Explainable AI for Credit Risk Management

Proof of Concept and Analysis

Bianca Fernandes, Dianni Estrada and Jiayi Wu
M2 Financial Technology and Data

May 12, 2025

Contents

1 Motivation

Credit risk assessment is a cornerstone of financial stability and institutional decision-making. Traditionally grounded in rule-based models and scorecards, the field has evolved with the advent of machine learning, which offers significantly improved predictive performance. However, these gains often come at the cost of transparency, raising concerns regarding model interpretability, fairness, and regulatory compliance.

This trade-off between accuracy and explainability presents a major research and practical challenge, particularly in domains where decisions must be both statistically sound and ethically defensible. As complex models such as ensemble methods and neural networks are increasingly deployed in credit scoring, the inability to interpret their outputs becomes a barrier to adoption and trust.

Explainable AI (XAI) has emerged as a promising framework to address this issue. By providing post hoc interpretability, XAI enables stakeholders to understand, audit, and potentially contest model decisions. In the context of credit risk management, this translates into the ability to identify the key drivers behind default predictions, support transparent client communication, and ensure alignment with regulatory expectations.

The aim of this project is to evaluate the practical benefits of XAI methods within a credit scoring pipeline. Specifically, we develop and validate a machine learning model for default prediction and apply leading XAI techniques—SHAP and LIME—to interpret the results. In addition, we incorporate Captum, a PyTorch-based interpretability library, to analyze feature attributions within a deep learning model. This allows us to explore how different types of XAI tools perform across model architectures, from tree-based ensembles to neural networks.

Beyond technical performance, our objective is to assess whether explainability contributes meaningfully to model usability and governance. This work contributes to ongoing academic discussions around responsible AI, model interpretability, and the integration of machine learning in high-stakes decision environments. By situating our analysis within a realistic credit risk context, we aim to provide both methodological insight and empirical evidence to inform the adoption of XAI in financial services.

To evaluate the applicability of Explainable AI (XAI) in credit risk management, we first needed to construct a reliable predictive model grounded in realistic financial data. This required selecting an appropriate dataset, preparing it for analysis, and ensuring its quality and consistency. These initial steps are critical not only for model performance, but also for the credibility of any interpretability techniques applied thereafter. The following section outlines our approach to dataset selection and preprocessing as the foundation for building a robust, explainable credit scoring model.

2 Data Exploration & Preprocessing

After evaluating the three proposed datasets, we selected the Credit Risk Dataset (approximately 33,000 observations) for its relevance, granularity, and breadth of financially meaningful features. The German Credit dataset, while historically used for benchmarking, was excluded due to its limited sample size (1,000 observations) and absence of critical financial indicators such as income and loan-to-income ratio. The Credit Card Approval dataset was also deemed unsuitable, as it lacked key variables required for robust risk modeling—particularly comprehensive credit histories and quantitative loan characteristics.

The selected dataset provides a solid foundation for credit risk analysis, including core features such as annual income, loan amount, interest rate, homeownership status, credit history, and default flag.

To ensure data integrity prior to model development, the following preprocessing steps were undertaken:

- Removal of Redundant Features: The variable loan-grade was dropped due to its strong collinearity (pearson correlation of 93%) with loan-int-rate, potentially introducing multicollinearity issues in certain models.
- Correction of Data Entry Errors: Five records reported implausible ages ranging from 123 to 144, exceeding human lifespans. These were presumed to be data entry anomalies and were removed.
- Filtering Inconsistent Employment Records: We identified and excluded 897 cases where the reported employment length exceeded the individual's age (e.g., employment lengths over 100 years for individuals under 30), indicating data inconsistencies.

These preprocessing actions were necessary to uphold the reliability and interpretability of subsequent model outputs, particularly in the context of explainable AI, where input feature quality directly influences the validity of interpretability metrics.

