isrjournals

International Journal of Advanced Research in Computer Networking,Wireless and Mobile Communications

Volume: 5 Issue: 3 Apr,2025,ISSN_NO: 2320-7248

MACHINE LEARNINGFOR BANK LOAN ELIGIBILITY PREDICTION: FOCUS ON HOME LOAN AND EDUCTION LOAN

D Shanmuga Prudvi[1], M Santhosh[2], Moulishwar S[3], Ms Edith Esther E[4*]
[123]UG Scholar–Dept CSE, GRT Institute of Engineering and Technology Tiruttani, India.
[4*]Professor- Dept CSE, GRT Institute of Engineering and Technology Tiruttani, India.
dshanmugaprudvi@gmail.com, kdsanthosh333@gmail.com, rishumouli201@gmail.com
[4*]Corresponding Author: edithesther.e@grt.edu.in, https://orcid.org/0009-0007-0506-1848

Abstract- Presently a day's individual approach or select bank credits to fulfill their needs, which are exceptionally common. [1]. This hone has been expanding day by day. The loan is one of the most important schemes of bank [2]. Banks typically offer loans to customers in accordance with their needs. However, unfortunately, some clients are unable to pay their debts on time or delay doing so because of their financial situation.[3] Several people take advantage and misuse the facilities given by the bank. In order to solve this problem Banks must employ certain approaches to assist in anticipating the loan repayment status.[4] Banking system always need accurate modelling system for large number of issues. The ability to predict credit defaulters is one of the most challenging tasks for any bank [5] However,by predicting the loan defaulters, the banks will undoubtedly be able to cut their loss by decreasing their non-profit assets, allowing the recovery of sanctioned loans to proceed without incurring any losses and acting as a contributing factor to the bank statement.

1. INTRODUCTION

In the modern banking sector, accurately determining loan eligibility is crucial for minimizing financial risk and improving service efficiency. With the increasing demand for financial assistance in the form of home loans and education loans, traditional manual assessment methods are proving to be time-consuming, inconsistent, and prone to human bias. To address these challenges, Machine Learning (ML) offers a powerful, data-driven approach for predicting loan eligibility. By analyzing historical loan data and identifying patterns, ML models can make informed predictions based on various factors such as income, credit score, employment status, educational background, and property details. This not only speeds up the loan approval process but also improves the accuracy and fairness of decisions.

This project focuses on using machine learning techniques to predict the eligibility of applicants specifically for home loans and education loans. With the rise of digital transformation in the financial sector, banks are increasingly adopting intelligent systems to enhance their decision-making capabilities. Among these, machine learning stands out for its ability to learn from large volumes of past data and improve over time. Its applications are vast, ranging from fraud detection to customer segmentation—but one of the most impactful uses lies in loan eligibility prediction.

Home loans often involve substantial amounts and long repayment periods, making it essential for banks to assess the financial stability and creditworthiness of the applicant with high precision. Similarly, education loans are pivotal in supporting the academic aspirations of students, yet they pose a unique challenge as applicants may lack steady income or strong credit histories. This makes traditional assessment models less effective in such cases.

Machine learning models can take into account a broader range of features—such as the applicant's family income, co-applicant status, academic records, employment prospects, existing liabilities, and more. Unlike rigid rule-based systems, ML algorithms can adapt to changing data trends and provide predictions with improved accuracy over time.

2. RELATED WORK

In recent years, the application of machine learning in financial services—particularly in loan prediction and credit risk analysis—has gained significant attention from researchers and industry professionals alike. Numerous studies have been conducted to develop and evaluate models that can predict loan eligibility with high accuracy, using various supervised learning techniques and real-world datasets.

Several researchers have explored the use of classification algorithms such as Logistic Regression, Decision Trees, Random Forest, and Naive Bayes to predict loan approval. These models typically consider features such as applicant income, loan amount, credit history, and employment status. Studies have shown that ensemble methods like Random Forest often outperform individual classifiers due to their ability.

