Integrating Predictive and Generative AI for Explainable Credit Risk

Ravi Kumar Jain[1] and Anju Goswami[2]

[1]Indian Institute of Foreign Trade, Delhi, India
[2]Indian Institute of Foreign Trade, Delhi, India

Abstract

Credit-risk models continue to improve in predictive accuracy, but their explanations often fail to show why the model behaves the way it does. Modern machine-learning methods can capture complex patterns, but many explanation tools only provide numerical scores that may not reflect the model's real reasoning. This paper proposes a predictive–explanatory framework that treats explainability as proposing explanations and testing whether they truly hold, using stability checks, counterfactual tests, and a controlled generative-AI layer that outputs only validated reasoning. The framework introduces a dual-selector method for stable feature attribution and applies three validation tests—rank stability, label randomisation, and counterfactual validity—to filter out unreliable explanations. Only explanations that pass these tests are converted into clear, human-readable narratives through a constrained generative-AI step. Using the German Credit dataset, we evaluate several classification models and apply the framework to assess the quality and stability of the explanations they generate. The results show that the explanations produced by this framework are stable, testable, and closely aligned with the underlying structure of each model. The approach supports explainable-AI practices suitable for Basel model-risk governance and can be extended to causal discovery, domain-specific fine-tuning, and multi-dataset evaluation.

Keywords: Credit-risk assessment; Explainable AI; Stability testing; Counterfactual validation; Feature attribution; Generative AI explanations; Basel model-risk governance.

1. Introduction

Credit-risk assessment is a core function within banking and contributes to institutional and systemic stability (Altman, 1968; Beaver, 1966). Early modelling efforts in credit

scoring adopted statistical methods aimed at producing interpretable and operationally consistent predictions; logistic regression was one widely used example within this broader class of approaches (Ohlson, 1980; Wiginton, 1980). Subsequent advances introduced a wider range of predictive algorithms, including support vector machines (Cortes & Vapnik, 1995), k-nearest neighbour classifiers (Cover & Hart, 1967), and artificial neural networks (Cybenko, 1989; Hornik et al., 1989; Rumelhart et al., 1986), each contributing incremental improvements in discriminatory power. More recently, ensemble methods such as Random Forests and Gradient Boosting (Breiman, 2001; Friedman, 2001) and deep-learning architectures (LeCun et al., 2015) have further expanded the modelling landscape, delivering notable gains in predictive accuracy. Recent advances—including transformer-based architectures, graph neural networks, and foundation-model adaptations for tabular credit data—have pushed this frontier even further (Borisov et al., 2022; Gorishniy et al., 2021; Huang et al., 2020).

As modelling complexity has increased, predictive architectures have become increasingly opaque and susceptible to overfitting—a concern well documented in the literature on black-box modelling (Louzada et al., 2016; Slack et al., 2020; Yeh & Lien, 2009). This opacity prevents analysts from determining whether predictions arise from genuine economic behaviour or artefactual correlations (Hassija et al., 2024; X. Wang et al., 2025), directly challenging the governance structures mandated by the Basel framework. Although regulators require that decision logic be transparent, stable, and empirically verifiable (Basel Committee on Banking Supervision, 2011; European Banking Authority, 2021), institutional practice often defaults to high-level documentation rather than concrete insight, creating a widening gap between compliance expectations and methodological reality. While substantial progress has been made in feature attribution and interpretable design, the field still lacks an integrated framework that validates explanatory fidelity alongside predictive performance (Caruana et al., 2015; Guidotti et al., 2018; Rudin, 2019).

To navigate this challenge, the field of Explainable Artificial Intelligence (XAI) has developed distinct pre-hoc and post-hoc interpretability paradigms. Pre-hoc approaches prioritise inherently transparent "glass-box" architectures—such as logistic regression, decision trees, or k-nearest neighbours—that reveal their decision logic directly (Doshi-Velez & Kim, 2017; Rudin, 2019). Yet in credit-risk contexts, this intrinsic interpretability often comes at a measurable cost in predictive capacity, a compromise that is difficult to justify in high-stakes portfolios where even small accuracy gains yield material economic benefits (Khandani et al., 2010; Lessmann et al., 2015). Conversely, post-hoc methods—ranging from visualisation tools like ICE and ALE (Apley & Zhu, 2020; Goldstein et al., 2015) to attribution frameworks such as LRP, LIME, and SHAP (Bach et al., 2015; Lundberg & Lee, 2017; Ribeiro et al., 2016; Wachter et al., 2018)—analyse trained models without altering their internal structure, thereby preserving predictive performance.

Although their adoption has grown rapidly due to their utility in identifying influential features and decision boundaries (Li & Wu, 2024; C. Wang et al., 2025; Yildirim & Kulekci, 2021), emerging evidence shows that these methods often provide descriptive summaries rather than robust explanations. Crucially, their outputs can exhibit high sensitivity to perturbations, adversarial manipulation, and model randomness (Adebayo et al., 2018; Agarwal et al., 2022; Alvarez-Melis & Jaakkola, 2018; Slack et al., 2020). This distinction—between describing a model's behaviour and explaining its underlying reasoning—remains a central unresolved challenge in the governance of modern credit-risk systems.

Against this backdrop, we introduce a predictive–explanatory framework designed to address the challenge left unresolved by existing XAI methods: determining whether a model's apparent reasoning reflects genuine economic structure or merely artefactual correlations (Louzada et al., 2016; Slack et al., 2020; Yeh & Lien, 2009). As a first step, the framework evaluates the reliability of model predictions themselves by subjecting paired performance estimates to statistical significance tests widely used in comparative credit-scoring research. These include the Wilcoxon signed-rank test for paired performance differences, McNemar's test for correlated classification outcomes, and DeLong's test for comparing AUCs (DeLong et al., 1988; Demšar, 2006; McNemar, 1947). Together, these tests establish whether observed gains stem from genuine improvements in predictive capacity rather than sampling variability or model noise.

Building on this predictive reliability layer, the framework then treats each explanatory claim as a hypothesis about the underlying credit-risk mechanism and exposes it to structured refutation through rank-stability checks, label-randomisation tests, counterfactual-validity assessments (Wachter et al., 2018), and model-sanity evaluations (Adebayo et al., 2018). These procedures are motivated by extensive evidence showing that commonly used attribution approaches can be unstable, easily manipulated, or sensitive to spurious structure (Agarwal et al., 2022; Alvarez-Melis & Jaakkola, 2018; Ribeiro et al., 2016). By enforcing falsifiability at the explanatory level, the framework distinguishes genuine domain-aligned reasoning from merely descriptive patterns.

Only after candidate explanations survive both layers of empirical scrutiny are they passed to a constrained generative-AI layer, whose role is limited to rendering the validated reasoning in a coherent, communicable, and audit-ready form rather than generating new inferential content (X. Wang et al., 2025). By grounding both prediction and explanation in explicit empirical tests rather than visual summaries or correlation-based attributions, the framework responds directly to long-standing governance expectations that credit-risk models exhibit transparent, stable, and evidence-aligned decision logic (Basel Committee on Banking Supervision, 2011; European Banking Authority, 2021).

Using four widely studied real-world credit datasets from the UCI Machine Learning Repository—the Bank Marketing, Credit Approval, German Credit, and Statlog (Aus-

tralian Credit Approval) datasets—we assess multiple classification models and evaluate the robustness, stability, and empirical validity of the explanations they produce. To ensure that predictive comparisons are not confounded by irrelevant or redundant variables, the analysis incorporates a feature-optimization procedure that systematically searches over selector–classifier combinations to obtain maximally informative feature subsets (Zeng et al., 2024). In settings where the target variable exhibits class imbalance, we additionally employ the Synthetic Minority Oversampling Technique (SMOTE) to construct balanced training distributions, thereby reducing bias toward majority-class predictions and improving the reliability of downstream model comparisons (Chawla et al., 2002; He & Garcia, 2009). These datasets span diverse credit environments, feature compositions, and class distributions, thereby enabling a broader and more reliable assessment of model behaviour across heterogeneous economic and demographic contexts. Prior studies demonstrate that evaluating models across multiple independent datasets provides stronger evidence of generalisability than relying on any single benchmark, particularly in credit-risk settings where feature interactions, sparsity, and population drift can differ substantially across domains (Quan & Sun, 2024).

By grounding our analysis in this multi-dataset structure, the framework supports a structured comparison of predictive performance and explanatory reliability rather than relying solely on accuracy from a single source. The results demonstrate that a falsifiability-driven approach to explainability can strengthen model-risk governance under Basel expectations while laying the groundwork for future research on causal-structure discovery, domain-specific generative-AI alignment, and cross-dataset validation in heterogeneous credit environments.

This paper establishes that the identity of the best-performing credit-risk model is not invariant under conventional performance metrics: when explanation stability and reliability constraints are imposed, model rankings are systematically altered, with models previously judged optimal frequently replaced by more stable alternatives. This effect is observed consistently across heterogeneous datasets, demonstrating that explanation stability is an independent and necessary criterion of model quality.

2. Literature Review

The evolution of credit-risk modelling is best understood through the lens of large-scale benchmarking studies that have systematically evaluated classifier performance over time. Early comparative research, most notably by (Baesens et al., 2003), established that while traditional statistical methods like logistic regression offered operational stability, they were increasingly outperformed by machine-learning algorithms capable of capturing nonlinear borrower behaviour. This finding was rigorously updated by (Lessmann et al., 2015), who conducted a comprehensive evaluation of 41 algorithms across multiple data-

sets. Their results confirmed that ensemble methods—specifically Random Forests and Gradient Boosting—had become the new state-of-the-art, consistently delivering superior predictive accuracy compared to individual classifiers. Subsequent systematic reviews have reinforced this dominance, highlighting that the field has shifted decisively toward these high-performing architectures, even as they introduce new challenges regarding opacity and interpretability (Louzada et al., 2016).

