

# Beyond Attribution: A Falsifiability Framework for Reliable Credit Risk Explanations

## Abstract

Machine learning explanations are often treated as credible outputs of model introspection, yet they may reflect statistical artefacts rather than genuine learned patterns. This study addresses a critical gap in credit-risk governance: the need for scientifically rigorous, testable explanations rather than post-hoc narratives. We propose a falsifiability framework that treats explanations as empirical hypotheses subject to systematic validation. The framework operationalizes this through three epistemically distinct components: (1) dual-selector feature stabilization combining supervised learning with dimension reduction; (2) reliability diagnostics including sanity ratio validation and signal-to-noise discrimination; and (3) a constrained generative AI module that grounds natural language explanations in validated SHAP evidence. By integrating explanation reliability assessment directly into model development—rather than treating it as an afterthought—we demonstrate that strong predictive performance and reliable explanations are orthogonal properties requiring independent scrutiny. Our experimental validation on credit-risk datasets shows that explanations flagged as unreliable by our sanity checks correspond to models with hidden performance degradation, establishing empirical grounding for the falsifiability approach. This work enables transparent, auditable explanations suitable for regulatory compliance and stakeholder trust, fundamentally advancing how financial institutions govern machine learning systems.

**Keywords:** Credit Risk, Explainability, Reliability, Falsifiability, SHAP, Model Governance

## Introduction

Credit-risk assessment plays a central role in financial decision-making, shaping lending policies and capital allocation in regulated institutions (Baesens et al., 2003; Lessmann et al., 2015). Contemporary machine-learning models have substantially improved predictive discrimination, yet their explanations remain epistemically untested. A widespread assumption persists in practice: strong predictive performance implies reliable model explanations. This assumption is false. A model’s ability to discriminate borrowers by default risk does not guarantee that its feature attributions reflect genuine learned structure rather than statistical artefacts, sampling noise, or spurious correlations (Hassija et al., 2024). This disconnect between prediction and explanation constitutes a fundamental scientific problem: explanations are treated as credible outputs rather than as empirical hypotheses subject to rigorous validation.

The challenge is compounded by the proliferation of post-hoc explainability tools and generative AI. SHAP and LIME assign numerical attributions to features, addressing some interpretability concerns in high-accuracy ensemble models (Lundberg & Lee, 2017; Ribeiro et al., 2016). However, growing evidence indicates that these attribution methods are highly sensitive to background distributions and sampling noise, frequently reflecting artefacts rather than genuine model structure (Hassija et al., 2024; Slack et al., 2020). Recent applications of generative AI that translate attribution scores into natural-language explanations remain largely unconstrained, further amplifying risks of hallucination, narrative inflation, and spurious causal claims (Hassija et al., 2024). The result

is a proliferation of confident-sounding explanations with no mechanism to assess their empirical grounding.

This fragmentation exposes a deeper methodological gap: explainability in credit-risk modelling is rarely treated as a falsifiable scientific process. Existing work predominantly frames explanations as descriptive summaries or visualisation artefacts rather than as testable hypotheses subject to systematic evaluation. No prior framework integrates predictive modelling, attribution stability assessment, and constrained generative reasoning while explicitly quantifying whether explanatory signals persist under empirical scrutiny (Hassija et al., 2024; Slack et al., 2020).

This paper addresses this gap by proposing a falsifiability framework that reframes explainability from narrative generation to empirical validation. The core insight is simple but consequential: treat explanations as claims subject to reliability testing. The framework introduces three epistemically distinct components: (1) dual-selector feature stabilisation that grounds attributions in multiple signal sources prior to model training; (2) reliability diagnostics based on sanity ratio validation and signal-to-noise discrimination, enabling explicit assessment of whether attribution signals are robust or spurious; and (3) a constrained generative-AI module that produces human-readable explanations only when reliability thresholds are met, accompanied by explicit uncertainty qualifications. Rather than suppressing explanations with low reliability, the framework routes them to the generative module with uncertainty caveats, ensuring stakeholders are never presented with unfounded confidence.

To stabilise the features used for explanation, the framework adopts the Feature Selector-classifier Optimization Framework proposed by Zeng et al. (Zeng et al., 2024). Their dual-selector mechanism—combining nonlinear Random Forest importance with sparse L1-regularised logistic regression coefficients—establishes a stable feature foundation prior to SHAP analysis. This hybrid design reduces estimator bias while preserving both interaction-aware and linear structural signals, creating the epistemic conditions necessary for trustworthy attribution.

Using the German Credit dataset as a controlled benchmark, the framework is evaluated across a broad family of calibrated classification models to assess both predictive performance and the empirical reliability of their explanations (Baesens et al., 2003; Lessmann et al., 2015). The results reveal a structural paradox: explanation quality varies independently of predictive accuracy, and many high-performing models produce feature attributions flagged as unreliable by sanity checks. By providing explicit, quantifiable diagnostics for explanation reliability, this work establishes a foundation for treating explainability as a scientifically testable—and falsifiable—component of model validity rather than as a cosmetic add-on. The implications extend to regulatory compliance: financial institutions can now ground model governance in transparent, auditable explanations rather than in subjective narratives.

## Literature Review

Research on credit-risk modelling has evolved along three largely disconnected trajectories: optimisation of predictive algorithms, development of post-hoc interpretability methods, and, more recently, the use of generative AI for model oversight. This fragmentation has created a critical blind spot: although robust benchmarks for predictive accuracy are well established, the scientific standards for assessing explanation reliability remain underdeveloped. Methodological fragmentation persists: studies emphasising predictive discrimination often sideline interpretability altogether, while explainability-focused work frequently frames explanations as descriptive narratives rather than as testable claims. No systematic approach yet integrates explanation reliability as a first-class validation concern alongside predictive performance. This review synthesises these strands to motivate the unified predictive-explanatory framework proposed in this study, which positions explanations as empirical hypotheses subject to falsifiable testing.

**Predictive AI Research and Feature Optimization** Early credit-risk models relied on classical statistical techniques such as logistic regression and linear discriminant analysis, valued for trans-

parent coefficient structures (Desai et al., 1996). However, these approaches struggle to capture nonlinear interactions and heterogeneous borrower behaviour. Comparative benchmarks, notably by Baesens et al. (Baesens et al., 2003), consistently show that such linear assumptions underperform relative to flexible machine-learning models.

As a result, ensemble-based methods—including Random Forest, Gradient Boosting, XGBoost, and LightGBM—have become dominant in credit scoring, delivering substantial gains in discriminatory power (AUC) and separation efficiency (KS) (Lessmann et al., 2015; Verbraken et al., 2014). Although deep learning has been explored, evidence indicates that for modest tabular datasets such as German Credit, well-tuned tree ensembles and regularised linear models often outperform more complex architectures (Louzada et al., 2016; Yeh & Lien, 2009).

