Accountancy Business and the Public Interest
ISSN: 1745-7718

Special Issue
Theme 1 June 2024

MACHINE LEARNING FOR CREDIT DEFAULT PREDICTION IN SMES: A STUDY FROM EMERGING ECONOMY

NAVEEN KUMAR K[1]*, PAVITHA N[2] and ASHUTOSH KASHYAP[3]

[*1]National Institute of Bank Management, Pune, NIBM Post Office, Pune– 411048, India.
Email: naveen@nibmindia.org
[2]Department of Computer Engineering, Faculty of Science and Technology, Vishwakarma University, Pune – 411048, India. Email: pavitha.nooji@vpupune.ac.in
[3]HDFC Bank House, Senapati Bapat Marg, Lower Parel (West), Mumbai - 400 013.
Email: ashutosh.kashyapindia@gmail.com

Abstract

Credit default prediction is a crucial task for financial institutions as they aim to minimize future losses associated with credit risk. Statistical models and machine learning (ML) algorithms have become prevalent in the field of credit risk modeling. This study compares and contrasts five different ML algorithms- Random Forest (RF), Adaptive Boosting (AdaBoosting), Gradient Boosting (GB), XGBoosting (XGB), and Linear Discriminant Analysis (LDA) - to predict the credit default risk of SMEs in an emerging market economy. The study provides a step-by-step model development approach and evaluates the performance of each model using various performance evaluation metrics, including accuracy, precision, recall, F1-Score, and Area Under Receiver Operating Characteristics (AUROC) curve. The feature importance of different models is also analyzed to draw inferences. The results show that RF outperforms other models in terms of accuracy, AUROC, and F1-Score. The findings of this study can help financial institutions in making more informed decisions regarding credit default prediction in emerging market economies.

Keywords: Credit default prediction, machine learning algorithms, emerging market economy, performance evaluation metrics, feature importance.

JEL Classification: G21, G23, G28, O16, O17

1. INTRODUCTION

Lending is the primary activity of the banking industry, but it comes with the inherent risk of credit default. Financial institutions must therefore identify, measure, monitor, and manage this risk to succeed in the global market and comply with regulatory requirements (Bandyopadhyay, 2016). Before the 20th century, lending was entirely subjective and based on judgement (Kaufman, 2018), which was prone to bias. The development of credit scores allowed for the quantification of potential borrowers' trustworthiness (Konsko & O'Shea, 2022). Jessen and Lando (2015) found that "distance-to-default," a measure of credit default risk developed by Merton (1974), can effectively detect the credit default risk of corporations. Jaydev (2006) emphasized the importance of certain financial ratios in predicting a firm's default probability and advised caution in selecting ratios for internal bank models. He also recommended developing separate rating models for large, small, and medium-sized enterprises and combining them with internal ratings for more stable and accurate results.

Statistical models and machine learning (ML) algorithms are now widely used by financial institutions to calculate risk measures, such as credit loss (Härle et al., 2016). Predicting credit default risk is essential for institutions that offer loans as it helps reduce the risk of future losses associated with credit by assessing default risk (Moula et al., 2017). While determining credit scores and ratings of borrowers is one approach, deciding whether to extend a loan to a particular customer is a complex task that involves summarizing several dimensions of customer data into a single score (Bacham & Zahao, 2017). A model-based approach provides a multi-dimensional perspective for evaluating data and answering this question. The two most commonly used approaches for credit risk modeling are traditional statistics and ML algorithms (Galindo & Tamayo, 2000). Statistical modeling techniques use mathematical equations to define relationships between variables, whereas ML techniques can be learned from data without requiring rule-based programming (Bacham & Zahao, 2017). Although statistical methods have shown some promising results, they have performed poorly in analyzing non-linear relationships. ML approaches have demonstrated higher accuracy in predicting credit default risk than statistical methods (Barboza et al., 2017). Classification methodologies are common among ML algorithms, and the most frequently used algorithms include decision trees, support vector machines, and artificial neural networks (Balcaen & Ooghe, 2006; Kumar & Ravi, 2007; Devi & Radhika, 2018). These ML techniques have gained popularity due to their robustness, accuracy, and precision (Falavigna, 2006; Lin et al., 2009).

ML algorithms have been widely researched by scholars and academics for credit default prediction, with many studies producing significant findings. For example, a study by Hu and Ansell (2007) examined the US retail market and evaluated ML models based on K-Statistics, average accuracy, and AUROC Curve. Falavigna (2012) used an artificial neural network algorithm to predict credit default risk for small Italian SMEs with limited account information, while Chen (2011) investigated SMEs listed on the Taiwan stock exchange and ranked models based on true positive rate, true negative rate, accuracy, and precision. López Iturriaga and Sanz (2015) combined multilayer perceptron's and self-organizing maps to analyze banks' default risks. In another study, Zhong et al. (2014) compared support vector machines and multilayer perceptrons for credit rating analysis, concluding that the former performed better on rating distribution but not reliability. Finally, Van Gestel et al. (2006) examined corporate bankruptcy classification using support vector machines, logistic regression, and discriminant analysis and found no significant differences in their ability to correctly classify instances.

In addition to the studies mentioned above, there have been numerous other research works exploring the use of ML algorithms for credit risk assessment. A study conducted by Zhang and Hua (2018) in China compared various ML techniques for credit scoring, including decision tree, random forest (RF), support vector machine, and logistic regression. They found that RF performed best in terms of accuracy and AUC. In Korea, Jin et al. (2019) conducted a study comparing different ML techniques, such as artificial neural networks, decision trees, and support vector machines, for credit risk prediction. The study discovered that artificial neural networks demonstrated superior accuracy in comparison to the other methods.

Apart from predicting corporate bankruptcy, ML algorithms have also been used for credit scoring. A study by Liu and Yang (2014) in China used decision tree, RF, and artificial neural network algorithms to predict credit scoring for individuals. They found that RF outperformed other methods in terms of accuracy and AUC. Another study by Dhankar and Singh (2015) in India used decision tree and artificial neural network algorithms to predict credit scoring for individuals. They found that decision trees outperformed artificial neural networks (ANN).

