

Digital and Intelligent Analysis and Operation Strategies for Purchase Conversion in Social E-Commerce Under the Intelligent Internet of Everything

Qifei Zhang^{1*}, Lingyan Gu¹, Zijun Qiang¹

¹ Department of Statistics, School of Management, Fudan University, Shanghai, China
Undergraduate Student

*Correspondence: Qifei Zhang

Received: 22 May 2026; Accepted: 19 June 2026; Published: 3 August 2026

Abstract:

Social e-commerce suffers from high traffic but low conversion in the digital economy era. This study aims to reveal purchase conversion patterns and improve operational efficiency using data analytics. Based on 100,000 user behavior records from a social e-commerce platform, we adopt K-Means clustering, Apriori association rule mining, and Random Forest to build an integrated analysis framework. Data preprocessing includes missing value filling, outlier correction, and SMOTE sampling. Empirical results show that the overall conversion rate is 44.98%. Three user groups are identified: discount-sensitive, content-interactive, and general users. The combination of like, add to cart, and claim coupon has the highest conversion confidence of 86.76%. The Random Forest model achieves an AUC of 0.7743, showing strong discrimination ability. This study provides data-driven strategies for user segmentation, conversion path optimization, and content operation. The findings help platforms allocate marketing resources efficiently and enhance commercial performance.

Keywords:

Social E-commerce, Purchase Conversion, User Segmentation, Random Forest, Association Rule Mining, Internet of Everything

1. Introduction

The Internet of Everything accelerates the digital transformation of social e-commerce. Social interaction, content communication, and online transactions are deeply integrated, forming a new consumption model. However, most platforms face low conversion rates and inefficient marketing due to unclear user characteristics and lack of data-driven decisions. Traditional operation strategies rely on experience rather than quantitative analysis, leading to resource waste and low return on investment.

Existing studies focus on single aspects such as user behavior or marketing strategies, lacking a systematic framework for conversion analysis. Few studies combine user segmentation, association rule mining, and machine learning prediction

to solve practical industrial problems. To fill this gap, this study uses large-scale real data to explore conversion mechanisms, identify key influencing factors, and build a prediction model.

The contributions of this study are twofold. First, it constructs a comprehensive analytical framework for social e-commerce conversion, integrating multiple data mining methods. Second, it provides actionable operational strategies to improve conversion efficiency and platform performance. This research offers a replicable paradigm for data-driven business analysis.

2. Materials and Methods

2.1. Data Description

This study uses 100,000 user behavior records from a mainstream social e-commerce platform. The dataset includes 31 features and 1 binary target variable (purchase or non-purchase). The sample covers demographic attributes, interaction behaviors, commodity information, and marketing incentives. The training set and test set are divided at a ratio of 7:3.

2.2. Data Preprocessing

Outliers are detected by 3σ criterion and boxplot. SMOTE algorithm is used to address sample imbalance. Feature engineering includes encoding, standardization, and multicollinearity test via VIF.

2.3. Analysis Methods

·K-Means Clustering: Used for user segmentation. The optimal cluster number is determined by elbow method and silhouette coefficient.

·Apriori Algorithm: Applied to mine association rules between user behaviors and purchase conversion.

·Random Forest: Used to build conversion prediction model. Hyperparameters are optimized by grid search. Model performance is evaluated by accuracy, precision, recall, F1-score, and AUC.

3. Results and Discussion

3.1. Basic Conversion Characteristics

The overall purchase conversion rate of the dataset reaches 44.98%, with a 95% confidence interval of [44.72%, 45.24%], which is significantly higher than the industry average level. Statistical tests indicate that user level, discount intensity, and social interaction depth are significantly positively correlated with conversion performance. Price, likes, comments, shares and other interactive variables present a significant right-skewed distribution, which verifies the existence of the blockbuster product effect in social e-commerce. (Table 1)

Table 1. Overview of the dataset and conversion performance.