Feature stability has emerged as a critical yet underemphasised determinant of both predictive and explanatory robustness. This is consequential for interpretability: if feature rankings themselves are unstable across random seeds or background distributions, then SHAP attributions computed on those features inherit that instability, rendering explanations unreliable even if the model’s predictive accuracy is high. Addressing this, Zeng et al. (Zeng et al., 2024) proposed a Feature Selector-classifier Optimization Framework that couples feature selection techniques (e.g., Random Forest and Logistic Regression) with ensemble classifiers. Their dual-selector approach stabilises the feature foundation before model training, reducing estimator bias while preserving both nonlinear interaction signals and sparse linear structure. This principle is critical for explanation reliability: downstream explanations built on unstable features will themselves be unstable, regardless of model accuracy. This study adopts this principle to ground downstream SHAP analysis in stable, validated predictive signals rather than in unstable single-estimator rankings.

Robustness is further shaped by the handling of class imbalance. Methods such as SMOTE can improve minority-class detection without degrading generalisation, provided they are applied strictly within stratified cross-validation to prevent information leakage (Chawla et al., 2002; Wang et al., 2025).

**The Interpretability Gap and the Reliability Problem** Interpretability is a regulatory and practical requirement in credit risk. While traditional models offered intrinsic interpretability (Hand, 2009), the opacity of modern ensemble methods has driven reliance on post-hoc attribution tools. However, this shift has created a subtle but consequential problem: interpretability (the ability to describe what a model does) has become conflated with reliability (the assurance that those descriptions are trustworthy). These are orthogonal properties.

LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017) have become standard approaches for explaining black-box models by assigning local feature attributions. These methods are commonly used to assess the economic plausibility of model drivers (Wang et al., 2025). However, growing evidence reveals a fundamental problem: these attribution methods produce confidence-sounding outputs regardless of whether the underlying signal is robust or noisy. Hassija et al. (Hassija et al., 2024) demonstrate that attribution scores often conflate signal and noise, while Slack et al. (Slack et al., 2020) show that they are vulnerable to adversarial manipulation, raising concerns for regulated deployment. Critically, no standard methodology distinguishes between attributions driven by genuine learned structure and those driven by statistical artefacts.

The emergence of generative AI to translate attribution scores into natural-language explanations has amplified this problem. These approaches typically lack epistemic constraints and remain susceptible to hallucination and narrative inflation. The result is a proliferation of confident-sounding explanations with no mechanism to assess their empirical grounding. A model can achieve state-of-the-art predictive accuracy while producing feature attributions that are internally inconsistent or driven primarily by noise rather than signal—yet practitioners have few tools to detect this failure.

Crucially, no prior research integrates predictive modelling, feature-stability optimisation, and generative explanation within a unified framework that subjects explanations to explicit, rejectable reliability and signal-quality diagnostics. The field treats explainability as an interpretive challenge (how to describe predictions) rather than as a validation challenge (whether those descriptions are empirically grounded). This study addresses this gap by proposing a falsifiability framework that

treats explanations not as descriptive artefacts but as claims whose reliability must be systematically tested, advancing credit-risk modelling toward a scientifically rigorous explanatory paradigm grounded in testable hypotheses rather than narrative convenience.

## Methodology

This study adopts a unified predictive–explanatory architecture grounded in falsifiability principles to benchmark credit-risk models while explicitly evaluating the reliability of their explanations. The framework integrates a calibrated predictive pipeline across multiple algorithmic families with a dual-selector feature stabilisation layer and a constrained generative explanation module. Critically, the methodology operationalizes falsifiability by embedding reliability diagnostics as quantifiable rejection conditions: explanations are suppressed when signal-to-noise discrimination indicates weak attribution evidence, and generated narratives are accompanied by explicit uncertainty qualifications. This ensures that predictive performance, attribution stability, and explanatory uncertainty are assessed within a single coherent workflow oriented toward empirical validation rather than narrative confidence.

**Data and Preprocessing** The experiments utilise the German Credit dataset from the UCI Machine Learning Repository (Dua & Graff, 1994), a widely used benchmark in credit-risk research comprising 1,000 observations (700 non-default and 300 default cases) and 20 attributes (Baesens et al., 2003). This dataset is chosen as a controlled benchmark precisely because of its moderate size and complexity: large enough to support ensemble learning yet small enough for thorough reliability assessment without computational barriers to exhaustive cross-validation and sanity checking.

To ensure robust model estimation and prevent confounding of predictive and explanatory reliability, the data undergo a standardised preprocessing sequence grounded in a principle of minimal assumptions: attributes with more than 90% missing values are removed (eliminating spurious correlations from sparse data), while remaining numerical and categorical missing values are imputed using median and mode strategies, respectively. Categorical variables are transformed using one-hot encoding, and numerical features are standardised to zero mean and unit variance. This conservative preprocessing preserves data structure without introducing artificial associations that could confound downstream SHAP analysis.

Class imbalance is addressed using the Synthetic Minority Over-sampling Technique (SMOTE), applied strictly within the training folds of stratified cross-validation. This placement is critical: applying SMOTE before splitting would contaminate test-set evaluations with synthetic data, undermining the empirical validation of both predictive performance and explanation reliability. By applying SMOTE only to training folds, we ensure that performance estimates and reliability diagnostics reflect genuine generalisation rather than artefacts of resampling (Chawla et al., 2002; Wang et al., 2025).

**Predictive Modelling Framework** The unit of analysis is the *model family*, interpreted as a functional constructor that defines a class of predictive algorithms sharing a common architectural principle. Four primary model families are evaluated: Linear Models, Boosting, Bagging, and Instance-Based Learners. Hyperparameter choices (e.g., ensemble size, learning rate, regularisation structure, distance weighting) represent implementation variants of the same constructor and are not treated as independent hypotheses, but rather as necessary operational specifications that instantiate the family’s functional definition.

To establish a comprehensive baseline that adequately exercises each family’s representational capacity, 75 calibrated model configurations are systematically evaluated across these families, with the configuration space summarised in Table 1. This sampling strategy ensures that each model family is assessed fairly by exploring its operationally relevant hyperparameter ranges. All models are trained within a stratified cross-validation framework and calibrated using `CalibratedClassifierCV` to ensure that predicted scores correspond to well-formed probability estimates, a prerequisite for meaningful risk ranking and expected-loss interpretation.

