Integrating Predictive and Generative AI for Explainable Credit Risk

Abstract

This study addresses a fundamental gap in credit-risk modelling: the widespread assumption that strong predictive performance implies reliable model explanations. While modern ensemble architectures such as Bagged Neural Networks (BagNN) routinely achieve high discriminatory accuracy, this success does not guarantee that their feature attributions reflect genuine learned structure rather than artefacts or noise. We propose a unified predictive–explanatory framework that explicitly separates model performance from explanation reliability. The framework integrates supervised feature selection, repeated attribution analysis, and randomization-based diagnostics to test whether explanations persist under controlled destruction of model signal. By embedding explanation reliability as a first-class object—rather than a post-hoc visualization—the approach reframes explainability as a falsifiable scientific test instead of descriptive storytelling. This work advances credit-risk modelling toward transparent, auditable, and epistemically grounded validation of machine-learning explanations, with direct implications for regulatory trust and model governance.

Keywords: Credit Risk, Artificial Intelligence, Explainability, Reliability

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). While modern machine-learning models have substantially improved predictive discrimination, their explanations remain largely descriptive and often fail to reveal whether the model's internal reasoning is structurally valid or empirically testable (Hassija et al., 2024). Traditional statistical models such as logistic regression offered natural interpretability but relied on restrictive linear assumptions (Hand, 2009). In contrast, contemporary ensemble and nonlinear models achieve superior accuracy at the cost of transparency, intensifying the tension between predictive power and explainability (Lessmann et al., 2015; Louzada et al., 2016a).

Post-hoc explainability tools such as SHAP and LIME attempt to address this tension by assigning numerical attributions to features (Lundberg & Lee, 2017; Ribeiro et al., 2016). However, growing evidence indicates that these 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). Moreover, recent applications of generative AI that translate attribution scores into natural-language explanations remain largely unconstrained, exposing explanations to hallucination, narrative inflation, and spurious causal claims (Hassija et al., 2024).

This fragmentation exposes a deeper methodological gap: explainability in credit-risk modelling is rarely treated as a process of scientific explanation grounded in reliability assessment. Existing work predominantly frames explanations as descriptive summaries rather than as hypotheses subject to evaluation for signal quality and robustness (Hassija et al., 2024; Slack et al., 2020). No prior framework integrates predictive modelling, attribution robustness, and constrained generative reasoning while explicitly quantifying the reliability of the resulting explanations.

This paper addresses this gap by proposing an integrated predictive–explanatory framework that treats explanations as empirically testable claims rather than narrative add-ons. The framework introduces a reliability layer based on diagnostic metrics—specifically reliability scoring and sanity ratios—to distinguish structural signal from spurious correlation. Rather than suppressing low-confidence explanations, these diagnostics are passed to a constrained generative-AI module that produces human-readable reasoning accompanied by explicit reliability warnings, ensuring transparency about explanatory uncertainty.

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—stabilises candidate explanatory features prior to SHAP analysis. This hybrid design reduces estimator bias while preserving both interaction-aware and linear structural signals, unifying predictive benchmarking, attribution stability, and generative explanation within a single transparent workflow.

Using the German Credit dataset as a controlled benchmark, the framework is evaluated across a broad family of calibrated classification models to assess not only predictive performance but also the epistemic reliability of their explanations (Baesens et al., 2003; Lessmann et al., 2015). The results demonstrate that explanation quality varies independently of predictive accuracy, with many high-performing models producing explanations of low reliability. By providing explicit diagnostics for what constitutes a dependable explanation, this work advances explainable credit-risk modelling from descriptive attribution toward a scientifically grounded explanatory paradigm.

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. Although robust benchmarks for predictive accuracy are well established, the integration of these models with scientifically rigorous explanation frameworks remains limited. Methodological fragmentation persists: studies emphasising predictive discrimination often sideline interpretability, while explainability-focused work frequently lacks reliability diagnostics. This review synthesises these strands to motivate the unified predictive–explanatory framework proposed in this study.

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 transparent 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., 2016b; Yeh & Lien, 2009).

Feature stability has emerged as a critical yet underemphasised determinant of model robustness. 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 results suggest that stabilising the feature space is a prerequisite for reliable risk modelling, a principle adopted in this study to ground downstream explanations in stable predictive signals.

