Star Hotels Group

Business Case

Background

A significant number of hotel bookings are called off due to cancellations or no-shows. The typical reasons for cancellations include change 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 impact a hotel on various fronts:

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

The increasing number of cancellations calls for a Machine Learning based solution that can help in predicting which booking is likely to be canceled. Star 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.

Objective

To analyze the data provided 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.

Data Information

The data contains information about the business problem

Variable	Description	Type of Variable
no_of_adults	Number of adults	int64
no_of_children	Number of Children	int64
no_of_weekend_nights	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel	int64
no_of_week_nights	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel	Int64
type_of_meal_plan	Type of meal plan booked by the customer	object
required_car_parking_space	Does the customer require a car parking space? (0 - No, 1- Yes)	Int64
room_type_reserved	Type of room reserved by the customer	Object
lead_time	Number of days between the date of booking and the arrival date	Int64
arrival_year	Year of arrival date	Int64
arrival_month	Month of arrival date	Int64
arrival_date	Date of the month	Int64
market_segment_type	Market segment designation	Object
repeated_guest	Is the customer a repeated guest? (0 - No, 1- Yes)	Int64
no_of_previous_cancellations	Number of previous bookings that were canceled by the customer prior to the current booking	Int64
no_of_previous_bookings_not_canceled	Number of previous bookings not canceled by the customer prior to the current booking	Int64
avg_price_per_room	Average price per day of the reservation; prices of the rooms are dynamic	float
no_of_special_requests	Total number of special requests made by the customer (e	Int64
booking_status	Flag indicating if the booking was canceled or not	Object

Observations	Variables
56,926	18

Manipulations to Raw Data:

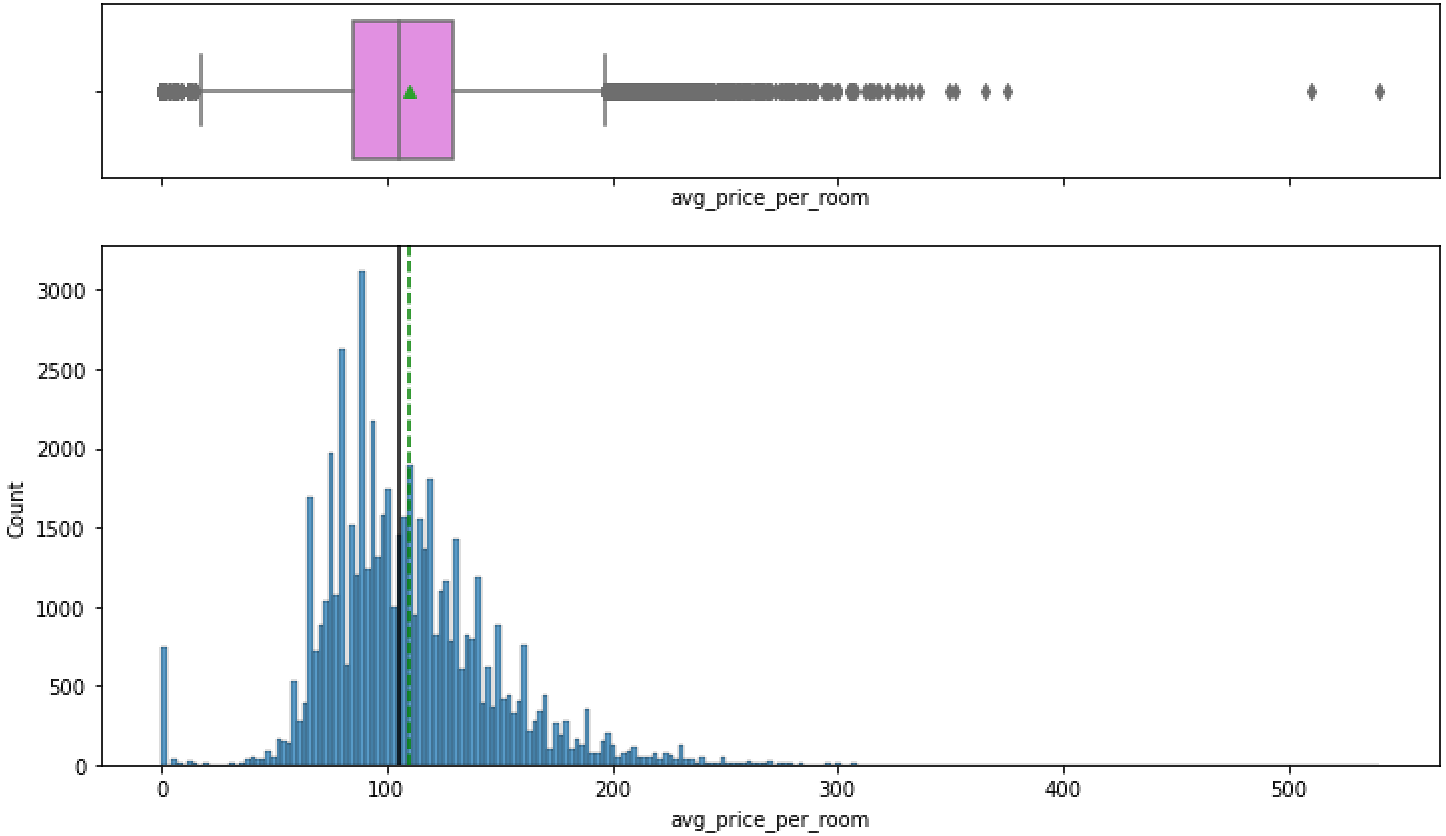
1.Object variables were converted to Category
2.replacing of arrived_month to month names

