Proceedings of ICFTBA 2026 Symposium: Strategic Management & Business Analytics: Data-Driven Performance Models
DOI: 10.54254/2754-1169/2026.LD36622

Robust Churn Retention for Mature E-commerce Customers: A Noise-Resilient Framework with Survival-Based Intervention Timing

Rong Li

School of Accounting, Beijing Wuzi University, Beijing, China
Guoy3604@gmail.com

Abstract. Current churn prediction research mainly emphasizes static accuracy on curated datasets, and they often overlook label noise arising from heuristic CRM rules, uncontrollable user attrition, and suboptimal retention timing. This study proposes a robust two-stage retention framework targeting e-commerce users with tenure of more than four years. Firstly, structural churn cases are filtered from the dataset. Then, 5% label noise is added to simulate real-world annotation errors, and benchmark evaluations are conducted across nine machine learning models. Ensemble methods, specifically GBM, XGBoost, and LightGBM, achieve F1 scores ranging from 0.43 to 0.44 with cross-validation standard deviations below 0.02. About the method, the analysis progresses from binary classification to Cox proportional hazards modeling, identifying a proactive intervention window between the fourth and fifth year of user tenure. Business-oriented threshold tuning enables GBM campaigns to attain a 345.7% return on investment. This significantly surpasses simple methods. Further analysis via SHAP values shows service call fatigue is the primary churn driver. The assumed impact of fee sensitivity lacks empirical support within this mature cohort. The proposed project effectively bridges statistical prediction and operational retention, providing an effective methodology for protecting high-value customer assets.

Keywords: Customer churn, Label noise, Survival analysis, E-commerce, Return on investment

1. Introduction

Customer churn prediction is a core revenue-retention instrument in e-commerce and finance, particularly as customer acquisition costs continue to climb. Mature customers constitute a high-value segment, with prior estimates placing their lifetime value at 5 to 10 times that of new customers. Although this group is essential, the long-term group remains underestimated by others, and this typically benchmarks on mixed-tenure populations. Most retention strategies sourced from this perform poorly in production; here are three interrelated reasons.

First, heuristic churn labels (e.g., the "180 days inactive" rule) are always thought of as no error, despite injecting systematic mislabeling into CRM datasets. Previous research noted in a systematic review of 240 churn studies that fewer than 12% evaluate under noisy labels and only 6% optimize

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2754-1169/2026.LD36622

for profit—leaving a pronounced lab–production gap [1]. Prior work categorized this CRM-induced mislabeling as asymmetric noise, estimating real-world rates at 3–8%, yet 91% of benchmarks reviewed still train on curated data [2]. The fragility this creates is observable even in robust pipelines: previous studies benchmarked DT, ANN, KNN, and SVM on an Iranian telco dataset and proposed a hybrid ensemble that pushed recall and precision above 95%, but conceded that performance was measured on sanitized labels—an optimism bias the present study explicitly corrects via 5% asymmetric noise injection [3].

Second, few frameworks separate structural churn from persuadable churn. Mixing the two diverts retention spend toward unreachable users. This conflation persists partly because e-commerce churn research remains sparse relative to telecom and banking. Previous research modeled a Brazilian e-commerce retailer via XGBoost and logistic regression with AUC/lift metrics, demonstrating that churn prediction in contract-free settings demands different feature constructions than telcos' 2–3% churn-rate regime—yet their framing stayed binary and did not partition structural attrition [4].

Third, the field remains tethered to static metrics (Accuracy, AUC, F1), which quantify snapshot performance but convey no timing signal. Survival analysis addresses this: prior work introduced the Proportional Hazards model for right-censored data, later applied to retail banking by subsequent research [5, 6]. E-commerce transpositions of survival modelling, however, remain uncommon and rarely paired with noisy-label stress tests. In addition, Interpretability tools are getting more mature —previous studies established SHAP, and later works validated its stability across 14 churn datasets [7, 8]. Recent telecom benchmarks combine five-tab learner benchmarks with nested-CV and SHAP. It reports a 92.28% accuracy and a tenure tipping point near 3 years. This study is built on a framework that includes ensemble machine learning, Cox proportional hazards modeling, and SHAP-based feature attribution under controlled label noise. It examines what makes churn among e-commerce customers with four or more years of platform tenure. It aims to identify a practical intervention opportunity instead of creating assumptions. It proves that fee sensitivity plays an important role in their attrition [8].

