**CUSTOMER SEGMENTATION RECOMMENDATION SYSTEM**

k.Sujith reddy

*computer science engineering -Data science*

*MLR Institute of Technology* Hyderabad, India

S charan

*computer science engineering -Data Science*

*MLR Institute of Technology* Hyderabad, India

Sneha Thakur

*computer science engineering -Data Science*

*MLR Institute of Technology* Hyderabad, India

**Abstract** — In this project, we explore the burgeoning field of online retail through a comprehensive analysis of transactional data from a UK-based retailer, sourced from the UCI Machine Learning Repository. This dataset, which captures transactions over a one-year period from 2010 to 2011, serves as a foundation for enhancing marketing effectiveness and driving sales growth through customer segmentation. Our primary objective is to reformat the transactional data into a customer-centric framework by engineering new features that allow us to identify distinct customer profiles.

Leveraging the K-means clustering algorithm, we categorize customers into various segments based on their purchasing behaviors and preferences. This segmentation enables us to understand the unique characteristics of each customer group, providing valuable insights into their preferences and buying patterns. Building on these insights, we develop a recommendation system tailored to each segment. This system suggests top-selling products to customers who have yet to purchase these items, ultimately optimizing marketing strategies and boosting sales. Through this data-driven approach, our project demonstrates the potential of customer segmentation and targeted recommendations to strengthen customer relationships and maximize retail performance.

1. **INTRODUCTION**

The rapid expansion of online retail has intensified the need for effective data-driven strategies to understand customer behavior and optimize marketing efforts. In this context, customer segmentation and recommendation systems have emerged as powerful tools for retailers to tailor their marketing strategies to specific consumer groups, ultimately driving engagement and boosting sales. By segmenting customers based on purchasing patterns and making personalized product recommendations, retailers can adopt a targeted approach that resonates with diverse customer types. This paper explores both customer segmentation and recommendation systems within the online retail space, utilizing a comprehensive dataset of transactional data from a UK-based retailer, sourced from the UCI Machine Learning Repository.

The dataset, publicly available from the UCI Machine Learning Repository, provides a rich view of customer transactions from 2010 to 2011, including purchase quantity, invoice date, unit price, and customer ID. Covering a full year of transactions, this dataset captures the seasonality, frequency, and diversity of customer purchases, making it a valuable resource for understanding customer behavior over time. Additionally, the dataset includes customer location data, enabling exploration of regional purchasing trends and preferences, which adds another layer of insight into customer behavior.

Our primary objective is to transform this transactional dataset into a customer-centric one by creating derived features, such as purchase frequency, average purchase value, and recency of purchase, which are essential for understanding customer behavior. Using these features, we segment customers into distinct groups through the K-means clustering algorithm, enabling us to develop detailed profiles that illuminate customer preferences and spending habits. Building on these insights, we proceed to design a recommendation system that suggests top-selling products to customers within each segment who have not yet purchased those items.

This integrated approach to customer segmentation and recommendation allows the retailer to deploy highly personalized marketing strategies, with the dual goals of enhancing customer engagement and increasing sales. Through this study, we aim to demonstrate the value of combining segmentation and recommendation systems to optimize retail marketing efforts, offering a replicable framework for leveraging transaction data to achieve improved business outcomes.

1. **LITERATURE SURVEY**

***[1] Author Gomes and Meisen (2022)*** provided a structured overview of various customer segmentation methods tailored for e-commerce applications in their paper, ***“A review on customer segmentation methods for personalized customer targeting in e-commerce use cases.”*** Through an extensive literature review of 105 publications spanning from 2000 to 2022, the authors emphasize the growing significance of customer-centric marketing and the evolution of segmentation techniques over the years. They identify a four-phase process for customer segmentation, including data collection, customer representation, segmentation through clustering, and customer targeting. K-means clustering emerges as the most frequently used segmentation method due to its simplicity and effectiveness, particularly in e-commerce contexts. Additionally, they analyze temporal trends and the applicability of these techniques to datasets with varying dimensionalities. However, the paper’s focus on e-commerce may limit the transferability of its findings to other industries, and the literature review spans a wide range of publications, which might make it challenging to draw specific, actionable insights.

***[2] Author Ranjan and Srivastava (2022***), in their work *“Customer segmentation using machine learning: A literature review,”* explore various machine learning techniques applied to customer segmentation, with an emphasis on understanding customer behavior and trust. They discuss the relevance of customer analytics for managing customer churn and enhancing customer experiences. Through a comparative analysis of different clustering algorithms, the authors highlight the potential of hybrid models—combinations of multiple clustering algorithms—to improve the effectiveness of segmentation compared to single-method approaches. While k-means clustering is acknowledged as one of the widely used methods, the review suggests that hybrid approaches could uncover deeper insights within customer data. The paper’s focus on machine learning may limit its inclusion of traditional methods, and there is a need for more empirical evidence on the practical implementation and success rates of the proposed segmentation methods.

