Predict Loan Eligibility for a Finance Company Report

Introduction

Finance companies in Sri Lanka offers home loans and have a presence in urban, semi-urban, and rural areas. When a customer wants to apply for a home loan, the finance company checks if they are eligible or identify applicants who may default on loans. To automate this process, company has to identify the customers segments, those are eligible for loan amount so that they can specifically target those customers. Here we are using a sample dataset which is almost similar to the finance company dataset.

Dataset Information

Variable	Description	Data Type
Loan_ID	Unique Loan ID	object
Gender	Male/ Female	object
Married	Applicant married (Y/N)	object
Dependents	Number of dependents	object
Education	Applicant Education (Graduate/ Under Graduate)	object
Self_Employed	Self employed (Y/N)	object
ApplicantIncome	Applicant income	int64
CoapplicantIncome	Coapplicant income	float64
LoanAmount	Loan amount in thousands	float64
Loan_Amount_Term	Term of loan in months	float64
Credit_History	credit history meets guidelines	float64
Property_Area	Urban/ Semi Urban/ Rural	object
Loan_Status	Loan approved (Y/N)	object

There are 614 observations and 12 Independent Variables and 1 Target Variable in the data.

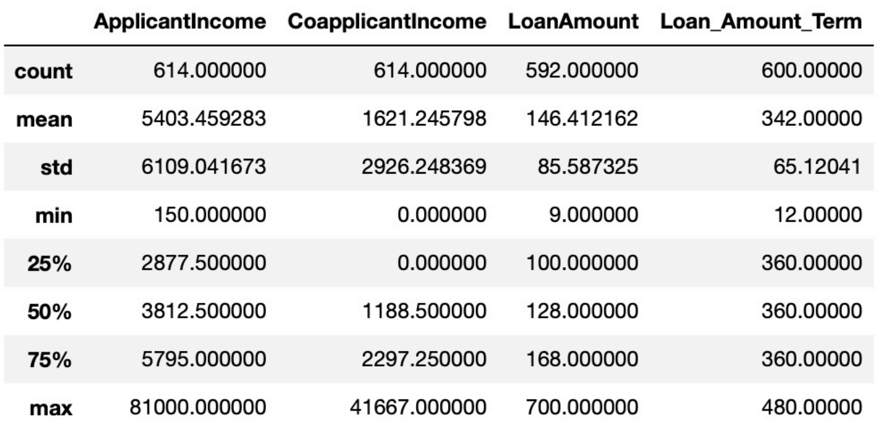
We can see there are three format of data types:

object: Object format means variables are categorical. Categorical variables in our dataset are: Loan_ID, Gender, Married, Dependents, Education, Self_Employed, Property_Area, Loan_Status.

int64: It represents the integer variables. ApplicantIncome is of this format.

float64: It represents the variable which have some decimal values involved. They are also numerical variables. Numerical variables in our dataset are: CoapplicantIncome, LoanAmount, Loan_Amount_Term, and Credit_History.

