TravelTide Project Report

Objective: The goal of the TravelTide project is to segment users based on their travel behavior, including bookings, cancellations, preferences, and demographics. The project uses machine learning techniques like KMeans clustering, PCA for visualization, and a Random Forest Classifier for predictive modeling. The key objective is to recommend personalized offers based on customer segments to maximize company profits through targeted advertising.

---

1. Data Loading and Initial Exploration:

The dataset is loaded from a Google Spreadsheet containing user information. Here's a snapshot of the data exploration process:

- **Initial Data Insights:**
  - Columns include: `user_id`, `gender`, `married`, `has_children`, `flight_booked`, `page_clicks`, `cancellation`, `seats`, `checked_bags`, `base_fare_usd`, `nights`, `rooms`, `hotel_per_room_usd`, `home_country`.
  - Gender distribution: The data includes two genders, and the category 'O' (other) was removed for consistency.
  - **Missing values**: Handled for features such as `base_fare_usd`, `seats`, and `checked_bags`.
  - New features are added such as `age, flight_hotel_booked, flight_duration_days, hotel_stay_duration, total_hotel_cost, active_days, cancellation_rate`

Preprocessing Actions:

- **Feature Encoding:** Categorical features (`gender`, `married`, `has_children`, `home_country`) are converted to numerical codes for model compatibility.
- **Missing Values:** Imputed with mean or median values where appropriate.

---

2. Data Preprocessing:

- **Scaling:** The dataset is scaled using `StandardScaler` to standardize numerical features, ensuring that no feature dominates due to differences in units.

---

3. KMeans Clustering:

- **Optimal Cluster Selection:** The silhouette score was calculated for different values of clusters (from 2 to 20), showing how well-separated the clusters are. The optimal number of clusters was found to be 6.

- **Cluster Assignments:** The dataset was assigned to 6 clusters based on their features, and the distribution of users in each cluster was explored.

---

4. PCA (Principal Component Analysis) for Dimensionality Reduction:

- PCA was performed to reduce the dimensionality of the data to 4 components.

- A scatter plot was generated using the first two PCA components, showing how users are clustered.

---

5. Cluster Analysis:

- **Cluster Means:** A detailed analysis of each cluster was conducted, calculating the average values for key features like `has_children`, `flight_booked`, `page_clicks`, and `base_fare_usd`.

- **Recommended Offers per Cluster:** Personalized offers were mapped to clusters based on user behavior:

    - **Cluster 0**: Frequent travelers - *"10% off next trip"*
    - **Cluster 1**: Frequent engagement, short trips - *"Discount at special events"*
    - **Cluster 2**: Budget-conscious travelers with families - *"Free child ticket"*
    - **Cluster 3**: Frequent travelers with high cancellations - *"Free meal"*
    - **Cluster 4**: Family travelers on short trips - *"Meal voucher"*
    - **Cluster 5**: Married with children - *"Free child ticket"*

---

6. Predictive Modeling with Random Forest Classifier:

- **Model Training:** A Random Forest Classifier was trained on the dataset, with features used to predict user clusters.

- **Model Evaluation:** The model's accuracy was evaluated, and the classification report was generated.

- **Feature Importance:** The importance of each feature in predicting user clusters was visualized through a bar plot, showing which features have the most significant impact on clustering.

---

7. Conclusions and Insights:

- **User Segments:** Six distinct user segments were identified, each with different preferences and behavior patterns.
- **Offers:** Personalized offers tailored to each cluster can drive higher engagement and increase the likelihood of repeat customers.
- **Predictive Modeling:** Random Forest Classifier offers a reliable way to predict future customer behavior, which can be leveraged to further personalize offers and optimize marketing efforts.

---

8. Next Steps:

- **Further Model Optimization:** Experiment with other machine learning models like Gradient Boosting or XGBoost for potentially better accuracy.
- **Real-Time Recommendations:** Integrate this model into the company’s customer-facing platform to provide real-time personalized offers.
- **Data Collection:** Further enrich the dataset by collecting more data on user interactions and expanding features.
