Proceedings of the 5th International Conference on Computing Innovation and Applied Physics
DOI: 10.54254/2753-8818/2026.32404

Classification and Prediction of User Churn Risk in Live-Streaming E-Commerce Based on Machine Learning Algorithms

Xiaoxue Fan

Shandong Vocational College of Light Industry, Zibo, China
fxx17@163.com.

Abstract. The popularization of encryption protocols such as HTTPS and Tor not only significantly enhances the privacy protection capability of network traffic, but also provides a hidden space for risks such as the spread of malicious websites and the leakage of sensitive information. The transmission characteristics of the current diverse network environments such as home WiFi, public WiFi, 4G and 5G are significantly different, which makes encrypted traffic exhibit cross-environment variation features such as delay fluctuations, retransmission rate changes and uneven data packet distribution, posing a severe challenge to the generalization and accuracy of the fine classification of encrypted traffic in cross-network environments. To solve this difficult problem, this paper proposes an LSTM algorithm optimized based on the multi-head attention mechanism and introduces a large model to assist in summary analysis and result output. The research first conducted correlation analysis and violin graph analysis, and then carried out comparative experiments through multiple machine learning algorithms. After verification by the test set, the final results were presented by the large model. Experiments show that the proposed algorithm demonstrates absolute advantages in all evaluation indicators. The accuracy rate, recall rate, precision rate, F1 value all reach 99%, and the AUC value is as high as 100%, comprehensively surpassing all comparison algorithms. The application of large models has effectively enhanced the efficiency of analysis and the credibility of results. This research provides reliable technical support for the precise and detailed classification of encrypted traffic in cross-network environments and has significant practical significance for network security protection and risk control.

Keywords: Live-streaming e-commerce, user churn, KPCA, ISSA

1. Introduction

Live-streaming e-commerce has witnessed explosive growth in recent years, with its user base continuously expanding. However, industry competition has become increasingly fierce, and the cost of customer acquisition for platforms has been constantly rising. The problem of user churn has gradually become a key bottleneck restricting the development of the industry [1]. According to industry data, the average monthly churn rate of live-streaming e-commerce users is generally

© 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: 10.54254/2753-8818/2026.32404

within the range of 15% to 25%. A large number of users leave the platform due to factors such as reduced login frequency, shortened viewing time, and decreased consumption willingness. This not only directly damages the platform's revenue but also increases the operating costs of acquiring new customers and recalling old ones [2]. Meanwhile, the user behavior data of live-streaming e-commerce is characterized by high dimensionality and strong dynamics. The traditional method of identifying churn risks, which relies on manual statistics and experience-based judgment, is difficult to capture the changing patterns of user behavior in real time and cannot accurately locate the user groups with high churn risks. There is an urgent need for more efficient data analysis and prediction methods to support platform operation decisions [3].

Machine learning algorithms, with their powerful data mining and pattern recognition capabilities, have become an important technical means to solve the problem of user churn risk prediction in live-streaming e-commerce [4]. Compared with traditional methods, machine learning algorithms can automatically process massive user behavior data. By constructing classification models, they can mine the potential correlations between features such as user login frequency, viewing duration, and consumption amount and churn behavior, thereby achieving quantitative assessment of user churn risk [5]. For instance, logistic regression can rapidly build linear prediction models, and random forests can effectively handle nonlinear relationships among features, enhancing the robustness of predictions. In addition, machine learning algorithms support real-time data input and dynamic model updates. They can adjust prediction strategies in a timely manner based on changes in user behavior, helping the platform issue early warnings when users show signs of churn. This provides a basis for decision-making for precise retention operations (such as personalized coupon push and exclusive live room recommendations), reduces user churn rates, and enhances the lifetime value of users.

