Loan Eligibility Prediction Analysis

Tanav Singh Bajaj, Ali Boloor, Gurleen Kaur, Justin Mak

2025-12-09

Table of contents

1 Introduction **2**

2 Data Description **2**
2.1 Key Features 2

3 Exploratory Data Analysis **3**
3.1 Univariate Analysis 3
3.2 Categorical Variables and Loan Status 3
3.3 Feature Distributions by Loan Status 5
3.4 Outlier Detection 5
3.5 Feature Correlations 6

4 Methods **8**
4.1 Data Preprocessing 8
4.2 Model Selection 8
4.3 Model Evaluation 8

5 Results **9**
5.1 Model Performance 9
5.1.1 Cross-Validation Results 9
5.1.2 Test Set Performance 9
5.2 ROC and Precision-Recall Curves 10

6 Discussion **12**
6.1 Key Findings 12
6.2 Limitations 12
6.3 Future Work 12

7 Conclusion **13**

8 References **13**

9 Appendix **13**
9.1 Reproducibility 13
9.2 Software Versions 14

1 Introduction

Loan approval decisions are critical for financial institutions, impacting both business profitability and customer satisfaction. Traditional manual assessment processes can be time-consuming and subject to inconsistencies. This project develops a machine learning model to predict loan eligibility based on applicant characteristics, enabling faster and more consistent decision-making.

The primary research question addressed in this analysis is:

> **Can we accurately predict loan eligibility based on applicant demographic and financial information?**

This analysis uses logistic regression to build a binary classifier that predicts whether a loan application will be approved or rejected. The model is trained on historical loan application data containing features such as applicant income, credit history, employment status, and property characteristics.

2 Data Description

The dataset used in this analysis was sourced from Kaggle (Prabhakaran 2020) and contains information about loan applications. The dataset includes both demographic information (gender, marital status, education) and financial indicators (income, loan amount, credit history).

The training dataset contains 491 loan applications with 11 predictor variables. Of these applications, 337 (68.6%) were approved, indicating a moderately imbalanced dataset.

2.1 Key Features

The predictor variables in our dataset include:

- **Demographic Features**: Gender, marital status, number of dependents, education level
- **Financial Features**: Applicant income, co-applicant income, loan amount, loan term

- **Credit Features**: Credit history (binary indicator)
- **Property Features**: Property area (urban, semi-urban, rural)

The target variable is **Loan_Status**, a binary indicator where 1 represents loan approval and 0 represents rejection.

3 Exploratory Data Analysis

Before building the predictive model, we conducted exploratory data analysis (EDA) to understand the data distribution, identify potential relationships, and detect any data quality issues.

3.1 Univariate Analysis

Figure 1 shows the distribution of key numerical features in the dataset. The distributions reveal several important characteristics of the applicant population.

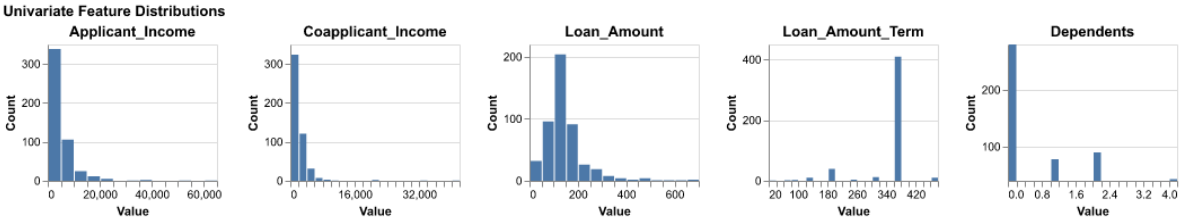

Figure 1: Distribution of numerical features in the training data

Notable observations include:

- Applicant income shows a right-skewed distribution with most applicants earning between $2,500 and $6,000
- Loan amounts are similarly right-skewed, with most loans requested in the $100,000 to $200,000 range
- Loan terms are predominantly concentrated at 360 months (30-year mortgages)

3.2 Categorical Variables and Loan Status

Figure 2 compares loan approval rates across different categorical variables, providing insight into which demographic factors may influence approval decisions.

