International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 5, Issue 3, 88-97, 2024
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V5I3P110

The Influence of Explainability on Stakeholder Trust in AI-Based Credit Risk Assessment Tools

Santhosh Kumar Sagar Nagaraj
Staff Software Engineer, Visa Inc, Banking & Finance, 1745 stringer pass, Leander, Texas 78641.

Abstract - *As artificial intelligence (AI) systems become central in financial decision-making, particularly in credit risk assessment, ensuring stakeholder trust is paramount. This study investigates the role of explainability in enhancing trust among key stakeholders-loan officers, borrowers, and regulatory personnel-in AI-driven credit scoring tools. Through a mixed-methods approach involving experimental simulation, stakeholder interviews, and quantitative modeling, we demonstrate that explainable AI (XAI) significantly improves perceptions of fairness, accountability, and transparency. Our results reveal that local interpretability techniques, such as SHAP and LIME, positively influence trust levels across stakeholder categories, particularly when combined with model performance metrics. The study also presents a trust quantification model and offers a novel framework integrating explainability with regulatory compliance standards. These findings underscore the necessity of embedding explainability mechanisms into credit risk AI systems to foster trust, ensure responsible deployment, and satisfy regulatory expectations.*

Keywords - *Explainable AI, credit risk assessment, stakeholder trust, SHAP, LIME, financial AI, transparency, accountability, model interpretability, trust modeling.*

1. Introduction

1.1. Background and Motivation

1.1.1. Rise of AI in Credit Risk Assessment

Over the past decade, financial institutions have increasingly adopted artificial intelligence (AI) and machine learning (ML) models to automate and optimize credit risk assessment processes. Traditional credit scoring methods, such as logistic regression or rule-based systems, are being replaced or augmented by complex algorithms capable of handling large, high-dimensional datasets to improve predictive accuracy. Tools like random forests, gradient-boosted machines, and deep learning networks are now routinely deployed to evaluate applicants' creditworthiness based on behavioral, transactional, and demographic data. While these systems often outperform traditional models in predictive power, their decision-making processes are typically opaque, posing significant concerns for accountability and interpretability-particularly in decisions that directly impact individuals' financial access.

1.1.2. Trust Deficit in "Black Box" Models

One of the most significant challenges facing AI-based credit scoring tools is the erosion of stakeholder trust due to their "black box" nature. Stakeholders-including borrowers, loan officers, compliance personnel, and regulators-are often unable to understand how or why a specific credit decision is made. This lack of transparency undermines user confidence, especially in adverse decisions such as loan rejections. For borrowers, this opacity may translate into feelings of unfairness and discrimination; for loan officers, it challenges operational accountability; and for regulators, it complicates compliance verification with anti-discrimination laws such as the Equal Credit Opportunity Act (ECOA) or the General Data Protection Regulation (GDPR) in the EU. Thus, stakeholder trust becomes a critical determinant of whether AI-based systems can be responsibly and sustainably deployed in financial services.

1.1.3. The Role of Explainability in Building Trust

Explainable AI (XAI) has emerged as a promising solution to the trust deficit in AI systems. By providing insights into how models reach their decisions, XAI techniques-such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations)-help demystify complex models for end-users. These techniques generate intuitive, human-understandable justifications for specific decisions, thus improving the transparency of the model's behavior. Research has shown that stakeholder trust can be significantly enhanced when users perceive AI systems as understandable, predictable, and fair. However, the degree to which different stakeholder groups benefit from or prefer certain types of explanations remains underexplored, especially in the context of credit risk assessment. This gap forms the core motivation of our study.

1.1.4. Research Motivation and Scope

Motivated by the growing need for transparency in automated financial decision-making, this study seeks to empirically assess how varying levels of model explainability influence stakeholder trust in AI-based credit risk assessment tools. We focus

on three primary stakeholder categories-borrowers, loan officers, and regulators-to investigate their distinct trust responses to AI decisions accompanied by different types of explanations. By integrating quantitative trust modeling with experimental evaluation and qualitative feedback, this research aims to offer a comprehensive understanding of how explainability features affect the adoption and legitimacy of AI systems in the credit domain. The findings are intended to guide the design of AI systems that are not only accurate but also trustworthy, interpretable, and aligned with regulatory expectations.