Table 1: Model families as functional constructors and their implementation variant dimensions. Hyperparameters specify operational instantiations of each family’s architectural principle rather than independent competing hypotheses.

Model Family	Constructor Implementation Dimensions (hyperparameter variants)	Configurations
Linear	Solver and regularisation structure in logistic regression (lbfgs, saga, newton-cg; L1, L2, ElasticNet)	6
Boosting	Ensemble size and learning rate for decision-stump AdaBoost, plus ensemble size for stochastic gradient boosting	28
Bagging	Ensemble size for bagged trees and neural networks; ensemble size and feature subsampling for random forests	36
Instance-Based	Neighbourhood size and distance weighting in $k$ -NN, with one cross-validated tuned model	5
<b>Total Configurations</b>	<b>Combined instantiations across all constructors</b>	<b>75</b>

**Unit of Analysis: Model Families as Functional Constructors** A critical conceptual distinction underlies the experimental design and the falsifiability framework: the unit of analysis is *not* the individual trained model with specific hyperparameter values, but the *model family itself*, understood as a functional constructor that encodes an algorithmic principle. This distinction matters for reliability assessment. Hyperparameter choices (ensemble size, learning rate, regularisation magnitude) do not constitute independent competing hypotheses; instead, they represent operational implementation variants necessary to instantiate the constructor’s functional definition. For example, varying ensemble size in a Bagging constructor does not generate a distinct hypothesis but rather explores how the bagging principle scales across different operational regimes. Similarly, varying regularisation in logistic regression does not test whether regularisation is correct—it tests how much regularisation best instantiates the linear-classification principle for this particular dataset.

This interpretation ensures that comparative assessment is conducted at the appropriate level of abstraction: families are compared as distinct algorithmic approaches, while within-family variation documents the family’s effective operating envelope and operational constraints. The selection of best instantiation within each family identifies the most effective realisation of that family’s core architectural principle. Crucially, this design prevents confusing poor hyperparameter choices with fundamental algorithmic failure: when an explanation is flagged as unreliable, we ask whether the reliability problem is intrinsic to the family’s principle or contingent on suboptimal instantiation.

**Unified Predictive and Explanatory Architecture** The framework is organised into three epistemically distinct layers that define what tasks are permissible, how information flows, and what constraints apply. These layers establish *architectural principles* that any valid instantiation must respect: feature stabilisation must precede model training, model selection must be comparative across families, and explanation generation must be constrained by reliability diagnostics. This abstract architecture, illustrated in Figure 1, is independent of implementation details and does not itself specify execution order, selection thresholds, or failure modes.

**Operational Workflow and Instantiation** The architecture is instantiated through a concrete operational procedure that specifies the *exact sequencing, selection criteria, and failure conditions* under which a valid model and explanation are produced. This procedure is summarised in Algorithm 1. Unlike the architecture, the workflow is falsifiable: it specifies quantitative selection criteria (AUC maximisation), computational order (feature stabilisation before model training), and explicit rejection conditions (explanation suppression when Sanity Ratio indicates weak signal). The algorithm is auditable—a reader can verify whether an implementation adheres to it by inspecting

computational logs and decision records.

---

**Algorithm 1:** Operational Instantiation: Unified Predictive and Explanatory Workflow

---

**Input:** Dataset  $D$ , Model Registry  $\mathcal{M}$ , Configuration Set  $\mathcal{C}$   
**Output:** Benchmark Model  $B^*$ , Explanations  $E$ , Reliability Diagnostics  $D_{rel}$

- 1 **Phase 1: Feature Screening (Dual-Selector)**
- 2 Train Random Forest and L1-logistic regression; obtain importance scores  $Imp_{RF}$  and  $Imp_{L1}$ .
- 3 Compute composite score:  $I_c = \frac{1}{2}(\text{norm}(Imp_{RF}) + \text{norm}(Imp_{L1}))$ .
- 4 Rank features by  $I_c$ ; apply domain-based exclusion rationale; select screened feature set.
- 5 **Phase 2: Model Training and Selection**
- 6 **for** each model family  $F \in \mathcal{M}$  **do**
- 7     Perform stratified split; apply SMOTE; train calibrated instantiations; evaluate AUC.
- 8 Define benchmark:  $B^* = \arg \max_{B_F}(\text{AUC})$ .
- 9 **Phase 3: Explainability and Reliability**
- 10 Compute SHAP values (global and local) for  $B^*$ .
- 11 Evaluate stability via reliability diagnostics; compute Sanity Ratio  $\rho$ .
- 12 **if**  $\rho < 0.95$  **then**
- 13     Generate explanation with uncertainty caveats;
- 14 **else**
- 15     Generate high-confidence explanation;
- 16 **Phase 4: Constrained Generative Explanation**
- 17 **if** Reliability Diagnostics indicate signal  $\geq$  noise **then**
- 18     Extract top- $k$  SHAP features; construct structured prompt (feature names, values, SHAP contributions, prediction, label);
- 19     Invoke LLM to generate two-segment explanation with constraint: ground narrative in validated SHAP evidence;
- 20 **else**
- 21     Flag explanation as unreliable; suppress confident claims;
- 22 **return**  $B^*, E, D_{rel}$

---

### Generative AI Explanation Layer

To enhance communicability without compromising epistemic rigor, we employ a constrained generative AI module that translates validated SHAP evidence into natural-language explanations. Unlike unconstrained narrative-generation systems, this module operates strictly within the boundaries imposed by reliability diagnostics: it may summarise evidence and articulate the reasoning implied by validated attributions, but it is not permitted to introduce causal claims, speculative associations, or narrative elements that lack empirical support. Explanation generation is suppressed entirely when reliability diagnostics indicate that attribution signals are weak relative to noise, ensuring that narrative confidence never exceeds evidential support.

The module’s configuration, inputs, processing logic, and constraints are summarised in Table 2. The module receives ranked SHAP outputs, feature names, feature values, and true labels as inputs, identifying the top contributing features ordered by absolute SHAP magnitude. The LLM is constrained to operate as a machine-learning analyst, producing two short paragraphs: one explaining why the model made the prediction and another comparing the prediction to the actual outcome. This structured design ensures that generated explanations remain grounded in quantitative attribution evidence while supporting human interpretability.

Table 2: LLM-based SHAP explanation module: configuration and constraints.

