

Customer Churn Prediction Using AdaBoost Classifier and BP Neural Network Techniques in the E-Commerce Industry

Xiancheng Xiahou, Yoshio Harada

Faculty of Business Management, Osaka Sangyo University, Osaka, Japan

Email: Kakou.osu@hotmail.com

How to cite this paper: Xiahou, X. C., & Harada, Y. (2022). Customer Churn Prediction Using AdaBoost Classifier and BP Neural Network Techniques in the E-Commerce Industry. *American Journal of Industrial and Business Management*, 12, 277-293. <https://doi.org/10.4236/ajibm.2022.123015>

Received: February 7, 2022

Accepted: March 14, 2022

Published: March 17, 2022

Copyright © 2022 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

In customer relationship management, it is important for e-commerce businesses to attract new customers and retain existing ones. Research on customer churn prediction using AI technology is now a major part of e-commerce management. This paper proposes a churn prediction model based on the combination of k-means clustering and AdaBoost classifier algorithm, allowing the segmentation of customers into three categories. Important customer groups can also be determined based on customer behavior and temporal data. Customer churn prediction was carried out using AdaBoost classification and BP neural network techniques. The results show that the research method of clustering before prediction can improve prediction accuracy. In addition, a comparative analysis of the results suggests that the AdaBoost model has better prediction accuracy than the BP neural network model. The research results of this paper can help B2C e-commerce companies develop customer retention measures and marketing strategies.

Keywords

Customer Churn, E-Commerce, K-Means Clustering, AdaBoost, BP Neural Network

1. Introduction

Customer relationship management (CRM) is crucial in marketing, as it is the core of enterprise information management (Kotler & Keller, 2016). In the past decade, enterprises have focused on CRM, especially the challenges of customer churn (Daoud et al., 2018). Customers are key to the success of enterprises. Companies can improve their market competitiveness and economic benefits (Bi, 2019)

with a large customer base. However, it is easy for customers to switch e-commerce service providers amid fierce competition. Therefore, customer churn has become an inevitable issue for many companies (Hadden et al., 2007). Businesses usually retain their customers through advertising or product optimization, but these retention measures that are not targeted at specific customers can lead to rising costs and potential waste of resources (Jahromi et al., 2014). Chung et al. (2016) found that the cost of acquiring new customers is five to six times higher than retaining existing ones. Thus, it is essential for enterprises to conduct customer churn prediction research and analyze the causes of customer churn to win back lost customers.

There have been numerous studies on customer churn in various industries, including telecommunications (Bock & De, 2021; Kozak et al., 2021; Alboukaey et al., 2020; Verbeke et al., 2012; Coussement et al., 2017), finance (Devriendt et al., 2021; Dumitrescu et al., 2022; Velez et al., 2020) and e-commerce (Gattermann-Itschert & Thonemann, 2021; De et al., 2021; Gordini & Veglio, 2017; O'Brien et al., 2020). Zhao et al. (2021) analyzed the causes and influencing factors of customer churn in the telecommunications industry through data mining algorithms, and found that logistic regression algorithms can accurately identify the causes of customer churn while customer lifetime value modeling was the best method to manage customer churn. Meanwhile, Devriendt et al. (2021) adopted a new logistic regression tree algorithm to study customer churn issues in the financial industry. The results showed that this method has an accurate prediction and can help the financial and credit industry improve risk efficiency. Moreover, Schaeffer & Sanchez (2020) predicted customer churn in the B2B industry using the support vector machine (SVM). The results indicated that SVM can help enterprises identify churned and non-churned customers in a timely manner, while saving time and money. In these studies, the Recency, Frequency, Monetary (RFM) model and various machine learning algorithms were adopted for customer churn prediction, such as logistic regression (LR), decision tree (DT), random forest (RF), support vector machine (SVM) and k-nearest neighbors (KNN). These studies can reduce customer churn and help companies devise effective marketing strategies.

However, churn prediction research in the aforementioned literature mainly focused on the telecommunications and financial industries. There were few studies addressing customer churn in e-commerce but only discussed customer churn in B2B. Technology has transformed online shopping, with e-commerce sites and portable mobile devices gaining traction. Online shopping is diverse and convenient among B2C customers. Customers' shopping behavior datasets usually include time of shopping, purchase readiness, purchase intention and customer satisfaction (Kotler & Keller, 2008). Chen et al. (2005) stated that customer demographics can be obtained directly from enterprises' data warehouses, while longitudinal behavior data of B2C customers is temporary and vary over the time of shopping. For example, e-commerce websites usually provide services and features such as product collection, shopping cart, evaluation manage-

ment, shopping reward points, time of delivery, time of receipt, payment methods, invoice management and product specifications. Such data are usually stored separately in transactional databases and characterized by longitudinal temporality and multi-dimensionality. These information variables may result in better customer churn predictions, but existing literature on e-commerce customer churn prediction often ignores longitudinal behavior and longitudinal temporality data (Eichinger et al., 2006; Orsenigo & Vercellis, 2010). In addition, the consumption behaviors of customers in the telecommunications or financial industry are different from online B2C customers, with the former grouped as contractual customers and the latter representing non-contractual customers. In terms of business management, the consumption behaviors of contractual customers are relatively straightforward as it is easier to ascertain customer churn based on the data variables of consumption behaviors. On the other hand, B2C enterprises find it difficult to identify and predict customer churn. Existing literature on B2C customer churn prediction model is incomplete, and such studies are also lagging behind. Hence, it is of great significance to conduct research on customer churn prediction in B2C contexts.

With the wide application of big data and machine learning, businesses can easily acquire information from consumption behavior data for analysis and prediction modeling. Based on the real data of an e-commerce website and analysis of shopping behavior information variables, this study will segment customers using k-means clustering before filtering features through a random forest, and a B2C customer churn prediction model is established through the AdaBoost classification algorithm. To verify the advantages of AdaBoost algorithm modeling, the results of BP neural network modeling are analyzed and discussed.

