INN Hotels
Business Presentation

By
Atchara Chantharak

Contents

- Business Problem Overview
- Data Overview
- Exploratory Data Analysis (EDA)
- Data Preparation
- Model Performance Summary
- Business Insights and Recommendations

Business Problem Overview and Solution Approach

- INN Hotels Group has a chain of hotels in Portugal and they are facing challenges with a significant number of cancellations or no-shows.

- To keep customers satisfaction, the hotel make it easily for guests to cancel their bookings with no fee or small fees. However, the hotels are taking losses in various ways.
  - Loss of revenue when the hotel cannot resell the room.
  - Additional costs when the hotel pays commissions for publicity to help sell these rooms.
  - Reducing revenue when lowering pricings last minute to get those room re-booked.
  - Utilizing human resources inefficiently: over-staff or under-staff because last minute cancellations or no-shows make it difficult to estimate the efficiency amount of human resources.
  - Loss of human resources to make arrangement for guests canceling guests.

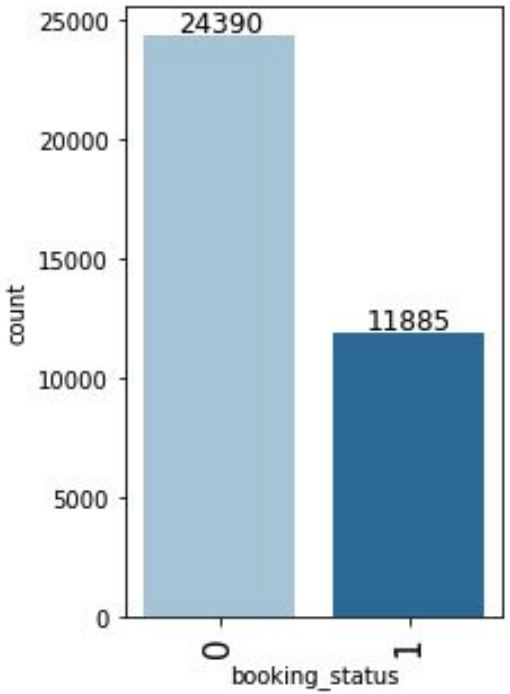
- The task is to analyze the data provided and develop a strategy to help in formulating profitable policies for cancellations and refunds using a logistic regression model and a decision tree model and identify factors that highly influence which booking is going to be canceled in advance.

Data Overview

- The data contains 36,275 bookings and 19 customers' booking details.
- Booking details include:

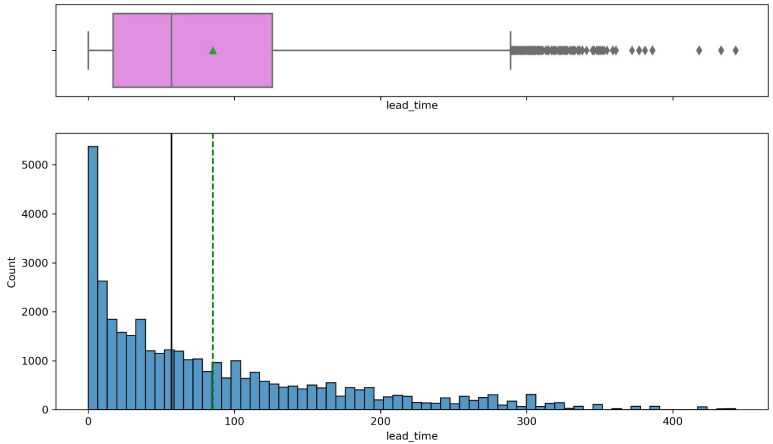
  booking status, average price per room, number of guests, no of weekday no of weekend, type of room, lead time from booking to arrival, arrival date/month/year, market segment type, repeated guests, no of special requests, and more.
- We will use booking status (canceled and not cancel) as target.
- There is missing value.

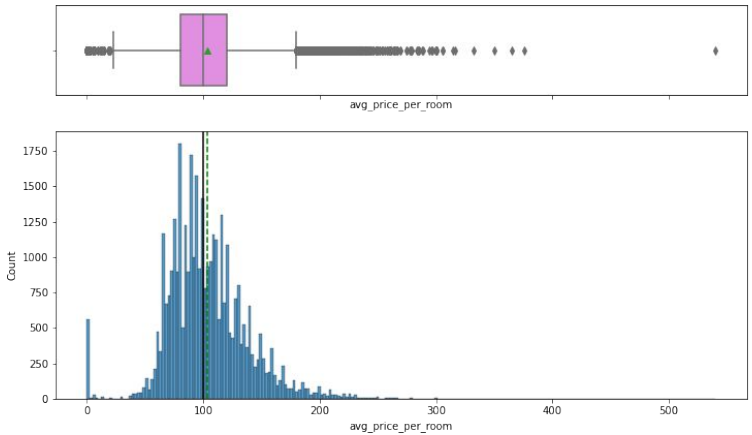
Exploratory Data Analysis (EDA)



- Booking Status: Not cancel booking as 0 and canceled as 1.
- There are 24,390 bookings (~63%) that did not cancel and 11,885 bookings (~33%) which has been canceled.
- It's a very significant number of cancellations that we will need to analyze to reduce resources and to improve revenue and profits of the hotels.

