualize the Impact of Loan Amount and Loan Amount Term on Status of Loan

****

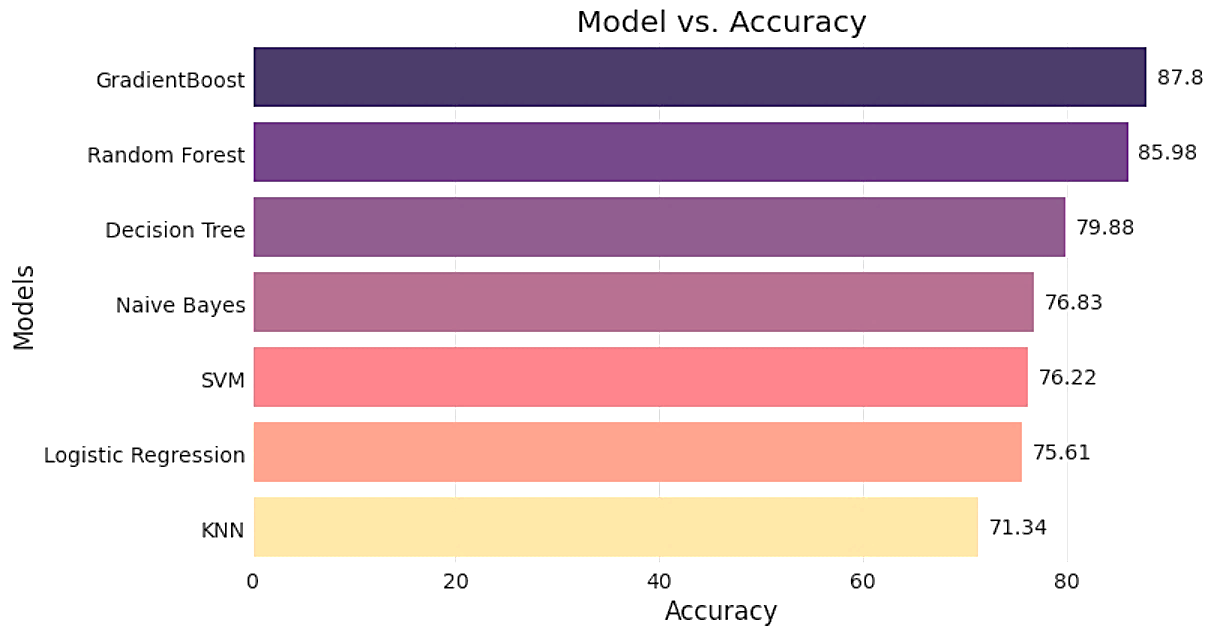
**Bivariate Analysis on Qualitative Features**

Visualize the relation between different categories of Data with Target variable (Loan\_Status)

****

**As we see in above chart the Credit history is high relation with Loan Status**

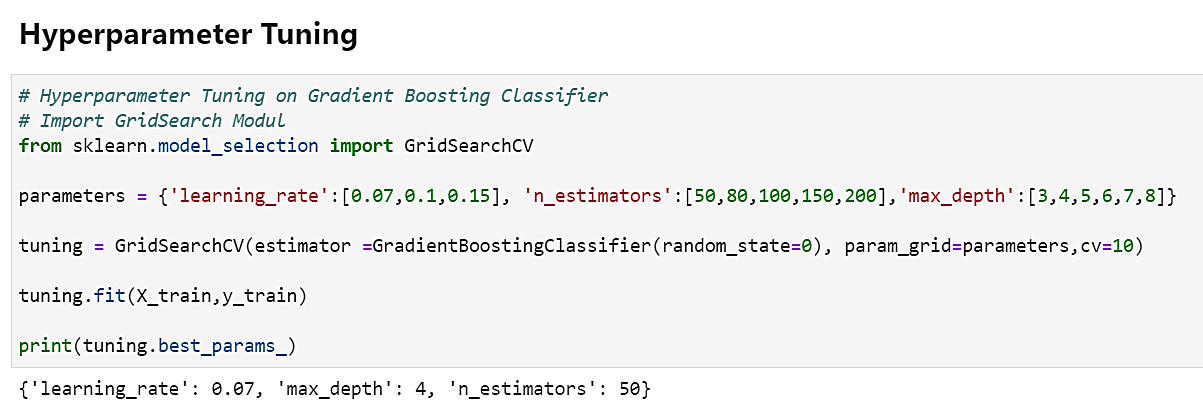
## **Machine Learning Models**

****

Fit our data on default parameters of different algorithms for binary classification. Surprisingly, Gradient Boost Classifier turned out to best in terms of validation set accuracy.

### **Hyper Parameter Tuning**

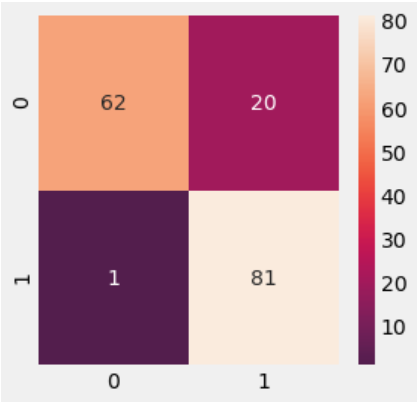
Hyper parameter tuning on Gradient Boosting Classifier



