Analysis of Customer Segmentation Clustering Techniques

1st Utsav Sharma
Amity University,
Noida, Uttar Pradesh
utsavsharma1006@gmail.com

2nd Aditi G
Amity University,
Noida, Uttar Pradesh
aditi.ganapathi.2000@gmail.com

3rd Nihar Ranjan Roy
Sharda University,
Noida, Uttar Pradesh
niharranjanroy@gmail.com

4th Shailendra Narayan Singh
Amity University,
Noida, Uttar Pradesh
snsingh36@amity.edu

***Abstract*: The growing customer base from all over the interconnected world, it has become vital for a company's marketing team to know and thus understand their target customer base in order to build an effective and appropriate customer-company relation. A practical implementation of analysis of customer data to extract valuable conclusion for companies is customer segmentation clustering techniques. Unsupervised machine learning algorithms plays a vital role in segmenting customers from all around the world. Due to various clustering techniques available today, businesses need the most effective and accessible clustering techniques for their application. This research paper tackles the comparison and evaluative problem, where an automobile company's unlabeled data is clustered into 3 predominant groups. It covenants with the questions that are previously conferred by the scholars and it spotlights the definition, methodology of market dissection and discerns which clustering technique betweenK-means and Hierarchical clustering is better with different number of cluster and different data size.**

Keywords-- Anthropoid intellect, Auto-Encoders, K-means clustering, elbow method

I. INTRODUCTION

High-tech scientific commotions for instance synthetic brainpower (more commonly known as AI), gigantic record systematic analysis (big data inquiries), internet of things (IoT), have propound electronic elucidations for alluring and magnetizing the patron garrison. Budding artificial automations stipulate a spirited benefit by expediating the clienteles 'merchandise and facility offering [1]. In the contemporary commerce picture, the ruthless antagonism and technological unsettling have rehabilitated the way officialdom's function. Worldwide clientele-mesial attitude single-minded on consumer desires plays a critical protagonist in organizational progression [2]. Artificial intelligence (AI) is an extensively expended incipient knowhow which benefits consortiums trail instantaneous statistics to scrutinize and retort promptly to habitué necessities [3]. AI propositions buyer intuition on buyer comportment elemental for buyer temptation and punter maintenance. AI spurs the habitué's ensuing gesture and delineates the boilersuit experience in an augmented manner. AI paraphernalia are handy to comprehend and understand client anticipations and traverse the potential track [4].

The term Artificial Intelligent conspicuously primes to public to ponder merely around programmed robots who grind for individuals since the public has only perceived the human-machine collaboration in cinemas or any performances barely through automatons [5]. Synthetic Intelligence concerns to any genus of contrivance that prerequisites to cogitate like a mortal stemming in unremitting comprehending and conceptualizing. These are the topographies of AI that marks it exclusive. Contrasting to anthropoid intellect, synthetic intellect (AI) is the astuteness exhibited by the contraptions. A co- ordination of smart proxy engines that senses the ecosystem to efficaciously accomplish its aim embodies artificial intelligence. Artificial intelligence portrays contraptions (computers) that feign perceptive and affective jobs of hominoid cognizance. The expansion of Synthetic intelligence is astounding, and specialists have toiled diligently to enhance AI conceptions since the last few decades. This exertion led to several chief innovations such as neural network applications in myriad sectors and big data and context.

Artificial Intelligence has unearthed its functions in distinctive milieu in modern day commercial setup. Experts and academics deem that Synthetic Acumen is the imminent technology of our guild. Amid the expansion of computerization, this planet has befitted a maze of interlocked webs. The technological realization leads to venture in knowledge engineering (or AI) for big data inquiry to engender market intelligence. AI is incessantly getting employed to profit sundry businesses. As the confederacies passage headlong near Industry 4.0, Synthetic Brainpower & erstwhile up-and-coming knowhows and automations are also sprouting concomitantly. Conversely, the enactment of AI within each sphere hasn't stood doable owed to numerous limitations, but researchers are effective on techniques which gratify to the philosophy of brain and self-awareness of the synthetically clever machines.
Currently the populace intermingles thru specific ritual of AI in circadian behaviours. For instance, the client relishes the programmed message sieving facet. In the latest mobile handset, the consumer may perhaps satiate out an almanac with Bixby or Siri or Cortana. The customer of a new automobile becomes aided whilst driving. Analogously, AI can systematize the professional course, absorb

discernments from previous records, and spawn customer and bazaar intuitions via the program-based set of rubrics. Engineering science such as, Deep Learning, Neural networks, Natural Language Processing and Machine Learning edify engines to analyse big data for the production of marketplace astuteness.

