International Conference on Machine Learning and Data Engineering (ICMLDE 2024)

A Comparative Performance Assessment for Prediction of Loan Approval in Financial Sector

Lopamudra Hota[a], Puneet Kumar Jain[a], Arun Kumar[a]

[a] *Department of Computer Science and Engineering, National Institute of Technology, Rourkela, 769008, India*

Abstract

Loans are the bank's assets since they generate income in terms of interest to banks. Lending a loan to a customer creates credit and liability for the bank and the customer. The profit and loss of a bank depend on the customer's ability to pay back the loan or not, i.e., defaulter or not. Therefore, predicting the probability of loan repayment becomes a crucial task. For this purpose, ensemble learning methods have been incorporated extensively, and studies have reported the superiority of these methods over conventional classification methods. This paper provides a comprehensive comparative performance assessment of various ensemble methods for predicting Loan approval in the banking sector. Ensemble algorithms, including bagging, boosting and stacking, are considered with the Neural Network (NN), Decision Tree (DT), Logistic Regression (LR), Naive Bayes (NB), and Support Vector Machine (SVM) as baseline classifiers, which are regarded as benchmarks. The quantitative analysis has been presented in terms of accuracy (ACC), Receiver Operating characteristic Curve (ROC), Area Under the Curve (AUC), Kolmogorov-Smirnov Statistic (KS), Cohen's Kappa Score (CKS), and Brier Score (BS). The experimental results affirm that ensemble learning performs better than individual learning. LR outperforms other baseline classifiers, whereas the RF (bagging DT) performs the best among the ensemble approaches, followed by XGB and LightGBM, respectively.

Keywords: Loan Approval; Machine Learning; Logistic Regression; XGBoost; SVM; KNN; Decision Tree; Random Forest

1. Introduction

The banking sector is a substantial pillar in the economic growth of any country. It caters to all sections of society by providing financial services, including deposits, withdrawals, and money lending. Based on population density and economic status of a particular place, various categories of banks exist to cater for these required services, including public sector banks, urban cooperative banks, rural cooperative banks, regional banks, and foreign banks. In India, the Bank of Hindustan was the first bank founded in 1770, while the Imperial Bank of India, now known as State Bank of India (SBI), is one of the largest commercial banks [1]. Reserve Bank of India (RBI) controls credits available to banks. It can increase and decrease the funds available by quantitative measures like Statutory Liquidity Ratio (SLR) and Cash Reserve Ratio (CRR).

*Corresponding author. Tel.: +91-661-246-2373; fax: +0-000-000-0000.
E-mail address: kumararun@nitrkl.ac.in

Banks lend loans to every sector for their growth and expansion, adhering to the country's overall development. The bank provides a loan by building a trust or guarantee that the recipient will repay the amount with certain added benefits as the interest rate. The three components of a loan include principal or the borrowed amount, rate of interest, and duration of availed loan [2]. When initiated by a borrower, a loan has to go through numerous procedures and checks to be sanctioned and approved. The first and foremost factor considered for loan approval is the borrower's credit score, which indicates the borrower's capability to repay the loan on time. Other factors are the borrower's income, employment status in case of job or business turnover, debt-to-income ratio in case of self-employed, Collateral and down-payment in case of previous loans. A few general factors include age, sex, marital status, family status, and relationship with a bank. Analysis of these factors predicts the borrower's position and eligibility for the loan issue. Assessment of Credit Score [3] is a climacteric issue pointed out by Banks nowadays to aid in evaluating whether a borrower is a defaulter.

Traditionally, banks employ highly trained professionals to evaluate applicants and ultimately decide a candidate's eligibility for a loan. A numerical credit score helps the authorities determine the likelihood that borrowers will repay the loan. Statistical algorithms [4] are implemented to determine credit scores based on a person's past payment history and other financial behaviour. However, with the explosion of data in today's society, the significance of statistical analysis models is hindered due to the assumptions made by these models while dealing with large data.

