Journal of Numerical Optimization and Technology Management

Vol. 1, No. 2, 2023

htttps://shmpublisher.com/index.php/jnotm

p-ISSN: xxxx-xxxx
e-ISSN: xxxx-xxxx

Home credit default risk assessment using embedded feature selection and stacking ensemble technique

Yosza Dasril[1], Yosy Arisandy[1], Shahrul Nizam Salahudin[1]
[1]Facuty of Technology Management and Business, Universiti Tun Hussein Onn Malaysia, Malaysia

Article Info

Article history:

Received September 2023
Revised September 2023
Accepted October 2023

Keywords:

Risk assessment
Stacking
Embedded technique
Feature selection
Home credit dataset

ABSTRACT

The objective of this study is to evaluate and compare the accuracy of typical credit assessment techniques, particularly Logistic Regression, Gradient Boosting, Random Forest (RF), Extra Gradient Boosting (XGB), Light Gradient Boosting Machine (LGBM), and Cat Boost. Furthermore, the study involves stacking ensemble learning with feature selection based on embedded techniques. This research utilized a data set sourced from Kaggle, namely the Home Credit Default Risk gathering data. The results of the study indicate that the reached accuracies were as follows: Logistic regression - 92.02%, XGB - 92.01%, LGBM - 92.09%, RF - 92.07%, and CB - 92.06%. Additionally, while stacking with XGB, RF, and LGBM models, and utilizing the final logistic regression estimator 92.01%, the accuracy does not show any improvement when compared to the usual algorithm. It is even lower than the LGBM accuracy results. However, the findings of this study demonstrate better rates of accuracy in comparison to other previous research conducted by researchers, regardless of that used the same dataset. However, study Mahmudi et al. in 2022 performs better than it in terms of accuracy using oversampling approaches. This finding provides evidence that the accuracy of the model is affected by the quantity of features that are examined. The level of accuracy will be better the more optimally chosen features are for examination.

Corresponding Author:

Yosy Arisandy,
Department of Production Operation Management,
Faculty of Technology Management and Business,
Universiti Tun Hussein Onn Malaysia,
Parit Raja, 86400, Batu Pahat, Johor, Malaysia.
Email: gp200002@student.uthm.edu.my
https://doi.org/10.00000/jnotm.0000.00.00.000

1. INTRODUCTION

Credit default risk is the possibility of a loss for a lender due to a borrower's failure to repay a loan [1]. Traditionally, credit analysts assess this risk by analyzing a borrower's capability to repay a loan [2].

However, with the advent of big data and machine learning, there has been a shift towards using these technologies to predict the probability of default [3] and assess credit risk [4].

Machine learning algorithms have enormous potential in the domain of credit risk assessment owing to their exceptional prediction capabilities and rapid processing capabilities [5]. Lenders can determine whether a borrower will default on a loan and estimate their probability of default by using machine learning [6], [7]. The dataset used for this purpose includes information on each borrower, as well as factors pertaining to each borrower [8]. Machine learning algorithms are employed to assess default risks through the analysis of various borrower characteristics [9], including income, age, gender, and other pertinent features.

Based on the findings of Lessmann et al. [10], it has been observed that the ensemble technique provides better performance compared to individual artificial intelligence and statistical methods. This study will additionally conduct a comparative analysis of multiple machine learning methods and evaluate the outcomes of experiments employing stacking ensemble methodologies on the Home Credit dataset.

2. METHOD

Data Collection and Pre-processing

Home credit obtained a dataset pertaining to loans granted in the South East Asian markets from a European lender, which was accessed through an internet repository. The credit data set used in this study is sourced from Kaggle.com, specifically the Home Credit Default Risk dataset.

There are 7 CSV files included in this competition. This study utilized the dataset application_train.csv for conducting experiments. Application_{train|test}.csv: Each row in this file is considered one loan, the file application_train.csv contains a target column, while application_test.csv does not contain a target column. The number of the clients in this file is 307511, and the number of the features is 124. The target variable defines whether the loan was repaid or not.

Feature Selection

The application of Personal Component Analysis (PCA) involves the reduction of data dimensions in the context of feature selection. The mapping of low-dimensional features does not yield any statistically meaningful impact. Principal Component Analysis (PCA) lacks the ability to discern the relative significance of features during the categorization procedure. The utilization of vital feature selection strategies is crucial, particularly when making decisions regarding loan disbursement or acquisition.

