

# Leveraging knowledge of Previous Baskets to Anticipate online Buyer Behaviour

Priyank Sirohi

Ph.D. Research Scholar

Shobhit Institute of Engineering and Technology (Deemed-to-be University), Meerut, India  
priyanksirohi01@gmail.com

Pradeep Kumar

Department of Computer Science and Engineering,  
JSS Academy of Technical Education  
Noida, Uttar Pradesh, India  
pradeep8984@jssaten.ac.in

Niraj Singhal

Director, Sir Chhotu Ram Institute of Engineering and Technology, Chaudhary Charan Singh University, Meerut, India  
drnirajsinghal@gmail.com

Mahboob Alam

Manav Rachna International Institute of Research and Studies Faridabad  
mahboobalam.set@mriu.edu.in

**Abstract**— Customer behavior prediction is an important task for any company to improve the sale of product. Retailer used Market Basket Analysis pattern of customer to identify the behavior of customers. On the basis of previous basket analysis one most important feature suggestion for customer for next purchasing. Market Basket Analysis was successfully applied to evaluate a large amount of data to understand customer behavior of purchasing patterns. To facilitate reordering and maintaining adequate product stock, this paper will explain how Instacart can use its consumer transaction history and concentrate on descriptive analysis of customer buying habits, objects frequently purchased around each other, and units purchased from of the shop. Additionally, to locate client subsets and cluster with similar purchasing habits and to visualize the data to make useful recommendations geared toward enhancing revenue and satisfaction using segment and forecasting framework. In this paper design a machine learning framework to forecast whichever previously purchased item will be in the customer future purchasing. Proposed model has high accuracy on the same data set is 74.16 percent.

**Keywords**— Customer behavior prediction, machine learning, Market Basket Analysis

## I. INTRODUCTION

The method of anticipating client interests is used in several business methods. Businesses invest significantly in tactics ranging from conducting customer reviews to employing machine learning to build sophisticated models to fully understand consumer behavior. One of the more popular methods is the market basket research, a data mining approach that identifies commodities with significant associations. It is important to analyze large data sets, such as payment records, to find categories and items that are most likely to be bought concurrently.

**Market Basket Research Categories:** There are two distinct kinds of market basket assessment. Predictive Market Basket Assessment: Predictive market basket study assesses the goods purchased to find cross-sell opportunities.

**A. Differential Market Basket Assessment:** This kind considers information from various shops and transactions made by various customer groups at various points throughout the day, month, or year. Investigators can identify the causes of an anomaly if a rule applies to one aspect (such as a store, time frame, or client category) yet not the other. Such discoveries may result in the latest product proposals that drive revenue.

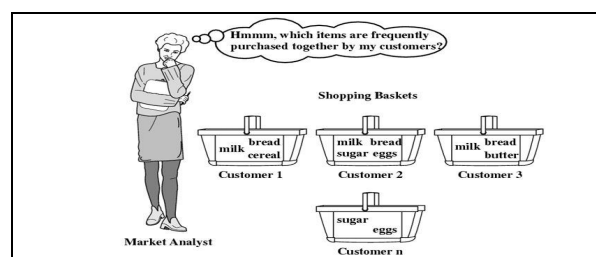


Fig. 1 Market Basket Analysis

To identify which products should be promoted or cross-marketed, market basket research looks at the products consumers usually buy combined. The phrase refers to the number of groceries that supermarket patrons load into their trolleys when out buying. When attempting to identify a relationship between various items in a collection or identify recurring patterns in a transaction database, database system, or other data store, association rules mining is used. Market Basket Analysis is a fundamental strategy in use by big businesses to evaluate customer buying behaviors by identifying links between the various things that customers place in their "shopping carts," which is how such trends are most frequently discovered. By supplying details on the items consumers frequently purchase together, the detection of these linkages may help businesses develop marketing strategies. The strategies could include

- 1) Modifying the layout of the business in accordance with tendencies.
- 2) Analyzing the behavior of customers.
- 3) Designing a catalogue.
- 4) Cross-selling at online retailers.
- 5) Personalized mailings that include add-on deals, etc.