With the development of machine learning models, analysts are now equipped with more insight into big data and produce outcomes better than the statistical models [6]. Some of the machine learning used for this purpose are Linear discriminant analysis (LDA), logistic regression (LR), Naive Bayes (NB), Support Vector Machines (SVM), Random Forests (RF), and Neural Networks (NN) [5]. In recent works, researchers are exploring the ensemble learning models (bagging, boosting, and random space) and integrated ensemble machine learning models (RS-boosting and multi-boosting) instead of individual models [7, 8, 9]. Ensemble learning, including XGBoost and LightGBM, has become a winning strategy for various Kaggle competitions since Chen and Guestrin's [10] proposed XGBoost in 2016. Ensemble learning is a technique for producing superior performance by fusing various base models with specific tactics. There are two types of ensemble learning: homogeneous and heterogeneous. Homogeneous ensemble learning combines the same sort of classifiers, whereas heterogeneous ensemble learning includes classifiers of various kinds [11]. AdaBoost's boosting and random forest's bagging are examples of homogenous integration in ensemble learning while stacking is an example of heterogeneous integration.

In light of the above-mentioned scenario, this paper aims to discuss different ensemble methods and determine the best approach to identify whom to lend money to based on credit score.

1.1. Problem Statement

Credit score prediction for loan approval is a crucial decision faced by financial and banking sectors analyzing eligible customers. An automated loan approval system tends to reduce human resources and person-hours with faster operations. These money-lending organizations must have a robust model for predicting customers' credit scores, thereby minimizing the risk of defaulters. This binary classification problem states approved (Y: Non-defaulter) or not approved (N: Defaulter). The loan status is the dependent or target variable, whereas all other variables are independent of feature variables. The main focus is computing the customer's credibility for repayment by the behavioural study of the features. The literature shows no research for loan prediction based on these techniques. Thus, our work is the first to compare the ensemble algorithms for the loan dataset.

1.2. Contribution

The paper's main contribution is to provide a comprehensive comparative performance assessment of various ensemble methods for predicting Loan approval in the banking sector. Moreover, this study focuses on income, education, gender, marital status, credit history, and credit score, whereas most traditional approaches focus only on credit score. The feature analysis clearly shows the significance of a particular feature for the prediction. Ensemble algorithms, including bagging, boosting and Stacking, are considered with the NN, DT, LR, NB, and SVM as baseline classifiers. This study frames a few ex-ante hypotheses: RF (Bagging DT) performs better than other methods because, first, ensemble learning models outperform conventional models; second, the application of Bagging will significantly improve DT; and third, the execution time of these ensemble approaches are much less than the traditional approaches.

1.3. Organisation of the paper

The rest of the paper is organized as follows. Section II describes the literature study, followed by the proposed work in section III, describing the methodologies, data processing and analysis, and performance metrics. The modelling of machine learning algorithms and result analysis is presented in section IV. Finally, the paper concludes with the conclusion and future work in section V.

2. Related Work

Loan prediction and approval are among the most debated topics in finance and banking. Primarily, the loan approval analysis is based on credit risk assessment factors. Pertaining to the high demand for loans, automating loan processing and prediction of loan eligibility is necessary, thereby providing real-time service. The following is a survey of the work proposed recently in the literature.

Zhu et al. [12] proposed a random forest algorithm for loan prediction assessment and provided a comparative analysis of random forest, decision tree, SVM, and logistic regression model. Authors in [13] used the Lending Club dataset to predict loan borrowers' credit scores. The dataset was prepared using SMOTE and preprocessed using discretisation, removing unnecessary features and outliers. They compare the two machine learning models, Decision Forest and Decision Jungle Platform, by AzureML. Z. Eriez has used the "optimal" algorithm from BigML [14] to find the best machine-learning algorithm for the given data. Logistic regression, Decision Forests, and Neural network models were compared based on accuracy, precision, recall, and F1 score, and observed that Logistic Regression gives the most accurate results. Breeden [15] compared LR, SVM, KNN, NN, and boosting models and stated that it could not be one model that can be treated as the best model. All methods have certain advantages and limitations for different applications and datasets. There should be proper statistical and logical analysis for designing a model based on a study of the dataset to develop the best model. Further, image processing and quantum computing methodologies can be implemented for better credit risk prediction as per recent developments.

