Predict Loan Eligibility

Data Science - Project

239306V - Banujan Chandrakanthan

Introduction

- A finance company in Sri Lanka offers home loans and operates in urban, semi-urban, and rural areas.
- Before approving a home loan application, the company checks the eligibility of the customer or identify applicants who may default on loans.
- The company aims to automate the loan eligibility assessment process by identifying eligible customer segments.
- This identification process will allow the company to specifically target eligible customers for loan amounts.

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

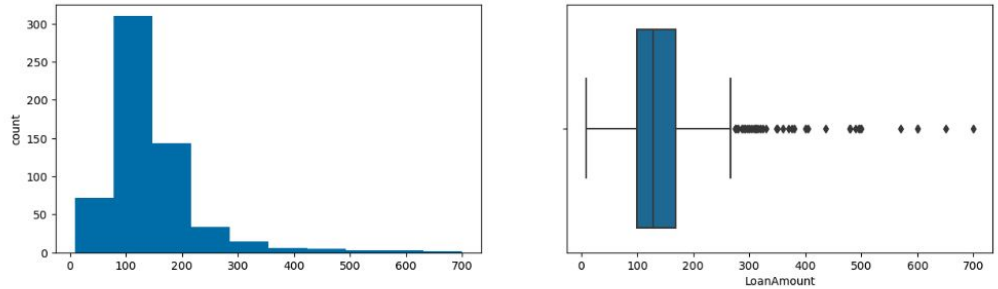
- 614 Rows
- 12 Independant Variables
- 1 Target Variable

Exploratory Data Analysis

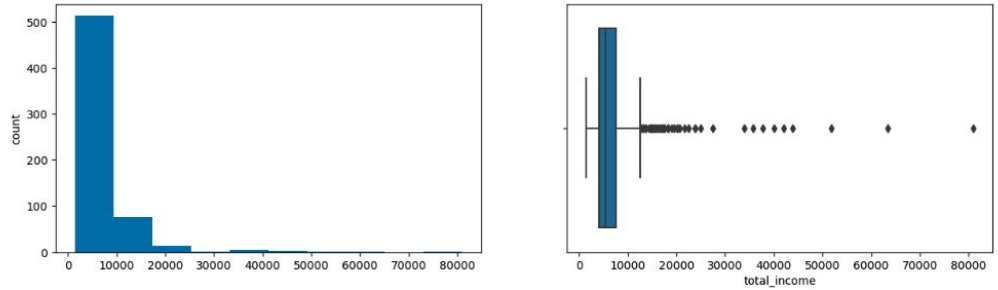
	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term
count	614.000000	614.000000	592.000000	600.00000
mean	5403.459283	1621.245798	146.412162	342.00000
std	6109.041673	2926.248369	85.587325	65.12041
min	150.000000	0.000000	9.000000	12.00000
25%	2877.500000	0.000000	100.000000	360.00000
50%	3812.500000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