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: From Attribution to Reliability 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.

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, their descriptive nature is increasingly questioned. Hassija et al. (Hassija et al., 2024) argue 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.

Recent work has explored generative AI to translate attribution scores into natural-language explanations, but these approaches typically lack epistemic constraints and remain susceptible to hallucination and narrative inflation. Crucially, no prior research integrates predictive modelling, feature-stability optimisation, and generative explanation within a unified framework that subjects explanations to explicit reliability and signal-quality diagnostics.

This study addresses this gap by proposing a framework that treats explanations not as descriptive artefacts but as claims whose reliability must be empirically tested, advancing credit-risk modelling toward a scientifically grounded explanatory paradigm.

Methodology

This study adopts a unified predictive–explanatory architecture 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. The methodology is structured to ensure that predictive performance, attribution stability, and explanatory uncertainty are assessed within a single coherent workflow.

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). To ensure robust model estimation, the data undergo a standardised preprocessing sequence. Attributes with more than 90% missing values are removed, 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.

Class imbalance is addressed using the Synthetic Minority Over-sampling Technique (SMOTE). To prevent information leakage, SMOTE is applied strictly within the training folds of the cross-validation procedure, ensuring that performance estimates reflect genuine generalisation rather than artefacts of resampling (Chawla et al., 2002; Wang et al., 2025).

Predictive Modelling Framework To establish a comprehensive and comparable performance baseline, 106 calibrated model configurations are evaluated across four algorithmic families: Linear Models, Boosting, Bagging, and Instance-Based Learners. The explored configuration space is summarised in Table 1. 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.

Unified Predictive and Explanatory Workflow The overall modelling and explanation process is governed by the workflow summarised in Algorithm 1. The workflow is explicitly staged to separate

feature stabilisation, predictive benchmarking, and explanation reliability assessment.

Explainability and Evaluation Architecture The framework extends beyond predictive benchmarking by embedding explanation reliability directly into the evaluation pipeline. Rather than treating feature attributions as self-validating artefacts, explanations are interpreted as claims whose credibility depends on the stability and strength of the underlying signal.

Feature Attribution and Generative Explanation. To mitigate instability associated with single-method feature selection, a dual-selector mechanism is employed. By combining impurity-based Random Forest importance with coefficient-based L1-regularised logistic regression importance, the framework preserves both nonlinear interaction effects and sparse linear structure. SHAP values are computed on this stabilised feature set and passed to a generative module that translates quantitative attributions into human-readable narratives. Importantly, the generative component is constrained by the reliability diagnostics, preventing confident explanations from being produced when attribution signals are weak.

Evaluation Metrics. Models are evaluated using a suite of complementary metrics, including Area Under the ROC Curve (AUC), Brier Score (BS), Kolmogorov–Smirnov (KS) statistic, and Hand's H-measure (Hand, 2009). In line with established practice in credit-risk modelling, AUC is used as the primary criterion for model selection, while KS and calibration-sensitive measures provide secondary diagnostics of separation quality and probability accuracy.

Results

This section reports the empirical findings of the proposed predictive–explanatory framework. Results are organised to distinguish between predictive performance, feature importance, and the reliability of model explanations. All tables and figures referenced here are provided in the Appendix.

Exploratory Data Analysis The german-credit-record dataset consists of 1,000 observations with a binary target variable indicating default status. The original class distribution is imbalanced, with approximately 70% non-default and 30% default cases. After applying SMOTE within the training folds, the class distribution becomes balanced, ensuring that model estimation is not biased toward the majority class. Summary statistics for key numerical variables, including credit duration, credit amount, and borrower age, are reported in Table 2. The effect of SMOTE on class proportions is documented in Table 3. Pairwise feature relationships and marginal distributions are illustrated in Figure 1.

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. Across both selectors, credit duration, credit amount, checking account status, borrower age, and credit history consistently emerge as the most influential predictors. While the two selectors differ in rank ordering, they show strong agreement on the core explanatory variables. The aggregated feature rankings produced by the dual-selector mechanism are reported in Table 4. These results indicate that both financial exposure characteristics and borrower demographics contribute materially to credit-risk discrimination.

