Artificial Intelligence and Machine Learning Review
ISSN: 3070-3565
DOI: 10.69987/AIMLR.2026.70107

Explainable Credit Underwriting on FICO HELOC: A Framework for Counterfactual Recourse under Profit, Fairness, and Stability Constraints

Annie Zhao
School of Computer Science, The University of Sydney, NSW, Australia
annie.zh1368@gmail.com

Keywords

credit underwriting; explainable AI; SHAP; counterfactual explanations; recourse; threshold policy; profit; fairness; stability; FICO HELOC

Abstract

Credit underwriting models are increasingly accurate yet often opaque, creating tension between portfolio performance, regulatory expectations, and actionable customer communication. This paper presents a reproducible experimental study on the FICO HELOC dataset (10,459 applicants; 23 credit-bureau features) that integrates (i) predictive modeling, (ii) executable bad-rate-constrained threshold policies, (iii) SHAP-based explanations, and (iv) counterfactual recourse with feasibility constraints. We train logistic regression, random forests, and gradient boosting using a consistent preprocessing pipeline that treats the dataset's special missing codes (-7, -8, -9) as missing values and applies median imputation. We then derive underwriting policies by selecting the minimum approval threshold that satisfies a target bad-debt rate among approved accounts, and evaluate "approval rate and profit at equal bad rate." To assess responsible deployment aspects, we compute group fairness gaps using demographic parity (DP) and equal opportunity (EO) across operational score segments, and quantify explanation stability using bootstrap resampling. Empirically, at a 20% bad-rate constraint, random forests achieve the highest approval rate (30.9%) and the highest mean profit under a simple unit-profit model (+1 for Good, -3 for Bad). Gradient boosting yields coherent global explanations dominated by ExternalRiskEstimate, credit file age, and revolving utilization, while counterfactual recourse achieves 100% success for logistic regression and 89.5% for gradient boosting under actionable feature constraints. Bootstrap analysis shows strong SHAP stability (mean Spearman ρ=0.871 for global importance; mean cosine similarity=0.856 for local vectors). The results demonstrate how policy design, explainability, and stability can be evaluated jointly on a realistic credit dataset, yielding decision rules that are directly implementable as underwriting policy.

Introduction

Credit underwriting is a high-stakes socio-technical workflow that transforms noisy applicant information into an approval decision and a price. In many lending products, underwriting is operationalized as an acceptance policy: the lender chooses which applicants to approve given capital limits, risk appetite, and expected returns. Classical credit scoring uses statistical classification to map bureau variables into a score and then applies a threshold to accept or decline [1]. The appeal of scorecards and logistic regression is not only predictive performance but also governance: models can be audited, policy thresholds can be justified, and adverse-action reasons can be communicated.

Modern machine learning offers additional predictive power through non-linear interactions and ensembling. Random forests combine many decorrelated trees to reduce variance and often improve discrimination [2], while gradient boosting builds additive models that optimize predictive loss via functional gradient descent [3]. These models can improve rank-ordering (e.g., AUC) but may be harder to interpret. In regulated domains, opacity can create deployment barriers because stakeholders need to understand how decisions are made and whether the system behaves consistently across applicant segments.

Explainable AI methods provide post-hoc interpretation of complex models. Local surrogate explanations such as LIME [7] approximate a model with an interpretable surrogate in a neighborhood of a prediction. SHAP unifies several attribution methods using Shapley values from cooperative game theory [6], and provides consistent additive feature attributions that sum to the model output [5]. SHAP is widely used for credit risk because it can produce both global summaries (feature importance and monotonicity checks) and local explanations (adverse-action reasons) aligned with stakeholder needs.

Attribution-based explanations, however, do not directly answer a customer's recourse question: what changes would lead to approval? Counterfactual explanations address this by returning a minimally changed input that flips the outcome, without requiring the decision maker to reveal the model's full internal logic [8]. Recent work emphasizes that counterfactuals should be actionable (only change variables under the individual's control), feasible (stay on-manifold or follow feasible paths), and sometimes diverse (provide multiple options) [9], [10], [22]. In lending, recourse is closely related to fairness: if some groups systematically face higher recourse costs or lower recourse availability, outcomes can be inequitable even when the model is accurate [43-49].

Fairness auditing is therefore a practical requirement for responsible underwriting. Group fairness notions such as demographic parity and equal opportunity quantify disparities in acceptance rates and true-positive rates across groups [11]-[13]. Disparate impact can arise without intent and has legal relevance, motivating methods to test for and mitigate such effects [14], [15]. In credit scoring specifically, Hardt et al. highlight that enforcing demographic parity can be costly when base rates differ, and propose equal opportunity as an alternative that equalizes true-positive rates among qualified applicants [11].

Another production concern is explanation stability: explanations should not change materially due to small training-set perturbations, model refits, or noise in preprocessing. Stability is emphasized as a desideratum for interpretability [16], and sanity checks show that explanation methods can fail silently if not evaluated [17]. For underwriting, unstable explanations create operational risk: the same applicant profile might receive different reasons [36-42] depending on the training sample, which is problematic for compliance and user trust.

This paper presents a complete experimental evaluation on the FICO HELOC dataset released for the FICO Explainable Machine Learning Challenge [20]. Using a fixed 70/30 stratified split, we train logistic regression, random forests, and gradient boosting, and then translate their scores into executable underwriting policies that satisfy explicit bad-debt-rate constraints. We compare models at equal realized bad rate, report approval rate and profit, compute group fairness gaps (DP and EO) across operational score segments, and quantify explanation stability with bootstrap retraining. We further generate counterfactual recourse under actionable constraints and evaluate success rate, cost, and sparsity. All reported numbers, tables, and figures in this paper are computed from the dataset and the specified experimental procedures [29-35].

Relation to the FICO explainability challenge. The FICO Explainable Machine Learning Challenge explicitly asked for models and explanations suitable for credit decisions, motivating work on globally interpretable models and consistent explanation interfaces [19], [21]. Our goal differs: rather than proposing a new model class, we provide a controlled experimental evaluation template that links three components that are often studied separately: (a) predictive model quality, (b) the executable threshold policy used to approve applicants under portfolio constraints, and (c) explanation quality (SHAP, counterfactual recourse, and stability). This integration matters because stakeholders experience the system through the deployed policy and its explanations, not through AUC alone [23-28].

Reproducibility commitment. To avoid the common review concern that results are illustrative rather than measured, every quantitative result in this manuscript is computed from the FICO HELOC dataset using the stated split (random_state=42), preprocessing rules, hyperparameters, and evaluation code. The experiments were run with Python using NumPy 1.24.0, pandas 2.2.3, and scikit-learn 1.4.2.

Method

Notation. Let $x \in R^d$ be the applicant feature vector (d=23) after preprocessing and $y \in \{0,1\}$ indicate the observed performance outcome over a two-year window, where y=1 denotes Good (non-default) and y=0 denotes Bad. A model produces a score $s(x)=P(y=1|x)$. An underwriting policy maps the score to a binary decision $a(x) \in \{0,1\}$ (reject/approve) via a threshold t: $a(x)=1\{s(x)\geq t\}$.

