Received: 9 April 2023 | Revised: 23 May 2023 | Accepted: 31 May 2023

DOI: 10.1002/eng2.12707

RESEARCH ARTICLE

Engineering Reports Open Access WILEY

Analyzing the impact of loan features on bank loan prediction using Random Forest algorithm

Debabrata Dansana[1] | **S Gopal Krishna Patro**[2] | **Brojo Kishore Mishra**[3] | **Vivek Prasad**[4] | **Abdul Razak**[5] | **Anteneh Wogasso Wodajo**[6]

[1]Department of Computer Science, Rajendra University, Balangir, Odisha, India

[2]Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India

[3]Department of CSE, GIET University, Gunupur, Odisha, India

[4]Department of CSE, Nirma University, Ahmedabad, India

[5]Department of Mechanical Engineering, P. A. College of Engineering (Affiliated to Visvesvaraya Technological University), Belagavi, Mangaluru, India

[6]Department of Automotive Engineering, College of Engineering and Technology, Dilla University, Dilla, Ethiopia

Correspondence
S Gopal Krishna Patro, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India.
Email: sgkpatro2008@gmail.com
Abdul Razak, Department of Mechanical Engineering, P. A. College of Engineering (Affiliated to Visvesvaraya Technological University), Belagavi, Mangaluru 574153, India.
Email: arkmech9@gmail.com
Anteneh Wogasso Wodajo, Department of Automotive Engineering, College of Engineering and Technology, Dilla University, Dilla, Ethiopia.
Email: wogasso.anteneh@gmail.com

Abstract

Loans are a crucial source of income for the financial sector, but they also come with significant financial risks. The interest on loans constitutes a significant portion of a bank's assets. The demand for loans is growing worldwide, and organizations are devising efficient business strategies to attract more clients. Every day, a large number of people apply for loans for various reasons, but not all of them can be approved due to the risk of loan default. It is not uncommon for people to default on their loans, causing significant losses to banks. The purpose of this article is to determine whether to grant loans to specific individuals or organizations. The Random Forest Regressor model has been utilized to measure performance and identify suitable customers for loan approval. The model suggests that banks should not only target affluent clients but also consider other customer characteristics that are critical in credit granting and predicting loan default. The research examines various loan approval parameters such as gender, educational qualification, employment type, business type, loan term, and marital status. Additionally, the study analyzes the number of approved, drawn, and rejected loans, which provides valuable insights into loan approval and prediction.

KEYWORDS

bank loan approval analysis, financial risk, loan prediction, machine learning, Random Forest regressor

1 | INTRODUCTION

Banking is a regulated industry in most countries due to its crucial role in ensuring a country's financial stability. Loans are a primary activity for most banks, and the profits made from interest on loans make up a significant portion of their assets. However, the loan approval process is currently a time-consuming and error-prone manual procedure that relies on individual bank managers to assess applicants' eligibility and risk of loan default. Loan defaults can result in significant losses for banks and even lead to banking collapses that affect the economy. Therefore, the objective of this article is to explore the use of machine learning approach in the loan taking process, particularly the Random Forest Regressor, to accurately identify eligible loan applicants and reduce credit risk. The classification model can predict whether a loan will be granted or not, providing a rapid and straightforward method for selecting meritorious applicants that offers the bank specific benefits, such as increased customer satisfaction and reduced operating expenses. The study includes a visual analysis of the factors that influence loan acceptance, which can inform the development of the ML model. Overall, the paper aims to provide a reliable and efficient loan approval process that reduces the risk of loan default and ensures the financial stability of the banking industry. At first, we bank loan acceptance analysis and a visual analysis of the elements has been performed that influence, enhance, or decrease a person's chances of obtaining a bank loan. Then classification ML model is applied that can predict whether a loan would be granted or not. It is a binary classification problem in which we must forecast either of the two classes given, that is, granted 1 or not granted 0.

2 | LITERATURE SURVEY

This section summarizes previous efforts in developing machine learning and deep learning models with various algorithms to enhance loan prediction processes and assist banking authorities and financial firms in selecting low-risk, qualified applicants. The author tested multiple machine learning algorithms on a dataset to determine which were best suited for analyzing bank credit data. Results indicated that, aside from Gaussian Naive Bayes and Nearest Centroid, the remaining algorithms performed well in terms of accuracy and other performance measures, ranging from 76% to over 80%.[1] The study also identified critical factors affecting customer trustworthiness and developed a predictive model using linear regression and significant characteristics. Additionally, the author employed a logistic regression technique to identify suitable loan recipients by analyzing their risk of default and other consumer qualities.[2] Another study implemented decision trees to construct and assess models for loan prediction, achieving an 81% accuracy rate on a public test set.[3] Finally, a comparative analysis of decision trees and random forest algorithms on the same dataset showed that random forest had significantly higher accuracy, achieving 80% compared to decision trees' 73%.[4]

Article[5] performed experiments using the C4.5 algorithm in decision trees, and the highest accuracy achieved was 78.08% with a 90:10 data partition, while the highest recall value was 96.4% with an 80:20 data partition. Consequently, the 80:20 partition was determined to be the best due to its high accuracy and recall values. In a separate study, the authors in Reference 6 conducted an exploratory data analysis to categorize and investigate the characteristics of loan applicants. They plotted seven distinct graphs, and the analyses showed that most loan applicants chose short-term loans[7] found that Support Vector Machines outperformed other models such as logistic regression and random forest in their comparative performance evaluation. They created decision tree, SVM, AdaBoost, Bagging, and Random Forest models and compared their prediction accuracy to a Logistic Regression model benchmark. The results indicated that AdaBoost and Random Forest performed better than other models, while SVM models performed poorly when using both linear and nonlinear kernels. Overall, these findings suggest that there is potential for businesses to develop default prediction models by experimenting with machine learning approaches.[8] This study focuses on using machine and deep learning models with real-world data to estimate loan default probability. The most significant features from multiple models are selected and used to compare the performance of Random Forest and decision tree classifiers. The author achieved a 78.64% effectiveness with the Random Forest classifier using parameter tuning, which is comparable to the decision tree classifier's prediction efficiency of 85.3%.[9] The study also applies the Random Forest method to create a loan default prediction model using customer loan data from Lending Club. The SMOTE technique is used to address the issue of imbalanced classes, and other procedures such as data cleaning and dimensionality reduction are performed. The experimental findings show that the Random Forest method outperforms other techniques such as logistic regression and decision trees in forecasting

default samples.[10] Additionally, Reference 11 use CNN to forecast loan default by analyzing time series data from customer transactions, while Ma et al. employ XGBoost, LightGBM, Random Forest, and Logistic Regression (LR) to build prediction models for determining the likelihood of loan default. Finally, Zhu et al.[10] use the Random Forest method to create a loan default prediction model and compare it with other algorithms, including LR, DT, and SVM, and Khan et al.[12] use predictive models based on LR, DT, and Random Forest to decrease the time and effort required for loan approval and filtering out the best loan applicants.[13] The predictive model is beneficial in terms of decreasing the time and efforts necessary to approve loans as well as filtering out the best applicants for granting loans. Further study can be found in the following works.[2,14-22]

This research article is organized as follows: literature survey is mentioned in Section 2. Materials and methods are presented in Section 3. Result analysis and future work with conclusion is depicted in Sections 4 and 5, respectively.

3 | MATERIALS AND METHODS

3.1 | Dataset representation

The machine learning model is trained on a training dataset containing 25 columns, as shown in Table 1. The new applicant's details filled in the application form are treated as a test dataset. The model predicts whether the new applicant is suitable for loan approval based on the inferences it draws from the training datasets after the testing operation.

3.2 | Methodology

During training, random forests (RF) generate numerous decision trees to create machine learning models. To predict outcomes, the trees' results are combined, either by selecting the mode of classes for classification or by calculating the mean prediction for regression. These models are referred to as ensemble techniques because they use a group of outcomes to make a final decision. The importance of a feature in RF models is computed by measuring the reduction in node impurity and weighting it by the likelihood of reaching that node. The probability of reaching a node is determined by dividing the number of samples that reach the node by the total number of samples. The feature's importance increases as its value increases. Scikit-learn, a popular Python library for machine learning, uses Gini Importance represent in Equation (1) to calculate the importance of a node in RF models, assuming that only binary trees are used.

$$Km_n = W_n T_n - W_{\text{left}(n)} T_{\text{left}(n)} - W_{\text{right}(n)} T_{\text{right}(n)} \tag{1}$$

where Km_n = importance of node n; W_n = weighted number of samples reaching node n; T_n= impurity value of node n; left(n) = child node from left split on node n; right(n) = child node from right split on node n.

$$F_m = \frac{\sum_{n:\text{node } n \text{ splits on feature } m} Km_n}{\sum_{k \in \text{all nodes}} Km_k} \tag{2}$$

where F_m = importance of feature m; Km_n = importance of node n.

TABLE 1 Dataset representation.