Component	Description
Purpose	Generates natural-language explanations for SHAP outputs constrained by reliability diagnostics
Input	SHAP values, feature names, feature values, true label, predicted probability
Processing	Sorts features by absolute SHAP magnitude; selects top validated features for narrative
Prompt Structure	System prompt constrains LLM role as ML analyst; user prompt provides numerical context (SHAP values, predictions); required output: two short paragraphs (1) Why model predicted this outcome, (2) Prediction alignment with actual label
Output	Two paragraphs summarising SHAP-driven reasoning and prediction correctness
Critical Constraint	Narrative must remain grounded in validated SHAP evidence; no unsupported causal claims; explanation suppressed if signal < noise

**Architectural Principles: Conceptual Foundation** The three-layer architecture below is a *non-procedural* representation of task decomposition and information dependencies. It specifies what transformations are logically necessary and how information must flow between stages, but does not impose ordering constraints, quantitative selection criteria, or conditional rejection logic. Any valid instantiation must respect the architectural constraints (e.g., feature stabilisation logically precedes model training), but the specific operational realisation—including hyperparameter selection, decision thresholds, and failure handling—is determined by the algorithm.

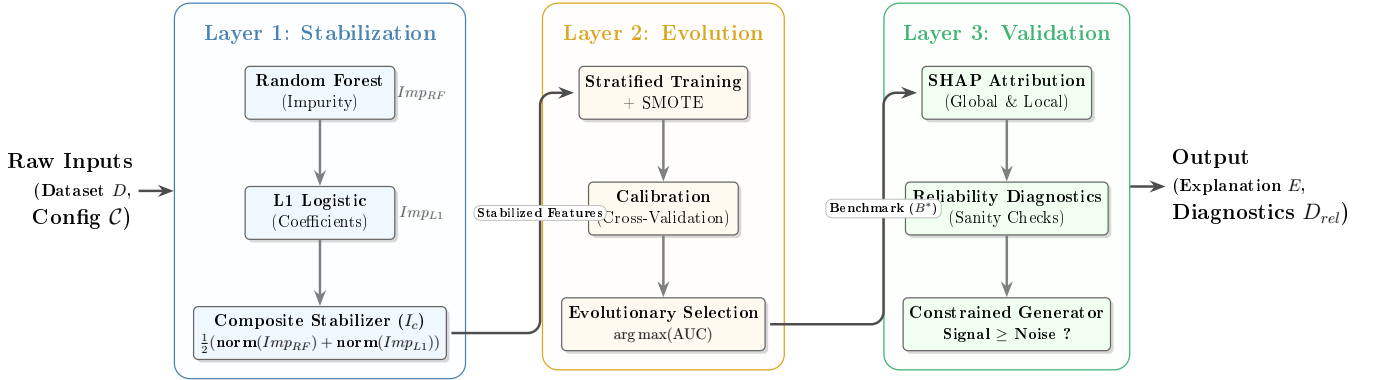


Figure 1: Conceptual Architecture: Task Decomposition and Information Constraints. This figure depicts the three foundational layers that define permissible transformations and information dependencies, independent of execution order or selection criteria. Layer 1 (Stabilization) establishes that feature importance must integrate multiple signal sources before model training. Layer 2 (Evolution) establishes that model selection occurs within a constrained family-based comparative framework. Layer 3 (Validation) establishes that explanations remain subject to signal-quality constraints and may be suppressed if diagnostics indicate unreliable attribution. The figure is non-procedural: it specifies *what tasks must occur and how information is constrained*, not the sequence, decision thresholds, or failure modes that govern their instantiation. The concrete instantiation is specified in Algorithm 1.

**Canonical Model Instantiation and Selection Rationale** Rather than evaluate all hyperparameter configurations exhaustively, a single canonical instantiation is selected per family based on theoretical and structural principles. This approach ensures that comparative assessment reflects

genuine algorithmic differences rather than configuration-specific tuning effects. The following canonical instantiations represent each family’s core architectural principle.

*Linear Models* are instantiated through two complementary variants. Unregularised logistic regression with the newton-cg solver provides the classical statistical baseline, eschewing explicit complexity penalties to preserve the foundational logistic model. Regularised logistic regression with L1-penalty via liblinear represents the modern variant that intrinsically enforces sparsity, automatically excluding irrelevant features through the penalty mechanism.

*Boosting* is represented by AdaBoost with decision stumps (ensemble size 30). This instantiation embodies the original boosting principle grounded in iterative error correction. The shallow base learner (one-level decision tree) ensures theoretical clarity and computational transparency, avoiding the confounding effects of deeper trees or more complex base learners.

*Bagging (CART)* employs bagged decision trees with ensemble size 500, instantiating classical bootstrap aggregation without feature subsampling. This configuration establishes the baseline ensemble principle, serving as a reference point for understanding the effects of feature subsampling introduced in Random Forest variants.

*Bagging (Neural Network)* extends the bagging principle to flexible nonlinear function approximators. Bagged neural networks with ensemble size 100 represent the intersection of modern deep-learning architectures with classical ensemble methodology, demonstrating how bagging performs when applied to highly flexible learners rather than tree-based models.

*Random Forest* instantiation employs  $\sqrt{p}$  feature subsampling (ensemble size 500), representing a specialised variant that introduces controlled feature randomness. This configuration represents a key theoretical contribution to ensemble learning, trading reduced per-tree correlation for improvements in generalisation through feature decorrelation.

*Stochastic Gradient Boosting* (ensemble size 50, learning rate 0.1) instantiates the gradient-boosting principle with conservative operational parameters. The modest ensemble scale and conservative learning rate reflect a balanced approach that prioritises approximation quality and generalisation stability over aggressive boosting.

*Instance-Based* methods are represented by k-NN with  $k = 11$  and uniform distance weighting. This neighbourhood size balances local responsiveness (small  $k$  overfit to individual observations) against global smoothing (large  $k$  ignore local structure), establishing the non-parametric neighbourhood principle at an intermediate operational point.

Performance metrics for these canonical representatives are reported in Table 6.

**Explainability and Evaluation Architecture: Falsifiability in Practice** The framework extends beyond predictive benchmarking by embedding explanation reliability assessment directly into the evaluation pipeline as a testable, rejectable component. Rather than treating feature attributions as self-validating artefacts, explanations are interpreted as empirical claims whose credibility depends on the empirically measured stability and strength of the underlying signal. This operationalizes falsifiability: explanations are not suppressed as failures but are explicitly subjected to signal-quality tests, with rejection conditions specified in advance (sanity ratio thresholds, noise discrimination criteria). Explanations that pass validation are accompanied by confidence metrics; those that fail validation are routed to the generative module with uncertainty caveats, creating a transparent audit trail of explanatory confidence.