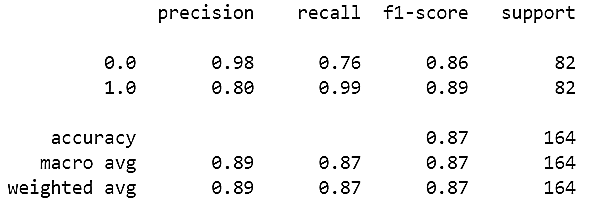
Check the Model Accuracy Report

### **Accuracy: 87.20%**

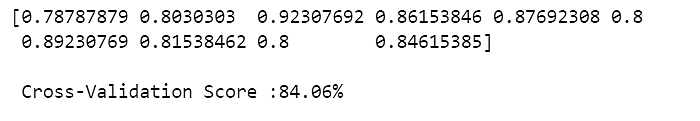
### **Confusion-Matrix:**

****

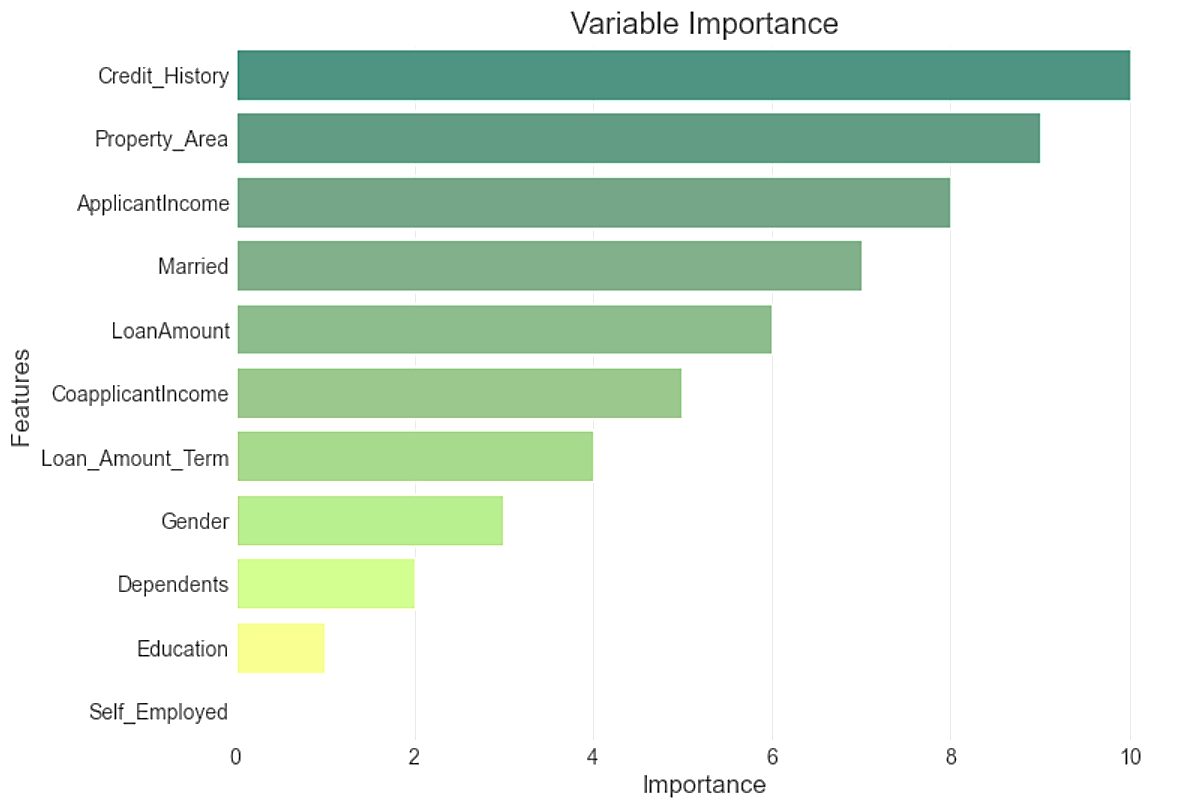
### **Classification Report:**

****

### **Cross-Validation score:**

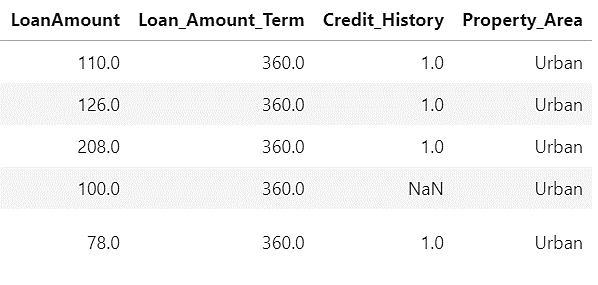
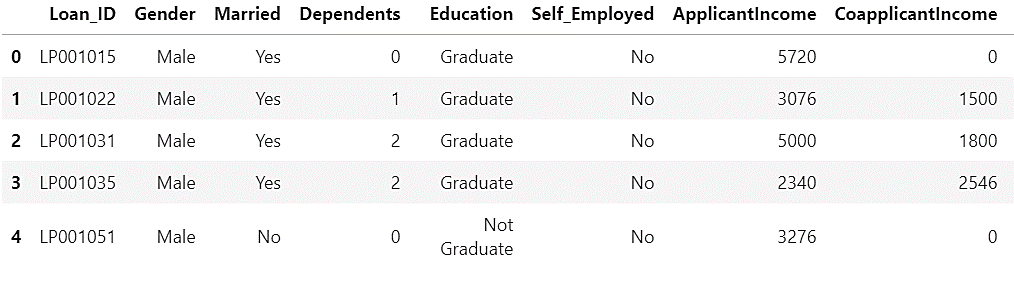
****

## **Feature Importance**



**As we see in above plot the Credit History, Property Area, Income, Married, Loan Amount are the most important features of data.**

## **Predict the Outcomes of Testing Data**



Fill the missing values of Test Data and convert to Numerical values to Categorical using functions which was created for Train Data.

Apply the Gradient Boosting Classifier and store result

## **Output Result**

The result is saved in CSV format using pandas in output directory.

It contains the Loan\_ID column and Loan\_Status Column

