Report

The purpose of this analysis was to build and evaluate several machine learning models to predict loan status (approved or not approved) based on financial information provided in the dataset.

The dataset contained various borrower attributes such as:

- loan_size
- interest_rate
- borrower_income
- debt_to_income
- num_of_accounts
- derogatory_marks
- total_debt

The target variable was loan_status, where:

- 0 = Loan not approved
- 1 = Loan approved

There was a severe class imbalance:

- 0: 75,036 instances
- 1: 2,500 instances

This imbalance made recall for class 1 a priority, since we care more about identifying applicants who should be approved.

Machine Learning Process

1. Data Cleaning: Verified no nulls or categorical encoding needed.
2. Exploratory Data Analysis (EDA): Identified strong correlations and multicollinearity between features.
3. Scaling: Used StandardScaler for all numerical features.

4. Train-Test Split: 75% training and 25% testing with stratified sampling to maintain class distribution.

5. Model Training & Evaluation: Evaluated multiple models: Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, XGBoost, LightGBM.

Key metrics included accuracy, precision, recall, and ROC AUC, with emphasis on recall for class 1.

Results

Machine Learning Model 1: Logistic Regression

- Accuracy: 99%
- Precision (class 1): 0.88
- Recall (class 1): 0.98
- AUC: 0.997

Machine Learning Model 2: K-Nearest Neighbors (K=7)

- Accuracy: 100%
- Precision (class 1): 0.88
- Recall (class 1): 0.99
- AUC: 0.997

Machine Learning Model 3: Random Forest

- Accuracy: 99%
- Precision (class 1): 0.88
- Recall (class 1): 0.88
- AUC: 0.997
- Note: Slight overfitting observed

Machine Learning Model 4: XGBoost

- Accuracy: 100%

- Precision (class 1): 0.88
- Recall (class 1): 0.99
- AUC: 0.998
- Note: Very strong results but feature importance is heavily skewed toward loan_size

Machine Learning Model 5: LightGBM

- Accuracy: 100%
- Precision (class 1): 0.88
- Recall (class 1): 0.99
- AUC: 0.998
- Note: Similar to XGBoost but with more balanced feature importance

Summary

All models performed very well in terms of accuracy and recall. However, given the imbalanced dataset and business objective (identifying who should be approved), recall for class 1 was the most important metric.

- Best Model Overall: LightGBM
    - Strong recall (0.99), high AUC (0.998), and better feature diversity than XGBoost
    - Generalizes well without overfitting
- Best for Explainability: Logistic Regression
    - Slightly lower performance but offers high interpretability and regularization to handle multicollinearity

Recommendation:

Use LightGBM for production if maximizing performance is the goal. Use Logistic Regression in contexts where explainability and transparency are required (e.g., financial compliance).