International Core Journal of Engineering Volume 5 Issue 9, 2019
ISSN: 2414-1895 DOI: 10.6919/ICJE.201908_5(9).0008

Research on E-commerce Coupon User Behavior Prediction Technology Based on Decision Tree Algorithm

Huawei Mei [a], Xinyao Li [b]

Department of Computer Science, North China Electric Power University, Baoding 071000, China.

[a]3094998@qq.com, [b]ncepuxy@163.com

Abstract

With the continuous development of mobile devices, China's O2O (Online to Offline) e-commerce field is developing rapidly, more and more enterprises are joining it, and the competition of major e-commerce platforms is fierce. Customer churn has become a problem that every major platform will encounter now. How to leave old customers and attract new customers is a problem that needs to be solved urgently. All major e-commerce platforms use large amounts of coupons to maintain old users and attract new customers. However, random coupons cause insignificant interference to most users. This topic examines how to push coupons more efficiently. This paper uses the O2O scene related data provided by Alibaba. Through feature engineering, it extracts features from users, merchants, coupons and other aspects, expands the data set, and generates the final data set for XGBoost model training. Through parameter tuning and evaluation and selection of the importance of features, the model is optimized to achieve an effective prediction of the user's use of coupons within a specified time. This article is implemented in Python. The Alibaba company's January-June data was used for processing analysis and modeling, and the use of coupons received by users in July was predicted.

Keywords

E-Commerce Coupon; Decision Tree; XGBoost; Feature Engineering.

1. Introduction

1.1 Research background

With the information technology revolution and the continuous development of mobile devices, O2O e-commerce as a new type of circulation has further affected everyone's production and consumption methods, which has brought about tremendous changes in the economic growth patterns of countries around the world. At the same time, China's online shopping and online payment have also developed rapidly. O2O e-commerce is favored by more consumers because of its low cost and high efficiency. It also enables more SMEs to find business opportunities and quickly occupy the market. Large companies have reduced costs and broadened sales channels.

In recent years, China's O2O e-commerce field has developed rapidly, and more and more enterprises have participated in it. The competition among major enterprises is fierce, and new types of e-commerce models have also joined in, making the share competition of e-commerce hot. . Customer churn has become a problem that every major platform will encounter. How to leave old customers and attract new customers is an urgent problem to be solved. In addition to strict quality control and improved website service capabilities, it is also a good way to carry out more discounts. With the continuous development of information technology in recent years, electronic coupons have gradually

ISSN: 2414-1895 DOI: 10.6919/ICJE.201908_5(9).0008

become a popular promotion means because of the use of the Internet as a medium of communication. All major e-commerce platforms have maintained large numbers of coupons to maintain old users. Attract new customers. However, random coupons cause insignificant interference to most users. For merchants, freely placing coupons not only increases the cost of marketing, but also has a certain impact on the brand's reputation. Therefore, in today's era of big data, we should use data mining technology to analyze the user's historical behavior record, establish a suitable model, predict the write-off behavior of coupons, and achieve personalized coupons and improve offers. The utilization rate of coupons not only can effectively reduce the cost of the merchants, increase the sales volume, but also enable consumers to get real benefits.

The recommendation system widely used by many commercial websites is to use the historical behavior data of customers to predict their preferences and make corresponding product recommendations. Data mining is more widely applied in the field of e-commerce.

1.2 Data mining method in O2O e-commerce

Data mining refers to the extraction of knowledge from large data sets [1]. Commercially speaking, data mining is a kind of commercial information processing technology. Its main feature is to extract, transform, analyze and other model processing of a large amount of business data in a commercial database, and extract key data for assisting business decision-making [2].

The main data mining methods used in the field of e-commerce are:

(1) Association rules. The association rule is to find the association between items in a given data set. By looking for some data associated with the data set, the correlation between the data is inferred [3].

(2) Classification rules. Classification analysis is the use of data objects of known categories to find models that describe and distinguish categories or concepts. It can be used to predict the category of unknown data objects [4]. By using data mining in O2O e-commerce, it is possible to provide decisions for different users and suppliers, and help suppliers to issue coupons more effectively. Classification methods include, for example, neural network algorithms, decision tree algorithms, SVM algorithms, K-nearest neighbor classification algorithms, and the like.

