Annals of Operations Research
https://doi.org/10.1007/s10479-022-04526-5

ORIGINAL RESEARCH

A PCA-AdaBoost model for E-commerce customer churn prediction

Zengyuan Wu[1] · Lizheng Jing[1] · Bei Wu[2] · Lingmin Jin[1]

Accepted: 3 January 2022

Abstract
Customer churn is detrimental to corporate revenue. Hence, accurate customer churn prediction is vital for enterprises to improve customer retention and corporate revenue. However, in e-commerce, there are challenges when predicting customer churn using traditional models. First, the conventional recency, frequency, and monetary (RFM) analysis based on single-source data can hardly accurately predict the e-commerce customer churn since such customers are often non-contractual. Second, in general, the data about e-commerce customers is high-dimensional and unbalanced, making traditional models even less effective. In this paper, we propose an improved prediction model by integrating data pre-processing and ensemble learning to solve these two problems. Specifically, two new features are first integrated into the RFM analysis to better capture customer behaviors. Second, the principal component analysis is adopted to reduce data dimensions. Third, adaptive boosting (AdaBoost) is employed to cascade multiple decision trees to minimize the impacts from the unbalanced data. For clarity, this model is called the PCA-AdaBoost model. We use an e-commerce dataset published on the Kaggle platform to demonstrate its effectiveness by conducting numerical experiments. We compare the performance of the PCA-AdaBoost model developed in this paper with several models proposed in the literature. Our results confirm that the PCA-AdaBoost model can achieve more accurate customer churn prediction and better overall stability.

Keywords E-commerce · Customer churn prediction · Ensemble learning · PCA-AdaBoost model

1 Introduction

With the further popularization of the Internet, online shopping has become an even more important way for people to purchase products and services (Tsai et al., 2009). According to "Digital 2020-Global Digital Overview," the world's internet users exceeded 4.5 billion in

✉ Bei Wu
wubeicoco@163.com

1 China Jiliang University, Hangzhou 310018, People's Republic of China

2 Zhejiang Gongshang University, Hangzhou 310018, People's Republic of China

2020; more than 90% of them shop online. The rapid expansion of online shoppers drives more enterprises to develop e-commerce platforms. As a result, their competition for customers becomes increasingly fierce.

Different from the customer relationship at traditional enterprises, the customer relationship at an e-commerce platform is often non-contractual. Termination of a contractual relationship with customers is predictable, while it is difficult to define and predict non-contractual relationships. Therefore, customer relationship management has become an important issue for an e-commerce platform (Azeem et al., 2017; Renjith et al. 2015). In particular, it is a challenge for an e-commerce platform to predict customers' purchase behavior. In the e-commerce environment, customers can switch from one provider to another at will, and customer churn is a widespread phenomenon. The cost of retaining an existing customer is often much lower than acquiring a new customer. Therefore, an e-commerce platform needs to minimize customer churn rates (Qi et al., 2008). To this end, managers can focus on the churners, communicate with them quickly, and take targeted strategies before losing them.

The existing research on customer churn prediction mainly focuses on selecting features and improving prediction methods. In terms of selecting features, there are two research streams. Scholars select features based on customer information data and consumption records in the first stream. For example, Muhammad (2017) built a churn prediction model for prepaid customers in the telecommunications industry. Scholars select features based on the RFM customer value analysis (Vo et al., 2021). They usually extract customer value features from customer consumption data and further improve the RFM analysis. Coussement et al. (2021) built a recommendation system by adding relationship variables to customer behavior in addition to the RFM variables. Martínez et al. (2021) presented the fuzzy linguistic 2-tuple RFM model combined with product hierarchy based on analytic hierarchy process (AHP). For non-contractual e-commerce customers, both transaction length (Chen et al., 2015) and customer satisfaction are important factors to be considered to avoid customer churn. However, most existing research omits these factors. Therefore, it is necessary to integrate these two features into the RFM analysis.

Traditional statistics-based methods are also commonly adopted to predict e-commerce customer churn. However, they are less effective for large-scale data. Therefore, it is necessary to employ ensemble learning methods based on artificial intelligence to predict e-commerce customer churn (Wang et al., 2016; Ji et al. 2017). As a branch of machine learning, ensemble learning is an effective tool for data mining. A powerful classifier can be constructed by cascading multiple weak classifiers, such as logistic regression (LR) and support vector machine (SVM). Boosting and Bagging algorithms are more widely used in ensemble learning, and they can effectively reduce prediction error and possess satisfactory generalization ability. Overall, ensemble learning usually performs better compared with a single classifier such as logistic regression, SVM, and Bayesian classifier (Talayeh et al., 2019). However, its training process cascading multiple base classifiers is more complex and time-consuming. Furthermore, an ensemble learning algorithm may not perform well in the presence of high-dimension data.

Therefore, this research focuses on the following question: how to devise an algorithm based on machine learning to predict e-commerce customer churn? Several related models have been proposed in other applications, which partly motivated our research. For example, Qu et al. (2021) developed an ensemble model for electricity theft detection based on genetic optimization, which integrated the synthetic minority oversampling technique (SMOTE) for resampling data, principal component analysis (PCA) for reducing dimensions, and an

ensemble deep learning network based on AdaBoost. Xiao et al. (2020) proposed a positioning the optimal occlusion area (POOA) algorithm for occlusion face detection, where PCA and AdaBoost were adopted to improve the detection precision. Chen et al. (2020) proposed a new algorithm for ultrasound image segmentation, which integrated the PCA method and AdaBoost algorithm. However, no applicable algorithms have been proposed in e-commerce customer churn prediction. Hence, this paper contributes to predicting e-commerce customer churn by proposing a novel model that integrates a pre-processing data method and an ensemble learning algorithm.

