Loan Approval Prediction Using Logistic Regression

Moavia Hassan 24-MS-DS-09

January 30, 2025

Abstract

In this project, I analyze factors influencing loan approval decisions using logistic regression. The dataset contains financial and demographic details of loan applicants, including income, loan amount, credit history, and categorical variables such as marital status, education level, and property location. My goal is to build a predictive model that helps us understand which factors play a crucial role in loan approvals.

I start by pre-processing and cleaning the data, handling missing values, encoding categorical features, and scaling numerical variables. Feature engineering techniques are used to improve model performance. A logistic regression model is trained and tuned to ensure that it produces consistent predictions. Model performance is evaluated using accuracy, precision, recall, and the F1 score. The best final model has an accuracy of around 78.86%, where credit history has turned out to be the most critical predictor followed by income and loan amount.

The results show the effectiveness of structured data analysis in helping financial institutions make sound lending decisions. The predictions of the model agree with the expectations in the real world, further underlining the importance of creditworthiness in the approval of loans. This research not only proves the efficiency of logistic regression in binary classification problems but also sets the ground for further research into automated loan risk assessment.

1 Introduction

1.1 Objective

The main goal of this project is to create a predictive model that predicts the probability of loan approval from applicant data. Using logistic regression, I seek to find prominent financial and demographic variables that are drivers of loan approval. The aim is to yield insights that can help financial institutions make more fact-based and equitable lending decisions.

1.2 Problem Statement

Loan approval is an important process within the financial industry, where faulty decisions can cause lenders to lose money or pass up opportunities for potential borrowers. Most loan approval procedures are based on manual evaluation, which is inefficient and subject to variability. The project solves the problem of loan approval automation and enhancement by creating a statistical model that forecasts loan eligibility from attributes of applicants such as income, credit history, and loan amount.

1.3 Scope

The project centers around supervised learning methods, specifically logistic regression, in predicting loan approval decisions. The data comprises real loan application histories, with numeric and categorical variables. The paper focuses on feature engineering, optimization, and explainability for facilitating practicality. The anticipated outcome is minimizing labor-intensive work during loan processing, enhanced decisional accuracy, and improved fairness of lending. .

1.4 Stakeholders

The findings of this project can benefit multiple stakeholders:

- **Financial Institutions:** Banks and lending agencies can utilize the predictive model to streamline loan approvals and minimize risks.
- **Loan Applicants:** Individuals seeking loans can gain a clearer understanding of eligibility criteria and factors influencing approval decisions.
- **Policy Makers:** Regulatory bodies can use insights from the model to assess biases in lending decisions and improve financial inclusion.
- **Data Scientists:** The study serves as a reference for professionals working on similar financial prediction problems.

By addressing these aspects, this project contributes to improving transparency and efficiency in the loan approval process.

2 Data Description

2.1 Data Sources

The dataset utilized for this project contains records of loan applications, collected from publicly accessible finance datasets. The data includes financial, demographic, and credit history information about the applicants. It is well structured and was collected in its initial form from lending organizations to carry out predictive models.

2.2 Data Types

The dataset primarily consists of structured data, with both numerical and categorical features. Key attributes include:

- **Numerical Data:** Applicant income, co-applicant income, loan amount, and monthly loan payment.
- **Categorical Data:** Credit history, education level, marital status, employment type, and property area.
- **Binary Encoded Data:** Gender, loan status (approved or not), and credit history status.

2.3 Data Preprocessing

To ensure the dataset was clean and suitable for modeling, several preprocessing steps were performed:

2.3.1 Missing Value Treatment

Missing values were identified and handled appropriately:

- Categorical features with missing values, such as credit history and employment type, were imputed using mode imputation.
- Numerical missing values, particularly in the loan amount variable, were filled using median imputation to reduce bias.