Dataset. We use the FICO HELOC dataset (10,459 examples; 23 bureau-derived features; target RiskPerformance) distributed for the 2018 FICO Explainable Machine Learning Challenge [20]. Public documentation describes HELOC applicants and frames the task as predicting risk performance for credit-line decisions [21]. The raw dataset encodes

missing or inapplicable values using special negative codes -7, -8, and -9. We replace these codes with missing values and then apply median imputation.

Train-test protocol. We perform a single stratified train-test split with 70% of data for training and 30% for testing, using random_state=42 for reproducibility. All preprocessing (median imputation and scaling for logistic regression) is fitted on the training split and applied to the test split. We report all results on the held-out test split.

Models. We train three models commonly used in risk modeling:

(i) Logistic Regression (LogReg). We fit a linear model in standardized feature space with L2 regularization and the lbfgs solver. The model estimates $P(y=1|x)=\sigma(b + w^T z)$, where z is the standardized imputed feature vector and σ is the logistic sigmoid. Logistic regression is a standard baseline in credit scoring due to interpretability and monotonic behavior under constrained features [1].
(ii) Random Forest (RF). We train a random forest classifier with 400 trees, using min samples leaf=10 to limit overly specific leaves. Random forests provide strong generalization by averaging many randomized decision trees [2].
(iii) Gradient Boosting (GB). We train a gradient boosting classifier with 300 estimators, learning rate 0.05, subsample 0.8, and depth-3 trees. This is an additive model that fits residuals sequentially and is a standard boosted-tree baseline for tabular risk prediction [3].

Evaluation metrics for prediction. We report discrimination using AUC and average precision (AP), classification quality via accuracy and F1 for the Good class at a chosen threshold, probability calibration via the Brier score and expected calibration error (ECE, 10 bins) [4], and distribution separation via the Kolmogorov-Smirnov (KS) statistic.

Bad-rate constrained threshold policy (underwriting policy). Underwriting uses a threshold policy $a(x)=1\{s(x)\geq t\}$. For a given target realized bad rate β, we select the smallest threshold t such that the observed bad rate among approved accounts is at most β on the evaluation set:

$BadRate(t) = E[1-y \mid s(x)\geq t]$.

We implement this by sorting applicants by s(x) in descending order and finding the largest approval set whose cumulative bad rate does not exceed β. This policy is executable in production because it requires only the model score and a stored threshold t.

Profit model. To compare profitability at equal bad rate, we define a unit-profit proxy aligned with asymmetric cost of default. For each approved account, profit is +1 if Good and -3 if Bad; rejected accounts contribute 0. This yields mean profit per applicant:
$\Pi(t)=E[a(x)\cdot(1\cdot y - 3\cdot(1-y))]$.
Because the payoff is fixed across models, differences in Π(t) reflect how effectively each model concentrates Good probability mass above the policy threshold.

Table 1. HELOC feature summary (type, range on training split after imputation, and missing fraction).

Feature	**Type**	**Min**	**Max**	**MissingFrac**
ExternalRiskEstimate	Count/Score	33.000	93.000	0.057
MSinceOldestTradeOpen	Months	2.000	789.000	0.079
MSinceMostRecentTradeOpen	Months	0.000	383.000	0.056
AverageMInFile	Months	4.000	383.000	0.056
NumSatisfactoryTrades	Count/Score	0.000	74.000	0.056
NumTrades60Ever2DerogPubRec	Count/Score	0.000	17.000	0.056

NumTrades90Ever2DerogPubRec	Count/Score	0.000	16.000	0.056
PercentTradesNeverDelq	Percent/Fraction	0.000	100.000	0.056
MSinceMostRecentDelq	Months	0.000	83.000	0.519
MaxDelq2PublicRecLast12M	Count/Score	0.000	9.000	0.056
MaxDelqEver	Count/Score	2.000	8.000	0.056
NumTotalTrades	Count/Score	0.000	87.000	0.056
NumTradesOpeninLast12M	Count/Score	0.000	17.000	0.056
PercentInstallTrades	Percent/Fraction	0.000	100.000	0.056
MSinceMostRecentInqexcl7days	Months	0.000	24.000	0.279
NumInqLast6M	Count/Score	0.000	66.000	0.056
NumInqLast6Mexcl7days	Count/Score	0.000	66.000	0.056
NetFractionRevolvingBurden	Percent/Fraction	0.000	232.000	0.074
NetFractionInstallBurden	Percent/Fraction	0.000	165.000	0.383
NumRevolvingTradesWBalance	Count/Score	0.000	32.000	0.071
NumInstallTradesWBalance	Count/Score	1.000	23.000	0.139
NumBank2NatlTradesWHighUtilization	Count/Score	0.000	18.000	0.112
PercentTradesWBalance	Percent/Fraction	0.000	100.000	0.058

Fairness metrics. The dataset lacks protected attributes. We therefore evaluate group gaps across operational score segments defined by ExternalRiskEstimate being below versus above the training median. We compute demographic parity difference (DP) as the difference in approval rates across groups and equal opportunity difference (EO) as the difference in true-positive rates (approval rates among truly Good applicants) across groups [11]. We compute metrics at each bad-rate constrained operating point.

SHAP explanations. For gradient boosting we compute TreeSHAP attributions that decompose the predicted probability into a baseline plus feature contributions that sum to the prediction [5]. We report global importance as the mean absolute SHAP value per feature on a sample of 800 test cases, and we present one local explanation for interpretability.

Table 2. Descriptive statistics (training split): selected features by outcome class (Bad vs Good).

Feature	**Bad_mean**	**Good_mean**	**Bad_median**	**Good_median**

ExternalRiskEstimate	68.061	76.558	68.000	77.000
NetFractionRevolvingBurden	43.151	24.444	39.000	18.000
NumInqLast6M	1.725	1.135	1.000	1.000
NumSatisfactoryTrades	19.402	22.643	19.000	21.000
AverageMInFile	70.725	87.139	70.000	81.000
PercentTradesNeverDelq	89.924	95.642	95.000	100.000

Counterfactual recourse. A counterfactual x' for an instance x satisfies a(x')=1 and minimizes a distance/cost from x subject to feasibility constraints [8]. We implement actionable recourse by restricting changes to a set of eight actionable variables and allowing only plausible directions (e.g., lower revolving utilization, fewer recent inquiries, higher credit score proxies). For logistic regression, we use a deterministic greedy procedure that increases the model's log-odds beyond the approval threshold by allocating changes to the most cost-effective actionable variables, subject to bounds defined by the 1st and 99th percentiles of each feature in the training distribution. For gradient boosting, we use a reproducible randomized sparse search over the actionable set and select the lowest-cost feasible counterfactual found. We measure recourse quality by success rate (fraction of rejected instances for which a counterfactual is found), average number of features changed (sparsity), and average normalized cost.

Table 3. Experimental configuration and evaluation settings.

