iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** III **Month of publication:** March 2025

DOI: https://doi.org/10.22214/ijraset.2025.67894

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue III Mar 2025- Available at www.ijraset.com

Comparative Study of Machine Learning Algorithms for Loan Default Prediction

Mrs. Swati Chandurkar[1], Mahesh Kakde[2], Manoj Jamble[3], Vinay Hivrale[4], Om Hedau[5]
Dept. Of Computer Engineering, Pimpri Chinchwad College Of Engineering, Pune

Abstract: ***This review paper compares machine learning algorithms for loan default prediction, focusing on preprocessing techniques, Random Forest, Gradient Boosting Machine (GBM), Naive Bayes, and visualization methods. It highlights the critical role of these algorithms in financial risk assessment. The study evaluates their performance, strengths, weaknesses, and suitability for different data types. It emphasizes the importance of selecting the right algorithm based on dataset characteristics. The findings emphasize advanced machine learning's significance in improving risk management and lending decisions in the financial sector. Ongoing research aims to enhance prediction model accuracy further.***

Keywords: prediction, machine learning, Preprocessing, Random Forest, Gradient Boosting Machine (GBM), Naive Bayes, Visualization Methods.

I. INTRODUCTION

Loan default prediction is crucial for risk assessment in finance, aiding institutions in informed decision-making. This paper delves into machine learning algorithms tailored for this purpose, emphasizing their role in financial risk assessment. The journey begins with collecting and preprocessing historical loan data, addressing missing values, encoding variables, and selecting key features like credit scores. Python's libraries, especially XGBoost and Random Forest, are pivotal in training accurate prediction models.

Model evaluation, gauging accuracy, precision, recall, and more, guides risk assessment and decision-making, such as adjusting loan terms. The paper compares Random Forest, GBM, and Naive Bayes, KNN evaluating their performance for loan default prediction. Visualization techniques like histograms and box plots help interpret data patterns, aiding in model selection and refinement. Leveraging advanced algorithms and visualization improves risk management and informs lending decisions, with ongoing research enhancing prediction accuracy in finance.

II. LITERATURE REVIEW

This section provides an overview of the existing work in the field of loan prediction using machine learning (ML) and deep learning (DL) models, focusing on various algorithms employed to enhance the loan prediction process and aid banking authorities and financial firms in selecting candidates with low credit risk. Loan prediction has become a pivotal subject in the banking and finance sectors, with credit scoring emerging as a crucial tool in a competitive financial landscape. The recent advancements in data science and artificial intelligence have propelled research interest in loan prediction and credit risk assessment. There is a growing demand for improved credit scoring models due to the increased need for loans, prompting researchers and banking authorities to explore machine learning algorithms and neural networks for credit scoring and risk assessment.

The Random Forest Algorithm adopted to develop models for loan default prediction [4] [5]. Studies concluded that the Random Forest algorithm exhibits superior accuracy (98%) compared to other algorithms like logistic regression (73%), decision trees (95%), and support vector machines (75%). The effectiveness of the Random Forest algorithm was attributed to its competitive classification accuracy and simplicity, as discussed in paper [5].

Other studies [7] reviewed credit scoring for mortgage loans and highlighted the importance of meeting specific requirements for credit approval to minimize the risk of non- payment. [8] Decision Trees was utilized for credit scoring, achieving an accuracy of 81% through rigorous data pre- processing and model evaluation.

Additionally, research demonstrated that Support Vector Machines [11] can outperform other models like logistic regression and random forest in terms of predictive performance. [14] leveraged the Naive Bayesian algorithm and supplementary techniques like k-NN and binning to enhance classification accuracy and data quality. Moreover, the local banks in certain regions predominantly use logit method-based models, with other methods like CART or neural networks [15] serving as supplementary tools in variable selection and model evaluation. In conclusion, the existing literature showcases a diverse array of methodologies and algorithms employed in loan prediction and credit risk assessment, emphasizing the significance of data quality, feature selection, and algorithm choice in developing robust and accurate models for loan approval processes.

