

The Fragility of Value in Customer Churn: Predict-then-Optimize vs. Predict-and-Optimize Under Dataset Shift

Shimanto Rahman^{1,2}, Bram Janssens¹, and Matthias Bogaert^{1,2}

¹ Ghent University, Ghent, Belgium

{shimanto.rahman, bram.janssens, matthias.bogaert}@UGent.be

² FlandersMake@UGent-corelab CVAMO, Ghent, Belgium

Abstract. Customer churn prediction (CCP) models are essential tools for identifying consumers at risk of attrition. However, traditional CCP approaches prioritize statistical accuracy over financial return, often ignoring the inherent value disparity among individual customers. This limitation has driven the adoption of cost-sensitive learning, broadly categorized into two approaches: *Predict-then-Optimize* (PtO), which trains standard probabilistic models and applies economic decision rules post-hoc, and *Predict-and-Optimize* (PaO), which directly integrates misclassification costs into the training objective. While recent literature has evaluated these approaches on static datasets, models in real-world business applications are typically deployed into non-stationary environments operating under static, unyielding decision policies. This research investigates the economic fragility of cost-sensitive CCP approaches under dataset shift. Utilizing a Gaussian copula architecture on an empirical telecommunications dataset, we decouple marginal behavior from feature covariance to systematically induce covariate and label drift. These statistical perturbations mirror real-world phenomena such as economic recessions and shifts in consumer pricing models. Our findings reveal a fundamental operational trade-off: while PtO maximizes in-distribution profitability, its reliance on probabilistic calibration makes it vulnerable to scale-based shifts. Conversely, PaO approaches establish a defensive moat, utilizing stable, uncalibrated score densities to maintain profitability during structural covariance drift. This study provides a comprehensive taxonomy of economic model decay, offering vital guidance for deploying cost-sensitive classifiers in volatile business applications.

Keywords: Predict-then-Optimize · Predict-and-Optimize · Cost-Sensitive Learning · Customer Churn Prediction · Dataset Shift.

1 Introduction

Customer Relationship Management (CRM) heavily relies on retaining existing clientele, as acquiring new customers incurs substantially higher costs than preserving current ones [10]. To facilitate this, organizations deploy Customer

Churn Prediction (CCP) models, which analyze historical behavior to identify individuals with a high probability of terminating their relationship with the company [17]. While traditional CCP approaches treat all customers equally by optimizing purely for statistical accuracy (e.g., maximizing the Area Under the ROC Curve), this assumption is fundamentally misaligned with the financial realities of business applications [18]. Because customers exhibit heterogeneity in their Customer Lifetime Value (CLV), ranking interventions solely based on attrition probability frequently results in suboptimal and unprofitable retention campaigns [11].

To resolve this misalignment, the literature has established two distinct cost-sensitive learning approaches [16]. The first is the **Predict-then-Optimize (PtO)** approach. In this framework, standard classification algorithms are trained using proper scoring rules (e.g., cross-entropy) to maximize predictive accuracy [6], and economic optimization is deferred to the decision-making stage. Theoretical decision rules, such as Bayes Minimum Risk thresholds, are then applied directly to the model’s calibrated probabilities [4].

The second approach is the **Predict-and-Optimize (PaO)** paradigm. This stream argues that training a model independently of its operational context is economically suboptimal [16]. Instead, PaO integrates asymmetric cost matrices directly into the model training phase to explicitly minimize the Average Expected Cost [3] or maximize profit metrics like the Expected Maximum Profit [18]. Because bespoke cost-sensitive loss functions are improper scoring rules, PaO models output uncalibrated economic scores rather than true conditional probabilities. This requires practitioners to abandon theoretical Bayesian bounds and instead deploy empirical decision thresholds tuned on historical validation data.

While the literature rigorously debates the in-distribution superiority of these two approaches, a critical operational vulnerability remains unaddressed: their robustness under dataset shift. In real-world business applications, retention models are deployed with static, fixed threshold policies. When macroeconomic shocks induce distribution drift, the financial viability of these immutable policies decays. To address this gap, we utilize a continuous Gaussian copula framework to rigorously decouple marginal behavioral drift from underlying covariance structures, allowing us to investigate the following research questions:

1. Do the cost-aligned representations learned by PaO models exhibit greater robustness to shifting feature dependencies than the purely statistical representations of PtO models?
2. Under a systematic decline in customer usage, which failure mechanism proves more detrimental: the degradation of probabilistic calibration in PtO models, or the misalignment of empirical score densities in PaO models?
3. Can either approach maintain profitability during a sudden macroeconomic shift in the prior probability of churn while operating under fixed deployment policies?

2 Related Work

Customer attrition is a critical concern in business analytics, prompting a shift from accuracy-driven metrics to cost-sensitive strategies [18]. Traditional CCP models implicitly assume symmetric misclassification costs, fundamentally ignoring the economic reality that the financial upside of successfully retaining a true churner vastly outweighs the cost of erroneously incentivizing a loyal customer [17]. Early efforts to resolve this imbalance introduced class-dependent costs, which aggregate these tradeoffs by estimating the average benefit of retention against the global average cost of a campaign intervention. While recognizing this baseline asymmetry provides significant structural advantages, treating all targeted customers identically remains detrimental in industries characterized by high customer value heterogeneity, such as Business-to-Business sectors [11]. In these volatile environments, it is highly beneficial to further individualize the economic penalties utilizing instance-dependent costs. To conceptualize the evolution and implementation of these cost-sensitive methodologies, Vanderschueren et al. [16] proposed a unifying two-stage framework that categorizes the literature into the Predict-then-Optimize and Predict-and-Optimize approaches.