The main contributions are as follows. First, the existing literature on customer churn prediction is limited to contractual customers and relies on data from a single source. In contrast, we focus on e-commerce customers, which are mostly non-contractual. Second, transaction length and customer satisfaction are important factors to consider to minimize customer churn. As such, we improve traditional customer value models by integrating these two features into the RFM analysis. Third, existing customer churn prediction models may not perform well in the presence of high-dimension and unbalanced data, and the training process of ensemble learning will be time-consuming. Our PCA-AdaBoost model can overcome this drawback, which can achieve more accurate and stable results.

The rest is organized as follows in this paper. Section 2 reviews the related literature on churn prediction and machine learning algorithms. In Sect. 3, our improved model is proposed. In Sect. 4, the data sources and data pre-processing process are described in detail. In Sect. 5, the results of cross-validation are shown. In Sect. 6, theoretical and practical contributions are discussed. In the last section, the conclusions are drawn, and ideas for future research are proposed.

2 Literature review

Customers can be divided into contractual and non-contractual customers (Zeineb et al., 2019). Customers in traditional industries are generally contractual ones who are bound by contracts and cannot quit at will (Coussement et al., 2008; Verbeke et al., 2012; Mahajan et al., 2020). Therefore, enterprises can identify churners with relative ease. However, e-commerce customers are often non-contractual. They can switch the consumption platform at will. In general, it is difficult for enterprises in the e-commerce industry to identify the churners effectively. They need to consider customer characteristics and choose appropriate forecasting methods. For non-contractual customers, the feature selection based on customer value is an important research issue in customer churn prediction (Montiel et al. 2020; Hughes et al. 2005; Dhote et al., 2020; Chen et al., 2015). Customer value refers to the revenue expectation created by customers for the enterprise. Hughes (2005) first proposes the RFM analysis, which is the basis of customer value assessment. This model can effectively measure customer values and identify their status. Researchers have made a series of improvements to the RFM analysis. For example, while studying customer churn in the logistics industry, Chen et al. (2015) extended the RFM analysis and constructed a prediction model called LRFMP. Their results showed that the transaction length was important. Zhou et al. (2021) introduced the RFMT analysis, where a new dimension called Interpurchase Time was incorporated. They used the RFMT analysis to parse consumers' online purchase sequences. Their model was shown to be capable of tracking and discerning customer purchasing behaviors during the entire shopping cycle.

Two major types of methods, including traditional statistical methods for prediction and machine learning prediction methods based on artificial intelligence, have been utilized to study customer churn. Traditional statistical prediction methods mainly use some algorithms to design more accurate classifiers, such as decision trees (Zacharis, 2018), logistic regression (Miguis et al., 2012), and Bayes classifier (Kisioglu et al. 2011; Hossain et al., 2019). These algorithms are highly interpretable and can also be highly accurate for small-scale data. However, these traditional statistical methods have to make multiple assumptions before establishing classification models (Vo et al., 2021). In reality, these assumptions may not be suitable for e-commerce customer churn data. As a result, these methods are less accurate. They can become even less accurate in the presence of large-scale data, diverse data dimensions, and nonlinear features.

Machine learning methods based on artificial intelligence have also been applied for customer churn prediction. In particular, in recent years, prediction algorithms based on deep learning (Martin et al. 2018) and ensemble learning (Guo et al. 2017; Li et al., 2020) have been developed. As a further development of traditional ANN, the deep neural network can better fit continuous numerical functions. Unlike image and voice data, e-commerce customer data is mostly discrete with pronounced nonlinear features. Therefore, these data are difficult to be fitted by the neural network, and their network structure cannot be determined. In order to solve these problems, scholars have refined traditional multi-layer perceptron models by introducing particle swarm optimization algorithms (Yu et al., 2018), stacked autoencoders, and entity embedding methods. As a branch of machine learning, ensemble learning improves the final learning effect by training and cascading multiple weak classifiers such as Logistic regression, SVM, and ANN (Demirer et al., 2017). Among them, Boosting and Bagging algorithms are widely used. The ensemble learning method can effectively reduce the prediction error and improve the generalization ability. Caigny et al. (2018) used a hybrid algorithm of logistic regression and decision tree to build a prediction model with satisfactory prediction accuracy and interpretability. Jayaswal et al. (2016) proposed an integrated and more accurate method to predict customer churn using the combination of Decision Tree, Gradient Boosted Decision Tree, and Random Forest.

Although progress has been made in customer churn prediction research after introducing ensemble learning, there are still several limitations. First, existing research on feature selection generally focuses on customer value features, such as demographics and customer purchase behavior. However, some new customer features are less studied. Second, existing research mainly relies on a single base classifier. Compared with a single base classifier, the ensemble learning method can improve the prediction performance for unbalanced data. However, it is still difficult for ensemble learning to predict the churners in the e-commerce industry, partly because the data of e-commerce customers can exhibit dual features of high-dimension and imbalance. Incorporating these features into the prediction model by simple cleaning will lead to high complexity, long training time, and poor prediction performance.

Overall, this paper takes e-commerce customers as the research object and contributes by addressing the gaps in existing research. First, we integrate two new features into RFM analysis during the process of feature selection. Second, we propose a PCA-AdaBoost prediction model based on the AdaBoost algorithm. Finally, we conduct experiments on a dataset of e-commerce customers in the Kaggle platform. In order to test the effectiveness, we compare the performance of the PCA-AdaBoost model with other algorithms, including SVM, Logistic regression, and AdaBoost algorithm.

3 PCA-AdaBoost model

This section proposes our PCA-AdaBoost model for customer churn prediction in e-commerce in detail. Specifically, the improved prediction model integrates a data preprocessing method and an ensemble learning algorithm to handle high-dimensioned and unbalanced datasets. Researches have shown that data mining applications in many scenarios, such as credit scoring (Koutanaei et al., 2015) and wind forecasting (Feng et al., 2017), benefit from feature selection algorithms and ensemble classifiers. According to the data processing flow, the proposed PCA-AdaBoost model includes three components, namely RFM analysis, PCA method, and AdaBoost algorithm. Among them, the PCA method and AdaBoost algorithm are the core of the proposed model. To facilitate understanding the principles of the proposed model clearly, we first introduce the AdaBoost algorithm and PCA method and then give an overview and summary of the entire model.