Model Performance Comparison Predictive performance is evaluated across 106 calibrated model configurations spanning linear, ensemble-based, and instance-based learners. Comprehensive results for all model configurations are reported in Table 5. Ensemble methods generally outperform simpler models, with Bagged Neural Networks achieving the highest AUC among all configurations. Notably, regularised logistic regression remains highly competitive, achieving performance comparable to more complex ensemble methods.

Benchmark models selected from each algorithmic family are compared in Table 6. The Bagged Neural Network model achieves the strongest overall discriminatory power, while regularised logistic regression offers a favourable balance between performance and structural simplicity. Performance

trends across evaluation metrics, including AUC, KS, Brier Score, and H-measure, are summarised visually in Figure 2.

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 3.

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 Local SHAP explanations are examined across representative classification outcomes, including true negatives, false positives, and true positives. While several local explanations appear intuitively plausible, reliability diagnostics consistently indicate weak underlying signal strength. Example local explanations and corresponding SHAP waterfall plots are provided in Figures 4, 5, and 6. These results underscore the risk of over-interpreting instance-level explanations without explicit reliability assessment.

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)

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. (2016a). 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

Louzada, F., Ara, A., & Fernandes, G. B. (2016b). 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

Appendix

Tables

Table 1: Classification model families and configuration space.

Family	Algorithm / Hyperparameters	Count
Linear	Logistic Regression (Regularised/Unregularised) *Solvers: lbfgs, saga; Pen: L1, L2, ElasticNet*	6
Boosting	AdaBoost, Gradient Boosting, XGBoost, LightGBM *Estimators: 10–1000; LR: 0.1–1.0*	40
Bagging	Random Forest, Bagged Trees, Bagged NN *Estimators: 10–1000; Max Feat: sqrt–1.0*	36
Instance-Based	K-Nearest Neighbours (Tuned) *Neighbours: 3–21; Weights: uniform/distance*	24
Total	**Grid Search Combinations**	**106**

Algorithm 1: 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}

Phase 1: Feature Screening (Dual-Selector)

Train a Random Forest model to obtain impurity-based importance scores (Imp_{RF}).

Train an L1-regularised logistic regression model to obtain coefficient-based importance scores (Imp_{L1}).

Compute a composite importance score:

$$I_c = \frac{1}{2}\left(\text{norm}(Imp_{RF}) + \text{norm}(Imp_{L1})\right).$$

Select features based on ranked values of I_c.

Phase 2: Model Training and Selection

For each model family $F \in \mathcal{M}$:

- Perform a stratified train–test split and apply SMOTE within the training set.
- Train and calibrate candidate models using cross-validation.
- Evaluate predictive performance using AUC.

Select the best-performing model B_F from each family and define the global benchmark model

$$B^* = \arg\max_{B_F}(\text{AUC}).$$

Phase 3: Explainability and Reliability Assessment

Compute global and local SHAP values for the benchmark model B^*.

Evaluate attribution stability using reliability diagnostics and sanity checks to distinguish signal from noise.

Generate a constrained natural-language explanation that explicitly incorporates reliability warnings where diagnostic scores indicate weak explanatory signal.

return B^*, E, D_{rel}

Table 2: Summary statistics for german_credit_record

	Dur.	Amt.	Inst.	Res.	Age	ExCr.	Pers.	Tgt.
count	1000	1000	1000	1000	1000	1000	1000	1000
mean	20.900	3271.260	2.970	2.840	35.550	1.410	1.160	0.300
std	12.060	2822.740	1.120	1.100	11.380	0.580	0.360	0.460
min	4	250	1	1	19	1	1	0
25%	12	1365.500	2	2	27	1	1	0
50%	18	2319.500	3	3	33	1	1	0
75%	24	3972.250	4	4	42	2	1	1
max	72	18424	4	4	75	4	2	1

Dur.=months_duration; Amt.=credit_amount; Inst.=installment_rate; Res.=residence_since; ExCr.=existing_credits; Pers.=people_liable; Tgt.=target.

