INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN , ISSN-L , Volume: 17, Issue: 6, Month: December, Year: 2022
Article Number: 4988, https://doi.org/10.15837/ijccc.2022.6.4988

CCC Publications
AGORA UNIVERSITY PRESS

A Feature Engineering and Ensemble Learning Based Approach for Repeated Buyers Prediction

M. Zhang, J. Lu, N. Ma, T.C.E. Cheng, G. Hua

Mingyang Zhang
Department of Management Science and Engineering, School of Economics and Management
Beijing Forestry University, China
No. 35 Qinghua East Road, Haidian District, Beijing 100083, China
mingyangzhang@bjfu.edu.cn

Jiayue Lu
1. National Science Library, Chinese Academy of Sciences
Address33 Beisihuan Xilu, Zhongguancun, Beijing 100190, China
lujiayue22@mails.ucas.ac.cn
2. School of Economics and Management, University of Chinese Academy of Sciences
No.19A Yuquan Road, Beijing 100049, China

Ning Ma*
Department of Management Science and Engineering, School of Economics and Management
Beijing Forestry University, China
No. 35 Qinghua East Road, Haidian District, Beijing 100083, China
*Corresponding author: maning@bjfu.edu.cn

T.C. Edwin Cheng
Department of Logistics and Maritime Studies
The Hong Kong Polytechnic University
M923, Li Ka Shing Tower, Hong Kong Special Administrative Region, China
edwin.cheng@polyu.edu.hk

Guowei Hua
Department of Logistics Management, School of Economics and Management
Beijing Jiaotong University, China
Siyuan East Building, Beijing Jiaotong University, Haidian District, Beijing 100044, China
gwhua@bjtu.edu.cn

Abstract

The global e-commerce market is growing at a rapid pace, but the percentage of repeat buyers is low. According to Tmall, the repurchase rate is only 6.1%, while research shows that a 5% increase in the repurchase rate can lead to a 25% to 95% increase in profit. To increase the repurchase rate, merchants need to predict potential repeat buyers and convert them into repurchasers. Therefore, it is necessary to predict repeat buyers. In this paper we build a prediction model of repeat purchasers using Tmall's dataset. First, we build high-quality feature engineering for e-commerce scenarios by manual construction and algorithmic selection. We introduce the synthetic minority oversampling technique (SMOTE) algorithm to solve the data imbalance problem and improve prediction performance. Then we train classical classifiers including factorization machine and logistic regression, and ensemble learning classifiers including extreme gradient boosting, and light gradient boosting machine machines. Finally, we construct a two-layer fusion model based on the Stacking algorithm to further enhance prediction performance. The results show that through a series of innovations such as data imbalance processing, feature engineering, and fusion models, the model area under curve (AUC) value is improved by 0.01161. Our findings provide important implications for managing e-commerce platforms and the platform merchants.

Keywords: feature engineering; ensemble learning; fusion model; repeat buyer prediction.

1 Introduction

The rapid development of e-commerce has had a profound impact on the global economy[23]. In the US, e-commerce sales reached US$960.4 billion in 2021, accounting for 15% of total US retail spending[48]. In China, 81.6% of Internet users have undertaken online shopping[49]. Global e-commerce platforms have run promotions on specific dates (e.g., Black Friday, "6.18"), which significantly stimulate the potential purchase value of new and existing users. In 2021, the Taobao "Double Eleven" promotion, one of China's largest e-commerce promotions, achieved an all-time high transaction volume of 540 billion RMB. However, many buyers that participated in the promotion were one-time users, making it difficult to generate long-term benefits [50]. Studies have shown that the repurchase rate for Tmall is only 6.1%. The cost of acquiring new users is 5-10 times higher than the cost of maintaining old users, while a 5% increase in customer retention can increase profit by 25% to 95% [4][10]. Thus ,if focusing solely on new user acquisition in the long term and ignoring low levels of repurchase rates„ it can cost the platform a great deal of money but not the desired return on investment. Increasing repurchase rates is imminent. A possible method is to predict the users with the tendency to repurchase and then implement precise marketing to lead them to become repurchase users.

Recently, data-driven analytics have been increasingly developed to help companies analyze user behavior and optimize decision making[18][21][31][43]. However, e-commerce scenarios are complex and varied, and their data has special characteristics. The question of how to use a data-driven approach in e-commerce repurchase prediction is worth thinking about. We try to answer the following questions: How to comprehensively and accurately portray the features in e-commerce scenarios? How to efficiently predict potential repurchase users?

We propose an AI-based data-driven approach to address thess problem. We use the Tmall's real user dataset from AliTianchi to build the model. These data are log data generated by users in e-commerce behavior, which are implicit feedback. We carry out a series of steps to process the special characteristics of e-commerce data, introduce the synthetic minority oversampling technique (SMOTE) algorithm to solve the data imbalance problem. We manually construct feature engineering from three perspectives: merchant, user, and merchant-user. Meanwhile, we use different models for prediction, compare model performance, and choose the best approach. Results show that fusion model can significantly improve the accuracy of prediction. Finally, we efficiently and accurately predicted the potential target users in the dataset with a AUC value of 0.68406.

Predicting repeat buyers can help e-commerce firms provide personalized services, such as accurate product recommendations, differentiated pricing, and demand management, to effectively improve the repurchase rate. We organize the rest of the paper as follows: In Section 2 we review and summarize the related work. In Section 3 we introduce the data and the data pre-processing process. In Section 4 we discuss the data-driven methodology, including data imbalance processing, feature engineering, and the prediction models. In Section 5 we present the experimental results and analysis. Finally, in Section 6, we conclude the paper, discuss the management implications of the research findings, and suggest topics for future research.

2 Literature Review

In this paper, we use user behaviour data to predict whether he/she will become a repeat buyer. In this section we review and summarize related work covering two research streams, namely user behaviour research and user purchase behaviour prediction, and discuss the motivation for this study.

2.1 User Behaviour Research

Users generate massive data during using the online platform, which can be used to analyze their behavior. Further, user data can be divided into. "explicit feedback" and "implicit feedback"[26]. Explicit feedback is a direct and quantifiable expression of users' preferences, while implicit feedback records users' natural behaviours when using a product[12]. Explicit feedback relies on users' active evaluation and is difficult to collect, so the related data are quite sparse, while implicit feedback is easy to obtain and a large amount of the related data is available. Thus, many researchers have used implicit feedback.