2.3.2 Data Cleaning

The dataset was examined for inconsistencies and redundant information:

- Duplicate entries were checked and removed.
- Outliers in income and loan amount were detected using interquartile range (IQR) analysis and addressed accordingly.

2.3.3 Feature Engineering

New features were derived to enhance predictive performance:

- Created a `Monthly_Loan_Payment` variable by dividing the loan amount by the loan term.
- Transformed the `Dependents` column from categorical values to numerical representations.
- Applied one-hot encoding to categorical variables such as property area, education level, and employment type.

2.3.4 Data Transformation

To improve model efficiency and performance, transformations were applied:

- Standardization was applied to numerical features (income, loan amount) to bring them to a common scale.
- Log transformation was tested on highly skewed variables to reduce variance.
- Min-max scaling was used for logistic regression to ensure optimal convergence.

These preprocessing steps were crucial in preparing the dataset for model training, ensuring better accuracy and interpretability in predictions.

3 Exploratory Data Analysis (EDA)

3.1 Data Summary

To understand the dataset, I first examined basic statistics and distributions of key features. Summary statistics revealed the following insights:

- The average applicant income is significantly higher than the co-applicant income, indicating that many applications were made individually rather than jointly.
- Loan amounts vary widely, with some extreme values suggesting potential outliers.
- The majority of applicants have a credit history, which could be a strong predictor of loan approval.
- Categorical variables like education level, marital status, and employment type exhibit imbalanced distributions, which needed to be addressed in modeling.

3.2 Visualization

To further explore patterns and relationships within the data, I utilized several visualizations:

- **Histograms:** Displayed the distribution of applicant income, loan amount, and monthly loan payment. The skewness in income and loan amounts suggested the need for log transformation.
- **Boxplots:** Helped identify outliers in numerical features, particularly in applicant income and loan amount.

- **Scatter Plots:** Examined the relationship between income levels and loan amount, revealing a weak correlation.
- **Correlation Matrix:** Showed that credit history had the strongest correlation with loan approval, reinforcing its importance as a predictor.
- **Count Plots:** Visualized categorical distributions, showing that married applicants and those with a positive credit history were more likely to get approved.

3.3 Key Insights

From the EDA, I derived several insights that influenced my modeling decisions:

- **Credit history is crucial:** A strong correlation with loan approval suggests it will be a key feature in model training.
- **Income and loan amount need scaling:** Due to high variance, standardization or normalization was necessary for better model performance.
- **Feature transformation improves interpretation:** Converting categorical variables to numerical and handling imbalances ensures better predictive accuracy.
- **Outliers may affect performance:** Extreme values in applicant income and loan amount required careful handling to prevent model distortion.

4 Methodology

4.1 Model Selection

For this project, I focused on a classification task to predict loan approval based on applicant and loan-related features. Initially, I considered multiple machine learning models, including Decision Trees, Random Forest, and Support Vector Machines. However, after evaluating their complexity, interpretability, and performance, I decided to use **Logistic Regression**.

Logistic Regression was chosen because:

- It provides a probabilistic interpretation of loan approval, which is useful in financial decision-making.
- It is computationally efficient and less prone to overfitting compared to complex models.
- The model coefficients allow for feature importance analysis, providing insights into the factors influencing loan approval.

4.2 Feature Selection

Selecting the right features was crucial for improving model performance. The process involved:

- **Correlation Analysis:** A correlation matrix was used to identify redundant features. Features highly correlated with each other were carefully analyzed to avoid multicollinearity.
- **Domain Knowledge:** Factors like credit history, income levels, and loan amount were prioritized based on their expected impact on loan approval.
- **One-Hot Encoding:** Categorical variables (e.g., property area, education level) were converted into binary features to make them usable in the model.
- **Scaling:** Since income and loan amounts had wide variations, I applied feature scaling to standardize these numerical variables.

