Global Journal of Modeling and Intelligent Computing (GJMIC)

Vol. 3, No. 1, January, 2023 ISSN: 2767-1917

LOAN ELIGIBILITY PREDICTION USING MACHINE LEARNING: A COMPARATIVE APPROACH

Sonali Kumari, Siksha 'O' Anusandhan University, India, (chaturvediswarna864@gmail.com)
Debasish Swapnesh Kumar Nayak, Siksha 'O' Anusandhan University, India (swapnesh.nayak@gmail.com)
Tripti Swarnkar, Siksha 'O' Anusandhan University, India (triptiswarnkar@soa.ac.in)

Abstract - Machine learning (ML) algorithms can bring revolution in the research field in almost all areas. Processes in numerous industries, including finance, real estate, security, and genomics, are being transformed by machine learning (ML) algorithms. One of the major impediments in the banking sector is the loan approval process. Modern tools like ML models help accelerate, streamline, and increase the precision of loan approval procedures. It will benefit both the client and the bank in terms of time and manpower required for loan eligibility prediction. The entire work is cantered on a classification problem and is a form of supervised learning in which it is important to determine whether the loan will be approved or not. Also, it is a predictive modeling problem where a class label is predicted from the input data for a specific sample of input data. In this work, we deployed various ML algorithms to identify the loan approval status and compare the performance of implemented models. The implemented models will attempt to predict our target column on the test dataset using information from the loan eligibility prediction dataset obtained from Kaggle, which includes features like loan amount, number of dependents, and education. The parameters like accuracy, confusion matrix, ROC curve, and precision are measured for specific models whose performance is significant.

Keywords: Machine Learning (ML), Supervised Learning, Loan Eligibility Prediction (LEP), Kaggle, Real estate.

1. INTRODUCTION

The banking sector is a crucial component of any economy. Banks serve as intermediaries between savers and borrowers and are responsible for providing financial services to individuals, businesses, and governments. One of the most important services provided by banks is loans. Loans are essential for the growth of the economy (King et al., 2001). They allow individuals and businesses to invest in new ventures, buy homes and cars, and make other purchases that they would not be able to afford without borrowing. Loans also help businesses to expand their operations, hire more employees, and ultimately contribute to the growth of the economy. In addition to providing access to credit, banks also play a critical role in managing risk. Banks carefully evaluate loan applications to determine the creditworthiness of the borrower and assess the risks associated with the loan (Samreen, 2012). By doing this, the risk of default is reduced because loans are given to borrowers who are more likely to repay them. Banks create lending policies, and according to such regulations, loans are approved based on the status of the applicant. Loan applications are reviewed in accordance with the standing of the applicant under the lending policies established by banks. Loans are frequently only approved by banks after a thorough evaluation of the applicant's condition, either through methodically examining submitted documents or through direct asset verification. However, there is no guarantee that the individual who was selected from all of the candidates is the best one (Sheikh et al., 2020).
Machine learning can help automate loan eligibility prediction by analyzing vast amounts of data and identifying patterns and trends to eliminate human error. One of the main benefits of machine learning is that it can learn from past data and use that knowledge to predict future outcomes (Srivastava et al., 2018). By training machine learning algorithms on historical loan data and outcomes, banks can develop models that can accurately predict loan eligibility based on various factors such as the qualification of an applicant, gender, employment status, and other relevant factors. The machine learning models can also help identify critical risk factors that can be used to refine the loan application process by identifying high-risk applications and prompting human review (Al Mamun et al., 2022). By making loan decisions more quickly and accurately, eliminating the need for manual review, and making personalized loan recommendations, machine learning can also enhance the overall client experience. Overall, machine learning can enhance the speed and accuracy of loan processing while also enhancing the applicants experience (Singh et al., 2021).

In this paper, we explored the capability of various machine-learning classification algorithms. The analysis of these algorithms leads to the suggestion of an original solution in this area. The deployment of various models and the model architecture is shown in Figure 1.

