RESEARCH-ARTICLE

Machine Learning–Based Random Forest Prediction of Online Shopping Behavior in the Digital Economy

TIAN WANG, Guangdong University of Foreign Studies South China Business College, Guangzhou, Guangdong, China

JIANBANG LIN, Nanfang College of Sun Yet-sen University, Guangzhou, Guangdong, China

YINGPENG ZHANG, Nanfang College of Sun Yet-sen University, Guangzhou, Guangdong, China

JIAJING ZHANG, Guilin Tourism University, Guilin, Guangxi, China

Published: 25 April 2025

DEIS 2025: 2025 International Conference on Digital Economy and Information Systems
April 25 - 27, 2025
Guangzhou, China

DEIS '25: Proceedings of the 2025 International Conference on Digital Economy and Information Systems (April 2025)
https://doi.org/10.1145/3745133.3745186
ISBN: 9798400714375

Machine Learning–Based Random Forest Prediction of Online Shopping Behavior in the Digital Economy

Tian Wang
Guangdong University of Foreign Studies South China Business College
Guangzhou, Guangdong, China
204220@gwng.edu.cn

Jianbang Lin*
Nanfang College Guangzhou
Guangzhou, Guangdong, China
linjb@nfu.edu.cn

Yingpeng Zhang
Nanfang College Guangzhou
Guangzhou, Guangdong, China
gguadd0@gmail.com

Jiajing Zhang
Guilin Tourism University
Guilin, Guangxi, China
lzzs@glit.cn

Abstract

Against the backdrop of the continuous development of the digital economy, consumers' online shopping behaviors have become increasingly complex. Predicting consumers' purchasing behavior has become a major challenge. Based on the e-commerce behavior data in February 2021, this study constructed a consumer online purchase prediction model using Random Forest. The performance of prediction model for out-of-sample data has been enhanced through rigorous data cleaning, feature engineering, and class balancing. The accuracy and recall exceed the 0.90, and the F1-score exceeds 0.80. Exploratory data analysis has demonstrated to reveal behavioral patterns related to time and price. Furthermore, this study has proposed significant strategies for the enhancement of conversion of shopping carts into purchases. Feature importance analysis in this study has found that the four important factors influencing purchase intention are as follows: cart_event_ratio, total_events, cart_per_product and view_count. This study proposes a practical, data-driven framework for e-commerce practitioners to predict consumers' online purchasing behavior more accurately. Furthermore, the conversion strategies for online shopping will be effectively improved in the highly competitive digital economy.

CCS Concepts

• **Applied computing** → Electronic commerce; Online shopping.

Keywords

Machine learning, E-commerce, Purchase Prediction, Random Forest algorithm

ACM Reference Format:
Tian Wang, Jianbang Lin, Yingpeng Zhang, and Jiajing Zhang. 2025. Machine Learning–Based Random Forest Prediction of Online Shopping Behavior in the Digital Economy. In *2025 International Conference on Digital Economy and Information Systems (DEIS 2025), April 25–27, 2025, Guangzhou, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3745133.3745186

*Corresponding author.

DEIS 2025, Guangzhou, China

ACM ISBN 979-8-4007-1437-5/2025/04
https://doi.org/10.1145/3745133.3745186

1 Introduction

The digital economy has developed rapidly, and this has led to changes in consumption patterns and business models. In particular, the well-developed mobile Internet and logistics infrastructure have greatly improved the efficiency and convenience of online shopping. According to a report by Shopify, global e-commerce sales will reach USD 6.09 trillion, up 8.4% from last year [1]. E-commerce has become one of the important pillars of the digital economy. However, consumers' online shopping behavior is a complex process that includes browsing, adding to cart, and purchasing. The accuracy of predicting consumers' online purchasing behavior has become more and more difficult. As a result, it becomes a most important for improving the efficacy of prediction models of consumers' online purchasing behavior.

