Loan Status Classification Report

Submitted by: Pragya Gupta

PGDM-Research & Business Analytics

Abstract

This project focuses on predicting the loan status of applicants using Logistic Regression and Decision Trees. Since the dataset was imbalanced, SMOTE (Synthetic Minority Oversampling Technique) was applied to balance the classes. The primary goal is to identify significant factors influencing loan approval decisions and evaluate the performance of classification models.

1. Introduction

Loan status prediction plays a crucial role in financial institutions as it helps reduce risks in lending. Applicants are evaluated on demographic, financial, and credit-related factors. The project applies machine learning classification techniques to build predictive models and improve decision-making accuracy.

2. Dataset Description

The dataset includes details such as applicant income, employment status, education, credit history, loan amount, and loan status. The target variable is the loan status (approved/not approved). The dataset was imbalanced, with fewer approved or non-approved cases, which required balancing using SMOTE.

3. Methodology

The following steps were taken to build the models:

- Data Preprocessing: Handled missing values, encoded categorical variables, and split data into train-test sets.
- Balancing Data: Applied SMOTE to address imbalance in target classes.
- Modeling: Trained Logistic Regression and Decision Tree models.
- Evaluation: Compared models using accuracy, precision, recall, F1-score, and ROC-AUC.

4. Results & Evaluation

The Logistic Regression model provided interpretability by highlighting significant features such as credit history and applicant income. Decision Trees captured non-linear relationships but required tuning to prevent overfitting. Evaluation metrics showed trade-offs between accuracy and recall across models.

5. Conclusion & Insights

The project demonstrated that both Logistic Regression and Decision Trees can be effective for loan status classification. Logistic Regression was better for interpretability, while Decision Trees were stronger at capturing complex patterns. Future work could include ensemble methods such as Random Forests or Gradient Boosting, and additional feature engineering to improve performance.