

A CatBoost Social Commerce Conversion Prediction Model Based on BGMM Population Stratification

Zhongrong Yan^{1*}, Lingrui Kong^{1*}

¹ School of Statistics and Data Science, Shanghai University of International Business and Economics, Shanghai, China

*Correspondence: Zhongrong Yan, Lingrui Kong

Received: 22 May 2026; Accepted: 8 July 2026; Published: 3 August 2026

Abstract:

Social e-commerce has shifted from traditional search-based comparison models to content-driven discovery, where user purchase behavior follows sequential pathways of content exposure, social interaction, trust establishment, and final conversion. However, the inherent characteristics of user behaviors—including high dimensionality, non-linearity, and significant heterogeneity—combined with long-tail transaction distributions, pose substantial challenges to conventional one-size-fits-all prediction approaches. Traditional methods such as logistic regression and standard clustering techniques often struggle with categorical feature processing and overfitting in long-tail scenarios. To address these limitations, this paper proposes a hybrid modeling approach that integrates unsupervised Bayesian Gaussian Mixture Model (BGMM) for adaptive user stratification with Categorical Boosting, a gradient boosting algorithm specifically designed to handle categorical features effectively. Bayesian Gaussian Mixture Model automatically determines the optimal number of user segments via Dirichlet process priors, thereby overcoming manual cluster specification and improving generalization. The stratified user groups are subsequently incorporated as categorical features into CatBoost for accurate conversion prediction, enabling refined operational strategies on social e-commerce platforms.

Keywords:

Bayesian Gaussian Mixture Model, Categorical Boosting, Social E-commerce, User Segmentation, Conversion Prediction

1. Introduction

Social e-commerce has gradually shifted from the traditional search-and-comparison model to a content-driven model, and user purchase behavior follows the link of content exposure, social interaction, trust establishment and decision-making conversion. Affected by user attributes, commodity characteristics and social interaction, user behavior presents high-dimensional, non-linear and heterogeneous characteristics, and the transaction data shows an obvious long-tail distribution. A few high-value users contribute most of the transaction volume, while most ordinary users have low conversion rates. The traditional one-size-fits-all operation mode is difficult to meet the needs of different groups, so it is urgent to carry out refined user

stratification and accurate conversion prediction. User segmentation has become a core strategy for brands to build close ties with consumers and enhance stickiness, as demonstrated by typical brand practice of HEYTEA [9]. The behavioural patterns of app users shaped by functionality interaction further highlight the necessity of adaptive segmentation frameworks rather than static grouping [10], and relevant exploration of community operation and user layering represented by Lululemon also provides useful reference.

At present, the common conversion prediction methods mainly include logistic regression, traditional decision trees and neural networks. However, these methods have limitations in dealing with categorical features, heterogeneous data and interpretability. Categorical Boosting has been widely applied in many practical scenarios, which can effectively solve the problem of categorical feature processing [6]. Related research in cardiovascular disease detection verifies the advantages of the model in heterogeneous data mining [5], and it also shows good performance in equipment fault early warning [7] and meteorological prediction research [8]. Meta-learning provides an effective debiasing idea for conversion prediction under insufficient unbiased data [11]. Bayesian Gaussian mixture models (BGMM) can realize automatic clustering without artificially presetting categories, which has been well applied in traffic law analysis [2] and sensor simulation modeling [4]. In view of this, this paper introduces BGMM to realize automatic and adaptive user stratification, and combines CatBoost, which is good at processing categorical features, to build a hybrid prediction model.

The research objectives are as follows: (1) Realize efficient stratification of heterogeneous users; (2) Mine key conversion paths and influencing factors; (3) Construct a high-performance conversion prediction model; (4) Put forward targeted operational suggestions for social e-commerce platforms. Previous research on social e-commerce has laid a solid foundation for these objectives: studies on opinion leader characteristics, consumer psychological interaction mechanism, and Z-generation platform preference behavior provide critical insights into the factors shaping purchase intentions and platform choices, which can be integrated into the stratified conversion prediction framework proposed in this paper.

2. Materials and Methods

2.1. Data Sources and Preprocessing

The experimental data simulates the real interaction logs of Xiaohongshu and Douyin, with a total of 100,000 pieces of data, including 31 feature variables and 1 target variable (whether to purchase). The features are divided into seven categories: user identification, user characteristics, commodity characteristics, social interaction, behavior sequence, derived features and target variables.

