Explainable Artificial Intelligence in Financial Decision-Making: A Scoping Review

Fateenah Farid[1]
College of Information Sciences and Technology, The Pennsylvania State University, Pennsylvania, USA
[1]njn5346@psu.edu

I. INTRODUCTION

In the past decade, the increasing adoption of machine learning (ML) models in financial services have transformed how financial high-stakes decisions including loan approvals and credit risk assessments are made. Financial institutions now rely on complex predictive ML models to evaluate borrower risk and automate approval decisions since these models often yield high predictive accuracy. However, oftentimes, these models operate as black boxes which makes it difficult for the financial institutions and its stakeholders to understand, justify, and evaluate how the model performs its individual decisions. This lack of transparency poses significant threats within financial institutions, where decision explainability is essential to ensure regulatory compliance and fairness and, most importantly, provide explainability and transparency with its customers.

Explainability in financial decision-making is more about legal and ethical concern rather than its technicalities. The Fair Credit Reporting Act requires its lenders to provide a credit decision justification capability as part of its regulatory requirement [1]. Without explainable artificial intelligence (XAI) that makes these models interpretable, financial institutions risk regulatory non-compliance and biased decision outcomes which could negatively impact its customer trust.

This paper conducts a scoping review of XAI methods in financial decision-making with a focus on loan approval and credit risk scoring. This scoping review analyzes existing research to examine the usage of explainability techniques that are applied in the real-world, specifically SHAP and LIME. This scoping review is guided by the following research questions:

- **RQ1:** How are explainable AI techniques applied to financial decision-making tasks such as loan approval and credit risk assessment?
- **RQ2:** How are explainability methods such as SHAP and LIME evaluated in financial contexts?
- **RQ3:** Are there any trade-offs between predictive performance, interpretability, and fairness are reported in existing studies? If so, what are they?

II. BACKGROUND

Machine learning has become a core component of modern financial systems, particularly in credit scoring and loan default prediction where automated approval processes are prevalent. Traditional approaches such as logistic regression and rule-based credit scoring models offer transparency but lack the flexibility to capture nonlinear relationships that exist in high-dimensional financial data. As a result, ensemble methods such as XGBoost and Random Forests and neural networks have been increasingly adopted in financial systems due to their superior performance.

However, despite the superior performance of ensemble methods and neural networks, these complex ML models create opacity which is a critical concern in finance. This is due to the fact that decisions that are related to credit approval have a direct impact on individuals' financial abilities and therefore, must be a transparent, auditable, and justifiable process. Hence, the development of XAI enables the financial industry to address this challenge by providing mechanisms to interpret model behavior without putting its predictive accuracy at significant risk.

XAI methods can be broadly classified into intrinsically interpretable models and post-hoc explanation techniques. In financial applications, post-hoc methods are especially common because they allow institutions to retain the use of black-box models that are often high performing while adding interpretability layers. SHAP (Shapley Additive

Explanations) provides both global and local feature attributions while LIME (Local Interpretable Model-agnostic Explanations) generates local surrogate models to explain individual predictions.

SHAP and LIME differ in how they approximate and attribute model behavior. SHAP assigns a contribution value to each feature based on its impact on the model's prediction across all possible feature combinations. This ensures consistency and local accuracy while providing global explanations of overall feature importance and local explanations for individual decisions [2]. In contrast, LIME explains individual predictions by perturbing the input data around a specific instance and uses the resulting predictions to fit a simple, interpretable surrogate model that captures how the original classifier behaves locally. The explanation is then derived from the surrogate model's feature weights which indicate the relationship between each feature on the prediction for that specific case. LIME offers intuitive, local-level explanations and is computationally efficient, hence why its outputs can be sensitive to sampling variability and lack the theoretical guarantees provided by SHAP [3].

These two explanation techniques have become standard tools for explaining financial ML systems.

III. RELATED WORK

This section focuses on three representative studies that are highly relevant to explainable AI in financial decision-making: explainable loan approval systems, explainable loan default prediction, and explainable credit risk scoring frameworks. These studies are closely aligned with the scope of this scoping review as they investigate how post-hoc explainability methods such as SHAP and LIME are integrated into ML models to achieve transparency, fairness, and regulatory compliance in lending systems.