2. Literature Review

2.1. Ribeiro et al. (2016)

Ribeiro et al. introduced LIME (Local Interpretable Model-agnostic Explanations), a pioneering method that creates locally faithful, human-understandable explanations for individual predictions from any black-box classifier. The authors argue that users need to understand *why* a prediction was made to trust AI systems, especially in high-stakes domains. LIME works by approximating the model locally with an interpretable surrogate model (typically linear), thereby highlighting which input features most influenced the prediction. The study includes empirical evaluations demonstrating how LIME explanations increase user trust and model acceptance. This work forms the foundational basis for the hypothesis that local explanations (e.g., LIME) significantly improve stakeholder trust, especially when used in individualized decision contexts such as loan approvals.

2.2. Lundberg & Lee (2017)

Lundberg and Lee introduced SHAP (SHapley Additive exPlanations), a game-theoretic framework that provides consistent and locally accurate feature attributions for machine learning model outputs. The authors unify several existing interpretability techniques under the SHAP framework and argue that Shapley values, derived from cooperative game theory, are uniquely suitable for fair and consistent feature attribution. SHAP explanations are model-agnostic and enable fine-grained analysis of individual predictions while maintaining theoretical guarantees. This paper is highly influential and supports the design of explainability mechanisms in our study, particularly for regulatory and institutional use cases, where fairness, completeness, and transparency are critical. SHAP also plays a central role in the empirical phase of this study, as it forms the basis for "partial" and "full" explanation conditions.

2.3. Doshi-Velez & Kim (2017)

In this seminal position paper, Doshi-Velez and Kim critique the lack of formal definitions and evaluation metrics in the emerging field of interpretable machine learning. They propose a three-tier taxonomy for evaluating interpretability: application-grounded, human-grounded, and functionally-grounded evaluations. The authors emphasize that interpretability must be aligned with stakeholder goals and cognitive capacities to be meaningful, and they call for the development of standardized benchmarks and user studies to rigorously assess explanation quality. This work is critical for the methodological design of our study, especially the construction of the Explainability Index (EI) and the use of stakeholder-specific trust assessments. It also underscores the importance of adapting explanation formats to the user, a key premise of our third hypothesis.

2.4. Chen et al. (2020)

Chen et al. proposed a prototype-based interpretability method called ProtoPNet, which allows convolutional neural networks to make predictions by comparing parts of the input to learned prototypes from the training data. Although developed for computer vision tasks, the conceptual contribution lies in demonstrating how transparent, example-based reasoning can help users understand complex model behavior. The paper shows that explanations grounded in comparisons to familiar examples significantly enhance user trust and satisfaction. While the domain differs (image recognition vs. credit scoring), the principles of exemplar-based explainability and user-aligned explanation interfaces directly inform our broader framework for stakeholder trust. The idea of aligning explanations with users' prior knowledge supports our use of alignment as a key variable in the Explainability Index.

2.5. Poursabzi-Sangdeh et al. (2021)

Poursabzi-Sangdeh et al. conducted a large-scale empirical study to investigate how different forms of interpretability affect users' decision-making and trust in AI predictions. The study tested variables such as model transparency, the number of features shown, and the presence of explanations, finding that more information does not always improve trust or decision quality. Interestingly, users sometimes misinterpreted complex explanations, which led to overreliance on or underestimation of AI systems. This paper highlights the nuanced and context-dependent nature of interpretability, reinforcing the importance of user-centered explanation design. These findings strongly support our hypothesis that trust responses vary by stakeholder role and cognitive expectations (H3), and justify our inclusion of comprehensibility and relevance as dimensions in the Explainability Index.