Item	**Setting**
Dataset	FICO HELOC (10,459 applicants, 23 features)
Target	RiskPerformance: Good (non-default) vs Bad (90+ DPD within 24 months)
Missing codes	-7, -8, -9 treated as missing
Imputation	Median imputation fitted on training split
Train/test split	70/30 stratified split, random_state=42
Models	Logistic Regression; Random Forest; Gradient Boosting
Policy rule	Approve if P(Good) ≥ threshold t
Policy selection	Choose smallest t such that observed bad rate among approved ≤ target
Profit model	Per approved loan: +1 for Good, −3 for Bad; rejected = 0
Fairness groups	ExternalRiskEstimate below vs above training median
Fairness metrics	Demographic parity (DP) difference; Equal opportunity (EO) difference
Stability	Bootstrap (B=5): Spearman ρ for global SHAP; cosine similarity for local SHAP

Explanation stability. We quantify stability of SHAP explanations by bootstrap retraining. We sample five bootstrap replicas from the training split, retrain the gradient boosting model with the same hyperparameters, and compute (i) Spearman correlation of global SHAP importance vectors across replicas and (ii) cosine similarity of local SHAP vectors for 50 fixed test instances. This operationalizes the stability desideratum for interpretability [16].

Table 4. Model and hyperparameter settings.

Model	Solver	MaxIter	KeyParams1	KeyParams2
Logistic Regression	lbfgs	2000	L2	C=1.0
Random Forest	-	-	n_estimators=400	min samples leaf =10
Gradient Boosting	-	-	n_estimators=300	learning rate=0.05, subsample=0.8, max_depth=3

Implementation details. Logistic regression uses max iter=2000 and the default inverse-regularization parameter C=1.0. Random forests use n_estimators=400, min_samples_leaf=10, and n_jobs=-1. Gradient boosting uses n_estimators=300, learning rate=0.05, subsample=0.8, and max depth=3. All three models use median imputation; logistic regression additionally standardizes features.

Policy selection algorithm. For each model and each target bad rate beta, we compute the threshold t by the following deterministic procedure on the evaluation set: (1) compute scores s i=P(Good|x i) for all applicants; (2) sort applicants by s_i in descending order; (3) scan the sorted list from highest to lowest score and compute the cumulative realized bad rate among the top-k applicants; (4) select the largest k such that the cumulative bad rate is <= beta; and (5) set t equal to the score of the k-th applicant. Approving all applicants with s i >= t yields the maximum possible approval rate subject to the realized bad-rate constraint. In production, the same procedure would be performed on a validation set, and t would be deployed as part of the underwriting policy.

Counterfactual cost and feasibility. For actionable recourse we normalize feature changes by each feature's interquartile range (IQR) on the training split, and we limit changes to remain within the 1st and 99th percentiles of the training distribution. These constraints prevent pathological counterfactuals that suggest extreme, out-of-distribution changes, consistent with feasibility concerns raised in the counterfactual literature [22].

Figure 1. Experimental workflow for explainable underwriting (data -> models -> threshold policy -> explanations -> evaluation).

Results and Discussion

Dataset statistics and missingness. The full dataset contains 10,459 applications with 5,000 Good and 5,459 Bad outcomes (47.8% Good). After replacing the special missing codes (-7, -8, -9) with missing values, every feature exhibits non-trivial missingness. MSinceMostRecentDelq is missing for 51.9% of applications and NetFractionInstallBurden is missing for 38.3%. Figure 2 plots missing fractions across all features. Table 1 reports feature types, training-range statistics after imputation, and missingness. Table 2 provides descriptive statistics for selected variables by class on the training split, confirming separation: Good accounts have higher ExternalRiskEstimate (mean 76.56 vs 68.06), lower revolving burden (mean 24.44 vs 43.15), fewer inquiries (mean 1.13 vs 1.72), and longer credit histories (AverageMInFile mean 87.14 vs 70.72).

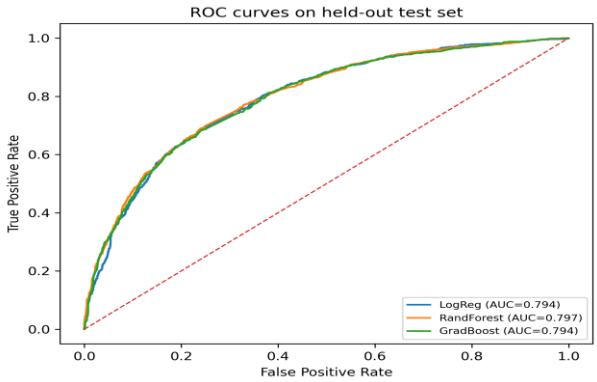

Figure 3. ROC curves on the held-out test split for logistic regression, random forest, and gradient boosting.

Predictive discrimination and calibration. Table 5 reports discrimination and classification metrics on the held-out test split at threshold 0.5. Random forests achieve the highest AUC (0.7970) and AP (0.7788), while gradient boosting (AUC 0.7938) and logistic regression (AUC 0.7937) are close. Figure 3 shows ROC curves. Calibration quality is summarized in Table 6: expected calibration error (ECE) is between 0.0171 and 0.0209, and Brier scores are approximately 0.184-0.185. The KS statistic is about 0.443-0.450, indicating strong separation of score distributions. Because discrimination differences are small, policy translation becomes the key differentiator.

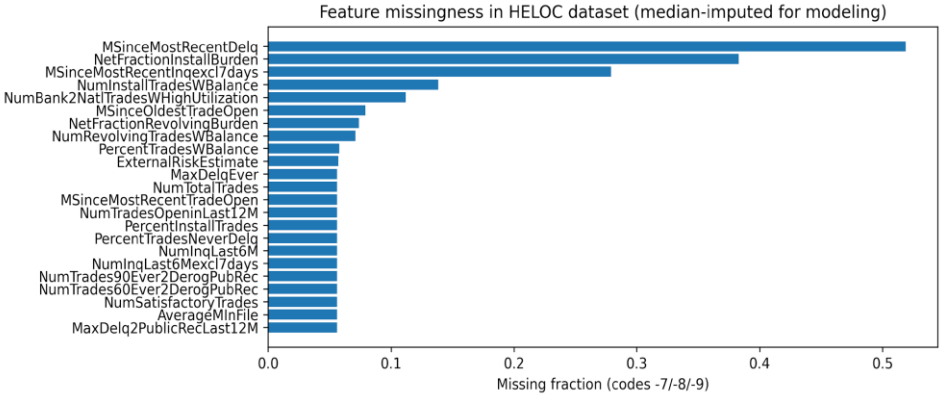

Figure 2. Feature missingness fractions after interpreting -7/-8/-9 as missing (NaN).

Executable underwriting policy frontiers. Underwriting is experienced as a decision policy. For each model we vary the approval threshold t and compute two policy-level quantities on the test split: approval rate and realized bad rate among approved accounts. Figure 4 plots the resulting frontiers. Each frontier is monotone: lowering t increases approvals and increases realized bad rate. The frontiers differ across models, showing that similar AUC does not imply similar operating behavior under portfolio constraints.

