Scholar Publishing Group
Socio-Economic Statistics Research
https://doi.org/10.38007/SESR.2026.070202
ISSN 2790-2722 Vol. 7, Issue 2: 10-18

Application of Data Mining Technology in Predicting User Churn and Retention Strategies in E-Commerce

Chengfeng Jin

Carey Business School, Johns Hopkins University, Baltimore, 21218, Maryland, United States

Keywords: Data mining; User churn prediction; E-commerce platform; User retention; Explainable machine learning

Abstract: Amid rising customer acquisition costs and intensified platform competition, e-commerce operations are shifting their focus from acquiring new customers to retaining existing ones. This paper constructs a data mining-driven framework for user churn prediction and retention strategies, based on behavioral data such as browsing, adding to cart, placing orders, complaints, coupon usage, and repurchase intervals. Combining English-language research from the past three years with publicly available e-commerce churn data, the paper analyzes issues such as sample imbalance, label noise, insufficient model interpretability, and strategy cost constraints, proposing a closed-loop path of "data governance—feature engineering—model training—interpretable analysis—risk stratification—strategy feedback." The study argues that churn prediction can only be transformed from risk identification into effective retention when linked to user value, intervention costs, and response probability.

1. Introduction

E-commerce platforms are currently in a phase of refined operations. The fragmentation of traffic sources, homogenized competition, and decreasing user migration costs make it increasingly difficult to rely solely on new customer acquisition to increase profits. Unlike customer acquisition, user retention relies more on fulfillment experience, product matching, price incentives, and service recovery. Therefore, identifying risks and taking intervention measures before users stop purchasing has become a key application scenario for e-commerce data mining.

In the past three years, English literature has placed user churn prediction at the intersection of three fields: machine learning, deep learning, and customer relationship management. Imani et al. believe that learning models have become the main means of identifying potential churned users and formulating intervention measures [1]. Zaghloul et al. used a combination of RFM, clustering, and deep sequence models to verify its value in online retail [2], and Boozary et al. compared the customer retention prediction effect of integrated models [3].

Current applications still suffer from gaps in their implementation. User behavior is non-linear, and short-term silence does not necessarily mean churn. Due to the lack of sample limitations, models are prone to bias. High-precision models are not interpreted by the business context, and retention actions are often simplified to a uniform coupon, ignoring the differentiated value of users

and the underlying reasons, while also taking into account the cost of intervention.

2. Current Status Analysis of the Research Topic

2.1 Data foundation for e- commerce user churn prediction

E-commerce data mainly includes transaction data, access data, interaction data, marketing data, and customer service data. Transaction data reflects the number of orders, amount, and repurchase cycle; access data reflects dwell time, device, and path; interaction data reflects favorites, adding to cart, reviews, and returns/exchanges; marketing data reflects coupon redemption and usage; and service data reflects complaints and satisfaction. Together, they form the feature space for churn prediction.

The publicly available e-commerce churn dataset contains 5,630 customer records and 20 features, with 948 churned users and a churn rate of 16.84% [4]. This structure illustrates that sample imbalance is a fundamental problem in e-commerce churn prediction. If the model only pursues accuracy, it will obtain high scores by predicting a large number of "retentions" while ignoring those users who really need to be retained.

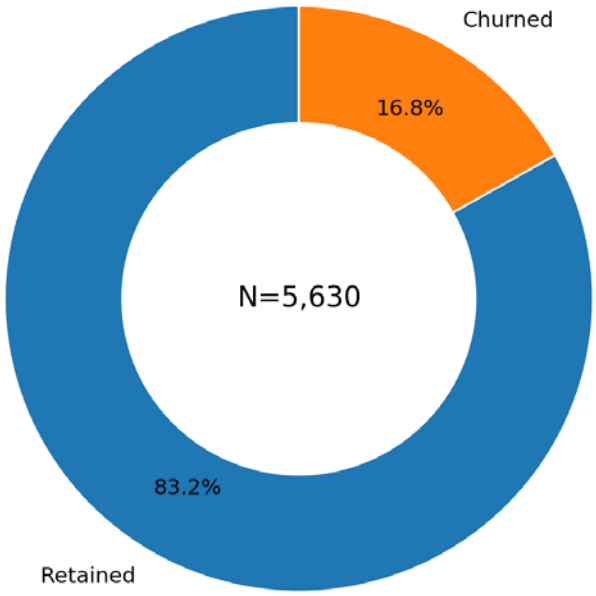

Figure 1. Statistics on the percentage of users churned and retained.