(3) Cluster analysis. It groups data objects into classes or clusters according to the attribute information of the data objects or the relationship between the objects, so that the objects in the same cluster have a high degree of similarity, and the objects in different clusters are different from each other [5]. As a basic data method, cluster analysis is widely used in similar search, customer segmentation, trend analysis, financial investment and other fields [6]. By clustering the coupons of merchants that are frequently purchased in the user behavior record, the coupons can be improved; clustering customers with similar purchasing habits, and discovering the characteristics of each type of customer can be targeted for specific products. And publicity. Using cluster analysis in e-commerce platforms can help operators develop and implement marketing strategies to provide customers with more suitable and satisfactory services. Commonly used clustering methods include: k-means algorithm [7], factor analysis, clustering algorithm based on probability model.

(4) Discovery of time series patterns. In time-ordered transaction sets, find other similar time-series events according to time, and find patterns with high probability of recurrence. Time series mode In e-commerce, it is to predict the customer behavior, find the user's interest sequence, and provide personalized service for each customer. Common findings using time series patterns include: Auto-Regressive Integrated Moving Average (ARIMA), long-momery time-series modeling, and autoregression.

(5) Prediction and evaluation. Data analysis and induction are carried out from historical data, and its regularity is obtained through data mining, so as to construct a model to predict and evaluate the development trend and results of new events.

ISSN: 2414-1895 DOI: 10.6919/ICJE.201908_5(9).0008

2. Related work

With the development of computer information technology, algorithms such as logistic regression, neural network, decision tree and random forest have become the research hotspots of user data analysis. Logistic regression is a generalized linear regression model based on linear regression and uses logic functions to predict classification problems [8]. Because the logistic regression algorithm has better interpretability on the two-classification problem, and can better fit the function relationship between the independent variable and the dependent variable, Peng Kai [9] et al. use logistic regression algorithm to evaluate the stability level for telecom customers. Yan Changyi [10] and others used the logistic regression algorithm to study the customer classification problem in the telecommunications industry to prevent customer loss.

Neural network is an intelligent information processing technology that imitates the processing process of human brain information. It has the characteristics of self-organization, self-adaptation and self-learning [11]. Potharst [12] used neural network algorithm to construct a model to predict the possibility of repeated purchases by customers, and found that neural network algorithms have strong learning ability for feature variables with complex nonlinear relationships.

The basic idea of the decision tree is to deduct the classification rules of the decision tree representation from a bunch of irregular and unordered instances according to a certain standard in a top-down recursive way [13]. Gordini [14] uses the decision tree to predict the customer's purchase behavior, which clearly and intuitively indicates the logical classification compared to other prediction methods.

Random forests refer to the establishment of a forest in a random manner with unrelated decision trees. After the forest is built, when the new sample is input, let all the decision trees in the forest judge the category of the sample separately, which category is selected the most, and the prediction result is which category [15]. Guelman [16] used random forests to retain old customers in the insurance industry. Liu [17] applied the fusion vector model of Perceptual Vector Machine, Logistic Regression and Random Forest to the project of e-commerce platform to predict customer purchase behavior, and achieved good results.

Like random forests, GBDT (Gradient Boosting Decision Tree) is also a combined model based on decision trees. Its idea is to construct a decision tree every time the existing model loss function is reduced. Zou Run [18] used GBDT to recommend personalized products to Tmall users, and achieved good results in the construction of feature engineering.

This article will use the real online and offline consumer behavior data between January 1, 2016 and June 30, 2016 provided by Alibaba. Through data cleaning, data integration and other data preprocessing methods, the data will be processed through feature engineering. The data set is divided, including the interval for extracting features and the training data interval, and extracting features from the feature interval, including user features, merchant features, coupon features, user and merchant association features, and other features. Based on the XGBoost (eXtreme Gradient Boosting) algorithm, a predictive model is constructed to predict whether users will use it within 15 days of receiving the coupon in July 2016.