Exploratory Data Analysis – Average Price per Room

This data contains the average room price

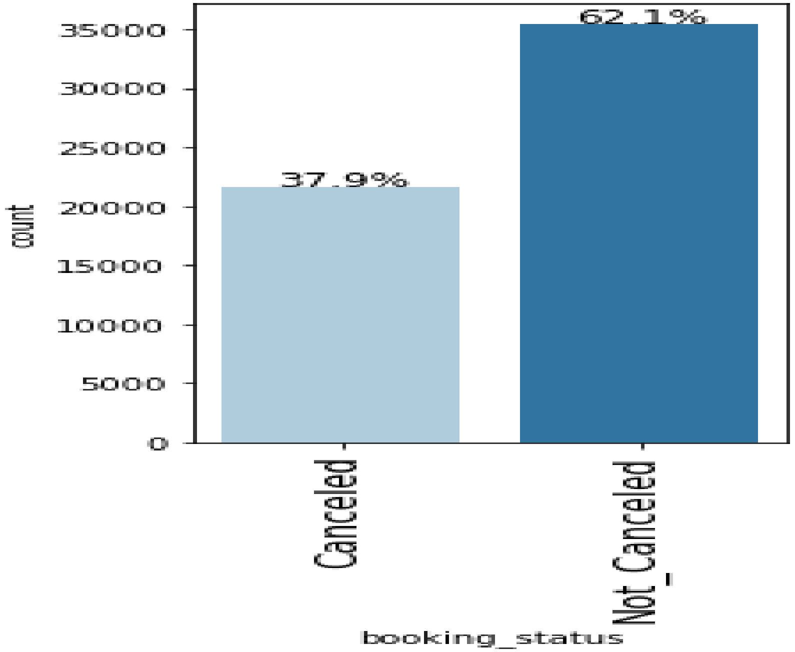
Observations:

1. The data is reasonably right skewed
2. The mean and median are very close to each other, both being slightly above €100

