HNSJ Humanities & Natural Sciences Journal
ISSN: (e) 2709-0833
www.hnjournal.net

Peer-Reviewed Journal
Indexed at: NSP - SJIF
Impact Factor 2022 = 4.91

RESEARCH TITLE

IMPROVING REPEAT PURCHASE PREDICTION IN DIGITAL MARKETING USING MACHINE LEARNING

Muthana HAMAD[1], Sefer KURNAZ[2]

[1]Department of Data Analytics, Altinbas University, Turkey, mu9.ha9@gmail.com (https://orcid.org/0009-0009-2040-3898)
[2]Department of Computer Engineering, Altinbas University, Turkey, sefer.kurnaz@altinbas.edu.tr (https://orcid.org/0000-0002-7666-2639)

HNSJ, 2025, 6(1); https://doi.org/10.53796/hnsj61/7

Received at 07/11/2024 **Accepted at 15/12/2024** **Published at 01/01/2025**

Abstract

This study presents a machine learning-based approach to predict repeat purchase behavior in digital marketing by integrating behavioral and demographic data. The study relied on a dataset of 10,000 customer records from an e-commerce platform, which was reduced to 5,741 records after data cleaning. Neural Networks, Random Forest, and AdaBoost algorithms were applied to analyze key features such as age, lifetime customer value, email conversion rate, and average order value. Established performance metrics CA(Classification Accuracy), F1-Score and AUC(Area Under the Curve) were calculated to evaluate models performances. Our results indicated that the Neural Network algorithm separated remarkable performance and also the characteristics of high stability in the rest of the criteria AUC- 0.886, followed by the Random Forests. The AdaBoost algorithm showed lower performance due to its sensitivity to the nature of the data. These findings support how crucial it is to combine various data sources and apply complex algorithms in order to enhance digital marketing tactics. The suggested strategy helps to improve customer targeting efforts and boost customer loyalty by precisely identifying the clients who are most likely to make repeat purchases.

Key Words: machine learning, digital marketing, prediction, repeat purchase.

1. Introduction

In digital marketing time, data is very important for bettering marketing plans and knowing customer behavior clearer. The success now relies on how good companies are at studying data to see what customers want and like, plus guessing what they will buy next. With so much more data around, from demographic details showing who customers are to behavioral info showing how they act online, there's a big need for smart tools that can put together and study all this data well. Machine learning tech is key in guessing trends, giving companies smart ways to make marketing plans based on data.

A potent tool in digital marketing, predictive analytics examines both past and present data to forecast future consumer behavior. Businesses can enhance their marketing efforts and strategies by using these analytics to find elements that influence consumer behavior and detect behavioral patterns (Shrivas, 2023). Businesses can forecast conversion rates and future purchasing behavior by examining customer data, including prior purchase trends, levels of engagement with advertising efforts, and client retention periods. This increases the efficacy of marketing techniques by enabling advertising efforts to be more precisely focused towards individuals who are more likely to make a purchase (Puri, Sehgal, & Sharma, 2013).

Predictive analytics has evolved over the past period and has emerged as an essential tool as a result of advances in data analysis and processing. Earlier, predictive analytics used traditional statistical methods to measure the behaviour of customers based on their data but as advanced technologies such as machine learning and artificial intelligence evolved, these analyses began to accommodate huge volume of data, resulting in an increase the accuracy towards predicting consumer behavior (Sasmal, 2024).

In the 1970s, predictive analytics relied primarily on traditional statistical modelsAs technology advanced in the 1990s, businesses started using powerful computers to analyze vast volumes of data. By the early 21st century, advances in machine learning and big data had dramatically improved the efficiency of predictive analytics. Businesses may, for instance, use real-time data analysis to make quicker, more effective marketing decisions and more precise forecasts of future consumer behavior (Alfian et al., 2019).

