International Journal of Enhanced Research in Science, Technology & Engineering
ISSN: 2319-7463, Vol. 13 Issue 12, December-2024

Comparative Analysis of Machine Learning Algorithms in Customer Segmentation for Improvement in the Performance of an Online Retail Store

Arun Kumar Mishra[1], Megha Sinha[2], Sudhanshu Kumar Jha[3]

[1]Ph.D. Research Scholar, Department of Computer Science and Engineering, Sarala Birla University, Ranchi-835103, Jharkhand, India
[2]Assistant Professor, Department of Computer Science and Engineering, Sarala Birla University, Ranchi-835103, Jharkhand, India
[3]Assistant Professor (CSE), Department of Electronics and Communication, Faculty of Science, University of Allahabad, Prayagraj-211002, Uttar Pradesh, India

ABSTRACT

Businesses are trying hard to survive and prosper in today's competitive environment by trying to satisfy consumers' increased expectations by optimizing their supply chains. Firms are making strategies for the success keeping consumer at the centre. One of the most important strategies for improving supply chain performance is customer segmentation. There are a number of segmentation techniques, such as RFM (Recency, Frequency, Monetary) segmentation, behavioral, psychographic, geographic, and demographic segmentation. Businesses can better address individual needs, optimize inventory management, enhance customer service, and cut costs by segmenting their customer base into discrete groups and customizing their supply chain strategy accordingly. Data gathering, criteria definition, segment profiling, plan development, and ongoing monitoring are all part of implementation. An effort is being made in this paper to segment the customer data into different clusters by using machine learning techniques on the online retail dataset. Further, comparison of the performance of three clustering algorithms viz. K-Means, DBSCAN and Hierarchical Clustering algorithm was also performed using Silhouette score as a performance metrics. The paper also discusses ethical issues and difficulties, with a focus on the significance of precise data analysis and protecting client privacy.

Keywords— Data Analytics, Customer Segmentation, Supply Chain Management, RFM Segmentation, K-Means,

INTRODUCTION

The business world is changing at a very fast rate. Over the years, data has increased exponentially. Digitization of data, storage in the cloud, online transactions, tracking of products, Social media, E-commerce has facilitated this growth. Businesses want to increase revenue, enhance customer experience and build sustainable practices for survival and growth in this competitive environment. As a human being, everyone wants better health, home and look. Having fun and overall well-being of individual has remained the primary concern for an enlightened society. The technologies like Augmented Reality (AR)/ Virtual Reality (VR), Cloud Computing, Internet of Things (IoT), Data Science, Artificial Intelligence, Machine Learning and 3D Printing are helping businesses to realize the business objectives.

The success of a firm is heavily influenced by the performance of its supply chain in a market that is becoming more and more competitive and dynamic. Customer segmentation is a useful tactic to improve supply chain efficiency. Businesses may customize their supply chain strategies to each segment's specific requirements and preferences by grouping clients according to predetermined criteria. This strategy lowers expenses, increases overall operational efficiency, and optimizes inventory management in addition to raising customer happiness. An effort is being made in this paper to segment the customer data into different clusters by using machine learning techniques on the online retail dataset. Further, comparison of the performance of three clustering algorithms viz. K-Means, Density-based spatial clustering of applications with

noise (DBSCAN) and Hierarchical Clustering algorithm was also performed using Silhouette score as a performance metrics. The paper also discusses ethical issues and difficulties, with a focus on the significance of precise data analysis and protecting client privacy.

LITERATURE REVIEW

Ref. [1] aimed to advance the understanding of how machine learning can be used to improve customer segmentation and support economic stability and growth by methodically examining various clustering algorithms to determine the most efficient method for correctly classifying credit card customers.

The K-Means clustering model successfully divided the credit card customer dataset, displaying excellent clustering quality. The financial sector's decision-making processes can be improved and made more responsive to changing consumer behaviors by using a dynamic approach to client segmentation, which would eventually promote economic growth and stability. The study [2] used unsupervised learning techniques to analyze customer segmentation and discovered that there were more female customers than male customers, and that the mean and median incomes of male consumers were higher than those of female customers. Women made up 56% of all patrons at the mall, making up a larger proportion of the customer base than men. The mean and median yearly income of male clients was higher than that of female customers. The mean spending score of female customers was higher than that of male customers. Five unique client segments were found by the study based on annual income and spending score. With a high accuracy rate of 95% in customer segmentation based on shared behaviors and characteristics, Ref. [3] explored the use of the k-means clustering algorithm in conjunction with RFM (Recency, Frequency, Monetary) analysis for effective customer segmentation. This could offer online retail companies valuable insights to develop targeted marketing strategies and enhance customer satisfaction and business performance. Combining RFM analysis with the k-Means clustering technique produced a high accuracy rate of 95% when it came to consumer segmentation based on shared behaviors and attributes.

