Loan Eligibility Prediction Using Machine Learning Models with SMOTE for Class Imbalance

Mohammod Tareq Aziz Justice
Dept of
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
mohammod.tareq.aziz@g.bracu.ac.bd

Sanjana Afroz Troyee
Dept of
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
sanjana.afroz.troyee@g.bracu.ac.bd

Tasfique Zaman Chowdhury Sifat
Dept of
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
tasfique.jaman.chwodhury@g.bracu.ac.bd

***Abstract*—Prediction of loan eligibility is an important activity in the banking industry because it assists banks in reducing risk and also providing equitable credit services. This article discusses how machine learning can be used to predict who is eligible to take out a loan, and the inefficiency of conventional methods of credit scoring. The research uses different machine learning algorithms such as Random Forest, Logistic Regression, AdaBoost, K-Nearest Neighbors, and Multilayer Perceptron on a loan eligibility dataset that was given by Dream Housing Finance Company. The models are trained and tested in terms of accuracy, precision, recall and F1 score. Data preprocessing included the treatment of missing data by mean and mode imputation, categorical data coding, elimination of outliers, and normalization. SMOTE was employed to deal with class imbalance. Findings show that the best predictions are offered by the Random Forest and MLP models with the former having the accuracy of 89.29 and the latter recording the highest recall among the eligible applicants.**

***Index Terms*—Machine Learning, SMOTE, Ensemble Methods, Random Forest, Logistic Regression, AdaBoost, K-Nearest Neighbors, Multilayer Perceptron.**

I. Introduction

The accelerating pace of the development of financial institutions into data-driven decision-making processes has highlighted the critical importance of machine learning (ML) as a tool to improve loan and credit prediction algorithms. Conventional credit scoring systems, which in many ways are constrained by their strictness in nature, have failed to cope with the dynamics of the current lending world. This paper discusses how ML approaches can be used to predict the eligibility of loans using a dataset offered by the Dream Housing Finance company. The research emphasizes the replacement of traditional classifiers by Random Forest, AdaBoost, and XGBoost that are more accurate and robust to imbalanced data. Based on preprocessing methods, such as SMOTE to balance classes and feature engineering, this study will be able to create a powerful model capable of automating the process of loan eligibility evaluation. The research presents several ML models such as Random Forest, Logistic Regression, AdaBoost, K-Nearest Neighbors, and Multilayer Perceptron and compared their effectiveness based on such metrics as accuracy, precision, recall, and F1-score. This practice aims at determining the most efficient model, where there are gaps in data representativeness and ethical issues, to enhance lending decisions and lower the default rates.

II. Literature Review

The continued shift to data-driven systems by financial institutions has led to a wide range of studies on the use of machine learning in loan and credit prediction. In all the works reviewed, the authors repeatedly highlight the weaknesses of conventional credit scoring systems and the opportunities of ML to increase the quality of predictions, decrease default rates, and simplify decision-making.

The first works on the prediction of loan eligibility concentrated on the use of simple classifiers on small, publicly accessible datasets. For instance, Zhang [3] analyzed the differences between urban, rural, and semi-urban areas with the help of the Logistic Regression approach, achieving 83.78% accuracy, and revealed how the geographic area influences creditworthiness. Similarly, Manglani and Bokhare [7] used Logistic Regression on a publicly available dataset and obtained 79.45% accuracy with a high recall (96.6%). Both papers emphasized the interpretability of logistic models, which are more frequently used by banks due to their transparency. Deepa et al. [6] further compared Logistic Regression with Random Forest and reported that Random Forest performed slightly better (86% versus 85%). Likewise, Tumuluru et al. [9] evaluated Logistic Regression, SVM, KNN, and Random Forest, concluding that Random Forest achieved the best accuracy (81%), though they noted that relying only on overall accuracy may neglect fairness or bias issues.

