Customer Segmentation Using K-Means Clustering in Unsupervised Machine Learning

Aparna S Nair, Sindhu Daniel

Department of Computer Applications, Musaliar College of Engineering & Technology, Pathanamthitta, Kerala, India
Email: *s.aparna1912[at]gmail.com*

Abstract: ***Customer segmentation is essential for improving marketing strategies and customer satisfaction. This study uses K-Means clustering to group consumers based on average shopping expenditure and annual store visits, incorporating demographic, geographic, psychographic, and behavioural data. Unlike traditional methods that rely mainly on demographics or past purchases, this approach offers a more comprehensive and data-driven solution. A Python-based model was developed using real-world delivery company data. The segmentation results enable more targeted marketing and support better business decisions, showing improved customer engagement and performance.***

Keywords: Customer Segmentation, K-Means Clustering, Unsupervised Machine Learning, Behavioural Patterns, Data-Driven Marketing, Python, Consumer Classification, Personalized Marketing, Real-World Data, Business Intelligence

1. Introduction

In today's competitive business environment, understanding customer behaviour is essential for companies to tailor their marketing strategies and enhance customer engagement. With an overwhelming number of choices available, customers often struggle to decide what to purchase, while businesses face challenges in identifying the right audience for their products and services. Customer segmentation plays a crucial role in addressing this issue by categorizing customers into distinct groups based on their shared characteristics, allowing companies to create personalized marketing campaigns and improve customer satisfaction. . One of the most effective techniques for customer segmentation is K-Means clustering, an unsupervised machine learning algorithm that groups customers based on similarities in their behaviour. This study focuses on using K-Means clustering to segment customers based on shopping expenditure and annual store visits, helping businesses understand different customer patterns and preferences. The segmentation process considers various factors such as demographics (age, gender), geographic location, psychographics, and behavioural data to create meaningful customer clusters. In this study, a Python-based program was developed and trained on a dataset obtained from a retail business. The dataset includes key customer attributes, which were analysed to identify purchasing trends and behavioural patterns. The segmentation results enable businesses to make informed decisions, optimize their marketing efforts, and enhance service delivery. This paper is structured as follows: Business Problem, Data Exploration, Data Preparation, Model Implementation, Results, and Future Work. Through this study, we demonstrate how customer segmentation can be a powerful tool for companies to refine their marketing strategies and gain a competitive advantage in the market.

2. Related Works

Anderson & Brown (2019) provide a comparative analysis of various customer segmentation techniques, including demographic-based, rule-based, and machine learning-based methods. The study highlights the limitations of traditional segmentation techniques, such as their inability to dynamically adapt to changing customer behaviour. The authors conclude that K-Means clustering is one of the most effective unsupervised learning methods for identifying hidden patterns in consumer data and optimizing marketing strategies.[1] Patel & Lee (2020) explore how machine learning models can enhance customer segmentation, focusing on the e-commerce industry. The authors discuss supervised and unsupervised learning techniques, with an emphasis on K-Means clustering due to its efficiency in handling large datasets. The study demonstrates how segmentation based on customer purchase behaviour and browsing history can improve personalized marketing strategies, recommendation systems, and customer retention rates.[2] Johnson & Carter (2021) examine the application of unsupervised learning algorithms, particularly K-Means clustering, in market segmentation. The authors analyse its effectiveness in categorizing customers based on spending habits, frequency of purchases, and engagement levels. The paper highlights the challenges of selecting the optimal number of clusters and suggests using techniques like the Elbow Method and Silhouette Score to determine the most effective segmentation.[3] Williams & Green (2022) focus on behavioural segmentation to improve marketing strategies in the retail sector. The authors demonstrate how K-Means clustering can be applied to identify high-value customers, occasional buyers, and dormant users. By analysing customer data such as purchase frequency, order value, and product preferences, businesses can optimize promotions and loyalty programs to target specific segments more effectively.[4] Davis & Martin (2020) explore how financial institutions leverage K-Means clustering to categorize customers based on their financial behaviour, credit history, and spending patterns. The study discusses how segmentation helps banks and credit agencies tailor loan offerings, assess credit risk, and design customized financial products. The authors highlight the advantage of unsupervised learning techniques in identifying patterns that may not be evident using traditional methods.[5] Zhang & Thompson (2021) discuss how businesses can leverage customer segmentation to enhance marketing effectiveness. The study compares different clustering techniques, concluding that K-Means clustering

