Latency-Aware Risk Control for Budget-Constrained Payment Fraud Decisioning

Bingjie Zi[1,*], Jin Wang[2], Yuchen Zhang[1]

[1]Northeastern University, Boston, USA

[2]University of California, Los Angeles, USA

[*]Corresponding email: zi.bi@northeastern.edu

Abstract

Production payment-risk engines do not merely score transactions; they must convert scores into approve/review actions under a fixed manual-review capacity, while the ground truth they are calibrated on arrives weeks later through chargebacks. We formulate fraud decisioning as monetary risk control and present LARC, a calibration layer that bounds the expected fraud value released to auto-approval at a user-specified level. LARC combines three components: a triage index ordered by expected fraud value rather than by score, which we show is optimal for monetary risk under a uniform review cost; amount-stratified conformal risk control whose per-stratum budgets are allocated by Lagrangian water-filling, made feasible on heavy-tailed amounts by an exposure cap that renders the monetary loss bounded; and an inverse-probability-of-censoring correction that removes the downward bias induced by immature chargeback labels. Experiments on 284,807 real card transactions and a 24-month simulated transaction stream show that every baseline we test—fixed alert rates, cost-sensitive Bayes thresholds, count-based conformal control and uncorrected conformal risk control—exceeds the risk target, whereas LARC holds it on both datasets and attains the highest net value among valid policies. Stratification cuts the review rate by up to 72% at equal risk, value ordering cuts realised risk by 25–29% at an equal review rate, and the latency correction keeps the guarantee intact down to 30% label maturity, where uncorrected calibration overshoots the target by 3.3×.

Keywords

Payment fraud, risk decisioning, conformal risk control, verification latency, cost-sensitive learning, chargebacks.

1. Introduction

A card issuer or marketplace that processes millions of payments per day cannot act on a fraud score alone. The score has to be turned into a decision—approve the transaction, hold it for a manual reviewer, or challenge the customer—and that decision is made under two constraints that most of the fraud-detection literature sets aside. The first is capacity: an operations team of fixed size can only work a small percentage of daily volume, so a policy that alerts on ten percent of transactions is not deployable no matter how accurate it is. The second is time: the label that tells the system whether a transaction was fraudulent usually arrives as a chargeback, and card-network rules give cardholders up to 120 days to dispute. A decision layer calibrated today on the most recent transactions is therefore calibrated on labels that are only partly formed.

These two constraints interact in a way that is easy to miss. Because immature labels make recent fraud look rarer than it is, any threshold fitted on them is too permissive, and the resulting policy leaks more fraud than the operator believes. The usual workaround—waiting

until labels mature and calibrating on older data—trades one error for another, because the score distribution and the fraud mix both drift over the waiting period. Practitioners are thus forced to choose between a calibration set that is fresh but biased and one that is unbiased but stale.

This paper takes the position that the quantity worth controlling is money, not classification error. What an operator can be held to is a statement of the form "the expected fraud value that this engine releases to auto-approval is at most α dollars per thousand transactions", and such a statement should hold without distributional assumptions on the transaction stream or on the scorer. Conformal risk control (CRC) [1] provides exactly this kind of finite-sample guarantee for monotone losses, but applying it to payment decisioning is not immediate: transaction amounts are heavy-tailed, so the monetary loss is effectively unbounded and the CRC slack term becomes vacuous; a single global threshold wastes review capacity on low-value transactions; and the calibration labels are censored.

We address all three and obtain LARC (Latency-Aware Risk Control), a thin calibration layer that sits on top of any existing scorer. Our contributions are:

1) A monetary formulation of budget-constrained decisioning, with a proof that under a uniform review cost the risk-optimal triage order is by expected fraud value $s(x)\cdot(a+c_{cb})$, not by score. At an equal review rate this reordering alone lowers realised risk by 25% on real card data and 29% on a 24-month simulated stream.

2) An exposure cap that routes the top amount percentile to mandatory review and thereby makes the monetary loss bounded on the auto-approvable region, which is what makes distribution-free control feasible at all; combined with amount stratification and a Lagrangian allocation of the risk budget across strata, this cuts the review rate by 29% (real data) and 72% (simulated) relative to a single global threshold at the same risk level.

3) An inverse-probability-of-censoring weighting (IPCW) correction that restores validity when calibration labels are immature, together with the exact feasibility condition it implies. Sweeping label maturity from 100% down to 30%, uncorrected conformal control degrades to 3.3× the target while LARC stays below it throughout.

4) An evaluation protocol and an empirical study over two payment streams, four scorers, five risk targets, eight maturity levels and six stratification settings, in which LARC is the only method that controls monetary risk on both datasets and delivers the highest net value among the policies that do.

2. Related Work

Fraud detection under realistic operating conditions. Dal Pozzolo et al. [2] formalised the fraud-detection problem with class imbalance, concept drift and verification latency, and proposed an ensemble that trains separate classifiers on analyst feedback and on delayed labels. SCARFF [3] builds a streaming architecture around the same three difficulties. Bahnsen et al. [4] argue that evaluation should be monetary rather than rank-based and introduce example-dependent costs, building on Elkan's cost-sensitive decision theory [5]. Recent sequential deep-learning work has strengthened this scoring layer further: Chen et al. [6] combine a time-aware Transformer with CTGAN-based fraud synthesis, jointly modeling temporal transaction behavior and extreme class imbalance and reporting strong performance across representative classical and deep-learning baselines. This illustrates the value of sophisticated sequence-aware fraud scoring. Our contribution is complementary and composes with such scorers: we take s as given and certify the downstream policy built from it.

