

Research on User Portrait Based on Xgboost and Logistic Ensemble Learning Methods

Jiafeng Hu^a, Guangyi Huang^b, Lili Wang^c

School of Anhui University of Technology, School of mathematics and big data, Anhui 232001, China

^a 605281999@qq.com, ^b 317626347@qq.com, ^c 64460112@qq.com

Abstract. In the era of big data, accurately analyzing users' purchasing behavior and specifying strategies for corresponding enterprises can not only reduce enterprise costs but also improve users' consumption habits, which is of great significance to the development of enterprises. In view of this, this paper analyzes the data of user behavior. Firstly, the original data are preprocessed, and then feature engineering is carried out based on correlation. The important features were selected as reference and 15 features with practical significance were extracted. Considering the imbalanced data, this paper adopts resampling to balance the data. After continuous experiments, a combined model based on three models is proposed: Based on the parallel combination of XGBoost and RandomForestClassifier, serial modeling is carried out with LogisticRegression, and the final model is established to predict user consumption behavior. It can be observed from the experimental results that the combinatorial model not only has the advantages of high classification accuracy in tree model, but also has the advantages of strong interpretation and high stability in logistic regression. Due to the great difference between the logic regression model and the tree model, the correlation of the results is low. Therefore, this method is feasible and scientific.

Keywords: Feature engineering, Resampling, Combination model, Consumption behavior, User portrait.

1. Introduction

With the recurrence of the epidemic, the global real economy is still in the doldrums, while the Internet economy has ushered in a larger scale of growth. According to China Internet Development Report (2021), the scale of China's big data industry will reach 71.87 billion yuan in 2020, with a year-on-year growth of 16.0%, leading the global big data market. In 2023, the market size of public cloud will break through 200 billion yuan. In this period, online training and network learning platform also ushered in a new round of spring. In 2020, the scale of China's Internet audiovisual network reached 241.2 billion yuan, and the scale of China's Internet audiovisual users continued to grow.

In the Internet era, the previous marketing model is no longer applicable, and many enterprises have found another way to establish their own big data platform. Using past user behavior data to analyze user behavior can get many unexpected results. User behavior analysis is an important means to study user habits. Understanding user habits aims to better recommend more valuable information to different users, so as to reduce unnecessary costs for enterprises, make users more satisfied and reduce customer churn. On the other hand, many algorithms have been widely used in the analysis of user behavior. In the case of large amount of data and large amount of features, the establishment of an appropriate analysis model requires multi-angle analysis and multi-dimensional processing of data. Read foreign relevant literature, its research is user-centered, from the user's perspective of all aspects of research. Including user demand, user heart and so on. Literature [1] studies the individuation of user portrait and constructs user portrait. Literature [2] carried out research based on mobile phone users. Literature [3] proposed to build a model based on user portraits of keywords, extract key information and label corresponding users. Literature [4] analyzes users' collection of commodities and designs user portrait algorithm based on users' interests. Literature [5] studies user interest, prediction model algorithm and recommendation algorithm. Literature [6] uses logistic regression

algorithm model to model and analyze user information. Literature [7] uses neural network algorithm to build user model, predict user interest tags, and construct dynamic user portraits.

Read relevant domestic literature. In 2016, Lu Yiqun and Yu Xiaobing collected and analyzed the dynamic information of customers, used the dynamic information to predict the number of users [8], and realized the dynamic management of users and the formulation of plans. In the same year, Wenwen Yi studied and analyzed the click rate of users, and realized the prediction of click by logistic regression. In 2016, Wang Meiji, Yang Xiumei and Sun Yong proposed a decision tree model to predict user preference information based on users' historical browsing information. This model effectively improves user satisfaction, and also solves the cold start phenomenon when facing new users [9]. In 2017, Sun Lin used regression model to predict the loss of users and carried out targeted telephone return visits to reduce invalid customer return visit tasks [10].

Through the analysis and understanding of relevant literatures at home and abroad, it is concluded that the research field of user portrait mainly focuses on three aspects: user attribute, user interest and user behavior. User attributes can be analyzed and obtained from the perspective of statistics, user interests can be obtained from user data, and user behavior can be mined and studied through a large amount of data analysis to analyze user behavior characteristics in all aspects, so as to further analyze the connections and differences between group users and individual users.

