**TEBONG ROLAND**

**DSC680-T301 Applied Data Science (2243-1)**

[**11.1 Project 3: Draft/Milestone 2**](https://cyberactive.bellevue.edu/webapps/assignment/uploadAssignment?content_id=_15084592_1&course_id=_522998_1&group_id=&mode=view)

**Hotel Booking Cancellation Prediction**

**1. Business Problem**

A significant number of hotel bookings are called off due to cancellations or no-shows. The typical reasons for cancellations include changes of plans, scheduling conflicts, etc. This is often made easier by the option to do so free of charge or preferably at a low cost which is beneficial to hotel guests but it is a less desirable and possibly revenue-diminishing factor for hotels to deal with. Such losses are particularly high on last-minute cancellations.

The new technologies involving online booking channels have dramatically changed customers' booking possibilities and behavior. This adds a further dimension to the challenge of how hotels handle cancellations, which are no longer limited to traditional booking and guest characteristics.

The cancellation of bookings potentially impacts a hotel on various fronts:

1. Loss of resources (revenue) when the hotel cannot resell the room.
2. Additional costs of distribution channels by increasing commissions or paying for publicity to help sell these rooms.
3. Lowering prices at last minute, so the hotel can resell a room, resulting in reducing the profit margin.
4. Human resources to make arrangements for the guests.

**2. Background/History**

The increasing number of cancellations calls for a Machine Learning based solution that can help in predicting which booking is likely to be canceled. INN Hotels Group has a chain of hotels in Portugal, they are facing problems with the high number of booking cancellations and have reached out to your firm for data-driven solutions. We have to analyze the data to find which factors have a high influence on booking cancellations, build a predictive model that can predict which booking is going to be canceled in advance, and help in formulating profitable policies for cancellations and refunds.

**3. Data Explanation**

The data contains the different attributes of customers' booking details. The detailed data dictionary is given below.

* Booking\_ID: The unique identifier of each booking
* no\_of\_adults: The number of adults
* no\_of\_children: The number of children
* no\_of\_weekend\_nights: The number of weekend nights (Saturday and Sunday) the guest stayed or booked to stay at the hotel
* no\_of\_week\_nights: The number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel
* type\_of\_meal\_plan: The type of meal plan booked by the customer:
  + Not Selected – No meal plan selected
  + Meal Plan 1 – Breakfast
  + Meal Plan 2 – Half board (breakfast and one other meal)
  + Meal Plan 3 – Full board (breakfast, lunch, and dinner)
* required\_car\_parking\_space: Does the customer require a car parking space? (0 - No, 1- Yes)
* room\_type\_reserved: The type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels.
* lead\_time: The number of days between the date of booking and the arrival date
* arrival\_year: The year of arrival date
* arrival\_month: The month of arrival date
* arrival\_date: The date of the month
* market\_segment\_type: Market segment designation.
* repeated\_guest: Is the customer a repeated guest? (0 - No, 1- Yes)
* no\_of\_previous\_cancellations: The number of previous bookings that were canceled by the customer before the current booking
* no\_of\_previous\_bookings\_not\_canceled: The number of previous bookings not canceled by the customer before the current booking
* avg\_price\_per\_room: The average price per day for the reservation; prices of the rooms are dynamic. (in euros)
* no\_of\_special\_requests: The total number of special requests made by the customer (e.g. high floor, view from the room, etc.)
* booking\_status: Flag indicating if the booking was canceled or not. The class 0 represents the Not\_Canceled whereas class 1 represents the Canceled label.

