Academic Journal of Management and Social Sciences
ISSN: 2958-4396 | Vol. 15, No. 2, 2026

Customer Segmentation and Churn Prediction in Express Logistics

Yulin Wu [a], Han Zhang [b], Hao Wang *

School of Information Science and Engineering, Dalian Polytechnic University, Dalian, Liaoning 116034, P.R. China
*** Corresponding author:** Hao Wang (Email: wanghao120911@outlook.com), [a] 230520854000567@xy.dlpu.edu.cn,
[b] zhang1506711279@outlook.com

Abstract: With the rapid growth of e-commerce and increasing competition in the express logistics industry, effective customer management has become critical for improving operational efficiency and revenue stability. This study proposes a data-driven framework for refined customer management based on real-world waybill data. Customer behavior is characterized using multi-dimensional features derived from aggregated transaction records. A hybrid segmentation approach combining a customer value four-quadrant model and K-means clustering is employed to identify customer value levels and behavioral patterns. Based on the segmentation results, a Random Forest model is developed to classify customers into churn and non-churn groups and identify high-risk customers. Experimental results show that the clustering model achieves a silhouette coefficient of 0.8797, while the Random Forest model outperforms other models with an accuracy of 0.825. The results demonstrate that the proposed framework effectively identifies high-value customers with elevated churn risk and supports more informed customer management decisions.

Keywords: Customer Segmentation; K-means Clustering; Customer Value Analysis; Express Logistics; Customer Churn Prediction.

1. Introduction

With the rapid development of e-commerce and the increasing complexity of supply chain systems[1][2], the express logistics industry has entered a stage characterized by intensified competition and shrinking profit margins[3]. In this context, customer management has become a critical factor influencing enterprise performance, revenue stability, and long-term competitiveness[4]. Customers in the express logistics sector often exhibit high shipment frequency, diverse service demands, and strong sensitivity to cost and service quality, making them prone to switching providers in a highly competitive environment[5]. However, many enterprises still rely on coarse-grained customer management strategies, which fail to effectively capture customer heterogeneity and often result in inefficient resource allocation, loss of high-value customers, and suboptimal service delivery[6]. To address these challenges, data-driven approaches have been increasingly adopted in customer analysis. Clustering methods, such as K-means, have been widely used to identify latent customer groups from behavioral data due to their simplicity and effectiveness. In addition, customer value evaluation models, including RFM-based frameworks, have been applied to support targeted marketing and resource allocation. Meanwhile, predictive models such as random forest and gradient boosting algorithms have demonstrated strong performance in customer churn prediction tasks. These approaches provide useful tools for understanding customer behavior and identifying potential risks[7][8][9][10].

Despite these advances, most existing studies focus on isolated analytical tasks, such as customer segmentation or churn prediction, without considering their integration[11]. As a result, the insights obtained from these models are often fragmented and difficult to translate into actionable strategies[12]. In particular, segmentation models can reveal customer value levels but cannot identify potential churn risk, while churn prediction models can estimate risk but lack contextual information about customer importance[13] [14] [15]. The absence of a unified analytical framework limits the effectiveness of data-driven decision-making in real-world logistics operations[16][17][18]. To overcome these limitations, this study proposes a unified data-driven framework for refined customer management in express logistics enterprises based on real-world operational data, as illustrated in Figure 1. The framework integrates data preprocessing, feature engineering, customer segmentation, and churn prediction into a unified process, enabling the identification of critical customer groups by jointly considering customer value and churn risk. Within this framework, customer segmentation, value modeling, and churn prediction are systematically integrated into a coherent analytical process. Customer behavior is first characterized through multi-dimensional features, and a hybrid segmentation strategy combining a customer value four-quadrant model with K-means clustering is employed to identify customer value levels and behavioral patterns. Building upon the segmentation results, a Random Forest-based churn prediction model is developed to identify high-risk customers. By jointly considering customer value and churn risk, the proposed approach enables a more comprehensive identification of critical customer groups. This integrated analysis enables a systematic approach to customer management, where segmentation provides a structured representation of customer value, and prediction identifies potential risk patterns, thereby supporting more informed and targeted decision-making in express logistics enterprises.