Based on this point, this paper starts from statistical analysis, feature engineering and model building. After data preprocessing based on correlation analysis, relationships between different indexes, again by distribution to extract important features, finally, the different algorithms together, complete with the result of user behavior prediction algorithm framework to build, model can dynamically update user data, can improve real-time behavior system, can greatly improve the user experience degrees.

2. Data Analysis

2.1 Data Preprocessing

1) Delete the missing value

Due to a large number of missing data in the city field, this field needs to be deleted when we conduct statistical analysis of this field in the following paragraphs to establish the model.

After data fields in the user_INFO table are processed, the result shown in Figure 1 is displayed.

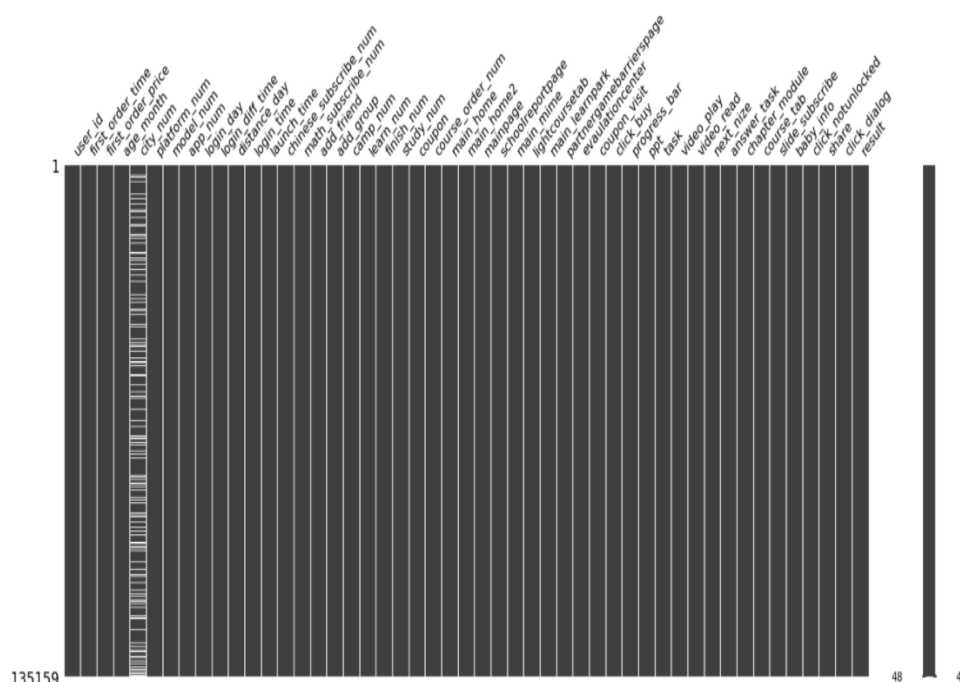


Figure 1. Missingno

2) On the sampling



Figure 2. Distribution of purchased and unpurchased formal courses

The original data is on the left, and the sampled data is on the right.

Set user purchase record to 1 and no purchase record to 0. By observing the above two figures, we find that the number of these two is extremely imbalanced, sample imbalance will affect the learning of classifier, therefore, we will use SMOTE method to solve the sample imbalance and ensure the data balance.

3) Unique thermal coding

One-hot encoding mainly uses n-bit status register to encode N states, each state is independent of its register bits, and only One is effective at any time. Independent thermal coding is a technology that uses 0 and 1 to represent some parameters and uses n-bit status register to encode N states.

In this paper, chinese_Subscribe_NUM and study_num fields in the login_day table are selected for unique hot coding.

2.2 Statistical analysis



Figure 3. Correlation map

(1) In terms of the number of learning lessons, the number of completing lessons and the number of repeated learning, our analysis results show that the proportion of users who have more lessons is higher, which is positively correlated with the final result of buying lessons, and the correlation among the three is high.

(2) From the analysis of the number of coupons received and the click-through number of home page ads, it is found that the number of coupons received and the click-through number of home page ads are highly correlated, and both the number of coupons received and the click-through number of home page ads are positively correlated with the final result.

