User online purchase behavior prediction based on fusion model of CatBoost and Logit

Wencho Cao¹,², Ke Wang¹, Hongcheng Gan¹ and Mingcheng Yang¹

¹Business School, University of Shanghai for Science and Technology, No.516, Junong Road, Yangpu District, Shanghai, China
²Corresponding author: qpee1464@163.com

Abstract. Analyzing users’ consumption preferences and willingness based on massive online behavior data can help optimize user experience and improve product purchase rate. For the online platform user data, feature engineering methods were used to preprocess, visualize, analyze and construct features. Then, a prediction and evaluation model of users’ online purchase behavior was developed based on the CatBoost algorithm, and the important features affecting users’ purchase behavior were quantitatively analyzed using the Logit Model. Compared with Random Forest, Support Vector Machine and XGBoost, the prediction model has better performance with prediction accuracy of 98.38% and F1 score of 0.775. The results show that user basic information, user access data and user login data have impact on user online purchase behavior. Especially the number of visits with or without coupons, unfinished orders of annual classes, number of coupons received, repeated learning of courses, public number following, days of login, city and number of learning classes sections are important features affecting purchase behavior.

1. Introduction
In the context of rapid development of e-commerce, online purchase brings great convenience to consumers and merchants, while effectively reducing costs of goods circulation and transaction. For merchants, obtaining consumers’ purchase intentions and analyzing their preferences and influencing factors can help them to provide consumers with personalized services and recommendations, increase market share and competitiveness, and promote the sustainable development of e-commerce platform. E-commerce platforms have plenty of user behavior data, such as click, visit, login and browsing data, etc. How to mine user preferences and purchase intentions based on user behavior data has become one of the key concerns of current researchers [1]. Recently, data-based users’ purchase behavior models have been developed, such as recommendation techniques, machine learning and integration algorithms [2-4].

Recommendation techniques construct similar product collections for recommendation based on user information characteristics and purchase preferences, including content-based recommendations, collaborative filtering, and hybrid recommendations [4]. Personalized recommendation technique is now an important research direction in the field of e-commerce recommendation because of the advantages in retaining users, increasing sales and saving costs [6]. However, user purchase behavior is affected by various factors such as online shopping platform, information collection ability, consumer psychology and consumption habits, etc. Recommendation techniques solve the user...
purchase prediction problem more from an algorithmic perspective to a certain extent, but may ignore the correlation between user purchase behavior and other information. Some statistical probability models or integrated models have achieved better performance for certain purchase behavior prediction problems, but there are still limitations for modeling nonlinear and high-dimensional data sets [7, 8].

With the development of data mining technology, more and more machine learning models are applied to the prediction of online purchase behavior. For example, some scholars have combined machine learning algorithms such as XGBoost [9], Deep Neural Network and Plain Bayes with user consumption behavior data to develop purchase behavior prediction models and achieved good results [10]. However, some machine learning algorithms do not deal well with the problem of category imbalance of actual purchase behavior data, and are also prone to overfitting or overgeneralization for sparse features of large dimensional data [10, 11]. Some machine learning algorithms such as BG/NBD and RMF have specific economic and market application contexts and limited applicability [12]. To solve the problem of data imbalance classification, Zhang et al. processed with sliding window and ICKMDS sample balancing algorithm to obtain balanced data and proposed a user purchase behavior prediction model FCVS based on balanced samples, feature engineering and integrated learning, and the F1 value of the model was significantly improved [13].

To effectively identify users' potential purchase behavior, Boosting algorithms are gradually being widely used. Boosting algorithm is an ensemble learning method that reduce bias of supervised learning and effectively improve the classification accuracy. And CatBoost algorithm, as an open source gradient enhancement algorithm, can effectively handle the classification features of large data sets, especially in the prediction of unbalanced data sets, and its performance is better than LightGBM and XGBoost algorithms [14-16]. This article combined user socioeconomic attributes and user online operation behavior data (including login and access), develop a user purchase behavior prediction and evaluation model based on CatBoost algorithm, and quantitatively analyzed the important feature factors affecting user online purchase behavior using Logit model.

