

Explainable AI (SHAP) on XGBoost for E-Commerce Customer Churn: An Exploratory Case Study of Interpretability Under Data-Quality Constraints

Tikno¹, Adisti Maulidya Jannah², Vanisa Eka Bela³, Alden Muzakky Tri Vicienza⁴, and
Nancy Tamariska⁵

^{1,2,3,4,5}Information Systems Department, Universitas Internasional Semen Indonesia, Gresik, Indonesia

*Corresponding author: tikno@uisi.ac.id

ABSTRACT

Customer churn remains a critical challenge in e-commerce, yet predictive performance is fundamentally bounded by the quality of the underlying data and its labels. This study applies the XGBoost algorithm together with SHAP (SHapley Additive exPlanations) to build a churn prediction model that balances predictive capability with managerial interpretability, using the secondary public dataset "E-Commerce Customer Behavior Analysis" from Kaggle, consisting of 250,000 transaction records aggregated into 49,673 unique customer profiles. Class imbalance was addressed through XGBoost's built-in class-weighting mechanism, with hyperparameters optimized using Optuna. The best configuration achieved a Recall of 0.9909 at the F1-optimal threshold of 0.5733, but this figure is misleading in isolation: the accompanying ROC-AUC of 0.5070 and PR-AUC of 0.2028 confirm near-random discrimination. The root cause is a critical data validity limitation — the churn label carries no documented derivation methodology, and all features exhibited Pearson correlations below $|r| = 0.01$ against the target, consistent with suspected synthetic or randomized label generation. Despite near-random model performance, SHAP was applied in an exploratory manner to surface dominant behavioral patterns: Tenure, Recency, and Age emerged as the most influential features, directionally consistent with RFM-based churn theory. These findings are reported as exploratory hypotheses rather than confirmed causal drivers, and were translated into a set of segmented, hypothesis-driven CRM retention ideas pending empirical validation. The primary contribution of this study is methodological: demonstrating that interpretability tools retain value for hypothesis generation in data-constrained settings, while affirming that high-quality, operationally derived churn labels are a prerequisite for actionable predictive output.

Keywords: Customer Churn, XGBoost, SHAP, Explainable AI, scale_pos_weight, CRM, E-Commerce, Data Quality

1. INTRODUCTION

The e-commerce industry faces a persistent threat in the form of customer churn — the condition in which customers cease transacting or migrate to a competing platform. This problem carries significant economic weight: the Customer Acquisition Cost (CAC) is, on average, five times higher than the cost of retaining existing customers (Xiahou and Harada, 2022; Nagaraj et al., 2023), meaning that each churned customer erodes not only short-term transactional revenue but also Customer Lifetime Value (CLV) — the total economic value a customer generates over the course of their relationship with the company. The ability to

predict churn early and identify its behavioral drivers has therefore become a strategic business necessity, not merely a technical exercise (Gattermann-Itschert and Thonemann, 2022).

Existing literature on churn prediction has developed along two largely separate tracks: studies focused on maximizing predictive accuracy using ensemble algorithms such as XGBoost, Random Forest, and Support Vector Machine (SVM), and studies applying Explainable AI (XAI) methods predominantly in medical and financial domains (Vimbi, Shaffi and Mahmud, 2024; Alkhanbouli et al., 2025). Gradient-boosted tree ensembles such as XGBoost remain a preferred choice for structured, tabular business data of this kind, consistently matching or outperforming deep-learning alternatives while requiring less tuning effort (Shwartz-Ziv and Armon, 2022; Bentéjac, Csörgő and Martínez-Muñoz, 2021; Pasaribu, Yudistira and Mahmudy, 2024), and prior work specifically confirms XGBoost's robustness on imbalanced tabular targets akin to churn labels (Zhang, Jia and Shang, 2022). Research that explicitly integrates XGBoost and SHapley Additive exPlanations (SHAP) for e-commerce churn — while generating per-customer local explanations translatable into managerial action — remains limited, and few such studies disclose what happens when the underlying predictive signal itself is weak. The majority of prior churn studies rely solely on aggregate feature importance metrics under the implicit assumption of adequate model discrimination, stopping short of examining whether SHAP explanations remain interpretable evidence once that assumption fails.

