Dean&Francis
ISSN 2959-6157

Customer Online Purchase Behavior Prediction and Performance Analysis Using Decision Tree and Random Forest

Wenjie Zheng

Department of Computer Technology, Fuzhou University, Fuzhou, China

maggiezheng@pitlc.com

Abstract:

In the era of online shopping, understanding customer behavior has become increasingly crucial. By analyzing the detailed factors that influence customer actions, businesses can gain deeper insights into their clientele and enhance their targeted marketing strategies. This study investigates the influence of various factors on online customer purchasing frequency using a comprehensive dataset. The research is divided into two phases. Initially, the decision tree and random forest algorithms were utilized to analyze all dataset features, establishing a baseline model and determining feature importance. Subsequently, the second phase delved into the impact of feature count on model efficacy by incrementally eliminating less significant features. The study revealed that a model incorporating half of the features—namely purchase amount, age, review rating, previous purchases, location, color, purchased item, and shipping type—achieved comparable performance to the full-feature model. This streamlined approach not only expedited computation time but also reduced memory usage during the training process, offering valuable insights for businesses to refine their marketing strategies and enhance customer engagement. The findings underscore the potential of data-driven methods to optimize marketing efforts in the e-commerce sector, making a fundament for the future analysis.

Keywords: Online behavior; frequency of purchase; decision tree; random forest.

1. Introduction

Post COVID-19 pandemic, the popularity of e-commerce is growing. An increasing number of people tend to buy products online. Based on statista's studies, global retail e-commerce sales reached an estimated $5.8 trillion in 2023 and are expected to hit $8 trillion by 2027 [1]. This result reflects that people's purchase behaviors have been already reshaped. With the tendency of online shopping, studying customer behavior on the Internet has become a hot topic in recent years. Rich collection of customer behavior analysis can not only enable businesses to formulate targeted marketing strategies, optimize product offerings, but also enhance overall customer satisfaction.

The field of behavior analysis includes a comprehensive framework for investigating and understanding human behavior [2]. Traditional research are basically related to social psychology such as Theory of Planned Behavior (TPB) [3] and Technology Acceptance Model (TAM) [4]. The emergence of data mining and artificial intelligence technology has resulted in the transformation of behavior analysis from a theory-driven psychological study to a data-oriented interdisciplinary study, like healthcare [5], robot simulation [6], and particularly within the realm of business [7]. The motivation of customer behavior in commerce are actually divided into two categories: The detection of behavioral patterns can give a fuller picture of customers, while the prediction of behaviors can allow businesses to develop more appropriate strategies for existing and potential customers [8]. For example, Ernawati et al. [9] proposed a novel framework that leveraged data mining methods and segmentation technologies to investigate and comprehend customer characteristics within a Geographic Information System (GIS) environment. Further, Arefin et al. employed a set of supervised machine learning models to predict Customer Lifecycle Value (CLV) of UK retail industry [10]. However, it is worth noting that these studies mainly work on existing traditional industry datasets, which are often measured by large organizations, and lack a focus on emerging industry datasets such as online shopping data. Furthermore, businesses often pay more attention to the exploration of customers they prefer, but ignore the growth of customers they have, no matter what industry they are in. This preference is also an important factor of uneven research content. For instance, in order to find high-value e-shop clients, Sakalauskas and Kriksciuniene proposed a new algorithm to measure customer engagement by utilizing clickstream data, including but not limited to time spent on the website and frequency of visits to e-shops [11].

Based on existing research, this study would pay more attention to the factors influencing the decision-making process of online customers, which will assist enterprises in refining their marketing strategies for inactive customers. To show different analysis from others, consumer behavior and shopping habits dataset [12] is ultimately chosen to explore the factors of frequency of purchases of customers. The rest structure of this paper is organized as follows: Section 2 introduces 2 referred methods: decision tree and random forest model used in this study. Section 3 provides results and discussion on feature importance and the influence of number of features in model performance. The conclusion of this work and future work is discussed in Section 4.

