Kalpa Publications in Computing

Volume 19, 2024, Pages 206–216

Proceedings of 6th International Conference
on Smart Systems and Inventive Technology

Improving Customer Engagement via Segmentation Empowered by Machine Learning

Mahendhiran P D[1], Harini M[2], Bavana S[2], Swetha D[2]

[1]Assistant Professor, Department of Computer Science and Business Systems, Sri Krishna College of Engineering and Technology, Coimbatore-641008. mahendhiranpd@gmail.com

[2]UG Scholar, Department of Computer Science and Business Systems, Sri Krishna College of Engineering and Technology, Coimbatore-641008. harini.kmanickam@gmail.com

Abstract

Customer segmentation is a critical component of any marketing strategy for contemporary companies that are ready to compete in a market that is fiercely competitive today. The goal is to use segmentation of customers methodologies to better decision-making, marketing tactics, and customer satisfaction levels in general. The project will get started by gathering and studying various client data, including comments, purchase history, online activity, and demographics. The consumer base will be segmented into various segments based on shared traits and preferences using data analytics as well as machine learning algorithms. The primary objective is to leverage these segments to optimize various facets of business operations, such as marketing campaigns, product development, and customer support. Enterprises can position themselves for enduring expansion and triumph in the constantly evolving and fiercely competitive commercial sphere by adeptly segmenting their consumer demographics and adjusting their approaches accordingly. Through these insights, businesses can achieve more efficient resource utilization and improved ROI (Return On Investment). Emphasizing the significance of customer segmentation as a strategic tool enhances business performance.

Keywords: Customer Segmentation - Customer Classification - Customer Analytics – Personalization - Customer Clustering

1 Introduction

In the dynamic landscape of modern business, the pursuit of effective customer segmentation strategies is integral to understanding and responding to the diverse and evolving needs of consumers. This research paper embarks on a comprehensive exploration of contemporary methodologies in customer segmentation, synthesizing insights from recent scholarly works to contribute to the ongoing discourse in this field. Recent studies have demonstrated a diverse array of approaches to customer

segmentation, showcasing the innovation and dynamism within this area of research explored the use of online product reviews as a tool for interpreting machine learning methods for segmenting customers, providing valuable insights for new product development[1].Ma et al. focused on stochastic-user-equilibrium perspective to address customer segmentation, lead time decisions and pricing offering a novel perspective that considers uncertainty in consumer behavior[2].The integration of recent technologies have become a prevalent theme in recent research focused on explainable AI in product development for customer segmentation, emphasizing the importance of interpretability in understanding segmentation outcomes[3]. A smart meter data analytics for customer segmentation, demonstrating the practical applications of data-driven approaches in understanding consumer behavior[4].Beyond traditional segmentation methodologies, employed the jobs-to-be-done theory to compare customer needs in both direct and through online fashion shopping. These findings offer a nuanced understanding of customer behavior across different retail channels[5]. Boehe and Becerra examined market entry strategies and the likelihood of firms imitating competitors' market presence, providing insights into competitive behaviors in international markets[6].The interplay between competitor strategies and firm-level decisions further investigated the relationship between alliance orientations and competitor in product innovativeness[7], The research sheds light on technological strategies influenced by competitive landscapes by employing association rule mining to find the competitors[8]. Moreover, studies also focused on modeling customer satisfaction through online reviews and heuristic clustering methods, respectively, offering alternative perspectives on understanding and segmenting customers[9][10]. This paper aspires to contribute a comprehensive framework that has customer segmentation techniques and competitor analysis, empowering businesses with actionable insights for strategic decision-making.

2 Related work

2.1 RFM and Customer Segmentation for E-commerce Data Analysis

An Ensemble Machine Learning Approach by Li et al. proposed a machine learning ensemble approach to segment ecommerce customers .It focuses on using the RFM (Recency, Frequency, Monetary) analysis framework to extract features from customer data. Various machine learning algorithms were then utilized to categorize users on the basis of their purchase pattern, including decision trees, gradient boosting machines and random forests. The ensemble approach combines the results of these algorithms to improve segmentation accuracy.

2.2 Customer Segmentation in Financial Services Using Unsupervised Machine Learning

Barros, Pereira presented a methodology for customer segmentation in the financial services industry using unsupervised machine learning algorithms. According to the authors, customers are segmented based on their financial characteristics and behavior using clustering algorithms such as K-means. They also use dimensionality reduction techniques to visualize and interpret the segmentation results.