The second part of this paper will discuss existing literature and the third part will address the research methods and introduce the basic principles of AdaBoost and BP neural network algorithms. Furthermore, the fourth part involves empirical research, including data preparation, data preprocessing, customer segmentation, variable screening and prediction evaluation indicators. The final part will conclude the findings and address future research directions.

2. Literature Review

Customer churn refers to a situation where customers stop using specific products or services in favor of another competitor's products or services (Amin et al., 2017). Studies on customer churn involved three aspects—churn prediction and identification, churn cause analysis and customer retention strategies. Churn prediction has become increasingly complex due to the different consumption characteristics of customers in various industries, making it difficult for enterprises to determine whether they are losing customers. Previous research showed that it is hard to obtain accurate predictions through the RFM model and temporal threshold method. Schmittlein et al. (1987) first adopted the Pareto/NBD models to predict customer behavior, before Fader et al. (2005) proposed the BG/NBD models. With the wide application and promotion of big data and data

mining, progress has been made on research in customer churn prediction. There are currently three types of algorithms for customer churn prediction: traditional statistics-based prediction, machine learning-based prediction and classification-based prediction.

The main traditional statistics-based prediction methods are the logit model, linear discriminant analysis and quadratic discriminant analysis. [Jahromi et al. \(2014\)](#) adopted the logit model to study customer churn issues of a B2B e-commerce platform in Australia, and compared the decision tree model with the Boosting algorithm. They found that the logit model can be used to predict customer churn, but the results were not as accurate as other prediction models. Moreover, [Nie et al. \(2011\)](#) developed a logit model for a bank's credit card customer data to identify potential customer churn, and compared the prediction through the decision tree model. The results suggested that the prediction of the logit model was more accurate.

Machine learning predictive models include decision tree (DT), support vector machine (SVM), artificial neural network (ANN) and other algorithms. [De et al. \(2018\)](#) conducted a comparative analysis on various datasets through the decision tree algorithm. The results showed that decision tree was deficient in dealing with linear relationships among variables. [Neslin et al. \(2006\)](#) believed that the decision tree algorithm can be applied as the base model for churn prediction. On the other hand, [Zhang & Zhang \(2015\)](#) conducted churn prediction for the short message services of telecommunication companies, and C5.0 decision tree predictive model was found to have high accuracy. [Farquad et al. \(2014\)](#) performed churn prediction for bank credit card customers and proposed a hybrid method to extract rules from SVM. Moreover, [Gordini & Veglio \(2017\)](#) conducted churn prediction for B2B e-commerce customers, and found that SVM was better in dealing with unbalanced and nonlinear data. [Tian et al. \(2007\)](#) adopted the 2-layer neural network to extract variables from data of telecommunication customers, and proposed a churn prediction model based on artificial neural network. The results indicated that the prediction of this method was more accurate than decision tree and Naive Bayes classifier.

Ensemble classification and prediction methods refer to the combining of some base models and transforming weak classifiers into strong ones through integration. The common ensemble methods are Boosting, Gradient Boosting, AdaBoost and XGBoost. With different base models and ensemble rules, ensemble methods include linear discriminant method ([Xie & Li, 2008](#)), decision tree ([Abbasimehr et al., 2014](#)), support vector machine ([Vafeiadis et al., 2015](#)), and neural network ([Gordini & Veglio, 2014](#)). [Wu & Meng \(2016\)](#) conducted churn prediction of e-commerce customers and improved the classification accuracy of the classifier by reducing the dataset size and combining it with the AdaBoost algorithm, which has a high prediction accuracy. Furthermore, [Ji et al. \(2021\)](#) studied the telecommunication customer dataset with temporal characteristics and adopted the XGBoost hybrid algorithm to filter the features of customer churn. The results showed that the XGBoost hybrid algorithm had good predic-

tion performance. [Ahmed & Maheswari \(2019\)](#) performed churn prediction of customers in the telecommunications industry and proposed a predictive model integrating heuristic algorithms. On the other hand, [Ying et al. \(2010\)](#) conducted churn prediction of bank customers and used integrated LDA and Boosting methods that produced accurate predictions. [Zhang et al. \(2014\)](#) found that an integrated model of CART and adaptive Boosting has a high prediction accuracy.

Most of the aforementioned literature is related to the telecommunications and financial industries, with predictive models producing inconsistent results. Therefore, it is necessary to develop a targeted churn prediction model in B2C, and take into account variables such as the features of the B2C environment—type of product, product collection, adding products to the shopping cart, product preferences and time of shopping. This paper will evaluate the churn prediction performance of AdaBoost ensemble classifier in the B2C environment. The AdaBoost model will also be analyzed in comparison with BP neural network model.

3. Research Method

Customer churn prediction is related to classification research, which involves churned and non-churned customers. Given that customers can shop anytime in a B2C context, factors such as time of shopping and behavioral tendencies in each period may be critical in identifying customer churn. This study uses e-commerce data, this data set contains multiple shopping behavior variables, and can reflect the time attribute of shopping, which is very suitable for the research content of this paper, which was first preprocessed before k-means clustering was adopted to classify customers. AdaBoost and BP neural network were then used for modeling to ascertain the prediction accuracy of the two models and make a comparative analysis. The basic research process is shown in [Figure 1](#).

