INN Hotels Project

06/01/2022

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Appendix

Executive Summary

- **Insights & recommendations**
  - **The most important feature affect the cancelation in order: [lead_time, market_segment_type_Online, no_of_special_requests, avg_price_per_room, no_of_adults, no_of_weekend_nights, arrival_month, required_car_parking_space, market_segment_type_Offline, no_of_week_nights, type_of_meal_plan_Not Selected, arrival_date.**
  - **Guests check all the previous features have high risk of cancelation therefor we need to come up with a plan to secure the reservations. E. g. 1. Online reservations without any request and with lead time more than 100 days in summer need to have fee for cancelation which increases with decreasing the days left to the check in date. 2. Prepaid with discount, this policy will make the cancelation much harder and increase the credibility and the profit. 3. booking for more than 7 day need to have extra fee for cancelation could be 50% of the total cost. Staying week-nights >5 need higher fee for cancelation than <5.**

Executive Summary

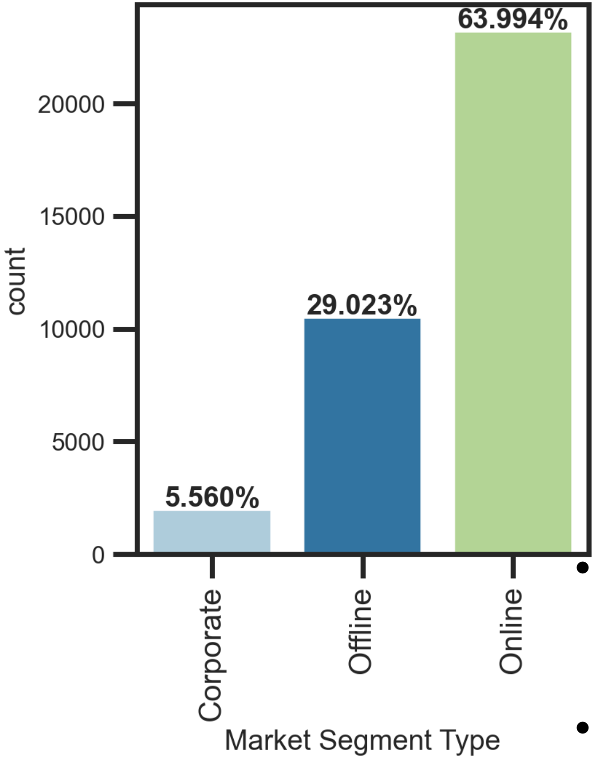
- **Insights & recommendations**
  - **Regular guests who make the order offline and asked for parking and other requests most likely are trusted guests. No need to over secure their booking, I recommend give them extra credit to your loyal guests.**
  - **Future work. We will be able to apply ensemble model such as xgboost which will improve our recent result. More data will help in improving the results as well. This dataset for INN Hotels Group has a chain of hotels in Portugal**

Business Problem Overview and Solution Approach

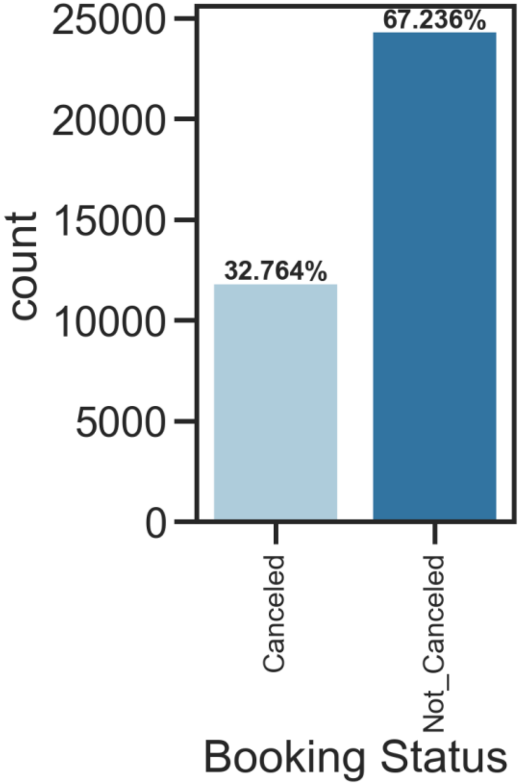
- **<u>The problem:</u>**
- **Cancelation of hotel booking. This problem has big impact on company.**
- **e.g. Loss of resources (revenue); Additional costs of distribution channels; Lowering prices last minute, more human resources to make arrangements .**
- **<u>The solution:</u>**
- **Building classification machine learning model that predict which booking is likely to be canceled in advance and help in formulating profitable policies for cancellations and refunds. This classification models are Decision Tree and Logistic Regression. The ML model predict the most important features impact the cancelation.**

