Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 5, No. 2, 2023

XGBoost Analysis based on Consumer Behavior

Kunpeng Cai, Maria Rosario Rodavia *

Graduate School, Angeles University Foundation, Angeles City, Philippines
* **Corresponding author**: Maria Rosario Rodavia (Email: rose.rodavia@gmail.com)

Abstract: With the rapid development of the Internet and e-commerce, a large amount of consumer data has become available, which allows us to better understand consumer preferences and purchasing trends. XGBoost (Extreme Gradient Boosting) is a powerful machine learning algorithm that performs well when dealing with large data sets and complex features. This article first introduces the basic principles of the XGBoost algorithm, and then discusses in detail how to apply it to consumer behavior analysis. This paper uses real consumer data set, including multi-dimensional information such as customer identification, product identification and customer purchasing behavior. By building the XGBoost model, we are able to identify important features, predict consumer purchase intentions, and provide personalized recommendations. In addition, the performance evaluation and optimization methods of the model are discussed to ensure its accuracy and practicability. Finally, we summarize the main findings of this study, highlighting the potential applications of XGBoost analytics based on consumer behavior in marketing and business decisions. By digging deeper into consumer behavior data, businesses can better meet customer needs, improve sales efficiency, and achieve sustainable competitive advantage, and this research provides strong support for the use of machine learning techniques to optimize market strategies.

Keywords: XGBoost; Consumer Behavior; Marketing; Business Decision.

1. Introduction

With the rapid development of the Internet and e-commerce, consumer behavior analysis has become a crucial part of marketing and business decisions. Instead of relying solely on traditional market research methods, businesses and organizations are actively adopting machine learning and data science techniques to better understand and predict consumer behavior. Among them, XGBoost (Extreme Gradient Boosting) algorithm, as a powerful machine learning tool, is widely used in the field of consumer behavior analysis to achieve more accurate consumer insight and personalized market strategies. This article aims to take an in-depth look at XGBoost analytics based on consumer behavior and its potential applications in marketing and business. First, we will briefly introduce the basic principles and features of the XGBoost algorithm, as well as its wide application in the field of machine learning. We will then explore in detail how XGBoost algorithms can be applied to consumer behavior analysis, including key steps such as data preparation, feature engineering, model building, and performance evaluation. Through a real-world case study, we will demonstrate how the XGBoost model can be used to predict consumer purchase intentions and provide personalized product recommendations. Finally, we will discuss the potential benefits of this approach and the direction of future research in this area. Chen, T. et al. introduced the XGBoost algorithm [1], which is a gradient lifting tree algorithm used to solve various machine learning problems. It excels in performance and accuracy and is widely used for data mining and prediction tasks. For the problem of how to use XGBoost algorithm to predict click-through rate in e-commerce, Li, H., & Guo, J. Describes and illustrates the application of XGBoost in real business such as AD click prediction [2]. Tong, J. et al. have focused on using gradient lift machines (including XGBoost) to predict customer purchase behavior in e-commerce, and this article provides a case study of using XGBoost in an e-commerce context, emphasizing its predictive power [3]. Rashid, T. et al. focused on using XGBoost to predict customer churn in the telecom industry [4]. Customer churn prediction is crucial for maintaining customer loyalty, and the application of XGBoost in this field is expected to improve the accuracy of prediction. Jia, X. Et al. described a customer behavior analysis and prediction model based on XGBoost algorithm [5], which highlighted the application of XGBoost in analyzing customer behavior and predicting trends, especially in the fields of management and education. Sanchez-Morales et al. studied the use of XGBoost to predict customer churn in the mobile telecom industry [6]. Customer churn is an important issue for telecom companies, and XGBoost is used to improve the prediction accuracy of customer churn. Chen, Y. et al. [7] introduce a hybrid model that combines XGBoost and Short-term memory networks (LSTM) to predict customer churn in e-commerce. Sun, X. et al., explore the use of XGBoost in e-commerce to predict customer churn [8], which highlights the practical application of data science and machine learning in e-commerce to improve customer retention and sales efficiency. Through the analysis of customer behavior data, Yang, W. introduced how to use XGBoost to identify different customer groups [9] and provide customized recommendations and promotions based on their needs and preferences. Wu, H. et al., by combining XGBoost with deep learning methods [10], demonstrated how to better capture complex sales trends and seasonal changes, thereby improving the performance of sales forecasting, which is very important for supply chain management and inventory optimization. The innovation of this paper is that XGBoost algorithm is used for the first time to build a consumer prediction model, and has achieved good results in the evaluation of indicators. The experiment proves that XGBoost algorithm model will have great application scenarios and practical significance in e-commerce applications.

