Integrating Predictive and Generative AI for Explainable Credit Risk

A Unified Framework for Epistemic Reliability and Model Validation

Author Names

Institution/Affiliation

Conference Name / Year

NotebookLM

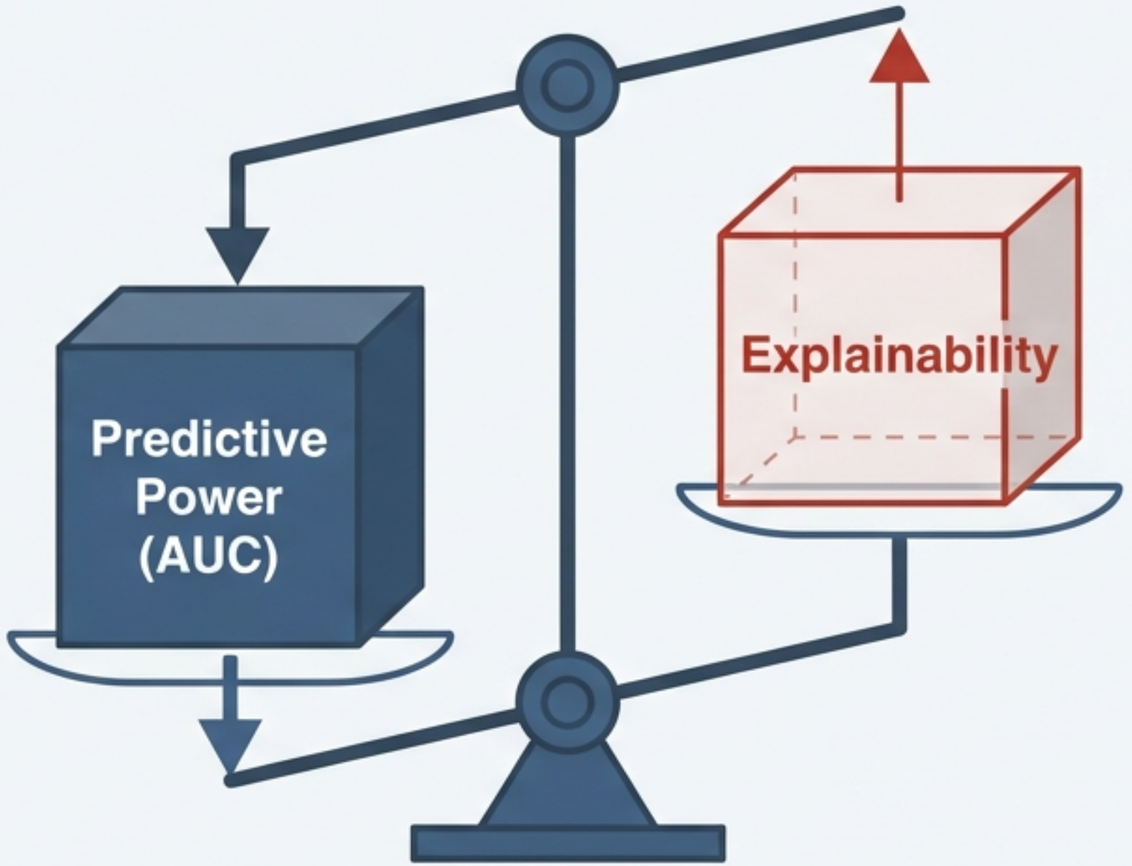
The Tension Between Predictive Power and Regulatory Transparency

Context: Credit-risk assessment dictates capital allocation and requires strict regulatory adherence (Basel accords).

Current State: Ensemble architectures like Bagged Neural Networks have replaced linear models due to superior discriminatory power.

The Gap: High predictive accuracy (AUC > 0.80) often sacrifices internal transparency.

Critical Issue: Models rank risk effectively, but their internal reasoning remains structurally opaque, creating compliance bottlenecks.