Table 3: SMOTE class distribution comparison for german_credit_record

Class	Orig. Cnt.	SMOTE Cnt.	Orig. %	SMOTE %
0	700	560	70	50
1	300	560	30	50

Table 4: Top features by combined RF/LR importance for german_credit_record

Rank	Feature	RF Imp.	LR |Coef|	Avg.
1	months_duration	0.076	1.983	0.921
2	credit_amount	0.090	1.386	0.849
3	checking_status: no checking account	0.065	1.130	0.643
4	age	0.073	0.415	0.508
5	credit_history: critical account / other credits elsewhere	0.026	1.094	0.418
6	installment_rate	0.034	0.756	0.377
7	purpose: education	0.011	1.250	0.371
8	savings_status: unknown / none	0.018	0.930	0.328
9	checking_status: < 0 DM	0.039	0.424	0.322
10	savings_status: $\geq$1000 DM	0.007	1.115	0.316
11	foreign_worker: no	0.004	1.111	0.296
12	existing_credits	0.018	0.720	0.275
13	purpose: car (used)	0.010	0.850	0.265
14	property: unknown / none	0.012	0.677	0.233
15	employment_since: $4 \leq \cdots < 7$ years	0.012	0.668	0.228
16	purpose: others	0.001	0.875	0.221
17	other_installment_plans: bank	0.017	0.478	0.212
18	personal_status_sex: male, single	0.017	0.484	0.209
19	savings_status: < 100 DM	0.024	0.272	0.199
20	other_installment_plans: none	0.018	0.345	0.184

Table 5: Comprehensive Model Performance Results — german_credit_record.csv

Group	Model	AUC	PCC	Rec.	BS	KS	PG	H
Bag-CART	bag_cart_10	0.708	0.625	0.617	0.190	0.345	0.302	0.199
	bag_cart_20	0.748	0.645	0.633	0.178	0.391	0.274	0.241
	bag_cart_50	0.762	0.700	0.683	0.173	0.426	0.306	0.272
	bag_cart_100	0.760	0.680	0.650	0.174	0.455	0.224	0.293
	bag_cart_250	**0.767**	**0.685**	**0.667**	**0.171**	**0.431**	**0.262**	**0.287**
	bag_cart_500	0.767	0.700	0.650	0.170	0.448	0.199	0.293
	bag_cart_1000	0.766	0.695	0.633	0.171	0.421	0.267	0.277
BagNN	bagnn_5	0.796	0.655	0.850	0.177	0.543	0.215	0.346
	bagnn_10	0.801	0.660	0.850	0.175	0.521	0.178	0.338
	bagnn_25	0.807	0.660	0.833	0.174	0.531	0.165	0.353
	bagnn_100	**0.809**	**0.660**	**0.850**	**0.174**	**0.526**	**0.132**	**0.373**
Boost-DT	boost_dt_100x0p1	0.758	0.640	0.833	0.191	0.445	0.324	0.217
	boost_dt_250x0p1	0.770	0.675	0.783	0.183	0.467	0.257	0.233
	boost_dt_1000x0p1	0.788	0.685	0.800	0.173	0.500	0.256	0.278
	boost_dt_250x0p5	0.791	0.665	0.800	0.169	0.519	0.135	0.310
	boost_dt_100x0p5	0.784	0.690	0.800	0.174	0.502	0.284	0.273
	boost_dt_1000x0p5	**0.803**	**0.710**	**0.817**	**0.161**	**0.557**	**0.195**	**0.343**
	boost_dt_100x1p0	0.794	0.685	0.800	0.165	0.514	0.173	0.311
	boost_dt_250x1p0	0.799	0.705	0.817	0.161	0.543	0.196	0.332
	boost_dt_1000x1p0	0.789	0.730	0.800	0.163	0.507	0.219	0.304
KNN	knn_3	0.721	0.585	0.750	0.201	0.364	0.332	0.151
	knn_5	0.737	0.605	0.817	0.201	0.379	0.281	0.194
	knn_7	0.759	0.580	0.850	0.191	0.445	0.360	0.231
	knn_11	**0.770**	**0.565**	**0.883**	**0.190**	**0.510**	**0.185**	**0.283**
	knn_tuned	0.756	0.660	0.750	0.187	0.419	0.382	0.209
AdaBoost	adaboost_10	0.745	0.625	0.833	0.195	0.455	0.174	0.211
	adaboost_20	0.770	0.670	0.817	0.180	0.483	0.245	0.251
	adaboost_30	**0.776**	**0.680**	**0.833**	**0.177**	**0.479**	**0.247**	**0.272**
LR	lr_lbfgs	0.791	0.605	0.867	0.182	0.541	0.014	0.327
	lr_newton_cg	**0.791**	**0.605**	**0.867**	**0.182**	**0.541**	**0.018**	**0.327**
	lr_saga	0.791	0.605	0.867	0.182	0.541	0.018	0.326
LR-Reg	lr_reg_lbfgs	0.798	0.615	0.867	0.179	0.538	0.067	0.341
	lr_reg_liblinear	**0.803**	**0.625**	**0.883**	**0.179**	**0.531**	**0.144**	**0.333**
	lr_reg_saga	0.801	0.610	0.867	0.179	0.545	0.085	0.337