Metrics used in customer behavior analysis

**Support:** It is the item's inherent attractiveness. The proportion of transactions product P1 to all purchases is mathematically related to the supporting of product P2.

**Confidence:** Probability of a buyer purchasing both P1 and P2. It is the proportion of transactions that involve both P1 and P2 to those that only involve P2.

$\text{Confidence}(P1 \Rightarrow P2) = \text{Support}(P1, P2) / \text{Support}(P2)$

Lift: P1 will sell more because of the sale of P2

$\text{Lift}(P1 \Rightarrow P2) = \text{Confidence}(P1, P2) / \text{Support}(P2)$

$\text{Lift}(P1 \Rightarrow P2) = 1$  indicates that there is no association among the elements in the collection.

$\text{Lift}(P1 \Rightarrow P2) > 1$  identifies a positive relationship inside the information set, i.e., indicating goods P1 and P2 in the information set are more likely to be purchased jointly.

$\text{Lift}(P1 \Rightarrow P2) < 1$  signifies the existence of a negative association among the goods, i.e., it is improbable that goods P1 and P2 would be purchased simultaneously.

## II. RELATED WORKS

Customer-generated content continues to be a significant source of data on customer opinions that is valuable for service assessment. On the basis of the history of customer purchasing, there are several related works available. In this section will discuss the available related work. Cihan H *et al.*[1] Evaluation of the buy forecasts and potential income gains from the framework created for price forecasting. Our suggested approach revealed an enhanced revenue-generating strategy featuring fewer mistakes in predicting client sales for only one similar product given at a set price for a particular set of buyers. The supervised learning will result in more detailed findings as time goes on and more data is gathered, which will be useful in figuring out the precise consequences for buying behaviour.

Wang XingFen *et al.*[2] proposed consumers' consumption behaviour model forecasts how consumers will behave in terms of buying a specific product at a particular time in the future. The purchase data of the selected product is included in the data file created by the user modelling training, which offers applicable technologies for precise advertising of big data analysis in the e-commerce system.

Bettina von Helversen *et al.*[3] proposed a framework for the impact of customer reviews on both elder and young consumers' online shopping choices. The results demonstrate not only that reviews and ratings have distinct effects on consumers to purchase online but also that older and younger persons place varying values on rankings and reviews. The older individuals in our group assigned minimal weight to these sorts of consumer data, in contrast to students, who were substantially affected by average buyer rankings and positive mood evaluations. Even though these were not typical of the customer reviews, actually effect bad review had a noticeable impact on both older and younger people.

Praphulla Kumar Jain *et al.*[4] Suggested User Suggestions Predictions are online tools that reduce the burden of online information burden both for customers and advertisers by understanding and recommending usage patterns. With the advent of machine learning and online customer recommender systems, the tourism sector has made an effort to provide customers with better experiences that are based on intelligent, information judgement methods. Airlines' marketing departments employ ML algorithms to gain a greater understanding of passengers' attitudes towards airline services and help them make the most of their individual tourism experience. The proper choice of these algorithms and efficient fine tuning are of vital importance to build reliable user suggestion guidance.

Meher Neger *et al.*[5]. Author focusing the factors that have a large impact on online purchasing decisions. The items, savings in time, payment, and management factors have a big impact on how individuals buy online. Whenever provided digitally, the calibre of the products and services being offered should be preserved. Additional convenient, affordable, and trusted payment options are required. Because most online customers hesitate to utilise payment methods such as debit or credit cards and are accustomed with electronic systems for payment, companies can accept money upon shipment. Lin Guo *et al.*[6] discussed A significant amount of actual network information is gathered to facilitate the study of customer behavior, as mentioned. This process produces precise information. Time-series data includes statistics on browsing habits. To analyse particular customer tastes and consumption behaviours prior to developing a forecast regarding future financial behaviours, this study will use the approach of time-series analysis. The outcomes of the predictions can be applied to promote marketing and product recommendations, for example. Using a range of data, the study illustrates how to predict consumer behaviour and verifies the viability and effectiveness of the SeqLearn methodology.

