Journal of Fintech and Business Analysis Vol.3 Issue 3 EWA Publishing
Available Online: 22 July 2026 DOI: 10.54254/3049-5768/2026.35516

Identifying impulse buying via clickstream data: the role of decision conversion efficiency

Yijun Zhang

School of Engineering, Nanfang College Guangzhou, Guangdong, China

zhangyj230052@stu.nfu.edu.cn

Abstract. With the rapid development of e-commerce, impulse purchases have become a significant driver of revenue growth for e-commerce platforms. Traditional research often relies on questionnaire surveys for analysis, which are susceptible to memory bias and social desirability bias, and commonly conflate the number of items added to a shopping cart with purchase intent. Based on 778 valid purchase sessions from the Retail Rocket public dataset, this paper constructs the core metric of the "add-to-cart-to-view ratio" to identify impulse purchases from the perspective of behavioral conversion efficiency. It also introduces the average dwell time per product to avoid circular reasoning and compares the predictive performance of logistic regression, random forest, and XGBoost models. The results indicate that the "add-to-cart-to-view ratio" is a core predictive indicator of impulse purchases, while the number of times an item is added to the cart alone has no significant impact; the XGBoost model performed best (AUC = 0.7576). This study provides data support for e-commerce platforms to accurately identify impulse buyers and implement targeted marketing strategies.

Keywords: impulse purchases, cart-view ratio, XGBoost, clickstream data

1. Introduction

With the development of internet technology and the widespread adoption of e-commerce, online shopping has become a major form of consumption. Compared to in-store shopping, e-commerce influences consumer decision-making through mechanisms such as personalized recommendations, live-stream marketing, and limited-time promotions, making consumers more prone to impulse purchases. Existing research primarily relies on questionnaire surveys or experimental designs to analyze the mechanisms underlying impulse buying from the perspectives of personality traits, emotional states, self-control, and marketing stimuli. However, due to the influence of subjective biases, such studies struggle to accurately reflect actual shopping behavior. As e-commerce platforms accumulate vast amounts of clickstream data, users' browsing, wishlist-adding, and purchasing behavior patterns can objectively characterize shopping decision-making, providing a new data-driven approach for impulse buying research.

This paper uses the Retail Rocket public dataset to construct a framework for identifying impulse purchases. It extracts features such as browsing depth, adding-to-cart behavior, and temporal context, and builds logistic regression, random forest, and XGBoost models for prediction and comparison. The significance of this study lies in the following: theoretically, it combines behavioral data with psychological mechanisms to explain impulse purchases, providing new evidence for empirical research; practically, it helps optimize recommendation systems, precision marketing, and intelligent operational strategies. This study addresses two questions:

(1) How do user behavioral characteristics influence impulse purchases?

(2) Is it possible to build effective predictive models based on behavioral data?

The core innovation of this study lies in the introduction of the "add-to-cart-to-view ratio" metric, which corrects the traditional misconception that "number of items added to cart = purchase intent" by examining behavioral conversion efficiency.

2. Literature review

2.1. Current research on impulsive buying behavior

Impulse buying is a key topic in consumer behavior research, referring to unplanned, immediate purchasing behavior driven by external stimuli or emotions. Rook argued that impulse buying is characterized by immediacy and emotional drive [1]; Beatty and Ferrell further emphasized the significant role of the shopping environment and emotional state [2]. With the development of e-commerce, personalized recommendations, promotional activities, and social commerce have continuously intensified purchasing stimuli, making the phenomenon of online impulse buying more prevalent. In recent years, relevant research has begun to utilize clickstream data and machine learning methods to analyze consumer behavior; however, existing studies have largely focused on predicting purchase conversions, while research on the identification and prediction of impulse buying remains relatively limited.