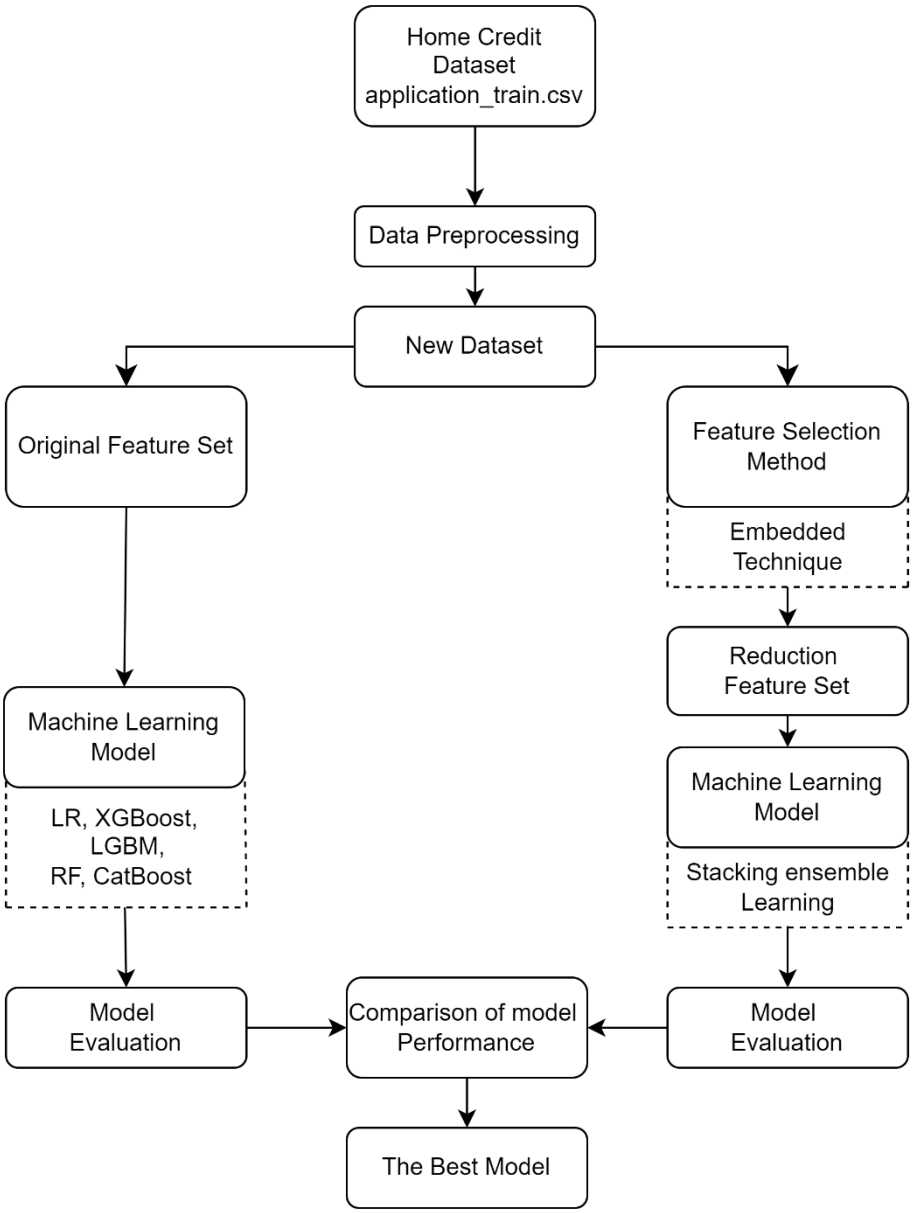

Figure 1. The procedural stages for implementing feature selection methods and a stacking ensemble learning model
Source: [7]

An alternate methodology involves the utilization of feature selection techniques. The embedded technique is considered as one of the approaches to feature selection. The process of feature selection in machine learning algorithm building is accomplished by the utilization of embedded approaches. Embedded methods are referred to as such because they engage in feature selection during the process of model training. The learning algorithm effectively utilizes its variable selection procedure to concurrently execute feature selection and classification or regression tasks. All embedded techniques operate by initially training machine learning models. The researchers subsequently derived the salient characteristics of this model, which quantifies the significance of these aspects in the context of prediction. Ultimately, the researchers eliminate extraneous characteristics by prioritizing essential child attributes.

