## Project Title : Home loan approval prediction using machine learning models.

# Home loan approval prediction using machine learning models.

Khagendra Khati
Anderson College of Business and Computing, Regis University
MSDS692 Data Science Practicum-I
Professor Dr. Kellen Sorauf
March 5, 2023

## Purpose

The purpose of this project is to predict the home loan approval of the client. The housing finance company wants to automate the loan eligibility process based on customer details provided while filling out the online application form. Attributes provided by the applicant include Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History, and others.This project has taken the data of previous customers of the bank to whom on a set of parameters loans were approved. So the machine learning model is trained on that record to get accurate results. Our main objective of this project is to predict the safety of loan, which in turn will help to protect the bank from loan defaulters.

### Background

Generally, loan prediction involves the lender looking at various background information about the applicant and deciding whether the bank should grant the loan. Parameters like credit score, loan amount, lifestyle, career, and assets are the deciding factors in getting the loan approved. If, in the past, people with parameters similar to yours have paid their dues timely, it is more likely that your loan would be granted as well.

Machine learning algorithms can exploit this dependency on past experiences and comparisons with other applicants and formulate a data science problem to predict the loan status of a new applicant using similar rules.

### Table of content

### Understanding and loading the data

The machine learning model is trained using the training data set. Every new applicant details filled at the time of application form acts as a test data set. On the basis of the training data sets, the model will predict whether a loan would be approved or not. We have 13 features in total out of which we have 12 independent variables and 1 dependent variable i.e. Loan_Status in train dataset and 12 independent variables in test dataset. The Loan_ID, Gender, Married, Dependents, Education, Self_Employed, Property_Area, Loan_Status are all categorical

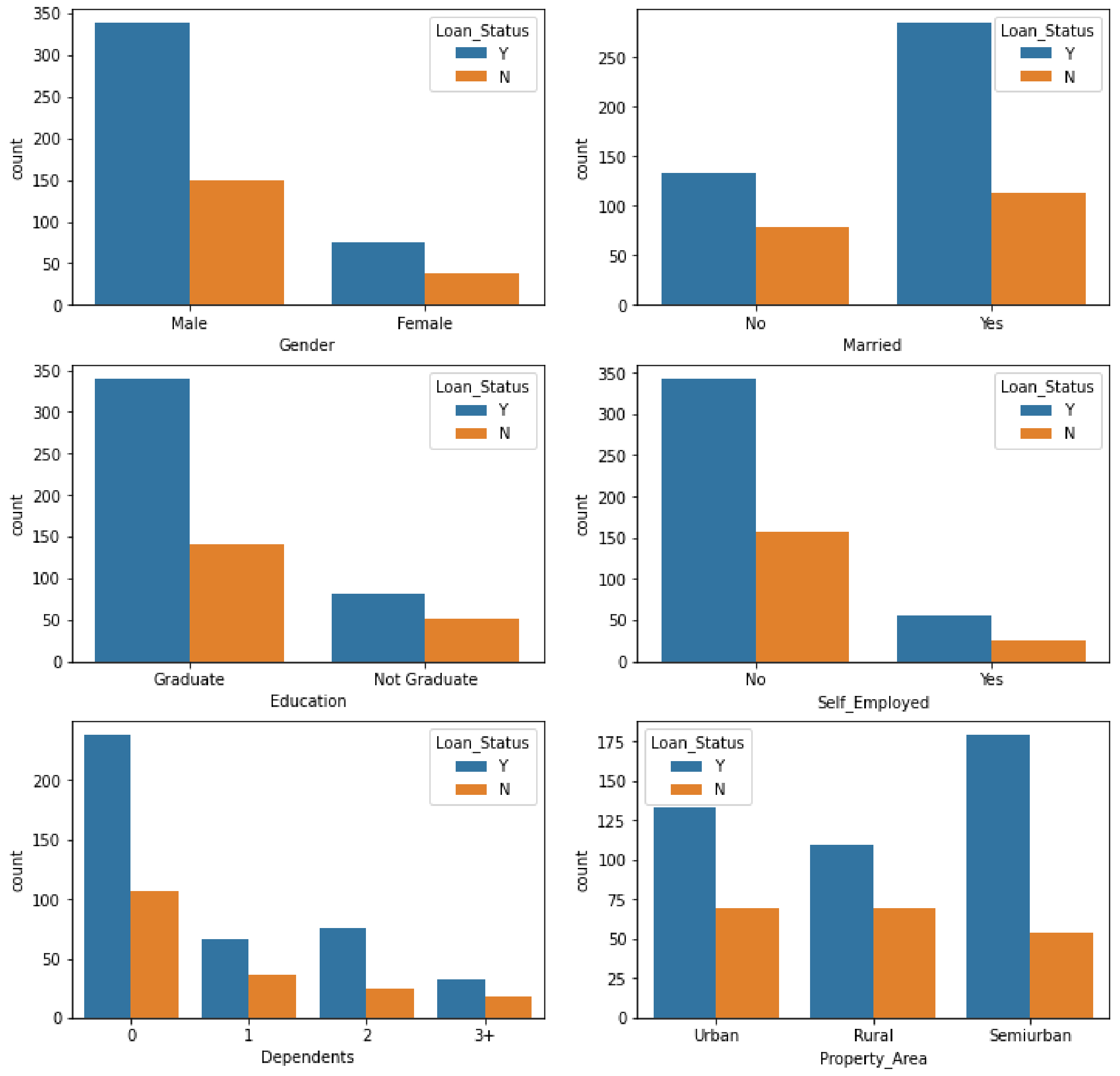
## Missing value and outlier treatment

```
Loan_ID 0
Gender 13
Married 3
Dependents 15
Education 0
Self_Employed 32
ApplicantIncome 0
CoapplicantIncome 0
LoanAmount 22
Loan_Amount_Term 14
Credit_History 50
Property_Area 0
Loan_Status 0
```

