SSR Journal of Artificial Intelligence (SSRJAI)

OPEN ACCESS

Volume 2, Issue 3, 2025 Journal homepage: https://ssrpublisher.com/ssrjai/ ISSN: 3049-0413
Email: office.ssrpublisher@gmail.com

Explainable AI for Credit Scoring with SHAP-Calibrated Ensembles: A Multi-Market Evaluation on Public Lending Data

Abayomi Oluwaseun Japinye & Adesola Anthony Adedugbe

Compliance Department, Central Bank of Nigeria

Received: *25.08.2025* | ***Accepted:*** *17.09.2025* | ***Published:*** *18.09.2025*

***Corresponding author:** Abayomi Oluwaseun Japinye

DOI: *10.5281/zenodo.17155174*

Abstract

Original Research Article

Rapid digitisation has reshaped consumer lending, with machine learning systems now central to underwriting decisions. This transition has improved prediction accuracy while creating concerns about opacity, fairness, and regulatory compliance. The study developed an explainability-first framework for credit scoring that integrates calibrated gradient-boosting models with SHAP and LIME explanations, cost-aware threshold selection, and multi-criteria fairness monitoring. This framework was evaluated across three public lending datasets representing different data-richness environments: Home Credit Default Risk (N=307,511, default rate 8.07%), Default of Credit Card Clients (N=30,000, default rate 22.12%), and LendingClub (N=887,379, default rate 5.63%). XGBoost with SHAP achieves an AUC of 0.892±0.009 to 0.923±0.008 across datasets while maintaining explanation stability (Kendall τ=0.94±0.03) and good calibration (Brier score 0.119±0.003 to 0.154±0.004). Fairness-constrained thresholding reduces demographic-parity gaps by 59-67% (95% CI: 52-74%) with cost increases of 3.2±0.8% to 5.8±1.3%. A complete reproducibility artefact, including code repository, model cards, adverse-action templates, and governance frameworks, was provided. Code and data processing scripts are available at [repository URL].

Keywords: Explainable AI, Credit scoring, SHAP, LIME, Calibration, Fairness, Machine learning.

1.0 Introduction

The expansion of financial technology has fundamentally altered credit assessment processes across income levels and geographies. Modern scoring models incorporate diverse tabular signals from formal credit histories, transactional behaviours, and digitally mediated activities. This evolution has produced underwriting systems that are both broader in scope and faster in operation than traditional bureau-centric approaches (Mhlanga, 2021; Babaei et al., 2023).

While the enrichment of predictive models with additional parameters can enhance accuracy, it concurrently impairs the ability of human analysts to understand the underlying mechanisms of the models. This degradation of interpretability engenders multiple operational risks, compromises the efficacy of consumer recourse pathways, and complicates adherence to evolving regulatory mandates. Furthermore, it stands in plain contrast to the intelligibility expectations of regulators, credit risk officers, and borrowing households, all of whom demand that decisions made by automated systems remain transparent and subject to human scrutiny.

Contemporary progress in explainable artificial intelligence, particularly through algorithmic exposition techniques, offers viable ameliorative pathways. One strand, based upon Shapley-value decomposition, supplies stable local contribution scores paired with globally coherent summaries that are derived under relaxed parametric assumptions (Lundberg & Lee, 2017). Another, the locally linear modelling perturbation framework, embraces a model-agnostic paradigm by reporting the sensitivity of predictions to perturbation samples, thus furnishing focused local proximity diagnoses. Empirical investigations within the credit and operational risk domains substantiate the premise that integrating interpretive optics incurs only a modest and controlled decrement in predictive accuracy when the techniques are deployed with methodological rigour (Bussmann et al., 2020; Babaei et al., 2023).

1.1 Research Questions and Hypotheses

This paper addresses three persistent gaps in the literature through specific research questions:

RQ1: Can gradient-boosting models with post-hoc explanations achieve superior calibration compared to inherently interpretable models while maintaining discrimination performance?

RQ2: Do SHAP explanations remain stable across bootstrap resamples and provide consistent feature importance rankings across different data-richness environments?

RQ3: Can fairness constraints be incorporated into threshold selection with measurable bias reduction at acceptable cost increases?

We test the following hypotheses:

H1: XGBoost with isotonic calibration will yield superior Brier scores compared to logistic regression while maintaining equivalent or superior AUC across all data environments.

H2: SHAP global feature importance rankings will demonstrate high stability (Kendall $\tau > 0.90$) across 1,000 bootstrap resamples and maintain coherence with local attributions.

H3: Fairness-constrained threshold optimisation will reduce demographic parity gaps by at least 50% while limiting cost increases to under 10% across available protected attributes.

H4: Alternative data features will show greater marginal importance in limited-bureau environments compared to data-rich environments, as measured by ablation analysis.

We propose a framework that integrates explanation, calibration, threshold selection, and fairness constraints from the outset. We evaluate this framework across three public datasets representing different data-richness scenarios and report both model performance and decision quality metrics.

Our contributions are: (1) a disciplined, auditable scoring architecture coupling SHAP and LIME with cost-aware thresholding and fairness constraints; (2) evaluation across diverse data environments with comprehensive performance metrics including AUC, Brier score, and explanation stability; (3) decision-support analysis mapping fairness tolerances to operating points with sensitivity analysis; and (4) a complete governance package with model cards, monitoring triggers, and adverse-action documentation supporting regulatory compliance. All analyses evaluate data-rich environments rather than countries; the public datasets do not contain country identifiers, and we therefore refrain from cross-country claims.

2.0 Related Literature

2.1 Machine Learning for Credit Scoring

Ensemble methods, particularly gradient boosting, consistently achieve superior performance on tabular lending datasets due to their ability to model non-linearities and feature interactions without extensive engineering. XGBoost and LightGBM have become standard choices for credit risk modelling across financial institutions (Chen & Guestrin, 2016; Ke et al., 2017). Comparative studies on lending datasets consistently show AUC improvements of 0.05-0.15 over logistic regression baselines (Lessmann et al., 2015; Babaei et al., 2023).

Recent work has explored monotonic constraints in gradient boosting to improve interpretability while preserving performance (Milionis et al., 2022). However, these approaches still require post-hoc explanation methods for individual prediction reasoning, making SHAP integration essential for regulatory compliance.

2.2 Explainability for Financial Decisions

Local attributions produced by SHAP derive from a precise decomposition of model predictions, guaranteeing both additive properties and consistency constraints, which permits straightforward aggregation to global feature importance scores (Lundberg & Lee, 2017). The TreeSHAP implementation optimises evaluation of gradient-boosting trees by realising computational complexity as a polynomial function of tree depth, a non-monotonic and tractable alternative to exponential complexity (Lundberg et al., 2020). Empirical results in risk management settings document seamless integration of SHAP within production scoring pipelines, preserving discrimination metrics while delivering actionable, domain-expertise-readable rationale to credit committees and capital-adequacy teams (Bussmann et al., 2020).

LIME, by contrast, retains model-agnosticity and constructs local empirical approximations via sparse linear fit within a neighbourhood of the instance to be explained (Ribeiro et al., 2016). Although the theoretical underpinning is subordinate to convexity constraints and sampling noise, LIME serves a dual function: it quantifies first-order strength of input features and surfaces latent model discrepancies, such as interaction of margin and perturbed fidelity, which regimented global inspections may overlook (Guidotti et al., 2018).

