Journal of Electronic Research and Application, 2022, Volume 6, Issue 3
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

# Prediction of Online Consumers' Repeat Purchase Behavior via BERT-MLP Model

**Junchao Dong[1], Tinghui Huang[2]*, Liang Min[3], Wenyan Wang[4]**

[1]School of Computer Engineering, Guilin University of Electronic Technology, Beihai 536000, Guangxi Province, China

[2]School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, Guangxi Province, China

[3]Department of Computer Science and Technology, Xi'an Jiaotong University City College, Xi'an 710018, Shaanxi Province, China

[4]School of Computer Sciences, Universiti Sains Malaysia, Gelugor 11800, Penang, Malaysia

***Corresponding author:** Tinghui Huang, glhth@guet.edu.cn



**Abstract:** It is an effective means for merchants to carry out precision marketing and improve ROI by using historical user behavior data obtained from promotional activities in order to build a model to predict the repeat purchase behavior of users after promotional activities. Most of the existing prediction models are supervised learning, which does not work well with a small amount of labeled data. This paper proposes a BERT-MLP prediction model that uses "large-scale data unsupervised pre-training + small amount of labeled data fine-tuning." The experimental results on Alibaba real dataset show that the accuracy of the BERT-MLP model is better than the baseline model.

**Keywords:** Data mining; Business intelligence; E-commerce; BERT; Multilayer perceptron

***Online publication:*** June 3, 2022

## 1. Introduction

The competition between e-commerce platforms is growing increasingly heated these days. In order to attract more consumers, e-commerce platforms have been inviting well-known figures and online celebrities to play a part in live sales promotions on special days, such as "Double 11" and "618." However, the majority of customers lured by these live sales promotions will only end up purchasing promotional products and will not become devoted anchor users or offer merchants long-term revenue. Therefore, to reduce promotion expenses and increase return on investment, it is necessary to forecast the users who will participate in promotional activities and identify among them those loyal customers who will make repeated purchases. The process of purchasing products on e-commerce platforms is accompanied by a variety of actions, including clicking, collecting, and adding to shopping carts. Large-scale user log data are left on the e-commerce platform after a promotion ends, and deep mining of this data can reveal users' preferences for a particular product and whether they are inclined to purchase it. This is an effective way to predict repeat purchase behavior.

In recent years, many scholars have conducted research and investigations on user purchase or repeat purchase behavior, and the existing research results mainly include individual models and ensemble models. The individual prediction model is a prediction model of users' repeat purchase behavior based on a single

machine learning algorithm. Multiple models such as logistic regression (LR) [1], support vector machine (SVM) [2], recurrent neural network (RNN), and multi-layer perceptron (MLP) [3] have been widely used in the prediction of users' repeat purchase behavior. With the continuous development of social commerce and the gradual expansion of user scale, the scale of users' historical behavior data has surged. Traditional machine learning algorithms cannot achieve ideal results by relying on the characteristics of numerous influential factors, thus necessitating a new machine learning algorithm that is tailored to large-scale data machine learning algorithms in order to improve prediction performance. Therefore, many researchers have proposed multiple integrated prediction models in combination with different individual prediction models, among which the most representative ones are Random Forest (RF), Gradient Boosting Decision Tree (GBDT), and XGBoost. Results have shown that the ensemble prediction model is superior to the individual prediction model in both accuracy and robustness [4]. Aiming at the prediction of online consumers' purchase behavior, Hu and Shi proposed a Long Short-Term Memory (LSTM)-RF prediction model based on the combination of LSTM and RF by analyzing the shopping data of consumers on an e-commerce platform, in order to improve the learning ability and prediction effect of the prediction model [5].

However, the aforementioned existing models have a common problem: they are all supervised learning, and the model effect depends on the quantity and quality of manually labeled labels. Hence, if the scale of the labeled data is small, the model cannot achieve the ideal prediction effect. Based on the massive public competition data left by the social commerce platform Alibaba during the "Double Eleven" sales, this paper analyzes the influencing factors of users' repeat purchase behavior, builds a BERT-MLP model for a small amount of labeled data, and predicts the after-promotion repeat purchase behavior of users who purchased goods during the promotion.

## 2. BERT-MLP prediction model of users' repeat purchase behavior

Using the idea of "unsupervised pre-training + fine-tuning of annotated data" in natural language processing, the BERT-MLP individual prediction model can maintain good prediction results even with a small amount of labeled data. Firstly, it performs unsupervised training on unlabeled user behavior sequence data to obtain the pre-trained BERT model [3]. Then, the pre-trained BERT model is fine-tuned with a small amount of labeled data [6]. Finally, the fine-tuned user behavior sequence output and user age and gender are cascaded through the MLP model and input to the classification layer [7] to perform the final classification output for users' repeat purchase behavior, as shown in **Figure 1**.

