ISSN: 2688-9323

Online Purchase Forecast based Entirely on Behavior Analysis Using CatBoost

Lei Shi[1, a], Jian Ke[1] and Jiaming Zhu[2]

[1]School of Management Science and Engineering, Anhui University of Finance and Economics, Bengbu, Anhui 233030, China

[2]Institute of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu, Anhui 233030, China

[a]mrwonderful@163.com

Abstract

Purchase forecast is one of the research hotspots in the field of customer data mining for online shopping platforms. This paper proposes a purchase forecast model based entirely on customer behavior features. The model does not use any customer's personal information, constructs a feature system with 76 feature variables completely based on behavior data, and completes the empirical analysis with the ensemble learning algorithm CatBoost using the real behavior data of a large multi-category online store. The prediction accuracy of the model reaches 91.3%. The effectiveness and progressiveness of this method are verified by the comparative test results. In addition, the top 15 feature variables that have an impact on the prediction are listed, and the possible reasons for the impact of these variables are analyzed.

Keywords

Purchase Forecast; Customer Behavior Analysis; Machine Learning; CatBoost.

1. Introduction

Customer data mining is one of the research hotspots in the field of e-commerce. It is an effective means for e-commerce enterprises to understand customer value and enhance enterprise competitiveness. It is of great significance to e-commerce enterprises. In recent years, with the enhancement of people's awareness of network privacy protection, it has been difficult for e-commerce enterprises to directly use customers' personal information, such as age, gender, occupation, location, and so on. Therefore, customer data mining gradually turns to customer behavior analysis-oriented mining. A. M. Hughes first proposed RFM model in his own monograph to analyze customer behavior from three dimensions: Recency, Frequency, and Monetary [1]. However, there are too few variables in the RFM model to meet the needs of the era of big data.

For the online shopping platform, customer behavior mainly includes product browsing, adding to favorites, adding to cart, and purchasing, as well as the time, frequency, and purchase amount of these behaviors. It is a new challenge to forecast purchases based entirely on behavioral data without using any customer's personal information.

The purchase forecast in this study refers to whether the customer will complete the purchase of the commodity after the behavior of “add-to-cart” according to the historical behavior features of the customer. In essence, it belongs to a binary classification problem. Machine learning methods to solve classification problems include decision trees, neural networks, SVM, Bayesian classifiers, ensemble learning, etc. L. Tang proposed a firefly algorithm-based SVM model to predict whether or not a customer makes a purchase during the next visit to the online

ISSN: 2688-9323

store based on the previous behaviors [2]. L. S. Chen attempted to define the potential factors influencing in-App purchases for game users using two feature selection methods, Neural Network Pruning and Chi-square test [3].

In recent years, boosting algorithm has been widely used by promoting multiple weak learners to strong learners. The specific implementation of boosting algorithm includes AdaBoost, XGBoost, LightGBM, etc. The working mechanism of this family of algorithms is similar: first train a basic learner from the initial training set, and then adjust the distribution of training samples according to the performance of the basic learner, so that the training samples wrong by the previous basic learner will receive more attention in the follow-up, and then train the next basic learner based on the adjusted sample distribution; This is repeated until the number of base learners reaches the value specified in advance, and finally, these base learners are weighted and combined. X. Dou proposed an online purchase behavior prediction model using ensemble learning and further improved the prediction accuracy [4].

In all boosting algorithms, CatBoost can directly support category features. Considering that customer behavior in this study contains category features, such as the hour and the day of the week, so CatBoost is chosen as the base algorithm for the forecast model of this study.

2. Data and Feature Engineering

The data set in this study comes from a multi-category online store [5], which contains nearly 100 million customer behavior records involving more than 5 million customers in two months (from January to February 2020). Each row in the data set represents an event. All events are related to products and users. Each event is like a many-to-many relationship between products and users. Each row contains 8 fields, see Table 1 for details.

