Research Article

Qiuping Lu*

E-commerce big data processing based on an improved RBF model

https://doi.org/10.1515/jisys-2023-0131
received August 17, 2023; accepted July 17, 2024

Abstract: In the dynamic landscape of China's booming economy, the surge in e-commerce customer volume presents both opportunities and challenges, notably in managing customer churn (CC). Addressing this critical issue, this study introduces an innovative approach employing a radial basis function neural network for predicting CC within the e-commerce sector. To enhance the model's performance in handling the vast and complex data inherent to e-commerce, the least absolute shrinkage and selection operator regression algorithm is employed, optimizing the model's predictive accuracy. By meticulously analyzing the customer lifecycle, this refined model adeptly predicts churn at various stages, enabling the identification of features most correlated with churn. Empirical results underscore the model's exceptional capability, achieving a prediction accuracy of 95% and a remarkably low loss rate of 3%. Furthermore, during the excavation, advanced, stable, and decline stages of the customer lifecycle, accuracy levels of 97.6, 93.1, 92.7, and 91.8% are attained, respectively, facilitating the precise selection of highly correlated customer features. Thus, the advanced churn prediction model proposed herein significantly contributes to the e-commerce domain, offering a robust tool for strategizing customer retention and mitigating churn.

Keywords: RBF model, customer churn, Lasso algorithm, big data, lifecycle

1 Introduction

In recent years, the transaction volume of e-commerce customers has been rapidly increasing, mainly due to the dividends brought by the era of digitization. In the era of big data, customers' preferences, consumption levels, and other information are monitored at any time to provide them with accurate services and improve their viscosity [1,2]. However, due to the varying preferences and consumption habits of customers, there is a certain turnover rate in the field of e-commerce. Customer churn (CC) can bring certain economic losses to e-commerce platforms, so predicting the risk of CC has certain research value. To predict the CC in the e-commerce field, this research proposes the use of a radial basis function (RBF) neural network to build a prediction model. The RBF has strong nonlinear fitting ability and can quickly find the optimal value. However, due to the large cardinality and high complexity of customer data in the e-commerce field, to improve the prediction accuracy of the model, the study also uses the least absolute shrinkage and selection operator (Lasso) regression algorithm to optimize the RBF model. Lasso has efficient feature classification ability [3]. The combined model combining the RBF and Lasso regression has several advantages. First, the RBF neural network is very suitable for dealing with complex and high-dimensional data due to its strong nonlinear fitting ability and global approximation ability, which can quickly find the optimal solution. Second, e-commerce data usually contain a large number of redundant and irrelevant features, which can lead to high computational complexity and easy overfitting of the model. Lasso regression, on the other hand, can effectively perform

* **Corresponding author: Qiuping Lu,** Faculty of Economics and Trade, Henan Polytechnic Institute, Nanyang, 473000, China, e-mail: 2008011@hnpi.edu.cn

feature selection and reduce the dimensionality of the data by introducing L1 regularization, thus optimizing the performance of the RBF model. Compared to the traditional algorithmic model data prediction research, this study innovatively combines the concept of customer lifecycle in the field of e-commerce. An improved L-RBF model was used to analyze the correlation of CC characteristics at different stages of the cycle, to facilitate e-commerce stores to develop the corresponding strategies to recover customers with churn risk. Adopting the lifecycle concept can help models analyze the characteristics of data and make accurate predictions. At the same time, this study also integrates the customer lifecycle concept into the analysis of e-commerce CC for the first time and provides an in-depth analysis of the churn characteristics of different customer lifecycle stages through the improved L-RBF model. This approach not only enhances the accuracy of predicting the risk of CC but also provides e-commerce stores with targeted strategies to recover potential churn customers.

To complete the above research content, this article will be divided into five parts for discussion. The first part is a brief introduction to the research direction and content of the article. The second part is the research status of customer prediction and the RBF algorithm. The third part is the study of the L-RBF model in e-commerce customer prediction, which is divided into two sections. The first section is the optimization research of Lasso regression algorithm on the RBF model, and the second section is the application research of combining the L-RBF model with customer lifecycle. The fourth part is an analysis of the performance and e-commerce applications of the improved L-RBF model. This part is divided into two sections. The first section is the performance analysis of the L-RBF model, and the second section is the application effect analysis of the improved L-RBF model in e-commerce. The fifth part is a summary of the research content and shortcomings of the entire article.

2 Related works

The RBF algorithm is a neural network algorithm that can quickly approximate the optimal value for evaluation. This algorithm is widely used in different fields due to its ability to avoid falling into the local extremum and convenient operation. In the field of big data, Li et al. proposed an ELM-RBF (ELM = extreme learning machine) model to effectively classify a large number of mixed data. This model could classify and recognize data by calculating the distance between the center points of the data and RBF, and then achieving data processing and model fusion operations by training the weights of the data. The experimental outcomes indicated that the $F1$ value of the model has increased by 2.37%, indicating excellent classification performance [4]. Han et al. used the GHC algorithm to optimize the RBF model to improve the computational efficiency of parallel technology networks. The optimized model could perform classification operations on different subtasks of the parallel technology network. The corresponding RBF models were set based on different subtasks to improve the overall operational efficiency of the parallel technology network. After verification, the model had certain advantages [5]. Yang et al. proposed using the RBF algorithm to construct a rock plasticity model to predict the creep behavior of rocks. First, experiments were conducted on the silty sand of rocks to extract characteristic data, and then the dataset was processed using the constructed model to predict the creep behavior of rocks. The research findings indicated that the prediction error of the rock plasticity model was only 0.41%, and it had high prediction stability [6]. Liu et al. proposed two methods to apply the RBF algorithm to predict the dynamic deviation of robot manipulations. One was to add a payload to the RBF algorithm to improve the approximate domain range of the algorithm. Another approach was to reduce the dimensionality of the robot to reduce the computational efficiency of the RBF model. After simulation experiments, both of the above methods have been verified to be feasible [7]. Gopi et al. focused on the field of sentiment analysis, using classification techniques for opinion mining and scoring sentiment on a scale of −5 to +5. In the article, movie review data from Twitter between 2003 and 2012 were collected, and sentiment was scored using WordNet lexicon. The results of the study showed that the improved RBF model achieved 98.8% in accuracy, which is significantly higher than the existing RBF classifiers and other models [8]. Sarina and Tanniewa aimed to perform user reviews of Indonesia's TikTok Shop Seller Center application on Google Play Store through the support vector machine (SVM) and text mining techniques to perform a sentiment analysis. The study collected

data of Indonesian reviews between May and July 2023, including ratings, review content, and review date. Sentiment analysis was used to categorize the reviews as positive or negative, and the sentiment was classified using SVMs with different cores. The results of the study show that the SVM with the RBF kernel has better classification performance. The study can contribute to the reference value of sentiment analysis in the context of e-commerce in Indonesia [9].

In any business field, CC will bring certain economic losses, so the prediction of CC has always been a widely studied hot topic. Many scholars have adopted different methods to construct CC prediction models and applied them to different fields. Gattermann-Itschert used deep learning classification algorithms to construct a CC prediction model, which improved the prediction accuracy by training multiple time slice datasets. The experimental outcomes indicated that the CC prediction model trained through multiple slices could achieve dynamic prediction of CC [10]. Bhattacharyya and Dash designed a network graph approach for the construction of a CC prediction model in the telecommunications industry, which collected and summarized the dissatisfaction and churn reasons of telecommunications users through social media. The prediction model was trained based on the collected cause dataset to improve its prediction accuracy. This study provided strong data support for the construction of CC prediction models in the telecommunications industry [11]. Al-Najjar et al. used five algorithms, namely, the Bayesian network, C5 tree, cardholder automatic interaction detection tree, classification and regression tree, and RBF neural network, to predict the CC rate for bank credit card processing. The aim was to analyze the CC prediction model with the best performance. The experimental findings expressed that the prediction model constructed using the C5 tree and RBF neural network had better comprehensive performance [12]. Routh et al. proposed a competitive risk approach to reduce the limitations of the dataset on customer prediction models. This method could compare the relationship between customer behavior characteristics and churn risk and modeled it to provide scientific data support for customer prediction. The experimental outcomes expressed that the application of this method improved the prediction accuracy of the customer prediction model by 20% [13].

