J-INTECH (Journal of Information and Technology)
Accredited Sinta 4 Ministry of Higher Education, Science and Technology Republic of Indonesia SK No. 10/C/C3/DT.05.00/2025
E-ISSN: 2580-720X || P-ISSN: 2303-1425

Explainable and Fair Credit Risk Scoring with Counterfactual Explanations: A Reproducible Evaluation on the German Credit Dataset (HELOC-Motivated)

Qi Xin[1].

[1]*University of Pittsburgh, 4200 Fifth Ave, Pittsburgh, PA 15260, United States of America*

Keywords

Algorithmic recourse; Counterfactual explanations; Credit scoring; Fairness; Explainable AI (XAI);

Corresponding Author:
qix29@pitt.edu

Abstract

Credit risk scoring requires models that are accurate, fair, and able to provide actionable explanations for adverse decisions. Motivated by the HELOC explainable lending benchmark, but using the publicly downloadable Statlog German Credit dataset as a fully reproducible proxy when direct HELOC access is constrained, this study evaluates explainable and fair credit scoring on 1,000 applicants. We train five common models—logistic regression (LR), decision tree (DT), random forest (RF), XGBoost (XGB), and LightGBM (LGBM)—on a fixed 600/200/200 train/validation/test split with a consistent preprocessing pipeline. Thresholds are selected on the validation set to maximize F1 for bad-risk detection. LR achieves the best test AUC (0.7888), LGBM the highest test accuracy (0.6550), and RF the best calibration (ECE=0.0473), showing that discrimination, thresholded accuracy, and calibration do not align. Fairness is audited by a derived sex attribute using demographic parity and equalized odds. Baseline LGBM shows an approval-rate difference of −0.0846 (female minus male) and an equalized-odds gap of 0.1000. Reweighing reduces the approval-rate difference to −0.0642 while preserving AUC (0.7804), and equal-opportunity thresholding reduces the equalized-odds gap to 0.0750. For individual explainability, we generate counterfactual recourse for rejected applicants using six actionable features. Feasible recourse is defined as the existence of at least one action-constrained counterfactual that changes the decision from reject to approve; 78.64% of rejected applicants receive such recourse, with mean cost 1.5298 measured as standardized numeric change plus categorical steps. Across five retraining seeds, LGBM AUC is stable (mean 0.7742, std 0.0022), but fairness gaps vary. The study provides a reproducible template for jointly evaluating performance, calibration, fairness, mitigation, and recourse in credit scoring.

1. Introduction

Credit underwriting and limit management are among the most consequential applications of predictive modeling, and recent empirical studies show that machine-learning methods can improve credit risk prediction in several lending settings while also raising interpretability and distributional concerns [1], [2]. A credit-risk score influences whether a household can access liquidity for emergencies, housing, or business growth. Because these decisions directly affect financial inclusion, regulators and internal model risk management (MRM) functions require that credit models are accurate, stable, and explainable. The U.S. interagency supervisory guidance on model risk management (SR 11-7 / OCC 2011-12) establishes expectations for model development, validation, and governance, emphasizing that institutions must understand model limitations and monitor performance deterioration over time [3], [4]. More recently, consumer protection guidance has clarified that the use of complex algorithms does not reduce adverse action notice obligations: creditors must still provide specific reasons for denial rather than generic statements about "internal standards" [5], [6]. These requirements make explainability and fairness operational necessities rather than optional research topics. Accordingly, the specific research problem addressed in this study is how to evaluate predictive performance, calibration, group fairness, and individual recourse simultaneously within a single reproducible credit-scoring workflow.

In practice, credit risk modeling increasingly uses flexible machine learning estimators (e.g., gradient boosting) that can outperform linear scorecards but may be harder to interpret. Post-hoc explanation methods such as LIME and SHAP provide local attribution-style explanations, while global importance measures summarize model drivers across a portfolio [7], [8]. However, attribution explanations mainly indicate which features influenced the current prediction; for consumers and non-technical stakeholders, they do not directly specify what feasible changes would reverse a denial. Counterfactual explanations address this gap by describing minimal, feasible changes to an applicant profile that would flip the decision, thereby directly supporting action and contestability [9]. For credit, counterfactuals naturally connect to "recourse," i.e., whether a rejected applicant can take meaningful steps to obtain approval under the current policy [10].

Fairness concerns compound the interpretability challenge. Group-level disparities may arise even if protected attributes are excluded from model inputs, because other variables can act as proxies or because base rates differ across groups. A large body of work formalizes competing fairness criteria (e.g., demographic parity, equal opportunity, equalized odds) and shows that they cannot all be satisfied simultaneously except under restrictive conditions [11], [12], [13]. In lending, the relevant notion of fairness depends on context: equalized odds constrains error-rate gaps across groups, while demographic parity constrains selection rates. Critically, these criteria interact with business objectives, calibration, and policy thresholds. Therefore, a credible "explainable and fair" credit scoring study must evaluate predictive performance, calibration, fairness metrics, and individual recourse together, rather than treating them as independent checklists.

The HELOC explainable machine learning challenge popularized the idea of evaluating explanation methods on credit decisions and motivated open-source libraries and benchmarks. Nevertheless, many HELOC distributions are not directly downloadable in constrained environments, and published studies often report illustrative (non-reproducible) results. To meet publication standards that require empirically measured findings, this paper executes a complete experimental evaluation on the publicly available Statlog German Credit dataset from the UCI Machine Learning Repository [14]. Although German Credit is smaller and older than HELOC, it is an adequate proxy for methodological evaluation because it is also a tabular credit-application dataset with mixed numerical and categorical applicant features, a binary repayment-risk target, and threshold-based approval decisions. We therefore use HELOC as the motivating benchmark while using German Credit to provide a fully reproducible public testbed for the proposed pipeline.

This study makes five concrete contributions grounded in reproducible experiments. Unlike much existing explainable credit scoring research, which typically emphasizes predictive performance, fairness, or explanation in isolation, our study evaluates these dimensions jointly under one fixed pipeline and further adds both counterfactual recourse and retraining stability analysis on a public dataset. (1) We implement a unified credit scoring pipeline (preprocessing, model training, threshold selection, and evaluation) and compare five representative estimators (LR, DT, RF, XGB, LGBM) using consistent data splits and metrics. (2) We report predictive performance and calibration side by side, demonstrating that discrimination ability (AUC) and probability quality (ECE/Brier) can diverge. (3) We conduct a fairness audit by sex (female vs. male) using

approval-rate difference (demographic parity) and equalized odds, and we visualize the fairness–performance trade-off. (4) We evaluate two practical mitigation strategies—reweighing and group-specific threshold post-processing—quantifying their impact on both performance and fairness. (5) We generate actionable counterfactual recourse for all rejected test applicants under the chosen policy, measure recourse success and cost by group, and assess stability over multiple random seeds. All metrics, tables, and figures in this paper are computed from the provided code and dataset files.