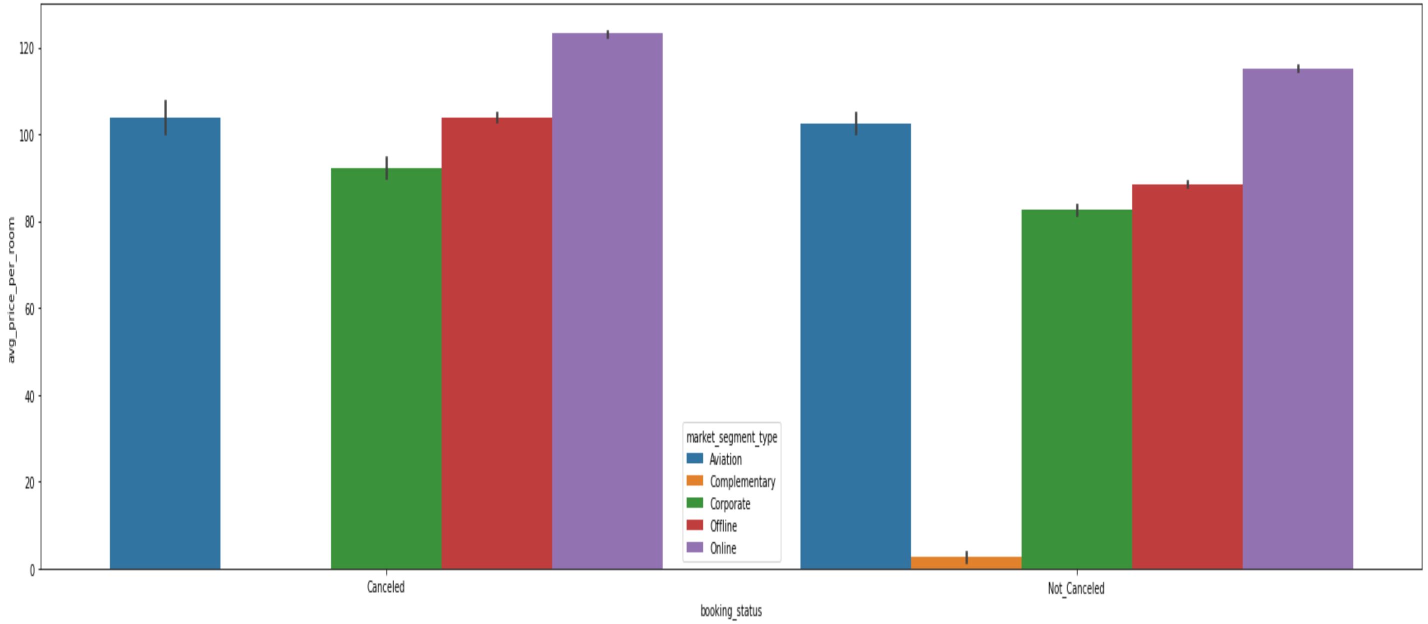


Exploratory Data Analysis – Booking Status

This data contains the Booking Status



Cancellation Status

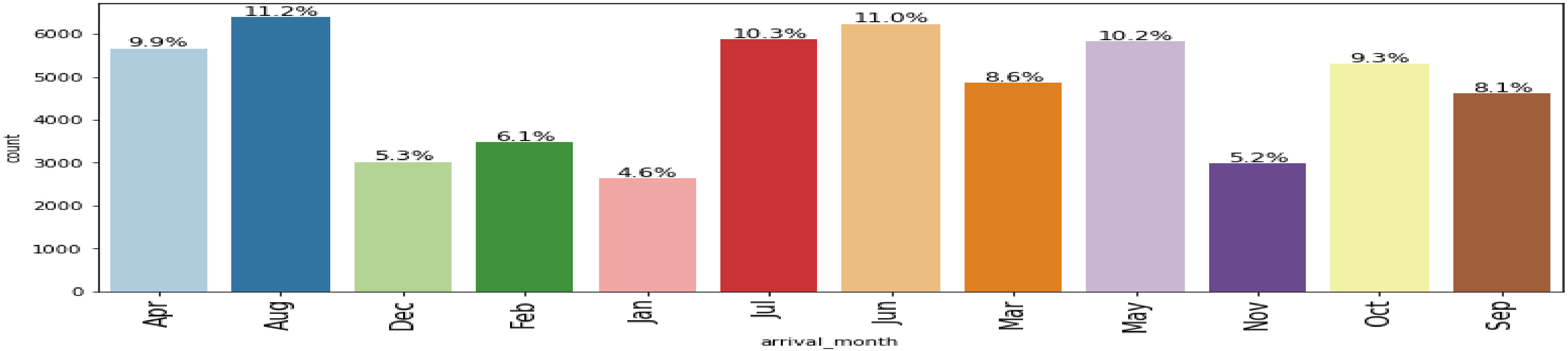


Observations:

1. The total number of cancelled bookings is 37.9% (~21,400)
2. The total number of non-cancelled bookings is 62.1% (~35,160)
3. This indicates that the total number of guests that keep bookings is significantly higher than those that default
4. Online has the most cancelled and non-cancelled (as they have the highest number of users)
5. However, complementary users have no cancellations as they are essentially free guests