Statistical Item	Value
Total Samples	100000
Purchased Samples	44983
Unpurchased Samples	55017

Statistical Item	Value
Purchase Conversion Rate	44.98%

This Figure 1 shows the average purchase conversion rate corresponding to social interaction indicators such as likes, comments, shares and commodity-related variables, reflecting the sensitivity of conversion to different behaviors.

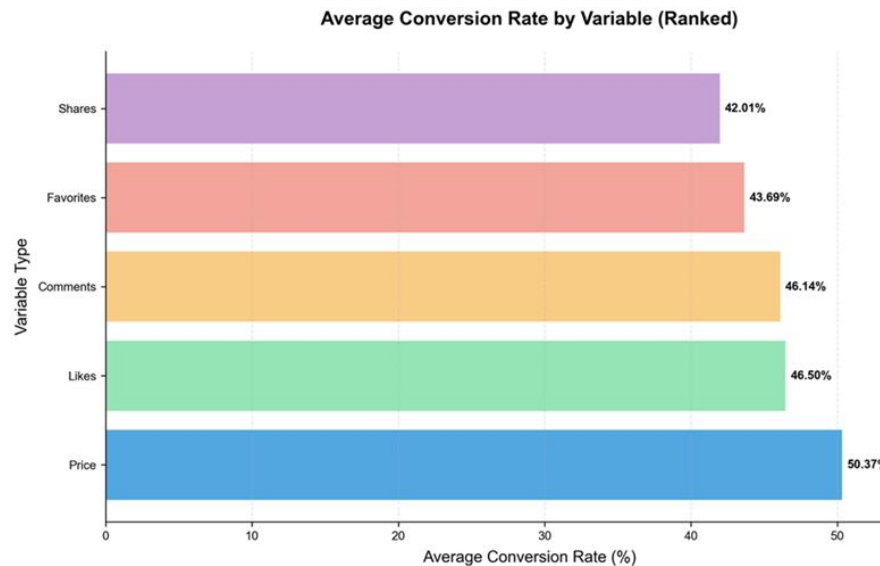


Figure 1. Average conversion rate across key behavioral variables.

