**MALL CUSTOMER SEGMENTATION USING CLUSTERING ALGORITHM**

**MAHENDRAN C MBA Business Analytics,**

**Mrs . R. Raajalakshmi Assistant Professor**

**ABSTRACT**

Customer segmentation is a separation of a market into multiple distinct groups of consumers who share similar characteristics. Segmentation of the market is an effective way to define and meet customer needs. Unsupervised Machine Learning Techniques, K-Means Clustering Algorithm, Minibatch K-Means and Hierarchical Clustering are used to perform Market Basket Analysis. Market Basket Analysis is carried out to predict the target customers who can be easily converged, among all the customers. In order to allow the marketing team to plan the strategy to market the new products to the target customers which are similar to their interests.

Key words: Target Customers, Clusters, Unsupervised Learning, K-Means, Minibatch K-Means, Hierarchical Clustering Segmentation, Market Basket Analysis

**1.INTRODUCTION**

Management and maintenance of customer relationships have always played a vital role to provide business intelligence to organizations to build, manage and develop valuable long term customer relationships. The importance of treating customers as an organization’s main asset is increasing in value in the present day and era. Organizations have an interest in investing in the development of customer acquisition, maintenance, and development strategies. The business intelligence has a vital role to play in allowing companies to use technical expertise to gain better customer knowledge and Programs for outreach. By using clustering techniques like k-means, customers with similar means are clustered together. Customer segmentation helps the marketing team to recognize and expose different customer segments that think differently and follow different purchasing strategies. Customer segmentation helps in figuring out the customers who vary in terms of preferences, expectations, desires, and attributes. The main purpose of performing customer segmentation is to group people who have similar interests so that the marketing team can converge in an effective marketing plan. Clustering is an iterative process of knowledge discovery from vast amounts of raw and unorganized data. Clustering is a type of exploratory data mining that is used in many applications, such as machine learning, classification, and pattern recognition.

**2. METHODOLOGY**

**Customer Segmentation**

Over the years, as there is very strong competition in the business world, the organizations have to enhance their profits and business by satisfying the demands of their customers and attract new customers according to their needs. The identification of customers and satisfying the demands of each customer is a very complex task. This is because customers may be different according to their demands, desires, preferences and so on. Instead of “one-size-fits-all” approach, customer segmentation clusters the customers into groups sharing the same properties or behavioral characteristics. According to [1] customer segmentation is a strategy of dividing the market into homogenous groups. The data used in customer segmentation technique that divides the customers into groups depends on various factors like, demographical conditions, data geographical conditions and economic conditions as well as behavioral patterns. The customer segmentation technique allows the business to make better use of their marketing budgets, gain a competitive edge over their rival companies, demonstrating a better knowledge of the needs of the customer. It also helps an organization in increasing their marketing efficiency, planning the marketing budget, determining new market opportunities, making better brand strategy, identifying customers retention.

According to [1], Decision makers use many variables to segment customers. Demographic variables such as age, gender, family, education level and income are the easiest and most common variables for segmentation. Socio- cultural, geographic, psychographic, and behavioral variables are the other major variables that are used for segmentation.

[2], presented various clustering algorithms considering the characteristics of Big Data such as size, noise, dimensionality, algorithm calculations, cluster shape and presented a brief overview of the various clustering algorithms grouped under partitioning, hierarchical, density, grid-based and model-based algorithms.

[4] explored the necessity of segmentation of the customers using clustering algorithms as the core functionality of CRM. The mostly used K-Means and Hierarchical Clustering were studied and the advantages and disadvantages of these techniques were highlighted. Finally, the idea of creating a hybrid approach is addressed by integrating the above two strategies with the potential to surpass the individual designs.

[5], Merged clustering of fuzzy c-means and genetic algorithms to cluster, steel industry customers, by using the LRFM variables (length, recency, frequency, monetary value) system, customers were divided into two clusters.