Exploratory Data Analysis – Number of Guests

This data contains the number of guests per month



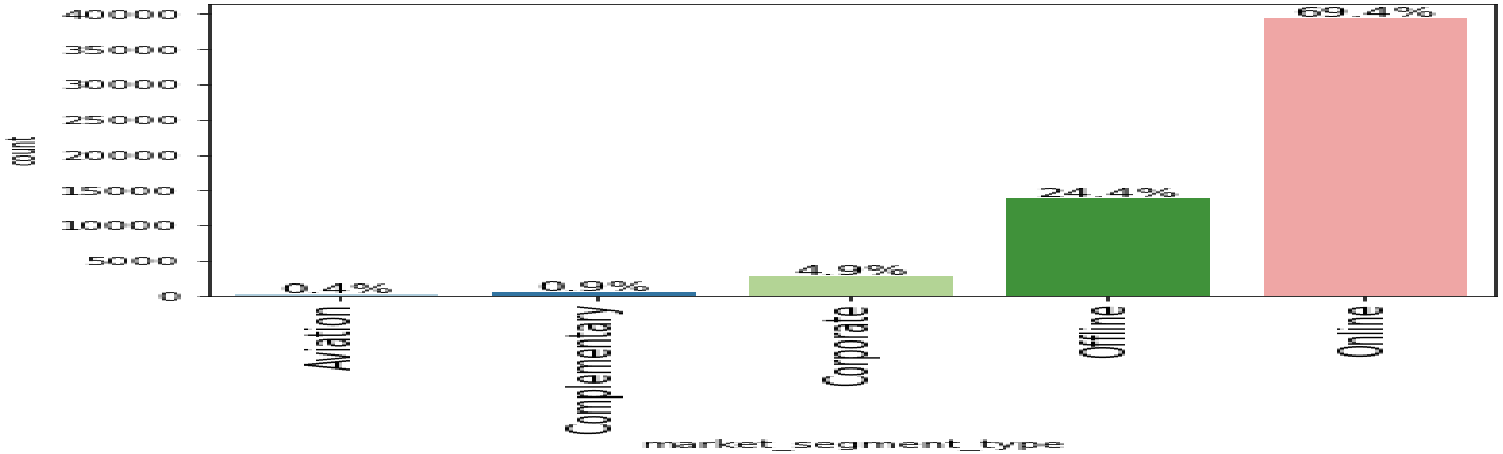
Observations:

1. The busiest month of the year is August with June coming a very close second
2. However, the trend shows that there is a general uptick in bookings from January with a peak in August then a downward trend from November which lasts till January with a pickup again in February
3. The 3 busiest months are between June to August which coincides with the Summer holidays

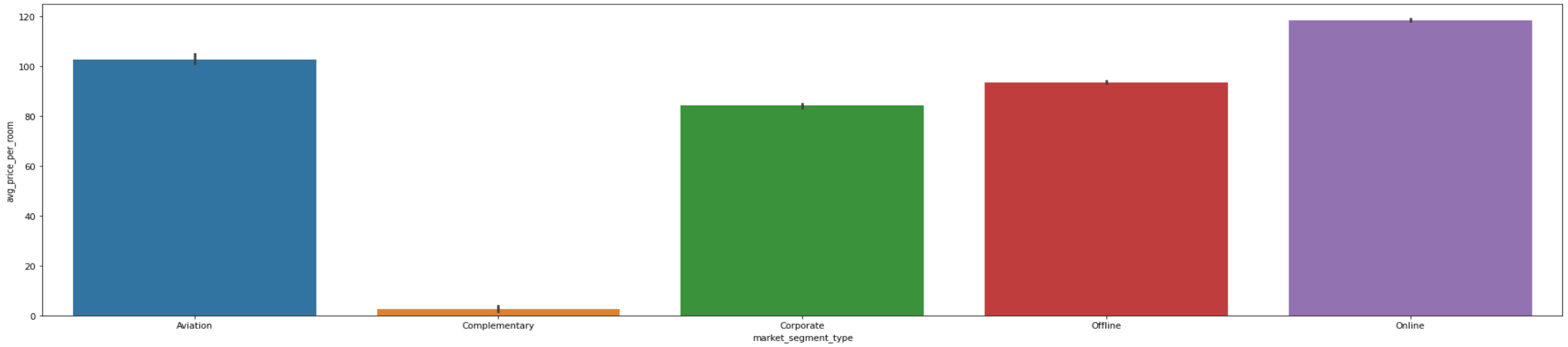
Exploratory Data Analysis – Market Segments

Information about the Market Segments

Percentage split of Market Segments



Market Segments wrt to Room Prices



Observations:

1. Online, with 69.4% has the highest number of guests by a large margin
2. Offline is second with about a quarter of the guests (~24%)
3. The other 3 segments contribute about ~8%

