An Ensemble Classifier-Based Approach for Feature Selection and Prediction of Loan Approval

Narendra Singh Ranawat [1]

Abstract

Purpose : The current research addressed the challenges confronted by banks and financial institutions in processing loan requests. With an increase in the number of loans being requested, the risk of non-payment also increases. Thus, an accurate and reliable method was needed that can evaluate the eligibility of the applicant and lessen limit losses. Therefore, this study aimed to implement machine learning (ML) methods to anticipate loan approvals and diagnose potential defaulters.

Methodology : An ensemble-based ML method that uses grid search-based optimized classifiers was proposed. Four different classifiers, namely LightGBM, XGBoost, CatBoost, and AdaBoost, were used in this study. At first, 12 features were utilized to train these models, and the models established their order importance. These features were arranged in descending order of importance, and feature sensitivity analysis was done to find the most optimal set of features for the particular classifier. The performance of the model was compared based on accuracy and F1 score.

Findings : The results depicted that all three ensemble classifiers except Adaboost achieved an accuracy of ~98% trained using only six best features. Furthermore, Catboost outperformed all other models in the prediction of loan approvals with an accuracy and F1 score of 98.68% and 92.632%, respectively. This reflected the importance of selecting optimal features and ensemble learning, which improves the model performance and mitigates the computational complexity.

Practical Implications : The findings recommended that utilizing ML algorithms can improve the loan approval method by reducing the risk of operating inefficiencies and defaults. They also highlighted the necessity of selecting the optimal features to make reliable ML models. The future scope of this research lies in exploring additional data sources, such as transactional and behavioral data, to improve the model's effectiveness.

Originality : The traditional methods for loan approval rely on subjective assessments and manual judgments. This study provides a data-driven method that can be automated with an improved decision-making process. Furthermore, this study used feature ranking methodology and ensemble classifiers that contribute to different aspects of loan analysis, offering a scalable and efficient alternative to banks and financial institutions.

Keywords : **loan approval, machine learning, feature ranking, LightGBM, XGBoost, CatBoost, AdaBoost**

JEL Classification Codes : A1, C45, G21, G24, G32, N2

Paper Submission Date : July 5, 2024 ; Paper sent back for Revision : February 25, 2025 ; Paper Acceptance Date : March 10, 2025 ; Paper Published Online : April 15, 2025

Banks and financial institutions play a key role in the economy. They act as intermediaries between borrowers and investors and even provide financial services to consumers, businesses, and the government (Mohapatra et al., 2015). Lending is among the vital solutions available for economic growth as it permits investment in housing, corporate expansion, education, and other otherwise impossible projects without borrowing (Kumar & Singh, 2017). Ranawat and Chakraborty (2023) observed that the major income source of banks is derived from the loans they offer to their customers. Thus, the distribution of the loan to the

[1] *Manager (Corresponding Author),* Shree Rameshwaram Minerals, Udaipur - 313 022, Rajasthan.
(Email : narendraranawat@outlook.com) ; ORCID iD : https://orcid.org/0000-0002-1973-9775

DOI : https://doi.org/10.17010/ijf/2025/v19i4/174947

customers is the key element in the bank's business model. These banks charge interest on the loans they extend. However, the loan approval process is complex and fraught with monetary risks, as banks must thoroughly evaluate the creditworthiness of applicants to minimize defaults (Chandra & Awasthi, 2019). Banks have established stringent guidelines and rigorous assessment procedures, but these do not guarantee that eligible applicants will repay their loans based on the assessment, which can lead to significant financial losses for the banks. This highlights the necessity for more effective and accurate systems to evaluate loan eligibility, thereby reducing risks (Kaur & Pasricha, 2019).

The current loan approval process is mainly based on traditional methods that rely on manual assessments and established lending procedures. These procedures are lengthy and are also susceptible to bias, human error, and lack of efficiency (Shukla, 2016). The bank reviews the applicants based on income or credit history as well as collateral. Yet, these approaches frequently fail to recognize the risk of default (Gupta & Jaiswal, 2020). Moreover, the substantial rise in loan applications further complicates the process. It creates the need to find an alternative solution using modern technology to improve the decision-making process (Nag & Shah, 2022). Past researchers have examined various aspects of loan forecasting; however, a significant gap remains in utilizing machine learning (ML) that is scalable to automate and improve this process with better performance and security (Wang et al., 2023).

This study seeks to bridge the gap by implementing an ML-based approach for predicting loan approval and identifying meritorious loan applicants. The study evaluates four different ensemble classifiers: Light Gradient Boosting Machine (LightGBM), eXtreme Gradient Boosting (XGBoost), CatBoost, and Adaptive Boosting (AdaBoost). The parameters of these classifiers are tuned using a grid-based search method. These classifiers are then used to rank the features according to their importance to select relevant features. This research aims to build a solid model that will help banks make well-informed lending decisions. The study focuses on how ML methods can effectively be used to improve the efficiency and accuracy of lending approval processes.

The potential contribution of this study lies in its innovative approach of integrating the feature selection process with the classifier for loan prediction. By achieving an extremely high level of accuracy and identifying the most critical factors that influence the approval of loans, the proposed approach can drastically lower the risk of operating inefficiencies and defaults for financial institutions and banks. In addition, this research will lay the foundation for further research to explore the use of other sources of data and sophisticated ML methods to improve loan prediction models.

Literature Review

In this section, various research articles on loan prediction are discussed. It explores several data mining techniques for loan prediction, utilizing multiple technologies. It also reviews research publications that explain the lowest necessary parameter for prediction. It recognized risk assessment and entropy as critical challenges in financial organizations like the banking system.

Predicting loan defaulters is one of the most challenging issues for any bank. On the other hand, banks can significantly reduce their losses by accurately forecasting loan defaulters (Nagaraju, 2014). Therefore, it became imperative to do a study on loan approval prediction. Abdelmoula (2015) modeled the short-term loan default risk for a Tunisian bank to assess the credit risk associated with the bank. Between 2003 and 2006, he utilized a database containing 924 credit files. The K-Nearest Neighbor (K-NN) classifier technique was used in the evaluation, and he experimented with several values of k (2, 3, 4, and 5). The reduction of the alarming risk rate served as the performance criterion. He constructed three different K-NN classifier types: the non-cash-flow model, the cash-flow model, and the complete information model. The primary findings indicate that, among the three models, the optimal K-NN with $k = 3$ achieved the best global classification rate, approximately 88.63%

(second-best classifier). Additionally, ROC curves were generated to assess the model's performance. The outcome demonstrates that the 95.6% range was met by the area under curve (AUC) criterion. However, this study does not reveal how these findings can be applied to Tunisia's implementation of Basel II or III.