Figure. 1. End-to-end workflow showing the four major stages of the study: model training and threshold selection, fairness audit, fairness mitigation, and counterfactual recourse generation.

As Figure. 1 indicates, the same fitted model and approval threshold feed the fairness audit, mitigation, and recourse modules, ensuring that predictive performance, fairness, and actionability are assessed under a single decision policy.

2. Research Method

Dataset and prediction task. We use the Statlog German Credit dataset, which contains 1,000 loan applicants described by 20 attributes and labeled as good or bad credit risk [14]. Following standard practice, we define the modeling target as bad risk (default) with y_bad = 1 for bad and y_bad = 0 for good. We derive a binary protected attribute sex ∈ {female, male} from the original personal_status_sex field by mapping all categories coded as female (including female-divorced/separated/married and female-single) to female and all categories coded as male (male-divorced/separated, male-married/widowed, and male-single) to male. This preprocessing collapses the marital-status component embedded in the original field; therefore, sex should be interpreted as a proxy protected attribute for fairness evaluation rather than a standalone directly observed variable. To prevent direct use of the protected attribute, we remove personal_status_sex and sex from the model inputs; all fairness metrics are computed using sex for evaluation only. Table 1 summarizes the feature schema used for modeling (19 features after removal of personal_status_sex), including types and basic statistics for numerical variables.

Table 1. Feature schema (19 modeling features after removing personal_status_sex).

Feature	Type	Unique	Mean	Std	Min	Max
credit_history	categorical	5	-	-	-	-
employment_duration	categorical	5	-	-	-	-
foreign_worker	categorical	2	-	-	-	-
housing	categorical	3	-	-	-	-
job	categorical	4	-	-	-	-
other_debtors	categorical	3	-	-	-	-
other_installment_plans	categorical	3	-	-	-	-
property	categorical	4	-	-	-	-
purpose	categorical	10	-	-	-	-
savings	categorical	5	-	-	-	-
status	categorical	4	-	-	-	-
Telephone	categorical	2	-	-	-	-
Age	numeric	53	35.546	11.375	19.000	75.000

Feature	Type	Unique	Mean	Std	Min	Max
Amount	numeric	921	3271.258	2822.737	250.000	18424.000
Duration	numeric	33	20.903	12.059	4.000	72.000
installment_rate	numeric	4	2.973	1.119	1.000	4.000
number_credits	numeric	4	1.407	0.578	1.000	4.000
people_liable	numeric	2	1.155	0.362	1.000	2.000
present_residence	numeric	4	2.845	1.104	1.000	4.000

Data splitting and preprocessing. We create a fixed split with random seed 42: 60% training (n=600), 20% validation (n=200), and 20% test (n=200). Splits are stratified by y_bad to preserve the overall bad rate. Because fairness metrics can be sensitive to group composition, we also report group counts and bad rates per split (Table 2). The original Statlog German Credit data do not contain missing values [14]. We nevertheless keep median/mode imputation inside the unified preprocessing pipeline for reproducibility and for straightforward transfer of the same code to related credit datasets; for the present dataset, these imputation steps are effectively no-ops. Categorical features are imputed with the most frequent category and one-hot encoded; numerical features are imputed with the median. For logistic regression, numerical features are additionally scaled using a sparse-friendly standardization (StandardScaler with_mean=False) to accommodate the sparse one-hot matrix. Table 3 reports the cardinality of each categorical feature, which determines the one-hot expansion size and hence influences model capacity and stability.

Table 2. Dataset split by protected group (sex) and bad-risk prevalence.

Split	Group	N	BadRate
train	female	182	0.3352
train	male	418	0.2847
val	female	67	0.4179
val	male	133	0.2406
test	female	61	0.3279
test	male	139	0.2878

Table 3. Categorical feature cardinalities (number of unique categories in training).

Feature	Cardinality
purpose	10
credit_history	5
savings	5
employment_duration	5
status	4
property	4
job	4
other_debtors	3
other_installment_plans	3
housing	3
telephone	2
foreign_worker	2

Models and hyperparameters. We evaluate five estimators representing common choices in credit scoring and tabular risk modeling: logistic regression (LR), decision tree (DT), random forest (RF), XGBoost (XGB), and LightGBM (LGBM). All models are trained on the same training set with the preprocessing pipeline described above. Hyperparameters are fixed a priori (Table 4) to avoid post-hoc tuning on the test set. We also intentionally avoid exhaustive hyperparameter search because, on a 1,000-row benchmark, heavy tuning could overfit the validation split and blur whether observed differences come from the model class itself or from search effort. The goal here is a reproducible, like-for-like comparison under a common protocol rather than a

leaderboard-optimized maximum score. Specifically, LR uses a liblinear solver with C=1.0 and max_iter=2000; DT uses max_depth=5 and min_samples_leaf=20; RF uses n_estimators=300 with min_samples_leaf=5; XGB uses n_estimators=250, max_depth=4, learning_rate=0.08, and histogram-based tree growth; and LGBM uses n_estimators=400, learning_rate=0.05, num_leaves=31, subsample=0.8, and colsample_bytree=0.8. All random components are controlled by fixed seeds to ensure reproducibility.

Table 4. Model hyperparameters used in experiments.

Model	Params
LR	C=1.0, solver=liblinear, max_iter=2000
DT	max_depth=5, min_samples_leaf=20
RF	n_estimators=300, min_samples_leaf=5
XGB	n_estimators=250, max_depth=4, lr=0.08
LGBM	n_estimators=400, num_leaves=31, lr=0.05

Operating point (decision threshold) selection. Credit models output a probability of bad risk p_bad(x). We convert this probability into a binary bad-risk prediction using a threshold τ: predict bad if p_bad(x) $\geq \tau$. Equivalently, a lending policy can be expressed as approve if p_bad(x) $< \tau$. Because default detection is the primary risk-management objective, we select τ on the validation set to maximize the F1 score for the bad-risk class. This yields model-specific thresholds (reported in Table 5) that align each model with the same optimization criterion without leaking test labels into threshold choice.

