BANK LOAN APPROVAL PREDICTION USING MACHINE LEARNING

PROJECT REPORT

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING - AI

NAME :- ROHAN KARURI

STUDENT CODE :- BWU/BTA/23/139

ROLL NO. :- 23010332130

REGISTRATION NUMBER :- 23013001743 of 2023-2024

COURSE NAME :-ARTIFICIAL INTELLIGENCE FOR REAL WORLD APPLICTION

COURSE ID :- PCC-CSM503

DATASET : https://www.kaggle.com/code/melissamonfared/bank-personal-loan-eda-classification

Team Members

Arpita Naskar

Rohan Karuri

Manas Kumar Ghosh

Abhra Purkait

Pritam Naskar

1. Abstract

The banking industry plays a crucial role in the economic development of a country. One of its primary functions is providing loans to individuals and organizations. However, loan approval is a sensitive and risky process, as incorrect decisions may lead to financial losses due to loan defaults. Traditionally, loan approval decisions are made manually by bank officials after reviewing applicant documents and financial history. This process is time-consuming, labor-intensive, and susceptible to human bias.

This project presents a Machine Learning–based Bank Loan Approval Prediction System that automates the decision-making process by analyzing historical loan data. Using supervised machine learning classification techniques, the system predicts whether a loan application should be approved or rejected. Algorithms such as Logistic Regression and Random Forest are used to train and evaluate the model.

Key Highlights of the Project

Automates loan approval decisions using data-driven intelligence

Reduces human bias and improves consistency in decision-making

Enhances accuracy and efficiency compared to manual methods

Supports banks in minimizing loan default risks

Provides a scalable and cost-effective solution

The proposed system improves efficiency, accuracy, and fairness in loan approval decisions, thereby helping banks reduce operational costs and financial risk while maintaining transparency and reliability.

2.Introduction

In recent years, the banking and financial sector has witnessed rapid digital transformation. With the availability of large volumes of customer data, banks are increasingly adopting data-driven approaches to improve their services. One of the most important applications of data analytics and machine learning in banking is credit risk assessment and loan approval automation.

Loan approval is a complex process that depends on multiple factors such as income, employment status, credit history, loan amount, and repayment capability. Manual evaluation of these factors is inefficient, especially when the number of loan applications is large. Moreover, human judgment may introduce inconsistencies, delays, and bias, which can negatively impact customer satisfaction and bank profitability.

Machine Learning (ML) provides an effective solution by learning patterns from historical data and making accurate predictions on new data. ML models can process large datasets, identify hidden relationships between variables, and deliver reliable outcomes in a short time.

Importance of Machine Learning in Loan Approval

Enables fast and automated decision-making

Handles large-scale data efficiently

Identifies complex patterns beyond human capability

Improves customer experience through quicker responses

Ensures objective and unbiased decisions

This project focuses on developing an ML-based loan approval predction system that assists banks in making faster, more accurate, and consistent loan sanction decisions. The system acts as a decision-support tool for banking professionals rather than replacing human judgment entirely.

3. Problem Definition

The objective of this project is to design and develop an automated decision-support system that predicts whether a loan application should be sanctioned or rejected. Loan approval is framed as a binary classification problem, where the output variable takes the value 1 (sanctioned) or 0 (not sanctioned).

The system must meet multiple real-world constraints such as high predictive accuracy, fairness across applicant groups, explainability of decisions, and low inference latency. These requirements are essential for deployment in large financial institutions like Tata Capital, where decisions must be fast, auditable, and compliant with regulatory standards.

4. Project Objectives

The primary objectives of this project are multifold. First, it aims to maximize predictive performance by using robust and scalable machine learning techniques. Second, the system is designed to provide explainable and auditable decisions, ensuring transparency for loan agents and regulatory bodies. Third, it reduces manual workload by assisting loan officers with data-driven recommendations. Finally, the solution is built to integrate seamlessly with existing loan processing workflows, making it practical for real-world adoption.

5. Motivation and Business Impact

Manual loan approval processes involve document verification, subjective judgment, and multiple approval layers, making them slow and inconsistent. Human bias and fatigue can further affect decision quality.

By introducing an AI-based loan approval system, banks can significantly reduce processing time, increase consistency, lower operational costs, and improve customer satisfaction. The model also enables better risk management, helping financial institutions minimize non-performing assets while maintaining regulatory compliance.

6. Dataset Overview

The dataset used in this project consists of approximately 50,000 historical loan records sourced from Tata Capital. Each record represents a loan application and includes applicant demographics, financial indicators, credit history, and collateral information.

The dataset is structured and labeled, making it well-suited for supervised machine learning. Its size and diversity allow the model to learn complex patterns and generalize well to unseen applications.

7. Data Sources and Collection

Data was collected from multiple internal systems. The primary sources include loan application platforms and credit bureau feeds, which provide real-time applicant and credit information. Secondary sources include derived behavioral metrics and historical repayment performance. Strict data governance measures were enforced, such as secure data transfer, anonymization, access control, and minimization of personally identifiable information (PII), ensuring data privacy and regulatory compliance.

8. Key Features Used

The model uses a diverse set of features capturing multiple aspects of loan risk. These include demographic variables such as age and employment type, financial indicators such as income, loan amount, and interest rate, credit metrics like credit score and previous defaults, and collateral-related features such as collateral value and loan-to-value ratio.

