IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15660

LEP-Model: Ascertain Loan Predictions Using ML Approach

Pratibha Deshmukh[1], Yogendra Chhetri[2], Harsh Nakti[3], Vinay Gupta[4], Shivam Patil[5]

Bharati Vidyapeeth's Institute of Management and Information Technology, Navi Mumbai, India,

University of Mumbai[1]

Dept. of Centre for Continuing Education, Indian Institute of Science (IISc), India[2]

MCA BVIMIT, Navi Mumbai, India[3-5]

Abstract: As a result of the banking industry's advancements, a large number of people are applying for bank loans. However, the bank can only approve a limited number of applicants due to its limited resources, so determining who will be a safer candidate for approval is a common procedure. We therefore attempt to lower the risk involved in choosing the safe individual in this study report in order to preserve numerous bank endeavors and assets. This is accomplished by taking information from the previous records of the borrowers and based on these records, the machine was trained using an ML and Python model to provide the most accurate results. Assigning a debt to a particular individual or not is the primary goal of this study article. With the Logistic Regression algorithm receiving the maximum score of 80.78%, our research result demonstrated good performance accuracy. Finding out if it will be safe to give a loan to a specific individual is the first priority. The principal objective of the research paper is to forecast the loan eligibility of the clients and ascertain the conditions that precluded them from obtaining a loan for the construction of their own home.

Index Terms: Loan Prediction, Machine Learning, Logistic Regression, Banking Sector.

I. INTRODUCTION

The banking sector is an important part of the financial industry that offers a range of financial products and services to individuals and businesses. Banks provide a variety of services such as deposit accounts, loans, credit cards, insurance, investment services, and foreign currency exchange. The banking sector is a vital part of the economy because it channels funds from depositors to borrowers efficiently. Banks mobilize customer deposits and extend credit to individuals, businesses and institutions for various productive purposes. These loans support for various sectors leads towards business development. Banks collect deposits from customers and use these funds to lend to businesses and individuals who need capital for their operations. Banks also play a key role in managing monetary policy, ensuring financial stability, and promoting economic growth.

The current situation calls for a bank representative to manually approve each loan, meaning that person is in charge of determining the borrower's eligibility as well as the loan's associated risk. Being a human-to-human process, it takes a lot of time and is prone to mistakes. Since interest is how banks make the majority of their earnings, failure to repay the loan would result in a loss for the bank. A banking crisis will occur if the banks experience excessive losses. This banking crisis has an impact on the national economy. Therefore, it is necessary to have a loan prediction model that can swiftly determine with the least amount of risk whether the loan can be approved or denied. In this paper, we compared several algorithms and used the best of them for prediction. We are using a logistic regression algorithm because it gives the best accuracy than other algorithms.

II. LITERATURE SURVEY

Predictions are commonly made to anticipate future events and they can be based on scientific calculations or simple guessing. As a subfield of advanced analytics, predictive analytics analyses and forecasts current data using a variety of methods including data mining, statistics, modeling, machine learning, and artificial intelligence. 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 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 84%. Interestingly, this model was slightly better because it included variables such as a customer's age, purpose, credit score, education, dependents, marital status, and credit duration instead of solely

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026
DOI: 10.17148/IJARCCE.2026.15660

relying on checking account information to determine a customer's wealth. A. Sarkar created a statistical model [1] by preprocessing, combining, and putting data into three machine learning models: Random Forest, Decision Tree, and Logistic Regression. With an accuracy rating of 80.78%, logistic regression proved to be the most reliable model for predicting loan acceptance. Nitesh Pandey et al.'s work [2] examined loan prediction by machine learning technique using data from prior customers from different banks. They used the most accurate machine learning technique Support Vector Machine (SVM) in addition to four other methods: Random Forest, Decision Tree, Logistic Regression, and Support Vector Machine. For evaluation, the confusion matrix, F1 score, precision, recall, and accuracy are used.

III. RESEARCH OBJECTIVES

This study paper's primary goal is to forecast an applicant's eligibility for a loan. Through the reduction of risk and default rate, these prediction systems benefit the bank as well as the applicant. The machine learning approach automates the entire procedure. Compiling pertinent data from multiple sources, including credit histories, financial records, demographic data, and economic indicators, handling outliers, inconsistent data, and missing values by cleaning and preparing the data. Using feature engineering, useful features that can improve the model's predictive capacity can be extracted. Analyzing variables including credit score, income, debt-to-income ratio, work status, and loan amount in order to assess the credit risk associated with each loan application. Calculating using the anticipated risk score the likelihood of default or delinquency. Giving lenders the tools they need to make well-informed judgments about whether to approve or refuse loans.

