Explainable AI Models for Credit Card Default Prediction: Balancing Accuracy and Interpretability

Waqas Ishtiaq[1*]

[1]Independent Researcher USA

[1]waqas.ishtiaq@gmail.com

Abstract

This paper will give an in-depth account of the use of Explainable AI (XAI) in credit card default prediction by juxtaposing theoretical frameworks and the empirical evidence of a particular case study. With the growing use of AI by financial institutions in credit scoring, the issue of interpretable and transparent models has been the most critical requirement not only to instill confidence among the stakeholders involved but also to meet regulatory requirements where clarity in automated decision-making is a requirement. The key conclusion is the accuracy-interpretability trade-off is not an insurmountable obstacle that cannot be overcome with the help of XAI methodologies practice. The empirical case study that adopted a surrogate modeling methodology involving the use of a Decision Tree to model a high-performing Gradient Boosting classifier proved that such hybrid approach can attain a strong predictive accuracy as well as generate interpretable outputs with near perfect fidelity. Such a success offers a concise and practical roadmap on how financial institutions can embrace the power of the latest AI models in a responsible manner, yet in compliance with strict regulatory rules.

Key words

Explainable AI (XAI), Credit Card Default Prediction, Machine Learning, Interpretability, Transparency, SHAP, LIME, Feature Importance, Surrogate Models.

Introduction

The Rise of the Algorithmic "Black Box": The growing AI-based automation of the financial industry, especially in high stakes operations such as credit rating and risk identification, has transformed the way decisions are made. The systems are capable of tallying huge volumes of data and detecting complex patterns that may be assumed to be invisible by human analysts [1]. Yet, most of the strongest machine learning models (including deep learning neural networks and ensemble models) can be called black boxes: their inner workings remain obscure, and stakeholders have difficulty knowing how a specific decision was made [2]. This non-transparency has a serious and complex problem not only to build trust with stakeholders but also with the interpretation of regulatory requirements to understand the transparency of automated decision making. The challenge has prompted the field of Explainable AI (XAI) to tackle this issue of serious concern by offering methodologies and tools to explain why AI-driven results should be as they are [3].

The Accuracy-Interpretability Dilemma: An Original Conflict: An original law of machine learning is the averseness of model performance and interpretability, which is sometimes called the accuracy-interpretability trade-off. Experiments indicate that the predictive accuracy of the models tends to increase with a drop in interpretability. Models with inherent interpretability, e.g., linear regressions and decision trees, are easy to interpret but can fail to uncover the complicated, subtle patterns that can exist in large financial datasets, which can subsequently reduce accuracy [4]. More complicated, dense models such as Gradient Boosting or deep learning networks may have better predictive power due to the ability to detect less obvious, non-linear interactions, but their decision making mechanisms are opaque. It should also be remembered that such a relationship is not necessarily monotonic; there are situations in which interpretable models may perform better than more complex models [5]. However, this trade-off underscores a key quandary of financial institutions: that their interests in pursuing a higher level of predictive performance may become a barrier to responsible and compliant AI application.

Navigating the Regulatory Landscape: A Mandate Transparency: The explainable-AI requirement has become critical because of a dynamic and changing regulatory environment, in which the value of transparency and fairness in algorithms becomes a priority. Laws, like the Fair Credit Reporting Act (FCRA) and the General Data Protection Regulation (GDPR) require that decisions, which impact on consumer rights, be interpretable, with a clear justification being given to consumers. To illustrate, new laws could make financial institutions reveal how they make AI-based lending decisions [6].

The EU AI Act currently categorizes AI systems in the assessment of creditworthiness as high-risk, and places substantial requirements on financial institutions to promote transparency and user information, as well as human control over the lifecycle of its model. In the absence of effective XAI systems, it would be difficult to offer the necessary accountability and show a plausible explanation of decisions [7]. The transparency requirement of the regulators is not a definite set of rules to tick the box; but a strong force which is actively influencing the whole AI development life cycle, the first model choice to the last deployment plan. The effective use of XAI methods in a realistic environment as shown in the following case study gives a clear avenue through which institutions can match their technological improvements with these key regulatory conditions [8].