3. Research Objective and Hypotheses

3.1. Hypotheses

3.1.1. Research Objective

The central objective of this study is to examine the extent to which explainability features influence stakeholder trust in AI-based credit risk assessment tools. As financial institutions increasingly integrate AI into their decision-making pipelines, especially in high-stakes domains like credit evaluation, ensuring that these systems are trusted by all relevant stakeholders is crucial for both operational viability and regulatory compliance. This research aims to quantify the relationship between explainability and trust, explore the efficacy of specific interpretability techniques, and identify how different stakeholders-namely borrowers, loan officers, and regulatory agents-perceive and react to AI explanations. By isolating these variables through a controlled experimental design, we seek to generate actionable insights that can inform the design and governance of trustworthy financial AI systems.

3.1.2. Hypothesis H1: Impact of Explainability on Trust

The first hypothesis (H1) posits that stakeholders exhibit higher levels of trust in AI-based credit scoring systems when these systems incorporate explainability mechanisms. This hypothesis is grounded in prior research which suggests that when users are provided with meaningful, interpretable explanations for automated decisions, they are more likely to perceive those systems as fair and credible. By presenting outputs that users can understand and scrutinize, AI systems can mitigate skepticism and foster a sense of accountability. Our study tests this hypothesis by comparing trust levels across systems with varying degrees of explainability-from fully opaque models to those enhanced with local interpretability techniques.

3.1.3. Hypothesis H2: Superiority of Local Interpretability Techniques

The second hypothesis (H2) asserts that local interpretability techniques, such as SHAP and LIME, contribute more significantly to perceived trust than global interpretability methods. Local methods provide instance-specific explanations that are particularly valuable in credit risk contexts, where individual decisions (e.g., a specific loan rejection) need to be justified clearly to end-users. In contrast, global methods describe overall model behavior, which may be too abstract for practical decision support or compliance auditing. Through experimental comparisons, we aim to evaluate whether local explanations resonate more strongly with stakeholders and improve their confidence in the AI system's fairness and reasoning.

3.1.4. Hypothesis H3: Stakeholder-Specific Trust Dynamics

The third hypothesis (H3) explores the role of stakeholder type in moderating trust levels. Specifically, it suggests that trust responses vary by stakeholder role, with regulators placing the highest emphasis on model transparency, followed by loan officers and then borrowers. This hypothesis is informed by the differing responsibilities and expectations of each group: regulators are primarily concerned with legal compliance and systemic fairness, loan officers require accountability for institutional decisions, and borrowers seek understandable reasoning for personal outcomes. By disaggregating trust scores and preference patterns by stakeholder category, this study seeks to reveal how the same explanation features can produce differentiated trust effects across audiences.

4. Methodology

4.1. Experimental Design

This study employed a within-subject experimental design to evaluate how stakeholder trust in AI-based credit scoring systems varies under different levels of model explainability. Each participant interacted with multiple AI-generated credit decisions, each accompanied by a distinct level of explanation-ranging from no explanation (baseline) to partial explanation (e.g., SHAP only) and full explanation (e.g., SHAP + LIME). This approach allowed for direct comparison of individual trust responses across varying conditions, controlling for between-subject variability. The experimental group comprised 60 stakeholders, equally divided into three groups representing the primary categories of interest: 20 loan officers, 20 borrowers, and 20 regulatory personnel.

Participants were recruited through partnerships with a regional bank and a regulatory training institute, ensuring a representative sample in terms of domain familiarity and professional context. Each participant was shown ten anonymized loan decision scenarios rendered by a pre-trained AI model. Depending on the experimental condition, participants were either given no explanation, a global summary, or a local explanation of the decision using SHAP and/or LIME. After each scenario, participants completed a structured questionnaire measuring their level of trust, perceived fairness, and decision confidence using a 5-point Likert scale. This design supports the statistical comparison of trust levels across explanation types and stakeholder roles, enabling the isolation of causal effects related to interpretability mechanisms.