Early researchers analyze implicit feedback data, such as click and search, to provide support for interface UI improvement and user behavior understanding[1][3][26]. Later studies have made richer use of user behavior data. Some scholars focus on the optimization of algorithmic models based on implicit feedback data. They combined implicit feedback data with the SVD algorithm in matrix decomposition[16][17], and applying collaborative filtering algorithms based on items and users[45][47] to improve algorithm accuracy. Further, more scholars have combined implicit feedback data with specific scenarios, such as e-commerce[34], finance[28], beauty[44],and medicine[24]. However, these studies only stay at analyzing the current behavior of users.

User behaviour analysis in the e-commerce context ultimately aims to predict subsequent user behaviour and raise the user purchase rate. The common methods used for user purchase prediction are statistical methods and machine learning methods. Statistical methods achieve prediction by modelling the relationships between input variables and output variables in advance[41]. However, under practical scenarios, the complex relationships between variables are often difficult to model. In addition, different models are based on different assumptions, rendering it difficult to achieve experimental prediction accuracy in reality[8]. Therefore, researchers have started to use machine learning, a method that does not require experimental simulations, for user purchase prediction. models that are often used for purchase prediction are decision trees[22], artificial neural networks[37]. However, all the related studies use a single prediction model, which has problems such as vulnerability to random factors and low generalization ability. To effectively exclude the interference of random factors in the single model and improve prediction accuracy, some researchers introduce the ideas of ensemble learning and fusion model into research on user behaviour prediction, such as GBDT[**?**], AdaBoost[27], CatBoost[6]. The above studies demonstrate that models using ensemble learning and fusion ideas have better prediction results than traditional single algorithm models.

In summary, early scholars' research stopped at behavioral analysis; however, some scholars have already started to make behavioral predictions. Implicit feedback is more suitable for behavioral analysis because of its large amount of data and high availability. Several scholars have attempted to predict e-commerce user behavior. They have confirmed the superiority of machine learning methods, and stronger ensemble models have emerged. Like these studies, this paper will also use user logs, an implicit feedback, for analysis. However, currently, a single machine learning model is mainly used, and this paper will try a more powerful integration approach.

2.2 Repeat Purchase Behaviour Prediction

In an increasingly competitive market, the prediction of e-commerce buying behavior has entered a new phase. For content, repeat purchase behavior is emphasized; for methodology, more advanced methods such as ensemble models are introduced.

Research on prediction of repeat buyers is much less abundant than research on topics such as user purchase behaviour prediction and user repurchase intention. Some researchers use non-machine learning methods such as interview method[33], game theory[29] and Buy till You Die (BTYD) models[9] to model and predict e-commerce users' repurchase behaviour. Other researchers have tried to use machine learning to improve the accuracy and robustness of prediction models. [39]applied the explanation method based on an improved decision tree algorithm to enable firms to explore the factors that drive customers' repurchases.[46]used the vote-stacking method to combine the prediction results of three separate models, namely DeepCatboost, DeepGBM, and DABiGRU, and found that the accuracy of fusion models is significantly higher than that of a single model. [13]proposed a BERT-MLP prediction model, with large-scale data unsupervised pre-training and small amount of labeled data fine-tuning, to predict repeat buyer.

However, the important step of feature engineering has been neglected in the above studies. While machine learning algorithms tend to be generic, feature engineering is specific, and good feature engineering determines the final prediction results.

The importance of feature engineering can be seen in the study of general user behavior prediction.[30] constructed feature engineering from the perspectives of users, merchants, products, brands, categories, and their interactions to improve prediction accuracy. Considering time-evolving features in feature engineering, [15] found that it could more realistically depict users' purchase intention. [**?**] dynamically updated user features monthly, characterizing customers in a given month, and achieved a prediction accuracy of 98%. Obviously, these studies have placed great emphasis on feature engineering.

In summary, methods such as game theory rely on strict assumptions, while the questionnaire method suffers from small sample size and under-representation, making it difficult for traditional statistical schemes to fully model user behavior. Therefore, we will use real data provided by Alibaba and use machine learning methods to prediction. Insufficient attention has been paid to feature engineering in existing studies, and we will manually construct a high-quality feature set for e-commerce scenarios. Finally, real data are prone to data imbalance and cheating users, and we will also give improvement measures.

2.3 Research Method

Ensemble learning is one of the frontiers in the field of computing. Its main principle is to train multiple learners and use special rules to combine their prediction results to improve the final prediction performance[36].

[11]introduced the concept of "ensemble learning" for the first time. [19]constructed an ensemble model based on a neural network model, and showed that the integrated model has lower absolute value of variance and superior generalization ability compared with the common neural network. [38]upgraded many weak performance classifiers by Boosting idea to obtain strong performance classification models. Since then, ensemble learning has jumped to be a popular research topic. Many new models have been born, such as hybrid expert models[20], stacked generalization models[42], bagging algorithms[5], and so on.

Ensemble learning has rich applications in e-commerce, such as e-commerce review mining[25][40], e-commerce product category labeling and recommendation[**?**][35], e-commerce security[7], among others. These studies confirm the superiority of ensemble learning approaches. In summary, the superiority of ensemble learning is undeniable and is widely used in e-commerce. We will use ensemble learning for repurchase prediction. In addition, are there methods that can combine ensemble learning with classical models to optimize the results? We will also explore this question.

3 Data Description and Pre-Process

In this section we introduce the dataset and discuss data pre-processing process.

3.1 Research Method

The data come from Alibaba's Tianchi platform. The data are provided by Tmall, which records the real data of 4,995 merchants and 424,170 new buyers in the 2014 Double Eleven shopping festival on the Tmall platform. The purpose of the experiment is to predict whether a new user that purchases a merchant's product on the Double Eleven Day will make a second purchase from that merchant

Table 1: User Behaviour Log

Field Name	Description
user_id	Unique ID code of the purchaser
item_id	Unique ID code of the merchandise
cat_id	Unique ID code for merchandise categories
merchant_id	Unique ID code for merchants
brand_id	Unique ID code for merchandise brands
time_tamp	Purchase time
action_type	Contains 0, 1, 2, 3. 0=Click, 1=Cart Add, 2=Purchase, 3=Bookmark to Favorites

Table 2: User Profile Table

Field Name	Description
user_id	Unique ID code of the purchaser
age_range	User age range. 1 for <18 years; 2 for [18,24]; 3 for [25,29]; 4 for [30,34]; 5 for [35,39]; 6 for [40,49]; 7 and 8 for > = 50; 0 and NULL for unknown
gender	User gender. 0 for female, 1 for male, 2 and NULL for unknown

within six months, and the new user is are a "repeat buyer" for each merchant. The dataset consists of three tables, namely the user behaviour log table, the user profile table, and the training set table.

The user behaviour log records the behaviours of all the "new users" on the day of Double Eleven and 6 months before Double Eleven, spanning 12 May - 11 November 2014. The field information is listed in Table 1.