Arun et al. (2016) focused on minimizing the workload of bank workers by proposing a model that utilizes various ML algorithms and identifies the approaches that yield accurate results. They employed a min–max normalization or the K-NN algorithm, which predicted loans with an accuracy of 75.08%. Madane and Nanda (2019) built a model using the decision tree approach to examine application requirements and the credit score of mortgage loans. Using a model they developed to determine the safety of loan sanctioning, they found that most low-income applicants had their loans authorized, as they had a higher repayment rate. This model achieved an accuracy of 85%. Supriya et al. (2019) developed a well-designed model to reduce the risk factors in selecting a reliable candidate for a loan, thereby saving the bank time and money. They utilized gradient boost-decision tree regression, logistic regression, and classification to predict the loans. The decision tree, with an accuracy of 82%, proved to be the best-performing model.

Tejaswini et al. (2020) proposed a reliable modeling method that decides to approve or reject a loan based on the applicant's previous credit and financial rating. They aimed to develop an efficient method for selecting the most suitable applicants. Three different ML techniques were used in this study to determine the likelihood of loan acceptance for customers: decision tree, logistic regression, and random forest. Compared to other ML algorithms, the decision tree ML approach has an accuracy of 82%. Behera et al. (2021) proposed a method to predict whether an applicant can repay the loan. They utilized the logistic regression method, which achieved an accuracy of 86% and an F1 score of 90%. Six different ML algorithms were evaluated by Park et al. (2021) based on accuracy, precision, and recall to predict whether a loan would be approved or not. In their investigation, random forest had the highest accuracy of 95.55%, whereas logistic regression had the lowest.

Al Mamun et al. (2022) employed various ML techniques, including K-NN, random forest, XGBoost, AdaBoost, LightGBM, and Decision Tree, to identify the most significant feature. Standard measures were used to compare and evaluate the aforementioned algorithms against one another. With an accuracy of 92%, logistic regression was the most accurate. Additionally, it was demonstrated to be the most effective model, outperforming other ML techniques by a substantial margin. Kumari et al. (2023) utilized seven ML algorithms: Linear

Table 1. ***Comparative Studies Amid Past Literature Regarding Bank Loan Approval Prediction Using Various ML Approaches***

Authors	Proposed Models	Classifiers' Performance
Abdelmoula (2015)	K-Nearest Neighbor (K-NN)	Accuracy: 88.63%
Arun et al. (2016)	K-NN	Accuracy: 75.08%
Madane & Nanda (2019)	Decision Tree	Accuracy: 85%
Supriya et al. (2019)	Logic regression, Decision Tree, and Gradient Boosting	Accuracy: 82%
Tejaswini et al. (2020)	Logistic Regression, Decision Tree, and Random Forest	Accuracy: 82%
		F1-Score: 75%
Park et al. (2021)	Random Forest and Logistic Regression	Accuracy: 95.55%
Behera et al. (2021)	Logistic Regression	Accuracy: 86%
		F1-Score: 90%
Al Mamun et al. (2022)	Random Forest, XGBoost, AdaBoost, K-NN, LightGBM, and Decision Tree	Accuracy: 92%
Kumari et al. (2023)	Linear Regression, K-NN, Logistic Regression, Decision Tree, SVM, and Naïve Bayes	Accuracy: 90.71%
Viswanatha et al. (2023)	Random Forest, Naïve Bayes, Decision Tree, and K-NN algorithms	Accuracy: 83.73%

Regression, K-NN, Logistic Regression, Decision Tree, SVM, and Naïve Bayes. They assessed the usefulness of these ML techniques in terms of accuracy. Random forest classifier outperformed all six other models, scoring the highest accuracy of 90.71%, followed by Naïve Bayes, which achieved an accuracy of 85.96%, whereas linear regression had the lowest accuracy. Viswanatha et al. (2023) proposed combining ML models and ensemble learning approaches to determine the probability of accepting individual loan requests. They employed four algorithms to predict a customer's loan approval status accuracy: Random Forest, Naïve Bayes, Decision Tree, and K-NN. With the Naïve Bayes algorithm as the best, they achieved an accuracy of 83.73%. Table 1 presents a comparative review of past literature on loan approval prediction using various ML methods. There is a need to increase the accuracy in predicting loan approval in the banking sector (Narayanaswamy & Muthulakshmi, 2016). Therefore, this paper presents a comparative study of various ensemble classifiers utilizing ML to enhance accuracy in contrast to the models above.

Methodology and Data Sources

In banking, loan eligibility approval systems can benefit the bank and its customers (Budhedeo & Pandya, 2018). Because ML techniques are efficient and time-saving, they can simplify the entire process. For the prediction of loan approval, this study evaluates four distinct ensemble classifiers — LightGBM, XGBoost, CatBoost, and AdaBoost. Furthermore, the importance of the attributes (features) in training these models is computed and utilized to rank these features. A systematic methodology for developing, training, and testing a predictive model to predict loan approval is illustrated in Figure 1.

