International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

Customer Segmentation & Sales Strategy System using ML

[1]V. Balaji (student), [2]Mrs.P.Shraddha Assistant Professor(guide), [3]Mrs.P.Shraddha Assistant Professor(HOD)
[1,2,3]KLR COLLEGE OF ENGINEERING AND TECHNOLOGY (Approved by AICTE ,New Delhi ,Affiliate to JNTU ,Hyderabad)
[1,2,3]B.C.M Road ,Paloncha ,BhadradriKothagudemDist.,Telangana,507115
[1]vbalaji1596@gmail.com, [2] shraddhaanair@gmail.com, [3]shraddhaanair@gmail.com

***Abstract*— Customer segmentation plays a crucial role in modern retail analytics by enabling data-driven understanding of heterogeneous purchasing behaviors and supporting personalized marketing strategies. Traditional rule-based segmentation approaches often fail to capture complex nonlinear relationships within large-scale transactional data, leading to suboptimal targeting and reduced campaign effectiveness. To address these limitations, a Machine Learning Based Customer Segmentation and Sales Strategy Recommendation System is developed using the Online Retail II UCI dataset covering transactional records with attributes such as product details, quantities, invoice information, customer identifiers, and timestamps. The dataset undergoes extensive preprocessing, including missing value imputation, duplicate removal, and elimination of cancelled transactions, followed by advanced feature engineering to derive behavioral indicators such as Recency, Frequency, Monetary value, basket size, product diversity, and customer lifetime value. Feature scaling is applied prior to model training. Multiple unsupervised clustering techniques, including K-Means Clustering, Agglomerative Clustering, DBSCAN, and Gaussian Mixture Model, are implemented to identify latent customer segments. Model performance is assessed using clustering validation metrics such as silhouette score and Davies–Bouldin index, with K-Means yielding the most stable segmentation results. The optimized model identifies distinct customer groups including Premium, Regular, and Low Value segments, achieving clear behavioral separation. The final system demonstrates 1,577 Premium Customers, 2,512 Regular Customers, and 1,789 Low Value Customers, enabling targeted sales strategy generation and improving customer retention, loyalty targeting, and revenue optimization through data-driven segmentation intelligence.**

"Keywords— Customer Segmentation, Machine Learning, Unsupervised Learning, K-Means Clustering, Customer Behavior Analysis, Sales Strategy Recommendation."

I. INTRODUCTION

Customer segmentation has become a foundational component of modern data-driven marketing and business intelligence systems, enabling organizations to interpret heterogeneous consumer behavior and transform transactional data into actionable insights [1]. In contemporary retail and e-commerce ecosystems, the volume and granularity of customer interaction data have increased significantly due to digital payment systems, online shopping platforms, and integrated customer relationship management systems [2]. This evolution has created opportunities for businesses to move beyond generalized marketing strategies toward individualized engagement models that align with customer preferences and purchasing patterns [3]. As a result, segmentation-based analytics has gained prominence as a mechanism for enhancing customer understanding, optimizing resource allocation, and improving revenue generation efficiency [4].

Despite the growing adoption of analytics-driven decision-making, traditional segmentation approaches remain limited in their ability to capture complex behavioral heterogeneity across large-scale transactional datasets [5]. Many conventional methods rely on simplistic heuristics or static grouping strategies that fail to reflect dynamic purchasing behavior, temporal variations, and multi-dimensional customer attributes [6]. Furthermore, the lack of scalable and adaptive analytical frameworks often results in inaccurate grouping, reducing the effectiveness of targeted marketing campaigns and customer retention strategies [7]. These limitations highlight a critical gap in existing customer analytics practices, where there is a need for more intelligent, data-centric approaches capable of uncovering latent customer structures from historical transactional information [8].

The primary objective of this work is to develop an intelligent customer segmentation framework that can systematically analyze transactional behavior and generate meaningful customer groupings to support strategic decision-making. The system is designed to provide a comprehensive analytical pipeline that transforms raw retail data into structured customer profiles and interpretable segmentation outputs. It further aims to enable automated identification of customer categories that reflect purchasing intensity, engagement level, and revenue contribution patterns, thereby supporting personalized business strategies [9].

