J. COMBIN. MATH. COMBIN. COMPUT. 127a (2025) 3057--3071

Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

A Study of Consumer Data-Driven Purchasing Behaviour Insight and Prediction in the Context of Digital Marketing

Lanlan Zhou[1,✉]

[1] *Sichuan University of Science & Engineering, Yibin, Sichuan, 644002, China*

ABSTRACT

Consumer data is an important support for analysing and observing consumer behaviours in the era of digital marketing, and constructing models to predict consumer purchasing behaviours. In this paper, we select the Retailrocket consumer behaviour dataset based on real shopping websites, analyse the distribution of various types of consumer behaviour over time and other data characteristics, and gain insights into the behavioural habits of consumers when shopping. Based on the XGBoost algorithm in machine learning, a prediction model of consumer behaviour is constructed, and the genetic algorithm is used to optimize and improve the XGBoost algorithm.The XGBoost prediction model has a significantly better prediction performance than the LSTM prediction model and the LR prediction model when facing the data under the under-sampling data balancing method and the improved random under-sampling method based on the K-means algorithm. . The performance of the GA-XGBoost prediction model optimised by the genetic algorithm is significantly improved compared to the XGBoost prediction model, and substantially better than the LSTM prediction model and the LR prediction model. The accuracy and F1 value of the GA-XGBoost prediction model in the data under the improved stochastic undersampling method are 0.90865 and 0.92435, respectively, which are improved by 14.69% and 17.26% relative to the XGBoost prediction model. Meanwhile, the stability of GA-XGBoost prediction model is also significantly improved compared to XGBoost prediction model.

Keywords: xGBoost algorithm, genetic algorithm, LSTM, K-means algorithm, purchase behaviour prediction

1. Introduction

With the continuous development of the Internet, e-commerce platforms are attracting more and more consumers to join the online consumption team with their time- and space-independent, stable and secure payment environment as well as diversified product services. Thanks to the development of

✉ Corresponding author.
E-mail address: zhoulan13785@163.com (L. Zhou).
Received 15 February 2024; Revised 25 March 2024; Accepted 15 December 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-173

data-driven technology, the implicit feedback data between e-commerce platforms and online consumers is constantly enriched, and huge amounts of user behaviour data are recorded and retained by e-commerce platforms. In-depth analysis of user behaviour and exploration of effective marketing strategies are of great significance for enterprises to grasp market opportunities and enhance competitiveness [1-3].

In the e-commerce environment, studying consumers' behavioural patterns and psychological mechanisms in the purchase decision-making process is the key to understanding consumers' needs, preferences, purchase motives, and decision-making process [4-5]. Since consumers' information processing, purchasing habits, and interactive feedback mechanisms on e-commerce platforms are different from those in traditional markets, this requires companies to re-examine and re-interpret multiple phases of consumer behaviour, such as information searching, choice evaluation, purchase decision-making, and post-purchase behaviours [6-9]. Especially in the context of digital and personalised marketing strategies, analysing user behavioural patterns using data-driven decision theory can help reveal consumer preferences and market trends. Data-driven decision theory refers to the use of data analysis to guide actions in the decision-making process, and the core of this theory lies in how to collect, process, and analyse a large amount of user data and how to formulate or adjust marketing strategies based on these data [10-12]. Data-driven decision-making theory not only provides a scientific approach to decision-making, but also emphasises a continuous learning and iterative process, whereby strategies are constantly optimised through data analysis to adapt to changes in the market and consumer behaviour [13-14].

Literature [15] outlines key algorithms for e-commerce platforms to provide consumers with personalised shopping services, pointing out that while personalised services can enable product recommendations and targeted promotions, excessive recommendations can also lead to negative outcomes such as resistance, and concern is raised about the ensuing privacy protection issues. Literature [16] points out that enterprises can use big data analytics to gain a deeper understanding of customer behaviours and industry trends, helping them to make the best decisions to improve sales, marketing, customer retention, etc., based on which they can design the best frame of reference to deal with online business, and build the most suitable solutions for their customers through a combination of various technologies. Literature [17] highlights the importance of big data analytics in the e-commerce market, based on behavioural expectancy theory and rational activity theory, the use of big data analytics to build a technical model to accomplish insights into consumer needs, prediction of behaviour, and matching of preferences, and the information collected collectively contributes to paradigm changes in the e-commerce industry. Literature [18] explored the role played by decision tree algorithms in predicting consumer purchasing behaviour, and found that the decision tree model of the random tree algorithm has a higher accuracy rate of user behaviour classification, and that re-adjusting the marketing plan based on the results of the analysis of this model will significantly improve the understanding of user behaviour, and increase the body of knowledge of the predictive analytics of e-commerce. Literature [19] constructed a multimodal learning framework incorporating multiple information sources to understand and predict consumer behaviour in e-commerce platforms, and the experimental results show that the multimodal learning approach has the performance of handling complex cross-border e-commerce data, and its effectiveness and superiority in capturing and analysing heterogeneous data. Literature [20] shows that it is vital to maintain a good relationship between customers and goods in e-commerce activities, in order to improve the accuracy of frequent itemset prediction and discovery of association rules in user personalised services, a data mining model is proposed, in addition to this, the Apriori algorithm is applied to provide recommendations based on the user's purchase history. Literature [21] constructed a consumer behaviour prediction model by using various machine learning and data analysis techniques, and illustrated how enterprises can guide their marketing strategies through the results of the model analysis through example analysis, which provides useful references and guidance for enterprises in the era of data-driven marketing. Literature [22] fused the logistic regression model and decision tree XGBoost model to construct an online shopping behaviour analysis and prediction system, and the experiment found that the hybrid model has a higher classification accuracy than a single model, and it can effectively screen out the user's behavioural characteristics.

