Proceedings of CONF-CDS 2026 Symposium: Data-Centric AI Security: Securing Models, Learning Agents, and Autonomous Systems
DOI: 10.54254/2755-2721/2026.CH35873

An Interpretable Machine Learning Framework for E-Commerce Customer Churn Prediction Using SHAP

Yunhao Leng

School of Computer Science, Taylor's University, Subang Jaya, Malaysia
0367520@sd.taylors.edu.my

Abstract. In the digital economy, the cost of customer acquisition for e-commerce platforms has been rising steadily. Customer retention has become an essential force for achieving sustainable profitability. Although the machine learning models exhibit superior performance in predicting customer churn, the black-box nature of some models limits their interpretability in practical applications, which in turn hampers their performance in real-world scenarios. This paper proposes a hybrid framework that combines advanced ensemble learning algorithms with interpretable artificial intelligence (XAI) techniques to bridge the gap caused by this limitation in machine learning. Multiple classification models were evaluated in this paper, including Logistic Regression, Random Forest, and XGBoost, in a high-fidelity synthetic e-commerce dataset. In order to bridge the gap between the model and business utility, the Shapley Additive exPlanations (SHAP) method is adopted to interpret the output of the optimal model. The results show that the proposed system can not only achieve stable predictive performance but also reveal the key drivers of customer churn, such as "engagement score" and "days since last purchase". Hence, the system can provide actionable, data-driven business strategies for customer retention.

Keywords: Customer Churn Prediction, E-commerce, Explainable AI, Shapley Additive exPlanations (SHAP)

1. Introduction

In the digital economy, e-commerce platforms face increasing competition and rising customer acquisition costs, making customer retention a key driver for being a sustainable and profitable platform. Retaining existing customers is strategically important because online customer acquisition can be expensive, and prior work on e-loyalty shows that even a small improvement in retention can substantially improve profitability [1]. Accurately predicting customer churn and understanding its underlying causes have therefore become important for effective decision-making.

Machine learning models are widely deployed for churn prediction because of their ability to capture complex patterns. Although the models exhibit strong predictive performance, their practical adoption remains limited. The limitation arises from two aspects: a lack of interpretability, and the gap between the predictive result and the business applicability. These high-performance models often cannot provide clear, understandable insights, which makes it difficult for stakeholders to understand the causes of customer churn and design effective intervention strategies. Previous churn

DOI: 10.54254/2755-2721/2026.CH35873

research has shown that methodological choices affect predictive performance [2], while later work emphasizes that churn models should also be comprehensible and useful for managers [3].

To solve these challenges, this study proposed a system that uses ensemble learning models combined with explainable artificial intelligence(XAI) techniques. Logistic Regression, Random Forest, and XGBoost are evaluated on a high-fidelity synthetic e-commerce dataset, and the Shapley Additive exPlanations (SHAP) method, which is a game-theoretic explanation method that attributes each prediction to individual feature contributions [4], is applied to evaluate and interpret the optimal model. The goal of this approach is to bridge the gap between predictive performance and real-world business utility, to enhance the practical value of customer churn prediction models in real-world e-commerce applications.

Unlike the traditional approaches that focus solely on predictive performance, this study emphasizes the integration of interpretability and business applicability, providing a practical decision support framework for real-world e-commerce platforms.

2. Literature review

Customer retention is a key challenge in the e-commerce industry because acquiring new customers is often more expensive than retaining existing ones. Early-stage studies have primarily focused on traditional statistical methods, such as RFM (Recency, Frequency, Monetary) analysis, but these methods often struggle to capture complex, nonlinear customer patterns, which limits their effectiveness in practical applications. Logistic Regression always remains a common baseline for binary classification because it estimates the probability of a dichotomous outcome using a linear combination of predictors [5].

