Customer Segmentation using Machine Learning

Varenya Gupta[1], Vikas Sani[2], Shobhit Tiwari[3], Priyansh Gaur[4], Sudhir Dawra[5]
[1,2,3,4]*Student, B.Tech. (CSE), Department of Data Science, Inderprastha Engineering College, Ghaziabad*
[5]*H.O.D., B.Tech. (CSE), Department of Data Science, Inderprastha Engineering College, Ghaziabad*

***Abstract*—Customer segmentation has become an essential strategy for businesses to enhance customer engagement, improve retention, and maximize profitability. This research delves into grouping customers from diverse organizations by analyzing behavioral traits, such as their spending habits and income levels. Behavioral segmentation proves to be a more effective method compared to other approaches, as it allows for a deeper understanding of customer preferences. Utilizing the K-means clustering algorithm, this study categorizes customers into clusters based on shared characteristics. These clusters enable organizations to craft personalized marketing strategies and targeted social media campaigns that align with individual customer interests, ultimately fostering greater engagement and driving revenue growth.**

***Index Terms* – Big Data, Classification, Clustering, Model**

I. INTRODUCTION

In today's digital age, businesses are increasingly focusing on establishing a strong online presence, making effective marketing strategies crucial for success. However, treating all customers the same by adopting a generalized marketing approach can lead to dissatisfaction and a lack of engagement. Customers expect to be treated as individuals, with marketing strategies tailored to their specific needs and preferences.

Customer segmentation has emerged as a practical solution to this challenge. This approach categorizes a company's clientele based on demographic factors, such as age, gender, and marital status, as well as behavioral aspects, including spending patterns, income levels, and product preferences. While demographic segmentation provides an overview, it often fails to capture individual customer nuances, as people of similar demographics may have vastly different interests. Behavioral segmentation addresses this limitation by focusing on individual characteristics, enabling businesses to create highly targeted marketing campaigns that resonate with customers on a personal level.

II. LITERATURE SURVEY

A. Customer Classification

In a competitive business environment, organizations aim to meet customer expectations, attract new clients, and streamline their operations. Catering to the unique requirements of each customer is challenging due to the wide diversity in preferences and behaviors. To overcome this, customer segmentation has emerged as a key strategy. This process divides customers into subgroups based on shared characteristics, allowing businesses to understand their needs better and tailor their services accordingly.

B. Big Data

Big Data, characterized by its vast volume, variety, and velocity, has revolutionized the way businesses handle information. Beyond these three core attributes, Big Data is often described using two additional Vs: veracity and value. Together, these five dimensions enable businesses to gain actionable insights. Big Data analytics supports better forecasting, cost reduction, and operational efficiency, benefiting sectors like finance, healthcare, disaster management, and national security. By leveraging Big Data, organizations can analyze customer behavior and preferences on an unprecedented scale, aiding in more precise segmentation and decision-making.

C. Data Repository

Data collection is the cornerstone of research across fields, including business and the sciences. It involves systematically gathering and measuring information to address specific questions or evaluate outcomes. For this study, the data was sourced from the UCI Machine Learning Repository, a widely used platform offering high-quality datasets for research and analysis.

D. Clustering Data

Clustering involves grouping data points in a dataset based on their similarities. Various clustering algorithms are available, each suited to different types of data and objectives. In this study, clustering was implemented using Python, leveraging its robust libraries for data analysis and visualization. Selecting the appropriate clustering algorithm is crucial for obtaining meaningful results.

E. K-means

The K-means clustering algorithm groups data into clusters based on centroids, which are dynamically adjusted as data points are assigned to them. By identifying hidden patterns in data, K-means offers valuable insights that aid decision-making. This study uses the elbow method to determine the optimal number of clusters, ensuring the results are both accurate and practical.

III. TECHNICAL INTRODUCTION

K-means Clustering Algorithm

The K-means algorithm groups data points into K clusters by measuring their similarities, typically using Euclidean distance. The process includes the following steps:

1. Randomly selecting K initial centroids.
2. Assigning each data point to the nearest centroid and recalculating the centroid positions.
3. Repeating the process until the centroids stabilize, ensuring all data points are correctly classified.

This analysis was conducted using Python 3.x within the Anaconda Jupyter Notebook environment. Python's extensive libraries for data manipulation, clustering, and visualization were employed to streamline the process.

IV. PROPOSED MODEL

A. Package and data importation:

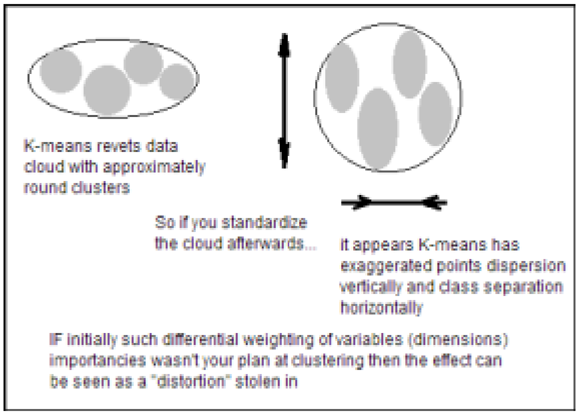
The analysis begins with importing the necessary Python packages and the dataset in Excel format. The dataset is sourced from the UCI Machine Learning Repository and stored in the same directory as the Jupyter notebook.

B. Data preparation:

Raw data often contains inconsistencies and noise, making it unsuitable for analysis in its initial state. Data cleaning and organization are essential steps to ensure meaningful insights.