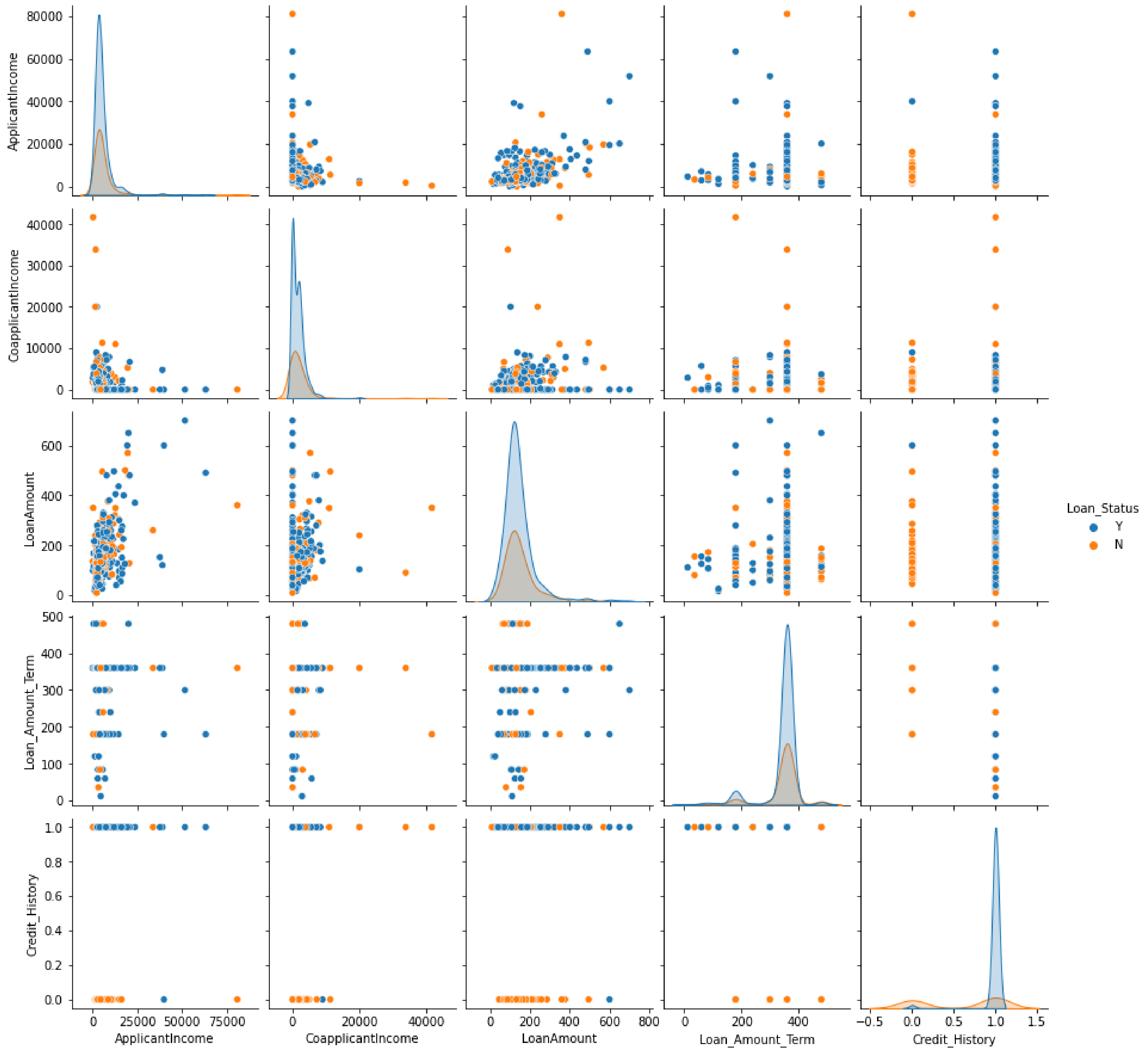
Here we can see that our dataset have some missing values that need to be treated. 7 columns have a non-zero number of NULL values, with Credit_History having the most (50). Given the size of our dataset, some of these columns have many empty fields, which we cannot handle by just removing the respective rows. Doing this will significantly decrease the size of the training dataset and adversely impact the model performance. Instead, we use null value treatment methods like replacing the values with the Mean or median of the column values. Here we treated the missing values using median. The categorical features such as Gender, Married, Self_Employed, Credit_History, and Loan_Status were OneHot Encoded, and the ordinal categorical features such as Dependents, Education and Property area were Label Encoded.

