A Comparative Study of Classification Algorithms For Loan Eligibility Prediction

Abrin Azad Era*
Department of Computer Science and Engineering,
Bangladesh University of Business and Technology, Dhaka, Bangladesh.
Email: abrinazadera15@gmail.com*

***Abstract*—The study focuses on assessing borrower eligibility for loan approval based on creditworthiness and risk scores in commercial loan lending. It proposes a machine learning-based approach to automate and enhance the loan validation process. The research integrates advanced data preprocessing techniques and a range of classification algorithms, including Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, Decision Tree, Gradient Boosting, XGBoost, LightGBM, CatBoost, AdaBoost, Gaussian Naive Bayes, Multinomial Naive Bayes, and Bernoulli Naive Bayes, as well as ensemble methods like Bagging and Voting. The goal is to improve prediction accuracy and expedite decision-making in the banking sector. After comparing models, Random Forest, XG Boost, and LightGBM achieved the highest accuracy rates of 83.65%, 82.42%, and 83.45%, respectively. Applying ensemble Bagging resulted in an accuracy of 82.98% from Random Forest, while Voting achieved 84.02%. After tuning, the voting parameters accuracy improved to 86.43%. Further applying 5-fold cross-validation and fine-tuning the Voting technique resulted in an accuracy of 87.18%. On the other hand, Naive Bayes algorithms performed the worst, with accuracy rates of 65.33%, 64.11%, and 62.22%.**

***Keywords*—Loan Approval, Machine Learning, Classification, Random Forest, XGBoost, LightGBM, Ensemble Model, Bagging, Voting**

I. Introduction

In the realm of commercial loan lending, assessing borrower creditworthiness presents a formidable challenge, pivotal for predicting credit default risks and ensuring banking industry stability. Traditional methods, reliant on manual evaluation and simplistic scoring models, need to address the complexity and scale of modern loan applications. This paper, titled "A Comparative Study of Classification Algorithms for Loan Eligibility Prediction," proposes a novel machine learning-based approach to automate and enhance the loan validation process. By integrating sophisticated data preprocessing techniques and advanced classification algorithms, including Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, Decision Tree, Gradient Boost, XG Boost, Light GBM, Cat Boosting, ADA Boosting, Gaussian Naive Bayes (GNB), Multinomial Naive Bayes(MNB), and Bernoulli Naive Bayes(BNB), as well as ensemble approaches like Bagging and Voting, our study not only aims to improve prediction accuracy but also to significantly expedite the decision-making process, aligning with the evolving needs of the banking sector.

Some significant goals of this paper are outlined as follows:

- Compare various machine learning models to identify the best model for analyzing loan eligibility.
- Determine the most effective ensemble technique for loan eligibility prediction, including Bagging and Voting.
- Improve overall accuracy of loan eligibility predictions.
- Balance an imbalanced dataset to ensure fair and accurate model performance.

This research's inspiration arises from expanding upon these research efforts to enhance loan approval prediction accuracy and efficiency. We aim to contribute by developing an ensemble machine learning model, addressing data imbalance, conducting comprehensive model evaluations, and comparing our approach with existing methods, ultimately streamlining the loan approval process for financial institutions and applicants. We aim to develop a classification model utilizing leading machine learning algorithms. This strategy can transform the loan approval process and speed up loan approvals for clients, ultimately providing advantages for both banks and loan seekers.

By leveraging a comprehensive dataset and employing rigorous evaluation metrics, this research aims to provide a detailed comparison of models and techniques, striving to develop a robust and reliable loan prediction system that can assist financial institutions in making informed lending decisions.

II. Literature Review

Several Machine Learning models and research on loan eligibility prediction and classification have been carried out till today which exceed the predictability of conventional loan approval methods used previously.

In [1], Adnan Alagic et al. proposed integrating mental health data into loan approval prediction algorithms to address concerns regarding biases in traditional loan evaluations. The paper employs various machine learning algorithms, including ensemble methods such as AdaBoost and Gradient Boost, to achieve high prediction accuracy by considering mental health data along with financial indicators. The key contributions include improving predictive accuracy, with ensemble methods achieving up to 85.7% accuracy. However, challenges may appear with large-scale dataset handling and the requirement of further generalization through cross-validation. The proposed approach has the potential to improve loan approval decisions by taking into account applicants' mental health status.

In [2], Muhammad Zunnurain Hussain et al. proposed a machine-learning approach to enhance the accuracy of

bank loan approvals using applicant data. The study employs classification models like Decision Trees, Logistic Regression, and a new stacking model to achieve an accuracy of up to 83.24% on the test set. The key contribution of the paper includes the implementation of clustering for predictive enhancement and the implementation of ensemble methods for robust predictions. However, scalability issues and reliance on extensive data preprocessing remain. The practical implications include a more reliable and versatile credit risk assessment system, which will assist banks in handling large amounts of applications more efficiently.

In [3], Nazim Uddin et al. developed an ensemble machine learning system to enhance loan approval predictions for financial institutions, emphasizing the implementation of ensemble models for improved accuracy. The paper utilizes machine learning models like Random Forest, Decision Trees, and Extra Trees, with a voting-based ensemble achieving an accuracy of 87.26%. Key contributions include the successful implementation of a voting ensemble and a user-friendly application for loan approval predictions. However, possible computational intensity and the requirement for real-time deployment capabilities pose challenges. This system's implication is a streamlined loan approval process that increases efficiency for both banks and customers.

In [4], Ugochukwu E. Orji et al. proposed the development of machine learning models to predict bank loan eligibility, addressing the tedious loan approval process in the banking industry. This paper utilizes six machine learning algorithms (Random Forest, Gradient Boost, Decision Tree, Support Vector Machine, K-Nearest Neighbor, and Logistic Regression), achieving high accuracy with Random Forest scoring 95.55%. The key contributions of this article are the development and comparison of various ML models to enhance loan approval accuracy. However, the study's reliance on a single dataset may not generalize across different banking environments. The real-life implication of the proposed models is the potential to significantly improve the efficiency and accuracy of loan approval processes in financial institutions.

In [5], Debabrata Dansana et al. proposed a study on the impact of loan features on bank loan prediction using the Random Forest algorithm. This paper utilized the Random Forest Regressor model to analyze various loan approval parameters and predict loan defaults, achieving significant accuracy. The article's key contributions are identifying critical customer characteristics for loan approval and developing a reliable loan prediction model. However, the size of the dataset used in this study is limited and requires further validation with larger datasets. The real-life implication of the proposed model is its potential to reduce financial risks for banks by improving the accuracy of loan approvals.

In [6], Spyridon D. Mourtas et al. proposed a bio-inspired neural network for credit and loan approval classification. This study develops an unexplored weights and structure determination (WASD) neural network, enhanced with the beetle antennae search (BAS) algorithm, to improve learning efficiency and accuracy. This article's key contributions include creating a BWASD algorithm for binary classification, optimizing neural network structure, and providing a MATLAB package for implementation. However, the paper fails to address the scalability of larger datasets. The real-life implication of the proposed model is its potential to significantly reduce the risk and resources involved in loan approval processes for banks.

In [7], Niwan Wattanakitrungroj et al. proposed enhancing supervised model performance in credit risk classification using sampling strategies and feature ranking. This paper utilized logistic regression, random forest, and gradient boosting, achieving over 99.92% accuracy, precision, recall, and F1 scores with gradient boosting. The key contributions are demonstrating effective sampling strategies, highlighting feature ranking's impact, and comparing machine learning techniques. However, the paper only focuses on the Lending Club dataset, which may affect generalizability. The real-life implication of the proposed methods is improved credit risk prediction, aiding lenders in minimizing non-performing loans.

