INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Identification of Most Preferable Machine Learning Technique for Prediction of Bank Loan Defaulters

Digambar B Uphade[1]*, Aniket A Muley[2], Swapnil V Chalwadi[3]

1 Department of Statistics, KRT Arts, BH Commerce and AM Science College, Nashik, 422002, Maharashtra, India
2 School of Mathematical Sciences, Swami Ramanand Teerth Marathwada University, Nanded, 431606, Maharashtra, India
3 School of Liberal Arts, Dr. Vishwanath Karad MIT World Peace University, Pune, 411038, Maharashtra, India

OPEN ACCESS

Received: 24-11-2023

Accepted: 28-12-2023

Published: 20-01-2024

Citation: Uphade DB, Muley AA, Chalwadi SV (2024) Identification of Most Preferable Machine Learning Technique for Prediction of Bank Loan Defaulters. Indian Journal of Science and Technology 17(4): 343-351. https://doi.org/10.17485/IJST/v17i4.2978

* **Corresponding author**.

dbuphade@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: In the current financial landscape, banks confront with the significant challenges in effectively managing credit risk and ensuring the stability of their loan portfolios. It is imperative for the banks to ensure an accurate assessment of loan default possibility as a critical aspect of their overall risk management process. The study aims to develop a predictive model that is suitable for accurately identifying potential defaulters. **Methods:** Investigation employs a diverse range of machine learning techniques, including Random Forest, Logistic Regression, Decision Tree, k-Nearest Neighbour, Support Vector Machine, XG Boost, Ada Boost, and Gradient Boosting Machines, to evaluate loan default probabilities in both balanced and imbalanced data environments. The study's methodology involved the application of these algorithms to datasets typically characterized by imbalance, a frequent occurrence in financial risk assessments. We addressed this challenge by implementing resampling techniques, thereby enhancing the representativeness and accuracy of findings. **Findings:** Findings of this study indicate that in imbalanced datasets, the Random Forest algorithm emerged as the most accurate, registering an impressive 0.91 accuracy score. Comparable efficacy was noted in Logistic Regression and SVM, each achieving 0.90 and 0.91 accuracy scores respectively. Remarkably, in balanced datasets, the Random Forest model demonstrated a perfect accuracy score of 1.00, surpassing other models. This model consistently excelled in precision, recall, and F1-score metrics across different data scenarios. **Novelty:** This study highlights the Random Forest classifier as an optimal tool for predicting loan defaults, marking a significant advancement over existing methodologies. The outcomes of this research provide crucial insights for financial institutions in enhancing their loan risk assessments, thus enabling more precise and informed decision-making in lending processes.

Keywords: Credit risk; Machine learning; Random forest; Loan defaulter; Classification

1 Introduction

The challenge of predicting loan defaults accurately remains a critical issue in financial risk management. Current methodologies often deal with the inherent imbalances in financial datasets, where default instances are less frequent but bear significant impact. This imbalance leads to a skew in predictive models, often favoring the majority class, and thereby affecting the reliability of risk assessments. Such limitations in traditional predictive models underscore a significant gap in the risk assessment domain, particularly in decision-making processes related to loan approvals, interest rates, and credit limits.

This study provides a thorough comparison of various models, intending to enhance the predictive tools available to financial institutions for assessing loan default risks. Improved predictive accuracy is crucial for minimizing financial losses and optimizing lending decisions. The findings of this research are aimed at guiding financial institutions toward more informed and efficient decision-making, thereby contributing to the overall stability and efficiency of the financial sector.

In the contemporary world, the banking industry produces considerable amount of data that contains crucial information. As a result, it becomes imperative to efficiently store, process, control, and analyze this data to obtain valuable insights that can enhance business profitability. As an important part of financial sector, banking industry plays a critical role in the economy, with customers serving as its primary asset [(1)]. The alarming rise of non-performing assets (NPAs) in the banking sector has brought down global economies. Therefore, it is vital to direct attention towards the challenges posed by NPA encountered by banks [(2)]. The primary objective of this study is to develop a model for identifying fraudulent customers in the loan sanctioning process, and also to minimize errors in classifying fraudulent and genuine customers. To achieve this, supervised machine-learning algorithms are applied for the purpose of classification.