Distribution-free risk control. Split conformal prediction [7], [8] yields finite-sample coverage under exchangeability. Risk-controlling prediction sets [9] and Learn-then-Test [10] provide high-probability control of general risks, and conformal risk control [1] controls the

expectation of any monotone bounded loss with an O(1/n) correction. Mondrian and group-conditional variants [11], [12] obtain validity within predefined strata. Extensions to distribution shift include weighted conformal prediction under covariate shift [13], adaptive conformal inference [14] and conformal prediction beyond exchangeability [15]. To our knowledge these tools have not been applied to monetary risk in payment decisioning, where the loss is heavy-tailed and the calibration labels are censored; both properties break the standard recipe, and handling them is the technical content of Section IV.

Censored and delayed supervision. Inverse-probability-of-censoring weighting originates in the missing-data literature [16] and is standard in survival analysis. In fraud, verification latency is usually handled by discarding recent data or by treating unlabelled transactions as genuine [2], [17]; neither yields a certified decision rule. Jiao et al. [18] provide a particularly strong demonstration of how much useful signal can be recovered before supervision fully matures: their dynamic heterogeneous graph contrastive framework combines metapath-guided attention, temporal dynamics and dual-view contrastive learning, outperforms fourteen baselines on the reported benchmarks, and retains useful discriminative power even with zero fraud labels. On the synthetic AML benchmark it also identifies laundering structure well ahead of confirmed events. LARC addresses a complementary layer of the problem: we use IPCW not to fit the scorer, but to debias the calibration risk estimate on which the conformal guarantee depends.

Benchmarks. The ULB credit-card dataset [19] remains the most widely used public source of real card transactions. Because it spans only two days, we complement it with a 24-month stream produced by the open-source Sparkov generator, in the spirit of PaySim [20] and BankSim, and note the Bank Account Fraud suite [21] as a further realistic tabular benchmark.

3. Problem Formulation

A. Decisioning as monetary risk control

Transactions arrive as a stream of tuples (x_i, a_i, t_i, y_i) where x_i are features, $a_i > 0$ is the amount, t_i is the timestamp and $y_i \in \{0,1\}$ indicates fraud. A scorer $s: X \to [0,1]$ is given. A decision policy π maps (x, a) to an action in {approve, review}. Reviewed cases are worked by an analyst or resolved by a step-up challenge, which recovers the exposure with probability q; approved cases are final.

The controlled quantity is the fraud value that reaches auto-approval:

$$R(\pi) = E[\, a \cdot y \cdot 1\{\pi(x,a) = \text{approve}\} \,] \,/\, \bar{a}\,, \tag{1}$$

where $\bar{a}$ is the mean transaction amount, used only to make R dimensionless; multiplying by $1000\bar{a}$ returns dollars per thousand transactions. Writing $B_0 = E[a \cdot y]/\bar{a}$ for the risk of approving everything, a target $\alpha = \alpha_{rel} \cdot B_0$ states that at most a fraction α_{rel} of the uncontrolled monetary loss may be released. The operator also faces a capacity constraint $E[1\{\pi = \text{review}\}] \le B$ and cares about net value

$$V(\pi) = E[\, 1\{\text{rev}\}(\, q \cdot y \cdot (a + c_{cb}) - c_r - (1-y)c_f \,) \,] - E[\, 1\{\text{app}\} \cdot y \cdot (a + c_{cb}) \,], \tag{2}$$

with c_{cb} a fixed dispute-handling fee, c_r the marginal cost of a review and c_f the friction cost of challenging a legitimate customer. Risk control and net value are distinct objectives: (1) is a constraint the operator must be able to certify, (2) is what the business optimises subject to it.

B. Verification latency

Let D_i be the delay between transaction i and the arrival of its chargeback. At a data cut-off the observed label is $\tilde{y}_i = y_i \cdot 1\{D_i \le u_i\}$, where u_i is the age of transaction i. Fraud labels are censored in one direction only: an undisputed fraud is indistinguishable from a genuine transaction, so $E[\tilde{y}_i] = y_i \cdot G(u_i)$ with G the delay CDF. Production sample layers typically materialise labels at a fixed age Δ, in which case $G(\Delta) = m$ is a single number we call the label maturity. Substituting $\tilde{y}$ for y in (1) understates R by exactly the factor m, so a threshold calibrated naively releases roughly $1/m$ times the intended fraud value—a 67% overshoot at $m = 0.6$ and a 3.3× overshoot at $m = 0.3$.

4. The LARC Calibration Layer

A. Value-ordered triage

Let g(x,a) be any index by which transactions are ranked for review, with the top-ranked cases sent to analysts. The following says that among all indices, expected fraud value is the one that buys the most risk reduction per unit of review capacity.

Proposition 1. Fix a review budget B and let $\eta(x) = P(y = 1 \mid x)$. Among all measurable review sets S with $P((x,a) \in S) \le B$, the set that minimises R is $S^* = \{(x,a) : a \cdot \eta(x) \ge \lambda\}$ for the λ that makes the constraint tight.

Proof. Reviewing (x,a) removes $a \cdot \eta(x)/\bar{a}$ from R at a capacity cost of dP; every element of S costs the same. The problem is a fractional knapsack with uniform weights, whose greedy solution ranks items by value density $a \cdot \eta(x)$. An exchange argument gives optimality: if $(x,a) \in S$ with $a \cdot \eta(x) < a' \cdot \eta(x')$ for some $(x',a') \notin S$, swapping them lowers R without changing capacity. ■

Since c_cb is constant, ranking by $a \cdot \eta(x)$ and by $\eta(x) \cdot (a + c_{cb})$ coincide; we use

$$g(x,a) = s(x) \cdot (a + c_{cb}) \tag{3}$$

so that the index is denominated in dollars at stake and is directly comparable to the economic quantities in (2). Ranking by s alone is optimal only when amounts are independent of fraud probability, which payment data plainly violate.

