IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 24, No. 5, October, 2019, pp. 525-530
Journal homepage: http://iieta.org/journals/isi

A Hybrid Prediction Model for E-Commerce Customer Churn Based on Logistic Regression and Extreme Gradient Boosting Algorithm

Xueling Li*, Zhen Li

Computer and Software College, Jincheng College of Sichuan University, Chengdu 611731, China

Corresponding Author Email: lixueling@scujcc.edu.cn

https://doi.org/10.18280/isi.240510

Received: 17 April 2019
Accepted: 23 July 2019

Keywords:
customer churn, logistic regression, E-commerce, extreme gradient boosting (XGBoost) algorithm, empirical analysis

ABSTRACT

Customer churn is an important problem in the field of e-commerce. Based on the real data of an e-commerce platform, this paper establishes a hybrid prediction model for customer churn based on logistic regress and extreme gradient boosting (XGBoost) algorithm. More than 20 key indices were selected through data mining of the real data, covering such dimensions as order information, customer profile, and aftersales situation. With these indices, the hybrid model was applied to predict the churn state of each customer in the sample data. The results show that our model achieved a greater-than-85% accuracy in the forecast of customer churn. The research findings provide an important guide for e-commerce enterprises to improve customer adhesiveness.

1. INTRODUCTION

Early warning of customer churn has long been a research hotspot in the field of e-commerce. The customers that might be lost should be identified accurately through data mining and data analysis, and retained in time with effective marketing measures. Even if the customer loss is inevitable, a prediction model of customer churn can still help e-commerce enterprises to identify the causes of the loss and reduce similar losses in future.

The traditional prediction models for e-commerce customer churn are simple and rough, due to budget constraint. The customers who have not logged in or purchased any goods are considered lost. Besides, only the data on the previous orders are analyzed by these models, because of the limited capacity of technologies for data collection and storage. Other types of data on customers, e.g. the profile, browsing path, favorites and comments, are rarely taken into account in the prediction process.

The most popular customer churn prediction model is the recency, frequency, and monetary value (FRM) model [1]. The model divides the customers to different categories based on three indices: recency (when did the customer make the last purchase?), frequency (how often the customer makes purchases?) and monetary value (how much the customer spends on each purchase?), and provides e-commerce enterprises corresponding operation measures to retain the target customers and maximize their profits. However, the order data has a certain time delay than the other types of data on customers, especially the data on browsing behavior. Besides, different customers have varied repurchase cycles. Thus, the analysis of the RFM model is too single and too simple.

In this paper, more types of data are introduced to predict the customer churn in e-commerce, making the prediction more comprehensive and accurate. Specifically, a hybrid prediction model for customer churn was established based on logistics regression and extreme gradient boosting (XGBoost) algorithm. Then, multiple indices were selected from various dimensions as the basis for prediction. The hybrid model was verified through a case study on an actual e-commerce platform.

2. LITERATURE REVIEW

The advent of big data and artificial intelligence (AI) has greatly promoted the prediction of customer churn. The customer churn problem is widely seen as a binary classification problem [2] and solved by supervised machine learning algorithms, such as clustering analysis [3], decision tree [4], association analysis [5], logistics regression [6], support vector machine (SVM) [7] and artificial neural network (ANN) [8]. Below are some representative studies on the prediction of customer churn.

Huang and Wang [9] improved the Iterative Dichotomiser 3 (ID3) algorithm with weighted entropy, constructed a decision tree model based on the improved algorithm, and proved the effectiveness and accuracy of the improved algorithm through empirical analysis. Ahmed et al. [10] relied on the genetic algorithm (GA) to identify eight indices of customer churn from basic and transaction data, combined the SVM and neural network (NN) into a hybrid prediction model, and confirmed that the hybrid model is more accurate and efficient than the SVM and NN along.

Ju et al. [11] analyzed the basic information and behaviors (e.g. shopping, check-in and sharing) of online customers, built a comprehensive prediction model for customer churn based on the NN, decision tree and C5.0 algorithm, and verified the extraordinary accuracy of the established model. Targeting the data of Internet financial platform, Chan and Misra [12] introduced social network factors into machine learning, and constructed a XGBoost-based prediction model for customer churn, with individual information and social

network activity as variables.