Missing value detection was conducted, confirming that all core features contained no missing entries. Regarding outlier treatment, reasonable long-tail anomalies originating from high-value (head) users were retained to preserve the natural distribution characteristics. Lastly, to eliminate dimensional effects across features, both Min-Max scaling and Z-score standardization were employed

2.2. Feature Engineering

2.2.1. Variable Transformation and Distribution Adjustment

Log transformation ($\log(x+1)$) was applied to severely right-skewed variables such as price and number of likes. The distribution plots (Figure 1) clearly reveal the pattern of higher conversion rates for “low-priced goods” and “high-interaction content.” The log-transformed distribution plots reveal core differences between purchasing and non-purchasing users in terms of price and social interaction:

Price Distribution Plot (left): Shows that the distribution density of purchasing users (orange) in the low-price range (1-10 yuan) is significantly higher than that of non-purchasing users (blue), while the difference between the two groups narrows in the high-price range (>100 yuan). This indicates that social commerce users are highly price-sensitive, and low-priced goods are more likely to prompt immediate purchase decisions.

Likes Distribution Plot (right): Shows that purchasing users have significantly higher density in the high-interaction range (100-1000 likes), while the low-interaction range (<10 likes) is dominated by non-purchasing users. This validates the social proof effect—users trust content that has been recognized by many others, and high interaction effectively enhances purchase confidence.

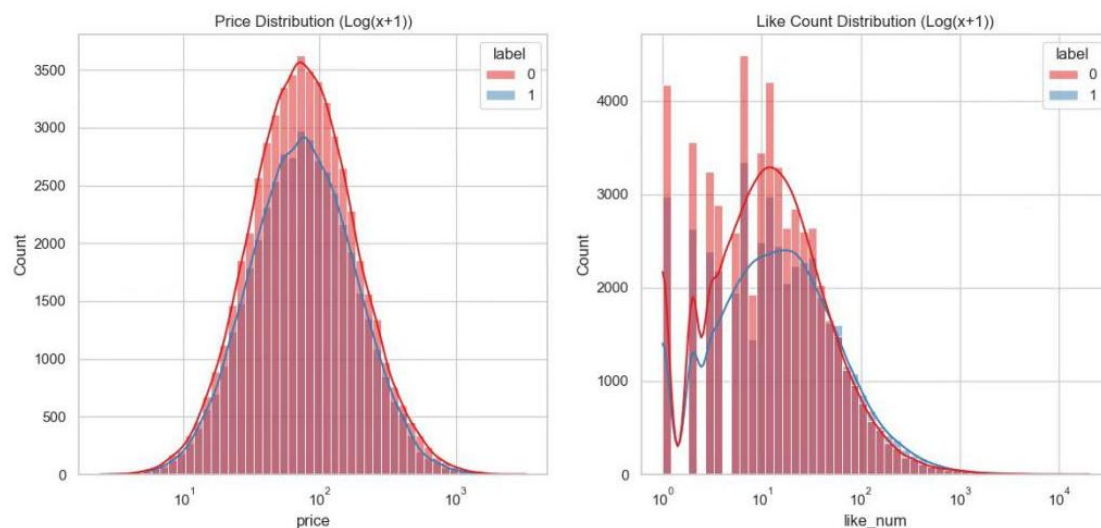


Figure 1. Price and Likes Distribution Plots.

2.2.2. Core Feature Selection and Correlation Analysis

From the 31 fields, 17 core features were selected, covering four categories: (1) User Characteristics: Age, User Level, Number of Purchases in Last 30 Days, Cumulative Consumption Amount; (2) Content and Product Characteristics: Price, Discount Rate, Title Sentiment Score, Number of Images, Whether Video Exists; (3) Social Interaction Features: Number of Likes, Number of Comments, Number of Favorites, Number of Shares, Whether Followed Author; (4) Behavioral Sequence Features: Whether Added to Cart, Whether Used Coupon, Time Since Last Click

Based on these features, a correlation coefficient matrix was calculated and a heatmap (Figure 2) was plotted to reveal linear correlations and collinearity with conversion.add2_cart has the highest correlation (0.36), being the strongest signal for purchase prediction; coupon_used correlation 0.20, is_follow_author correlation 0.13, both positive predictors; last_click_gap is negatively correlated (-0.07), shorter gaps lead to higher conversion rates; price shows a weak negative correlation (-0.01), low-priced goods are more easily converted

Collinearity among features: like_num, comment_num, and collect_num are highly correlated (0.6-0.8), reflecting the same dimension of "content popularity," allowing redundant information to be merged.

Therefore, add to cart, coupon usage, and following authors are the three core conversion signals; intervention strategies should be designed around these three actions. Likes, comments, and favorites can be merged into an interaction rate indicator.