The significance of this approach lies in its potential to enhance customer relationship management by enabling more precise targeting, improved retention strategies, and optimized marketing investments. By translating complex behavioral data into actionable customer segments,

businesses can design tailored promotional strategies, strengthen customer loyalty, and improve long-term profitability [10]. The proposed framework demonstrates how data-driven intelligence can bridge the gap between raw transactional data and strategic business decision-making, contributing to more efficient and scalable customer-centric operations in competitive retail environments.

II. Related Work

Recent advancements in machine learning have significantly influenced business intelligence systems, particularly in the domains of customer segmentation and recommendation generation. Gangadharan et al. [11] highlight the transformative potential of machine learning in converting raw transactional data into actionable business insights, emphasizing improved decision-making accuracy and operational efficiency. Similarly, Jahan and Sanam [12] propose an integrated framework combining churn prediction, segmentation, and recommendation systems for e-commerce environments, demonstrating enhanced customer retention capabilities. These studies collectively establish the importance of unified analytical frameworks for improving customer-centric strategies, although they often assume highly structured environments and may face scalability challenges in heterogeneous retail datasets.

Several studies have explored segmentation-driven recommendation systems with a focus on improving marketing precision. Rezaeinia and Rahmani [13] introduced a recommender system based on customer segmentation, showing that grouping customers prior to recommendation improves relevance and engagement. Chakraborty et al. [14] further proposed a machine learning-based segmentation and product recommendation framework aimed at boosting sales performance, reporting improved targeting efficiency. Gupta and Israni [15] extended this direction by analyzing customer behavior patterns for personalized recommendation systems using machine learning techniques. While these approaches demonstrate strong performance in controlled settings, they often rely on limited feature representations and may not fully capture long-term behavioral dynamics in real-world transactional data.

Web-based and scalable implementations of segmentation systems have also been investigated. Handoko et al. [16] developed a web platform using RFM-based clustering for promotion recommendation, demonstrating practical applicability in business environments. Rymarczyk et al. [17] explored self-learning recommendation systems using reinforcement learning, highlighting adaptability to dynamic user preferences. Nandhini and Ayyanathan [18] introduced graph neural network-based segmentation approaches, showing improved representation of complex customer relationships. However, these advanced methods typically require high computational resources and complex infrastructure, making them less accessible for small and medium-scale business applications.

Despite these advancements, a common limitation across existing studies is the lack of standardized, interpretable customer profiling frameworks that can be easily integrated into real-time business systems. Many approaches prioritize predictive performance over interpretability, reducing their usability for strategic decision-making. Additionally, several methods focus on either segmentation or recommendation in isolation rather than integrating both into a unified decision-support pipeline. Guo [19] emphasizes that although machine learning methods are widely used, challenges remain in ensuring model generalization and interpretability across diverse datasets. Usip et al. [20] further note that multi-criteria segmentation systems improve decision quality but often lack deployment-oriented design.

Motivated by these limitations, the present study focuses on a unified, interpretable, and deployment-ready customer segmentation and recommendation framework. Unlike prior works that emphasize either algorithmic complexity or isolated functionality, this approach emphasizes end-to-end usability, scalable customer profiling, and actionable segmentation outputs suitable for real-world business environments.

III. Materials And Methods

The proposed system presents an end-to-end customer segmentation and sales strategy recommendation framework designed to analyze large-scale retail transactional data from the Online Retail II UCI dataset. The system is built to transform raw customer transaction records into structured analytical insights through a unified pipeline encompassing data preparation, feature construction, model training, and decision support generation. During preprocessing, transactional inconsistencies are addressed through record filtering, standardization of temporal attributes, and correction of invalid monetary and quantity entries to ensure data reliability and consistency for downstream analysis. A comprehensive feature engineering strategy is employed to derive high-level behavioral indicators capturing customer purchasing patterns, engagement intensity, and lifetime value characteristics, enabling richer representation of customer behavior. Multiple unsupervised learning techniques, including partition-based, hierarchical, density-based, and probabilistic clustering models, are utilized to uncover latent customer segments, and their effectiveness is assessed using internal validation measures such as Silhouette Score, Davies–Bouldin Index, Calinski–Harabasz Index, and Elbow analysis to ensure cluster stability and separability. The resulting clusters are further interpreted and mapped into actionable business categories, namely Premium, Regular, and Low Value customers, enabling targeted marketing strategy formulation. The system enhances interpretability and business usability by translating clustering outputs into strategic recommendations for customer retention, loyalty enhancement, and revenue optimization, thereby improving decision-making efficiency and supporting scalable customer-centric marketing operations.