Research focusing on home loan eligibility has emphasized the importance of including features related to property value, loan term, and the applicant's repayment capacity. Deep learning approaches and neural networks have also been explored, though their complexity often makes them less interpretable compared to traditional models—an important consideration in the banking sector.

Comparatively fewer studies have focused specifically on education loans, but existing work highlights the unique challenges posed by applicants with limited or no credit history. In such cases, models that incorporate alternative data—such as academic performance, admission status, and future income potential—have shown promise in improving eligibility assessments.

Many studies have conducted comparative analysis of different machine learning models based on performance metrics such as accuracy, precision, recall, and F1-score. These comparisons help identify the most effective model for specific loan types and applicant demographics. Techniques such as cross-validation and grid search have been used to fine-tune hyperparameters and avoid model overfitting.

Given the critical nature of loan decisions, recent work has also focused on improving the explainability of ML models. Techniques like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) have been used to make model decisions more transparent to both bank officials and customers.

Overall, the existing literature demonstrates the potential of machine learning to revolutionize loan eligibility prediction. However, there remains a need for domain-specific models that cater to the unique characteristics of different loan types, such as home and education loans. This study builds upon prior research by developing and evaluating machine learning models tailored specifically for these two loan categories.

Another important area of research has been the integration of alternative data sources to improve model robustness. For example, behavioral data from mobile usage, social media activity, and spending patterns have been used in some studies to supplement conventional financial data, especially in cases where applicants lack sufficient credit history. This is particularly relevant for education loans, where students may not have formal financial footprints.

Several researchers have also examined the impact of data preprocessing techniques on model. This is particularly relevant for education loans, where students may not have formal financial footprints

3. PROPOSED SYSTEM

The proposed system is designed to utilize machine learning techniques to predict the eligibility of applicants for bank loans, with a specific focus on home loans and education loans. The system leverages historical loan application data to build predictive models that can automate and streamline the loan approval process. By analyzing various applicant attributes such as income, credit history, employment status, and other relevant financial indicators, the model will classify whether an applicant is eligible for a specific type of loan. Separate models or feature considerations may be applied for home loans and education loans to reflect the distinct nature and criteria associated with each loan type.

The system begins by collecting and preprocessing data, which involves handling missing values, encoding categorical variables, and scaling numerical features. Important features like applicant and co-applicant income, loan amount, loan term, credit history, property value (for home loans), and education performance (for education loans) are considered. Feature selection techniques are employed to identify the most influential variables that impact loan eligibility. The data is then split into training and testing sets to build and validate the machine learning models.

4. MODULES DESCRIPTION

4.1 DATA COLLECTION

The dataset collected for foretelling loan failure clients is foretold into Training set and testing set. Generally8020 proportion is applied to dissociate the training set and testing set. The data model which was created using Decision tree is applied on the training set and hung on the test take fineness, Test set forecasting is done. Some of the attributes in our dataset are Loan-id Gender, Dependents, Education, self-employed, Applicant Income, Coapplicant Income, LoanAmount, Loan_Amount _term, Credit_history etc.

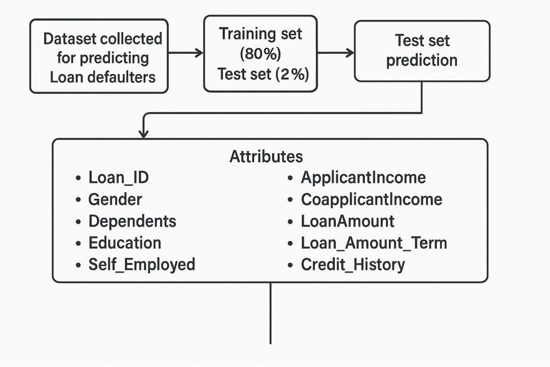


Fig. 4.1 Data Collection

4.2 DATA PREPROCESSING

The collected data may contain missing values that may lead to inconsistency. To gain better results data need to be preprocessed and so it'll better the effectiveness of the algorithm. We should remove the outliers and we need to convert the variables. In order to flooring these issue we use chart function.