2. Method

2.1 Dataset Preparation

In this study, a consumer behavior and shopping habits dataset from Kaggle [12] was used. The dataset provided comprehensive insights into consumer preferences and purchasing behaviors, such as their age, gender and previous purchases etc. The original dataset contained 3,900 customer data, each containing 18 items of personal information. In order to investigate the factors influencing customers' purchasing frequency, column named "Frequency of Purchases" was utilized as a label, which indicates how often the customer engages in purchasing activities. Fig. 1 shows the distribution of labels. The remaining columns (except for the "Customer ID" column) were used as features to train the model to find a relationship with the frequency of customer purchases. Fig. 2 shows an example of correlation between column aged "Gender" and "Frequency of Purchases".

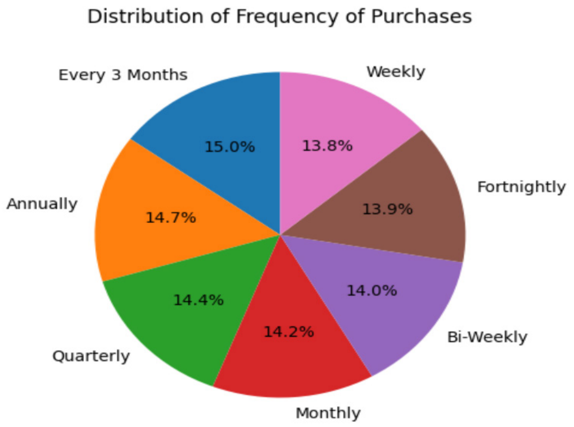

Fig. 1 The distribution of Frequency of Purchases (Photo/Picture credit: Original)

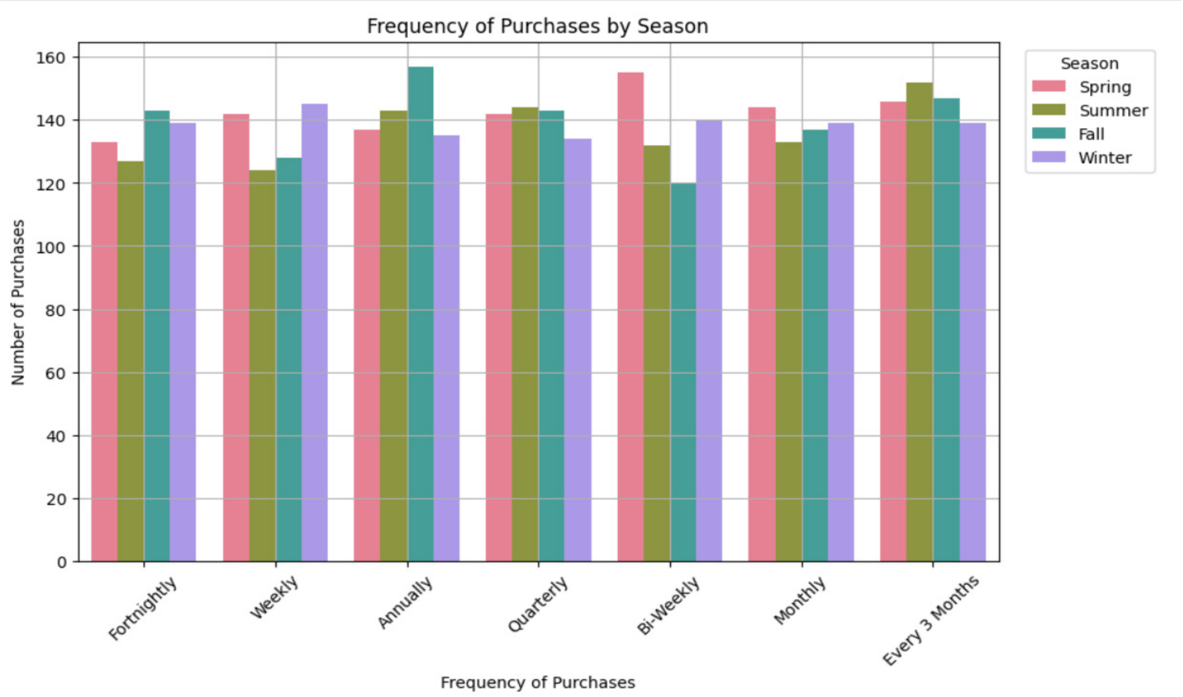

Fig. 2 The distribution of Frequency of Purchases by Season (Photo/Picture credit: Original)

