Research on Improved Xgboost Algorithm for Big Data Analysis of e-Commerce Customer Churn

Li Li
Department of Economics & Management, Sichuan TOP IT Vocational Institute, Chengdu, China

***Abstract*—With the increasing cost of acquiring new users for e-commerce enterprises, it has become an important task for e-commerce enterprises to actively carry out customer churn management. Therefore, based on the distributed gradient enhancement library algorithm (XGBoost), this research proposes a big data analysis study on e-commerce customer churn. First, it conducts an evaluation analysis on e-commerce customer segmentation and combines the random forest algorithm (RF) to build an RF XGBoost prediction model based on customer churn. Finally, it verifies the performance of the prediction model. The results show that the area under receiver operating characteristic curve (AUC) value, prediction accuracy, recall rate, and F1 value of the RF-XGBoost model are significantly better than those of the RF, XGBoost, and ID3 decision trees to build an e-commerce customer churn prediction model; The average output error of RF-XGBoost model is 0.42, and the average output error is relatively good, indicating that the model proposed in this study has a smaller error and higher accuracy. It can make a general assessment of the customer churn of e-commerce enterprises, and then provide data support for the customer maintenance work of e-commerce enterprises. It is helpful to analyze the relevant factors affecting customer churn, to Equationte targeted customer service programs, thus improving the economic benefits of e-commerce enterprises.**

***Keywords*—E-commerce; customer churn; random Forest; XGBoost; big data**

I. INTRODUCTION

At present, domestic e-commerce enterprises and websites are also gradually increasing, and people are getting used to buying goods and services online [1]. Most traditional industries have built their e-commerce platforms, and the opportunities for customers to choose products are gradually increasing [2]. For e-commerce companies, they must face fierce competition for customer resources from other companies, and companies that fail to compete will face elimination from the market. Therefore, finding a suiTABLE method to more accurately predict the purchase behavior and churn of users, and taking effective measures to reduce the loss rate has become a hot topic in related industries. In the absence of new customer resources, some e-commerce companies began to focus on old customers, and used data mining-related technologies to analyze the shopping behavior of old customers to mine useful data. By formulating marketing plans based on these data, enterprise can avoid the loss of old customers and increase more economic benefits [3]. Some scholars have pointed out that enterprises will accumulate massive amounts of data in the daily operation process, from which valuable information can be mined and customers who are about to be lost can be accurately predicted [4]. For e-commerce companies, when they find customers who are about to be lost, taking corresponding retention measures can effectively reduce the loss rate of corporate customers and make customers active on the company's e-commerce platform [5]. Therefore, this research takes e-commerce enterprise customers as the research object, analyzes the relevant behavioral data recorded by the background when customers purchase online, subdivides them according to customer value, combines the distributed gradient enhancement library algorithm (XGBoost) with the random forest algorithm (RF), and constructs a big data analysis model for e-commerce customer loss.

II. RELATED WORK

Janabi S and his team proposed an integrated system, in recent years, under the global economic downturn under the new crown epidemic; all walks of life are competing for customer resources. Therefore, the analysis of big data for enterprise customer churn has attracted people's attention. Janabi S and his team proposed an integrated system to help telecom companies achieve intelligent prediction of customer churn, use training data to build an intelligent predictor of customer churn, and use a genetic algorithm (GA) algorithm to group customers for decision-making. The integrated system is superior to the system built by traditional methods and has high accuracy [6]. Li W et al. used feature extraction technology to analyze lost customers and proposed a customer loss prediction method. The online test shows that the proposed customer churn prediction method can effectively predict customer churn, which is superior to the traditional prediction technology [7]. Gu Y et al. studied the effect of the decision tree algorithm in customer churn prediction. Each performance index of this algorithm is superior to other traditional algorithms. After applying this method, it can help enterprises retain old customers and avoid the loss of old customers [8]. Jiao G and his team predicted telecommunication customer churn and compared four prediction models on a telecommunication dataset, and found that the model built by XGBoost and RF algorithm has the best prediction ability among many algorithms [9]. Eria K's studio studied the relevant characteristics of customer data, and on this basis proposed a customer churn prediction method, that reduces the data according to missing values and irrelevant variables, which helps reduce churn analysis. The computational cost is incurred by teachers using big data in churn prediction and analysis [10].

Sharma T and other researchers propose an improved XGBoost algorithm as a model to address customer churn in the telecom industry. Compared with non churning customers,

the previous model paid more attention to prediction accuracy, and the proposed model correctly classified all churning customers, having better performance [11]. Swetha P et al. designed a data characteristic function model to judge and predict the loss rate of enterprise customers. In the process of research, the loss function is constructed and combined with the XG_ Boost method. The proposed model is helpful to identify correct and wrong classification examples of South Asia's global mobile communication system service providers [12]. Scholars such as Naser AM built a customer churn prediction model to determine whether each customer is a churned customer and get more opinions about serving customers and use the extreme gradient to boost the classification model of "XGBoost", which is added to the original network by computing A centrality measure for the new attribute of the dataset to evaluate model performance. Experiments show that the proposed clustering-based churn detection method combining social influence and web content significantly improves the prediction accuracy of telecom datasets [13].