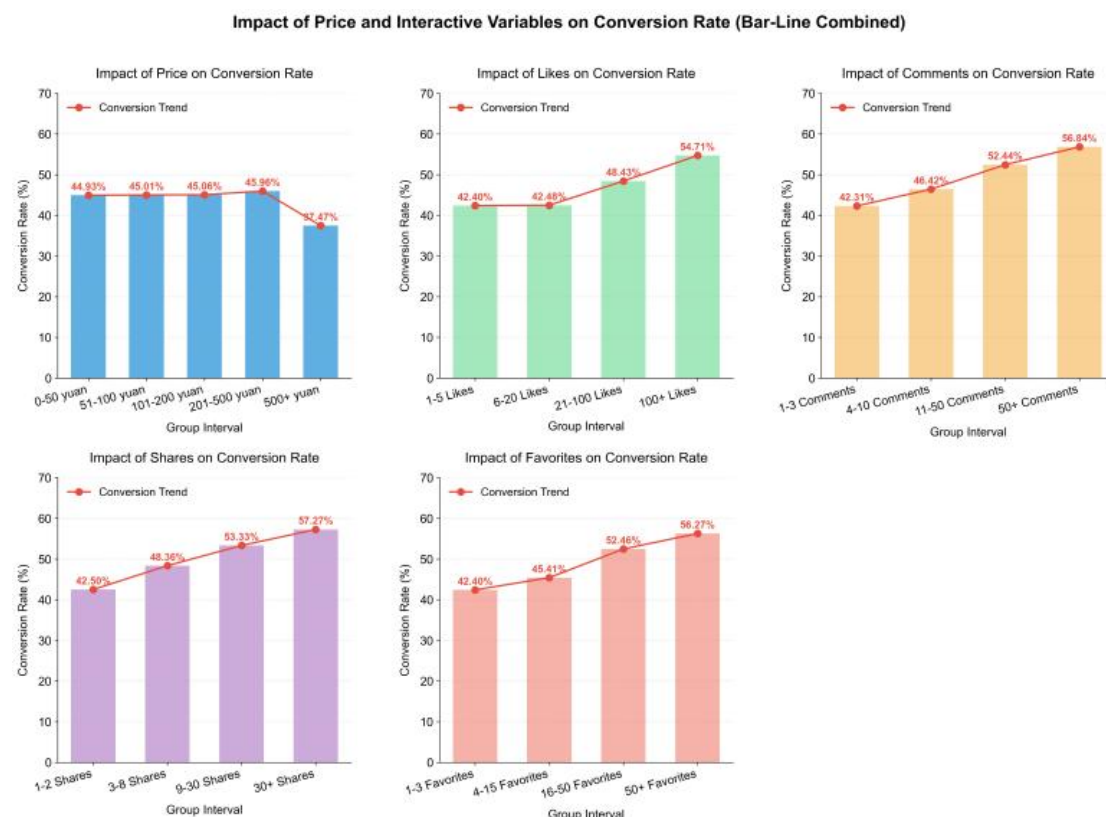


Figure 2. Impact of variable intensity on purchase conversion rate.

The combined bar-line plots illustrate conversion rate changes across tiered intervals of price and interactive variables. For likes, comments, shares, and favorites, conversion rises steadily with higher interaction intensity, peaking at 54.71%, 56.84%,

57.27%, and 56.27% respectively in the highest tier. Price follows an inverted U-shaped trend, with conversion peaking at 45.96% in the 201–500 yuan range and declining to 37.47% for products over 500 yuan.

Right-Skewed Distribution of Price & Interaction Variables (Strong Right-Skew)

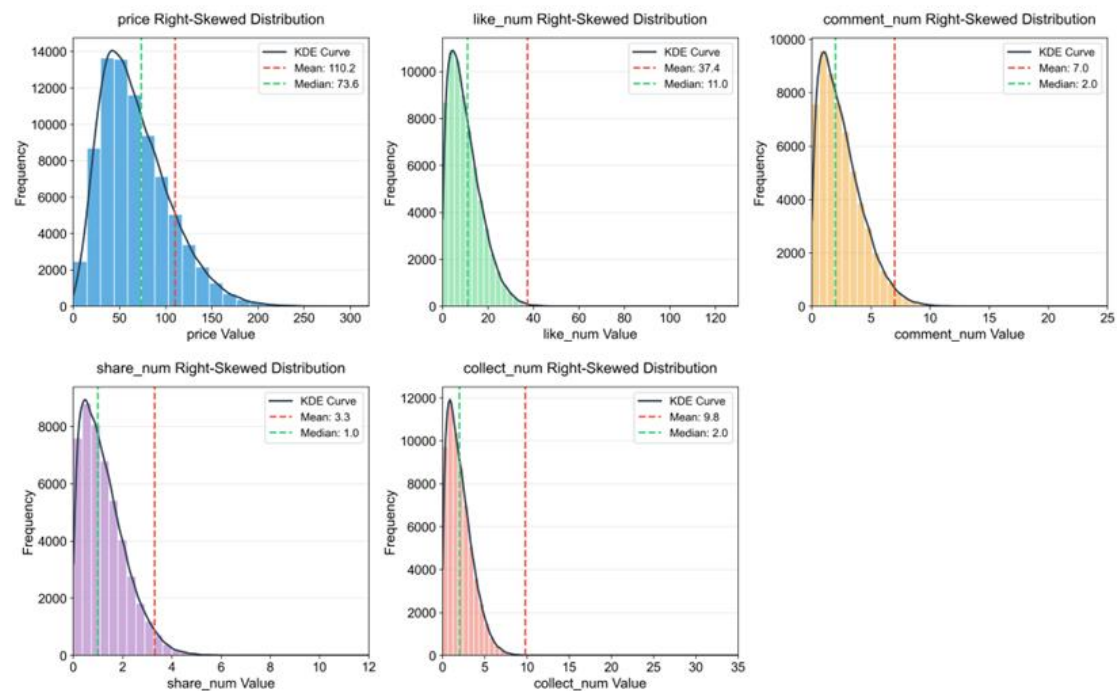


Figure 3. Right-skewed distribution of price and interactive variables.