However, the performance of these predictive algorithms is heavily contingent on the quality of the input data and the rigorous handling of feature spaces. Credit-risk datasets are characteristically imbalanced, a structural property that can bias standard classifiers toward the majority class. To mitigate this, techniques such as the Synthetic Minority Oversampling Technique (SMOTE) have become standard practice for constructing balanced training distributions that improve minority-class detection (Chawla et al., 2002; He & Garcia, 2009). Beyond balancing, recent research emphasizes that predictive gains in ensemble learning are maximized when paired with systematic feature optimization. As demonstrated by (Zeng et al., 2024), optimizing feature subsets not only reduces computational complexity but also enhances model stability by eliminating redundant or noisy variables, a critical step often overlooked in standard pipelines.

Building on these foundations, the modern modelling landscape has expanded to include not only tree-based ensembles but also deep-learning architectures adapted for tabular data. While algorithms like Random Forests and Gradient Boosting Machines remain the industry workhorses (Breiman, 2001; Friedman, 2001), recent innovations have introduced specialized deep architectures, such as TabTransformer and adapted ResNets, which aim to leverage representation learning for structured financial data (Gorishniy et al., 2021; Huang et al., 2020). Comprehensive surveys suggest that while deep neural networks are closing the performance gap, tree-based ensembles often retain an edge in terms of training efficiency and robustness on tabular tasks (Borisov et al., 2022). Consequently, a robust methodology must evaluate these diverse algorithmic families side-by-side to determine which architecture offers the optimal balance of performance and reliability for a given credit portfolio.

Evaluating these models requires a multidimensional framework that extends beyond simple accuracy. In credit-risk governance, the ability to rank borrowers correctly is paramount, yet standard metrics like the Area Under the ROC Curve (AUC) have faced theoretical criticism. Hand (2009) argues that AUC implicitly assumes incoherent misclassification costs, proposing the H-measure as a more theoretically sound alternative for cost-sensitive financial applications. Furthermore, reliance on point estimates of performance can be misleading; rigorous comparison requires statistical validation. Techniques such as DeLong's test for comparing correlated ROC curves and McNemar's test for classification disagreement are essential for distinguishing genuine predictive improvements from sampling variability (DeLong et al., 1988; McNemar, 1947).

Despite advances in predictive power and rigorous evaluation, the "black box" nature of modern ensembles remains a critical barrier to adoption. To address this, the field of Explainable AI (XAI) developed post-hoc attribution methods such as SHAP and LIME, which approximate complex models to provide local feature contributions (Lundberg & Lee, 2017; Ribeiro et al., 2016). While these tools are widely adopted, a growing body of critical literature warns that they function as descriptive summaries rather than stable explanations. Research by (Rudin, 2019) and (Slack et al., 2020) demonstrates that these explanations can be misleading or even manipulated by adversarial perturbations. More recently, Agarwal et al. (2022) provided empirical evidence that feature attributions often exhibit high variance, implying that a "top predictor" identified by SHAP might change simply due to random initialization or minor data shifts, rendering such explanations unsuitable for high-stakes regulatory compliance without secondary validation.

To bridge the gap between technical explanations and human understanding, emerging approaches have begun to integrate Generative AI to narrate model reasoning. These systems aim to translate quantitative attribution scores into coherent natural-language summaries, potentially enhancing the auditability of complex models (X. Wang et al., 2025). However, the use of generative models introduces its own risks, particularly the potential for "hallucinated" reasoning where the AI invents plausible but structurally unsupported causal claims. This necessitates a constrained approach where generative outputs are strictly bounded by empirically validated evidence, ensuring alignment with the rigorous governance expectations for transparency and stability mandated by banking supervisors (Basel Committee on Banking Supervision, 2011; European Banking Authority, 2021).

3. Methodology

This section outlines the predictive modelling pipeline, the dual-selector attribution mechanism, the falsifiability layer for testing explanation validity, and the constrained generative explanation module. The overall objective is to establish a unified predictive–explanatory architecture that produces explanations which are not only informative but also scientifically testable and aligned with model-governance requirements.

Datasets Overview

To rigorously evaluate the efficacy of the Factorization Machine (FM) model, this study utilizes four real-world credit risk datasets sourced from the UCI Machine Learning Repository (Quan & Sun, 2024). These datasets are established benchmarks in the literature, routinely employed to validate classification algorithms ranging from traditional logistic regression to advanced ensemble methods (Baesens et al., 2003; Lessmann et al., 2015).

The German Credit dataset is a primary standard for comparative credit scoring studies (Lessmann et al., 2015). Comprising 1,000 instances with a 70:30 class split, its 20 features provide a mixed-attribute environment essential for testing how well models capture feature interactions in the presence of categorical variables (Quan & Sun, 2024).

The Bank Marketing dataset is utilized to assess performance under conditions of high sparsity and class imbalance (Quan & Sun, 2024). Containing 45,211 records from telemarketing campaigns, it is heavily skewed with a minority class of only 11.7%, requiring robust performance in detecting rare positive instances in binary classification tasks (Louzada et al., 2016).

The Credit Approval dataset serves as a test bed for handling data quality issues. The subset of 300 applications contains missing values and anonymized mixed-type attributes. It is frequently used to evaluate the robustness of classifiers against data sparsity and the effectiveness of preprocessing techniques (Quan & Sun, 2024).

The Statlog (Australian Credit Approval) dataset provides a rigorous test of model flexibility (Baesens et al., 2003). With 690 samples and a relatively balanced distribution, it features 14 attributes that mix continuous financial metrics with nominal demographic codes. This diversity makes it a standard for verifying a model's capacity to generalize across different feature representations (Quan & Sun, 2024).

Table 1 summarizes the structural characteristics of these datasets.

Table 1: Summary of Credit Risk Datasets Utilized in Experiments

German Credit Dataset	**Bank Marketing Dataset**
Records: 1,000	Records: 45,211
Features: 20 (7 num, 13 cat)	Features: 16 attributes
Target: Good (700) / Bad (300)	Target: Good (39,922) / Bad (5,289)
Credit Approval Dataset	**Statlog (Australian Credit)**
Records: 300 (Selected subset)	Records: 690
Features: 15 (Cat, Int, Real)	Features: 14 (6 cont, 8 cat)
Target: Good (211) / Bad (89)	Target: Good (383) / Bad (307)

End-to-End Workflow

Figure 1 presents the complete workflow that links data ingestion, feature screening, model training, benchmark selection, global explainability, and local counterfactual analysis into a unified computational pipeline. The process begins with dataset ingestion and diagnostic checks, which establish baseline data quality, missing-value patterns, feature types, and class distributions. This is followed by feature screening through a composite selector that integrates Random Forest and L1-logistic ranking, together with analyst overrides where required. Preprocessing steps—such as imputation, encoding,

scaling, and optional imbalance handling using SMOTE—are applied uniformly to ensure consistency across datasets. The training phase then evaluates multiple algorithm families using stratified splits, model calibration, and a common set of evaluation metrics, producing a structured output of model performance across datasets.

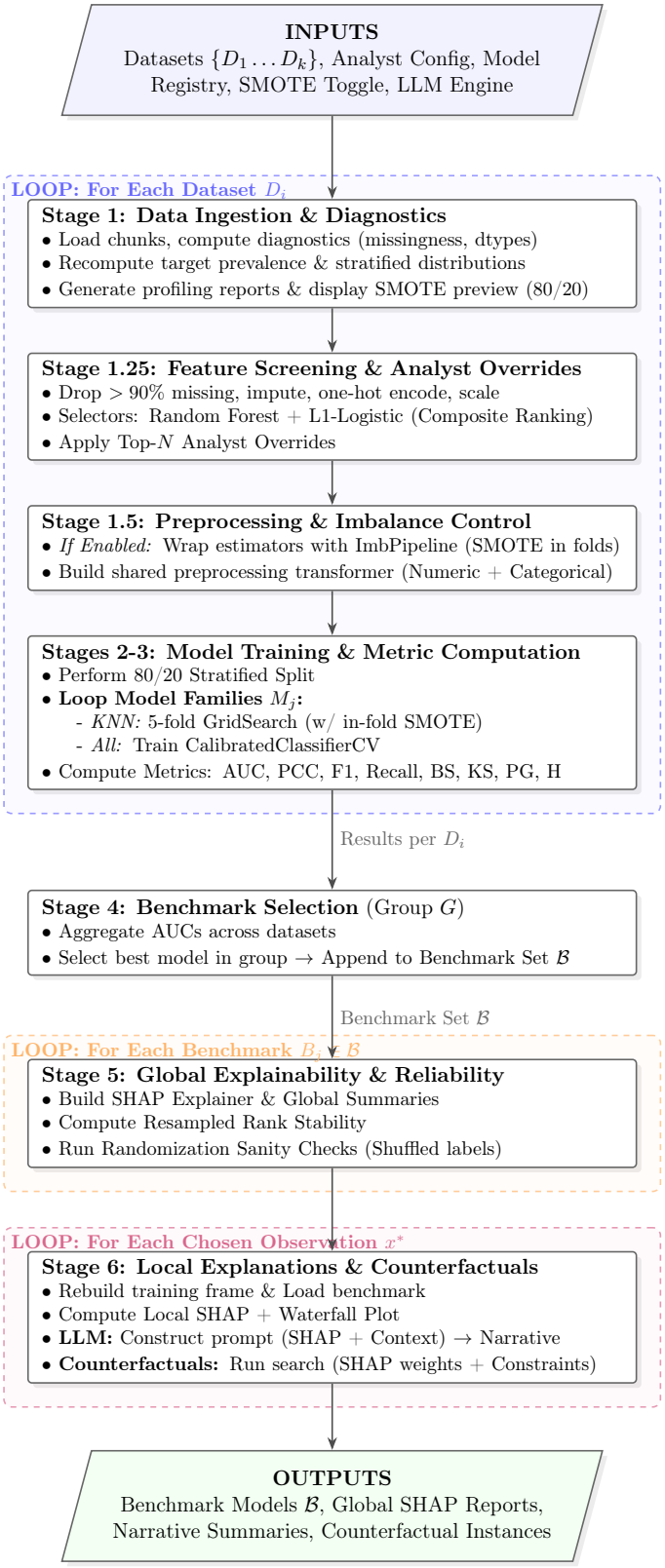

Figure 1: End-to-End Workflow for Predictive Modelling, Explainability, and Counterfactual Analysis

The subsequent stages derive global and local interpretability layers from this predictive foundation. Benchmark models are selected by aggregating performance across

datasets, and these models serve as the basis for global SHAP analyses, stability checks, and randomisation tests that assess reliability and robustness. For selected observations, the workflow computes local SHAP explanations, generates constrained natural-language narratives using the LLM module, and performs counterfactual search under feature-constraint rules. The final outputs include the benchmark model set, global and local explanation artefacts, narrative summaries, and counterfactual instances, providing an integrated, traceable, and reproducible pipeline for model evaluation and interpretability.

