IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Advance Customer Segmentation

[1]Swanand Kulkarni, [2]Vipul Lokhande, [3]Nishant Bhalerao, [4]Yash Tajane, [5]Jyoti Kharat

[1-4]Dept. of Computer Engineering, JSPM's Narhe Technical Campus, Pune, Maharashtra, India

[5]Assistant Professor, Dept. of Computer Engineering, JSPM's Narhe Technical Campus, Pune, Maharashtra, India

Abstract: In today's world, the number of competitors in the E-commerce market has grown, it is only inevitable that there would be fierce rivalry. Organizations must analyze their consumers in order to stay competitive and operate at their best. Customers may also leave customer reviews, which allow them to express their personal thoughts on the product. Customers have the ability to offer their opinions about the product.

Thus, most companies should have sufficient market knowledge to forecast or predict which consumer segments are their most productive, the individuals of those businesses realize that scaling the job is not best left to guesswork or instinct. It is also important to understand how each segment will be targeted and what they need from the organization.

Segmentation is a process of dividing customers into groups based on characteristics such as age, income level or gender. Segmentation can be done by using different methods like demographic, psychographics and behavioral factors. This paper analyses 3 prominent clustering algorithms (K-means, Agglomerative and Meanshift) to identify the target audience for segmentation. These algorithms are used in order to identify potential customers are likely to buy the product or service.

Index Terms **– Clustering, K-means algorithm, Agglomerative algorithm, Meanshift algorithm.**

I. INTRODUCTION

Business-customer connections have become increasingly crucial as technology has advanced. Managing this link is crucial for the company's future success. Customer Relationship Management is the process of managing relationships between businesses and their customers (CRM) [1]. CRM plays an important function in the corporate world. Companies, on the other hand, can determine the customer's habits, features, and so on, as well as who the most lucrative customers are. Customer segmentation is the division of a huge customer database into smaller groups. Members of the sub-parts share traits that are distinct from those of members of other sub-parts. Customer segmentation is based on demographic factors such as age, gender, religion, family size, and more.

Behavioral parameters include clustering on the basis of recency and frequency of purchases. Recency is how recent was the last purchase of customer and frequency is how often the purchase happens. Despite the simplicity of these parameters, the clustering on this basis gives classes of customers which can be then handled differently leading to boost sales. According to the Pareto's rule, only 20% of the customers contribute to 80% of the sales of the organization. So as per [3], the best as well as the weakest cluster obtained from behavioral clustering, can be targeted differently to gain maximum profit to the businesses.

Automated merchandising is one of the most effective techniques to target customers with marketing efforts. This notion entails giving clients with relevant and personalized recommendations, which may be accomplished with the help of recommender systems. Recommender systems are divided into three categories: content-based, collaborative, and hybrid. In content-based recommender systems, the goods that are recommended to a user are comparable to those that the user has previously purchased or is actively researching. Similarities between users are detected in collaborative recommender systems based on their purchases and preferences.

The following is a breakdown of the paper's structure. An exhaustive literature survey outlines all related work in this subject in Section II. In Section III, the suggested approach is shown and explained and the results are provided in Section IV. Finally, Section VI concludes the paper

II. RELATED WORK

A significant number of papers have been examined in relation to segmenting clients in various sectors. A wide range of approaches and objectives have been identified. Methods for producing segmentation predictions appear to be based on the assistance of several technologies. The related work for customer segmentation using different clustering is presented here.

For a variety of purposes, several researchers employed the consumer segmentation approach. The banking industry's top aim is to increase customer segmentation and incorporate it into the creation and marketing of new products. In this study [3] introduce

the loyalty program, which comprises the issuing of several sorts of cards for such clients. The number of people using the internet to bank is continuously increasing. Customer segmentation may be done using Internet banking data. Create clustering models based on customer profile data and their use of XYZ bank's Internet Banking.

K-means Clustering:

Customer segmentation has been proved to be effective using clustering. Clustering is a type of unsupervised learning that involves the ability to detect clusters in unlabeled data. K-means is a straightforward unsupervised learning system. It is the most basic clustering technique based on the partitioning concept.

It classifies data using the Euclidian Distance technique. The data is divided into k sections using this procedure. The algorithm is sensitive to the initialization of the centroid position. The cluster requirements are calculated depending on the user's preferences The number of K (centroids) is calculated using the elbow method (discussed later), after which data points are assigned to the closest centroid, forming the cluster; after the cluster is formed, the centers are calculated using the mean and this process is repeated until there is no change in centroid position.

