



Outcome Definition Before Model Choice in Credit Risk Machine-Learning Models versus Ratio-Based Scorecards

Saharsh Jangaon

Polygence Research Program · August 2026

Abstract

Machine-learning credit models are usually judged against ratio-based scorecards using ROC-AUC, a measure of ranking alone that is silent on whether the outcome being ranked was defined correctly. This study asks what that comparison overlooks, using 432,248 U.S. Small Business Administration 7(a) loan records — approvals from fiscal 2020 onward, outcomes observed through June 30, 2026 — plus two consumer-credit benchmarks and controlled simulations.

On identical SBA data a five-ratio scorecard reached an AUC of 0.924 against 0.946 for a gradient-boosted model (paired difference 0.022; 95% CI 0.017 to 0.028). The larger gains lie elsewhere: precision-recall AUC rises from 0.151 to 0.412 and expected calibration error falls 45%, from 0.0087 to 0.0048, letting the model fund 94.8% of an approved pool against 87.9% at the same realized loss rate (+6.9 points, CI 5.6 to 8.1).

The larger effect is not the model but the definition of the outcome. Restricting analysis to loans that have reached a terminal status gives an FY2022 charge-off rate of 13.24%; a maturity-matched 36-month design on the same portfolio gives 1.91%. On the same loans the mis-specified model has the higher AUC — 0.954 versus 0.946 — while overstating risk fourfold and, under a 3% decline policy, declining 14,674 borrowers whose loans did not charge off within the window against 2,145. A discrete-time hazard model on 2.03 million loan-quarters varied by only 0.028 in AUC across 12- to 48-month horizons, against 0.597 for separately fitted classifiers whose 12-month AUC was 0.353 (CI 0.268–0.448).

Three further results follow the same pattern: calibration decays under regime shift while ranking holds, in simulation and across real fiscal years; protected-group membership is recoverable at AUC 0.907 yet the simulated approval gap matches an oracle scoring true risk; and simple scorecards, not complex models, produce the more homogeneous decisions. Define the outcome first, monitor calibration rather than ranking alone, and route borderline cases to human review, which recovers 47% of errors on 5% of volume.

Keywords: AI lending, algorithmic fairness, credit risk, discrete-time hazard model, model calibration, outcome-window bias, ratio-based scorecards

1. Introduction

Traditional underwriting relies on a small, standardized set of measures: a bureau score, a debt-to-income ratio, an appraisal, a record of past borrowing. Their virtue is legibility — borrower, regulator, and loan officer read the same number and understand roughly what it



means — and that narrowness is also their weakness, since a bureau score is backward-looking, updates slowly, and may be unavailable for consumers with limited credit histories. Financial technology firms built their underwriting around that gap, and machine learning made wider data usable: transaction records, digital footprints, mobile-phone usage.

How these systems are evaluated has not changed with them: almost any applied credit-scoring paper or industry model card reports the same headline number. AUC is scale-free, threshold-free, and easy to compare across models and datasets, and those conveniences are why it hides so much. Depending only on the order of predicted scores, it is unchanged by errors in the absolute level of predicted risk and by errors in how the outcome was defined. Those two failures are what this paper measures.

The study began from two failure modes in alternative-data lending and identified an issue preceding both. Using real federal loan-level data, we found that the most consequential modelling decision was not the algorithm, the fairness constraint, or the feature set, but how the outcome was defined: which loans count as observed, and how long a loan is watched before being labelled defaulted or not. That choice changed estimated portfolio risk sevenfold and reversed which of two models appeared better on ROC-AUC, the mis-specified model scoring higher while making substantially worse lending decisions.

1.1 Research question

When a lender replaces ratio-based underwriting with a machine-learning credit model, what measurably improves, what measurably breaks, and which design choices repair the breaks?

Everything in this study puts a number on one of those three clauses; four hypotheses organize it.

H1 (Outcome definition). How the outcome variable is defined — which loans are eligible, and the length of the performance window, meaning the fixed period after origination during which a loan is observed before being labelled default or non-default — changes estimated risk, calibration, and resulting accept/reject decisions by more than algorithm choice, while commonly used discrimination metrics such as ROC-AUC may fail to reveal the problem.

H2 (Calibration drift). When the economy moves away from training conditions, ranking remains roughly stable while absolute probability estimates decay. Because approval cutoffs and risk-based pricing consume absolute probabilities, this matters even when the ranking looks fine.

H3 (Complexity is not free accuracy). Flexible models will not meaningfully outperform well-specified logistic regression on small structured credit files.

H4 (Proxy discrimination). Features containing no protected attribute will allow a model to reconstruct protected-group membership, and a model trained on them will produce an approval gap larger than the true difference in repayment rates warrants.

1.2 How the four hypotheses relate

The hypotheses are neither independent nor equally consequential, and their ordering determines how the paper is built. Because H1 concerns the outcome definition on which every other measurement depends, it is tested first, in Section 4.2, on the full SBA portfolio; H2 and H3 are then tested against that corrected outcome, in Sections 4.5 and 4.6.1; H4, which requires knowing true risk, is deferred to the controlled environment of Section 4.4.

Reversing that order would audit fairness and calibration against a target that may itself be mis-defined — the error this paper measures, and the evaluation-stage selection problem Hand and Adams (2014) and Kozodoi et al. (2025) identify.

H1 and H2 are one failure in two settings, absolute probabilities decaying while ranking holds; H3 serves as a control.

H4 is hardest to test, because establishing that a model discriminates beyond what genuine risk implies requires knowing what genuine risk is, which observational data never supplies.

2. Background

The rationale for alternative data predates the technology: Akerlof (1970) showed that asymmetric information can unravel a market, and Stiglitz and Weiss (1981) extended this to credit, where lenders facing hidden borrower risk may ration credit rather than raise rates because a higher rate attracts a worse pool. Alternative data shrinks that gap — Berg et al. (2020) reported an AUC of 0.696 for digital-footprint variables alone against 0.683 for a bureau score, with the combination reaching 0.736, and Björkegren and Grissen (2020) showed mobile-phone usage predicting repayment where borrowers have no formal financial history. The harder question is what else those signals encode.

Three literatures bear on this paper. The first is survival analysis in credit scoring, a mature field established in regulatory stress testing and IFRS 9 provisioning (Bellotti & Crook, 2014; Dirick et al., 2017; Djeundje & Crook, 2019; Medina-Olivares et al., 2023; Wang et al., 2024; Botha & Verster, 2026); the hazard model in Section 4.3 is not offered as a methodological contribution. What is new is the head-to-head measurement of what the windowed alternative costs on the same data, its application to federally guaranteed small-business lending rather than the consumer and mortgage portfolios this literature almost always uses, and the finding that loan term in a windowed model is substantially a proxy for time at risk. The second literature concerns this paper's central variable. The performance window — the period after origination over which a loan is observed before being labelled good or bad — is a standard step in scorecard construction (Siddiqi, 2017); what it does not supply is a measurement of what that choice costs when made badly on a portfolio in which most loans are

still outstanding. This paper supplies that measurement; it does not claim to have discovered the problem. Lessmann et al. (2015), the reference benchmark across eight credit datasets, report differences among well-specified classifiers that are real but small — which is what makes a first-order design choice such as the outcome window worth isolating.

The third literature addresses sample selection. Reject inference asks how to handle applicants who were denied and whose repayment is therefore unobservable (Banasik & Crook, 2007; Hand & Adams, 2014; Kozodoi et al., 2025). The problem here is different: every loan in this analysis was approved and disbursed, and the missing outcomes belong to loans still running. Denial censors by decision; maturity censors by clock. The remedies differ accordingly — reject inference imputes labels for a population never observed, while this problem is solved by not discarding observations that are simply incomplete. Kozodoi et al. (2025) find that correcting selection bias at the evaluation stage matters more than correcting it during training, consistent with Section 4.2.