In this paper, the Retailrocket open source dataset is selected as the consumer behaviour dataset for insight and prediction of consumer purchasing behaviour. The Retailrocket dataset is analysed in depth and the results are visualised to understand the overall distribution of the data, explore the distribution of various types of consumer behaviours over time, and explore the operating habits of consumers before they make purchasing decisions on shopping websites. We exclude dates that do not contain all-day data from the consumer behaviour data, and exclude consumers with less than 30 behaviours based on their activity level, to form the final dataset for subsequent research. A regular term is introduced into the loss function of the GBDT algorithm to construct the XGBoost algorithm, and the prediction model of consumer purchase behaviour is constructed based on this algorithm. Using the unique advantage of genetic algorithm to find the optimal solution to the XGBoost algorithm for hyper-parameter optimisation, forming the final model used for the prediction of consumer purchasing behaviour in this paper. Based on the insight results of the unbalanced characteristics of consumer behaviour data, this paper analyses the model prediction effect under the undersampled data balancing method and the improved random sampling method based on the K-means algorithm. Evaluation indexes such as accuracy rate and F1 value are used to analyse the prediction performance of different models in depth, and the comparison highlights the superior performance and smoothness of the GA-XGBoost prediction model of this paper in the prediction of consumer purchase behaviour.

2. Consumer data selection and processing

2.1. Introduction to the data set

The empirical part of this paper adopts the open source dataset Retailrocket, which is based on the consumer behaviour dataset of a real shopping website, and contains three underlying data tables, namely, the events table that records consumer behaviour, the category_tree table that records category information, and the item_properties table that records product properties. These data contain a total of 2870349 records of consumer behaviours and product-related information from August 1st to December 30th. Because of its large data volume and reliable data source, it has become a classic dataset in the field of consumer purchase behaviour insight and prediction.

2.2. Data analysis

In order to understand the overall distribution of the data, explore the distribution of various types of consumer behaviour over time, and dig deeper into the operating habits of consumers before they make a purchase decision on a shopping website, this section carries out an in-depth analysis of the Retailrocket dataset and presents the results visually.

The Retailrocket dataset records a total of 2870349 behavioural events of 1503847 unique visitors to the products. Figure 1 shows the frequency of the three types of behaviours in the dataset. B1-B3 represent the browsing, adding to cart and purchasing behaviours of consumers, respectively, and contain 2750182 browsing behaviours, 63293 adding to cart behaviours and 56874 purchasing behaviours. Browsing, adding to cart and purchasing behaviours account for 95.81%, 2.21% and 2.03% of the total behaviours respectively, with an average of about 48 times of browsing to produce a purchase. This shows that the purchase rate of users is extremely low, and there is an obvious imbalance in the data.

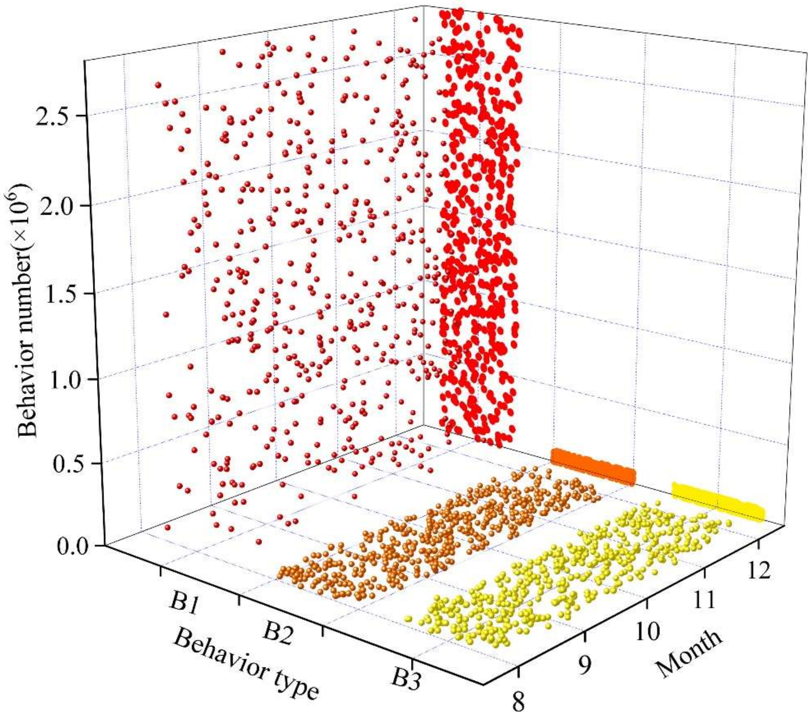

Fig. 1. Behavioral frequency statistics

Further analysis of the consumers' purchase profiles revealed that out of the 1503847 unique visitors included in the dataset, only 12,381 consumers had made a purchase. Figure 2 shows the distribution of the number of purchases made by consumers who had at least 1 purchase, the size of the circle in the figure indicates how much the percentage is, and 1-5 indicates the types of behaviours of purchasing 1, 2, 3, 4 and more than 4 times, respectively. Among all the consumers who have made purchases, the number of consumers who have only made 1 purchase is as high as 79.36%, and only about 20.64% of consumers have made 2 or more purchases. There are a large number of consumers in the data set who have never made a purchase.

