



Evaluating Credit Score Explanations in Lending Decisions and Risk Management

Serkan Yilmaz Kandir^{1,*} Murat Ismet Haseki¹

¹ Faculty of Economics and Administrative Sciences, Cukurova University, Adana, Turkey

Emails: skandir@cu.edu.tr · mhaseki@cu.edu.tr

Received: August 11, 2026 Revised: September 10, 2026 Accepted: October 06, 2026 ★ Corresponding author

ABSTRACT

Credit score explanations can influence lending reviews, exception handling, and customer communication, yet an explanation may change under resampling, recommend an infeasible action, or fail to illuminate the approval decision. This structured review examines the evidence needed to use explanations in lending and risk management. Thirty-eight journal studies published from 2023 through January 2026 were coded by method family, information setting, validation design, and decision connection. The evidence spans interpretable scores and rules, post hoc attributions, counterfactuals, stability tests, fairness audits, and decision-oriented comparators. It does not support a single ranking of methods: studies examine different properties on different applicant populations, and few identify the incremental effect of an explanation on an underwriter's action or an applicant's outcome. Four comparisons organize the findings. Predictive discrimination does not establish explanation fidelity; plausible local reasons may not survive changes in training data; a score-changing counterfactual may be unattainable; and similar attributions do not imply equal lending outcomes. We develop a validation framework linking temporal performance, explanation perturbations, feasible action, group analysis, and policy value. Additional analyses map the evidence to decision stages, distinguish estimable quantities from missing outcomes, and specify reporting standards for management use. The review contributes an operational basis for deciding when an explanation can be treated as evidence in underwriting rather than as a model display.

Keywords: Business intelligence ▪ Credit scoring ▪ Explainable machine learning ▪ Underwriting ▪ Counterfactual recourse ▪ Algorithmic fairness

1. INTRODUCTION

Credit underwriting links a forecast of repayment to an allocation of capital. A lender chooses an approval threshold, loan terms, and an exposure limit before learning whether an accepted borrower will repay. A default score matters because it changes those choices. A reason supplied for the score can matter too: it may support a review, reveal a data error, guide an applicant's next step, or expose an unjustified pattern across groups. These purposes place different demands on the same explanation. A feature ranking that

describes a fitted model is not automatically a reason that a loan officer can act on, still less proof that changing a feature will cause a safer loan.

The value of a credit decision depends on more than the ordering of predicted defaults. Interest income, funding, servicing, expected loss, and the use of scarce review capacity all enter a lender's operating result. An explanation could help a reviewer identify a mistaken input before a high-value application is declined, or it could consume review time without changing a justified choice. These possibilities have different financial implications. They also suggest a business test: the

lender should specify which decision or control an explanation is meant to improve and measure that change against its cost.

Explanations also become part of the institution's decision record. An acquisition team, model owner, credit committee, and customer-facing employee may see different versions of a reason at different times. If the reference population, feature definition, or approval rule changes without an auditable record, the same wording may no longer describe the decision that was actually made. This is a problem of information governance and accountability as much as model design. It motivates a review focused on the path from statistical output to a documented lending action.

Recent studies illustrate the divide. An intrinsically interpretable optimal tree can make the branch followed by an applicant inspectable [1], whereas an additive boosting model displays feature-specific functions and selected interactions [2]. A SHAP value can be attached to an opaque prediction [3, 4]; a counterfactual can describe a model input that would cross the threshold [5, 6]. These objects differ even if a paper calls each one "explainable." Their tests also differ. Ballegeer et al. [7] directly examine how cost-sensitive training and class imbalance alter explanation stability, while Idbenjra et al. [8] compare discrimination, top-decile lift, and profit for a segmented score. An apparent disagreement about the best method may therefore be a disagreement about the question being answered.

This review asks a specific question: *What evidence establishes that an explanation of a credit score is reliable for an underwriting decision?* The unit of analysis is the study's validation claim. We do not rank algorithms by a pooled accuracy statistic, because population, target, prevalence, acceptance policy, and observation window vary. We instead distinguish explanatory properties that can be tested, identify which are measured in the recent literature, and propose a protocol for checking whether an explanation remains useful as the lending environment changes. This connects the journal's interests in business intelligence, data analytics, and operations research to a consequential business decision: a model must be judged at the point where the lender commits capital and the customer receives a decision.

2. METHODOLOGY

2.1 Boundary and selection

The review covers 38 distinct journal articles whose final journal dates fall between January 2023 and January 2026, inclusive. Searches combined *credit scoring*, *credit underwriting*, *credit rating*, and *lending* with *explainable*, *interpretable*, *SHAP*, *LIME*, *counterfactual*, *fairness*, *stability*, and *decision*. Publisher and institutional article records were used to verify titles, author order, final issue dates, and DOI identifiers. A study was retained if it evaluated a credit-risk or credit-approval model and made interpretation, explanation, an explanation-relevant fairness test, or a downstream underwriting comparison central to its contribution. Three predictive papers and two decision-objective papers serve as explicit comparators because a validation claim cannot be interpreted without a reference model or approval objective. This is a structured, purposive review, not a census of every

article returned by a database or a PRISMA systematic review. The accompanying CSV records each inclusion and coding choice.

The evidence windows differ in substantive ways. Consumer loans, credit cards, SME panels, mortgages, and e-commerce lending are represented. Published experimental designs include public benchmark data, proprietary or institutional histories, simulated changes, and model comparisons. We classify the principal contribution of each article into one of eight families: intrinsic interpretation (10), attribution (12), counterfactual recourse (3), fairness (4), stability (3), prediction comparator (3), decision objective (2), and governance (1). These counts describe the coding file and are not field-wide publication frequencies. A paper may discuss several methods, but is counted once under its main test. No quantitative effects from different datasets are pooled.

2.2 Coding the strength of a claim

For each paper, we asked four questions. What output is explained: a default probability, an approval decision, a credit rating, or a conditional survival prediction? What can change: training sample, background distribution, correlated features, time period, threshold, group, or feasible applicant action? What is observed: a ranking, rule, cost, fairness statistic, reviewer response, or realized financial outcome? Finally, which intervention is warranted by that observation? These questions prevent a display of explanatory variables from being described as proof of causal recourse. Tables 2–5 give the study-level map. A blank validation dimension in the synthesis below means that the reviewed design did not establish that dimension; it is not a judgment about the researchers' other contributions.

2.3 Analytical framework

Let X_t be the information available when an application arrives, $Y_{t+h} = 1$ a default over a specified horizon, and $f_t(X_t)$ a fitted default probability. The lender's action is $A_t = \pi(f_t(X_t), L_t, C_t)$, where L_t represents loan terms and capacity and C_t represents the current loss and revenue structure. An explanation $E_t = g(f_t, X_t, B_t)$ also depends on a background or reference specification B_t for many post hoc methods. The business outcome depends on the chosen action and subsequent repayment, rather than on E_t alone. This separation matters because an explanation can be coherent while the score is miscalibrated, or the score can be predictive while its explanation is unstable.