Carver et al. [13] empirically demonstrated that XGBoost-based prediction model outperforms logistic regression, the SVM and random forest model, and the prediction performance can be further improved after incorporating social network variables. Focusing on customer behavior, Fathian et al. [14] set up a decision tree model with three feature attributes, i.e. recency, frequency and monetary value in the past half a year, and successfully predicted 90% of the customer churn with the model.

Using the basic and behavior attributes of customers, Sharma and Panigrahi [15] established a model coupling classification and regression tree (CART) and adaptive boosting algorithm, and manifested the high accuracy of the model in customer churn forecast through simulation experiments.

Keramati et al. [16] analyzed the current and historical changes of customer consumption in telecommunications, created a prediction model for customer churn based on logistic regression, and found that the model can achieve the accuracy of 93.17%. Su et al. [17] derived over 300 new variables from 49 basic variables to reflect the dynamic changes of customers, and screened the variables by logistic regression, laying the basis for customer churn prediction.

3. PRELIMINARIES

Data mining is a technique to process and analyze a large number of data, and provide decision-makers with important and meaningful information. In this paper, data mining is adopted to screen the features on e-commerce customers, aiming to facilitate the prediction of customer churn.

3.1 Data mining methods

Depending on the learning tasks, data mining methods (Figure 1) generally falls into two categories: supervised learning methods and unsupervised learning methods.

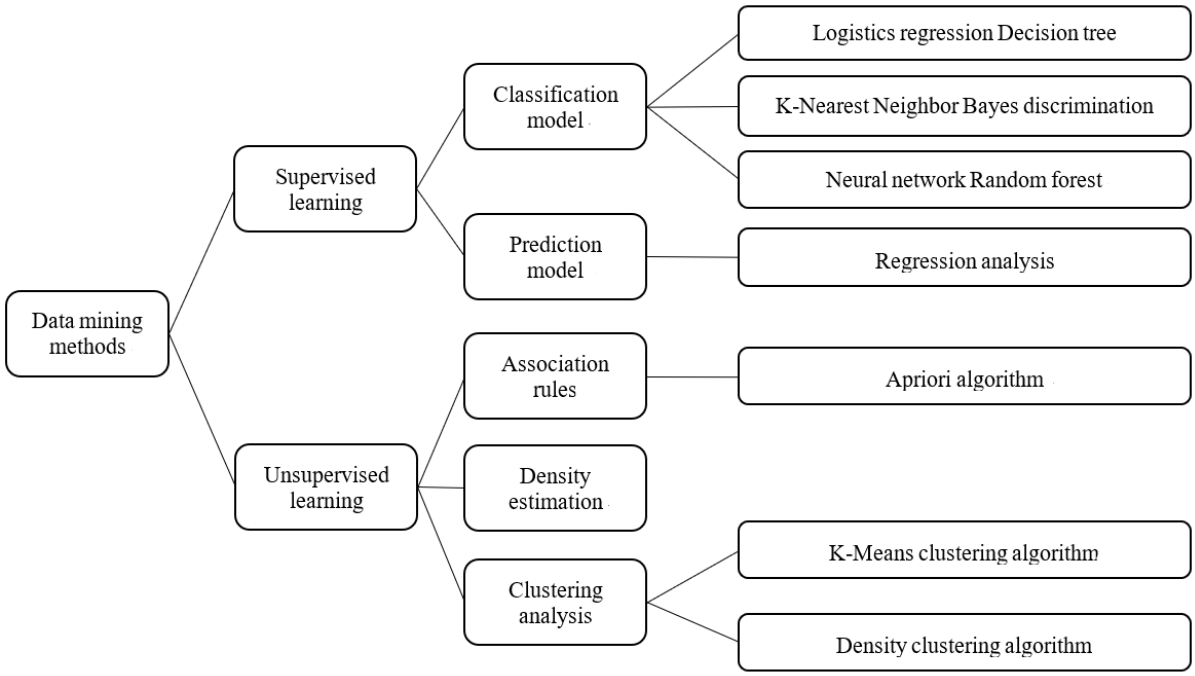

Figure 1. The common data mining methods

Supervised learning infers a function from labeled training data consisting of a set of training examples, which can correctly determine the class labels for unseen instances. Supervised learning methods can be further divided into classification and prediction methods. The most frequently used supervised learning methods include logistics regression, decision tree, k-nearest neighbor (kNN), Bayes discrimination, etc.