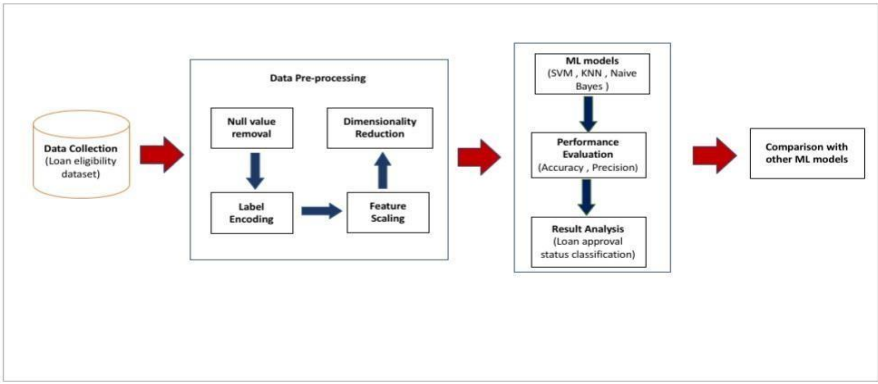

Figure 1: Proposed pipeline of the Model

2. RELATED WORK

C. Naveen Kumar et al., (2022), discussed the traditional loan approval process and how they are unsafe for the bank. They discussed data mining and machine learning approaches to automate the loan approval process so that it will save time and resources. This work aims to replace the traditional loan eligibility approval process in the banking industry.

Park et al., (2021), compared six different machine-learning algorithms based on precision, recall, and accuracy. The objective of this work is to predict if a loan will be approved or not. Random Forest displayed the most accuracy in their study, at 95.55%, whereas Logistic Regression displayed the lowest accuracy.

Mohankumar et al., (2016), discussed how feature selection is important in a predictive model. They used several methods and approaches for the same. They also suggested that a linearneural network can be used instead of a linear regression model to take advantage of the vivid powerof an artificial neural network.

Awodele et al., (2022), discussed that the cascade of Deep Learning networks and Support vector machines improves accuracy by 9%. In this study, the authors employed DNN to convert input data from a lower dimension to output features in a higher dimension, which wasfurther used to upskill a support vector machine-based classification model.

Pidikiti et al. (2019), the main objective was to reduce the risk involved with choosing a reliableapplicant to sanction the loan that will many resources including efforts and cost for the bank.In this article, they used machine learning techniques like classification, logistic regression, decision trees, and gradient boosting to forecast loan data. The decision tree algorithm performed best when compared to other algorithms, with an accuracy of 82 percent. It was effective since the classification problem results were improved.

Foretelling loan non-payers is the most burdensometask for every bank, according to Pandey et al. (2010). However, banks can drastically cut their losses by minimizing non-profit assets by foreseeing loan defaulters. The study of loan approval prediction thus became essential. Machine learning algorithms are very significant and helpful in the prediction of this kind of data. This study deployedfour machine learning algorithms and later compared them. With a high accuracy of 79.67%, Support Vector Machine proved to be the most efficient model for predicting loan acceptance in this case. From various banks that had supported several boundary advances, they collected a list (dataset) of former clients' information.

An assertion regarding what one anticipates will happen in the future is known as a prediction. Every day, predictions are made. Some are serious and founded on mathematical calculations, while others are only educated guesses. Predicting what will happen in the future, whether it be in a few months, a year, or ten years, can help us in several ways. In their 2016 study, Kumar Arun et al. experimented the ways to predict what are the different proceduresin which any bank can sanction loan applications. They deployed a model that makes use of machine learning tools like neural networks and SVM.