***[3]Author Joni Salminen, Mekhail Mustak, Muhammad Sufyan, and Bernard J. Jansen,*** in their systematic review titled ***“How can algorithms help in segmenting users and customers?”*** explore key questions in customer segmentation research and practice. Analyzing 172 articles, the authors identify 46 segmentation algorithms and 14 evaluation metrics, with K-means clustering emerging as the most frequently utilized method. They advocate for using multiple algorithms and evaluation metrics to improve segmentation accuracy, proposing seven primary goals and three practical implications to guide further research and application. Despite these contributions, the study’s emphasis on algorithmic methods may overlook traditional segmentation approaches that could offer additional insights, the limited practical applicability due to its academic focus, and the reliance on literature review methods could lead to biases based on the selected studies and the researchers' interpretations.

1. **EXPERIMENTAL SETUP**
   1. **Motivation**

The motivation for this project lies in the growing demand for personalized marketing strategies within online retail, driven by the need to understand and cater to diverse customer behaviors in an increasingly competitive e-commerce environment. Traditional broad-based marketing approaches have become less effective as customer preferences shift rapidly, making segmentation and targeted recommendations essential for engaging customers and fostering loyalty. By leveraging customer segmentation, this project aims to optimize marketing efforts, enabling retailers to create tailored campaigns and recommendations that resonate with specific customer groups, ultimately enhancing customer satisfaction, reducing churn, and driving sales. Through a data-driven approach, the project aspires to unlock insights that will improve both marketing efficacy and the overall customer experience, offering a framework for sustainable business growth in the retail sector.

**3.2 Design**

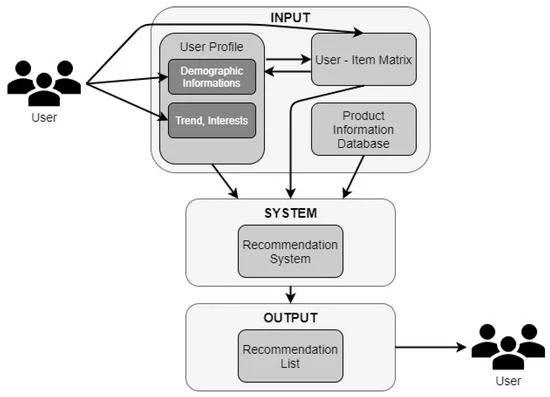


Fig 1 : A high-level overview of a recommendation system framework. The framework is divided into three main sections: Input, System, and Output.

* In the Input section, user-related data such as user profiles, demographic information, trends, and interests are collected. This information is integrated with the User-Item Matrix and the Product Information Database to create a comprehensive dataset.
* The System section represents the core Recommendation System, which processes the input data to generate recommendations.
* In the Output section, the system produces a Recommendation List, which is tailored to meet user preferences and is then presented to the users.

This framework illustrates how user data and product information are used to generate personalized recommendations.

* 1. **Requirements**

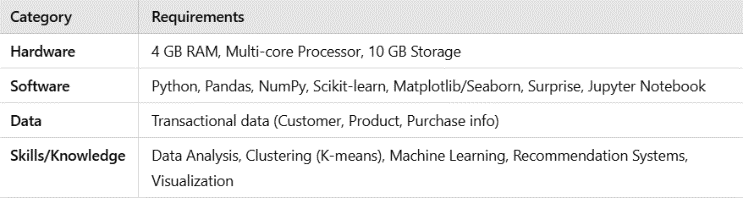
****

Fig 2 : Requirements of the Project

* 1. **Dataset**

The dataset used in your project is from the ***UCI Machine Learning Repository***, specifically the **Online Retail** dataset, which is a transactional dataset provided by a UK-based retailer. This dataset contains detailed transaction records for online sales between 2010 and 2011 and is widely used for data mining, customer behavior analysis, and recommendation system research.

***Information About the Dataset***

**Source:**

**Dataset Name**: Online Retail Dataset

**Source**: The dataset is available at the

[**UCI Machine Learning Repository**](https://archive.ics.uci.edu/ml/datasets/Online+Retail)**.**

* 1. **K-means Algorithm**

The K-means algorithm is one of the most widely used clustering methods in customer segmentation due to its simplicity, efficiency, and ability to create well-defined clusters. It is an unsupervised machine learning algorithm designed to partition a dataset into kkk clusters, where each data point belongs to the cluster with the nearest mean. Here is a detailed explanation of the K-means algorithm, including its mathematical foundation and why it is an effective choice for customer segmentation.

**How K-Means Works:**

The K-means algorithm operates in the following steps:

1. Initialize centroids: Select kkk initial centroids randomly from the dataset. These centroids represent the centers of the clusters.
2. Assignment step: Each data point in the dataset is assigned to the nearest centroid based on the Euclidean distance.
3. Update step: Calculate the new centroid of each cluster by taking the mean of all data points assigned to that cluster.
4. Repeat: The assignment and update steps are repeated iteratively until the centroids stabilize (i.e., there is no significant change in their positions) or until a predefined number of iterations is reached.

