Loan Eligibility Status Prediction

Swapnil Gavit
swapnilgavit19@gmail.com |+91 8177931221

Description:

1. Prediction of Loan Eligibility for Dream Housing Finance company is a Hackathon project on Datahack. (link)
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
Download Data Sets:

- Training Data Set
- Testing Data Set

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**:** CSV containing 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:

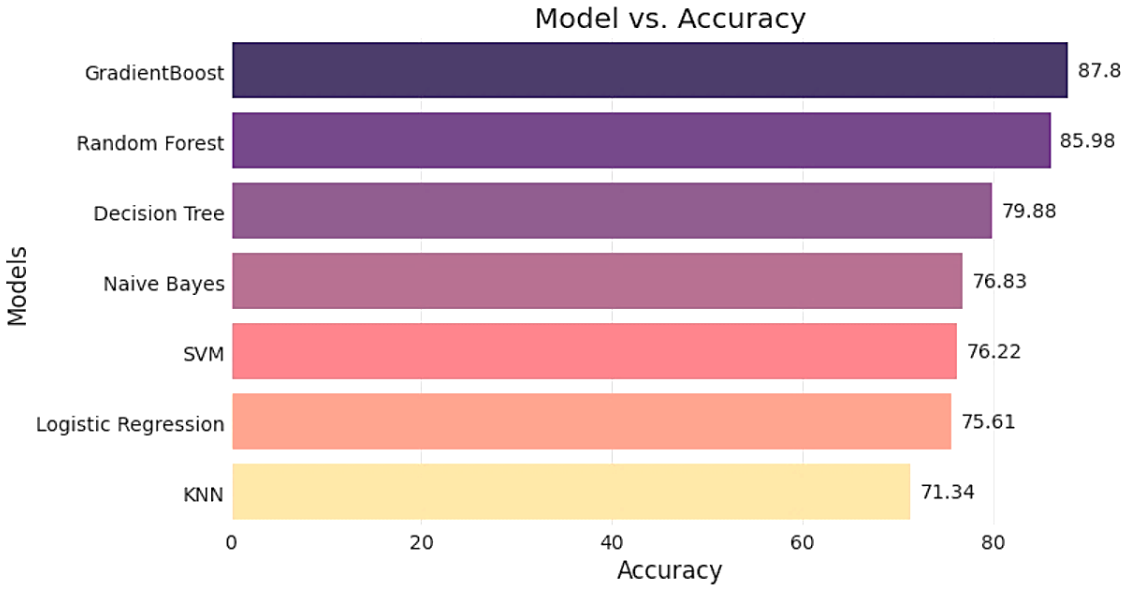
The final submission file format should be in following manner. The Loan_ID column contains unique loan IDs and the Loan_Status column contain predicted result of loan status.
The final submission file should be in CSV format