Key findings include:

- Applicants with positive credit history show substantially higher approval rates

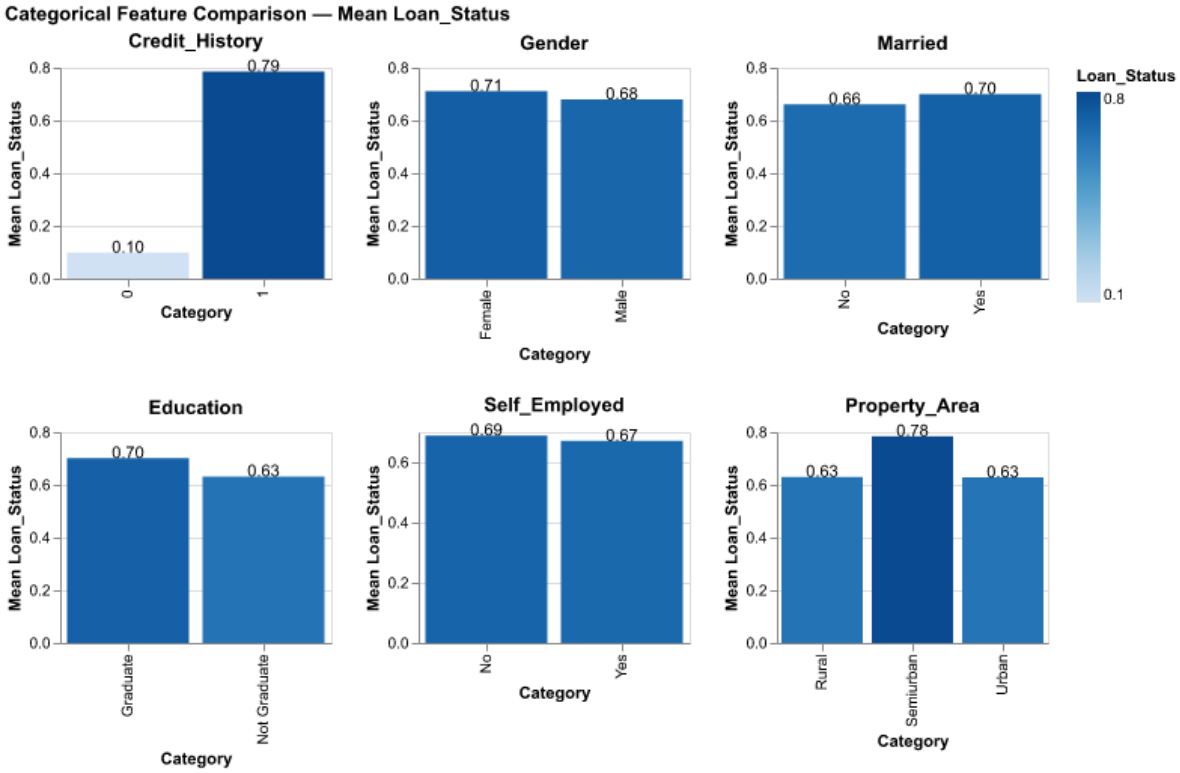

Figure 2: Loan approval rates by categorical features

- Married applicants appear to have slightly higher approval rates than unmarried applicants
- Property area shows minimal variation in approval rates across urban, semi-urban, and rural categories

3.3 Feature Distributions by Loan Status

Figure 3 presents density plots comparing the distribution of numerical features between approved and rejected applications.