As research matured, scholars broadened the scope of algorithms and datasets. Haque and Hassan [1] tested AdaBoost, GaussianNB, Random Forest, Decision Tree, and SVM on a massive Kaggle dataset of more than 148,000 entries, finding that AdaBoost performed almost flawlessly (99.99%), illustrating the strength of ensemble methods in capturing complex interactions. Meenaakumari et al. [2] explored health loan eligibility prediction using personal characteristics (e.g., BMI, smoking, region), where Random Forest again outperformed others with 91% accuracy. Naveen Kumar et al. [5] complemented these findings by empirically testing AdaBoost with Decision Trees, Random Forest, SVM, and KNN on

614 records, where AdaBoost once more yielded the highest accuracy (84%), reinforcing the superiority of ensemble techniques. Altogether, these studies demonstrate the field's transition from single-model testing toward ensemble-based approaches, though concerns regarding data representativeness remain significant.

To address the persistent issue of class imbalance, ensemble learning and hybrid approaches were increasingly adopted. Orji et al. [10] compared six algorithms (Random Forest, Gradient Boost, Decision Tree, SVM, KNN, Logistic Regression) while applying SMOTE, finding Random Forest to be the most accurate (95.55%). Similarly, Muhammad et al. [12] combined multiple classifiers such as XGBoost and LightGBM with SMOTE and one-hot encoding, reporting that XGBoost surpassed others with more than 95% accuracy. Lakshmi Narasimha et al. [4] went further by proposing a hybrid pipeline for home and education loans that integrated supervised learning, clustering, and anomaly detection, highlighting a multidimensional approach to loan prediction.

Beyond traditional loan approval, other works have extended ML applications to credit card defaults and special lending contexts. B. D. S. et al. [8] analyzed a dataset of 30,000 Taiwanese credit card users to forecast defaults, where ensemble models consistently outperformed individual classifiers such as Logistic Regression, Random Forest, SVM, KNN, and Gradient Boosting. This line of research not only demonstrated the importance of model choice but also underscored the role of feature engineering and ethical considerations in applying ML to sensitive financial areas.

An important methodological evolution can also be observed in evaluation metrics. While many early studies (e.g., Zhang [3], Manglani and Bokhare [7]) relied solely on accuracy, more recent works (e.g., Orji et al. [10], Muhammad et al. [12]) incorporated metrics such as precision, recall, F1-score, ROC-AUC, and confusion matrices. This progression reflects a growing recognition that accuracy alone is insufficient, particularly in the presence of imbalanced datasets where false negatives may have severe financial consequences.

In summary, early literature confirmed the practicality of basic classifiers such as Logistic Regression due to their interpretability, whereas later studies repeatedly demonstrated the superior predictive performance of ensemble approaches including Random Forest, AdaBoost, Gradient Boost, and XGBoost (Haque and Hassan [1]; Meenaakumari et al. [2]; Orji et al. [10]; Muhammad et al. [12]). Despite these advances, considerable gaps remain. Future work could focus on incorporating alternative signals such as transaction history, social behaviors, and macroeconomic indicators to enhance robustness, while simultaneously addressing critical concerns of data fairness, privacy, and transparency.

III. Methodology

A. Dataset Description

The data offered by Dream Housing Finance company holds data on the applicants to home loan. It is aimed at assisting in the automatization of the loan eligibility process on the basis of the information the applicants fill in their online application forms. These features include the applicants gender, marital status, level of education, number of dependants, income, loan amount, credit history and the area of the property. This information is useful in determining the eligibility of a person with regard to a home loan. The dependent variable in the data is LoanStatus that shows whether the loan has been granted (Yes, N).

Every row in the data sets is a single loan applicant and the columns provide the personal data and the loans specifics of the applicants. In the columns, such as ApplicantIncome and CoapplicantIncome, one can find the income of the applicant and Coapplicant, and in the LoanAmount one can find the amount asked by the applicant. Others like creditHistory and property area give details of the financial background of the applicant and the area of the property location. This information will be useful in identifying the segments of customers more likely to be sanctioned to take loans, thus the company can focus on these customers in future marketing and services.