4.3 Evaluation Metrics

To assess model performance, I relied on multiple evaluation metrics:

- **Accuracy:** The proportion of correctly classified instances out of all predictions.
- **Precision:** The proportion of correctly predicted positive cases (loan approvals) among all predicted positive cases.
- **Recall:** The proportion of correctly predicted loan approvals out of all actual loan approvals.
- **F1-score:** The harmonic mean of precision and recall, providing a balanced measure of model performance.
- **Confusion Matrix:** A breakdown of true positives, false positives, true negatives, and false negatives to analyze misclassifications.

5 Model Building

5.1 Training and Testing Split

To ensure the model's generalizability, I split the dataset into training and testing sets. The training set was used to learn patterns from the data, while the testing set evaluated how well the model performed on unseen instances. A common 80-20 split was chosen, ensuring that the model had sufficient data for learning while maintaining a reliable test set for evaluation.

5.2 Model Implementation

I implemented a logistic regression model due to its simplicity, interpretability, and effectiveness for binary classification problems. The data was preprocessed, ensuring that categorical variables were encoded, and numerical features were appropriately scaled. Using the scikit-learn library, I trained the model on the training data, optimizing it to differentiate between approved and rejected loan applications based on key financial and demographic factors.

5.3 Hyper-parameter Tuning

To enhance the model's predictive performance, I fine-tuned hyper-parameters such as the regularization strength using Grid Search Cross-Validation. This approach systematically explored a range of values to identify the optimal configuration that balanced bias and variance. Ultimately, the best-performing model was selected based on accuracy and other relevant evaluation metrics.

6 Model Evaluation

6.1 Performance Metrics

To assess the effectiveness of the logistic regression model, I evaluated its performance using accuracy, precision, recall, and F1-score. The confusion matrix provided insights into classification errors, revealing that while the model achieved an overall accuracy of 78.86%, it struggled with classifying certain instances of the minority class.

6.2 Comparison of Models

Since logistic regression was the only model considered, no direct comparisons with other algorithms were made. However, different preprocessing and feature selection techniques were tested to optimize its performance.

6.3 Validation

I implemented cross-validation to ensure that the model's performance was stable across different subsets of the data. This helped in reducing the risk of overfitting and ensured robustness in predictions.

6.4 Error Analysis

Examining misclassified cases from the confusion matrix, I observed that false negatives were minimal, which is crucial for loan approval decisions. However, the false positive rate was relatively high, indicating that some non-eligible applicants were classified as eligible, which could be a concern for financial institutions.

7 Results and Discussion

7.1 Final Model Performance

The logistic regression model achieved an accuracy of 78.86%, with a high recall for the positive class. This suggests that the model effectively identifies applicants who are likely to get loan approval.

7.2 Key Findings

- Income-related features had a significant impact on loan approval predictions. - Credit history played a crucial role in determining loan eligibility. - Feature scaling and preprocessing improved model performance.

7.3 Limitations

- The model assumes a linear relationship between predictors and the target variable. - Imbalanced class distribution may have influenced the model's ability to generalize well. - External economic factors, which could impact loan approvals, were not included in the dataset.

7.4 Business Implications

The findings from this project can help financial institutions refine their loan approval processes by focusing on key applicant attributes. Improving data collection on financial history and employing more advanced models could further enhance decision-making accuracy.

8 Conclusion

8.1 Summary of Findings

The project demonstrated that logistic regression is effective for loan approval prediction.

8.2 Recommendations

Financial institutions should incorporate additional applicant details to enhance predictions.

8.3 Future Work

Future improvements could include exploring deep learning techniques and additional features.

9 Appendix A: Additional Tables and Figures

Here, I provide supplementary materials that support the analysis but were too detailed for the main report.

Table 1 shows additional statistical insights.

Feature	Mean	Standard Deviation
ApplicantIncome	5400.5	6100.3
CoapplicantIncome	1625.4	2390.8
LoanAmount	146.2	85.6
Monthly_Loan_Payment	1.18	0.62

Table 1: Additional summary statistics