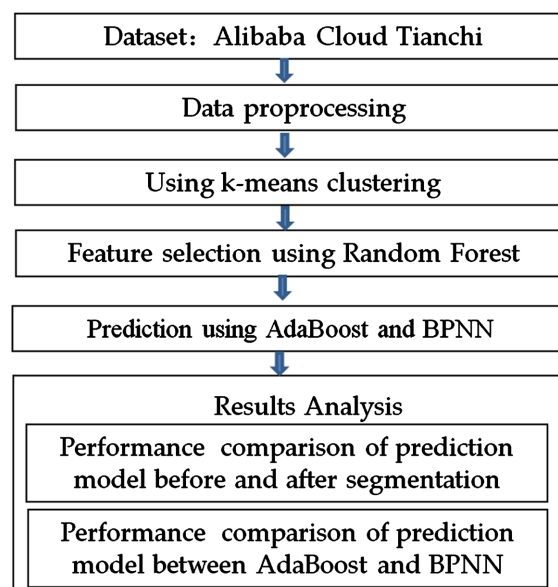


Figure 1. Process of AdaBoost and BP NN for customer churn prediction.

The principles of AdaBoost and BP neural network algorithms are discussed below.

3.1. AdaBoost

AdaBoost is an iterative algorithm first proposed by Yoav Freund and Robert Schapire (Freund & Schapire, 1996). Its core idea is to train different classifiers in the same training set, i.e., weak classifiers that are then combined to form a stronger classifier. The training set sample is:

$$T = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}, \quad w_{li} = \frac{1}{m}; \quad i = 1, 2, \dots, m \quad (1)$$

The output weight of the k^{th} weak learner in the training set is:

$$D(k) = (w_{k1}, w_{k2}, \dots, w_{km}) \quad (2)$$

Churned and non-churned customers are binary classifications with outputs of $\{-1, 1\}$, and the weighted error rate of the k^{th} weak classifier $G_k(x)$ in the training set is:

$$e_k = P(G_k(x_i) \neq y_i) = \sum_{i=1}^m w_{ki} I(G_k(x_i) \neq y_i) \quad (3)$$

The weight coefficient of the k^{th} weak classifier $G_k(x)$ is:

$$\alpha_k = \frac{1}{2} \log \frac{1 - e_k}{e_k} \quad (4)$$

Suppose that the sample set weight coefficient of the k^{th} weak classifier is $D(k) = (w_{k1}, w_{k2}, \dots, w_{km})$, then the sample set weight coefficient of the corresponding $k + 1^{\text{st}}$ weak classifier is:

$$w_{k+1,i} = \frac{w_{ki}}{Z_k} \exp(-a_k y_i G_k(x_i)) \quad (5)$$

AdaBoost classifier uses a weighted voting technique, and the final strong classifier is:

$$f(x) = \text{sign} \left(\sum_{k=1}^K a_k G_k(x) \right) \quad (6)$$

AdaBoost is an advanced ensemble algorithm with a high detection rate and is less prone to overfitting. In each iteration, a weak classifier is first trained in the sample set. As each sample has many attributes, training the optimal weak classifier from a large number of features requires intensive computation.

3.2. BP Neural Network

The Sigmoid transfer function (Minka, 2004) in the BP neural network is a non-linear transformation function. The Sigmoid function is defined as:

$$f(x) = \frac{1}{1 + e^{-x}} \quad (7)$$

The domain of the function is a set of real numbers, with a range between $[0, 1]$.

When x is large enough, it can be regarded as two types of problems: 0 or 1. If x is higher than 0.5, it can be regarded as a Type 1 problem, or Type 0 problem if it is lower.

Using the three-layer perceptron as an example, an output error E can occur when the neural network is not giving the expected output. The output error E is defined as:

$$E = \frac{1}{2}(D - O)^2 = \frac{1}{2} \sum_{k=1}^l (d_k - o_k)^2 \quad (8)$$

After expanding the aforementioned error definition to the hidden layer, then

$$E = \frac{1}{2} \sum_{k=1}^l [d_k - f(\text{net}_k)]^2 \quad (9)$$

The network input error is the function of the weights W_{jk} and V_{ij} of each layer. This means that Error E can be changed by adjusting the weights. Adjusting the network's weights can keep reducing the error. Therefore, the weights should be proportional to the gradient of the error, which is:

$$W_{jk} = -\eta \frac{\partial E}{\partial W_{jk}}, \quad j = 0, 1, 2, \dots, m; \quad k = 1, 2, \dots, l \quad (10)$$

The BP neural network's function and derivatives are continuous, making it easier in processing. Training is fast, and the computation of classification is only related to the number of features, allowing easy interpretation of results for continuous and categorical variables.

4. Experimental Study

The datasets published on the Alibaba Cloud Tianchi platform (Aliyun TIANCHI, 2021) were used in this study. From Nov. 23, 2017 to Dec. 4, 2017, a total of 987,994 customers conducted online shopping. The behavior data of customers included five types of indicators—User ID, Product ID, Item ID, Type of Behavior and Timestamp. There were four types of behavior, including the page views of an item or similar to an item click, buy (purchase an item) cart (adding an item to shopping cart), favorite (adding an item to the favorite list). The original data was preprocessed, and the dates of shopping behaviors were divided into two stages—observation and validation periods. The observation period included the first six days, while and the last six days were used in the validation period. Churned and non-churned customers were defined as follows: Customers who make purchases several times during the observation period and did not make any purchases during the validation period are defined as churned, as expressed by 1. Customers who make purchases several times during the observation period and validation period are defined as non-churned, which is expressed by 0. The customers were first grouped based on User ID, before the number of purchases each customer made within the observation and validation periods were computed, and customers who met the filtering criteria were retained. Some 95,388 pieces of data of 8156 customers were retained, including 7576 churned

customers, accounting for 92.8% of the total and 580 non-churned customers, making up 7.2% of the total.