EDA Results

Univariate Analysis

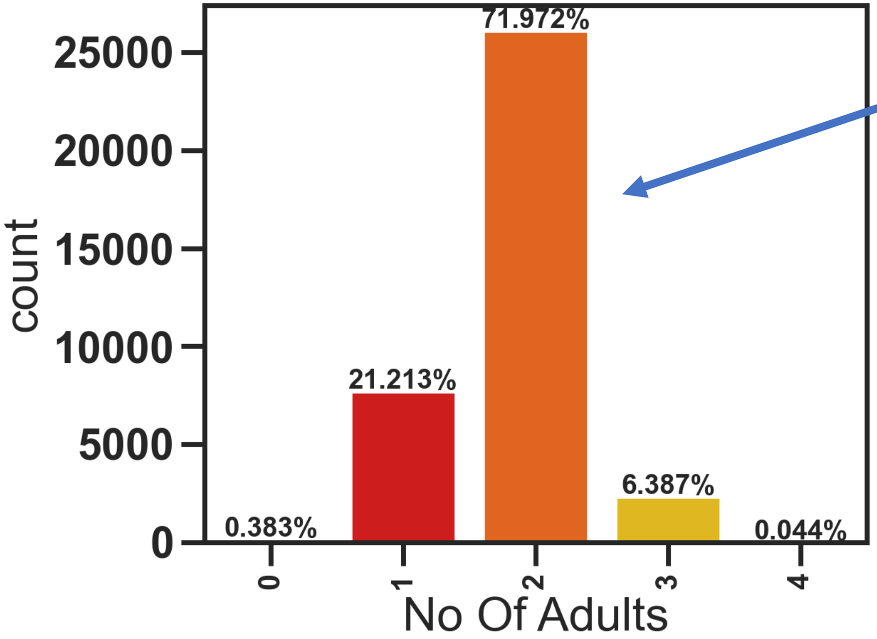


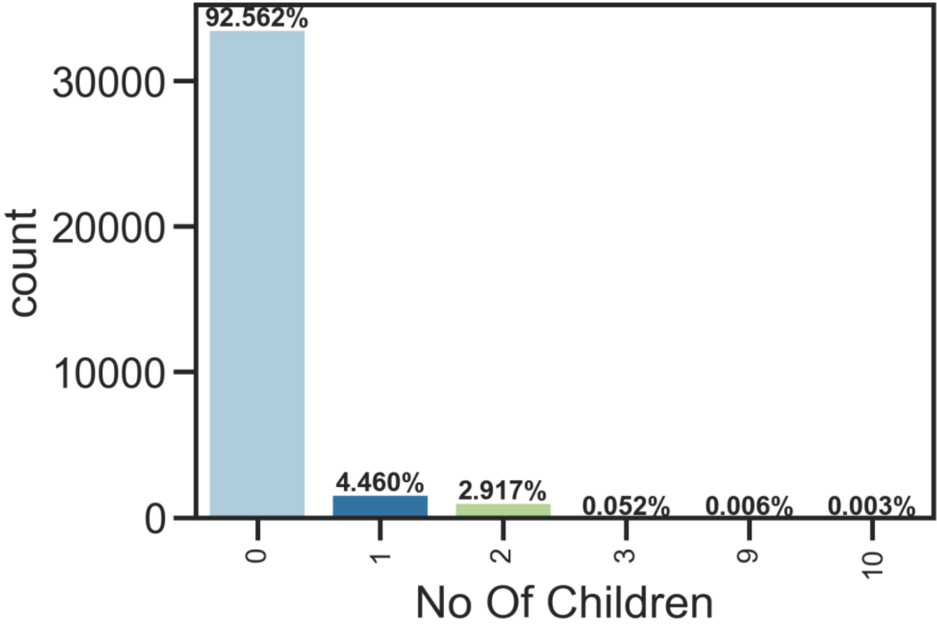
- **Online is the most segment the guests come from.**
- **32% of booking order was canceled.**



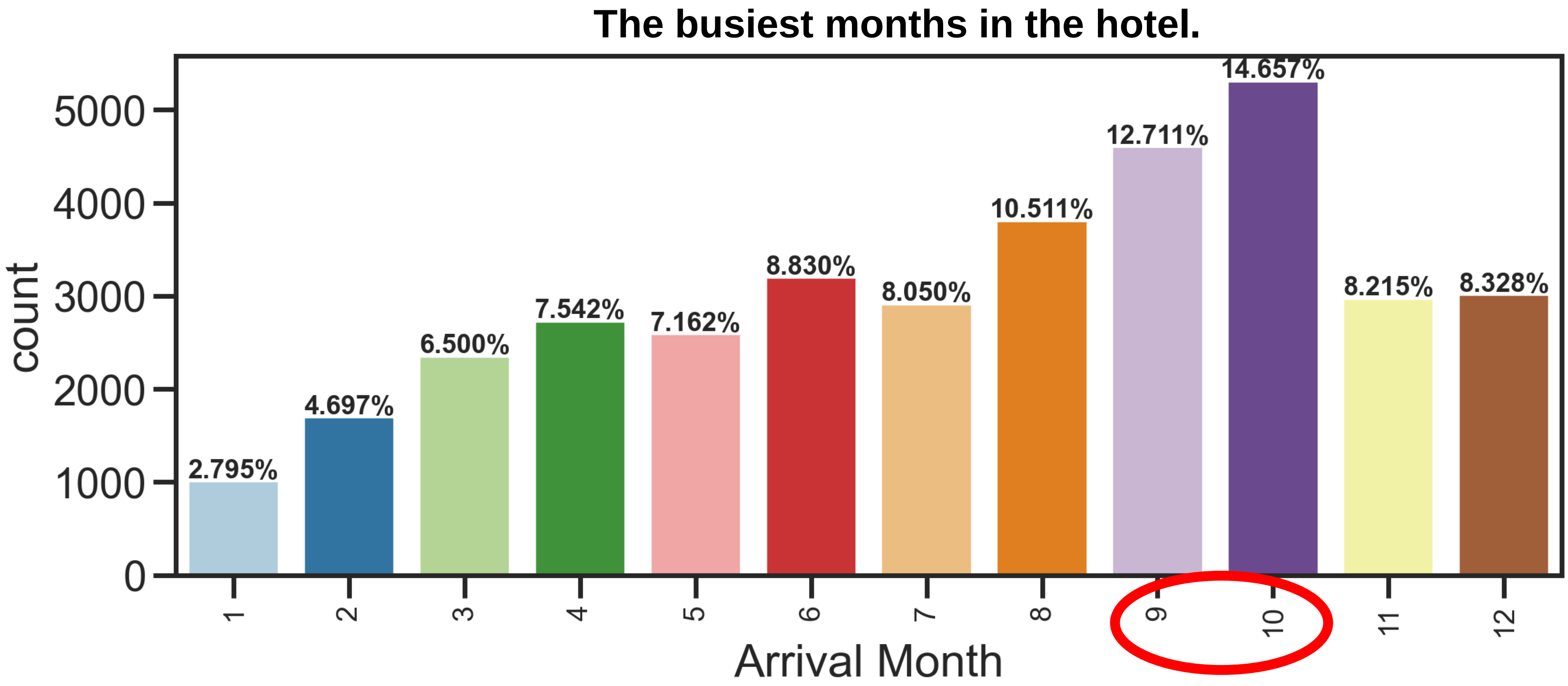
Link to Appendix slide on data background check