The user profile table records the demographic information. Table 2 lists the field information.

The training set table records whether a specific user makes repeat purchases at a specific merchant. The field information is listed in Table 3.

3.2 Data Pre-Processing

Data pre-processing includes data cleaning, data integration, data transformation, data imputation etc. In this paper we use real behaviour log data, and there is inevitably data noise. Data quality will directly affect the accuracy and universality of the prediction model. Therefore, we combine the characteristics of the dataset and the special attributes of e-commerce to pre-process the data to improve data quality.

3.2.1 Data Integration

We use three datasets, where the two fields of user id (user_id) and merchant id (merchant_id) are common fields. To improve the efficiency of feature engineering, we use user id and merchant id as the primary keys for data integration. We show the structure of the integrated user purchase behaviour information table in table 4.

3.2.2 Missing Value Processing

The age and gender data of users in the dataset are missing 0.52% and 1.52%, respectively because they are category attributes with significant differences in repurchase behaviour, the missing values are filled by the plural of the whole data. For the brand id, on the one hand, it is difficult to replace it by plenary features or other attributes of the dataset; on the other hand, it has no impact on the subsequent feature engineering, so the missing data are excluded.

Table 3: Training Set Table

Field Name	Description
user_id	Unique ID code of the purchaser
merchant_id	Unique ID code of a merchant
label	Repeat purchase user identifier. Contains 0, 1. 1 for a repeat buyer, 0 for a non-repeat buyer.

Table 4: User Purchase Behaviour Information

Field Name	Description
user_id	Unique ID code of a purchaser
merchant_id	Unique ID code of a merchant
item_id	Unique ID code of an item
cat_id	Unique ID code of a merchandise category
brand_id	Unique ID code of a merchandise brand
time_tamp	Purchase time
action_type	Contains 0, 1, 2, 3. 0=Click, 1=Cart Add, 2=Purchase, 3=Bookmark to Favorites
age_range	User age range. 1 for <18 years; 2 for [18,24]; 3 for [25,29]; 4 for [30,34]; 5 for [35,39]; 6 for [40,49]; 7 and 8 for > = 50; 0 and NULL for unknown
gender	User gender. 0 for female, 1 for male, 2 and NULL for unknown
label	Repeat purchase user identifier. Contains 0, 1. 1 for a repeat buyer, 0 for a non-repeat buyer.

3.2.3 Abnormal User Identification

In the e-commerce environment, there are phenomena such as crawlers and swipers, whose behaviors are different from the normal purchase behaviour, and who belong to abnormal users. If a user has a large amount of product browsing behaviour in a period but the purchase behaviour is 0, then the user is likely to be a “crawler user”. If a user has a large amount of product purchase behaviour in a period but little or no browsing behaviour, then the user is likely to be a “crawler user”. After identifying the abnormal users, we delete their records.

4 Methodology

In this section we present methodologies related to data imbalance processing, feature engineering, predictive models, and evaluation metrics.

4.1 Data Imbalance Processing

In our dataset, the positive sample, i.e., the percentage of repeat users is only 6.1%, and the data are severely imbalanced. When the data are unbalanced, the minority class samples in the overlapping region between classes will be misclassified in large batches, and the class interval surface will move to the side with sparse sample distribution, thus interfering with the classification accuracy of the model for the minority classes. Existing research deals with data imbalance in three aspects, namely data pre-processing, features, and algorithms. In this paper we use SMOTE for imbalanced data processing(Figure 1). The process is as follows:

– Each sample x in the minority class is obtained as its k-nearest neighbour based on its distance to all the samples in the minority class set (generally using the Euclidean distance).

– The sampling multiplicity N is determined by the degree of sample positive and negative class imbalance; many samples are arbitrarily selected from the k nearest neighbours of each minority class sample x. The selected nearest neighbour is assumed to be x_{new}.

– To the arbitrarily selected nearest neighbour x_new, construct its new sample about the original sample as follows:

$$x_{new} = x + rand(0,1) * (\widetilde{x} - x). \tag{1}$$

4.2 Feature Engineering

Feature construction refers to the use of manual methods to select meaningful data in the initial dataset, or to combine and deform the initial data to obtain new features. To better characterize e-commerce behaviour, we focus on two main subjects in the e-commerce context, namely users and merchants, and construct features manually in three dimensions, namely user portraits, merchant portraits, and user-merchant interaction portraits. Besides basic features (gender, age), different

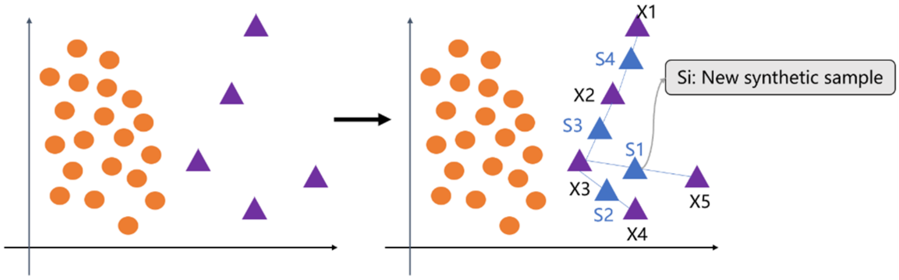


Figure 1: Principle of SMOTE

statistical methods are used to construct features, including counting and ratio features, aggregation features (mean, median etc), and temporal features.

The construction of features is based on the features of e-commerce and the subjective judgement of researchers, which are prone to developing invalid features and feature redundancy. If we use all the features, the model will be inefficient. Therefore, we keep the strong features and drop the useless features to avoid the loss of important information.

Feature selection mainly addresses three types of problems, namely dimensional catastrophe, overfitting, and noise, which can not only reduce the model complexity and computation but also improve the final prediction of the model by including high-quality features. There are three types of mainstream methods for feature selection as follows:

– Filtering: Feature selection and model training are independent of each other. First, select the experimental data features, then train the model using the filtered data set, and determine the weights by the scores.

– Wrapping: Feature selection and model training are correlated with each other. A subset of all the features is selected for model training and compared each time, and the best features are selected based on the classifier results.

– Embedding: Feature selection is carried out together with model training. The dataset is trained on the model, the weights of the features are obtained by model fitting, and feature selection is performed in the order from highest to lowest.

There is no uniform way for feature selection, so we use four methods for feature selection including random forest, ANOVA, recursive feature elimination, and L1 regularization, and compare the results and retain the best set of features.

4.3 Prediction Models

We use classical machine learning models, ensemble models, and fusion models to predict repeat buyers and compare the prediction results of different models.

