Analysis of Behavioural Data of Customer for the E-Commerce Platform by using Machine Learning Approach

Ayush Maurya (IEEE, Member)
Department of Computer Science and Engineering
Shambhunath Institute of Engineering and Technology, Prayagraj, U.P
aayushmaurya312@gmail.com

Saurabh Pratap (IEEE, Member)
Department of Mechanical Engineering, Indian Institute of Technology (BHU),
Varanai, U.P
s.pratapiitkgp@gmail.com

Prabal Pratap
Defence Research and Development Organisation, Kanpur, U.P,
India,
prabalpr@gmail.com

Ashish Dwivedi
Jindal Global Business School,
Sonepat, Haryana-131001
adwivedi@jgu.edu.in

2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES) | 979-8-3503-2391-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/CISES58720.2023.10183475

***Abstract*— This research analyses the effect of numerous factors that allow E-commerce platforms to obtain previous knowledge about customer purchase tendencies. This research improves customer behaviour, product discoverability, and warehouse maintenance. Demanding product categories assist maintain the warehouse and generate sales ideas by offering an appropriate discount to attract consumers and promote same-day delivery. The system uses an e-commerce company database to store consumer purchase information. Our model analyses the data to classify customers and commodities. Our study includes descriptive, predictive, and prescriptive analytics to anticipate e-commerce sales. The descriptive study includes data cleaning, preprocessing, and visualization to analyze gender, city tier, spending money, age, city, marital status, financial position, brand cost, and product categories these are the attributes of dataset. To advance to predictive analytics, machine learning techniques such as naive Bayes classification, support vector classification, logistical regression, decision tree, KNN classification, and random forest classification are applied to the dataset to anticipate product sales. This proposed research reduces e-waste and promotes sustainable development by anticipating product sales.**

Keywords- E-commerce Platform; Recommender system; Machine Learning; Customer behaviors

I. INTRODUCTION

Online sales have been enlarged rapidly in modern years. With increased competition, online shoppers are eager to expand the effectiveness of their e-commerce platforms by predicting sales and customer choice. The Paper gives an understanding of circumstances features and customer choice-related features that can help online shops` marketing strategies as well as bring a better user experience through personalized offers and discounts based on users' early purchase predictions.[1] described that stimulates prospective buyers, enlarging cross-trade and building customer allegiance through analysis of previous data by applying machine learning to amplify e-commerce. The product fulfilled information and data with different shapes and is usually fundamental for counsel making. Standard category allocation is the most common content information of e-commerce products mentioned by the e-commerce seller. In the spite of standard systematic attributes, algorithms merge both systematic and social tagging information[ibid].[2] The use of data mining models helps sellers to the rapid growth of their sales by analyzing their consumer's online deportment and etiquette. In e-commerce behavioral and activity data become progressively paramount to understanding consumers' penchant and needs. Products browsed, bought, or added to the cart are incorporated in the data concerning the session branch in the e-commerce platform. These algorithms and models can predict consumers' purchase objectives, which can be used for product recommendation and product pricing systems or tactics.

In order to best serve consumers, buy prediction and recommender systems rely heavily on methodologies that require a thorough understanding of consumer preferences. The most crucial factor in developing effective customization strategies is awareness of the customer's intention to purchase from the moment they first land on the e-commerce site. Today, machine learning and artificial intelligence play a significant role in analysing customer behaviours concerning new product launches and anticipated increases in product demand in the future. By analysing data provided to the consumer even before they engage with any goods, sophisticated approaches based on machine learning could be effective tools for the development of purchase objectives. According to one of the findings, the first step in the process of making a purchase for a consumer is considering the lay, which involves the consumer evaluating numerous characteristics of the product, such as its price, quality, and material, among other things, before moving on to the next step, which involves being influenced by information, advertising, and marketing offers [1]. Various e-commerce models for C2C, B2B, and C2B transactions have developed moderately over time. Some emerging business ideas include selling through social media platforms like Instagram, WhatsApp, and mobile phones.