3.1 AdaBoost algorithm

The AdaBoost algorithm, short for Adaptive Boosting, was proposed by Freund and Schapire (1999). It is a boosting technique that transforms weak learners or predictors into strong predictors to solve classification problems. The AdaBoost algorithm is initially applied to online distribution and weighted voting research as a concrete realization of Boosting's ensemble learning method. Its core idea is to increase the weight of misclassified samples after each round of iteration so that the next sub-classifier will focus on the samples that are more likely to be misclassified. In this way, an improved base classifier can be obtained. Then, the final classifier is integrated by weighted voting. In this work, we take binary classification as an example to explain the training process of the AdaBoost algorithm.

Input: training dataset $K = \{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$, where $x_i (i = 1, 2, 3 \cdots, m)$ is an n-dimensional column vector in the input space X, y_i is the label of the i th sample in the output category Y, and $Y = \{-1, +1\}$.

(1) Initialize the weights in the training dataset:

$$D^1(i) = \frac{1}{N}(1 \le i \le N). \tag{1}$$

(2) Set number of iterations t, $t = 1, 2, 3 \cdots, T$,

(a) Use D^T to train base classifier $h : t \to Y$;
(b) Calculate weight parameters β_t;
(c) Update sample weights:

$$D^{(t+1)}(i) = D^t(i) \exp(-\beta_t y_i h_t(x_i))/Z_t. \tag{2}$$

$$Z_t = \sum D^t(i) \exp(-\beta_t y_i h_t(x_i)). \tag{3}$$

Output: the final strong classifier is obtained as.

$$H(x) = sign\left(\sum_{t=1}^{T} \beta_t h_t(x)\right). \tag{4}$$

In the AdaBoost algorithm, boosting is used to reduce bias as well as the variance for supervised learning. It works on the principle of learners growing sequentially, and each

subsequent classifier is grown from previously grown classifiers except for the first classifier. In other words, weak learners are converted into strong ones. Therefore, the AdaBoost algorithm can be used to boost the performance of this prediction task which is based on binary classification problems.

3.2 PCA Method

Principal Component Analysis (PCA) is a dimensionality reduction method that is often used to reduce the dimensionality of a data set consisting of a large number of interrelated variables while retaining as much as possible the variation present in the data set. This is achieved by transforming a large set of variables into a new set of variables, called the principal components (PCs), which are uncorrelated and ordered so that the first few variables retain most of the variation present in all of the original variables.

The core idea of the PCA algorithm is the Karhunen–Loeve (K-L) transformation theory. Let Y be an n-dimensional random vector, and it is expressed as.

$$Y = \sum_{i=1}^{n} \beta_i \varphi_i = \Phi_w. \tag{5}$$

where $w = (\beta_1, \beta_2, \cdots \beta_n)$ is a weighting factor, and $\Phi = (\varphi_1, \varphi_2, \cdots \varphi_n)$ is an n-dimensional vector. Vector β yields

$$\beta_i \beta_j = \begin{cases} \lambda_i, \ i = j, \\ 0, \ i \neq j. \end{cases} \tag{6}$$

i.e.,

$$ww^T = C_\lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}. \tag{7}$$

Suppose $\Phi^T \Phi = I$ and P is the autocorrelation matrix of vector Y. It follows from (5) that.

$$w = \Phi^T y. \tag{8}$$

$$P = yy^T = \Phi w w^T \Phi^T = \Phi C_\lambda \Phi^T. \tag{9}$$

$$P\Phi = \Phi C_\lambda \Phi^T = \Phi C_\lambda. \tag{10}$$

$$P\varphi_i = \lambda_i \varphi_i. \tag{11}$$

where λ_i is the eigenvalue of the autocorrelation matrix P, and φ_i is the corresponding feature vector. The implementation steps of the PCA algorithm are as follows. First, it follows from (5) that the vector y can be expressed as.

$$y = \sum_{i=1}^{m} \beta_i \varphi_i + \sum_{i=m+1}^{n} \beta_i \varphi_i. \tag{12}$$

Then the error vector Δy can be calculated as.

$$\Delta y = y - \hat{y} = \sum_{i=m+1}^{n} \beta_i \varphi_i. \tag{13}$$

$$\varepsilon^2 = \Delta y^2 = \sum_{i=m+1}^{n} \beta_i^2 = \sum_{i=m+1}^{n} \lambda_i. \tag{14}$$

According to Eq. (14), to minimize the mean square error ε^2, the sum of the last $n - m$ eigenvalues of the matrix P should be minimized. We use Φ_m to replace Φ. Φ_m is composed of eigenvectors corresponding to the first m eigenvalues of the autocorrelation matrix P. Then,

$$w_m = \Phi_m^T y. \tag{15}$$

where the vector w_m is the principal component after dimensionality reduction. The original vector is mapped to the new coordinate system through the above steps.

3.3 Overview of PCA-AdaBoost model

As mentioned above, the proposed PCA-AdaBoost model is mainly divided into two phases. In the first phase, the PCA method is used to reduce the data dimensions. In the second phase, the AdaBoost algorithm is employed to cascade multiple decision trees to minimize the impacts from the unbalanced data. We use the PCA algorithm to map the initial dataset to a new feature space in model training. The AdaBoost algorithm can achieve better performance when the base classifier is established in a new feature space. Since the extracted principal components contain the most information in the original dataset, the classification accuracy can be improved. In addition, because the original dataset contains many complex features, we employ the PCA method to map the original dataset to the low-dimensional space, which greatly reduces the running time of the AdaBoost algorithm. Thus, the PCA-AdaBoost model can achieve higher accuracy of customer churn prediction as well as higher efficiency.