In recent years, China's cross-border e-commerce has been developing rapidly and is widely recognized as a trend in the country's foreign trade development. However, the accuracy of the most existing forecasting methods is usually limited due to the complexity of multiple influencing factors. Zhong et al. proposed a joint forecasting method combining the back propagation neural network (BPNN) model and the SVM. A case study using publicly available data from Hangzhou, China, shows that the relative error of this joint prediction method is less than 1%, which is smaller than the relative error obtained by using BPNN or SVM prediction alone [14]. Facing the challenge of CC, Huda et al. proposed a data mining model aimed at assisting e-commerce companies in developing customer retention strategies through customer behavior prediction. The model classifies customers by segmenting and categorizing them based on several variables such as session, application interaction, and purchase behavior. A clustering technique based on the recent frequency and amount model is used, focusing on the last visit time, visiting frequency, and total amount spent by customers. Comparing the two algorithms, decision tree and SVM, this method performs the best in terms of customer categorization accuracy at 87% [15]. To gain insight into foreign trade characteristics and the export volume calculation method, Dai used BPNN for prediction. BPNNs have unique and advanced advantages in solving nonlinear problems and are well suited for solving forecasting and decision-making problems related to nonlinear financial systems. By building multi-factor and single-factor export forecasting models, this study predicted the export volume of a Chinese city in recent years and compared it with the actual export volume. The results of the study show that the prediction accuracy of the designed model is more than 30% higher than that of the traditional prediction method, and the accuracy of the application is also about 15% higher than that of the traditional method [16].

In summary, while the RBF algorithm has been extensively employed across various domains due to its remarkable optimization and classification capabilities, its application within the realm of e-commerce CC prediction remains significantly underexplored. This study bridges this gap by leveraging an improved RBF model, which ingeniously integrates the Lasso regression algorithm, enhancing the model's predictive accuracy and efficiency in processing large-scale e-commerce data sets. Notably, the inclusion of the customer lifecycle concept represents a pioneering approach within this field, offering a nuanced understanding of churn dynamics across different customer stages. This innovative integration enables the identification of

stage-specific churn predictors, thereby providing e-commerce platforms with tailored strategies for customer retention. The contributions of this research lie not only in the advancement of churn prediction methodologies through the application of the improved RBF model but also in the novel application of lifecycle segmentation, enriching the analytical depth and operational relevance of churn prediction models in the e-commerce context.

3 E-commerce CC prediction model based on the L-RBF model

There are various types of e-commerce platforms, with a huge customer base and a wide range of features. To control the CC rate of e-commerce platforms, this study proposes using the RBF model to predict the CC risk of e-commerce platforms. However, the original customer dataset contains a large number of irrelevant features. To improve the prediction accuracy of the RBF model, the Lasso regression algorithm is used to improve the RBF model. At the same time, considering the lifecycle of e-commerce customers, the characteristics of the lifecycle are integrated into the L-RBF model for accurate prediction of CC, aiming to provide a basis for e-commerce enterprises to develop retention strategies.

3.1 Construction of the RBF model improved by the Lasso regression algorithm

In the field of e-commerce, the diversity, dynamics, and large-scale characteristics of customer data can significantly affect the performance and accuracy of the model. To better predict CC in e-commerce, this study builds an e-commerce CC prediction model using an improved RBF network. The impact of data diversity on model performance is first considered. E-commerce customer data cover a wide range of consumption behaviors, preferences, and transaction histories. This data diversity requires the model to be able to effectively process different types of data and extract valuable information from them. The proposed L-RBF model aims to better adapt to this diversity through a flexible network structure and optimization algorithms, thus improving the accuracy of prediction. Next, the impact of data dynamics on model performance is considered. E-commerce data change dynamically rapidly and their customer behavior and market trends evolve continuously [17]. The L-RBF model is optimized using Lasso regression, which aims to enhance the model's adaptability to new data and learning speed, so as to respond to the dynamic changes in the data more effectively. Finally, the large-scale characterization of data is considered. In e-commerce, various types of models are often required to handle large-scale customer data, which challenges the computational efficiency and processing capability of the models. The L-RBF model optimizes the computational process and reduces unnecessary data processing with a view of effectively reducing time complexity while maintaining high prediction accuracy.

The RBF, as an efficient feed-forward neural network, has a strong global approximation capability and can accurately represent any nonlinear function. Therefore, this neural network is suitable for e-commerce domains with large amount of data [18]. In addition, RBF has strong generalization ability and can map higher features, so it can also be used for the dynamic prediction of e-commerce CC. The standard RBF model has issues with overfitting, high computational complexity, parameter selection, and sensitivity to noise. The study introduced least absolute reduction and Lasso regression to optimize it. First, the RBF model often overfits with high-dimensional data due to its complexity, leading to poor test data generalization. Second, calculating distances for each sample point to all RBF centers is computationally heavy and time-consuming. Third, selecting key parameters is difficult and typically requires multiple cross-validation trials, adding complexity. Finally, RBF models are sensitive to noisy data, common in e-commerce, which reduces the prediction accuracy. The general structure of RBF is shown in Figure 1.

Figure 1 shows the structural diagram of the RBF. It is divided into three layers: input, hidden, and output layers. The amount of nodes in the input layer depends on the dimensionality of the input sample features. The function of the hidden layer is to perform radial transformation on the input feature data, transform the

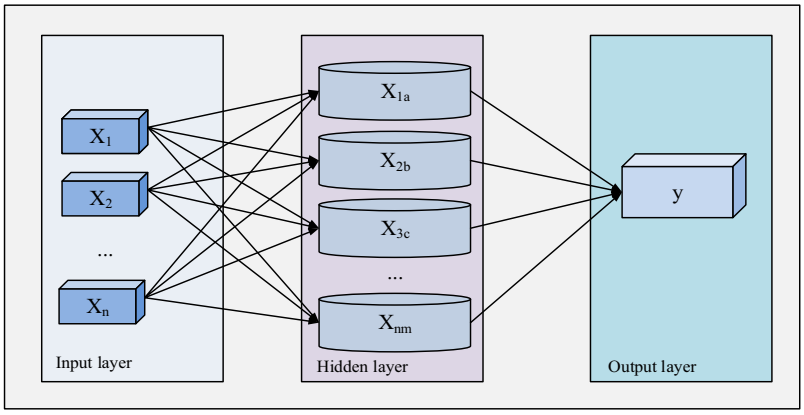

Figure 1: Structure diagram of the RBF neural network.

spatial dimension of the features, and transform them into linearly separable states in high-dimensional space. In the input layer, it randomly inputs data X_n into the input layer, then randomly selects m center points, and map the input data to the hidden layer. Finally, it calculates the output calculation result through the hidden layer function and outputs the output layer output result y. In the hidden layer, the RBF of neurons is a Gaussian kernel function, and its expression is shown in formula (1) [19].

$$G(\|X - C\|) = \exp\left(-\frac{\|X - C\|^2}{2\sigma^2}\right). \tag{1}$$

In formula (1), C is the central value of RBF. σ means the expansion coefficient. $G(*)$ denotes a Gaussian kernel function. X indicates the input characteristic value. Since the operation process of RBF neural network is actually the weighted operation of hidden layer neurons, to avoid the extreme phenomenon of RBF, the definition expression of the expansion coefficient σ is shown in formula (2).

$$\sigma = \frac{d_{\max}}{\sqrt{2 \times ck}}. \tag{2}$$

In formula (2), $d_{\max}$ is the maximum distance of the center point of the RBF. ck expresses the amount of center points. The final layer of the RBF neural network is the output layer, which is used to weigh and calculate the neurons in the hidden layer. Its weight values are generally within the range of [0,1]. The output value of the RBF network is indicated in formula (3).

$$y_i = \sum_{i=1}^{k} w_{ij} \exp\left(-\frac{1}{2\sigma^2} \|x_p - c_j\|^2\right). \tag{3}$$

In formula (3), y_i is the ith neuron of the output layer's output feature. k is the number of input features. w_{ij} is the weight between the ith and jth neurons. x_p is the pth sample feature of the input. c_j is the jth center point. From formula (3), the operational efficiency of RBF is influenced by factors such as input feature numbers, center points, and expansion coefficients.

The data characteristics of e-commerce customers are relatively complex, including much data not related to CC prediction, coupled with the fact that the traditional RBF network faces the challenges of complex operation and difficult parameter selection when dealing with large-scale e-commerce data. To solve this problem, this study introduces the Lasso regression algorithm to improve the traditional RBF network and finally designs the L-RBF model. Lasso regression is able to effectively reduce the complexity of the model by introducing a regularization term and helps to prevent overfitting. In the L-RBF model, the centroid selection and weight allocation of the RBF network are optimized by the Lasso algorithm, which improves the model's ability to process big data and its prediction accuracy. In addition, the L-RBF model utilizes the sparse nature of Lasso regression to automatically select the most informative features, which further improves the efficiency and accuracy of the model in large-scale e-commerce data processing.