III. THEORY

The push toward complete automation is evident across various sectors globally, with ongoing developments in concepts and methods to achieve this objective. One field that has particularly captured the interest and enthusiasm of scientists, researchers, and technologists is Artificial Intelligence (AI). AI involves creating computer systems or machines capable of simulating human-like intelligence and behavior [16]. This concept traces back to the early days of computing and has since branched into diverse areas such as Machine Learning (ML), Neural Networks, and Natural Language Processing (NLP) [16].

Machine Learning (ML) is a pivotal concept that empowers machines to learn from real-world interactions and observations, mimicking human learning and enhancing their capabilities using input data [11]. In recent years, ML has garnered significant attention from researchers and technologists, leading to the implementation of various ML models and algorithms across different sectors. For instance, in fields like banking and finance, ML models have been instrumental in detecting patterns and drawing conclusions related to credit card frauds and loan default predictions, streamlining processes and improving accuracy.

The ML models mentioned in this context are based on a variety of ML methods, making it challenging to compile an exhaustive list. Typically, the nomenclature of a model reflects a combination of factors such as data structure, design, estimator, ensemble mechanism, and more [6]. In this paper, the focus is on two prominent algorithms in the ML domain: Random Forest and Decision Trees. Decision Trees are versatile algorithms used for classification and regression tasks [7]. They are widely favored for classification purposes and comprise nodes like branches, leaf nodes, and root nodes, forming a tree-like structure through a Recursive Partitioning Algorithm (RPA) [1]. Each leaf node represents a class label, while internal nodes represent test results based on attributes. On the other hand, Random Forest, belonging to the supervised learning algorithm category, is also utilized for classification and regression tasks. It involves building a predictor ensemble using multiple decision trees that expand within randomly selected data subspaces [12]. Employing Random Forest offers several advantages over other ML algorithms, such as immunity to overfitting, accurate classification or regression, and improved efficiency with large databases.

IV. DATASET

The dataset under consideration contains information related to socioeconomic factors and risk profiles of individuals. It consists of 252,000 observations and comprises 13 attributes, each providing insights into various aspects of an individual's demographic and professional background. The attributes include:

Id: A unique identifier for everyone.

Income: The annual income of the individual. Age: The age of the individual in years.

Experience: The number of years of work experience. Married/Single: The marital status of the individual. House_Ownership: The type of housing owned by the individual.

Car_Ownership: The ownership status of a car. Profession: The occupation or profession of the individual. CITY: The city of residence.

STATE: The state of residence.

CURRENT_JOB_YRS: The number of years in the current job.

CURRENT_HOUSE_YRS: The number of years in the current residence.

Risk_Flag: A binary variable indicating the risk profile of the individual, where 0 represents low risk and 1 represents high risk.

This dataset offers a comprehensive view of socioeconomic indicators and their association with risk profiles, making it suitable for various analytical tasks such as exploratory data analysis, predictive modeling, and comparative studies across different algorithms.

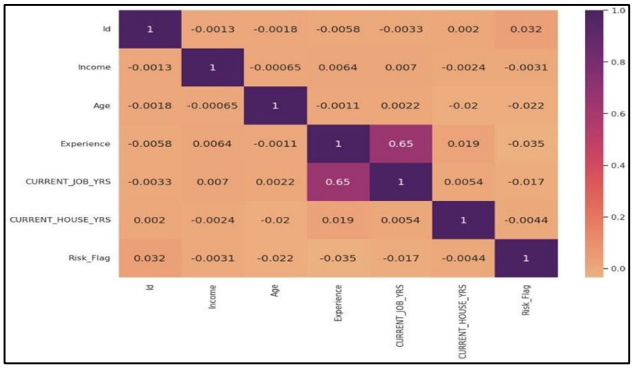


Fig. 1. Correlation Matrix for dataset.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue III Mar 2025- Available at www.ijraset.com

V. PROPSED MODEL

In our quest to devise a robust and accurate solution for loan default prediction, we propose a hybrid ensemble model that amalgamates the strengths of various machine learning algorithms. This approach aims to leverage the diversity of individual models to mitigate the weaknesses inherent in any single algorithm, thereby enhancing overall predictive performance and reliability.