EDA Results


count
25000
20000
15000
10000
5000
0
0.383%
21.213%
71.972%
6.387%
0.044%
0
1
2
3
4
No Of Adults

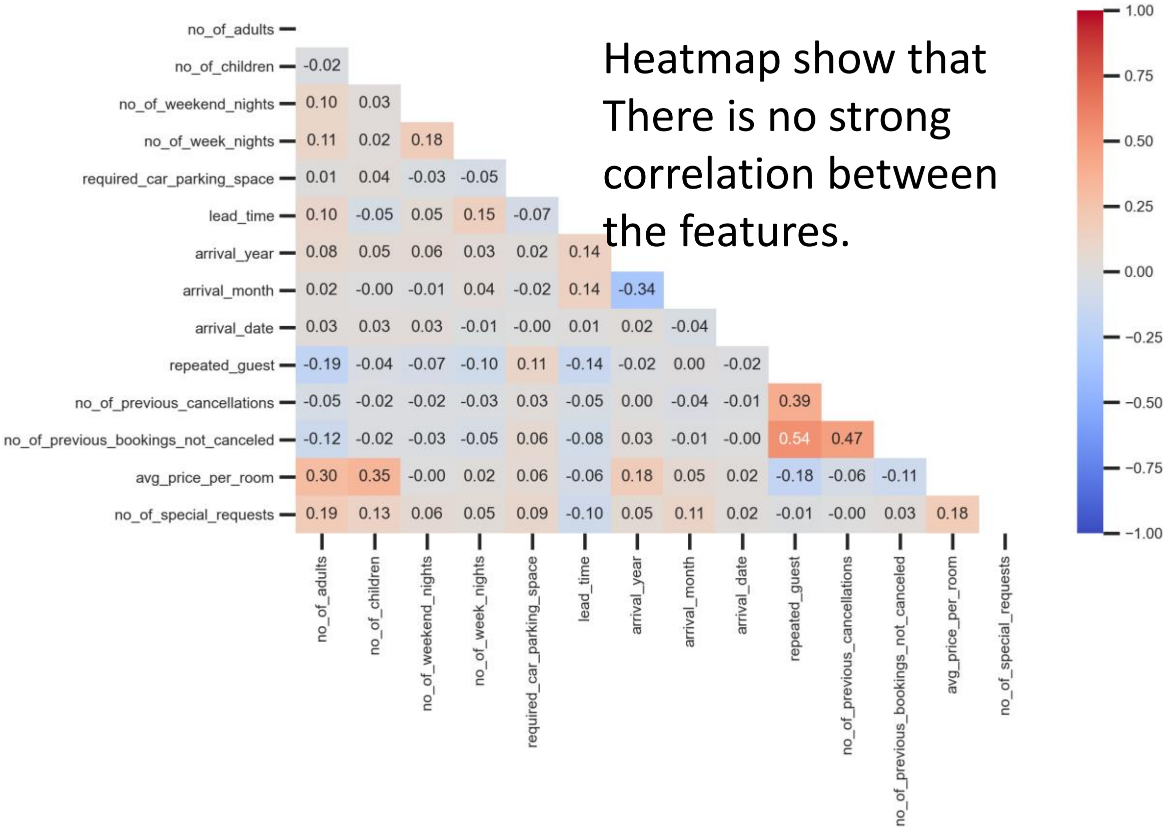

count
30000
20000
10000
0
92.562%
4.460%
2.917%
0.052%
0.006%
0.003%
0
1
2
3
9
10
No Of Children

EDA Results


The busiest months in the hotel.
count
5000
4000
3000
2000
1000
0
2.795%
4.697%
6.500%
7.542%
7.162%
8.830%
8.050%
10.511%
12.711%
14.657%
8.215%
8.328%
1
2
3
4
5
6
7
8
9
10
11
12
Arrival Month