The preprocessing used in this study consists of four steps. First, the column named “Customer ID” was dropped, as it was deemed superfluous for data analysis purposes. Second, text standardization was applied to deal with all columns. In this dataset, there were 12 categorical columns e.g. Gender used to describe customer behavior. To make samples suitable for training models, class “OrdinalEncoder” method is utilized to standardize such categorical data by converting text into numerical values. For instance, after data standardization, “male” of “Gender” is represented as 1 while “female” of “Gender” is represented as 2. Also, this study preprocessed the remaining 4 non-numerical columns e.g. Age into numerical columns. Third, the dataset was divided into training set and test set: The number of samples in the training set is 2, 613 while the number of samples in the test set is 1, 287. Finally, a StandardScaler is chosen to normalize the dataset. Initially, the training data is transformed to establish a standard, and subsequently, the test data is normalized based on this standard.

2.2 Machine Learning-based Classification

This paper chose decision tree and random forest model from scikit-learn, two classical Machine Learning (ML) algorithms, as the proposed methods. The ratio of training set to test set was 8:2, and the random state was set to 42. After training, evaluation metrics including accuracy, and F1-score were utilized to measure the results of machine learning models. Mathematically, the relationship of these measures can be written as below [13]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precison = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - score = \frac{2}{\frac{1}{Precison} + \frac{1}{Recall}} = \frac{2TP}{2TP + FP + FN} \quad (4)$$

Where TP represents true positive, which describes the number of records that are correctly classified; TN represents true negative, which describes the number of the correct rejection of records that have been classified; FP represents false positive, which describes the number of records incorrectly classified; FN represents false negative, which describes the number of incorrect rejection of records that have been classified [14].

2.2.1 Decision tree

Decision Tree (DT) is a flowchart-like structure in the field of machine learning that is used for classification and regression [15]. The concept of DT was derived from the conventional tree structure, characterized by a central node and multiple leaves nodes and branches. It breaks down a dataset into smaller subsets by Recursive Portioning Algorithm (RPA) while at the same time, sub-trees are incrementally developed [16]. Each internal node in the tree corresponds to a feature attribute, each branch represents a decision rule, and each leaf node represents an outcome.

In this paper, pre-pruning was utilized as the method for decision tree simplification and overfitting mitigation. After that, data groups were divided based on their features, following the decision rules mentioned above. The process terminated when all records in the current subset are classified into the same class [17].

2.2.2 Random forest

Random Forest (RF) is an ensemble learning method that operates by constructing a multitude of decision trees randomly at training time and predicts by using an equally weighted majority vote. During the process of model training, each decision tree uses two-thirds of the initial dataset at random [18]. Regarding optimal segmentation points, a feature subset is selected at random and then the optimal feature is selected in the feature subset for segmentation [17]. Compared with DT, RF corrects for decision trees' habit of overfitting to their training set, therefore, are widely used due to their robustness, ease of use, and ability to handle large datasets with high dimensionality.

During model learning, data groups were divided based on their features, following the decision rules mentioned above. It is noted that RF also provides results with the feature importance of the dataset after every model training. Regarding this paper, the feature importance was quite important as it provides empirical evidence for adjusting each feature set during training.

3. Results and Discussion

This study primarily examines the impact of customer shopping behavior on purchasing frequency, which is conducted in two stages. To begin with, Section 3.1 compares the classification performance of DT and RF using all features. Following the selection of an appropriate ML model, Section 3.2 analyzes the impact of the number of features, considering feature importance in the model.

3.1 Classification Performance of DT and RF Based on all Features

By fine-tuning the hyperparameters of the DT model, a series of experiments were conducted, yielding varying results. Among these hyperparameters, "max_depth" stands out as the most critical, as it controls the maximum depth of the DT, thereby determining the complexity of the decision tree. Fig. 3 illustrates the accuracy and F1-score of the model across different maximum depths of the trees. As depicted, the model achieved a superior performance when the maximum depth was set to 8, with an accuracy of 16.01% and a F1-score of 15.60%.