In the past, scholars have used various methods such as Logistic Regression, Decision Tree, and Artificial Neural Network to predict consumer behavior [2-4]. However, as data complexity increases, these methods suffer from limitations such as overfitting and low interpretability. In contrast, the ensemble methods of machine learning have stronger robustness and adaptability. Among them, the Random Forest algorithm is notable for its ability to effectively integrate multiple decision trees, reduce overfitting, and perform well on heterogeneous datasets. Empirical studies have confirmed its advantages in improving the accuracy of consumer behavior prediction, and it becomes widely used in online purchase prediction [5-6].

This study proposes a consumers' online purchasing behavior prediction model by using Random Forest. The innovation of this study is a feature engineering process based on the intensity and temporal dynamics of consumer behavior, and a model optimization process in order to solve the problem of high-dimensional data and data imbalance. This study enables the implementation of a highly accurate predictive model for consumers' online purchasing behavior within the e-commerce industry. The main contributions are the practice of systematic feature engineering, the deployment of prediction model optimization, and the enhancement of the digital economy.

2 Literature Review

Most of the past research on predicting consumers' online purchasing behavior has been widely adopted by modeling methods such as Logistic Regression, Decision Tree or Artificial Neural Network [7–9]. While these methods can accomplish the construction of prediction models, they have certain limitations. For example, the application of Logistic Regression can be limited by the linear divisibility of the data, while a single Decision Tree is prone to overfitting [10]. Despite the high prediction accuracy of Artificial Neural Network, it often falls into a black box and cannot easily explain consumer behavioral characteristics. As behavioral data in e-commerce becomes more complex, it becomes harder for these methods to provide better prediction performance [11].

To improve the aforementioned limitations, recent researches have started to apply machine learning methods to construct prediction models, especially ensemble learning. Random forests have been proven to reduce variance, handle missing values, and perform well on heterogeneous datasets [12]. Chaubey et al. demonstrated that the ensemble model is better than independent classifiers [13]. Zhang's hybrid GBDT-logistic model improves performance on imbalanced datasets [14]. Meanwhile, Random Forest consistently provides highly accurate prediction of consumers' online shopping behavior, and it is a suitable choice for behavioral modeling and precision marketing [15]. These findings demonstrate the potential of the Random Forest method in dealing with increasingly complex e-commerce data.

3 Methods

3.1 Research Design

The goal of this study is to construct a high-precision online shopping purchase behavior prediction model. The machine learning modeling workflow is shown in Figure 1. First, the research problem is defined through a literature review. Next, the representative raw data is obtained from an e-commerce platform. A comprehensive exploratory data analysis and cleaning process is then performed to extract key information. Then, the feature engineering and data balancing processes are applied to address feature redundancy, variable transformation, and class imbalance. Next, the prediction models are constructed separately using Random Forest and Decision Tree algorithms. The both prediction models are compared using a confusion matrix. Finally, the study identifies the optimal model and verifies its applicability and accuracy in practical applications.

3.2 Data and Feature Engineering

The dataset for this study is derived from the open-source "eCommerce events history in electronics store". This is provided by the REES46 marketing platform under the CDP project. The data records the online shopping behavior of consumers on the e-commerce platform from February 1 to February 14, 2021. After cleaning the data and removing variables with high missing rates, seven valid original variables with a total of 82,688 samples were kept in this study. The dataset has a sufficient sample size, high-quality data, and stable behavior to provide strong support for model development and validation.