Maria Nicola *et al.*[7] Because of worries about an emerging economy and financial meltdown, times like this call for powerful and powerful leadership in the medical field, in business, in government, and in wider society. For those who could slip between the cracks, immediate adaptation assistance must be implemented. To maintain equilibrium and regenerate the company following this calamity, preparation for both the short and long term is required. Strong and long-lasting business plans must be accompanied by a complete sustainable livelihood plan that incorporates sector objectives and an environment that encourages innovation if one is to be successful. In order to ensure that the "what it takes" pledge is kept, officials and financial institutions should closely assess the situation on a regular basis. Maheswari *et al.*[8] Data mining methods were utilised to forecast the behaviour of consumers on the internet by utilising a number of classifiers and Support Vector Machines. The linear kernel is used to create SVM models. According to the trial findings, younger customers were drawn to online shopping in past years. Clients will spend more when there are special deals.

Shengyu Gu *et al.*[9] The survey included consumers of internet services from the best ten countries for online

shopping development. The toolset of measurements used to examine the dynamism of online shoppers' behaviour over the research time gave researchers the possibility to spot significant patterns and track behavioral shifts. It reveals the key elements that have the most impact on what they buy. According to the report, the COVID-19 epidemic caused the normal changes in online consumer buying habits. Nowadays, customer service and comprehension are more important. Online shoppers are increasingly spiced up, that has impacted the way that they behave when making purchases. The current research showed how different pandemic-related online consumer purchasing factors have an influence. The capacity of consumers to make snap decisions grows more important while developing web-based services and product purchasing.

Weiwei Zhang *et al.*[10] Throughout this study, we propose an enhanced dense forest model to forecast e-commerce customers repurchase behaviour. Its classifier analyses the customer attributes, product attributes, and characteristics of consumer interacting behaviour in comparison to the current repurchase behaviour prediction. Additionally, by contrasting with conventional machine learning algorithms, the enhanced deep forest's accuracy in forecasting e-commerce customers' purchasing behaviour is confirmed. The utilisation of behavioural data alone is not sufficient to support consumers' research, which is a complex topic. In order to create featured projects with more valuable features, several customer personality traits, including such age, age, degree of spending, user profiling, etc., can be taken into account.

Jagdish Sheth [11] Lockdown and social exclusion procedures implemented in response to the COVID-19 virus have had a major negative impact on consumer behaviour. Every application has a specific time and place. Customers have learned how to change in distinctive and inventive ways since time is flexible while the place is fixed. More and more individuals are conducting work from home, learning from residence, and taking breaks from their jobs, blurring the line between work and personal life. The store has to greet the consumers as they do not have the means to travel to the establishment. Buyers are inclined to endorse modern technology, which renders employment, education, and shopping more enjoyable if they become accustomed to spending a large portion of their lives in confinement. It's possible that using technology will alter the way people live.

Tao Chen *et al.*[12]. The present research used visual tracking equipment to investigate how online product reviews influence consumers' purchase decisions. Their results showed that consumers, especially female consumers, paid far more attention to unfavourable comments than to positive ones. The study also discovered a high correlation between consumers' visual surfing behaviours and their likelihood to purchase. Additionally, it was shown that users were unable to detect fake comments. First, the study that is now available indicates how gender affects this influence and explains it from the standpoint of cognitive prejudice, which is essential for the consumer behaviour assumption. It provides a full knowledge of the mechanism by which review sites affect purchasing choices.

especially, it's clear from the study that consumers' sensitivity to both positive and negative feedback is moderated by their gender, with female customers giving much more weight to adverse than to good input.

Kiran Chaudhary *et al.*[13]. Big data has been applied to the processing and analysis of information in order to forecast customer experiences on social media. Depending on a few metrics and characteristics, we examined customer behaviour on social media sites. We looked at how consumers felt and acted around social media sites. We use a variety of data preparation techniques to identify outliers, noise, errors, and redundant records to produce high-quality results. Utilizing machine learning, we created mathematical models to forecast user behaviour on social media. The shopping on the social media site can be predicted using this model, which is prescriptive. 20% of the information is utilized for assessment, while 80% are utilized for learning.

