Customer Segmentation & Predictive Analysis

1st Keerthi Sri Cherukuri
Department of Info Science & Tech
University of Missouri S&T
Rolla,USA
kc6bd@mst.edu

2nd Ganesh Guddanti
Department of Computer Science
University of Missouri S&T
Rolla,USA
gg7hz@mst.edu

3rd Kishore Kumar Jami
Department of Computer Science
University of Missouri S&T
Rolla,USA
kjrdc@mst.edu

4th Venkata Mokshagna Nadella
Department of Info Science & Tech
University of Missouri S&T
Rolla,USA
vnqrd@mst.edu

5th Shruthi Shinde
Department of Info Science & Tech
University of Missouri S&T
Rolla,USA
ssd7n@mst.edu

6th Sasidhar Reddy Velkuri
Department of Info Science & TEch
University of Missouri S&T
Rolla,USA
svdfy@mst.edu

***Abstract*—Understanding customer preferences in the current competitive business landscape is essential to ensure efficient marketing strategies. Businesses can use customer segmentation to tailor their marketing strategy to specific client groups, resulting in personalized product suggestions and greater sales. For instance, companies might target customers with similar products to increase sales if a top-selling product is found inside a particular client segment. By combining predictive analysis with consumer segmentation, firms can predict future behavior, spot patterns, and anticipate client demands. By optimizing the potential of best-selling products, this integration not only increases consumer engagement and loyalty but also promotes long-term success.**

***Index Terms*—K-means, PCA, Recommendation Systems, Predictive Analysis**

I. Introduction

Customer segmentation is a key strategy used in e-commerce that helps companies to target and interact with different client segments. Using transactional data, companies can learn more about the preferences, behavior, and buying habits of their customers. A thorough customer-centric dataset is created by adding new features to the dataset, such as purchase history, interaction data, and customer demographics. Then, to improve this dataset's usability and accuracy in identifying pertinent consumer segments, feature scaling and dimensionality reduction techniques are applied.

Customers are divided into various categories according to similarities in their demographics, purchasing habits, and preferences using the well-liked unsupervised machine learning technique K-means clustering. Businesses can customize their product offers and marketing methods to target certain client segments thanks to this segmentation, which results in more specialized and focused approaches. Businesses may better serve their consumers' requirements and preferences by knowing the distinctive traits and behaviors of each segment, which will eventually increase customer happiness and loyalty.

An important factor in determining the segmentation process's efficacy is cluster analysis and evaluation. Businesses can obtain important insights into the most profitable or responsive consumer categories by analyzing the discovered clusters and their features. By ensuring that the segmentation is meaningful and actionable, this review helps organizations manage resources wisely and make well-informed decisions.

Additionally, the consumer segmentation process is further improved by the integration of predictive analysis tools and recommendation systems. Recommendation systems improve user experience and increase engagement by using algorithms to offer consumers personalized suggestions based on their past interactions and behavior. Conversely, predictive analysis uses statistical modeling techniques and historical data to predict future patterns and behaviors, allowing organizations to anticipate client needs.

Overall, companies can better understand their client base, and target consumer categories, and drive growth in the competitive marketplace of e-commerce by utilizing analytical approaches and procedures in customer segmentation. Businesses may enhance client relationships, maximize marketing efforts, and ultimately succeed in the long run in the digital marketplace by using this strategic approach to segmentation.

II. Literature Review

Utilizing data science and machine learning methods to enhance sales prediction models and product recommendation systems has attracted much attention in recent years. These developments are essential for companies looking to improve their comprehension of customer behavior and maximize their marketing tactics. Below is a summary of important research in this field:

Yıldız et al. (2023) [1] introduced a personalized recommendation system tailored for the fashion retail sector. Their approach emphasizes customer segmentation to provide

Identify applicable funding agency here. If none, delete this.

highly relevant product recommendations, potentially leading to increased sales.

Zhao and Keikhosrokiani (2022) [2] proposed a model that leverages user behavior analytics for sales prediction and product recommendations. By analyzing user interactions, this model offers insights into purchasing patterns, aiding businesses in inventory management and marketing decision-making.

Nikose et al. (2022) [3]focused on predicting best-selling products and categories through sales analysis. Their study contributes to understanding market trends and consumer preferences, which is valuable for businesses looking to optimize their product offerings and promotional strategies.

Satheesan et al. (2020) [4]developed a product recommendation system specifically designed for supermarkets. By employing machine learning techniques, their system offers personalized recommendations, thereby enhancing the shopping experience for customers.