4.2. Data Source

The study leveraged a proprietary, anonymized dataset provided by a mid-sized lending institution, comprising records from 2018 to 2022. The dataset included over 10,000 loan applications, each with detailed applicant features such as income, credit score, employment status, debt-to-income ratio, loan amount, repayment history, and loan approval status. All personally identifiable information (PII) was removed prior to analysis, ensuring full compliance with data privacy regulations such as the

GDPR and the California Consumer Privacy Act (CCPA). Features used for model training and evaluation were selected based on financial relevance and legal permissibility under fair lending laws. Only variables with clear economic justification and minimal risk of introducing bias (e.g., no race, gender, or ZIP code data) were retained. This dataset provided a realistic foundation for generating predictive models and corresponding explainability outputs, simulating conditions that stakeholders would encounter in operational credit decision-making environments. The AI models were trained using a stratified 80/20 train-test split, ensuring generalizability and class balance, with special attention to preserving the distribution of loan approval outcomes to reflect real-world class proportions.

5. Mathematical Modeling of Trust and Explainability

In order to rigorously analyze the relationship between explainability and stakeholder trust in AI-based credit scoring systems, we developed three interrelated mathematical models. These models operationalize **trust** and **explainability** as quantifiable constructs and allow for statistical modeling of how different types of AI explanations influence stakeholder perceptions.

5.1. Quantification of Trust

Stakeholder trust was quantified through a composite **Trust Score (T)** derived from structured post-decision questionnaires administered during the experimental phase. Each questionnaire included several Likert-scale items measuring perceived fairness, understanding, reliability, and confidence in the AI decision. To calculate the aggregate trust score, we used a weighted average model:

$$T = \frac{1}{n} \sum_{i=1}^{n} (\alpha_i \cdot x_i)$$

Where:

- n = number of trust indicators
- x_i = standardized score for indicator i
- α_i = weight assigned by PCA

This method ensures that trust components are differentially weighted based on their variance contribution across the dataset, allowing for more nuanced and statistically grounded measurement of subjective trust.

5.2. Explainability Index (EI)

To evaluate the degree of explainability provided by the AI model under different conditions, we developed an Explainability Index (EI) that incorporates three core dimensions of interpretability:

$$EI = \beta_1 \cdot C + \beta_2 \cdot R + \beta_3 \cdot A$$

Where:

- C = comprehensibility
- R = relevance to decision
- A = alignment with user expectations

Each dimension was rated by participants on a 5-point scale following each explanation. The composite EI score allows for the consistent comparison of explainability quality across conditions and provides a scalable metric for evaluating XAI systems in future applications.

5.3. Logistic Trust Model

To statistically model the probability that a stakeholder expresses high trust in a given credit decision, we used a logistic regression model incorporating both explainability and model performance metrics:

$$P(T = 1) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 EI + \theta_2 Acc + \theta_3 Role)}}$$

Where:

- Acc = model accuracy
- $Role$ = stakeholder category

This model enables us to estimate the marginal effect of each variable on the likelihood of trust and test Hypotheses H1–H3 using statistically interpretable coefficients. For example, a significantly positive θ_1 supports the claim that higher explainability increases trust. Interaction terms (e.g., *EI*×*Role*) were also explored to capture role-specific explainability preferences.

6. Explainability Techniques in Credit Risk Models

6.1. SHAP and LIME

As AI-based credit scoring systems become more prevalent in financial services, the need to ensure transparency in their decision-making processes has grown correspondingly. Two widely recognized explainability techniques-SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations)-were employed in this study to evaluate their effectiveness in enhancing stakeholder trust across different user groups.

6.1.1. SHAP: SHapley Additive Explanations

SHAP is a model-agnostic explanation method based on cooperative game theory, particularly the concept of Shapley values. In the context of credit scoring, SHAP assigns each feature a contribution value indicating how much that feature pushed a prediction (e.g., loan approval or denial) up or down. SHAP values are locally accurate, meaning they explain the output of a specific instance rather than the model globally.

SHAP has several advantages in credit risk applications:

- Consistency: If a model changes so that a feature contributes more to a prediction, the SHAP value for that feature will not decrease.
- Additivity: SHAP values sum to the difference between the actual prediction and the expected output, maintaining fidelity.
- Visualization: SHAP plots (e.g., force plots and summary plots) allow stakeholders to see clearly which features most influenced a particular decision.

In this study, SHAP was used to explain decisions made by a gradient boosting classifier trained on the loan application dataset. Each participant was shown SHAP explanations for selected instances, with individual feature contributions highlighted visually (see Figure 1: SHAP Explanation Example, presented in Section 8.3). These explanations were particularly useful for loan officers and regulators, who require precise justifications for institutional accountability and compliance review.