– For classical machine learning, we use the common classification predictors such as logistic regression, factorization machine, decision tree, and support vector machine for experiments, and use prediction accuracy as the baseline in this paper.

–Ensemble models combine multiple learners into one stronger learner and can crack problems that cannot be solved by a single model. Therefore, using integrated learning can yield higher prediction accuracy and more reliable prediction results than a single model[14]. We use ensemble learning models such as XGBoost and LightGBM, which are commonly used for dichotomous prediction, to make predictions.

–Different models have unique advantages, and fusing models by certain methods can construct stronger classifiers and greatly improve the prediction results, so we fuse different models to improve prediction accuracy. Stacking is one strategy of model fusion, which is usually a two-layer construction[42]. Its framework is shown in Figure 2. First, the original dataset is divided into several sub-datasets and input to the n base learners in the first layer in turn; the output training results become the input to the second layer learners, which are trained to output the final results.

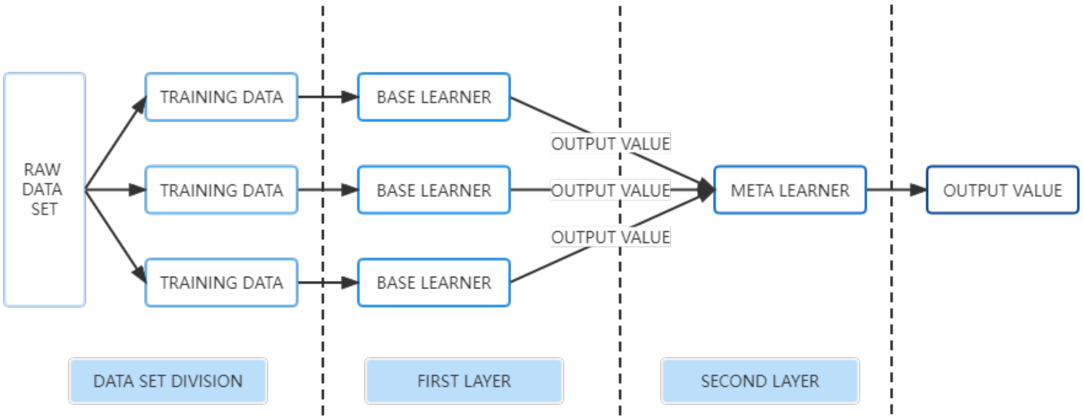
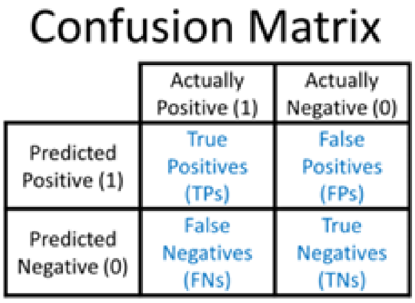


Figure 2: The Framework of Stacking

The above model contains numerous parameters and optimizing the parameters can effectively improve the model's performance. Gradient method, genetic algorithm, and other common parameter optimization methods converge faster, but with more than two parameters, the parameters will affect one and other and interfere with the results. The grid search method combines the parameters and performs the optimization search at the same time, avoiding the local optimum, improving efficiency, and obtaining the best generalization ability. In this paper we use grid search with cross-validation (GSCV) algorithm for model parameter optimization. This method combines cross-validation and grid search. In the grid search, the parameters are gradually adjusted by the learning rate in a limited range, the learner is trained, and the results are continuously compared. The scores of the model on the test set are calculated, and the final score is averaged over k times.

4.4 Model Evaluation Metrics

We use the area under curve (AUC) value as the model classification ability evaluation metric. The problem in this paper is dichotomous, i.e., whether a user is a repeat buyer, there are the positive class (1) and negative class (0), and the positive class is the repeat buyer. The positive class and the negative class of the real type and the predicted type constitute the confusion matrix, as shown in Figure 3.



Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figure 3: Confusion Matrix.

Four major categories are derived from the confusion matrix, namely TP, FN, FP, and TN, representing the number of samples in the four categories of true positive, false negative, false positive, and true negative, respectively. The accuracy rate, which is the ratio of the number of samples with correct classification results to the total number of samples is calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}. \tag{2}$$

In addition, two indicators can be derived, namely the true positive rate ($TPRate$) and the false positive rate ($FPRate$), which are calculated as follows:

$$TPRate = \frac{TP}{TP + FN}. \tag{3}$$

$$FPRate = \frac{FP}{FP + TN}. \tag{4}$$

$TPRate$ indicates the probability that a sample with true category 1 is predicted to be class 1, and FPRate indicates the probability that a sample with true category 0 is predicted to be class 0. The curve formed by taking $FPRate$ as the horizontal axis and $TPRate$ as the vertical axis is the receiver operating characteristic (ROC) curve, as shown in Figure 4. In general, $AUC = 1$ means the perfect classifier, $1 > AUC > 0.5$ means fair performance, and $0.5 > AUC > 0$ means poor performance.

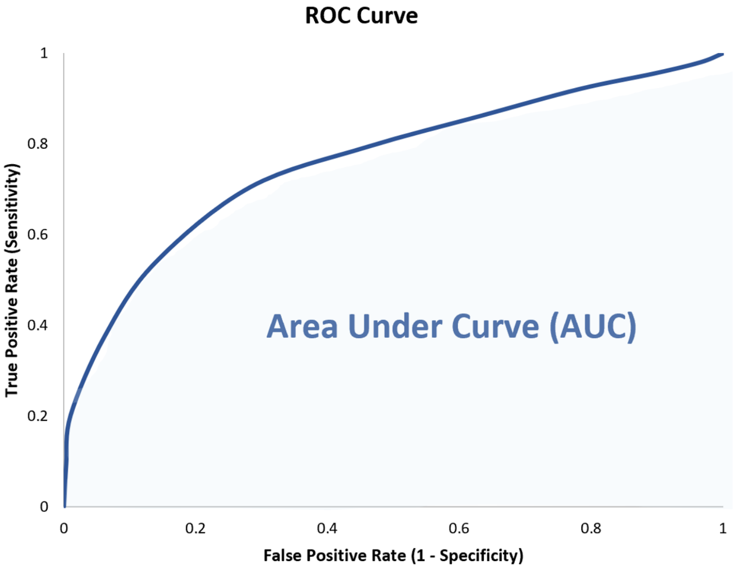


Figure 4: Confusion Matrix.

When the sample is imbalanced between the positive and negative classes, the AUC value can measure the classification accuracy when the true type is positive (=1) and negative (=0) at the same time, which can avoid the impact of the sample imbalance on the evaluation of the classifier.