The kernel density plots demonstrate that price, likes, comments, shares, and collection counts are all right-skewed with heavy tails, reflecting the “blockbuster effect” common in social e-commerce behavior.

3.2. User Segmentation Results

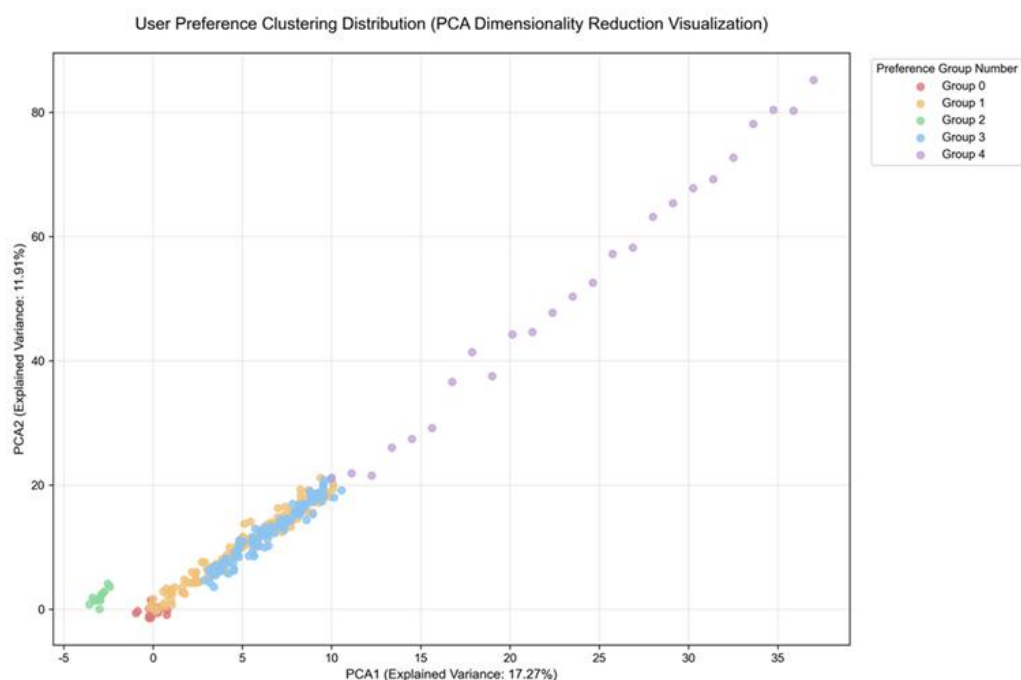


Figure 4. PCA visualization of user clustering results.

Based on the K-Means clustering algorithm, users are divided into three core groups with significant behavioral differences: discount-sensitive users (10.75%), content-interactive users (59.38%), and general users (29.86%). The optimal number of clusters is determined by the elbow method, and the validity of clustering is verified by PCA dimensionality reduction visualization.

The two-dimensional projection of clustered user data shows well-defined boundaries between the three identified groups, with distinct clusters for discount-sensitive, content-interactive, and general users.

Table 2. User stratification and targeted operational strategies.

User Group	Proportion	Conversion Rate	Operational Priority	Core Actions
Discount-Sensitive	10.75%	72.70%	Highest	Targeted coupons & cart reminders
Content-Interactive	59.38%	50.00%	High	Content push & interaction activation
General Users	29.86%	42.30%	Medium	Small coupons & lightweight content
Low-Potential	—	12.70%	Low	Reduce unnecessary pushes

3.3. Association Rule Mining Results

The Apriori algorithm mined high-confidence conversion rules. The combination {Like $\cap$ Add to Cart $\cap$ Claim Coupon} $\rightarrow$ Purchase achieves the highest confidence of 86.76% and lift of 1.93.