TABLE I  
RELATED WORK

Study	Methodology	Key Findings
Wang <i>et al.</i> (2015)	Association rule mining	Discovered frequent item sets for personalized recommendations, but lacked predictive power for future
Chen <i>et al.</i> (2017)	Collaborative filtering with matrix factorization	Successfully predicted future purchases by considering user-item interactions but struggled with cold-start problems.
Rendle <i>et al.</i> (2010)	Factorization Machines	Achieved state-of-the-art performance in predicting user preferences and improving recommendation quality
Koren <i>et al.</i> (2009)	Matrix factorization with bias	Demonstrated the effectiveness of matrix factorization in enhancing recommendation accuracy by capturing latent user and item characteristics.
Xie <i>et al.</i> (2021)	Sequential recommendation with recurrent neural networks	Leveraged sequential behavior data to model temporal dependencies in user behavior, resulting in improved recommendation accuracy.
Zhang <i>et al.</i> (2018)	Deep learning-based approach (DeepFM)	Combined the strengths of factorization machines and neural networks, achieving better accuracy and generalization in recommendation systems.

### III. PROBLEM STATEMENT

Customer behaviour prediction in advance is a major task for any organization. The aim how to predict accurate and efficiently customer behaviour prediction on the basis of

past behavior of customer. A model is proposed based on machine learning has been proposed.

#### A. Machine learning based model

A machine learning model is used to predict the behavior of customers by leveraging anonymized transactional data from past customer transactions to predict whether previously purchased items would be included in a user's forthcoming transaction. It would corroborate a user's recommendation of the goods. To create this kind of prototype, information from past orders must first be extracted to comprehend customer purchasing behavior and the level of popularity of a given product. I take the features listed below from the customer's transactional data. Create a data frame using the characteristics that display together all goods the customer has previously purchased, user level features, product level characteristics, account and departmentalization characteristics, user-product level characteristics, and details about the existing order, such as the date of the week, minutes of the day, etc. The goal is to show the number of already own things the customer ordered these times by reordering. Proposed machine learning model based on the XGBoost algorithm because it manages large data easily. AUC Rating is used to compare the XG Boost-based models with the neural network-based model on the basis of Evaluation metrics while altering the threshold. Following a study of the two algorithms using the Confusion Matrix, ROC curve, and categorization summary, the results are displayed below. To comprehend significant features that aid in predicting item reordering, the characteristic critical graph from of the XGBoost model is also displayed. These models work about equally well, with XGBoost marginally outperforming the other with respect to ROC-AUC.

### IV. METHODOLOGY

The following essential phases are part of the methodology:

- A. *Data Collection*: Gather comprehensive data on customer transactions, including information on products purchased, order timestamps, and customer demographics. This dataset serves as the foundation for our predictive model.
- B. *Feature Engineering*: Extract relevant features from the transaction data, including customer purchase history, product associations, and temporal patterns. These features are crucial for training our XGBoost model.
- C. *XGBoost Model Training*: Employ the XGBoost algorithm, a powerful gradient boosting technique, to develop a predictive model. By training on historical data, the model learns to identify patterns and relationships that indicate future purchase Behaviour.
- D. *Model Evaluation*: Rigorously evaluate the performance of our XGBoost model using appropriate metrics such as accuracy, precision, recall, and F1-score. Cross-validation techniques are also applied to ensure robustness.
- E. *Recommendation System*: Once the model is validated, it is integrated into an online retail platform to provide real-time product recommendations. This personalized approach aims to increase sales and customer Satisfaction.

- F. *Continuous Improvement*: Emphasize the importance of an iterative process, continually updating and fine-tuning the model with new data to adapt to changing customer preferences and market trends.