The proposed PCA-AdaBoost model works as the following steps. Assuming that the obtained original dataset K contains N samples with l attributes, the dataset can be represented by a $N \times l$ matrix X.

Step 1:

(a) Use PCA to reduce the dimensionality of the original dataset;
(b) Calculate the covariance matrix P of the sample features;
(c) Calculate and select the eigenvalue combination λ_m that minimizes the mean square error to obtain the transformation matrix Φ_m;
(d) Use K' to denote the converted dataset $(X\Phi_m, y)$.

Step 2: Initializing the weight of the training sample K', which yields.

$$D^1(i) = \frac{1}{N}. \tag{16}$$

Step 3: For each iteration $t \in [1, 2, 3 \cdots, T]$:

(a) Train a weak classifier on the sample distribution $\left\{K', D^t\right\}$:

$$h_t : x \to \{-1, +1\}. \tag{17}$$

(b) Calculate classification error rate:

$$\varepsilon_t = \sum_{i=1}^{N} D^t(i) I(h_t(x_i) \neq y_i). \tag{18}$$

(c) If the classification error rate is greater than 0.5, proceed to the next iteration. If the classification error rate is less than or equal to 0.5, switch to step (d).

(d) Assume.

$$\alpha_t = \frac{1}{2} ln \frac{1-\varepsilon_t}{\varepsilon_t}. \tag{19}$$

(e) Update the sample weight as follows:

$$D^{(t+1)}(i) = D^t(i) \exp(-\alpha_t (I(h_t(x_i) = y_i)))/Z_t. \tag{20}$$

$$\sum_i D^{(t+1)}(i) = 1. \tag{21}$$

where Z_t is the normalization operator.

Step 4: Output the final classifier.

$$H(x) = sign\left(\sum_{t=1}^{T} \beta_t h_t(x)\right). \tag{22}$$

4 Data and pre-processing

4.1 Defining Churned Customers

In this study, to form a complete training dataset, we need to extract predictive features such as length, recency, frequency, monetary consumption, and customer satisfaction from the original dataset. Furthermore, we need to obtain the classification labels of churned and non-churned customers.

The classification criteria between churned and non-churned customers are based on the customer's life cycle. An e-commerce customer's life cycle has four stages: new customers, active customers, inactive customers, and churners. New customers are those who have never used the e-commerce platform. After a customer completes the first online shopping, they become an active customer. After that, if an active customer does not shop on the platform over a special time span (the transition period), the customer's status will be switched to inactive. If an inactive customer does not shop on the platform after a long period of time (which is called the forecast period), we then conclude that the customer is churned. We divide the customer's life cycle into three time periods, including the observation period, the transition period, and the forecast period. According to the estimation of SMC activity on the e-commerce platform (Wübben and Wangenheim 2008), we set the transition period and the forecast period to be three months in this article. If a customer has shopped on the platform at least once during the transition period, his label is marked as 1. Otherwise, his label is marked as 0. In the forecast period, we mark it in the same way. Therefore, if a customer has not used the platform in both the transition period and forecast period, they will be defined as a churner. The classifications of churned and non-churned customers are shown in Table 1.

Table 1 The consumption states of churned customers and non-churned customers

Transition period	Forecast period	Churned customers
0	0	Yes
0	1	No
1	0	No
1	1	No

Table 2 Fields and their descriptions

Field	Field description
Order purchase timestamp	Show purchase timestamp
Customer ID	This is the key to the ordered dataset. It is unique for each order
Payment value	Trading price
Order ID	The unique identifier of the order
Review score	Customer satisfaction (with rating scales from 1 to 10)

4.2 Data description

The dataset used in this paper comes from customer consumption data published on the Kaggle competition platform. The dataset contains information about 100,000 orders from 2016 to 2018. This platform allows viewing orders from multiple dimensions: order status, price, payment, shipping cost, customer locations, product attributes, and customer reviews. These fields are listed in Table 2. We find a strong correlation between consumer behavior and the month when pre-processing the dataset. Therefore, in order to reduce the influence of other potential interfering factors on the prediction results, the observation period is set to be one year in this paper. There are 50,713 samples. We found 43,031 churned customers and 7682 non-churned customers by applying the discriminant method. The ratio between churned customers and non-churned customers is about 6: 1. Hence, this is an imbalanced dataset.

4.3 Data pre-processing

Data pre-processing in this paper mainly includes data cleaning and organizing, feature extracting, one-hot encoding, normalization, and PCA for dimension reducing.

(1) *Data cleaning and organizing*
First, we deal with missing values and outliers of the dataset, such as the empty value. Second, we conduct data deduplication. For example, a small number of customers repeatedly buy or browse the same products within one hour. In actual processing, similar data are deleted. Third, we deal with the consistency of the data. The features L and R involved in this paper are in the unit of days. To facilitate the calculation of time, we convert the Timestamp field in the form of year, month, and day. Finally, to reduce the imbalance rate in the dataset, we use the random sampling function embedded in python to perform random oversampling on the dataset before extracting features.

(2) *Feature extracting*

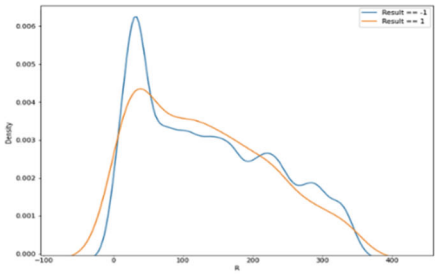
We add two new features based on the RFM analysis: transaction length(L) and customer satisfaction(S). This analysis framework is called RFMLS in this paper. Partial data after feature extracting are illustrated in Table 3.

