CUSTOMER SEGMENTATION

[1]Disha Maini, [2]Muskan Aggarwal, [3]Prof. Ankur Jain
[1,2]Student, [3]Assistant Professor
Department of Electronics and Communication Engineering
Mahavir Swami Institute of Technology, Sonipat, India

Abstract: - Customer segmentation plays an important role in customer relationship management. It allows companies to design and establish different strategies to maximize the value of customers. Customer segmentation refers to grouping customers into different categories based on shared characteristics such as age, location, spending habits and so on. Similarly, clustering means putting things together in such a way that similar types of things remain in the same group. In this study, a machine learning (ML) hierarchical agglomerative clustering (HAC) algorithm is implemented in the python programming language to perform customer segmentation on credit card data sets to determine the appropriate marketing strategies. Customer segmentation divides customers into groups based on common characteristics, which is useful for banks, businesses, and companies to improve their products or service opportunities. The analysis explores the applications of the K-means, the Hierarchical clustering, and the Principal Component Analysis (PCA) in identifying the customer segments of a company based on their credit card transaction history. The dataset used in the project summarizes the usage behavior of 8950 active credit card holders in the last 6 months, and our aim is to perform customer segmentation in the most accurate way using clustering techniques. The project uses two approaches for customer segmentation: first, by considering all variables in the clustering algorithms using the Hierarchical clustering and the K-means. Second, by applying the dimensionality reduction through Principal Component Analysis (PCA) to the dataset, then identifying the optimal number of clusters, and repeating the clustering analysis with the updated number of clusters. Results show that the PCA can effectively be employed in the clustering process as a check tool for the K-means and Hierarchical clustering.

1. INTRODUCTION

Companies that deploy customer segmentation are under the notion that every customer has different requirements and require a specific marketing effort to address them appropriately. Companies aim to gain a deeper approach to the customer they are targeting. Therefore, their aim has to be specific and should be tailored to address the requirements of each and every individual customer. Furthermore, through the data collected, companies can gain a deeper understanding of customer preferences as well as the requirements for discovering valuable segments that would reap them maximum profit. This way, they can strategize their marketing techniques more efficiently and minimize the possibility of risk to their investment. The technique of customer segmentation is dependent on several key differentiators that divide customers into groups to be targeted. Data related to demographics, geography, economic status as well as behavioral patterns play a crucial role in determining the company direction towards addressing the various segments.

2. LITERATURE SURVEY

2.1 Customer Classification

Over the years, the commercial world has become more competitive, as organizations such as these have to meet the needs and wants of their customers, attract new customers, and thus improve their businesses. The task of identifying and meeting the needs and requirements of each customer in the business is a very difficult task. This is because customers may vary according to their needs, wants, demographics, shapes, taste and taste, features and so on. As it is, it is a bad practice to treat all customers equally in business. This challenge has led to the adoption of the concept of customer segmentation or market segmentation, where consumers are divided into subgroups or segments where members of each subcategory exhibit similar market behaviors or features. Accordingly, customer segmentation is the process of dividing the market into indigenous groups.

2.2 Big Data

Recently, Big Data research has gained momentum. defines big data as - a term that describes a large number of formal and informal data, which cannot be analyzed using traditional methods and algorithms. Companies include billions of data about their customers, suppliers, and operations, and millions of internally connected sensors are sent to the real world on devices such as mobile phones and cars, sensing, creating, and communicating data. the ability to improve forecasting, save money, increase efficiency and improve decisionmaking in various fields such as traffic control, weather forecasting, disaster prevention, finance, fraud control, business transactions, national security, education, and healthcare. Big data is seen mainly in the three Vs namely: volume, variability and speed. There are other 2Vs available - authenticity and value, thus making it 5V.

2.3 Data Collection

Data collection is the process of collecting and measuring information against targeted variations in an established system, enabling one to answer relevant questions and evaluate results. Data collection is part of research in all fields of study including physical and social sciences, humanities and business. The purpose of all data collection is to obtain quality evidence that allows analysis to lead to the creation of

convincing and misleading answers to the questions submitted. We collected data from the UCI Machine Learning Repository.

