LOAN ELIGIBILITY PREDICTION USING MACHINE LEARNING

Ramaswamy Vikram ReddyStudent
Department Of Information TechnologyBV Raju institute of technology Affiliated to JNTUH
Vishnupur,Narspur,Medak,TelanganaState,India.

Nalabolu Jyoshna
Student
Department Of Information TechnologyBV Raju institute of technology Affiliated to JNTUH
Vishnupur,Narspur,Medak,TelanganaState,India.

Jayalaxmi.k
Assistant Professor
Department Of Information Technology
BV Raju institute of technology Affiliated to JNTUH
Vishnupur,Narspur,Medak,TelanganaState,India.

Abstract—**Banks are making major part of profits through loans. Loan approval is a very important process for banking organizations. It is very difficult to predict the possibility of payment of loan by the customers because there is anincreasing rate of loan defaults, and the banking authorities arefinding it more difficult to correctly access loan requests and tackle the risks of people defaulting on loans.The applications of this project are: There will be a shorter loan sanctioning period. The entire procedure will be automated, preventinghuman error. A loan will be approved for a qualified applicant right away. Employees of the bank personally verify the applicant's information before awarding loans to those whoqualify. It takes a long time to review every applicant's information. We created automatic loan prediction using machine learning approaches to solve the issue. With previously collected data, we will train the machine. So that the machine may examine and comprehend the process. Then the system will look for a qualified applicant and inform us of the results. In this project, two algorithms are used such as Random Forest algorithm, Logistic Regression algorithm to predict the loan approval of customers. These two algorithms are going to be used on the same dataset and going to find the algorithm with maximum accuracy to deploy the model.**

Keywords-LoanEligibility,Prediction,Logistcregression,Maximum accuracy.

1 .Introduction

In our banking system, banks have many products to sell but main source of income of any banks is on its credit line. So, they can earn from interest of those loans which they credit. A bank's profit or a loss depends to a large extent on loans i.e., whether the customers are paying back the loan or defaulting. By predicting the loan defaulters, the bank can reduce its Non- Performing Assets. This makes the study of this phenomenon very important. Previous research in this era has shown that there are so many methods to study the problem of controlling loan default. But as the right predictions are very important for the maximization of profits, it is essential to study the nature of the different methods and their comparison.A very important approach in predictive analytics is used to study the problem of predicting loan defaulters:

The Logistic regression model. The data is collected from the Kaggle for studying and prediction. Logistic Regression models have been performed and the different measures of performances are computed.While the mentioned features can be used for loan eligibility prediction, it is essential to note that including certain features such as gender or marital status may lead to biased predictions. Therefore, it is crucial to ensure that the model is not discriminating against any group and that the selected features are relevant to the loan approval process.

Furthermore, the use of a deadline for loan approval may not be feasible in some cases as loan processing times can vary based on the institution's policies, regulations, and other factors. It is important to consider the implications of such a system on the customer experience and the institution's reputation.

II. SURVEY OF LITERATURE

A. Predictions are commonly made to anticipate future events, and they can be based on scientific calculations orsimple guessing. Predictive analytics is an advanced analytics branch that uses various techniques such as datamining, statistics, modeling, machine learning, and artificial intelligence to analyze current data and make predictions. Previous research has explored using machine learning technologies like SVM and neural networks to forecast how banks will approve loans. For instance, a study by Kumar Arun et al. in 2016 investigated loan approval prediction using these techniques. Similarly, Mohammad et al. (2010) proposed a study that aimed to predict whether a bank would grant a loan to a customer using Logistic Regression with a sigmoid function. In conclusion, predictive analytics can assist in identifying suitable customers for loan approval, and logistic regression can be a useful approach in determining the probability of loan default correctly.

Their study used a dataset from Kaggle that included training and testing data. Before model development, data cleansing was performed to avoid missing values. The model's performance was assessed using sensitivity and specificity measures, and the final results showed an accuracy of 81%. Interestingly, this model was slightly better because it included variables such as a customer's age, purpose, credit history, credit amount, and credit duration instead of solely relying on checking account information to determine a customer's wealth.

III.Proposed Work

Data collection is the first step in the suggested methodology and then we moved to the data pre-processing. Using the standard hold-out approach, the selected classifiers such as XGBoost, AdaBoost, LighGBM, Random Forest, Decision Tree, and K-Nearest Neighbor are then trained and tested on the provided dataset. To establish the best effective Bank Loan eligibility prediction method, the findings are computed and analyzed. Figure 1 depicts the overview of the proposed strategy.

A. Dataset Collection

The dataset used in this study was obtained from Kaggle's online website and contains 10,128 instances with 23 attributes, of which one is a class attribute and the remaining 23 attributes are predictive. The attributes were selected based on their relevance to describing the eligibility criteria for bank loans. These predictive attributes relate to a person's age, gender, education, property ownership, financial status, income sources, credit card information, and more. The class attribute used for the study is bank loan eligibility prediction.