On fairness, Fuster et al. (2022) found that when U.S. mortgage underwriting moves from scorecards to machine learning, Black and Hispanic borrowers are disproportionately less likely to gain; their decomposition is frequently misreported, since they attribute the widening disparity primarily to the models' greater functional flexibility, with triangulation of excluded characteristics secondary. Bartlett et al. (2022) found FinTech rate disparities similar to non-FinTech lenders on GSE mortgages and lower only for specific FHA products and vintages, narrower than the widely circulated claim that algorithmic lending reduces discrimination generally. Barocas and Selbst (2016) supply the mechanism: a model never given a protected attribute can still discriminate, because protected status is redundantly encoded across the features it does see.

3. Data and Methods

3.1 Data

SBA 7(a) loan-level records. The Small Business Administration publishes loan-level records for its 7(a) guaranteed-lending program under FOIA (U.S. Small Business Administration, 2026). The extract used here covers FY2020 onward, contains 432,248 records and 42 fields, and carries an as-of date of June 30, 2026. Origination fields recorded at approval include loan amount and guarantee, pricing, term, processing method, industry, geography, business type and age, collateral and revolver indicators, and jobs supported; outcome fields are status, charge-off date, and paid-in-full date. The dataset is central because it has what academic benchmarks lack: a dated longitudinal outcome structure spanning a macroeconomic shock. That shock is the COVID-19 pandemic, and the federal response is part of the data-generating process. Under Section 1112 of the CARES Act, SBA made six months of principal, interest, and fee payments on 7(a) loans approved on or before September 27, 2020, and the Economic Aid Act extended a further six months to loans approved between February 1 and September 30, 2021 (CARES Act, 2020; U.S. Small Business Administration, 2021). Nearly every FY2020 training loan therefore had part of its early debt service paid federally, as did the later two-thirds of the FY2021 calibration cohort, while the FY2022 test cohort received none. Charge-off rates in the earlier cohorts are therefore suppressed relative to borrower risk alone,

so any comparison across these years crosses a policy boundary as well as an economic one. Paycheck Protection Program lending, published separately, is not in this file.

Loan status takes one of five values in this vintage: PIF (paid in full), CHGOFF (charged off), CANCLD (canceled), COMMIT (committed), and EXEMPT (exempt). PIF and CHGOFF are terminal — once a loan reaches either, its credit outcome is final. CANCEL and COMMIT never became loans at risk and were excluded. EXEMPT marks a disbursed loan still outstanding with its detailed status withheld: 61.9% of the file, roughly 267,600 records. After these exclusions the FY2020–FY2022 cohorts used throughout contain 123,393 loans under the maturity-matched design and 56,716 under the resolved-only design (Table 3). For most loans the question "did this default?" has no answer yet, and that is the starting point for this analysis.

Consumer benchmarks. The Statlog German Credit file (Hofmann, 1994) and its corrected counterpart, South German Credit (Grömping, 2019), each contain 1,000 loans with a 30% bad rate. They serve as replication benchmarks, not a basis for substantive conclusions, and the analysis uses the UCI original so the corrected version can be compared directly. Grömping documented that eight attributes in the classic file have mislabelled levels — under the original coding, creditworthiness appears to worsen when credit history changes to "all credits paid back duly" — and the UCI page carries no warning about this and does not link to the correction.

Two simulations. Sections 4.4 and 4.5 use a generated lending environment of 30,000 applicants per replication, Appendix A an advisory environment of 30,000 households. Both exist because measuring whether a model discriminates beyond what legitimate risk implies, or whether a policy chooses well, requires knowing the truth the model approximates, which observational data never supplies. Group membership was excluded from both outcome equations, and every number from them is reported as simulation.

3.2 Which predictors are investigated, and why

Every predictor in Table 1 is observable at origination, before any repayment behaviour exists, since a predictor reflecting how a loan is performing would trivially predict whether it fails. The outcome is charge-off, dated; loan status, charge-off amount, and paid-in-full date are excluded as post-origination fields that would leak it. Protected attributes appear only in the consumer-benchmark audit and are never model inputs.

Table 1. Predictors used on the SBA data and the reason each is included.

Predictor	Why it is investigated
Term in months	Loan structure; Section 4.3 shows part of its weight is a time-at-risk artifact.
Gross approval (log)	Loan size; small loans historically charge off more often.
Guarantee ratio	Share of risk retained by the private lender.
Initial interest rate	The lender's own priced-risk signal.
Jobs supported	Proxy for business scale, otherwise absent.



Predictor	Why it is investigated
NAICS sector (2-digit)	Industry cyclicity.
Project state	Tests the digital-redlining channel (13.4% of attribution, Section 5.2).
Business age	Strongest predictor in the small-business survival literature.
Business type	Legal structure sets liability and owner exposure.
Collateral indicator	Secured versus unsecured; classical underwriting variable.
Revolver status	Revolving credit behaves differently from term debt.
Processing method	Underwriting channel; proxy for scrutiny received.
Fixed or variable rate	Exposure to rate movements over the loan's life.
Loan age (quarter)	Time at risk. Added for the hazard model; fixed-window models saw it only through term.
Calendar quarter	Macro regime, separating loan age from the economy it ages in.

3.3 Two cohort designs

The methodological core compares two ways of defining the analysis cohort, both defensible on their face.

Design A (resolved-only) keeps loans that have reached PIF or CHGOFF and labels CHGOFF. This is the approach the structure of the file invites: 61.9% of records carry a withheld status, and discarding them leaves a dataset in which every label is certain. It is not offered as best practice, and this paper does not document how often it is used in production; the results below establish the size of the error it produces, not its prevalence.

Design B (maturity-matched) requires that every disbursed loan be approved early enough for a full 36 months to have elapsed, and labels charge-offs occurring within 36 months of approval. A loan still performing at 36 months is a non-default.

The difference is whether outcome availability depends on the outcome. Under Design A it does: a loan that fails quickly reaches CHGOFF and enters the cohort, while an identical loan performing normally sits in EXEMPT and is discarded. Both designs use the same origination-only features, model families, and strictly out-of-time protocol — trained on FY2020 approvals, calibrated on FY2021, evaluated on FY2022 — so no loan appears at more than one stage and every result is measured on approvals the model has never seen. Cohort sizes are in

Table 3; the FY2020 training cohort contains 36,464 loans under Design B and 22,892 under Design A.

3.4 The ratio-scorecard baseline

Comparing machine learning to traditional underwriting requires a defined baseline, and the choice influences the comparison. The baseline here is a ratio scorecard of five origination ratios discretized into quintiles and combined additively by logistic regression; quintile binning is deliberate, because it is how underwriting sheets were built and it constrains the model to the coarse monotone judgment a human could reproduce. A second variant adds business type, business age, and collateral status.

This is generous in one respect and harsh in another — generous because it is fitted on the same 36,464 loans as the machine-learning model, whereas real scorecards use less data and are updated rarely; harsh because a human underwriter also sees documents, statements, and the applicant. It measures what each method extracts from identical structured data, not whether humans underwrite well.

3.5 Models, evaluation, and uncertainty

Three model families are used throughout: L2-regularized logistic regression, random forest (Breiman, 2001), and gradient-boosted trees — LightGBM (Ke et al., 2017) on SBA data, XGBoost (Chen & Guestrin, 2016) on consumer files.

Every imputation, scaling, and encoding is fitted inside the resampling loop on training data only: within each fold the transformer is fitted on the training partition and applied unchanged to the held-out partition, so no evaluation data influences the fitted parameters. Fitting transformers before splitting leaks test-set information; on 1,000-row files the effect is small, but every result here was produced with the ordering corrected.

