Stage by stage E- Ecommerce market database analysis by using machine learning models

Narendra Ryali[1*] , Nikita Manne[2], A. Ravisankar[3], Mano Ashish Tripathi[4,] Ravindra Tripathi[5], M.Venkata Naresh [6]

[1]K. L. Business School, Koneru Lakshmaiah Education Foundation, Greenfields, Vaddeswaram, Tadepalli Mandal, Guntur District, Andhra Pradesh
[2]Emerging Technologies- Computers Science and Information Technology, CVR College of Engineering, Hyderabad, India.,
[3]Department of Management Studies, Erode Sengunthar Engineering College- Autonomous, Erode-638057
[4]Senior Research Fellow, Department of Humanities and Social Sciences, Motilal Nehru National Institute of Technology Allahabad
[5]Department of humanities and social sciences, Motilal Nehru National Institute Technology Allahabad
[6]Department of ECE, Mohan Babu University (Erstwhile Sree Vidyanikethan Engineering College), Tirupati, Andhrapradesh, India

Abstract

In the recent era, advertising strategies are far more sophisticated than those of their predecessors. In marketing, business contacts are essential for online transactions. For that, communication needs to develop a database; this database marketing is also one of the best techniques to enhance the business and analyze the market strategies. Businesses may improve consumer experiences, streamline supply chains, and generate more income by analyzing E-Commerce market datasets using machine learning models. In the ever-changing and fiercely competitive world of e-commerce, the multi-stage strategy guarantees a thorough and efficient use of machine learning. Analyzing the database can help to understand the user's or industry's current requirements. Machine Learning models are developed to support the marketing sector. This machine learning model can efficiently operate or analyze e-commerce in different stages, i.e., systematic setup, status analysis, and model development with the implementation process. Using these models, it is possible to analyze the marketing database and create new marketing strategies for distributing marketing objects, the percentage of marketing channels, and the composition of marketing approaches based on the analysis of the marketing database. It underpins marketing theory, data collection, processing, and positive and negative control samples. It is suggested that e-commerce primarily adopt the database marketing method of the model prediction. This is done by substituting the predicted sample into the model for testing. The issue of unequal marketing item distribution may be resolved by machine learning algorithms on the one hand, and prospective customer loss can be efficiently avoided on the other. Also, a proposal for an application approach that enhances the effectiveness of existing database marketing techniques and supports model prediction is made.

*Corresponding author. Email: ryalinarendra@gmail.com

1. Introduction

The study's focus is on electronic commerce. The implementation of a novel data collection and analysis strategy may have far-reaching consequences for an organization, both good and negative. The data centres of e-commerce platforms gather and retain a vast quantity of information. Data and its trends over time are not being used as a commercial opportunity by them. Customers'

registration information, search histories, purchase records, and conversation transcripts, for instance, are all being collected and kept on their server for use solely in the event of a data error. Even though they may be reluctant to share their data with other companies due to privacy concerns, they can build their analytics team, which may have a positive financial impact.

Among the top 50 industries with the most data storage, e-commerce ranks highly. With so much information, they can revolutionize the online shopping landscape. This sector has invested millions of dollars in marketing, social media, data sorting, and other initiatives to increase sales. Still, they have to be entirely aware of the competitive advantage that machine learning may provide. Data mining, AI, AR, and prediction are just a few of the subfields that fall under the umbrella of machine learning. In this study, machine learning is used for forecasting purposes.

When used in e-commerce, machine learning algorithms' predictive power will help to spot previously unseen trends, outliers, POIs, etc. This will enable online businesses to recognise critical information in any context accurately. To manage and grow their business, they may analyse information like total sales, product categories bought, payment methods used, interest rates, shipping times, and client locations.

1.1 Introduction to Database Marketing Technique:

An organization can maximize earnings by using database marketing, which stores relevant data in a database. By doing so, they can provide better and more complete goods and services to their clients. Marketing needs technical support to organize the database and marketing tools used to develop the corporate business. Recent IT tools are used to analyze marketing statistics in the micro-level view [13]. Database marketing should first analyse consumers of different types, such as previous customers' buying data, to determine their consumption direction and preference to build long-term relationships. Database marketing may aid firms in four ways:
1. Customer service
2. Improving brand marketability
Revenue growth
4. Simplifying sales

Because of its excellence and appeal, database marketing has been a smashing success in theory and practice.