The k-Means clustering algorithm's effectiveness in precisely segmenting and classifying clients into discrete clusters based on their commonalities was demonstrated by the high purity value of 0.95. The algorithm's ability to accurately segment customers made it possible to implement personalized and targeted marketing campaigns. Ref. [4] showed that Informed, data-driven marketing decisions were made possible by DeepLimeSeg, a customer segmentation method that integrates explainable AI and deep learning inside a mathematical framework. The DeepLimeSeg model had great prediction accuracy and the capacity to explain the variance in customer spending scores, as evidenced by its low mean square error of 0.9412 and high R-squared value of 0.94152. Targeted marketing campaigns could be benefitted from the precise and understandable client segmentation data that DeepLimeSeg's explainable AI and deep learning algorithms produced. In order to identify three important client clusters—new, best, and intermittent— Ref. [5] presented a customer profiling and segmentation approach utilizing RFM analysis and clustering algorithms. It also offered insights to help digital start-ups maximize customer engagement and foster long-term business growth.

The development of a customer profile and sales prediction to support decision-makers in making strategic marketing decisions was the study's primary output. The research's scientific innovation was the framework for client profiling that was created utilizing AI techniques. In addition to providing useful insights for businesses looking to enhance customer retention and benefit upgrades through data-driven customer segmentation, the study examined consumer segmentation and the role of AI in understanding customer behavior. In order to assess the state of artificial intelligence (AI) in fashion e-commerce, comprehend how AI boosts business profitability, and pinpoint areas in need of further research, Ref. [6] undertook a thorough analysis of the literature. The study's primary goals were to assess the state of the art and the effects of AI on the fashion e-commerce industry, comprehend how AI could boost business profitability, and pinpoint areas in need of further research. The study provided an objective and thorough analysis of the body of literature by using the narrative literature review (NLR) approach. AI technology is being used by e-commerce behemoths to streamline their platforms and increase their competitiveness. New AI approaches and technologies in the fashion e-commerce space are encouraging and backed by continuous research advancements. In addition to leveraging deep neural networks and feature engineering with an auto-encoder and a deep clustering algorithm, Ref. [7] presented a novel approach to consumer segmentation. Based on their supermarket purchasing habits, four unique client categories were found by the deep clustering method. Cluster 0 consumers primarily purchase fresh food, groceries, and daily necessities at the supermarket, which they use in place of traditional wet markets. Customers in Cluster 1 spend primarily on clothing, accessories, and makeup; they do not purchase groceries or fresh food.

Customer segmentation is the process of breaking down a client base into groups based on shared traits, habits, or requirements. This may depend on a number of variables, including value to the business, purchasing patterns, demography, and location. Segmentation in the context of supply chain management enables companies to better inventory

management [8], optimize resource allocation [9], increase customer service [10] and improve demand forecasting [11]. Demands for product types, volumes, and delivery schedules vary widely throughout consumer sectors. Businesses could minimize holding costs and optimize inventory levels by comprehending these distinctions. It could improve customer happiness and loyalty by offering more personalized services through segmentation, such as tailored delivery options or targeted discounts. Businesses could more effectively devote resources to the client categories with the greatest revenue generation or growth potential by identifying high-value consumers. Demand patterns unique to a certain segment could improve forecasts and lower the chance of stockouts or overstocking. Ref. [11] also talked about challenges associated with the customer segmentation. He opined the significant role of ethics. Businesses must make sure that their segmentation tactics do not result in unfair business practices or invasions of client privacy. All segmentation initiatives should uphold transparency and equity. Businesses must make sure that their procedures for gathering and analyzing data are reliable and accurate. Furthermore, segmentation criteria need to be evaluated and changed frequently to account for shifting consumer behavior and market dynamics.