Although a large number of studies have focused on enterprise customer churn, there are still few studies on e-commerce enterprise customer churn. And the prediction accuracy needs to be improved, and the application of the XGBoost loss prediction model is not flexible enough. Therefore, based on the improved XGBoost algorithm, this study proposes a prediction method for e-commerce customer churn, to contribute to the development of e-commerce enterprises.

III. E-Commerce Customer Segmentation Evaluation and Customer Churn Prediction Model Construction

A. E-commerce Customer Value Segmentation Evaluation

The "28" law proposed by the Italian economist Pareto shows that in any group of things, the most important part is only a small part, accounting for about 20%, and the remaining 80%, although the majority, is secondary. Its core content is that 80% of the results in life almost all come from 20% of activities. According to the famous "28" law, nearly 80% of the benefits of an enterprise are brought by 20% of the main customers [14]. Therefore, in the current situation that resources are getting smaller and smaller, e-commerce companies need to further subdivide customers, and at the same time customize different service plans for different customers, to avoid the loss of customers. Based on the traditional customer behavior analysis model (RFM), this study analyzes the customer value evaluation indicators of e-commerce companies [15]. The period indicator (T) of the first online activity and the last online activity in the observation period is introduced through the RFM model, and the RFM model (RFMT) based on the time factor is constructed. It mainly includes four key indicators in the observation period, namely the time interval (R) between the last time the customer had consumption behavior and the observation point of analysis, the total consumption times (F), the total consumption amount (M) and T .

The customer segmentation process is shown in Fig. 1. The first step is to build an index system, mainly using expert scoring to evaluate four key indicators; the second step is to evaluate customer value (V), by calculating the normalized index weight value used to obtain the corresponding V value; the third step is to classify the customer groups. The V value is analyzed through K-means, customer types are divided, and the indicators of each type of center point are analyzed to get the category labels of various customers, and set the value threshold, so that the enterprise can further identify the customers with special value.

According to the actual situation of the research, and according to the method in the literature, a single-layer indicator system for e-commerce customer segmentation is proposed, the target layer is "customer value evaluation", and the indicator layer is four key indicators [16]. Then sort the four key indicators proposed, calculate the weight of each indicator, and realize the calculation of customer value V .

According to the single-level indicator system of e-commerce customer segmentation, the four key indicators are compared with each other in pairs and graded. From the comparison results, the feature discriminant matrix can be constructed $X = (x_{ij})_{n \times n}$, as shown in Eq. (1).

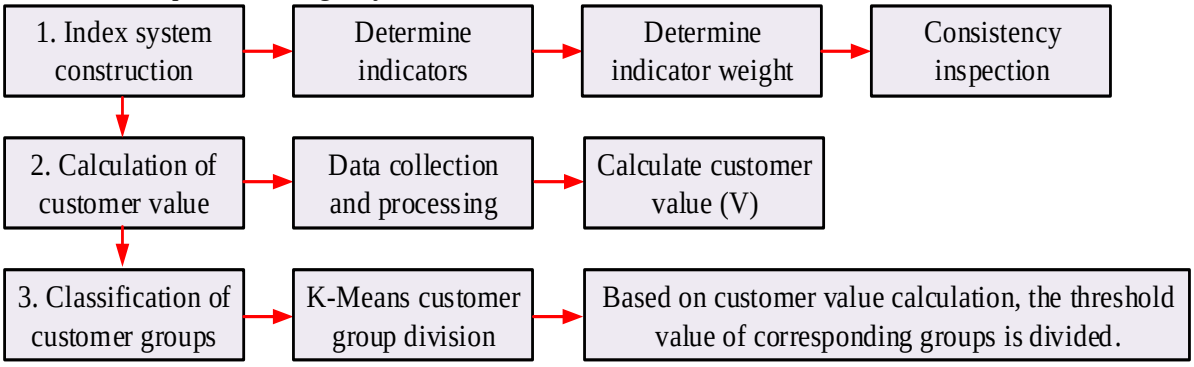


Fig. 1. Customer segmentation model framework

$$X=\begin{vmatrix} x_{ij} & X_1 & X_2 & ... & X_n \\ X_1 & x_{11} & x_{12} & ... & x_{1n} \\ X_2 & x_{21} & x_{22} & ... & x_{2n} \\ ... & ... & ... & ... & ... \\ X_n & x_{n1} & x_{n2} & ... & x_{nn} \end{vmatrix} \quad (1)$$

It can be seen from Eq. (1) that x_{ij} the contrast between the $x_{ij}=1/x_{ji}$ representation feature i and the feature exists at the same time j. According to the characteristics of the matrix in Eq. (1), to obtain the weight of each index, it is necessary to use the expert scoring method to score X the eigenvectors. Then, the consistency test is performed by the largest eigenroot $\lambda_{\max}$ pair X, as shown in Eq. (2).

$$CI=(\lambda_{\max}-n)/(n-1) \quad (2)$$

To be able to better evaluate X the consistency, this study introduces an index X related to RI. Use the CR value to judge X whether it passes the test, if $CR>0.1$ it fails the test, otherwise it passes the test, see Eq. (3).