1. Branch Code	2. Application In-take Date	3. Application Input Date	4. Applied Loan Amount	5. Applied Loan Tenor
6. Loan Purpose	7. Title	8. Gender	9. Age	10. Marital Status
11. Education Level	12. Residential Status	13. Monthly Housing/Rental	14. Contract Staff (Y/N)	15. Contract End Date
16. Employment Type	17. Nature of Business	18. Job Position	19. Monthly Income	20. Office (Area)
21. Date (Full Doc)	22. Date (Pending Doc)	23. Date (Pending Approval)	24. Final Status	25. Indicators

To obtain values between 0 and 1, these can be normalized by dividing each feature importance value by the sum of all the feature importance values represented in Equation (3).

$$NF_m = \frac{F_m}{\sum_{n \in all\ nodes} F_n} \tag{3}$$

To obtain the final feature importance, the importance values of each feature across all trees are averaged. This involves computing the sum of the feature's importance value in each tree and then dividing by the total number of trees is represent in Equation (4)

$$RF_m = \frac{\sum_{n \in all\ trees} NF_{mn}}{To} \tag{4}$$

where RF_m= importance of feature; NF_{mn}= normalized feature; To = total number of trees.

3.2.1 | Random Forest

Random Forest uses ensemble learning to construct multiple decision trees for classification, regression, and other tasks. The model selects random samples and features to build several decision trees, which are then combined to obtain the mode or mean prediction of the individual trees (Figure 1).[5] This approach helps to prevent overfitting by constructing smaller subtrees and randomizing subsets of features.

3.2.2 | The Random Forest classifier

An ensemble method called random forest employs multiple decision trees that operate in unison. Each tree produces a class prediction, and the most commonly predicted class is chosen as the model's output,[9] as depicted in Figure 2. The model's output is determined by aggregating the class predictions of all "n" trees.

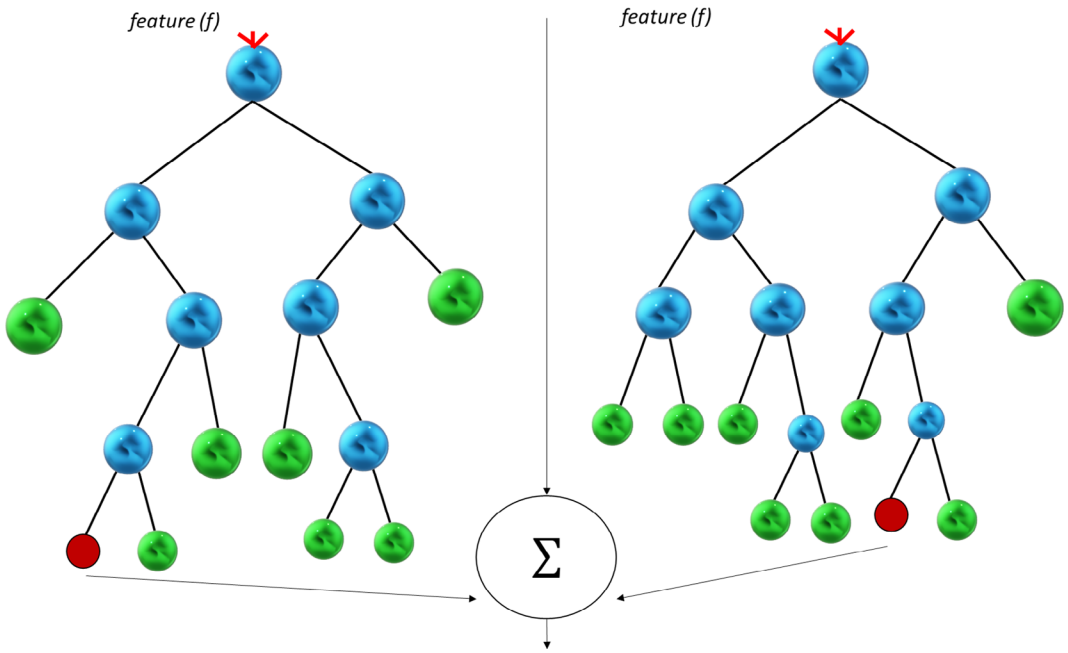

FIGURE 1 Random Forest tree.[5]

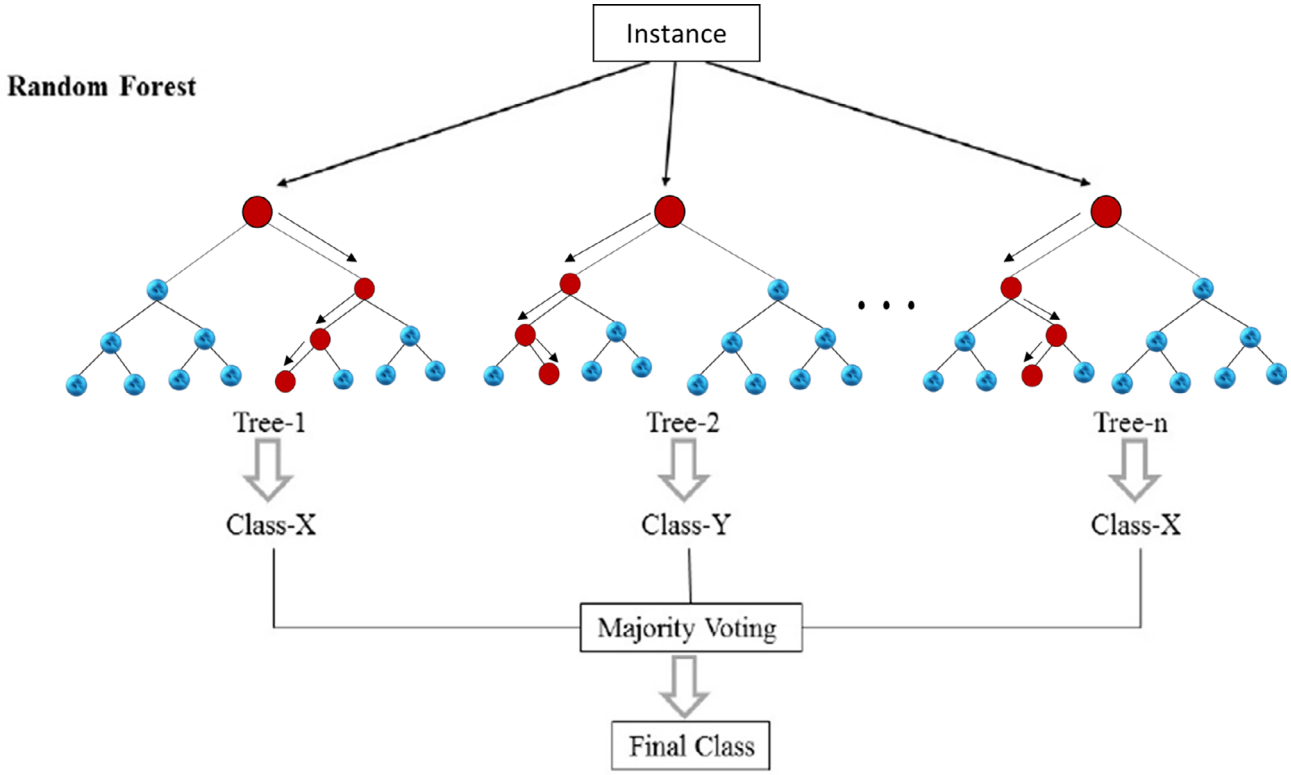

FIGURE 2 Random Forest classifier.[9]

3.3 | Working model

The $n_{\text{estimators}}$ parameter is set to 100 for the Random Forest model used in this case, indicating the number of decision trees used in the process. The model makes predictions by obtaining results from each tree for the chosen data samples and selecting the best solution by voting. Additionally, the model provides a reliable estimate of feature importance. The model works in the following manner:

- Randomly selecting samples from the dataset.
- Creating a decision tree for each sample and obtaining a prediction result for each tree.
- Voting is performed to predict the outcome.
- The final prediction is made by selecting the result with the most votes.

Figure 3 depicts the workings of a decision tree in the Random Forest model, which constructs decision trees for each sample and obtains a prediction result from each decision tree.[5]

4 | RESULT AND ANALYSIS

Figure 4 represent the idea regarding the distribution of loan acceptations, rejections and approval pending. From the data we have, we can conclude that rejected number of loan applications is 6278 and number of approval drawdown are 5345 and the number of approval pending are 46.

