Journal of Business and Management Studies
ISSN: 2709-0876
DOI: 10.32996/jbms
Journal Homepage: www.al-kindipublisher.com/index.php/jbms

JBMS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Credit Risk Prediction Using Explainable AI

Sarder Abdulla Al Shiam[1] ✉ Md Mahdi Hasan[2], Md Jubair Pantho[3], Sarmin Akter Shochona[4], Md Boktiar Nayeem[5], M Tazwar Hossain Choudhury[6] and Tuan Ngoc Nguyen[7]

[124]Department of Management, St Francis College, Brooklyn, NY, USA
[3]Department of Computer Science and Engineering, University of Tennessee at Chattanooga, Chattanooga, TN, USA
[56]Department of Graduate and Professional Studies, Trine University, Angola, IN, USA
[7]VNDirect Securities, 97 Lo Duc, Hai Ba Trung, Hanoi, Vietnam

Corresponding Author: Sarder Abdulla Al Shiam, **E-mail**: sshiam@sfc.edu

| ABSTRACT

Despite advancements in machine-learning prediction techniques, the majority of lenders continue to rely on conventional methods for predicting credit defaults, largely due to their lack of transparency and explainability. This reluctance to embrace newer approaches persists as there is a compelling need for credit default prediction models to be explainable. This study introduces credit default prediction models employing several tree-based ensemble methods, with the most effective model, XGBoost, being further utilized to enhance explainability. We implement SHapley Additive exPlanations (SHAP) in ML-based credit scoring models using data from the US-based P2P Lending Platform, Lending Club. Detailed discussions on the results, along with explanations using SHAP values, are also provided. The model explainability generated by Shapely values enables its applicability to a broad spectrum of industry applications.

| KEYWORDS

Credit Risk, Machine Learning, Shapely Value, Explainable AI

| ARTICLE INFORMATION

ACCEPTED: 01 March 2024 **PUBLISHED:** 18 March 2024 **DOI:** 10.32996/jbms.2024.6.2.6

1. Introduction

Credit risk management is one of the prime challenges that banks and lenders face as a large number of borrowers fail to meet their loan repayment obligations. This risk poses a threat to a lender's income stream, as a surge in credit defaults could lead to expenses that may render the lender insolvent. Since bank failure can result in widespread economic adversity, as seen during the global financial crisis, they are legally obligated to develop credit default prediction models that guide capital requirements to absorb potential losses. While the need for effective model performance is apparent, the primary focus lies on interpretability and explainability, given the heavily regulated nature of these models.

Deploying credit default prediction models demands transparency and interpretability, ensuring that lay users can understand the reasons behind the predictions. These models not only determine regulatory capital requirements but also influence whether potential borrowers are granted credit. Legal jurisdictions have long recognized the right to explanation, emphasizing the necessity for transparent and accessible model explanations. Logistic Regression has traditionally been favored for its interpretability, even though more advanced machine learning models exist. The reluctance to adopt these models stems from their opaque nature, termed "Black-Box." This research aims to explore whether credit default prediction can be enhanced through explainable AI while maintaining adequate model interpretability. Utilizing an openly available dataset, the study constructs a credit default prediction model using high-performance black-box methods. The optimal model is then transparently explained using Shapely values to meet a predefined standard of interpretability.

The document details the process undertaken to build an explainable credit default model, starting with a literature review in section 2, which examines previous attempts to enhance credit default prediction. Section 3 introduces the methodology, whereas Section 4 outlines the key findings of the research. Concluding section 5 determines whether the research objectives are met and outlines implications for future work on the topic.

2. Literature Review

The subsequent exploration of existing literature delves into the ongoing efforts to advance credit default prediction and underscores the critical importance of ensuring these models are explainable. The review initiates by scrutinizing research that demonstrates pathways to enhance credit default prediction, traversing through historical perspectives to underscore the compelling need for model explainability. Subsequently, it navigates through alternative methodologies that have been employed in an effort to render credit default models more interpretable.

In light of banking portfolios often reaching astronomical values in the billions, even marginal improvements are deemed consequential. This acknowledgment propels the pursuit of enhancing model performance in credit default prediction, a pursuit that carries substantial merit. Past studies have showcased successful implementations of credit default prediction through the utilization of less interpretable machine learning methods, particularly in the domain of corporate bankruptcy prediction. Exemplars such as Moscatelli et al. (2020), Barboza et al. (2017), and Guegan and Hassani (2018) highlight the superiority of tree-based methods like Random Forest and Gradient-boosted trees over Logistic Regression. Additionally, Fitzpatrick and Mues (2016) provide evidence of the effectiveness of tree-based methods in predicting mortgage defaults. While machine learning methods promise performance enhancements, it is crucial to recognize that credit default prediction models must not only exhibit robust performance but also embody transparency and explainability, given their sensitive role in automating decisions on loan applications.