Ramachandra et al. [16] designed an automated loan approval system using the AWS cloud platform for affordability, ease of use, and scalability. The prediction is based on machine learning models, logistic regression, random forests, and decision trees. The accuracy is 86% by combining the decision tree model and logistic regression for confusion matrix analysis. Karthiban et al. [17] presented a comparative study of classification models like Naive Bayes, Linear Model, LR, DT, RF, and Gradient Boost (GBoost). The comparison predicts the Gradient Boost model to perform better than compared models with an accuracy of 98.6% followed by Decision Tree with 98.5%. Barboza et al.[18] compared SVM, bagging, boosting, and RF against various conventional approaches (discriminant analysis, LR, and NN) for predicting bankruptcy. Ma et al. [19] evaluated the performance of XGBoost and a prediction of P2P loan default using data cleaning methods of "multiple observation" and "multiple dimensional" respectively, and found that the LightGBM result based on multiple observational datasets performs the best. Alazzam et al. [20] demonstrated that boosting demonstrated a superior performance than bagging by predicting problematic modules of software systems utilising bagging, boosting, and stacking. Weighted RF and AdaBoost achieved the best accuracy, according to Jhaveri et al. [21], which used a range of classification and boosting algorithms to predict the Kickstarter campaign. Authors in [22] have taken real-time bank data and implemented machine learning to predict a customer's eligibility. The loan grant for a customer is dependent on the credit score. Various classification algorithms like KNN, Neural Networks, Naive Bayes, and Linear Regression are used for training and testing datasets.

From those mentioned above, it appears that various datasets produce various results for comparison studies. Our research, therefore, adds to this discussion and is anticipated to give a solid foundation for credit rating and default prediction. Compared to the existing literature, our work focuses on a comparative study of ensemble approaches like bagging, boosting and stacking, along with traditional approaches like NN, NB, SVM, LR, and DT. The methodologies implemented are EDA, data pre-processing, PCA, missing value analysis, and feature extraction. The performance evaluation is based on accuracy, AUC, KS, CKS, BS and execution time.

3. Proposed methodology

As shown in Fig. 1, the following steps are performed while experimenting with various methods for the loan prediction: Data collection, Exploratory Data Analysis (EDA), data pre-processing, model design, and evaluation of the designed model.

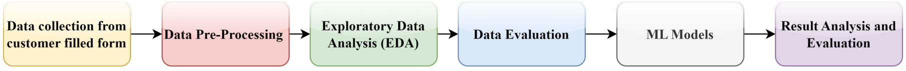
Data collection from customer filled form
Data Pre-Processing
Exploratory Data Analysis (EDA)
Data Evaluation
ML Models
Result Analysis and Evaluation

Figure 1: Flow of Implementation

3.1. Hypothesis

A regular hypothesis based on the data collected to validate customers for loan approval is listed below. This hypothesis may not be entirely intuitive, and machine learning models can validate it further.

- Education: The chance of loan approval for applicants with a higher education level, i.e. graduate level, should be increased.
- Income: Loan approval chance is higher for applicants with higher income.
- Loan amount: Loan approval is chance high in case of less loan amount.
- Loan term: The chance of loan approval is high for loans with a shorter period.
- Previous credit history: Loan approval chance is higher for applicants who have repaid their earlier debts.
- Monthly instalment amount: Loan approval chance is high if the monthly instalment amount is low.

3.2. Data Pre-Processing

The data generated from real-time is inconsistent and incomplete, lacks trends and patterns, and may contain unwanted erroneous data. Therefore, data pre-processing is required to formulate and develop clean data for analysis. The encoding mechanism using Wavelet Transformation (WT) and Principal Component Analysis (PCA) reduces the data size. Missing values in data and outliers may lead to model performance degradation and imperfect analysis. Thus, it is necessary to impute the missing values before data processing. We have used missing value attribution and outlier treatment [23] for data pre-processing. Mean and median for numerical values and mode for categorical values are used to fill in the missing values of the dataset. Similarly, having outliers in the dataset impacts statistical computations like standard deviation and mode and thus affects the data distribution. Log Transformation cuts down the larger values to smaller ones like Loan_Amount to get a normal uniform distribution. Logarithmic Transformation is a standard mathematical technique used in feature engineering. This mechanism handles skewed data and decreases outliers to achieve data normalization. The One Hot encoding is performed to generate numerical format and groups categorical data using the LabelEncoder library, which is a part of sklearn.preprocessing library.