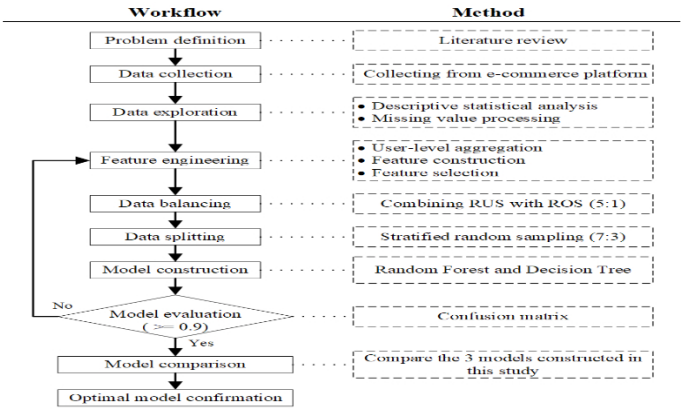

Figure 1: Research methodology flowchart

Feature engineering is used to improve predictive performance by constructing or selecting significant variables. First, event-level records are aggregated by user_id into 38,361 user-level profiles. Exploratory data analysis provides insights into consumer activity, time intensity, and purchase intent. The new meaningful predictor variables for the model are initially constructed. Then, in the feature selection stage, correlation analysis, mutual information, and SelectKBest (f_classif, $p < 0.05$) are applied to select the influential predictor variables [16]. Finally, the dependent variable (purchase) is binary (where view and cart events are categorized as non-purchase). The selected characteristics serve as independent variables (see Table 1).

3.3 Data Balancing and Data Splitting

The dataset is clearly imbalanced. The non-purchase sample is far outnumbering the purchase sample. The non-purchase sample is 94.49%. The purchase sample is 5.51%. If this problem isn't solved efficiently, the model will favor the majority of categories. This leads to a seemingly high accuracy in predicting purchase behavior, but a poor prediction in reality. To address the imbalance, random under-sampling (RUS) and over-sampling (ROS) were used to adjust the ratio of non-purchases to purchases to 5:1 [17-18]. Finally, the modeling dataset for this study consisted of 10,251 samples. There are 8,137 non-purchases and 2,114 purchases.

To ensure the robustness of the prediction model in real-world environments, the dataset is split into a training set and a testing set in a ratio of 7:3 in this study. The training set is mainly used to construct a prediction model of consumers' online shopping purchase behavior. The testing set is used to validate the performance of the prediction model on out-of-sample data. The split method can reduce the potential bias of the prediction model. Furthermore, it has been shown to enhance the generalizability of the prediction model and to confirm its external validity and deployment adaptability.

3.4 Random Forest Algorithm

The Random Forest method was proposed by Breiman (2001). It is an ensemble learning technique that has proven to be robust, noise tolerance, and highly generalizable. This makes it well suited

Table 1: Variables used for model construction

Variable	Description
purchase	purchase or non-purchase
total_events	the consumers' total number of events
unique_products	number of different products of interaction
unique_categories	number of different product categories
active_hours	hours between the first and last event
view_count	number of product view events
cart_count	number of add-to-cart events
session_count	total number of sessions
avg_session_duration	average session duration (in minutes)
cart_event_ratio	ratio of cart events to total events
cart_per_product	average cart additions per viewed product

Table 2: Confusion matrix

	Predicted: No	Predicted: Yes
Actual: No	TN	FP
Actual: Yes	FN	TP

to the classification prediction of high-dimensional data [12]. The efficacy of prediction performance is enhanced by two fundamental randomization strategies. First, a number of single decision trees are constructed by bootstrap sampling different training subsets. This process reduces the model's dependence on a single sample. Secondly, at the point of each decision tree node split, Random Forest randomly selects features to prevent the influence of certain features on the model's construction. These two stochastic processes collaborate to provide Random Forest with stable and efficient predictive capability, thereby effectively prevent overfitting.

In this study, the parameters of the random forest model were set based on the literature and the characteristics of e-commerce user behaviors. The number of decision trees is set to 100 (n_estimators = 100) to ensure sufficient model capacity. For each node split, the number of features is set as the square root of the total number of features (max_features = sqrt) to increase feature diversity. The minimum number of samples per leaf was set to 1 (min_samples_leaf = 1) to maintain model flexibility, and bootstrap sampling was enabled (bootstrap = True) to maintain the core integration structure [19]. This configuration achieves a balance between prediction accuracy and generalization ability. This enhances the stability and applicability of the model in real e-commerce scenarios.

3.5 Model Evaluation

The confusion matrix is used in this study to evaluate the performance of the prediction model (see Table 2). It calculates the difference between the predicted and actual results to provide a quantitative measure of the model's accuracy [20]. True positive (TP) is the correct prediction of purchase. True negative (TN) is the correct prediction of non-purchase. False positive (FP) is to wrongly classify non-purchase as purchase. False negative (FN) is to wrongly classify actual purchase as non-purchase.