The legal landscape, exemplified by acts such as the Fair and Accurate Credit Transactions Act and the US Fair Credit Reporting Act in the United States, along with the General Data Protection Regulation in the European Union, emphasizes the right to an explanation for loan application rejections. Ethical concerns surrounding potential discriminatory biases underscore the imperative for model explainability, as demonstrated by Munnell et al. (1996) and Charles and Hurst (2002), who shed light on racial disparities in lending success. Addressing this need, the European Banking Authority (EBA) establishes a clear minimum standard for model explainability, demanding both understandability by humans and justifications for the primary factors influencing the model's output.

Despite the progress made, previous research has predominantly focused on model performance, with scant attention given to explainability, rendering such models unsuitable for practical industrial deployment. While some studies, such as Fitzpatrick and Mues (2016), acknowledge the interpretability of Logistic Regression, feature importance measures derived by Moscatelli et al. (2020) and Fitzpatrick and Mues (2016) fall short in substantiating model predictions. Chen et al. (2021) argue that machine learning models primarily aim for prediction, leaving explanation to statistical models like Logistic Regression. Nevertheless, the study advocates for harnessing the performance benefits of machine learning methods while upholding a pre-defined standard of explainability.

Recent endeavors to elucidate "black-box" credit default prediction models are gaining momentum, with the SHapley Additive exPlanations (SHAP) method proposed by Lundberg and Lee (2017) standing out. This model-agnostic approach offers feature importance measures. Bussmann et al. (2021) employ SHAP values to explain an XGBoost model's predictions, revealing superior performance compared to Logistic Regression. However, the study falls short of meeting the EBA's standard, lacking sufficient justification for the model's predictions. Alternative methods, such as rule extraction exemplified by Prentzas et al. (2019) and counterfactual approaches endorsed by Keane and Smyth (2020) and Fernandez et al. (2020), strive to justify predictions from black-box models. These approaches aim to satisfy GDPR's right to explanation and claim to provide clearer insights than feature importance methods, according to Fernandez et al. (2020).

3. Methodology

3.1 Data

In this section, we delve into the specifics of the credit risk dataset utilized in our research and provide a detailed examination of the black box machine learning/deep learning classifiers designed to differentiate between various risk classes. Our experimentation leverages the Lending Club dataset, sourced from Kaggle, which encompasses records from over 2.2 million peer-to-peer loans facilitated through the Lending Club platform. Each loan within this dataset has a uniform 3-year term, and crucially, the outcome of each loan is discernible, allowing us to determine whether it was fully paid or charged off as a loss. The primary aim of our analysis is straightforward: to develop a classifier capable of accurately discerning between two distinct classes, namely default and non-default. To achieve this, we deploy four machine learning-based classifiers, including Decision Tree, Light GBM, Random Forests, and XGBoost.

The original Lending Club dataset comprises a comprehensive array of 145 features, encompassing diverse aspects ranging from consumer demographic details to indicators capturing consumer payment behavior. A glimpse into some of these features is provided in Table 1. Our objective, based on the available dataset, is to identify the optimal set of features that yield the highest F1 and Receiver Operating Characteristic (ROC) scores on the test dataset. However, it's important to note that the dataset from Kaggle presents a significant challenge due to its highly imbalanced nature. We resolve this issue by adjusting the weights of the two types of observations.

3.2 Machine Learning Algorithms

In the exploration of machine learning algorithms, our initial step involves empirically validating the association between current and future interest rates, along with other economic indicators, and stock prices. To substantiate this, we conduct a linear regression of all the features against stock prices, reporting relevant coefficients. Once the relevance of our proposed variables is established in explaining stock prices, we proceed to train several machine learning algorithms, both with and without our proposed variables. The models include the Decision Tree, Light Gradient Boosting Model (LGBM), Extreme Gradient Boosted Model (XGBM), and Random Forests model. We systematically compare the performances of each machine learning model trained on datasets with and without the inclusion of our proposed variables.

3.2.1 Decision Tree Model

The Decision Tree Model operates on a tree-based algorithm, where the predictor space is divided into distinct and non-overlapping regions. These regions are determined by selecting the predictor and cut point to minimize the Residual Sum of Squares (RSS). For each observation falling within a specific region, the response value is set as the mean of all observed response values within that region. This model provides a clear structure, representing predictors as splitting rules within a tree.