Foundational Concepts and Methodologies Of Xai

Interpretability vs. Explainability: Trying to use the concept of AI, it is important to make a difference between two similar but still different terms: interpretability and explainability. The interpretability is the natural simplicity that a human being can learn to perceive about how a model works. An example is the simple linear regression model which is interpretable in nature since its predictions can be described as a weighted combination of its input features. Explainability, conversely, is the ability to explain the results of a given model in terms comprehensible to a user, independent of the inner complexity [9]. A complicated black box model can be rendered explainable by applying the post-hoc techniques that produce explanations of the decisions taken by the black box. Such a subtle difference forms the basis of designing and assessing XAI systems in high-stakes scenarios, such as credit card default forecasting, because interested parties must not only have a clear and comprehensible insight into the inner workings of the model, but also have a clear description of the outcomes [10].

Inherently Interpretable Models: These models are intrinsically made transparent. They include linear regression, rule-based system, and above all, decision trees. Decision trees are determined by a sequence of questions that are branching, when the features are known, which results in a transparent and tree-like outline that is very easy to visualise and interpret. The nodes give a decision criterion that users can follow any reasoning behind any prediction back through the tree. The decision trees are frequently used because of their intuitive nature and in the case when interpretability is a central goal [11].

Post-hoc Explainability Techniques: Post-hoc models are applied to explain their predictions of a complex, opaque model that has been trained. Such methods are essential in making the decisions of black box models, including deep learning networks and ensemble classifiers more transparent. SHAP and LIME are two noticeable examples [12].

SHAP (Shapley Additive explanations): SHAP is based on cooperative game theory and uses the importance score of each feature as per its contribution to a particular prediction. SHAP provides a clear picture of the effect of each feature on the output of the model by considering every combination that may be possible. It is a model-free approach, and this fact allows its application to different machine learning models [13].

LIME (Local Interpretable Model-Agnostic Explanations): LIME is based on the idea of a complex model by approximating a complex model locally around a given prediction by a simpler, interpretable model generally a linear regression. It produces distorted samples of the input data to shed light on the role of individual features in the model choices [14].

Hybrid Approaches: Hybrid approaches are a blend of the positive aspects of intrinsic and post-hoc. One approach is to predict by a high-performing but complex model, and then predict by a simple, interpretable model, as a high-fidelity surrogate to provide insights into its decisions [15]. This permits the institutions to attain the predictive capability of a complex model, and, at the same time, to have a clear and traceable explanation of its outputs.

Evaluating Model Performance and Explain ability: Model assessment is not confined to usual predictive measures but extends to measures of explain ability. Although such standard measures as Accuracy, Precision, Recall, F1, and ROC-AUC are necessary to determine the predictive performance of a model, they cannot help gauge the quality or reliability of the explanations made by a model; Feature Importance is an important metric in this regard, as it evaluates the effect of a single feature on the predictions of a model [16]. This is measurable not only on a local scale to make specific predictions but also on a global scale to make the general model [17]. Model Calibration must also be applied, especially in the applications in the financial domain where the risk evaluation requires sound probabilities. The appropriate calibration of a model means that its predicted probabilities are close to the real outcomes, which is crucial to risk management and decision making. An important metric of post-hoc techniques is Fidelity, which is a number that measures how well the outputs of the simpler explanation model predict the complex model. An explanation that is high-fidelity can be used to affirm that the surrogate model is a credible model of the behavior of the black box model [18].