According to the clients historical financial and credit scores, Tejaswini et al. (2020) proposed a vigorous predictiveapproach to decide whether an applicant should be accepted or rejected. This study set out to provide a quick, easy, and efficient method for selecting qualified applications. The dataset used was collected from different commercial organizations. The dataset was divided into training and testing, the training dataset was used to train the model whereas the testing dataset was used to validate the model. Three machine learning algorithms were used in this study to forecast customer loan approval. The evaluation outcomes suggestthat the Decision Tree algorithm has outperformed all models with the highest accuracy of 82%.

A Solid machine-learning model to forecast loan acceptance was put forth by the authors of Shrishti et al. (2018). Fast loan approval for applicants was the main objective of this program. Logistic Regression, Decision Trees, and Random Forest were the three machine algorithm types they employed. The Random Forest algorithm was found to have the highest accuracy of all the models after examining the data sets for various models.

Author Ndayisenga, T., (2021) bestowed the study with financial banks to foretell the nature of debtors byunfolding and assessing the efficacy of numerous models using data from a Bank in Kigali. Training and test datasets were produced by splitting the dataset into the ratio of 70:30. Association was utilized to determine the most efficient machine learning models for predicting defaulters for any bank. The two most effective models for predicting loan nonpayers were discovered to be gradient boost (accuracy 80.40%) and XGBoost, while decision tree, logistic regression, and random forest did not perform well.

3. METHODOLOGY

Loan eligibility approval systems in the banking sector can be beneficial for both customers and the bank. Machine learning approaches can ease the whole process as they will be cost-effective, and efficient and will save time. In this work, important characteristics that are necessary for predicting loan eligibility are discussed (Reddy et al., 2022).The dataset used in this study was gathered from a public repository. Training and testing datasets are created after pre-processing the data to increase its overall quality (Gupta et al., 2020). Machine learning models are developed using training data, while test data are used to assess the model's performance (Nayak et al., 2022). Decision tree, random forest, support vector machine, K-nearest neighbor, Naïve Bayes, logistic regression, and linear regression are deployed here and their effectiveness in predicting loan eligibility is evaluated (Kadam et al., 2021).

4. IMPLEMENTATION

A common Python distribution used for data science and scientific computing, Anaconda comes with a graphical user interface (GUI) named Anaconda Navigator (https://docs.anaconda.com/free/navigator/index.html). Jupyter Notebook is only one of the tools and environments that can be managed and launched with ease using Anaconda Navigator.Jupyter Notebook is a potential implementation option (https://jupyter.org/). It is interactive and features a user-friendly environment for creating machine learning models.

4.1 Dataset

The dataset for predicting loan eligibility is gathered from the Kaggle public repository. Thirteen (13) characteristics and 614 records build up the dataset. Loan id, Gender of the applicant, marital status of the applicant, dependents in the family, qualification, employment status, applicant's income, co-applicant income, loan amount and term, credit worthiness, property area, and status of loan are the features present in the dataset. The table below shows all the features with their description.

Table 1: features and their description

Attributes	**Description**
Loan_id	A unique ID
Gender	Male/Female
Married	Married/unmarried
Education	Graduate/Undergraduate
Self_employed	Yes/No
Dependents	Number of dependents
Applicant_income	Applicant's income

Coapplicant_income	Co-applicants income
Loan_amount	Loan amount
Loan_term	Loan's repayment period
Credit_history	Creditworthiness
Property_area	Urban/rural/semi-urban
Loan_status	Yes/No

4.2 Data Pre-processing

For the models to perform well, the dataset needs to be pre-processed. In this work, we used four Pre-processing techniques which are as follows.

4.2.1 Handling missing values: Our dataset has many missing values and to handle that we used two functions namely bfill() and fill() to fill all the null values so that our dataset will work efficiently (Yuvarani et al., 2023).

4.2.2 Label encoding: Our dataset had both categorical variables as well as numerical variables so to bring uniformity we used labelencoder() to convert all the categorical variables into numerical format (Wang et al., 2022).

4.2.3 Feature scaling: Feature scaling is a Pre-processing technique that is used to transform the dataset into a certain range. There are two scaling methods, one is normalization and another one is standardization. We have used standard scalar to standardize our dataset. It removes the mean and scales the variables to the unit variance.