3.2.2 Light Gradient Boosting Machine

The Light Gradient Boosting Machine (LGBM) algorithm, a variant of the Gradient Boosting Decision Tree, differs by expanding vertically (leaf-wise) instead of horizontally. It approximates loss functions using second-order Taylor approximation and trains a decision tree to minimize this approximation. LGBM introduces efficiency-improving techniques, such as Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). GOSS allows excluding data instances with small gradients, estimating information gain using the remaining instances. EFB bundles mutually exclusive features, reducing their number without compromising the accuracy of split points. LGBM proves to be an effective choice for feature reduction without sacrificing model accuracy.

3.2.3 Extreme Gradient Boosting

Moving on to the Extreme Gradient Boosting Model (XGB), it is an implementation of a gradient-boosting decision tree algorithm designed to predict a target variable accurately. This algorithm combines the estimates of simpler models to prevent overfitting, introducing LASSO (L1) and Ridge (L2) regularization to penalize more complex models. The objective function comprises the deviation of the model and a regularization term, determining prediction accuracy by considering deviation and variance. The training process iteratively adds new trees, predicting residuals of prior trees and combining them for the final prediction.

3.2.4 Random Forests

Random Forests operate by constructing multiple decision trees during training and outputting the mode (classification) or mean (regression) prediction of the individual trees. Each decision tree in the forest is built based on a randomly selected subset of the training data and a randomly selected subset of features. This randomness helps to reduce overfitting and improves the generalization ability of the model. During the training process, each tree in the forest is trained independently, using a technique called bagging (bootstrap aggregating), which involves sampling the training data with replacement. This creates diverse trees that collectively form a robust model capable of capturing complex patterns in the data. To make predictions, new data points are passed through each individual tree in the forest, and the predictions from all trees are aggregated to produce the final prediction. In classification tasks, the mode (most frequently occurring class) of the individual tree predictions is taken as the final prediction, while in regression tasks, the mean of the individual tree predictions is computed.

4. Results and Discussion

In Table 1, we conduct a comprehensive evaluation of model performance using key metrics, including Accuracy, Precision, Recall, and Area Under the ROC Curve (AUC). The focus of our analysis revolves around four tree-based machine-learning models: Decision Tree, Light GBM, Random Forests, and XGBoost. The comparison of results across these models is essential for discerning their effectiveness in predicting credit default.

Beginning with the Decision Tree model, its performance metrics showcase an Accuracy of 71.21%, Precision of 17.23%, Recall of 71.23%, and an AUC value of 74.37%. Moving on to LightGBM, this model demonstrates an Accuracy of 72.45%, Precision of 18.87%, Recall of 71.94%, and an AUC value of 72.05%. Similarly, Random Forests exhibit an Accuracy of 70.93%, a Precision of

19.78%, a Recall of 70.24%, and an AUC value of 70.11%. Concluding with, the XGBoost model, it surpasses the others with an Accuracy of 74.55%, Precision of 22.03%, Recall of 75.27%, and an impressive AUC value of 81.29%.

Upon an overarching comparison of these models, it becomes evident that while their performances are relatively similar, XGBoost emerges as the most proficient for this specific dataset. Notably, a consistent trend is observed across all models, where accuracy, recall, and AUC metrics surpass precision. This observation aligns with expectations, considering the highly imbalanced nature of the credit default data. The models face the challenge of learning from a limited number of records with default cases, influencing the precision values across the board.

Table 1: Results of Model Performances

	Accuracy	**Precision**	**Recall**	**AUC**
Decision Tree	71.21%	17.49%	71.23%	74.37%
LightGBM	72.45%	18.87%	71.94%	72.05%
Random Forest	70.93%	19.78%	70.24%	70.11%
XGBoost	74.55%	22.03%	75.27%	81.29%

While results presented in Table 1 highlight prediction performances, they don't provide any information regarding how the predictions are made. We use Shapely values to interpret the model and examine the contribution of each feature that contributes to these results. Shapely values utilize the Collaborative Game theory approach to provide desirable properties and are widely used in the literature for explaining computational intelligence models. Figure 1 highlights the important features identified by Shapely values that have the strongest contribution to model predictions.