B. Exposure capping

CRC requires the loss to be bounded, and the bound enters the threshold rule additively. Transaction amounts are heavy-tailed—the largest amount in our real dataset is 291 times the mean—so using the empirical maximum makes the correction term larger than any risk target an operator would set, and the procedure degenerates to reviewing everything. We therefore introduce an exposure cap A_cap, set to the 99th percentile of historical amounts, and route every transaction above it to mandatory review. This is not only a technical device: high-value step-up is standard practice in payments, and it costs about 1% of volume.

Lemma 1. Under the mandatory-review rule $a > A_{cap}$, the per-transaction loss $\ell = a \cdot y \cdot 1\{\text{approve}\}/\bar{a}$ satisfies $0 \le \ell \le A_{cap}/\bar{a}$ almost surely, and the bounded loss coincides with the true monetary loss on the auto-approved region, so no approximation is introduced.

The proof is immediate: approval implies $a \le A_{cap}$, on which $\min(a, A_{cap}) = a$. Capping the loss would normally bias the risk estimate downward; pairing the cap with mandatory review removes the bias exactly.

C. Conformal monetary risk control

Given a calibration sample of size n with losses ℓ_i and indices g_i, and writing $\hat{R}_n(\tau) = n^{-1} \sum_i \ell_i \, 1\{g_i < \tau\}$, we choose the largest auto-approve threshold that satisfies

$$\left(n \cdot \hat{R}_n(\tau) + \mathrm{Bnd} \right) / (n + 1) \le \alpha \,, \quad \mathrm{Bnd} = A_{cap}/\bar{a} \,. \tag{4}$$

$\hat{R}_n$ is non-decreasing in τ, so the rule is an instance of conformal risk control with $\lambda = -\tau$, and Theorem 1 of [1] gives $E[\ell(x_{n+1}, y_{n+1}; \hat{\tau})] \le \alpha$ whenever the calibration points and the test point are exchangeable. Recent financial evidence reinforces the practical value of coupling predictive models with conformal calibration: Chen et al. [22] integrate recurrent forecasting with volatility-adaptive and temporal conformal mechanisms, attaining 95.2% empirical coverage at a 95% target while maintaining competitive point accuracy and strong extreme-event detection. Their results provide a compelling financial precedent for separating predictive modeling from calibrated uncertainty control. LARC carries that principle to the decision layer, where the controlled object is expected monetary loss rather than a prediction interval. The guarantee here is on the expectation of the realised risk, not on its per-deployment realisation; because a single large fraud can dominate one window, we report both the mean realised risk and the frequency with which individual replicates exceed α.

D. Amount stratification and budget allocation

A single global τ is wasteful. Scorers are not equally calibrated across the amount range: Fig. 4 (left) shows the ratio of mean score to empirical fraud rate varying by a factor of six across amount strata on real data. A global threshold therefore over-reviews the bands where the scorer is pessimistic and under-protects those where it is optimistic.

We partition amounts into K quantile strata and give each its own threshold τ_k with its own bound $\mathrm{Bnd}_k = \min(\mathrm{edge}_{k+1}, A_{cap})/\bar{a}$. If the per-stratum risks satisfy $R_k \le \alpha_k$ then $R = \Sigma_k p_k R_k \le \Sigma_k p_k \alpha_k$, so any allocation with $\Sigma_k p_k \alpha_k \le \alpha$ preserves marginal validity. The allocation is chosen to minimise total review load:

$$\min_{\tau_1..\tau_K} \Sigma_k p_k \cdot \mathrm{Rev}_k(\tau_k) \quad \text{s.t.} \quad \Sigma_k p_k \cdot \hat{R}_k(\tau_k) \le \alpha \,. \tag{5}$$

Each stratum contributes a finite frontier of achievable (risk, review-rate) pairs—one per calibration point—so (5) is a multiple-choice knapsack. We solve its Lagrangian relaxation $\min_k [\mathrm{Rev}_k + \mu \cdot R_k]$ and bisect on μ, which equalises the marginal review cost of risk reduction across strata (water-filling). The result, shown in Fig. 4 (right), is that review capacity is withdrawn almost entirely from the low-amount strata and concentrated on the high-exposure band.

E. Latency-corrected calibration

Let $w_i = 1/G(u_i)$ and define the corrected loss $\ell_i = w_i \cdot a_i \cdot \tilde{y}_i / \bar{a}$. Since $E[\tilde{y}_i] = y_i \cdot G(u_i)$, we have $E[\ell_i] = E[\ell_i]$: the corrected empirical risk is unbiased for the true monetary risk even though every calibration label is censored. Substituting ℓ for ℓ in (4) and inflating the bound to Bnd/G_{min}, where G_{min} is the smallest maturity retained, preserves the CRC guarantee because ℓ is again a bounded monotone loss. Under a fixed-age snapshot this is simply $w = 1/m$; under variable ages we retain calibration cases with $G(u_i) \ge G_{min} = 0.5$ and weight them individually.

The correction has a price, and it is worth stating precisely. The additive term in (4) becomes $\mathrm{Bnd}/(G_{min}(n+1))$, so a target is attainable only if

$$\alpha \ge A_{cap} / (\bar{a} \cdot G_{min} \cdot (n + 1)) \,. \tag{6}$$

This is a sharp and practically useful statement: the tightest monetary risk a deployment can certify is set by the ratio of its exposure cap to its calibration size, divided by the label maturity. On our real dataset with $n = 85{,}442$ and $m = 0.6$ the floor is \$21.4 per thousand transactions,

which is why targets below about 11% of the uncontrolled loss are unreachable there and why the same relative target is comfortably feasible on the larger-loss second dataset.