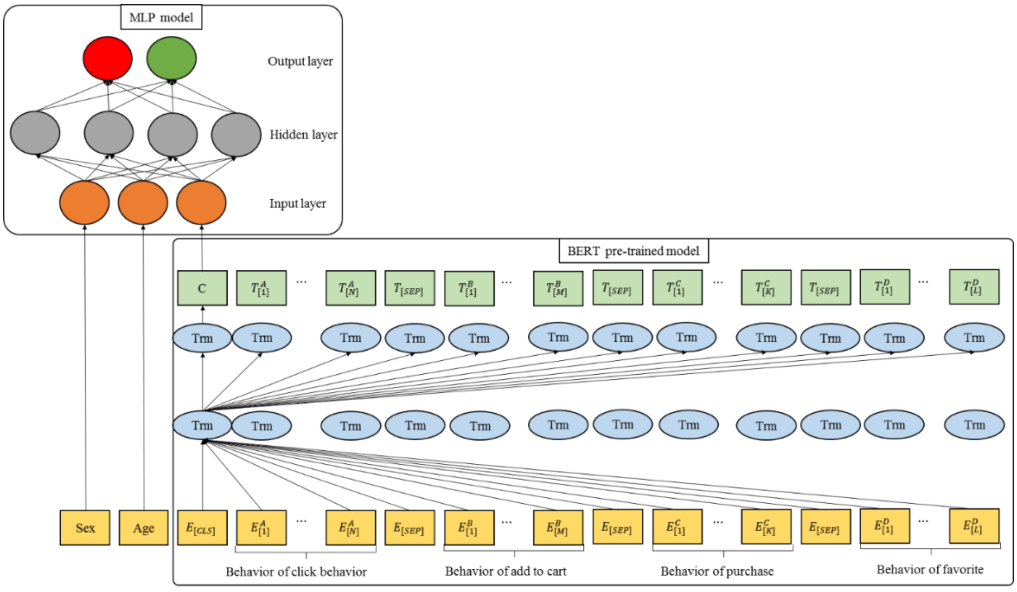


**Figure 1.** BERT-MLP individual prediction model framework

### 2.1. Pre-trained BERT model

BERT is a general natural language processing pre-trained model proposed by Google for multiple natural language processing (NLP) downstream tasks, such as sentiment analysis, named entity recognition, reading comprehension, and intelligent dialogue [8]. The model can dynamically encode word vectors based on contextual information, solve the problem of multiple meanings in natural language coding, and encode the logical relationship of each sub-clause in long sentences. Since user behavior sequences are highly similar to sentence sequences in natural language processing, a pre-trained BERT model has been built for the prediction of users' repeat purchase behavior by borrowing the current effective pre-trained BERT model in NLP. The model mainly includes a sequence encoding layer and a sequence information extraction layer.

#### 2.1.1. Sequence encoding layer

The main function of the sequence encoding layer is to encode users' behavior sequence, which is convenient for the input sequence information extraction layer. Sequence encoding is achieved by the masking method and swap order. The masking method involves randomly masking 15% of the specific moment of behavior data and allowing the model to make predictions based on the information of the sequence of behaviors before and after that moment. There are three ways to achieve this.

(1) Null value substitution
    80% masked behavior sequence is replaced with null values.
(2) Original value replacement
    The behavior sequence for the 10% mask is replaced with the original value.
(3) Remaining value substitution
    The sequence of behaviors for the 10% mask is replaced with the remaining values.

Swap order embodies the commutation method. First, the sequence of user purchase behavior is divided into two sub-sequences, and then the order of the two sub-sequences is exchanged. The specific implementation method is as follows: first, the user behavior sequence is divided into two subsequences by a special token [SEP], and then a subsequence unchanged behavior sequence and a subsequence swapped behavior sequence are input to the sequence encoding layer according to the 1:1 ratio, which facilitates the subsequent sequence information extraction layer to learn the logical relationship of the behavior sequence. The final output of the sequence encoding layer is the sum of token encoding, segment encoding, and position encoding. The segmented coding can divide the user behavior sequence into two subsequences and four types of user behavior sequences. The outputs of the four behavior sequences (click, add to cart, purchase, and favorite) in the sequence encoding layer are as follows: $E_A = \{e_a^1, e_a^2, \cdots, e_a^N\}$, $E_B = \{e_b^1, e_b^2, \cdots, e_b^M\}$, $E_C = \{e_c^1, e_c^2, \cdots, e_c^K\}$, and $E_D = \{e_d^1, e_d^2, \cdots, e_d^L\}$.