2. Prediction and evaluation model

2.1. CatBoost

CatBoost is a gradient boosting library which is capable of handing categorical data. It was develo ped by Yandex company in 2017. CatBoost does not use binary substitution of categorical values, instead it performs a random permutation of the dataset and calculates the average label value for the example with same category value placed before the given one in the permutation[16, 17]. CatBoost reduces overfitting by using more effective strategies, and reduce the loss function in each iteration, which effectively improves the generalization ability of the model and realizes the full utilization of data information.

2.2. Evaluation Indicators

The evaluation metrics are important for testing the model’s prediction performance. Considering the possible category imbalanced characteristics of the user purchase behavior data, accuracy, precision, recall and F1 score were used to evaluate the performance of prediction models. F1score is the weighted summed average of the precision and recall rates, and truly reflect model’s prediction performance for category imbalanced data. The calculation formulas are respectively:

\[
\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

\[
F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]

where the precision and recall are calculated as follows:
where real types in the example and predicted type were combined and divided into true positive (TP), false positive (FP), true negative (TN), and false negative (FN) as shown in Table 1. TP is the number of samples where the model predicts that a purchase will occur correctly, FP and FN denote the number of samples where the model predicts that a purchase will occur incorrectly and the number of samples where it does not occur, respectively, as shown in the confusion matrix below.

| Table 1. Confusion Matrix. |
|---------------------------|
| Predicted                |
| Actual | Positive | Negative |
| True   | TP       | TN       |
| False  | FP       | FN       |

3. Online purchase behavior prediction

3.1. Data processing and analysis based on feature engineering

Data quality has a large impact on algorithm performance, so data needs to be processed before analysis and modeling. In this paper, the raw data are processed by feature engineering methods to improve the data quality and ensure the model performance.

The data processing based on feature engineering includes data collection, cleaning, exploratory analysis, feature construction and selection, where data cleaning mainly solves the problems of data noise, missing and redundancy; exploratory analysis mainly analyzes the data structure and prepares for feature construction; feature construction mines meaningful features through observation; feature selection aims to remove irrelevant features and improve model accuracy.

The original set of data contains: user basic information (user_info), user login data (login_day), user visit data (visit_info) and user purchase result (result), containing 135968, 135617, 135617 and 4639 user data respectively.

The data cleaning mainly checks the dataset for data duplication, missing, redundancy, noise and outliers, etc. The processing methods are as follow.

(1) For missing data, if there are few missing and they are continuous variables, we use attribute averages to fill them; if they are categorical variables, the dataset is filtered with the user id as the unique corresponding value.

(2) For noisy data, use statistical analysis to find noisy data and deal with it.

(3) For abnormal data, the actual situation is considered and filtered.

The specific processing methods are shown in Table 2. After data cleaning and data integration, there are 79581 valid data.

| Table 2. Pre-processing Methods for Relevant Variable. |
|-----------------------------|
| Variables | Treatment |
| user_id (users’id) | delete duplicated items |
| age_month (age) | delete values other than 10-80 |
| city_num (city) | delete missing values |
| login_day (number of days logged) | delete negative values |
| login_diff_time (login interval) | delete negative values |
| distance_day (last landing days to end of period) | delete negative values |

In order to explore the difference of users’ online purchasing behavior under different city
development levels, cities are divided into six categories according to GDP: first-tier cities (Q1), new first-tier cities (Quasi Q1), second-tier cities (Q2), third-tier cities (Q3), fourth-tier cities (Q4) and fifth-tier cities (Q5). Considering whether or not purchased and the frequency, we have counted the users separately for cities with different development levels. The results are shown in Figure 1.

![Figure 1. Number of users and purchasing behavior in cities at different levels of development distribution.](image)

The number of users in the new first-tier cities is the largest, and the number of users in the first-tier cities has the lowest percentage. Considering the proportion of users purchasing, the highest proportion of users purchasing is in first-tier cities.