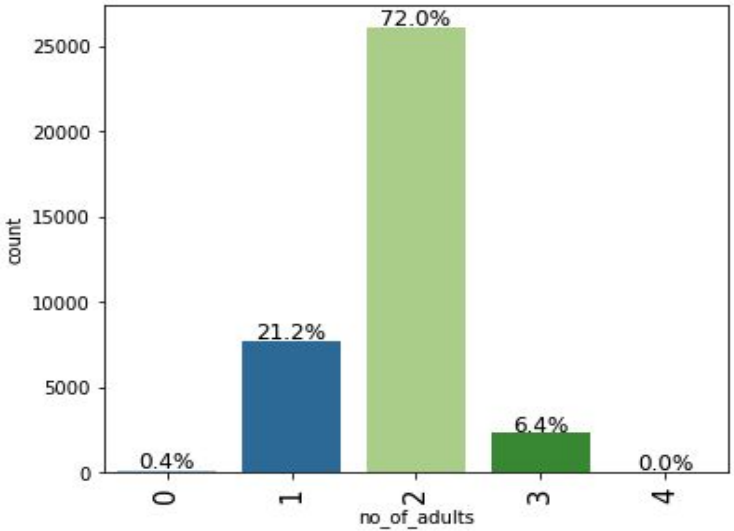
Exploratory Data Analysis (EDA)

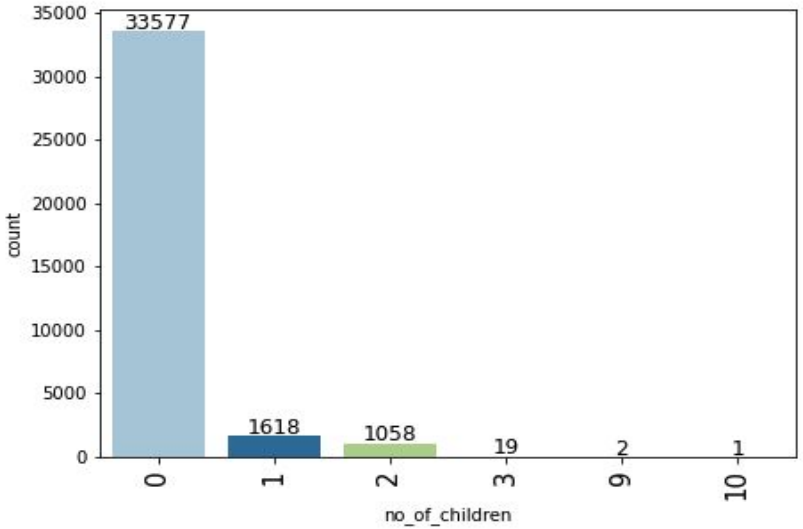




- ***Lead time***, number of day from booked to arrival date, has a heavily right-skewed with lot of outliers. More than 5,000 booking were made on the day of or a few days before an arrival date. Maximum lead time is 443 days and median is 57 days.
- ***Average price per room*** is a right-skewed distribution with lots of outliers. Mean and Median are around 100 dollars. There are more than 500 booking counts that 0 dollars. Its upper whisker is 179.55 dollars.

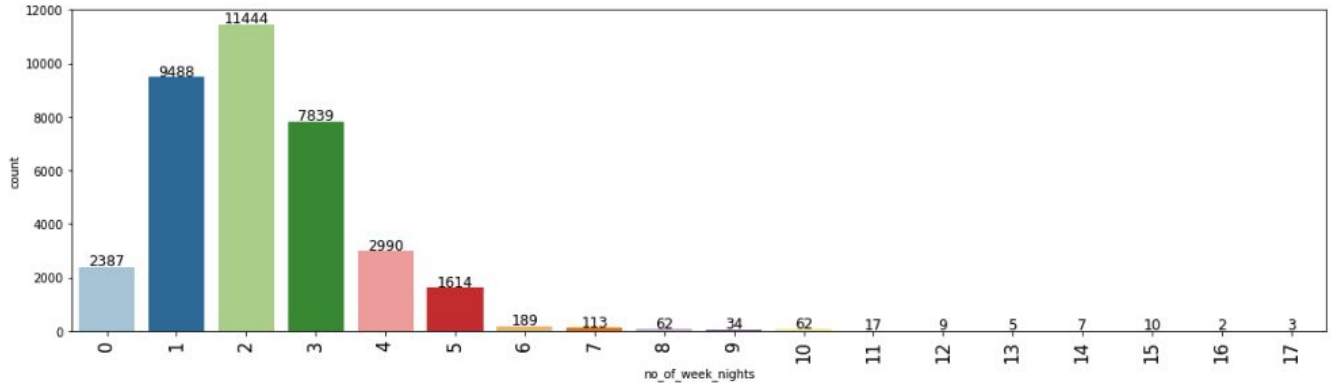
Exploratory Data Analysis (EDA)

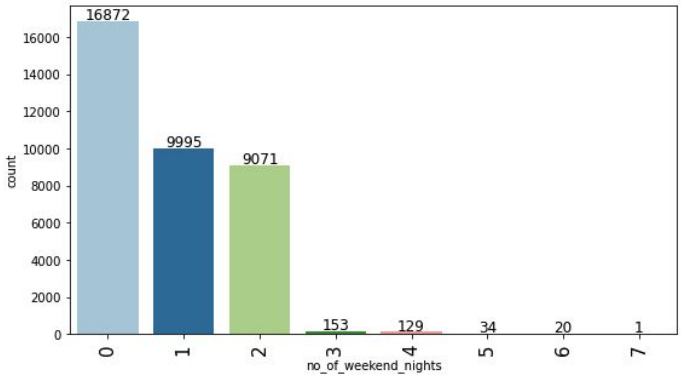




- 72% of Bookings was made for 2 adults and 21% was for 1 adults.
- Very small amount of bookings included children