K-means will algorithm can be implemented as follows:

1) Determine the number of clusters, which is a precondition for k-means clustering.
2) Select the first centroids at random.
3) Find the Euclidean distance between each data point and the centroid, which is the square of the distance between each data point and the centroid.
4) Assign each data point to the cluster with the shortest Euclidean distance.
5) From the clusters found, calculate the new centroids.
6) Iterate until you have the appropriate number of ideal clusters depending on recency and frequency. [4] [5]

Agglomerative Clustering:

Besides K-means, Agglomerative Clustering is among the most commonly used clustering algorithms, because its conceptual simplicity and its low computational complexity. Their biggest disadvantage is their slowness.

Agglomerative Clustering is based on the formation of a hierarchical structure represented by dendrograms. The dendrogram serves as a memory for the algorithm, allowing it to tell how clusters are generated. The clustering process begins with the formation of N clusters for N data points, followed by the merging of the nearest data points in each step, such that the current step has one fewer cluster than the previous one. [6]

Mean shift clustering:

This non-parametric iterative clustering approach works by treating all data points in the feature space as empirical probability density functions. The algorithm clusters each data point by enabling them to converge to an area of local maximum, which is accomplished by establishing a window around each data point, calculating the mean, then adjusting the window to the mean, and repeating the steps until all of the data points converge, producing clusters.

Elbow method:

Elbow method is used for finding optimal value of K for K-means clustering algorithm. This method works by finding the SSE (Sum of squared error) of each data point with its nearest centroid with different values of K. As value of K increases the SSE will decrease and at a particular value of K where there is most decline in the SSE is the elbow, the point at which we should stop dividing data further.

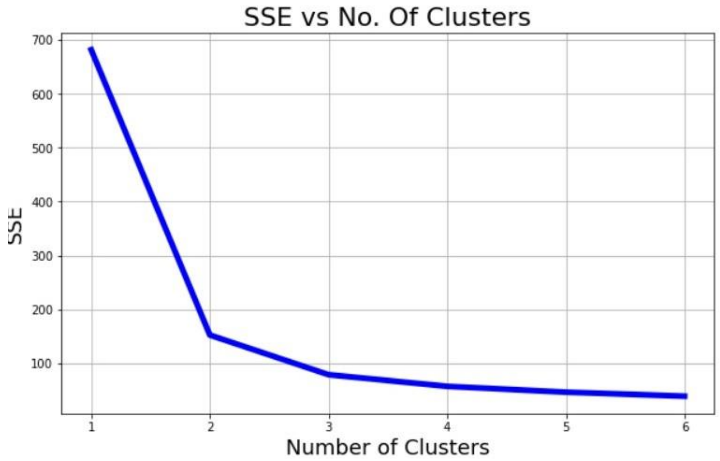

Fig 1: Graph for values of SSE vs Number of cluster

Clustering is a part of unsupervised learning. Artificial intelligence, bioinformatics, pattern recognition, segmentation, and machine learning are just a few of the domains where it may be used. On the basis of the scenarios based on accuracy and efficiency, the best clustering method should be chosen. A similar customer segmentation was implemented on anonymized Instacart Grocery Shopping Dataset 2017 containing a purchase information of over 3 million orders of grocery from more than 200,000 users of Instacart. [7]

Cluster No. .	Cluster (Category)	Description
1	Most profitable	Customers who have just purchased and who purchase frequently. As a result, they have a strong relationship with the company and may be described as loyal and profitable clients.
2	Potential Loyalist	Customers who buy from Instamart on a regular basis. If appropriate goods are marketed to this cluster, it may be able to join them most profitable cluster.
3	At Risk	Customers who have not purchased in a long time and do not purchase frequently. This group might include even first-time purchasers.

TABLE 1: Description of clusters formed

A Recommender System was also developed after the analysis done using above method. The effectiveness of recommendations is measured in two ways: if the suggested product was added to the user's basket or whether the product was seen by the consumer after being recommended. A customer's session can be retained if the former strategy is used. The client is not obligated to click on the recommended product immediately after it is suggested.
It can be regarded successful targeting if the product IDs of the products ultimately purchased by the consumer match the product ids of the products advised.

Limitations:

The dataset's data was gathered using a manual manner. The needed data, such as the time of purchase, product details, aisle, and department information, must be acquired directly from the customer's billing information. This, on the other hand, may be both laborious and incorrect. As a result, the study has a restriction. The inclusion of electronics for data collecting, however, can overcome this constraint.

Comparison between algorithms:

The algorithms can be evaluated using *Silhouette Score*. [2]

Silhouette Score:

It's a metric for determining how successfully a data item was clustered into the right cluster.

First Step:

a = Average distance between a cluster's centroid and the data points entangled inside it.

Second Step:

b = Average distance between the data point and the cluster data points nearest to it.