Then, we visualized and analyzed some of the user login features, and selected login_day, login_diff_time, last login distance_day and login_time respectively. login_day is a count of the cumulative number of user login days, login_diff_time indicates the time interval between consecutive user logins to the platform, distance_day is the number of days between the last user login and the end of the course, and login_time indicates the length of user login time. The statistical results of the four features are shown in Table 3.

| login_day | login_diff_time | distance_day | login_time |
|-----------|-----------------|--------------|------------|
| Min       | 1               | 0            | 0          | 0          |
| Max       | 101             | 85.33        | 6588       | 1480       |
| Average   | 4.41            | 1.26         | 141.58     | 43.09      |
| Median    | 4.00            | 1.00         | 86.00      | 22.00      |
| SD        | 3.01            | 2.70         | 137.93     | 68.94      |

Figure 2 shows the trend of the distribution of the four characteristics with histogram of the frequency distribution.

The four login features have significant trend of concentrated distribution, i.e., login_day and login_diff_time are concentrated between 0-10, and login_time is mainly distributed between 0-200. Among them, distance_day has a significant decreasing trend in the interval of 0-200 and a sudden increase in frequency between 350-400.

Before feature selection, the dataset is divided into four feature dimensions: basic information, login information, user access, and purchase results. Based on the original variable fields of the feature dimensions, the relevant features are processed univariately, including dimensionlessization, continuous variable discretization, split-box operation, and coding, to achieve effective processing of the features.
Figure 2. Frequency Distribution of User Login Attributes.

Feature selection is mainly to reduce the number of model features by eliminating irrelevant variables, improve the generalization ability of the model, and reduce the risk of overfitting. According to the data features and the actual research, two features, platform_num and app_num, were selected to be excluded. Through the above operation, a total of 45 variables were identified, including category features city_num(city), platform_num(device), chinese_subscribe_num(follow public platform1), math_subscribe_num(follow public platform 2), add_friend(whether add sales friends), add_group(whether to join a group), study_num(whether repeat study), click_buy(whether click on the purchase button) and result(whether to purchase), and the rest are considered as continuous features. As shown in Table 4.

Table 4. List of variables and category characteristics.

| Feature            | Category features | Continuous feature                                                                 |
|--------------------|-------------------|------------------------------------------------------------------------------------|
| basic user information | city_num;platform_num | first_order_price;age_month;                                                       |
| chinese_subscribe_num; math_subscribe_num; add_friend; add_group; study_num; | login_day;login_diff_time;distance_day;login_time;launch_time;                     |
| login information  |                   | camp_num;learn_num;finish_num;coupon;course_order_num;                              |
| visit information  | click_buy;        | main_home;main_home2;mainPage;schoolReportPage;main_mime;lightCourseTab;main_learnPark;partnerGameBarrierPage;evaluationCenter;coupon_visit;progress_bar;ppt;task;video_play;video_read;Next_nize;Answer_task;Chapter_module;course_tab;slide_subscribe;buy_info;click_dialog |
| purchase results   |                   | result                                                                             |
3.2. Model result and evaluation

Python programming language was used to implement CatBoost algorithm, the main parameters are shown in Table 5.

| Main parameter   | Values |
|------------------|--------|
| iteration        | 300    |
| depth            | 5      |
| Learning_rate    | 0.05   |
| loss_function    | Logloss|
| subsample        | 0.7    |
| eta              | 0.5    |

Figure 3. gives the trend of loss function changes in the training and test sets during the model iteration, and the model reaches the optimum when Iteration is 298. CatBoost can score and rank the importance of data features during the modeling process, which helps to select the more valuable features for analysis. The importance of model features is shown in Figure 4.