Exploratory Data Analysis

LoanAmount
Skew : 2.68

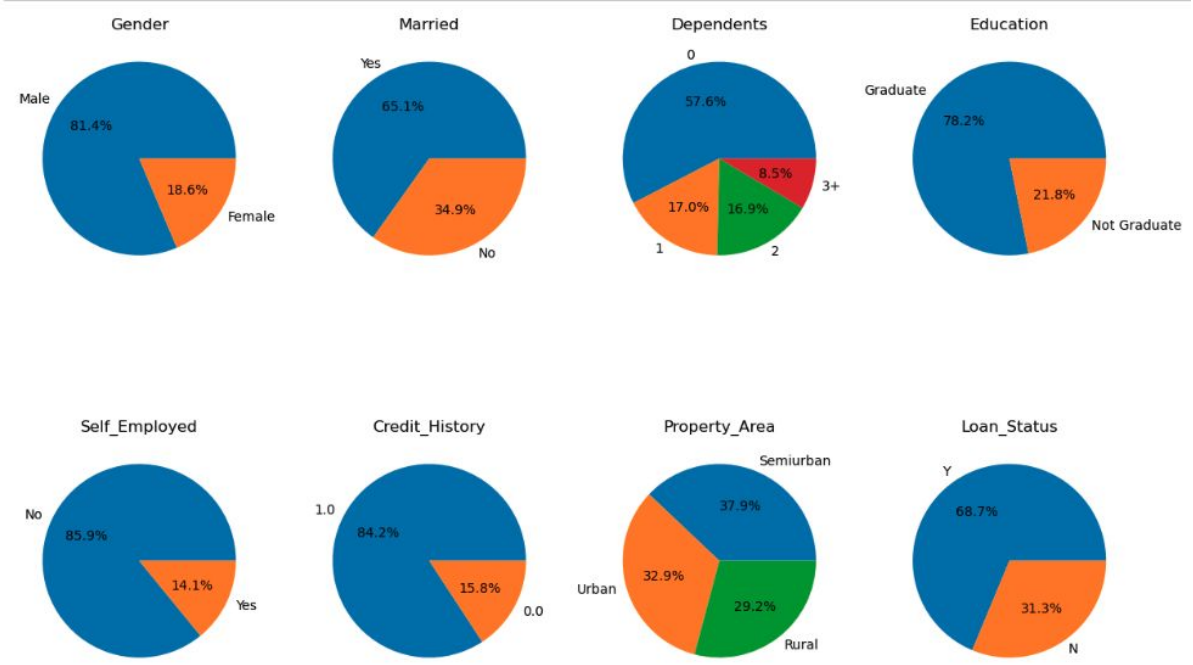


total_income
Skew : 5.63

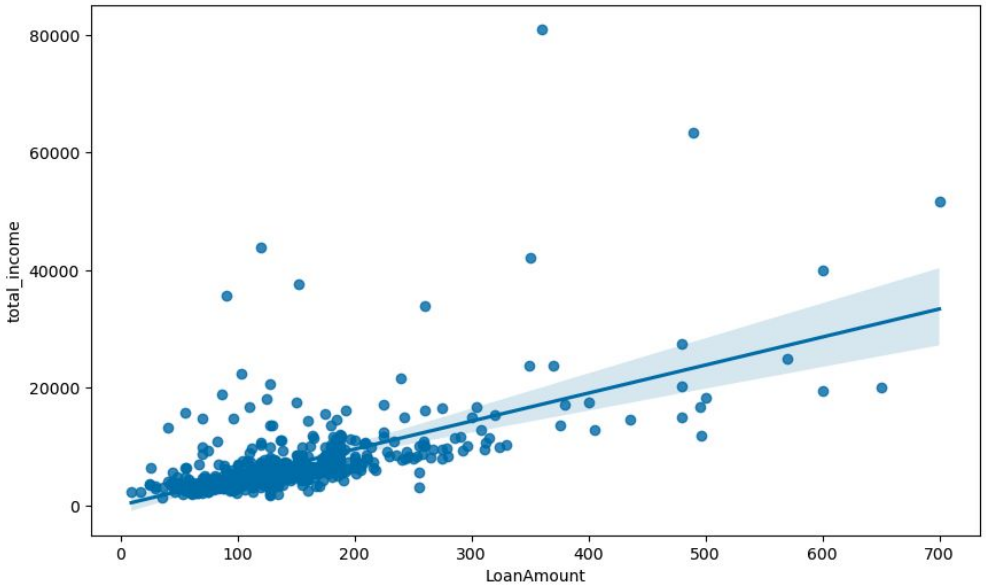


Distribution and outliers for each columns

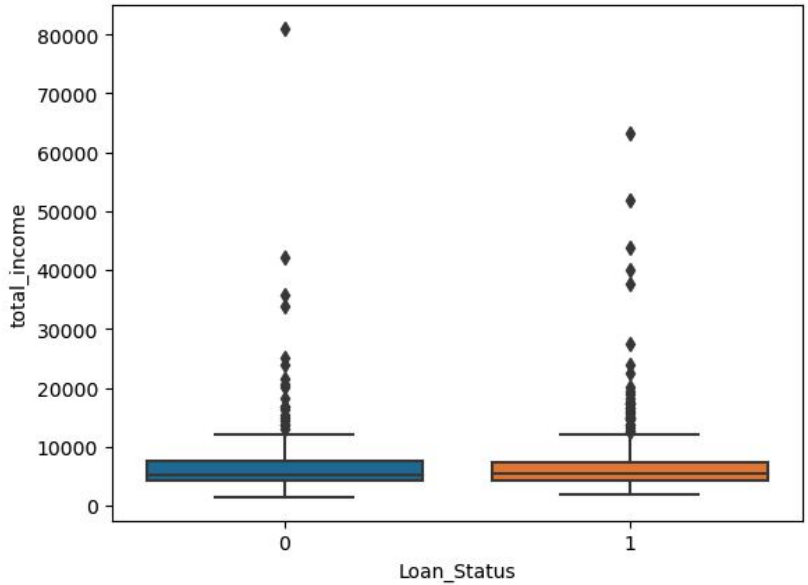
Exploratory Data Analysis



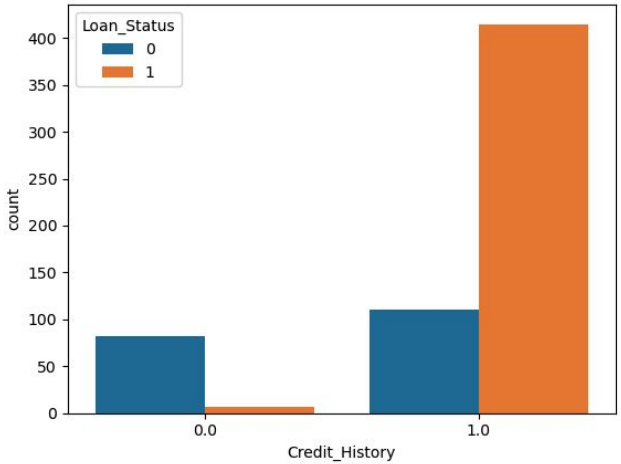
Exploratory Data Analysis

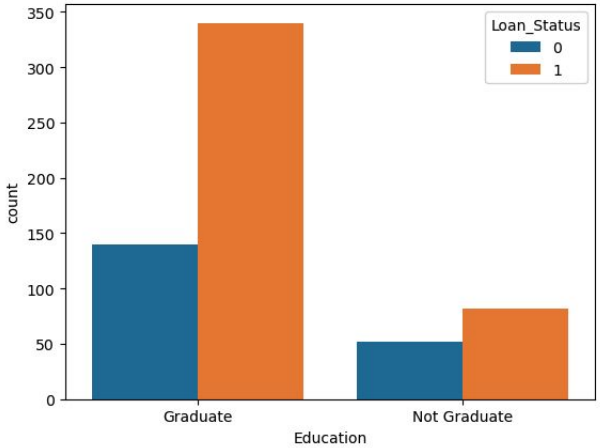


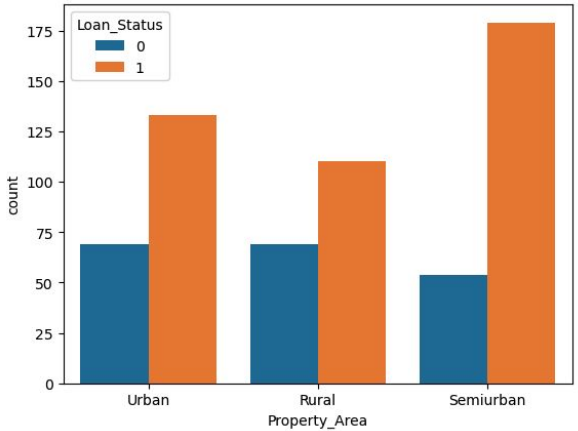
Exploratory Data Analysis



Exploratory Data Analysis





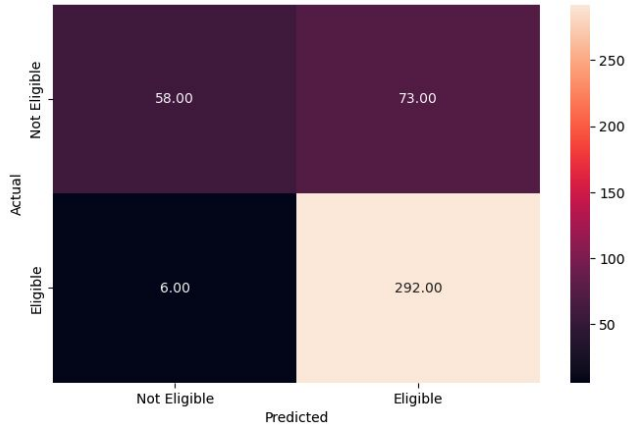


Methodology - Logistic Regression

	odds