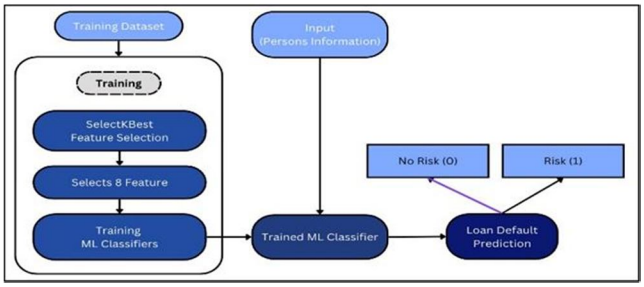


Fig.2 .Flowchart

A. Random Forest Classifier:

Random Forest is selected as the cornerstone ofour ensemble model due to its inherent robustness and versatility in handling complex, high-dimensional datasets. By aggregating predictions from multiple decision trees, Random Forest effectively captures nonlinear relationships and interactions between features, making it well-suited for loan default prediction tasks For each decision tree, Scikit-learn calculates a nodes importance using Gini Importance, assuming only two child nodes (binary tree):

$$ni_j = w_j C_j - W_{left(j)}C_{left(j)} - w_{right(j)}C_{right(j)}$$

- ni sub(j)= the importance of node j
- w sub(j) = weighted number of samples reaching nodej
- C sub(j)= the impurity value of node j
- left(j) = child node from left split on node j
- right(j) = child node from right split on node j
- sub() is being used as subscript isn't available in Medium

The importance for each feature on a decision tree is then calculated as:

$$fi_j = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k \in all\ nodes}\ ni_k}$$

- fi sub(i)= the importance of feature i
- ni sub(j)= the importance of node j

These can then be normalized to a value between 0 and 1 by dividing by the sum of all feature importance values:

$$normfi_j = \frac{fi_i}{\sum_{j \in all\ features} fi_j}$$

The final feature importance, at the Random Forest level, is its average over all the trees. The sum of the feature's importance value on each trees is calculated and divided by the total number of trees:

$$RFfi_i = \frac{\sum_{j \in all\ trees}\ norm\ fii_{ij}}{T}$$

- RFfi sub(i)= the importance of feature i calculated from all trees in the Random Forest model
- normfi sub(ij)= the normalized feature importance for i in tree j
- T = total number of trees

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue III Mar 2025- Available at www.ijraset.com

B. Gradient Boosting Machine (GBM):

Gradient Boosting Machine is incorporated to complement Random Forest by focusing on iterative refinement of the model's predictive capability.

Through sequential construction of decision trees that emphasize the residual errors of previous trees, GBM excels in capturing intricate patterns and nuances present in the data. The key Formulas involved are:

1) Objective Function

XGBoost tries to minimize the Objective function. It combines the traveling loss and regularization term.

$$Objective(T) = Loss + Regularization$$

For XGBoost the objective function is given as:

$$Objective(T) = \sum l(yi, y_{pred}, i) + \sum \Omega(f)$$

Where:

- T represents the ensemble of decision trees
- l(y, y_pred) is a differentiable convex loss function that measures the — difference between the true output (y) and the predicted output (y_pred)
- yi is the true output, for instance i y_pred,i is the predicted output for instance i
- Ω(f) is the regularization term applied to each tree (f) in the ensemble (T)

2) Additive Training

It learns in additive manner creating an iterative ensemble of decision trees (weak learners) that gradually optimizes the objective function.