This study addresses that gap by combining the predictive power of XGBoost with SHAP-based interpretability in an e-commerce churn context. Three research questions guide the inquiry: (1) how effective is XGBoost in predicting customer churn, measured through Recall, PR-AUC, and ROC-AUC; (2) which behavioral features most significantly drive churn decisions based on SHAP value analysis; and (3) what CRM-level hypotheses can be derived from exploratory SHAP analysis even under conditions of limited model discriminative power.

The proposed approach constructs an XGBoost churn prediction model with class imbalance handling and hyperparameter optimization through Optuna, followed by global and local SHAP interpretation. Class imbalance — where the proportion of churning customers is substantially smaller than the majority class — is a structural characteristic of e-commerce transaction data that, if unaddressed, biases models toward the majority class and inflates the False Negative Rate at precisely the segment most critical for intervention (Chawla et al., 2002; Elreedy, Atiya and Kamalov, 2024). Two imbalance-handling strategies were empirically compared for this purpose — Synthetic Minority Over-sampling Technique (SMOTE) and XGBoost's built-in class-weighting parameter — with the better-performing approach retained for the final model (see Methods).

The primary contribution of this research is a decision support framework that is simultaneously transparent and empirically grounded: XGBoost answers who will churn, while SHAP explains why that particular customer will churn, at both aggregate and individual levels. Unlike prior studies that report model accuracy without contextual interpretation, this work

translates SHAP-derived findings into testable CRM hypotheses and explicitly acknowledges dataset limitations that constrain predictive performance — a degree of transparency rarely reported in the e-commerce churn literature.

2. METHODS

This study follows the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework as seen on Figure 1, progressing iteratively through six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment with interpretation. The full pipeline was implemented in Python 3.x using the xgboost, scikit-learn, optuna, imbalanced-learn, and shap libraries, with a fixed random seed (42) applied uniformly across the split, resampling, cross-validation, and model training; exact package versions were not archived separately and will accompany the code release upon publication.

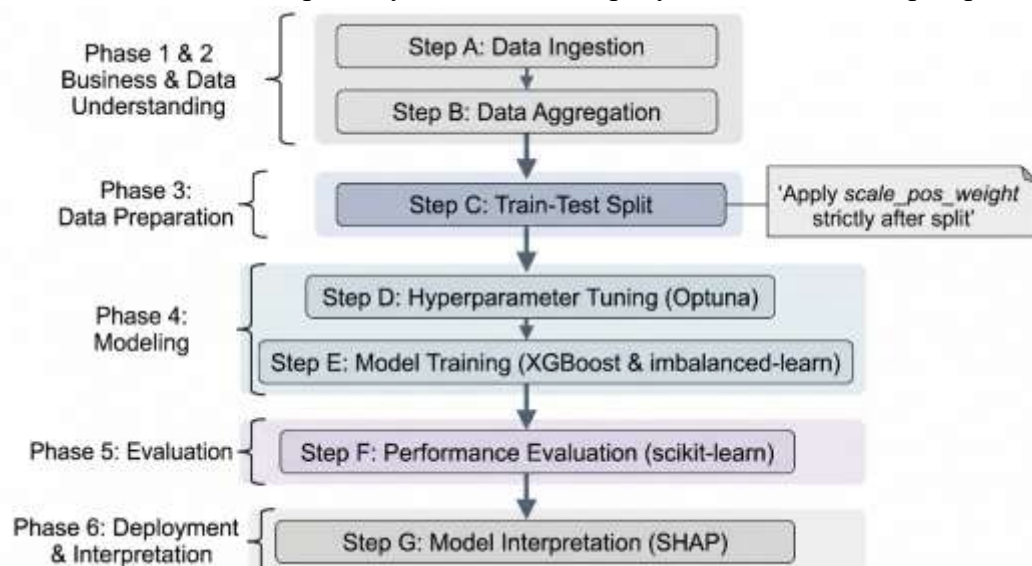


Figure 1. Research methodology

Business and Data Understanding

The study uses a publicly available secondary dataset, "E-Commerce Customer Behavior Analysis," sourced from the Kaggle platform (Jagtap, n.d.). The raw dataset contains 250,000 transaction rows with 13 variables, spanning customer demographics (age, gender), transaction behavior (purchase date, product category, price, quantity, total spending, payment method), and interaction records (returns as a proxy for dissatisfaction). After customer-level aggregation, the dataset yields 49,673 unique customer profiles. The target variable is a pre-assigned binary churn label (0 = non-churn, 1 = churn), with class distribution of approximately 80.05% non-churn and 19.95% churn. A critical limitation is acknowledged: the churn label carries no documented derivation methodology — no inactivity window, cancellation event, or