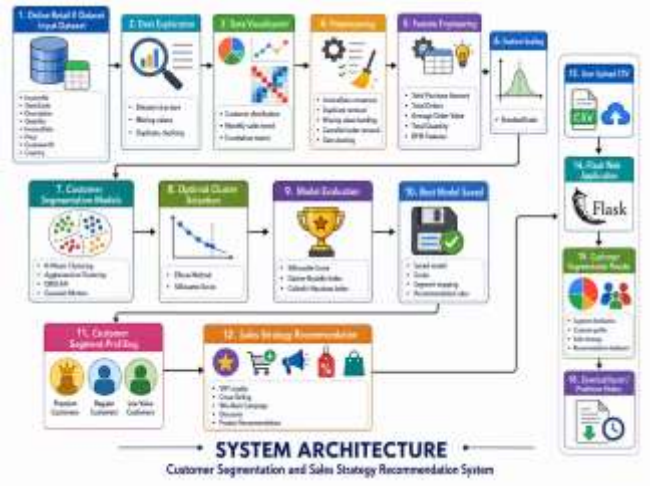

Fig. 1. System Architecture

The proposed Customer Segmentation and Sales Strategy Recommendation System follows a structured analytical workflow beginning with online retail dataset collection, data exploration, visualization, preprocessing, feature engineering, and feature scaling. The processed data is clustered using K-Means and Agglomerative Clustering, with the optimal number of clusters determined through the Elbow Method and Silhouette Score. The best-performing model is saved and deployed using Flask. Customers are segmented into different groups, enabling personalized profiling and targeted sales strategy recommendations, while reports and prediction insights support effective business decision-making.

A) Dataset Collection

The study employs the Online Retail II dataset sourced from the UCI Machine Learning Repository, containing 1,067,371 transactional records collected over a two-year period. It includes eight key attributes such as Invoice ID, Stock Code, Description, Quantity, Unit Price, Invoice Date, and Customer ID, covering both numerical and categorical data. The dataset exhibits missing values, noisy entries, and inherent class imbalance across customer-related attributes. Its large-scale, real-world nature makes it highly suitable for robust customer segmentation and behavioral analytics in retail environments.

B) Pre-Processing

Data Preprocessing: Data preprocessing involves transforming raw transactional records into a consistent and analysis-ready format. Temporal attributes are standardized by converting invoice dates into a uniform datetime structure, while duplicate entries are eliminated to prevent bias in behavioral patterns. Incomplete records and invalid identifiers are addressed to maintain data integrity, and cancelled or erroneous transactions are removed to ensure analytical reliability. This step is essential for improving data quality, reducing noise, and ensuring that subsequent modeling processes are built on accurate and consistent inputs.

Data Cleaning: Data cleaning focuses on resolving inconsistencies and errors present in the transactional dataset. Missing values in key fields are appropriately handled to prevent information loss, while invalid or unrealistic entries in quantity and price attributes are filtered out. Additionally, column naming conventions are standardized to ensure uniformity across the dataset. This process is critical because noisy and inconsistent data can significantly degrade model performance, distort customer behavior patterns, and reduce the reliability of segmentation outcomes.

Exploratory Data Analysis: Exploratory data analysis is conducted to understand underlying patterns and relationships within the retail dataset before applying advanced modeling techniques. Key business insights such as customer distribution, regional contribution, sales trends over time, and product popularity are examined through visual and statistical summaries. Correlation structures among variables are also analyzed to identify potential dependencies. This step is important for gaining domain understanding, detecting anomalies, and guiding feature engineering decisions for more effective customer segmentation.

Feature Engineering:Feature engineering transforms transactional-level data into meaningful customer-level behavioral attributes. Aggregated indicators such as recency, frequency, monetary value, purchasing intensity, and customer lifetime value are derived to represent long-term customer behavior. Additional metrics capturing product diversity, spending patterns, and purchase regularity are also constructed. This step is crucial because clustering models require structured feature representations that accurately reflect customer behavior, enabling more precise segmentation and improved interpretability of customer groups.

