International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 06 | June – 2025 DOI: 10.55041/ISJEM04570
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

# TOWARDS TRANSPARENT ARTIFICIAL INTELLIGENCE: A COMPARATIVE STUDY OF EXPLAINABLE AI MODELS FOR DECISION-MAKING IN FINANCIAL RISK ASSESSMENT

Deepak Kumar Patel
Research Scholar
Artificial Intelligence, Kalinga University Naya Raipur

## 1. Introduction

In recent years, the financial industry has witnessed a transformative shift driven by the integration of Artificial Intelligence (AI) into risk management systems. Machine learning algorithms now power a broad spectrum of financial applications, including credit scoring, fraud detection, and loan default prediction. These systems offer unprecedented speed and predictive accuracy, enabling institutions to process complex datasets and uncover risk patterns that traditional statistical methods might miss (Khandani, Kim & Lo, 2010; Liu et al., 2022). However, the rising reliance on AI-based decision-making has simultaneously introduced significant challenges—chief among them is the issue of transparency. Many high-performing AI models, particularly ensemble methods and deep neural networks, function as black-box systems with limited interpretability (Doshi-Velez & Kim, 2017). In domains like finance, where decisions have legal, ethical, and economic consequences, the inability to understand or audit the rationale behind model outputs poses a substantial barrier to trust and regulatory compliance (Samek et al., 2019; Barredo Arrieta et al., 2020). This growing concern has catalyzed research into Explainable Artificial Intelligence (XAI), a subfield of AI aimed at developing methods and tools that render AI decisions understandable to humans without compromising performance (Adadi & Berrada, 2018). The need for explainability is further underscored by global regulatory mandates, such as the General Data Protection Regulation (GDPR), which enshrine the "right to explanation" for automated decisions affecting individuals (Goodman & Flaxman, 2017). Despite the proliferation of XAI methods—ranging from model-agnostic techniques like SHAP (Lundberg & Lee, 2017) and LIME (Ribeiro, Singh & Guestrin, 2016) to inherently interpretable models such as decision trees and rule-based learners—there remains a critical gap in domain-specific evaluations. Current literature often lacks empirical, comparative analysis tailored to the unique constraints and requirements of financial risk assessment, where both interpretability and predictive performance are vital (Chen et al., 2023; Du et al., 2021). This study aims to address that gap by evaluating and comparing several prominent XAI techniques in the context of financial risk decision-making. We investigate how different models balance the trade-offs between transparency, accuracy, and computational efficiency when applied to real-world financial datasets. The primary contributions of this research are twofold: (1) a systematic comparison of selected XAI models applied to financial risk assessment tasks, and (2) practical insights into the strengths, limitations, and suitability of these models for adoption in regulatory-sensitive financial environments.

## 2. Materials and Methods

### 2.1 Data Collection and Preprocessing

This study utilizes publicly available and institutionally vetted financial datasets, including the German Credit Dataset, Give Me Some Credit (Kaggle), and the LendingClub Loan Data, each comprising consumer financial attributes and labeled outcomes such as loan default, credit risk classification, or repayment status.

 

These datasets reflect real-world financial decision-making scenarios and are widely used benchmarks for credit scoring and risk modeling research (Khandani et al., 2010; Liu et al., 2022). The raw datasets underwent preprocessing to ensure suitability for machine learning modeling. This included:

- **Handling Missing Values:** Imputation using median or mode depending on data distribution.
- **Encoding Categorical Features:** One-hot encoding and ordinal encoding were used as appropriate.
- **Feature Engineering:** Domain-specific feature construction (e.g., debt-to-income ratio, credit utilization) was applied to enhance model learning capabilities.
- **Normalization and Scaling:** Continuous variables were standardized using z-score normalization to ensure comparability across models, particularly those sensitive to feature scales like neural networks.

To preserve fairness and mitigate data leakage, preprocessing steps were executed within cross-validation folds, ensuring the test data remained unseen during transformation.

### 2.2 Model Selection

To evaluate the interplay between performance and interpretability, two categories of models were employed:

#### Black-Box Models

These are high-performing predictive models often used in financial risk scoring but inherently lack transparency:

- **Random Forest** – An ensemble of decision trees known for robustness and performance.
- **XGBoost** – A gradient-boosting framework offering high accuracy and resilience to multicollinearity.
- **Multilayer Perceptron (Neural Network)** – A deep learning model capturing non-linear relationships in high-dimensional data.

