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

Oluwafemi Azeez
Pastel Africa
femi@pastel.africa

Samson Tontoye
Pastel Africa
samsont@pastel.africa

Olorunleke White
Pastel Africa
olorunleke@pastel.africa

Abuzar Royesh
Pastel Africa
abuzar@pastel.africa

Abstract

Although loan defaults continue to cause substantial financial losses, this study focuses on improving how AI credit risk models are explained. Beyond developing a predictive model based on the demographics of the borrower, the attributes of the loan, and the credit history, the core contribution lies in introducing and comparing explanation methods. Specifically, we evaluated two ways to provide explanations. One method is a module that integrates SHAP values and GPT-4 to generate human-friendly narratives, a second is a rule-based logic explanation. This approach aims to enhance interpretability and trust, offering a clearer understanding of model predictions than traditional explanation techniques.

1. Introduction

1.1. Literature Review

Model explainability is central to promoting transparency in machine learning applications, especially within high-stakes domains like finance. Several techniques have been proposed in the literature to demystify black-box models. Lundberg and Lee [1] introduced SHAP, a unified framework based on game-theoretic Shapley values that attributes the contribution of each feature to a model's output. This technique has become a standard for post hoc interpretability across tree-based models, including XGBoost. Ribeiro et al. [2] proposed LIME, which builds locally linear interpretable models around each prediction. Although effective, its sensitivity to perturbations often limits its robustness. SHAP overcomes this by ensuring consistency and local accuracy. More recently, attention has shifted to language-based explanations. Tools such as LLMExplainer [4] and GPT-4-based methods demonstrate how large language models (LLMs) can augment feature-based explanations with human-readable justifications. In the financial domain, explainability tools have seen application in loan and credit risk modeling [3]. These efforts highlight the growing importance of visual and textual explanations in improving end-user trust, regulatory compliance, and auditability.

1.2. Background

Financial institutions face significant losses due to loan defaults, which occur when borrowers fail to meet repayment obligations. Traditional rule-based credit scoring systems struggle to adapt to nonlinear borrower behavior and may misclassify borrowers with atypical but reliable profiles. Machine learning (ML) models, such as gradient-boosting trees, have recently improved predictive performance for loan default detection. However, their black-box nature remains a major barrier to adoption in highly regulated domains such as finance. Stakeholders, including loan officers, compliance teams, and regulators, require clear justifications for automated decisions. This has led to the rise of explainable artificial intelligence (XAI), which aims to provide transparency in ML predictions. Prominent XAI methods include SHapley Additive exPlanations (SHAP), which assign feature-level attributions to model outputs. Additionally, recent advances in large language models (LLMs) such as GPT-4 enable the generation of human-readable natural language explanations. In this study, we propose and compare two methods:

1. Rule-based logic to capture high-level decision heuristics

2. local SHAP visualizations to attribute feature contributions at the individual prediction level, and GPT-generated textual rationales to translate explanations into business-friendly language.

2. Methodology

2.1. Data Preprocessing

The dataset comprised anonymized records of loan applicants and their repayment behaviors. The key attributes included demographic data, employment details, financial history, and credit bureau characteristics. Initial data cleaning steps involved; Removing a negligible fraction of records with missing values, encoding categorical variables using frequency encoding to retain ordinal relationships, and conserving numerical column scales, as XGBoost inherently handles unscaled data effectively.

2.2. Feature Engineering

To capture non-linear signals, several derived features were introduced and features selected by removing highly correlated variables (Pearson > 0.9) and low-variance columns.

2.3. Model Training

An XGBoost classifier was employed for its robustness and ability to handle missing values and non-linear interactions in tabular data. A key challenge was class imbalance: loan defaults were significantly less frequent than non-defaults. To address this, the $scale\_pos\_weight$ parameter was computed as follows:

$$weight = \frac{\sum_{y=0}}{\sum_{y=1}} \tag{1}$$

The model evaluation used 5-fold stratified cross-validation to preserve class proportions across splits. The final training was conducted on the complete training set. Performance was evaluated, and this yielded discrimination capability and calibration between predicted probabilities and actual default outcomes.

2.4. Model Validation

The model achieved an AUC of 0.74, with good balance between precision and recall. Confusion matrix analysis indicated that the classifier maintained conservative decision boundaries to minimize false positives, a crucial metric in risk-sensitive applications.

2.5. Post-Processing

The explanation module Predictions and probability scores from the XGBoost model were fed into the explanation module to derive rationale for each instance, enhancing transparency and trustworthiness of automated decisions.