In order to observe the distribution of features more intuitively, we draw the kernel density estimation curve. The kernel density estimation curve can reflect the distribution of each feature in the dataset. Since the features F and S are discrete variables, the kernel density estimation curves of R, L, and M are presented in this paper (Fig. 1).

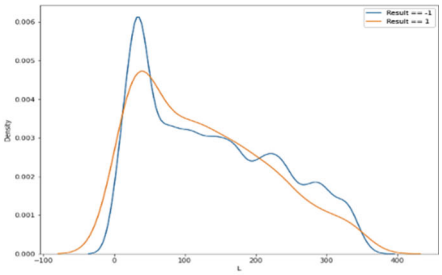
In Fig. 1, the curve corresponding to "result = -1" is the distribution of the churners. The curve corresponding to "result = 1" is the distribution of the non-churners. In Fig. 1, the

Table 3 Partial data after feature extracting

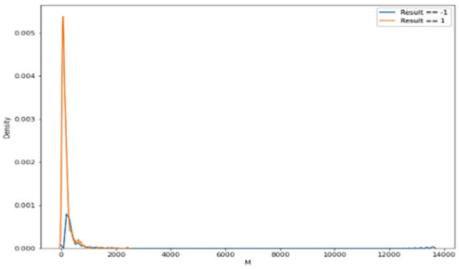
Customer_unique_id	R	F	M	L	S
0000…7064	296	1	86.22	296	3
000b…d6d3	258	2	162.28	272	10
000d…d855	35	2	423.86	294	9
…	…	…	…	…	…
fe8e…fcd2	21	1	84.58	21	4
ff70…6164	327	1	112.46	327	5
ffff…cbeb	243	1	71.56	243	5



(a) Kernel density estimation of feature R



(b) Kernel density estimation of feature L



(c) Kernel density estimation of feature M

Fig. 1 The kernel density estimation curves

Table 4 Partial data after one-hot encoding

Customer_unique_id	R	M	L	F1	…	F22	S1	…	S52
0000…7064	296	86.22	296	1		0	0		0
000b…d6d3	258	162.28	272	1		0	1		0
000d…d855	35	423.86	294	1		0	0		0
…	…	…	…	…	…	…	…	…	…
fe8e…fcd2	21	84.58	21	1		0	1		0
ff70…6164	327	112.46	327	1		0	1		0
ffff…cbeb	243	71.56	243	1		0	0		0

distributions of churners and non-churners on these three features are different. Thus, we can conclude that the five features selected in this paper do differ between the two types of customers. This indicates that the multi-source dataset based on RFMLS can make better predictions.

(3) *One-hot encoding*

Data can be divided into discrete and continuous data. The continuous data can be processed directly by a machine learning algorithm. However, discrete data cannot be processed directly. In our research, features F and S are discrete data and should be transformed into continuous data. We do it by one-hot encoding, whose principle is to use the Euclidean space to expand the discrete data. After encoding, each observation can be normalized to a number between −1 and 1. This helps to reduce the experimental error caused by the different units between data.

In this paper, 78 features are obtained after one-hot encoding, and partial data after one-hot encoding is illustrated in Table 4. From Table 4, we can see that feature F is converted into 22 continuous features, and feature F is converted into 52 features. Therefore, together with M, R, and L, a total of 77 features are obtained. At this point, the features M, R, and L are dimensional data, while the features F and S are non-dimensional data. In order to reduce the large differences within the data, we need to conduct normalization.

(4) *Normalization*

The function of normalization is to eliminate the error caused by different units. The features involved in this paper contain amount and time, which have different units. If normalization is not carried out, it may lead to serious errors and affect the effect of the prediction model. The normalization is implemented as follows:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{23}$$

This method can scale the original data equally. In (23), X_{norm} is the normalized data. X is the original data. The maximum value is described by X_{max} is, and the minimum value is described by X_{min}.

In order to facilitate the subsequent iterative calculation, we normalize the data of e-commerce customers by converting the value of each sample to a number within the range of [0,1]. Partial data after normalization processing is shown in Table 5.

As shown in Table 5, all the data describing features F and S are converted into data in the range of [0, 1]. There are 74 features in total. If all these features are directly input

Table 5 Partial data after normalization processing

Customer_unique_id	R	M	L	F1	…	F22	S1	…	S52
0000…7064	0.82222	0.00558	0.82222	1		0	0		0
000b…d6d3	022,223	0.00246	0.22222	1		0	1		0
000d…d855	0.00556	0.01369	0.13056	1		0	0		0
…	…	…	…	…	…	…	…	…	…
fe8e…fcd2	0.57222	0.15068	0.57222	1		0	1		0
ff70…6164	0.90833	0.00230	0.90833	1		0	1		0
ffff…cbeb	0.675	0.00450	0.675	1		0	0		0

Table 6 Partial data after PCA

Customer_unique_id	0	1	…	17	18	19
0000…7064	0.42962	0.13428		−0.00082	0.00096	−0.00012
000b…d6d3	0.78927	−0.49680		−0.00039	−0.00017	0.00022
000d…d855	−0.51916	0.05106		0.00013	−0.00021	0.00011
…	…	…	…	…	…	…
fe8e…fcd2	0.79325	−0.42743		−0.00028	−0.00030	0.00030
ff70…6164	−0.53805	−0.27813		−0.00025	0.00041	−0.00039
ffff…cbeb	−0.53238	−0.17939		−0.00022	0.00021	−0.00013

into the prediction model, there might be excessive computation time and unsatisfactory prediction accuracy. Therefore, it is necessary to simplify the data. This paper employs the PCA algorithm to extract the principal component before customer churn prediction.

(5) *PCA for dimension reducing*

PCA is a principal component analysis method whose function is to reduce data dimension. The main idea of PCA is to map the original data to the new feature space. The newly produced features are also called the principal components, which refer to the main information contained in the original data features.