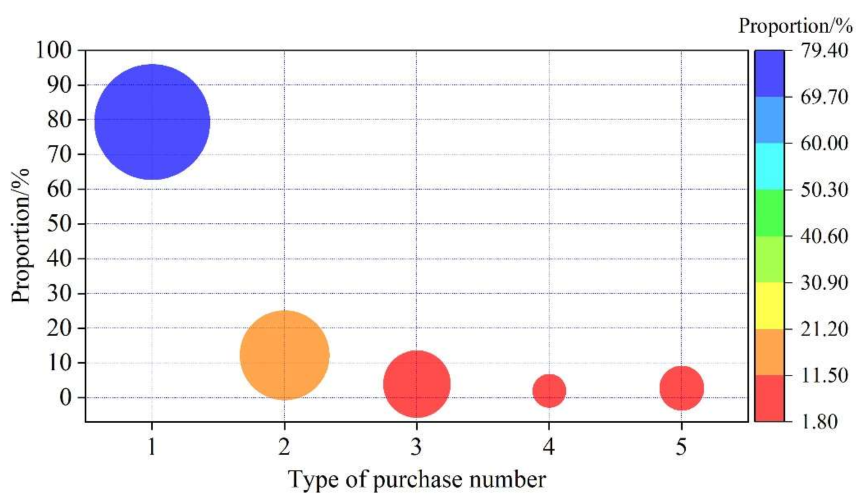

Fig. 2. Consumer purchase frequency distribution

In order to explore the distribution of the number of consumer behaviours over time, the distribution of the number of consumer behaviours per day is shown in Figure 3. The figure shows the number of consumer behaviours per day on the shopping website from 1 August to 31 December, with the horizontal coordinate on the time axis and the vertical coordinate on the number of behaviours. Since the number of daily browsing behaviours is much higher than other behaviours, the distribution of daily browsing behaviours is plotted separately. The graph reflects the daily traffic of the shopping website, and it can be seen that the number of browsing, adding to cart, and purchasing behaviours all show wave-shaped changes over time, and in a cycle of 10 days. In addition, the number of browsing, shopping cart adding and purchasing behaviours almost show the same relationship of increasing and decreasing, that is, the more the number of browsing behaviours and shopping cart adding behaviours, the more the number of purchasing behaviours.

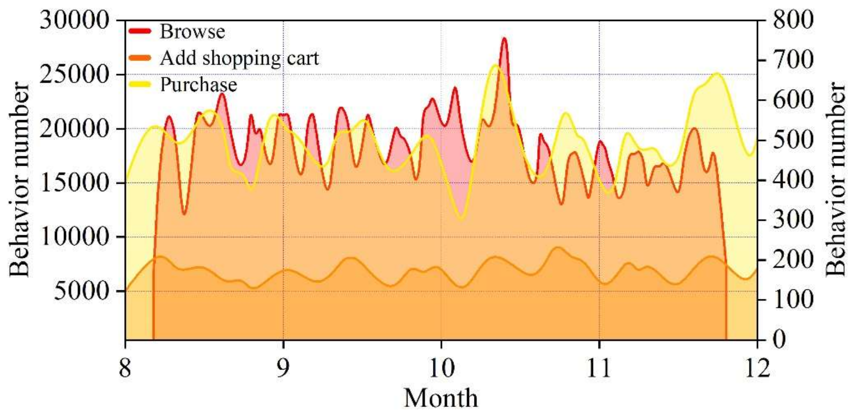

Fig. 3. Consumer daily behavior distribution

To further explore the behavioural habits of consumers before making a purchase decision, statistics on the number of times consumers browse a particular product before making a purchase decision on that product, the results are shown in Figure 4, where the size of the circle indicates the percentage of how many times, and 1-5 indicates the type of behaviour of browsing 1 time, 2 times, 3 times, 4 times and more than 4 times, respectively. As can be seen from Fig. 4, about 52.46% of the purchasing behaviours occurred after consumers browsed the products for 1 time, and only 10.41% of the purchasing behaviours were completed after more than 4 times of browsing. Therefore, we should pay attention to explore consumers' short-term interests and preferences.

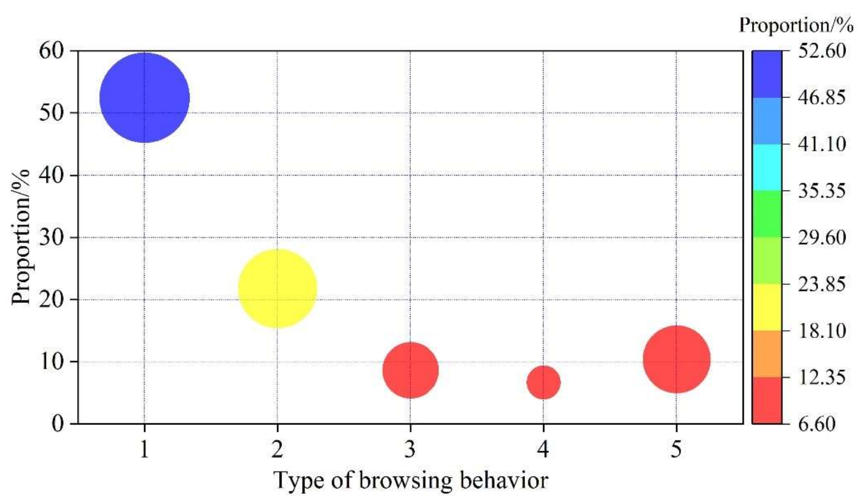

Fig. 4. The number of browsing behaviors before consumers buy goods

2.3. Data pre-processing

In the original data, the time of occurrence of consumer behavior is in the form of timestamps, which are converted to a common time-specific format, accurate to the second, for subsequent use. The data in this paper contains three data tables, each table is related to each other, so we need to merge the data to link the fields that contain valid information. As the events table records the behavioral logs of consumers, it is used as the main table of this paper, and the "itemid" field is associated with the consumer behavior events table and the category_tree table to obtain the relevant fields of the product category.

Observing the events table, we find that the data collection start and end dates are from August 1 to December 31, and we find that the consumer behavior data of August 1 and December 31 do not include the 24-hour data, and the data volume is less compared with other dates, so we exclude the data of these two days. After the exclusion, the dataset contains all-day data from August 2nd to December 30th.

The total number of consumer behaviors is counted, and the results are shown in Figure 5, where the size of the circle indicates the percentage, and 1-5 indicates the types of consumer behaviors with 1, 2, 3, 4 and more than 4 interactions, respectively. It can be seen that 73.38% of consumers have only 1 interaction during the data counting period, and about 6.07% of consumers have total number of behaviors more than 4 times, there are a large number of consumers with very low number of behaviors in the data set.