In Figure 5, loan approval on the basis of gender and marital status of the customer is represented, it can be sheening the about 68% of loan application is given by male. It can be observed that around 62% loan applications are given by married. Around 32% of loan applicants are single and people who have taken divorced and taking loan are around 4%. In case of loan applicants' distribution based on marital status around 62% of loan applications are given by married, 32%

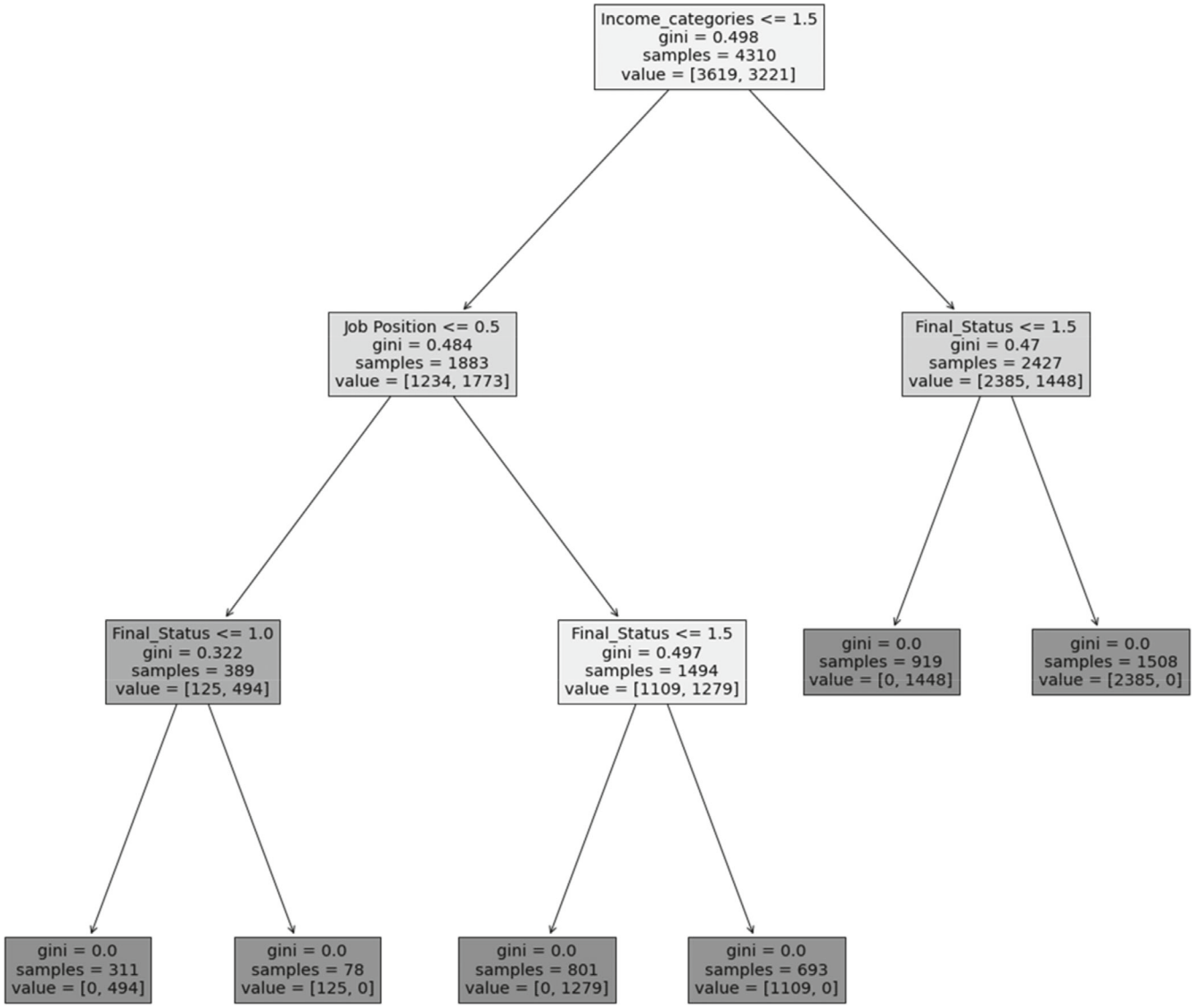

FIGURE 3 Working decision tree.[5]

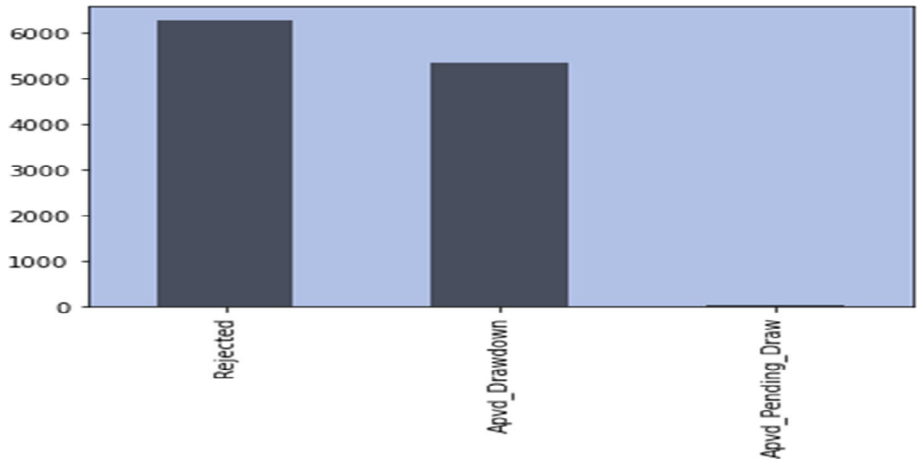

FIGURE 4 Counting rejected and accepted applicants.

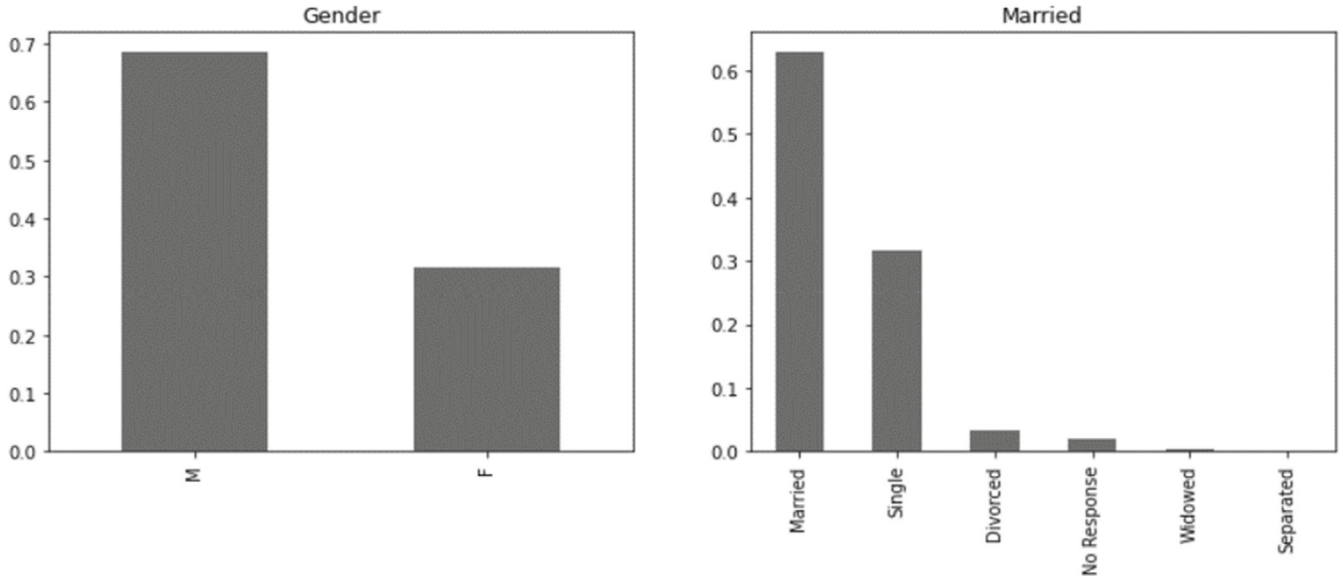

FIGURE 5 Loan approval on the basis of gender and marital status.

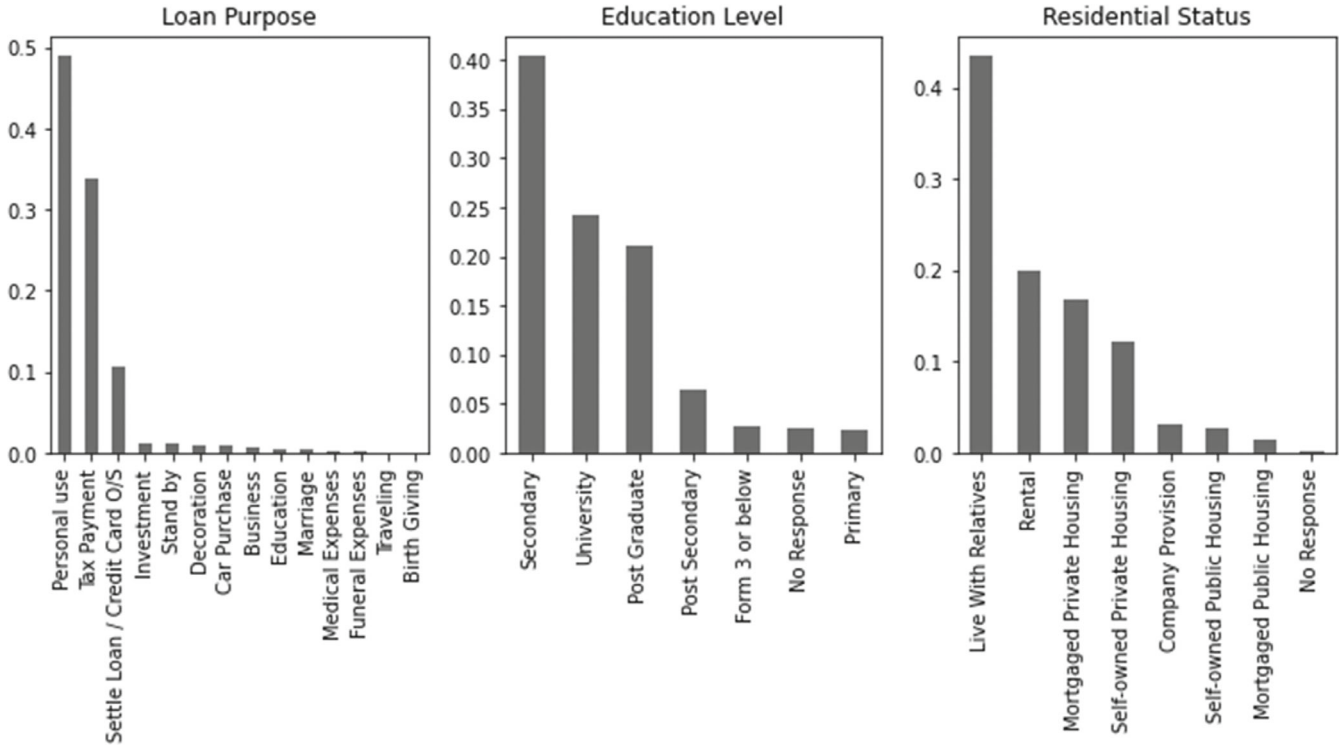

FIGURE 6 Loan approval on the basis of loan purpose, education level, and residential status.

of loan applicants are single, people who have taken divorced and taking a loan are around 4% and around 2%–3% of the loan applicants have not given any response on their marital status. By this we can say that around 2%–3% of the people do not want to talk about their marital status. One percent of the loan applicants are widowed.