1.2 Machine learning models:

Machine learning is used to summarize, identify, and forecast unknown outcomes or unobservable data, algorithms analyze vast amounts of sample data to learn specific rules or characteristics. Develop the concept in recent works on statistical learning. On the other hand,

traditional statistical learning theory primarily offers crucial grounding for learning from a small number of samples. Its impact could be more optimal when the number of data to analyze is enormous. This study chooses the most well-known machine learning algorithms for tackling classification issues, mainly consisting of logistic regression (LR), random forest (RF), support vector machine (SVM), and Gradient-boosted decision trees (GBDT). The logistic regression model is popular in classification learning because of its ease of use, rapidity, and flexibility in adapting to new data. In machine learning, the random forest model stabilizes the model and mitigates overfitting using an average decision tree. Due to their superior generalization capabilities, support vector machines have quickly become one of the most popular and successful classifiers. The GBDT model is well-known for the accuracy of its predictions and the versatility with which it can handle different kinds of data. As there is no universally best algorithm, testing various algorithms before settling on one is essential.

Literature review:

Accurately estimating future sales is a critical factor in guiding strategic business decisions for online retailers. Companies can better manage their staff and enhance their supply chain management system if they can anticipate sales for the e-commerce platform and better grasp their financial situation. In light of [1] and [2], an e-commerce platform may make more informed decisions about inventory levels, pricing, and promotional timing with a sales forecast. According to [3], it is possible to learn about the e-commerce platform's development, stability, and decline via sales forecasting and the impact of short-term product objectives like promotion, price, season, and online ranking on long-term sales.

The authors of [1] found that the convolution neural network (CNN) method effectively predicted online store sales. This study was conducted to address the method's recognized shortcoming of requiring laborious, time-consuming, and highly specialised case-by-case manual feature engineering for certain circumstances. The purpose of this study, however, was to determine whether or not this method could automatically extract the relevant elements and then give sales predictions based on those qualities. The CNN algorithm was used as a primary tool for making accurate sales forecasts. The study has picked the ARIMA, DNN, TL, and WD algorithms to compare their performance and determine which produces the most reliable sales forecasts. Experiments have shown that the researcher may significantly increase the predicting accuracy using sample weight decay and the transfer learning approach. While the ARIMA model has the highest average value according to the MST boxplot, the CNN algorithm has succeeded in automatically extracting the practical features and using them to predict sales.

Based on their research efforts, both [2] and [3] have settled on a neural network method. Researchers in

2018 used a neural network algorithm called a nonlinear autoregressive network (NARNN), while researchers in 2019 used a neural network algorithm called a recurrent neural network (RNN) and an extended short-term memory network (LSTM). Researchers have utilised this method of approaching the algorithm to anticipate e-commerce sales and demand. Researchers have raised similar concerns about distinguishing between cross-product demand/sales patterns and the available correlations. Both papers set out to provide some systematic pre-processing or forecasting framework to help with problems that arise in e-commerce. ARIMA (time series analysis) was utilised to evaluate and contrast these studies. According to the discussion of the 2018 research's findings, NARNN had a lower error rate than the ARIMA, with a prediction error of 0.1016 compared to 0.1389 for ARIMA. The mean and median for 2019 research are likewise lower for LSTM than for ARIMA.

Methodology- Machine learning algorithm:

1. Data collection process: This stage needs to gather the data related to the consumer issues/ needs, product details, and competitive company.

A. Primary Information: This needs to collect primary customer-related data. Customer-related data includes registration information, customer behaviour like product priority, amount of interactions/ frequency of purchases and customer relationship with storekeepers/ sales executives. Primary information is a source to develop the e-marketing sector.

B. Product-related information: it includes distribution/ sales, various types of advertisement, stock availability, and relevant store information.

C. Feedback system: Gather the customer feedback on the product, product/ customer care service-related improvement, and quality-related data need to know by direct/online/ third party feed. Competitor company-created feedback also needs to be analysed. Convert total feedback on a statistical basis to enrich the businesses.