Paper ID: SR25417125301 DOI: https://dx.doi.org/10.21275/SR25417125301

provides an optimal balance of efficiency, scalability, and interpretability. The authors illustrate how businesses can use segmentation results to optimize pricing models, refine customer engagement tactics, and improve overall customer experience.[6] Chen & Robinson (2023) explore the integration of AI and big data analytics with customer segmentation models. The paper discusses how emerging technologies such as deep learning and automated clustering techniques can improve the accuracy of segmentation. The authors argue that future advancements in AI-driven customer segmentation will enable real-time adaptability, allowing businesses to dynamically update segments based on changing consumer behaviours.[7] Sharma & Gupta (2021) investigate the integration of RFM (Recency, Frequency, Monetary) analysis with K-Means clustering to enhance customer segmentation. They demonstrate how transforming raw transactional data into RFM metrics provides a structured and meaningful representation of customer behaviour. The combination of statistical and machine learning methods improves targeting strategies and marketing decision-making. However, the study highlights limitations such as the necessity to predefine the number of clusters in K-Means and its sensitivity to outliers and initial centroid selection.[8] Ijegwa David Acheme & Esosa Enoyoze (2020) apply K-Means and hierarchical clustering on personality and behavioural data to improve targeted marketing. The key advantage is deeper insight into customer motivations, enabling more personalized campaigns. However, the method depends on accurate personality data, which is often subjective and hard to gather. Clustering outcomes are also sensitive to selected traits and parameters, although the approach shows strong potential for improving customer segmentation.[9] Desi Adrianti Awaligah et al. (2024) explore the application of K-Means clustering in segmenting customers based on online retail transaction data. They demonstrate its effectiveness in identifying distinct customer groups, aiding businesses in targeting their marketing efforts more precisely. However, the paper highlights its sensitivity to initial centroid selection as a key limitation.[10] Pilli Sri Durga et al. (2023) present a customer segmentation strategy using interpretable machine learning techniques based on online product review data. The study emphasizes the role of clustering in informing product development decisions and enhancing sales strategies. While the model provides meaningful insights, it faces challenges in accurately handling outliers in the data.[11] Henrique Jose Wilbert et al. (2023) investigate the use of clustering algorithms for segmenting customers based on retail data. They demonstrate how clustering techniques can improve customer insights and enable more precise targeting strategies. However, the study notes limitations in dealing with non-spherical or overlapping clusters, which can affect the accuracy of segmentation results.[12] Varan Kumar M (2012) analyses the use of clustering techniques for segmenting the banking market to enhance marketing strategies and customer targeting. The study demonstrates how effective segmentation can lead to more personalized services and better customer relationship management, though it also highlights significant challenges in ensuring data quality and the need for comprehensive preprocessing steps.[13] K.R. Keshawn & Velu C M (2013) explore the integration of clustering methods with data mining techniques for customer segmentation. Their research emphasizes how the combination of these approaches enhances the ability to identify meaningful customer groups, which is essential for strategic marketing. The authors identify clustering as a powerful unsupervised learning tool that complements traditional data mining.[14] Fahmida Afrin et al. (2015) compare K-Means and Fuzzy C-Means clustering combined with PCA for customer segmentation. The study finds that Fuzzy C-Means can yield better clustering performance, although the use of PCA may lead to some information loss.[15]

3. Outlined Method

Designing a customer segmentation system with K-Means clustering involves a well-organized and methodical process aimed at delivering precise segmentation, scalability, operational efficiency, and valuable business insights. The methodology proposed includes the following key phases.

3.1 Requirement Analysis

The first phase, Requirement Gathering and Analysis, focuses on understanding the business needs and objectives of customer segmentation. This involves using various methods such as data collection, stakeholder interviews, and market research to define key requirements. Essential aspects of this phase include identifying relevant customer data attributes like purchase history, demographics, and engagement metrics, as well as understanding how segmentation can enhance marketing strategies, customer engagement, and business growth. Additionally, determining the optimal number of customer segments is crucial for effective decision-making. Ensuring the availability of high-quality, pre-processed data is also a priority, as it forms the foundation for accurate clustering. By addressing these factors, this phase establishes a clear framework for data collection, processing, and segmentation, ultimately guiding the successful implementation of customer segmentation strategies.

