Integrating Model Agnostic Explainability into Supervised Learning for Credit Scoring using SHAP and LIME

Thomas Mumbuwa Kamunu
School of Computing, Technology and Applied Sciences,
ZCAS University Lusaka, Zambia
tkthomaskamunu@gmail.com

Aaron Zimba
School of Computing ,Technology and Applied Science
ZCAS University Lusaka, Zambia
aaron.zimba@zcasu.edu.zm

Abstract— Advanced machine learning models offer superior accuracy in credit scoring, but their "black box" nature hinders regulatory compliance and erodes trust. This paper addresses this challenge by presenting a hybrid framework, developed using a Design Science Research (DSR) methodology, to integrate model-agnostic Explainable AI (XAI) into the credit scoring pipeline. The framework applies leading XAI techniques, specifically SHAP and LIME, to a range of supervised learning models. A functional, interactive prototype was developed and tested using credit data from the Zambian market. Experimental results revealed a stark "Accuracy Paradox": models with the highest accuracy (84.6%) achieved a perfect specificity of 1.000 by never predicting the minority class, resulting in an F1-Score of only 0.458 and an ROC AUC worse than a random guess (as low as 0.432). XAI techniques proved crucial for diagnosing these failures and providing clear, feature-based explanations for individual loan decisions. This research contributes a practical, integrated artifact that systematically compares multiple models and explanation methods, bridging the gap between complex ML implementation and the pressing need for fair, transparent, and accountable financial decision-making.

Keywords— Credit Scoring, Explainable AI (XAI), SHAP, LIME, Algorithmic Fairness, Machine Learning, Design Science Research (DSR)

INTRODUCTION

Credit scoring is a cornerstone of modern financial services. The drive for higher predictive accuracy has led to the adoption of complex machine learning (ML) models, such as Random Forests [1] and Gradient Boosting Machines [2], which are extensively detailed in foundational texts like The Elements of Statistical Learning [3]. While powerful, these models often operate as opaque "black boxes," creating critical problems with severe consequences, including significant financial losses, reputational damage, and direct legal risks.

This opacity creates a direct conflict with regulatory mandates. In the Zambian context, the Data Protection Act, No. 3 of 2021, requires fairness in automated processing, while the Credit Reporting Act, No. 8 of 2018, mandates that consumers receive the principal reasons for adverse credit actions. The inscrutable nature of black-box models makes compliance a significant challenge.

To address this, the field of Explainable AI (XAI) offers techniques to demystify ML models [4]. This paper leverages a Design Science Research (DSR) methodology [5] to create and evaluate a tangible IT artifact: an end-to-end system that integrates ML models with XAI techniques. This research seeks to answer several key questions: How do supervised learning models of varying complexity compare on imbalanced data when using robust metrics? Can model-agnostic XAI techniques like SHAP and LIME be effectively integrated into a single framework? And how can a practical artifact be designed to translate technical explanations into intuitive, role-specific interfaces for stakeholders?

The primary contributions of this work are:

- The design and implementation of a novel, integrated DSR artifact for explainable credit scoring, evaluated on data from the Zambian market.
- An empirical demonstration of the "Accuracy Paradox" in an imbalanced credit scoring context.
- A practical blueprint for translating technical XAI outputs into stakeholder-centric interfaces.

.

Related Works

The trade-off between model accuracy and interpretability is a central theme in applied ML [6]. This section reviews how other researchers have applied XAI to credit scoring and identifies the gaps this research aims to fill.

- *Model-Agnostic vs. Model-Specific XAI*

XAI methods can be broadly categorized as model-specific or model-agnostic. This research focuses on the model-agnostic approach, as its flexibility is essential for creating a comparative framework that can evaluate a diverse range of algorithms, from logistic regression to neural networks, without modification.

