Home Loan Application Prediction

Sheyda Saadat

Web Mining

May 08, 2024

Problem

Manual assessment of home loan applications has led to long processing times and uncertainty with stakeholders. By using machine learning and predictive models, lenders can assess the risks with each application and streamline the process of loan approval.

Introduction

This project aims to build a predictive model that categorizes loan applicants as approved or denied based on their credit profile. The dataset consists of 614 instances of home loan applications with information about the applicant's profile such as gender, marital status, credit history, and more. This report will explain the steps taken towards building a strong model by data cleaning, preprocessing, visualization, model training, and checking the performance of these models.

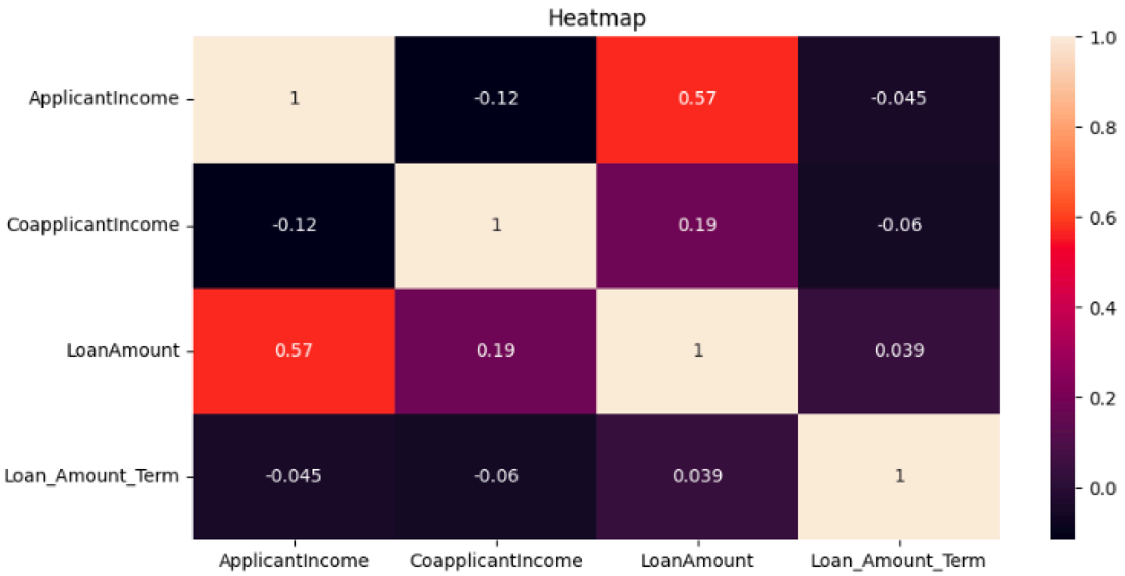
Libraries

Libraries used in this project include: pandas, matplotlib, seaborn, numpy, and sklearn.