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

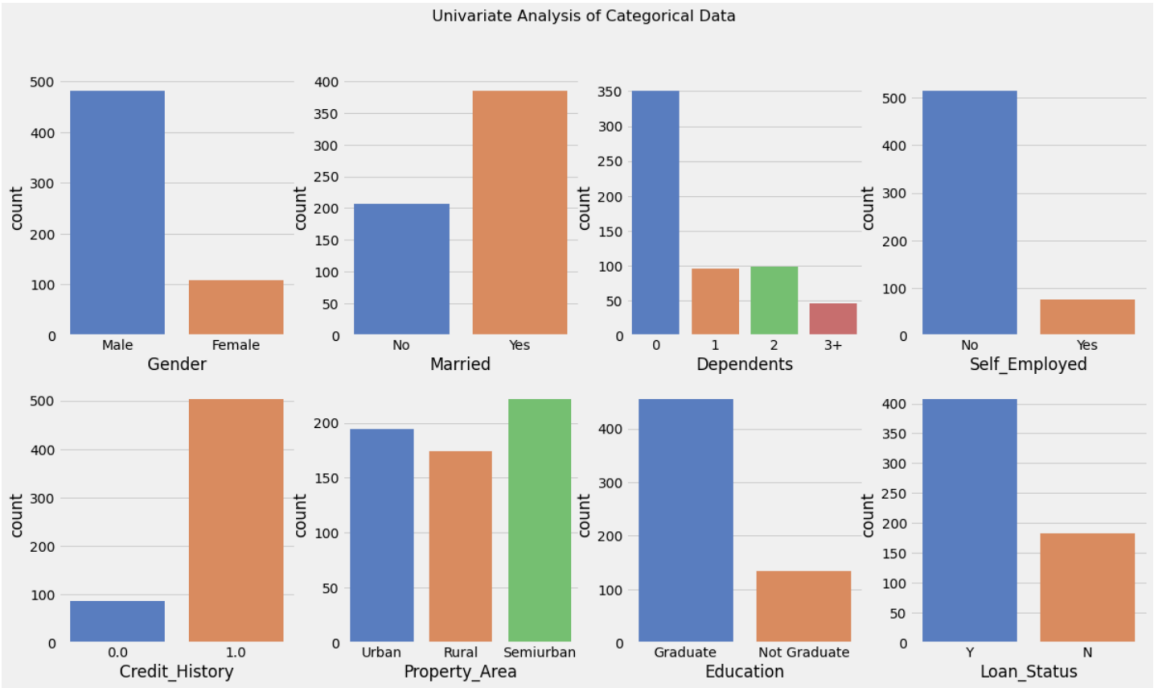
	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome
0	LP001002	Male	No	0	Graduate	No	5849	0.0
1	LP001003	Male	Yes	1	Graduate	No	4583	1508.0
2	LP001005	Male	Yes	0	Graduate	Yes	3000	0.0
3	LP001006	Male	Yes	0	Not Graduate	No	2583	2358.0
4	LP001008	Male	No	0	Graduate	No	6000	0.0

LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
NaN	360.0	1.0	Urban	Y
128.0	360.0	1.0	Rural	N
66.0	360.0	1.0	Urban	Y
120.0	360.0	1.0	Urban	Y
141.0	360.0	1.0	Urban	Y