4.1. Data Preprocessing

The first step of data processing is to convert the timestamp of each raw data item. The converted time format is “year/month/day” and “hour/minute/second”. Throughout the day, customers’ shopping may generate behavioral data such as “Product collection”, “Add to shopping cart”, “Favorites” and “Click”. These behavioral data are rarely mentioned in existing literature, even though such data can play an important role in churn prediction (Cao, 2008). Thus, customer behavioral segmentation was done based on different times in this study. The time periods were divided into: 00:00-06:00 (daybreak), 06:00-12:00 (am), 12:00-18:00 (pm) and 18:00-00:00 (night). The shopping behavior of each customer in these four periods were computed.

There were four types of behavior data, including the PV number, Buy number, Cart number and Favorite number. After classification, 17 types of data variables were used for churn prediction—Items of categories, Daybreak PV, Daybreak Buy, Daybreak Cart and Daybreak Favorite; Am PV, AM buy, AM Cart, AM Favorite; PM PV, PM Buy, PM Cart, PM Favorite; Night PV, Night Buy, Night Cart, Night Favorite.

4.2. Customer Segmentation

Many previous studies did not conduct the classification process before churn prediction in a B2C environment, which may lead to low prediction accuracy and precision. This paper asserts that the segmentation of customers can improve the accuracy of customer churn prediction. Companies can devise targeted and effective marketing strategies based on the shopping behavior of various customer groups, which can identify key customers, general customers and churn-prone customers. Relevant studies (Pham et al., 2004; Chen et al., 2001) showed that k-means clustering is simple and easy for implementation, and has been widely adopted. Thus, this paper uses k-means clustering for customer segmentation. The aforementioned 17 clustering variables were used. The number of clusters, i.e. K is determined in advance for a given sample dataset, so that the samples in the clusters are distributed as close as possible and the distance between clusters is as long as possible. We tested from $K = 2$ to $k = 8$ one by one. When $k = 3$, the distance between clusters was the longest. Thus, the number of clusters is 3, i.e., the customers were segmented into 3 categories—Cluster I, Cluster II and Cluster III. Churned and non-churned customers were further analyzed. According to the collated results, Cluster I involved 4935 customers, including 484 non-churned customers, accounting for 9.8% of the total in Cluster I, and the churn rate was 90.2%. Cluster II had 2697 customers, including 83 non-churned customers, accounting for 3.1% of the total in Cluster II, and the churn rate was 96.9%. Cluster III included 524 customers, with 13 non-churned

customers that make up 2.5% of the total in Cluster III, and the churn rate was 97.5%.

4.3. Feature Selection

The basic method of churn prediction is to incorporate variables into the model as data features for prediction. An excessive number of features will lead to data redundancy and may affect the model's prediction performance (Verbeke et al., 2012). All 17 variables were included as features in the model for prediction, leading to a decrease in prediction accuracy, or even prediction failure. Therefore, the 17 features were filtered first to select those that are suitable for prediction. Random forest is an effective feature selection algorithm with high classification accuracy and good generalization (Breiman, 2001). The key part in the feature selection process is how to select the optimal number of features (M). The out-of-bag (OOB) error (Breiman, 1996) was adopted first, and the number of features was determined based on the minimum OOB error rate before computing the importance of features. The selection of the number of features is shown in Table 1, which indicated that the OOB error rate is the lowest at 0.081 when the number of features is 4. The Gini index can discern the importance of features. The higher the Gini value is, the higher the importance of features (Goldstein et al., 2011). As shown in Table 2, "Night Buy", "PM Buy", "Night PV" and "PM PV" were selected as features for churn prediction. The data in

Table 1. OOB error.

Number of features	4	6	8	10	12
OOB error	0.081	0.089	0.098	0.104	0.104

Table 2. Importance of random forest variable selection.

Variable	Mean Decrease Gini	Mean Decrease Accuracy
Night Buy	135.2431	23.645784
PM Buy	120.5565	26.333866
Night PV	114.7848	-2.905496
PM PV	110.5329	-5.03892
AM PV	102.2129	-6.711931
AM Buy	87.24348	21.823613
Categories	68.46213	16.58795
Daybreak PV	47.48254	-2.350555
Night Cart	41.98615	-3.251536
Daybreak Buy	37.84351	12.998789
PM Cart	31.62181	-2.414922
AM Cart	25.06281	-1.886356

Table 2 below is obtained by the author according to random forest algorithm.

4.4. Evaluation Metrics

Three main metrics were used to evaluate the predictive model’s performance—Accuracy, Recall and Precision. Once the receiver operating characteristic (ROC) curve was drawn, the model can be evaluated based on the area under the curve (AUC) (Fan & Ke, 2010). The formulae of the three metrics are as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \tag{11}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{13}$$

where, True Positive (TP) and True Negative (TN) indicate the number of correct predictions (TP and TN), respectively.

False Positive (FP) and False Negative (FN) indicate the number of incorrect predictions (FP and FN), respectively.

5. Results and Discussion

Data was input into the AdaBoost and BP neural network models for prediction, and iterations were repeated until convergence is achieved. The 10-fold cross-validation method was used to divide the data into 10 parts, where 9 were used as the training set and 1 as the test set. The average value obtained from the 10 tests was used as the final evaluation result of the AdaBoost and BP neural network models. The confusion matrix was obtained after the AdaBoost and BP neural network models were applied to the test set data. The confusion matrices before and after segmentation are shown in **Tables 3-6**. **Figure 2** and **Figure 3**

Table 3. AdaBoost confusion matrix before segmentation

	Predicted		Accuracy	Recall	Precision	AUC
	Predicted Positive (0)	Predicted Negative (1)				
Actual Positive (0)	718	36	0.9446	0.9524	0.9372	0.9931
Actual Negative (1)	48	710				