**Clustering**

Clustering is one of the most common methods used in exploring data to obtain a clear understanding of the data structure. It can be characterized as the task of finding the subtitles and subgroups in the complete dataset. Similar data is clustered in many subgroups. A cluster refers to a collection of aggregated data points due to some similarities. Clustering is used in Market basket analysis used to segment the customers based on their behaviors and transactions.

**RESEARCH DESIGN:**

The research design employed for this project is **Descriptive Research**. This design is appropriate as the primary objective is to describe and segment customer behavior based on their credit card transaction data. Descriptive research aims to provide an accurate representation of characteristics, behaviors, or phenomena related to the research subject (in this case, credit card customers).

**Justification:**

Descriptive research is well-suited for this project because it allows for the analysis and interpretation of existing data (credit card transactions) to identify patterns, trends, and relationships. By employing descriptive research methods, such as data analysis and clustering techniques, the study can effectively segment customers based on their transactional behavior, providing valuable insights for targeted marketing and customer relationship management strategies.

We are going to aim to cluster a data set that is about behavior of the customers having credit card using many unsupervised algorithms.

**Our research question is "How many clusters can we distinguish the customers according to their transactions or behaviors?"**

**3. MODELLING**

**K Means Clustering Algorithm**

K Means Clustering is the most common and simplest Machine learning algorithm and it follows an iterative approach which attempts to partition the dataset into different “k” number of predefined and non-overlapping subgroups where each data point belongs to only one subgroup according to their similar qualities.

**Hierarchical Clustering Segmentation**

[7]The output of this model is a set of visualized clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other in features.

**4. ANALYSIS**

**K-Means**

K Means Clustering is the most common and simplest Machine learning algorithm and it follows an iterative approach which attempts to partition the dataset into different “k” number of predefined and non-overlapping subgroups where each data point belongs to only one subgroup according to their similar qualities.

The silhouette value measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation).

The values of silhouette score are close to each other in range 6 to 8. In the circumstances, Let's look at another metric. The metric is Davies Bouldin that is defined as the average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances. The minimum score is zero, with lower values indicating better clustering.

Unlike Davies Bouldin, we want to be high of Silhouette score. Hence, when we evaluate both Elbow technique and Silhouette score, optimal cluster numbers are 7 according to K-Means Algorithm. So, I have determined 7 as the k values of the K-means model.

Now, Let's visualize "CC GENERAL" dataset in three-dimensional space. Hence, we should apply to PCA before.

Comparing clustering algorithms...

 Algorithms Davies Bouldin Silhouette Score

0 K-Means 1.354323 0.237578

1 MiniBatch K-Means 1.644649 0.203657

2 Hierarchical Clustering 2.017857 0.162692

Finally, we have tried three algorithms. K-Means has the best Silhouette and Hierarchical has best Davies Bouldin score.

However, we are going to 3 algorithms for make conclusion separately.



**K-Means model conclusion**

Based on the characteristics of different K-Means clusters, here's an analysis of each cluster:

Cluster 0: This cluster exhibits high balance, high purchase frequency, and a tendency to make purchases in installments. Customers in this cluster tend to have higher credit limits and are long-term customers.

Cluster 1: The bar charts show that customers in this cluster have moderate account balances, high purchase frequency, and a high tendency to make purchases in installments. They have moderate credit limits and are long-term customers.

Cluster 2: This cluster has moderate balance, moderate purchase frequency, and a moderate tendency to make purchases in installments. These customers have moderate credit limits and shorter tenures as customers.

Cluster 3: Customers in this cluster have moderate balance, low purchase frequency, and no tendency to make purchases in installments. They have moderate credit limits and are long-term customers.

Cluster 4: This cluster exhibits high balance, low purchase frequency, a low tendency to make purchases in installments, and a high tendency for cash advances. They have higher credit limits than other clusters and are long-term customers.

Cluster 5: The bar charts indicate that this cluster has low balance, high purchase frequency, and a high tendency to make purchases in installments. They have moderate credit limits and are long-term customers.

Cluster 6: This cluster shows low balance, low purchase frequency, low tendency to make purchases in installments, and no tendency for cash advances. They have moderate credit limits and are long-term customers.