In recent years, machine learning algorithms have become a mainstream method for customer churn prediction. Ensemble learning methods, such as Random Forest and Gradient Boosting, have demonstrated powerful performance in a variety of applications. Breiman [6] shows that Random Forest can reduce variance and improve generalization by aggregating multiple weakly correlated trees. Chen and Guestrin [7] introduced XGBoost as a scalable tree boosting system with regularization and optimization techniques that improve predictive performance and computational efficiency.

These models can handle high-dimensional data and capture complex nonlinear relationships between customer attributes, which makes them more effective and powerful than traditional statistical methods.

Although the ensemble learning methods and deep learning models have high accuracy, their "black box" nature remains a major limitation in practical applications. The lack of interpretability makes it hard for businesses to trust and act on model predictions. Ribeiro et al. [8] argue that explanations are important when users must decide whether to trust a model prediction. SHAP provides a unified additive feature attribution framework based on Shapley values, assigning each feature a contribution value for a prediction [4]. In particular, for tree-based models, Tree SHAP provides an efficient and consistent way to compute feature attributions [9].

3. Methodology

3.1. Overall framework

This study proposes a hybrid and interpretable customer churn prediction framework that integrates machine learning models with Explainable Artificial Intelligence (XAI) techniques.

DOI: 10.54254/2755-2721/2026.CH35873

The framework consists of the following stages:
(1) Data preprocessing
(2) Model training (Logistic Regression, Random Forest, XGBoost)
(3) Model evaluation
(4) Model interpretation using SHAP