B. Dataset pre-processing

To prepare the dataset for analysis, we performed several pre-processing steps, including feature extraction, datacleaning, handling missing values, and transforming categorical variables.

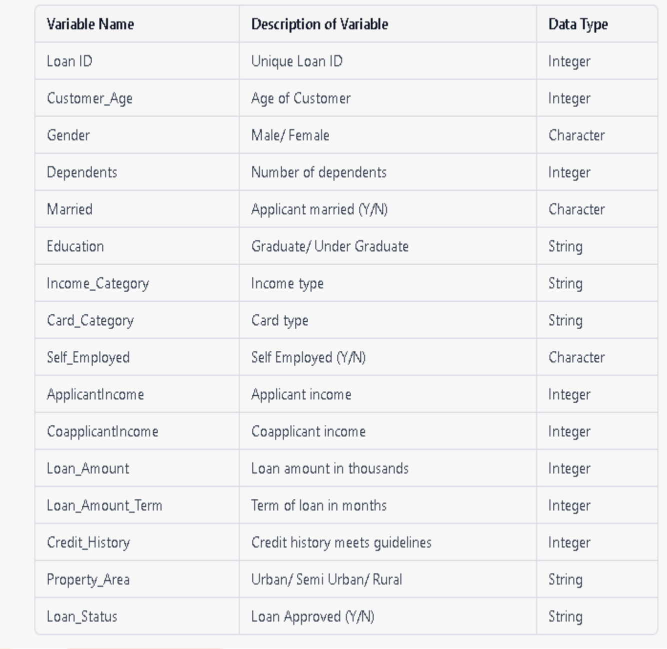
C. Validation process:

The selection of an appropriate validation process is crucial for obtaining accurate results from a dataset. One effective method for validation is the hold-out approach, where 70% of the data is used for training and 30% for testing [12]. We used this approach to evaluate the performance of each machine learning technique, analyzing the results through the confusion matrix and computing measures such as accuracy, precision, recall, area under the curve (AUC), and F1-score.

D. Dataset Descriptions and Pre-processing:

This study utilized a dataset comprising 10,128 applicant records with 23 attributes, including information such as education, marital status, income, and assets. These attributes contain a combination of categorical and numerical data, which are presented in Table 2. Prior to applying the dataset to a machine learning algorithm, we performed pre-processing and feature engineering to address missing values and normalize the data. The dataset was then split into training and testing sets, and the model was trained using machine learning methods, followed by evaluation of its performance on the testing set. Details regarding the model and its results are outlined in the subsequent section.

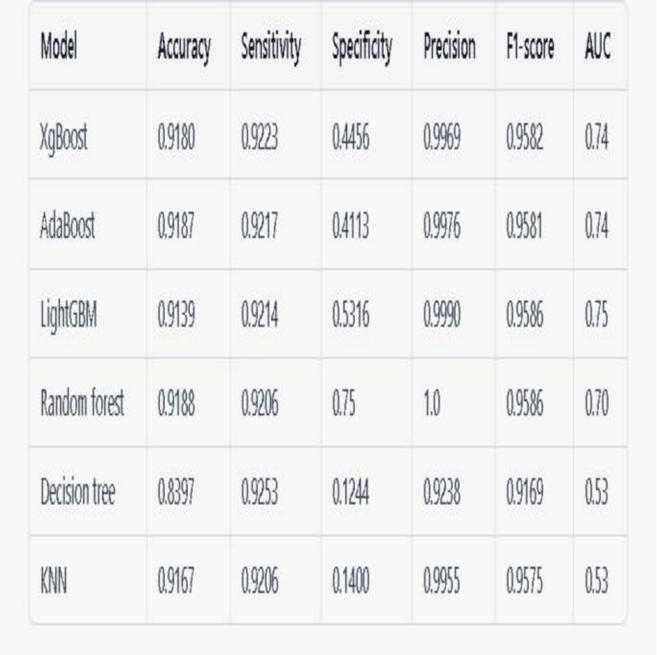
Table 1. Some of dataset attribute names



Variable Name	Description of Variable	Data Type
Loan ID	Unique Loan ID	Integer
Customer_Age	Age of Customer	Integer
Gender	Male/ Female	Character
Dependents	Number of dependents	Integer
Married	Applicant married (Y/N)	Character
Education	Graduate/ Under Graduate	String
Income_Category	Income type	String
Card_Category	Card type	String
Self_Employed	Self Employed (Y/N)	Character
ApplicantIncome	Applicant income	Integer
CoapplicantIncome	Coapplicant income	Integer
Loan_Amount	Loan amount in thousands	Integer
Loan_Amount_Term	Term of loan in months	Integer
Credit_History	Credit history meets guidelines	Integer
Property_Area	Urban/ Semi Urban/ Rural	String
Loan_Status	Loan Approved (Y/N)	String

E. Result Analysis

Table presents the mean performance of various machine learning classifiers, namely XGBoost, LightGBM, Adaboost, Decision Tree, Random Forest,and KNN. Subsequently, we examined the outcomesof these models by referring to Figures To evaluate the efficacy of the models, we have provided their Accuracy, Precision, Recall, F1Score, and AUC metrics in Table .