Figure 1. ***Research Methodology Flowchart***

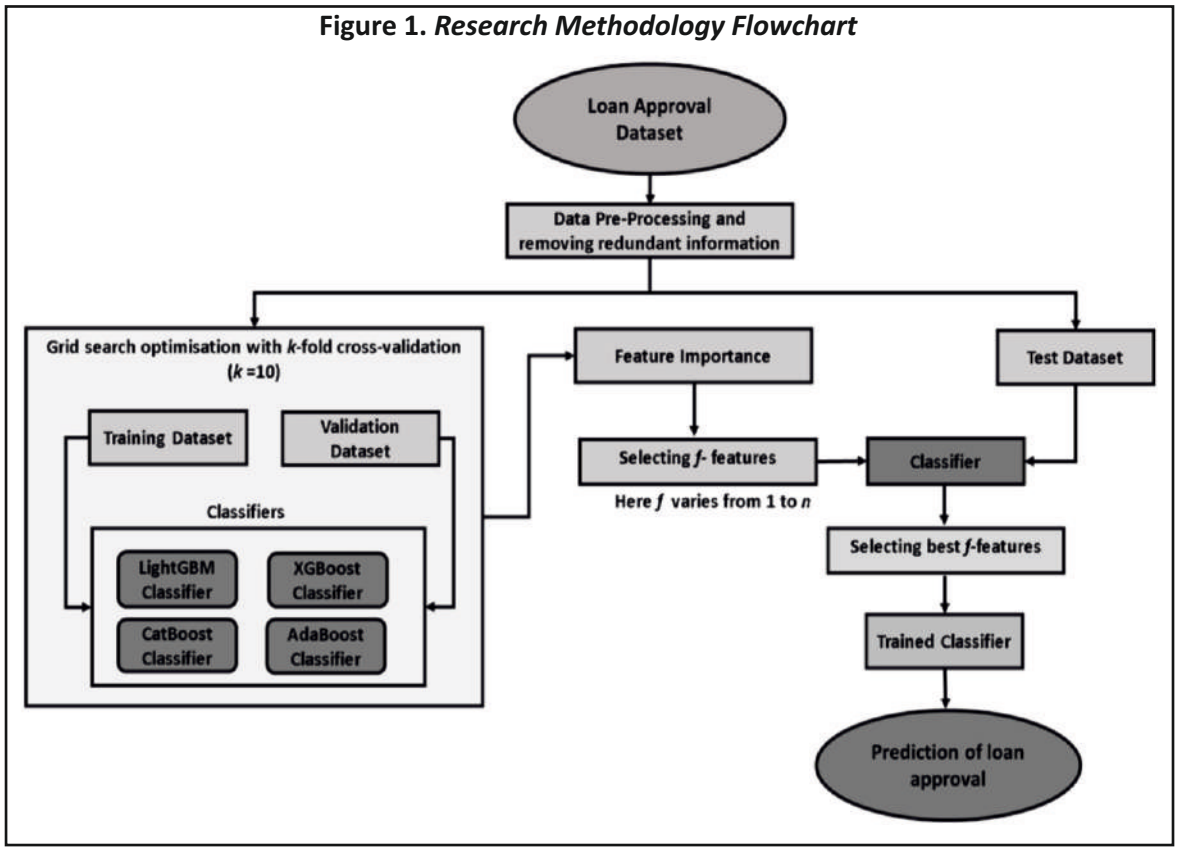


Table 2. *Dataset Description*

Name of Features (*f*)	Description of Features	Data Type
ID*	Customer ID	Integer
Age	Age of customer	Integer
Experience	Customer's experience	Integer
Income	Income of the customer	Integer
ZIP Code	Customer's residence Zip code/Location	Integer
Family	Number of family members of the customer	Integer
CC Avg.	Credit card average score	Integer
Education	Education of the customer	String
Mortgage	Mortgage taken or not taken by the customer (Yes/No)	Character
Securities Account	Having or not having a Securities Account (Yes/No)	Character
CD Account	Having or not having a certificate of deposit account (Yes/No)	Character
Online Banking	Having or not having online banking (Yes/No)	Character
Credit Card	Having or not having a credit card (Yes/No)	Character
Personal Loan**	0 = No personal loan given, 1 = personal loan given	String

Note. ** Target variable cannot be assigned while ranking features.

* These features were removed during data cleaning.

Dataset Collection

The dataset used in this study was compiled from the Kaggle repository, a publicly accessible source. There are a total of 5,000 instances comprising loan approvals and loan rejections. A total of 13 attributes were used to determine the approval of the loan. In this study, predicting the loan approval is considered the target class. Thus, the attributes must be selected judiciously to characterize the eligibility for loan approval. Table 2 shows all the features with their descriptions.

Dataset Pre-Processing

Data pre-processing was accomplished through data cleaning and transformation of categorical variables. It was performed to normalize the data and get all features into numerical form. For instance, the target variable "Personal Loan" was previously in categorical (string) form, and it becomes essential to convert it into numerical (integer) form to perform ML algorithms over it. Therefore, one-hot coding was used to convert them into numerical form (0 = No personal loan given, 1 = personal loan given).

Ensemble Classifiers

Ensemble classifiers combine the predictions of different models used to enhance overall performance. These approaches leverage the diversity of multiple models for their unique strengths to compensate for their weaknesses rather than relying on one model (Cozarenco & Szafarz, 2018). Ensemble classifiers provide more precise and reliable results by combining the predictions of several models, especially when one model struggles to understand specific aspects of the datasets (Beade et al., 2024). Ensemble methods excel in generalizing patterns in data. The variety of models allows them to capture a wider array of attributes, thus improving the ability

of models to adapt to different datasets (Liu et al., 2022). This study uses four different classifiers, namely LightGBM, XGBoost, CatBoost, and AdaBoost.

↳ ***LightGBM :*** LightGBM is an ensemble classifier of different weak learners, mainly decision trees, using the boosting method to create an overall robust model. It is built on the framework of gradient boosting. Unlike other tree-based algorithms that expand tree level-wise, it grows leaf-wise instead. Therefore, it uses a histogram-based decision tree that is highly optimized, unlike XGboost, which uses a sorted-based decision tree algorithm. It improves the efficiency of the model, and the consumption of the system's memory is also low. It runs faster while maintaining a high level of accuracy as it uses gradient-based one-sided sampling and exclusive feature bundling. It can be effectively used for regression tasks, classification tasks, and even for computing feature importance on large datasets (Rodney & Mincey, 2020).

↳ ***XGBoost :*** The XGBoost also ensembles different weak learners, comprising a decision tree to create an overall strong learner. The weak learners are taught in a sequence, each one focusing on fixing mistakes made by previous ones. Regularization techniques are introduced in XGBoost that reduce the risk of overfitting (Ranawat, Prakash, Miglani et al., 2023). XGBoost adjusts the model parameters while training using a second-order gradient-based optimization method. This technique for optimization helps speed up the process of convergence during the training process. For each feature of the data set, XGBoost assigns a feature importance score. The score reflects the importance of each component of the feature to the prediction model. It helps to deepen the understanding of various aspects that affect the decision-making ability of the model.

↳ ***CatBoost :*** Similar to LightGBM and XGBoost, CatBoost is also built on a gradient-boosting framework. Several weak learners, typically the decision tree, can be joined to form a powerful ensemble model. One of CatBoost's strengths is its capacity to handle categorical characteristics without requiring lengthy pre-processing procedures like one-hot encryption. CatBoost makes training data preparation easier by efficiently managing categories of variables (Wilson, 2016). During training, it will automate the removal of missing values from the data set, removing the need to compute manually. It comes with built-in cross-validation that trains and tests the model using different parts of the data to assess its performance more precisely (Wu, 2022). It also includes a vital feature measure that allows for understanding how each feature influences the models' predictions.

↳ ***AdaBoost :*** The AdaBoost is an effective ensemble learning technique that helps build a strong model by combining weak learners (Pandey et al., 2021). Selecting weak learners to balance the model's complexity and ability to adapt is vital. Adjusting the weights of an instance in response to errors in classification changes them during training. Emphasizing cases that have been misclassified frequently starts with equal weights and enhances efficiency. A weighted mix of learner predictions generates an end-to-end model. AdaBoost can withstand the test of time because of its flexibility and focus on dealing with complex situations. However, it could become sensitive to extreme outliers.

Grid search optimization is a tuning of hyperparameters method that determines a model's optimal range of hyperparameters. Hyperparameters are the settings that define the model, which is not learned from data; however, they must be set prior to learning (Prakash & Kankar, 2020). Grid search optimization based on cross-validation *k*-fold tunes the hyperparameters used in the above-mentioned ML models. It assists in determining the best combination of hyperparameter values to provide the most excellent performance. In this case, *k*-fold cross-validation (k =10) is used to evaluate the model's performance across multiple subsets of the training data. Table 3 presents the hyperparameters various classifiers use and the corresponding parameters space.