#### 2.1.2. Sequence information extraction layer

The sequence information extraction layer adopts the Funnel-Transformer encoder [9]. Compared with the original Transformer encoder in the BERT model, the Funnel-Transformer adds a pooling operation to each block, which has higher computational efficiency and lower space consumption; in addition, it is more suitable for information extraction of user behavior sequences with longer sequences. The Funnel-Transformer encoder includes a Transformer encoder and pooling operation. The Funnel-Transformer encoder consists of N identical Transformer basic blocks stacked with pooling operations. A Transformer basic block consists of two sub-blocks, multi-head attention, and a feed-forward neural network. Residual connections and layer normalization connections are used between the two sub-blocks, as shown in **Figure 2**.

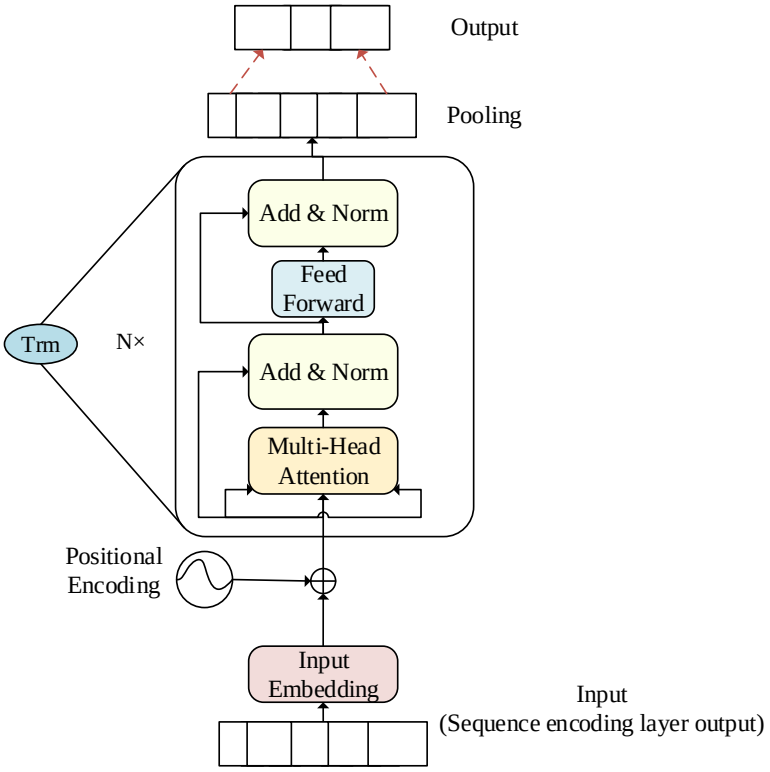


**Figure 2.** Funnel-Transformer encoder basic block

In order to deal with the low efficiency of Transformer in long behavior sequences, a pooling operation is added to the output vector of each Transformer basic block, except the output vector of the special token. The dimensions of Transformer output vectors in the direction of behavior sequences are compressed to reduce the number of model parameters and improve the efficiency of the model in long behavior sequences, as shown in Equation 1.

$$E^{'} \leftarrow Poooling(E^{*}) \tag{1}$$

$E^{*}$ is the output vector of the Transformer basic block, $Poooling(\cdot)$ is the pooling operation, and $E^{'}$ is the output vector after pooling. At this point, the output of the Transformer basic block is shown in Equation 2.

$$E^{''} \leftarrow LN(E^{*} + \text{Self-Attion}(Q^{'},K,V)) \tag{2}$$

$E^{*}$ is the output vector of the Transformer basic block, $\text{Self-Attion}(\cdot)$ is the self-attention mechanism, and $Q^{'}=E^{'}$ is the output vector after pooling. K and V are the same as the K and V of the base Transformer block before pooling.