A novel hybrid econometric-machine learning approach for estimating multi-period corporate default probabilities has proposed [(3)]. The comparative assessment of the effectiveness demonstrated by five distinguished classifiers the Naïve Bayesian model, Logistic regression, Random Forest, decision tree, and K-nearest neighbor classifiers, which are essential in the field of machine learning and are employed for credit scoring purposes. These classifiers are central to the field of machine learning and are utilized for credit scoring purposes. The findings of experiments indicate that random forest performs better in precision, recall, AUC, and accuracy than other methods [(4)].

To improve model performance, decision trees are useful for creating categorical variables [(5)]. The data science algorithms have advanced significantly in the domains of deep learning models, ensemble models, hybrid deep learning models, and hybrid machine learning models. The increasing use of hybrid models can be attributed to their higher predictive accuracy [(6)].

The undirected and directed volatility networks of global stock market based on simple pair-wise correlation and system-wide connectedness of national stock indices using a vector auto-regressive model. This study has revealed the significance of network indicators as additional tools in predicting global stock market trends and regional relative directions [(7)].

A cluster-based classification model comprised of two stages improved k-means clustering and a fitness-scaling chaotic genetic ant colony algorithm (FSCGACA) based classification model. The algorithm proposed by the researchers was implemented on a set of three benchmark datasets, consisting of the qualitative bankruptcy dataset, Weislaw dataset, and Polish dataset. Through the course of their study, it was discovered that the financial crisis prediction (FCP) model, which was presented by the researchers, proved to be superior to other classification models based on various measures.

Furthermore, it was observed that the FCP model was more suitable for datasets with diverse characteristics [8]. [9] Adopted the Kitchenham methodology to extract, synthesize, and report the results. The supervised algorithms were employed more than unsupervised approaches viz., clustering.

The study aims to classify loan defaulters by using different machine learning techniques. We compared the performance of different machine learning algorithms (Random forest, logistic regression, and decision tree, XG Boost, Ada Boost, K Nearest Neighbour, SVM, GNB) and identified the most accurate model for loan defaulter prediction. Thus, financial institutions can mitigate financial losses and also assess the risk involved in lending money to borrowers by accurately predicting individuals who are likely to default on their loans. This predictive capability enables banks to make more accurate decisions regarding loan approvals, interest rates, and credit limits. Therefore, financial institutions will be able to make more informed lending decisions.

2 Methodology

A distinctive feature of our study is its comprehensive and nuanced analysis of classifier performance in both imbalanced and balanced datasets, a dimension that is often ignored in existing literature. An extensive and detailed evaluation of a variety of classifiers including Decision Tree, Random Forest, Logistic Regression, XG Boost, Ada Boost, K Nearest Neighbor, SVM, and GNB, for both imbalanced and balanced datasets has been analyzed. This approach contrasts sharply with previous studies, which typically concentrate on a limited selection of models. For example, our analysis revealed the superior performance of the Random Forest model in dealing with imbalanced data. This particular insight offers a significant enhancement over previous research [10] [11], which predominantly focus on the efficacy of models such as LightGBM and XGBoost. Our study fills a critical gap in existing research by providing comprehensive perspective on the capabilities and limitations of various classifiers, especially relevant in the area of bank loan default prediction. This comprehensive approach not only provides valuable insights for model selection and tuning in practical applications but also opens up new paths for future research in the field.

The data for this study was obtained from the Kaggle platform, which is a renowned online platform that enables individuals' collaboration interested in the field of data science and machine learning (Loan Default Prediction Dataset). The study utilized the "Bank Loan Defaulter Prediction" dataset, which contains 67,464 rows and 35 columns, with 29 of them being features, including 23 numerical and 6 categorical ones. The prime focus of the study is on the Loan Status, which is a crucial variable for indicating whether a loan is classified as a defaulter (1) or a non-defaulter (0).