3. Method

3.1 Data preprocessing and feature engineering

3.1.1 Factors affecting users' consumption of coupons

Through the in-depth study of O2O business model and people's consumption habits, feature extraction can be better. For the coupons to be written off within the specified time, this paper analyzes the following possible influencing factors:

(1) Coupon factor. The coupon rate will inevitably affect the customer's consumption and the possibility of using the coupon. And different discount methods will have different degrees of impact. The preferential rate is necessarily the higher the preferential rate, the greater the possibility that

ISSN: 2414-1895 DOI: 10.6919/ICJE.201908_5(9).0008

customers use coupons. In the case of a certain discount rate, direct discounts will attract customers to use coupons more than the full reduction.

(2) User factors. For users with higher consumption times, users with higher consumption times are more likely to use coupons; the average user who uses smaller coupons will be more likely to use coupons. Higher. Users who receive the same number of coupons are more likely to use their purchases.

(3) Merchant factors. For merchants, merchants with a higher total consumption and a larger proportion of coupon consumption will have higher merchants who use coupons to collect coupons. .

(4) Other characteristics. The total number of coupons received by the user, the number of times the user has received the coupon, and the number of times the user has consumed the coupon will have a certain impact on the user's write-off of the coupon.

3.1.2 Data analysis

Before data processing, the first step is to fully analyze and understand the original data set. The purpose of this phase is to understand the relationship between the field attributes in the data tables in the data set and the different tables for the actual online and offline user consumption behavior data provided by Alibaba. This data set contains two tables: the user offline consumption and coupon collection behavior table, and the user O2O offline coupon usage prediction sample table.

Table 3-1 User offline consumption and coupon collection behavior table

Field name	Field description	Field description
User_id	User ID	Customer unique identifier
Merchant_id	Merchant ID	Merchant unique identifier
Coupon_id	Coupon ID	Coupon unique identifier
Discount_rate	Discount rate	x \in [0,1] represents the discount rate; x:y represents full x minus y.
Distance	User distance from merchant	The user's activity location is closest to the merchant's store distance is x*500 meters (if it is a chain store, the nearest store is taken), x\in[0,10]; null means no such information, 0 means less than 500 meters, 10 means More than 5 kilometers;
Date_received	Receive coupon date	Receive coupon date
Date	Date of consumption	If Date=null & Coupon_id != null, the record indicates that the coupon was received but not used, ie a negative sample; if Date!=null & Coupon_id = null, the normal consumption date; if Date!=null & Coupon_id != null , indicating the date of consumption of the coupon, that is, a positive sample;

ISSN: 2414-1895 DOI: 10.6919/ICJE.201908_5(9).0008

Table 3-2 User O2O offline coupon use forecast sample table

Field name	Field description	Field description
User_id	User ID	Customer unique identifier
Merchant_id	Merchant ID	Merchant unique identifier
Coupon_id	Coupon ID	Coupon unique identifier
Discount_rate	Discount rate	x \in [0,1] represents the discount rate; x:y represents full x minus y. Unit is yuan
Distance	User distance from the merchant	The user's activity location is closest to the merchant's store distance is x*500 meters (if it is a chain store, the nearest store is taken), x\in[0,10]; null means no such information, 0 means less than 500 meters, 10 means More than 5 kilometers;
Date_received	Receive coupon date	Receive coupon date

3.1.3 Data cleaning

Data cleaning is the processing of missing and smoothed noise data (dirty data) in the data. The basic methods of data cleaning are as follows:

(1) Vacancy value processing: There are many methods of processing, such as filling the gap value with a uniform constant, or filling the gap value with the average value of the attribute, or classifying all the data according to certain attributes, and then classifying the data. The data in the data is populated with the average value under this attribute. It is more common to use the most likely value to fill, using regression, decision tree or Bayesian, to guess the vacancy value from other existing data.

(2) Noise data processing: Noise is a random error or deviation in the data set. In practice, there is basically no noise data in the data, including the wrong value and the isolated point value that deviates from the expectation. Noise data is processed to reduce the distortion of the data, making the data more representative of the general characteristics of the data.