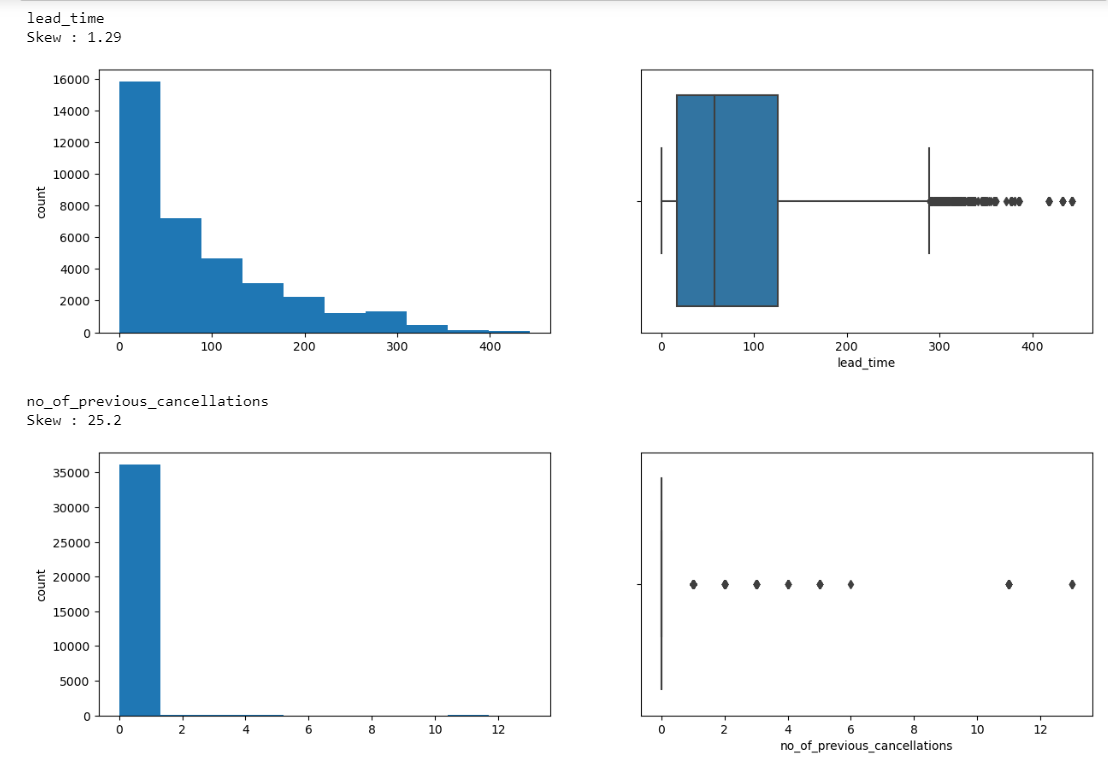
**4. Methods & Analysis**

**4.1 Exploratory Data Analysis (EDA):**

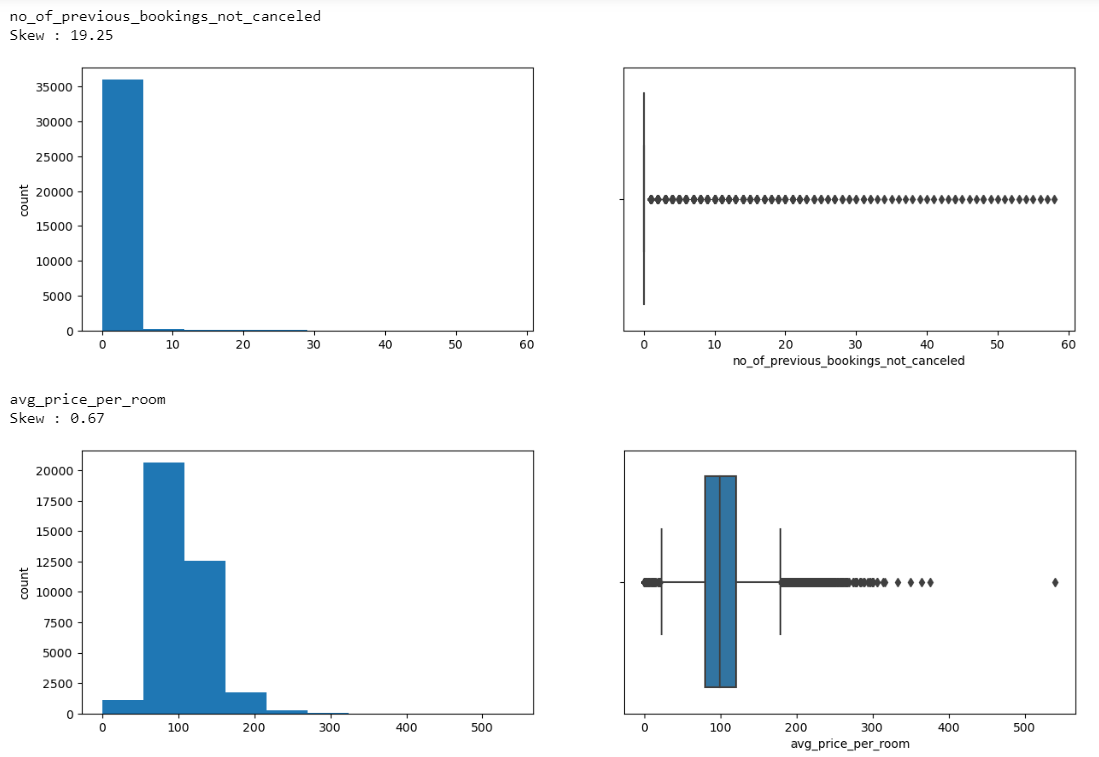
Let's check the statistical summary of the data.