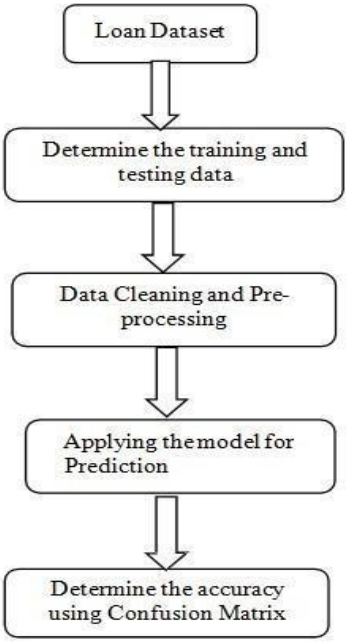
IV. METHODOLOGY

This study was carried out using Kaggle's Jupyter Notebook cloud environment. Where Python was employed for data analysis and model development. The proposed model utilizes the available dataset. Additionally to predict a customer's loan eligibility based on relevant input features. As shown in Table I, the selected attributes from the dataset are provided to the model. Thereafter it generates an eligibility prediction for loan approval. The following section explains the data cleaning and preprocessing steps.

A. Collection of Data

A historical dataset named Loan Eligible Dataset available on Kaggle was used as the primary data source. The dataset is published under the Database Contents License (DbCL) v1.0. Ensuring authorized use for analytical and research purposes. Table I presents an overview of the key characteristics and attributes contained in the dataset.

TABLE I BANK CUSTOMER DATASET



IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15660

Variable Name	Description	Data Type
Loan_ID	Loan reference number (Unique I.D.)	Numeric
Gender	Applicant gender	Categorical
Married	Applicant marital status	Categorical
Dependents	Number of family members	Numeric
Education	Applicant educational qualification (graduate or not graduate)	Categorical
Self_Employed	Applicant employment status (yes for self-employed, no for employed/others)	Categorical
Applicant_Income	Applicant's monthly salary/income	Numeric
Coapplicant_Income	Additional applicant's monthly salary/income	Numeric
Loan_Amount	Loan amount	Numeric
Loan_Amount_Term	The loan's repayment period (in days)	Numeric
Credit_History	Records of applicant's credit history (0: bad credit history, 1: good credit history)	Numeric
Property_Area	The location of the applicant's home (Rural/Semi-urban/Urban)	Categorical
Loan_Status	Status of loan (Y: accepted, N: not accepted)	Categorical

Fig. 1 Variable Relationships Overview

B. Data Preprocessing and Analysis

The below mentioned methods are used to evaluate and prepare the data for modeling for optimum performance:
One is the One-hot encoding Method which is used to make categorical variables in a dataset more understandable to the machine learning model. Other one is the Normalization which used to transforming characteristics and making sure they are all on the same scale. Moreover they are the objectives of normalizing data for ML models. Addition to this is the Exploratory Data Analysis (EDA) method - The process of examining a dataset to find patterns, trends, and anomalies. At this point, the dataset was additionally cleaned to eliminate or manage incomplete or missing data by data imputation.

V. DESIGN

Loan Dataset: The Loan Dataset is a valuable resource for our system's more accurate outcome prediction. The loan dataset is utilized by the system to automatically determine which customer loans should be approved and which should be rejected. The loan application form will be accepted by the system as input. An application form in a justified format must be provided as an input to be processed.

Determine the Training and Testing Data: The majority of the data in this dataset is used for training, and a smaller piece is used for testing. Usually, the system divides a dataset into training and testing sets. The prediction is made by the system against the test set following its processing of the training set.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15660

Data Cleaning and Processing: Data cleaning is the process of recognizing and fixing corrupt or inaccurate records from databases. It involves determining whether sections of the data are incomplete, wrong, or irrelevant, and then replacing, updating, or identifying the coarse or dirty data. The system used in data processing transforms data from one form to another, making it more relevant and instructive and much more usable.