Making precise and successful marketing judgments is based on the data used in predictive analytics in digital marketing. This data is usually divided into two main categories: demographic data, which includes information on age, gender, income, and location, and behavioral data, which includes information about website visits, ad engagement levels, and purchase patterns. By analyzing this data, businesses may better understand their target population and develop more successful and customized marketing campaigns (Wang, 2023).

However, behavioral and demographic data must be combined in order to create effective predictive analytics models. While behavioral data reveals the factors impacting how customers engage with companies and make decisions about what to buy, demographic data gives you a broad picture of the people you are targeting.

Machine learning, a crucial element of predictive analytics, enables the manipulation of vast volumes of data to forecast customer behavior and improve marketing strategies. Machine learning relies on the analysis of past and behavioral data to find patterns in customer behavior and get accurate forecasts about future behavior. Businesses may now use the data being processed to more accurately target their customers and increase the efficacy of their marketing initiatives thanks to the development of AI technology.

Since machine learning techniques allow for the prediction of future purchase rates based on

historical consumer data, they are useful tools for studying repeat purchase behavior. These analyses use a collection of advanced algorithms, including XGBoost, LightGBM, and Logistic Regression, to find patterns in behavior that suggest the possibility of repeat purchases (Zhang, Lu, Ma, Cheng, & Hua, 2022).

Although machine learning techniques can offer significant advantages in predicting purchasing behavior, it is not without its challenges. One such challenge is the lack of data, as missing or incomplete data means that predictive models will be less accurate, and therefore predictions will be unreliable (Emmanuel et al., 2021). Furthermore, combining information from diverse or heterogeneous sources creates a challenge because the quality and structure of the data vary, making it difficult to integrate the data and complicating the analysis process (Hai et al., 2023).

Previous studies have focused on analyzing purchasing behavior using customer data and machine learning algorithms, with variations in the types of data and algorithms used, as well as in the focus and research objectives. Some of these studies include:

Zhang et al. (2022) studied the integration of behavioral and demographic data to develop a "Stacking Fusion" model, which performed well with an AUC of 0.68406, showing that data integration can improve performance, but the study mainly focused on repurchase and did not cover long-term patterns of repeat purchases.

Deniz and Bülbül (2024) relied on a variety of algorithms such as Random Forest and XGBoost, achieving high accuracy of up to 94%. However, the study focused on general purchasing behavior without delving into repeat purchasing behavior, which makes its results less targeted to studying long-term customer engagement with a brand.

Kuric, Puskas, Demcak, and Mensatorisova (2024) focused on low-level engagement data and used algorithms such as Gradient Boosting and Random Forest, with the Gradient Boosting model achieving an accuracy of up to 89%. Although this study provided strong results, it neglected to incorporate behavioral and demographic data together.

Studies such as Lyu (2023) and Li (2024) have provided analyses of repeat purchase behavior using algorithms such as ISSA-SSVM-XGBoost, as well as combined models such as Boosting and Stacking. The former focused on improving prediction accuracy and achieved 97.92% using only demographic data, while the latter combined models to improve the reliability of predictions by focusing on behavioral and interaction data.

Finally, Zhao, Takasu, Yahyapour, and Fu (2019) found that combined models such as LightGBM and XGBoost achieved superior results in repurchase analysis, but the study focused on repurchase of a specific product and did not address long-term repeat purchase that highlights customers' ongoing association with the brand.

This study, as indicated above, seeks to build a predictive model of repeat purchase in the context of e-marketing by integrating behavioral and demographic data. The focus is on optimizing the prediction of such behavior with the help of basic features like Customer Lifetime Value and Email Conversion Rate, which are vital in predicting purchasing behavior. In addition, it studies the impact of advanced algorithms such as neural networks and the AdaBoost algorithm, compared to traditional algorithms such as the random forest algorithm, in this field.