Chkoniya (2020) [5] discussed the challenges involved in using data science approaches to decode consumer behavior. This review highlights the complexities of analyzing consumer data and emphasizes the importance of addressing privacy concerns and data biases.

Rodrigues and Ferreira (2016) [6] proposed a recommendation system based on shared customer behavior. Their approach leverages collective intelligence to improve recommendation accuracy, particularly in scenarios with limited user-item interactions.

García et al. (2016) [7] provided an overview of big data preprocessing techniques. Their study discusses various methods for data cleaning, transformation, and reduction, which are essential for preparing large datasets for analysis in consumer behavior studies.

Montesinos López et al. (2022) [8] explored multivariate statistical machine learning methods for genomic prediction. While not directly related to consumer behavior, their work underscores the diverse applications of machine learning techniques in predictive modeling.

Wong and Wei (2018) [9] proposed an integrated model for online retail customer segmentation and service prediction, aiming to enhance satisfaction and engagement. Published in the International Journal of Retail & Distribution Management, their study underscores the importance of data analytics in tailoring services to improve online retail strategies.

Chen, D., Sain, S., and Guo, K. (2012) [10] conducted a case study on data mining for online retail, focusing on RFM model-based customer segmentation. Their research, published in the Journal of Database Marketing & Customer Strategy Management, Volume 19, pages 197-208, highlights the significance of RFM models in enhancing customer segmentation strategies for online retail.

III. Proposed Method

A. *Initial Data Analysis:*

First, importing the necessary packages to analyze the data. The dataset used for this initial data analysis was collected from the UCI Machine Learning Repository. It comprises original transactions from the UK spanning from 2010 to 2011. The dataset is stored in an Excel spreadsheet file, denoted by the extension .xlsx.

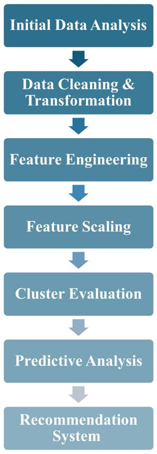


Fig. 1. Proposed Methodolgy

B. *Data Cleaning and Transformation:*

After importing the packages and the data. We can find a lot of errors in the data such as missing values, duplicates, cancelled transactions, misplaced stock codes. So in order to processing this data first step is cleaning the data and organizing the data such a way it is error free.

C. *Feature Engineering:*

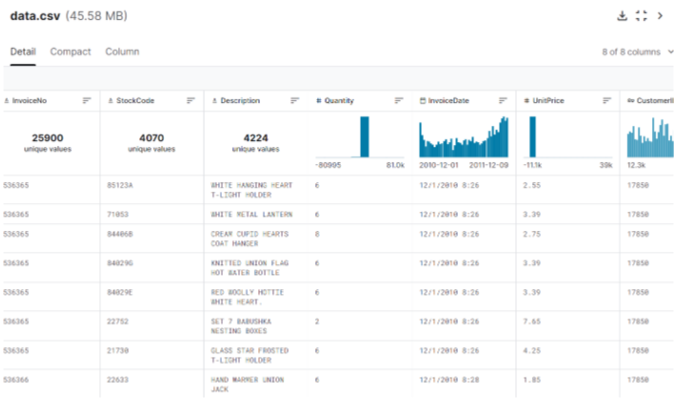
We engineer RFM metrics alongside features like Unique Products Purchased, Average Days Between Purchases, and Favorite Shopping Day/Hour to create a customer-centric dataset. These features facilitate personalized marketing, targeted promotions, and optimized product recommendations, leveraging insights from customer behavior and preferences.

D. *Feature Scaling:*

Feature scaling is essential for K-means clustering and PCA to ensure accurate results. Unevenly scaled features in K-means can skew distance calculations, impacting cluster formation and finding the optimal k using the elbow point method, while in PCA, larger-scale features may dominate principal components, distorting data patterns and Visualization of Customer Clusters in PCA Space. Scaling ensures fair representation of features, enhancing the accuracy of clustering and dimensionality reduction.

