E-Commerce User Purchase Prediction Analysis Based on Data Mining

Qixian Rui, Xinyu Tan

Northeastern University, Shenyang 110819, China.

Abstract: In the Internet era, various e-commerce platforms are gradually rising, people's consumption has gradually changed from offline to online. How to predict users' purchasing behavior based on big data has become an important challenge for e-commerce platform. Based on the knowledge of e-commerce, this paper constructs a funnel model to analyze the user conversion from each link, so as to locate the link with low conversion rate. The data mining method is used to process the user history data of an international e-commerce company. And through the importance of features,we can know from which aspects to promote the transformation of users, so as to take targeted measures to users.

Keywords: Data Mining; Funnel Analysis; Logistic Regression; Decision Tree; Random Forest; Feature Importance

1. Introduction

As an online market, e-commerce has completely overturned the traditional business relationship and brought unprecedented convenience and choice to users. Through online shopping, users can browse a large number of catalogs, compare product attributes, build a shopping wish list, and finally make purchase choices based on personal preferences. The historical interaction behavior of the users can reflect the demand of the users in the near future. At present, the competition in e-commerce market is very fierce, when the service demand of consumers cannot be satisfied, they can easily transfer to another e-commerce platform. Therefore, in order to improve the competitiveness of e-commerce platform, e-commerce platform needs to understand the user's portrait of the crowd, and through user behavior path optimization page, but also need to predict whether users will buy.

2. Analysis of Data pre-processing and explorations

2.1 Data Presentation

Data from an international e-commerce platform user data. This is an international e-commerce market covering China, the United States, Germany, the United Kingdom, users can use a mobile phone APP or browser to access the site for transactions. There are six data sets which includes user_table, listing_page, product_page, payment_page, payment_confirmation_page.

2.2 Data pre-processing

Data pre-processing is an essential part of task modeling. The quality of the training set directly determines the accuracy and generalization ability of the model. This section therefore focuses on the following processing of the dataset.

1) Missing value processing. Some of the users in the original data had missing values for age, gender, user origin, and operating system. For categorical attributes such as age and gender, the missing values are filled with the age and gender modes of all data samples. For user sources, the grouping shows that the old users are mainly from direct and the new users are mainly from SEO, so when the user is a new user, the missing values of source fill in the SEO. For user operating systems, the grouping shows that the mobile primary operating system is ios and the desktop primary operating system is windows, so

when the user is using desktop, the operating system fills in windows.

(2) Outlier handling. The analysis found that there were users in the data set who were over 100 years old, and removed this part of the data.

2.3 User portraits

User attributes include gender, age, new customer, user source, country, login device, operating system, number of pages viewed. This section analyzes the purchasing behavior of users from these dimensions, which is helpful for e-commerce platform to build user portrait and provide specific recommendations for different user groups. Through the analysis, we found that among the customers who placed orders, the turnover rate of female customers was higher than that of male customers, accounting for 80% ; the turnover rate of old customers was higher than that of new customers, accounting for 83% ; and the turnover rate of users gradually decreased with the growth of age, college students aged 17-25 accounted for the largest share of users. E-commerce platforms are more attractive to young people, mainly targeting 17-25-year-olds. The more pages a user views, the more likely they are to place an order on the e-commerce platform, which is the same as the reality. When the number of pages viewed exceeds 10, the conversion rate increases significantly with the number of pages visited. However, page views in the range of 4 to 7 have the largest number of users, indicating the need to continue to optimize the site. A further subdivision will find that Germany has the highest purchase rate of old users, but Germany accounts for the smallest proportion of users in the e-commerce platform, indicating that the German user quality is very high, the platform can continue to expand and dig into the German market.

2.4 Funnel model

From entering the platform to the final completion of the purchase, users will go through the home page, product list page, product details page, order settlement page, payment completion confirmation page these links, there is a certain amount of user churn on different pages. The conversion rate of home-> list pages was 0.73, list pages-> product pages was 0.68, product pages-> payment pages was 0.13, and payment pages-> payment completion pages was 0.33. On the device dimension, desktop had a 4% higher conversion rate from the payment page-> payment completion page than mobile. Mobile may have a hard nut to crack compared to desktop, and the payment button needs to be improved.