By achieving these objectives, this study contributes to bridging the research gap represented by the lack of studies that rely on building models using advanced algorithms such as neural networks and AdaBoost to analyze data that combines customer demographic and behavioral features and characteristics, with the aim of understanding long-term repeat purchase behavior patterns. This approach helps companies design more effective and personalized marketing strategies, which enhances customer loyalty and improves retention rates.

2. Methodology

In this section, we will explain the methodological steps that were followed during the work to analyze the data and develop the predictive model for repeat purchase behavior. The methodology includes several main stages starting with describing and cleaning the data, through selecting the algorithms used, and ending with evaluating the performance of the models using the approved basic criteria. The study is based on a dataset taken from the Kaggle platform, where machine learning techniques were applied to analyze behavioral and demographic data using the tools available on the Orange platform.

The Orange platform was used as a primary tool in this study due to its easy-to-use, interactive visual interface that allows for the implementation of various required operations such as data cleaning, feature selection, predictive model performance evaluation, and other operations without the need for specialized programming expertise.

2.1. Dataset Description

A dataset from Kaggle titled "E-commerce Customer Engagement and Demographics Dataset" prepared by Subashanan Nair was used. This dataset is available under the MIT license. The original dataset contains 10,000 records distributed between demographic, behavioral, and procedural data, and describes multiple aspects of customer behavior on an e-commerce platform (Nair, 2024).

After cleaning and preparation, the number of records was reduced to 5,741, focusing on features that showed high importance in the repeat purchase analysis. Table 1. below shows the selected features and their description:

Table 1. Selected features for analyzing repeat purchase behavior.

Variable Name	Category	Data Type	Description
RepeatCustomer	Behavioral (Target)	Categorical ("Yes", "No")	Indicates if the customer is a repeat purchaser.
CustomerLifetimeValue	Behavioral	Numerical (e.g., 181.72, 256.68, 540.31)	Represents the estimated lifetime value of the customer.
EmailConversionRate	Behavioral	Numerical (e.g., 0.11, 0.16, 0.34)	Represents the conversion rate from email marketing campaigns.
AverageOrderValue	Behavioral	Numerical (e.g., 15.88, 177.86, 199.16)	Represents the average value of orders placed by the customer.
Age	Demographic	Numerical (e.g., 18, 34, 65)	Represents the age of the customer.

These features were selected based on their importance in improving the accuracy of predictive models, as they reflect long-term purchasing behavior and demographic factors that influence purchasing decisions.

2.2. Data Preprocessing

To ensure the quality of the data used and improve the accuracy of the predictive models, a set of cleaning and pre-processing steps were implemented as follows:

Dealing with missing values: Records with missing values in selected features, such as Customer Lifetime Value and Age, were excluded to avoid negative impact on the analysis. This step reduced the number of records from 10,000 to 7,733.

Handling illogical values: Records containing illogical values such as negative values in numeric features such as Average Order Value and Customer Lifetime Value, and age values under 18 years were removed, as this age group may not reflect stable purchasing behavior. Due to this processing, the data size was reduced to 6055 records.

Handling Outliers: Using the Box Plot tool in Orange platform, outliers were identified in numeric features such as Customer Lifetime Value and Average Order Value. The Select Rows tool was then used to identify the highest acceptable values in these features and trim the outliers. After this step, the data size was reduced to 5830 records.

Data Type Verification: Feature labels were reviewed to ensure they matched the correct data types (e.g., numeric, categorical, etc.), and no identification or classification errors were detected.

Table 2. below shows the data cleaning stages and the distribution of records through each step:

Table 2. Data cleaning and record distribution stages.

Stage	Number of Records	Number of Records Removed
Before Cleaning	10000	-
After Removing Missing Values	7733	2267
After Removing Unreasonable Values	6055	1678
After Restricting Outliers	5741	314

Final data preparation: Descriptive statistics of the final data after cleaning, such as mean, median, range, etc., are presented to illustrate the distribution of the features used. Table 3. below provides a summary of the descriptive statistics of the data:

Table 3. Basic statistics of features after cleaning.

