**Data Preprocessing and Feature Engineering:**

The dataset contains important features that mostly are correlated with chances of booking. First, some columns that contain high number of null values (such as: **occupation\_id\_desc**, **last\_nps\_segment**, **flag\_bkd\_cxl\_last3y**, and **num\_cruises\_last3y**) and date columns are dropped. The different strategies are performed to handle remaining NULL values with the aim to lose data in the minimum level.

* Keeping NULL values in categorical columns. In this way, when the encoding is done, the NULL will be considered as one category, and the reality of feature distribution will not be changed.
* Segmenting and transforming some numerical features with NULL values to categorical features to not fulfill the NULL values (such as: **sent\_last\_month** and **days\_after\_cruise\_end**).
* Fulfilling some numerical features NULL values with mode and median based on their distribution (such as: **flag\_app, last\_cruise\_bkg\_anticpiation**, and **last\_cruise\_duration**).

The idea is to use all the features in the feature engineering and model training to differentiate insightful features from redundant ones. After performing different analysis, the following features are selected as input of algorithm:

**had\_a\_flight\_in\_last\_cruise**, **BKG\_Channel**, **rfm\_segment**, **period\_after\_cruise\_end**, **last\_paid\_cabin\_meta\_cat\_code**, **category\_open\_last\_month**, **category\_sent\_last\_month**, **AGE**, **LOYALTY\_TOTAL\_SCORE**, **is\_loyalty**, **NUMBER\_OF\_CRUISES**, **pcc\_flag**, **flag\_app**, **last\_cruise\_duration**, **last\_cruise\_bkg\_anticipation**.

**Model**:

Another problem to handle is the imbalanced target column which causes the inability of models to predict the positive class (which are potential customers). For fixing this problem, while trying different classification models with hyperparameters tunning, **SMOTE** and **Class Weight** is used. The best model is **RandomForestClassifer** using Class Weight to balance the target classes. 15% of dataset (13500 rows) is used for testing. The confusion metrics is as follows:

A chart of a blue yellow and purple box

Description automatically generated with medium confidence

Finding the customers that want to book is the most important KPI. So, it was the baseline for optimization of algorithms. The final version could predict 57% of clients that booked (97 out of 170). The metrics can be improved by some new feature engineering approaches.

**Deployment:**

A simple deployment is done using **Streamlit** which accepts a .CSV file with the same columns as furnished dataset and gives a table with **ID** and **predicted booking probability** as output.