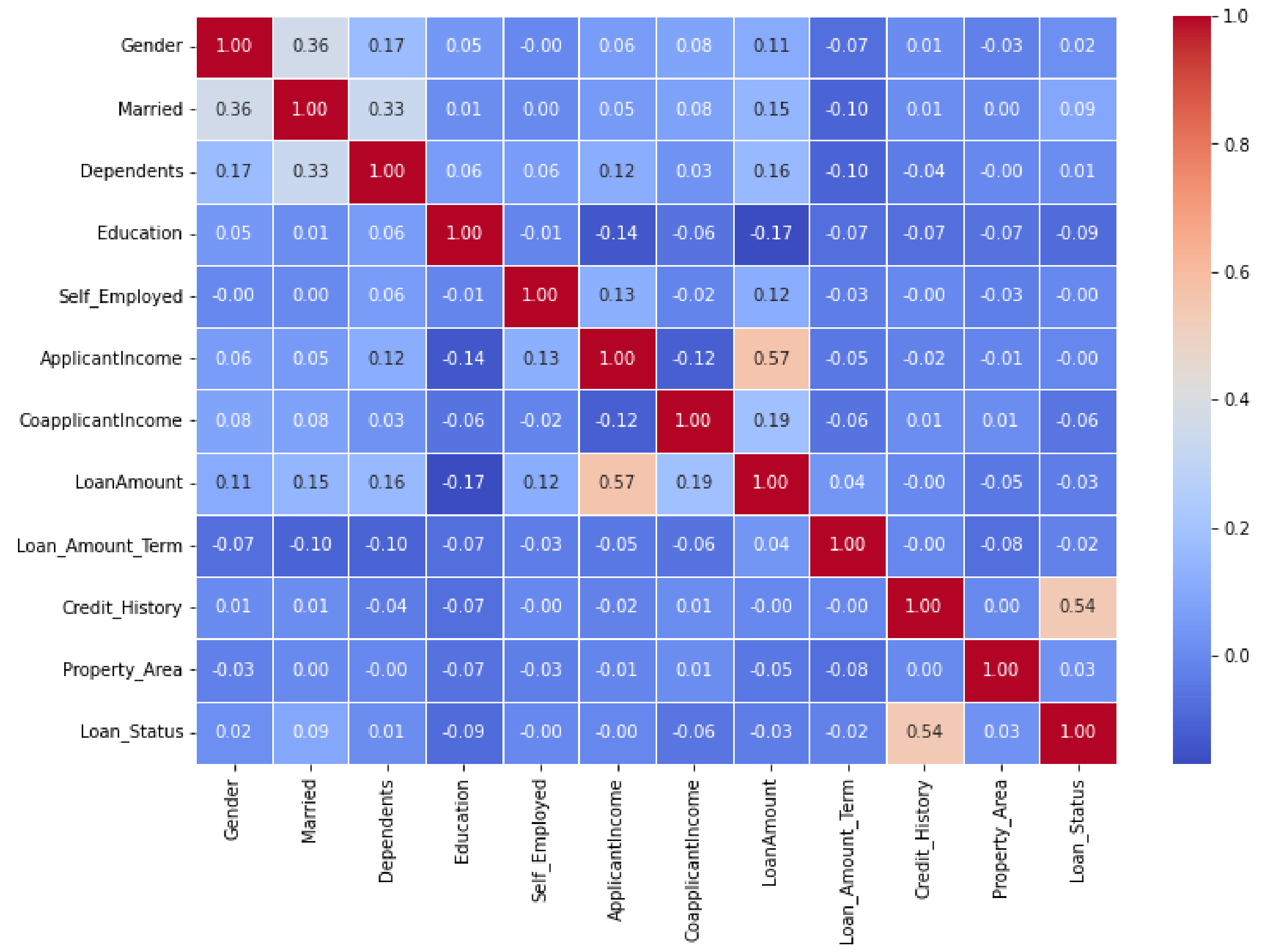
## Exploratory Data Analysis

Exploratory data analysis is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations. Following are some EDA that depicts the trend and pattern of our datasets.