3.1.4 Data integration

With the continuous development of e-commerce, the amount of data is getting larger and larger, and the collected data is more and more complicated, so the data will be distributed, which inevitably requires data integration and merges the required information so that Processing and analysis of data. During the feature engineering process, various features are extracted, and the features need to be merged for model training.

3.1.5 Feature engineering

The user offline consumption and coupon collection behavior table includes user information, merchant information, shopping coupon information, discount information, and consumption information. The characteristics can be extracted according to the relationship between these attributes and attributes, and the model is predicted to predict the user under the O2O line. Whether the coupons in the coupon usage forecast sample form will be written off within the next 15 days.

Based on the data from January 1 to June 15 in the offline consumption and coupon collection behavior table of the user, the original data set is expanded according to the basic attributes and the

ISSN: 2414-1895 DOI: 10.6919/ICJE.201908_5(9).0008

basic attributes of the existing relationship, and the final data set is expanded to generate a final The data set for training of subsequent models.

(1) User characteristics

The user characteristics include: the number of times the user spends using the coupon; the total number of times the user consumes; the total number of coupons received by the user; the time interval from the receipt of the coupon to the use of the coupon; the proportion of the user after receiving the coupon; the user uses The ratio of coupon consumption to total consumption; the average distance between users and merchants; the farthest distance between users and merchants; the closest distance between users and merchants.

(2)Merchant Features

Merchant characteristics include: total number of merchants sold; number of coupons used in sales; total number of coupons placed by merchants; coupon usage rate in merchant consumption; usage rate of coupons issued by merchants; average distance between merchants and customers; The farthest distance from the customer; the closest distance between the merchant and the customer.

(3) Coupon Features

The coupon features include: discount rate for discount coupons; discount for full discount coupons; reduction for coupons for full reduction; coupons for full reduction; coupons for the first few days of the week; the time to receive the coupon is several months.

(4) Other features

Other features include: the total number of coupons the user received this month; the number of coupons the user received this month; whether the user received the last coupon in the same coupon this month; whether the user received the same coupon this month One; the number of coupons the user has ever received on a certain day; the same number of coupons the user gets on a certain day.

3.2 Model

Experiment environment

The experimental environment used in this experiment is shown in Table 3-3:

Table 3-3 Experimental environment

CPU	1.6 GHz Intel Core i5
RAM	8 GB 1600 MHz DDR3
Operating System	macOS High Sierra 10.13.4

Experiment procedure

The key part of this experiment is the establishment of the model. According to the predicted target, the relevant features are extracted from the data set, and the relationship between the data is modeled. Through multiple parameter adjustments and selection of important features, each time The training results are evaluated and the model with the best generalization ability is selected as the final model.

Firstly, the data set is analyzed, and the data set is divided into training set and test set. The sliding window method is used to divide the data into three data sets. The first data set is the data extraction feature from January 1st to April 13th, 2016, using the test set from April 14th to May 14th; the second data set will be February 1st. As a data set extraction feature on May 14th, using May 15th to June 15th as a test set; the third data set is a data set extraction feature from March 15th to June 30th; The data from July 1st to July 31st is the test set.

Secondly, through feature engineering, the characteristics of the data set are extracted and integrated, and the original data set is finally used as the data set used for model training. After each model is trained, the features are evaluated and the contribution to the training model will be extended. Larger features are selected to eliminate features that are less contributing. Through repeated training, the

ISSN: 2414-1895 DOI: 10.6919/ICJE.201908_5(9).0008

parameters of the model are modified and adjusted to obtain a model with strong generalization ability, and the training time complexity is reduced to avoid over-fitting.

3.3 Evaluation

The model evaluation is mainly to select a better model for the generalization ability of the new sample. Through the model evaluation, we can select the appropriate model type, model parameters, features that improve the generalization ability, and can be judged more intuitively. The pros and cons of the model. There are many methods for model evaluation. For classification models, AUC, Accuracy, F1-score, Precesion, Recall, etc. are the most commonly used. This paper mainly uses AUC as the main evaluation method, and the rest is the auxiliary evaluation method.