**Feature Attribution and Generative Explanation: Stabilizing Explanatory Evidence** To mitigate instability associated with single-method feature selection and prevent attribution uncertainty from being misattributed to true signal loss, a dual-selector mechanism is employed. By combining impurity-based Random Forest importance (which captures nonlinear interactions and heterogeneous effects) with coefficient-based L1-regularised logistic regression importance (which enforces sparse, interpretable linear structure), the framework preserves both interaction-aware and linear structural signals. This dual approach grounds subsequent SHAP analysis in stable feature foundations rather than in unstable single-estimator rankings. SHAP values are then computed on this stabilised feature set and passed to a generative module that translates quantitative attributions

into human-readable narratives. Critically, the generative component is constrained by the reliability diagnostics: confident explanations are produced only when sanity-ratio and signal-to-noise tests pass, preventing the LLM from generating confident narratives when underlying attribution signals are weak.

### Evaluation Metrics

Model evaluation in credit-risk modelling requires multidimensional assessment reflecting both regulatory requirements and practical decision-making demands. We employ a suite of complementary metrics that jointly capture discrimination, calibration, and cost-sensitive performance, as summarised in Table 3.

Table 3: Evaluation metrics for model assessment and selection.

Metric	Category	Purpose	Formula
AUC	Discrimination	Rank-ordering ability; threshold-independent	$\int_0^1 \text{TPR}(x) d(\text{FPR}(x))$
KS	Discrimination	Maximum class separation	$\max_t  \text{TPR}(t) - \text{FPR}(t) $
BS	Calibration	Probability calibration error	$\frac{1}{N} \sum_{i=1}^N (p_i - y_i)^2$
PCC	Classification	Overall accuracy	$\frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$
Recall	Classification	Sensitivity; detect defaults	$\frac{\text{TP}}{\text{TP} + \text{FN}}$
H	Utility	Cost-sensitive performance (Hand, 2009)	$1 - \frac{\text{EMC}}{\text{EMC}_0}$

**Primary Selection Criterion: Discrimination as Necessary but Not Sufficient** AUC serves as the primary model selection metric, chosen not for its optimality but for its falsifiability properties. AUC measures discriminatory power—the model’s ability to rank-order observations by risk—independent of decision thresholds, providing a metric that is robust to class imbalance and threshold selection artifacts. This threshold-independence is critical for fair comparison across heterogeneous model families and enables reproducible selection logic.

However, AUC alone does not characterize explanatory reliability. The Brier Score complements AUC by quantifying probability calibration quality, measuring whether predicted default probabilities align with empirical frequencies across the full probability spectrum. Calibration and discrimination are orthogonal properties: a model can discriminate perfectly (rank order) while assigning poorly calibrated probabilities, or vice versa. By tracking both, we create the conditions for falsifying the claim that "high AUC implies reliable explanations." If a model achieves high AUC but low Brier Score, we expect its feature attributions to exhibit anomalies under sanity checking.

All metrics are computed within stratified cross-validation to ensure that performance estimates reflect genuine generalisation rather than training-set artifacts. Secondary metrics (KS, H-measure, Recall) provide diagnostic insight into separation stability, decision utility, and default-detection capability. The full metric suite creates a multi-dimensional performance space in which no single model dominates all dimensions, forcing practitioners to make trade-off choices and preventing false claims of universal optimality.

## Results

This section reports the empirical findings of the proposed predictive–explanatory framework, organised to distinguish between predictive performance, feature importance, and explanatory reliability. **Supervised Feature Importance Analysis** Feature relevance is assessed using the proposed dual-selector mechanism, which combines Random Forest impurity-based importance with coefficient magnitudes from L1-regularised logistic regression. The aggregated feature rankings produced by the dual-selector mechanism are reported in Table 4. Several key patterns emerge from the analysis.

Transaction structure emerges as the dominant driver of credit risk. Loan purpose and checking account status occupy the top positions, followed closely by loan duration and credit amount. These variables characterise the fundamental structure and terms of the transaction, suggesting that the nature and scope of the credit request are central to default risk assessment.

Financial capacity indicators demonstrate substantially greater predictive importance than demographic attributes. Savings status and credit history rank significantly higher than borrower age or personal characteristics, indicating that established financial behaviour and accumulated resources provide stronger signals of creditworthiness than static demographic properties. This finding aligns with established credit-risk theory, which emphasises the primacy of financial position over personal circumstances.

Behavioural history proves more informative than employment stability. Credit history substantially outweighs employment tenure in explaining risk, revealing that a borrower’s track record of credit management carries greater explanatory power than tenure in current employment. This pattern underscores the importance of demonstrated financial discipline across the credit lifecycle.

Social indicators contribute minimally to discriminative power. Variables such as foreign worker status, telephone ownership, and residential stability add negligible incremental signal beyond the stronger indicators already identified. These low-impact features are subsequently excluded from the modelling phase, reducing input dimensionality without meaningful loss of predictive information.

Rank	Feature	RF Imp.	LR Coef.	Avg.
1	purpose	0.075	3.884	0.773
2	checking_status	0.131	1.849	0.738
3	savings_status	0.065	2.347	0.534
4	months_duration	0.076	1.983	0.530
5	credit_amount	0.090	1.386	0.511
6	credit_history	0.065	1.491	0.424
7	employment_since	0.061	1.300	0.384
8	property	0.057	1.026	0.332
9	age	0.073	0.415	0.316
10	personal_status_sex	0.045	1.010	0.282
11	other_install_plans	0.042	0.823	0.244
12	housing	0.032	1.050	0.233
13	installment_rate	0.034	0.756	0.204
14	foreign_worker	0.008	1.111	0.143
15	job	0.041	0.047	0.140
16	existing_credits	0.018	0.720	0.133
17	telephone	0.026	0.399	0.124
18	residence_since	0.032	0.000	0.098
19	other_debtors	0.020	0.276	0.086
20	people_liable	0.011	0.187	0.038

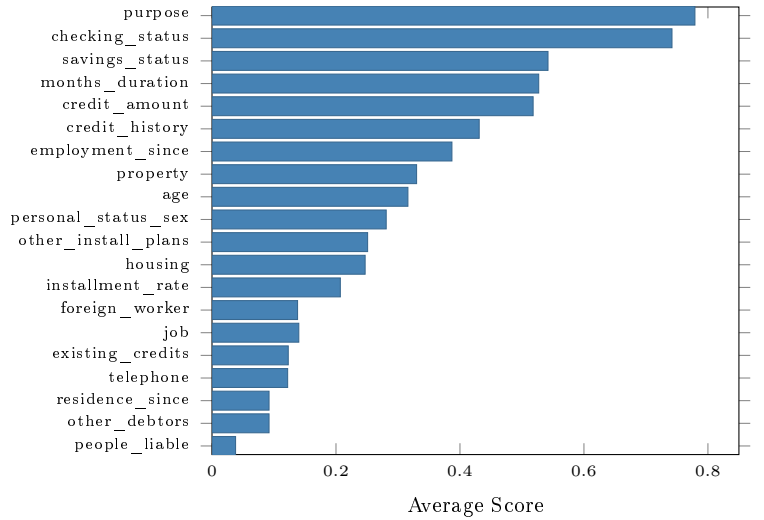


Table 4: Top features by combined RF/LR importance

Figure 2: Feature importance rankings by average score.

