# Empirical Analysis of User Repurchase Behavior Prediction, Channel Attribution and Causal Effect under Multichannel Marketing

Yangshangjia Hu

Faculty of Humanities and Arts, Macau University of Science and Technology, Avenida Wai Long, Taipa, Macau, China

1230034851@student.must.edu.modu

**Abstract.** Against the backdrop of homogeneous e-commerce competition, logistics experience has become a key factor influencing user repurchase, yet its quantitative value and causal effect remain unclear. Based on 115,034 valid user data from a large e-commerce platform, this study constructs an integrated analytical framework of "repurchase prediction-channel attribution- causal verification". Logistic Regression and Random Forest are adopted for repurchase prediction, Shapley value is used to quantify channel contributions, and propensity score matching is employed to verify causal effects. The results show that the Random Forest model achieves the optimal performance (AUC=0.825, F1=0.631); logistics timeliness contributes as high as 77.0% to repurchase, serving as the core driving channel; the core channel combination consisting of the Southeast China Region, credit card payment, and logistics delivery within 3 days can significantly increase the repurchase rate by 3.9 percentage points, with statistically significant results. This study fills the research gap of the disconnect between logistics and multi-channel analysis, providing a scientific basis for enterprise resource allocation.

**Keywords:** E-commerce, Repurchase Prediction, Channel Attribution, Logistics Timeliness, Shapley Value.

## 1. Introduction

Amid the increasingly convergent competition in the e-commerce industry, product and price advantages can no longer sustain user loyalty in the long run, and the repurchase rate has gradually become a core growth indicator for enterprises (Lu, 2025) [8]. In addition, the multi-channel layout has exacerbated conflicts in enterprises' resource allocation. Current surveys show that practitioners generally recognize the importance of logistics experience for repurchase, and further studies have confirmed that logistics service quality is a key antecedent of e-commerce users' repurchase intention, with customer satisfaction further strengthening this impact (Chen, 2023) [1]. One key question is still unanswered. People don't know the exact measurable value logistics optimization brings to customer repurchases. This uncertainty puts enterprises in two tough spots. They might undervalue logistics optimization. They might lack sound scientific grounds when assigning resources to logistics and marketing channels alike.

This real-world problem has not been solved by existing academic research. Most repurchase prediction studies focus on users' clickstream and purchase history data. They always ignore logistics as a core factor driving repurchase behavior (Chou, 2021; Kim, 2021) [2][4]. Some research has used attribution methods like the Shapley value in certain scenarios. These scenarios include supply chain profit distribution and forest product e-commerce logistics (Liu, 2025) [7]. These studies do not treat logistics as an independent channel. They fail to measure its direct influence on end-users' repurchase decisions (Li, 2022) [5].

Another point needs more attention. Research on the full "prediction — attribution — causal verification" chain is still disconnected. Research findings end up being fragmented. Traditional analytical methods face a tough issue. They handle large-scale imbalanced datasets. They can hardly balance accuracy and strict methodological standards. The actionability of research results in real-world scenarios gets further limited (Park, 2023; Yılmaz Benk, 2022) [11][14]. Recent research has

made new attempts. It uses multimodal Shapley values. It aims to boost analytical accuracy in e-commerce settings (Obayemi, 2024) [9]. These attempts only apply to product classification tasks. They don't get extended to repurchase prediction or channel attribution. The analytical framework leaves out logistics. It becomes hard to match enterprises' actual resource allocation needs. This study addresses these interrelated problems. It builds a single analytical framework. The framework connects "repurchase prediction — channel attribution — causal verification" stages. It takes logistics timeliness as a core analytical dimension for the first time. The study uses more than 100,000 valid sample records. It seeks to find key predictors of repurchase behavior. It measures the unique contributions of different channels in concrete terms. It verifies the causal effects of core channels. The study puts forward a data-driven and reliable resource allocation plan. The plan fills the existing gaps in academic research. And it offers practical guidance for enterprise daily operations.