EDA Results

Bivariate Analysis



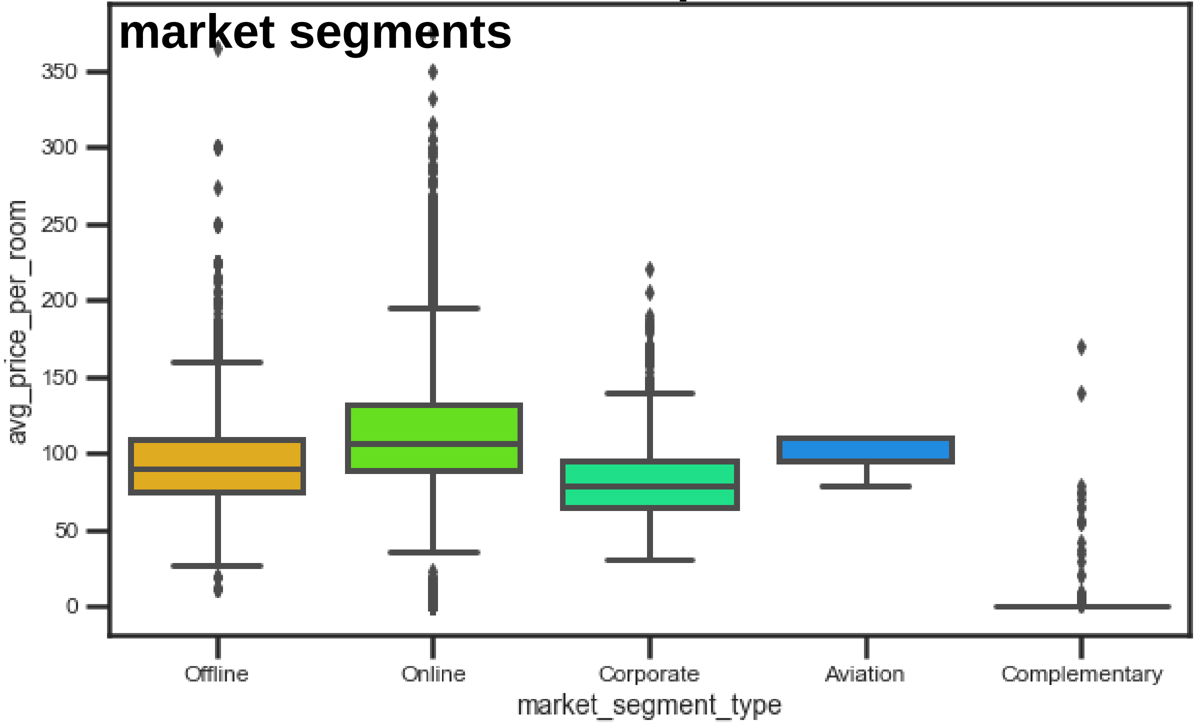
	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	required_car_parking_space	lead_time	arrival_year	arrival_month	arrival_date	repeated_guest	no_of_previous_cancellations	no_of_previous_bookings_not_canceled	avg_price_per_room	no_of_special_requests
no_of_adults														
no_of_children	-0.02													
no_of_weekend_nights	0.10	0.03												
no_of_week_nights	0.11	0.02	0.18											
required_car_parking_space	0.01	0.04	-0.03	-0.05										
lead_time	0.10	-0.05	0.05	0.15	-0.07									
arrival_year	0.08	0.05	0.06	0.03	0.02	0.14								
arrival_month	0.02	-0.00	-0.01	0.04	-0.02	0.14	-0.34							
arrival_date	0.03	0.03	0.03	-0.01	-0.00	0.01	0.02	-0.04						
repeated_guest	-0.19	-0.04	-0.07	-0.10	0.11	-0.14	-0.02	0.00	-0.02					
no_of_previous_cancellations	-0.05	-0.02	-0.02	-0.03	0.03	-0.05	0.00	-0.04	-0.01	0.39				
no_of_previous_bookings_not_canceled	-0.12	-0.02	-0.03	-0.05	0.06	-0.08	0.03	-0.01	-0.00	0.54	0.47			
avg_price_per_room	0.30	0.35	-0.00	0.02	0.06	-0.06	0.18	0.05	0.02	-0.18	-0.06	-0.11		
no_of_special_requests	0.19	0.13	0.06	0.05	0.09	-0.10	0.05	0.11	0.02	-0.01	-0.00	0.03	0.18	

Heatmap show that There is no strong correlation between the features.

ık to Appendix slide on data background check

EDA Results

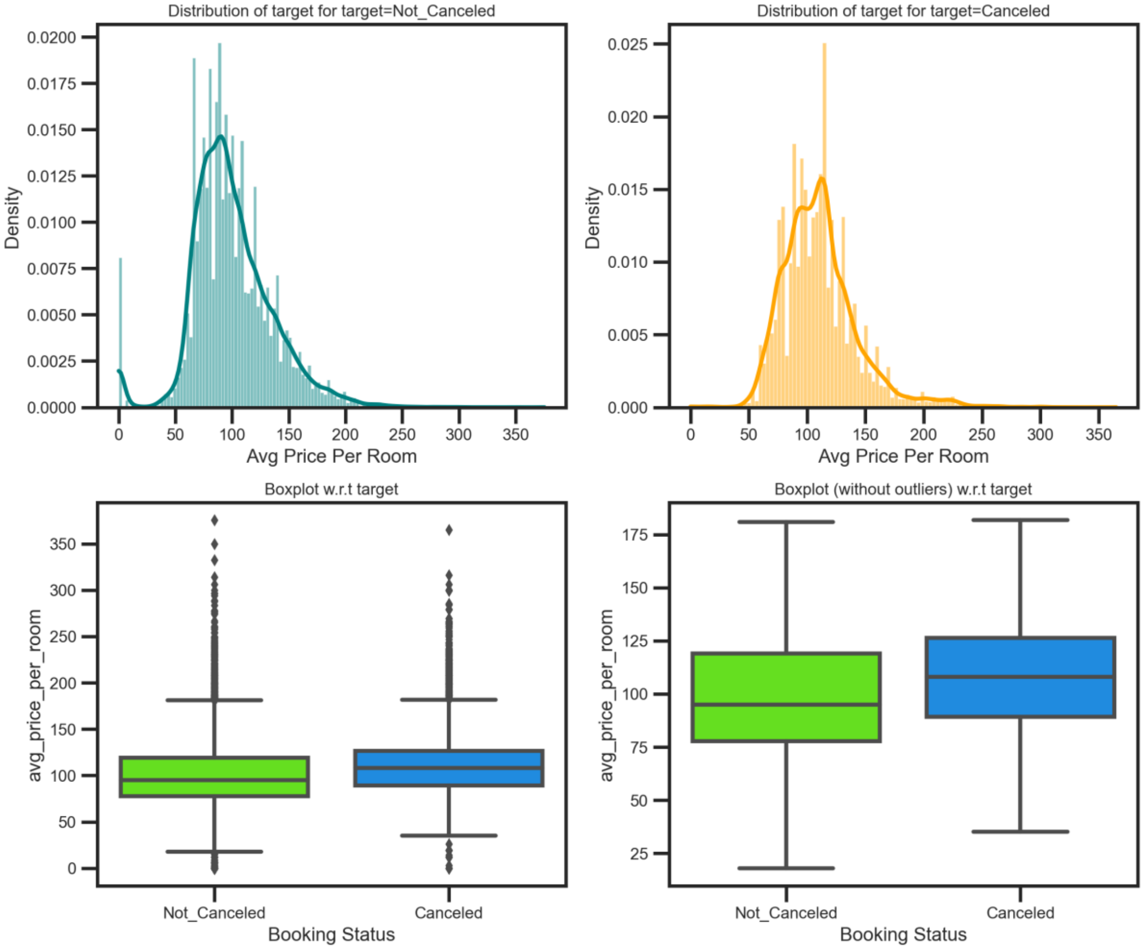
The differences in room prices in different market segments