Firstly, we have started with data pre-processing. Then, we applied clustering algorithms. After comparing these clustering models then, we decided to use K-Means as the first model. Then, we divided the data into seven clusters, because seven clusters can be easily used to determine the behaviors of customers. However, each of the clusters has their own characteristics.

Overall, the K-Means clustering algorithm has effectively segmented the customers into distinct groups based on their credit card usage patterns, account balances, credit limits, and tenure as customers.

**Hierarchical Clustering Segmentation model**



**Hierarchical Clustering Segmentation model conclusion**

Based on the bar charts of Hierarchical Clustering, here's an analysis of each cluster:

Cluster 0: This cluster exhibits low balance, low purchase frequency, low tendency to make purchases in installments, and low tendency for cash advances. Customers in this cluster tend to have lower credit limits and are long-term customers.

Cluster 1: The bar charts show that customers in this cluster have moderate balance, high purchase frequency, and a high tendency to make purchases in installments. They have moderate credit limits and are long-term customers.

Cluster 2: This cluster has high balance, moderate purchase frequency, moderate tendency to make purchases in installments, and a moderate tendency for cash advances. These customers have moderate credit limits and shorter tenures as customers

**5. RESULTS AND DISCUSSION**

**K-Means Clustering (7 clusters):**

Cluster 0 (1865 customers):

- High account balance

- High purchase frequency

- High tendency for installment purchases

- Higher credit limits

- Long tenure customers

Cluster 1 (1653 customers):

- Moderate account balance

- High purchase frequency

- High tendency for installment purchases

- Moderate credit limits

- Long tenure customers

Cluster 2 (1551 customers):

- Moderate account balance

- Moderate purchase frequency

- Moderate installment purchases

- Moderate credit limits

- Shorter tenure customers

Cluster 3 (1305 customers):

- Moderate account balance

- Low purchase frequency

- No installment purchases

- Moderate credit limits

- Long tenure customers

Cluster 4 (1150 customers):

- High account balance

- Low purchase frequency

- Low installment purchases

- High cash advance tendency

- Higher credit limits

- Long tenure customers

Cluster 5 (772 customers):

- Low account balance

- High purchase frequency

- High installment purchases

- Moderate credit limits

- Long tenure customers

Cluster 6 (654 customers):

- Low account balance

- Low purchase frequency

- Low installment purchases

- No cash advances

- Moderate credit limits

- Long tenure customers

**Hierarchical Clustering (3 clusters):**

Cluster 0 (3380 customers):

- Low account balance

- Low purchase frequency

- Low installment purchases

- Low cash advance tendency

- Lower credit limits

- Long tenure customers

Cluster 1 (2802 customers):

- Moderate account balance

- High purchase frequency

- High installment purchases

- No cash advances

- Moderate credit limits

- Long tenure customers

Cluster 2 (2768 customers):

- High account balance

- Moderate purchase frequency

- Moderate installment purchases

- Moderate cash advance tendency

- Moderate credit limits

- Shorter tenure customers

The findings provide a detailed profile of each customer segment based on key transactional features like account balance, purchase patterns, payment methods, credit limits, and customer tenure. This can aid in developing targeted marketing strategies and personalized offerings tailored to the unique characteristics and behaviors of each segment .

1. Data Enrichment: Consider incorporating additional data sources or features that may enhance the clustering results and provide a more comprehensive understanding of customer behavior. This could include demographic information, purchase history from other channels, or customer feedback data.

2. Feature Engineering: Explore advanced techniques for feature selection and engineering to identify the most relevant features for customer segmentation. This can lead to more interpretable and actionable clusters.

3. Ensemble Clustering: Investigate the potential of using ensemble clustering methods, which combine multiple clustering algorithms to leverage their strengths and potentially improve the accuracy and stability of the results.

4. Temporal Analysis: Extend the analysis to incorporate temporal dynamics by examining how customer behavior and clusters evolve over time. This can provide insights into customer lifecycle management and targeted marketing strategies for different stages.

