**1. Background**

The hotel industry encompasses a wide range of establishments that provide accommodation, meals and various services to guests or travelers. It plays a crucial role in the global economy and tourism sector. In recent years, there has been a growing emphasis on sustainability and eco-friendly practices within the hotel industry. Many hotels are implementing initiatives to reduce energy consumption, minimize waste, promote responsible tourism, and support local communities.

**2. Business Problem**

The hotel industry experiences demand from both frequent business travelers and leisure tourists. Our objective is to utilize hotel booking data to forecast potential cancellations. We aim to address the following business inquiries:

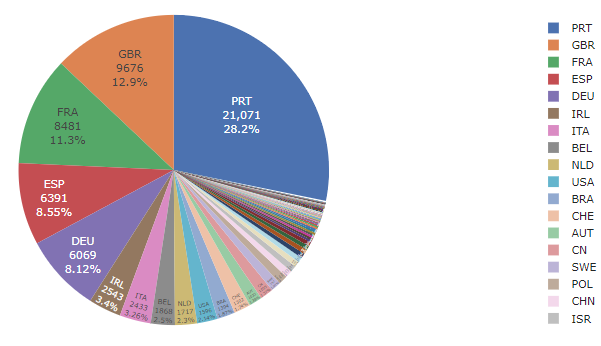
1. Would a customer cancel his booking?
2. What are the primary factors influencing booking cancellations?

**3. Data Overview**

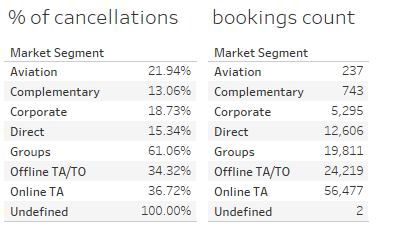
The data for this study was sourced from Kaggle: Hotel booking demand datasets. The dataset primarily consists of two types of hotels: the resort and the city hotel. It comprises 119,391 rows and 32 features.( Label: Is\_Cancelled, Predictors: Hotel, Lead time, Arrival date year, Adults, Children, Babies, Meal, Country and 24 others) The dataset encompasses bookings by customers scheduled to arrive between 1st July 2015 and 31st August 2017, and includes information on whether the booking was canceled or not. Additionally, It provides details such as the duration of the stay, the number of adults, the nationality of the guests, hotel amenities like parking and meal plans and the arrival date of the booking.

**4. Methodology**  
**4.1 Exploratory Data Analysis**

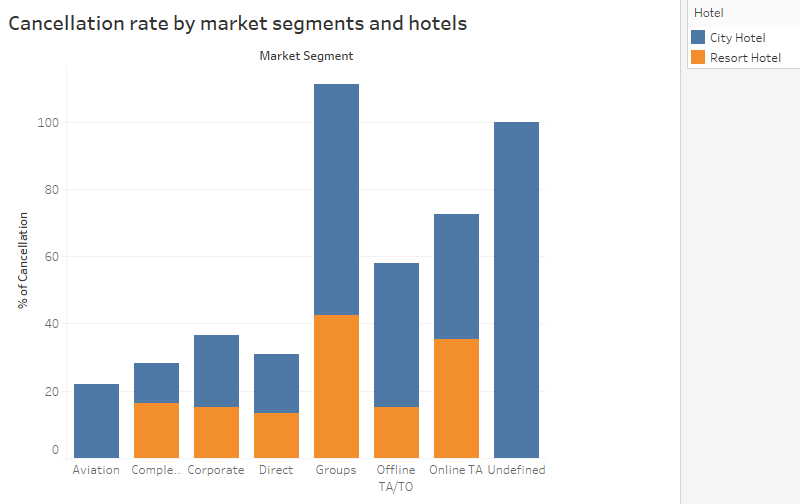
To gain deeper insights into the data, we utilized Python for visualization purposes. This enables us to understand the relationship between the predictors and the label more effectively. Through this analysis, we developed hypotheses to determine whether each feature could potentially be an influential factor in the context of cancellations.