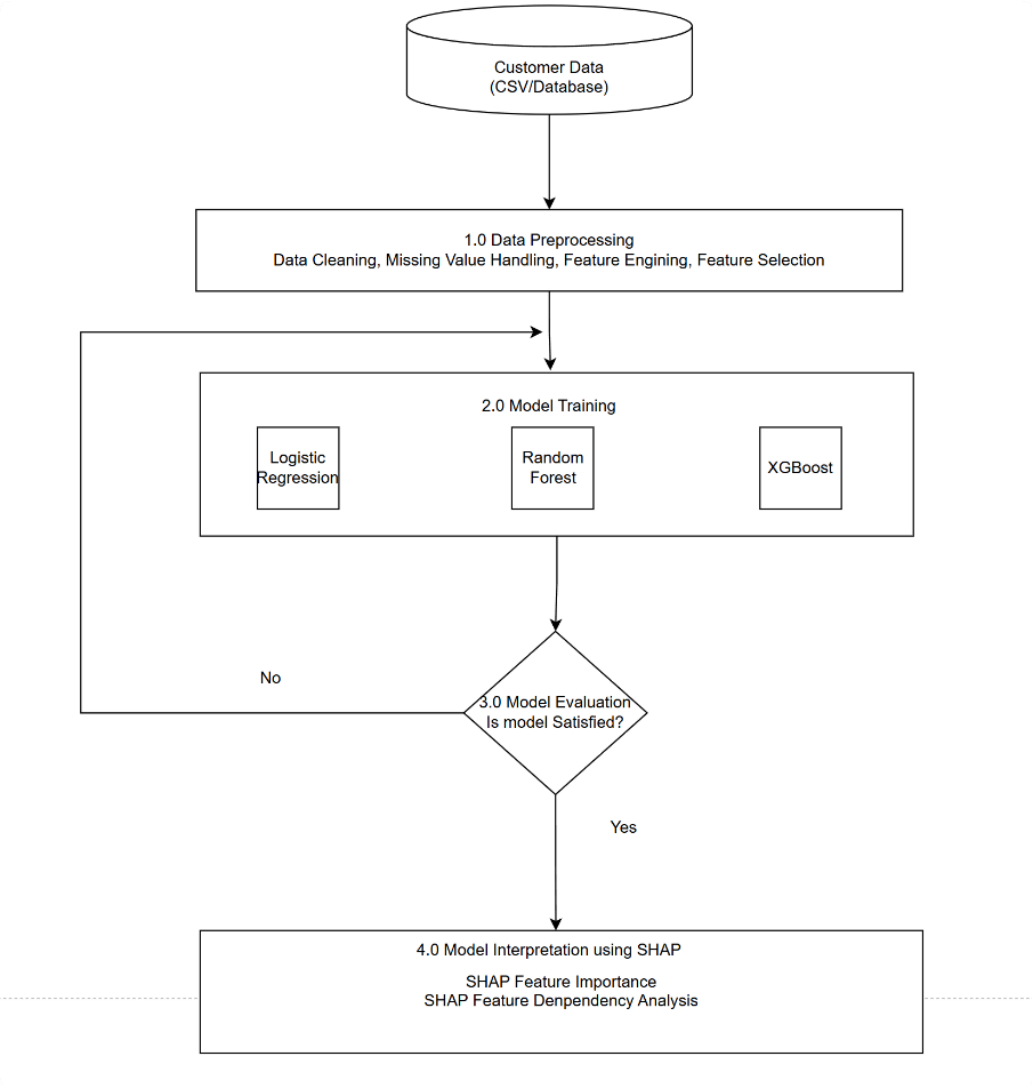

Figure 1. Working flow

This framework aims to achieve high predictive performance and interpretability, thereby bridging the gap between machine learning models and real-world business applications.

The recent studies have highlighted the importance of combining machine learning models with interpretability techniques such as SHAP to improve transparency and decision-making in customer churn prediction tasks.

3.2. Dataset description

The dataset used in this study is a high-fidelity synthetic e-commerce dataset that simulates real customer behavior.

DOI: 10.54254/2755-2721/2026.CH35873

Each record represents a customer and includes:

Behavioral features: Indicators that reflect users' daily operational habits on the platform, e.g., "browsing frequency" and "engagement score".

Transactional features: Quantitative metrics derived from users' consumption behaviors to measure purchasing capacity, e.g., "average order value", "total score"

Customer interaction features: Data generated from user-service communication reflecting service experience, e.g., "support tickets" and "satisfaction score".

The target variable is binary: 0 is Non-churn, 1 is Churn.

3.3. Data preprocessing

Data preprocessing is crucial for ensuring model stability, preventing data leakage and improving data quality. Based on the nature of the customer behavior dataset, this study designed and implemented the following preprocessing workflow:

3.3.1. Data integration and feature selection

First, the feature dataset and the target dataset are merged using an inner join based on the unique identifier (Customer_ID). Then, the Customer_ID column is removed from the feature space because it serves as a differentiating identifier and has no predictive power regarding customer churn.

3.3.2. Data encoding and missing value imputation

Categorical variables with binary outcomes (such as loyalty_member and the target variable churned) are mapped to numerical values, with "Yes" mapped to 1 and "No" mapped to 0, to fulfill the input requirements of machine learning algorithms. In order to handle the missing data without introducing bias from outliers, missing values in continuous features are imputed using the median.

3.3.3. Stratified data splitting

The dataset was divided into training and testing sets with an 80:20 ratio. To address the potential class imbalance problem common in churned datasets, this study applied stratified sampling based on the target variable (churned). This step ensures that the distribution of churn and non-churned customers remains consistent across the training and testing sets, thus providing a reliable foundation for evaluating the model's generalization performance.

3.3.4. Feature standardization

In order to eliminate the influence of different dimensional scales among the features, this study applied Z-score standardization. Each feature was scaled to a mean of 0 and standard deviation of 1 using the following formula:

$$z = (x - \mu)/\sigma \tag{1}$$

The important thing is, the scaler was fitted only to the training set (μ and σ are computed from the training data) and then applied to transform both the training set and the test set. This approach can prevent test set data from leaking into the model training phase.

DOI: 10.54254/2755-2721/2026.CH35873

3.4. Machine learning models

In this study, three classification models were implemented and compared for customer churn prediction. These models were selected to represent different learning paradigms, including linear models, bagging-based ensemble methods, and Boosting-based ensemble methods.