Models are judged on AUC, precision-recall AUC (PR-AUC), Brier score (Brier, 1950), and expected calibration error (ECE; Naeini et al., 2015; Guo et al., 2017), reflecting this paper's premise that no single measure suffices. Calibration used isotonic regression (Zadrozny & Elkan, 2002) fitted on a held-out fiscal year, never the test period; being non-parametric, it requires only that the mapping from predicted score to realized default rate be monotone, which suits a setting where the shape of the miscalibration is unknown. On the consumer files, evaluation uses ten repetitions of stratified five-fold cross-validation with thresholds selected by inner cross-validation on training folds only, since a single five-fold split of 1,000 cases at a 30% bad rate cannot separate differences of order 0.01 AUC. Simulation results report means with 2.5th–97.5th percentile intervals and the fraction of replications in which each directional claim holds. Protected attributes were never model inputs, and proxy-reconstruction AUC was computed out of sample because the in-sample version overstates the channel badly: in this study's own audit, 0.918 against 0.746 (Section 4.4).

4. Results

4.1 Ratio scorecards against machine learning

Three approaches were fitted on the same FY2020 loans, calibrated on FY2021, and evaluated on the same 41,764 FY2022 loans (Table 2). Ranking barely separates them — 0.924 for both scorecards against 0.946 for the machine-learning model — while precision-recall and calibration separate them sharply: PR-AUC rises from 0.151 to 0.412 and expected calibration error falls from 0.0087 to 0.0048.

Table 2. Three underwriting approaches on identical data

Approach	Inputs	AUC	PR-AUC	Brier	ECE	Speed (μs/app)
Ratio scorecard	5	0.924	0.151	0.0172	0.0087	0.6
Scorecard + judgment fields	8	0.924	0.164	0.0173	0.0093	1.6
Machine learning (full)	13	0.946	0.412	0.0136	0.0048	16.5

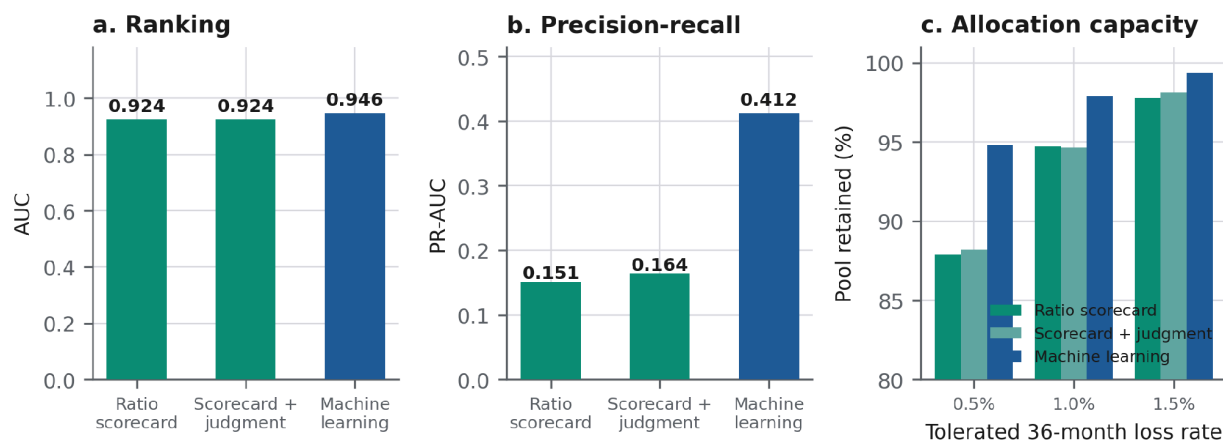


Figure 1. Ratio-scorecard and machine-learning approaches for ranking (a), precision-recall (b), and allocation capacity (c), the last being the share of the pool retainable within a stated loss budget.

On the most-reported metric, the area under the receiver operating characteristic curve (ROC-AUC), the difference is modest (Figure 1a): AUC rises from 0.924 to 0.946 in moving from five hand-binned ratios to a gradient-boosted ensemble on thirteen fields, and adding a loan officer's judgment fields to the scorecard adds 0.0003.



The precision-recall gap is the substantively larger finding (Figure 1b). The realized 36-month charge-off rate of the FY2022 cohort is 1.91% (Table 3); because only about one loan in fifty charges off, a model can rank well overall while scattering most of those rare failures through the middle of its ranking, and that scattering is what the fourfold PR-AUC gap measures. The scorecard ranks borrowers nearly as well but estimates absolute risk less accurately.

This complicates the two claims most often made about machine learning in credit scoring: that it substantially outperforms traditional scorecards, and that whatever gain it delivers is visible in ROC-AUC. It did not outperform substantially, and the gains it did deliver, in rare-event precision and calibration, are exactly the ones the headline metric hides. One caveat governs the comparison: the three approaches differ in model class and feature set at once — five, eight, and thirteen fields — so the gap is not a clean estimate of what flexibility alone buys. Section 4.2 shows that removing a single field, loan term, costs the model most of its precision-recall advantage, suggesting much of the difference is carried by the feature set rather than by model capacity. A matched design fitting every family on both feature sets would separate the two effects; that is the natural next step and is not something this cohort can settle.

Scoring speed is not a differentiator either: 0.6 microseconds per application against 16.5, a 28-fold gap that is operationally negligible. Faster turnaround comes from reducing human review, which is where automation removes the chance for unusual but legitimate cases to be reconsidered (Section 5.3).

The clearest advantage for machine learning is allocation (Figure 1c). Sorting applicants from lowest to highest predicted risk until the realized loss rate reached a budget, the model retained 94.8% of the pool at a 0.5% budget against 87.9% for the scorecard — 6.9 points, roughly 2,900 businesses — narrowing to 1.6 points at a 1.5% budget, where nearly everyone is approvable. Because every loan was approved and disbursed, this is a counterfactual within an accepted pool, not evidence of expanded access.

These differences exceed test-set noise. Resampling the FY2022 cohort 1,000 times with the same indices applied to every approach, the model's paired advantage over the ratio scorecard was +0.022 AUC (95% CI: 0.017 to 0.028), +0.261 PR-AUC (0.234 to 0.287), -0.0039 expected calibration error (-0.0049 to -0.0026), and +6.9 points of retained pool at a 0.5% loss budget (5.6 to 8.1); every interval excludes zero. Adding judgment fields is a different matter: +0.0003 AUC, interval -0.0016 to 0.0021.

4.2 The outcome window dominates everything else

This is the paper's central result: the two cohort designs of Section 3.3 describe the same portfolio but produce charge-off rates differing sevenfold, while ROC-AUC does not reveal it.

Table 3. Cohort size and observed charge-off rate under each design

Approval FY	Design A		Design B	
	n	rate	n	rate
FY2020 (train)	22,892	6.38%	36,464	0.83%
FY2021 (calibrate)	20,938	6.10%	45,165	0.73%

	Design A		Design B	
Approval FY	n	rate	n	rate
FY2022 (test)	12,886	13.24%	41,764	1.91%

Design A suggests charge-offs among FY2022 approvals reached 13.24%, more than double FY2020 and enough to justify much tighter credit; Design B describes the same loans but puts the FY2022 36-month rate at 1.91% against 0.83% in FY2020 (Table 3). The direction is the same; the level differs sevenfold.

The mechanism is straightforward. Under Design A, a loan enters the cohort only after resolving, and where most loans are still outstanding, early resolution is strongly associated with failure: a FY2022 loan that is charged off in 2024 is included, while an otherwise similar ten-year loan still performing is not. The distortion therefore grows in more recent cohorts, and Design A measures risk among loans that finished early rather than the risk of the portfolio. Design B avoids this by construction, since eligibility turns only on whether 36 months have elapsed since approval — fixed at origination, independent of how the loan turned out. Design A conditions the sample on the outcome; Design B conditions it on the clock, and that difference separates 13.24% from 1.91%.