* The number of adults booking the hotel ranges from 0 to 4, which is usual. At least 75% of bookings have 2 adults.
* Majority of the bookings have no children. This indicates that the most of the bookings might be from single people or young couples. The maximum value in the number of children column is 10, which is a bit unusual and might be an outlier.
* The maximum values for the number of weeks and weekend nights are 7 weekends or 17 nights. These might be very long stays. We can explore these values further.
* At least 75% of the customers do not require a car parking space. This indicates that the majority of the customers are traveling by personal vehicles.
* On average, the customers book 85 days in advance. There's also a very huge difference in 75th percentile and the maximum value which indicates that there might be outliers present in this column.
* The arrival year shows that we have the data from two years - 2017 and 2018.
* At least 75% of the customers are not repeating customers.
* At least 75% of bookings have no previous cancellations. The number of previous cancellations has a maximum value of 13.
* The mean of the column avg\_price\_per\_room is 103 euros. There's a huge difference between the 75th percentile and the maximum value which indicates there might be outliers present in this column. There are also 0 values in this column, let's check these values.

Let’s check the distribution and outliers for numerical columns in the data:



* The distribution of **lead time is right-skewed** implies the majority of customer make bookings close to the arrival date. Many customers have made the booking on the same day of arrival as well. There are many outliers, **some customers made booking more than 400 days in advance**.
* **Very few customers have more than one cancellation**. Some customers canceled more than 12 times.



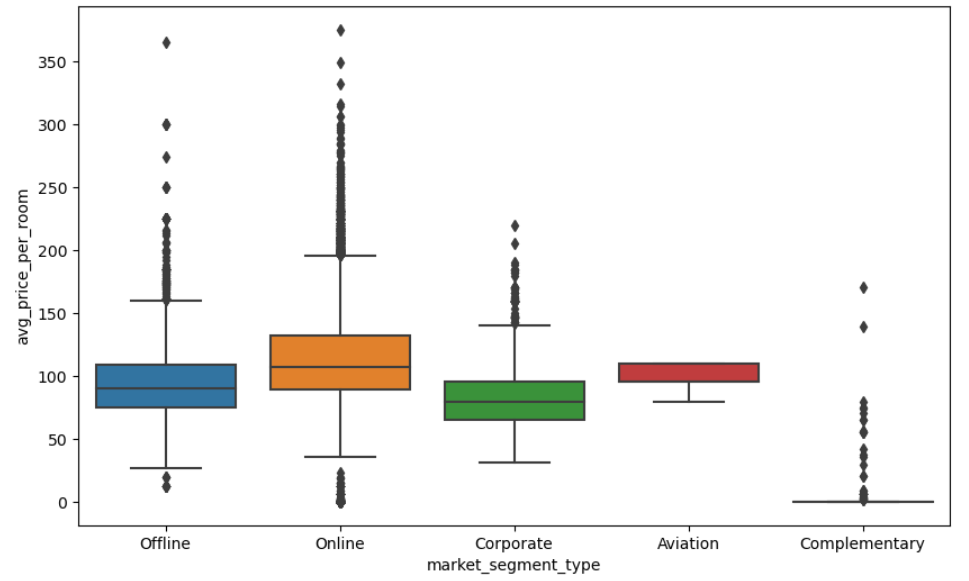
* **Very few customers have more than 1 booking not canceled previously**.
* **The distribution of average price per room is skewed to right**. The boxplot shows that there are outliers on both sides. The median price of a room is around ~100 euros. There is 1 observation where the average price of the room is more than 500 euros. This observation is quite far away from the rest of the values.