3.4.1. Logistic regression

Logistic Regression is a traditional statistical model widely used for classification tasks. In this study, Logistic Regression was employed as the baseline model due to its simple structure, high computational efficiency and good interpretability. It predicts the probability of a binary classification result by applying the Sigmoid function to a linear combination of input features.

$$P\left(y=1\middle|X\right)=\frac{1}{1+e^{-\left(w^{T}X+b\right)}} \tag{2}$$

where w represents the weight vector, and b is the bias term. Although the model is essentially linear, it provides significant feature interpretation capabilities as a baseline model.

3.4.2. Random forest

Random Forest is an ensemble learning method based on the Bagging technique. During the training process, it constructs multiple decision trees and performs classification prediction through majority voting. This method can effectively reduce model variance and overfitting by introducing randomness into sample selection (bootstrapping) and feature selection.

3.4.3. Extreme gradient boosting (XGBoost)

XGBoost is an efficient implementation of the gradient boosting algorithm. It continuously improves the overall predictive ability by progressively building decision trees, with each new tree used to fit the residuals of the previous model. This sequential learning mechanism enables it to capture complex nonlinear relationships.

The prediction of XGBoost can be represented as an additive model:

$$\hat{y}_i=\sum_{k=1}^{K}f_k\left(x_i\right),\quad f_k\in\mathscr{F} \tag{3}$$

where indicates the customer i 's predicted output, K is the number of decision trees, represents the k-th tree, and is the space of the regression tree. The objective function of XGBoost consists of two parts: training loss and regularization term:

$$Obj=\sum_{i=1}^{n}l\left(y_i,\hat{y}_i\right)+\sum_{k=1}^{K}\Omega\left(f_k\right) \tag{4}$$

where measures the difference between actual and predicted values, and the $\Omega\left(f_k\right)$ is to measure the complexity of the model. This regularization mechanism can help reduce overfitting and improve the model's generalization ability.

For implementation, XGBoost was trained using the same training dataset. After training, the model generates class labels and churn probabilities. The predicted class labels are used to calculate classification metrics such as accuracy and F1 score, which is the harmonic mean of precision and

DOI: 10.54254/2755-2721/2026.CH35873

recall and is useful when both false positives and false negatives matter [10], while the predicted probabilities are used to construct the ROC curve and calculate the AUC value. The ROC analysis evaluates classifier behavior across thresholds, and AUC summarizes discriminative ability [11].

4. Evaluation and results

4.1. Evaluation metrics

In order to evaluate the performance of the proposed customer churn prediction model, a variety of classification metrics were used, including accuracy, precision, recall, F1-score, confusion matrix and area under the receiver operating characteristic curve (AUC). These metrics comprehensively evaluate the model's performance from multiple perspectives.

Accuracy measures the proportion of correctly classified samples out of all test samples:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{5}$$

where TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives.

Precision measures the percentage of customers correctly predicted as churn out of all customers predicted as churn:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{6}$$

Recall measures the proportion of actual churned customers correctly identified by the model:

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{7}$$

The F1-score is the harmonic mean of precision and recall. It is useful when evaluating customer churn prediction because it balances the model's ability of correctly identify churned customers with its ability to minimize false positive.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}} = \frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}} \tag{8}$$

In addition, the ROC curves are used to assess the trade-off between true positives and false positives under different classification thresholds. The AUC value summarizes the model's discriminative ability, the higher the AUC, the stronger the model's ability to distinguish between churned and non-churned customers

4.2. Model performance comparison

The performance of Logistic Regression, Random Forest and XGBoost was compared using the same testing dataset. The F1-score shows that all three models achieved good predictive performance, with all scores above 0.88, but XGBoost still had the highest F1-score among the three models.