Figure 1 is a pie chart created using publicly available e-commerce churn datasets to illustrate the percentage of users in each state. Retained users far outnumber churned users, therefore class imbalance must be addressed before model training. The chart indicates that platforms should use recall, AUC, and net retention gains as metrics to measure model value, rather than solely relying on overall accuracy to judge model performance.

2.2 Evolution of Data Mining Methods

E-commerce churn prediction has evolved from the initial process of rule identification, statistical modeling, machine learning, and deep learning to the current methods. Rule-based methods are intuitive but have limited coverage; Logistic Regression and Survival Analysis can

provide probabilities but cannot handle complex nonlinearities; tree models and ensemble learning improve the fit of tabular data; sequence models are suitable for characterizing repurchase rhythms and browsing paths.

In the machine learning stage, Decision Tree, Random Forest, XGBoost, LightGBM and SVM are widely used. Huda et al. used RFM and classification models to identify e-commerce grouping and churn. The results showed that decision trees can achieve an accuracy of 87% in some cases, and purchase behavior, conversation, claim and discount behavior in decision trees make a significant contribution to the discrimination[5].

In the deep learning phase, LSTM, GRU and attention mechanisms are used to process continuous behavior logs. Zaghloul et al. validated the performance of deep sequence models on Olist, Instacart and Online Retail datasets, where LSTM achieved 99.7% and 99.9% on Olist and Instacart respectively, and GRU achieved 99.5% on Online Retail[2].

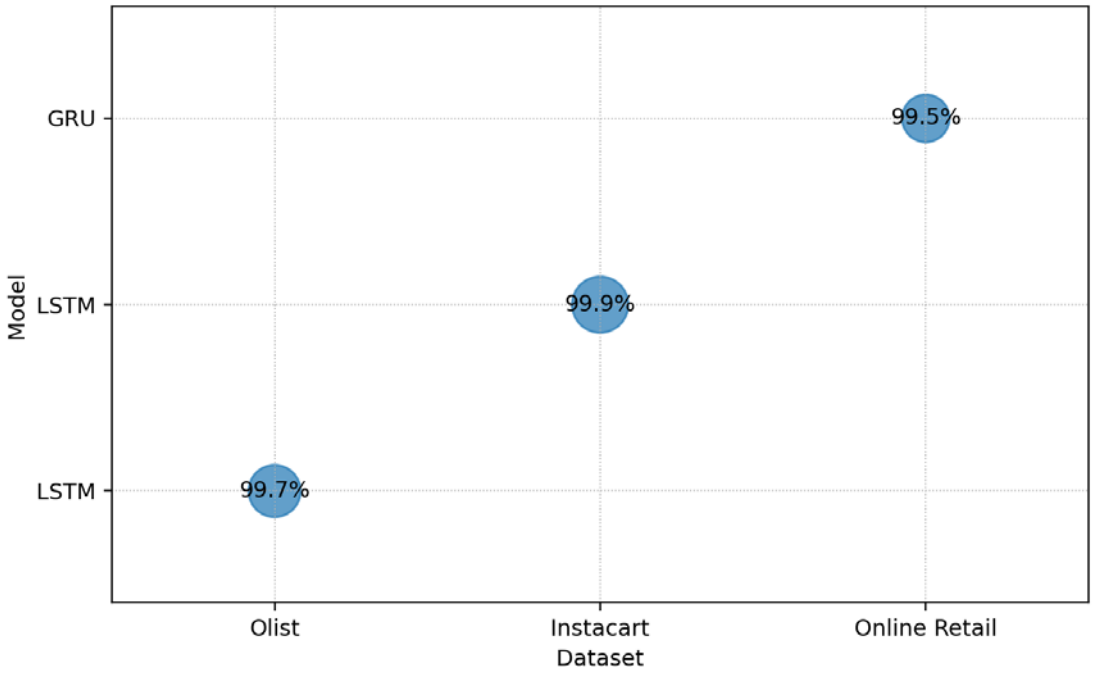

Figure 2. Comparison of prediction accuracy of sequence models on different datasets