Table 3. ***Hyperparameters and Parameter Space of Respective Classifier***

Hyperparameters	Parameter Space of Ensembled Classifiers			
	LightGBM	XGBoost	CatBoost	AdaBoost
Num_leaves	[31, 50, 100]	[31, 50, 100]	[31, 50, 100]	–
Learning_rate	[0.05, 0.1, 0.2]	[0.05, 0.1, 0.2]	[0.05, 0.1, 0.2]	[0.01, 0.1, 0.5, 1]
n_estimators	[50, 100, 200]	[50, 100, 200]	[50, 100, 200]	[50, 100, 200]

Feature Importance

Ensemble classifiers evaluate feature importance by assessing the contribution of each feature to the model's overall performance. For LightGBM and XGBoost, feature importance is often measured using coverage, gain, split value, and split gain metrics. Here, gain represents the improvement in accuracy brought about by a feature, while coverage measures the relative quantity of observations affected by a feature (Ranawat, Miglani et al., 2023). LightGBM presents the split value and gains to provide a more detailed understanding of the importance of features. Additionally, split value and gain are crucial in determining the structure of the decision tree (Chen et al., 2024). The algorithms look for feature space splits that maximize the model's potential for improving prediction accuracy. CatBoost, on the other hand, calculates importance by considering the contribution of each feature during the training process, focusing on how much a feature contributes to the model's performance improvement (Hancock & Khoshgoftaar, 2020). AdaBoost assigns weights to features based on their impact on accuracy in weak learners, emphasizing features that consistently help correct model errors (Shahraki et al., 2020).

Selection of f - Feature Based on Feature Importance

The importance of the features is measured based on their contribution to training the respective ensemble models. Then, the feature vector is modified, and the features are arranged in descending order of importance. For selecting the best features, the ensemble classifiers are again trained on *f* features where *f* varies from 1 to *n* in an incremental order. Here, $f = 1$ signifies that the model is trained only on the first feature from the feature vector. Similarly, $f = 5$ signifies that it is trained on the top five features from the modified feature vector, whereas $f = n$ signifies that the model is trained on all features. In other words, the model is trained on *f* features *n* times and evaluated based on performance. At last, the best *"f"* features are selected at the instance where the model achieves the best accuracy, and the remaining features are discarded.

Analysis and Results

The primary objective of this study is to identify the suitable features necessary for training the ML model to predict loan approval. Features ranked by four different classifiers are used to train the respective ML model to achieve the objective. The ranked features are used to train the model using a training set, and validation is performed on the test set to evaluate the performance of feature-ranking techniques. Thus, as discussed in the section above, the entire dataset is divided into training (80%) and testing (20%) data. This section presents the results of various feature selection techniques applied to the test data.

Grid search optimization was employed with 10-fold cross-validation ($k = 10$) to train and optimize an ensemble classifier to improve the model's performance by systematically exploring various hyperparameter combinations. The predictions from multiple base classifiers were combined into an ensemble classifier. Grid search was utilized to fine-tune the hyperparameters, substantially impacting the model's behavior. Therefore, a

grid of hyperparameter values is specified, and the model is then trained and assessed for every possible combination of these values across all datasets. The model is trained 10 times once the dataset is split into 10 folds. Onefold was used as the test set, and the remaining folds were used as the training set for each iteration. Ten iterations of this technique are carried out, using a different fold as the test set each time. The best model from the ensemble classifier is trained and optimized using a grid search and tenfold cross-validation, and then its performance is evaluated.

Thereafter, the features are arranged in descending order of importance scores to reorder the feature vector. The best-trained ensemble model is then used to get feature importance scores, as shown in Figure 2. The features are then sorted in descending order of importance scores, so the more essential features appear first in the updated feature vector.

Figure 2. ***Feature Importance of (a) LightGBM, (b) XGBoost, (c) CatBoost, and (d) AdaBoost Classifiers***