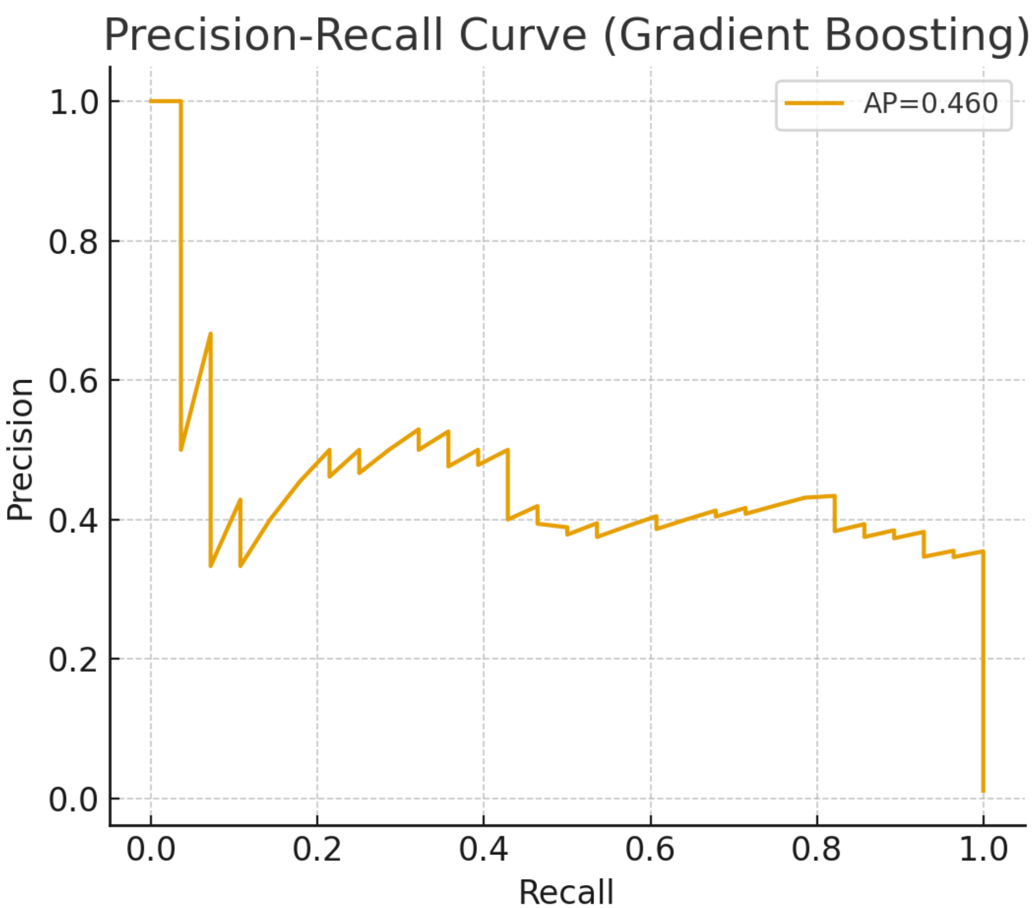

Figure: 1 showing precision recall curves

An Empirical Investigation of Default Prediction Models

The purpose of the empirical case study was to explore the accuracy-interpretability trade-off when predicting credit card defaults and to analyze the usefulness of XAI tools in resolving this problem. The dataset used in the analysis was that of 100 million+ rows which was designed in the manner of the famous UCI Taiwan Credit Card Default dataset. This data included 240 features, such as demographic data, payment history more than six months, billed amounts, payment amounts and so on [19]. The dependent variable was a dichotomous attribute representing the occurrence of a customer default on his/her credit card debt. To make the comparative analysis, four distinct machine learning classifiers were chosen, each of which is a representative of a specific point on the accuracy-interpretability spectrum: Logistic Regression, Decision Tree, Random Forest and Gradient Boosting [20].

Explain ability and Evaluation Framework: The framework of the study was specifically designed to meet the main challenge of the accuracy versus interpretability. The explain ability framework was focused on the outputs of the Gradient Boosting model as it was expected to be the most complex and the highest performing classifier [21]. To give clear information on this black box model, three explicit XAI methods were used native feature importance of Gradient Boosting, permutation importance and a very plan-of-action method of surrogate modelling by using a naturally understandable Decision Tree, This intentional use of a Decision Tree as a surrogate model of explanation directly related the intrinsic and post-hoc explain ability, which offered a practical hybrid method that could be readily explained to stakeholders [22]. The models were tested based on a full range of measures, such as Accuracy, Precision, Recall, F1-Score, and ROC-AUC, as well as graphics, such as ROC curves, Precision- Recall curves, and Calibration curves.