Figure 2 uses a bubble chart to represent the accuracy results for different datasets and sequence models, preventing the discrete experimental comparisons from being obscured by a line chart. Bubble size and labeled values together reflect model performance. The results show that deep sequence models have a strong fitting ability for continuous transaction data; however, interpretability, cost, and business feasibility need to be considered before deployment.

2.3 Current Status of the Integration Between Predictive Models and Retention Strategies

Recent studies have placed greater emphasis on turning prediction results into retention actions. Jahan and Sanam combined CatBoost, K-means, and recommendation modules to form an e-commerce customer retention framework [6]; Shobana et al. integrated SVM prediction with a hybrid recommendation strategy to predict potential churned users and provide them with discounts and service configurations [7]; Ardhani and Tania used the SHAP interpretation perspective to explain that the model results should be able to reflect the causes of risk [8]; Boukrouh and Azmani

emphasized that interpretable models are conducive to turning risk identification into business decisions [9].

Table 1. Data Variable System for E-commerce User Churn Prediction

Data domain	Core variables	Mining purpose	Retention implication
Transaction	OrderCount, OrderValue, PaymentMode	Measure purchase intensity	Identify value decline
Behavior	HourSpendOnApp, LoginDevice, Session	Detect engagement pattern	Optimize channel experience
Relationship	Tenure, DaySinceLastOrder	Capture lifecycle stage	Trigger early warning care
Service	Complain, SatisfactionScore	Capture negative experiences	Prioritize service recovery
Promotion	Coupon Used, Cashback Amount	Measure incentive sensitivity	Allocate differentiated offers
Logistics	WarehouseToHome, DeliveryDelay	Link fulfillment to churn	Improve delivery reliability

Table 1 presents the e-commerce churn prediction data variable system. The five dimensions—transaction, behavior, relationship, service, and marketing—correspond to value decline, decreased activity, lifecycle changes, damaged experience, and increased sensitivity to incentives, respectively. The variables in the table can serve as both characteristic variables of the model and explanatory variables for operations, helping to determine which 环节 (stage/stage) should be addressed for improvement.

3. Problem Statement: Key Bottlenecks in E-commerce Churn Prediction

3.1 Inconsistent definitions of churn lead to tag noise

Non-contract e-commerce platforms do not have specific cancellation policies; churn tags are generally based on the number of days without placing an order. However, 30 days without an order is a risk for daily necessities users, but normal for home appliance users. Fixed thresholds can lead to noise in tagging; therefore, dynamic windows should be set based on three criteria: category cycle, historical frequency, and user value.

To standardize the monitoring metrics, we can first perform periodic churn rate statistics. Let Nchurn be the number of users churned during the observation period, and Ntotal be the number of active users at the beginning of the period. Then the churn rate is:

$$Churn\ Rate = \frac{N_{churn}}{N_{total}} \times 100\% \quad (1)$$

In equation (1), Nchurn represents the number of users identified as churned during the observation period, and Ntotal represents the total number of active users initially observable. This indicator is suitable for macro-level monitoring, but before modeling, factors such as product cycle and individual repurchase habits need to be added to the tags.

3.2 Imbalanced sample size reduces the ability to identify high-risk individuals

There are generally fewer churned samples than retained samples. If the imbalance is not addressed, the model will learn the bias of "the majority of users retaining". The data layer can use

SMOTE, undersampling, or combined sampling; the algorithm layer sets class weights and cost-sensitive loss; and the evaluation layer focuses on recall, F1-score, AUC, and lift, rather than accuracy alone.