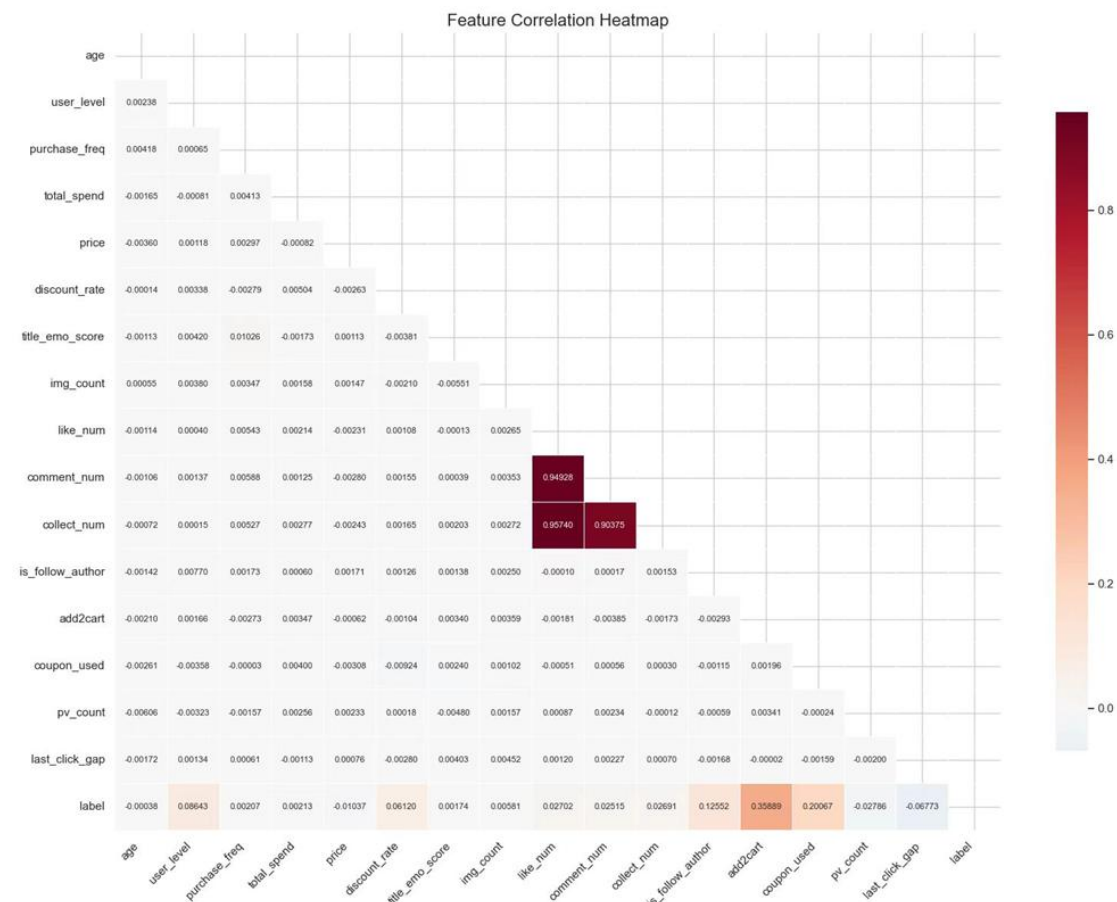


Figure 2. Feature Correlation Heatmap.

2.2.3. Consumption Distribution Skewness and Long-Tail Effect

This histogram (Figure 3) visually presents the distribution shape of total_spend. total_spend exhibits a typical right-skewed distribution (positive skew). This means that most users' consumption amounts are concentrated in the lower range. As the consumption amount increases, the frequency drops sharply, forming a very long "right tail," with a small number of samples still present in the high-consumption range of 30,000-50,000 yuan.

This distribution shape profoundly reflects the "long-tail effect" of user consumption behavior. The vast majority of users are ordinary consumers, with annual consumption concentrated in a relatively low range, constituting the platform's base. However, although the long-tail part has a small sample size, it represents high-value users. Their single consumption amounts are extremely high, contributing significantly to the platform's revenue.

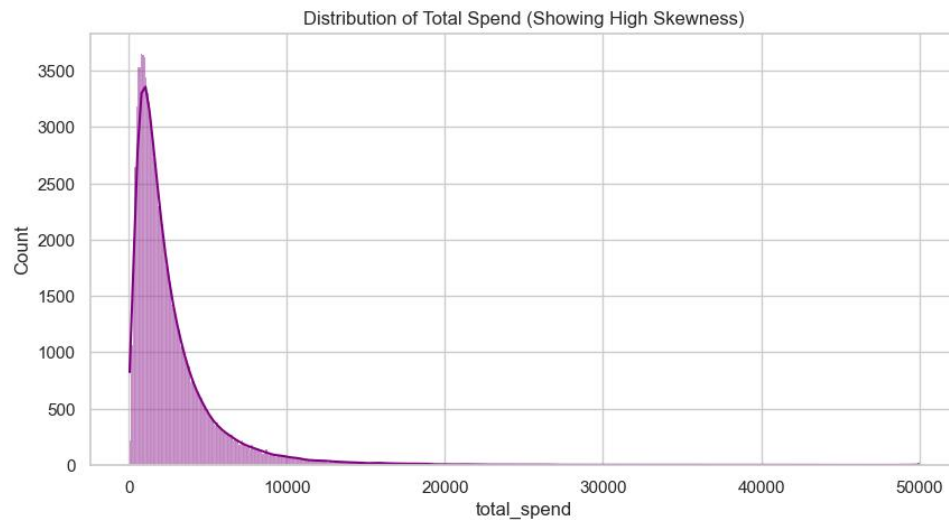


Figure 3. Right-skewed Distribution Characteristic of Total Consumption Amount.