$$CR=CI/RI\ (2) \quad (3)$$

In the classification of customer groups, the scores of each customer in the RFMT model are compared with the average of the indicators, and those higher and lower than the average are marked as "+" and "-" respectively. Based on this, customers are divided into 16 sub-categories, and then through expert analysis, the e-commerce customers are finally divided into four groups, namely core customers, important customers, important development customers, and general customers. After the customer is preprocessed, the customer's score on the four indicators will be obtained, and the weight value of each indicator will be obtained by using the expert scoring method. Finally, the customer's V value will be calculated by the Eq. (4). Among them, W represents the feature vector obtained by expert scoring.

$$V=W_R\times\frac{1}{R}+W_F\times F+W_M\times M+W_T\times T \quad (4)$$

B. Improved XGBoost Prediction Model

First of all, we need to measure the data characteristics of customers. This research is implemented through random forests (RF). When the RF algorithm randomly selects the training subset applied to the decision tree algorithm, the unselected data are out-of-bag data [17]. The error rate of the out-of-bag data evaluation model is named E1; then random interference terms are added to the out-of-bag data, x and the evaluation model error is $E2$. If $E2>E1$, then the importance of feature x (I) is higher, and I, the higher the value, the more important the feature is. Assuming that the number of training decision trees generated by the RF algorithm is k, then I the calculation Eq. is shown in Eq. (5).

$$I=\frac{\sum(E1-E2)}{k} \quad (5)$$

The introduction of the feature importance measure is convenient for the RF algorithm to select a better feature set [18]. In the process of screening important features through the RF algorithm, a tree-building process will be carried out at the same time. This process mainly includes four processes: one is I to calculate the value of the I features in the data set by Eq. (5). Sort by size, and output the I values of all features; the third is to select the previous x_i feature in the output to regroup into multiple feature sets, input each combined feature set into the RF model, and then calculate the out-of-bag error rate of each feature set, among them $i=1,2,3...$, $x_1<x_2<x_3....$; Fourthly, all out-of-bag error rates are obtained by comparison, and the feature set whose out-of-bag error rate does not change continuously is selected as the optimal feature set, so as to facilitate subsequent empirical analysis and research.

The XGBoost algorithm belongs to a classic ensemble algorithm, which implements part of the generalized linear machine learning algorithm based on the boosted decision tree (GBDT) [19]. XGBoost model can use CPU multithreading to realize parallel operation and obtain the best parameters. See Eq. (6) for the calculation of its objective function.

$$Obj(\theta)=L(\theta)+\Omega(\theta) \quad (6)$$

In Eq. (6), θ is the model parameter and $L(\theta)$ is the loss function. The higher the value, the higher the model accuracy. However, the model is easy to regard noise as a learning sample, which leads to overfitting. Overfitting refers to the phenomenon that the model predicts the known data well, but the prediction effect on the unknown data is poor, which will reduce the robustness of the model.

The regularization term ($\Omega(\theta)$) can improve the generalization ability of the model and the complexity of the evaluation model. Let the number of ensemble trees in the XGBoost model be k.

$$y_i=\sum_{k=1}^{K}f_k(x_i),\ f_k\in\rho \quad (7)$$

In Eq. (7), f_k is a base classifier, K and ρ respectively represents the number and space of the base classifier. If y_i represents the class i mark, the calculation of the objective function is shown in Equation (8).

$$Obj=\sum_{i=1}^{n}L(y_i,\hat{y}_i)+\sum_{k=1}^{K}\Omega(f_k) \quad (8)$$

The XGBoost model is a boosted tree model with a special form, and its essence is an additive model. To solve the objective function of the model, it is necessary to obtain each tree f_k. However, all trees f_k cannot be obtained at one time, so it is necessary to train according to the model that has been iterated several $t-1$ times to obtain the first t tree. After iteration t times, the loss function of the model is expanded by second-order Taylor expansion, and then the constant term is removed, and a regular term is added to constrain the loss function, to control the complexity of the model. Based on the above content, t the objective function of the first iteration

can be obtained, as shown in the Eq. (9).

$$Obj^{(t)} \approx \sum_{i=1}^{n}\left[g_i f_t(x_i)+\frac{1}{2}h_i f_t^2(x_i)\right]+\Omega(f_t) \quad (9)$$

For the objective function, g_i and h_i are their first and second derivatives respectively. After minimizing the objective function, the model can achieve the best classification effect, thus improving the prediction accuracy [20]. The loss function when not optimized is drawn in Fig. 2. From the analysis of the curve gradients in the graph, it is found that the gradients of the first two types of curves are consistent, indicating that the losses of the two types of errors are regarded as the same, and the positive examples are misclassified more than the negative ones after optimization, which is consistent with the actual situation.