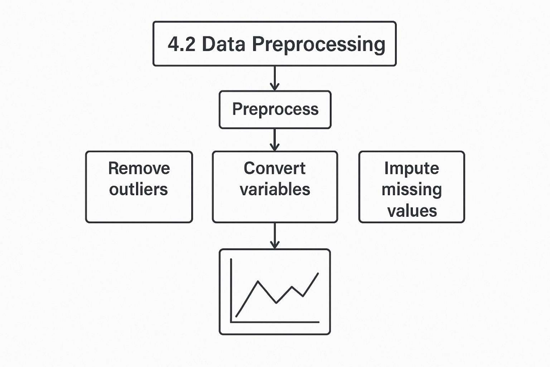


Fig.4.2. Data Preprocessing

4.3. BUILDING A MODEL

After the Pre-processing we will do the exploratory analysis to analyse the data set and to get a clear idea of the characteristics of the data. After the Exploratory Data Analysis is complete it can be used for developing Supervised and Unsupervised learning models. We initially make severalhypotheses by looking at the data before we hit the modelling. EDA helps in confirming and validating the hypotheses we make. For the most part, the Exploratory Information Review is carried out using the accompanying Univariate Analysis Strategies, which outlines the insights of each field in the raw information index and the Bivariate Analysis.

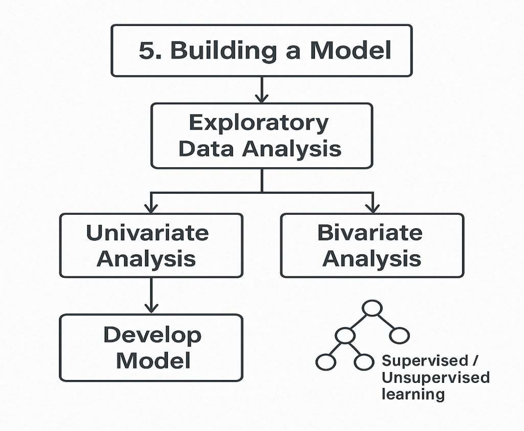


Fig.4.3. Building A Model

4.4. TRAIN MODEL ON TRAINING DATA SET

Now we should train the model on the training dataset and make sooth sayings for the test dataset. We can divide our train dataset into two tract train and testimony. We can train the model on this training part and using that make soothsayings for the testimony part. In this way, we can validate our sooth sayings as we've the true sooth sayings for the testimony part (which we don't have for the test dataset)

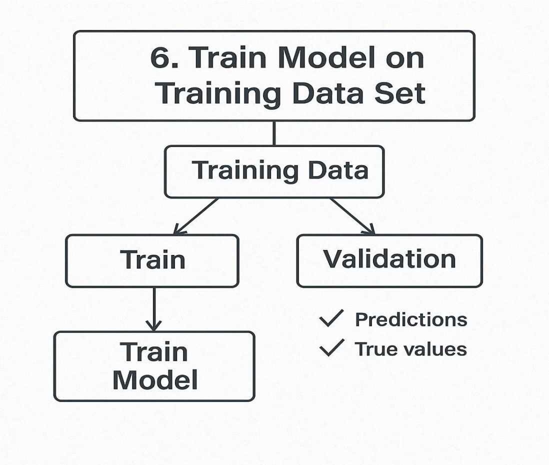


Fig.4.4. Train Model OnTraining Data Set

4.5. LIGHT GBM ALOGORITHM

LightGBM is a gradient boosting framework based on decision trees to increases the efficiency of the model. It uses two novel techniques: Gradient-based One Side Sampling and Exclusive Feature Bundling

(EFB) which fulfills the limitations of histogram-based algorithm that is primarily used in all GBDT (Gradient Boosting Decision Tree) frameworks. The two techniques of GOSS and EFB described below form the characteristics of Light GBM Algorithm. They comprise together to make the model work efficiently and provide it a cutting edge over other GBDT. Different data instances have varied roles in the computation of information gain. The instances with larger gradients (i.e., under-trained instances) will contribut GOSS keeps those instances with gradients (e.g., larger than a predefined threshold, or among the top percentiles), and only randomly those instances with small gradients to retain the accuracy of information gain estimation. This treatment can lead to a accurate gain estimation than uniformly random sampling, with the same target especially when the value of information gain has a large range.

