Proceedings of Business and Economic Studies, 2024, Volume 7, Issue 6
http://ojs.bbwpublisher.com/index.php/PBES
Online ISSN: 2209-265X
Print ISSN: 2209-2641

Machine Learning Models for Predicting Order Returns in Cross-Border E-Commerce

Jia Cai[1,2], **Ronaldo Juanatas**[1]*, **Apollo Portez**[1], **Jonan Rose Montaña**[1]

[1]Technological University of the Philippines, Manila 1000, Philippines
[2]Henan Finance University, Zhengzhou 450014, Henan Province, China

Corresponding author: Ronaldo Juanatas, ronaldo_juanatas@tup.edu.ph

Abstract: This study investigates the application of machine learning models to address after-sales service issues in cross-border e-commerce, focusing on predicting order returns to reduce return costs and optimize customer experience. Using H cross-border e-commerce company as a case study, the research employs Random Forest and XGBoost models to identify high-risk return orders. By comparing the performance of these two models, the study highlights their respective strengths and weaknesses and proposes optimization strategies. The findings provide a valuable reference for e-commerce companies to refine their business models, reduce return rates, improve operational efficiency, and enhance customer satisfaction.

Keywords: Random Forest Model; XGBoost Model; After-sales issues; Prediction

Online publication: December 23, 2024

1. Introduction

With platforms such as Amazon, Alibaba, and eBay offering global shipping and logistics solutions, consumers can now purchase products from international sellers as easily as from local vendors. According to Statista, approximately 56% of manufacturers and retailers globally will engage in online overseas product sales in 2023, with e-commerce revenue expected to grow at a compound annual growth rate of 12.9% from 2017 to 2029 [1]. Cross-border e-commerce is poised to become a vital sales channel in the future.

In cross-border e-commerce, after-sales service challenges are more complex and frequent due to factors such as language barriers, cultural differences, logistics inefficiencies, and time zone discrepancies [2]. A key challenge faced by global cross-border merchants in 2023 is the high proportion of return costs, accounting for 26.6%, while return-related factors contribute to 6% of global cross-border online shopping customer complaints. These issues significantly impact customer retention and brand reputation. If enterprises fail to address these problems effectively, they risk losing a substantial customer base and hindering business growth.

Machine learning, a subset of artificial intelligence, offers a promising solution to these challenges by enabling businesses to identify potential after-sales issues, predict return likelihood, and take proactive measures. It can help reduce operational costs, enhance personalized customer experiences, optimize logistics, and improve after-sales efficiency [3].

This study focuses on H cross-border e-commerce company as a case study, utilizing machine learning models to predict order return scenarios and identify high-risk return orders. These predictions enable businesses to implement preventive measures, reducing return-related costs. The research employs two machine learning models, Random Forest and XGBoost, to explore the complexity and breadth of data features and patterns involved in predicting return issues. The performance of these models is analyzed and compared, with proposed optimization measures aimed at improving the accuracy and effectiveness of return predictions.

2. Literature review

In recent years, the application of machine learning technology in the e-commerce field has expanded significantly, with extensive research focusing on leveraging various machine learning models to analyze customer data and predict potential after-sales service issues. Common machine learning models include random forest, XGBoost, logistic regression, and deep learning algorithms. Traditional machine learning techniques, such as random forests, are widely applied to classification and regression tasks to identify key factors influencing after-sales problems [4].

The random forest model is particularly effective in handling large volumes of heterogeneous data from e-commerce platforms, such as order characteristics, customer behavior, and product attributes [5]. By constructing multiple decision trees and employing a voting mechanism, it significantly enhances the accuracy of predicting after-sales service issues [6]. Furthermore, random forests improve the prediction accuracy of minority class problems through techniques such as adaptive sample weighting, oversampling, and undersampling [7]. For example, the model can identify features strongly correlated with return rates, including customer behavioral history and product categories, which aids platforms in reducing unnecessary return costs [8]. Researchers have emphasized the applicability of random forests in addressing complex after-sales service challenges within cross-border e-commerce environments, highlighting their robustness in predicting such issues [9].

The XGBoost (Extreme Gradient Boosting) model, on the other hand, offers efficient gradient-boosting algorithms and superior predictive capabilities. For instance, some studies have proposed a return prediction model that incorporates feature engineering during data preprocessing, significantly enhancing the model’s ability to identify return behavior and improving the recall rate by analyzing order characteristics, such as purchase frequency and logistics time [10]. XGBoost is particularly adept at capturing patterns of return behavior, optimizing predictive performance, and managing high-dimensional data effectively for return predictions in e-commerce [11].