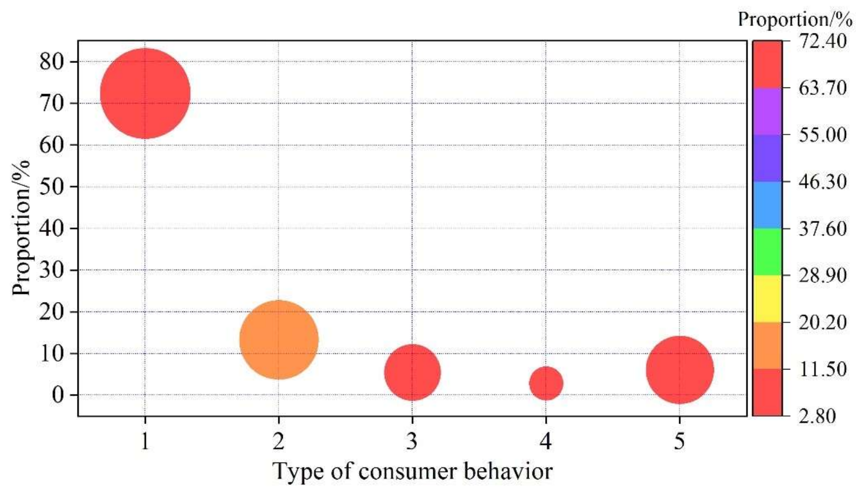

Fig. 5. The total frequency of consumer behavior is distributed

Generally speaking, the total number of consumer behaviours can be used to measure the degree of consumer activity: the higher the total number of consumer behaviours, the higher the degree of consumer activity, and vice versa, the lower the degree of activity. The total number of consumer behaviours is compared with the classification threshold, if the total number of consumer behaviours is greater than the threshold, then it is considered to be a highly active consumer, otherwise it is a consumer leap user, and the browsing and purchasing conversion rates of low and highly active consumers are calculated respectively, and the results are shown in Figure 6. Comprehensive Figure 5 and Figure 6 can be seen, there are a large number of low active consumers in the dataset, the purchase rate of these consumers is extremely low, it is not very meaningful to predict the purchase behaviour of these consumers. Therefore, this paper excludes the consumers whose total number of behaviours is lower than 30 times, and retains the behavioural records of 6714 independent consumers after exclusion.

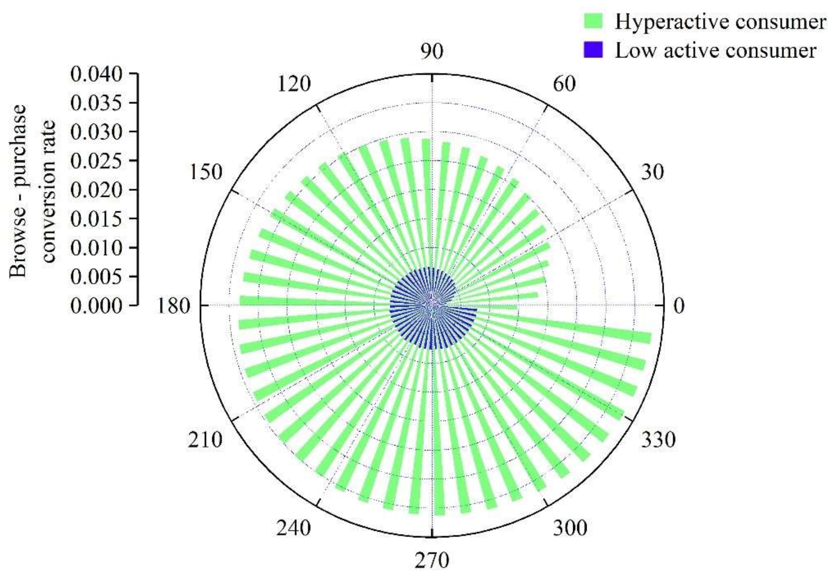

Fig. 6. Browse-purchase conversion rate for different activity

3. Purchase behaviour prediction algorithms

By selecting and visualising and analysing consumer behaviour data, this paper initially completes the work of insight and analysis of consumer buying behaviour based on consumer data. Next, this paper will focus on the prediction of consumer purchase behaviour with the help of the XGBoost algorithm in machine learning and its optimization using genetic algorithm.

3.1. XGBoost Algorithm

XGBoost is an improved algorithm of GBDT.The objective function of GBDT is a loss function and the objective function of XGBoost is a loss function plus a regular term, which is used to reduce the complexity of the model and to prevent model overfitting [23]. The principle is that given n instance, m features in the case of set $D=\{(x_i,y_i)\}\left(|D|=n; x_i \in R^m, y_i \in R\right)$, the integrated model of XGBoost is equation (1):

$$\hat{y}_i=\sum_{k=1}^{K} f_k(x_i), f_k \in \mathrm{F} \tag{1}$$

where $f_k(x_i)$ is the predicted value of the sample instance and the objective function of XGBoost is a loss function plus a regular term:

$$Obj=\sum_{i=1}^{n} l(y_i,\hat{y}_i)+\sum_{k=1}^{K}\Omega(f_k) \tag{2}$$

Rewrite the objective function as equation (3):

$$Obj^{(t)}=\sum_{i=1}^{n} l\left(y_i,\hat{y}_i^{(t-1)}+f_i(x_i)\right)+\Omega(f_t)+cons\tan t \tag{3}$$

XGBoost uses the Taylor series approximation of the objective function to minimise the objective function, so the objective function is approximated as equation (4), and finally we can get the new objective function equation (5):

$$Obj^{(t)} \square \sum_{i=1}^{n}\left[l\left(y_i,\hat{y}_i^{(t-1)}\right)+gf_i(x_i)+\frac{1}{2}hf_t^2(x_i)\right]+\Omega(f_t)+cons\tan t \tag{4}$$

$$Obj^{(i)} \approx \sum_{i=1}^{n}\left[gf_i(x_i) + \frac{1}{2} hf_i^2(x_i) \right] + \Omega(f_i)$$
$$g = \partial_{ij} l\left(y_i, \hat{y}_i^{(t-1)}\right), h_i = \partial_i^2 L\left(y_i, \hat{y}_i^{(t-1)}\right) \tag{5}$$

