Journal of Applied Mathematics and Computation, 2024, 8(1), 68-72
https://www.hillpublisher.com/journals/JAMC/
ISSN Online: 2576-0653
ISSN Print: 2576-0645

User Repeat Purchase Behavior Prediction Based on Accuracy-Weighted Stacking Algorithm

Yibing Huang[1,2], Sisi Zheng[2,*]

[1]School of Mathematical Information, Shaoxing University, Shaoxing, Zhejiang, China.
[2]School of Mathematics and Statistics, Huizhou University, Huizhou, Guangdong, China.

How to cite this paper: Yibing Huang, Sisi Zheng. (2024) User Repeat Purchase Behavior Prediction Based on Accuracy-Weighted Stacking Algorithm. *Journal of Applied Mathematics and Computation*, 8(1), 68-72.
DOI: 10.26855/jamc.2024.03.008

Received: February 28, 2024
Accepted: March 27, 2024
Published: April 26, 2024

***Corresponding author:** Sisi Zheng, School of Mathematics and Statistics, Huizhou University, Huizhou, Guangdong, China.

Abstract

In this paper, an accuracy-weighted stacking fusion model based on Borderline-SMOTE is developed to predict the repeat purchase behavior of users on the Tmall platform. GBDT feature selection algorithm was utilized to obtain the importance ranking of features, filter redundant features, and construct a more efficient feature system. By calculating the error of the base learner's predicted results, weight coefficients were constructed to weigh the output. Subsequently, the new training set was used as input for the meta-learner. Through accuracy weighting, learners with better performance received higher weights, enabling weighted heterogeneous integration of the model and improving overall performance. The experiments verify that the accuracy weighting stacking algorithm proposed in this paper, based on Borderline-SMOTE, achieves high fitting accuracy. The AUC, F1, recall, and precision have all increased, with the AUC value reaching 0.93. The prediction effectiveness surpassed that of the XGBoost, LightGBM, CatBoost, Random Forest, and other models. This research will provide superior prediction models for e-commerce platform merchants, assisting them in capturing potential new users and retaining loyal customers, ultimately generating greater business value.

Keywords

Accuracy-weighted, stacking fusion model, Borderline-SMOTE, repeat purchase, GBDT

1. Introduction

In recent years, the competition among e-commerce platforms has become increasingly fierce. In order to attract more users for consumption, merchants often launch large-scale promotional activities on specific dates. While these promotions can attract a large number of new buyers, many of them do not continue to make purchases after the promotional period, posing a significant challenge for merchants. Additionally, the diverse range of products on e-commerce platforms requires users to spend a considerable amount of time and effort searching for desired items, which, to some extent, impacts repeat purchase behavior [1]. Therefore, merchants need to study how to identify consumers worth investing in for repeat purchases and enhance the overall shopping experience to increase user stickiness and loyalty. To address these challenges, merchants can analyze users' online behavior data to identify potential repeat buyers. By recognizing users with the potential for repeat purchases and delivering targeted advertisements to these potential loyal customers, merchants can optimize their marketing efforts. This not only significantly reduces promotion costs but also improves return on investment. Furthermore, analyzing online behavior data allows merchants to assist users in quickly finding suitable products, saving browsing time, and enhancing the overall shopping experience, creating a win-win situation [2]. To efficiently leverage this online behavior data, merchants can analyze user behavior patterns, predict repeat purchase behavior, and provide feasibility references and recommendations for operational decision-making. This presents both opportunities and challenges for merchants on e-commerce platforms [3].

2. Materials and methods

2.1 Application algorithm

This article employed a variety of machine learning classification and regression algorithms, including the XGBoost algorithm, LightGBM algorithm, CatBoost algorithm, Random Forest algorithm, and logistic regression algorithm [4]. Among these, the first four algorithms were used to compare predictive performance with the ensemble model, serving as the base learners for the ensemble model. Additionally, the logistic regression algorithm was utilized as the meta-learner for the ensemble model [5].

2.1.1 XGBoost algorithm

XGBoost (eXtreme Gradient Boosting) is a further improvement on GDBT and can effectively enhance the performance of machine learning models. The XGBoost algorithm pre-prunes each tree, which is equivalent to using Newton's method to optimize the loss function. Additionally, it combines multi-threading, data compression, sharding, and other technologies to maximize computing speed. This algorithm has made a significant breakthrough in computing power and can train models faster and more efficiently.