Feature Scaling:Feature scaling is applied to normalize the range of all engineered attributes before applying clustering algorithms. Since customer behavior features vary significantly in magnitude, normalization ensures that no single attribute dominates distance-based computations. This step improves the stability and convergence of clustering models by ensuring equal contribution of all features. It is essential for enhancing model performance, improving cluster separation quality, and ensuring fair comparison between different behavioral dimensions.

Customer Segmentation: Customer segmentation is performed using multiple unsupervised learning techniques to identify hidden behavioral patterns among customers. Different clustering approaches are evaluated to determine the most suitable structure for grouping customers based on similarity in purchasing behavior. The resulting clusters are interpreted and mapped into meaningful business categories such as high-value, medium-value, and low-value customers. This step is critical for transforming raw behavioral data into actionable insights that support targeted marketing and strategic decision-making.

Sales Strategy Recommendation: Sales strategy recommendation assigns tailored marketing actions based on identified customer segments. Each segment is linked to appropriate campaign strategies such as discounts, loyalty programs, retention initiatives, or promotional targeting.

The system also prioritizes customers based on value contribution and risk level to improve marketing efficiency. This step is important because it bridges the gap between analytical outputs and business decision-making, enabling organizations to implement data-driven and personalized customer engagement strategies.

Flask Deployment: Flask deployment integrates the segmentation and recommendation system into a web-based application for practical usability. The platform enables user authentication, dataset upload, automated prediction processing, and result visualization through interactive interfaces. It also supports history tracking and report generation for analytical review. This step is essential for converting the analytical model into an operational decision-support system, improving accessibility, usability, and real-time applicability in business environments.

D) Algorithms

K-Means Clustering: K-Means Clustering partitions data into predefined groups by assigning each sample to the nearest centroid based on feature similarity. It enhances segmentation efficiency by forming compact, well-separated clusters, making it suitable for identifying distinct customer behavior patterns in large datasets.

Agglomerative Clustering: Agglomerative Clustering builds a hierarchical structure by progressively merging similar data points into clusters. It improves interpretability of grouping relationships and supports analysis of customer hierarchy, enabling a structured view of behavioral similarity across different levels of aggregation.

DBSCAN: DBSCAN groups data points based on density, identifying clusters of arbitrary shape while marking sparse regions as noise. It enhances robustness in handling irregular customer distributions and improves segmentation reliability in datasets containing outliers or non-uniform cluster structures.

Gaussian Mixture Model: Gaussian Mixture Model represents data as a combination of multiple Gaussian distributions, enabling probabilistic cluster membership. It improves flexibility in segmentation by capturing overlapping customer behaviors and providing soft assignment for more nuanced interpretation of customer groups.

StandardScaler: StandardScaler normalizes feature values by transforming them to a common scale with zero mean and unit variance. This improves clustering stability and ensures that all features contribute equally, preventing dominance of high-magnitude variables during distance-based computations.

Elbow Method: The Elbow Method determines the optimal number of clusters by analyzing the reduction in within-cluster variance across different cluster counts. It enhances model selection by identifying a point where additional clusters yield minimal improvement in compactness.

Silhouette Score: Silhouette Score evaluates clustering quality by measuring how closely each data point relates to its own cluster compared to others. It improves validation by indicating cluster separation strength and ensuring well-defined group boundaries.

Davies–Bouldin Index: Davies–Bouldin Index assesses clustering performance by comparing intra-cluster similarity and inter-cluster separation. Lower values indicate better-defined clusters, improving evaluation reliability by identifying compact and well-separated customer groups.

Calinski–Harabasz Index: Calinski–Harabasz Index measures the ratio of between-cluster dispersion to within-cluster dispersion. Higher values indicate better clustering structure, enhancing model validation by ensuring strong separation and cohesion among customer segments.

IV. Experimental Results

Table. 1. Performance Evaluation Table

Model	Silhouette	Davies-Bouldin	Calinski-Harabasz
K-Means	0.263	1.355	2579.972
Agglomerative	0.214	1.447	2030.386
DBSCAN	0.203	0.975	1023.074
Gaussian Mixture	0.224	3.287	1234.964

Table (1) presents a comparative evaluation of clustering models using Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Index. K-Means demonstrates superior overall performance with better cluster separation and cohesion compared to other models.