3.3. Exploratory Data Analysis

This mechanism analyzes the dataset, summarizes its characteristics or features, and visualises data using graphical, statistical, or visualization tools. Understanding the dataset and its attributes before designing any model is essential. The data sets have three data types: object (categorical values), int64, and float64. Data type analysis is vital to differential categorical data from continuous data—this analysis aids in selecting an accurate model for analysis. The proposed work has implemented EDA with univariate and bivariate analysis on the datasets. The univariate analysis considers individual values for analysis, taking categorical and numerical features. Bi-variate is used to explore the target variable after univariate analysis and analyze the hypotheses generated.

3.4. Machine Learning Models

- **NB:** The NB is based on estimating the posterior probability according to the prior probability of a feature in the training data, so no significant parameter tuning is carried out.

- **NN model:** The NN model is built using a single hidden layer of four neurons with a "sigmoid" activation function, chosen on a trial-and-error basis. The model was trained for 1000 epochs.

- **SVM:** The two most important hyperparameters that can be used to train the best SVM model using an RBF kernel are gamma and C values. A training's influence is determined by the gamma parameter, with low values denoting "far" and large values denoting "near." The C parameter controls how tolerant the model is regarding incorrect classification. A model with a higher value of C will have very high accuracy. The SVM model uses a Gaussian Radial Basis Function with a gamma value of 0.01 and a control error (C) value set to 1.0. The computed accuracy is 71.79%

- **DT:** It splits the population into two or more homogeneous sets based on the most significant attributes or variables to create distinct groups. For DT, the impurity evaluation is done per the Gini Diversity Index, and the maximum depth is set to 5, along with the minimum sample leaf size set to 1. The accuracy is found to be 64.58% which is much less than Logistic Regression.

- **LR:** It predicts the probability of the existence of a certain model or class. It is used to estimate discrete binary values (0/1) based on sets of the independent variable(s). The values other than numerical values are changed to numerical values, like in the case of Gender, let Male be '1' and Female be '0'.

- **RF:** A trademark term for ensemble decision trees. It is a collection of decision trees that classify objects based on attributes, each tree classification. The classification with the highest votes compared to all trees in the forest is chosen. RF model is built considering n_estimators and max_features of the dataset. The number of trees for the RF model is set as 105, and the number of attributes is taken to be 10. The accuracy is computed to be 78.39%.

- **XGBoost:** It is a decision tree-based ensemble algorithm used to solve unstructured data prediction problems [24]. For XGBoost, learning_rate = 0.2, n_estimators = 10, objective = "multi:softmax", nrounds = 10, gamma = 1, min_child_weights = 2, and eval_metric = "auc". The accuracy is found to be 81.12%.

- **LightGBM:** Learning_rate = 0.3, n_estimators = 66, max_depth = 2, feature_fraction = 0.9, bagging_fraction = 0.6, and num_leaves = 31. Finally, after repeated combination experiments, two learners (RF, LightGBM) are selected as first-level classifiers for Stacking [9].

- **AdaBoost:** learning_rate = 0.2, n_estimators = 40, max_depth = 4. For Stacking, the two learners' models selected on repeated iterative combinations are RF and LightGBM as first-level classifiers and LR as second-level classifiers.

XGBoost and LightGBM were carried out using the Python libraries "xgboost" and "lightgbm", respectively. The other classifiers were run using the Python library "sklearn," where-as NN used the "keras." To avoid the contingency, the "5-fold" cross-validation accuracy is applied as a target to streng-then the robustness of the model and overcome the implications of overfitting and pick 80% samples as the training set, the remaining 20% samples as the testing set to assess the performance of the classifier. This stratification process rearranges data to ensure each fold represents the whole. It deals with variance and bias.

4. Results and discussion

The experiments with various base learners and ensemble techniques are conducted on a publicly available dataset and compared based on multiple quantitative parameters. The following subsections describe the dataset, performance metrics, and the results obtained.