NotebookLM

Problem Statement: The Unreliability of Post-Hoc Explanations

The Core Conflict

Predictive performance ≠ Explanatory reliability. Accurate models may rely on spurious correlations or noise.

Methodological Deficit

Explainability is currently treated as descriptive storytelling rather than a falsifiable scientific test.

Tool Limitations

Instability: SHAP/LIME are sensitive to background distributions.
Hallucination: Unconstrained GenAI creates false causal narratives from weak signals.

Research Question

How can we decouple predictive performance from explanation reliability to ensure auditable validation?

NotebookLM

Research Objectives and Contributions

Primary Goal: Establish a unified framework that embeds explanation reliability as a first-class object.

Contribution 1: Feature Stabilization. Implementation of a ‘Dual-Selector’ mechanism (Random Forest + L1 Logistic Regression) to stabilize feature sets.

Contribution 2: Reliability Diagnostics. Introduction of ‘Sanity Ratios’ to quantify the distinction between structural signal and random noise.

Contribution 3: Constrained GenAI. A pipeline where Generative AI produces narratives only when attribution signals are empirically robust.

NotebookLM

Methodology: A Unified Predictive–Explanatory Workflow

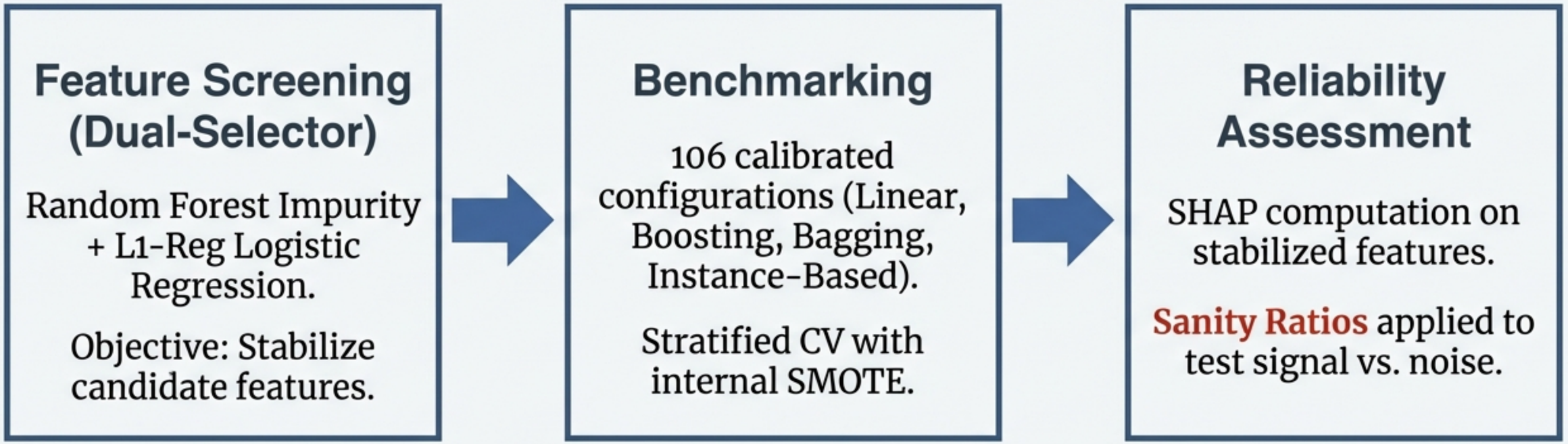


Figure 1: The integrated pipeline ensuring features are stabilized and models are calibrated before reliability testing.

NotebookLM

Data Specification and Experimental Setup

Dataset Specification

- Source: German Credit Dataset (UCI Repository)
- Observations: 1,000 (700 non-default / 300 default)
- Attributes: 20 (Financial exposure & demographics)

Preprocessing & Metrics

- Preprocessing: >90% missing dropped; mode/median imputation; One-Hot Encoding; Standardization.
- Imbalance Handling: SMOTE applied to training folds only.
- Predictive Metrics: AUC (primary), Brier Score, KS Statistic.
- Diagnostic Metrics: Sanity Ratios, Attribution Stability Scores.

