Lookalike Model Development Report

TASK 2

Overview of the Model Development:

In this project, I built a Lookalike Model that recommends similar customers based on both their customer profiles and transaction histories. The model calculates similarity scores between customers by utilizing their demographic and transaction data, and then identifies the top 3 most similar customers (lookalikes) for each of the first 20 customers (C0001 - C0020) from the provided dataset.

Step 1:

Data Pre-processing:

- I began by loading and merging the Customers.csv, Products.csv, and Transactions.csv datasets.
- The data was cleaned and merged based on common identifiers, ensuring the integration of customer information with their product purchase history.

Step 2:

Feature Engineering Overview:

- Feature engineering is the process of transforming raw data into meaningful features for machine learning models.

Steps Performed:

- YearsWithBusiness: Calculated the number of years each customer has been with the business by converting the SignupDate to a numeric value (years).
- Total Spend: Aggregated total spending for each customer, reflecting how much they've spent over their lifetime with the business.
- Average Quantity Purchased: Calculated the average quantity of products purchased per transaction, indicating a customer's typical purchase behaviour.
- Most Frequent Category: Identified the product category that each customer buys the most, providing insight into their product preferences.
- Product Diversity: Measured the variety of products purchased by each customer, showing the range of items they're interested in.

Purpose:

- These features (YearsWithBusiness, TotalSpend, AvgQuantityPurchased, ProductDiversity, and MostFrequentCategory) enhance the customer profile matrix, making it more informative for machine learning models.

Step 3:

- Output of the lookalike recommendation system for the first 20 customers (CustomerID: C0001 - C0020). The table lists each customer along with their top 3 lookalikes and the similarity scores. These recommendations were generated using cosine similarity based on various customer attributes like Region, YearsWithBusiness, TotalSpend, AvgQuantityPurchased, ProductDiversity, and MostFrequentCategory.

Key Observations:

- Identical Lookalikes: In some cases, the top 3 lookalikes are identical, which indicates that the most similar customer shares an extremely close profile with others. For example:

- C0146 has C0056 as all three of its top lookalikes with a similarity score of 0.9732.
- C0127 has C0172 as all three of its top lookalikes with a similarity score of 0.9867.
- Similarity Score Range: The similarity scores for the top lookalikes typically range from 0.85 to 0.99, which is a good indication of how closely the customers' profiles align with each other. A higher score (close to 1) means that the customers are very similar.

Detailed Customer Lookalike Example:

C0199:
- Lookalike:
- C0060 with score 0.9676
- C0025 with score 0.9431
- C0025 (again) with score 0.9431
- Observation: C0025 appears twice, which suggests that this customer is highly similar to the second lookalike in both instances.

C0087:

- Lookalikes:
- C0046 with score 0.9282 (appears three times)
- Observation: C0046 has been recommended as the most similar lookalike across all three positions, showing a consistent similarity.

C0070:
- Lookalikes:
- C0074 with score 0.9273 (appears three times)
- Observation: Similar to the previous case, the lookalike list consists of identical customers, reinforcing the strong similarity between them.

CONCLUSION

The lookalike analysis using cosine similarity identified the top 3 similar customers for each of the first 20 customers based on their purchasing behaviour and attributes. High similarity scores, such as 0.9895, indicate nearly identical profiles, suggesting key customer segments. Repeated lookalikes imply that certain customers represent a central group with shared characteristics. This segmentation allows for more personalized marketing strategies and targeted promotions. Overall, the findings help businesses optimize customer engagement and retention efforts.