4.1. Dataset

The dataset used for the experiments is taken from the Analytics Vidhya site, which is also taken from Kaggle [25]. The dataset consists of 614 customer entries with 13 data columns (Loan_ID, Gender, Marital_Status, Dependents, Education, Employment_Status, Income, Loan_amount, coApplicantIncome, Credit_History, Property, Loan_Status, LoamAmount_Term). The common patterns are extracted from the dataset, and the model is trained based on these extracted patterns. The data set consists of 13 columns and 614 rows in the training set and 12 columns and 367 rows in the test set.

4.2. Exploratory Data Analysis

EDA (univariate and bivariate analysis) is used for data analysis and feature understanding.

4.2.1. Univariate Analysis

There are five feature variables (Gender, Married, Self_Emplo-yed, that are categorically depicted in graphs of Fig. 2.

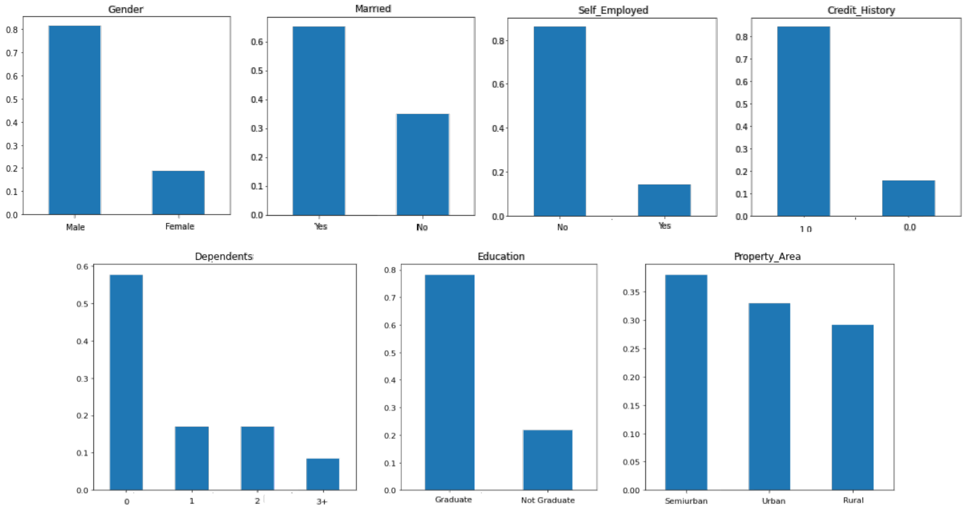

Figure 2: Loan Status Analysis

It is inferred from Fig. 2, that 80 % of applicants are male, 65 % are married, 15 % are self_employed, and 85 % have credit_history 1 for whom loans can be approved. Two features are ordinal variables in the categorical variable, i.e. (Dependents, Education and Property_Area). More than 50% of the applicants do not have any dependents, more than 75% are graduates, and most of them are from the semi-urban region.

4.2.2. Bivariate Analysis

Bivariate analysis is carried out to explore the relationship between the target and the categorical independent variable.

In Fig. 3, the Bivariate Analysis graph infers that the male and female proportion is nearly identical for approval of the loan. Loan status for applicants with one or three dependents is similar. No specific inference was drawn from employment vs status. The percentage of loans approved for graduates is higher. People with credit history 1 are likely to get their loan approved. Semi-urban areas loans are approved more than loans in rural and urban areas.

The relation between the target and the numerical independent variable is depicted in Fig. 3. Analysis of the mean income of people for whom the loan is approved vs those for whom the loan is not approved. Data Binning is done here to reduce the effect of minor observation errors. It mitigates bias in the model to transform data into a uniform distribution form. It is also known as quantization. It analyses the occurrence of quantitative data grouped into categories that are in the range of possible values. Bins for applicant income variables are computed based on values, and then the loan status in each bin is analysed.

The loans getting approved for applicants with low Total_Income is less than applicants with Average, High, and Very High Income depicted in Fig. 3. This result shows consistency with the hypothesis.

After binning the Loan_Amount variable, the loan approved is higher for low and average Loan_Amount than for high Loan_Amount, which goes with our hypothesis depicted in Fig. 3.

The bins for exploration parts are dropped. The values of dependents changed from +3 to 3 to make it numerical. All 'Y' is replaced by '1' and 'N' by '0'. Male by '1' and Female by '0' in the Gender variable. The correlation matrix is then plotted from these numerical values.