2.3 Customer Segmentation in E-commerce using Convolutional Neural Networks

Teixeira et al. in their study explored the advantage of convolutional neural networks (CNNs) for users classification in e-commerce. The authors preprocess customer data, including user behavior and interactions, into image-like representations. They then used CNNs to extract features from these representations and segment customers based on their browsing and purchasing patterns. The study demonstrates the effectiveness of CNNs in capturing complex patterns in customer data for segmentation purposes.

2.4 Customer Segmentation Using Machine Learning Techniques in Retail Banking

Ghosh et al. discussed on the application of machine learning techniques for customer segmentation in retail banking. It preprocess customer data, including transaction history and demographic information, to extract relevant features. Consequently, they divided users according to their banking habits and preferences utilizing range of machine learning algorithms, such as clustering and classification algorithms. The study highlights the importance of personalized marketing strategies based on customer segmentation in the retail banking industry.

2.5 Customer Segmentation in Retail Industry Using Machine Learning Algorithms

This study evaluated the effectiveness of various machine learning algorithms for segmenting customers within the retail sector .Yadav et.al, preprocess customer data, including purchase history and demographic information, to prepare it for segmentation. A variety of algorithms were applied to segment the customers based on their shopping preferences and behaviors, including DBSCAN, hierarchical and K-means clustering. The study evaluates the effectiveness of each algorithm in producing meaningful customer segments for targeted marketing campaigns.

2.6 Customer Segmentation in Telecom Industry Using Machine Learning Techniques

This study focused on customer segmentation in the telecom industry using machine learning techniques. Kumar et al. preprocess customer data, including call records and usage patterns, to extract features. Subsequently, they utilized clustering algorithms like DBSCAN and K-means to categorize customers according to their patterns of usage and preferences. The study evaluated the effectiveness of these algorithms in identifying distinct customer segments for targeted marketing campaigns.

2.7 A Comparative Study of Customer Segmentation Techniques Using Machine Learning Algorithms

A. Sharma et al. compared different machine learning algorithms for customer segmentation. The authors preprocess customer data, including transaction history and demographic information, to

prepare it for segmentation. Following that, they employed algorithms like support vector machines (SVMs), hierarchical clustering and K-means clustering, to divide customers into segments. The study assessed each algorithm's performance by considering both computational efficiency and segmentation accuracy.

2.8 Utilizing Machine Learning Techniques for Customer Segmentation in Online Retail

The primary focus of this study was on utilizing machine learning techniques for customer segmentation within the online retail sector .Jain et al. preprocess customer data, including browsing history and purchase patterns, to extract features. They then applied clustering algorithms to segment customers based on their online shopping pattern. The study evaluates the effectiveness of these algorithms in identifying distinct customer segments for personalized marketing strategies.

2.9 Applying Machine Learning to Segment Customers in Healthcare

Chen et al. explored how machine learning can be applied to segment customers within the healthcare sector. It preprocess patient data, including medical history and treatment records, to prepare it for segmentation. They then applied clustering algorithms to segment the patients on the basis of their health profiles. The study demonstrates the potential of machine learning in improving patient care and treatment outcomes through personalized segmentation strategies.

2.10 Customer Segmentation for Personalized Marketing Using Machine Learning Algorithms

This study focused on customer segmentation for personalized marketing using machine learning algorithms by Das et al. preprocess customer data, including purchase history and demographic information, to extract features. They then applied algorithms such as K-means clustering, decision trees, and random forests to segment the users. The study highlights the importance of customer segmentation in enhancing marketing effectiveness and customer satisfaction.

3 Methodology

3.1 Data Interpretation

The first crucial step is the interpretation of the available data. This involves gaining a deep understanding of the dataset, its structure, and the nature of the information it contains. Exploratory Data Analysis (EDA) techniques are employed to uncover patterns, trends, and potential insights. By comprehensively interpreting the data, lay the foundation for informed decision-making throughout the subsequent steps of the methodology. The datasets used are online datasets which contains information about customers, their orders, and various attributes related to their purchasing behavior. For example,

- customer: Unique identifier for each customer.
- order: Order number or identifier.