The chart ranks rules by confidence, with the combination of interaction, cart addition, and coupon claiming as the most effective driver.

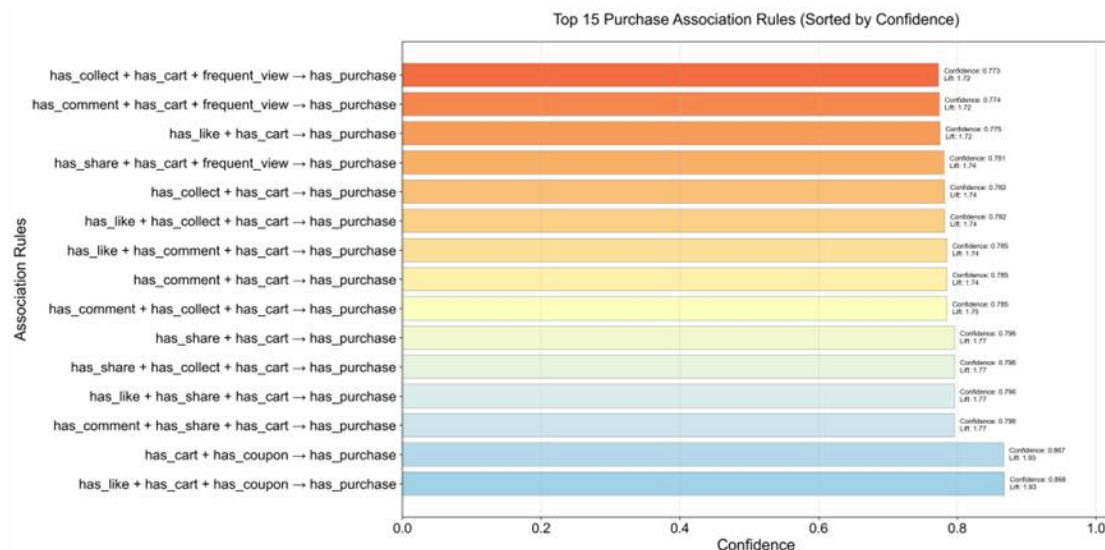


Figure 5. Top 15 association rules for purchase conversion.

3.4. Prediction Model Performance

To quantitatively identify user purchase propensity and support refined operational decision-making, this study constructs a purchase conversion prediction model based

on the Random Forest algorithm. The model is optimized for social e-commerce's imbalanced data distribution and high-dimensional behavioral features, with hyperparameters tuned via grid search to balance fitting performance and generalization ability. The key parameters are set as follows: tree depth = 8, number of decision trees = 100, and minimum sample split = 20. This configuration effectively avoids overfitting and ensures stable prediction results in real business scenarios.

Feature engineering is conducted to extract high-information-value variables, and feature importance is calculated to quantify the contribution of each indicator to conversion prediction. The top 10 influential features are screened out, among which the top five core variables account for more than 70% of the total importance weight, forming a highly discriminative feature system for purchase conversion.

Table 3. Top 10 feature importance for purchase conversion.

Feature Rank	Feature Name	Importance	Statistical Interpretation	Business Connection
1	Add-to-Cart Behavior	0.247	Strongest predictor	Key conversion action
2	Purchase Intent Score	0.215	Comprehensive intent indicator	User value pre-judgment
3	Coupon Usage	0.109	Price incentive signal	Promotion strategy adaptation
4	Follow Author	0.087	Trust driver factor	Core of trust marketing
5	Browsing Intensity	0.068	Interest depth indicator	Basis for content matching
6	Discount Rate	0.052	Price sensitivity signal	Pricing strategy reference
7	Coupon Claiming Behavior	0.049	Indirect reflection of intent	Basis for promotion reach
8	Number of Comments	0.038	Deep interaction indicator	Content quality evaluation
9	Number of Shares	0.032	Social transmission signal	Viral potential prediction
10	User Level	0.029	Trust endorsement indicator	Basis for customer stratification



Figure 6. Top 10 feature importance word cloud.