3. Model prediction

3.1 Logistic regression

Logistic regression is actually a categorical model, which essentially converts the linear regression result into a probability between 0 and 1 by a sigmoid function[1]. The probability is 1 when the probability p is greater than 0.5, when the probability p is less than 0.5, the prediction result is 0. The Sigmoid function is

$$p = \frac{1}{1+e^{-x}}$$

Using 80% of the data set as a training set and 20% as a test set. By using logistic regression model to predict whether users will place an order or not, there are 18080 users in the test set, 386 users actually place an order, and 263 users predict placing an order. The precision rate is 0.749, recall recall is 0.51 and F1 is 0.6.

3.2 Decision tree

Decision tree is a kind of tree structure, which consists of nodes and edges. In essence, the construction of decision tree model is mining effective classification rules and presenting them in the form of tree[2]. The construction of decision tree includes three parts: feature selection, tree generation and pruning. According to the location of nodes, the constructed decision tree can be divided into internal nodes and leaf nodes, in which the internal nodes are features or feature sets, and leaf nodes are the final classification results of model training.

The decision tree model was used to predict whether users would place an order or not. There were 18080 users in the test set, 386 users actually placed an order, and 273 users predicted placing an order. The precision was 0.723, rec recall is 0.52 and F1 is 0.6.

3.3 Random forest

Random Forest algorithm construction process: from the data provided by the random sampling of different subsets, used to build multiple different decision trees, and according to the Bagging rules to integrate the results of a single decision tree, they are classified according to the principle that the minority is subordinate to the majority[3].

The random forest model was used to predict whether users would place orders. There were 18,080 users in the test set, 386 users actually placed orders, and 267 users predicted whether they would place orders. The precision rate was 0.76, the recall rate is 0.52 for recall and 0.62 for F1.

3.4 Feature importance

Through the importance of features can know from which aspects to promote the transformation of users, thus targeted measures to users[4]. Through the ranking of the importance of features by logistic regression, the number of pages visited by users, their age and whether they are new users are ranked in the top three, the number of pages viewed by users, whether they are new users, and their age rank in the top three, while the number of pages viewed by users, whether they are new users, and their age rank in the top three by random feature importance ranking. According to the Order of three importance, the number of pages visited by users has the greatest impact on their orders.

4. Research and conclusions

Based on the data set of an international e-commerce platform, this paper conducts a deep-level exploration and research on the issue of future users placing orders. The following is a summary of the main work of this article.

(1) through the exploratory analysis of the data set, we can get the true distribution of the different attributes of the users, and dig out the common consumption habits and purchasing rules of the users.

(2) through the funnel model to explore the e-commerce user behavior path, from each link analysis user transformation.

(3) the best results are obtained by using machine learning model, logistic regression, decision tree and random forest to predict whether users will place an order or not, its accuracy, recall, and F1 values are the best-performing of the three models.

Through the analysis of the following conclusions and recommendations: first, the e-commerce platform in Germany under the highest rate of old users, combined with the current German users less total situation, it is suggested that targeted measures should be taken to encourage the conversion of users in Germany. Second, the more pages users browse, the more likely they are to place an order in e-commerce platform. At the same time, through modeling, we also know that the number of pages users browse has the greatest impact on their orders. Therefore, it is necessary to optimize the page design, to increase the number of pages users browse. Third, the conversion rate of Chinese users in each funnel is lower than that in other countries, presumably due to improper UI design or website translation problems. Fourth, the accuracy rate, recall rate and F 1 value of forecasting user's order by random forest are the highest, and the model effect is the best.

References

[1] Zhao TY, Liu SC, Xu J, et al. Comparative analysis of seven machine learning algorithms and five empirical models to estimate soil thermal conductivity[J]. Agricultural and Forest Meteorology, 2022, 323.

[2] Yan HW, Huo XR, Liu JJ. Research And Development of Suitable Aging Platform Based on Middle-Aged And Elderly Online Shopping Market Research[J]. International Journal of Education and Teaching Research, 2022, 3(2).

[3] Hou J, Li QM, Liu YZ, Zhang SN. An Enhanced Cascading Model for E-Commerce Consumer Credit Default Prediction[J]. Journal of Organizational and End User Computing (JOEUC), 2021, 33(6).

[4] Chang H, Yang SQ. Research on Commodity Mixed Recommendation Algorithm[J]. International Journal of Advanced Network, Monitoring and Controls, 2020, 5(3).