### V. IMPLEMENTATION

#### A. Description of the data

**Aisles**: This file comprises numerous aisles and has a variety of 134 unique aisles.

**Departments**: This file contains information about 21 various departments in all.

**Orders**: All the purchases made by various users are contained in this database. This is what we may infer from the study below:

- 3421083 orders have been placed by a total of 206209 clients.
- Prior, Train, and Test are the three sets of guidelines that are accessible. Despite being similar, the ordered structures in the Training and Testing groups diverge from the order distribution in the previous group.
- A Customer total number of purchases can vary from zero to hundred.
- Based solely on the "Orders versus Day of Week" plot, we may denote one and zero as Sunday and Saturday respectively, presuming that almost all consumers buy products on Saturday.
- During the day, the vast bulk of transactions are made.
- According to the surges at seven, fourteen, and thirty inside the "Orders versus Days since" Last Buy" table, customers typically make one transaction per week.
- The "Days of the week" and "Time of the day" hotspots indicate that Saturday lunchtime and Sunday mornings are the busy hours for purchasing.

**Products** This file includes a list of all 49688 goods, together with information on each aisle and division. Various aisles and sections have varying numbers of merchandise.

- 1) *Order products prior*: This file contains details on the items that were purchased and the order in which they were put in the shopping basket. It also lets us know whether the goods were purchased again. This file contains details just on 3214874 purchases that total of 49677 goods were purchased online. According to the "Count VS Things in Basket" graph, most customers only purchase 1 to 15 things, with orders including a total of 145 products. Throughout this collection, 58.97% of the total order is repeated.
- 2) *Order Products Train*: File contains details on the items that were purchased as well as the order in that they were put in the shopping basket. It also lets us know whether the goods were purchased again. This document includes details about all 131209 purchases that resulted in purchases for 39123 various items. Based on the "Count versus Things with Basket" graph, most customers only order 1 to 15 things, with orders including a total of 145

products. Throughout this set, 59.86% of the objects are repeats.

Customer Segmentation: The first two principal components of the grouping can be seen here. The grouping produces 5 tidy groups, and after examining the most prevalent commodities there, we can draw the concluding remarks:

- 5428 buyers in Group 1 show a very strong affinity for the bubbling water and water-based seltzer sector.
- Group 2 generates 55784 buyers who primarily buy fresh vegetables and fruits.
- 7948 buyers from Group 3 are found to primarily purchase bagged produce and fruits and veggies.
- 37949 buyers fall under Group 4 and they strongly like fruits, preceded by fresh veggies.
- 99100 buyers from Cluster 5 place orders for goods across numerous aisles. The average number of orders is modest in comparison to certain other clusters, indicating that they are likely infrequent Instacart customers or new buyers with little purchases yet to place.

**Weighted avg** 0.90 0.73 0.78 2118666  
**Accuracy Score:** 0.7251921728106271  
**F1 Score:** 0.3592928105783494  
**Area under curve:** 0.8344086677839623

C. *Confusion matrix and ROC Curve for neural Network based model*

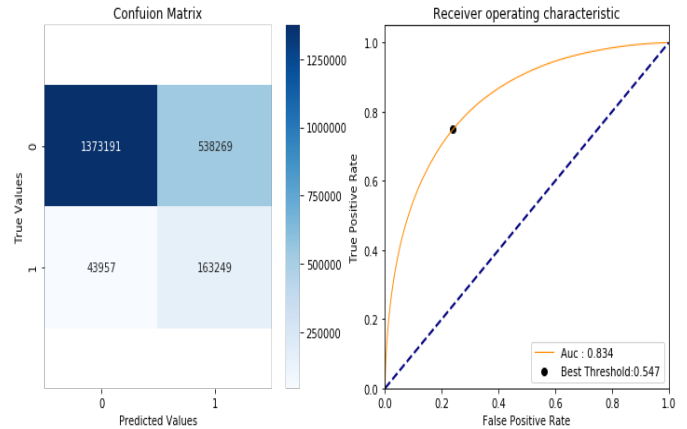


Fig. 4 Confusion matrix and ROC Curve of ANN

The Analysis found that 85% of customers only purchased 10,000 out of 49688 products in the below graph of total combined customers per product vs. products. Only certain 10,000 items if shelf efficiency is what we're after. Here, I'm assuming that the other 39688 products' margins won't be very great. When considering products with high income, high reorder rates, and strong overall product sales, prices for these products should be recognized.