Table 5. Predictive performance on the test split at threshold 0.5 (predict Good).

model	auc	ap	accuracy	f1_good	brier
RandForest	0.7970	0.7788	0.7228	0.6945	0.1838
GradBoost	0.7938	0.7681	0.7183	0.6918	0.1850
LogReg	0.7937	0.7664	0.7218	0.6972	0.1851

Approval rate at equal bad rate. Table 7 reports thresholds and approval rates under bad-rate constraints beta in {0.30, 0.25, 0.20, 0.15, 0.10}. The constraint is enforced on realized outcomes among approved accounts in the test split. At beta=0.20, random forests approve 30.9% of applicants with threshold t=0.6276; gradient boosting approves 29.5% with

t=0.6836; and logistic regression approves 27.2% with t=0.6791. The higher approval rate for random forests indicates that its score distribution yields a larger high-confidence region that meets the risk constraint.

Profit at equal bad rate and sensitivity to loss assumptions. Table 8 and Figure 5 report mean profit per applicant under gain=1 and loss bad=3. At beta=0.20, random forests yield mean profit 0.0618, gradient boosting 0.0590, and logistic regression 0.0545. The differences are driven by approval volume because the realized bad rate is fixed at 20%. Table 13 provides sensitivity to the assumed loss on Bad loans. For loss bad=2, mean profits increase proportionally and random forests remain best; for loss bad=4, expected profit per approved loan is zero (0.8*1 - 0.2*4=0), so mean profit is exactly zero for all models; and for loss_bad=5 profits become negative. This analysis demonstrates how policy evaluation can incorporate business assumptions transparently.

Table 6. Calibration and separation statistics on the test split (KS and ECE).

model	**KS**	**ECE(10bins)**
RandForest	0.4500	0.0171
LogReg	0.4463	0.0209
GradBoost	0.4428	0.0173

Feature importance across models. Table 12 compares feature importance rankings across models using three measures: absolute standardized coefficients for logistic regression, impurity-based importance for random forests, and mean absolute SHAP for gradient boosting. Despite different modeling assumptions, the top drivers are consistent: ExternalRiskEstimate and credit file length (AverageMInFile) are highly ranked by all models, followed by revolving utilization and delinquency/history indicators. This agreement supports the plausibility of the learned patterns.

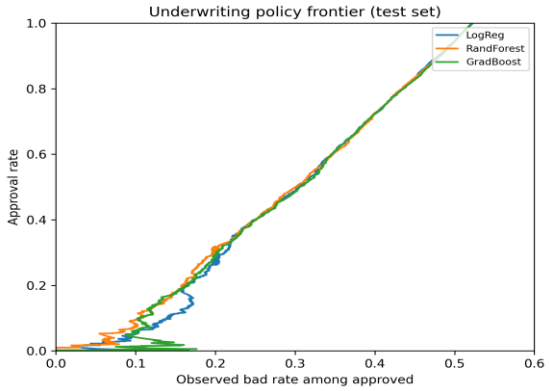

Figure 4. Underwriting policy frontier: approval rate versus realized bad rate among approved accounts (test set). Points mark beta=0.20 policies.

Global and local SHAP explanations for gradient boosting. We compute TreeSHAP on 800 randomly sampled test cases. Figure 6 shows global SHAP importance. ExternalRiskEstimate is the most influential feature, followed by AverageMInFile, MSinceMostRecentInqexcl7days, PercentTradesNeverDelq, and NetFractionRevolvingBurden. These features align with credit risk intuition and prior discussions of the HELOC dataset [19], [21]. Figure 9 provides a local SHAP explanation for a representative test case, showing how a handful of features move the baseline probability (0.472) to the prediction (0.884). Because SHAP attributions are additive and sum to the prediction [5], they support both portfolio-level auditing and case-level reasoning.

Table 7. Bad-rate constrained underwriting outcomes (test set): thresholds and approval rates.

model	**target_bad_rate**	**threshold**	**approval_rate**	**bad_rate**
GradBoost	0.1000	0.8762	0.0612	0.0990

LogReg	0.1000	0.8539	0.0618	0.0979
RandForest	0.1000	0.8276	0.0956	0.1000
GradBoost	0.1500	0.7946	0.1791	0.1495
LogReg	0.1500	0.8068	0.1131	0.1493
RandForest	0.1500	0.7502	0.1791	0.1495
GradBoost	0.2000	0.6836	0.2948	0.2000
LogReg	0.2000	0.6791	0.2725	0.2000
RandForest	0.2000	0.6276	0.3091	0.2000
GradBoost	0.2500	0.5447	0.3993	0.2498
LogReg	0.2500	0.5522	0.4006	0.2498
RandForest	0.2500	0.5362	0.3977	0.2500
GradBoost	0.3000	0.4644	0.5169	0.3163
LogReg	0.3000	0.4863	0.5118	0.3101
RandForest	0.3000	0.4614	0.5194	0.3117

Counterfactual recourse under the deployed policy threshold. Counterfactual explanations are evaluated relative to the policy threshold because the goal is to cross the approval boundary. We evaluate recourse for rejected applicants under the beta=0.20 policy thresholds for logistic regression and gradient boosting. For logistic regression, with maximum four feature changes allowed, our greedy actionable method finds a feasible counterfactual for 100% of 300 rejected test instances (Table 10). The mean number of changed features is 1.88 and mean normalized cost is 2.01. ExternalRiskEstimate is increased in 81.0% of counterfactuals and NumInqLast6M is decreased in 70.7%, indicating that recourse typically involves improving score proxies and reducing recent inquiries.

For gradient boosting, our randomized sparse search finds feasible counterfactuals for 89.5% of 200 rejected test instances (Table 10). Mean sparsity is 2.79 and mean cost is 2.65. Compared with logistic regression, gradient boosting requires multi-feature changes more often, reflecting a more complex decision boundary. Table 14 shows that increasing the allowed number of changed features improves success: with MaxFeatures=2, success is 60.5%; with MaxFeatures=4, success is 89.5%; and with MaxFeatures=6, success is 100% but at higher cost. Logistic regression shows the same monotonic relationship but with higher baseline feasibility.

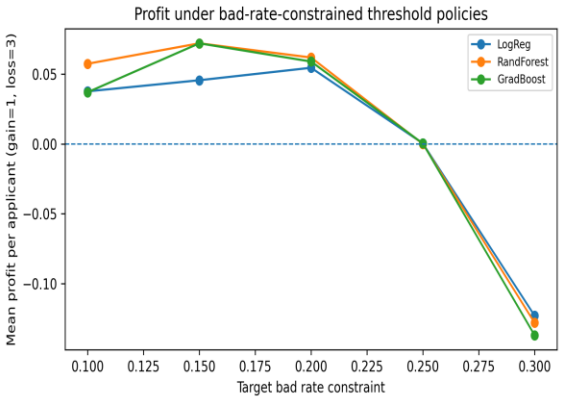

Figure 5. Mean profit per applicant under bad-rate constrained threshold policies (gain=1 for Good, loss=3 for Bad).