- Complementary is the less room price. The online segment has highest median.

EDA Results

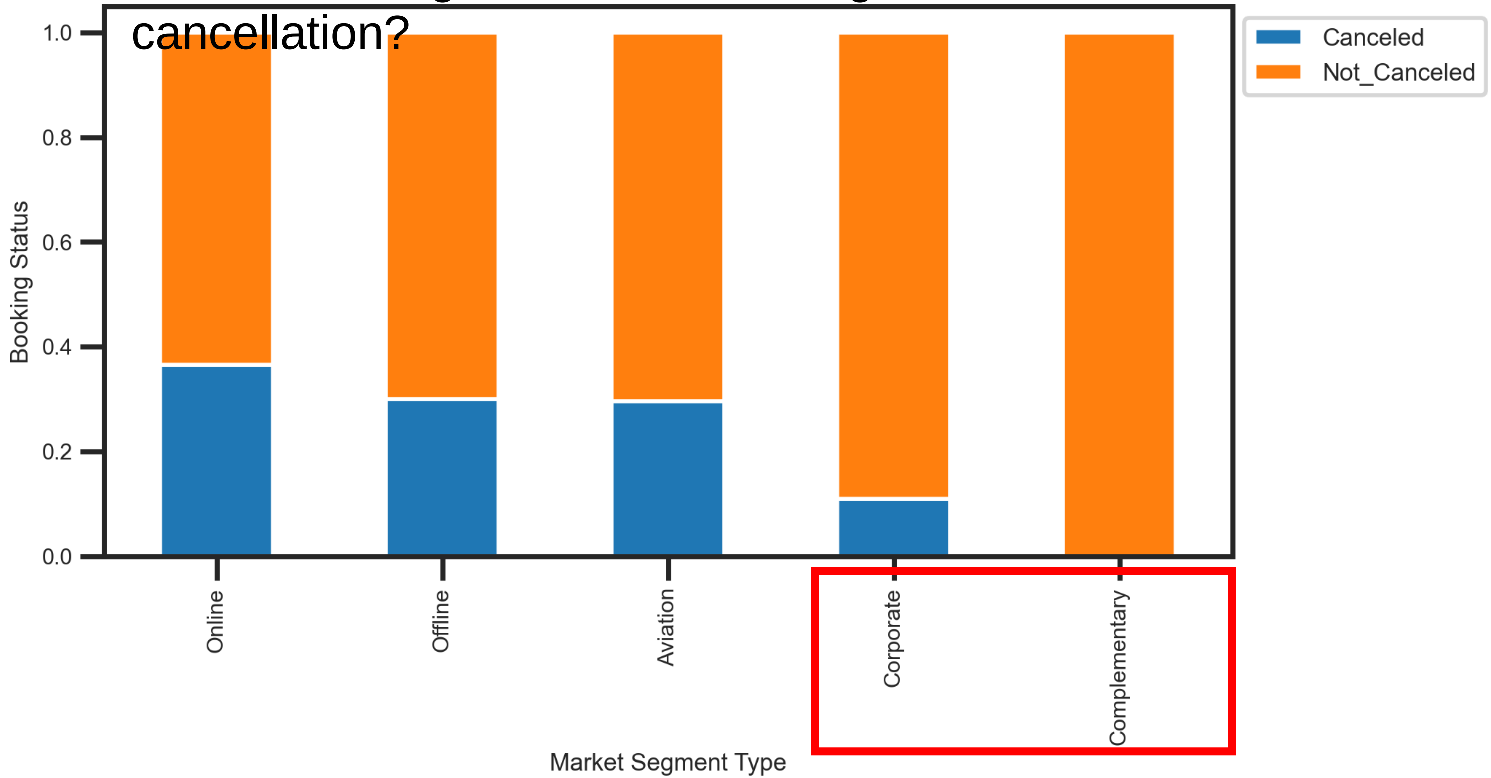
- Rooms with higher median were canceled more than rooms with lower median.