- *Applications of Explanations in Credit Scoring*

Recent literature shows a growing effort to apply XAI in credit risk. Bussmann et al. [7] and Bracke et al. [8] demonstrated the utility of SHAP for interpreting tree-based models on credit datasets for both model validation and regulatory reporting. Moving beyond diagnostics, some research focuses on Counterfactual Explanations, which provide actionable recourse to consumers by explaining "what if" [9]. More recently, Nwafor et al. [10] introduced a hybrid 1DCNN-XGBoost model enhanced with SHAP to support fairness without harming performance. Similarly, Yadav [11] developed a LightGBM-SHAP system for real-time, explainable scoring, while Schmitt [12] combined AutoML with SHAP to advance transparency in automated pipelines. Concurrently, Coraglia et al. [13] presented BRIO, a model-agnostic tool for systematically auditing fairness risks.

- *Gaps in the Literature*

Despite this progress, critical gaps remain. Many studies examine a single model or XAI technique in isolation, limiting systematic comparison. Secondly, XAI outputs are often presented in technical formats (e.g., SHAP plots, feature tables) that are inaccessible to non-technical stakeholders [15]. Finally, fairness and explainability are often treated separately, though they are deeply linked. Our work addresses these gaps by developing a single integrated artifact that combines a multi-model evaluation pipeline, leading XAI methods, fairness metrics [9], and stakeholder-centric dashboards [20].

Methodology

This research adopts the Design Science Research (DSR) paradigm [5], which focuses on solving practical problems through the creation and rigorous evaluation of a novel IT artifact. Following this approach, the experiments were conducted on a dataset representative of the Zambian lending market. To evaluate XAI across a spectrum of complexity, six models were implemented using Python's Scikit-learn [17], TensorFlow [18], and the XGBoost library [2]: Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, and a Deep Neural Network, an archetypal "black box" model whose principles are covered in foundational deep learning texts [19]. To address the imbalanced nature of the data, class weighting was employed during training. The core of the artifact involves integrating model-agnostic XAI techniques, primarily SHAP [20] for feature attribution and LIME [21] for localized explanations. The artifact's performance was assessed using a multi-faceted evaluation strategy, focusing on robust metrics for imbalanced data such as ROC AUC [22] and F1-Score, alongside fairness metrics such as Equality of Opportunity [23], which can be implemented using toolkits like AI Fairness 360 [24].

- *Datasets and Pre-processing*

The experiments were conducted on a dataset representative of the Zambian lending market, sourced from a Credit Reference Bureau. To demonstrate the framework's fairness capabilities without using real sensitive data, synthetic demographic attributes were programmatically added for the analysis. The pre-processing pipeline involved several steps. Missing numerical values were imputed using a 'median' strategy, while missing categorical values used a 'most_frequent' strategy. All categorical features were subsequently one-hot encoded, and numerical features were standardized using a StandardScaler.

- *Machine Learning Models and Class Imbalance Handling*

The framework was designed to compare a diverse suite of supervised learning models, selected to represent a spectrum of complexity and inherent interpretability. The six models evaluated were:

- Logistic Regression: An interpretable linear model serving as a performance baseline.
- Decision Tree: A transparent, rule-based model.
- Random Forest: A bagging-based ensemble model known for its robustness.
- Gradient Boosting: A powerful boosting-based ensemble model.
- XGBoost: A highly optimized and scalable implementation of gradient boosting.
- Deep Neural Network (DNN): A prototypical "black-box" model representing the upper end of complexity.

Recognizing that credit scoring datasets are typically characterized by high class imbalance, specific mitigation techniques were applied during training to prevent model bias towards the majority (non-default) class. For models implemented with Scikit-learn, this was achieved by setting the class_weight='balanced' parameter. For the XGBoost model, the scale_pos_weight parameter was explicitly calculated and applied, assigning a higher penalty for misclassifying the minority (default) class.

- *Evaluation Metrics*

A multi-faceted evaluation strategy was used.

