PREDICTING CUSTOMER PREFERENCES AND FORECASTING DEMAND IN E-COMMERCE LEADING TO BETTER SUPPLY CHAIN MANAGEMENT USING MACHINE LEARNING TECHQNIQUES

Dr. M. Hemalatha [1], Dr. V. Nirmala Devi [2], Dr. N. Selvakumar [3], Dr. J. Jeyaboopathi Raja [4], Dr. Princess Mariajohn [5], Dr. C. Suresh [6] and Dr. S. B. Mahalakshmi [7]

[1] Associate Professor, PG & Research Department of Computer Science, Sri Ramakrishna College of Arts &Science, Coimbatore. Email: mhemalatha@srcas.ac.in
[2] Associate Professor & Head, PG & Research Department of Commerce, Sri Ramakrishna College of Arts &Science, Coimbatore. Email: nirmaladevi@srcas.ac.in
[3] Associate Professor, PG & Research Department of Commerce, Sri Ramakrishna College of Arts &Science, Coimbatore. Email: selvakumar@srcas.ac.in
[4] Assistant Professor, PG & Research Department of Computer Science, Sri Ramakrishna College of Arts &Science, Coimbatore. Email: jeyaboopathiraja@srcas.ac.in
[5] Assistant Professor, Department of Computer Applications, Hindusthan College of Engineering and Technology, Coimbatore. Email: princess.mariajohn@gmail.com
[6] Assistant Professor, PG & Research Department of Commerce, Sri Ramakrishna College of Arts &Science, Coimbatore. Email: suresh.c@srcas.ac.in
[7]Assistant Professor, Department of Computing AI & M, Coimbatore Institute of Technology, Coimbatore. Email: mahalakshmi@cit.edu.in

DOI: 10.5281/zenodo.11484947

Abstract

The advent of e-commerce and online shopping has led to a need for predictive analytics to understand and forecast customer behavior. This study aims to analyze e-commerce customer behavior and predict demand for better supply chain management. The study proposes using machine learning algorithms such as the K-Nearest Neighbor algorithm and Naïve Bayes Classifier algorithms to anticipate customer needs. The dataset used for the study includes a set of variables from every sales done during the period such as main category name, subcategory name, brand name, product code, product name, quantity purchased, color name, festival flag, festival name, size name, gender, locality purchased, age group, month, season, price, price group, and rating. The study analyzes the dataset using mean ranking analysis to evaluate the main categories and subcategories. The proposed system will use KNN algorithms to evaluate age group and gender variables to identify the most preferred products and brands by customers, leading to increased sales and customer satisfaction. The Naïve Bayes Classifier algorithm will be used to group variables to predict and forecast data based on historical results, leading to informed decisions about product inventory, marketing strategies, and promotions. The study aims to provide insights into the importance of forecasting and how it serves as a necessary aid to planning. The use of machine learning algorithms makes this study innovative and important in the context of e-commerce. The proposed system will be developed using advanced machine learning techniques to analyze vast datasets and extract useful insights efficiently.

Keywords: E-Commerce, Customer Behavior, Demand Forecasting, Predictive Analytics, Data Analysis.

1. INTRODUCTION

In today's world, e-commerce has become an essential part of our lives. It has revolutionized the way we shop and has opened up numerous opportunities for businesses to reach a wider audience. However, with the increase in competition, it has become crucial for e-commerce businesses to understand their customers and predict their behaviour to stay ahead of the game. This is where the power of prediction comes in. By analysing customer behaviour and forecasting demand, e-commerce businesses can make informed decisions about product development, pricing, and marketing strategies.

One of the primary benefits of predicting customer behaviour is the ability to offer personalized recommendations. By analysing customer purchase history and browsing behaviour, businesses can suggest products that are more likely to appeal to individual customers. This has been shown to increase customer satisfaction and loyalty, leading to higher conversion rates and revenue for businesses. In a study by McKinsey, it was found that personalized recommendations can increase sales by up to 30%. [1] In addition to personalized recommendations, predicting customer behaviour can also help businesses identify which products are likely to be popular in the future. This can inform product development and marketing strategies, ensuring that businesses are offering products that are in high demand. By using predictive analytics to analyse historical sales data and market trends, businesses can identify emerging trends and respond accordingly. This can help businesses stay ahead of the competition and remain relevant in a constantly evolving market [2].