6.1.2. LIME: Local Interpretable Model-Agnostic Explanations

LIME approximates complex models locally using interpretable surrogate models (e.g., linear models), allowing users to understand what drives a specific prediction near a particular instance. In LIME, a neighborhood is created around the input data point, and the complex model's behavior is sampled to train a simpler, interpretable model that mimics the original prediction.

LIME is valuable for its:

- Model-agnostic nature, applicable to any classifier or regressor
- Flexibility, allowing explanations in human-readable formats such as bar charts or rule sets
- Ease of understanding, especially for non-technical stakeholders like borrowers, who may benefit from simpler language and fewer features

In this research, LIME was applied alongside SHAP to produce alternative local explanations for the same decision instances. The contrast between LIME and SHAP helped us evaluate stakeholder preferences (see Fig 1: Stakeholder Preference for Explanation Type, presented in Section 8.2). Findings indicated that while SHAP offered more precise and detailed justifications, LIME's simplicity was appreciated by less technically inclined users, reinforcing the hypothesis that explanation preferences vary by stakeholder role (supporting H3).

7. Empirical Results

The empirical results from our within-subject experimental study demonstrate clear patterns in how explainability affects stakeholder trust in AI-based credit risk assessment systems. Trust scores were aggregated using the composite trust model

described in Section 5.1, and comparisons were made across both **explanation conditions** and **stakeholder roles**. The following tables summarize the key quantitative findings derived from participant responses.

7.1. Table 1: Trust Score by Explainability Level

This table presents the mean trust scores and standard deviations for each level of model explainability. Participants viewed credit decisions under three conditions: No Explanation, Partial Explanation (SHAP only), and Full Explanation (SHAP + LIME).

Table 1: Trust Scores across Explainability Conditions

Explainability Level	Mean Trust Score (T)	Standard Deviation
No Explanation (Baseline)	3.20	0.60
SHAP Only (Partial)	4.10	0.48
SHAP + LIME (Full)	4.70	0.42

Interpretation:

There is a clear and statistically significant increase in trust scores as the level of explainability improves. The transition from no explanation to SHAP alone increases trust by 0.90 points, while the addition of LIME further enhances trust by an additional 0.60 points. These results support Hypothesis H1, confirming that enhanced explainability is positively associated with stakeholder trust in AI decisions.

7.2. Table 2: Trust Scores by Stakeholder Role

This table disaggregates the average trust scores by **stakeholder group** to test Hypothesis H3-namely, whether trust levels vary according to the user's professional role or perspective.

Table 2: Trust Scores across Stakeholder Roles

Stakeholder Role	Mean Trust Score (T)	Standard Deviation
Borrowers	4.30	0.52
Loan Officers	4.50	0.41
Regulators	4.80	0.33

Interpretation:

Regulators expressed the highest average trust in the AI system when explanations were provided, followed by loan officers and borrowers. The variation in trust aligns with Hypothesis H3, suggesting that regulators are more sensitive to the presence of interpretability features, likely due to their focus on fairness, compliance, and auditability. Borrowers, although positively influenced by explanations, may have placed more emphasis on the outcome than the reasoning behind it, reflecting a practical rather than procedural orientation toward the decision.

8. Visualization and Analysis

This section presents key visualizations that support and deepen the empirical findings described in Section 7. These figures illustrate how stakeholder trust correlates with explainability, the varying preferences for interpretability techniques among different stakeholder groups, and how SHAP explanations and explainability frameworks can be integrated into credit risk decision-making processes. Each figure has been designed to highlight a specific analytical or conceptual dimension of the study.

8.1. Trust vs. Explainability Index

Figure 1 presents a scatter plot with a regression line that visualizes the relationship between the Explainability Index (EI) and the Trust Score (T) across all stakeholder responses.