Applying the Model for Prediction: Clean up and preprocess the data, deal with missing values and outliers, and scale the features to prepare your dataset. Create training and testing sets from the data. Select a machine learning model (classification, regression, etc.) that is appropriate for the prediction task. To understand patterns and correlations, train the model with the training set of data. Analyze the model's performance in terms of accuracy and generalizability using the testing data. Based on the evaluation results, modify and refine the model as necessary.

Determine the Accuracy Using Confusion Matrix: When a classification model is evaluated for accuracy, its predictions are summarized into four outcomes: true positive (TP), true negative (TN), false positive (FP), and false negative (FN) in the confusion matrix. Divided by the total number of forecasts in terms of accuracy is determined as (TP + TN). The confusion matrix helps to clarify classification errors. Also offers a thorough analysis of the model's performance across many classes. Interpreting these metrics in conjunction with other performance indicators like precision, recall and F1-score is crucial for an assessment.

VI. IMPLEMENTATION

This research helps the banking sector for their loan approval and selecting the right customer for a loan. To find the right customer who is eligible for a loan, researcher used a loan prediction system. In this system, one can choose the right customer who can repay the loan amount to the bank in the given period. The model predicts the right customer using the customer credit score, monthly income, dependents, loan amount, loan duration and assets.

Checks the missing values of the data and correct them.
Preprocess the customer data for prediction.
After that, it uses label encoding for every column. Numerical values in columns can be obtained by applying a technique called label encoding. By doing this, incremental integer labels are used to encode the values. Table columns such as "Credit Score," "Dependants" and "Gender" for instance, can have the corresponding encodings of 0, 1, and 2.

Fill the null values of the data. We use the fillna() function to replace the null values with some value of their own.
Creating the heatmap for correlation. A heatmap visualization of the correlation matrix of the Data frame, allowing you to quickly identify patterns and relationships between variables in your dataset.
Splitting the data into X and Y.

train_X = data.iloc[:614,] all the data in X (Train set)
train_y = Loan_status Loan status will be our Y

We calculate the different machine learning model accuracy for our prediction and we will use the best of them. After calculating the result, logistic regression gives the highest accuracy among them, therefore we used logistic regression for our model.

This model works well for problems where there are two possible outcomes, like loan approval. In loan approval the outcome is: Approved or rejected. The model also helps us see how different things affect the decision. Logistic regression gives us reliable results when choosing the right customer for a loan. It helps us understand loan approval better. The model is good at providing reasons for its predictions. Logistic regression is useful for loan approval.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15660