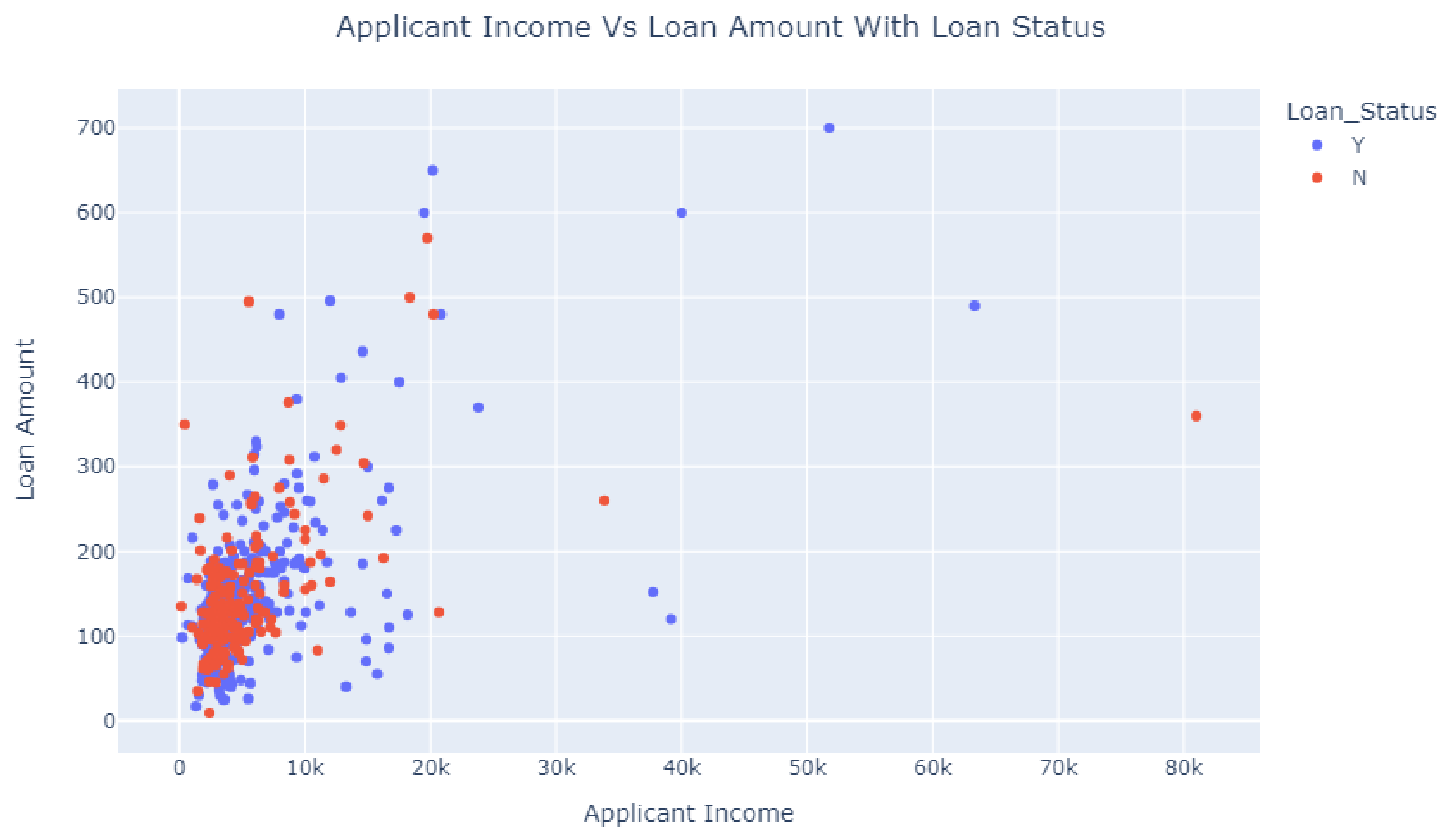


We can see that people who haven't graduated are far less likely to have their loans approved, also there exists a biad towards married couples over those who haven't married





A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. It is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses. From above correlation plot,we see