The Lasso regression algorithm is based on least squares and solves the problem of linear overfitting by adding a norm. This algorithm can simplify excess feature items to 0 based on the obtained feature coefficients, reducing the computational complexity of data operations. It assumes that there is a dataset called (X_i, y_c), where the expression for X_i is shown in formula (4).

$$X_i = (x_{i1}, \ x_{i2}, \ \ldots, \ x_{iq})^T \quad i = 1, \ 2, \ \ldots, \ N. \tag{4}$$

Here, N stands for the amount of samples. y_c in (X_i, y_c) refers to the output value of different category features, where c has a value range of $[1, M]$. M means the amount of output categories of sample features. The linear regression model expression of y is shown in formula (5).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q + \varepsilon. \tag{5}$$

In formula (5), β_0 expresses a constant. $\beta_1, \ \beta_2, \ \ldots, \ \beta_q$ means the linear regression coefficient. ε refers to the penalty term. It performs the Lasso algorithm estimation operation on it, and the expression is shown in formula (6).

$$\hat{\beta}^{\text{Lasso}} = \arg\min \sum_{i=1}^{N} \left(y_c - \beta_0 - \sum_{i=1}^{p} x_{ip} \beta_q \right)^2. \tag{6}$$

In formula (6), β stands for the average of all linear regression coefficients. Extracting non-zero coefficients with high stability in formula (6) can obtain feature points related to the target data, achieving the effect of dimensionality reduction on their features. At the same time, the regression formula of Lasso can be obtained as shown in formula (7).

$$y^{\text{Lasso}} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k. \tag{7}$$

In formula (7), $\beta_0, \ \beta_1, \ \ldots, \ \beta_k$ is a non-zero constant. According to the Lasso regression algorithm mentioned above, target features can be extracted from the original e-commerce customer dataset to obtain the dimensionality reduced dataset. The dimensionality reduced dataset can also optimize the parameter selection process of RBF, obtain suitable center points, and improve the operational accuracy of RBF models.

Figure 2 shows the operation steps of the RBF model under the Lasso regression algorithm. Lasso regression enhances the model performance by using L1 regularization to select important features and reduce redundancy. In e-commerce, customer behavior data are complex with many features, but only some significantly impact churn prediction. Lasso regression filters critical features like purchase frequency, spending amount, and browsing time, which more accurately reflect the customer behavior and potential churn risk.

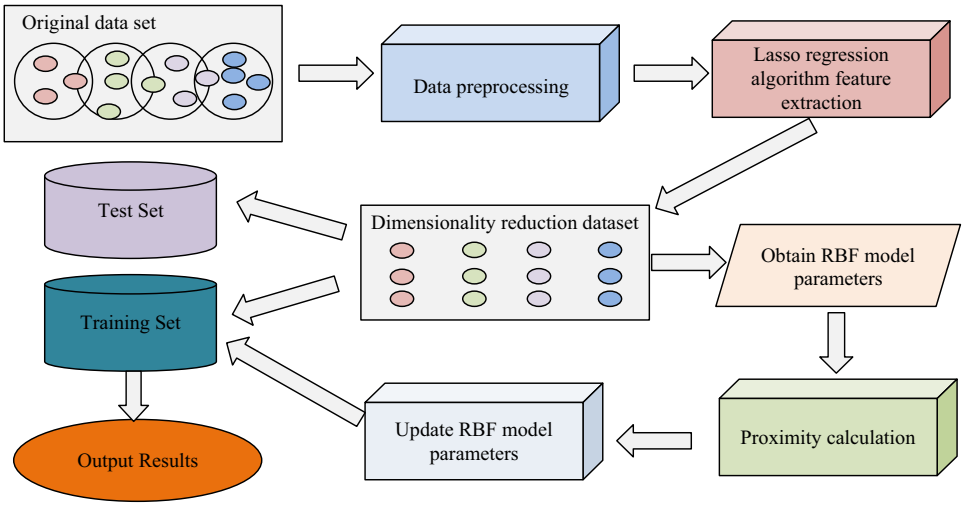

Figure 2: RBF model operation steps under the Lasso regression algorithm.

This improves the model's predictive accuracy and robustness. There are multiple features in the original dataset. They are input into the Lasso regression algorithm as input values, and their dimensions are reduced to obtain a dataset containing only the target features. By substituting this dataset into the RBF for operation, the operational parameters of the RBF, including the amount of center points in the hidden layer, can be obtained. The proximity to a predicted value of 1 can be calculated based on the center point of the hidden layer. The higher the proximity, the higher the fitting accuracy of the Gaussian kernel function at the center point, and the more accurate the operation results of the RBF. The expression for the proximity of the center point is shown in formula (8).

$$p = \frac{1}{1 + e^{-(y^{\text{Lasso}(c)})}}. \tag{8}$$

In formula (8), p is the proximity of the center point to the predicted value 1. After obtaining the appropriate operating parameters of the RBF, the RBF model is updated. It trains and calculates the dimensionality reduced training set based on the updated RBF model to obtain feature output values. The output value is compared with the expected value, and if the output value is less than the expected value, the predicted result is loss. If the output value is greater than the predicted value, it is predicted as non-loss.

3.2 L-RBF e-commerce CC prediction model based on lifecycle

The purchasing needs of customers in e-commerce enterprises vary over different periods, resulting in dynamic changes in their feature datasets. To improve the accuracy of L-RBF prediction, this study combines the concept of customer lifecycle with the L-RBF prediction model. The combined model can dynamically predict CC rates in different periods based on changing customer feature datasets [20,21]. The customer lifecycle refers to the development process from the beginning of consumption to the end of consumption by customers on e-commerce platforms.

Figure 3 shows a schematic diagram of the customer lifecycle of an e-commerce platform. On e-commerce platforms, the customer's lifecycle shows a trend of segmented changes. The first stage is the exploration stage, where the customer type is potential customers who have not yet collaborated but are willing to consume. The second stage is the advanced stage, where potential customers are successfully discovered and become new customers for e-commerce enterprises. The third stage is the stable stage, which refers to the customer's consumption level and willingness in the e-commerce enterprise being relatively stable. The fourth stage is the recession stage, which refers to customers losing their willingness to consume on the platform. The fifth stage is the churn stage, where customers lose interest in the platform's products and no longer pay attention to the company's products. So, to reduce the proportion of lost customers, the risk of customer loss should be predicted before the loss stage.

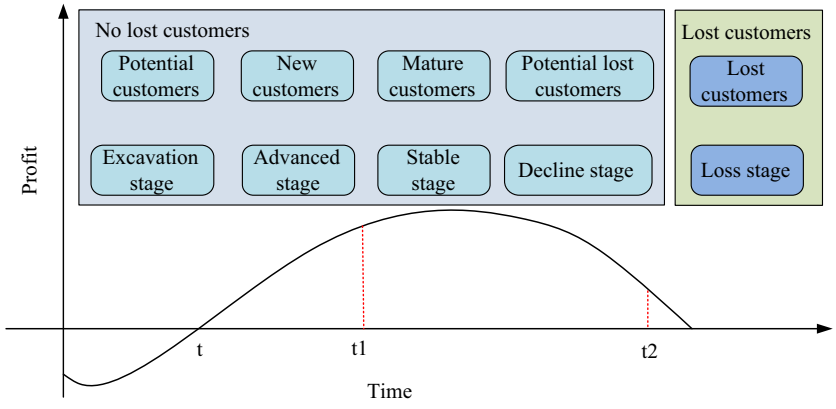

Figure 3: Customer lifecycle diagram of the e-commerce platform.

Figure 4 shows a schematic diagram of L-RBF e-commerce CC prediction based on lifecycle. First, it obtains the customer's dataset from the data management department of the e-commerce enterprise platform and then divides the dataset into cycles based on its lifecycle characteristics. Datasets from different periods are utilized as input values for the L-RBF model prediction model and then it determines whether customers have a tendency to churn based on the output results. As customers have already undergone periodic classification, when analyzing the reasons for churn, classification analysis can also be conducted based on the characteristics of customers in different periods, to develop the corresponding strategies to retain customers at risk of churn. It assumes that the original data set of customers in the exploration, advanced, stable, and declining stages is as shown in formula (9) [22,23].