2. Methodology

To visually compare the contribution of this study and the written literature, Table 1 makes a comparison of characteristics of previous churn prediction with the proposed methodology in five different dimensions. The comparison highlights the change from clean-label assumptions to noise injection, from binary classification to survival analysis, and from accuracy metrics to ROI-driven optimization.

Table 1. Comparison of existing literature and the proposed framework

Dimension	Existing Work	The paper Contribution
Data Assumptions	Clean labels, static snapshots	5% label noise injection simulating CRM heuristics
Modeling Task	Binary Classification	Survival Analysis (Cox Proportional Hazards)
Evaluation Metrics	Accuracy, F1, AUC	Net Profit, ROI

DOI: 10.54254/2754-1169/2026.LD36622

Table 1. (continued)

Interpretability	Post-hoc SHAP for feature ranking	Hypothesis Testing
Business Alignment	Generic retention targeting	Two-Stage Filtering

2.1. Data and feature engineering

The empirical analysis uses the E-commerce Customer Churn Dataset. It provides an overall view of customers across a global subscription platform [9]. The raw file has 50,000 customer records, including 25 different behavioral, demographic, and transactional features.

To select the high-value demographic of people, a stratified sample of 13,085 mature-user records is extracted by filtering for users with four or more years of continuous membership. This group indicates a 21.4% churn rate at the conclusion of the observation window. The reason for narrowing the scope to long-term users is fairly straightforward. Most churn models in the current literature are built on datasets that lump everyone together — people who signed up last month alongside those who have been around for years. The problem is, these two groups behave very differently, and a model tuned on the average of both tends to perform poorly on the high-value, long-tenure segment specifically. This study focuses on users with at least four years of history, which is a population that has not received much dedicated attention.

On the data side, one thing that became obvious early on was the shape of the Average Order Value distribution. It was not even close to symmetric. A handful of unusually large transactions — think bulk purchases or corporate accounts — dragged the mean well above what most orders actually looked like. Feeding raw AOV numbers into a model under those conditions would have been misleading, so the preprocessing step included a log transformation to tame that skew before anything else was done. Before model training, we also exclude users who are effectively unreachable through any retention intervention, separating behavioural attrition (reversible) from structural churn (irreversible). Structural churn is defined by two exclusion rules:

$$(CreditBalance = 0 \cap DaysSinceLastPurchase > 365) \cup (MembershipYears \geq 4 \cap TotalPurchases = 0) \quad (1)$$

In words, rule (1) flags users with neither credit balance nor a purchase in over a year, while rule (2) captures long-tenure members with zero purchases — e.g., a country no longer serviced. Both groups would only add noise to the training set.

These unreachable users are excluded because no retention measure — discount coupons, re-engagement emails, or otherwise — can recover them; keeping them in the training set would bias the model.

To keep things concrete, structural churn is identified through two exclusion rules, described below:

$$Fee\ Sensitivity\ Ratio = \frac{AOV}{MembershipYears} + \varepsilon \quad (2)$$

$$Service\ Call\ Fatigue = \frac{ServiceCalls}{TotalPurchases} + \varepsilon \quad (3)$$

DOI: 10.54254/2754-1169/2026.LD36622

$$LTV\ Per\ Year = \frac{LifetimeValue}{MembershipYears} + \varepsilon \tag{4}$$

$$Churn\ Urgency = \frac{DaysSinceLastPurchase}{MembershipYears \times 365} + \varepsilon \tag{5}$$

2.2. Noise injection and model setup

For the survival analysis component, the Cox Proportional Hazards (CPH) model is employed to model time-to-event distributions. Prior to model fitting, three features exhibiting strong temporal collinearity—namely Membership Years, Churn Urgency, and LTV Per Year—are excluded from the covariate set. This preprocessing step mitigates multicollinearity, which would otherwise bias the estimation of hazard ratios. The CPH model estimates the hazard function as:

$$h(t \mid X) = h0(t)exp(\beta\top X) \tag{6}$$

In this formulation, t denotes the specific time point measured in days since user registration. The term X represents the vector of covariates retained for model training, which includes behavioral metrics such as Service Call Fatigue and Cart Abandonment Rate. The vector β corresponds to the coefficients estimated via partial likelihood maximization; these coefficients quantify the directional and magnitude effects of each covariate on churn risk. Specifically, h0(t) defines the baseline hazard function, which captures the aggregate churn risk over the user tenure in the absence of explanatory variables. Consequently, the term exp(β⊤X) acts as a scaling factor applied to the baseline hazard.