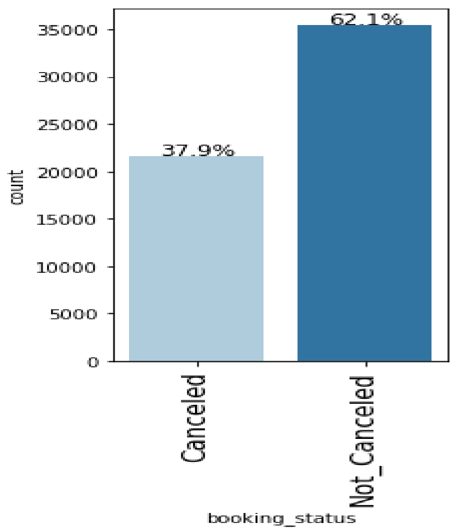
Observations:

1. Online has the highest room prices. Despite this it has the highest number of guests as shown in the previous graph
2. Aviation has the second highest room price
3. Complementary has the lowest room price which makes sense as the segment name suggests it is free or almost free

Exploratory Data Analysis – Booking Status

Information about the Booking Status

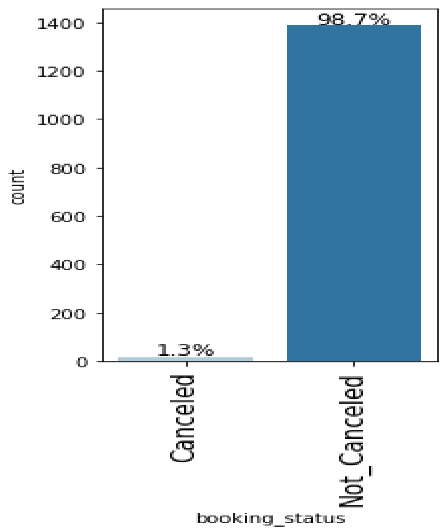
wrt to total number of guests



Observations:

62% of the guests do not cancel which is a high percentage and gives reasonable certainity to the hotel about their bookings

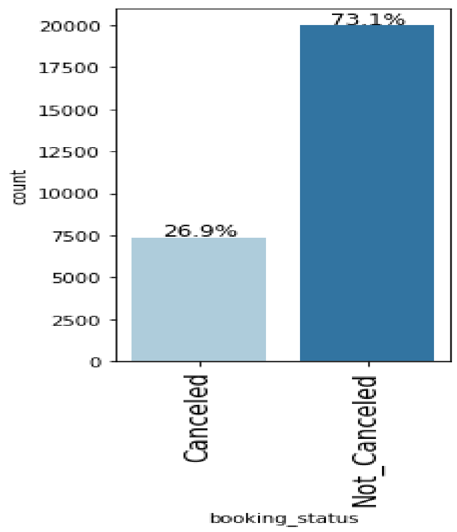
wrt to Repeated customers



Observations:

1. The number of guests under consideration is 1,404
2. For repeating guests, only 1.3% or about 18 guests cancel, which means the repeating guests genrally like the hotel

