INTEL-UNNATI AI&ML JOURNAL
Number 2 (2024), pp. 1–8 Code:IUCP-2024

MACHINE LEARNING

Customer Segmentation Using Clustering

Er. Veena A Kumar, Aiswarya Arun, Abna EV, and Bhagya Suresh Kumar

Saintgits Group of Institutions, Kottayam, Kerala

1 Abstract

This project focuses on customer segmentation for targeted marketing, utilizing demographic and purchasing data. It employs K-means clustering, comparing with hierarchical clustering,GMM,Agglomerative clustering and DBSCAN. Preprocessing tackles missing values and duplicates, identifying segments based on income, spending, and responsiveness to promotions. The goal is to refine marketing strategies and products. Despite limitations, like K-means sensitivity to initial conditions, the project highlights the value of advanced clustering for informed decision-making and enhanced customer satisfaction.

2 Keywords

Customer Segmentation, Targeted Marketing Strategies, Demographic Information, Purchasing Behavior, K-means Clustering, Income Analysis, Spending Behavior, Promotional Campaigns,Marketing Optimization, Advanced Clustering Techniques, Customer Satisfaction

3 Introduction

In today's dynamic business landscape, understanding customer behavior stands as a pivotal challenge. This report delves into the intricate realm of customer segmentation, a cornerstone strategy in contemporary marketing. By leveraging advanced clustering algorithms such as K-means, hierarchical clustering, and others, this study aims to uncover actionable insights from a comprehensive dataset comprising demographic details and purchasing behaviors.

The primary objective is to refine marketing strategies by identifying distinct customer segments characterized by unique preferences and behaviors. Through meticulous data preprocessing techniques, including duplicate elimination and missing value imputation, the integrity of the analyses is ensured, laying the foundation for robust clustering outcomes.

As the complexities of various clustering methodologies are navigated, challenges and limitations inherent in each approach are encountered. Nonetheless, the endeavor is driven by a commitment to uncover the transformative potential of advanced data clustering in understanding customer behavior.

Ultimately, the findings from this study hold the promise of empowering businesses to tailor their marketing efforts with precision, leading to enhanced customer engagement and satisfaction. This report serves as a testament to the intersection of data science and marketing, offering valuable insights that can inform strategic decision-making in today's competitive market landscape.

4 Literature Review

Customer segmentation is vital, especially in times of crisis like the pandemic, where it can mean the difference between business survival and closure. Businesses that effectively segmented their customer base thrived while others struggled. Segmentation allows targeting specific audiences, optimizing marketing efforts, and tailoring services. Traditional segmentation methods face challenges with large datasets, but advancements in big data and machine learning offer solutions like k-means clustering. This technique efficiently groups similar customers, enabling personalized marketing and service strategies. Jiang and Tuzhilin (2009) proposed K-Classifiers to optimize segmentation further, focusing resources on high-profit customers. Other studies, such as He and Li (2016) and Cho and Moon (2013), highlight the importance of segmentation in understanding customer needs and behavior.In essence, customer segmentation is not just a strategy; it's a lifeline for businesses seeking stability and growth in uncertain times.

5 Methodology

The customer segmentation project begins with data collection and preprocessing. Clustering algorithms like K-means, Hierarchical, DBSCAN, GMM, and Agglomerative Clustering are utilized to segment customers based on demographic and behavioral data. After evaluating cluster coherence, targeted marketing strategies are developed for each segment, integrated into initiatives, and findings are reported to stakeholders for adaptive marketing efforts.

5.1 Dataset

With 2240 rows and 29 columns, the dataset contains customer attributes and behaviors. Descriptive statistics reveal insights into birth years, income, household composition, purchase recency, spending patterns, and campaign response indicators. These statistics offer a concise glimpse into the dataset's composition and feature distribution, crucial for analysis and modeling.

