Project Report: Loan Approval Prediction Analysis

Prepared By

Name: Muhammed Shafi EK
Institution: Learnlogic AI Manjeri
Phone: 8075224593

Executive Summary

This project focuses on building and evaluating machine learning models to predict loan approval status based on customer and loan-related features. The dataset provided included demographic, financial, and loan-related attributes. Several machine learning algorithms were evaluated, and their performance was compared to identify the most accurate and efficient model.

1. Introduction

1.1 Problem Statement

Loan approval prediction is critical for financial institutions to minimize risks and optimize decision-making. This project aims to build a predictive model that can efficiently classify loan applications as approved or denied based on historical data.

1.2 Objectives

- Perform exploratory data analysis (EDA) and preprocessing.
- Engineer features for better prediction accuracy.
- Compare the performance of multiple machine learning algorithms.
- Identify the best-performing model for deployment.

2. Data Overview

2.1 Dataset Description

The dataset contains features such as:

- **Demographic Information**: Age, income, gender, education level.
- **Loan Details**: Loan amount, loan purpose, interest rates.
- **Credit History**: Credit score, past loan defaults.

2.2 Preprocessing Steps

- Handling missing values and duplicates.
- Encoding categorical variables using ordinal and label encoding.
- Balancing the dataset using SMOTE to address class imbalance.
- Removing multicollinearity by dropping correlated features.
- Standardizing numerical features for better model performance.

---

3. Exploratory Data Analysis (EDA)

3.1 Data Visualization

1. **Class Distribution**: Visualized using count plots to confirm class imbalance.
2. **Feature Correlation**: Heatmap analysis identified highly correlated features for removal.
3. **Box Plots**: Detected and addressed outliers using Interquartile Range (IQR).
4. **Log Transformation**: Reduced skewness in features like income and loan amounts.

---

4. Methodology

4.1 Models Evaluated

1. **Logistic Regression**
2. **Random Forest Classifier**
3. **XGBoost Classifier**
4. **Gradient Boosting Classifier**
5. **K-Nearest Neighbors (KNN)**

4.2 Evaluation Metrics

- Accuracy
- Confusion Matrix
- Precision, Recall, and F1-Score (for class imbalance insights)

---

5. Results and Discussion

5.1 Accuracy Comparison

Model	Accuracy (%)	Observation
XGBoost Classifier	93.00%	Highest accuracy achieved
Gradient Boosting Classifier	93.00%	Highest accuracy achieved
Random Forest Classifier	90.00%	Tied for third-highest accuracy
K-Nearest Neighbors (KNN)	90.00%	Tied for third-highest accuracy
Logistic Regression	88.00%	Lowest accuracy

5.2 Observations

- **XGBoost Classifier** and **Gradient Boosting Classifier** delivered the best performance with 93% accuracy.
- **Random Forest** and **KNN** performed well but were slightly less accurate.
- **Logistic Regression** had the lowest accuracy due to its linear nature, which may not capture complex relationships in the data.

---

6. Model Performance Visualization

1. **Confusion Matrices**: Visualized using heatmaps to understand true positives, false positives, false negatives, and true negatives.
2. **Accuracy Box Plot**:
   Displays the accuracy range of all models for comparison.

---

7. Conclusion and Recommendations

7.1 Key Findings

- Ensemble models (**XGBoost** and **Gradient Boosting**) were the most effective, leveraging their ability to handle complex relationships in the data.
- Balancing the dataset with SMOTE significantly improved model performance.
- Feature scaling and multicollinearity reduction were crucial preprocessing steps.

7.2 Recommendations

- Deploy the **XGBoost Classifier** model in production due to its robustness and scalability.
- Periodically retrain the model with new data to ensure continued accuracy.
- Explore hyperparameter tuning and advanced techniques like stacking to further enhance performance.

---

8. Future Work

- Incorporate additional features such as customer behavioral data for improved predictions.
- Implement explainable AI techniques to interpret model predictions.
- Develop a user-friendly dashboard for stakeholders to visualize predictions.

Colab link:https://colab.research.google.com/drive/1QS8S8n6TrBQu2i1ctTwDuBtgTDpiM7Yh?usp=sharing