Unsupervised learning is a self-organized learning strategy that looks for previously unknown patterns in dataset without pre-existing labels. It helps to solve the internal relationship and similarity between features in unlabeled data. The common methods of unsupervised learning are association rules, density estimation and clustering analysis.

3.2 Logistics regression

Logistic regression is extended from linear regression. However, the target variable of logistic regression is a discrete variable, rather than a numerical value in linear regression. Logistic regression is a desirable way to deal with binary classification problems. During the regression process, each feature is multiplied by a regression coefficient, then a sigmoid function is introduced, and finally a value is outputted in the interval [0, 1] through linear regression. If the value is greater than 0.5, the label is classified as class 1; otherwise, the label is classified as class 0.

The keys steps of logistic regression include setting up the prediction function $g_{\vartheta}(x)$ to project the judgement based on the input data, constructing the loss function about the deviation of predicted output from the class of training data, and obtaining the regression function with minimal loss.

The sigmoid function plays a significant role in logistic regression. Suppose there exists a binary classification problem, with $y \in \{0,1\}$ as the output of the target variable and $z = \vartheta^T x + \vartheta_0$ as the predicted output of linear regression

(the real value). Then, a step function *f*(z) is needed to convert z into 0/1:

$$f(\mathrm{z}) = \begin{cases} 0 \;\; if\; z < 0 \\ 0.5\; if\; z = 0 \\ 1 \;\; if\; z > 0 \end{cases} \quad (1)$$

However, *f*(z) is discontinuous rather than differentiable and monotonic. The sigmoid function provides a tool with good differentiability and monotonicity:

$$y = \frac{1}{1+e^{-z}} \quad (2)$$

If $z>0$ and $y>0.5$, y is always positively correlated with z; if $z<0$ and $y<0.5$, y is always negatively correlated with z.

According to the sigmoid function, the prediction function can be constructed as:

$$g_\vartheta(x) = \frac{1}{1+e^{-\vartheta^T x}} \quad (3)$$

The probabilities for an input x to be allocated to class 1 or class 0 can be respectively described as:

$$\mathrm{P}(y_i = 1|x;\vartheta) = g_\vartheta(x) \quad (4)$$

$$\mathrm{P}(y_i = 0|x;\vartheta) = 1 - g_\vartheta(x) \quad (5)$$

Considering the dichotomy of the dependent variables in logistic regression model, the loss function can be solved by the maximum likelihood estimation.

According to Eqns. (4) and (5), the probability function can be established as:

$$\mathrm{P}(\mathrm{y}|\mathrm{x};\vartheta) = g_\vartheta(x)^y(1 - g_\vartheta(x))^{1-y} \quad (6)$$

Assuming that the samples are independent of each other, the likelihood function can be obtained as:

$$L(\vartheta) = \prod_{i=1}^{s} P(y_i|x_i;\vartheta) = \prod_{i=1}^{s} g_\vartheta(x_i)^{y_i}(1 - g_\vartheta(x_i))^{1-y_i} \quad (7)$$

Then, the log likelihood function can be obtained as:

$$l(\vartheta) = \log\big(L(\vartheta)\big) = \sum_{i=1}^{s}(y_i log g_\vartheta(x_i) + (1 - y_i)\log\,(1 - g_\vartheta(x_i))) \quad (8)$$

Maximum likelihood estimation aims to maximize $l(\vartheta)$. The estimation can be implemented by gradient ascend method, and the ϑ value thus obtained must be the best parameter.

The logistics regression modelling includes the following steps:

Step 1. Initialization: Determine the dependent and independent variables according to the objective, and initialize the regression equation.

Step 2. Coefficient estimation: Estimate the regression coefficient for the model.

