IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

CUSTOMER SEGMENTATION FOR BUSINESS INTELLIGENCE USING K- MEANS CLUSTERING

[1]ANUBHAV SANKET, [2]SHIPRA SINHA, [3]YAMINI MANDAGIRI, [4]Dr. SHIVAPRASAD A C, [5]Mrs. SOWBHAGYA M P

[1]UG Student, [2]UG Student, [3]UG Student, [4]Asst. Professor (Dept. of AIML), [5]Asst. Professor (Dept. of AIML)

[1]DEPT. OF Artificial Intelligence and Machine Learning

[1]Dayananda Sagar Academy of Technology and Management

Abstract *The paper emphasizes the importance of customer segmentation in today's marketing landscape, where understanding and catering to individual customer needs are paramount. It highlights customer service management and customer relationship management as key elements in achieving revenue growth and profitability. The authors discuss the use of data mining and clustering algorithms, specifically the k-means clustering technique, to segment customers based on their value. By analysing customer data and clustering them into distinct groups, businesses can better target their marketing efforts and tailor their strategies to specific customer segments. The paper also introduces a unique aspect by considering the duration and trend of customer value changes. By examining the historical behaviour of customers, businesses can improve the accuracy of their forecasts and make more informed decisions based on past customer interactions.*

Keywords: *K-mean algorithm, Customer relationship management, Data mining, Customer segmentation, Business intelligence*

I. INTRODUCTION

Consumer segmentation involves dividing a market into distinct customer groups with similar characteristics. This strategy helps uncover unmet customer needs, allowing companies to differentiate themselves from competitors by offering unique products and services. As technology has advanced, business-customer relationships have become increasingly important. By understanding customer actions, preferences, and other factors, businesses can identify their most profitable customers.

Segmenting customers involves breaking down a large customer database into sub-parts based on various factors such as age, gender, religion, family size, buying habits, social choices, and lifestyle preferences. This segmentation can provide several benefits, including improved customer service, stronger customer relationships, increased efficiency, enhanced distribution channels, improved brand recognition, and optimized pricing.

Some studies advocate the use of k-means clustering in customer segmentation, particularly in the financial industry. As customer needs evolve, segmentation can help businesses in the telecommunications sector identify new services to meet those needs. Combining consumer segmentation with a k-means clustering algorithm can yield accurate segmentation results, making accuracy a crucial aspect of the process.

In machine learning, there are two types of designs: supervised and unsupervised learning. Businesses commonly segment their consumer base using demographic facts, which involve breaking down the population into subgroups based on statistics such as marital status, age, race, gender, work status, income, nationality, and political orientation. Geographic information is also used, ranging from specific cities for localized businesses to broader regions for larger companies.

Behavioural consumer categorization is based on observed customer behaviours, including preferred brands and spending patterns, which can be used to predict future events. Psychographic analysis explores emotional and cognitive factors influencing consumer behaviour, combining actions, interests, and psychographic data to understand why people behave in certain ways.

II. METHODOLOGY

A. Block Diagram

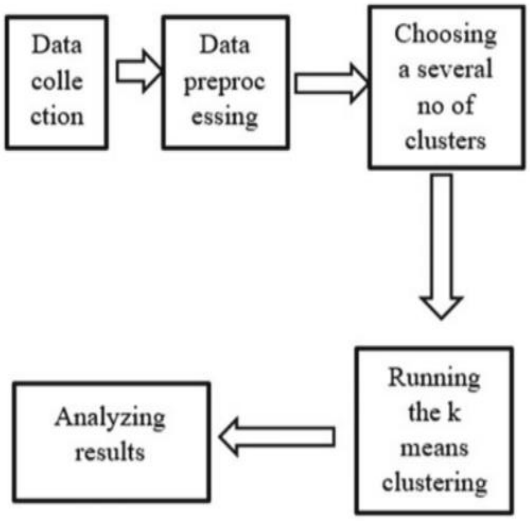

Figure. 1 Block Diagram

B. K-Means Clustering

The K-means method is indeed commonly used for clustering features. The process of K-means clustering involves several steps. Initially, the desired number of clusters, denoted as 'k,' needs to be determined. Then, k objects are randomly selected to form the initial clusters. The centroid, or mean, of each cluster is calculated by measuring the Euclidean distance between the data points and the centroid. The data points are then assigned to the nearest centroid.

After the initial clustering, the centroids are updated by recalculating the mean of the data points within each cluster. The distances are re-evaluated, and the data points are reassigned to the new centroids. These two steps of centroid updating and data reassignment are iteratively repeated until a stopping criterion is met, such as a maximum number of iterations or when the centroids no longer change significantly.