3.3.1 Sub-section Headings

AUC (Area Under ROC Curve) is the area under the AOC curve. In the two-category problem, according to the actual classification of the sample and the classification of the model prediction, the results are divided into the following types: true positive, false positive, true negative, False negatives are represented by TP, FP, TN, and FN, respectively.

Table 3-4 Confusion matrix

The true situation	Forecast result	
	Positives	Negatives
Positives	TP (True positives)	FN (False negatives)
Negatives	FP (False positives)	TN (True negatives)

The abscissa of the ROC curve is “false positives rate” and the ordinate is “true positives rate”, which are respectively defined as:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN+FP}}. \tag{3-1}$$

$$\mathrm{TPR} = \frac{TP}{TP+FN}. \tag{3-2}$$

In the ROC curve, the closer the point on the curve is to the lower right, the higher the probability of error in the positives under the current threshold, the lower the accuracy. If the point on the curve is closer to the upper left corner, the lower the probability of predicting the positives. The higher the accuracy [21]. AUC is the area under the ROC curve, and the range of the horizontal and vertical coordinates of the ROC curve is [0, 1]. Generally, the value of AUC is within [0.5, 1], and the larger the better.

3.3.2 Accuracy

Accuracy: Accuracy refers to the proportion of correctly classified samples to the total number of samples [22]. To a certain extent, it is possible to preliminarily judge whether a classifier is effective from the accuracy rate, but sometimes the effect is not good, and it cannot evaluate the model well from the test data.

3.3.3 Precision (Precision)

Precision: (Precision): It should be noted that this is not the same as Accuracy, which is the ratio of the true positives in the sample that is predicted to be positive. Accuracy is also known as precision.

3.3.4 Recall rate (Recall)

Recall rate: The recall rate refers to the proportion of samples that are predicted to be positive in the real positives, also known as the recall rate [23].

3.3.5 F1 score (F1-score)

F1 score (F1-score):. F1 is defined as the harmonic mean of the accuracy and recall rate [24].

ISSN: 2414-1895 DOI: 10.6919/ICJE.201908_5(9).0008

3.4 XGBoost algorithm

The XGBoost parameters used in this article are shown in Table 3-5:

Table 3-5 XGBoost parameters

Parameter name	Value
booster	gbtree
objective	rank:pairwise
eval_metric	auc
gamma	0.3
min_child_weight	1.1
max_depth	5
lambda	10
subsample	0.7
colsample_bytree	0.7
colsample_bylevel	0.7
eta	0.01

The XGBoost model construction process first reads the extended data set after feature engineering from the csv file, divides the data set into training set and verification set, and trains the XGBoost model. Through parameter tuning, the generalization ability is found. A strong model ultimately tests the model on the prediction set.

At the end of the training, the AUC value was 0.8223, the Accuracy value was 0.9206, the Precision was 0.7767, the Recall was 0.2753, and the F1-score was 0.3814. The ratings for all features are as follows:

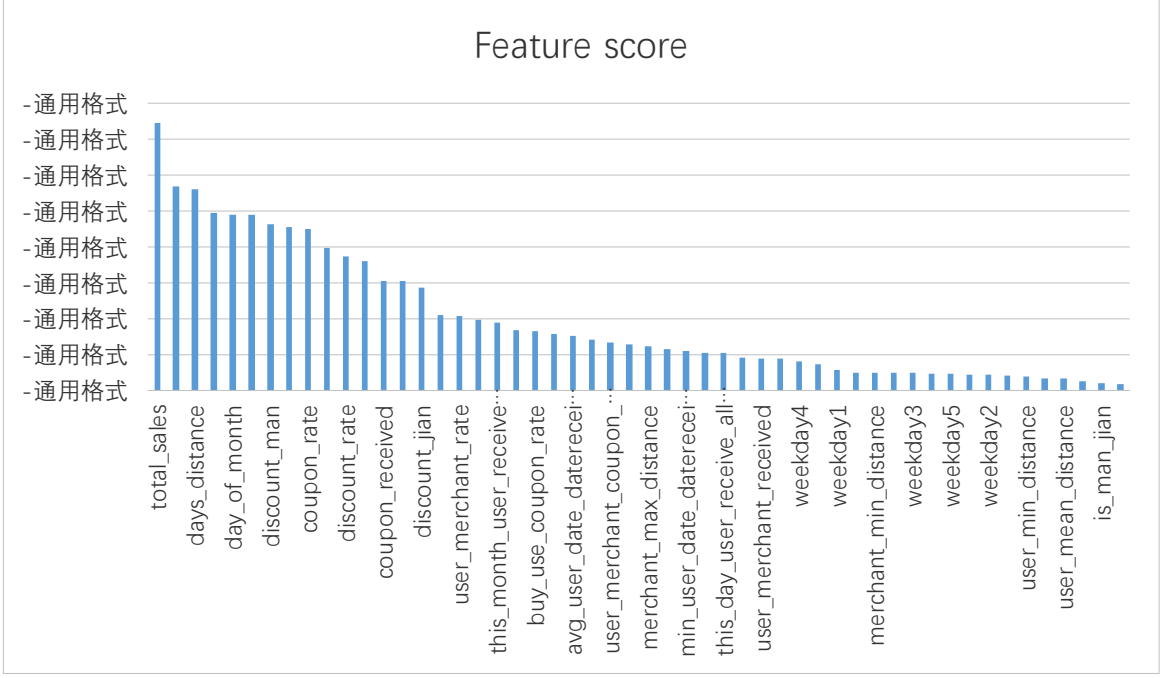

Fig. 1 Phase analysis of arrangement

ISSN: 2414-1895 DOI: 10.6919/ICJE.201908_5(9).0008

3.5 Model optimization

For the optimization of the XGBoost model, it is mainly through parameter tuning. The most influential parameter is max_depth, which is determined step by step using the dichotomy in the range [1,7]. Finally, when the value is 5, the value of auc is the largest. Then adjust the gamma parameter, which is determined from the range of [0, 0.4], and finally obtain the optimal value when gamma=0.1. The subsample and colsample_bytree parameters are in steps of 0.05 and take values around 0.7. The final optimal value is still 0.7. The final optimal parameters are shown in the following table:

Table 3-6 XGBoost optimal parameter list

Parameter name	Value
booster	gbtree
objective	rank:pairwise
eval_metric	auc
gamma	0.1
min_child_weight	1.1
max_depth	5
lambda	10
subsample	0.7
colsample_bytree	0.7
colsample_bylevel	0.7
eta	0.01

4. Conclusion

This paper studies the issue of the low effectiveness of Internet O2O corporate coupons, and uses the relevant data of O2O scenarios provided by Alibaba. Based on the decision tree, it predicts whether users will use the corresponding coupons within the specified time. To provide an analytical basis for future coupon push decisions. The main contents of this paper are as follows:

(1) For the problem of user coupon write-off, this paper analyzes the current challenges of O2O e-commerce and emphasizes the importance of data mining. Several data mining methods commonly used in e-commerce are introduced, and the decision tree and XGBoost algorithm are introduced.

(2) Introduced the data preprocessing method and feature engineering. In this paper, data cleaning and data integration are carried out for the data required for model training. According to the analysis of the historical behavior of users' verification of shopping vouchers, four aspects of user characteristics, merchant characteristics, coupon characteristics and other features are extracted through feature engineering. .

(3) Introduced the steps of constructing the XGBoost model, and optimized the model through parameter tuning, and used AUC, Accuracy, Precision, Recall, F1-score to evaluate the model. Each training model, XGBoost model for feature engineering The eigenvectors in the score are scored, and the features that contribute little to the model are eliminated, leaving a feature that is more optimized for the model.

The model trained by the feature engineering proposed in this paper has a higher score in the model evaluation, and has achieved ideal results for both the training set and the test set. It proves that the feature engineering and model proposed in this paper are very suitable for predicting whether the user can write off the coupon.

ISSN: 2414-1895 DOI: 10.6919/ICJE.201908_5(9).0008

The model based on XGBoost algorithm proposed in this paper can predict whether users can write off coupons, which is feasible. However, for some data, there are still unsatisfactory situations, and the model is relatively simple and has room for improvement. And the optimization of the model is only in the parameter tuning, only a small increase. To further improve the AUC value, there are several aspects that can be improved:

(1) The feature engineering can be further enriched. Although the feature engineering in this paper extracts more kinds of data information, the number is still less, and more features can be extracted to extend the data set.

