Risk-Aware Adaptive Verification for Digital Onboarding: A Cost-Sensitive Contextual Bandit Framework with Calibrated Uncertainty

Tianjun Mo[1], Enmiao Feng[1,*], Gaike Wang[2], Pengfei Li[1]

[1] Duke University, Durham, NC 27708, USA

[2] New York University, New York, NY 10012, USA

[*]Corresponding Email: enmiaofeng@ieee.org

Abstract

Digital onboarding forces a bank to route every account-opening application through a ladder of identity-verification treatments whose operating costs, interdiction power and customer friction differ by orders of magnitude. The routing policy must be learned from bandit feedback-only the outcome of the treatment actually applied is ever observed-under extreme class imbalance, heavy-tailed monetary stakes, hard analyst-capacity limits and adversaries that adapt. We formulate risk-aware adaptive verification as a cost-sensitive contextual bandit with a structured reward decomposition and propose CARVE, which (i) turns a gradient-boosted risk score into an interval-valued calibrated probability using a bootstrap-aggregated Venn–ABERS predictor, (ii) pools treatment-effect regressions across the verification ladder so that the few hundred interdiction observations available at 1% fraud prevalence become usable, (iii) replaces the parametric LinUCB confidence radius with a block-aggregated conformal radius that measures epistemic rather than aleatoric uncertainty, and (iv) treats the two uncertainty sources asymmetrically-optimism over learnable treatment effects to drive exploration, calibrated pessimism over fraud risk to bound exposure-while a Lagrangian dual variable holds the manual-review rate at a prescribed capacity. On the NeurIPS-2022 Bank Account Fraud data, CARVE reduces expected cost per application by 19.6% relative to a tuned banded policy and by 9.7% relative to an offline cost-sensitive policy ($p<0.01$, 15 paired streams), closing to within 6.3% of a treatment-effect oracle, whereas unstructured LinUCB and Thompson sampling incur twice its cost. The conclusions transfer to card-not-present step-up authentication on a 285k-transaction real card log and hold across twelve perturbations of the response model.

Keywords

Contextual bandits, cost-sensitive learning, conformal prediction, probability calibration, identity verification, account-opening fraud, off-policy evaluation, concept drift.

1. Introduction

Opening a bank account no longer requires a branch visit. The consequence is that the identity of a remote applicant must be established by software, in seconds, and at a unit cost that is small enough to be spent millions of times per year. Banks therefore operate a verification ladder: an application may be approved instantly, subjected to an automated document and device check, escalated to biometric liveness capture, referred to a human analyst, or declined outright. Each rung buys interdiction power with two currencies. The first is operating expenditure, which ranges from fractions of a cent for a device fingerprint to tens of dollars for an analyst review.

The second, and usually the larger, is customer friction: every additional step increases the probability that a legitimate applicant abandons the funnel, and the lost lifetime value of that customer never appears on the fraud-loss ledger.

The prevailing industrial design is a banded policy. A supervised model produces a risk score, and fixed thresholds map score bands onto rungs of the ladder. This design has three structural weaknesses. First, it conflates a probability with a decision: the thresholds are tuned against population-average assumptions about how effective each treatment is and how much friction it creates, even though both quantities vary strongly across applicants-a mobile-only applicant with a free e-mail address and no prior address history is neither as checkable nor as friction-tolerant as an established customer. Second, the parameters of the ladder are not observable from historical data in the way the risk model is: because the incumbent policy is deterministic, the logs contain almost no information about what would have happened had a different rung been chosen. Third, the environment is adversarial and non-stationary. When synthetic-identity and deepfake toolkits defeat an automated document check, its interdiction rate collapses, but a frozen offline policy keeps prescribing it.

These three weaknesses point to the same formal object: a cost-sensitive contextual bandit. The context is the application, the arms are the rungs of the ladder, the loss is a monetary cost that mixes fraud exposure, operating expenditure and forgone customer value, and the feedback is bandit feedback because only the chosen rung is ever executed. What makes this instance unusual, and what off-the-shelf bandit algorithms handle poorly, is the statistical geometry of the loss. Fraud prevalence in account opening is around 1%; the loss of an undetected fraudulent account is three orders of magnitude larger than the cost of a document check; and interdiction feedback is observed only on the intersection of the fraudulent minority with the verified subset. A direct regression of realised cost on context and action-the standard LinUCB or linear Thompson sampling construction-must estimate a mean whose per-observation standard deviation exceeds the differences it is asked to resolve by two orders of magnitude. We show empirically that such learners never leave the exploration regime.

This paper argues that the way out is to exploit the accounting structure of the problem and to be careful about which uncertainty is used for what. Our contributions are as follows.

- We formulate risk-aware adaptive verification as a budget-constrained cost-sensitive contextual bandit and give a reward decomposition that separates the calibrated fraud probability, which is estimable from abundant offline labels, from the treatment-effect functions, which are only estimable from bandit feedback (Section III).
- We propose CARVE, a decision layer with four components: a bootstrap-aggregated Venn–ABERS calibrator that returns an interval-valued risk; treatment-effect heads that pool context slopes across the ladder so that the scarce interdiction observations are usable; a block-aggregated conformal confidence radius that calibrates the exploration bonus to epistemic rather than aleatoric variability; and an asymmetric uncertainty rule that is optimistic about treatment effects and pessimistic about risk. A dual variable enforces an analyst-capacity constraint online (Section IV).
- We evaluate on two real datasets—the NeurIPS-2022 Bank Account Fraud suite and the ULB credit-card transaction log—under an explicitly documented semi-synthetic response model, and report a full component ablation, a risk-aversion frontier, capacity control, a twelve-regime sensitivity study, a differential-friction audit across a protected attribute, and an off-policy evaluation study that asks whether the policy could have been screened before deployment (Sections V–VI).
- We report negative results as well: unstructured cost regression is statistically infeasible in this regime, and the soft capacity constraint converges too slowly when the budget is very tight (Section VII).

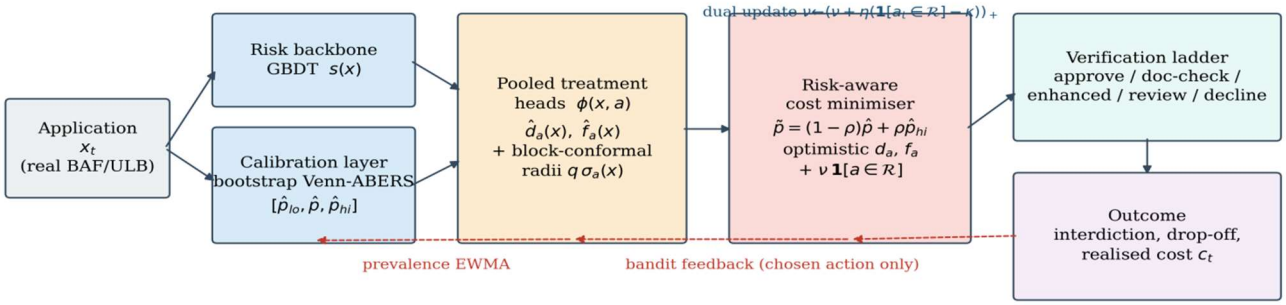


Fig. 1 CARVE architecture. Solid arrows: the online decision path for one application. Dashed arrows: bandit feedback, which updates the pooled treatment-effect heads, the block-conformal radii and the running prevalence estimate. Only the outcome of the executed rung is ever observed.