2. Research Basis

2.1. Algorithm Overview

XGBoost is an ensemble learning algorithm that combines multiple weak learners (usually decision trees) into a powerful model. Its core idea is to train weak learners iteratively and then combine them into a more powerful model to improve the accuracy of predictions. Based on the idea of Boosting trees, where each Tree is trained on the mistakes of the previous tree. XGBoost uses a custom loss function, usually one that combines prediction error and model complexity. It includes regularization terms to limit the complexity of the tree and prevent overfitting. It uses L1 and L2 regularization to control the complexity of the model to improve generalization, and employs an Early Stopping strategy to avoid overfitting during training. By monitoring the performance of the validated data, training can be stopped before the model starts to overfit.

2.2. Principle of Algorithm

Suppose the problem we are trying to solve is a regression problem, and the goal is to learn a prediction function f(x) where x represents the input features and f(x) represents the predicted value of the output. Our goal is to minimize the loss function, usually using the Mean Squared Error to measure the difference between the predicted value and the true value:

Loss function:

$$L(y, f(x)) = (y - f(x))^2 \quad (1)$$

Where y represents the true target value.

XGBoost uses decision trees as a weak learner, and each decision tree is a regression tree. The goal of the regression tree is to fit the residual of the training data (that is, the difference between the true value and the predicted value of the current model). The key idea here is that in each iteration, we train a new regression tree, which is then added to the model to gradually reduce the residual. The goal of XGBoost is to minimize the following weighted loss functions:

Objective function:

$$Obj(\Theta) = \Sigma[LF\ y\_i, \hat{y}\_i) + RT(\Theta\_t)] \quad (2)$$

Among them,

- Θ Indicates the parameters of the model, including the structure of the tree and the value of the leaf node.
- i represents each training sample.
- Enclosure ŷ_i indicates the predicted value of the current model.
- The loss function is used to measure how well the model fits the training data.
- Regularization terms are used to control the complexity of the model, including L1 and L2 regularization terms.

XGBoost uses a gradient lift method to optimize the objective function. In each iteration, it calculates the negative gradient of the objective function and then fits this negative gradient with a new regression tree. That's why it's called Gradient Boosting.

Where, the negative gradient:

$$g\_i = \partial L(y\_i, \hat{y}\_i) / \partial \hat{y}\_i \quad (3)$$

$$H\_i = \partial^2 L(y\_i, \hat{y}\_i) / \partial \hat{y}\_i^2 \quad (4)$$

Where g_i represents the gradient of the loss function with respect to the predicted value, and H_i represents the second derivative of the loss function. Specific steps:

a. Initialize the model: To begin, initialize a simple model as the initial predicted value, usually a constant, such as the average of all the training samples.
b. Iterative training: Repeat the following steps:
    a) - Calculate the gradients and second derivatives of the current model.
    b) - Train a new regression tree to fit the negative gradient to reduce the loss function.
    c) - Use linear search to find the best tree structure (split point and leaf node values).
    d) - Update model parameters, including the structure of the tree and the values of the leaf nodes.
c. Regularization: After each iteration, regularization terms are applied to control the complexity of the model.
d. Stop early: We can monitor the performance of the validated data and stop training when the performance is no longer improving to avoid overfitting.

2.3. Experimental Object

The overall goal is to classify potential consumers according to different characteristics, and further data analysis on the basis of classification, on this basis, we use xgboost algorithm to classify consumers.

- Get some clean data through data preprocessing.
- Group consumers with different characteristic values.
- In the classification results, and specific analysis of the results.