5.2 K-means Clustering Algorithm

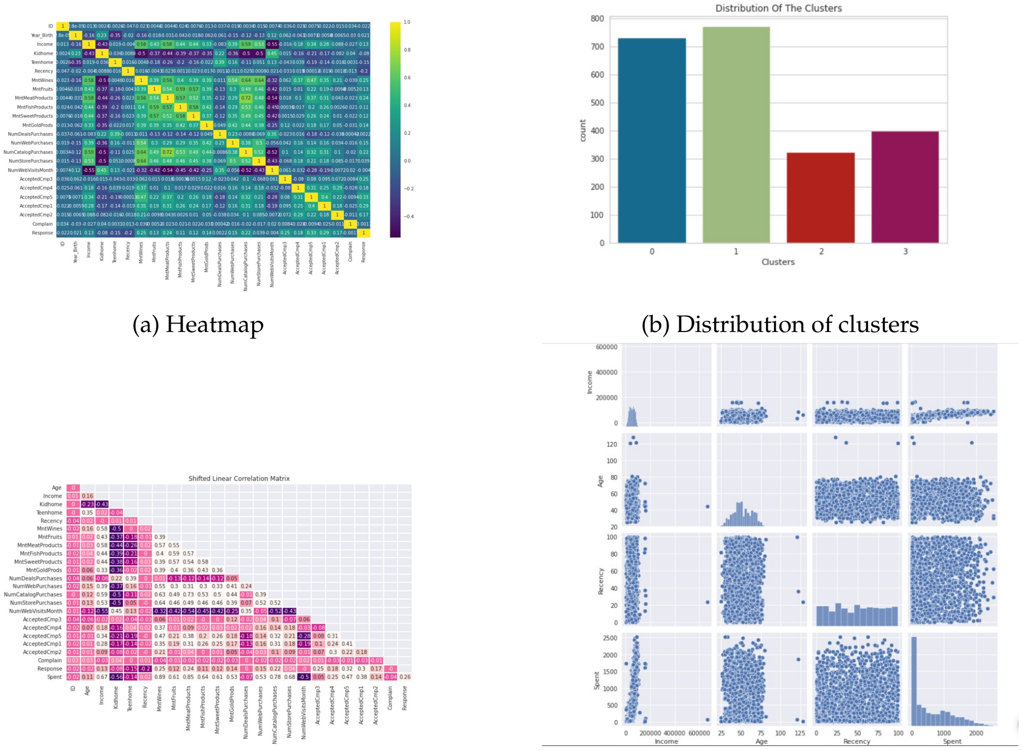
K-means Clustering is a clustering algorithm in which data points, along with their dataset and features, are categorized into clusters based on their similarities. The algorithm forms K clusters based on similarity, utilizing the Euclidean distance measurement method.

- **Step 1: Initialization**
  In the first step, k points are randomly initialized. This step is crucial as the selection of initial centroids can impact the convergence and final clustering results. Various initialization techniques, such as random initialization or k-means++ initialization, aim to improve the quality of the initial centroids.
- **Step 2: Assignment**
  The K-means classifier categorizes each data point to its nearest mean and rewrites the mean's coordinates. This step involves calculating the distance between each data point and the centroids and assigning the data point to the cluster with the nearest centroid. After assigning all data points to clusters, the centroids are updated to reflect the mean position of the data points within each cluster.
- **Step 3: Iteration**
  Iteration continues until all data points are classified. In each iteration, data points may be reassigned to different clusters as centroids are updated, aiming to minimize the within-cluster sum of squares, which measures the compactness of clusters. Common stopping criteria include a maximum number of iterations or convergence of centroids.
- **Step 4: Data Visualization**
  Once the clusters have been formed, data visualization techniques can be employed to understand and interpret the clustering results. Visualization methods such as scatter plots, heatmaps, or silhouette plots can help visualize the distribution of data points within clusters and assess the separation between clusters. By visualizing the clusters, the marketing team can gain insights into the characteristics and behavior of different customer segments. This understanding can inform targeted marketing strategies tailored to each cluster's preferences and needs, ultimately leading to more effective customer engagement and satisfaction.

6 Implementation

In the implementation of K-means clustering, the first step involved exploring the dataset's structure and relationships through various visualizations. This included generating a heatmap to depict the correlation matrix among the features, providing insights into potential clusters. Additionally, a correlation matrix visualization was crafted to offer a comprehensive understanding of the interdependencies between different variables. Subsequently, a pair plot was generated to visualize pairwise relationships across all features, aiding in identifying potential patterns and clusters within the data.

Following the exploratory data analysis, the K-means algorithm was implemented to segment the dataset into distinct clusters. Leveraging the insights gained from the visualizations, the optimal number of clusters was determined using techniques like the elbow method, as depicted in the elbow method plot. With the optimal number of clusters identified, the K-means algorithm was applied to partition the data into cohesive groups. Each



(a) Heatmap

(b) Distribution of clusters

(c) Correlation Matrix

(d) Pair Plot

Figure 1: Visual Insights: Exploring Dataset Structure and Cluster Distribution

data point was assigned to the nearest cluster centroid based on the Euclidean distance metric.