wrt to special requirements

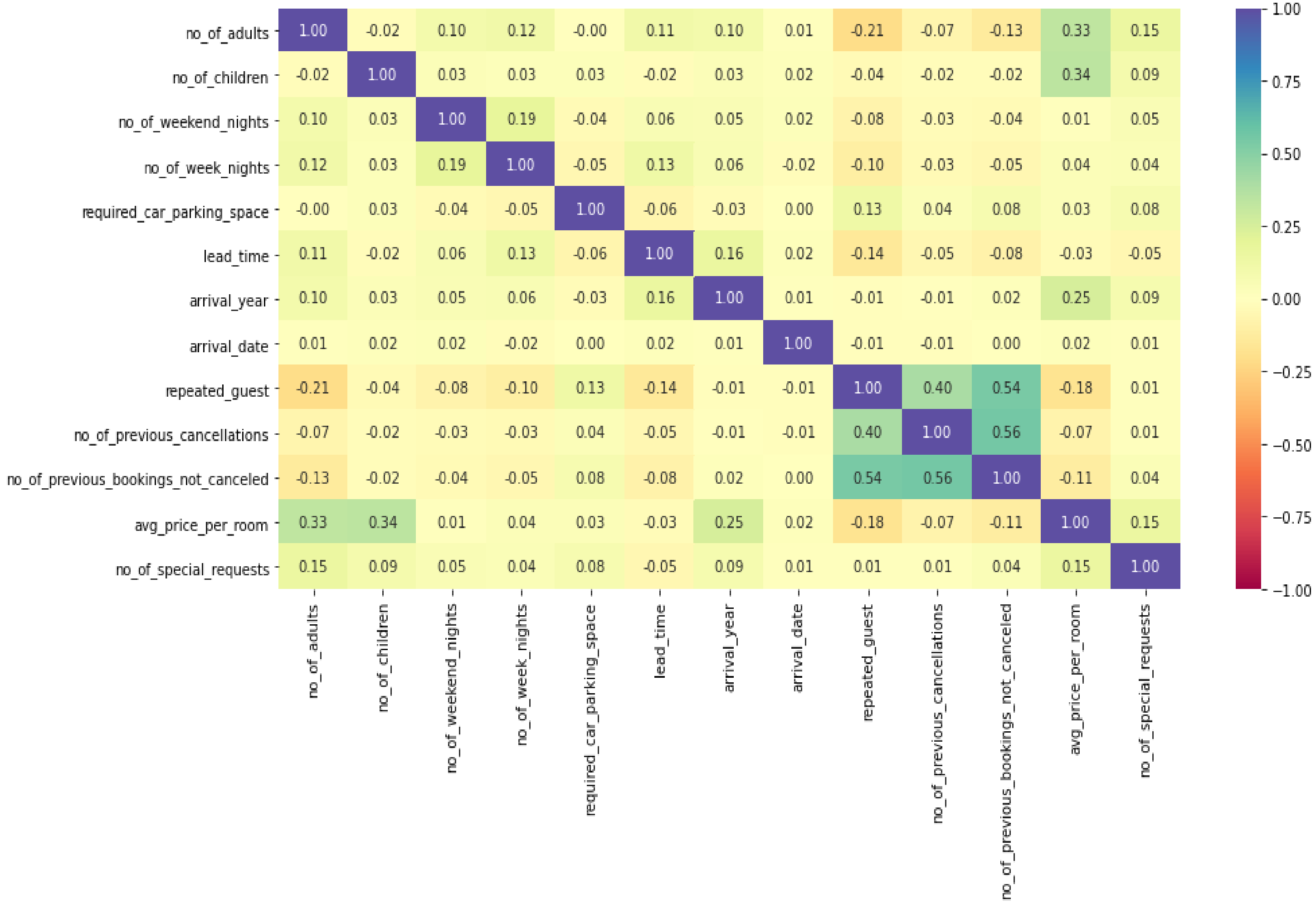


Observations:

1. The number of guests under consideration is 27,352
2. For guests with special requests, about 27% or ~7,300 guests cancel, which means the special requirements are not a major factor that affect cancellation for most of the guests that come to the hotel

Exploratory Data Analysis – Correlation of Variables

This is a correlation heatmap of the various numerical variables of the data



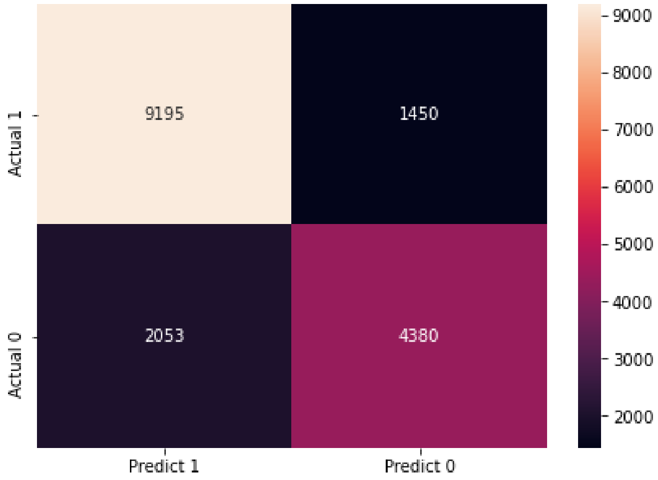
Observations:

There are no major correlations between any of the variables

Prediction Model

Logistic Regression was used to build the model due to the fact that it is able to find the relationship between dependent (booking_status) variables and independent (all other variables) variables.

