Data Science Project Report: Customer Segmentation and Perk Optimization in TravelTide

1. Business Background & Strategic Objectives

Problem Definition

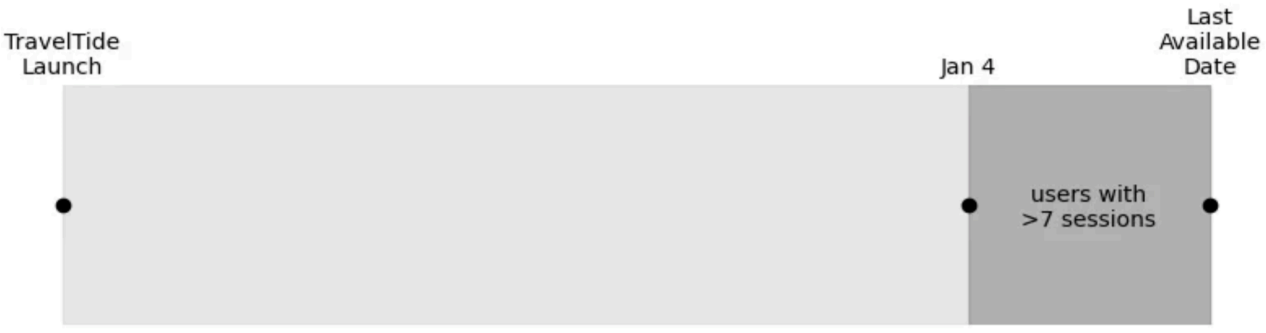
TravelTide, a fast-growing e-booking startup that launched in April 2021 at the tail end of the COVID-19 pandemic, has built its reputation on a best-in-class data aggregation and search engine, offering the largest travel inventory available online. However, despite this competitive advantage, customer retention has lagged. CEO Kevin Talanick has emphasized retention, and newly appointed Head of Marketing Elena Tarrant has been tasked with designing a personalized rewards strategy to drive repeat business.

It was built an end-to-end data pipeline that extracts, cleans, and transforms travel session and booking data, performs detailed exploratory analysis and customer segmentation, and assigns a likely favorite perk to each customer. Our objectives include:

- **Validating the hypothesis** that a subset of customers has a particular affinity for specific rewards (e.g., “no cancellation fees”).
- **Providing data-driven, personalized perk recommendations** for targeted marketing.

Cohort Selection

Customers who signed up recently have very few interactions with the platform. Including them in the same analysis as long-term customers would skew results. This issue extends to comparing customers with 6 months on the platform vs. 24 months. To control for the influence of time on platform usage, we select a defined cohort.



2. Data Extraction, Aggregation, and Exploratory Analysis

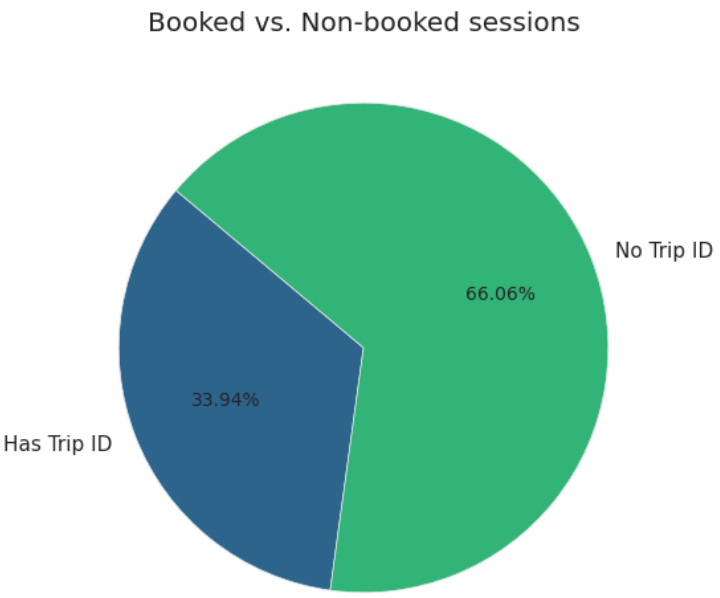
Raw Data Overview

The dataset consists of raw, unprocessed data collected from TravelTide's platform, including:

- **User Demographics:** Gender, age, marital status, presence of children, and geographic location.
- **Session Data:** Number of visits to the platform, page clicks, session duration, device type, and referral source.
- **Booking Data:** Flight and hotel booking history, payment method, frequency of cancellations, lead time before departure.
- **Travel Preferences:** Preferred destinations, flight class selection (economy, business, first class), and hotel star ratings.
- **Engagement and Spending Patterns:** Discounts used, fare prices, total spend per session, and booking conversion rates.
- **Historical Behavior:** Repeat customers vs. first-time users, preferred airlines, and hotel chains.