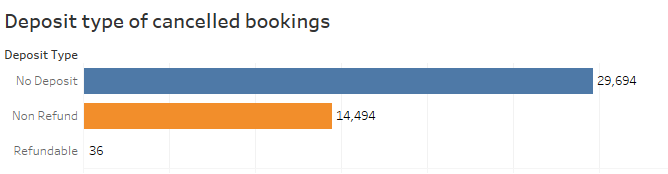
The pie chart illustrates that most visitors are from Europe, namely Portugal, UK and France.



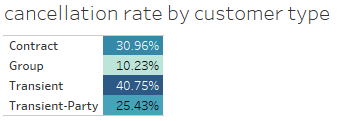
The above table presents that the top segment is the online travel agent(OTA) segment and the least is from the Aviation segment. The cancellation rate is lowest for the Complementary segment and reaches highest from the Groups segment at 61%. The reason for the potentially high cancellation rate in this segment is attributed to the bulk bookings made for events, conferences, and similar occasions. Consequently, if an event is canceled, a substantial number of bookings associated with it would also be canceled.



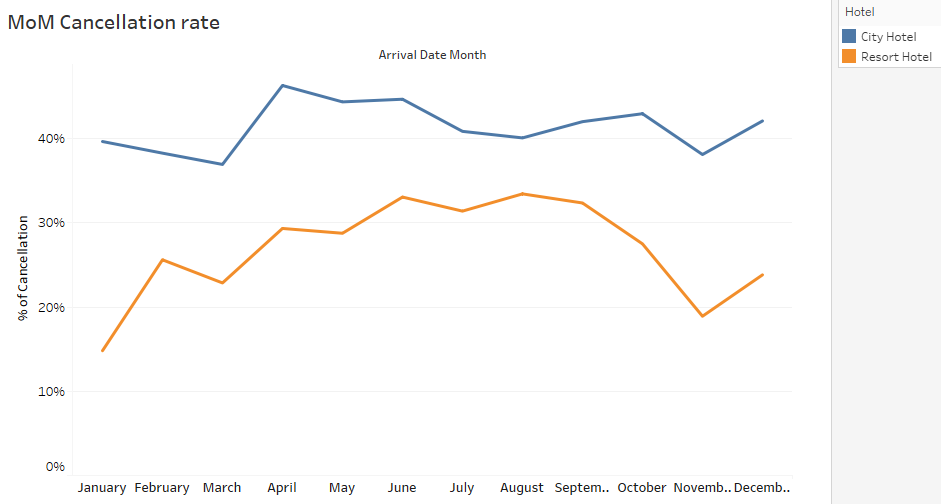
There are high rates of cancellations at both Resort and City hotels. Compared to the Resort Hotel, the City Hotel has a higher cancellation rate. This reasoning is logical since resort hotels typically attract leisure travelers who have carefully planned their vacations. As a result, the likelihood of trip cancellations is significantly lower compared to city hotels, which are primarily booked by business travelers.



The above chart shows that the majority of the canceled reservations did not require a deposit.



It was observed from the above table that transient customers account for the highest cancellation rate while the group guests have the lowest. This makes sense since it is not a pre-planned event for transient travelers and the risk of getting canceled is high.



The line chart presents that city hotels have more cancellation rates compared to resort hotels in all months. It is noted that the cancellation rate of city hotels is almost equal to other months even in winter. While the cancellation rate in winter months is low for resort hotels.

**4.2 Data Preprocessing**

Since data types are various, so we use the following steps to clean them:

* Handle missing values: We replaced missing values of children, country, agent and company features with ‘0’ and ‘Unknown’ values.
* Correlation analysis: Check if it consists of variables highly correlated
* Feature Engineering: Convert reservation status date to ‘year’, ‘month’ and ‘day’
* One Encoding: We encoded categorical variables into numerical independent variables for models such as logistic regression.

**5. Model Building**

The methods I used to answer these questions are models used for classification:

Logistic Regression – This is a classification model used to allocate observation to binary classes by providing a probability value.

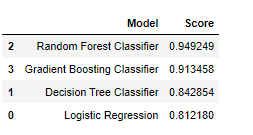
Decision Tree – This is a supervised machine learning algorithm that can be used for both classification and regression tasks. It represents a flowchart-like structure where each internal node represents a feature or attribute and each branch represents a decision based on that feature.