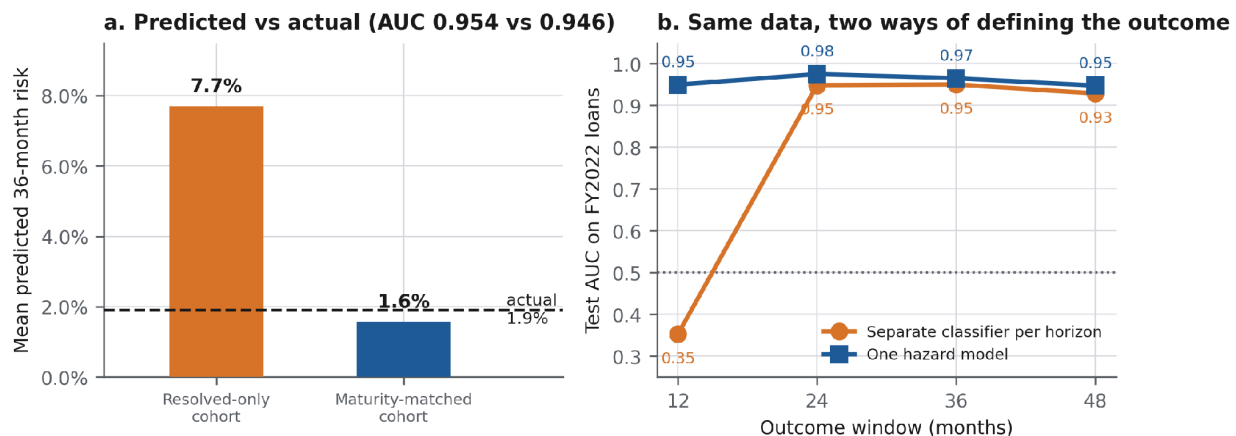


Figure 2. (a) The two models on the same 41,764 FY2022 loans and the same 36-month outcome, only the training cohort differing. (b) Test AUC against outcome-window width, for separately fitted classifiers and for the single hazard model of Section 4.3.

A possible objection is that the distortion may not matter if the model still ranks borrowers correctly and can be recalibrated. Both models therefore scored the same 41,764 FY2022 loans against the same 36-month outcome and were isotonicly recalibrated on a held-out fiscal year, only the training cohort differing. The resolved-only model predicts an average 36-month charge-off risk of 7.70% against a realized 1.91% — a fourfold overstatement after calibration — while the maturity-matched model predicts 1.58%; expected calibration error was 0.058 against 0.005. Under a policy declining applicants above 3% predicted risk, the resolved-only model declines 37.0% of the cohort against 6.7%, turning away 14,674 borrowers whose loans did not charge off within the window against 2,145; at a 10% policy, 3,899 against 1,180.

The mis-specified model still had the higher AUC — 0.954 versus 0.946. The gap is small and the argument does not rest on its size; what matters is its direction and what it conceals. On the metric a lender is most likely to consult, the mis-specified model is not merely indistinguishable from the correct one but marginally preferred, while the two disagree on 12,529 declines. Selecting on discrimination would therefore choose the model that wrongly declines those borrowers. PR-AUC does reveal the problem, at 0.286 against 0.412, supporting its use in rare-event settings; it is not, however, the metric most often reported.

An ablation identifies where the apparent performance comes from. Removing the interest rate changed nothing, leaving AUC at 0.950; removing loan term reduced AUC to 0.747 and PR-AUC from 0.433 to 0.057, an 87% drop, so most of the model's performance rests on one administrative field describing loan structure rather than borrower creditworthiness. At a 12-month window the same pipeline, features, and code produce an AUC of 0.353 on a cohort with only 27 charge-offs among 41,764 loans: the model performs worse than chance at 12 months though it looks excellent at 36, and only the outcome definition changed.

4.3 A hazard model addresses the windowing problem

The instability in Section 4.2 is not caused by a particular algorithm but by forcing a time-to-event process into a binary label at an arbitrary cutoff: loans default at different times or remain under observation, and a time-to-event model represents that structure directly instead of requiring an outcome window.

Discrete-time hazard models are standard in credit risk (Section 2); the question here is what is lost when fixed-window classifiers are used instead. Following the person-period formulation of Singer and Willett (1993), we restructured the SBA data so that each loan contributes one row per quarter it survives, up to 24 quarters, producing 2,030,904 loan-quarter observations across 123,393 loans with 4,171 charge-off events. Writing $h(t | x)$ for the probability that a loan charges off during quarter t given that it is active at its start, the model is a logistic regression $\text{logit } h(t | x) = a(t) + b'x$, with $a(t)$ a flexible baseline in loan age and x the origination covariates of Table 1 plus calendar quarter. Cumulative default by horizon T is one minus the product of $(1 - h(t | x))$ across $t = 1$ to T , which is how Table 4 comes from one fitted model rather than four. Loans still performing when the data ends simply stop contributing risk periods.

The baseline hazard explains the earlier instability. Quarterly charge-off risk is near zero for the first 15 months, then rises steadily to about 0.34% per quarter by year five (Figure 3a), so a 12-month window observes very few events while a 36-month window observes borrowers at different points on that curve depending on loan structure. Loan term was partly acting as a proxy for time at risk: once loan age enters explicitly, term's share of attribution falls from 31.1% to 23.7%, while loan age and calendar quarter together account for 34.9%.

Table 4. One hazard model, queried at four horizons, against four separately fitted classifiers.

Horizon	Observed rate	Hazard model: predicted	Hazard AUC	Fixed-window AUC
12 months	0.07%	0.01%	0.950	0.353
24 months	0.67%	0.59%	0.975	0.949
36 months	1.91%	1.74%	0.965	0.950
48 months	4.89%	3.75%	0.947	0.929

Note. The fixed-window column refits a classifier at each horizon and discards loans without a full look-forward; the hazard column is one model, fitted once on FY2020 loan-quarters, with survival probabilities chained to each horizon.

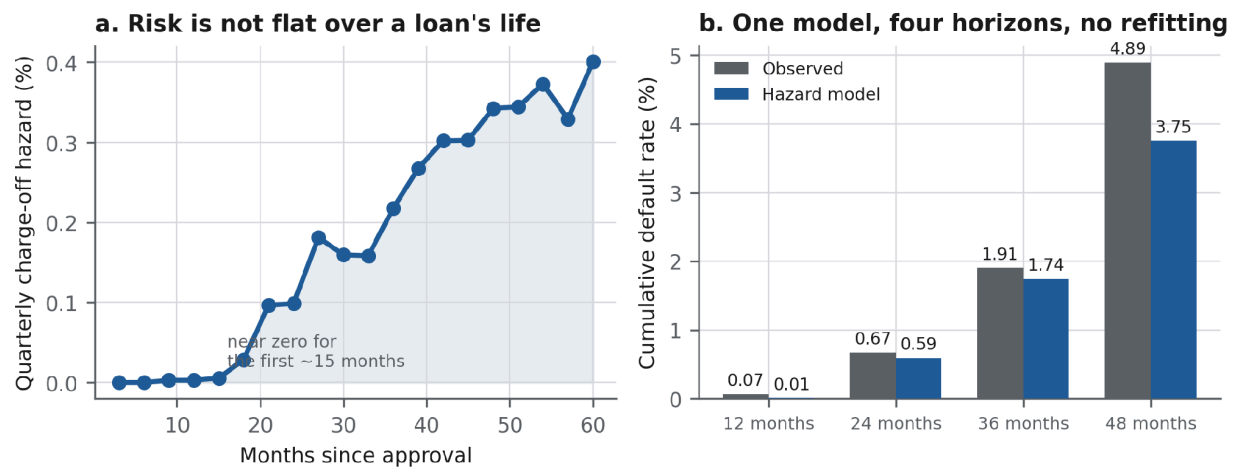


Figure 3. (a) Estimated quarterly charge-off hazard by loan age, FY2020 cohort. (b) Cumulative default predicted by the single hazard model against realized rates at four horizons.