The following four metrics are widely used to evaluate the classification performance of the model. The formulas are given below:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

$$Precision = TP/(TP + FP) \quad (2)$$

$$Recall = TP/(TP + FN) \quad (3)$$

$$F1 - score = 2 \times TP/(2 \times TP + FP + FN) \quad (4)$$

Accuracy measures the proportion of correctly categorized samples and reflects the model's performance in predicting purchase events. Precision indicates the proportion of samples that are actually purchases out of those predicted as purchases by the model. Recall represents the proportion of predicted purchase samples among all actual purchased samples. The F1-Score is a harmonic mean of precision and recall. It is used to comprehensively assess the balance between precision and recall of the model, even in the presence of class imbalance. In this study, an accuracy benchmark of 0.90 is adopted to initially select candidate models. Due to the categorical imbalance in this study, the model with the optimal prediction performance is identified through the F1-score. This metric is used to ensure the effectiveness of consumers' purchasing behavior prediction model in e-commerce applications.

4 EMPIRICAL RESULTS

4.1 Exploratory Data Analysis

Before the model construction, exploratory analysis for the time of the event and the price of the event are conducted in this study. As illustrated in Figure 2, there is a clear daily pattern in the consumer purchasing behavior. The data show that activity increases steadily from 6:00. The highest point of the curve is reached between 10:00 and 12:00, and a significant decline is observed between 00:00 and 5:00. It has been observed that viewing events often exceed the number of carts and purchases. Carts often follow peaks in viewing

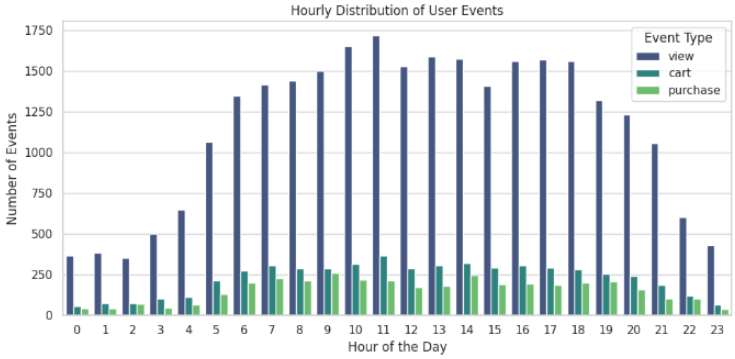

Figure 2: Hourly distribution of user events

Figure 3: Price distribution by event type

and purchases follow cart activity. This suggests that consumers usually evaluate and compare items before making a purchase. As a result, e-commerce platforms can push notifications prior to peak hours. It is recommended that businesses offer time-limited discounts between 10:00-12:00 to maximize conversion potential.

An analysis of the price distribution (see Figure 3) reveals that viewing events are associated with lower-priced products (0-100), cart events occur more frequently in the 200-400 range, and purchases tend to favor lower-priced products. These findings imply that low-cost items tend to elicit impulsive purchasing. High-priced items require additional comparisons, considerations or external incentives before purchase. In practice, e-commerce platforms can offer instant discounts on low-priced items to stimulate quick decision-making. For high-priced items, it is recommended to offer installment payment options, comparison tools or personalized promotions to increase the conversion rate of the shopping cart. This differentiated pricing strategy can facilitate e-commerce platforms to effectively match consumer decisions across various product categories and price levels.

4.2 Model Evaluation and Comparison

As shown in Table 3, both Random Forest and Decision Tree models achieve an accuracy of more than 0.90 on the testing set, and both can be used as candidate models for this study. Relying on accuracy alone may mask the model's ability to recognize a small number of events, so the F1-score is further used in this study to obtain a more comprehensive assessment. Random Forest outperformed Decision Tree in terms of accuracy (0.9168 vs. 0.9083), precision (0.7339 vs. 0.7032), and F1 score (0.8225 vs. 0.8120). It can be seen that Random Forest provides a more balanced and robust performance in both overall classification and minority category identification. Random Forest is the optimal prediction model in this study.

4.3 Feature Importance Analysis

Figure 4 shows the relative importance scores assigned to the 10 behavioral features based on the optimal random forest prediction model.