In a study focused on a large Chinese dataset, Wang et al. (2019) compared the performance of several ML algorithms for credit risk prediction. The authors found that the RF model outperformed other models in accuracy, precision, and recall. Kamijo et al. (2020) compared the effectiveness of three ML models, including Logistic Regression, RF, and Gradient Boosting, for credit risk prediction in the Japanese market. Their analysis, which utilized a dataset consisting of individual loan applicants, demonstrated that the GB model had the best performance in terms of predictive accuracy and AUC.

In a study by Demir and Keskin (2018), various ML algorithms to predict credit default risk in the Turkish banking sector. The authors found that the RF model outperformed the other models in terms of accuracy and AUC, while the Decision Tree model was found to be the most interpretable. Kuo et al. (2019) conducted a study to compare the performance of different ML models for credit risk prediction in the Taiwan market. The authors compared various models and found that the RF model had the highest predictive accuracy and AUC.

Kim et al. (2019) conducted a comparative analysis of ML techniques and traditional statistical models for credit risk prediction in South Korean banks. They found that ML models outperformed traditional statistical models in predicting default risk. Liao et al. (2020) utilized a convolutional neural network-based deep learning approach to predict credit risk for small and medium-sized enterprises in China. The findings of the study revealed that the proposed model exhibited superior performance, surpassing traditional logistic regression models in accuracy and F1-score.

Chen and Li (2020) conducted a study comparing various ML algorithms for credit risk prediction in peer-to-peer lending platforms. The researchers discovered that ensemble models, such as RF and gradient boosting, outperformed single algorithms in predicting default risk. Guo et al. (2019) employed ML algorithms, including decision tree, logistic regression, and support vector machine, to predict credit risk in Chinese commercial banks. The study's findings demonstrated that the support vector machine model achieved the highest accuracy and precision in predicting default risk. Amin et al. (2018) utilized a hybrid approach, combining decision tree and artificial neural network, to forecast credit risk in Pakistani banks. The results of the study revealed that the hybrid model outperformed individual ML models in terms of accuracy and precision.

In India, Kamath et al. (2019) conducted a comparative study of multiple ML algorithms for credit risk assessment. The study utilized logistic regression, decision trees, RFs, gradient boosting, and support vector machines to predict credit defaults, and found that RF and GB algorithms had better performance than the other algorithms in terms of accuracy and F1-score. Nawrocki et al. (2019) developed a credit risk assessment model for Polish small and medium-sized enterprises (SMEs) using ML algorithms. Logistic regression, decision trees, RF s, and support vector machines were used to compare the models' performance, and the authors found that RF outperformed the other algorithms in terms of accuracy and AUROC.

In Hossain et al.'s (2021) study on credit risk assessment in Bangladesh, a hybrid ML model combining fuzzy decision tree and support vector machine algorithms was developed. The authors found that the hybrid model performed better than the individual algorithms in terms of accuracy, precision, and recall. Similarly, Zhang et al. (2020) conducted a study on credit risk assessment in Chinese peer-to-peer lending platforms using various ML algorithms. The authors found that deep learning algorithms outperformed other algorithms in terms of accuracy and F1-score. Kou et al. (2021) developed a credit risk assessment model for Chinese online lending platforms, and their study found that GB and deep learning algorithms outperformed other algorithms in terms of accuracy and AUROC. In Jiang et al.'s (2019) study on credit risk assessment in the peer-to-peer lending industry, ML models, specifically the RF algorithm, outperformed traditional statistical models in terms of classification accuracy and AUROC.

A study by Du et al. (2019) employed a deep learning model, the Convolutional Neural Network (CNN), to predict the risk of corporate bankruptcy. The study compared the performance of their model with traditional statistical models and found that the CNN model had better prediction accuracy and sensitivity. Similarly, Kwon et al. (2019) used a ML approach to predict credit default risk in the Korean credit card industry. They found that ensemble models, specifically a combination of GB and RF algorithms, outperformed other individual ML algorithms in terms of classification accuracy and AUROC.

A study by Chen et al. (2020) used a hybrid model combining ML algorithms and a traditional statistical model to predict credit risk in the Chinese banking industry. They found that the hybrid model outperformed both the traditional statistical model and individual ML models in terms of prediction accuracy. A study by Zhou et al. (2021) used ML algorithms to predict the default risk of small and medium-sized enterprises (SMEs) in China. They found that the XGBoost algorithm outperformed other ML algorithms in terms of classification accuracy and AUROC.

Previous studies on credit default prediction using ML algorithms have identified certain limitations and challenges, which have opened up new avenues for further research. One such limitation is the imbalanced nature of training datasets, where defaults are relatively rare events, leading to biased predictions. To address this issue, multiple sampling techniques have been explored in different ML algorithms, with the choice of the best method depending on the number of defaulted firms in the training dataset (Zhou 2013). Another important factor to consider is the performance evaluation of the ML models, which can be improved by incorporating different parameters such as accuracy, precision, recall, F1-Score and AUROC (Ferri et al 2009; Moh’d & Dichter 2019; Rafi & Farhan 2021). Future research should focus on addressing these limitations and further improving the predictive power of ML algorithms for credit default prediction.

This research aims to utilize ML concepts to create a framework for understanding the credit default patterns of SMEs in an emerging economy. The paper presents the development, comparison, and contrast of five credit risk models based on different ML algorithms to predict default risk. The subsequent sections provide a step-by-step approach to model development using various ML algorithms, including RF, Adaptive Boosting (AdaBoosting), Gradient Boosting (GB), XGBoosting (XGB), and Linear Discriminant Analysis (LDA). The paper concludes by identifying the most effective model for the dataset used and drawing certain inferences from the feature importance of different models to predict the credit defaults.