3 Model Development & Evaluation Metrics

We implemented and evaluated five classification algorithms commonly used in credit risk modeling: Logistic Regression, Random Forest (RF), XGBoost (XGB), Support Vector Machine (SVM), and a basic Neural Network architecture. All models were trained on

the cleaned and preprocessed dataset, with the objective of predicting a binary target variable indicating loan default.

To assess model performance, we reserved a held-out test set and employed three key evaluation metrics:

- Recall, to measure the model's ability to correctly identify actual defaulters—crucial for minimizing credit losses.
- Precision, to evaluate how many of the predicted defaulters were indeed true defaulters—important to limit false positives and avoid unnecessarily rejecting creditworthy applicants.
- F1-Score, the harmonic mean of precision and recall, offering a balanced measure that is particularly useful in the presence of class imbalance.

This multi-metric evaluation framework ensures a comprehensive understanding of each model's predictive behavior, especially in a domain like credit scoring where the cost of misclassification varies asymmetrically.

Table 1: Model 1

Model	**Recall**	**Precision**	**F1-Score**
Logistic Regression	0.448	0.727	0.554
Random Forest	0.672	0.933	0.781
XGBoost	0.677	0.956	0.793
SVM	0.321	0.844	0.465
Neural Networks	0.606	0.742	0.667

3.1 Threshold Tuning for Recall Optimization

In credit risk modeling, missing a true defaulter (false negative) can lead to significant financial loss. Therefore, our primary objective was to maximize recall—the ability to correctly identify risky clients. To achieve this, we adjusted the classification threshold from the default 0.5 to 0.2, making the models more conservative in granting credit. This change was motivated by the goal of reaching at least 80% recall, ensuring the model captures the majority of potential defaulters, even if it means increasing false positives.

Among the models tested, Random Forest (RF) and XGBoost (XGB) were the only ones to achieve recall values above the 80% threshold, making them suitable candidates for our objective of minimizing missed defaulters. RF obtained the highest recall (0.826), but with moderate precision (0.605), resulting in an F1-score of 0.699. XGBoost, on the

other hand, achieved a slightly lower recall (0.807) but compensated with a significantly higher precision (0.688) and the highest F1-score (0.743) among all models, indicating a better tradeoff between capturing defaulters and limiting false positives. Based on its balanced performance, XGBoost was selected as the final model for further explainability and business evaluation.

This tradeoff has direct business implications. By prioritizing recall, the model reduces the likelihood of approving loans for clients likely to default, thus protecting the institution from credit losses. However, a lower precision means some creditworthy clients may be incorrectly classified as high-risk. While this could lead to missed business opportunities and customer dissatisfaction, the decision is justified in contexts where risk minimization outweighs volume growth. Furthermore, using Explainable AI tools alongside the model can help advisors justify rejections to clients, improving transparency and trust.

Table 2: Model 2

Model	**Recall**	**Precision**	**F1-Score**
Logistic Regression	0.779	0.444	0.566
Random Forest	0.826	0.605	0.699
XGBoost	0.807	0.688	0.743
SVM	0.781	0.444	0.566
Neural Networks	0.782	0.509	0.617

4 Explainable AI (XAI) Implementation

To assess the interpretability of our selected models—particularly the XGBoost classifier and a basic neural network—we employed three prominent Explainable AI (XAI) frameworks: SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), and Captum, a library specialized for interpreting neural networks in PyTorch. These methods offer complementary insights into model behavior: SHAP provides consistent global and local attributions; LIME offers instance-level approximations; and Captum, via Integrated Gradients, delivers gradient-based sensitivity analysis tailored to deep learning architectures.

4.1 SHAP Analysis – Global Feature Attributions

SHAP assigns each feature an additive contribution to a specific prediction by computing Shapley values, a concept from cooperative game theory that ensures fair allocation of "credit" among all features. One of SHAP's key strengths is consistency: if a model

changes such that a feature contributes more to predictions, its SHAP value will not decrease.

