

**PDF REPORT:**

**Siddharthan.R**

**Data science assignment:  
ecommerce transaction  
dataset**

# 1. Introduction:

In the field of customer relationship management (CRM), identifying similar customers can drive personalized marketing strategies, improve customer retention, and optimize product recommendations. This task involves building a Lookalike Model to recommend 3 similar customers for the first 20 customers in the dataset based on both their profile and transaction history. The similarity score indicates how closely the recommended customers match the input customer.

Model Development:

To solve this task, we utilize both customer profile information and transaction history. The core idea is to compute similarity between customers based on shared characteristics and purchasing behavior. A combination of customer demographics (age, location, etc.) and transaction data (product purchases, frequency, and spending) are used for similarity calculation.

## 2. Data Preprocessing:

- **Customer Profile Data:** The Customers.csv file contains essential demographic information, such as age, gender, and location. This data is first cleaned and normalized.
- **Transaction Data:** The Transactions.csv file provides historical transaction information for each customer. It includes transaction amount, products purchased, and purchase frequency. This data is also preprocessed and transformed to capture valuable features like average spend, product categories, and total number of transactions.

## 3. Feature Engineering:

- We create the following features from the raw data:
- Customer Demographics: Age, gender, and location.
- Transaction Patterns: Total expenditure, frequency of purchases, preferred product categories, and time-based features (e.g., seasonality of purchases).
- These features are then standardized for uniformity and to enhance the model's performance.

## 4. Recommendation Algorithm:

- For each of the first 20 customers (CustomerID: C0001 to C0020), we compute the cosine similarity with all other customers in the dataset.
- We then recommend the top 3 most similar customers based on the highest similarity scores.
- A threshold of similarity can be used to filter out customers that have a low similarity score.

## 5. Output: Lookalike.csv

The result of the model is a CSV file (Lookalike.csv) where each row maps a customer ID to a list of three similar customers, along with their respective similarity scores.

CustomerID, Lookalike1\_ID, Similarity\_Score1, Lookalike2\_ID, Similarity\_Score2, Lookalike3\_ID, Similarity\_Score3

C0001, C0025, 0.89, C0033, 0.85, C0017, 0.83

C0002, C0041, 0.91, C0020, 0.88, C0053, 0.82

Evaluation Criteria

The model is evaluated based on:

- 1.Accuracy and Logic: The model's ability to accurately find similar customers based on demographic and transaction information.
- 2.Quality of Recommendations: The relevance of the recommended customers (i.e., customers with similar interests and behaviors).
- 3.Similarity Scores: The strength of the similarity score, ensuring that the top recommended customers are genuinely similar.

## 4. Conclusion:

The Lookalike Model effectively identifies and recommends the most similar customers based on their profile and transaction data. By leveraging cosine similarity, the model provides meaningful recommendations that can be used for targeted marketing campaigns, product recommendations, and customer retention strategies.