3. Explanation Methods

The explanation module is designed to provide intelligible justifications for individual loan default predictions using a combination of SHAP (SHapley Additive exPlanations), rule-based logic and GPT-4o. This modular design enables explainability in both structured numerical formats and human-understandable narratives.

3.1. SHAP Explainer

We applied SHAP TreeExplainer to the trained XGBoost model to calculate the local feature attributions for each prediction. SHAP values provide additive feature importance scores for each input instance. These values are visualized using waterfall plots, where positive and negative contributors to the predicted probability are clearly distinguished. This graphical representation enables analysts to rapidly identify why a borrower was classified as high or low risk.

3.1.1 Prompt Generator

The prompt generator converts the top SHAP-ranked features into a domain-specific natural language prompt. It selects the top 3 to 5 contributing features, attaches their directionality. This structured prompt forms the basis for the next stage.

3.1.2 Language Generator (GPT-4o)

Each prompt is asynchronously sent to OpenAI's GPT-4o via the openai Python SDK. GPT-4o processes the inputs and returns a short, explainable narrative describing why the customer was flagged as a high or low default risk. The model runs with a low temperature (temperature=0.4) to ensure conciseness and factual consistency. Outputs are saved to the dataframe alongside SHAP plots for each record.

3.1.3 Integration Pipeline

The entire explanation module runs in asynchronous batches (typically 100–200 records) to optimize throughput while respecting rate limits. The final output includes; SHAP plots saved as .png for visual explanation, Textual justifications appended to each record, Audit-ready logs for compliance and decision traceability

Age	Income	LoanAmount	CreditScore
36	80846	179949	347

MonthsEmployed	InterestRate	DTIRatio	Education
20	23.96	0.9	PhD

Table 1. A Test Dataset sample for prediction

3.2. Rule-Based Logic Method

We create an histogram of categorical variables normalized across the target variable class and plotted together e.g figure 2 and KDE plot of numerical variables grouped by target class e.g figure 1. The visualization is then used to determine boundary variables and business logical rules that could be used to provide explanations after a prediction is made. e.g

```
if row["Age"] < 40:
    explanations.append("Young age may
    indicate lack of financial experience.")
```

If any of these business rules are triggered, a tag is added to the explanation record, reinforcing the decision from both a statistical and deterministic perspective. This enhances model reliability and auditability by aligning predictions with institutional underwriting policies.

The various approach statistical(SHAP), rule-based, and language-based ensures robust, interpretable and human aligned explanations for credit risk predictions.

4. Results

1. **gpt_explanation**: 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:

    (a) **Interest Rate**: The interest rate on the loan is quite high at 23.96%. This significantly increases the cost of borrowing, making it more challenging for the customer to manage their monthly payments. The high SHAP impact of 0.95 indicates that this factor is a strong contributor to the default risk.

    (b) **DTI Ratio**: The debt-to-income ratio of 0.9 suggests that the customer's monthly debt payments are very high relative to their income. However, the negative SHAP impact of -0.38 implies that, in this context, the DTI ratio is somewhat mitigating the default risk, possibly because the model expects even higher DTI ratios for high-risk cases.

    (c) **Months Employed**: The customer has been employed at their current job for 20 months. While this is a moderate duration, the positive SHAP impact of 0.32 suggests that the model views this employment length as a slight risk factor, possibly due to the lack of longer-term job stability.

2. **rule_explanation**: 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

    (a) Short employment duration may indicate job instability.

    (b) High interest rate increases financial burden, raising risk.

    (c) Low credit score indicates high risk of default.

    (d) Young age may indicate lack of financial experience.

    (e) High debt-to-income ratio indicates financial strain.

    (f) High loan amount increases risk of default.

5. Conclusion

In financial applications of AI, especially those that involve risk-sensitive tasks such as credit scoring, the ability to generate understandable and trustworthy explanations is crucial. This paper introduced two explanation methods designed to improve the interpretation of the model at the individual prediction level.

The first method leverages local SHAP values to identify the most influential features in a prediction, and then utilizes GPT-based natural language generation to produce human-readable explanations. By pairing the importance of quantitative characteristics with qualitative descriptions, this approach allows contextual and user-friendly interpretations of the behavior of the model.

The second method focuses on the extraction of rules through descriptive statistics. By analyzing feature distributions (e.g., via histograms) across classes, simple yet effective logical rules can be derived. These rules are used to construct transparent, rule-based explanations that can be applied post-prediction to help clarify why a particular decision was made.