$$F(x) = Fm - 1(x) + fm(x)$$

Where:

- Fm(x) is the prediction after adding m trees
- Fm-1(x) is the prediction up to m-1 trees
- fm(x) is the new tree added in the m-th iteration

3) Gradient and Hessian Calculation

$$Gradient(g): \nabla l(y, y_{pred}) = \frac{d(l)}{dy_{pred}}$$

$$Hessian(h): \nabla^2 l(y, y_{pred}) = \frac{d^2(l)}{(dy^2 \quad i)}$$

4) Tree Construction

Best tree that minimizes the objective function, using (g) and (h) is formulated in the m_th iteration

$$Gain = \frac{1}{2} * \left[\frac{Gl^2}{(Hl + \lambda)} + \frac{Gr^2}{(Hl + \lambda)} - \frac{G^2}{(Hl + \lambda)}\right] - \gamma$$

Where,

- Gl and Gr are the sums of gradients in the left and right regions of the split
- Hl and Hr are the sums of Hessians in the left and right regions of the split
- G = Gl + Gr, the sum of gradients for the entire node
- H = Hl + Hr, the sum of Hessians for the entire node
- λ, the L2 regularization term
- γ, the minimum loss reduction required for a split (another regularization term)

The final prediction for the model is obtained by summing up the predictions from all the models.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue III Mar 2025- Available at www.ijraset.com

C. K-Nearest Neighbors (KNN):

K-Nearest Neighbors (KNN) algorithm is included in our ensemble to exploit local patterns and relationships within the data. By classifying a data point based on the majority class among its nearest neighbors, KNN provides a simple yet powerful approach for loan default prediction, especially in scenarios with localized decision boundary Manhattan Distance metric is been used where the total distance traveled by the object instead of the displacement. This metric is calculated by summing the absolute difference between the coordinates of the points in n-dimensions.

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

D. Naive Bayes:

Naive Bayes algorithm is integrated into our ensemble for its simplicity and efficiency in handling categorical data and making probabilistic predictions.

Despite its simplistic assumptions of feature independence, Naive Bayes can provide valuable insights and complement the ensemble's predictive capabilities.

Naive Bayes can provide valuable insights and complement the ensemble's predictive capabilities.

$$P(Class|Features) = \frac{P(Featues|Class) * P(Class)}{P(Features)}$$

- P(Class | Features): Probability of a class given a set of features
- P(Features | Class): Probability of features given a class (often estimated from training data)
- P(Class): Prior probability of a class
- P(Features): Total probability of the feature set
- (normalization factor)

E. XGBoost (Extreme Gradient Boosting)

XGBoost is included as an additional boosting algorithm known for its scalability, speed, and high performance.

With advanced regularization techniques and optimization algorithms, XGBoost can effectively handle complex dataset and capture nonlinear relationships, further enhancing the ensemble model's predictive accuracy.

F. Model Evaluation and Validation

The proposed ensemble model undergoes comprehensive evaluation using cross-validation techniques and is validated on independent test sets. Performance metrics including accuracy, precision, recall, F1-score are computed to assess the model's efficacy and robustness across various evaluation scenarios. The proposed hybrid ensemble approach represents a sophisticated yet practical solution for loan default prediction, harnessing the collective intelligence of multiple machine learning algorithms to enhance predictive accuracy and reliability. By leveraging a diverse set of models, our approach equips financial institutions with a powerful tool to effectively manage credit risk and make informed lending decisions.

VI. METHODOLOGY

A. Experimental Methodology

This section outlines the experimental procedures for evaluating the performance of machine learning models in predicting possible loan defaults and conducting a comparative analysis. It covers dataset details, data preprocessing, models used, evaluation metrics, and the overall experimental methodology.