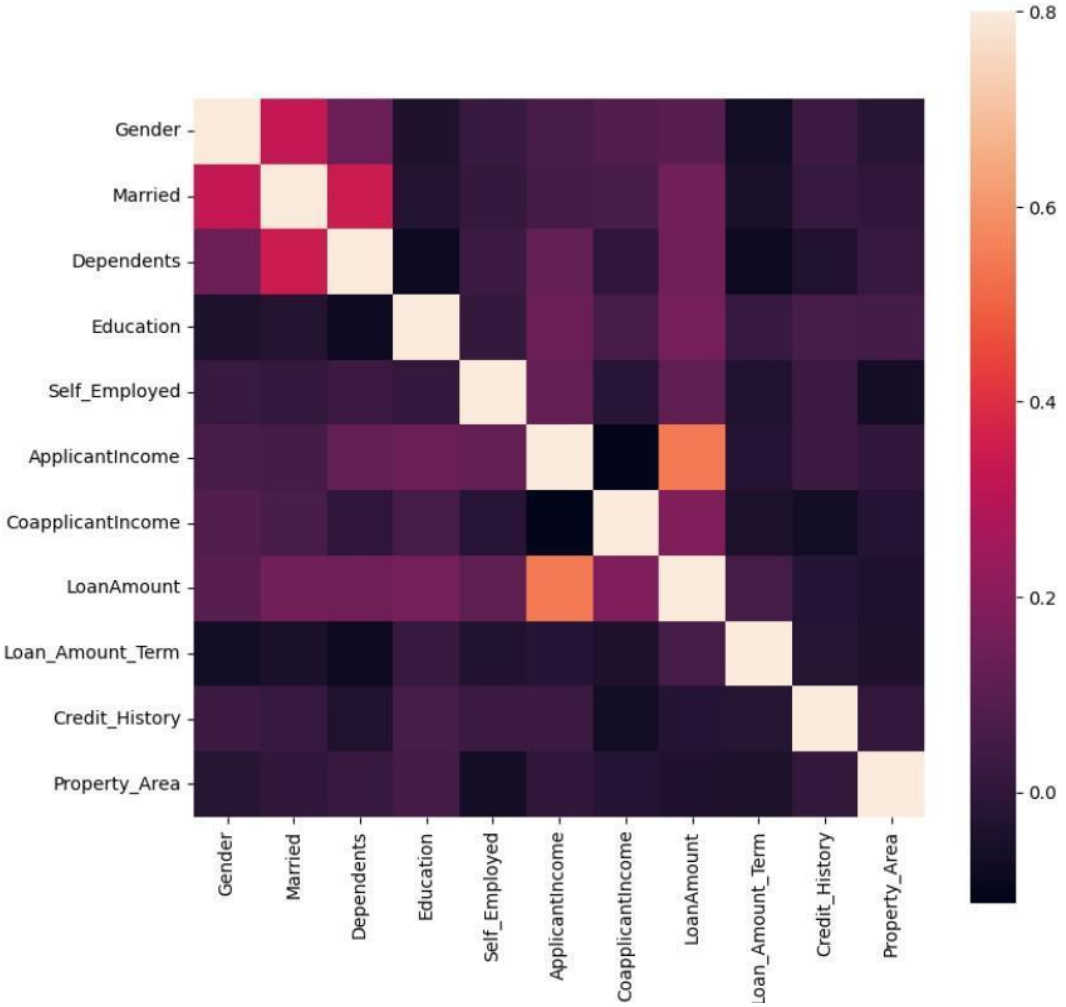


Fig. 2 Workflow Diagram

VII. MODEL AND ALGORITHM USED

A. Logistic Regression Model

The logistic regression model is one of the most frequently used models for binary classification. The model that is displayed has the equation:

$$P(Y = 1) = \frac{1}{1 + e^{-(b_0 + b_i X_i)}}$$

B. Naive Bayes (NB) Model

The Bayes Theorem (BT) is the basis of the NB model. The input features often referred to as predictors are taken to be independent according to BT. The model that is displayed has the equation:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where, A and B are events and P(B) ≠ 0.

C. Random Forest (RF) Model

A supervised learning strategy is Random Forest. The model that is displayed has the equation for a new instance x. The prediction of the Random Forest model is calculated based on the predictions of all individual trees:

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15660

Classification: Majority Vote (Mode)

$$\hat{Y} = \arg\max_i \sum_{k=1}^{n} I\left(h_k(x) = i\right)$$

Regression: Mean (Average)

