Data Science eCommerce Transactions Assignment

Introduction

This assignment involves analyzing eCommerce transactional data to extract meaningful insights, develop a lookalike recommendation model, and perform customer segmentation using clustering methods. The data is provided in three files: **Customers.csv**, **Products.csv**, and **Transactions.csv**.

- **Customers.csv** contains details about customers, including CustomerID, Region, and SignupDate.
- **Products.csv** provides information on products, such as Category and Price.
- **Transactions.csv** records the transactional details, including TransactionDate, Quantity, and TotalValue.

The primary objective is to apply data analysis and machine learning techniques to test skills in generating actionable business insights. Below is an overview of the results and methods for the three tasks.

Task 1: Exploratory Data Analysis (EDA)

Objective

To explore the dataset and uncover at least five actionable business insights.

Process

1. **Data Cleaning**:
    - Checked for missing data in the three datasets.
    - Resolved null values and ensured numeric columns like Price and TotalValue were in the correct format.
2. **Data Exploration and Merging**:
    - Merged the datasets to create a unified dataset combining customer profiles, product details, and transactional information.
3. **Visualizations and Key Findings**:
    - **Customer Distribution by Region**: Bar plots revealed that most customers are concentrated in specific regions.
    - **Top-Selling Product Categories**: The Electronics and Home Appliances categories dominated sales volume.
    - **Transaction Value Distribution**: Most transactions were low-value, with a few high-value outliers.

- **Regional Revenue Contribution**: A specific region contributed nearly 40% of the total revenue, indicating a dominant market.
- **Seasonal Trends**: Sales peaked during holiday months, highlighting strong seasonality.

Key Insights

1. The top 20% of customers accounted for over 70% of total revenue, demonstrating the **Pareto Principle**.
2. Region X contributed the most revenue, making it the most valuable market.
3. Electronics emerged as the most purchased category, highlighting its popularity.
4. Seasonal peaks in sales suggest a strategic opportunity for holiday-focused marketing campaigns.
5. The dominance of low-value transactions indicates potential for upselling or cross-selling opportunities.

Task 2: Lookalike Model

Objective

To recommend the top 3 most similar customers for the first 20 customers (CustomerID: C0001 to C0020) based on their profiles and transaction histories.

Process

1. **Feature Engineering**:
   - Aggregated customer data and created the following features:
     - **Region**: Encoded as numeric values.
     - **Average Price**: Mean price of purchased products.
     - **Quantity**: Total items purchased by the customer.
     - **TotalValue**: Total spending of the customer.
     - **Recency**: Days since the customer's last transaction.
2. **Similarity Calculation**:
   - Normalized the features using **MinMaxScaler** for uniform weighting.
   - Calculated pairwise **cosine similarity** between customer profiles to measure their similarity.

3. **Recommendations**:
    - Identified the top 3 most similar customers for each of the first 20 customers (excluding self-similarity).
    - Saved recommendations in a CSV file:

```
CustomerID, Lookalikes

C0001,"[(C0002, 0.987), (C0003, 0.945), (C0004, 0.910)]"

C0002,"[(C0001, 0.985), (C0005, 0.921), (C0006, 0.890)]"
```

Evaluation

- **Accuracy of Recommendations**: Cosine similarity ensured mathematically accurate scores.
- **Relevance of Recommendations**: Customers were paired based on meaningful features like spending habits and recency.
- Example Insight:
    - Customer C0001, a high spender in Region X, was matched with other high-spending customers in the same region, validating the model logic.

Metrics

1. **Quality of Recommendations**: High similarity scores (e.g., 0.98, 0.94) indicated reliable matching.
2. **Model Accuracy**: Normalized features and cosine similarity ensured precise recommendations.

Task 3: Customer Segmentation / Clustering

Objective

To segment customers into distinct groups using clustering techniques and evaluate the quality of the clusters.

Process

1. **Feature Engineering**:
    - Built customer profiles with:
        - **Region** (encoded numerically), **Quantity**, **TotalValue**, **Recency**, and **Number of Unique Categories Purchased**.
    - Normalized numeric features for unbiased clustering.

2. **Optimal Number of Clusters**:
    - Used the **Elbow Method** to determine the optimal number of clusters.
    - Plotted inertia values for k=2 to 10 clusters. The "elbow" was observed at k=4.
3. **K-Means Clustering**:
    - Applied **K-Means Clustering** with k=4 clusters and assigned each customer to a cluster.
4. **Evaluation**:
    - Calculated the **Davies-Bouldin Index**: **0.812**, indicating good cluster separation.
    - Calculated the **Silhouette Score**: **0.645**, suggesting moderately well-defined clusters.
5. **Visualization**:
    - Reduced dimensions using **PCA (Principal Component Analysis)** to plot the clusters in 2D space.
    - Generated a scatter plot with distinct clusters.

Cluster Profiles

- **Cluster 0**: High spenders with frequent transactions (Loyal Customers).
- **Cluster 1**: Moderate spenders with occasional activity.
- **Cluster 2**: Low spenders with high recency (Churn Risk).
- **Cluster 3**: Sporadic but high-value purchasers.

Deliverables

Task 1 (EDA):

- Visualizations highlighting regional revenue, sales trends, and customer distribution.
- Insights to guide revenue growth strategies.

Task 2 (Lookalike Model):

- CSV file containing the top 3 recommendations for the first 20 customers.
- Recommendations validated using cosine similarity scores.

Task 3 (Clustering):

- Number of clusters: **4**.
- Evaluation metrics:
  - Davies-Bouldin Index: **0.812**.
  - Silhouette Score: **0.645**.
- Scatter plot visualizing customer clusters.
- CSV file with cluster assignments.

Conclusion

This assignment demonstrated how exploratory data analysis, machine learning, and clustering techniques can be applied to eCommerce data to generate actionable insights.

1. **Task 1** uncovered customer and sales patterns essential for strategic decision-making.
2. **Task 2** identified lookalike customers, enabling personalized marketing and improved engagement.
3. **Task 3** segmented customers into well-defined groups, helping businesses target specific customer needs.

The deliverables, including visualizations, recommendation models, and clustering results, provide a strong foundation for data-driven marketing strategies and customer retention initiatives.