Comparative studies have shown that XGBoost consistently outperforms traditional models, including random forests, in terms of accuracy and F1 scores [12]. By integrating user behavior characteristics with order information, XGBoost has been demonstrated to substantially improve the accuracy of return predictions in e-commerce settings [13]. Compared to alternative machine learning methods, such as the K-nearest neighbor algorithm (KNN) and support vector machine (SVM), XGBoost exhibits greater accuracy, stability, and computational efficiency, establishing it as a preferred model for e-commerce return predictions [14]. Researchers often optimize XGBoost performance by tuning hyperparameters, such as tree depth and subsampling rate, to achieve optimal prediction results for datasets from various e-commerce platforms [15].

3. Methodology

This study employs two machine learning models, random forest and XGBoost, to predict the return status of orders from H E-commerce Company. The methodology includes analyzing influencing factors, evaluating model performance, and proposing optimization measures. The approach aims to provide an intuitive understanding of constructing predictive models for e-commerce return issues and selecting relevant indicators. By overcoming the limitations of traditional regression analysis methods, the models capture the complexity and breadth of data features and patterns, thereby improving the effectiveness and accuracy of e-commerce return predictions.

The data for this study were extracted from H E-commerce Company's "Shopee" cross-border platform and covered monthly transactions from September 30, 2024, to October 31, 2022. The dataset is compiled from six interconnected tables through one-to-one correspondence of primary keys, as detailed below. The experimental process consists of the following steps.

3.1. Data preprocessing

Multiple features related to e-commerce orders were collected to construct a predictive model for e-commerce returns. The dataset encompasses diverse dimensions, including orders, logistics, and customer behavior, as shown in **Table 1**. Data preprocessing steps were undertaken to ensure high-quality data and consistency of features during model training, thereby enhancing predictive performance.

Table 1. Comprehensive features for e-commerce return prediction

Feature category	Attribute	Description
Product information	Product ID	Unique identifier for a product
	Category	Product category
	Price (USD)	Price of the product in USD
	User rating (5-point scale)	Average user rating of the product
Order information	Order ID	Unique identifier for an order
	Sales country	Country where the order was placed
	Order amount (USD)	Total value of the order in USD
	Quantity of purchased goods	Number of items purchased in the order
	Category	Product category
	Product ID	Identifier linking order to a product
	Purchasing date	Date when the order was placed
	Customer ID	Identifier linking order to a customer
Customer information	Customer ID	Unique identifier for a customer
	Customer age	Age of the customer
	Customer gender	Gender of the customer
	Purchase frequency	Frequency of purchases by the customer
	Historical after-sales frequency	Number of after-sales issues reported previously

Table 1 (Continued)

Feature category	Attribute	Description
Logistics information	Logistics company	Name of the logistics service provider
	Delivery duration (days)	Number of days taken to deliver the order
	Delivery delay frequency	Number of instances of delayed deliveries
	Delivery area	Geographic location of delivery
Customer characteristics	Browsing time (minutes)	Total browsing time spent by the customer
	Shopping cart dwell time (hours)	Time spent on the shopping cart page
	Customer review (5-point scale)	Review score given by the customer
Other features	Customer location	Geographic location of the customer
	Customer device type	Type of device used to access the platform
	Follow-up behavior	Post-purchase engagement activities by the customer

3.2. Dataset partitioning

To effectively train and evaluate the e-commerce return prediction models, the dataset was randomly divided into a training set (70%) and a testing set (30%). Before partitioning, preprocessing steps were conducted, including handling missing values, addressing category imbalances, and feature standardization, to ensure data quality and consistency.

3.3. Model training and optimization

Hyperparameter optimization was performed for both random forest and XGBoost using grid search. For the random forest model, parameters such as "max_depth", "min_samples_leaf", "min_samples_split", and "n_ estimators" were tuned to improve prediction accuracy and generalization. For the XGBoost model, parameters such as "learning_rate", "max_depth", "n_estimators", and "subsamples" were adjusted to address class imbalances and enhance predictive accuracy for returned items.

3.4. Model evaluation

Model evaluation focused on metrics such as accuracy, precision, recall, and F1 score, using the test dataset. A confusion matrix was employed to demonstrate classification performance for both "returned" and "non-returned" categories. Additionally, feature importance analysis was conducted to identify key factors influencing model predictions.