Demand forecasting is another important aspect of prediction in e-commerce. By analysing historical sales data and market trends, businesses can forecast future demand for products. This can help businesses plan inventory and production levels, ensuring that they have the right amount of stock to meet customer demand [3]. Demand forecasting can also help businesses optimize pricing strategies by identifying when prices should be adjusted based on predicted demand. One of the challenges of predicting customer behaviour and forecasting demand is the vast amount of data that needs to be analysed. However, advances in machine learning and artificial intelligence have made it easier for businesses to process and analyse large data sets [4]. By using machine learning algorithms, businesses can identify patterns and insights in data that would be difficult to identify using traditional methods.[5]

Another challenge of prediction in e-commerce is the need to ensure data privacy and security. Businesses must ensure that customer data is protected and that they are complying with relevant regulations such as the General Data Protection Regulation (GDPR).[6] Failure to do so can result in severe consequences such as fines and reputational damage [7].In conclusion, the power of prediction is crucial for e-commerce businesses to stay ahead of the competition. By analysing customer behaviour and forecasting demand, businesses can make informed decisions about product development, pricing, and marketing strategies. Personalized recommendations, identifying emerging trends, and optimizing inventory and pricing strategies are just some of the benefits of prediction in e-commerce. However, businesses must also ensure data privacy and security to avoid potential consequences. With advances in machine learning and artificial intelligence, the potential for prediction in e-commerce is vast, and businesses that embrace it are likely to be more successful in the long run.[8]

2. LITERATURE REVIEW

The article, "Rapidly forecasting demand and adapting commercial plans in a pandemic" by McKinsey & Company, highlights the importance of demand forecasting and adaptation in response to a crisis such as the COVID-19 pandemic. [9] The article emphasizes the need for businesses to have agile and responsive strategies that can

ISSN 1462 2815

quickly adapt to changing market conditions. The authors suggest that businesses can use predictive analytics and scenario planning to forecast demand, adjust pricing strategies, and optimize their commercial plans. The article also emphasizes the importance of cross-functional collaboration, as businesses must work together to coordinate their response to the crisis.

In his chapter "On Forecasting the Demand for E-Commerce," Ward provides a detailed analysis of the unique challenges involved in forecasting demand for e-commerce businesses.[10] The author emphasizes the importance of understanding the factors that influence consumer behaviour and how they interact with technology. The chapter covers different methods for forecasting e-commerce demand, including regression analysis and time series modelling. The author also discusses the importance of data quality and the need for continuous monitoring and updating of forecasting models.

In her article, "Role of Artificial Intelligence in Shaping Consumer Demand in E-Commerce," Khrais highlights the increasing role of AI in analysing and predicting consumer behaviour in the e-commerce industry. The author discusses how AI can be used to analyse consumer data and generate personalized recommendations, as well as its potential to forecast demand and optimize pricing strategies. [11]. Zhang and Zhao's paper "Study on Consumer Behaviour Predict in E-commerce Based on Multi-Agent" proposes a novel approach to predicting consumer behaviour in e-commerce using multi-agent systems. The authors discuss the potential benefits of using multi-agent systems to model complex consumer behaviour and improve demand forecasting accuracy. The paper provides valuable insights into the potential of multi-agent systems in e-commerce and its implications for business performance.[12]

In their paper "E-Commerce Customers Behaviour Research Using Cohort Analysis: A Case Study of COVID-19," Fedushko and Ustyianovych collect data from an e-commerce website with a total of 149,045 customers. They use a convenience sampling technique to select the sample for the study. The authors use Google Analytics and R programming language to analyse customer data and conduct cohort analysis.[13] In their study, "An improved deep forest model for prediction of e-commerce consumers' repurchase behaviour," Zhang and Wang collect data from an e-commerce platform with a sample size of 40,600 customers. They use a random sampling technique to select the sample for the study. The authors use an improved deep forest model to predict e-commerce customers' repurchase behaviour. The results of the study show that the deep forest model outperforms other machine learning models in predicting repurchase behaviour. The study provides valuable insights into the use of advanced machine learning techniques for customer behaviour prediction in e-commerce.[14]