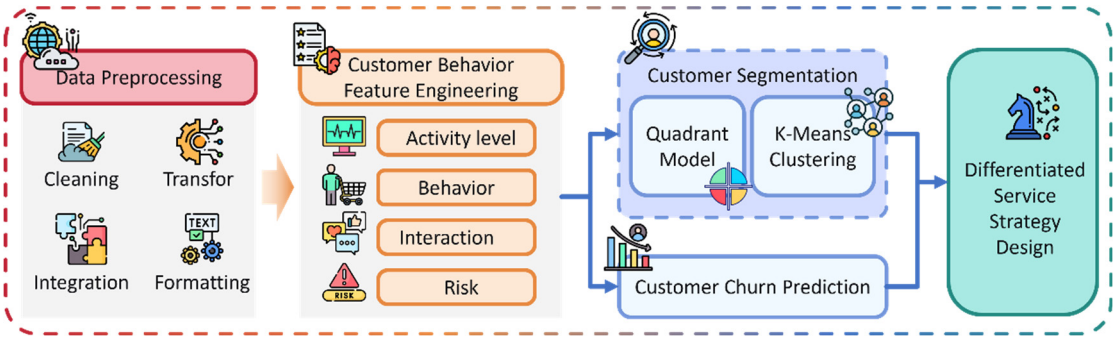


Figure 1. Overall framework of the proposed customer analysis approach

2. Materials and Method

2.1. Dataset construction

The dataset used in this study was constructed based on real-world waybill records provided by a large express logistics enterprise. To enable customer-level analysis, raw transaction data were aggregated using a unique customer identifier, transforming fragmented waybill records into structured customer profiles. This aggregation allows customer behavior to be represented from multiple perspectives, including value contribution, shipment frequency, cargo characteristics, and spatial distribution. Specifically, a set of primary aggregated features was first extracted, including total revenue, number of shipments, total cargo quantity, and billing weight. These variables directly reflect the scale and intensity of customer interactions with the logistics system. Based on these primary indicators, additional derived features were further constructed to capture more detailed behavioral patterns, such as average revenue per shipment, average billing weight, average number of items per shipment, revenue per unit weight, proportion of bulky shipments, and regional concentration. In total, approximately 30 variables were included to form a comprehensive feature set describing customer behavior.

All numerical features were normalized or standardized to eliminate scale differences and ensure comparability across variables. This dataset construction strategy provides a structured and comprehensive representation of customer behavior, forming a solid foundation for subsequent customer segmentation, value modeling, and churn prediction.

2.2. Customer Segmentation and Value Modeling

To achieve refined customer segmentation, this study develops a hybrid framework that integrates a rule-based Customer Value Four-quadrant Model with a data-driven K-means clustering approach. The former provides an interpretable segmentation based on key business indicators, while the latter enables fine-grained grouping in a multi-dimensional feature space.

From a business perspective, customer value and activity are two fundamental dimensions for describing customer behavior in express delivery services. Accordingly, a Customer Value Four quadrant Model is first constructed using total revenue and shipment frequency. Let V_i denote the total revenue generated by customer i, and F_i denote the corresponding shipment frequency. To improve robustness and mitigate the influence of extreme values, the median values of these two variables, denoted as M_V and M_F, are adopted as segmentation thresholds. Based on their relative positions, customers are partitioned into four groups: high-value high-frequency customers, highvalue low-frequency customers, low-value high-frequency customers, and low-value low-frequency customers.

A two-dimensional coordinate system is then constructed with total revenue and shipment frequency as the axes, and the segmentation thresholds M_V and M_F are used to divide the space into four quadrants for visualization and interpretation. To further capture the heterogeneity of customer behavior, a K-means clustering model is introduced. In this study, three representative features are selected, including total revenue, shipment frequency, and total shipment volume, which jointly characterize customer value, activity level, and operational scale. To eliminate the influence of different feature scales, Min-Max normalization is applied to all features. The K-means algorithm partitions the dataset into K clusters by minimizing the within-cluster sum of squared distances:

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} \| x_i - \mu_k \|^2$$

where C_k denotes the k-th cluster and μ_k represents its centroid. The clustering process iteratively updates cluster assignments and centroids until convergence. The optimal number of clusters is determined using a combination of the elbow method and silhouette coefficient, which evaluates the compactness and separability of clusters.