This paper strives to identify and study the difference in clustering quality of k-means and hierarchical clustering by first, with reference to the data size and second, with reference to the number of clusters using the same dataset for both techniques respectively. This applied research model normalizes the sales records of an automobile company for discerning co-relation between the attribute of data to apply auto-encoding for predicting the optimal number of clusters and thus uses K-means and Hierarchical clustering techniques.

In the remainder of this research paper, the next section-II provides literature evidence and reason for analyzing K-means and hierarchical clustering among various other techniques available while also delineating which among different models have been looked upon during this research. After taking up the two above mentioned clustering algorithms, the methodology taken to implement these techniques has been described following which are results clearly stating which among these two is better along with conclusion at last.

II. Literature Review

Unsupervised Learning, customer segmentation and the influence of AI in customer segmentation were intricately analyzed in previous studies.

Due to the mass availability of unlabeled data over labeled data, this research has taken unlabeled data for implementation in an effort to make this study widely accessible and genial to a large audience. Since unlabeled data has been used, a choice between unsupervised machine learning and the nascent semi-supervised machine learning algorithms are taken into consideration. While semi-supervised learning is better than the conventional unsupervised learning in most application, semi-supervised learning methods doesn't scale well for the large amount of data [5] especially, graph based semi-supervised methods takes cubic time complexity O (n^3). As a consequence of this and for making comparative analysis over a large dataset, SSL did not deem fit for application. Further along, for companies to describe their customers as valuable, intangible assets that directly correspond to financial outcomes [6] Customer Relationship Management CRM techniques are used for customer segmentation. CRM is a philosophy that not only improves customer loyalty and value for a business but also involves clienteles-focused policies to gain and retain more and more customers [7].

Earlier in their research report, Cheng and Chen [8] stated that pervasive achievement of CRM progressions necessitates the effectual application of IT paraphernalia. The application of arithmetic algorithms and qualitative techniques applied on consumer records to undertake segregated or customized advertising evaluations is predominantly associated with Analytical CRM which relies over quantitative analysis. Bahari and Elayidom [9] in their research constructs that investigating buyer statistics with data mining methods is crucial to cognize consumers, to instrument a modest CRM model, and to upsurge client worth. Liu [10] stated that Analytical CRM for focused publicizing is fundamental to contrive consumer driven stratagems, and superlatively entails aiming on consumer records that incorporate the all-inclusive consumer voyage. With this concern in mind, consumer record relies on innumerable dealings instituted with clienteles. Conversely, most communications do not exhibit clients' buying intents as accurate as the preceding procurement registers.

Among many of the CRM models, the most famous one is the Recency Frequency Monetary RFM model. Recency, Frequency, and Monetary Technique is extensively acknowledged as modus to recognize the features of clients vis-à-vis procurement patterns. This system scrutinizes prior silo puchases and ascertains consumer clusters that encompass consumers who mimic or echo one another in respect to transaction choices. With reference to the previously mentioned statement, RFM can be categorized as a comportment

segmentation archetypal that could stipulate noteworthy evidence in publicizing and other selling verdicts.

In the model mentioned as RFM [11], the consumption deeds of customers are denoted as an amalgamation of three different dimensions which essentially dangle upon the summation and timing of purchases done by customers. Meticulously, the transaction record is probed with an evaluation over the previously mentioned dimensions, and every customer is appraised correspondingly.