PRESENT WORK

In this paper, Machine Learning is used for segmentation of online retail customer dataset [12]. The segmentation was done on the basis of Geographical locations, Recency, Frequency and Monetary Value. First of all data was collected and then it was prepared for segmentation exercise. Then, segmentation was performed using k-means clustering algorithm. Different graphs were produced to visualize the segments. Finally, evaluation of the work was done. Python was used as a programming language to perform all the tasks [13]. The flow of work is being shown in the Fig. 1.

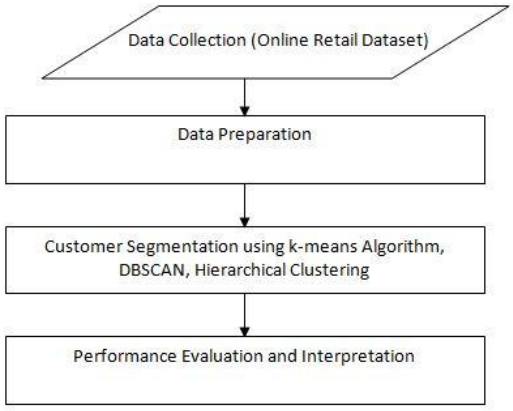

Fig. 1. Work Flow

Online retail dataset was read in a data frame named data. It contains eight columns viz. Invoice No, Stock Code, Description, Quantity, Invoice Date, Unit Price, Customer ID and Country. Sample dataset has been shown in Fig. 2.

Fig. 2. Sample dataset

Then the data was prepared for the customer segmentation. The null values present in the dataset were dropped, duplicate values were removed and only those rows were kept in which value of quantity is more than zero. Then the data set was classified on the basis of customers' country and a graph was plotted to demonstrate the same.

The k-means, DBSCAN and Hierarchical clustering algorithms were applied to form the segments of the customers based on Recency, Frequency and Monetary value of the transactions. K-means clustering algorithm is a type of unsupervised machine learning algorithm. The process of training an algorithm to work on unlabeled, unclassified data without human oversight is known as unsupervised machine learning. The machine's task in this scenario is to arrange unsorted data based on parallels, patterns, and variances without any prior data training. K stands for clustering [14], which divides data points into K clusters based on how far apart they are from each other's centers. The cluster centroid in the space is first randomly assigned. Next, each data point is assigned to a cluster according to how far it is from the cluster centroid. Following the assignment of each point to a cluster, new cluster centroids are designated. Iteratively, this procedure continues until it finds a good cluster. Elbow method is being used to identify the number of clusters. Dense areas in the data space are called clusters, and they are divided by areas with a lower point density. This logical concept of "clusters" and "noise" serves as the foundation for the DBSCAN algorithm. The main principle is that there must be a minimum number of points in the vicinity of a particular radius for every point in a cluster [15]. Known alternatively as hierarchical cluster analysis, or HCA, hierarchical clustering is another unsupervised machine learning approach that groups the unlabeled datasets into clusters. The dendrogram is the structure created by this algorithm, which develops the hierarchy of clusters like a tree [16]. For performance metrics 'Silhouette analysis' was utilized. The score ranges from -1 to 1. The score close to 1 is considered as good score for clustering.

RESULTS AND DISCUSSIONS

First of all the geographical segmentation was performed giving us the following results.

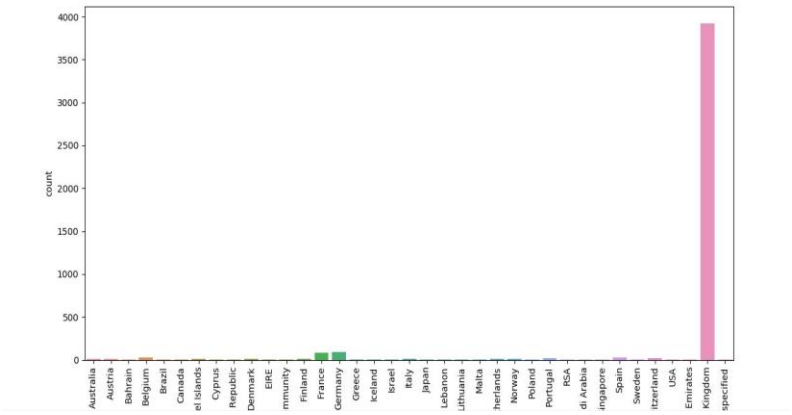

Fig. 3. Customer Segmentation on the basis of geographical locations (Country)

Then, clustering was performed using k-means, DBSCAN and Hierarchical clustering algorithms. The silhouette scores of these algorithms were compared by plotting following graph.