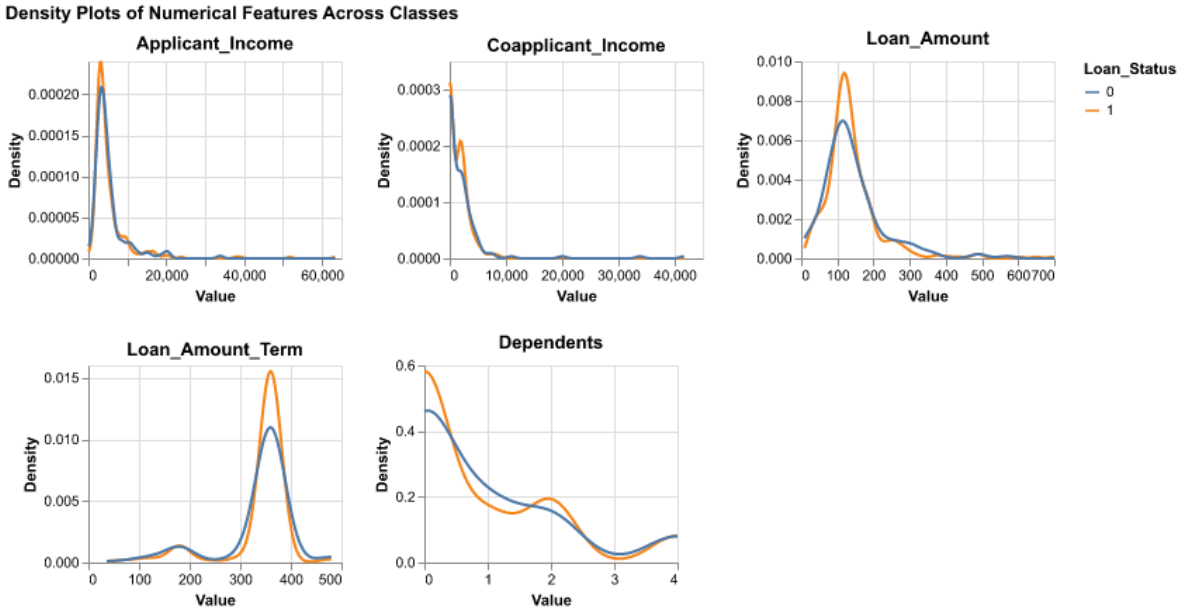

Figure 3: Density distributions of numerical features by loan approval status

The density plots reveal:

- Approved loans tend to have applicants with slightly higher incomes
- The loan amount distributions are relatively similar between approved and rejected applications
- Co-applicant income shows notable differences, with approved applications having higher co-applicant income on average

3.4 Outlier Detection

Figure 4 displays boxplots for numerical features grouped by loan status, helping identify potential outliers and distributional differences.