Dual-Selector Feature Attribution

Attribution methods based solely on a single estimator risk introducing modelling biases or instability. To address this issue, we employ a **dual-selector** mechanism that integrates both nonlinear and linear attribution signals. The procedure combines Random Forest importance with L1-regularised logistic regression to form a robust, bias-reduced feature selection strategy. Random Forest provides a nonlinear, interaction-aware estimate of feature relevance derived from impurity-based importance metrics, while the L1-regularised logistic regression model contributes a sparse, linearly grounded measure based on the magnitude of penalised coefficients.

Both sets of importance values are normalised to the interval $[0, 1]$ and then averaged to produce a composite importance score that reflects the strengths of each method while reducing dependence on any single modelling assumption. This hybridised attribution structure yields a consistent set of candidate explanatory variables, which subsequently feed into SHAP computation, stability testing, and counterfactual validation. The overall process is summarised in Table 2, and it ensures that downstream explanation layers operate on a foundation of empirically supported and methodologically diverse attribution signals.

The dataset summary tables provide an overview of the structural characteristics of each credit-risk dataset used in the analysis. They report the number of variables, sample size, extent of missingness, and distribution of variable types, enabling a transparent comparison of data complexity across sources. These descriptive statistics help clarify why certain preprocessing steps—such as imputation, encoding, and scaling—are necessary prior to model training. Presenting these details ensures that differences in data composition are accounted for when evaluating model performance and the stability of feature-importance results.

The workflow table summarises the procedure used to compute combined feature-importance scores from Random Forest and L1-regularised logistic regression. It outlines the preprocessing operations, data-splitting approach, model-training configuration, and the method used to merge importance values into a unified ranking. Describing the process in this structured way clarifies how predictive models are prepared and how attribution scores are generated, while also making explicit the sequence of steps that

influence interpretability. This supports reproducibility and provides a consistent basis for comparing importance patterns across datasets and modelling approaches.

Table 2: Feature-importance computation workflow (Step 4).

Step	Details
1. Preprocessing	- Drop columns with >90% missing values - Impute numerical/categorical values (mean / most frequent) - One-hot encode categorical variables - Min–Max scale numerical fields
2. Split & Balance	- Train–test split (75–25) - Random oversampling on the training set for balance
3. Model Training	- Random Forest $\Rightarrow$ impurity-based feature importance - Logistic Regression (L1) $\Rightarrow$ coefficient magnitude
4. Merge Results	- Normalise importance scores to $[0, 1]$ - Compute average of RF and LR importances - Rank features based on the combined score
5. Output	- Individual RF and LR attribution tables - Merged composite importance DataFrame - Final ranks, normalised scores, and metadata

Predictive Modelling Pipeline

To ensure robustness across modelling families, we evaluate a total of 392 calibrated model configurations grouped into four broad algorithmic families: linear models, boosting methods, bagging ensembles, and instance-based learners (Table 3). Within each family we vary key hyperparameters (e.g., solvers, penalties, number of estimators, learning rates, neighbourhood sizes) to obtain a rich design space for benchmarking both predictive performance and explanation stability.

Table 3: Classification model families used in the predictive pipeline.

Category	**Algorithm**	**Versions**	**Model Count**
Linear	Logistic Regression	solvers: lbfgs, saga, newton-cg	3
	Regularized Logistic Regression	penalties: elasticnet, L2, L1	3
Boosting	AdaBoost	estimators: 10, 20, 30	3
	Boosted Decision Trees	estimators: 10–1000 × learning rate (0.1, 0.5, 1.0)	18
	Stochastic Gradient Boosting	estimators: 10–1000	7
	XGBoost	estimators: 100–300 × learning rate (0.1, 0.3)	6
	LightGBM	estimators: 100–300 × learning rate (0.1, 0.3)	6
Bagging	Bagged Decision Trees (CART)	estimators: 10–1000 trees	7
	Bagged Neural Networks	estimators: 5–100 networks	4
	Random Forest	estimators: 100–1000 × max features (sqrt–1.0)	25
Instance-Based	K-Nearest Neighbours (tuned)	neighbours: 3–21 × weights (uniform, distance) × distance metric (1, 2) (24 candidate configurations explored within a single tuned model)	1
Deep Learning	Feedforward Neural Network (MLP)	single architecture, probability-calibrated	1
	Temporal Convolutional Network (TCN)	single architecture, probability-calibrated	1
	Transformer-based Classifier	single architecture, probability-calibrated	1
Total	—	—	109

Evaluation Metrics

Model evaluation in credit-risk modelling requires more than a single accuracy score. The regulatory context demands a multidimensional assessment of discrimination, calibration, and cost-sensitive performance. Accordingly, we employ a suite of complementary metrics summarised in Table 4. Together, these metrics reflect ranking ability, probability reliability, class-separation behaviour, and utility under asymmetric misclassification costs. Each measure contributes a distinct diagnostic perspective, ensuring that model assessment captures both statistical performance and governance-relevant behaviour.

Table 4: Summary of evaluation metrics used for benchmarking and performance assessment.

Metric (Full Form)	**Usage**	**Formula**
AUC – Area Under ROC Curve	Ranking ability; sensitivity to score ordering	$\text{AUC} = \int_0^1 \text{TPR}(x)\, d(\text{FPR}(x))$
PCC – Prevalence Corrected Classification	Accuracy adjusted for empirical class prevalence	$\text{PCC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$
F1 – F1 Score	Balances precision and recall; useful under class imbalance	$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$
Recall – Sensitivity	Ability to detect defaults (true positives)	$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$
BS – Brier Score	Measures probability calibration error	$\text{BS} = \frac{1}{N} \sum_{i=1}^{N} (p_i - y_i)^2$
KS – Kolmogorov–Smirnov Statistic	Maximum class separation between default and non-default distributions	$\max_t \lvert\text{TPR}(t) - \text{FPR}(t)\rvert$
PG – Partial Gini	Local discrimination in conservative decision regions	$\text{PG} = 2 \times \text{AUC}_{(p \le b)} - 1$
H – Hand's H-measure	Utility-based, cost-sensitive alternative to AUC	$H = 1 - \frac{\text{EMC}}{\text{EMC}_0}$

The decision threshold in this study is not fixed at 0.5; instead, it is dynamically aligned with empirical class prevalence within each training fold to ensure that binary predictions reflect the true population base rate. Following Lessmann et al. (2015), the H-measure provides a theoretically grounded utility-based alternative to AUC and is particularly suited to imbalanced financial datasets. The Brier Score complements these metrics by quantifying probability-level accuracy, while KS and Partial Gini capture the extent of distributional separation and local discriminatory power—both crucial in credit-risk environments where stability across scoring bands is required.

Although a wide range of metrics is reported for completeness, AUC is adopted as the primary benchmark measure. Its threshold independence avoids distortions introduced by shifting class prevalence, and it enables consistent comparison across heterogeneous datasets and model families. Moreover, AUC directly measures ranking quality, which

is fundamental in probability of default modelling and aligns closely with regulator expectations for discriminatory power. For these reasons, AUC serves as the unifying basis for selecting benchmark models across algorithmic groups, with the remaining metrics providing complementary diagnostic insight.

Benchmarking and Model Selection Across Algorithm Groups

To evaluate predictive consistency across modelling families and datasets, we implement a structured benchmarking and model-selection workflow. The process ensures that model comparison is fair, reproducible, and grounded in performance profiles aggregated across datasets rather than single-run outcomes. The six-stage process is summarised in Table 5. It includes the collection of model-level metrics, aggregation across datasets, computation of group-wise averages, selection of benchmark models, and consolidated comparison across algorithm families.

Table 5: Benchmarking and model selection workflow across algorithm groups.

Step	**Description**
1. Collect Results	Gather AUC, F1, KS and related metrics for each dataset and algorithm group
2. Aggregate Metrics	Combine recorded results and organise performance scores within each algorithm family across datasets
3. Compute Averages	Compute the mean AUC for each model across datasets to measure robustness and consistency
4. Identify Best Model	Select the model with the highest average AUC within each algorithm family
5. Build Summary Table	Compile benchmark models with full metric profiles (AUC, PCC, F1, Recall, BS, KS, PG, H) for each dataset
6. Display Comparison	Present the average AUC comparison by algorithm group and the consolidated benchmark summary across datasets

Falsifiability Layer: Testing Explanation Validity

A central contribution of this research is the introduction of a falsifiability-based diagnostic layer designed to determine whether explanations correspond to genuine structure in the underlying data. In this framework, explanations are treated as hypotheses that must withstand perturbation, noise, and counterfactual scrutiny. The objective is to shift explanation assessment from descriptive reporting toward empirical testing, ensuring that only reasoning supported by evidence is considered valid.