We distinguish five tests. *Fidelity* asks whether an explanation accurately characterizes the model in a defined neighborhood. *Stability* asks whether materially similar applications and reasonable re-estimations receive materially similar reasons. *Feasibility* asks whether a proposed counterfactual respects immutable attributes, feature dependencies, time, and affordability. *Equity* asks whether explanations and resulting actions behave acceptably across relevant groups after accounting for selection into observed repayment. *Decision value* asks whether showing the explanation changes a defensible underwriting decision or its auditability. Table 1 converts these into falsifiable comparisons.

One way to report stability is $S = 1 - \mathbb{E}[d(E_t(X), E'_t(X))]$, where E'_t is obtained after a prespecified, plausible pertur-

Table 1. Validation tests for a credit-scoring explanation

Test	Object held fixed	Perturbation or comparison	Observable failure	Business interpretation
Fidelity	Fitted score and decision time	Local surrogate versus original score	Reasons contradict local model behavior	Review reason is misleading
Stability	Applicant or matched cohort	Retraining, background, seed, period	Top reasons or signs reverse	Audit trail is fragile
Feasibility	Applicant and valid action set	Constrained candidate changes	Action violates timing or dependencies	Recourse advice cannot be used
Equity	Eligibility policy and horizon	Group and intersectional comparison	Explanation or error disparity persists	One reason may not serve all groups
Decision value	Score, threshold and exposure	Explanation shown versus withheld	No improvement in justified decisions	Interpretation has no measured operational return

bation and d is a normalized distance between explanation vectors or ranked reasons. The analyst must state the perturbation distribution and distance before reporting S ; otherwise the number has no portable meaning. A counterfactual x' should solve

$$\min_{x' \in \mathcal{A}(x,t)} c(x, x') \quad \text{s.t.} \quad \pi(f_t(x'), L_t, C_t) \neq \pi(f_t(x), L_t, C_t), \quad (1)$$

where $\mathcal{A}(x, t)$ contains feasible changes at the relevant time and c represents applicant effort or expense. The constraint is on the *decision*, not merely a score threshold imagined without loan terms. A useful follow-up also verifies that the proposed change can occur and that a later score or approval responds as predicted. Equation 1 is a review benchmark, not a claim that the surveyed studies implement identical optimization problems.

Finally, for an explanation interface $z \in \{0, 1\}$, define a policy-value comparison $\Delta V = \mathbb{E}[U(A_t(z = 1), Y_{t+h}) - U(A_t(z = 0), Y_{t+h})]$, with the same information available at application time. Measuring ΔV requires a design that separates the effect of the explanation from the score and identifies outcomes under changed approvals. Historic defaults are observed principally among accepted applications. This selective-label problem means that a retrospective profit calculation cannot by itself establish what would have happened to rejected borrowers.

3. RELATED WORK

3.1 Intrinsically interpretable models

Intrinsic methods make their explanation a property of the scoring function. The optimal hyperplane-split trees of Tu and Wu [1], additive functions of Zou et al. [2], and segmented logit structure of Idjenjra et al. [8] give an analyst inspectable components before a post hoc explainer is run. Teng et al. [9] use rescaling and clustering before local logistic regression; Chen et al. [10] reserve a shallow neural model for cases where a simpler linear explanation is insufficient. These choices change the object to be audited: a branch, function, cluster membership, or selective route. Their advantage is inspectability, but a local rule still depends on how boundaries and reference populations were learned.

Table 2 includes ten intrinsically interpretable studies. A fuzzy classifier [11] and an interpretable three-way decision for Chinese SMEs [12] illustrate that explanation can be expressed as graded membership or a deferment decision rather than a single feature list. The boundary-sample analysis

in Zhang et al. [13] directs attention to applicants near a decision frontier. The collaborative model of Wang et al. [14] raises a different question: when features are distributed across institutions, an importance report must reflect both restricted information access and the policy under which the score is used. These approaches should be compared on the decisions they induce as well as their accuracy. In Idjenjra et al. [8], reported AUC, top-decile lift, and profit supply a more business-facing comparison than AUC alone; none of these results establishes that a customer would understand a displayed reason.

3.2 Post hoc attribution

Feature attributions answer how a model's output is allocated over a chosen representation and baseline. They do not automatically identify a causal driver of repayment. The open-banking text study by Hjelkrem and de Lange [3], the ensemble comparison of Liu et al. [4], and the bank credit-worthiness analysis of Chang et al. [16] use such explanations for different data types. Hlongwane et al. [17] translate Shapley values into a scorecard presentation; a published correction to that paper clarifies the second author's name, a reminder that bibliographic records themselves require checking. The corrected citation is used here [18]. Darvish et al. [19] incorporate SHAP into a classification error-correction design, which makes the explanation part of model construction rather than a decorative final chart.

The 12 attribution studies in Table 3 also differ in their sampling and outcome processes. Oversampling in Han and Jung [20] and Han et al. [21] addresses rare defaults, but a distribution synthesized for training is not necessarily the distribution against which an applicant's reason should be stated. Shahee and Patel [22] combine adaptive synthetic sampling and focal loss with local and global SHAP. Chen et al. [23] extends local interpretation to survival modeling with time-varying covariates, where an explanation at one horizon cannot simply be carried to another. Wang et al. [24] addresses deep models in e-commerce lending. These are contributions to explanation coverage, not interchangeable evidence of explanation reliability. Whenever rankings are compared, the background sample, correlated predictors, fitted model, and reporting horizon need to be recorded.

3.3 Counterfactual recourse and fairness

A counterfactual explanation answers a conditional model question: what input change would alter an output? The sparse corporate-rating changes studied by Wang et al. [5] are

Table 2. Intrinsic interpretation studies ($n = 10$)

Study	Explanation or model	Evaluation reported	Evidence setting
Tu and Wu [1]	Model structure	classification and rule clarity	four public datasets
Zou et al. [2]	Model structure	shape functions and prediction	four credit datasets
Zhang et al. [15]	Model structure	feature interpretation	consumer credit
Idbenjra et al. [8]	Model structure	AUC, lift and profit	65,536 active customers
Chen et al. [10]	Model structure	selective model comparison	two credit datasets
Teng et al. [9]	Model structure	segment model comparison	two datasets
Baser et al. [11]	Model structure	fuzzy rule comparison	credit data
Pang et al. [12]	Model structure	three-way rule and error	Chinese SMEs
Wang et al. [14]	Model structure	privacy, cost and importance	multi-party credit data
Zhang et al. [13]	Model structure	boundary cases and interpretation	credit risk benchmarks

“Evaluation reported” is a topic coded from the published record, not a standardized effect estimate.