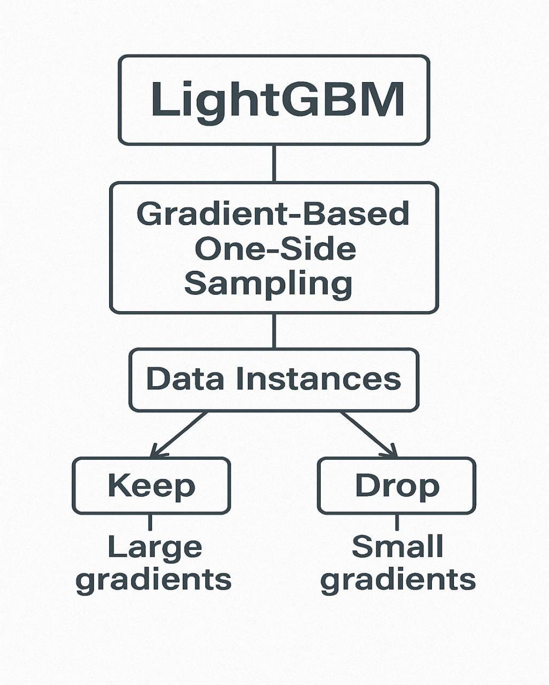


Fig.4.5. Light GBM Alogorithm

4.6. PREDICTING THE OUTCOMES

Initially Import all the required python modules andImport the database for both TESTING and TRAINING, after that Check any NULLVALUES are exists, If NULLVALUES exits, fill the table with corresponding coding Exploratory Data Analysis for all ATTRIBUTES from the table then Plot all graphs using MATPLOTLIB module after that Build the LIGHT GBM Model for the coding Finally we can predict the proper output by using the predicted model.

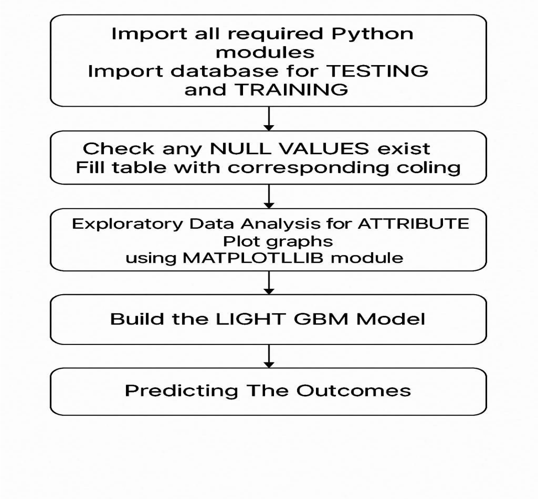

Import all required Python modules
Import database for TESTING and TRAINING
Check any NULL VALUES exist
Fill table with corresponding coling
Exploratory Data Analysis for ATTRIBUTE
Plot graphs
using MATPLOTLLIB module
Build the LIGHT GBM Model
Predicting The Outcomes

Fig.4.6 Predicting The Outcomes

5.EXPERIMENTAL RESULTS

This Result discusses about the This system would be able to determine the status of the loan whether it would get approved or denied swiftly in real-time the below Fig.5.1., Fig.5.2., Fig.5.3. and Fig.5.4. shows the implementation of Homeloan and Education loan