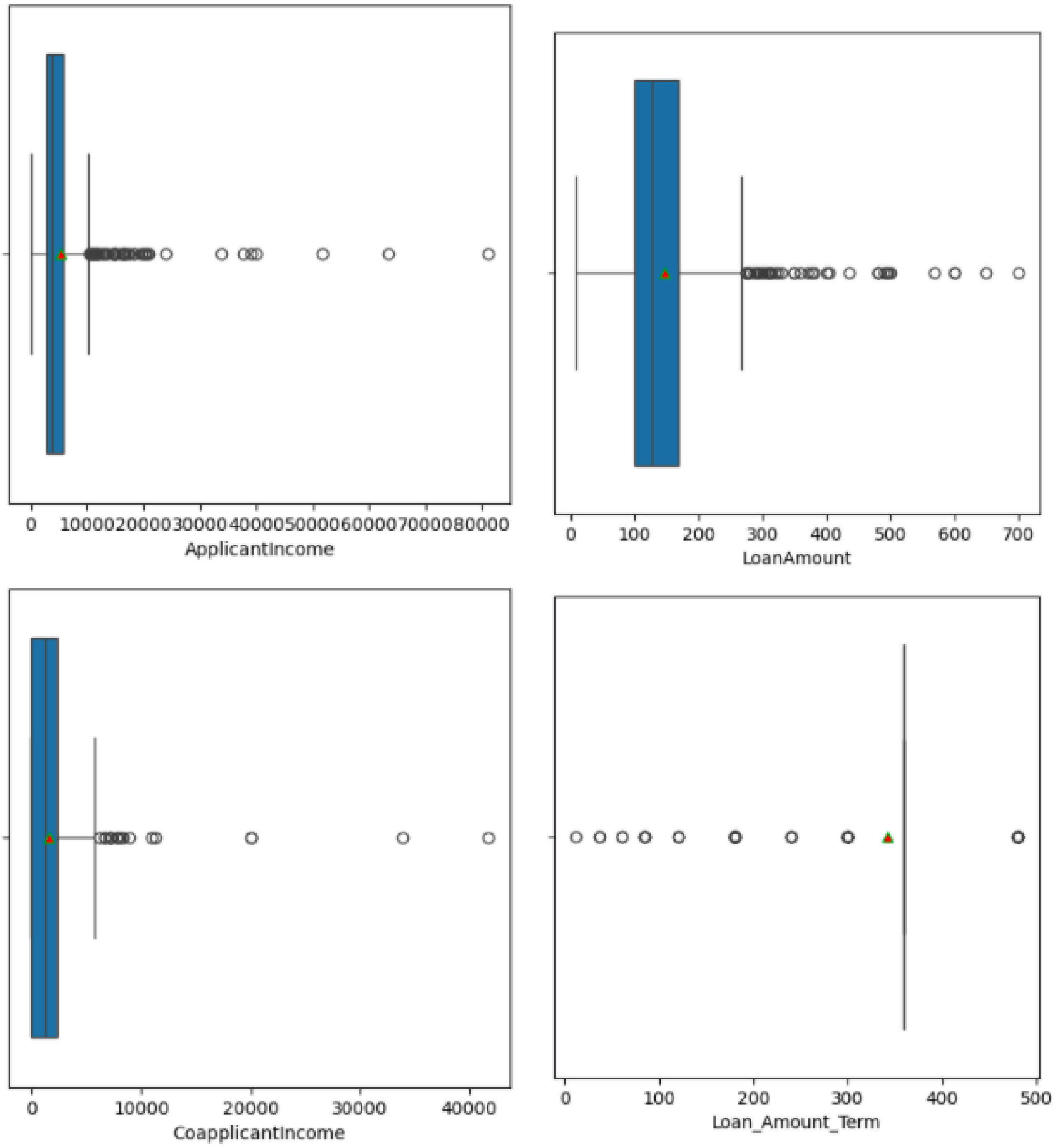
Exploratory Data Analysis (EDA)

To better understand the distribution of features, find patterns and relationships between features, and compare categorical data, data visualization tools such as box plots, heatmaps, and histograms are used.

The **heatmap** is used to get a general idea of the relationships between numeric variables. Based on this heatmap, LoanAmount has a positive correlation with Loan_Amount_Term, CoApplicantIncome, and most of all ApplicantIncome.