In the Predict-then-Optimize (PtO) approach, the training phase relies on proper scoring rules to estimate true conditional probabilities, deferring economic trade-offs to the downstream deployment stage [16]. Because the outputs of PtO models are assumed to be well-calibrated probabilities, the optimal business policy can be determined analytically post-hoc. For instance, applying theoretical Bayes Minimum Risk thresholds to standard classifier probabilities drastically improves financial savings compared to arbitrary, accuracy-based cut-offs [4]. However, the operational success of the PtO approach is inextricably linked to the model’s probabilistic calibration remaining perfectly intact upon deployment.

Conversely, the Predict-and-Optimize (PaO) paradigm asserts that training a model independently of its ultimate operational context yields suboptimal economic boundaries [16]. This approach integrates misclassification costs directly into the machine learning algorithm. Because true positive benefits and false positive costs in customer churn are inherently instance-dependent—varying drastically based on a customer’s individual Customer Lifetime Value [11]—algorithms such as Cost-Sensitive Logistic Regression [8] and Cost-Sensitive Decision Trees [2] alter their internal splitting criteria or gradients to explicitly minimize expected financial loss. Similarly, other methodologies directly integrate the Expected Maximum Profit Measure for Customer Churn (EMPC) into the training process utilizing evolutionary learning [15,9]. A profound theoretical consequence of direct PaO optimization is the loss of probability calibration. Even when practitioners attempt to make standard proper scoring rules (such as cross-entropy) cost-sensitive via asymmetric weighting, the optimization landscape fundamentally shifts to minimize financial risk rather than to elicit the true conditional probability. Because these bespoke economic loss functions deviate from standard maximum likelihood estimation and thus become mathematically improper, PaO models inherently output uncalibrated risk scores, forcing practitioners to rely on empirical decision policies tuned on historical validation data.

While both approaches aggressively optimize for economic value, existing literature overwhelmingly evaluates them under the assumption that the deployment data is drawn from the identical distribution as the training data. However, empirical studies in general churn management demonstrate that this static assumption is operationally flawed. For instance, Risselada et al. [14] established that the “staying power” of churn prediction models is notoriously low; predictive performance deteriorates rapidly post-estimation due to temporal shifts in customer behavior and changing parameter significance, fundamentally necessitating frequent model retraining. Despite this recognized reality, research addressing structural covariate drift in the feature space remains remarkably scarce within the cost-sensitive literature. To rigorously benchmark policy decay without destroying the complex dependencies of empirical churn data, modern methodologies employ Gaussian copulas [12,13]. This architecture allows for the precise, isolated induction of distinct drift typologies reflecting real-world business disruptions. By isolating marginal drift (representing macroscopic usage decline of customers, such as an economic recession) from covariance drift (representing the decoupling of customer behaviors, such as shifts to flat-rate pricing models), this framework provides a controlled environment to answer a critical operational question: under dataset shift, what fails first—the theoretical probabilistic calibration of a PtO model, or the empirical score density of a PaO model?

3 Methodology

To systematically evaluate the operational fragility of cost-sensitive classifiers under dataset shift, we ground our experimental framework in customer churn prediction. Churn management provides an ideal empirical setting due to its highly asymmetric financial consequences—wasting retention incentives on loyal customers versus forfeiting the lifetime value of actual defectors [17]. This section formalizes the economic parameters of this proactive retention scenario, details the translation of predictive scores into operational threshold policies, and outlines the synthetic drift architecture utilized to simulate adverse macroeconomic shocks.

3.1 Cost in Churn Management

We model a standard proactive campaign where an enterprise attempts to prevent attrition by offering predicted churners ($y_i = 1$) a financial incentive to remain. The economic consequences of a model’s interventions are formalized in a cost matrix relative to a baseline of inaction (Table 1). This relative formulation isolates the net impact of the campaign, treating financial gains as negative costs.

Adopted from Janssens et al. [11], the cost of a False Positive ($C_{FP,i}$) represents wasted expenditure comprising an individualized incentive cost (d_i) and a

Table 1. The instance-dependent cost matrix for a proactive customer retention campaign. Financial benefits (retained customer value) are expressed as negative costs. For class-dependent costs, the individualized parameters (CLV_i and d_i) are replaced by their global averages.

	Actual Positive ($y = 1$)	Actual Negative ($y = 0$)
Pred. Positive ($\hat{y} = 1$)	$C_{TP,i} = (1 - \gamma)f$ $-\gamma(CL V_i - d_i - f)$	$C_{FP,i} = d_i + f$
Pred. Negative ($\hat{y} = 0$)	$C_{FN,i} = 0$	$C_{TN,i} = 0$

fixed administrative contact cost (f). Conversely, a True Positive ($C_{TP,i}$) represents the expected net gain of successfully retaining a churner, explicitly factoring in their Customer Lifetime Value (CLV_i) and their probability of accepting the offer (γ). Following established specifications [18,11], we set $f = 1$, a static acceptance rate of $\gamma = 0.3$, and scale the retention incentive proportionally to individual value ($d_i = 0.05 \cdot CLV_i$).