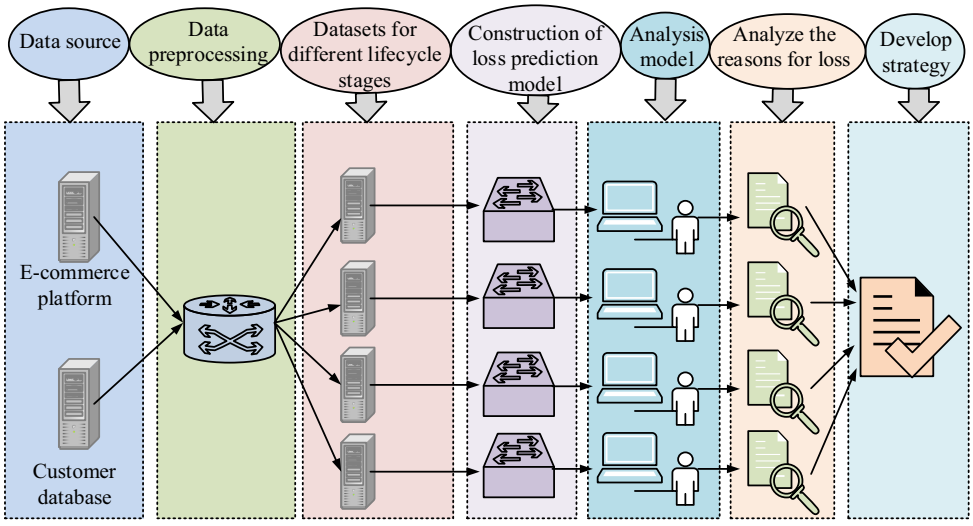

Figure 4: Schematic diagram of L-RBF e-commerce CC prediction based on lifecycle.

$$Z = \{z_1,\ z_2,\ \ldots,\ z_n\}. \tag{9}$$

In formula (9), n is the data feature dimension, and the original feature stages of the four stages are the same. However, the customer demand characteristics in the four stages are inconsistent, so there are certain differences in the regression formulas constructed by using the Lasso algorithm to extract features. The Lasso regression formula for the four lifecycles is denoted in formula (10).

$$\begin{cases} y_A = a_0 + a_1x_{a1} + a_2x_{a2} + \ldots + a_ix_{ai}, \\ y_B = b_0 + b_1x_{b1} + b_2x_{b2} + \ldots + b_jx_{bj}, \\ y_C = c_0 + c_1x_{c1} + c_2x_{c2} + \ldots + c_mx_{cm}, \\ y_D = d_0 + d_1x_{d1} + d_2c_{d2} + \ldots + d_hx_{dh}. \end{cases} \tag{10}$$

In formula (10), y_A, y_B, y_C, and y_D are customer dimensionality reduction datasets for the mining, advanced, stable, and declining stages, respectively. i, j, J, and h are the number of features, with values less than or equal to the data feature dimension n. a_i, b_j, c_m, and d_h are non-zero regression coefficients. The above content can extract customer features from different lifecycles, and different L-RBF prediction models can be constructed based on the feature set.

Figure 5 shows the schematic diagram of the L-RBF model training process based on the lifecycle. The Lasso regression algorithm can extract features based on the characteristics of different lifecycles. The L-RBF model can obtain the center points of different models based on different feature sets and construct different prediction models based on the obtained center points. According to the L-RBF prediction process, customers with loss risk can be predicted. To deeply analyze the reasons for CC, this study introduces the correlation coefficient R to represent the correlation between the features and churn degree, which is expressed in formula (11).

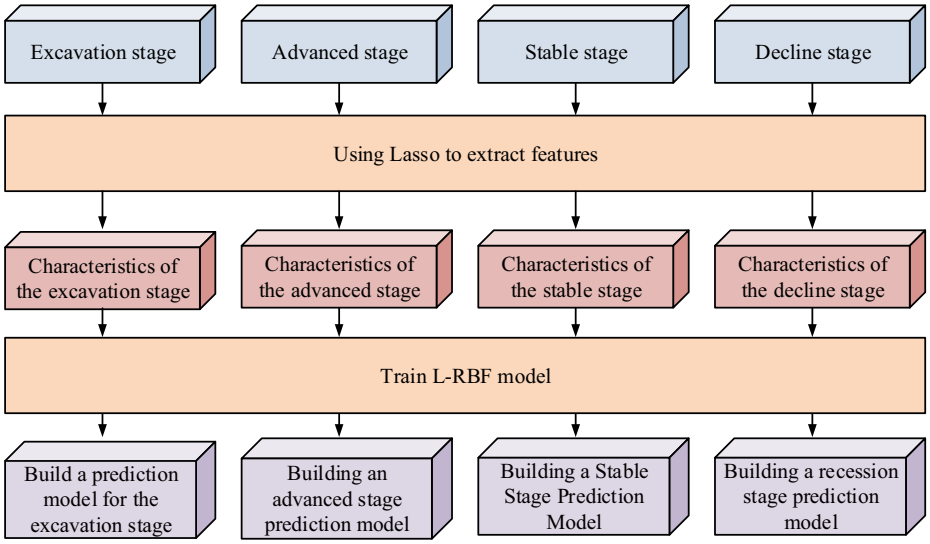

Figure 5: Schematic diagram of the L-RBF model training process based on lifecycle.

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}. \tag{11}$$

In formula (11), $\bar{x}$ and $\bar{y}$ are the average values of the loss rates of the original feature set and the feature set, respectively. The criteria for determining correlations are indicated in formula (12).

$$R = \begin{cases} 0 \le |R| \le 0.2, \\ 0.2 \le |R| \le 0.4, \\ 0.4 \le |R| \le 0.6, \\ 0.6 \le |R| \le 0.8, \\ 0.8 \le |R| \le 1.0. \end{cases} \tag{12}$$

In formula (12), $|R|$ means extremely weak, weak, moderate, strong and extremely strong correlations between the ranges of [0, 0.2], [0.2, 0.4], [0.4, 0.6], [0.6, 0.8], and [0.8, 1.0], respectively. E-commerce platforms can analyze the reasons for CC based on their characteristics and the degree of correlation with CC and develop retention strategies for customers at risk of churn.

When making predictions about e-commerce data, attention should also be paid to customer privacy and the handling of sensitive personal data. This study employs four measures to ensure security and compliance when handling sensitive personal data in e-commerce. The first is data anonymization and de-identification. When processing customer data, data anonymization and de-identification are performed first to ensure that personal information cannot be identified directly or indirectly, a step that helps reduce the risk of personal privacy breaches. This is followed by compliance with data protection regulations. When conducting predictive analytics on CC data, there is a need to strictly comply with relevant data protection laws and regulations to ensure that all data processing activities comply with legal requirements. Next is the need to restrict data access. It should strictly limit access to sensitive data to authorized researchers and necessary system administrators. In addition, encryption is used to protect stored and transmitted data. Finally, regular security audits are conducted. Regular security audits and privacy impact assessments are conducted to identify and address any potential privacy and security issues.

In this study, ethical and legal issues also need to be given high priority when using AI and machine learning techniques for e-commerce customer data analysis. First, strict compliance with data protection laws and regulations, such as the EU General Data Protection Regulation, during data collection and processing, ensures the protection of customer privacy and sensitive information. Second, considering the issues of bias and discrimination that may arise from machine learning algorithms, this study takes appropriate measures to

identify and mitigate these potential risks during the model design and implementation stages. In addition, this study focuses on the transparency and interpretability of the models so that users and regulators can understand the decision-making process of the models. Finally, this study emphasizes the continuous monitoring and evaluation of model performance in real-world applications to ensure that it meets ethical standards and legal requirements. Through the above measures, the study aims to ensure that the application of AI and machine learning in e-commerce customer data analytics is both efficient and responsible while complying with ethical and legal requirements.

4 Performance and application analysis of the e-commerce CC prediction model based on improved L-RBF

According to the characteristics of big data of e-commerce customers, this article proposed the optimization of the RBF prediction model with the Lasso regression algorithm to improve the prediction performance of this model in e-commerce customers. To verify the performance of the model, this study used the other three algorithms as comparative algorithms for model performance analysis. At the same time, the study also applied the model to predict churn at different stages of the customer lifecycle and analyzed the actual application performance of the model based on prediction accuracy and feature correlation.

