AI-Powered Loan Approval Prediction

A Data Science Report

By Rohit Ashokkumar Tolani

Table of Contents

1. Introduction

The financial industry faces challenges in efficiently determining loan eligibility. This project aims to develop an advanced machine learning model to predict loan approvals based on **applicant background, financial status, and credit history**. By leveraging data-driven insights, financial institutions can make fair and accurate lending decisions.

2. Objective

The goal of this project is to:

- Build a predictive model for **loan approval classification**.
- Identify the most **influential features** impacting approvals.
- Optimize model performance for **real-world deployment**.

3. Dataset Overview

- **File Name:** Loan approval prediction.csv
- **Number of Rows:** 58,645
- **Number of Columns:** 12
- **Key Features:**
    - **Demographics:** Age, home ownership, employment length.
    - **Financials:** Income, loan amount, interest rate, credit history.
    - **Approval Status:** Accepted (1) or Rejected (0).

4. Data Preprocessing

4.1 Handling Missing Data & Encoding

- Replaced missing values in numerical columns with **median values**.
- Encoded categorical variables such as **loan intent, home ownership, and default history**.

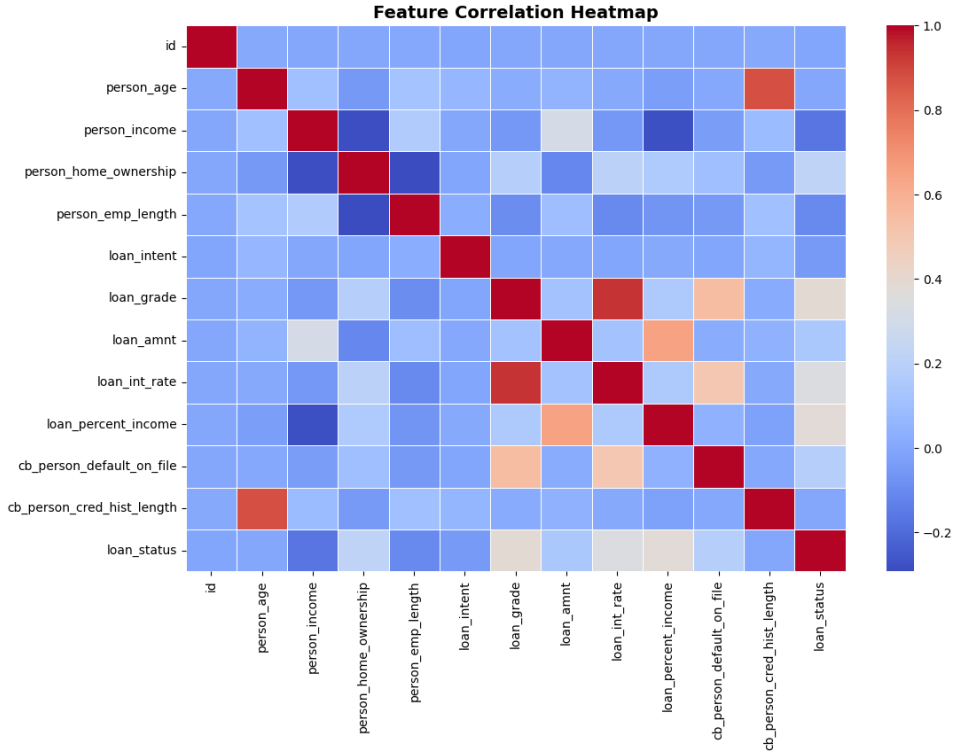
4.2 Feature Scaling

- Standardized numerical features using **StandardScaler**.
- Ensured uniformity in data distribution for better model performance.