It typically includes details such as who the competitors are, the size of their business, the types of products they offer, the number of people who use the brand, how many people joined the membership, and how consumers respond. It is essential to keep data organized and up-to-date throughout data gathering and guarantee its legitimacy and accessibility.

2. Data Storing process: Must store the data obtained for various reasons and through multiple channels securely and efficiently; use the databases for this purpose. First, the database service business safely keeps all the obtained customer data in a dedicated database. Then, the data fusion channel brings in data about products and competitors, which will save the database. May use the Databases to sift through mountains of data, identify trends, and make educated guesses about consumer behaviour to generate additional data for decision-making.

3. Data processing: Data that was initially disorganized, nonstandard, and had varying properties using data processing technology; so that businesses may better respond to shifting customer demand, adapt quickly to new market possibilities and fulfil the needs of all of their departments. Because different companies keep the original data in various ways outside marketing, including R&D, product enhancement, and customer relationship management, database analysis is becoming more critical. Hence, following data processing, it's important to have centralized control while still maintaining data autonomy, data uniformity, and applicability.

4. New customer-finding process: Analyze the customer likes, income, and product preferences. This analysis needs to conclude the customer characteristics or relevant attributes, and purchase motivation, and list the targeted customers via different marketing technologies.

5. Database utilization: Depending on the company's nature, it can use the data in various ways, including assessing customers' spending habits. Based on their purchase histories, tailoring marketing campaigns to appeal to customers based on their demographics and maximizing customer loyalty and repeat business by optimizing the store's membership system and selecting the appropriate demographic segments. This study focuses on using feature vectors and models constructed from e-commerce data to improve the ability to forecast the characteristics of the targeted consumers.

6. Database Improvement process: The database develops and improves in three ways: first, as the number of users increases in everyday business operations; second, collect more data by collecting feedback, giving more offers, and collecting the primary data at the time of billing etc., and third, by using the different advertising platform and database expansion by using many marketing techniques. In summary, everyday transactions, cross-platform data fusion, and data from several domains increase data volume, diversity, and database quality. A complete database targets ideal consumer.

This research primarily develops a model using a machine learning algorithm to foresee buyers' actions on Taobao during a specific time frame. Analyze Brand Data Bank data from the last year for this purpose; this data primarily consists of customer profiles, purchase histories, web surfing records, and product information. Figure 1 depicts many different conceptual models used in database marketing.
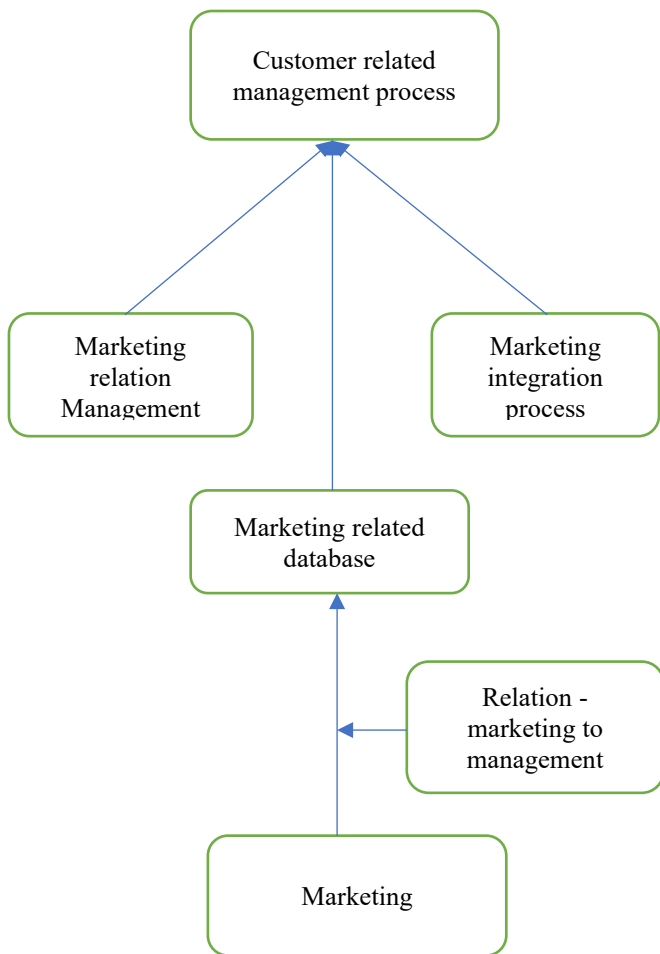
Figure 1: Marketing Database model.