The user churn probability in a binary classification model is represented by the Logistic function. The user feature vector is denoted by xi, and the parameter vector by β; therefore, we have:

$$P(y_i=1 \mid x_i)=\frac{1}{1+e^{-(\beta_0+\beta^T x_i)}} \tag{2}$$

In equation (2), yi=1 indicates that a user has churned, xi is the feature vector and intercept term of various aspects such as transactions, behaviors, services, and marketing, and β is the weight coefficient of the features. This model has good interpretability and can be used as a baseline model, but it is poor at characterizing high-dimensional nonlinear behavioral relationships.

3.3 Insufficient model explanation hinders business implementation

Operations personnel not only need to give risk scores, but also need to know the sources of risk. If the model only outputs a churn probability of 0.82, then it is impossible to determine what kind of discount should be given, what kind of after-sales service should be repaired, or what new products should be recommended. Explainable machine learning transforms variable contributions into business leads. Ardhani and Tania found that metrics such as relationship duration, complaints, and cashback have a significant effect on churn determination [8]; Boukrouh and Azmani believe that e-commerce forecasting should consider both the ability to identify and interpret [9].

The interpretable model can decompose the prediction result for a single user into feature contributions. Let the model output be f(x), and the contribution of the j-th feature be φj. Then the model's loss function can be written as...

$$f(x)=\phi_0+\sum_{j=1}^{m}\phi_j \tag{3}$$

In the formula, phi₀ represents the average risk of the entire sample population, the number of features is represented by m, and phi_j indicates whether the j-th feature has a positive or negative impact on the current user's risk. phi_j > 0 indicates that the feature has a positive impact on the current user's risk, and vice versa. This formula can explain the impact of factors such as complaints, silence time, and unused coupons on risk.

3.4 Retention strategies lack cost-benefit constraints

High-risk users should not be offered substantial discounts. Excessive subsidies can lead to decreased profits and may not be effective for users dissatisfied with the experience; frequent push notifications can also cause user churn. Therefore, retention strategies must consider four factors: risk probability, user value, intervention costs, and response probability.

Let pi be the probability of user churn, CLVi be the user lifetime value, Ci be the intervention cost, and ri be the strategy response probability. Then, the expected retention benefit is:

$$E(G_i)=p_i\times r_i\times CLV_i-C_i \tag{4}$$

In equation (4), Ei represents the expected return for user retention intervention. Intervention is only economically justifiable when this value is greater than 0. This formula changes the platform's focus from "who has the greatest risk" to "who is most worth saving."

4. Problem Solving and Retention Strategy Design

4.1 Constructing a technical process of "data governance - feature engineering - model training - policy delivery"

The first step in the technical process is data governance. The platform needs to manage user IDs, order IDs, device IDs, and marketing IDs uniformly, and distinguish between naturally missing and abnormally missing values. An unused coupon might simply be due to not receiving it, and missing delivery distance might be due to an incomplete address. Outlier handling must also be guided by business implications; high-priced or long-term customers should not be arbitrarily excluded.

Feature engineering should start with the original variables and then construct metrics such as repurchase interval, number of views in the past 30 days, change rate of amount in the past 90 days, coupon usage rate, repurchase interval after complaints, number of times added to cart but not purchased, and category concentration. RFM remains the benchmark for judging user value and can be combined with clustering, classification, and sequence modeling.

Model training should employ stratified sampling, cross-validation, and calibration. For scenarios with high interpretability requirements, Logistic Regression and Decision Tree can be used; for scenarios with high accuracy requirements, Random Forest, XGBoost, LightGBM, CatBoost, or deep sequence models can be used. Model selection should comprehensively consider recall, AUC, deployment cost, and operational readability.