Stacking Ensemble Model

The stacking method was introduced by Wolpert and David [11], [12] as an ensemble algorithm that is distinct from bagging, random forests (RF), and boosting, stacking employ heterogeneous learners. Another study [13] has found that the stacking technique in ensemble learning demonstrates distinct advantages, particularly when dealing with imbalanced data.

Stacking refers to the combination of multiple models of diverse natures, employing the concept of a meta-learner [14]. Stacking is an example of an ensemble technique. Ensemble learning is a methodology in machine learning that involves the utilization of multiple learners to address a common problem. Ensemble methods aim to create a collection of models and merge them, in contrast with typical machine learning approaches that aim to construct a model based on training data [15], [16].

During the training phase, the base classification algorithm(s) and training data are typically utilized as inputs. The process of model generation involves training the algorithm using a dataset and subsequently generating models. In the stacking implementation, it is common for the "model generation" phase to produce n models by employing a single algorithm and randomly selecting sub datasets. In cases where the implementation allows for the utilization of multiple algorithms, it is customary for the training set to undergo training with each of these algorithms. Once the models are generated, they produce the predicted labels.
The initial stage of the training phase consists of two layers. The input to the second layer consists of the predicted labels from each model. In the subsequent layer, a classification algorithm, commonly referred to as the combiner method, is employed to produce a conclusive model. Throughout this process, the initial labels are retained for the purpose of annotating the newly acquired training data. The prediction phase of the stacking method comprises two layers. The initial layer utilizes input data and previously generated models to generate a prediction, while the subsequent layer employs the model generated in the second layer of the training phase. The Stacking ensemble method comprises the incorporation of original (training) data, primary level models, primary level prediction, a secondary level model, and the ultimate prediction. The fundamental structure of stacking can be depicted in the manner illustrated below the accompanying visual representation.

Figure 2. The fundamental structure of stacking

a. Original data: The data is partitioned into n-folds and is referred to as either test data or training data.
b. Base models, also known as level-0 models, are the subject of discussion. These models utilize a set of training data to generate compiled predictions, which are then outputted as level-0 predictions.
c. Initial Predictions: Each base model is activated using a set of training data and generates distinct predictions, referred to as level-0 predictions.
d. The Meta Model is a linguistic tool used in the field of psychology and communication to identify and challenge the underlying assumptions and generalizations made in a person's language. It aims The architectural design of the stacking model comprises a single meta-model that facilitates the optimal fusion of predictions generated by the base models. The meta-model is alternatively referred to as the level-1 model.
e. Level-1 Prediction involves the meta-model acquiring the optimal method of combining predictions from the base models. This is achieved by training the meta-model using predictions generated by individual base models. Specifically, the meta-model is fed with data that was not used during the base models' training phase. Predictions are made using this data, and these predictions, along with the expected outputs, form the input and output pairs of the training dataset used to train the meta-model.

This study employed three distinct algorithms, which are XGBoost [17], Random Forest [18], and LGBM [19], as combiner methods, with logistic regression[20] supporting as the final estimator.

3. RESULTS AND DISCUSSIONS

.

This section discusses the first performance metrics, followed by the results of all experiments conducted and their discussion.

Proposed Classification Models

In order to evaluate the effectiveness of the suggested methodology, several performance metrics have been employed, including accuracy, precision, recall, and f1-score. The following definitions are provided:

a. Accuracy score

Accuracy refers to the proportion of correct predictions in relation to the total number of forecasts made. The significance of this parameter in obtaining precise outcomes is widely acknowledged. The accuracy score can be represented mathematically as depicted in equation (1):

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \times 100 \quad (1)$$

b. Precision

The statistic being referred to is the true positive rate, which is defined as the proportion of properly predicted true positive instances out of the total number of true positive instances. The accuracy metric can be mathematically represented as depicted in equation (2):

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

c. Recall

The certainty of positive predictions in each class is determined by the ratio of correctly predicted positive instances to the total number of observations in that class. A recall is commonly referred to as sensitivity in academic literature. The recall metric can be represented mathematically as demonstrated in equation (3):

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

d. F1-Score

The F1 Score can be defined as the arithmetic mean of Precision and Recall, with each metric being given equal weight. The F1-Score incorporates both false positives and false negatives to quantify the performance of a classification model. The F1-Score can be represented mathematically as demonstrated in equation (4):

$$F1 - Score = 2\frac{Recall \times Precision}{Recall+Precision} \quad (4)$$

A false positive (FP) refers to a positive outcome that exhibits a significant percentage of instances that are not accurate. In the equation, TP is the variable denoting True Positive, which signifies the quantity of cases that have been accurately detected. The quantity of accurately identified examples is shown as TN, while FN represents the quantity of examples that have been categorized wrongly.