In terms of the importance of model features, first_order_price (the price of the experience course), city_num, login_day (the number of days of login), distance_day (last landing days to end of period), chinese_subscribe_num (follow public platform1), learn_num (the number of study sessions), finish_num (the number of completed sessions), study_num (the number of repeat study), coupon (the number of coupons), course_order_num (the number of unfinished orders), coupon_visit (the number of visit with or without coupons), etc. have significant influence on whether users choose to purchase. The most important factor was found to be coupon followed by distance_day, which means that the number of coupon visits has a greater impact on whether to purchase or not, and the user's login status and class learning situation affect the user's purchase decision.

To explore the influence of the above important features on users' purchasing behavior, a binary logit model was used for quantitative analysis, and the results are shown in Table 6.

|              | Coef. | Std.Err. | P>|z| |
|--------------|-------|----------|-----|
| Intercept    | -3.379| 0.092    | 0.000|
| answer_task  | -0.016| 0.002    | 0.000|
| task         | 0.037 | 0.002    | 0.000|
| ppt          | -0.003| 0.001    | 0.015|
| coupon_visit | -0.486| 0.013    | 0.000|
Pseudo R2 is 0.399, which indicates that the model fits well. From the parameter calibration results, it can be seen that answer_task (the number of visit for answer resolution), ppt (the number of next ppt visit), coupon_visit (the number of visit with or without coupons), learn_num (the number of study sessions), camp_num (the number of open sessions), distance_day (last landing days to end of period), and login_day (the number of days of login) all have asignificant negative impact on user purchase behavior. With the increase of the number of visit to the answer resolution, visit to the next ppt, visit with or without coupons, and the number of courses studied, courses started and days of logged in, the high the users’ willingness to purchase. The more unfinished orders, the number of coupons received, the more repeat learning of the course in the current year. Users are also more likely to choose to purchase if they follow the public platform or live in cities with a high level of development.

The results of the quantitative analysis of the logit model tend to be consistent with feature importance of the variables, and there are differences in the magnitude of the feature importance and the variable coefficients, so there is no direct comparability between the variable coefficients and feature importance.

To evaluate the prediction performance of CatBoost algorithm, F1 score and Accuracy of CatBoost algorithm were calculated to compared with those of support vector machine, random forest and XGBoost algorithm. The model evaluation indexes are shown in Table 7.

F1 score can effectively evaluates the model performance using unbalanced data, and CatBoost algorithm has the greatest prediction accuracy and F1 score indicating that CatBoost algorithm has the best user purchase prediction performance compared to other algorithm.

**Table 7. Model predictive performance metrics.**

|        | F1 Score | Accuracy |
|--------|----------|----------|
| SVM    | 0.3519   | 0.9668   |
| Random Forest | 0.6437 | 0.9779   |
| XgBoost | 0.7643   | 0.9834   |
| CatBoost | 0.7753  | 0.9838   |

For binary classification problem, the selection of decision threshold directly affects the model accuracy. The ROC-AUC image can visually reflect the classification sensitivity accuracy trend of the model under different thresholds. AUC value is the area formed by the ROC curve and the x-axis, which can objectively evaluate the classification accuracy of the model. As presented in ROC-AUC curves of the four algorithms in Figure 5, CatBoost algorithm has the highest model classification
Figure 5. AUC-ROC curve of purchase prediction.

4. Conclusion
This paper takes online course learning platform user data as the research object, uses the feature engineering method to pre-process, visualize and analyze the data and construct feature, develops a model for predicting and evaluating users’ online purchase behavior based on CatBoost algorithm, and uses logistic model quantitatively analyze the important feature factors.

Compared to Support Vector Machines, Random Foresta, and XGBoost, CatBoost has a better performance and practicality with a prediction accuracy of 98.38% and F1 score of 0.7753. The AUC value of CatBoost is the largest (0.84), indicating that CatBoost has higher prediction sensitivity. The quantitative analysis of important characteristic factors suggest that, (1) an appropriate increase of the price of experiential courses may generate sunk costs and promote the purchase of courses. The higher the development level of users’ city, the greater the demand and users are more likely to make a purchase. (2) For login data, continuous login reduces users’ willingness to purchase; the promotion and maintenance of social platforms facilitate users’ purchase. (3) For access data, improving coupon distribution and optimizing the user access interface experience significantly increases users’ willingness to purchase.

