International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Predicting Bank Loan Eligibility Using Machine Learning

Mrs. Y. Swathi[1], Karnati Srilakshmi Pujitha[2],

Neelaboina Siva Jyothi[3], Nandhyala Navya Sri[4], Mogilicharla Poojitha[5]

[1]Associate Professor, Department of Computer Science and Engineering, Tirumala Engineering College
[2,3,4,5]Student, Department of Computer Science and Engineering, Tirumala Engineering College

---

Abstract - With the increasing needs of people, the demand for bank loans also rises. Each day, banks receive numerous loan applications from various customers and individuals, however, not all applicants are approved. Banks typically assess an applicant's eligibility before processing a loan application, a process that can be time-consuming and complex. In evaluating loan applications and making credit decisions, banks often rely on their credit scoring and risk assessment systems. Nevertheless, there are still cases where some applicants default on their payments annually, resulting in significant financial losses for financial institutions.
In this research, we utilized machine learning algorithms to analyze a dataset of approved loans and predict which applicants are most deserving of a loan. The study incorporates customers' historical data, including factors such as age, income type, loan annuity, last credit bureau report, organization type, and length of employment. Various machine learning methods like Random Forest, XGBoost, Adaboost, Lightgbm, Decision tree, and K-Nearest Neighbor were utilized to identify the most influential features that impact prediction outcomes. These algorithms were then compared using standard metrics, with Logistic Regression achieving the highest accuracy rate of 92%.

Key Words: Bank Loans, Logistic Regression, Credit Report, Risk Assessment

1.INTRODUCTION

More and more people are opting to apply for loans online due to the growing amount of data in the financial industry as it transitions to digitalization. Artificial intelligence, commonly known as AI, is gaining popularity for its ability to analyze data effectively. Professionals in different fields are turning to AI algorithms to tackle industry-specific issues. Banks are struggling to handle the high influx of loan applications on a daily basis, leading to potential errors. The loan distribution process is a crucial aspect of a bank's operations, directly influencing its profits. Any errors in this process could result in significant financial setbacks for a bank.

The main aim of the banking industry is to ensure the safety of their funds. Nowadays, banks and financial institutions are offering loans after a thorough verification and validation process. However, there is no certainty that the selected applicant is the most deserving among all applicants. We have developed a method to predict whether an applicant is trustworthy or not, and the validation process is automated using machine learning techniques. This Loan Prediction system benefits both bank employees and applicants (Kumar et al., 2016).

This paper aims to offer a fast, simple, and effective way to choose eligible candidates. It could offer special advantages to the bank. The Loan Prediction System automatically calculates the importance of each factor in loan processing and processes the same factors based on their importance in new test data. Applicants may be given a deadline to find out if their loan will be approved. The Loan Prediction System enables you to quickly assess and prioritize specific applications. This method allows you to prioritize applications that should be approved first. Gender.

Factors such as marriage status, number of dependents, level of education, self-employment status, applicant's income, coapplicant's income, loan amount, loan term, credit history, property area, and loan status play a crucial role in the prediction process. This report is structured in six sections, starting with a review of the literature, followed by an analysis of the dataset. A machine learning approach is proposed in the next section, with a discussion on the algorithms used to construct the model. The findings will be briefly analyzed and discussed, ultimately leading to a conclusion.

2. RELATED WORK

A prediction is a statement about what someone thinks will happen in the future. Predictions are common and can range from serious scientific calculations to simple guesses. They help us anticipate future outcomes, whether it's in the short term, long term, or even over decades. Predictive analytics is a field of advanced analysis that examines current data to make predictions using various techniques like data mining, statistics, modeling, machine learning, and artificial intelligence. In a study by Kumar Arun et al. (2016), they demonstrated how to predict a bank's loan approval using machine learning tools like SVM and neural networks.