The word cloud intuitively presents the contribution weight of core variables, highlighting add-to-cart behavior and purchase intent as the dominant driving factors for conversion prediction.

The prediction model is evaluated using a comprehensive set of metrics to verify its effectiveness in distinguishing purchasing and non-purchasing users. All indicators demonstrate that the model achieves reliable performance in the imbalanced sample scenario of social e-commerce.

Table 4. Performance metrics of the Random Forest model.

Evaluation Metric	Value	Data Interpretation	Practical Significance
Accuracy	0.7164	The overall prediction accuracy of the model is 71.64%, showing high fault tolerance in the imbalanced sample scenarios of social e-commerce.	Reduces marketing resource waste and improves the effectiveness of strategy coverage.
Precision	0.6996	Among users predicted to “purchase”, actual purchases account for 69.96%; the precision meets commercial application requirements.	Reduces invalid reach and improves marketing ROI.
Recall	0.6474	Among actual “purchasing” users, 64.74% are identified by the model, indicating good coverage.	Avoids missing high-potential users and maximizes the capture of conversion opportunities.
F1 Score	0.6725	The harmonic mean of precision and recall, comprehensively reflecting the model’s classification performance.	Balances the core demands of “precise reach” and “comprehensive coverage”.
AUC Score	0.7743	Area under the ROC curve; the model has a strong ability to distinguish between “purchase” and “non-purchase” users.	Provides a scientific basis for stratifying high-potential users and supports refined operations.
Specificity	0.7727	The probability of correctly identifying “non-purchase” users, effectively filtering out low-value groups.	Reduces disturbance to low-potential users and improves overall user experience.

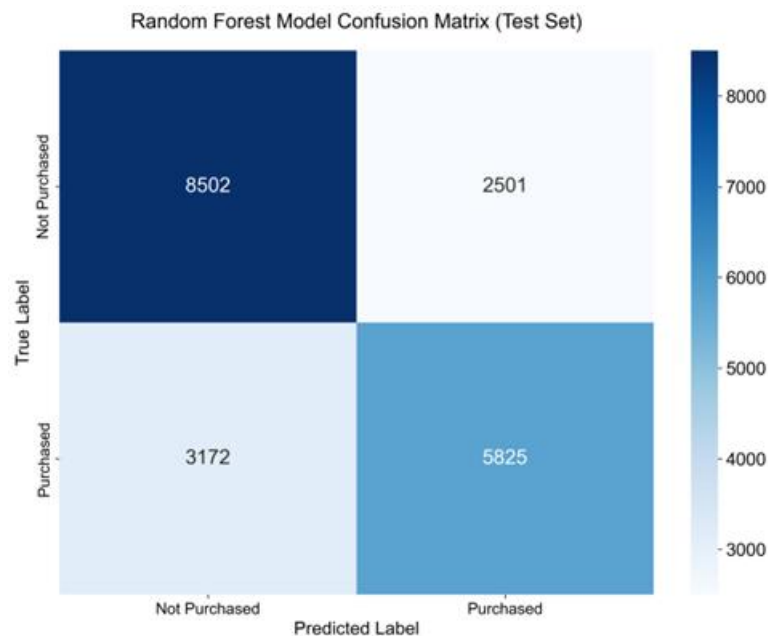


Figure 7. Confusion matrix of the Random Forest model.

The matrix presents the classification performance on the test set, showing stable identification ability for both purchasing and non-purchasing users with low misclassification rate.