Hartley and Kevin (2025) present a comprehensive XAI framework for automated loan approval in FinTech platforms that integrates predictive modeling, post-hoc explanation techniques, and a decision-facing interface. Their system combines gradient-boosted decision trees and deep neural networks with SHAP and LIME to provide both global and local-level explanations of credit decisions. The framework is designed to support transparency for multiple stakeholders which includes risk or financial analysts as well as its end users through a dashboard-based explanation interface. A key contribution of this work is the incorporation of fairness-aware training into the loan approval process. Hartley and Kevin introduce a regularized loss function that balances predictive accuracy with fairness constraints such as demographic parity in order to mitigate bias in credit decisions. The framework is evaluated using both public and proprietary credit datasets. As a result, the framework demonstrates that fairness regularization can significantly reduce discriminatory outcomes while maintaining strong predictive performance. The study also evaluates explainability and computational efficiency. SHAP is shown to provide stable and globally consistent feature attributions, while LIME Offers faster and case-specific explanations that are more suitable for real-time decision making. The authors highlight trade-offs between interpretability and performance to emphasize that explainable method selection should depend on operational constraints.

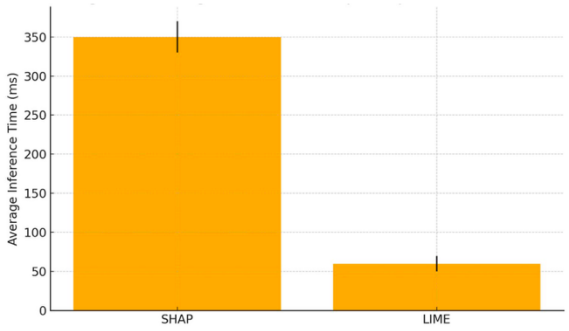

Figure 1. Average Inference Time in milliseconds per Explainable Method

Note. Reprinted from Hartley and Kevin (2025).

Overall, this paper demonstrates how explainable AI can be practically deployed in real-world loan approval systems to enhance transparency, regulatory compliance, and trust in automated financial decision-making.

Taneja (2021) explores uses of explainable machine learning for loan default prediction using pre-2021 LendingClub's peer-to-peer lending dataset. This dataset contains borrower attributes such as loan terms and credit history which are prevalent indicators to credit risk modeling. The paper frames default prediction as a binary classification problem. The paper performs standard preprocessing practices by imputing missing values, encoding categorical variables, and scaling numerical features. The paper compares several predictive models: Random Forest, XGBoost, and Neural Networks. The study reports that higher capacity models achieve stronger discrimination performance and that this is insufficient for deployment in real-life financial contexts, hence the paper applies post-hoc explanation methods, SHAP and LIME, to

bridge this gap. These two explanation methods provide an understanding of how the model characterizes features such as interest rate, debt-to-income ratio, loan amount, income, and employment history that contribute to default loan prediction. SHAP is presented as supporting both global and local-level interpretability, while LIME is used to generate localized explanations for individual borrower cases that may be easier to communicate in operational financial settings. The main takeaway of this paper is the trade-off between performance and interpretability across model families. The paper argues that although neural networks and gradient-boosted models can improve predictive performance, they require explanation layers to provide sufficient transparency that are expected in regulated domains, such as financial contexts. On the other hand, intrinsically interpretable models such as decision trees offer better clarity but risk the complex lending data to underfit. To sum up, this paper positions SHAP and LIME as practical tools for increasing transparency in credit risk modeling and at the same time highlights the fact that model explainability should be treated as a deployment requirement rather than an optional add-on.

Renner (2025) discusses XAI approaches for credit risk assessment in U.S. peer-to-peer lending platforms. The paper builds a framework using the publicly available LendingClub dataset which contains over one million peer-to-peer lending records that spans various borrower and loan types. This study formulates credit risk prediction as a binary classification task by grouping loan outcomes into default and on-default categories to address the class imbalance that exists in lending data. The framework uses Extreme Gradient Boosting (XGBoost) as its primary predictive model as XGBoost is known for its strong performance on structured financial data. The study performs synthetic minority oversampling (SMOTE) and class-weighted loss functions to mitigate the class imbalance which resulted in XGBoost outperforming logistic regression and random forest baselines. In the context of model interpretability, the study integrates SHAP and LIME as its post-hoc explainability methods. SHAP is used to generate both global feature importance rankings and local explanations for individual loan predictions and LIME serves as a comparative baseline for local-level interpretability. Results show that SHAP produces more stable and consistent explanations with higher transparency to the underlying model compared to LIME. Beyond interpretability and predictive performance, the study performs subgroup analyses across borrower segments: income level, employment status, and home ownership. These analyses reveal that feature importance vary across segments, however factors such as employment length and home ownership heavily influence self-employed borrowers.