Table 3. Post hoc attribution studies ($n = 12$)

Study	Explanation or model	Evaluation reported	Evidence setting
Babaei et al. [25]	Feature attribution	model selection and explanation	over 100,000 SMEs
Darvish et al. [19]	Feature attribution	error correction	German and Australian benchmarks
Hlongwane et al. [17]	Feature attribution	scorecard construction	credit scorecard data
Han and Jung [20]	Feature attribution	imbalance and classification	public credit benchmarks
Han et al. [21]	Feature attribution	oversampling and prediction	credit benchmarks
Chang et al. [16]	Feature attribution	feature ranking	supply-chain finance
Talaat et al. [26]	Feature attribution	prediction and local explanation	credit-card default
Hjelkrem and de Lange [3]	Feature attribution	transaction text explanation	open banking
Shahee and Patel [22]	Feature attribution	imbalance and SHAP	credit assessment benchmarks
Chen et al. [23]	Feature attribution	longitudinal explanation	Freddie Mac mortgages
Wang et al. [24]	Feature attribution	deep model explanation	e-commerce lending
Liu et al. [4]	Feature attribution	ensemble and SHAP	credit datasets

The table records the main published evaluation; no causal interpretation of feature attribution is assumed.

conceptually different from an applicant’s feasible intervention: a firm’s balance-sheet movements across quarters may be observed without being immediately controllable. Dastile and Celik [27] explicitly address multiple counterfactual properties, including plausibility and actionability. Wang et al. [6] combine SHAP-guided search and a genetic algorithm to generate changes for credit benchmarks. Each narrows a different part of the gap between a mathematically valid point and usable advice. None permits the general assertion that every suggested change would cause a future approval in a live portfolio.

Fairness tests require similar precision. Agarwal et al. [28] use model-agnostic interpretations in a study of racial discrimination in lending. Chen et al. [29] compare Shapley–Lorenz contributions across population groups in a multi-country SME panel, whereas Babaei et al. [30] use Shapley values to examine how conditioning on requested loan amount can reveal a Simpson’s-paradox reversal in credit-lending fairness. Nwafor et al. [31] study a hybrid model and removal of potentially discriminatory inputs. Removing age or gender alone cannot establish the absence of proxy information in correlated variables. Table 4 deliberately places the counterfactual and group-audit studies together: both can be misread as a guarantee to an individual when they establish a narrower property of a fitted model or observed data.

3.4 Stability and decision comparators

The remaining nine papers in Table 5 explain why a stability or decision test cannot be bolted onto a conventional performance table after the fact. Ballegeer et al. [7] vary imbalance while comparing explanation rankings of instance-dependent cost-sensitive models; lower financial loss and steadier reasons need not move together. Sun et al. [32] combine several explainers and clustering in a credit-scoring framework. Prediction comparators [33–35] show that rare-event handling and feature representation change the fitted function whose behavior will later be explained. A reason for a model trained on one borrower mix may have a different meaning after an acquisition strategy changes that mix.

The underwriting objective is also mutable. Krivorotov [36] compare risk-based and profit-oriented credit-card scoring in a modeled portfolio; Kazemi et al. [37] optimize classification thresholds. Neither can be reduced to choosing the highest AUC. Jammalamadaka and Itapu [38] discusses responsible automated credit scoring and organizational accountability. Such work defines the question to be asked of an explanation even when it does not estimate the causal effect of presenting one to a loan officer. The study map is therefore organized by the *test of the explanation*, rather than a chronological sequence of algorithms.

Table 4. Counterfactual and fairness studies ($n = 7$)

Study	Explanation or model	Evaluation reported	Evidence setting
Babaei et al. [30]	Fairness audit	Shapley-based fairness and Simpson's paradox	credit lending
Chen et al. [29]	Fairness audit	Shapley-Lorenz group comparison	SME credit panel
Agarwal et al. [28]	Fairness audit	Shapley and lending audit	New York mortgage applications
Nwafor et al. [31]	Fairness audit	attribute removal and SHAP	peer-to-peer lending
Wang et al. [6]	Counterfactual	SHAP-informed counterfactuals	three public datasets
Wang et al. [5]	Counterfactual	sparsity and rating transitions	US firm statements
Dastile and Celik [27]	Counterfactual	plausibility and actionability	credit scoring benchmarks

Studies are grouped by their primary validation question; individual and group claims require different follow-up evidence.

Table 5. Stability, prediction, decision, and governance comparators ($n = 9$)

Study	Explanation or model	Evaluation reported	Evidence setting
Ballegeer et al. [7]	Stability check	resampling and cost	four public credit datasets
Chen et al. [39]	Stability check	imbalance perturbation	mortgage and two public datasets
Krivorotov [36]	Decision objective	profit versus risk score	credit-card underwriting
Jammalamadaka and Itapu [38]	Governance	responsible AI analysis	automated credit scoring
Kazemi et al. [37]	Decision objective	threshold optimization	credit scoring
Xia et al. [33]	Comparator model	imbalance benchmark	fintech lending
Xiao et al. [34]	Comparator model	imbalance benchmark	internet finance
Bhandary and Ghosh [35]	Comparator model	performance baseline	credit-card default
Sun et al. [32]	Stability check	multiple explanation methods	public credit datasets

Comparator studies are retained because explanation validity depends on the score and policy being explained.

4. RESULTS

4.1 Comparative evidence

4.1.1 Scope of observed claims

Table 6 compares claims that are often put on the same axis. An explanation is useful only relative to its intended recipient and action. The model developer may want a debugging signal; a credit analyst may need an audit trail for a marginal rejection; a borrower may need a feasible and truthful next step; a risk committee may ask whether explanations change across cohorts. It is possible for one artifact to meet one purpose and fail another. The literature is strongest on model-side demonstrations and benchmark comparisons. The reviewed records supply fewer direct observations of applicant follow-through, altered officer choices, or realized profitability due specifically to an explanation. This is an inference from the coded sample, not a field-wide frequency estimate.

The most consequential measurement choice is the target population. An accepted-loan dataset is selected by an earlier policy. A fairness comparison on this data can miss the rejected applicants whose outcomes were never observed. A similar limitation affects explanations of defaults if the model is used to expand approval to a new population. The empirical contexts in the corpus are heterogeneous by product, geography, time, and institution. For example, the SME panel used by Chen et al. [29] and the open-banking text in Hjelkrem and de Lange [3] are not alternative samples of the same estimand. A meta-analysis of their reported scores would be numerically tidy and substantively misleading.

4.1.2 Strength of validation evidence

Table 7 proposes a staged reading of a result. Passing an earlier stage is useful but does not certify a later stage. The sequence also explains why a more transparent model can be a stronger operational baseline than an elaborate explainer layered on a poorly calibrated score. The appropriate endpoint is not always higher approvals. A lender may rationally keep a manual-review region for cases with unstable predictions or weak reasons, as the selective architecture of Chen et al. [10] suggests. Conversely, a stable reason can be consistently wrong if its reference distribution excludes the intended applicant population.

4.2 Cross-study synthesis

4.2.1 Explanation boundaries

An explanation is anchored to three boundaries: a representation of the applicant, a reference population, and a policy threshold. Changing any one can change the managerial meaning of the reason without changing its prose. Consider debt-to-income ratio. If a score uses the ratio and income separately, an attribution to the ratio can share predictive information with income. A local feature ranking will depend on how those variables are treated jointly in the reference distribution. If the lender changes the approval threshold while holding the model fixed, the same ranking may accompany a different action. If a borrower reduces debt after a recommended change, the lender may also update income or loan terms. A counterfactual that moves only the ratio can be valid in a feature vector and impossible in the real data-generating process. These are design questions to document, not defects uniquely attributable to one algorithm.