2.3. Customer Churn Prediction Model

After completing customer segmentation, the characteristics and value levels of different customer groups have been clearly identified. However, segmentation alone only reflects the current state of customer behavior and cannot capture future changes, particularly the risk of customer churn. Therefore, further analysis is required to evaluate the potential churn risk of customers within different groups. A customer churn prediction model is developed based on historical behavioral data to identify customers with potential churn risk. The objective of the model is to learn the differences between churned and retained customers and to classify customers into churn and non-churn groups. The input features used in this model are constructed as described in Section 2.1, covering multiple aspects of customer behavior, including value contribution, operational scale, and service usage patterns. The churn prediction problem is formulated as a binary classification task, where the target variable indicates whether a customer has churned. The overall modeling process is illustrated in Figure 2.

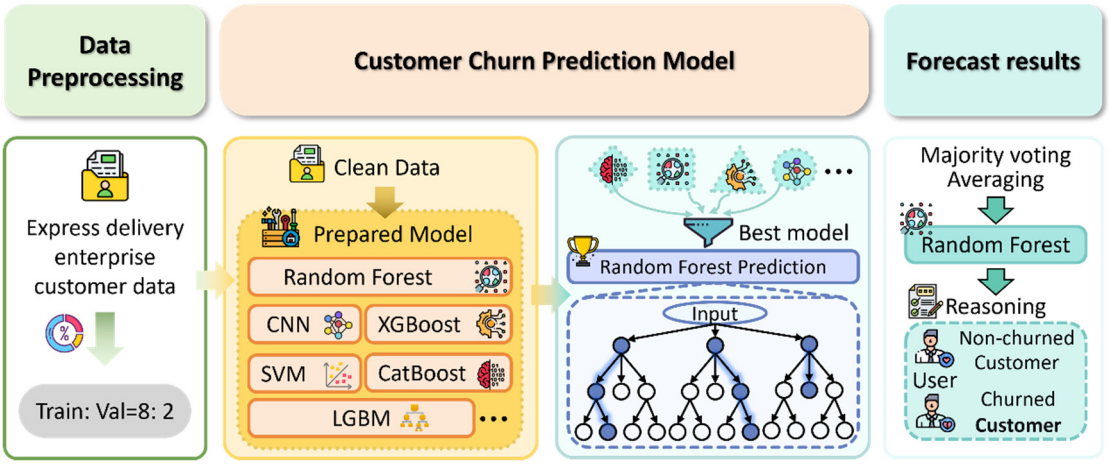


Figure 2. Workflow of the customer churn prediction model

To ensure the robustness of the modeling framework, multiple machine learning and deep learning models are constructed and compared, including Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest, Extreme Gradient Boosting (XGBoost), LightGBM, Categorical Boosting (CatBoost), and a one-dimensional Convolutional Neural Network (1D-CNN). All models are trained using the same feature set and data representation to ensure comparability. After model training, the performance of each model is evaluated on the test set using multiple evaluation metrics, including accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC). These metrics provide a comprehensive assessment of classification performance, particularly under class imbalance conditions.

The evaluation metrics are defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

where TP (True Positive) represents the number of correctly predicted positive samples, FP (False Positive) denotes the number of samples incorrectly predicted as positive, FN (False Negative) refers to the number of positive samples incorrectly predicted as negative, and TN (True Negative) represents the number of correctly predicted negative samples. The F1-score is defined as the harmonic mean of precision and recall:

$$\text{F1} - \text{score} = \frac{2 \cdot (\text{Precision} \cdot \text{Recall})}{\text{Precision} + \text{Recall}}$$

In addition, the AUC metric is used to evaluate the overall discriminative ability of the model across different classification thresholds. AUC values range from 0 to 1, where higher values indicate better classification performance. Compared with accuracy, AUC provides a more reliable evaluation for imbalanced classification problems, as it reflects the model's ability to distinguish between classes under varying decision thresholds. Therefore, AUC is considered the primary evaluation metric in this study.