- total_items: Total number of items purchased in the order.
- discount%: Discount percentage applied to the order and so on

3.2 Selection of Samples

The initial stage of the proposed methodology involves the meticulous selection of samples from the dataset. This step is crucial to ensure that the dataset is representative of the diverse customer base. Various factors, including demographic information, transaction history, and behavioral patterns, are considered to construct a comprehensive dataset that captures the richness and heterogeneity of the customer population. In sample selection a bar plot (Fig 1)is created for easy visualization of the principal component.

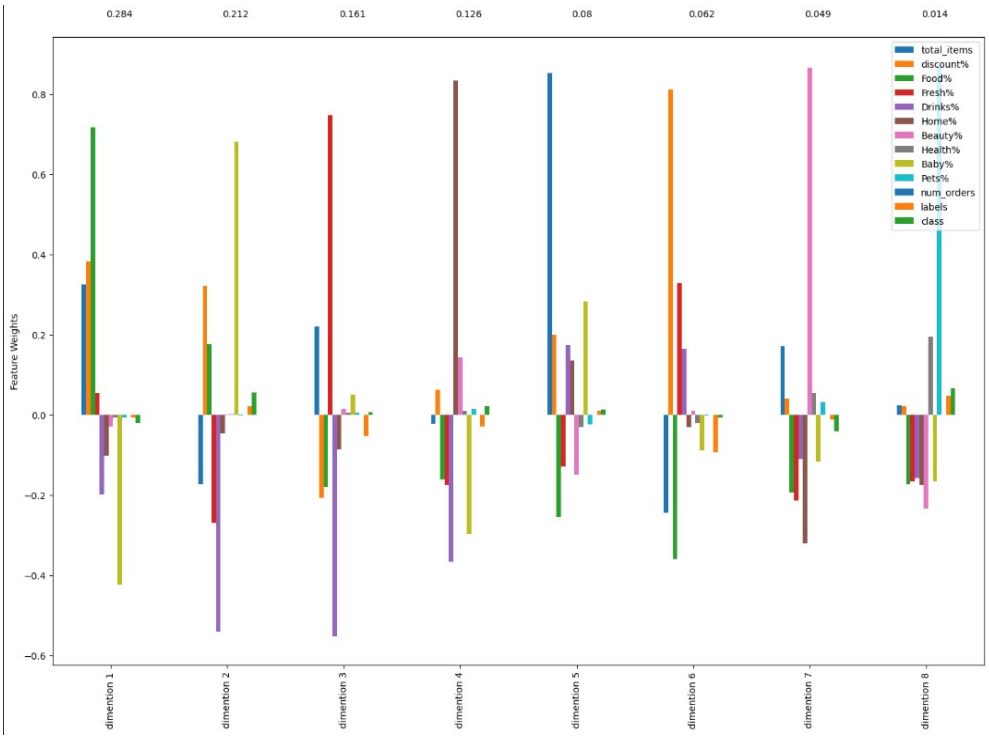

Figure 1: Bar plot to visualize the principal component

PCA is applied to reduce the dimensionality of the dataset df which contains various features related to customer behavior and characteristics. By using methods like fit() [example: pca.fit(df.values)], PCA learns the principal components that capture the most significant variance in the data, effectively reducing the dimensionality of the feature space. The transformed data, containing fewer dimensions (components) than the original dataset, is used for subsequent analysis. After making a bar plot visualization, a 2D scatter plot is viewed based on the results of Principal Component Analysis (PCA). This provides a simplified view of the dataset, allowing for visual inspection of patterns, clusters, or separations in the data. Then a function named (pca_2d_plot_arrow) that extends the 2D scatter plot from the previous

function (pca_2d_plot) by adding arrows representing the direction and magnitude of the principal components. This function takes a PCA object (pca) and a DataFrame (df) as input and plots the data points in a 2D space with arrows indicating the direction of the principal components. Next to the scatter plot using sns.pairplot(df) which creates the pair plot using Seaborn for the DataFrame df and displays it in a matplotlib figure. Seaborn's pair plot function to create a grid of scatterplots and histograms for each pair of variables in the DataFrame df. This allows for a visual inspection of relationships between different pairs of variables, revealing potential patterns or correlations in the data. Then by using function named (plot_corr_matrix) that generates a correlation matrix plot for the DataFrame df. This type of plot is useful for visually inspecting the pairwise correlations between variables in a dataset. It is used to create a visually informative plot that illustrates the pairwise correlations between variables in the DataFrame df. The color-coded correlation matrix provides an overview of relationships between variables based on their strength and direction. This type of visualization is valuable for identifying potential multicollinearity, understanding variable associations, and guiding feature selection in statistical analyses or machine learning tasks.