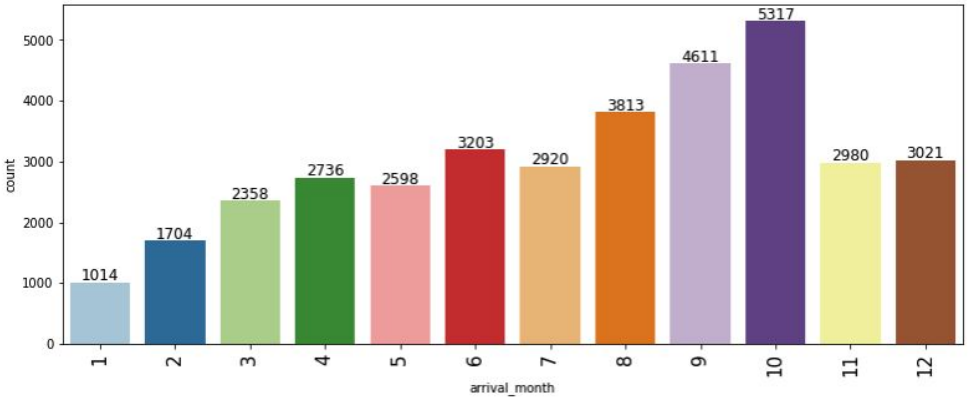
Exploratory Data Analysis (EDA)_3

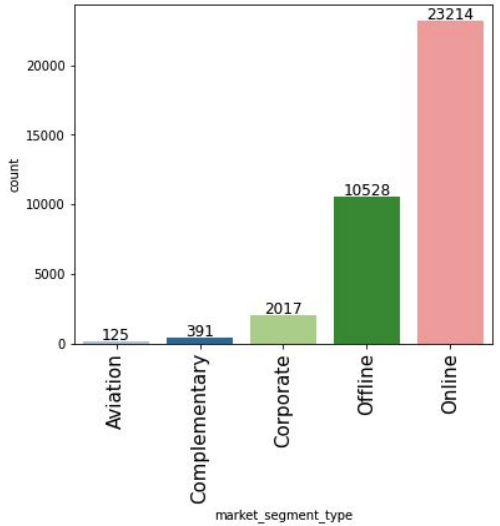




- ***No of weekday nights*** has a range from 0 to 17 days with 2 nights as the most frequent bookings, following by 1 days. The bookings that have 0 weekday night, assuming that they are leisure travels.
- ***No of weekend nights*** has a range from 0 to 7 days. The most frequent bookings is 0 nights, assuming customers booked rooms for business. Following by 1 nights and 2 nights.
- A customers will most likely book a room for business.

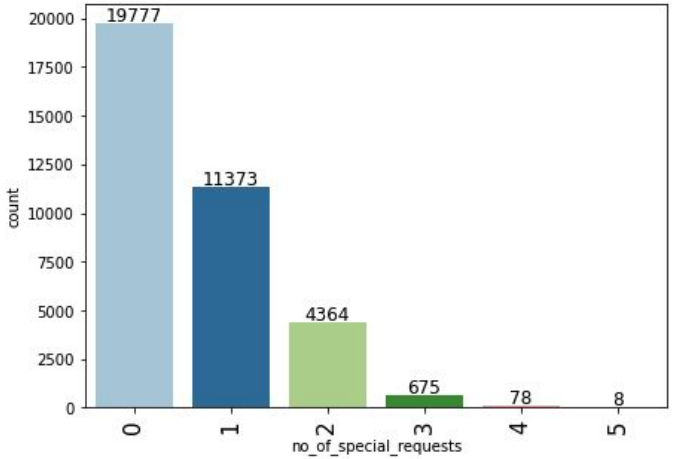
Exploratory Data Analysis (EDA)

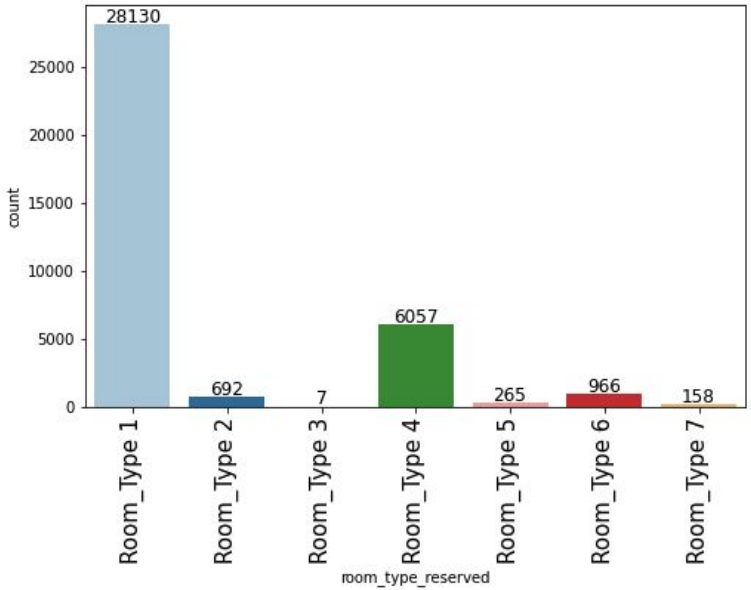




- The top 3 ***arrival months*** are October at 5,317 bookings, September at 4,611 bookings, and August 3,813 bookings. Holidays season in November and December have similar booking counts
- There are 5 type of ***market segments***. Customers made room reservations via online, most convenience. Following by offline, corporate, complementary and aviation.

Exploratory Data Analysis (EDA)