**Why K-Means is Effective for Customer Segmentation**

1. Interpretability: K-means provides clusters with distinct centroids, making it easy to interpret each group of customers. This clear delineation helps marketers understand specific customer segments.
2. Efficiency and Scalability: K-means is computationally efficient, especially with a large number of data points, making it a good fit for customer segmentation tasks that involve large transactional datasets.
3. Non-Hierarchical Structure: K-means is a non-hierarchical method, allowing for faster clustering without the need for a predefined hierarchy. This flexibility is helpful when working with a variety of customer data features that don't naturally follow a hierarchical structure.
4. Ability to Handle High-Dimensional Data: While K-means works best with lower-dimensional data, it can still be effective for high-dimensional datasets by using techniques like Principal Component Analysis (PCA) to reduce dimensionality before clustering.
5. Quick Convergence: The iterative process of K-means usually converges quickly, making it suitable for exploratory data analysis in customer segmentation tasks.

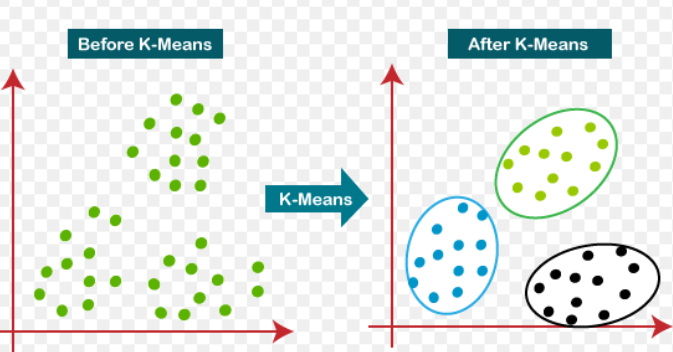


Fig 3 : K-means Algorithm Diagram

1. **IMPLEMENTATION**

The implementation of the project follows a structured approach to preprocess, analyze, and apply machine learning techniques to the dataset. First, we perform data cleaning and transformation, addressing issues like missing values, duplicates, cancelled transactions, and correcting anomalies such as zero unit prices and stock codes. This ensures the dataset is accurate and ready for analysis. Next, we proceed with feature engineering, where we calculate key customer attributes like Recency, Frequency, and Monetary (RFM), along with additional features such as product diversity, seasonality, and geographic insights to enrich the dataset and capture important customer behaviors.

Following this, we perform outlier detection and treatment to ensure that extreme values do not skew the analysis. Then, correlation analysis is carried out to identify relationships between features, helping us to understand the data better. Afterward, we apply feature scaling to standardize numerical features, ensuring that all features contribute equally to the clustering process. Dimensionality reduction using PCA is then performed, where we visualize the cumulative variance to determine the optimal number of principal components to retain for subsequent analysis.

The next step involves K-Means clustering, where we use methods like the Elbow and Silhouette techniques to determine the optimal number of clusters. We then apply the K-Means algorithm to segment customers into distinct groups. The clustering results are evaluated through 3D visualizations and cluster distribution visualizations, along with performance metrics to assess the quality of the clustering. Cluster analysis and profiling follow, using tools like radar and histogram charts to derive detailed insights into the characteristics of each customer segment.

Finally, the project culminates in the development of a recommendation system, which leverages customer segmentation to suggest personalized products to customers based on their purchase behaviors and preferences, enhancing marketing strategies and driving sales.

1. **CONCLUSION**

This project successfully applied K-means clustering and Principal Component Analysis (PCA) to segment customers based on their purchasing behavior, enabling personalized marketing strategies. By cleaning the data, engineering relevant features, and reducing dimensionality, we were able to gain valuable insights into customer profiles and recommend products tailored to each segment. The clustering and recommendation system, evaluated through various metrics and visualizations, demonstrated the potential to enhance sales by providing targeted suggestions, ultimately improving the overall marketing efficiency of the retail business.

1. **REFERENCES**
2. Gomes, M. A., & Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. Information Systems and e-Business Management, 21, 527-570  
   https://doi.org/10.1007/s10257-023-00640-4
3. Ranjan, A., & Srivastava, S. (2022). Customer segmentation using machine learning: A literature review. AIP Conference Proceedings, 2481, 020036.

https://doi.org/10.1063/5.0103946

1. Salminen, J., Mustak, M., Sufyan, M., & Jansen, B. J. (2023). How can algorithms help in segmenting users and customers? A systematic review and research agenda for algorithmic customer segmentation. Journal of Marketing Analytics,11,677-692.

<https://doi.org/10.1057/s41270-023-00235-5>

1. Sukru Ozan, "A Case Study on Customer Segmentation by using Machine Learning Methods",IEEE, Year: 2018.
2. Jayant Tikmani, Sudhanshu Tiwari, Sujata Khedkar “Telecom customer segmentation based oncluster analysis An Approach to Customer Classification using k-means”, IJIRCCE,Year: 2015.
3. Chinedu Pascal Ezenkwu, Simeon Ozuomba, Constance kalu Electrical/Electronics and ComputerEngineering Department, University of Uyo, Akwa Ibom State, Nigeria “Application of K-MeansAlgorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services”,IJARAI,Year: 2015.
4. Potharaju, S. P., Sreedevi, M., Ande, V. K., & Tirandasu, R. K. (2019). Data mining approach for

accelerating the classification accuracy Of

cardiotocography. Clinical Epidemiology & Global

Health, 7(2), 160-164