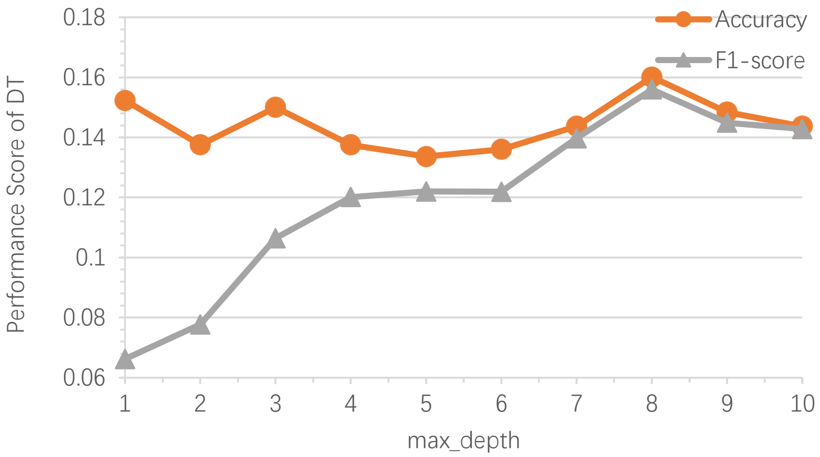

Fig. 3 Accuracy and F1-score of DT with different maximum depth (Photo/Picture credit: Original)

As depicted in Fig. 3, there are 10 experiments conducted with varying maximum depths, ranging from 1 to 10. F1-score shows an upward trend from trials 1 to 8. At the same time, there is a slight increase in accuracy compared to the initial trial when maximum depth equals 8. It is noted that increasing the maximum depth of the DT allows the model to learn more high-level representations. However, if the depth continues to increase without restraint, the structure of the decision tree becomes increasingly complex. This can lead to overfitting, where the model learns the noise in the training data, which may in turn reduce the generalizability and performance of the decision trees.

To ensure a fair comparison between the Decision Tree and Random Forest models, it is crucial to configure the trees in the RF model with a structure similar to that of the DT model. Consequently, a series of experiments were conducted by fine-tuning the hyperparameters of the RF model in order to match the "max_depth" of 8 from the DT model. Among these hyperparameters, "n_estimators" is particularly significant because it determines the number of trees in the RF, which directly influences the model's complexity. Fig. 4 illustrates the accuracy and F1-score of the model across various numbers of decision trees. As shown, the model performed better when the number of trees was set to 14, achieving an accuracy of 17.44% and a F1-score of 17.41%.

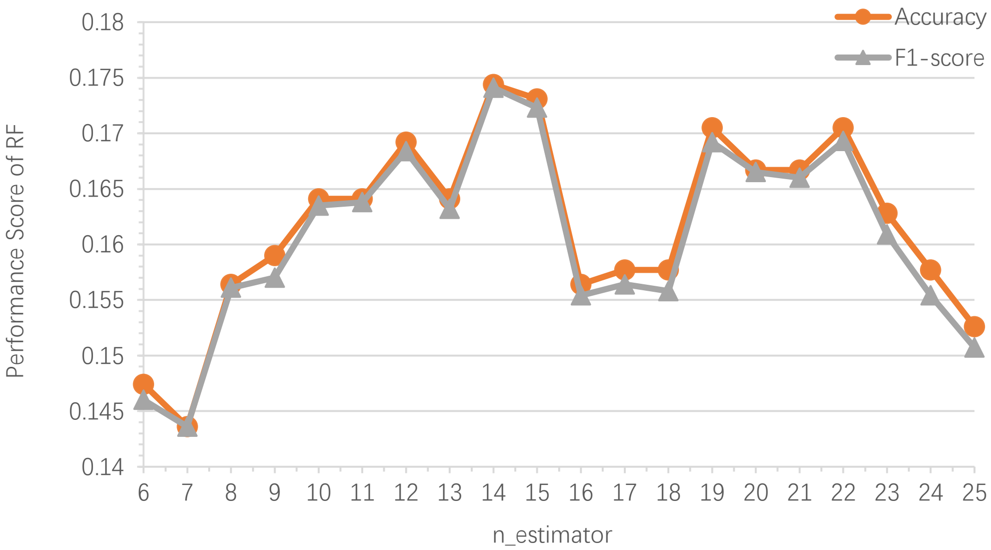

Fig. 4 Accuracy and F1-score of RF with different number of trees (Photo/Picture credit: Original)