5. References
[1] Wen, Y. T., Yeh, P. W., Tsai, T. H., Peng, W. C and Shuai, H. H. (2018, February). Customer purchase behavior prediction from payment datasets. In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining pp. 628-636.
[2] Kim, E., Kim, W and Lee, Y. 2003. Combination of multiple classifiers for the customer's purchase behavior prediction. Decision Support Systems, vol.34, no.2, pp.167-175.
[3] Helfert, M and Bezbradica, M. (2020, May). Customer purchase behavior prediction in e-commerce: A conceptual framework and research agenda. In New Frontiers in Mining Complex Patterns: 8th International Workshop, NFMCP 2019, Held in Conjunction with ECML-PKDD 2019, Würzburg, Germany, September 16, 2019, Revised Selected Papers (Vol.
Saranya, G., Gopinath, N., Geetha, G., Meenakshi, K and Nithya, M. (2020, December). Prediction of Customer Purchase Intention Using Linear Support Vector Machine in Digital Marketing. In *Journal of Physics: Conference Series* (Vol. 1712, No. 1, p. 012011). IOP Publishing.

Balabanović, M and Shoham, Y. (1997). Fab: content-based, collaborative recommendation. *Communications of the ACM*, vol.40, no.3, pp.66-72.

Lyu, Z., Dong, Y., Huo, C and Ren, W. (2020, April). Deep match to rank model for personalized click-through rate prediction. In *Proc. of the AAAI Conference on Artificial Intelligence*, vol.34, no.01, pp.156-163.

Kong, H., Lin, S., Wu, J and Shi, H. (2019, March). The Risk Prediction of Mobile User Tricking Account Overdraft Limit based on Fusion Model of Logistic and GBDT. In 2019 *IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)* (pp. 1012-1016). IEEE.

Peña-García, N., Gil-Saura, I., Rodríguez-Orejuela, A and Siqueira-Junior, J. R. (2020). Purchase intention and purchase behavior online: A cross-cultural approach. *Heliyon*, 6(6), e04284.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y and Cho, H. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2, 1(4)*, p.1-4.

Hu, X., Yang, Y., Chen, L and Zhu, S. (2020, May). Research on a Prediction Model of Online Shopping Behavior Based on Deep Forest Algorithm. In 2020 *3rd Int. Conf. on Artificial Intelligence and Big Data (ICAIBD)* (pp. 137-141). IEEE.

Hou, C., Chen, C and Wang, J. (2018). Tree-based feature transformation for purchase behavior prediction. *JEICE TRANSACTIONS on Information and Systems*, vol.101 no.5, pp.1441-1444.

Saranya, G., Gopinath, N., Geetha, G., Meenakshi, K and Nithya, M. (2020, December). Prediction of Customer Purchase Intention Using Linear Support Vector Machine in Digital Marketing. In *Journal of Physics: Conference Series* (Vol. 1712, No. 1, p. 012011). IOP Publishing.

Zhang, H and Dong, J. (2020). Application of sample balance-based multi-perspective feature ensemble learning for prediction of user purchasing behaviors on mobile wireless network platforms. *EURASIP Journal on Wireless Communications and Networking*, 2020(1), p.1-26.

Dou, X. (2020, April). Online purchase behavior prediction and analysis using ensemble learning. In *2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)* (pp. 532-536). IEEE.

Ibrahim, A., Muhammed, M. M., Sowole, S. O., Raheem, R and Abdulaziz, R. O. Performance of CatBoost classifier and other machine learning methods.

Jhaveri, S., Khedkar, I., Kantharia, Y and Jaswal, S. (2019, March). Success prediction using random forest, catboost, xgboost and adaboost for Kickstarter campaigns. In *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 1170-1173). IEEE.

Dorogush, Ershov, Gulin. “CatBoost: gradient boosting with categorical features support”. Yandex.