Lookalike Model Development Report

eCommerce Transactions Dataset

---

1. Report Overview

- **Date**: 27/01/2024
- **Prepared by**: Anbhi Thakur
- **Objective**: Develop a lookalike model to recommend similar customers based on profile and transaction behavior, aimed at enhancing marketing strategies and customer engagement.

---

2. Data Preparation

To build the lookalike model, we utilized the following datasets:

- **Customers.csv**: Contains customer profiles, including **CustomerID**, **Region**, and **SignupDate**.
- **Products.csv**: Includes product details such as **ProductID**, **Category**, and **Price**.
- **Transactions.csv**: Records customer transactions with **ProductID**, **Quantity**, and **TotalValue**.

The datasets were merged into a consolidated view. Missing values were handled appropriately, and categorical variables (like **Region** and **Product Category**) were one-hot encoded. Transaction-based features, including **total expenditure**, **average transaction value**, and **product preferences**, were derived to enhance the customer profiles.

---

3. Model Development

3.1 Feature Engineering

- **Profile Features**:
  - One-hot encoded **Region** and **converted SignupDate to tenure**.
- **Transaction Features**:
  - Calculated **total expenditure**, **number of transactions**, **product category preferences**, and **average transaction value**.

3.2 Similarity Calculation

- **Metric**:
  - **Cosine similarity** was used to measure pairwise similarity between customers based on their profiles and transaction behaviors. This metric captures the relative similarity by measuring the cosine of the angle between two feature vectors.

3.3 Recommendation Logic

- For each customer, the similarity scores were calculated with all other customers.
- The top three customers with the highest similarity scores were selected as lookalikes.
- Results were structured in the following format:
    - **Map<CustomerID, List<CustomerID, SimilarityScore>**

---

4. Results

For the first 20 customers (C0001 - C0020), three lookalike recommendations were made with corresponding similarity scores. These recommendations were validated for diversity and relevance, ensuring that transaction behavior and regional preferences were adequately considered.

---

5. Challenges and Solutions

5.1 Imbalanced Data

- Some customers had fewer transactions, which led to skewed similarity scores.
- **Solution**:
    - Normalized **average transaction values** to mitigate this imbalance.

5.2 High Dimensionality

- Feature dimensions were initially large, which may impact model performance.
- **Solution**:
    - Although **Principal Component Analysis (PCA)** was considered, it was deemed unnecessary as feature dimensions were manageable.

---

6. Future Enhancements

- **Advanced Techniques**:
    - Incorporate **deep learning** models for better similarity measurements.
- **Temporal Trends**:
    - Include **temporal transaction trends** to enable dynamic, evolving lookalike recommendations.