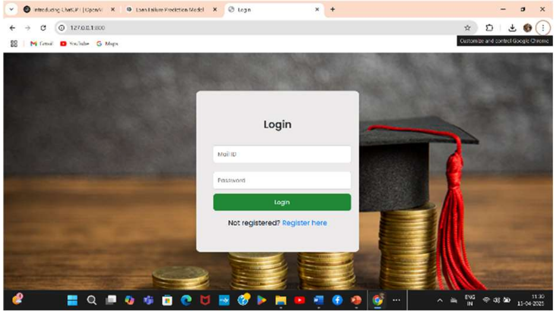
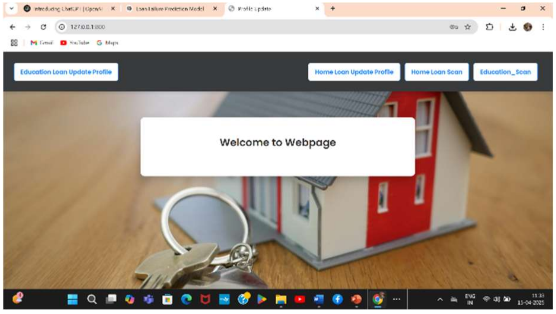
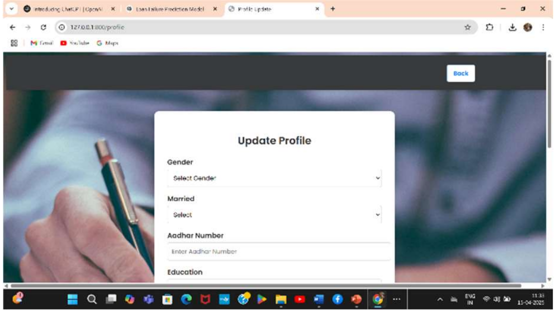


Fig.5.1. Login Page



Welcome to Webpage

Fig.5.2. Home Page



Update Profile

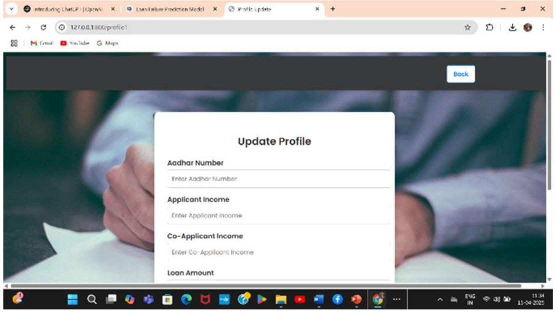
Gender
Select Gender

Married
Select

Aadhar Number
Enter Aadhar Number

Education

Fig.5.3. Home Loan Data Entry



Update Profile

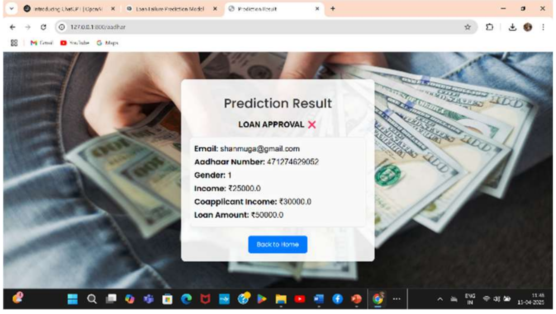
Aadhar Number
Enter Aadhar Number

Applicant Income
Enter Applicant Income

Co-Applicant Income
Enter Co-Applicant Income

Loan Amount

Fig.5.4. Education Loan Data Entry

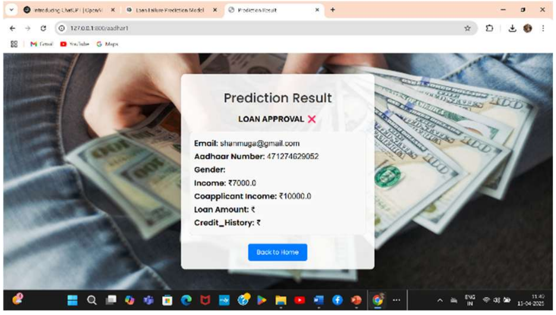


Prediction Result

LOAN APPROVAL ✗

Email: shanmuga@gmail.com
Aadhaar Number: 471274629052
Gender: 1
Income: ₹25000.0
Coapplicant Income: ₹30000.0
Loan Amount: ₹50000.0