F. Algorithm

Calibration (offline, per refresh): compute g_i from (3) with g = ∞ for a_i > A_cap; form ℓ_i; for each stratum k sort by g and build the cumulative (risk, review-rate) frontier; bisect on μ to solve (5); emit τ_1..τ_K. Serving (online, per transaction): review if a > A_cap or g(x,a) ≥ τ_{k(a)}, else approve. Calibration is O(n log n); serving is a comparison against one threshold and adds no measurable latency to an authorisation path.

5. Experimental Setup

A. Datasets

Table I. Dataset and Deployment Characteristics

	D1 (ULB)	D2 (Sparkov)
Transactions	**284,807**	**272,297**
Frauds (rate)	**492 (0.173%)**	**1,441 (0.529%)**
Span	**2 days**	**731 days**
Mean / max amount	**$88.35 / $25,691**	**$70.93 / $18,149**
Train / pool size	**113,922 / 142,404**	**108,918 / 136,149**
Scorer	**logistic reg.**	**LightGBM**
AUC / AP	**0.890 / 0.615**	**0.886 / 0.198**
Exposure cap A_cap	**$1,098**	**$540**
Mandatory review	**0.85%**	**1.11%**
Uncontrolled loss B_0	**$201.12 / 1k txn**	**$2,577.23 / 1k txn**
Risk target α	**$40.22 / 1k txn**	**$128.86 / 1k txn**

D1 (ULB) contains 284,807 real card transactions made by European cardholders over two days in September 2013, of which 492 (0.173%) are fraudulent [19]. Features are 28 PCA components plus the amount; the mean amount is $88.35 and the maximum $25,691.

D2 (Sparkov) is a 24-month stream of 272,297 transactions from 150 simulated cardholders generated with the open-source Sparkov generator, containing 1,441 frauds (0.529%). It supplies what D1 cannot—calendar-time structure over 731 days, merchants, categories and geography—so that chargeback delays can be expressed in days rather than as a fraction of a short window. Monthly fraud rates range from 0.066% to 0.95%, giving genuine non-stationarity. We derive 22 features including per-card expanding amount statistics, 1-hour and 24-hour velocity counts, and the haversine distance between cardholder and merchant, all computed causally.

B. Protocol

For each dataset the first 40% of the stream by time trains the scorer, the next 10% forms a fully matured but older calibration reservoir used by the stale-calibration baseline, and the final 50% is the deployment pool. The pool is split at random into 60% calibration and 40% test; we repeat this over 10 splits and 3 independent label-snapshot draws, for 30 replicates per configuration. Training and calibration labels are censored at maturity m (default 0.6); test evaluation uses the eventually observed labels. Economics are c_cb = $15, c_r = $2.50, c_f = $0.50, q = 0.90. The exposure cap is the 99th percentile of training amounts ($1,098 on D1, $540 on D2, implying 0.85% and 1.11% mandatory review), K = 4 strata, and the default risk targets are α_rel = 0.20 on D1 and 0.05 on D2, the tightest values that (6) admits at these calibration sizes. Scorers are ℓ2-regularised logistic regression with balanced class weights on D1 (AUC 0.890, AP 0.615) and LightGBM [23] on D2 (AUC 0.886, AP 0.198).

C. Baselines

Top-k alert rate fixes the review rate at 2% and is the dominant industry practice. Cost-sensitive Bayes applies Elkan's rule [5], reviewing whenever the expected recovery exceeds the review cost, with no capacity constraint. Count-CRC applies conformal risk control to the number of leaked frauds rather than their value. CRC (score order) and CRC (value order) apply (4) globally without stratification. CRC (stale mature calibration) calibrates the full LARC procedure on the older reservoir, whose labels are fully mature but whose distribution is stale. Two matched-budget variants rank by score or by value and are given exactly the review rate LARC chose in the same replicate, isolating the effect of the ordering. The oracle calibrates LARC on the true, uncensored labels and bounds what any latency correction could achieve.

We report the risk ratio (mean realised risk divided by α; at most 1 means the constraint holds), the fraction of replicates exceeding α, the review rate, dollar recall, and net value per thousand transactions. Significance is assessed with paired Wilcoxon signed-rank tests over the 30 replicates.

6. Results

A. Main comparison

Table II. Main Comparison at the Default Risk Target

Method	Risk / α	Viol. %	Rev. %	$Rec. %	NV / 1k
D1 (ULB), α = $40.22 / 1k txn					
Top-k (score)	**1.20 ± 0.53**	**57**	**2.00**	**74.0**	**84.7**
Top-k (value)	**1.07 ± 0.38**	**53**	**2.00**	**76.8**	**76.5**
Cost-sens. Bayes	**0.03 ± 0.05**	**0**	**35.99**	**99.4**	**−890.1**
Count-CRC	**1.89 ± 0.75**	**87**	**1.07**	**60.0**	**84.9**
CRC (score)	**1.20 ± 0.80**	**57**	**4.30**	**74.7**	**15.8**
CRC (value)	**1.21 ± 0.66**	**50**	**2.03**	**74.0**	**70.2**
CRC (stale mature)	**0.29 ± 0.22**	**0**	**28.96**	**93.5**	**−689.5**
Top-k (score) @ LARC	**0.94 ± 0.53**	**43**	**3.67**	**79.6**	**45.0**
Top-k (value) @ LARC	**0.70 ± 0.45**	**27**	**3.69**	**84.5**	**40.9**
LARC − latency corr.	**1.18 ± 0.70**	**53**	**1.96**	**74.9**	**74.3**
LARC (ours)	**0.61 ± 0.42**	**13**	**3.68**	**86.5**	**46.0**
Oracle thresholds	**0.76 ± 0.48**	**27**	**2.55**	**83.6**	**73.3**
D2 (Sparkov), α = $128.86 / 1k txn					
Top-k (score)	**2.91 ± 0.30**	**100**	**2.02**	**85.2**	**1922.2**
Top-k (value)	**2.37 ± 0.43**	**100**	**2.01**	**87.9**	**1987.3**
Cost-sens. Bayes	**2.17 ± 0.56**	**100**	**2.37**	**89.0**	**2001.2**
Count-CRC	**0.46 ± 0.20**	**0**	**24.38**	**97.7**	**1558.9**
CRC (score)	**1.61 ± 0.41**	**90**	**8.08**	**91.8**	**1902.1**
CRC (value)	**1.63 ± 0.39**	**93**	**4.59**	**91.8**	**2002.8**
CRC (stale mature)	**0.82 ± 0.19**	**17**	**4.03**	**95.9**	**2115.8**
Top-k (score) @ LARC	**2.36 ± 0.37**	**100**	**4.11**	**88.0**	**1928.5**
Top-k (value) @ LARC	**1.68 ± 0.33**	**100**	**4.12**	**91.5**	**2010.2**
LARC − latency corr.	**1.43 ± 0.41**	**80**	**2.89**	**92.8**	**2075.2**
LARC (ours)	**0.89 ± 0.33**	**37**	**4.09**	**95.5**	**2104.8**
Oracle thresholds	**0.90 ± 0.27**	**40**	**4.01**	**95.4**	**2106.1**