Most conspicuously, Migros Türk, Cooil et al. [12] the three dimensions considered are Recency (duration since the time of previous transaction), Frequency (total count of transactions) and Monetary (average amount purchased). Building an approach parallel to this, yet not completely similar, this research paper worked on clustering customers who were most profitable based on the dimensions given by RFM. Further along, clustering in modern world has various algorithms and techniques while not many are most accessible or easily available, neither can easily be understood by common non-technical related background marketers.

While clustering is a ubiquitous and has been developed enough to crack multifarious tribulations across detailed arenas, nevertheless, there is no clustering algorithm which can unanimously or pervasively be expended to decipher all obstacles [13]. "It has been extremely challenging to cultivate an integrated skeleton for analyzing about clustering on a technical magnitude, and severely distinct methodologies to clustering" [14], as substantiated by an impossibility theorem. Even after having no definite solution, clustering broadly has 4 stipulated steps including feature selection and extraction, cluster algorithm design and selection, cluster validation and result implementation which have been done carefully during the implementation of this research paper.

Clustering has a prolonged antiquity, with an ancestry courting nether to Aristotle [15]. General references on clustering techniques include [14], [15], [16], [17], [18],

[19], [20]. Imperative investigation papers on clustering methods additionally are present in this literature. Preliminary to a statistical configuration identification stance, Jain, Murty, and Flynn reexamined the clustering models and supplementary imperative questions allied to clustering examination [16], while Hansen and Jaumard explained the bundling complications with a arithmetical indoctrination outline [17]. Kolatch scrutinized clustering algorithms applications for three-dimensional record organizations [18] and evidence recovery [20], correspondingly.

Berkhin additional enlarged the subject to the unabridged arena which is of data mining [4]. Murtagh stated the expansions in hierarchical clustering models [21] and Baraldi gauged numerous algorithms for neural networks clustering and fuzzy logic [20]. In accumulation with other review papers, quantitative and evaluative research on clustering techniques is also imperative. Rauber along with research associates offered pragmatic outcomes for five mainstream clustering techniques [8]. Wei and others sited the weight on the contrast of steadfast algorithms for sizeable databanks [280]. Scheunders equated various clustering algorithms in the arena of color image quantization with giving stress on computing time and the prospect of locating global optima [20]. Functions and assessments of diverse clustering techniques for the investigation of gene expression figures from DNA microarray experimentations were pronounced in [21]. Steinbach, Karypis, and Kumar [16] in their study gave an investigational appraisal on record clustering algorithms, grounded on K-means and hierarchical clustering.

After scrutinizing from major research done previously as mentioned above, it can be synthesized that K-means and Hierarchical clustering is one of the basic yet crucial clustering techniques which can easily and most accessibly be used today. Thus, this research paper performs a comparative study upon these two most famous clustering techniques, K-means and Hierarchical clustering techniques.

III. PROPOSED METHODOLOGY

The simulated steps performed in implementation of this proposed methodology is given in fig 1. Distinctive commencing points and conditions typically prime to diverse taxonomies in clustering techniques and algorithms. But in this research implementation, one basic yet universal set of rules and steps have been followed to implement both K-means and hierarchical clustering. These steps are described in this section hereafter:

A. Data Acquisition

The data obtained had 2823 sales history entries for the last 2.5 years of an automobile company which has 25 attributes relating to the order history and details of the product as well as the customer's purchase habit has been delineated. This data was taken from an open-source database known as Kaggle on the web which consists of sales customer data.