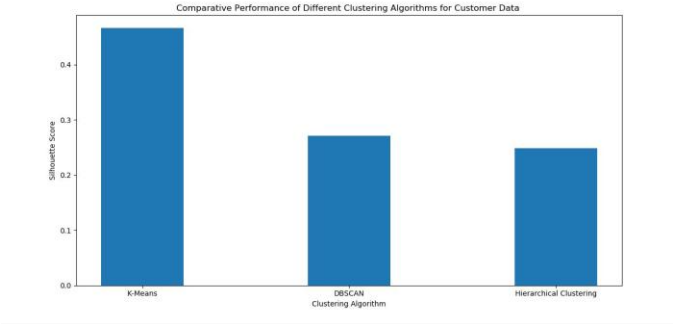

Fig. 4.Comparative Performance of Different Clustering Algorithms for Customer Data

It can be seen from the above figure that K-Means clustering algorithm outperformed DBSCAN and Hierarchical clustering algorithms on the basis of Silhouette Score. Then, clusters obtained by K-Means algorithm on the basis of recency, frequency and monetary value were plotted as shown in Fig. 5, 6 and 7 and then results were analyzed

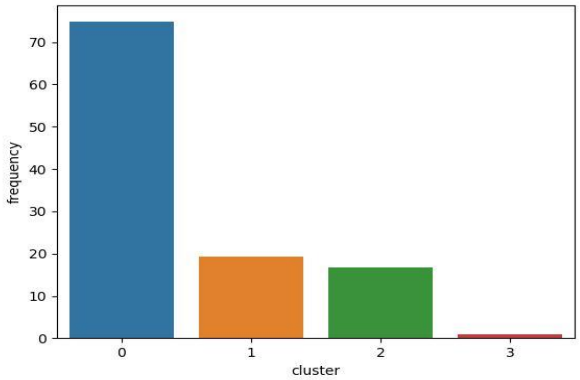

Fig. 5. Customer Segmentation on the basis of Frequency

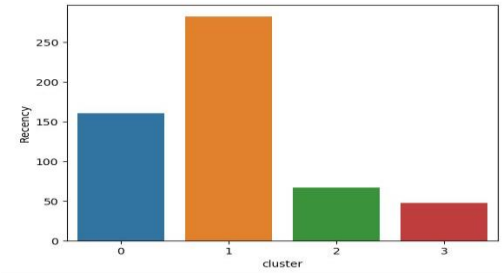

Fig. 6. Customer Segmentation on the basis of Recency

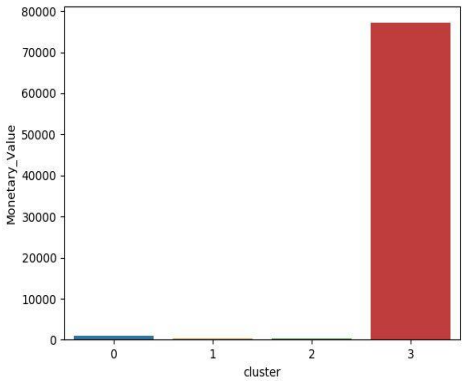

Fig. 7. Customer Segmentation on the basis of Monetary value

The implementation of k-means algorithm on this dataset scored 0.4662824029069976 in Silhouette analysis, which is a reasonable score as per previous discussion. As it can be seen from the results, the majority of the customers for this online retail store are from United Kingdom. Further, customers have been classified into 4 clusters. Cluster 0 represents those customers, who are frequently and recently visiting the store but with minimum spending. Cluster 1 represents the group of customers who recently visited the store with moderate frequency and minimum spending. Cluster 2 represents minimum spending, moderate in recency and frequency customers; and Cluster 3 represents those customers who are not frequent, not recently visited the store but who have done heavy spending.

