Khoury College of Computer Science

CS 6220 - Data Mining

Home Loan Approval Prediction

Author:
Jiawei Tong

Date 04/16/2023

Table of Contents

List of Figures

List of Tables

1 Introduction

Home loan, or a mortgage, is a type of loan provided by financial institutions, such as banks or housing finance companies, to help individuals or families purchase a house. They are typically long-term loans, ranging from 5 to 30 years. Home loan approval can be a complicated process for individuals due to various factors that financial institutions consider when assessing loan applications. Besides, a home loan application can affect your credit score. When you apply for a home loan, the lender will typically request a copy of your credit report from one or more credit bureaus. This hard inquiry may cause your score to drop slightly in a short period of time. And multiple hard inquiries in a short period, such as applying for multiple home loans simultaneously, can have a more significant impact on the score.

The main objective of this project is to develop a model that can predict the eligibility of an individual for a home loan. The model will take into account various factors that are typically used by financial institutions to determine the eligibility of a borrower and provide predictions of the likelihood of a borrower being approved for a home loan, allowing borrowers to better understand their chances of obtaining financing for a home purchase.

2 Background

As the demand for home loans is increasing significantly, home loan approval prediction can assist lenders in making more thoughtful choices about loan applications, which is one of its main advantages. Lenders can find patterns and trends that might affect loan acceptance rates by looking at previous data on loan applications. And they can predict whether or not their application will be approved or not based on their situations. In order to more accurately evaluate risk and decide which loans to approve, they can focus on the variables that are most important of loan failure and improve their chances of loan approval. This predictive models can help them in mitigating risks of loan rejection or overborrowing.

3 Data Analysis and Preprocessing

The data is getting from Kaggle:
https://www.kaggle.com/datasets/rishikeshkonapure/home-loan-approval

Data Description:

Loan_ID: Unique Loan ID

Gender: Male/ Female

Married: Applicant married or not(Y/N)

Dependents: Number of dependents

Education: Applicant education level (Graduate/ Undergraduate)

Self_Employed: Self employed or not (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:

Loan approved or not(Y/N)

The dataset is imbalanced. 2/3 of the dataset has Approved Loan Status and 1/3 has Denied Loan Status.

4 Approach Overview

The data will be divided into training and testing sets for model development and validation. It will utilize different models and select one model with the highest accuracy.

Data Analysis: Explore the relationships within the dataset using data visualization to help us understand the data better. Methods include correlation analysis, scatter plots, heat maps, etc.

Data Cleaning: Fill missing values and incorrect values. Remove outliers.

Data Transformation: Normalization and standardization.

Classification Algorithms: Logistic Regression, Naive Bayes, Random Forest

Evaluation: Models will be evaluated based on its accuracy, recall, precision, F1-score .

5 Feature Engineering Implementation Specifics

- Remove null values
- Using log transformation to convert the skewed distribution of ApplicantIncome and TotalIncome into a more symmetrical one, which results in a more uniform spread of the data and can make it easier to model and analyze the data, as well as reduce the impact of outliers.
- Changing the string values in several columns to numerical values, which can be more useful in statistical analysis and modeling. The columns that are modified include Gender, Married, Self_Employed, Education, Loan_Status, and Dependents.
- Creating dummy variables for the categorical variable 'Property_Area'.
- Standardizing the scale of the numerical features in the dataset.

6 Results

(1) Logistic Regression Report:

	precision	recall	f1-score	support
0	0.88	0.40	0.55	35
1	0.74	0.97	0.84	61
accuracy			0.76	96
macro avg	0.81	0.68	0.69	96
weighted avg	0.79	0.76	0.73	96

```
Accuracy Score (Train): 0.8203125
Accuracy Score (Test): 0.7604166666666666
```

(2) Naive Bayes Report:

	precision	recall	f1-score	support
0	0.78	0.40	0.53	35
1	0.73	0.93	0.82	61
accuracy			0.74	96
macro avg	0.75	0.67	0.67	96
weighted avg	0.75	0.74	0.71	96

```
Accuracy Score (Train): 0.828125
```

```
Accuracy Score (Test): 0.7395833333333334
```

(3)Random Forest Report:

	precision	recall	f1-score	support
0	1.00	0.79	0.88	113
1	0.92	1.00	0.96	271
accuracy			0.94	384
macro avg	0.96	0.89	0.92	384
weighted avg	0.94	0.94	0.94	384

```
Accuracy Score (Train): 0.9375
Accuracy Score (Test): 0.7604166666666666
```

7 Conclusions

In conclusion, this project aimed to predict the approval of home loans based on all features of the dataset. Feature engineering techniques were applied to preprocess and transform the data. The accuracy of the models was improved after feature engineering. But the accuracy may be affected by the imbalanced dataset. It was observed that the imbalanced dataset exists, with class 1 (approved loans) having more instances than class 0 (rejected loans).

Three models were used for the prediction: Logistic Regression, Naive Bayes, and Random Forest. The accuracy scores on the training set for the Logistic Regression and Naive Bayes models were 82%, while the Random Forest model outperformed both with an accuracy score of 94%. The Random Forest model also had higher precision, recall, and f1-score values for both classes. Therefore, it can be inferred that the Random Forest model is the best model for predicting the approval of home loans based on the given features. Overall, the results of this project suggest that using machine learning models can be a useful tool for predicting the approval of home loans based on various features.