4.2.4 Linear discriminant Analysis: Linear discriminant Analysis is a Pre-processing step that reduces the dimensionality of the dataset and transforms the dataset from 2 dimensional to 1 dimensional. It also reduces the curse of dimensionality.

4.2.5 Feature correlation: Making data-driven judgements and seeing patterns requires an understanding of the relationships between characteristics. Figure 2 displays the association between various features in the dataset we used (https://www.kaggle.com/code/vinodkumargr/loan-prediction).

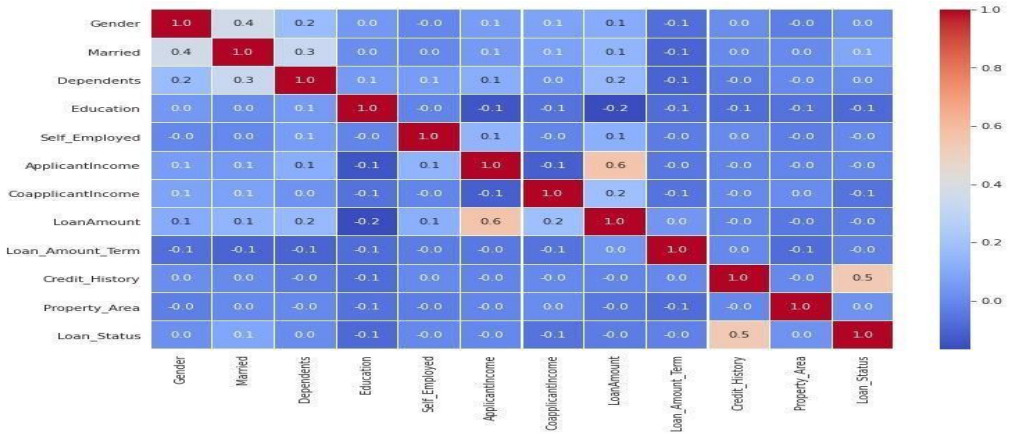

Figure 2: The correlation among all the attributes in the dataset (https://www.kaggle.com/code/vinodkumargr/loan-prediction)

4.3 Machine Learning Models

We employed seven machine learning methods in this study: Linear Regression, Logistic Regression, Support Vector Machine, Naïve Bayes, Decision Tree, K-nearest Neighbour, and Random Forest.

4.3.1 Linear Regression: Linear regression is one of the easiest and drastically used Machine Learning algorithms. It is a manner of performing predictive assessment through the use of statistics. Linear regression generates forecasts for continuous/real/numeric variables (Orji et al., 2022).

4.3.2 Logistic Regression: Logistic regression is one of the Machine Learning algorithms this is maximum often hired within the Supervised Learning category. It is used to forecast the specific based variable and the usage of a special set of impartial variables. Logistic regression is used to be expecting the output for a based variable this is expressed. The final results have to consequently be a discrete or express value. It gives the probabilistic values that lie between zero and 1 in preference to the perfect values between zero and 1. It may be either True or False, zero or 1, or Yes or No. Logistic regression and linear regression are pretty similar,

apart from how they're used. While logistic regression is used to deal with type problems, regression problems are addressed with the aid of using linear regression

4.3.3 Support Vector Machine: The extreme vectors and points that help create the hyperplane are chosen via SVM. The SVM approach is based on support vectors, which are utilized to represent these extreme situations (Bansal et al., 2022).

4.3.4 Naïve Bayes: Naïve Bayes classifiers are a subset of classification algorithms based on the Bayes theorem. Instead of being a single approach, it is a family of algorithms that all work by the same guiding principle: that each pair of features is being classed stand-alone (Orji et al., 2022).