2.3. ROI-driven threshold optimization

Rather than tuning the decision threshold for F1 or AUC, the threshold here is optimized for something more practical: net profit from retention campaigns. The logic is that a churn model with 0.92 AUC is worthless if the retention team can only afford to call 30% of the flagged users. Classification accuracy and business value are not the same thing, and this study optimizes for the latter.

The objective function:

$$Net\ Profit(\tau) = (TP \times 0.3 \times LTV) - Costs \tag{7}$$

A grid search over [0,1] in 0.01 steps finds the best τ. At each candidate threshold, true positives (TP) and false positives (FP) are computed based on the binarized prediction rule $(\hat{y} = \mathbb{I}(\widehat{P}(y = 1) \geq \tau))$. One hard constraint: no more than 30% of the user base can be targeted for outreach. This is not an arbitrary number — it reflects the actual headcount ceiling of a mid-size customer success team. Without it, the optimizer would happily "intervene" on 80% of users, which looks great on paper and is completely unworkable in practice.

3. Results

The framework is tested on 13,085 mature e-commerce customers. Nine models are benchmarked under the 5% label noise regime described earlier, survival analysis pinpoints when intervention is most effective, and ROI is computed for each model post-threshold optimization.

DOI: 10.54254/2754-1169/2026.LD36622

3.1. Model robustness

The data is split 70/30 (train/test), stratified on the churn label so the class ratio stays intact in both halves. All preprocessing sits inside Column Transformer pipelines — skip that, and the scaling statistics leak from the test fold back into training, which quietly invalidates every cross-validation number in the report.

Under 5% noise the ensembles hold up; the linear and kernel baselines do not. LightGBM posts the best single F1 (0.4326), XGBoost the most stable one (0.4429, CV standard deviation 0.0124), and GBM the tightest fold-to-fold spread of all nine (variance 0.0072). A paired t-test between LightGBM and XGBoost returns $p = 0.4931$, so declaring a winner between them would be over-reading the data. The three are statistically indistinguishable, which is exactly what the "ns" bracket over the leftmost bars in Figure 1(a) is marking. GBM ends up as the deployment pick not because it tops any leaderboard, but because it matches the other two on quality while costing less to train and serve.

Figure 1 puts the gap on display. Panel (a) plots test F1 with 5-fold CV error bars; panel (b) does the same for ROC-AUC, with the random-baseline line drawn in for reference. The ensemble trio clusters at the top of both panels. Naive Bayes and Decision Tree sit a clear step below, and DT in particular carries a huge error bar in panel (a), the visual signature of a model that is guessing differently every fold. Random Forest and SVM are absent from Figure 1 because both yielded F1 = 0.0000.

The two failures have different causes. Under corrupted labels, Random Forest's bootstrap sampling causes errors to accumulate across trees rather than cancel, driving the ensemble vote toward the wrong class. SVM over-regularizes, memorizing noisy training points and collapsing on clean test data. Either route lands at the same place — a model that cannot separate a churned user from an active one.