However, deciding which metrics to use when assessing the model on imbalanced data can be challenging. Confusion matrix is a useful tool to calculate Recall, Precision, Specificity, Accuracy, G-Mean, F1, and Area Under the ROC

Curve, and it consists of a matrix of four different combinations (TN, FN, FP, TP) of predicted and actual values.

		Predicted values	
		Class (0)	Class (1)
Actual values	Class (0)	TN	FP
	Class (1)	FN	TP

Figure 3. The components of 2x2 confusion matrix

Table 1. Confusion matrix calculation

Algorithm	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.92018	0.39837	0.01003	0.019573
XGBoost	0.92015	0.46699	0.03911	0.07217
LGBM	**0.92095**	0.56111	0.02068	0.03989
Random Forest	0.92069	0.75	0.00184	0.00368
CatBoost	0.92059	0	0	0

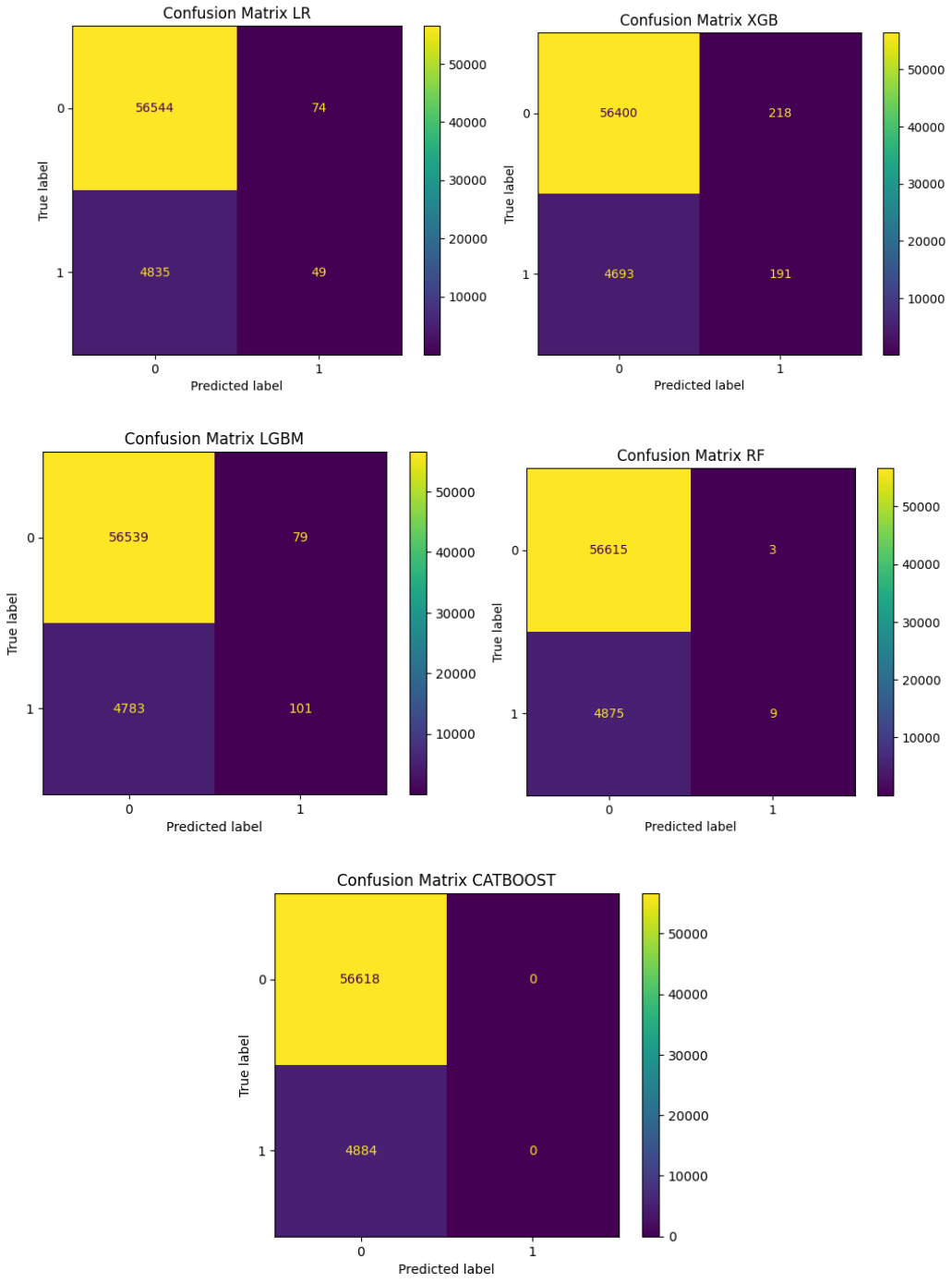

Figure 4. Confusion matrix of original algorithm

XGBoost

According to the data shown in Table 1. The LGBM algorithm demonstrates the highest level of accuracy, with a notable accuracy rate of 92.09%.

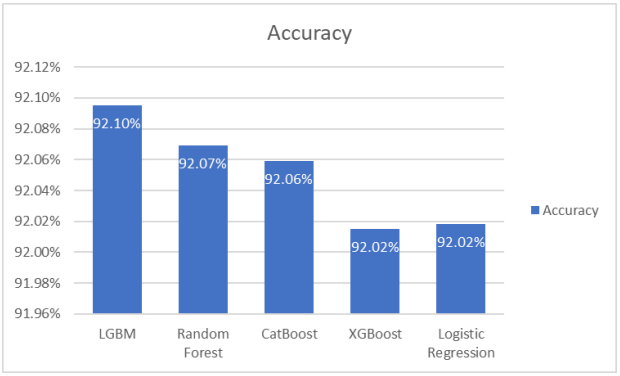

Figure 5. Persentage accuracy of each measurement algorithm