3.3 Outlier Detection

Prior to employing the suggested algorithm, detecting and addressing outliers is vital as they can significantly impact the clustering process. Outliers can distort the clusters and lead to less meaningful segmentation. Turkey's method, also known as the Tukey method, is a statistical technique used to identify outliers in a dataset based on the interquartile range (IQR). Outliers are detected by calculating the IQR, which is the difference between the third quartile (Q3) and the first quartile (Q1) of the data distribution. Any data points that fall outside the range [Q1 - 1.5* IQR, Q3 + 1.5 * IQR] are considered outliers. Outlier detection is performed on the feature 'total_items' in the dataframe df.The turkey_outlier_detector function is defined to detect outliers using Turkey's method. Within the function, for each column, the first quartile (Q1), third quartile (Q3), and IQR are calculated using np.percentile(). Then, the upper and lower bounds for outlier detection are determined as Q3 + 1.5 * IQR and Q1 - 1.5 * IQR, respectively.Data points that fall outside these bounds are identified as outliers, and their indices are stored in the outlier_indices dictionary for each column. Visualization of total_item after log transformation is shown in Figure 2.

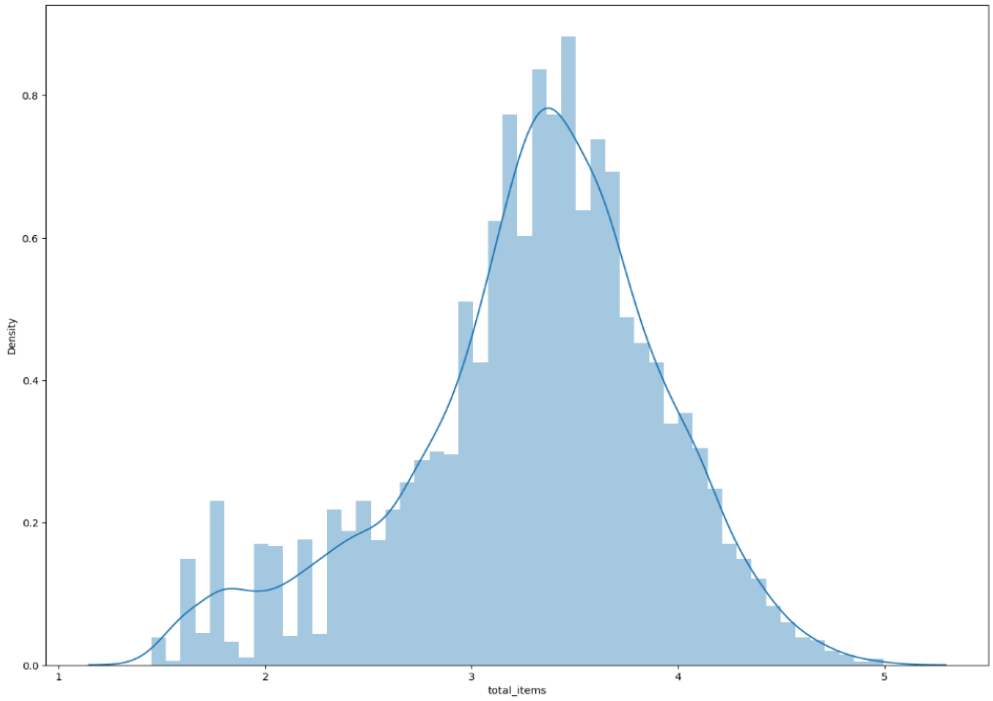

Figure 2: Visualization of total_item after log transformation

3.4 Implementing the K-Means Algorithm

The central aspect of the methodology entails applying the algorithm for the clustering. Pre-processed samples, devoid of outliers, are fed into the algorithm. This process includes the initialization of centroids, iterative assignment of data points to the nearest centroid, and the subsequent updating of centroids until convergence. The goal is to group customers with similar characteristics into distinct segments. Number of clusters are got based on the data set . To analyse the clusters got from K-means clustering next clustering and interpretation has been done. The below dataset is a sample that shows the most preferred product by customers. In this sample data drinks is most preferred product by customers. The table 1 displays the first few rows of a DataFrame.