5 Experimental Results and Analysis

In this section we present the results after data processing and modelling based on the methodology presented in the previous section, including the feature sets, prediction model parameters and results, and important features.

5.1 Feature Construction and Selection

The user portrait dimension includes four aspects:

– Basic information. During the initial exploration of the data, we find that user repurchase differs in different ages and genders, so these attributes should be added.

–Time information. It mainly describes the user's activity of "shopping" and fluctuations of the activity over time.

–Preference information. It mainly describes the user's favorite items/categories/shops/brands, the pattern of the user's various operations, and the comparison between a user's value and the average value of all the users.

Table 5: Results of Feature Selection.

Method	No Select AUC	Feature Select AUC	Feature dimension before selection	Feature dimension after selection
Random forest-based feature selection	0.582	0.587	147	94
ANOVA	0.582	0.585	147	79
Recursive feature elimination	0.582	0.585	147	74
L1 regularization	0.582	0.569	147	34

–Behaviour information. It mainly describes the frequency, extensiveness, and recent repurchase behaviour of users' various operations. A total of 83 features are constructed.

The merchant portrait dimension includes three aspects:

–Basic information. In the initial exploration of the data, we find that there are differences in the repurchase situations of stores, and that items, brands, and categories are the most basic attributes of an e-commerce merchant.

–Time information. It mainly describes the frequency and status of the merchant in operation.

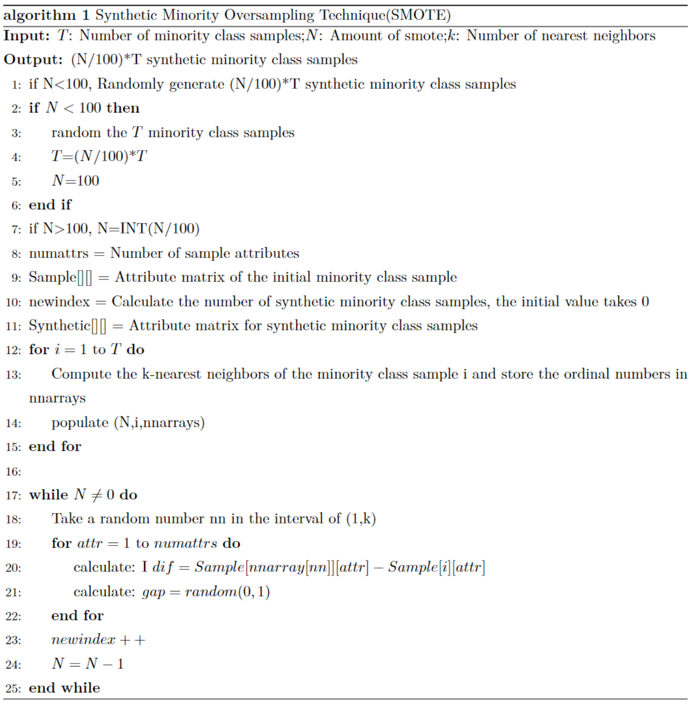
–Strength and popularity information. It mainly describes the merchant's popularity with users, the merchant's strength, and the merchant's repurchase situation. A total of 41 features are constructed.

The user-merchant interaction portrait dimension is the most important dimension to show the repurchase characteristics and relationship between different users and merchants, and describes the interaction information between merchants and users in terms of frequency, time, and status through user-merchant matching. A total of 23 features are constructed.

We successively use four methods for feature selection based on random forest, ANOVA, recursive feature elimination, and L1 regularization. The results of the AUC values and feature dimensionality before and after the four methods are shown in Table 5. Finally, we pick the random forest-based feature selection as the best, and retain the selected 94 features as the feature set.

5.2 Data Preparation

We first address the data imbalance problem by using the SMOTE algorithm, which is implemented using the SMOTE interface in the imblearn library in python. The pseudo-code is shown in Figure 5. To avoid high generalization error, the dataset is often split by the self-help method, which will change



algorithm 1 Synthetic Minority Oversampling Technique(SMOTE)

Input: T: Number of minority class samples;N: Amount of smote;k: Number of nearest neighbors

Output: (N/100)*T synthetic minority class samples

```
if N<100, Randomly generate (N/100)*T synthetic minority class samples
if $N < 100$ then
    random the $T$ minority class samples
    $T$=($N$/100)*$T$
    $N$=100
end if
if N>100, N=INT(N/100)
numattrs = Number of sample attributes
Sample[][] = Attribute matrix of the initial minority class sample
newindex = Calculate the number of synthetic minority class samples, the initial value takes 0
Synthetic[][] = Attribute matrix for synthetic minority class samples
for $i = 1$ to $T$ do
    Compute the k-nearest neighbors of the minority class sample i and store the ordinal numbers in
nnarrays
    populate (N,i,nnarrays)
end for

while $N \neq 0$ do
    Take a random number nn in the interval of (1,k)
    for $attr = 1$ to $numattrs$ do
        calculate: I $dif = Sample[nnarray[nn]][attr] - Sample[i][attr]$
        calculate: $gap = random(0, 1)$
    end for
    $newindex{+}{+}$
    $N = N - 1$
end while
```

Figure 5: Pseudo-code for the SMOTE Algorithm.

Table 6: Best Parameters and Prediction Results of a Single Model.

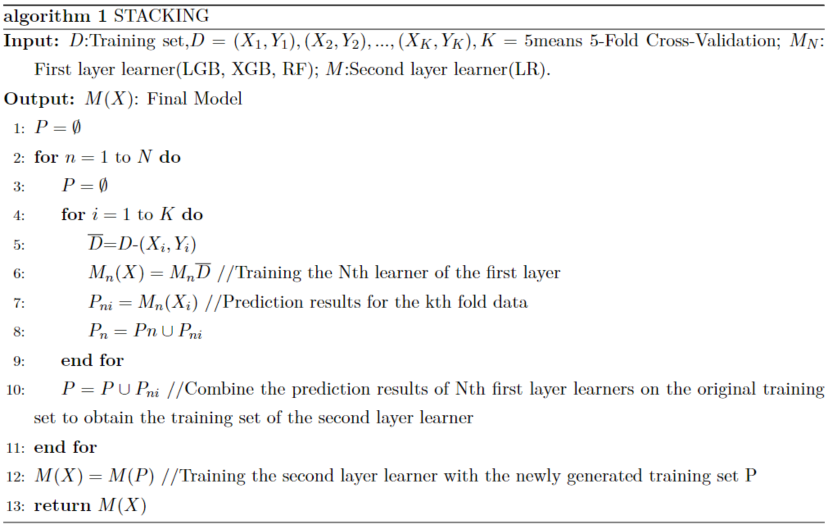
Model Best	parameter	AUC score
LR	Penalty: L2; intercept_scaling: liblinear; C: 0.05; class_weight: None; max_iter: 100	0.67305
FM	n_iter: 0; l2_reg_w: 0.1; rank: 4	0.67245
XGBoost	n_estimator: 2000; learning_rate: 0.01; max_features: 9; subsample: 0.5; max_depth: 8; min_samples_split: 1000; min_samples_leaf: 30; scale_pos_weight: 0.061	0.67956
LightGBM	max_depth: 7; num_leaves: 80; colsample_bytree: 1; learning_rate: 0.01; reg_alpha: 0	0.67991