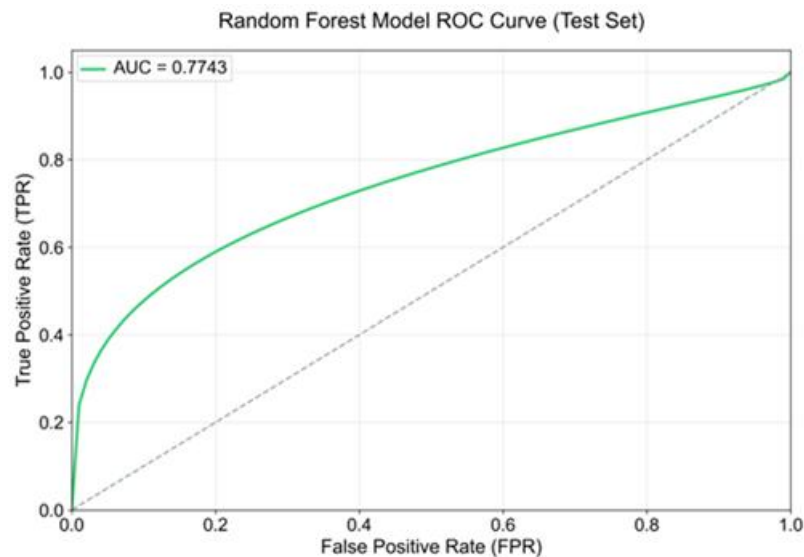


Figure 8. ROC curve of the Random Forest model.

The Random Forest model achieves an AUC of 0.7743, indicating moderate discriminatory power to distinguish purchasing and non-purchasing users. According to mainstream machine learning evaluation standards, AUC ranging from 0.7 to 0.85 represents moderate classification performance, while values above 0.85 are defined as strong discriminative ability.

Based on the predicted purchase probability, users are divided into four tiers with distinct operational priorities. This stratification mechanism enables platforms to allocate marketing resources rationally and implement targeted intervention strategies, thereby improving conversion efficiency and marketing ROI. The probability threshold, actual conversion rate, and core operational actions for each tier are formulated according to empirical results and business practicality.

Table 5. User stratification based on prediction probability and operational strategies.

User Stratification	Probability Threshold	Actual Conversion Rate	Corresponding User Group	Operational Priority	Core Operational Actions
Extremely High-Potential Users	>0.8	89.30%	Discount Sensitive core users	Highest	Simplify the order process and provide membership services (priority shipping, exclusive customer service), no additional price incentives needed.
High-Potential Hesitant Users	0.6-0.8	76.50%	Content Interactive main users	High	Push time-limited targeted coupons (discount $\geq 20\%$), activate the "Add to Cart + Claim Coupon" association rule, and strengthen decision-making momentum.
Potential Cultivation Users	0.3-0.6	42.30%	Ordinary User potential group	Medium	Issue small general coupons to guide add-to-cart behavior, push high-quality short video content, and cultivate purchasing habits.

Low-Potential Users	<0.3	12.70%	Low-activity users with no preference	Low	Reduce active pushes, only push general benefits during major promotions to avoid disturbing users.
---------------------	------	--------	---------------------------------------	-----	---

The prediction model and stratification system not only validate the statistical laws derived from previous analysis but also form a closed data-driven operational loop from feature identification to strategy implementation. This framework provides a quantifiable and replicable decision-making tool for social e-commerce platforms to enhance conversion performance.

4. Conclusions

Under the Internet of Everything and digital economy transformation, social e-commerce is shifting from extensive traffic operation to refined data-driven governance. Focusing on the industry pain point of “high interaction but low conversion”, this study uses 100,000 user behavior records and builds an integrated framework with statistical testing, K-Means, Apriori and Random Forest. It systematically explores the driving mechanisms, heterogeneous patterns and optimization paths of purchase conversion, providing quantitative evidence for understanding user behavior and designing precise operation strategies.

The core conclusions of this study can be summarized as follows. First, the overall purchase conversion rate of the sample reaches 44.98%, which is considerably higher than the industry average. Key driving factors include add-to-cart behavior, purchase intent score, coupon usage, follow-author behavior, and browsing intensity, while price and social interaction variables present a significant right-skewed distribution that supports the existence of a “blockbuster product” effect. Second, user behavioral heterogeneity is prominent, and users can be clearly divided into three interpretable groups: discount-sensitive users, content-interactive users, and general users. Behavioral preferences rather than demographic attributes serve as the fundamental basis for user segmentation. Third, the combination of “like + add to cart + claim coupon” forms the most effective conversion trigger with high confidence and lift value, indicating that conversion is jointly promoted by social trust, decision-making action, and price incentive. Fourth, the Random Forest model achieves reliable predictive performance with an AUC of 0.7743, which can effectively distinguish high-potential users and support stratified operation.