2. Related Work

2.1. Cost-sensitive Learning for Financial Fraud

Cost-sensitive classification replaces the 0–1 loss with a cost matrix; Elkan formalised the threshold-shifting view for class-dependent costs and the associated reductions to weighted classification [1]. Financial applications quickly moved to example-dependent costs, because the loss of a missed fraud is the exposed amount rather than a constant: cost-sensitive decision trees [2], [3], Bayes-minimum-risk decision rules on calibrated scores [4] and cost-sensitive feature engineering [5] all optimise savings rather than accuracy. Ding et al. push the cost matrix itself into the learning loop: their GT-CSAT framework derives asymmetric misclassification penalties from a two-player zero-sum payoff structure and adapts them online with a multiplicative, Nash-inspired rule, which lets a cloud intrusion detector hold a Pareto-superior robustness–accuracy frontier instead of trading one objective for the other [6]. We find their central design choice persuasive and adopt it in a different guise: the price of a scarce resource should be a control variable the system adjusts online, not a hyper-parameter an engineer guesses (Section IV-F). A parallel line of work widens the evidence base rather than the loss function, and Ping et al.'s TriAudit is an elegant example: it fuses structured accounting ratios, narrative semantics and document-layout geometry in one trimodal attention network and, unusually for a fraud detector, emits a contrastive evidence chain that traces each alert back to the specific ratio, sentence and page region responsible for it [7]. Their finding that layout deviation carries fraud signal independently of textual content is a useful corrective to the habit of treating a financial record as a flat feature vector. This literature is nevertheless confined to a binary accept/investigate decision with known costs. Our setting has five actions whose effects are themselves unknown functions of the context and must be learned from partial feedback.

2.2. Contextual Bandits and Off-policy Learning

LinUCB [8] and linear Thompson sampling [9] are the canonical algorithms for linearly parameterised contextual bandits, with regret guarantees developed by Chu et al. [10] and surveyed by Lattimore and Szepesvári [11]; regression-oracle reductions [12] extend the machinery to general function classes. Empirical comparisons show that the practical ranking of these methods is dominated by how the confidence radius is set rather than by the underlying theory [13], and that greedy or nearly greedy behaviour is often optimal when contexts are sufficiently diverse [14]. Non-stationarity is handled by discounting or sliding windows [15]. Learning and evaluating policies from logged bandit feedback relies on inverse propensity scoring, doubly robust estimation [16] and counterfactual risk minimisation [17].

Resource and knapsack constraints have been studied in the linear contextual setting [18]; our dual-variable capacity control is a practical instance of that idea. We depart from this literature in exploiting a known decomposition of the loss and in calibrating the confidence radius empirically.

2.3. Calibration and Distribution-free Uncertainty

Post-hoc calibration by Platt scaling [19], isotonic regression [20] and beta calibration [21] is standard practice; Niculescu-Mizil and Caruana document which learners need it and how it interacts with the Brier decomposition [22]. Venn–ABERS predictors [23] are a special case of Venn prediction that returns a pair of probabilities and is provably calibrated under exchangeability alone, with efficient inductive and cross variants [24]; the broader conformal framework [25], [26] extends distribution-free guarantees to prediction sets and, more recently, to expected-loss control [27]. Closest in spirit to our calibration layer is the volatility-adaptive conformal framework of Chen et al., which normalises nonconformity scores by a learned multi-scale volatility estimate and widens intervals when an anomaly ensemble signals market stress, holding empirical coverage at the nominal level where fixed-width conformal intervals drift badly [28]. Their demonstration that the right normalisation restores validity under strong heteroscedasticity is precisely the argument we make in Section IV-C, transposed from the volatility of a return to the sampling noise of an estimated mean. We use Venn–ABERS not merely to sharpen the score but as the source of a decision-relevant risk band, and we import the conformal idea into the bandit confidence radius. Prior shift is corrected with the classical odds-rescaling procedure of Saerens et al. [29]; the distinction between aleatoric and epistemic uncertainty that motivates our asymmetric treatment is discussed by Hüllermeier and Waegeman [30].

2.4. Selective Prediction, Deferral and Risk Measures

Routing a subset of cases to a human is studied as learning to defer [31], [32]; realistic benchmarks with synthetic analyst teams have recently appeared for exactly the account-opening task [33]. That literature typically assumes a binary automate/defer choice and a fixed expert. Our ladder is a five-way choice in which every rung, including full automation, has its own interdiction and friction profile. Finally, risk-sensitive objectives in bandits-mean–variance and coherent risk measures [34], with conditional value at risk as the standard financial functional [35]-motivate our tunable pessimism parameter, and the pessimism-for-safety principle mirrors its use in offline reinforcement learning [36]. Concept drift in evolving fraud streams is surveyed in [37]. Fan et al. argue convincingly that in fraud this drift is not exogenous noise but a strategic response, and they act on that premise: their GT-ACGL framework trains a temporal graph detector against a learned generator that anticipates the adversary's next structural move, and couples it with a motif-based replay memory that holds the forgetting rate below six points while still adapting [38]. We adopt their diagnosis—that a frozen detector is permanently one attack behind-as the motivation for the deployment-time regime change in our protocol, although our change point is exogenous by construction rather than endogenously generated.

2.5. Data

We use the Bank Account Fraud suite [39], a privacy-preserving generative resample of a real bank's account-opening log that preserves temporal dynamics and class imbalance, and the ULB credit-card transaction dataset released with the undersampling-calibration study of Dal Pozzolo et al. [40], [41]. Both are among the few public resources with realistic prevalence and, in the latter case, real monetary amounts.

3. Problem Formulation

3.1. The Verification Ladder

Applications arrive as a stream. At round t the decision maker observes a feature vector x_t in R^d describing the application, the applicant and the session, and must choose one action a_t from A = {0,1,2,3,4}, interpreted as instant approval, automated document and device verification, enhanced verification with biometric liveness, manual analyst review, and decline. A latent label y_t in {0,1} indicates whether the application is fraudulent. Two exposures are attached to each application: L(x_t), the loss the bank suffers if a fraudulent application is onboarded undetected, and V(x_t), the value forgone if a legitimate applicant abandons the funnel or is declined.

Each action a has an operating cost c_a, an interdiction function d_a(x) = Pr[fraud detected | y=1, a, x] and a drop-off function f_a(x) = Pr[abandon | y=0, a, x]. Approval interdicts nothing and creates essentially no friction; decline interdicts everything and loses every legitimate applicant it touches. The realised cost of a decision is

$$c(x,y,a) = c_a + y\,(1 - B_d)\,L(x) + (1 - y)\,B_f\,V(x), \tag{1}$$

where B_d ~ Bernoulli(d_a(x)) and B_f ~ Bernoulli(f_a(x)). Taking expectations gives the quantity the policy must minimise,

$$C(a \mid x, p) = c_a + p\,(1 - d_a(x))\,L(x) + (1 - p)\,f_a(x)\,V(x), \tag{2}$$

with p = Pr[y=1 | x]. Equation (2) is the accounting identity that we exploit: it factorises the loss into a risk term, which offline labelled data identify well, and treatment-effect terms, which offline data identify badly.

3.2. Feedback Structure

After acting the decision maker observes the realised cost and its components: whether an impostor was interdicted (only if y=1 and a>0) and whether a genuine applicant dropped out (only if y=0). The label itself becomes known eventually, either at interdiction or through chargebacks and first-party-fraud recovery, and we follow common practice in abstracting away the reporting delay. Crucially, d_a(x) is observable only on the intersection of the fraudulent minority with the verified subset. With 1% prevalence and a realistic exploration budget this intersection contains on the order of 10^2 observations, which is the binding statistical constraint of the problem.

3.3. Capacity Constraint and Objective

Manual review consumes analyst hours, which are fixed in the short run. Let R be the subset of actions that consume review capacity. Over a horizon T the policy must satisfy a budget on the long-run review rate, so the problem is

$$\min E\left[\frac{1}{T}\sum_t c(x_t, y_t, a_t)\right] \quad \text{s.t.} \quad \frac{1}{T}\sum_t 1[a_t \in R] \le \kappa. \tag{3}$$

We report performance as expected cost per application and as cumulative regret against a treatment-effect oracle that knows d_a and f_a exactly but sees the same calibrated risk score as the learners; this isolates the cost of not knowing the action model, which is what the bandit layer can actually recover.

Table 1. Verification Ladder: Population-Level Response Parameters (Baf). Context Slopes Modulate The Log-Odds Of Interdiction And Drop-Off Through Indices Built From Real Covariates.