4. Results and discussion

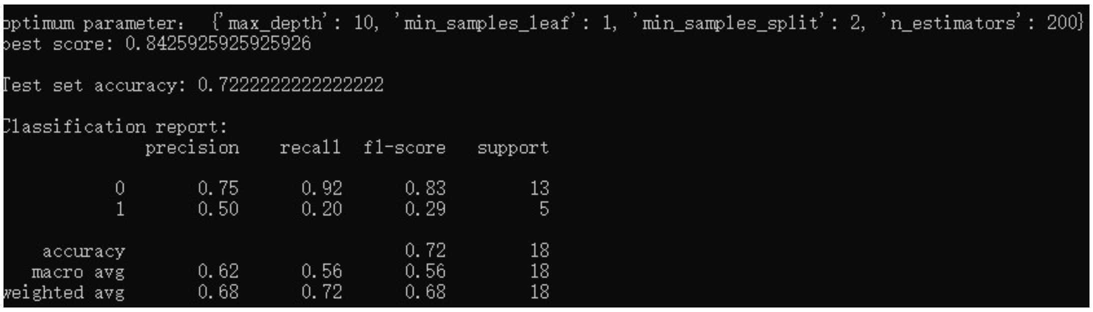
The experimental results highlight notable differences in the performance of the random forest and XGBoost models for return prediction. A detailed comparison is as follows.

4.1. Experiment results

The key experimental outcomes are presented in **Figures 1** and **2**.

(1) Random Forest Model:

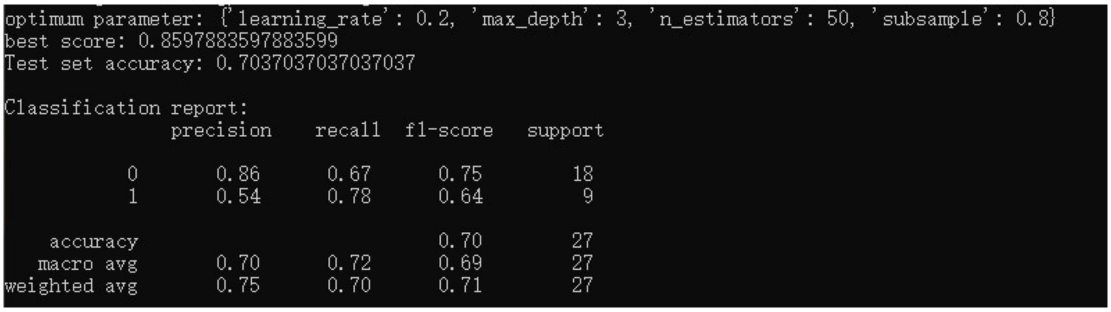
Best Parameters: “max_depth=10”, “min_samples_leaf=1”, “min_samples_split=2”, “n_estimators=200”
Test Set Accuracy: 0.722
Classification Metrics:
Category 0 (Non-Return): Precision = 0.75, Recall = 0.92, F1 Score = 0.83
Category 1 (Return): Precision = 0.50, Recall = 0.20, F1 Score = 0.29



```
ptimum parameter:  {'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}
est score: 0.8425925925925926

est set accuracy: 0.7222222222222222

lassification report:
             precision    recall  f1-score   support

           0       0.75      0.92      0.83        13
           1       0.50      0.20      0.29         5

    accuracy                           0.72        18
   macro avg       0.62      0.56      0.56        18
eighted avg       0.68      0.72      0.68        18
```

Figure 1. Random Forest Model experiment result

(2) XGBoost Model:
Best Parameters: “learning_rate=0.2”, “max_depth=3”, “n_estimators=50”, “subsample=0.8”
Test Set Accuracy: 0.704
Classification Metrics:
Category 0 (Non-Return): Precision = 0.86, Recall = 0.67, F1 Score = 0.75
Category 1 (Return): Precision = 0.54, Recall = 0.78, F1 Score = 0.64



```
optimum parameter: {'learning_rate': 0.2, 'max_depth': 3, 'n_estimators': 50, 'subsample': 0.8}
best score: 0.8597883597883599
Test set accuracy: 0.7037037037037037

Classification report:
              precision    recall  f1-score   support

           0       0.86      0.67      0.75        18
           1       0.54      0.78      0.64         9

    accuracy                           0.70        27
   macro avg       0.70      0.72      0.69        27
weighted avg       0.75      0.70      0.71        27
```

Figure 2. XGBoost Model experiment result

4.2. Model performance

The random forest model achieves an accuracy of 72.2%, outperforming XGBoost’s accuracy of 70.4%. Its precision for non-return orders (Category 0) is 92%, with an F1 score of 83%, indicating strong predictive performance for non-return samples. The recall rate of 0.92 further underscores its ability to accurately identify non-return orders, making it effective in reducing false alarms. However, its recall for return orders (Category 1) is only 20%, highlighting a weakness in identifying return samples, which could lead to missed predictions of return risks.