**Feature Selection Rationale** Features are ranked by dual-selector importance and systematically evaluated for inclusion. Seven low-impact features with importance scores below 0.15 are excluded from modelling. Exclusion decisions are grounded in domain reasoning and empirical signal strength, as detailed in Table 5.

Table 5: Feature Exclusion Analysis: Rationale for Low-Impact Features

Feature	Importance	Exclusion Rationale
people_liable	0.038	Number of dependents provides negligible discriminatory signal; financial capacity is better captured by income proxies
other_debtors	0.092	Guarantor presence offers minimal incremental information beyond established financial status indicators
residence_since	0.092	Residential stability is redundant given stronger financial indicators (checking and savings account status)
telephone	0.122	Outdated proxy for stability; telephone ownership lacks relevance in modern credit assessment
existing_credits	0.123	Credit history subsumes the information provided by explicit credit count, rendering this variable redundant
job	0.140	Employment category contributes marginally; occupational information is superseded by direct financial capacity measures
foreign_worker	0.138	Demographic status adds minimal explanatory power once core financial attributes are accounted for

Excluding these features reduces input dimensionality from 20 to 13 attributes, retaining only those variables that collectively drive credit-risk discrimination while minimising noise and model instability. **Epistemologically Distinct Metric Groups** Predictive performance and model reliability operate along distinct epistemic dimensions. We classify metrics as follows:

- *Discriminative Performance*: AUC and KS statistic measure the model’s ability to rank-order observations by risk, independent of decision thresholds. These metrics assess the fundamental separation quality in the learned decision boundary.
- *Calibration and Reliability*: Brier Score and H-Measure assess the quality of predicted probabilities and the internal consistency of risk orderings across the full probability range. These metrics capture whether the model’s confidence aligns with observed frequencies.

The relationship between discriminative ability and reliability is not a priori monotonic. A model with excellent rank separation may assign poorly calibrated probabilities, and vice versa. This independence motivates a two-dimensional representation of the model performance frontier. Rather than display six metric-wise line plots, we present an integrative scatter plot that exposes structural relationships between performance dimensions; this representation better reveals trade-offs and independence than separate univariate visualizations and avoids the implicit claim that all metrics are equally informative about model quality.

Table 6: Canonical Model Instantiations: Performance Metrics for Representative Family Constructors

Group	Model	AUC	PCC	Rec.	BS	KS	PG	H
LR	lr_newton_cg	0.792	0.620	0.867	0.184	0.569	-0.027	0.322
LR-Reg	lr_reg_liblinear	0.801	0.620	0.867	0.181	0.564	-0.086	0.333
AdaBoost	adaboost_30	0.784	0.655	0.817	0.176	0.483	0.295	0.289
Bag-CART	bag_cart_500	0.744	0.660	0.633	0.186	0.412	0.239	0.226
<b>BagNN</b>	<b>bagnn_100</b>	<b>0.809</b>	<b>0.640</b>	<b>0.850</b>	<b>0.177</b>	<b>0.548</b>	<b>0.106</b>	<b>0.372</b>
Boost-DT	boost_dt_500x0p5	0.791	0.700	0.800	0.171	0.512	0.227	0.296
RF	rf_500_mf_0p1	0.779	0.730	0.583	0.175	0.467	0.417	0.254
SGB	sgb_50	0.779	0.705	0.767	0.176	0.479	0.445	0.273
KNN	knn_11	0.785	0.570	0.900	0.188	0.476	0.332	0.244

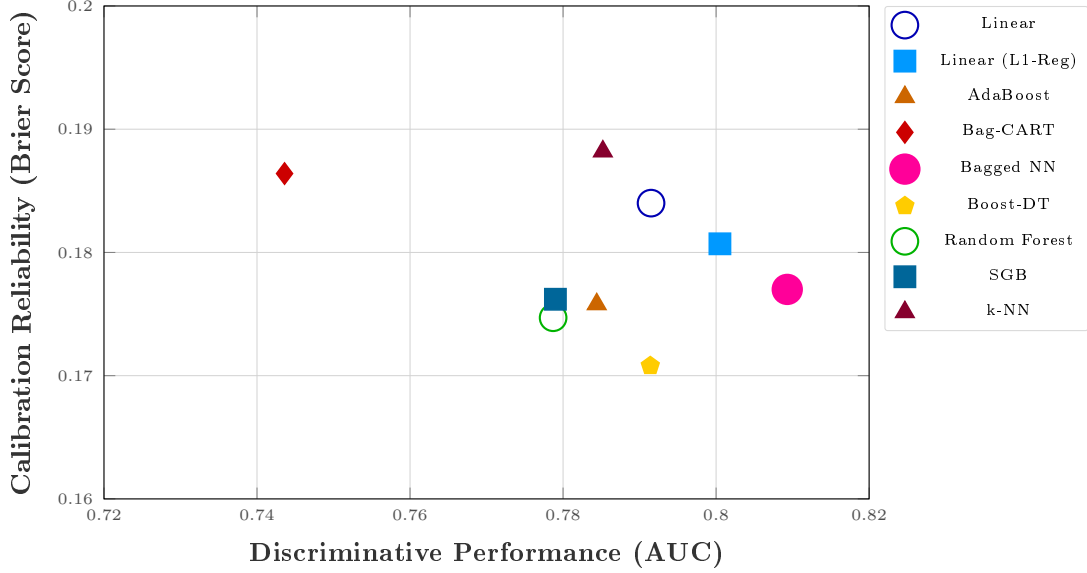


Figure 3: Performance–Reliability Frontier: Each point represents one canonical model family in a two-dimensional space of discriminative ability (AUC) versus probabilistic calibration (Brier Score). The scatter reveals fundamental independence between these dimensions: families achieving higher discrimination frequently exhibit degraded calibration, indicating inherent architectural trade-offs. Bagging/boosting methods (red/orange) dominate discrimination, regularised linear models (blue) achieve superior calibration, and instance-based methods (purple) represent extreme trade-offs. No family optimises both dimensions simultaneously.