3. Results and Discussion

3.1. Customer Segmentation Results

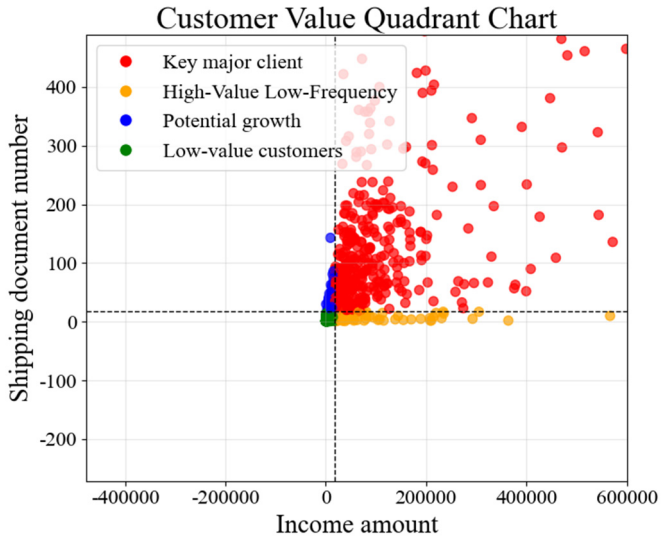
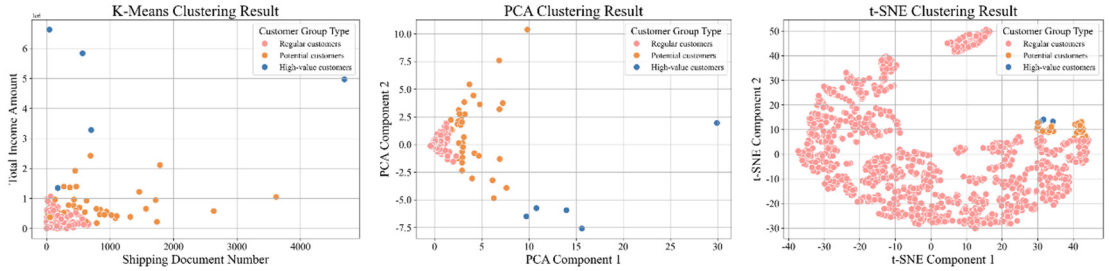


Figure 3. Customer value quadrant distribution based on revenue and shipment frequency

The customer segmentation results obtained from the proposed hybrid framework are presented in this section. To provide an interpretable overview of customer value distribution, a preliminary segmentation is first conducted using the Customer Value Four-quadrant Model. As shown in Figure 3, customers are distributed unevenly across the four

quadrants defined by total revenue and shipment frequency. A relatively small number of customers are concentrated in the high-revenue and high-frequency region, indicating that a limited group contributes disproportionately to overall business performance. In contrast, a large proportion of customers are located near the origin, reflecting low activity levels and limited contribution. This distribution reveals a clear hierarchical structure of customer value and provides an intuitive baseline for subsequent analysis.

To further capture the heterogeneity of customer behavior, the K-means clustering model is applied with the optimal number of clusters set to K=3. The clustering results are shown in Figure 4A, where customers are grouped into high-value customers, potential customers, and regular customers. The high-value customer group is primarily located in the region with both high revenue and high shipment frequency, while potential customers exhibit relatively high activity but moderate value. Regular customers are mainly concentrated in the low-value and low-frequency region. This distribution is consistent with the pattern observed in Figure 1, indicating that the data-driven clustering results align well with the business-based segmentation. To further evaluate the robustness of the clustering structure, dimensionality reduction techniques are applied. As shown in Figure 4B and Figure 4C, the clustering results are projected into lower-dimensional spaces using PCA and t-SNE, respectively. In both representations, the three customer groups form clearly separated regions with minimal overlap. In particular, high-value customers appear as a compact and well-isolated cluster, indicating strong intra-cluster similarity, while potential and regular customers remain distinguishable in both projection spaces. Quantitatively, the clustering achieves a silhouette coefficient of 0.8797, indicating strong intra-cluster cohesion and clear inter-cluster separation. Combined with the visual evidence in Figure 2, this result confirms the effectiveness of the proposed segmentation approach in capturing the underlying structure of customer behavior.



A. K-means clustering result B. PCA clustering result C. t-SNE clustering result

Figure 4. Visualization of customer segmentation results using K-means clustering and dimensionality reduction

Overall, the results demonstrate that the hybrid framework provides both interpretable and data-driven insights into customer segmentation. The identified customer groups not only reflect practical business patterns but also establish a reliable basis for subsequent churn prediction and targeted customer management.

3.2. Customer Churn Prediction Results

The performance of different models on the customer churn prediction task is evaluated using multiple metrics, including accuracy, precision, recall, F1-score, and AUC. The quantitative results are summarized in Table 1, and the corresponding visualization is presented in Figure 5. Among all models, the Random Forest model achieves the best overall performance, with an accuracy of 0.8249, an AUC of 0.7669, and an F1-score of 0.2146. Compared with other models, Random Forest demonstrates a more balanced performance across multiple evaluation metrics. In contrast, gradient boosting models such as XGBoost and LightGBM show competitive accuracy and AUC but relatively lower F1-scores, while CatBoost and Logistic Regression exhibit higher recall but weaker precision and overall stability. Other models, including SVM, KNN, and CNN, demonstrate comparatively lower performance across most metrics.