The first component of this layer is the rank-stability test, which evaluates whether attribution methods produce consistent results under resampled background distributions. SHAP values are recomputed repeatedly, and a valid explanation is expected to preserve both the ranking and relative magnitude of its most influential features across replications. When top-ranked features fluctuate substantially, or when attribution magnitudes vary widely, the resulting explanations are deemed unstable. Such instability indicates

that the explanation is sensitive to sampling noise and therefore cannot be regarded as a reliable reflection of model behaviour.

The second component is the label-randomisation test, which distinguishes true structural signal from spurious correlations. The model is retrained on data in which the target labels have been randomly permuted. Under these conditions, a meaningful explanation should effectively collapse: attribution mass should become diffuse, incoherent, and substantially reduced in magnitude. If strong attributions persist despite randomised labels, the conclusion is that the original explanations were artefactual—driven by methodological biases rather than genuine predictive structure.

The third component is the counterfactual validity test, which examines whether an explanation can produce plausible minimal-change counterfactuals. Counterfactual construction is formulated as a constrained optimisation problem over a weighted Gower distance metric that accommodates heterogeneous data types. The optimisation seeks the nearest feasible counterfactual by applying minimal perturbations while satisfying plausibility constraints and assigning importance-weighted distances to feature changes. A valid explanation must support interpretable and domain-consistent feature modifications capable of flipping the model's prediction. When no reasonable minimal-change counterfactual can be generated, the explanation fails to represent an actionable or structurally meaningful decision boundary.

Generative AI Explanation Layer

To enhance communicability without compromising epistemic rigor, we introduce a constrained generative AI module that translates validated SHAP evidence into concise and interpretable natural-language explanations. Unlike unconstrained narrative-generation systems, this module operates strictly within the boundaries imposed by the falsifiability layer. It may summarise evidence and articulate the reasoning implied by validated attributions, but it is not permitted to introduce causal claims, speculative associations, or any narrative element that lacks empirical support.

The generative component produces a structured explanatory output consisting of two short interpretive segments. The first is a causal narrative that summarises why the model produced the predicted class, emphasising only those SHAP features that have demonstrated high impact and stability under the falsifiability tests. The second segment provides a reflective commentary comparing the model's prediction to the observed outcome, highlighting areas of agreement or divergence and noting conditions under which deviations occurred. Together, these components aim to balance interpretability with methodological discipline, ensuring that narrative clarity never overrides evidential constraints.

The configuration, inputs, processing logic, prompt structure, and output format of this module are summarised in Table 6. By formalising these elements, the design ensures

that generated explanations remain reproducible, testable, and aligned with transparent model-governance expectations under the Basel Framework. This structure also supports auditability by making explicit the boundaries within which generative reasoning is allowed to operate.

The LLM-based explanation module translates SHAP attribution values into short, structured natural-language summaries. It operates on the ranked SHAP outputs from the predictive models, taking the feature names, feature values, true label, and predicted probability as inputs. To ensure consistency, the module first identifies the top contributing features ordered by absolute SHAP magnitude and restricts the explanation to a curated subset of validated variables. This step ensures that the generated narrative reflects the most influential components of the model's decision process.

The prompt design enforces a controlled structure that limits speculation while maintaining interpretability. The system prompt constrains the LLM to operate as an analytical explainer, whereas the user prompt supplies the exact numerical context required for faithful explanation. The model is instructed to produce two short paragraphs: one describing how the top features drive the prediction, and another comparing the prediction to the actual label. This configuration enhances transparency by providing an interpretable narrative while respecting the boundaries set by SHAP, thereby reducing the risk of unsupported causal claims or hallucinations. Error-handling routines ensure that invalid inputs or API issues are surfaced clearly, supporting reproducibility and operational reliability.

Table 6: LLM-based SHAP explanation module configuration and workflow.

Step	**Description**
Purpose	- Generates natural-language explanations for SHAP outputs using OpenAI models
Input	- Receives SHAP values, feature names, feature values, true label, and predicted probability
Processing	- Sorts features by absolute SHAP magnitude - Selects top 10 validated features for narrative explanation
Prompt Used	- System prompt defines the role as an ML analyst explaining SHAP to business users - User prompt contains: model prediction, actual value, and top-N feature contributions - Required output: two short paragraphs 1. Why the model made the prediction 2. Whether the prediction aligns with the actual value
LLM Model	- Calls `gpt-4o-mini` via `chat.completions.create`
Output	- Returns two paragraphs summarising SHAP-driven reasoning and prediction correctness
Error Handling	- Produces clear diagnostics for missing/invalid API keys or malformed input data

The LLM-based explanation module converts SHAP attribution values into concise, natural-language summaries that complement the quantitative feature-importance analysis. It receives the ranked SHAP outputs along with feature names, feature values, the model prediction, and the true label. The module first identifies the most influential features by ordering them according to absolute SHAP magnitude and restricting the explanation to a validated subset. This ensures that the narrative remains grounded in the model's most important drivers and avoids the inclusion of low-impact or noisy variables.

The structured prompting framework guides the generation of explanations. The system prompt constrains the LLM to act as a machine-learning analyst, while the user prompt provides the numerical and contextual details required for faithful summarisation. The model is instructed to produce two short paragraphs: the first explains how the top features influence the prediction, and the second evaluates whether the prediction is consistent with the actual outcome. This design supports transparency while limiting the risk of unsupported causal claims or hallucinated reasoning. Built-in error-handling routines surface issues such as invalid API credentials or malformed inputs, thereby maintaining reliability and ensuring the module operates within well-defined boundaries.

Reliability Scoring Through Stability-Weighted Explanatory Metricss

A central challenge in credit-risk modelling is determining not merely *what* a model predicts, but *why* it produces a given probability of default (PD) and whether that explanation is stable under perturbations. Conventional SHAP analyses describe local

feature contributions but do not evaluate the explanatory reliability of those contributions. Prior research has demonstrated that post hoc attribution methods such as SHAP may be sensitive to sampling variation, background distribution choices, and adversarial or randomisation-based perturbations (**dimanov2020_conceal**; Adebayo et al., 2018; Agarwal et al., 2022; Slack et al., 2020). This instability is especially problematic in regulated financial settings, where supervisory frameworks emphasise reproducibility, transparency, and robustness of model explanations (**pra2023_modelrisk**; **sr11_7_fed**; European Banking Authority, 2021).

The proposed approach combines three complementary metrics derived from repeated perturbation runs: average rank, rank standard deviation, and the sanity ratio. Average rank captures a feature's expected explanatory priority across trials, while rank standard deviation operationalises the observation in (Agarwal et al., 2022) that unstable feature attributions reflect fragile or easily perturbed explanations. The sanity ratio extends the logic of randomisation-based validation introduced in (Adebayo et al., 2018; Hooker et al., 2019) to quantify whether genuine predictive structure dominates noise features constructed through label-shuffling or synthetic baselines. Together, these components address three epistemic tasks: explanatory prioritisation, stability assessment, and signal–noise discrimination.

Reliability Scoring Formulation. To quantify the trustworthiness of model explanations, we define a composite Reliability Score (RelScore). This metric integrates local feature stability with global signal verification. Let R_i be the average rank of feature i, σ_i its rank standard deviation across perturbations, and S the sanity ratio derived from label-randomisation tests. The formulation is detailed in Table 7.

Table 7: Mathematical formulation of the Stability-Weighted Reliability Score.

Component	Mathematical Definition	Description
1. Feature Stability Weight	$W_i = \left(\frac{1}{1+R_i}\right) \times \left(\frac{1}{1+\sigma_i}\right)$	*Penalises features with low importance or high variance.*
2. Aggregated Robustness	$ER = \frac{1}{K}\sum_{i=1}^{K} W_i$	*Average stability across the top-K features.*
3. Signal Adjustment	$W_{\text{signal}} = \frac{\min(S,3)}{3}$	*Caps confidence based on signal-to-noise ratio (S).*
4. Final Reliability Score	$\text{RelScore} = ER \times W_{\text{signal}}$	*Composite score* $\in [0,1]$.

This formulation ensures that RelScore $\in [0,1]$, where high values indicate an explanation that is both internally stable (low σ_i) and structurally distinct from noise (high S).

In practice, the score admits four interpretable regimes—Reliable, Moderately Reliable, Questionable, and Unreliable—which characterise the epistemic quality of the PD estimate. A high score indicates that the model's prediction arises from a reproducible explanatory structure, whereas a low score indicates that the explanation is easy to vary and therefore unsuitable for high-stakes inference. This methodology reframes explanation assessment not as a descriptive exercise but as an inquiry into the stability of the underlying knowledge the model expresses. By operationalising explanatory robustness, the proposed score provides a principled way to assess whether a PD estimate should be trusted, interpreted cautiously, or rejected as artefactual, while remaining open to falsification if future evidence reveals systematic mismatches between stable attributions and realised outcomes.