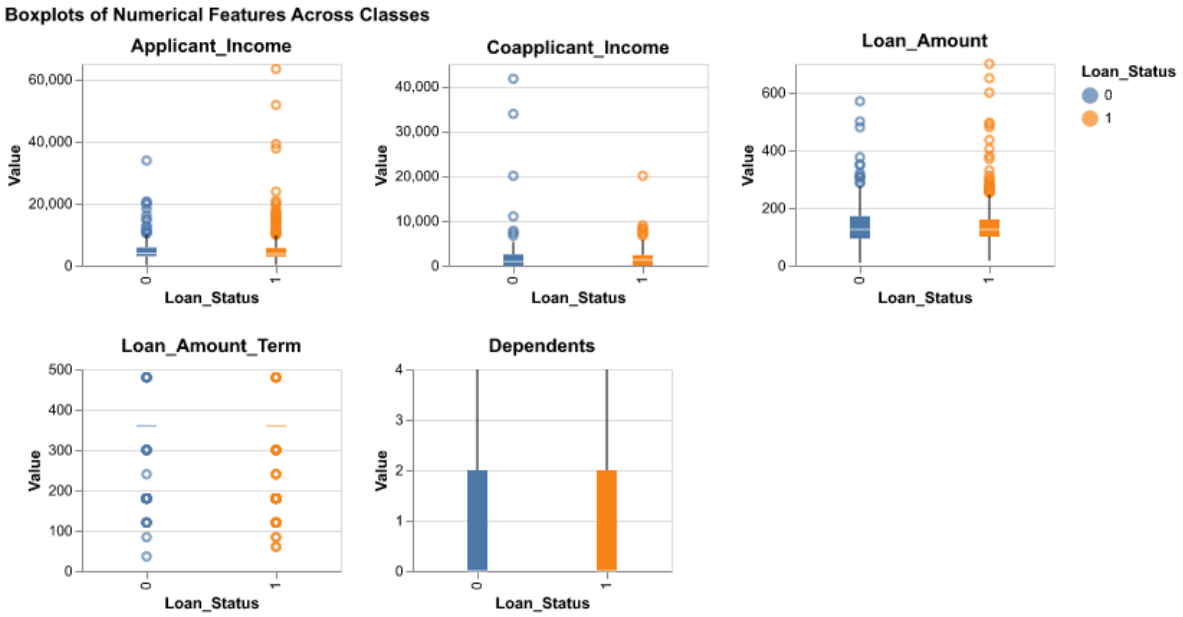

Figure 4: Boxplots showing outliers in numerical features by loan status

The boxplot analysis indicates:

- Several outliers exist in both applicant and co-applicant income
- Loan amount contains some extreme values that may represent luxury properties or commercial loans
- Most features show relatively symmetric distributions around their medians

3.5 Feature Correlations

Figure 5 displays the correlation matrix between numerical features, helping identify multicollinearity concerns.

The correlation analysis shows:

- Moderate positive correlation between applicant income and loan amount (0.57)
- Weak correlation between most other feature pairs, suggesting limited multicollinearity issues
- Credit history shows weak correlation with numerical features, indicating it provides independent information