Action a	Description	opex c_a (USD)	interdiction d_a	drop-off f_a	slope β_d	slope β_f
a0	instant approval	0.05	0.001	0.005	0.00	0.50
a1	automated document + device check	0.60	0.42	0.030	0.90	0.90
a2	enhanced verification (liveness + doc)	2.20	0.72	0.095	0.70	1.10
a3	manual analyst review	11.00	0.90	0.210	0.45	1.30
a4	decline	0.00	0.999	0.999	0.00	0.00
post-shift a1	document check defeated (month 7)	0.60	0.22	0.030	0.90	0.90
post-shift a2	improved UX (month 7)	2.20	0.72	0.060	0.70	1.10

Exposure L(x) = 0.85 × 3 × requested credit limit; forgone value V(x) = max(25, 0.09 × limit × (0.6 + income)). ULB uses opex (0.01, 0.10, 0.30, 3.00, 0), interdiction (0.001, 0.35, 0.68, 0.90, 0.999), drop-off (0.002, 0.015, 0.070, 0.220, 0.999), L(x) = amount and V(x) = 0.015 × amount + 1.5.

4. Method: Carve

CARVE (Calibrated Adaptive Risk-aware VErification) instantiates (2) with estimated components and a specific discipline about uncertainty. Fig. 1 shows the architecture.

4.1. Interval-valued Calibrated Risk

A gradient-boosted decision tree is trained on historical applications to produce a score s(x). Scores are calibrated on a held-out temporal window with an inductive Venn–ABERS predictor (IVAP): for a test score s, the isotonic regression of the calibration set is refitted twice, once with the pair (s,0) appended and once with (s,1), yielding a probability pair whose midpoint is the point estimate and which is provably calibrated under exchangeability [23], [24]. We precompute the pair on a quantile grid of the calibration scores so that prediction costs one binary search.

IVAP intervals are extremely narrow: they reflect only the granularity of the isotonic map, and under heavy imbalance their mean width is below 5x10^-3 (Table 2). That is far too tight to be useful as a hedge, because the dominant uncertainty is the finite-sample uncertainty of the calibration map itself-our entire calibration month contains only 35 fraudulent applications. We therefore resample the calibration set B times, refit an IVAP on each resample, and report

$$p_hat = median_b\ mid_b(s),\quad p_lo = Q_\{delta/2\}(\ p0_b(s)\ ),\quad p_hi = Q_\{1-delta/2\}(\ p1_b(s)\ ). \tag{4}$$

The resulting band widens where the calibration data are thin, which is exactly where a bank is exposed. It is used only through its upper end, as described in Section IV-D.

4.2. Pooled Treatment-effect Heads

Fitting an independent regression of d_a on x for each action wastes the little interdiction data there is. We instead fit a single ridge model on the augmented feature map

$$phi(x, a) = [\ onehot(a)\ ;\ x\ ], \tag{5}$$

so that each rung of the ladder keeps its own intercept while the context slopes are shared. The implicit assumption is that the covariates which make an identity claim easy to check-valid mobile and landline numbers, an established banking relationship, a non-free e-mail domain, a plausible name/e-mail match, address history-act in the same direction on every rung, differing in magnitude. This is a bias–variance trade: pooling introduces a small bias when the true slopes differ across rungs, and removes a large variance. The ablation in Section VI-C shows that the trade is strongly favourable, pooling being the single most valuable component of the method. A second, identically structured head models the drop-off functions f_a; because drop-off is observed on the legitimate majority, this head is comparatively data-rich.

Both heads are maintained by recursive least squares with Sherman–Morrison inverse updates, so a round costs O(m^2) with m = |A| + d and no matrix factorisation. A running residual variance estimate s2 supplies the noise scale, and the epistemic standard error of the mean prediction is

$$sigma_a(x) = sqrt(s2 * phi(x,a)' A^{-1} phi(x,a)). \quad (6)$$

4.3. Block-conformal Confidence Radius

Optimism-based bandits add alpha*sigma_a(x) to the estimate, and alpha is a tuning parameter with a loose theoretical value. The natural conformal alternative—take the empirical quantile of normalised residuals |r - mu| / sigma—fails badly here, and understanding why is instructive. The target is Bernoulli, so |r - mu| is of order one whatever the sample size, whereas sigma shrinks like n^{-1/2}. The normalised score therefore diverges, the quantile grows without bound, and the inflated interdiction estimate saturates at one: the policy concludes that verification is free and verifies everybody. The residual quantile is measuring aleatoric noise, but the exploration bonus needs epistemic uncertainty.

We fix this by aggregating before calibrating. The learner keeps a rolling window of pairs (residual, standard error) per head. Periodically the window is partitioned at random into blocks of size m, and for each block the standardised statistic

$$u = | sum_i r_i | / sqrt(m * sum_i sigma_i^2) \quad (7)$$

is formed. Under correct scaling u behaves like the absolute value of a standard normal variate, so q = Q_{1-alpha}(u) is a data-driven width multiplier that concerns the mean rather than a single draw. The radius is q*sigma_a(x). This removes the exploration constant from the hyper-parameter list, adapts automatically as the heads improve, and, unlike the parametric radius, tightens when the linear model happens to fit well. The construction is a discrete-feedback analogue of scale-normalised conformal prediction, in which dividing a residual by a local dispersion estimate makes the nonconformity score approximately exchangeable across regimes [28]; our block statistic plays the same role, except that the quantity we must render comparable across rounds is the sampling noise of a mean rather than the conditional volatility of a return.

4.4. Asymmetric Uncertainty: Optimism for Exploration, Pessimism for Safety

Two kinds of uncertainty coexist. Uncertainty about d_a and f_a is reducible: acting resolves it, so the classical argument applies and the decision maker should be optimistic, choosing the action whose cost lower bound is smallest. Uncertainty about p is not reducible by acting on this applicant, and it is the uncertainty attached to the largest term in the cost; here optimism is simply gambling. CARVE therefore uses the two in opposite directions. It forms an optimistic view of the treatment effects,

$$d_a^+ = \mathrm{clip}(\hat{d}_a(x) + q_d \sigma_a^d(x)), \quad f_a^- = \mathrm{clip}(\hat{f}_a(x) - q_f \sigma_a^f(x)), \tag{8}$$

and a pessimistic view of the risk,

$$\tilde{p} = (1 - \rho)\,\hat{p} + \rho\, p_{hi}, \quad \rho \in [0,1], \tag{9}$$

and minimises the resulting index

$$J(a \mid x) = c_a + \tilde{p}\,(1 - d_a^+)\,L(x) + (1 - \tilde{p})\, f_a^-\, V(x) + \nu\, \mathbb{1}[a \in R]. \tag{10}$$

Setting $\rho = 0$ recovers a purely optimistic learner; $\rho = 1$ evaluates every application at the upper end of its risk band. Section VI-D shows that ρ traces a frontier along which fraud exposure falls by a fifth at essentially unchanged mean cost, which is the shape a risk manager wants: the parameter buys tail protection almost for free.

4.5. Prior-shift Correction

Fraud prevalence in the deployment stream drifts (in our data it doubles between the sixth and eighth month), which miscalibrates $\hat{p}$ even when the ranking is unaffected. CARVE maintains an exponentially weighted estimate π_t of the observed prevalence and rescales the odds of both $\hat{p}$ and p_{hi} by the ratio of current to training prevalence odds [29]. This is a one-line correction and removes 3.1% of total cost in the ablation.

4.6. Capacity Control by Dual Ascent

The review-rate constraint in (3) is enforced by a projected dual ascent step taken after every decision,

$$\nu \leftarrow \max(0,\ \nu + \eta\,(\,\mathbb{1}[a_t \in R] - \kappa\,)), \tag{11}$$

with ν entering the index in (10) as a shadow price on analyst time. The multiplicative cost-weight rule of [6] is the closest antecedent we know of-there the adapted quantity is the penalty for a missed attack, here it is the price of an analyst hour-and both rest on the observation that a static penalty either binds too hard or not at all once the environment moves. The multiplier is a deployment-time control and is reset to zero after warm start, since the historical logs were generated under a different capacity. Because the constraint is enforced in expectation rather than pathwise, tightness is achieved asymptotically; Section VII discusses the observed slack under very tight budgets.

4.7. Algorithm and Complexity

Algorithm 1 summarises the procedure. The offline stage fits the risk backbone and the calibrator and warm-starts the heads from historical logs; the online stage is a single pass. Per round the policy evaluates $|A| = 5$ quadratic forms and performs one rank-one update, giving $O(|A|\, m^2)$ time and $O(m^2)$ memory with $m = 33$ in our configuration. A stream of 3×10^4 applications is processed in under two seconds on one CPU core, so the decision layer adds no material latency to an onboarding flow.