On social media platforms, online questionnaires are distributed that increase response rates, feedback, and promptitude, and reduce cost by the satisfaction sampling method. Google forum and disseminated through social media platforms used for the online questionnaire. On satisfaction samples with a willingly available list of targets socioeconomics as e-commerce customers [3]). Becoming less dependent on third-party distribution channels and connecting with customers online may allow small manufacturers to speed up their international processes. When involving proficiency and heuristics associated with e-

commerce, companies cultivate exercises, mental outlines, and inter-functional interactivity which in the end feed into their OMCs. The compound nature and situational character of these potentials make them unique and, thus, difficult to transfer and imitate. Overcoming liabilities, acquiring knowledge and understanding, and increasing performance are the important of physical presence in foreign markets. Optimizing international growth among small and medium enterprises digital market and e-commerce can be strong vehicles and interactions with remaining customers and to develop new customer relationships in the drowsing customer section [4]. A larger company that is in the authority of substantial resources can have a deep impact on markets. It is also crucial to protuberance that a sublimate degree of impetuous online shopping increases environmental waste; reckless purchasing decisions lead consumers to acquire material things that do not need, and then, as a result, cause consequences for sustainable commerce [5]. The appraisal inscription the following research Question:(Qs):

Q1: What are the ML algorithms used for the classification of product categories?

Q2: What are the best fit ml algorithms used for the classification of product categories?

Q3: How does the prediction of which product category on high sale impact e-commerce businesses?

II. Literature Review

The online e-commerce manager plays a significant role in balancing the pressure of increasing their sales with an inclination and need to act appropriately and ethically and in avoiding the impulsive shopping tendencies that are associated with negative consumer-side effects such as overconsumption, challenges for sustainability, and other similar issues. This is a difficult task that requires the manager to strike a delicate balance. The authors Mohammed and Tejay [6] contend that impulsive buying leads to a challenge for the development of sustainable practises, as well as financial repercussions and the waste of items.

Affective and cognitive states are impacted by virtual interactions of customers which also affects shopping behaviors. To appeal to the user's e-commerce platform to create, design, and manage virtual shopping environments used a huge number of financial resources. Enchanting of the e-commerce platform plays a crucial role in making sales growth and visits to the customer on a website which places a role in a recommendation system for better placement of the product on the user web page such that customers may achieve a flow state where they become deeply engaged in online wayfinding.

Gupta and Pathak [7] focused on how the involvement of machine learning and artificial intelligence affect e-commerce and how they help in understanding purchase intention by analyzing data that defines the customer even before they interact with any product and early purchase prediction the main aim is a method involving early purchase intention prediction. Product quality can throw back the benefits brought by the product itself to customers, while the brand value will empower customers to obtain auxiliary products. Requisite in shopping for online shoppers and e-commerce platforms.[8] there is a lot of research topic in this field like business-to-business, strategy, and business-to-consumer with involvement in machine learning and adoption of technology. It provides a general understanding of e-commerce as an important research system, management information system, and international business research [8].

A huge amount of data on items purchased online by users brought resonance of e-commerce. There are extremely unnecessary, reclaim efficiency of users. It results in vanquishing the satisfaction of users as the burden to e-commerce websites. Many e-commerce researchers have dedicated more attention to searching the items that interest users effectively and accurately out of massive data. Social networks generate a large amount of data that includes people's connections, locations, interests, etc. With swift development, data science research takes a huge interest in the study of social recommendation systems [9]. The incorporation of trust and relationship information provides more accuracy and personalized recommendation results. Social media or networks perform a crucial role in establishing an e-commerce market and it`s an easy way to advertise their product to the open world of the market. There is a giant media platform that offers a business account that enables us small and medium enterprises.

[10] Predicting sale for such information support planning the inventory at the warehouse and point of sales, as well strategic decisions during manufacturing processes for a company. [11] studied the sequel of online behavioral advertisements inferior on advertiser-controlled factors and consumer-controlled factors. Machine learning embraces numerical, approximation, and round-off errors, which are trained in predicting and forecasting models. The dependability of prediction by including information with a physical meaning increases by integrating machine learning with data analytics [12]. [13] analysis of the relationship between the presence of customer reviews on product pages, review informativeness fully mediated the influence of product elucidation readability and product sales are trivial.