## 2. Literature Review

Building up e-commerce user loyalty needs one key step first. Pinpoint the users with the highest repurchase chances in a precise way. Early research relied a lot on simple models. The models are easy to understand, like Logistic Regression. A problem exists with these models. They work poorly with complex user behaviors. They often lack enough accuracy. They lack the predictive power for real business situations. Machine learning methods have grown in use in new studies. The methods include Random Forest.

Some scholars have tried combining these algorithms with the classic RFM model. The model assesses customer value by recency, frequency and monetary. They do this to polish repurchase predictions (Verma, 2025) [12]. Other scholars have taken the S-O-R framework from marketing studies. The framework stands for Stimulus-Organism-Response. They use it to predict online shopping behaviors. The approach maps out external triggers' influence on user psychology and actions. It brings new theoretical perspectives to repurchase research (Csoban-Mirka, 2024) [3]. The "buy-till-you-die" model has also been paired with machine learning. The model tracks customer lifecycles as a marketing tool. It creates insights across different research fields (Chou, 2021) [2]. All these prediction models have the same blind spot. They focus all attention on users' past clicks and purchase records. They ignore logistics timeliness. It is a key factor that directly affects customers' feelings about delivery experiences.

Figuring out which users will repurchase is only half the work. Resource investment decisions need fair measurement of each channel's contribution. The old method of crediting the last ad click for a conversion fails in today's complex business environment. The Shapley value method has become a more dependable option in practice. The method comes from game theory. Its application scope is surprisingly wide. It can divide profits in e-commerce supply chains (Zhou, 2024) [16]. It can settle profit-sharing conflicts between logistics partners (Liu, 2025; Zhang, 2022) [7] [15]. Game theory-based studies reveal another fact. Suppliers choose sales platforms mostly based on the platform's own logistics abilities (Wang, 2022) [13]. This finding highlights the importance of logistics. It also points out an existing gap. No current attribution research regards logistics as an independent channel. No research measures its direct effect on users'repurchase decisions.

This study also needs to confirm one thing. The relationships between channels and repurchases must be truly causal. Some methods are made for this very purpose. Propensity Score Matching (PSM) is one such method. They filter out external factors. They help check if there is a real cause-and-effect link. One study used PSM in practice. It proved that launching e-commerce operations can greatly increase cooperative profits (Liang, 2021) [6]. Look at the full process of "user prediction - channel attribution - causal verification". The research in this area is disconnected. The different parts don't fit together well. The biggest issue is logistics. It is missing from every part of this process.

This study builds a single framework to fix this problem. The framework connects all these steps into one. This study adds logistics-related features to the repurchase prediction model. And it clearly defines logistics as an independent channel. It uses the Shapley value method to calculate its

contribution. It does this along with different marketing channels. This study uses causal inference methods. It checks the real impact of these core channels. This framework solves the fragmentation problem in existing research. It gives e-commerce platforms a better decision-making way. The way is more consistent and useful.

## 3. Data and Research Methodology

### 3.1. Data Resources

This study's data comes from a big cross-border e-commerce platform in Brazil. It includes 24 months of real operational data (Olist, 2018) [10]. The data covers users'basic details. It has order transaction records. It includes payment method choices. It also has full-cycle logistics and delivery information. This study cleaned the data first. This study dealt with missing values. This study removed outlier data. Finally, this study got 115,034 valid sample records. The samples'overall repurchase rate is 29.8%. The observation window is 6 months. The data covers many types of information. It has users' consumption behaviors. It includes channel touchpoints. It also has logistics experience. This data provides a reliable empirical basis for the study.

### 3.2. Variable Definition

The dependent variable is user repurchase behavior (1: Repurchased within 6 months; 0: No repurchase). Independent variables include four categories of features: demographic characteristics, consumption behavior, channels, and logistics. The grouping variable is core channel exposure. Specific definitions are shown in Table 1.