To rigorously evaluate the impact of this economic granularity, we contrast this instance-dependent formulation against a class-dependent baseline. In the class-dependent approach, the individualized CLV_i is replaced by the global average Customer Lifetime Value ($\overline{CLV}$) across the training population. This aggregation yields static, uniform penalty terms (C_{TP} and C_{FP}) that treat all successful retentions and wasted incentives symmetrically, irrespective of a customer’s specific economic profile.

To optimize the models in the PaO approach, the objective function minimizes the Average Expected Cost (AEC). Given N instances and a continuous model output $\hat{y}_i \in [0, 1]$, the empirical risk is evaluated using the cost parameters corresponding to the chosen granularity (utilizing either the individualized $C_{TP,i}$ and $C_{FP,i}$, or the static C_{TP} and C_{FP}):

$$\text{AEC} = \frac{1}{N} \sum_{i=1}^N [y_i \hat{y}_i C_{TP,i} + (1 - y_i) \hat{y}_i C_{FP,i}] \quad (1)$$

3.2 The Two Decision Approaches: Theoretical vs. Empirical Thresholds

Because the PtO and PaO approaches optimize fundamentally different loss functions during training, they output distinct types of prediction scores. This architectural divergence necessitates specific operational policies to translate these scores into binary business actions (Figure 1).

Predict-then-Optimize (Theoretical Policy) PtO models minimize standard log-loss during training, outputting calibrated conditional probabilities ($\hat{p}_i$).

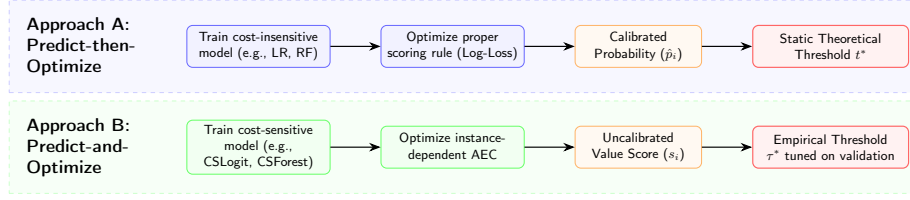


Fig. 1. The architecture of threshold policies. PtO models leverage proper scoring rules to establish theoretical thresholds. Conversely, PaO models output uncalibrated scores, mandating the use of empirical thresholds derived from historical validation data.

This enables the use of exact theoretical decision rules. By applying Bayes Minimum Risk [4], we first calculate the optimal, instance-specific theoretical threshold for each customer in the validation set:

$$t_i^* = \frac{C_{FP,i}}{-C_{TP,i} + C_{FP,i}} \quad (2)$$

However, to deploy a fixed, operational policy on unseen test data, we extract the median of these individual boundaries to establish a single static campaign threshold, denoted as t^* . An intervention is subsequently executed in the deployment environment strictly if the calibrated prediction $\hat{p}_i > t^*$.

Predict-and-Optimize (Empirical Policy) PaO models explicitly minimize the Average Expected Cost (AEC) during training. Because the AEC is an improper scoring rule, the resulting model outputs uncalibrated economic value scores ($s_i \in [0, 1]$) rather than probabilities. Consequently, applying the Bayesian formula (Eq. 2) to these scores is mathematically invalid. Instead, the PaO approach requires an empirical policy: the uncalibrated scores are systematically swept over a historical validation fold to identify the static, empirical threshold (τ^*) that minimizes the total AEC in-distribution.

3.3 Base Models and Empirical Dataset

To ensure a robust comparative analysis across varying geometric algorithmic properties, we select four standard machine learning architectures: Logistic Regression (LR), Decision Trees (DT), Random Forests (RF), and LightGBM (GBM). These algorithms serve as the cost-insensitive baselines within the PtO approach. They are evaluated directly against their specialized, cost-sensitive counterparts within the PaO approach: Cost-Sensitive Logistic Regression (CSLogit), Cost-Sensitive Decision Trees (CSTree), Cost-Sensitive Random Forests (CSForest), and Cost-Sensitive Gradient Boosting (CSBoost) [8,2].

We utilize the Iranian Churn Dataset [1], encompassing 3,150 customers and 13 attributes over a 12-month period. This specific dataset was selected because it provides continuous financial variables necessary for computing precise

CLV—an absolute prerequisite for instance-dependent cost-sensitive optimization. Furthermore, it contains highly granular continuous usage features (e.g., call durations, SMS frequency) required to effectively simulate structural feature decoupling and scale-based data drifts. To mirror proactive retention cycles, predictive features are aggregated over the first 9 months. The target label designates the churn state at the end of the 12th month, reflecting an operational 3-month intervention window.

3.4 Experimental Setup and Synthetic Drift Architecture

To guarantee statistical robustness and systematically benchmark policy decay, the experimental framework utilizes a continuous 13-dimensional Gaussian copula [12]. This generative architecture allows us to decouple the marginal cumulative distribution functions (CDFs) of empirical features from their multivariate dependency structure, creating a controlled sandbox to test out-of-distribution resilience.

As illustrated in Figure 2, the evaluation lifecycle is executed across 10 independent iterations, proceeding through two distinct chronological phases.