4.1 Performance analysis of the L-RBF CC prediction model

To improve the overall performance of e-commerce customer prediction models, this study proposed using the Lasso regression algorithm to optimize the RBF model. To address the model bias and discrepancy, several steps can be taken. First, comprehensive pre-processing, including data cleaning, outlier detection, and normalization, reduces noise and improves model robustness and accuracy. Second, Lasso regression for feature selection filters out the most influential features, excluding redundant and irrelevant ones. Finally, cross-validation divides the dataset into training and validation sets, allowing effective model evaluation and avoiding bias from uneven data distribution.

To verify the performance of the optimized L-RBF model, the traditional RBF model, the original logistic regression algorithm, and the gradient boosting decision tree (GBDT) were used as comparison algorithms. Among them, the original logistic regression algorithm is an iterative classification algorithm used to solve the optimal value, and the GBDT is an integrated learning algorithm used to reduce the deviation, which can classify and predict the data according to the weight. The above three methods have varying degrees of application in the field of loss prediction. The data set used in this experiment was the customer information set of a clothing store in an e-commerce platform, which contained 100,000 pieces of data. The above four

Table 1: Experimental environment details

Type	Parameter	Specifications
Hardware environment	CPU version	4 Intel(R) Xeon(R) CPU E7-4850 v3@2.20 GHz
	CPU cores	5
	Memory	32G
	Capacity	320G
Software environment	Operating system	Windows 10
	System bits	64位
	Editing tools	Python3.6
	Language type	C++

algorithms were used to analyze the data set, and the performance of the algorithm was compared and analyzed according to the accuracy, recall rate, receiver operating characteristic (ROC) area, and iteration error value of the algorithm.

Table 1 shows the experimental environment details. The experimental computing environment was 64 bit Windows 10. The customer prediction model was constructed using the L-RBF, RBF, original logistic regression algorithm, and GBDT algorithm. The performance of the algorithm was analyzed according to the performance of the computing process and the results.

Figure 6 shows the PR curves under different algorithms. From Figure 6, it can be observed that RBF, original logistic regression, and GBDT algorithms had different cross positions, which showed that at different recall rates, the accuracy of the three algorithms had their own advantages. The L-RBF model proposed in the study outperformed the other three algorithms in accuracy at any recall rate, achieving a maximum accuracy of 95%. The comprehensive performance of the L-RBF algorithm was superior to the other three algorithms. Therefore, the L-RBF algorithm proposed in the study had better comprehensive performance.

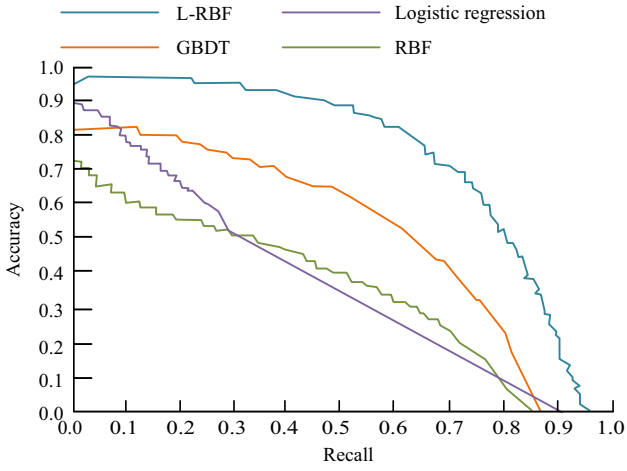

Figure 6: PR curves under different algorithms.

Figure 7 shows the ROC under different algorithms. The models used in Figure 7(a)–(d) were L-RBF, GBDT, original logistic regression, and RBF models. From Figure 7, the ROC of the L-RBF model was closer to the upper left corner, and its AUC was larger, followed by the GBDT model, then the original logistic regression model, and finally the RBF model. It showed that the L-RBF model could better classify the features related to CC from the original data set, which was conducive to the subsequent prediction of CC. In summary, the L-RBF model had better performance in classifying CC features.

Figure 8 shows the loss rate curves of prediction models under different algorithms. From Figure 8, as the amount of customer data processed by the four algorithms increased, the overall loss rate of the algorithms was on a downward trend, and finally tended to stabilize. Among them, the overall loss rate of the L-RBF algorithm was smaller and tended to stabilize the fastest. When the amount of customer data reached 100,000, the loss rate of L-RBF algorithm started to converge. The final loss rate was 3%, lower than 4% of GBDT algorithm, 5% of original logistic regression algorithm, and 6% of RBF algorithm. In summary, the L-RBF algorithm had a higher algorithm loss rate and could more effectively predict CC.

Figure 9 shows the time complexity of prediction models under different algorithms. As the number of processed samples increased, the time complexity of all four models increased. Among them, under the same sample size, the time complexity of the L-RBF model was the lowest among the four models. When the number of processed samples was 20,000, the L-RBF model's $O(2^n)$, $O(n^3)$, $O(n^2)$, $O(n \log 2^n)$, $O(n)$, $O(\log 2^n)$, and $O(1)$ were 142, 86, 47, 34, 30, 17, and 3, respectively, followed by the original logistic regression model, RBF model, and GBDT model. Because the Lasso regression algorithm in the L-RBF model improved the RBF model by

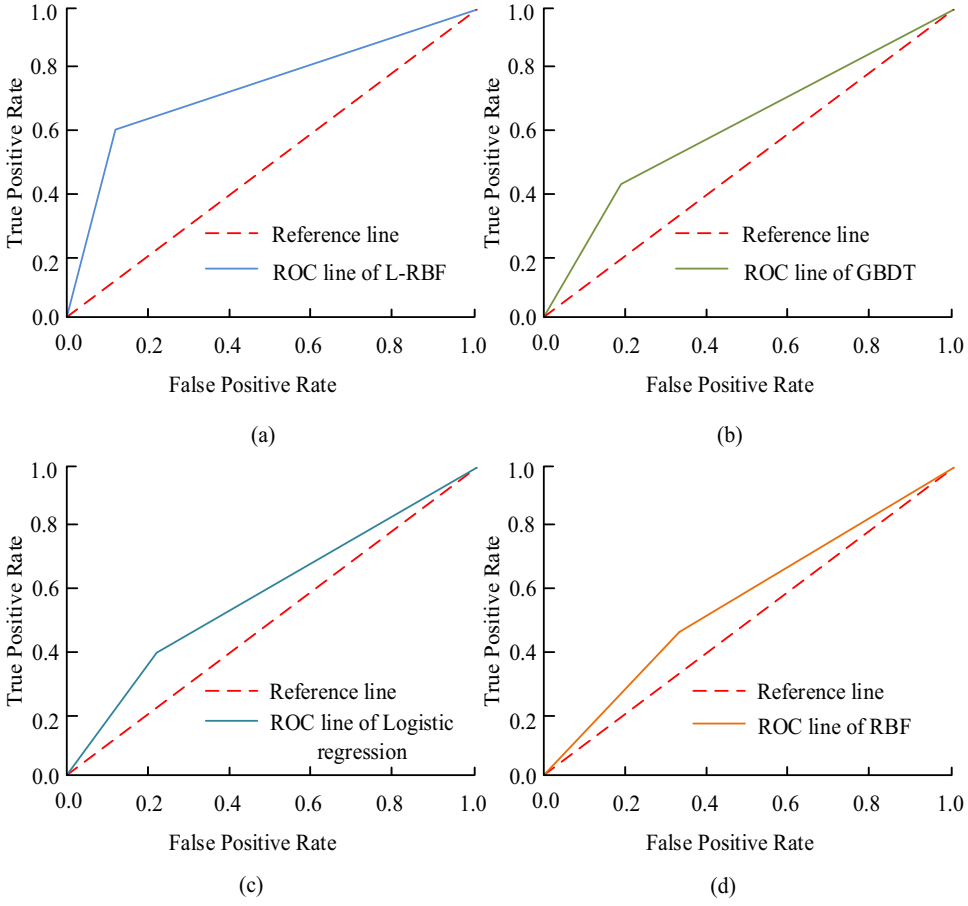

Figure 7: ROC under different algorithms. (a) ROC of L-RBF algorithm, (b) ROC of GBDT algorithm, (c) ROC of logistic regression algorithm, and (d) ROC of RBF algorithm.