Table 1 shows the three models on the same testing dataset. All the models have strong performance, with F1-scores above 0.88 and AUC values above 0.99. XGBoost achieves the best F1-score, precision, recall, and accuracy under the selected classification threshold. Logistic

DOI: 10.54254/2755-2721/2026.CH35873

Regression has a slightly higher AUC, indicating that its probability ranking is also highly competitive even though its default-threshold F1-score is lower than that of XGBoost.

Table 1. Model performance comparison on the testing dataset

Model	Accuracy	Precision	Recall	F1-score	AUC
Logistic Regression	0.966	0.914	0.860	0.886	0.994
Random Forest	0.967	0.924	0.855	0.888	0.992
XGBoost	0.970	0.936	0.866	0.899	0.993

As shown in Figure 2, the three models perform similarly, but XGBoost records the highest F1-score at 0.899. This suggests that the boosting-based ensemble may be better able to balance precision and recall for the churn class. In churn prediction scenarios, this balance is important because marketing resources should be directed toward truly high-risk customers while also minimizing missed churn opportunities.

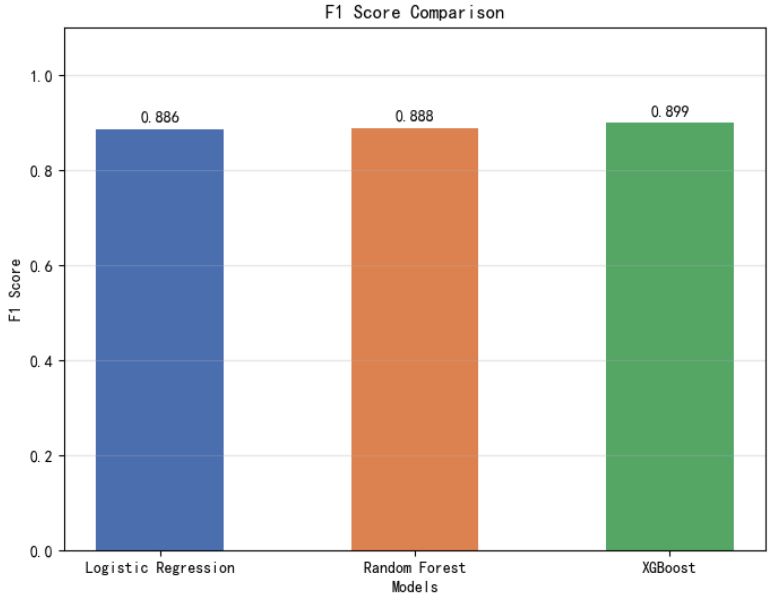

Figure 2. F1-score comparison across Logistic Regression, Random Forest, and XGBoost

4.3. Confusion matrix analysis

The confusion matrices present the details of classification performance (as shown in Figure 3). Logistic Regression correctly identifies 999 non-churn customers and 160 churn customers, with 15 false positives and 26 false negatives. Random Forest correctly identifies 1001 non-churn customers and 159 churn customers, with 13 false positives and 27 false negatives. XGBoost has the best performance in the confusion matrix, correctly identifying 1003 non-churn customers and 161 churn customers, while reducing false positives to 11 and false negatives to 25.

The XGBoost result is particularly useful for retention management. Compared with the other models, it mismatches fewer non-churn customers incorrectly and misses fewer actual churners. This means the platform can allocate retention incentives more efficiently while also reducing the risk of missing customers who are likely to leave.