2.4 Clustering data

Clustering is the process of grouping the information in the dataset based on some similarities. There are a number of algorithms which can be chosen to be applied on a dataset based on the situation provided. However, no universal clustering algorithm exists that's why it becomes important to opt for appropriate clustering techniques. In this paper, we have implemented three clustering algorithms using python sklearn library.

2.5 K-Mean

K- means that an algorithm is one of the most popular classification algorithm. This clustering algorithm depends on the centroid where each data point is placed in one of the overlapping K clusters pre-programmed into the algorithm, The clusters are created that correspond to the hidden pattern in the data that provides the information needed to help decide the execution process. There are many ways to make k-means assembling; we will use the elbow method

3. PREPROCESSING

3.1 CHALLENGES OF PERFORMING ANALYSIS

The benefits of customer segmentation analysis are clear. By having a stronger understanding of their consumer base, retailers can properly allocate resources to collect and mine relevant information to boost profits. However, getting to the point of performing high-level customer segmentation analysis is more difficult than originally thought for many retailers. Many retailers may have the rights to the necessary data to perform the analysis, but do not have either the ability to access it in a user-friendly manner or have an employee that has the skills to work with it. The lack of proper personnel or equipment to handle the necessary volume of data is perhaps the biggest hindrance to smaller firms being able to perform such analysis. The popularity of open source programming software such as R or Python has certainly helped make this type of analysis more accessible, but it still would require retailers having someone on their team who can code in either of those languages. Additionally, some retailers are simply unaware of either the extent of their data collection or are not yet inspired to dig into it. Nevertheless, retailers that have not fully adopted customer segmentation analysis are likely not doing so simply because they cannot afford to spend the time, money, or labor to perform the analysis. Therefore, it is an aim of this paper to show that this rich analysis can be performed cheaply and efficiently.

However, there is a far subtler but still consequential reason why retailers do not implement customer segmentation analysis: it is too complicated to understand. When compared to traditional demographic segmentation or RFM analysis, high-level customer segmentation analysis requires far more precise knowledge of machine learning and the mathematics that describe how the algorithms work. In addition, traditional marketing analysts are not equipped with the math or programming skills necessary to successfully implement customer segmentation analysis with machine learning methods; similarly, programmers and data analysts are not well- suited to handle marketing tasks. This poses another conundrum as it involves transforming a typical marketing assignment—segmenting customers based on purchasing behaviors— into a purely programming one, which means the marketing team does not have the skills to code it up themselves but the programming team does not have the marketing skills to interpret the results. Hence, there is a necessity for a hybrid role that involves knowledge of the business, programming, and marketing. In modern workspaces, this role is dubbed the data scientist or information specialist. In sum, customer segmentation analysis is the process of trying to understand a consumer base by splitting it up into segments. While traditional analysts found some success with demographic or RFM analysis, these models simply do not have the technological capabilities to provide rich insight into more specific details regarding the customers. On the other hand, customer segmentation analysis that is combined with machine learning methods has the ability to transform the way a retailer thinks about their data. As such, retailers are trying to find cheap, easy ways to implement and communicate how clustering can be used to segment their customers. Now that there has been plenty of introduction into customer segmentation analysis, it is time to take a look under the hood of some clustering algorithms before finally engaging in discussion of the analysis.

4. DATASET COLLECTION

➔ DATASET 1

	ID	Gender	Ever_Married	Age	Graduated	Profession	Work_Experience	Spending_Score	Family_Size	Var_1	Segmentation
0	462809	Male	No	22	No	Healthcare	1.0	Low	4.0	Cat_4	D
1	462643	Female	Yes	38	Yes	Engineer	NaN	Average	3.0	Cat_4	A
2	466315	Female	Yes	67	Yes	Engineer	1.0	Low	1.0	Cat_6	B
3	461735	Male	Yes	67	Yes	Lawyer	0.0	High	2.0	Cat_6	B
4	462669	Female	Yes	40	Yes	Entertainment	NaN	High	6.0	Cat_6	A

➔ DATASET 2

	Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Defaulted	Address	DebtIncomeRatio	Segmentation
0	1	41	2	6	19	0.124	1.073	0.0	NBA001	6.3	A
1	2	47	1	26	100	4.582	8.218	0.0	NBA021	12.8	A
2	3	33	2	10	57	6.111	5.802	1.0	NBA013	20.9	A
3	4	29	2	4	19	0.681	0.516	0.0	NBA009	6.3	A
4	5	47	1	31	253	9.308	8.908	0.0	NBA008	7.2	A