Table 4. BP neural network confusion matrix before segmentation

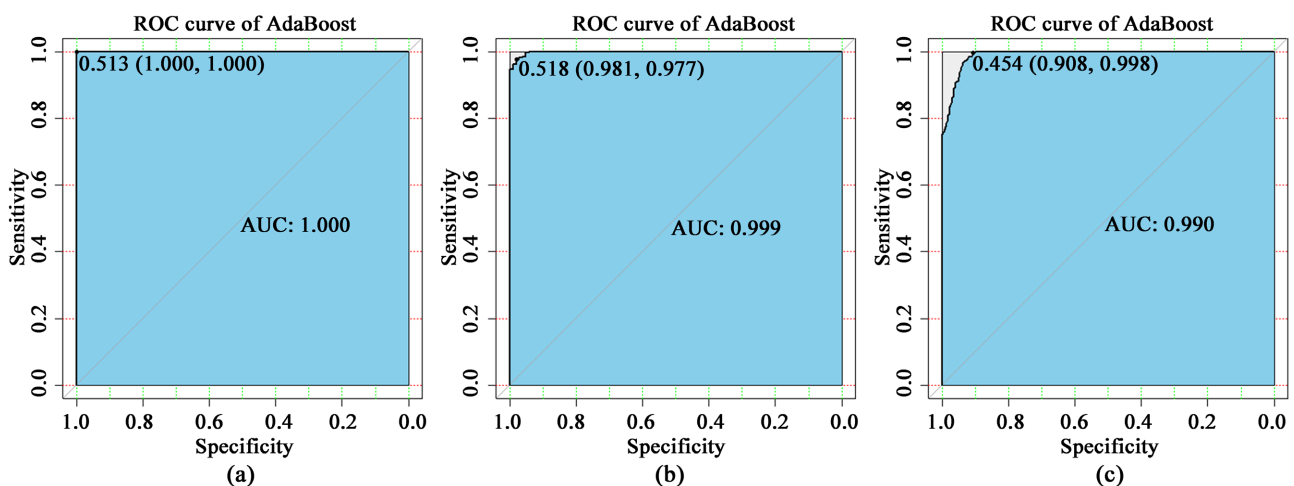
	Predicted		Accuracy	Recall	Precision	AUC
	Predicted Positive (0)	Predicted Negative (1)				
Actual Positive (0)	749	5	0.9082	0.9937	0.9229	0.9374
Actual Negative (1)	134	624				

Table 5. AdaBoost confusion matrix after segmentation.

		Predicted		Accuracy	Recall	Precision	AUC
		Predicted Positive (0)	Predicted Negative (1)				
Cluster I	Actual Positive (0)	11	1	0.9604	0.9145	0.9719	0.9939
	Actual Negative (1)	1	50				
Cluster II	Actual Positive (0)	100	7	0.9663	0.9346	0.9792	0.9956
	Actual Negative (1)	5	256				
Cluster III	Actual Positive (0)	601	34	0.9399	0.9457	0.9303	0.9911
	Actual Negative (1)	31	414				
Avg.				0.9555	0.9316	0.9064	

Table 6. BP neural network confusion matrix after segmentation.

		Predicted		Accuracy	Recall	Precision	AUC
		Predicted Positive (0)	Predicted Negative (1)				
Cluster I	Actual Positive (0)	11	2	0.9287	0.8764	0.9404	0.940
	Actual Negative (1)	3	48				
Cluster II	Actual Positive (0)	105	2	0.9164	0.9766	0.8916	0.963
	Actual Negative (1)	28	233				
Cluster III	Actual Positive (0)	633	2	0.9052	0.9973	0.9940	0.913
	Actual Negative (1)	101	344				
Avg.				0.9167	0.9501	0.942	

**Figure 2.** ROC curve of AdaBoost model after segmentation. (a) Cluster I; (b) Cluster II; (c) Cluster III.

show the ROC curves predicted through the AdaBoost and BP neural network models before and after segmentation. The classification data and prediction index data in the following tables are obtained by computer operation with Python

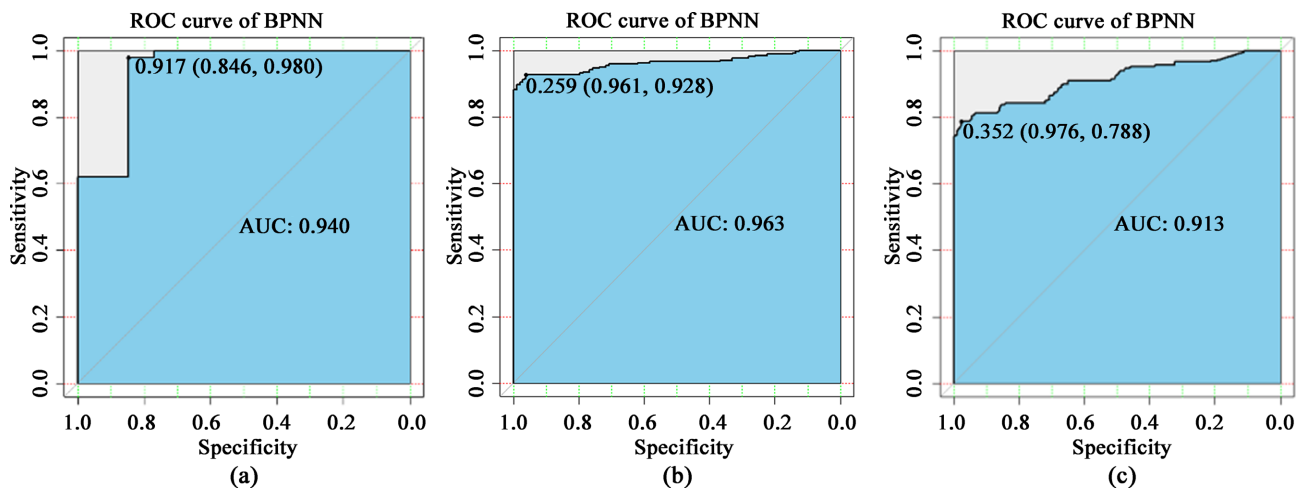


Figure 3. ROC curve of BP neural network model after segmentation. (a) Cluster I; (b) Cluster II; (c) Cluster III.

program in this paper.