Variable Name	Mean	Mode	Median	Dispersion	Min	Max	Missing (%)
Age	38.34	35	37	0.32	18	88	0%
AverageOrderValue	91.12	1.55	57.77	1.32	1.55	2458.06	0%
CustomerLifetimeValue	431.95	0.02	273.21	1.05	0.02	2584.98	0%
EmailConversionRate	0.1979	0.00088	0.1764	0.61	0.00088	0.7830	0%
RepeatCustomer	0.288	Yes	-	-	-	-	0%

Data Splitting: The data was split using the Cross Validation technique, where the 10-Fold Cross Validation method was applied to evaluate the performance of the models. This technique involved dividing the data into 10 equal parts, the model is trained on 9 parts and tested on the remaining part, with the process repeated 10 times. This method aims to achieve

a comprehensive and accurate evaluation of the models and reduce the bias resulting from using a specific set for training or testing.

2.3. Machine Learning Models

The following machine learning algorithms were employed to build predictive models and analyze repeated purchase behavior:

Random Forest: The basic idea behind Random Forest is that it creates a number of random decision trees and combines their output in order to get better predictive accuracy. Its multi-data handling and stable performance render it appropriate for evaluating behavioral patterns.

Neural Networks: It has the unique ability to learn complex relations between the variables and non-linear ones. It is composed of many layers of neurons that learn how different features interact, enabling the models to predict more accurately.

AdaBoost: It is an algorithm for cumulative performance improvement that combines the least performing baseline models to obtain a final high-accuracy model. This approach facilitates incremental improvement in the accuracy of models and this makes it suitable for binary class prediction and supporting predictive analysis.

2.4. Model Evaluation Metrics

we evaluated the performance of the applied models using several standard measures. These measures are (AUC - Area Under the Curve) which is useful in determining the model's ability to distinguish between different classes, CA (Classification Accuracy) which measures the percentage of correct predictions out of the total predictions, and the F1-Score was also used which balances between precision which is the percentage of correct positive predictions out of the total positive predictions, and recall which measures the model's ability to detect true positive cases.

3. Results

3.1. Performance of Algorithms

Machine learning algorithms (Random Forest, Neural Networks, and AdaBoost) were applied to analyze the data and predict repeat customer purchase behavior. The performance of each algorithm was evaluated using the metrics described above, and the results are summarized in the following tables:

Random Forest: This algorithm showed strong performance in terms of prediction accuracy and balance between different metrics. Table 4. presents the main results.

Table 4. Performance metrics of the random forest model.

Performance Metric	AUC	CA	F1-Score	Precision	Recall
Random Forest	0.870	0.930	0.963	0.937	0.990

The ROC (Receiver Operating Characteristic) curve of the Random Forest model shows high efficiency in distinguishing between classes, as evidenced by the large area under the curve, as shown in Fig. 1.

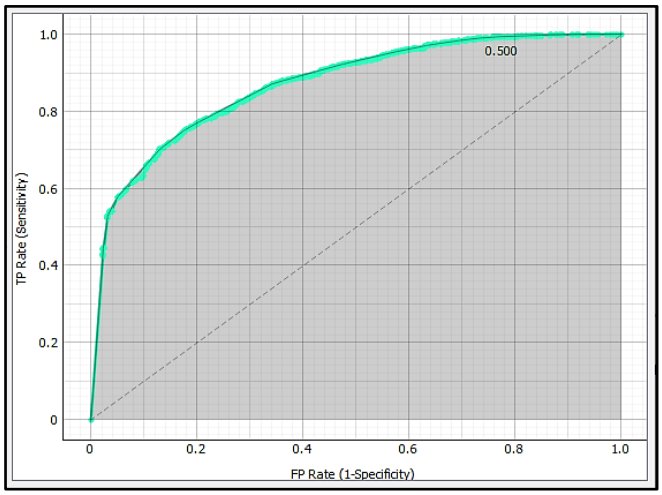

Figure 1. ROC curve for the random forest model.