Before predicting customer churn, we use the PCA method to extract principal components. Partial data after PCA processing are shown in Table 6. After PCA for dimension reducing, 17 new features are obtained. Therefore, together with M, R, and L, a total of 20 features are obtained.

5 Cross-validation

The experiments are done to test the effectiveness of our proposed model in this paper. In customer churn prediction, there are two kinds of errors. The first kind of error comes from classifying "churner" as "non-churner," while the second kind of error comes from classifying "non-churner" as "churner." The second error increases customer retention costs for an enterprise, while the first error will lead to a significant risk of customer churn.

Table 7 Confusion matrix for binary classification problems

	Non-churned customers (predicted)	Churned customers (predicted)
Non-churned customers (actual)	TP	FN
Churned customers(actual)	FP	TN

Considering the difference between the two types of errors, we should try to minimize the first error in particular.

5.1 Validation index

We adopt overall classification accuracy rates, the G-mean value, the accuracy rate of non-churned customers, and the recall rate as the evaluation indicators. These indicators can be calculated using the confusion matrix, which is shown in Table 7.

As an evaluation indicator for the classification accuracy, the larger the G-mean value, the better the classification effect. The G-mean value is computed as follows:

$$G - mean = \sqrt{TPR \times TNR}. \tag{24}$$

where

$$TPR = \frac{TP}{TP + FN}. \tag{25}$$

$$TNR = \frac{TN}{TN + FP}. \tag{26}$$

In practice, due to the randomness of samples, the finite number of samples, and the prediction algorithm's inability to weight samples, we cannot control the probability of the two types of errors simultaneously. Most existing research uses accuracy rate as the evaluation indicator and ignores customer maintenance cost produced by the second error, which may lead to unsatisfactory performance. The precision and recall of non-churned customers are also introduced to measure the model's classification accuracy for non-churned customers, which are given by:

$$precision = \frac{TP}{TP + FP}. \tag{27}$$

$$recall = \frac{TP}{TP + FN}. \tag{28}$$

5.2 Cross-validation

A ten-fold cross-validation method is used to extract the training set and testing set. We randomly divide the dataset into ten subsets. Nine of them are used as the training set, and one subset is used as the testing set. Ten trials were conducted. Finally, the average values of the evaluation indicators obtained from the ten trials are used as the final values of the model accuracy.

Table 8 The performance of different models

	Overall accuracy	G-mean	Recall	Precision
Logistic regression	0.9788	0.9831	0.9771	0.9793
SVM	0.6751	0.6335	0.8912	0.6197
AdaBoost	0.9737	0.9714	0.9766	0.9669
PCA-AdaBoost	0.9898	0.9897	0.9917	0.9880

5.2.1 Comparison of model accuracy

To test the effectiveness of the PCA-AdaBoost model proposed in this paper, we compare it with Logistic regression, SVM, and AdaBoost algorithms. The same dataset with the same features is used. The overall accuracy rate, G-mean value, recall rate, and precision are calculated through numerical experiments, as shown in Table 8.

From Table 8, we can see the following results. First, the PCA-AdaBoost model performs best in terms of overall accuracy rate and G-mean value. Compared with the AdaBoost model, the PCA-AdaBoost model can significantly improve the performance by mapping the original features into a new feature space. It is advantageous in processing imbalanced datasets. Second, SVM has the lowest accuracy rate compared with the other three models. This result shows that SVM is more suitable for processing small-scale datasets. Third, the recall rate and precision rate of the PCA-AdaBoost model are better than the other three models. This confirms that the model proposed in this paper works well in predicting non-churned customers.

5.2.2 Comparison of model stability

We test the stability of our model in two ways. First, we calculate the maximum, minimum, and standard deviations to evaluate the stability. The results show that the standard deviation of the PCA-AdaBoost model is the smallest. Second, in order to compare the stability more intuitively, we draw the line chart of both the overall accuracy rate and G-mean value. In Figs. 2 and 3, the number of experiments is represented in the horizontal axis, and the value of the evaluation indicator is represented in the vertical axis. Overall, these two figures show

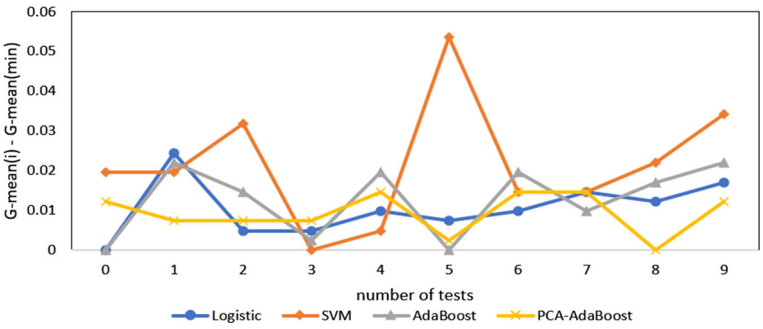


Fig. 2 Line chart of overall accuracy using ten-fold cross-validation

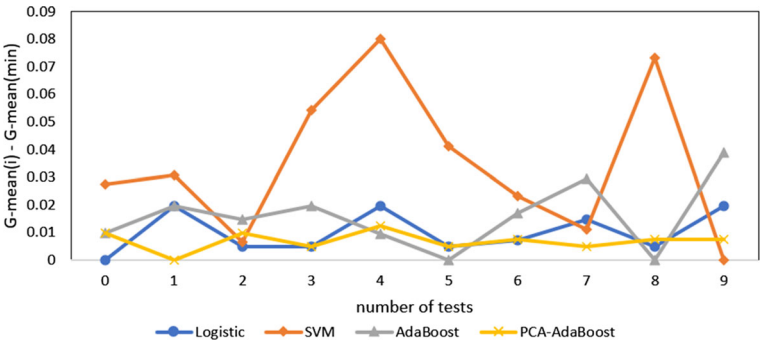


Fig. 3 Line chart of G-mean value using ten-fold cross-validation

that the PCA-AdaBoost model proposed in this paper achieves better stability than the other classification algorithms.