As shown in Fig. 4, the co-relation matrix infers that ApplicantIncome to LoanAmount has a correlation coefficient of 0.57, Credit_History to Loan_Status is 0.56 and Loan_Amount to CoApplicantIncome as 0.19 and so on.

The data then goes through Outlier Treatment and filling up of missing values with mean, median, and mode computation.

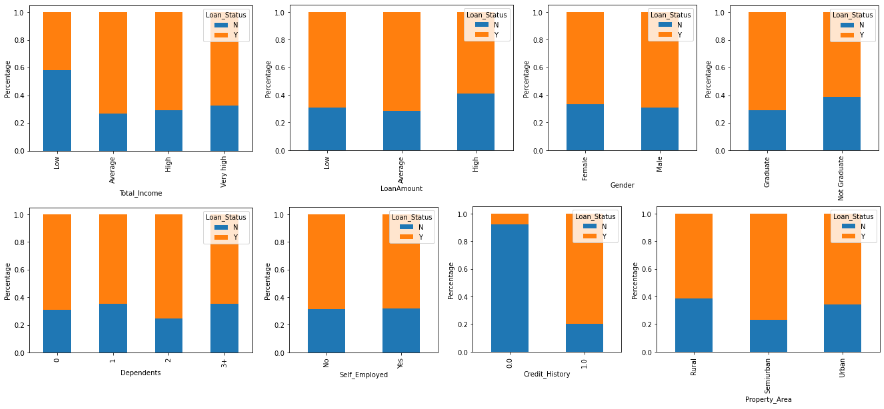

Figure 3: Bivariate Analysis

4.3. *Feature Importance*

Fig. 5 depicts the importance of various features for loan approval. This shows Credit_History is the most important feature, followed by Total Income Log, Total Income, Balance Income, Property Area Suburban, EMI. This is achieved by the mechanism of Feature Engineering, including imputation, PCA, univariate and bivariate selection.

4.4. *Results obtained using various models*

This study compares the performance of traditional and ensemble models—ANN, SVM, LR, DT, NB, AdaBoost, XGBoost, LightGBM, RF, and Stacking. The dataset was split into 80:20 ratio for training and test dataset, and then "5-fold" cross-validation was conducted.

4.4.1. *Performance Metrics*

The model-building procedure begins with model evaluation. The metrics considered for performance evaluation are Accuracy (ACC), precision, recall, specificity, F1-score, (Area Under the ROC Curve (AUC), Kolmogorov-Smirnov Statistic (KS), Cohen's Kappa Score (CKS), and Brier Score (BS).

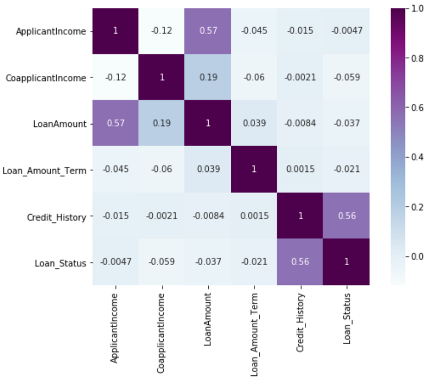

Figure 4: Co-relation Matrix

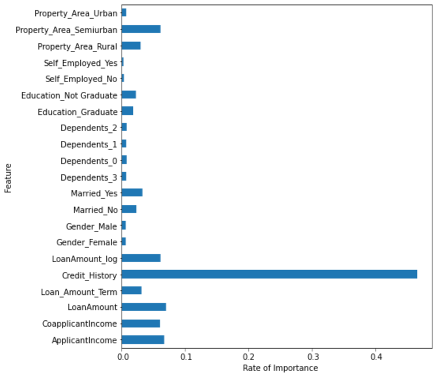

Figure 5: Feature Analysis

Table 1: Results obtained using various models

Model Name	ACC	AUC	KS	CKS	BS	Execution Time
NB	0.75	0.72	0.29	0.39	0.136	0.31s
LR	0.83	0.76	0.48	0.19	0.122	1.32s
NN	0.82	0.72	0.42	0.51	0.123	112s
SVM	0.73	0.66	0.32	0.41	0.121	421s
DT	0.77	0.71	0.33	0.46	0.122	2.34s
RF	**0.88**	**0.87**	**0.51**	**0.57**	**0.113**	0.82s
XGBoost	0.81	0.79	0.49	0.45	0.137	3.32s
AdaBoost	0.82	0.78	0.48	0.43	0.138	2.87s
LightGBM	0.81	0.79	0.49	0.44	0.137	1.27s
Stacking	0.80	0.73	0.46	0.43	0.139	23.57s