On the other hand, the XGBoost model demonstrates a higher recall rate for return orders (Category 1) at 0.78 compared to the random forest model’s 0.20. This indicates XGBoost’s superior ability to identify return risks, which is essential for addressing business challenges related to returns. Despite this, its overall classification performance, including accuracy and precision for non-return samples, is slightly lower than that of the random forest model.

Random Forest strengths: Higher accuracy, precision, and F1 score for non-return samples, making it well-suited for scenarios requiring reduced false alarms.

XGBoost strengths: Higher recall for return samples, aiding in the identification of return risks, though at the expense of slightly lower overall accuracy and weaker performance for non-return predictions.

4.3. Confusion matrix analysis

The confusion matrix provides insights into the model's ability to predict "no return" and "return" outcomes.

(1) Random Forest Model: **Figure 3** illustrates the confusion matrix for the random forest model. The model correctly predicted 10 instances as category 0 (non-return) but misclassified 3 instances of category 0 as category 1 (return). For category 1, the model misclassified 5 instances as category 0 and failed to correctly predict any instances of category 1. While the model's overall accuracy is relatively high, its performance in predicting returns is significantly weaker. The recall rate for category 1 is only 0.20, demonstrating a bias toward non-return samples. This limitation leads to a large number of return orders being misclassified as non-return, which highlights the model's inadequacy in identifying return samples. Despite its strong accuracy in predicting non-return orders, the random forest model struggles to correctly identify return orders, resulting in a low recall rate for category 1.

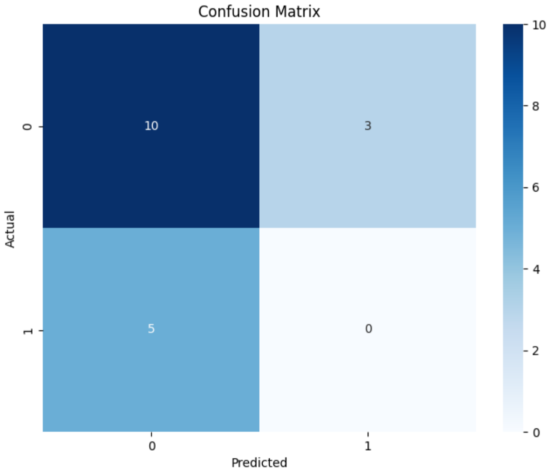


Figure 3. Confusion matrix analysis by Random Forest Model

(2) XGBoost Model: **Figure 4** displays the confusion matrix for the XGBoost model. The model correctly predicted 12 non-return instances but misclassified 6 non-return instances as returns. For category 1 (returns), the model correctly identified 7 instances but misclassified 2 return instances as non-return. The XGBoost model demonstrates a stronger ability to identify return samples, achieving a recall rate of 0.78. However, its recognition ability for non-return samples is slightly inferior, which slightly lowers its overall accuracy. The confusion matrix indicates that XGBoost performs better in identifying return orders, making it more suitable for scenarios that prioritize recognizing potential returns. Although its overall accuracy is marginally lower than that of the random forest model, its higher recall rate for return samples makes XGBoost particularly valuable for addressing return risks in cross-border e-commerce.

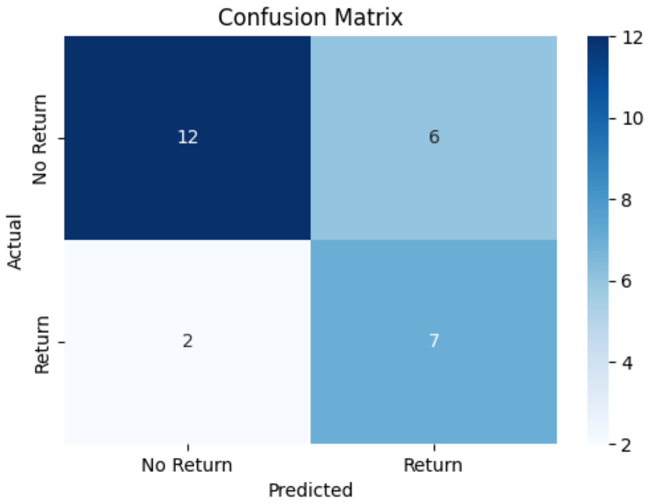


Figure 4. Confusion matrix analysis by XGBoost Model

4.4. Feature importance analysis

Feature importance analysis highlights the contribution of individual features to model predictions.