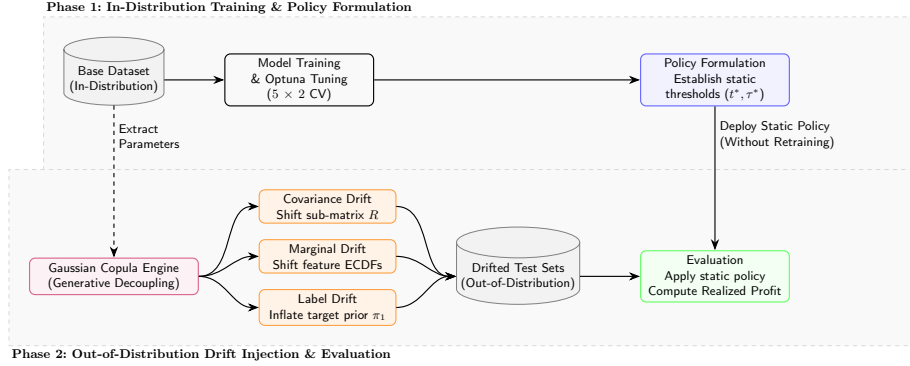


Fig. 2. The longitudinal evaluation lifecycle. Base models and operational policies are strictly optimized and established on the in-distribution dataset. The Gaussian Copula engine subsequently spawns out-of-distribution test sets to benchmark the economic decay of those static policies.

Phase 1: In-Distribution Training and Policy Formulation. First, the copula engine generates an empirical base dataset representing stable, in-distribution historical data. For each base dataset, hyperparameter optimization is conducted across all model architectures via the Tree-structured Parzen Estimator (TPE) [5]. To align with the operational deployment environment, the objective function strictly minimizes the Average Expected Cost (AEC) using a 5×2 cross-validation procedure.

Crucially, the final operational policies are explicitly calculated and established on this base dataset prior to out-of-distribution deployment. For PtO

models, the theoretical Bayes threshold (t^*) is derived from the calibrated baseline probabilities. For PaO models, the empirical thresholds (τ^*) are tuned via cross-validation over the baseline economic scores.

Phase 2: Out-of-Distribution Drift Injection and Evaluation. Once the static operational policies are established, the models are subjected to dataset shift. Utilizing the mathematical decoupling properties of the Gaussian copula, we spawn multiple out-of-distribution test sets from the original base distribution, systematically inducing three specific types of macroeconomic drift:

1. **Covariance Drift (Structural Shock):** We simulate the sudden decoupling of customer usage behaviors, such as shifts to flat-rate pricing models. We hold the marginal distributions constant and isolate the sub-matrix R of highly correlated usage and financial features within the copula’s correlation structure. We apply a Fisher z-transformation to negatively scale the off-diagonal Pearson correlation coefficients by a shift strength of 0.80, utilizing Higham’s alternating projections algorithm to ensure the modified matrix remains positive semi-definite [7].
2. **Marginal Drift (Scale Shock):** We simulate an absolute macroeconomic decline, such as an economic recession causing a global contraction in telecommunications usage. We hold the correlation dependency matrix R entirely constant. Instead, we isolate the continuous usage variables and scale their absolute margins by a shift strength of 0.80, systematically shifting the Empirical CDFs without altering internal feature relationships.
3. **Label Drift (Base-Rate Shock):** We simulate a sudden influx of churners caused by aggressive competitor marketing. We artificially inflate the global prior probability of the target class (π_1) by a factor of 1.20. We resample the test set to reflect this new prior distribution while leaving the conditional feature distributions $P(X|Y)$ mathematically untouched.

To evaluate true economic resilience, the static policies formulated in Phase 1 are deployed directly onto these drifted test sets.

3.5 Evaluation Metrics

To evaluate the true economic resilience of the competing operational policies under adverse dataset shift, our primary metric is the Realized Profit. Subsequently, to empirically disentangle the underlying failure mechanisms driving any observed financial decay, we track specific diagnostic metrics tailored to the mathematical vulnerabilities of both the PtO and PaO approaches.

1. Realized Profit: Ultimately, the success of a proactive churn campaign is measured by its net financial return. Let $\mathcal{T}$ represent the subset of customers selected for intervention by the fixed deployment policy (i.e., customers exceeding the theoretical t^* or the empirical τ^*). For each drifted test set, we simulate a deterministic business outcome by sampling a true underlying campaign acceptance vector $a_i \sim \text{Bernoulli}(0.3)$ for $i \in \mathcal{T}$. The Realized Profit aggregates the net financial gains and wasted expenditures generated across the targeted

portfolio by explicitly separating accepted offers from rejected ones:

$$\text{Realized Profit} = \sum_{i \in \mathcal{T}} [y_i a_i (CLV_i - d_i - f) - y_i (1 - a_i) f - (1 - y_i) C_{FP,i}] \quad (3)$$

While Realized Profit establishes the macroscopic financial impact, we utilize the following three diagnostic metrics to isolate the exact architectural failures driving that profit loss:

2. Brier Score (Calibration Loss): The PtO approach relies on absolute probabilistic accuracy to validate its theoretical Bayes threshold. We compute the Brier Score—the mean squared difference between the predicted probabilities $\hat{p}_i$ and the actual outcomes y_i —to precisely measure the destruction of calibration under drift.

$$BS = \frac{1}{N} \sum_{i=1}^N (\hat{p}_i - y_i)^2 \quad (4)$$

3. Wasserstein Distance (Density Misalignment): The PaO approach relies on its uncalibrated score density remaining stable so the fixed empirical threshold (τ^*) remains valid. We calculate the 1D Wasserstein Distance (Earth Mover’s Distance) between the validation scores S_{val} and the drifted test scores S_{test} . A high Wasserstein distance proves the score distribution has shifted, rendering the static threshold obsolete.