Random Forest – The ensemble nature of this tree model enhances the accuracy of the results. Combining multiple individual models leverages the strength of each component. Random selection of variables for each model reduces correlation among the trees, thus improving the overall performance.

Gradient Boosting – Similar to the Random Forest model, this model is an ensemble learning method that combines multiple weak prediction models. But gradient boosting builds an ensemble of weak models sequentially, with each model learning to correct the mistakes made by the previous models.

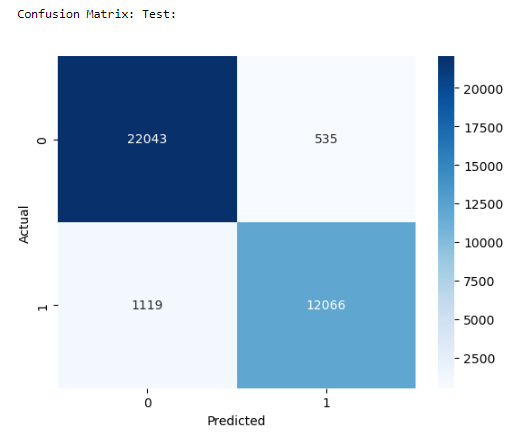
I split the data into test and train data sets – 30%, 70% respectively.

Next, I evaluated the accuracy score of the test data and compared all four models. The results are shown as follows:

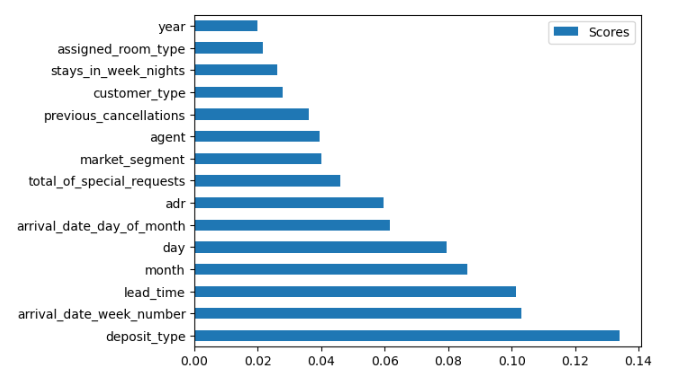


It is evident that the Random Forest model achieved the highest accuracy score( around 95%) and highest Precision, Recall and F1-Score. On the other hand, the Logistic Regression obtained the lowest score due to its linear nature.

From the confusion matric of the best model, we can find:



Out of the total observations, 1119 were wrongly classified as non-canceled when they were actually canceled, whereas 535 observations were incorrectly labeled as canceled when they were not.

Following that, we can examine the feature importance using the best model, as displayed below. The variables that had a significant impact on the model can be observed: deposit\_type, arrival\_date\_week\_number, lead\_time, month, adr, total\_of\_special\_requests, etc. 

**6. Recommendation**

Based on the analysis of the Random Forest results and the insights obtained from the feature importance values, we propose the following recommendations to hotel business owners:

* Utilize the model to obtain the predicted number of cancellations per day, and based on this, calculate the optimal overbooking limit that maximizes returns while considering the capacity and booking limit.
* The probability of cancellation decreases with an increase in the total number of special requests. Therefore, being accommodating in terms of offerings appears to incentivize customers to proceed with payment. Utilizing our model to calculate the true demand would enable us to better accommodate these requests.
* A significant proportion of reservations were made without a deposit, suggesting that guests are more likely to cancel a reservation when a deposit is not required. So the hotel should consider revising the cancellation policy to introduce some restrictions. Additionally, the hotel can conduct a brief survey to gather feedback from customers when they cancel a reservation, aiming to identify the underlying reasons for the cancellations. (especially for winter)
* When considering the market segment, it has been observed that Online Travel Agents (TA) and Tour Operators (TO) have demonstrated the strongest performance. So the utilization of these marketings could be encouraged by offering a special promotion or incentive.
* In terms of feature importance, rank details within each feature group should be observed in further analysis.