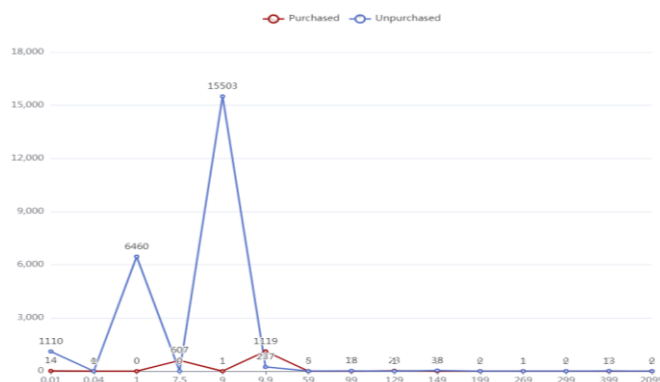


Figure 4. The relationship between the order amount of experience courses and the purchase of formal courses

In terms of the purchase price of experience lessons, we find that users who buy paid experience lessons or those who buy experience lessons at a high price have a better proportion of purchasing lessons in the final result. Those who buy experience lessons account for the total data, and those who buy experience lessons account for the total data in the final result of buying lessons. It can be seen that users who buy paid experience lessons are more likely to buy lessons.

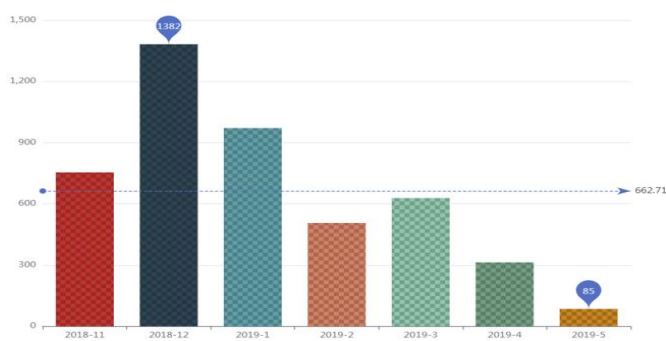


Figure 5. The relationship between the number of formal courses purchased in each period

The possible reason for this situation is that in December, when the final examination is approaching, parents or students buy courses online to consolidate and review the knowledge, and in March, the decline starts again. The main reason is that students go to school shortly after the school starts. To sum up, it is strongly suggested that the company should increase publicity at the end of each semester, which can greatly improve the company's profits.

Analyze the number of course purchasers in city_num(city) in the attached user_INFO table, and draw the geographical location distribution map of course purchasers after removing the blank field, as shown in Figure 6.



Figure 6. Users who purchase formal courses are distributed in cities

The final purchase of formal courses in major cities is shown in Figure 7.

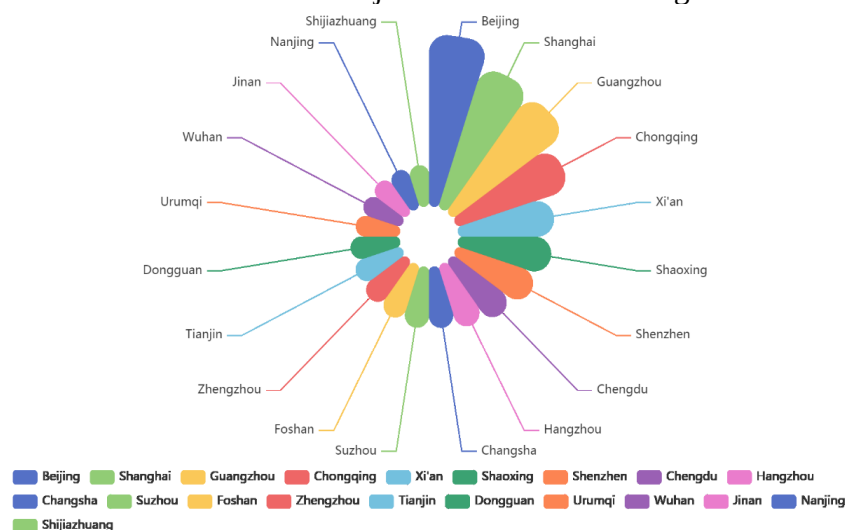


Figure 7. Major cities eventually purchase official course distribution

People who buy courses are mainly concentrated in developed areas such as Beijing, Shanghai, Guangzhou, Shenzhen and Chongqing, which is also related to the local Internet penetration rate and economic development. Through the analysis of the data in the table, it can be concluded that the number of experience courses sold in developed cities is significantly higher than that in other underdeveloped cities.