It's worth noting that this method relies on simple architectures, which may not always deliver optimal performance. This limitation is recognized in the paper mentioned. In the specific context of the paper, the K-means method is applied for document clustering, which involves grouping similar documents together.

C. Algorithm

K-means clustering is a popular unsupervised learning algorithm that can be used to cluster data points into groups. The algorithm works by first randomly selecting k data points to serve as the initial cluster centroids. The algorithm then iteratively assigns each data point to the cluster with the nearest centroid. After all data points have been assigned to clusters, the centroids are recalculated as the mean of the data points in each cluster. This process is repeated until the centroids no longer change significantly.

In the example you provided, the data set contains 200 tuples and five attributes: age, gender, expenditure, early income, and customer ID. The algorithm would first randomly select k = 4 data points to serve as the initial cluster centroids. The algorithm would then assign each of the remaining 196 data points to the cluster with the nearest centroid. The centroids would then be recalculated as the mean of the data points in each cluster. This process would be repeated until the centroids no longer change significantly.

This process would result in 4 clusters of data points. The data points in each cluster would be similar to each other in terms of their values for the five attributes. For example, all of the data points in one cluster might be young women with high expenditures and low incomes. K-means clustering can be used for a variety of tasks, such as customer segmentation, market research, and fraud detection. It is a simple and effective algorithm that can be used to cluster data points into groups.

Some of the benefits of using k-means clustering are, it is simple and easy to understand. It is computationally efficient and it can be used to cluster data points of any type. Here are some of the limitations of using k-means clustering. It requires the number of clusters to be known in advance. It is sensitive to the initial choice of cluster centroids. It can be unstable for large data sets.

Overall, k-means clustering is a powerful tool that can be used to cluster data points into groups. It is simple and easy to understand, and it is computationally efficient. However, it is important to be aware of its limitations, such as the need to know the number of clusters in advance and the sensitivity to the initial choice of cluster centroids.

In the data pre-processing stage, it is essential to check for null values and detect outliers in the dataset. Null values refer to missing or undefined data points, while outliers are data points that significantly deviate from the normal distribution of the dataset

To address null values, you can perform the following steps:

I. Identify null values: Check each attribute or column in the dataset and determine if any values are missing or marked as null.

II. Handle null values: There are several ways to handle null values, depending on the nature of the data. Some common approaches include:

- Deleting rows or columns with null values: If the null values are limited and do not significantly affect the dataset, you can choose to remove the corresponding rows or columns.
- Imputing null values: If the null values are few, you can fill them in with appropriate values. This can be done by taking the mean, median, or mode of the non-null values in the attribute or using more advanced imputation techniques.

III. Regarding outlier detection and treatment, the following steps can be followed:

- Identify outliers: Use statistical methods or visualization techniques to detect data points that deviate significantly from the majority of the dataset. Common methods include box plots, z-scores, or the interquartile range (IQR) method.
- Handle outliers: Outliers can be treated in different ways, depending on the specific context and the impact of the outliers on the analysis. Some approaches include:
  i. Removing outliers: If the outliers are due to data entry errors or measurement issues, it may be appropriate to remove them from the dataset. However, this should be done with caution, as removing too many outliers can result in a biased analysis.
  ii. Transforming outliers: Instead of removing outliers, you can transform their values using techniques such as winsorization or log transformation. This approach can help mitigate the impact of outliers while still retaining their influence on the analysis.

Once null values and outliers have been addressed in the data pre-processing stage, the next step mentioned is to choose the number of clusters using k-means clustering. The k-means algorithm aims to partition the dataset into k clusters based on the similarity of data points within each cluster.

It is important to note that the sentence "The clustering algorithm used above is shortest clustering algorithm" is incomplete and does not provide sufficient information. If you have further details or context about the clustering algorithm mentioned, please provide more information for a more accurate response.

The provided information explains the utilization of the elbow method in determining the optimal number of clusters (k) in the k-means clustering algorithm. Here is a more detailed summary of the steps involved:

i. Initialization: The algorithm begins by initializing the positions of the k centroids.

ii. Determining the number of clusters (k): The elbow method is employed to determine the appropriate value for k. The main objective of this method is to identify the k value that provides a good balance between minimizing within-cluster variation and avoiding excessive fragmentation.

iii. Euclidean distance calculation: Using the Euclidean distance metric, the algorithm calculates the distances between each data point and the centroids.

iv. Assigning data points to clusters: Each data point is assigned to the cluster whose centroid it is closest to based on the calculated distances. This results in the formation of initial clusters.

v. Updating centroids: After the initial clusters are created, the algorithm calculates the barycenter (mean) of each cluster, which serves as the new centroid for that cluster.

vi. Repeating the process: The steps of recalculating distances, reassigning data points, and updating centroids are repeated iteratively until there is minimal change in the centroid positions or a predefined stopping criterion is met.