The dataset contains numerical and categorical features.

- **Description of Features**
  - Loan Amount: loan amount applied.
  - Funded Amount: loan amount funded.
  - Funded Amount Investor: loan amount approved by the investors.
  - Term: term of loan (in months).
  - Batch Enrolled: batch numbers to representatives.
  - Interest Rate: interest rate (%) on loan.
  - Grade: grade by the bank.
  - Sub Grade: sub-grade by the bank.
  - Debit to Income: ratio of representative's total monthly debt repayment divided by self-reported monthly income excluding mortgage.
  - Delinquency- two years: number of 30+ days delinquency in past 2 – years.
  - Inquires - six months: total number of inquiries in last 6 months.
  - Open Account: number of open credit line in representatives - credit line.
  - Public Record: number of derogatory public records.
  - Revolving Balance: total credit revolving balance.
  - Revolving Utilities: amount of credit a representative is using - relative to revolving-balance.
  - Total Accounts: total number of credit lines available in - representative's credit line.
  - Initial List Status: unique listing status of the loan - W(Waiting), F(Forwarded).
  - Total Received Interest: total interest received till date.
  - Total Received Late Fee: total late fee received till date.
  - Recoveries: post charge off gross recovery.
  - Collection Recovery Fee: post charge off collection fee.
  - Collection 12 months Medical: total collections in last 12 months - excluding medical collections.

- Application Type: indicates when the representative is an individual or joint.
- Last week Pay: indicates how long (in weeks) a representative has paid EMI after batch enrolled.
- Accounts Delinquent: number of accounts on which the representative is delinquent.
- Total Collection Amount: total collection amount ever owed.
- Total Current Balance: total current balance from all accounts.
- Total Revolving Credit Limit: total revolving credit limit.
- Loan Status: 1 = Defaulter, 0 = Non-Defaulter.

2.1 Pre-processing of data

In pre-processing cleaning data is an essential step in preparing it for machine learning tasks. The steps involved in cleaning data for machine learning are handling missing values; dealing with outliers, data transformation, handling imbalanced classes, feature selection etc.

The features "ID", "Accounts Delinquent", "Loan Title", "Batch Enrolled", "Sub Grade", "Payment Plan" we delete this features from our dataset because, these features are not effective or relevant for predicting loan status.

In this study, the data is imbalanced so, first of all we have applied classification technique on imbalanced data and find performance measure. Thereafter, resampling technique is used to convert the data into balanced data. Further, classification techniques were applied on it and then evaluated performance measure for it. Finally, based on the performance measures we have tried to identify the most preferred model for classification.

2.2 Resampling Techniques

Resampling technique is used to convert Imbalanced data to balance. It can be done through two ways: under-sampling and oversampling. In under-sampling, we reduce the size of the majority class by randomly removing samples from it but it may discard potentially useful information. We increased the representation of the minority class using oversampling with either synthetic or replicated samples. This is done through the techniques viz., random oversampling, SMOTE (Synthetic Minority Over-Sampling Technique).[12] Suggested that SMOTE is the best technique for balancing a dataset.

2.3 Classification Technique

2.3.1 Logistic regression

A logistic function is used to model the relationship between the input variables and the probability of belonging to a specific class[13]. Through methods like softmax regression it can be extended to handle multi-class classification problems through techniques[14].

Logistic Regression-Sigmoid function is $y = \frac{e^{b_0+b_1X}}{1+e^{b_0+b_1X}}$

$X = input\ value,\ y = predicted\ output,\ b_0 = bias\ or\ intercept\ term,\ b_1 = coefficient\ of\ input\ (x)$

2.3.2 Decision trees

It is a methodology that takes sequential decisions based on the unique attributes of input data. It divides the data based on attribute values and creates a tree-like structure with if-else conditions[15]. Decision trees possess the attribute of interpretability and able to process both categorical as well as numerical data. The most widely used decision tree algorithms are ID3, CART, and Random Forests.

