DEPARTMENT OF

COMPUTER SCIENCE AND ENGINEERING

(CSHO634DAP)

Data Science at Production Scale

Automating Loan Approval Decision using Traditional Machine Learning Techniques

B. Tech – Computer Science and Engineering (AIML)

School of Engineering and Technology,

CHRIST (Deemed to be University),

Kumbalagodu, Bengaluru-560074

March -2025

1. INTRODUCTION

The loan approval process is a pivotal activity in the banking sector, establishing the qualification of individuals or enterprises to obtain credit based on their history, income, and other variables. Historically, this has been a labor-intensive, time-consuming, and error-prone manual process that resulted in inefficiencies and bias. With machine learning, there is potential to automate and improve this process so that decisions are faster, more accurate, and equitable. Automating loan approval decisions not only makes financial institutions more operationally efficient but also improves customer satisfaction through quicker turnaround times and clear results. This research investigates the use of conventional machine learning methods to automate loan approval decisions to address the increasing demand for scalable and reliable solutions in the financial sector.

The emphasis of this research is to utilize classical machine learning models, including Logistic Regression, Decision Trees, Random Forests, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN), to forecast loan approval results based on past data. Through feature analysis of applicant age, income, credit score, loan amount, and interest rate, these models are designed to categorize applicants into approved or rejected groups. The research stresses the need for data preprocessing, such as missing value handling, outliers, and categorical encoding, to make the models more robust. Furthermore, the research compares the performance of these algorithms in terms of accuracy and interpretability, with implications for their applicability in real-world applications.

In spite of the progress made in machine learning, there remain a number of challenges in automating loan approval processes. One key challenge is the imbalance of loan datasets, where the number of approved loans tends to be much larger than the number of rejected loans, resulting in biased models. Outliers and missing values in financial datasets can also negatively impact model performance. Even though complex methods such as deep learning have been found to be useful, they are often not explainable, with financial institutions facing challenges in defending decisions to regulators and customers. Previous studies have focused mainly on standalone algorithms, without filling the need for end-to-end comparisons of various traditional machine learning methods used in loan approval automation.

This work overcomes these challenges by following a systematic method of data preprocessing, such as outlier removal, normalization, and label encoding, to provide high-quality input data. Contrary to the existing literature focusing on one algorithm, this work presents a comparative study of several conventional machine learning models, indicating their strengths and limitations

in loan approval. Through the evaluation of performance metrics such as accuracy, precision, recall, and F1-score, the work presents an overall picture of model performance. In addition, the application of interpretable models such as Decision Trees and Logistic Regression provides transparency, which is essential for regulatory purposes and customer trust.

The remainder of this paper is organized as follows: Section 2 discusses the dataset and the preprocessing steps undertaken to prepare the data for modeling. Section 3 provides an overview of the machine learning algorithms used in the study and their theoretical foundations. Section 4 presents the experimental setup, including the evaluation metrics and results. Section 5 discusses the findings, comparing the performance of the models and their implications for real-world applications. Finally, Section 6 concludes the paper with a summary of the key insights and suggestions for future research directions in automating loan approval decisions using machine learning.

2. Literature Review

Credit risk modeling and loan default prediction have been extensively studied in financial research. The following studies represent significant contributions to this field:

1. Loan Default Prediction with Machine Learning Techniques:

This research was done for the 2020 International Conference on Computer Communication and Network Security (CCNS) concerning the loan default prediction using machine learning algorithms. The study was aimed at analyzing several machine learning models such as decision tree, support vector machine (SVM), and artificial neural network (ANN) to assess their impact on the evaluation of credit risk. Results showed that ensemble learning methods outperformed the rest of the models in terms of accuracy. This study also demonstrates that choosing appropriate features has a great impact on model effectiveness and trustworthiness.

2. Loan Default Prediction Using Spark Machine Learning Algorithms

Carried out by Aiman Muhammad Uwais and Hamidreza Khaleghzadeh at the University of Portsmouth, this research explores the scalability and performance of Spark-based machine learning algorithms on loan default predictions. It addresses the problem of imbalanced datasets in credit scoring by proposing several resampling methods to reduce bias in model training. This research also shows that, through the use of distributed processing, Spark MLlib improves the speed and the range of processing with the increasing amount of the data, which is ideal for large financial datasets.

3. Research on loan default prediction based on logistic regression, randomforest, XGBoost and ADABoost

Jinchen Lin of Guangdong University of Technology analyzes and compares multiple machine learning techniques, including logistic regression, random forest, XGBoost, and AdaBoost, in forecasting loan defaults. The research notes the merits and drawbacks of each model, maintaining that tree-based ensemble techniques outperform the others in managing non-linear and intricate interactions among financial variables. The research also offers insights on hyperparameter tuning that enhances model accuracy.