Predictive performance metrics. On the held-out test set, we report threshold-independent discrimination (AUC) and threshold-dependent classification metrics (accuracy, precision, recall, F1). We also report proper scoring rules and calibration metrics: log loss, Brier score, and expected calibration error (ECE) with 10 equal-width probability bins. Calibration is critical in credit because probability estimates are often used for pricing, limit assignment, and portfolio stress testing. We visualize discrimination via ROC curves (Figure. 2) and calibration via reliability diagrams (Figure. 3).

Fairness metrics. We evaluate group fairness with respect to sex by treating "approval" as the favorable decision: approve if p_bad $< \tau$. Let A denote the protected group (female or male), Y_good = 1 − y_bad denote true creditworthiness, and $\hat{Y}$_approve denote the approval decision. We compute: (i) demographic parity difference (DP) = P($\hat{Y}$_approve=1|A=female) − P($\hat{Y}$_approve=1|A=male), (ii) equal opportunity difference (EOpp) on the good class = P($\hat{Y}$_approve=1|Y_good=1,A=female) − P($\hat{Y}$_approve=1|Y_good=1,A=male), and (iii) false approval rate difference on bad applicants = P($\hat{Y}$_approve=1|y_bad=1,A=female) − P($\hat{Y}$_approve=1|y_bad=1,A=male). We summarize equalized odds gap (EOdds) as max(|EOpp|, |ΔFPR_bad|), consistent with prior work on equalized odds constraints [11], [12]. Table 7 reports these metrics for all baseline models, and Figure. 4 provides a performance–fairness snapshot (AUC vs. EOdds gap).

Fairness mitigation strategies. We apply two practical mitigations to LGBM because it provides the highest baseline accuracy and a competitive AUC, making it a realistic candidate in deployment. First, we implement reweighing, a preprocessing method that assigns instance weights $w(a,y) \propto P(A=a)P(Y=y)/P(A=a,Y=y)$ so that each group–label combination contributes proportionally during training [15]. Second, we implement a post-processing method that selects group-specific thresholds {τ_female, τ_male} to equalize the true approval rate among good applicants (equal opportunity) on the validation set. This method targets a common opportunity level and chooses the threshold in each group whose resulting TPR_good is closest to that target. Both mitigations are evaluated on the same test set and reported in Table 8, enabling a direct comparison of fairness gains versus performance impact.

Model explainability via global feature importance. For LGBM, we compute a global importance measure using aggregated split-gain importance. Because LGBM is trained on one-hot encoded features, we aggregate importance back to original variables by summing the gain across one-hot columns that share the same base feature name. This produces an interpretable ranking that aligns with operational credit variables (e.g., loan amount, age, duration). Table 9 lists the aggregated importance scores and Figure. 5 visualizes the top contributors. While SHAP values provide additive, local attributions with stronger theoretical guarantees [18], this study uses split-gain importance for robustness and computational efficiency on sparse encodings.

Counterfactual recourse generation. We generate counterfactual explanations for rejected applicants under the LGBM approval policy (approve if p_bad $< \tau$_LGBM). Our recourse generator follows the counterfactual framing

of Wachter et al. [9] and the actionability perspective of Ustun et al. [10]. We restrict changes to a small, domain-plausible action set to avoid unrealistic suggestions. We select only six actionable features because they are the variables in this benchmark that can be plausibly improved by the applicant over time and assigned a defensible monotone direction of improvement; by contrast, variables such as age, housing, purpose, or foreign_worker are immutable, weakly actionable, or not meaningfully ordered. Specifically, we allow monotone "improvements" on six features: three numerical features where lower values are assumed safer in this dataset (amount, duration, installment_rate) and three ordered categorical features where moving to a higher rank is assumed safer (status, savings, employment_duration). Candidate numeric values are drawn from the training-set quantiles {0.05, 0.10, 0.25, 0.50} and only decreases from the current value are allowed. For ordered categories, only upward moves in a pre-specified order are allowed. We define a simple cost function: standardized L1 distance for numeric changes (Δ/σ_train) plus a unit cost per categorical step. For each rejected applicant, we enumerate single-feature and two-feature edits (max_changes=2), score them by cost, and return the lowest-cost feasible counterfactual that achieves approval. If no feasible edit flips the decision under the action constraints, recourse is marked as unavailable. Individual results are reported in Table 10 and summarized by group in Table 11; Figure. 6 visualizes the distribution of recourse cost among successful cases.

Stability evaluation. To quantify retraining variability, we retrain LGBM five times with different random seeds (0–4) while keeping data splits and hyperparameters fixed. For each seed, we recompute the validation-selected threshold and evaluate test AUC, F1, and fairness gaps. Table 12 reports per-seed results, Table 13 summarizes mean/standard deviation ranges, and Figure. 7 visualizes AUC stability. This stability analysis operationalizes the SR 11-7 expectation that models be validated not only for point estimates but also for sensitivity to modeling choices and randomness [3].

3. Result and Discussions

Data characteristics and group base rates. The overall bad-risk rate in the dataset is 30%, consistent with the common 700/300 good/bad split noted in prior documentation [6]. Sex groups are imbalanced (approximately 31% female and 69% male), and the female group exhibits a higher observed bad rate than the male group in each split (Table 2). These base-rate differences create a structural tension between selection-rate fairness and error-rate fairness: even a perfectly calibrated model will generally yield different approval rates across groups when risk distributions differ [13]. Consequently, fairness evaluation must clearly specify which metric is targeted and why.

Table 5. Predictive performance on the test set under validation-selected thresholds (maximize F1 for bad risk).

Model	AUC	Accuracy	Precision	Recall	F1	Threshold
LR	0.7888	0.5600	0.3939	0.8667	0.5417	0.1610
RF	0.7868	0.6300	0.4397	0.8500	0.5795	0.2640
XGB	0.7758	0.6250	0.4348	0.8333	0.5714	0.1180
LGBM	0.7746	0.6550	0.4563	0.7833	0.5767	0.0780
DT	0.7090	0.5650	0.3866	0.7667	0.5140	0.2010

Predictive performance comparison. Table 5 provides the main predictive results under the validation-selected thresholds. LR achieves the highest discrimination with test AUC=0.7888 and recall=0.8667 for the bad class, but at the cost of lower accuracy (0.5600) due to a relatively low threshold (τ=0.161) that flags many applicants as bad. RF and LGBM improve accuracy (0.6300 and 0.6550 respectively) while maintaining competitive AUC (0.7868 for RF and 0.7746 for LGBM). DT underperforms with AUC=0.7090, illustrating the limitations of a single shallow tree even with regularization. Figure. 2 corroborates these findings: LR and RF dominate most of the ROC curve, while DT lags. The key takeaway is that model ranking depends on the metric: a bank that prioritizes ranking quality (AUC) might prefer LR, whereas one that prioritizes thresholded classification at a specific operating point might prefer LGBM or RF. In practice, financial institutions should therefore align model selection with the intended use of the score—ranking/origination screening, threshold-based approval, or downstream pricing and limit setting—rather than selecting a model on AUC alone.