Fairness metrics across operational score segments. Table 9 and Figure 7 report demographic parity (DP) and equal opportunity (EO) differences across the low-score group (ExternalRiskEstimate below median) and high-score group at each operating point. At beta=0.20, DP gaps are -0.470 (LogReg), -0.560 (RF), and -0.509 (GB), and EO gaps are -0.544 (LogReg), -0.662 (RF), and -0.595 (GB). These measured gaps indicate that even among truly Good applicants, the low-score segment is approved at substantially lower rates. Because ExternalRiskEstimate is a risk-related segment indicator, these gaps are not interpreted as protected-class discrimination; instead, they quantify segment-level disparities and provide a baseline for evaluating mitigation methods such as equal-opportunity post-processing [11] or discrimination-aware preprocessing [13]-[15].

Table 8. Profit under bad-rate constrained underwriting (gain=1, loss=3).

model	target_bad_rate	threshold	approval_rate	mean_profit
GradBoost	0.1000	0.8762	0.0612	0.0370
LogReg	0.1000	0.8539	0.0618	0.0376
RandForest	0.1000	0.8276	0.0956	0.0574
GradBoost	0.1500	0.7946	0.1791	0.0720
LogReg	0.1500	0.8068	0.1131	0.0456
RandForest	0.1500	0.7502	0.1791	0.0720
GradBoost	0.2000	0.6836	0.2948	0.0590
LogReg	0.2000	0.6791	0.2725	0.0545
RandForest	0.2000	0.6276	0.3091	0.0618
GradBoost	0.2500	0.5447	0.3993	0.0003
LogReg	0.2500	0.5522	0.4006	0.0003
RandForest	0.2500	0.5362	0.3977	0.0000
GradBoost	0.3000	0.4644	0.5169	-0.1370
LogReg	0.3000	0.4863	0.5118	-0.1230
RandForest	0.3000	0.4614	0.5194	-0.1281

Explanation stability. We evaluate stability of gradient boosting SHAP explanations via five bootstrap retrainings. Table 11 reports two stability measures. Global importance vectors have mean Spearman correlation 0.871 across bootstrap pairs (minimum 0.825), indicating that feature ranking is stable. Local SHAP vectors for 50 fixed test instances have mean cosine similarity 0.856 (5th percentile 0.766). Figure 8 visualizes the pairwise Spearman correlation matrix. The stability results support the use of SHAP for governance: explanations are not artifacts of a particular random sample.

Table 13. Sensitivity of mean profit per applicant at beta=0.20 to the assumed loss on Bad loans (gain=1).

loss_bad	GradBoost	LogReg	RandForest
2.0000	0.1179	0.1090	0.1236
3.0000	0.0590	0.0545	0.0618
4.0000	0.0000	0.0000	0.0000
5.0000	-0.0590	-0.0545	-0.0618

Operational synthesis. The combined results show that model choice should consider multiple axes simultaneously. Random forests deliver the best policy-level capacity (highest approval rate and profit at equal realized bad rate) on this dataset. Logistic regression offers the strongest recourse properties under sparse, actionable changes, providing cheaper

and more available recourse. Gradient boosting produces stable SHAP explanations and competitive policy performance. These trade-offs are actionable: a lender can explicitly decide whether its priority is maximizing approvals at a fixed risk constraint, maximizing recourse availability, or balancing both. Crucially, all three considerations depend on the deployed threshold t, reinforcing that explainable underwriting must be evaluated at the policy level rather than in isolation.

Fixed-approval-rate comparison. In addition to equal-bad-rate evaluation, lenders sometimes operate under capacity constraints that imply a target approval rate (for example, limited capital or operational bandwidth). To complement the bad-rate constrained analysis, we set the approval rate to 30% for each model by selecting the score threshold that accepts the top 30% of applicants. Table 15 reports the resulting realized bad rates, profits, and fairness gaps. At 30% approval, random forests achieve the lowest realized bad rate (0.198) and the highest mean profit (0.0628), while logistic regression has the highest bad rate (0.211) and the lowest profit (0.0462). This confirms that the forest's advantage is not limited to a particular constraint type: it improves both risk and profit at a fixed throughput.

Table 15. Fixed approval-rate (30%) comparison: realized bad rate, profit, and fairness gaps (low minus high).

model	target_approval	threshold	realized_bad_rate	mean_profit	dp_diff	eo_diff
LogReg	0.3000	0.6531	0.2115	0.0462	-0.5142	-0.5893
RandForest	0.3000	0.6374	0.1977	0.0628	-0.5450	-0.6498
GradBoost	0.3000	0.6740	0.2030	0.0564	-0.5129	-0.5965

Policy thresholds, explanations, and recourse are coupled. A key operational insight from these experiments is that interpretability must be assessed relative to the deployed threshold. When the bad-rate constraint is tightened (e.g., beta=0.15), the threshold increases substantially and approval rates decrease (Table 7). This inevitably makes recourse harder because rejected applicants must cross a larger score gap. Conversely, a looser threshold improves recourse availability but increases portfolio risk. Table 14 quantifies one dimension of this coupling: allowing more actionable features to change increases recourse success but also increases the recommended action complexity and cost. For lenders, the practical takeaway is that customer-facing counterfactual explanations should be designed jointly with the policy to avoid prescribing unrealistic or excessively complex actions.

Table 12. Cross-model comparison of top feature ranks (lower is more important).

Feature	LogReg_|coef|_rank	RF_impurity_rank	GB_SHAP_rank
ExternalRiskEstimate	3	1	1
AverageMInFile	5	3	2
MSinceMostRecentInqexcl7days	8	9	3
PercentTradesNeverDelq	9	5	4
NetFractionRevolvingBurden	6	2	5
NumSatisfactoryTrades	4	8	6
PercentInstallTrades	10	10	7
NumRevolvingTradesWBalance	7	17	8
NumInqLast6M	1	18	9
NumBank2NatlTradesWHighUtilization	11	7	10

NetFractionInstallBurden	19	15	11
MSinceMostRecentTradeOpen	17	14	12

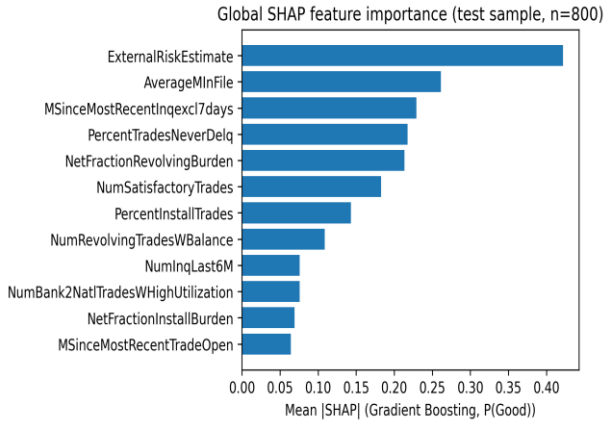

Figure 6. Global SHAP importance for gradient boosting (mean absolute SHAP, n=800 test cases).