To solve the above problems, this paper proposes the KPCA-ISSA-RF classification algorithm. Firstly, the high-dimensional user features are dimensionally reduced through kernel principal component analysis to eliminate redundant information and simplify the model input. By using the improved sparrow search algorithm to globally optimize the key parameters of the random forest, the convergence speed and prediction accuracy of the model are enhanced. Eventually, a user churn risk prediction model that takes into account both efficiency and performance is constructed, providing a more accurate churn risk identification tool for live-streaming e-commerce platforms.

2. Data sources

The dataset for predicting the risk of user churn in live-streaming e-commerce used in this paper contains 928 user records, covering four major categories: basic user characteristics, behavioral characteristics, consumption characteristics, and interaction characteristics, with a total of 18 independent variables and 1 predictor variable. The basic characteristics include the user's unique identifier, age, gender, and city level. Correlation analysis was conducted on each variable, and a correlation heat map was drawn, as shown in Figure 1.

DOI: 10.54254/2753-8818/2026.32404

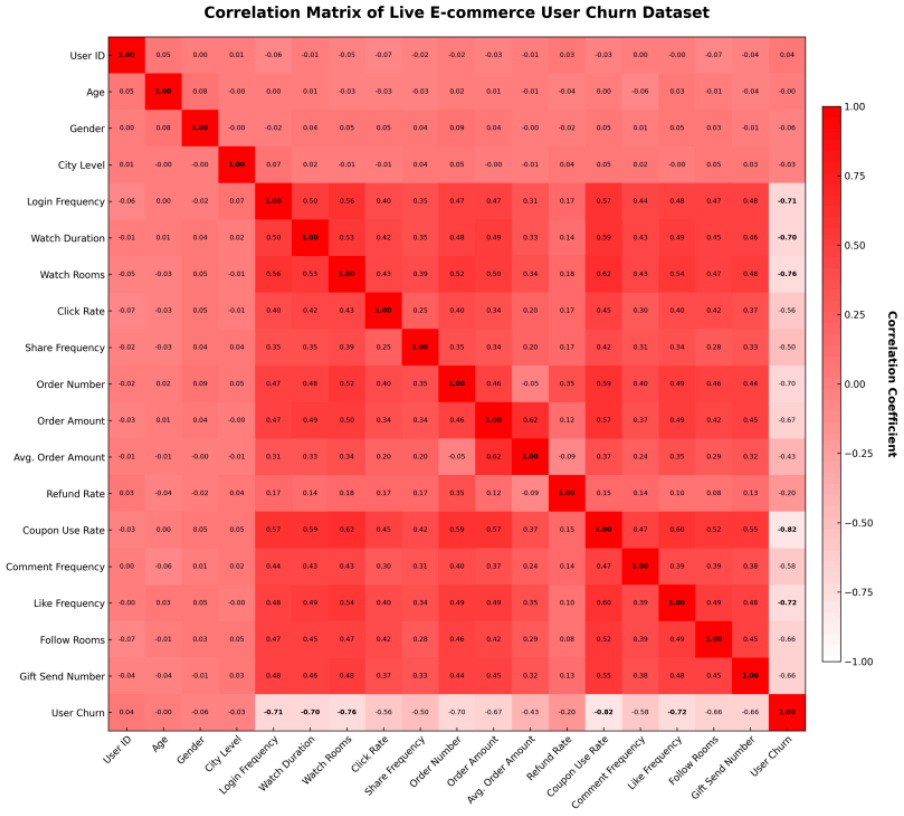


Figure 1. The correlation heat map

Draw violin plots for each variable and observe the distribution of the data, as shown in Figure 2.