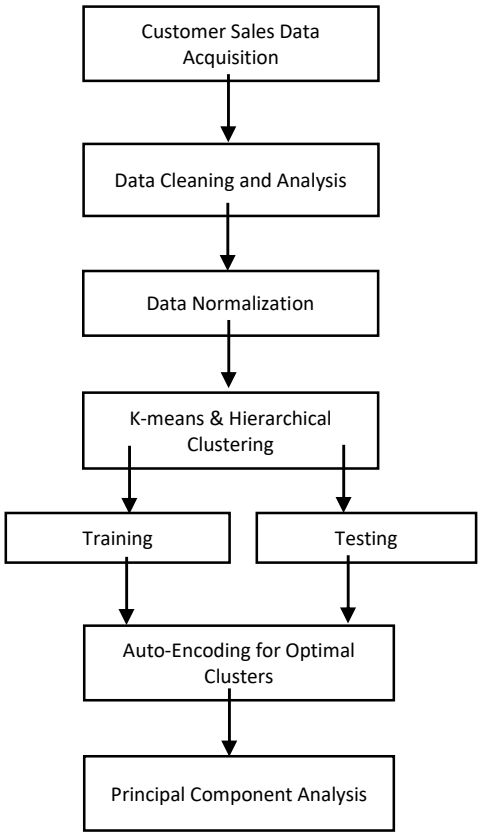

Fig 1: Flow Process of Proposed Methodology

B. Data Cleaning and analysis

The considered customer sales dataset contains information too deep about the customer that renders itself irrelevant such as address, postal code, state, name of customer and order number which are thus dropped. Order status is dropped due to high imbalance Attributes with null and non-unique is again removed before the next step as well as orderdate is changed to date-time format so as to analyze the peak sales season for the company. Dataset is visualized and grouped according to countries.

C. Data Normalization

The converted and cleaned sales dataset now undergoes one-hot encoding and further matrix correlation is done to intricately understand and visual the normalized data. The relation between variables in the dataset is performed using pair plots and thus possible conclusion are drawn out to analyze and cluster the data.

D. Clustering

Before moving to the K-means and hierarchical clustering techniques to cluster the data under unsupervised machine learning, the optimal number of clusters are chosen. Elbow method of finding optimal number is put under application. The WCSS, Euclidian distance is used to draw a graph between scores and number of clusters to delineate a point where the graph forms an elbow at which point the optimal number is figured out. The formula for Within-Cluster-

Sum-of-Squares is given by:

$$WCSS = \sum_{C_k}^{C_n} \left(\sum_{d_i \, in \, c_i}^{d_m} distance(d_i, C_k)^2 \right) \quad \text{(I)}$$

Where,
C is the cluster centroids and d is the data point in each Cluster

E. *Auto-Encoding*

While analyzing each feature based on clusters, dense layers are created for auto-encoding based on co-related data found by pair plots. The auto-encoders are used to predict sales values for further finding the optimal number of clusters. The formula for the bottle-neck layer of autoencoders is:

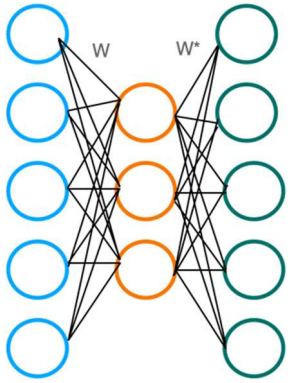

Fig 2. Auto-Encoder Architecture

Encoder:

$$h(x) = sigmoid(W * x + b) \quad \text{(II)}$$

Decoder:

$$\hat{x} = sigmoid(W^* * h(x) + c) \quad \text{(III)}$$

where x is the input dataset, Wand W* are the tied weights which are weights from input to hidden layer and weights from hidden layer to output layer and $\hat{x}$ is the output encoded data.

F. *PCA*

Principal component analyses performed on the data for dimensionality reduction of a multidimensional into only 3 dimensions so as to easily visualize the data and analyse it using scatter plot.

IV. RESULT

Auto-encoding for predicting the optimal number of clusters works best when the data is efficiently corelated. The data, after being cleaned and normalized or balanced can be seen in Fig: 3. The diagonal row in the pair plots are identical which are delineated for each attribute being related to others and the identical observation confirms the appropriate relativeness of attributes in the dataset, that now clustering can efficiently be done on the dataset. Hence using auto-encoders for creating dense layer to find the optimal number of clusters, which came out 3 for this used case, namely (cluster-0, cluster-1, cluster-2) for grouping the customer in a particular automobile market into 3 groups for marketers to do target marketing and target advertising.