The elbow method specifically focuses on determining the optimal number of clusters. It calculates the sum of squared distances between data points and their associated cluster centroids (within-cluster sum of squares or WCSS) for different values of k. The method suggests that as the number of clusters increases, the WCSS tends to decrease. However, there is a point where the marginal reduction in WCSS starts to level off significantly, resembling an elbow shape in the plot.

To apply the elbow method, the algorithm is run with various values of k, typically ranging from 1 to a predefined maximum value (e.g., 10). For each value of k, the WCSS is computed, and the values are plotted. The plot of WCSS versus the number of clusters (k) is then analysed to determine the point of the elbow, indicating the optimal number of clusters.

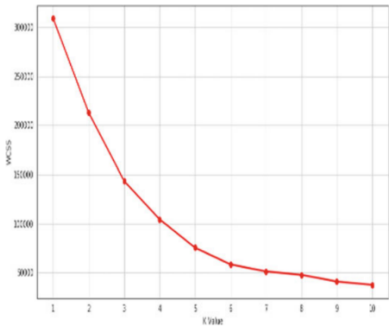

Figure 2 Elbow method

III. RESULTS

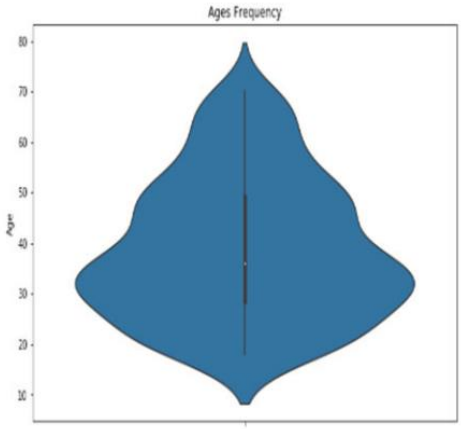

Figure 3: The id column has been removed as it is deemed irrelevant for the analysis. The figure represents the age distribution of customers. This visualization helps in understanding the age demographics of the customer base.

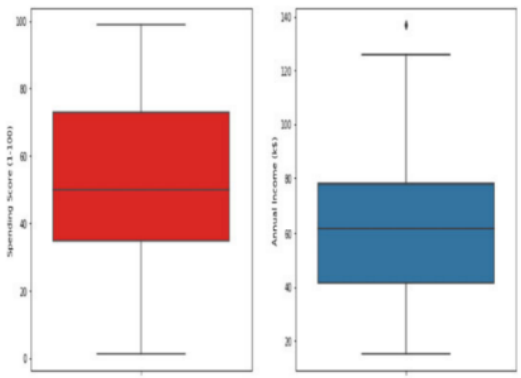

Figure 4: A box plot is generated to visualize the distribution range of the spending score and annual income. It is observed that the range of annual income is significantly higher than the range of the spending score. This plot provides insights into the spread and distribution of these two variables. In addition, a bar plot was created to display the number of clients based on their annual income. The bulk of clients earns between $60,000 and $90,000 each year.

$$\mathrm{WCSS} = \sum_{i \in n} (X_i - Y_i)^2$$

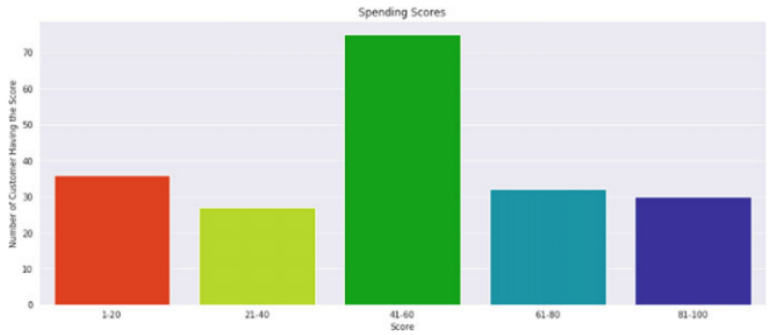

Figure 5: A bar plot is created to illustrate the number of clients based on their spending scores. The plot shows that the majority of clients fall within the spending score range of 41-60. This visualization helps in understanding the concentration of customers within different spending score ranges.