Credit_History	20.620598
Property_Area_Semiurban	2.274152
Married_Yes	1.208542
Dependents_2	1.167358
Dependents_3+	1.120508
total_income	0.999990
LoanAmount	0.999526
Property_Area_Urban	0.993578
Loan_Amount_Term	0.936248
Self_Employed_Yes	0.906185
Gender_Male	0.899425
Dependents_1	0.864392
Education_Not Graduate	0.611551

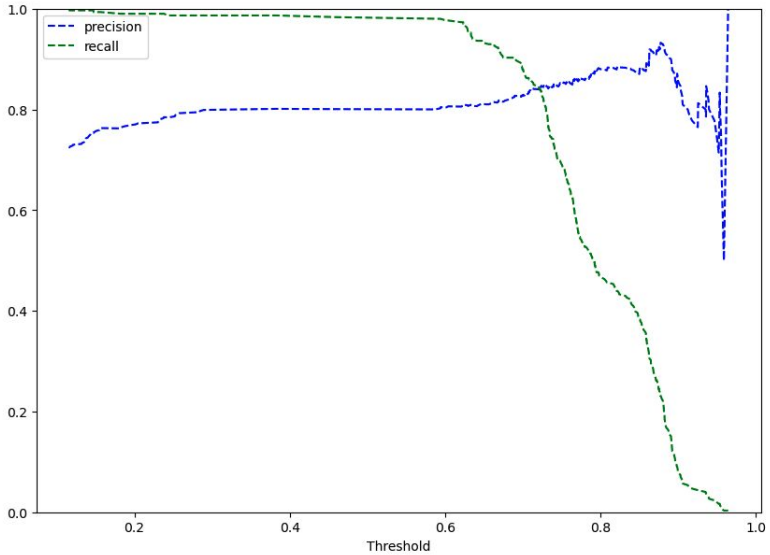
Odds calculated from the logistic regression model coefficients

	precision	recall	f1-score	support
0	0.91	0.44	0.59	131
1	0.80	0.98	0.88	298
accuracy			0.82	429
macro avg	0.85	0.71	0.74	429
weighted avg	0.83	0.82	0.79	429