#### Explainable AI (XAI) Models

To extract interpretability, both model-agnostic and intrinsically interpretable approaches were used:

- **SHAP (SHapley Additive exPlanations)** – Provides feature-level additive explanations grounded in cooperative game theory (Lundberg & Lee, 2017).
- **LIME (Local Interpretable Model-agnostic Explanations)** – Approximates local decision boundaries of black-box models using interpretable surrogates (Ribeiro et al., 2016).
- **RuleFit** – Learns sparse rule-based models from tree ensembles, providing global and local explanations (Friedman & Popescu, 2008).
- **Interpretable Decision Sets (IDS)** – Uses rule-based classifiers that ensure simplicity, non-redundancy, and transparency in decision logic (Lakkaraju et al., 2016).

All models were implemented using Python libraries including scikit-learn, XGBoost, SHAP, and imodels. Hyperparameters were optimized using grid search with 5-fold cross-validation.

### 2.3 Evaluation Criteria

- **Predictive Accuracy**
  - *Metrics:* Area Under the Curve (AUC), F1-score, precision, and recall were computed to evaluate discriminative power on imbalanced datasets.
  - *Goal:* Ensure baseline predictive performance is competitive across models.
- **Explainability**
  - *Fidelity:* The degree to which explanations approximate the original model's behavior.
  - *Interpretability Score:* Human-centered evaluation based on complexity (e.g., number of rules or features involved in the explanation).
  - *Stability:* Consistency of explanations across similar instances or perturbed inputs (Rudin, 2019; Alvarez-Melis & Jaakkola, 2018).
- **Computational Efficiency**
  - *Metrics:* Execution time and memory usage during both inference and explanation generation.
  - *Goal:* Identify models feasible for deployment in time-sensitive, high-volume financial systems.

All evaluations were performed on a standardized computing environment (Intel i7 CPU, 32 GB RAM, Python 3.9) to ensure consistency and reproducibility.

## 3. Experiments and Results

### 3.1 Implementation Details

All experiments were conducted using Python 3.9 on a system equipped with an Intel Core i7 processor, 32 GB RAM, and an NVIDIA RTX 3080 GPU. The following tools and libraries were used:

- **Scikit-learn** for data preprocessing and classical machine learning models
- **XGBoost** for gradient boosting classification
- **TensorFlow/Keras** for building and training the neural network model
- **SHAP**, **LIME**, and **imodels** for explainability techniques
- **Pandas**, **NumPy**, and **Matplotlib/Seaborn** for data handling and visualization

Each model was trained and evaluated using **5-fold stratified cross-validation**. For black-box models (Random Forest, XGBoost, Neural Network), hyperparameter tuning was performed via grid search. All explanation techniques were applied post hoc, except RuleFit and Interpretable Decision Sets, which are intrinsically interpretable.

### 3.2 Comparative Performance Analysis

#### Quantitative Evaluation

The following table summarizes the average performance across all folds, comparing both predictive accuracy and explainability-related metrics:

**Table 1: Model Performance Comparison (German Credit Dataset)**

| Model | AUC-ROC | F1-Score | Fidelity (to original) | Explanation Complexity[1] | Inference Time (ms) |
|---|---|---|---|---|---|
| **Random Forest** | 0.842 | 0.721 | N/A | N/A | 13.2 |
| **+ SHAP** | 0.842 | 0.721 | 0.97 | Medium | 85.6 |
| **+ LIME** | 0.842 | 0.721 | 0.89 | High | 61.7 |
| **XGBoost** | 0.861 | 0.735 | N/A | N/A | 14.9 |
| **+ SHAP** | 0.861 | 0.735 | 0.98 | Medium | 93.5 |
| **+ LIME** | 0.861 | 0.735 | 0.90 | High | 69.2 |
| **Neural Network** | 0.847 | 0.715 | N/A | N/A | 23.5 |
| **+ SHAP** | 0.847 | 0.715 | 0.95 | High | 107.4 |
| **+ LIME** | 0.847 | 0.715 | 0.86 | Very High | 74.1 |
| **RuleFit** | 0.823 | 0.707 | 1.00 | Low | 10.4 |
| **Interpretable Decision Sets** | 0.812 | 0.695 | 1.00 | Low | 8.7 |

[1] *Explanation complexity refers to the number of features, rules, or tokens involved in each explanation. Lower values imply easier interpretability.*

**Observations:**

- **XGBoost** achieved the best predictive performance overall but required post hoc explanation tools.
- **SHAP** generally provided higher fidelity than LIME across all models but incurred greater computational cost.
- **RuleFit** and **Interpretable Decision Sets**, while slightly less accurate, offered transparent and consistent explanations with low complexity and high stability.
- Post hoc methods introduced trade-offs between explanation clarity and computational overhead.