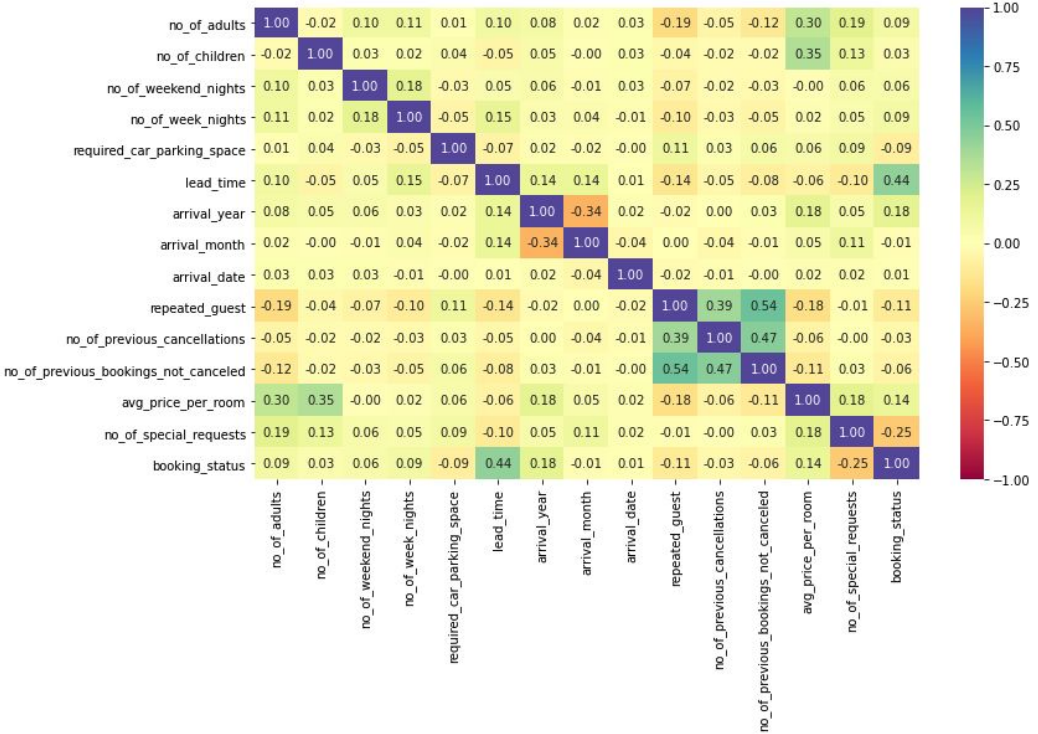


- ***No of special requests*** has a range from 0-5 requests. Most bookings, 19,777 bookings, have no special request, following by 1 request, 2 requests, and the rest.
- ***Room type reserved*** has 7 types of rooms. Most of customers choose room type 1 as 28,130 bookings.

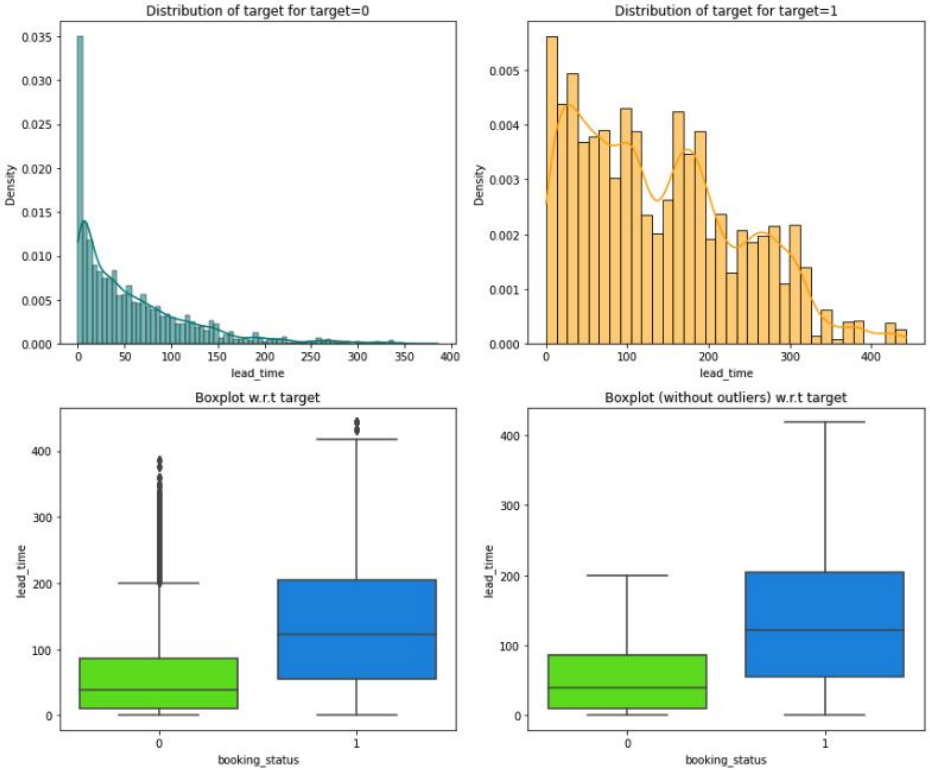
Exploratory Data Analysis (EDA)

Correlation

- ***Lead_time*** has the highest correlation with booking_status at 0.44. Following by no_of_special_request at -0.25.
- ***repeated_guest*** has 0.54 correlation with no_of_previous_bookings_not_canceled.
- ***ave_price_per_room*** has 0.30 correlation with no_of_adult and 0.35 correlation with number of children.
- ***booking_status*** has 0.14 correlation with ave_price_per_room and -0.11 with repeated-guest
- We will investigate more of the relationship between lead_time and booking_status and more.