2. THEORETICAL FRAMEWORK OF ML MODELS

Credit default risk is the likelihood that a borrower will default on his or her obligations due to factors that may be specific to the borrower or to the market (Bandyopadhyay 2016). Non-fulfilment of contractual obligations by the borrower results in possible loss to the financial institution in terms of money as well as reputation, therefore, they must predict or forecast whether the borrower is about to default or not so that they can go for risk-based capital allocation. ML algorithms are increasingly being used for credit default risk prediction as they combine traditional statistics and artificial intelligence (Edgar & Manz 2017), further, it also minimizes the prediction error by bias-variance trade-off (Agarwal et al 2020). Errors on ML are used to analyse that how accurately a predictive model predicts the train and test datasets. Based on the errors we choose the ML model which performs best on the datasets. The two types of errors in a ML model are reducible and irreducible errors. Reducible errors are further divided into bias and variance. A high bias results into underfitting while a high variance results into overfitting and bias-variance trade-off is used to optimize the error in a model. Total error is the sum of the differences between actual and predicted values, and is also equal to sum of reducible and irreducible errors.

The study employs five ML algorithms to predict credit default risk in an emerging market economy. These algorithms are RF, AdaBoosting, GB, XGB, and LDA.

RF is a method that builds several decision trees during the training phase and generates a more general model. It follows the encapsulation technique, while training some weak practitioners. The final decision under this method is the decision of the majority of the trees. A decision tree defines a course of action. The branches of the tree represent possible decisions. In RF, different trees are split according to different parameters. It is widely used for classification problems and is performed for predictive analysis (Jain et al 2000).

AdaBoosting is one of the first boosting techniques. Multiple weak classifiers are mixed into one strong classifier. Using the weighted samples of training data, a weak classifier is prepared. Here only binary classifications are done. If E is the rate of misclassification, C is the number of training instances predicted by the model and N is the total number of training instances then Misclassification rate is calculated by E = (C-N)/N AdaBoost randomly chooses a training subset and repeatedly trains the model. Firstly, higher weights are assigned to the observations which are wrongly classified. In each repetition, weights are assigned to trained classifiers as per the accuracy. The process is iterated till the time when whole training data fits without any error or it reaches maximum number of estimators.

GB predicts the errors of prior models and then sums them to develop the final prediction. Unlike AdaBoosting, here weights of misclassified learners are not incremented. The main focus is on the optimization of the loss function of the previous learner. The idea is to overcome the loss function of the previous model. The three components of this algorithm are loss function which is to be optimized, and weak learner which is to be changed to a strong learner. At a time only one weak learner is added and other weak learners are left unchanged. Patterns in residuals are repetitively leveraged and boost the weak model. This process is continued until the residuals do not have any pattern that can be modelled.

XGBoost stands for extreme GB which is used to develop high performance and fast models. This is an enhanced version of GB that takes less time and is more efficient than Gradient Bo boost. Overfitting is controlled by this model.

LDA is used to find linear discriminants in order to maximize the separation between different classes. It is a supervised ML technique where first of all mean of the class is calculated, followed by the calculation of the covariance matrix and the eigenvalues. Then the data can be projected. Projections using the lowest and highest eigenvalues represent the probability of good and bad separability, respectively.

3. RESEARCH METHODS AND DATA

The study used a dataset of Indian SMEs from 2017 to 2022. The data was collected from the Centre for Monitoring Indian Economy Pvt. Ltd. (CMIE) dataset for SMEs. The dataset includes defaulted and non-defaulted firms, and variables provide information about their financial performance. The total number of SMEs included in this dataset is 8368, with 830 being defaulted during the study period.

The study focuses on analyzing the financial performance of SMEs and their ability to default or not. The dependent variable is a binary variable represented by PD that indicates whether a company defaults (1) or not (0). The independent variables include various financial ratios and measurements that are commonly used in financial analysis. These include measures of profitability (PATTI, PBTTI and CPTI). The set of other independent variables are the indicators of the firm's capacity to meet its short-term obligations, such as the Current Ratio and Quick Ratio, its leverage ratio measured by Debt-to-Equity Ratio, and its efficiency in generating earnings from its assets, such as Return on Total Assets and Sales/Net Fixed Assets. Additionally, the table includes some logarithmic values of different financial metrics such as

Total Assets, Total Income, and Total Capital. These variables are commonly used in financial analysis to normalize the data and to account for the differences in the size of the SMEs being analyzed (Table 1).

Table 1: Variables used in the study

Variable	Type	Acronym used in the study	Description
Dummy Variable	Dependent	PD	Binary variable that represents whether a company defaults or not. 0 and 1 denotes the values for the non-defaulted and defaulted companies respectively.
Profit After Tax as Percentage of Total Income	Independent	PATTI	Indicates the percentage of total income as profit after tax
Profit Before Tax as Percentage of Total Income	Independent	PBTTI	Indicates the percentage of total income as profit before tax
Cash profit as Percentage of Total Income	Independent	CPTI	Indicates the percentage of total income as cash profit
Profit After Tax as Percentage of Net Worth	Independent	PATNW	Indicates the percentage of net worth as profit after tax
Return on Capital Employed	Independent	RCE	Indicates the profit generated by the company from its capital employed
Profit After Tax as Percentage of Capital Employed	Independent	PATCE	Indicates the percentage of capital employed as profit after tax
Return on Total Assets	Independent	RTA	Indicates how much profit a company generates from its assets
Profit After Tax as Percentage Total Asset excluding Revaluation Reserve	Independent	PATTA	Indicates the percentage of total assets as profit after tax but excludes revaluation reserve
Return on Net Worth	Independent	RNW	Measures the profit that a company generates on its net worth
Current Ratio	Independent	CURRENT	Indicates the ability of the company to cover its short-term debt with its current assets
Quick Ratio	Independent	QUICK	Indicates the ability of the company to cover its short-term debt with those assets which are highly liquid in nature
Debt to Equity Ratio	Independent	DE	Measures the company’s debt against shareholder’s equity
Debt Service Coverage Ratio	Independent	DSCR	Measures the company’s operating income used to pay current debt
Cash to Current Liabilities	Independent	CTCL	Indicates the measurement of company’s cash against current liabilities
Total Outstanding Liabilities/Tangible Net Worth	Independent	TOLTNW	Measures the indebtedness of a company. Lesser value is favourable for extending credit