3.1 State-of-the-Art Benchmarks

Table 8: Best Reported Results by Dataset and Paper

Dataset	**Paper**	**Best Model**	**Interpretability**	**AUC**
AC	Baesens et al., 2003	Ensemble (NN + Trees)	Feature Sensitivity	93.1
	Quan & Sun, 2024	Factorization Machine	–	89.28
CA	Quan & Sun, 2024	Factorization Machine	–	90.53
CR	Zeng et al., 2024	Optimised XGBoost	–	86.61
GCD	Quan & Sun, 2024	Factorization Machine	–	81.65
	L. Wang et al., 2025	Two-Stage XGBoost	SHAP, LIME	80.28
GC	Baesens et al., 2003	GASEN Ensemble	–	80.7
HELOC	X. Wang et al., 2025	Explainable Neural Network	LIME-based	81.40
HMEQ	L. Wang et al., 2025	Two-Stage XGBoost	SHAP, LIME	90.71
TH02	Baesens et al., 2003	GASEN Ensemble	–	64.4
PAKDD	Baesens et al., 2003	HCES-Bagging	–	65.2
GMSC	Baesens et al., 2003	Top-T Ensemble	–	86.5

Table 9: Best Reported Results by Dataset and Paper

Dataset	Paper	Best Model	Interpretability	AUC
AC	Baesens et al., 2003	Ensemble (NN + Trees)	Feature Sensitivity	93.1
	Quan & Sun, 2024	Factorization Machine	–	89.28
CA	Quan & Sun, 2024	Factorization Machine	–	90.53
CR	Zeng et al., 2024	Optimised XGBoost	–	86.61
GCD	Quan & Sun, 2024	Factorization Machine	–	81.65
	L. Wang et al., 2025	Two-Stage XGBoost	SHAP, LIME	80.28
GC	Baesens et al., 2003	GASEN Ensemble	–	80.7
HELOC	X. Wang et al., 2025	Explainable Neural Network	LIME-based	81.40
HMEQ	L. Wang et al., 2025	Two-Stage XGBoost	SHAP, LIME	90.71
TH02	Baesens et al., 2003	GASEN Ensemble	–	64.4
PAKDD	Baesens et al., 2003	HCES-Bagging	–	65.2
GMSC	Baesens et al., 2003	Top-T Ensemble	–	86.5

4. Dataset Foundations

Table 10: Unified Dataset Summary

Dataset	Samples	Features (N/C/B)	Default (%)
credit approval dataset	690	15 (6/6/3)	0.0
german credit record	1,000	21 (7/11/3)	0.0
Australian Credit	690	14 (10/0/4)	0.0
hmeq	5,960	12 (10/2/0)	0.0
TH02	1,225	14 (11/2/1)	0.0
MSME Credit Data by 30S-CR	1,707	87 (79/3/5)	0.0

Note: N/C/B denotes the number of Numeric, Categorical, and Binary features respectively.

The datasets exhibit substantial variation in scale, feature composition, and outcome distributions, defining heterogeneous credit environments. This diversity establishes the scope of the claims in this study and motivates the need for stability- and reliability-based model evaluation, as performance-optimal models in one dataset may rely on fragile correlations that do not generalise across domains.

5. Results

Model selection based solely on predictive performance is insufficient for deployment. This section introduces a reliability-constrained selection framework that enforces three sequential criteria: (1) performance-equivalence, (2) statistical robustness, and (3) explanation reliability. Multiple models may achieve statistically indistinguishable predictive performance; reliability constraints resolve this selection problem.

5.1 Predictive Performance Frontier

The initial performance frontier identifies the best-performing model per dataset under conventional AUC metrics. However, raw performance rankings do not account for statistical uncertainty or explanation trustworthiness.

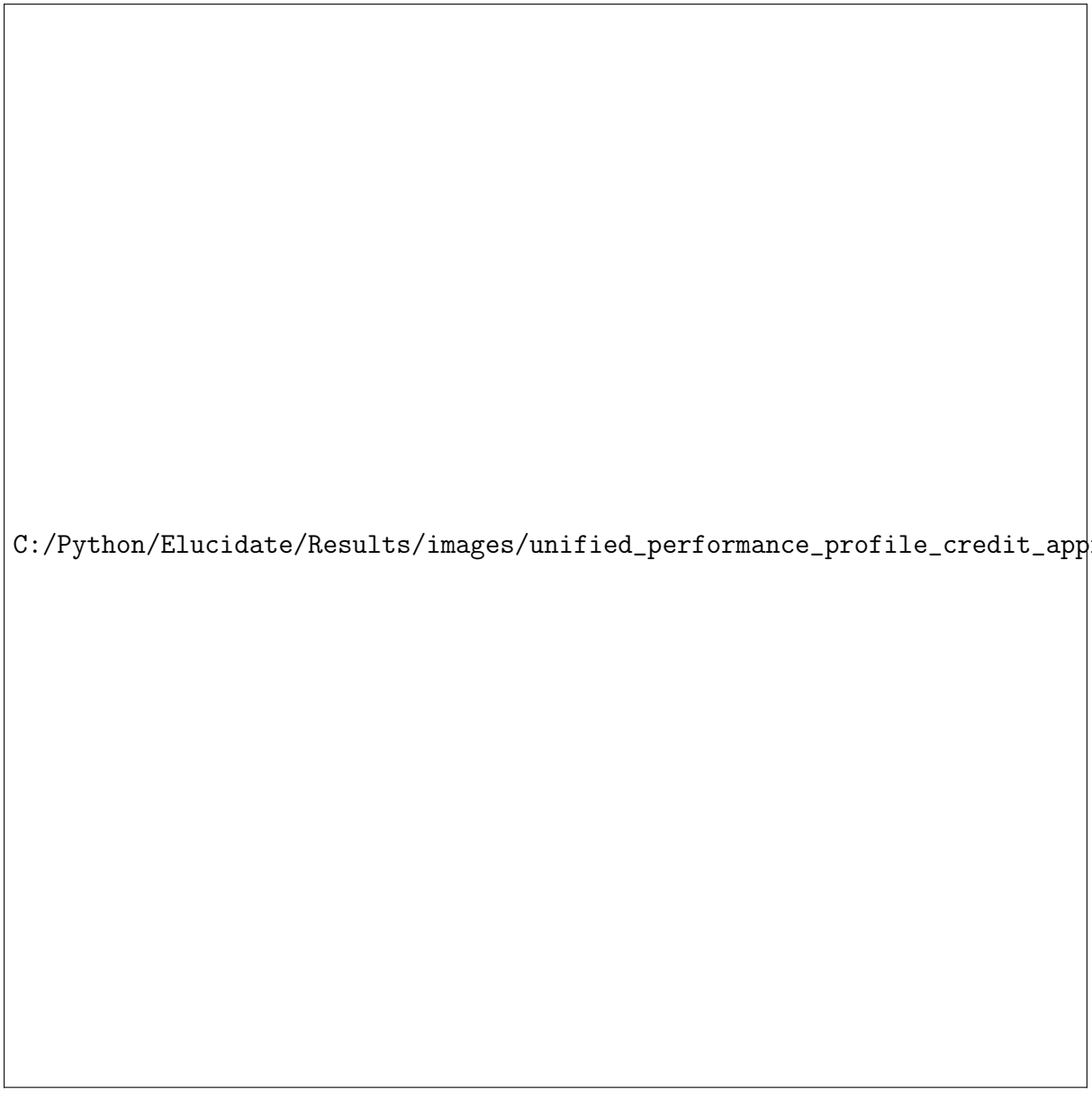

Figure 2: Unified Performance Profile: Best models from each group across datasets. Red circles indicate the overall best model for each dataset.

Table 11: Performance-Optimal Models by Dataset

Dataset	**Best Model**	**AUC**
credit_approval_dataset	Torch-Transformer-64	0.9596
german_credit_record	GB-500-0.05	0.8213
Australian Credit	GB-50-0.1	0.9383
hmeq	LGBM-500-0.1	0.9651
TH02	RF-100-sqrt	0.6548
MSME Credit Data by 30S-CR	BAG-MLP-25	0.8229

Note: Full performance frontier showing all models is provided in Appendix C.

5.2 Performance-Equivalence Candidate Models

To ensure architectural diversity in the candidate set, we apply a **group-wise selection rule**. For each model group (e.g., Gradient Boosting, Random Forest, Neural Networks), we identify the group-best model M_g^* (highest AUC within group g) and include it as a candidate. This prevents a single high-capacity family from dominating the selection pool and ensures that reliability filtering operates over structurally diverse architectures.

Table 12: Performance-Equivalence Candidate Models (Group-wise)

Dataset	Model Group	Group-Best Model	AUC
lightgray credit_approval_dataset	PyTorch Neural Networks	Torch-Transformer-64	0.9596
	Gradient Boosting	GB-50-0.1	0.9582
	Regularized Logistic Regression	LR-Reg-SAGA	0.9579
	XGBoost	XGB-500-0.1	0.9562
	Logistic Regression	LR-SAGA	0.9559
	AdaBoost (Decision Tree Stumps)	AdaBoost-20	0.9552
	LightGBM	LGBM-500-0.1	0.9548
	Random Forest	RF-250-0.25	0.9531
	Bagging (CART)	BAG-CART-10	0.9518
lightgray german_credit_record	Gradient Boosting	GB-500-0.05	0.8213
	PyTorch Neural Networks	Torch-MLP-64-32	0.8155
	Bagging (MLP)	BAG-MLP-5	0.8138
lightgray Australian Credit	Gradient Boosting	GB-50-0.1	0.9383
	XGBoost	XGB-10-0.3	0.9356
	PyTorch Neural Networks	Torch-MLP-64-32	0.9353
	Random Forest	RF-250-0.25	0.9347
	LightGBM	LGBM-10-0.3	0.9340
	AdaBoost (Decision Tree Stumps)	AdaBoost-10	0.9337
	Bagging (CART)	BAG-CART-100	0.9334
lightgray hmeq	LightGBM	LGBM-500-0.1	0.9651
	Random Forest	RF-100-sqrt	0.9619
	XGBoost	XGB-100-0.3	0.9599
lightgray TH02	Random Forest	RF-100-sqrt	0.6548
lightgray MSME Credit Data by 30S-CR	Bagging (MLP)	BAG-MLP-25	0.8229

Note: Each row shows the best-performing model within its group. Models must be within 1% AUC of the global best to qualify as candidates. Group-wise selection ensures exactly one representative per model family, preventing architectural homogeneity in the candidate set. Global best model per dataset highlighted in grey.

By selecting exactly one representative per model family, the framework ensures that final deployment decisions reflect both predictive performance and architectural diversity. Each group-best model advances to statistical robustness and explanation reliability evaluation.