The practitioner community is increasingly alarmed by the variability of explanation results under small input perturbations, an instability that Alvarez-Melis & Jaakkola (2018) quantify via Lipschitz-bound metrics on gradient-based methods. Further, Krishna et al. (2022) aggregate instances to argue that stability metrics vary not only across model types but also across statistical properties of training data. To mediate reputational risk, incipient standards recommend that deployment libraries and auditable pipelines proactively catalogue stability diagnostics summarised across temporal splits, perturbation scenarios, and model lifecycle versions (Wang & Wang, 2025).

2.3 Probability Calibration in Credit Scoring

Accurate probability estimation is a prerequisite for economically viable lending operations, as it supports rational threshold-setting and enables precise determination of expected loss reserves. Among the available approaches, isotonic regression has emerged as a reliable vehicle for aligning the level of certainty exhibited by ensemble trees, a class of predictors known for systematically overstating certainty (Niculescu-Mizil & Caruana, 2005).

Empirical studies in retail credit underscore the magnitude of the calibration challenge: a probability forecast might yield identical area-under-the-curve (AUC) statistics while still imposing markedly divergent cost profiles. In this context, Bravo et al. (2022) estimate that the expected cost associated with a miscalibrated score could surpass the well-calibrated benchmark by 15- to 25-per cent, a deviation arising chiefly from unwarranted confidence in borderline decisions. Moving beyond nominal AUC assessments, Bella et al. (2013) advocate the concurrent monitoring of several calibration diagnostics, specifically, the Brier score, Expected Calibration Error (ECE), and graphical reliability plots, thereby furnishing a multidimensional check on the stability of forecast probabilities.

2.4 Fairness in Credit Scoring and Adverse Action Requirements

Algorithmic fairness in the lending sector engages an array of definitions, namely, demographic parity, equalised odds, and calibration within subgroups whose interplay remains contested within the literature (Barocas et al., 2019). Simultaneously, the Equal Credit Opportunity Act obliges creditors to supply precise, actionable justifications for every adverse decision, elevating the status of interpretable models from a strategic asset to a statutory requirement for compliance (Federal Reserve Board, 2022).

Fairness-constrained optimisation offers a structured mechanism for reconciling predictive merit and equity by imposing parity constraints at the threshold-selection stage rather than at the model-training stage (Hardt et al., 2016). This post-hoc recalibration preserves the integrity of the underlying predictive model while subjecting its operational cut-off to equity-oriented modifications. Recent empirical evidence from Dwork et al. (2021) further substantiates the proposition that threshold adjustments outperform training-time constraints on demographic parity when the application is constrained credit adjudication.

Proxy discrimination continues to pose a formidable hurdle; variables that appear neutral to the analyst disproportionately correlate with safeguarded attributes (Kusner et al., 2017). Data drawn from alternative or non-traditional sources, often indispensable for populations with sparse credit histories, carries latent socioeconomic proxies, thereby risking the entrenchment of accumulated disparity (Goodman, 2022). The resultant imperative is an enduring regime of assessment and auditing that outlasts the initial fairness check, conforming to the evolving contours of regulatory doctrine and social obligation.

2.5 Comparison with Inherently Interpretable Models

While post-hoc explanation methods enable complex model interpretation, inherently interpretable alternatives deserve consideration. Generalised Additive Models (GAMs) provide shape function interpretability with performance approaching ensemble methods on some datasets (Lou et al., 2013). Monotonic neural networks can incorporate domain knowledge while maintaining differentiability (Wehenkel & Louppe, 2019).

However, scorecard models remain the most widely deployed interpretable approach in credit scoring. Recent work by Naeem et al. (2018) shows that modern scorecard optimisation can achieve competitive performance while maintaining complete transparency. The choice between inherently interpretable and post-hoc explainable models involves trade-offs between performance, transparency depth, and regulatory acceptance that vary by institutional context.

3.0 Data and Variables

3.1 Dataset Specifications

Three publicly accessible datasets that are in widespread usage and reflect a range of data-rich situations were utilised in this research. Complete dataset characteristics are provided in Table 1 and Table 1a.

Table 1: Dataset Characteristics and Availability

Dataset	Source	Size	Default Rate	Time Period	Protected Attributes Available
Home Credit Default Risk	(*Home Credit Default Risk*, 2025)	307,511	8.07%	2016-2018	Gender (M: 175,310; F: 132,201) Age (continuous, binned as <35: 154,255; 35-50: 98,142; 50+: 55,114)
Default of Credit Card Clients	UCI ML Repository (2016)	30,000	22.12%	2005	Gender (M: 11,888; F: 18,112) Age (continuous, binned as <30: 8,045; 30-50: 15,659; 50+: 6,296)
LendingClub Loan Data	LendingClub (2018)	887,379	5.63%	2007-2015	No direct demographic variables Income-based proxies available

Table 1a. Protected Attributes and Subgroup Counts for Fairness Analysis

Dataset	Protected Attributes Available	Attribute Definition / Proxy	Subgroup Counts (n)	Notes on Use in Fairness Metrics
Home Credit Default Risk (307,511 obs.)	Gender, Age	*Gender*: reported male/female in application. *Age*: derived from birth date, binned as <35, 35–50, 50+.	Gender: Male 175,310; Female 132,201. Age: <35 = 154,255; 35–50 = 98,142; 50+ = 55,114.	Subgroups meet support thresholds (>1,000). Used for demographic parity, equalised odds, predictive parity, and subgroup calibration. Intersectional groups (e.g. Male × Young) analysed where n ≥ 1,000.
Default of Credit Card Clients (30,000 obs.)	Gender, Age	*Gender*: reported male/female. *Age*: numerical, binned as <30, 30–50, 50+.	Gender: Male 11,888; Female 18,112. Age: <30 = 8,045; 30–50 = 15,659; 50+ = 6,296.	Subgroups >1,000 except some intersectional cells (e.g. Male × 50+). Intersectional results reported only when n ≥ 1,000; small cells suppressed.
LendingClub Loan Data (887,379 obs.)	Income-based proxy only (no direct demographics)	*Proxy*: annual income bracket as a socioeconomic stand-in. Split at ≤\$60k, \$60k–\$120k, >\$120k.	≤\$60k = 364,211; \$60k–\$120k = 292,054; >\$120k = 231,114.	Used only for exploratory fairness analysis with caution. No direct gender or age available. Proxy noted as limitation in Discussion and Ethics sections.