The hazard model's AUC ranges from 0.947 to 0.975 across the four horizons, a spread of 0.028, against 0.353 to 0.950 — a spread of 0.597 — for the four separately fitted classifiers; at the horizon where the fixed-window approach fails completely the hazard model scores 0.950. Predicted cumulative default tracks the realized rate at every horizon, with expected calibration error from 0.0005 at twelve months to 0.016 at 48 months, where the model under-predicts because it must extrapolate the calendar-quarter covariate beyond its training range. Expected calibration error is absolute, however, and at these base rates it flatters every model, so it should be read alongside the ratio of predicted to realized risk. On that measure the twelve-month column is the weakest rather than the strongest: 0.01% predicted against 0.07% realized understates sevenfold, while 3.75% against 4.89% at 48 months falls short by a quarter. An absolute error of 0.0005 against a base rate of 0.0007 shows only that both quantities are small — which is why decision-level counts are reported alongside the metric in Section 4.2.

The 12-month figure rests on 27 charge-offs, so it deserves an interval. Bootstrapping that cohort 2,000 times gives AUC 0.353 (95% interval 0.268 to 0.448), with 99.9% of resamples

below 0.5; the 24-, 36- and 48-month intervals are 0.934–0.962, 0.942–0.958, and 0.919–0.938. The collapse at twelve months is a property of the design, not a thin event count.

The baseline recovered here — flat for 15 months, then rising steadily — is the small-business analogue of the flexible baselines Djeundje and Crook (2019) found to improve on fixed forms, and is not a shape any parametric assumption would have supplied. The calendar-quarter term plays the role of the time-varying macroeconomic covariates in Bellotti and Crook (2014), carrying 16.3% of attribution — a large share for a variable describing the environment rather than the borrower.

The practical implication is that a lender need not choose an outcome window. One hazard model, read at whatever horizon the business question requires, gives internally consistent answers, uses every loan including those still outstanding, and makes the shape of risk over time an output rather than a hidden assumption. The costs are real but bounded: restructuring into loan-periods multiplies the row count roughly sixteenfold here, a few minutes of computation; the person-period form assumes censoring is uninformative once covariates are conditioned on; and horizons beyond the observed range require extrapolating the calendar-quarter covariate, which is where the 48-month estimate under-predicts. Against these, the fixed-window alternative offers only simplicity, bought by fixing an arbitrary horizon and discarding every loan that has not reached it.

4.4 The proxy audit tests the wrong thing

The standard fair-lending diagnostic asks whether protected-group membership can be reconstructed from a model's features; if it can, the features are proxies and the usual remedy is to delete them. This is narrower than the criteria the fairness literature has formalised — equalised odds and equality of opportunity constrain error rates across groups rather than recoverability of group membership (Hardt et al., 2016) — and the two can move independently. We tested both halves of that logic where group membership is excluded from the outcome equation by construction, so any disparity beyond what genuine risk implies must come from the model.

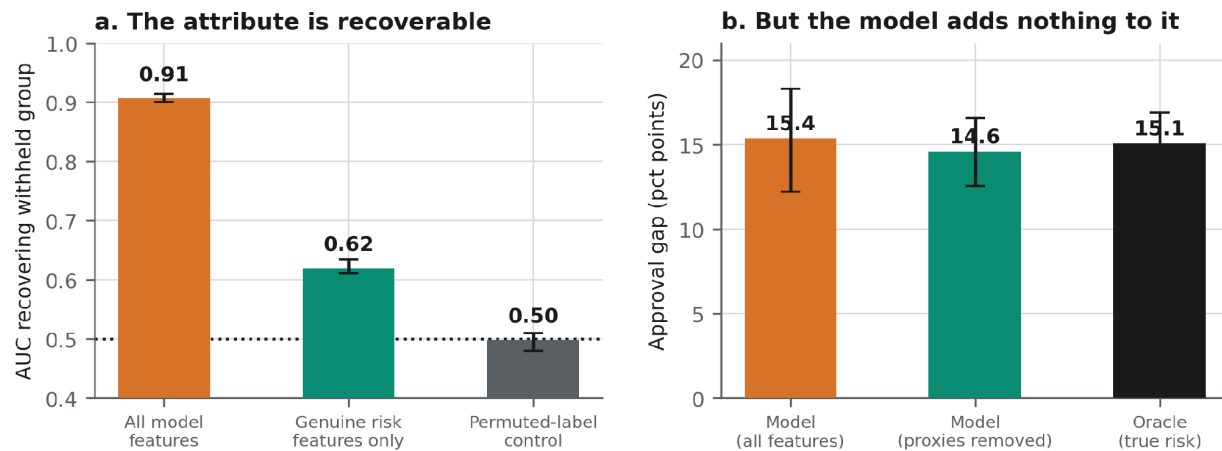


Figure 4. Simulation, 25 replications of 30,000 applicants. (a) Recovering the withheld group labels. (b) Approval gap at a fixed 70% approval rate against an oracle scoring true default probability.

The first half holds emphatically. A classifier trained on exactly the features the lending model uses recovers group membership at AUC 0.907 (95% interval 0.901–0.914); restricted to the six genuine risk features, with the designed proxy channels removed, it still reaches 0.620, while a randomly permuted label returns 0.498, confirming the result is not an artifact. Redundant encoding is real, strong, and not confined to features one would flag as suspicious.

The second claim is not supported here, and the qualification carries weight: because group membership is excluded from the outcome equation by construction, the result shows that a model fitted to a process containing no excess discrimination reproduces none — a consistency check on the design as much as a test of the hypothesis, and no evidence about how often this holds in real lending. At a fixed 70% approval rate the model produces a 15.4-point approval gap (12.2–18.3) against an oracle using true default probability at 15.1 points (13.2–16.9); the excess attributable to the model is +0.3 points (–2.1 to +3.5), and the model's gap exceeds the oracle's in 60% of replications, close to chance.

Deleting the two proxy features reduces the gap only from 15.4 to 14.6, helps in 72% of replications and backfires in the rest, costs 0.0001 AUC, and leaves reconstruction at 0.620. Flagging proxy features, deleting them, and rerunning the same audit could therefore appear to address the issue while changing little in the decision gap.

The consumer benchmark adds a methodological warning. Cross-validated reconstruction AUC for foreign-national status is 0.746 on the original file and 0.747 on the corrected one, with every approval gap spanning zero on a comparison group of 37 people; fitted in sample, as a quick audit usually is, those figures are 0.918 and 0.925. An audit without a held-out split overstates the proxy channel badly.

4.4.1 The same test, on real loans

A simulation can only show that two quantities are separable in principle; the SBA data allows the question empirically. SHAP attributes 13.4% of this model's decision weight to project

state, and because geography is a classical redlining channel it is the natural real-data analogue of a proxy feature. Removing it and refitting produces Table 5.

Table 5. Removing the geographic feature from the machine-learning model, FY2022 test cohort

Model	AUC	PR-AUC	Brier	ECE	Mean state excess
With project state	0.946	0.412	0.0136	0.0048	0.0064
Project state removed	0.949	0.401	0.0137	0.0047	0.0061

Note. State excess is the predicted 36-month risk for a state minus that state's realized 36-month rate, over the 38 states with at least 250 test loans; a model treating every state as its own record warrants would show excess near zero. The state's realized rate plays the role the oracle plays in the simulation.

Both halves of the simulated result are reproduced. Removing the feature costs nothing — AUC rises slightly to 0.949, precision-recall falls slightly to 0.401 — and achieves nearly nothing: mean absolute state-level excess moves from 0.0064 to 0.0061, a reduction of 0.00027 whose bootstrap interval, -0.00095 to $+0.00027$, comfortably spans zero. Pruning helped in 81.5% of resamples, which sounds encouraging until the magnitude is read alongside it: at a 3% decline policy the spread in state decline rates narrows only from 14.2 to 12.5 percentage points, so a disparity of over 12 points survives deleting the only explicitly geographic input.

The reason is the same as in the simulation: industry mix, loan size, term structure, and processing channel all vary regionally, so the model reconstructs most of the geography it was denied. This is the paper's most contrarian claim, and the one we most expected the real data to overturn. It did not.

