Concept for Customer retention study

What kind of model should be estimated

It should be a binary classification model.

What is the dependent variable and how is it calculated

The dependent(target) variable is a parameter that will estimate if a customer is a repeater or has churned. It is labeled as Retention in the notebook and takes two values : **Repeater: 1, Churn : 0**.
The calculation of the dependent variable can be summarized as follows:

- Find the **time period(in months)** between the first booking date and the last booking date for each customer.
- Calculate the average time(in months) required for 1 booking which is the **time period(in months) / no of bookings**.
- Compute the **time period of inactivity** for each customer: **current date(the latest date in the booking dates column) - last booking date of a customer**
- Now a business and a domain-specific rule needs to be defined such as a frequency metric f: how many frequencies can be skipped, such that we can define a customer as churned.

```
if average time(in months) * f < time period of inactivity then
    Churned
else
    Repeating
end if
```

It's a little bit hard to define the value of the frequency parameter f. A customer cannot be labeled as churn because they stopped booking for some time although they were quite active before. There might be many reasons such as seasonal, traveling less frequently, customer's health and economic situation, etc.

Example: Let's say a customer books a trip in 2 months on average and we choose the skipping frequency to be 2. So, if the customer doesn't book for 4 months in a row, we define that customer as churned.

What would be the structure of the final data frame used to estimate the model

It will have 22 feature columns(after feature selection and creation) and 1 target variable.

What could be possible features, how could they be created

The possible features could be gender, country, age group, subscription to advertising or not, has a club program membership or not, number of booked nights on the trip, the distance of the trip, number of additional services taken during the trip, number of bookings in the past.

A few features can be modified such as country. Since a majority of the people are from Germany DE, we can label the remaining people as 'Outside DE'.

Price per customer for the booked nights in a trip can be created since the total price and no of persons in a trip is given.

Which quality checks should be made to check the quality of the variables

- The first check would be to remove duplicate samples since the data does have a lot of duplicates.
- Data Drift: The age group of a repeating customer might change after a while. So the data needs to be updated regularly. The same goes for the features such as a subscription to advertising or not, a club program membership or not, and Length of stay(Intended vs Actual). These can change due to seasonality and change in consumer preference.
- The dependent variable(retention) can change since it depends on the business rule (frequency metric) defined before.

How would you test if the model has a good performance

The data in this case study is quite imbalanced (churn: repeat ratio is 90:10). Since we are implementing a NBO system that predicts which of the previous customers have a high probability of returning in the next year, we want to correctly identify the minority class(repeat customer). Hence, using precision, recall, and F-1 scores is more feasible. For this specific case, precision seems to be more important since we want to correctly classify the minority class.

KPIs

- Customer retention rate: The company would want to increase the retention rate since most of the revenue comes from these customers.
- Customer churn rate: The company also wants to monitor the churn rate to reduce marketing costs. To actually improve retention, the company needs to find out what's really causing people to leave.
- Customer lifetime value: If the retention rate is increasing, this will increase over time and can be used as a good indicator.

Next Best destination

Each customer-destination pair will have a certain probability associated with it. This can be estimated by the intersection of the customer's travel history and the destination's data from other customers. A model such as collaborative filtering can be built to predict the next best destination for that particular customer.

Next Booking date

Several features can be created to build this model such as :

- Average no of days between bookings
- Minimum no of days between two bookings
- Maximum no of days between two bookings
- No of Days from the first booking
- No of Days from the last booking
- Day/month when the customer is more likely to book
- Total revenue from a customer

Now we can build a regression model which can output the number of days to the customer's next booking. We can fine-tune this by looking at the distribution of bookings for each month in a year and selecting an appropriate time period to send an additional incentive at the right moment.