Based on the findings, this study will explore two stacking combinations: one using LGBM, RF, and CatBoost, and another including LGBM, RF, and XGBoost, with LR serving as the ultimate estimator for both combinations. Here are the results:

1. Stacking LGBM, RF, and XGBoost with final estimator LR
   Accuracy: **0.91999**
   Precision: 0.46783
   Recall: 0.055078
   F1 Score: 0.09855

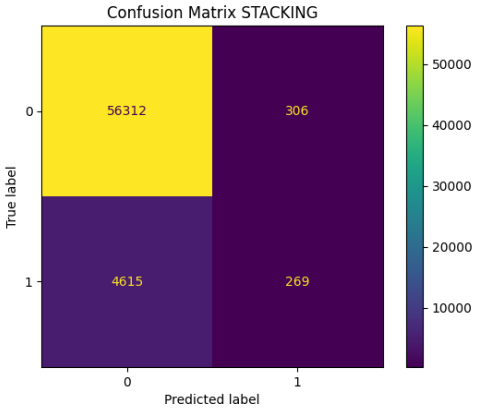

Figure 6. Confusion matrix stacking LGBM, RF, and XGBoost

2. Stacking LGBM, RF, and CatBoost with final estimator LR
   Accuracy: **0.92012**
   Precision: 0.47723
   Recall: 0.06224
   F1 Score: 0.11013

.

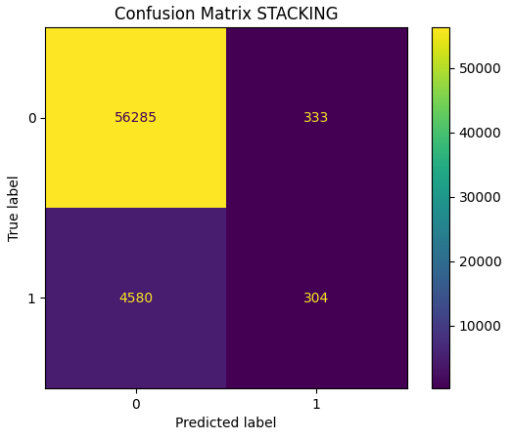

Figure 7. Confusion matrix stacking LGBM, RF, and CatBoost

Accuracy with a combination of stacking between LGBM, RF, and CatBoost is higher than stacking with a combination of LGBM, RF, and XGBoost. Although the difference is not very specific. However, it is important to note that the effectiveness of the stacking ensemble method depends on various factors such as the choice of base models, the quality of the data, and the hyperparameters of the models(Ture et al., 2023).