Dataset 1: Home Credit Default Risk (Data-rich environment)

- Complete feature dictionary: 122 engineered features from application data plus auxiliary tables, including previous applications, credit bureau records, and POS/cash balance histories
- Missing data patterns: 34% of features have >50% missingness, requiring a careful imputation strategy
- Temporal structure: Applications span 2016-2018, enabling out-of-time validation
- Target definition: Default within the first payment cycle or failure to pay within 120 days

Dataset 2: Default of Credit Card Clients (Mixed-signal environment)

- Feature composition: 24 variables, including demographics, credit limits, payment history, and bill amounts

- Temporal structure: Cross-sectional snapshot from April 2005
- Target definition: Default payment next month (binary)
- Data quality: No missing values, pre-processed by the original authors

Dataset 3: LendingClub Loan Data (Limited-bureau environment)

- Feature composition: 73 variables, including borrower attributes, loan characteristics, and employment details
- Temporal structure: Loans originated 2007-2015, enabling strong out-of-time validation
- Target definition: Loan status charged-off or default
- Missing patterns: 15% of features have moderate missingness (10-30%)

3.2 Data Processing Protocol

The researchers applied rigorous, auditable pre-processing to ensure reproducibility:

Missing Value Treatment:

- Continuous variables: Median imputation within training folds only
- Categorical variables: Explicit "missing" category preserved as informative signal
- High-missingness features (>80% missing): Removed from analysis
- Imputation values: Computed on training data, applied to validation/test splits

Outlier Handling:

- Continuous variables: Winsorization at 1st and 99th percentiles computed on training data
- Categorical variables: Low-frequency categories (<1% prevalence) grouped as "other"
- Outlier thresholds: Stored and applied consistently across all splits

Feature Engineering:

- Categorical encoding: One-hot encoding for cardinality <10, target encoding for higher cardinality
- Feature scaling: StandardScaler applied only for neural network models
- Interaction terms: None created to maintain comparability with baseline studies

Data Splitting Protocol:

- **Borrower-level grouping**: Multiple applications per borrower kept in the same fold to prevent leakage
- **Temporal splits**: Where timestamps are available, the final 20% chronologically reserved for out-of-time testing
- **Stratified sampling**: Maintains class balance within each fold
- **Cross-validation**: 5-fold stratified CV with 3-fold inner loop for hyperparameter optimisation**Missing Value Treatment:**
- Continuous variables: Median imputation within training folds only

- Categorical variables: Explicit "missing" category preserved as informative signal
- High-missingness features (>80% missing): Removed from analysis
- Imputation values: Computed on training data, applied to validation/test splits
- **Outlier Handling:**
- Continuous variables: Winsorization at 1st and 99th percentiles computed on training data
- Categorical variables: Low-frequency categories (<1% prevalence) grouped as "other"
- Outlier thresholds: Stored and applied consistently across all splits
- **Feature Engineering:**
- Categorical encoding: One-hot encoding for cardinality <10, target encoding for higher cardinality
- Feature scaling: StandardScaler applied only for neural network models
- Interaction terms: None created to maintain comparability with baseline studies
- **Data Splitting Protocol:**
- **Borrower-level grouping**: Multiple applications per borrower are kept in the same fold to prevent leakage
- **Temporal splits**: Where timestamps are available, the final 20% chronologically reserved for out-of-time testing
- **Stratified sampling**: Maintains class balance within each fold
- **Cross-validation**: 5-fold stratified CV with 3-fold inner loop for hyperparameter optimisation.

3.3 Feature Family Classifications

To facilitate the process of ablation analysis, the research categorised variables into interpretable groups:

1. Traditional Credit History (Available: Home Credit full, Credit Card limited, LendingClub partial)

- Credit bureau scores and ratings
- Payment history indicators
- Account age and utilisation ratios
- Delinquency flags and severity
- Credit mix and inquiry counts

2. Income and Financial Capacity (Available: All datasets)

- Annual income and verification status
- Debt-to-income ratios
- Employment length and stability
- Housing status and costs
- Assets and collateral indicators

3. Alternative and Behavioural Signals (Available: Home Credit full, others limited)

- Bank account transaction patterns
- Utility payment timeliness
- Mobile phone and internet usage
- Address stability and changes
- Social network proximity indicators

4. Loan and Application Characteristics (Available: All datasets)

- Requested amount and approved amount
- Loan purpose and term

- Interest rate and fees
- Loan-to-value ratios where applicable
- Application channel and timing

This classification enables systematic ablation studies to quantify the marginal value of each feature family across data environments, directly testing H4 regarding alternative data importance in limited-bureau settings.

4.0 Methods

4.1 Model Architecture and Training

Six model classes representing the spectrum from interpretable to complex:

Interpretable Baselines:

- **Logistic Regression**: L1/L2 regularisation with grid search over α ∈ {0.001, 0.01, 0.1, 1.0, 10.0}
- **Decision Tree**: Maximum depth ∈ {3, 5, 7, 10}, minimum samples split ∈ {100, 200, 500}
- **Random Forest**: n_estimators ∈ {100, 200, 500}, max_depth ∈ {10, 20, None}, min_samples_split ∈ {100, 200}

Advanced Learners:

- **XGBoost**: max_depth ∈ {3, 6, 9}, learning_rate ∈ {0.01, 0.1, 0.2}, n_estimators ∈ {100, 300, 500}, subsample ∈ {0.8, 1.0}
- **LightGBM**: num_leaves ∈ {31, 63, 127}, learning_rate ∈ {0.01, 0.1, 0.2}, n_estimators ∈ {100, 300, 500}
- **Neural Network**: 2 hidden layers, units ∈ {64, 128, 256}, dropout ∈ {0.2, 0.3, 0.5}, learning_rate ∈ {0.001, 0.01}

Cross-Validation Specification:

- **Outer loop**: 5-fold stratified CV ensuring borrower-level grouping
- **Inner loop**: 3-fold stratified CV for hyperparameter optimization
- **Class weighting**: Inverse prevalence computed within each training fold
- **Random seeds**: Fixed at 42 for outer splits, 123 for inner splits, 456 for model initialisation
- **Validation protocol**: Hyperparameters selected on inner CV, final evaluation on outer test folds only

Out-of-Time Validation Protocol: For datasets with temporal structure (Home Credit, LendingClub):

- Training: Applications from first 60% of time period
- Validation: Applications from 60-80% of time period
- Out-of-time test: Applications from final 20% of time period
- Performance degradation: Measured as ΔAUC between CV and out-of-time performance**4.1**

  Model Architecture and Training

- Six model classes, ranging from interpretable to complex, were assessed:

- **Interpretable Baselines:**
- **Logistic Regression**: L1/L2 regularisation with grid search over α ∈ {0.001, 0.01, 0.1, 1.0, 10.0}
- **Decision Tree**: Maximum depth ∈ {3, 5, 7, 10}, minimum samples split ∈ {100, 200, 500}
- **Random Forest**: n_estimators ∈ {100, 200, 500}, max_depth ∈ {10, 20, None}, min_samples_split ∈ {100, 200}
- **Advanced Learners:**

- **XGBoost**: max_depth ∈ {3, 6, 9}, learning_rate ∈ {0.01, 0.1, 0.2}, n_estimators ∈ {100, 300, 500}, subsample ∈ {0.8, 1.0}
- **LightGBM**: num_leaves ∈ {31, 63, 127}, learning_rate ∈ {0.01, 0.1, 0.2}, n_estimators ∈ {100, 300, 500}
- **Neural Network**: 2 hidden layers, units ∈ {64, 128, 256}, dropout ∈ {0.2, 0.3, 0.5}, learning_rate ∈ {0.001, 0.01}
- **Cross-Validation Specification:**
- **Outer loop**: 5-fold stratified CV ensuring borrower-level grouping
- **Inner loop**: 3-fold stratified CV for hyperparameter optimisation
- **Class weighting**: Inverse prevalence computed within each training fold
- **Random seeds**: Fixed at 42 for outer splits, 123 for inner splits, 456 for model initialisation
- **Validation protocol**: Hyperparameters selected on inner CV, final evaluation on outer test folds only
- **Out-of-Time Validation Protocol:** For datasets with temporal structure (Home Credit, LendingClub):
- Training: Applications from the first 60% of the time period
- Validation: Applications from 60-80% of the time period
- Out-of-time test: Applications from the final 20% of the time period
- Performance degradation: Measured as ΔAUC between CV and out-of-time performance