2.3. BGMM-based User Segmentation and Profile Analysis

2.3.1. User Heterogeneity Stratification Based on Bayesian Gaussian Mixture Model

To overcome the limitations of the traditional K-Means algorithm in handling long-tail heterogeneous data, this study introduces the Bayesian Gaussian Mixture Model (BGMM). A standard Gaussian Mixture Model assumes that the data is generated by a linear combination of K multi-dimensional normal distributions, with its probability density function expressed as:

$$p(x) = \sum_{k=1}^K \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

where π_k is the mixing weight of the k -th component (satisfying $\sum \pi_k = 1$), and μ_k and Σ_k are its mean vector and covariance matrix, respectively. However, the standard GMM requires a fixed number of components K to be specified manually, making it prone to overfitting in complex long-tail distributions.

To address this, BGMM introduces a Bayesian inference framework, assigning prior distributions to the model parameters. For the mixing weights π , the model introduces a Dirichlet Process Prior. During the solution process using Variational Inference, by maximizing the Evidence Lower Bound (ELBO), the model can automatically shrink the weights π_k of redundant components that are not allocated sufficient samples towards zero, given an initial upper bound on the number of clusters.

During model initialization, this study set a relatively broad hyperparameter upper bound of $K_{\max} = 10$. After iterative training, the weight distribution results of the model are shown in Figure 4. It can be observed intuitively that the convergence mechanism based on the Dirichlet process strongly prunes the weights of redundant clusters. The weights of 2 component clusters were pruned to approach zero (both below the empirical effective threshold of 0.05) during iteration. Thus, mathematical convergence from 10 clusters to 8 clusters was initially achieved. These 8 clusters constitute the effective Gaussian distribution components intrinsic to this high-dimensional dataset.

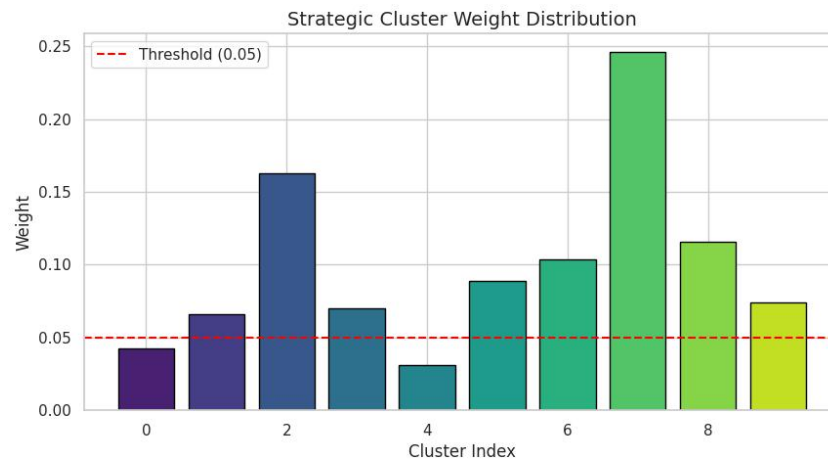


Figure 4. Posterior mixing weight distribution of BGMM components after variational inference.

2.3.2. Deconstruction of Multi-dimensional Statistical Profiles of Heterogeneous Groups

Further examining the 8 core groups identified through BGMM clustering and commercial space dimensionality reduction, this study performs a reverse deconstruction and calculates statistical means for the multi-dimensional behavioral characteristics of each group. Due to the vast differences in absolute values across different features (e.g., cumulative consumption amount, number of likes), to prevent a single numerical value from visually dominating the radar chart, this study uses Min-Max Scaling to linearly map each indicator to a standard interval of [0, 1]:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Within this standardized space, this study plotted a multi-dimensional statistical profile radar chart for the eight core groups (as shown in Figure 5).