The SHAP summary plot for the XGBoost model reveals the distribution and magnitude of feature impacts across the entire dataset. The most influential variables include:

- Loan-to-Income Ratio: This feature exhibited the highest SHAP values. High ratios (depicted in red) were strongly associated with higher default probabilities, reflecting increased financial burden—a fundamental risk indicator in credit underwriting.
- Elevated interest rates contributed positively to the default prediction, consistent with lending models where riskier clients are offered higher rates to compensate for credit risk.
- Lower values of income (represented in blue) significantly increased the likelihood of default. This inverse relationship reinforces standard credit assessment logic.

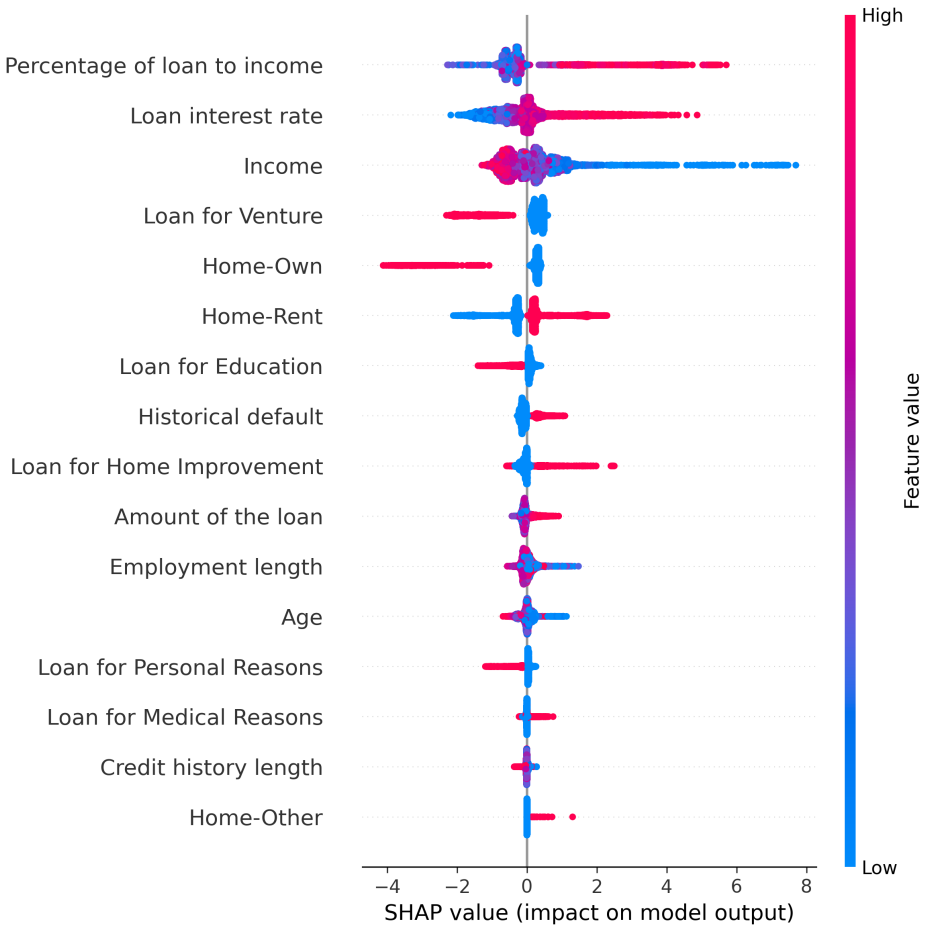

Figure 1: SHAP Summary Plot

4.2 LIME Analysis – Instance-Level Interpretability

While both SHAP and LIME can provide local, individual-level explanations, SHAP does so with a strong theoretical foundation and consistency guarantees. However, due to its computational cost, practitioners sometimes turn to LIME for faster, more interpretable, surrogate-model-based explanations by perturbing inputs and observing output changes.