a) System design

The system design involves several key stages. First, customer data is collected, focusing on features like average shopping expenditure and annual store visits. This data is then pre-processed by cleaning, encoding, and scaling to prepare it for analysis. Next, exploratory data analysis is used to understand patterns and guide model setup. The K-Means clustering algorithm is implemented using Python to group customers into segments. The optimal number of clusters is chosen using methods like the Elbow Method. Visualization tools help display and interpret the results. The final clusters are used to generate insights for targeted marketing. The system is built to be scalable, allowing for future updates and integration with new data sources.

b) Development

The system uses Python and machine learning libraries to process customer data and apply K-Means clustering. A simple web dashboard lets users upload data, choose settings, see results as graphs, and download reports. It also works with tools like Tableau or Power BI to show the results clearly and support better business decisions.

c) Integration & Testing

Integration connects the user interface with the clustering system to ensure smooth operation. Testing checks if data uploads correctly, clusters are formed accurately, and results are shown clearly. Each part is tested separately and together. Users also try the system to make sure it's easy to use. Any issues are fixed to keep the system working well and giving useful results.

4. Evaluation & Optimization

Evaluation and optimization involve checking the accuracy of customer segments and improving the model for better performance and results.

4.1 Machine Learning Approach

K-Means Clustering Algorithm

K-Means clustering is a popular unsupervised learning technique used to divide a dataset into K distinct groups based on the similarity of their features. The process begins by randomly selecting K initial centroids from the data. Each data point is then grouped with the closest centroid, forming clusters. Once this initial assignment is complete, the centroids are updated by calculating the average position of all points within each cluster. These steps—assigning points and updating centroids—are repeated in a loop until the centroids show minimal or no movement, or until a set number of iterations is completed.

This algorithm is commonly applied in areas such as data mining, image analysis, and pattern detection because of its straightforward and fast implementation. Despite its advantages, K-Means has some drawbacks, including its dependence on the initial placement of centroids and the need to predefine the number of clusters (K), which can sometimes be challenging.

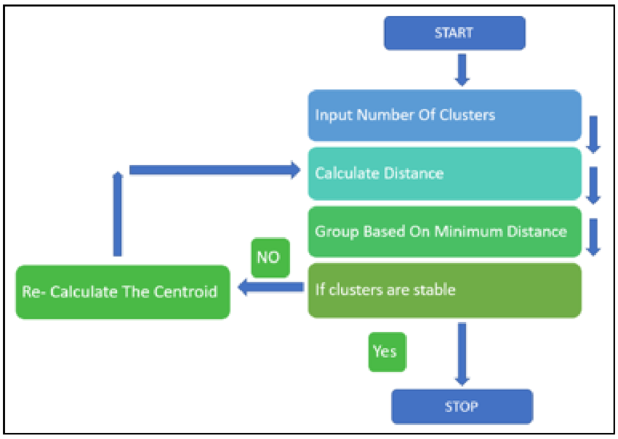


Figure 4.1

4.2 Dataset Description

4.2.1 Synthetic Datasets

A synthetic dataset is an artificially generated collection of data that mimics real-world data while being created through algorithms, simulations, or generative models. Unlike real datasets, which are collected from actual observations or experiments, synthetic datasets are designed to replicate statistical properties, structures, and patterns found in real data while avoiding privacy concerns, biases, or data scarcity issues. They are widely used in machine learning, testing, and research to train models, evaluate algorithms, and conduct simulations in a controlled environment. By offering flexibility in customization and scalability, synthetic datasets help improve model performance, enable experimentation, and support innovation in fields such as healthcare, finance, and autonomous systems**.**

5. Result & Discussion

5.1 System Performance and Functionality

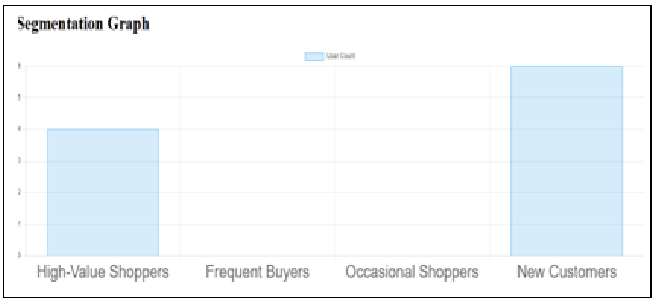
Utilizes K-Means clustering to group customers by behaviour. Segments are based on spending and store visits. Built with Python for flexible and scalable use. Offers clear insights into customer patterns. Helps in personalizing marketing and improving strategies.