Result with discussion:

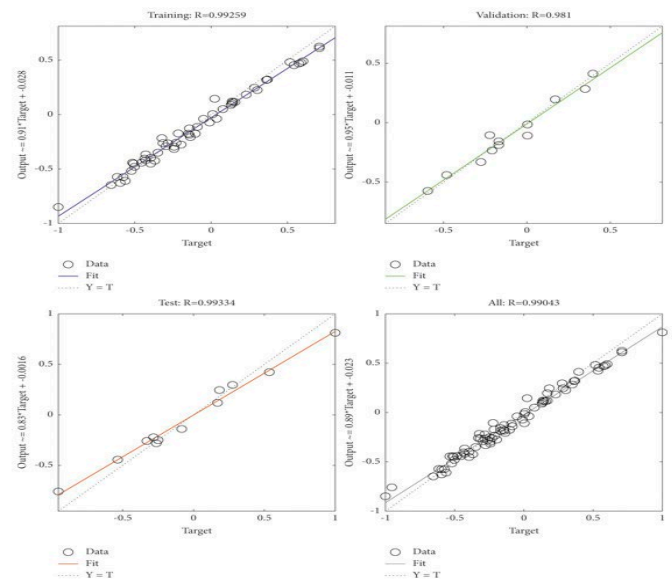**Figure 2:** Sample prediction result

Customer information is known as "consumer picture data," It often falls into two groups: demographic details and purchase behaviour data. Fifteen categories may be used to categorise people, including age, gender, income, monthly expenditure, region, and phone kind. Label the other groups by industry, and it mainly consists of the yearly consumption quantity, consumption frequency, and consumer preference of the industry. A purchaser's personality traits may be thoroughly investigated by gathering customer picture data.

To be clear, the "historical transaction data," is just the records of customers' buying behaviours on the Tmall platform over a given period. Amount, quantity, route, and time of each transaction are all included, allowing for a more in-depth examination of customer spending habits and the likeliness of repeat purchases.

The consumer browsing log records all the requests made by the client to the server when using the Tmall platform to browse, shop, and do other business. Visitor IP addresses, page URLs, visit durations, and frequency may all be gleaned from these logs.

The term "product data" refers to information on a product's features, benefits, costs, and other relevant aspects. May use Product data to infer consumers' preferences and evaluate consumers' buying power.

Figure 2 displays the predicted images of the test samples GBDT model which got the best effect in all the ways by comparing the different machine learning models, and the GBDT model provides the highest performance. More than half of the population has a prediction accuracy score in the 80-99 range (consisting of purchasers; positive samples), whereas more than three-quarters of the population has a score in the 0-9 range (consisting of non-purchasers; negative examples). The model effect comparison further confirms that the GBDT3 model chosen is superior. More than 64% of anticipated sample non-purchasing users are in the prediction score range of 0-9, similar to the percentage of negative examples. The model's generalization ability and the predicted sample's predicting power may be evaluated based on the similarity of the curves between the negative and the indicated samples. Hence, in the real business world, the input prediction population falls into the application through the platform and the prediction sample's threshold value falls between 84 and 99.

At all phases, but notably during the epidemic, the crowd selection's return on investment (ROI) by model prediction marketing strategy is better than 1. Next, (e performance during the warm-up phase, but the ROI may reach 3.5 under heavy consumption, and the impact is also excellent. However, while the presale period's return on investment is somewhat lower than that of the other two phases, it delivers a greater rate of collection and purchase, which plays an enormous role in the subsequent transaction transformation. Figure 3 displays the predicted impact of population growth or decline throughout a range of epochs, as the model indicates.
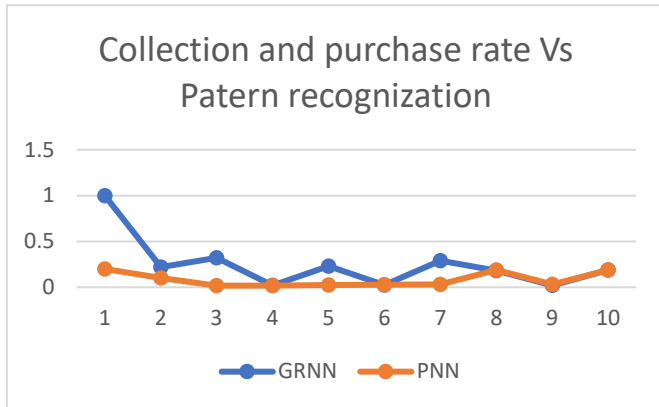
Figure 3: the predicted impact of population