2.1.2 LightGBM algorithm

LightGBM (Light Gradient Boosting Machine) is a distributed gradient boosting framework based on the decision tree algorithm. It can be used for regression, classification, sorting, and other tasks. LightGBM has been further optimized and improved based on gradient decision trees. This framework has the advantages of distribution and high performance. As a result, it is widely used in machine-learning tasks on various large-scale datasets.

2.1.3 CatBoost algorithm

CatBoost (Gradient Boosting + Categorical Features) is implemented by using a gradient-boosting tree-based learner. CatBoost has fewer parameters and can support categorical variables, while still providing high-precision results. Unlike traditional GBDT frameworks, CatBoost discourages the use of one-hot encoding because it can handle categorical variables more efficiently. The CatBoost model uses a sorting boosting algorithm, which can effectively solve the overfitting problem caused by gradient deviation.

2.1.4 Random forest algorithm

Random forest is an ensemble learning model that is based on bagging ideas and random feature selection. It achieves more accurate and stable predictions by building multiple decision trees and integrating them. During the construction process of the random forest, each decision tree is trained using a randomly selected part of the data and features, so that each decision tree has a certain degree of difference, thus improving the generalization ability and robustness of the entire random forest.

2.1.5 Logistic regression algorithm

Logistic regression is a binary classification algorithm used to predict the probability of an event. The main idea of logistic regression is to map the results of linear regression to the range of 0 to 1, representing the probability of an event occurring. Typically, samples with a probability greater than 0.5 are predicted as the positive class, while samples with a probability less than 0.5 are predicted as the negative class.

2.2 Improved stacking algorithm

The traditional stacking algorithm involves assigning equal weights to predictions from various base learners, which are then used as input for the meta-learner. However, for the chosen base learners in our case, their predictive performances are not consistent. The equal-weighted aggregation method assumes equal performance among all base learners, overlooking performance disparities between them [6]. In reality, different base learners may exhibit varied capabilities and performances; some learners might excel in predicting certain samples, while others may perform better on different samples. Neglecting these performance differences may lead to a degradation in the overall model performance.

In this study, we improved the stacking algorithm by introducing a performance-weighted approach. This method associates the weights of base learners with their performances on the training set, allowing better-performing learners to receive higher weights. The performance-weighted approach allocates weights based on the performance of base learners on the training set, ensuring that better learners receive higher weights, thus enhancing the overall model performance [7]. The pseudocode is provided below:

Algorithm 1: Improved stacking algorithm

Input: training data
Output: final prediction result
$D \leftarrow \{(x_{i1}, x_{i2}, \ldots, x_{im}, y_i)\}$, $i = 1,2,\ldots,n$;
trained base learner $\leftarrow h_1, h_2, \ldots, h_t$;
base learner error $\leftarrow e_{it}$, $i = 1,2,\ldots,n$;
base learner weight $\leftarrow \rho_{it}$, $i = 1,2,\ldots,n$;
meta-learner algorithm $\leftarrow A$;
meta-learner $\leftarrow L$;
for $i = 1,2,\ldots,n$ **do**
  for $t = 1,2,\ldots,T$ **do**
    $\text{pred}_{it} \leftarrow h_t(D)$;
    $\rho_{it} \leftarrow (\sum e_{it} - e_{it}) / \sum e_{it}$;
    meta-learner-train $\leftarrow \text{pred}_{it} * \rho_{it}$;
  end
  meta-learner dataset $\leftarrow$ {(meta-learner-train,y)} ;
  training meta-learner $L \leftarrow A$(meta-learner dataset) ;
  fusion model $H(D) \leftarrow L(h_1(D), h_2(D), \ldots, h_t(D))$;
end

Figure 1. Improved stacking algorithm.

3. Experimentation

3.1 GBDT feature selection

GBDT (Gradient Boosting Decision Tree) The principle of feature selection is based on the characteristics of the GBDT model. GBDT is an ensemble learning algorithm based on decision trees. It trains multiple weak classifiers through iterative iterations and ultimately combines them to form a strong classifier [8]. During the training process, GBDT ranks features based on their importance, with higher-ranked features considered more crucial for predicting the target variable.