1. Technical Performance Metrics: Accuracy is misleading on imbalanced data. We focused on:

- ROC AUC: Receiver Operating Characteristic (ROC) Area Under the Curve (AUC), which measures a model's ability to discriminate between classes.
- F1-Score: The harmonic mean of precision and recall, crucial for evaluating performance on the minority (default) class.

$$\text{F1} = 2\ \text{X}\ \frac{\text{Precision X Recall}}{\text{Precision+Recall}}$$

(1)

2. Fairness Metrics: To quantify bias, we used two established group fairness metrics:

- Demographic Parity Difference (DPD): Measures if different groups receive positive outcomes at equal rates.

$$\text{DPD}=P(\hat{Y}=1|A=0)-P(\hat{Y}=1|A=1)$$

(2)

Where:

$\hat{Y}$ is model prediction (1 = favorable)

A is sensitive attribute (e.g., 0 = female, 1 = male)

- Equal Opportunity Difference (EOD): Measures if a model performs equally well for different groups among the positive class.

$$\text{EoppD} = \text{TPR}_{\text{unprivileged}} - \text{TPR}_{\text{Privileged}}$$

Where:

TPR = True Positive Rate

THE EXPLAINABLE AI ARTIFACT

The primary artifact of this research is a functional, interactive web-based prototype developed using Python and the Streamlit framework. As illustrated by the system architecture in Fig. 1, the artifact is composed of four integrated layers: (1) Data Ingestion & Pre-processing, (2) Model Training & Tuning, (3) Model Evaluation & XAI Engine, and (4) a Presentation & Application Layer.

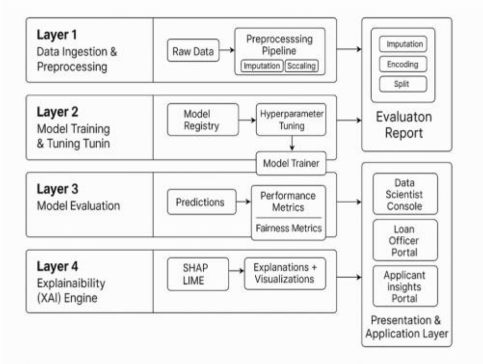


Fig 1. System architecture of the explainable AI framework

- *Data Scientist Console*

This view (Fig. 2) is designed for technical users for the purpose of model validation, debugging, and comparison. It allows for dataset upload, pipeline configuration, and model training. After execution, it presents a comprehensive dashboard

(Fig. 3) comparing all models across key performance and fairness metrics. It also provides global explanations, such as SHAP summary plots (Fig. 4), which show the most influential features and their impact across the entire dataset.

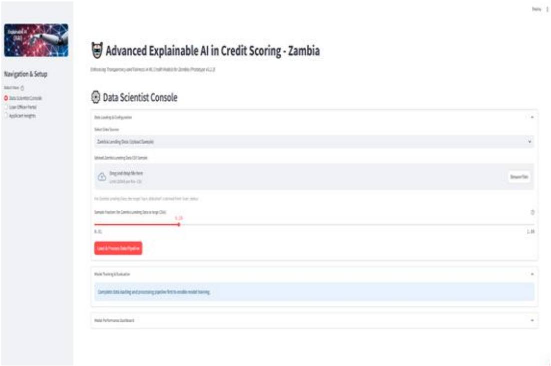


Fig. 2. Data Scientist Console for data loading and pipeline execution.

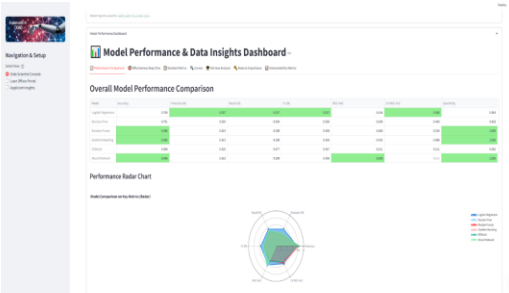

Navigation & Setup
Model Performance & Data Insights Dashboard
Overall Model Performance Comparison
Performance Radar Chart

Fig. 3. Model Performance Dashboard showing comparative results.