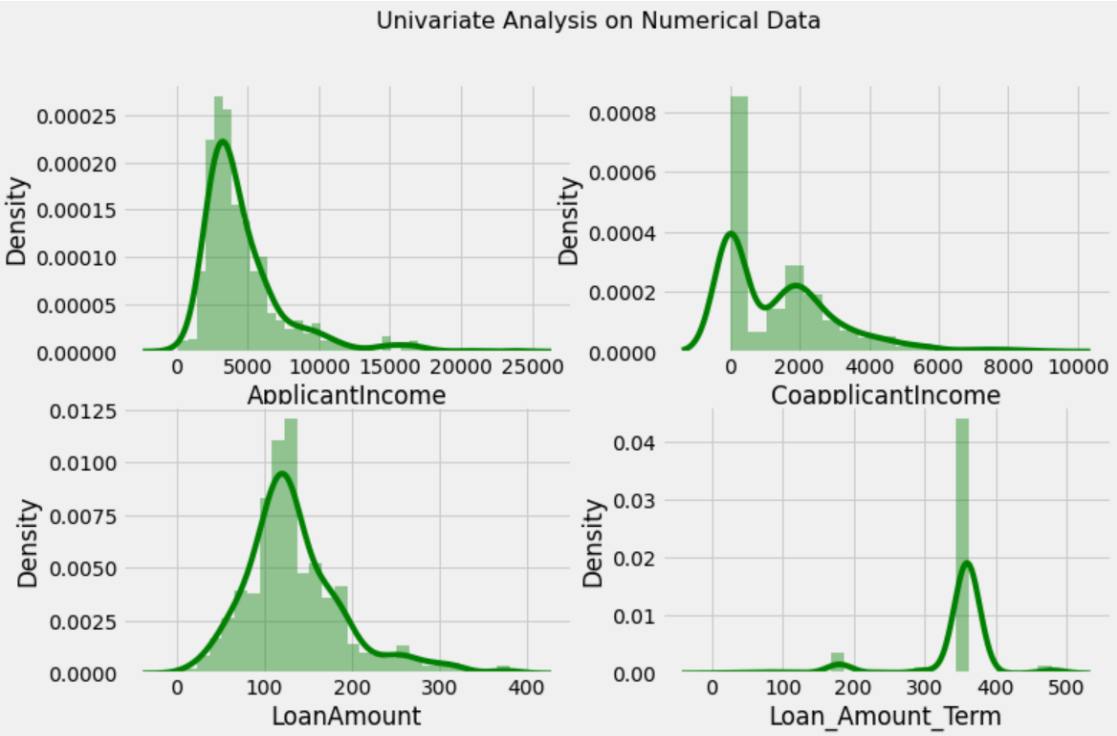
Exploratory Data Analysis

Univariate Data Analysis

1. On Qualitative Features of Data



2. On Quantitative Features of Data

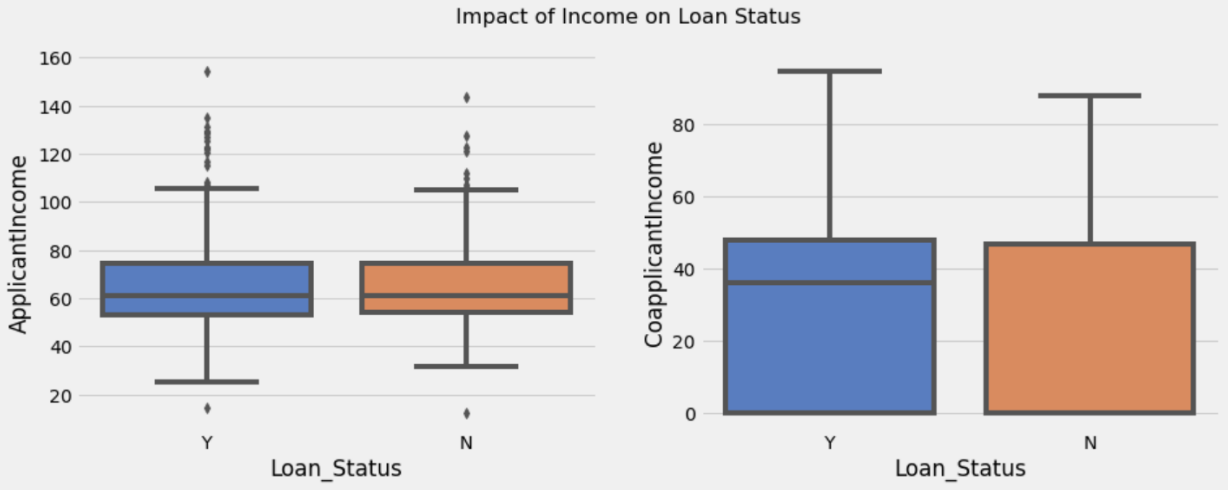


As we see in above plot the data is highly skewed.