In [8], Asfand Ali et al. proposed a machine learning-based approach to predict loan defaults among applicants in financial organizations. This article utilizes logistic regression, K Nearest Neighbor, and Decision Tree algorithms to study historical customer data, achieving high prediction accuracy and moderate training and validation data loss. This article's notable contributions are automating the loan eligibility process and improving decision-making in lending. However, the paper does not address the potential biases in the dataset. The real-life implication of the proposed system is more efficient and accurate loan approval procedures, benefiting both banks and applicants.

In [9], Hemachandran et al. proposed a machine learning algorithm to predict loan repayment capabilities for financial institutions, addressing the increased risk of loan defaults post-pandemic. This paper utilizes K-nearest neighbor, decision tree, support vector machine, and logistic regression for data classification, achieving the highest accuracy with SVM at 79.6%. This paper's pivotal contributions are developing a cost-effective method to predict loan repayment behavior and aiding bank officers in loan approval decisions. Regardless, its reliance on a specific dataset may not generalize across different financial environments. The real-life implication of the proposed algorithm is enhanced fraud detection and streamlined loan approval processes for banks.

In [10], Haokun Dong et al. proposed a comparative study of five machine learning algorithms for financial risk evaluation. This paper develops models using K-nearest neighbor, Naïve-Bayes, Decision Tree, Logistic Regression, and Random Forest to predict loan defaults. The key contributions are evaluating model performance using accuracy, precision, recall, F1 scores, and ROC-AUC. However, the study's reliance on specific datasets may not be widespread in all financial contexts. The real-life implication of the proposed models is enhanced accuracy in predicting loan defaults, potentially reducing financial losses for lending organizations.

In [11], Nguyen Ly et al. explored predictive models for peer-to-peer (P2P) loan defaults, focusing on non-personal

TABLE I: Comparison between previous research works

Ref	Research Purpose	Used Methods/Techniques	Result Evaluation	Research Gaps
[1]	Integrate mental health data into loan approval predictions to reduce biases	AdaBoost, Gradient Boost	Achieved 85.7% accuracy	Challenges with large datasets, need for further generalization
[2]	Enhance bank loan approval accuracy using applicant data	Decision Trees, Logistic Regression, Stacking model	Achieved up to 83.24% accuracy	Scalability issues, reliance on extensive preprocessing
[3]	Improve loan approval prediction using ensemble methods	Random Forest, Decision Trees, Extra Trees with voting ensemble	Achieved 87.26% accuracy	Computational intensity, need for real-time capabilities
[4]	Predict bank loan eligibility to streamline approval process	Random Forest, Gradient Boost, Decision Tree, SVM, K-Nearest Neighbor, Logistic Regression	Achieved 95.55% accuracy with Random Forest	Limited generalizability due to reliance on a single dataset
[5]	Analyze impact of loan features on loan prediction	Random Forest Regressor, Decision Tree	Achived 80% accuracy with random forest	Small dataset size, need for larger dataset validation
[6]	Classify credit and loan approvals with bio-inspired neural network	WASD neural network, BAS algorithm	Achieved a 95% accuracy	Scalability concerns with larger datasets
[7]	Enhance credit risk classification through sampling strategies and feature ranking	Logistic Regression, Random Forest, Gradient Boosting	Achieved over 99.92% accuracy	Limited generalizability due to focus on Lending Club dataset
[8]	Predict loan defaults in financial organizations	Logistic Regression, KNN, Decision Tree	Achieved an accuracy of 81.3% with Logistic Regression model	Did not address dataset bias
[9]	Predict loan repayment capabilities post-pandemic	KNN, Decision Tree, SVM, Logistic Regression	Achieved highest accuracy 79.6% with SVM	Limited generalizability due to specific dataset reliance
[10]	Compare algorithms for financial risk evaluation	KNN, Naïve Bayes, Decision Tree, Logistic Regression, Random Forest	Achieved highest 98.5% accuracy with random forest.	Limited applicability across financial contexts
[11]	Predict P2P loan defaults using non-personal data	Random Forest, SVM, Decision Tree, Logistic Regression, Naïve Bayes, XGBoost	Achieved 83% accuracy	Reliance on financial data may reduce predictive accuracy
[12]	Predict corporate loan defaults using financial ratios	Logistic Regression, Neural Networks	Achieved 89.1% accuracy	Focuses only on Estonian firms, limiting applicability
[13]	Assess credit risk in auto loans using PSO-XGBoost model	PSO-XGBoost, Smote-Tomek link	Achieved an accuracy of 83.11%	Model complexity limits real-time applicability
[14]	Predict loan defaults through ensemble modeling	SentiNet model, ensemble learning	Improved prediction accuracy to 76.8%	Moderate accuracy indicates room for improvement
[15]	Assess loss-given default (LGD) using a multi-stage model	Hybrid model combining clustering and oversampling	Achieved an accuracy of 92.3%	Model complexity limits scalability
[16]	Improve loan default prediction in rural banks	Logistic Regression, ANN, Entropy method	Achieved an accuracy of 91.08%	Requires additional validation with varied datasets
[17]	Predict loan defaults in P2P lending	Blending (Logistic Regression, Random Forest, CatBoost)	Achieved 76.8% accuracy	Model complexity may hinder real-time deployment
[18]	Support credit decisions for cash loans	Random Forest with correlation-based feature selection	Achieved an accuracy of 82.8%	Challenges with interpretability due to complex classifiers
[19]	Predict loan eligibility using classification algorithms	Logistic Regression, Random Forest, Decision Tree, SVM, KNN	Achieved 98% accuracy with Random Forest	Overfitting observed with KNN
[20]	Credit scoring in micro-lending markets with limited history	Random Forest, XGBoost, AdaBoost	Achieved approximately 80% accuracy with ensembles	Observed overfitting in some models

data due to privacy concerns. This study used a dataset with financial attributes, employing Random Forest, SVM, Decision Tree, Logistic Regression, Naïve Bayes, and XGBoost models. The best-performing model achieved 83% accuracy. The primary contribution is a model that respects privacy while predicting loan defaults accurately. However, the reliance on financial data alone may limit predictive accuracy. This approach has implications for enhancing P2P platforms by offering lenders improved tools for evaluating creditworthiness without compromising user privacy.

In [12], Keijo Kohv et al. performed a comparative analysis to predict corporate loan defaults using financial ratios, tax arrears, and annual report submission delays. Using logistic regression and neural networks on data from an Estonian bank, they found tax arrears to be the strongest predictor, achieving 89.1% accuracy when combined with all variables. The study contributes by introducing novel predictive variables like tax arrears, which outperform traditional financial ratios. However, it focuses solely on Estonian firms, which may limit its applicability elsewhere. This model could help banks make informed credit decisions by highlighting tax arrears as a critical indicator of default risk.

In [13], C Rao et al. developed a credit risk assessment model for personal auto loans based on the PSO-XGBoost model to manage the risk of defaults in auto loans. The study uses data from Kaggle, balanced using the Smote-Tomek link method, and applied an enhanced feature selection process. The PSO-XGBoost model outperformed XGBoost, RF, and LR in accuracy and precision metrics, making it highly effective. The main contribution is a more reliable credit risk assessment model for auto finance. However, the model's complexity may hinder its real-time applicability. This model helps financial institutions minimize bad debt by accurately evaluating auto loan risks.