Algorithm 1 CARVE

Input: risk backbone $s(\cdot)$; calibration window; historical logs; ladder parameters $\{c_a, d_a^0, f_a^0\}$; $\rho, \alpha, \kappa, \eta$.

Offline: fit bootstrap Venn–ABERS on the calibration window $\rightarrow (\hat{p}_{lo}, \hat{p}, \hat{p}_{hi})$; warm-start the pooled heads on the historical logs; reset $\nu \leftarrow 0$.

for t = 1, 2, … do

observe x_t; get (p_hat, p_hi) ← calibrator(s(x_t)); apply prior-shift correction with π_t.

p̃ ← (1 − ρ) p_hat + ρ p_hi.

for each a: σ_a ← head standard error; $d_a^{+} \leftarrow \mathrm{clip}(\hat{d}_a + q_d\, \sigma_a^{d})$; $f_a^{-} \leftarrow \mathrm{clip}(\hat{f}_a - q_f\, \sigma_a^{f})$.

a_t ← argmin_a $c_a + \tilde{p}(1 - d_a^{+})L(x_t) + (1 - \tilde{p})\, f_a^{-}\, V(x_t) + \nu \cdot 1[a \in R]$.

execute a_t; observe realised cost, interdiction / drop-off indicator and (eventually) y_t.

update the corresponding pooled head, its block-conformal multiplier, π_t and $\nu \leftarrow (\nu + \eta(1[a_t \in R] - \kappa))_{+}$.

end for

Table 2. Calibration On The Real Baf Deployment Window (Months 5–7, N = 6581, 1.23% Fraud). Lower Is Better Except Auc.

Calibrator	ECE ×10⁻³	MCE ×10⁻³	Brier ×10⁻³	NLL	ROC-AUC
Uncalibrated	5.87	39.25	11.464	0.0557	0.8405
Platt	4.20	23.59	11.396	0.0545	0.8405
Beta	4.29	22.09	11.342	0.0546	0.8405
Isotonic	3.93	9.24	11.540	0.0738	0.8140
Venn–ABERS	**3.51**	**8.80**	**11.729**	**0.0556**	**0.8380**

Risk band	bin agreement	mean width	width, top decile	ECE ×10⁻³
IVAP	0.92	0.0047	0.024	3.51
Bootstrap Venn–ABERS	**1.00**	**0.0162**	**0.112**	**4.51**

Bin agreement: fraction of score bins in which the predicted band overlaps a Wilson 95% interval on the observed fraud frequency (12 bins).

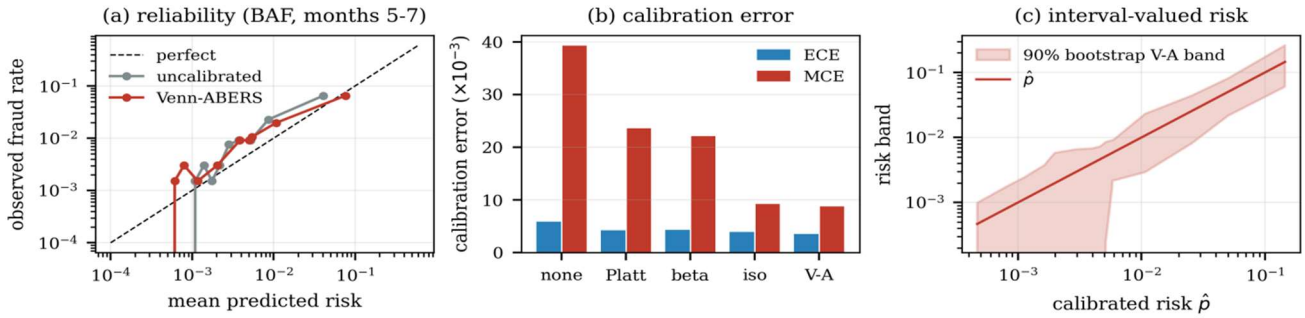


Fig. 2 Calibrated risk on the real BAF deployment window. (a) Reliability curves on log axes: the raw gradient-boosted score is over-confident in the upper deciles, which the Venn–ABERS map corrects. (b) Expected and maximum calibration error by calibrator. (c) The 90% bootstrap Venn–ABERS band widens by more than an order of magnitude in the high-risk tail, where the calibration set is thin; only its upper end enters the decision rule.

5. Experimental Setup

5.1. Datasets

BAF. The Bank Account Fraud suite [39] is a privacy-preserving generative resample of an anonymised real account-opening log from a large European consumer bank, released at NeurIPS 2022 with a datasheet. We use the public sample of the Base variant: 20,000 online applications, 31 predictive features (application, applicant, device and session attributes), a fraud label, and a month index spanning eight months. Fraud prevalence is 1.11% overall and

rises from 0.96% in month 6 to 1.93% in month 7, reproducing the temporal dynamics the suite was designed to expose. Months 0–3 (10,869 applications) train the risk backbone, month 4 (2,550) calibrates it and supplies the historical verification logs, and months 5–7 (6,581) form the deployment stream. All features and labels are real; only the response of alternative verification treatments is modelled (Section V-B).

ULB. The credit-card dataset released with [40] contains 284,807 real card-not-present transactions over two days with 492 frauds (0.172%), 28 PCA components and the true transaction amount. We use it as a second domain-step-up authentication rather than onboarding-because the amount field provides genuinely heterogeneous, heavy-tailed monetary stakes. Because the log spans only two days and prevalence is strongly non-stationary within it, a chronological split would confound the transfer test with prior shift; we therefore use a random 45/15/40 split and inject a controlled treatment-effect change instead.

5.2. Response Model and what is Real

We state plainly what is measured and what is modelled. All covariates, all fraud labels, all discrimination and calibration results (Section VI-A), and the exposures L and V-derived from the real proposed credit limit and, on ULB, the real transaction amount-come from the data. No public dataset contains counterfactual outcomes of alternative verification treatments, because no bank randomises its onboarding funnel at scale. The functions d_a and f_a are therefore specified by an explicit structural model, following the standard practice of converting supervised datasets into bandit problems [13], [16].

Population-level parameters are listed in Table 1 and are anchored on published identity-verification pricing and drop-off ranges: document verification and device intelligence cost cents to low single-digit dollars per check, biometric liveness slightly more, and an analyst review an order of magnitude more, while each added step raises abandonment from a fraction of a percent to tens of percent. Heterogeneity is a deterministic function of real covariates. Interdiction is modulated by a checkability index built from phone validity, existing card holdings, banking tenure, e-mail freeness, foreign-request and device-e-mail-multiplicity flags, name/e-mail similarity and address history; drop-off is modulated by a friction-sensitivity index built from age, session length, session persistence, income, requested limit, device family and channel. Exposure is $L(x) = LGD \times m \times limit$ with loss given default 0.85 and a bust-out multiplier $m = 3$, and forgone value is $V(x) = \max(25, 0.09 \times limit \times (0.6 + income))$. Deployment introduces a regime change at the start of month 7: automated document verification loses interdiction power (0.42 -> 0.22), reflecting synthetic-identity and injection attacks, while enhanced verification becomes less abrasive (drop-off 0.095 -> 0.060) after a user-experience release. On ULB the 3-D Secure challenge degrades from 0.68 to 0.38 in the final 40% of the stream, reflecting one-time-passcode phishing.

Because these constants are not identifiable from public data, every headline comparison is repeated under twelve systematic perturbations of them in Section VI-F. We regard the direction and robustness of the comparison, not the absolute dollar figures, as the result.