E. *Cluster Evaluation:*

After identifying the optimal number of clusters (3), I'll evaluate their quality. This involves visualizing top principal components in 3D to gauge cluster separation, assessing cluster distribution, and utilizing evaluation metrics like Silhouette, Calinski Harabasz, and Davies Bouldin scores to ensure coherent and distinct clusters.



data.csv (45.58 MB)

Detail Compact Column — 8 of 8 columns

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerI
25900 unique values	4070 unique values	4224 unique values	-80995 81.0k	2010-12-01 2011-12-09	-11.1k 39k	12.3k
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850
536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850
536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12/1/2010 8:26	7.65	17850
536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	12/1/2010 8:26	4.25	17850
536366	22633	HAND WARMER UNION JACK	6	12/1/2010 8:28	1.85	17850

Fig. 2. Data set

F. *Predictive Analysis:*

Our prediction analysis aims to forecast customer spending over the next three or more months using predictive modeling and historical transaction data. This allows us to anticipate individual purchasing habits, enabling proactive marketing campaigns and tailored recommendations based on expected needs and preferences.

G. *Recommendation system:*

Developing a recommendation system that makes product recommendations based on user cluster purchase patterns, thereby improving the online shopping experience. I'll examine cleaned customer data to determine the best-selling items in each cluster after separating outliers, allowing for tailored suggestions.

IV. Experiments

A. *Dataset*

E-commerce datasets are often proprietary, making them challenging to access publicly. However, the UCI Machine Learning Repository offers a dataset named ”Online Retail,” containing real transactions from 2010 and 2011. This dataset is freely available on their website, facilitating research and analysis in the e-commerce domain.

B. *Implementation*

1) *Elbow Method:*

- **KMeans Clustering:**The KMeans algorithm is chosen for its simplicity and efficiency in clustering large datasets.Multiple iterations of KMeans clustering are performed with different values of k (number of clusters) to determine the optimal number of clusters for customer segmentation.
- **Evaluation Metrics:**Two primary metrics are used to assess the quality of clustering: The Elbow Score is plotted against different values of k to identify the point where adding more clusters ceases to significantly reduce within-cluster inertia. Conversely, the Silhouette Score measures cluster cohesion and separation, with a higher score indicating more distinct clusters.

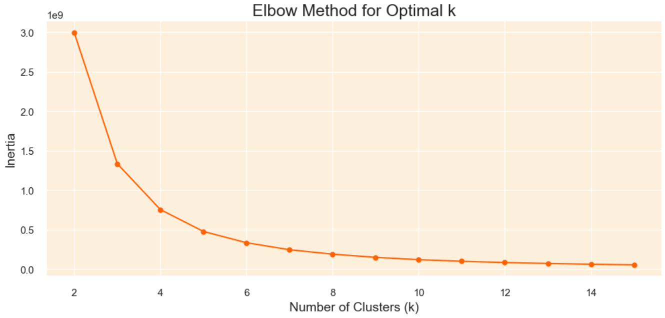


Fig. 3. Elbow Method for Optimal k

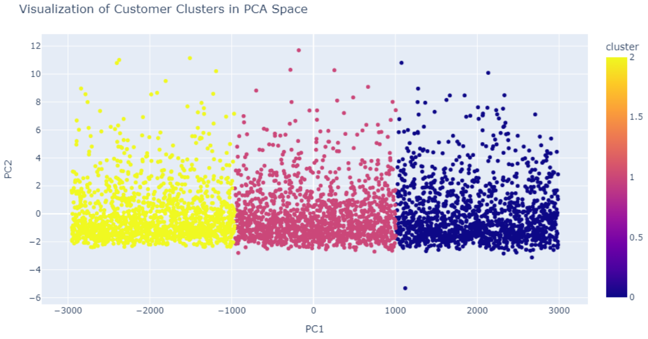


Fig. 4. Visualization of Customer Clusters in PCA Space

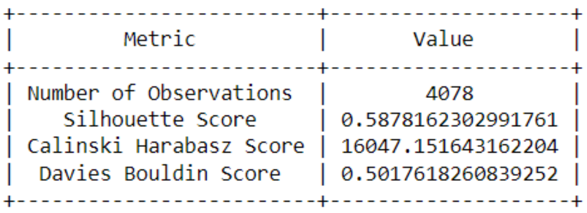
2) *Clustering Evalution::*

- **Optimal K Determination:**Utilizing the Elbow Method and Silhouette Score, the optimal number of clusters (k=3) is identified, ensuring meaningful segmentation without overfitting.
- **Clustering Analysis:**KMeans clustering is applied with the determined optimal k value to partition the customer data into three distinct clusters.The algorithm iteratively assigns customers to the nearest cluster centroid based on their feature similarities.
- **Visualization:**Visual representation of the clustered data in a two-dimensional space using PCA-transformed features.Each point on the scatter plot represents a customer, colored according to their assigned cluster.