$$\hat{Y} = \frac{1}{n} \sum_{k=1}^{n} h_k(x)$$

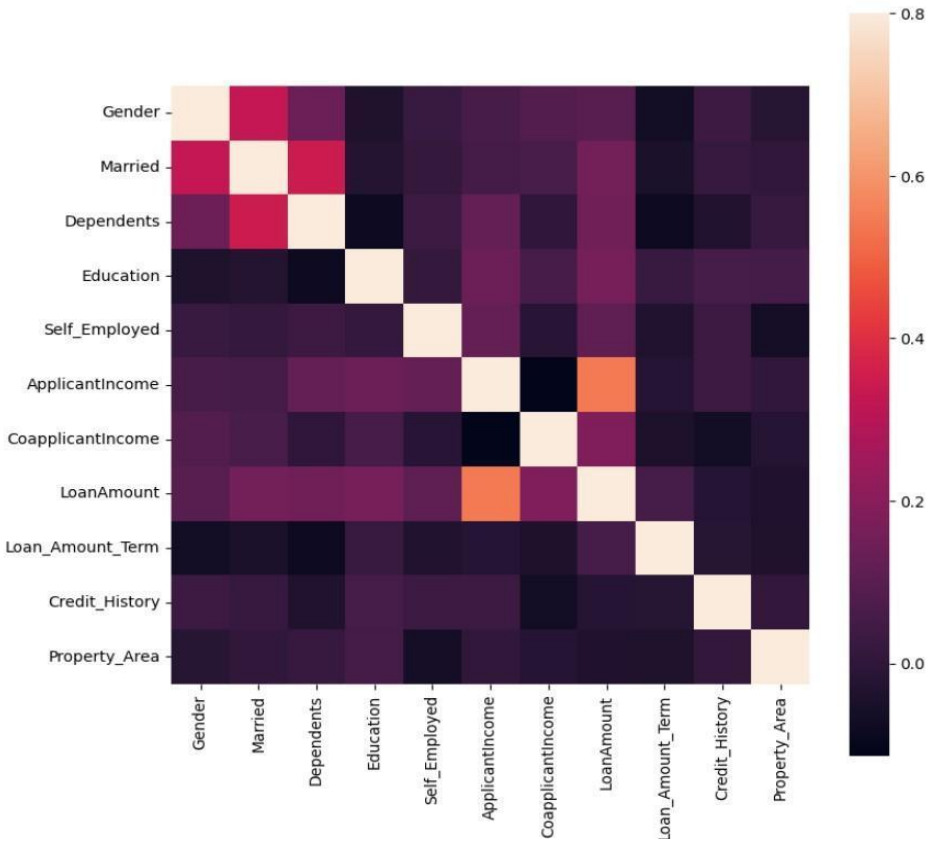


Fig. 3 Heatmap Visualization of the Correlation Matrix

TABLE II MODEL PERFORMANCE COMPARISON

MODEL	ACCURACY (%)
Logistic Regression	80.78
Naive Bayes	79.02
Random Forest	78.35
Support Vector Machines	77.56
Decision Tree	76.89
Gradient Boosting	76.54
Linear Discriminant Analysis	75.67
K-Nearest Neighbour	75.01

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15660

VIII. RESULTS AND DISCUSSION

Table II presents an overview of the performance of several machine-learning algorithms in terms of their accuracy. Of all these algorithms tested for predicting loan eligibility, Logistic Regression exhibited the best performance (80.78%); therefore, Logistic Regression is the best-fitting algorithm for predicting loan eligibility. Logistic Regression's strong performance demonstrates how well suited an algorithm is to binary classification tasks and its reliability and interpretability for binary classification tasks. Naive Bayes and Random Forest also performed reasonably well but were not as good at predicting loan eligibility as Logistic Regression. The implementation of data preprocessing, missing value handling and categorical variable encoding toward enhancing model performance. It indicated that income credit score and loan amount had a significant effect on the prediction of the outcome variable. Therefore, Logistic Regression was found to be an efficient and pragmatic method for predicting loan approvals.

IX. CONCLUSION

The proposed LEP-Model demonstrates machine learning techniques can significantly improve the efficiency and reliability of loan prediction. Additionally it explores the comparative analysis of multiple algorithms. Where the Logistic Regression method provides the most suitable balance between predictive accuracy and interpretability in loan eligibility assessment. The study highlights that although several models produce effective results, Logistic Regression remains highly practical for real-world banking applications. Because of its transparent decision-making capability and consistent performance. Moreover the findings suggest that ML-based loan prediction systems can support financial institutions in reducing credit risk. Future work may focus on integrating advanced ensemble techniques and deep learning models. Which can work with borrower behavioral attributes and macroeconomic indicators to further enhance predictive performance.

REFERENCES

[1]. A. Sarkar, "Loan Prediction using Machine Learning Algorithms," *International Journal of Advanced Research in Computer Science*, vol. 8, no. 5, pp. 123-127, May 2017.
[2]. Nitesh Pandey, Shruti Bhattacharya, and Meghana Srinivas, "Machine Learning Techniques for Predicting Loan Approval," *Proceedings of the International Conference on Data Science and Engineering (ICDSE)*, pp. 45-50, 2018.
[3]. Kaggle, "Loan Eligibility Dataset," [Online]. Available: https://www.kaggle.com/datasets/
[4]. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.
[5]. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.
[6]. J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189-1232, 2001.
[7]. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 785-794, 2016.
[8]. D. G. Altman, *Practical Statistics for Medical Research*. Chapman and Hall/CRC, 1993.
[9]. R. Kaur and A. Goyal, "Using the ensemble model that is a combination of two or more algorithms for Prediction," 2016.
[10]. P. Rawat and S. Tiwari, "Random forests," *Machine Learning*, vol. 45, no. 1, 2017.
[11]. S. Vimala and K. C. Sharmili, "Prediction of Loan Risk using NB and Support Vector Machine," *Int. Conf. on Advancements in Computing Technologies (ICACT 2018)*, vol. 4, no. 2, pp. 110-113, 2018.