The stacking ensemble technique does not always improve accuracy when compared to the accuracy of the individual models. It could be due to the selection of the combination model, the base model. Because with a combination that produces different accuracy. The combination of CatBoost RF, LGBM 92.00, while the combination of XGB, RF and LGBM 92.01 both use the final estimator LR. Despite this, the utilization of LGBM for measurements yielded the most optimal accuracy compared to other measurement techniques, specifically achieving a value of 0.92095 that information is presented in Table 2. This finding aligns with the research conducted by K.S. Naik(Naik, 2021) and Rabia(Aziz et al., 2022), which indicates that LGBM demonstrates notable benefits in effectively handling bigger volumes of data and improving efficiency.

Table 2. Accuracy measurement results

Algorithm	Accuracy
Logistic Regression (LR)	0.92018
XGBoost (XGB)	0.92015
LGBM	**0.92095**
Random Forest (RF)	0.92069
CatBoost (CB)	0.92059
Stacking (XGB, RF, LGBM) final estimator LR	**0.92012**
Stacking (CB, RF, LGBM) final estimator LR	0.91999

The results of this study obtain higher accuracy than some previous studies using the same dataset, namely the home credit data set. However, it is no better than research Hafizullah (Mahmudi et al., 2022) which obtains an accuracy of 0.98472, using feature techniques and SMOTE and ADASYN oversampling techniques. This proves that feature engineering and oversampling techniques provide advantages to the model to improve unbalanced dataset performance, reduce dataset complexity such as the Home Credit dataset (Ghorbani & Ghousi, 2020; Yan et al., 2019).

Table 3. Research comparison

Research	Best Accuracy
[19]	DeepGBM 0.755832
[21]	LGBM 0.778
[22]	LGBM 0.79304
[23]	CatBoost 0.7792
This Research	Stacking **0.92012**, LGBM **0.92095**
[24]	XG Boost **0.98472**

4. CONCLUSION

The application of Light Gradient Boosting Machine (LGBM) for measurements resulted in the highest level of accuracy compared to alternative measurement approaches, notably obtaining a value of 0.92095. The findings suggest that LGBM has high efficiency and can significantly improve accuracy when applied to datasets with huge volumes. The accuracy achieved by combining LGBM, RF, and CatBoost in a stacking ensemble is higher compared to combining LGBM, RF, and XGBoost, with a value of 0.92012. The stacking ensemble technique may not consistently enhance accuracy in comparison to the accuracy achieved by the individual models. The potential cause for this outcome may be attributed to the use of the combination

model as the base model. In order to address the issue of imbalanced datasets, such as the Home Credit dataset, it is imperative to employ feature engineering and oversampling approaches.

REFERENCES

[1] Z. Li, K. Li, X. Yao, and Q. Wen, "Predicting Prepayment and Default Risks of Unsecured Consumer Loans in Online Lending," *Emerg. Mark. Financ. Trade*, vol. 55, no. 1, pp. 118–132, 2019, doi: 10.1080/1540496X.2018.1479251.

[2] W. Wu, D. Xu, Y. Zhao, and X. Liu, "Do consumer internet behaviours provide incremental information to predict credit default risk?," *Econ. Polit. Stud.*, vol. 8, no. 4, pp. 482–499, Oct. 2020, doi: 10.1080/20954816.2020.1759765.

[3] P. M. Addo, D. Guegan, and B. Hassani, "Credit risk analysis using machine and deep learning models," *Risks*, vol. 6, no. 2, pp. 1–20, 2018, doi: 10.3390/risks6020038.

[4] S. A. and M. Dowling, *Machine Learning and AI for Risk Management*. Digital Business & Enabling Technologies, 2019. doi: 10.1007/978-3-030-02330-0_3.

[5] S. B. S. Simão, "Machine Learning applied to credit risk assessment: Prediction of loan defaults." 2023.

[6] M. Moscatelli, F. Parlapiano, S. Narizzano, and G. Viggiano, "Corporate default forecasting with machine learning," *Expert Syst. Appl.*, vol. 161, p. 113567, 2020.