**Qualitative Evaluation**

To assess user-centric interpretability, a **panel of three financial analysts** evaluated model outputs using a Likert scale (1–5) for:

- **Clarity of Explanation**
- **Relevance to Financial Logic**
- **Trust in Decision**

Results averaged over multiple predictions:

**Table 2: Human Evaluation of Explanation Quality**

| Model + XAI Tool | Clarity | Financial Relevance | Trust Score |
|---|---|---|---|
| **Random Forest + SHAP** | 4.2 | 4.4 | 4.5 |
| **XGBoost + SHAP** | 4.5 | 4.6 | 4.7 |
| **Random Forest + LIME** | 3.6 | 3.9 | 3.8 |
| **Neural Net + LIME** | 3.1 | 3.4 | 3.2 |
| **RuleFit** | 4.7 | 4.8 | 4.9 |
| **Interpretable Decision Sets** | 4.6 | 4.7 | 4.8 |

Rule-based models were consistently rated higher for their alignment with financial reasoning and trustworthiness, despite slightly lower predictive metrics.

## 4. Discussion

### 4.1 Trade-offs Between Performance and Transparency

The results underscore the intrinsic tension between predictive performance and interpretability—an ongoing challenge in explainable artificial intelligence (XAI). Black-box models such as XGBoost and neural networks consistently yielded superior accuracy (e.g., AUC > 0.85) across datasets, supporting their continued dominance in high-stakes domains like credit scoring and loan default prediction (Chen & Guestrin, 2016; Bahnsen et al., 2014). However, these gains come at the cost of model transparency, necessitating post hoc explanation tools.

While **SHAP** achieved high fidelity to original predictions and provided feature-level insights grounded in cooperative game theory (Lundberg & Lee, 2017), its computational overhead and explanation complexity limited its real-time applicability in some use cases. **LIME**, though faster, was more unstable and less faithful to model behavior under perturbation, echoing findings from Alvarez-Melis & Jaakkola (2018).

By contrast, **RuleFit** and **Interpretable Decision Sets (IDS)** offered directly interpretable models with modest sacrifices in accuracy (2–5% lower AUC on average). These models provided clear, rule-based rationales for predictions, facilitating ease of audit and alignment with regulatory requirements—highlighting their practicality for institutions prioritizing explainability over marginal accuracy gains (Rudin, 2019).

### 4.2 Model Suitability for Financial Risk Use-Cases

In real-world financial risk assessment, model selection depends not solely on statistical performance but also on contextual factors such as compliance, user trust, and operational efficiency. The results suggest a tiered suitability framework:

- **Compliance-Driven Environments (e.g., loan approvals, regulatory audits):** Interpretable models like RuleFit and IDS are ideal, offering transparency, auditability, and justifiability—critical under legal frameworks such as the EU's GDPR "right to explanation" (Goodman & Flaxman, 2017).

- **Operational Risk Scoring or Internal Credit Monitoring:** Black-box models with post hoc explanations (e.g., XGBoost + SHAP) are better suited when predictive accuracy is paramount, provided sufficient infrastructure exists to support their interpretability pipelines.
- **Decision Support for Analysts:** Hybrid setups that combine high-performing models with reliable explanation layers can enhance analyst decision-making, offering both precision and actionable insights.

The human-centered evaluation further affirmed that trust and usability are not merely by-products of accuracy but are tightly coupled with how explanations align with financial reasoning—a result consistent with contemporary literature on interpretable machine learning in finance (Liu et al., 2022; Ribeiro et al., 2016).

### 4.3 Limitations of the Study

#### Dataset Generalizability

This study employed publicly available datasets such as the German Credit dataset and LendingClub loan data, which, while widely accepted in academic research, may not fully reflect the heterogeneity, feature richness, or domain-specific nuances present in proprietary banking datasets. As such, generalizing results to more complex or region-specific financial products should be done cautiously. Additionally, these datasets may suffer from historical biases, sampling limitations, or outdated financial behaviors, which can affect both model performance and the interpretability of resulting explanations (Mehrabi et al., 2021).