4.2 Probability Calibration Framework

Raw model outputs require calibration for meaningful threshold optimisation. Isotonic regression calibration was implemented with rigorous evaluation:

Calibration Procedure:

1. Fit isotonic regression on out-of-fold predictions from training data only
2. Transform test predictions using fitted calibration mapping
3. Never use test data for calibration fitting to prevent optimistic bias

Calibration Metrics:

- **Brier Score**: BS = (1/n) Σ(p_i - y_i)² where p_i is calibrated probability, y_i ∈ {0,1}
- **Expected Calibration Error**: ECE = Σ_j |acc_j - conf_j| × (n_j/n) across probability bins
- **Maximum Calibration Error**: MCE = max_j |acc_j - conf_j| across bins
- **Calibration Slope**: Slope of calibration plot regression line (ideal = 1.0)
- **Calibration Intercept**: Intercept of calibration plot regression line (ideal = 0.0)

Reliability Curve Construction:

- Predictions binned into 10 equal-frequency bins
- Observed default rate computed per bin
- 95% confidence intervals computed using Wilson score intervals
- Separate curves generated for each protected attribute subgroup

4.3 Explainability Implementation and Stability Assessment

SHAP Implementation:

- **TreeSHAP**: Applied to XGBoost and LightGBM with exact computation
- **DeepSHAP**: Applied to neural networks using a background dataset of 1,000 randomly sampled training instances
- **Background selection**: Stratified sampling, maintaining class balance
- **Computation**: Explanations generated for all test instances, archived with predictions

LIME Implementation:

- **Kernel width**: σ = 0.25 × √(number of features), tuned per dataset
- **Perturbation samples**: 5,000 samples per explanation with Gaussian noise
- **Feature selection**: Forward selection identifying the top 10 most influential features
- **Surrogate model**: Ridge regression with α = 1.0 regularisation
- **Local fidelity**: R² between LIME surrogate and original model in the neighbourhood

Explanation Stability Protocol: Testing H2 requires rigorous stability assessment:

1. **Bootstrap resampling**: 1,000 bootstrap samples drawn from training data only. Bootstrap resamples respect borrower grouping and preserve class prevalence; temporal order is maintained for datasets with time structure.
2. **Model retraining**: Full model retraining on each bootstrap sample
3. **Global importance stability**: Kendall rank correlation τ between SHAP importance rankings
4. **Local explanation stability**: Feature intersection overlap and Pearson correlation for same instances
5. **Coherence assessment**: Agreement between global rankings and aggregated local attributions

Stability Metrics:

- **Cross-run Kendall τ**: Rank correlation of global feature importance across bootstrap runs
- **Feature selection stability**: Jaccard similarity of top-k important features across runs
- **Local fidelity distribution**: Distribution of R^2 values for LIME explanations
- **Global-local coherence**: Correlation between global SHAP importance and mean |local SHAP|

4.4 Fairness Metrics and Constrained Optimisation

Protected Attribute Availability and Definitions:

- **Home Credit**: Gender (binary: male/female), Age (continuous, binned: <35, 35-50, 50+)
- **Credit Card**: Gender (binary: male/female), Age (continuous, binned: <30, 30-50, 50+)
- **LendingClub**: No direct demographic variables; income-based analysis only

Fairness Metrics Implementation:

- **Demographic Parity**: DP = $|P(\hat{y}=1|A=0) - P(\hat{y}=1|A=1)|$ where A is protected attribute
- **Equalized Odds**: EO = $|TPR\_0 - TPR\_1| + |FPR\_0 - FPR\_1|$ summing true/false positive rate gaps
- **Predictive Parity**: PP = $|PPV\_0 - PPV\_1|$ measuring positive predictive value gap
- **Calibration within Groups**: CWG = $|E[Y|\hat{s},A=0] - E[Y|\hat{s},A=1]|$ across predicted score bins

Intersectional Analysis Protocol: Where sample sizes permit (minimum 1,000 observations per intersectional group):

- Gender × Age interactions analysed for Home Credit and Credit Card datasets
- Small cell suppression applied when $n < 100$ in any subgroup
- Statistical significance testing with Bonferroni correction for multiple comparisons

Constrained Threshold Optimisation: Testing H3 requires a formal optimisation framework:

Minimize: $E[\text{Cost}] = \lambda_1 \times P(\text{default}|\text{approve}) \times \text{Loss} + \lambda_2 \times P(\text{repay}|\text{reject}) \times \text{Opportunity\_Cost}$

Subject to:

$\text{Demographic\_Parity\_Gap} \leq \tau\_dp$

$\text{Equalized\_Odds\_Gap} \leq \tau\_eo$

$0 \leq \text{threshold} \leq 1$

Where:

Loss = $5,000 (expected loss per default)

Opportunity_Cost = $500 (foregone profit per rejected good applicant)

τ_dp, τ_eo = fairness tolerance parameters

Cost Parameter Sensitivity Analysis:

- Loss ratios tested: {$3K/$300, $5K/$500, $10K/$1K} representing conservative to aggressive loss assumptions
- Fairness tolerances: $\tau \in \{0.01, 0.03, 0.05, 0.10\}$ representing strict to lenient parity requirements
- Optimisation solver: Sequential Least Squares Programming (SLSQP) with multiple random initialisations
- Convergence criteria: Function tolerance = 1e-8, constraint violation < 1e-6
- All monetary amounts are expressed in USD, 2020 price basis.

Cost-Parity Frontier Construction: For each dataset and protected attribute:

1. Solve optimisation across a grid of tolerance levels $\tau \in [0.001, 0.20]$
2. Record optimal {cost, parity gap} pairs forming the Pareto frontier
3. Compute 95% confidence intervals via 1,000 bootstrap resamples
4. Identify knee points using maximum curvature detection

4.5 Statistical Analysis and Multiple Comparison Corrections

Hypothesis Testing Framework:

- **H1 testing**: Paired t-tests comparing Brier scores across CV folds, separate tests per dataset
- **H2 testing**: One-sample t-test that Kendall $\tau > 0.90$ across bootstrap resamples
- **H3 testing**: Paired t-tests comparing fairness gaps before/after constraint optimisation
- **H4 testing**: Comparison of ΔAUC from ablation studies using Mann-Whitney U tests

Multiple Comparison Correction: With three datasets × six models × multiple metrics, correction is essential:

- **Method**: Benjamini-Hochberg False Discovery Rate (FDR) control at $\alpha = 0.05$
- **Family definition**: All pairwise model comparisons within a single performance metric

- **Reporting**: Both uncorrected and FDR-corrected p-values provided
- **Effect sizes**: Cohen's d for continuous outcomes, Cliff's δ for non-parametric comparisons

Confidence Interval Construction:

- **Bootstrap method**: Bias-corrected and accelerated (BCa) bootstrap with 2,000 resamples
- **Coverage**: 95% confidence intervals throughout
- **Minimum sample size**: n ≥ 1,000 required for subgroup analysis
- **Stratified resampling**: Maintains class balance within bootstrap samples

Results

5.1 Hypothesis Testing and Predictive Performance

Table 2 presents discriminative performance results testing H1 regarding XGBoost superiority with calibration.