III. Proposed Framework

The study suggests taking decisions for e-commerce businesses. The structure uses data analytics and machine learning. It helps e-commerce companies to analyze customer preference for which product categories have high sales which help businesses to maintain their warehouse. Data analytics helps to make the data sensible for individuals or organizations. The suggested structure is shown in the figure.

1. Descriptive analytics
2. Predictive analytics
3. Prescriptive analytics

In the descriptive analysis, methods are utilized to find usable columns or features such as gender, tier, spend money, age, city, marital status financial status, and cost over brand and product categories. The information allows critical business and analytical decisions. Predictive analytics utilizes models to predict classification. KNN, Navies Bayes, support vector, logistics regression, decision tree, and random forest are trained and tested. In prescriptive analysis results of all models are compared through the F1 score and analysis the best models from them by changing different attributes.

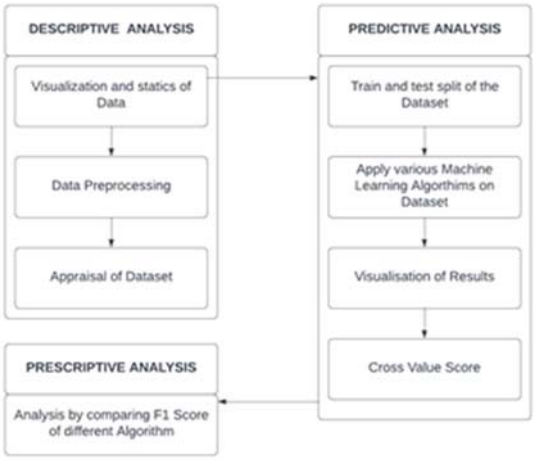

Fig. 1. Step by Step procedure for Data Analysis

IV. Result Analysis

The data collection of users that visited the website of e-commerce for the purchase of products. This includes age, gender, current state, tier, marital status, financial status, cost over the brand, product category, time to buy, shopping monthly frequently, average money spends, and sub-products. The experiment results consist of different classification algorithms applied to the dataset and pick out the best f1 score algorithms by comparing different models. Seaborn module makes visualizations and sci-kit-learn library used for performing machine learning algorithms on datasets.

A. Descriptive Analytics

We have conducted an extensive study of our obligatory datasets, prepared with attribute optimal, organized, and understandable data format. First, we performed data pre-processing by using python libraries such as pandas, and NumPy.

1) Data pre-processing and initialization

Data pre-processing modifies raw data into treasured data. This step is important for machine learning because we prepare data for processing. Before analysis, the data is filtered. It includes removing unrelated and missing data from the dataset and handling categorical data which is done by using either panda (get dummies) or one hot encoder which convert categorical data into binary data, replacement of unusable value from columns, and drops non-impactful columns.

This reduces the amount of time needed to train the classification model, boosts creativity, and reduces overfitting.

V. Dataset Statistics And Visualization

The complete dataset includes 255 different review submissions. Python's panda's library is utilised to generate the resulting description of the processed dataset. Information regarding the complete initial dataset that was utilised for predictive analysis can be found in Table I. The number of attribute entries, the count, the mean, the standard deviation, as well as the minimum and maximum values are included in the table.

TABLE I. SHOW THE COUNT, MEAN, STANDARD DEVIATION, MINIMUM OF AGE AND MONTHLY SHOPPING FREQUENCY

	Age	Shopping monthly Freq
count	254.000000	254.000000
mean	32.200787	2.543307
std	13.664350	1.370242
min	15.000000	0.000000