5.3. Protocol

The deployment stream is replayed with bootstrap resampling inside each month, preserving month order, to length $T = 30{,}000$; $B = 15$ independent streams are drawn and all policies see identical streams and identical outcome draws, so comparisons are paired. Historical logs consist of 40,000 decisions generated by the incumbent banded policy with a 10% uniform exploration budget-deliberately not a randomised log, because banks do not randomise. Every online learner is warm-started from exactly the same logs, so the comparison measures adaptation quality, not initialisation luck. Hyper-parameters were selected on three tuning streams disjoint from the fifteen evaluation streams: rho = 0.10 for CARVE, and, for the

baselines, the best of alpha in {0.1, 0.3, 1, 3} for LinUCB, v in {0.05, 0.2, 0.5} for Thompson sampling and epsilon in {0.01, 0.05, 0.15} for epsilon-greedy.

5.4. Baselines

- Fixed-action policies: ApproveAll, DocCheckAll, ManualAll.
- StaticThresh: the incumbent design. Three thresholds on the calibrated score, fitted offline by exhaustive grid search using population-average treatment parameters.
- PlugIn-Nominal: minimises (2) per applicant using the calibrated score and the nominal vendor-quoted treatment parameters, without learning. A strong and very cheap baseline.
- DM-Offline: direct-method cost regression per action, fitted on the historical logs and frozen; the unstructured counterpart of PlugIn-Nominal.
- OfflineCSC: the structured cost model of (2) with treatment-effect heads fitted on the historical logs and then frozen; no adaptation at deployment.
- Greedy, EpsGreedy, LinUCB [8], LinTS [11]: online learners that regress realised cost on context per action, i.e. without the decomposition (2).
- Oracle-Theta: knows d_a and f_a exactly, sees the same calibrated score as the learners. Regret is measured against it.

5.5. Metrics

We report expected cost per application decomposed into uncaught fraud loss, verification opex and friction/lost value; fraud interdiction rate; abandonment rate among legitimate applicants; verification, review and decline rates; per-round regret against Oracle-Theta; cost during the drifted month; and a differential-friction audit across the BAF protected attribute (age >= 50). Calibration is assessed by equal-mass expected and maximum calibration error, Brier score and negative log-likelihood. Significance is assessed with the paired Wilcoxon signed-rank test over the fifteen streams.

Table 3. Online Adaptive Verification On The Baf Deployment Stream. Mean Over 15 Paired Streams Of T = 30 000 Applications; ± Is The Across-Stream Standard Deviation. All Differences From Carve Are Significant (Paired Wilcoxon, P < 0.01).

Policy	cost (USD/app)	fraud loss	opex	friction	interdict.	verify %	review %	regret/rd	drift-month cost
ApproveAll	24.340 ± 1.73	24.032	0.050	0.258	0.1%	0.0	0.0	15.289	38.639
DocCheckAll	21.530 ± 1.74	19.176	0.600	1.753	21.8%	100.0	0.0	12.448	35.910
ManualAll	26.083 ± 0.89	3.841	11.000	11.242	85.1%	100.0	100.0	16.831	27.603
DM-Offline (unstructured)	27.411 ± 1.79	21.523	2.319	3.570	25.7%	70.3	12.2	18.315	39.203
Greedy	19.197 ± 0.82	14.426	1.514	3.258	46.2%	52.5	6.5	10.164	24.708
EpsGreedy (ε=0.01)	19.245 ± 0.86	14.329	1.539	3.377	46.4%	53.1	6.8	10.248	24.722
LinUCB (α=0.1)	19.156 ± 0.70	14.312	1.576	3.268	46.6%	54.3	7.0	10.182	24.663
LinTS (v=0.05)	19.062 ± 0.86	14.273	1.520	3.269	46.7%	52.5	6.6	10.150	24.301
StaticThresh (incumbent)	11.922 ± 1.29	9.923	0.790	1.208	46.3%	15.1	4.7	2.741	16.773
OfflineCSC (frozen)	10.620 ± 0.96	8.329	0.509	1.782	48.2%	25.3	1.8	1.535	14.676
PlugIn-Nominal	9.908 ± 0.96	7.444	0.654	1.810	52.3%	34.4	2.7	0.812	14.542
CARVE (ours)	**9.588 ± 0.92**	**6.485**	**0.783**	**2.320**	**57.3%**	**46.0**	**3.3**	**0.673**	**13.301**
Oracle-Θ (benchmark)	*9.017 ± 0.92*	*6.296*	*0.836*	*1.885*	*57.2%*	*28.5*	*4.5*	*0.000*	*11.285*

Cost is decomposed into uncaught fraud loss, verification opex and friction (lost value from abandonment and declines); components sum to the total. Regret is per round against Oracle-Θ, which knows the true interdiction and drop-off functions but sees the same calibrated risk score as the learners.

6. Results

6.1. Risk Backbone and Calibrated Uncertainty

On the real BAF deployment window the tuned gradient-boosted model [42] reaches ROC-AUC 0.841, PR-AUC 0.145, recall 0.469 at 5% false-positive rate and 0.235 at 1%; logistic regression and random forests are within one point of AUC (0.841 and 0.818) but weaker in the high-specificity region that matters operationally. These numbers are consistent with published results on the full one-million-row Base variant and confirm that the sample retains the difficulty of the task.

Table 2 compares calibrators. Venn–ABERS attains the lowest expected calibration error (3.5x10^-3, a 40% reduction over the raw score) and the lowest maximum calibration error (8.8x10^-3, a 4.5-fold reduction), while retaining almost all discrimination; isotonic regression matches it on MCE but loses 2.7 AUC points and inflates the negative log-likelihood by a third, and the parametric calibrators sit in between. Fig. 2(a) shows the reliability curves. The interval comparison in Table 2 confirms the argument of Section IV-A: the raw IVAP band has mean width 4.7x10^-3 and 2.4x10^-2 in the top risk decile, whereas the bootstrap-aggregated band widens to 1.6x10^-2 and 1.1x10^-1 respectively and agrees with a Wilson interval on the observed frequency in every score bin. Fig. 2(c) plots the band. Monthly diagnostics show AUC between 0.823 and 0.896 with calibration error rising from 4.0x10^-3 to 1.0x10^-2 in the drifted month, which is what the prior-shift correction is for.

6.2. Online Adaptive Verification

Table 3 is the main result. CARVE attains 9.588 +/- 0.920 USD per application against 11.922 for the incumbent banded policy (-19.6%), 10.620 for the frozen offline cost-sensitive policy (-9.7%), 9.908 for the nominal plug-in rule (-3.2%) and 24.340 for instant approval (-60.6%); all differences are significant at $p < 0.01$ under the paired Wilcoxon test ($p = 6.1\text{x}10^{-5}$ except against PlugIn-Nominal, $p = 3.4\text{x}10^{-3}$). Per-round regret against the treatment-effect oracle is 0.673 USD, i.e. CARVE recovers 75.5% of the gap that the incumbent policy leaves open and lands 6.3% above the oracle in absolute cost. Fig. 3(a) shows cumulative regret; CARVE's curve is the flattest and, unlike the frozen policies, does not kink upward at the regime change.

The decomposition in Fig. 3(b) explains where the saving comes from. Relative to StaticThresh, CARVE raises the verification rate from 15.1% to 46.0% while cutting the manual-review rate from 4.7% to 3.3%: it substitutes cheap, well-targeted automated checks for expensive analyst time. Fraud interdiction rises from 46.3% to 57.3% and uncaught fraud loss falls from 9.923 to 6.485 USD, at the price of friction cost rising from 1.208 to 2.320 USD-a trade the cost model endorses because L exceeds V by roughly an order of magnitude. Under the drifted month the ordering widens: 13.301 USD for CARVE against 16.773 for StaticThresh and 14.676 for OfflineCSC.

The unstructured learners are the informative negative result. Greedy, epsilon-greedy, LinUCB and linear Thompson sampling all land between 19.06 and 19.25 USD-twice CARVE's cost-despite identical warm-start data, tuned exploration constants and identical streams. Their per-observation cost signal has a standard deviation near 140 USD (fraud prevalence 1.2% times exposure of order 10^3), while the differences they must resolve between adjacent rungs are of order 1 USD; with the few thousand observations per action available they cannot separate the arms and settle on over-verification with a 53% verification rate and only 46.7% interdiction. DM-Offline, the frozen unstructured counterpart, is worse than doing nothing (27.4 USD). Exploiting the accounting structure of the loss is not an optimisation detail here; it is what makes the problem learnable.