Let’s check the percentage of each category for categorical variables.

* ~72% of the bookings were made for 2 adults. There are some observations with 0 adults as well, it is possible that an adult made the booking for children.
* 93% of the customers didn't make reservations for children. There are some values in the data where the number of children is 9 or 10, which is unusual. We will replace these values with the maximum value before 9, i.e., 3 children.
* Most bookings are made for 2 nights (~31.5%) followed by 1 night (26.1%). A very less proportion of customers made the booking for 8 or more days.
* 46.5% of the customers do not plan to spend the weekend in the hotel. The percentage of customers planning to spend 1 or 2 weekends in the hotel is almost the same.
* ~97% of the customers do not require a car parking space.
* Most of the customers prefer meal plan 1 that is only breakfast. ~14% of the customers didn't select a meal plan.
* ~77% of the customers booked Room\_Type 1 followed by ~17% of the customers booking Room\_Type 4.
* October is the busiest month for the hotel followed by September. ~14.7% of the bookings were made in October. It would be interesting to see whether months with a higher number of bookings have higher cancellations as well or not.
* ~64% of the hotel bookings were made online followed by 29% of the bookings which were made offline.
* ~54.5% of the customers generally do not make any special requests while booking a hotel room.
* ~32.8% of the bookings were canceled by the customers. We can encode Canceled bookings as 1 and Not\_Canceled as 0 for further analysis.

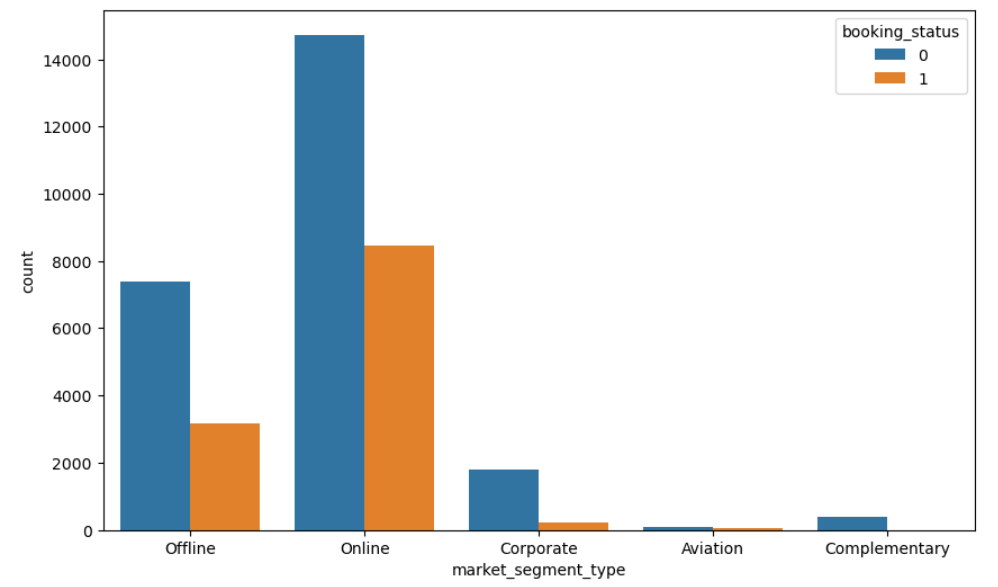
We are done with univariate analysis and data preprocessing. Let's explore the data a bit more with bivariate analysis.

Let's check the relationship of market segment type with the average price per room.



* **Rooms booked online have the highest variations in prices.**
* The distribution for offline and corporate room prices are almost similar except for some outliers.
* Complementary market segment gets the rooms at very low prices, which makes sense.

Let's see how booking status varies across different market segments. Also, how lead time impacts booking status.



* **Online bookings have the highest number of cancellations.**
* Bookings made offline are less prone to cancellations.
* Corporate and complementary segment also show a very low number of cancellations.

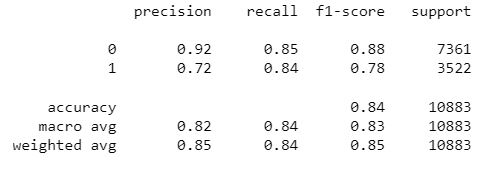
**4.2 Preparing the data for modeling:**

1. Models cannot take non-numeric inputs. So, we will first create dummy variables for all the categorical variables.
2. We will then split the data into train and test sets.