In "Analysis Model of Consumer Sentiment Tendency of Commodities in E-Commerce," Yao proposes an analysis model based on online customer reviews to predict consumer sentiment tendencies towards e-commerce products. The study uses a sample size of 100,000 product reviews collected from an e-commerce platform. The author uses a combination of text mining and sentiment analysis tools to analyse the data. The results of the study demonstrate the effectiveness of the proposed model in predicting consumer sentiment tendencies in e-commerce. The study highlights the importance of sentiment analysis in understanding consumer behaviour and making informed business decisions in the e-commerce industry.[15]

3. METHODOLOGY

3.1 Dataset Description

The HistoricalDataset is an online sales dataset containing sales data collected between 2019 and 2022. The dataset has a primary key column named "Sales_no" and includes features such as sales date, MainCategory name, SubCategory name, Brand name, Product name, Qty Purchased, Color Name, Festival name, Size name, Gender, Locality Purchased, Age Group, Month, Season, Price, Price Group, Rating, age, and Festival. The dataset also includes demographic features such as gender, age group, and locality purchased, and can be used for various data analysis and modelling tasks related to sales and marketing.

3.2 Pseudo-code of k-NN algorithm for the proposed solution:

The pseudo-code for the k-NN algorithm for a classification problem is as follows:

- Step 1: Instantiate the database connection and initialize it with the dbconnectDataContext class.
- Step 2: Check if the page is being loaded for the first time (not a postback) in the Page_Load event.
- Step 3: Create two methods, maincategorycode and subcategorycode, to perform the k-NN algorithm for the main and subcategories, respectively.
- Step 4: In both methods, calculate the value of k by counting the number of distinct MainCategory/SubCategory names in the HistoricalDatasets table.
- Step 5: Delete all existing data from the knn_filter table.
- Step 6: Retrieve the data from the HistoricalDatasets table that falls between the given date range.
- Step 7: Loop through the retrieved data and calculate the Euclidean distance between the input variables and the variables of the data point.
- Step 8: Store the Sales_no, distance, variable_heading (which is the criteria selected by the user), and variable_name (which is either the MainCategory_name or SubCategory_name of the data point) in a new knn_filter object.
- Step 9: Insert the new knn_filter object into the knn_filters table.
- Step 10: Delete all existing data from the variableTable.
- Step 11: Retrieve the top K rows (as calculated in step 4) from the knn_filters table, sorted by distance in ascending order.
- Step 12: Loop through the retrieved rows and update the corresponding variable_count in the variableTable.
- Step 13: If a variable does not exist in the variableTable, create a new variableTable object and insert it into the table.
- Step 14: Bind the retrieved data to the GridView1 control and display the top K closest data points.

ISSN 1462 2815

- Step 15: Retrieve the variableTable data, sorted by variable_count in descending order.
- Step 16: Bind the retrieved data to the GridView2 control and display the most frequent MainCategory/SubCategory names.
- Note: Steps 7 and 8 may vary depending on the number and type of input variables, as well as the distance metric used.

3.3 Pseudo-code of Naïve Bayes Classifier algorithm for the proposed solution:

- // Step 1: Loop through distinct values in Input_Variable column of HistoricalDatasets table
- for each value in distinct values of Input_Variable column in HistoricalDatasets table {
- // Step 1a: Create new Naive_baye object
- Naive_baye obj = new Naive_baye();
- // Step 1b: Set InputVariable_XVariable property
- obj.InputVariable_XVariable = value;
- // Step 1c: Calculate proportion of records in HistoricalDatasets table with current value of Input_Variable
- double proportion = calculateProportion(value);
- obj.Input_Variable_AB = proportion;
- // Step 1d: Calculate conditional probabilities for each input variable
- obj.AgeGroup_Prob = calculateConditionalProb("Age Group", value);
- obj.Gender_Prob = calculateConditionalProb("Gender", value);
- obj.LocalityPurchased_Prob = calculateConditionalProb("Locality Purchased", value);
- obj.Season_Prob = calculateConditionalProb("Season", value);
- // Step 1e: Calculate Naive Bayes probability for current value of Input_Variable
- double naiveBayesProb = calculateNaiveBayesProb(proportion, obj.AgeGroup_Prob, obj.Gender_Prob, obj.LocalityPurchased_Prob, obj.Season_Prob);
- obj.Naive_bayes = naiveBayesProb;
- // Step 1f: Insert object into Naive_bayes table
- insertIntoNaiveBayesTable(obj);
- // Step 1g: Retrieve all records from Naive_bayes table and bind to GridView control
- GridView.bind(getAllNaiveBayesRecords());
- }