**Table 1.** Variable definitions and measurement methods.

| Variable type | variable type | Variable definition and measurement method |
|---|---|---|
| Dependent variable | Repurchase behavior | Binary variable,<br>1: Repurchased within 6 months,<br>0: No repurchase |
| Consumption behavior characteristics | Total consumption amount | Continuous variable, cumulative consumption amount (USD) during the observation period, winsorized at the 99th percentile (upper limit: 1596 USD) |
| Consumption behavior characteristics | Number of purchases | Continuous variable<br>total number of orders in the observation period |
| Channel characteristics | Payment channel | Categorical variable, including four types: credit card, debit card, voucher, boleto |
| Logistics experience characteristics | Logistics timeliness | Continuous variable, the number of days from order placement to receipt, truncated at 0-30 days |
| Grouping variable | Core channel reach | Binary variable,<br>1 = southeast region + credit card + logistics < 3 days,<br>0 = failure to meet the combined conditions |

### 3.3. Data Preprocessing

First, missing value handling was conducted: 2,567 records with missing values in the logistics timeliness field were deleted to ensure sample integrity.

Second, outlier handling was performed: total consumption was truncated at the 99th percentile (with an upper limit of USD 1,596), and logistics timeliness was truncated at a 30-day upper limit to avoid interference from extreme values. For sample balance processing, the parameter class_weight='balanced' was adopted in model training to mitigate the sample imbalance bias caused by a 29.8% repurchase rate. For variable encoding and standardization, categorical features were encoded via LabelEncoder, and predictive features were standardized with StandardScaler to ensure model convergence. In terms of unifying variable calibers, the same sample set and variable

definitions were used for prediction, attribution and causal analysis to guarantee the consistency of analysis results.

### 3.4. Exploratory Data Analysis (EDA)

To identify data patterns through descriptive statistics and visualization, a variety of visual charts were constructed in this paper to compare the differences in feature distribution between repurchasing and non-repurchasing users, with a focus on verifying the negative correlation between logistics timeliness and repurchase behavior. The results of Figure 1 show that the correlation coefficient between the two is -0.35. Box plots were used to analyze the distribution differences of core features between the two user groups, and as shown in Figure 2, the logistics timeliness of repurchasing users is somewhat shorter than that of non-repurchasing users, which further confirms the negative correlation between them.

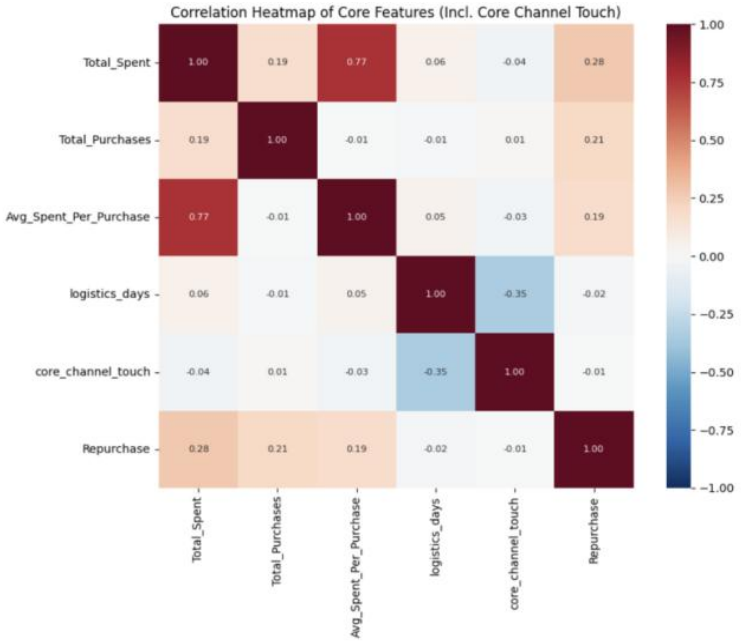



**Figure 1.** Correlation heatmap of core features (picture credit: original).




**Figure 2.** Feature comparison of repurchasing and non-repurchasing users (picture credit: original).

## 4. Research Methodology

### 4.1. Repurchase Prediction Model

This study picks two well-used machine learning methods. They are for building the repurchase prediction model. The methods are Logistic Regression and Random Forest.