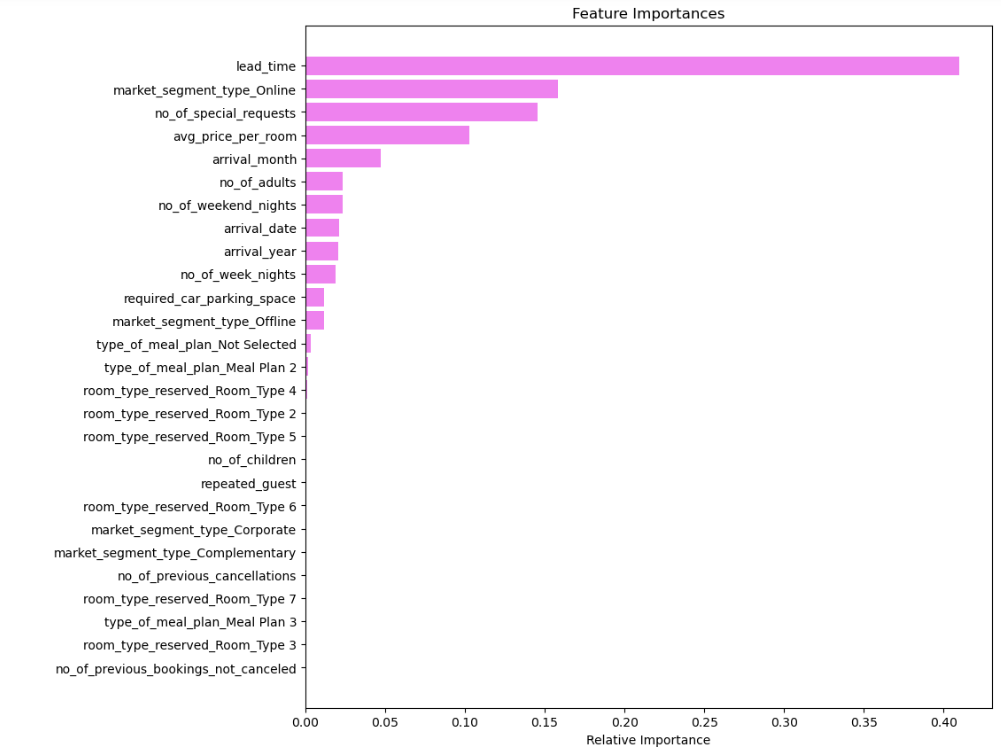
**4.3 Modeling**

Let’s build Decision Tree classifier and Random forest Classifier.

**Decision Tree Classifier:** After fitting the training data into the Decision tree classifier model the model has been tested on the test dataset, which was splitted above in the above section. After testing the model the results are quite promising as shown below.

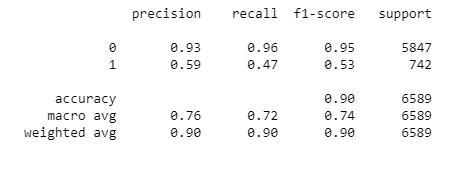


**Let's look at the feature importance** of the tuned decision tree model:



* The lead\_time, market\_segment\_type\_Online, no\_of\_special\_requests and avg\_price\_per\_room are the most important features followed by arrival\_month.
* The rest of the variables have no impact in this model except no\_of\_adults, no\_of\_weekend\_nights, arrival\_date, arrival\_year, no\_of\_week\_nights, required\_car\_parking\_space, and market\_segment\_type\_Offline while deciding whether a hotel booking will be canceled or not.

**Random forest Classifier:** After fitting the training data into the random forest classifier model the model has been tested on the test dataset, which was splitted above in the above section. After testing the model the results are quite promising as shown below.



**5. Assumptions**

Our approach to developing the predictive model relies on several foundational beliefs. These include expectations of consistent booking behaviors over time, dependable data quality, and uniformity across hotels within our chain, the potential for our model to generalize beyond the provided dataset, and the assumption that identified features indeed influence booking cancellations causally.

