Topic: Loan Approval Prediction using Logistic Regression

Name: ANAND KUMAR
Roll No.: 2201AI05
Course: CS502- Advanced Pattern Recognition

Abstract

This project focuses on predicting loan approval status using a logistic regression model. The model was trained and evaluated on a synthetic dataset containing 45,000 records and 14 variables, inspired by credit risk data. After preprocessing, which included handling missing values, encoding categorical variables, and feature scaling, the logistic regression model was trained. The model achieved a **test accuracy of 89.01%**, demonstrating its effectiveness in distinguishing between approved and rejected loan applications. The performance was further analyzed using a classification report and a confusion matrix.

1. Introduction

The objective of this project is to build a predictive model to determine whether a loan application will be approved or rejected. Using a synthetic dataset enriched with financial risk variables, a logistic regression classifier was developed. The model's performance was measured using standard classification metrics, including accuracy, precision, recall, and F1-score, to assess its real-world applicability.

2. Dataset

The dataset is a synthetic version inspired by the original Credit Risk dataset, containing **45,000 records and 14 variables**. It includes a mix of categorical and continuous features such as applicant age, income, employment experience, loan amount, and credit history. The target variable loan_status is binary, where `1` indicates an approved loan and `0` indicates a rejected loan.

3. Methodology

The project methodology involved exploratory data analysis (EDA), data preprocessing, model training, and evaluation.

3.1 Exploratory Data Analysis

Initial analysis was performed to understand the data distribution and relationships between variables. This included checking for missing values (none were found) and visualizing the

distribution of the target variable loan_status as well as its relationship with features like gender, education, and income.

3.2 Data Preprocessing

The data was prepared for modelling through the following steps:

- **Categorical Encoding:** Object-type categorical features were converted into numerical format using `Label Encoder`.
- **Feature Scaling:** All numerical features were standardized using Standared scaling to ensure they have a mean of 0 and a standard deviation of 1, which is important for logistic regression.

3.3 Model Training

A **Logistic Regression** classifier from the Scikit-learn library was used for the classification task. The dataset was split into training (80%) and testing (20%) sets. The model was trained on the pre-processed training data with a max_iteration of 1000.

4. Results

The model's performance was evaluated on the unseen test set.

4.1 Performance Metrics

The model achieved an overall **accuracy of 89.01%** on the test data. The detailed classification report below provides further insight into its performance for each class:

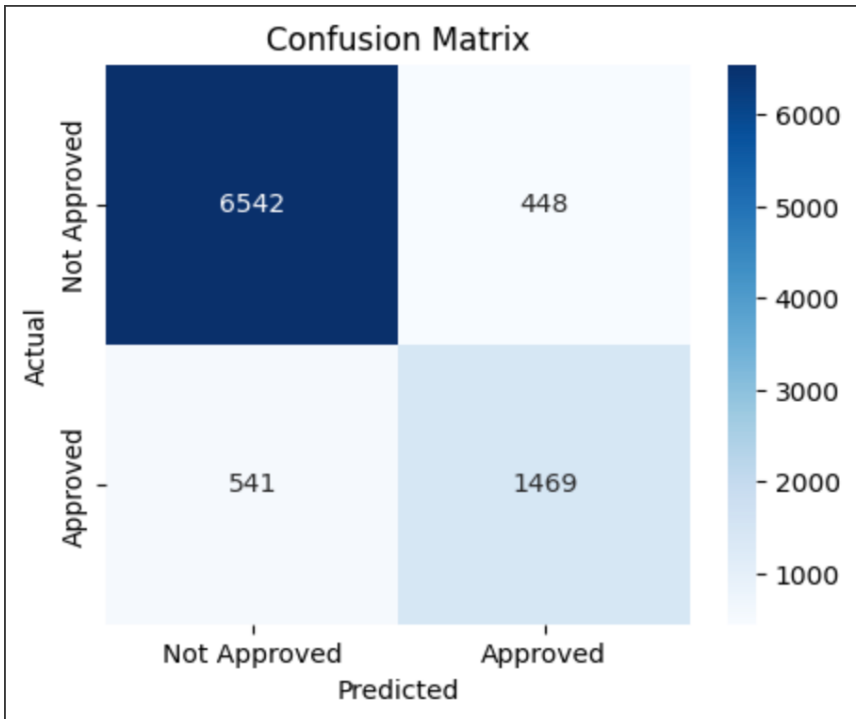
```
Accuracy: 0.8901111111111111

Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.94      0.93      6990
           1       0.77      0.73      0.75      2010

    accuracy                           0.89      9000
   macro avg       0.84      0.83      0.84      9000
weighted avg       0.89      0.89      0.89      9000
```

Confusion matrix showing true vs. predicted labels for loan approval.



4.2 Qualitative Analysis

Below are sample predictions from the test set, which illustrate the model's performance on individual applicant profiles:

- **Applicant Profile 1:** An applicant with a low loan-to-income ratio (0.2) and a high credit score (702) applied for a personal loan.
    - **Actual Status:** Not Approved
    - **Predicted Status:** Not Approved
- **Applicant Profile 2:** An applicant with a very low loan-to-income ratio (0.02) and a good credit score (670) applied for an education loan but had a history of previous defaults.
    - **Actual Status:** Not Approved
    - **Predicted Status:** Not Approved
- **Applicant Profile 3:** An applicant with zero employment experience and a moderate credit score (603) applied for a debt consolidation loan.
    - **Actual Status:** Approved
    - **Predicted Status:** Approved

5. Conclusion

The logistic regression model demonstrated strong predictive capability, achieving an accuracy of **89.01%** on the test set. The model showed high precision and recall for predicting non-approvals (class 0) but was less effective for approvals (class 1), as indicated by the lower F1-score of 0.75. This suggests that while the model is reliable for identifying low-risk applicants, it might misclassify some creditworthy candidates. Future work could involve exploring more complex models or using techniques to address the class imbalance to improve prediction accuracy for approved loans.