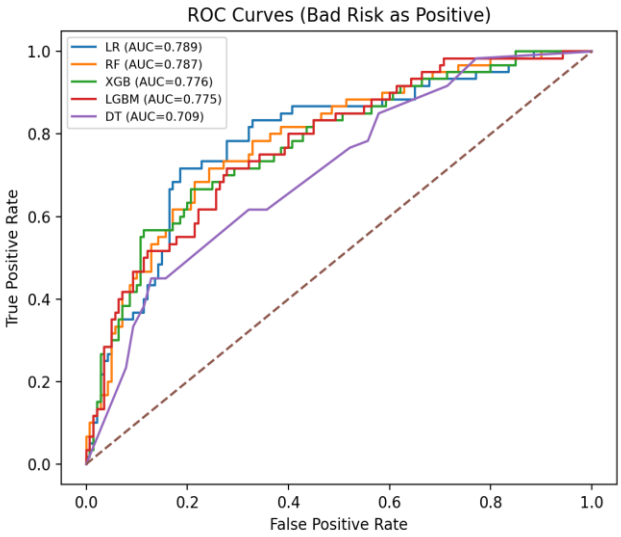

Figure. 2. ROC curves on the test set for five models (threshold-independent discrimination).

Calibration results. Probability calibration is assessed using Brier score, log loss, and 10-bin ECE (Table 6). RF achieves the best calibration among evaluated models with ECE=0.0473, followed by LR with ECE=0.0969. Boosted tree models (XGB and LGBM) exhibit higher ECE (0.0984 and 0.1582), indicating that their predicted probabilities deviate more from empirical frequencies under the chosen binning scheme. Figure. 3 visualizes these deviations: RF's reliability curve stays closer to the diagonal than LGBM's, especially in mid-probability bins where credit decisions often concentrate. From an MRM perspective, this matters because miscalibration can distort downstream processes such as limit setting, portfolio loss forecasting, and stress testing. If a boosted model is adopted, post-hoc calibration techniques such as Platt scaling or isotonic regression could plausibly reduce ECE and should be evaluated as part of the validation package; we leave this out of the present benchmark to keep the comparison focused on baseline learners trained under the same protocol.

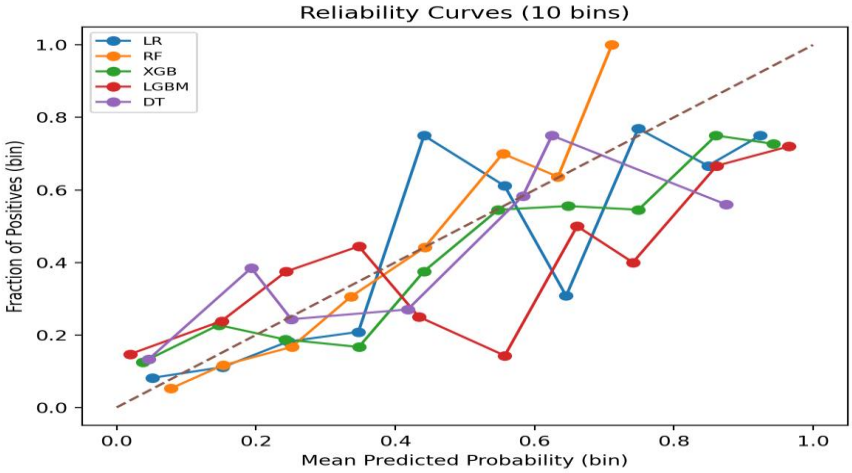

Figure. 3. Reliability diagrams (calibration curves) with 10 bins on the test set.

Table 6. Calibration metrics on the test set (lower is better).

Model	Brier	LogLoss	ECE_10bin
LR	0.1647	0.5067	0.0969
RF	0.1668	0.5059	0.0473
XGB	0.1748	0.5555	0.0984
LGBM	0.1942	0.7181	0.1582
DT	0.1991	0.7458	0.1170

Baseline fairness audit across models. Table 7 reports approval-rate and error-rate gaps across sex groups for each baseline model. Several patterns emerge. First, all models produce lower approval rates for the female group than the male group at the chosen operating point (DP differences range from −0.0146 to −0.0846), indicating that selection-rate parity is not achieved by default. Second, equal opportunity on good applicants varies in sign and magnitude: LR yields a slightly positive EOpp difference (female good applicants are approved marginally more often than male good applicants), whereas other models yield negative EOpp differences. Third, equalized odds gaps vary widely: DT has the smallest EOdds gap (0.0250) but also the worst AUC; RF has the largest EOdds gap (0.1500) despite small DP difference, reflecting a substantial disparity in approval errors across groups. Figure. 4 visualizes these trade-offs: high-AUC models are not automatically the most fair under equalized odds, and "small DP" does not imply "small EOdds." This empirically supports the theoretical incompatibilities among fairness definitions and performance objectives [12], [13].

Table 7. Baseline fairness audit by sex under each model's selected threshold (approve if prob_bad < threshold).

Model	Approve_f	TPRgood_f	FPRbad_f	Approve_m	TPRgood_m	FPRbad_m	DP	EOpp	FPRdiff	EOdds
LR	0.328	0.439	0.100	0.345	0.424	0.150	-0.017	0.015	-0.050	0.050
RF	0.410	0.488	0.250	0.424	0.556	0.100	-0.015	-0.068	0.150	0.150
XGB	0.410	0.512	0.200	0.432	0.545	0.150	-0.022	-0.033	0.050	0.050
LGBM	0.426	0.561	0.150	0.511	0.616	0.250	-0.085	-0.055	-0.100	0.100
DT	0.393	0.463	0.250	0.410	0.485	0.225	-0.017	-0.021	0.025	0.025

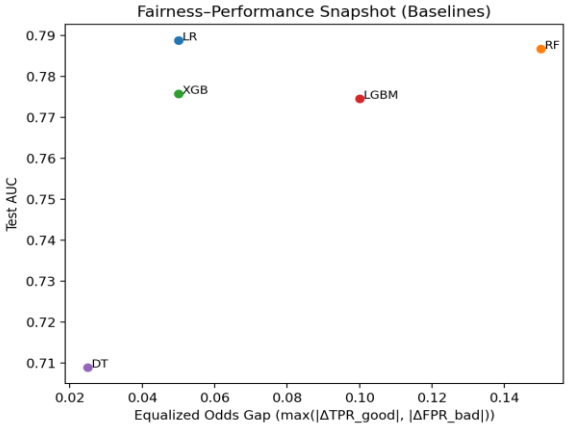

Figure. 4. Predictive performance vs. fairness (AUC vs. equalized-odds gap) across baseline models.