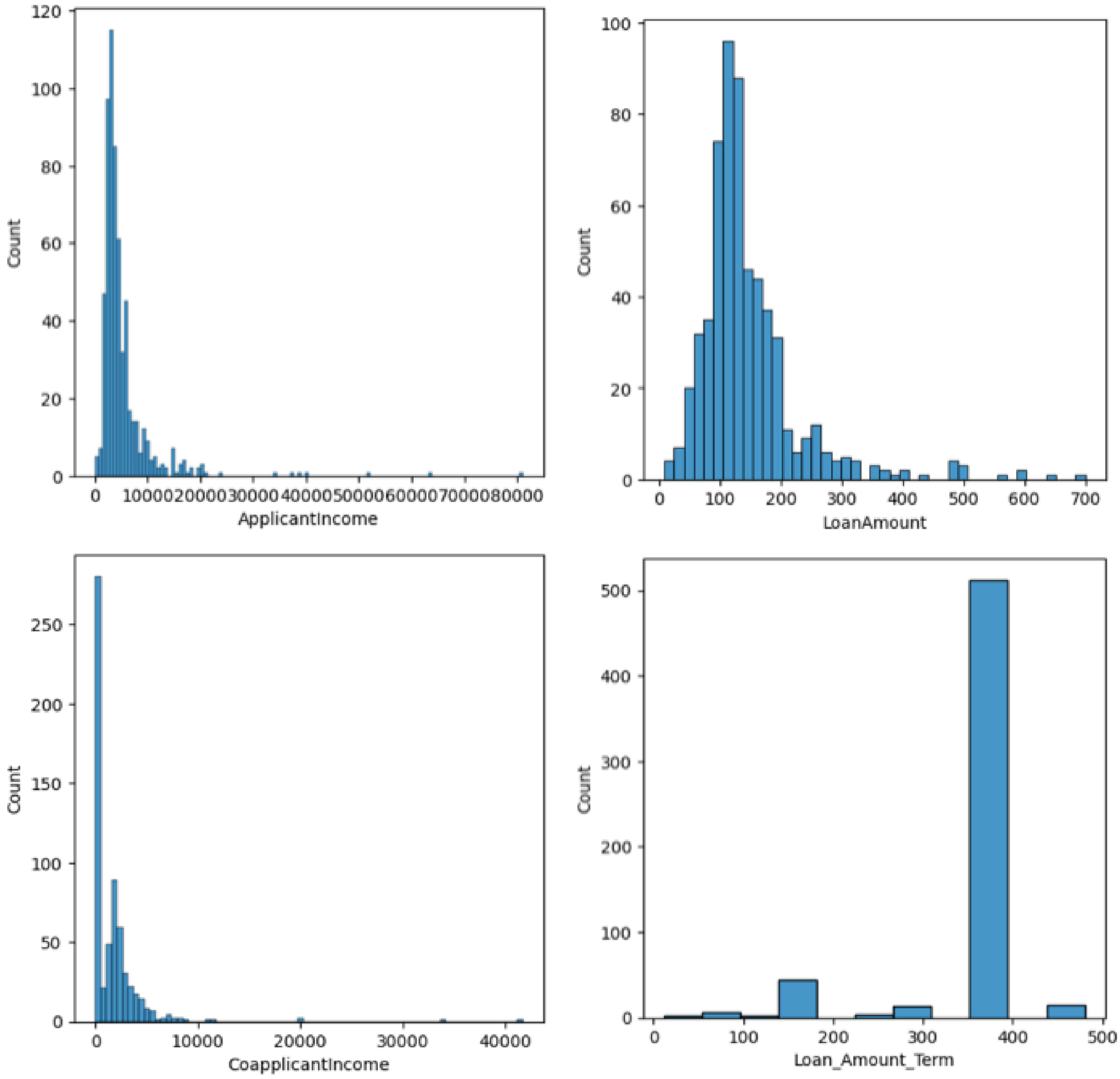


Boxplots are used to visualize the distribution of numeric data and detect any outliers. It can be seen that ApplicantIncome and LoanAmount have the most outliers.

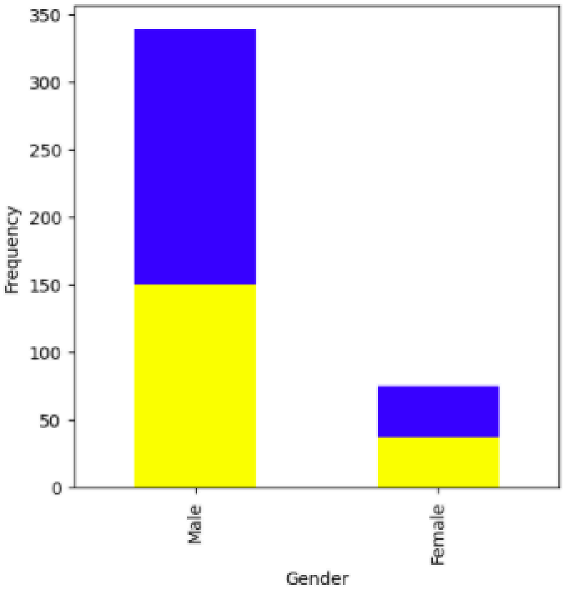

0
10000
20000
30000
40000
50000
60000
70000
80000
ApplicantIncome
0
100
200
300
400
500
600
700
LoanAmount
0
10000
20000
30000
40000
CoapplicantIncome
0
100
200
300
400
500
Loan_Amount_Term