2.1 Feature selection

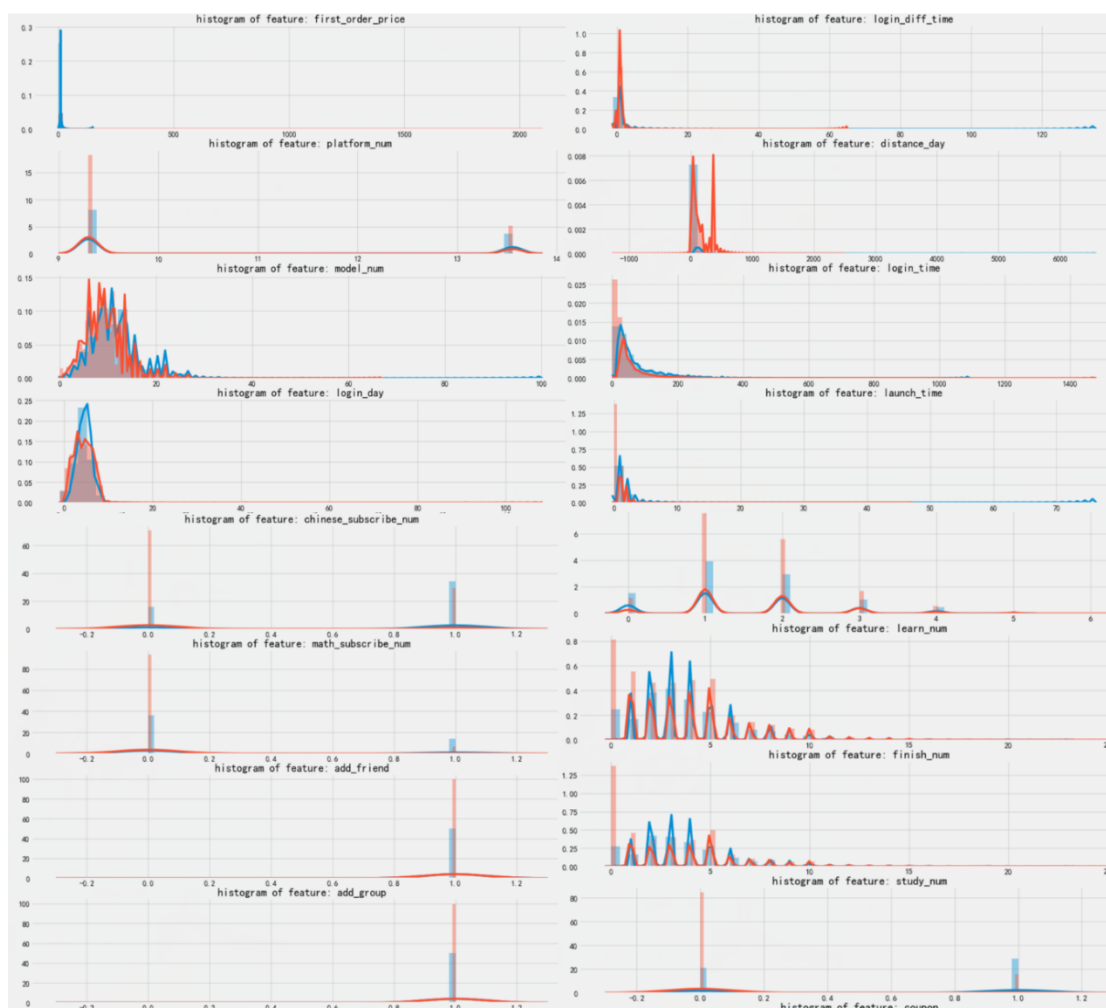


Figure 8. Distribution of partial fields between purchased and unpurchased courses

The yellow for formal courses in figure 8, for not to purchase the blue into formal curriculum, the yellow and blue graphics were observed by above with similar trends, explain the characteristic value is not obvious effects on whether to buy the formal curriculum, should give up, if observed at yellow and blue graphic trend has certain differences, then the characteristic value impact on whether to purchase course, Its eigenvalues should be retained. See Table 1.