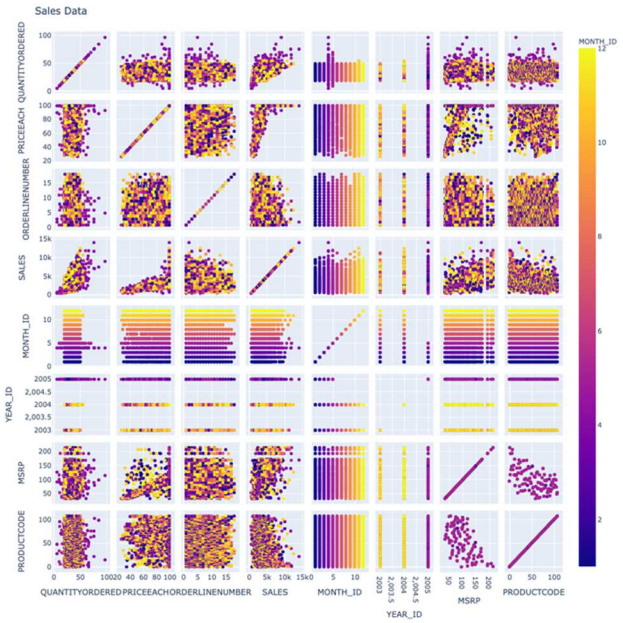

Fig.3. Normalized Data Co-relation establishment

Furthermore, the elbow method provides the optimal number of clusters at 3, giving a bend at around 3 as shown in Fig :4. It is necessary to compare the different clustering technique discussed in order to identify which clustering algorithm should use in which situation. The use of K-means and hierarchical clustering on different sizes of dataset and the number of optimal clusters chosen alters the quality(accuracy) of each algorithm.

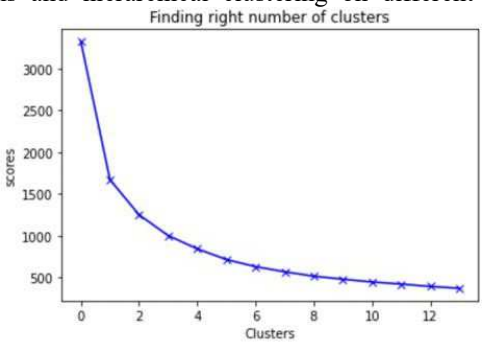

Fig. 4. Elbow method depiction for optimal number of clusters

As clearly shown in table 1, K-means is a better clustering algorithm when the number of clusters increase as even though its efficiency decreases with the increase in number of clusters, yet its quality(accuracy) is much more than hierarchical clustering algorithm in statistical terms. The

measure of precision is not possibly logical in customer segmentation issue since this isn't a classification algorithm or regression model. So, instead the efficient of both clustering algorithms is mentioned in terms of their relative quality(accuracy) as the number of clusters and data decreases and its impact on the efficiency of the entire model.

TABLE I. RELATION BETWEEN QUALITY(ACCURACY) OF ALGORITHMS AND NUMBER OF CLUSTER

Number of clusters	K-means	Hierarchical
8	1111	1080
16	1090	950

The impact of how efficient the K-means and hierarchical clustering algorithm are when the data size increases is given by table 2.

TABLE II. RELATION BETWEEN QUALITY(ACCURACY) OF ALGORITHMS AND DATA SIZE [17]

Data Size	K-means	Hierarchical
36000	909	855
40000	96	92

Additional PCA gives a genial and user-friendly output for non-technical background marketers to figure out their target customers and benefit from sales data.

While K-means clustering seems more efficient in Market segmentation issue as depicted, Hierarchical clustering also provides better results out of the many other clustering techniques available. Between the two compared here, K-means gives better market segmentation clustering than hierarchical clustering algorithm in this case.