Table 1. Field Description of Raw Data

Column	Type	Description
event_time	datetime	the time of the event (accurate to seconds)
event_type	enumeration	including view, favor, cart, purchase
product_id	string	unique identifier of the product
category_id	string	unique identifier of the product category
brand	string	brand name of the product (if present)
price	float	product price
user_id	string	unique identifier of the customer (permanent)
user_session	string	identifier of the session in which the current event occurred

2.1. Descriptive Statistics

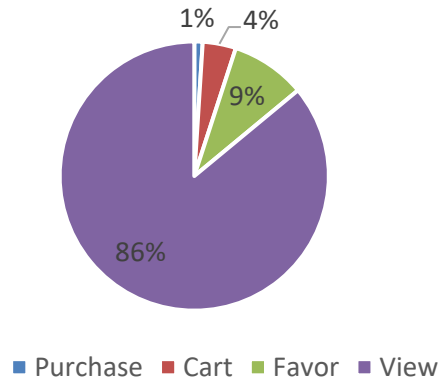

Figure 1. Proportion of Four Behaviors

ISSN: 2688-9323

The event type field in the data set contains four types of behavior: product browsing, adding to favorites, adding to cart, and purchasing, their respective proportions are shown in Figure 1. Although behaviors such as product browsing, adding to favorites, and adding to cart do not directly create value for e-commerce enterprises, these behaviors account for 99% of all behaviors and provide important data support for customer data mining, especially purchase forecast. Therefore, it is necessary to incorporate these behaviors into feature construction.

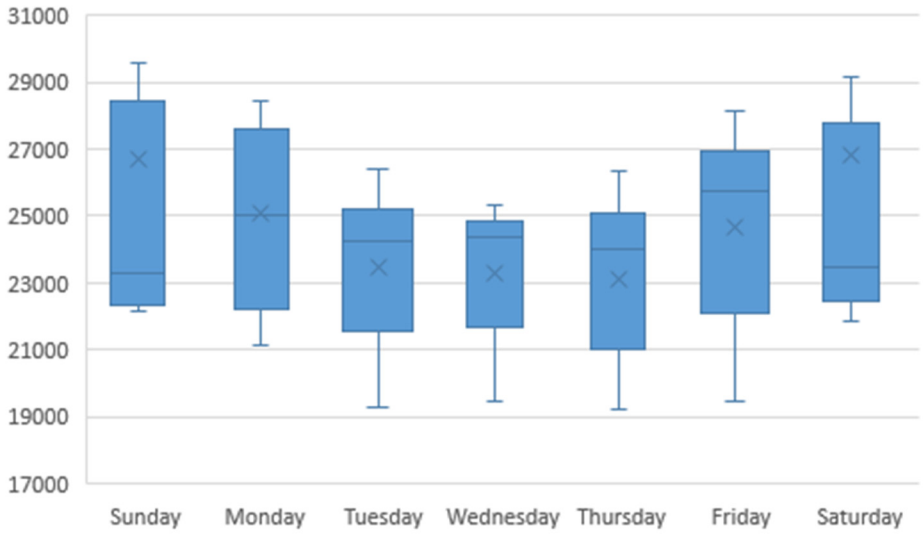

Figure 2. Daily Distribution of Purchase Behavior by Week

Figure 2 shows the daily distribution of customers' purchase behavior by week in the form of box plot. It can be seen that the purchase behavior of customers has obvious rules in time distribution. Within a week, the number of purchase behaviors on weekends was significantly higher than that on weekdays; Followed by Friday and Monday, because these two days are adjacent to the weekend; The number of purchase behaviors from Tuesday to Thursday is the least. The distribution of purchase time in one single day is further analyzed, and the results show that purchase behavior mainly occurs between 8 a.m. and 8 p.m. Therefore, it is necessary to consider the time factor of behavior in the purchase forecast. [6]

2.2. Feature Engineering

The original data cannot be directly used for purchase forecasting, so it is necessary to build the interactive features of the "user-product" based on the original data, so as to reflect the information of the original data from different angles. Good feature construction plays an important role in the model establishment, training and prediction.

To predict whether customers will purchase, it needs to start from the aspects of customers, products, categories, and brands, and first build independent customer behavior features, product behavior features, category behavior features, and brand behavior features. These features include the respective counts of product browsing, adding to favorites, adding to cart, and purchasing, as well as the conversion rate, that is, the ratio of purchase count to browsing count. Then consider the interactive features of customers and products, such as customer-product pair behavior features, customer-category pair behavior features, and customer-brand pair behavior features. The time span of the above features needs to be considered when constructing, that is, the features need to reflect the information of different time spans such as short-term, medium-term, and long-term. Due to the high timeliness requirements of purchase forecast tasks, the above features are divided into three time spans: the last day, the last week,

ISSN: 2688-9323

and the last month. In addition, we should also consider the features of the current "add-to-cart" behavior, such as the time (including the hour and the day of the week), the number of behaviors in the current session, and the price of the product.