As shown in Fig. 4, 20 experiments were conducted with varying numbers of trees, ranging from 6 to 25. Despite some fluctuations, both accuracy and the F1-score exhibit an upward trend from trials 6 to 14. This trend aligns with theoretical expectations: as the number of trees increases, the model's performance tends to improve. However, a different pattern emerges when the number of trees exceeds 14. This deviation could be due to the relatively small size of the dataset. As the model complexity grows with the addition of more trees, identifying optimization points becomes challenging, potentially resulting in decreased performance.

Table 1 shows the comparison of performance of DT and RF based on the optimal architecture using all features.

Table 1. The performance of DT and RF based on all features

Model	Performance			
	Training Accuracy	Training F1-score	Testing Accuracy	Testing F1-score
Decision Tree	0.3131	0.3124	0.1641	0.1579
Random Forest	0.6186	0.6176	0.1744	0.1741

As indicated in Table 1, the random forest outperformed the decision tree model on both the training and testing sets. In terms of the training set, the performance of RF was approximately double that of DT, indicating that the RF model has a significantly stronger learning capability than the original dataset. Regarding the testing set, the result of RF was slightly higher than that of DT, demonstrating that the RF has a better generalization ability on unseen data.

3.2 The influence of Number of Features in RF Based on Feature Importance

After applying random forest algorithm, the importance of all features can be precisely calculated. The impact of feature count on the performance of an RF classifier is a critical aspect of model optimization. Fig. 5 presents a bar chart that ranks the importance of each feature from the highest to the lowest. According to this distribution, the feature "Purchase Amount" holds the greatest proportion, accounting for 11.44%. Conversely, the feature "Promo Code Used" has the least impact on the frequency of customer purchases, with a mere 0.88%.

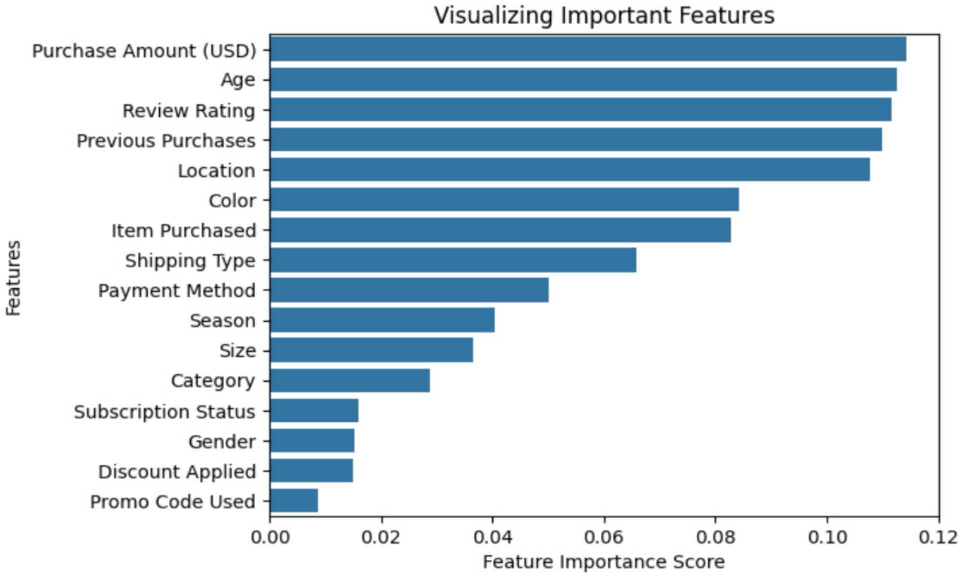

Fig. 5 Feature Importance by RF (Photo/Picture credit: Original)

To investigate the impact of the number of features, various feature sets were constructed for the experiments. Following the feature importance rankings depicted in Fig. 5, the least significant feature was consistently excluded from each subsequent feature set. The outcomes of these experiments are presented in Fig. 6.