4. Ensemble-Based Machine Learning Algorithm for Loan Default Risk Prediction

This was a joint study carried out by scholars from Sheffield Hallam University, Edge Hall University, and the University of Brighton. The study suggests an ensemble approach for predicting the risk associated with loan defaults. It aims at creating a more robust prediction model by combining multiple classifiers into a single complex hybrid model. The study assesses how stacking and boosting could reduce the prediction error rate. The findings reveal that ensemble models accomplish more complex tasks than traditional statistical models when working with financial data; these methods actually use the complex interdependencies within the financial data. The study elaborates on the application of ensemble strategies in real finances to improve their practicality.

5. Attention-based Dynamic Multilayer Graph Neural Networks for Loan Default Prediction

Sahab Zandi, Kamesh Korangi, Maria Oskarsdottir, Christophe Mues, and Cristian Bravo proposed a new method of loan default prediction using deep learning and graph neural networks (GNNs) in this study. This study applies attention mechanisms in order to construct financial networks of borrowers and their relations. The model improves predictive accuracy by using multilayered graphs to capture complex relationships within credit information. The authors claim that GNNs can be better utilized in financial risk modeling, presenting a refreshing approach towards the analysis of complex financial data.

3. Methodology:

1. Logistic Regression

Logistic regression is a widely used statistical algorithm for binary classification problems. Unlike linear regression, which predicts continuous values, logistic regression estimates the probability of a class using the sigmoid function. The equation is as follows:

$$P(Y=1)=1/(1+e^{-(\beta 0+\beta 1X1+\beta 2X2+...+\beta nXn)})$$

Similar to predicting loan defaulting, logistic regression is frequently employed to estimate whether an applicant will pay back the loan (1) or will default (0) based on certain criteria detailing the applicant's income, credit score, and debt-to-income ratio. Whereas, With its straightforward calculation process and ease of understanding, the relative simplicity of logistic regression also has a largely unsolvable weakness: the assumption of linear separability (however, linear decision boundaries are common most of the time).

2. Support Vector Machine (SVM)

An SVM classification is an algorithm that classifies data using an optimal hyperplane. In separating boundaries, it seeks to maximize the gap between data points of different classes so that there is greater generalization.

SVM can also be employed in loan prediction models to classify applicants into high and low risk based on their credit score, outstanding debts, and repayment history. This technique is advantageous in cases with high dimensionality, but it can be costly in terms of computing time when working with large datasets. SVMs apply kernel tricks to separate non-linear data by moving it to a higher dimensional space.

3. XGBoost

XGBoost, an advanced version of ensemble learning incorporates gradient boosting and is referred to as Extreme Gradient Boosting. It constructs several weak decision trees one by one while minimizing errors for every iteration in order to enhance performance.

XGBoost is commonly applied in financial domains, such as predicting loaning risks because of its effectiveness in dealing with imbalanced datasets and complex feature interactions. It provides excellent accuracy and efficiency at all levels while avoiding overfitting with the use of regularization techniques like L1 and L2 penalties. However, the XGBoost complexity makes hyperparameter tuning a difficult task.

4. K-Nearest Neighbors (KNN)

KNN is considered a very simple classification algorithm owing to the fact that a class is automatically assigned to data points via a majority vote from nearest neighbor points. KNN is helpful when the boundary is not uniform across the graph.

KNN does not have strong inhibitions on the data distribution, which makes it highly versatile. It does not, however, work in an efficient manner for larger data sets. KNN is highly inefficient due to its need to measure distance to every point.

5. Random Forest

The Random Forest is an ensemble learning method that consists of aggregating the outputs from numerous decision trees, each of which attempts to maximize the predictive accuracy and minimize overfitting. To enhance variation in decision-making, every tree is trained for a random subset of features and data.

To evaluate the creditworthiness of an applicant, Random Forest can analyze several financial factors including and not limited to the income's stability, prior repayment behaviors, and remaining debt. Random Forest is less sensitive to overfitting than individual decision trees and is more powerful as a result. However, when applied to large datasets, it can become very resource intensive.

Proposed Algorithm

The proposed algorithm follows a structured machine learning pipeline to predict loan approval decisions accurately. The steps involved are as follows:

1. Data Preprocessing

To ensure data quality and improve model performance, we apply the following preprocessing steps:

- **Handling Missing Values:** Missing values are identified and imputed using statistical methods (mean, median) or dropped if necessary.
- **Removing Duplicate Records:** Any redundant entries in the dataset are removed to prevent bias in model training.
- **Outlier Detection and Removal:** Outliers in numerical features (such as income, loan amount, and credit score) are detected using the **Interquartile Range (IQR)** method and handled appropriately.

- **Encoding Categorical Variables:** Categorical attributes (such as home ownership, education, and loan intent) are converted into numerical values using **Label Encoding**.
- **Feature Scaling:** Standardization using **StandardScaler** is applied to normalize feature distributions and improve model convergence.
- **Dimensionality Reduction: Principal Component Analysis (PCA)** is applied to reduce the dataset to **five principal components**, retaining essential variance while reducing complexity.