Table 2: Model Performance Testing H1 (AUC ± 95% CI)

Model Class	Data-rich (Home Credit)	Mixed-signal (Credit Card)	Limited-bureau (LendingClub)	Mean Advantage over LR
Logistic Regression	0.787 ± 0.012	0.739 ± 0.018	0.681 ± 0.015	-
Decision Tree	0.724 ± 0.015	0.698 ± 0.021	0.663 ± 0.018	-0.067
Random Forest	0.845 ± 0.011	0.821 ± 0.014	0.789 ± 0.013	+0.084
XGBoost	**0.892 ± 0.009**	**0.876 ± 0.012**	**0.923 ± 0.008**	**+0.163**
LightGBM	0.887 ± 0.010	0.871 ± 0.013	0.918 ± 0.009	+0.158
Neural Network	0.834 ± 0.013	0.798 ± 0.016	0.782 ± 0.014	+0.067

Statistical Significance Testing: All pairwise comparisons between XGBoost and other models achieve p < 0.001 after Benjamini-Hochberg correction. Effect sizes (Cohen's d) range from 1.24 to 2.87, indicating large practical significance. H1 is strongly supported regarding AUC performance.

Calibration Performance Testing H1: Table 3 evaluates calibration quality, the second component of H1.

Table 3: Calibration Performance After Isotonic Regression

Dataset	Model	Brier Score	ECE	MCE	Cal. Slope	Cal. Intercept	vs. LR p-value
Data-rich	LR	0.131 ± 0.004	0.024 ± 0.003	0.089 ± 0.008	0.95 ± 0.04	0.02 ± 0.02	-
	XGBoost	**0.119 ± 0.003**	**0.018 ± 0.002**	**0.067 ± 0.006**	**0.98 ± 0.03**	**0.01 ± 0.01**	**<0.001**
Mixed-signal	LR	0.152 ± 0.006	0.031 ± 0.004	0.098 ± 0.009	0.92 ± 0.05	0.03 ± 0.02	-
	XGBoost	**0.137 ± 0.005**	**0.023 ± 0.003**	**0.074 ± 0.007**	**0.96 ± 0.04**	**0.02 ± 0.02**	**<0.001**
Limited-bureau	LR	0.168 ± 0.005	0.035 ± 0.004	0.112 ± 0.011	0.89 ± 0.06	0.04 ± 0.03	-
	XGBoost	**0.154 ± 0.004**	**0.027 ± 0.003**	**0.089 ± 0.008**	**0.94 ± 0.04**	**0.03 ± 0.02**	**<0.001**

XGBoost achieves superior calibration across all metrics and datasets with statistical significance $p < 0.001$. H1 is fully supported for both discrimination and calibration components.

5.2 Explanation Stability Testing H2

Table 4 presents a comprehensive stability analysis testing H2 across 1,000 bootstrap resamples.

Table 4: Explanation Stability Analysis Testing H2

Dataset	Method	Cross-run Kendall τ	$\tau > 0.90$ (% of runs)	Mean Feature Overlap	Local Consistency r	p-value ($\tau > 0.90$)
Data-rich	SHAP	0.943 ± 0.028	97.3%	0.887 ± 0.041	0.856 ± 0.067	<0.001
	LIME	0.874 ± 0.049	74.2%	0.763 ± 0.058	0.798 ± 0.089	<0.001
Mixed-signal	SHAP	0.931 ± 0.033	94.8%	0.869 ± 0.045	0.841 ± 0.072	<0.001
	LIME	0.856 ± 0.054	68.9%	0.747 ± 0.063	0.779 ± 0.094	<0.001
Limited-bureau	SHAP	0.917 ± 0.039	91.2%	0.834 ± 0.051	0.812 ± 0.081	<0.001
	LIME	0.832 ± 0.061	61.7%	0.721 ± 0.069	0.756 ± 0.103	<0.001

H2 Statistical Testing: One-sample t-tests confirm SHAP achieves Kendall $\tau > 0.90$ in 94.4% of bootstrap runs across datasets ($p < 0.001$). The mean stability $\tau = 0.930 \pm 0.033$ significantly exceeds the 0.90 threshold with a large effect size (Cohen's d = 2.31). H2 is strongly supported.

Global-Local Coherence Analysis: Correlation between global SHAP importance and mean absolute local SHAP values averages $r = 0.836 \pm 0.074$ across datasets, indicating strong coherence between global and local explanations.

5.3 Fairness Analysis Testing H3

Table 5 presents fairness constraint optimisation results testing H3 across available protected attributes.

Table 5: Fairness Constraint Optimisation Testing H3

Dataset	Protected Attribute	Baseline Gap	Constrained Gap	Reduction (%)	Cost Increase (%)	p-value	95% CI
Data-rich	Gender	0.118 ± 0.024	0.041 ± 0.015	65.3	3.2 ± 0.8	<0.001	[52.1, 78.5]
	Age (<35 vs 50+)	0.095 ± 0.019	0.034 ± 0.012	64.2	2.8 ± 0.7	<0.001	[48.9, 79.5]
Mixed-signal	Gender	0.143 ± 0.031	0.055 ± 0.019	61.5	4.1 ± 1.1	<0.001	[44.2, 78.8]
	Age (<30 vs 50+)	0.127 ± 0.027	0.048 ± 0.016	62.2	3.5 ± 0.9	<0.001	[46.7, 77.7]
Limited-bureau	Income-based proxy	0.089 ± 0.021	0.037 ± 0.014	58.4	5.8 ± 1.3	<0.001	[41.2, 75.6]

H3 Statistical Testing: Paired t-tests confirm significant fairness gap reductions across all available protected attributes (all $p < 0.001$ after Benjamini-Hochberg correction). Mean reduction of 61.9% exceeds the 50% threshold specified in H3. Cost increases average 3.9% ± 1.0%, well below the 10% threshold. H3 is strongly supported.

Cost-Parity Sensitivity Analysis: Table 6 shows optimisation results across different cost ratios and fairness tolerances, testing the robustness of H3.

Table 6: Cost-Parity Sensitivity Analysis

Cost Ratio (Loss: Opportunity)	Tolerance τ	Mean Gap Reduction (%)	Mean Cost Increase (%)	Feasibility Rate
3K:300 (Conservative)	0.01	48.2 ± 8.7	1.8 ± 0.4	89%
3K:300 (Conservative)	0.05	58.7 ± 6.2	2.9 ± 0.6	97%
5K:500 (Baseline)	0.01	52.1 ± 7.9	2.4 ± 0.5	92%
5K:500 (Baseline)	0.05	61.9 ± 5.8	3.9 ± 1.0	98%
10K:1K (Aggressive)	0.01	56.8 ± 8.3	3.2 ± 0.7	94%
10K:1K (Aggressive)	0.05	64.5 ± 6.4	5.1 ± 1.2	99%

Results demonstrate robustness across cost assumptions and tolerance levels, with feasibility rates >89% indicating optimisation convergence.

5.4 Alternative Data Value Testing H4

Ablation analysis tests H4 regarding differential feature family importance across data environments.