4.5 Calibration decays; ranking does not

Under a simulated shock lifting the realized default rate from 26.2% to 42.6%, the model's AUC moves from 0.726 to 0.721 — a decline of 0.005, within the run-to-run variation of the training procedure itself — so any monitoring system watching discrimination would report the model as healthy. Over the same shock the Brier score rises from 0.168 to 0.237 and expected calibration error reaches 0.161: an applicant assigned 20% predicted default risk corresponded to realized risk closer to 36%.

Credit decisions use absolute probabilities, not rankings alone — approval cutoffs, risk-based pricing, and loss provisioning all depend on estimated risk levels — and because losses arrive months later a model can keep ranking borrowers well while its probabilities drift. In the simulation, isotonic regression fitted on 1,500 labelled post-shift cases reduces expected calibration error from 0.161 to 0.026 and Brier score from 0.237 to 0.211 in all 25 replications,

without retraining the model and with AUC essentially unchanged. The failure is calibration; the usual monitoring metric is ranking.

4.5.1 The same drift, across real fiscal years

FY2020 to FY2022 spans a real macroeconomic disturbance, so the simulated shock has an empirical counterpart. This is the stress test referred to in Section 5.4: a calibration-based check that registers a failure ROC-AUC does not report. A model was fitted on 70% of FY2020 approvals, its calibration map on the held-out 30%, and both were frozen and applied forward. Because Section 1112 payments covered the FY2020 cohort and much of FY2021 but not FY2022 (Section 3.1), the drift is partly a policy effect: the model learns a subsidised default process and is applied to an unsubsidised one. That does not weaken the operational point — a lender underwriting in 2022 faced exactly this shift — but the magnitude is a combined economic and policy transition.

Table 6. One model and one calibration map, both fitted in FY2020 and left untouched.

Cohort	n	Realized	Predicted	AUC	ECE
FY2020 (held-out slice)	10,940	0.89%	0.89%	0.974	0.000
FY2021 (+1 year)	45,165	0.73%	0.58%	0.968	0.0019
FY2022 (+2 years)	41,764	1.91%	1.25%	0.948	0.0080
FY2022, recalibrated on FY2021	41,764	1.91%	1.52%	0.948	0.0058

Note. Expected calibration error is zero by construction in the first row: the isotonic map was fitted on exactly that held-out slice. That row is the baseline against which subsequent drift is measured, not an independent performance claim.

The pattern the simulation produced is the pattern the real portfolio produces. Over two fiscal years the model's ranking degrades by 0.026 of AUC, from 0.974 to 0.948 — a decline any dashboard would treat as noise — while its probabilities drift from exactly calibrated to an expected calibration error of 0.0080, understating realized charge-off risk by a factor of 1.52 by FY2022. A lender pricing or provisioning off those numbers would have been short by half again, while every discrimination metric said the model was healthy.

The repair behaves as in simulation: refitting only the calibration map on FY2021, the model itself untouched, cuts expected calibration error by 28.2% and costs 0.0004 of AUC. It is smaller here than under the simulated shock, as should be expected, since a real recalibration window is never drawn from the regime it corrects for — a limitation of the remedy worth stating alongside it.

4.6 Simple models produce greater decision homogeneity in this experiment

A common concern about automated underwriting is that it removes a second opinion: if lenders use similar models, a borrower rejected by one may be rejected by all. We split FY2020 approvals into disjoint halves, trained the same method on each, and measured how often the

two lenders rejected the same borrowers; under independent judgment the overlap would equal the decline rate. The result runs opposite to the usual concern about complexity.

Table 7. Agreement between two lenders using the same method trained on disjoint data.

Approach	Score correlation	Joint decline at 10%	Joint decline at 20%
Ratio scorecard	0.965	89.6%	88.2%
Scorecard + judgment fields	0.955	83.4%	89.6%
Machine learning (full)	0.616	59.3%	62.7%
Independent judgment (reference)	—	10.0%	20.0%

The simple models produce the most similar decisions: two independently trained scorecards correlate at 0.965 and jointly decline 89.6% of each other's rejects, while two independently trained machine-learning models correlate at 0.616, with a joint decline rate of 59.3%. In this experiment the simpler and more interpretable method produces greater decision homogeneity.

This bears on the literature on algorithmic monoculture. Kleinberg and Raghavan (2021) show that convergence on a single algorithm can reduce collective decision quality even when that algorithm is individually more accurate, and Bommasani et al. (2022) argue that shared components, especially training data, make decisions more homogeneous. Here the lenders share no training data, yet the scorecards still agree closely: shared components can cause homogenization but are not required for it, so capacity and data sharing are separate sources of concentration risk.

The mechanism is straightforward. A five-ratio scorecard has little room to vary, and when trained on large samples from the same population it recovers similar coefficients on the same quintiles, while a flexible model trained on different data can learn different interactions and treat boundary cases differently. Model capacity is therefore not only an accuracy property; it also affects decision diversity. This is not an endorsement of complexity — with a shared vendor score much of that diversity may disappear, and 59.3% joint rejection remains far above independent judgment. Concentration risk should be measured in the full scoring system, not inferred from model class.

4.6.1 Model families on the two consumer files

H3 predicted that flexible models would not meaningfully outperform a well-specified logistic regression on small structured credit files. Table 8 tests it on both German Credit versions under ten repeats of stratified five-fold cross-validation. On ranking the prediction holds: random forest leads logistic regression by 0.0112 AUC on the original file and 0.0127 on the corrected one, but the paired 95% intervals across the fifty folds are $[-0.0043, 0.0266]$ and $[-0.0026, 0.0277]$, both crossing zero, as do XGBoost's smaller margins. The trees win most

folds — 92.6% and 94.7% for random forest — without winning by an amount these files can resolve. Calibration is the one unambiguous difference: expected calibration error falls from 0.135 to 0.097 on the original file and from 0.136 to 0.098 on the corrected one, about 28%. That is the division the SBA data shows in Section 4.1: on files this size, flexibility buys probability quality rather than ranking.

Table 8. Three model families on both German Credit files, ten repeats of five-fold cross-validation.

Dataset	Model	AUC	PR-AUC	Brier	ECE
German Credit (original)	Logistic L2	0.780	0.587	0.191	0.135
	Random forest	0.791	0.625	0.176	0.110
	XGBoost	0.786	0.620	0.176	0.097
South German (corrected)	Logistic L2	0.773	0.595	0.193	0.136
	Random forest	0.785	0.618	0.178	0.112
	XGBoost	0.782	0.612	0.178	0.098

Note. Means over fifty folds. Paired differences against logistic regression, 95% intervals across folds: random forest +0.0112 [−0.0043, 0.0266] and XGBoost +0.0065 [−0.0091, 0.0221] on the original file; random forest +0.0127 [−0.0026, 0.0277] and XGBoost +0.0098 [−0.0055, 0.0244] on the corrected file. All four contain zero. Threshold and cost results under the 5:1 cost matrix are in the replication package.

5. Discussion: A Priority Framework for Mitigations

Four interventions are common in responsible-AI-in-finance discussions. We implemented and measured each rather than recommending them in principle; their effects differ substantially, and those receiving most attention are not those producing the largest measured benefit here.

5.1 Fairness-constrained optimization

The constrained advisory policy of Appendix A shows a clear benefit: a hard suitability constraint costs 0.04% of utility and removes all violations. Feature pruning also costs little but changes little. A constraint acts directly on the decision, while pruning removes inputs and assumes the decision will follow; when protected status is redundantly encoded, as in Section 4.4, it does not. Constrain decisions directly where possible.

5.2 Explainable AI

Regulation B requires an adverse-action notice that either states the specific reasons for a decision or discloses the applicant's right to obtain them, and any reasons given must reflect the factors actually considered, so explanation remains a compliance requirement (Equal Credit Opportunity Act, 2024). SHAP attribution (Lundberg & Lee, 2017) over the FY2022 cohort is concentrated: loan term 31.1%, loan size 17.2%, project state 13.4%, industry sector 11.6%,

interest rate 9.8%. Three fields carry 61.7% of attribution and 17 of 108 encoded features carry 80%, so the main drivers are directly identifiable.