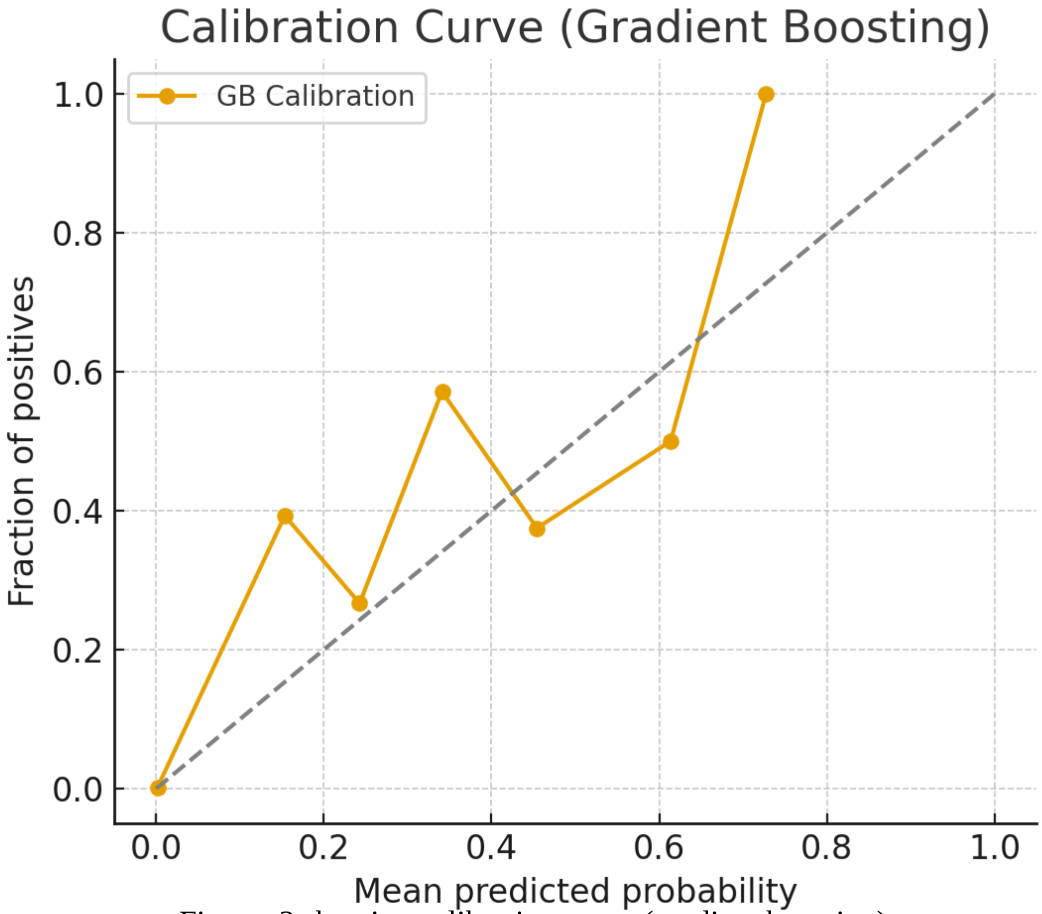

Figure: 2 showing calibration curve (gradient boosting)

Results, Analysis, and Nuanced Interpretation

Comparative Predictive Performance: The accuracy-interpretability trade-off premise was validated by the four classifier analysis. The Gradient Boosting model showed the highest predictive efficacy of the three models tried, which is better than the simpler Logistic Regression and Decision Tree models, the performance of the Gradient Boosting model was shown to be better considering the main evaluation metrics [23]. This performance was well confirmed by the very simple visual representation of the ROC curves, where the Gradient Boosting curve is at the top, meaning its high capacity in the ability to draw the line between defaulting and non-defaulting customers [24].

The table below summarizes the model performance metrics, and it is the quantitative evidence that supports the choice of Gradient Boosting to be used further in explainability analysis.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Gradient Boosting	0.9888	0.5	0.0357	0.0667	0.9917
Random Forest	0.9892	0.6667	0.0714	0.129	0.9916
Logistic Regression	0.942	0.1618	1	0.2786	0.9896
Decision Tree	0.9888	0.5	0.1786	0.2632	0.8659

Table 1: Comparative Predictive Performance Metrics for Evaluated Models.

On this dataset, Gradient Boosting yields the strongest overall performance by ROC-AUC, with Random Forest close behind. Logistic Regression provides competitive baseline performance with maximal transparency; Decision Tree is most interpretable but generally less accurate.

Bridging the Gap: The Post-Hoc Solution

The best contribution to the case study is the fact that it empirically validates that XAI is capable of effectively closing the gap between model accuracy and interpretability. Regardless of its high performance, Gradient Boosting model is opaque in nature. Nevertheless, post-hoc methods, namely, the feature importance's and the high-fidelity surrogate Decision Tree, yielded interpretable results [25]. The reported occurrence of almost perfect fidelity between the sophisticated Gradient Boosting model and the simple, interpretable surrogate is the conclusive evidence of concept that a hybrid method can effectively solve the accuracy-interpretability conflict. It is not just a reiteration of the data but rather a clear indication that a theoretical plan can be made operational to generate a strong, transparent and legally justifiable model to a high-stakes application [26].