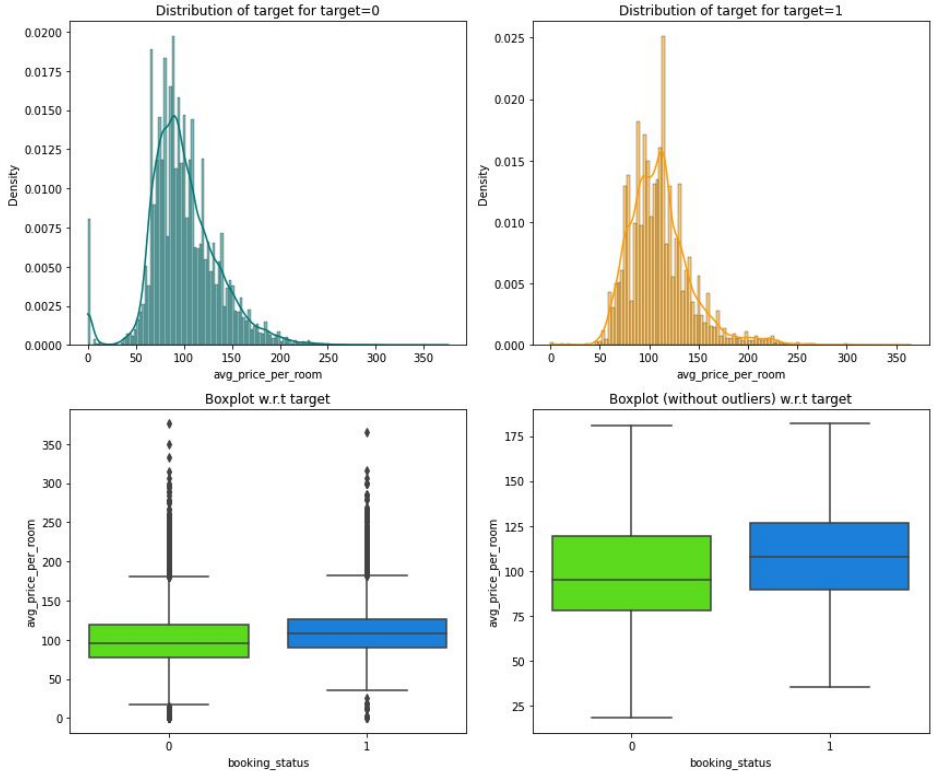


Exploratory Data Analysis (EDA)



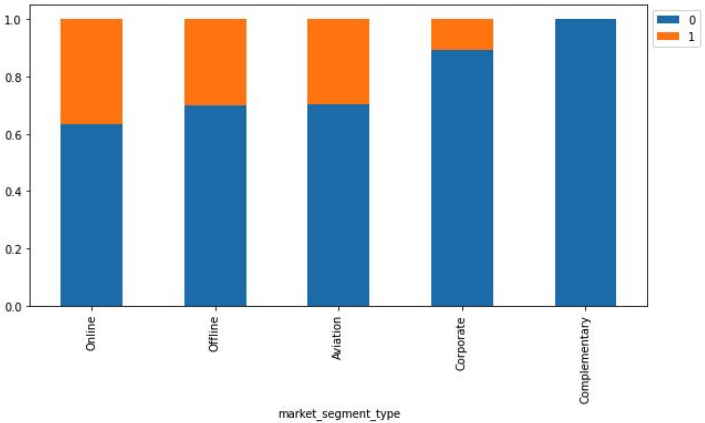
- The median of ***lead_time*** of not cancel booking is around 50 days.
- The median of lead_time of canceled booking is around 120 days.
- Without outliers, the range for lead_time of not cancel booking is 0-200 days and for lead_time of canceled booking is 0-400 days.
- From this observation, the longer the lead_time of booking, the higher chance for cancellations.

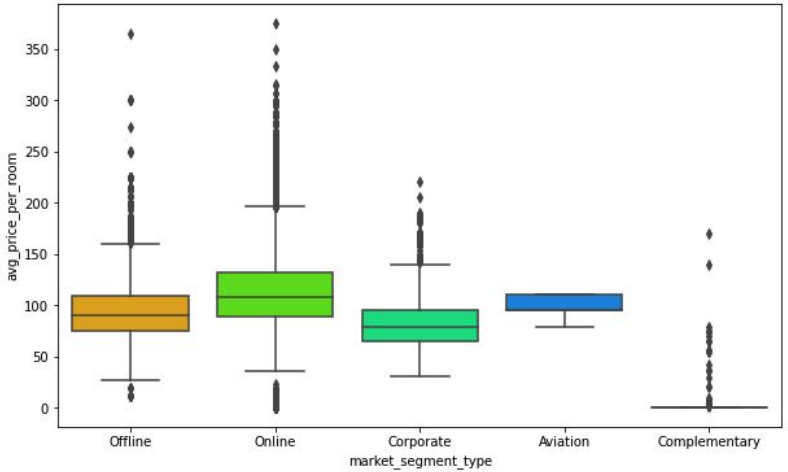
Exploratory Data Analysis (EDA)



- With and without outliers, the ***average price per room*** for canceled booking is ~110 dollars and for not canceled bookings is ~95 dollars.
- Customers who canceled their booking may find a more affordable room from different hotels.
- Customers who did not cancel their booking are satisfied with their room prices and see it as good prices.

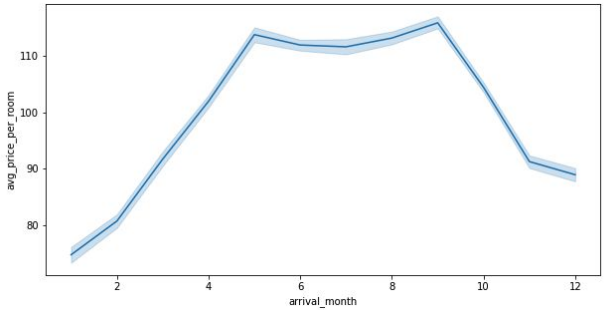
Exploratory Data Analysis (EDA)

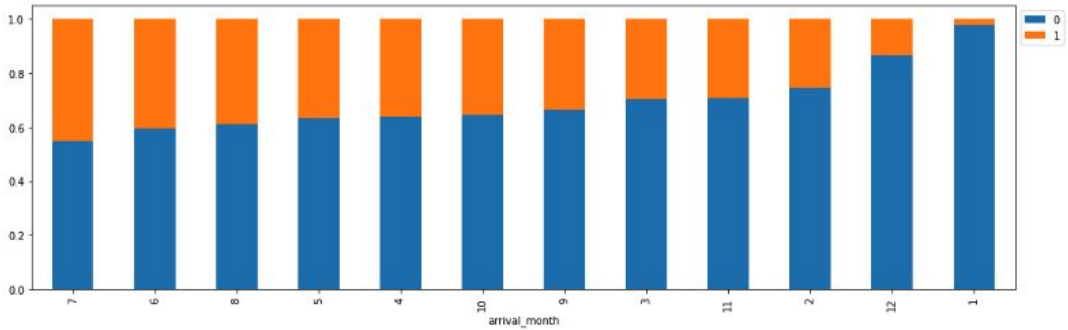




- Observation on ***market segment type***, Online booking has the highest cancellations (~40%), following by Offline, Aviation, Corporate and Complementary (~0%). Aviation cancellations might due to flight delays or flight canceled.
- ***Average price per room*** for online bookings has the highest price (over 100 dollars), following aviation bookings, offline bookings and complementary (data shown that a large number of bookings is free).