C. Data normalization:

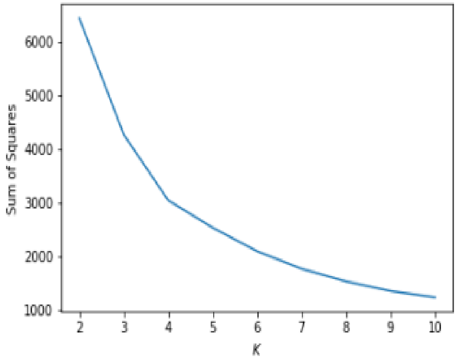
Normalization ensures that all variables contribute equally to the clustering process. For example, scaling income and spending scores to the same range prevents one variable from dominating the clustering results.


K-means revets data
cloud with approximately
round clusters
So if you standardize
the cloud afterwards...
it appears K-means has
exaggerated points dispersion
vertically and class separation
horizontally
IF initially such differential weighting of variables (dimensions)
importancies wasn't your plan at clustering then the effect can
be seen as a "distortion" stolen in

D. Determining the ideal cluster count:

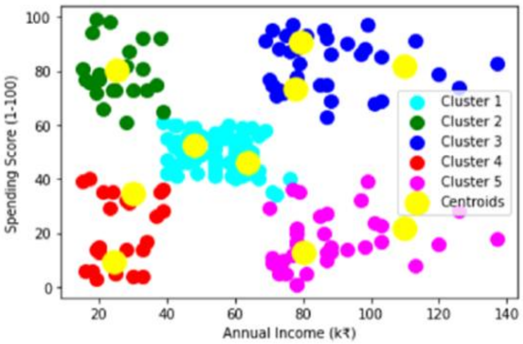
Two methods are employed to identify the optimal number of clusters:

1. Silhouette Analysis: This technique evaluates how well each data point fits within its assigned cluster compared to neighboring clusters.
2. Elbow Method: The sum of squared errors (SSE) is plotted for different cluster counts, and the "elbow" of the graph indicates the optimal number.



V. RESULTS

The analysis segmented customers into distinct clusters based on their spending habits and income levels. These clusters serve as a foundation for developing personalized marketing strategies and enhancing customer engagement. By addressing specific customer needs, businesses can increase their effectiveness and profitability.



VI. DRAWBACKS OF SYSTEM

1. Increased marketing expenses due to the need for personalized campaigns.
2. Smaller customer segments may lead to challenges in production and inventory management.

VII. CONCLUSION

This study successfully applied the K-means clustering algorithm to segment customers based on behavioral attributes. By prioritizing spending patterns and income levels, businesses can better understand their clientele and design targeted marketing strategies. The results demonstrate that K-means clustering is an effective method for addressing the complexities of customer segmentation, enabling businesses to enhance customer engagement and retention.

REFERENCES

[1] Blanchard, Tommy. Bhatnagar, Pranshu. Behera, Trash. (2019). Marketing Analytics Scientific Data: Achieve your marketing objectives with Python's data analytics capabilities. S.l: Packt printing is limited

[2] [2] Griva, A., Bardaki, C., Pramatari, K., Papakiriakopoulos, D. (2018). Sales business analysis: Customer categories use market basket data. Systems Expert Systems, 100, 1-16.

[3] Hong, T., Kim, E. (2011). It separates consumers from online stores based on factors that affect the customer's intention to purchase. Expert System Applications, 39 (2), 2127-2131.

[4] Hwang, Y. H. (2019). Hands-on Advertising Science Data: Develop your machine learning marketing strategies… using python and r. S.l: Packt printing is limited

[5] Puwanenthiren Premkanth, - Market Classification and Its Impact on Customer Satisfaction and Special Reference to the Commercial Bank of Ceylon PLC.‖ Global Journal of Management and Business Publisher Research: Global Magazenals Inc. (USA). 2012. Print ISSN: 0975-5853. Volume 12 Issue 1.

[6] Puwanenthiren Premkanth, - Market Classification and Its Impact on Customer Satisfaction and Special Reference to the Commercial Bank of Ceylon PLC.‖ Global Journal of Management and Business Publisher Research: Global Magazenals Inc. (USA). 2012. Print ISSN: 0975-5853. Volume 12 Issue 1.

[7] Sulekha Goyat. "The basis of market segmentation: a critical review of the literature. European Journal of Business and Management www.iiste.org. 2011. ISSN 2222-1905 (Paper) ISSN 2222-2839 (Online). Vol 3, No.9, 2011

[8] By Jerry W Thomas. 2007. Accessed at: www.decisionanalyst.com on July 12, 2015.

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

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

[11] Jean Yan. - Big Data, Big Opportunities- Domains of Data.gov: Promote, lead, contribute, and collaborate in the big data era. 2013. Retrieved from http://www.meritalk.com/pdfs/bdx/bdxwhitepaper-090413.pdf July 14, 2015.

[12] A.K. Jain, M.N. Murty and P.J. Flynn.‖ Data Integration: A Review‖. ACM Computer Research. 1999. Vol. 31, No. 3.

[13] Vishish R. Patel1 and Rupa G. Mehta. MpImpact for External Removal and Standard Procedures for JCSI International International Science

Issues Issues, Vol. 8, Appeals 5, No 2, September 2011 ISSN (Online): 1694-0814

[14] Jayant Tikmani, Sudhanshu Tiwari, Sujata Khedkar "Telecom Customer Classification Based on Group Analysis of K-methods", JIRCCE, Year: 2015.