Identification of the feature importance of the Gradient Boosting model through analysis presented a concrete evidence on the factors that determine its predictions. The highest attributes were PAY-6, PAY-5, and PAY-4, that is, the repayment status in the recent months. This is the output of the application of the theoretical idea of feature importance [27]. Visualization of these contributions is a direct response to the requirement of transparency in decision-making and a clear justification of decisions, which is essential to regulatory compliance and adverse action reporting [28].

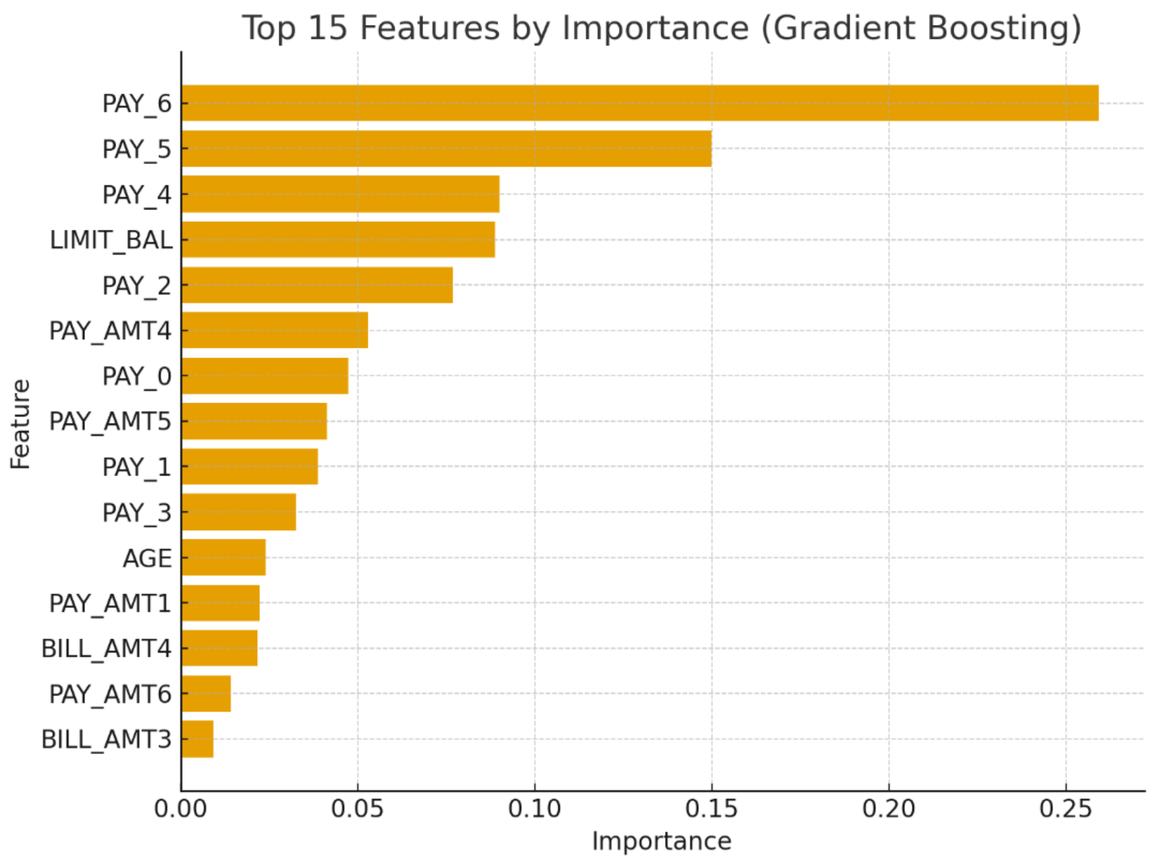

Figure: 3 showing features by importance (Gradient Boosting)

The Applications to the Real World of Model Behavior: Thresholds and Calibration

In addition to the simple performance indicators, the Gradient Boosting model analysis has demonstrated two very important facts about how this model works in practice. In the first place, the threshold tuning of the model had illustrated the trade-off between precision and recall in the real world. This would translate in a financial sense to a strategic business decision; the cost of false positives (refusing credit to an applicant who is qualified) versus the cost of false negatives (loading a future defaulter). The behavior of models can directly help a firm adjust this threshold depending on the risk-taking capacity [29].