4.3.5 Decision Tree: Classification and regression troubles may be resolved with the usage of the supervised studying approach called a choice tree, but this technique is often preferred. It is a tree-dependent classifier, wherein every leaf node represents the category's final results and internal nodes constitute the functions of a dataset. The nodes in a choice tree are the Decision Node and Leaf Node. Decision nodes are used to make choices and feature many branches, while Leaf nodes are the results of choices and do now no longer have any extra branches (Bansal et al., 2022).

4.3.6 K-nearest Neighbour: The KNN technique locations the brand-new example inside the class that resembles the modern-dayclasses the maximum, presuming that the brand-new case and the preceding instances are comparable. After all of the preceding statistics have been recorded, a brand new statistics factor is labelled the use of the KNN set of rules primarily based totally on similarity. This suggests that with the KNN approach, new statistics can be reliably and unexpectedly labelled. The KNN method may be used for regression even though category issues are wherein its far maximum normally applied (Bansal et al., 2022).

4.3.7 Random Forest: The random forest bases its prediction of the final outcome on the predictions that garnered the most overall votes rather than using just one decision tree's forecast. To increase the projected accuracy of the input dataset, the Random Forest classifier averages the results from multiple decision trees used on various subsets of the input dataset. Higher accuracy and overfitting are prevented by the larger number of trees in the forest.

5. RESULT

In the banking sector, the loan has always been the greatest source of income for a bank but the loan approval process consumes a lot of time if done manually. To automate this whole procedure, we accustomed seven algorithms of machine learning. We assessed the efficacy of seven machine learning techniques in terms of accuracy. Linear regression has the lowest accuracy of just 51%, while the Random Forest classifier outscored all six other models, scoring the greatest accuracy of 90.71%. Table 2 summarizes the overall machine learning models' average accuracy.

Table 2: Accuracy of all the implemented models

Classifier	**Accuracy**
Linear Regression	51.21%
KNN	54.09%
Logistic regression	74.83%
Decision Tree	78.62%
SVM	82.23%
Naïve Bayes	85.96%
Random Forest	***90.71%***

6. FUTURE WORK AND CONCLUSION

This work concluded that the dataset is incomplete and still lacks some feature vectors. No classifier was able to perform better than 90.71% because the subspace of the enter area that we have been seeking to generalize has unknown greater dimensions (Random Forest classifier). More characteristic vectors ought to be produced inside the destiny if similar studies are completed to provide the dataset applied to this observationso that the classifiers can construct higher expertise of the difficulty at hand. For improved accuracy and performance in the future, our work aims to employ a machine learning model with some deep learning techniques like CNN.

The deployment of DL models may also reduce computational time as it requires fewer manual pre-processing tasks.

REFERENCE

Al Mamun, M., Farjana, A., &Mamun, M. (2022): Predicting Bank Loan Eligibility Using Machine Learning Models and Comparison Analysis. 7th North American International Conference on Industrial Engineering and Operations Management, Orlando, Florida, USA, June 12-14.

Awodele, O., Alimi, S., Ogunyolu, O., Solanke, O., Iyawe, S., &Adegbie, F. (2022, November). Cascade of Deep Neural Network And Support Vector Machine for Credit Risk Prediction. In 2022 5th Information Technology for Education and Development (ITED) (pp. 1-8). IEEE.

Bansal, M., Goyal, A., &Choudhary, A. (2022). A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. Decision Analytics Journal, 3, 100071.

Gupta, A., Pant, V., Kumar, S., & Bansal, P. K. (2020, December). Bank Loan Prediction System using Machine Learning. In 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART) (pp. 423-426). IEEE.

https://jupyter.org/ [Last accessed on 15-06-2023]

https://docs.anaconda.com/free/navigator/index.html [Last accessed on 10-05-2023]

https://www.kaggle.com/code/vinodkumargr/loan-prediction [Last accessed on 01-05-2023]

Kadam, A. S., Nikam, S. R., Aher, A. A., Shelke, G. V., &Chandgude, A. S. (2021). Prediction for loan approval using machine learning algorithm. International Research Journal of Engineering and Technology (IRJET), 8(04).