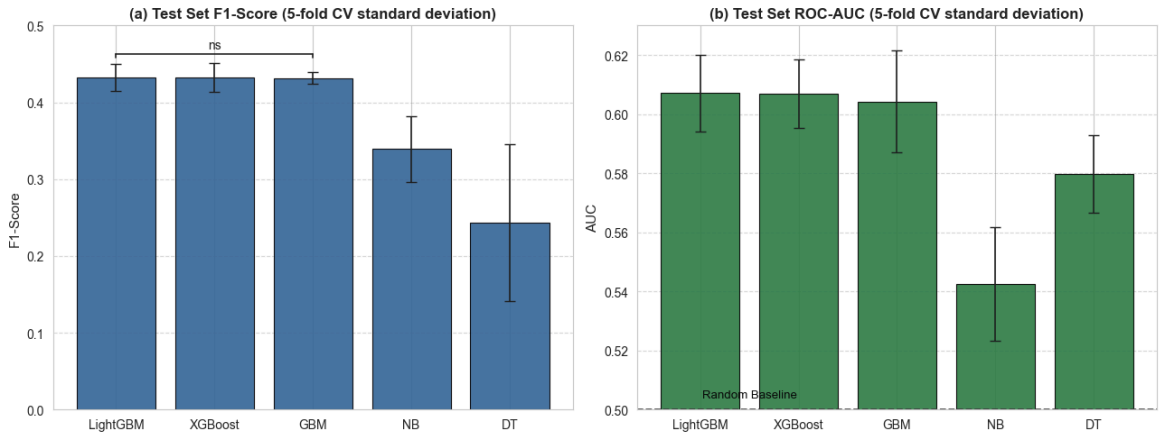

Figure 1. Model performance under 5% label noise (photo/picture credit: original)

3.2. ROI and survival analysis

Tuning the threshold for profit instead of for F1 pays off — literally. GBM yields a headline ROI of 345.7% ($3.5 net profit per $1 outreach cost). This superiority stems not from abstract classification power, but from its ability—under the 30% capacity cap—to prioritize high-value, at-risk users over indiscriminate discounting. The mechanism shows up cleanly in the false-positive rate among high-LTV customers: a false positive here is a coupon handed to someone who was never going to churn,

DOI: 10.54254/2754-1169/2026.LD36622

money burned on a save that was not needed. GBM leaks fewer of those, so its budget goes further. The full breakdown under a $50 per-user outreach cost is in Table 2, which lists net profit, retained-user counts, and ROI side by side for all nine models, so the gap between "statistically tidy" and "actually profitable" is easy to read off the page (Figure 2).

Table 2. Business ROI assessment

Model	Intervened	Retained	Cost ($)	Earnings ($)	Net Profit ($)	ROI (%)
GBM	785	111	39,250	174,938	135,688	345.7
RF	785	111	39,250	172,992	133,742	340.7
XGBoost	785	111	39,250	172,265	133,015	338.9
LightGBM	785	109	39,250	171,729	132,479	337.5
NB	785	90	39,250	75,332	36,082	91.9
LR	785	91	39,250	126,628	87,378	222.6
SVM	785	84	39,250	124,576	85,326	217.4
KNN	785	84	39,250	107,364	68,114	173.5
DT	2208	218	110,400	296,615	186,215	168.7