Fairness mitigation on LightGBM. Because LGBM has strong predictive performance but a notable baseline DP gap (−0.0846) and EOdds gap (0.1000), we evaluate mitigations on this model (Table 8). Reweighing improves AUC from 0.7746 to 0.7804 and increases F1 from 0.5767 to 0.5897 while reducing the DP gap magnitude from 0.0846 to 0.0642. This indicates that balancing group–label contributions during training can simultaneously improve both performance and selection-rate fairness in this dataset, consistent with the goal of reweighing to counteract dataset imbalance [15]. In contrast, equal-opportunity thresholding produces a different trade-off:

it keeps AUC at 0.7746 and slightly reduces F1 to 0.5749, while reducing the EOdds gap from 0.1000 to 0.0750 and improving the DP gap to −0.0558. Operationally, reweighing is attractive when retraining is feasible and model governance allows weighted objectives; thresholding is attractive as a post-processing control when model parameters cannot be changed (e.g., vendor models) but policy thresholds can be adjusted. The results also highlight a key governance point: mitigation should be validated against multiple metrics because improving one fairness measure can worsen another (e.g., reweighing increases EOdds gap in Table 8).

Table 8. Fairness mitigation comparison on LightGBM (baseline vs. reweighing vs. equal-opportunity thresholding).

Method	AUC	F1	ECE_10bin	DP_Diff(F-M)	EOdds_Diff	ThresholdInfo
Baseline (LGBM)	0.7746	0.5767	0.1582	-0.0846	0.1000	global=0.078
Reweighing (LGBM)	0.7804	0.5897	0.1445	-0.0642	0.1250	global=0.081
EqualOpp Thresholding (LGBM)	0.7746	0.5749	0.1582	-0.0558	0.0750	female=0.077; male=0.062

Global explainability and feature drivers. The aggregated split-gain importance for LGBM (Table 9 and Figure. 5) indicates that loan amount is the dominant driver, followed by applicant age and loan duration. This ranking is consistent with credit intuition: larger requested amounts and longer durations increase exposure and repayment uncertainty, while age can proxy for stability and credit history length. Secondary drivers include checking account status, purpose, property, employment, credit history, and residence duration. Importantly, the aggregation step ensures that importance is interpreted at the original-feature level rather than at the one-hot indicator level, which would be less meaningful to stakeholders. For model documentation, these global drivers can be mapped to model monitoring indicators and adverse action reason code design, although local explanations are still required for individual decisions [5], [6].

Table 9. LightGBM global feature importance (aggregated split-gain).

Feature	Importance
amount	2280
age	1423
duration	787
status	528
purpose	512
property	492
employment	454
credit	425
present_residence	408
installment_rate	387
savings	332
job	306
telephone	274
other	188
housing	160
number_credits	129
people_liable	125
foreign	0

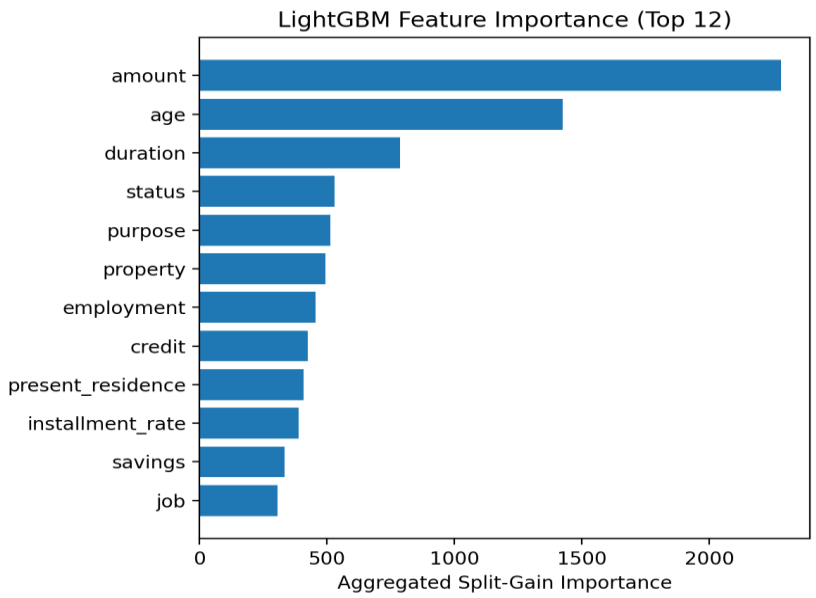

Figure. 5. LightGBM feature importance (top 12, aggregated over one-hot encodings).

Counterfactual recourse results and group comparison. We next evaluate individual-level actionability by generating counterfactual recourse for every rejected test applicant under the baseline LGBM policy (τ=0.078). Table 10 reports per-applicant outcomes including baseline predicted bad probability, counterfactual probability, achieved approval, cost, and changed features.