CONCLUSION

Businesses may improve customer happiness, cut expenses, and optimize their supply chain operations by comprehending and meeting the specific needs of various consumer segments. So we can think of customer segmentation as one of the effective strategies for enhancing supply chain performance. Although segmentation techniques need to be carefully planned and carried out, the advantages greatly exceed the difficulties. Businesses, which use customer segmentation, well will be in a better position to compete in the fast-paced market of today. The paper compared the performance of three

clustering algorithms viz. K-Means, DBSCAN and Hierarchical clustering on the online retail dataset. Clearly, on the basis of performance metrics i.e. Silhouette Score in this case, K-Means outperformed other two clustering algorithms. Applying K-Means For the online retail store, we got 4 clusters of customers having different values for attributes like recency, frequency and monetary value. The management may look for making strategies based on this analysis. For cluster 1, management of online store may make strategy for more customer engagement and enhance shopping experience. Cluster 0 customers may be target for new products. Management may be interested in finding the reasons for cluster 2 customers' response and may create customized market plans to encourage them to purchase from the store. Cluster 3 customers are identified as heavy spending but low and not very recent visit to store. The management may introduce unique heavy value products as per their liking. Further, special attention may be provided to these customers.

REFERENCES

[1]. Qiu, Yujuan & Wang, Jianxiong. (2024). A Machine Learning Approach to Credit Card Customer Segmentation for Economic Stability. 10.4108/eai.27-10-2023.2342007.

[2]. Saxena, Arpit & Agarwal, Ashi & Pandey, Binay & Pandey, Digvijay. (2024). Examination of the Criticality of Customer Segmentation Using Unsupervised Learning Methods. Circular Economy and Sustainability. 1-14. 10.1007/s43615-023-00336-4.

[3]. Sarkar, M., Puja, A. R., & Chowdhury, F. R. (2024). Optimizing Marketing Strategies with RFM Method and K-Means Clustering-Based AI Customer Segmentation Analysis. *Journal of Business and Management Studies*, *6*(2), 54-60.

[4]. Talaat, F. M., Aljadani, A., Alharthi, B., Farsi, M. A., Badawy, M., & Elhosseini, M. (2023). A mathematical model for customer segmentation leveraging deep learning, explainable AI, and RFM analysis in targeted marketing. *Mathematics*, *11*(18), 3930.

[5]. Kasem, M. S., Hamada, M., & Taj-Eddin, I. (2024). Customer profiling, segmentation, and sales prediction using AI in direct marketing. *Neural Computing and Applications*, *36*(9), 4995-5005.

[6]. Goti, A., Querejeta-Lomas, L., Almeida, A., de la Puerta, J. G., & López-de-Ipiña, D. (2023). Artificial Intelligence in Business-to-Customer Fashion Retail: A Literature Review. *Mathematics*, *11*(13), 2943.

[7]. Nguyen, S. P. (2021). Deep customer segmentation with applications to a Vietnamese supermarkets' data. *Soft Computing*, *25*(12), 7785-7793.

[8]. Christopher, Martin, 2016. *Logistics & Supply Chain Management*. 5th ed. Harlow, United Kingdom: Pearson Education.

[9]. Wang, G., Gunasekaran, A., Ngai, E. W., & Papadopoulos, T. (2016). Big data analytics in logistics and supply chain management: Certain investigations for research and applications. *International journal of production economics*, *176*, 98-110.

[10]. Chopra, Sunil and Meindl, Peter. 2016. *Supply Chain Management : Strategy, Planning, and Operation*. 6th ed. Boston, Mass.: Pearson.

[11]. Binns, R. (2018, January). Fairness in machine learning: Lessons from political philosophy. In *Conference on fairness, accountability and transparency* (pp. 149-159). PMLR.

[12]. Online Retail. (2015). UCI Machine Learning Repository. https://doi.org/10.24432/C5BW33.

[13]. https://www.python.org/doc/essays/blurb/. Python Software Foundation. 2023. "What Is Python? Executive Summary." Python. 2023

[14]. GeeksforGeeks. "K Means Clustering - Introduction - GeeksforGeeks." *GeeksforGeeks*, 30 May 2019, www.geeksforgeeks.org/k-means-clustering-introduction/.

[15]. Dey, Debomit. "DBSCAN Clustering in ML | Density Based Clustering." GeeksforGeeks, 6 May 2019, www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/.

[16]. "Hierarchical Clustering in Machine Learning - Javatpoint." *Www.javatpoint.com*, www.javatpoint.com/hierarchical-clustering-in-machine-learning.