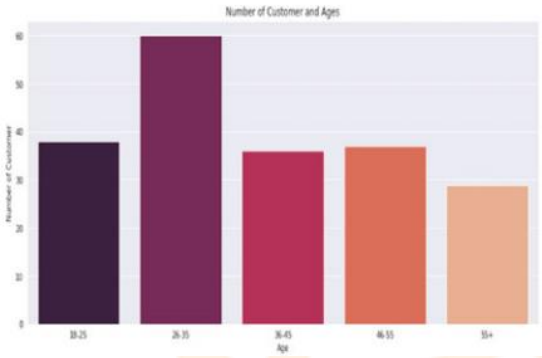

Figure 6: A bar plot is generated to examine the distribution of clients across different age groups. It is observed that the age group of 26-35 has the highest number of customers compared to other age groups. This plot provides insights into the age demographics of the customer base.

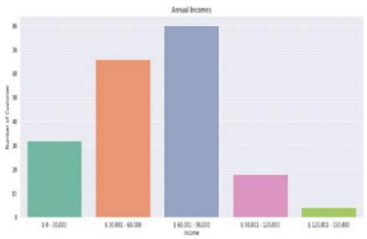

Figure 7: Additional details about this figure are not provided, but it likely pertains to the distribution or characteristics of customers based on some other variable or feature.

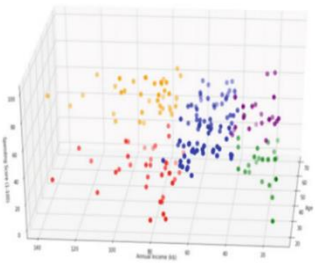

Figure 8: Visualisation of 3D cluster

The provided information describes the utilization of the Within-Clusters Sum of Squares (WCSS) metric and the creation of a 3D plot to visualize the relationship between consumers' expenditure scores and their annual revenue. Here is a summary of the mentioned points:

WCSS Calculation: WCSS is calculated by summing the squared distances of each observation from its cluster centroid. The formula provided represents the calculation of WCSS, where Yi represents the centroid of the observation Xi. By computing the WCSS for different values of k (number of clusters), the goal is to determine the appropriate number of clusters that minimizes the within-cluster variation.

Determining the Number of Clusters: To determine the appropriate number of clusters, WCSS is plotted against the number of clusters (k value). The plot helps in identifying the point where adding more clusters does not significantly reduce the WCSS. This point, often referred to as the "elbow point," is considered as the optimal number of clusters.

3D Plot of Expenditure Score and Annual Revenue: A 3D plot is created to depict the consumers' expenditure score in proportion to their annual revenue. The data points are categorized into five classes, which are represented by various colours in the 3D visualization. This plot helps in understanding the relationship between expenditure score and annual revenue and how different consumer segments are distributed within the data.

These techniques aid in exploring the optimal number of clusters for the data and visualizing the relationship between expenditure score and annual revenue. The insights gained from these analyses can inform decision-making processes, such as customer segmentation strategies and targeted marketing efforts.

IV. CONCLUSION

In conclusion, this study focused on the application of the k-means clustering algorithm for customer segmentation. The dataset used was unlabelled, meaning that the clustering was performed without relying on external data or labels. Internal clustering validation measures were used to evaluate the quality of the clustering results. By choosing the optimal number of clusters and applying the k-means algorithm, the data set was successfully segmented into meaningful groups. The k-means clustering algorithm is widely used in various industries for data structuring and customer segmentation. It helps businesses gain a better understanding of their customers and ultimately improve their revenue. The results of this study demonstrated high accuracy compared to previous analyses, indicating a successful clustering process. The findings provide valuable insights into customer segmentation and can be utilized to enhance marketing strategies and decision-making processes in businesses. Overall, the application of k-means clustering and the achieved accuracy in this study contribute to a better understanding of customer behaviour and effective utilization of customer segmentation techniques.

REFERENCES

[1] P. Anitha, M. Patil, RFM model of customer purchasing behaviour using *K*-Means algorithm. Computer Science Inf. J. King Saud Univ. (2019)

[2] Chinedu, Efficient segmentation using the *K*-Means methodology, segmentation: a plan for targeted customer services. Int. J. Surv. Mach. Intell. (IJARAI) 4(10) (2015); P. Ezenkwu, O. Simeon, K. Constance, Framework of the *K*-Means technique for efficient customer groups: a plan for directed customer care (2015)

[3] W.T. Jerry, Segmentation of Market (2007), Recovered from www.decisionanalyst.com on 12 July 2015

[4] Y. Kushwaha, D. Prajapati, Customer segmentation using the *k*-Means algorithms (2014)