In [14], P Pathak et al. proposed the SentiNet model, focusing on loan default prediction through ensemble modeling. This paper utilizes data processing and ensemble learning on the Lending Club dataset, improving prediction accuracy to 76.8% compared to baseline models. Key contributions include enhancing model robustness and optimizing loan approval rates for better credit access. However, its moderate accuracy indicates potential for improvement. This method supports financial institutions in balancing risk and credit accessibility through an optimized loan approval process.

In [15], M Fan et al. developed a multi-stage hybrid model (HMS) for assessing loss-given default (LGD) bank loans. The study combines supervised and unsupervised learning models with clustering and oversampling, showing that the HMS model outperformed others on mean absolute error (MAE) and mean squared error (MSE) metrics. Key contributions include a multi-stage framework that enhances LGD prediction accuracy and supports capital requirement determinations. However, the model's complexity may hinder its scalability. This model can assist banks in effectively managing credit risk and regulatory compliance.

In [16], Yiheng Li et al. proposed a combined model for improving loan default prediction in rural commercial banks in China. This paper uses a logistic regression algorithm and an artificial neural network model, combined with an entropy method, to enhance predictive performance. The key contributions are developing a novel credit scoring model (LNN-Entropy model) and validating it against a state-of-the-art approach, stacking. However, more datasets of different sizes and attributes are needed to validate the model. The implication of the proposed model is its potential to improve loan evaluation processes and reduce default rates in rural commercial banks.

In [17], Xingyun Li et al. developed a multi-model fusion approach to improve loan default prediction in online P2P lending environments. This paper uses Adaptive Synthetic Sampling (ADASYN) to handle data imbalance. It used Blending to combine Logistic Regression, Random Forest, and CatBoost models, achieving higher accuracy than individual models, with an accuracy of approximately 76.8%. The key contributions include an effective ensemble model for default risk reduction in P2P platforms. However, the model's complexity may impact real-time deployment. This approach supports online loan platforms in managing credit risk and enhancing investor confidence.

In [18], P Ziemba et al. proposed a credit decision support system using machine learning algorithms to assess cash loan risk. This paper developed a framework that includes feature selection, resampling, discretization, and classification, concluding that Random Forest with correlation-based feature selection was the most effective method. The main contribution is an efficient model framework for credit scoring using cash loan data. However, the model may struggle with interpretability, given the use of complex classifiers. This pipeline can aid financial institutions in making informed loan-granting decisions, particularly under economic challenges like the COVID-19 pandemic.

In [19], Krishanu Agarwal et al. proposed a comparative study of classification algorithms for predicting loan eligibility. This paper utilizes machine learning techniques such as logistic regression, random forest, decision tree, support vector classifier, and k-nearest neighbors to analyze bank loan data, achieving the highest accuracy of 98% with the random forest algorithm. The key contributions are providing a comprehensive comparison of different algorithms regarding accuracy, precision, recall, and ROC-AUC. However, the limitation of the paper is the overfitting observed in the k-nearest neighbors algorithm. The proposed study can be implemented to enhance the efficiency of loan approval processes by accurately predicting potential loan buyers.

In [20], Apostolos Ampountolas et al. developed a machine-learning approach for micro-credit scoring to address the lack of recorded credit history in micro-lending markets. This paper compares machine learning algorithms such as Random Forest, XGBoost, and AdaBoost on micro-lending data, achieving an accuracy of approximately 80% for ensemble classifiers. The key contributions are as follows: providing a comprehensive comparison of different algorithms for classifying borrowers into credit categories. However, overfitting has been observed in some models. The paper aims to assist micro-lending institutions with a cost-effective and reliable method for assessing creditworthiness when credit history is lacking.

Many different types of loan eligibility prediction model were developed in the earlier work. Some of these current systems are tabulated in Table I.

III. Proposed Model

The methodology for analyzing and processing the bank loan data is a structured approach that encompasses a series of critical steps, each designed to ensure thorough data preparation, effective model training, and precise evaluation of outcomes. Below is a comprehensive breakdown of this research's detailed methodology:

- **Load Dataset:** The initial step involves importing the bank loan dataset for training and testing the models, which serves as the foundation for the analysis. This dataset will subsequently be utilized for training various models and for testing their performance.
- **Data Preprocessing and Handle Imbalance with SMOTE:** Data preprocessing is vital for preparing the dataset for analysis. This involves cleaning the data by removing duplicates and fixing any inconsistencies, normalizing numerical values to a common scale, and transforming categorical variables into a suitable format. To address class imbalance within the dataset—where one class of outcomes may be underrepresented—this research applies SMOTE (Synthetic Minority Over-sampling Technique). This innovative approach generates synthetic instances of the minority class, thus aiding in creating a more balanced dataset that enhances model training effectiveness.