- // Step 2: Construct message string
- string message = "Input values: " + ageGroup + ", " + gender + ", " + localityPurchased + ", " + season + ".\nPredicted result: ";
- // Step 3: Retrieve record with highest Naive Bayes probability and append to message string
- Naive_baye highestProbObj = getRecordWithHighestNaiveBayesProb();
- message += highestProbObj.InputVariable_XVariable;
- // Step 4: Display message string
- displayMessage(message);

4. RESULTS AND DISCUSSION

4.1 k-Nearest Neighbors (k-NN):

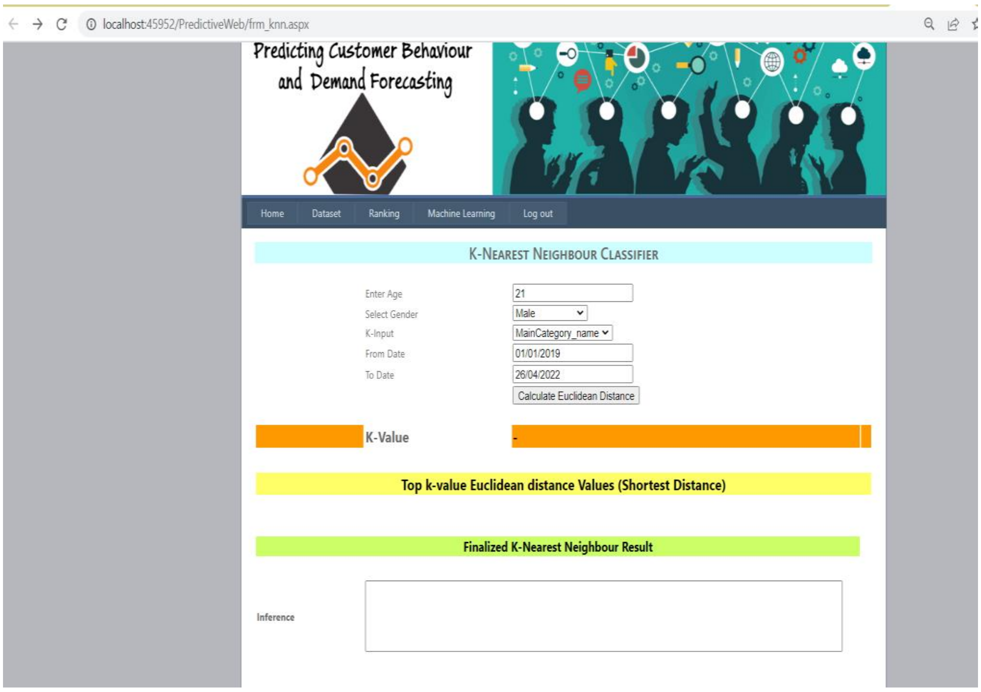


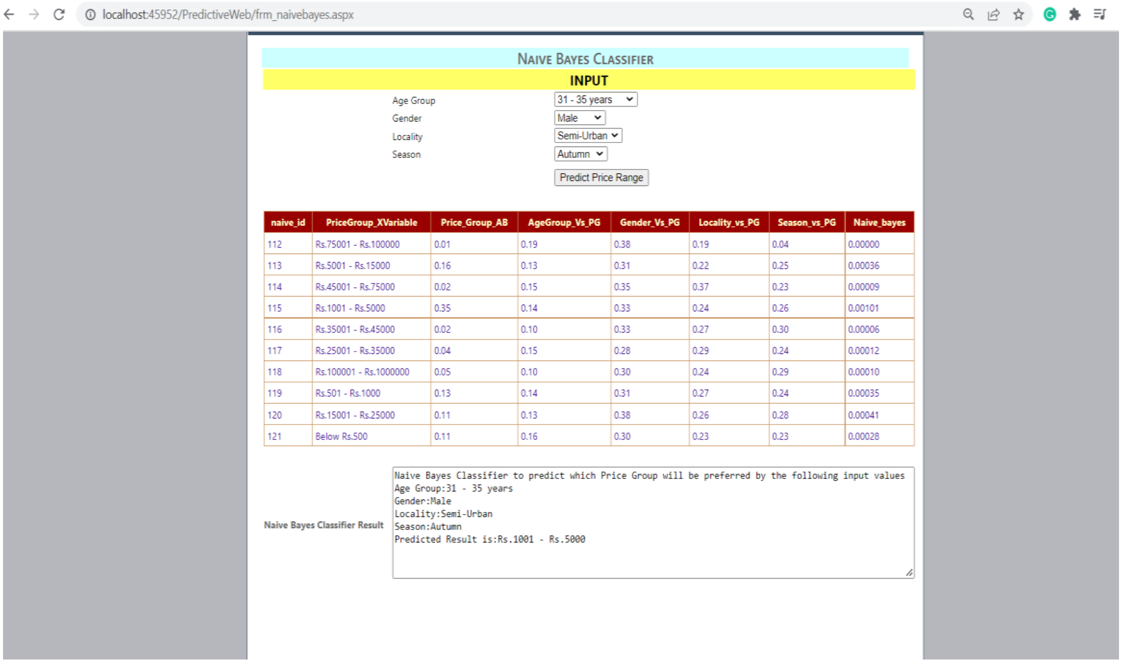
Figure 4.1: K-NN Result

From the above figure 4.1, the K-nearest Neighbour algorithm was applied on the given input data, which included the age and gender of a person, Main Category name, and the period between January 1, 2019, to April 26, 2022. The algorithm used a dataset with instances and their corresponding labels, and calculated the Euclidean distance between the input data point and all instances in the dataset. The algorithm then selected the top k=8 nearest neighbours based on the distance metric and displayed their sales numbers, Euclidean distances, and Main Category name feature. The algorithm then used the k nearest neighbours to determine the class of the input data point by taking a majority vote among the k data points. The table provided in the results shows that the algorithm determined the best fit for the input data point based on the number of instances in each category among the k nearest neighbours. The algorithm determined that the input data point was most likely to belong to the categories of dress or home appliances, with each having a count of 3 among the k

ISSN 1462 2815

nearest neighbours. The categories of book and electronic had a count of 1 each and were less likely to be the class of the input data point. Overall, the K-nearest Neighbour algorithm is a powerful classification algorithm that can be applied to a wide range of datasets and can provide valuable insights into the patterns and trends in the data.