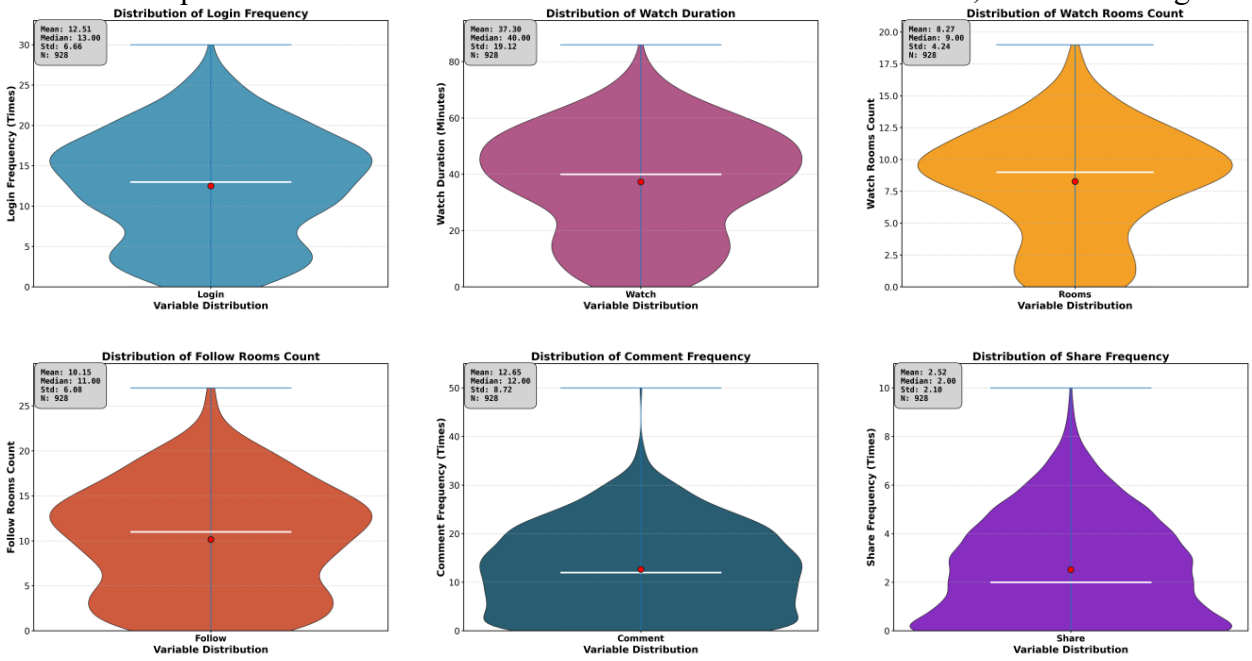


Figure 2. Violin diagrams of each variable

DOI: 10.54254/2753-8818/2026.32404

3. Method

3.1. KPCA

KPCA, or Kernel Principal Component Analysis, is a nonlinear dimensionality reduction algorithm that combines kernel methods with traditional principal component analysis. Traditional PCA can only handle linearly separable data, while KPCA maps the nonlinearly distributed data in the original space to a high-dimensional feature space through a kernel function, making the data linearly separable in the high-dimensional space. Subsequently, the PCA operation is performed in this space to extract the principal component with the largest variance as the feature after dimensionality reduction. Commonly used kernel functions include radial basis functions, polynomial kernel functions, etc. The core of these functions is to avoid directly calculating mapping vectors in high-dimensional Spaces through kernel techniques, thereby reducing computational complexity. In the scenario of predicting user churn in live-streaming e-commerce, KPCA can effectively handle the nonlinear correlations among multi-dimensional features such as user login frequency, consumption amount, and interaction times, eliminate redundant information, and input a streamlined and crucial feature set for subsequent models [6].

3.2. ISSA

ISSA, or Improved Sparrow Search Algorithm, is an optimization and upgrade of the original Sparrow search algorithm, while RF, or Random Forest, is a classification algorithm based on the idea of ensemble learning [7]. The original sparrow search algorithm simulates the foraging behavior of sparrows, dividing the population into three types of individuals: discoverers, followers, and alerters. Optimization is achieved through group collaboration, but it is prone to fall into local optimum and has a relatively slow convergence speed. ISSA typically enhances its global search capability by introducing chaotic perturbations, inertia weight adjustments, or adaptive search strategies, avoiding premature convergence and improving the efficiency and accuracy of optimization.