5.1. Customer Segmentation Analysis

Segmentation of customers was conducted to identify important and general customers. Enterprises formulate marketing strategies and enhance their products based on the importance of their customer base, while matching products with customer preferences to retain customers and improve marketing performance. According to the results in Section 4.2, the churn rate of Cluster I customers reached 90.2%, but the non-churn rate was 9.8% higher compared with Cluster II and III customers. Cluster III had the lowest number of customers at 524. The proportion and non-churn rate of the three types of customers indicated that Cluster I customers may be important for the enterprise. Hence, the company should further analyze Cluster I customers to develop personalized marketing plans and prevent these customers from churning. This result also showed the effectiveness of segmentation before prediction, which was more targeted. These insights can be applied to data analysis, customer classification and predictive modeling for B2C e-commerce enterprises.

5.2. Performance of Predictive Model

A comparative experiment on AdaBoost and BP neural network models was conducted. The Accuracy, Recall and Precision values of each category were computed based on the confusion matrix to evaluate the performance of the three categories in the dataset. **Table 5** and **Table 6** respectively show the experimental results predicted by AdaBoost and BP neural network models after customer segmentation. It can be seen that the AdaBoost model's prediction accuracy for the three types of customers was higher than the BP neural network model, suggesting that the AdaBoost model was more accurate. All three metrics—Accuracy, Recall and Precision, are important in evaluating the prediction performance of models. On average, the three metrics of the AdaBoost model after customer segmentation were 0.9555, 0.9316 and 0.9604, while those of the BP neural net-

work model were 0.9167, 0.9501 and 0.942. Hence, the prediction performance of the AdaBoost model is more accurate than the BP neural network model.

In addition, the generalization of the models was evaluated using ROC and AUC, where the former can detect the effect of any threshold on the generalization performance of learners. The ROC of the AdaBoost model after segmentation in **Figure 2** demonstrated that when the thresholds of Cluster I, II and III are 0.513, 0.518 and 0.454, the sensitivity values are 1.0, 0.977 and 0.998, the specificity values are 1.0, 0.981 and 0.908 and the AUCs are 1.0, 0.999 and 0.99, respectively. Similarly, the performance metrics of the BP neural network model is shown in **Figure 3**. These experimental data proved that the AdaBoost model has better generalization and prediction performance. Thus, the AdaBoost model is recommended for predicting customer churn in B2C e-commerce over the BP neural network model.

5.3. Customer Management

In B2C e-commerce marketing, customer churn prediction aims to improve enterprises' customer relationship management and increase profits. There is no doubt that developing a model to accurately predict customer churn can have a positive impact on the management and finances of companies. The results of this study will help boost enterprises' customer relationship management. Accurate clustering of customers and identification of important customers will help enterprises maintain their core customer base. Segmentation of churned and non-churned customers allows companies to carry out targeted marketing activities and formulate effective strategies. However, misclassification can occur where non-churned customers might be classified as churned, or churned customers might be classified as non-churned. According to [Coussement \(2014\)](#) and [Viaene & Dedene \(2005\)](#), incorrect identification of churned customers will have a negative impact on enterprises' customer retention measures and profits. Therefore, it is a long-term undertaking for enterprises to accurately classify and predict customer churn, which is particularly important in the highly competitive B2C market. Based on this research, the AdaBoost model is recommended for B2C e-commerce enterprises to accurately identify churned customers and develop effective customer retention measures to reduce administrative costs ([Jahromi et al., 2014](#)).

6. Conclusion

This paper used a set of B2C e-commerce data to conduct customer churn prediction. After data cleaning, the data samples were discretized and 17 variables were collated. Customers were segmented into three categories through k-means clustering before customer churn prediction was performed. The Accuracy, Recall, Precision and AUC of the three categories were computed and the performances of the predictive models were evaluated. The AdaBoost ensemble classification model and BP neural network can effectively predict customer data with a large number of features. Furthermore, adding customer behavior data and

temporal data to the RFM model can better reflect the shopping behavior of B2C customers in the prediction. The results confirmed the importance of this method in customer churn prediction and marketing decision-making, and showed the promising role of AdaBoost model in establishing an effective early warning model for customer churn management in the B2C industry. However, the results may have some limitations due to dataset issues and generality of predictive models. This paper did not study the persistence of predictive models, which refers to the prediction performance of models over a specific number of periods after the estimation period. If the persistence of a prediction model is limited, the model cannot be used for a long time in the B2C environment. Therefore, it may be necessary to establish new models for the prediction of customer churn over extended periods.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