5.3 Statistical Robustness of Candidate Models

Performance-optimal models (reference models) are compared against second-best candidates using three statistical tests: Wilcoxon Signed-Rank Test, McNemar's Test, and DeLong's Test. Each test evaluates whether performance differences are statistically significant ($p < 0.05$). Strong evidence (all 3 tests significant) confirms structural superiority. When models are not significantly different, explanation reliability assessment determines which model provides more trustworthy predictions for deployment.

Table 13: Statistical Robustness of Model Comparisons Across Datasets

Dataset	Model 1	Model 2	Wilcoxon	McNemar	DeLong	Verdict
credit_approval_dataset	Torch-Transformer-64	AdaBoost-20	0.1250	0.3816	0.7510	N.S.
	Torch-Transformer-64	BAG-CART-10	0.6250	$4.82e-07^{***}$	$4.02e-07^{***}$	Moderate
	Torch-Transformer-64	GB-50-0.1	0.0625	$1.20e-04^{***}$	$2.35e-06^{***}$	Moderate
	Torch-Transformer-64	LGBM-500-0.1	0.8125	$1.60e-12^{***}$	$8.49e-08^{***}$	Moderate
	Torch-Transformer-64	LR-Reg-SAGA	0.1250	0.1002	$9.78e-04^{***}$	Weak
	Torch-Transformer-64	LR-SAGA	0.8125	0.1698	0.1263	N.S.
	Torch-Transformer-64	RF-250-0.25	0.3125	$1.02e-11^{***}$	$6.89e-08^{***}$	Moderate
	Torch-Transformer-64	XGB-500-0.1	0.3125	$2.62e-11^{***}$	$1.51e-07^{***}$	Moderate
german_credit_record	GB-500-0.05	BAG-MLP-5	0.6250	0.3019	$1.94e-06^{***}$	Weak
	GB-500-0.05	Torch-MLP-64-32	0.3573	$3.43e-08^{***}$	$0.00e+00^{***}$	Moderate
Australian Credit	GB-50-0.1	AdaBoost-10	0.0625	$1.82e-05^{***}$	$3.55e-08^{***}$	Moderate
	GB-50-0.1	BAG-CART-100	0.1441	$1.94e-04^{***}$	$5.78e-04^{***}$	Moderate
	GB-50-0.1	LGBM-10-0.3	0.6250	$8.00e-05^{***}$	0.0056^{**}	Moderate
	GB-50-0.1	RF-250-0.25	1.0000	$2.31e-06^{***}$	$9.06e-04^{***}$	Moderate
	GB-50-0.1	Torch-MLP-64-32	0.1250	$4.69e-05^{***}$	$1.07e-07^{***}$	Moderate
	GB-50-0.1	XGB-10-0.3	0.0625	0.7103	0.0601	N.S.
hmeq	LGBM-500-0.1	RF-100-sqrt	0.0625	0.1556	0.6950	N.S.
	LGBM-500-0.1	XGB-100-0.3	1.0000	0.0027^{**}	0.0010^{**}	Moderate

Note: Significance levels: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. Tests: Wilcoxon Signed-Rank, McNemar's, DeLong's. Verdict: Strong = all 3 tests significant; Moderate = 2 tests significant; Weak = 1 test significant; N.S. = no tests significant.

Statistical analysis shows mixed evidence. 0 datasets demonstrate strong evidence and 4 show moderate evidence, while 0 show weak evidence and 0 show no significant differences. **Moderate evidence:** credit_approval_dataset, german_credit_record, Australian Credit, hmeq.

5.4 Explanation Reliability of Candidate Models

Statistically robust candidates must also provide reliable explanations for deployment. Explanation reliability combines attribution stability (SHAP rank perturbation) and noise sensitivity (label randomization). Only models with Reliable verdicts are deployment-ready.

Table 14: Explanation Reliability Assessment

Dataset	Model	Sanity Ratio	Test Config	Stability (S/M/U)	Noise Sensitivity	Verdict
credit_approval_dataset	RF-250-0.25	0.905	3 / 20 / 15	15 / 0 / 0	Moderate	Fragile
credit_approval_dataset	Torch-Transformer-64	0.858	3 / 20 / 15	11 / 3 / 1	Moderate	Fragile
credit_approval_dataset	LR-Reg-SAGA	1.006	3 / 20 / 15	7 / 4 / 4	Low	Fragile
credit_approval_dataset	LR-SAGA	1.160	3 / 20 / 15	9 / 4 / 2	Very Low	Fragile
credit_approval_dataset	AdaBoost-20	0.985	3 / 20 / 15	15 / 0 / 0	Low	Reliable
credit_approval_dataset	BAG-CART-10	1.813	3 / 20 / 15	15 / 0 / 0	Very Low	Reliable
credit_approval_dataset	GB-50-0.1	1.014	3 / 20 / 15	15 / 0 / 0	Low	Reliable
credit_approval_dataset	LGBM-500-0.1	0.937	3 / 20 / 15	14 / 1 / 0	Moderate	Fragile
credit_approval_dataset	XGB-500-0.1	0.954	3 / 20 / 15	14 / 1 / 0	Low	Reliable
german_credit_record	GB-500-0.05	1.023	3 / 20 / 21	20 / 1 / 0	Low	Reliable
german_credit_record	BAG-MLP-5	0.993	3 / 20 / 21	8 / 9 / 4	Low	Fragile
german_credit_record	Torch-MLP-64-32	1.077	3 / 20 / 21	6 / 10 / 5	Low	Fragile
Australian Credit	GB-50-0.1	1.001	3 / 20 / 14	9 / 5 / 0	Low	Reliable
Australian Credit	XGB-10-0.3	1.016	3 / 20 / 14	10 / 4 / 0	Low	Reliable
Australian Credit	AdaBoost-10	1.022	3 / 20 / 14	13 / 1 / 0	Low	Reliable
Australian Credit	Torch-MLP-64-32	1.008	3 / 20 / 14	8 / 5 / 1	Low	Fragile
Australian Credit	LGBM-10-0.3	1.018	3 / 20 / 14	10 / 4 / 0	Low	Reliable
Australian Credit	RF-250-0.25	0.977	3 / 20 / 14	12 / 2 / 0	Low	Reliable
Australian Credit	BAG-CART-100	0.998	3 / 20 / 14	13 / 1 / 0	Low	Reliable
hmeq	RF-100-sqrt	1.001	3 / 20 / 12	9 / 3 / 0	Low	Reliable
hmeq	LGBM-500-0.1	1.003	3 / 20 / 12	10 / 2 / 0	Low	Reliable
hmeq	XGB-100-0.3	0.988	3 / 20 / 12	9 / 1 / 2	Low	Fragile
MSME Credit Data by 30S-CR	BAG-MLP-25	0.757	3 / 20 / 87	14 / 11 / 62	High	Fragile

Note. Explanation verdict is now binary: a model is **Reliable** iff sanity ratio $\geq$ 0.95 AND unstable features = 0; otherwise **Fragile**. Test Config: Randomization trials / Background sample size / Feature count. Stability (S/M/U): Stable / Moderate / Unstable feature counts. Noise Sensitivity: Very Low (sanity $\geq$ 1.10), Low (0.95–1.10), Moderate (0.85–0.95), High ($<$ 0.85).

Explanation reliability is mixed. 13 of 23 models show reliable explanations, while 10 are fragile (either sanity ratio $<$ 0.95 or unstable features $>$ 0). Only models with Reliable verdicts should be considered for deployment.

Most Reliable Models per Dataset: credit_approval_dataset: BAG-CART-10 (sanity ratio = 1.813, Very Low noise sensitivity); german_credit_record: GB-500-0.05 (sanity ratio = 1.023, Low noise sensitivity); Australian Credit: AdaBoost-10 (sanity ratio = 1.022, Low noise sensitivity); hmeq: LGBM-500-0.1 (sanity ratio = 1.003, Low noise sensitivity);

Appendix A: Explanation Diagnostics

This appendix contains detailed SHAP feature importance tables and label randomization test results for each dataset. These provide the raw diagnostic evidence supporting the consolidated Explanation Reliability Assessment in Section 3.3.

credit_approval_dataset

SHAP global importance data not found.

german_credit_record

SHAP global importance data not found.

Australian Credit

SHAP global importance data not found.

hmeq

SHAP global importance data not found.

TH02

No SHAP analysis results available.

MSME Credit Data by 30S-CR

SHAP global importance data not found.

Appendix B: Statistical Test Details

This appendix contains detailed pairwise comparison tables for all statistical tests, organized by dataset. Each dataset includes Wilcoxon Signed-Rank Test, McNemar's Test, and DeLong's Test results. These raw comparisons support the consolidated statistical robustness assessment in Section 3.2.

credit_approval_dataset

Statistical significance tests were conducted to compare model performance. The following tests assess whether observed performance differences are statistically meaningful:

- **Wilcoxon Signed-Rank Test**: Non-parametric test comparing paired AUC scores across cross-validation folds
- **McNemar's Test**: Tests for significant differences in classification errors between two models
- **DeLong's Test**: Compares ROC curves by testing equality of AUC values

Significance level: $\alpha = 0.05$. A result is significant if $p < 0.05$.

5.5 Wilcoxon Signed-Rank Test Results

Table 15: Wilcoxon signed-rank test results comparing model performance on credit_approval_dataset. Tests compare AUC scores across cross-validation folds. Effect size (Cohen's d) quantifies the magnitude of difference.

Model 1	Model 2	Statistic	p-value	Effect Size	Significant
Torch-Transformer-64	GB-50-0.1	0.00	0.0625	—	×
Torch-Transformer-64	LR-Reg-SAGA	1.00	0.1250	—	×
Torch-Transformer-64	AdaBoost-20	1.00	0.1250	—	×
Torch-Transformer-64	XGB-500-0.1	3.00	0.3125	—	×
Torch-Transformer-64	RF-250-0.25	3.00	0.3125	—	×
Torch-Transformer-64	BAG-CART-10	5.00	0.6250	—	×
Torch-Transformer-64	LR-SAGA	6.00	0.8125	—	×
Torch-Transformer-64	LGBM-500-0.1	6.00	0.8125	—	×

5.6 McNemar's Test Results

Table 16: McNemar's test results comparing classification errors on credit_approval_dataset. Tests whether the two models make significantly different types of errors.