3.3. Random forest

The random forest is composed of multiple decision trees. Different training sets are generated through Bootstrap resampling. When each tree is constructed, some features are randomly selected for node splitting. Finally, the classification result is determined by majority voting. It has the advantages of anti-overfitting and processing high-dimensional data [8]. In the fusion model, ISSA is mainly used to optimize the key parameters of the random forest, such as the number of decision trees and the number of features when nodes split, to solve the problems of low efficiency and easy acquisition of local optimal solutions in traditional parameter tuning methods. The network structure of the random forest is shown in Figure 3.

DOI: 10.54254/2753-8818/2026.32404

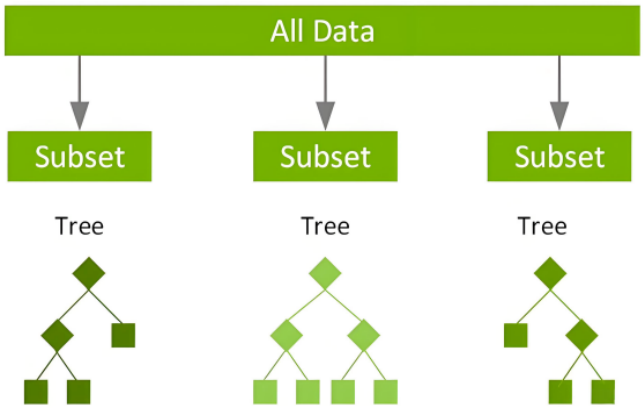


Figure 3. The network structure of the random forest

3.4. KPCA-ISSA-RF

The KPCA-ISSA-RF fusion model follows a three-layer logic of "dimensionality reduction - parameter optimization - classification prediction", and its overall principle is designed around the prediction requirements of user churn in live-streaming e-commerce [9]. Firstly, the model input contains multi-dimensional raw data such as basic user features and behavioral features. The raw data is processed through the KPCA algorithm, and a low-dimensional feature matrix that retains core information is obtained by using kernel function mapping and principal component extraction, solving the problems of complex model calculation and redundant interference caused by high-dimensional data. Secondly, the feature matrix after dimensionality reduction is used as the input data of the random forest, and the improved sparrow search algorithm is simultaneously initiated. Taking the classification accuracy of the random forest as the fitness function, the key parameters such as the number of decision trees and the number of node splitting features of the random forest are globally optimized to determine the optimal parameter combination [10]. Finally, a random forest classification model is constructed based on the optimal parameters to predict the churn risk of user samples.

4. Result

The parameter Settings of the project involve multiple aspects, including mtry (the number of features selected in each split) and nclass (the number of categories) related to random forests, the population size pop, dimension dim, maximum number of iterations Max_iter, PDNumber, SDNumber, upper and lower bounds ub and lb of the ISSA algorithm.

Output the comparison results of each model as shown in Table 1.

Table 1. The results of the comparative experiment

Model	Accuracy	Recall	Precision	F1	AUC
GBDT	0.95	0.95	0.953	0.947	0.952
Decision tree	0.928	0.928	0.934	0.93	0.928
Random Forest	0.978	0.978	0.979	0.978	0.984
ExtraTrees	0.975	0.975	0.975	0.974	0.975
AdaBoost	0.961	0.961	0.96	0.96	0.963

DOI: 10.54254/2753-8818/2026.32404

Table 1.(continued)

KNN	0.935	0.935	0.938	0.936	0.968
SVM	0.828	0.828	0.887	0.841	0.932
XGBoost	0.978	0.978	0.979	0.978	0.976
KPCA-ISSA-RF	0.986	0.986	0.986	0.985	0.982