Table 6. Contrasting the evidentiary reach of common designs

Design	Directly observed	Inference supported	Inference needing another design
AUC or default comparison	Separation on a defined test sample	Predictive ranking in that sample	Reason fidelity or change in approvals
Attribution display	Decomposition of fitted output	Model-relative feature contribution	Causal effect of applicant action
Retraining comparison	Variation in stated reasons	Sensitivity to specified perturbation	Stability under untested economic shocks
Counterfactual search	Feasible point under stated constraints	Model decision flips at that point	Applicant can enact it; repayment improves
Group fairness audit	Differences under a chosen metric	Disparity on observed labels and sample	Equal opportunity on rejected loans
Underwriter experiment	Change in human choices with interface	Decision effect in the tested workflow	Long-run portfolio value elsewhere

Table 7. Evidence hierarchy for an underwriting explanation

Stage	Minimum study design	Reported quantity	Remaining question
1	Temporally valid score comparison	Discrimination and calibration by cohort	Which reasons describe the score?
2	Fixed model and defined explanation reference	Local fidelity and feature dependence	Do reasons survive reasonable changes?
3	Prespecified retraining, background, and time shifts	Rank, sign, and magnitude stability	Can a person use the reason?
4	Constrained recourse and group audit	Feasibility, cost, coverage, subgroup error	Does the interface alter decisions?
5	Controlled workflow or credible policy evaluation	Overrides, adverse actions, utility, audit errors	Does value persist under rollout?

The reviewed work addresses these boundaries unevenly. Hlongwane et al. [17] focus on interpretable scorecard construction from Shapley values; Hjelkrem and de Lange [3] explain words in transactions and thus confront a different, high-dimensional representation. Chen et al. [23] model endogenous time-varying covariates, making the explanation horizon explicit. Wang et al. [6] and Dastile and Celik [27] formulate counterfactual changes, which move the question from “which features mattered” to “what could be different.” A study claiming the latter must check feature feasibility and the relevant decision rule. The concept of a local explanation should never conceal that locality is selected by a researcher: a neighborhood defined by random perturbations is different from a neighborhood of eligible applicants who can realistically change their circumstances.

The distinction between explanation stability and prediction stability is equally important. A model may maintain AUC while assigning a different mix of reasons to borderline cases, especially if highly correlated predictors substitute for one another. Conversely, a feature ranking can be remarkably stable while the default probabilities become miscalibrated in a downturn. Ballegeer et al. [7] are notable for testing explanation rankings under class imbalance and cost-sensitive training rather than inferring stability from a predictive metric. Sun et al. [32] build an explanation ensemble, but agreement among explanation tools on one dataset is a narrower test than persistence through a later cohort. Stability designs should specify whether they perturb observations, labels, training windows, reference distributions, or decision thresholds. A single number without its perturbation design cannot be transported to another bank.

Table 8 makes the boundaries operational. A review reader can use it as a checklist when interpreting any reported “transparent” credit model. The final two rows emphasize measurement outside the classifier: repayment labels are selectively observed, and the audience for a reason may change its value.

These are business analytics concerns, not merely technical properties of an explainer.

4.2.2 Comparisons across method families

First, an intrinsic model and a post hoc explanation of an opaque model should be compared at equal information and a common decision threshold. Otherwise a richer feature set or a different sample can create an apparent trade-off between accuracy and interpretability that is actually a data trade-off. The structural approaches in Tu and Wu [1], Zou et al. [2], Idbenja et al. [8], and Teng et al. [9] supply concrete alternatives to a single global logistic score. Their different architectures matter: a hyperplane split can be written as a rule but may combine many coefficients; an additive shape function can be visualized but may conceal an interaction unless it is included; segmentation can offer simple within-group coefficients but shift the interpretive burden to cluster assignment. The relevant question is which complete decision process an underwriter can inspect and contest. Reporting model size alone would not answer it.

Second, the attribution literature tests a model-relative notion of reason. SHAP or LIME can reveal how a score changes around a selected reference, but two applications with the same final risk can receive different attributions because one is compared with a different baseline or occupies a different local region. This is useful for detecting anomalies. It is not a stand-alone fairness criterion. Chen et al. [29] explicitly connect attributions to group comparisons; Agarwal et al. [28] use interpretation methods in a lending-discrimination question. Those designs are closer to the fairness objective than simply removing a protected column and showing the same AUC. The latter test in Nwafor et al. [31] is informative about the tested feature removal, but a lender would still need to examine correlated proxies, calibration, errors, and actual approvals. Fairness metrics can conflict, particularly under differing base rates; an audit must identify its target and the

Table 8. Boundaries of an explanation claim and the corresponding audit

Boundary	Change that tests it	Quantity to disclose	Misleading conclusion if omitted	Related evidence
Feature representation	Correlated features encoded jointly or separately	Dictionary, dependencies, missingness	“This variable caused default”	[3, 23]
Reference population	Training versus current applicants	Baseline and cohort dates	“The reason is applicant-specific”	[7, 39]
Class balance	Original versus oversampled data	Base rate and resampling stage	“The explanation is representative”	[20–22]
Policy threshold	Refit or revise cutoff and terms	Score, cutoff, reason, action	“The reason explains rejection”	[36, 37]
Feasible action	Restrict immutable and linked variables	Change cost and time horizon	“The applicant can comply”	[6, 27]
Observed labels	Extend or restrict accepted cohort	Acceptance rule and missing outcomes	“Disparity is absent”	[28–30]
User and workflow	Show explanation at review stage	Recipient, time, overrides	“Interpretation improved decisions”	[8, 38]

acceptable trade-off before a favorable statistic is presented. Third, credit-score counterfactuals and underwriting recourse are not equivalent. An optimized point may cross $f_i(x')$ while violating a causal dependency or a lender's current loan product rules. A formal solver can guarantee feasibility only for constraints it has encoded. The sparse changes in Wang et al. [5] offer an intelligible comparison of corporate ratings; the multiple-property design in Dastile and Celik [27] moves toward realistic candidate points; Wang et al. [6] ties feature importance to search. A definitive applicant-level recourse study would go further: it would present a feasible option, record whether the applicant acted, and evaluate the later decision and cost. The same intervention might be sensible for an applicant but unprofitable for a lender, or appear cheap to the lender while imposing significant borrower costs. Both perspectives need to be recorded.

Fourth, the economic objective changes what constitutes a valuable explanation. A credit score can be evaluated by discrimination, but the bank acts under margins, funding costs, expected loss, and capacity. Krivorotov [36] compare risk-oriented and profit-oriented portfolios; Kazemi et al. [37] optimize the classification threshold; and Idbenjra et al. [8] include a profit metric alongside AUC and lift. This is a stronger setting for studying explanation value than evaluating feature importance in isolation. The required counterfactual is what the bank and its officers would have done with the same score but without the reason. An explanation can improve an audit and correction process without raising the approval rate, and it can raise approval volume while worsening loss. A well-posed study should define the objective, constraints, and outcome horizon before presenting model interpretability as business value.

Table 9 states the head-to-head designs needed to decide among these claims.