<u>**Confusion Matrix**</u>



The model score is 0.793 or 79.3%

Observations:

1. True Positives (TP): we correctly predicted that they did not cancel 9,195
2. True Negatives (TN): we correctly predicted that they canceled 4,380
3. False Positives (FP): we incorrectly predicted that did not cancel (a "Type I error") 2,053 Falsely predict positive Type I error
4. False Negatives (FN): we incorrectly predicted that they cancelled (a "Type II error") 1,450 Falsely predict negative Type II error

Decision Tree

In order to build the decision tree, a number of permutations were run through so as to get the decision tree that neither overfit nor underfit. In the end, the recall was the most important factor used to determine the viability of the tree

<u>Comparison between Test and Training Data of the Various Decision Trees</u>

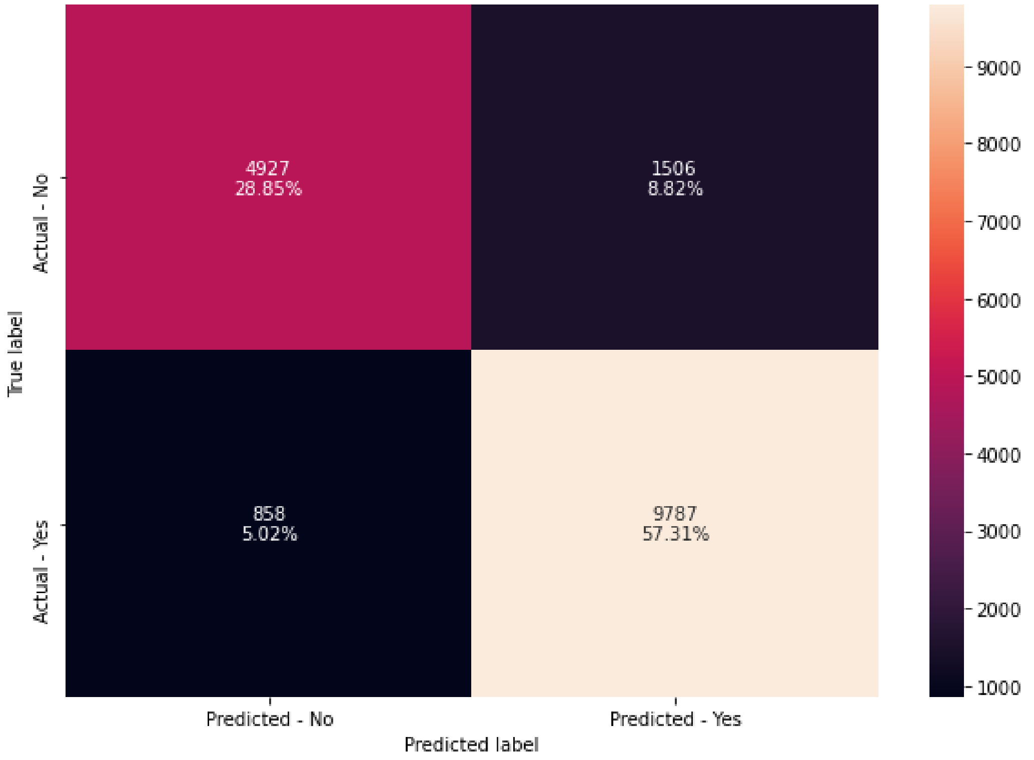
	Recall on Training Data	Recall on Test Data
Initial Decision Tree Model	0.99	0.88
Decision Tree with Reduced Maximum Depth	0.86	0.86
Decision Tree with Hyperparameter Tuning	1	1
Decision Tree with Post-Pruning	0.94	0.92