Firstly, $f_k(x_i)$ is reduced to the corresponding leaf node of each tree multiplied by the weight of the corresponding leaf node, i.e., $\omega_q(x_i)$. Then Eq. (6) is substituted into Eq. (5) to obtain Eq. (7), and finally Eq. (7) for traversal of the samples is transformed into Eq. (8) for traversal of the leaf nodes:

$$\Omega(f_i) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega^2 \tag{6}$$

$$Obj \approx \sum_{i=1}^{n}\left[g\omega_q(x_i) + \frac{1}{2} h_i \omega_q^2(x_i) \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{7}$$

$$Obj = \sum_{j=1}^{T}\left[\left(\sum_{i=1} g_i \right)\omega_j + \frac{1}{2}\left(\sum_{i<h} h_i + \lambda \right)\omega_j^2 \right] + \gamma T \tag{8}$$

Where γ represents the penalty strength, T represents the number of leaf nodes, ω represents the leaf weights, q represents the tree structure of each tree mapping strength samples to the corresponding leaf nodes, and $q(x_i)$ represents the number of leaf nodes.

The objective function is rewritten as a one-quadratic function with respect to the leaf weights ω and the vertex formula is used to find the optimal ω and the value of the objective function, see Eqs. (9) and (10):

$$\omega_j^* = \frac{G_i}{H_j + \lambda} \tag{9}$$

$$Obj = -\frac{1}{2}\sum_{j=1}^{T} \frac{G_i^2}{H_j + \lambda} + \gamma T \tag{10}$$

3.2. GA-XGBoost algorithm

Genetic algorithm is a stochastic adaptive global search algorithm that draws on Darwinian evolutionary theory and has the unique advantage of finding optimal solutions, which is effective in the field of algorithmic parameter search [24]. The process of genetic algorithm for XGBoost hyperparameter search is shown in Fig. 7. There are selection operator, crossover operator and variation operator in genetic algorithm. The selection operator is to select the chromosomes in the population according to certain logic or rules to get a population that can be crossed. Crossover operator is to determine the genetic rules of two crossover chromosomes and to determine the genetic sequence of the chromosomes of the progeny of a cross between two chromosomes. The mutation operator is to change certain chromosomal genes with a certain probability and pass them directly to the next generation of the population. The method allows the generation of new individuals and provides conditions for finding the optimal solution.

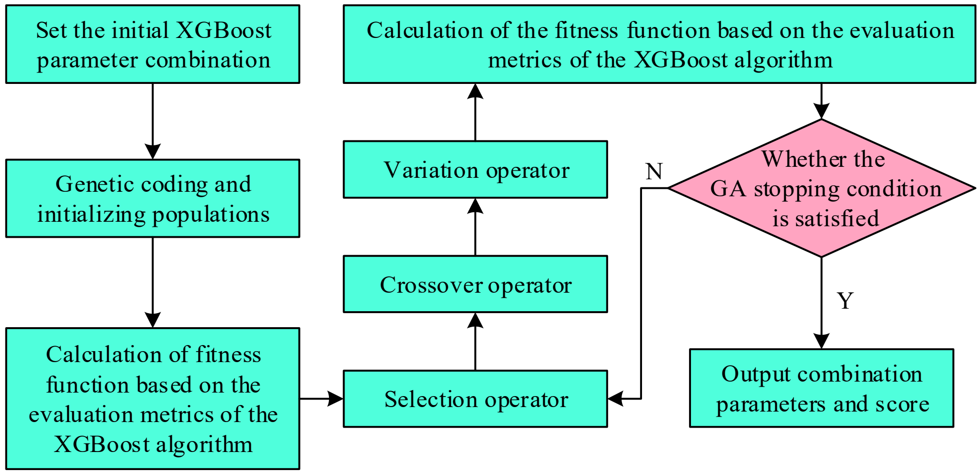

Fig. 7. GA optimizes the XGBoost hyperparameter

In this paper, three hyperparameters commonly used in XGBoost (respectively: learning_rate, number of trees n_estimators, and maximum depth max_depth) are selected as population inputs to the genetic algorithm, and the evaluation metric AUC of the classification algorithm is used as the fitness function for population evolution through selection, crossover, and mutation. The fitness function is recalculated to determine whether the loop termination condition is satisfied. Since the upper bound of the fitness function is difficult to set, the number of iterations is used to terminate the algorithm cycle.

4. Consumer purchasing behaviour forecasting studies

4.1. Experimental environment

The experimental environment of this paper contains hardware and software environments, in which the hardware environment is: (1) CPU: Intel(R) Core(TM) i5-3230M,2.6GHz.(2) Graphics card: NVIDA(R) GERORCE(R) 740M.(3) Memory: 12GB.(4) Software environment is: (1) Operating system: Windows 10. (2) Development language: Python. (3) Development environment: Python 3.6.5, Anaconda3, Spyder. (4) Toolkit: Pandas, scikit-learn, numpy.

4.2. Indicators for model evaluation

In the evaluation of the model prediction effect is generally selected as the evaluation index of accuracy, but in the multi-classification prediction scenario is more concerned about the prediction effect of one of the classes. For example, in the prediction of consumer purchasing behaviour, this paper pays more attention to the effect of consumer purchasing behaviour prediction, while the effect of non-purchasing behaviour prediction can be disregarded, so this paper takes the precision rate, the recall rate, and the F1 value of the reconciliation average of the two as the criteria for model prediction evaluation.