Exploratory Data Analysis (EDA)

Several key insights were derived from EDA:



- **Trip Motive Distribution**: The majority of travelers fall into leisure and family categories, while VIP and host travelers form a small fraction.
- **Booking Conversion Rate**: Approximately 33.94% of sessions result in a successful booking, while 66.06% do not.
- **Gender Distribution**: The majority of our customers are women. Interestingly, despite having children, they tend to travel alone rather than with family.
- **International vs. Domestic Travel**: The dataset reveals that a significant proportion of trips are international, highlighting a strong preference for overseas travel among users.
- **Session Activity & Engagement**: Users generally engage with the platform across multiple sessions before completing a booking. Repeat visitors show

a higher conversion rate, suggesting that users who interact more frequently with the platform are more likely to finalize their travel plans.

- **Data Cleaning & Outliers**:
    - **Negative Durations**: Instances where stay durations were recorded as negative were corrected or removed.
    - **Duplicate Bookings**: Canceled bookings that appeared multiple times were removed to ensure accuracy.
    - **Luxury Outliers**: Users with extremely high spending patterns were identified as outliers and later reassigned to premium perk categories.

These insights played a crucial role in refining our customer segmentation model and optimizing the assignment of perks.

3. Feature Engineering and Customer Segmentation

Feature Engineering

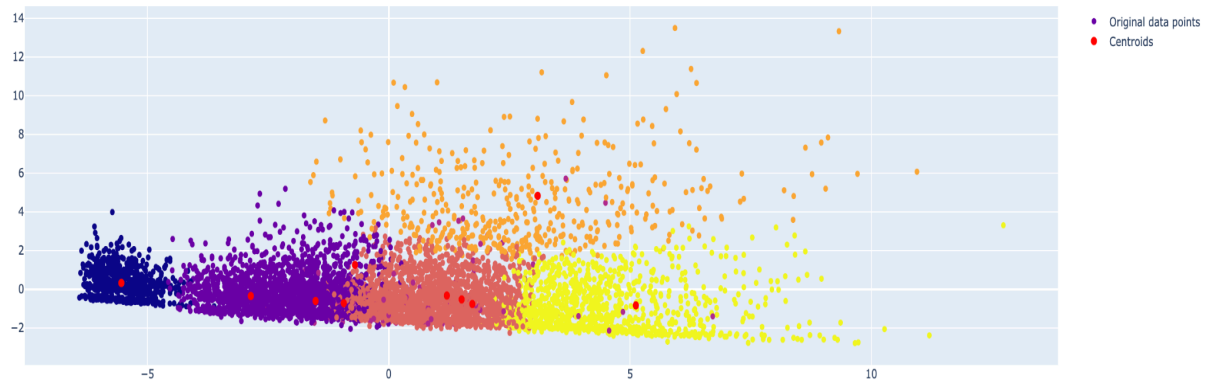
Based on raw data extraction, several new features were engineered to improve segmentation and perk prediction:

- **User Demographics Features:** Categorical encoding of age groups, marital status, and presence of children.
- **Travel Behavior Metrics:** Aggregation of total flights booked, number of hotel stays, and travel spending over time.
- **Spending Efficiency Metrics:** Calculation of average cost per kilometer traveled, fare per seat, and discount utilization.
- **Trip Motive Classification:** Labeling users based on inferred trip purpose (leisure, business, budget-conscious, etc.) using behavioral indicators.

Clustering Approach

To segment customers based on spending and travel patterns, we applied:

- **K-Means clustering**: To categorize users into six groups based on total trip cost, airfare, and discount usage.



- **t-SNE visualization**: To validate and interpret the separation between different clusters.

The resulting clusters provide insights into different customer behaviors:

