Approaches to Customer Segmentation:

A Comprehensive Review of Retail Applications

Chiranjib Muduli, Subrat Dash, Aayushi Awasthi, Eric Minz, Souravi Das, Shivshakti Vashist, Diya Manna

Group 8, DMDW CS_6 (Dr. Suchismita Das)

Department of Computer Engineering, KIIT University

INTRODUCTION

The digital revolution has significantly impacted the retail sector, with the rise in big data analysis to comprehend customers' needs. The emergence of supermarkets and hypermarkets all over the world has led to the demand of effective customer segmentation for all the retail industries to effectively sustain their business. In the past, segmentation was primarily based on simple demographic information. However, the advent of big data and machine learning has introduced more advanced techniques. These new methods enable businesses to utilize extensive customer data, such as purchase patterns, browsing activities, and social media engagement, to create more precise and useful segments.

People have diverse needs, depending on their social background, the needs are different based on age, gender, income etc. Identifying unique customer groups and tailoring their requirements is essential for every store. The Customer Relationship Management (CRM) department segments customers to better understand them and enhance sales and the effectiveness of marketing strategies. The more precise the customer segmentation, the more successful the corresponding marketing strategies and personalized advertisement campaigns will be for the business.[1]

Sales transaction data can be segmented for better perceptions of what such customer segments are like. It allows companies to tailor their marketing strategies accordingly by classifying their customers according to their purchase behaviour. Clustering is among the frequently utilized methods for this, which arranges data based on the resemblance of different clients. Apart from clustering, Market Basket Analysis (MBA) making use of Association Rule Mining is also employed. MBA identifies trends in products bought from sales transaction data. The main aim is to enable retailers have an insight on how customers behave so they can make informative decisions and enhance their marketing techniques [2][3][4]. In addition to clustering and MBA, the most widely applicable input variables are Lifetime Value (LTV), Recency, Frequency, and Monetary (RFM) [5]. RFM model is a statistical method, that has been employed in many practical scenarios which includes financial and nonprofit organizations such as banking industry [6], on-line industries [7], and telecommunication industries [8].

Despite its promise for customer segmentation, machine learning faces several challenges including algorithm selection and results interpretation. The type of algorithm affects the quality of segmentation substantially with different algorithms yielding different results for similar data sets. Some models such as deep learning in machine learning are very dark boxes making it difficult for companies to know how segmentation decisions were taken. This review will examine recent research on use of machine learning in retail customer segmentation focusing on methods employed, results obtained, and obstacles encountered.

METHODOLOGY

k-MEANS

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean (cluster centres or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. [9]

The algorithm starts with selecting k centroids randomly, followed by an iterative process of assigning each data point to the nearest centroid and recalculating the centroids based on the current cluster members. This process continues until the centroids stabilize, and the cluster assignments no longer change.

Integration of K-means Clustering and RFM Model
Often in market segmentation, the K-means clustering is combined with RFM (Recency, Frequency and Monetary model) to give rise to more precise segments of customers. Within the RFM model, recency of last purchase, frequency of purchases and amount spent are put into consideration when evaluating customers. The RFM model is based on three quantitative factors:

1. **Recency**: How recently a customer has made a purchase
2. **Frequency**: How often a customer makes a purchase
3. **Monetary value**: How much money a customer spends on purchases

From these factors, an RFM score can be arrived at which then gives way for clustering.

For instance, in the research done by Maraghi et al in 2022 combined K-means algorithm with RFM model so that they could classify their customers into eight diverse groups. The authors adjusted the RFM variables and clustered their data based on these scores. The number of clusters was established through previous studies while Self-Organizing Map (SOM) technique served as a validation method. Afterwards, Average Silhouette Score helped assess how well each data point fits into its cluster.[1]

Likewise, Turkmen (2022) employed the K-means clustering technique to classify online retail customers. Using elbow method, they identified five clusters and used Silhouette scores to evaluate customer distribution across various levels of income. Each consumer category was thus managed in such a way that pricing strategies could be tailored for them allowing adverts to be geared towards certain shopping behaviours among other things.[10]

DBSCAN and Hierarchical Clustering

DBSCAN (Density-based spatial clustering of applications with noise) is a non-parametric clustering algorithm introduced by Ester, Kriegel, Sander, and Xu in 1996. It groups closely packed points while marking isolated points in low-density regions as outliers. DBSCAN is widely used and frequently cited in clustering applications.[11]. Unlike k-means, DBSCAN does not require the number of clusters to be specified beforehand. Instead, it identifies clusters based on the density of data points, which makes it particularly suitable for datasets with varying densities and complex shapes.