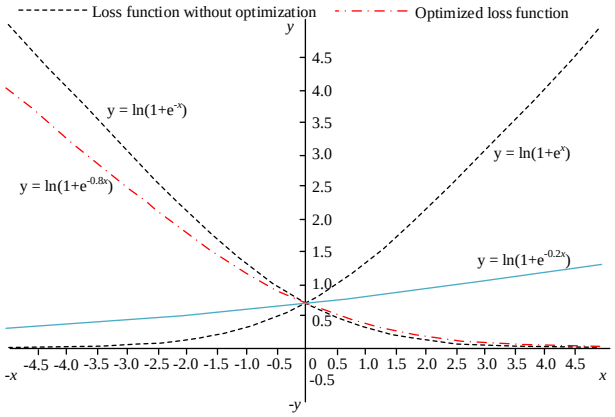


Fig. 2. Comparison diagram of loss function before and after optimization

Due to the serious loss of e-commerce customers, this study judges the loss of customers by studying the basic data and behavioral data of customers and uses the RF feature selection method combined with the XGBoost algorithm to build a prediction model, to better handle the relationship between e-commerce companies and customers. This research combines RF and XGBoost to build a prediction model, and its process is shown in Fig. 3. In classifier 1, if the customer is not lost, the output is 0; if the customer is lost, the output is 1. In classifier 2, the customers who have not lost the e-commerce company are predicted. If there is a possibility of losing customers, the output result is 1; otherwise, the output is 0.

Firstly, the model should be trained through the data test set, and each parameter in XGBoost should be tuned to obtain the optimal parameters of the algorithm, to make the model prediction effect in the later stage the best. In addition, each tree structure in the algorithm needs to be made simpler. For the detailed process of training the objective function, see Eq. (10).

$$\begin{cases} \hat{y}_l^{(0)} = 0 \\ \hat{y}_l^{(1)} = f_1(x_i) = \hat{y}_l^{(0)} + f_1(x_i) \\ \hat{y}_l^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_l^{(1)} + f_2(x_i) \\ \cdots\cdots \\ \hat{y}_l^{(t)} = \sum_{k=1}^{t} f_1(x_i) = \hat{y}_l^{(t-1)} + f_t(x_i) \end{cases} \quad (10)$$

The process of training the model and obtaining the optimal parameters proposed in this study mainly includes four steps: first, obtaining a new decision tree, second, deriving the loss function, third, obtaining the predictive value of the sample, and fourth, returning to the first step until the result of parameter optimization is optimal.

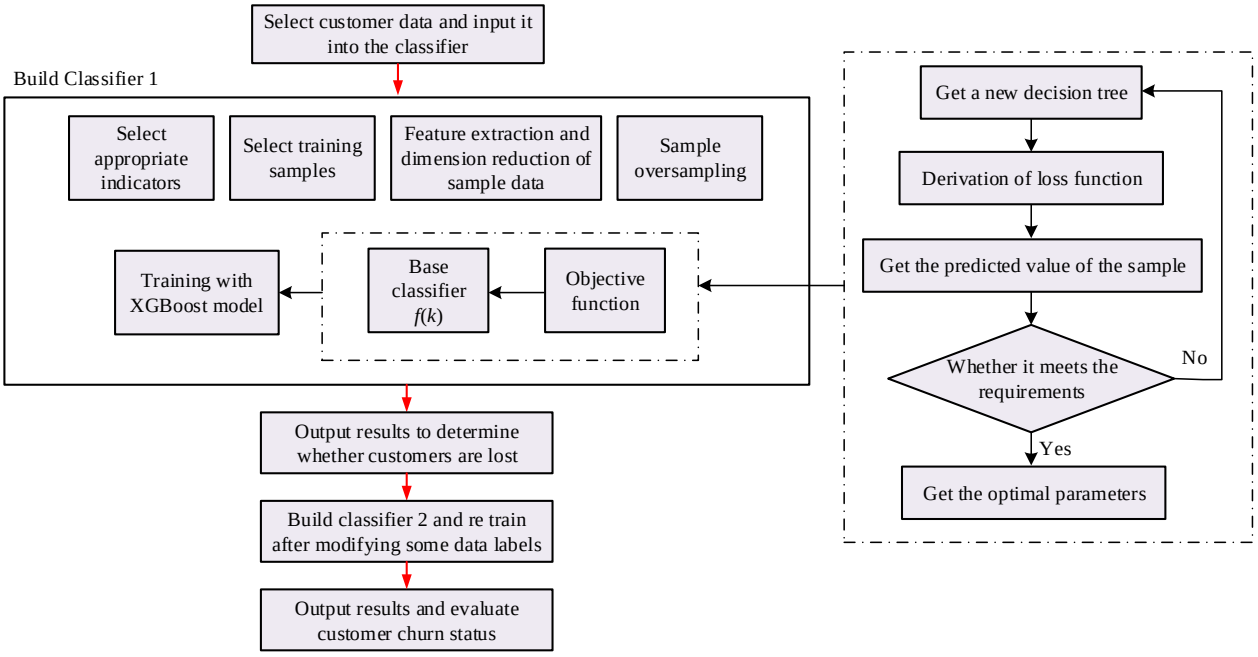


Fig. 3. Customer churn prediction model proposed in this study

IV. Model Performance Analysis

This study selects the data of an e-commerce platform from April 1, 2020, to April 1, 2022, and analyzes the performance of the model constructed in this study. Take the data from April 1, 2020, to February 19, 2022, as the training set, and the data from the rest of the dates as the test set. The data mainly includes the original sample skeleton data, advertising information attribute description, user information data, and user behavior log data. After the cleaning and conversion of the selected data with missing values and abnormal values, we selected 121949 effective users, a total of 4488519 user behavior data for customer churn prediction research. Based on the above customer segmentation results, Table I shows the churn rate of each type of customer.