In Figure 6, different levels of loan approval has been analyzed, from the loan purpose it can be found that around 50% of the loan applicants take the loan for personal use, the second reason for which people have applied for a loan is to pay their taxes in the last we see that around 10% of the loan applicants apply for loan to pay their credit card bills.

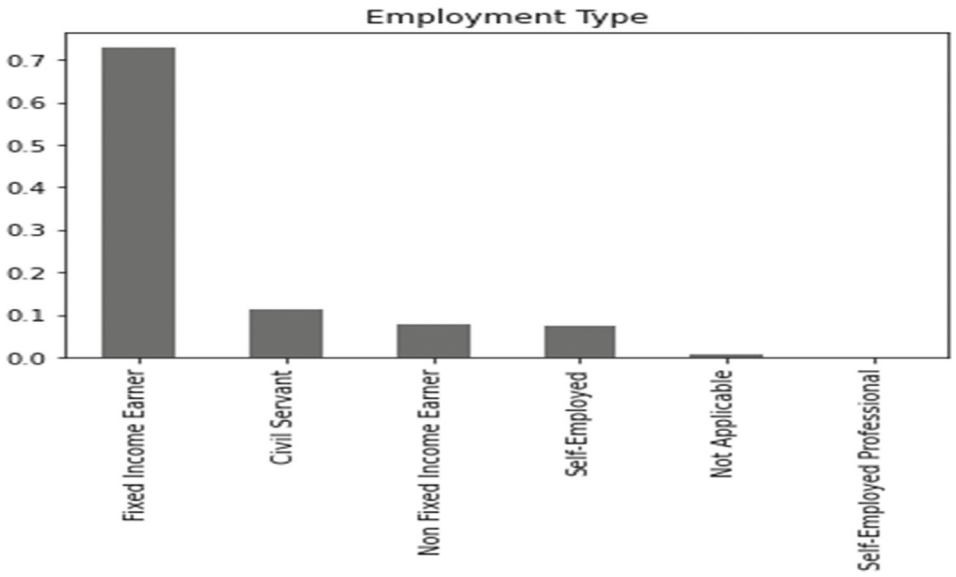

FIGURE 7 Loan approval on the basics of employment type.

When the education level is considered into account for loan approval around 40% of the loan applicants have done their education till secondary schooling, 25% of the loan applicants who have applied for loan are university students or university Pass out. Around 22% of the loan applicants who have applied for loan are post-graduate students or have done post-graduation and 4% of the loan applicants who have applied for loan are post-secondary students or have done post-secondary education level.

When residential status is taken account for loan approval it is found that around 45% of the loan applicants live with their relatives, 20% of the loan applicants live at a rented house, 18% of the loan applicants live at mortgaged private housing and 13% of the loan applicants live at self-owned private houses.

Loan approval on the basis of employment type from the above analysis in Figure 7, it can be found that 72% of the loan applicants are fixed income earner, 12% of the loan applicants are civil servants and around 9%–8% are non-fixed income earners or self-employed.

Figure 8 represented loan analysis is done on the basis of nature of the business of person it can be found around 28% as highest of the loan applicants are manager because they are having high salary can repay their loan in time, 18% are office workers, 13% of the loan applicants are in the service sector, 10% of the loan applicants are executive, and 7% are owner of business.

The loan approval analysis based on the monthly income of the customers is illustrated in Figure 9. The distribution of applicant monthly income is found to be mostly concentrated towards the left, indicating a non-normal distribution. The distribution is positively skewed or right-skewed. The presence of a significant number of outlier values is confirmed by the box plot. Income disparity in society can be one of the reasons for this. It is also possible that the disparity may be due to different education levels among the people being analyzed.

Figure 10 represents loan application on the basis of educational level of a student, student's higher number of post graduates and university people with very high incomes, which are appearing to be the outliers. So, we can say that people with higher education level apply more for bank loan.

Figure 11 represented the distribution of applied loan amount by the loan applicants this shows that majority of the people who apply for loan have their salary between 10,000 and 20,000.

4.1 | Distribution of loan tenor variable

In Figure 12, the distribution of loan tenor variable in term of *loan_amount_term* is analyzed, customer choosing 1 year term, that is, 12 months maximum percentage as compare to other term like 24 months, 60 months, etc.

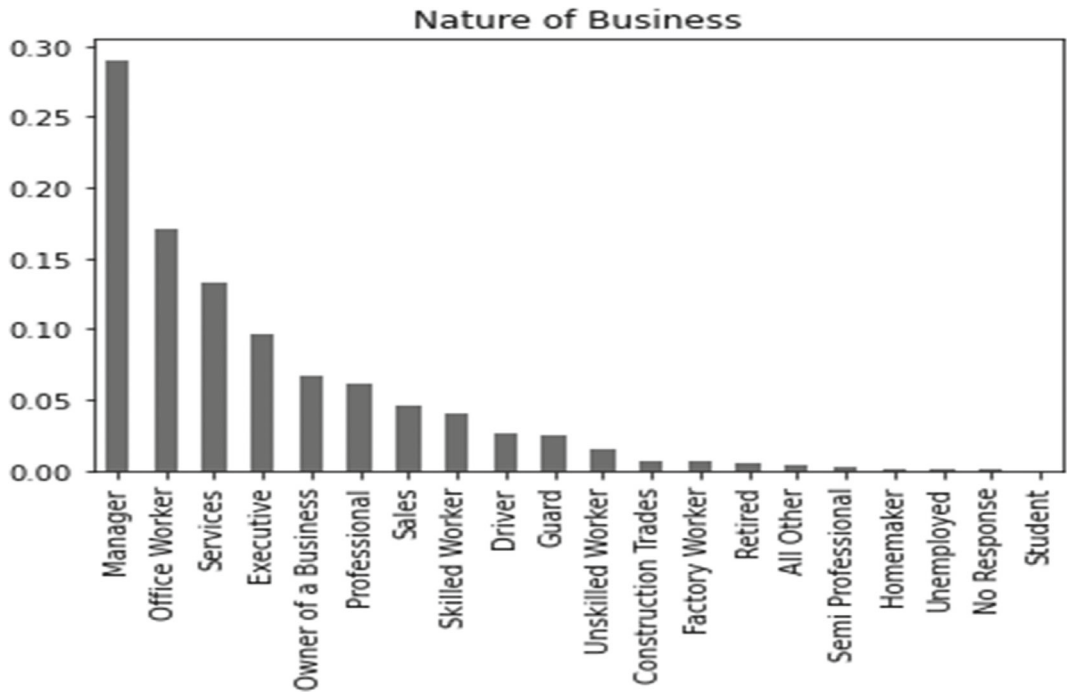

FIGURE 8 Loan analysis on the basics of nature of business.