Link to Appendix slide on data background check

EDA Results

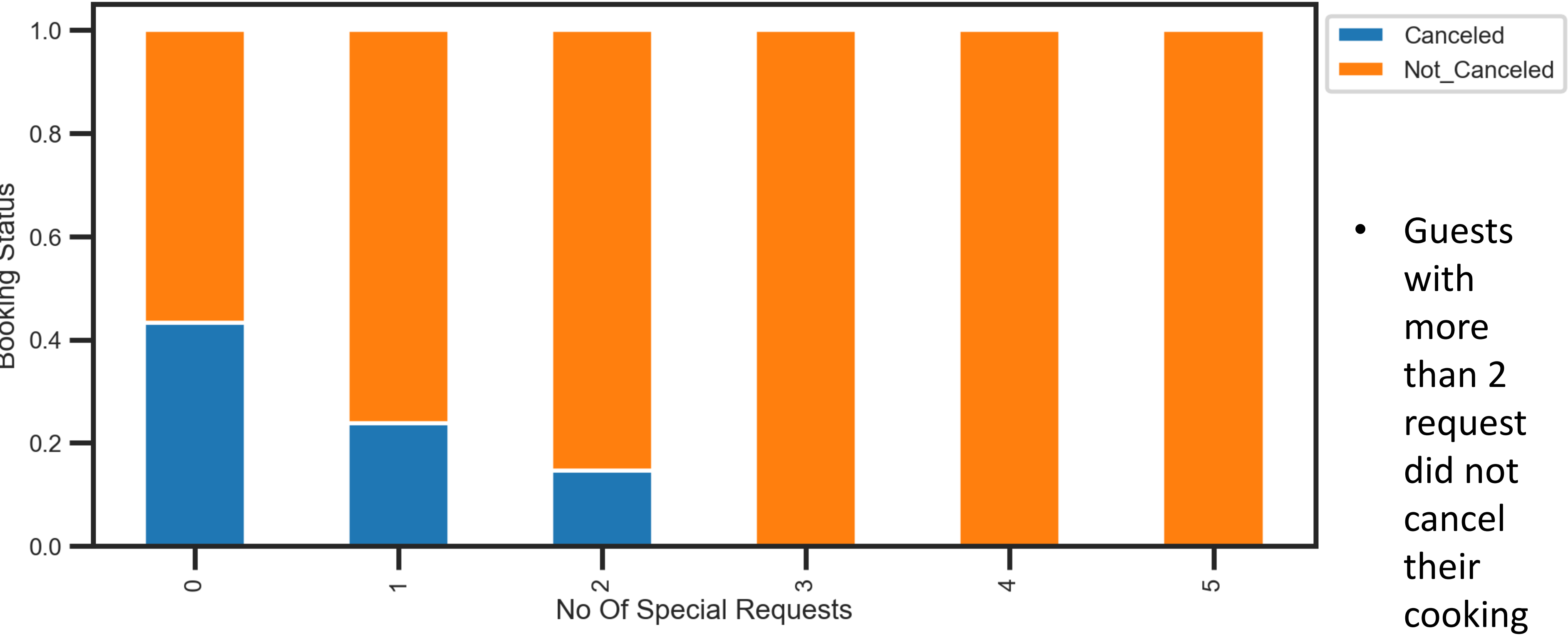
Do market segment affect booking cancellation?



ackground check

EDA Results

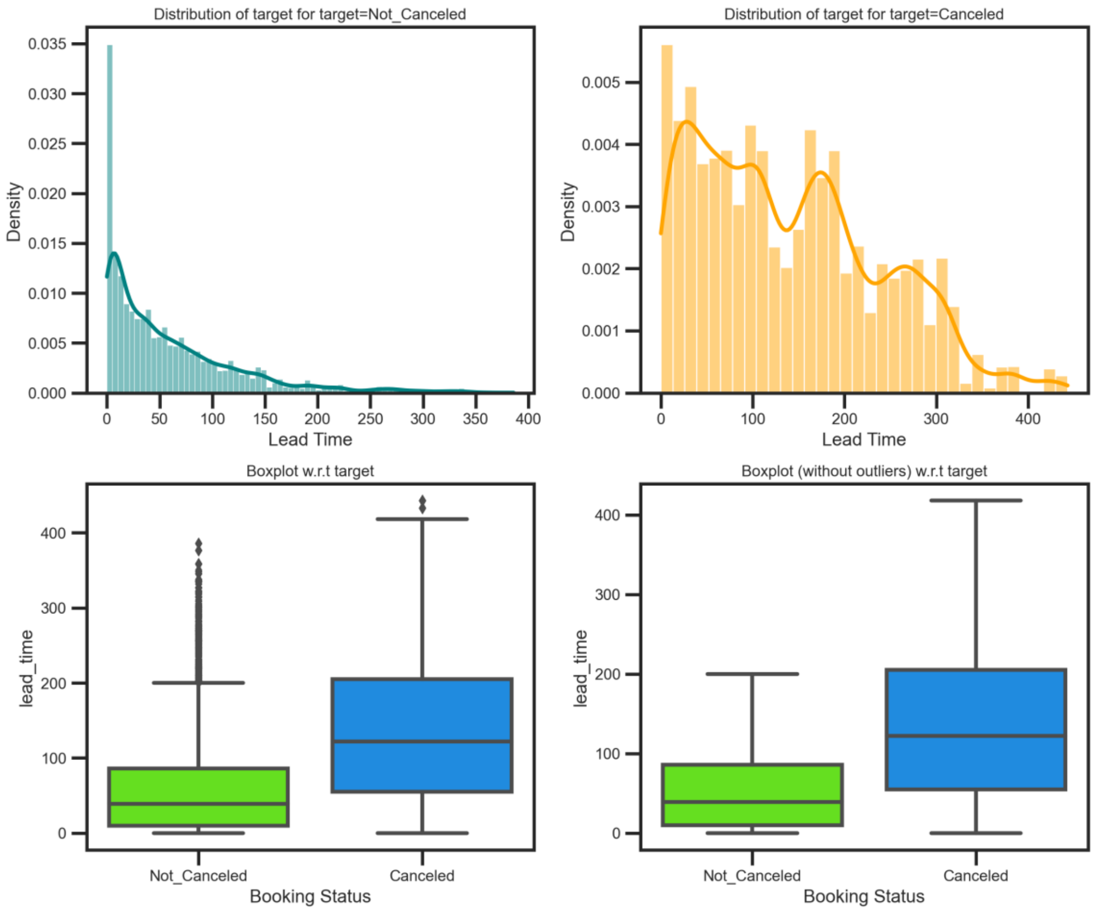
- Do guest requirements affect booking cancellation?