Secondly, examine the impact of the model prediction by comparing different advertising approaches. Nevertheless, the lowest return on investment (ROI) can be seen in the third category, suggesting that although it may achieve a specific volume of traffic via this channel, it comes at a steep price. In contrast, the other group's most excellent CTR was 9.62%. However, they only saw an ROI of 1.78. (They may further refine their positioning by adding tags. Because of the abundance of highly correlating product categories available at the Lister flagship location, cross-category marketing consumption is extreme. In this scenario, the ROI can be maintained at 2.37, suggesting that it may still employ cross-category marketing but with a cumbersome implementation method. The rate of collecting and purchasing of similarly situated individuals is most excellent, while the turnover rate is typical. The associate group's conversion rate and return on investment (ROI) are the highest of any marketing strategy. This is because associate marketing targets existing consumers with a foundational grasp of the brand and its offerings. The model estimated ROI at 4.62, second only to the whole population. Initially, start the group attack to list out the existing customers; after that, compare the product sales concerning the population. In the end, understand the expected sales of the product. The ROI is used to determine the different people. Figure 4 states that various advertisements are available in the market.
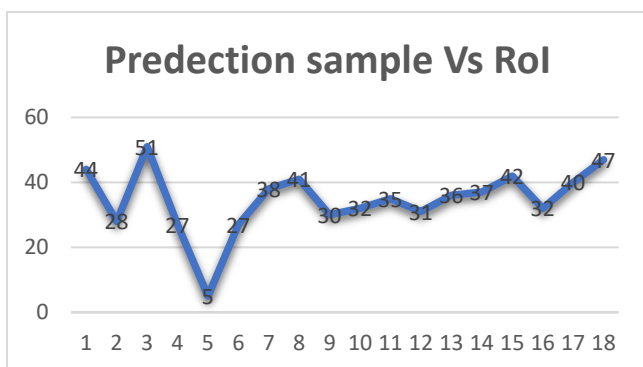
Figure 4: Various marketing methods with comparison

Conclusion:

This article summarizes the essential findings and concludes with its research into using machine learning models in the database marketing of an online retailer.

Before selecting a machine learning algorithm as a foundation for the models, one should get as much knowledge as possible regarding database marketing methods. Understanding database marketing may be summed up on three levels: connotation, advantages, and foundational procedures. Most introductory material on machine learning focuses on the LR, RF, SVM, and GBDT models.

Second, analyse the current status of an online store. This article begins with a high-level summary of the importance and standing of an online brand shop in its industry before detailing the store's present situation from the perspective of the distribution of marketing objects, marketing channels, and marketing approaches across three tiers. The article then provides a summary of the issues that have occurred and suggests potential remedies via the use of a database marketing approach.

To tackle the sales forecasting issue of e-commerce items with limited sample data, the authors of this study used text mining. They integrated learning approaches to create a multi-dimensional index system impacting e-commerce product sales. In doing this, a transfer learning model is built to estimate sales of new products when there are few data points to work with or when index data is unavailable. E-commerce firms may use the study's results as a yardstick for demand-side choices and a basis for forecasting product sales.