4.2.3 Limits of cross-study ranking

A tempting review result would list the highest reported AUC for each approach and declare a winner. That ranking would mix different default horizons, sampling frames, class prevalences, and stages of the decision. The German and Australian benchmark datasets recur in interpretability work [6, 19]; they are useful for controlled algorithm comparisons but cannot represent all products or contemporary applicant behavior. Public data also allow a method to be tuned repeatedly against

known benchmarks. In contrast, the active-customer dataset of Idbenjra et al. [8], the SME panel in Chen et al. [29], and the Freddie Mac mortgage records analyzed by Chen et al. [23] carry different operational constraints and labels. Neither a public benchmark nor a proprietary field dataset automatically yields superior causal evidence: the former facilitates replication, while the latter may better represent a deployed workflow but limit external checking.

The dates at which information becomes available can reverse an apparent result. A repayment-related feature measured after the loan decision must not be used to explain a score supposedly computed at origination. A text description in an open-banking transaction stream can contain information unavailable to a bureau-only lender. A credit-card model trained on an earlier default regime may be impressive in random cross-validation and less useful in a later economic period. Reviewers should therefore distinguish random held-out tests from chronological and institution-level transport tests, and ask whether the explanation was recomputed after refitting. These questions apply to intrinsic and post hoc methods alike.

Finally, the reviewed studies differ in their primary outputs. A three-way decision may defer a borderline application [12]; a selective learning system may route it to a model with a different complexity [10]. A binary AUC comparison loses that deferment or routing option. In a constrained bank, the costly case may be a false approval, a needless denial, an expensive manual review, or an unauditable override. The choice of metric should follow that cost structure. A serious synthesis preserves these distinctions rather than manufacturing precision from incomparable reported numbers.

4.3 Evidence at the lending decision

The 38 records can be read along two axes: what the study makes visible and which business action its design can evaluate. The coding yields 12 attribution papers, 10 intrinsic-model papers, four fairness papers, three each on recourse, stability, and prediction comparators, two decision-objective papers, and one governance paper. These are mutually exclusive *primary* categories imposed for this review; they do not measure the prevalence of techniques in the full literature. Table 10 crosses each family with the lending stage for which its output could be relevant. The assignments are analytical interpretations of the reported designs, not claims that every paper in a row measured the stated downstream outcome.

Table 9. Comparison designs that would adjudicate contested claims

Claim under test	Models or methods	Conditions kept common	Evidence that would resolve it
Intrinsic is more usable	Rules, additive score, post hoc attribution	Information set, approval rule, user task	Correct decisions and explanation comprehension
Attribution is reproducible	SHAP, LIME, scorecard translation	Applicant, score, horizon	Perturbation-specific rank and sign stability
Recourse is actionable	Sparse and constrained counterfactuals	Applicant and loan products	Feasible action rate and later approval
Fairness explanation is adequate	Group attribution and outcome audit	Eligibility, time, protected-group definition	Joint disparities and sensitivity to missing labels
Complexity has business value	Baseline score versus enhanced model	Tuning budget, terms, exposure, costs	Out-of-time policy value and audit burden

An intrinsic rule may help a model committee inspect the basis of a score, but a rule alone does not reveal whether an officer would correct an error or a borrower would act differently. Attribution studies explain a fitted output relative to a representation and reference distribution; they become business evidence only when the explanation is tested against a defined review task. The three recourse papers move closer to applicant action, but a proposed feasible point and a realized change are separate endpoints. Fairness papers connect reasons to group comparisons, yet a group statistic on accepted loans does not identify outcomes for applicants the prior policy rejected. Decision-objective work sets the bank's payoff or threshold and thus tells us what an explanation would have to improve; it does not measure the incremental effect of displaying an explanation. This distinction is visible in the empirical contrast between interpretable profit reporting [8], explicit portfolio objectives [36], and explanation stability under cost-sensitive fitting [7].

The matrix also shows why “more explainability” is not a business outcome. A bank can improve the visual clarity of a dashboard while making no change to eligibility, error correction, review duration, or losses. Conversely, a concise reason that exposes a wrongly entered income could improve a particular decision even if it scores poorly on a generic explanation-similarity measure. The relevant denominator is the set of applications at which a reason can change an action. An evaluation should report both the population-wide result and the result among cases near the actual cutoff, while preserving the original eligibility and pricing rules. A method that changes only decisions far from the cutoff may be describing variation that the bank does not use.

4.4 Estimands and missing outcomes

At application time, the bank observes characteristics X and chooses $A \in \{0, 1\}$, where $A = 1$ denotes approval. Let $Y(1)$ denote repayment or loss if credit is extended; it is generally observed for approved accounts and missing for rejected applicants. The observed outcome is therefore $Y = AY(1)$ when a zero placeholder is used for missing rejected outcomes. That placeholder is a bookkeeping device, not a measured repayment result. A policy comparison over a common eligible population requires a value such as

$$V(\pi) = \mathbb{E}[\pi(X)\{R(X) - L(Y(1), X)\} - C_{\text{review}}(X, \pi)], \quad (2)$$

where R is revenue net of funding and servicing, L is loss conditional on an extended loan, and C_{review} is review cost.

The form is deliberately general: actual contracts, exposure, prepayment, and collections would determine the components in a particular lender. An explanation is valuable in this formulation if it changes the policy, error correction, or review process in a way that improves a prespecified objective or audit outcome. Model AUC alone supplies none of the revenue, cost, or counterfactual repayment terms in Equation 2.

Two comparisons need to be kept apart. A *score comparison* changes the scoring model and often the decisions it generates. An *explanation comparison* holds the score and permissible action set fixed and changes only what the decision maker sees. In the second case, the estimand is the difference between policies induced by explanation display and non-display, including review time and override behavior. A randomized display experiment could identify an immediate workflow effect for eligible reviewers, assuming assignment is implemented and measured as designed. It would not by itself identify long-run loss on additional rejected borrowers unless the resulting approvals create outcome observations or credible identification assumptions are supplied. Thus an experiment with a clean officer-level effect can still have an incomplete portfolio estimate.

The accepted-only record presents a second difficulty. Suppose a new explanation persuades officers to approve applicants whom the historical rule denied. Their potential losses are absent from the historical approved sample, so applying the old observed default rate to these new approvals is an extrapolation. If the old policy was deterministic for a stratum, inverse-propensity weighting cannot recover an outcome there because the approval probability was zero. Studies can instead report value bounds under explicit assumptions about missing losses, analyze policy overlap where it exists, or collect prospective outcomes under a controlled policy change. A narrow estimate conditional on previously approved borrowers is legitimate if it is labeled as such. It should not be presented as the value of a new approval policy for all applicants.

