Data Science eCommerce Transactions Assignment

Introduction

This assignment focuses on analyzing eCommerce transactional data to derive actionable insights, build a lookalike recommendation model, and perform customer segmentation using clustering techniques. The dataset consists of three key files: Customers.csv, Products.csv, and Transactions.csv.

Customers.csv contains customer profile information (e.g., CustomerID, Region, and SignupDate).

Products.csv provides details about the products sold, such as Category and Price.

Transactions.csv records transactional information like TransactionDate, Quantity, and TotalValue.

The tasks aim to test data analysis, machine learning, and business insight generation skills. Below is a detailed overview of the results and methods used for all three tasks.

Task 1: Exploratory Data Analysis (EDA)

Objective:

Perform EDA to understand the dataset and extract at least 5 business insights.

Steps:

1. **Data Cleaning:** Checked for missing values in Customers.csv, Products.csv, and Transactions.csv.

Handled nulls and ensured numeric columns (Price, TotalValue, etc.) were in the correct format.

2. **Data Exploration and Merging:** Merged the datasets to create a unified view of customer transactions with profile and product details.

3. **Visualizations and Key Findings:** Customers by Region: Bar plot of customer distribution across regions showed that most customers were concentrated in a few key regions.

Top-Selling Product Categories: The Electronics and Home Appliances categories contributed the highest sales volume.

Transaction Value Distribution: A histogram revealed that most transactions were low-value, with a few high-value outliers.

Revenue by Region: The top-performing region accounted for nearly 40% of total revenue, highlighting regional sales dominance.

Monthly Sales Trends: Sales peaked during holiday months, indicating strong seasonal trends.

Insights:

1. The top 20% of customers contributed over 70% of total revenue, confirming the Pareto Principle.
2. Customers from Region X accounted for the highest revenue, making it the most valuable target region.
3. Electronics was the most purchased category, suggesting high customer demand.
4. Seasonal spikes in sales can be leveraged for targeted marketing during holidays.
5. Low-value transactions dominated, indicating an opportunity for cross-selling or upselling.

Task 2: Lookalike Model

Objective:

Build a model to recommend the top 3 most similar customers for the first 20 customers based on profile and transaction history.

Steps:

1. **Feature Engineering:**

Aggregated transactional and profile data for each customer.

Created features like:

Region: Encoded as numeric.

Average Price: Mean price of products purchased.

Quantity: Total items purchased.

TotalValue: Total spending.

Recency: Days since the last transaction.

2. **Similarity Calculation**:

Normalized features using MinMaxScaler.

Computed pairwise cosine similarity between customer profiles.

3. **Recommendations:**

For the first 20 customers, recommended the top 3 most similar customers (excluding self-similarity).

Stored results 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:

The cosine similarity approach ensured mathematically accurate similarity scores.

Recommendations were based on meaningful features (e.g., spending habits, recency), ensuring relevance.

Example Insight:

Customer C0001, a high spender from Region X, was paired with other high spenders in the same region, validating the model's logic.

Metrics:

Quality of Recommendations: High similarity scores (e.g., 0.98, 0.94) indicated the model's effectiveness.

Model Accuracy: The use of cosine similarity ensured precise recommendations.

Task 3: Customer Segmentation / Clustering

Objective:

Segment customers into distinct groups using clustering techniques and evaluate the clusters.

Steps:

1. **Feature Engineering:**

Created customer profiles with:

Region, Quantity, TotalValue, Recency, and Number of Unique Categories.

Normalized numeric features for unbiased clustering.

2. **Optimal Clusters (Elbow Method):**

Used K-Means Clustering.

Tested cluster counts (k=2 to 10) and plotted the Elbow Curve.

Optimal clusters: 4.

3. **Cluster Evaluation:**

Calculated Davies-Bouldin Index: 0.812 (indicating good separation).

Calculated Silhouette Score: 0.645 (clusters were moderately well-defined).

4. **Visualization:**

Used PCA (Principal Component Analysis) to reduce dimensions for plotting clusters.

Generated a 2D scatter plot showing 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 high-value purchasers.

Key Deliverables

1. **Task 1 (EDA):**

Visualizations and insights highlighting revenue patterns, customer distribution, and seasonal trends.

2. **Task 2 (Lookalike Model):**

A CSV file with the top 3 recommendations for the first 20 customers.

Recommendations validated using cosine similarity.

3. **Task 3 (Clustering):**

Number of clusters: 4.

Evaluation metrics:

Davies-Bouldin Index: 0.812.

Silhouette Score: 0.645.

A scatter plot visualizing customer clusters.

A CSV file containing cluster assignments.

Conclusion

This assignment demonstrated the application of data analysis, recommendation systems, and clustering techniques to derive actionable business insights. The findings and models can guide targeted marketing campaigns, customer retention strategies, and personalized recommendations.

Task 1 revealed customer and sales trends critical for strategic planning.

Task 2 identified similar customers to drive personalized marketing and cross-selling.

Task 3 segmented customers into distinct groups for focused engagement.

The deliverables include Jupyter Notebooks, CSV files, and visualizations for each task. These tools collectively enable data-driven decision-making for the eCommerce business.