Together, these approaches strike a balance between statistical rigor and semantic clarity, providing a pathway toward more interpretable and actionable AI systems in finance. Future work may involve validating the

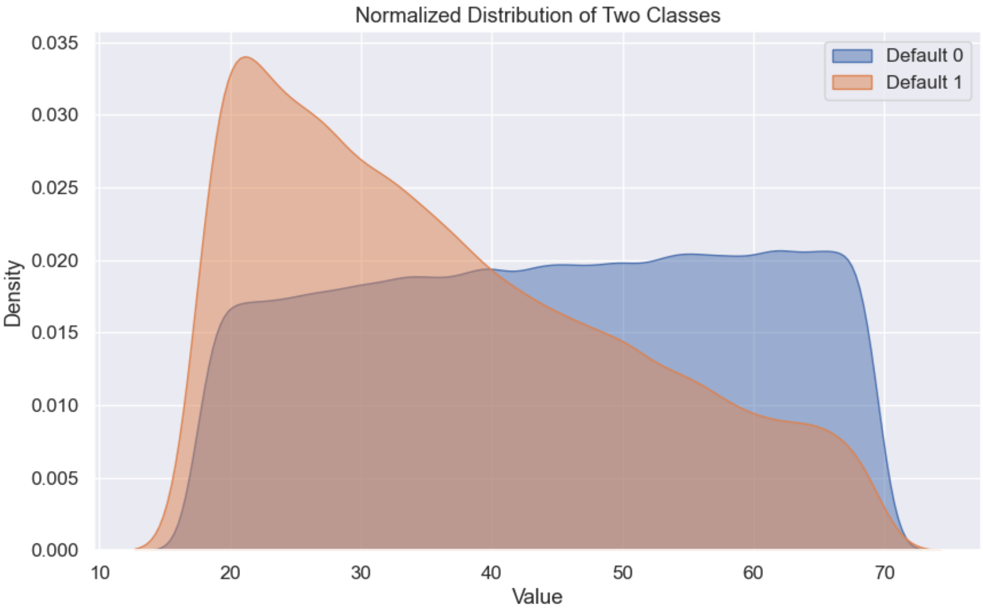

Figure 1. KDE Plot of Age (continuous) for default 0 and default 1 class.

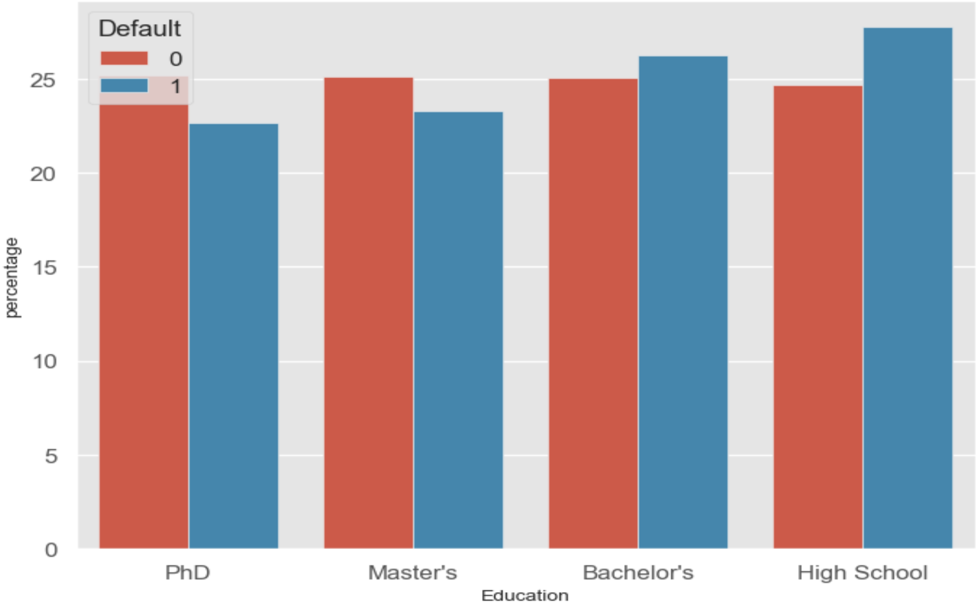

Figure 2. Histogram Plot of Education (categorical) for default 0 and default 1 class.

effectiveness of these explanations through user studies and expanding the logic framework to support more complex feature interactions.

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.