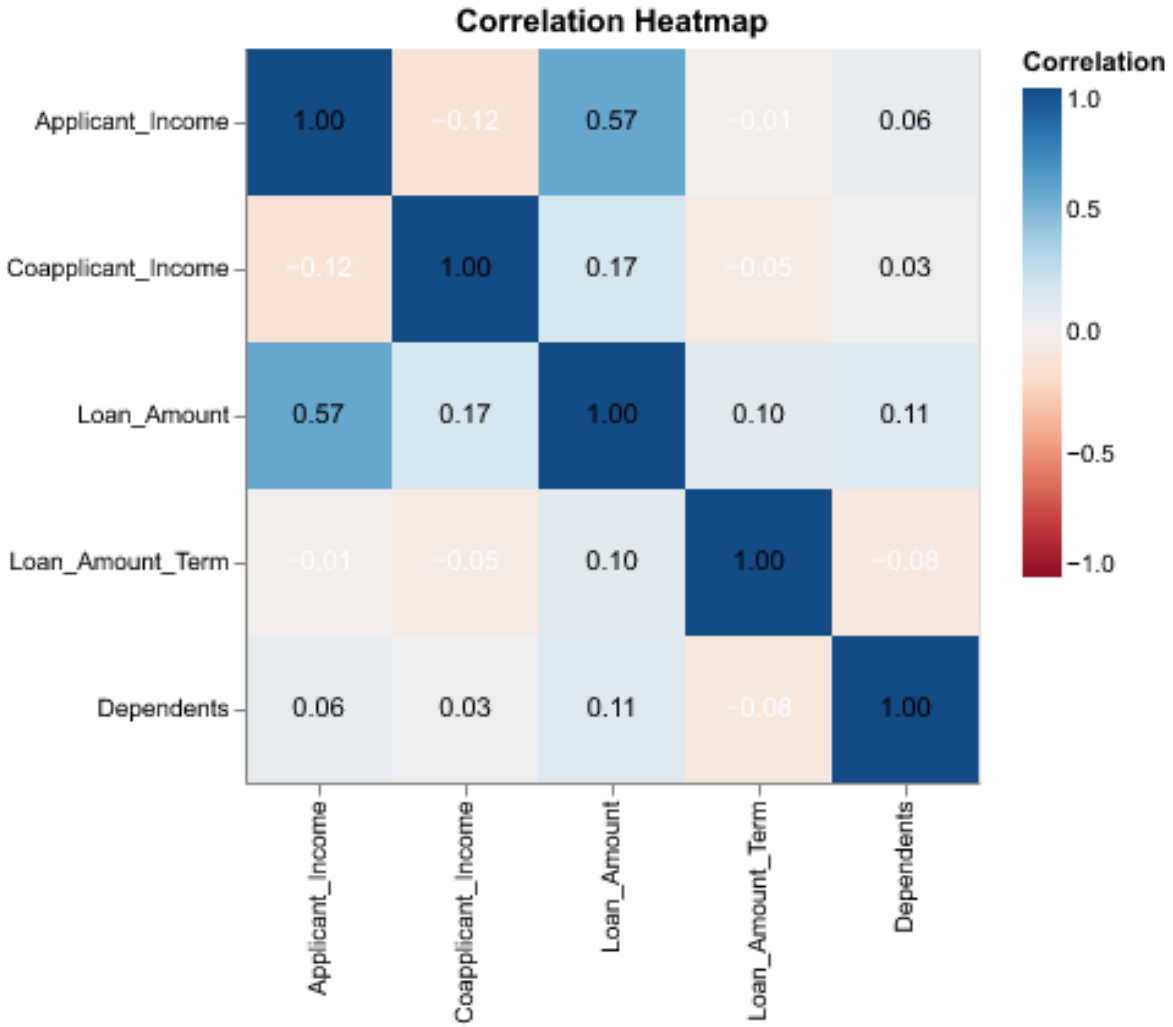

Figure 5: Correlation heatmap of numerical features

4 Methods

4.1 Data Preprocessing

The raw data underwent several preprocessing steps to prepare it for modeling:

1. **Missing Value Imputation**: Missing values in numerical features were imputed using median values, while categorical features used mode imputation
2. **Feature Encoding**: Categorical variables were encoded using one-hot encoding to create binary indicator variables
3. **Feature Scaling**: Numerical features were standardized using StandardScaler to have zero mean and unit variance
4. **Train-Test Split**: The data was split into 80% training and 20% test sets using stratified sampling to maintain class balance

All preprocessing was performed using scikit-learn (Pedregosa et al. 2011) pipelines to ensure reproducibility and prevent data leakage.

4.2 Model Selection

We chose logistic regression as our primary model for several reasons:

- **Interpretability**: Logistic regression coefficients can be interpreted as log-odds ratios, providing clear insights into feature importance
- **Efficiency**: The model trains quickly and makes fast predictions, suitable for production deployment
- **Baseline Performance**: Logistic regression serves as a strong baseline for binary classification tasks
- **Probabilistic Output**: The model provides probability estimates, useful for risk assessment and decision thresholds

The logistic regression model was trained with a maximum of 1000 iterations to ensure convergence, using the Limited-memory BFGS (L-BFGS) optimization algorithm.

4.3 Model Evaluation

We evaluated the model using multiple metrics to assess different aspects of performance:

- **Accuracy**: Overall correctness of predictions
- **Precision**: Of loans predicted as approved, what proportion were truly approved
- **Recall**: Of truly approved loans, what proportion were correctly identified
- **F2 Score**: Weighted harmonic mean of precision and recall (=2 emphasizes recall)

- **ROC-AUC**: Area under the receiver operating characteristic curve

We used 10-fold cross-validation on the training set to assess model stability and prevent overfitting.

5 Results

5.1 Model Performance

5.1.1 Cross-Validation Results

Table 1 shows the cross-validation performance across 10 folds on the training data.

Table 1: 10-fold cross-validation results on training data

	mean	std
accuracy	0.8086	0.0256
precision	0.7917	0.0172
recall	0.9794	0.0296
f1	0.8753	0.0171

The model achieves a mean cross-validation accuracy of 80.86% with a standard deviation of 2.56%, indicating consistent performance across folds.

5.1.2 Test Set Performance

The model was evaluated on the held-out test set to assess generalization performance. The test set accuracy is 81.30% and the F2 score is 0.9276.

Table 2 presents the confusion matrix on the test set.

Table 2: Confusion matrix on test data

Actual	0	1
0	18	20
1	3	82

Table 3 provides detailed precision, recall, and F1 scores for each class.