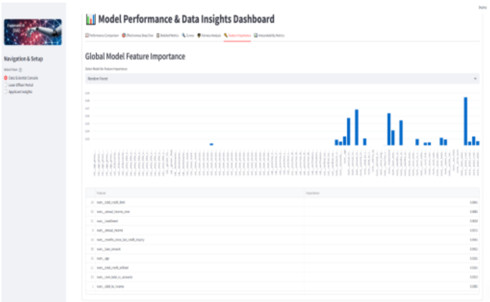

Navigation & Setup
Model Performance & Data Insights Dashboard
Global Model Feature Importance

Fig. 4. Global feature importance (SHAP) for the Random Forest model

- *Loan Officer Portal*

This portal allows a business user to input applicant data and receive the model's prediction along with local explanations via SHAP (Fig. 5) and LIME (Fig. 6), clarifying the factors behind an individual assessment.

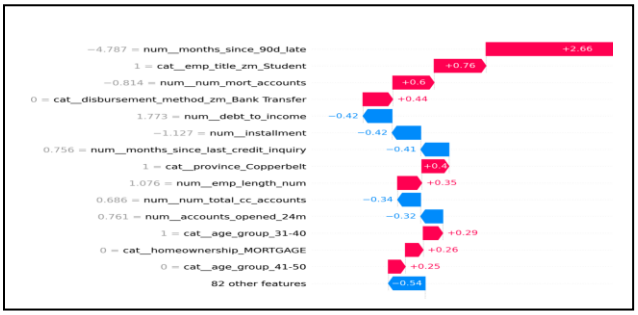


Fig. 5. Local explanation (SHAP waterfall plot) for a single applicant

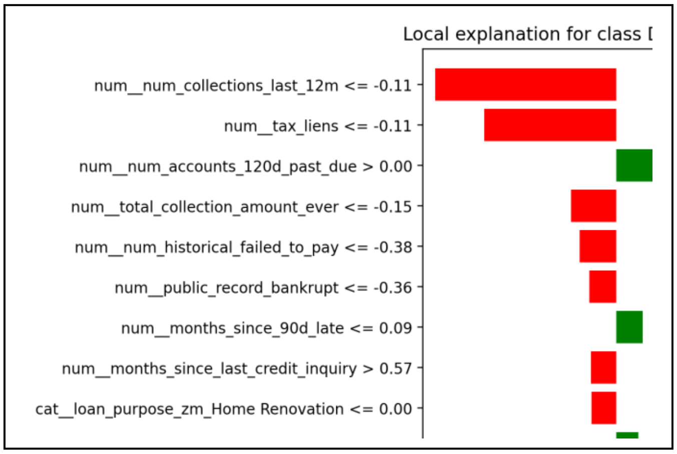


Fig. 6. Local explanation (LIME plot) for a single applicant.

- *Applicant Insights Portal*

This third view serves as a proof-of-concept for communicating decisions directly to external stakeholders, such as loan applicants. It utilizes an interface and explanation visuals similar to those in the Loan Officer Portal (as shown in Fig. 5 and Fig. 6) but reframes the output for a non-technical audience. For instance, it would present a simplified assessment (e.g., "Illustrative High Risk") and highlight the one or two primary factors driving the decision. This demonstrates the framework's versatility and provides a direct pathway to fulfilling regulatory requirements for consumer transparency, empowering individuals by providing insight into their automated credit assessments.