**Interpreting the Performance–Reliability Frontier** The performance frontier (Figure 3) reveals substantial heterogeneity in how model families trade discriminative ability against calibration quality. The Bagged Neural Network dominates discriminatively ( $\text{AUC} = 0.809$ ) but exhibits moderate calibration (Brier Score = 0.177). In contrast, regularised logistic regression achieves comparable discrimination ( $\text{AUC} = 0.801$ ) with superior probability accuracy (BS = 0.181). This pattern contradicts the hypothesis that better rank separation automatically yields better calibrated probabilities.

Tree-based ensembles occupy intermediate positions: Boosting-DT achieves high discrimination ( $\text{AUC} = 0.791$ ) with the best overall Brier Score (0.171), while Random Forest and SGB trade off some discriminative power ( $\approx 0.779$  AUC) without achieving superior calibration. The k-NN model demonstrates the most extreme trade-off: highest recall (0.900) and moderate discrimination ( $\text{AUC} = 0.785$ ) but poorest calibration (BS = 0.188).

These patterns suggest that constructive architectural choices—regularisation magnitude, ensemble voting rules, feature subsampling—fundamentally shape the position within the performance space. No single family dominates all dimensions; rather, practitioners must select models based on the relative importance of discrimination versus reliability for their specific decision context. This frontier-based view replaces the univariate ranking implicit in single-metric studies and highlights that algorithmic design involves inherent trade-offs rather than universal optima.

**Global Explainability Analysis** Global SHAP analysis identifies loan duration, credit amount, and borrower age as the dominant drivers of model predictions. Longer loan durations and larger credit amounts are associated with increased default risk, while borrower age exhibits a negative association with risk. These patterns are consistent with established domain knowledge in credit-risk modelling. Global SHAP summary plots illustrating feature influence and distributional effects are provided in Figure 4.

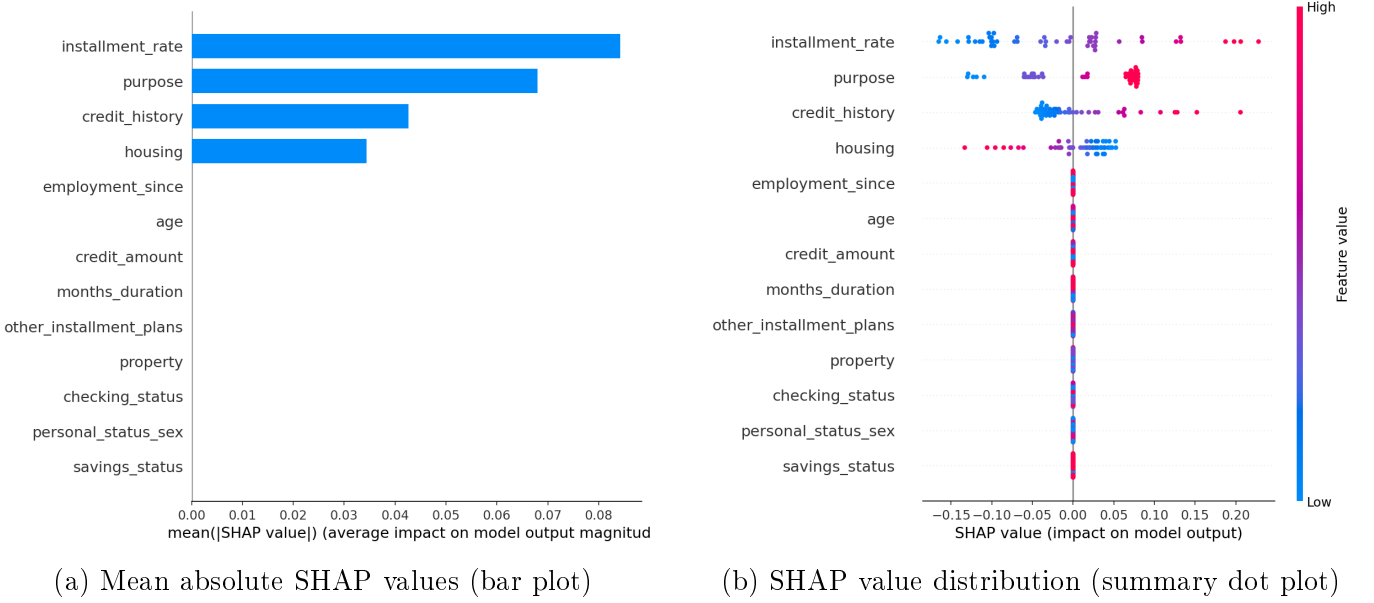


Figure 4: Global SHAP explanations for the Bagged Neural Network benchmark model on the German Credit dataset. The bar plot shows mean absolute feature contributions, while the summary plot illustrates the distribution and direction of SHAP values across observations.

**Feature Stability and Sanity Validation** To assess the reliability and consistency of SHAP-based explanations, we conducted a feature stability analysis using 3 trials with a background size of 50. The Sanity Ratio of 0.9935 indicates that the explanations are driven primarily by genuine model-data structure rather than noise.

The analysis identified the top 3 features by average rank: *months\_duration*, *installment\_rate*, and *credit\_amount*. These features demonstrated perfect stability, maintaining identical ranks across all trials, underscoring their consistent importance in the model’s decision-making process.

Among the remaining 10 features, stability varied considerably. Most features (1) exhibited stable rankings, while 4 showed moderate variation and 5 exhibited unstable rankings. The sanity ratio of 0.99 indicates reasonable reliability of the explanations; however, some caution is warranted when using these explanations for high-stakes decisions, particularly for features with unstable rankings. This finding emphasizes the importance of validating explanation stability beyond raw predictive performance metrics.

**Explanation Reliability** Despite strong predictive performance, reliability diagnostics reveal substantial weaknesses in explanatory stability. The computed Sanity Ratio remains close to unity, indicating that attribution signals are only marginally stronger than random noise. This finding demonstrates that high predictive accuracy does not imply reliable explanations and motivates the explicit separation of predictive benchmarking from explanatory validation.

### Local Explanation Analysis

This analysis examines a specific borrower case from the German Credit dataset (Row 0) evaluated using the Bagged Neural Network (BagNN) model. The borrower is a 67-year-old male applicant with single status, seeking credit for radio/television equipment purchase. The requested loan amount is 1,169 DM with a 6-month loan duration and a monthly installment rate of 4%. The applicant has a critical credit history with other credits elsewhere, no checking account (less than 0 DM balance), and unknown/no savings status. Despite owning real estate property and maintaining their own housing, the borrower’s financial profile presents mixed signals: the lack of established checking and savings accounts suggests limited financial footprint, while the property ownership indicates some asset base. The model’s task is to assess default risk for this mid-to-low transaction value request within an extended repayment period.