Table 1. Fields after feature selection

name	meaning
coupon	Coupon redemption number
distance_day	Number of days since the end of term
course_order_num	Some classes have not completed orders
login_diff_time	Login interval
login_day	Log in number of days
model_num	Mobile phone models
age_month	age
login_time	The login time
study_num0	Lesson repeat _ Unique heat coding 1
camp_num	Number of classes
video_Read	Interface continue access number
learn_num	Learn the number of lessons
chinese_subscribe_num1	Pay attention to the public number 1_ unique heat code 1
main_home2	Number of homepage Visits
study_num1	Heat coding 2
chinese_subscribe_num2	Concern public number 1_ unique heat code 2
finish_num	To complete the course

3. Model establishment and solution

1) Gboost algorithm

The objective function of XGBoost is shown in Equation (1):

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

The loss function is represented by $L(y_i, \hat{y}_i)$. the constant term is represented by C, and the regular term is represented by $\omega(F_T)$. The model will overfit and the complexity of the model will be increased by the loss function. To solve this problem, the regular term is added into the objective function, and the Taylor expansion is shown in (2)

$$\begin{aligned} g_i &= \partial_{\hat{y}^{(t-1)}} L(y_i, \hat{y}^{(t-1)}) \\ h_i &= \partial_{\hat{y}^{(t-1)}}^2 L(y_i, \hat{y}^{(t-1)}) \end{aligned} \quad (2)$$

Substitute Formula 2 into Formula 1 to calculate the second-order Taylor expansion in formula 1, and the final expression to be solved is as shown in (3):

$$\text{Obj}^{(t)} \cong \sum_{i=1}^n [L(y_i, \hat{y}^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) + c \quad (3)$$

In order to further obtain the accurate final prediction results of colleges and universities, the objective function of solving the process, for the reciprocal error only need to solve the first and second order.

2) Logistic regression algorithm

Logistic regression is a kind of generalized linear regression. The linear regression model has reliable statistical theoretical basis and high robustness. As it belongs to the white box model, it can

provide the expression of each indicator and the default situation, and can intuitively reflect the interpretation of each indicator to the default, so as to obtain reasonable and correct prediction.

$$\log i t(p) = X\beta + \varsigma, \varsigma \sim N(0, I\sigma^2) \quad (4)$$

$$\hat{p} = \frac{\exp(X\hat{\beta})}{1+\exp(X\hat{\beta})} \hat{p} \text{ is the probability of default} \quad (5)$$

However, when the model is trained on over-sampled samples, the prediction probability will be biased. At this point, the model is:

$$\log i t(p^*) = \ln\left(\frac{\rho_1\pi_0}{\rho_0\pi_1}\right) + X\beta + \varsigma, \varsigma \sim N(0, I\sigma^2) \quad (6)$$

Where, P^* is the posterior probability of biased samples, and is, so the adjusted posterior probability is:

$$\bar{p} = \frac{p^* p_0 \pi_1}{(1-p^*) p_1 \pi_0 + p^* p_0 \pi_1} \quad (7)$$

3) User portrait model based on Stacking policies

Step1: Divide 150,000 pieces of data into 10 pieces, among which 8 pieces are training set T and 2 pieces are verification set M

Step2: Fold the training set T by 50%. The first layer model is random forest and Xgboost. After training each feature, integrate the results of each model into a new training set T' by Stacking strategy.

Step3: Make a 50% discount on the verification set M, and then make a prediction in the first-layer model, and the results are used as the second verification data set M' after average processing.

Step4: The second layer model is logistic regression. T' in the second step and M' in the third step are put into the second layer model, and finally the task of user behavior prediction and classification is realized

4. Experimental Analysis

4.1 The model results

To make the model have better prediction effect and generalization ability, it is necessary to optimize the parameters of the model. Taking Xgboost as an example, the model predicts whether the customer will buy the product or not, so it is binary. Therefore, binary: Logistic and ETA are selected as the learning rate, which can improve the robustness of the model by reducing the weight of each step. Min_child_weight is the sum of the weight of the minimum leaf node sample. The smaller the value, the easier it is to overfit. Colsample_bytree Selects the characteristics of the spanning tree to prevent overfitting. Indicators of reg regularization. Finally, the following parameters are selected under GridSearchCV test, as shown in table 2, 3.

