Loan Approval Prediction Using Logistic Regression

Muhammad Ghulam Ali

June 23, 2025

Contents

1 Introduction

Loan approval is a vital task in banking and finance. With the rise of machine learning, banks can improve their decision-making by using historical data to predict whether a loan should be approved. This project aims to build a binary classification model using logistic regression to automate the prediction of loan approvals.

2 Objective

The main objective is to create a supervised learning model that classifies loan applications as **Approved** or **Not Approved**, based on applicant and loan-related features.

3 Dataset Overview

- **Dataset Name**: `loan_approval_dataset.csv`
- **Features**:
  - Gender, Marital Status, Dependents
  - Education, Self_Employed
  - ApplicantIncome, CoapplicantIncome
  - LoanAmount, Loan_Amount_Term
  - Credit History, Property Area
- **Target Variable**: `Loan_Status` (Approved / Not Approved)

4 Tools and Libraries

This project uses the following Python libraries:

- `pandas`, `numpy` – data analysis
- `matplotlib`, `seaborn` – visualization
- `scikit-learn` – machine learning

5 Data Preprocessing

5.1 Handling Missing Values

Missing values were treated using statistical imputation (mean or mode based on context).

5.2 Encoding

- Binary categories were label-encoded.
- Nominal variables were one-hot encoded.

5.3 Feature Scaling

Numerical columns were standardized using `StandardScaler`.

5.4 Train-Test Split

The dataset was split into:

- **Training set**: 80%
- **Test set**: 20%

6 Model Development

6.1 Algorithm

Logistic Regression was selected due to its efficiency in binary classification problems.

6.2 Training

The model was trained using the following code:

```
from sklearn.linear_model import LogisticRegression
model = LogisticRegression()
model.fit(X_train, y_train)
```

Listing 1: Training the Logistic Regression Model

7 Model Evaluation

7.1 Metrics

- Accuracy Score
- Confusion Matrix
- Classification Report (Precision, Recall, F1-score)
- ROC AUC Score

7.2 Evaluation Code

```
from sklearn.metrics import classification_report, confusion_matrix
y_pred = model.predict(X_test)
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

Listing 2: Evaluating Model Performance

8 Results

The logistic regression model achieved high accuracy. The ROC curve demonstrated the model's ability to differentiate between approved and non-approved loan applications effectively.

9 Conclusion

Logistic regression is an effective approach for this loan classification problem. It provides a good baseline and interpretable results, making it suitable for decision-making in financial systems.

10 Future Work

- Experiment with advanced models like Random Forest, XGBoost
- Apply hyperparameter tuning using GridSearchCV
- Create a web-based interface using Flask/Django for deployment

11 Appendix

```
from sklearn.metrics import roc_curve, auc
fpr, tpr, thresholds = roc_curve(y_test, model.predict_proba(X_test)
    [:,1])
roc_auc = auc(fpr, tpr)
```

Listing 3: Plotting ROC Curve