3.2 Borderline-SMOTE

In classification problems, an imbalance in the number of samples across different classes can lead to poor predictive performance for the minority class, a situation known as the data imbalance problem. Upsampling and downsampling are techniques used to address data imbalance. In this study, the Borderline-SMOTE algorithm was selected for oversampling.

The sampling process of Borderline-SMOTE is as follows:

(1) Determine the minority class and majority class samples.

(2) For each minority class sample, calculate the number of majority class samples of its k nearest neighbors and label them as boundary samples or noise samples.

(3) For each boundary sample, randomly select a majority class sample from its k nearest neighbors, calculate the difference between them, and then randomly generate a synthetic sample between the two samples.

(4) Add synthetic samples to the data set.

(5) Repeat steps 2-4 until the gap between the majority class and the minority class is narrowed.

(6) Use the generated new data set to train and test the classification model.

The main advantage of the Borderline-SMOTE algorithm is its ability to increase the number of minority class samples without losing useful information. Since Borderline-SMOTE selectively chooses boundary samples to generate synthetic samples, it preserves the distribution and structure of the minority class in the original dataset, thereby enhancing the performance of the classifier [9]. In addition, this algorithm can adapt to different datasets and classifiers by adjusting parameters, demonstrating a certain degree of flexibility.

3.3 Model parameter configuration

Using a single model as part of the stacking ensemble model, while also serving as a benchmark for the comparison of the ensemble model, helps assess whether the performance improvement of the ensemble model is significant. This evaluation aids in determining whether to adopt the strategy of using an ensemble model. The hyperparameter settings of the model directly impact its predictive accuracy. In this study, we employed random hyperparameter tuning to determine the model's hyperparameters.

Compared to traditional methods like grid search, random hyperparameter tuning allows for trying more hyperparameter combinations in less time, making it more efficient in finding the optimal hyperparameter combination. Additionally, random tuning samples hyperparameters randomly within the hyperparameter space, covering the entire space more comprehensively. This approach mitigates the risk of focusing on a specific region, thereby increasing the likelihood of finding the optimal hyperparameter combination.

3.4Application of improved stacking algorithm

Due to the requirement of ensuring good performance and significant diversity among the base learners in the Stacking ensemble model, an analysis of the strengths and weaknesses of various classification algorithms is conducted. Several algorithms are selected based on their effectiveness in fitting data features and substantial differences in principles. The improved Stacking algorithm is then applied [10]. A heterogeneous ensemble is performed on XGBoost, LightGBM, CatBoost, and Random Forest, further enhancing the predictive accuracy of user repeat purchase behavior.

3.5 Analysis of model evaluation metrics

Based on the experimental results in Table 1, it is evident that before model fusion, the prediction results of the CatBoost, Random Forest, XGBoost, and LightGBM classifiers have already reached a favorable state. Utilizing XGBoost, LightGBM, CatBoost, and Random Forest as base learners, with logistic regression as the meta-learner, we constructed a multi-classifier stacking ensemble model. As shown in the table, the stacking ensemble model's predictive results are improved compared to individual models [11]. The accuracy-weighted Stacking ensemble model based on the Borderline-SMOTE algorithm not only overcomes issues like overfitting and underfitting present in single models but also effectively utilizes the diversity of individual models by constructing weight coefficients. This enhances the model's generalization ability and robustness.

Table 1. Prediction performance of each classifier

Model (Borderline-SMOTE)	Precision	Recall	F1-score	AUC
Logistic regression	0.61	0.24	0.35	0.59
XGBoost	0.90	0.81	0.85	0.89
LightGBM	0.75	0.87	0.80	0.88
CatBoost	0.93	0.74	0.84	0.86
Random forest	0.85	0.84	0.85	0.89
Stacking	0.92	0.84	0.87	0.90
Accuracy-weighted stacking	0.94	0.87	0.91	0.93

Borderline-SMOTE is an enhanced method of SMOTE. In order to achieve more accurate predictions, most classification algorithms strive to learn the boundaries of each class during the training process. Examples located at or near the boundary are more prone to misclassification compared to examples far from the boundary. As a result, these boundary examples carry greater importance for classification tasks. Borderline-SMOTE addresses this by selectively choosing boundary samples from the minority class to generate new samples. By doing so, it increases the sample size while preserving the minority class distribution and structure within the original dataset, thus enhancing the classifier's performance. In this study, both Borderline-SMOTE and SMOTE techniques were employed to sample the data. The model's prediction performance is presented in Table 2. From the table, it is evident that Borderline-SMOTE outperforms SMOTE in terms of prediction performance for this dataset.