(1) Random Forest Model: **Figure 5** depicts the feature importance ranking derived from the random forest model. Features such as "number of delivery delays," "purchase date," "logistics company," and "order amount (USD)" are identified as the most influential. This suggests that delivery performance and order characteristics significantly impact return predictions. Specifically, "number of delayed deliveries" emerges as the most critical factor, followed by "purchase date," "logistics company," and "order amount." These findings imply that optimizing delivery reliability, selecting dependable logistics partners, and implementing strategic pricing can improve prediction accuracy and enhance operational efficiency.

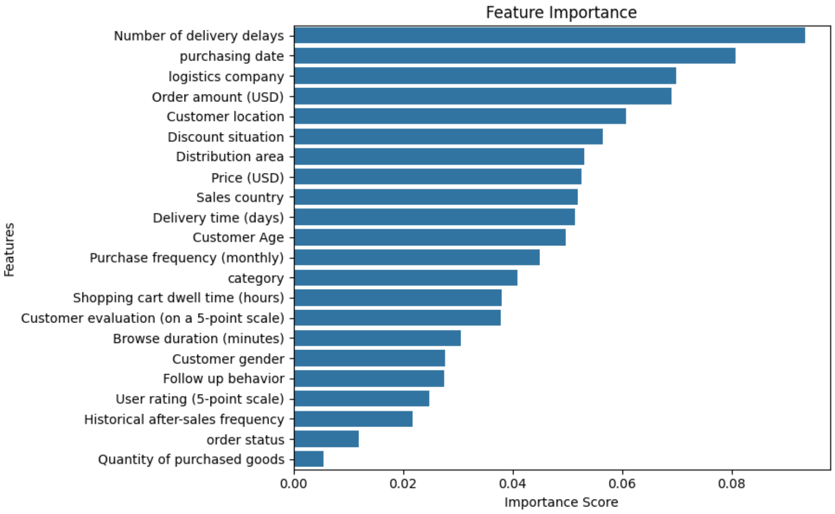


Figure 5. Feature importance analysis by Random Forest Model

(2) XGBoost Model: **Figure 6** presents the feature importance analysis for the XGBoost model. Unlike the random forest model, the most influential feature is "product category," followed by "discount situation,"

"number of delivery delays," and "purchase frequency (monthly)." In contrast, features like "shopping cart dwell time" and "price (USD)" have relatively minor impacts on predictions. This indicates that product attributes and purchasing conditions play pivotal roles in predicting return behavior. The results suggest that focusing on high-impact factors such as product categories and delivery performance can help reduce return rates. Businesses can utilize these insights to refine their return management strategies and improve operational outcomes.

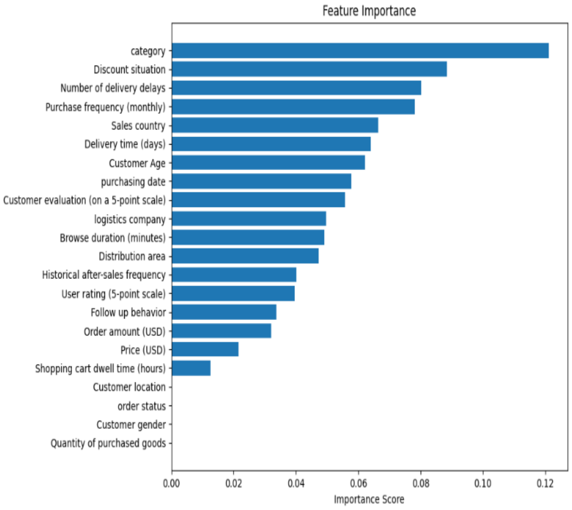


Figure 6. Feature importance analysis by XGBoost Model

4.5. Model optimization suggestions

The experimental results demonstrate that the random forest model excels in identifying the "non-return" category, while the XGBoost model shows superior performance in recognizing "return" samples. To achieve a more balanced classification performance, integrating the strengths of both models is recommended. This integration could combine the high accuracy of random forests with the strong recall rate of XGBoost for return samples.