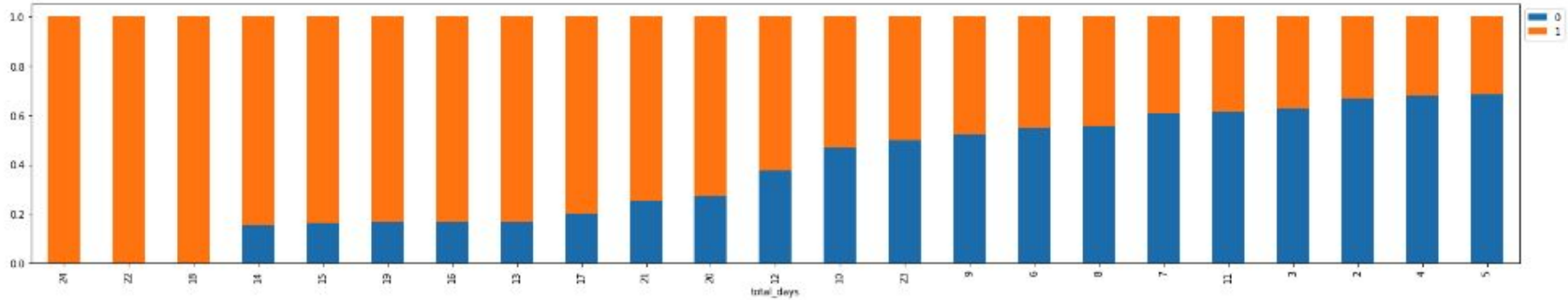
Exploratory Data Analysis (EDA)





- The most expensive months are September, May, August, June and July . The average price in these months are around 110 dollars or higher. Correlating to the most cancellations months.
- The least expensive months are January, February, March, December, and November. Those prices are around 90 dollars or less. Correlating to the less cancellations months.

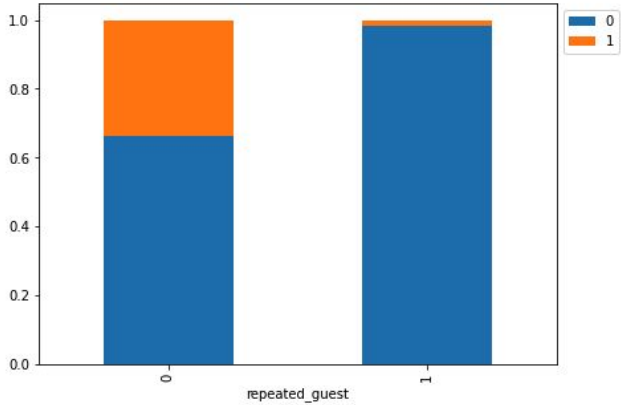
Exploratory Data Analysis (EDA)

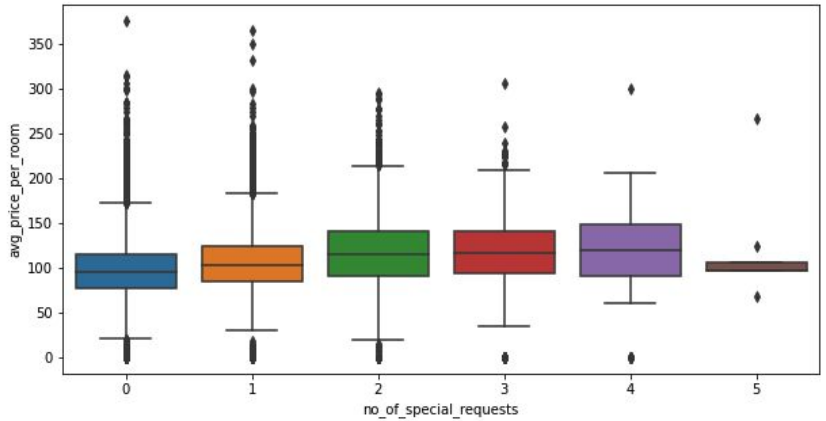


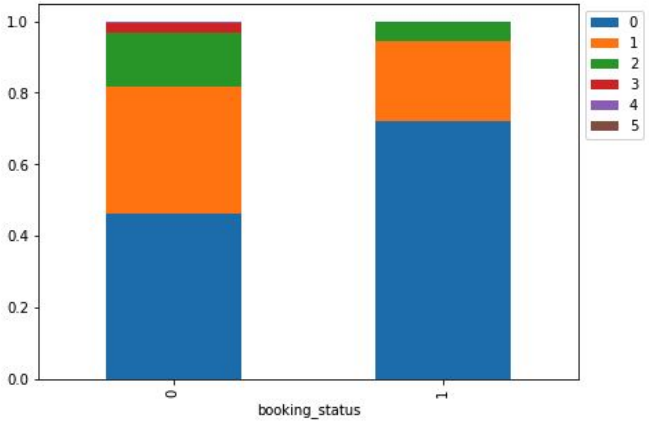
- There are 17094 bookings that stay longer than 2 days.
- The range is from 2 days to 24 days.
- The longer days of stay, the high percentage of cancellations.
- Bookings for 22-24 days of stay have almost 100 percent chance that the bookings will be canceled.
- Bookings for 2-5 days of stay have the least chance of those bookings to be canceled, ~30%.

Exploratory Data Analysis (EDA)

- Bookings of ***repeated customers*** is almost 100% that bookings did not cancel and for not repeated customers has ~65% cancellations. For
- Top 3 of median price of Bookings with ***No of special requests*** is 2, 3, and 4, which is around 120 dollars. For 0 request bookings price is around 90 dollars. All of them have outliers.







Data Preparation

Before we proceed to build a model, we will:

- Drop Booking_ID column. We decided to group booking ID because they are unique numbers.