C. *Number of Product purchased Versus Product*

A threshold effect was seen in the accompanying plot of repurchase % and number of product purchases. Several individuals only ever test a new product once before giving up on it. Additionally, some customers routinely purchase products. Buyers. Add-to-cart orders and the mean reorder % are shown in a graph. Findings show that the percentage of repeat orders is bigger the smaller the add-to-cart order is. This really is understandable given how frequently we purchase daily necessities initially.

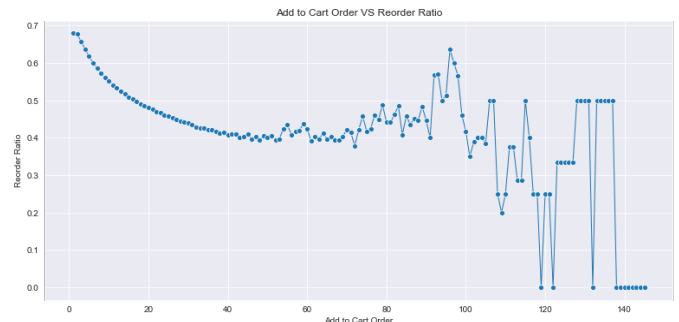


Fig.5 Reorder Ratio versus Add to cart Order



Fig. 2 Distribution of order in various groups.

B. *Confusion matrix and ROC Curve for Xg Boost algorithm*

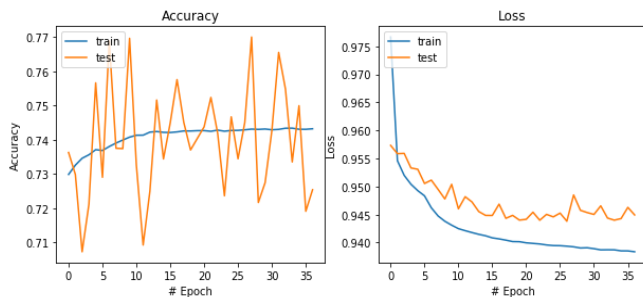


Fig3 Artificial Neural Network based Accuracy and loss analysis

TABLE II

CLASSIFICATION REPORT

	Precision	Recall	f1-score	Support
0.0	0.97	0.72	0.83	1911460
0.1	0.23	0.79	0.36	207206

**Accuracy** 0.73 2118666  
**Macro avg** 0.60 0.75 0.59 2118666

#### D. Confusion matrix and ROC Curve for Xg Boost algorithm

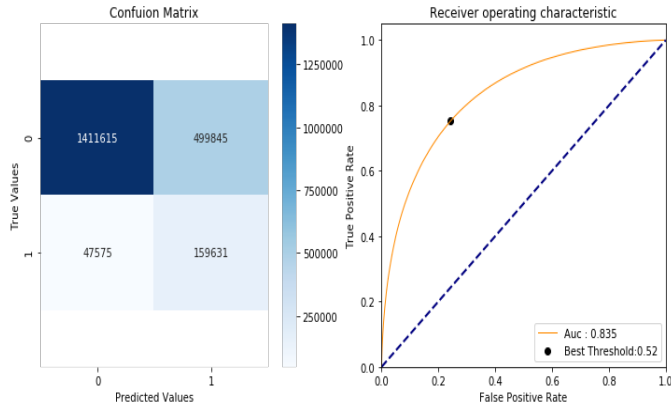


Fig.6 Confusion matrix and ROC Curve of Xg boost

TABLE III  
CLASSIFICATION REPORT

	Precision	Recall	f1-score	Support
0.0	0.97	0.74	0.84	1911460
0.1	0.24	0.77	0.37	207206

Accuracy 0.74 2118666  
Macro avg 0.60 0.75 0.60 2118666  
Weighted avg 0.90 0.74 0.79 2118666  
Accuracy Score: 0.7416204347452595  
F1 Score: 0.3683727134058397  
Area under curve: 0.8347162433873154

#### VI. CONCLUSION AND FUTURE SCOPE

Customer behavior prediction is the most important activity for any e-commerce company for the highest sale of any product and commodity. Here in this proposed model utilized Xgboost machine learning algorithm to predict the customer behaviour prediction on the bases of past behaviour of customer purchasing. For the analysis of proposed model utilized 3 million orders from more than 2 lakh Instacart customer orders. Here for the comparison of proposed model same data have to apply on the neural network-based model. After the analysis, it was found that the accuracy of the neural network-based model was 72.51 percent. Proposed model has high accuracy on the same data set is 74.16 percent. So the proposed model has higher accuracy. For the future point of view utilize collaborative filter to predict the customer behavior.