Logistic Regression is a linear binary classification tool. Its core logic works like this. It turns linear combinations of features into values between 0 and 1. It uses the Sigmoid function to do this. This lets this study give a specific number. The number shows users' repurchase probability. For the research, this study uses L1 regularization. It uses it to pick out key features. This helps cut down interference. The interference comes from multicollinearity.

Random Forest is an ensemble learning method. Its way of working is simple. First, it builds a group of independent decision trees. Then it gets the final prediction result. It uses a voting mechanism to do this. This design helps this study capture non-linear correlations between features. This is just right for the complex influencing mechanism. The mechanism is behind repurchase behavior. This study adjusts key hyperparameters carefully. The hyperparameters include the number of decision trees and maximum tree depth. It uses grid search to do this.

Both models use class_weight='balanced'. This eases the problem of sample imbalance. It keeps in line with the data processing logic.

Baseline Model: Logistic Regression, with the core formula as follows:

$$P(y=1|X)=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n)}} \tag{1}$$

Where P(y=1|X) represents the user's repurchase probability, $\beta_0$represents the intercept term, $\beta_i$ represents the coefficient of each feature, and $X_i$ represents the standardized feature value.

Core Model: Random Forest, with the core formula for feature importance calculation as follows:

$$Importance(X_i)=\frac{\sum_{t=1}^{T}\left(Gini_t-Gini_{t\,left}-Gini_{t\,right}\right)}{T} \tag{2}$$

Where T represents the number of decision trees, $Gini_t$ represents the Gini coefficient before node splitting, and $Gini_{t\,left}$ and $Gini_{t\,right}$ represent the Gini coefficients of the left and right child nodes after splitting, respectively.

Grid Search Parameters: Optimize C=[1,10] and penalty=['l1'] for Logistic Regression; optimize n_estimators=300, max_depth=10, etc. for Random Forest.

### 4.2. Channel Attribution Method

Based on the optimal Random Forest model from the previous step, the Shapley value method is adopted to calculate the marginal contribution of each channel. During attribution, core features such as total consumption amount and average consumption amount per order are incorporated as control variables to reduce confounding interference. Three types of channel features—region, payment, and logistics—are selected as the attribution objects, and their relative contribution ratios are quantified by the mean absolute Shapley values.

The core calculation formula of the Shapley value is as follows:

$$\phi_i=\frac{1}{n!}\sum_{S\subseteq N\{i\}}|S|!\,(n-|S|-1)!\left(v\left(S\bigcup i\right)-v(S)\right) \tag{3}$$

Where $\phi_i$ represents the contribution degree of the i-th channel, S represents the channel subset, $v(S)$ represents the repurchase prediction contribution value of subset S, and n represents the total number of channels.

### 4.3. Causal Inference Method

Propensity Score Matching (PSM) was adopted to verify the causal effect of core channel reach. First, in terms of the matching strategy, core channel reach was taken as the treatment variable, and four covariates including total consumption amount were selected to construct a Logistic Regression model for predicting propensity scores. A 1:1 nearest neighbor matching strategy was adopted with the matching caliper set at 0.03, which was adjusted to 0.05 in the subsequent robustness test. For the balance test, the matching effect was verified by the standardized mean difference (SMD<0.1) and

Levene's Test for Homogeneity of Variances. For effect evaluation, the Average Treatment Effect on the Treated (ATT) was calculated, and its significance was verified via the independent samples t-test.