4.5 Sensitivity of an explanation claim

The claim that a reason is reliable should specify the changes under which it remains reliable. Table 11 separates shifts in the score, explainer, population, policy, and observation process. Each row holds a different object fixed, so its failure cannot be repaired by quoting success in another row. An attribution that remains stable across random seeds may change under a later reference cohort; a stable reason may accom-

Table 10. Primary study families mapped to a lending decision stage

Family	<i>n</i>	Immediate output	Potential decision stage	Further evidence needed for business use
Intrinsic	10	Rule, function, or segment	Model selection and case review	Common-data comparison of policy value and correct human use
Attribution	12	Relative local or global contribution	Error diagnosis and reason communication	Fidelity, reference sensitivity, and task-based comprehension
Recourse	3	Alternative input or rating	Applicant advice and reconsideration	Feasible uptake, later approval, and applicant cost
Fairness	4	Group contrast or audit statistic	Policy oversight and exception review	Selection-aware outcomes and prespecified group criteria
Stability	3	Change in reasons under perturbation	Monitoring and model refresh	Perturbations linked to actual cohort and policy shifts
Prediction comparator	3	Performance under an alternative model	Baseline model procurement	An explanation test on the same applicant population
Decision objective	2	Threshold or portfolio comparison	Pricing, approval, and exposure policy	Incremental policy value attributable to the explanation
Governance	1	Organizational account	Documentation and accountability	Traced actions and measured control effectiveness

Counts are from the 38-record coding file; each article appears in one primary family. Stage and evidence columns are this review's interpretation.

pany a score with deteriorating calibration; and a feasible counterfactual under one product may become unavailable when terms or affordability constraints change. The sampled studies motivate these tests: imbalance and cost sensitivity [7, 39], alternative feature representations [3, 23], threshold choices [37], and constrained recourse [6, 27] each move a different boundary.

The final row is especially important for group analysis. A comparison of error rates on accepted loans may be informative about the portfolio actually booked, but it says less about equal access to credit among all eligible applicants. Explanations can also affect who applies again, what documents are supplied, and which cases officers escalate. Those behavior changes alter the population on which future models are trained. Validation should therefore retain both the decision-time snapshot and subsequent events, with the interpretation conditioned on the observation process. In the studies reviewed here, no single paper establishes every link from a displayed reason through applicant behavior, policy change, and realized portfolio performance; this is a statement about the joint evidence needed, not a dismissal of the narrower results.

Taken together, these analyses imply a reporting unit larger than a single accuracy or explanation score. A lending study should report the applicant cohort and information cutoff, score and explanation versions, operative threshold and terms, intended audience, perturbation or comparison, selection into observed outcomes, and cost horizon. A failure on any one dimension can change the interpretation of the others. Publishing the joint record would allow business researchers to distinguish an explanation that helps inspect a model from one that changes a lending decision, and to replicate the latter claim under a later portfolio without relying on an unreported baseline.

5. DISCUSSION

5.1 Implications for lending operations

The immediate management decision is not whether to buy an explanation package. It is what failure an explanation is expected to prevent. A credit committee might want to

detect data leakage; an operations manager might want fewer costly manual reviews; an applicant-facing team might need clear adverse-action reasons. Those are separable outcomes. Before deployment, the bank should define the decision point, the audience, and the error that the explanation is intended to reduce. It can then compare an intrinsic model, an opaque model with a post hoc explainer, and the current policy under the same data cutoff and financial objective. A method should not receive credit for a gain produced by a different feature set or a different approval threshold.

Version control should cover more than the classifier weights. A reproducible adverse-action reason requires a frozen feature dictionary, preprocessing code, missing-value rules, explanation baseline, explainer version, and policy rule. The bank should retain the score and explanation as they were shown at the decision time, rather than reconstructing them months later from an updated model. A reason that changes during reconstruction can make a sound historical decision look arbitrary. This point is especially acute for local methods whose reference sample or perturbation generator can be revised independently of the score. An audit record should also distinguish an explanation displayed to an officer from one displayed to an applicant; those audiences may require different wording and permissible detail.

The decision boundary deserves disproportionate attention. Very safe and very risky applications may receive the same action under a broad range of models; small score changes near the threshold can reverse a loan decision. A validation sample should therefore report overall performance and a prespecified band around the operative cutoff. Within that band, the lender can compare the stability of stated reasons, calibration, incidence of manual review, and financial effect of overrides. The study of cost-sensitive explanation stability [7] warns that the very training objective used to reduce losses can change explanation behavior. Threshold optimization [37] confirms that the cutoff is itself a policy choice. An explanation study that reports only population-average rankings can miss the applicants whose actions actually change.

A credible deployment test would randomize the display of a reason while keeping the score and terms fixed for a bounded group of trained reviewers. Primary outcomes would be error

Table 11. Prespecified sensitivity checks for explanation validation

Source of change	Hold fixed	Re-estimate or vary	Report together with the reason
Model fitting	Applicant and feature definitions	Training window, seed, loss weighting	Score change, calibration, reason sign and top- <i>k</i> overlap
Explanation reference	Fitted score and applicant	Background cohort and correlations	Attribution magnitude, direction, and reason wording
Applicant mix	Product and eligibility rule	Later cohort, channel, or institution	Coverage, subgroup error, reason drift at cutoff
Decision policy	Fitted score and test cohort	Threshold, terms, manual-review band	Approval changes, expected margin, override volume
Recourse constraints	Original applicant and product	Time, affordability, linked features	Feasible share, cost, and verified later action
Outcome observation	Eligibility and outcome horizon	Rejected-loan loss assumptions	Value or fairness bounds, overlap, unidentified strata

The rows specify proposed audits, not estimates reported for the entire corpus.

detection and policy-compliant choices; secondary outcomes would include processing time, overrides, applicant complaints, and subsequent portfolio losses. The randomization should never conceal legally required information from applicants. Where direct randomization is unavailable, phased rollout and well-specified comparison cohorts are possible, but concurrent policy changes must be documented. The objective is to identify the marginal contribution of the explanation interface, not to attribute all changes in lending performance to the model. This design makes business intelligence accountable to a measurable workflow.

5.2 Research challenges

5.2.1 Changing the reference population

Post hoc explanations can change when the background sample changes, even with the same applicant and fitted model. A publication that reports one feature ranking but omits the reference population makes the explanation hard to reproduce. Class imbalance increases the stakes: a lender may oversample defaults to train a classifier [20–22], then explain it against a synthetic training distribution rather than an application population. Research should vary the background design and compare signs, top-*k* membership, and decisions around the threshold, rather than reporting only an average similarity score. The controlled resampling in Ballegeer et al. [7] and imbalance analysis in Chen et al. [39] show why a perturbation must be tied to a plausible operational shift.

5.2.2 Observational labels and policy feedback

A loan approval rule changes who supplies a repayment label. If the explanation persuades an officer to override the rule, it also changes future training data. Ex post comparisons that treat the resulting labels as if drawn at random can credit the explanation for a selection effect. A credible design should log the score, threshold, explanation version, original recommendation, override, terms, and later outcome, with the reason for missing labels. Where experimentation is infeasible, a policy evaluation needs overlap assumptions and a sensitivity analysis for rejected cases. The profit and threshold papers [36, 37] clarify the objective, but do not remove this identification problem.

5.2.3 Actionable advice and time

For a rejected applicant, “reduce debt” may be understandable yet unattainable in the timeframe of a new application. Some

features are immutable, others are downstream consequences of income or employment, and still others can be changed only jointly. Counterfactual search should encode those dependencies in $\mathcal{A}(x, t)$, offer more than one feasible option when possible, and check what happened after advice was given. The multiple-property counterfactual study [27] and SHAP-guided recourse design [6] are relevant starting points. Corporate credit-rating transitions [5] illustrate why an observed transition need not map to an immediately executable action.