2.2. A study on user behavior and e-commerce conversion

User behavior logs objectively record the entire process from browsing to purchase and are widely used in purchase conversion analysis. Research indicates that characteristics such as browsing depth, dwell time, click frequency, and adding-to-cart behavior are significantly correlated with purchasing decisions [3]. Among these, adding-to-cart behavior is regarded as a crucial intermediate step linking browsing and purchasing, while behavioral paths and decision-making efficiency reflect the extent of users' information search and their propensity to purchase. Although existing research has validated the value of behavioral data for purchase prediction, there remains significant room for expansion in studies that utilize real clickstream data to identify and predict impulse purchasing behavior [4].

2.3. Related theoretical foundation

This paper introduces the Stimulus–Organism–Response (SOR) model to explain the relationship between user behavioral characteristics and impulse buying [5]. Its basic logic can be summarized as follows: Stimulus → Organism → Response

This model posits that External Stimuli (Stimulus) influence final Behavioral Responses (Response) through an individual's Cognition and Emotions (Organism). In an e-commerce environment, product categories (category_level), purchase time (purchase_hour, weekday), and platform interaction stimuli constitute external stimuli; behavioral characteristics (cart_view_ratio, unique_items_viewed, avg_time_per_item) serve as objective representations of internal psychological states and are incorporated into the Organism (Organism) level; User purchase intent, information search intensity, decision-making efficiency, and cognitive engagement constitute the organism; impulse buying behavior serves as the final response. Although behavioral data cannot directly measure psychological states, it can indirectly reflect changes in cognitive engagement and purchase intent. Based on this framework, this paper constructs an impulse buying prediction model using real clickstream data to analyze the impact of different stimuli on impulse buying.

3. Research methods and data sources

3.1. Data sources and notes

This paper conducts an empirical analysis using the Retail Rocket e-commerce dataset published on Kaggle [6], which records users' real-world behavioral trajectories, including browsing, adding to cart, and purchasing. To ensure label accuracy, the original data was rigorously filtered: only sessions containing complete purchase behaviors were retained, while records with abnormal timestamps or no substantive interactions were excluded. This resulted in 778 valid samples, of which impulse purchase samples accounted for 20.82%. To mitigate the impact of class imbalance on model training, the Synthetic Minority Over-sampling Technique (SMOTE) method was employed to oversample the minority class [7].

In terms of research methodology, this paper constructs logistic regression, random forest, and XGBoost models for comparison. Among these, XGBoost was selected as the core predictive model because it can effectively handle complex nonlinear relationships and provides feature importance analysis.

3.2. Variable definitions

3.2.1. Dependent variable (impulse buying behavior)

This paper uses impulsive purchasing as the dependent variable. Drawing on clickstream data, it measures the tendency toward impulsive purchasing by examining the time interval between a user's last page view and their purchase: $view\_to\_purchase\_sec = purchase\_time - view\_time$.

Based on this, this paper defines impulse buying as follows:

If a user completes a purchase within 60 seconds of viewing a product for the last time, or makes a purchase without having viewed the product beforehand, that is: $view\_to\_purchase\_sec < 60$.

it is defined as an impulse purchase (impulsive_purchase = 1); otherwise, it is defined as a non-impulse purchase (impulsive_purchase = 0).

3.2.2. Independent variable (user behavior characteristics)

This paper constructs three types of features—behavioral interactions, temporal contexts, and product attributes—based on user behavior logs.

(1) cart_view_ratio

Used to measure the conversion efficiency from browsing to adding to cart; the formula (1) is:

$$cart\_view\_ratio = add\_to\_cart\_count \div (view\_count + 1) \tag{1}$$

The higher this metric is, the more likely users are to add items to their cart after minimal browsing, indicating that they make decisions more quickly and are more prone to impulse purchases.

(2) avg_time_per_item

Used to measure the average time users spend processing information about a single product (formula (2)):

$$avg\_time\_per\_item = session\_duration \div (unique\_items\_viewed + 1) \tag{2}$$

In this context, $unique\_items\_viewed$ represents the number of distinct products a user views. A lower value indicates that the user processes information about individual products more quickly and superficially. For example, a consumer might spend as long as two hours in a single session but view 100 products, spending an average of just one second on each before completing a purchase; conversely, a higher value suggests that the user has thoroughly compared information and made a well-considered decision. By distinguishing between decision-making speed (time spent on a single item) and overall session duration (total session length), this metric operationalizes the theoretical perspective that "impulse purchases are driven by rapid decision-making rather than brief total browsing time", thereby providing a more theoretically grounded and granular explanatory variable.