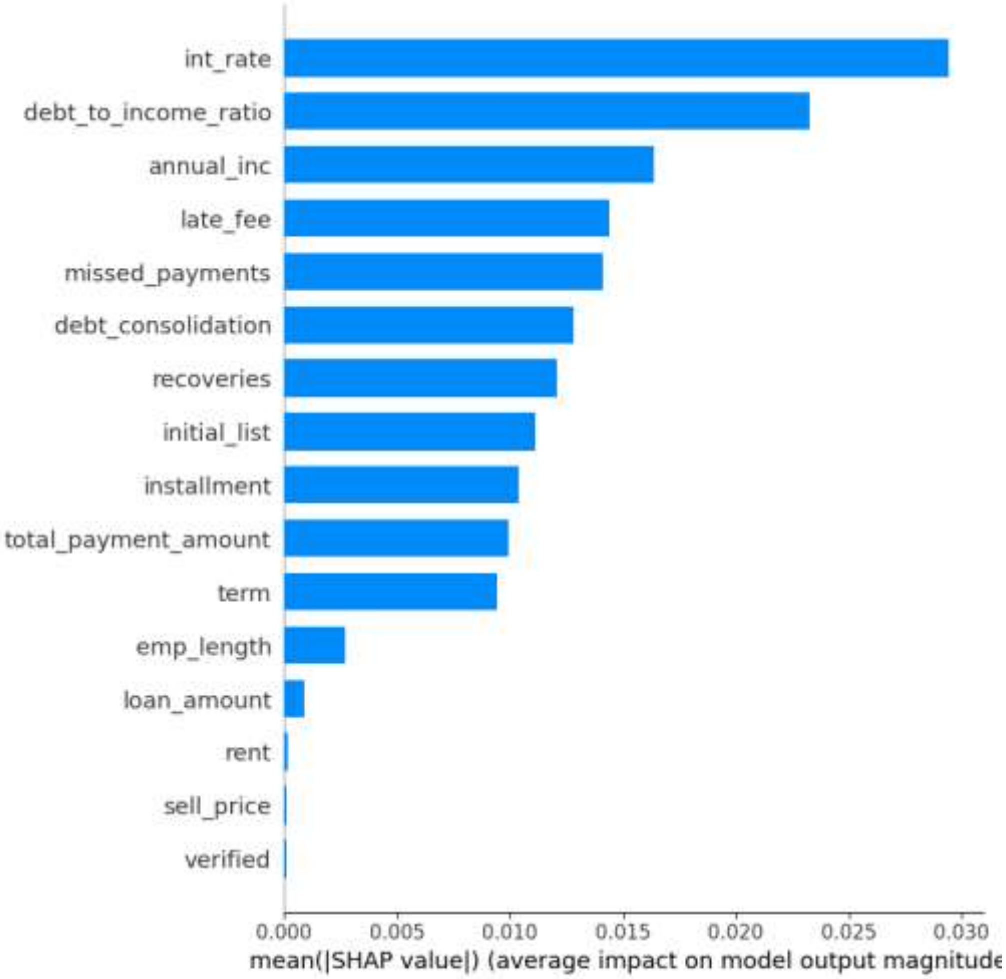

Figure 1: Average Impact of Features on Model Output Magnitude

5. **Conclusion**

In this research article, we address the critical need for accurate credit default prediction models that are also explainable, given the implications of loan default on financial institutions and borrowers alike. Despite the advancements in machine learning techniques, traditional methods like Logistic Regression have been favored due to their interpretability. However, this study explores the potential of tree-based ensemble methods in enhancing both model performance and explainability.

The methodology involves using openly available datasets and employing machine learning algorithms such as Decision Trees, Light Gradient Boosting Machine (LGBM), Random Forests, and XGBoost to construct credit default prediction models. These models are then evaluated based on key metrics like accuracy, precision, recall, and Area Under the ROC Curve (AUC). Additionally, the study includes an explanation of the best fitted model and identifies the features that contribute to these results. Results indicate that while all models perform relatively similarly, XGBoost stands out as the most proficient for the dataset under consideration. Furthermore, the discussion emphasizes the challenges posed by the imbalanced nature of credit default data, affecting the precision values of the models. Finally, our research also emphasizes model explainability by utilizing Shapely values. By meeting a predefined standard of explainability, our proposed model holds promise for practical industrial deployment, addressing regulatory requirements and ethical concerns surrounding discriminatory biases in lending decisions. Future work in this domain could further refine and validate the proposed methods on diverse datasets, ultimately advancing the field of credit risk prediction with explainable AI.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations or those of the publisher, the editors and the reviewers.

References

[1] Barboza, F., Kimura, H. and Altman, E. (2017). Machine learning models and bankruptcy prediction, *Expert Systems with Applications 83*: 405–417.

[2] Bhuiyan, M. S. (2024). The Role of AI-Enhanced Personalization in Customer Experiences. *Journal of Computer Science and Technology Studies*, *6*(1), 162-169.

[3] Chowdhury, O. S., & Baksh, A. A. (2017). IMPACT OF OIL SPILLAGE ON AGRICULTURAL PRODUCTION. *Journal of Nature Science & Sustainable Technology*, *11*(2).