Table 2. Performance of improved stacking model with different sampling methods

Model	Precision	Recall	F1-score	AUC
Accuracy-weighted stacking (SMOTE)	0.93	0.83	0.88	0.90
Accuracy-weighted stacking (Borderline-SMOTE)	0.94	0.87	0.91	0.93

4. Summary and prospects

The Borderline-SMOTE algorithm employed in this study interpolates only those samples located on the class boundaries, thereby better preserving the original sample distribution. In cases where noise data may exist within minority class samples, the SMOTE algorithm tends to interpolate these noise data points, thereby reducing the model's accuracy. In contrast, the Borderline-SMOTE algorithm, by exclusively interpolating samples situated on class boundaries, avoids interference with noise data, thereby enhancing model accuracy. The accuracy-weighted stacking algorithm assigns weights to individual base models based on their accuracy, aiming to improve the overall accuracy and stability of the model and enable more precise predictions of the target variable [12]. Experimental results indicate a significant advantage of the accuracy-weighted stacking algorithm over other single models and ensemble learning algorithms in enhancing predictive accuracy. This underscores the promising application prospects of the model in predicting user repeat purchases, helping merchants accurately recommend products to users, thereby enhancing user experience and benefiting the business.

Funding

This paper is supported by the 2022 Provincial and Municipal Joint Fund Project of Guangdong Basic and Applied Basic Research Foundation (grant number: 2022A1515110437), the 2018 Higher Education of Guangdong Key Platforms and Scientific Research Projects: Young Innovative Talents Program (grant number: 2018KQNCX251), and the 2017 Scientific Research Project of Guangdong Ocean University (grant number: 101502/R18014).

References

[1] Zhang Ning, et al. The impact of consumer perceived value on repeat purchase intention based on online reviews: By the method of text mining [J]. Data Science and Management, 2021: 22-32.

[2] Esmeli R, Bader-El-Den M, Abdullahi H. Towards early purchase intention prediction in online session based retailing systems [J]. Electronic Markets, 2021, 31: 697-715.

[3] Martínez A, Schmuck C, Pereverzyev Jr S, et al. A machine learning framework for customer purchase prediction in the non-contractual setting [J]. European Journal of Operational Research, 2020, 281(3): 588-596.

[4] Lee J, Jung O, Lee Y, et al. A comparison and interpretation of machine learning algorithm for the prediction of online purchase conversion [J]. Journal of Theoretical and Applied Electronic Commerce Research, 2021, 16(5): 1472-1491.

[5] Chaudhuri N, Gupta G, Vamsi V, et al. On the platform but will they buy? Predicting customers' purchase behavior using deep learning [J]. Decision Support Systems, 2021, 149: 113622.

[6] Ghasemian A, Hosseinmardi H, Galstyan A, et al. Stacking models for nearly optimal link prediction in complex networks [J]. Proceedings of the National Academy of Sciences, 2020, 117(38): 23393-23400.

[7] Ma L, Sun B. Machine learning and AI in marketing—Connecting computing power to human insights [J]. International Journal of Research in Marketing, 2020, 37(3): 481-504.

[8] Kshatri S S, Singh D, Narain B, et al. An empirical analysis of machine learning algorithms for crime prediction using stacked generalization: an ensemble approach [J]. Ieee Access, 2021, 9: 67488-67500.

[9] Kuric E, Puskas A, Demcak P, et al. Effect of Low-Level Interaction Data in Repeat Purchase Prediction Task [J]. International Journal of Human-Computer Interaction, 2023: 1-19.

[10] Xu J, Wang J, Tian Y, et al. SE-stacking: Improving user purchase behavior prediction by information fusion and ensemble learning [J]. Plos one, 2020, 15(11): e0242629.

[11] Park C, Kim D, Yang M C, et al. Click-aware purchase prediction with push at the top [J]. Information Sciences, 2020, 521: 350-364.

[12] Chen S, Wang X, Zhang H, et al. Customer purchase prediction from the perspective of imbalanced data: A machine learning framework based on factorization machine [J]. Expert Systems with Applications, 2021, 173: 114756.