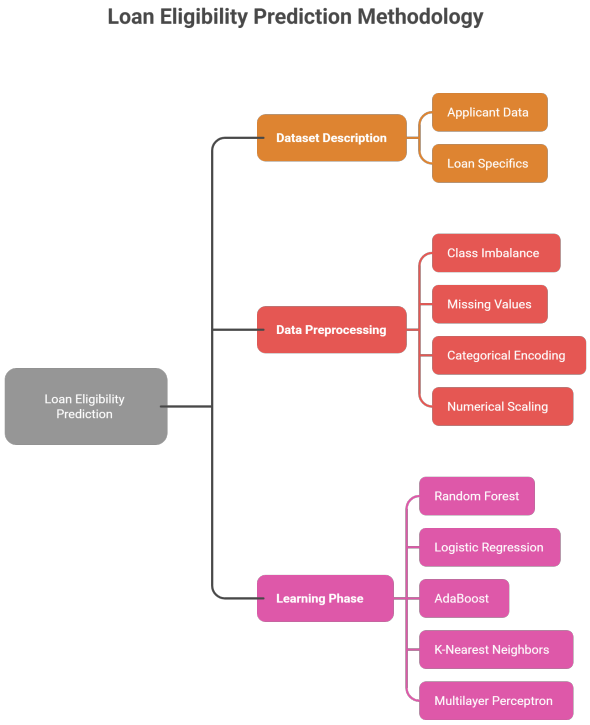

Fig. 1. Methodology At a Glance

B. Data Preprocessing

The issue of class imbalance, where a single class (such as 'LoanStatus = Yes') is much underrepresented compared to the other (such as 'LoanStatus = No') is an important step in the dataset preprocessing. This has a potential impact on machine learning model performance. The SMOTE technique is employed to solve this. It operates by creating artificial specimens

of the minority group and not just by copying them. It achieves this through the process of picking examples which are in close proximity to each other in the feature space, and drawing a line between them and generating synthetic examples along. This makes the distribution of classes balanced and the model does not get skewed in favor of the major class.

Once the dataset is loaded and missing values are dealt by replacing missing values in the LoanAmount column with the median, two things are left to do: first, the categorical variables such as Gender, Married, and Education have to be encoded. This is carried out through one-hot encoding. In the case of numerically defined features such as ApplicantIncome, CoapplicantIncome and LoanAmount, scaling methods like normalization are used to reduce all values to one comparable range.

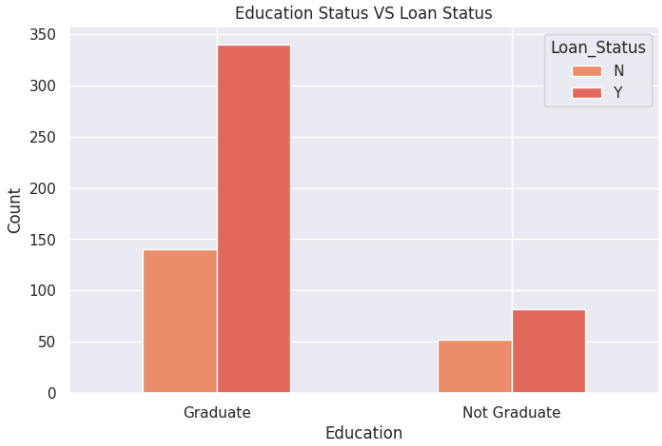

Fig. 2. Imbalanced Education Class

SMOTE is implemented on the training data after partitioning the dataset into training and testing sets. This will guarantee that the model will be trained on a balanced dataset that can ensure better prediction of both classes.