In the paper by Monalisa et al. (2023), the DBSCAN algorithm is utilized for customer segmentation by clustering customers based on their Recency, Frequency, and Monetary (RFM) values. The study determined the optimal epsilon and MinPts values using a k-dist graph, leading to the identification of

distinct customer clusters. The DBSCAN results were further validated to assess the quality and number of clusters, ensuring the algorithm's effectiveness in the segmentation process.[12]

Abdulhafedh A. (2021) applied hierarchical clustering to group customers with similar traits using the Agglomerative hierarchical clustering method. They created dendrograms to see how different linkage techniques—single, complete, and average—shape cluster formation. The dendrograms show the cluster structure and help decide the number of clusters by finding where to slice the dendrogram. They worked out the space between clusters with the Euclidean distance measure, and standardized the variables to make sure they could compare things on different scales.[13]

FUZZY C MEANS

Fuzzy c-means (FCM) clustering, developed by J.C. Dunn in 1973 and improved by J.C. Bezdek in 1981, is like k-means but allows data points to belong to multiple clusters with varying degrees. The algorithm involves choosing several clusters, assigning random coefficients to each data point, and iteratively refining these coefficients and centroids until convergence.[14]

Snehalatha et al. (2023) put forward a mixed method that brings together DBSCAN and Fuzzy C-Means (FCM) to group e-commerce customers. DBSCAN finds tight clusters and deals with outliers, while FCM fine-tunes these groups by giving fuzzy membership scores to customers. This combined model boosts clustering precision and tackles complex messy data well. The research reveals better results than old-school methods, with measures like the Silhouette Index backing up how well it works. Looking ahead, the authors suggest adding the model to changing systems to get real-time insights.

Rusdiana et al. (2021) test how well Fuzzy C-Means (FCM) works for grouping customers. They do this by looking at different ways to measure distance such as Euclidean, Manhattan, Chebyshev, and Minkowski. The research shows that the way you measure distance has a big effect on how good the grouping is. They found that Manhattan and Minkowski work the best when you try different numbers of groups. To check how well it's working, they use things like Partition Coefficient and RMSE. The paper points out that it's important to pick the right way to measure distance if you want FCM to work well when grouping customers.

Features	**Execution Time**	**Accuracy of Clustering**	**Number of Clusters Used**
K-means	Less	Less	Large
Fuzzy C-Means	More	More	Large and Small

OUR METHODOLOGY:

We're going to use the K-means clustering algorithm along with the Apriori algorithm to improve how we group customers and look at what they buy together in the Online Retail dataset. To start, K-means will put customers into groups based on how they shop looking at things like how much they buy and spend from the data we have. After that, we'll use the Apriori algorithm to find products that are often bought together in each group and figure out the connections between them. By using these two methods together, we'll get a better understanding of how customers behave and what they tend to buy.

CONCLUSION

In this comprehensive discussion of various customer segmentation methods, we have explored the history of techniques used in retail, beginning with the transition from old-fashioned demographic-based methods to up-to-date machine learning algorithms. The combination of big data and advanced analytics has brought a revolution in consumer segmentation, which has given more accurate insights into customer behaviour.

Our review began by describing the traditional ways to carry out customer segmentation, that is the demographic profiling and the RFM clustering method. Although these strategies provided the base requirements for understanding, they often missed the granularity necessary for modern retail use cases. Incorporation of clustering algorithms, such as k-means and hierarchical clustering, devised in the RFM data also prolonged the ability to create more exact customer segments. For example, k-means clustering combined with RFM variables has demonstrated in the segment accuracy significant success through the incorporation of recency in computing clustering processes.

We have also discussed the use of the more advanced clustering algorithms such as DBSCAN and Fuzzy C-Means. DBSCAN's ability to identify clusters that may have various shapes and size and without use of predefined clusters has now been quite useful in detecting consumer sectors with complex behaviours. FCM's adaptability which lets the data points to belong to more than one cluster with different degrees of membership has also made the segmentation more accurate by addressing the shortcomings of hard boundaries among clusters.

The review outlines the ways that by combining k-means and market basket analysis (MBA) with FCM, a deeper understanding of customer behaviour and purchase patterns can be achieved. Integrating these methods does not only improve the accuracy of segmentation but also offers actionable insights for personalized marketing. For instance, a combination of the Apriori algorithm and k-means to evaluate purchasing patterns can result in more precise targeting and better marketing campaigns.