Table 2. Xgboost and random forest model parameters

Xgboost		Random forests	
parameter	Selected value	parameter	Selected value
objective	binary:logistic	criterion	gini
n_estimators	300	n_estimators	500
max_depth	4	splitter	best
eta	0.1	oob_score	False
min_child_weight	3	bootstrap	True
colsample_bytree	0.8	min_samples_split	2
reg_alpha	0.2	min_samples_leaf	1
subsample	0.5	min_weight_fraction_leaf	0

Table 3. Parameters of logistic regression model

Logistic regression	
parameter	Selected value
C	10
penalty	L2
class_weight	auto

When selecting the base model of the first layer, the GRU deep learning model with good performance in the prediction field and XGBoost, one of the strongest machine learning algorithms, should be selected first. Meanwhile, several other models with excellent performance should be considered as the base learner, because the base model with strong learning ability is conducive to the improvement of the overall prediction effect. RF and gradient boosting Decision Tree (CDBT) have been widely applied in various fields, and logistic regression has good practical application effect due to its mature theory and efficient training.

In the second layer, the model with strong generalization ability should be selected to induce and correct the bias of multiple learning algorithms to the training set, so as to prevent the over-fitting phenomenon by means of set. Therefore, LR and SVM are initially selected for the base learner of the Stacking integration model in the first layer, RF is selected as the meta-learner in the second layer, and the resulting model architecture is shown in Figure 9:

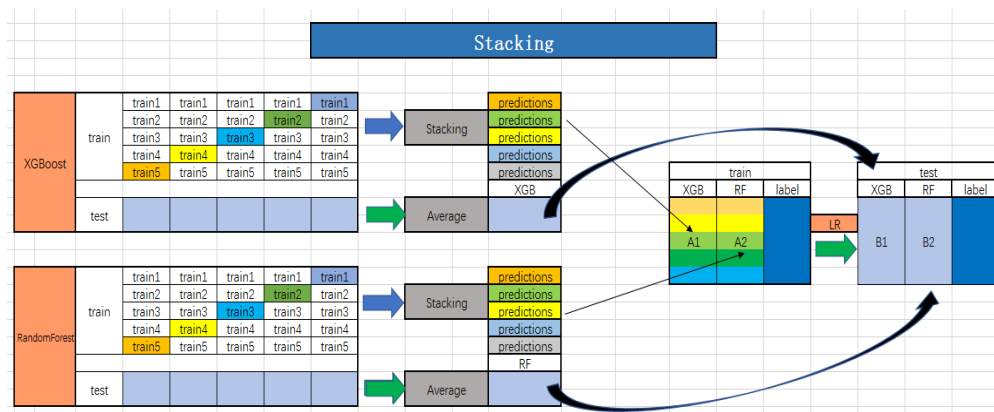


Figure 9. Model architecture diagram

4.2 Results contrast

Table 4. XGBoost model evaluation

	precision	recall	F1-score	support
0.0	0.973	0.963	0.968	39218
1.0	0.963	0.973	0.968	39094
accuracy			0.968	78312
Macro avg	0.968	0.968	0.968	78312
Weighted avg	0.968	0.968	0.968	78312

Table 5. Logistic regression model evaluation

	precision	recall	f1-score	support
0.0	0.92	0.91	0.92	39218
1.0	0.91	0.92	0.92	39094
accuracy			0.92	78312
Macro avg	0.92	0.92	0.92	78312
Weighted avg	0.92	0.92	0.92	78312

Table 6. Portfolio model evaluation

	precision	recall	f1-score	support
0.0	0.99	1.00	1.00	129808

1.0	1.00	0.86	0.93	5351
accuracy			0.99	135159
macro avg	1.00	0.93	0.96	135159
weighted avg	0.99	0.99	0.99	135159

As you see, a combination of Staking strategies is superior to a simple algorithm, with better accuracy, precision and completeness of prediction results, and a higher F1 score, reflecting a more robust model.

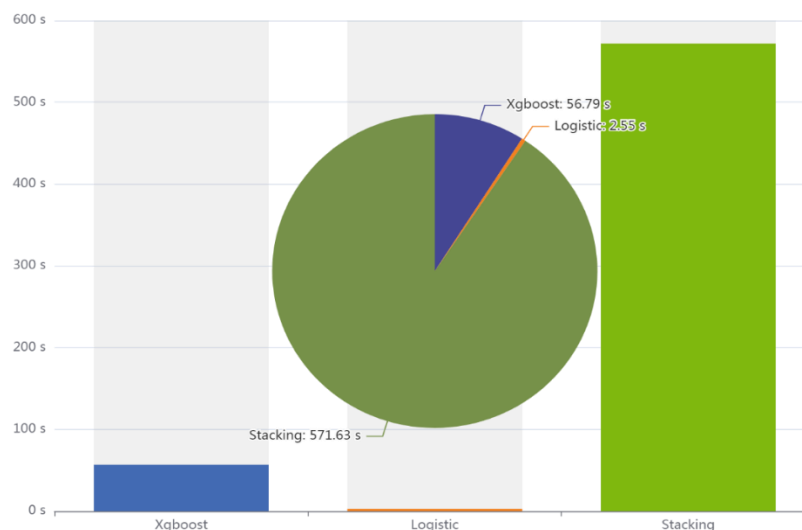


Figure 10. Comparison of time spent by different models

Stacking models takes much more time than other models in terms of time, and if we need a compromise between time and robustness of the model, Xgboost clearly fits our needs.