the distribution of data, so we do not apply it to this study. As the amount of data used in this paper is sufficient, we use the simple and efficient leave-out method, dividing the training set by 1:1, with 50% as the training set and 50% as the test set. We apply five-fold cross-validation, i.e., we divide the training set into five equal and arbitrarily divided subsets, where the data sets are mutually exclusive. Each time, we randomly chose a subset as the validation set and the other four as the training set.

5.3 Model Construction and Parameters

We first use the classical models such as logistic regression (LR), factorization machine (FM), decision tree (DT), and support vector machine (SVM), finding that LR and FM perform the best. Meanwhile, the good prediction accuracy achieved by the classical algorithms confirms the effectiveness of our study in dealing with data imbalance and constructing feature engineering. Seeking better prediction performance, we apply the ensemble learning models to construct XGBoost, and LightGBM for prediction. The parameter settings and AUC values of a single model are shown in Table 6. The optimal parameters here are obtained by the grid search with cross-validation (GSCV) algorithm.

To achieve the best prediction possible, we introduce the Stacking method to construct a two-layer fusion model. We consider the following factors in the selection of the first layer base learner: a) Performance: A strong learner should be selected to achieve the effect of combining advantages; otherwise, it will affect the efficiency of the whole model; there should be differences in the structure of each learner to fuse and learn from different perspectives. b) Number: The number n of the first layer model determines the feature dimension n+1 of the second layer model input, and the dimension should be at least three, so at least two models are selected. After comprehensive consideration and trials, we select LightGBM, XGBoost, and RF as the base learners in the first layer. The second layer generally uses weak learners, and we choose the LR model, which is commonly used in machine learning studies. The LR model has strong generalizability and can avoid the risk of overfitting in the Stacking algorithm. The pseudo-code of the Stacking fusion model is shown in Figure 6. The



```
algorithm 1 STACKING
Input: $D$:Training set,$D = (X_1, Y_1), (X_2, Y_2), ..., (X_K, Y_K), K = 5$means 5-Fold Cross-Validation; $M_N$:
    First layer learner(LGB, XGB, RF); $M$:Second layer learner(LR).
Output: $M(X)$: Final Model
P = ∅  →  $P = \emptyset$
```

Figure 6: Pseudo-code for the Stacking Fusion Model.

Table 7: Comparison of Model Results.

Model AUC	score	Ranking	Gain	Other studies
FM	0.67245	Top 9.7%	-	-
LR	0.67305	Top 9.5%	+0.0006	0.617[51]
XGBoost	0.67956	Top 5.6%	+0.00651	0.6774[52]
LightGBM	0.67991	Top 5.5%	+0.00035	0.6797[53]
Stacking fusion model	0.68406	Top 2.8%	+0.00415	0.6232[54]

Table 8: Top Ten Features.

Rank Feature	
1	Number of purchases made by users at merchants
2	The number of repeat purchases made by users before "Double Eleven"
3	The click-to-purchase conversion rate of users to merchants
4	Number of repeated purchases made by users before "Double Eleven"
5	Number of store item categories
6	Difference between number of times repurchase is made to a merchant before "Double Eleven" and the average value of all the merchants
7	Difference between the user click-to-purchase conversion rate and the average value of all the users
8	Variance of user active days
9	Number of user clicks as a percentage of the number of all user actions
10	Difference between a merchant's click-to-purchase conversion rate and the average of all the merchants

modelling process comprising four steps is as follows:

– Divide the training dataset D into five random and uniform copies to obtain D_{1-5};

– For the first learner M_1 in the first layer, four sub-datasets are randomly taken as the training data, the remaining copy is used as test data, and the new learner generated by learning from the training data is used to predict the test data to obtain the prediction results. Specifically, the first-fold cross-training of learner M_1 takes D_{1-4} as the training data and D_5 as the test data to obtain the prediction result P_{11} for the first-fold;

– Complete five-fold cross-validation for each first-layer trainer based on Step 2. Combine the results of the first layer to obtain the input P of the second layer;

– Perform five-fold cross-validation for the second layer of learners M according to Step 2, using the dataset as the output training set P of the first layer to obtain the final prediction model and results.

5.4 Comparison of Model Results

The AUC values of the classical, ensemble, and fusion models are shown in Table 7 (ranked in ascending order of the AUC values). On the one hand, comparing the five models used in this paper, we see that the Stacking fusion model has a significant improvement in the AUC value compared with the general model, with an increase of 0.01161 in absolute amount and 6.9% in the ranking ratio, which can reach top 2.8% of all the participants in Alibaba Tianchi, confirming that the fusion model constructed in this paper is effective. On the other hand, comparing the results of this paper with other studies using the same model, based on the same model, we achieve a higher AUC value, which confirms the effectiveness of our approach that embraces feature engineering and data processing. From the official baseline of 0.704954, there is still room for improvement in the results of this paper. However, after incorporating the innovations of feature engineering and data imbalance processing, even the classical model can reach the top 10% of the ranking.

5.5 Important Features

The prediction of repeat buyers can bring insights to merchants and platforms. Focusing on the features of potential repeat buyers further helps merchants and platforms understand user behaviour and adjust their business actions. We obtain the importance score of each feature through the LightGBM model, and the top ten features are shown in Table 8. The top ten features mainly concern the user profile (six features), merchant profile (two features), and user-merchant interaction profile (two features). These features provide guidance for management. Among them, user purchase features such as “number of purchases made by users at merchants”, “number of repeat purchases made by users before “Double Eleven””, and preference information such as “number of days users are active”, “proportion of user clicks to the number of all user actions”, and other characteristics reflect the user’s habit of using the e-commerce platform. Users with these characteristics may be potential repeat buyers, and managers should give timely and targeted measures. The characteristics of merchants such as “number of store categories” and “difference between the merchant’s click-to-purchase conversion rate and the average value of all the merchants” also affect user repurchase. So merchants should optimize their stores based on these key features, improve their strength and competitiveness to attract repurchase users.