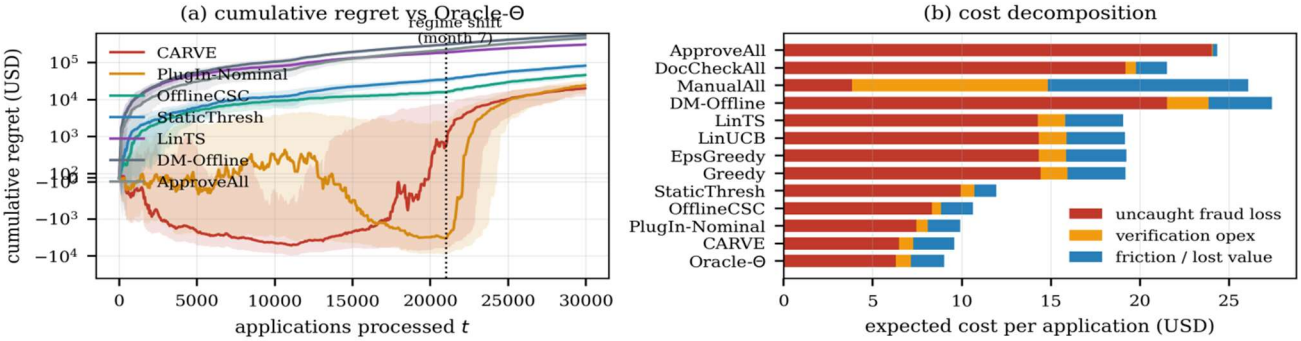


Fig. 3 (a) Cumulative regret against Oracle-Θ (symmetric-log scale, mean ± 1 s.d. over 15 streams). The dotted line marks the treatment-effect regime change at the start of month 7, after which the frozen policies steepen and CARVE does not. (b) Cost decomposition per application: CARVE substitutes cheap targeted automated checks for analyst time and accepts more friction in exchange for a much smaller uncaught-fraud term.

6.3. Component Ablation

Table 4 removes one component at a time on six streams. Every component contributes and the ordering is interpretable. Removing slope pooling is the most damaging (+8.1% cost): with independent per-action heads the interdiction estimates are so noisy that the optimism term drives the verification rate from 44.8% to 54.3% and opex from 0.753 to 1.017 USD. Removing exploration entirely (greedy on the structured model) costs +4.9% and inflates regret by 64%, confirming that the residual value of information is real even in this data-poor regime. Removing context from the treatment-effect heads costs +3.2% and removing the prior-shift correction +3.1%, the latter concentrated in the drifted month (12.836 -> 13.573 USD). Replacing the block-conformal radius with a tuned parametric radius costs +1.1%, and removing risk aversion +0.8% in mean cost but +3.6% in uncaught fraud loss.

6.4. The Risk-aversion Frontier

Fig. 4(a,b) sweeps rho. Uncaught fraud loss falls monotonically from 6.436 to 5.173 USD (-19.6%) as rho goes from 0 to 1, while total cost varies only between 9.079 and 9.431 USD, a band of 3.9%. Interdiction rises from 56.2% to 63.1% and the verification rate from 42.6% to 58.1%. Calibrated pessimism is thus close to a free hedge over a wide range: an institution can buy a fifth of its fraud exposure for a low single-digit percentage of total cost, and the exchange rate is set by an interpretable dial rather than by re-tuning thresholds. We stress that rho = 0.10 was fixed on tuning streams; the frontier is reported for interpretation, not selection.

Table 4. Component Ablation (6 Paired Streams, T = 30 000).

Configuration	cost	Δ cost	fraud loss	verify %	regret/rd	drift cost
CARVE (full)	**9.199**	**—**	**6.213**	**44.8**	**0.664**	**12.836**
– risk aversion (ρ = 0)	9.269	+0.8%	6.436	42.6	0.679	12.933
– block-conformal radius	9.299	+1.1%	6.730	33.6	0.754	12.927
– prior-shift correction	9.484	+3.1%	6.667	45.0	0.693	13.573
– context in heads	9.491	+3.2%	6.658	39.4	0.776	12.646
– exploration (greedy)	9.646	+4.9%	6.838	24.8	1.091	12.649
– slope pooling	9.942	+8.1%	6.678	54.3	0.907	14.023
Oracle-Θ	*8.680*	*–5.6%*	*6.004*	*28.4*	*0.000*	*11.037*

The parametric-radius variant uses a tuned LinUCB confidence width in place of the block-conformal one; the pooling ablation fits one independent ridge head per action.

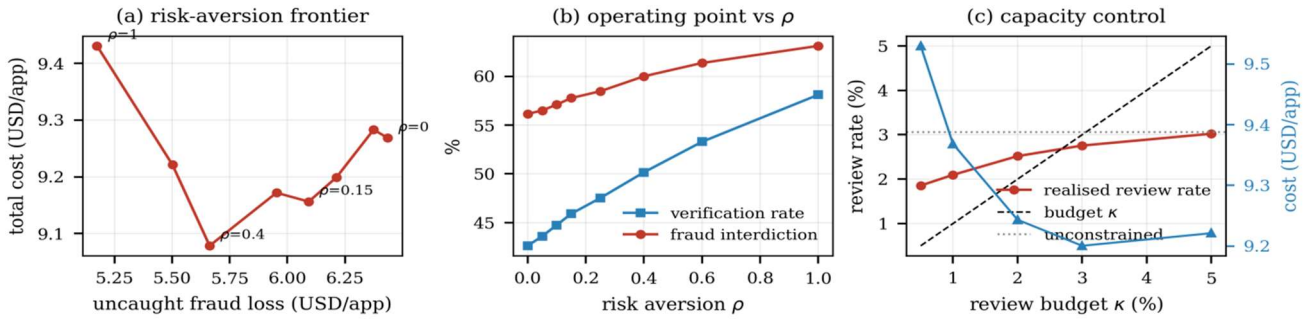


Fig. 4 (a) Risk-aversion frontier: uncaught fraud loss falls 19.6% as ρ goes from 0 to 1 while total cost varies within 3.9%. (b) The operating point moves monotonically with ρ. (c) The Lagrangian dual tracks the analyst-capacity budget κ; the constraint is met for κ ≥ 2% and slack remains for tighter budgets (Section VII).

6.5. Capacity Control

Fig. 4(c) and the accompanying sweep show the dual variable tracking the review budget: as kappa tightens from 5% to 0.5% the realised review rate falls monotonically from 3.02% to 1.86% and cost rises from 9.221 to 9.530 USD, giving a directly usable capacity–cost curve. The constraint is satisfied for kappa >= 2%; for kappa <= 1% the realised rate exceeds the target, an honest limitation discussed in Section VII.

6.6. Sensitivity to the Response Model

Fig. 5(a) repeats the head-to-head comparison under twelve perturbations: loss multiplier 1, 2, 3, 5; operating costs halved and doubled; weak and strong interdiction ladders; high and low friction; doubled customer value; and no regime change. CARVE beats the incumbent banded policy in every regime, by 8.1% (doubled customer value, where verification is least attractive) to 30.1% (loss multiplier 5, where it is most attractive), with a mean saving of 17.9%; it beats the frozen offline policy in every regime by 1.0% to 7.2%. The gap to the oracle ranges from 5.3% to 24.1%. The conclusion is insensitive to the constants we could not identify from data.