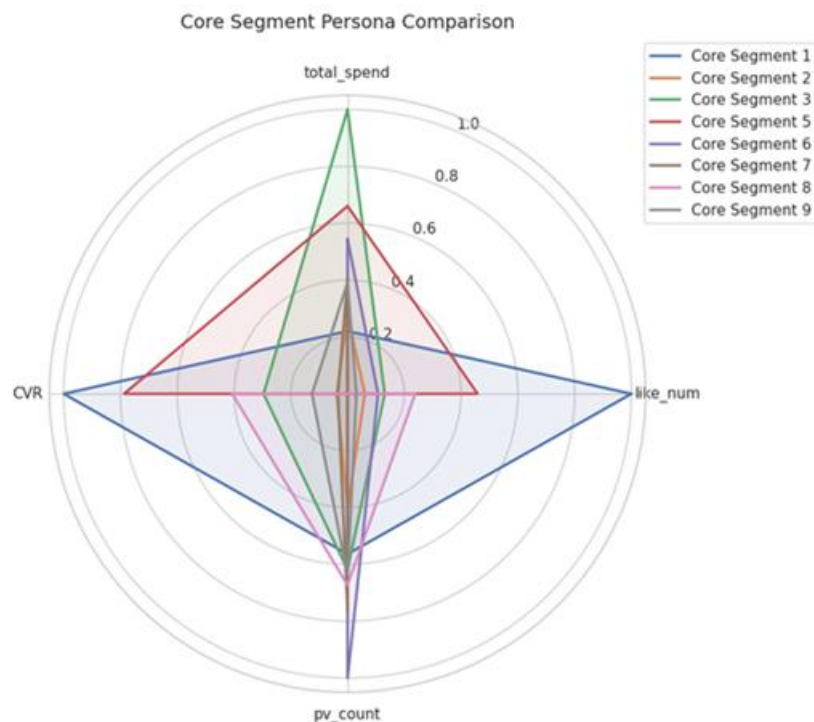


Figure 5. Radar chart of multi-dimensional statistical profiles for the eight core heterogeneous groups.

This radar chart compares the characteristic differences of core user segments across four dimensions: number of likes (like_num), number of views (pv_count), conversion rate (CVR), and total consumption (total_spend): Core Segment 3 (green) leads significantly in the total consumption dimension, representing the platform's high-net-worth core paying users; Core Segment 5 (red) performs strongest in the conversion rate dimension, representing the most effective conversion population with the clearest purchase intention; Core Segment 1 (blue) performs balanced in likes/interaction and browsing depth, being a core contributor to platform traffic and social activity, while also possessing good conversion potential.

To further examine the distribution shape of these 8 groups in real business dimensions, this study projected them onto a two-dimensional space (as shown in Figure 6) using users' "cumulative consumption amount (total_spend)" and "interaction likes count (like_num)" as core feature axes, on a log-log scale. The full long-tail user set presents a highly characteristic L-shaped distribution in logarithmic space. The boundaries of the 8 groups remain clearly separated and regularly intertwined within the "consumption-interaction" 2D space. Notably, the very small-weight micro-clusters are precisely separated from the majority population by BGMM. This classification result, which retains high-resolution group details, achieves an optimal fit to the data.

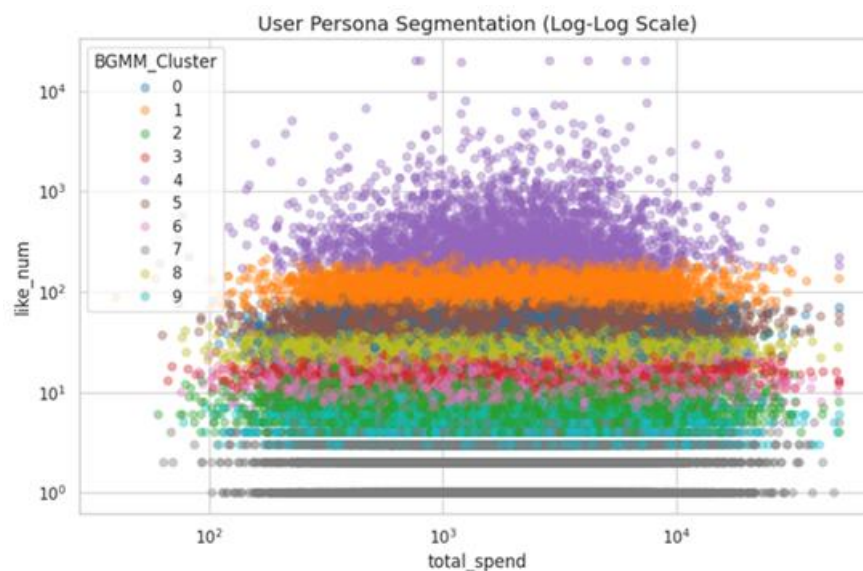


Figure 6. Log-coordinate 2D projection distribution of core heterogeneous groups.

Finally, based on the joint probability distribution output by BGMM and the feature mean vector of each group, combined with consumer behavior theory, this study analyzed the commercial profiles of these 8 fine-grained groups: Strong Conversion - High Average Order Value Group; Price-Sensitive-High Browsing Group; High Interaction - Low Conversion Group; Socially Dependent-Impulse Decision Group; Goal-Oriented-Low Lurking Group; Moderate and Balanced Mass Group; Marginal Micro-cluster A (Abnormal Deal-Seeker); Marginal Micro-cluster B (Low-Activity Long-Tail Dormant Group)

2.4. Modeling Process and Framework

2.4.1. Feature Construction

To enhance the model's ability to capture user behavior, the following derived features are constructed:

Add-to-Cart Conversion Rate Feature:

$$\text{cart_per_pv} = \frac{\text{add2cart}}{\text{pv_count} + 1}$$

Intent Speed Feature:

$$\text{intent_speed} = \frac{\text{purchase_intent}}{\text{last_click_gap} + 1}$$

Impulse Buyer Indicator

$$\text{is_impulse_buyer} = \begin{cases} 1, & \text{add2cart} > 0 \text{ and last_click_gap} < 2 \\ 0, & \text{otherwise} \end{cases}$$

Additionally, for long-tail variables such as total spend and follower count, a logarithmic transformation is applied: $x' = \log(1 + x)$

2.4.2. Unsupervised Population Stratification

(1) Standardization

Prior to clustering, numerical features are standardized: $z = \frac{x - \mu}{\sigma}$

(2) Bayesian Gaussian Mixture Model

Bayesian Gaussian Mixture Model assumes the data is generated by a mixture of multiple Gaussian distributions: $p(x) = \sum_{k=1}^K \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$

Where: K is the maximum number of clusters; π_k is the mixing weight; μ_k, Σ_k are the mean and covariance, respectively

A diagonal covariance structure is adopted:

$$\Sigma_k = \text{diag}(\sigma_1^2, \dots, \sigma_d^2)$$

(3) Cluster Feature Generation

Cluster labels are assigned using maximum posterior probability:

$$\text{cluster_tag} = \arg \max_k P(z_i \in k)$$

2.4.3. CatBoost Classification Model

(1) Model Structure

Categorical Boosting is a gradient-boosted decision tree model with the following prediction function:

$$F_M(x) = \sum_{m=1}^M \eta f_m(x)$$

Where: M is the number of trees; η is the learning rate;

$f_m(x)$ is the m -th tree

For binary classification, the output probability is:

$$P(y = 1|x) = \frac{1}{1 + e^{-F_M(x)}}$$

(2) Loss Function

The model optimizes the log-loss function:

$$L = - \sum_{i=1}^n [y_i \log p_i + (1 - y_i) \log (1 - p_i)]$$

(3) Categorical Feature Encoding Mechanism

A key advantage of CatBoost lies in its encoding of categorical features. For a given category c , ordered target statistics encoding is used:

$$\text{Enc}(c) = \frac{\sum_{j < i} y_j \cdot \mathbb{I}(x_j = c) + aP}{\sum_{j < i} \mathbb{I}(x_j = c) + a}$$

Where: P is the global mean; a is the smoothing parameter

(4) Model Parameters

Learning rate: 0.03; Tree depth: 6; Number of iterations: 3000; L2 regularization: 12; Random perturbation strength: 1.5; Early stopping rounds: 150

3. Results and Discussion

The learning curve (Figure 7) further validates the training dynamics. The overfitting detector was triggered at the 181st iteration, automatically stopping training and reverting to the optimal state. This phenomenon demonstrates the model's good adaptive capability during training.

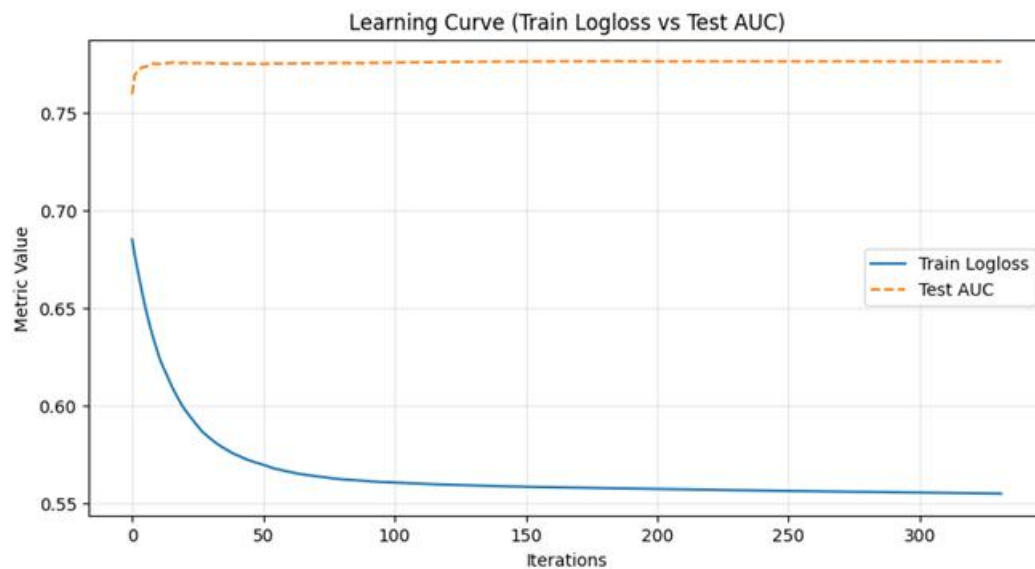


Figure 7. Learning curve chart.