1) Data Preprocessing

Before initiating model training, the dataset underwent meticulous preprocessing to ensure data quality and consistency. The following steps were executed:

a) *Handling Missing Data:* Records with missing values were identified, and appropriate strategies, such as imputation or removal, were employed to address missing data while preserving data integrity.

b) *Outlier Detection and Treatment:* Outliers, if present, were identified and treated using robust statistical techniques to mitigate their influence on model performance.

c) *Feature Scaling:* Numerical features were scaled to a uniform range to prevent dominance by features with larger magnitudes during model training.
d) *Categorical Encoding:* Categorical variables were encoded using techniques such as one-hot encoding or label encoding to facilitate their incorporation into machine learning algorithms.
e) *Feature Engineering:* Additional features were engineered to capture complex relationships and enhance model performance. This involved transformations, interactions, and aggregations of existing features.
f) *Train-Test Split:* The preprocessed dataset was partitioned into training and testing subsets using a standard split ratio, typically 80% for training and 20% for testing. Additionally, cross-validation techniques were applied during model training to ensure robustness and generalizability.

2) Model Used

In the loan default prediction module, a diverse set of machine learning algorithms was employed to model the relationship between borrower attributes and loan default likelihood. The selected algorithms include:

a) Logistic Regression
b) Random Forest
c) KNN
d) Gradient Boosting

Each algorithm offers distinct advantages in capturing complex patterns and nonlinear relationships within the dataset.

3) Evaluation Metrics

To evaluate the predictive performance of the machine learning models, a suite of evaluation metrics was employed, including:

a) *Accuracy*

Proportion of correctly predicted loan defaults.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

b) *Precision*

Ability to correctly identify loan defaults among predicted defaults.

$$Precision = \frac{TP}{TP + FP}$$

c) *Recall:*

Proportion of actual loan defaults correctly identified by the model.

$$Recall = \frac{TP}{TP + FN}$$

VII. RESULTS

Algorithm	Accuracy		
	Accuracy	Recall	Precision
Random Forest	90.10	34.56	53.39
GBM	87.60	50.83	81.81
Naive Bayes	87.825	49.63	52.634
KNN	88.839	50.733	54.47069
XGBoost	87.95	22.44	43.93

TABLE I. RESULTS

VIII. FUTURE SCOPE

Implementing advanced deep learning models like recurrent neural networks (RNNs) and long short-term memory (LSTM) networks for loan default prediction to capture temporal dependencies and improve predictive accuracy. Exploring ensemble methods that combine the strengths of multiple machine learning algorithms, such as Random Forest, GBM, and Naive Bayes, to further enhance prediction performance and robustness. Incorporating additional features and data sources, such as social media activity, transaction histories, and external economic indicators, to enrich the dataset and improve the model's predictive capabilities. Conducting research on explainable AI (XAI) techniques to enhance model interpretability and provide insights into the factors influencing loan default predictions, aiding stakeholders in making more informed decisions. Collaborating with financial institutions and regulatory bodies to validate and deploy machine learning models in real-world scenarios, ensuring compliance with industry standards and regulatory requirements. Exploring the integration of alternative data sources, such as alternative credit scoring data, IoT devices, and block chain technology, to further refine risk assessment models and adapt to evolving financial landscapes. Investigating the impact of demographic shifts, economic trends, and global events on loan default rates and developing dynamic prediction models that can adapt to changing environments and mitigate emerging risks effectively.