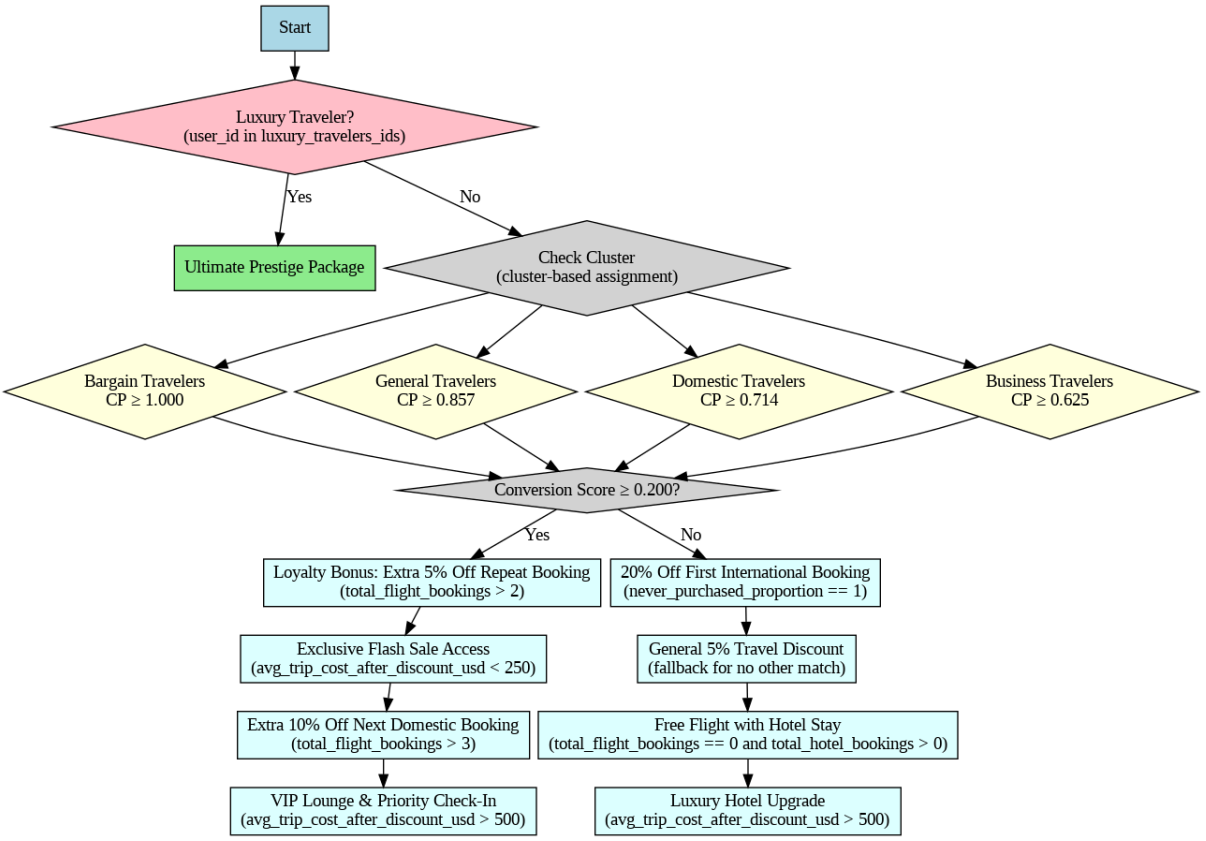
- **Budget Travelers**: Low-cost trips with minimal spending.
- **Occasional Travelers**: Moderate spending with infrequent travel.
- **Business Travelers**: High spending on premium flights and hotels.
- **Luxury Travelers**: Very high spending on exclusive services.

4. Perk Assignment Using Decision Trees

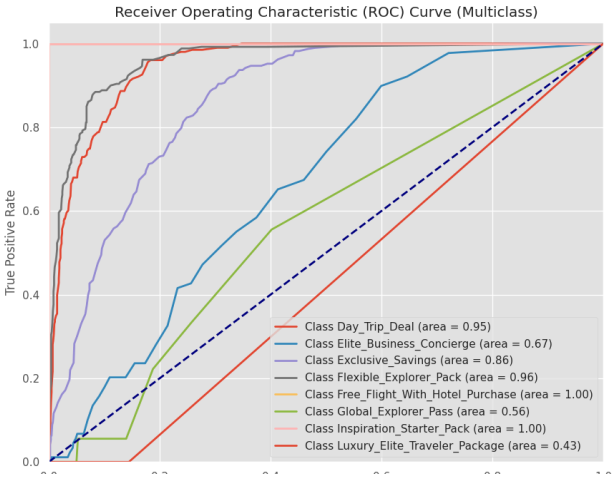
A decision tree model was developed to assign perks based on user spending, booking patterns, and cluster assignment. The flowchart below illustrates the perk assignment logic:

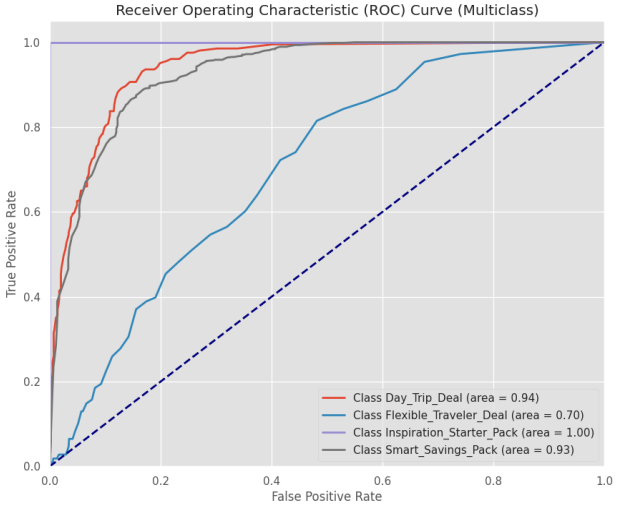
The perks include:

- **General 5% Travel Discount** for fallback cases.
- **Exclusive Flash Sale Access** for budget-conscious travelers.
- **Luxury Hotel Upgrades** for high-spending travelers.
- **VIP Lounge & Priority Check-In** for premium users.
- **Free Flight with Hotel Stay** for first-time high-value customers.



Model Performance Evaluation





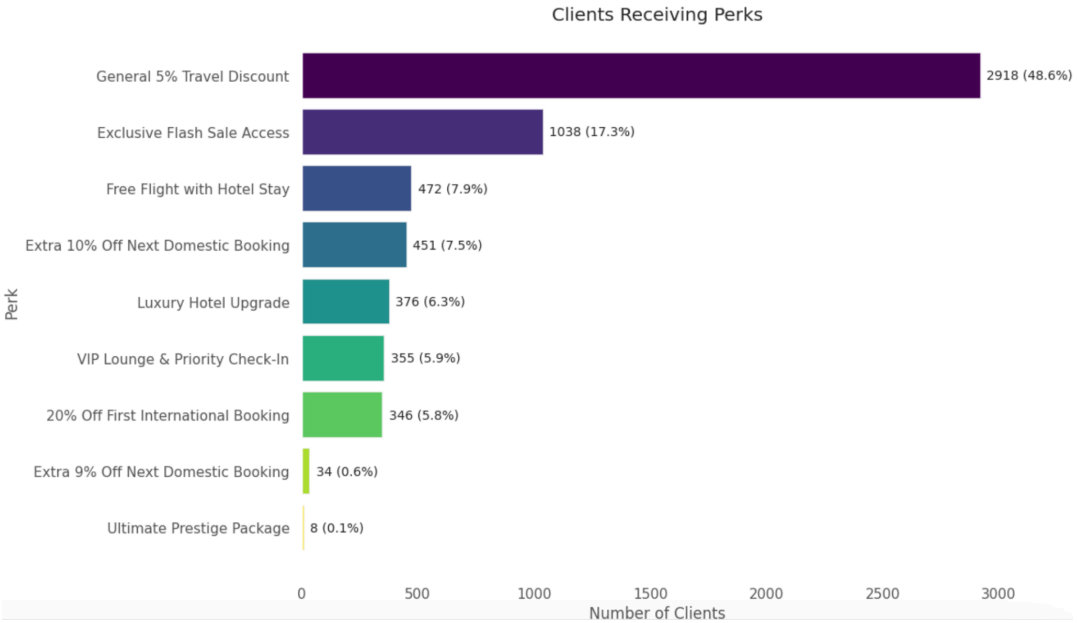
To assess the effectiveness of the perk assignment model, we analyzed the **Receiver Operating Characteristic (ROC) Curves** for different perk categories:

- **ROC Curve for 8 Perks**: This curve evaluates the model's performance across multiple perks, showcasing which classes are predicted with high confidence.
  ROC Curve for 4 Major Clusters: This visualization helps understand the distinction between key customer segments.

These results highlight that perks like **Day Trip Deal** and **Flexible Explorer Pack** are predicted with high confidence, whereas perks for luxury travelers have lower predictive performance due to smaller sample sizes.

5. Results and Insights

- **Perk Distribution**: The majority of users received a general 5% discount, while high-spending customers benefited from premium perks.



- **User Clusters Distribution**: Budget-savvy travelers and casual occasional travelers form the largest segments, indicating a need for budget-friendly perks.

6. Conclusions and Recommendations

Key takeaways:

- **Validated perk affinity**: Customers exhibit distinct preferences for specific rewards, confirming the hypothesis.
- **Targeted marketing strategies**: Personalized perks will enhance engagement and retention.
- **Scalability**: The clustering model can adapt over time with additional data.

By leveraging advanced data science techniques, TravelTide is poised to enhance its customer retention strategy through a highly personalized rewards program.

Prepared by

Svitlana Kovalivska, PhD
14.02.2025