[4] Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics, 57*(1), 203-216.

[5] Charles, K. K. and Hurst, E. (2002). The transition to home ownership and the black-white wealth gap, *The Review of Economics and Statistics 84*(2): 281–297.

[6] Chen, S., Guo, Z. and Zhao, X. (2021). Predicting mortgage early delinquency with machine learning methods, *European Journal of Operational Research 290*(1): 358–372.

[7] Fitzpatrick, T. and Mues, C. (2016). An empirical comparison of classification algorithms for mortgage default prediction: evidence from a distressed mortgage market, *European Journal of Operational Research 249*(2): 427–439.

[8] Guegan, D. and Hassani, B. (2018). Regulatory learning: How to supervise machine learning models? An application to credit scoring, *The Journal of Finance and Data Science 4*(3): 157–171.

[9] Haque, M. S. (2023). Retail Demand Forecasting Using Neural Networks and Macroeconomic Variables. *Journal of Mathematics and Statistics Studies, 4*(3), 01-06.

[10] Haque, M. S., Amin, M. S., & Miah, J. (2023). Retail demand forecasting: a comparative study for multivariate time series. arXiv preprint arXiv:2308.11939.

[11] Islam, M. T., Ayon, E. H., Ghosh, B. P., MD, S. C., Shahid, R., Rahman, S., ... & Nguyen, T. N. (2024). Revolutionizing Retail: A Hybrid Machine Learning Approach for Precision Demand Forecasting and Strategic Decision-Making in Global Commerce. *Journal of Computer Science and Technology Studies*, *6*(1), 33-39.

[12] Jewel, R. M., Linkon, A. A., Shaima, M., Sarker, M. S. U., Shahid, R., Nabi, N., ... & Hossain, M. J. (2024). Comparative Analysis of Machine Learning Models for Accurate Retail Sales Demand Forecasting. *Journal of Computer Science and Technology Studies, 6*(1), 204-210.

[13] Jewel, R. M., Chowdhury, M. S., Al-Imran, M., Shahid, R., Puja, A. R., Ray, R. K., & Ghosh, S. K. (2024). Revolutionizing Organizational Decision-Making for Stock Market: A Machine Learning Approach with CNNs in Business Intelligence and Management. *Journal of Business and Management Studies*, *6*(1), 230-237.

[14] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree, Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, p. 3149–3157.

[15] Keane, M. T. and Smyth, B. (2020). Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai). 163– 178.

[16] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

[17] Moscatelli, M., Parlapiano, F., Narizzano, S. and Viggiano, G. (2020). Corporate default forecasting with machine learning, *Expert Systems with Applications 161*: 113567.

[18] Munnell, A. H., Tootell, G. M. B., Browne, L. E. and McEneaney, J. (1996). Mortgage lending in boston: Interpreting hmda data, *The American Economic Review 86*(1): 25– 53.

[19] Nath, F., Asish, S., Debi, H. R., Chowdhury, M. O. S., Zamora, Z. J., & Muñoz, S. (2023, August). Predicting hydrocarbon production behavior in heterogeneous reservoir utilizing deep learning models. In *Unconventional Resources Technology Conference, 13–15 June 2023* (pp. 506-521). Unconventional Resources Technology Conference (URTeC).

[20] Nath, F., Chowdhury, M. O. S., & Rhaman, M. M. (2023). Navigating Produced Water Sustainability in the Oil and Gas Sector: A Critical Review of Reuse Challenges, Treatment Technologies, and Prospects Ahead. *Water*, *15*(23), 4088.

[21] Prentzas, N., Nicolaides, A., Kyriacou, E., Kakas, A. and Pattichis, C. (2019). Integrating machine learning with symbolic reasoning to build an explainable ai model for stroke prediction, pp. 817–821.

[22] Rana, M. S., Hossain, M. M., Jewel, R. M., & Islam, M. R. (2017). Evaluating Customers Satisfaction of Electronic Banking: An Empirical Study in Bangladesh. *The SIJ Transactions on Industrial, Financial & Business Management*, *5*(03), 07-12.

[23] Uddin, B., Al Mamun, A., Haque, A., & Jewel, R. M. (2017). FACTORS INFLUENCING SELECTION OF HIGHER LEARNING INSTITUTES: AN EMPIRICAL INVESTIGATION IN BANGLADESH. *Aktual'ni Problemy Ekonomiky= Actual Problems in Economics*, *196*, 27-101.

[24] Zhang, M., & Shahid, R. (2024). Enlightening Bangladesh: Navigating power sector challenges through PPP excellence. *Journal of Infrastructure, Policy and Development*, *8*(3).