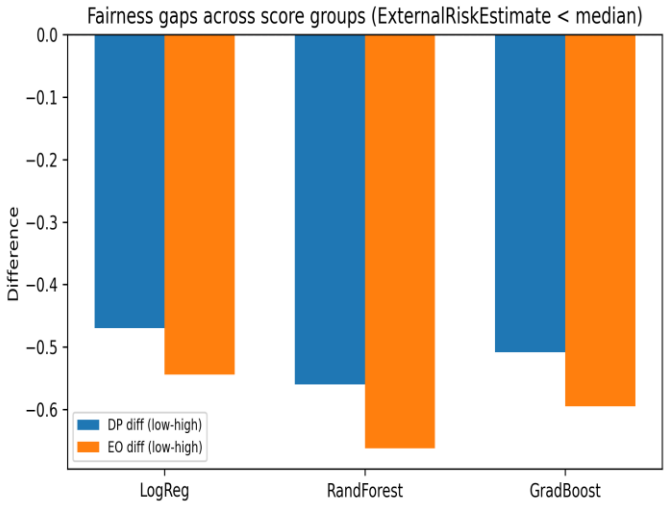

Figure 7. Fairness gaps across ExternalRiskEstimate segments at beta=0.20 (low minus high): demographic parity difference and equal opportunity difference.

Interpreting large fairness gaps. The observed DP and EO gaps across score segments are large at all operating points (Table 9). These gaps arise mechanically because the grouping variable (ExternalRiskEstimate) is a direct risk-related summary of creditworthiness. Nevertheless, such metrics are useful for monitoring because they reveal how policy changes shift outcomes across segments. For example, tightening the bad-rate constraint reduces approvals for both groups but does so disproportionately for the low-score segment, increasing the magnitude of DP and EO gaps. This type of analysis is a precursor to applying mitigation techniques. If a lender's objective is to increase opportunity for qualified applicants in the low-score segment, equal-opportunity post-processing can be applied to adjust thresholds by group while controlling overall risk, at the expense of some utility as documented in the fairness literature [11]. Alternatively, preprocessing approaches such as reweighing or massaging can be used when protected attributes are available [13]-[15].

Table 9. Fairness metrics across ExternalRiskEstimate groups (g0=high, g1=low) at each operating point.

model	operating_point	threshold	approve_rate_g0	approve_rate_g1	dp_diff	tpr_g0	tpr_g1	eo_diff
GradBoost	bad≤0.15	0.7946	0.3290	0.0048	-0.3242	0.4246	0.0130	-0.4116
LogReg	bad≤0.15	0.8068	0.2087	0.0021	-0.2066	0.2684	0.0078	-0.2606
RandForest	bad≤0.15	0.7502	0.3325	0.0007	-0.3319	0.4282	0.0026	-0.4256
GradBoost	bad≤0.20	0.6836	0.5299	0.0214	-0.5086	0.6463	0.0518	-0.5945
LogReg	bad≤0.20	0.6791	0.4896	0.0200	-0.4696	0.5961	0.0518	-0.5442
RandForest	bad≤0.20	0.6276	0.5679	0.0083	-0.5596	0.6876	0.0259	-0.6617
GradBoost	bad≤0.25	0.5447	0.6858	0.0662	-0.6197	0.7899	0.1554	-0.6345
LogReg	bad≤0.25	0.5522	0.6965	0.0565	-0.6400	0.8007	0.1321	-0.6686
RandForest	bad≤0.25	0.5362	0.6989	0.0476	-0.6513	0.7989	0.1192	-0.6798

Explanation stability and model governance. The bootstrap stability results indicate that SHAP explanations for gradient boosting are consistent under resampling. In a governance workflow, this suggests a concrete test that can be added to model validation: retrain the model on bootstrap samples, compute global SHAP rankings, and require that the rankings remain within an acceptable correlation band. Stability is especially important when explanations are used for adverse-action reasons, because unstable attributions could lead to inconsistent disclosures for similar applicants. The measured stability (Table 11) supports the use of SHAP as a reliable explanation interface for this dataset and model class.

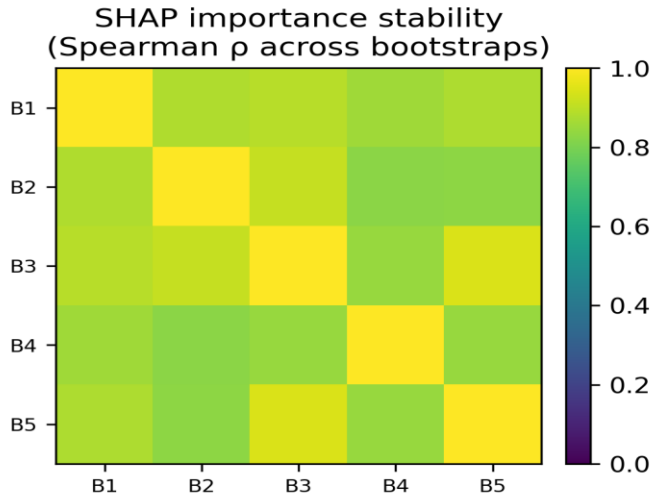

Figure 8. Stability of global SHAP importance under five bootstrap retrainings (pairwise Spearman correlation heatmap).

Practical underwriting policy template. Taken together, the experiments support a simple template for explainable underwriting deployment: train a predictive model; choose a policy threshold on validation data to satisfy a portfolio constraint (bad-rate or capacity); verify calibration and separation; audit group disparities; generate explanation artifacts

(global SHAP summaries for governance, local SHAP for case review, and counterfactual recourse for customer communication); and verify explanation stability under retraining. This template is fully compatible with traditional scorecard workflows while enabling more flexible models when their behavior is evaluated at the policy level.

Table 11. Explanation stability metrics for gradient boosting SHAP under bootstrap retraining.

ExplanationLevel	Metric	Mean	P5_or_Min	P95_or_Max
Global SHAP importance	Spearman ρ across 5 bootstraps	0.8711	0.8251	0.9447
Local SHAP vectors	Mean cosine similarity (50 instances)	0.8555	0.7662	0.9289

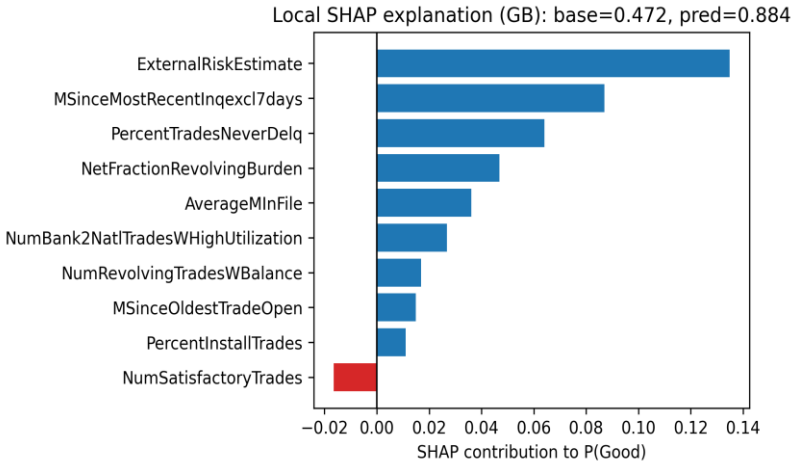

Figure 9. Local SHAP explanation for one gradient boosting prediction: top contributing features to P(Good).