**6. Limitations**

Despite our efforts, there are inherent constraints to our predictive modeling approach. These limitations encompass issues like data volume and diversity, potential gaps in feature representation, complexities in model building, oversight of temporal dynamics, and the inability to account for external factors beyond our dataset.

**7. Challenges**

Throughout the predictive modeling process, we encounter various hurdles that demand creative solutions. These challenges encompass balancing accuracy with interpretability, identifying relevant features without overfitting, selecting appropriate evaluation metrics, ensuring scalability, and navigating ethical considerations such as bias, privacy, and fairness.

**8. Future Uses/Additional Applications**

Looking ahead, our predictive models hold promise for diverse applications beyond predicting booking cancellations. These include dynamic pricing strategies, optimizing resource allocation, delivering personalized customer service, managing risks in other sectors, and enhancing customer retention efforts.

**9. Recommendations**

1. The lead time can play a key role in identifying if a booking will be canceled or not. We observed that the bookings where a customer has made the booking well before the date of arrival are more likely to be canceled.
   * The hotel can contact such customers before the arrival date for re-confirmation of their bookings. The response given by the customer will give the hotel ample time to re-sell the room or make preparations for the customer's requests.
2. Stricter cancellation policies can be adopted by the hotel.
   * The bookings where the average price per room is high, and there were special requests associated should not get a full refund as the loss of resources will be high in these cases.
   * Ideally the cancellation policies should be consistent across all market segments but as noticed in our analysis, high percentage of bookings done online are canceled. The booking canceled online should yield less percentage of refund to the customers.

The refunds, cancellation fees, etc. should be highlighted on the website / app before a customer confirms their booking to safeguard guest's interest.

1. The length of stay at the hotel can be restricted.
   * We saw in our analysis that bookings, where the total length of stay was more than 5 days, had higher chances of getting canceled.
   * Hotel can allow bookings up to 5 days only and then customers should be asked to re-book if they wish to stay longer. These policies can be relaxed for corporate and Aviation market segments. For other market segments, the process should be fairly easy to not hamper their experience with the hotel.

Such restrictions can be strategized by the hotel to generate additional revenue.

1. In December and January, the cancellation to non-cancellation ratio is low. Customers might travel to celebrate Christmas and New Year. The hotel should ensure that enough human resources are available to cater to the needs of the guests.
2. October and September saw the highest number of bookings but also high number of cancellations. This should be investigated further by the hotel.
3. Post-booking interactions can be initiated with the customers.
   * Post-booking interactions will show the guests the level of attention and care they would receive at the hotel.
   * To give guests a personalized experience, information about local events, nearby places to explore, etc. can be shared from time to time.
4. Improving the experience of repeated customers.
   * Our analysis shows that there are very few repeat customers and the cancellation among them is very less which is a good indication as repeat customers are important for the hospitality industry as they can help in spreading the word of mouth and require almost no marketing cost. Attracting new customers is tedious and costs more as compared to a repeated guest.
   * A loyalty program that offers special discounts, access to services in hotels, etc. for these customers can help in improving their experience.

**10. Implementation Plan**

To effectively integrate our predictive models into our operational workflow, we have devised a comprehensive implementation plan. This involves continuous data collection, iterative model development, seamless integration with existing systems, staff training, and ongoing performance monitoring.

**11. Ethical Assessment**

We carefully evaluate the ethical implications of our predictive modeling endeavors. This entails mitigating biases, safeguarding customer privacy, ensuring transparency and accountability, obtaining informed consent, and considering broader societal impacts to uphold ethical standards throughout our model lifecycle.

**Questions:**

1. What are the primary reasons for hotel booking cancellations according to the provided data?
2. How do changes in lead time influence the likelihood of booking cancellations?
3. Which market segment type exhibits the highest rate of booking cancellations?
4. What are the most influential factors in predicting booking cancellations based on the decision tree classifier?
5. How does the distribution of the number of adults booking the hotel vary in the dataset?
6. What is the average lead time for hotel bookings in the provided data?
7. Which month shows the highest number of hotel bookings according to the dataset?
8. How does the average price per room impact the likelihood of booking cancellations?
9. What percentage of customers require a car parking space based on the data?
10. How do offline bookings compare to online bookings in terms of cancellation rates?