In order to more objectively reflect the forecast effect of the model, this study will forecast the purchase behavior in February based on the data in January. Therefore, all behavior feature variables are calculated based on January data only. Finally, a feature system containing 76 behavior feature variables is obtained, as shown in Table 2.

Due to the dimensions of these feature variables are not uniform, it is necessary to deal with the eigenvalues dimensionless in order to avoid affecting the feature weight. In this study, min-max normalization is adopted, as shown in Formula (1).

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

Table 2. Behavior Feature Variables

Feature categories	Features	Variable count
Customer behavior features	the separate count of each behavior of the customer in 1 day, 1 week, and 1 month before the survey	12
	the customer's conversion rate in 1 day, 1 week, and 1 month before the survey	3
	the average purchase amount of the customer in 1 day, 1 week, and 1 month before the survey	3
Product behavior features	the separate count of each behavior of the product in 1 day, 1 week, and 1 month before the survey	12
	the conversion rate of the product in 1 day, 1 week, and 1 month before the survey	3
Category behavior features	the separate count of each behavior of the product category in 1 day, 1 week, and 1 month before the survey	12
	the conversion rate of the product category in 1 day, 1 week, and 1 month before the survey	3
Brand behavior features	the separate count of each behavior of the brand in 1 day, 1 week, and 1 month before the survey	12
	the conversion rate of the brand in 1 day, 1 week, and 1 month before the survey	3
Interactive features	the total count of all behaviors of the "customer-product" pair in 1 day, 1 week, and 1 month before the survey	3
	the total count of all behaviors of the "customer-category" pair in 1 day, 1 week, and 1 month before the survey	3
	the total count of all behaviors of the "customer-brand" pair in 1 day, 1 week, and 1 month before the survey	3
Features of the current "add-to-cart" behavior	the hour when the "add-to-cart" behavior occurred	1
	the day of the week when the "add-to-cart" behavior occurred	1
	the count of behaviors in this session	1
	price of the product	1
Total		76

ISSN: 2688-9323

3. Model

CatBoost, also known as categorical boosting, is an open-source ensemble learning algorithm framework based on gradient boosting developed by Yandex in 2017 [7]. Similar to XGBoost and LightGBM, CatBoost is also an improved implementation under the framework of GBDT (Gradient Boost Decision Tree) algorithm. CatBoost can be used for both regression and classification problems. The main feature of CatBoost is that it can handle category features efficiently and reasonably, which can be noticed from its name.

The reasons for choosing CatBoost model are as follows:

a. CatBoost embeds an innovative algorithm to automatically process category features into numerical features. It first counts the category features, calculates the frequency of a category feature, and then adds super parameters to generate new numerical features. In this study, the hour and the day of the week when customers' "add-to-cart" behavior occurred belong to category features, which are very suitable to be processed with CatBoost model.
b. It uses the ordered boosting method to combat the noises in the training set, so as to avoid the problems of gradient bias, and then solve the problem of prediction shift, so as to reduce the occurrence of over fitting and improve the accuracy and generalization ability of the algorithm.
c. It implements oblivious decision trees (binary tree in which the same features are used to make the left and right split for each level of the tree) thereby restricting the features split per level to one, which helps in decreasing prediction time.
d. It also uses combined category features, which can take advantage of the relationship between features, which greatly enriches the feature dimension.
e. It has effective usage with default parameters thereby reducing the time needed for parameter tuning.
f. It has certain extensibility to support user-defined loss function.
g. It has excellent performance and supports GPU acceleration.
h. It is easy to use with packages in R and Python.