  We cannot use it for pattern recognitions.
- Treat outliers
  - Avg_price_per_room: There are only outliers above upper whisker.
    - Calculated upper whisker which is 179.55 dollars
    - Assigned 179.55 dollars to outliers greater or equal to 500 dollars.
  - No_of_children: We used 3 children to replacing bookings with 9 or 10 children

- Encode categorical features: Type_of_meal_plan, Room_type_reserved,

  Market_segment_type, and Booking_status
- Split the data into train (70%) and test (30%) to evaluate the model that we build on the train data.

Model Performance Summary

- We want to predict which bookings will be canceled.
- Model can make wrong prediction as ***false negative*** (predicting a booking to be canceled when it does not) and ***false positive*** (predicting a booking to not cancel when it does).
- We decided that both false negative and false positive are important.
    - ***False negative***: the hotel might not be able to provide satisfactory services to the customer by assuming that this booking will be canceled. This might damage the brand equity.
    - ***False positive***: the hotel will lose resources and will have to bear additional costs of distribution channels try to resell the room.
- We want to reduce the losses by want ***F1 Score*** to be maximized for higher the chances of minimizing False Negatives and False Positives.
- We will use ***Logistic Regression Model*** and ***Decision Tree Model*** for prediction.

Model Performance Summary

- The model that provides the highest F1 is Decision Tree (Post-Pruning) with great numbers for Accuracy, Recall, and Precision as well.

Accuracy	Recall	Precision	F1
0.86888	0.85576	0.76634	0.80858

- Top 5 features (with relative important values) to predict which bookings will be canceled
  - Lead time (~0.4)
  - Market segment type online (~0.15)
  - Average price per room (~0.14)
  - No of special requests (~0.13)
  - Arrival month (~0.7)
- The rest of the important features are arrival date, no of weekday, no of weekend nights, and no of adult

Business Insights and Recommendations

- Lead time, market segment type online, average price per room, no of special requests, and arrival month (in order) are the most important variables in determining which bookings will be canceled.
- A booking with a lead time 120 days or more is mostly likely to be canceled
- Booking a room through online market segment is mostly likely to be cancel. A booking with corporation and complimentary are less likely to be canceled
- A booking with 0 or 1 special request has a higher chance to not cancel the booking. The more special requests, the higher chance for a customer to cancel the booking.
- A booking is June, July, and August has a higher chance to cancel the booking. A booking in January, February, March, November, and December are less likely to be canceled.
- A booking with 2 adults has a small chance for cancellation. A booking with 3 or more guests is more likely to cancel.

Business Insights and Recommendations

- Improving human resources efficiency and maximizing revenue by creating a matrix factoring lead time, lead time, market segment, no special requests, arrival months, and no of adults to predict a percentage of bookings that might be canceled. Weekly the matrix will be updated and reviewed by managers before scheduling staff.
- Reducing number of cancellations by
    - No RSVP 1 year or more in advance policy and making a revision on the hotel online reservation platform.
    - A cancellation fee for any cancellations less than 24 hours and a full price for no-shows. Need to investigate more for the right amount of cancellation fee.
    - Comparison between a price with no-cancellation option (higher price) and a price with a cancellation option (lower price).
    - Promotion of booking a room for Aug-Oct (highest number of bookings and highest cancellations) and get a one night stay complimentary to be redeemed in the next 6 months for free.

Business Insights and Recommendations

- Increasing revenue and brand equity by creating a membership program to improve number of repeat customers. The membership benefits such as a priority check-in, a free breakfast, a free parking, an access to a gym, a complimentary gift card to purchase items in the hotel gift shops or stores, etc.
- Reducing cost and other resources by removing Meal Plan 3 which is the least popular with only 5 orders.
- Increasing available room inventory by converting the unpopular floor plans such as room type 3, 5, and 7 to the most popular plan, room type 1. If the hotels have a renovation budget and if the building floor plans are permitted.

greatlearning
Power Ahead

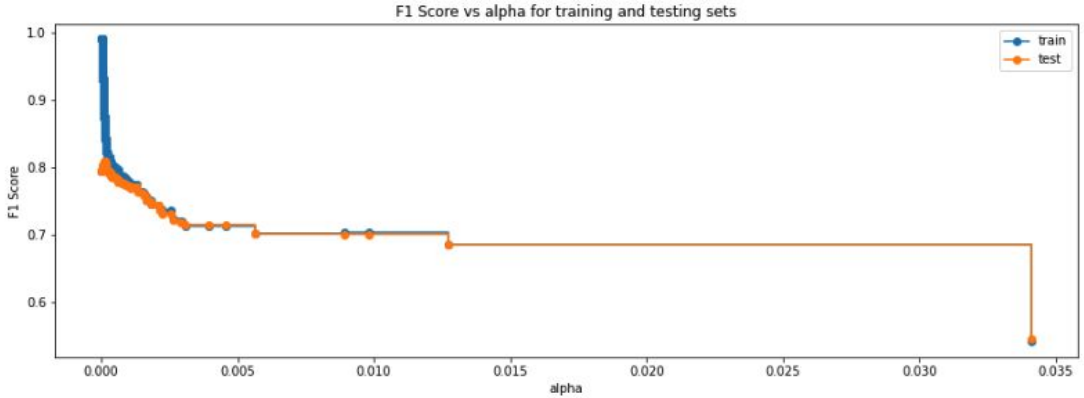
Happy Learning !

Appendix: Model Performance Comparison

	Logistic Regression Models			Decision Tree Models		
	default Threshold	0.37 Threshold	0.42 Threshold	Decision Tree sklearn	Pre-Pruning	Post-Pruning
Accuracy	0.80465	0.79555	0.80345	0.87118	0.83497	0.86888
Recall	0.63089	0.73964	0.70358	0.81175	0.78336	0.85576
Precision	0.72900	0.66573	0.69353	0.79461	0.72758	0.76634
F1	0.67641	0.70074	0.69852	0.80309	0.75444	0.80858

Decision Tree (Post-Pruning) has the highest F1 Score and Recall Score with the second highest Precision and Accuracy.