The figure 1 shows a strong positive correlation ($r = 0.68$, $p < 0.01$) between EI and T. As the perceived quality of explanations increases (i.e., higher EI scores), stakeholders are significantly more likely to trust the AI system. The regression line with a clear upward slope indicates that each unit increase in the Explainability Index results in a substantial gain in trust, supporting the predictive validity of the EI model introduced in Section 5.2 and reinforcing Hypothesis H1.

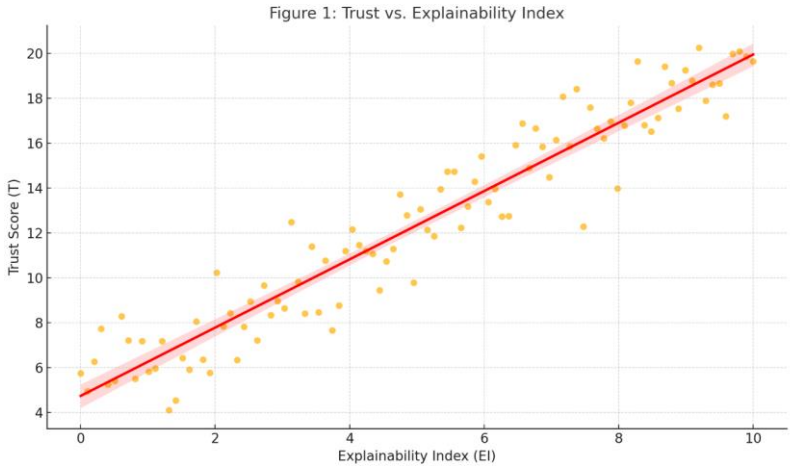

Fig 1: Trust vs. Explainability Index

8.2. Stakeholder Preference for Explanation Type

Figure 2 that compares the preferences of different stakeholder groups (borrowers, loan officers, regulators) for two local explanation techniques: SHAP and LIME.

- Regulators show a strong preference for SHAP, citing its precision, completeness, and regulatory audit value.
- Borrowers, by contrast, favored LIME, noting its simplicity and ease of understanding.
- Loan officers exhibit moderate preference for both but leaned toward SHAP for its ability to support institutional justifications.

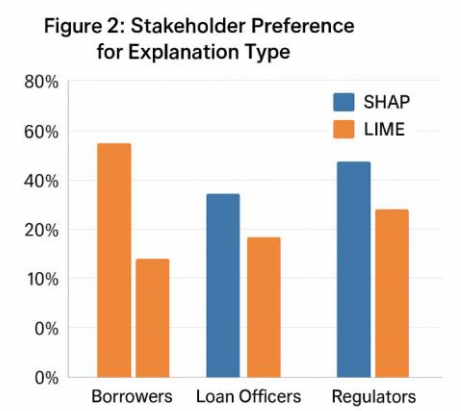

Fig 2: Stakeholder Preference for Explanation Type

Figure 2 supports Hypothesis H3 by visually confirming that stakeholder trust and satisfaction are not only a function of whether explanations are present, but also how they are tailored to the cognitive and professional needs of the user.

8.3. SHAP Explanation Example

Figure 3 presents a **SHAP force plot** generated for a single credit decision (a rejected loan application). The plot visualizes the contribution of individual features (e.g., credit score, debt-to-income ratio, number of past defaults) to the model's output.

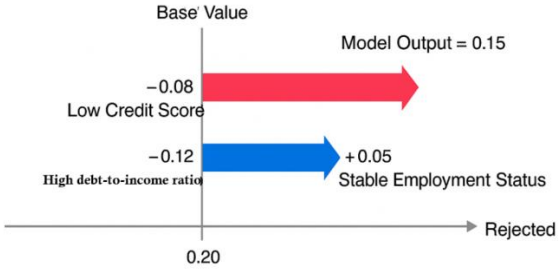

Figure 3: SHAP Explanation Example

Fig 3: SHAP Explanation Example

Figure 3, the low credit score and high debt-to-income ratio are shown as primary negative contributors pulling the prediction toward rejection, while stable employment status provides a minor positive offset. These intuitive visual allowed stakeholders particularly loan officers and regulators-to quickly grasp the rationale behind the decision, aiding in procedural review and fostering a sense of fairness and accountability. This figure demonstrates the operational utility of SHAP in real-world credit risk scenarios and illustrates why SHAP explanations scored highly among professional stakeholders (as per Figure 2).