6 Conclusion

The development of e-commerce platforms has so far resulted in a low percentage of new users in the market, so it is the future direction to focus on existing users of the platform and carry out targeted marketing strategies for repeat buyers. We use Tmall real user behaviour log data, and apply machine learning algorithms such as ensemble learning and fusion model to carry out repurchase prediction. Introducing feature engineering based on e-commerce characteristics and imbalance data processing, we construct a repurchase prediction model that contain a feature set of 94 features. The model attains an AUC value as high as 0.68406 after incorporating the fusion model, realizing efficient and accurate repeat buyer prediction.

We also make the following findings: (1) Compared with traditional machine learning models, the ensemble models and fusion models can improve the predictive effect of models, allowing us to predict repeat buyers more accurately and efficiently. (2) In the e-commerce context, especially when using real, large data sets, problems such as data imbalance and anomalous samples are inevitable, and solving these problems is important to attain accuracy in model prediction results. (3) It is necessary to adopt advanced and popular models, but the application of feature engineering should also be emphasized, and appropriate feature engineering can help bring the best model results.

We deduced important management insights from the study: (1) The study confirms the possibility of predicting repeat users, and by predicting it also captures important characteristics of potential users. Platforms should pay attention to users with these characteristics and guide them to repurchase through product personalized recommendations and SMS alerts; (2) the prediction model also contains merchant features. for store improvement, the current store operations can be checked based on the important features, and optimized and upgraded to better convert users into repeat buyers, such as diversifying their product range. (3) the merchant can develop separate strategies for repeat buyers and new users based on the prediction results, implementing differentiated marketing and pricing, and (4) predictions of repeat users indicate future orders, and merchants can adjust their product generation and inventory management based on the predictions. For example, when an item has a large number of potential repeat customers, production should be increased.

Due to the single data source and limited technology, our study has shortcomings. From a management perspective, future research can be improved in the following ways:(1) real-time modelling analysis should be conducted using real-time datasets from more platforms to broaden the platform sources, thus enhancing the generalizability of the model across different platforms and uncovering the differences in repeat buyers between platforms. (2) machine learning methods should be introduced into the feature construction process to improve the explanatory power of features to the problem.

Funding

Supported by Natural Science Foundation of China(71901027); Beijing Forestry University 2021 Course Ideological and Political Teaching, Research and Teaching Reform Project, "Management Model and Basic Decision-making" project(2021KCSZXY010)

Author contributions

The authors contributed equally to this work.

Conflict of interest

The authors declare no conflict of interest.

References

[1] Abel, F.; Gao, Q.; Houben, G. J.; Tao, K. (2011). Analyzing user modeling on twitter for personalized news recommendations. *International Conference on User Modeling, Adaptation, And Personalization*, 1-2, 2011

[2] Belem, F. M.; Silva, R. M.; de Andrade, C. M.; Person, G.; Mingote, F.; Ballet, R.; Alponti, H.; de Oliveira, H. P.; Almeida, J. M.; Goncalves, M. A. (2020). "Fixing the curse of the bad product descriptions"–Search-boosted tag recommendation for E-commerce products. *Information Processing Management*, 57(5), 102289, 2020

[3] Benevenuto, F.; Rodrigues, T.; Cha, M.; Almeida, V. (2009). Characterizing user behavior in online social networks. *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*, 49-62, 2009

[4] Bhattacharya, C. B. (1998). When customers are members: Customer retention in paid membership contexts. *Journal of The Academy of Marketing Science*, 26(1), 31-44, 1998

[5] Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2): 123-140, 1996

[6] Cao, W.; Wang, K.; Gan, H.; Yang, M. (2021). User online purchase behavior prediction based on fusion model of CatBoost and Logit. *Journal of Physics: Conference Series*, 2003(01), 012011, 2021

[7] Carta, S.; Fenu, G.; Recupero, D. R.; Saia, R. (2019). Fraud detection for E-commerce transactions by employing a prudential Multiple Consensus model. *Journal of Information Security and Applications*, 46, 13-22, 2019

[8] Chen, S.; Wang, J. Q.; Zhang, H. Y. (2019). A hybrid PSO-SVM model based on clustering algorithm for short-term atmospheric pollutant concentration forecasting. *Technological Forecasting and Social Change*, 146, 41-54, 2019

[9] Chou, P.; Chuang, H. H. C.; Chou, Y. C.; Liang, T. P. (2022). Predictive analytics for customer repurchase: Interdisciplinary integration of buy till you die modeling and machine learning. *European Journal of Operational Research*, 296(2), 635-651, 2022

[10] Daly, J. L. (2002). *Pricing for profitability: Activity-based pricing for competitive advantage.* John Wiley & Sons, 2002.

[11] Dasarathy, B. V.; Sheela, B. V.(1979). A composite classifier system design: Concepts and methodology. *Proceedings of the IEEE*, 67(5): 708-713, 1979

[12] Deng, Z. H.; Huang, L.; Wang, C. D.; Lai, J. H.; Philip, S. Y. (2019). Deepcf: A unified framework of representation learning and matching function learning in recommender system. *Proceedings of The AAAI Conference on Artificial Intelligence*, 33(01), 61-68, 2019

[13] Dong, J.; Huang, T.; Min, L.; Wang, W. (2022). Prediction of Online Consumers' Repeat Purchase Behavior via BERT-MLP Model. *Journal of Electronic Research and Application*, 6(3), 12-19, 2022

[14] Dong, X.; Yu, Z.; Cao, W.; Shi, Y.; Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, 14(2), 241-258, 2020

[15] Dong, Y.; Jiang, W. (2019). Brand purchase prediction based on time-evolving user behaviors in e-commerce. *Concurrency and Computation: Practice and Experience*, 31(1), e4882, 2019

[16] Enrich, M.; Braunhofer, M.; Ricci, F. (2013). Cold-start management with cross-domain collaborative filtering and tags. *International Conference on Electronic Commerce and Web Technologies* 101-112, 2013

[17] Fernández-Tobías, I.; Cantador, I. (2014). Exploiting Social Tags in Matrix Factorization Models for Cross-domain Collaborative Filtering. *Proceedings of the 1st Workshop on New Trends in Content-based Recommender Systems*, 34-41, 2014

[18] Gajsek B.; Dukic G.; Kovacic M.; Brezocnik M. (2021). A Multi-Objective Genetic Algorithms Approach for Modelling of Order Picking. *Int. Journal of Simulation Modelling*, 20(4), 719-729, 2021

[19] Hansen, L. K.; Salamon, P. (1990). Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 12(10): 993-1001, 1990