Mean ± s.d. over 30 replicates. Risk / α ≤ 1 means the monetary constraint holds; Viol. % is the share of individual replicates above α. NV is net value in dollars per thousand transactions relative to approving everything. "@ LARC" rows are given LARC's realised review rate.

Table II reports the default operating point. On both datasets LARC is the only method other than the oracle and the two degenerate over-review policies that keeps the mean realised risk below the target, and it does so while reviewing 3.7% (D1) and 4.1% (D2) of volume. All differences in risk ratio against LARC are significant at $p < 10^{-4}$ except for the oracle on D2 ($p = 0.35$) and stale calibration on D2 ($p = 0.27$).

The failures are instructive and distinct. A fixed 2% alert rate releases 1.20× and 2.91× the target because it has no mechanism linking the alert rate to a monetary objective. Cost-sensitive Bayes is the mirror image on D1: ignoring capacity, it reviews 36% of transactions, drives risk almost to zero and destroys $890 of net value per thousand transactions. Count-CRC controls the number of leaked frauds but not their value—on D1 it still overshoots the dollar target by 1.89× and on D2 it meets the dollar target only by reviewing 24% of volume, a six-fold-larger workload than LARC for lower net value. Uncorrected conformal control, whether stratified or not, overshoots by 1.18–1.63× because its calibration labels are immature.

Stale-but-mature calibration is the most interesting failure because it is the workaround practitioners most often adopt. Its error is not systematic but unpredictable: on D2 it lands at 0.82, while on D1 it becomes so conservative that it reviews 29.0% ± 20.1% of transactions and turns a positive net value into −$689 per thousand. Waiting for labels to mature removes the censoring bias but replaces it with a distribution shift whose direction and size are not known at calibration time.

Among the policies that satisfy the constraint, LARC has the highest net value on D1 ($46.0 per thousand versus −$689 and −$890 for the two over-reviewing alternatives) and is within 0.1% of the oracle on D2 ($2,104.8 versus $2,106.1). Its dollar recall—the share of fraud value routed to review—is the highest of all non-degenerate policies at 86.5% and 95.5%.

B. Sensitivity to the risk target

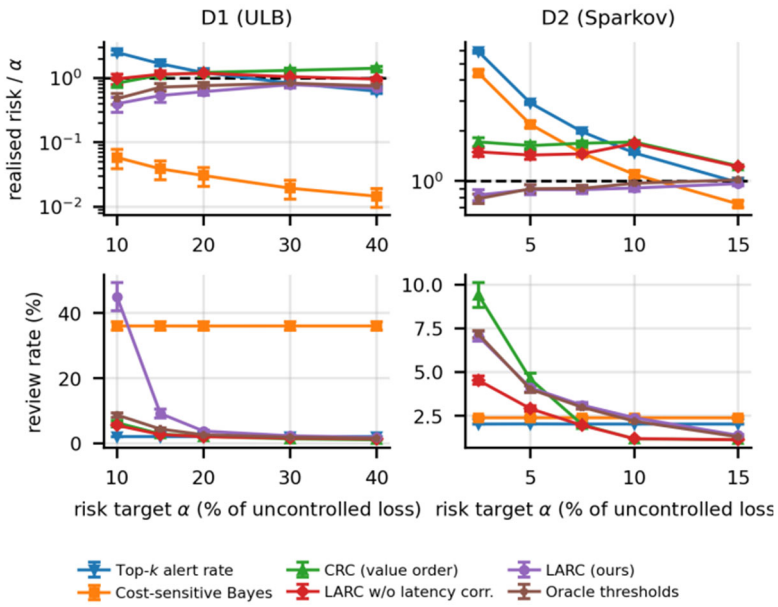

Figure 1. Realised risk relative to the target (top) and review rate (bottom) as the risk target is tightened. Points below the dashed line satisfy the constraint. LARC tracks the oracle across the whole range; baselines cross the line as soon as the target becomes demanding. The rise in LARC's review rate at the far left of D1 is the feasibility boundary of (6).

Fig. 1 sweeps α_rel. The pattern is consistent across both streams: loose targets are met by almost anything, and the baselines separate from LARC exactly as the target tightens. On D2, LARC holds the constraint at every target from 2.5% to 15% of the uncontrolled loss, while the fixed alert rate rises to 2.9× and global CRC to 1.6×. The left-most D1 point illustrates condition

(6) in action: at α_rel = 0.10 the target falls below the certifiable floor of $21.4 per thousand and LARC responds by reviewing 45% of volume rather than issuing a threshold it cannot justify—conservative, but never silently invalid.