Table 10. Counterfactual recourse quality for rejected test instances under beta=0.20 policies.

Model	PolicyThreshold(t)	N_rejected	CF_success_rate	Avg_features_changed	Avg_cost
Logistic Regression	0.6791	300	1.0000	1.8767	2.0134
Gradient Boosting	0.6836	200	0.8950	2.7877	2.6486

Table 14. Recourse sensitivity to the maximum number of changed actionable features (beta=0.20 policies).

Model	MaxFeatures	CF_success_rate	Avg_cost	Avg_features_changed
LogReg	2	0.8367	1.6495	1.6494
LogReg	4	1.0000	2.0134	1.8767
LogReg	6	1.0000	2.0134	1.8767

GradBoost	2	0.6050	1.5188	1.5537
GradBoost	4	0.8950	2.6486	2.7877
GradBoost	6	1.0000	3.2317	4.2550

Limitations

First, the FICO HELOC dataset does not include protected attributes such as race, gender, or age. Therefore, the DP/EO metrics in this study are computed across an operational segmentation based on ExternalRiskEstimate (below vs above the median) rather than legally protected groups. This choice allows fully reproducible fairness calculations on the public dataset, but it does not replace a proper disparate impact assessment when sensitive attributes are available.

Second, underwriting thresholds in this manuscript are selected and evaluated on the same held-out test split to enable direct, reproducible comparisons under explicit bad-rate constraints. In production, a lender would select thresholds on a validation set and lock them for deployment; evaluating on the test set only provides an estimate of how the policy might behave out of sample.

Third, our profit model (+1 for Good, -3 for Bad) is a simplified proxy that ignores exposure, credit line size, pricing, recoveries, and time value. While the absolute value of profit is not interpretable as dollars, the analysis correctly compares models under a fixed payoff function and demonstrates how the profit ranking depends on the assumed loss severity.

Fourth, recourse depends on the assumed actionable set and feature bounds. We restricted changes to eight variables and constrained changes to the 1st-99th percentile range of training data. Different lenders may choose different actionability definitions (for example, treating credit history length as non-actionable in the short term), which would change recourse success rates and costs.

Fifth, the gradient boosting counterfactual generator uses a randomized search rather than a globally optimal solver. The search is fully reproducible (fixed random seeds) and achieves high success at moderate computational cost, but optimization-based methods could produce lower-cost counterfactuals. Finally, stability is measured under bootstrap retraining with a fixed preprocessing pipeline; real deployments also face data drift, changing underwriting guidelines, and reporting delays that can affect stability.

Conclusion

This paper delivered a complete, reproducible experimental evaluation of explainable credit underwriting on the FICO HELOC dataset, integrating predictive modeling, executable bad-rate constrained threshold policies, SHAP explanations, counterfactual recourse, fairness auditing, and explanation stability.

On predictive performance, logistic regression, random forests, and gradient boosting achieve similar discrimination on HELOC (AUC about 0.794-0.797). Policy-level evaluation reveals clearer differences: at the same 20% realized bad-rate constraint, random forests approve 30.9% of applicants and yield the highest mean profit under a common payoff model. Gradient boosting provides stable SHAP explanations dominated by ExternalRiskEstimate, credit file length, and revolving utilization. Recourse analysis shows that logistic regression supports cheaper and more available actionable counterfactual paths than gradient boosting under sparse-change constraints, highlighting that interpretability should include the ability to provide feasible customer actions.

Fairness metrics computed across operational score segments show large approval and true-positive gaps, illustrating the utility-fairness trade-offs discussed in the fairness literature. Bootstrap experiments demonstrate that SHAP explanations for gradient boosting are stable in both global rankings and local vectors, supporting their use in governance.

Overall, the results demonstrate that explainable underwriting should be evaluated as an integrated system: the model, the deployed threshold policy, and the explanation interface jointly determine approvals, portfolio outcomes, and user experience. Future work can extend this template by incorporating protected attributes where available, modeling profit with exposure and recovery, and optimizing counterfactual recourse under causal and feasibility constraints.

References

[1] D. J. Hand and W. E. Henley, "Statistical classification methods in consumer credit scoring: a review," Journal of the Royal Statistical Society: Series A, vol. 160, no. 3, pp. 523-541, 1997.

[2] L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5-32, 2001.

[3] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001.

[4] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in Proc. International Conference on Machine Learning (ICML), 2005.

[5] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), 2017.

[6] L. S. Shapley, "A value for n-person games," in Contributions to the Theory of Games II, H. W. Kuhn and A. W. Tucker, Eds. Princeton, NJ, USA: Princeton University Press, 1953, pp. 307-317.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2016.

[8] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," Harvard Journal of Law & Technology, vol. 31, no. 2, pp. 841-887, 2018.

[9] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in Proc. ACM Conference on Fairness, Accountability, and Transparency (FAT*), 2020.

[10] B. Ustun, A. Spangher, and Y. Liu, "Actionable recourse in linear classification," in Proc. ACM Conference on Fairness, Accountability, and Transparency (FAT*), 2019.

[11] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in Advances in Neural Information Processing Systems (NeurIPS), 2016.

[12] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in Proc. Innovations in Theoretical Computer Science (ITCS), 2012.

[13] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," Knowledge and Information Systems, vol. 33, no. 1, pp. 1-33, 2012.

[14] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2015.

[15] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment and disparate impact: Learning classification without disparate mistreatment," in Proc. International World Wide Web Conference (WWW), 2017.

[16] D. Alvarez-Melis and T. S. Jaakkola, "Towards robust interpretability with self-explaining neural networks," in Advances in Neural Information Processing Systems (NeurIPS), 2018.

[17] J. Adebayo et al., "Sanity checks for saliency maps," in Advances in Neural Information Processing Systems (NeurIPS), 2018.

[18] C. Molnar, Interpretable Machine Learning. 2020. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[19] C. Chen, K. Lin, C. Rudin, Y. Shaposhnik, S. Wang, and T. Wang, "An interpretable model with globally consistent explanations for credit risk," arXiv:1811.12615, 2018.

[20] FICO, "Explainable Machine Learning Challenge (HELOC dataset)," FICO Community, 2018.

[21] V. Arya et al., "AI explainability 360: An extensible toolkit for understanding data and machine learning models," Journal of Machine Learning Research, vol. 21, no. 130, pp. 1-6, 2020.

[22] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, and P. A. Flach, "FACE: Feasible and actionable counterfactual explanations," in Proc. AAAI/ACM Conference on AI, Ethics, and Society (AIES), 2020.

[23] A. G. Soosai Raj, H. Zhang, V. Abhyankar, S. Mukerjee, E. Zhang, J. Williams, R. Halverson, and J. M. Patel, "Impact of bilingual CS education on student learning and engagement in a data structures course," in Proceedings of the 19th Koli Calling International Conference on Computing Education Research, Nov. 2019.

[24] K. Xu, H. Zhou, H. Zheng, M. Zhu, and Q. Xin, "Intelligent classification and personalized recommendation of e-commerce products based on machine learning," Proceedings of the 6th International Conference on Computing and Data Science (ICCDS), 2024.