3) *Evaluation Metrics::*

- **Silhouette Score:** Measures the compactness and separation of clusters.
- **Calinski-Harabasz Score:** Evaluates the between-cluster dispersion and within-cluster dispersion ratio.
- **Davies-Bouldin Score:** Quantifies the average similarity between each cluster and its most similar cluster.

4) *Radar Chart Approach:* Initially, I'll generate radar charts to visually depict the centroid values of each cluster across diverse features. These charts offer a rapid comparison of different cluster profiles. Before constructing the radar charts, it's crucial to compute the centroid for each cluster, which denotes the mean value across all features within the



```
+---------------------------+---------------------+
|          Metric           |        Value        |
+---------------------------+---------------------+
|  Number of Observations   |         4078        |
|     Silhouette Score      |  0.5878162302991761 |
|  Calinski Harabasz Score  |  16047.151643162204 |
|   Davies Bouldin Score    |  0.5017618260839252 |
+---------------------------+---------------------+
```

Fig. 5. Evaluation Metrics

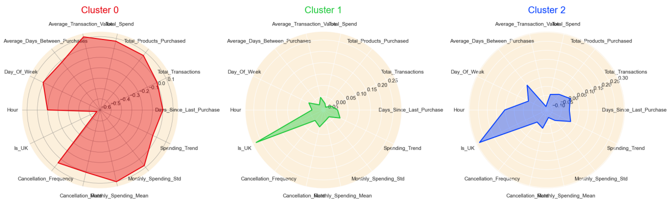


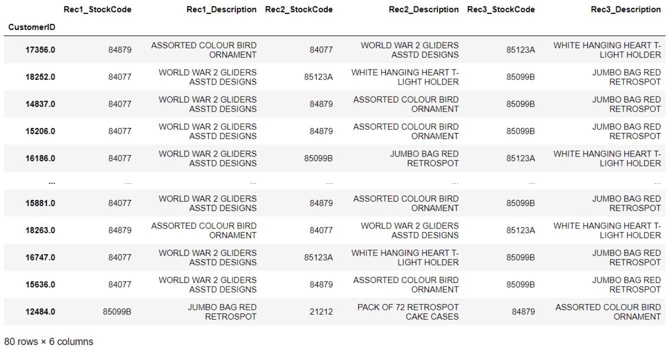
Fig. 6. Radar Charts Comparing Customer Cluster Profiles

cluster. These centroids will then be showcased on the radar charts, providing a clear visualization of the central tendencies of each feature across the clusters.

V. Results & Analsyis

As the project nears its conclusion, our attention turns to creating a recommendation system that will improve the online buying experience. We seek to identify top-selling goods by examining the purchase trends within each cluster of the core 95% client group. By utilizing this knowledge, the system will provide users with tailored suggestions, recommending the top three items that are popular within their specific clusters but have not yet been bought. This customized strategy improves targeted marketing efforts and enriches the entire buying experience, which may result in more sales. Furthermore, one first tactic for the outlier group may be to suggest random goods in order to spark interest and reveal possible preferences.

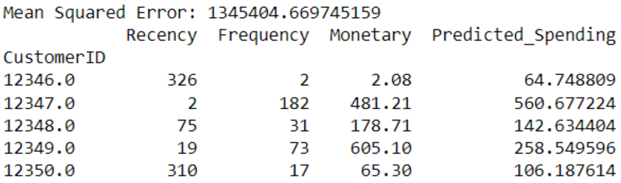
The implementation of an RFM (Recency, Frequency, Monetary) analysis and a predictive model to forecast future spending behavior of customers. It begins by calculating the RFM metrics for each customer based on their transaction history. Next, it splits the dataset into training and testing sets, employing a Linear Regression model to predict monetary spending based on recency and frequency of purchases. The model's performance is evaluated using Mean Squared Error (MSE), providing insights into its accuracy. Finally, future spending is predicted for each customer, and the results are displayed alongside their original RFM metrics. This comprehensive approach enables businesses to understand customer behavior and forecast potential revenue streams effectively.