5.2.4 Recipients and workflow

An attribution vector used for developer diagnosis may overwhelm a credit officer and offer an applicant no lawful or useful reason. A defensible interface should specify who receives what, at which step, and how errors can be contested. It should test whether the recipient understands the explanation, detects planted model or data errors, and makes a better decision under a fixed time budget. The organizational treatment of responsible scoring [38] gives a governance setting; the corpus does not establish a general effect of explanations on actual officer behavior. Calling an algorithm “human in the loop” without recording overrides and outcomes would not resolve that gap.

6. FUTURE RESEARCH DIRECTIONS

A focused agenda follows from these failure modes. Table 12 gives six research designs with an estimand or measurable outcome and a falsification condition. These are recommendations, not results attributed to the 38 studies. They prioritize explanatory validity at the margins where business consequences are largest: threshold cases, new applicant cohorts, and borrowers receiving concrete advice.

Preregister the outcome horizon, applicant eligibility, decision costs, calibration measure, and explanation perturbations.

The evaluation should report negative results when a reason is unstable and keep the policy threshold fixed during explanation-only experiments. If a richer model and explainer improve AUC but worsen audit consistency, show the trade-off as a frontier rather than compressing it into one trust score. If a simple scorecard meets the bank’s operating constraints, it remains the relevant benchmark. If complex modeling helps a thin-file cohort, validate its explanation for that cohort rather than borrowing from average customers.

Table 12. Testable research directions for credit-scoring explanation validity

Question	Proposed design	Primary outcome and failure test	Decision level
Does a reason survive retraining?	Refit on rolling historical windows; preserve applicant cases and background definition	Rank/sign agreement at approval margin; fail on prespecified reversal rate	Model
Can an applicant use recourse?	Randomize feasible advice among eligible rejects with consent and follow-up	Verified actions and later decisions; fail if action or approval is not realized	Applicant
Are reasons equitable with selective labels?	Audit accepted and rejected cohorts with explicit label sensitivity bounds	Subgroup explanation and error differences; fail if bounds cross policy tolerance	Portfolio
Does explanation improve review?	Randomize explanation display to trained officers with fixed scores and time	Correct error detection, overrides, and decision cost; fail if no net improvement	Workflow
Will results transport?	Freeze model and explainer, test a later cohort or institution	Calibration, reason drift, and constrained utility; fail under prespecified degradation	Deployment
What is the simplest adequate model?	Compare intrinsic and post hoc models under equal information and tuning budgets	Joint frontier of value, fidelity, stability, and operating cost	Procurement

7. LIMITATIONS

The 38 studies were selected to provide a focused comparison of explanation validation rather than an exhaustive count of all credit-scoring research. Search engines and bibliographic databases differ in coverage, online-first dates may differ from issue dates, and proprietary lending studies may remain unpublished. We resolved the date boundary using final journal issues where available and retained one article appearing in the January 2026 issue. We checked the named correction to a 2024 PLOS ONE article and used its corrected author spelling. The coding assigns each study to one principal family even when it discusses multiple techniques. Counts therefore depend on that rule and should not be read as an estimate of a literature-wide trend.

Most reported evaluations describe a model on data available to the authors. We did not obtain the borrower-level records or rerun those algorithms. Accordingly, statements about a study's design and evidence boundary are based on its article record and reported methods, not a replication claim. We avoid averaging performance figures from incomparable settings. The review also does not adjudicate legal compliance across jurisdictions; fairness metrics and explanation duties depend on specific products and rules. A stronger future review could preregister a database search, dual-screen full texts, publish a risk-of-bias assessment, and restrict comparisons to studies with compatible outcome horizons and applicant populations. Such a synthesis would complement, rather than replace, the decision-centered validation proposed here.

8. CONCLUSION

This review examines whether a credit-score explanation supports a lending decision that a bank can justify, execute, and monitor. Across 38 recent journal studies, the methods range from interpretable rules and additive scores to attributions, counterfactuals, stability tests, fairness audits, and decision-oriented comparators. Their findings cannot be collapsed into one preferred technique because the tests measure different objects and populations. The decision-stage matrix identifies where each family can inform underwriting and which further observation is required before a model property becomes a business claim.

For business analytics and risk management, the next step is to evaluate explanations as components of an underwriting

policy. A reliable reason must describe the deployed score, withstand plausible data changes, respect feasible applicant actions, and be examined for group effects. Its contribution to review quality, exception management, or portfolio value needs a separate decision-level test. The proposed estimands, sensitivity checks, and prospective designs make those requirements measurable while accounting for outcomes missing among rejected applicants. This purposive review does not estimate industry-wide effects or certify regulatory compliance. It provides a common basis for comparing claims and for deciding whether an explanation merits operational use.

REFERENCES

- [1] J. Tu and Z. Wu, "Inherently interpretable machine learning for credit scoring: Optimal classification tree with hyperplane splits," *European Journal of Operational Research*, vol. 322, no. 2, pp. 647–664, 2025.
- [2] Y. Zou, M. Xia, and X. Lan, "Interpretable credit scoring based on an additive extreme gradient boosting," *Chaos, Solitons & Fractals*, vol. 194, Art. no. 116216, 2025.
- [3] L. O. Hjelkrem and P. E. de Lange, "Explaining deep learning models for credit scoring with SHAP: A case study using open banking data," *Journal of Risk and Financial Management*, vol. 16, no. 4, Art. no. 221, 2023.
- [4] Y. Liu, F. Huang, L. Ma, Q. Zeng, and J. Shi, "Credit scoring prediction leveraging interpretable ensemble learning," *Journal of Forecasting*, vol. 43, no. 2, pp. 286–308, 2024.
- [5] D. Wang, Z. Chen, I. Florescu, and B. Wen, "A sparsity algorithm for finding optimal counterfactual explanations: Application to corporate credit rating," *Research in International Business and Finance*, vol. 64, Art. no. 101869, 2023.
- [6] L. Wang, Z. Yu, J. Ma, X. Chen, and C. Wu, "A two-stage interpretable model to explain classifier in credit risk prediction," *Journal of Forecasting*, vol. 44, no. 7, pp. 2132–2150, 2025.
- [7] M. Ballegeer, M. Bogaert, and D. F. Benoit, "Evaluating the stability of model explanations in instance-dependent cost-sensitive credit scoring," *European Journal of Operational Research*, vol. 326, no. 3, pp. 630–640, 2025.
- [8] K. Idbenjra, K. Coussement, and A. de Caigny, "Investigating the beneficial impact of segmentation-based modelling for credit scoring," *Decision Support Systems*, vol. 179, Art. no. 114170, 2024.