Model 1	Model 2	Statistic	p-value	Significant
Torch-Transformer-64	LGBM-500-0.1	49.92	< 0.001	✓
Torch-Transformer-64	RF-250-0.25	46.29	< 0.001	✓
Torch-Transformer-64	XGB-500-0.1	44.44	< 0.001	✓
Torch-Transformer-64	BAG-CART-10	25.33	< 0.001	✓
Torch-Transformer-64	GB-50-0.1	14.79	< 0.001	✓
Torch-Transformer-64	LR-Reg-SAGA	2.70	0.1002	×
Torch-Transformer-64	LR-SAGA	1.88	0.1698	×
Torch-Transformer-64	AdaBoost-20	0.77	0.3816	×

5.7 DeLong's Test Results

Table 17: DeLong's test results comparing ROC curves on credit_approval_dataset. Tests whether the AUC values of two ROC curves are significantly different.

Model 1	Model 2	Statistic	p-value	Significant
Torch-Transformer-64	RF-250-0.25	-5.394	< 0.001	✓
Torch-Transformer-64	LGBM-500-0.1	-5.356	< 0.001	✓
Torch-Transformer-64	XGB-500-0.1	-5.252	< 0.001	✓
Torch-Transformer-64	BAG-CART-10	-5.068	< 0.001	✓
Torch-Transformer-64	GB-50-0.1	-4.721	< 0.001	✓
Torch-Transformer-64	LR-Reg-SAGA	3.297	< 0.001	✓
Torch-Transformer-64	LR-SAGA	1.529	0.1263	×
Torch-Transformer-64	AdaBoost-20	0.317	0.7510	×

german_credit_record

Statistical significance tests were conducted to compare model performance. The following tests assess whether observed performance differences are statistically meaningful:

- **Wilcoxon Signed-Rank Test**: Non-parametric test comparing paired AUC scores across cross-validation folds
- **McNemar's Test**: Tests for significant differences in classification errors between two models
- **DeLong's Test**: Compares ROC curves by testing equality of AUC values

Significance level: $\alpha = 0.05$. A result is significant if $p < 0.05$.

5.8 Wilcoxon Signed-Rank Test Results

Table 18: Wilcoxon signed-rank test results comparing model performance on german_credit_record. Tests compare AUC scores across cross-validation folds. Effect size (Cohen's d) quantifies the magnitude of difference.

Model 1	Model 2	Statistic	p-value	Effect Size	Significant
GB-500-0.05	Torch-MLP-64-32	2.50	0.3573	—	×
GB-500-0.05	BAG-MLP-5	5.00	0.6250	—	×

5.9 McNemar's Test Results

Table 19: McNemar's test results comparing classification errors on german_credit_record. Tests whether the two models make significantly different types of errors.

Model 1	Model 2	Statistic	p-value	Significant
GB-500-0.05	Torch-MLP-64-32	30.45	< 0.001	✓
GB-500-0.05	BAG-MLP-5	1.07	0.3019	×

5.10 DeLong's Test Results

Table 20: DeLong's test results comparing ROC curves on german_credit_record. Tests whether the AUC values of two ROC curves are significantly different.

Model 1	Model 2	Statistic	p-value	Significant
GB-500-0.05	Torch-MLP-64-32	9.985	< 0.001	✓
GB-500-0.05	BAG-MLP-5	4.760	< 0.001	✓

Australian Credit

Statistical significance tests were conducted to compare model performance. The following tests assess whether observed performance differences are statistically meaningful:

- **Wilcoxon Signed-Rank Test**: Non-parametric test comparing paired AUC scores across cross-validation folds
- **McNemar's Test**: Tests for significant differences in classification errors between two models
- **DeLong's Test**: Compares ROC curves by testing equality of AUC values

Significance level: $\alpha = 0.05$. A result is significant if $p < 0.05$.

5.11 Wilcoxon Signed-Rank Test Results

Table 21: Wilcoxon signed-rank test results comparing model performance on Australian Credit. Tests compare AUC scores across cross-validation folds. Effect size (Cohen's d) quantifies the magnitude of difference.

Model 1	Model 2	Statistic	p-value	Effect Size	Significant
GB-50-0.1	XGB-10-0.3	0.00	0.0625	—	×
GB-50-0.1	AdaBoost-10	0.00	0.0625	—	×
GB-50-0.1	Torch-MLP-64-32	1.00	0.1250	—	×
GB-50-0.1	BAG-CART-100	1.00	0.1441	—	×
GB-50-0.1	LGBM-10-0.3	5.00	0.6250	—	×
GB-50-0.1	RF-250-0.25	7.00	1.0000	—	×

5.12 McNemar's Test Results

Table 22: McNemar's test results comparing classification errors on Australian Credit. Tests whether the two models make significantly different types of errors.

Model 1	Model 2	Statistic	p-value	Significant
GB-50-0.1	RF-250-0.25	22.32	< 0.001	✓
GB-50-0.1	AdaBoost-10	18.37	< 0.001	✓
GB-50-0.1	Torch-MLP-64-32	16.57	< 0.001	✓
GB-50-0.1	LGBM-10-0.3	15.56	< 0.001	✓
GB-50-0.1	BAG-CART-100	13.88	< 0.001	✓
GB-50-0.1	XGB-10-0.3	0.14	0.7103	×

5.13 DeLong's Test Results

Table 23: DeLong's test results comparing ROC curves on Australian Credit. Tests whether the AUC values of two ROC curves are significantly different.

Model 1	Model 2	Statistic	p-value	Significant
GB-50-0.1	AdaBoost-10	5.512	< 0.001	✓
GB-50-0.1	Torch-MLP-64-32	5.315	< 0.001	✓
GB-50-0.1	BAG-CART-100	-3.442	< 0.001	✓
GB-50-0.1	RF-250-0.25	-3.318	< 0.001	✓
GB-50-0.1	LGBM-10-0.3	2.771	0.0056	✓
GB-50-0.1	XGB-10-0.3	1.880	0.0601	×

hmeq

Statistical significance tests were conducted to compare model performance. The following tests assess whether observed performance differences are statistically meaningful:

- **Wilcoxon Signed-Rank Test**: Non-parametric test comparing paired AUC scores across cross-validation folds
- **McNemar's Test**: Tests for significant differences in classification errors between two models
- **DeLong's Test**: Compares ROC curves by testing equality of AUC values

Significance level: $\alpha = 0.05$. A result is significant if $p < 0.05$.

5.14 Wilcoxon Signed-Rank Test Results

Table 24: Wilcoxon signed-rank test results comparing model performance on hmeq. Tests compare AUC scores across cross-validation folds. Effect size (Cohen's d) quantifies the magnitude of difference.

Model 1	Model 2	Statistic	p-value	Effect Size	Significant
LGBM-500-0.1	RF-100-sqrt	0.00	0.0625	—	×
LGBM-500-0.1	XGB-100-0.3	7.00	1.0000	—	×

5.15 McNemar's Test Results

Table 25: McNemar's test results comparing classification errors on hmeq. Tests whether the two models make significantly different types of errors.

Model 1	Model 2	Statistic	p-value	Significant
LGBM-500-0.1	XGB-100-0.3	9.03	0.0027	✓
LGBM-500-0.1	RF-100-sqrt	2.02	0.1556	×

5.16 DeLong's Test Results

Table 26: DeLong's test results comparing ROC curves on hmeq. Tests whether the AUC values of two ROC curves are significantly different.

Model 1	Model 2	Statistic	p-value	Significant
LGBM-500-0.1	XGB-100-0.3	3.277	0.0010	✓
LGBM-500-0.1	RF-100-sqrt	-0.392	0.6950	×

TH02

Statistical significance tests were conducted to compare model performance. The following tests assess whether observed performance differences are statistically meaningful:

- **Wilcoxon Signed-Rank Test**: Non-parametric test comparing paired AUC scores across cross-validation folds
- **McNemar's Test**: Tests for significant differences in classification errors between two models
- **DeLong's Test**: Compares ROC curves by testing equality of AUC values

Significance level: $\alpha = 0.05$. A result is significant if $p < 0.05$.
No statistical test results available.

MSME Credit Data by 30S-CR

Statistical significance tests were conducted to compare model performance. The following tests assess whether observed performance differences are statistically meaningful:

- **Wilcoxon Signed-Rank Test**: Non-parametric test comparing paired AUC scores across cross-validation folds
- **McNemar's Test**: Tests for significant differences in classification errors between two models
- **DeLong's Test**: Compares ROC curves by testing equality of AUC values

Significance level: $\alpha = 0.05$. A result is significant if $p < 0.05$.
No statistical test results available.

Appendix C: Full Predictive Performance Frontier

This appendix contains the complete predictive performance frontier showing AUC scores for all evaluated models across all datasets. This detailed table supports the summary performance table presented in Section 3.1.