The training set Logloss (blue line) continuously decreases with iterations, showing the model constantly enhances its fit to the training data and possesses strong learning ability. Meanwhile, the test set Logloss (red line) decreases significantly in the early stages, indicating the model quickly learns effective features. After approximately 181 rounds, it gradually stabilizes and shows a slight upward trend, which is typically an

early sign of overfitting. By stopping training early at the 'bestTest' point and reverting the model, overfitting to the training data is effectively avoided, thereby improving generalization performance on unseen data.

SHAP analysis (Figure 8) provides an interpretable analysis of the model, clearly revealing the contribution of each feature to the prediction results. Among the original behavioral features, add2cart (add-to-cart count) is the most critical predictive factor, directly reflecting the user's purchase intention. Furthermore, the custom-engineered feature cart_per_pv (add-to-cart conversion rate) ranks high in importance, suggesting that "behavioral efficiency" has more predictive value than the mere frequency of behavior. This feature effectively captures the user's conversion tendency during browsing, distinguishing between "high-intent users" and "inefficient browsers." Additionally, timeliness and impulse-related features (such as is_impulse_buyer and last_click_gap) significantly impact the model's predictions. Specifically, users who act within a short period are more likely to convert, while prolonged inactivity significantly reduces the purchase probability.

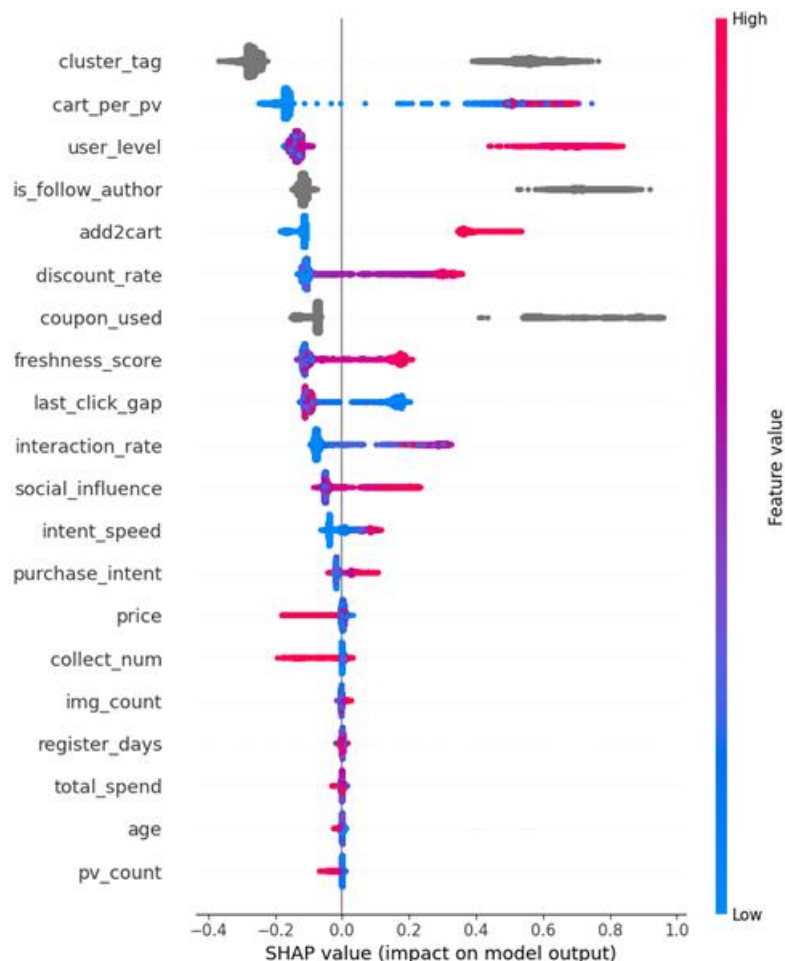


Figure 8. SHAP feature importance plot.

4. Conclusions

This paper proposes a hybrid modeling approach that combines unsupervised learning with gradient boosting trees. By using Bayesian Gaussian Mixture Model for population stratification and Categorical Boosting for classification prediction, the model effectively enhances its ability to capture complex user behaviors.

Experimental results show that this method performs well in both predictive performance and interpretability, making it suitable for complex recommendation and conversion prediction scenarios such as social e-commerce.

4.1. Strengthen Real-time Conversion Strategies for High-Intent Users

For users identified as having low browsing frequency but frequent add-to-cart actions (high `cart_per_pv`), it is recommended to immediately push information such as limited-time discounts or low-stock alerts upon triggering their behavior to accelerate decision-making. These users have short decision paths and strong conversion intent, making them the core audience for improving overall precision.

4.2. Implement Windowed Exposure Strategies for Impulsive Users

For users flagged as `is_impulse_buyer = 1`, significantly increase the exposure frequency of relevant products or categories within the golden 2-hour window following their behavior. Simultaneously, reduce interference from cross-comparison content to guide them towards completing a purchase quickly, thereby maximizing short-term conversion efficiency.