Cart_event_ratio (27.23%) is the most influential factor. This indicates that the higher the percentage of "add-to-cart" behaviors, the higher the likelihood of completing a purchase. Total_events (19.69%) ranked second. This highlights the key role of overall user engagement in predicting conversions. The third and fourth most important variables are cart_per_product (16.01%) and view_count (12.18%). This suggests that repeated interactions with the same product and extensive viewing are strong signals of purchase intent. Cart_count (11.53%) further validates the impact of adding a shopping cart on purchasing behavior. Active_hours (6.28%) and avg_session_duration (6.11%) suggest that longer interaction

Table 3: Confusion matrix for the testing set

	Accuracy	Precision	Recall	F1-score	Selected
Random Forest	0.9168	0.7339	0.9353	0.8225	V
Decision Tree	0.9083	0.7032	0.9605	0.8120	

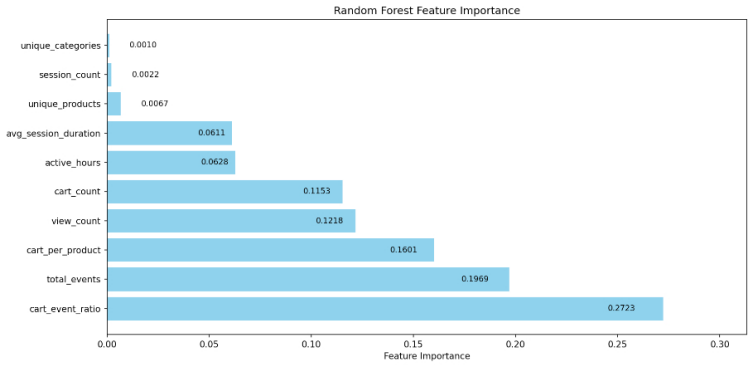

Figure 4: Feature importance in the Random Forest model

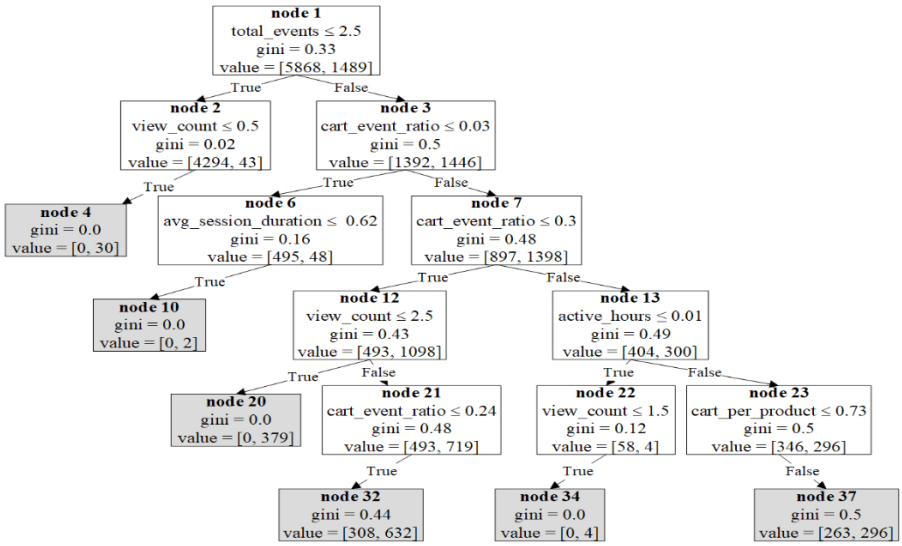

Figure 5: Simplified Random Forest Model with Terminal Nodes Representing Purchases

times help differentiate purchasers from non-purchasers. Although the importance of session_count (0.22%), unique_products (0.67%), and unique_categories (0.10%) is relatively low, they still provide valuable additional information. Overall, these findings not only reaffirm the critical "add-to-cart → purchase" behavioral pathway. They also emphasize the need to integrate behavioral intensity and temporal dimensions.

4.4 Random Forest-Based Purchase Prediction

This study aims to predict consumers' online purchasing behavior using Random Forest. The model has a total of 37 nodes across 6 layers. Due to space constraints, Figure 5 provides a simplified illustration of the model. It focuses on the terminal nodes that represent purchasing behavior.