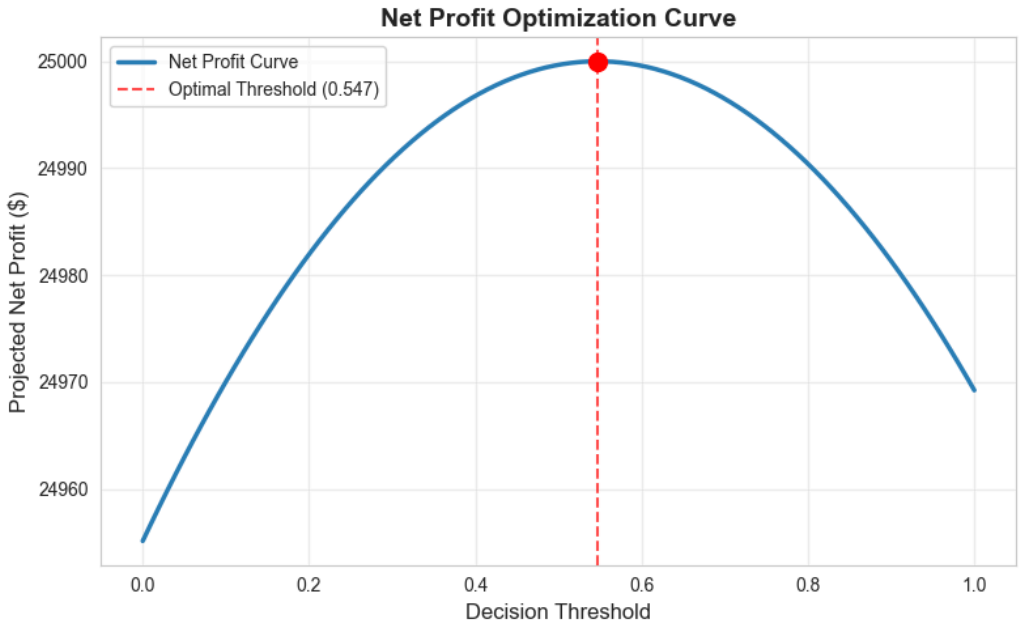

Figure 2. Threshold optimization for net profit (photo/picture credit: original)

Figure 3 identifies the 4–5 year membership window as the optimal point for proactive retention intervention. Higher service call fatigue is associated with a steeper decline in survival probability, and the performance gap between high-fatigue and low-fatigue user groups widens noticeably between 1,460 and 1,825 days of tenure.

DOI: 10.54254/2754-1169/2026.LD36622

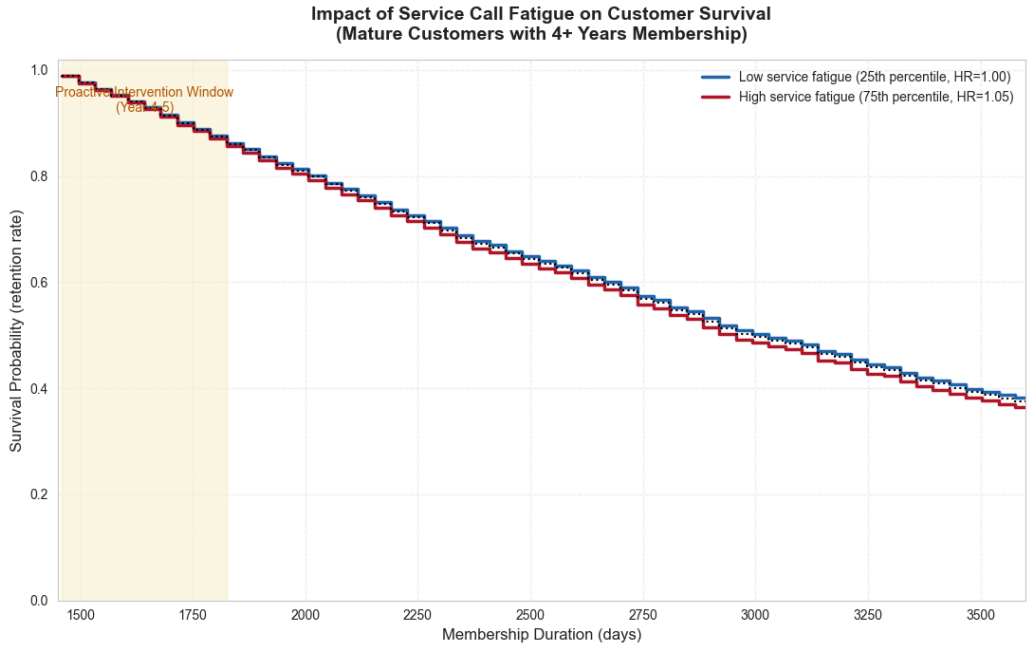

Figure 3. Survival curves for mature customers (shaded region: 4–5 year intervention window) (photo/picture credit: original)