- Abbasimehr, H., Setak, M., & Tarokh, M. J. (2014). A Comparative Assessment of the Performance of Ensemble Learning in Customer Churn Prediction. *The International Arab Journal of Information Technology*, 11, 599-606.
- Ahmed, A. A. Q., & Maheswari, D. (2019). An Enhanced Ensemble Classifier for Telecom Churn Prediction Using Cost Based Uplift Modeling. *International Journal of Information Technology*, 11, 381-391. <https://doi.org/10.1007/s41870-018-0248-3>
- Alboukaey, N., Joukhadar, A., & Ghneim, N. (2020). Dynamic Behavior Based Churn Prediction in Mobile Telecom. *Expert Systems with Applications*, 162, Article ID: 113779.
- Aliyun TIANCH (2021). *Alibaba Cloud Tianchi Data Sets*. <https://tianchi.aliyun.com/dataset/dataDetail?dataId=649>
- Amin, A., Anwar, S., Adnan, A., Adnan, A., Nawaz, M., Alawfi, K., Hussain, A. et al. (2017). Customer Churn Prediction in the Telecommunication Sector Using a Rough Set Approach. *Neurocomputing*, 237, 242-254.
- Bi, Q. Q. (2019). Cultivating Loyal Customers through Online Customer Communities: A Psychological Contract Perspective. *Journal of Business Research*, 103, 34-44.
- Bock, K. W. D., & De, C. A. (2021). Spline-Rule Ensemble Classifiers with Structured Sparsity Regularization for Interpretable Customer Churn Modeling. *Decision Support Systems*, 150, 1-14.
- Breiman, L. (1996). Bagging Predictors. *Machine Learning*, 24, 123-140. <https://doi.org/10.1007/BF00058655>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5-32. <https://doi.org/10.1023/A:1010933404324>
- Cao, L. (2008). Behavior Informatics and Analytics: Let Behavior Talk. In F. Bonchi, B. Berendt, F. Giannotti, D. Gunopulos, F. Turini, C. Zaniolo, N. Ramakrishnan, & X.D. Wu (Eds.), *2008 IEEE International Conference on Data Mining Workshops* (pp. 87-96). IEEE Computer Society Press. <https://doi.org/10.1109/ICDMW.2008.95>
- Chen, M. C., Chiu, A. L., & Chang, H. H. (2005). Mining Changes in Customer Behavior in Retail Marketing. *Expert Systems with Applications*, 28, 773-781.

- Chen, N., Chen, A., & Zhou, L. (2001). An Effective Clustering Algorithm in Large Transaction Databases. *Journal of Software*, *12*, 476-484.
- Chung, B. D., Park, J. H., Koh, Y. J., & Lee, S. (2016). User Satisfaction and Retention of Mobile Telecommunications Services in Korea. *International Journal of Human-Computer Interaction*, *32*, 532-543.
<https://doi.org/10.1080/10447318.2016.1179083>
- Coussement, K. (2014). Improving Customer Retention Management through Cost-Sensitive Learning. *European Journal of Marketing*, *48*, 477-495.
<https://doi.org/10.1108/EJM-03-2012-0180>
- Coussement, K., Lessmann, S., & Verstraeten, G. (2017). A Comparative Analysis of Data Preparation Algorithms for Customer Churn Prediction: A Case Study in the Telecommunication Industry. *Decision Support Systems*, *95*, 27-36.
- Daoud, R. A., Amine, A. E., Bouikhalene, B., & Lbibb, R. (2018). Clustering Prediction Techniques in Defining and Predicting Customers Defection: The Case of E-Commerce Context. *International Journal of Electrical and Computer Engineering*, *8*, 2367-2383.
<https://doi.org/10.11591/ijece.v8i4.pp2367-2383>
- De, C. A., Coussement, K., Verbeke, W., Idbenjra, K., & Phan, M. (2021). Uplift Modeling and Its Implications for B2B Customer Churn Prediction: A Segmentation-Based Modeling Approach. *Industrial Marketing Management*, *99*, 28-39.
- De, C. A., Coussement, K., & De Bock, K. W. (2018). A New Hybrid Classification Algorithm for Customer Churn Prediction Based on Logistic Regression and Decision Trees. *European Journal of Operational Research*, *269*, 760-772.
- Devriendt, F., Berrevoets, J., & Verbeke, W. (2021). Why You Should Stop Predicting Customer Churn and Start Using Uplift Models. *Information Sciences*, *548*, 497-515.
- Dumitrescu, E., Hue, S., Hurlin, C., & Tokpavi, S. (2022). Machine Learning for Credit Scoring: Improving Logistic Regression with Non-Linear Decision-Tree Effects. *European Journal of Operational Research*, *297*, 1178-1192.
- Eichinger, F., Nauck, D. D., & Klawonn, F. (2006). Sequence Mining for Customer Behaviour Predictions in Telecommunications. In S. Bickel (Ed.), *The Workshop on Practical Data Mining: Applications, Experiences and Challenges* (pp. 3-10). Springer.
- Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005). "Counting Your Customers" the Easy Way: An Alternative to the Pareto/NBD Model. *Marketing Science*, *24*, 275-284.
<https://doi.org/10.1287/mksc.1040.0098>
- Fan, X., & Ke, T. (2010). Enhanced Maximum AUC Linear Classifier. In M. Z. Li, Q. L. Liang, L. P. Wang, & Y. B. Song (Eds.), *International Conference on Fuzzy Systems and Knowledge Discovery* (pp. 1540-1544). Institute of Electrical and Electronics Engineers.
- Farquard, M. A. H., Ravi, V., & Raju, S. B. (2014). Churn Prediction Using Comprehensive Support Vector Machine: An Analytical CRM Application. *Applied Soft Computing*, *19*, 31-40.
- Freund, Y., & Schapire, E. R. (1996). Experiments with a New Boosting Algorithm. In L. Saitta (Ed.), *Machine Learning: The Thirteenth International Conference* (pp. 1-9). Morgan Kaufmann.
- Gattermann-Itschert, T., & Thonemann, U. W. (2021). How Training on Multiple Time Slices Improves Performance in Churn Prediction. *European Journal of Operational Research*, *295*, 664-674.
- Goldstein, B. A., Polley, E. C., & Briggs, F. B. S. (2011). Random Forests for Genetic Association Studies. *Statistical Applications in Genetics and Molecular Biology*, *10*, 32.