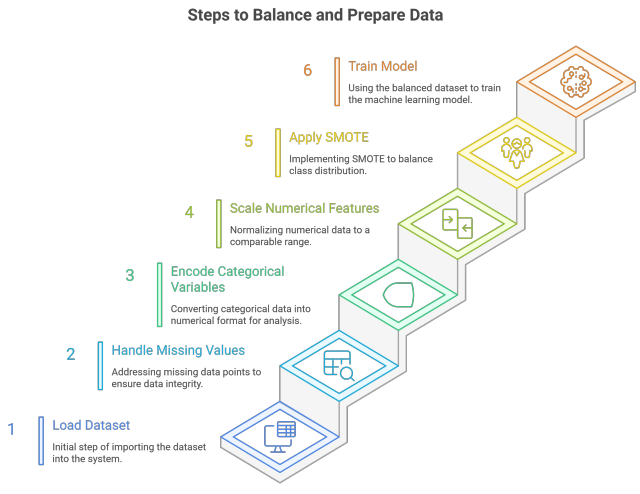

Fig. 3. Data Preprocessing Steps

C. Learning Phase

During the learning stage, various machine learning models were used to extend loan eligibility on different attributes of the applicants. The data were preprocessed and trained on these data and then tested on the test set with the aim of classifying applicants as eligible (Y) or ineligible (N) to take a loan.

1) Random Forest (RF): Random Forest represents an ensemble learning algorithm, which builds several decision trees and aggregates their outputs to enhance prediction accuracy. The model in this is constructed by randomly selecting a subset of the data to construct each tree and the overall prediction is made by summing all the outputs of these trees. This method assists in the minimization of over-fitting and enhances the strength of the model and is especially useful with high-dimensional and complicated data like loan approval.

2) Logistic Regression (LR): The popular statistical method of binary classification is the Logistic Regression. It has been used in this project in an attempt to predict the probability of success of a loan given different characteristics including income and credit history and education. The model employs the use of a logistic, thus generating results between 0 and 1, which may be described as loan-approval probability. It is a useful instrument in the interpretation of the relationship between features because it is very simple and easy to interpret.

3) AdaBoost (Adaptive Boosting): AdaBoost AdaBoost is an ensemble method of learning that integrates several weak learners together to form a strong classifier. AdaBoost was applied in this project with 100 estimators being trained on the dataset in iterative mode. The same method enhances the model in terms of classifying challenging cases. AdaBoost proves to be valuable when the base model is weak and also when one wants to achieve more with simple models.

4) K-Nearest Neighbors (KNN): K-Nearest Neighbors algorithm is one of the non-parametric algorithms that classify data points as belonging to majority one of their nearest neighbors. The KNN model in this instance categorized applicants against the characteristics of the nearest data points in the value of k. KNN is most effective in the situation that non-linear relationship between data needs to be captured and when the boundaries of the decision between classes are complicated.

5) Multilayer Perceptron (MLP): A Multilayer Perceptron is a form of networks with neurons in more than one layer. MLP uses the backpropagation to modify the network weights in the process of training in order to learn intricate patterns in the data. The model can model non-linear relationships and thus it is very effective in classification such as loan eligibility. The capability of MLP in learning complex patterns in great quantities of data makes it an effective tool in precise predictions.

These models were trained and evaluated to determine the best performing model for the prediction of loan eligibility.