4.2 Naive Bayes Classifier



naive_id	PriceGroup_XVariable	Price_Group_AB	AgeGroup_Vs_PG	Gender_Vs_PG	Locality_vs_PG	Season_vs_PG	Naive_bayes
112	Rs.75001 - Rs.100000	0.01	0.19	0.38	0.19	0.04	0.00000
113	Rs.5001 - Rs.15000	0.16	0.13	0.31	0.22	0.25	0.00036
114	Rs.45001 - Rs.75000	0.02	0.15	0.35	0.37	0.23	0.00009
115	Rs.1001 - Rs.5000	0.35	0.14	0.33	0.24	0.26	0.00101
116	Rs.35001 - Rs.45000	0.02	0.10	0.33	0.27	0.30	0.00006
117	Rs.25001 - Rs.35000	0.04	0.15	0.28	0.29	0.24	0.00012
118	Rs.100001 - Rs.1000000	0.05	0.10	0.30	0.24	0.29	0.00010
119	Rs.501 - Rs.1000	0.13	0.14	0.31	0.27	0.24	0.00035
120	Rs.15001 - Rs.25000	0.11	0.13	0.38	0.26	0.28	0.00041
121	Below Rs.500	0.11	0.16	0.30	0.23	0.23	0.00028

Figure 4.2 Naives Bayes Classifier Result

The result of a Naive Bayes classifier from the figure 4.2 predicts which Price Group will be preferred by a customer with the input values of Age Group (31-35 years), Gender (male), Locality (Semi-urban), and Season (Autumn). The table shows the probabilities of each price group based on the given input values. The predicted result of the Naive Bayes classifier is Rs.1001 - Rs.5000, which has the highest probability of 0.35 among all the price groups. This indicates that customers with the given input values are most likely to prefer products in this price range during the Autumn season. The Naive Bayes classifier is a probabilistic algorithm that assumes the independence of the input features. The classifier calculates the probability of each class given the input features and selects the class with the highest probability as the predicted output. In this case, the classifier has used the input features of age group, gender, locality, and season to predict the preferred price group.

Overall, the Naive Bayes classifier has provided a useful prediction of the preferred price group based on the given input values, which can be used for marketing and sales strategy decisions.