References:

[1] Nie Chen, "Research on E-Commerce Database Marketing Based on Machine Learning Algorithm", Computational Intelligence and Neuroscience, vol. 2022, Article ID 7973446, 13 pages, 2022. https://doi.org/10.1155/2022/7973446

[2] Liu C-J, Huang T-S, Ho P-T, Huang J-C, Hsieh C-T (2020) Machine learning-based e-commerce platform repurchase customer prediction model. PLoS ONE 15(12): e0243105. https://doi.org/10.1371/journal.pone.0243105

[3] Sarvjeet Kaur Chatrath, G.S. Batra, Yogesh Chaba, Handling consumer vulnerability in e-commerce product images using machine learning, Heliyon, Volume 8, Issue 9, 2022, e10743, ISSN 2405-8440, https://doi.org/10.1016/j.heliyon.2022.e10743. (https://www.sciencedirect.com/science/article/pii/S24058 4402202031X)

[4] Shanthan Kandula, Srikumar Krishnamoorthy, Debjit Roy, A prescriptive analytics framework for efficient E-commerce order delivery, Decision Support Systems, Volume 147, 2021, 113584, ISSN 0167-9236, https://doi.org/10.1016/j.dss.2021.113584. (https://www.sciencedirect.com/science/article/pii/S01679 23621000944)

[5] "REAL TIME SENTIMENT ANALYSIS OF E-COMMERCE WEBSITES USING MACHINE LEARNING ALGORITHMS." (2018).

[6] Ritu Punhani et al 2021 J. Phys.: Conf. Ser. 1714 012026 DOI 10.1088/1742-6596/1714/1/012026

[7] Peng Gao, Liang Zhao, "Study on Deep Learning Technology to Construct E-Commerce Industry Marketing Promotion Model", Security and Communication Networks, vol. 2022, Article ID 9958398, 11 pages, 2022. https://doi.org/10.1155/2022/9958398

[8] Li Yan and Mohammad Ayoub Khan. 2022. Predictive Analysis of User Behavior Processes in Cross-Border E-Commerce Enterprises Based on Deep Learning Models. Sec. and Commun. Netw. 2022 (2022). https://doi.org/10.1155/2022/1560017

[9] Ren, X., He, J. & Huang, Z. RETRACTED ARTICLE: An empirical study on the behaviour of e-commerce strategic planning based on a deep learning algorithm. Inf Syst E-Bus Manage (2021). https://doi.org/10.1007/s10257-021-00504-9

[10] Kaur, G., Sharma, A. A deep learning-based model using a hybrid feature extraction approach for consumer sentiment analysis. J Big Data 10, 5 (2023). https://doi.org/10.1186/s40537-022-00680-6

[11] AlGhamdi, Rayed & Alfarraj, Osama & Bahaddad, Adel. (2014). How do Retailers at Different Stages of E-Commerce Maturity Evaluate Their Entry into E-Commerce Activities? Journal of Computer Science and Information Technology. 2. 37-71.

[12] Y. -S. Fang and L. -C. Fang, "A Review of Chinese E-Commerce Research: 2001–2020," in IEEE Access, vol. 10, pp. 49015-49027, 2022, doi: 10.1109/ACCESS.2022.3172433.

[13] Wenlong Zhu, Jian Mou, Morad Benyoucef, Exploring purchase intention in cross-border E-commerce: A three-stage model, Journal of Retailing and Consumer Services, Volume 51, 2019, Pages 320-330, ISSN 0969-6989, https://doi.org/10.1016/j.jretconser.2019.07.004. (https://www.sciencedirect.com/science/article/pii/S09696 98919306083)

[14] Gaikar Vilas Bhau, Radhika Gautamkumar Deshmukh, T. Rajasanthosh Kumar, Subhadip Chowdhury, Y. Sesharao, Yermek Abilmazhinov, IoT based solar energy monitoring system, Materials Today: Proceedings, 2021, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2021.07.364. (https://www.sciencedirect.com/science/article/pii/S22147 85321052238)

[15] Singh, Rahul Kumar, Pardeep Singh, and Gaurav Bathla. "User-Review Oriented Social Recommender System for Event Planning." Ingénierie des Systèmes d Inf. 25.5 (2020): 669-675.

[16] Yudiana, Wayan Agus, Maya Ariyanti, and Andry Alamsyah. "Wisdom of the Crowd" as Personalized Music Recommendation Model for Langit Musik Service." 2019 International Conference on Information Management and Technology (ICIMTech). Vol. 1. IEEE, 2019.

[17] Zeeshan, Syed, Olumide Euba, and Andrey Abadzhiev. "Internet and Intermediaries in the Tourism Distribution Channel-Study of Swedish, Bulgarian and Online Travel Agencies." rapport nr.: Masters Thesis 2005 (2006).