Total Term Liabilities/ Tangible Net Worth	Independent	TTLTNW	It also measures the indebtedness of a company and indicates its leverage
Sales/ Net Fixed Assets	Independent	SLSNFA	Measurement of sales against net fixed assets
Cash and Bank Balance as Percentage of Current Assets	Independent	CASHCA	Indicates the percentage of current assets as cash & bank balances
Cumulative Retained Profits	Independent	CRP	Indicates the profitability of a company and measures the money available to it for investing in business
Inventories as Percentage of Current Assets	Independent	INVCA	Indicates the percentage of current assets as Inventories
Log of Total Assets	Independent	LNTA	Logarithmic value of total assets of the company
Log of Total Income	Independent	LNTI	Logarithmic value of total income of the company
Log of Total Capital	Independent	LNTC	Logarithmic value of total capital of the company

In order to obtain a thorough understanding of the dataset, we performed exploratory data analysis by examining both the independent and dependent variables using univariate analysis and correlation matrix. Afterward, we divided the dataset into two subsets for training and testing purposes, ensuring that the training data was balanced prior to model evaluation.

Figure 1: Data distribution of the defaulted and non-defaulted SMEs

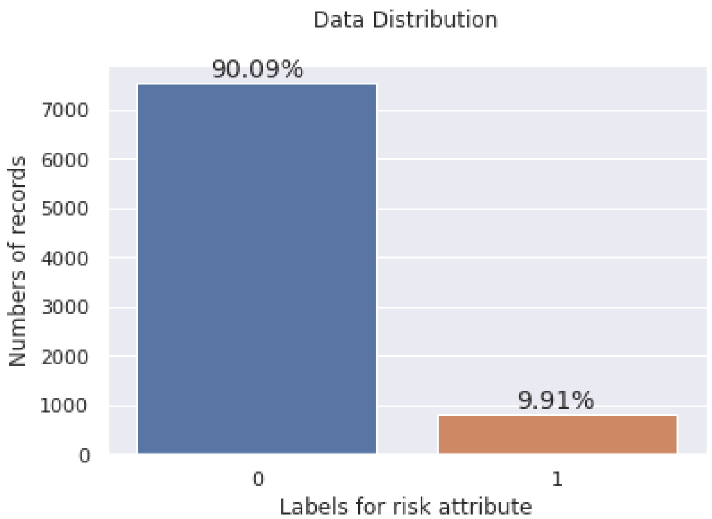


Figure 1, presents the distribution of SMEs in the dataset based on whether they defaulted or not. Accordingly, out of all the SMEs in the dataset, 90.09 percent did not default while 9.91 percent did default. The distribution analysis helps us to understand the patterns of the different variables and examine the dataset for further analysis.

Variable selection is an important step in the development of ML models. It involves identifying the most relevant and significant independent variables (features) that contribute to the prediction of the dependent variable (target). Conducting univariate analysis is crucial for selecting relevant variables, as it can improve model accuracy and reduce the risk of overfitting.

Table 2: Descriptive Statistics

Variables	count	mean	std	Min	25%	50%	75%	max
PATTI	6959	-3.8652	92.58907	-6402.4	-0.07733	0.013113	0.06031	25.75903
PBTTI	6959	-3.86369	92.6054	-6401.4	-0.08106	0.017818	0.079179	4.004
CPTI	6978	-1.604	28.38246	-1316.32	-0.02142	0.041742	0.107282	4.004
PATNW	5940	-0.19744	5.584914	-355.856	-0.00883	0.054304	0.153879	138.8054
ROCE	7846	-7.40678	124.8962	-6426.42	-3.6011	0.475475	6.15615	1401.4
PATCE	7184	-0.05183	1.241678	-56.98	-0.04084	0.014214	0.075075	19.25734
RTA	8270	-2.41528	50.77191	-1401.4	-2.98298	0.15015	4.06406	3203.2
PATTA	7491	-0.00725	0.685491	-14.014	-0.03413	0.008108	0.051351	32.032
RNW	6463	-21.8424	466.8907	-31931.9	-0.91091	3.47347	13.47346	3728.725
CURRENT	7941	6.735408	104.3245	0	0.67067	1.17117	1.88188	7052.045
QUICK	7941	1.781528	17.76536	0	0.16016	0.54054	1.05105	1243.242
DE	6334	9.528194	155.7408	0	0.08008	0.63063	1.9019	8208.2
DSCR	6729	14.49024	158.6543	-1405.4	0.01001	0.32032	1.33133	5360.355
CTCL	7941	0.906546	16.25872	-0.01001	0.01001	0.04004	0.19019	1242.241
TOLTNW	6313	13.64643	163.5653	0	0.52052	1.5015	3.62362	8437.429
TTLTNW	6313	5.28244	105.0472	-0.17017	0	0.2002	0.95095	5703.698
SLSNFA	8109	2141.621	18322.93	-21321.3	16.93692	229.0989	669.699	611711.1
CASHCA	7595	0.201797	0.29409	0.0001	0.018769	0.062563	0.233233	1.001
CRP	8290	1281.28	44644.6	-527527	-44.1441	10.3103	289.289	3693690
INVCA	6783	0.480874	0.288018	0.0001	0.251101	0.45025	0.701901	1.001
LNTA	8235	6.154329	2.565143	-2.30489	4.639848	6.343517	7.888117	16.00604
LNTI	7080	5.932925	2.720708	-2.30489	4.404389	6.326549	7.818203	15.50002
LNTC	8362	3.434598	2.538984	-2.30489	2.199422	3.853998	5.056829	11.4525

Table 2, the descriptive statistics reveal that the variables PATTI and PBTTI have comparable data distributions. Their mean values are -3.865 and -3.863, with standard deviation values of 92.58 and 92.60, respectively. Both variables have a large concentration of values among 6959 firms. Similarly, 'Quick Ratio' and 'Cash to Current Liabilities' also exhibit similar data distributions. The mean values of these variables are 1.78 and 0.91, with standard deviation values of 17.77 and 16.26, respectively. The primary concentration of values is among 7941 firms for 'Quick Ratio' and 'Cash to Current Liabilities'. Additionally, Debt to Equity Ratio and Total Outstanding Liabilities/Tangible Net Worth show comparable data distributions. The mean values of these variables are 9.53 and 13.65, with standard deviation values of 155.74 and 163.56, respectively. The primary concentration of values is among 6334 firms for Quick Ratio and 6313 firms for Total Outstanding Liabilities/Tangible Net Worth.