NotebookLM

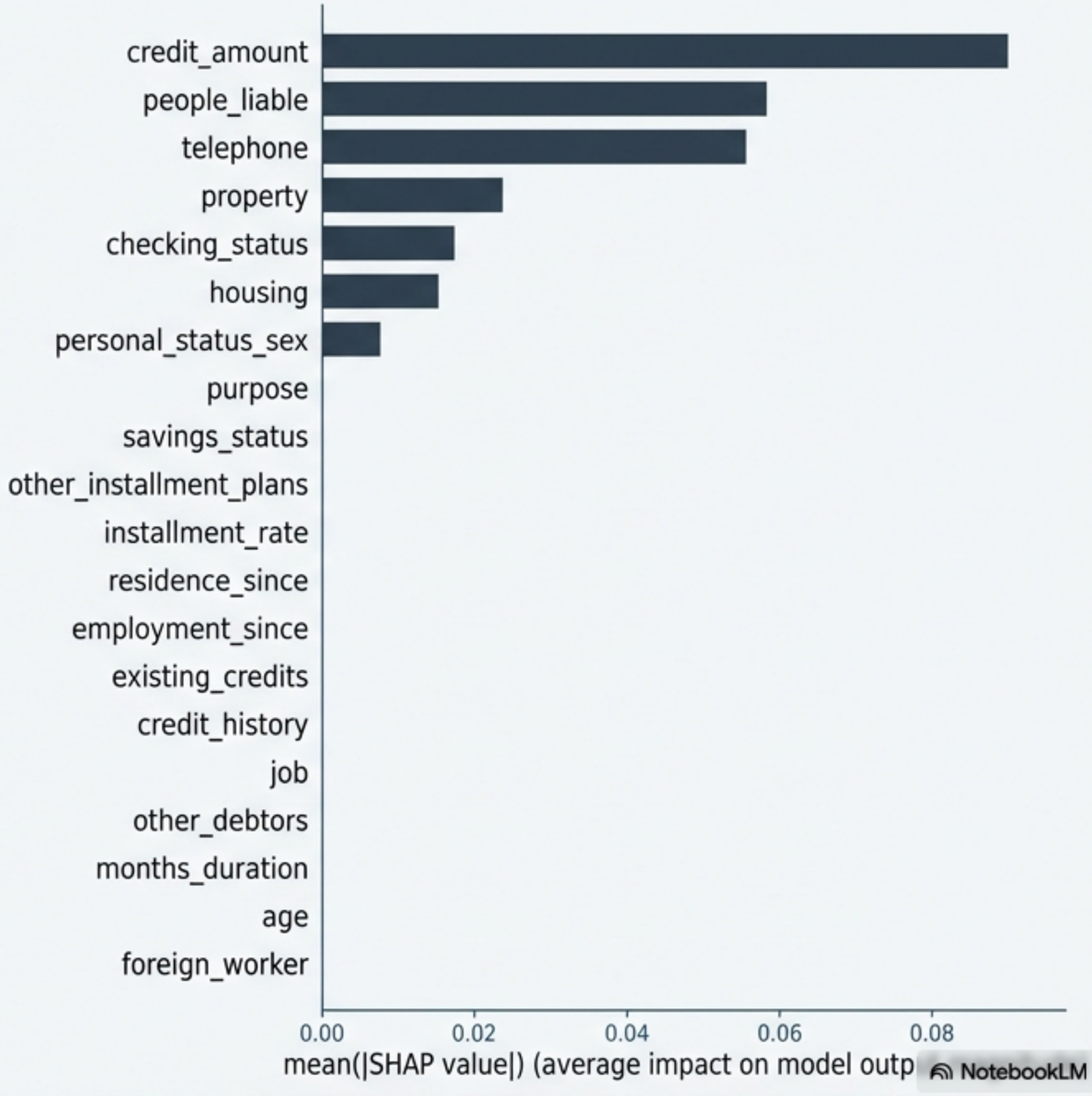
Feature Importance Analysis via Dual-Selector Mechanism

Mechanism: Aggregation of Random Forest impurity and L1-regularized logistic regression coefficients.