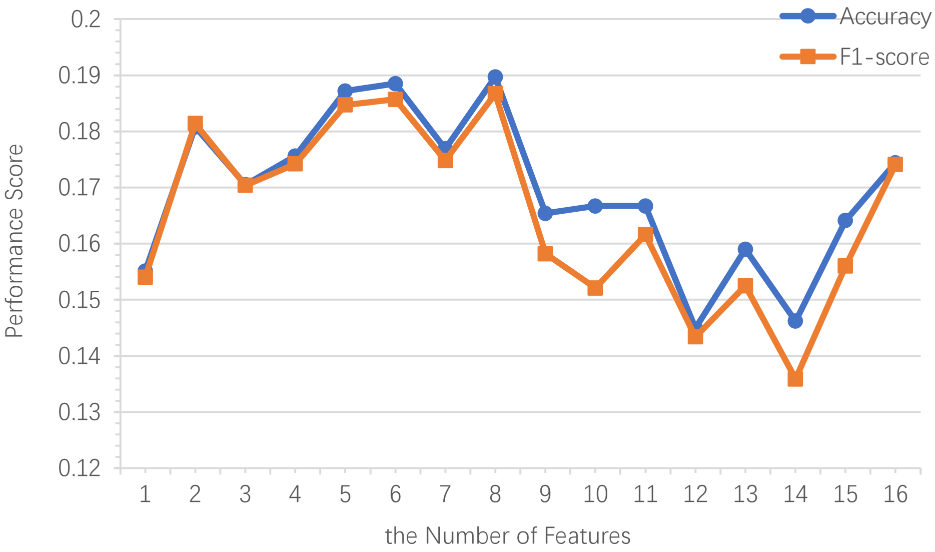

Fig. 6 Accuracy and F1-score of RF with different feature sets (Photo/Picture credit: Original)

Fig. 6 illustrates the figures for RF performance, namely accuracy and F1-score, both starting at approximately 17.44% and 17.41%, respectively. During the number of features decreases from 16 to 9, there is a decline in performance, suggesting that attributes such as age, review rating, previous purchases, location, color, item purchased, and shipping type contribute positively, albeit a little, to the model's predictive power. However, a significant improvement in performance is observed once the "payment method" feature is excluded, with the accuracy and F1-

score peaking at 18.97% and 18.67%, respectively. This indicates that the payment method, which includes six distinct methods in the dataset, may act less contribution to learning customer's frequency of purchases. This also leads to the conclusion that the model achieves optimal performance through 8 features while reducing resource consumption by half. In addition, the continued removal of features, excluding purchased items and review rating led to improvements in model performance, suggesting that these features can also be considered disturbing variables.

Furthermore, this study highlights the critical role of feature selection in the field of customer behavior analysis. Traditionally, researchers have analyzed customer behavior without feature filtering, a practice that is not only unreasonable but also inefficient. It is expected that the application of the aforementioned group of features will lead to enhanced optimization outcomes in subsequent research.

4. Conclusion

This study aimed to identify the determinants in the online consumer decision-making process, thereby enabling businesses to tailor their marketing strategies more effectively. Initially, Decision Tree and Random Forest were both employed to model all available features. The findings indicated that RF outperformed DT under various hyperparameters settings. Once an optimal method was decided, a focused discussion on the impact of varying feature counts ensued, guided by feature importance which was calculated by RF. Notably, models utilizing 8 features surpassed the benchmark of 16 features across different hyperparameter configurations, achieving this with a 50% reduction in training duration and resource consumption. Moreover, the payment method was found to have minimal impact on customer purchase frequency. It is suggested that the feature group with purchase amount, age, review rating, previous purchases, location, color, purchased item and shipping type might affect more on customers' behavior. Future studies will explore other data sources with the above 8 features to serve more customer behavior learning tasks.