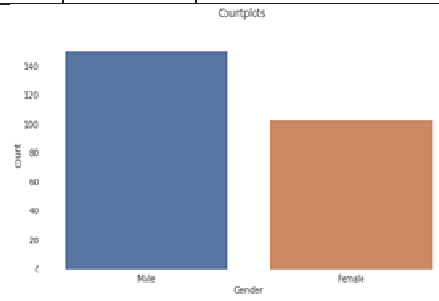

Fig. 2. Count plot of gender versus count value

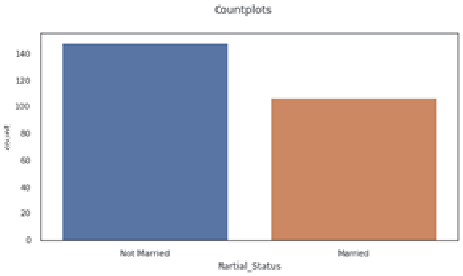

Fig. 3. Count plot of marital status contra to count value

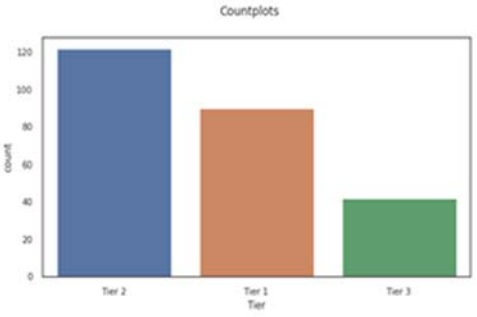

Fig. 4. Count plot of Tier of city

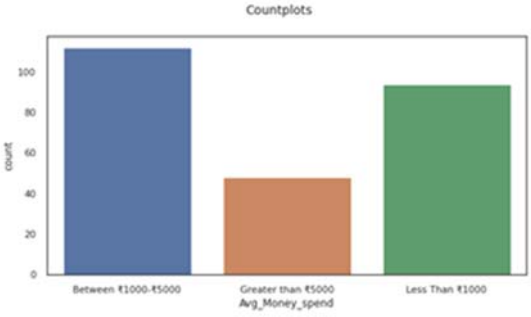

Fig. 5. count plot of average salary

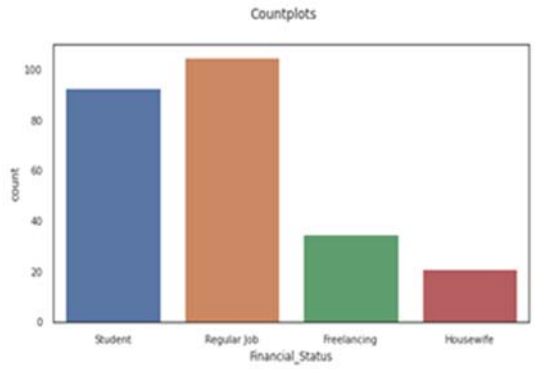

Fig. 6. count plot of financial status of users

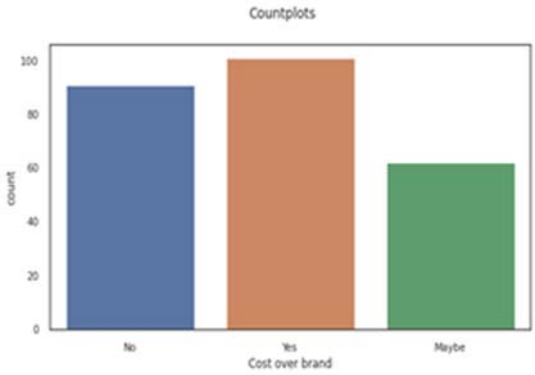

Fig. 7. Count plot of cost over brand

In figure 4 it is noticed that the Tier 2 cities are the most in our dataset. From figure 7 it shows that cost over brand is almost equally distributed so brand may or may not be a big deal. People salary on an average between 1000 -5000 rupees,

Average Money Spend in Different Product Categories

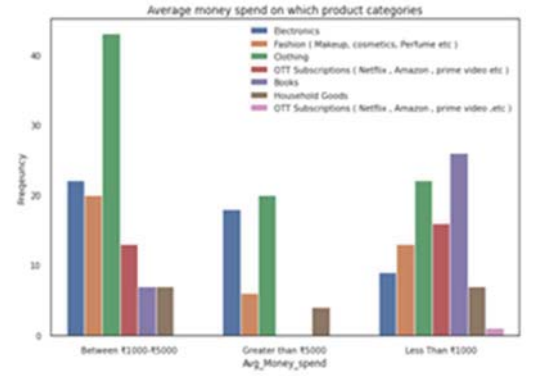

Fig. 8. Count plot for frequency counter to average money spent by product categories

Figure 8 shows the average amount of money spend on different product categories which gives ideas about the frequency of product categories purchased in which category of spend amount. Money spends between 1000-5000 mostly on clothing followed by electronics. That money spends less than 1000 most likely to spend on books.