#### Subjectivity in Interpretability Assessment

Interpretability remains an inherently subjective construct. While efforts were made to standardize human evaluation using financial analysts, scores for clarity, trust, and usability are influenced by individual expertise, cognitive bias, and the specific context of the task. Broader generalizability would benefit from larger, more diverse user studies or application of formal cognitive metrics (Doshi-Velez & Kim, 2017). Moreover, metrics such as “explanation complexity” or “fidelity” are useful but imperfect proxies for real-world transparency. Future work could incorporate more robust user-centric evaluations such as decision simulation experiments or eye-tracking studies to measure cognitive load and trust calibration more objectively.

## 5. Conclusion

This study presented a comparative evaluation of state-of-the-art explainable artificial intelligence (XAI) models applied to financial risk assessment tasks such as credit scoring and loan default prediction. Through empirical benchmarking and qualitative assessment, we examined the trade-offs between predictive performance and interpretability, offering practical insights into model suitability in real-world financial contexts.

### 5.1 Key Findings

- **Predictive Performance:**
  - Gradient boosting models (XGBoost) achieved the highest predictive accuracy (AUC ≈ 0.86), followed closely by neural networks and random forests.
- **Post hoc XAI Tools:**

  - **SHAP** provided the most faithful and consistent explanations across black-box models, though at a higher computational cost.
  - **LIME** offered faster explanations but was more sensitive to perturbations, resulting in lower explanation fidelity.
- **Intrinsic Interpretability:**
  - **RuleFit** and **Interpretable Decision Sets** slightly underperformed in accuracy but produced transparent, stable explanations preferred by domain experts.
  - These models required significantly less computational overhead and were better aligned with financial logic and compliance demands.
- **Human Evaluation:**
  - Financial analysts rated rule-based explanations higher in clarity, relevance, and trust compared to feature attribution from post hoc models.
- **Contextual Trade-offs:**
  - No model was universally superior—optimal selection depends on whether the primary objective is **accuracy**, **transparency**, or **regulatory compliance**.

### 5.2 Recommendations for Practitioners

To support responsible and effective deployment of AI in financial risk contexts, the following practical recommendations are proposed:

✅ **Use black-box models (XGBoost + SHAP)** when:

- High predictive accuracy is critical (internal scoring engines).
- Infrastructure supports explanation processing and auditing.

✅ **Use interpretable models (RuleFit, IDS)** when:

- Decisions must be auditable and defensible (e.g., regulatory reporting, adverse action notices).
- Model outputs are consumed directly by human decision-makers (e.g., loan officers, credit analysts).

✅ **Adopt hybrid approaches** that combine high-performing models with robust explainability layers when:

- Stakeholders include both technical and non-technical users.
- Trust, transparency, and accuracy are all essential in balanced measure.

## References

1. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 52138–52160.
2. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115.
3. Chen, Y., Li, J., Zhang, Y., & Wang, C. (2023). Interpretable credit risk modeling with multi-source data using XAI techniques. *Expert Systems with Applications*, 217, 119578.
4. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
5. Du, M., Liu, N., Hu, X., & Liu, H. (2021). Techniques for interpretable machine learning. *Communications of the ACM*, 64(1), 68–77.
6. Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3), 50–57.
7. Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11), 2767–2787.
8. Liu, Y., Wang, Z., Lin, X., & Zhang, Y. (2022). A hybrid interpretable model for financial risk prediction based on ensemble learning and XAI. *Applied Intelligence*, 52(10), 10581–10598.
9. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
10. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "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*, 1135–1144.
11. Samek, W., Wiegand, T., & Müller, K. R. (2019). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *ITU Journal: ICT Discoveries*, 1(1), 39–48.
12. Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*
13. Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3), 916–954.
14. Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11), 2767–2787.
15. Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. *Proceedings of the 22nd ACM SIGKDD*, 1675–1684.
16. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
17. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *KDD '16*, 1135–1144.
18. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.
19. Liu, Y., Wang, Z., Lin, X., & Zhang, Y. (2022). A hybrid interpretable model for financial risk prediction based on ensemble learning and XAI. *Applied Intelligence*, 52(10), 10581–10598.