Figure 5 illustrates the model's structure. Each node is classified by a specific behavioral characteristic. Each node has a number, and the Gini index measures the level of impurity within the node's samples. The Gini index takes values in the range of 0 to 1, with smaller values indicating purer nodes. The value of each node is represented as an array in the form [x, y], where x is the number of non-purchasers and y is the number of purchasers. Each node's value is represented as an array [x, y]. x is the number of non-purchasers and y is the number of purchasers. For example, Figure

5 shows that Node 32 is a terminal node with a Gini value of 0.44 and a value of [308, 632]. Since the number of purchasers (632) exceeds the number of non-purchasers (308), the terminal node is categorized as "Purchase" with a purchase probability of 67.23%.

To deepen Random Forest's understanding of predicting purchase behavior, this study conducts an interpretable analysis of the terminal nodes. For example, the purchasing behavior population of node 32 has the following characteristics. "Total_events > 2.5" indicates above-average consumer engagement; "0.03 < cart_event_ratio ≤ 0.3" means that "add to cart" behaviors accounted for 3-30% of interactions. It suggesting that consumers make real choices rather than clicks; "view_count > 2.5" shows consumers who choose to browse at least 3 different items. This reflects meaningful product comparisons; "cart_event_ratio ≤ 0.24" ensures consistency in adding to cart.

The above rules emphasize the key function of total_events in predicting consumers' online purchasing behavior. It indicates that the larger the total_events (>2.5) value is, the stronger the user interaction will be and the higher the purchase intention will be. In this group, cart_event_ratio further distinguishes the purchase possibility. A low ratio (≤0.03) indicates low stickiness, and a high ratio (>0.3) indicates high purchase intention. For consumers with total_events > 2.5 and cart_event_ratio between 0.03 and 0.3, view_count plays a key role. These findings align with the feature importance analysis in Section 4.3, reinforcing the significance of these features in predicting purchase behavior.

5 Conclusion

This study addresses the increasing complexity and diversity of e-commerce consumer behavior by constructing a Random Forest prediction model of consumers' online purchasing behavior. The model's prediction performance on out-of-sample data is enhanced by rigorous data cleaning, feature engineering, and class balancing. It results in accuracy and recall exceeding 0.90 and F1-score greater than 0.80. Through a comprehensive analysis of the existing data, this study has found that the four important factors influencing purchase intention are as follows: cart_event_ratio, total_events, cart_per_product and view_count.

In fact, these findings emphasize the value of data-driven consumer insights for e-commerce platforms seeking to enhance conversions, personalize recommendations, and optimize precision marketing. The modeling framework not only provides high prediction accuracy and generalizability, but also lays a foundation for future research. The integration of blockchain technology with an extensive array of data sources enables the acquisition of comprehensive behavioral insights, thereby enhancing the predictive accuracy of models. This study offers a comprehensive and replicable methodology for predicting online purchasing behavior. It provides actionable guidance for stakeholders operating in an increasingly competitive digital economy.

Acknowledgments

This work is supported by the project grant from 2021 Guangdong Key Discipline Research Ability Improvement Project (Grant No. 2021ZDJS129), Department of Education of Guangdong Province, Research Team on Digital-real Integration and Management Innovation (Grant No. 2023WCXTD023), and "Interdisciplinary Integration and Innovation" Project and the First-Batch Research Project of the Doctoral Workstation from Guangdong University of Foreign Studies South China Business College (Grant No. 2024XJZ02).

