White Box Finance: Interpreting AI Decisions in Finance through Rules and Language Models

Oluwafemi Azeez[1 2] Samson Tontoye[1] Olorunleke White[1] Abuzar Royesh[1]

[1]Pastel Africa [2]AI Saturdays Lagos

Motivation

- Loan defaults → major financial losses.
- ML models (e.g., XGBoost) improve prediction, but are black-boxes.
- Finance requires transparent, auditable explanations for regulators, loan officers, and customers.

Research Goal/Methodology

Enhance interpretability and trust in AI credit risk models by creating and comparing:

- SHAP + GPT-4 → feature-based + natural language explanations.
- Rule-based logic → transparent, business-aligned decision rules.

Experimental Setup

- Dataset: Anonymized loan applicant records containing demographics, employment, credit history, and repayment behavior.
- Preprocessing: Missing values removed (<1%), categorical variables frequency-encoded, numerical features preserved.
- Model: XGBoost classifier trained with 5-fold stratified cross-validation. Class imbalance addressed using scale_pos_weight.
- Evaluation Metrics: Area Under the Curve (AUC), Precision, Recall, F1-score, and Confusion Matrix analysis.

Explanation Modules

Two complementary explanation pipelines were applied to model predictions:

- SHAP + GPT-4: Local feature attributions → top 3–5 contributors → converted into business-friendly textual narratives.
- Rule-Based Logic: Categorical histograms and KDE plots used to derive interpretable decision rules aligned with institutional underwriting heuristics.

Age	Income	LoanAmount	CreditScore
36	80846	179949	347

MonthsEmployed	InterestRate	DTIRatio	Education
20	23.96	0.9	PhD

A Test Dataset sample for prediction

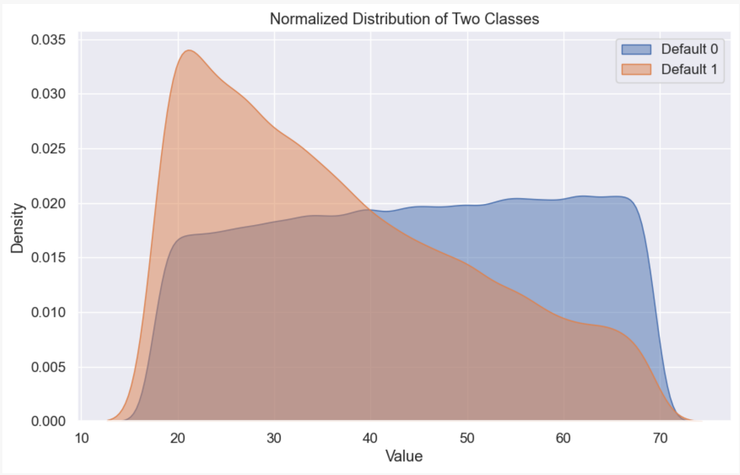
Result

The sample in table 1 was passed into the model and it predicted it as loan default. We then passed it through the explanation pipeline and got the following results.

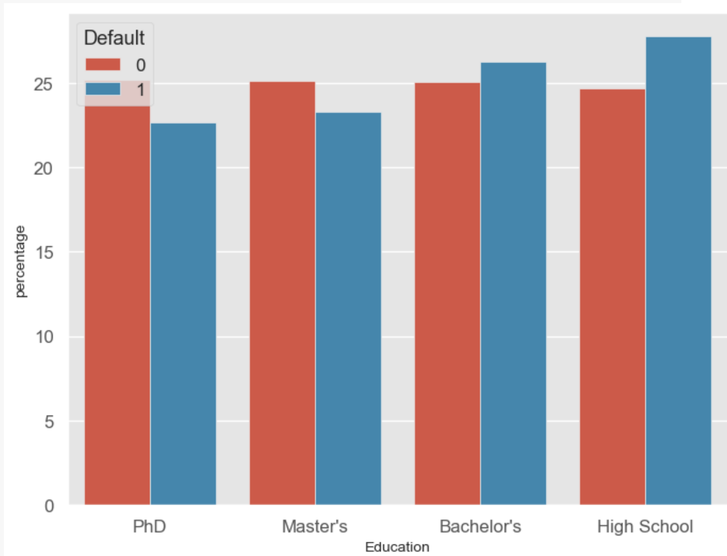
- **GPT-Supported Explanations (SHAP + GPT-4 Pipeline)** Based on the provided information, the risk of this customer defaulting on their loan can be explained by examining the key factors and their respective SHAP impacts: The interest rate on the loan is quite high at 22.75%. This significantly increases the cost of borrowing, making it more challenging for the customer to manage their monthly payments. The high SHAP impact of 0.71 indicates that this factor is a strong contributor to the default risk
- **Rule-Based Explanations (KDE + Business Logic Rules)** From the KDE plot in Figure 1 for example, you would notice that you can visually create a business logic on numerical variable age based on the boundary of 40, The age variable in the table 1 is 36 and less than 40 so a good explanation about young age listed below would be reasonable
  1. Short employment duration may indicate job instability.
  2. High interest rate increases financial burden, raising risk.
  3. Low credit score indicates high risk of default.
  4. Young age may indicate lack of financial experience.
  5. High debt-to-income ratio indicates financial strain.
  6. High loan amount increases risk of default.

Comparative Insights

- GPT Explanations → Rich, nuanced, human-friendly
- Rule-Based Explanations → Transparent, audit-ready, regulatory aligned
- Hybrid Approach = Best of both worlds: trust + compliance



KDE Plot of Age (continuous) for default 0 and default 1 class.



Histogram of Education levels comparing default (1) vs non-default (0) borrowers.

References

[1] Scott Lundberg and Su-In Lee.
A unified approach to interpreting model predictions, 2017.

[2] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin.
"why should i trust you?": Explaining the predictions of any classifier, 2016.

[3] Shreya and Harsh Pathak.
Explainable artificial intelligence credit risk assessment using machine learning, 2025.

[4] Jiaxing Zhang et al.
LLMExplainer: Large Language Model Based Bayesian Inference for Graph Explanation Generation.
arXiv preprint arXiv:2407.15351, jul 2024.