3. Methods

First of all, the data is obtained from Taobao official, and then some characteristics and dimensions of the data set are learned by reading the attributes of the data set. Secondly, the data set is cleaned and preprocessed. Some null values and indexes will be cleaned, and the cleaned data will be saved as a local CSV file. After data processing is completed, xgboost algorithm is introduced, and different models are drawn according to different eigenvalues, and specific analysis is made according to these specific models. The results show that the xgboost algorithm can effectively classify consumers. The data comes from Taobao Alibaba Tianchi data set, the official name is "UserBehavior.csv", the data set has 5 variables. The five different variables are user_id, item_id, category_id, behavior, and timestamp. user_id is the number of each consumer, numbered in Arabic numerals; item_id indicates the item ID (Arabic digit). category_id refers to a category of merchandise. behavior refers to customer consumption behavior; timestamp refers to the time spent by the customer.

4. Experiment and Verification

The experiment system is Win10, python language environment, using anaconda software. For the convenience of subsequent statistical data, the original data must be preprocessed first, null and duplicate values removed, and the timestamp converted to the correct time format. The result after processing is shown in Figure 1.

	user_id	item_id	category_id	behavior	timestamp	datetime	date	hour
0	1	2268318	2520377	pv	1511544070	2017-11-25 01:21:10	2017-11-25	01
1	1	2333346	2520771	pv	1511561733	2017-11-25 06:15:33	2017-11-25	06
2	1	2576651	149192	pv	1511572885	2017-11-25 09:21:25	2017-11-25	09
3	1	3830808	4181361	pv	1511593493	2017-11-25 15:04:53	2017-11-25	15
4	1	4365585	2520377	pv	1511596146	2017-11-25 15:49:06	2017-11-25	15

Figure 1. Correct timestamp format

There are four kinds of consumer behaviors, including click (pv), buy (buy), add to cart (cart) and like (fav). After

preliminary statistics, it is found that most behaviors of consumers are click behaviors, and there are few samples of purchase behaviors, and the proportion of positive and negative samples is seriously unbalanced. In machine learning models, unbalanced ratios of positive and negative samples can lead to overfitting of training results, which means that the prediction results may be more accurate in the category with more data. In order to improve the prediction accuracy, it is necessary to delete some special data: 1. Impulse consumption or malicious brushing suspected data, that is, consumers make purchases when there is no clicking behavior on the product. Such consumer behaviors belong to impulse consumption behaviors without considering the brushing behavior. 2 Low frequency single operation of a single product, the consumer has only one click on a category of goods, that is, the consumer only views one of the goods of this category, and the number of operations is too few, limited to browsing behavior, and there is no other behavior to add to the cart or like to show the consumer's interest in the product, the number of behaviors is too small. It is difficult to judge consumer behavior habits through such behaviors. 3. Low-frequency single operation of multiple goods, view of multiple goods under different categories, the number of behaviors are too small, and the type of behavior is limited to click behavior, and there is no discernable behavior of liking goods and adding goods to the cart. The result after processing is shown in Figure 2:

	user_id	item_id	category_id	behavior	timestamp	datetime	date	hour
0	1	2268318	2520377	pv	1511544070	2017-11-25 01:21:10	2017-11-25	1
1	1	2333346	2520771	pv	1511561733	2017-11-25 06:15:33	2017-11-25	6
2	1	2576651	149192	pv	1511572885	2017-11-25 09:21:25	2017-11-25	9
4	1	4365585	2520377	pv	1511596146	2017-11-25 15:49:06	2017-11-25	15
6	1	230380	411153	pv	1511644942	2017-11-26 05:22:22	2017-11-26	5

Figure 2. Preliminary data cleaning

user_id	0	1	2	3	4	5	6	7	8	9
1	0	1	1	1	0	1	1	1	1	1
13	0	0	0	0	0	0	0	0	0	1
19	0	0	0	0	0	0	1	0	1	1
21	0	0	0	1	1	1	1	1	1	1
100	0	0	0	2	0	0	2	1	1	1
...	...	...	...	...	...	...	...	...	...	...
1017990	0	1	0	0	0	0	0	0	0	0
1017994	0	0	0	0	0	0	0	0	2	2
1017997	1	1	0	0	0	0	0	0	0	2
1018000	0	0	0	0	0	0	0	1	1	1
1018011	0	0	0	0	0	0	0	0	1	0