5. Summary

With the rapid development of big data technology, how to deal with massive data has brought new opportunities and challenges to many enterprises. In this paper, conclusions and suggestions are given from the analysis results of a series of characteristics, such as the number of repeated learning, the number of receiving papers and the number of clicking to share visits.

1) Free products are more likely to drive consumer spending

In terms of the number of lessons learned, number of lessons completed and number of repeated lessons learned, the results of our analysis show that the results of the three are similar: users with more lessons will buy more lessons, which is positively correlated with the final result of buying lessons, and the purchasing rate of users will increase.

2) Coupons can attract more users

Based on the analysis of the number of coupons, it can be found that the proportion of users who receive coupons or those who receive more coupons is higher than that of those who receive coupons. Therefore, the platform can distribute some free coupons for a limited time to active users, which has a positive impact on the final result of buying courses.

3) Strengthen multimedia operation

From into the group and pay attention to the public micro signal to buy class proportion, pay attention to the public account user final consumption of about, into the group. Therefore, increasing the strength of multimedia operation can allow more users to consume.

This paper analyzes user behavior in the context of big data, and predicts whether users will consume according to their behavior. Comparing the performance of different algorithms, it is found that Stacking model based on Xgboost and Logistic combination can be used to predict customer consumption. There are still some deficiencies in this paper.

1) The paper has a shallow understanding of the field background, and the city field with serious deficiency was deleted in the final modeling process. However, these city fields can be used for classification modeling in real business.

2) Due to the limitations of data and software itself, the model cannot be widely applied in different user behavior analysis fields. In the future, more user behavior data from different software will be selected to make the model more representative.

Supported by: Anhui Provincial Teaching Reform Research Project (2020JYXM0443, 2020JYXM0441), Anhui Provincial Ideological and Political Work Innovation Project (2020ZDXSJG094) natural Science Foundation of Anhui Province (2008085QD178); Excellent Talents Support Program of Anhui Universities (No. GxyqZD2020020)

References

- [1] Adomavicius G, Tuzhilin A. Personalization technologies [J]. Communications of the ACM, 2015, 48 (10): 83-90.
- [2] Barabasi A L, Albert R. Emergence of Scaling in Random Networks. Science 286, 509-512 [J]. Science, 1999, 286 (5439): 509-512.
- [3] Stanescu A, Nagar S, D Caragea. A Hybrid Recommender System: User Profiling from Keywords and Ratings [C] // IEEE/WIC/ACM International Joint Conferences on Web Intelligence. ACM, 2013.
- [4] Pazzani M, Billsus D. Learning and Revising User Profiles: The Identification of Interesting Web Sites [J]. Machine Learning, 1997, 27 (3): 313-331.
- [5] Wernli, Karen, J, et al. Screening for Skin Cancer in Adults: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force. [J]. JAMA: Journal of the American Medical Association, 2016, 316 (4): 436-447.
- [6] Slanzi G, Balazs J, Velasquez J D. Predicting Web User Click Intention Using Pupil Dilation and Electroencephalogram Analysis [C] // IEEE/WIC/ACM International Conference on Web Intelligence. ACM, 2017.
- [7] Andrejkova, Gabriela, Kuzma. Predicting user's preferences using neural networks and psychology models [J]. Applied Intelligence the International Journal of Artificial Intelligence Neural Networks & Complex Problem Solving Technologies, 2016.
- [8] Yu Xiaobing, Lu Yiqun. Systems Engineering, 2016, 34 (9): 7.
- [9] Yang Xiumei, SUN Yong, Wang Meiji, et al. Application of Computer Systems, 2016 (3): 4.
- [10] Sun Lin. Building customer churn Forecasting Model based on logistic regression algorithm [J]. Information Systems Engineering, 2017 (6): 1.