verifiable business rule. Given that all features exhibit extremely weak linear correlation with the label (highest $|r| < 0.01$), synthetic or randomized label generation cannot be excluded. This provenance issue directly constrains the achievable predictive performance of any model trained on this data.

Data Preparation

Missing values in the Returns column (~19%) were imputed with 0, consistent with the domain logic that an absent return record implies no product was returned. Categorical encoding applied One-Hot Encoding to the Gender variable. Row-level transactions were aggregated into customer profiles, generating Recency, Frequency, and Monetary (RFM) features. Five additional engineered features were constructed: Avg_Order_Value (total spending divided by transaction count), Return_Rate (returns divided by total transactions), Tenure (span between first and last transaction), N_Categories (count of distinct product categories purchased), and Avg_Days_Between_Trans (mean inter-purchase interval).

Modeling

Two imbalance-handling strategies were evaluated: SMOTE (50:50 and 60:40 ratios; standard formulation of Chawla et al., 2002, rather than adaptive variants such as Kosolwattana et al., 2023) applied on training data only, and scale_pos_weight as XGBoost's built-in class-weighting parameter. SMOTE was ultimately excluded from the final model because its combination with scale_pos_weight introduced synthetic noise that degraded generalization on the hold-out set. The final configuration relies solely on scale_pos_weight, with its value co-optimized during hyperparameter tuning.

Hyperparameters were tuned using Optuna (40 trials, default Tree-structured Parzen Estimator sampler) with 5-fold Stratified Cross-Validation (shuffled, seed = 42), maximizing mean ROC-AUC on the original (non-resampled) training distribution to avoid synthetic-data overfitting. The search space was: n_estimators [50, 300]; max_depth [2, 6]; learning_rate [0.01, 0.2] (log-uniform); subsample [0.6, 1.0]; colsample_bytree [0.5, 1.0]; reg_alpha [0.1, 20] (log-uniform); reg_lambda [1.0, 20] (log-uniform); min_child_weight [5, 30]; gamma [0, 5]; and scale_pos_weight [1.0, 6.0]. The resulting best configuration is: n_estimators = 259, max_depth = 2, learning_rate = 0.0130, subsample = 0.9591, colsample_bytree = 0.5435, reg_alpha = 12.3878, reg_lambda = 1.4707, min_child_weight = 7.

Evaluation

The dataset was split into training and hold-out test sets using stratified sampling to preserve class proportions. Model performance was assessed using Recall, Precision-Recall Area Under Curve (PR-AUC), and ROC-AUC — metrics appropriate for imbalanced classification where the minority class carries disproportionate business cost. The default

decision threshold of 0.5 was replaced by an F1-optimal threshold of 0.5733, identified from the Precision-Recall Curve computed on the hold-out test set itself — a methodological limitation rather than a validated procedure, since it introduces a degree of optimistic bias (evaluation leakage) into the reported Precision, Recall, and F1-Score at this threshold. The underlying ROC-AUC and PR-AUC, both threshold-independent, are unaffected and already establish near-random discrimination on their own. A three-way train/validation/test split for unbiased threshold calibration is identified as a priority for future replication of this pipeline.

Deployment and Interpretation

SHAP.TreeExplainer was used to compute SHAP (SHapley Additive exPlanations) values for all test-set predictions. Interpretation was conducted at two levels: (1) global analysis via summary and bar plots, ranking features by mean absolute SHAP contribution and visualizing the direction of influence across the dataset; and (2) local analysis via waterfall plots, decomposing predictions for individual customers by showing each feature's deviation from the base value. The two local case-study customers were selected deliberately rather than at random, to illustrate the model's most confident correct predictions in each direction: the churned-customer example is the true positive with the highest predicted churn probability, and the retained-customer example is the true negative with the lowest predicted churn probability. An explicit disclaimer applies: because model discrimination in this study approaches near-random levels, SHAP patterns may partly reflect statistical noise rather than genuine behavioral signals. All SHAP-derived findings are therefore treated as exploratory — intended for CRM hypothesis generation rather than as confirmed causal explanations of churn.