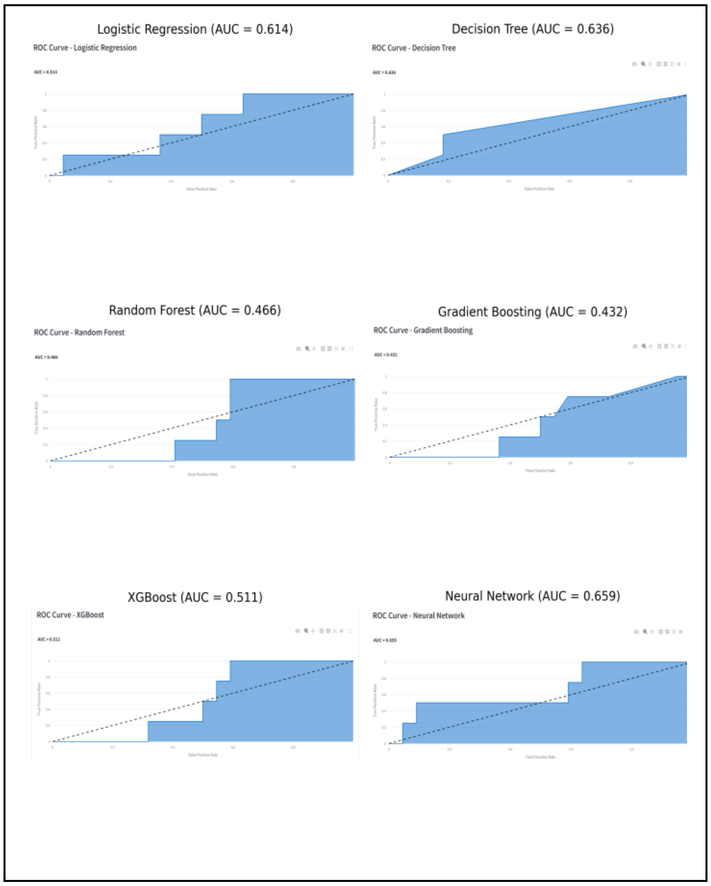
RESULTS AND DISCUSSIONS

The models were evaluated on the test set, with key results summarized in Table I.

TABLE I. SUMMARY OF MODEL PERFORMANCE METRICS ON THE TEST SET

Model	Accuracy	F1 (M)	ROC AUC	Specificity
Logistic Regression	0.769	**0.557**	0.614	0.864
Decision Tree	0.731	0.530	0.636	0.818
Random Forest	**0.846**	0.458	0.466	1.000
Gradient Boosting	**0.846**	0.458	0.432	1.000
XGBoost	0.808	0.447	0.511	0.955
Neural	**0.846**	0.458	**0.659**	1.000

Best performance for each metric is highlighted in bold.