38840 rows × 10 columns

Figure 3. Sequence length 10 schematic diagram

After data statistics, it is found that the number of behaviors generated in the process of consumer activities is different, and there may be differences in consumer behaviors under different operation times. The next step is to layer the data, which uses strings for recording consumer behavior, which is not easy to train in xgboost. Therefore, after analyzing the sequence of consumer behaviors, the three behaviors are defined as integer 1, 2 and 3 respectively, and the purchase behavior is separately defined as integer 4 according to the order of clicking behavior, liking behavior and adding to shopping cart behavior, that is, by analyzing the degree of interest in the products shown by different behaviors of consumers. Suppose we initialize a 10-bit vector at this point, that is, the consumer behavior sequence length is 10, and the execution result is shown in Figure 3.

In the above, we defined consumer behavior as integers 1, 2, 3 and 4. Next, in order to better integrate the data into the prediction model, this paper will mark data 0 and 1 for each user according to whether there is a purchase behavior, that is, if the consumer has a purchase behavior, the data will be marked as 1, otherwise it will be 0. The marked data is shown in Figure 4:

user_id	0	1	2	3	4	5	6	7	8	9	label
1	0	1	1	1	0	1	1	1	1	1	0
13	0	0	0	0	0	0	0	0	0	1	0
19	0	0	0	0	0	0	1	0	1	1	0
21	0	0	0	1	1	1	1	1	1	1	0
100	0	0	0	2	0	0	2	1	1	1	1
...	...	...	...	...	...	...	...	...	...	...	...
1017990	0	1	0	0	0	0	0	0	0	0	0
1017994	0	0	0	0	0	0	0	0	2	2	0
1017997	1	1	0	0	0	0	0	0	0	2	1
1018000	0	0	0	0	0	0	0	1	1	1	0
1018011	0	0	0	0	0	0	0	0	1	0	1

38840 rows × 11 columns

Figure 4. Behavior marker graph

If we start training the data at this point, it is clear that the data has too few feature points, and we need to build a feature project for this data. Taking the data in this study as an example, the basic features include the ID of the consumer, the ID of the product category, the ID of the product, the category of the behavior, and the timestamp of the behavior. The goal of the study is to predict whether a consumer will buy a certain product, so it is necessary to analyze which characteristics will affect the probability of a consumer's purchase. In order to achieve this goal, the study summarized the degree of influence of different characteristics on purchasing behavior by analyzing the influence of various characteristics. Finally, the research decides to carry out feature statistics and extraction from three aspects, which are consumer characteristics, commodity characteristics and interaction characteristics. In order to better understand consumer behavior patterns, the study conducted a secondary treatment of consumer behavior according to different time spans. This means that the study considers consumer behavior over different time periods to obtain data related to the time interval. This operation not only increases the dimension of features, but also increases the complexity of data processing. Consumer characteristics are the description of consumers' behavior patterns for commodities, so as to capture consumers' personalized behavior rules and thus reveal individual characteristics. This paper mainly statistics consumers' behavior information characteristics. Product characteristics study the influence of product characteristics on consumer behavior from the perspective of product, and understand the popularity of the product in the same period from the behavior of the product being clicked, liked, added to the shopping cart and liked, so as to reflect the degree of favor and popularity of the product among similar consumers.

This paper mainly calculates the conversion rate of the product from non-purchase behavior to purchase behavior. Interactive characteristics refer to the various behaviors of consumers for a particular product. This paper analyzes the conversion rate of consumers' purchase with various time intervals. After establishing the feature process, part of the results after data processing are shown in Figure 5:

user_id	0	1	2	3	4	5	6	7	8	9	...	afternoon_actions	evening_actions	night_total_ra
1	0	1	1	1	0	1	1	1	1	1	...	25	0.0	0.5000
13	0	0	0	0	0	0	0	0	0	1	...	4	0.0	0.0000
19	0	0	0	0	0	0	1	0	1	1	...	10	0.0	0.0000
21	0	0	0	1	1	1	1	1	1	1	...	22	0.0	0.4285
100	0	0	0	2	0	0	2	1	1	1	...	21	0.0	0.2857
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1017990	0	1	0	0	0	0	0	0	0	0	...	4	0.0	1.0000
1017994	0	0	0	0	0	0	0	0	2	2	...	11	0.0	0.0000
1017997	1	1	0	0	0	0	0	0	0	2	...	13	0.0	0.5000
1018000	0	0	0	0	0	0	0	0	1	1	...	10	0.0	0.0000
1018011	0	0	0	0	0	0	0	0	0	1	...	5	0.0	0.0000