V. CONCLUSION

Finally, as a vital instrument for data examination, cluster analysis scrutinizes unlabeled data and this procedure comprises of a chain of stages, extending from preprocessing the data and algorithm formation to result evaluation. Each of these steps are firmly interconnected with one other and wields prodigious issues to scientific disciplines. In this paper, we dwell our emphasis on two of the most famous and accessible clustering techniques after reviewing a comprehensive multiplicity of algorithms showing in literature. These two techniques advance from diverse exploration areas, aiming to crack distinctive hitches, and devour their own pros and cons. Although we have previously seen countless instances of efficacious uses of clustering exploration, yet there are many open hitches that continue to remain because of the continuation of numerous inherent indefinite influences in this field. These hitches have previously drawn traction and will endure more in the future to attract exhaustive exertions from extensive disciplines. We summarize and conclude this survey by mentioning that while there is no clustering algorithm that can unanimously or globally be applied to application for solving all problems yet in our problem statement with the available dataset, it is safe to conclude that in this implementation of clustering analyses which is inherently a supremely subject area of study, K-means performs better than hierarchical clustering only for this respective dataset used. With evolving time, many more effective clustering algorithms have been developed which are subject to future research and study.

REFERENCES

[1] Tynan, A. Caroline, and Jennifer Drayton. "Market segmentation." Journal of marketing management 2.3 (1987): 301-335.

[2] Wind, Yoram, and Richard N. Cardozo. "Industrial market segmentation." Industrial Marketing Management 3.3 (1974): 153-165.

[3] Yanelovich, Dniel, and Daavid Meaer. "Rediscovering market segmentation." Harvard business review 84.1 (2006): 122.

[4] Berkhin, Pavel. "A survey of clustering data mining techniques." Grouping multidimensional data. Springer, Berlin, Heidelberg, 2006. 25-71.

[5] KABASAKAL, İnanç. "Customer segmentation based on recency frequency monetary model: A case study in E-retailing." Bilişim Teknolojileri Dergisi 13.1 (2020): 47-56.

[6] Smith, Wendell R. "Product differentiation and market segmentation as alternative marketing strategies." Journal of marketing 21.1 (1956): 3-8.

[7] Dolnicar, Sara, Bettina Grün, and Friedrich Leisch. Market segmentation analysis: Understanding it, doing it, and making it useful. Springer Nature, 2018.

[8] Cheng, Ching-Hsue, and You-Shyang Chen. "Classifying the segmentation of customer value via RFM model and RS theory." Expert systems with applications 36.3 (2009): 4176-4184..

[9] Bahari, T. Femina, and M. Sudheep Elayidom. "An efficient CRM-data mining framework for the prediction of customer behaviour." Procedia computer science 46 (2015): 725-731.

[10] Liu, Ying, et al. "Multicriterion market segmentation: a new model, implementation, and evaluation." Marketing Science 29.5 (2010): 880-894.

[11] Leu, Yeng, et al. "Multicriterion market segmentation: a new model, implementation, and evaluation." Marketing Science 29.5 (2010): 880-894.

[12] Cooil, Bruce, Lerzan Aksoy, and Timothy L. Keiningham. "Approaches to customer segmentation." Journal of Relationship Marketing 6.3-4 (2008): 9-39.

[13] Xu, Rui, and Donald Wunsch. "Survey of Clustering Algorithms IEEE Transactions On Neural Networks, Vol. 16 (3)." (2005): 645-678..

[14] Anderberg, Michael R. Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks. Vol. 19. Academic press, 2014.

[15] Weinstein, Art. Handbook of market segmentation: Strategic targeting for business and technology firms. Routledge, 2013.

[16] Jain, Anil K., Robert P. W. Duin, and Jianchang Mao. "Statistical pattern recognition: A review." IEEE

Transactions on pattern analysis and machine intelligence 22.1 (2000): 4-37.
[17] A. Jain, R. Duin, and J. Mao, “Statistical pattern recognition: A review,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 1, pp. 4–37, 2000.
[18] Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn. "Data clustering: a review." ACM computing surveys (CSUR) 31.3 (1999): 264-323.
[19] A. Krogh, M. Brown, I. Mian, K. Sjölander, and D. Haussler, "Hidden Markov models in computational biology: Applications to protein modeling." Journal of molecular biology 235.5 (1994): 1501-1531.
[20] Scheunders, Paul. "A comparison of clustering algorithms applied to color image quantization." Pattern Recognition Letters 18.11-13 (1997): 1379-1384..
[21] Kaufman, Leonard, and Peter J. Rousseeuw. Finding groups in data: an introduction to cluster analysis. Vol. 344. John Wiley & Sons, 2009.