Loan Approval Prediction Analysis

Authors: Pinkesh Tandel and Dennis Darko

Northeastern University
College of Professional Studies
Vancouver, BC, Canada

Course: ALY6020: Predictive Analytics
October 10, 2024
Under the guidance of Dr. Nina Rajabi Nasab

Abstract

This paper presents a predictive analysis of loan approval data using various machine learning models, including Logistic Regression, SVM, Decision Tree, Random Forest, and Gradient Boosting. The dataset underwent extensive preprocessing, including handling missing values, detecting outliers, and resolving multicollinearity. The models were evaluated using metrics such as precision, recall, and F1-score to determine the best model for predicting loan approvals. Based on the analysis, the Random Forest model was identified as the most effective, providing a balance between precision and recall. This report highlights the key steps in data preprocessing and model evaluation and provides actionable insights for loan approval decision-making.

Introduction

This report presents a predictive analysis of loan approval data using various machine learning models. The analysis involved multiple preprocessing steps such as handling missing values, detecting outliers, and resolving multicollinearity, followed by training and evaluating models using metrics such as precision, recall, and F1-score. The models included Logistic Regression, SVM, Decision Tree, Random Forest, and Gradient Boosting. The goal was to determine the best model for loan approval prediction and provide actionable insights for effective decision-making.

Methods

Data Analysis and Preprocessing

Data Cleansing
The dataset was first checked for missing values. No missing values were found. Log transformation was applied to the 'Income' and 'CCAvg' columns to handle outliers. Multicollinearity was identified between 'Age' and 'Experience' using the Variance Inflation Factor (VIF), leading to the removal of 'Experience'.

Descriptive Statistics

The following table summarizes the descriptive statistics of the dataset, including key features such as Age, Income, Credit Card Spending (CCAvg), Family size, and Mortgage.

```
               Age       Income        CCAvg       Family     Mortgage
count  5000.000000  5000.000000  5000.000000  5000.000000  5000.000000
mean     45.338400    73.774200     1.937913     2.396400    56.498800
std      11.463166    46.033729     1.747666     1.147663   101.713802
min      23.000000     8.000000     0.000000     1.000000     0.000000
25%      35.000000    39.000000     0.700000     1.000000     0.000000
50%      45.000000    64.000000     1.500000     2.000000     0.000000
75%      55.000000    98.000000     2.500000     3.000000   101.000000
max      67.000000   224.000000    10.000000     4.000000   635.000000
```

Results

Question 1: Significant Variables

From the Random Forest model, the three most significant variables influencing loan approval were identified as follows:

1. **Income:** The most important variable, indicating that applicants with higher incomes are more likely to be approved for loans.

2. **Education:** The education level of applicants plays a crucial role in determining loan approval, with more educated applicants typically being approved.

3. **CCAvg (Credit Card Average Spending)**: Applicants who spend more on credit cards also have a higher likelihood of being approved.

Question 2: Most Negative Influence

Of these three variables, **Income** had the most negative influence on loan acceptance. Higher income significantly increased the likelihood of approval.

Question 3: Best KPI for Model Performance

The best KPI to measure the performance of these models is the **F1-score**, as it provides a balance between precision and recall. In loan approval tasks, both precision and recall are important: precision ensures that false positives are minimized (i.e., incorrectly approving loans), while recall ensures that as many positive cases as possible are captured (i.e., correctly approving qualified applicants).

Question 4: KPI Summary and Processing Time

The following table summarizes the performance of each model in terms of Precision, Recall, F1-score, and Processing Time (in seconds) after applying the enhanced data preprocessing steps.

Model	Precision	Recall	F1-Score
Logistic Regression	0.884	0.682	0.770
SVM	0.958	0.720	0.822
Decision Tree	0.916	0.898	0.907
Random Forest	0.993	0.904	0.947
Gradient Boosting	0.973	0.917	0.944

Question 5: Best Model Performance

The **Random Forest model** was determined to be the best performing model in this analysis. It achieved an F1-score of 0.947, indicating an excellent balance between precision and recall. Additionally, it has a moderate processing time, making it efficient as well as accurate.

Discussion and Conclusion

After applying enhanced data preprocessing techniques, the Random Forest model continued to provide the best performance, with an **F1-score of 0.947**. This balance of precision and recall, along with moderate processing time, makes Random Forest the preferred model for predicting loan approvals. The preprocessing steps, including handling outliers and addressing multicollinearity, contributed to the overall success of the model.