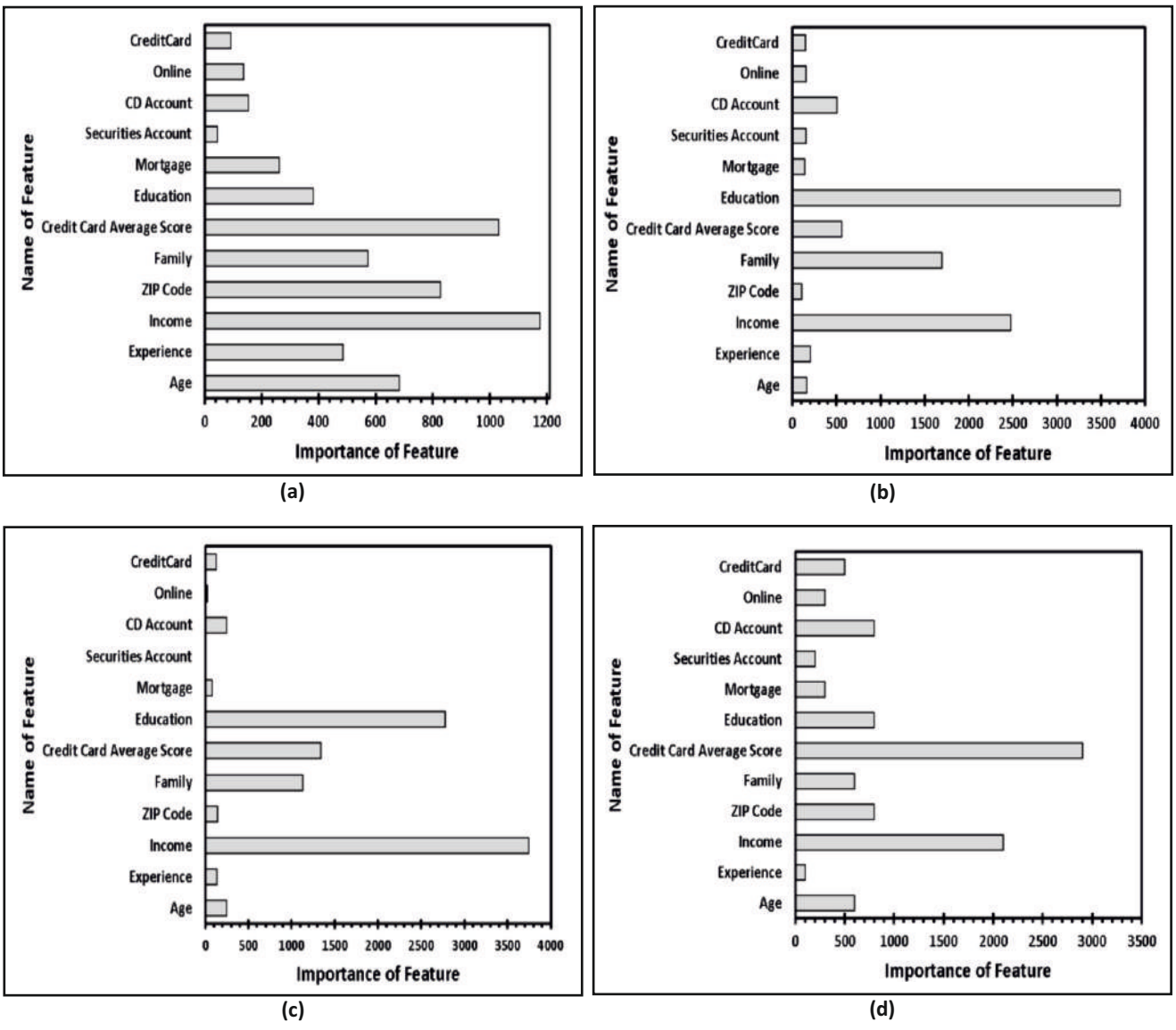


Table 4. ***Comparative Performance Evaluation of Classifiers***

	LightGBM		XGBoost		CatBoost		AdaBoost	
Features (f)	***Accuracy***	***F1-Score***	***Accuracy***	***F1-Score***	***Accuracy***	***F1-Score***	***Accuracy***	***F1-Score***
f 1	90.40%	35.37%	90.27%	34.17%	90.58%	31.50%	91.20%	38.37%
f 2	92.15%	54.32%	92.10%	53.22%	91.85%	48.32%	91.13%	45.38%
f 3	92.58%	56.85%	92.32%	55.61%	92.70%	54.28%	91.80%	48.90%
f 4	92.35%	55.83%	92.40%	55.31%	92.48%	52.60%	91.98%	51.00%
f 5	96.65%	80.88%	96.72%	81.19%	96.85%	81.47%	95.98%	76.51%
f 6	98.43%	91.17%	98.53%	91.79%	**98.68%**	**92.63%**	96.65%	80.00%
f 7	98.60%	92.15%	98.38%	91.02%	98.48%	91.45%	96.62%	79.87%
f 8	98.43%	91.20%	98.48%	91.51%	98.55%	91.89%	96.68%	80.13%
f 9	98.20%	89.94%	98.33%	90.74%	98.53%	91.70%	96.43%	78.89%
f 10	98.25%	90.17%	98.30%	90.60%	98.30%	90.45%	96.40%	79.05%
f 11	98.40%	91.11%	98.40%	91.14%	98.40%	90.98%	96.28%	78.37%
f 12	98.48%	91.46%	98.40%	91.13%	98.38%	90.87%	96.48%	79.70%

Note. CatBoost outperformed other classifiers, achieving a maximum accuracy of 98.68% and an F1-score of 92.63% for predicting loan approval.

After arranging features based on their importance, a feature selection process is by iteratively training models with different subsets of these features, starting from the most essential feature to the least important one; in each iteration, the accuracy and F1 score of the model is recorded to assess the change in performance of the model based on the feature's inclusion or exclusion. This technique is also known as "feature importance-based feature selection." Table 4 shows that the four classifiers—LightGBM, XGBoost, CatBoost, and AdaBoost—performed well with the top six features only. The top six features for LightGBM are income, credit card average score, location, age, family, and customer experience. Similarly, the top six features for XGBoost are education, income, family, average credit card scores, CD accounts, and customer experience. The top six features for CatBoost are income, education, average credit card score, family status, CD account, and customer age. The top six features for AdaBoost are credit card average score, income, education, CD account, and customer location. However, CatBoost outperformed other classifiers, achieving a maximum accuracy of 98.68% and an F1-score of 92.63% for predicting loan approval.

CatBoost can work directly with categorical variables, so it is more convenient than XGBoost, which often requires one-hot encoding. Additionally, it features a built-in mechanism to address overfitting, which can sometimes lead to improved generalization on specific datasets compared to XGBoost. Correspondingly, CatBoost also helps automate hyperparameter tuning, making finding the optimal set of hyperparameters for a particular dataset easier. It validates CatBoost's best performance compared to other classifiers. Here, Figures 3 and 4 present the graphical representation of accuracy and the F1-score for each classifier.

Figures 5(b) and (c) of the confusion matrix show that the XGBoost and CatBoost classifiers performed accurately, as the number of instances misclassified is equal, i.e., nine. Additionally, LightGBM performed decently, misclassifying only 10 cases (Figure 5(a)). However, AdaBoost did not perform satisfactorily in accurately predicting instances, misclassifying 35 instances for loan approval (Figure 5(d)).

