Question 3 Report

Rohan Gore

~rmg9725

Files attached:

- midterm_q3.ipynb

Question 3.1: Customer Value Segmentation

Detailed Approach

1. **Recency Calculation**:
   - Filter completed orders.
   - Group by customer_id.
   - Calculate the difference between the current date (April 13, 2025) and the most recent order date for each customer.
2. **Frequency Calculation**:
   - Filter completed orders.
   - Group by customer_id.
   - Count the number of orders for each customer.
3. **Monetary Value Calculation**:
   - Calculate total amount for each order by multiplying quantity and unit price.
   - Join with completed orders.
   - Calculate average order value for each customer.
   - Round the average value to two decimal places.
4. **RFM Combination**: Join the recency, frequency, and monetary value dataframes on customer_id.
5. **Segmentation Function**:
   - Implement the provided `segment_customer` function as a User-Defined Function (UDF).
   - This function assigns scores to recency, frequency, and monetary value.
   - Calculates a total score and assigns a segment based on the score.
6. **Segmentation Application**: Apply the UDF to the combined RFM dataframe to assign segments to each customer.
7. **Result Preparation**:
   - Join the segmentation results with customer information.
   - Select only the required columns: customer_id, name, tier, and segment.
   - Show the customer segmentation results.

Real-Life Use Cases

1. **Personalized Marketing Campaigns**: Businesses can tailor their marketing strategies based on customer segments. For example, they might offer exclusive deals to "High Value" customers to maintain their loyalty, while providing incentives to "Medium Value" customers to increase their engagement and potentially move them to the "High Value" segment.
2. **Resource Allocation**: Companies can optimize their customer service and support resources by prioritizing "High Value" customers, ensuring they receive premium support to maintain their satisfaction and loyalty. This segmentation also helps in identifying "Low Value" customers who might need different strategies to increase their value or reduce service costs.

Question 3.2: Top 3 Categories

Detailed Approach

1. **Data Loading**: Load the customers and orders JSON datasets.
2. **Category Extraction**:
    - Filter out cancelled orders.
    - Use the `explode` function to transform the categories array into individual rows, creating a separate row for each category in an order.
3. **Category Counting**:
    - Group by customer_id and category.
    - Count the occurrences of each category for each customer.
4. **Category Ranking**:
    - Use a window function to rank categories for each customer based on purchase count.
    - The ranking is done in descending order of purchase count.
5. **Top 3 Selection**: Filter to keep only the top 3 ranked categories for each customer.
6. **Result Preparation**:
    - Join the top categories with customer information.
    - Select relevant columns: customer_id, name, category, purchase_count, and rank.
7. **Result Display**:
    - Show the top 3 categories for each customer.
    - Optionally, display a more detailed, customer-by-customer breakdown of the top categories.

Real-Life Use Cases

1. **Personalized Product Recommendations**: E-commerce platforms can use this analysis to create tailored product recommendations for each customer. By knowing a customer's top 3 categories, the platform can prioritize showing products from these categories or related categories, potentially increasing the likelihood of additional purchases and improving the customer's shopping experience.
2. **Inventory Management and Stock Optimization**: Retailers can use this information to optimize their inventory levels and stock placement. Understanding which categories are most popular among their customer base allows them to ensure adequate stock of high-demand items and potentially reduce overstocking in less popular categories, leading to more efficient inventory management and reduced carrying costs.