Exploratory Data Analysis

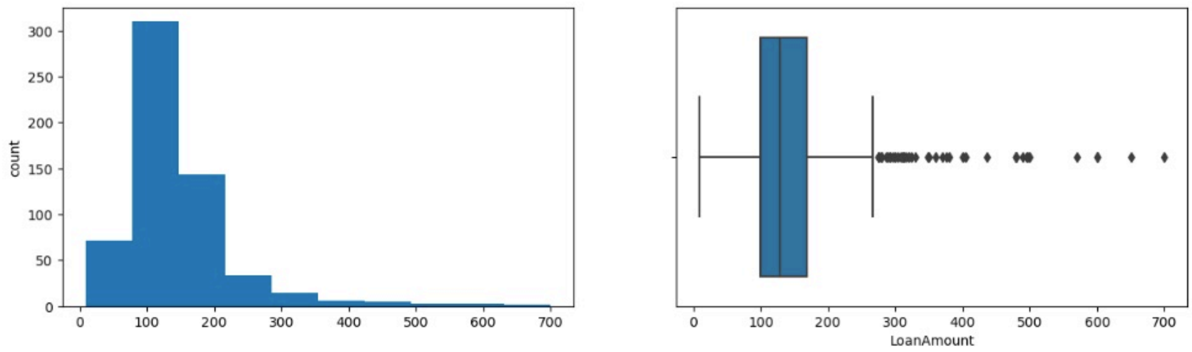


	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term
count	614.000000	614.000000	592.000000	600.00000
mean	5403.459283	1621.245798	146.412162	342.00000
std	6109.041673	2926.248369	85.587325	65.12041
min	150.000000	0.000000	9.000000	12.00000
25%	2877.500000	0.000000	100.000000	360.00000
50%	3812.500000	1188.500000	128.000000	360.00000
75%	5795.000000	2297.250000	168.000000	360.00000
max	81000.000000	41667.000000	700.000000	480.00000

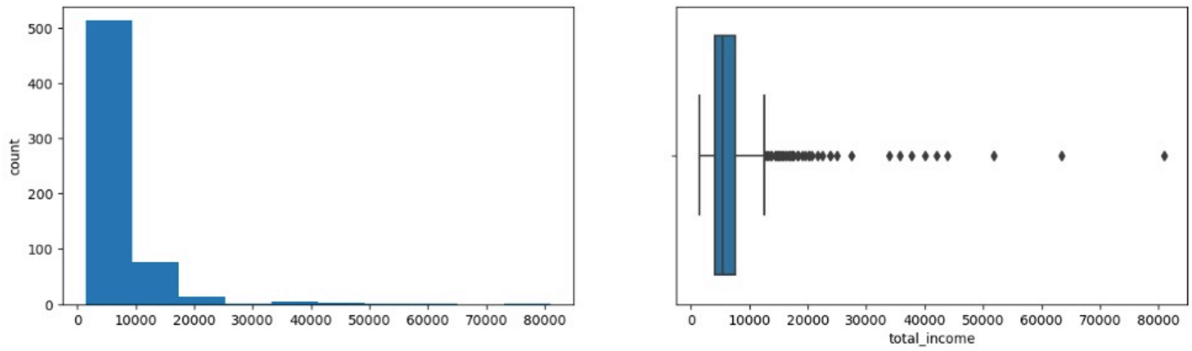
- The mean applicant income is around 5403.46
- The minimum value for applicant income is 150 which is very small and could be an outlier since it is very far from both the mean and the median.
- There are a good number of co applicants with no income, as can be observed by the fact that the 25% quantile of coapplicants have an income of 0

Distribution and Outliers for LoanAmount and total_income column in the data

```
LoanAmount
Skew : 2.68
```