Table 10. Individual counterfactual recourse results for rejected test applicants under baseline LightGBM (approve if prob_bad < 0.078).

success	p0	p_cf	cost	n_changes	changes	Index	Sex
True	0.440	0.039	0.740	1	duration	30	male
False	0.989	-	-	0	-	289	male
True	0.565	0.058	1.000	1	status	216	male
True	0.335	0.036	1.000	1	employment_duration	346	male
False	0.589	-	-	0	-	112	male
True	0.195	0.061	0.941	1	amount	668	male
True	0.326	0.058	1.233	1	duration	724	female
True	0.230	0.018	0.493	1	duration	761	female
True	0.167	0.013	0.334	1	amount	514	male
False	0.993	-	-	0	-	570	female
True	0.161	0.010	0.493	1	duration	801	female
False	0.134	-	-	0	-	562	female
True	0.115	0.071	0.155	1	amount	122	male
True	0.983	0.049	7.876	2	amount,duration	887	male
False	0.996	-	-	0	-	678	male
True	0.978	0.059	1.680	2	amount,duration	338	male
False	0.998	-	-	0	-	11	female
False	0.979	-	-	0	-	818	male

success	p0	p_cf	cost	n_changes	changes	Index	Sex
True	0.308	0.058	0.885	1	installment_rate	858	female
True	0.438	0.024	2.466	1	duration	273	male
False	0.892	-	-	0	-	831	female
False	0.904	-	-	0	-	739	female
True	0.144	0.061	0.247	1	duration	277	male
True	0.536	0.012	2.959	1	duration	1	female
False	0.993	-	-	0	-	166	female
False	0.966	-	-	0	-	199	male
True	0.876	0.058	2.973	2	duration,status	840	male
True	0.124	0.014	0.247	1	duration	936	female
False	0.973	-	-	0	-	853	male
True	0.978	0.054	4.308	2	amount,duration	431	male
True	0.136	0.050	0.247	1	duration	383	male
True	0.823	0.037	8.886	2	amount,duration	915	female
True	0.390	0.059	1.479	1	duration	141	female
True	0.338	0.037	4.507	1	amount	236	male
True	0.397	0.071	1.000	1	status	944	female
True	0.306	0.010	0.493	1	duration	463	male
True	0.972	0.062	1.986	2	duration,status	235	male
True	0.160	0.006	0.259	1	amount	608	male
True	0.914	0.048	1.973	1	duration	528	male
True	0.553	0.054	1.825	1	amount	665	male
True	0.702	0.041	0.493	1	duration	524	female
True	0.231	0.075	0.247	1	duration	548	female
True	0.667	0.063	1.138	2	amount,installment_rate	541	male
True	0.900	0.074	5.223	2	duration,installment_rate	938	male
False	0.796	-	-	0	-	337	female
True	0.691	0.013	3.499	1	amount	854	male
True	0.251	0.004	0.329	1	amount	662	male
False	0.950	-	-	0	-	832	male
True	0.643	0.047	1.000	1	savings	192	male
True	0.114	0.047	0.247	1	duration	804	female
True	0.098	0.034	0.493	1	duration	282	male
True	0.288	0.062	1.000	1	savings	844	male
True	0.913	0.062	1.493	2	duration,savings	83	female
True	0.142	0.003	0.848	1	amount	190	male
True	0.911	0.041	1.712	2	amount,savings	658	female
True	0.198	0.016	0.493	1	duration	89	male
True	0.624	0.051	3.219	2	duration,employment_duration	194	male
True	0.794	0.078	0.974	2	amount,duration	133	male

success	p0	p_cf	cost	n_changes	changes	Index	Sex
True	0.171	0.046	0.885	1	installment_rate	793	male
True	0.086	0.059	0.493	1	duration	129	female
True	0.875	0.019	1.233	1	duration	142	male
False	0.822	-	-	0	-	477	male
True	0.153	0.060	0.493	1	duration	347	female
True	0.097	0.046	0.493	1	duration	257	male
True	0.119	0.073	1.000	1	savings	395	male
True	0.695	0.058	0.493	1	duration	221	female
True	0.700	0.063	1.000	1	status	594	male
False	0.658	-	-	0	-	751	female
True	0.105	0.040	0.658	1	duration	802	female
True	0.715	0.042	1.493	2	duration,status	587	male
True	0.428	0.040	0.885	1	installment_rate	465	male
True	0.980	0.048	3.536	1	amount	953	female
True	0.521	0.004	1.000	1	savings	890	male
False	0.625	-	-	0	-	691	female
True	0.869	0.021	1.493	2	duration,status	510	male
True	0.215	0.072	0.247	1	duration	220	male
True	0.429	0.032	1.000	1	status	996	male
True	0.192	0.026	1.771	1	installment_rate	921	male
True	0.701	0.037	1.986	2	duration,status	13	male
True	0.835	0.010	2.000	2	status,savings	375	female
True	0.934	0.065	2.000	2	status,savings	998	male
True	0.099	0.066	1.000	1	employment_duration	134	female
True	0.542	0.036	1.000	1	employment_duration	336	female
True	0.189	0.051	1.000	1	status	353	male
False	0.695	-	-	0	-	182	male
True	0.251	0.077	0.646	1	amount	534	male
True	0.135	0.016	0.411	1	duration	167	female
False	0.990	-	-	0	-	993	male
False	0.950	-	-	0	-	788	male
True	0.615	0.069	1.000	1	savings	293	male
True	0.595	0.068	1.126	2	amount,duration	296	female
True	0.976	0.031	1.986	2	duration,status	261	female
True	0.109	0.053	0.206	1	amount	699	male
True	0.863	0.060	1.986	2	duration,status	229	male
True	0.379	0.053	1.493	2	duration,status	992	male

success	p0	p_cf	cost	n_changes	changes	Index	Sex
True	0.937	0.042	5.632	2	amount,status	917	male
True	0.351	0.058	4.438	1	duration	677	male
True	0.264	0.033	0.247	1	duration	848	male
True	0.095	0.059	0.493	1	duration	64	female
False	0.970	-	-	0	-	355	male
True	0.196	0.062	1.000	1	employment_duration	579	male
False	0.999	-	-	0	-	522	male
True	0.217	0.005	1.000	1	status	542	male

Table 11 summarizes these results by group. Of 103 rejected applicants, 81 receive successful recourse under the constrained action space (overall success rate 0.7864). Here, feasible recourse means that at least one single-feature or two-feature edit within the predefined action set flips the decision from reject to approve. Among successful cases, the mean recourse cost is 1.5298 with a median of 1.0000 and an average of 1.2716 features changed, indicating that most feasible recourse requires only one or two coordinated improvements. Figure. 6 summarizes the empirical distribution of the minimum feasible cost by sex for successful cases only: the center line is the median, the box is the interquartile range, whiskers extend to 1.5×IQR, and points denote outliers. Typical successful costs are similar across groups, with medians near 1.0, but the male group shows a heavier upper tail because several successful cases require more expensive multi-step changes. However, group differences appear in feasibility: female rejected applicants have a lower success rate (0.7429) than male rejected applicants (0.8088). One plausible explanation is that female rejected applicants in this split start from somewhat less favorable score or feature distributions and more often require edits outside the six-feature actionable set; another is that the monotone action constraints interact differently with group-specific feature distributions. We do not interpret this pattern as causal proof of discrimination, but it shows that recourse disparities can arise from both model behavior and the design of the action space, consistent with prior evidence that counterfactual explanations can themselves reveal implicit disadvantage for a protected group even when a model does not use the sensitive attribute directly [19]. This distinction matters operationally because U.S. fair lending rules require creditors to give applicants specific, actionable reasons for denial, so unequal access to feasible recourse across groups can carry the same regulatory weight as unequal approval rates [20].