Appendix: Decision Tree (Post-Pruning)

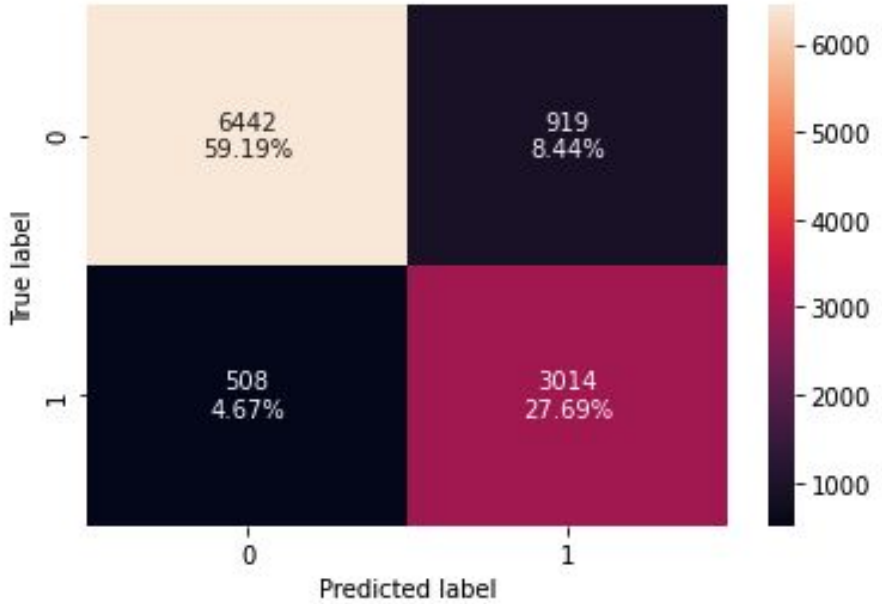


- **The best model**: alpha =0.00012267633155167 06, class_weight='balanced', and random_state = 1
- **F1 Score**: the Training set and the Test set are very close. This model is a great model for that this data set.
- The best **alpha** is very small that indicates a very complex model.

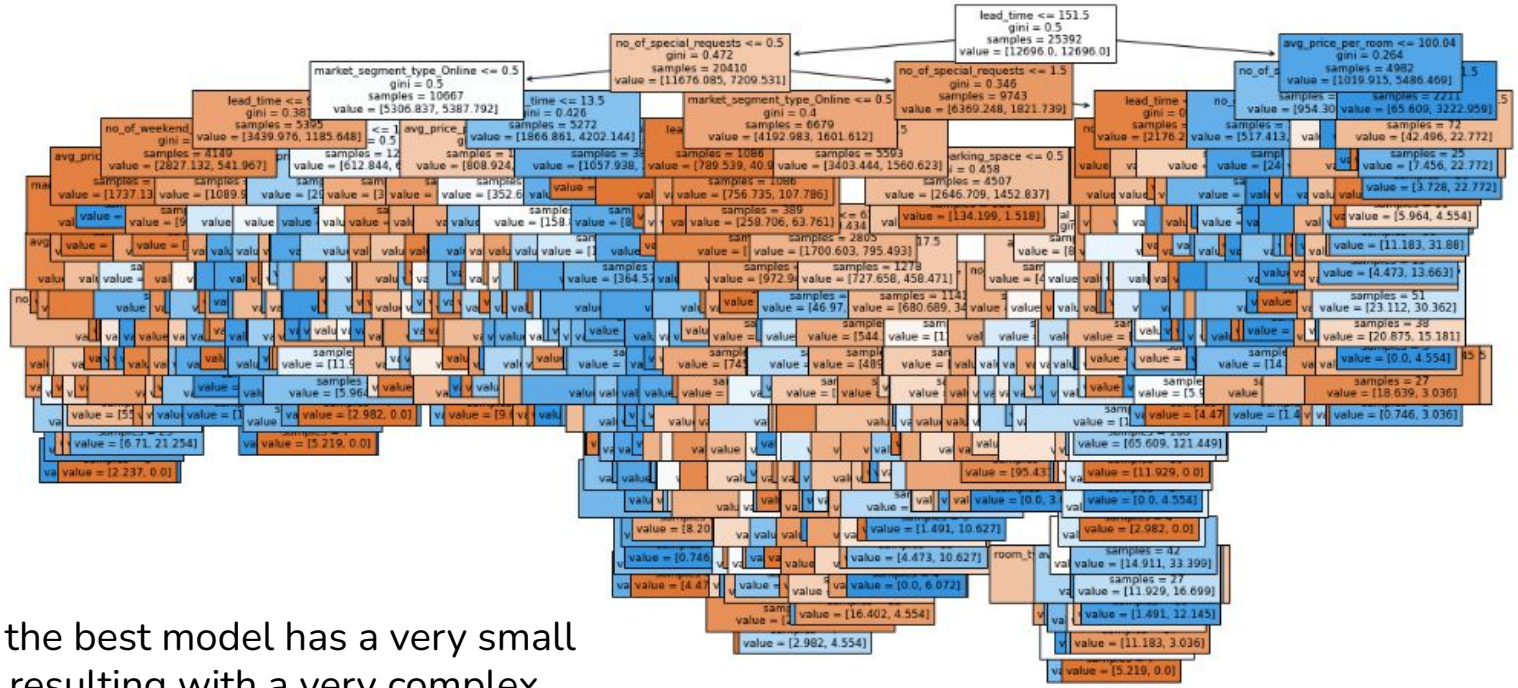
Appendix: Decision Tree (Post-Pruning)

Confusion Matrix

- We want F1 Score to be maximized, minimizing False Negatives and False Positives.
- False Negatives: 508 bookings (4.67%)
- False Positives: 919 bookings (8.44%)



Appendix: Decision Tree (Post-Pruning)



Since the best model has a very small alpha resulting with a very complex tree.