2.3.3 Random Forest

The technique of Random Forest is a well-known method in ensemble learning that effectively combines multiple decision trees to produce predictions. Each tree is constructed by a random selection of features and average prediction is determined by taking the majority vote or average of predictions from all trees. As a result, we can improve model's robustness and generalization.

2.3.4 Naive Bayes

In order to predict the class of an instance, conditional probabilities are used instead of assuming that there are any relationships between the input features[16]. Naive Bayes classifiers work effectively, particularly with large data sets. In general, they are used for classifying text or filtering out spam.

2.3.5 Support Vector Machine (SVM)

One of the key benefits of utilizing a Support Vector Machine (SVM) is the ability to procure an optimal hyperplane that effectively divides differing classes through the maximization of the margin between them. It transforms the input data into a high-dimensional feature space and classifies instances based on their position relative to the hyperplane[17]. By using different kernel functions it can handle both linear and non-linear classification problems.

2.3.6 K-Nearest Neighbors (k-NN)

It is a non-parametric classification algorithm that assigns a class label to an instance based on the majority vote of its k- Nearest Neighbors in the training data[18]. The distance metric is used to determine the neighbors. It is simple and intuitive but can be sensitive to the choice of k and the distance metric.

2.3.7 Gradient Boosting

It is a machine learning technique that combines numerous weak learners to create an accurate predictive model. The subsequent models have been trained to correct any inaccuracies made by their previous counterparts. The most popular gradient boosting algorithms are AdaBoost, Gradient Boosting Machines (GBM), XGBoost, and Light GBM[19].

A flow diagram (Figure 1) can help in understanding the sequential steps involved in selecting the most effective algorithm for predicting bank loan defaults for a given dataset. This is attributed to its exploratory nature and comprehensible presentation of the procedure. Figure 1 outlines a process for examining an imbalanced dataset via machine learning methodologies. This process requires pre-processing the dataset, constructing visual representations, and executing diverse machine learning techniques for classification. To address the issue of imbalanced data, resampling techniques are employed to convert the imbalanced data into a balanced dataset. The machine learning techniques are then employed on the balanced dataset, and their performance is evaluated. The objective is to identify the most suitable model based on its performance on the balanced dataset.

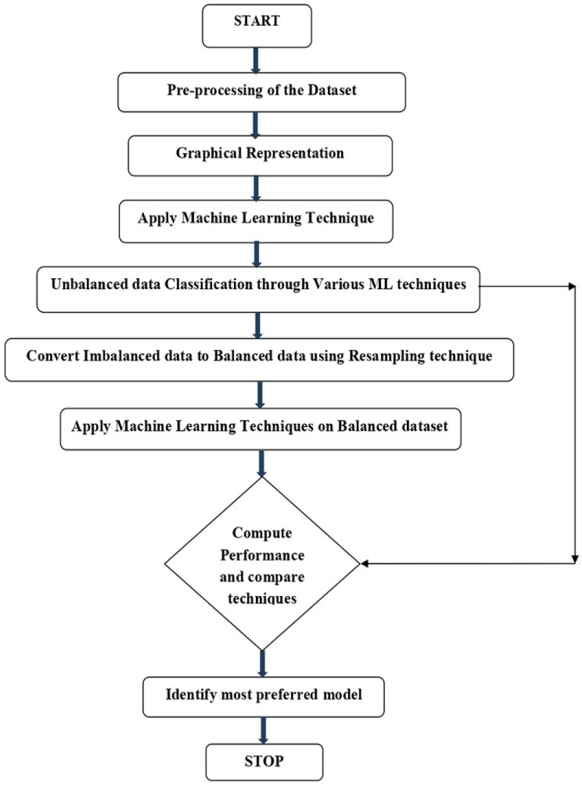

Fig 1. Flow Chart

3 Results and Discussion

3.1 Imbalanced Data

In machine learning, imbalanced data refers to a situation where the number of instances belonging to one class is significantly different from the number of instances belonging to another class [20]. In other words the distribution of classes or categories in the dataset is highly uneven, with one class being much more important than the other.