4: VISUALISATION

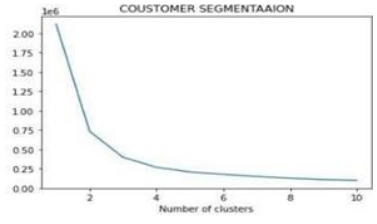
4.1 VISUALIZING THE OPTIMAL NUMBER OF CLUSTERS

4.1.1 ELBOW GRAPH

K-means is a simple unsupervised machine learning algorithm that groups data into a specified number (k) of clusters

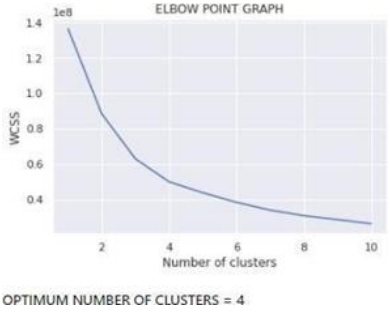
The elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters.

➔ DATASET 1



OPTIMUM NUMBER OF CLUSTER = 3

➔ DATASET 2



OPTIMUM NUMBER OF CLUSTERS = 4

4.2 K-MEANS ALGORITHM FOR CLUSTER

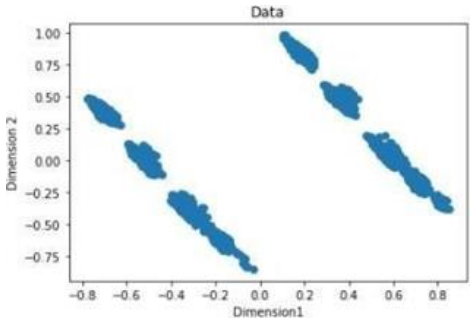
While using the k-means clustering algorithm, the first step is to indicate the number of clusters
(k) that we wish to produce in the final output. The algorithm starts by selecting k objects from dataset randomly that will serve as the initial centers for our clusters. These selected objects are the cluster means, also known as centroids. Then, the remaining objects have an assignment of the closest centroid. This centroid is defined by the Euclidean Distance present between the object and the cluster mean. We refer to this step as "cluster assignment". When the assignment is complete, the algorithm proceeds to calculate new mean value of each cluster present in the data. After the recalculation of the centers, the observations are checked if they are closer to a different cluster. Using the updated cluster mean, the objects undergo reassignment. This goes on repeatedly through several iterations until the cluster assignments stop altering. The clusters that are present in the current iteration are the same as the ones obtained in the previous iteration.

5. PLOTS

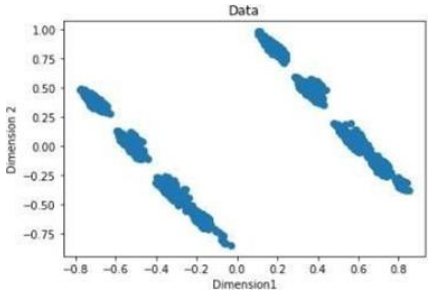
5.1 VISUALIZING THE CLUSTERING RESULTS USING THE FIRST TWO PRINCIPAL COMPONENTS INTRODUCTION

Principal Component Analysis is an unsupervised learning algorithm that is used for dimensionality reduction in machine learning. It is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation. These new transformed features are called the Principal Components. It is one of the popular tools that is used for exploratory data analysis and predictive modeling. It is a technique to draw strong patterns from the given dataset by reducing the variances. PCA generally tries to find the lower-dimensional surface to project the high-dimensional data. PCA works by considering the variance of each attribute because the high attribute shows the good split between the classes, and hence it reduces the dimensionality. Some real-world applications of PCA are image processing, movie recommendation systems, and optimizing the power allocation in various communication channels. It is a feature extraction technique, so it contains the important variables and drops the least important variable.