that the most correlated variables are (ApplicantIncome — LoanAmount) and (Credit_History — Loan_Status). LoanAmount is also correlated with CoapplicantIncome.



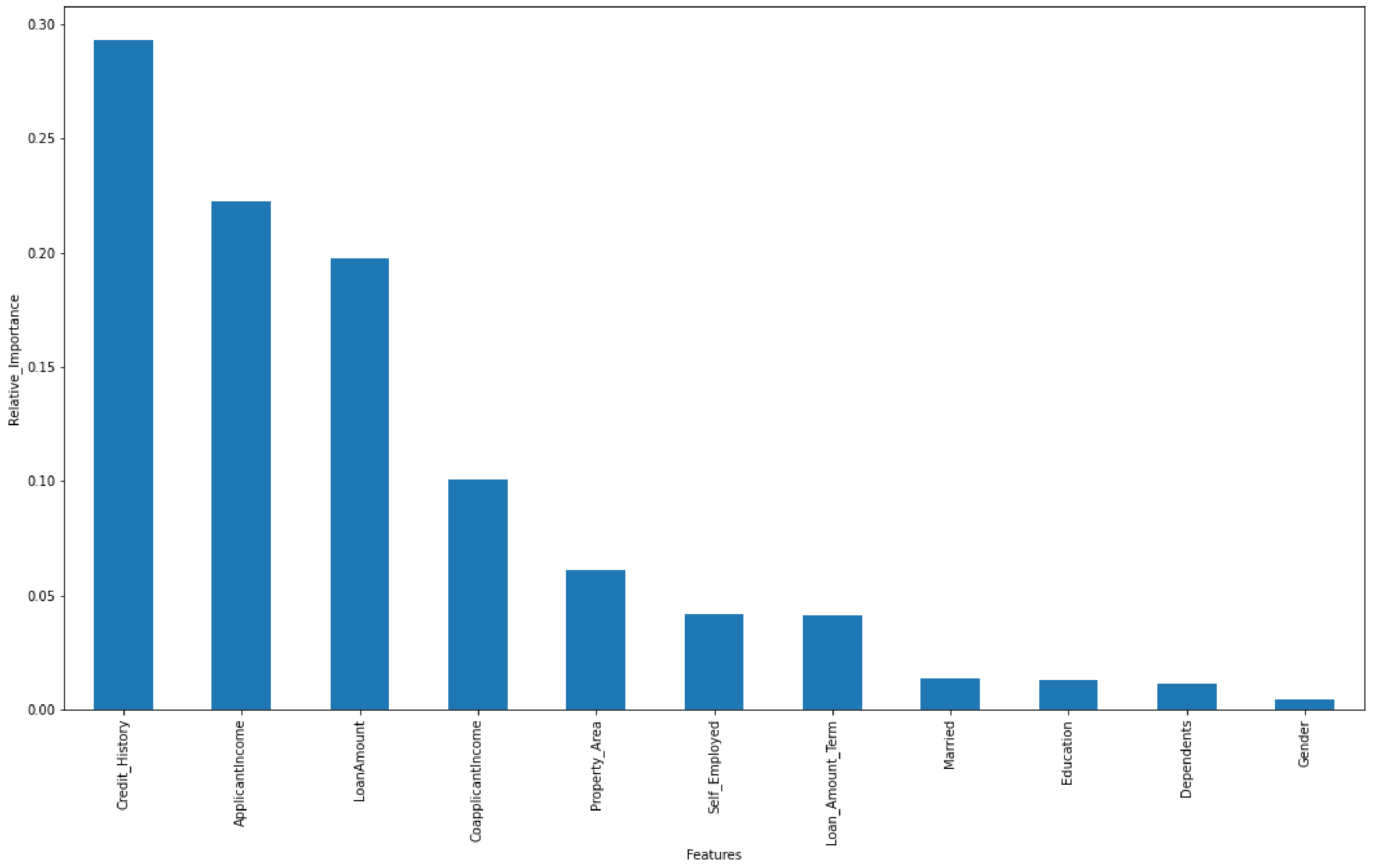
Above plot is the scatter plot that shows the relation between loan amount and applicant income.Here we can see that applicants having low income gets the low loan amount.