4. Target Volume Drift (ΔV): This metric translates theoretical density shifts into concrete operational terms. Let $\mathcal{T}$ be the subset of targeted customers. ΔV measures the directional percentage-point deviation in the proportion of the customer base selected for intervention between the validation and test phases:

$$\Delta V = \frac{|\mathcal{T}_{test}|}{N_{test}} - \frac{|\mathcal{T}_{val}|}{N_{val}} \quad (5)$$

This metric tracks whether threshold staleness causes a model to erroneously inflate its total campaign volume (a positive ΔV) or collapse its targeted cohort (a negative ΔV).

4 Results

The empirical evaluation highlights a fundamental operational trade-off in cost-sensitive classification: the modeling approach that maximizes initial financial return in-distribution exhibits the highest sensitivity to structural dataset shift.

To establish a clear macroscopic understanding of these decay mechanisms, the initial analysis presents an aggregated view. Specifically, the first two subsections evaluate the overarching performance of the Predict-then-Optimize (PtO) and Predict-and-Optimize (PaO) paradigms by averaging outcomes across all underlying base algorithms (e.g., Logistic Regression, Random Forests, Gradient Boosting) and cost granularities (class-dependent versus instance-dependent optimization). Subsequent sections systematically disentangle these dimensions to isolate the structural robustness of specific algorithmic architectures and the exact impact of cost granularity.

4.1 Predictive Performance and Pipeline Degradation

To establish the baseline differences between the two pipelines, Table 2 reports the average realized profit under in-distribution conditions across all algorithms compared to the three out-of-distribution drift scenarios.

Table 2. Average realized profit in units per customer across operational pipelines. Values in parentheses indicate the absolute drop or gain relative to the baseline.

Pipeline	Baseline	Covariance (0.8)	Marginal (0.8)	Label (1.2)
PtO (Cost-Insensitive)	4.48	2.96 (-1.52)	1.78 (-2.70)	5.17 (+0.69)
PaO (Cost-Sensitive)	3.28	2.92 (-0.36)	2.16 (-1.12)	4.23 (+0.95)

Under strictly in-distribution conditions, the PtO pipeline yields a superior average baseline profit of 4.48 units per customer, outperforming the PaO pipeline (3.28 units). Because PtO relies on proper scoring rules to generate calibrated probabilities, applying theoretical Bayes Minimum Risk thresholds yields mathematically optimal decisions when the test data matches the training distribution.

However, this reliance on exact calibration introduces severe vulnerabilities under distribution shift. When subjected to structural covariance drift (a shift strength of 0.80), PtO models suffer an average profit reduction of 1.52 units. In contrast, PaO models demonstrate substantial robustness, degrading by only 0.36 units. By explicitly integrating asymmetric costs into the loss function during Empirical Risk Minimization, PaO representations prioritize economically vital feature interactions, maintaining their ordinal ranking even when peripheral correlations decouple.

This vulnerability is most pronounced under marginal drift. When the absolute scale of continuous features (e.g., usage metrics) contracts, the degradation of the PtO approach is so severe that its profit drops below that of the PaO approach (1.78 vs. 2.16 units). Conversely, when macroeconomic base-rates change (Label Drift of 1.20), both approaches exhibit an absolute increase in realized profit. It is important to note that this aggregate increase is structurally driven by the evaluation framework: because average profit is normalized across the entire customer base, an artificially inflated proportion of churners inherently raises the total pool of retainable value. Nevertheless, PtO models successfully capture significantly more of this upside. Because their outputs are probabilities, they naturally scale with base-rate influxes, allowing the Bayes threshold to instantly expand the targeted cohort and generate 5.17 units of profit, outpacing the PaO approach.

4.2 Mechanisms of Policy Decay

To isolate why these approaches fail, we track diagnostic metrics tailored to their respective operational policies (Table 3). Because the PaO approach optimizes an improper scoring rule, its outputs are uncalibrated economic scores. Consequently, evaluating it with a Brier score is mathematically invalid. Instead, we evaluate PtO degradation via calibration loss (Brier Score), and PaO degradation via threshold staleness (Wasserstein Distance and Target Volume Drift).

Table 3. Diagnostic metrics isolating the failure mechanisms of each approach. PtO decay is measured via calibration loss (Brier Score). PaO decay is measured via score density shift (Wasserstein Distance) and target volume drift (ΔV).