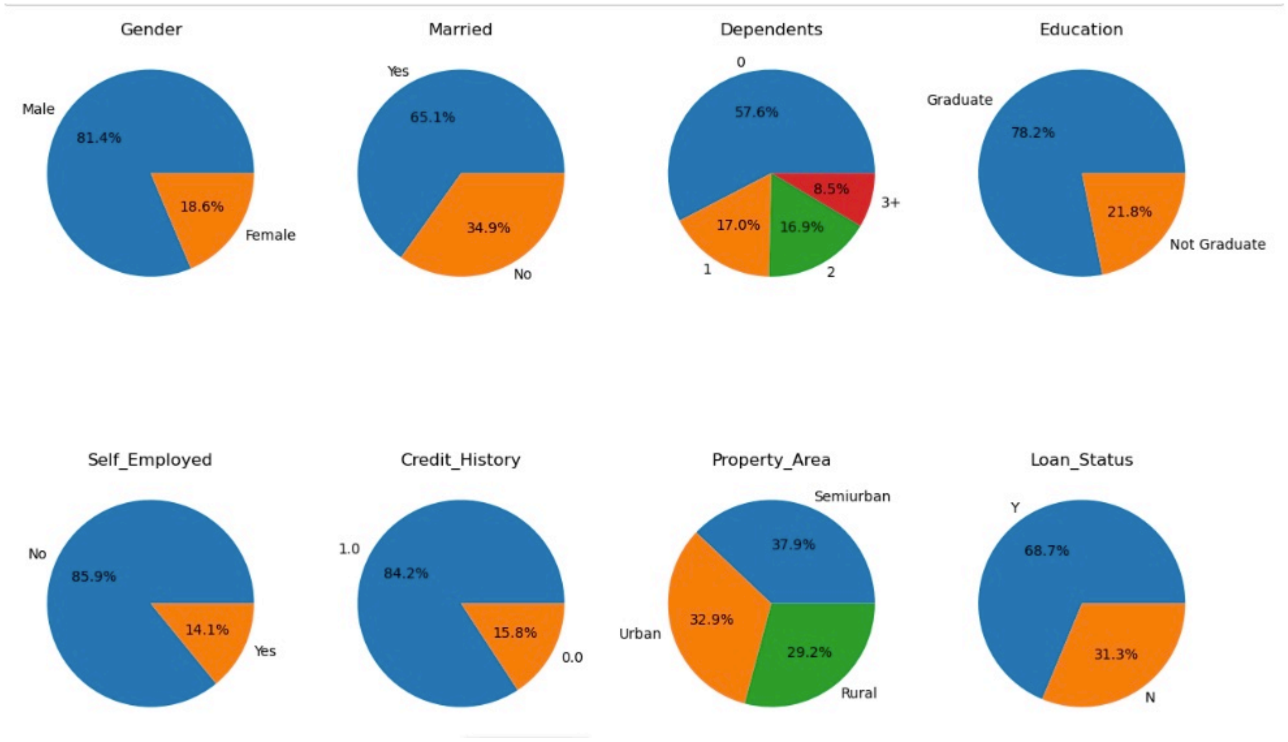


```
total_income
Skew : 5.63
```

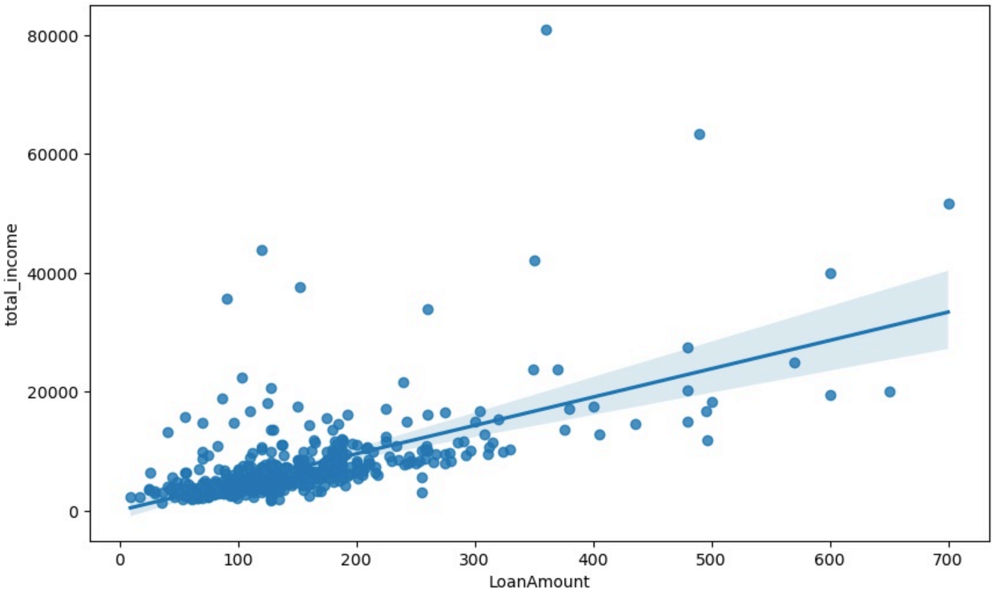


- Both the variables are highly skewed to the right and have many outliers which can be expected as the data contains different types of areas - Rural, Urban & Semi-Urban.
- We can observe from the histogram that majority of values for total income are less than 10,000.