References

[1] Statista Retail e-commerce sales worldwide from 2014 to 2027, 2023, https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/

[2] Sundararaj V, Rejeesh MR. A detailed behavioral analysis on consumer and customer changing behavior with respect to social networking sites. Journal of Retailing and Consumer Services. 2021 Jan 1;58:102190.

[3] Pillai SG, Kim WG, Haldorai K, Kim HS. Online food delivery services and consumers‘ purchase intention: Integration of theory of planned behavior, theory of perceived risk, and the elaboration likelihood model. International journal of hospitality management. 2022 Aug 1;105:103275.

[4] Islam H, Jebarajakirthy C, Shankar A. An experimental based investigation into the effects of website interactivity on customer behavior in on-line purchase context. Journal of Strategic Marketing. 2021 Feb 17;29(2):117-40.Amjady N. Short-term hourly load forecasting using time series modeling with peak load estimation capability. IEEE Transactions on Power Systems, 2001, 16(4): 798-805.

[5] Ntoumanis N, Ng JY, Prestwich A, Quested E, Hancox JE, Thøgersen-Ntoumani C, Deci EL, Ryan RM, Lonsdale C, Williams GC. A meta-analysis of self-determination theory-informed intervention studies in the health domain: Effects on motivation, health behavior, physical, and psychological health. Health psychology review. 2021 Apr 3;15(2):214-44.

[6] Wang Q, Jiao W, Wang P, Zhang Y. Digital twin for human-robot interactive welding and welder behavior analysis. IEEE/CAA Journal of Automatica Sinica. 2020 Nov 24;8(2):334-43.

[7] Wibowo A, Chen SC, Wiangin U, Ma Y, Ruangkanjanases A. Customer behavior as an outcome of social media marketing: The role of social media marketing activity and customer experience. Sustainability. 2020 Dec 28;13(1):189.

[8] Mach-Król M, Hadasik B. On a certain research gap in big data mining for customer insights. Applied Sciences. 2021 Jul 29;11(15):6993.

[9] Ernawati E, Baharin SS, Kasmin F. A review of data mining methods in RFM-based customer segmentation. InJournal of Physics: Conference Series 2021 Apr 1 (Vol. 1869, No. 1, p. 012085). IOP Publishing.

[10] Arefin S, Parvez R, Ahmed T, Ahsan M, Sumaiya F, Jahin F, Hasan M. Retail Industry Analytics: Unraveling Consumer Behavior through RFM Segmentation and Machine Learning. In2024 IEEE International Conference on Electro Information Technology (eIT) 2024 May 30 (pp. 545-551). IEEE.

[11] Sakalauskas V, Kriksciuniene D. Personalized Advertising in E-Commerce: Using Clickstream Data to Target High-Value Customers. Algorithms. 2024 Jan 10;17(1):27.

[12] Kaggle, Consumer Behavior and Shopping Habits Dataset, 2023, https://www.kaggle.com/datasets/zeesolver/consumer-behavior-and-shopping-habits-dataset?select=shopping_behavior_updated.csv

[13] Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics. 2020 Dec;21:1-3.

[14] Jehad R, Yousif SA. Fake news classification using random forest and decision tree (j48). Al-Nahrain Journal of Science. 2020 Nov 30;23(4):49-55.

[15] Madane N, Nanda S. Loan prediction analysis using decision tree. Journal of the Gujarat Research society. 2019 Dec;21(14):214-21.
[16] Aslam U, Tariq Aziz HI, Sohail A, Batcha NK. An empirical study on loan default prediction models. Journal of Computational and Theoretical Nanoscience. 2019 Aug 1;16(8):3483-8.
[17] Ali J, Khan R, Ahmad N, Maqsood I. Random forests and decision trees. International Journal of Computer Science Issues (IJCSI). 2012 Sep 1;9(5):272.
[18] Tariq A, Yan J, Gagnon AS, Riaz Khan M, Mumtaz F. Mapping of cropland, cropping patterns and crop types by combining optical remote sensing images with decision tree classifier and random forest. Geo-Spatial Information Science. 2023 Jul 3;26(3):302-20.