DOI: 10.54254/2755-2721/2026.CH35873

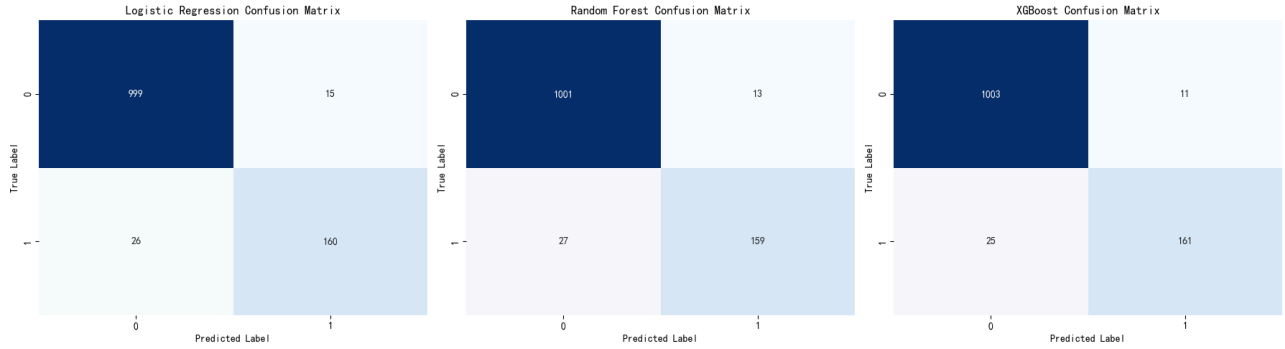

Figure 3. Confusion matrices for Logistic Regression, Random Forest, and XGBoost

4.4. ROC-AUC analysis

Figure 4 shows that all three ROC curves are close to the top-left corner, indicating strong discrimination between churned and non-churned customers. Logistic Regression obtains the highest AUC of 0.994, followed by XGBoost with 0.993 and Random Forest with 0.992. Because the differences are small, all three models have excellent ranking ability. However, XGBoost remains the preferred model in this study because it achieves the strongest threshold-based results, especially the highest F1-score and the lowest number of false negatives.

The small gap between AUC and F1-score rankings also demonstrates why multiple evaluation metrics should be reported. A model can rank customers very well across thresholds but still produce a less favorable classification result at a specific operating threshold. For business deployment, the threshold should therefore be selected based on retention budget, campaign capacity, and the relative costs of false positives and false negatives.

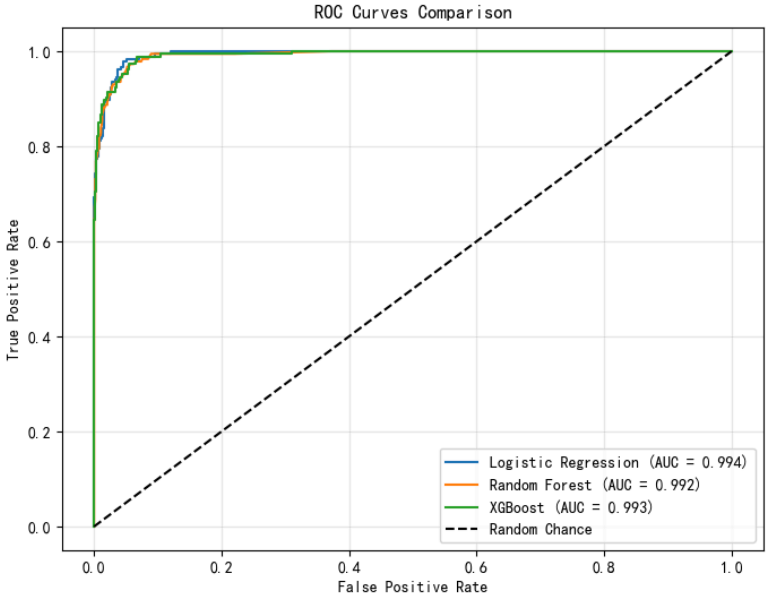

Figure 4. ROC curve comparison across the three models

DOI: 10.54254/2755-2721/2026.CH35873

4.5. SHAP explainability results

Figure 5 presents the SHAP summary plot for the XGBoost model. Features are ranked by their average absolute SHAP value, with those at the top exerting greater influence on model predictions. The most influential feature is engagement_score, followed by days_since_last_purchase, satisfaction_score, loyalty_member, return_rate, cart_abandonment_rate, account_age_months, customer_support_tickets, avg_order_value, and product_review_score_avg.