#### REFERENCES

- [1] R. Gupta and C. Pathak, "A machine learning framework for predicting purchase by online customers based on dynamic pricing," *Procedia Comput. Sci.*, vol. 36, no. C, pp. 599–605, 2014, doi: 10.1016/j.procs.2014.09.060.
- [2] X. Wang, X. Yan, and Y. Ma, "Research on User Consumption Behavior Prediction Based on Improved XGBoost Algorithm," *Proc. - 2018 IEEE Int. Conf. Big Data, Big Data 2018*, pp. 4169–4175, 2019, doi: 10.1109/BigData.2018.8622235.

- [3] B. von Helversen, K. Abramczuk, W. Kopeć, and R. Nielek, "Influence of consumer reviews on online purchasing decisions in older and younger adults," *Decis. Support Syst.*, vol. 113, no. March, pp. 1–10, 2018, doi: 10.1016/j.dss.2018.05.006.
- [4] P. K. Jain, E. A. Yekun, R. Pamula, and G. Srivastava, "Consumer recommendation prediction in online reviews using Cuckoo optimized machine learning models," *Comput. Electr. Eng.*, vol. 95, no. August, p. 107397, 2021, doi: 10.1016/j.compeleceng.2021.107397.
- [5] Meher Neger and Burhan Uddin, "Factors Affecting Consumers' Internet Shopping Behavior During the COVID-19 Pandemic: Evidence From Bangladesh," *Chinese Bus. Rev.*, vol. 19, no. 3, pp. 91–104, 2020, doi: 10.17265/1537-1506/2020.03.003.
- [6] L. Guo, B. Zhang, and X. Zhao, "A Consumer Behavior Prediction Model Based on Multivariate Real-Time Sequence Analysis," *Math. Probl. Eng.*, vol. 2021, 2021, doi: 10.1155/2021/6688750.
- [7] M. Nicola, Z. Alsafi, C. Sohrabi, A. Kerwan, and A. Al-Jabir, "Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information," *Int. J. Surg. J.*, vol. 78, no. January, pp. 185–193, 2020.
- [8] K. Maheswari and P. P. A. Priya, "Predicting customer behavior in online shopping using SVM classifier," *Proc. 2017 IEEE Int. Conf. Intell. Tech. Control. Optim. Signal Process. INCOS 2017*, vol. 2018-Febru, pp. 1–5, 2018, doi: 10.1109/ITCOSP.2017.8303085.
- [9] S. Gu, B. Ślusarczyk, S. Hajizada, I. Kovalyova, and A. Sakhbieva, "Impact of the COVID-19 pandemic on online consumer purchasing behavior," *J. Theor. Appl. Electron. Commer. Res.*, vol. 16, no. 6, pp. 2263–2281, 2021, doi: 10.3390/jtaer16060125.
- [10] W. Zhang and M. Wang, "An improved deep forest model for prediction of e-commerce consumers' repurchase behavior," *PLoS One*, vol. 16, no. 9 September, pp. 1–16, 2021, doi: 10.1371/journal.pone.0255906.
- [11] J. Sheth, "Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information," *J. Bus. Res.*, no. January, pp. 280–283, 2020.
- [12] T. Chen, P. Samaranayake, X. Y. Cen, M. Qi, and Y. C. Lan, "The Impact of Online Reviews on Consumers' Purchasing Decisions: Evidence From an Eye-Tracking Study," *Front. Psychol.*, vol. 13, no. June 2022, doi: 10.3389/fpsyg.2022.865702.
- [13] K. Chaudhary, M. Alam, M. S. Al-Rakhani, and A. Gumaei, "Machine learning-based mathematical modelling for prediction of social media consumer behaviour using big data analytics," *J. Big Data*, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00466-2.