Key Drivers:

- Financial: ‘credit_amount’, ‘checking_status’
- Demographic: ‘people_liable’, ‘telephone’.

Finding: The Dual-Selector stabilizes input variables, ensuring downstream explanations focus on robust signals rather than noise.



NotebookLM

Predictive Performance Benchmarking

Model Family	Configuration	AUC	KS Statistic
Bagging (Ensemble)	Bagged Neural Networks (BagNN)	0.81	0.53
Boosting	Boosted Decision Trees	0.80	0.56
Linear	Reg. Logistic Regression	0.80	0.53
Instance-Based	KNN (Tuned)	0.77	0.51

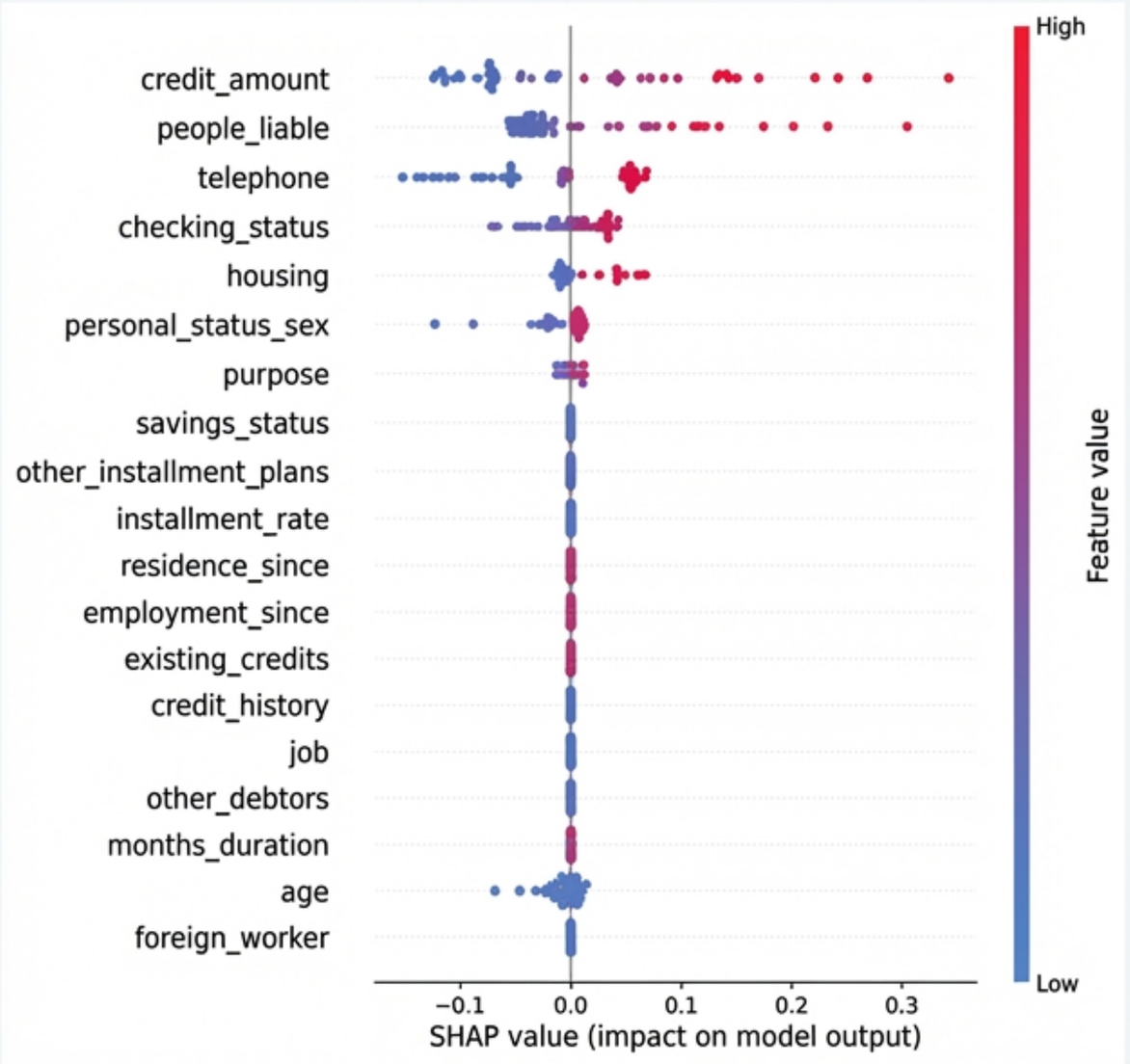
Top Performer: Bagged Neural Networks achieved highest discriminatory power.