Loyalty membership appears to reduce churn risk, suggesting that loyalty programs may be associated with stronger retention. Higher return rates, cart abandonment rates, and more customer support tickets tend to increase churn risk, indicating that product dissatisfaction, purchase friction, and service problems are important warning signals. These findings demonstrate the value of SHAP explanations: they convert a black-box churn model into a prioritized list of actionable business factors.

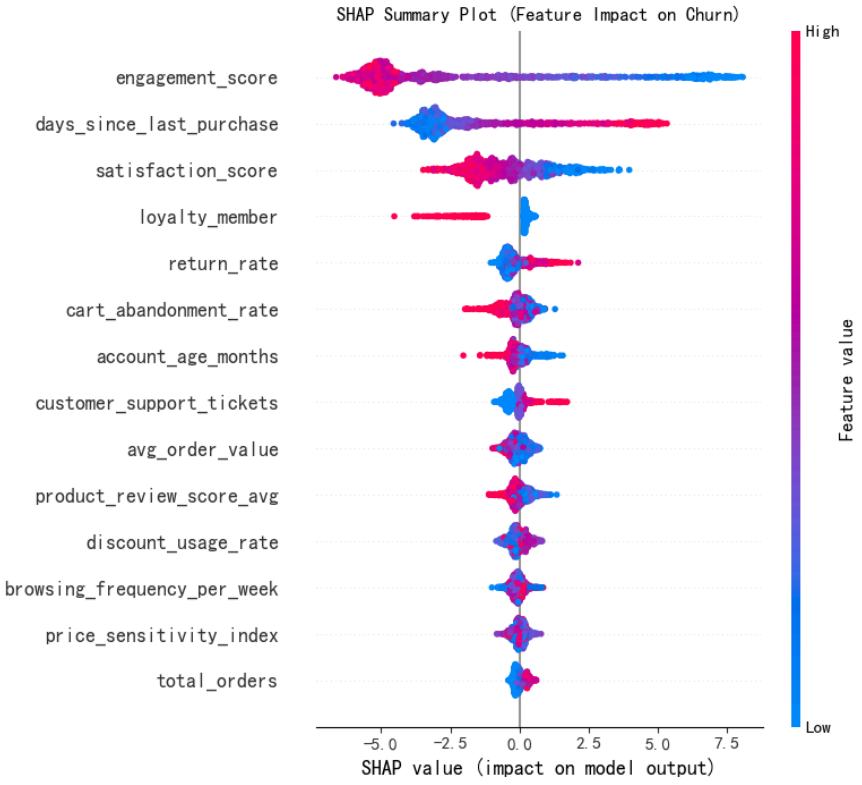

Figure 5. SHAP summary plot showing feature impact on churn prediction [4, 9]

5. Discussion

5.1. Predictive performance

The results show that the machine learning model can effectively distinguish between churned and non-churned customers. The high AUC indicates that the selected features contain strong signals related to customer churn. The XGBoost F1 score meets expectations, demonstrating that the gradient boosting algorithm can capture the nonlinear interactions between behavioral, transactional, and satisfaction-related variables.

DOI: 10.54254/2755-2721/2026.CH35873

The differences among the three models are not large. One possible reason is that the dataset has relatively clear churn patterns that can be captured by both linear and nonlinear methods. For practical deployment, the final model should not be chosen based on a small metric difference. Model complexity, interpretability, computational cost, ease of maintenance, and business requirements should also be considered.

5.2. Interpretability and business value

The SHAP analysis provides an explanatory perspective that is important for business. A churn score can tell managers which customers are at risk, and SHAP values can show why those customers are at risk. This is important because different churn drivers require different actions. Customers with low engagement may need personalized recommendations or reactivation campaigns, while customers with many support tickets may need service recovery actions.