There are two limits. Project state is the third-largest driver, and geography is a classic redlining concern; SHAP can show that geography matters but not whether the pattern reflects genuine regional differences in business risk or historical lending patterns. And SHAP does not reveal the outcome-window problem, the largest defect in this study. Explainability describes the model that was fitted; it cannot establish that the target was defined correctly.

5.3 Hybrid human–AI review

The value of human review depends on whether model errors concentrate near the decision threshold. In this cohort they do: we ranked applications by distance from the 3% cutoff and deferred the most ambiguous (Table 9).

Table 9. Concentration of decision errors near the cutoff.

Share deferred	Errors captured	Error rate in band	Error rate outside	Lift
2%	24.0%	66.0%	4.3%	15.5×
5%	47.1%	51.8%	3.1%	16.9×
10%	60.2%	33.1%	2.4%	13.6×
20%	86.2%	23.7%	0.9%	25.0×

Routing 5% of applications to a person captures 47% of the model's decision errors; within that band the model is wrong 51.8% of the time against 3.1% outside it. No human reviewer is simulated, so the claim is only about where review is most valuable: under a symmetric error-cost assumption, human accuracy above 50% in this band would improve the system. Among the four interventions tested, this gives the largest measured benefit per unit of review effort.

5.4 Robustness testing

The stress test in Section 4.5 detects a failure AUC misses, so the recommendation is to monitor calibration as well as discrimination: a reliability curve on a rolling window, paired with an expected-calibration-error threshold that triggers recalibration, detects this drift without retraining, where an AUC-only dashboard would stay flat.

5.5 The ordering

Across these experiments the priority ordering is: hybrid deferral first, calibration monitoring and recalibration second, outcome-constrained fairness rules third, input-side feature pruning last. Explainability sits outside the ranking, being a compliance requirement rather than an optimization tool.

The paper's contribution is therefore an ordering of design priorities rather than a new algorithm: a lender gains more from fixing how the outcome is defined and monitored than from any choice among model families, and the measurements here put a size on each step.

6. Limitations

The SBA analysis is not an approval model. It estimates charge-off risk among loans approved, disbursed, and federally guaranteed, and says nothing about who was denied credit; the decline-policy, allocation-capacity, and systemic-exclusion results are counterfactuals about how each model would behave if used for approval, computed on an already-approved pool.

EXEMPT is treated as non-default within the window, but those loans are outstanding, not necessarily healthy: Design B and the hazard model both estimate charge-off rather than distress and are lower bounds on the latter. Federal debt relief widens that gap unevenly, since Section 1112 payments kept loans current that were not necessarily sound, so the cohort-over-cohort deterioration reported throughout mixes credit and policy in a way this design cannot separate.

The hazard model extrapolates at long horizons: its 48-month predictions require the calendar-quarter covariate to run beyond its training range and under-predict there (3.75% against 4.89%), so long-horizon figures should be read as ordering rather than level.

The simulations are mechanism studies. Both generating processes were specified by the authors, so two findings reflect those specifications in part: that the model adds nothing to the oracle's gap, and that safety costs two orders of magnitude less than legibility. They establish that these quantities can come apart, not how often they do so in practice.

The scorecard baseline is a stand-in, fitted on far more data than a real scorecard would use and seeing no documents, statements, or applicants. The consumer benchmarks are too small for fairness conclusions: the minority category contains 37 people, and sex is only partly recoverable from the personal-status field. Of these limitations the first is binding — nothing here measures access for applicants who were declined, which would require application-level data the FOIA extract does not contain. The most easily repaired is the long-horizon extrapolation: two further fiscal years of training data, or a spline in place of the linear calendar-quarter term, would let the 48-month estimate rest on observed rather than extrapolated regimes.

7. Conclusion

Machine learning outperformed ratio underwriting here, but the advantage depended on the metric, and the largest defect in the study is tied to neither model class: a model trained on the wrong cohort produces the highest AUC in the paper yet would decline roughly fourteen thousand borrowers whose loans did not charge off within the window. The metric is not wrong; it answers a narrower question than the lending decision requires.

Modelling the outcome as a time-to-event process addresses the windowing problem directly: one hazard model spans four horizons with an AUC spread of 0.028 against 0.597 for four separately fitted classifiers, while retaining outstanding loans that fixed-window designs must discard. Outcome design therefore comes before the fairness and explainability interventions. Two further results deserve restating. The simulated regime shift and the real FY2020-to-FY2022 drift produce the same signature — ranking holds while calibration decays — with an AUC decline any dashboard would treat as noise while probabilities drift far enough to understate charge-off risk by half again. And two independently trained scorecards agreed on

89.6% of their declines against 59.3% for two machine-learning models, so concentration risk cannot be inferred from model class alone.

The remaining recommendations follow: send the cases closest to the decision threshold to human review, monitor calibration rather than ranking alone, and constrain what the system does rather than only what it sees. These steps begin after the model is built, yet they matter as much as model selection itself.

Refining these systems is a different task from adopting them, and the refinement step deserves more attention than it receives.

8. Declarations

Funding. No funding was received.

Competing interests. None declared.

Data availability. The SBA 7(a) FOIA extract is public at <https://data.sba.gov/dataset/7a-504-foia>; the German Credit files are at UCI (doi:10.24432/C5NC77, doi:10.24432/C5X89F). Analysis code, machine-readable results, and figures are in the replication package archived, which records the exact vintage, the cohort-construction scripts, and the random seeds used for every bootstrap and simulation reported here.

Use of AI. Anthropic's Claude assisted with analysis code, figure generation, reference verification, and language editing. All results were produced by executing the code in the replication package. The author reviewed and edited all content and takes full responsibility for it.

References

- Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84(3), 488–500. <https://doi.org/10.2307/1879431>
- Banasik, J., & Crook, J. (2007). Reject inference, augmentation, and sample selection. *European Journal of Operational Research*, 183(3), 1582–1594. <https://doi.org/10.1016/j.ejor.2006.06.072>
- Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104(3), 671–732. <https://doi.org/10.15779/Z38BG31>
- Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics*, 143(1), 30–56. <https://doi.org/10.1016/j.jfineco.2021.05.047>
- Bellotti, T., & Crook, J. (2014). Retail credit stress testing using a discrete hazard model with macroeconomic factors. *Journal of the Operational Research Society*, 65(3), 340–350. <https://doi.org/10.1057/jors.2013.91>
- Berg, T., Burg, V., Gombović, A., & Puri, M. (2020). On the rise of FinTechs: Credit scoring using digital footprints. *Review of Financial Studies*, 33(7), 2845–2897. <https://doi.org/10.1093/rfs/hhz099>
- Björkegren, D., & Grissen, D. (2020). Behavior revealed in mobile phone usage predicts credit repayment. *World Bank Economic Review*, 34(3), 618–634. <https://doi.org/10.1093/wber/lhz006>

