Lookalike Model Insight Report

Prepared by: Murra Pranai Kumar Reddy
Project: eCommerce Transactions Dataset - Lookalike Model

Objective

The objective of the lookalike model is to identify and recommend **3 similar customers** for a given customer based on their **profile (region, signup date, etc.)** and **transaction history (purchase behavior, quantity, total value, etc.)**. This information can help the business better target customers with personalized campaigns and improve marketing strategies.

Steps in Building the Lookalike Model

1. Data Preparation

- **Customers Dataset**: Information such as region, signup date, and customer name was used as profile data.
- **Transactions Dataset**: Features such as total value, quantity purchased, and product preferences were aggregated for each customer to capture purchasing behavior.
- **Products Dataset**: Product category and price were combined with transaction data to better understand customer preferences.

Key Features Selected for Modeling:

1. **Region**: Customer's geographical location.
2. **Average Transaction Value**: Indicates spending habits.
3. **Total Quantity Purchased**: Represents purchase frequency.
4. **Category Preference**: Most purchased product category.
5. **Signup Recency**: Days since signup.

---

2. Similarity Calculation

- **Approach**: The model computes the **similarity score** between customers using their feature vectors.
- **Similarity Metric**: Cosine similarity was chosen because it is effective for high-dimensional feature spaces and captures the relative importance of features.
- **Scaling**: Features were standardized to ensure all attributes contributed equally to the similarity score.

---

3. Recommendations

For each customer in C0001–C0020, the top 3 similar customers were identified based on the similarity score. The results were saved in the Murra_Pranai_Kumar_Reddy_Lookalike.csv file.

Insights and Observations

1. Spending Patterns Drive Similarity

- **Insight**: Customers with high average transaction values tend to have higher similarity scores with other high-value customers, even across different regions. This indicates that spending habits are a critical determinant of customer similarity.

Recommendation: Target high-spending customers with premium offers, regardless of their geographical region, to enhance their lifetime value.

---

2. Regional Segmentation Plays a Role

- **Insight**: Customers from the same region showed higher similarity scores compared to those from different regions, likely due to shared product preferences or pricing policies.

Recommendation: Develop region-specific marketing strategies to cater to the preferences of customers in different regions.

---

3. Product Preferences Align Similar Customers

- **Insight**: Customers with overlapping product categories (e.g., "Electronics" or "Furniture") tend to have higher similarity scores. Electronics buyers also showed overlap with customers purchasing related accessories.

Recommendation: Bundle products or cross-sell complementary items to similar customers within the same product category.

---

4. Recency of Transactions Impacts Recommendations

- **Insight**: Customers who recently signed up (low signup recency) tend to form a separate cluster and are often compared to other new users. Their behavior may not align with older customers.

Recommendation: Provide personalized onboarding experiences and targeted offers to newer customers to retain their engagement early on.

Key Challenges

1. **Data Sparsity**: Customers with limited transaction data were harder to compare accurately.
    - **Solution**: Imputed missing data by assigning default values based on regional averages.

2. **Scalability**: Calculating similarity scores for large datasets can become computationally expensive.
    - **Solution**: Optimized similarity calculations using efficient libraries like sklearn.

Results

The following is an example of the lookalike output for customer C0001:

Customer ID	Recommended Customers	Similarity Scores
C0001	C0050, C0032, C0045	0.92, 0.89, 0.87

This format was replicated for all customers in the range C0001–C0020 and saved in the CSV file.