Table 1 depicts the obtained results using various models regarding various performance metrics. RF method is producing the best result in terms of ACC (88%), AUC (87%), and KS (51%). LR, NN, XGBoost, AdaBoost, and LightGBM produce satisfactory results with ACC 80-83%. While other models exhibit a subtle difference, RF continues to achieve the best results in CKS and BS. Except for AdaBoost, ensemble learning methods perform better overall than conventional models. Considering the classifiers and all performance metrics, LR has consistently been a common choice model and has outperformed most evaluation measures. Due to the subtractive nature of time, NN is the second-best classifier. The third-best models are SVM and DT, primarily attributed to DT's high variability and SVM's lack of a set standard for selecting an appropriate kernel function. The fact that the prediction result for NB is based on prior probability and that the assumed prior model may impact the outcome may be the cause of its poor performance as a classifier model.

Notably, AdaBoost has performed worse than some individual learners, even when, for loan prediction modelling, the ensemble learning can compete effectively against individual classifiers. Additionally, RF has achieved outstanding results, and the use of bagging has significantly improved DT. Since RF outperforms base learner DT in all the performance matrices, it can be claimed that ensemble learning is superior to an individual learning model.

5. Conclusion

The paper presents a comprehensive study of the machine learning methods to analyze the bank customers' dataset to predict the loan approval's worthiness. Various algorithms were employed to determine the best model for loan approval. The process of loan approval prediction started with data analysis by pre-processing, EDA, data-cleaning, imputation of missing values, outlier detection, and finally, designing models from train and test datasets. The dataset analysis focuses on not only the customer's credit history but also other customer details. Although the feature analysis shows that credit history is one of the essential features, other features also play a significant role, including income, EMI, loan amount, gender, employment, etc. The main insight of the paper is the empirical comparison of machine learning models. The comparison is done considering ensemble models like XGBoost, AdaBoost, LightBoost, and Stacking and some traditional models like NB, LR, NN, DT, SVM, and RF. The overall result analysis states that the ensemble approach outperforms the traditional approaches. To be more specific, RF shows the best result compared to all other models in terms of ACC, AUC, KS, CKS, BS and execution time. LR can be considered as the best classifier model with an accuracy of 83%. To sum up, LR, RF, XGBoost, and LightGBM can be considered the best choice for financial organizations for loan approval prediction based on defaulter detection.

In future, we plan to implement neural network models (ANN, CNN) and deep learning models and test their accuracy for loan approval prediction. As hyper-parameter tuning is essential, some more hyper-parameter optimization methods can be considered, like random search and Bayesian, to enhance the models' performance.

References

[1] Kumaran R. 2020 Study on Financial Inclusion Status in India, Doctoral dissertation.

[2] G Calcagnini, R Cole, G Giombini, and G Grandicelli 2016 Hierarchy of Bank Loan Approval and Loan Performance. *Econ Polit, Springer*, 35, 935–954.

[3] P Golbayani, I Florescu, R Chatterjee 2020 A Comparative Study of Forecasting Corporate Credit Ratings using Neural Networks Support Vector Machines, and Decision Trees. arXiv.

[4] A Karimi 2014 Evaluation of the Credit Risk with Statistical Analysis. *International Journal of Academic Research in Accounting, Finance and Management Sciences*, vol.4, issue 3.

[5] A M A M Al-Sartawia, K Hussaineyband, and A Razzaque 2022 The Role of Artificial Intelligence in Sustainable Finance *Journal of Sustainable Finance & Investment* DOI: 10.1080/20430795.2022.2057405.

[6] X Ma, J Sha, D Wang, Y Yu, Q Yang, X Niu 2018 Study on a Prediction of P2P Network Loan Default based on the Machine Learning LightGBM and XGboost Algorithms According to Different High Dimensional Data Cleaning. *Electron. Commer. Res. Appl*, 31, 24–39.