Drift Scenario	PtO Diagnostics		PaO Diagnostics	
	Base Brier	Drift Brier	Wasserstein	Target Vol. Drift (ΔV)
Covariance (0.8)	0.037	0.076	0.006	0.000
Marginal (0.8)	0.037	0.079	0.015	0.018
Label (1.2)	0.037	0.046	0.015	0.020

Under covariance drift, the source of the PtO approach’s profit collapse is evident: its Brier score more than doubles from 0.037 to 0.076, rendering the theoretical Bayes thresholds invalid. In contrast, the PaO models exhibit a Wasserstein distance of only 0.006, indicating that the uncalibrated score density remained highly stable. This stability results in zero target volume drift ($\Delta V = 0.000$), meaning the campaign size neither erroneously inflated nor collapsed.

Under marginal drift, both approaches exhibit significant density warping. For the PaO approach, this misalignment causes the static empirical threshold to erroneously inflate the campaign volume ($\Delta V = +0.018$), effectively wasting retention budget on false positives and driving the mutual collapse in profitability observed in Table 2. Conversely, under label drift, the positive volume shift ($\Delta V = +0.020$) reflects a structurally valid and necessary expansion of the campaign to accommodate the macroeconomic influx of true churners.

4.3 Algorithmic Variance and Structural Robustness

Disaggregating the approaches reveals that economic robustness depends heavily on the geometric properties of the underlying base algorithms, as visualized in Figure 3.

LightGBM (GBM) exhibits the highest variance. Standard GBM achieves the maximum baseline profit (5.83) but collapses entirely under marginal drift (-2.91 absolute reduction). Cost-Sensitive Boosting (CSBoost) mitigates this volatility, trading peak baseline performance to restrict losses under covariance drift (reducing the drop from -2.58 to -0.85).

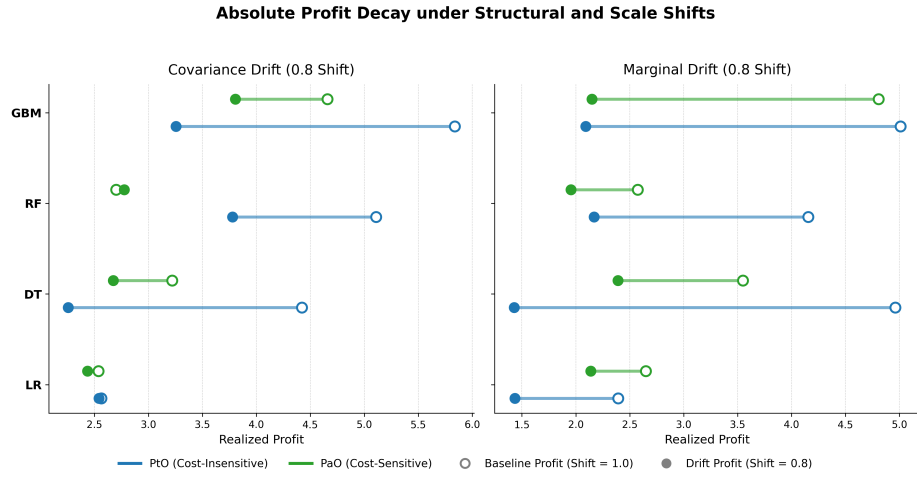


Fig. 3. Absolute profit decay under negative dataset shift across base algorithms. Hollow markers denote baseline profitability, while solid markers denote realized profit under shift. The length of the line visualizes the severity of the degradation.

Cost-Sensitive Random Forests (CSForest) exhibit the highest resistance to structural shock. While a standard Random Forest’s profit drops by -1.33 under covariance drift, CSForest maintains consistent performance, effectively neutralizing the impact of feature decoupling. Conversely, standard Decision Trees produce brittle step-function probabilities that are highly sensitive to absolute feature scaling, resulting in a severe profit reduction (-3.53) under marginal drift. Implementing the Cost-Sensitive Tree (CSTree) resolves this vulnerability by adjusting the splitting criterion, mitigating the loss to -1.15 . This heightened resilience across both CSForest and CSTree can likely be attributed to their architectural divergence from gradient-based methods. While CSBoost penalizes misclassifications by altering continuous gradients, the tree-based models fundamentally alter the recursive partitioning process by replacing standard metrics (such as Gini impurity) with a direct cost impurity measure. This discrete, non-gradient optimization mechanism appears structurally more robust to the continuous spatial warping induced by dataset shift.

4.4 Impact of Cost Granularity: Class vs. Instance Optimization

The analysis identifies a significant interaction effect regarding the granularity of the cost parameters utilized during model training (Figure 4).

During scale-based shocks (Marginal Drift), Instance-Dependent optimization demonstrates clear superiority. Relying on average class costs caused algorithms to disproportionately drop customers below the operational threshold as raw usage metrics contracted. Integrating the exact, individualized Customer Lifetime Value (CLV) acts as a mathematical anchor. The elevated gradient