- Bommasani, R., Creel, K. A., Kumar, A., Jurafsky, D., & Liang, P. (2022). Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? In *Advances in Neural Information Processing Systems* 35. <https://arxiv.org/abs/2211.13972>
- Botha, A., & Verster, T. (2026). Approaches for modelling the term-structure of default risk under IFRS 9: A tutorial using discrete-time survival analysis. *International Journal of Data Science and Analytics*, 22(1), Article 67. <https://doi.org/10.1007/s41060-026-01032-w>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD* (pp. 785–794). <https://doi.org/10.1145/2939672.2939785>
- Coronavirus Aid, Relief, and Economic Security Act, Pub. L. No. 116-136, § 1112 (2020).
- Dirick, L., Claeskens, G., & Baesens, B. (2017). Time to default in credit scoring using survival analysis: A benchmark study. *Journal of the Operational Research Society*, 68(6), 652–665. <https://doi.org/10.1057/s41274-016-0128-9>
- Djeundje, V. B., & Crook, J. (2019). Dynamic survival models with varying coefficients for credit risks. *European Journal of Operational Research*, 275(1), 319–333. <https://doi.org/10.1016/j.ejor.2018.11.029>
- Equal Credit Opportunity Act (Regulation B), 12 C.F.R. § 1002.9 (2024). <https://www.ecfr.gov/current/title-12/part-1002/section-1002.9>
- Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *Journal of Finance*, 77(1), 5–47. <https://doi.org/10.1111/jofi.13090>
- Grömping, U. (2019). South German credit data: Correcting a widely used data set (Report 04/2019). Beuth University of Applied Sciences Berlin.
- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th ICML (PMLR 70)*, pp. 1321–1330.
- Hand, D. J., & Adams, N. M. (2014). Selection bias in credit scorecard evaluation. *Journal of the Operational Research Society*, 65(3), 408–415. <https://doi.org/10.1057/jors.2013.55>
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems* 29 (pp. 3315–3323). <https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Aabstract.html>
- Hofmann, H. (1994). Statlog (German Credit Data) [Data set]. UCI Machine Learning Repository. <https://doi.org/10.24432/C5NC77>
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *NeurIPS* 30 (pp. 3146–3154).
- Kim, E., Garg, A., Peng, K., & Garg, N. (2025). Correlated errors in large language models. In *Proceedings of the 42nd International Conference on Machine Learning (PMLR 267)*, pp. 30038–30066). <https://proceedings.mlr.press/v267/kim25e.html>
- Kleinberg, J., & Raghavan, M. (2021). Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22), e2018340118. <https://doi.org/10.1073/pnas.2018340118>
- Kozodoi, N., Lessmann, S., Alamgir, M., Moreira-Matias, L., & Papakonstantinou, K. (2025). Fighting sampling bias: A framework for training and evaluating credit scoring models.

- European Journal of Operational Research, 324(2), 616–628.
<https://doi.org/10.1016/j.ejor.2025.01.040>
- Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124–136. <https://doi.org/10.1016/j.ejor.2015.05.030>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *NeurIPS 30* (pp. 4765–4774).
- Medina-Olivares, V., Calabrese, R., Crook, J., & Lindgren, F. (2023). Joint models for longitudinal and discrete survival data in credit scoring. *European Journal of Operational Research*, 307(3), 1457–1473. <https://doi.org/10.1016/j.ejor.2022.10.022>
- Naeini, M. P., Cooper, G. F., & Hauskrecht, M. (2015). Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), 2901–2907. <https://doi.org/10.1609/aaai.v29i1.9602>
- Siddiqi, N. (2017). *Intelligent credit scoring: Building and implementing better credit risk scorecards* (2nd ed.). Wiley. <https://doi.org/10.1002/9781119282396>
- Singer, J. D., & Willett, J. B. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics*, 18(2), 155–195. <https://doi.org/10.3102/10769986018002155>
- Stiglitz, J. E., & Weiss, A. (1981). Credit rationing in markets with imperfect information. *American Economic Review*, 71(3), 393–410.
- U.S. Small Business Administration. (2021). 7(a) and 504 Section 1112 payment extension (Procedural Notice 5000-20079). <https://www.sba.gov/document/procedural-notice-5000-20079-7a-504-section-1112-payment-extension>
- U.S. Small Business Administration. (2026). 7(a) & 504 FOIA [Data set]. <https://data.sba.gov/dataset/7a-504-foia>
- Wang, H., Bellotti, T., Qu, R., & Bai, R. (2024). Discrete-time survival models with neural networks for age–period–cohort analysis of credit risk. *Risks*, 12(2), 31. <https://doi.org/10.3390/risks12020031>
- Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD* (pp. 694–699). <https://doi.org/10.1145/775047.775151>

Appendix A. Advisor AI in personal banking

The same tensions appear on the retail side, where learned policies decide what to tell a household to do with its next dollar, though the structure differs: a lending model predicts an outcome, while an advisory system chooses an action and never observes what the rejected actions would have produced. The setup is a contextual bandit with eight context variables and three actions — build a cash reserve, pay down high-interest debt, invest — each with a known true expected utility, so regret is computable. Training is a log of 60,000 randomly assigned actions with noisy utility; evaluation uses 30,000 fresh households over 15 replications.

Table A1. Four advisory policies. Regret is measured against the best available action.

Policy	Utility	Mean regret	95th pct regret	Unsuitable
Oracle (true best action)	1.622	0.000	0.000	1.51%
Transparent suitability rule	1.558	0.063	0.297	0.00%
Unconstrained learned policy	1.620	0.0015	0.000	1.19%
Constrained learned policy	1.620	0.0021	0.000	0.00%

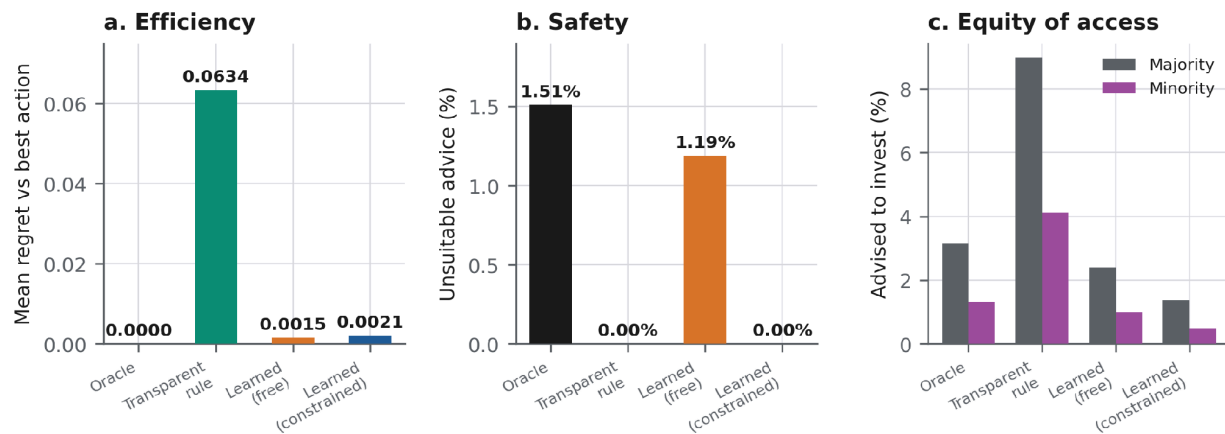


Figure A1. Advisory policies across 15 replications of 30,000 households; panel c reports the share of each group advised to invest.

The learned policy has mean regret of 0.0015 against an oracle that knows the true utility function, beating the transparent rule in every replication; the transparent rule's mean regret is 0.063, forty-two times higher, with a 95th-percentile regret of 0.297, so for the worst-served 5% of households the difference is economically meaningful.

The more important trade-off is between safety constraints and simplifying the policy for transparency: enforcing the suitability standard costs 0.0006 of utility — 0.04% — and removes unsuitable recommendations in all 15 replications, while replacing the learned policy with the transparent rule costs 3.8%. The safety constraint is nearly free; simplifying the decision rule is not, especially in the tail.

One limitation is that the oracle itself violates the suitability waterfall 1.51% of the time, blocking the highest-utility action for about one household in sixty-six, so such constraints are better treated as a minimum safety standard with a documented override process than as an absolute rule.

On equity, the oracle advises 3.16% of majority and 1.32% of minority households to invest, a 1.84-point gap reflecting differences in liquidity and debt within the simulation; the transparent rule advises investing more often (8.98% and 4.12%), widening the absolute gap to 4.86 points, while the constrained policy is most conservative at 1.37% and 0.49%. Ratio and absolute-gap measures therefore rank these policies differently, which is why a single fairness statistic is not enough. The transparent rule also has the only materially non-zero service-quality



gap (-0.029), favouring minority households because its conservative defaults fit lower-slack households better.