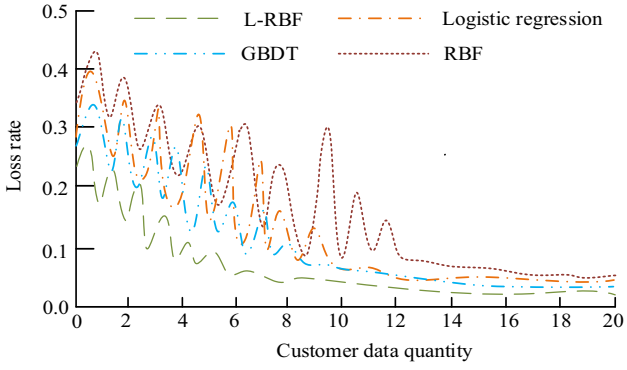

Figure 8: Loss rate of prediction models under different algorithms.

simplifying the feature coefficients and reducing the computational complexity of data operations, the time complexity during the operation process would also be correspondingly reduced. Therefore, the improved L-RBF model in this study had better time complexity.

In order to demonstrate the statistical significance of the L-RBF model in benchmarking performance tests, the study statistically analyzed the results of several repeated experiments, which are shown in Table 2.

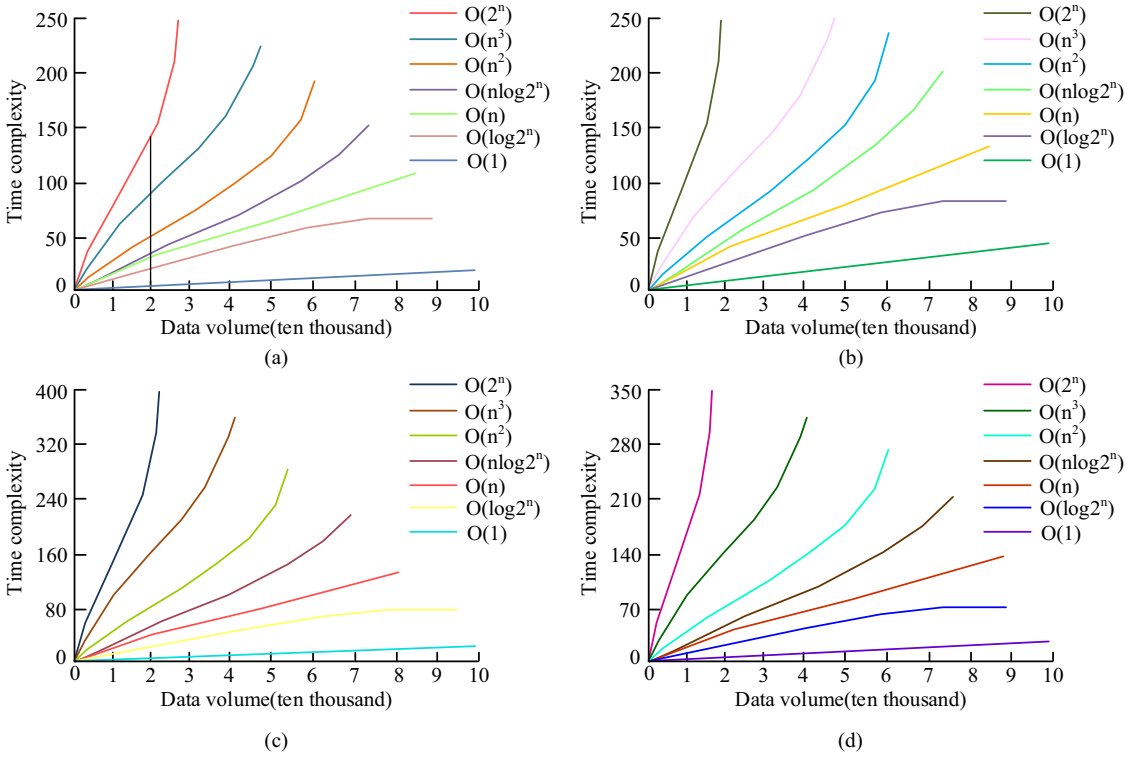

Figure 9: Time complexity of prediction models under different algorithms. (a) L-RBF, (b) logistic regression, (c) GBDT, and (d) RBF.

The statistical test results of the four models under multiple experiments are given in Table 2. As can be seen from Table 2, L-RBF has the best performance in average accuracy, average AUC, and average $F1$ score compared to the other three models, which are as high as 95.0%, 0.96, and 0.94, respectively. In addition, paired t-tests were also conducted to test the performance metrics of the L-RBF model against the other models, and the results show that the L-RBF model is statistically significantly better than the other models ($P < 0.05$). Through the above statistical tests, the validity and robustness of the L-RBF model under multiple repetitive experiments were finally verified.

4.2 Performance analysis of improved L-RBF's e-commerce CC model application

To make the L-RBF prediction model applicable to the field of e-commerce, this study further improved the L-RBF model by combining customer lifecycle concepts. To verify the application performance of the L-RBF model combined with the concept of customer lifecycle in e-commerce CC, the research would still use the dataset in Section 3.1, with the feature dimensions shown in Table 3.

Table 3 shows the feature dimensions of the experimental customer dataset for clothing shops, which include the physiological index, daily consumption level, dressing style, platform activities, and shop purchase. Sensitivity analysis reveals that daily consumption level has the highest importance (0.89), followed by shop purchase (0.85), platform activities (0.78), physiological index (0.72), and dressing style (0.64). Physiological index refers to the match between customers' height and weight with clothing style. Daily consumption level assesses the similarity to shop prices. Dress style consistency, platform usage frequency, and shop purchase frequency also impact customer stickiness to the shop, influencing their life cycle characteristics. Because the number of features was 5, which means that the dimension of the data set was 5, the number of input neurons of the L-RBF model was determined to be 5. There were 1,524 missing data in the data set, and 70,000 data were randomly selected from the remaining non-missing data for detection. Because the number of core points was

Table 2: Statistical test results of different models

Model	Average accuracy (%)	Average AUC	Average *F*1 score	Confidence interval for mean accuracy (%)	Confidence interval for mean AUC	Confidence interval for mean *F*1 score	*P*
RBF	82.2	0.84	0.83	81.3–82.5	0.81–0.85	0.80–0.84	<0.05
Logistic regression	84.5	0.87	0.85	83.4–84.7	0.85–0.89	0.82–0.86	>0.05
GBDT	87.8	0.91	0.87	86.8–87.9	0.88–0.92	0.85–0.88	>0.05
L-RBF	95.0	0.96	0.94	94.5–95.5	0.95–0.98	0.93–0.95	>0.05

Table 3: Characteristic dimensions of experimental customer dataset in the cloth shop

Tag name	Feature name	Description of correlation degree	Sensitivity score
S1	Physiological index	Mainly whether the height, weight, and age match the clothing style of the store	0.72
S2	Daily consumption level	The correlation between the average consumption level of the customer on the platform and the clothing prices in the store	0.89
S3	Dressing style	The similarity between the main clothing styles purchased by customers on the platform and the clothing design styles of the store	0.64
S4	Platform activities	How often customers use this platform	0.73
S5	Shop purchase	The number of times customers have made purchases in this store	0.85

usually 10% of the sample data, there were 14,000 core points in this L-RBF model. The study would use an L-RBF model that combined the customer lifecycle to construct a corresponding cycle churn prediction model for analysis. The application performance of the model would be analyzed based on the prediction accuracy and error rate of the four models, as well as the correlation between features and CC at different stages.