Figure 3. ***Accuracy of Each Classifier for the Number of f-Features***

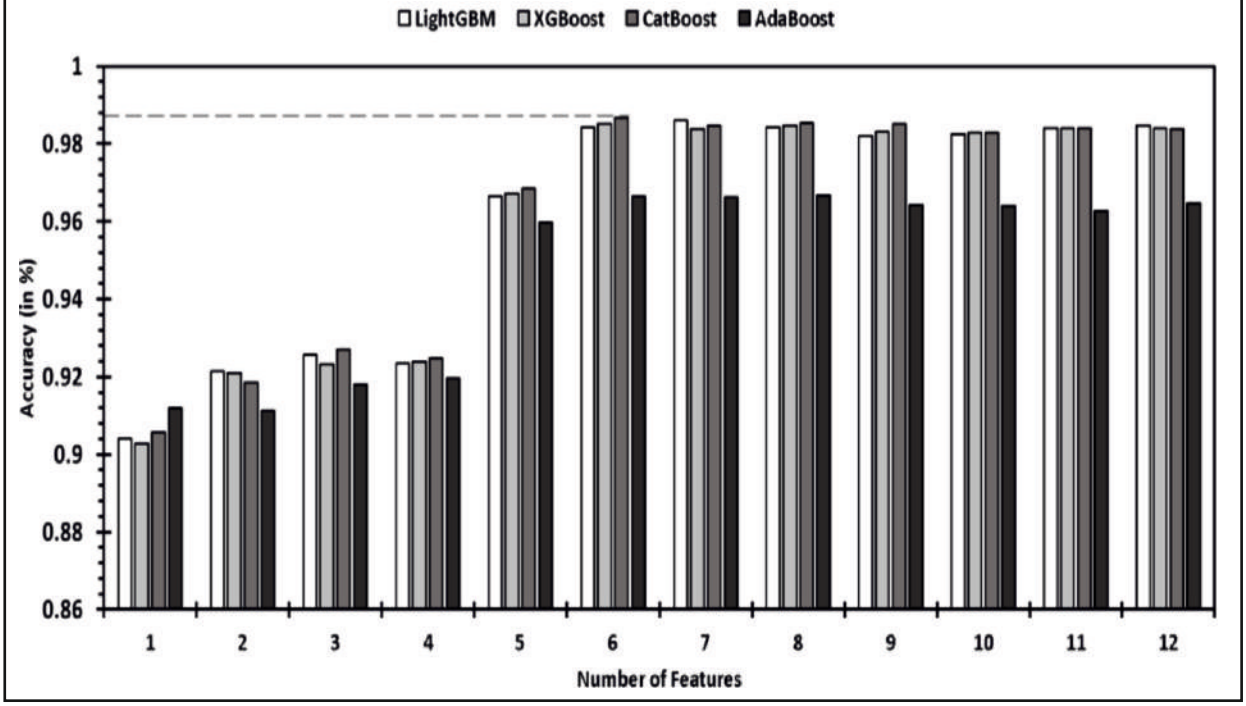


Figure 4. ***F1 Score of Each Classifier for Numbers of f-Features***

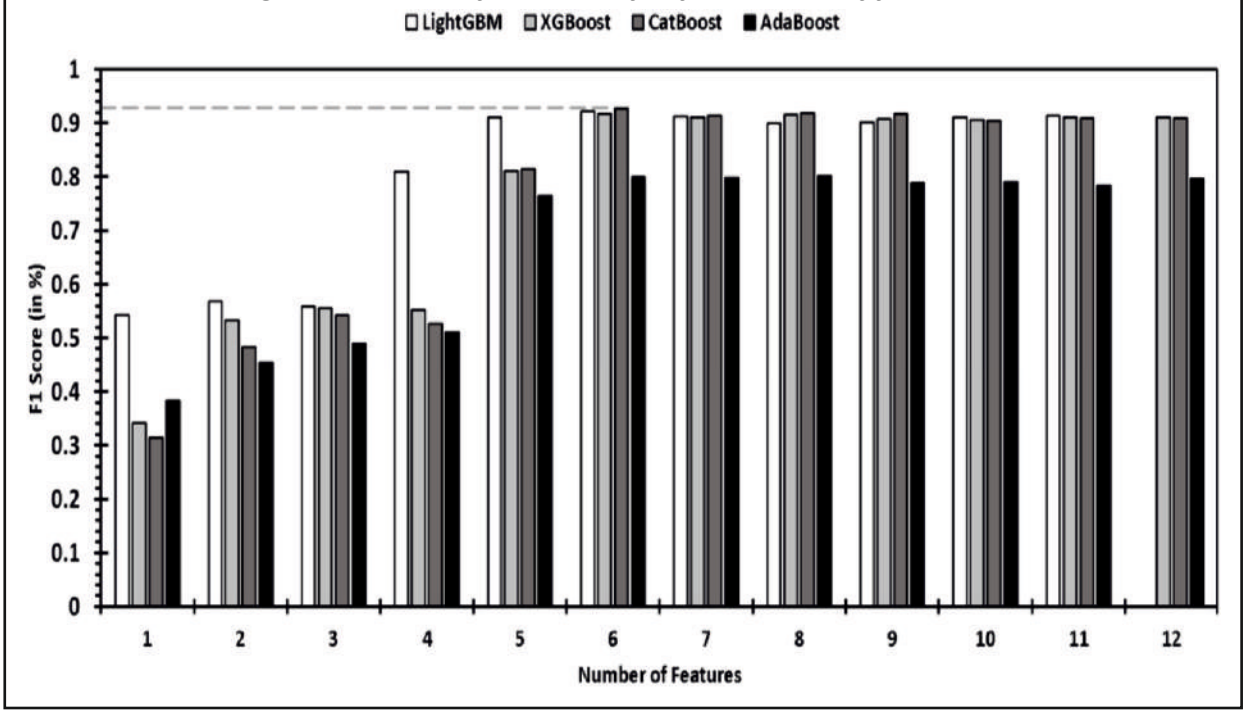
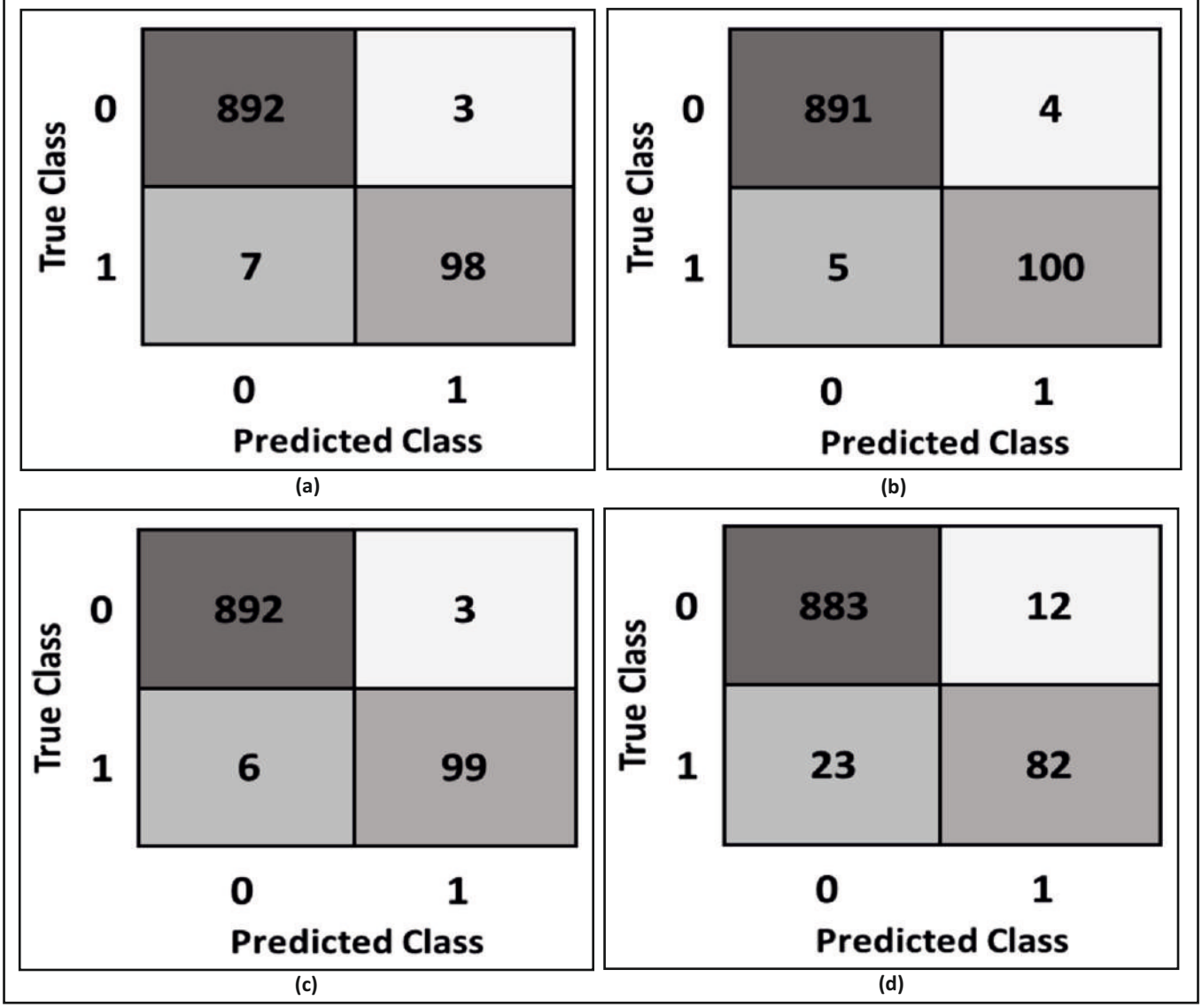


Figure 5. ***Confusion Matrix for (a) LightGBM, (b) XGBoost, (c) CatBoost, and (d) AdaBoost Classifiers***



Implications

Managerial Implications

This study provides banking institutions with a robust, data-driven framework for loan approval, achieving ~98% accuracy using ensemble classifiers, such as CatBoost. Banks can minimize defaults, enhance decision-making, and improve customer trust by identifying critical features and reducing computational complexity. The approach enhances operational efficiency, enabling managers to balance risk and growth opportunities.