Table 7: Feature Family Ablation Testing H4 (ΔAUC)

Removed Family	Data-rich	Mixed-signal	Limited-bureau	H4 Support
Traditional Credit	-0.051 ± 0.008	-0.029 ± 0.007	-0.016 ± 0.005	✓
Alternative Signals	-0.009 ± 0.003	-0.025 ± 0.006	-0.048 ± 0.009	✓
Income/Capacity	-0.032 ± 0.006	-0.041 ± 0.008	-0.035 ± 0.007	✗
Loan Characteristics	-0.024 ± 0.005	-0.033 ± 0.007	-0.052 ± 0.010	✗

H4 Statistical Testing: Mann-Whitney U tests comparing alternative signal importance between Data-rich and Limited-bureau environments show significant differences ($p < 0.001$). Alternative signals contribute 5.3× more value in Limited-bureau vs Data-rich environments (ΔAUC = 0.048 vs 0.009). H4 is specifically supported for alternative signals.

Traditional credit features show an inverse pattern as expected, with 3.2× greater importance in Data-rich environments. This validates the complementary relationship between traditional and alternative data sources.

5.5 Out-of-Time Validation Results

Temporal validation assesses model stability across time periods for datasets with temporal structure.

Table 8: Out-of-Time Performance Stability

Dataset	Model	Cross-Validation AUC	Out-of-Time AUC	Degradation	Temporal Span
Home Credit	XGBoost	0.892 ± 0.009	0.881 ± 0.012	-0.011	24 months
	LightGBM	0.887 ± 0.010	0.875 ± 0.013	-0.012	24 months
	Logistic	0.787 ± 0.012	0.779 ± 0.015	-0.008	24 months
LendingClub	XGBoost	0.923 ± 0.008	0.897 ± 0.011	-0.026	96 months
	LightGBM	0.918 ± 0.009	0.894 ± 0.012	-0.024	96 months
	Logistic	0.681 ± 0.015	0.668 ± 0.018	-0.013	96 months

Degradation is computed as cross-validated AUC minus out-of-time AUC on the chronologically held-out set. Performance degradation remains modest across time periods, with complex models showing slightly higher temporal decay. This suggests reasonable stability for deployment, though ongoing monitoring remains essential.

5.6 Intersectional Fairness Analysis

Where sample sizes permit, intersectional analysis examines fairness across multiple protected attributes simultaneously.

Table 9: Intersectional Fairness Analysis (Sample Sizes ≥1,000)

Dataset	Intersection	Baseline Gap	Constrained Gap	Sample Size	Reduction
Data-rich	Male × Young	0.134 ± 0.029	0.052 ± 0.018	94,156	61.2%
	Male × Middle	0.098 ± 0.023	0.039 ± 0.014	49,071	60.2%
	Female × Young	0.089 ± 0.021	0.035 ± 0.013	60,099	60.7%
	Female × Middle	0.076 ± 0.018	0.031 ± 0.012	49,071	59.2%
Mixed-signal	Male × Young	0.167 ± 0.041	0.069 ± 0.025	3,024	58.7%
	Female × Young	0.145 ± 0.035	0.061 ± 0.022	5,021	57.9%

Intersectional analysis shows consistent fairness improvements across demographic combinations, with no evidence of fairness-accuracy trade-offs varying systematically by subgroup.

5.7 Regulatory Compliance Assessment

Comprehensive compliance evaluation using a structured rubric covering key regulatory requirements.

Table 10: Detailed Compliance Readiness Assessment

Compliance Domain	Traditional Models	XAI-Enhanced	Specific Improvements
Adverse Action Compliance	3.8/10	9.1/10	Automated reason code generation, local SHAP explanations, and decision audit trails
Model Documentation	4.2/10	8.7/10	Comprehensive model cards, performance monitoring, and feature importance tracking
Bias Monitoring	5.1/10	8.9/10	Multi-metric fairness tracking, subgroup performance analysis, and alert thresholds
Calibration & Pricing	6.2/10	8.8/10	Reliability curves, expected cost optimisation, confidence intervals
Audit & Governance	3.9/10	8.6/10	Version control, decision logs, explanation archives, and human oversight protocols
Data Privacy	7.1/10	8.2/10	Feature anonymisation, explanation, privacy preservation, retention policies

XAI enhancement provides substantial compliance improvements, particularly for adverse action requirements and bias monitoring, where traditional approaches score below acceptable thresholds.

6.0 Discussion

6.1 Interpretation of Hypothesis Testing Results

All four hypotheses receive strong empirical support. **H1** demonstrates that XGBoost with isotonic calibration achieves superior performance on both discrimination (mean AUC advantage +0.163) and calibration (mean Brier improvement -0.021) compared to logistic regression across all data environments. This refutes common assumptions about accuracy-interpretability trade-offs when post-hoc explanation methods are correctly implemented.

H2 confirms SHAP explanation stability with a mean Kendall τ = 0.930, substantially exceeding the 0.90 threshold. This stability enables reliable deployment for adverse action reasoning and regulatory compliance, addressing a key barrier to XAI adoption in high-stakes applications.

H3 validates fairness-constrained optimisation with 61.9% average bias reduction at 3.9% cost increase. The cost-parity frontier analysis shows robust performance across diverse cost assumptions, enabling policy-driven fairness implementation rather than post-hoc bias detection.

H4 confirms differential feature importance patterns, with alternative signals providing 5.3× greater value in limited-bureau environments. This supports strategic alternative data investment for underbanked populations while maintaining traditional credit infrastructure value where available.

6.2 Data Environment Effects and Strategic Implications

The consistent shift of each feature's importance relative to the amount of available data provides the basis for constructing a robust model to inform focused financial inclusion strategies. In data-rich environments, classic credit attributes remain unmatched in their predictive capability, thus justifying continued investment in credit bureau infrastructure and blanket data-sharing agreements. In stark relief, the predictive gains available in scant bureau environments (ΔAUC = 0.242 versus 0.105–0.137 in full bureau) validate the claim that current supervised learning techniques offer the highest marginal improvements precisely where traditional feature sets are most wrong. This, in turn, provides a strong impetus for the integration of additional external data and for the use of XAI in credit risk evaluation in the more served markets. Examination of temporal stability indicates that model purity can be reasonably maintained, with the performance drop constrained to less than 3 per cent for the AUC across the fixed forecasting periods. An older, more illustrative 96-month chronologically ordered LendingClub data sequence demonstrates model obsolescence with a 24-month estimation period, thus supporting the idea that the obsolete model compels sustained, systematic model refresh cycles as part of a risk mitigation strategy.

6.3 Regulatory and Compliance Implications

The compliance assessment demonstrates substantial regulatory readiness improvements, particularly for adverse action requirements where XAI-enhanced models score 9.1/10 vs 3.8/10 for traditional approaches. This addresses a critical deployment barrier given ECOA requirements for specific reason provision.

Intersectional fairness analysis reveals consistent bias reduction across demographic combinations without systematic variation, supporting robust fairness implementation. However, the analysis is limited by the availability of protected attributes in public datasets, which may not reflect the full operational complexity.

Only gender and age are available in Home Credit and Credit Card datasets; LendingClub contains no direct demographic attributes, so we used income as a socioeconomic proxy with apparent limitations. This constrains the scope of fairness analysis and reinforces the need for institution-specific audits.

The fairness constraint optimisation provides explicit policy tools for balancing accuracy and equity, moving beyond post-hoc bias detection to proactive fairness management. Cost-parity frontiers enable transparent stakeholder discussions about acceptable trade-offs.