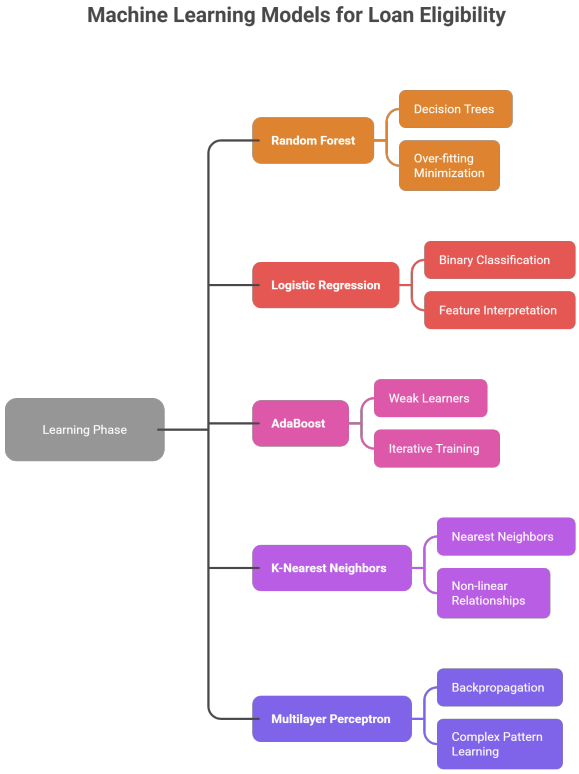

Fig. 4. Models Trained

IV. Results and Discussion

We present the findings of training and testing several machine learning models to estimate loan eligibility. Five models trained on the given dataset are the Random Forest, Logistic Regression, AdaBoost, K-Nearest Neighbors, and Multilayer Perceptron. All the models were analyzed according to the accuracy, precision, recall, and F1 score and the outcomes were compared to identify the most productive model to predict loan eligibility.

A. Feature Importance

We used correlation matrix as a means to understand the relations between the features in the dataset since it reveals the strength of association between various variables. Values of -1 are deemed as negative correlations and values near to 1 are thought to be positive correlations. Based on the matrix, ApplicantIncome is positively correlated to LoanAmount (0.57), implying that, the higher the income, the more the applicant is approved to take a larger loan. LoanAmount has almost no correlation with CreditHistory (-0.0084) indicating that good credit history might have negligible impact on loan approval. LoanAmount and CoapplicantIncome has a moderate correlation with (0.19) meaning that increasing the amount of coapplicant income may result in higher loan amounts.

Conversely, such features as Gender, Married and SelfEmployed had low correlations with LoanStatus, which means

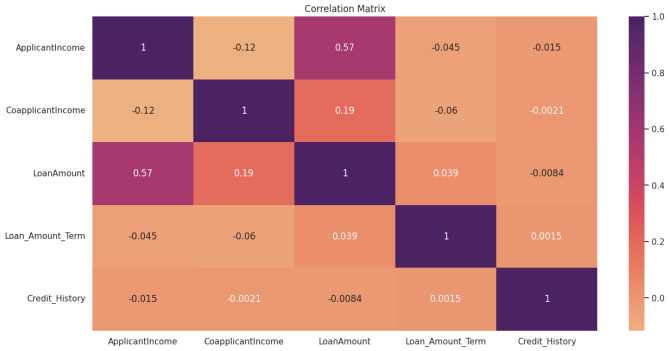

Fig. 5. Heatmap of the Dataset

that they contribute less to loan eligibility. All in all, ApplicantIncome, LoanAmount, and CreditHistory are the most significant financial characteristics to predict loan eligibility.

B. Performance Metrics

The models were compared according to various important metrics, which are accuracy, precision, recall and F1 score. Random Forest was the best in the accuracy of 89.29% with great success in identifying eligible applicants. It also scored high in F1 score of 0.94 on the class of True implying that it was effective in lending in predicting the loan approval. Multilayer Perceptron (MLP) was also a good participant, with a 89.29% accuracy and an excellent recall of 0.96 in the True class demonstrating that it can recognize loan-qualified applicants. But it performed poorly with the False class with 0.00 precision and recall, which means that it has implications in recognizing ineligible applicants.