[18] Ferdousi, Zannatul. Design and development of a real-time gesture recognition system. Tennessee State University, 2008.

[19] Pondel, Maciej, et al. "Deep learning for customer churn prediction in e-commerce decision support." Business Information Systems. 2021.

[20] Schafer, J. Ben, Joseph A. Konstan, and John Riedl. "E-commerce recommendation applications." Data mining and knowledge discovery 5 (2001): 115-153.

[21] Moazzam, Ali, et al. "Customer Opinion Mining by Comments Classification using Machine Learning." International Journal of Advanced Computer Science and Applications 12.5 (2021).

[22] Mykhalchuk, Taras, et al. "Development of recommendation system in e-commerce using emotional analysis and machine learning methods." 2021 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS). Vol. 1. IEEE, 2021.

[23] Zheng, Xiaolin, Shuai Zhu, and Zhangxi Lin. "Capturing the essence of word-of-mouth for social commerce: Assessing the quality of online e-commerce reviews by a semi-supervised approach." Decision Support Systems 56 (2013): 211-222.

[24] Saleem, Hussain, et al. "Data science and machine learning approach to improve E-commerce sales performance on social web." International Journal of Computer Science and Network Security (IJCSNS) 19 (2019).

[25] Gubela, Robin, et al. "Conversion uplift in e-commerce: A systematic benchmark of modelling strategies." International Journal of Information Technology & Decision Making 18.03 (2019): 747-791.

[26] Gupta, Rajan, and Chaitanya Pathak. "A machine learning framework for predicting purchase by online customers based on dynamic pricing." Procedia Computer Science 36 (2014): 599-605.

[27] Monil, Patel, et al. "Customer Segmentation Using Machine Learning." International Journal for Research in Applied Science and Engineering Technology (IJRASET) 8.6 (2020): 2104-2108.

[28] Shen, Boyu. "E-commerce Customer Segmentation via Unsupervised Machine Learning." In The 2nd International Conference on Computing and Data Science, pp. 1-7. 2021.

[29] Singla, Zeenia, Sukhchandan Randhawa, and Sushma Jain. "Sentiment analysis of customer product reviews using machine learning." 2017 international conference on intelligent computing and control (I2C2). IEEE, 2017.

[30] Zhao, B., Takasu, A., Yahyapour, R., & Fu, X. (2019, November). Loyal consumers or one-time deal hunters: Repeat buyer prediction for e-commerce. In 2019 International Conference on Data Mining Workshops (ICDMW) (pp. 1080-1087). IEEE.

[31] Fu, Min, et al. "DAliM: Machine learning based intelligent lucky money determination for large-scale e-commerce businesses." Service-Oriented Computing: 16th International Conference, ICSOC 2018, Hangzhou, China, November 12-15, 2018, Proceedings 16. Springer International Publishing, 2018.

[32] Liu, Liping. "e-commerce personalized recommendation based on machine learning technology." Mobile Information Systems 2022 (2022).

[33] Nosratabadi, S., Mosavi, A., Duan, P., Ghamisi, P., Filip, F., Band, S. S., ... & Gandomi, A. H. (2020). Data science in economics: a comprehensive review of advanced machine learning and deep learning methods. Mathematics, 8(10), 1799.

[34] Zhang, Mingyang, et al. "A Feature Engineering and Ensemble Learning Based Approach for Repeated Buyers Prediction." INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL 17.6 (2022).

[35] Viamonte, Maria João, et al. "Learning User Preferences Models and Business Strategies for E-Commerce." Proceedings of the EChallenges E-2004 Conference on E-business and E-work. 2004.

[36] Piskunova, Olena, and Rostyslav Klochko. "Classification of e-commerce customers based on Data Science techniques." CEUR Workshop Proc. Vol. 2649. 2020.

[37] Mazumdar, Bireshwar Dass, and Shubhagata Roy. "Multi-Agent Paradigm for B2C E-Commerce." Artificial Intelligence and Machine Learning in Business Management. CRC Press, 2021. 29-52.

[38] Khrais, Laith T. "Role of artificial intelligence in shaping consumer demand in E-commerce." Future Internet 12.12 (2020): 226.