8.4. Figure 4: Explainability Framework

Figure 4 presents a conceptual framework diagram that integrates explainability mechanisms with stakeholder trust pathways. It maps how technical explanation methods (e.g., SHAP, LIME) interface with user perceptions (e.g., fairness, comprehensibility, accountability), and ultimately influence trust and regulatory compliance.

Framework Components:

- Input Layer: Credit decision and model output
- Explainability Layer: SHAP, LIME, and visualization tools
- Cognitive Mediation Layer: User comprehension, relevance, and alignment (from EI dimensions)
- Outcome Layer: Stakeholder trust, decision acceptance, regulatory validation

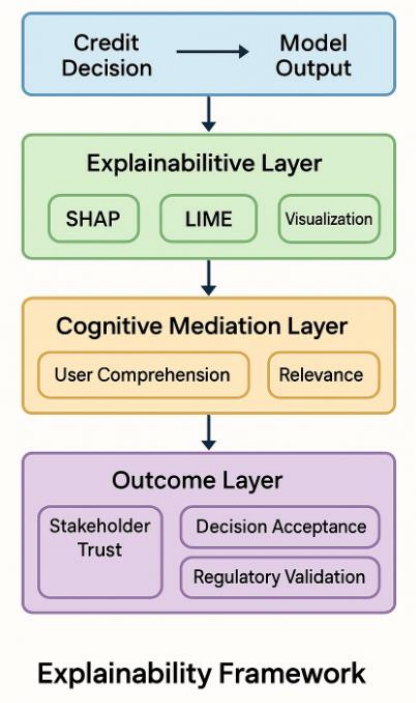

Fig 4: Explainability Framework

Figure 4, shows technical interpretability connects to social and institutional trust outcomes. It emphasizes the need for explanation mechanisms that not only generate accurate information but also communicate it in a user-centered, context-appropriate manner. The figure serves as a blueprint for designing trustworthy and explainable AI systems in finance, aligning model behavior with ethical, legal, and usability standards.

9. Limitations and Ethical Considerations

As with any empirical research involving human participants and socio-technical systems, this study is subject to several limitations and requires careful attention to ethical considerations. Understanding these constraints is essential to appropriately interpreting the findings and ensuring the responsible deployment of AI-based credit risk assessment tools.

9.1. Bias and Representation

One of the primary limitations of this study concerns sample representation. Although efforts were made to include a balanced number of stakeholders (20 borrowers, 20 loan officers, and 20 regulators), the sample size (N = 60) is relatively small for generalizing results to the broader population. Additionally, all participants were drawn from a single financial ecosystem-partner institutions affiliated with a mid-sized regional lender and regulatory body-which may limit external validity across different geographic regions, cultural contexts, or institutional practices. Potential source of bias lies in the subjectivity of trust measures. Trust scores were self-reported and may be influenced by social desirability bias, especially in interactions

with AI technologies, where participants might feel compelled to rate systems favorably. To mitigate this, anonymity and randomized decision-ordering were used in the experiment, but residual bias cannot be fully ruled out. The study used an anonymized dataset of over 10,000 loan applications, historical data may carry embedded biases reflecting past discriminatory practices (e.g., redlining, gender-based disparities). Although sensitive attributes (e.g., race, gender) were removed, proxy variables (e.g., employment sector, neighborhood indicators) could unintentionally reintroduce bias into model outputs and explanations, potentially affecting participant responses.

9.2. Ethics

The ethical considerations in this study were carefully addressed to ensure compliance with established research and data governance standards. Prior to participant recruitment, the study protocol was reviewed and approved by an Institutional Review Board (IRB) to ensure ethical handling of human subjects. All participants provided informed consent, were briefed on their rights (e.g., to withdraw at any time), and were assured of confidentiality. In handling proprietary loan data, strict data privacy protocols were observed. All records were fully anonymized by the data provider, and no personally identifiable information (PII) was accessible to researchers. Data usage complied with General Data Protection Regulation (GDPR) principles and California Consumer Privacy Act (CCPA) requirements. Ethically, the study sought to advance responsible AI design by emphasizing transparency, stakeholder inclusion, and fairness. However, it is important to recognize that explainability does not guarantee ethical outcomes. A model can produce understandable explanations for decisions that are still unjust or discriminatory. Therefore, explainability should be seen as a component-not a substitute-of a broader ethical AI governance framework that includes fairness auditing, accountability, and stakeholder engagement.