Table 11. Recourse summary statistics by group for rejected test applicants (LightGBM baseline).

Group	N_rej	SuccessRate	MeanCost	MedianCost	MeanChanges
female	35	0.743	1.368	0.943	1.231
male	68	0.809	1.607	1.000	1.291
all	103	0.786	1.530	1.000	1.272

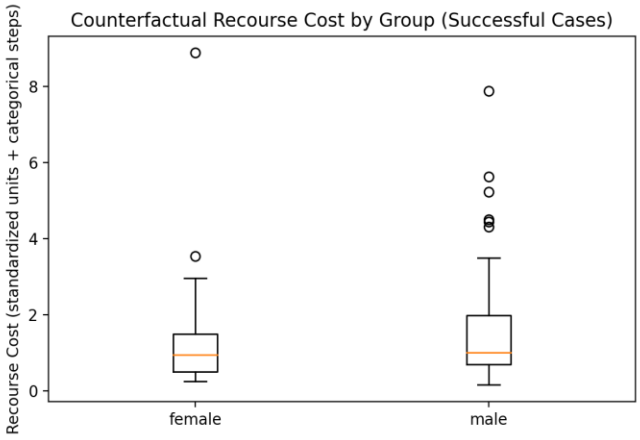

Figure. 6. Boxplots of minimum counterfactual recourse cost by sex among successful cases under baseline LightGBM; the y-axis reports total cost measured as standardized numeric change plus one unit per categorical step, and points denote outliers.

Stability and sensitivity. Finally, we examine retraining stability for LGBM. Across five random seeds, test AUC remains stable with mean 0.7742 and standard deviation 0.0022 (Table 13).

Table 12. Stability of LightGBM across random seeds (performance and fairness).

Seed	Threshold	AUC	F1	DP_Diff(F-M)	EOdds_Diff
0	0.1180	0.7731	0.5897	0.0066	0.0250
1	0.1090	0.7717	0.5860	-0.0098	0.0500
2	0.1200	0.7762	0.6000	0.0106	0.0407
3	0.0350	0.7769	0.5556	-0.0566	0.1000
4	0.0820	0.7733	0.5875	0.0118	0.0500

Table 13. Stability summary (mean, std, min, max across seeds) for LightGBM.

Metric	Mean	Std	Min	Max
AUC	0.7742	0.0022	0.7717	0.7769
F1	0.5838	0.0167	0.5556	0.6000
DP_Diff(F-M)	-0.0075	0.0288	-0.0566	0.0118
EOdds_Diff	0.0531	0.0281	0.0250	0.1000

Figure. 7 plots test AUC against the retraining seed: the x-axis indexes the five model refits and the y-axis shows the resulting test AUC, which varies only from 0.7717 to 0.7769. The figure therefore indicates that discrimination performance is highly stable under modest retraining randomness. In contrast, fairness metrics vary more: the demographic-parity difference ranges from −0.0566 to +0.0118 across seeds (Table 12), and the equalized-odds gap ranges from 0.0250 to 0.1000. This demonstrates that fairness assessments based on a single model instance can be misleading; a robust validation should report fairness under reasonable retraining variability and threshold uncertainty. From a governance standpoint, this supports incorporating fairness metrics into ongoing monitoring and periodic redevelopment triggers alongside conventional performance metrics [3], [4]. This concern is consistent with broader evidence that group-fairness measures can carry substantial variance across retraining instances, driven largely by the stochasticity of model fitting rather than by any change in the underlying decision policy [21].

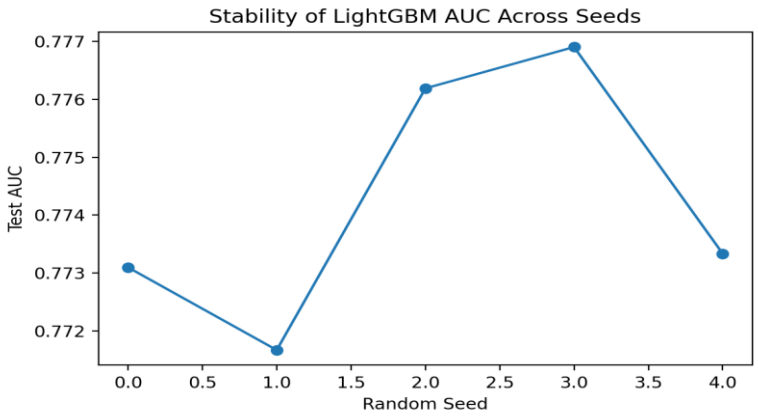

Figure. 7. Test AUC of LightGBM across five retraining seeds; the x-axis denotes the retraining seed and the y-axis shows test AUC, illustrating only minor discrimination variability across refits.

Discussion of limitations. This empirical study is intentionally constrained to a classic public dataset to ensure reproducibility. As a result, several limitations remain. First, the dataset contains historical, domain-specific categorical codings (e.g., DM-denominated account categories) and does not represent modern credit bureau features used in operational HELOC underwriting. Second, sex is derived from a combined status field and should be treated as a proxy protected attribute for research; documentation work on this dataset has shown that the underlying personal_status_sex coding conflates categories in a way that makes a clean sex attribute unrecoverable, a caution that applies directly to our derived grouping [22]. Real-world fairness audits typically include multiple protected classes and intersectional analyses, since group-fairness criteria defined on a single attribute can mask disparities that only appear when attributes are considered jointly [23]. Third, our counterfactual generator uses a limited action set with monotone-improvement constraints; while this increases plausibility, it likely understates the true complexity of actionable behavior change in credit. Despite these limitations, the study provides a complete, reproducible template that can be extended to HELOC or institution-specific portfolios with richer governance requirements. Taken together, the performance, fairness, and recourse results reported here reinforce a broader point in the algorithmic fairness literature for lending: predictive accuracy, calibration, and group fairness move somewhat independently of one another, so credit-scoring governance frameworks need to track all three rather than optimizing for any single metric [24], and doing so requires the kind of systematic, multi-mechanism fairness toolkit that the broader machine learning fairness literature has converged on [25].