5. CONCLUSION AND FUTURE WORK

In conclusion, this study highlights the significance of predictive analytics in the e-commerce industry to understand and forecast customer behavior. The dataset used in the study includes various features, including demographic variables such as age group, gender, and locality. The study proposes the use of machine learning algorithms, specifically the K-Nearest Neighbors algorithm and Naïve Bayes Classifier, to predict customer preferences and demand, leading to better supply chain management, increased sales, and customer satisfaction. The KNN algorithm was used to identify the most preferred product categories for a given input, while the Naïve Bayes classifier predicted the preferred price group based on the input values. The proposed system aims to use advanced machine learning techniques to analyze vast datasets and extract useful insights efficiently. Overall, this study provides valuable insights into the importance of forecasting and the role of machine learning algorithms in the e-commerce industry.

References

1) E. Lindecrantz, M. T. P. Gi, and S. Zerbi, "Personalizing the customer experience: Driving differentiation in retail.," *McKinsey Insights*, no. March, pp. 1–7, 2020.

2) F. Tiedemann, "Demand-driven supply chain operations management strategies – a literature review and conceptual model," *Prod. Manuf. Res.*, vol. 8, no. 1, pp. 427–485, Jan. 2020, doi: 10.1080/21693277.2020.1856012.

3) N. Martin, "How AI Is Revolutionizing Digital Marketing," *Forbes*, 2018. https://www.forbes.com/sites/nicolemartin1/2018/11/12/how-ai-is-revolutionizing-digital-marketing/?sh=53bcee611f62 (accessed Apr. 24, 2023).

4) S. Akter and S. F. Wamba, "Big data analytics in E-commerce: a systematic review and agenda for future research," *Electron. Mark.*, vol. 26, no. 2, pp. 173–194, May 2016, doi: 10.1007/s12525-016-0219-0.

5) I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput. Sci.*, vol. 2, no. 3, p. 160, May 2021, doi: 10.1007/s42979-021-00592-x.

6) J. Frankenfield, "General Data Protection Regulation (GDPR) Definition and Meaning," 2020. https://www.investopedia.com/terms/g/general-data-protection-regulation-gdpr.asp (accessed Apr. 24, 2023).

7) Chen, Chiang, and Storey, "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS Q.*, vol. 36, no. 4, p. 1165, 2012, doi: 10.2307/41703503.

8) A. Uzialko, "How Artificial Intelligence Is Transforming Business - businessnewsdaily.com," 2023. https://www.businessnewsdaily.com/9402-artificial-intelligence-business-trends.html (accessed Apr. 24, 2023).

9) C. Becdach, B. Brown, F. Halbardier, B. Henstorf, and R. Murphy, "Rapidly forecasting demand and adapting commercial plans in a pandemic," *McKinsey Co.*, no. April, pp. 1–9, 2020, [Online]. Available: https://disclosure.bursamalaysia.com/FileAccess/apbursaweb/download?id=190786&name=EA_DS_ATTACHMENTS.

10) M. R. Ward, "On Forecasting the Demand for E-Commerce," in *Forecasting the Internet*, Boston, MA: Springer US, 2002, pp. 81–103.

11) L. T. Khrais, "Role of artificial intelligence in shaping consumer demand in e-commerce," *Futur. Internet*, vol. 12, no. 12, pp. 1–14, 2020, doi: 10.3390/fi12120226.

12) Y. R. Zhang and Z. J. Zhao, "Study on Consumer Behavior Predict in E-commerce Based on Multi-Agent," *Int. J. u- e-Service, Sci. Technol.*, vol. 7, no. 6, pp. 403–412, Dec. 2014, doi: 10.14257/ijunesst.2014.7.6.35.

13) S. Fedushko and T. Ustyianovych, "E-Commerce Customers Behavior Research Using Cohort Analysis: A Case Study of COVID-19," *J. Open Innov. Technol. Mark. Complex.*, vol. 8, no. 1, p. 12, Jan. 2022, doi: 10.3390/joitmc8010012.

14) W. Zhang and M. Wang, "An improved deep forest model for prediction of e-commerce consumers' repurchase behavior," *PLoS One*, vol. 16, no. 9, p. e0255906, Sep. 2021, doi: 10.1371/journal.pone.0255906.

15) H. Yao, "Analysis Model of Consumer Sentiment Tendency of Commodities in E-Commerce," *Front. Psychol.*, vol. 13, Jun. 2022, doi: 10.3389/fpsyg.2022.887923.