Through examining existing research, we were able to conduct our own research and create a reliable model for predicting bank loans. In a study conducted by Mohammad et al. in 2010, they aimed to predict whether a bank would approve a loan for a customer. They utilized Logistic Regression with a sigmoid function to achieve classification in their model. The data for their study was sourced from Kaggle and included both training and testing data sets. To ensure accuracy in their analysis, the data was cleaned to remove any missing values. Performance metrics such as sensitivity and specificity were then used to evaluate and compare the models. The final results show that the model achieved an accuracy rate

of 81%. This performance was slightly improved due to the incorporation of various variables, such as age, purpose, credit history, credit amount, and credit duration, in addition to checking account information. Considering these factors is essential for accurately predicting the likelihood of loan default. By utilizing a logistic regression approach, it becomes easier to identify suitable customers to target for loan approval.

In 2019, Pidikiti and colleagues created a model with the aim of reducing risk when selecting individuals to receive loans, ultimately saving time and money for the bank. The paper consisted of four main sections: data collection, comparing machine learning models with the collected data, training the system with the best model, and testing. Various machine learning algorithms were used to forecast loan data, including classification, logistic regression, Decision Tree, and gradient boosting. The decision tree method proved to be the most accurate among these algorithms, achieving an 82 percent accuracy rate. The AI model achieved success by delivering better outcomes in classifying data. It was highly user-friendly, easy to set up, and offered results that were easy to understand.

In a study by Pandey et al. (2010), it was found that identifying loan defaulters poses a significant challenge for banks. However, by accurately predicting loan defaulters, banks can greatly reduce their losses and minimize non-profit assets. This has made research on predicting loan approvals essential. Machine learning techniques play a crucial role in analyzing this type of data. The study utilized four classification-based machine learning algorithms - Logistic Regression, Decision Tree, Support Vector Machine, and Random Forest. Among these, the Support Vector Machine method proved to be the most accurate, achieving a high accuracy rate of 79.67% in predicting loan acceptance. A list (dataset) of previous clients' information was compiled from various banks that had supported a range of cutting-edge developments.

In their study, Ndayisenga et al. (2021) partnered with commercial banks to analyze borrower behavior by creating and evaluating various models using data from Bank of Kigali. The data was split into training and test sets, with training data making up 70% and test data making up 30% of the total. Through the use of ensembles, they identified the most effective machine learning techniques for predicting bank loan defaults. Gradient Boosting proved to be the top model for predicting loan defaults, boasting an accuracy of 80.40%. XGBoosting also showed promise, while decision trees, random forest, and logistic regression performed poorly in comparison. In 2018, Shrishti and her team introduced a reliable machine learning system designed to forecast loan approval quickly. Their primary objective was to expedite the loan approval process for applicants. They employed three different machine algorithms: Logistic Regression, Decision Tree, and Random Forest. Upon analyzing the dataset for each model, they found that the Random Forest algorithm boasted the highest accuracy among all models.

3. METHODOLOGY

3.1. Data Collection:

The initial stage in the proposed methodology involves gathering data, followed by data pre-processing. Various classifiers including XGBoost, AdaBoost, LighGBM, Random Forest, Decision Tree, and K-Nearest Neighbor are selected and trained using the standard hold-out approach on the dataset. The results are then analyzed to determine the most effective method for predicting Bank Loan eligibility.

a. Dataset Used:

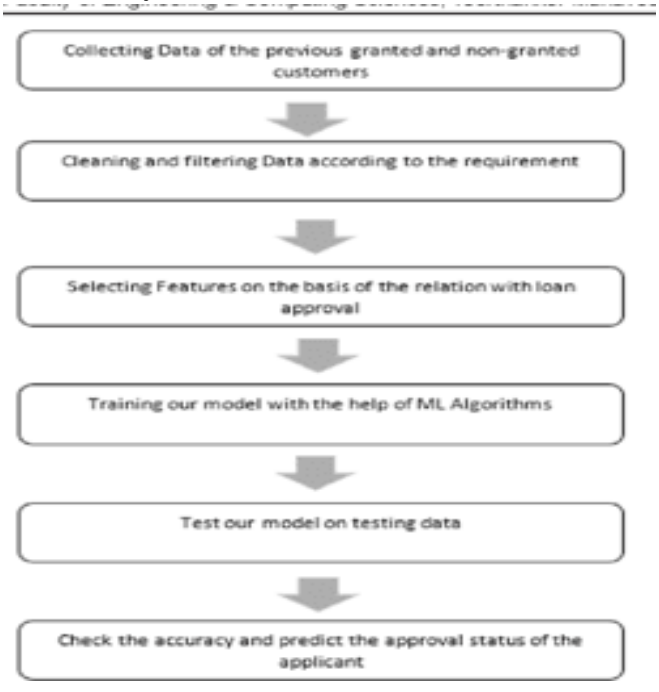
In this study, the dataset used was gathered from the Kaggle online platform. The dataset consists of 10,128 instances with 23 attributes, of which 1 is the class attribute and the remaining 23 are predictive attributes. The goal was to accurately predict Bank Loan eligibility using these attributes, which include information such as age, gender, education, ownership, financial status, income sources, credit card details, and more. The class attribute relates to the prediction of bank loan eligibility.