We analyzed a specific instance flagged as high-risk by the XGBoost model. LIME identified the following key contributors:

- Interest Rate higher than 2.65% and Non-Homeowner Status were the most influential positive drivers of the default prediction, echoing the global SHAP findings.
- On the contrary, High Income Levels (log higher than 11.29) and Low Loan-to-Income Ratio acted as mitigating factors, decreasing the predicted risk.
- Context-specific features such as loan purpose exhibited nuanced contributions, with certain categories (e.g., educational or medical loans) contributing differently depending on the presence of stabilizing financial attributes.

These localized insights are vital for generating case-specific justifications, especially in borderline scenarios or disputes, where model transparency must translate into client-understandable language.

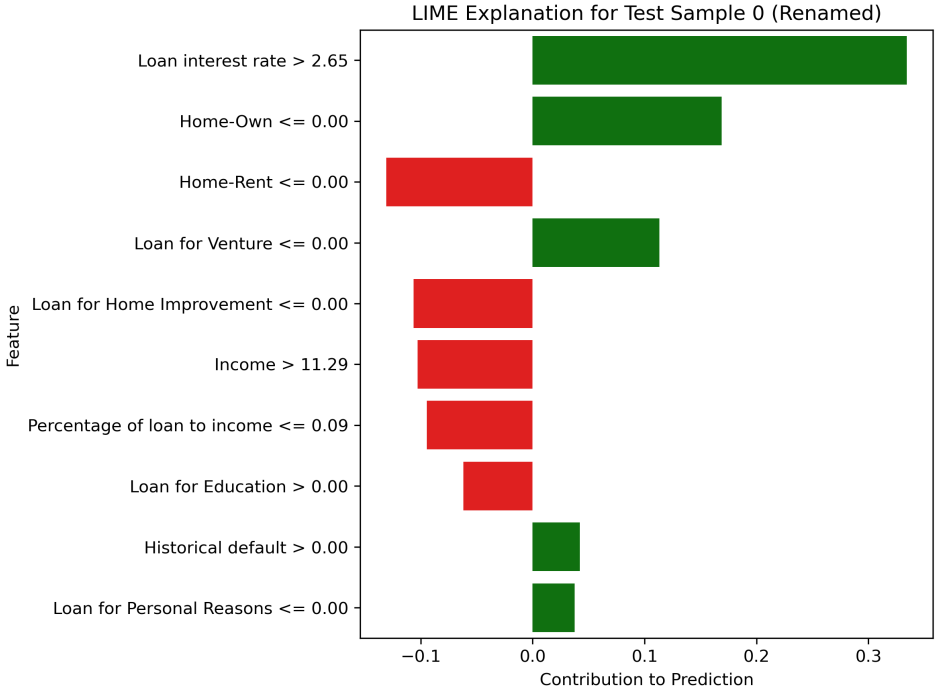

Figure 2: LIME Local Explanation for a Single Prediction

4.3 Captum – Gradient-Based Attribution for Neural Networks

To further explore interpretability in the context of neural networks, we applied Captum using the Integrated Gradients method. Integrated Gradients compute feature attributions by integrating the model's gradients along a path from a baseline input to the actual observation, offering a principled view of how small changes in input affect the model's output.

Our Captum results indicated that:

- Loan interest rate and income were once again dominant predictors of risk, with both positive and negative attributions depending on their values.
- Loan amount and educational loan purpose also showed non-trivial contributions.
- Some features deemed important by SHAP (e.g., loan-to-income ratio) had lower attributions here, reflecting possible differences in model architecture, feature interactions, or the choice of baseline.

This divergence highlights the value of using multiple XAI tools: while SHAP provides feature-agnostic consistency, Captum's gradient-based sensitivity offers deeper insight into how a neural network internally weights inputs.