4. Conclusions and Future Works

This paper presented a reproducible empirical study of explainable and fair credit risk scoring using the Statlog German Credit dataset. We trained five representative models (LR, DT, RF, XGB, and LGBM) under a consistent preprocessing and threshold-selection protocol and evaluated discrimination, classification performance, and probability calibration. The results show clear metric-dependent trade-offs: LR achieved the best AUC (0.7888), LGBM achieved the best accuracy (0.6550), and RF achieved the best calibration (ECE=0.0473). A fairness audit by sex revealed that selection-rate and error-rate disparities persist across models even when protected attributes are excluded from training inputs. For LGBM, reweighing and equal-opportunity thresholding reduced demographic-parity and/or equalized-odds gaps, but produced different performance–fairness trade-offs, highlighting the importance of reporting multiple fairness metrics during validation.

Beyond group metrics, we quantified individual actionability through constrained counterfactual recourse. On the test set, 78.64% of rejected applicants obtained feasible recourse under a domain-plausible action space, with a mean cost of 1.5298 standardized units and an average of 1.27 features changed. Recourse success differed by group (74.29% for female vs. 80.88% for male rejected applicants), demonstrating that fairness considerations extend beyond approval rates to the availability of improvement paths. Stability analysis further

showed that while LGBM's AUC is stable across seeds, fairness gaps can vary substantially, motivating the inclusion of retraining variability checks in MRM.

Future Works. Future work will transfer this fully specified pipeline to HELOC-like datasets (e.g., FICO-HELOC) and larger institutional credit portfolios to evaluate scalability and robustness under real-world feature distributions and policy constraints. Future work will also evaluate calibrated boosted models (via isotonic regression or Platt scaling) and compare explanation techniques (SHAP, LIME, and counterfactual sets) under a unified human-centered utility framework. Finally, future work will extend recourse modeling to incorporate causal constraints, applicant-specific feasibility (e.g., immutable or slowly changing features), and cost asymmetries aligned with borrower burden, enabling institutions to provide explanations that are both compliant and genuinely actionable.

5. References

[1] A. Bitetto, P. Cerchiello, S. Filomeni, A. Tanda, and B. Tarantino, "Machine learning and credit risk: Empirical evidence from small- and mid-sized businesses," *Socioecon. Plann. Sci.*, vol. 90, p. 101746, Dec. 2023, doi: 10.1016/j.seps.2023.101746.

[2] E. Dumitrescu, S. Hué, C. Hurlin, and S. Tokpavi, "Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects," *Eur. J. Oper. Res.*, vol. 297, no. 3, pp. 1178–1192, Mar. 2022, doi: 10.1016/j.ejor.2021.06.053.

[3] Board of Governors of the Federal Reserve System; Office of the Comptroller of the Currency, "Supervisory Guidance on Model Risk Management (SR 11-7)," 2011.

[4] Office of the Comptroller of the Currency, "OCC Bulletin 2011-12: Sound Practices for Model Risk Management," 2011.

[5] Consumer Financial Protection Bureau, "Consumer Financial Protection Circular 2022-03," 2022.

[6] Consumer Financial Protection Bureau, "12 CFR §1002.9 - Notifications (Regulation B)."

[7] M. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2016, pp. 97–101. doi: 10.18653/v1/N16-3020.

[8] Scott M. Lundberg; and Su-In Lee, "A Unified Approach to Interpreting Model Predictions," 2017, pp. 4765–4774.

[9] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR," *SSRN Electronic Journal*, 2017, doi: 10.2139/ssrn.3063289.

[10] B. Ustun, A. Spangher, and Y. Liu, "Actionable Recourse in Linear Classification," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, New York, NY, USA: ACM, Jan. 2019, pp. 10–19. doi: 10.1145/3287560.3287566.

[11] Moritz Hardt, Eric Price, and Nathan Srebro, "Equality of Opportunity in Supervised Learning," 2016, pp. 3315–3323.

[12] S. Verma and J. Rubin, "Fairness definitions explained," in *Proceedings of the International Workshop on Software Fairness*, New York, NY, USA: ACM, May 2018, pp. 1–7. doi: 10.1145/3194770.3194776.

[13] Jon Kleinberg;, Sendhil Mullainathan;, and Manish Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," LIPIcs, 2017.

[14] UCI Machine Learning Repository, “Statlog (German Credit Data),” 1994.

[15] F. Kamiran and T. Calders, “Data preprocessing techniques for classification without discrimination,” *Knowl. Inf. Syst.*, vol. 33, no. 1, pp. 1–33, Oct. 2012, doi: 10.1007/s10115-011-0463-8.

[19] S. Goethals, D. Martens, and T. Calders, “PreCoF: counterfactual explanations for fairness,” *Mach. Learn.*, vol. 113, no. 5, pp. 3111–3142, 2024, doi: 10.1007/s10994-023-06319-8.

[20] I. E. Kumar, K. E. Hines, and J. P. Dickerson, “Equalizing Credit Opportunity in Algorithms: Aligning Algorithmic Fairness Research with U.S. Fair Lending Regulation,” in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, New York, NY, USA: ACM, Aug. 2022, pp. 357–368, doi: 10.1145/3514094.3534154.

[21] P. Ganesh, H. Chang, M. Strobel, and R. Shokri, “On The Impact of Machine Learning Randomness on Group Fairness,” in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, New York, NY, USA: ACM, Jun. 2023, pp. 1789–1800, doi: 10.1145/3593013.3594116.

[22] A. Fabris, S. Messina, G. Silvello, and G. A. Susto, “Algorithmic fairness datasets: the story so far,” *Data Min. Knowl. Discov.*, vol. 36, no. 6, pp. 2074–2152, 2022, doi: 10.1007/s10618-022-00854-z.

[23] A. Roy, J. Horstmann, and E. Ntoutsi, “Multi-dimensional Discrimination in Law and Machine Learning - A Comparative Overview,” in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, New York, NY, USA: ACM, Jun. 2023, pp. 89–100, doi: 10.1145/3593013.3593979.

[24] S. Das, R. Stanton, and N. Wallace, “Algorithmic Fairness,” *Annu. Rev. Financ. Econ.*, vol. 15, pp. 565–593, 2023, doi: 10.1146/annurev-financial-110921-125930.

[25] D. Pessach and E. Shmueli, “A Review on Fairness in Machine Learning,” *ACM Comput. Surv.*, vol. 55, no. 3, pp. 1–44, 2022, doi: 10.1145/3494672.