Step 3. Equation checking: Test the significance of the regression equation by the F-value and P-value in the analysis of variance (ANOVA) table. If p-value is below the significance level, the equation passes the test; otherwise, the regression equation needs to be rebuilt based on new variables. Note that even if the regression equation passes the significance test, it does not mean that each independent variable has a significant influence on the dependent variable.

Step 4. Variable checking: Test the significance of each independent variable, remove the unimportant and insignificant variables, and rebuild the equation until the whole model and regression variable pass the significance test.

3.3 XGBoost algorithm

The XGBoost is a distributed gradient boosting algorithm based on the CART decision tree. The basic principle is to build a powerful high-precision classifier by improving the iterative performance of weak classification algorithm.

In the CART algorithm, each leaf node is allocated an input attribute, and then output a score. The output scores of all leaves are added up, forming the predicted result for each sample. Suppose K trees are available for prediction. Then, the prediction function can be established as:

$$y_i = \sum_{k=1}^{K} l_k(x_i)\,, l_k \in F, i \in [1, n] \quad (9)$$

where, n is the number of samples; F is the set of all regression trees; l_k is a function of F.

Then, the objective function can be written as:

$$\mathrm{Obj}(\vartheta) = \mathrm{L}(\vartheta) + \delta(\vartheta) \quad (10)$$

where, $L(\vartheta)$ is the loss function to fit training data and evaluate the deviation between the model prediction and sample data in the training set; $\delta(\vartheta)$ is the regularization term to measure the complexity of the model.

The objective function of XGBoost can be assumed as:

$$\mathrm{Obj(t)} = \sum_{i=1}^{N} L(y_i) + \sum_{i=1}^{M} \delta(y_i) \quad (11)$$

This function represents the maximum reduction on the target for a specific tree structure. The function value is called a structure score. The smaller the structure score, the better the tree structure.

4. HYBRID PREDICTION OF CUSTOMER CHURN

The previous sections have introduced the basic principles of logistics regression and XGBoost algorithm. In this section, the two techniques are combined into a hybrid prediction model of customer churn for e-commerce platforms.

The features of lost customers were mined out from the real data of an e-commerce platform. In total, there are nearly 300,000 customer samples, of which 165,825 (52%) belong to lost customers. The number of lost customers is roughly the same as that of return customers.

The original data samples were cleaned first to remove outliers and missing values, leaving 293,272 valid samples. Among them, 162,051 (55.2%) samples belong to lost customers.

The order behavior of the customers on the platform was analyzed, revealing that nearly 80% of customers, who have placed an order in the first three months, will make a repurchase in the fourth or fifth month. In other words, the platform may have lost a customer, if he/she has placed an order in the first three month, but does not place an order again in the following two months.

Therefore, the first quarter was taken as the observation

period, and the following two months as the verification period. Any customer that placed an order in Q1 was marked as a lost customer, if he/she did not place an order again in April and May, and as return customer, if otherwise.

4.1 Index selection

Through data analysis, 25 indices (Table 1) were empirically selected to predict the customer churn. These indices are correlated with each other in business logic, covering dimensions like order information, customer profile, preference, aftersales situation, adhesiveness and churn state.

In the dimension of order information, six indices were selected, namely, order days, order quantity, spending, product diversity, product quantity, and brand diversity. The order days refer to the number of days that a customer places an order in the observation period. The order quantity equals the total number of orders placed and paid by a customer in the observation period. The spending reflects the total amount paid by a customer in the observation period. The product diversity describes the number of categories of the purchased products in the observation period. The product quantity measures the total number of purchased products in the observation period. The brand diversity shows the number of brands of the purchased products in the observation period.

In the dimension of customer profile, there are a total of nine indices: customer level, gender, age, marital status, education level, registration recency, first order recency, purchase power and promotion sensitivity. The customer level refers to the value of a customer to the platform, which is identified based on his/her consumption, credit and other behaviors on the platform. Registration recency and first order recency are defined as the number of months since the registration and the first order placed by a customer, respectively. The purchase power was determined through clustering analysis, which compares the price range of the products purchased by a customer and the price range of the categories for the purchased products. The promotion sensitivity was obtained by clustering of all the order information of a customer.