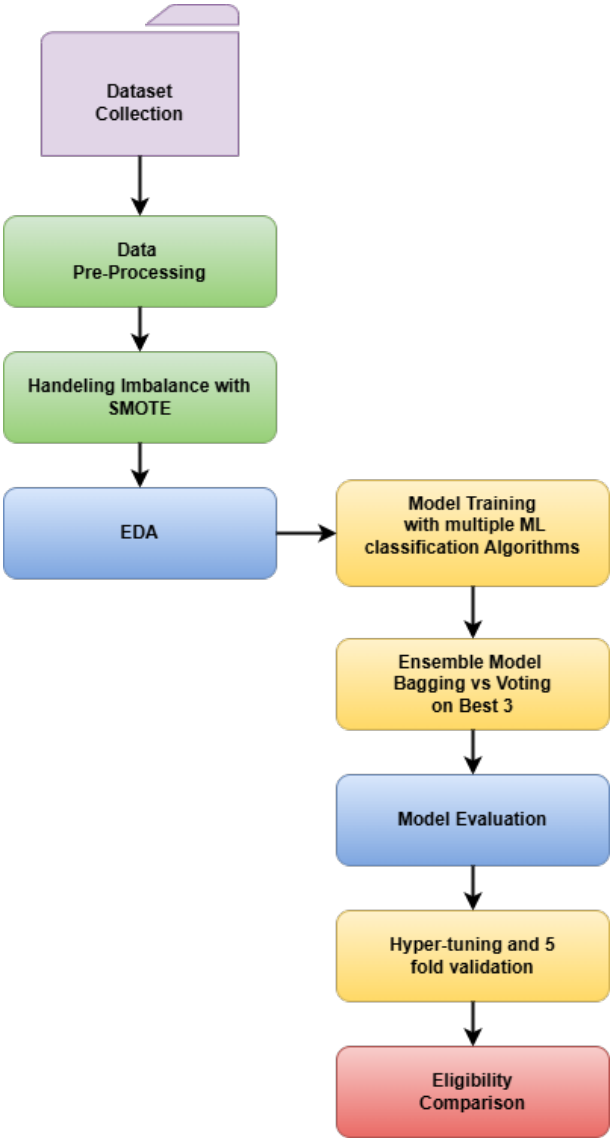


Fig. 1: Proposed Model

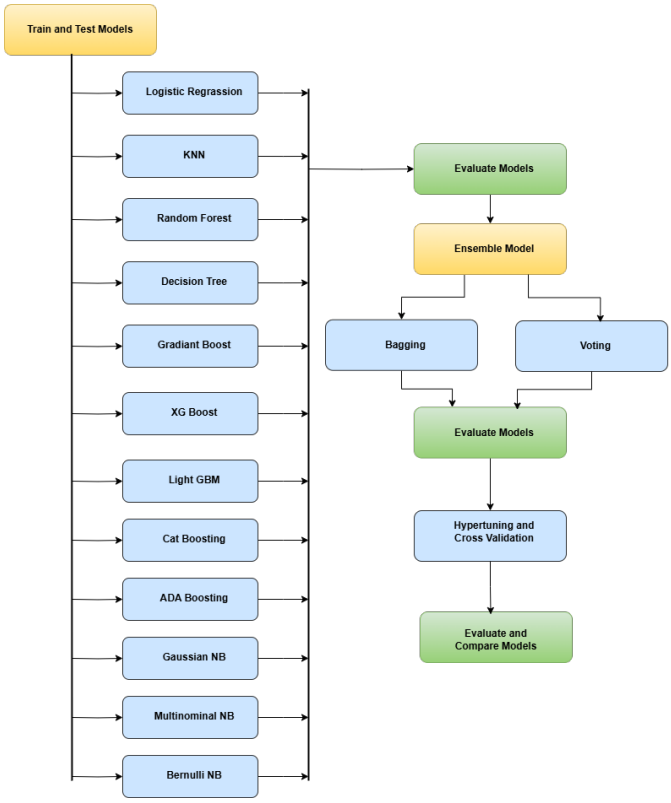


Fig. 2: Model Architecture

- **Train and Test Models: a. Logistic Regression:** This research deploys this statistical model, which utilizes a logistic function to estimate the probability of a binary outcome based on one or more predictor variables. This technique is particularly useful for understanding relationships between variables while predicting outcomes.
  b. KNN (K-Nearest Neighbors): This non-parametric method facilitates classification and regression by measuring the distance between data points. It predicts the class of a data point based on the classifications of its nearest neighbours, making it simple yet effective.
  c. Random Forest: This research leverages this ensemble learning method, which constructs multiple decision trees during training and merges their outputs to produce a more accurate and stable prediction. This method notably reduces the risk of overfitting while improving overall accuracy.

d. Decision Tree: This intuitive model mimics human decision-making through a tree-like structure, where each node represents a decision based on certain attributes. Its straightforward interpretability and ability to handle both numerical and categorical data make it a popular choice.
e. Gradient Boosting: This ensemble technique builds models sequentially; each new model corrects the errors of its predecessor. It effectively enhances the predictive accuracy of previously weaker models.
f. XG Boost: Known for its exceptional speed and performance, XG Boost is an optimized distributed gradient boosting library that stands out for its efficiency and flexibility in handling large datasets.
g. Light GBM: This gradient boosting framework employs tree-based learning algorithms, recognized for their speed and efficiency, making it particularly effective for large datasets where computational resources may be constrained.
h. Cat Boosting: Distinctive for its ability to automatically handle categorical features, this gradient boosting algorithm minimizes the need for extensive preprocessing, thereby enhancing model performance.
i. ADA Boosting: This ensemble method combines multiple weak classifiers into a single strong classifier, focusing on misclassified instances from previous models to improve overall performance.
j. Gaussian Naive Bayes (NB): This probabilistic classi-

fier implements Bayes' theorem with strong independence assumptions, making it effective for high-dimensional datasets.

k. Multinomial NB: Specifically designed for classification problems involving discrete features, this variant of Naive Bayes is frequently employed in text classification tasks.

l. Bernoulli NB: This variant is tailored for binary or boolean features, making it particularly suitable for datasets with binary outcomes.

- **Evaluate Models:** To assess the performance of each model, this research utilizes a variety of metrics, including recall, precision, F1 score, accuracy, log-loss, confusion matrix, and ROC curve. These metrics allow for gauging the effectiveness of each model comprehensively.
- **Ensemble Model:** This research explores the potential of combining multiple models to enhance overall performance.

  Bagging: This ensemble technique increases the stability and accuracy of machine learning algorithms by averaging the outputs of multiple models. It significantly mitigates variance and helps prevent overfitting.

  Voting: In this ensemble approach, diverse models cast votes regarding the output, and the majority decision is taken as the final prediction. This method synergizes the strengths of various models to improve accuracy.
- **Re-evaluate Models:** This research fine-tunes the hyperparameters of the models to optimize performance. This involves systematically adjusting the parameters within the algorithm to find the most effective configuration that yields the best results during training. Cross-validation is employed as a technique to validate the model's effectiveness by splitting the dataset into multiple iterations of training and validation sets. This process helps in assessing how well the model is likely to generalize to an independent dataset, reducing the risk of overfitting.
- **Hypertuning and Cross-Validation:** Optimize the hyperparameters of the models and validate them using cross-validation techniques. This ensures that the models are fine-tuned for the best performance.

  1. Random Search: A technique for hyperparameter optimization that randomly samples from the hyperparameter space.

  2. Grid Search: A technique for hyperparameter optimization that exhaustively searches through a specified subset of the hyperparameter space.

  3. 5-Fold Cross Validation: A technique where the dataset is divided into 5 subsets, and the model is trained and validated 5 times, each time using a different subset as the validation set and the remaining subsets as the training set.
- **Final Model Selection:** Based on the evaluation metrics and results from the re-evaluation, this research selects the final model or ensemble of models that demonstrate the highest predictive performance. The selection is informed by the balance between accuracy, recall, precision, and F1 score, as well as the model's interpretability and computational efficiency.

IV. Performance Evaluation

A. Correlation:

Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It's a common tool for describing simple relationships without making a statement about cause and effect.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

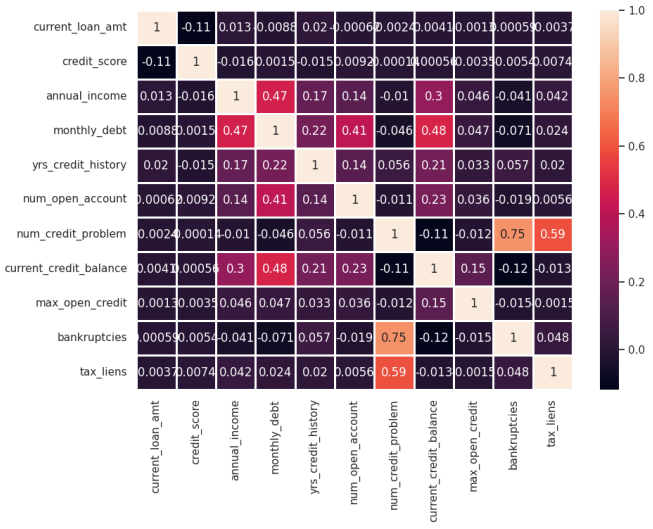


Fig. 3: Correlation Matrix

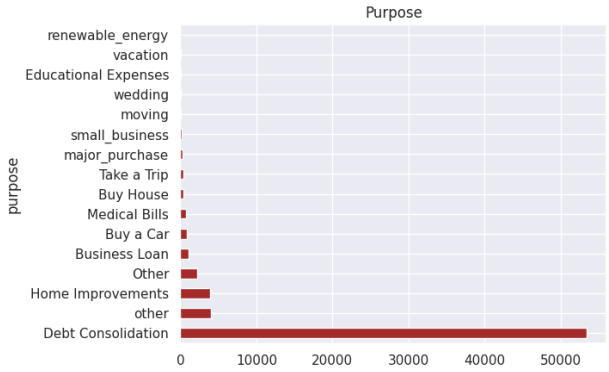


Fig. 4: Loan Purpose

B. Evaluation matrix:

In this study, we used several evaluation metrics to assess the performance of our machine-learning models. These metrics provide a comprehensive understanding of how well the models are performing in predicting loan eligibility.