### 2.2. Multilayer perceptron and model training

The multilayer perceptron (MLP) consists of three layers of nodes: an input layer, a hidden layer, and an output layer. The task of the input layer is to cascade the encoding information of user behavior sequence, user age, and user gender. The hidden layer includes multiple fully connected activation function neurons, and the hidden neurons are nonlinearly transformed through the Rectified Linear Unit (ReLU) activation function, as shown in Equation 3.

$$H = F(VW_h + b_h)W_0 + b_0 \tag{3}$$

$V = E^{''}_{(n)} + E_{Age} + E_{gender}$ is the concatenated vector of the final user behavior sequence output vector, user age vector, and user gender vector of n Transformer-Pooling blocks; $F(\cdot)$ is the ReLU activation function, $W_0$ is the weight vector, and $b_0$ is the bias term. The main task of the output layer is to classify whether the output users have repeat purchase behaviors, as shown in Equation 4.

$$y = \text{softmax}(x) = \frac{\exp(x)}{\sum_{k=1}^{c} \exp(x)} \tag{4}$$

$y \in R^c$ represents users' repeat purchase behavior category; $c$=2 represents the two behavior categories of users {repeat purchase, non-repeat purchase}. The model adopts the cross-entropy loss function, and the parameters are trained and optimized through the backpropagation algorithm.

## 3. Experiments and analysis

### 3.1. Dataset

The data in this paper are derived from the complete behavior data of more than 400,000 users and millions of commodity information provided by Alibaba's "Tianchi Big Data Competition" [10]. The data consist of two parts: user behavior data and user information data. The aim of this study is to predict whether users will repeatedly purchase products from the same merchant within half a year after the end of the promotion based on the historical behavior data of users on the "Double Eleven" sales and the first half-year of the social commerce platform. The test dataset is similar to the training dataset, except that the label field is Null.

### 3.2. Experimental environment and hyperparameters

The detailed configuration of the experimental environment in this paper is shown in **Table 1**.

**Table 1.** Experimental environment configuration list

| Type | Name | Configuration |
|---|---|---|
| Hardware | Central processing unit | Intel(R) Xeon(R) CPU E3-1226 v3 @ 3.30GHz |
| | Memory | 32GB |
| | Graphics processing unit | NVIDIA GTX1080Ti 11GB |
| Software | Operating system | Windows 10 Enterprise LTSC x64 1809 |
| | Programming language | Python 3.7 |
| | Machine learning framework | Sklearn 0.23.1 |
| | Natural Language Processing Framework | Gensim 4.0.1 |
| | Deep Learning Framework | PyTorch LTS 1.8.1 |

The recommended hyperparameter values of existing research results are referred to, in order to reduce the number of parameter combinations and improve the efficiency of parameter selection. The final results of the main hyperparameter selection of the BERT-MLP model are shown in **Table 2**.

**Table 2.** BERT-MLP main parameter configuration list

| Parameter name | Parameter value |
|---|---|
| layer | 12 |
| hidden | 768 |
| heads | 12 |
| parameters | 110M |
| loss_functioSn | Adam |
| Max_seq_length | 128 |
| masked_lm_prob | 0.15 |
| random_seed | 12345 |
| dupe_factor | 5 |
| learning rate | 1e-4 |
| hidden_layer_sizes | 256 |

### 3.3. Evaluation index

The prediction of users' repeat purchase behavior is a two-class problem in machine learning; hence, the three most commonly used evaluation indicators for two-class machine learning models have been chosen: accuracy, F1 score, and AUC.

### 3.4. Experimental results and analysis

In order to verify the effectiveness and practicability of the BERT-MLP model, it is compared with five baseline models (LR, KNN, RF, GBDT, and XGBoost) and two ablation models (MLP and BERT) under the same dataset and experimental environment. In order to ensure the accuracy and objectivity of the experimental results, ten-fold cross-validation is used to train each model, and the average value of accuracy, F1 score, and AUC are obtained as the final statistical results of the model, as shown in **Table 3**.

**Table 3.** Experimental results of the FCV-Stacking ensemble model, four individual models, and five baseline models

| Category | Model | Accuracy | F1 score | AUC |
|---|---|---|---|---|
| Baseline model | LR | 0.8685 | 0.7825 | 0.5720 |
| | KNN | 0.9225 | 0.7863 | 0.5981 |
| | RF | 0.9355 | 0.7833 | 0.6145 |
| | GBDT | 0.8685 | 0.7991 | 0.5830 |
| | XGboost | 0.9305 | 0.7833 | 0.6179 |
| Ablation model | MLP | 0.9362 | 0. 7952 | 0.6279 |
| | BERT | 0.9350 | 0.7983 | 0.6220 |
| | BERT-MLP | 0.9355 | 0.7991 | 0.6232 |