3. RESULTS AND DISCUSSION

Preprocessing and Descriptive Statistics

The preprocessing pipeline produced 49,673 unique customer profiles aggregated from 250,000 raw transactions. Missing values in the Returns column (47,596 rows; 19.03%) were imputed with 0 per domain logic. Five engineered features (Avg_Order_Value, Return_Rate, Tenure, N_Categories, Avg_Days_Between_Trans) were constructed alongside RFM variables. Pearson correlation analysis revealed that no engineered feature exceeded $|r| = 0.01$ against the churn label — a finding that foreshadowed the model's limited discriminative ceiling and is consistent with the suspected synthetic or randomized provenance of the churn label.

Model Performance

The final XGBoost model, configured with Optuna-tuned hyperparameters and `scale_pos_weight` balancing (without SMOTE), was evaluated on a stratified hold-out set of 9,935 customers (see Figure 2). At the F1-optimal threshold of 0.5733, performance on the minority churn class was: Precision = 0.2011, Recall = 0.9909, F1-Score = 0.3343, ROC-AUC = 0.5070, and PR-AUC = 0.2028.

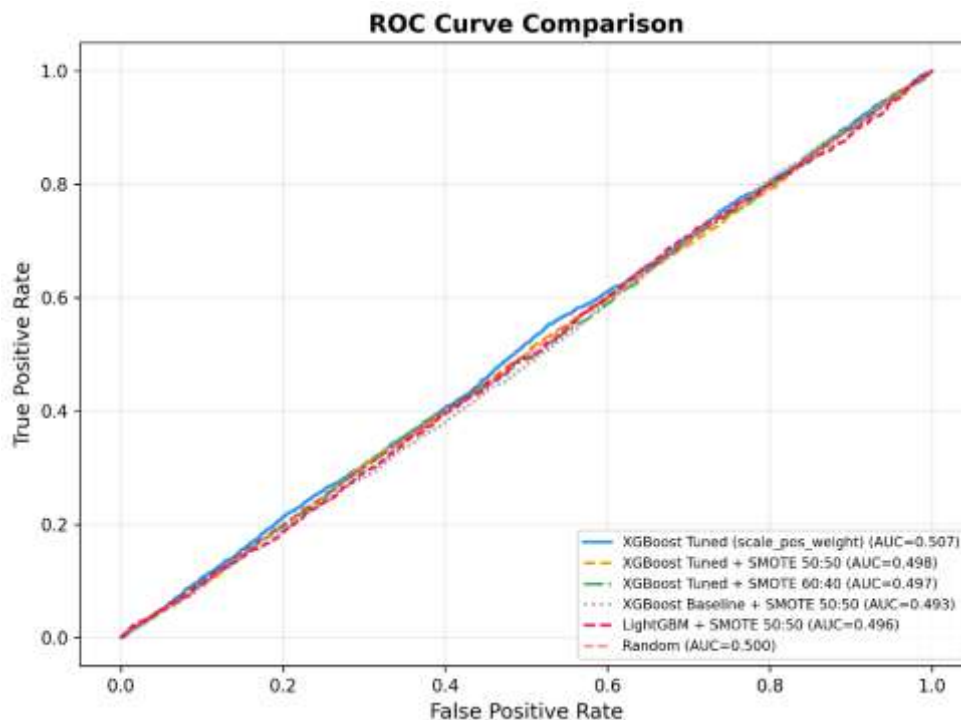


Figure 2. ROC curve comparison

The near-perfect Recall (0.9909) warrants explicit interpretation: it does not indicate strong churn detection. Rather, the model's learned probability distribution is nearly uniform across both classes, such that at the selected threshold, the model classifies almost all customers as churners. This produces 1,970 True Positives alongside 7,827 False Positives — an extreme precision cost. The ROC-AUC of 0.5070 in Table 1 confirms near-random discrimination, marginally above the 0.50 baseline, with all five evaluated model variants (including SMOTE variants) operating within the 0.49–0.51 range. The Precision-Recall Curve (Figure 3) similarly shows all variants tracking the random baseline at Precision ≈ 0.20 , equivalent to the class prior. Note that the single threshold of 0.5733, tuned for the primary (`scale_pos_weight`) model, was applied uniformly to all five variants in Table 1 for comparability; the zero Recall and Precision for the SMOTE-trained variants therefore reflects a threshold mismatch with each variant's own probability distribution rather than complete predictive failure — their ROC-AUC and PR-AUC, which are threshold-independent, remain in the same near-random range regardless. These results collectively confirm that the feature space constructed from the