In addition, this paper introduces features such as the number of unique items viewed (unique_items_viewed), purchase time (purchase_hour), weekday (weekday), and product category level (category_level) to characterize differences in user browsing depth, temporal context, and product attributes.

3.3. Research methods

This paper constructs three types of models for comparative analysis:

A. Explanatory logistic regression (statsmodels.Logit, no regularization): Used to analyze the statistical significance and direction of influence of variables, incorporating three core variables: add-to-cart-to-view ratio, number of views, and number of adds to cart;

B. Baseline logistic regression (sklearn.LogisticRegression, L2 regularization): Serves as a benchmark for comparison with machine learning models;

C. XGBoost and Random Forest: To capture nonlinear relationships among variables, with XGBoost serving as the primary predictive model [8].

All models use SMOTE oversampling to address class imbalance, and 5-fold stratified cross-validation is employed to ensure the stability of the results. AUC and recall are prioritized as evaluation metrics to more accurately measure the models' ability to identify the minority class (impulse purchases).

4. Empirical analysis

4.1. Analysis of user behavior characteristics

As shown in Figure 1, impulse buyers account for only 20.82%. Compared to non-impulse buyers, impulse buyers view fewer pages and fewer products, have a higher ratio of items added to cart to items viewed, and make decisions more quickly.

As shown in Figure 2, the correlation heatmap indicates that there is no severe multicollinearity among the variables.