Consumer purchase behaviour prediction is actually a binary classification problem, and the final prediction result is that the consumer buys or the user does not buy. Table 1 shows the form of the confusion matrix for the prediction results of the binary classification problem, where the columns (Positive/Negative) correspond to the results predicted by the model, while the rows (True/False) indicate the true values of the data. Explanations for the prediction scenarios in this paper: the TR is the number that the model predicted as purchased and actually purchased as well, FP is the number that the model predicted as purchased and actually did not purchase, TN is the number that the model predicted as not purchased but actually purchased, and FN is the number that the model predicted as not going to purchase and actually did not purchase as well.

Table 1. Confusion matrix

	Positive	Negative
True	TP	TN
False	FP	FN

Precision rate is the rate of correct consumer purchases predicted by the model. Recall is the ratio of correctly predicted consumer purchases to all actual purchases, which reflects the hit rate of the model's predictions. Equations (11) and (12) show how precision rate and recall rate are calculated:

$$\Pr ecision = \frac{TP}{TP+FP} \tag{11}$$

$$\operatorname{Re} call = \frac{TP}{TP+TN} \tag{12}$$

Precision and recall are two important indicators for evaluating a model, and a good model should have both precision and recall at a relatively high level. Equation (13) shows the calculation of the F1 value of the reconciled mean of precision and recall:

$$F1 = 2 * \frac{\Pr ecision * \operatorname{Re} call}{\Pr ecision + \operatorname{Re} call} \tag{13}$$

4.3. Analysis of model prediction effects

4.3.1. Effectiveness of model prediction under undersampled data balancing methods. Through the analysis of the experimental dataset found that there is a serious imbalance in the proportion of positive and negative samples in the training set, in the analysis of the training dataset after the data division found that the original data samples without adding the data balancing strategy of the data of the positive and negative sample ratio of 0.0031, the training set of positive and negative samples is seriously imbalanced, for this reason this paper uses the undersampling balancing method to balance the data found that the positive and negative samples of 0.008 when the model prediction The effect is better. Table 2 shows the prediction effect of XGBoost prediction model, LSTM prediction model and LRR prediction model under the undersampling method. Comparing the prediction effect of each model under the undersampling data balancing method, the LR prediction model is more sensitive to the imbalance of sample category distribution, and the prediction F1 value of XGBoost model is 0.72408, which is significantly higher than that of the other two prediction algorithms, which indicates that the algorithm based on the integration of boosting is very good for the data imbalance of the dataset.

Table 2. The model prediction effect of under-sampling data equilibrium method

Index	XGBoost	LSTM	LR
TP	4586	2308	986
FP	5378	5629	5814
TN	4736	5172	5483
ACC	0.70849	0.63278	0.60396
Recall	0.65874	0.56531	0.52085
F1	0.72408	0.65702	0.61443

4.3.2. Model Prediction Effect under Improved Stochastic Under-Sampling Methods. Random sampling method is the simplest and most widely used sampling technique to solve the current data imbalance problem [25]. As the randomness of deleting samples in random undersampling method will lead to the loss of information of most classes of samples, for the serious problem of data information loss in random undersampling data balancing method, this paper proposes an improved random undersampling method based on K-means algorithm on the basis of random undersampling method. In order to find out the best k value of the improved random undersampling data balancing method based on K-means algorithm, the data balancing effect of the improved random undersampling balancing method based on K-means algorithm with different k values is compared. Fig. 8 shows the predicted F1 values of each model for different k values of K-means algorithm based undersampling balancing method. From Fig. 8, it can be found that the data balancing effect is significantly improved by adding the K-means algorithm for improvement in the stochastic undersampling balancing method. Comparing the F1 values of XGBoost prediction model, LSTM prediction model and LR prediction model for the improved stochastic undersampling balancing method with different values of k, the predicted F1 values of each model are higher when the value of k is 35. Table 3 shows the prediction results of each model after adding the improved stochastic undersampling balancing method with a k value of 35. From the table, it can be seen that the F1 of each model is improved after applying the improved stochastic undersampling method, in which the F1 value of the XGBoost model is improved from 0.72408 under the undersampled data balancing method to 0.78794, and the F1 value of the XGBoost model remains the highest among the three models.

Table 3. Prediction effect of models under the improving random sampling method

Index	XGBoost	LSTM	LR
TP	5736	2587	1104
FP	4573	5064	5538
TN	4096	4487	4952
ACC	0.76852	0.75748	0.72317
Recall	0.78953	0.77089	0.73463
F1	0.78794	0.76243	0.71879