A. Gender vs Product Categories

Figure 9 shows gender vs product categories. It visualized that men and women go for which product most. By observing figure 9 it seems like male customers ascendant for electronics category, similarly fashion presiding by female customers.

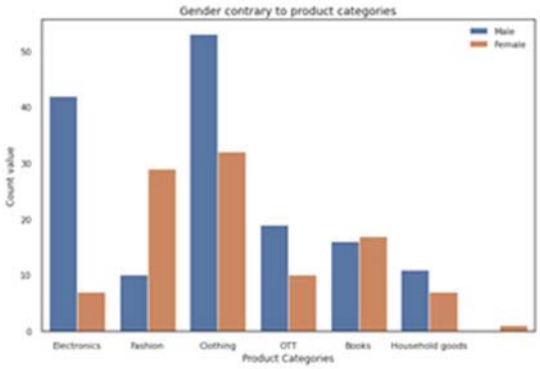

Fig. 9. Count plot for contrary to product categories by gender. Here the first bar appears for male and the second for female.

B. City Tier VS Product categories

Figure 10 visualizes the city tier vs product categories. It demonstrates how the sale of different product categories affects the tier of the city. It visualized that all tiers' maximums spend money on cloth products. Figure 10 shows that very few people are interested in buying fashion product from tier-3 city and Tier1 & Tier2 presides all categories specially clothing and electronics,

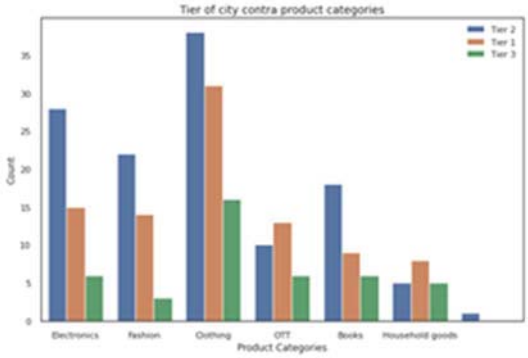

Fig. 10. Count plot for count versus product categories by Tier

Product category vs Monthly shopping frequency (on a scale of 1 to 5, shop customers in the selected product category monthly). Figure 11 show that mean shopping frequency is around 3, Household seem to be the most frequently bought item in a month, shopping frequency of electronics and books appear to be less than the others, and OTT subscriptions are the most frequently bought categories.

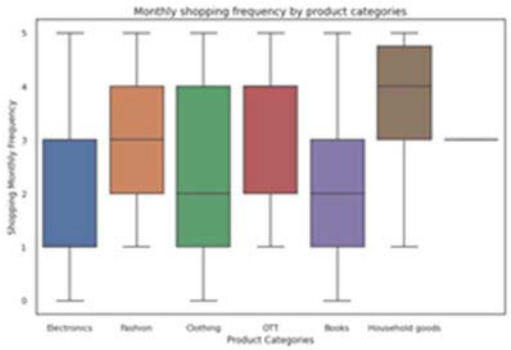

Fig. 11. Boxplot for shipping monthly frequency with product categories.

C. State Region vs Product Categories

Figure shows Product Category's sale vs state regions It provides the relation of a different region of the state on product categories.