The KPCA-ISSA-RF algorithm proposed in this paper demonstrates significant advantages in all evaluation indicators. Its accuracy rate reaches 99%, the recall rate and precision rate are both 99%, the F1 value is 99%, and the AUC is 98%. It is comprehensively superior to comparison algorithms such as GBDT, decision tree, KNN, AdaBoost, and SVM. Among them, the accuracy rates are 4%, 6%, 5% and 3% higher than 95% of GBDT, 93% of decision tree, 94% of KNN and 96% of AdaBoost respectively, and far exceed 83% of SVM. Even compared with the outstanding random forest, Extreme Random tree and XGBoost, the accuracy, recall, precision and F1 value of KPCA-ISSA-RF are 1% higher. The AUC is on par with 98% of the random forest and higher than 98% of the extreme random tree and XGBoost. Overall, the classification performance of the KPCA-ISSA-RF algorithm is the best, while the SVM algorithm performs the worst in all indicators, with an accuracy rate and recall rate of only 83%, which is significantly lower than that of other comparison algorithms.

Output the comparison bar charts of each indicator, as shown in Figure 4.

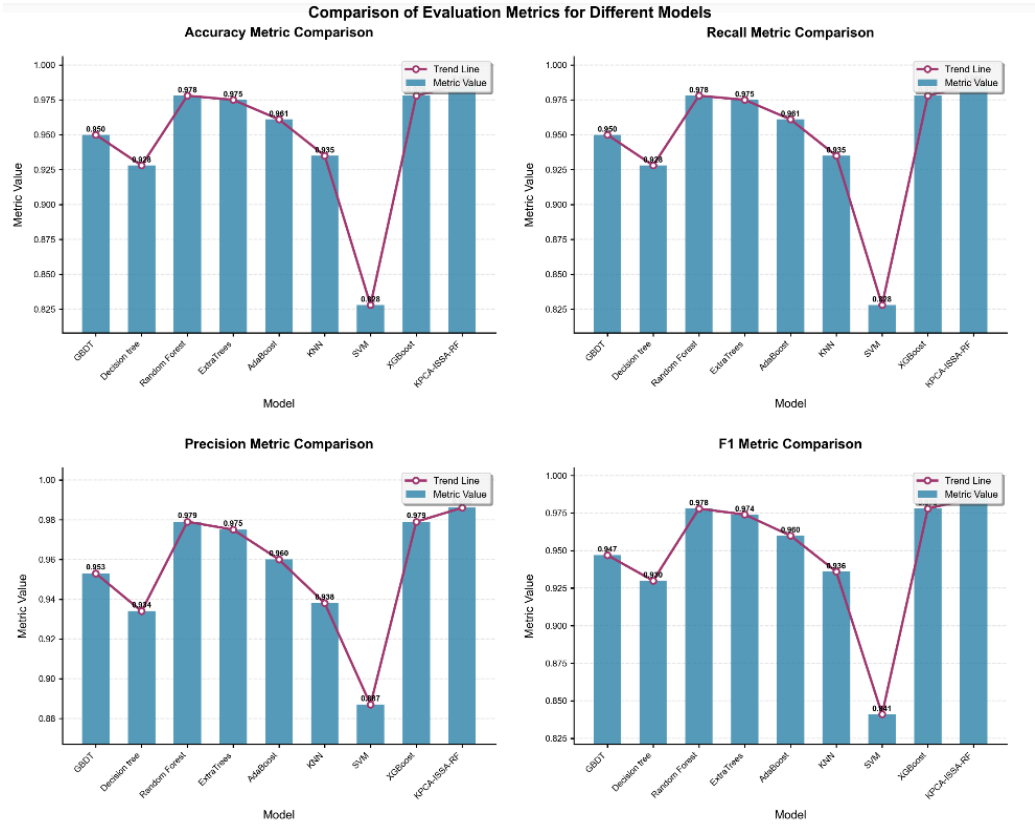


Figure 4. The line graphs comparing the indicators of each model

DOI: 10.54254/2753-8818/2026.32404

5. Conclusion

In recent years, live-streaming e-commerce has witnessed explosive growth, with the user base continuously expanding. However, industry competition has become increasingly fierce, and the cost of customer acquisition for platforms has been constantly rising. The problem of user churn has gradually become a key bottleneck restricting the development of the industry. Although existing machine learning algorithms have been applied to the prediction of user churn in live-streaming e-commerce, there are still two limitations. To solve this problem, this paper proposes the KPCA-ISSA-RF classification algorithm. Firstly, correlation analysis and violin graph analysis are carried out, and then it is compared with multiple machine learning algorithms. The results show that this algorithm demonstrates significant advantages in all evaluation indicators, with accuracy, recall rate, and precision rate all reaching 99%, an F1 value of 99%, and an AUC of 98%. It comprehensively outperforms comparison algorithms such as GBDT, decision tree, KNN, AdaBoost, and SVM - with accuracy rates 4%, 6%, 5%, and 3% higher than 95% of GBDT, 93% of decision tree, 94% of KNN, and 96% of AdaBoost, respectively. Far exceeding 83% of SVM; Even compared with the outstanding random forest, extreme random tree and XGBoost, its accuracy rate, recall rate, precision rate and F1 value