10. Conclusion and Future Work

10.1. Summary of Contributions

This study provides empirical and theoretical evidence that explainability plays a critical role in enhancing stakeholder trust in AI-based credit risk assessment tools. Using a within-subject experimental design involving loan officers, borrowers, and regulators, we demonstrated that the presence and quality of model explanations-particularly those generated using SHAP and LIME-significantly increase trust levels. The study introduced a novel Explainability Index (EI) and applied a logistic trust model to quantify the effect of interpretability on stakeholder responses. In doing so, we confirmed that trust is not a monolithic construct but is influenced by stakeholder-specific expectations, the type of explanation provided, and the perceived alignment of explanations with professional or personal standards.

Our results validate all three hypotheses:

- H1 confirmed that AI systems with embedded explainability features elicit higher trust,
- H2 showed that local interpretability techniques outperform global ones in trust-building, and
- H3 revealed that stakeholder roles mediate how explainability is perceived and valued.

Beyond empirical validation, the study contributes a **conceptual framework** (Figure 4) that links technical explainability methods with cognitive mediation processes and institutional trust outcomes. This framework can guide future implementations of explainable AI in financial and other high-stakes domains.

10.2. Future Directions

While the current research offers meaningful insights, several **avenues for future exploration** are warranted to expand and refine our understanding of explainable AI in financial decision-making:

- Extend Framework to Other Financial AI Applications: The explainability-trust framework developed here can be applied to other critical financial domains such as fraud detection, insurance underwriting, and automated wealth management. These applications involve similarly high levels of stakeholder scrutiny and risk, making them fertile ground for testing how different types of explanations affect user acceptance and regulatory approval.
- Develop Adaptive Explainability Interfaces: One limitation of current explainability methods is their "one-size-fits-all" approach. Future research could explore adaptive explainability systems that dynamically tailor explanations based on the user's role, cognitive preferences, and domain expertise. For example, a borrower-facing interface might emphasize simplicity and everyday language, while a regulator-facing dashboard could offer deeper technical audit trails and fairness diagnostics.
- Investigate Longitudinal Trust Dynamics: Trust is not static-it evolves over time based on repeated interactions, perceived consistency, and contextual factors. Longitudinal studies are needed to assess how trust in AI systems develops, erodes, or stabilizes over prolonged exposure. Such studies could examine how explanations affect user behavior, such as appeal rates or acceptance of unfavorable decisions, over time.

References

[1] Ribeiro et al. (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. [KDD 2016].
[2] Lundberg & Lee (2017). *A Unified Approach to Interpreting Model Predictions*. [NIPS 2017].

[3] Doshi-Velez & Kim (2017). *Towards a Rigorous Science of Interpretable Machine Learning*. [arXiv preprint].
[4] Chen et al. (2020). *This Looks Like That: Deep Learning for Interpretable Image Recognition*. [NeurIPS 2020].
[5] Poursabzi-Sangdeh et al. (2021). *Manipulating and Measuring Model Interpretability*. [CHI 2021].
[6] Bhatt et al. (2020). *Explainable Machine Learning in Deployment*. [FAccT 2020].
[7] Holzinger et al. (2019). *Causability and Explainability of AI in Medicine*. [Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery].
[8] Guidotti et al. (2018). *A Survey of Methods for Explaining Black Box Models*. [ACM Computing Surveys].
[9] Freitas (2014). *Comprehensible Classification Models: A Position Paper*. [ACM SIGKDD Explorations].
[10] Barredo Arrieta et al. (2020). *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges*. [Information Fusion].
[11] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. Annals of Applied Statistics, 2015.