In order to realize the conversion processing between different life stages, the L-RBF model incorporates the following improvements into the life cycle features. First, the dynamic feature update, the improved L-RBF model will periodically update the customer feature data to identify changes in customer behavior. By monitoring customer behavior in real time, the model can dynamically adjust the feature weights to ensure that the model can adapt to new data patterns during the life cycle transition. Next is adaptive learning. The improved L-RBF model adopts an adaptive learning mechanism, which is able to adjust itself during the

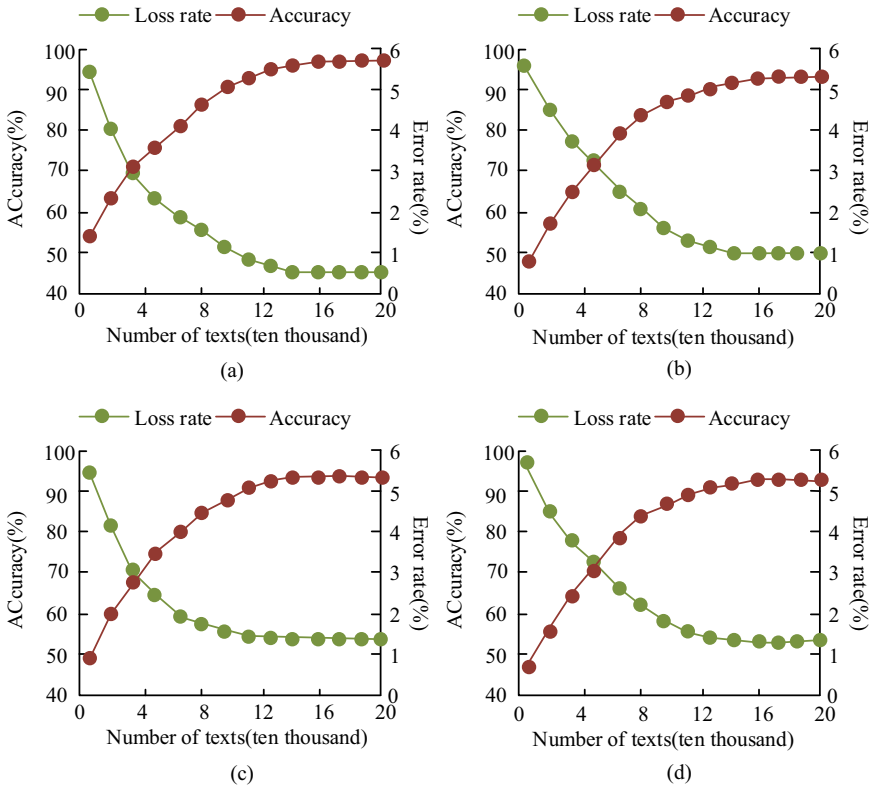

Figure 10: Improved L-RBF algorithm for predicting the accuracy and loss rate curves in different customer lifecycles. (a) Improving the prediction accuracy and error rate of L-RBF model in the excavation stage, (b) improving the prediction accuracy and error rate of L-RBF model in the advanced stage, (c) improving the prediction accuracy and error rate of L-RBF model in the stable stage, and (d) improving the prediction accuracy and error rate of L-RBF model in the decline stage.

customer life cycle transition. Through continuous learning and updating, the model is able to quickly adapt to new customer behavioral characteristics, ensuring that the prediction accuracy is not affected by the life cycle transition. Through the above methods, the improved L-RBF model can effectively handle the transitions between customer life cycle stages and ensure the prediction accuracy at different stages, and the specific results are shown in Figure 10.

Figure 10 shows the prediction accuracy and loss rate curves of the improved L-RBF algorithm in different customer lifecycles. Figure 10(a)–(d) represents the stages of customer lifecycle excavation, advanced, stable, and decline, respectively. The improved L-RBF model had high prediction accuracy at different customer lifecycle stages, and both accuracy and loss rates tended to stabilize with the increase of customer data quantity. The improved L-RBF model could achieve stable accuracy rates of 97.6, 93.1, 92.7, and 91.8% in the excavation, advanced, stable, and decline stages, respectively, with stable loss rates of 0.5, 1, 1.3, and 1.3%, respectively. Therefore, the improved L-RBF model could accurately predict the loss of customers at different stages of their lifecycle.

Figure 11 shows the correlation between different feature types and CC. From Figure 11, different types of features had different correlations with CC at different stages of the customer lifecycle. In the stages of excavation, advanced, stable, and decline, the characteristic indicators with a higher correlation with CC were physiological indicators, dressing style, platform activity, and store purchase frequency, with correlation indicators of 0.75, 0.58, 0.64, and 0.83, respectively. According to formula (12), the correlation degrees of the above correlation indicators were strong, moderate, strong, and extremely strong correlations, respectively. Therefore, based on the improved L-RPF model, features related to CC at different stages could be obtained, which helped platform stores develop corresponding strategies based on the obtained correlation features.

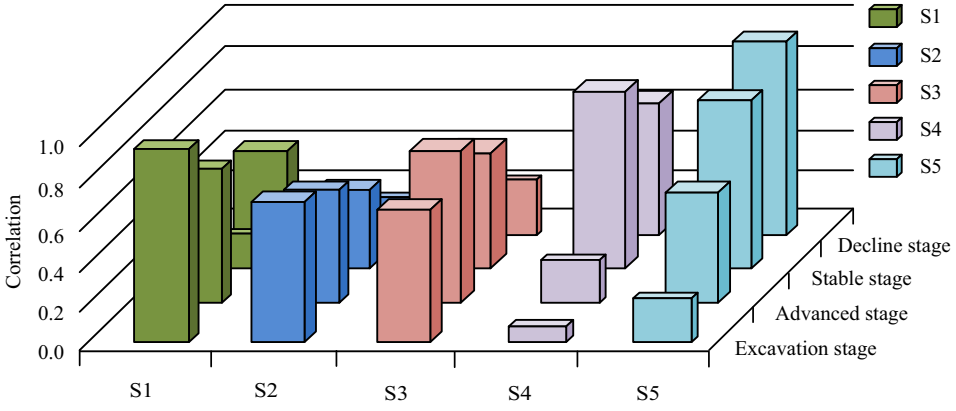

Figure 11: Correlation between different feature types and CC.

To further demonstrate the effectiveness of the L-RBF model incorporating the concept of customer lifecycle in e-commerce CC prediction, this study applied L-RBF, RBF, logistic regression, and GBDT to the actual e-commerce CC prediction task, respectively. A home-made large and diverse real-time e-commerce dataset was used for testing, and the prediction results of the four models for CC were further obtained, as shown in Table 4.

Table 4: Effectiveness of different e-commerce CC prediction models in practice

Model	Average prediction accuracy	Average prediction time (s)	Mean square error
L-RBF	0.93	0.13	1.87
RBF	0.84	1.06	8.98
Logistic regression	0.82	0.87	11.15
GBDT	0.87	0.58	5.24

The performance of the four e-commerce CC prediction models in real applications is given in Table 4. As shown in Table 4, the average prediction accuracy, average prediction time, and mean square error values of L-RBF were 0.93, 0.13 s, and 1.87 in the actual prediction task, those of RBF were 0.84, 1.06 s, and 8.98, those of Logistic regression were 0.82, 0.87 s, and 11.15, and those of GBDT were 0.82, 0.87 s, and 11.15 time, respectively. By combining the three indexes, it can be found that the L-RBF model proposed in this study showed a better performance of the practical application.

5 Discussion

In this study, an improved L-RBF model was constructed for the CC in the field of e-commerce. This model combined the characteristics of customer's whole lifecycle and deeply analyzed the reasons for CC. Merchants can make recovery strategies according to the reasons for CC. At present, there are also related studies to explore the customer purchasing tendency in marketing, and Chou and others have constructed a BG/BB model. The model combined Lasso regression and can predict the customer's buying behavior hypothesis through low-dimensional input data. In the experiment, 100 predictors were selected to predict the customer behavior, and the experimental results showed that the prediction accuracy of the model was about 90% [24]. Zhai et al. used an LSTM network to build a CC prediction model. The model predicted the interest preference according to the historical behavior of customers, and the prediction accuracy was also about 90% [25]. Therefore, the model constructed in this study showed excellent performance. At the same time, studies of Gattermann-Itschert and Thonemann, Bhattacharyya and Dash, AL-Najjar et al., and Routh et al. [10–13], respectively, showed the performance of CC prediction models under various algorithms such as the deep learning classification algorithm, network diagram, Bayesian network, C5 tree, chi-square automatic interactive detection tree, classification and regression tree, and RBF neural network. The experimental results showed that the RBF neural network had better performance. The minimum accuracy of the L-RBF model proposed in this study was 91.8%, which had better performance. Because the traditional RBF algorithm had the disadvantage of complex operation, the improved L-RBF algorithm based on Lasso regression could simplify the operation process and reduce the probability of error by calculating the distance of the fitting curve and sorting the data. Therefore, the L-RBF model proposed in this study had better performance.

Deploying this model in real e-commerce platforms presents challenges. First, distributed computing and cloud storage can improve the data processing efficiency through data slicing and parallel processing. Second, optimizing the computational efficiency and using faster algorithms and hardware can meet real-time requirements. Adhering to data privacy regulations is essential, ensuring data security through encryption, access control, and regular audits. Finally, as the e-commerce environment changes rapidly, models need regular updates to adapt to new trends and behaviors. Establishing an automated model updating and validation mechanism ensures continuous learning, maintaining the model accuracy and robustness. To integrate the L-RBF model with existing e-commerce systems, several steps are needed: data interface development, data pre-processing, model deployment, management, and real-time prediction and feedback. Three challenges are expected. First, developing flexible data conversion modules to handle different e-commerce data formats. Second, optimizing data interfaces and processing pipelines to maintain a stable performance under high load. Finally, adhering to data privacy regulations to ensure secure data transmission, storage, and processing.