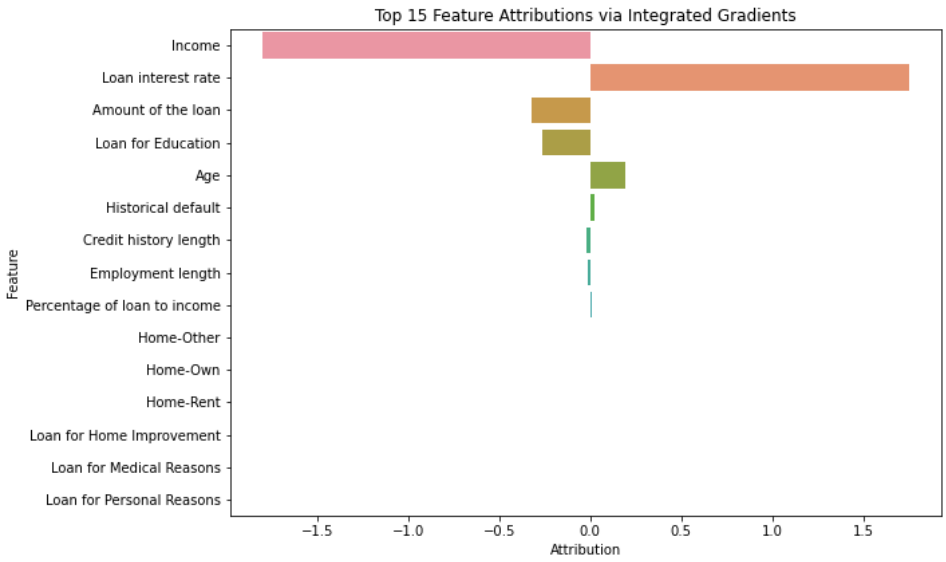

Figure 3: CAPTUM summary plot

4.4 Interpretation and Business Relevance

The application of SHAP, LIME, and Captum has enabled a multifaceted analysis of model behavior, offering both theoretical and empirical insights into the interpretabil-

ity of complex machine learning algorithms used in credit scoring. The convergence of findings across these frameworks—particularly the consistent importance of features such as loan interest rate, income, and loan amount—suggests that the models are learning functionally valid relationships aligned with established credit risk assessment principles.

From a methodological perspective, the triangulation of interpretability techniques enhances the robustness of our conclusions. SHAP's axiomatic foundation provides confidence in the fairness and consistency of feature attributions. LIME complements this with instance-level fidelity, useful for investigating localized model behavior. Captum's gradient-based approach, particularly through Integrated Gradients, offers valuable insights into the sensitivity of deep learning models to input variations, thus enriching our understanding of neural network decision boundaries.

Importantly, these interpretability tools also serve a critical role in advancing the broader goals of responsible AI. By enabling stakeholders to inspect and rationalize model outputs, they contribute to:

- Model transparency: Facilitating internal audit, validation, and documentation of algorithmic decision-making processes.
- Regulatory compliance: Supporting requirements for explainability in automated credit decisions, as stipulated in frameworks such as the EU's General Data Protection Regulation (GDPR).
- Algorithmic accountability: Allowing institutions to detect potential biases or unintended dependencies on socio-economic proxies, thereby contributing to fairer decision-making systems.

While each method presents its own strengths and limitations, their combined use provides a comprehensive interpretability framework that enhances both epistemic trust (trust in model logic and validity) and procedural trust (trust in how decisions are generated and justified). In high-stakes financial contexts such as credit risk management, this dual layer of trust is essential—not only for institutional governance, but also for sustaining public confidence in automated decision systems.

5 Conclusion

This project examined the integration of Explainable AI (XAI) techniques into credit risk modeling, combining predictive performance with interpretability. Through SHAP, LIME, and Captum, we assessed the internal logic of an XGBoost model and a neural

network, finding consistent and economically coherent patterns in the key drivers of default.

Our results confirm that XAI methods not only enhance transparency but also support regulatory compliance and model governance—critical considerations in financial decision-making. The alignment of model attributions with domain knowledge strengthens both trust and accountability in automated credit decisions.

In light of these findings, we conclude that investing in XAI for credit risk management is both technically justified and operationally beneficial, providing a foundation for responsible and interpretable use of machine learning in lending contexts.