available dataset does not provide sufficient signal to reliably distinguish churners from non-churners. As a further limitation, per-fold cross-validation variability and confidence intervals around these point estimates were not retained from the original experimental run and are not reported here; we flag this as a gap for future replications of this pipeline rather than substitute it with unverified figures.

Table 1. Model Comparison (Threshold = 0.5733)

Model	ROC-AUC	PR-AUC	Recall	Precision	F1-Score
XGBoost Tuned (scale_pos_weight)	0.5070	0.2028	0.9909	0.2011	0.3343
XGBoost Tuned + SMOTE 50:50	0.4976	0.1986	0.0000	0.0000	0.0000
XGBoost Tuned + SMOTE 60:40	0.4967	0.1995	0.0000	0.0000	0.0000
XGBoost Baseline + SMOTE 50:50	0.4933	0.1953	0.0000	0.0000	0.0000
LightGBM + SMOTE 50:50	0.4957	0.1990	0.0000	0.0000	0.0000

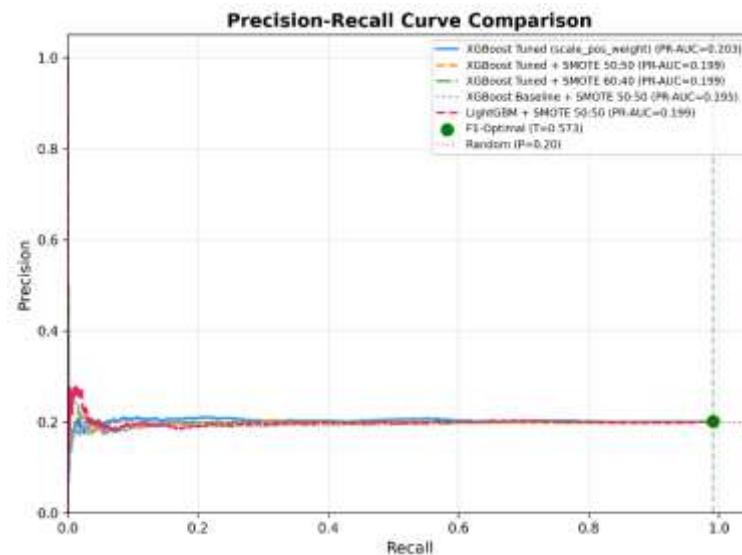


Figure 3. Precision-Recall curve comparison

Global SHAP Analysis

Despite near-random model discrimination, SHAP values were extracted to surface exploratory patterns in the model's internal decision logic as shown in Figure 4. Tenure ranked as the strongest predictor by mean absolute SHAP value: longer-tenured customers received

lower churn probability contributions, suggesting a weak stabilizing signal consistent with relationship duration theory in CRM literature (Gattermann-Itschert and Thonemann, 2022). Recency ranked second, with higher values (longer intervals since last transaction) pushing predictions toward churn — directionally aligned with the behavioral hypothesis that purchase inactivity precedes disengagement (Xiahou and Harada, 2022). Age showed non-trivial but non-directionally consistent contributions. Behavioral features including Avg_Days_Between_Trans, Avg_Order_Value, Return_Rate, Monetary, and Frequency contributed marginally; N_Categories and N_Payment_Methods registered near-zero influence.