3.2. Data Preprocessing

Data preprocessing for the dataset involved performing tasks such as extracting features, cleaning the data, handling missing values, and transforming categorical variables.

3.3. Validation

Validation is an important step in the process of assessing a dataset. The hold-out validation method is a commonly used approach for obtaining reliable results. In our study, we implemented a hold-out validation process where 70% of the data was used for training and 30% for testing. Through this validation process, we evaluated the performance of different machine learning techniques using metrics such as accuracy, precision, recall, area under the curve (AUC), and F1-Score, as determined by the confusion matrix.



4. ALGORITHMS USED:

- **Decision tre-es:**
  Decision trees make choices based on many factors. Say a bank wants to predict if a loan will get approved. The tree looks at age, income, and credit score. Simple rules decide the outcome.

- **Support Vector Machines (SVMs):**
  Support Vector Machines (SVMs) find the ideal dividing line between two groups. For loans, it may predict if a borrower will repay. Features include salary, credit history, and loan amount. The line separates payers from non-payers.

- **Random Forest:**
  Random forests combine many small decision trees. Their predictions are averaged for accuracy. To guide loan repayment, lenders may use income level, credit records, and loan size. Multiple simple models work together.

- **Naive Bayes:** This method calculates the danger of a characteristic set fitting a specific elegance using Bayes' theorem. For instance, to forecast mortgage approval in banks using information like gender, marital reputation, income, and loan size.

- **K-Nearest Neighbors (KNN):** This approach categorizes a new data factor primarily based on close-by factors. For instance, in bank loan predictions, the use of details like income, credit document, and mortgage length.

- **Gradient Boosting:** This technique creates a chain of weak models to shape a robust version. For example, in financial institution loan predictions, it could forecast mortgage reimbursement using elements like income, credit score document, and loan size.

5. FUTURE IMPROVEMENTS

There are many potential future works that are improving bank loan prediction by using machine learning. The first potential work is using deep learning models such as neural networks that can be used to predict loans using a large and complex data set. Since deep learning models can learn and extract features automatically instead of predefined features, the accuracy and performance are improved significantly. Furthermore, deep learning can find non-linear relationships and high-dimensional data perfectly, and in contrast, conventional machine learning is not good at it.

Moreover, ensemble learning techniques might be utilized as an option. This solution implies using various machine learning models for the accurate bank loan prediction. By reducing variance and bias, this approach can minimize the error of individual models, enhancing overall performance. Some of the techniques include bagging, boosting, and stacking along with others.

Use explainable AI to better understand the model. SHAP and LIME as the methods of explainable AI could be used to provide such properties. Such properties will lead to the formation of trust in the process of decision-making from the model's side and enhance its transparency; therefore, it can be used safely without any subjectivity or interpretability; furthermore, it can also help to catch some errors and fight with them. By using such processes as stream processing and event driven models that data might be processed in the real-time regime.

By leveraging AI technology, banks can streamline loan processing and enhance customer satisfaction. Real-time data analysis enables quicker and more precise decision-making, minimizing loan default risks and enhancing portfolio performance. Techniques like data synthesis and transfer learning address data scarcity and imbalance in loan prediction, boosting training data quality and machine learning model effectiveness.