Competitive Baselines: Regularized Logistic Regression is highly competitive with complex ensembles on tabular data.

NotebookLM

Global Explainability Analysis (SHAP)

- **Observation:** Global trends align with economic intuition.
- **Risk Drivers:** High ‘credit_amount’ (red dots on right) increases predicted risk.
- **Protective Factors:** Older ‘age’ shows a negative association with default risk.
- **Conclusion:** At a global level, the model appears to behave rationally.



NotebookLM

The Reliability Paradox: Signal vs. Noise

AUC ≈ 0.81

Predictive Power

Sanity Ratio ≈ 1.0

Explanation Reliability

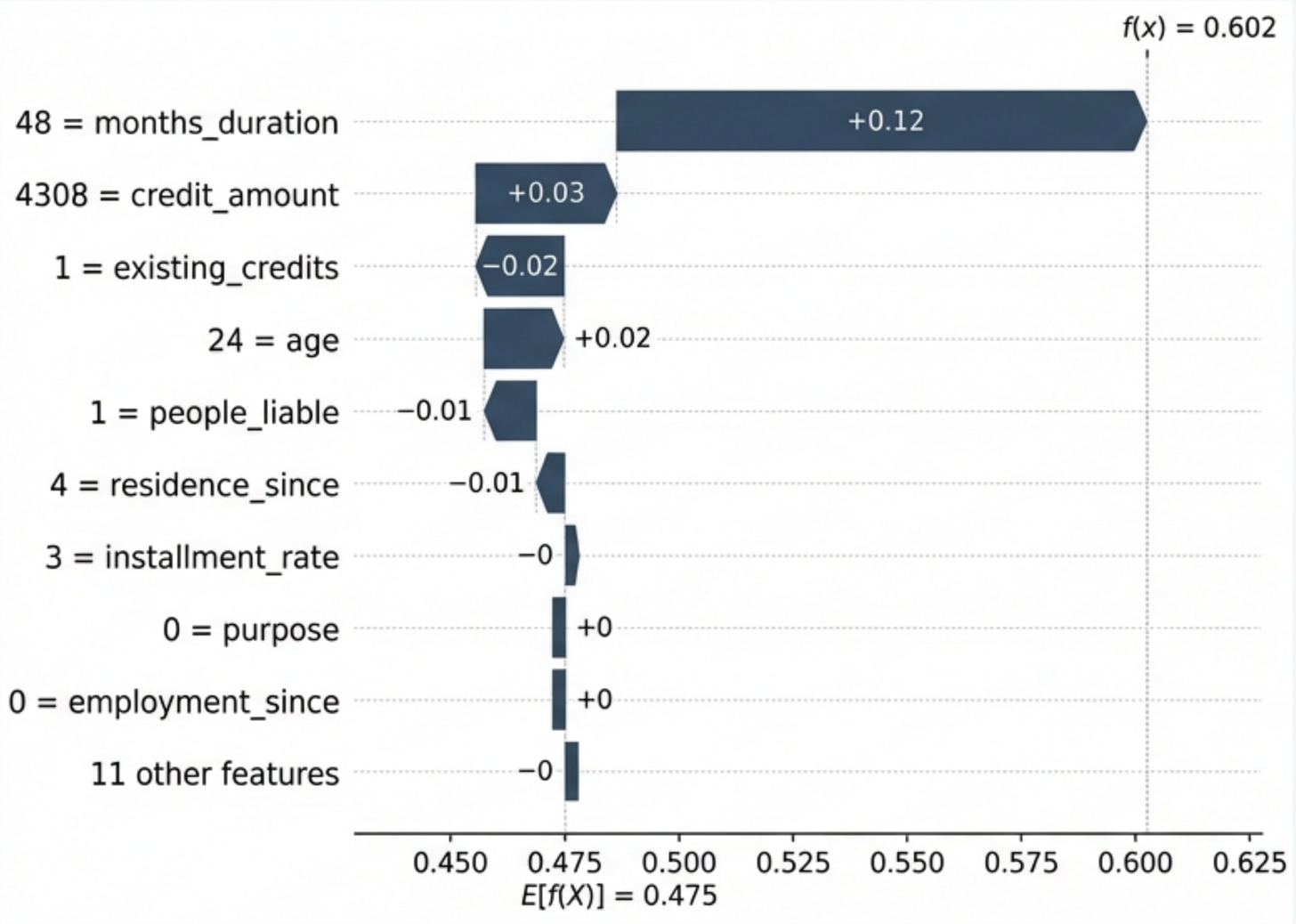
The Disconnect: Explanation quality varies independently of predictive accuracy.