Table 1. Comparison of Experimental Results of Different Models

Model	Accuracy	Precision	Recall	F1-score	AUC
Random Forest	**0.825**	**0.137**	0.707	**0.215**	**0.767**
XGBoost	0.808	0.118	0.451	0.187	0.717
CatBoost	0.303	0.060	0.896	0.112	0.708
LightGBM	0.798	0.124	0.424	0.192	0.706
CNN	0.579	0.078	0.456	0.141	0.705
KNN	0.821	0.113	0.499	0.181	0.703
SVM	0.485	0.063	0.683	0.115	0.633
Logistic Regression	0.255	0.052	0.832	0.099	0.545

Clear differences can be observed across models in terms of evaluation metrics, with the heatmap providing a visual comparison of their performance trade-offs. The Random Forest model maintains consistently strong performance across most metrics, whereas other models exhibit noticeable imbalances, such as high recall but low precision. Based on the classification results, customers can be divided into churn and non-churn groups, which enables the identification of customers with higher churn risk. When combined with the customer segmentation results in Section 3.1, these

classification results make it possible to identify critical customer groups, particularly high-value customers that are simultaneously associated with churn status. Overall, the results demonstrate that the proposed approach provides a reliable basis for customer risk identification and supports more targeted customer management.

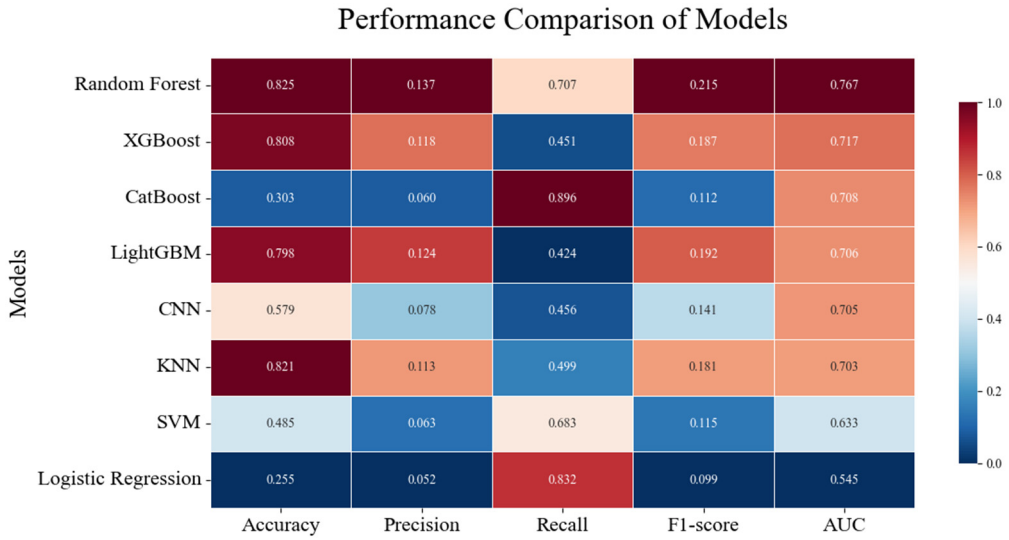


Figure 5. Heatmap visualization of performance comparison across different models

The results of this study demonstrate that integrating customer segmentation with churn prediction provides an effective framework for refined customer management in express logistics enterprises. The combination of the Customer Value Four-quadrant Model and K-means clustering enables a structured identification of customer value levels and behavioral patterns, revealing a clear hierarchical structure in which a small proportion of high-value customers contributes disproportionately to overall revenue, while potential customers exhibit high activity but relatively lower value contribution. Building upon this segmentation, the Random Forest-based churn prediction model further identifies customers with elevated risk levels and demonstrates strong robustness in handling complex and imbalanced data. Notably, the model is capable of detecting not only low-value customers at risk of churn, but also high-value customers showing early signs of disengagement. By jointly considering customer value and churn risk, the proposed approach enables a more comprehensive identification of critical customer groups, providing a more informative basis for decision-making than traditional single-perspective analysis.