Table 27: Predictive Performance Frontier: AUC Scores Across Datasets

Model Group	Model	credit_approval_dataset	german_credit_record	Australian Credit	hmeq	TH02	MSME Credit Data by 30S-CR
AdaBoost (Decision Tree Stumps)	AdaBoost-10	0.9547	0.7595	0.9337	0.8881	0.6140	0.7759
	AdaBoost-20	0.9552	0.7874	0.9320	0.8910	0.6249	0.7614
	AdaBoost-30	0.9527	0.7919	0.9318	0.8955	0.6410	0.7630
Bagging (CART)	BAG-CART-10	0.9518	0.7691	0.9237	0.9267	0.6101	0.7826
	BAG-CART-100	0.9472	0.7818	0.9334	0.9354	0.6065	0.7853
	BAG-CART-1000	0.9462	0.7833	0.9311	0.9347	0.6150	0.7809
	BAG-CART-20	0.9474	0.7742	0.9266	0.9319	0.6002	0.7874
	BAG-CART-250	0.9478	0.7843	0.9332	0.9340	0.6100	0.7722
	BAG-CART-50	0.9487	0.7825	0.9311	0.9328	0.6049	0.7902
	BAG-CART-500	0.9466	0.7831	0.9325	0.9337	0.6120	0.7770
Bagging (MLP)	BAG-MLP-10	0.9467	0.8133	0.9183	0.9349	0.6464	0.8182
	BAG-MLP-25	0.9493	0.8095	0.9194	0.9356	0.6420	lightgray0.8229
	BAG-MLP-5	0.9436	0.8138	0.9159	0.9314	0.6304	0.7970
Gradient Boosting	GB-10-0.1	0.9512	0.7541	0.9364	0.8949	0.6285	0.6968
	GB-100-0.1	0.9566	0.8142	0.9286	0.9312	0.6113	0.7409
	GB-1000-0.01	0.9570	0.8127	0.9306	0.9295	0.6131	0.7453
	GB-50-0.1	0.9582	0.7968	lightgray0.9383	0.9223	0.6291	0.7367
	GB-500-0.05	0.9571	lightgray0.8213	0.9188	0.9456	0.6135	0.7269
K-Nearest Neighbors	KNN-10	0.9310	0.7454	0.9109	0.8452	0.5950	0.5623
	KNN-3	0.9229	0.7402	0.8962	0.8974	0.6004	0.5387
	KNN-5	0.9255	0.7482	0.9092	0.8868	0.6007	0.4429
	KNN-7	0.9305	0.7474	0.9098	0.8712	0.6052	0.4546
LightGBM	LGBM-10-0.3	0.9513	0.7960	0.9340	0.9287	0.6228	0.7688
	LGBM-100-0.3	0.9505	0.7915	0.9131	0.9613	0.6282	0.2299
	LGBM-50-0.3	0.9517	0.7863	0.9167	0.9562	0.6379	0.2257
	LGBM-500-0.1	0.9548	0.7924	0.9165	lightgray0.9651	0.6241	0.2194
Logistic Regression	LR-LBFGS	0.9455	0.7982	0.9134	0.7719	0.6088	0.5616
	LR-Newton	0.9474	0.7982	0.9134	0.7719	0.6089	0.3122
	LR-SAGA	0.9559	0.7954	0.9131	0.7719	0.6087	0.7745
PyTorch Neural Networks	Torch-MLP-64-32	0.9550	0.8155	0.9353	0.8408	0.6386	0.7536
	Torch-TCN-64	0.9206	0.7873	0.9054	0.9201	0.6007	0.6766
	Torch-Transformer-64	lightgray0.9596	0.8014	0.9152	0.9274	0.5926	0.7503
Random Forest	RF-100-sqrt	0.9505	0.8033	0.9336	0.9619	lightgray0.6548	0.7545
	RF-250-0.25	0.9531	0.7992	0.9347	0.9598	0.6430	0.7806
	RF-500-0.5	0.9476	0.7966	0.9309	0.9448	0.6319	0.7808
Regularized Logistic Regression	LR-Reg-LBFGS	0.9567	0.8018	0.9187	0.7726	0.6133	0.7779
	LR-Reg-Liblinear	0.9570	0.7994	0.9177	0.7727	0.6140	0.7504
	LR-Reg-SAGA	0.9579	0.8009	0.9183	0.7727	0.6150	0.7623
XGBoost	XGB-10-0.3	0.9460	0.7845	0.9356	0.9224	0.6058	0.6725
	XGB-100-0.3	0.9535	0.7877	0.9240	0.9599	0.6269	0.4223
	XGB-50-0.3	0.9521	0.7963	0.9260	0.9545	0.6131	0.4834
	XGB-500-0.1	0.9562	0.7896	0.9240	0.9585	0.6333	0.5651

Note: Best performing model for each dataset is highlighted in grey.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity Checks for Saliency Maps. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 9505–9515. https://proceedings.neurips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b873d-Abstract.html

Agarwal, C., Nguyen, A., Phan, M.-T., & Abbasi-Asl, R. (2022). On the Stability of Feature Attributions. *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence (UAI)*, 41–51. https://proceedings.mlr.press/v180/agarwal22b.html

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, 23(4), 589–609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the Robustness of Interpretability Methods. *Proceedings of the 2018 ICML Workshop on Human Interpretability in Machine Learning*, 66–71. https://arxiv.org/abs/1806.08049

Apley, D. W., & Zhu, J. (2020). Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4), 1059–1086. https://doi.org/10.1111/rssb.12377

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7), e0130140. https://doi.org/10.1371/journal.pone.0130140

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

Basel Committee on Banking Supervision. (2011, June). *Principles for the Sound Management of Operational Risk* (tech. rep.). Bank for International Settlements. https://www.bis.org/publ/bcbs195.pdf

Beaver, W. H. (1966). Financial Ratios As Predictors of Failure. *Journal of Accounting Research*, 4(1), 71–111. https://doi.org/10.2307/2490171

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., & Kasneci, G. (2022). Deep Neural Networks and Tabular Data: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1), 4–26. https://doi.org/10.1109/TNNLS.2022.3228326

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730. https://doi.org/10.1145/2783258.2788613

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

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3), 273–297. https://doi.org/10.1007/BF00994018

Cover, T. M., & Hart, P. E. (1967). Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, 13(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Cybenko, G. (1989). Approximation by Superpositions of a Sigmoidal Function. *Mathematics of Control, Signals and Systems*, 2(4), 303–314. https://doi.org/10.1007/BF02551274

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves. *Biometrics*, 44(3), 837–845. https://doi.org/10.2307/2531595

Demšar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7, 1–30. https://www.jmlr.org/papers/volume7/demsar06a/demsar06a.pdf

Doshi-Velez, F., & Kim, B. (2017). Towards a Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608*. https://arxiv.org/abs/1702.08608

European Banking Authority. (2021). *Discussion Paper on Machine Learning for IRB Models* (tech. rep. No. EBA/DP/2021/04). EBA. https://www.eba.europa.eu

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5), 1189–1232. https://www.jstor.org/stable/2699986

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics*, 24(1), 44–65. https://doi.org/10.1080/10618600.2014.907095

Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2021). Revisiting Deep Learning Models for Tabular Data. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 18932–18943. https://proceedings.neurips.cc/paper/2021/hash/9d86d8dce5936cfa745633836181e434-Abstract.html

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys (CSUR)*, 51(5), 1–42. https://doi.org/10.1145/3236009

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

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Hooker, S., Erhan, D., Kindermans, P.-J., & Kim, B. (2019). A Benchmark for Interpretability Methods in Deep Neural Networks. *Advances in Neural Information Processing Systems (NeurIPS)*. https://proceedings.neurips.cc/paper/2019/hash/fe4b8556000d0f0cae99dadd8d363008-Abstract.html

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer Feedforward Networks Are Universal Approximators. *Neural Networks*, 2(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Huang, X., Khetan, A., Cvitkovic, M., & Karnin, Z. (2020). TabTransformer: Tabular Data Modeling Using Contextual Embeddings. *arXiv preprint arXiv:2012.06678*. https://arxiv.org/abs/2012.06678

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer Credit Risk Models via Machine-Learning Algorithms. *Journal of Banking & Finance*, 34(11), 2767–2787. https://doi.org/10.1016/j.jbankfin.2010.06.002

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature*, 521, 436–444. https://www.nature.com/articles/nature14539

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

Li, X., & Wu, Y. (2024). Advanced Post-Hoc Interpretability in Financial Modelling. *Journal of Financial Data Science*, 6(1), 10–25. https://jdsa.org/

Louzada, F., Ara, A., & Fernandes, G. B. (2016). Binary Classification Methods for Credit Scoring: A Systematic Review and Empirical Analysis. *Expert Systems with Applications*, 59, 117–136. https://doi.org/10.1016/j.eswa.2016.02.039

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

McNemar, Q. (1947). Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika*, 12(2), 153–157. https://doi.org/10.1007/BF02295996

Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, 18(1), 109–131. https://www.jstor.org/stable/2490395

Quan, J., & Sun, X. (2024). Credit Risk Assessment Using the Factorization Machine Model with Feature Interactions [doi:10.1057/s41599-024-02700-7]. *Humanities and Social Sciences Communications*, 11(234). https://doi.org/10.1057/s41599-024-02700-7

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

Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-y

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323, 533–536. https://www.nature.com/articles/323533a0

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

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2), 841–887. https://jolt.law.harvard.edu/articles/pdf/v31/31HarvJLTech841.pdf

Wang, C., Zhang, K., & Wang, H. (2025). Interpretable Credit Risk Modelling: Foundations, Challenges, and Future Directions. *Decision Support Systems*, 181, 113902. https://doi.org/10.1016/j.dss.2023.113902

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

Wang, X., Li, Y., & Zhang, Q. (2025). Explainable Deep Credit Scoring Under Regulatory Constraints [Forthcoming]. *Decision Support Systems (Forthcoming)*.

Wiginton, J. C. (1980). A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior. *Journal of Financial and Quantitative Analysis*, 15(3), 757–770. https://doi.org/10.2307/2330626

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

Yildirim, G., & Kulekci, M. O. (2021). Graph Neural Networks for Credit Risk Analysis. *Expert Systems with Applications*, 186, 115822. https://doi.org/10.1016/j.eswa.2021.115822

Zeng, G., Su, W., & Hong, C. (2024). Ensemble Learning with Feature Optimization for Credit Risk Assessment. *Research Square Preprint*. https://doi.org/10.21203/rs.3.rs-4665987/v1