3.3. SHAP feature importance

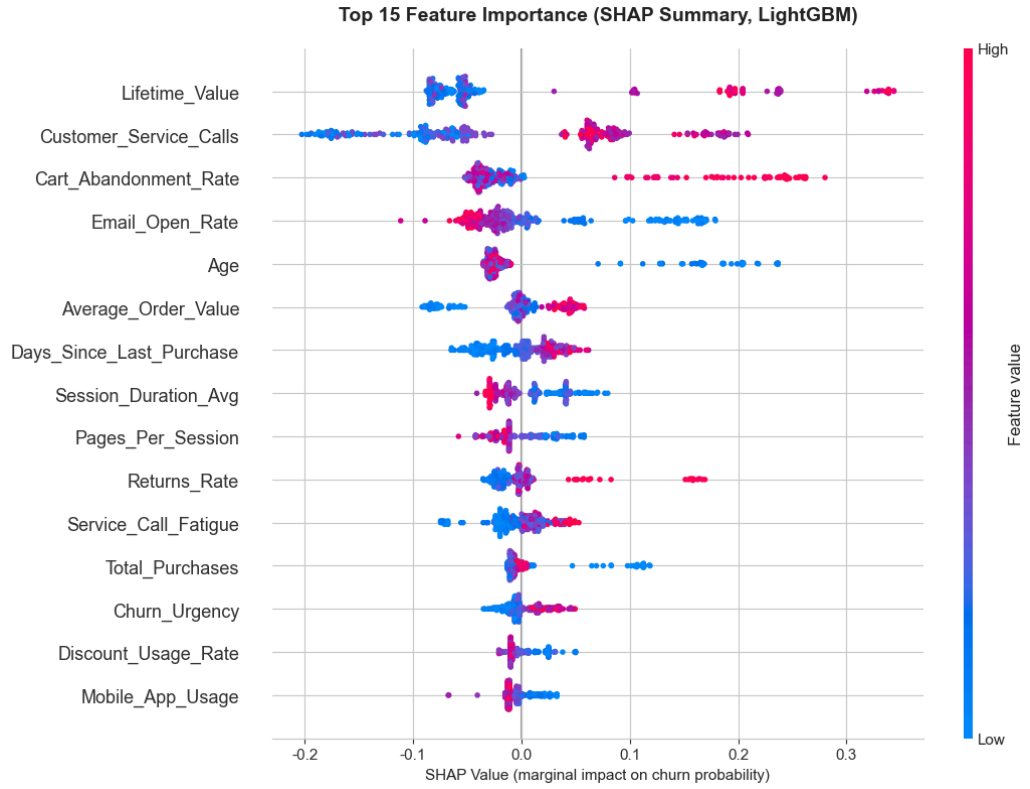

Figure 4. SHAP summary plot for GBM model (photo/picture credit: original)

DOI: 10.54254/2754-1169/2026.LD36622

SHAP analysis on the fitted model pulls out three features that genuinely move the churn probability, and the beeswarm in Figure 4 is where those effects stop being abstract. Among the levers a retention desk can actually pull, service call fatigue comes out strongest: high-fatigue customers get shoved into positive SHAP values, low-fatigue ones the opposite. Total customer service calls lean the same way, and cart abandonment rate closes the trio with the plot's most lopsided tail, high-abandonment users stretching out toward +0.3.

Conversely, price is negligible. The Fee Sensitivity Ratio is absent from Figure 4's top 15 features, and the proxy Discount Usage Rate shows no significant impact. This refutes the fee-sensitivity hypothesis for this mature cohort. For long-tenure users, the model is quietly ignoring the price tag and weighting service quality instead — a more useful takeaway than any extra decimal of accuracy, since it names the one knob worth turning.

4. Discussion

The framework keeps tenured, habit-driven shoppers from leaving — and that grip loosens the instant it meets a segment that actually shops on price. Less a bug than a side effect of behavioural inertia: long-tenure users just do not flinch at price the way a price-elastic newcomer would, so a model raised on them inherits the blindness. Out of that economic context, the feature ranking unravels, the same out-of-distribution collapse noted in [1]. The honest fix is to stop asking who churns and ask who churns over price, and only if price moves — a causal question, not a predictive one, which is the gap that heterogeneous treatment-effect models exist to fill. Future work should let a Generalized Random Forest split the retention budget on the fly: a kind word for the customers a kind word saves, a discount for the ones only a price cut will budge [9].

Five per cent label noise is survivable; past eight it is not, and the gap between them is a curve bending the wrong way. That decay is roughly what bagging theory predicts at its edge — feed an ensemble enough bad labels and the crowd starts agreeing on the wrong thing [2]. Cross-entropy will not fix it, since it trusts every label the same. Better to teach the model distrust: a noise-adaptive loss on one side, a self-supervised pass on the other, BYOL-style, where a consistency term holds the representation steady while the supervision lies. That unpins the features from the corrupted heuristic labels and claws back the headroom extreme noise otherwise eats.