	order	total_items	Food%	Drinks%	Beauty%
count	71.000000	71.000000	71.000000	71.000000	71.000000
mean	15030.845070	18.661737	8.817190	13.491423	4.445704
std	9313.156356	8.347328	10.947138	15.022799	9.187633
Min	551.000000	7.000000	0.000000	0.000000	0.000000
25%	0005.50000	11.500000	0.000000	0.000000	0.000000
50%	16497.000000	18.000000	4.220000	8.240000	0.000000
75%	22323.500000	24.000000	15.620000	23.190000	4.372167
max	29791.00000	46.000000	36.260000	53.580000	47.250000

Table 1: Sample dataset on clustering that identifies the most preferred product

3.5 Clustering and Interpretation

Upon successful segmentation, each data point is assigned a cluster label. As a next step is to interpret the clusters by analyzing the characteristics and behaviors of customers within each segment. This involves calculating cluster statistics, such as mean values for various features, and understanding the unique attributes that define each cluster. All the analysis so far suggests there could be around 10 clusters in the data, let's now manually examine and try to interpret the meaning of these clusters. For each clusters(labels) interpretation has been done and get to know the potential products. Histogram visualization of customers after segmentation is shown in Figure 3.

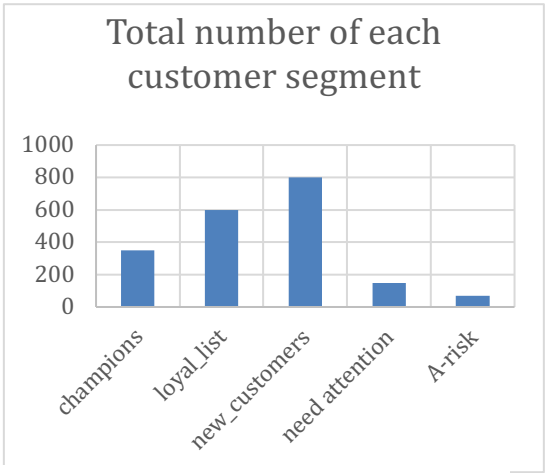

Figure 3: Histogram visualization of customers after segmentation

The data interpretation process involves understanding and extracting insights from the segmented customer data.

3.5.1 Clustering Analysis

After clustering customers using K-means, each cluster is assigned a label or class based on the characteristics of the customers within that cluster. For example, clusters might represent different types of customers such as "fresh_regulars", "home_decorators", "grocery_shoppers", etc., based on their purchasing behavior and preferences.

3.5.2 Interpretation of Clusters

Each cluster's characteristics and behaviors are analyzed to interpret the underlying patterns and preferences. Descriptive statistics such as mean values for different features within each cluster can provide insights into the purchasing habits of customers in that cluster.

3.5.3 Labeling Clusters

Based on the analysis, clusters are labeled with descriptive names that reflect the dominant characteristics of the customers within each cluster. For example, a cluster with customers who predominantly purchase fresh produce might be labeled as "fresh_regulars", while a cluster with customers focused on home decor items might be labeled as "home_decorators".

3.5.4 Visualization

Visualizations such as scatter plots, bar charts, or pie charts can be generated to visualize the distribution of customers across different clusters and understand the differences in their characteristics. These visualizations aid in interpreting the clusters and communicating the findings effectively.

3.6 Competitor analysis

For competitor analysis first set the number of clusters and perform K-Means clustering. Next to that predicting the cluster for a specific data point has been done. Following this extracting unique customer types from a DataFrame is made and then mapping the predicted cluster label to a corresponding customer type is done. To know the competitor ,filter competitors based on predicted customer type and extract unique brand from filtered competitors to know the competitor of a particular company/organization. For example the result displayed will be like 'tata', 'libre', 'mtechra'.

In all the above mentioned steps plots and visualizations are generated to illustrate the segmented groups, aiding in interpretation and insight generation.

4 Result and Discussion

The aim of the project is to predict the valuable competitors of a company based on customers preference. An examination of customer purchasing patterns within the specified market was conducted

using the K-means clustering algorithm. By analyzing customer preferences, behaviors, and characteristics, distinct segments of customers were identified.By using K-means algorithm the most preferred products of the company will be detected (the steps followed are mentioned in section 3 - Methodology) based on this the competitors are predicted.The following correlation matrix (Fig 4)gives the relationships between different customer attributes or features.