[20] Jacobs, R.; Jordan, M.; Nowlan, S.; Hinton G. (2014). Adaptive mixtures of local experts. *Neural Computation*, 3(1): 79-87, 1991

[21] Janekova J.; Fabianova J.; Kadarova J. (2021). Selection of Optimal Investment Variant Based on Monte Carlo Simulations. *Int. Journal of Simulation Modelling*, 20(2), 279-290, 2021

[22] Kagan, S.; Bekkerman, R. (2018). Predicting purchase behavior of website audiences. *International Journal of Electronic Commerce*, 22(4), 510-539, 2018

[23] Knezevic, B.; Skrobot, P.; Pavic, E. (2021). Differentiation of e-commerce consumer approach by product categories. *Journal of Logistics, Informatics and Service Science*, 8(1), 1-19, 2021

[24] Kocheturov, A.; Pardalos, P. M.; Karakitsiou, A. (2019). Massive datasets and machine learning for computational biomedicine: trends and challenges. *Annals of Operations Research*, 276(1), 5-34, 2019

[25] Koehn, D.; Lessmann, S.; Schaal, M. (2020). Predicting online shopping behaviour from clickstream data using deep learning. *Expert Systems with Applications*, 150, 113342, 2020

[26] Koren, Y. (2008). Factorization meets the neighborhood: a multifaceted collaborative filtering model. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 426-434, 2008

[27] Kumar, A.; Kabra, G.; Mussada, E. K.; Dash, M. K.; Rana, P. S. (2019). Combined artificial bee colony algorithm and machine learning techniques for prediction of online consumer repurchase intention. *Neural Computing and Applications*, 31(2), 877-890, 2019

[28] Kyriakou, I.; Mousavi, P.; Nielsen, J. P.; Scholz, M. (2021). Forecasting benchmarks of long-term stock returns via machine learning. /emphAnnals of Operations Research, 297(1), 221-240, 2021

[29] Li, X.; Hitt, L. M.; Zhang, Z. J. (2011). Product reviews and competition in markets for repeat purchase products. *Journal of Management Information Systems*, 27(4), 9-42, 2011

[30] Liu, X.; Li, J. (2016). Using support vector machine for online purchase predication. Emph2016 International Conference on Logistics, Informatics and Service Sciences, 1-6, 2016

[31] Ma X. Y.; Lin Y.; Ma Q. W. (2021). Data-Driven Robust Model for Container Slot Allocation with Uncertain Demand. *Int. Journal of Simulation Modelling*, 20(4), 707-718, 2021

[32] Martínez, A.; Schmuck, C.; Pereverzyev Jr, S.; Pirker, C.; Haltmeier, M. (2020). A machine learning framework for customer purchase prediction in the non-contractual setting. *European Journal of Operational Research*, 281(3), 588-596, 2020

[33] Moriuchi, E.; Takahashi, I. (2022). An empirical study on repeat consumer's shopping satisfaction on C2C e-commerce in Japan: the role of value, trust and engagement. *Asia Pacific Journal of Marketing and Logistics*, ahead-of-print, 2022

[34] Ni, Y.; Chen, X.; Pan, W.; Chen, Z.; Ming, Z. (2021). Factored heterogeneous similarity model for recommendation with implicit feedback. *Neurocomputing*, 455(2021), 59-67, 2021

[35] Oyewole, S. A.; Olugbara, O. O. (2018). Product image classification using Eigen Colour feature with ensemble machine learning. *Egyptian Informatics Journal*, 19(2), 83-100, 2018

[36] Sagi, O.; Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), e1249, 2018

[37] Sakar, C. O.; Polat, S. O.; Katircioglu, M.; Kastro, Y. (2019). Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Computing and Applications*, 31(10), 6893-6908, 2019

[38] Schapire, R. E.; Freund, Y. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1): 119-139, 1997

[39] Shen, Y.; Xu, X.; Cao, J. (2020). Reconciling predictive and interpretable performance in repeat buyer prediction via model distillation and heterogeneous classifiers fusion. *Neural Computing and Applications*, 32(13), 9495-9508, 2020

[40] Tripathi, P.; Singh, S.; Chhajer, P.; Trivedi, M. C.; Singh, V. K. (2020). Analysis and prediction of extent of helpfulness of reviews on E-commerce websites. *Materials Today: Proceedings*, 33, 4520-4525, 2020

[41] Van Nguyen, T.; Zhou, L.; Chong, A. Y. L.; Li, B.; Pu, X. (2020). Predicting customer demand for remanufactured products: A data-mining approach. *European Journal of Operational Research*, 281(3), 543-558, 2020

[42] Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, 5(2): 241-259, 1992

[43] Wu P. J., Yang D. (2021). E-Commerce Workshop Scheduling Based on Deep Learning and Genetic Algorithm. *Int. Journal of Simulation Modelling*, 20(1),192-200,2021

[44] Xu, J.; Kim, H.K. (2021). A study on the factors influencing consumers' purchase intention towards Chinese beauty industry: focusing on SNS characteristic elements. *Journal of Logistics, Informatics and Service Science*, 8(2), 47-64, 2021

[45] Yin, X. C.; Liu, C. P.; Han, Z. (2005). Feature combination using boosting. *Pattern Recognition Letters*, 26(14), 2195-2205, 2005

[46] Zhang, H.; Dong, J. (2020). Prediction of repeat customers on E-commerce platform based on blockchain. *Wireless Communications and Mobile Computing*, 2020(8841437), 2020

[47] Zhang, Z.; Zeng, D. D.; Abbasi, A.; Peng, J.; Zheng, X. (2013). A random walk model for item recommendation in social tagging systems. *ACM Transactions on Management Information Systems* 4(2), 1-24, 2013

[48] [Online]. Available: https://www.census.gov/retail/index.html

[49] [Online]. Available: https://www.cnnic.net.cn/n4/2022/0401/c88-1131.html

[50] [Online]. Available: https://tianchi.aliyun.com/competition/entrance/231576/introduction

[51] [Online]. Available: https://github.com/huiminren/RepeatBuyersPrediction

[52] [Online]. Available: https://github.com/leowang7553/repeatBuyersPrediction

[53] [Online]. Available: https://github.com/Ashitemaru/DM-Tmall-prediction

[54] [Online]. Available: https://github.com/DatAvalon/RepeatBuyersPrediction

Cite this paper as:
Zhang, M.; Lu, J.; Ma, N.; Cheng, T.C.E.; Hua G. (2022). A Feature Engineering and Ensemble Learning Based Approach for Repeated Buyers Prediction *International Journal of Computers Communications & Control*, 17(6), 4988, 2022.
https://doi.org/10.15837/ijccc.2022.6.4988