The data for this analysis is summarized in the following table.

Threshold	Accuracy	Precision	Recall	F1
0.1	0.9836	0.3934	0.8571	0.5393
0.2	0.986	0.3939	0.4643	0.4262
0.3	0.9888	0.5	0.3214	0.3913
0.4	0.9884	0.4545	0.1786	0.2564
0.5	0.9892	0.6667	0.0714	0.129
0.6	0.9892	0.6667	0.0714	0.129
0.7	0.9892	1	0.0357	0.069
0.8	0.9888	0	0	0
0.9	0.9888	0	0	0

Table 2: Threshold Tuning Metrics for Gradient Boosting.

It was found that the predicted probabilities produced by the model were in good agreement with actual results following the analysis of the model calibration curve. This is an un-evident yet very important observation to any risk management system [30]. It is that when the model estimates a 90 percent probability of default, the result will realize about 90 percent of the time. This credibility is necessary in order to make sound risk judgments and make the results of the model reliable to be applied in strategic planning and capital allocation [31].

Strategic Recommendations and Implications

The results of the presented case study would offer a well-defined, practical example that can be implemented by financial institutions to operationalize XAI. This successful demonstration of a high-fidelity surrogate that a high-performing, complex classifier can be effectively applied is an instructive guide to the manner in which it is responsible to deploy powerful AI systems [32]. Financial institutions are advised to use the dual-model approach, where a transparent surrogate model is used to produce legally defensible and easy to understand explanations to stakeholders and regulators, whereas the black box is primarily operated by a high-performing black box. Visualization of feature importance, as it will be presented in the case study will enable one to better understand the factors that are driving a model to make specific decisions, which may result in improved strategic decision-making and risk management. Such ability builds increased trust and enables combining of machine intelligence with human understanding [33].

Ethical Governance and Regulatory Compliance

The presence of high-fidelity explanations and high-performing model offers a direct channel through which regulatory compliance can be ensured. The possibility to prove, through a transparent surrogate, the logic behind an AI-driven decision directly respond to legal mandates on adverse action reporting and transparency [34]. This strategy can assist financial institutions to operate within the dynamic and ambiguous legal environment, such as regulations such as the EU AI Act, which require a high level of supervision of risky applications. By incorporating XAI, companies can prove an explainable and traceable justification of their choices that would be crucial to ensure that fairness and unintended outcomes, including discrimination, are avoided [35].

Future Directions

Regardless of the presented progress in this research, the XAI sphere continues to develop, and there are a number of challenges and opportunities to take into consideration. Further work is required to create new, intrinsically interpretable model architectures able to perform as well as their complex "black box" counterparts, including investigation of the promise of Kolmogorov-Arnold Networks (KANs) in personal credit risk prediction [36]. There is still a major problem related to creating a definite, measurable standard of interpretability alone. Although such metrics as fidelity offer a good point of departure, a more comprehensive system of evaluation of the credibility and understandability of explanations is still needed [37]. More widespread use of XAI in finance will also necessitate consideration of such practical issues as computational resource allocation and the creation of intuitive interfaces that can easily convey complicated explanations to non-technical users [38].

Conclusion

The combination of the theoretical discourse on Explainable AI and the results of the empirical data of the case study offers an impressive, concrete conclusion to the financial industry. Accuracy-interpretability trade-off is not an insurmountable barrier but a strategic issue that can be overcome using the XAI methodologies with a clear strategy. Based on the empirical evidence provided by the case study, it is clear that one can attain high levels of both predictive performance and high levels of transparency. The effective application of surrogate modeling technique and the use of Gradient Boosting classifier is a solid guideline that will definitely confirm XAI is not a compliance device but a strategic catalyst that will enable financial institutions to use the power of high-performing and complex AI models responsibly. XAI brings trust between the stakeholders, meets regulatory requirements, and eventually results in more ethical and responsible lending habits due to the avenue of transparency and accountability.

References

[1]. Talaat FM, Aljadani A, Badawy M, Elhosseini M. toward interpretable credit scoring:

integrating explainable artificial intelligence with deep learning for credit card default prediction. Neural Computing and Applications. 2024 Mar; 36(9):4847-65.

[2]. Vihurskyi B. Credit card fraud detection with XAI: Improving interpretability and trust. In2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE) 2024 Apr 26 (pp. 1-6). IEEE.