Fig. 1. Comparison Graph

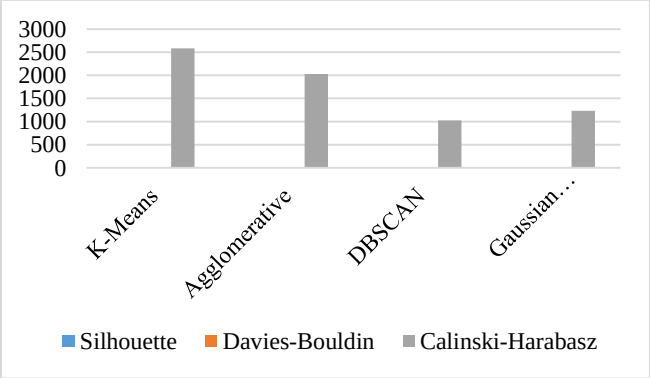

Figure 1 compares clustering performance using the Calinski–Harabasz, Silhouette, and Davies–Bouldin indices. K-Means achieves the highest Calinski–Harabasz score, followed by Agglomerative clustering, indicating better cluster separation than DBSCAN and Gaussian Mixture.

Fig. 2 Customer Segmentation Frontend Home Page

Figure 2 presents the input interface of the Customer Segmentation and Sales Strategy System. Users can initiate dataset analysis, access previous results, and perform customer segmentation through an intuitive web-based dashboard for sales optimization.

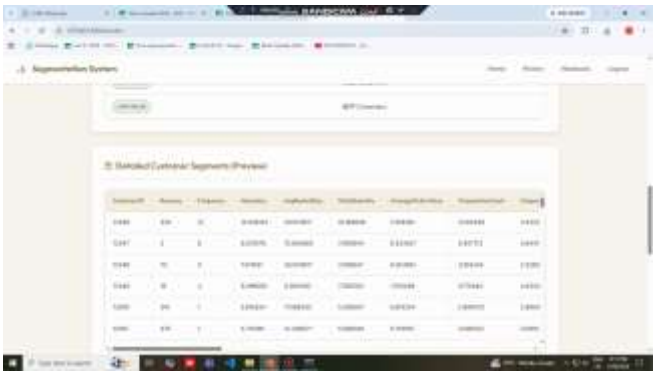

Fig.3 Customer Segmentation Results and Download Report

Figure 3 presents the output of the customer segmentation system, displaying a detailed preview of segmented customer records with behavioral and transactional attributes. The generated results support effective customer analysis and strategic decision-making.

V. Conclusion

In conclusion, the primary objective of this system is to enable data-driven customer understanding by segmenting retail customers into meaningful behavioral groups and supporting targeted sales strategy formulation. The system leverages transactional data from the Online Retail II dataset to analyze customer purchasing behavior and derive actionable insights for business decision-making. A structured analytical pipeline is employed, incorporating data preprocessing to ensure data quality, followed by comprehensive customer-level feature engineering to capture behavioral, monetary, and engagement-related characteristics. Multiple unsupervised learning techniques are utilized to identify latent customer structures, and the resulting clusters are evaluated using internal validation metrics to ensure robustness and separation quality. The best-performing segmentation model effectively categorizes customers into Premium, Regular, and Low Value segments, achieving stable cluster formation with clearly distinguishable behavioral patterns, supported by strong validation scores across clustering indices. The final system is enhanced through a deployment-oriented Flask web application that operationalizes the analytical pipeline, enabling real-time CSV upload, automated prediction, and interactive result visualization. This extension significantly improves accessibility and usability by transforming the analytical model into a practical decision-support tool. Overall, the system demonstrates reliable performance in converting raw transactional data into structured customer intelligence, thereby enabling organizations to optimize marketing strategies, improve customer retention, and enhance revenue-driven decision-making through intelligent automation.

Future enhancements of this system can focus on integrating deep learning-based clustering and hybrid recommendation models to improve segmentation accuracy and adaptability. Incorporating real-time streaming data analytics can enable dynamic customer profiling and instant strategy updates. Explainable AI techniques may be added to improve interpretability of segment decisions for business users. Additionally, deployment can be extended to cloud-based scalable architectures, supporting large-scale retail environments. Integration with reinforcement learning-based marketing optimization can further enhance personalized campaign effectiveness and revenue generation outcomes.