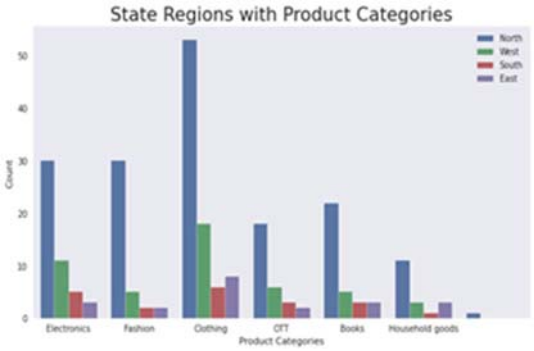

Fig. 12. Bar plot for count vs product categories

D. Predictive Analysis

Scikit learn is an open-source machine-learning library for python. It attributes various algorithms like support vector machines, random forests, and k-neighbors. It provides efficient tools for statistical modeling including classification, clustering, and prediction. The connection between independent and dependent features in ML models is familiarly acknowledged. In this section of the study, product categories are treated as dependent or target features, while other columns are considered independent features. In consequence, the first step is to train our ML model using independent features. The dataset was divided into 90 percent training data and 10 percent test data. This stage can be divided into two parts: the first involves making the structure of our proposed model, and the second involves gaining the model performance.

1) Results

After completing the training on various models, the results are presented in the next section.

F1 weighted score of Various Machine Learning Models (Classification)

The result of different machine learning algorithms is given with an F1 score (It is the measure of a model`s accuracy on a dataset and also balancing precision on the positive class). Table 2 shows that the Decision Tree and Random Forest have better results than others with a higher F1 score of 83 percent.

TABLE II. SHOWS THE F1SCORE OF DIFFERENT APPROACHES APPLIED ON DATASET

Approaches	F1 score
Navies Bayes	0.503
Logistic Regression	0.594
Support Vector Machines	0.715
K-Nearest Neighbors	0.574
Decision Trees	0.831
Random Forest	0.845s

The figure shows the result of every algorithm used on the dataset or a comparison of the results of different algorithms.

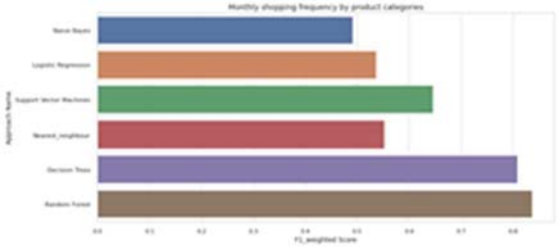

Fig. 13. Bar plot show the best approach in the form of F1 score

a) Number of Trees Vs F1 Weighted Score

The figure shows the Number of Trees vs F1 Weighted Score, figure 14 helps to analysis of relationship exist between these two attributes. The given figure states that there is somewhat linear relationship as the value of F1_Score decreases as the value of number of trees get increase. We get the value of F1_score equal to 0.844 when the value of number of trees equal to 100 and the value of F1_score equal to 0.845 when the number of trees equal to 300 which is the best result out of all others.

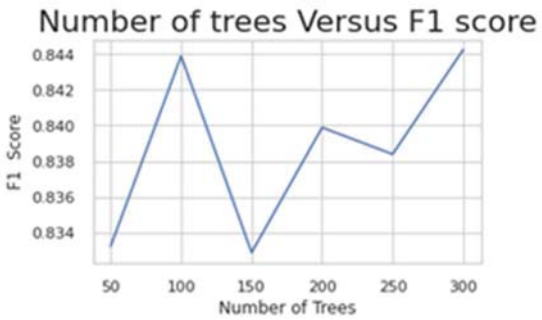

Fig. 14. Diagram show the variation of number of trees in random forest approach show how much effect on F1 score

VI. Conclusions And Future Scope

In this paper, we respectively summarized various machine learning developments in the field of e-commerce. Machine learning helps to make e-commerce uncomplicated for small vendors such that way of selling goods and services online adopted by small vendors. Predicting the future requirement of the customer provides sake for the company to plan for inventory at the warehouse. It helps the company better acknowledge customer demand and provides a discount in a less purchased category, which is beneficial for both. Analysis of customers by analyzing their behaviour data give dominant over market and idea about different factors that affect flow of goods, by applying various machine learning approach to the dataset after analysis, we conclude that decision tree and random forest dominant others. From figure 13 which shows that random Forest classification shows a better F1 score of 84.5 percent followed by a decision tree which has 83.1 percent. We perform analysis on F1 score by changing the value of number of trees on random forest tree and conclude that by increasing value of number of trees the result of F1 score increase gradually but not a linear relationship between both.