4.3. Implement Differentiated Operational Strategies Based on User Clusters

Leverage the user segmentation characteristics reflected by `cluster_tag` to execute refined operations. For price-sensitive users, highlight coupon advantages and final discounted prices. For social-driven users, strengthen expert recommendations, user reviews, and social influence endorsements. Matching differentiated content to distinct user profiles enhances overall conversion effectiveness.

4.4. Activate Long-tail Users to Improve Recall

The current model's recall for the positive class is 0.61, indicating that a significant proportion of potential converting users remain unidentified. It is recommended to re-engage less active users with large `last_click_gap` through secondary touches such as SMS or App Push. Additionally, dynamically adjust strategies based on changes in their `intent_speed` to gradually awaken silent users and improve overall recall capabilities.

Conflicts of Interest

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

Funding

This work was supported by the 6th Shanghai University Open Big Data Analysis Challenge.

Acknowledgments

The authors gratefully acknowledge the organizing committee of the 6th Shanghai University Open Big Data Analysis Challenge for providing the dataset and research platform. Thanks are also extended to Shanghai University of International Business and Economics for academic support. We sincerely appreciate the valuable guidance from Supervisor Xin Yan throughout this study.

References

- [1] Fateen, E.; Nosier, S.S.; Radwan, A.M.; Mohammed, E.E.A. Clinical, biochemical, and molecular characterization of a cohort of Egyptian patients with Sanfilippo B syndrome (MPS IIIB): Bayesian Gaussian mixture model. *Mol. Biol. Rep.* 2025, 52, 1009. DOI: <https://doi.org/10.1007/S11033-025-11012-1>.
- [2] Chen, X.; Saidi, S.; Sun, L. Understanding bus delay patterns under different temporal and weather conditions: A Bayesian Gaussian mixture model. *Transp. Res. Part C Emerg. Technol.* 2025, 171, 105000. DOI: <https://doi.org/10.1016/J.TRC.2025.105000>.
- [3] Mohammed, E.E.A.; Fayez, A.G.; Abdelfattah, N.M.; Fateen, E. Novel gene-specific Bayesian Gaussian mixture model to predict the missense variants pathogenicity of Sanfilippo syndrome. *Sci. Rep.* 2024, 14, 12148. DOI: <https://doi.org/10.1038/S41598-024-62352-0>.
- [4] Walenta, K.; Genser, S.; Solmaz, S. Bayesian Gaussian Mixture Models for Enhanced Radar Sensor Modeling: A Data-Driven Approach towards Sensor Simulation for ADAS/AD Development. *Sensors* 2024, 24, 2177. DOI: <https://doi.org/10.3390/S24072177>.
- [5] Sen, J.; Bhattacharya, S. An explainable hybrid framework for early detection of cardiovascular diseases using Categorical Boosting and Bees algorithm. *Sci. Rep.* 2025, 15, 45748. DOI: <https://doi.org/10.1038/S41598-025-28514-4>.
- [6] Fan, Z.; Gou, J.; Wang, C. An error complementarity-based iterative learning approach via categorical boosting for student performance prediction. *Eng. Appl. Artif. Intell.* 2025, 161, 112192. DOI: <https://doi.org/10.1016/J.ENGAPPAL.2025.112192>.
- [7] Liu, K.; Wang, G.; Mao, D.; Huang, J. A Hybrid Fault Early-Warning Method Based on Improved Bees Algorithm-Optimized Categorical Boosting and Kernel Density Estimation. *Processes* 2025, 13, 1460. DOI: <https://doi.org/10.3390/PR13051460>.
- [8] Ding, Y.; Wang, Y.; Li, Z.; Zhao, L.; Shi, Y.; Xing, X.; Chen, S. Improving Solar Radiation Prediction in China: A Stacking Model Approach with Categorical Boosting Feature Selection. *Atmosphere*, 2024, 15, 1436. DOI: <https://doi.org/10.3390/ATMOS15121436>.
- [9] Yang, P. From User Segmentation to Brand Loyalty: HEYTEA's Path to Building Close Ties with Consumers. *Acad. J. Manag. Soc. Sci.* 2025, 13, 65-71. DOI: <https://doi.org/10.54097/M15PBG70>.
- [10] Ooi, L.Y.; Ting, C.Y.; Zakariah, H.; Chandar, E. Behavioural user segmentation of app users based on functionality interaction patterns. *Cogent Eng.* 2024, 11. DOI: <https://doi.org/10.1080/23311916.2024.2430430>.
- [11] Li, P.; Tong, X.; Wang, Y.; Zhang, Q. Meta doubly robust: Debiasing CVR prediction via meta-learning with a small amount of unbiased data. *Knowl.-Based Syst.* 2025, 310, 112898. DOI: <https://doi.org/10.1016/J.KNOSYS.2024.112898>.



© 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/>)