➔ DATASET 1



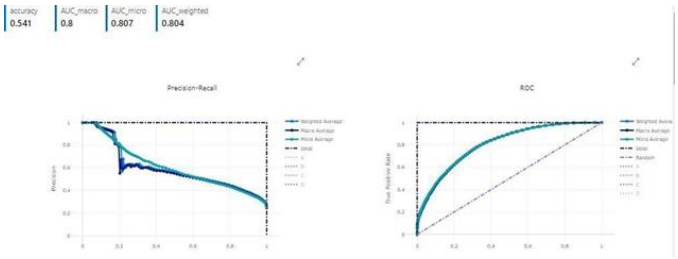
➔ DATASET 2



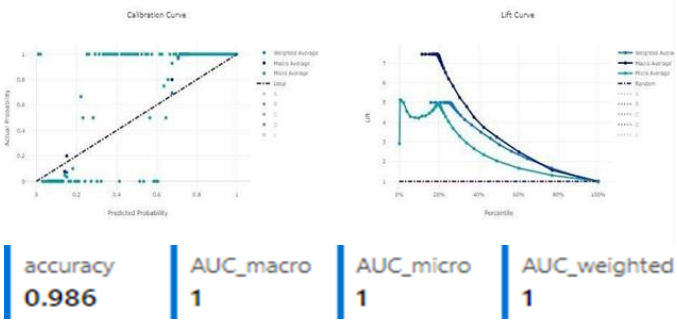
6. COMPARATIVE STUDY

PARAMETERS	DATASET 1	DATASET 2
accuracy	0.5411514071509841	0.985882352941176
precision_score_weighted	0.540228615767184	0.986824670532924
f1_score_macro	0.5290983530013584	0.982375814880148
recall_score_weighted	0.5411514071509841	0.985882352941176

➔ DATASET 1



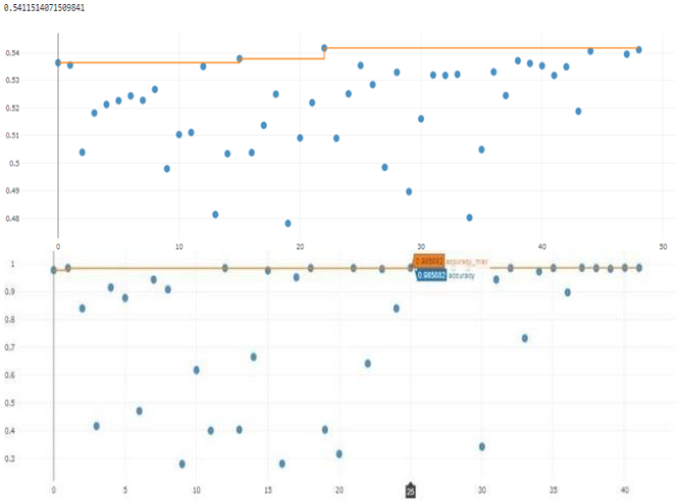
➔ DATASET 2



7. RESULTS AND CONCLUSION

7.1 RESULT

Crucial implications that accompany these results, particularly with regard to the structure of the clusters. To begin, analysis of both algorithms indicates that the optimal number of clusters to choose is somewhere between five and six. With only five or six clusters, it is straightforward to separate out the clusters to uncover patterns. However, a lower number of clusters might be easier to separate, but the clusters will be far less informative and too general to make accurate predictions. This is often a problem with traditional customer segmentation analysis. Since there is motivation to keep analysis simple and not expand features unless necessary, traditional customer segmentation analysis often leads to oversimplification of the clusters and thus complicates managerial action. Regardless, the decision to cluster with five or six segments is present throughout customer segmentation analysis research. In one sense, this implies that clustering cannabis retail data, even with cannabis-specific variables, may not be different from clustering other retail data. In turn, applying methods performed with other types of retail data to cannabis retail data is not only applicable, but perhaps even recommended as both the size and complexity of the data evolve