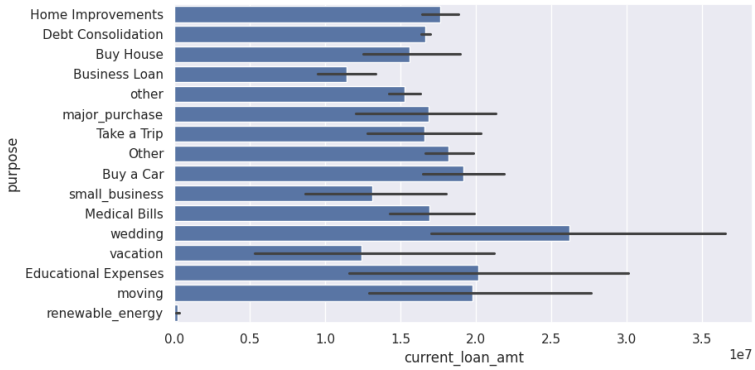


Fig. 5: Loan Amount Based on Purpose

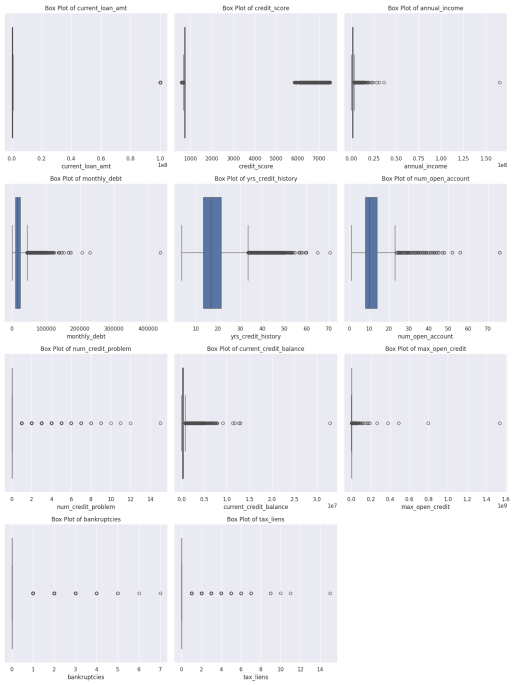


Fig. 6: Attributtes before trimming outliers

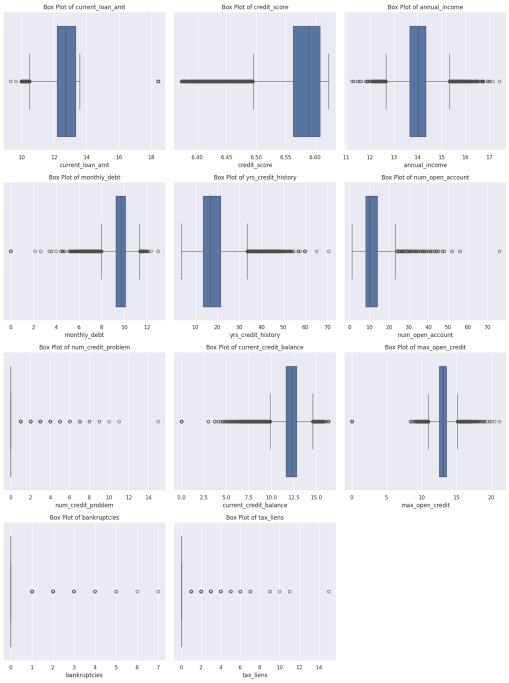


Fig. 7: Attributtes after trimming outliers

1. Accuracy: Measures the proportion of correctly predicted instances out of the total instances. It is a general indicator of the model's performance.

2. Precision: Indicates the proportion of true positive predictions out of all positive predictions made by the model. It is useful for understanding the model's performance in identifying relevant instances.

3. Recall: Measures the proportion of true positive predictions out of all actual positive instances. It helps in understanding the model's ability to capture all relevant instances.

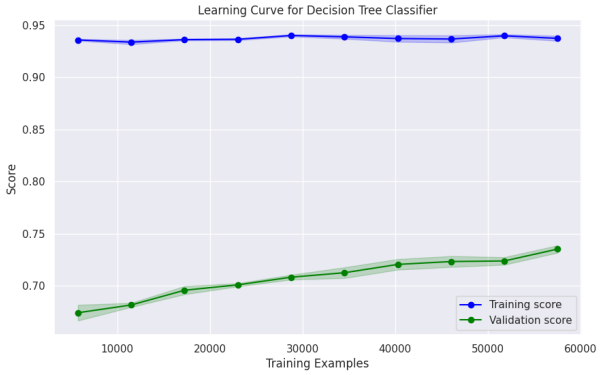
4. F1 Score: The harmonic mean of precision and recall. It provides a balance between precision and recall, which is especially useful when dealing with imbalanced datasets.

5. Log-Loss: Measures the performance of a classification model where the prediction is a probability value between 0 and 1. It penalizes false classifications more heavily. We calculated log-loss for both training and testing data to evaluate model performance.

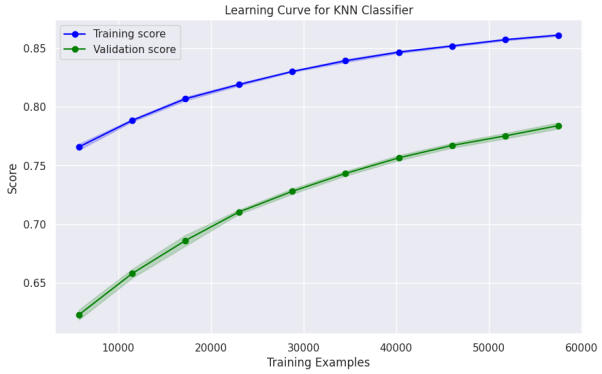
6. Confusion Matrix: A table used to describe the performance of a classification model. It shows the true positives, true negatives, false positives, and false negatives, providing a detailed breakdown of the model's performance.

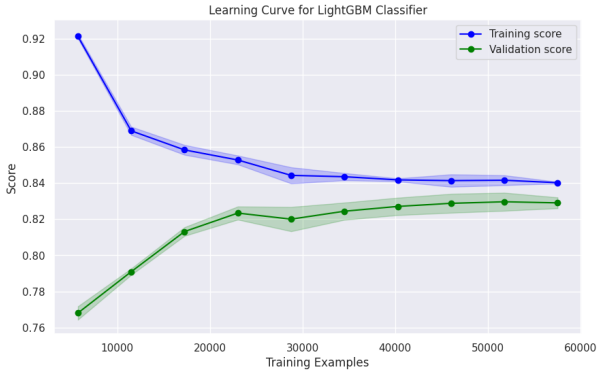
7. ROC Curve: A graphical representation of the model's performance across different threshold values. It plots the true positive rate against the false positive rate, helping to visualize the trade-off between sensitivity and specificity.