Table 1. Classification Result for Imbalanced data

Classifier	Accuracy Score	Confusion Matrix	Metrics Classification Report				
			Score	Precision	Recall	F1-Score	Support
Decision Tree	0.81	[16251 2148 1625 215]	0	0.91	0.88	0.90	18399
			1	0.09	0.12	0.10	1840
Random Forest	0.91	[18399 00 1840 00]	0	0.91	1.00	0.95	18399
			1	0.00	0.00	0.00	1840
Logistic Regression	0.90	[18399 00 1840 00]	0	0.91	1.00	0.95	18399
			1	0.00	0.00	0.00	1840
XG Boost	0.91	[18372 27 1835 05]	0	0.91	1.00	0.95	18399
			1	0.16	0.00	0.01	1840
Ada Boost	0.91	[18399 00 1839 01]	0	0.91	1.00	0.95	18399
			1	1.00	0.00	0.00	1840
K Nearest Neighbour	0.90	[18283 116 1831 09]	0	0.91	0.99	0.95	18399
			1	0.07	0.00	0.01	1840
SVM	0.91	[18399 00 1840 00]	0	0.91	1.00	0.95	18399
			1	0.00	0.00	0.00	1840
GNB	0.91	[18399 00 1840 00]	0	0.91	1.00	0.95	18399
			1	0.11	0.01	0.01	1840

The results in Table 1 reveal that the accuracy of all classification techniques is mostly satisfactory. However, it is observed that all models are inclines to favor the majority class, which is more predominant in the data. A significant proportion of algorithms have a tendency to prioritize the majority class, which has a negative impact on the performance of minority class [21] [22]. The minority class has significantly lower precision, recall, and F1 scores. To achieve fair and reliable findings, it is essential to resolve data imbalance through balancing approaches before applying classification methods.

3.2 Resampling techniques to Balanced Data

Table 2. Classification Result for balanced data

Classifier	Accuracy Score	Confusion Matrix	Metrics Classification Report				
			Score	Precision	Recall	F1-Score	Support
Decision Tree	0.93	[15934 2399 04 18397]	0	1.00	0.87	0.93	18333
			1	0.88	1.00	0.94	18401
Random Forest	1.00	[18333 00 04 18397]	0	1.00	1.00	1.00	18333
			1	1.00	1.00	1.00	18401
Logistic Regression	0.51	[9776 8557 9535 8866]	0	0.51	0.53	0.52	18333
			1	0.51	0.48	0.49	18401
XG Boost	0.80	[14008 4325 2893 15508]	0	0.83	0.76	0.80	18333
			1	0.78	0.84	0.81	18401
Ada Boost	0.55	[10276 8057 8487 9914]	0	0.55	0.56	0.55	18333
			1	0.55	0.54	0.55	18401
K Nearest Neighbour	0.83	[12385 5948 360 18041]	0	0.97	0.68	0.80	18333
			1	0.75	0.98	0.85	18401
SVM	0.51	[11230 7103 10975 7426]	0	0.51	0.61	0.55	18333

Continued on next page

Table 2 continued

			1	0.51	0.40	0.45	18401
GNB	0.51	[15751 2582]	0	0.50	0.86	0.64	18333
		[15492 2909]	1	0.53	0.16	0.24	18401

The oversampling methods effectively address the challenges posed by imbalanced datasets, leading to enhanced classifier performance in such scenarios[23,24]. We have applied the oversampling method to convert the imbalanced data in to balanced data and the results are presented in Table 2. The results showed an increase in accuracy for random forest and decision tree. However, the accuracy scores for other classification methods have reduced. We have observed an improvement in precision, recall, and F1 scores specifically for instances belonging to score 1.