7.2 CONCLUSION

For the most part, the cannabis industry is in its nascent stages. The intense federal criminalization of cannabis for years totally hampered professional research into all facets of cannabis, from cultivation to retail to consumption. As a result, dispensaries are learning how to navigate not just a thicket of regulations and other constraints, but also an unclear road of consumer behavior. Conducting direct research with consumers and products is not possible, so retailers must look inward to uncover the behaviors of their customers. Despite many retailers outside of cannabis have had tremendous success with traditional customer segmentation analysis, the supply of skilled analysts willing and capable to serve the cannabis industry is far smaller. To compound this, there is plenty of data in the cannabis industry— due to the enforcement of a traceability system— but few ways to access it. Although dashboards and elaborate interfaces have easened the responsibility of finding patterns or commonalities in retail data, none of them provides statistics or data that is rich enough to make advanced insights such as customer segmentations. As a result, it is necessary to bring in tools that are specifically built for situations such as this: machine learning. With ample data, cannabis-specific domain knowledge, and a background in machine learning, developing a set of scripts to cluster the raw data was possible. After engineering relevant features and reformatting the data, it was possible to perform customer segmentation analysis with two different clustering algorithms: K-Means and Agglomerative. Even though the algorithms used different numbers of clusters in their clusterings, they essentially convey the same three pieces of information. First, flower and vape consumption were the defining characteristics of the largest clusters, which hints at the importance of these two products to a dispensary's success. Second, both algorithms generated a cluster of ultra-frequent consumers, with average visits and total spent significantly higher than the rest of the clusters. Lastly, the tables also show that older consumers tend to enjoy edibles and topicals more than other consumers; on the flip side, younger consumers tend to enjoy vapes and concentrates more. Regardless of the information provided, the results provide actionable ways for retailers to employ a marketing campaign or similar segmentation for their consumers.

Despite the usefulness of the analysis as-is, there are numerous routes for improvement and growth. While there was motivation to keep the number of features low, adding a separate feature to account for the recency of the consumer would provide clearer details on whether certain purchase profiles are more common now than in the store's past. On a similar note, finding ways to cluster a customer quicker (such as in one or two visits rather than three) could generate insights into not only the evolutionary aspect of the clustering but potentially also the leakage of customers. Finally, attempting the same analysis with numerous other clustering algorithms such as Gaussian Mixture Models or deep learning would bring about insight into the stability of cluster formation.

REFERENCES

1. V. Sharma, H. K. Saxena and A. K. Singh, "Docker for Multi-containers Web Application," 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), 2020

2. Research on Cloud Data Storage Technology and Its Architecture Implementation (2012) Liu, Kun & Dong, Long-jiang.

3. An Introduction to Docker and Analysis of its Performance (2017). Babak Bashari Rad, Harrison John Bhatti, Mohammad Ahmadi.

4. Bridging the Missing Link of Cloud Data Storage Security in AWS (Jan, 2010) J. Feng, Y.

5. Liu, D., & Zhao, L. (2014). The research and implementation of cloud computing platform based on docker. 2014 11th International Computer Conference on Wavelet Active Media Technology and Information Processin (ICCWAMTIP).

doi:10.1109/iccwamtip.2014.7073453

6. A Survey on Docker Container and its Use Cases Bellishree P , Dr. Deepamala. N, Department of CSE, RVCE, Bengaluru ,Associate Professor, Department of CSE, RVCE, Bengaluru

7. Zhu Li-juan, "Research and application of SQL Server in the trade management system," 2011 3rd International Conference on Computer Research and Development, 2011, pp. 209-213, doi: 10.1109/ICCRD.2011.5764282.

8. An Overview of Data Storage in Cloud Computing by Issac Odun-Ayo, Olasupo Ajayi, Boladele Akanle, Ravin Ahuja (2017)

9. An analysis of the Server Characteristics and Resource Utilisation in Google Cloud by Peter Garraghan, Paul Townend, Jie Xu (2013)

10. Del L. Hawkins, Roger J. Best, Kenneth A. Coney. Customers' Behaviors(seventh edition)

11. Management Science: A Comparative Research on the Methods of Customer Segmentation Based on Consumption Behavior. 2003.2, Vol.16.

12. Bernard J. Jansen, Soon-gyo Jung, Dianne Ramirez Robillos, Joni Salminen. (2021) Too few, too many, just right: Creating the necessary number of segments for large online customer populations. Electronic Commerce Research and Applications

13. Agresti, A. 2002. Categorical data analysis, Hoboken, New Jersey: Wiley

14. Bock, T. and Uncles, M. 2002. A taxonomy of differences between consumers for market segmentation. International Journal of Research in Marketing, 19: 215–224.

15. Bodapati, A. V. and Gupta, S. 2004. The recoverability of segmentation structure from store-level aggregate data. Journal of Marketing Research, 41: 351–364.