38840 rows × 25 columns

Figure 5. Feature engineering coding map

Next, we need to deal with the problem mentioned above, that is, the serious imbalance in the proportion of positive and negative samples. Although we have deleted some special data in the above, the amount of positive and negative samples is still unbalanced. After statistics, it is found that the number of positive samples is about 12,459, and the number of negative samples is about 26,381. SMOTE was oversampling in this paper, and after this step, the number of positive and negative samples reached 26,381. The pre-processed data set is put into the prediction model for the experiment. Before the experiment, the model evaluation index used in this paper needs to be introduced. Since the problem studied in this paper is a binary classification problem, in order to better describe the classification effect of the model, confusion matrix and ROC curve will be used to describe the classification results in the analysis, so as to evaluate the accuracy and other classification effects of the model. Through the confusion matrix, this paper needs to calculate its secondary indexes Accuracy, Precision and Recall, tertiary indexes F1 and ROC curve area AUC. After model training, the results of secondary indexes are shown in Figure 6, and the ROC curve is shown in Figure 7.

XGBoost Model:
Accuracy: 0.59, Precision: 0.62, Recall: 0.51, F1: 0.56, AUC: 0.59

Figure 6. Secondary index

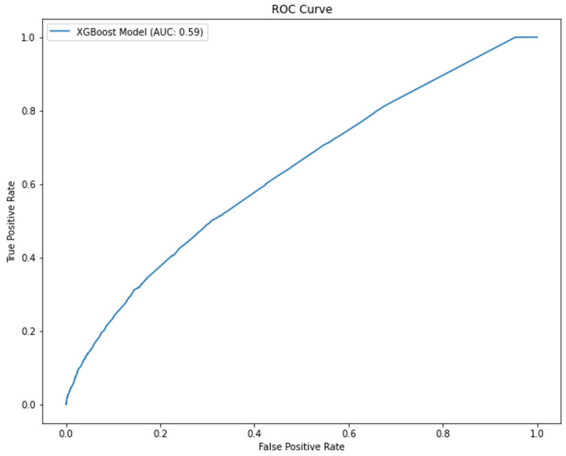

Figure 7. ROC Curve

It is not difficult to see from FIG. 6 and FIG. 7 that the prediction result of the model is not good, but the accuracy of the prediction is a little higher than that of the random model. As mentioned above, when the sequence length is 10, the AUC value is 0.59. Starting from just below 10 (say, 7) and increasing the length of the consumer behavior sequence, we can find the maximum AUC value. In this paper, the consumer behavior sequence value was set from 7 to 30, and after a series of uninterrupted experiments, all secondary and tertiary indicators and ROC curve area AUC were finally obtained, and the curve was generated as shown in Figure 8.

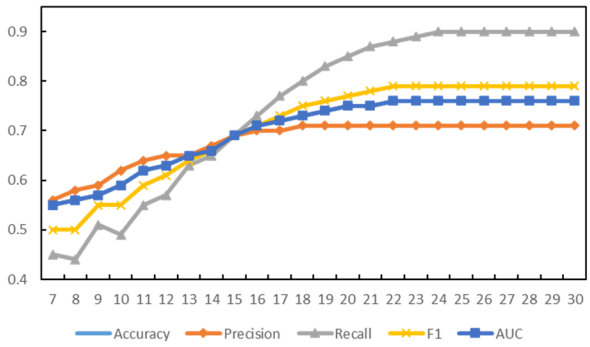

Figure 8. Index comparison chart

The precision value remains stable at 0.71 in length of the consumer behavior sequence is 18, indicating no further changes. Similarly, for a consumer behavior sequence length of 22, accuracy, F1 score, and AUC stabilize at values of 0.76, 0.79, and 0.76 respectively. Additionally, with a consumer behavior sequence length of 24, the precision value stabilizes at values0.71, 0.79 and 0.76 respectively; while recall remains constant at a value of 0.9 throughout these lengths.