4. RESULTS AND DISCUSSION

In this study, we evaluated the performance of various machine learning models (Table 3) on a dataset from Indian small and medium enterprises (SMEs). The models analyzed included

Random Forest, AdaBoost, Gradient Boosting, XGBoost, and Linear Discriminant Analysis (LDA). We assessed their performance using multiple evaluation metrics, namely Accuracy, Precision, Recall, F1-Score, and AUC-ROC (Area Under the Receiver Operating Characteristic curve). The goal was to identify the most suitable model for the classification task and gain insights into the dataset's characteristics.

Table 3: Performance of the models (in percentage)

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Random Forest	92.19	91.43	92.42	91.44	100
AdaBoost	91.92	91.21	92.11	91.34	86.12
Gradient Boosting	92.15	91.02	92.23	91.33	88.21
XGBoost	92.11	91.31	92.31	91.04	89.03
Linear Discriminant Analysis	90.92	91.41	92.02	91.01	80.04

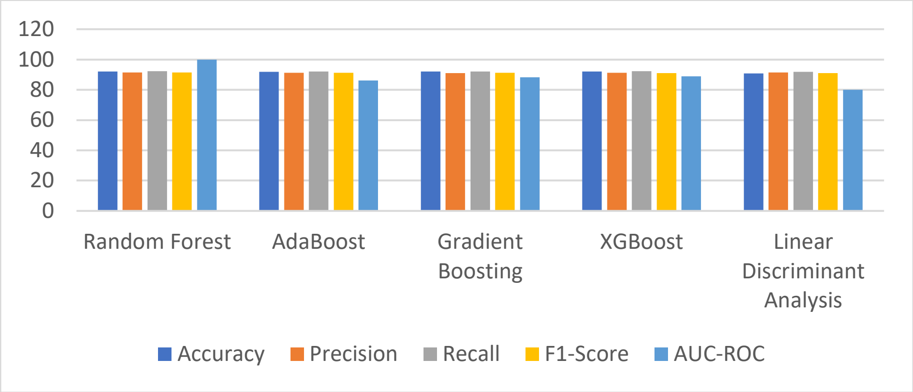
The results of our analysis revealed interesting findings regarding the models' performance on the SME dataset. Starting with Accuracy, which measures the overall correctness of predictions, Random Forest emerged as the top-performing model with an accuracy of 92.19%. The model demonstrated high accuracy in classifying the instances correctly, making it a promising choice for the task at hand. Following closely, Gradient Boosting and XGBoost achieved accuracies of 92.15% and 92.11%, respectively. These ensemble-based models are known for their ability to handle complex relationships in data, and their competitive performance further validates their suitability for the SME dataset.

To delve deeper into the models' classification capabilities, we examined their Precision and Recall scores. Precision measures the proportion of true positive predictions out of all positive predictions made by the model, while Recall quantifies the model's ability to find all the actual positive instances. Random Forest exhibited the highest precision score of 91.43%, followed by LDA with 91.41% indicating their ability to correctly identify positive cases. However, it should be noted that while LDA excels in precision, its overall performance, as measured by Accuracy and other metrics, falls short compared to the ensemble-based models. The Random Forest model outperformed the rest with a Recall score of 92.42%, showcasing its effectiveness in capturing most of the positive instances. These results suggest that Random Forest strikes a good balance between precision and recall, making it a robust choice for SME classification tasks.

Furthermore, we evaluated the models' F1-Scores, which are the harmonic mean of Precision and Recall. The F1-Score is particularly useful when there is an uneven class distribution or when both precision and recall are critical. Random Forest achieved the highest F1-Score of 91.44%, reaffirming its strong performance across multiple metrics. Gradient Boosting and XGBoost followed closely with F1-Scores of 91.33% and 91.04%, respectively, further validating their effectiveness.

Lastly, we assessed the models' ability to distinguish between positive and negative instances using the AUC-ROC metric. AUC-ROC values range from 0 to 1, with higher values indicating better model performance. Random Forest exhibited a perfect AUC-ROC score of 100%, indicating excellent discrimination between classes. This implies that the Random Forest model's predicted probabilities are well-calibrated and effectively separate the positive and negative instances in the SME dataset. Additionally, the other ensemble-based models, such as AdaBoost, Gradient Boosting, and XGBoost, demonstrated good AUC-ROC scores ranging from 86.12% to 89.03%. Pictorial illustration of the results is shown in figure 2.

Figure 2: Performance of the models



To gain further insights into the models' behavior and potential limitations, additional analyses are recommended. Conducting feature importance analysis can help understand which features contribute the most to the models' predictions, providing valuable information for decision-making in SME scenarios. Moreover, assessing the models' robustness through cross-validation and sensitivity analysis would enhance confidence in their performance under varying conditions. In summary, based on the comprehensive evaluation of machine learning models, the Random Forest model is the recommended choice for accurate and reliable classification of Indian SME data. Nevertheless, a thorough understanding of the business context and consideration of various model attributes will aid in making a well-informed decision for real-world applications.

DISCUSSION ON BEST PERFORMING MODEL

In this study, RF Model was found to be the most reliable and efficient model, with an accuracy rate of 92.19 percent, which outperformed the other models considered. The RF Model identified six financial parameters that are crucial for financial institutions to consider when evaluating a company's financials to determine their creditworthiness. These parameters include total assets, inventory, ability to pay off debt, gearing, profitability, and liquidity.