2. Model Selection and Training

We evaluate multiple supervised learning models for loan approval prediction:

- **Logistic Regression:** A baseline linear model suitable for binary classification.
- **Random Forest Classifier:** An ensemble-based model that improves decision boundaries by aggregating multiple decision trees.
- **XGBoost Classifier:** A gradient-boosting algorithm optimized for high accuracy and performance.
- **Support Vector Machine (SVM):** A model that finds the optimal hyperplane for classification.
- **K-Nearest Neighbors (KNN):** A distance-based algorithm for classification.

Each model is trained using an **80-20 train-test split**, and performance is compared based on accuracy, precision, recall, and F1-score.

3. Hyperparameter Optimization

To improve model performance, we apply **GridSearchCV** for hyperparameter tuning:

- **Logistic Regression:** Optimizing solver types, regularization strength (C), and max iterations.
- **XGBoost:** Tuning number of estimators, learning rate, max depth, and subsampling ratio.
- **Random Forest:** Finding the best number of estimators, depth, and minimum sample split.
- **SVM:** Tuning kernel types, regularization parameter (C), and gamma values.
- **KNN:** Selecting the optimal number of neighbors, distance metrics, and weighting strategies.

4. Model Evaluation and Interpretation

The trained models are evaluated using multiple metrics:

6. **Accuracy Score:** Measures overall correct predictions.
7. **Precision and Recall:** Important for minimizing false positives and false negatives.
8. **F1-Score:** Provides a balance between precision and recall.
9. **Confusion Matrix:** Visualizes classification errors and model reliability.
10. **Feature Importance Analysis:** Identifies the most influential features affecting loan approval.

11. Model Evaluation

As for assessing the trained models on loan approval predictions, We conducted an accuracy assessment using a multi-classification approach that included precision, recall, and F1 score as metrics. There were five machine learning models employed: Logistic Regression, XGBoost, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). After doing some hyperparameter searching to better the model, We cross-validated them to enable generalizability.

Performance Metrics

The models were evaluated based on the following key performance indicators:

- **Accuracy**:The ratio of true positives and negatives outcomes to total outcomes. .
- **Precision**: The ratio of true positives to the sum of true positives and false positives, indicating the reliability of positive classifications.
- **Recall**: The ratio of true positives to the sum of true positives and false negatives, representing the model's ability to identify all relevant instances.
- **F1-score**: The harmonic mean of precision and recall, balancing both metrics to provide an overall measure of model performance.

Table 1 summarizes the performance of each model on the test dataset:

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	88.97%	0.75	0.69	0.72
XG Boost	93.02%	0.86	0.78	0.82
Random Forest	92.25%	0.87	0.73	0.79
SVM	90.75%	0.83	0.69	0.75
KNN	89.78%	0.79	0.68	0.73

Results Analysis

XGBoost demonstrated the greatest predictive accuracy among all evaluated models, achieving an accurate rate of 93.02%. Additionally, its recall score for the positive class (approved loans) was the highest at 78%, meaning that it was able to identify true approvals with commendable precision at 86%. It was also outstandingly effective at distinguishing between positive and negative instances of approved loans. Its accuracy was closely followed by the Random Forest classifier at 92.25%. However, the Random Forest Classifier's recall was slightly lower than XGBoost.

The SVM and KNN classifiers did reasonably well, achieving accuracies of 90.75% and 89.78% respectively. Their significantly lower recall rates suggest a higher likelihood of misclassifying approved loans as rejections. Logistic Regression, as expected, served as the baseline model, but also displayed the lowest recall rate at 69%, suggesting that it struggles the most with accurately recognizing approved loans.

Implications and Future Improvements

The outcomes emphasize the effectiveness of the ensemble learning methods, especially XGBoost and Random Forest, in dealing with the intricacies of loan approval forecasting. Nonetheless, the recall for class 1 continues to be an issue for all models. Subsequent studies could focus on class balancing strategies using Synthetic Minority Over-sampling Technique (SMOTE), cost-sensitive methods, or additional hyperparameter tuning to enhance recall at the cost of some precision.

Also, feature impact assessment from SHAP (Shapley Additive Explanations) analyses would allow loan approvals to be better understood and interpreted, ensuring effective application and decision-making in practice.

In general, the assessment reveals the efficiency of applying machine learning approaches in forecasting loan approvals, where XGboost stands out as the best candidate for the intended use.

12. Conclusion and Future Work

The analysis of different classifiers for predicting loan approvals has shown the effectiveness of machine learning models in these tasks. Ensemble techniques XGBoost demonstrated the best results for accuracy and recall among all tested classifiers. The results suggest that ensemble based models achieved better results than traditional approaches based on classification, which makes them more convenient for use in financial decision making systems. Still, some challenges exist as models performed well, but recall of approved loans was not sufficient and would result in high rates of missed loans.

Future work can concentrate on improving the performance and explainability of the model on a few set objectives. First, more powerful models or techniques like deep learning and hybrid models could be employed to increase classification accuracy. Second, adding feature domains such as other financial aids can improve model generalization. Finally, provided guarantee for transparency and fairness in decision-making by financial institutions enable use of explainable methods like SHAP values.

Addressing these challenges helps to develop accurate and explanatory based AI systems for loan approval, which is a step for improving benefits for lenders and borrowers, enabling the optimization of futures work.