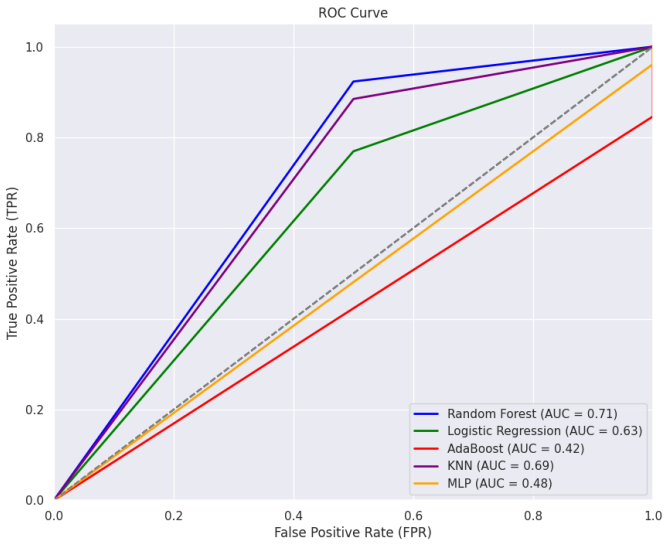

Fig. 6. ROC Curves of the Models

KNN followed with an accuracy of 85.71% with a high-precision of 0.96 of the True class and low-recall of 0.50 of the False class followed. AdaBoost had an accuracy of 78.57% and had poor accuracy and recall in particular in the False

group where the model did not predict any negative cases, which gave it a precision and recall of 0.00 on ineligible applicants. The weakest performance was registered by Logistic Regression (accuracy 75%). Although it did reasonably well on predicting the "True" class, its precision and recall on the "False" class was very low, indicating it had a problem predicting ineligible applicants. On balance, the best in terms of accuracy was the Random Forest and MLP, but the models should be refined to work with imbalanced classes to predict the two classes.

TABLE I
PERFORMANCE TABLE

Model	Accuracy	Precision (F)	Precision (T)	Recall (F)	Recall (T)
Random Forest	0.89	0.33	0.96	0.50	0.92
Logistic Regression	0.75	0.14	0.95	0.50	0.77
AdaBoost	0.79	0.00	0.92	0.00	0.85
KNN	0.86	0.25	0.96	0.50	0.88
MLP	0.89	0.00	0.93	0.00	0.96

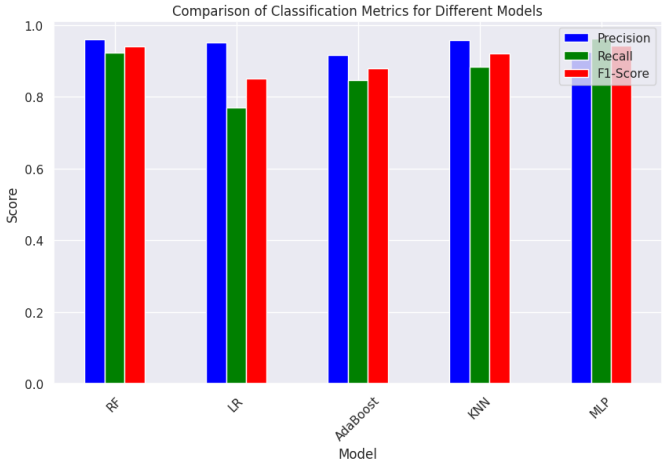

Fig. 7. Classification Metrics of Different Models

V. CONCLUSION

To sum up, this research shows the effectiveness of machine learning in the prediction of the eligibility of loans where Random Forest and Multilayer Perceptron came out as the leading models with an accuracy of 89.29. The imbalance in classes was resolved successfully using SMOTE and made the model more robust and, the correlation matrix indicated that ApplicantIncome, LoanAmount, and CreditHistory are the most important predictors. Nevertheless, there are issues, especially in the accurate recognition of ineligible applicants, which are shown by low precision and recall in the False class of most models. The results are congruent to the literature tendencies preferring the ensemble techniques to the conventional classifiers such as the Logistic Regression that scored the lowest at 75% precision. Further studies need to incorporate other sources of data, e.g. transactions and macroeconomic indicators, yet maintaining privacy. Optimising models to deal with imbalance classes and include measures of fairness will further improve predictive performance and adherence to ethics. Finally, this piece of work is a basis through which Dream Housing Finance can automate the process of loan approvals to enhance efficiency and minimize risks associated with funding in a data-intensive lending environment.