CustomerID	Rec1_StockCode	Rec1_Description	Rec2_StockCode	Rec2_Description	Rec3_StockCode	Rec3_Description
17356.0	84879	ASSORTED COLOUR BIRD ORNAMENT	84077	WORLD WAR 2 GLIDERS ASSTD DESIGNS	85123A	WHITE HANGING HEART T-LIGHT HOLDER
18252.0	84077	WORLD WAR 2 GLIDERS ASSTD DESIGNS	85123A	WHITE HANGING HEART T-LIGHT HOLDER	85099B	JUMBO BAG RED RETROSPOT
14837.0	84077	WORLD WAR 2 GLIDERS ASSTD DESIGNS	84879	ASSORTED COLOUR BIRD ORNAMENT	85099B	JUMBO BAG RED RETROSPOT
15206.0	84077	WORLD WAR 2 GLIDERS ASSTD DESIGNS	84879	ASSORTED COLOUR BIRD ORNAMENT	85099B	JUMBO BAG RED RETROSPOT
16186.0	84077	WORLD WAR 2 GLIDERS ASSTD DESIGNS	85099B	JUMBO BAG RED RETROSPOT	85123A	WHITE HANGING HEART T-LIGHT HOLDER
...	...	...	...	...	...	...
15881.0	84077	WORLD WAR 2 GLIDERS ASSTD DESIGNS	84879	ASSORTED COLOUR BIRD ORNAMENT	85099B	JUMBO BAG RED RETROSPOT
18263.0	84879	ASSORTED COLOUR BIRD ORNAMENT	84077	WORLD WAR 2 GLIDERS ASSTD DESIGNS	85123A	WHITE HANGING HEART T-LIGHT HOLDER
16747.0	84077	WORLD WAR 2 GLIDERS ASSTD DESIGNS	85123A	WHITE HANGING HEART T-LIGHT HOLDER	85099B	JUMBO BAG RED RETROSPOT
15636.0	84077	WORLD WAR 2 GLIDERS ASSTD DESIGNS	84879	ASSORTED COLOUR BIRD ORNAMENT	85099B	JUMBO BAG RED RETROSPOT
12484.0	85099B	JUMBO BAG RED RETROSPOT	21212	PACK OF 72 RETROSPOT CAKE CASES	84879	ASSORTED COLOUR BIRD ORNAMENT

80 rows × 6 columns

Fig. 7. customer data with recommendations



```
Mean Squared Error: 1345404.669745159
            Recency  Frequency  Monetary  Predicted_Spending
CustomerID
12346.0         326          2      2.08           64.748809
12347.0           2        182    481.21          560.677224
12348.0          75         31    178.71          142.634404
12349.0          19         73    605.10          258.549596
12350.0         310         17     65.30          106.187614
```

Fig. 8. Predictive Spending

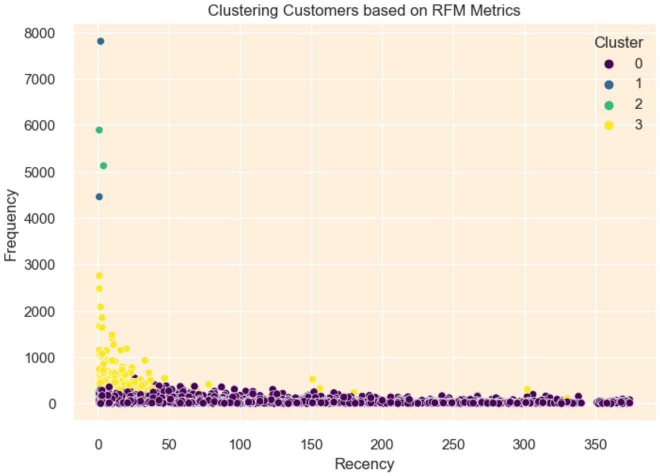


Fig. 9. Clustering Customers based on RFM Metrics

$$\mathrm{MSE}(y, \hat{y}) = \frac{\sum_{i=0}^{N-1} (y_i - \hat{y}_i)^2}{N} \tag{1}$$

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{2}$$

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \tag{3}$$

Using client data, RFM analysis and K-means clustering are used to divide up the consumer base according to their buying habits. Histograms are used to display the monetary, frequency, and recency distributions, and RFM metrics are computed. Based on frequency and recency, K-means clustering divides clients into four clusters that are presented on a scatter plot. Targeted marketing techniques are made easier with the help of the resultant clustered RFM dataframe, which shows allocated cluster labels for every consumer.