King, T., &Frishberg, I. (2001). Big Loans, Bigger Problems: A Report on the Sticker Shock of Student Loans.

Kumar, C. N., Keerthana, D., Kavitha, M., &Kalyani, M. (2022, June). Customer Loan Eligibility Prediction Using Machine Learning Algorithms in Banking Sector. In 2022 7th International Conference on Communication and Electronics Systems (ICCES) (pp. 1007-1012). IEEE.

Mohankumar, M., Amuthakkani, S., &Jeyamala, G. (2016). Comparative analysis of decision tree algorithms for the prediction of eligibility of a man for availing bank loan. Age, 19, 60.

Nayak, D. S. K., Routray, S. P., Sahooo, S., Sahoo, S. K., & Swarnkar, T. (2022, August). A Comparative Study using Next Generation Sequencing Data and Machine Learning Approach for Crohn's Disease (CD) Identification. In 2022 International Conference on Machine Learning, Computer Systems and Security (MLCSS) (pp. 17-21). IEEE.

Ndayisenga, T. (2021). Bank loan approval prediction using machine learning techniques (Doctoral dissertation).

Orji, U. E., Ugwuishiwu, C. H., Nguemaleu, J. C., &Ugwuanyi, P. N. (2022, April). Machine Learning Models for Predicting Bank Loan Eligibility. In 2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON) (pp. 1-5). IEEE.

Park, M. S., Son, H., Hyun, C., & Hwang, H. J. (2021). Explainability of machine learning models for bankruptcy prediction. IEEE Access, 9, 124887-124899.

Pandey, N., Gupta, R., Uniyal, S., & Kumar, V. (2021). Loan approval prediction using machine learning algorithms approach. International Journal of Innovative Research in Technology, 8(1), 898-902.

Reddy, C. S., Siddiq, A. S., &Jayapandian, N. (2022, June). Machine Learning based Loan Eligibility Prediction using Random Forest Model. In 2022 7th International Conference on Communication and Electronics Systems (ICCES) (pp. 1073-1079). IEEE.

Samreen, A., & Zaidi, F. B. (2012). Design and development of credit scoring model for the commercial banks of Pakistan: Forecasting creditworthiness of individual borrowers. International Journal of Business and Social Science, 3(17).

Sheikh, M. A., Goel, A. K., & Kumar, T. (2020). An approach for prediction of loan approval using machine learning algorithm. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 490-494). IEEE.

Singh, V., Yadav, A., Awasthi, R.,&Partheeban, G. N. (2021). Prediction of modernized loan approval system based on machine learning approach. In 2021 International Conference on Intelligent Technologies (CONIT) (pp. 1-4). IEEE.

Supriya, P., Pavani, M., Saisushma, N., Kumari, N. V., &Vikas, K. (2019). Loan prediction by using machine learning models. International Journal of Engineering and Techniques, 5(2), 144-147.

Srivastava, S., GARG, A., SEHGAL, A., & KUMAR, A. (2018). Analysis and Comparison of Loan Sanction Prediction Model using Python. International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR), 8, 1-8.

Tejaswini, J., Kavya, T. M., Ramya, R. D. N., Triveni, P. S., &Maddumala, V. R. (2020). Accurate loan approval prediction based on machine learning approach. Journal of Engineering Science, 11(4), 523-532.

Wang, C., Yang, N., Xu, W., Wang, J., Sun, J., & Chen, X. (2022, July). Research on a text data preprocessing method suitable for clustering algorithm. In 2022 3rd International Conference on Information Science, Parallel and Distributed Systems (ISPDS) (pp. 340-344). IEEE.

Yuvarani, P., Bharani, P., Dharun, B., & Dinesh, P. (2023, March). Time Series Forecasting of Ethereum Price by FB-Prophet. In 2023 4th International Conference on Signal Processing and Communication (ICSPC) (pp. 272-277). IEEE.