Table 2 Applicability of different data mining models in e-commerce churn prediction

Model	Main advantage	Main limitation	Suitable scenario	Business readability
Logistic Regression	Transparent probability	Weak nonlinearity	Baseline scoring	High
Decision Tree	Rule extraction	Unstable split	Small feature set	High
Random Forest	Robust performance	Weak local explanation	Mixed tabular data	Medium
XGBoost	Strong discrimination	Parameter complexity	High-dimensional features	Medium
CatBoost	Categorical handling	Training cost	Many categorical variables	Medium
LSTM/GRU	Sequence learning	Data demand	Behavior logs	Low-medium

Table 2 shows the advantages, disadvantages, and application areas of common models. Logistic regression and decision trees are easy to interpret and suitable as baselines; Random Forest, XGBoost, and CatBoost are suitable for complex tabular data; LSTM and GRU are suitable for continuous behavioral logs. In practice, the method used depends on the amount of data, the content to be explained, and the deployment cost.

4.2 Establish a user risk stratification and strategy matching mechanism

After the model outputs the churn probability, risk stratification should be performed. Low-risk users should receive regular recommendations and membership growth maintenance. Medium-risk users should be maintained with repurchase reminders and lightweight benefits. High-risk users should be provided with discounts, service compensation, or category recommendations based on the reason. Extremely high-risk users with high value should be subject to manual follow-up and

exclusive benefits.

Risk probability should be considered in conjunction with user value. High-risk, high-value users should be prioritized for re-engagement, while high-risk, low-value users require low-cost, automated outreach. Low-risk, high-value users require controlled intrusion and enhanced member care, and low-risk, low-value users can only maintain activity through basic recommendations.

Risk stratification can be accomplished using a comprehensive scoring function. Ri is the comprehensive churn risk, pi is the model probability, Si is the negative score of the service, Di is the standardized metric for recent non-purchase, and Vi is the value protection coefficient. Therefore, Ri = pi * Si * Di * Vi

$$Risk_i=w_1p_i+w_2S_i+w_3R_i-w_4V_i \tag{5}$$

In equation (5), w1, w2, w3, and w4 are weight parameters, Si is the severity of the service problem, Di is the degree of abnormal silence time, and Vi is the user's past value. The negative sign indicates that high-value users should be given priority protection. This function integrates model risk, service experience, purchase interval, and user value into the strategy decision-making process.

Table 3 User Churn Risk Strategies and Retention Strategy Configuration

Segment	Risk range	Main churn driver	Retention action	Evaluation metric
Low risk	0.00-0.25	Stable purchase	Membership reminder	Repurchase rate
Medium risk	0.25-0.50	Activity decline	Personalized recommendation	Click-to-order rate
High risk	0.50-0.75	Price or experience issue	Targeted coupon and service recovery	Conversion uplift
Critical risk	0.75-1.00	Complaint and long silence	Manual care and exclusive benefits	Net retained value

Table 3 translates predicted probabilities into retention behaviors. Different risk levels cannot use the same intervention methods. Automated operations are more suitable for medium-risk users, while high-risk users require more precise restoration of discounts and services. Extremely high-risk users should focus more on the value of net retention. This table illustrates that the model must enter a strategy closed loop.

4.3 Introducing interpretable analytics to improve operational availability

Interpretable analytics should consider both service platform strategies and individual interventions. Global interpretations identify the main reasons for churn, such as increased complaint rates, worsened delivery times, and the disappearance of discounts, while individual interpretations describe the risks associated with a particular user, supporting the creation of triggering messages.

The operations interface should translate the explanation results into business language. For example, "Complain contributes +0.18" can be expressed as "Recent complaints have significantly increased the risk of churn, and after-sales follow-up is recommended"; "CouponUsed contributes -0.07" can be expressed as "Users are still responding to the offer and can use a moderate discount".

4.4 Validate the effectiveness of retention strategies through experiments.

Retention strategies need to be validated through online trials. The platform will conduct A/B tests on high-risk users, with the experimental group using model-based recommendations and the control group using traditional recommendations. Metrics should include repurchase rate, cost of discounts, changes in complaints, net revenue increase, and long-term retention rate; short-term conversion should not be prioritized at the expense of profit.

The intervention results are then used to rewrite the user profile. If a user doesn't respond to the offer but does respond to a customer service follow-up, the system will lower the weight of the price sensitivity tag and increase the priority of service restoration. De Alwis et al. argue that explanation, risk modeling, and segmentation should be used together to support personalized retention strategies, suggesting that churn prediction should shift from a one-time scoring method to continuous feedback.