(2) Use other kinds of algorithms to build models, such as random forests, neural networks, and so on.

(3) Perform model fusion. The improvement effect of a single model is limited. If you want to improve further, you need to try to integrate with other models. Through the prediction of multiple algorithms, the results of all algorithms are combined to be the final result.

References

[1] Mao Guojun. The concept, system structure and method of data mining [J]. Computer Engineering and Design, 2002 (08): 13-17.

[2] Shi Yong, Zhang Yue. Let the data speak: the commercial value of data mining technology and its application in the financial industry [J]. Capital Markets, 2007 (11): 80-83.

[3] Zhu Huiyun. Application of Data Mining in E-commerce [D]. Hohai University, 2003.

[4] Wang Huilian. Application of decision tree algorithm in fault diagnosis of thermal power units [D]. North China Electric Power University (Beijing), 2006.

[5] Wang Lizhen, Zhou Lihua, Chen Hongmei. Principles and Applications of Data Warehouse and Data Mining [M]. Beijing: Science Press, 2005: 64-68.

[6] Shen Hongchao. Application of Data Mining Technology in E-commerce [D]. Jiangnan University, 2009.

[7] Wu Liping. Research on traffic anomaly detection based on local wave decomposition [D]. University of Electronic Science and Technology of China, 2011.

[8] Glosemeyer D. Fitting Generalized Linear Models [J]. 2017.

[9] Peng Kai, Qin Yongbin, Xu Daoyun. Customer Stability Modeling Based on Logistic Regression[J]. Computer Engineering, 2011, 37(9): 12-15.

[10] Yan Changwei, Hu Jianhua, Zhou Haihe. Classification of Telecom Customers Based on Clementine Logistic Regression[J]. Science and Technology Information, 2008(36): 94-94.

[11] Deng Shufang. Construction of a personal credit evaluation combination model based on decision tree-neural network [D]. Hunan University, 2012.

[12] Potharst, Rob, Kaymak, U, Pijls, Wim. Neural Networks for Target Selection in Direct Marketing [J]. Erim Report, 2001.

[13] Xing Yuankai. Research and application of neural network based on decision tree and genetic algorithm [D]. Zhejiang University, 2010.

[14] Gordini N, Veglio V. Customer relationship management and data mining: a classification decision tree to predict customer behavior in global markets[M]// Handbook of Research on Novel Soft Computing Intelligent Algorithms: Theory and Practical Applications. 2013:1 -40.

[15] Yan Kai. Research on feature selection and model optimization algorithm of random forest [D]. Harbin Institute of Technology, 2008.

[16] Guelman L, Guillén M, Pérezmarín A M. Random Forests for Uplift Modeling: An Insurance Customer Retention Case [J]. Lecture Notes in Business Information Processing, 2012, 115:123-133.

[17] Liu G, Nguyen T T, Zhao G, et al. Repeat Buyer Prediction for E-Commerce [C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 155-164.

[18] Zou Run. Research on User Personalized Recommendation Based on Model Combination Algorithm [D]. Nanjing University, 2014.

ISSN: 2414-1895 DOI: 10.6919/ICJE.201908_5(9).0008

[19] Quinlan R. Introduction of Decision Trees [J]. Machine Learning, 1986, 1(1): 81-106.

[20] Xiang Dong. Research on wood texture classification algorithm based on LBP-DEELM [D]. Northeast Forestry University, 2016.

[21] Zhai Jianrong. Research on telecom customer churn prediction method based on hybrid model [D]. University of Electronic Science and Technology of China, 2009.

[22] Song Longgao. Research on Network Service Flow Recognition Based on Decision Tree [D]. Nanjing University of Posts and Telecommunications, 2015.

[23] Wang Wei. Unbalanced data classification based on upsampling and integrated learning [D]. Xiamen University, 2017.

[24] Wang Na. Research on methods of human behavior prediction in video [D]. Hunan University, 2016.