6 Discussions

In this paper, there are two significant theoretical contributions. First, we focus on non-contractual customer churn prediction, where we integrate transaction length and customer satisfaction into the RFM analysis to achieve higher accuracy. Several existing studies in the literature on customer churn have focused on contractual customers (Coussement et al., 2008; Verbeke et al., 2012). In particular, some scholars conducted the RFM analysis to predict contractual customer churn (Coussement et al., 2021; Martínez et al., 2021). However, in this paper, we focus on non-contractual customer churn prediction. Our work demonstrates the generalizability of the RFMLS analysis and other similar improved RFM analyses (Chen et al., 2015). Second, our numerical results show that the proposed PCA-AdaBoost model can perform better than Logistic Regression, SVM, and AdaBoost algorithms. Several existing studies mainly deploy a single base classifier to predict customer churn (Caigny et al., 2018; Coussement & Poel, 2008). And several scholars employ ensemble learning algorithms to improve the predicting performance for unbalanced data (Dhote et al., 2020; Jayaswal et al., 2016). However, ensemble learning algorithms may still be inefficient for high-dimension and unbalanced data. In this paper, our proposed algorithm improves the accuracy rate of customer churn prediction by integrating a pre-processing technique and an ensemble learning algorithm. In this process, we employ PCA to overcome the high-dimensional challenge of data.

Furthermore, we use the random sampling function to perform random oversampling on the dataset to reduce the imbalance rate. As confirmed in Table 8, the proposed PCA-AdaBoost model performs the best overall accuracy rate and G-mean value. In short, the PCA-AdaBoost model can help managers improve customer retention rates.

7 Conclusions, limitations, and future research

With the continuous development of the Internet and mobile technologies, the e-commerce industry has expanded rapidly. Since most e-commerce customers are non-contractual, customer churn often occurs. The features from a single data source are often selected to predict customer churn in the existing literature. Furthermore, single base classifiers and AdaBoost algorithms have been employed to predict customer churn. However, the existing algorithms have difficulties in processing high-dimension and unbalanced data. We developed an improved PCA-AdaBoost model to solve these problems for predicting customer churn. We conduct numerical experiments based on 50,513 customer samples from an e-commerce platform. Our results show that our proposed PCA-AdaBoost model performs better than the other algorithms developed in the literature.

Three conclusions can be drawn. First, we integrate two more features, transaction length and customer satisfaction, into the RFM analysis. As a result, this RFMLS analysis is shown to be effective in predicting non-contractual customer churn in the e-commerce context. Second, compared with logistic regression, SVM, and AdaBoost algorithm, the PCA-AdaBoost model has higher classification accuracy according to cross-validation results. This confirms that our model is suitable to process high-dimension and unbalanced data. Third, compared with the other algorithms, our proposed model is more stable according to ten-fold cross-validation results.

However, there are still several potential shortcomings in the RFMLS analysis and PCA-AdaBoost model developed in this paper. First, the RFMLS is not the only effective analytical framework for predicting e-commerce customer churn. In the future, more customer behavior features can be incorporated into customer churn prediction. Second, the computation time of our model is not the shortest. Our model integrates 50 decision-tree (DT) classifiers. Our model's training and testing time are longer than the single DT algorithm. Third, we use DT as the base classifier of ensemble learning. We can use other classifiers such as BP neural network in the future. Fourth, we employ the PCA algorithm to reduce data dimension. In future research, we can compare the performance of different algorithms, such as linear discriminant analysis and locally linear embedding, for dimension reduction.

Funding Zengyuan Wu was supported by the Natural Science Foundation of Zhejiang Province [Grant No. LY20G010008], and Bei Wu was supported by China's National Natural Science Foundation [Grant No. 71472169].

References

Azeem, M., Usman, M., & Fong, A. C. M. (2017). A churn prediction model for prepaid customers in telecom using fuzzy classifiers. *Telecommunication Systems, 66*(4), 603–614.

Caigny, A. D., Coussement, K., & Koen, W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research, 269*(2), 760–772.

Chen, K., Hu, Y. H., & Hsien, Y. C. (2015). Predicting customer churn from valuable B2B customers in the logistics industry: A case study. *Information Systems and E-Business Management, 13*(3), 475–494.

Chen, T., Tu, S. X., Wang, H. L., Liu, X. S., Li, F. H., Jin, W., Liang, X. W., Zhang, X. Q., & Wang, J. (2020). Computer-aided diagnosis of gallbladder polyps based on high resolution ultrasonography. *Computer Methods and Programs in Biomedicine, 185*, 105118.

Coussement, K., De Bock, K. W., & Geuens, S. (2021). A decision-analytic framework for interpretable recommendation systems with multiple input data sources: A case study for a European e-tailer. *Annals of Operations Research*. https://doi.org/10.1007/s10479-021-03979-4

Coussement, K., & Poel, V. D. V. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications, 34*(1), 313–327.

Demirer, R., Pierdzioch, C., & Zhang, H. (2017). On the short-term predictability of stock returns: A quantile boosting approach. *Finance Research Letters, 22*(3), 35–41.

Dhote, S., Vichoray, C., Pais, R., Baskar, S., & Shakeel, P. M. (2020). Hybrid geometric sampling and AdaBoost based deep learning approach for data imbalance in E-commerce. *Electronic Commerce Research, 20*(2), 259–274.

Feng, C., Cui, M., Hodge, B. M., & Zhang, J. (2017). A data-driven multi-model methodology with deep feature selection for short-term wind forecasting. *Applied Energy, 190*, 1245–1257.

Freund, Y., & Schapire, R. E. (1999). A decision-theoretic generalization of online learning and an application to Boosting. *Journal of Computer and System Sciences, 55*(1), 119–139.