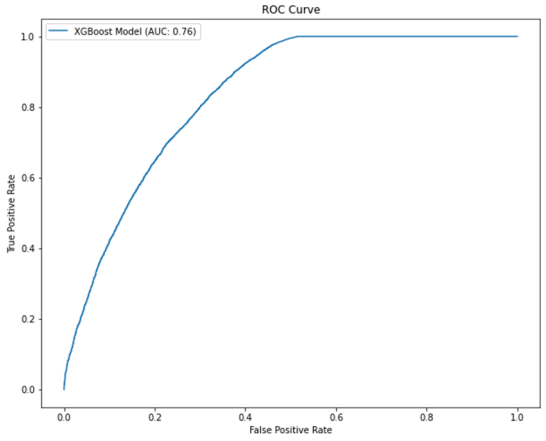

Figure 9. Final ROC curve

5. Conclusion

The predictive model in this study aims to predict consumers' purchasing behavior based on the sequence of operational behaviors generated in an e-commerce platform. We analyze the behavior sequence of each consumer and the related time series by commodity, and divide the sequence into different levels according to the number of behaviors. Using Python, we built the XGBoost predictive model, which is designed to use implicit behavioral information about consumers to predict their future behavior. Through the evaluation of the model, we found that the XGBoost model performs well in real-world e-business scenarios. Its F1 Score reaches 0.79 and AUC value is 0.76, which indicates that the

model can effectively analyze the purchasing trend of consumers. This study has important guiding significance for the relevant leaders of e-commerce platforms, which can help them better understand consumer behavior, to develop more effective sales strategies. In addition, the study provides valuable insights into how to maximize sales of e-commerce products. These results provide practical insights and methods for decision making in the field of e-commerce.

Acknowledgments

I would like to thank my advisor, Maria Rosario Rodavia, who has always supported my work, as well as the Angeles University Foundation, which has given me a platform to show my talents.

References

[1] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16).

[2] Li, H., & Guo, J. (2018). XGBoost Model for E-commerce Click-Through Rate Prediction. In Proceedings of the 1st International Workshop on Deep Learning Practice for High-Dimensional Sparse Data (DLP-KDD '18).

[3] Tong, J., Li, L., & Lu, Y. (2020). Predicting Customer Purchase Behavior with Gradient Boosting Machines: Evidence from E-commerce. Electronic Commerce Research, 1-19.

[4] Rashid, T., Li, L., & Wei, W. (2019). Customer Churn Prediction in Telecommunication Using XGBoost. IEEE Access, 7, 95127-95134.

[5] Jia, X., Gao, L., & Dai, Z. (2021). A Customer Behavior Analysis and Prediction Model Based on XGBoost Algorithm. In Proceedings of the 8th International Conference on Management, Education, Information and Control (MEICI '21).

[6] Sánchez-Morales, V., Sandoval-Oliva, D., & Martínez-González, M. Á. (2019). Predicting Customer Churn in Mobile Telecommunications Industry Using XGBoost. IEEE Access, 7, 38296-38305.

[7] Chen, Y., & Tang, W. (2020). A Hybrid Model for Customer Churn Prediction in E-commerce Using XGBoost and LSTM. IEEE Access, 8, 10582-10592.

[8] Sun, X., & Wu, J. (2019). Customer Churn Prediction in E-commerce Using XGBoost. In Proceedings of the 3rd International Conference on Data Science (ICDS '19).

[9] Yang, W., & Chen, L. (2022). XGBoost-based Customer Segmentation for Personalized Marketing in E-commerce. Information Sciences, 661, 37-53.

[10] Wu, H., Zhang, L., & Zhou, X. (2023). Enhancing Online Retail Sales Prediction with XGBoost and Transformer-based Models. Expert Systems with Applications, 213, 118799.