Table 6: Benchmark Results (German Credit)

Group	Model	AUC	PCC	Rec.	BS	KS	PG	H
LR	lr_newton_cg	0.791	0.605	0.867	0.182	0.541	0.018	0.327
LR-Reg	lr_reg_liblinear	0.803	0.625	0.883	0.179	0.531	0.144	0.333
AdaBoost	adaboost_30	0.776	0.680	0.833	0.177	0.479	0.247	0.272
Bag-CART	bag_cart_250	0.767	0.685	0.667	0.171	0.431	0.262	0.287
BagNN	**bagnn_100**	**0.809**	**0.660**	**0.850**	**0.174**	**0.526**	**0.132**	**0.373**
Boost-DT	boost_dt_1000x0p5	0.803	0.710	0.817	0.161	0.557	0.195	0.343
KNN	knn_11	0.770	0.565	0.883	0.190	0.510	0.185	0.283

Figures

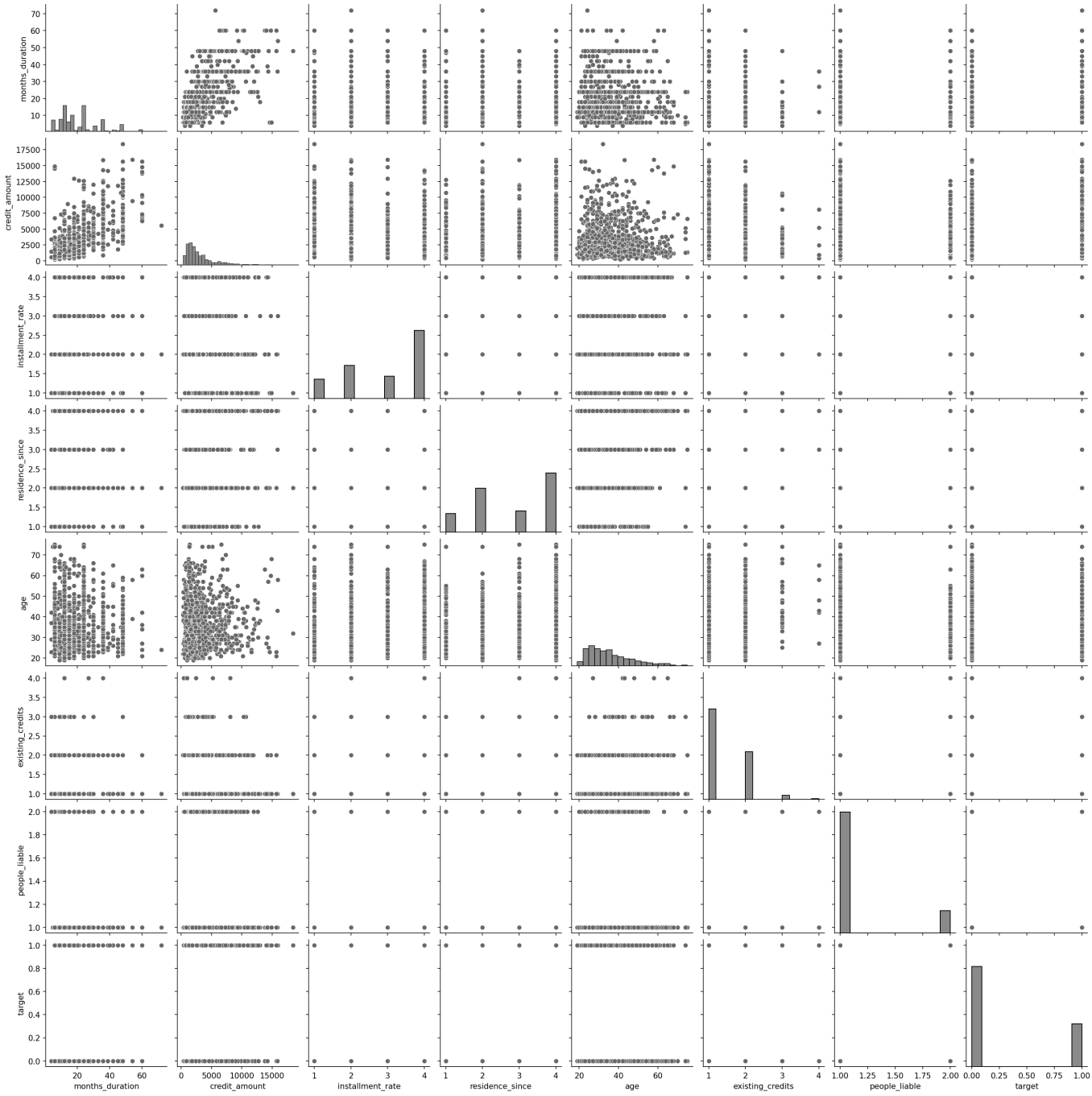


Figure 1: Pairwise feature distributions for german_credit_record