This study offers both theoretical and practical contributions. Theoretically, it integrates multiple data mining methods into a unified framework for social e-commerce conversion analysis, enriching relevant empirical research and providing a replicable quantitative paradigm. Practically, it delivers actionable data-driven strategies for platform operation, helping reduce marketing costs, improve ROI, and enhance long-term competitiveness.

Several limitations of this study should be acknowledged. First, the data are derived from a single social e-commerce platform, which may restrict the generalizability of conclusions to other platforms or industries. Second, this study focuses on correlation analysis rather than causal inference, meaning that the internal causal mechanisms of conversion require further verification. Third, external factors such as macro consumption environment and platform policy changes are not included in the analytical framework.

Future research can be extended in three directions. First, multi-platform and multi-scenario data can be incorporated to enhance the universality of conclusions. Second, causal inference methods such as propensity score matching and instrumental variables can be applied to explore the causal relationships between user behavior and conversion. Third, deep learning models can be combined with temporal dynamic data to construct real-time prediction systems, further improving the timeliness and accuracy of precision operations.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Acknowledgments

This work was supported by the 6th Shanghai University Open Big Data Analysis Challenge. The authors thank the instructor Lin Tian for guidance and support.

References

- [1] Lin, Y.; Zhang, Q.; Liu, H. Social interaction and purchase intention in social commerce. *Journal of Electronic Commerce Research*, 2021, 22, 189-206.
- [2] Zhang, H.; Li, S. User segmentation for social e-commerce based on K-Means clustering. *Journal of Business Analytics*, 2022, 5, 112-128.
- [3] Wang, Y.; Chen, L.; Zhao, J. Purchase conversion prediction in social commerce using machine learning. *Applied Soft Computing*, 2023, 142, 110356.
- [4] Han, J.; Pei, J.; Kamber, M. Data Mining: Concepts and Techniques. Morgan Kaufmann, Boston, MA, USA. 2011.
- [5] Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News*, 2002, 2, 18–22.
- [6] Chen, X.; Wang, H. Driving factors of purchase conversion in social e-commerce: An empirical analysis. *Electronic Commerce Research and Applications*, 2020, 42, 100968.
- [7] Li, J.; Wu, Y. Association rule mining for user behavior in social commerce platforms. *Computational Intelligence and Neuroscience*, 2022, 2022, 1-12.
- [8] Zhou, M.; Yang, Z. Data-driven user segmentation and personalized marketing for social e-commerce. *Information Systems Frontiers*, 2021, 23, 1019-1035.
- [9] Guo, Y.; Liu, S. Imbalanced data processing in e-commerce conversion prediction using SMOTE and ensemble learning. *Computers & Industrial Engineering*, 2023, 175, 108829.
- [10] Park, S.; Kim, S. The effect of social interaction and price incentive on purchase decision in social commerce. *Journal of Retailing and Consumer Services*, 2020, 55, 102132.
- [11] Agarwal, R.; Srikant, R. Fast algorithms for mining association rules in large databases. *Proceedings of the 20th International Conference on Very Large Data Bases*, 1994, 487-499.

- [12]Sun, Y.; Li, J. Feature importance analysis for e-commerce purchase prediction based on tree-based models. *Applied Intelligence*, 2022, 52, 15104-15118.
- [13]Zhang, Y.; Wei, L. Digital transformation and operational efficiency in social e-commerce. *Journal of Business Research*, 2021, 136, 403-412.



© 2026 by the author(s); licensee International Technology and Science Publications (ITS), this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (<http://creativecommons.org/licenses/by/4.0/>)