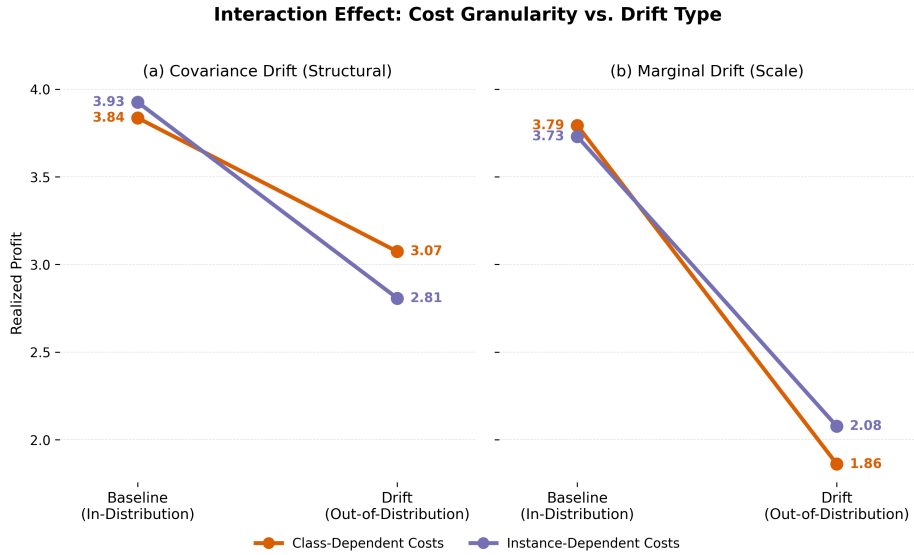


Fig. 4. The interaction effect between cost granularity and drift type. Panel (a) demonstrates the stabilizing effect of Class-Dependent costs under Covariance drift. Panel (b) illustrates the robustness provided by Instance-Dependent costs under Marginal drift.

penalty for misclassifying a high-value customer forces the model to maintain targeted interventions, systematically reducing financial losses compared to Class-Dependent models.

However, during structural shocks (Covariance Drift), Class-Dependent optimization yields greater stability for high-variance architectures such as Trees and Forests. When internal feature correlations decouple, optimizing for hyper-specific individual CLVs introduces excess variance, causing the algorithms to overfit to the disrupted feature space. In this context, utilizing a global average cost provides a macro-level smoothing effect that restricts the algorithm from capturing spurious interactions, thereby preserving out-of-distribution profitability.

5 Discussion

The results of this study recontextualize the consensus surrounding cost-sensitive classification. Recent large-scale benchmarking, such as Vanderschueren et al. (2022) [16], concludes that the Predict-then-Optimize (PtO) approach is the universally dominant strategy, positing that task-specific loss functions (Predict-and-Optimize) do not yield better downstream decisions.

While our baseline results align with this in-distribution supremacy, our longitudinal evaluation reveals a critical operational blind spot. PtO’s profitability is strictly contingent upon perfect probabilistic calibration—a property highly

susceptible to environmental decay. Under structural dataset shifts, the prerequisite for PtO’s success (the validity of the theoretical Bayes threshold) collapses. Consequently, the true value of the PaO approach lies not in maximizing in-distribution returns, but in establishing a robust, uncalibrated score density that defends against structural uncertainty.

5.1 Revisiting Class vs. Instance Optimization

Prior literature strongly advocates integrating instance-dependent rather than class-dependent costs to minimize overall expected costs [16]. Our findings validate but significantly amend this recommendation depending on the geometry of the environmental shock.

Under scale-based shocks (Marginal or Label Drift), instance-dependent optimization is superior. The individualized Customer Lifetime Value (CLV) provides a mathematical anchor; the heightened gradient penalties associated with misclassifying high-value individuals force the model to maintain targeted interventions even as overall usage metrics systematically decline.

Conversely, under structural shocks (Covariance Drift), we identify a critical divergence from the established literature. When historical feature correlations decouple, optimizing for hyper-specific individual CLVs introduces excess variance, causing high-capacity algorithms to overfit the disrupted feature space. In these highly volatile environments, class-dependent costs prove superior. It provides a macro-level smoothing effect that restricts algorithms from capturing spurious interactions within decoupled feature sets.

5.2 Strategic Implications for Model Deployment

Selecting an operational approach must transcend static performance benchmarking and instead align with the anticipated volatility of the deployment environment. For instance, when external market dynamics—such as a competitor launching an aggressive acquisition campaign—alter target class prevalence without fundamentally changing conditional feature distributions, PtO architectures maintain a distinct advantage. Because these models output proper probabilities, they naturally scale with shifting priors. This mathematical property allows the Bayes Minimum Risk threshold to dynamically expand the targeted cohort to capture the macroeconomic influx of true churners. Conversely, PaO models rely on static empirical thresholds that artificially constrain intervention volumes under base-rate inflation, resulting in substantial forfeited revenue.

However, the probabilistic calibration that makes PtO models resilient to base-rate fluctuations becomes a severe liability when the environment undergoes structural disruptions. If an industry-wide transition, such as moving from pay-per-use to flat-rate subscription models, breaks the historical correlations between predictive features, PtO calibration severely degrades. To preserve profitability under this covariance drift, practitioners must shift toward PaO approaches utilizing Cost-Sensitive Ensembles (e.g., CSForest). Furthermore, applying class-dependent costs during training provides a necessary macro-level

regularization effect, preventing high-capacity algorithms from overfitting to the newly decoupled feature space.

This regularization strategy must be deliberately inverted when the environmental shock manifests as a scale contraction rather than feature decoupling. When populations experience systematic contractions in absolute feature scales—such as a macroeconomic downturn causing widespread reductions in discretionary service usage—standard models generate artificially suppressed risk scores, effectively halting proactive campaigns. Enterprises must counter this marginal drift by deploying PaO models trained specifically with instance-dependent costs. The individualized financial penalty functions as an operational safeguard, forcing the algorithm to continue targeting high-value customers despite the macroscopic decay in their behavioral metrics.