The Cox model earns its keep by pinning the intervention window to four or five years — real temporal precision. The price is the proportional-hazards assumption, doing quiet work it cannot support for customers nearing end-of-lifecycle, where risk spikes in ways no constant multiplier tracks; such violations are well catalogued [10]. The next step is to drop the parametric straitjacket and let a network learn the hazard shape. DeepSurv-style neural survival models take time-varying covariates and non-linear risk in stride, so the late-stage acceleration lands on the curve instead of being smoothed flat [11].

A static structural filter is a fair-weather instrument — fine on stable tenure, useless the week the market sneezes, or the platform ships a breaking change, the brittleness critics of fixed feature engineering keep naming [7]. What it lacks is a loop; it decides once and trusts itself forever. A contextual-bandit layer closes that loop, reading the live reward signal and quietly rewriting its own thresholds as conditions drift — optimising for long-term equity, not for whatever the history happened to look like.

DOI: 10.54254/2754-1169/2026.LD36622

5. Conclusion

Acquisition keeps climbing, and the churn literature lags — three gaps mostly. It pretends the labels are clean, which no real CRM export ever is; it lumps the win-back-able together with the gone-for-good; and it names who is at risk without saying when to call. For members four-plus years in, the work runs one pipeline: a classifier that does not fall over on dirty labels, a survival model, and a threshold tuned on money, not accuracy.

On dirty labels, the tree ensembles do not blink — GBM holds at 5% noise, a live CRM's everyday mess, built for the field, not the benchmark. The two-stage filter crosses off the unreachable, so budget lands on the persuadable few, not spread thin over ghosts. Survival analysis names the window, the fourth to fifth year, risk climbing fast while the customer still picks up — the last honest chance. Tune the cutoff on return, and it reads 345.7% ROI: meaningless to a confusion matrix, everything to whoever signs the cheques. And the feature read tops service call fatigue with fee sensitivity flat, quietly bucking the pricing-first reflex the trade keeps reaching for.

None of this is e-commerce-only. Banking, telecoms, any subscription shop with a high-LTV base worth keeping can bolt the pipeline on; the two obvious next moves are recalibrating on a CRM's real annotation errors, not the synthetic 5%, and handing the threshold calls to a reinforcement-learning loop. The upshot is less a new accuracy score than a change of subject — churn prediction stops being a stats exercise and starts reading like strategy, on a workflow already stress-tested against real data's mess.

References

[1] Imani, M., Joudaki, M., Beikmohammadi, A., & Arabnia, H. R. (2025). Customer churn prediction: A systematic review of recent advances, trends, and challenges in machine learning and deep learning. *Machine Learning and Knowledge Extraction, 7*(3), 712–745.

[2] Han, B., Yao, Q., Liu, T., Niu, G., Tsang, I. W., Kwok, J. T., & Sugiyama, M. (2021). A survey of label-noise representation learning: Past, present and future. *arXiv*. https://arxiv.org/abs/2011.04406

[3] Keramati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadian, I., Mozaffari, M., & Abbasi, U. (2014). Improved churn prediction in telecommunication industry using data mining techniques. *Applied Soft Computing, 24*, 994–1012.

[4] Matuszelański, K., & Kopczewska, K. (2022). Customer churn in retail e-commerce business: Spatial and machine learning approach. *Journal of Theoretical and Applied Electronic Commerce Research, 17*(1), 165–198.

[5] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological), 34*(2), 187–202.

[6] Van den Poel, D., & Larivière, B. (2004). Customer attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research, 157*(1), 196–217.

[7] De Caigny, A., De Bock, K. W., & Verboven, S. (2024). Hybrid black-box classification for customer churn prediction with segmented interpretability analysis. *Decision Support Systems, 181*, Article 114217.

[8] Esmail, Z. (2023). E-commerce customer churn dataset [Data set]. *Kaggle*. https://www.kaggle.com/datasets/zakariaesmail/ecommerce-customer-churn-dataset

[9] Athey, S., & Imbens, G. W. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences, 113*(27), 7353–7360.

[10] Therneau, T., & Grambsch, P. (2000). *Modeling survival data: Extending the Cox model*. Springer Science & Business Media.

[11] Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology, 18*(1), Article 24.