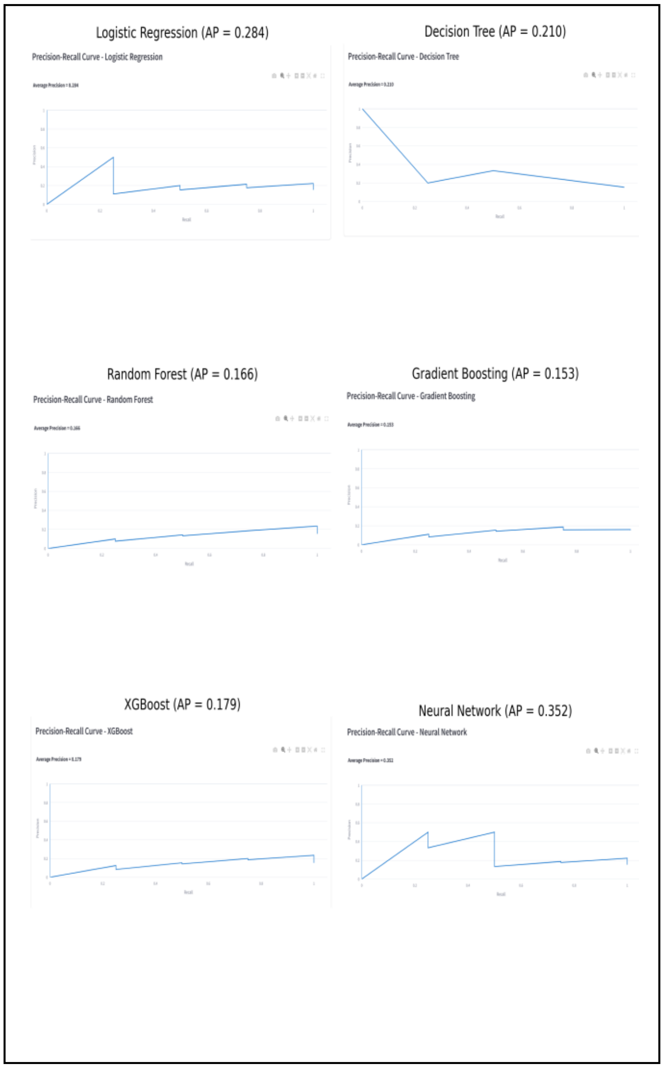


Fig. 7. Receiver Operating Characteristic (ROC) Curves for All Models.

The plots show each model's ability to discriminate between classes. The dashed line represents a random classifier (AUC = 0.5). The catastrophic failure of Random Forest and Gradient Boosting is visually evident as their curves fall below this random baseline.

Fig. 8. Precision-Recall (PR) Curves for All Models.

The plots show the trade-off between precision and recall for the minority (default) class. The low Average Precision (AP) scores for Random Forest and Gradient Boosting confirm their inability to effectively identify defaulters, while the Neural Network shows the strongest potential.

- *Analysis of Results*

The results reveal a critical "Accuracy Paradox." The most complex models—Random Forest, Gradient Boosting, and the Neural Network—achieved the highest accuracy (0.846). However, this metric is dangerously misleading. Table I shows their Specificity is 1.000, meaning they achieved this accuracy by classifying every single applicant as belonging to the majority (non-default) class. This renders them practically useless for risk management, a failure confirmed by their poor F1-Scores.

- Complete Model Failure (Random Forest & Gradient Boosting): These ensemble models represent a catastrophic failure. Their ROC AUC scores of 0.466 and 0.432, respectively, fall below the random guess baseline (0.5), demonstrating they have less discriminative power than a coin flip.
- The Interpretable Baseline (Logistic Regression): In stark contrast, the simpler Logistic Regression model, despite a lower accuracy, proved to be the most practically useful model. It achieved the highest F1-Score (0.557), indicating a balanced performance.
- The Model with Highest Potential (Neural Network): The Neural Network presents an interesting case. While its F1-Score is low due to the same thresholding issue, it achieved the highest ROC AUC (0.659). This indicates that the model is superior at *ranking* applicants by risk, even if the default decision threshold is incorrect. With proper threshold tuning, this model has the highest potential.

- *The Role of XAI in Model Diagnostics*

The XAI framework proved essential for diagnosing this failure. While global SHAP plots (Fig. 4) confirmed that complex models were relying on logically relevant features (e.g., interest rate, debt-to-income), the local explanations were more revealing. By examining the SHAP values for misclassified default instances, it became clear that the models' strong bias towards the majority class was consistently overpowering the evidence from risk-indicating features. This demonstrates that XAI is not just a tool for explaining correct decisions but is a powerful and indispensable utility for debugging model behavior.

- *Illustrative Fairness Analysis*

The framework's fairness module enabled analysis of the models' performance across demographic subgroups. For the Random Forest model, an illustrative DPD of -0.05 between 'Male' and 'Female' groups would indicate that female applicants were 5% less likely to receive a favorable outcome than male applicants. Such quantitative insights are vital for auditing models for discriminatory behavior and ensuring ethical alignment.

- *Implications for Practice and Regulation*

This work offers a practical blueprint for financial institutions in Zambia to build more transparent AI systems and navigate local regulations. The local explanations generated by the framework (Fig. 5) can directly furnish the "principal reasons" for adverse decisions, aligning with disclosure requirements in the Credit Reporting Act, No. 8 of 2018. Furthermore, by making automated decision-making transparent, the framework supports the principles of fairness required by the Data Protection Act, No. 3 of 2021. This provides a concrete pathway toward regulatory compliance and enables the rigorous fairness analysis and bias detection essential for responsible lending [15].