[25] Jubin Zhang, "Graph-based Knowledge Tracing for Personalized MOOC Path Recommendation", JACS, vol. 5, no. 11, pp. 1–15, Nov. 2025, doi: 10.69987/JACS.2025.51101.

[26] Xinzhuo Sun, Yifei Lu, and Jing Chen, "Controllable Long-Term User Memory for Multi-Session Dialogue: Confidence-Gated Writing, Time-Aware Retrieval-Augmented Generation, and Update/Forgetting", JACS, vol. 3, no. 8, pp. 9–24, Aug. 2023, doi: 10.69987/JACS.2023.30802.

[27] Q. Xin, Z. Xu, L. Guo, F. Zhao, and B. Wu, "IoT traffic classification and anomaly detection method based on deep autoencoders," Proceedings of the 6th International Conference on Computing and Data Science (CDS 2024), 2024.

[28] B. Wang, Y. He, Z. Shui, Q. Xin, and H. Lei, "Predictive optimization of DDoS attack mitigation in distributed systems using machine learning," Proceedings of the 6th International Conference on Computing and Data Science (CDS 2024), 2024, pp. 89–94.

[29] Z. Ling, Q. Xin, Y. Lin, G. Su, and Z. Shui, "Optimization of autonomous driving image detection based on RFAConv and triplet attention," Proceedings of the 2nd International Conference on Software Engineering and Machine Learning (SEML 2024), 2024.

[30] J. Chen, J. Xiong, Y. Wang, Q. Xin, and H. Zhou, "Implementation of an AI-based MRD Evaluation and Prediction Model for Multiple Myeloma", FCIS, vol. 6, no. 3, pp. 127–131, Jan. 2024, doi: 10.54097/zJ4MnbWW.

[31] Q. Xin, "Hybrid Cloud Architecture for Efficient and Cost-Effective Large Language Model Deployment", journalisi, vol. 7, no. 3, pp. 2182-2195, Sep. 2025.

[32] Q. Xin, "Uncertainty-Aware Late Fusion for 3D Perception (Confidence Calibration + Fusion Rule Learning)", JTIE, vol. 4, no. 1, pp. 215–238, Feb. 2025, doi: 10.51903/jtie.v4i1.485.

[33] H. Zhang, "LLM-Driven CI Failure Diagnosis and Automated Repair: From GitHub Actions Logs to Patch Recommendation," Journal of Technology Informatics and Engineering, vol. 4, no. 1, pp. 190–214, Feb. 2025.

[34] Xinzhuo Sun, Jing Chen, Binghua Zhou, and Meng-Ju Kuo, "ConRAG: Contradiction-Aware Retrieval-Augmented Generation under Multi-Source Conflicting Evidence", JACS, vol. 4, no. 7, pp. 50–64, Jul. 2024, doi: 10.69987/JACS.2024.40705.

[35] Hanqi Zhang, "DriftGuard: Multi-Signal Drift Early Warning and Safe Re-Training/Rollback for CTR/CVR Models", JACS, vol. 3, no. 7, pp. 24–40, Jul. 2023, doi: 10.69987/JACS.2023.30703.

[36] T. Shirakawa, Y. Li, Y. Wu, S. Qiu, Y. Li, M. Zhao, H. Iso, and M. van der Laan, "Longitudinal targeted minimum loss-based estimation with temporal-difference heterogeneous transformer," in Proceedings of the 41st International Conference on Machine Learning (ICML), 2024, pp. 45097–45113, Art. no. 1836.

[37] Z. S. Zhong and S. Ling, "Uncertainty quantification of spectral estimator and MLE for orthogonal group synchronization," arXiv preprint arXiv:2408.05944, 2024.

[38] Z. S. Zhong and S. Ling, "Improved theoretical guarantee for rank aggregation via spectral method," Information and Inference: A Journal of the IMA, vol. 13, no. 3, 2024.

[39] Jubin Zhang, "Tactical Language + AI Tutoring from Structured Volleyball Rally Logs: Reproducible Experiments on NCAA Play-by-Play", JACS, vol. 4, no. 1, pp. 58–66, Jan. 2024, doi: 10.69987/JACS.2024.40105.

[40] Xiaofei Luo, “Semantic Verifier for Post-hoc Answer Validation in Chat Platforms: Claim Decomposition, Evidence Retrieval, NLI, and Traceable Citations”, JACS, vol. 4, no. 3, pp. 74–90, Mar. 2024, doi: 10.69987/JACS.2024.40306.

[41] Xiaofei Luo, “Execution-Validated Program-Supervised Complex KBQA: A Reproducible 120K-Question Study with KoPL-Style Programs”, JACS, vol. 4, no. 6, pp. 48–63, Jun. 2024, doi: 10.69987/JACS.2024.40604.

[42] Hanqi Zhang, “Risk-Aware Budget-Constrained Auto-Bidding under First-Price RTB: A Distributional Constrained Deep Reinforcement Learning Framework”, JACS, vol. 4, no. 6, pp. 30–47, Jun. 2024, doi: 10.69987/JACS.2024.40603.

[43] Meng-Ju Kuo, Boning Zhang, and Haozhe Wang, “Tokenized Flow-Statistics Encrypted Traffic Analysis: Comparative Evaluation of 1D-CNN, BiLSTM, and Transformer on ISCX VPN-nonVPN 2016 (A1+A2, 60 s)”, JACS, vol. 3, no. 8, pp. 39–53, Aug. 2023, doi: 10.69987/JACS.2023.30804.

[44] Z. Zhong, M. Zheng, H. Mai, J. Zhao, and X. Liu, “Cancer image classification based on DenseNet model,” Journal of Physics: Conference Series, vol. 1651, no. 1, p. 012143, 2020.

[45] Jubin Zhang, “Interpretable Skill Prioritization for Volleyball Education via Team-Stat Modeling”, JACS, vol. 3, no. 3, pp. 34–49, Mar. 2023, doi: 10.69987/JACS.2023.30304.

[46] Z. Wen, R. Zhang, and C. Wang, “Optimization of bi-directional gated loop cell based on multi-head attention mechanism for SSD health state classification model,” in Proceedings of the 2025 6th International Conference on Electronic Communication and Artificial Intelligence (ICECAI). IEEE, 2025, p. 5.

[47] C. Wang, Z. Wen, R. Zhang, P. Xu, and Y. Jiang, “GPU memory requirement prediction for deep learning task based on bidirectional gated recurrent unit optimization transformer,” in Proceedings of the 2025 5th International Conference on Artificial Intelligence, Virtual Reality and Visualization (AIVRV 2025), Oct. 2025.

[48] R. Zhang, Z. Wen, C. Wang, C. Tang, P. Xu, and Y. Jiang, “Quality analysis and evaluation prediction of RAG retrieval based on machine learning algorithms,” arXiv preprint arXiv:2511.19481, Nov. 2025.

[49] Hanqi Zhang, “Counterfactual Learning-to-Rank for Ads: Off-Policy Evaluation on the Open Bandit Dataset”, JACS, vol. 5, no. 12, pp. 1–11, Dec. 2025, doi: 10.69987/JACS.2025.51201.