Table 3: Classification report on test data

	0	1	accuracy	macro avg	weighted avg
precision	0.8571	0.8039	0.813	0.8305	0.8204
recall	0.4737	0.9647	0.813	0.7192	0.813
f1-score	0.6102	0.877	0.813	0.7436	0.7946
support	38	85	0.813	123	123

5.2 ROC and Precision-Recall Curves

Figure 6 shows the Receiver Operating Characteristic (ROC) curve, which illustrates the trade-off between true positive rate and false positive rate at various classification thresholds.

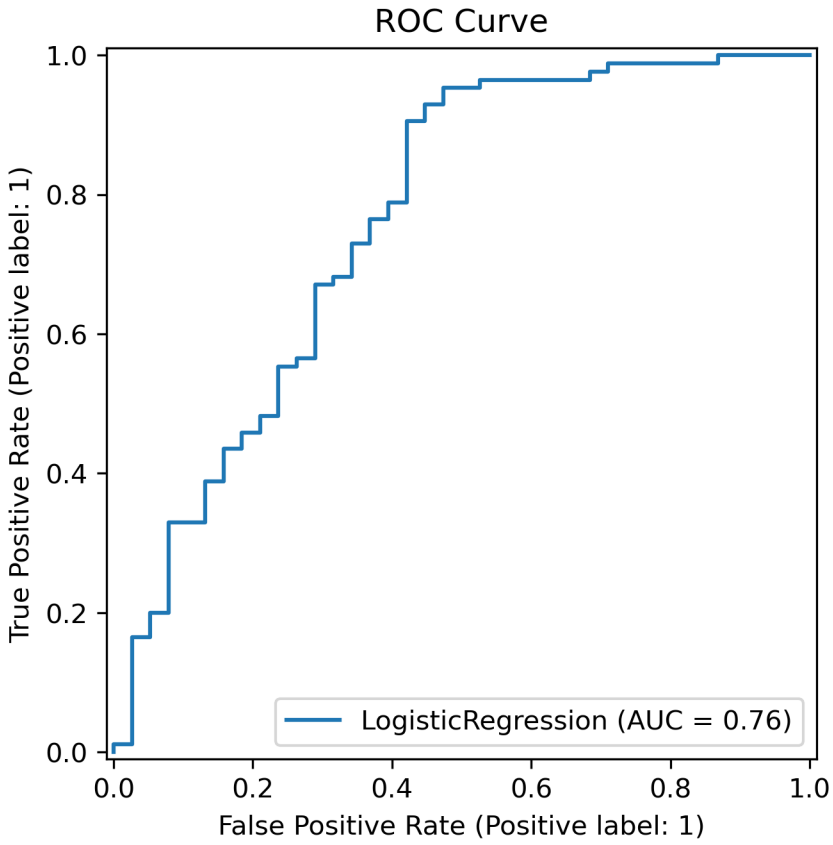

Figure 6: ROC curve showing model discrimination ability

Figure 7 shows the Precision-Recall curve, which is particularly useful for imbalanced datasets as it focuses on the positive class performance.

