Technical Report: TravelTides Customer Segmentation Project

Overview

TravelTide, an e-booking startup, has faced challenges in customer retention despite its steady growth. To address this, the company plans to implement a personalized rewards program. The goal is to validate the proposed perks and tailor them to customer preferences. This report details the data analysis process, including data cleaning, exploratory data analysis, feature engineering, clustering, and the final assignment of perks to customers.

1. Data Cleaning and Preparation

Importing and Loading Data

The dataset was first extracted from the database as CSV file using SQL (Postgres).

It was loaded and the necessary libraries were imported, including pandas, numpy, matplotlib, seaborn, and several scikit-learn modules.

Data Overview

The dataset contains various customer-related attributes such as user ID, sign-up date, demographic information, and booking details.

Handling Outliers

Outliers were identified in key variables like `base_fare_usd` and `page_clicks`. The Z-score method was used to cap outliers instead of removing them completely. Histograms and box plots confirmed the presence and handling of these outliers.

Handling Missing Values

Missing values were replaced with appropriate values, such as zero for numerical fields where it made sense.

2. Exploratory Data Analysis (EDA)

Univariate Analysis

- **Gender Distribution:** Majority of the customers are female (88.37%).
- **Marital Status**: Nearly 56% are unmarried.
- **Age Distribution**: A histogram showed a diverse age range among customers.

Other Univariate Analysis

Key Insights from Histograms

- **Page Clicks:** Most sessions have a small number of clicks, indicating quick engagements.
- **Base Fare**: Highly skewed with some extremely high fares.
- **Total Flights and Hotels Booke**d: Most users book few flights and hotels, indicating simpler travel plans.
- **Discounts**: Skewed towards lower values, suggesting a conservative discount strategy.

Bivariate Analysis

Correlation heatmaps highlighted relationships between various features:

- High correlation between `total_flights_booked` and `total_hotels_booked`.
- Moderate correlation between `page_clicks` and booking behaviors.
- Weak correlation between discounts and booking behaviors, indicating potential areas for improvement in promotional strategies.

3. Feature Engineering

New features were created to enhance the analysis:

- **Avg_checked_bags**: Average number of checked bags.
- **Avg_base_fare:** Average base fare paid.
- **Length_of_stay**: Derived from check-in and check-out times.
- **Flight_distance_km**: Calculated using the Haversine formula for distance between airports.

4. Scaling and Dimensionality Reduction

Scaling

Features were scaled using `**StandardScaler**` to prepare for dimensionality reduction.

PCA

Principal Component Analysis (PCA) reduced the dimensionality to three principal components, making the dataset more manageable and interpretable.

5. Clustering

Elbow Method

The optimal number of clusters was determined to be four based on the Elbow method.

KMeans Clustering

KMeans clustering was performed with four clusters. The clusters were visualized in a 3D scatter plot, and each customer was assigned to a cluster.

Cluster Analysis

Each cluster was analyzed to understand its characteristics and to suggest appropriate perks:

- **Cluster 0**: Budget-conscious customers - Suggested perk: Free checked bags.
- **Cluster 1**: Moderate spenders - Suggested perk: One night free hotel with flight.
- **Cluster 2**: High spenders and frequent travelers - Suggested perk: No cancellation fee.
- **Cluster 3**: Customers valuing comprehensive packages - Suggested perk: Free hotel meals.

6. Perk Assignment and Fuzzy Segmentation

Perk Mapping

Customers were mapped to perks based on their cluster. An additional perk, "Exclusive Discounts," was assigned using threshold conditions on metrics like `bargain_hunter_index` and `avg_session_length`.

Fuzzy Segmentation

For precise perk assignment, customers were ranked across different perks, and the perk with the highest affinity was assigned to each customer.

7. Results

The final distribution of customers per perk was visualized:

Free Checked Bags: 11,391 customers

No Cancellation Fee: 11,249 customers

Exclusive Discount: 10,838 customers

Free Hotel Meals: 7,399 customers

One Night Free Hotel: 4,629 customers

Conclusion

The data analysis validated the hypothesis about customer interest in the proposed perks and enabled the personalized assignment of perks. This approach is expected to enhance customer retention by catering to individual preferences more effectively.

The detailed steps, visualizations, and insights provide a robust framework for implementing the rewards program, ensuring that TravelTide can better meet the needs of its diverse customer base.

APPENDIX

The detailed visualizations, chats and codes can be found below:

C:\Users\David Alaofin\TravelTides_Customer_Segmentation_Project.ipynb