Our study focused on evaluating the performance of various classifiers on imbalanced and balanced data sets, a critical aspect often encountered in machine learning applications, especially within the financial sector. The classifiers tested included Decision Tree, Random Forest, Logistic Regression, XG Boost, Ada Boost, K Nearest Neighbor, SVM, and GNB. The evaluation metrics employed were accuracy scores, confusion matrices, precision, recall, F1-score, and support.

For imbalanced data, our findings revealed varying degrees of classifier effectiveness. The accuracy scores ranged from 0.81 to 0.91 across different models. Notably, the Random Forest classifier exhibited a high accuracy score of 0.91, with a confusion matrix, precision for class 0 at 0.91, recall at 1.00, and F1-score at 0.95. These results are in alignment with the findings of[11] who noted the strong performance of traditional models like Random Forest and Decision Tree in handling imbalanced datasets.

In the scenario of balanced data, the accuracy scores were equally noteworthy. The Decision Tree model, for instance, showed an enhanced accuracy of 0.93, a significant improvement over its performance on imbalanced data. This aligns with the work of Zhu et al., who highlighted the superior predictive abilities of complex models like LightGBM and XGBoost over traditional logistic regression and decision tree models in balanced data scenarios.

Our results offer a comprehensive comparative analysis across a spectrum of models. This is particularly relevant when considering the findings of[11] which emphasized the importance of employing techniques like resampling and cost-sensitive learning to address the challenges posed by imbalanced datasets. Our study corroborates these findings by demonstrating the varied effectiveness of classifiers under different data conditions.

4 Conclusion

In this research, two distinct perspectives were taken into consideration for the purpose of classifying loan defaulter data set. Initially, the available dataset was dealt with for classification using different machine learning techniques. The classification results obtained through performance measures have given poor results. As a result, a second approach was employed that deals with imbalanced dataset. In order to achieve data balance, the oversampling technique was utilized followed by the implementation of machine learning methodologies. The findings of this study indicate that the second approach produces more precise outcomes as compared to the first approach. The findings of this study, as well as those of[22] suggest that oversampling techniques produce highly reliable results, regardless of the dataset characteristics.

In view of the above, it is observed that the random forest technique produces highly accurate results with a precision, recall, and F1-score of 100%. We can conclude that it is the most effective method for classifying loan defaulter datasets using machine learning. Further research could be directed towards investigating the effectiveness of oversampling methods in machine learning and exploring alternative techniques to examine their efficacy in handling imbalanced datasets. Therefore, our study can establish a foundation for future research in this area.

The empirical evidence indicates that the random forest algorithm is an exceptionally reliable classifier in the context of predicting loan defaulters. The decision tree represents a feasible alternative to the algorithm discussed above, as evident from the data collected.

Financial institutions or banks can effectively regulate their loan portfolios and minimize the financial losses caused by loan defaults by embracing these techniques. One limitation is that the present study explores usefulness of oversampling technique; future research could be beneficial to explore other methods. Furthermore, it would be useful to study various datasets from different financial organizations to examine the relevance of the findings.

This study presents new insights into the performance of various classifiers in handling imbalanced and balanced datasets, specifically in the context of bank loan default prediction. The uniqueness of our research lies in the comprehensive evaluation of a wide range of classifiers, including Decision Tree, Random Forest, Logistic Regression, XG Boost, Ada Boost, K Nearest Neighbor, SVM, and GNB, across different data scenarios.

Unlike previous studies that primarily focused on a limited set of models, our research extended the analysis to a broader range of classifiers. For instance, we observed that the Random Forest classifier achieved an accuracy of 0.91 on imbalanced

data, outperforming its counterparts significantly in this scenario. This finding enriches the existing literature, which has often highlighted the effectiveness of models like LightGBM and XGBoost.

This study also explores into the specifics of classifier performance on imbalanced datasets, a critical aspect often overlooked. We provided detailed metrics such as precision, recall, and F1-scores, revealing, for example, that the Decision Tree model showed a marked improvement in balanced data with an accuracy of 0.93, enhancing the understanding of classifier behavior in different data conditions.