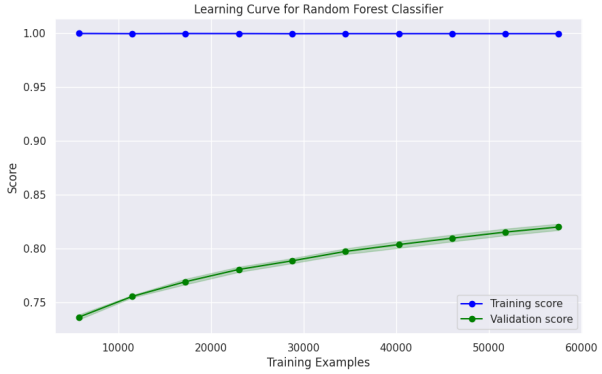
8. Learning Curve: A plot that shows the model's performance on the training set and the validation set over a varying number of training instances. It helps in understanding how the model's performance improves with more data and whether the model is overfitting or underfitting.

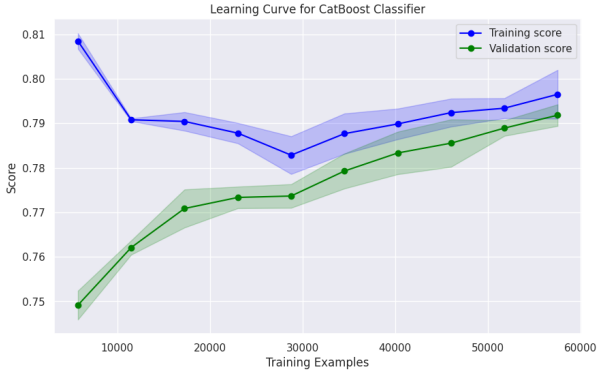


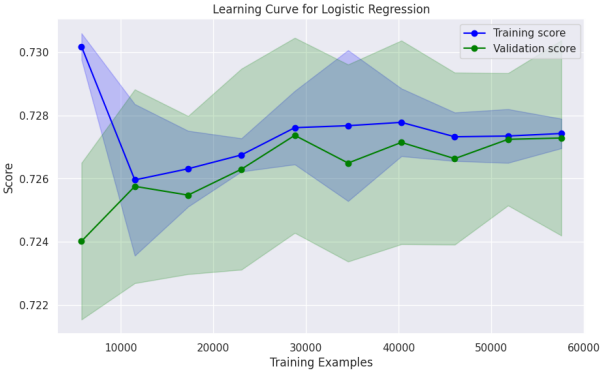
By using these evaluation metrics, we ensured a thorough assessment of our models, allowing us to identify the best-

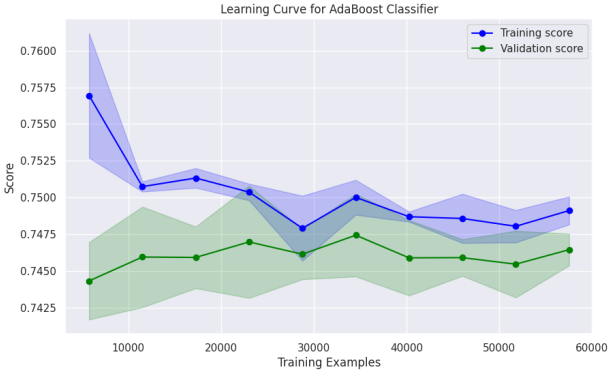


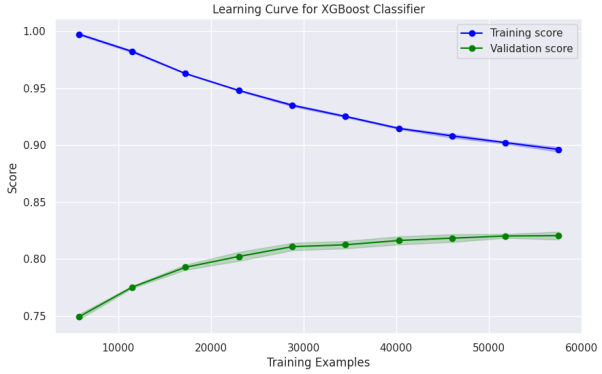


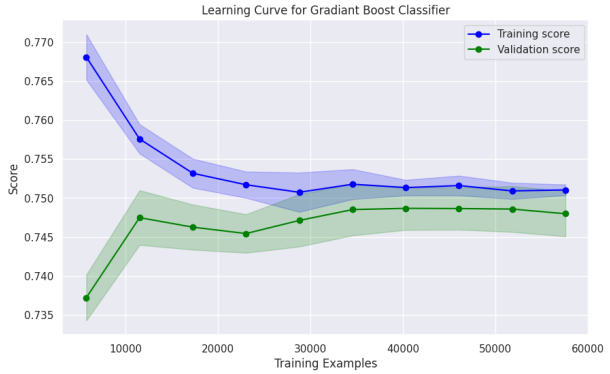












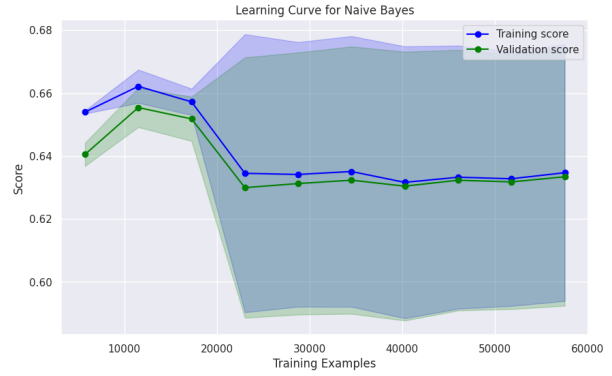
performing algorithms for loan eligibility prediction.

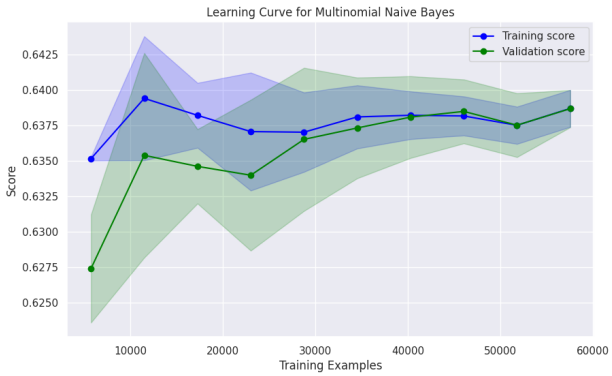
C. Ensemble Model (Bagging Vs Voting):

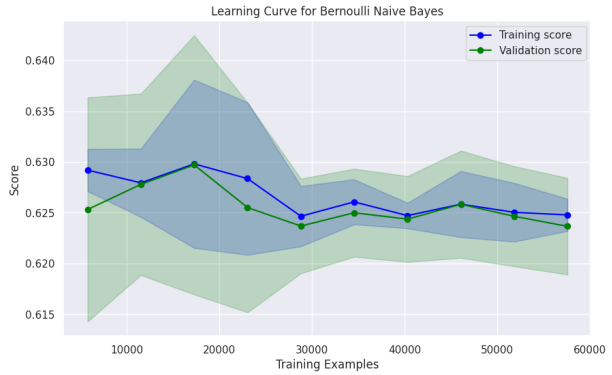
The analysis of model performance revealed key insights into various algorithms. Random Forest with Bagging achieved

TABLE II: Model Performance Comparison with Testing Data (Demo Version)