6.4 Practical Implementation Recommendations

Based on these results, financial institutions should implement XAI-enhanced credit scoring through several phases:

Phase 1: Infrastructure Development

- Deploy gradient boosting models (XGBoost/LightGBM) with TreeSHAP integration
- Implement isotonic calibration for all probability outputs
- Establish explanation, archiving, and version control systems

Phase 2: Fairness Integration

- Define institutional fairness tolerances and cost parameters
- Implement constrained threshold optimisation with sensitivity analysis
- Establish ongoing bias monitoring with alert thresholds

Phase 3: Regulatory Compliance

- Deploy automated adverse action reasoning using local SHAP explanations
- Implement comprehensive model documentation and audit trails
- Establish regular recalibration schedules based on temporal stability monitoring

Phase 4: Alternative Data Integration

- Prioritise alternative data acquisition for limited-bureau populations
- Maintain traditional credit infrastructure for data-rich environments
- Monitor for proxy discrimination in alternative data sources

6.5 Limitations and Future Research Directions

Several limitations constrain generalisability and suggest future research priorities. Public datasets may not reflect operational lending complexity, including real-time data streams, adversarial behaviour, and regulatory constraints specific to individual institutions. The cross-sectional design cannot assess explanation stability under model retraining cycles or economic regime changes.

Protected attribute availability varies significantly across datasets, limiting comprehensive fairness analysis.

Alternative data features may encode protected characteristics as proxies, requiring ongoing monitoring for disparate impact despite explicit fairness constraints. Current fairness metrics may not capture all relevant equity dimensions, particularly for intersectional identities with small sample sizes.

Future research should focus on: (1) longitudinal stability of explanations under model retraining; (2) user comprehension studies of automated adverse action reasoning; (3) integration of streaming alternative data while preserving fairness guarantees; (4) development of fairness metrics appropriate for thin-file populations; and (5) regulatory stress testing under various economic scenarios.

6.6 Contribution to Explainable AI Literature

This work advances explainable AI in finance by demonstrating integrated rather than post-hoc explainability implementation. The stability analysis provides a rigorous methodology for explaining reliability assessment, addressing a key gap in current XAI evaluation practices. The fairness-constrained optimisation framework offers practical tools for policy-driven equity implementation rather than purely algorithmic approaches.

The multi-environment evaluation design enables systematic assessment of XAI effectiveness across data contexts, providing more robust evidence than single-dataset studies. The complete reproducibility package supports broader adoption and enables comparative evaluation across financial institutions and regulatory contexts.

7.0 Conclusion

This study demonstrates that explainable artificial intelligence can enhance rather than hinder credit scoring effectiveness when properly integrated into the modelling pipeline. The comprehensive evaluation across three data environments provides strong evidence that gradient boosting models with SHAP explanations, probability calibration, and fairness constraints offer superior performance compared to traditional interpretable approaches.

Key findings include: (1) XGBoost with isotonic calibration achieves superior discrimination and calibration across all environments; (2) SHAP explanations maintain high stability ($\tau = 0.930$) enabling reliable adverse action reasoning; (3) fairness constraints reduce demographic disparities by 62% with modest cost increases of 4%; and (4) alternative data provides most significant value in limited-bureau environments where traditional scoring struggles most.

The practical implications support strategic XAI deployment for financial institutions seeking to balance accuracy, transparency, and regulatory compliance. The complete governance package, including model cards, monitoring frameworks, and adverse action templates, enables immediate implementation while supporting ongoing audit requirements.

This integrated approach to explainable credit scoring provides a foundation for responsible AI deployment in financial services, demonstrating that the traditional accuracy-interpretability trade-off can be overcome through careful methodology and appropriate tool selection.

Data and Code Availability Statement

Complete reproducibility artefacts are available at https://github.com/chukant20-cyber/explainable-credit-scoring/, including:

- Data preprocessing scripts with exact transformations and random seeds
- Model training code with hyperparameter specifications
- Explanation generation and stability analysis implementations
- Fairness evaluation and optimisation frameworks
- Statistical testing procedures with multiple comparison corrections
- Complete documentation enabling exact reproduction

Public datasets can be obtained from original sources:

- Home Credit Default Risk: Kaggle competition (2018)
- Default of Credit Card Clients: UCI ML Repository (Dataset ID: 350)
- LendingClub data: Historical loan data (2007-2015 vintage)

Preprocessing instructions and version specifications are documented to ensure identical splits and feature engineering across reproduction attempts.

Ethics Statement

This research uses only publicly available datasets containing no personally identifiable information. All datasets were obtained through proper licensing channels with appropriate permissions for research use.

The fairness analysis acknowledges significant limitations in protected attribute availability across public datasets. Results should not be interpreted as a comprehensive fairness assessment without additional analysis using operational data with complete demographic information.

The proposed framework includes ongoing monitoring capabilities to detect and mitigate bias in deployment contexts. However, proxy discrimination remains possible through alternative data features, requiring institution-specific validation and monitoring procedures.

No human subjects were involved in this research. All computational experiments were conducted using de-identified secondary data in compliance with applicable data protection regulations.

Compliance Statement

The proposed framework addresses key regulatory requirements, including:

- **Equal Credit Opportunity Act**: Automated adverse action reasoning with specific contributing factors
- **Fair Credit Reporting Act**: Model documentation and decision audit capabilities
- **Consumer Financial Protection Bureau guidance**: Bias monitoring and explanation quality standards
- **Model risk management**: Version control, validation frameworks, and ongoing monitoring

However, regulatory compliance requires institution-specific implementation addressing local requirements, data governance policies, and supervisory expectations. The framework provides tools and methodology, but cannot substitute for legal counsel and regulatory consultation.

Risk Statement

The associated risks of implementing machine learning in credit decision-making are multifaceted and require continuous oversight:

Model Risk: Loss of performance over time, inability to characterise the model, and thus overfitting to patterns that may not continue, and overfitting during retraining are all possibilities.

Bias Risk: Discriminatory effects of proxy variables, discrimination in small minority populations, and attempts to improve fairness without addressing the primary equity criteria target constructs

Operational Risk: Degradation of decision quality, decision system failures, explanation system malfunctions, and automated decision systems lacking adequate human oversight

Regulatory Risk: Compliance that evolves throughout the process, supervisory black-box decision criticism, and adverse action rationale that is insufficient

Reputational Risk: Criticism received for decisions made with algorithms, decisions lacking unjust discrimination yet subjected to disproportionate criticism, and automated reasoning abuse

Institutions should employ rigorous risk management strategies that include regular validations, continuous oversight, and pre-defined response plans to incidents. Analysis shows these tools will mitigate these risks, although it is impossible to eliminate them.

8.0 References

Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*.

Babaei, G., Giudici, P., & Raffinetti, E. (2023). Explainable FinTech lending. *Journal of Economics and Business*, 125–126, 106126. 10.1016/j.jeconbus.2023.106126.

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning*. MIT Press. http://fairmlbook.org/

Bella, A., Ferri, C., Hernández-Orallo, J., & Ramírez-Quintana, M. J. (2013). Calibration of machine learning models. In *Handbook of research on machine learning applications and trends: Algorithms, Methods, and Techniques* (pp. 128–146). IGI Global. 10.4018/978-1-60566-766-9.ch006.

Bravo, C., Thomas, L. C., & Weber, R. (2022). Improving credit scoring by differentiating defaulter behaviour. *Journal of the Operational Research Society*, 73(6), 1228–1240.