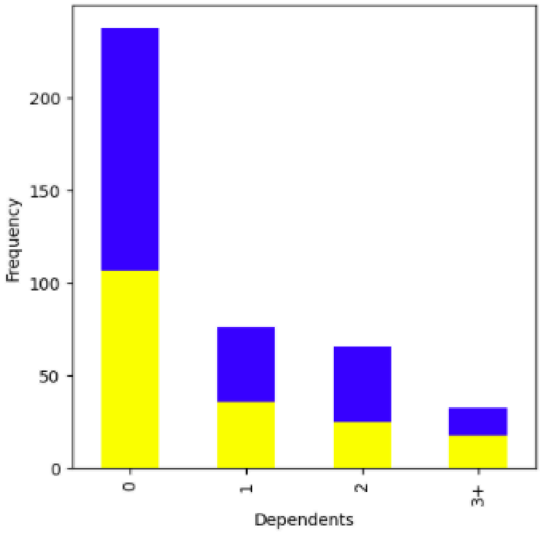
Histograms are used on numeric data to look at the distribution and skew of the data. Some outliers are still visible for ApplicantIncome and LoanAmount. It can also be seen that most applications do not have a co-applicant, the majority of loan terms are 12 months, and the majority of loan amounts are between 90-200 thousand dollars.

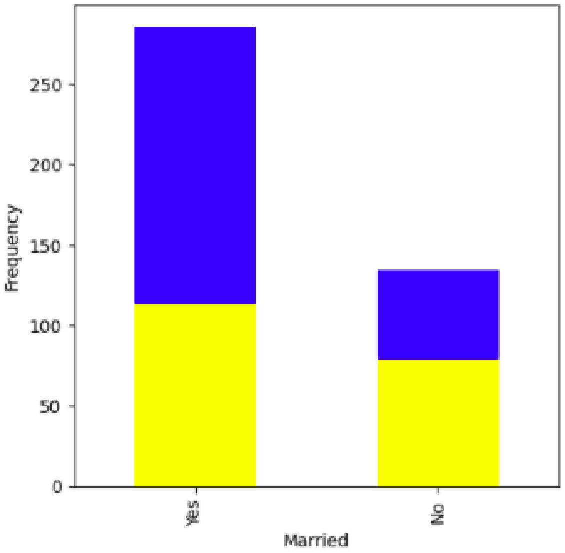
To remove the outliers, the standard deviation method is used which defines a threshold (three times the standard deviation) and removes data points outside of this range.

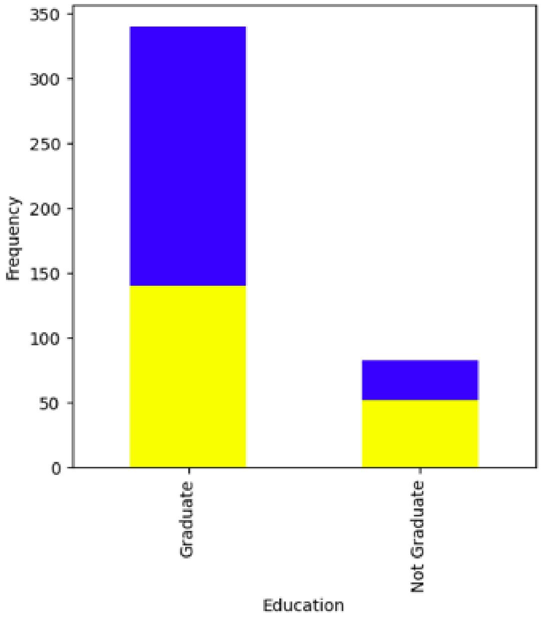


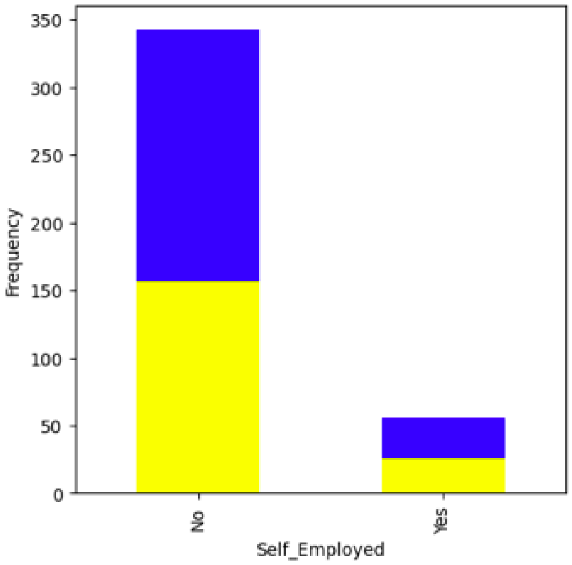
To compare categorical data and to visualize the acceptance rate between categories, **barcharts** are used. Based on these charts while there are more male applicants, the acceptance rate for both genders is about 50%, but we see more acceptance for married applicants rather than not married. We can also see that applicant's who are graduates are more likely to apply and get accepted for loans. Those with a bad credit history are not accepted at all.

