Data Science Assignment: eCommerce Transactions Dataset

Task 2: Lookalike Model

Customer Lookalike Analysis Report

In this analysis, I aimed to identify customers who are similar to each other based on their transaction behaviours and demographic features. The process involved several key steps, from loading and merging datasets to calculating customer similarities and generating lookalike recommendations. Below is an overview of how I approached the analysis and the results obtained.

Data Preparation

The first step was to load the necessary datasets: **Customers**, **Products**, and **Transactions**. Each dataset contained valuable information:

- **Customers**: Customer details like ID, region, and signup date.
- **Products**: Product details such as category and ID.
- **Transactions**: Transaction information, including customer purchases and quantities.

I merged these datasets to create a unified dataset where each transaction was associated with the relevant customer and product details. This gave me a comprehensive view of each customer's purchasing history, along with their demographic information.

Feature Engineering

After merging the datasets, I focused on feature engineering. This is where I created new, meaningful variables that would help in identifying similarities between customers. For each customer, I calculated several metrics:

- **Total Value**: The total amount spent by the customer.
- **Quantity**: The total number of products purchased.
- **Transaction Count**: The number of transactions made by the customer.

- **Signup Date**: The date when the customer first signed up.
- **Region**: The geographic region the customer belongs to.
- **Category**: The most frequently purchased product category by each customer.

I also added a **DaysSinceSignup** feature, which represents how long it has been since the customer signed up, providing a measure of recency.

Data Preprocessing

To prepare the data for analysis, I performed a few preprocessing steps:

- **One-hot Encoding**: I one-hot encoded the categorical features, such as Region and Category. This transformation allows me to work with these variables in a format that can be used for similarity calculations.
- **Normalization**: Using StandardScaler, I normalized the data to ensure all features were on the same scale. This is crucial because some features, like TotalValue, may have much larger ranges than others, and I didn't want any single feature to dominate the similarity calculation.

Similarity Calculation

Once the data was preprocessed, I calculated the **cosine similarity** between customers. Cosine similarity is a metric that measures how similar two customers are based on their purchasing patterns and demographic features. A score closer to 1 means the customers are very similar, while a score closer to 0 means they are quite different.

Generating Lookalike Recommendations

With the similarity scores calculated, I focused on generating lookalike recommendations. For the first 20 customers (C0001 to C0020), I identified the top 3 most similar customers. To do this, I excluded the customer from their own similarity calculation and selected the customers with the highest similarity scores.

Results

I compiled the top 3 lookalikes for each customer into a dictionary and then converted it into a DataFrame. This DataFrame contains the CustomerID and a list of their top 3 similar customers, along with the similarity scores. The results were then saved as a CSV file called Lookalike.csv.

	CustomerID	Lookalikes
0	C0001	[(C0184, 0.9530107530794122), (C0192, 0.9350887623338706), (C0112, 0.9080078858444088)]
1	C0002	[(C0106, 0.9321228466590178), (C0134, 0.919797480408065), (C0040, 0.9111090028038117)]
2	C0003	[(C0052, 0.9735448208255614), (C0076, 0.9697062419450831), (C0031, 0.9668832348221438)]
3	C0004	[(C0165, 0.9752419725283313), (C0155, 0.9414271268559732), (C0169, 0.9369836459901905)]

Conclusion

This analysis provides a powerful way to identify customers who are similar to each other. The lookalike recommendations can be used for targeted marketing, personalized product recommendations, and other customer engagement strategies. By understanding the similarities between customers, businesses can improve customer retention and satisfaction by offering more relevant products and services.

Overall, I believe this analysis can be a valuable tool for businesses looking to enhance their customer relationship management (CRM) efforts and drive growth through data-driven insights.