The target variable is the loan sanction status, represented as a binary outcome.

9. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to understand the statistical properties of the dataset. Distributions of income and loan amounts were found to be skewed, while variables such as credit score and debt-related ratios showed strong correlation with loan approval outcomes.

EDA also helped identify missing values, outliers, and class distribution, enabling informed preprocessing and feature engineering decisions.

10. Feature Engineering

Feature engineering played a crucial role in improving model performance. Domain-specific features such as debt-to-income ratio, credit utilization, and employment duration buckets were derived to better capture repayment capacity.

Log transformation was applied to income to reduce skewness, and historical delinquency metrics were aggregated to represent applicant behavior over time.

11. Feature Selection

To avoid overfitting and improve interpretability, feature selection was performed using correlation analysis, mutual information scores, and model-based importance metrics from XGBoost.

Special care was taken to prevent data leakage by excluding features that would not be available at the time of loan decision.

12. Data Preprocessing

Data preprocessing ensured consistency and quality of input data. Missing numerical values were imputed using the median to reduce sensitivity to outliers, while categorical values were imputed using the mode.

Outliers were handled through winsorization. One-hot encoding was used for nominal categorical variables, and ordinal encoding was applied where an inherent order existed.

13. Handling Class Imbalance

Although the dataset was relatively balanced, minor class imbalance was addressed using techniques such as SMOTE, class weighting, and stratified sampling.

Stratified cross-validation ensured that both training and validation sets preserved the original class distribution, preventing biased evaluation.

14. Model Selection and Justification

The XGBoost Classifier was selected due to its superior performance on structured tabular data. It effectively captures nonlinear relationships and feature interactions while offering built-in regularization to prevent overfitting.

Additionally, XGBoost provides fast inference, making it suitable for real-time decision systems, and supports interpretability through feature importance and SHAP explanations.

15. Model Architecture and Hyperparameters

The model architecture consists of an ensemble of gradient-boosted decision trees. Key hyperparameters such as learning rate, maximum tree depth, number of estimators, subsampling ratios, and regularization parameters were carefully tuned.

Early stopping was employed to halt training when validation performance stopped improving, ensuring optimal generalization.

16. Training Strategy and Validation

The dataset was divided into 80% training and 20% testing sets. To ensure robustness, Stratified K-Fold Cross-Validation (k = 5) was applied during training.

The best-performing model was selected based on validation ROC-AUC and evaluated on the unseen test set to assess real-world performance.

17. Evaluation Metrics

Multiple evaluation metrics were used to assess model performance comprehensively. Accuracy measured overall correctness, while precision and recall evaluated error trade-offs. The F1-score provided a balance between precision and recall. ROC-AUC measured ranking capability, and confusion matrices helped analyze misclassifications.

Cohort-wise metrics ensured fair performance across applicant segments.

18. Model Performance Results

The final XGBoost model achieved strong results, with approximately 91% accuracy, 89% precision, 93% recall, 91% F1-score, and 0.96 ROC-AUC.

These results demonstrate the model's high predictive capability and strong generalization to unseen data.

19. Error and Result Analysis

Detailed error analysis was conducted to study false positives and false negatives. The model's behavior was examined across income levels, employment types, and credit score bands.

This analysis helped identify areas where additional safeguards or human review may be necessary.

20. Bias, Fairness, and Ethical Considerations

Fairness audits were conducted using demographic parity and equal opportunity metrics. Bias mitigation strategies included threshold adjustments, reweighting techniques, and human-in-the-loop overrides.

All decisions are logged and traceable, ensuring ethical use of AI and compliance with financial regulations.

21. Deployment, Monitoring, and Future Enhancements

The model can be deployed as a REST API using FastAPI and Docker. Continuous monitoring tracks data drift, prediction latency, and performance degradation. Periodic retraining ensures model freshness.

Future enhancements include A/B testing, production-grade feature stores, advanced explainability interfaces, and privacy-preserving machine learning techniques.

22.Conclusion

This project successfully demonstrates the application of Machine Learning techniques in automating and enhancing the bank loan approval process. By leveraging historical loan data and advanced classification algorithms, the proposed system is able to accurately predict whether a loan application should be sanctioned or rejected. The problem was effectively formulated as a binary classification task, and a robust solution was developed to meet real-world banking requirements such as accuracy, fairness, explainability, and low decision latency.

The use of the XGBoost classifier proved to be highly effective for structured financial data, achieving strong performance across multiple evaluation metrics including accuracy, precision, recall, F1-score, and ROC-AUC. Feature engineering and careful preprocessing played a crucial role in improving model performance, while cross-validation and proper evaluation strategies ensured good generalization on unseen data. The incorporation of explainability techniques such as SHAP enhanced transparency, making the model suitable for regulatory and auditing purposes.

Furthermore, the system demonstrates significant business value by reducing manual effort, minimizing human bias, improving decision consistency, and increasing operational efficiency. Ethical considerations and fairness checks were integrated into the model development process to ensure responsible AI usage. The proposed deployment architecture allows seamless integration into existing loan processing workflows with provisions for monitoring and periodic retraining.

In conclusion, the AI-based loan approval prediction system provides a scalable, reliable, and efficient decision-support solution for modern financial institutions. With further enhancements such as real-time integration, advanced fairness controls, and continuous learning mechanisms, the system has strong potential for real-world deployment and long-term impact in the banking sector.