IX. CONCLUSION

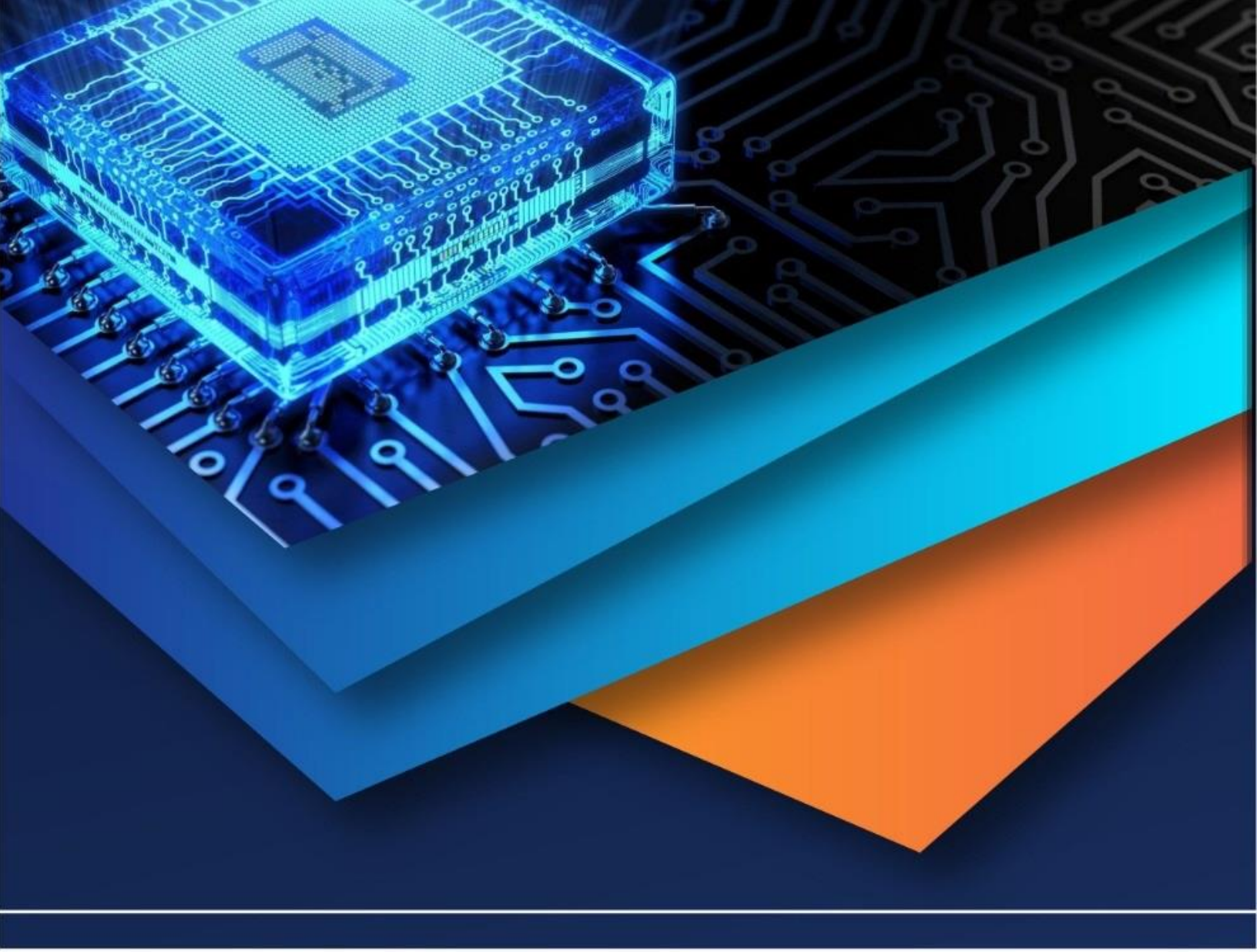
The analysis conducted in this review paper points to Random Forest as the top-performing algorithm for loan default prediction, boasting an accuracy rate of 90.10%. In comparison, Gradient Boosting Machine and Naive Bayes trailed slightly behind, achieving accuracies of 87.60% and 87.83%, respectively.

However, it's crucial to note that the most suitable algorithm choice hinges on specific dataset attributes and problem intricacies. For instance, Random Forest excels with large and complex datasets, while decision trees may be preferable for interpretable predictions. Additionally, computational resources must be considered, especially for training resource-intensive algorithms like Random Forest and Gradient Boosting Machine. To sum up, this review paper highlights machine learning's effectiveness in loan default prediction and stresses the importance of tailored algorithm selection based on dataset characteristics and interpretive requirements. Financial institutions can optimize risk assessment and decision-making processes by strategically evaluating and choosing algorithms to suit their operational needs.