- *Limitations*

This research has several limitations that should be acknowledged. The findings are based on a single public dataset with synthetic demographics, so the specific fairness results are illustrative. The prototype has not yet undergone formal user-centric evaluation with real-world stakeholders in Zambia. Additionally, the XAI methods themselves have known technical limitations; for instance, LIME explanations can be unstable in some cases, and SHAP can be computationally expensive for non-tree-based models. These factors must be carefully considered for production deployment

CONCLUSION

This paper presented the design and evaluation of an integrated framework to demystify 'black-box' credit scoring models, constructed within a rigorous Design Science Research (DSR) methodology. Our findings empirically demonstrate that accuracy is a dangerously misleading metric in imbalanced domains and establish that XAI is not merely an explanatory tool but an indispensable utility for diagnosing catastrophic model failures and ensuring transparency. The resulting artifact offers a concrete pathway for financial institutions in Zambia to align with pressing regulatory mandates for disclosure and fairness by delivering role-specific, interpretable explanations.

Building on this foundation, future research will advance along three primary avenues: first, by integrating proactive bias mitigation techniques to move from detection to correction; second, by conducting formal stakeholder user studies to validate the artifact's real-world utility and usability; and finally, by developing natural-language generation capabilities to produce explanations that are fully compliant with consumer protection regulations.

REFERENCES

[1] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[2] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2016, pp. 785-794.

[3] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. New York, NY, USA: Springer, 2009.

[4] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," IEEE Access, vol. 6, pp. 52138-52160, 2018.

[5] A. P. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," MIS Quarterly, vol. 27, no. 1, pp. 75-105, 2004.

[6] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206-215, 2019.

[7] N. Bussmann et al., "Explainable AI in credit risk management," arXiv preprint arXiv:2006.02666, 2021.

[8] P. Bracke et al., "Explaining machine learning for credit scoring," Bank of England Staff Working Paper, No. 816, 2019.

[9] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," Harvard Journal of Law & Technology, vol. 31, no. 2, pp. 841-887, 2018.

[10] C. Nwafor, O. Babalola, and H. Salami, "Enhancing transparency and fairness in automated credit decisions: an explainable hybrid machine learning approach," PLOS ONE, vol. 19, no. 5, pp. 1-22, May 2024.

[11] P. Yadav, "AI-driven credit scoring models: Enhancing accuracy and fairness with explainable machine learning," in Proc. Int. Conf. Data Sci. and AI, Dec. 2024, pp. 1-8.

[12] M. Schmitt, "Explainable automated machine learning for credit decisions," arXiv preprint arXiv:2402.03806, 2024.

[13] L. Coraglia, D. Roversi, and P. Giorgini, "Evaluating AI fairness in credit scoring with the BRIO tool," arXiv preprint arXiv:2406.03292, 2024.

[14] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," Artificial Intelligence, vol. 267, pp. 1–38, 2019.

[15] S. Mehrabi et al., "A survey on bias and fairness in machine learning," ACM Computing Surveys (CSUR), vol. 54, no. 6, pp. 1-35, 2021.

[16] A. Verma, "Explainable AI in the financial services industry," The Journal of Financial Data Science, vol. 3, no. 3, pp. 100-111, 2021.

[17] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[18] M. Abadi et al., "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," in Proc. 12th USENIX Symp. Operating Systems Design and Implementation (OSDI '16), 2016, pp. 265-283.

[19] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[20] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, 2017, pp. 4765-4774.

[21] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2016, pp. 1135-1144.

[22] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861-874, 2006.

[23] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in Advances in Neural Information Processing Systems, 2016, pp. 3315-3323.

[24] R. K. E. Bellamy et al., "AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," arXiv preprint arXiv:1810.01943, 2018.