In the dimension of preference, six indices were selected, including favorite stores, favorite products, good comments, bad comments, total comments and posting lists. The favorite stores and favorite products refer to the number of stores and products favored by a customer in the observation period, respectively. Good comments and bad comments stand for the number of favorable and unfavorable comments left by a customer in the observation period, respectively. The total comments and posting lists mean the total number of comments and posting lists of a customer in the observation period, respectively.

In the dimension of aftersales situation, two indices were selected: complaints and aftersales orders. The former refers to the number of complaints filed by a customer, and the latter, the number of orders a customer complained about. If the former is greater than the latter, it means the complaint about an order is not solved in time, leading to repeated complaints.

In the dimension of adhesiveness, login days and sign-in days were selected as the prediction indices. The former means the number of days a customer logs onto the platform and the latter, the number of days a customer signs in on the platform.

In the dimension of churn state, an index of the same name was selected. This index is discrete and has two values: 0 means the customer is not lost and 1 means he/she is lost.

Table 1. Name and description of indices

Dimensions	Name	Description
Order information	Oder days	The number of days that a customer places an order
	Order quantity	The total number of orders placed and paid by a customer
	Spending	The total amount paid by a customer
	Product diversity	The number of categories of the purchased products
	Product quantity	The total number of purchased products
	Brand diversity	The number of brands of the purchased products
Customer profile	Customer level	User level: 1~4; level 1 is the lowest, and level 4 is the highest.
	Gender	Gender: 1 or 2; 1 is male, 2 is female.
	Age	Age: 1~5; level 1 is the youngest, and level 5 is the oldest.
	Marital status	Marital status: 0 or 1; 0 is unmarried and 1 is married.
	Education level	Education level: 1~4; level 1 is the lowest, and level 4 is the highest.
	Registration recency	The number of months since registration
	First order recency	The number of months since the first order placed by a customer
	Purchase power	Purchase power: 1~4; level 1 is the weakest, and level 4 is the strongest.
	Promotion sensitivity	Promotion sensitivity: 1~4; level 1 is the least sensitive, and level 4 is the most sensitive.
Preference	Favorite stores	The number of stores favored by a customer
	Favorite products	The number of products favored by a customer
	Good comments	The number of favorable comments left by a customer
	Bad comments	The number of unfavorable comments left by a customer
	Total comments	The number of all comments left by a customer
	Posting lists	The number of posting lists of a customer
Aftersales situation	Complaints	The number of complaints filed by a customer
	Aftersales orders	The number of orders complained by a customer
Adhesiveness	Login days	The number of days a customer logs onto the platform
	Sign-in days	The number of days a customer signs in on the platform
Churn state	Churn state	Churn state: 0 or 1; 0 is no loss, and 1 is loss

In logistic regression, the modeling variables should not be multicollinear. If two variables are closely correlated, i.e. $0.85 \leq |r| \leq 1$, one of them must be removed. Judging by the Spearman's rank correlation coefficient, order days and order quantity, both in the dimension of order information, are closely correlated, with a Spearman's rank correlation

coefficient of 0.938. Therefore, only the order days was retained. Similarly, only one of the following pairs of indices was kept for further analysis: registration recency and first order recency; favorite stores and favorite products; complaints and aftersales orders; login days and sign-in days. Finally, twenty nonredundant indices were obtained for our prediction model

4.2 Model construction

The objective of our prediction task is to evaluate whether a customer will be lost. Hence, the customer churn problem is a binary classification problem, and should be solved by supervised learning methods. Here, our prediction model is constructed based on logistic regression and XGBoost algorithm. The XGBoost was introduced to enhance the prediction accuracy of the logistic regression technique.

Table 2. Confusion matrix of logistic regression

	Predicted non-churn (0)	**Predicted churn (1)**
Actual non-churn (0)	12,887	6,571
Actual churn (1)	4,021	20,512

Firstly, the valid samples were split into three parts by the ratio of 0.70: 0.15: 0.15, which in turn serve as the training set, the verification set and the test set. Then, the remaining twenty indices were tested repeatedly to remove those with little impact on the prediction, i.e. the indices with P-value greater than 0.05. After that, the customer churn prediction model was preliminarily set up based on the remaining indices. Next, the preliminary model based on logistic regression was evaluated, using the confusion matrix of the verification set (Table 2).