Neural Networks: It achieved excellent results in almost all metrics, outperforming Random Forest in some points. Table 5. shows the results.

Table 5. Performance metrics of the neural network model.

Performance Metric	AUC	CA	F1-Score	Precision	Recall
Neural Network	0.886	0.932	0.964	0.937	0.993

The ROC curve of Neural Network recorded the highest accuracy in distinguishing between classes compared to the rest of the models, as the area under the curve is the largest. This performance is clearly shown in Fig. 2.

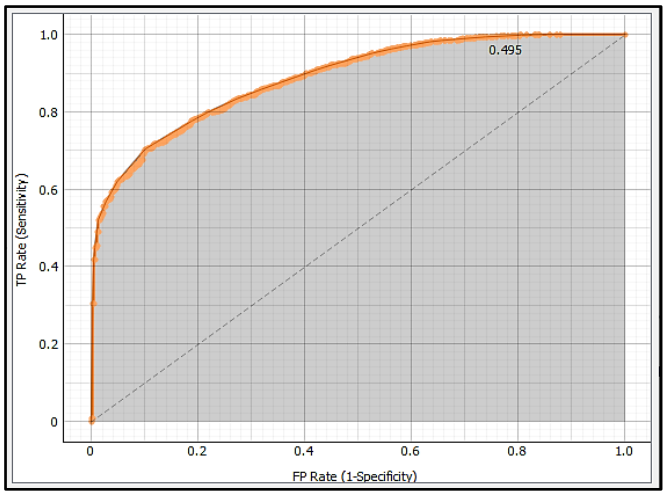

Figure 2. ROC curve for the neural network model.

AdaBoost: Its results were lower than expected compared to other algorithms, especially in terms of AUC. Table 6. shows the results.

Table 6. Performance metrics of the AdaBoost model.

Performance Metric	AUC	CA	F1-Score	Precision	Recall
AdaBoost	0.653	0.888	0.939	0.943	0.935

The ROC curve of the AdaBoost model reflects a relatively poor performance in class discrimination, with the curve approaching the random line compared to the Random Forest and Neural Network models. This is clearly shown in Fig. 3.

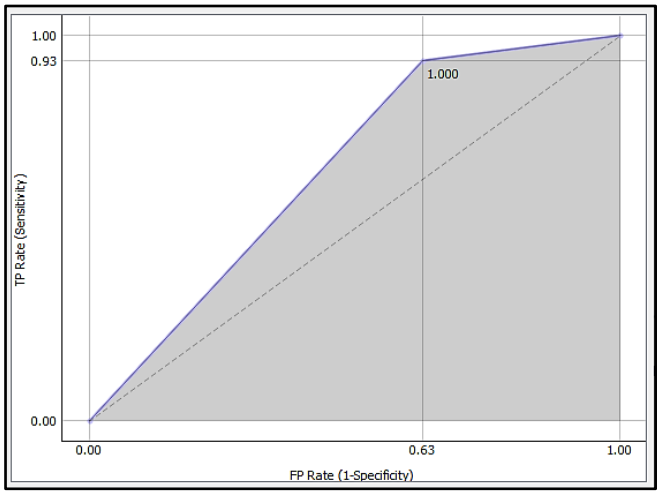

Figure 3. ROC curve for the AdaBoost model.

3.2. Comparison of Models

To make a comprehensive comparison between the three models, the results from the previous tables were analyzed with a focus on the key metrics. Table 7. shows the comparison of the models' performance.

Table 7. Models performance comparison.