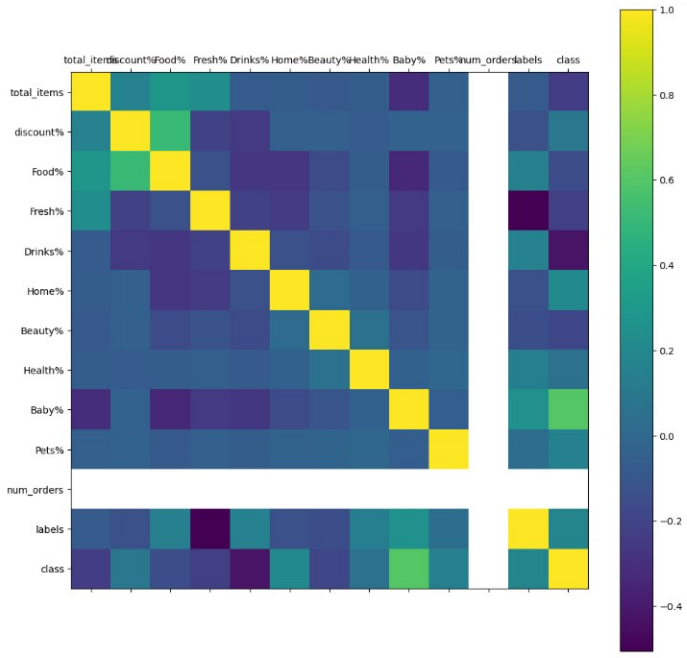

Figure 4: Correlation matrix

Following this clustering and interpretation has been done . Based on this interpretations using K-means algorithm competitors are predicted.

The below mentioned figure 5 is the final result which predicts the competitor for a particular company

```
Out[91]: array(['beautyHm', 'uliver', 'otohom', 'beautyhm'], dtype=object)
```

Figure 5: Result

5 Conclusion

Leveraging machine learning for customer segmentation presents a transformative approach to improving customer engagement. The adoption of advanced algorithms enables businesses to delve

deeper into their customer data, unveiling nuanced patterns and insights that traditional segmentation methods might overlook. This not only refines the understanding of customer behavior but also empowers businesses to deliver more personalized and targeted experiences. The implementation of machine learning in customer segmentation facilitates dynamic and real-time adjustments to engagement strategies. Through personalized recommendations, predictive analytics, and automated customer journeys, businesses can create a more responsive and adaptive framework. This, in turn, fosters stronger customer relationships, increases satisfaction, and enhances overall engagement.

References

1. Boehe, Dirk Michael, and Manuel Becerra. "Market entry into new export markets: When are firms more likely to imitate their competitors' market presence?." International Business Review 31.5 (2022): 102012.

2. Chen, Yen-Chun, et al."Understanding the interplay between competitor and alliance orientations in product innovativeness: An integrative framework." Technological Forecasting and Social Change 175 (2022): 121358.

3. Darko, Adjei Peter, and Decui Liang. "Modeling customer satisfaction through online reviews: A FlowSort group decision model under probabilistic linguistic settings." Expert Systems with Applications 195 (2022): 116649.

4. Hu, Xin, et al. "Explainable AI for customer segmentation in product development." CIRP Annals (2023).

5. Joung, Junegak, and Harrison Kim. "Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews." International Journal of Information Management 70 (2023): 102641.

6. Komatsu, Hidenori, and Osamu Kumaran. "Customer segmentation based on smart meter data analytics: Behavioral similarities with manual categorization for building types." Energy and Buildings 283 (2023): 112831.

7. Kullak, Franziska S., Daniel Baier, and Herbert Woratschek. "How do customers meet their needs in in-store and online fashion shopping? A comparative study based on the jobs-to-be-done theory." Journal of Retailing and Consumer Services 71 (2023): 103221.

8. Ma, Jun, Barrie R. Nault, and Yiliu Paul Tu. "Customer segmentation, pricing, and lead time decisions: A stochastic-user-equilibrium perspective." International Journal of Production Economics 264 (2023): 108985.

9. Sun, Zhao-Hui, et al. "GPHC: A heuristic clustering method to customer segmentation." Applied Soft Computing 111 (2021): 107677.

10. Wu, Yingwen, and Yangjian Ji. "Identifying firm-specific technology opportunities from the perspective of competitors by using association rule mining." Journal of Informetrics 17.2 (2023): 101398.