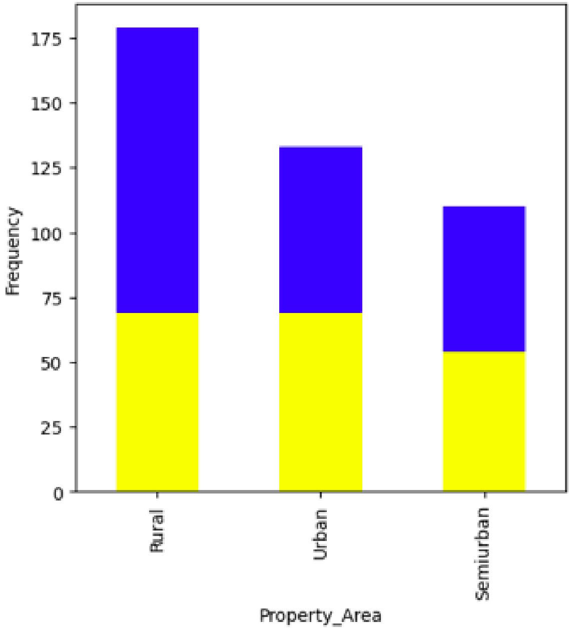

350
300
250
200
150
100
50
0
Frequency
Male
Female
Gender

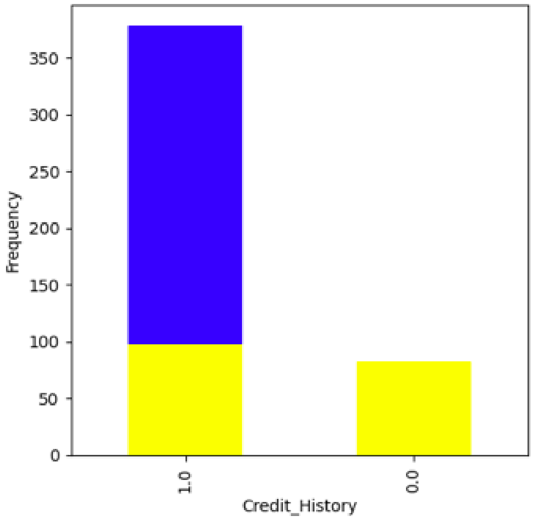

200
150
100
50
0
Frequency
0
1
2
3+
Dependents


250
200
150
100
50
0
Frequency
Yes
No
Married


350
300
250
200
150
100
50
0
Frequency
Graduate
Not Graduate
Education







Preprocessing

After exploring the dataset and understanding the features better, data preprocessing is done which includes data cleaning, dealing with missing values, and checking for duplicated rows.

The findings show that the dataset does not have any duplicated rows, but a significant amount of missing values were found which means deleting these values would not be beneficial to our model, so the missing values are filled by **imputing** them with mode for categorical features and mean for numeric ones. After imputing these values, the number of missing values is reduced to zero.

To ensure the algorithm can process our variables, categorical variables are encoded to convert them into numerical format. The numeric data is also scaled at this point.

Then, the data is split into a training set and a testing set for modeling and prediction, with 70% of the data in the training set and 30% of the data in the testing set to check the performance of our models.

The Models

The models I have chosen to explore in this project are:

- Logistic Regression
- Random Forest
- Support Vector Machine (SVM)
- Gaussian Naive Bayes

After predicting some of the test data, the following performance metrics are found for each model:

	Logistic Regression	Random Forest	SVM	Naive Bayes
Accuracy	86%	79%	82%	82%
Precision	84%	82%	82%	82%
Recall	100%	92%	97%	96%

Because of the nature of the problem, true positive and true negatives in the confusion matrix are explored as well, because the model aims to predict the correct amount of negatives to avoid giving a loan to a bad applicant as well as the correct amount of positives to avoid refusing a loan to a good applicant.

The Logistic Regression model has the most true negatives which is the most important at 125 counts and the highest true positive at 23 counts.

The **Logistic Regression** model is the best model for our problem with the most accuracy, precision, recall, true negative, and true positive.