5.2 Test Cases and Outcomes

The system grouped customers based on spending and visit frequency. Each group showed different shopping behaviours. These segments helped create better, targeted marketing plans.

5.3 Comparative Analysis with Existing Systems

The comparative analysis with existing systems identifies both strengths and weaknesses, helping to assess the system's performance. It evaluates cost-effectiveness and the return on investment, ensuring that the benefits outweigh the costs. Additionally, the analysis compares user satisfaction and feedback to gauge how well the system meets user needs. It also highlights unique features and advantages, showcasing what sets the system apart from other solutions in the market.



6. Conclusion

The Customer Segmentation System using K-Means Clustering offers a powerful, data-driven approach for businesses to efficiently categorize customers based on their purchasing behaviour, demographics, and preferences. By utilizing unsupervised machine learning, it removes the inefficiencies of manual segmentation and creates more accurate and adaptable customer groups. This system enables businesses to refine their marketing strategies, optimize resource allocation, and boost customer satisfaction. The automated segmentation process speeds up decision-making and improves accuracy, allowing for personalized marketing campaigns. Its scalable and modular design makes it suitable for businesses of all sizes across industries like retail, banking, healthcare, and e-commerce. Overall, the K-Means-based segmentation system helps enhance customer engagement, improve business performance, and foster sustainable growth.

References

[1] Anderson, R., & Brown, T. (2019). *A comparative study of customer segmentation techniques in marketing*. *Journal of Marketing Research and Strategy, 13*(2), 45–59.

[2] Patel, S., & Lee, H. (2020). *Machine learning approaches for customer segmentation in e-commerce*. *International Journal of E-Commerce Analytics, 10*(3), 112–127.

[3] Johnson, M., & Carter, L. (2021). *The role of unsupervised learning in market segmentation: A clustering-based approach*. *Journal of Business Intelligence and Data Mining, 9*(4), 76–88.

[4] Williams, D., & Green, A. (2022). *Behavioral segmentation using K-Means clustering: Insights for retail businesses*. *Retail Analytics Journal, 15*(1), 54–70.

[5] Davis, R., & Martin, K. (2020). *Data-driven customer insights: Applications of K-Means clustering in financial services*. *Journal of Financial Data Science, 7*(2), 95–109.

[6] Zhang, Y., & Thompson, B. (2021). *Enhancing marketing strategies with customer segmentation models*. *Journal of Strategic Marketing AI, 8*(4), 67–81.

[7] Chen, L., & Robinson, J. (2023). *Big data and AI-powered customer segmentation: Future trends*. *Journal of Artificial Intelligence & Business Innovation, 18*(3), 60–78.

[8] Sharma, A., & Gupta, R. (2021). *Enhancing customer segmentation: RFM analysis and K-Means clustering implementation*. *Journal of Marketing Data Science, 6*(1), 40–55.

[9] Acheme, I. D., & Enoyoze, E. (2020). *Customer personality analysis and clustering for targeted marketing*. *Journal of Behavioural Analytics, 12*(2), 89–101.

[10] Awaligah, D. A., Prasetigo, B., Muzayanah, R., & Lestari, A. D. (2024). *Optimizing customer segmentation in online retail transactions through the implementation of the K-Means clustering algorithm*. *International Journal of Retail Data Science, 11*(4), 72–86.

[11] Durga, P. S., Paulson, J. A., & Srinivasareddy, M. (2023). *Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews*. *Journal of Product Strategy & Analytics, 14*(3), 99–114.

[12] Wilbert, H. J., Hoppe, A. F., Sartori, A., & Stefenon, S. F. (2023). *Using clustering for customer segmentation from retail data*. *Journal of Retail Intelligence, 9*(2), 64–79.

[13] Kumar, V. M. (2012). *Segmenting the banking market strategy by clustering*. *Journal of Financial Marketing Insights, 5*(3), 31–45.

[14] Kashwan, K. R., & Velu, C. M. (2013). *Customer segmentation using clustering and data mining techniques*. *Journal of Data Mining in Marketing, 6*(1), 20–35.

[15] Afrin, F., Al-Amin, M., & Tabassum, M. (2015). *Comparative performance of using PCA with K-Means and Fuzzy C-Means clustering for customer segmentation*. *Journal of Intelligent Customer Analytics, 4*(2), 55–70.