The results in Table I show that corresponding to the characteristics of each type of customer after customer segmentation, the order of customer churn rate from high to low is: general customers, important development customers, important customers, and core customers. Moreover, the churn rate of general customers exceeds 90%, which is much higher than that of core customers, indicating that general customers may just consume on the platform by accident, and can be identified as temporary customers, which is very easy to lose. The core customer churn rate is the lowest, indicating that such customers have the highest loyalty to the company and stronger customer stickiness. The difference between the churn rate of important customers and core customers is 6.46%, and the difference between the churn rate of important development customers and important customers is 4.06%, indicating that the difference in churn rate between them is relatively stable, but the churn rate of general customers is more important The churn rate of development customers has increased by 43.47%, which makes managers see the fundamental difference between general customers and the other three types of customers, and reminds managers to give up properly and invest more energy in this part of customers.

Build two prediction models (Model A and Model B) based on RF-XGBoost. Among them, the training of model A is mainly through the pre-processed data; Model B uses the original data set for training. See Fig. 4 for the training iteration process. For the accuracy rate and loss curve, the curve of model A is relatively stable, while model B has large fluctuations. It shows that data preprocessing can improve the stability of the model and prediction accuracy.

TABLE I. Customer Churn Rate of Customer Segments

/	Customer volume	The number of customers lost	Loss rate
Core customers	23975	8983	37.47%
Key customers	24316	10683	43.93%
Important development customers	26812	12865	47.99%
General customers	46705	42723	91.46%

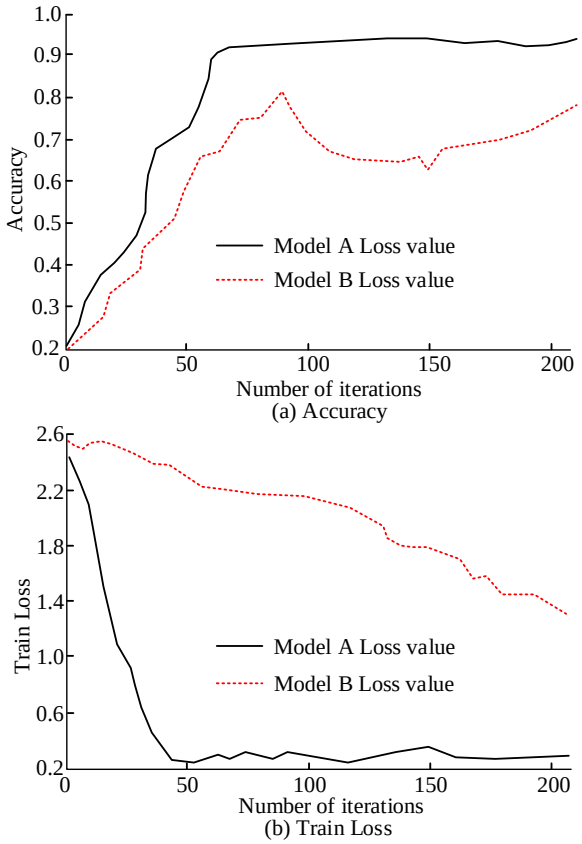


Fig. 4. Data preprocessing effect

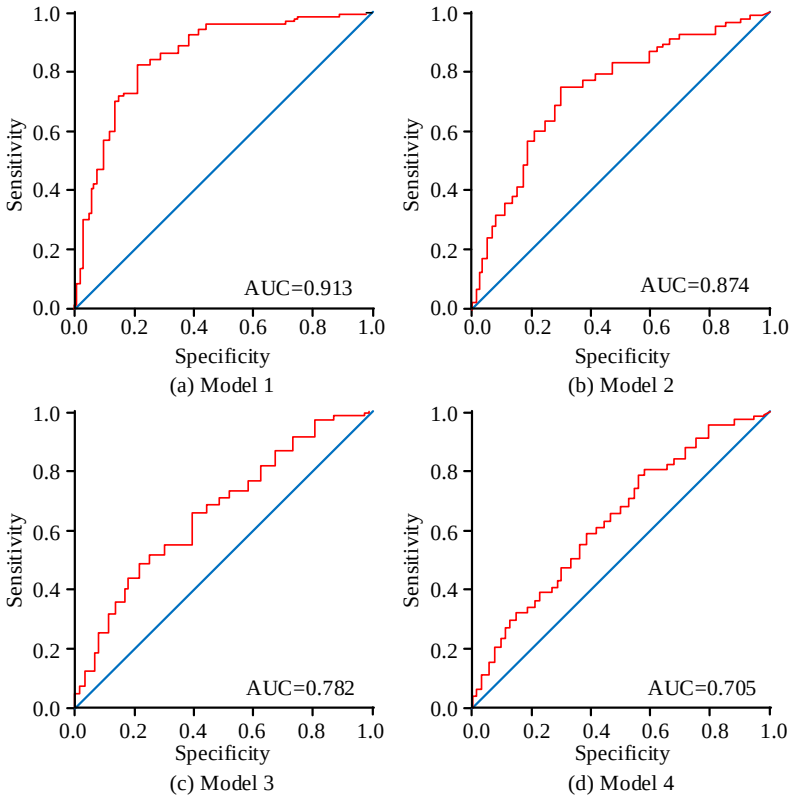


Fig. 5. ROC curve of four models

The prediction models are built based on RF-XGBoost, RF, XGBoost, and ID3 decision trees, which are represented as models 1 to 4 respectively. The performance of the built multiple models is evaluated by the receiver operating characteristic curve (ROC), as shown in Fig. 5. The results showed that model 1 had the largest AUC value, significantly exceeding the other three models. It shows that the performance of the model proposed in this study is better than several traditional prediction models, so it has high practicality and can provide accurate and scientific data support for enterprise customer churn management.