Percentage of each category for categorical variables

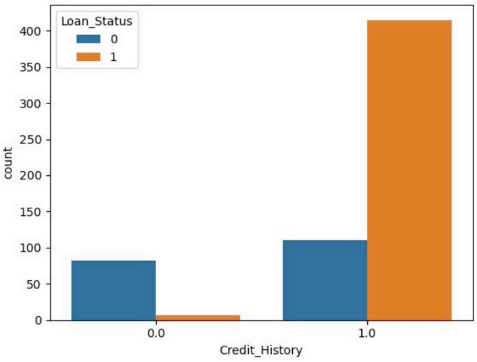


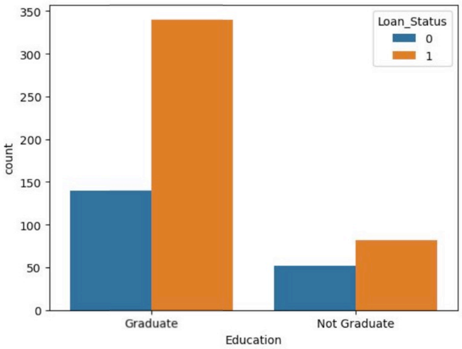
- The vast majority of applicants are male (81%).
- 65% of applicants are married, and 57% of them have no dependents
- 78% of the applicants are graduates
- 85% of the applicants are self employed
- 84% of the applicants have credit histories that meet the required guidelines
- The property area among the applicants is roughly evenly split across semiurban, urban, and rural, with semi-urban having a slightly higher portion (37.9%)

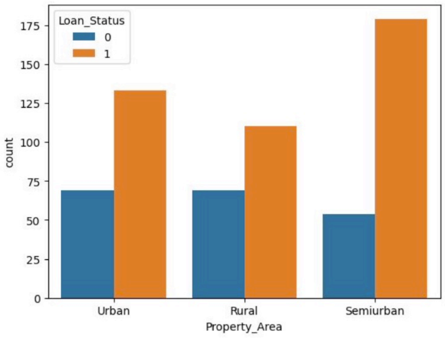


The plot shows that the loan amount is positively correlated with total income. This implies that the loan amount for higher-income applicants is progressively higher.

There are some outliers visible, showing applicants with low income having been given loans of a higher amount.





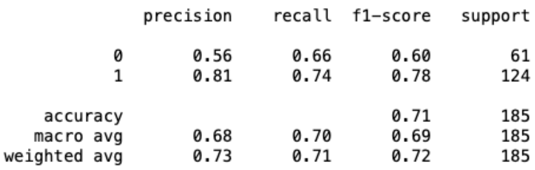


- The plot shows that credit history is an important factor while checking loan eligibility. There are very few customers whose loan was approved even when their credit history did not meet required guidelines.
- We can see that graduate customers are more likely to get loans
- The plot shows that more loans are approved for properties in semi-urban areas.
- This could be due to several reasons. The bank might be charging higher interest rates for semi-urban areas or the current customer base of the company from semi-urban areas might actually be more eligible for home loans based on loan applicant features. We cannot be certain as we don't have the data to support this claim.