On this basis, the accuracy, precision and recall were calculated to measure the effect of the preliminary model. Accuracy is the ratio of the number of customers whose churn state is correctly predicted to the total number of customers. Precision is the ratio of the number of actual lost customers to the number of customers predicted to be lost. Recall is the ratio of the number of customers correctly predicted to be lost to the number of actual lost customers. The three evaluation metrics of the preliminary model are computed as:

$$Accuracy = \frac{12887 + 20512}{12887 + 6571 + 4021 + 20512} = 75.9\%$$

$$Precision = \frac{20512}{6571 + 20512} = 75.7\%$$

$$Recall = \frac{20512}{4021 + 20512} = 83.6\%$$

Thus, the preliminary model achieved an accuracy of 75.9%, i.e. the model correctly predicted the churn state of 75.9% customers; the precision of the preliminary model was 75.7%, i.e. 76 out of every 100 customers predicted to be lost are indeed lost; the recall of the preliminary model was 83.6%, indicating that 83.6% of the actually lost users are correctly identified by the preliminary model.

Next, the XGBoost algorithm was introduced to process the valid samples again, producing the importance of each index. It can be seen that order days, first order recency, promotion sensitivity, purchase power and login days are the key indices to predict the churn state of each customer. The most important 10 indices are presented in Figure 2 below.

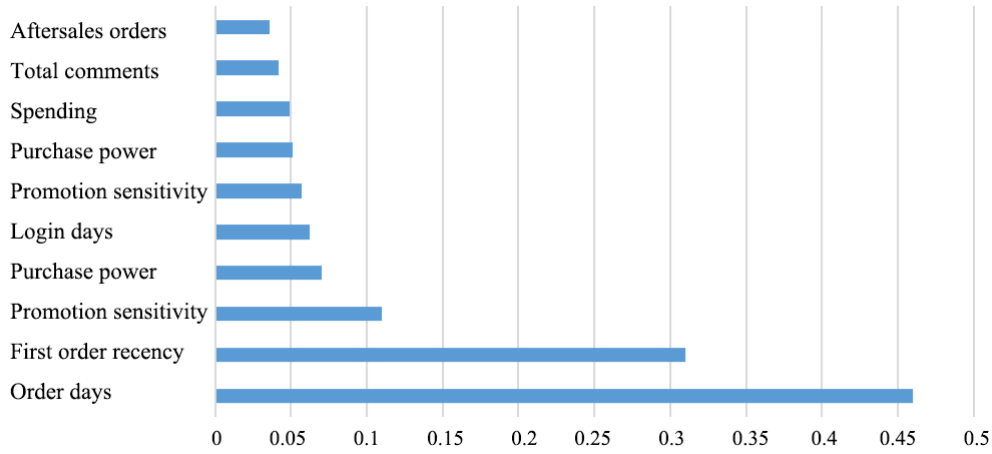

Figure 2. The ten most important indices

Table 3. Confusion matrix of XGBoost

	Predicted non-churn (0)	**Predicted churn (1)**
Actual non-churn (0)	13,123	6,402
Actual churn (1)	3,874	20,592

Table 3 provides the confusion matrix of the XGBoost algorithm.

Through the confusion matrix, the accuracy, precision and recall were computed as:

$$Accuracy = \frac{13123 + 20592}{13123 + 6402 + 3874 + 20592} = 76.6\%$$

$$Precision = \frac{20592}{6402 + 20592} = 76.3\%$$

$$Recall = \frac{20592}{3874 + 20592} = 84.2\%$$

Obviously, the XGBoost algorithm improved the prediction accuracy from the level of the preliminary model.

5. CONCLUSIONS

This paper proposes a hybrid prediction model for customer churn based on logistics regression and XGBoost algorithm model. More than twenty indices were selected from the dimensions like order information and customer profile as the independent variables for the hybrid model. The model was applied to predict the churn state of customers of an actual e-commerce platform. Judging by accuracy, precision and recall, it can be seen that the hybrid model can predict customer churn more accurately than logistic regression.