Even though some progress has been made, challenges are still faced when it comes to algorithm choices and the interpretation of the results. Unlike the case when you use different models in similar datasets, these models can give you dissimilar outputs. In addition, the decision-making process in models, like deep learning, is usually unclear. The forthcoming research should center on the production of models of higher transparency and interpretability, as well as on the implementation of real-time analytics to respond to the fluctuations in customer behaviour.

Globally, there has been a positive trend in customer segmentation technology's popularity, which has led retailers to brandish more sophisticated instruments for their customer data. This not only increases the efficiency of the marketing department but also provides more individualized customer experiences. As the technology continues to advance, the prevalence of innovative analytical approaches will strengthen and bring customer segmentation closer to the best practices, thus, the anger of the retail sector will be minimized.

REFERENCES

[1] Maraghi, M., Adibi, M. A., & Mehdizadeh, E. (2020). Using RFM model and market basket analysis for segmenting customers and assigning marketing strategies to resulted segments. *Faculty of Industrial and Mechanical Engineering, Islamic Azad University, Qazvin Branch*.

[2] Auliasari, K., & Kertaningtyas, M. (2019). Penerapan algoritma K-Means untuk segmentasi konsumen menggunakan R. *Jurnal Teknologi & Manajemen Informatika, 5*(1).

[3] Marwazia Shaliha, K., Angelyna, A., Nugraha, A. A., Wahisyam, M. H., & Kurnia Sandi, T. (2021). Implementasi K-Means clustering pada online retail berdasarkan recency, frequency, dan monetary (Implementation of K-Means clustering in online retail based on recency, frequency, and monetary). *Gunung Djati Conference Series, 3*. Retrieved from https://conferences.uinsgd.ac.id/gdcs

[4] Kaur, M., & Kang, S. (2016). Market basket analysis: Identify the changing trends of market data using association rule mining. *Procedia Computer Science*, *78*, 78–85. https://doi.org/10.1016/j.procs.2016.05.180

[5] UCI Machine Learning Repository. (2020). *Online retail data set*. Archive.ics.uci.edu. Retrieved from https://archive.ics.uci.edu/ml/datasets/Online+Retail

[6] Cheng, Y., & Chen, H. (2009). Classifying the segmentation of customer value via RFM model and RS theory. *Expert Systems with Applications, 36*(3), 4176–4184. https://doi.org/10.1016/j.eswa.2008.04.003

[7] McCarty, J., & Hastak, M. (2007). Segmentation approaches in data mining: A comparison of RFM, CHAID, and logistic regression. *Journal of Business Research, 60*(6), 656–662. https://doi.org/10.1016/j.jbusres.2006.06.015

[8] Li, Y., Lin, C., & Lai, K. (2010). Identifying influential reviewers for word-of-mouth marketing. *Electronic Commerce Research and Applications, 9*(4), 294–304. https://doi.org/10.1016/j.elerap.2010.02.004

[9] https://en.wikipedia.org/wiki/K-means_clustering#:~:text=k%2Dmeans%20clustering%20is%20a,a%20prototype%20of%20the%20cluster.

[10] Turkmen, B. (2022). Customer segmentation with machine learning for online retail industry. *The European Journal of Social and Behavioural Sciences, 31*(2). https://doi.org/10.15405/ejsbs.316

[11]https://en.wikipedia.org/wiki/DBSCAN#:~:text=Density%2Dbased%20spatial%20clustering%20of,and%20Xiaowei%20Xu%20in%201996.

[12] Monalisa, S., Juniarti, Y., Saputra, E., Muttakin, F., & Ahsyar, T. K. (2023). Customer segmentation with RFM models and demographic variable using DBSCAN algorithm. *TELKOMNIKA Telecommunication Computing Electronics and Control, 21*(4), 742-749. https://doi.org/10.12928/TELKOMNIKA.v21i4.22759

[13] Abdulhafedh, A. (2021). Incorporating K-means, Hierarchical Clustering, and PCA in Customer Segmentation. *University of Missouri, USA*.

[14] https://en.wikipedia.org/wiki/Fuzzy_clustering

[15] Rusdiana, U., Falih, N., Ernawati, I., & Arista, A. (2021). Comparison of distance metrics on fuzzy c-means algorithm through customer segmentation. *Proceedings of the 2021 International Conference on Informatics, Multimedia,Cyber and Information System (ICIMCIS)*. Retrieved from https://www.upnvj.ac.id

[16] Snehalatha, N., Kumar, S. M., & Kachroo, V. (2023). Customer segmentation and profiling for e-commerce using DBSCAN and fuzzy c-means. *Proceedings on Engineering Sciences, 05*(3), 539-544. https://doi.org/10.24874/PES05.03.016