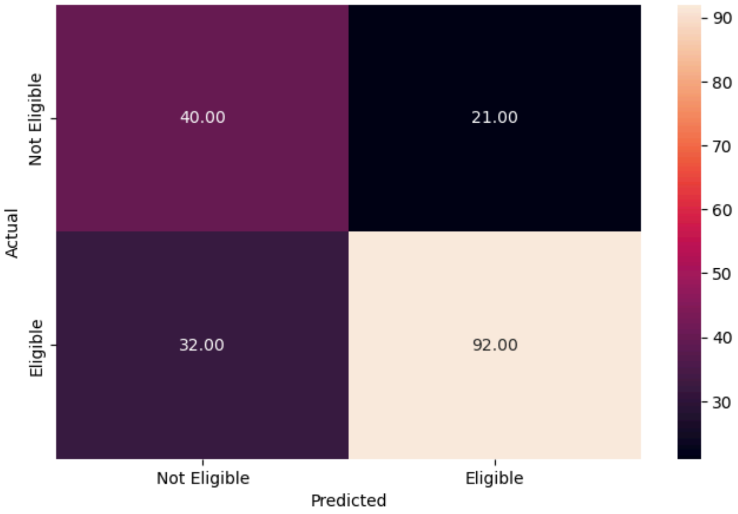
Building Classification Models

Logistic Regression

Performance of the model on the training set



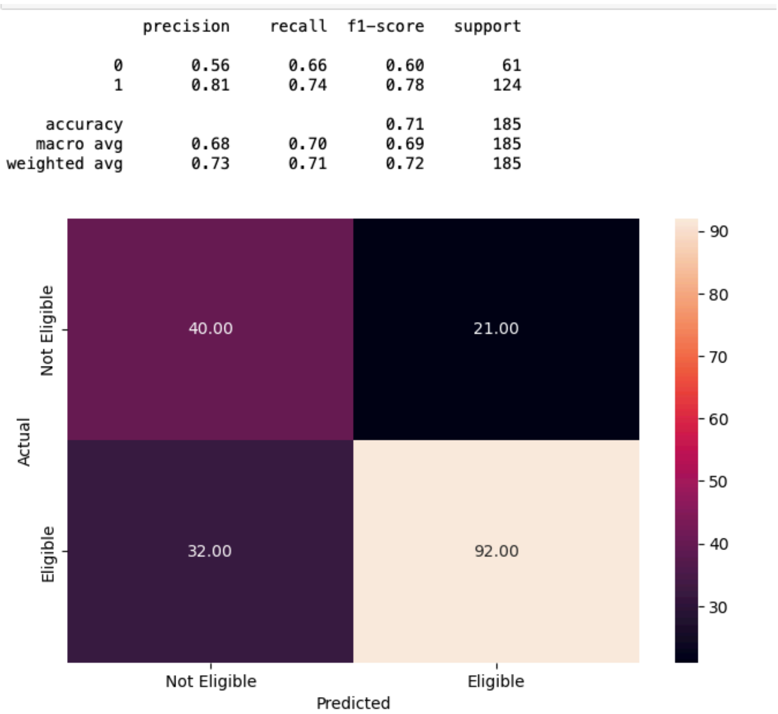
```
              precision    recall  f1-score   support

           0       0.56      0.66      0.60        61
           1       0.81      0.74      0.78       124

    accuracy                           0.71       185
   macro avg       0.68      0.70      0.69       185
weighted avg       0.73      0.71      0.72       185
```



- We see around 82% accuracy on our training set.

- The recall score is only 44% for class 0 which is low, considering we want to get a strong recall value for our specific problem. Thus, this model will not perform well for us.

Performance on the testing data



```
              precision    recall  f1-score   support

           0       0.56      0.66      0.60        61
           1       0.81      0.74      0.78       124

    accuracy                           0.71       185
   macro avg       0.68      0.70      0.69       185
weighted avg       0.73      0.71      0.72       185
```

We can see that even though the precision dropped for class 0, the recall score is much higher after using the optimal threshold, and now our model is a lot more appropriate after improvement.

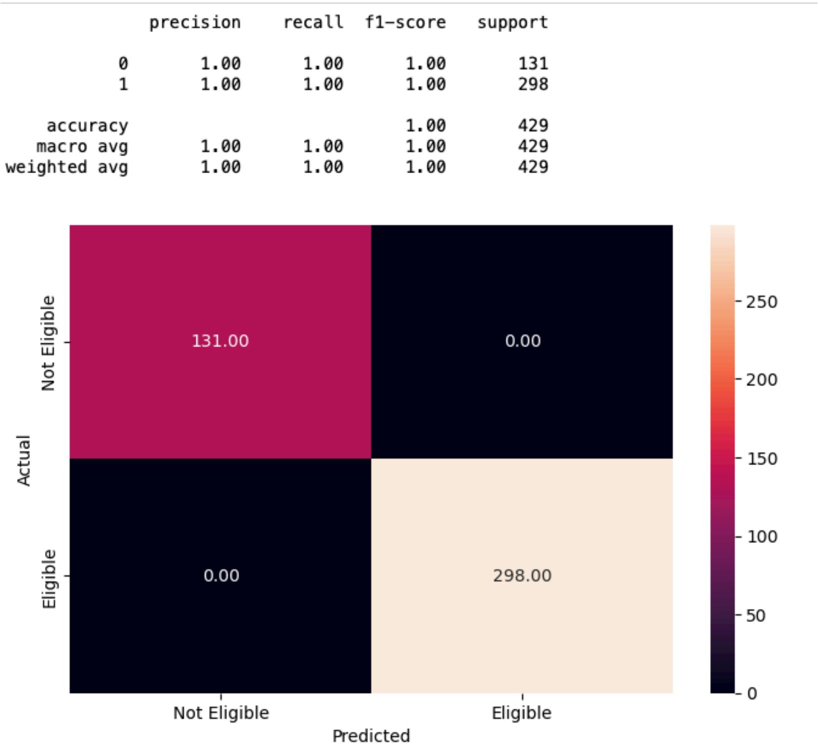
Similar performance on both the training data and the test data