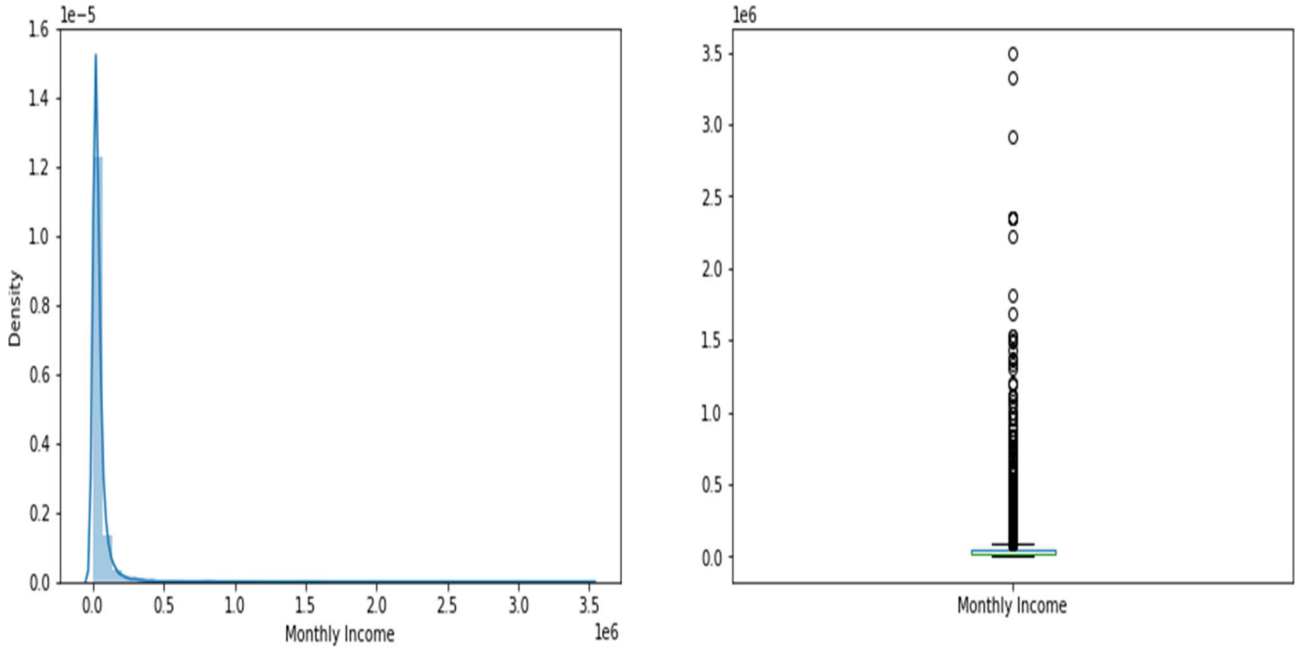

FIGURE 9 Loan approval analysis on the basics of monthly income.

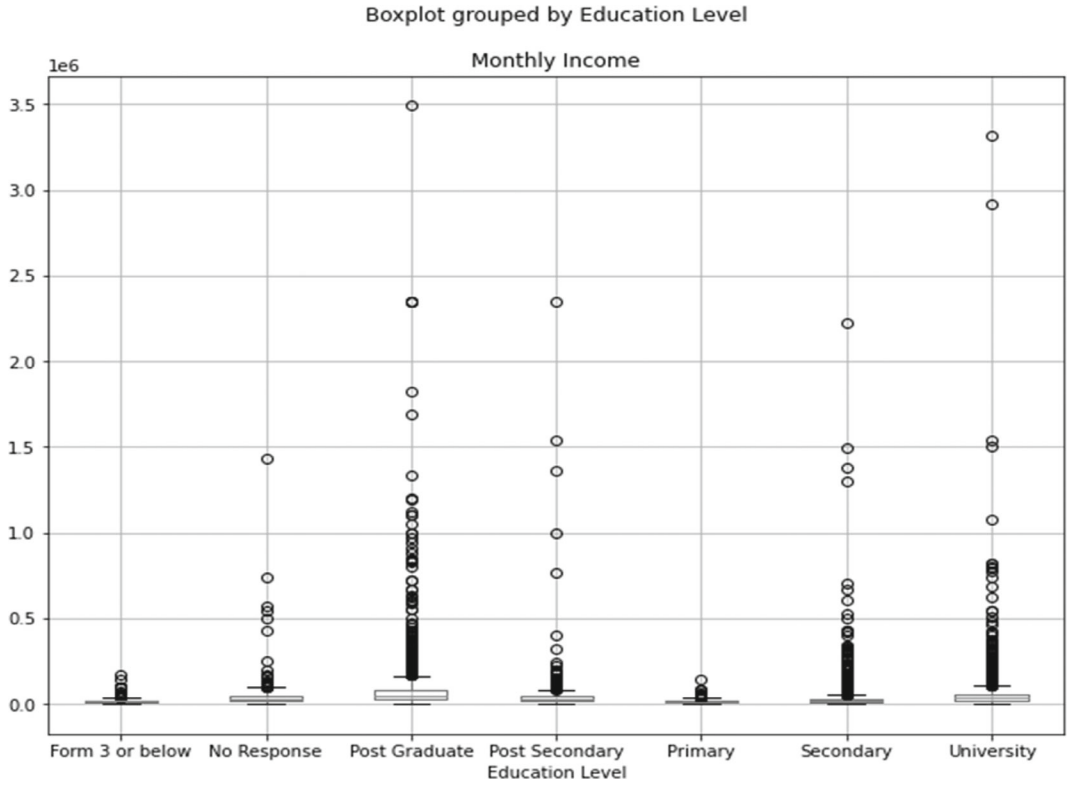

FIGURE 10 Loan application analysis on the basics of education level.

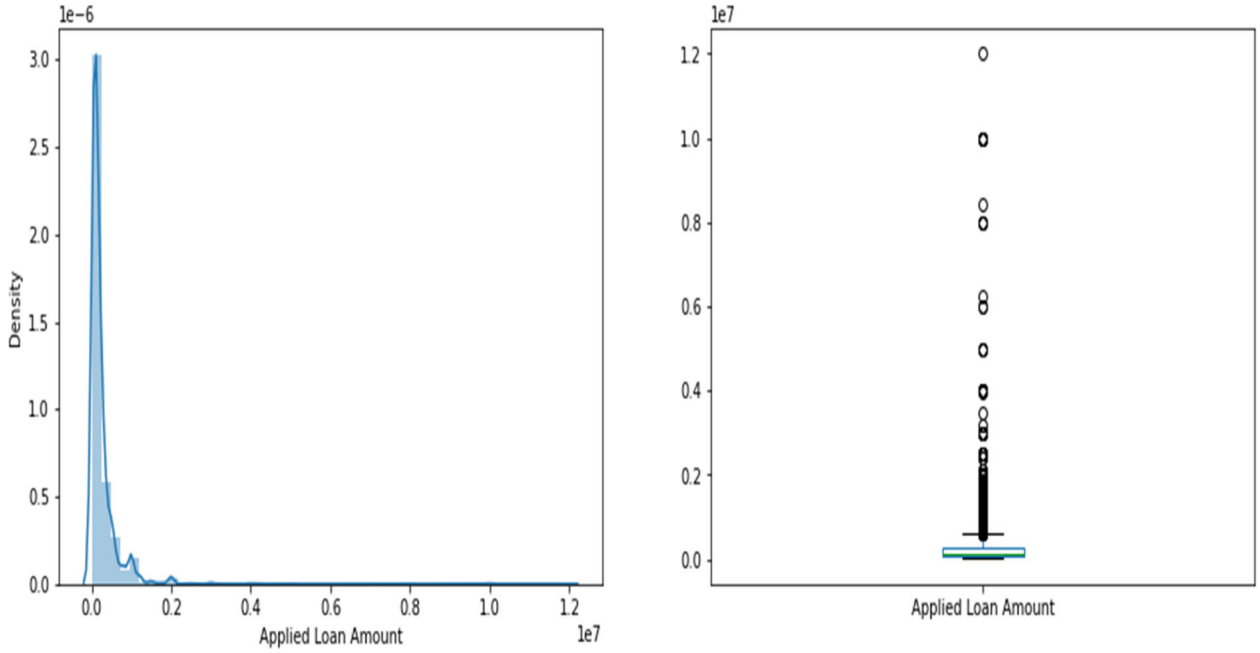

FIGURE 11 Loan application analysis on the basics of applied loan amount.

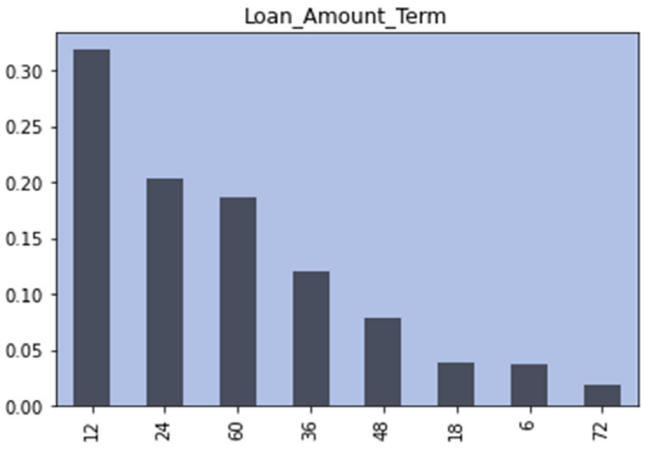

FIGURE 12 Loan amount term.