## 5. Construction of Repurchase Prediction Model, Channel Attribution and Causal Effect Test

Based on the repurchase prediction model constructed in Section 4.1, a 7:3 stratified sampling strategy was adopted to divide the dataset into training and test sets for empirical analysis. The results of model performance evaluation and feature importance analysis are presented as follows.

### 5.1. Data Processing, Model Construction and Performance Comparison

A 7:3 stratified sampling method was used to split the data into a training set of 80,523 records and a test set of 34,511 records. The repurchase rate for both sets was 29.8%, avoiding sampling bias that could affect model evaluation. There are 7 categories of predictive features.

The test set performance comparison of the two models is shown in Table 2. The Random Forest outperformed Logistic Regression on most metrics, with an AUC of 0.825 and an F1-score of 0.631, representing improvements of 11.3% and 17.7% respectively compared to Logistic Regression. Therefore, the Random Forest was selected as the base model for the subsequent attribution analysis.

**Table 2.** Performance comparison of repurchase prediction models on the test set.

| Model type | Test set f1 | Test set precision | Test set recall | Test set AUC |
|---|---|---|---|---|
| Logistic regression | 0.536 | 0.596 | 0.487 | 0.741 |
| Random forest | 0.631 | 0.595 | 0.671 | 0.825 |

### 5.2. Feature Importance Analysis

The feature importance ranking of the Random Forest model shows that total consumption amount (0.377), average consumption amount per order (0.261), and number of purchases (0.189) are the top three core features. Logistics timeliness ranks fifth with an importance score of 0.025, which is significantly higher than region code (0.011) and logistics code (0.005), verifying the predictive value of logistics timeliness at the model level. The feature coefficients of the Logistic Regression model indicate that total consumption amount (0.715) and number of purchases (0.574) have a positive impact on repurchase behavior, while logistics timeliness (-0.100) has a negative impact—meaning the shorter the logistics timeliness, the higher the repurchase probability.

### 5.3. Core Results of Shapley Value Attribution

The results of Shapley value attribution show that the logistics channel contributes 76.99% to repurchase behavior, while both the regional and payment channels account for 11.50% each. This finding indicates that logistics timeliness is the core channel driving repurchase behavior, with its contribution significantly exceeding that of the regional and payment channels, which reflects users' core demand for the receipt experience, as shown in Figure 3.