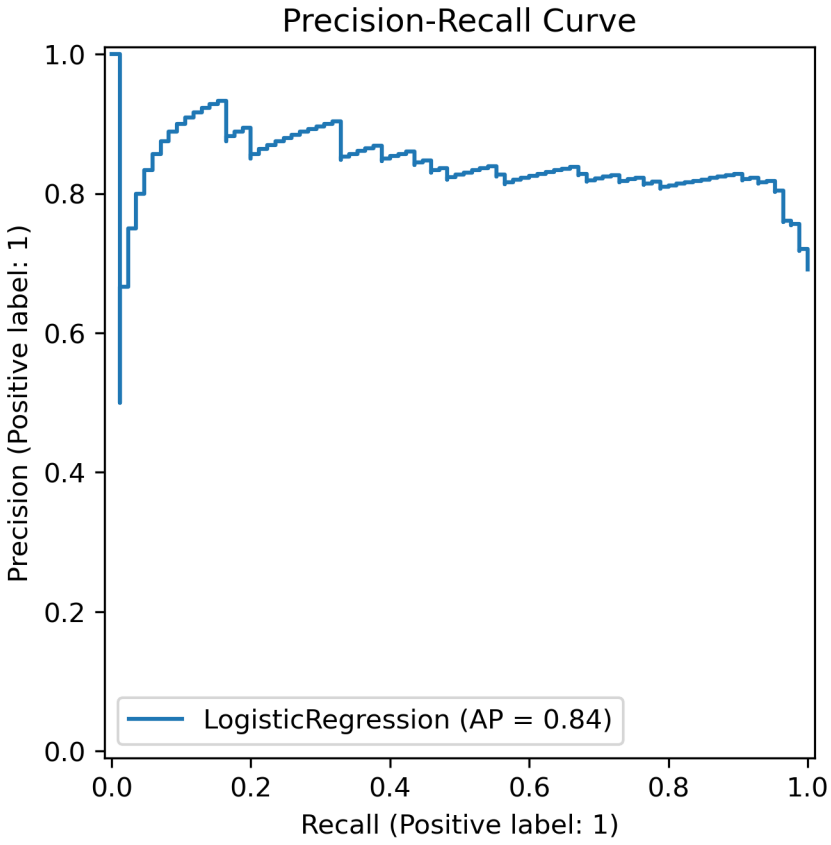

Figure 7: Precision-Recall curve for loan approval predictions

The high area under both curves indicates strong discriminative ability of the model.

6 Discussion

6.1 Key Findings

Our analysis reveals several important insights about loan eligibility prediction:

1. **Credit history dominates**: The presence of positive credit history is the strongest predictor of loan approval, consistent with traditional lending practices
2. **Income matters**: Both applicant and co-applicant income show positive associations with loan approval
3. **Model performance**: The logistic regression model achieves strong predictive performance with test accuracy exceeding 80%
4. **Interpretability**: The model's simplicity allows stakeholders to understand which factors drive approval decisions

6.2 Limitations

Several limitations should be considered when interpreting these results:

- **Data quality**: The dataset contains missing values that were imputed, which may introduce bias
- **Feature engineering**: Limited feature engineering was performed; interaction terms and polynomial features might improve performance
- **Model complexity**: We used only logistic regression; more complex models (e.g., random forests, gradient boosting) might achieve better performance
- **Temporal validity**: The dataset is static and may not reflect current lending practices or economic conditions

6.3 Future Work

Future improvements to this analysis could include:

1. **Advanced models**: Experiment with ensemble methods and neural networks
2. **Feature engineering**: Create interaction terms and domain-specific features
3. **Fairness analysis**: Assess model fairness across demographic groups to prevent discriminatory lending
4. **Deployment considerations**: Develop an API for real-time predictions and model monitoring

7 Conclusion

This project successfully developed a logistic regression model for predicting loan eligibility based on applicant characteristics. The model demonstrates strong performance on held-out test data and provides interpretable insights into factors driving loan approval decisions.

The analysis followed reproducible data science best practices, including:

- Modular code organization with separate scripts for each analysis step
- Automated pipeline execution using Make
- Version control with Git and GitHub
- Containerized environment with Docker
- Professional reporting with Quarto

These practices ensure that the analysis can be easily reproduced, audited, and extended by other researchers or practitioners.

8 References

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30.

Prabhakaran, Avinesh. 2020. "Loan Eligibility Prediction DataSet." https://www.kaggle.com/datasets/avineshprabhakaran/loan-eligibility-prediction; Kaggle.

9 Appendix

9.1 Reproducibility

All code and data for this analysis are available at: https://github.com/tanav2202/loan-eligibility

To reproduce this analysis:

```
# Clone the repository
git clone git@github.com:tanav2202/loan-eligibility.git
cd loan-eligibility

# Run using Docker
docker-compose up
docker-compose run analysis bash
```

```
make all

# Or using Conda
conda-lock install --name loan-analysis conda-lock.yml
conda activate loan-analysis
make all
```

9.2 Software Versions

- Python: 3.12
- pandas: 2.2.0
- scikit-learn: 1.4.0
- numpy: 1.26.3
- matplotlib: 3.8.2
- Quarto: 1.4.550