5. Domain Knowledge Integration: Collaborate with domain experts or marketing professionals to validate the clustering results and interpret the identified customer segments in the context of the specific industry or business domain.

6. Business Impact Analysis: Conduct a thorough analysis of the potential business impact and return on investment (ROI) associated with implementing targeted marketing strategies based on the identified customer segments.

7. Deployment and Monitoring: Develop a plan for deploying the customer segmentation model into production systems and establish processes for monitoring its performance, updating the model with new data, and continuously refining the segmentation strategy.

8. Privacy and Ethical Considerations: Ensure that the data collection, processing, and usage of customer information comply with relevant privacy regulations and ethical guidelines, maintaining transparency and respecting customer privacy.

By incorporating these suggestions, the research project can potentially deliver more robust and actionable insights, enabling businesses to develop effective customer relationship management strategies and drive growth through targeted marketing initiatives.

**6. CONCLUSION**

Based on the analysis and findings presented in the report, the following conclusions can be drawn:

1. Customer segmentation using unsupervised machine learning clustering algorithms, such as K-Means, Mini-Batch K-Means, and Hierarchical Clustering, can effectively identify distinct groups of customers based on their credit card transaction patterns and behaviors.

2. The K-Means algorithm performed well in segmenting customers into 7 clusters, providing insights into different customer profiles based on account balances, purchase frequencies, installment purchases, cash advance tendencies, credit limits, and tenure.

3. The Mini-Batch K-Means algorithm divided customers into 6 clusters, with similar characteristics to the K-Means clusters but with some variations in the specific cluster compositions.

4. The Hierarchical Clustering algorithm segmented customers into 3 broad clusters, capturing high-level distinctions based on account balances, purchase patterns, and cash advance tendencies.

5. Each clustering algorithm identified customer segments with varying levels of granularity, allowing businesses to tailor marketing strategies and personalized offerings based on the unique characteristics and behaviors of each segment.

6. The analysis highlighted the importance of data preprocessing, feature selection, and evaluation metrics (such as silhouette scores and Davies-Bouldin scores) in optimizing the clustering results and determining the optimal number of clusters.

7. The findings demonstrate the potential of leveraging unsupervised machine learning techniques, particularly clustering algorithms, for customer segmentation and targeted marketing initiatives in the retail and financial sectors.

8. The study provides a comprehensive framework for analyzing customer transaction data, applying clustering algorithms, interpreting the resulting segments, and deriving actionable insights to enhance customer engagement and business strategies.

Overall, the report presents a valuable approach to customer segmentation using clustering algorithms, offering businesses insights into their customer base and enabling tailored marketing efforts to improve customer satisfaction and drive business growth.

**7. REFERENCE**

|  |  |
| --- | --- |
| [1]  | D. P. Yash Kushwaha, “Customer Segmentation using K-Means Algorithm,” 8th Semester Student of B.tech in Computer Science and Engineering.  |
| [2]  | E. A. Onur DOĞAN1, “CUSTOMER SEGMENTATION BY USING RFM”.  |
| [3]  | C. M. S. R. a. K. V. N. T. Sajana, “A Survey on,” in *Indian Journal of Science and Technology, Volume 9, Issue 3,*, Jan 2016..  |
| [4]  | A. B. P. E. Shreya Tripathi, “Approaches to,” in *International Journal of Engineering and Technology, Volume 7*, 2018.  |
| [5]  | A. R. Azarnoush Ansari, ““Customer Clustering Using a,” in *International Journal of Business and Management, Volume 11, Issue 7,*, 2016.  |
| [6]  | “mini-batch-k-means-clustering-algorithm,” [Online]. Available: https://www.geeksforgeeks.org/ml-mini-batch-k-means-clustering-algorithm/.  |
|  |  |
| [7]  | “/hierarchical-clustering-customer-segmentation,” [Online]. Available: https://www.coursera.org/projects/hierarchical-clustering-customer-segmentation.  |