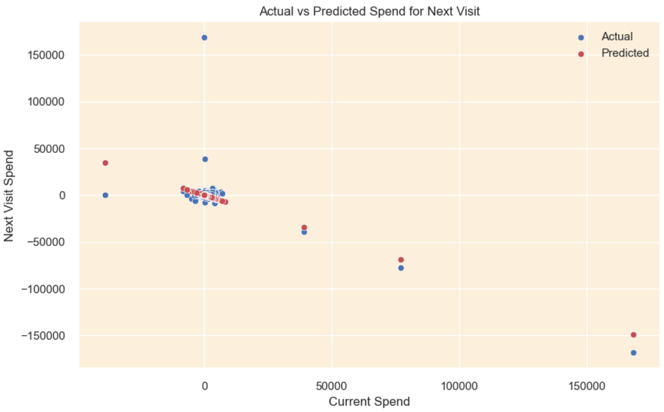


Fig. 10. Actual VS Predicted Spending Analysis

To use linear regression to forecast, from their present spending patterns, how much clients would spend on their next visit. It first adds to the dataset by calculating the total amount spent on each transaction. To maintain chronological order, the data is then sorted according to client ID and invoice date. By moving each customer's spend data to a new column called "SpendNextVisit," the expenditure on each customer's subsequent visit is represented. The dataset is divided into training and testing sets after missing values in the "SpendNextVisit" column are removed.

Next, using the training data, a linear regression model is instantiated and trained. Using the current spend as the input feature, the trained model is then applied to forecast the spending for the subsequent visit. The forecasts are kept in a new column called "PredictedSpendNextVisit." Finally, a scatter plot is created to show the difference between the actual and anticipated costs for the upcoming visit. This figure helps assess how well the linear regression model predicts the purchasing habits of its customers.

Conclusion

- Employed K-means clustering for customer segmentation, enabling targeted marketing strategies and personalized experiences.
- Developed a recommendation system to suggest top-selling products to customers within each segment, enhancing the shopping experience and increasing sales potential.
- Utilized predictive analysis to forecast customer spending and anticipate future trends, showcasing the value of data-driven approaches in driving business growth in online retail.

References

[1] E. Yıldız, C. Güngör Şen, and E. E. Işık, "A Hyper-Personalized Product Recommendation System Focused on Customer Segmentation: An Application in the Fashion Retail Industry," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 18, no. 1, pp. 571–596, Mar. 2023, doi: 10.3390/jtaer18010029.

[2] X. Zhao and P. Keikhosrokiani, "Sales prediction and product recommendation model through user behavior analytics," *Computers, Materials and Continua*, vol. 70, no. 2, pp. 3855–3874, 2022, doi: 10.32604/cmc.2022.019750.

[3] Ms. Archana Nikose, Tejal Mungale, Minal Shelke, Rohini Shelote, and Priyal Solanke, "Best Selling Product and Category Prediction Using Sales Analysis," *International Journal of Advanced Research in Science, Communication and Technology*, pp. 805–811, Mar. 2022, doi: 10.48175/ijarsct-2970.

[4] P. Satheesan, P. S. Haddela, and J. Alosius, "Product Recommendation System for Supermarket," in *Proceedings - 19th IEEE International Conference on Machine Learning and Applications, ICMLA 2020*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020, pp. 930–935. doi: 10.1109/ICMLA51294.2020.00151.

[5] V. Chkoniya, "Challenges in Decoding Consumer Behavior with Data Science," *European Journal of Economics and Business Studies*, vol. 6, no. 3, p. 77, Dec. 2020, doi: 10.26417/897ovg79t.

[6] F. Rodrigues and B. Ferreira, "Product Recommendation based on Shared Customer's Behaviour," in *Procedia Computer Science*, Elsevier B.V., 2016, pp. 136–146. doi: 10.1016/j.procs.2016.09.133.

[7] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, "Big data preprocessing: methods and prospects," *Big Data Anal*, vol. 1, no. 1, Dec. 2016, doi: 10.1186/s41044-016-0014-0.

[8] O. A. Montesinos López, A. Montesinos López, and J. Crossa, *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Springer International Publishing, 2022. doi: 10.1007/978-3-030-89010-0.

[9] Wong, E., & Wei, Y. (2018). Customer online shopping experience data analytics: Integrated customer segmentation and customised services prediction model. *International Journal of Retail & Distribution Management*, *46*(4), 406–420. doi.org/10.1108/IJRDM-06-2017-0130

[10] Chen, D., Sain, S., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*, *19*(2), 197–208.