C. Verification latency

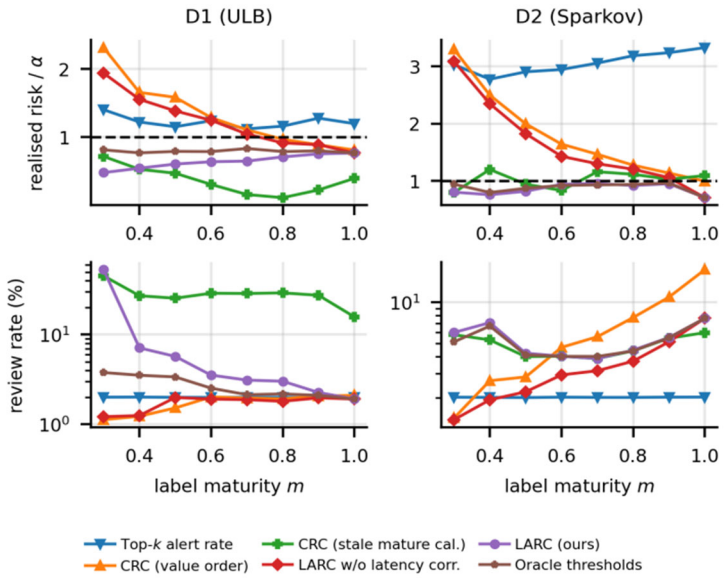

Figure 2. Effect of label maturity m. Uncorrected conformal control degrades monotonically as labels become less mature, closely following the predicted 1/m law; LARC remains below the target throughout. At m = 1 the three curves for LARC, LARC without correction and the oracle coincide, as they must.

Fig. 2 is the central experiment. As maturity falls from 1.0 to 0.3, global CRC degrades from 0.81 to 2.31 on D1 and from 1.00 to 3.29 on D2. The theory of Section III-B predicts a 1/m inflation, i.e. 3.33× at m = 0.3; the measured D2 value of 3.29 matches it. LARC stays within 0.48–0.77 (D1) and 0.71–0.96 (D2) over the entire range. At m = 1 the corrected and uncorrected procedures and the oracle produce identical numbers (0.709 on D2), a consistency check on the implementation.

The cost of the correction is visible in the lower panels: LARC reviews more than its uncorrected counterpart, and the gap widens as labels become less mature, because the bound inflation in (6) grows as 1/m. This is the honest exchange rate between label freshness and review capacity, and it is precisely what an operator needs in order to decide how long to wait before refreshing thresholds.

D. Where the gains come from

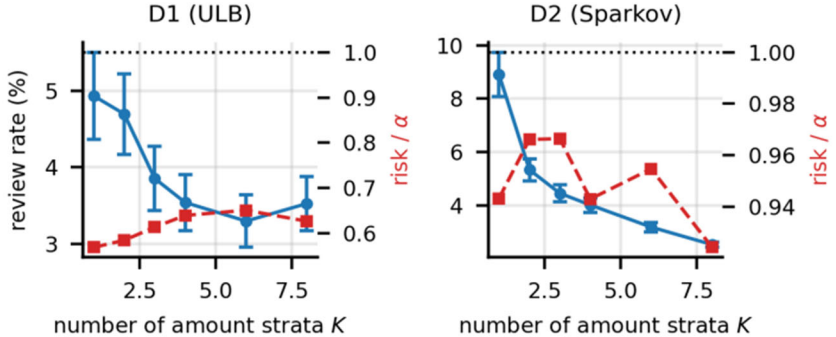

Figure 3. Review rate (left axis, solid) and risk ratio (right axis, dashed) as the number of amount strata grows. Validity is maintained throughout while the required review load falls steadily.

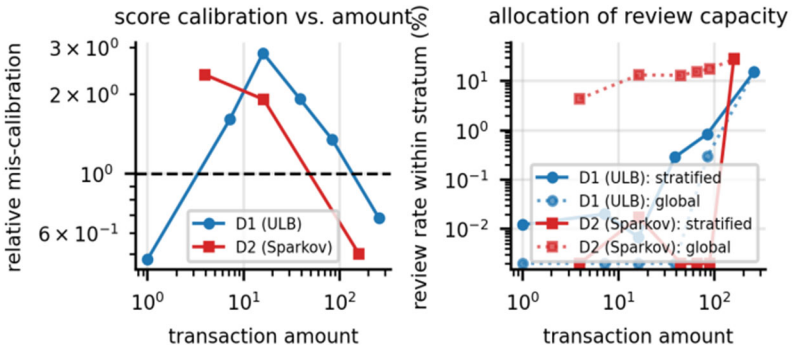

Figure 4. Left: the ratio of mean score to empirical fraud rate, normalised by its global value, varies by roughly a factor of six across amount strata—the scorer is not equally calibrated across the amount range. Right: the review capacity that the Lagrangian allocation assigns to each stratum, against a single global threshold.

Two ablations isolate the first two contributions. Ordering: at exactly the review rate LARC selected, ranking by expected fraud value instead of by score reduces the risk ratio from 0.94 to 0.70 on D1 (−25%) and from 2.36 to 1.68 on D2 (−29%), confirming Proposition 1 on data.

Stratification: Fig. 3 varies K from 1 (a single global threshold) to 8 at fixed α. On D2 the review rate falls from 8.91% to 2.53%, a 72% reduction in analyst workload, while the risk ratio stays between 0.92 and 0.97; on D1 it falls from 4.93% to 3.53% (−29%) with the ratio between 0.57 and 0.65. Validity is never traded away for efficiency. Fig. 4 explains the mechanism: a global threshold reviews between 4.4% and 17.7% within every amount stratum, whereas the stratified allocation withdraws capacity from the five low-amount strata almost entirely and spends 28.7% of the top stratum, which is where the dollars are.