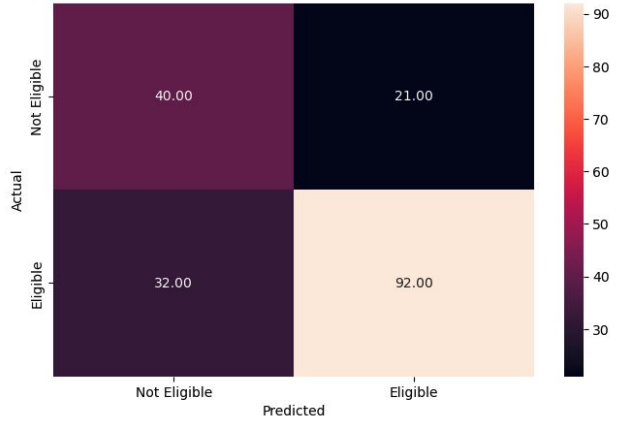
Performance of the model on the training set

Methodology - Logistic Regression



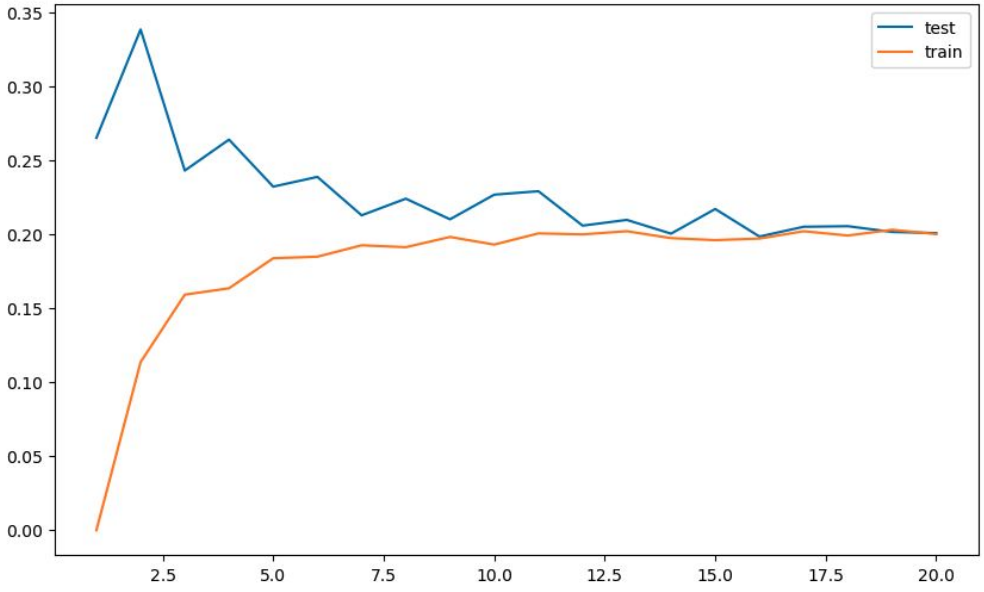
Precision-Recall curve for Logistic Regression

	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



Performance of the model on the training set

Methodology - K - Nearest Neighbors (KNN)



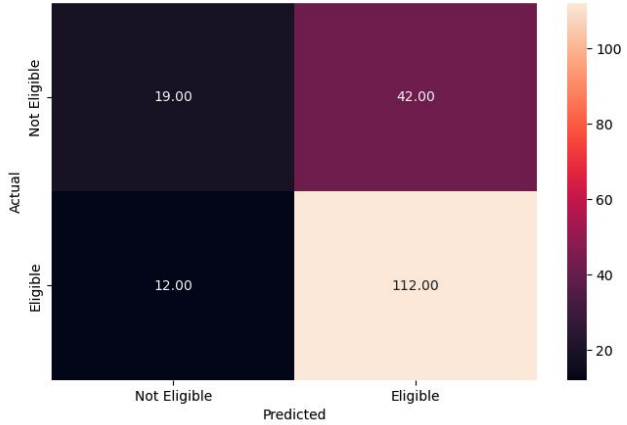
Extracting the train and the test error for each k in a list

Methodology - K - Nearest Neighbors (KNN)

Performance of the model on the training and testing data

	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

	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



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.