Model	AUC	CA	F1-Score	Precision	Recall
Random Forest	0.870	0.930	0.963	0.937	0.990
Neural Network	0.886	0.932	0.964	0.937	0.993
AdaBoost	0.653	0.888	0.939	0.943	0.935

Key metrics analysis:

AUC (area under the ROC curve): Neural Networks achieved the highest AUC value of 0.886, indicating its high ability to discriminate between classes. Random Forest came in second with a value of 0.870, indicating its strong performance. AdaBoost, on the other hand, recorded the lowest AUC value of 0.653, reflecting its poor ability to discriminate.

Accuracy (CA): Neural Networks and Random Forest achieved very high accuracy ratios of 0.932 and 0.930, respectively, with a slight advantage for neural networks. AdaBoost's performance was significantly lower at 0.888.

F1-Score: Neural Networks achieved the highest balance between precision and recall at 0.964, followed by Random Forest with a slight difference at 0.963. AdaBoost's score was relatively lower at 0.939.

ROC curve for the three models:

Fig. 4. shows the ROC curves of the three models together, allowing us to evaluate the overall performance of each model. The ROC curves of Neural Network and Random Forest are seen to be close to the top left corner of the graph, indicating their high ability to discriminate

between classes. In contrast, the ROC curve of AdaBoost appears to have a less steep slope, reflecting its limited performance in this context compared to the other two models.

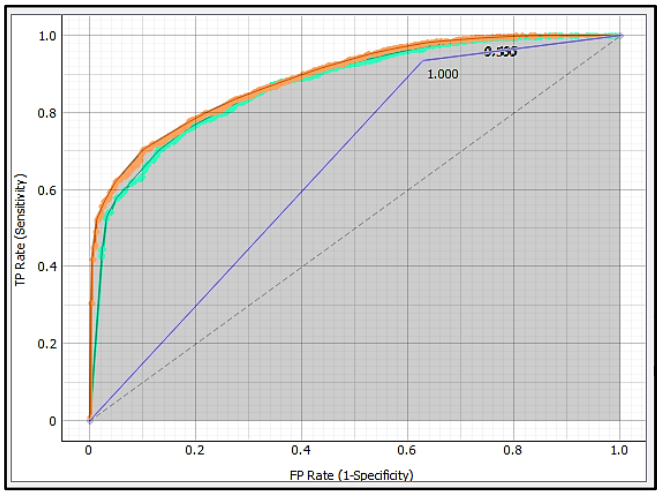

Figure 4. ROC curve for all models.

4. Discussion

The results indicated that using behavioral and demographic data in a predictive model improves the quality of predicting repeat purchase behavior. They also show that analyzing the complex patterns that arise from the interaction between customer behavior and their demographic characteristics is important in practice. Neural networks as one of the advanced algorithms showed good effectiveness in dealing with these patterns, which shows their role in discovering non-linear relationships between features.

On the other hand, the random forest algorithm showed good performance that was close to the performance of neural networks and can be considered a reliable option for predictive analysis. While the AdaBoost algorithm was less accurate, especially in distinguishing between classes; it may be the characteristics of the data or the sensitivity of the algorithm to noise that reduces its effectiveness for this type of analysis.

These results reflect the achievement of the study objectives, as the predictive model demonstrated its efficiency in identifying customers with a higher probability of repeat purchase. Selected features such as Customer Lifetime Value and Email Conversion Rate also highlight their importance in understanding the dynamics of customer engagement with digital platforms. In addition, the results confirm that the use of predictive models in digital marketing can enhance retention strategies and increase customer loyalty, which supports more accurate and effective marketing decisions.

5. Conclusion

In this study, a novel predictive model for repeat purchase behavior analysis requiring behavioral and demographic data has been developed by employing some advanced machine learning algorithms. Results suggest that neural networks were best for the model and therefore the most efficient in replicating repeat purchase behavior. Random forests exhibited very good performance, near that of neural networks, so it is trustworthy selection for the analysis, whereas AdaBoost performance is less than expected, probably due to the nature of the data adversely affecting algorithm efficiency.