E. Variable label age and scorer robustness

Table III. Variable Label Age (Weibull Chargeback Delays)

Method	Risk / α	Rev. %	NV / 1k	Viol. %
D1 (ULB), α_rel = 0.40				
Top-k (score)	**0.60**	**2.00**	**84.7**	**7**
CRC (value)	**1.33**	**1.14**	**43.8**	**77**
CRC (stale mature)	**0.70**	**1.53**	**79.5**	**10**
LARC – latency corr.	**1.01**	**1.34**	**62.8**	**37**
LARC (ours)	**0.51**	**2.06**	**76.8**	**3**
Oracle thresholds	**0.72**	**1.52**	**77.6**	**13**
D2 (Sparkov), α_rel = 0.05				
Top-k (score)	**2.91**	**2.00**	**1922.2**	**100**
CRC (value)	**1.59**	**4.59**	**2005.2**	**90**
CRC (stale mature)	**0.82**	**4.02**	**2115.8**	**17**
LARC – latency corr.	**1.33**	**3.08**	**2081.0**	**73**
LARC (ours)	**0.63**	**4.98**	**2109.3**	**7**
Oracle thresholds	**0.90**	**4.01**	**2106.1**	**40**

Each calibration case carries its own maturity G(u_i); LARC keeps cases above a 50% maturity floor. D1 uses a looser target because condition (6) makes α_rel = 0.20 unattainable once the bound is inflated.

Table III relaxes the fixed-age snapshot so that each calibration case carries its own maturity G(u_i), drawn from a Weibull chargeback-arrival model. LARC retains cases above a 50% maturity floor and weights them individually. On D2 it reaches a risk ratio of 0.63 with the highest net value of any policy; on D1, condition (6) forces a looser target, and at α_rel = 0.40 LARC again controls risk at 0.51 while uncorrected control sits at 1.01 and global CRC at 1.33.

Table IV swaps the scorer for logistic regression, LightGBM, a random forest and an MLP. Scorer quality varies widely—AUC from 0.740 to 0.954—yet LARC keeps the risk ratio below 1 in all eight dataset-scorer combinations (0.49–0.64 on D1, 0.87–0.97 on D2), while the fixed alert rate and global CRC violate the constraint in seven and eight of them respectively. This model-agnostic stress test is especially relevant in light of recent adaptive fraud-scoring systems such as the drift-gated stable–plastic temporal GNN of Yue et al. [24], which uses delayed prequential feedback and selective online updates under a strict chronological predict-then-update protocol and demonstrates effective adaptation to evolving fraud patterns. LARC is designed to benefit directly from such advances without tying its validity to any one scorer family. The price of a weak scorer is paid in review capacity, not in validity: with the worst scorer on D1 (AUC 0.740) LARC still controls risk but needs 19.1% review, against 3.9% with the best. This separation is the practical point of a calibration layer—the guarantee does not depend on the model being good, only the cost of honouring it does.

F. Discussion

Table IV. Risk Ratio Across Scorers

Scorer	AUC	Top-k	CRC (val.)	LARC −lat.	LARC	Rev.%
D1 (ULB)						
Logistic reg.	**0.927**	**1.34**	**1.41**	**1.39**	**0.64**	**3.9**
LightGBM	**0.740**	**2.64**	**1.11**	**0.91**	**0.49**	**19.1**
Random forest	**0.954**	**0.97**	**1.18**	**1.02**	**0.49**	**15.7**
MLP	**0.910**	**1.46**	**1.34**	**1.26**	**0.62**	**11.7**
D2 (Sparkov)						
Logistic reg.	**0.954**	**2.57**	**1.65**	**1.50**	**0.88**	**2.1**
LightGBM	**0.906**	**2.97**	**1.69**	**1.48**	**0.97**	**4.1**
Random forest	**0.895**	**1.71**	**1.71**	**1.63**	**0.95**	**2.1**
MLP	**0.897**	**2.51**	**1.57**	**1.50**	**0.87**	**4.9**

Rev.% is LARC's review rate. LARC holds the constraint for every scorer; a weaker scorer is paid for in review capacity, not in validity.

Three observations generalise beyond these datasets. First, the certifiable risk floor (6) means that monetary risk control is a large-calibration-set technique: halving the exposure cap or doubling the calibration window each halve the floor, and both are levers an operator controls. Second, the conformal guarantee is an expectation, so individual windows can and do exceed α; operators who need per-window guarantees should expect to pay additional review capacity for a high-probability variant. Third, the exchangeability assumption between the calibration and deployment windows is an approximation under drift; the stale-calibration results quantify what happens when it is violated badly, and the fact that LARC calibrated on fresh data outperforms it suggests that correcting censoring is the better trade than waiting for maturity. The upstream scoring literature is also becoming more robust to strategic change: Fan et al. [25] introduce game-theoretic anticipatory continual graph learning with adversarial motif memory, reporting materially stronger resilience to adaptive structural attacks while keeping forgetting low on the studied dynamic-graph benchmarks. Such progress makes a downstream model-agnostic risk-control layer especially valuable: stronger adaptive scorers can reduce the operational review burden required to honor the same monetary risk target.

7. Conclusion

We treated payment fraud decisioning as a monetary risk-control problem and built a calibration layer that certifies how much fraud value an engine releases, under a review budget and with labels that have not finished arriving. The three ingredients—value-ordered triage, amount-stratified conformal control made feasible by an exposure cap, and an IPCW correction for verification latency—each carry measurable weight: 25–29% less risk at equal review capacity, up to 72% less review capacity at equal risk, and a guarantee that survives down to 30% label maturity where uncorrected calibration overshoots threefold. Across two payment streams, four scorers and five risk targets, LARC is the only method tested that holds its monetary target while remaining economically competitive. The layer is model-agnostic, costs one threshold comparison at serving time, and gives an operator a statement about dollars that can be put in a control document.