Model	Accuracy	Recall	Precision	F1 Score	Log Loss	TP	FP	FN	TN
Decision Tree	75.01%	0.714	0.770	0.741	5.776	11,000	3,289	4,416	12,127
K-Nearest Neighbors (KNN)	79.94%	0.662	0.913	0.767	1.365	10,204	973	5,212	14,443
Random Forest	83.65%	0.802	0.861	0.831	0.405	12,369	1,995	3,047	13,421
Logistic Regression	72.79%	0.725	0.729	0.727	0.497	11,182	4,156	4,234	11,260
XGBoost	82.42%	0.838	0.815	0.827	0.370	12,922	2,925	2,494	12,491
LightGBM	83.46%	0.852	0.823	0.837	0.374	13,132	2,816	2,284	12,600
CatBoost	79.05%	0.775	0.800	0.787	0.422	11,941	2,983	3,475	12,433
AdaBoost	74.73%	0.657	0.801	0.722	0.673	10,136	2,512	5,280	12,904
Gradient Boosting	74.90%	0.650	0.810	0.721	0.495	10,024	2,348	5,392	13,068
Gaussian Naive Bayes	65.34%	0.649	0.655	0.652	1.053	10,003	5,274	5,413	10,142
Multinomial Naive Bayes	64.12%	0.601	0.653	0.626	0.647	9,266	4,914	6,150	10,502
Bernoulli Naive Bayes	62.14%	0.577	0.633	0.604	0.651	8,898	5,156	6,518	10,260







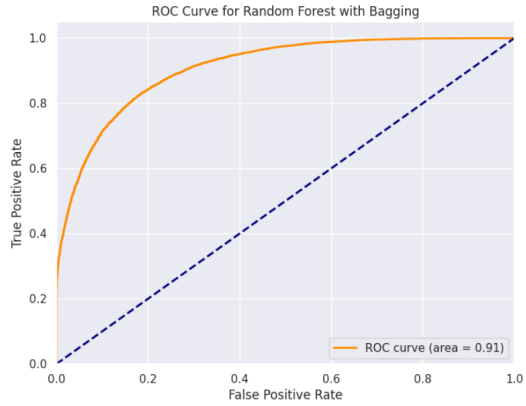


Fig. 8: Ensemble Bagging

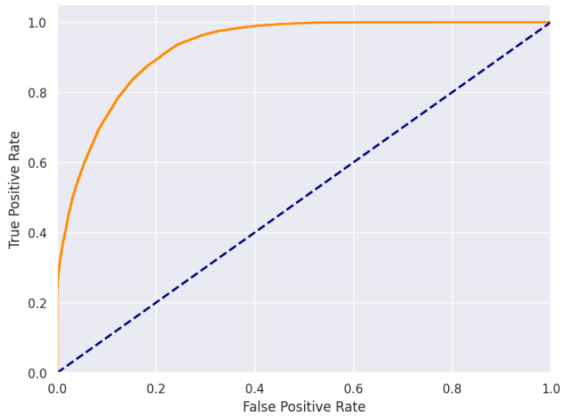


Fig. 9: Ensemble Voting

high training accuracy (97.69%) and F1 score (97.69%), but showed significant decline on testing data (accuracy: 81.67%, F1: 80.74%), indicating potential overfitting. XGBoost with Bagging displayed more balanced results, with training accuracy at 88.90% and testing accuracy at 82.98%, suggesting a better trade-off between bias and variance. LightGBM with Bagging had slightly higher testing accuracy (82.57%) and F1 score (82.61%) than XGBoost, but faced issues with lower recall and higher log loss.

The Voting Classifier outperformed the others with the highest testing accuracy (84.02%) and a competitive F1 score

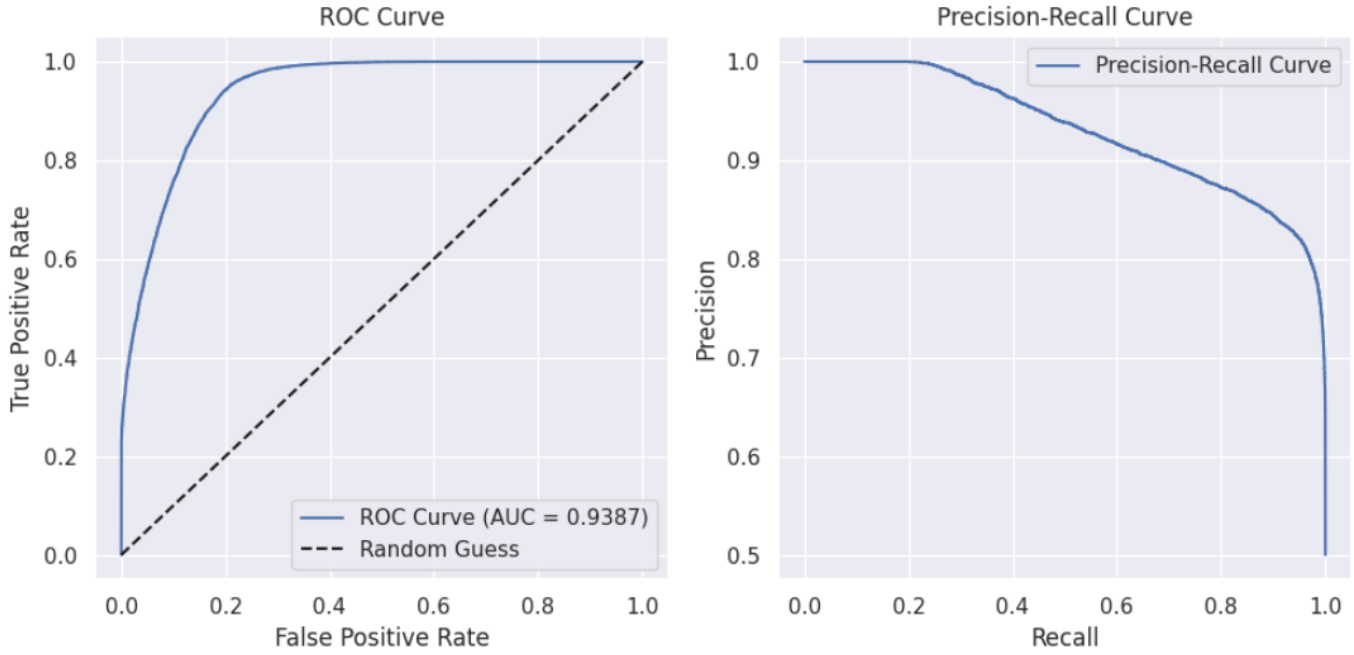


Fig. 10: Ensemble Voting after 5 fold cross validation

(83.85%), effectively harnessing the strengths of individual models for reliable predictions.

D. Tuning And Cross Validation:

The Voting Classifier achieved excellent performance, demonstrating high generalization ability and robustness. On the training set, it exhibited near-perfect metrics with an Accuracy of 98.90%, F1 Score of 98.91%, and Log Loss of 0.1714, suggesting effective learning without significant overfitting. On the testing set, it maintained strong results with an Accuracy of 86.43%, an F1 Score of 86.72%, and a Log Loss of 0.3336, reflecting its capability to generalize well to unseen data. Cross-validation confirmed consistent performance across folds with a mean accuracy of 85.82%, underscoring its stability.

After threshold adjustment, the testing metrics improved further, achieving an Adjusted Accuracy of 87.18%, an Adjusted F1 Score of 88.00%, and a significantly higher Recall of 93.99%, which enhances the model's sensitivity to identify positive cases correctly. The adjusted confusion matrix showed fewer false negatives, demonstrating its effectiveness in minimizing critical classification errors. Overall, the Voting Classifier is a highly reliable ensemble method with optimized threshold tuning for better recall and balanced performance.