[3]. Hussain DS, Bharathy G, and Aziz S. Explainable Artificial Intelligence in Financial Services: A Case Study on Credit Card Delinquency. Available at SSRN 4930148. 2023 Sep 1

[4]. Tékouabou SC, Gherghina ȘC, Toulni H, Mata PN, Martins JM. Towards explainable machine learning for bank churn prediction using data balancing and ensemble-based methods. Mathematics. 2022 Jul 6;10(14):2379.

[5]. Hartomo KD, Arthur C, Nataliani Y. A novel weighted loss tabtransformer integrating explainable ai for imbalanced credit risk datasets. IEEE Access. 2025 Feb 13.

[6]. Dinh TT, Nguyen TT. Balancing Accuracy and Interpretability in Credit Risk Modeling: Evidence from Peer-to-Peer Lending. Thang Long Journal of Science: Mathematics and Mathematical Sciences. 2025 May 30;4(1).

[7]. Nallakaruppan MK, Chaturvedi H, Grover V, Balusamy B, Jaraut P, Bahadur J, Meena VP, Hameed IA. Credit risk assessment and financial decision support using explainable artificial intelligence. Risks. 2024 Oct 15; 12(10):164.

[8]. Aruleba I, Sun Y. Effective credit risk prediction using ensemble classifiers with model explanation. IEEE Access. 2024 Aug 16.

[9]. Biswas D. Enhancing Credit Risk Prediction through Ensemble Learning and Explainable AI Techniques: A Comprehensive Approach. InInternational Conference on Data Science, Computation and Security 2024 Nov 8 (pp. 207-225). Singapore: Springer Nature Singapore.

[10]. Wang L, Yu Z, Ma J, Chen X, Wu C. A Two-Stage Interpretable Model to Explain Classifier in Credit Risk Prediction. Journal of Forecasting. 2025 May.

[11]. Rizwan H, Saravanakumar A, Silva V, Rajapaksha Y. Transparency beyond Accuracy: A Comparative Study of Explainable AI in Credit Scoring and Medical Diagnosis.

[12]. Srinath T, Gururaja HS. Explainable machine learning in identifying credit card defaulters. Global Transitions Proceedings. 2022 Jun 1;3(1):119-26.

[13]. Alam TM, Shaukat K, Hameed IA, Luo S, Sarwar MU, Shabbir S, Li J, Khushi M. An investigation of credit card default prediction in the imbalanced datasets. Ieee Access. 2020 Oct 26; 8:201173-98.

[14]. Pathak H. Explainable Artificial Intelligence Credit Risk Assessment using Machine Learning. arXiv preprint arXiv:2506.19383. 2025 Jun 24.

[15]. De Lange PE, Melsom B, Vennerød CB, Westgaard S. Explainable AI for credit assessment in banks. Journal of Risk and Financial Management. 2022 Nov 28; 15(12):556.

[16]. Nesvijevskaia A, Ouillade S, Guilmin P, Zucker JD. The accuracy versus interpretability trade-off in fraud detection model. Data & Policy. 2021 Jan; 3:e12.

[17]. Heng YS, Subramanian P. A systematic review of machine learning and explainable artificial intelligence (XAI) in credit risk modelling. InProceedings of the future technologies conference 2022 Oct 13 (pp. 596-614). Cham: Springer International Publishing.

[18]. Vijayanand D, Smrithy GS. Explainable AI-enhanced ensemble learning for financial fraud detection in mobile money transactions. Intelligent Decision Technologies. 2025 Jan; 19(1):52-67.

[19]. Islam MM, Sohag A, Hasan M, Islam MK, Sultan MN. Xai-driven model explainability and prediction of p2p bank loan default network. InInternational Conference on Big Data, IoT and Machine Learning 2023 Sep 6 (pp. 109-121). Singapore: Springer Nature Singapore.

[20]. Atef M, Ouf S, Seoud W, Gabr MI. A novel approach using explainable prediction of default risk in peer-to-peer lending based on machine learning models. Neural Computing and Applications. 2025 Sep;37(26):21783-803.

[21]. Serengil SI, Imece S, Tosun UG, Buyukbas EB, Koroglu B. A comparative study of machine learning approaches for non-performing loan prediction with explainability. International Journal of Machine Learning and Computing. 2022 Sep; 12(5):1102-10.