Fig.5.5. Home Loan Approval page



Prediction Result

LOAN APPROVAL ✗

Email: shanmuga@gmail.com
Aadhaar Number: 471274629052
Gender:
Income: ₹7000.0
Coapplicant Income: ₹10000.0
Loan Amount: ₹
Credit_History: ₹

Fig.5.6. Education Loan Approval Page

6. CONCLUSION

Displays accuracy with various algorithms. We have compared the LGBM Classifier algorithm to other algorithms, XGB Classifier Algorithm, Decision Tree Algorithm, and Random Forest Classifier Algorithm.

However, of all the algorithms, LGBM Classifier algorithm has the highest accuracy. Also, it can fill the missing values of the datasets, treat categorical values, scalability problems, overfitting problems, and provide a good visualization of the data using a confusion matrix.

For further research, applicants' Age, past health records, as well as the type of occupation they have will be utilized to evaluate the ambiguity factor of paying debts, and possible defaults of corporate loans for businesses and startups can be forecasted.

Another method could be developed to forecast defaulters on different types of loans as well. We used a medium-sized data set to train our model, which may have influenced the outcome; therefore, a big and well-defined data set is required for more accurate results

7. REFERENCE

[1] C. Liu, Y. Ming, Y. Xiao, W. Zheng and C. -H. Hsu, "Finding the Next Interesting Loan for Investors on a Peer-to-Peer Lending Platform," in *IEEE Access*, vol. 9, pp. 111293-111304, 2021, doi: 10.1109/ACCESS.2021.3103510.

[2] M. S. Park, H. Son, C. Hyun and H. J. Hwang, "Explainability of Machine Learning Models for Bankruptcy Prediction," in IEEE Access, vol. 9, pp. 124887-124899, 2021, doi: 10.1109/ACCESS.2021.3110270.

[3] Z. Li, D. Fu and H. Li, "Dynamic Forecasting for Systemic Risk in China’s Commercial Banking Industry Based on Sequence Decomposition and Reconstruction," in IEEE Access, vol. 11, pp. 132068-132077, 2023, doi: 10.1109/ACCESS.2023.3335609

[4] T. Nguyen et al., "Multi-Swarm Optimization for Extracting Multiple-Choice Tests FromQuestion Banks," in IEEE Access, vol. 9, pp. 32131-32148, 2021, doi: 10.1109/ACCESS.2021.3057515.

[5] I. Met, A. Erkoç and S. E. Seker, "Performance, Efficiency, and Target Setting for Bank Branches: Time SeriesWith Automated Machine Learning," in IEEE Access, vol. 11, pp. 1000-1010, 2023, doi: 10.1109/ACCESS.2022.3233529.

[6] S. K. Depren and M. T. Kartal, Prediction on the volume of non-performing loans in Turkey using multivariate adaptive regression splines approach, International Journal of Finance & Economics, vol. 26, no. 4, pp. 6395–6405, 2021.

[7] T. Bozdoğan, ‘‘Analysis of financial performance of foreign banks having branches in Turkey by TOPSIS and ELECTRE methods,’’ Alanya Akademik Bakış, vol. 5, no. 2, pp. 1049–1067, Mar. 2021

[8] K. G. No, S. Niroomand, H. Didehkhani, and A. Mahmoodirad, ‘‘Modified interval EDAS approach for the multi-criteria ranking problem in banking sector of Iran,’’ J. Ambient Intell. Humanized Comput., vol. 12, no. 7, pp. 8129–8148, Jul. 2021.

[9] J. Reig-Mullor and J. M. Brotons-Martinez, ‘‘The evaluation performance for commercial banks by intuitionistic fuzzy numbers: The case of Spain,’’ Soft Comput., vol. 25, no. 14, pp. 9061–9075, Jul. 2021.

[10] G. Szepannek and K. Lübke, ‘‘Facing the challenges of developing fair risk scoring models,’’ Frontiers Artif. Intell., vol. 4, pp. 1–9, Oct. 2021