To further enhance model performance, the following optimization strategies can be considered:

(1) Addressing class imbalance: Applying upsampling or downsampling techniques can improve the model's ability to recognize minority classes, such as return samples. Introducing additional interactive features from existing data, such as customer return history or regional return rates, may also significantly impact the accuracy of return predictions.

(2) Model fusion strategy: Implementing a model fusion strategy, such as weighted averaging of the prediction results from random forest and XGBoost, could achieve better predictive balance. This approach leverages the high accuracy of random forest for non-return samples and the strong recall rate of XGBoost for return samples, leading to more stable and reliable predictions.

(3) Parameter fine-tuning and threshold optimization: Further optimization of XGBoost parameters, particularly min_child_weight and gamma, can help suppress overfitting and improve model performance. Adjusting the probability threshold for return prediction can make the model more sensitive to return samples, enhancing its applicability in business contexts.

(4) Business applications and model interpretability: Utilizing feature importance analysis can inform

specific business optimization strategies, such as improving the quality inspection processes for high-risk product categories and optimizing logistics services to reduce return rates. Interpretable methods, such as Shapley Additive Explanations (SHAP) values, can help business personnel understand the influence of key features on predictions. This understanding facilitates the translation of model outputs into actionable operational strategies.

By implementing these optimization measures, the predictive models can be further refined, contributing to more effective management of returns in cross-border e-commerce and enhanced overall business performance.

5. Conclusion

This study provides an in-depth analysis of cross-border e-commerce return issues using machine learning models, focusing on random forest and XGBoost to predict influential features. The experimental results indicate that the random forest model performs well in the “non-return” category, achieving high precision and overall accuracy. In contrast, XGBoost demonstrates advantages in identifying the “return” category, effectively capturing potential return orders. These findings suggest that selecting or combining different models based on specific business needs can lead to improved prediction outcomes.

Through feature importance analysis, key factors influencing return predictions were identified, including delivery delays, order amounts, and product categories. These insights offer actionable optimization strategies for cross-border e-commerce enterprises, enabling them to reduce return rates, enhance operational efficiency, and improve customer satisfaction.

In summary, this study highlights the application value of machine learning models in addressing cross-border e-commerce return predictions. Future research could explore advanced techniques, such as complex data augmentation, category imbalance handling, and deep learning methods, to further enhance the adaptability and predictive accuracy of models in complex business scenarios.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Statista, 2024, Cross-Border E-Commerce.

[2] Chen S, 2023, Study on Customer Satisfaction of Cross-Border Import E-Commerce Logistics Service Based on Online Review, dissertation, Tongji University.

[3] Xie W, 2024, Prediction of Cross-Border E-Commerce Sales Volume Based on Machine Learning, dissertation, Jilin University.

[4] Qian Y, 2023, Research on User Behavior for E-commerce Platforms Based on Machine Learning, dissertation, Tianjin University of Commerce.

[5] Shi Y, 2021, Research on E-Commerce Customer Value Identification Based on Improved RFM Model, dissertation, Harbin University of Commerce.

[6] Yang S, 2023, 2023, Study on Customer Satisfaction of Cross-Border Import E-Commerce Logistics Service Based on Online Review, dissertation, Donghua University.

[7] Zhao M, Wang Z, 2020, Prediction of Customer Returns in E-Commerce Based on Random Forest. Journal of Retailing and Consumer Services, 55: 102070.

[8] Song J, 2023, Prediction and Analysis of B2C User Purchase Behavior Based on Machine Learning, dissertation, Northeast Normal University.

[9] Li M, 2023, E-Commerce User Purchase Behavior Prediction Based on Machine Learning Methods, dissertation, Dongbei University of Finance and Economics.

[10] Liu Y, 2023, E-Commerce Customer Loss Prediction Based on Machine Learning, dissertation, Southwest University.

[11] Li T, 2022, Research on the Improvement of Customer Satisfaction of Home Appliances Based on E-Commerce Data Mining, dissertation, Capital University of Economics and Business.

[12] Shi W, Liu X, Wang J, 2021, Addressing Class Imbalance in E-Commerce Returns Prediction Using Resampling Techniques and XGBoost. Journal of Business Research, 127: 145–153.

[13] Zhou Y, Chen F, Li Q, 2022, Enhancing E-Commerce Prediction Models Through XGBoost and User Behavior Analytics. Journal of E-Commerce Research, 9(3): 202–215.

[14] Yang Y, Zou X, Li C, 2023, User Purchase Behavior Prediction Method Based on XGBoost. Electronics, 12(9).

[15] Chen T, Guestrin C, 2016, XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.