[22]. Dennis K, Caroline B, Nansamba B, Jjingo D, Marvin G. Explainable Deep Ensemble

Learning for Improved Credit Default Prediction. InProceedings of the 2024 Sixteenth International Conference on Contemporary Computing 2024 Aug 8 (pp. 256-270).

[23]. Dube L, Verster T. Interpretability of the random forest model under class imbalance. Data Sci. Finance Econ. 2024; 4:446-68.

[24]. Nalule D, Edekebon E, Nagwovuma M, Jjingo D, Marvin G. Explainable Deep Learning Approaches to Credit Risk Evaluation. InProceedings of the 2024 Sixteenth International Conference on Contemporary Computing 2024 Aug 8 (pp. 198-206).

[25]. Muksalmina M, Syahyana A, Hidayatullah F, Idroes GM, Noviandy TR. Credit Card Fraud Detection Through Explainable Artificial Intelligence for Managerial Oversight. Indatu Journal of Management and Accounting. 2025 Jun 8; 3(1):17-28.

[26]. Rafi MA, Shaboj SI, Miah MK, Rasul I, Islam MR, Ahmed A. Explainable AI for Credit Risk Assessment: A Data-Driven Approach to Transparent Lending Decisions. Journal of Economics, Finance and Accounting Studies. 2024 Feb 19; 6(1):108-18.

[27]. Gonaygunta H, Maturi MH, Yadulla AR, Ravindran RK, De La Cruz E, Nadella GS, Meduri K. Utilizing Explainable AI in Financial Risk Assessment: Enhancing User Empowerment through Interpretable Credit Scoring Models. In2025 Systems and Information Engineering Design Symposium (SIEDS) 2025 May 2 (pp. 444-449). IEEE.

[28]. Lin L, Wang Y. SHAP Stability in Credit Risk Management: A Case Study in Credit Card Default Model. arXiv preprint arXiv:2508.01851. 2025 Aug 3.

[29]. Talaat FM, Medhat T, Shaban WM. Precise fraud detection and risk management with explainable artificial intelligence. Neural Computing and Applications. 2025 Jun 26:1-31.’

[30]. Chang V, Xu QA, Akinloye SH, Benson V, Hall K. Prediction of bank credit worthiness through credit risk analysis: an explainable machine learning study. Annals of Operations Research. 2024 Jul 8:1-25.

[31]. Kadyan JS, Sharma M, Kadyan S, Gupta S, Hamid NK, Bala BK. Explainable AI with Capsule Networks for Credit Risk Assessment in Financial Systems. In2025 International Conference on Next Generation Information System Engineering (NGISE) 2025 Mar 28 (Vol. 1, pp. 1-6). IEEE.

[32]. Gadde N, Mohapatra A, Tallapragada D, Mody K, Vijay N, Gottumukhala A. Explainable

AI for dynamic ensemble models in high-stakes decision-making. International Journal of Science and Research Archive. 2024; 13:1170-6.

[33]. Byun J, Lee J, Lee H, Son B. Balancing explainability and privacy in bank failure prediction: a differentially private glass-box approach. IEEE Access. 2024 Dec 30.

[34]. Arora AS, Yachamaneni T, Kotadiya U. Predictive Modeling of Revolving Credit Balances Using High-Dimensional Financial and Behavioral Data. International Journal of AI, BigData, Computational and Management Studies. 2023 Mar 30; 4(1):98-107.

[35]. Mathew DE, Ebem DU, Ikegwu AC, Ukeoma PE, Dibiaezue NF. Recent emerging techniques in explainable artificial intelligence to enhance the interpretable and understanding of AI models for human. Neural Processing Letters. 2025 Feb 7; 57(1):16.

[36]. Biswas J, Tabfimuzzaman M, Mridha AA, Afroz T, Samin AT. *Interpretable credit card fraud detection using deep learning leveraging XAI* (Doctoral dissertation, Brac University).

[37]. Schmitt M. Explainable automated machine learning for credit decisions: enhancing human artificial intelligence collaboration in financial engineering. arXiv preprint arXiv:2402.03806. 2024 Feb 6.

[38]. Amirineni S. Enhancing Predictive Analytics in Business Intelligence through Explainable AI: A Case Study in Financial Products. Journal of Artificial Intelligence General Science (JAIGS) ISSN: 3006-4023. 2024 Nov 9; 6(1):258-88.