[7] M. Munsarif and M. Sam'ansafuan, "Peer to peer lending risk analysis based on embedded technique and stacking ensemble learning," *Bull. Electr. Eng. Informatics*, vol. 11, no. 6, pp. 3483–3489, 2022, doi: 10.11591/eei.v11i6.3927.

[8] Z. Wang, C. Jiang, H. Zhao, and Y. Ding, "Mining semantic soft factors for credit risk evaluation in peer-to-peer lending," *J. Manag. Inf. Syst.*, vol. 37, no. 1, pp. 282–308, 2020.

[9] L. Barbaglia, S. Manzan, and E. Tosetti, "Forecasting loan default in Europe with machine learning," *J. Financ. Econom.*, vol. 21, no. 2, pp. 569–596, 2023.

[10] S. Lessmann, B. Baesens, H. V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," *Eur. J. Oper. Res.*, vol. 247, no. 1, pp. 124–136, 2015, doi: 10.1016/j.ejor.2015.05.030.

[11] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.

[12] E. Martin, "Stacked generalization," *Encycl. Mach. Learn.*, pp. 912–912, 2011, doi: 10.1007/978-0-387- 30164-8_778.

[13] X. Yin, Q. Liu, Y. Pan, X. Huang, J. Wu, and X. Wang, "Strength of stacking technique of ensemble learning in rockburst prediction with imbalanced data: Comparison of eight single and ensemble models," *Nat. Resour. Res.*, vol. 30, pp. 1795–1815, 2021.

[14] S. Susan, A. Kumar, and A. Jain, "Evaluating heterogeneous ensembles with boosting meta-learner," in *Inventive Communication and Computational Technologies: Proceedings of ICICCT 2020*, Springer, 2021, pp. 699–710.

[15] L. Liu and M. T. Özsu, *Encyclopedia of database systems*, vol. 6. Springer New York, NY, USA:, 2009.

[16] N. Demir and G. Dalkiliç, "Modified stacking ensemble approach to detect network intrusion," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 26, no. 1, pp. 418–433, 2018, doi: 10.3906/elk-1702-279.

[17] L. Guntay, E. Bozan, U. Tigrak, T. Durdu, and G. E. Ozkahya, "An Explainable Credit Scoring Framework: A Use Case of Addressing Challenges in Applied Machine Learning," in *2022 IEEE Technology and Engineering Management Conference: Societal Challenges: Technology, Transitions and Resilience Virtual Conference, TEMSCON EUROPE 2022*, 2022, pp. 222–227. doi: 10.1109/TEMSCONEUROPE54743.2022.9802029.

[18] A. Safiya Parvin and B. Saleena, "An Ensemble Classifier Model to Predict Credit Scoring-Comparative Analysis," in *Proceedings - 2020 6th IEEE International Symposium on Smart Electronic Systems, iSES 2020*, 2020, pp. 27–30. doi: 10.1109/iSES50453.2020.00017.

[19] X. Chen, X. Liu, Z. Liu, P. Song, and M. Zhong, "A deep learning approach using DeepGBM for credit assessment," *ACM Int. Conf. Proceeding Ser.*, pp. 774–779, 2019, doi: 10.1145/3366194.3366333.

[20] G. Cheng, "Financial Evaluation Model and Algorithm Based on Data Mining," in *ACM International Conference Proceeding Series*, 2021, pp. 151–155. doi: 10.1145/3510858.3510914.

[21] Z. Qiu, Y. Li, P. Ni, and G. Li, "Credit risk scoring analysis based on machine learning models," *Proc. - 2019 6th Int. Conf. Inf. Sci. Control Eng. ICISCE 2019*, pp. 220–224, 2019, doi: 10.1109/ICISCE48695.2019.00052.

[22] E. Al Daoud, "Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit Dataset," vol. 13, no. 1, pp. 6–10, 2019.

[23] Y. Tounsi, H. Anoun, and L. Hassouni, "CSMAS: Improving multi-agent credit scoring system by integrating big data and the new generation of gradient boosting algorithms," in *Proceedings of the 3rd international conference on networking, information systems & security*, 2020, pp. 1–7.

[24] H. Mahmudi, R. Bhargava, and R. Das, "Evaluation of Gradient Boosting Algorithms on Balanced Home Credit Default Risk," *2022 Int. Conf. Trends Quantum Comput. Emerg. Bus. Technol. TQCEBT 2022*, pp. 1–6, 2022, doi: 10.1109/TQCEBT54229.2022.10041584.

.