**Table 3** shows that the accuracy, F1 score, and AUC of the FCV-stacking model are better than the comparison model. This indicates that the FCV-Stacking integrated prediction model is more effective in the selection and fusion strategy of individual models. The main reasons are as follows: based on the tree model, the DeepGBM individual model in the FCV-Stacking ensemble model not only effectively processes the dense numerical features in users' historical behavior data, but also the large-scale sparse classification features, thus further enriching the feature types and improving the performance of the model

compared with the RF, GBDT, and XGboost ensemble learning models,; DeepCatBoost leverages the advantages of layer-by-layer feature learning in deep learning and has better feature learning capabilities than RF, GBDT, and XGboost ensemble learning models. At the same time, this paper not only considers DeepGBM and DeepCatBoost individual models based on the tree model but also constructs DABiGRU based on neural network and the BERT-MLP model based on attention mechanism when building the ensemble model. This greatly increases the differences between the individual models in the FCV-Stacking integrated model, fully exploits the excellent feature learning ability of the tree model and the better data learning ability of the neural network, as well as synergizes with their respective advantages to improve the model effect and generalization ability. In addition, the experimental results show that the two linear models, LR and KNN, have poor performance and are not suitable for the repeat purchase behavior prediction task with large data scales and complex structures. By adding the MLP layer to the BERT-MLP individual model, the model effectively combines the two key influencing factors of user gender and user age with the original BERT representation of behavioral sequence information, thus improving the overall effectiveness of the model.

## 4. Conclusion

This paper takes the prediction of social business users' repeat purchase behavior as the research goal, conducts research on Alibaba real dataset, and proposes a BERT-MLP prediction model, which has been evaluated in three metrics: accuracy, F1 value, and AUC. It achieved better experimental results than the selected baseline models on all metrics. In order to further improve the F1 value and AUC, knowledge base will be integrated into the BERT-MLP model in the future. The research results are of great significance for improving users' purchase experience and for merchants to enhance their live broadcast strategy in a targeted manner, which is conducive to maintaining the healthy and orderly development of e-commerce platforms.


## Funding

Shaanxi Provincial Education Science Regulations "Fourteenth Five-Year Plan" Project "Research on the Application of Educational Data Mining in Applied Undergraduate Teaching: A Case Study of 'Computer Application Technology' Course" (Project Number: SGH21Y0403)

The 2020 Bureau of Shaanxi Provincial Sports Regular Project (Project Number: 2021392)

The Special Research Project of Xi'an Jiaotong University City College (Project Number: KCSZ01005)


## Disclosure statement

The authors declare no conflict of interest.

## References


[1] Dong Y, Jiang W, 2019, Brand Purchase Prediction Based on Time-Evolving User Behaviors in E-Commerce. Communications in Nonlinear Science and Numerical Simulation, 31(1): e4882.

[2] Tsur O, Rappoport A, 2012, Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, February 8-12, 2012: What's in a Hashtag? Content Based Prediction of the Spread of Ideas in Microblogging Communities. Association for Computing Machinery, New York, NY, United States, 643-652.

[3] Song HS, 2017, Comparison of Performance Between MLP and RNN Model to Predict Purchase Timing for Repurchase Product. Journal of Information Technology Applications and Management, 24(1): 111-128.

[4] Liu G, Nguyen T, Zhao G, et al., 2016, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 13-17, 2016: Repeat Buyer Prediction for E-Commerce. Association for Computing Machinery, New York, NY, United States, 155-164.

[5] Hu W, Shi Y, 2020, Proceedings of the 2020 5th International Conference on Communication, Image and Signal Processing, November 13-15, 2020: Prediction of Online Consumers' Buying Behavior Based on LSTM-RF Model, IEEE, 224-228.

[6] Sun C, Qiu X, Xu Y, et al., 2019, Proceedings of the 18th China National Conference, October 18-20, 2019: How to Fine-Tune Bert for Text Classification?. Springer-Verlag, Berlin, Heidelberg, 194-206.

[7] Tolstikhin I, Houlsby N, Kolesnikov A, et al., 2021, MLP-Mixer: An All-Mlp Architecture for Vision. arXiv, 2105.01601 (preprint).

[8] Qiu X, Sun T, Xu Y, et al., 2020, Pre-Trained Models for Natural Language Processing: A Survey. Science China Technological Sciences, 63: 1872-1897.

[9] Dai Z, Lai G, Yang Y, et al., 2020, Funnel-Transformer: Filtering Out Sequential Redundancy for Efficient Language Processing. arXiv, 2006.03236 (preprint).

[10] Alibaba Cloud Tianchi, 2021, Repeat Buyers Prediction-Challenge the Baseline. Alibaba Cloud. https://tianchi.aliyun.com/competition/entrance/231576/information.