REFERENCES

[1] Chen, D., Sain, S.L., Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. Journal of Database Marketing & Customer Strategy Management, 19(3): 197-208. https://doi.org/10.1057/dbm.2012.17

[2] Chang, H., Jamin, S., Willinger, W. (2001). Inferring AS-level Internet topology from router-level path traces. International Society for Optics and Photonics, Scalability and traffic control in IP networks, 4526: 196-207. https://doi.org/10.1117/12.434395

[3] Lee, N., Kim, J.M. (2010). Conversion of categorical variables into numerical variables via Bayesian network classifiers for binary classifications. Computational Statistics & Data Analysis, 54(5): 1247-1265. https://doi.org/10.1016/j.csda.2009.11.003

[4] Shaker, A., Senge, R., Hüllermeier, E. (2013). Evolving fuzzy pattern trees for binary classification on data streams. Information Sciences, 220: 34-45. https://doi.org/10.1016/j.ins.2012.02.034

[5] Unler, A., Murat, A. (2010). A discrete particle swarm optimization method for feature selection in binary classification problems. European Journal of Operational Research, 206(3): 528-539. https://doi.org/10.1016/j.ejor.2010.02.032

[6] Ebenuwa, S.H., Sharif, M.S., Alazab, M., Al-Nemrat, A. (2019). Variance ranking attributes selection techniques for binary classification problem in imbalance data. IEEE Access, 7, 24649-24666. https://doi.org/10.1109/ACCESS.2019.2899578

[7] Shao, Y.H., Chen, W.J., Deng, N.Y. (2014). Nonparallel hyperplane support vector machine for binary classification problems. Information Sciences, 263: 22-35. https://doi.org/10.1016/j.ins.2013.11.003

[8] Xu, Z., Watada, J., Wu, M., Ibrahim, Z., Khalid, M. (2014). Solving the imbalanced data classification problem with the particle swarm optimization based support vector machine. IEEJ Transactions on Electronics, Information and Systems, 134(6): 788-795. https://doi.org/10.1541/ieejeiss.134.788

[9] Huang, Y., Wang, Y. (2012). Decision tree classification based on naive Bayesian and ID3 algorithm. Computer Engineering, 38(14): 41-43.

[10] Ahmed, M., Afzal, H., Majeed, A., Khan, B. (2017). A survey of evolution in predictive models and impacting factors in customer churn. Advances in Data Science and Adaptive Analysis, 9(3): 1750007. https://doi.org/10.1142/S2424922X17500073

[11] Ju, C.H., Lu, Q.B., Guo, F.P. (2013). E-commerce customer churn prediction model combined with individual activity. Systems Engineering-Theory & Practice, 33(1): 141-150.

[12] Chan, K.K., Misra, S. (1990). Characteristics of the opinion leader: A new dimension. Journal of Advertising, 19(3): 53-60. https://doi.org/10.1080/00913367.1990.10673192

[13] Carver, T., Harris, S. R., Berriman, M., Parkhill, J., McQuillan, J.A. (2011). Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics, 28(4): 464-469. https://doi.org/10.1093/bioinformatics/btr703

[14] Fathian, M., Hoseinpoor, Y., Minaei-Bidgoli, B. (2016). Offering a hybrid approach of data mining to predict the customer churn based on bagging and boosting methods. Kybernetes, 45(5): 732-743. https://doi.org/10.1108/K-07-2015-0172

[15] Sharma, A., Panigrahi, P.K. (2013) A neural network based approach for predicting customer churn in cellular network services. International Journal of Computer Applications, 27(11): 26-31. https://doi.org/10.5120/3344-4605

[16] Keramati, A., Azadeh, A., Mohammadi, M., Rostami, H. (2011). Identification of customer churn determinants using censored log file data in the Iranian mobile telecommunications service industry. International Journal of Electronic Customer Relationship Management, 5(2): 111-129. https://doi.org/10.1504/IJECRM.2011.041261

[17] Su, Q., Shao, P., Ye, Q. (2012). The analysis on the determinants of mobile VIP customer churn: A logistic regression approach. International Journal of Services Technology and Management, 18(1-2): 61-74. https://doi.org/10.1504/IJSTM.2012.049016