<u>Important Variables used in Model Building</u>

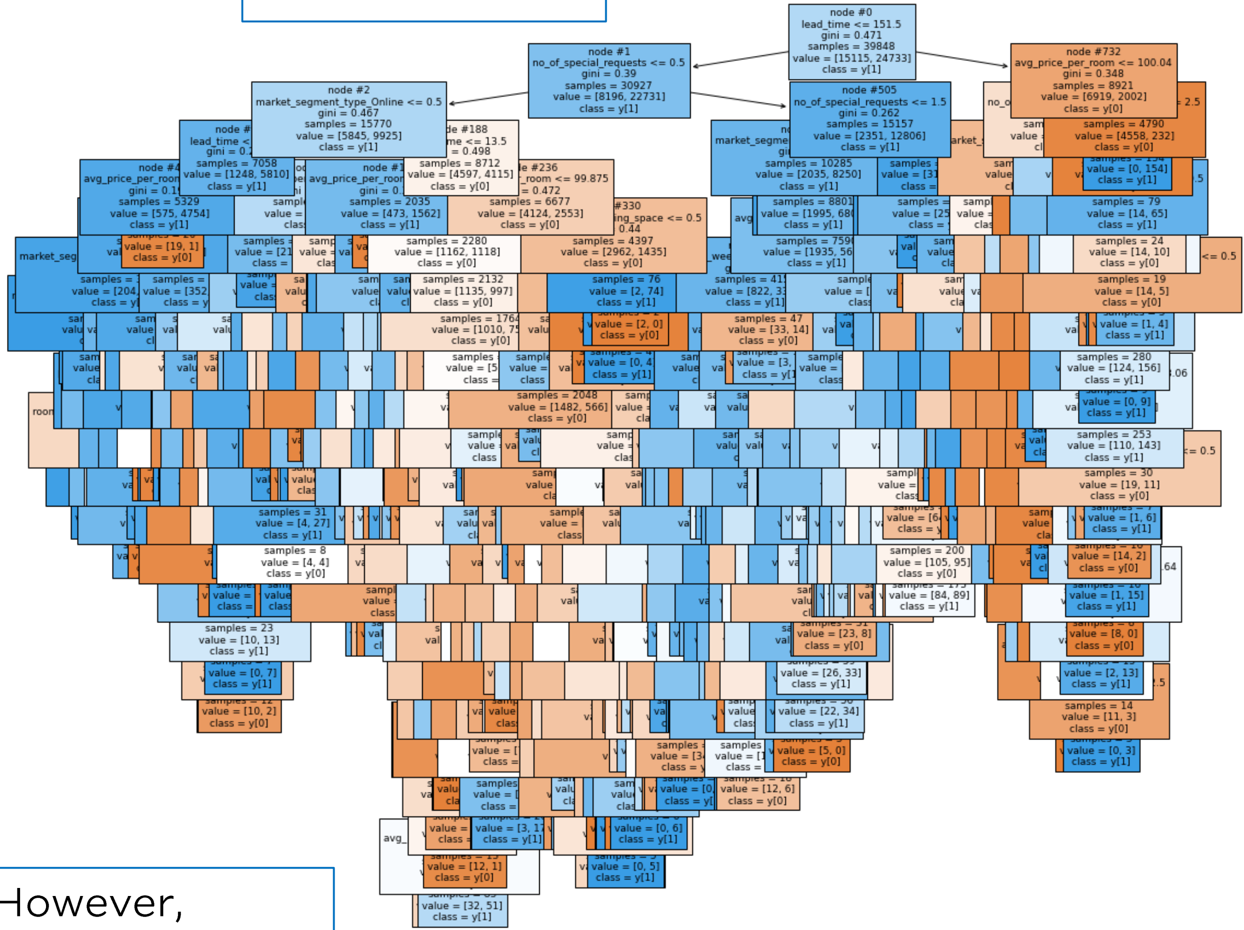
1. lead_time,
2. avg_price_per_room,
3. arrival_date,
4. no_of_special_requests
5. market_segment_type_Online

Decision Tree Output

Confusion Matrix



Decision Tree



1. The decision tree looks like it displays signs of overfitting. However, considering the number of variables involved and with comparison to the first decision tree, it is at a reasonable number of nodes
2. *lead_time* is still the most important variable for prediction
3. Recall is very high at 0.92

Conclusion

I analysed the variables, keeping *booking_status* as the dependent variable:

- After analysis of the data, and using different techniques and using a Decision Tree classifier to build a predictive model for same data
- The model built can be used to predict if a customer is going to cancel a booking or not
- We visualised different decisio trees and their confusion matrixes to get a better understanding of the model.
- We established that the most important variables were *lead_time*, *avg_price_per_room*, *arrival_date*, *no_of_special_requests* and *market_segment_type_Online*
- We established the importance of hyper-paramaters/pruning
- The decision tree model chosen – ***Decision Tree with post-pruning*** – *has the best recall score short of overfitting*

Recommendation

Based on the analysis, there are following recommendations that can help the business retain bookings:

- Focus on transforming customers to repeat customers. This can be done by
  - offering special discounts to customers after they have made a certain number of reservations and
  - if they hit a high enough number, offering complementary rooms
- During peak months (which coincide) with the Summer holidays, the hotels need to make sure that a wide variety of activities are available and that they are willing to cater to all requests
- From the catering to all requests, the hotel needs to take a tally of the most common special requests and try to address them all using the 80/20 rule (the requests that affect 80% of the customers that make special requests)
- As most customers seem to book online, the hotel needs to make sure that it's website is easily navigable and easy to use
- Also, a tiny percentage could be taken off the room prices of all the customers that actually show up. This fact would need to be heavily communicated and promoted across all media platforms