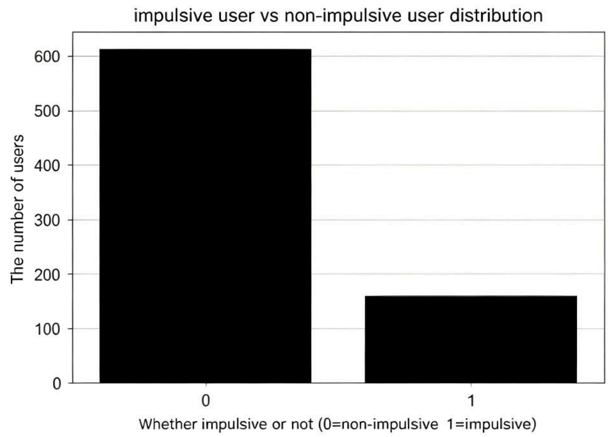

Figure 1. Distribution of impulsive and non-impulsive purchases

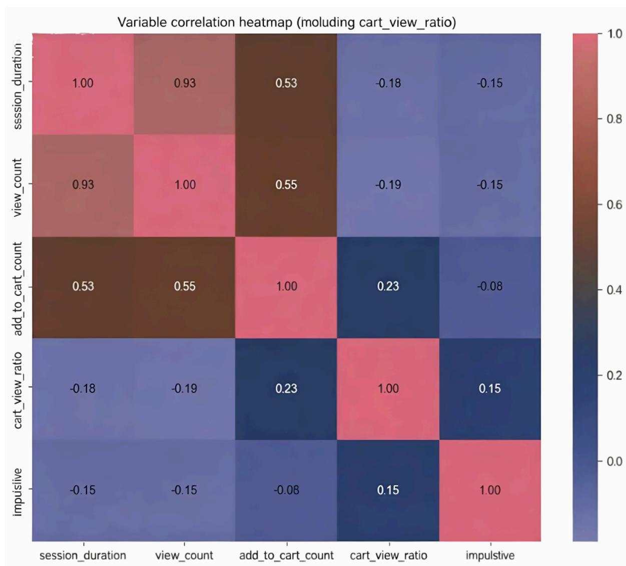

Figure 2. Heatmap of correlations in user behavior characteristics

4.2. Analysis of factors influencing impulse purchases

4.2.1. Analysis of the complete logistic regression model

Logistic regression (statsmodels Logit, no regularization, n = 544) was used to analyze the influencing factors. The final model included view_count, add_to_cart_count, and cart_view_ratio. The results show that cart_view_ratio is significantly positively correlated (Coef = 1.045, OR = 2.84, p = 0.0075), view_count is marginally negatively correlated (Coef = -0.037, p = 0.0585), and add_to_cart_count is not significant (p = 0.8577). This indicates that the cart-to-view ratio is the core predictive indicator of impulse purchases, while the number of times an item is added to the cart alone has no explanatory power. This finding overturns conventional wisdom, suggesting that "interaction frequency" is not synonymous with impulsivity; conversely, the cart-to-view ratio (cart_view_ratio) exhibits extremely high significance (p < 0.01). This indicates that the key factor determining impulse purchases lies not in "how many times a user adds items to the cart", but rather in "how efficiently the user converts views into cart additions".

Behavioral explanation: The essence of impulse buying is the "need for cognitive closure" [9]. A high cart_view_ratio indicates that users have made a commitment quickly with minimal information-searching effort, which is a classic example of the "what you see is what you get" mentality; conversely, a low ratio suggests that users are engaging in rational evaluation of options. The results of the explanatory comprehensive logistic regression model are shown in Table 1.

Table 1. Regression results from an explanatory full logistic regression model (statsmodels, MLE without regularization)

Statsmodels Summary (*p*-values)						
						Results: Logit
==						
Model:	Logit		Method:		MLE	
Dependent Variable:	impulsive		Pseudo R-squared:		0.051	
Date:	2026-05-26 21:55		AIC:		535.2980	
No. Observations:	544		BIC:		552.4938	
Df Model:	3		Log-Likelihood:		-263.65	
Df Residuals:	540		LL-Null:		-277.94	
Converged:	1.0000		LLR *p*-value:		2.7387e-06	
No. Iterations:	7.0000		Scale:		1.0000	
--						
Coef. Std.Err.	z		$p > \|z\|$		[0.025 0.975]	
--						
const	-1.4099	0.1901	-7.4150	0.0000	-1.7826	-1.0372
view_count	-0.0370	0.0196	-1.8922	0.0585	-0.0753	0.0013
add_to_cart_count	-0.0113	0.0632	-0.1794	0.8577	-0.1353	0.1126
cart_view_ratio	1.0452	0.3908	2.6745	0.0075	0.2792	1.8112
==						
p-values:						
const	1.216404e-13					
view_count	5.846396e-02					
add_to_cart_count	8.576563e-01					
cart_view_ratio	7.485070e-03					
dtype:	float64					

4.2.2. Simplifying logistic regression model analysis

Table 2. The results of the simplified model

Statsmodels Summary (Simplified Model)						
						Results: Logit
==						
Model:	Logit		Method:		MLE	
Dependent Variable:	impulsive		Pseudo R-squared:		0.028	
Date:	2026-05-26 21:56		AIC:		544.0932	
No. Observations:	544		BIC:		552.6911	
Df Model:	1		Log-Likelihood:		-270.05	
Df Residuals:	542		LL-Null:		-277.94	
Converged:	1.0000		LLR *p*-value:		7.0821e-05	
No. Iterations:	6.0000		Scale:		1.0000	
--						
Coef. Std.Err.	z		*p* > |z|		[0.025 0.975]	
--						
const	-1.7778	0.1641	-10.8364	0.0000	-2.0994	-1.4563
cart_view_ratio	1.2533	0.3335	3.7584	0.0002	0.5997	1.9069
==						
p-values						
const	2.314432e-27					
cart_view_ratio	1.709777e-04					
dtype:	float64					

To test for robustness, we constructed a simplified model containing only cart_view_ratio, with a coefficient of 1.253 (OR = 3.50, $p = 0.0002$). A high cart_view_ratio indicates that users make quick decisions with minimal information-searching effort, reflecting a need for cognitive closure. The results of the simplified model are shown in Table 2.