CatBoost model also has some shortcomings, for example, the processing of category features requires a lot of memory and time, and the setting of different random numbers has a certain impact on the prediction results of the model. The loss function and scoring function of CatBoost are shown in formulas (2) and (3).

$$CrossEntropy = -\sum_{i=1}^{N} w_i (t_i \log(p_i) + (1 - t_i) \log(1 - p_i)) / \sum_{i=1}^{N} w_i \tag{2}$$

$$Score' = Score \cdot \prod_{f \in S} W_f - \sum_{f \in S} P_f \cdot U(f) - \sum_{f \in S} \sum_{x \in L} EP_f \cdot U(f, x) \tag{3}$$

4. Experiment

4.1. Experiment Design

The main goal of the experiment is to forecast the purchase behavior in February based on the behavior data in January. Considering that the customers, products and other entities involved in these two months are not exactly the same, first filter out the records of those entities that both exist in January and February. The behavior feature system is constructed using January's data, including customers, products, categories, brands, and interactive behavior features. Then, the records with "add-to-cart" behavior are filtered out from these records. If the customer later purchased the product, mark this record as a positive example, and mark it as a negative

ISSN: 2688-9323

example if he didn't buy it. Then, 10000 samples are randomly sampled from the records in January as the training set of the model, including 5000 positive examples and 5000 negative examples, in order to maintain sample balance. 2000 samples are divided from the training set according to the ratio of 8:2 as the verification set. Lastly, 2000 samples are sampled from the records in February as the test set of the model using the same method.

To verify the effectiveness and advancement of the model, four sets of comparative experiments were designed, including SVM model, BP neural network model, XGBoost model, and CatBoost model. The main parameters of CatBoost model are shown in Table 3.

Table 3. Parameter Setting of CatBoost Model

Parameter	Value
eval_metric	AUC
task_type	GPU
learning_rate	0.01
iterations	10000
depth	8
l2_leaf_reg	10
early_stopping_rounds	500

4.2. Evaluation Criteria

The precision (*P*), recall (*R*), *F*1 measure, and *accuracy* are used as the evaluation criteria of the model. See formulas (4) to (7) for their calculation formulas.

$$P = \frac{TP}{TP + FP} \tag{4}$$

$$R = \frac{TP}{TP + FN} \tag{5}$$

$$F1 = \frac{2 * P * R}{P + R} \tag{6}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

TP stands for true positive, *TN* stands for true negative, *FP* stands for false positive, and *FN* stands for false negative.

4.3. Experiment Results

Table 4. Results of Comparative Experiments

Model	*P* (%)	*R* (%)	*F*1 (%)	*Accuracy*
SVM	78.62	69.53	73.80	75.48
BP neural networks	80.07	86.39	83.11	84.23
XGBoost	85.51	90.77	88.06	89.43
CatBoost	87.25	89.32	90.18	91.30

As can be seen from Table 4, the forecast effect of CatBoost model is the best, thanks to its good support for category features; The XGBoost model, which is also an ensemble learning model,

ISSN: 2688-9323

follows closely, and has good performance; The forecast effect of BP neural networks model is slightly weak; The worst performance is SVM model, it seems that SVM has been unable to solve the classification problem with too many feature variables.