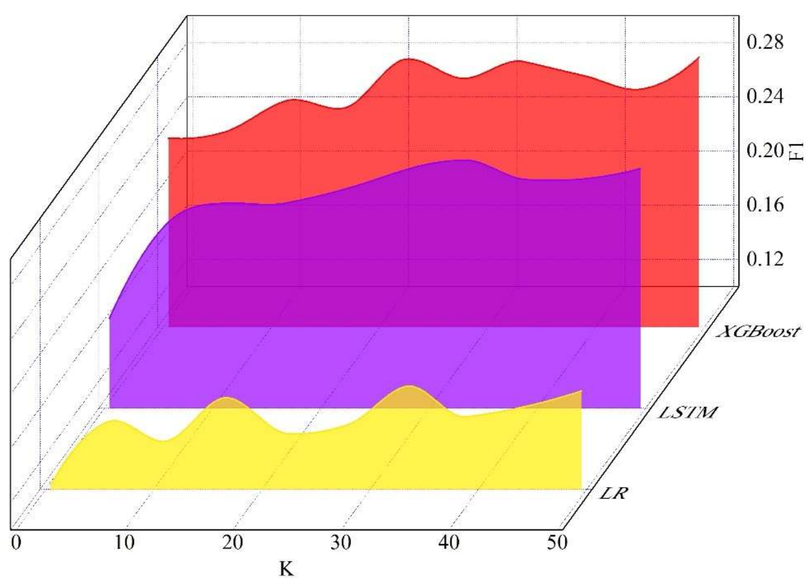

Fig. 8. F1 values of the model under different k values

4.3.3. GA-XGBoost model prediction effect. In order to improve the accuracy of the prediction model in predicting consumer purchasing behaviour. In this paper, we optimize the XGBoost model with the unique advantage of the optimal solution of the genetic algorithm, and use the GA-XGBoost model to predict the consumer's purchasing behaviour under the improved random undersampling data balancing method based on the K-means algorithm. Figure 9 shows the prediction effect of GA-XGBoost model and XGBoost model, LSTM model and LR model, and the lines in the figure reflect the trend of the evaluation indexes of each model.The prediction effect of GA-XGBoost model has been greatly improved relative to the other models, and the accuracy and F1 value of the GA-XGBoost model are 0.90865 and 0.92435, respectively. Which were improved by 14.69% and 17.26% relative to the XGBoost prediction model.

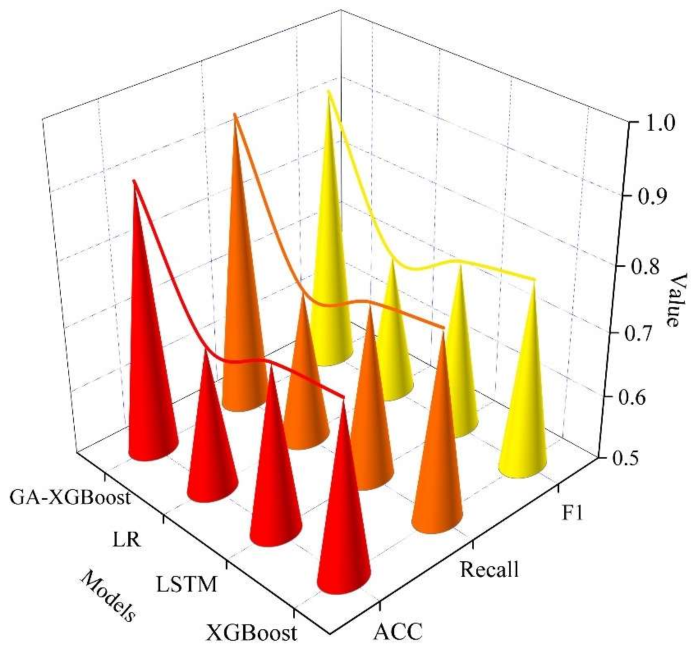

Fig. 9. Model prediction effect contrast

In order to further compare the stability of each model, five test datasets were divided to test each prediction model separately for the five predicted months of the test dataset, and the prediction results are shown in Fig. 10. The best prediction effect of the three prediction models before optimisation is the XGBoost prediction model, but the prediction F1 value of the XGBoost prediction model for different test sets varies greatly, while the prediction stability of the LSTM prediction model and the LR prediction model is better. Comparing the prediction effect of XGBoost prediction model, GA-XGBoost model has better prediction effect and prediction stability than XGBoost prediction model.

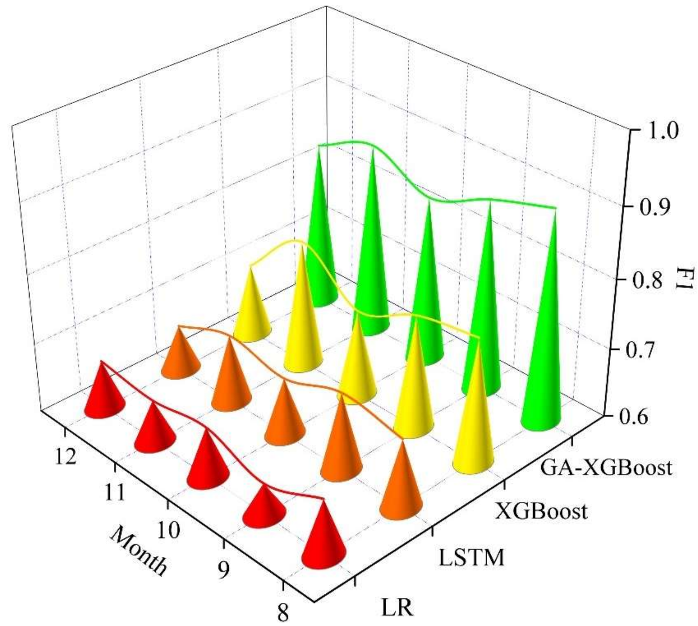

Fig. 10. Model prediction stability contrast

5. Conclusion

In this paper, the consumer behaviour dataset based on real shopping websites is selected to gain insight into the characteristics of consumer purchasing behaviour. With the help of genetic algorithm, XGBoost algorithm is optimized with hyperparameters to construct the consumer purchase behaviour prediction model in this paper. The F1 value of the XGBoost prediction model under the undersampled data balancing method is 0.72408, which is significantly higher than that of the LSTM prediction model and the LR prediction model. It indicates that the algorithm based on boosting integration performs better for data unbalanced datasets. The improved random sampling method based on K-means significantly improves the prediction performance of each model. The F1 value of the XGBoost prediction model increases from 0.72408 under the under-sampled data balancing method to 0.78794 with a k value of 35, which is 8.82%, and the prediction performance of the LSTM prediction model and the LR prediction model are also improved. The GA-XGBoost prediction model has the best performance in the data under the improved random sampling method, with the accuracy and the F1 value of the model prediction. Accuracy and F1 value of 0.90865 and 0.92435, respectively, which are improved by 14.69% and 17.26% relative to the XGBoost prediction model. In addition, the stability of the GA-XGBoost prediction model is equally superior to other models. It shows that the consumer purchase behaviour prediction model constructed in this paper can fully meet the needs of practical applications.

References

[1] Riwayat, A. A. P., Susilawati, A. D., & Naqiah, Z. (2024). Purchasing Patterns Analysis in E-commerce: A Big Data-driven Approach and Methodological. International Journal Software Engineering and Computer Science (IJSECS), 4(1), 148-164.