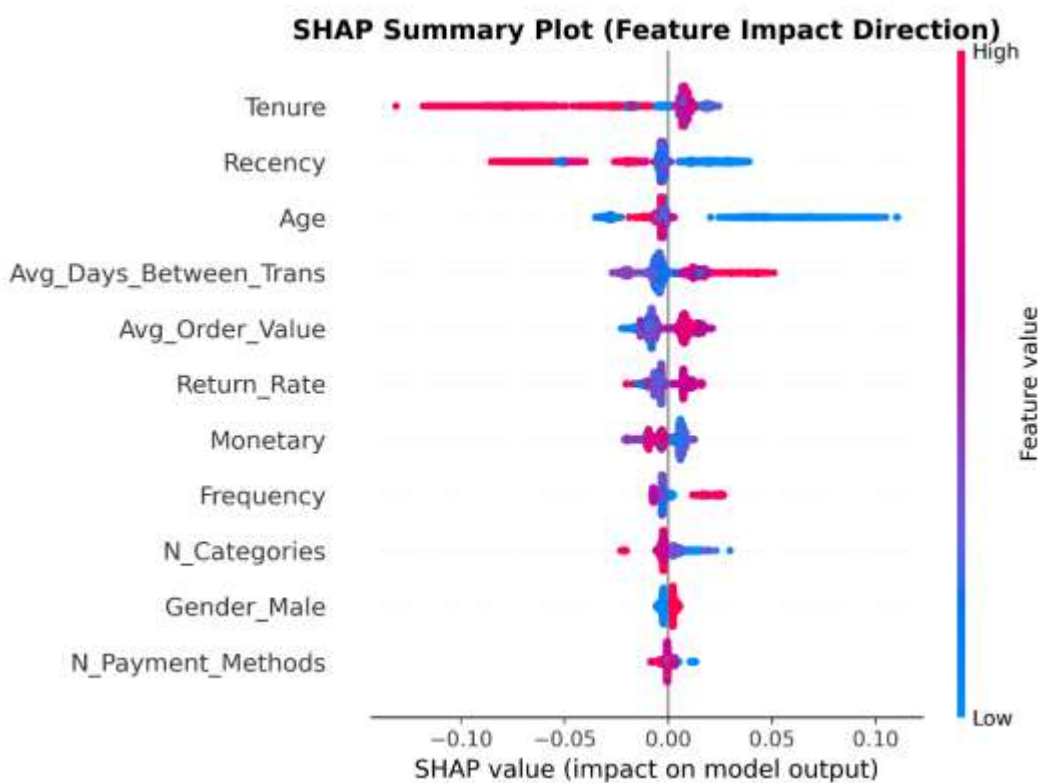


Figure 4. SHAP summary plot

Local SHAP Analysis

Local waterfall plots were generated for two representative customers (Figure 5 and 6 respectively). For a high-risk customer (ID: 38331), Age, Recency, and Tenure jointly elevated the prediction score toward the churn range. For a low-risk customer (ID: 20553), the same features in opposing directions collectively suppressed churn probability below threshold. These individual-level explanations are internally consistent with the model's logic, but must be interpreted as the model's current hypothesis about each customer — not as a validated behavioral diagnosis — given the low overall model confidence.