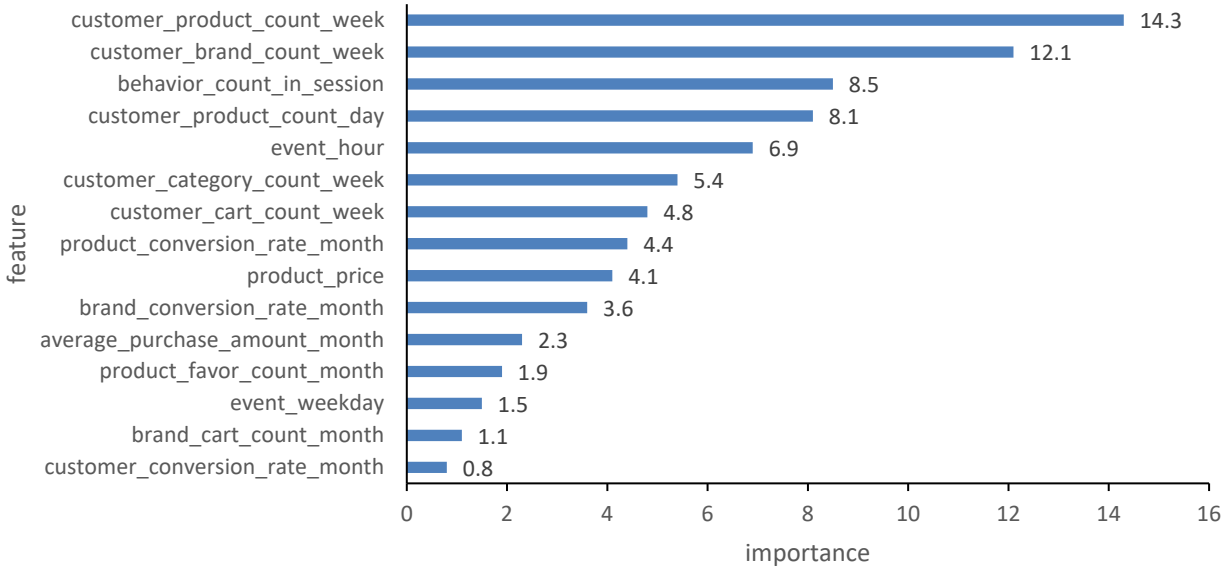

Figure 3. Top 15 importance characteristic variables

Figure 3 shows the importance of the top 15 feature variables in the CatBoost model to the forecast results. Although there are 76 feature variables in total, the top 15 feature variables have nearly 80% impact on the model, which basically determines the forecast results of the model. It can be seen from the figure that among the features with high importance ranking, four features belong to interactive features, because they directly reflect the interactive relationship between customers and commodities, customers and brands. The number of behaviors in the current session is also of high importance. The possible reason is that the more customers' behaviors in a session, the more likely they are to purchase. Among the time features, the importance of event hour is significantly higher than event day. The possible explanation is that people's shopping hour in the day is likely to be fixed, such as lunch break or before going to sleep at night, while people's shopping on the day of the week is not too fixed. It is also found that the feature variables of monthly span are more than that of weekly span, while there are almost no daily span variables. Because any forecast model needs the support of historical data, generally speaking, the larger the time span, the greater the impact of historical data on the forecast. It is worth noting that in the behavior category, the characteristic variables related to browsing behavior do not appear in the top 15. This is not because browsing behavior is not important, but it works indirectly through the conversion rate variable. Because the base of conversion rate is the count of browsing behavior. It is worth noting that the feature variables related to product browsing behavior do not appear in the top 15. This is not because browsing behavior is not important, but it plays an indirect role through the variable of conversion rate, and the base of conversion rate is the count of browsing behaviors.

5. Conclusion

This study confirms that customers' purchase behavior can be accurately predicted based entirely on behavior data. With the support of CatBoost model, the accuracy of the forecast model can reach 91.3%. Through the analysis of the model, 15 key behavior features affecting

customers' purchase behavior are found, which provides an important reference for the CRM of e-commerce enterprises. Whether we can further improve the accuracy of the forecast depends on Feature Engineering, that is, whether we can build a more perfect behavior feature system. The maximum time span of the behavior features in this study is one month. A larger time span may bring higher forecast accuracy, but it will also bring a greater amount of data and greater computational overhead. Therefore, researchers need to grasp the distance between theoretical research and practical operation to promote the further integration of industry, university, and research.

Acknowledgments

The work is supported by the university social science research project funded by the Department of education of Anhui Province, China. (Project No.:SK2021A0235)

References

[1] A. M. Hughes: Strategic database marketing: the masterplan for starting and managing a profitable, customer-based marketing program, volume 12. (McGraw-Hill New York, U.S. 1994).

[2] L. Tang, A. Wang, Z. Xu, et al: Online-purchasing behavior forecasting with a firefly algorithm-based SVM model considering shopping cart use, Eurasia Journal of Mathematics Science and Technology Education, Vol.13 (2007) No.12.

[3] L. S. Chen, M. R. Lin, Y. T. Pan: Find crucial factors of in-game purchase using neural networks, 8th International Conference on Awareness Science and Technology (iCAST) (2017).

[4] X. Dou: Online Purchase Behavior Prediction and Analysis Using Ensemble Learning, IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA) (2020).

[5] Information on https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store.

[6] L. Yong, Y. Zhuang: Research model of churn prediction based on customer segmentation and misclassification cost in the context of big data, Journal of Computer & Communications, Vol.03 (2015) No.6, p.87-93.

[7] L. Ostroumova, G. Gusev, A. Vorobev, et al: CatBoost: unbiased boosting with categorical features, Neural Information Processing Systems (NeurIPS) (2017).