## Data Preprocessing and Modelling

Data preprocessing involves label encoding, handling missing values, selecting appropriate columns, normalization, and more. We completed handling all the missing values,converted all the categirical variables into numeric value and normalized the data. Using the popular train_test_split function from sklearn and a split ratio of 80:20, we create the train and test data sets.

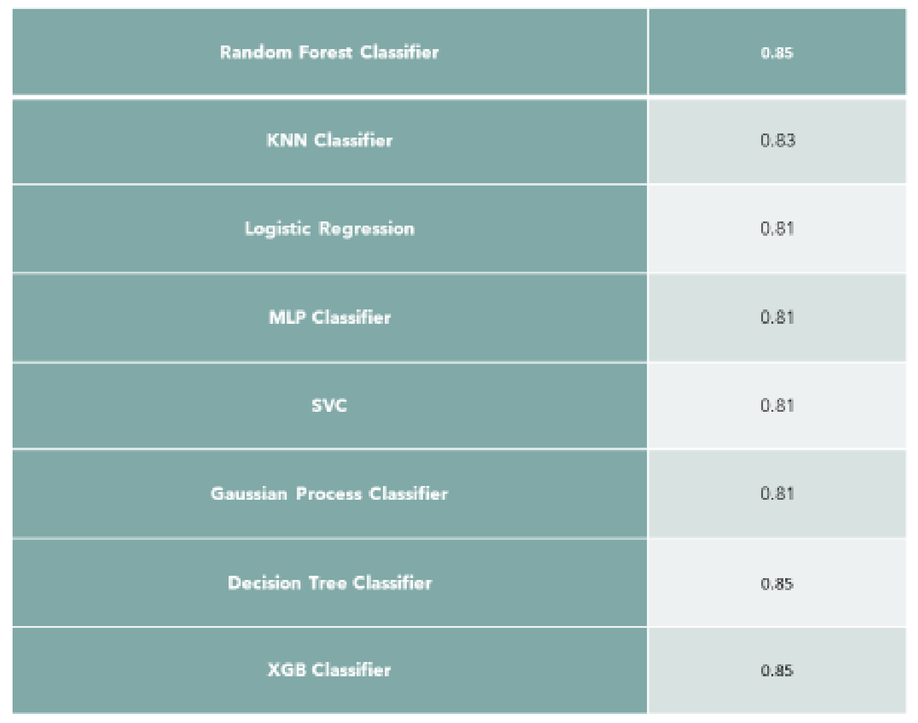
## Feature Selection

Feature selection is one of the crucial processes in any Machine Learning project, in which we reduce the number of input features to the predictive model.With the help of feature importance we can recognize the features that helped the most or were the most significant in adding information or decreasing the overall entropy of the tree.



Above feature importance plot depicts the credit history, applicants income and loan amount are the most important features for predicting loan approval.

## Results and Conclusion



| | |
|---|---|
| Random Forest Classifier | 0.85 |
| KNN Classifier | 0.83 |
| Logistic Regression | 0.81 |
| MLP Classifier | 0.81 |
| SVC | 0.81 |
| Gaussian Process Classifier | 0.81 |
| Decision Tree Classifier | 0.85 |
| XGB Classifier | 0.85 |

For predicting loan approval. We implemented various machine learning classifiers which are mentioned above and from above table we cam