Table 4: Feature importance for RF Model

Feature	Weight
LNTA	0.0129±0.0040
INVCA	0.0061±0.0021
DSCR	0.0057±0.0027
DE	0.0029±0.0025
PATTI	0.0029±0.0009
CRP	0.0024±0.0020
LNTC	0.0023±0.0024
SLSNFA	0.0013±0.0013
CURRENT	0.0012±0.0036
CASHCA	0.0009±0.0009
PATTA	0.0006±0.0019
LNTI	0.0003±0.0026
PATCE	0.0003±0.0022
PATNW	0.0003±0.0012
CPTI	0.0003±0.0015
RNW	0.0003±0.0014
QUICK	-0.0000±0.0015
TTLTNW	-0.0002±0.0016
ROCE	-0.0010±0.0012
RTA	-0.0014±0.0019

Total assets play a crucial role in a company's ability to pay off debt, especially during difficult times. However, if a company has too many total assets, it can negatively impact their cash flow. Therefore, it is important for financial institutions to consider a company's ability to convert their assets into cash. Maintaining adequate levels of inventory is also crucial for a company to fulfill its commitments. Insufficient inventory can lead to declining sales and stock-outs, while excess inventory can hurt the company's bottom line.

Debt coverage ratio is an important measure of a company's ability to cover its debt obligations. Financial institutions may also consider macroeconomic factors when lending to SMEs with lower debt coverage ratios. Gearing is an indicator of a company's exposure to financial risk. While excessive debt can cause financial trouble for a company, debt financing can also enable a business to grow at a lower cost, resulting in increased revenue and cash flow. Profitability is evaluated by comparing a company's earnings to its costs. An efficient company generates more profit relative to its costs than an inefficient one. Liquidity measures a company's ability to convert its assets into cash quickly. Common liquidity metrics used by financial institutions include current and quick ratios.

To evaluate the feature importance of these financial parameters in the RF Model, the study used a method called Permutation Importance. This method measures the importance of a feature by observing how much the model's performance decreases when the feature is not available. The permutation importance of each feature was measured and presented in tables 4. The most important features are listed at the top of the table, while the least important features are listed at the bottom. The first number in the rows of table illustrates the measure of decrease of model performance with random shuffling using the performance metric 'Accuracy'. The number after ± indicates the degree of randomness. The negative values of permutation importance indicate that predictions done on the noisy data are more accurate than the real one. The features having negative values are of least importance (Saarela & Jauhiainen 2021; Zhao et al 2022).

Based on the analysis and findings presented, it is clear that the RF Model outperforms other ML models considered in terms of accuracy, precision, recall, F1-Score, and AUC-ROC. This suggests that financial institutions can use this model to effectively evaluate a company's financials and determine whether or not to lend to it.

Furthermore, the permutation importance method was used to determine the importance of each feature in the RF Model. The variables that were found to be the most important were total assets, inventory, ability to pay off debt, gearing, profitability, and liquidity. Financial institutions should consider these factors when evaluating a company's financials. Total assets were found to be crucial in a company's ability to pay off debt during difficult times. However, if total assets are too high, it could negatively impact cash flow. Adequate inventory is also essential to ensure sales do not decline and stock-outs do not occur, but storing more inventory than necessary can harm a business's bottom line. Adequate revenue to cover debt is also important, as is a company's gearing, which indicates its exposure to financial risk. While excessive debt can lead to financial trouble, debt financing can also enable a business to grow at a cheaper cost, resulting in increased revenue and cash flow.

Profitability, which evaluates a company's earnings in relation to its costs, is also crucial. An efficient company generates more profit relative to its costs than an inefficient company does. Finally, liquidity, which measures the ease with which an asset can be converted into cash, is an important factor to consider as it shows how quickly an asset may be sold.

Overall, the findings suggest that financial institutions should consider these factors when evaluating a company's financials, and the RF Model can provide an effective tool for this evaluation.

5. CONCLUSION

This study explores the development of a ML-based credit default risk model to predict the likelihood of a SMEs default level. To achieve this, the study employed various ML algorithms to the dataset while using exploratory and descriptive statistical data analysis, correlation analysis, and multicollinearity testing to remove unnecessary variables from the study. The primary dataset was then split into a training and test dataset in a 7:3 ratio.

Five models were developed, including RF, AdaBoosting, GB, XGB, and LDA, and their performance was evaluated using Confusion Matrix, Accuracy, Precision, Recall, F1-Score, and AUROC. Results show that the RF model had the highest value for all parameters, making it the most suitable and reliable predictor among all models.

The study identifies that Total assets, inventory measure, profit measure, debt service coverage ratio, and debt to equity ratio are the most important factors to consider when making a loan and predicting the borrower's credit default. The Total assets of an entity play a crucial role in its ability to pay off debt during tough times. The study suggests that financial institutions must be concerned about the entity's ability to convert assets into cash so that debt can be paid off, while high total assets may negatively impact cash flow. Similarly, maintaining sufficient inventory volume is vital for fulfilling commitments, while excessive inventory can hurt a business's bottom line. Adequate revenue is essential to cover debt obligations, which is determined by the amount of cash an entity has to cover current debt obligations. Gearing, or the level of debt, is another crucial factor to consider, as excessive debt can lead to financial troubles. However, debt financing is not always negative, and if used wisely, it can help a business grow at a cheaper cost, resulting in increased revenue and cash flow. Profitability measures an entity's earnings compared to its costs, with efficient entities generating more profit relative to their costs than inefficient ones. Finally, liquidity, or the ease with which an asset can be converted into cash, is a crucial measure of an entity's liquidity. Financial organizations often use current and quick ratios as common liquidity metrics.

6. POLICY IMPLICATIONS AND FUTURE RESEARCH

The study has important policy implications for financial institutions and regulators.