Model	Accuracy	Sensitivity	Specificity	Precision	F1-score	AUC
XgBoost	0.9180	0.9223	0.4456	0.9969	0.9582	0.74
AdaBoost	0.9187	0.9217	0.4113	0.9976	0.9581	0.74
LightGBM	0.9139	0.9214	0.5316	0.9990	0.9586	0.75
Random forest	0.9188	0.9206	0.75	1.0	0.9586	0.70
Decision tree	0.8397	0.9253	0.1244	0.9238	0.9169	0.53
KNN	0.9167	0.9206	0.1400	0.9955	0.9575	0.53

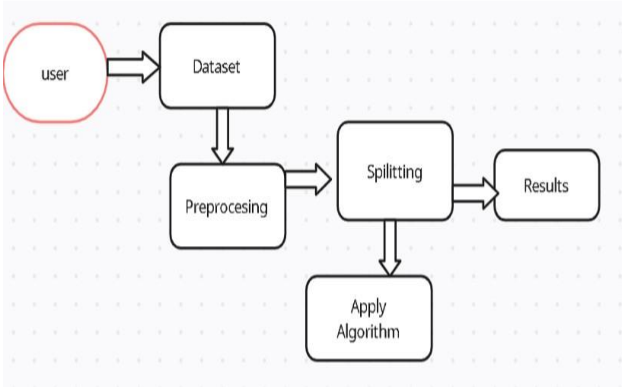


Fig 1.Architecture daigram

1. To create a loan eligibility prediction system, the raw data is initially processed and cleaned to ensurethat the testing and training phases are performed accurately.
2. Once the data is cleaned, the training data is used to train a classification model to evaluate the system's accuracy.
3. Subsequently, the test data is used to test the machine learning model. Based on the information obtained from the classification model and the training data, the machine learning model makes predictions about the eligibility of the loan applicant.

.

IV Logistic Regression .

The Scikit-learn library in Python offers the Logistic Regression model, which serves both as a statistical model and a classification algorithm. Its purpose is to predict a binary outcome by analyzing a set of independent variables. For this particular project, only the variables that have a direct impact on an applicant's loan eligibility, such as Credit History, Education, Self-Employment, and Property Area, are taken into account. However, not all variables are used at the same time to avoid overfitting, wherein the model becomes too specific to the dataset and does not generalize well to new situations. Including more attributes to the model increases the risk of overfitting and may lead to inaccurate predictions for new scenarios.

V CLASSIFICATION MODEL

Logistic regression is a statistical technique used to predict the likelihood of a binary outcome based on one or more independent variables, which can be continuous or categorical. It estimates the relationship between the predictor variables and the probability of the binary outcome using a logistic function. The model is trained using maximum likelihood estimation, and once trained, it can be used to classify new observations based on the predicted probability of the outcome. Logistic regression is commonly applied in fields such as medical research, marketing, and credit risk assessment.

V REFERENCES

- Gupta, Anshika, et al. "Bank Loan Prediction System using Machine Learning." 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART). IEEE,2020.
- Kumar, Arun, Garg Ishan, and Kaur Sanmeet. "Loan approval prediction based on machine learning approach." IOSR J. Comput. Eng 18.3, 18- 21, 2016.
- M. A. Sheikh, A. K. Goel and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learni ng Algorithm," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 490-494, 2020.
- Supriya, P. Usha et al. “Loan Prediction by using Machine Learning Models.” ,2019.
- Loan Approval Prediction using Machine Learning Algorithms Approach. 2021 [Ebook]. Retrieved fromhttps://ijirt.org/master/publishedpaper/IJIRT151769_PAPER.pdf Ndayisenga, Theoneste. Bank Loan Approval Prediction Using Machine Learning Techniques. Diss. 2021. Tejaswini, J., et al. "Accurate loan approval prediction based on machine learning approach." Journal of Engineering Science vol. 11, no.4, pp. 523-532. 2020.

VI CONCLUSION AND FUTURE WORK

We can observe that relying solely on CIBIL scores is not sufficient for accurately assessing a borrower's credibility. Additional parameters must also be considered, but manually analyzing all these parameters can be time-consuming and inefficient.Our research proposes a solution to this problem by developing a Logistic Regression model based on Machine Learning techniques. This model takes into account all the relevant parameters required toevaluate a client's creditworthiness. After being trained to produce satisfactory accuracy, the model can accurately determine whether a borrower should be granted a loan or not without requiring any tedious manual work.