Theoretical Implications

The research advances the literature by demonstrating the superiority of ensemble classifiers in loan prediction,

achieving 98.68% accuracy and a 92.63% F1 score. It introduces a novel approach to feature selection and incremental training, challenging traditional models by achieving high accuracy with fewer features. This study also emphasizes the importance of selecting relevant features in the ML model. This research provides the foundation for further research into ensemble models and using other data sources, including transactional and behavioral data.

Policy Implications

The research findings support utilizing data-driven methodologies that are fair, transparent, unbiased, and align with the regulatory requirements for ethical lending. This research can serve as a baseline for the policymaker to develop procedures and guidelines to implement fair and responsible use of artificial intelligence (AI) in finance, ensuring compliance and promoting financial inclusion. Moreover, collaboration between the industry and academia can refine these models to attain global best practices.

Summary and Conclusion

Loan approval is a critical challenge for banking institutions due to the associated potential risk in lending money. This risk of default can be lessened if the creditworthiness of loan applicants is assessed accurately by the banks at the time of loan disbursement. If banks approve the loan to the applicant and later fail to repay, it can lead to financial losses for the bank. On the other hand, rejecting the loan to creditworthy applicants can result in lost business opportunities for the bank. Therefore, maintaining the financial solidity and strength of the banking institution requires a prominent balance between granting loans to avoid risky borrowers and approving loans to deserving candidates. ML models, particularly ensemble classifier-based approaches, could aid banks by predicting the likelihood of loan approvals using various features related to the applicants. Since ensemble approaches can efficiently manage data with imbalances and ensure that the model is not biased toward most applicants, they also provide the ability to interpret by delivering a feature importance score. It helps in understanding which *f-features* are critical in making loan approval decisions. Therefore, this paper presents a comparative study of four different ensemble classifiers—LightGBM, XGBoost, CatBoost, and AdaBoost—to enhance the accuracy of loan approval processes. This ML model achieves a maximum accuracy of 98.68% and an F1-Score of 92.63% in predicting loan approval using only six features.

Limitations of the Study and Scope for Further Research

The study demonstrates significant improvements in loan approval prediction using ML, achieving ~98% accuracy with only six features, which highlights the effectiveness of ensemble classifiers, such as CatBoost. However, some limitations exist, including dependence on a specific dataset with 12 features, which may not encompass all relevant factors, such as behavioral or transactional data. Here, transactional data pertains to the purchases and transfers conducted by customers. On the other hand, behavioral data examines how customers engage with banking services, such as app usage and website visits, and this type of data also offers insights into their preferences and behaviors beyond mere transactions. Although the models demonstrate high accuracy, enhancing their interpretability could improve transparency for banks and financial institutions. Future research could focus on addressing these gaps by incorporating additional data sources, such as behavioral and transactional information, to enhance predictive capabilities and additionally, developing explainable AI techniques to improve model interpretability and fairness, optimizing computational efficiency for real-time use,

and examining how external factors, such as economic downturns, on model performance are promising areas for exploration. Expanding the methodology to other financial services, like credit scoring or insurance underwriting, and creating ethical and regulatory frameworks for AI-driven loan approvals could also enhance its relevance. These advancements would build on the study's strengths, offering more robust, fair, and scalable solutions for the banking sector.

Author's Contribution

Mr. Narendra Singh Ranawat has authored a comprehensive research paper titled "An Ensemble Classifier-based Approach for Feature Selection and Prediction of Loan Approval." It involved conceiving the idea and developing the design for an empirical study. He conducted a thorough literature review by extracting reputable research papers and filtering them based on relevant keywords. He generated concepts and codes from these papers pertinent to the study design. Additionally, he verified the analytical methods used in the study and oversaw its execution.

Conflict of Interest

The author certifies that he has no affiliations with or involvement in any organization or entity with any financial or non-financial interest in the subject matter or materials discussed in this manuscript.

Funding Acknowledgment

The author received no financial support for this article's research, authorship, and/or publication.

References

Abdelmoula, A. K. (2015). Bank credit risk analysis with k-nearest-neighbor classifier: Case of Tunisian banks. *Accounting and Management Information Systems*, *14*(1), 79–106. https://online-cig.ase.ro/jcig/art/14_1_4.pdf

Al Mamun, M., Farjana, A., & Mamun, M. (2022). Predicting bank loan eligibility using machine learning models and comparison analysis. In *Proceedings of the 7th North American International Conference on Industrial Engineering and Operations Management*. IEOM Society International. https://doi.org/10.46254/NA07.20220328

Arun, K., Ishan, G., & Sanmeet, K. (2016). Loan approval prediction based on machine learning approach. *IOSR Journal of Computer Engineering*, *18*(3), 79–81. https://doi.org/10.9790/0661-1803017981

Beade, Á., Rodríguez, M., & Santos, J. (2024). Multiperiod bankruptcy prediction models with interpretable single models. *Computational Economics*, *64*(3), 1357–1390. https://doi.org/10.1007/s10614-023-10479-z

Behera, A. P., Rautaray, S. S., Pandey, M., & Gourisaria, M. K. (2021). Predictive loan approval model using logistic regression. In S. C. Satapathy, V. Bhateja, M. N. Favorskaya, & T. Adilakshmi (eds.), *Smart computing techniques and applications. Smart innovation, systems and technologies* (Vol. 225, pp. 715–722). Springer. https://doi.org/10.1007/978-981-16-0878-0_69

Budhedeo, S. H., & Pandya, N. P. (2018). Financial performance of public sector banks in India: A post-reform analysis. *Indian Journal of Finance, 12*(10), 7–20. https://doi.org/10.17010/ijf/2018/v12i10/132491

Chandra, S., & Awasthi, R. (2019). Insolvency risk: Issues and challenges for public sector commercial banks in India. *Indian Journal of Finance, 13*(12), 19–33. https://doi.org/10.17010/ijf/2019/v13i12/149266

Chen, C. M., Tso, G. K., & He, K. (2024). Quantum-optimized cost-based feature selection and credit scoring for mobile micro-financing. *Computational Economics*, *63*(2), 919–950. https://doi.org/10.1007/s10614-023-10365-8

Cozarenco, A., & Szafarz, A. (2018). Gender biases in bank lending: Lessons from microcredit in France. *Journal of Business Ethics*, *147*, 631–650. https://doi.org/10.1007/s10551-015-2948-y