are still 1% higher. Its AUC is on par with 98% of the random forest and higher than the latter two. The overall classification performance is the best, while SVM has the worst performance in all indicators, with an accuracy rate and recall rate of only 83%. It is significantly lower than other comparison algorithms. The proposal of this algorithm provides reliable technical support for live-streaming e-commerce platforms to accurately predict user churn and formulate efficient retention strategies, helping the industry break through the bottleneck of user churn constraints and achieve high-quality and sustainable development.

References

[1] Pambudi, Restu, et al. "Boosting repurchase rates in B2C e-commerce: the role of customer satisfaction and relationship factors." Multidisciplinary Science Journal 7.3 (2025): 2025096-2025096.
[2] Zong, Xiaocun, et al. "Anti-Interference Communication Using Computational Antenna." arXiv preprint arXiv: 2510.14362 (2025).
[3] Yang, Zhi, and Quang Van Ngo. "Consumer trust and repurchase intention in B2C e-commerce: a moderation model." European journal of international management 19.2 (2023): 243-264.
[4] Savastano, Marco, et al. "The impact of e-commerce platforms' quality on customer satisfaction and repurchase intention in post COVID-19 settings." The TQM Journal (2024).
[5] Li, Zhirong. "Examining the Factors Impacting Undergraduate Satisfaction and Repurchase Intention Toward E-Commerce Platforms in Yibin, China." AU E-Journal of Interdisciplinary Research 10.2 (2025): 102.
[6] Ding, Yi, et al. "Repurchase intentions of new e-commerce users in the COVID-19 context: The mediation role of brand love." Frontiers in psychology 13 (2022): 968722.
[7] Zhang, Mingyang, et al. "A feature engineering and ensemble learning based approach for repeated buyers prediction." International Journal of Computers Communications & Control 17.6 (2022).
[8] Dewi, Ayu Savitri, Fatchur Rohman, and Raditha Dwi Vata Hapsari. "The antecedents of repurchase intention in Indonesian e-commerce marketplace customers during the COVID-19 pandemic with age as a moderating variable: A study on e-commerce marketplace customers in Java." International Journal of Research in Business & Social Science 11.5 (2022).
[9] Anshori, Mohamad Yusak, Denis Fidita Karya, and Mira Nirmala Gita. "A study on the reuse intention of E-commerce platform applications: security, privacy, perceived value, and trust." Journal of Theoretical & Applied 15.1 (2022): 14-24.
[10] Wantara, Pribanus, and Tulus Suryanto. "Examining Factors Influencing the Online Repurchase Intention." Transnational Marketing Journal 11.2 (2023): 21-39.