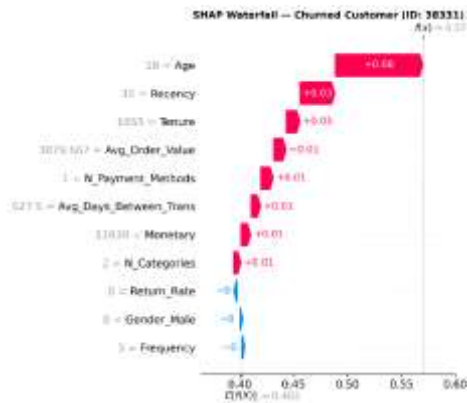


Figure 5. SHAP waterfall : churned customer

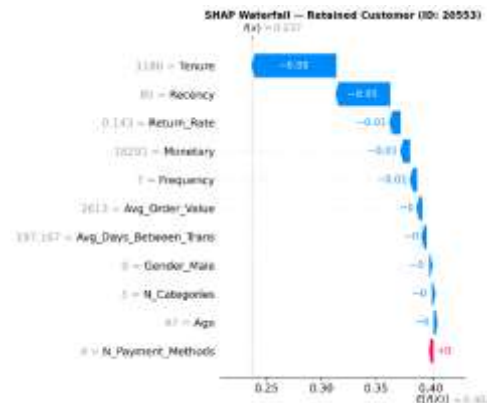


Figure 6. SHAP waterfall : retained customer

Discussion

The results address the three stated research questions. First, the XGBoost model's predictive effectiveness, measured through Recall, PR-AUC, and ROC-AUC, is severely constrained by the absence of genuine class separability in the dataset. The near-random ROC-AUC corroborates findings from studies on synthetic datasets, where label noise fundamentally limits ensemble model performance regardless of algorithmic sophistication (Elreedy, Atiya and Kamalov, 2024). Second, Tenure and Recency emerged as the dominant SHAP-identified features — a finding directionally consistent with RFM-based churn literature (Shobana et al., 2023), though their significance here reflects weak statistical patterns rather than confirmed behavioral mechanisms. Third, exploratory SHAP analysis supports the generation of testable CRM hypotheses (Table 2): long-tenured customers with declining activity as candidates for loyalty deepening programs; customers with Recency exceeding 30 days as targets for re-activation automation; high Return_Rate customers for proactive service recovery. These map directly onto established CRM retention frameworks (Pondel et al., 2021), but require A/B testing validation before operational deployment given the current model's near-random baseline.

Table 2. CRM Retention Strategy Framework Based on SHAP Findings

Dominant Feature (SHAP)	At-Risk Customer Profile	Recommended CRM Retention Strategy
-------------------------	--------------------------	------------------------------------

High Tenure	Long-tenured customers with declining activity	Loyalty Deepening: Exclusive access to new product lines and early-member benefits to reinforce emotional attachment.
High Recency	Customers who have not transacted for >30 days (1 shopping cycle median)	Re-activation Automation: Automated message triggers (Push Notification / WhatsApp) with personalized 'Welcome Back' coupon offers based on product category history.
High Age	Older customer segments with behavioral patterns flagged by SHAP	Segment-Specific Engagement: Tailored product assortments and communication channels aligned with generational preferences.
Low Monetary	Customers with total spending below the 3rd decile (low-spenders)	Tiered Loyalty Program: Double points or free shipping incentives with minimum spending thresholds to increase switching costs and encourage transaction value growth.
High Return_Rate	Customers with return rate >20% of total transactions	Proactive Service Recovery: Priority CS team escalation for product quality audit and compensation provision (discount, free shipping on next transaction).

The primary theoretical contribution is demonstrating that SHAP-based transparency does not compensate for insufficient predictive signal — a nuance underreported in prior XAI-churn literature that treats interpretability as independently valuable regardless of model quality. Practically, the study provides a replicable blueprint for SHAP-augmented CRM pipelines, with the critical caveat that high-quality, operationally derived churn labels are a prerequisite for actionable output.

4. CONCLUSION

This study examined XGBoost combined with SHAP interpretability for e-commerce customer churn prediction, with a central finding that departs from the accuracy-optimistic reporting common in prior literature: model discrimination remained essentially random, traceable not to algorithmic choice but to a churn label with no verifiable derivation methodology and features with negligible correlation to the target. This constitutes a first-order validity limitation — the dependent variable itself is suspect, rendering all downstream inferences, including SHAP rankings, contingent rather than conclusive.

Under this constraint, Tenure, Recency, and Age emerged as the dominant SHAP contributors, with directions consistent with established RFM-based churn theory. These patterns are theoretically coherent, but because the model approximates a random classifier, SHAP may be surfacing structured noise rather than genuine behavioral mechanisms; they are therefore reported as exploratory hypotheses for CRM hypothesis generation rather than confirmed causal drivers — a distinction most prior XAI-churn studies leave implicit. The corresponding segmented retention ideas (re-activation automation for high-Recency customers, loyalty deepening for high-Tenure customers, and proactive service recovery for high-Return_Rate customers) accordingly remain proposals for A/B validation rather than deployable strategies.

Four limitations should be weighed alongside these findings: an unverified, likely synthetic churn label; near-random discrimination that would incur substantial false-positive cost in any deployment; a feature space that omits behavioral signals such as session activity, login frequency, and review sentiment; and the absence of field validation for the proposed CRM ideas.