Step three:

Silhouette Score = $(b\text{-}a)/\max(b,a)$

For a data point to be firmly grounded in its cluster, 'b' must be large and 'a' must be small, with as little variation as possible between the two. To normalize the silhouette score 'max (b, a)' is added. The higher the score, the more likely the data point is part of that cluster.

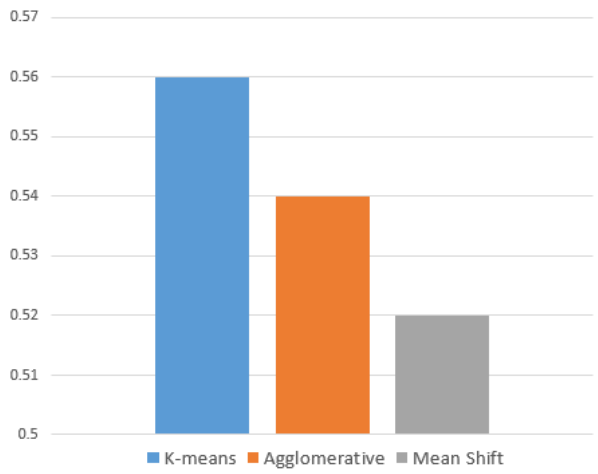

Fig 2: Comparison of Silhouette Score

III. METHODOLOGY

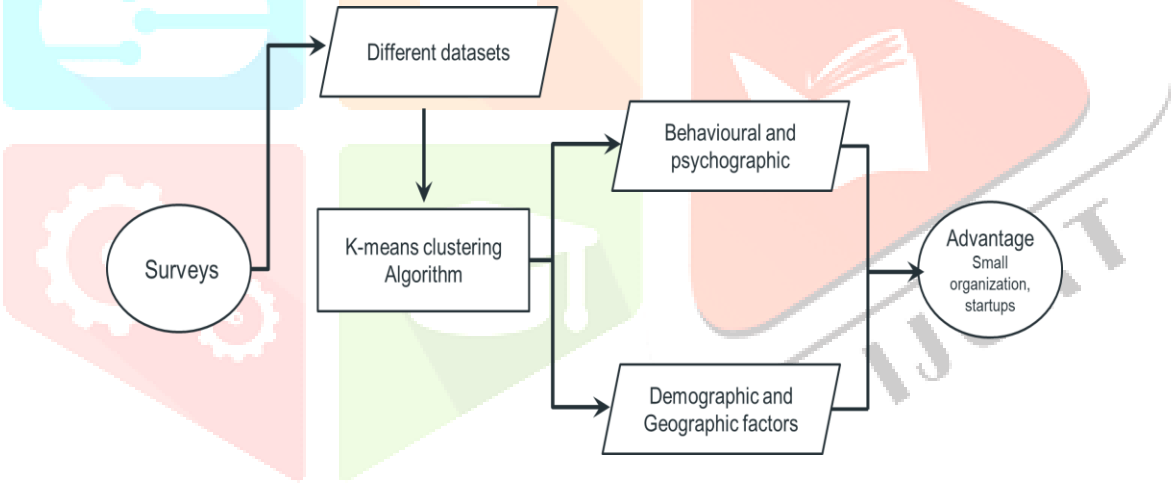

Fig 3: Proposed Methodology for Advance Customer Segmentation

As previously stated, current segmentation is limited. It is mostly based on demographic and geographic variables. Advance segmentation is a type of segmentation that is more in-depth and based on psychographic and behavioural characteristics. It is more effective and gives firms with a more tailored strategy, giving them a competitive advantage. Advanced segmentation is based on the principle of focusing on a certain sector of a company's market.

Advance segmentation can widely be used in the commercial world, rather than typical consumer segmentation, to obtain additional benefits. It takes a standard segmentation result and refines it using machine learning tools and artificial intelligence. Algorithms with AI (Artificial Intelligence) enhancements, such as RFM Analyses and segmentation for a specific goal. Advance segmentation is a more ground level kind of segmentation in which clients are divided into several independent parts. In the fields of information technology, business, and marketing, as well as other fields.

Advance segmentation is a technique that can be employed. Its findings might help businesses better understand the demands of their target clients and supply each of their target clients with offers, new goods, and effective services

Behavioural Segmentation:

The process of segmenting a customer dataset into micro-segments based on their activities is known as behaviour segmentation. This method of segmentation may be applied to a variety of situations.

Researchers are looking for methods to utilize this behaviour segmentation approach to help firms grow. Policyholders have distinct features in the insurance market, particularly throughout the claims procedure. The behaviour segmentation approach was

employed in this study [1] to segment the health insurance claim dataset. Consider the claiming dates, claiming frequencies, and claiming fees, as well as different elements pertaining to each policyholder in the insurance company.