The quantitative comparison of various models under both imbalanced and balanced conditions is a significant advancement. The accuracy improvement from 0.81 to 0.91 across different classifiers provides a new perspective on model selection and optimization in loan default prediction, a domain where data imbalance is a prevalent challenge.

Our findings have significant implications for practitioners in the financial sector, offering them a comprehensive guide for selecting and tuning models based on the specific nature of their datasets. Future research could build on our work by exploring the integration of these classifiers into ensemble methods, potentially leading to even more robust predictions in loan default scenarios.

In a nutshel, this study contributes new insights into the field of bank loan default prediction by providing a comprehensive analysis of various classifiers under different data scenarios. The detailed metrics and comparative analysis presented in our study not only complement but also extend the scope of existing literature, offering new perspectives and tools for researchers and practitioners in this domain.

5 Acknowledgement

DBU acknowledges Dr. D.D. Pawar, Director, School of Mathematical Science, SRTM University, Nanded for consistent encouragement in this research work.

References

1) Obiora SC, Zeng Y, Li Q, Liu H, Adjei PD, Csordas T. The effect of economic growth on banking system performance: An interregional and comparative study of Sub-Saharan Africa and developed economies. *Economic Systems*. 2022;46(1):100939. Available from: https://doi.org/10.1016/j.ecosys.2022.100939.
2) Kaur B, Kaur R, Sood K, Grıma S. Impact of Non-Performing Assets on the Profitability of the Indian Banking Sector. *Contemporary Studies of Risks in Emerging Technology, Part A*. 2023;p. 257–269. Available from: https://doi.org/10.1108/978-1-80455-562-020231017.
3) Sigrist F, Leuenberger N. Machine learning for corporate default risk: Multi-period prediction, frailty correlation, loan portfolios, and tail probabilities. *European Journal of Operational Research*. 2023;305(3):1390–1406. Available from: https://doi.org/10.1016/j.ejor.2022.06.035.
4) Wang Y, Zhang Y, Lu Y, Yu X. A Comparative Assessment of Credit Risk Model Based on Machine Learning - a case study of bank loan data. *Procedia Computer Science*. 2020;174:141–149. Available from: https://doi.org/10.1016/j.procs.2020.06.069.
5) Kristóf T, Virág M. EU-27 bank failure prediction with C5.0 decision trees and deep learning neural networks. *Research in International Business and Finance*. 2022;61:1–17. Available from: https://doi.org/10.1016/j.ribaf.2022.101644.
6) Nosratabadi S, Mosavi A, Duan P, Ghamisi P, Filip F, Band SS, et al. Data Science in Economics: Comprehensive Review of Advanced Machine Learning and Deep Learning Methods. *Mathematics*. 2020;8(10):1–25. Available from: https://doi.org/10.3390/math8101799.
7) Lee TK, Cho JH, Kwon DS, Sohn SY. Global stock market investment strategies based on financial network indicators using machine learning techniques. *Expert Systems with Applications*. 2019;117:228–242. Available from: https://doi.org/10.1016/j.eswa.2018.09.005.
8) Uthayakumar J, Metawa N, Shankar K, Lakshmanaprabu SK. Intelligent hybrid model for financial crisis prediction using machine learning techniques. *Information Systems and e-Business Management*. 2020;18:617–645. Available from: https://doi.org/10.1007/s10257-018-0388-9.
9) Ashtiani MN, Raahemi B. Intelligent Fraud Detection in Financial Statements Using Machine Learning and Data Mining: A Systematic Literature Review. *IEEE Access*. 2021;10:72504–72525. Available from: https://doi.org/10.1109/ACCESS.2021.3096799.
10) Zhu X, Chu Q, Song X, Hu P, Peng L. Explainable prediction of loan default based on machine learning models. *Data Science and Management*. 2023;6(3):123–133. Available from: https://doi.org/10.1016/j.dsm.2023.04.003.
11) Uddin N, Ahamed MKU, Uddin MA, Islam MM, Talukder MA, Aryal S. An ensemble machine learning based bank loan approval predictions system with a smart application. *International Journal of Cognitive Computing in Engineering*. 2023;4:327–339. Available from: https://doi.org/10.1016/j.ijcce.2023.09.001.
12) Chittora P, Chaurasia S, Chakrabarti P, Kumawat G, Chakrabarti T, Leonowicz Z, et al. Prediction of Chronic Kidney Disease - A Machine Learning Perspective. *IEEE Access*. 2021;9:17312–17334. Available from: https://doi.org/10.1109/ACCESS.2021.3053763.
13) Bisong E, &bisong E. Building Machine Learning and Deep Learning Models on Google Cloud Platform. A Comprehensive Guide for Beginners. 1st ed. CA, USA. Apress Berkeley. 2019. Available from: https://doi.org/10.1007/978-1-4842-4470-8.
14) Kavitha M, Yudistira N, Kurita T. Multi instance learning via deep CNN for multi-class recognition of Alzheimer's disease. In: 2019 IEEE 11th International Workshop on Computational Intelligence and Applications (IWCIA). IEEE. 2020;p. 89–94. Available from: https://doi.org/10.1109/IWCIA47330.2019.8955006.
15) Bhavani TT, Rao MK, Reddy AM. Network Intrusion Detection System Using Random Forest and Decision Tree Machine Learning Techniques. In: First International Conference on Sustainable Technologies for Computational Intelligence;vol. 1045 of Advances in Intelligent Systems and Computing. Springer, Singapore. 2020;p. 637–643. Available from: https://doi.org/10.1007/978-981-15-0029-9_50.
16) Saritas MM, Yasar A. Performance analysis of ANN and Naive Bayes classification algorithm for data classification. *International journal of intelligent systems and applications in engineering*. 2019;7(2):88–91. Available from: https://doi.org/10.18201//ijisae.2019252786.

