TravelTide – Projektbericht von Lev Marchenko

1. Projektziel und Datenbasis:
Das Ziel des Projekts TravelTide ist die Analyse von Kundendaten, um ein zukünftiges Prämienprogramm optimal auf verschiedene Kundensegmente zuzuschneiden. Die Marketing-Leiterin Elena Tarrant hat sich für das Prämienprogramm fünf attraktive Boni überlegt: exklusive Discounts, freies Gepäck, kostenloses Essen, eine extra Nacht im Hotel und keine Stornogebühren. Basierend auf der Analyse sollten passende Kundengruppen identifiziert werden, um in den personalisierten E-Mails gezielt auf diese Vorteile einzugehen. Dafür wurden Daten aus vier vom Kunden bereitgestellten Tabellen (User-Session-Verhalten, Demografie, Reiseverhalten) untersucht. Ausgangspunkt der Analyse waren über 1 Million Benutzer und 5 Millionen Sessions.

2. Datenfilterung mit SQL:
Zunächst wurde eine Vorselektion durchgeführt, um nur relevante Datensätze zu berücksichtigen. Es wurden nach Absprache mit Elena ausschließlich aktive Benutzer einbezogen, die seit Januar 2023 mehr als 7 Sessions hatten. Dadurch reduzierte sich die Anzahl der Benutzer auf 5998 und die Anzahl der Sessions auf 49211. Diese Filterung erfolgte mittels SQL-Abfragen, um Daten effizient zu bereinigen und die Grundlage für die nachfolgende Analyse zu schaffen.

3. Berechnung von Metriken in Python:
Nach der SQL-Filterung wurden in Python verschiedene Metriken berechnet, die das Verhalten und die demografischen Merkmale der Benutzer quantifizieren. Dazu gehörten:

- **Alter** (*age*)
- **Nutzungsdauer seit Registrierung** (*user_since*)
- **Gesamtzahl der Reisen** (*reisen_gesamt*)
- **Stornierte Reisen** (*stornierte_reisen*)
- **Durchschnittliche Zimmerbuchungen pro Reise** (*rooms_avg*)
- **Durchschnittliche Aufenthaltsdauer in Nächten** (*nights_avg*)
- **Durchschnittlicher Hotelpreis pro Nacht** (*hotel_price_avg*)
- **Anzahl der gebuchten Hotels** (*hotel_count*)
- **Durchschnittliche Anzahl der gebuchten Sitzplätze pro Reise** (*seats_avg*)
- **Durchschnittlicher Flugpreis pro Reise** (*flight_price_avg*)
- **Anzahl der Flugbuchungen** (*flight_count*)
- **Durchschnittliche Anzahl aufgegebenes Gepäck pro Reise** (*checked_bags_avg*)
- **Durchschnittliche Sitzungsdauer in Sekunden** (*session_dauer_avg_sec*)

4. Datenvisualisierung und Analyse:
Zur besseren Erkennung von Mustern wurden die berechneten Metriken mit Seaborn visualisiert. Es wurden Histogramme, Scatterplots, Pairplots und eine Korrelationsmatrix erstellt. Dies half dabei, Zusammenhänge zwischen den Variablen zu identifizieren und Trends im Reiseverhalten der Kunden zu erkennen.

5. Datenbereinigung und Skalierung:
Während der Analyse wurden fehlende Werte (*NaN*) untersucht und, falls notwendig, mit *dropna()* Funktion entfernt oder *mit fillna()* durch 0 ersetzt. Ein Sonderfall betraf 105 Datensätze, in denen die Anzahl der Übernachtungen negativ war. Diese Werte wurden auf positive Zahlen korrigiert. Anschließend wurden alle numerischen Werte mit dem StandardScaler normalisiert, um eine einheitliche Skalierung für die Clusteranalyse sicherzustellen.

6. Clusterbildung mit PCA und K-Means:
Zur Gruppierung der Kunden wurde eine Hauptkomponentenanalyse (PCA) durchgeführt, um die Dimensionalität der Daten zu reduzieren und die wichtigsten Merkmale hervorzuheben. Danach wurde der K-Means-Algorithmus mit fünf Clustern angewendet. Die resultierenden Gruppen wurden detailliert analysiert und ergaben folgende Kundensegmente:

Rentner

- Viel Zeit, aber wenig Einkommen
- Längste Sessiondauer
- Reisen überdurchschnittlich viel
- Stornieren wenig
- **Prämie:** Exklusive Discounts, damit sie bei ihren langen Sessions attraktive Angebote finden und auf der Seite bleiben.

Junge Leute mit minimalem Reiseverhalten

- Reisen am wenigsten, minimale Anzahl an Hotels und Sitzplätzen
- Höchste Stornierungsrate
- Wenig Einkommen und wenig Zeit
- Kürzeste Sessiondauer
- **Prämie:** Keine Stornogebühren, um stressfreie und bedenkenlose Reisebuchungen zu ermöglichen. Alternativ eine extra Nacht im Hotel, um sie zu einer (ersten) Kurzreise zu motivieren.

Frau mit Kind

- Überdurchschnittliche Anzahl an Sitzplätzen und Kinder
- Maximale Anzahl an Gepäckstücken
- Andere Metriken im Durchschnittsbereich
- Größte Gruppe mit 2,4K von 6K
- **Prämie:** Ein kostenloses Gepäckstück zur Erleichterung der Reiseplanung.

Business Traveller

- Ausschließlich männliche Reisende
- Reisen viel
- Überdurchschnittliche Anzahl an Hotel- und Flugbuchungen
- **Prämie:** Kostenloses Essen im Flugzeug, damit sie auf Geschäftsreisen Zeit sparen und nicht extra essen gehen müssen.

Langbleiber

- Maximale Aufenthaltsdauer von 9 Nächte (bei 3 als Durchschnitt)
- Höchste Ausgaben für Hotels
- Sehr wenige Flüge
- Kurze Sessiondauer
- **Prämie:** Eine zusätzliche kostenlose Nacht im Hotel, um ihren Aufenthalt noch attraktiver zu machen.

7. Ergebnisse und Nutzen:

Der erste Entwurf der Segmentierung kann noch ungenau sein. Eine Verbesserung kann durch Anpassung der Auswahlkriterien erfolgen, indem mehr Reisende in die engere Auswahl genommen werden, beispielsweise durch eine Reduzierung der Mindestanzahl an Sessions oder eine Ausweitung des analysierten Zeitraums. Dadurch könnten weitere Kundengruppen entstehen, wie z. B. Langzeitreisende, Kurztrip-Reisende, Städtereisende oder Geburtstagsreisende. Zudem wäre eine Kundenbefragung sinnvoll, um weitere potenzielle Prämien zu identifizieren und das Programm weiter zu optimieren.