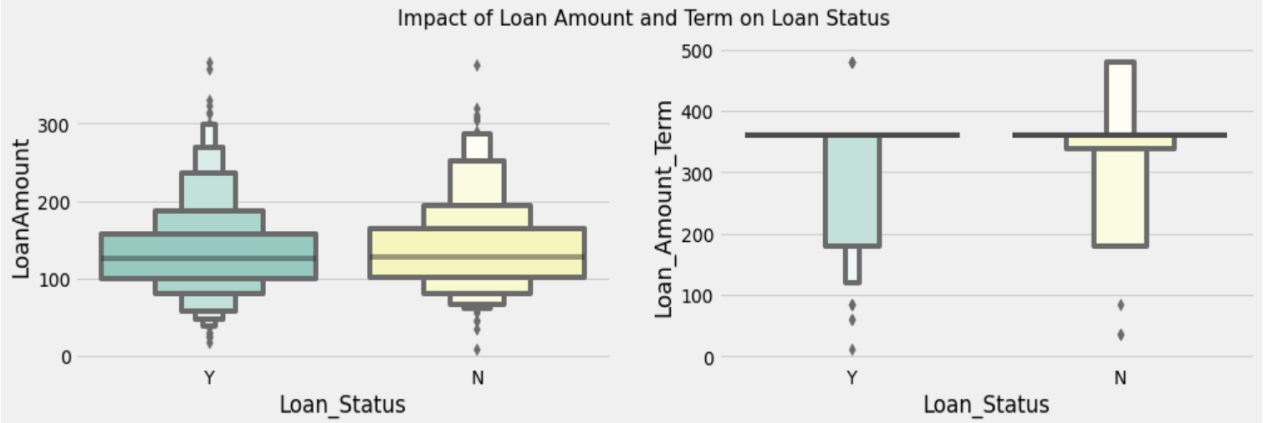
Bivariate Data Analysis

Bivariate Analysis on Quantitative Features

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

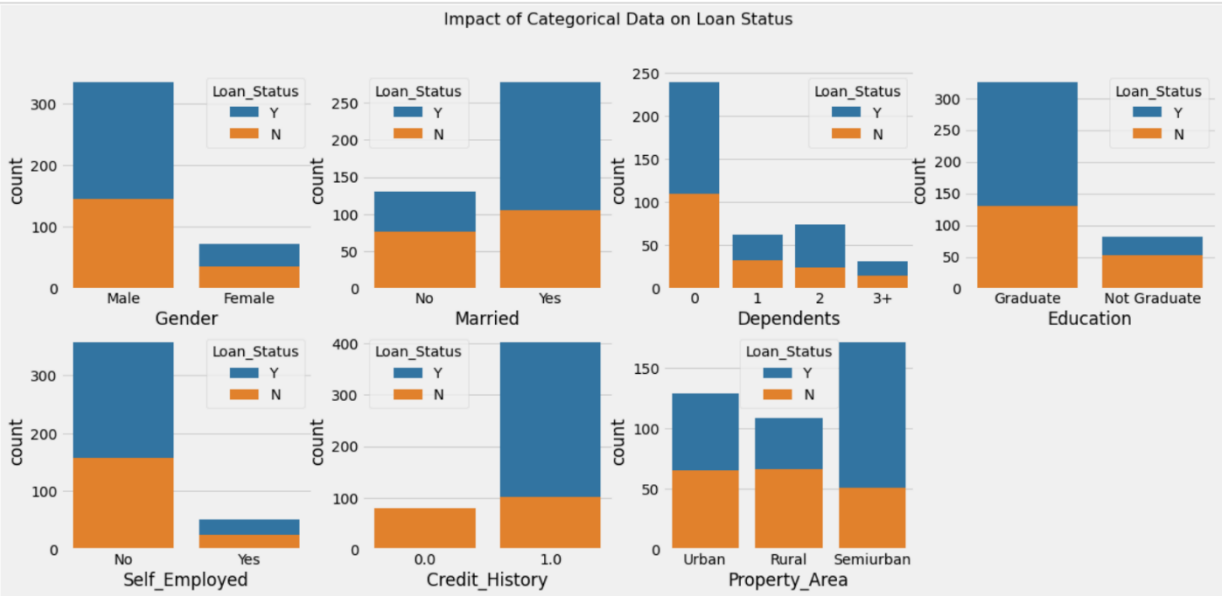


Visualize the Impact of Loan Amount and Loan Amount Term on Status of Loan



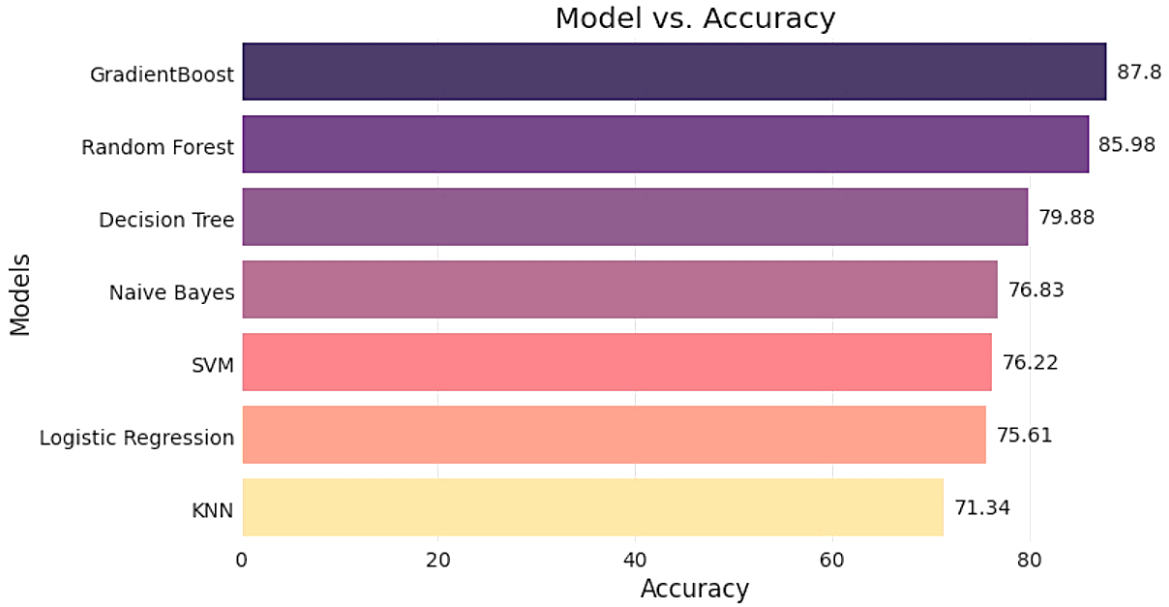
Bivariate Analysis on Qualitative Features

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

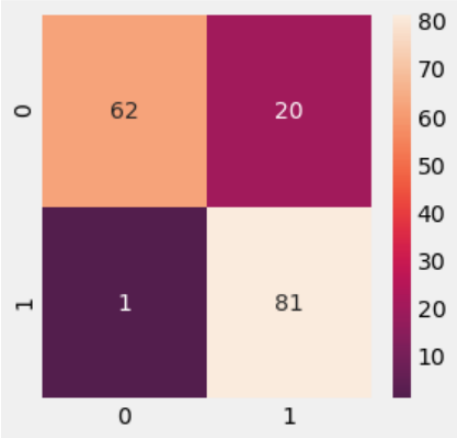
```
# Hyperparameter Tuning on Gradient Boosting Classifier
# Import GridSearch Modul
from sklearn.model_selection import GridSearchCV

parameters = {'learning_rate':[0.07,0.1,0.15], 'n_estimators':[50,80,100,150,200],'max_depth':[3,4,5,6,7,8]}

tuning = GridSearchCV(estimator =GradientBoostingClassifier(random_state=0), param_grid=parameters,cv=10)

tuning.fit(X_train,y_train)

print(tuning.best_params_)
```

```
{'learning_rate': 0.07, 'max_depth': 4, 'n_estimators': 50}
```

Check the Model Accuracy Report

Accuracy: 87.20%

Confusion-Matrix:



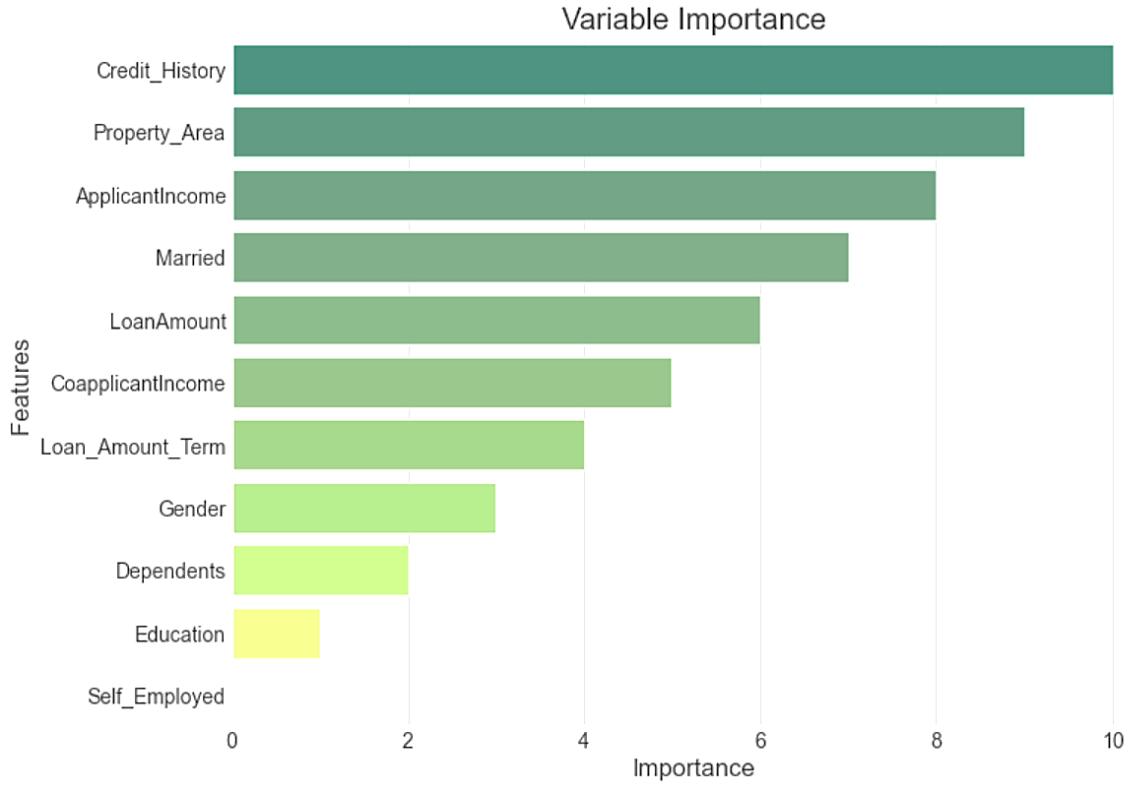
Classification Report:

	precision	recall	f1-score	support
0.0	0.98	0.76	0.86	82
1.0	0.80	0.99	0.89	82
accuracy			0.87	164
macro avg	0.89	0.87	0.87	164
weighted avg	0.89	0.87	0.87	164

Cross-Validation score:

```
[0.78787879 0.8030303  0.92307692 0.86153846 0.87692308 0.8
 0.89230769 0.81538462 0.8        0.84615385]

 Cross-Validation Score :84.06%
```

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

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome
0	LP001015	Male	Yes	0	Graduate	No	5720	0
1	LP001022	Male	Yes	1	Graduate	No	3076	1500
2	LP001031	Male	Yes	2	Graduate	No	5000	1800
3	LP001035	Male	Yes	2	Graduate	No	2340	2546
4	LP001051	Male	No	0	Not Graduate	No	3276	0

LoanAmount	Loan_Amount_Term	Credit_History	Property_Area
110.0	360.0	1.0	Urban
126.0	360.0	1.0	Urban
208.0	360.0	1.0	Urban
100.0	360.0	NaN	Urban
78.0	360.0	1.0	Urban

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

Loan_ID	Loan_Status
LP001015	Y
LP001022	Y
LP001031	Y
LP001035	Y
LP001051	Y