Future work should prioritize operationally derived churn labels, richer behavioral features, causal-inference approaches that move beyond associative SHAP patterns, a three-way train/validation/test split for unbiased threshold calibration, and randomized controlled trials to measure actual retention lift. More broadly, openly reporting near-random performance alongside exploratory SHAP analysis illustrates both the value and the limits of interpretability tools: they support hypothesis generation even in data-constrained settings, but cannot substitute for data quality — a point too often left implicit in applied XAI-churn literature.

5. REFERENCES

- Alkhanbouli, R. *et al.* (2025) “The role of explainable artificial intelligence in disease prediction: a systematic literature review and future research directions,” *BMC Medical Informatics and Decision Making*, 25(1), p. 110. Available at: <https://doi.org/10.1186/s12911-025-02944-6>.
- Bentéjac, C., Csörgő, A. and Martínez-Muñoz, G. (2021) “A comparative analysis of gradient boosting algorithms,” *Artificial Intelligence Review*, 54(3), pp. 1937–1967. Available at: <https://doi.org/10.1007/s10462-020-09896-5>.
- Chawla, N.V. *et al.* (2002) “SMOTE: Synthetic Minority Over-sampling Technique,” *Journal of Artificial Intelligence Research*, 16, pp. 321–357. Available at: <https://doi.org/10.1613/jair.953>.
- Elreedy, D., Atiya, A.F. and Kamalov, F. (2024) “A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning,” *Machine Learning*, 113(7), pp. 4903–4923. Available at: <https://doi.org/10.1007/s10994-022-06296-4>.

- Gattermann-Itschert, T. and Thonemann, U.W. (2022) "Proactive customer retention management in a non-contractual B2B setting based on churn prediction with random forests," *Industrial Marketing Management*, 107, pp. 134–147. Available at: <https://doi.org/10.1016/j.indmarman.2022.09.023>.
- Shobana, J. *et al.* (2023) "E-commerce customer churn prevention using machine learning-based business intelligence strategy," *Measurement: Sensors*, 27, p. 100728. Available at: <https://doi.org/10.1016/j.measen.2023.100728>.
- Jagtap, S. (n.d.) *E-Commerce Customer Behavior Analysis* [Data set]. Kaggle. [Full dataset URL to be confirmed by authors prior to submission.]
- Kosolwattana, T. *et al.* (2023) "A self-inspected adaptive SMOTE algorithm (SASMOTE) for highly imbalanced data classification in healthcare," *BioData Mining*, 16(1), p. 15. Available at: <https://doi.org/10.1186/s13040-023-00330-4>.
- Nagaraj, P. *et al.* (2023) "E-Commerce Customer Churn Prediction Scheme Based on Customer Behaviour Using Machine Learning," *2023 International Conference on Computer Communication and Informatics (ICCCI). 2023 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India: IEEE, pp. 1–6. Available at: <https://doi.org/10.1109/ICCCI56745.2023.10128498>.
- Pasaribu, J., Yudistira, N. and Mahmudy, W.F. (2024) "Tabular Data Classification and Regression : XGBoost or Deep Learning with Retrieval-Augmented Generation," *IEEE Access*, pp. 1–1. Available at: <https://doi.org/10.1109/ACCESS.2024.3518205>.
- Pondel, M. *et al.* (2021) "Deep Learning for Customer Churn Prediction in E-Commerce Decision Support," *Business Information Systems*, pp. 3–12. Available at: <https://doi.org/10.52825/bis.v1i.42>.
- Shwartz-Ziv, R. and Armon, A. (2022) "Tabular data: Deep learning is not all you need," *Information Fusion*, 81, pp. 84–90. Available at: <https://doi.org/10.1016/j.inffus.2021.11.011>.
- Vimbi, V., Shaffi, N. and Mahmud, M. (2024) "Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection," *Brain Informatics*, 11(1), p. 10. Available at: <https://doi.org/10.1186/s40708-024-00222-1>.
- Xiahou, X. and Harada, Y. (2022) "B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM," *Journal of Theoretical and Applied Electronic Commerce Research*, 17(2), pp. 458–475. Available at: <https://doi.org/10.3390/jtaer17020024>.
- Zhang, P., Jia, Y. and Shang, Y. (2022) "Research and application of XGBoost in imbalanced data," *International Journal of Distributed Sensor Networks*, 18(6), p. 155013292211069. Available at: <https://doi.org/10.1177/15501329221106935>.