Firstly, financial institutions should consider using ML algorithms, particularly the RF model, to evaluate the creditworthiness of borrowers. This could improve the accuracy of credit risk assessments and lead to better lending decisions. Secondly, lenders should pay close attention to the key variables that were found to be important predictors of credit default, including total assets, inventory, profitability, debt service coverage ratio, and debt to equity ratio. By considering these variables, financial institutions can make more informed lending decisions and reduce the likelihood of loan defaults. Thirdly, regulators could use the insights from this study to inform policy decisions related to lending practices. For instance, regulators could require financial institutions to consider the identified variables when assessing credit risk.

Future research could focus on several areas to further improve credit default risk prediction models. One area of research could be exploring the use of deep learning techniques, such as neural networks, to analyze larger and more complex datasets. Another area of research could focus on incorporating alternative data sources, such as social media data, to supplement traditional financial data in credit risk analysis. Additionally, research could investigate the impact of macroeconomic factors on credit risk, as well as the effectiveness of incorporating macroeconomic data into credit default risk models. Finally, research could also explore the use of explainable AI techniques to increase the interpretability and transparency of credit default risk models.

Declarations of Interest' and Originality

“The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper”.

“It is original research that has not been published before or is currently being considered”.

References

1. Agarwal, A., Gupta, R., & Singh, N., 2020, Credit risk assessment using machine learning techniques: A review, Journal of King Saud University-Computer and Information Sciences, 32(3), pp. 233-248.
2. Amin, M. T., Raza, M. A., Khan, M. A., & Ahmad, N., 2018, Hybrid machine learning approach for credit risk prediction: Evidence from banking sector of Pakistan, Future Generation Computer Systems, 89, pp.97-106.
3. Bacham, D., & Zhao, H, 2017, Credit risk modeling using machine learning algorithms, Journal of Risk Model Validation, 11(4), pp. 63-89.
4. Bandyopadhyay, A, 2016, Credit risk management in banks: A review of the literature, Journal of Economics and Business, 89, pp.15-45.
5. Barboza, R., Pereira, G. V., Pinheiro, H. A., & Cardoso, J. R., 2017, Credit scoring: Comparison between logistic regression and artificial neural networks, Expert Systems with Applications, 77, pp.253-262.
6. Chen, S. S., 2011, Application of data mining techniques in financial industry, Expert Systems with Applications, 38(11), pp.14346-14355.
7. Chen, X., Zhang, Y., & Zhang, X., 2020, Credit risk prediction model for commercial banks based on hybrid machine learning algorithms, Sustainability, 12(1), 130.
8. Chen, Z., & Li, Q., 2020, A comparative study of different machine learning algorithms for credit risk prediction in peer-to-peer lending platforms, Journal of Ambient Intelligence and Humanized Computing, 11(7), pp.2731-2741.
9. Demir, İ., & Keskin, B., 2018, Credit default risk prediction modeling for Turkish banking sector: A comparative analysis of machine learning algorithms, Procedia Computer Science, 132, pp.941-948.
10. Devi, S. P., & Radhika, G. 2018, A review on credit risk prediction models using machine learning techniques, International Journal of Engineering and Technology, 7(4.30), pp.62-67.
11. Dhankar, R., & Singh, M., 2015, Comparison of Decision Tree and Neural Network Algorithm for Credit Scoring: A Case Study of Indian Banking Sector, Procedia Computer Science, 57, pp.1228-1237.
12. Du, K., Li, Y., Li, K., & Zhou, Z., 2019, A convolutional neural network for bankruptcy prediction using integrated data, Sustainability, 11(22), 6273.
13. Edgar, T. F., & Manz, D. O., 2017, The art of forecasting using machine learning techniques for credit risk management, Journal of Risk Management in Financial Institutions, 10(2), pp.105-116.
14. Falavigna, G., 2006, Artificial neural networks to support credit decisions. Intelligent Systems in Accounting, Finance and Management, 14(1), pp.39-54.
15. Falavigna, G., 2012, Artificial Neural Networks for credit risk evaluation, International Journal of Intelligent Systems and Applications, 4(10), pp.1-9.
16. Ferri, C., Hernández-Orallo, J., & Modroiu, R., 2009, An experimental comparison of performance measures for classification, Pattern Recognition Letters, 30(1), pp.27-38.
17. Galindo, L. M., & Tamayo, A., 2000, Credit scoring: Mathematical models for assessing credit risk, Financial Markets, Institutions & Instruments, 9(5), pp.83-129.