6 Conclusion

This research demonstrates that the evaluation of cost-sensitive classifiers must extend beyond static, in-distribution performance metrics to encompass structural robustness over time. By rigorously decoupling dataset shift into marginal, covariance, and label drift utilizing a continuous Gaussian copula architecture, we map the distinct failure mechanisms of the two dominant operational approaches.

Our findings confirm that there is no universally dominant strategy for cost-sensitive classification. The Predict-then-Optimize paradigm maximizes potential financial return under stable conditions but suffers severe economic degradation when its probabilistic calibration is compromised by structural or scale-based shocks. Conversely, the Predict-and-Optimize paradigm sacrifices peak baseline performance to establish a resilient, uncalibrated score density that withstands feature decoupling. Ultimately, this research provides a strategic framework to guide practitioners in selecting the optimal modeling approach based on anticipated changes in their competitive environment. Rather than relying on a singular methodology, organizations can proactively safeguard their retention campaigns by deliberately aligning the algorithmic properties and cost-granularity of their classifiers with the specific macroeconomic and behavioral shifts anticipated in production.

7 Limitations and Future Work

While this study isolates the mechanisms of policy decay by evaluating static operational policies, it inherently assumes a worst-case deployment scenario where models remain unmonitored. In mature operational environments, thresholds are periodically audited. Future research should investigate the financial viability of dynamic thresholding techniques under dataset shift—specifically, determining whether continuously updating the empirical threshold (τ^*) of a PaO model using recent batches of unlabelled data can successfully restore profitability without necessitating complete model retraining. Additionally, extending this

experimental framework to other financial domains characterized by asymmetric misclassification costs, such as credit scoring and fraud detection, warrants further empirical investigation to establish the generalizability of these findings.

References

1. Iranian Churn. UCI Machine Learning Repository (2020), DOI: <https://doi.org/10.24432/C5JW3Z>
2. Bahnsen, A.C., Aouada, D., Ottersten, B.: Example-dependent cost-sensitive decision trees. *Expert Systems with Applications* **42**(19), 6609–6619 (2015)
3. Bahnsen, A.C., Aouada, D., Ottersten, B.: A novel cost-sensitive framework for customer churn predictive modeling. *Decision Analytics* **2**(1), 5 (2015)
4. Bahnsen, A.C., Stojanovic, A., Aouada, D., Ottersten, B.: Cost sensitive credit card fraud detection using bayes minimum risk. In: 2013 12th international conference on machine learning and applications. vol. 1, pp. 333–338. IEEE (2013)
5. Bergstra, J., Bardenet, R., Bengio, Y., Kégl, B.: Algorithms for hyper-parameter optimization. *Advances in neural information processing systems* **24** (2011)
6. Hand, D.J.: Measuring classifier performance: a coherent alternative to the area under the roc curve. *Machine learning* **77**(1), 103–123 (2009)
7. Higham, N.J.: Computing the nearest correlation matrix—a problem from finance. *IMA journal of Numerical Analysis* **22**(3), 329–343 (2002)
8. Höppner, S., Baesens, B., Verbeke, W., Verdonck, T.: Instance-dependent cost-sensitive learning for detecting transfer fraud. *European Journal of Operational Research* **297**(1), 291–300 (2022)
9. Höppner, S., Stripling, E., Baesens, B., vanden Broucke, S., Verdonck, T.: Profit driven decision trees for churn prediction. *European journal of operational research* **284**(3), 920–933 (2020)
10. Jahromi, A.T., Stakhovych, S., Ewing, M.: Managing b2b customer churn, retention and profitability. *Industrial Marketing Management* **43**(7), 1258–1268 (2014)
11. Janssens, B., Bogaert, M., Bague, A., Van den Poel, D.: B2boost: Instance-dependent profit-driven modelling of b2b churn. *Annals of Operations Research* **341**(1), 267–293 (2024)
12. Nelsen, R.B.: An introduction to copulas. Springer (2006)
13. Patki, N., Wedge, R., Veeramachaneni, K.: The synthetic data vault. In: 2016 IEEE international conference on data science and advanced analytics (DSAA). pp. 399–410. IEEE (2016)
14. Risselada, H., Verhoef, P.C., Bijmolt, T.H.: Staying power of churn prediction models. *Journal of Interactive Marketing* **24**(3), 198–208 (2010)
15. Stripling, E., vanden Broucke, S., Antonio, K., Baesens, B., Snoeck, M.: Profit maximizing logistic model for customer churn prediction using genetic algorithms. *Swarm and Evolutionary Computation* **40**, 116–130 (2018)
16. Vanderschueren, T., Verdonck, T., Baesens, B., Verbeke, W.: Predict-then-optimize or predict-and-optimize? an empirical evaluation of cost-sensitive learning strategies. *Information sciences* **594**, 400–415 (2022)
17. Verbeke, W., Dejaeger, K., Martens, D., Hur, J., Baesens, B.: New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European journal of operational research* **218**(1), 211–229 (2012)
18. Verbraken, T., Verbeke, W., Baesens, B.: A novel profit maximizing metric for measuring classification performance of customer churn prediction models. *IEEE transactions on knowledge and data engineering* **25**(5), 961–973 (2012)