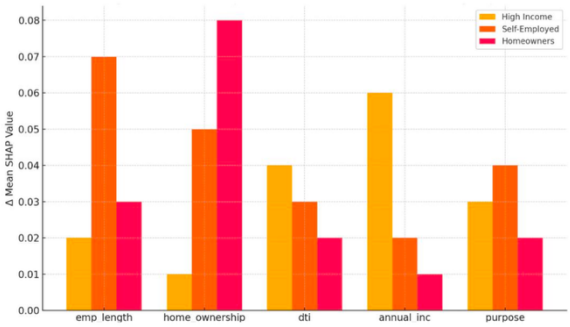

Figure 2. Feature Impact Variation Across Borrower Subgroups

Note. Reprinted from Renner (2025).

The study also discusses explainability within a fairness and regulatory framework. The subgroup analyses reveal modest disparities in model behavior across income and employment categories and to quantify these effects, the author evaluates group fairness using metrics such as equal opportunity and predictive parity. Small but consistent differences were observed across categories that require monitoring in regulated lending environments. Overall, the paper illustrates how SHAP-based explanations can support regulatory compliance in real-world financial markets as it provides concrete and model-derived justifications for adverse credit decisions.

IV. METHODOLOGY

This paper conducts a scoping review to systematically identify, screen, and synthesize existing research on the use of XAI in financial decision-making with a focus on credit scoring and loan approval.

All literature searches were conducted using Google Scholar, an academic database. Searches were performed using combinations of keywords related to XAI, financial decision-making, loan applications. The search strategy was iteratively refined to balance coverage and specificity. The search strings are summarized in Table 1.

Search Strings	No. of Results
("explainable AI" OR "XAI" OR "interpretable machine learning") AND ("financial decision making")	1,400
("explainable AI" OR "XAI") AND ("credit risk" OR "credit scoring") AND ("loan approval")	1,200

("explainable AI" OR "XAI") AND ("credit scoring") AND ("loan approval") AND ("SHAP" OR "LIME")	572
After filtering ("Since 2021")	537

Table 1. Search Strings and Number of Retrieved Results

Results were filtered to include only publications from 2021 onward to focus on recent and relevant developments. Following the search process, about 13 papers was selected for brief analysis based on their relevance, methodological clarity, and representation of XAI practices in financial decision-making. Particular attention was placed on studies that applied post-hoc explainability methods such as SHAP and LIME to real-world credit datasets. Ultimately, three papers were selected as representative examples of the application of explainable AI in credit scoring and loan approvals:

- Explainable AI for loan approval decisions in FinTech platforms [4]
- Explainable machine learning for loan default prediction [5]
- Explainable AI for credit risk scoring on loan platforms [6]

Across these studies, relevant information such as the financial task addressed, machine learning models used, explainability techniques used, evaluation criteria, reported trade-offs, and key findings were extracted to identify recurring themes, methodological patterns, and differences across studies. Overall, the analyses of these paper highlights how explainability is operationalized in real-world financial context, which explainable techniques are preferred for specific tasks, and what limitations remain in its operational financial context.

V. REFERENCES

1. Federal Trade Commission. (n.d.). Fair Credit Reporting Act. https://www.ftc.gov/legal-library/browse/statutes/fair-credit-reporting-act
2. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*. Retrieved from https://arxiv.org/pdf/1705.07874
3. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should I Trust You?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144). Retrieved from https://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf
4. Hartley, E., & Kevin, L. (2025). Explainable AI for Loan Approval Decisions in FinTech Platforms. *Journal of Computer Science and Software Applications*, *5*(6). Retrieved from https://www.mfacademia.org/index.php/jcssa/article/view/231
5. Taneja, A. (2021). Explainable Machine Learning for Loan Default Prediction: Enhancing Transparency in Banking. *International Journal of AI, BigData, Computational and Management Studies, 2*(1), 57-65. Retrieved from https://doi.org/10.63282/3050-9416.IJAIBDCMS-V2I1P106
6. Renner, T. (2025). Explainable AI for Credit Risk Scoring on Loan Platforms. *Transactions on Computational and Scientific Methods, 5*(6). Retrieved from https://doi.org/10.5281/zenodo.15703838