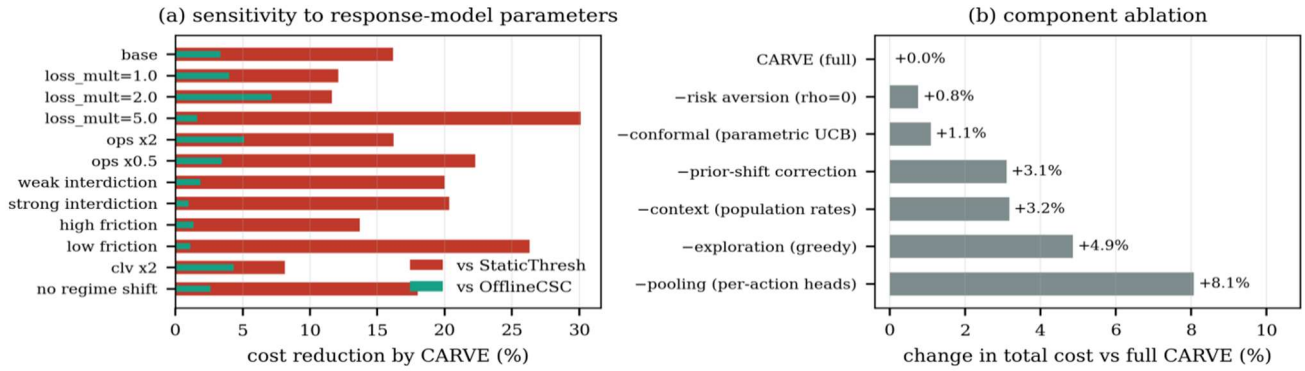


Fig. 5 (a) Sensitivity: CARVE beats the incumbent banded policy in all twelve perturbations of the response model, by 8.1% to 30.1%. (b) Component ablation; every component contributes, and slope pooling across the verification ladder is the most valuable one.

6.7. Differential Friction Across a Protected Attribute

BAF ships age band as a protected attribute, and in our deployment window applicants aged 50 and over carry 2.29% fraud prevalence against 0.84% for younger applicants. Any prevalence-following policy therefore imposes more friction on the older group, and Fig. 6(a) quantifies it: CARVE verifies 68.9% of protected-group applicants against 41.3% of the rest, a 27.6-point gap, compared with 9.0 points for StaticThresh (22.6% vs 13.6%) and 32.0 points for the oracle. Realised abandonment among legitimate applicants is much closer (3.18% vs 3.65%), because

the older group is also less friction-sensitive in our response model, but the burden of proof is unequally distributed. We report this as a finding, not a solution: minimising expected cost does not control disparity, the disparity is inherited from the label distribution rather than created by the learner, and a deployment would need an explicit parity constraint of the kind studied in [43], [44] on top of the cost objective.

6.8. Can the Policy Be Screened before Deployment?

A bank cannot cheaply A/B-test an onboarding funnel, so a candidate policy must be screened from logged data. We log the deployment stream under an epsilon-greedy incumbent (epsilon = 0.20) and estimate the value of four target policies with IPS, self-normalised IPS, the direct method and doubly robust estimation, comparing against the on-policy truth obtained by actually running each target (Fig. 6(b), Table 5). Doubly robust estimation is the most accurate on the policy that matters, estimating CARVE at 8.451 against a true 9.241 USD (8.5% optimistic bias), and gives the best rank correlation with the truth (Spearman 0.90 versus 0.80 for the others). IPS and SNIPS are accurate on the policy closest to the logging distribution (StaticThresh, effective sample size 25,205 of 30,000) and degrade to 12% error where the overlap is poor (ApproveAll, ESS 6,519). All four estimators rank CARVE first. Pre-deployment screening is therefore feasible but should be read as a ranking device with an optimistic bias, not as a cost forecast.

Table 5. Off-Policy Evaluation Of Four Candidate Policies From Logs Of An E-Greedy Incumbent (E = 0.20, 6 Streams). Estimated Cost Per Application Against On-Policy Ground Truth.

Target policy	truth	IPS	SNIPS	DM	DR	ESS
CARVE (frozen)	**9.241**	**8.216**	**8.187**	**10.082**	**8.451**	**2341**
StaticThresh	11.558	11.240	11.239	9.153	11.240	25205
DocCheckAll	20.909	21.894	21.833	19.091	20.175	1209
ApproveAll	23.602	20.685	20.797	21.748	21.346	6519
mean |error|	—	*1.311*	*1.276*	*1.730*	*1.025*	—
Spearman ρ with truth	—	*0.80*	*0.80*	*0.80*	*0.90*	—

ESS is the effective sample size of the importance weights out of 30 000 rounds. DR is the most accurate on the policy of interest and gives the best ranking; IPS and SNIPS degrade where the overlap with the logging policy is poor.

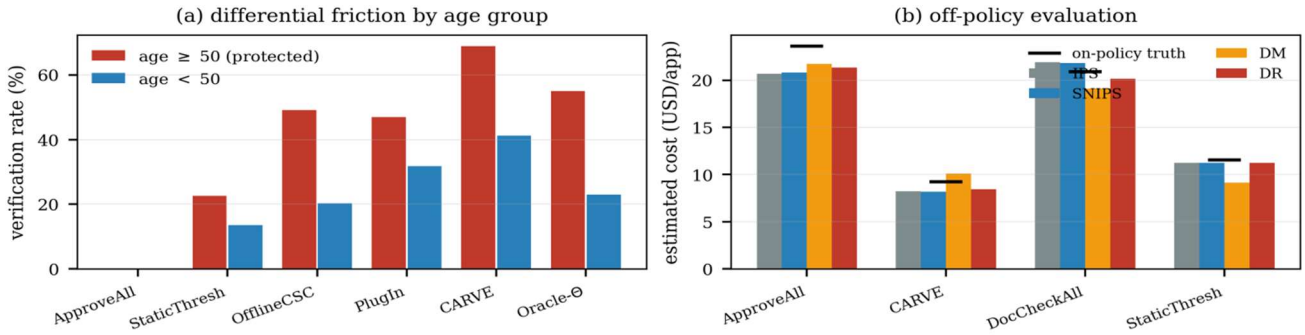


Fig. 6 (a) Differential friction: because fraud prevalence in the protected age band is 2.7× higher, every prevalence-following policy verifies that group more often; CARVE's gap (27.8 points) is wider than the incumbent's (9.0) and narrower than the oracle's (32.0). (b) Off-policy value estimates against on-policy truth (black dashes).

6.9. Transfer to Card-Not-Present Step-Up Authentication

Table 6 repeats the study on 60,000 real transactions with the amount as exposure. The backbone reaches ROC-AUC 0.977 and PR-AUC 0.845 at 0.16% prevalence. CARVE attains 0.1097 USD per transaction against 0.1342 for the banded policy (-18.3%), 0.1211 for the frozen offline policy (-9.4%), 0.1197 for the nominal plug-in (-8.4%) and 0.2057 for frictionless approval (-46.7%), landing 2.1% above the oracle (0.1074). It challenges only 3.8% of transactions and refers 0.01% to analysts. The unstructured learners again fail in the same way, challenging 39–43% of traffic for 2.2–2.4x the cost. The pattern established on onboarding data reproduces on a different action ladder, a different exposure distribution and a 7-fold lower prevalence.

Table 6. Transfer To Card-Not-Present Step-Up Authentication (Ulb, 60 000 Real Transactions, 0.162% Fraud, Exposure = True Amount). Mean Over 4 Paired Runs.

Policy	cost (USD/txn)	fraud loss	opex	friction	interdict.	challenge %	regret/rd
ApproveAll	0.2057	0.1889	0.0100	0.0067	0.0%	0.0	0.0961
Challenge-All (3-DS)	0.7132	0.1501	0.3000	0.2630	17.5%	100.0	0.5922
LinUCB (α=0.1)	0.2595	0.0885	0.1190	0.0519	83.5%	42.8	0.1471
LinTS (v=0.05)	0.2457	0.0885	0.1090	0.0481	83.5%	39.4	0.1333
StaticThresh	0.1342	0.0906	0.0282	0.0155	50.3%	0.6	0.0242
OfflineCSC (frozen)	0.1211	0.0362	0.0109	0.0740	83.8%	2.2	0.0089
PlugIn-Nominal	0.1197	0.0256	0.0115	0.0827	79.9%	2.5	0.0089
CARVE (ours)	**0.1097**	**0.0486**	**0.0131**	**0.0481**	**76.0%**	**3.8**	**0.0063**
Oracle-Θ	*0.1074*	*0.0203*	*0.0129*	*0.0742*	*80.4%*	*2.4*	*0.0000*

Risk backbone: ROC-AUC 0.977, PR-AUC 0.845. The 3-D Secure challenge degrades from 0.68 to 0.38 interdiction in the final 40% of the stream.