Based on the policyholders' gender, age, and sickness status, this can assist determine the best policyholders, faithful policyholders, high spender policyholders, almost lost policyholders, lost policyholders, and worst policyholders. Health insurance firms can use this information to pay more attention to certain consumers and make strategic judgments about how to handle their claims habits.

IV. RESULTS AND DISCUSSION

Rather than targeting a whole industry, most organisations employ demographic segmentation to find the distinctive resources in the target group in a given sector. Demographic consumer segmentation is a useful tool for intelligent advertising since it allows people to classify based on location.

If such segmentation is combined with psychographic and behavioral characteristics, then it can be used to identify the segments that are most likely to respond positively, then the results can be used to identify potential customers and tailor marketing messages accordingly

K-means clustering:

After comparing 3 algorithms- K-means clustering, Agglomerative clustering and Mean shift clustering, K-means algorithm was found out to be the most efficient of three. It also depends on the type of data we are dealing with. It will also depend on the scenario in which we are applying the algorithm. K-means can be used to segment large datasets. Following table explains different characteristics of K-means clustering algorithm.

Parameters	**K-Means Clustering**
Time Complexity	$O(n^2)$
Speed	Faster for unlabeled dataset
Accuracy	More Accurate
Dataset	Suitable for medium and big datasets.

V. FUTURE SCOPE

Many new innovations are developed as time passes, yet the fundamentals remain the same regardless of whatever new inventions are made. The notion of customer segmentation is similar. There may be different sophisticated approaches for performing Customer Segmentation in the future, but this practise will continue. Companies have been gathering consumer data for years in order to better understand who their customers are and how they interacted with their brand. When we integrate client demographics with behavioural data, we can figure out precisely what our customers desire.

A corporation may make their consumers feel appreciated and acquire their trust by collecting ongoing feedback. Companies should keep the data they create in a Data Warehouse, which will aid them in making strategic decisions in the future. As long as a firm takes this strategy, it can be assured that its future is in excellent hands, and that it will continue to grow its market share and client base over time. Thus, as future work, we would like to carry out experimentation for real-world examples where a upcoming business can be benefitted by such model and can lead organization or business to make better decisions.

VI. CONCLUSION

For E-commerce businesses, customer satisfaction is critical to their success. When businesses do adequate analysis, they may accomplish it. The behaviour of clients in the market is monitored as part of this investigation. Customer segmentation is one of the most important methods that may assist a company in determining the loyalty of its existing clients. In comparison to the acquisition of new consumers, it is less expensive for a firm to maintain its present client base.

This approach aids businesses in making better decisions, allowing them to expand their consumer base while also ensuring customer satisfaction. E-commerce companies also give a venue for their clients to provide feedback in the form of reviews and complaints. This project is critical because if clients are dissatisfied with the service offered by a company, they may switch to one of the many other businesses available in the market. Customer sentiment analysis aids firms in providing better service. The company will only succeed if the customers are happy. As a result, we can claim that customer segmentation is one of the most important components for every business.

VII. ACKNOWLEDGMENT

We would like to express our gratitude to Prof. J. S. Kharat, our project guide, for her advice and assistance during the project. Her great advice and recommendations were quite beneficial.

REFERENCES

[1] E.Y.Nandapala, K.P.Jayasena,M.Rathnayaka, " Micro-Segmentation Approach for Health Insurance Industry", IEEE (International Conference on advancement in Computing Techniques) (10 Dec 2020)

.

[2] Kalyani Bhade, Vedanti Gulalkari, Nidhi Harwani, Sudhir N. Dhage ,**"**A Systematic Approach to Customer Segmentation and Buyer Targeting for Profit Maximization", 9th ICCCNT 2018 July 10-12, 2018, IISC, Bengaluru, India

[3] V. Mihova and V. Pavlov, "A customer segmentation approach in commercial banks," AIP Conference Proceedings, vol. 2025, no. October 2018, 2018.

[4] OJ Tanupriya Choudhury, Vivek Kumar, Darshika Nigam, "Intelligent Classification & Clustering Of Lung & Oral Cancer through Decision Tree & Genetic Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering,2015

[5] Tanupriya Choudhury, Vivek Kumar, Darshika Nigam, "An Innovative and Automatic Lung and Oral Cancer Classification Using Soft Computing Techniques", International Journal of Computer Science & Mobile Computing,2015

[6] Omar Kettani, Faycal Ramdani, Benaissa Tadili, "An Agglomerative Clustering Method for Large Data Sets", International Journal of Computer Applications (0975 – 8887) Volume 92 – No.14, April 2014

[7] Tushar Kansal, Suraj Bahuguna, Vishal Singh, TanupriyaChoudhury, "Customer Segmentation using K-means Clustering", 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)