[2] Painuly, S., Sharma, S., & Matta, P. (2021, July). Big data driven e-commerce application management system. In 2021 6th international conference on communication and electronics systems (ICCES) (pp. 1-5). IEEE.

[3] Saleem, H., Uddin, M. K. S., Habib-ur-Rehman, S., Saleem, S., & Aslam, A. M. (2019). Strategic data driven

approach to improve conversion rates and sales performance of e-commerce websites. International Journal of Scientific & Engineering Research, 10(4), 588-593.

[4] Al-Ebrahim, M. A., Bunian, S., & Nour, A. A. (2023). Recent Machine-Learning-Driven Developments in E-Commerce: Current Challenges and Future Perspectives. Engineered Science, 28, 1044.

[5] Necula, S. C. (2023). Exploring the impact of time spent reading product information on e-commerce websites: A machine learning approach to analyze consumer behavior. Behavioral Sciences, 13(6), 439.

[6] Abakouy, R., En-naimi, E. M., Haddadi, A. E., & Lotfi, E. (2019, October). Data-driven marketing: How machine learning will improve decision-making for marketers. In proceedings of the 4th international conference on Smart City Applications (pp. 1-5).

[7] Shrirame, V., Sabade, J., Soneta, H., & Vijayalakshmi, M. (2020, July). Consumer behavior analytics using machine learning algorithms. In 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT) (pp. 1-6). IEEE.

[8] Zineb, E. F., Najat, R. A. F. A. L. I. A., & Jaafar, A. B. O. U. C. H. A. B. A. K. A. (2021). An intelligent approach for data analysis and decision making in big data: a case study on e-commerce industry. International Journal of Advanced Computer Science and Applications, 12(7).

[9] Balyemah, A. J., Weamie, S. J., Bin, J., Janda, K. V., & Joshua, F. J. (2024). Predicting Purchasing Behavior on E-Commerce Platforms: A Regression Model Approach for Understanding User Features that Lead to Purchasing. International Journal of Communications, Network and System Sciences, 17(6), 81-103.

[10] Fabra, J., Álvarez, P., & Ezpeleta, J. (2020). Log-based session profiling and online behavioral prediction in E–Commerce websites. IEEE Access, 8, 171834-171850.

[11] Rofi'i, Y. U. (2023). Analysis of E-Commerce Purchase Patterns Using Big Data: An Integrative Approach to Understanding Consumer Behavior. International Journal Software Engineering and Computer Science (IJSECS), 3(3), 352-364.

[12] Tan, K. S., & Subramanian, P. (2019). Proposition of machine learning driven personalized marketing approach for E-commerce. Journal of Computational and Theoretical Nanoscience, 16(8), 3532-3537.

[13] Sirohi, P., Singhal, N., Rizvi, S. V. A., & Kumar, P. (2023, July). A Reactive Approach for High-Accuracy and Data-Driven Customer Behaviour Analysis and Prediction. In International Conference on Artificial-Business Analytics, Quantum and Machine Learning (pp. 55-66). Singapore: Springer Nature Singapore.

[14] Li, L., Ma, S., Han, X., Zheng, C., & Wang, D. (2021). Data-driven online service supply chain: a demand-side and supply-side perspective. Journal of Enterprise Information Management, 34(1), 365-381.

[15] Wei, L., & Xia, Z. (2022). Big Data-Driven Personalization in E-Commerce: Algorithms, Privacy Concerns, and Consumer Behavior Implications. International Journal of Applied Machine Learning and Computational Intelligence, 12(4).

[16] Ghandour, A. (2015). Big Data driven e-commerce architecture. International Journal of Economics, Commerce and Management, 3(5), 940-947.

[17] Lv, H. (2022). Intelligent e-commerce Framework for Consumer Behavior Analysis Using Big Data Analytics. Advances in Data Science and Adaptive Analysis, 14(03n04), 2250007.

[18] Mustakim, N. A., Aziz, M. A., & Rahman, S. A. (2024). Predicting Consumer Behavior in E-Commerce Using Decision Tree: A Case Study in Malaysia. Information Management and Business Review, 16(3 (I)), 201-209.

[19] Zhang, X., & Guo, C. (2024). Research on Multimodal Prediction of E-Commerce Customer Satisfaction Driven by Big Data. Applied Sciences, 14(18), 8181.

[20] Alghanam, O. A., Al-Khatib, S. N., & Hiari, M. O. (2022). Data mining model for predicting customer purchase behavior in e-commerce context. International Journal of Advanced Computer Science and Applications, 13(2).

[21] Wan, L., & Xu, J. (2024). Data-driven marketing strategy and consumer behavior prediction model construction. Information Systems and Economics, 5(2), 73-79.

[22] Liu, C. J., Huang, T. S., Ho, P. T., Huang, J. C., & Hsieh, C. T. (2020). Machine learning-based e-commerce platform repurchase customer prediction model. Plos one, 15(12), e0243105.

[23] Shijie Qian,Tian Peng,Zihan Tao,Xi Li,Muhammad Shahzad Nazir & Chu Zhang. (2024). An evolutionary deep learning model based on XGBoost feature selection and Gaussian data augmentation for AQI prediction. Process Safety and Environmental Protection(PA),836-851.

[24] Li Xiangcheng,Wang Jialong,Geng Zhirui,Jin Yang & Xu Jiawei. (2023). Short-term Wind Power Prediction Method Based on Genetic Algorithm Optimized XGBoost Regression Model. Journal of Physics: Conference Series(1),

[25] Anoop Kumar,Shashi Bhushan,Manahil SidAhmed Mustafa,Ramy Aldallal,Hassan M. Aljohani & Fatimah A. Almulhim. (2024). Novel imputation methods under stratified simple random sampling. Alexandria Engineering Journal236-246.