Critical Finding: A Sanity Ratio of 1.0 indicates feature attributions are statistically indistinguishable from **noise**.

Implication: The model is **'accurate'** but its **internal reasoning cannot be distinguished** from random artifacts.

NotebookLM

Local Attribution Analysis and Evaluation



Instance Analysis (True Positive):
- High "months_duration" and "credit_amount" push prediction toward default.
- Visually plausible and intuitive.

Diagnostic Context: Despite the **visual clarity**, the reliability layer flags these specific attributions as **weak signals.**

Risk: Without the **reliability flag**, a human auditor would accept this explanation as **absolute truth.**

NotebookLM

Analysis: The Epistemic Shift in Model Validation

Signal vs. Noise

High-accuracy models leverage high-frequency noise that post-hoc explainers struggle to decompose. Feature stability is a prerequisite for reliable explanation.

Constrained GenAI

Without reliability filters, Generative AI translates weak attribution scores into confident claims. Our framework restricts output to only robust signals.

Paradigm Shift

Validation must move from verifying "economic plausibility" (does it look right?) to verifying "statistical robustness" (is it real?).

NotebookLM

Implications for Regulatory Compliance and Governance

Auditable AI

Explicitly flags uncertainty to human reviewers, aligning with Model Risk Management (MRM).

Rejection Criteria

Reject high-performing 'black boxes' if they fail reliability diagnostics (Sanity Ratios), regardless of AUC.

Operational Risk

Prevents narrative inflation in automated decision-support reports.

Trust

Founding regulatory trust on falsifiable evidence rather than visual intuition.

NotebookLM

Conclusion and Future Directions

- **Summary:** Predictive success is an insufficient proxy for explanatory truth.
- **Core Achievement:** A unified framework that filters out 'noise-based' explanations before they reach the user.
- **Future Work:** Extending reliability layers to deep learning architectures and adversarial robustness testing.
- **Final Takeaway:** Explainability must be redefined as a scientific test of signal quality, not a cosmetic add-on.

NotebookLM