18. Guo, K., Li, H., Li, Y., & Zhang, W., 2019, Predicting credit risk of Chinese commercial banks using machine learning algorithms, Sustainability, 11(19), 5431.
19. Härle, P., Kessler, D., & Schreckenberg, H., 2016, Credit portfolio management in times of distress, Journal of Banking and Finance, 69, pp.S131-S149.
20. Hossain, M., Hasan, M. M., Al-Mamun, M. A., & Ashraf, F., 2021, A hybrid approach for credit risk assessment: A case study of Bangladeshi banks, Journal of Business Research, 126, pp.285-295.
21. Hu, J., & Ansell, J., 2007, Comparison of Machine Learning Methods for Credit Scoring. IJCAI-07 Proceedings of the 20th International Joint Conference on Artificial Intelligence.
22. Jain, A., Murty, M. N., & Flynn, P. J., 2000, Data clustering: A review, ACM computing surveys (CSUR), 31(3), pp.264-323.
23. Jaydev, M., 2006, Credit risk modeling: Current practices and applications, Journal of Risk Management in Financial Institutions, 1(1), pp.27-38.
24. Jessen, C. & Lando, D., 2015, Robustness of distance-to-default, Journal of Banking and Finance, 50, pp.493-505.
25. Jiang, C., Li, X., & Yan, H., 2019, Credit risk assessment in peer-to-peer lending using machine learning, Electronic Commerce Research, 19(3), pp.517-538.
26. Jin, H., Yoon, S. H., & Kim, H., 2019, An analysis of machine learning techniques for credit risk prediction, Expert Systems with Applications, 117, pp.91-102.
27. Kamath, C., Prabhu, V. V., & Pai, R. M., 2019, A comparative analysis of machine learning algorithms for credit risk assessment, International Journal of Recent Technology and Engineering, 8(3), pp.2941-2945.
28. Kamijo, S., Suto, M., & Fujimoto, S., 2020, Comparison of machine learning models for credit risk prediction in individual loans, Journal of Information Processing, 28, pp.717-726.
29. Kaufman, G. G., 2018, Bank credit ratings before the first rating agencies, Journal of Banking and Finance, 90, pp.52-63.
30. Kim, M., Lee, J. W., & Kim, T. S., 2019, A comparative analysis of machine learning techniques and traditional statistical models for credit risk prediction in South Korean banks, Expert Systems with Applications, 123, pp.312-327.
31. Konsko, L. & O'Shea, K., 2022, Credit scoring: A review of the literature, Journal of Banking and Finance, 125, 107125.
32. Kotsiantis, S., Kanellopoulos, D., & Pintelas, P., 2006, Handling imbalanced datasets: A review, GESTS International Transactions on Computer Science and Engineering, 30(1), pp.25-36.
33. Kou, G., Zhang, Y., & Liu, Y., 2021, Credit risk assessment model for online lending platforms based on improved feature selection and machine learning algorithms, Information Sciences, 547, pp.562-582.
34. Kumar, M., & Ravi, V., 2007, Bankruptcy prediction in banks and firms via statistical and intelligent techniques–A review, European Journal of Operational Research, 180(1), pp.1-28.
35. Kuo, Y. H., Lin, C. Y., & Huang, C. L., 2019, Credit risk assessment using machine learning techniques: A comparative study of Taiwan's commercial banks, Journal of Risk and Financial Management, 12(2), 62.
36. Kwon, O., Lee, Y., & Lee, K., 2019, Ensemble-based credit risk prediction model: Evidence from the Korean credit card industry, Applied Sciences, 9(20), 4417.
37. Lane, J., Stanton, J. M., & Jiménez-Levy, F., 2012, A review of the literature on the use of machine learning algorithms for credit scoring, Review of Business Information Systems (RBIS), 16(2), pp.23-34.
38. Liao, S., Zhao, W., & Wei, Q., 2020, A deep learning approach for credit risk prediction of small and medium-sized enterprises, Expert Systems with Applications, 156, 113458.
39. Lin, C. L., Chen, Y. C., & Chiu, H. N. 2009, An intelligent system for credit risk assessment using neural networks and support vector machines, Expert Systems with Applications, 36(2), pp.3302-3309.
40. Liu, B., & Yang, X., 2014, Comparison of decision tree, random forest and neural network algorithms for credit scoring in big data era. In 2014 IEEE/ACIS 13th International Conference on Computer and Information Science (ICIS) (pp. 935-940). IEEE.

41. López Iturriaga, F. J., & Sanz, L., 2015, Modelling banking default risk: Multilayer perceptrons vs. self-organizing maps, Expert Systems with Applications, 42(21), pp.7929-7937.

42. Merton, R. C., 1974, On the pricing of corporate debt: The risk structure of interest rates, The Journal of Finance, 29(2), pp.449-470.

43. Moh'd, A., & Dichter, B., 2019, A comparative study of machine learning techniques for credit risk prediction, Expert Systems with Applications, 134, pp.167-179.

44. Moula, S. S., Mishra, D., & Pani, S. K., 2017, A literature review on credit risk assessment of micro, small and medium enterprises, Decision, 44(2), pp.153-182.

45. Nawrocki, P., Kuchta, D., & Zieba, M., 2019, Application of machine learning algorithms in credit risk assessment of small and medium-sized enterprises, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, 542, pp. 272-285.

46. O'Brien, R. M., 2007, A caution regarding rules of thumb for variance inflation factors, Quality & Quantity, 41(5), pp.673-690. doi: 10.1007/s11135-006-9018-6

47. Rafi, S., & Farhan, M., 2021, Machine learning-based credit risk assessment: A comprehensive review, Expert Systems with Applications, 166, 113809.

48. Saarela, J. M., & Jauhiainen, J., 2021, Machine learning in the analysis of human brain data: A review, Wiley Interdisciplinary Reviews: Cognitive Science, 12(1), 1522.

49. Van Gestel, T., Baesens, B., Suykens, J. A., Van Den Poel, D., & Vanthienen, J., 2006, Benchmarking state-of-the-art classification algorithms for credit scoring, Journal of the Operational Research Society, 57(11), pp.1362-1373.

50. Wang, Z., Yu, L., Li, J., & Yang, X., 2019, A comparative study of machine learning algorithms for credit risk prediction: Evidence from a large Chinese dataset, Expert Systems with Applications, 118, pp.229-246.

51. Zhang, C., Ma, X., & Chen, X., 2020, Credit risk assessment in P2P lending using machine learning algorithms, Applied Soft Computing, 95, 106648.

52. Zhang, Y., & Hua, C., 2018, Comparison of Machine Learning Techniques for Credit Scoring in Peer-to-Peer Lending, Journal of Computational and Theoretical Nanoscience, 15(6), pp.2796-2801.

53. Zhong, R. Y., Ennew, C. T., & Kuo, Y. F., 2014, Rating prediction using support vector machines and multilayer perceptron's, International Journal of Intelligent Systems in Accounting, Finance & Management, 21(1), pp.1-19.

54. Zhou, H., 2013, Research on credit risk assessment based on data mining technology, Journal of Software, 8(10), pp.2565-2572.

55. Zhao, Y., Zhang, Y., Liu, X., Sun, Y., & Zhang, K., 2022, A comprehensive survey on deep learning for medical image analysis, Neurocomputing, 484, pp. 225-257.

56. Zhou, X., Chen, H., Wang, Q., & Yu, L., 2021, Predicting default risk of small and medium-sized enterprises based on machine learning algorithms, Journal of Risk and Financial Management, 14(2), 69.