[7] Y Zhu, C Xie, G J Wang, and X G Yan 2017 Comparison of Individual, Ensemble and Integrated Ensemble Machine Learning Methods to Predict China's SME Credit Risk in Supply Chain Finance. *Neural Comput. Appl.*, 28, 41–50.

[8] O Sagi, L Rokach 2018 Ensemble learning: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 8, 1–18.

[9] W Liang, S Luo, G Zhao, and H Wu 2020 Predicting Hard Rock Pillar Stability using GBDT, XGBoost, and LightGBM Algorithms. *Mathematics*, 8, 765.

[10] T Chen, C Guestrin 2016 Xgboost: A Scalable Tree Boosting System. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 13–17, pp. 785–794.

[11] Y Li, and W Chen 2019 Entropy Method of Constructing a Combined Model for Improving Loan Default Prediction: A Case Study in China. *J. Oper. Res. Soc.*, 1–11.

[12] L Zhu, D Qiu, D Ergu, C Ying, and K Liu 2019 A Study on Predicting Loan Default based on the Random Forest Algorithm *The 7th International Conference on Information Technology and Quantitative Management (ITQM)*, Vol. 162, pp 503–13.

[13] K Alshouiliy, A. AlGhamdi, and D P Agrawal 2020 AzureML based Analysis and Prediction Loan Borrowers Creditworthy.*3rd International Conference on Information and Computer Technologies (ICICT)*, San Jose, CA, USA, 302-306.

[14] Z Ereiz 2019 Predicting Default Loans Using Machine Learning (OptiML). *27th Telecommunications Forum (TELFOR)*, Belgrade, Serbia, 1-4.

[15] J L Breeden 2020 Survey of Machine Learning in Credit Risk. Available at SSRN: https://ssrn.com/abstract=3616342.

[16] Ramachandra H V, Balaraju G, Divyashree R, Harish Patil 2021 Design and Simulation of Loan Approval Prediction Model using AWS Platform. *International Conference on Emerging Smart Computing and Informatics (ESCI)*, AISSMS Institute of Information Technology, Pune, India. Mar 5-7.

[17] R Karthiban, M Ambika, K E Kannammal 2019 A Review on Machine Learning Classification Technique for Bank Loan Approval. *International Conference on Computer Communication and Informatics (ICCCI)*, 23 – 25, Coimbatore, INDIA.

[18] F Barboza, H Kimura, E Altman 2017 Machine learning models and bankruptcy prediction. *Expert Syst. Appl.*, 83, 405–417.

[19] X Ma, J Sha, D Wang, Y Yu, Q Yang, X Niu 2018 Study on a Prediction of P2P Network Loan Default based on the Machine Learning LightGBM and XGboost Algorithms According to Different High Dimensional Data Cleaning. *Electron. Commer. Res. Appl.*, 31, 24–39.

[20] Alazzam I, Alsmadi I, Akour M 2017 Software Fault Proneness Prediction: A Comparative Study between Bagging, Boosting, and Stacking Ensemble and Base Learner Methods. *Int. J. Data Anal. Tech. Strateg.*, 9, 1.

[21] S Jhaveri, I Khedkar, Y Kantharia, S Jaswal 2019 Success Prediction using Random Forest, Catboost, XGboost and Adaboost for Kickstarter Campaigns *In Proceedings of the 3rd International Conference Computing Methodologies and Communication (ICCMC)*, Erode, India, 27–29, pp. 1170–1173.

[22] A S Aphale and S R Shinde 2020 Predict Loan Approval in Banking System Machine Learning Approach for Cooperative Banks Loan Approval. *International Journal of Engineering Research & Technology (IJERT)*, Vol. 9 Issue 8.

[23] Md A Sheikh, T Kumar 2020 An Approach for Prediction of Loan Approval using Machine Learning Algorithm. *Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC)*, ISBN: 978-1-7281-4108-4.

[24] C Zhang and Y Ma 2012 Ensemble machine learning: methods and applications. *Springer Science & Business Media.*

[25] Burak Ergün, Loan Data Set. *https://www.kaggle.com/burak3ergun/loan-data-set*, Accessed on 22-08-2023.