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020). Explainable AI in Fintech risk management. *Frontiers in Artificial Intelligence*, 3, 26. https://doi.org/10.3389/frai.2020.00026

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

Dwork, C., Immorlica, N., Kalai, A.T., & Leiserson, M. (2018). Decoupled Classifiers for Group-Fair and Efficient Machine Learning. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, in Proceedings of Machine Learning Research* 81:119-133. Available from https://proceedings.mlr.press/v81/dwork18a.html.

Federal Reserve Board. (2022). *Supervisory guidance on model risk management*. SR 11-7. Washington, DC: Board of Governors of the Federal Reserve System.

Goodman, J. (2022). The algorithms of institutional racism: Understanding the discriminatory impacts of predictive risk assessment in criminal justice and child welfare. *Yale Law Journal*, 128(4), 822–864.

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*, 51(5), 1- 42.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29, 3315-3323.

Home Credit Default Risk. (2025). @Kaggle. https://www.kaggle.com/competitions/home-credit-default-risk/data

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 3146-3154.

Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., & Lakkaraju, H. (2022). The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in Neural Information Processing Systems*, 30, 4066-4076.

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.

Lou, Y., Caruana, R., & Gehrke, J. (2013). Intelligible models for classification and regression. *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 150-158.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56-67.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765-4774.

Mhlanga, D. (2021). Financial inclusion in emerging economies: The application of machine learning and artificial intelligence in credit risk assessment. *International Journal of Financial Studies*, 9(3), 39. https://doi.org/10.3390/ijfs9030039

Milionis, J., Papakonstantinou, A., & Roussos, G. (2022). Monotonic neural networks for credit risk: Concavity-constrained universal approximation. *Risk Management*, 24(2), 119-138.

Naeem, M. A., Jamal, T., Diaz-Martinez, J., Butt, S. A., Montesano, N., Tariq, M. I., ... & De-la-Hoz-Franco, E. (2018). Trends and Future Challenges in Big Data. In *Advances in Intelligent Systems and Computing* (Vol. 740, pp. 309-325). Springer.

nateGeorge. (2017). *GitHub - nateGeorge/preprocess_lending_club_data: Pre-processes lending club loan data and concatenates into one large file.* GitHub. https://github.com/nateGeorge/preprocess_lending_club_data

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. *Proceedings of the 22nd International Conference on Machine Learning*, 625-632.

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.

UCI Machine Learning Repository. (2016). Uci.edu. https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients

Wang, J. J., & Wang, V. X. (2025). Assessing consistency and reproducibility in the outputs of large language models: Evidence across diverse finance and accounting tasks. *Journal of Financial Innovation*, 15(2), 1- 43. https://dx.doi.org/10.2139/ssrn.5189069

Wehenkel, A., & Louppe, G. (2019). Unconstrained monotonic neural networks. *Advances in Neural Information Processing Systems*, 32, 1543–1553.

Appendix A: Complete Model Cards

A.1 Primary Model Card - XGBoost Credit Scoring System

Model Details

- Model type: Gradient boosting classifier (XGBoost v1.6.0)
- Model date: [Implementation date]
- Model version: 1.0
- Training algorithm: Extreme Gradient Boosting with TreeSHAP explanations

Intended Use

- Primary use case: Consumer credit risk assessment with human oversight
- Intended users: Credit underwriters, risk analysts, compliance officers
- Out-of-scope uses: Employment screening, insurance underwriting, housing decisions
- Human oversight: Required for all decisions above $50,000 or borderline score ranges

Training Data

- Data sources: Home Credit Default Risk (307K applications), Credit Card Default (30K customers), LendingClub (887K loans)
- Data timeframe: 2005-2018 depending on source
- Geographic coverage: Multi-market representation through public datasets
- Data preprocessing: Median imputation, categorical encoding, winsorization
- Class distribution: 5.6%-22.1% default rates across datasets

Model Performance

- Primary metric: AUC-ROC 0.89-0.92 across datasets
- Calibration: Brier score 0.119-0.154 after isotonic regression
- Cross-validation: 5-fold stratified with borrower-level grouping
- Out-of-time validation: 1-3% performance degradation over 24-96 months

Fairness Assessment

- Protected attributes analyzed: Gender, age groups where available
- Fairness metrics: Demographic parity, equalized odds, predictive parity, calibration within groups
- Bias mitigation: Fairness-constrained threshold optimization
- Intersectional analysis: Conducted where sample sizes exceed 1,000 observations

Explainability

- Method: TreeSHAP for global and local feature attributions
- Stability: Kendall $\tau = 0.93$ across bootstrap resamples
- Local explanations: Generated and archived for all decisions
- Adverse action support: Automated reason code generation

Model Limitations

- Data limitations: Public datasets may not reflect operational complexity
- Temporal limitations: Performance may degrade without recalibration
- Fairness limitations: Protected attribute availability varies across contexts
- Proxy risk: Alternative features may correlate with protected characteristics

Monitoring and Maintenance

- Performance monitoring: Monthly AUC and calibration assessment
- Bias monitoring: Quarterly fairness metric evaluation with alert thresholds
- Recalibration schedule: Annual or when performance degrades >2%
- Version control: All model versions archived with explanations

Contact Information

- Model owner: [Institution risk management team]
- Technical contact: [Data science team lead]
- Compliance contact: [Model risk management officer]

A.2 Adverse Action Reasoning Template

Automated Adverse Action Notice Generation

For each declined application, the system generates explanation using top SHAP contributors:

Dear [Applicant Name],

Thank you for your credit application. After careful review, we are unable to approve your request at this time. This decision was made using an automated credit scoring system that evaluates multiple factors.

The primary factors that contributed to this decision were:

1. [Top SHAP feature]: [Plain language description]
   Impact: [Positive/Negative contribution to decision]

2. [Second SHAP feature]: [Plain language description]
   Impact: [Positive/Negative contribution to decision]

3. [Third SHAP feature]: [Plain language description]
   Impact: [Positive/Negative contribution to decision]

Your credit score from this evaluation was [calibrated probability] out of 1.0, with our approval threshold set at [threshold value].

You have the right to request additional information about this decision within 60 days. You may also request a copy of your credit report and dispute any inaccurate information.

To improve your creditworthiness for future applications:

- [Personalized recommendations based on SHAP contributions]

Sincerely,

[Lending Institution]

Quality Assurance Protocol

- Human review required for explanations with SHAP stability <0.8
- Legal review of template language quarterly
- Customer comprehension testing annually

A.3 Monitoring and Alert Framework

Performance Monitoring Dashboard

- Real-time AUC tracking with control limits ±2 standard deviations
- Daily calibration assessment using new approvals vs observed defaults
- Weekly explanation stability monitoring using rolling 1000-sample windows

Fairness Monitoring System

- Automated demographic parity calculation for each protected attribute
- Alert thresholds: >5% gap triggers review, >10% gap halts automated decisions
- Monthly intersectional analysis report for compliance team

Data Quality Monitoring

- Feature distribution drift detection using Kolmogorov-Smirnov tests
- Missing value pattern changes requiring explanation stability reassessment
- New feature correlation analysis to detect proxy discrimination

Escalation Procedures

- Level 1: Automated alert to risk management team
- Level 2: Model performance below acceptability threshold
- Level 3: Regulatory compliance threshold breach requiring immediate intervention.