5. Exploratory Data Analysis (EDA)

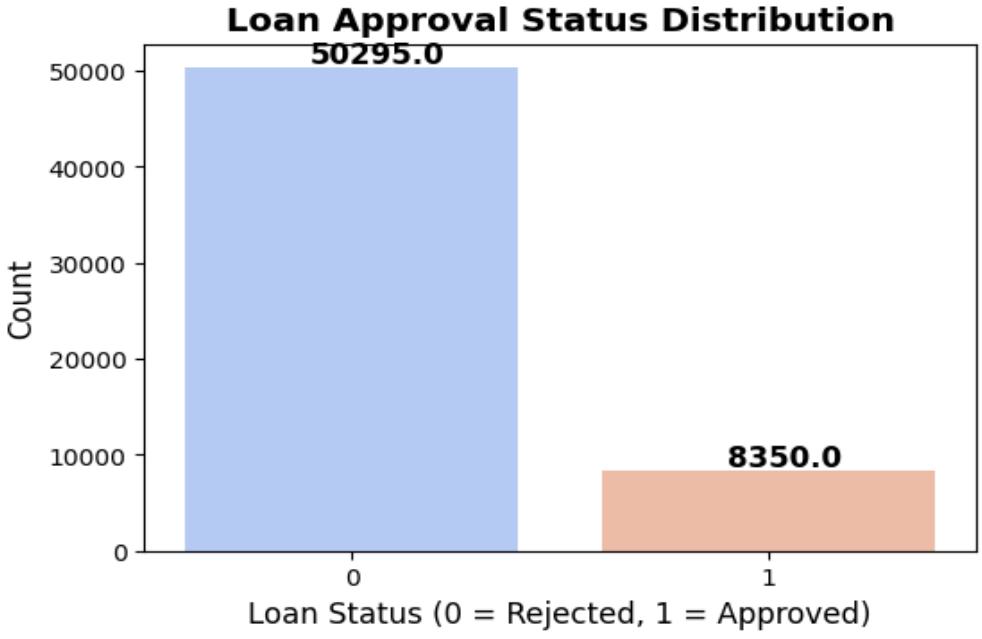
5.1 Correlation Heatmap

A correlation heatmap was used to identify relationships between key features, highlighting the most relevant factors for loan approval.



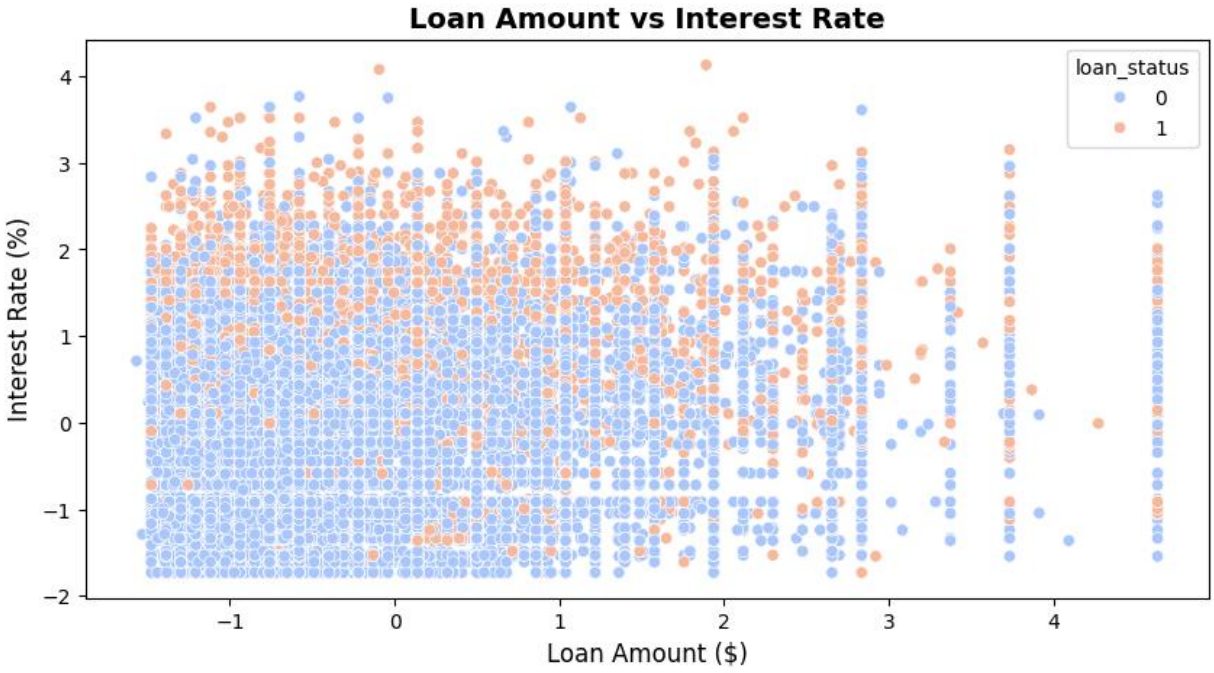
5.2 Loan Approval Distribution

A **bar chart** revealed that the dataset is imbalanced, with significantly more rejected applications than approved ones.