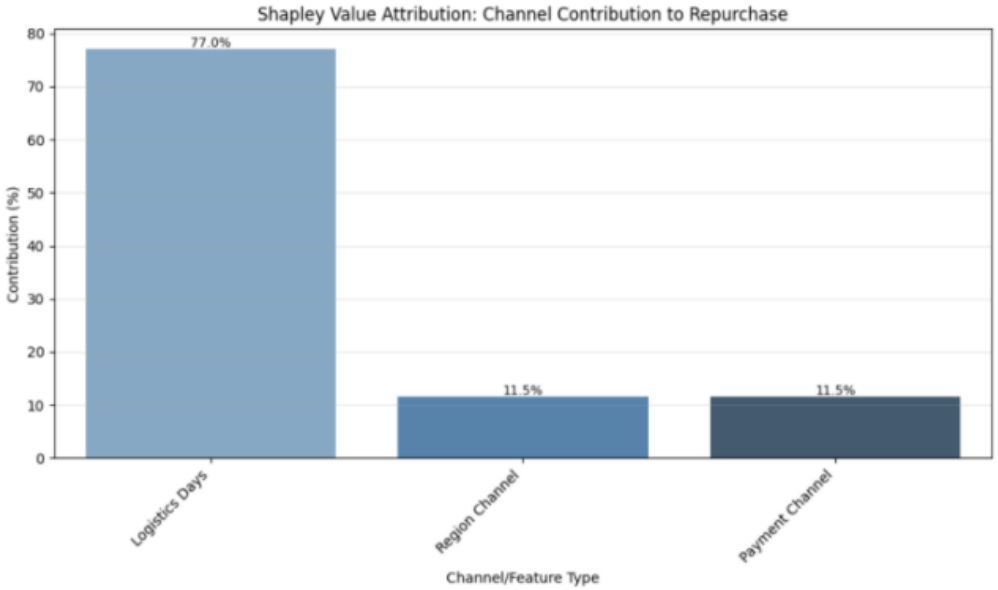



**Figure 3.** Bar chart of channel attribution based on shapley values (picture credit: original).

### 5.4. Recommendations for Channel Resource Allocation

Based on the above attribution results, it is recommended that e-commerce enterprises allocate 77% of their channel optimization budget to improving logistics timeliness, and stably control the average logistics timeliness within 3 days by optimizing warehouse layout and enhancing distribution efficiency. The remaining 23% of the budget should be evenly allocated to the regional and payment channels, so as to strengthen service coverage in the Southeast Region and optimize the credit card payment process.

### 5.5. Results of Propensity Score Matching Implementation

Before matching, the treatment group (i.e., users reached via the core channel) included 7,968 records and the control group 107,066 records, with the repurchase rate standing at 28.8% for the treatment group and 24.9% for the control group. After 1:1 nearest neighbor matching, 7,968 samples were retained for each group, and the SMD of all covariates was less than 0.1, meeting the balance requirement and eliminating systematic differences between the two groups.

### 5.6. ATT Results

The post-matching ATT calculation shows that core channel reach can increase the repurchase rate by 3.9 percentage points (28.8% for the treatment group, compared with 24.9% for the control group). The results of the t-test ($t=5.510$, $p<0.001$) indicate that the causal effect is highly significant.

### 5.7. Robustness Explanation of the Research Results

This study verified the robustness of the research results through two methods. First, channel attribution analysis was re-conducted after adjusting the core parameters of the Random Forest model, and the Shapley value contribution of the logistics channel still reached 88.0%, remaining the most contributory factor among all channels and being consistent with the core conclusions of the original model. Second, after adjusting the parameters of Propensity Score Matching, the average treatment effect of core channel reach remained at 3.9 percentage points and was significant at the 1% statistical level.

The results of the above two robustness tests show that the core conclusions—that the logistics channel is the core driving factor for repurchase and that core channel reach has a significant positive effect on repurchase—will not change with the adjustment of model and matching parameters, indicating that the research results have strong stability and reliability.

## 6. Conclusions

By constructing an integrated analytical framework of "repurchase prediction - channel attribution-causal verification" and conducting multi-dimensional empirical analysis based on actual e-commerce operation data, this study draws the core conclusions as follows: Logistics timeliness and the characteristics of users' consumption behavior jointly constitute the key factors influencing e-commerce users' repurchase behavior. Among them, the Random Forest model can more accurately capture the complex influencing mechanism of repurchase behavior and becomes the optimal model for repurchase prediction; the logistics channel is the core channel driving users' repurchase behavior, and its contribution to repurchase is higher than that of regional and payment channels, which confirms the core value of logistics experience in the multi-channel operation of e-commerce; in addition, the core channel combination consisting of the Southeast Region, credit card payment and logistics within 3 days has a significant positive causal effect on users' repurchase rate. This conclusion is reliable after robustness tests and also provides a clear practical direction for the collaborative optimization of multi-channels.

Merchants should take the optimization of logistics experience as the core starting point, increase investment in logistics resources, control the average logistics timeliness within 3 days, and allocate multi-channel resources scientifically, allocating 77% of the budget to logistics and 23% evenly to regional and payment channels. In addition, precise operation should be carried out based on core features, and customized logistics services should be provided for high-value users. For example, next-day delivery, exclusive distribution and other services can be offered to high-value users with a cumulative consumption amount of more than USD 1,596 to stimulate the conversion of potential repurchasing users.

Cross-sectional data is adopted, which fails to capture the dynamic changes of repurchase behavior; psychological characteristics such as user satisfaction and brand loyalty are not included, nor is the impact of emerging channels such as live-streaming e-commerce and social e-commerce taken into account; the division of core channel combinations does not consider more possibilities.