Conclusion

The value of e-commerce user churn prediction lies not simply in improving classification accuracy, but in identifying changes in various data behaviors, explaining the causes of risks, and formulating actionable retention strategies. Transaction, behavioral, service, and marketing data together form the foundation of prediction, while RFM, clustering, ensemble learning, sequence models, and interpretable machine learning can respectively improve value identification, risk assessment, and strategy explanation capabilities.

In practice, this requires a closed loop encompassing seven stages: data governance, feature engineering, model training, interpretation and analysis, risk stratification, policy delivery, and effect feedback. Future development could incorporate causal inference, reinforcement learning, and federated learning to enhance privacy protection and improve the efficiency of intervention allocation. Finally, long-term experiments would determine the impact of each retention action on user lifetime value.

References

[1] *Imani M, Joudaki M, Beikmohammadi A, Arabnia H R. Customer Churn Prediction: A Systematic Review of Recent Advances, Trends, and Challenges in Machine Learning and Deep Learning[J]. Machine Learning and Knowledge Extraction, 2025, 7(3):105.*

[2] *Zaghloul M, Barakat S, Rezk A. Enhancing customer retention in Online Retail through churn prediction: A hybrid RFM, K-means, and deep neural network approach[J]. Expert Systems with Applications, 2025, 290:128465.*

[3] *Boozary P, et al. Enhancing customer retention with machine learning: A comparative analysis of ensemble models for accurate churn prediction[J]. Machine Learning with Applications, 2025.*

[4] *Verma A. Ecommerce Customer Churn Analysis and Prediction[DB/OL]. Kaggle, 2024.*

[5] *Wu, Y. (2026). Research on Interpretable Confidence Rule Base Modeling Method Integrating Data-Driven and Constrained K-Means Optimization. Procedia Computer Science, 279, 612-619.*

[6] *Sun, J. (2026). Automated Feature Engineering and Screening System for Large-Scale Factor Libraries. Procedia Computer Science, 281, 1282-1290.*

[7] *Hou, Y. (2026). Collaborative Regulation for Stable Data Center Operation under Energy Efficiency Constraints. Procedia Computer Science, 281, 612-620.*

[8] Zhang, Z. (2026). Research on Performance Optimization Methods for Resource-Aware Model Services in AI Systems. Procedia Computer Science, 281, 1310-1317.

[9] Zhang, Z. (2026). Research on Performance Monitoring and Data-driven Optimization Mechanisms for AI Systems Oriented Toward Decision Support. Advances in Computer and Communication, 7(2).

[10] Wu, Y. (2026). Research on Interpretable Confidence Rule Base Modeling Method Integrating Data-Driven and Constrained K-Means Optimization. Procedia Computer Science, 279, 612-619.

[11] Liang, Q. (2026). Research on Low-Energy Space Organization Methods in the Context of Building System Integration. Global Journal of Science & Innovation, 3(1), 9-15.

[12] Gao, Y. (2026). Governance Structures and Checks and Balances in Private Equity Funds: Practice, Problems, and Improvement Paths. Financial Economics Insights, 3(2), 82-91.

[13] Liu, Z. (2026). Research on Measuring User Behavior Response Differences Supported by Propensity Scoring Method. European Journal of AI, Computing & Informatics, 2(2), 47-53.

[14] Chang, C. W. (2026). Privacy Infrastructure Vulnerability Mining and Automated Framework Based on Multimodal AI. Procedia Computer Science, 279, 702-711.

[15] Wu, L. (2026). Construction and Evolutionary Analysis of a Game Model for Supply Chain Finance Funding Based on Blockchain Technology. Procedia Computer Science, 282, 2004-2012.

[16] Ding, J. (2025, December). Research on Logistics Cost Control and Optimization in Automotive Manufacturing Supply Chain Based on DMAIC Model. In 2025 IEEE 1st International Conference on Recent Trends in Computing and Smart Mobility (RCSM) (pp. 1-7). IEEE.