References

[1] Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., & Schuster, T. (2024). Conformal risk control. In Proceedings of the International Conference on Learning Representations.

[2] Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2018). Credit card fraud detection: A realistic modeling and a novel learning strategy. IEEE Transactions on Neural Networks and Learning Systems, 29(8), 3784–3797. https://doi.org/10.1109/TNNLS.2017.2766186

[3] Carcillo, F., Dal Pozzolo, A., Le Borgne, Y.-A., Caelen, O., Mazzer, Y., & Bontempi, G. (2018). SCARFF: A scalable framework for streaming credit card fraud detection with Spark. Information Fusion, 41, 182–194. https://doi.org/10.1016/j.inffus.2017.09.001

[4] Correa Bahnsen, A., Aouada, D., Stojanovic, A., & Ottersten, B. (2016). Feature engineering strategies for credit card fraud detection. Expert Systems with Applications, 51, 134–142. https://doi.org/10.1016/j.eswa.2015.12.030

[5] Elkan, C. (2001). The foundations of cost-sensitive learning. In Proceedings of the 17th International Joint Conference on Artificial Intelligence (pp. 973–978).

[6] Chen, J., Liang, Y., Liu, J., & Zhou, M. (2026). Temporal transformer with conditional tabular GAN for credit card fraud detection: A sequential deep learning approach. Mathematics, 14(7), Article 1183. https://doi.org/10.3390/math14071183

[7] Vovk, V., Gammerman, A., & Shafer, G. (2005). Algorithmic learning in a random world. Springer.

[8] Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. Journal of the American Statistical Association, 113(523), 1094–1111. https://doi.org/10.1080/01621459.2017.1319823

[9] Bates, S., Angelopoulos, A. N., Lei, L., Malik, J., & Jordan, M. I. (2021). Distribution-free, risk-controlling prediction sets. Journal of the ACM, 68(6), 1–34. https://doi.org/10.1145/3478530

[10] Angelopoulos, A. N., Bates, S., Candès, E. J., Jordan, M. I., & Lei, L. (2022). Learn then test: Calibrating predictive algorithms to achieve risk control. In Proceedings of the 39th International Conference on Machine Learning (Vol. 162, pp. 986–1011). PMLR.

[11] Sadinle, M., Lei, J., & Wasserman, L. (2019). Least ambiguous set-valued classifiers with bounded error levels. Journal of the American Statistical Association, 114(525), 223–234. https://doi.org/10.1080/01621459.2018.1424625

[12] Papadopoulos, H., Proedrou, K., Vovk, V., & Gammerman, A. (2002). Inductive confidence machines for regression. In Proceedings of the European Conference on Machine Learning (pp. 345–356).

[13] Tibshirani, R. J., Foygel Barber, R., Candès, E. J., & Ramdas, A. (2019). Conformal prediction under covariate shift. In Advances in Neural Information Processing Systems (Vol. 32, pp. 2530–2540).

[14] Gibbs, I., & Candès, E. J. (2021). Adaptive conformal inference under distribution shift. In Advances in Neural Information Processing Systems (Vol. 34, pp. 1660–1672).

[15] Foygel Barber, R., Candès, E. J., Ramdas, A., & Tibshirani, R. J. (2023). Conformal prediction beyond exchangeability. The Annals of Statistics, 51(2), 816–845. https://doi.org/10.1214/23-AOS2276

[16] Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association, 89(427), 846–866. https://doi.org/10.1080/01621459.1994.10476818

[17] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), Article 44. https://doi.org/10.1145/2523813

[18] Jiao, Y., Fan, H., Yue, X., Ping, W., Sun, T., & Wang, J. (2026). Dynamic heterogeneous graph contrastive learning for uncovering collusive financial fraud. Scientific Reports. https://doi.org/10.1038/s41598-026-58938-5

[19] Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. In Proceedings of the IEEE Symposium Series on Computational Intelligence (pp. 159–166).

[20] Lopez-Rojas, E. A., Elmir, A., & Axelsson, S. (2016). PaySim: A financial mobile money simulator for fraud detection. In Proceedings of the 28th European Modeling & Simulation Symposium (pp. 249–255).

[21] Jesus, S., Pombal, J., Alves, D., Cruz, A., Saleiro, P., Ribeiro, R. P., Gama, J., & Bizarro, P. (2022). Turning the tables: Biased, imbalanced, dynamic tabular datasets for ML evaluation. In Advances in Neural Information Processing Systems (Vol. 35).

[22] Chen, Z., Wang, M., Zeng, Z., & Ping, W. (2026). Uncertainty-aware financial forecasting: Leveraging conformal prediction for risk-adjusted models. IEEE Access, 14, 90053–90077. https://doi.org/10.1109/ACCESS.2026.3702666

[23] 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 Advances in Neural Information Processing Systems (Vol. 30, pp. 3146–3154).

[24] Yue, X., Yang, J., Zhu, R., & Xu, Q. (2026). Adaptive temporal graph neural networks for real-time financial fraud detection under concept drift. Highlights in Business, Economics and Management, 69, 239–250.

[25] Fan, H., Jiao, Y., Wang, M., Chen, J., & Yang, S. (2026). Fraud learns too: Continual graph learning under strategic adversarial drift in dynamic networks. Scientific Reports. https://doi.org/10.1038/s41598-026-60997-7

[26] Dal Pozzolo, A., Caelen, O., Le Borgne, Y.-A., Waterschoot, S., & Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. Expert Systems with Applications, 41(10), 4915–4928. https://doi.org/10.1016/j.eswa.2014.02.026

[27] Vovk, V. (2012). Conditional validity of inductive conformal predictors. In Proceedings of the Asian Conference on Machine Learning (Vol. 25, pp. 475–490). PMLR.