K - Nearest Neighbors (KNN)

Performance of the model

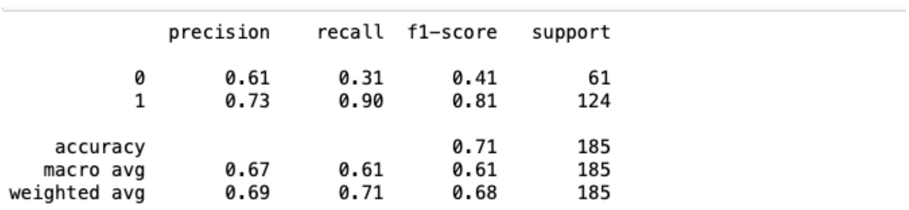
Predicting on train data:

The performance on the training set is very strong, as it predicts perfectly who is eligible and who isn't. The precision, recall, and and accuracy are all optimal at 100%

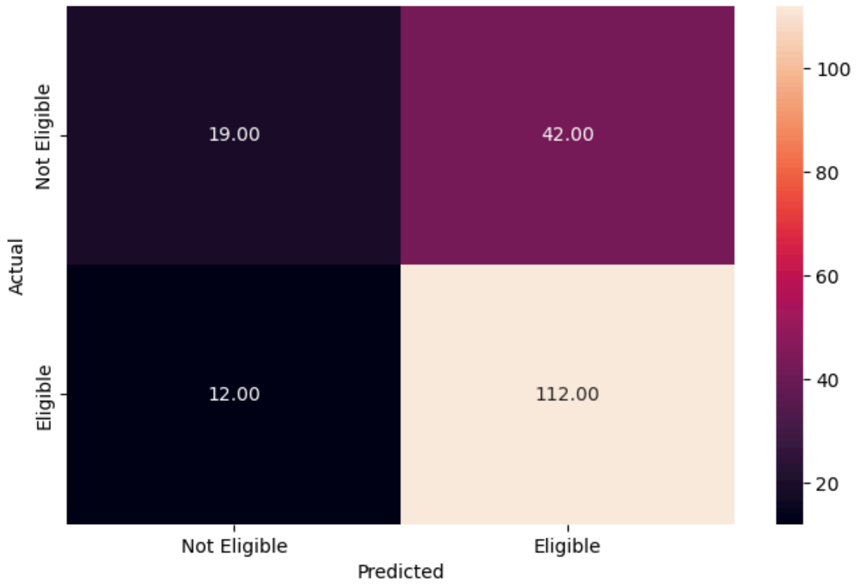


```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       131
           1       1.00      1.00      1.00       298

    accuracy                           1.00       429
   macro avg       1.00      1.00      1.00       429
weighted avg       1.00      1.00      1.00       429
```

Predicting on test data



```
              precision    recall  f1-score   support

           0       0.61      0.31      0.41        61
           1       0.73      0.90      0.81       124

    accuracy                           0.71       185
   macro avg       0.67      0.61      0.61       185
weighted avg       0.69      0.71      0.68       185
```



We see a weak performance on the scaled testing data, as the recall score is only 0.31 for the 0. There were 42 people that our Knn model predicted to be eligible who are actually not eligible, which is concerning given that we want to minimise this value as much as possible.

Conclusion

Through the use of multiple models, EDA, and visualization, we identified the main key factor affecting loan application acceptance is credit history.

We aimed to maximize the recall score in our Logistic regression model as it measures the ability to identify applicants who may default on loans. So that finance company can avoid targeting applicants who cannot repay them, as this could harm the company. As a result, the Logistic regression model provided the highest recall score.

Github link - https://github.com/Banujan/UOM_DataScience_Project