The training results of each model are compared as shown in Fig. 6. The number of iterations required for Model 1 to reach the best accuracy is 45, which is significantly less than the other three models. After iteration to the best accuracy, the error is 0.08, which is the best among all models. Therefore, the proposed model has better convergence and training effects, which is better than the other three models.

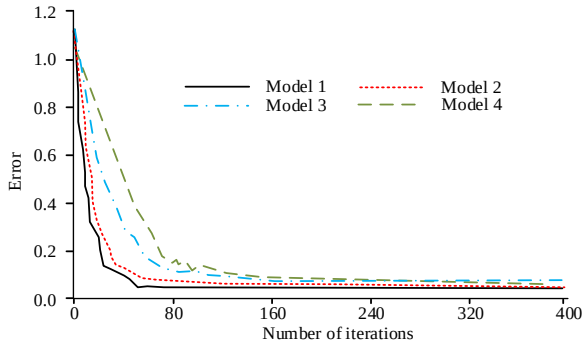


Fig. 6. Training results of each model

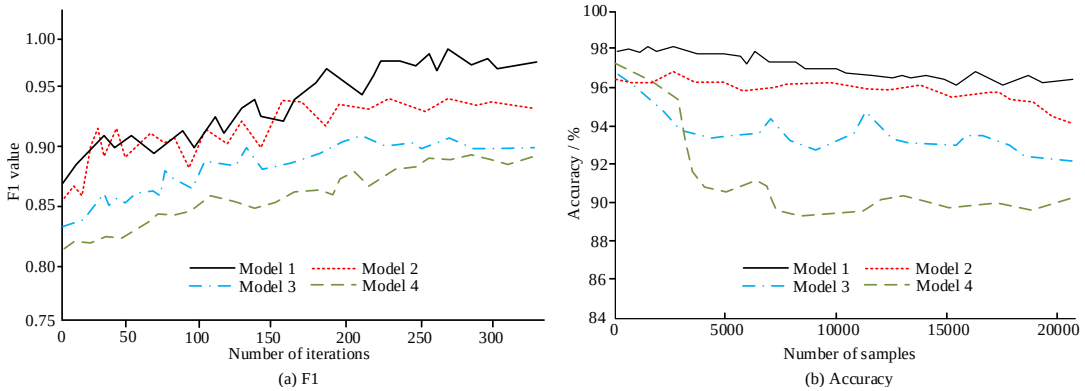
Test sets are used to test the performance of each model, as shown in Table II. For the accuracy rate, Model 1 has the highest prediction accuracy, reaching 92.15%. For the recall rate, Model 1 has the highest recall rate, reaching 0.93. For the F1 value, model 1 is the highest, reaching 0.91. In addition, the prediction time of model 1 is the shortest, 12.7s.

Further, we analyze the changes of F1 indicators and prediction accuracy of the four models in the iterative process, and the results are shown in Fig. 7. In Fig. 7 (a), after 300 iterations of model 1, the value of F1 reaches the maximum, and the value of F1 in the whole iteration process also remains optimal. In all models, this study proposes that the value of F1 in the model is optimal. In Fig. 7 (b), the accuracy of all models will decrease with the increase of test data. On the whole, no matter whether the number of samples increases or decreases, the prediction accuracy of model 1 is the best. It is said that the model proposed in this study has excellent prediction accuracy, which can accurately predict customer churn based on user data, thus guiding e-commerce enterprises to Equationte customer maintenance plans.

Divide the selected test set data into 25 pieces, test the four models respectively, and compare the average error between various models and the real output results. After 25 tests on four models, see Fig. 8. The average output error of model 1 is 0.42, which is the lowest among all comparison models, suggesting that the model proposed in this study has a smaller error and higher accuracy.

TABLE II. PERFORMANCE TEST RESULTS OF FOUR MODELS

Model	Index			
	Accuracy / %	Recall / %	F1	Detection time / s
1	92.15	0.93	0.91	12.7
2	85.34	0.89	0.84	33.5
3	86.19	0.87	0.82	48.3
4	83.48	0.80	0.76	36.9