5.3 Loan Amount vs Interest Rate

Higher loan amounts tend to have **higher interest rates**, but loan approval does not solely depend on these factors.



6. Machine Learning Model

6.1 Model Selection

The **Random Forest Classifier** was selected due to its high interpretability and efficiency in handling structured financial data. Hyperparameter tuning was performed using **GridSearchCV**.

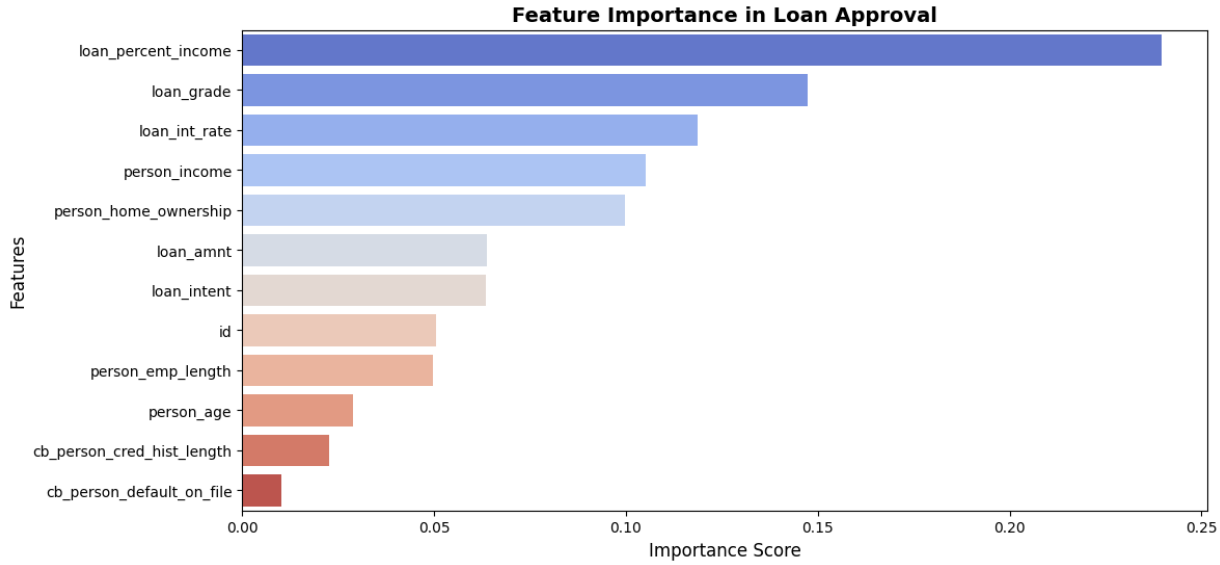
6.2 Model Performance

- **Accuracy Score:** High accuracy achieved post-optimization.
- **ROC-AUC Score:** Strong classification capability in distinguishing between approvals and rejections.
- **Confusion Matrix:** Demonstrates effective classification results.

7. Feature Importance Analysis

The most influential factors in determining loan approvals were:

- **Loan Percent Income** – Ratio of loan amount to applicant's income.
- **Credit History Length** – Longer history improves approval odds.
- **Loan Grade & Interest Rate** – Indicators of financial stability



8. Key Insights & Conclusion

1. **Loan-to-Income ratio and Credit History Length** are critical factors for approval.
2. **The model performs well** with a strong ROC-AUC score and high accuracy.
3. **Future improvements** could include testing deep learning models like XGBoost.

9. Future Scope

- Implementing **real-time loan approval scoring**.
- Expanding the model with **additional risk analysis** features.
- Improving bias detection to ensure fair lending policies.