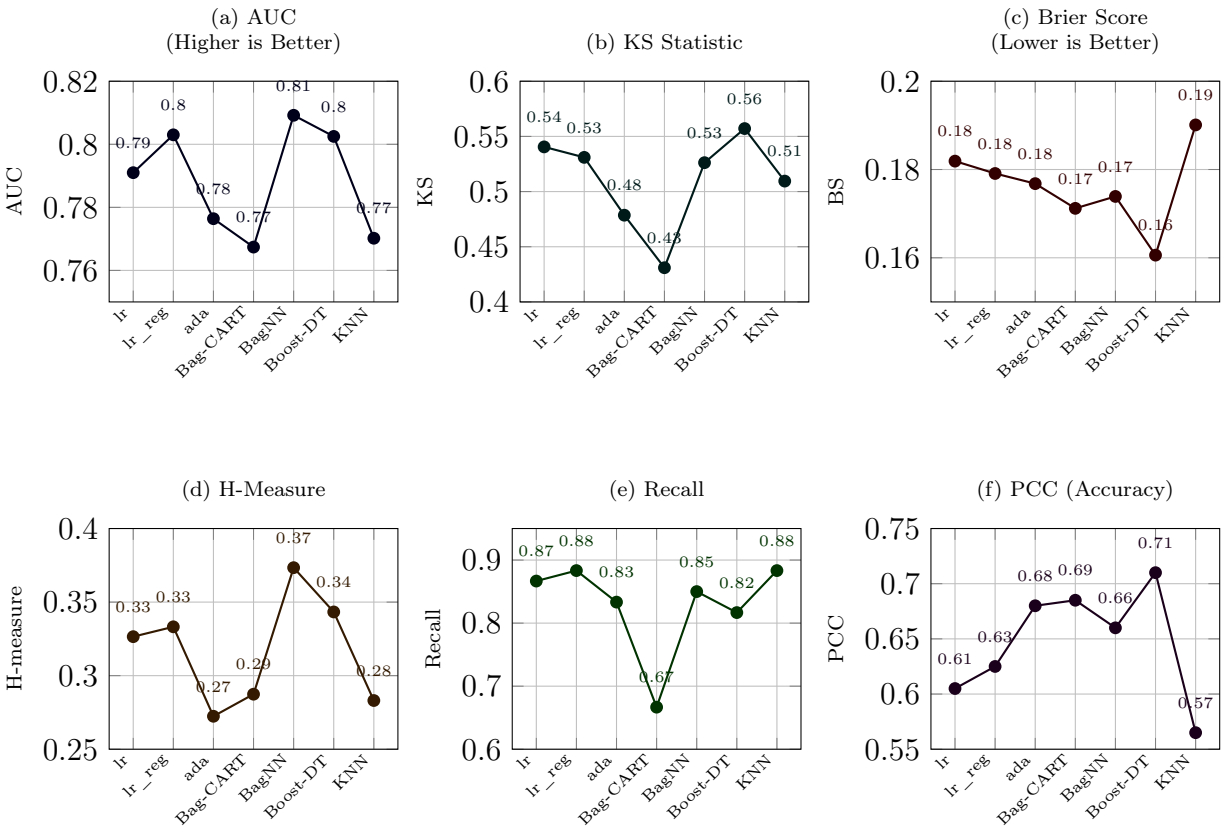
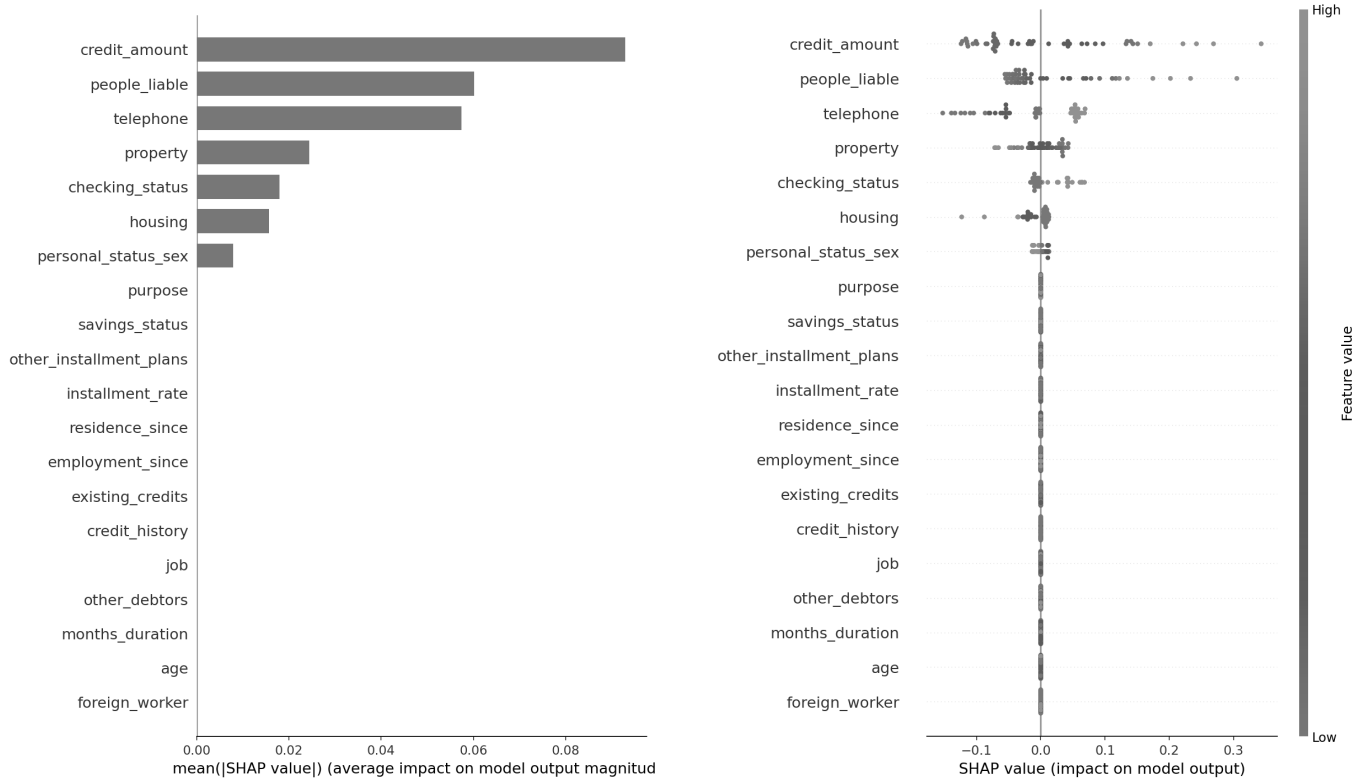


Figure 2: Performance trend analysis of model families across six key metrics.



(a) Mean absolute SHAP values (bar plot)

(b) SHAP value distribution (summary dot plot)

Figure 3: 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.