This study underlines the effect of some attributes, like customer lifetime value and email conversion rate, on enhancing prediction in customer behaviour. Results offer companies a new way to enhance their digital marketing efforts, improve customer segmentation, and boost loyalty.

These results provide a solid foundation for future research that attempts to improve prediction models by experimenting with more complex algorithms or applying the same approach to other types of data. The study could be expanded to evaluate the impact of more features in a more comprehensive analysis of customer behavior.

AUTHOR STATEMENT

This study does not involve experiments on humans or animals, personal data collection, or clinical trials. It is based only on analyzing publicly available datasets using machine learning techniques. Therefore, no ethical approval was required for this study.

References

Alfian, G., Ijaz, M. F., Syafrudin, M., Syaekhoni, A., Fitriyani, N. L., & Rhee, J. (2019). Customer behavior analysis using real-time data processing: A case study of digital signage-based online stores. *Asia Pacific Journal of Marketing and Logistics, 31*(1), 265–290. https://doi.org/10.1108/APJML-03-2018-0088

Deniz, E., & Bülbül, S. (2024). Predicting customer purchase behavior using machine learning models. *ADBA Information Technology and Publishing Limited Company, 1*. https://doi.org/10.69882/adba.iteb.2024071

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data, 8*(1), Article 140. https://doi.org/10.1186/s40537-021-00516-9

Hai, R., Koutras, C., Ionescu, A., Li, Z., Sun, W., van Schijndel, J., Kang, Y., & Katsifodimos, A. (2023). Amalur: Data integration meets machine learning. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, 3729–3739. https://doi.org/10.1109/ICDE55515.2023.00301

Kuric, E., Puskas, A., Demcak, P., & Mensatorisova, D. (2024). Effect of low-level interaction data in repeat purchase prediction task. *International Journal of Human–Computer Interaction, 40*(10), 2515–2533. https://doi.org/10.1080/10447318.2023.2175973

Li, S. (2024). Machine learning-based prediction mechanism of repeated purchase behavior of e-commerce customers. In *Proceedings of the 2024 IEEE 4th International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB)* (pp. 522–526). https://doi.org/10.1109/ICEIB61477.2024.10602662

Lyu, Q. (2023). The construction of a model for predicting users' repeat purchase behavior and its impact on the economic efficiency of enterprises. *WSEAS Transactions on Computer Research, 11*, 303–315. https://doi.org/10.37394/232018.2023.11.28

Nair, S. (2024). *E-commerce customer engagement and demographics dataset*. Kaggle. Retrieved from https://www.kaggle.com/datasets/noir1112/e-commerce-customer-engagement

Puri, S., Sehgal, V., & Sharma, V. (2013). Customer centricity with predictive analytics in Indian retailing. *International Journal of Intercultural Information Management, 3*(3), 207–218. https://doi.org/10.1504/IJIIM.2013.057738

Sasmal, S. (2024). Predictive analytics in data engineering: An AI approach. *International Research Journal of Engineering and Applied Sciences, 12*(1), 13–18. https://doi.org/10.55083/irjeas.2024.v12i01003

Shrivas, G. (2023). A study of impact and applications of predictive analytics in sales forecasting. *International Journal for Research in Applied Science and Engineering Technology, 11*, 1109-1116. https://doi.org/10.22214/ijraset.2023.57535

Wang, S. (2023). Research on predicting the impact of promotional activities on consumer behavior in omnichannel retailing. *Advances in Economics and Management Research, 7*(1), 148. https://doi.org/10.56028/aemr.7.1.148.2023

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). https://doi.org/10.15837/ijccc.2022.6.4988

Zhao, B., Takasu, A., Yahyapour, R., & Fu, X. (2019). Loyal consumers or one-time deal hunters: Repeat buyer prediction for e-commerce. In *2019 International Conference on Data Mining Workshops (ICDMW)*, 1080–1087. https://doi.org/10.1109/ICDMW.2019.00158