7. Discussion and Limitations

Three findings generalise beyond the specific application. First, when the loss has a known accounting structure, learning its components separately is not a refinement but a precondition: the variance of the aggregate loss can exceed the signal by two orders of magnitude, and every unstructured contextual bandit we tried remained stuck in exploration. Second, conformal calibration of a confidence radius must be applied to a statistic of the mean; the naive normalised-residual quantile measures aleatoric noise and, with Bernoulli feedback, produces a radius of order one that saturates the estimate. Third, epistemic and irreducible uncertainty should enter a decision rule with opposite signs, and the pessimism half of that rule is close to free in mean cost while buying a substantial reduction in tail exposure.

The limitations are real. (i) The treatment-response functions are modelled, not measured; the sensitivity study bounds how much this matters for the comparison but cannot validate the absolute figures, and the only genuine remedy is a randomised exploration slice in a production funnel. (ii) We use the public 20,000-row sample of BAF rather than the full million rows, because the full suite is not reachable from our compute environment; the sample reproduces the published prevalence and drift pattern, but the risk backbone is trained on roughly 100 positive examples and a larger sample would sharpen every number reported here. (iii) Labels are assumed to be observed without delay. In production, first-party fraud surfaces months later, so the treatment-effect heads would update on a lag; a delayed-feedback bandit formulation is the natural extension. (iv) The capacity constraint is enforced in expectation, and for budgets below 1% the dual ascent is too slow relative to the horizon; a self-tuning step size or a hard per-window quota would fix it. (v) The treatment-effect heads are linear in a fixed feature map, and pooling assumes that context effects have a common direction across rungs;

a hierarchical model with shrunken per-rung deviations would relax this at quadratic cost in the number of features. (vi) Adversarial adaptation is modelled as an exogenous change point rather than as a response to the policy itself; endogenising the adversary inside the training loop, as in the Stackelberg formulation of [38], is the natural next step and is out of scope here. (vii) The differential-friction result shows that cost minimisation alone does not control disparity, and any deployment would require an explicit fairness constraint and human oversight.

8. Conclusion

We recast the routing of digital-onboarding applications through a verification ladder as a budget-constrained cost-sensitive contextual bandit, and showed that the combination of a structured cost decomposition, treatment-effect heads pooled across the ladder, a block-aggregated conformal exploration radius and an asymmetric use of calibrated uncertainty closes most of the gap to a treatment-effect oracle in a regime where standard contextual bandits fail outright. On real account-opening data the method reduces expected cost per application by 19.6% relative to a tuned banded policy while cutting analyst load by 30%, the result holds across twelve perturbations of the response model, and it reproduces on card-not-present step-up authentication. The components are simple, run in constant time per decision, and expose two interpretable operating dials-a risk-aversion parameter and a review budget-which is what makes the approach deployable. Code and experimental artefacts accompany the paper.

References

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

[2] Sahin, Y., Bulkan, S., & Duman, E. (2013). A cost-sensitive decision tree approach for fraud detection. Expert Systems with Applications, 40(15), 5916–5923.

[3] Correa Bahnsen, A., Aouada, D., & Ottersten, B. (2015). Example-dependent cost-sensitive decision trees. Expert Systems with Applications, 42(19), 6609–6619.

[4] Correa Bahnsen, A., Stojanovic, A., Aouada, D., & Ottersten, B. (2014). Improving credit card fraud detection with calibrated probabilities. In Proceedings of the SIAM International Conference on Data Mining (pp. 677–685).

[5] 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.

[6] Ding, J., Shen, Z., & Liu, W. (2026). Game-theoretic cost-sensitive adversarial training for robust cloud intrusion detection against GAN-based evasion attacks. Applied Sciences, 16(8), Article 3944.

[7] Ping, W., Jiao, Y., Fan, H., & Zhang, X. (2026). Multimodal fraud detection in financial statements: A trimodal attention network with contrastive evidence chain construction. IEEE Access, 14, 80456–80468.

[8] Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In Proceedings of the 19th International Conference on World Wide Web (pp. 661–670).

[9] Agrawal, S., & Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In Proceedings of the 30th International Conference on Machine Learning (pp. 127–135).

[10] Chu, W., Li, L., Reyzin, L., & Schapire, R. E. (2011). Contextual bandits with linear payoff functions. In Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (pp. 208–214).

[11] Lattimore, T., & Szepesvári, C. (2020). Bandit algorithms. Cambridge University Press.

[12] Foster, D. J., & Rakhlin, A. (2020). Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In Proceedings of the 37th International Conference on Machine Learning (pp. 3199–3210).

[13] Bietti, A., Agarwal, A., & Langford, J. (2021). A contextual bandit bake-off. Journal of Machine Learning Research, 22(133), 1–49.

[14] Bastani, H., Bayati, M., & Khosravi, K. (2021). Mostly exploration-free algorithms for contextual bandits. Management Science, 67(3), 1329–1349.

[15] Garivier, A., & Moulines, E. (2011). On upper-confidence bound policies for switching bandit problems. In Proceedings of the 22nd International Conference on Algorithmic Learning Theory (pp. 174–188).

[16] Dudík, M., Langford, J., & Li, L. (2011). Doubly robust policy evaluation and learning. In Proceedings of the 28th International Conference on Machine Learning (pp. 1097–1104).

[17] Swaminathan, A., & Joachims, T. (2015). Counterfactual risk minimization: Learning from logged bandit feedback. In Proceedings of the 32nd International Conference on Machine Learning (pp. 814–823).

[18] Agrawal, S., & Devanur, N. R. (2016). Linear contextual bandits with knapsacks. In Advances in Neural Information Processing Systems (Vol. 29, pp. 3450–3458).

[19] Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. J. Smola, P. Bartlett, B. Schölkopf, & D. Schuurmans (Eds.), Advances in large margin classifiers (pp. 61–74). MIT Press.

[20] Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 694–699).

[21] Kull, M., Silva Filho, T. M., & Flach, P. (2017). Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (pp. 623–631).

[22] Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In Proceedings of the 22nd International Conference on Machine Learning (pp. 625–632).

[23] Vovk, V., & Petej, I. (2014). Venn–Abers predictors. In Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (pp. 829–838).

[24] Vovk, V., Petej, I., & Fedorova, V. (2015). Large-scale probabilistic predictors with and without guarantees of validity. In Advances in Neural Information Processing Systems (Vol. 28, pp. 892–900).

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

[26] Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning, 16(4), 494–591.

[27] Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., & Schuster, T. (2024). Conformal risk control. In Proceedings of the 12th International Conference on Learning Representations.

[28] 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.

[29] Saerens, M., Latinne, P., & Decaestecker, C. (2002). Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. Neural Computation, 14(1), 21–41.

[30] Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Machine Learning, 110(3), 457–506.

[31] Madras, D., Pitassi, T., & Zemel, R. (2018). Predict responsibly: Improving fairness and accuracy by learning to defer. In Advances in Neural Information Processing Systems (Vol. 31, pp. 6147–6157).

[32] Mozannar, H., & Sontag, D. (2020). Consistent estimators for learning to defer to an expert. In Proceedings of the 37th International Conference on Machine Learning (pp. 7076–7087).

[33] Alves, J., Jesus, S., Sousa, L., Saleiro, P., & Bizarro, P. (2023). FiFAR: A fraud detection dataset for learning to defer. arXiv. https://arxiv.org/abs/2312.13218

[34] Sani, A., Lazaric, A., & Munos, R. (2012). Risk-aversion in multi-armed bandits. In Advances in Neural Information Processing Systems (Vol. 25, pp. 3275–3283).

[35] Rockafellar, R. T., & Uryasev, S. (2000). Optimization of conditional value-at-risk. Journal of Risk, 2(3), 21–41.

[36] Jin, Y., Yang, Z., & Wang, Z. (2021). Is pessimism provably efficient for offline RL? In Proceedings of the 38th International Conference on Machine Learning (pp. 5084–5096).

[37] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), Article 44.

[38] 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

[39] Jesus, S., Pombal, J., Alves, D., Cruz, A. F., 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), Datasets and Benchmarks Track.

[40] 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).

[41] 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.

[42] 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).

[43] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems (Vol. 29, pp. 3315–3323).

[44] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In Proceedings of the 8th Innovations in Theoretical Computer Science Conference (pp. 43:1–43:23).