REFERENCES

[1] Aslam, Uzair, et al. "An empirical study on loan default prediction models." Journal of Computational and Theoretical Nanoscience 16.8 (2019): 3483-3488.
[2] Li, Yu. "Credit risk prediction based on machine learning methods." 2019 14th international conference on computer science & education (ICCSE). IEEE, 2019.
[3] Ahmed, MS Irfan, and P. Ramila Rajaleximi. "An empirical study on credit scoring and credit scorecard for financialinstitutions." International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) 8.7 (2019): 2278-1323.
[4] Zhu, Lin, et al. "A study on predicting loan default based on the random forest algorithm." Procedia Computer Science 162 (2019): 503-513.
[5] Ghatasheh, Nazeeh. "Business analytics using random forest trees for credit risk prediction: a comparison study." International Journal of Advanced Science and Technology 72.2014 (2014): 19-30.
[6] Breeden, Joseph. "A survey of machine learning in credit risk." Journal of Credit Risk 17.3 (2021).
[7] Madane, Nikhil, and Siddharth Nanda. "Loan prediction analysis using decision tree." Journal of the Gujarat Research society 21.14 (2019): 214-221.
[8] Supriya, Pidikiti, et al. "Loan prediction by using machine learning models." International Journal of Engineering and Techniques 5.2 (2019): 144-147.
[9] Amin, Rafik Khairul, and Yuliant Sibaroni. "Implementation of decision tree using C4. 5 algorithm in decision making of loan application by debtor (Case study: Bank pasar of Yogyakarta Special Region)." 2015 3rd International Conference on Information and Communication Technology (ICoICT). IEEE, 2015.
[10] Jency, X. Francis, V. P. Sumathi, and Janani Shiva Sri. "An exploratory data analysis for loan prediction based on nature of the clients." International Journal of Recent Technology and Engineering (IJRTE) 7.4 (2018): 17-23.
[11] Shoumo S Z H, Dhruba M I M, Hossain S, Ghani N H, Arif H and Islam S 2019 "Application of machine learning in credit risk assessment: a prelude to smart banking" TENCON 2019 – 2019 IEEE Region 10 Conf.
(TENCON) pp 2023–8
[12] Addo P M, Guegan D and Hassani B 2018 "Credit risk analysis using machine and deep learning models" Risks
6 p 38
[13] Hamid A J and Ahmed T M 2016, " Developing prediction model of loan risk in banks using data mining Machine Learning and Applications: An Int. Journal" (MLAIJ) 3 pp 1–9
[14] Kacheria A, Shivakumar N, Sawkar S and Gupta "A 2016 Loan sanctioning prediction system" Int. Journal of Soft Computing and Engineering (IJSCE) 6 pp 50–3
[15] Vojtek M and Kocenda E 2006 "Credit scoring methods Finance a uver" - Czech Journal of Economics and Finance 56 pp 152–167
[16] Russel S and Norvig P 1995 "Artificial intelligence - a modern approach"
[17] Alshouiliy K, Alghamdi A and Agrawal D P 2020 "AzureML based analysis and prediction loan borrowers creditworthy" The 3rd Int. Conf. on Information and Computer Technologies (ICICT) 1 pp 302–6
[18] Li M, Mickel A and Taylor S 2018, "Should this loan be approved or denied?": a large dataset with class assignment guidelines Journal of Statistics Education

26 pp 55–66
[19] Vaidya A 2017 “Predictive and probabilistic approach using logistic regression: application to prediction of loan approval” The 8th Int. Conf. on Computing, Communication and Networking Technologies (ICCCNT) 1 pp 1–6
[20] Murphy K P 2012 “Machine learning: a probabilistic approach”



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY