Name: Likhith K Raj
LinkedIn: LinkedIn Profile Link
GitHub: GitHub Profile Link
Portfolio: Portfolio Link

Zeotap Data Science Intern Assignment

Task 2: Lookalike Model

Google Colab Link:
https://colab.research.google.com/drive/1MHs2rP_Hkcs3IMrURgBBFk2PzUo_BKw?usp=drive_link

Steps for solving the problem:

Step 0: Data Processing
Point: Merged the Customers.csv, Products.csv, and Transactions.csv files using common keys (CustomerID and ProductID). Cleaned the data by handling missing values, correcting data types, and removing redundant columns.

Step 1: Feature Selection
Point: Selected relevant features such as Quantity, TotalValue, and Price to capture customer purchasing behavior. Categorical data like Region was retained for one-hot encoding.

Step 2: Data Scaling
Point: Standardized the numerical features using StandardScaler to ensure all variables have equal importance when calculating similarity.

Step 3: Cosine Similarity Calculation
Point: Calculated the cosine similarity between customers based on their scaled numerical features and encoded categorical data.

Step 4: Identify Top Lookalikes
Point: For each customer, identified the **Top 3 Lookalikes** (excluding themselves) based on the highest similarity scores from the cosine similarity matrix.

Step 5: Save Lookalike Results
Point: Saved the lookalike results into a file Lookalike.csv with columns CustomerID, Lookalike_1, Lookalike_2, and Lookalike_3, including similarity scores for each lookalike.

Predications (Customized):

Here is the 2 case of predication:

- **Case 1: For customer id C0003.**

```
Enter CustomerID (e.g., C0001): C0003
Top 3 Lookalikes for Customer C0003:
Lookalike 1: C0190, Similarity Score: 0.9546214849405968
Lookalike 2: C0091, Similarity Score: 0.9086281452471642
Lookalike 3: C0174, Similarity Score: 0.9044670576598652
Mean Squared Error (MSE) between actual and predicted
lookalikes: 0.0006458004230247932
```

- **Case 2: For customer id C0199.**

```
Enter CustomerID (e.g., C0001): C0199
Top 3 Lookalikes for Customer C0199:
Lookalike 1: C0073, Similarity Score: 0.9850059276267502
Lookalike 2: C0132, Similarity Score: 0.974943257219227
Lookalike 3: C0019, Similarity Score: 0.9196573480364263
Mean Squared Error (MSE) between actual and predicted
lookalikes: 0.0009143822517061226
```

Top 3 lookalikes with their similarity scores for the first 20 customers:

|CustomerID|Lookalike_1|Lookalike_2|Lookalike_3|

|-----|---------------|------------|-------------|

|C0001|C0076:0\.9446|C0011:0\.9432|C0137:0\.9298|

|C0002|C0025:0\.8750|C0157:0\.8581|C0121:0\.8524|

|C0003|C0190:0\.9546|C0091:0\.9086|C0174:0\.9045|

|C0004|C0175:0\.9434|C0109:0\.9369|C0101:0\.9319|

|C0005|C0186:0\.9127|C0103:0\.8282|C0131:0\.8167|

|C0006|C0171:0\.9533|C0107:0\.9446|C0168:0\.9300|

|C0007|C0078:0\.9935|C0146:0\.9875|C0092:0\.9763|

|C0008|C0109:0\.9596|C0034:0\.9291|C0084:0\.9177|

|C0009|C0061:0\.9843|C0167:0\.9547|C0128:0\.9283|

|C0010|C0111:0\.9570|C0121:0\.9393|C0027:0\.9352|

|C0011|C0001:0\.9432|C0076:0\.8773|C0184:0\.8756|

|C0012|C0102:0\.9505|C0099:0\.9108|C0113:0\.9108|

|C0013|C0148:0\.9869|C0163:0\.9490|C0096:0\.8855|

|C0014|C0097:0\.9675|C0119:0\.9660|C0063:0\.9580|

|C0015|C0058:0\.9776|C0042:0\.9404|C0092:0\.9381|

|C0016|C0050:0\.9721|C0079:0\.9191|C0125:0\.8953|

|C0017|C0041:0\.9770|C0124:0\.9729|C0075:0\.9613|

|C0018|C0068:0\.9630|C0065:0\.8847|C0008:0\.8441|

|C0019|C0073:0\.9708|C0172:0\.9479|C0199:0\.9197|

|C0020|C0015:0\.9315|C0128:0\.9302|C0058:0\.9201|

Please consider my profile:

Strong Data Science Background:
Extensive experience in data analysis, machine learning, and NLP, demonstrated through internships, projects, and certifications.

Proven Impact on Business Outcomes:
Delivered a 15% revenue improvement at Leucine and developed predictive models with high accuracy for stock trading and sentiment analysis.

Research and Innovation Focus:
Published papers on machine learning applications and fine-tuned models like Gemma 2 for Kannada, blending technical skills with research.

Technical Proficiency and Continuous Learning:
Skilled in Python, SQL, Tableau, and TensorFlow, with 100+ LeetCode problems solved and active participation in Kaggle competitions.