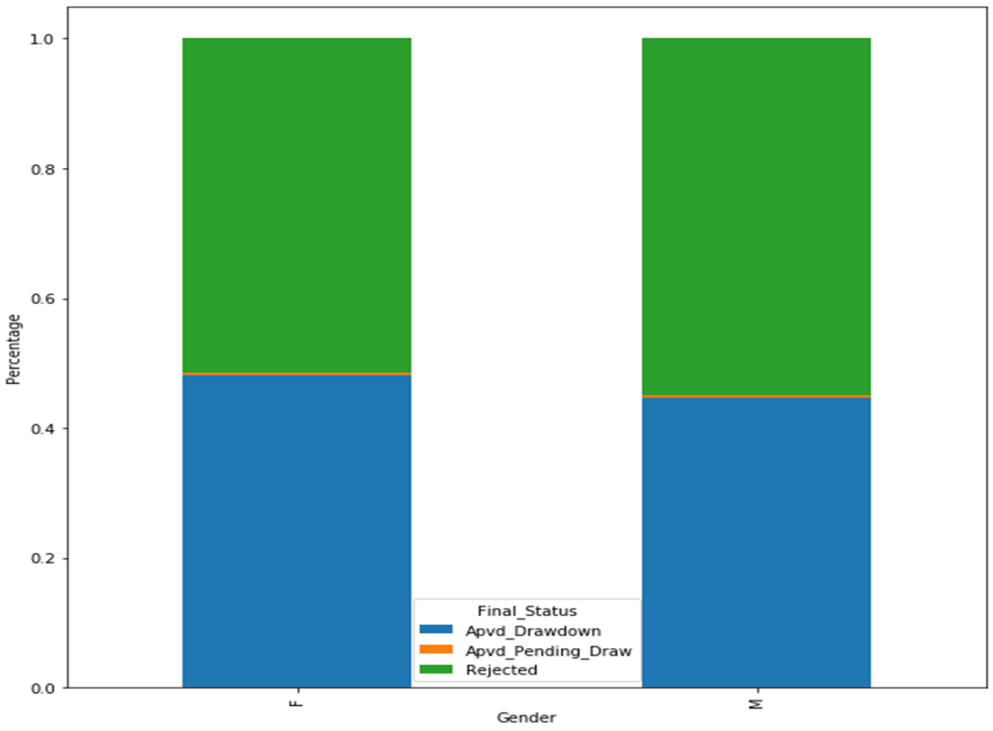

FIGURE 13 Final loan status varies with gender.

Figure 13 represents the final status varies with gender, that is, male and female. Male and female have quite similar chance of getting the loan. Loan approval rejection for male respondents is more than female respondents.

Figure 14 represents the final loan approval status varies with marriage. Married people have 50% chance of getting the loan approved. Loan approval rejection for divorced respondents is more than widowed and married respondents. The status of loan approval for single and widowed respondents is almost close to each other.

Figure 15 represented the final loan status with employment type. Civil servants have the most chance of getting the loan approved. People with Government/Semi-government have 70% chance of getting their Loan Approved. In business line people who are executive and professional have the most chance of getting the loan that is around 70%. People who take loan for tax payment purpose have the major chance of getting the loan that is around 72%.

Figure 16 represent the final loan status varies with educational qualification of a person. People who are post graduate or have a university degree have the most chance of getting loan approved. People who are primary and secondary degree

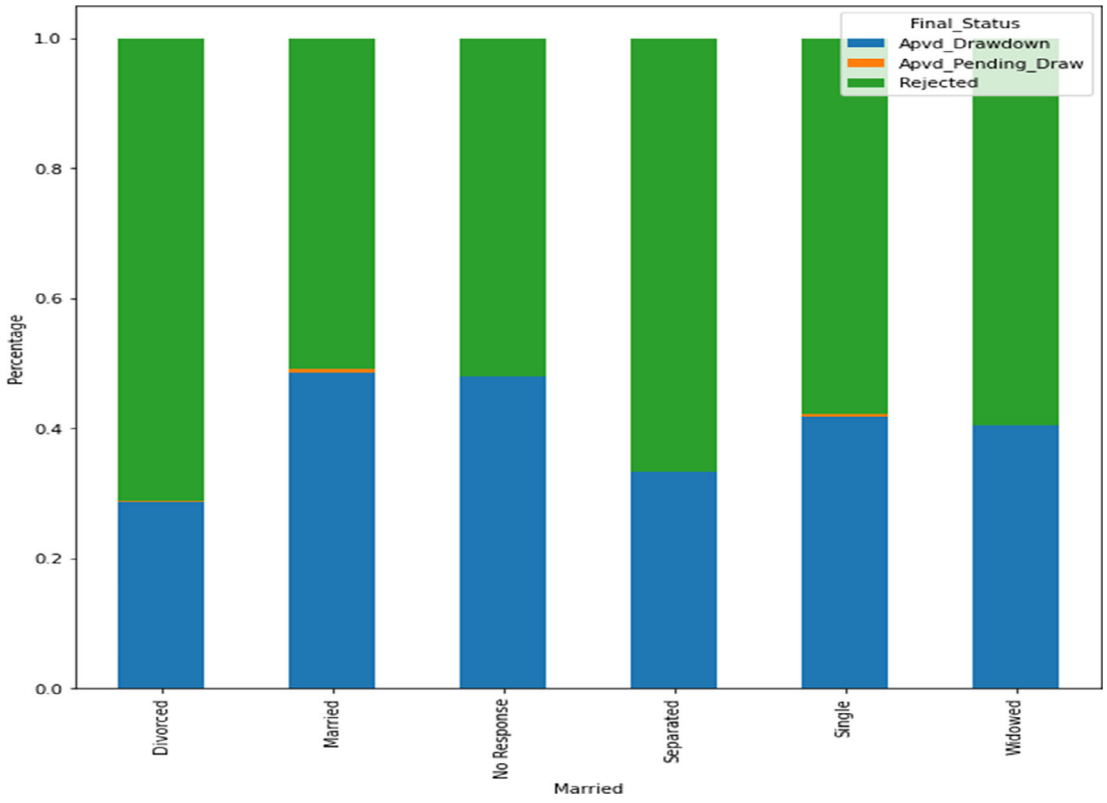

FIGURE 14 Final loan status varies with marriage.

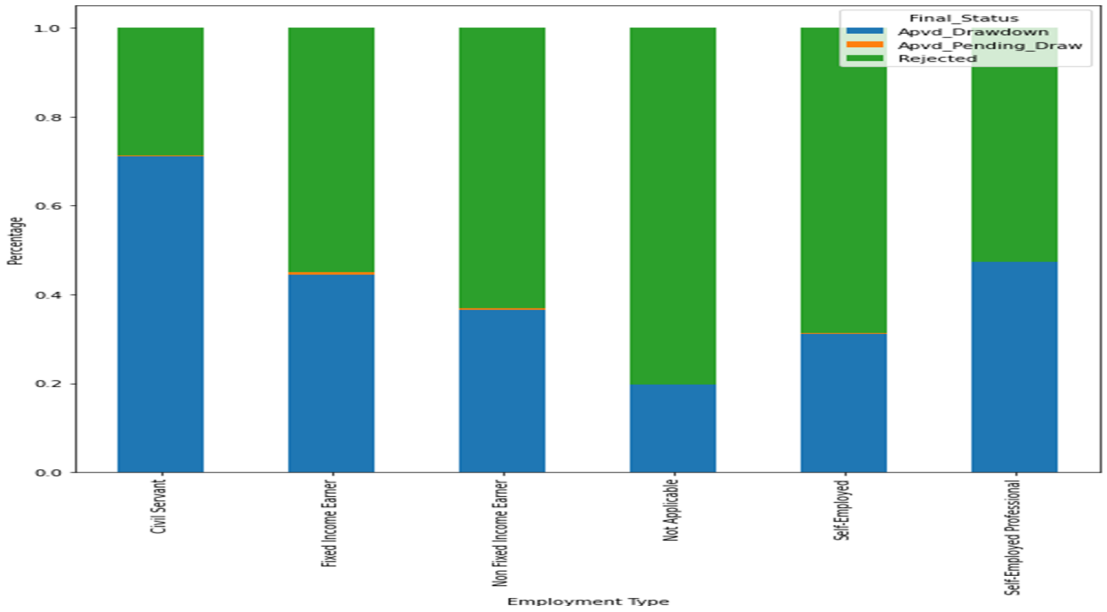

FIGURE 15 Final loan status varies with employment type.

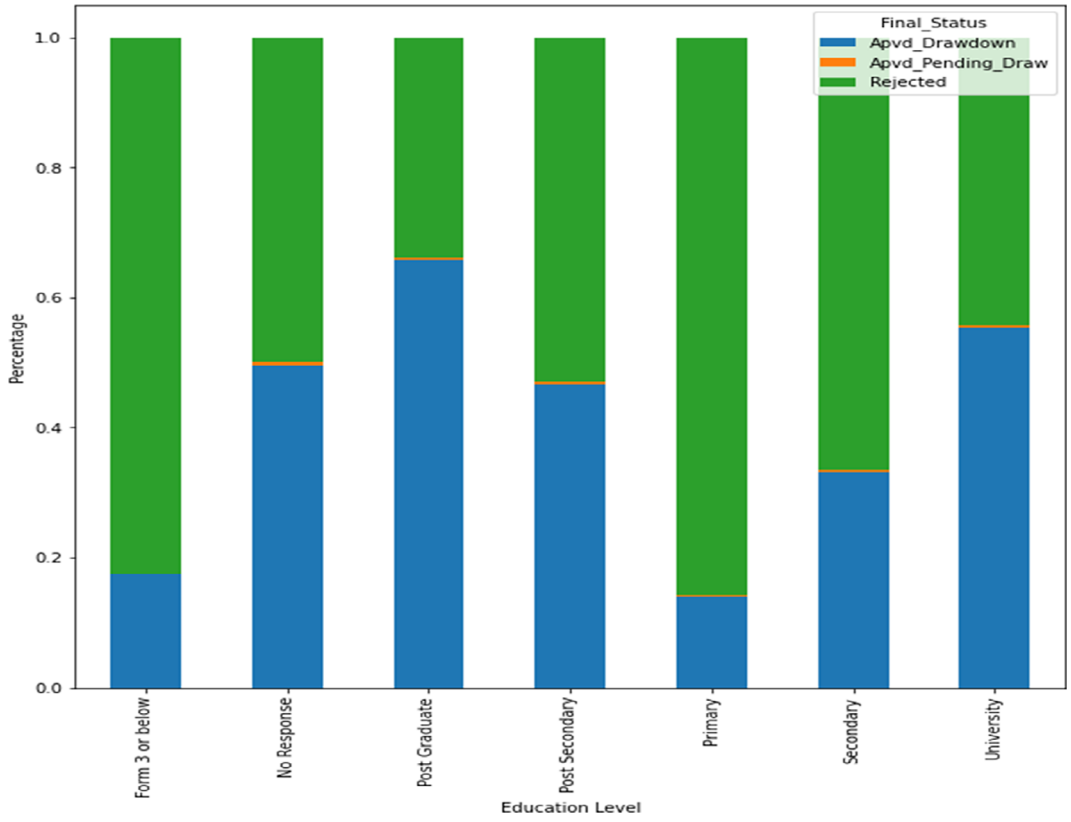

FIGURE 16 Final loan status varies with educational qualification.