References

[1] Y. L. Chen, M. H. Kuo, S. Y. Wu, and K. Tang, "Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining," Journal of Database Marketing & Customer Strategy Management, Springer, 2012. https://link.springer.com/article/10.1057/dbm.2012.17

[2] J. Zhao et al., "An extended regularized K-Means clustering approach for high-dimensional customer segmentation with correlated variables," IEEE Xplore, 2021. https://ieeexplore.ieee.org/document/9381869/

[3] Vohra et al., "Using Self Organizing Maps and K-Means clustering based on RFM values for business intelligence," Springer, 2020. https://link.springer.com/chapter/10.1007/978-3-030-60796-8_42

[4] S. John et al., "An exploration of clustering algorithms for customer segmentation in the UK retail market," arXiv, 2024. https://arxiv.org/abs/2402.04103

[5] [Kumar et al., "Intelligent customer segmentation: Unveiling consumer behaviour using RFM analysis and clustering," Springer, 2025. https://link.springer.com/article/10.1007/s43995-025-00180-7

[6] Karulkar et al., "Unravelling consumer patterns with K-Means and fuzzy logic," Springer, 2025. https://link.springer.com/article/10.1007/s44199-025-00143-w

[7] Agus et al., "Customer segmentation using an enhanced RFM-K-Means clustering approach," International Journal of Intelligent Information Systems, 2024. https://ijiis.org/index.php/IJIIS/article/view/289

[8] P. Bellini et al., "Multi-clustering recommendation system for fashion retail," Multimedia Tools and Applications, Springer, 2021. https://link.springer.com/article/10.1007/s11042-021-11837-5

[9] G. Vianna Filho et al., "A graph-based approach to customer segmentation using the RFM model," arXiv, 2025. https://arxiv.org/abs/2505.08136

[10] Flask Documentation, Pallets Projects. https://flask.palletsprojects.com/

[11] Gangadharan, K., Purandaran, A., Malathi, K., Subramanian, B., Jeyaraj, R., & Jung, S. K. (2025). From data to decisions: The power of machine learning in business recommendations. IEEE Access, 13, 17354-17397.

[12] Jahan, I., & Sanam, T. F. (2026). A comprehensive framework for customer retention in E-commerce using machine learning based on churn prediction, customer segmentation, and recommendation: I. Jahan, TF Sanam. Electronic Commerce Research, 26(1), 1-44.

[13] Rezaeinia, S. M., & Rahmani, R. (2016). Recommender system based on customer segmentation (RSCS). Kybernetes, 45(6), 946-961.

[14] Chakraborty, A., Dey, K., Ghosh, S., Chakraborty, R., Mitra, I., & Nandy, P. (2023, March). A novel approach for customer segmentation and product recommendation to boost sales using machine learning. In 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 97-103). IEEE.

[15] Gupta, S., & Israni, D. (2024, October). Machine Learning based Customer Behavior Analysis and Segmentation for Personalized Recommendations. In 2024 2nd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS) (pp. 654-660). IEEE.

[16] Handoko, K. F., Nugraha, N., & Sujjada, A. (2026). Customer Segmentation-Based Promotion Recommendation Using RFM and K-Means Clustering on a Web Platform. bit-Tech, 8(3), 3661-3672.

[17] Rymarczyk, P., Smutek, T., Stefańczak, D., Cwynar, W., & Zupok, S. (2024). Self-learning recommendation system using reinforcement learning.

[18] Nandhini, C., & Ayyanathan, N. (2026). Graph Neural Networks for Enhanced Customer Segmentation in Next - Generation Recommendation Systems. Next - Generation Recommendation Systems: A Comprehensive Guide to Enabling Technologies and Tools and their Business Benefits, 465-485.

[19] Guo, Y. (2025). Machine Learning Methods in Customer Segmentation and Recommendation Systems. In SHS Web of Conferences (Vol. 218, p. 02012). EDP Sciences.

[20] Usip, P. U., Gibson¹, U. E., & Udoh, S. S. (2025, February). A Multiple Criteria-Based Customer Segmentation and Recommender System. In Artificial Intelligence: Towards Sustainable Intelligence: Second International Conference, AI4S 2024, Alcala de Henares, Spain, October 3–4, 2024, Proceedings (p. 27). Springer Nature.