Guo, H., Li, Y., Jennifer, S., Gu, M., Huang, Y., & Gong, B. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications., 73*, 220–239.

Hossain, N. U. I., Nur, F., Jaradat, R., Hosseini, S., Marufuzzaman, M., Puryear, S. M., & Buchanan, R. K. (2019). Metrics for assessing overall performance of inland waterway ports: A Bayesian network based approach. *Complexity*. https://doi.org/10.1155/2019/3518705

Hughes, A. M. (2005). *Strategic database marketing*. McGraw-Hill Pub. Co.

Jayaswal, P., Tomar, D., Agarwal, S., & Prasad, B. R. (2016). An ensemble approach for efficient churn prediction in telecom industry. *International Journal of Database Theory and Application, 9*(8), 211–232.

Ji-fan, R. S., Fosso, W. S., Akter, S., Dubey, R., & Childe, S. J. (2017). Modelling quality dynamics, business value and firm performance in a big data analytics environment. *International Journal of Production Research, 55*(17), 5011–5026.

Kisioglu, P., & Topcu, Y. I. (2011). Applying Bayesian belief net-work approach to customer churn analysis: A case study in the telecom industry of Turkey. *Expert Systems with Applications, 38*(6), 7151–7157.

Koutanaei, F. N., Sajedi, H., & Khanbabaei, M. (2015). A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring. *Journal of Retailing and Consumer Services, 27*, 11–23.

Li, H., Qu, Y., Guo, S. K., Gao, G. F., Chen, R., & Chen, G. (2020). Surprise bug report prediction utilizing optimized integration with imbalanced learning strategy. *Complexity*. https://doi.org/10.1155/2020/8509821

Mahajan, P. D., Maurya, A., Megahed, A., Elwany, A., Strong, R., & Blomberg, J. (2020). Optimizing predictive precision in imbalanced datasets for actionable revenue change prediction. *European Journal of Operational Research, 285*(3), 1095–1113.

Martín, C. A., Torres, J. M., Aguilar, R. M., & Diaz, S. (2018). Using deep learning to predict sentiments: Case study in tourism. *Complexity*. https://doi.org/10.1155/2018/7408431

Martínez, R. G., Carrasco, R. A., Sanchez-Figueroa, C., & Gavilan, D. (2021). An RFM model customizable to product catalogues and marketing criteria using fuzzy linguistic models: Case study of a retail business. *Mathematics, 9*(16), 1836.

Miguis, V. L., Poel, V. D. V., Camanho, A. S., & Cunha, J. F. E. (2012). Modeling partial customer churn: On the value of first produce-category purchase sequences. *Expert Systems with Applications, 39*(12), 11250–11256.

Montiel, M. A. D., & Lopez, F. (2020). Spatial models for online retail churn: Evidence from an online grocery delivery service in Madrid. *Papers in Regional Science, 99*(6), 1643–1665.

Muhammad, A., Muhammad, U., & Fong, A. C. M. (2017). A churn prediction model for prepaid customers in telecom using fuzzy classifiers. *Telecommunication Systems, 66*(4), 603–614.

Qi, J., Zhang, L., Liu, Y., Li, L., & Li, H. (2008). ADTreesLogit model for customer churn prediction. *Annals of Operations Research, 168*(1), 247–265.

Qu, Z., Liu, H., Wang, Z., Xu, J., Zhang, P., & Zeng, H. (2021). A combined genetic optimization with AdaBoost ensemble model for anomaly detection in buildings electricity consumption. *Energy and Buildings*. https://doi.org/10.1016/j.enbuild.2021.111193

Renjith, S. (2015). An integrated framework to recommend personalized retention actions to control B2C E-commerce customer churn. *International Journal of Engineering Trends and Technology, 27*(3), 152–157.

Talayeh, R., Ilya, S., Joseph, E., Ehsan, S., & John, D. S. (2019). Predictive models for bariatric surgery risks with imbalanced medical datasets. *Annals of Operations Research, 280*, 1–18. https://doi.org/10.1007/s10479-019-03156-8

Tsai, C. F., & Lu, Y. H. (2009). Customer churn prediction by hybrid neural networks. *Expert Systems with Applications, 36*(10), 12547–12553.

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*(1), 211–229.

Vo, N. N. Y., Liu, S., Li, X., & Xu, G. (2021). Leveraging unstructured call log data for customer churn prediction. *Knowledge-Based Systems, 212*(4), 106586.

Wang, G., Gunasekaran, A., Ngai, E. W., & Papadopoulos, T. (2016). Big data analytics in logistics and supply chain management: Certain investigations for research and applications. *International Journal of Production Economics, 176*, 98–110.

Wübben, M., & Wangenheim, F. (2008). Instant customer base analysis: Managerial heuristics often "get it right." *Journal of Marketing., 72*(3), 82–93.

Xiao, Y. H., Cao, D. L., & Gao, L. P. (2020). Face detection based on occlusion area detection and recovery. *Multimedia Tools and Applications, 79*(2), 16531–16546.

Yu, R., An, X., Jin, B., Shi, J., Move, O. A., & Liu, Y. (2018). Particle classification optimization-based BP network for telecommunication customer churn prediction. *Neural Computing and Application, 29*(2), 707–720.

Zacharis, N. Z. (2018). Classification and regression trees (cart) for predictive modeling in blended learning. *International Journal of Intelligent Systems and Applications, 10*(3), 1–9.

Zeineb, A., & Rania, H. K. (2019). Forecast bankruptcy using a blend of clustering and MARS model: Case of US banks. *Annals of Operations Research, 281*(1–2), 27–64.

Zhou, J., Wei, J., & Xu, B. (2021). Customer segmentation by web content mining. *Journal of Retailing and Consumer Services, 61*(11), 102588.