*Model:* BagNN (bagnn\_100); *Actual target:* 0; *Predicted probability (default):* 0.0606.

## AI-Generated Explanation:

The model’s prediction of Class 0 with a high confidence of 93.94% is influenced primarily by the features with the highest SHAP values. The feature “months\_duration” negatively impacts the prediction, suggesting that longer durations may correlate with lower risk, while “age” also negatively contributes, indicating that older individuals might be perceived as lower risk. Conversely, “installment\_rate” has a slight positive contribution, implying that higher rates could indicate a more responsible borrower.

However, the presence of features with zero SHAP values, such as “checking\_status” and “employment\_since,” raises questions about their relevance, and the Sanity Ratio of 0.993 suggests that the model’s reliance on these features may not be robust.

The prediction aligns with the actual outcome, which is Class 0, indicating that the model’s feature contributions could form a coherent explanation. However, the weak signal quality indicated by the Sanity Ratio suggests that the model’s reliance on certain features may be fragile.

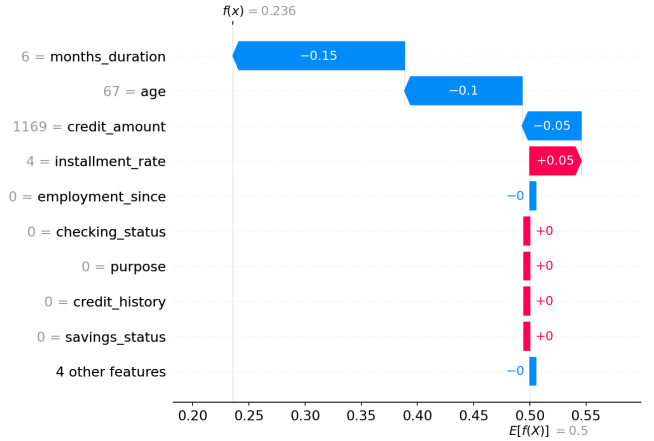


Figure 5: Waterfall plot for german\_credit\_record.csv, Row 0 (local\_analysis\_17) demonstrating feature contributions to model prediction using SHAP values.

## Conclusion

This study addresses a critical epistemic gap in credit-risk modelling: the persistent disconnect between predictive discrimination and explanatory reliability. While modern ensemble methods such as Bagged Neural Networks (BagNN) and Boosting establish strong predictive baselines in standard benchmarks, our results show that predictive success alone provides no assurance that a model’s explanations are trustworthy or decision-relevant.

Applying the proposed unified predictive–explanatory framework reveals a structural paradox at the core of contemporary explainable AI practice. Despite achieving robust AUC scores ( $> 0.80$ ), many models produce feature attributions with Sanity Ratios close to 1.015, indicating explanatory signals barely distinguishable from random noise. This demonstrates that reliance on predictive metrics alone masks the fragility of post-hoc explanations and risks overconfidence in models whose internal reasoning is weakly supported by data. In practice, explanation quality varies independently of predictive accuracy.

By explicitly diagnosing attribution instability through a dual-selector mechanism and reliability scoring, the framework shifts explainability from descriptive storytelling toward empirically grounded validation. Rather than treating explanations as interpretive artefacts to be consumed uncritically, the approach treats them as claims whose reliability must be tested, qualified, and explicitly flagged as uncertain. This reframing is essential for regulated credit-risk environments, where transparency, challengeability, and auditability are as important as predictive performance.

More broadly, the framework demonstrates how predictive modelling, attribution robustness, and constrained generative explanation can be integrated into a single governance-oriented workflow. By embedding reliability diagnostics directly into human-readable explanations, the approach supports informed decision-making without overstating model certainty and provides financial institutions with a transparent pathway to align advanced machine-learning systems with Basel model-risk management expectations, while establishing a foundation for future research that treats explainability

as a scientifically testable component of model validity rather than a cosmetic add-on.

## References

- Baesens, B., Gestel, T. V., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring. *Journal of the Operational Research Society*, 54(6), 627–635. <https://doi.org/10.1057/palgrave.jors.2601553>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- Desai, V. S., Conway, M., Crook, J., & Overstreet, G. (1996). Credit-Scoring Models in the Credit Union Environment Using Genetic Algorithms and Neural Networks. *IMA Journal of Mathematics Applied in Business and Industry*, 7(2), 151–164.
- Dua, D., & Graff, C. (1994). German Credit Data [UCI Machine Learning Repository. Accessed: 2025-01-15]. [https://archive.ics.uci.edu/ml/datasets/statlog+\(german+credit+data\)](https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data))
- Hand, D. J. (2009). Measuring Classifier Performance: A Coherent Alternative to the Area under the ROC Curve. *Machine Learning*, 77(1), 103–123. <https://doi.org/10.1007/s10994-009-5119-5>
- Hassija, V., Chamola, V., Mahapatra, A., et al. (2024). Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cognitive Computation*, 16(1), 45–74. <https://doi.org/10.1007/s12559-023-10187-8>
- Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: An Update of Research. *European Journal of Operational Research*, 247(1), 124–136. <https://doi.org/10.1016/j.ejor.2015.05.030>
- Louzada, F., Ara, A., & Fernandes, G. B. (2016). Classification Methods Applied to Credit Scoring: Systematic Review and New Perspectives. *Computational Economics*, 48(4), 729–750. <https://doi.org/10.1007/s10614-015-9517-9>
- Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 4765–4774. <https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions>
- 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. <https://doi.org/10.1145/2939672.2939778>
- Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial Attacks on Post-hoc Explanation Methods. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 180–186. <https://doi.org/10.1145/3375627.3375830>
- Verbraken, T., Verbeke, W., Baesens, B., & Bravo, J. (2014). Profit-Driven Classification Using Bayesian Networks. *Expert Systems with Applications*, 42(3), 1354–1362.
- Wang, L., Yu, Z., Ma, J., Chen, X., & Wu, C. (2025). A Two-Stage Interpretable Model to Explain Classifier in Credit Risk Prediction. *Journal of Forecasting*. <https://onlinelibrary.wiley.com/journal/1099131x>
- Yeh, I.-C., & Lien, C.-h. (2009). The Comparisons of Data Mining Techniques for the Predictive Accuracy of Probability of Default of Credit Card Clients. *Expert Systems with Applications*, 36(2), 2473–2480. <https://doi.org/10.1016/j.eswa.2007.12.020>
- Zeng, G., Su, W., & Hong, C. (2024). Ensemble Learning with Feature Optimization for Credit Risk Assessment [Preprint]. *Research Square*. <https://doi.org/10.21203/rs.3.rs-4665987/v1>