Combining different types of data, like merging and integrating information, can be beneficial for predicting bank loans. This approach can enhance the accuracy and efficiency of machine learning models by including a wider range of data. Additionally, it can minimize bias, and enhance fairness and transparency in loan prediction systems.

Continuous monitoring and evaluation methods, like A/B testing and model monitoring, can assess the performance of machine learning models and confirm they are current and efficient. By constantly monitoring and evaluating, potential issues and errors in the models can be identified and corrected, enhancing the accuracy and reliability of predictions. These processes also ensure that the loan prediction system aligns with business objectives and regulatory standards.

6. CONCLUSION

To sum up, the system you created helps make the loan application process easier by allowing users to submit information directly on the homepage. The data is then sent to the data layer for analysis. The Random Forest algorithm is used to analyze the data and predict whether a consumer will be approved for a loan. This algorithm has shown to be very accurate in predicting loan approval ratings, making it a reliable tool for making decisions on loan approvals.

Moreover, the system takes a thorough risk assessment approach that considers various factors when determining the risk of a home loan, minimizing the likelihood of mistakes in decision-making. This method improves the precision of loan approval and offers a more dependable way to evaluate the creditworthiness of borrowers.

Additionally, the Random Forest algorithm is highly flexible and scalable, making it perfect for handling large amounts of data. This allows the system to adjust to different data inputs and improve the efficiency of loan approval processes. Its

capacity to manage intricate information and deliver precise forecasts is a crucial advantage in the current financial sector, where data-guided decision-making is becoming more crucial. To summarize, the application of the Random Forest algorithm in the system for predicting loan approvals provides a strong and dependable approach for evaluating loan requests. This method produces precise predictions and reduces mistakes in the decision-making phase. The system's adaptability and expandability contribute to its value as a tool for streamlining loan approval procedures and improving the precision of credit evaluations.

7. ACKNOWLEDGEMENT

We want to express our gratitude to Mrs. Y. Swathi for her amazing guidance and support throughout our project. Her knowledge and motivation were crucial to our achievements, and we are thankful for her commitment. We would also like to extend our appreciation to the professors in the Computer Science and Engineering Department at Tirumala Engineering College for giving us the opportunity to participate in this important research project, which has been a valuable learning experience for us.

8. REFERENCES

1. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
2. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).
3. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.
4. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.
5. Verma, P., & Srivastava, S. (2020). Loan eligibility prediction using machine learning algorithms. In 2020 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI) (pp. 401-405). IEEE.
6. Zhang, H., & Patel, V. M. (2018). Deep learning for remote sensing data: A technical tutorial on the state of the art. IEEE Geoscience and Remote Sensing Magazine, 5(2), 8-36.
7. Brownlee, J. (2021). Introduction to machine learning with Python: A guide for data scientists. Machine Learning Mastery.
8. Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. Informatica, 31(3), 249-268.
9. Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, 23(4), 589-609.
10. Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.
11. Hand, D. J., & Yu, K. (2001). Idiot's Bayes—not so stupid after all? International Statistical Review, 69(3), 385-398.
12. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). Springer.
13. Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.
14. Paudyal, D., & Bohara, A. K. (2020). A comparative study of machine learning techniques for loan default prediction. Journal of Big Data, 7(1), 1-25.
15. Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81-106.
16. Ravi, V., & Ravi, V. (2017). A survey on opinion mining and sentiment analysis: Tasks, approaches, and applications. Knowledge-Based Systems, 89, 14-46.
17. Tan, P. N., Steinbach, M., & Kumar, V. (2006). Introduction to data mining. Pearson Addison Wesley.
18. Witten, I. H., Frank, E., & Hall, M. A. (2011). Data mining: Practical machine learning tools and techniques. Morgan Kaufmann.
19. Gupta, Anshika, et al. "Bank Loan Prediction System using Machine Learning." 2020 9th International Conference
20. https://www.semanticscholar.org/paper/Bank-Loan-Prediction-System-using-Machine-Learning-Gupta-Pant/e11f3a848c1d99564db6b536f108cba05cc17a9b
21. System Modeling and Advancement in Research Trends (SMART). IEEE, 2020.
22. Algorithm," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 490-494, 2020.