17) Zhou Y, Uddin MS, Habib T, Chi G, Yuan K. Feature selection in credit risk modeling: an international evidence. *Economic Research-Ekonomska Istraživanja*. 2021;34(1):3064–3091. Available from: https://doi.org/10.1080/1331677X.2020.1867213.
18) Moorthy RS, Pabitha P. Optimal Detection of Phising Attack using SCA based K-NN. *Procedia Computer Science*. 2020;171:1716–1725. Available from: https://doi.org/10.1016/j.procs.2020.04.184.
19) Sharma P, Bora BJ. A Review of Modern Machine Learning Techniques in the Prediction of Remaining Useful Life of Lithium-Ion Batteries. *Batteries*. 2023;9(1):1–17. Available from: https://doi.org/10.3390/batteries9010013.
20) Park J, Kwon S, Jeong SP. A study on improving turnover intention forecasting by solving imbalanced data problems: focusing on SMOTE and generative adversarial networks. *Journal of Big Data*. 2023;10(1):1–16. Available from: https://doi.org/10.1186/s40537-023-00715-6.
21) Viloria A, Lezama OBP, Mercado-Caruzo N. Unbalanced data processing using oversampling: Machine Learning. *Procedia Computer Science*. 2020;175:108–113. Available from: https://doi.org/10.1016/j.procs.2020.07.018.
22) Wei Z, Zhang L, Zhao L. Minority-prediction-probability-based oversampling technique for imbalanced learning. *Information Sciences*. 2023;622:1273–1295. Available from: https://doi.org/10.1016/j.ins.2022.11.148.
23) Younas F, Usman M, Yan WQ. A deep ensemble learning method for colorectal polyp classification with optimized network parameters. *Applied Intelligence*. 2023;53(2):2410–2433. Available from: https://doi.org/10.1007/s10489-022-03689-9.
24) Kovács G. An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. *Applied Soft Computing*. 2019;83:105662. Available from: https://doi.org/10.1016/j.asoc.2019.105662.