4.3. Analysis of prediction model results

4.3.1. Comparison of model prediction performance

To evaluate the predictive capabilities of different models for impulse buying behavior, this paper constructed Logistic Regression, Random Forest, and XGBoost models, and used accuracy, recall, F1 score, and AUC as evaluation metrics. The results of the performance comparison between the different models are shown in Table 3.

Table 3. Comparison of model performance

Model	Accuracy	Recall	F1	AUC
Logistic Regression	0.56	0.683	0.582	0.53
Random Forest	0.805	0.721	0.643	0.7482
XGBoost	0.812	0.749	0.678	0.7576

The results show that all three models are capable of identifying users' impulse buying behavior to some extent, but there are significant differences in their overall performance.

Logistic regression serves as a simple baseline model, but its predictive performance is relatively limited. In contrast, the random forest model effectively captures the nonlinear relationships in user behavior data by combining multiple decision trees, resulting in an AUC of 0.7482 and a recall of 0.721, indicating that the model is capable of identifying more genuine impulse buyers.

Among all models, XGBoost performed the best. It achieved an accuracy of 81.2%, a recall of 74.9%, an F1 score of 0.678, and an AUC of 0.7576, outperforming both the logistic regression and random forest models across all metrics. The results indicate that XGBoost is more effective at learning the complex relationships among user behavior variables and demonstrates stronger predictive power for impulse buying behavior. A comparison of the ROC curves for the three models is shown in Figure 3.

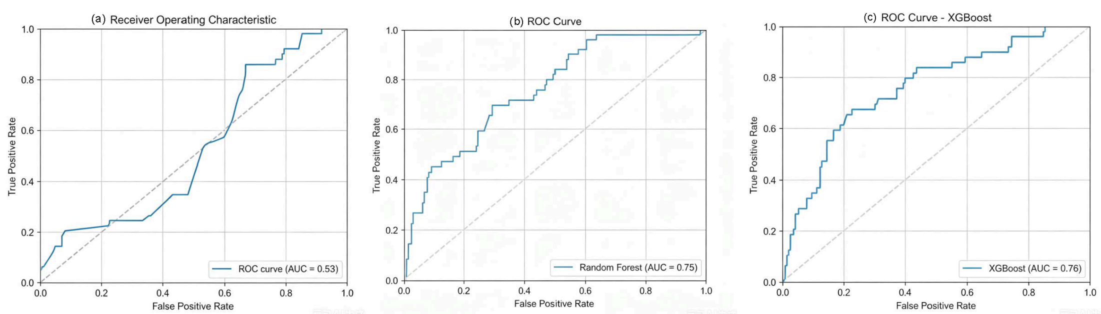

Figure 3. The comparison of the ROC curves for the three models. (a) Logistic Regression ROC curve; (b) Random Forest ROC curve; (c) XGBoost ROC curve

As shown by the ROC curve, the XGBoost curve generally lies above those of the other models and has the largest area under the curve, further validating its superior classification performance.

4.3.2. Analysis of threshold optimization results

Since the positive samples account for only 20.82%, we optimized the classification thresholds for Random Forest and XGBoost; the optimal thresholds were 0.35 and 0.32, respectively (logistic regression remained at 0.5). Lowering the threshold improves recall, which better aligns with the e-commerce platform's business requirement to "reduce the number of impulsive users who slip through the cracks".

4.3.3. Feature importance analysis

To gain a deeper understanding of the model's predictive mechanisms, this paper analyzes the feature importance results from the XGBoost model, as shown in Figure 4.