6 Conclusions

Due to the fact that CC can bring certain economic losses to e-commerce enterprises, predicting the risk of CC has certain research significance. To predict e-commerce CC, this research proposed the use of the RBF algorithm to build a prediction model and combined the characteristics of e-commerce big data and used the Lasso regression algorithm to optimize the RBF model to achieve dynamic prediction of CC. At the same time, to facilitate the development of corresponding recovery strategies for e-commerce platforms, the

research would also combine the characteristics of customer lifecycle to conduct application research on the model, to screen out customer characteristics that were highly related to CC. The experimental findings indicated that the L-RBF model could achieve an accuracy of 95% with a loss rate of only 3%. In the stages of customer lifecycle excavation, advanced, stable, and decline, the predictive accuracy of this model could reach 97.6, 93.1, 92.7, and 91.8%, respectively, and the loss rate could reach 0.5, 1, 1.3, and 1.3%, respectively. It could also successfully screen the features with the highest correlation with CC in different periods. Therefore, the improved L-RBF model proposed in this study could achieve high prediction accuracy in the field of e-commerce and provide a basis for the platform to develop recovery measures. However, compared with the amount of big data in e-commerce, the number of experimental sets used in the study was small, and the amount of data can be expanded for in-depth analysis later.

7 Future work

Although the improved L-RBF model proposed in this research demonstrated high prediction accuracy in the field of e-commerce, there are still many challenges and opportunities in terms of practical applications and further optimization of the model performance. Future work will be centered on the following aspects:

(1) Real-world deployment and application challenges: Future work will focus on how to deploy and apply L-RBF models in real e-commerce platforms, including real-time updating of the models, optimization, and integration with platform-specific features. In addition, the user-friendliness and ease of maintenance of the model will be further improved so that e-commerce operation teams can use and maintain the model efficiently.
(2) Model customization for different platforms: In the future work, the unique characteristics of different e-commerce platforms (e.g., B2B, B2C, and C2C) and business models (e.g., online retail and social e-commerce) will be analyzed in depth, and the structure and algorithmic parameters of the L-RBF model will be adjusted accordingly. First, the customer behaviors, transaction modes, product categories, and market strategies of different e-commerce platforms are analyzed in depth. Second, considering that different platforms may employ different data collection and processing techniques, the model may need to integrate specific data preprocessing or feature extraction methods to accommodate these differences. Finally, due to the rapid changes in the e-commerce environment, models need to be able to update and adapt to new market trends and customer behavior patterns in real time.
(3) Improvements and new research directions: To improve the accuracy and efficiency of churn prediction models, ways to improve existing models through more efficient data processing methods, algorithm optimization and new feature selection techniques will be explored. Meanwhile, utilization of the latest machine learning and artificial intelligence techniques, such as the most popular convolutional neural network and its derivative networks, will be considered to further enhance the performance of the models.
(4) Authenticity and representativeness of the dataset: In the future, the L-RBF model needs to be further validated and optimized by obtaining more representative and authentic e-commerce platform data. Meanwhile, the data dimensions are extended to capture the multifarious factors of CC more comprehensively.

These future works aim to further improve the performance of the L-RBF model in CC prediction and ensure its wide applicability and practical value in different e-commerce environments.

Funding information: Not applicable.

Author contribution: The author confirms sole responsibility for the following: study conception and design, analysis and interpretation of results, and manuscript preparation.

Conflict of interest: The author reports that there are no competing interests to declare.

Data availability statement: Data will be made available on reasonable request.

References

[1] Xu X, Lockwood J. What's going on in the chat flow? A move analysis of e-commerce customer service webchat exchange. Engl Specif Purp. 2021;61(3):84–96.
[2] Guo L, Hu X, Lu J, Ma L. Effects of customer trust on engagement in live streaming commerce: mediating role of swift guanxi. Internet Res. 2021;31(5):1718–44.
[3] Liu Y, Liu F, Feng H, Zhang G, Wang L, Chi R, et al. Frequency tracking control of the WPT system based on fuzzy RBF neural network. Int J Intell Syst. 2022;37(7):3881–99.
[4] Li Q, Xiong Q, Ji S, Yu Y, Wu C, Yi H. A method for mixed data classification base on RBF-ELM network. Neurocomputing. 2021;431(28):7–22.
[5] Han Z, Qian X, Huang H, Huang T. Efficient design of multicolumn RBF networks. Neurocomputing. 2021;450(25):253–63.
[6] Yang Y, Lai X, Luo T, Yuan K. Study on the viscoelastic–viscoplastic model of layered siltstone using creep test and RBF neural network. Open Geosci. 2021;13(1):72–84.
[7] Liu Q, Li D, Ge SS, Ji R, Ouyang Z, Tee KP. Adaptive bias RBF neural network control for a robotic manipulator. Neurocomputing. 2021;447(4):213–23.
[8] Gopi AP, Jyothi RNS, Narayana VL, Sandeep KS. Classification of tweets data based on polarity using improved RBF kernel of SVM. Int J Inf Technol. 2023;15(2):965–80.
[9] Sarina S, Tanniewa AM. Implementasi algoritma support vector learning Terhadap Analisis Sentimen Penggunaan Aplikasi Tiktok shop seller center. Pros SISFOTEK. 2023;7(1):165–70.
[10] Gattermann-Itschert T, Thonemann UW. How training on multiple time slices improves performance in churn prediction. Eur J Oper Res. 2021;295(2):664–74.
[11] Bhattacharyya J, Dash MK. An investigation of customer churn Insights and intelligence from social media: A Netnographic research. Online Inf Rev. 2020;45(1):174–206.
[12] Al-Najjar D, Al-Rousan N, Al-Najjar H. Machine learning to develop credit card customer churn prediction. J Theor Appl Electron Commer Res. 2022;17(4):1529–42.
[13] Routh P, Roy A, Meyer J. Estimating customer churn under competing risks. J Oper Res Soc. 2021;72(5):1138–55.
[14] Zhong H, Zhang J, Zhang S. A combined prediction model of cross-border e-commerce export volume based on BP neural network and SVM. Int J Technol, Policy Manag. 2023;23(3):310–28.
[15] Huda I, Suhendra AA, Bijaksana MA. Design of prediction model using data mining for segmentation and classification customer churn in e-commerce Mall in Mall. JOIV: Int J Inform Vis. 2023;7(4):2280–9.
[16] Dai C. A method of forecasting trade export volume based on back-propagation neural network. Neural Comput Appl. 2023;35(12):8775–84.
[17] Bohnsack R, Liesner MM. What the hack? A growth hacking taxonomy and practical applications for firms – ScienceDirect. Bus Horiz. 2019;62(6):799–818.
[18] Matsuoka K. A framework for variance analysis of customer equity based on a Markov chain model. J Bus Res. 2021;129(5):57–69.
[19] Jin Y, Qin X. Significance of TP53 mutation in treatment and prognosis in head and neck squamous cell carcinoma. Biomarkers Med. 2021;15(1):15–28.
[20] Zhang X, Wang Z, Xiao B, Ye Y. A neural network learning-based global optimization approach for aero-engine transient control schedule. Neurocomputing. 2022;469(16):180–8.
[21] Hills E, Woodward TJ, Fields S, Brandsen B. Comprehensive mutational analysis of the Lasso Peptide Klebsidin. ACS Chem Biol. 2022;17(4):998–1010.
[22] Toraya H. Finding the best-fit background function for whole, owder, attern fitting using LASSO combined with tree search. J Appl Crystallography. 2021;54(2):427–38.
[23] Hidayat I, Ali MZ, Arshad A. Machine learning-based intrusion detection system: an experimental comparison. J Comput Cognit Eng. 2022;2(2):88–97.
[24] Chou P, Chuang HC, Chou YC, Liang TP. Predictive analytics for customer repurchase: Interdisciplinary integration of buy till you die modeling and machine learning. Eur J Oper Res. 2022;296(2):635–51.
[25] Zhai CY, Zhang MM, Xia XL, Liao YY, Chen H. Customer churn prediction model based on user behavior sequences. J Donghua Univ (Engl Ed). 2022;39(6):597–602.