(a) F1 (b) Accuracy

Fig. 7. F1 value and accuracy change curve of four models

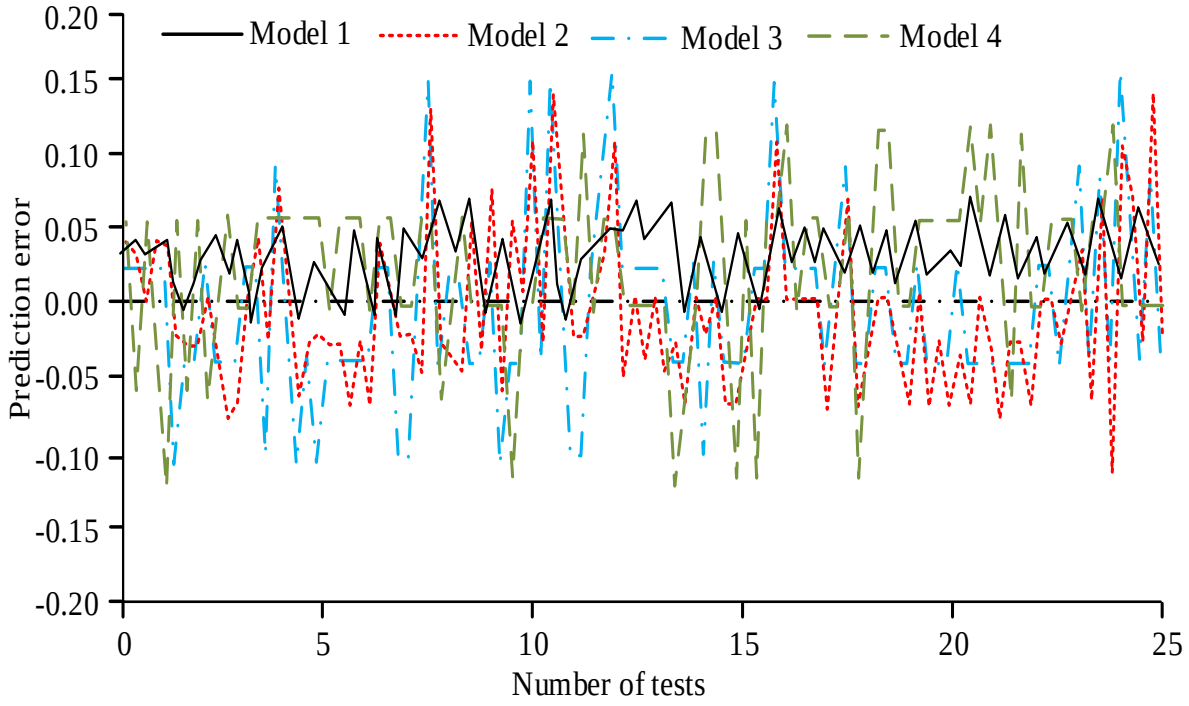


Fig. 8. Comparison of output errors of various models

The prediction model proposed in this study has excellent performance, including prediction accuracy, stability, and convergence. It can make a rough assessment of customer churn, thus providing data support for e-commerce enterprises' customer maintenance work, helping to analyze the relevant factors affecting customer churn, to Equationte targeted customer service programs, thus improving the enterprise's revenue.

In view of the importance of the model proposed in this study to other models in the application process, the ROC curve is selected for comparative study of model performance. It can be seen from the Fig. 9 that the AUC area of the optimized model proposed in this experiment is larger than that of the non-optimized model, which can significantly improve the accuracy of the method.

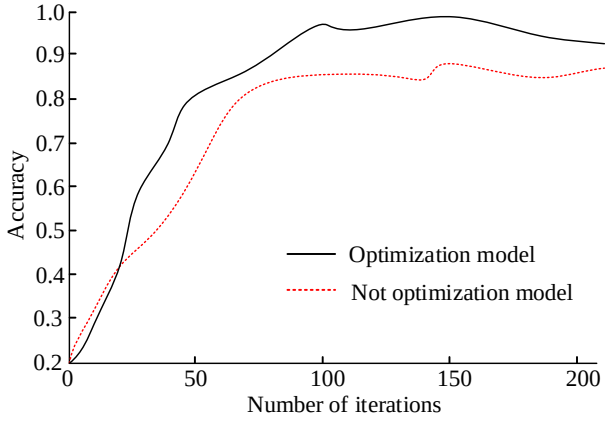


Fig. 9. Performance comparison between different models

V. Conclusion

For e-commerce enterprises, they must face fierce competition for customer resources from other enterprises. While enterprises that fail in the competition will be eliminated from the market. This research takes e-commerce enterprise customers as the research object, analyzes the relevant behavioral data recorded by the background when customers purchase online, subdivides them according to customer value, combines the distributed gradient enhancement library algorithm (XGBoost) with the random forest algorithm (RF), and constructs a big data analysis model for e-commerce customer churn. The results show that the AUC value of model 1 is the largest, which shows that the performance of the model proposed in this study is significantly better than several traditional prediction models, so it has high practicality and can provide accurate and scientific data support for enterprise customer churn management. The number of iterations required for Model 1 to reach the best accuracy is 45, which is significantly less than the other three models. After iteration to the best accuracy, the error is 0.08, which is the best among all models. Therefore, the proposed model has better convergence and better training effect. For the accuracy rate, Model 1 has the highest prediction accuracy, reaching 92.15%; for the recall rate, Model 1 has the highest recall rate, reaching 0.93; For F1 value, model 1 is the highest, reaching 0.91; In addition, the prediction time of model 1 is the shortest, 12.7s. After 300 iterations of model 1, the value of F1 reaches the maximum, and the value of F1 in the whole iteration process also remains optimal. In all models, this study proposes that the value of F1 in the model is optimal. The prediction accuracy of model 1 is the best, indicating that the model proposed in this study has excellent prediction accuracy. Since the data obtained in this study are preliminary data of e-commerce customers, deeper

customer behavior data such as evaluation, sharing, return and exchange cannot be obtained, which leads to the too one-sided prediction of customer behavior. We will try to get more customer behavior data.