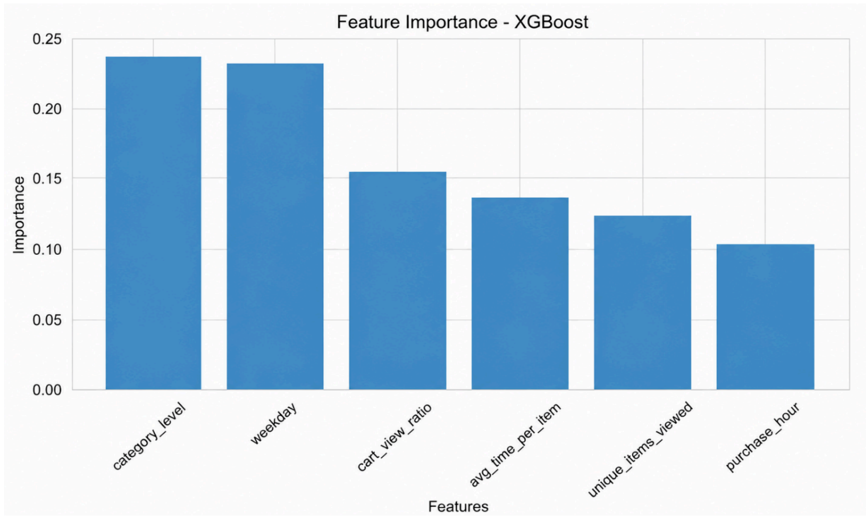

Figure 4. XGBoost feature importance ranking chart

The results indicate that there are significant differences in the extent to which different behavioral variables influence impulse buying.

In terms of feature importance ranking, 'cart_view_ratio' is the most important variable and serves as the core predictor of impulse buying behavior, followed by 'unique_items_viewed', indicating that the number of items viewed by a user makes a significant contribution to identifying impulse buying. In addition, temporal features such as 'purchase_hour' and 'weekday' also demonstrate a certain level of importance, suggesting that user purchasing behavior may be influenced by temporal contextual factors. In contrast, 'avg_time_per_item' has relatively lower importance but still contributes to the model's predictions.

Overall, behavioral conversion features (cart_view_ratio), browsing depth features (unique_items_viewed), and temporal features (purchase_hour, weekday) played a relatively significant role in predicting impulse purchases. However, there were marked differences in the importance of these features, indicating that impulse purchasing behavior is influenced by a combination of multidimensional user behavior factors.

4.3.4. Summary of this chapter

The comprehensive comparison results indicate that machine learning models outperform traditional logistic regression models overall. Among them, XGBoost achieved the best performance across metrics such as accuracy, recall, F1 score, and AUC, demonstrating greater effectiveness in identifying impulse buying behavior.

At the same time, feature importance analysis reveals that users' browsing-to-purchase conversion efficiency, browsing breadth, and temporal context are key factors influencing impulse buying behavior. This indicates that clickstream data can effectively capture the consumer decision-making process, providing data support and a basis for decision-making to help e-commerce platforms identify potential impulse buyers.

5. Results and discussion

5.1. Key findings

Based on the Retail Rocket dataset, this paper analyzes the key factors influencing impulse purchases in e-commerce from the perspective of behavioral characteristics.

Research shows that there is a significant negative correlation between the number of products viewed and impulse buying: the more products users view, the more thoroughly they search for and compare information, and the more rational their decision-making becomes. Non-impulse buyers follow the standard "exploration—evaluation—decision" path, while impulse buyers skip the stage of thorough information search and commit to a purchase immediately.

The "Cart-view ratio" is the core "behavioral fingerprint" of impulse buying: when users quickly add items to their cart after minimal browsing, the probability of impulse buying increases significantly. For these users, the decision-making mind operates in "immediate response" mode; adding an item to the cart is not a "pause button" for further consideration, but rather an impulsive outlet to "reserve it first". This finding validates the mechanism in the SOR model where "stimuli directly trigger responses, weakening cognitive mediation", and also reflects impulsive users' strong tendency toward instant gratification and their need for cognitive closure—they tend to quickly conclude decisions rather than engage in prolonged deliberation.

The time context plays a moderating role in impulse buying: users' self-control tends to decline late at night or on weekends, making them more likely to seek immediate gratification. The explanatory power of the average dwell time per item is limited, as it may be influenced by behaviors such as "browsing quickly and then returning to make a purchase" and variations in the amount of information on product detail pages.