Longitudinal data will be used to track the dynamic evolution of repurchase behavior; the analytical dimensions of users' psychological characteristics and emerging channels will be expanded.

## References


[1] Chen, Y., Zhang, J., 2023. The impact of logistics service quality on e-commerce customer repurchase intention: The mediating role of customer satisfaction. Journal of Business Research. 156, 113289.

[2] Chou, P., Chuang, H. H.-C., Chou, Y.-C., Liang, T.-P., 2021. Predictive analytics for customer repurchase: Interdisciplinary integration of buy till you die modeling and machine learning. European Journal of Operational Research. 294 (3), 1043–1055.

[3] Csoban-Mirka, E., Esqueda-Henríquez, S., Ríos, A., 2024. Predicción del comportamiento de compra online: una aplicación del modelo S-O-R. Retos Revista de Ciencias de la Administración y Economía. 14 (27), 21–33.

[4] Kim, J., Ji, H., Oh, S., Hwang, S., Park, E., del Pobil, A. P., 2021. A deep hybrid learning model for customer repurchase behavior. Journal of Retailing and Consumer Services. 59, 102356.

[5] Li, M., Wang, Y., Zhang, L., 2022. Channel attribution in e-commerce: A comparative study of Shapley value and traditional methods. Electronic Commerce Research and Applications. 56, 101185.

[6] Liang, C., Xingqing, Y., Ma, J., 2021. Can e-commerce improve the profitability of agricultural co-operatives? Transformations in Business & Economics. 20 (2A), 684–700.

[7] Liu, J., Liu, M., Hong, L., Lin, Q., 2025. Improved Shapley Value with Trapezoidal Fuzzy Numbers and Its Application to the E-Commerce Logistics of the Forest Products. Applied Sciences. 15 (1), 123.

[8] Lu, T., Zhang, X., Zhuang, G., 2025. Managing multichannel conflict: the role of interactions enabled by interorganizational information system and public social media. Industrial Marketing Management. 108, 104123.

[9] Obayemi, A., Nguyen, K., 2024. Leveraging Multimodal Shapley Values to Address Multimodal Collapse and Improve Fine-grained E-Commerce Product Classification. Proceedings of the IEEE International Conference on Data Mining (ICDM). pp. 1–6.

[10] Olist., 2018. Olist Brazilian E-Commerce Public Dataset. Kaggle. Retrieved from https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce/data.

[11] Park, E., 2023. CRNet: a multimodal deep convolutional neural network for customer revisit prediction. Journal of Big Data. 10 (1), 15.

[12] Verma, A., Singh, P., Kumar, R., 2025. Enhancing customer repurchase prediction: Integrating classification algorithms with RFM analysis. Journal of Consumer Behaviour. 24 (2), 123–135.

[13] Wang, X., Yao, S., Wang, J., Li, Y., Yu, B., 2022. A Study of Cooperation Between Suppliers and E-Commerce Platforms Based on Biform Game. Journal of Theoretical and Applied Electronic Commerce Research. 17 (4), 1456–1475.

[14] Yılmaz Benk, G., Badur, B., Mardikyan, S., 2022. A New 360° Framework to Predict Customer Lifetime Value for Multi-Category E-Commerce Companies Using a Multi-Output Deep Neural Network and Explainable Artificial Intelligence. Information.

[15] Zhang, H., Hou, Y., He, F.-F., 2022. A Decision Method for Benefit Distribution Mechanism of Shared End Distribution on Shapley Value. Mathematical Problems in Engineering. pp. 1–11.

[16] Zhou, X., Chi, L., Li, J., Xing, L., Yang, L., Wu, J., Meng, H.,2024. A Study on Revenue Distribution of Chinese Agricultural E-Commerce Supply Chain Based on the Modified Shapley Value Method. Sustainability. 16 (20), 9023.