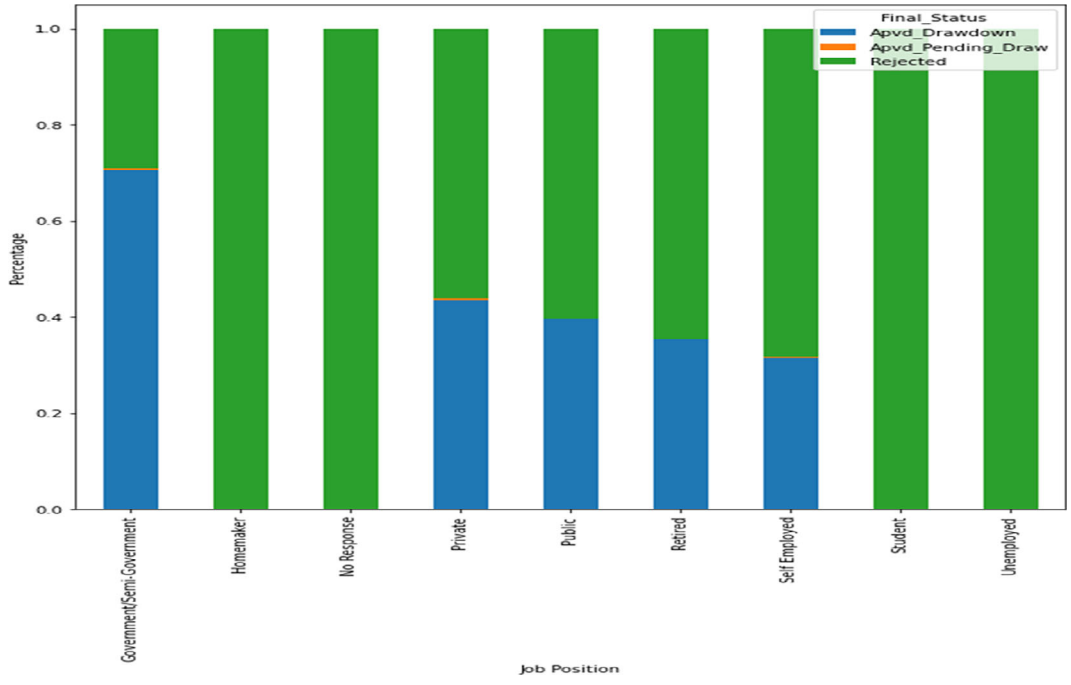

FIGURE 17 Final loan status varies with job type.

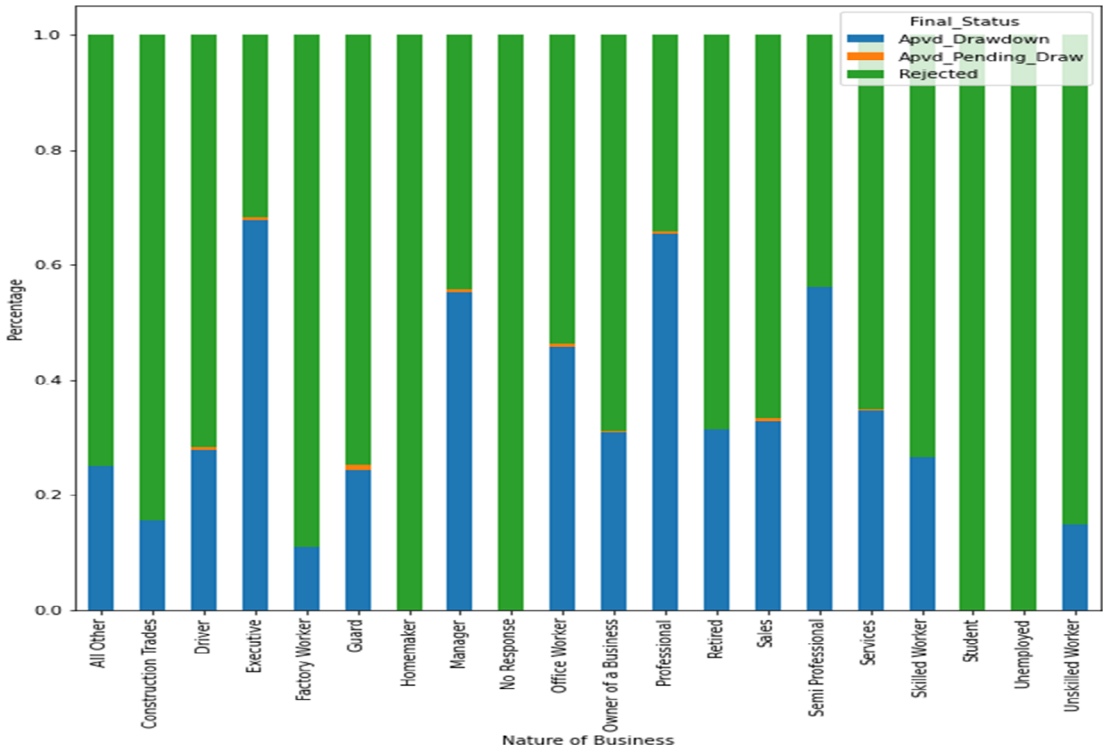

FIGURE 18 Final loan status varies with business.

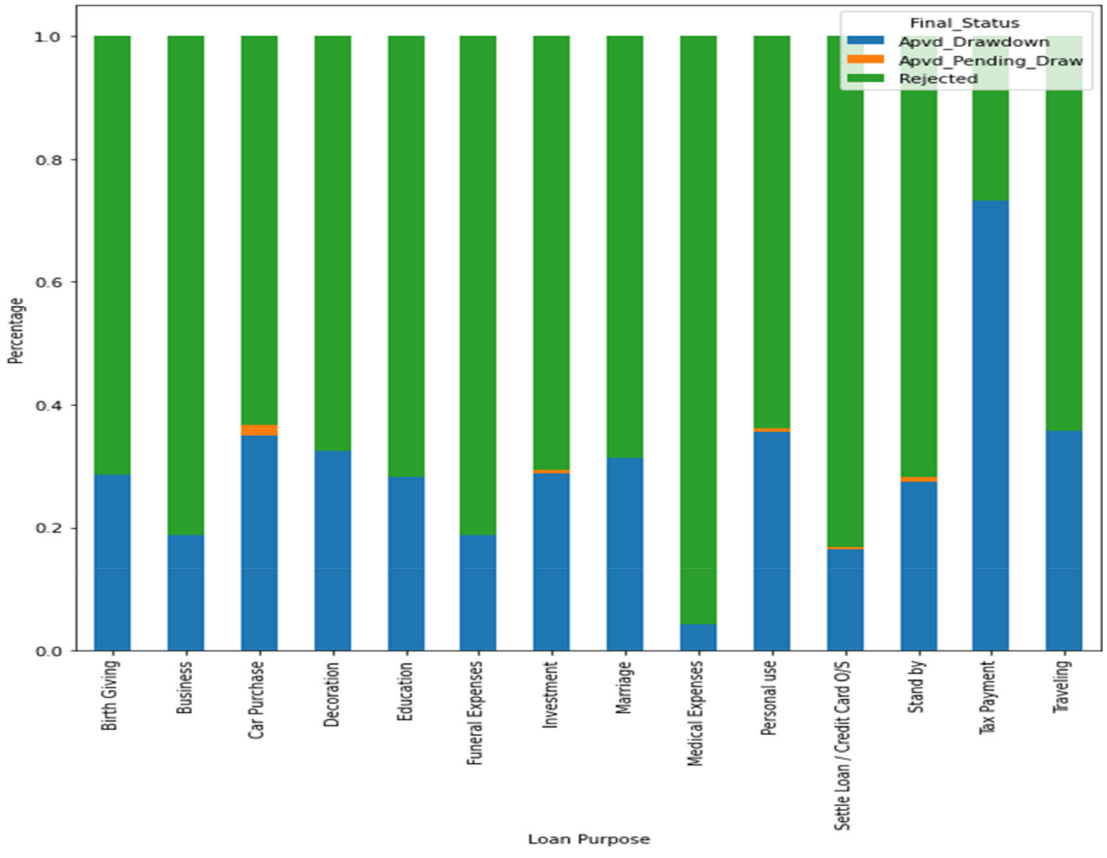

FIGURE 19 Final loan status varies with loan purposes.