Gupta, P., & Jaiswal, K. K. (2020). Analysis of financial performance of selected public and private sector banks. *Indian Journal of Finance, 14*(1), 45–57. https://doi.org/10.17010/ijf/2020/v14i1/149856

Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: An interdisciplinary review. *Journal of Big Data, 7*(1), Article no. 94. https://doi.org/10.1186/s40537-020-00369-8

Kaur, S., & Pasricha, J. S. (2019). Problems faced by bank employees in implementing financial inclusion schemes. *Indian Journal of Finance, 13*(12), 34–49. https://doi.org/10.17010/ijf/2019/v13i12/149267

Kumar, J., & Singh, R. (2017). Fiscal decentralization and urban local finance in India: A case study of Haryana. *Arthashastra Indian Journal of Economics & Research, 6*(6), 38–50. https://doi.org/10.17010/aijer/2017/v6i6/120116

Kumari, S., Nayak, D. S., & Swarnkar, T. (2023). Loan eligibility prediction using machine learning: A comparative approach. *Global Journal of Modeling and Intelligent Computing*, *3*(1), 48–54. https://www.researchgate.net/publication/372656643

Liu, Z., Du, G., Zhou, S., Lu, H., & Ji, H. (2022). Analysis of internet financial risks based on deep learning and BP neural network. *Computational Economics*, *59*(4), 1481–1499. https://doi.org/10.1007/s10614-021-10229-z

Madane, N., & Nanda, S. (2019). Loan prediction analysis decision tree. *Journal of the Gujrat Research Society*, *21*(14s). http://www.gujaratresearchsociety.in/index.php/JGRS/article/view/1351

Mohapatra, S., Sahoo, D., & Kesharwani, A. (2015). Outsourcing of information technology: An empirical study in the Indian banking industry. *Indian Journal of Finance, 9*(7), 7–21. https://doi.org/10.17010/ijf/2015/v9i7/72350

Nag, A. K., & Shah, J. (2022). An empirical study on the impact of Gen Z investors' financial literacy to invest in the Indian stock market. *Indian Journal of Finance, 16*(10), 43–59. https://doi.org/10.17010/ijf/2022/v16i10/172387

Nagaraju, T. (2014). An analysis of profitability and marketability efficiencies of Indian public and private banks. *Indian Journal of Finance, 8*(1), 15–28. https://doi.org/10.17010/ijf/2014/v8i1/71981

Narayanaswamy, T., & Muthulakshmi, A. P. (2016). Productivity and cost efficiency of commercial banks in India. *Indian Journal of Finance, 10*(1), 8–27. https://doi.org/10.17010/ijf/2016/v10i1/85839

Pandey, N., Gupta, R., Uniyal, S., & Kumar, V. (2021). Loan approval prediction using machine learning algorithms approach. *International Journal of Innovative Research in Technology*, *8*(1), 898–902. https://ijirt.org/Article?manuscript=151769

Park, M. S., Son, H., Hyun, C., & Hwang, H. J. (2021). Explainability of machine learning models for bankruptcy prediction. *IEEE Access*, *9*, 124887–124899. https://doi.org/10.1109/ACCESS.2021.3110270

Prakash, J., & Kankar, P. K. (2020). Health prediction of hydraulic cooling circuits using a deep neural network with an ensemble feature ranking technique. *Measurement, 151,* Article ID 107225. https://doi.org/10.1016/j.measurement.2019.107225

Ranawat, N. S., & Chakraborty, A. (2023). The impact of third-party financial products on the consumer loan services market in the banking sector: An analysis of sales progress and consumer behavior. *Asia-Pacific Financial Markets, 31*, 367–387. https://doi.org/10.1007/s10690-023-09419-6

Ranawat, N. S., Prakash, J., Miglani, A., & Kankar, P. K. (2023). Performance evaluation of LSTM and Bi-LSTM using non-convolutional features for blockage detection in a centrifugal pump. *Engineering Applications of Artificial Intelligence, 122*, Article ID 106092. https://doi.org/10.1016/j.engappai.2023.106092

Ranawat, N. S., Miglani, A., & Kankar, P. K. (2023). Blockage detection in centrifugal pump using butterfly optimization-based feature selection. *Mapan - Journal of Metrology Society of India, 38*(2), 499–509. https://doi.org/10.1007/s12647-022-00616-8

Rodney, A., & Mincey, K. (2020). Understanding the relationship between student loan debt and stress among female students at an HBCU. *Journal of African American Studies*, *24*, 269–275. https://doi.org/10.1007/s12111-020-09468-5

Shahraki, A., Abbasi, M., & Haugen, Ø. (2020). Boosting algorithms for network intrusion detection: A comparative evaluation of Real AdaBoost, Gentle AdaBoost and Modest AdaBoost. *Engineering Applications of Artificial Intelligence, 94*, Article ID 103770. https://doi.org/10.1016/j.engappai.2020.103770

Shukla, S. (2016). Performance of the Indian banking industry: A comparison of public and private sector banks. *Indian Journal of Finance, 10*(1), 41–55. https://doi.org/10.17010/ijf/2016/v10i1/85843

Supriya, P., Pavani, M., Saisushma, N., Kumari, N. V., & Vikas, K. (2019). Loan prediction by using machine learning models. *International Journal of Engineering and Techniques*, *5*(2), 144–148. http://www.ijetjournal.org/volume5/issue2/IJET-V5I2P28.pdf

Tejaswini, J., Kavya, T. M., Ramya, R. D., Triveni, P. S., & Maddumala, V. R. (2020). Accurate loan approval prediction based on machine learning approach. *Journal of Engineering Science*, *11*(4), 523–532. https://jespublication.com/upload/2020-110471.pdf

Viswanatha, V., Ramachandra, A. C., Vishwas, K. N., & Adithya, G. (2023). Prediction of loan approval in banks using machine learning approach. *International Journal of Engineering and Management Research, 13*(4), 7–19. https://doi.org/10.31033/ijemr.13.4.2

Wang, D., Hong, D., & Wu, Q. (2023). Prediction of loan rate for mortgage data: Deep learning versus robust regression. *Computational Economics*, *61*(3), 1137–1150. https://doi.org/10.1007/s10614-022-10239-5

Wilson, F. (2016). Making loan decisions in banks: Straight from the gut? *Journal of Business Ethics*, *137*, 53–63. https://doi.org/10.1007/s10551-014-2515-y

Wu, Z. (2022). Using machine learning approach to evaluate the excessive financialization risks of trading enterprises. *Computational Economics*, *59*(4), 1607–1625. https://doi.org/10.1007/s10614-020-10090-6

About the Author

Mr. Narendra Singh Ranawat is currently working as a Manager at Shree Rameshwaram Minerals, Udaipur, India. Previously, he worked as an Assistant Manager – Operations Department at Axis Bank Limited, India. He pursued an MBA and a B. Tech from Rajasthan Technical University, Kota, India. He has authored five research articles, two of which have been published, and the remaining three are currently undergoing peer review. His current research includes econometrics, machine learning, and HR analytics.