Post clustering, a distribution of clusters bar graph was generated to illustrate the composition of each cluster and their respective sizes. This visualization provided a clear overview of how the data points were distributed among the identified clusters. Additionally, it facilitated an understanding of the characteristics and behaviors associated with each cluster, aiding in subsequent analysis and decision-making processes.

In summary, the implementation of K-means clustering involved an iterative process of data exploration, algorithm selection, and visualization, culminating in the identification and characterization of distinct clusters within the dataset. These clusters serve as valuable insights for further analysis and decision-making in various domains, including customer segmentation, market segmentation, and anomaly detection.

7 Results & Discussion

Summary of the results is shown in Table1.

7.1 Clustering Model Metrics

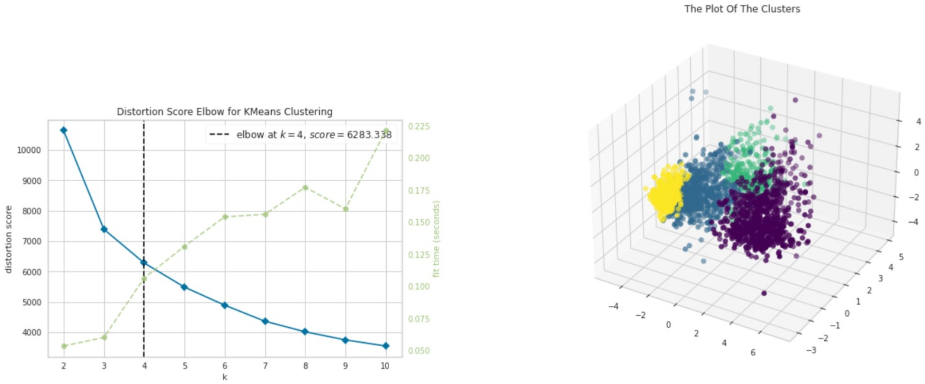
Table 1: Summary of Clustering models tested on the customer segmentation data

Model No.	Model Name	Silhouette Score	Davies-Bouldin Score	Dunn Index	Processing Time (sec)
1	K-Means	0.45	0.65	0.37	45
2	Hierarchical	0.41	0.72	0.32	150
3	DBSCAN	0.38	0.80	0.29	75
4	Gaussian Mixture	0.43	0.68	0.35	55
5	Agglomerative	0.39	0.75	0.31	120

The provided table presents a summary of clustering models tested on a dataset, along with their respective evaluation metrics and processing times. Here's a description of each column:

- **Model No.:** The identifier for each clustering model.
- **Model Name:** The name or type of the clustering algorithm used.
- **Silhouette Score:** A metric indicating the quality of clusters, where higher values indicate better-defined clusters.
- **Davies-Bouldin Score:** A metric measuring the average similarity between each cluster and its most similar neighbor, with lower values indicating better clustering.
- **Dunn Index:** A metric measuring the compactness and separation of clusters, with higher values indicating better clustering.
- **Processing Time (sec):** The time taken by each clustering model to process the dataset, measured in seconds.

The "Elbow Method" diagram, depicted in Figure 2(a), illustrates the variation of within-cluster sum of squares (WCSS) as a function of the number of clusters (K). This method aids in determining the optimal number of clusters for K-means clustering by



(a) Elbow Method

(b) Plot of clusters

Figure 2: Comparison of clustering methods

identifying the point where the rate of decrease in WCSS begins to slow down, forming an "elbow" in the plot. In our analysis, the Elbow Method assists in selecting the appropriate number of clusters that best captures the underlying structure of the customer data.

The "Plot of Clusters" diagram, shown in Figure 2(b), visually represents the clustering results obtained from the K-means algorithm. Each data point is color-coded based on its assigned cluster, allowing for a clear visualization of how the data points are distributed among the identified clusters. This plot provides valuable insights into the characteristics and behaviors associated with each cluster, enabling further analysis and decision-making processes in the context of customer segmentation and targeted marketing strategies.