From a practical perspective, the integration of segmentation and prediction supports more targeted customer management strategies. High-value customers can be prioritized for retention through proactive monitoring and personalized services, while potential customers can be guided toward value enhancement through appropriate incentives. In addition, the results suggest that linking customer-level analysis with regional characteristics can further improve resource allocation and operational efficiency. Despite these advantages, several limitations remain. The current model relies on static aggregated features and does not capture temporal dynamics of customer behavior, and external factors such as market competition and economic conditions are not considered. Future work may incorporate time-series data and external contextual variables to develop more dynamic and adaptive customer analysis models.

4. Conclusion

This study investigates customer segmentation and churn prediction for express delivery enterprises and proposes a data-driven framework for refined customer management. A hybrid segmentation approach combining a customer value four-quadrant model with K-means clustering is employed to achieve both interpretable and data-driven customer grouping. The results demonstrate that the proposed method effectively identifies customer value levels and partitions customers into high-value, potential, and regular groups, with a silhouette coefficient of 0.8797, further supported by PCA and t-SNE visualizations. For churn prediction, the Random Forest model achieves the best overall performance among eight comparative models, with an accuracy of 0.825, enabling effective identification of customers with elevated churn risk. Overall, the proposed framework integrates value-based segmentation, clustering analysis, and churn prediction into a unified approach, offering both methodological support and practical implications for customer management in the express logistics industry. Future work may incorporate temporal behavioral features and external factors to further improve model performance and adaptability.

References

[1] Gunasekaran A, Subramanian N, Papadopoulos T. Information technology for competitive advantage within logistics and supply chains: A review[J]. Transportation Research Part E: Logistics and Transportation Review, 2017, 99: 14-33.

[2] E-Logistics: Managing your digital supply chains for competitive advantage[M]. Kogan Page Publishers, 2016.

[3] Ivanov D, Dolgui A. A digital supply chain twin for managing the disruption risks and resilience in the era of Industry 4.0[J]. Production Planning & Control, 2021, 32(9): 775-788.

[4] Haenlein M, Verhoef P C, Donkers B, et al. Customer Relationship Management: Concept, Strategy, and Tools[J]. International Journal of Research in Marketing, 2013.

[5] Hübner A, Holzapfel A, Kuhn H. Distribution systems in omni-channel retailing[J]. Business Research, 2016, 9(2): 255-296.

[6] Payne A, Frow P. A strategic framework for customer relationship management[J]. Journal of marketing, 2005, 69(4): 167-176.

[7] Verbeke W, Dejaeger K, Martens D, et al. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach[J]. European journal of operational research, 2012, 218(1): 211-229.

[8] Akter S, Bandara R, Hani U, et al. Analytics-based decision-making for service systems: A qualitative study and agenda for future research[J]. International Journal of Information Management, 2019, 48: 85-95.

[9] Tsai C F, Lu Y H. Customer churn prediction by hybrid neural networks[J]. Expert Systems with Applications, 2009, 36(10): 12547-12553.

[10] Wamba S F, Gunasekaran A, Akter S, et al. Big data analytics and firm performance: Effects of dynamic capabilities[J]. Journal of business research, 2017, 70: 356-365.

[11] Ngai E W T, Xiu L, Chau D C K. Application of data mining techniques in customer relationship management: A literature review and classification[J]. Expert systems with applications, 2009, 36(2): 2592-2602.

[12] Dibb S. Market segmentation: Conceptual and methodological foundations[J]. Journal of Targeting, Measurement and Analysis for Marketing, 2000, 9(1): 92-93.

[13] Fader P S, Hardie B G S, Lee K L. RFM and CLV: Using iso-value curves for customer base analysis[J]. Journal of marketing research, 2005, 42(4): 415-430.

[14] Chen T, Guestrin C. Xgboost: A scalable tree boosting system [C]// Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.

[15] Huang B, Kechadi M T, Buckley B. Customer churn prediction in telecommunications[J]. Expert Systems with Applications, 2012, 39(1): 1414-1425.

[16] Vafeiadis T, Diamantaras K I, Sarigiannidis G, et al. A comparison of machine learning techniques for customer churn prediction[J]. Simulation Modelling Practice and Theory, 2015, 55: 1-9.

[17] Raeisi S, Sajedi H. E-commerce customer churn prediction by gradient boosted trees[C]//2020 10th International Conference on Computer and Knowledge Engineering (ICCKE). IEEE, 2020: 055-059.

[18] Wu J, Shi L, Yang L, et al. User value identification based on improved RFM model and K-means++ algorithm for complex data analysis[J]. Wireless Communications and Mobile Computing, 2021, 2021(1): 9982484.