REFERENCES

[1] B. Senthilnayaki, M. Swetha, D. Nivedha, "Customer Churn Prediction." IARJSET, Vol. 8. No. 6, pp. 527-531, 2021.

[2] A. Amin, S. Anwar, A. Adnan, M. Nawaz, K. Alawfi, A. Hussain, K. Huang, "Customer churn prediction in the telecommunication sector using a rough set approach." Neurocomputing, Vol. 237, NO. (MAY10), pp. 242-254, 2017.

[3] R. Yu, X. An, B. Jin, J. Shi, O.A. Move, Y. Liu, "Particle classification optimization-based BP network for telecommunication customer churn prediction." Neural Computing and Applications, Vol. 29, No. 3, pp. 707-720, 2018.

[4] I. Alshourbaji, N. Helian, Y. Sun, M. Alhameed, "Anovel HEOMGA Approach for Class Imbalance Problem in the Application of Customer Churn Prediction." SN Computer Science, Vol. 2, No. 6, pp.1-12, 2021.

[5] A. Sk, B. Hl. "Customer Churn Prediction in Influencer Commerce: An Application of Decision Trees." Procedia Computer Science, Vol. 199, pp. 1332-1339, 2022.

[6] S. Janabi, F. Razaq, "Intelligent Big Data Analysis to Design Smart Predictor for Customer Churn in Telecommunication Industry," Springer, Cham. Springer, Cham, Vol. 53, pp. 4659-4676, 2018.

[7] W. Li, C. Zhou, "Customer churn prediction in telecom using big data analytics." IOP Conference Series: Materials Science and Engineering, Vol. 768, No. 5, pp. 052070, 2020.

[8] Y. Gu, T. D. Palaoag, J. Cruz, "Comparison of Main Algorithms in Big Data Analysis of Telecom Customer Retention." IOP Conference Series: Materials Science and Engineering, Vol. 1077, No. 1, pp. 012045, 2021.

[9] G. Jiao, H. Xu, "Analysis and Comparison of Forecasting Algorithms for Telecom Customer Churn." Journal of Physics: Conference Series, Vol. 1881, No. 3, pp. 032061 (6pp), 2021.

[10] K. Eria, B. P. Marikannan, "Significance-Based Feature Extraction for Customer Churn Prediction Data in the Telecom Sector." Journal of Computational and Theoretical Nanoscience, Vol. 16, No. 8, pp. 3428-3431, 2019.

[11] T. Sharma, P. Gupta, V. Nigam, M. Goel, "Customer Churn Prediction in Telecommunications Using Gradient Boosted Trees." Vol. 1059, pp. 235–246, 2020.

[12] P. Swetha, B. Dayananda, "Improvised_XgBoost Machine learning Algorithm for Customer Churn Prediction." EAI Endorsed Transactions on Energy Web, Vol. 7, No. 30, pp. 164854, 2018.

[13] A. M. Naser, Al-Shamery E S. Churners Prediction Based on Mining the Content of Social Network Taxonomy. International Journal of Recent Technology and Engineering, 8(Issue-2S10):341-351, 2020.

[14] A. Vignesh, T. Y. Selvan, G. G. Krishnan, A.N. Sasikumar, A.V.D. Kumar. "Efficient Student Profession Prediction Using XGBoost Algorithm // International Conference on Emerging Current Trends in Computing and Expert Technology." Springer, Cham, Vol. 35, pp. 140-148, 2020.

[15] P. Lalwani, M. K. Mishra, J. S. Chadha, P. Sethi, "Customer churn prediction system: a machine learning approach." Computing, Vol. 104, No. 2, pp. 271-294, 2021.

[16] D. Liu, X. Zhang, Y. Shi, H. Li, "Prediction of Railway Freight Customer Churn Based on Deep Forest." Vol. 12837, pp. 479-489,2021.

[17] M. Panjasuchat, Y. Limpiyakorn, "Applying Reinforcement Learning for Customer Churn Prediction." Journal of Physics: Conference Series Vol. 1619, NO. 1, pp. 012016 (5pp), 2020.

[18] H. Wei, Q. T. Zeng, "Research on sales Forecast based on XGBoost-LSTM algorithm Model." Journal of Physics: Conference Series, Vol. 1754, No. 1, pp. 012191 (6pp), 2021.

[19] M. Li, X. Fu, D. Li, "Diabetes Prediction Based on XGBoost Algorithm." IOP Conference Series: Materials Science and Engineering, Vol. 768, No. 7, pp. 072093 (7pp), 2020.

[20] Z. Liu, Q. Kong, L. Yang, "Power consumption prediction with K-nearest-neighbours and XGBoost algorithm." International Journal of Wireless and Mobile Computing, Vol. 15, No. 4, pp. 374-381, 2018.