A comparison of the models shows that XGBoost performed best (AUC = 0.7576), indicating that impulse buying exhibits strong nonlinear characteristics and requires the integration of multidimensional behavioral features for accurate identification.

5.2. Comparison with previous research

Most existing studies on impulse buying rely on questionnaire surveys and focus on psychological factors such as emotions and promotions, making them susceptible to subjective bias. Based on real clickstream data, this paper validates the classic characteristics of impulse buying—namely, "low-information search and rapid decision-making" [10]. Unlike previous research, this study reveals the core predictive value of the "add-to-cart-to-view ratio" from the perspective of behavioral conversion efficiency and demonstrates the superiority of XGBoost in identifying impulse buying.

5.3. Practical recommendations

Based on the findings of this study, the following recommendations are proposed.

First, e-commerce platforms can develop a "real-time decision-making pattern recognition system". When the system detects a significant increase in a user's "add-to-cart-to-view ratio" coupled with a short dwell time on individual products, it determines that the user is in an "impulse-prone state". For such users, the recommendation system should prioritize displaying immediate incentives—such as limited-time offers and one-click purchase options—rather than extensive product specifications and comparison information, in order to align with the user's psychological need for rapid decision-making.

Second, for rational users who conduct in-depth browsing and have longer decision-making cycles, the platform should provide product comparison tools, user reviews, and decision-support information to help them reduce search costs and facilitate the transition from "continuous comparison" to "purchase execution".

Third, by integrating the XGBoost prediction model into the recommendation system, we can achieve real-time prediction of user purchasing tendencies and implement dynamic interventions. The value of real-time prediction lies not only in identifying "who the impulse buyers are", but also in capturing "when users are in a state of impulse-prone vulnerability", thereby enabling a precise alignment of strategies with users' current mental states.

6. Conclusion

Based on clickstream data from the Retail Rocket e-commerce platform, this paper constructs a framework for identifying and predicting impulse purchases, analyzes the influence of user behavioral characteristics, and compares the performance of three types of models. The study found that the add-to-cart-to-view ratio is the core predictive indicator for impulse purchases, while the number of add-to-cart actions alone has no significant explanatory power. The number of products viewed is negatively correlated with impulse purchases, and the time context plays a moderating role. The XGBoost model performed best (AUC = 0.7576), validating the strong nonlinear characteristics of impulse purchases.

This study provides additional empirical evidence on impulse buying from the perspective of behavioral conversion efficiency, offering insights for precision marketing and intelligent recommendations on e-commerce platforms. Due to limitations in sample size and the use of a single general-purpose e-commerce platform, the generalizability of these conclusions remains to be verified; future research could explore cross-platform studies incorporating scenarios such as live-streaming e-commerce and vertical e-commerce.

References

[1] Rook, D. W. (1987). The buying impulse. *Journal of Consumer Research, 14*(2), 189–199.

[2] Beatty, S. E., & Ferrell, M. E. (1998). Impulse buying: Modeling its precursors. *Journal of Retailing, 74*(2), 169–191.

[3] Verhagen, T., & Van Dolen, W. (2011). The influence of online store beliefs on consumer online impulse buying: A model and empirical application. *Information & Management, 48*(8), 320–327.

[4] Montgomery, A. L., Li, S., Srinivasan, K., & Liechty, J. C. (2019). Modeling online clickstream data: A review and future directions. *Marketing Science, 38*(4), 537–554.

[5] Mehrabian, A., & Russell, J. A. (1974). *An approach to environmental psychology*. MIT Press.

[6] Retail Rocket. (2016). *Retail Rocket e-commerce dataset*. Kaggle. https://www.kaggle.com/retailrocket/ecommerce-dataset

[7] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

[8] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM.

[9] Kruglanski, A. W. (1989). The psychology of being "right": The problem of accuracy in social perception and cognition. *Psychological Bulletin, 106*(3), 395–409.

[10] Akram, U., Hui, P., Khan, M. K., & Hashim, M. (2018). Factors affecting online impulse buying: Evidence from Chinese social commerce environment. *Sustainability, 10*(3), 721.