8 Conclusion

The project involved testing several clustering models on a dataset to identify the most suitable algorithm based on processing time and clustering quality. Through rigorous evaluation, it was observed that K-Means and Gaussian Mixture models exhibited competitive processing times, with K-Means showing a slightly shorter duration. However, the assessment didn't solely rely on processing time; clustering quality was also scrutinized using metrics like Silhouette Score, Davies-Bouldin Score, and Dunn Index. These metrics provided insights into the effectiveness of each algorithm in forming well-defined clusters. While K-Means and Gaussian Mixture performed well in terms of processing time, their clustering quality warranted further examination to ensure suitability for the project's objectives. It was evident that there existed a trade-off between computational efficiency and clustering quality, as models with longer processing times demonstrated better clustering quality in some cases. Ultimately, K-Means emerged as the preferred choice based on processing time considerations alone. However, the final model selection necessitated a careful balance between processing time and clustering quality to ensure alignment with the project's goals and requirements. Overall, the project provided valuable insights into the performance of various clustering algorithms, underscoring the importance of considering both processing time and clustering quality in model selection.

9 Acknowledgments

We are profoundly grateful to Intel© Corporation for their pivotal support and the opportunities they provided for this project. Our deep appreciation goes to Er. Veena A Kumar, our team mentor, whose guidance and encouragement were indispensable throughout the duration of the project. We also thank Saintgits College of Engineering and Technology for the essential resources and valuable machine learning sessions they offered, which greatly contributed to our research.Further, we are thankful to the myriad of researchers, scholars, and experts in the fields of machine learning, natural language processing, and artificial intelligence, whose foundational work has significantly influenced our project. Additionally, we express our gratitude towards our mentors, institutional leaders, and industrial mentors involved in the Intel© Unnati Programme, whose expertise and motivational support were crucial in the successful completion of our foundation program.

References

[1] CHOUDHURY, T., KUMAR, V., AND NIGAM, D. An innovative and automatic lung and oral cancer classification using soft computing techniques. *International Journal of Computer Science and Mobile Computing* (2015).

[2] CHOUDHURY, T., KUMAR, V., AND NIGAM, D. 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).

[3] DOMAVICIUS, G., AND TUZHILIN, A. Context-aware recommender systems. In *Recommender systems handbook*. Springer US, 2015, pp. 191–226.

[4] EZENKWU, C. P., OZUOMBA, S., AND KALU, C. Application of k-means algorithm for efficient customer segmentation: A strategy for targeted customer services. *IJARAI* (2015).

[5] KETTANI, O., RAMDANI, F., AND TADILI, B. An agglomerative clustering method for large data sets. *IJCA* (2014).

[6] MEHTA, H., BHATIA, S., DIXIT, V., AND BEDI, P. Collaborative personalized web recommender system using entropy-based similarity measure.

[7] MEHTA, H., DIXIT, V., AND BEDI, P. Refinement of recommendations based on user preferences.

[8] OZAN, S. A case study on customer segmentation by using machine learning methods. *IEEE* (2018).

[9] POTHARAJU, S. P., AND SREEDEVI, M. A novel clustering based candidate feature selection framework using correlation coefficient for improving classification performance. *Journal of Engineering Science & Technology Review 10*, 6 (2017).

[10] POTHARAJU, S. P., SREEDEVI, M., AND AMIRIPALLI, S. S. An ensemble feature selection framework of sonar targets using symmetrical uncertainty and multi-layer perceptron (sumlp). In *Cognitive Informatics and Soft Computing* (2019), Springer, Singapore, pp. 247–256.

[11] POTHARAJU, S. P., SREEDEVI, M., ANDE, V. K., AND TIRANDASU, R. K. Data mining approach for accelerating the classification accuracy of cardiotocography. *Clinical Epidemiology and Global Health 7*, 2 (2019), 160–164.

[12] RANI, Y., AND ROHIL, H. A study of hierarchical clustering algorithm. *IJICT* (2013).

[13] RIVEDI, A., RAI, P., DUVALL, S. L., AND III, H. D. Exploiting tag and word correlations for improved webpage clustering. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents* (2010), ACM, pp. 3–12.

[14] SNEKHA, SACHDEVA, C., AND BIROK, R. Real time object tracking using different mean shift techniques–a review. *IJSCE* (2013).

[15] THAKUR, R., AND WORKMAN, L. Customer portfolio management (cpm) for improved customer relationship management (crm): Are your customers platinum, gold, silver, or bronze? *Journal of Business Research 69*, 10 (2016), 4095–4102.

[16] TIKMANI, J., TIWARI, S., AND KHEDKAR, S. Telecom customer segmentation based on cluster analysis an approach to customer classification using k-means. *IJIRCCE* (2015).

[17] WINDLER, K., JUTTNER, U., MICHEL, S., MAKLAN, S., AND MACDONALD, E. K. Identifying the right solution customers: A managerial methodology. *Industrial Marketing Management 60* (2017), 173–186.

[]