REFERENCES

[1] Haque, and M. M. Hassan, "Bank Loan Prediction Using Machine Learning Techniques," arXiv (Cornell University), Oct. 2024, doi: https://doi.org/10.48550/arxiv.2410.08886.

[2] M. Meenaakumari, P. Jayasuriya, N. Dhanraj, S. Sharma, G. Manoharan, and M. Tiwari, "Loan Eligibility Prediction using Machine Learning based on Personal Information," IEEE Xplore, Dec. 01, 2022, doi: https://doi.org/10.1109/IC3I56241.2022.10073318

[3] Z. Zhang, "Loan Eligibility Prediction: An Analysis of Feature Relationships and Regional Variations in Urban, Rural, and Semi-Urban Settings," Highlights in business, economics and management, vol. 21, pp. 688–697, Dec. 2023, doi: https://doi.org/10.54097/hbem.v21i.14739.

[4] T.Lakshmi Narasimha, T.V.S Chandra Rao, P.S Yashwanth Roy, Dr.A.Vinoth Kumar, and Dr.T.Kumanan, "Machine Learning For Bank Loan Eligibility Prediction:Focus on Home Loan and Education Loan," International Journal on Science and Technology, vol. 16, no. 1, Mar. 2025, doi: https://doi.org/10.71097/ijsat.v16.i1.2810.

[5] Ch. Naveen Kumar, D. Keerthana, M. Kavitha, and M. Kalyani, "Customer Loan Eligibility Prediction using Machine Learning Algorithms in Banking Sector," 2022 7th International Conference on Communication and Electronics Systems (ICCES), Jun. 2022, doi: https://doi.org/10.1109/icces54183.2022.9835725.

[6] M Deepa, S. Pal, and Vaishnavi Prashant Ghusey, "Monetary Loan Eligibility Prediction using Logistic Regression Algorithm," Feb. 2024, doi: https://doi.org/10.1109/ic-etite58242.2024.10493584.

[7] R. Manglani and A. Bokhare, "Logistic Regression Model for Loan Prediction: A Machine Learning Approach," 2021 Emerging Trends in Industry 4.0 (ETI 4.0), May 2021, doi: https://doi.org/10.1109/eti4.051663.2021.9619201.

[8] D. B. S, V. Kumar, Ashwini Kodipalli, and T. Rao, "Default credit card scoring using ML," pp. 44–48, Apr. 2024, doi: https://doi.org/10.1109/ciscsd63381.2024.00022.

[9] P. Tumuluru, L. R. Burra, M. Loukya, S. Bhavana, H. M. H. CSaiBaba, and N. Sunanda, "Comparative Analysis of Customer Loan Approval Prediction using Machine Learning Algorithms," 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), Feb. 2022, doi: https://doi.org/10.1109/icais53314.2022.9742800.

[10] Ugochukwu. E. Orji, Chikodili. H. Ugwuishiwu, Joseph. C. N. Nguemaleu, and Peace. N. Ugwuanyi, "Machine Learning Models for Predicting Bank Loan Eligibility," 2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON), Apr. 2022, doi: https://doi.org/10.1109/nigercon54645.2022.9803172.

[11] P. B. R, A. K, A. Kumar, B. Rao, P. S. K, and S. A. P, "An Approach to Predict Loan Eligibility using Machine Learning," IEEE Xplore, Dec. 01, 2022, doi: https://doi.org/10.1109/AIDE57180.2022.10059881

[12] F. Muhammad, J. C. Halim, H. Lucky, and Derwin Suhartono, "Loan Eligibility Prediction Using Ensemble Machine Learning Techniques and SMOTE," pp. 102–107, Oct. 2024, doi: https://doi.org/10.1109/icset63729.2024.10775270.