References

[1] Ying Lin. **2024**. Global Ecommerce Sales Growth Report. Retrieved April 10, 2025 from https://www.shopify.com/blog/global-ecommerce-sales.
[2] Saed Alizamir, Kasun Bandara, Ali Eshragh, and Foaad Iravani. **2022**. A Hybrid Statistical-Machine Learning Approach for Analysing Online Customer Behavior: An Empirical Study. ArXiv, abs/2212.02255.
[3] Hui Xu. **2022**. GBDT-LR: A Willingness Data Analysis and Prediction Model Based on Machine Learning, In Proceedings of IEEE International Conference on Advances in Electrical Engineering and Computer Applications(AEECA'22) IEEE, Dalian, China, 396-401.
[4] Abdul Aziz M, Mustakim N A, Abdul Rahman S. **2024**. Decision tree and rule-based classification for predicting online purchase behavior in Malaysia. Malaysian Journal of Computing (MJoC), 2024, 9(2), 1905-1915.
[5] Ruiqin Wang, Zongda Wu, Jungang Lou, and Yunliang Jiang. **2022**. Attention-based dynamic user modeling and Deep Collaborative filtering recommendation. Expert Systems with Applications, 188, 116036.
[6] Po Abas Sunarya, Untung Rahardja, Shih Chih Chen, Yung-Ming Lic, and Marviola Hardini. **2024**. Deciphering digital social dynamics: A comparative study of logistic regression and random forest in predicting e-commerce customer behavior. Journal of Applied Data Sciences, 5(1), 100-113.
[7] Wei-yu K. Chiang, Dongsong Zhang, and Lina Zhou. **2006**. Predicting and explaining patronage behavior toward web and traditional stores using neural networks: a comparative analysis with logistic regression. Decision Support Systems, 41(2), 514-531.
[8] J. R. Quinlan. **1996**. Improved use of continuous attributes in C4.5. Journal of artificial intelligence research, 4, 77-90.
[9] Laura Maria Badea. **2014**. Predicting consumer behavior with artificial neural networks. Procedia Economics and Finance, 15: 238-246.
[10] Shashi Pal Singh, Ajai Kumar, Neetu Yadav, and Rachna Awasthi. **2018**. Data Mining: Consumer Behavior Analysis, In Proceedings of 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT'18) IEEE, Bangalore, India, 1917-1921.
[11] Hua Wang, Lingwei Wang, Fuyu Zhu. **2024**. E-Commerce User Behavior Analysis and Prediction Based on Artificial Neural Network and Data Mining," 2024 IEEE 7th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chongqing, China, 583-586
[12] Leo Breiman. **2001**. Random forests. Machine learning, 45, 5-32.
[13] Gyanendra Chaubey, Prathamesh Rajendra Gavhane, Dhananjay Bisen, and Siddhartha Kumar Arjaria. **2023**. Customer purchasing behavior prediction using machine learning classification techniques. Journal of Ambient Intelligence and Humanized Computing, 14(12), 16133-16157.
[14] Xuan Zhang. **2022**. Research on Consumer Purchase Intention Prediction Based on Data Mining. Master's thesis, University of International Business and Economics.
[15] Aravind K. Kalusivalingam, Amit Sharma, Neha Patel, and Vikram Singh. **2020**. Enhancing Predictive Business Analytics with Deep Learning and Ensemble Methods: A Comparative Study of LSTM Networks and Random Forest Algorithms. International Journal of AI and ML, 1(2), 1-23.
[16] Rofik Rofik and Nurul Hidayat. **2023**. Improving the accuracy of the logistic regression algorithm model using SelectKBest in customer prediction based on purchasing behavior patterns. Future Computer Science Journal. 1, 1, 9-17.
[17] Małgorzata Bach, Aleksandra Werner, and Mateusz Palt. **2019**. The proposal of undersampling method for learning from imbalanced datasets. Procedia Computer Science. 159, 125–134.
[18] Roweida Mohammed, Jumanah Rawashdeh, and Malak Abdullah. **2020**. Machine learning with oversampling and undersampling techniques: overview study and experimental results. In Proceedings of the 11th International Conference on Information and Communication Systems (ICICS 2020). IEEE, Los Alamitos, CA, 243-248.
[19] Philipp Probst, Marvin N. Wright, and Anne-Laure Boulesteix. **2019**. Hyperparameters and tuning strategies for random forest. Wiley interdisciplinary reviews. Data mining and knowledge discovery. 9(3), e1301.
[20] Jiqi Yan. **2024**. Purchase Behavior Prediction Analysis Based on Online Shopping Users. Master's thesis. Chongqing University of Technology, Chongqing, China.