V. Future Research Direction

For future smart irrigation management, several issues must be addressed as follows:

- **Incorporating Additional Features:** Explore the inclusion of more relevant features, such as socio-economic indicators or real-time contextual data, to improve the model's predictive power and adaptability across different datasets or domains.
- **Advanced Hyperparameter Optimization:** Implement advanced optimization techniques, such as Bayesian Optimization or Genetic Algorithms, to fine-tune hyperparameters of individual classifiers within the Voting Classifier for further performance gains.
- **Real-World Deployment and Feedback Loop:** Deploy the model in a real-world setting and integrate a feedback loop to continuously learn from misclassifications, adapt to evolving patterns, and improve its robustness in practical applications.

VI. Conclusion

The study focused on the estimation of loan default risks using machine learning classification models in the banking sector. The choice of algorithm plays a crucial role in loan decision management, as this helps determine the likelihood of loan default by clients. Initially, we performed the dataset cleansing, which involved removing variables with a high proportion of missing data. We addressed unbalanced data and outliers issues before feeding the data to machine learning algorithms. We explored multiple machine learning algorithms, including LR, DT, RF, KNN, XGB, LightGBM, Cat Boost, AdaBoost, GB, and Naive Bayes. We discovered the three top-performing models and proposed an effective ensemble classifier along with the best hyper-tuning parameters and 5-fold cross-validation. Our comprehensive evaluation confirmed that our approach effectively enhances the accuracy and reliability of loan default predictions. The findings of this study might contribute to the advancement of risk management practices

in the banking industry and provide valuable insights for future research in this domain. Although our proposed models have showcased enhanced performance rates, we have certain limitations. Additionally, the potential benefits of employing unsupervised machine learning models due to the unlabeled nature of real-world data and ensemble techniques like the Stack ensemble have not been investigated. To address these limitations and enhance the applicability of our bank loan prediction models in real-life scenarios, our future work will encompass the collection of more diverse and extensive real-life data, hyperparameter tuning, a wider range of data balancing approaches, exploration of unsupervised learning methods, and the utilization of ensemble techniques for building more reliable models for banking and financial institutions.

ACKNOWLEDGMENT

We would like to acknowledge the support of the Bangladesh University of Business & Technology and the Data Mining lab for their suggestion and resource sharing.

REFERENCES

[1] A. Alagic, N. Zivic, E. Kadusic, D. Hamzic, N. Hadzajlic, M. Dizdarevic, and E. Selmanovic, "Machine learning for an enhanced credit risk analysis: A comparative study of loan approval prediction models integrating mental health data," *Machine Learning and Knowledge Extraction*, vol. 6, no. 1, pp. 53–77, 2024.

[2] M. Z. Hussain, S. Ejaz, E. Batool, M. Z. Hasan, M. Mustafa, A. Khalid, U. Hussain, Z. Khan, A. Javaid, M. F. Ashraf *et al.*, "Bank loan prediction system using machine learning models," in *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*. IEEE, 2024, pp. 1–5.

[3] N. Uddin, M. K. U. Ahamed, M. A. Uddin, M. M. Islam, M. A. Talukder, and S. Aryal, "An ensemble machine learning based bank loan approval predictions system with a smart application," *International Journal of Cognitive Computing in Engineering*, vol. 4, pp. 327–339, 2023.

[4] U. E. Orji, C. H. Ugwuishiwu, J. C. Nguemaleu, and P. N. Ugwuanyi, "Machine learning models for predicting bank loan eligibility," in *2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON)*. IEEE, 2022, pp. 1–5.

[5] D. Dansana, S. G. K. Patro, B. K. Mishra, V. Prasad, A. Razak, and A. W. Wodajo, "Analyzing the impact of loan features on bank loan prediction using r andom f orest algorithm," *Engineering Reports*, vol. 6, no. 2, p. e12707, 2024.

[6] S. D. Mourtas, V. N. Katsikis, P. S. Stanimirović, and L. A. Kazakovtsev, "Credit and loan approval classification using a bio-inspired neural network," *Biomimetics*, vol. 9, no. 2, p. 120, 2024.

[7] N. Wattanakitrungroj, P. Wijitkajee, S. Jaiyen, S. Sathapornvajana, and S. Tongman, "Enhancing supervised model performance in credit risk classification using sampling strategies and feature ranking," *Big Data and Cognitive Computing*, vol. 8, no. 3, p. 28, 2024.

[8] A. Ali, A. Irfan, A. Raza, and Z. Memon, "Banking in the digital age: Predicting eligible customers through machine learning and aws," in *2024 IEEE 1st Karachi Section Humanitarian Technology Conference (KHI-HTC)*. IEEE, 2024, pp. 1–6.

[9] K. Hemachandran, R. V. Rodriguez, R. Toshniwal, M. Junaid, and L. Shaw, "Performance analysıs of different classıfıcatıon algorıthms for bank loan sectors," in *Intelligent Sustainable Systems: Proceedings of ICISS 2021*. Springer, 2021, pp. 191–202.

[10] H. Dong, R. Liu, and A. W. Tham, "Accuracy comparison between five machine learning algorithms for financial risk evaluation," *Journal of Risk and Financial Management*, vol. 17, no. 2, p. 50, 2024.

[11] L. Nguyen, M. Ahsan, and J. Haider, "Reimagining peer-to-peer lending sustainability: unveiling predictive insights with innovative machine learning approaches for loan default anticipation," *FinTech*, vol. 3, no. 1, pp. 184–215, 2024.

[12] K. Kohv and O. Lukason, "What best predicts corporate bank loan defaults? an analysis of three different variable domains," *Risks*, vol. 9, no. 2, p. 29, 2021.

[13] C. Rao, Y. Liu, and M. Goh, "Credit risk assessment mechanism of personal auto loan based on pso-xgboost model," *Complex & Intelligent Systems*, vol. 9, no. 2, pp. 1391–1414, 2023.

[14] P. Pathak, A. Jain, M. Bansal, and P. S. Rana, "Sentinet: Empowering robust loan default prediction through ensemble modeling," in *2023 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI)*. IEEE, 2023, pp. 1–6.

[15] M. Fan, T.-H. Wu, and Q. Zhao, "Assessing the loss given default of bank loans using the hybrid algorithms multi-stage model," *Systems*, vol. 11, no. 10, p. 505, 2023.

[16] Y. Li and W. Chen, "Entropy method of constructing a combined model for improving loan default prediction: A case study in china," *Journal of the Operational Research Society*, vol. 72, no. 5, pp. 1099–1109, 2021.

[17] X. Li, D. Ergu, D. Zhang, D. Qiu, Y. Cai, and B. Ma, "Prediction of loan default based on multi-model fusion," *Procedia Computer Science*, vol. 199, pp. 757–764, 2022.

[18] P. Ziemba, J. Becker, A. Becker, A. Radomska-Zalas, M. Pawluk, and D. Wierzba, "Credit decision support based on real set of cash loans using integrated machine learning algorithms," *Electronics*, vol. 10, no. 17, p. 2099, 2021.

[19] K. Agarwal, M. Jain, and A. Kumawat, "Comparing classification algorithms on predicting loans," in *Information Systems and Management Science: Conference Proceedings of 3rd International Conference on Information Systems and Management Science (ISMS) 2020*. Springer, 2022, pp. 240–249.

[20] A. Ampountolas, T. Nyarko Nde, P. Date, and C. Constantinescu, "A machine learning approach for micro-credit scoring," *Risks*, vol. 9, no. 3, p. 50, 2021.