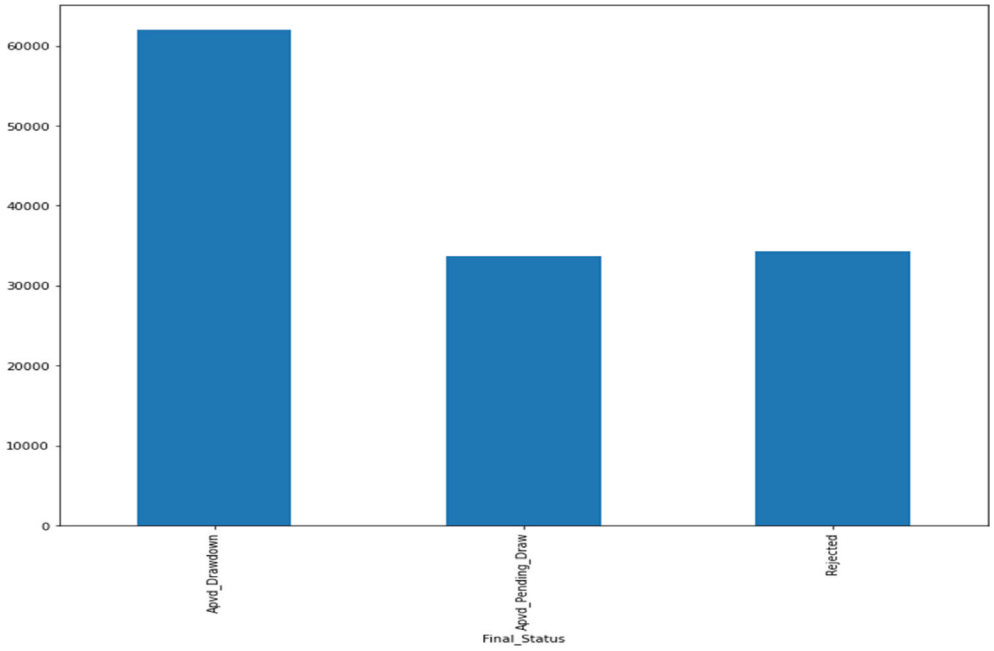

FIGURE 20 Final status of loan approval and loan rejected.

have the least chance of getting loan approval. People whose educational qualification is only post-secondary have around 48% chance of getting the loan approval.

Figure 17 represents the final loan status varies with job type where government/semi-government employees have the most chance of getting the loan approved. Homemaker and unemployed people have 0% chance of getting their loan Approved. The status of loan approval for retired and self-employed persons is almost close to each other.

Figure 18 represents final loan status varies with business Executive and professional persons have the most chance of getting the loan approved. Manger people have 55% chance of getting their loan approved. Factory workers have 10% chance of getting their loan approved. The status of loan approval for driver and skilled workers is almost close to each other.

Figure 19 represents final loan status varies with loan purpose of a person; people who take loan for tax payment purpose have the major chance of getting the loan approved. People who take loan for personal use and traveling purpose have 38% loan approval. People who take loan for business, decoration, education, marriage and birth giving have the least chance of getting the loan approved around less than 40%.

In Figure 20, the mean income of applicants, who were approved, for a loan is compared to the mean income of those who were not approved. It can be observed that the majority of applicants with an income above 60,000 have their loan approved, while applicants with salaries between 30,000–40,000 mostly have their loan applications rejected or pending.

5 | CONCLUSION AND FUTURE SCOPES

Banking is a crucial industry in many nations as it plays a vital role in determining a country's economic stability. The loan approval and prediction have been investigated in this research study based on different parameters and aspects such as gender, educational qualifications, type of employment, type of business, loan term, and marital status. Additionally, the number of loan approvals, disbursements, and rejections was analyzed. In future work, deep learning algorithms should be used to forecast loan repayment status. Moreover, with a larger dataset, we can have more training samples, which can help in resolving large variance issues and improve the validity of the analysis. Currently, the loan industry is rapidly growing, and many individuals request loans for various reasons. However, some fail to repay the loan amount, leading to significant financial losses for banks. Therefore, if an efficient technique to classify loan applicants in advance is developed, it would substantially reduce financial losses.

AUTHOR CONTRIBUTIONS

Debabrata Dansana: Data curation (equal); formal analysis (equal). **S Gopal Krishna Patro:** Software (equal); supervision (equal). **Brojo Kishore Mishra:** Validation (equal); writing – original draft (equal). **Vivek Prasad:** Methodology (equal); writing – original draft (equal). **Abdul Razak**: supervision. **Anteneh Wogasso Wodajo**: supervision and co-ordination.

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1002/eng2.12707.

DATA AVAILABILITY STATEMENT

All data used to support the findings of this study are included within the article.

ORCID

Abdul Razak https://orcid.org/0000-0001-7985-2502

REFERENCES

1. Aphale AS, Shinde SR. Predict loan approval in banking system machine learning approach for cooperative banks loan approval.
2. Sheikh MA, Goel AK, Kumar T. An approach for prediction of loan approval using machine learning algorithm. Paper presented at: 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 490–494. IEEE. 2020, July.
3. Supriya P, Pavani M, Saisushma N, Vimala N, Vikas K. Loan prediction by using machine learning models. *Int J Eng Tech*. 2019;5(22):144-148.
4. Madaan M, Kumar A, Keshri C, Jain R, Nagrath P. Loan default prediction using decision trees and random forest: a comparative study. *IOP Conf Ser Mater Sci Eng*. 2021;1022:012042.
5. Amin RK, Sibaroni Y. Implementation of decision tree using C4. 5 algorithm in decision making of loan application by debtor (case study: Bank pasar of Yogyakarta special region). Paper presented at: 2015 3rd International Conference on Information and Communication Technology (ICoICT), pp. 75–80, IEEE. 2015, May.
6. Jency X, Sumathi VP, Sri J. An exploratory data analysis for loan prediction based on nature of the clients. *Int J Recent Technol Eng*. 2019;7:176-179.
7. Shoumo SZH, Dhruba MIM, Hossain S, Ghani NH, Arif H, Islam S. Application of machine learning in credit risk assessment: a prelude to Smart banking. Paper presented at: TENCON 2019–2019 IEEE Region 10 Conference (TENCON), pp. 2023–2028, IEEE. 2019, October.
8. Aniceto MC, Barboza F, Kimura H. Machine learning predictivity applied to consumer creditworthiness. *Future Business J*. 2020;6(1):1-14.
9. Gautam K, Singh AP, Tyagi K, Kumar MS. Loan prediction using decision tree and Random Forest. 2008.
10. Zhu L, Qiu D, Ergu D, Ying C, Liu K. A study on predicting loan default based on the random forest algorithm. *Procedia Comput Sci*. 2019;162:503-513.
11. Kvamme H, Sellereite N, Aas K, Sjursen S. Predicting mortgage default using convolutional neural networks. *Expert Syst Appl*. 2018;102:207-217.
12. Khan A, Bhadola E, Kumar A, Singh N. Loan approval prediction model a comparative analysis. 2021.
13. Ma X, Sha J, Wang D, Yu Y, Yang Q, Niu X. Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. *Electron Commer Res Appl*. 2018;31:24-39.
14. Tejaswini J, Kavya TM, Ramya RDN, Triveni PS, Maddumala VR. Accurate loan approval prediction based on machine learning approach. *J Eng Sci*. 2020;11(4):523-532.
15. Arun K, Ishan G, Sanmeet K. Loan approval prediction based on machine learning approach. *IOSR J Comput Eng*. 2016;18(3):18-21.
16. Rath GB, Das D, Acharya B. Modern approach for loan sanctioning in banks using machine learning. *Advances in Machine Learning and Computational Intelligence*. Springer; 2021:179-188.
17. Gupta A, Pant V, Kumar S, Bansal PK. Bank loan prediction system using machine learning. Paper presented at: 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART), pp. 423–426, IEEE. 2020, December.
18. Gupta K, Chakrabarti B, Ansari AA, Rautaray SS, Pandey M. Loanification-loan approval classification using machine learning algorithms. 2021.
19. Karthiban R, Ambika M, Kannammal KE. A review on machine learning classification technique for Bank loan approval. Paper presented at: 2019 International Conference on Computer Communication and Informatics (ICCCI), pp. 1–6. IEEE. 2019, January.
20. Odegua R. Predicting Bank loan default with extreme gradient boosting. *arXiv Preprint*. 2020;arXiv:2002.02011.

21. Rao MS, Sekhar C, Bhattacharyya D. Comparative analysis of machine learning models on loan risk analysis. *Machine Intelligence and Soft Computing*. Springer; 2021:81-90.
22. Dushimimana B, Wambui Y, Lubega T, McSharry PE. Use of machine learning techniques to create a credit score model for airtime loans. *J Risk Finan Manag*. 2020;13(8):180.

How to cite this article: Dansana D, Patro SGK, Mishra BK, Prasad V, Razak A, Wodajo AW. Analyzing the impact of loan features on bank loan prediction using Random Forest algorithm. *Engineering Reports*. 2024;6(2):e12707. doi: 10.1002/eng2.12707