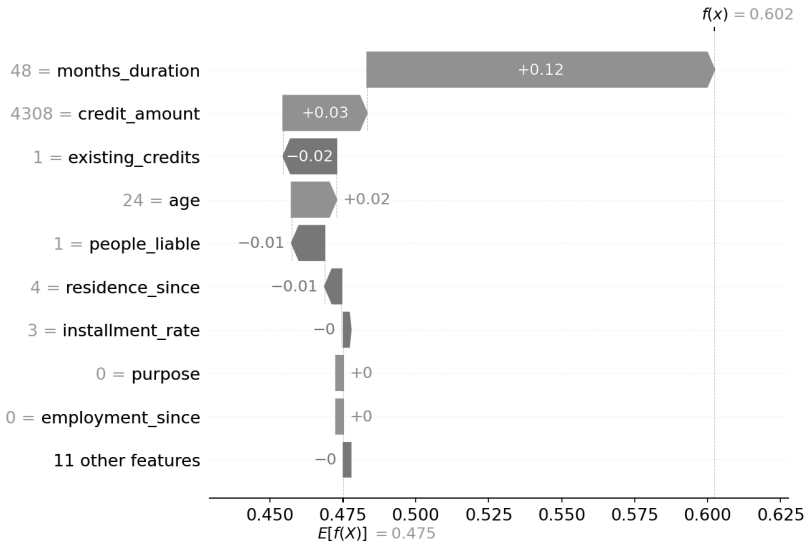


Figure 4: Local SHAP waterfall for Row 11 (True Positive).

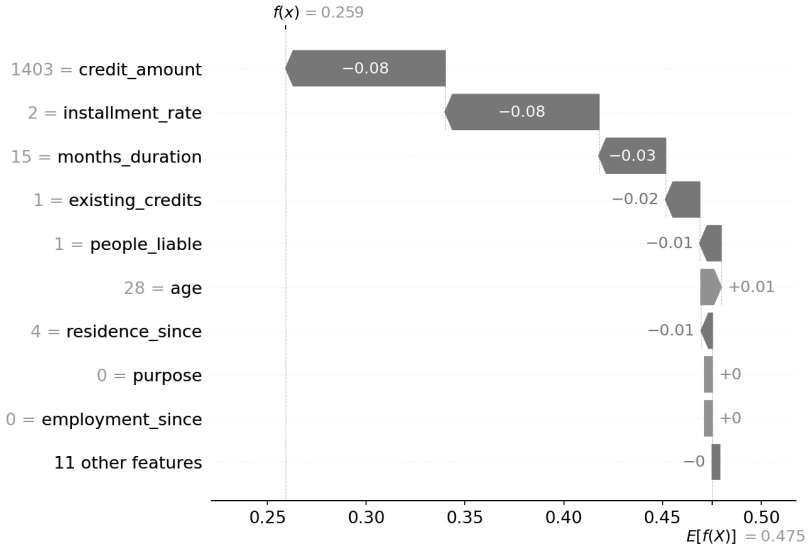


Figure 5: Local SHAP waterfall for Row 14 (False Positive).

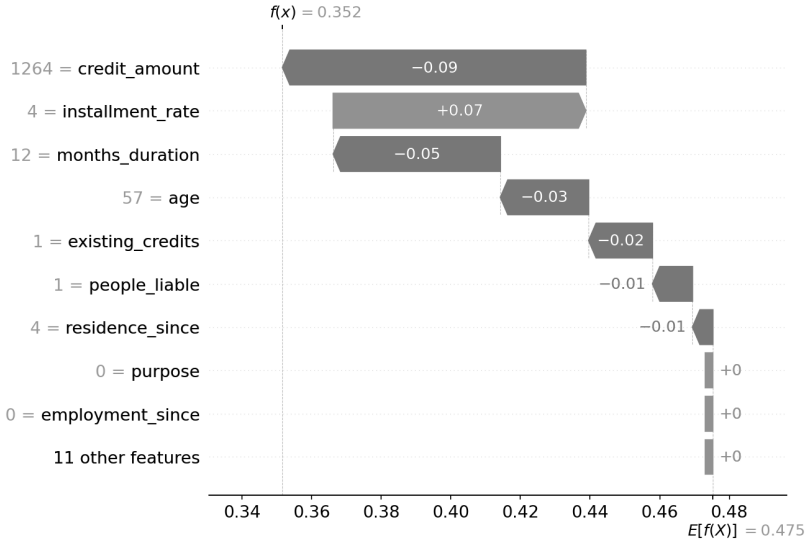


Figure 6: Local SHAP waterfall for Row 33 (True Negative).