The most important features identified by SHAP are also meaningful. Engagement score and days since last purchase reflect customer activity and recency. Satisfaction score and support tickets reflect customer experience. Return rate and cart abandonment rate reflect product fit and purchase friction. Loyalty membership reflects relationship strength. These patterns indicate that churn prediction should be integrated with customer relationship management rather than treated as a standalone scoring task.

5.3. Managerial implications

An interpretable churn model enables differentiated interventions tailored to specific churn drivers rather than applying uniform discounts. Specifically, for low-engagement customers, personalized content and product recommendations are deployed to drive reactivation; meanwhile, for those with long purchase gaps, retention programs and limited-time coupons serve as timely reminders. In addition, customers with low satisfaction or frequent support tickets are routed into a service recovery process, whereas high cart abandonment cases receive checkout optimization, free shipping, or simplified payment options. Furthermore, loyalty strategies differentiate actions even more precisely: non-members at high churn risk are offered membership registration discounts, while existing members receive tiered benefits to sustain engagement. Ultimately, this approach ensures each intervention addresses the root cause of churn, thereby maximizing retention efficiency.

6. Conclusion

This paper proposes an interpretable machine learning framework for e-commerce customer churn prediction. Logistic Regression, Random Forest, and XGBoost were trained and evaluated under a consistent preprocessing pipeline. The results show that all three models achieve strong performance, while XGBoost provides the best F1-score and confusion matrix results under the selected threshold. ROC analysis confirms that all models have excellent discriminative ability. SHAP interpretation further identifies engagement score, days since last purchase, satisfaction score, loyalty membership, and return rate as the key churn drivers.

In conclusion, churn prediction should combine accuracy with explainability. Accurate models help identify customers at risk, while explanations help managers understand the reasons behind churn and design targeted retention strategies. For e-commerce platforms, this combination can enhance the effectiveness of marketing campaigns, service recovery, and loyalty management.

DOI: 10.54254/2755-2721/2026.CH35873

This study has several limitations. First, the dataset is synthetic, so the findings may not represent the noise, seasonality, missingness, and behavioral diversity of real world. Second, the evaluation uses a single train-test split. Future work should use cross-validation and temporal holdout validation to better estimate generalization performance. Third, the classification threshold is not optimized according to business costs. In deployment scenarios, the optimal threshold should reflect retention budget, expected customer lifetime value, and the relative cost of contacting customers incorrectly versus missing actual churners.

Fourth, the current framework focuses on churn prediction rather than causal impact. A model may identify customers who are likely to churn, but it does not prove which intervention will prevent churn. Therefore, retention campaigns should be evaluated using controlled experiments or uplift modeling. Finally, privacy and ethical considerations should be addressed when using customer behavioral data in operational decision-making.

Future work can extend this study in several directions. First, the framework should be validated on real-world transactional and behavioral data. Second, temporal features and sequence-based models can be added to capture changes in customer behavior over time. Third, probability calibration and cost-sensitive threshold optimization can make the model more suitable for operational deployment. Fourth, uplift modeling and A/B testing can be used to evaluate which retention actions actually reduce churn. Finally, model monitoring should be implemented to detect data drift and maintain long-term performance after deployment.

References

[1] Reichheld, F. F., & Schefter, P. (2000). E-loyalty: Your secret weapon on the web. *Harvard Business Review, 78*(4), 105-113.

[2] Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research, 43*(2), 204-211. https://doi.org/10.1509/jmkr.43.2.204

[3] Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications, 38*(3), 2354-2364. https://doi.org/10.1016/j.eswa.2010.08.023

[4] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems, 30*, 4765-4774.

[5] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression (3rd ed.)*. Wiley. https://doi.org/10.1002/9781118548387

[6] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

[7] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). ACM. https://doi.org/10.1145/2939672.2939785

[8] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135-1144). ACM. https://doi.org/10.1145/2939672.2939778

[9] Lundberg, S. M., Erion, G. G., Chen, H., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*, 56-67. https://doi.org/10.1038/s42256-019-0138-9

[10] Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies, 2*(1), 37-63.

[11] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010