References

[1] R. Esmeli, M. Bader-El-Den, and H. Abdullahi, “An analyses of the effect of using contextual and loyalty features on early purchase prediction of shoppers in e-commerce domain,” J Bus Res, vol. 147, pp. 420–434, Aug. 2022, doi: 10.1016/j.jbusres.2022.04.012.

[2] M. Mao, S. Chen, F. Zhang, J. Han, and Q. Xiao, “Hybrid ecommerce recommendation model incorporating product taxonomy and folksonomy,” Knowl Based Syst, vol. 214, Feb. 2021, doi: 10.1016/j.knosys.2020.106720.

[3] B. T. Khoa, “Dataset for the electronic customer relationship management based on S-O-R model in electronic commerce,” Data Brief, vol. 42, Jun. 2022, doi: 10.1016/j.dib.2022.108039.

[4] D. Tolstoy, E. R. Nordman, and U. Vu, “The indirect effect of online marketing capabilities on the international performance of e-commerce SMEs,” International Business Review, vol. 31, no. 3, Jun. 2022, doi: 10.1016/j.ibusrev.2021.101946.

[5] M. B. Gulfraz, M. Sufyan, M. Mustak, J. Salminen, and D. K. Srivastava, “Understanding the impact of online customers’ shopping experience on online impulsive buying: A study on two leading E-

commerce platforms," Journal of Retailing and Consumer Services, vol. 68, Sep. 2022, doi: 10.1016/j.jretconser.2022.103000.

[6] Z. A. Mohammed and G. P. Tejay, "Examining privacy concerns and ecommerce adoption in developing countries: The impact of culture in shaping individuals' perceptions toward technology," Comput Secur, vol. 67, pp. 254–265, Jun. 2017, doi: 10.1016/j.cose.2017.03.001.

[7] R. Gupta and C. Pathak, "A machine learning framework for predicting purchase by online customers based on dynamic pricing," in Procedia Computer Science, 2014, vol. 36, no. C, pp. 599–605. doi: 10.1016/j.procs.2014.09.060.

[8] Y. Bai and H. Li, "Mapping the evolution of e-commerce research through co-word analysis: 2001–2020," Electron Commer Res Appl, vol. 55, Sep. 2022, doi: 10.1016/j.elerap.2022.101190.

[9] I. Belkhadir, E. D. Omar, and J. Boumhidi, "An intelligent recommender system using social trust path for recommendations in web-based social networks," in Procedia Computer Science, 2019, vol. 148, pp. 181–190. doi: 10.1016/j.procs.2019.01.035.

[10] A. Martínez, C. Schmuck, S. Pereverzyev, C. Pirker, and M. Haltmeier, "A machine learning framework for customer purchase prediction in the non-contractual setting," Eur J Oper Res, vol. 281, no. 3, pp. 588–596, Mar. 2020, doi: 10.1016/j.ejor.2018.04.034.

[11] S. C. Boerman, S. Kruikemeier, and F. J. Zuiderveen Borgesius, "Online Behavioral Advertising: A Literature Review and Research Agenda," J Advert, vol. 46, no. 3, pp. 363–376, Jul. 2017, doi: 10.1080/00913367.2017.1339368.

[12] C. Buizza et al., "Data Learning: Integrating Data Assimilation and Machine Learning," J Comput Sci, vol. 58, Feb. 2022, doi: 10.1016/j.jocs.2021.101525.

[13] X. Cai, J. Cebollada, and M. Cortiñas, "Impact of seller- and buyer-created content on product sales in the electronic commerce platform: The role of informativeness, readability, multimedia richness, and extreme valence," Journal of Retailing and Consumer Services, vol. 70, p. 103141, Jan. 2023, doi: 10.1016/j.jretconser.2022.103141.