- [9] H.-W. Teng, M.-H. Kang, I.-H. Lee, and L.-C. Bai, "Bridging accuracy and interpretability: A rescaled cluster-then-predict approach for enhanced credit scoring," *International Review of Financial Analysis*, vol. 91, Art. no. 103005, 2024.
- [10] D. Chen, J. Ye, and W. Ye, "Interpretable selective learning in credit risk," *Research in International Business and Finance*, vol. 65, Art. no. 101940, 2023.
- [11] F. Baser, O. Koc, and A. S. Selcuk-Kestel, "Credit risk evaluation using clustering based fuzzy classification method," *Expert Systems with Applications*, vol. 223, Art. no. 119882, 2023.
- [12] M. Pang, F. Wang, and Z. Li, "Credit risk prediction based on an interpretable three-way decision method: Evidence from Chinese SMEs," *Applied Soft Computing*, vol. 157, Art. no. 111538, 2024.
- [13] R. Zhang, I. Li, and Z. Ding, "An interpretable credit risk assessment model with boundary sample identification," *PeerJ Computer Science*, vol. 11, Art. no. e2988, 2025.
- [14] Z. Wang, Y. Tian, S. Li, and J. Xiao, "A secure cross-silo collaborative method for imbalanced credit scoring," *European Journal of Operational Research*, vol. 326, no. 2, pp. 357–373, 2025.
- [15] R. Zhang, I. Li, Z. Ding, and T. Zhu, "IWSL model: A novel credit scoring model with interpretable features for consumer credit scenarios," *Journal of Forecasting*, vol. 44, no. 7, pp. 2230–2251, 2025.
- [16] V. Chang, Q. A. Xu, S. H. Akinloye, V. Benson, and K. Hall, "Prediction of bank credit worthiness through credit risk analysis: An explainable machine learning study," *Annals of Operations Research*, vol. 354, no. 1, pp. 247–271, 2025.
- [17] R. Hlongwane, K. Ramaboa, and W. Mongwe, "A novel framework for enhancing transparency in credit scoring: Leveraging Shapley values for interpretable credit scorecards," *PLOS ONE*, vol. 19, no. 8, Art. no. e0308718, 2024.
- [18] R. Hlongwane, K. Ramaboa, and W. Mongwe, "Correction: A novel framework for enhancing transparency in credit scoring: Leveraging Shapley values for interpretable credit scorecards," *PLOS ONE*, vol. 20, no. 8, Art. no. e0329901, 2025.
- [19] E. Darvish, M. Jahangoshai Rezaee, and M. Abbaspour Onari, "A hybrid error correction approach for prediction of credit approval: An explainable artificial intelligence approach," *Engineering Applications of Artificial Intelligence*, vol. 144, Art. no. 110140, 2025.
- [20] S. Han and H. Jung, "NATE: Non-parametric approach for Explainable credit scoring on imbalanced class," *PLOS ONE*, vol. 19, no. 12, Art. no. e0316454, 2024.
- [21] S. Han, H. Jung, P. Yoo, A. Provetti, and A. Cali, "NOTE: Non-parametric oversampling technique for explainable credit scoring," *Scientific Reports*, vol. 14, Art. no. 26070, 2024.
- [22] S. A. Shahee and R. Patel, "An explainable ADASYN-based focal loss approach for credit assessment," *Journal of Forecasting*, vol. 44, no. 4, pp. 1513–1530, 2025.
- [23] Y. Chen, R. Calabrese, and B. Martin-Barragan, "JointLIME: An interpretation method for machine learning survival models with endogenous time-varying covariates in credit scoring," *Risk Analysis*, vol. 45, no. 12, pp. 4135–4156, 2025.
- [24] C. Wang, Y. Li, S. Wang, and Q. Wu, "Demystifying deep credit models in e-commerce lending: An explainable approach to consumer creditworthiness," *Knowledge-Based Systems*, vol. 312, Art. no. 113141, 2025.
- [25] G. Babaei, P. Giudici, and E. Raffinetti, "Explainable FinTech lending," *Journal of Economics and Business*, vol. 125–126, Art. no. 106126, 2023.
- [26] F. M. Talaat, A. Aljadani, M. Badawy, and M. Elhosseini, "Toward interpretable credit scoring: Integrating explainable artificial intelligence with deep learning for credit card default prediction," *Neural Computing and Applications*, vol. 36, no. 9, pp. 4847–4865, 2024.
- [27] X. Dastile and T. Celik, "Counterfactual explanations with multiple properties in credit scoring," *IEEE Access*, vol. 12, pp. 110713–110728, 2024.
- [28] S. Agarwal, C. B. Muckley, and P. Neelakantan, "Countering racial discrimination in algorithmic lending: A case for model-agnostic interpretation methods," *Economics Letters*, vol. 226, Art. no. 111117, 2023.
- [29] Y. Chen, P. Giudici, K. Liu, and E. Raffinetti, "Measuring fairness in credit ratings," *Expert Systems with Applications*, vol. 258, Art. no. 125184, 2024.
- [30] G. Babaei, P. Giudici, and P. Neelakantan, "Explainability, fairness and the Simpson's paradox in credit lending," *Physica A: Statistical Mechanics and its Applications*, vol. 680, Art. no. 131030, 2025.
- [31] C. N. Nwafor, O. Nwafor, and S. Brahma, "Enhancing transparency and fairness in automated credit decisions: An explainable novel hybrid machine learning approach," *Scientific Reports*, vol. 14, Art. no. 25174, 2024.
- [32] X. Sun, J. Liu, and Y. Zhang, "Enhancing credit risk prediction through an ensemble of explainable model," *Journal of Systems Science and Systems Engineering*, vol. 34, no. 5, pp. 619–640, 2025.
- [33] Y. Xia, Z. Han, Y. Li, and L. He, "Credit scoring model for fintech lending: An integration of large language models and FocalPoly loss," *International Journal of Forecasting*, vol. 41, no. 3, pp. 894–919, 2025.
- [34] J. Xiao, Y. Zhong, Y. Jia, Y. Wang, R. Li, X. Jiang, and S. Wang, "A novel deep ensemble model for imbalanced credit scoring in internet finance," *International Journal of Forecasting*, vol. 40, no. 1, pp. 348–372, 2024.
- [35] R. Bhandary and B. K. Ghosh, "Credit card default prediction: An empirical analysis on predictive performance using statistical and machine learning methods," *Journal of Risk and Financial Management*, vol. 18, no. 1, Art. no. 23, 2025.
- [36] G. Krivorotov, "Machine learning-based profit modeling for credit card underwriting - implications for credit risk," *Journal of Banking & Finance*, vol. 149, Art. no. 106785, 2023.
- [37] H. R. Kazemi, K. Khalili-Damghani, and S. Sadi-Nezhad, "Estimation of optimum thresholds for binary classification using genetic algorithm: An application to solve a credit scoring problem," *Expert Systems*, vol. 40, no. 3, Art. no. e13203, 2023.
- [38] K. R. Jammalamadaka and S. Itapu, "Responsible AI in automated credit scoring systems," *AI and Ethics*, vol. 3, no. 2, pp. 485–495, 2023.
- [39] Y. Chen, R. Calabrese, and B. Martin-Barragan, "Interpretable machine learning for imbalanced credit scoring datasets," *European Journal of Operational Research*, vol. 312, no. 1, pp. 357–372, 2024.