<https://doi.org/10.2202/1544-6115.1691>

- Gordini, N., & Veglio, V. (2014). Using Neural Networks for Customer Churn Prediction Modeling: Preliminary Findings from the Italian Electricity Industry. In *Convegno Annuale della Società Italiana Marketing: "Smart Life. Dall'Innovazione Tecnologica al Mercato* (pp. 1-13). Università degli Studi di Milano-Bicocca.
- Gordini, N., & Veglio, V. (2017). Customers Churn Prediction and Marketing Retention Strategies, an Application of Support Vector Machines Based on the AUC Parameter-Selection Technique in B2B E-Commerce Industry. *Industrial Marketing Management*, 62, 100-107.
- Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2007). Computer Assisted Customer Churn Management: State-of-the-Art and Future Trends. *Computers & Operations Research*, 34, 2902-2917.
- Jahromi, A. T., Stakhovych, S., & Ewing, M. (2014). Managing B2B Customer Churn, Retention and Profitability. *Industrial Marketing Management*, 43, 1258-1268.
- Ji, H., Ni, F., & Liu, J. (2021). Prediction of Telecom Customer Churn Based on XGB-BFS Feature Selection Algorithm. *Computer Technology and Development*, 31, 21-25.
- Kotler, P., & Keller, K. (2008). *Marketing Management: International Edition*. Prentice Hall.
- Kotler, P., & Keller, K. (2016). *Marketing Management* (15nd ed., pp. 89-120.). Pearson Education Ltd.
- Kozak, J., Kania, K., Juszczuk, P., & Mitreġa, M. (2021). Swarm Intelligence Goal-Oriented Approach to Data-Driven Innovation in Customer Churn Management. *International Journal of Information Management*, 60, 16 p.
- Minka, T. P. (2004). *A Comparison of Numerical Optimizers for Logistic Regression*. Technical Report (Mathematics), 18 p.
- Neslin, S. A., Gupta, S., Kamakura, W., Lu, J. X., & Mason, C. H. (2006). Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models. *Journal of Marketing Research*, 43, 204-211. <https://doi.org/10.1509/jmkr.43.2.204>
- Nie, G. L., Rowe, W., Zhang, L. L., Tian, Y. J., & Shi, Y. (2011). Credit Card Churn Forecasting by Logistic Regression and Decision Tree. *Expert Systems with Applications*, 38, 15273-15285.
- O'Brien, M., Liu, Y., Chen, H. Y., & Lusch, R. (2020). Gaining Insight to B2B Relationships through New Segmentation Approaches: Not All Relationships Are Equal. *Expert Systems with Applications*, 161, 1-11.
- Orsenigo, C., & Vercellis, C. (2010). Combining Discrete SVM and Fixed Cardinality Warping Distances for Multivariate Time Series Classification. *Pattern Recognition*, 43, 3787-3794.
- Pham, D. T., Dimov, S. S., & Nguyen, C. D. (2004). Selection of K in K -Means Clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 219, 103-119. <https://doi.org/10.1243/095440605X8298>
- Schaeffer, S. E., & Sanchez, S. V. R. (2020). Forecasting Client Retention-A Machine-Learning Approach. *Journal of Retailing and Consumer Services*, 52, Article ID: 101918. <https://doi.org/10.1016/j.jretconser.2019.101918>
- Schmittlein, D. C., Morrison, D. G., & Colombo, R. (1987). Counting Your Customers: Who Are They and What Will They Do Next? *Management Science*, 33, 1-24. <https://doi.org/10.1287/mnsc.33.1.1>
- Tian, L., Qiu, H., & Zheng, L. (2007). Telecom Churn Prediction Modeling and Application Based on Neural Network. *Computer Applications*, 27, 2294-2297.

- Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A Comparison of Machine Learning Techniques for Customer Churn Prediction. *Simulation Modelling Practice and Theory*, 55, 1-9. <https://doi.org/10.1016/j.simpat.2015.03.003>
- Velez, D., Ayuso, A., Perales-González, C., & Rodriguez, J. T. (2020). Churn and Net Promoter Score Forecasting for Business Decision-Making through a New Stepwise Regression Methodology. *Knowledge-Based Systems*, 196, Article ID: 105762. <https://doi.org/10.1016/j.knsys.2020.105762>
- Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New Insights into Churn Prediction in the Telecommunication Sector: A Profit Driven Data Mining Approach. *European Journal of Operational Research*, 218, 211-229. <https://doi.org/10.1016/j.ejor.2011.09.031>
- Viaene, S., & Dedene, G. (2005). Cost-Sensitive Learning and Decision Making Revisited. *European Journal of Operational Research*, 166, 212-220. <https://doi.org/10.1016/j.ejor.2004.03.031>
- Wu, X., & Meng, S. (2016). E-Commerce Customer Churn Prediction Based on Customer Segmentation and AdaBoost. In *13th International Conference on Service Systems and Service Management* (pp.). Institute of Electrical and Electronics Engineers.
- Xie, Y. Y., & Li, X. (2008). Churn Prediction with Linear Discriminant Boosting Algorithm. In *International Conference on Machine Learning and Cybernetics* (pp. 228-233). Institute of Electrical and Electronics Engineers.
- Ying, W. Y., Lin, N., Xie, Y. Y., & Li, X. (2010). Research on the LDA Boosting in Customer Churn Prediction. *Journal of Applied Statistics & Management*, 29, 400-408.
- Zhang, W., Yang, S., & Liu, T. (2014). Customer Churn Prediction in Mobile Communication Enterprises Based on CART and Boosting Algorithm. *Chinese Journal of Management Science*, 22, 90-96.
- Zhang, Y., & Zhang, Z. M. (2015). A Customer Churn Alarm Model Based on the C5.0 Decision Tree-Taking the Postal Short Message as an Example. *Statistics & Information Forum*, 30, 89-94.
- Zhao, M., Zeng, Q. J., Chang, M., Tong, Q., & Su, J. F. (2021). A Prediction Model of Customer Churn Considering Customer Value: An Empirical Research of Telecom Industry in China. *Discrete Dynamics in Nature and Society*, 2021, Article ID: 7160527.