- Guests with more than 2 request did not cancel their cooking

EDA Results

- The more lead time, the more likely the booking will be canceled.
- Lead time less than ~80 most likely will not be canceled .



Distribution of target for target=Not_Canceled

Distribution of target for target=Canceled

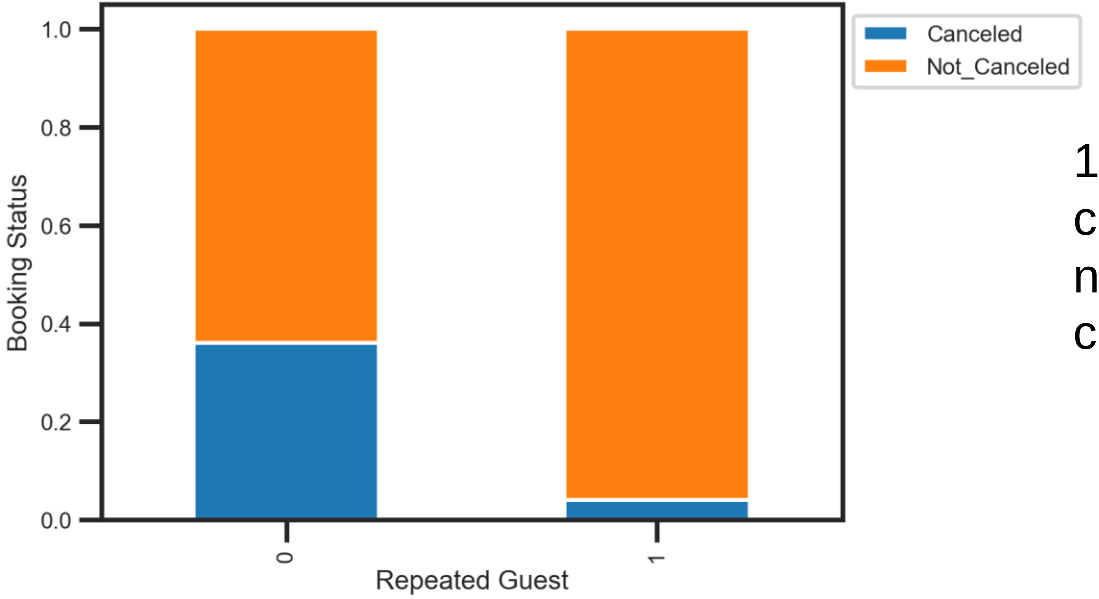
Boxplot w.r.t target

Boxplot (without outliers) w.r.t target

[Link to Appendix slide on data background check]

EDA Results

- Booking Statues with Repeated Guest.

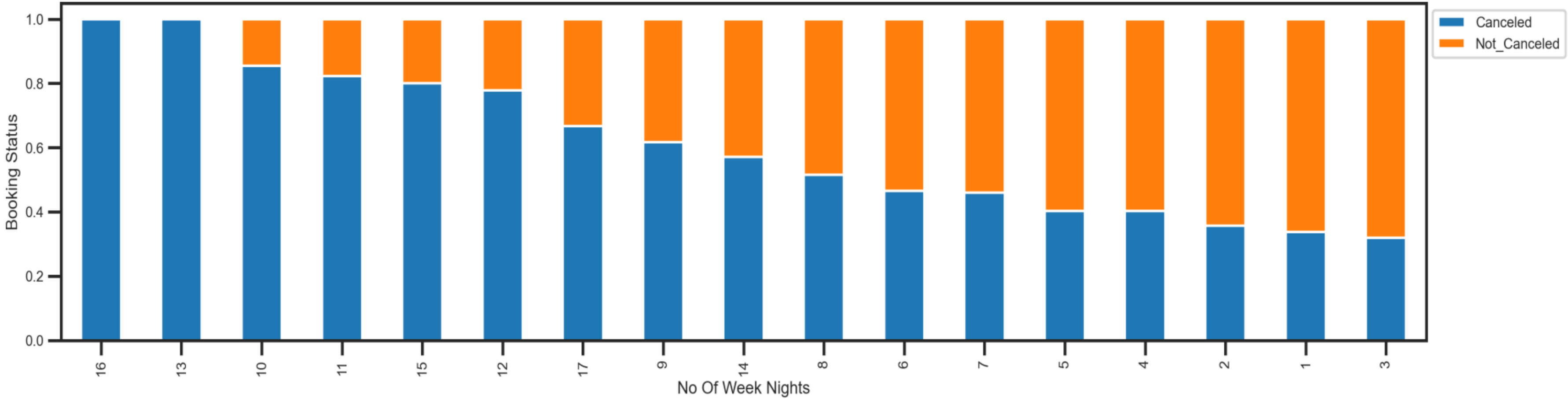


10% repeating guests cancel while ~40% of non repeated guest was canceled

Link to Appendix slide on data background check

EDA Results

- Week-night versus booking statues:

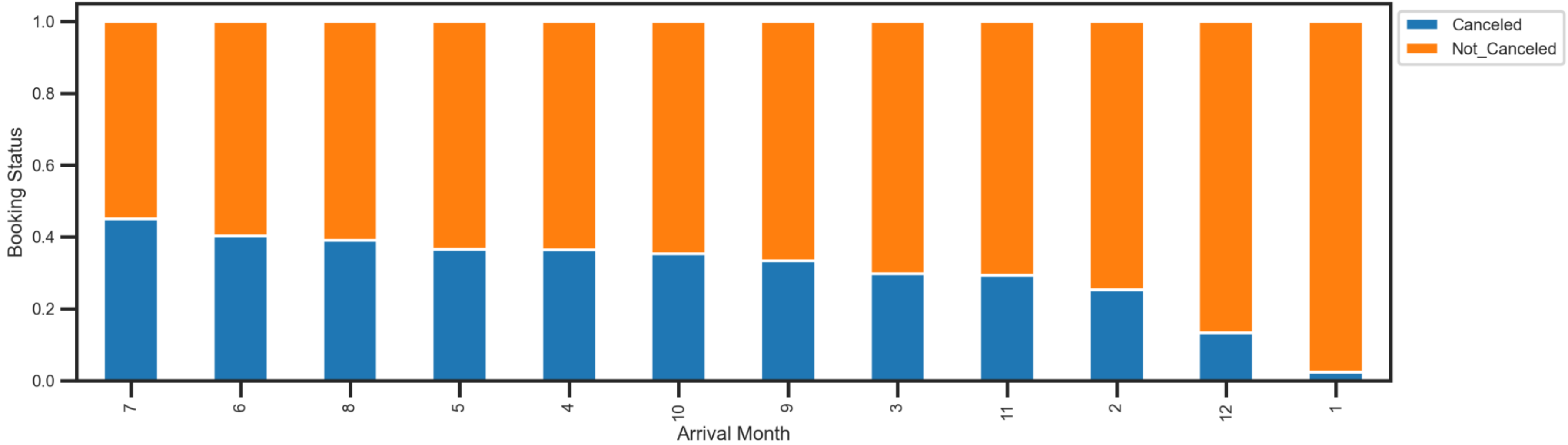


7 nights and less have 50% probability of cancelation comparing to >7 nights. The less weeknights the more likely to not cancel the booking.

Link to Appendix slide on data background check

EDA Results

- Variation the booking statues with the arrival month

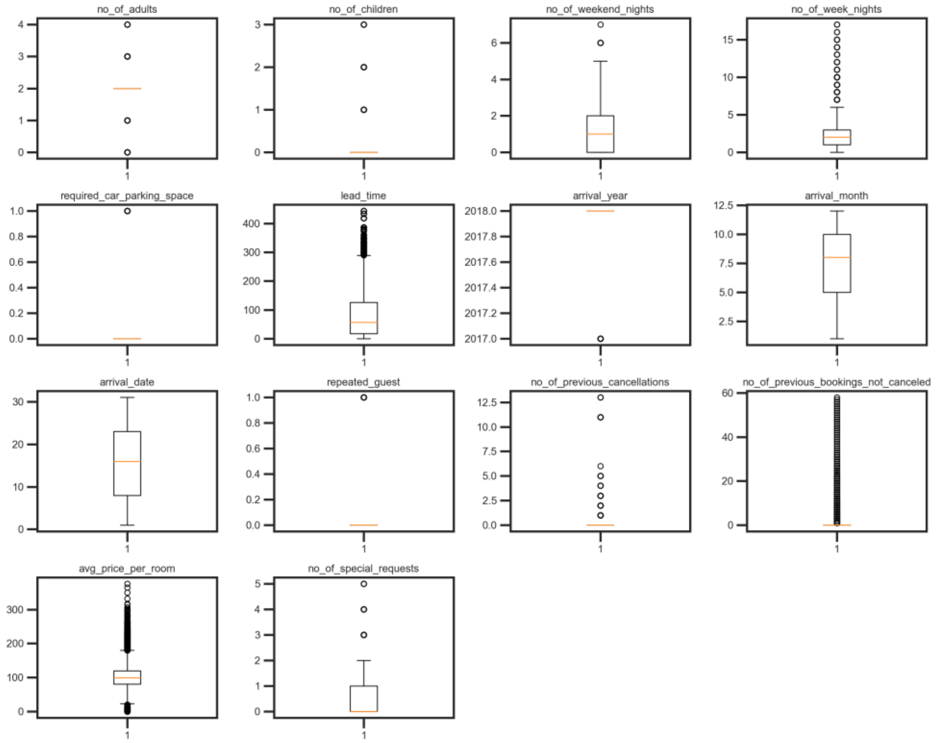


Cancelation is much less during the Winter months (Jan, Dec, February) comparing to summer months (July, June)

[Link to Appendix slide on data background check](#)

Data Preprocessing

- Duplicate value check
  - No duplicate data
- Missing value treatment
  - No missing values
- Outlier check (treatment if needed)
  - We did not remove outlier
- Feature engineering
- Data preparation for modeling
  - One hot encoding for categorical features

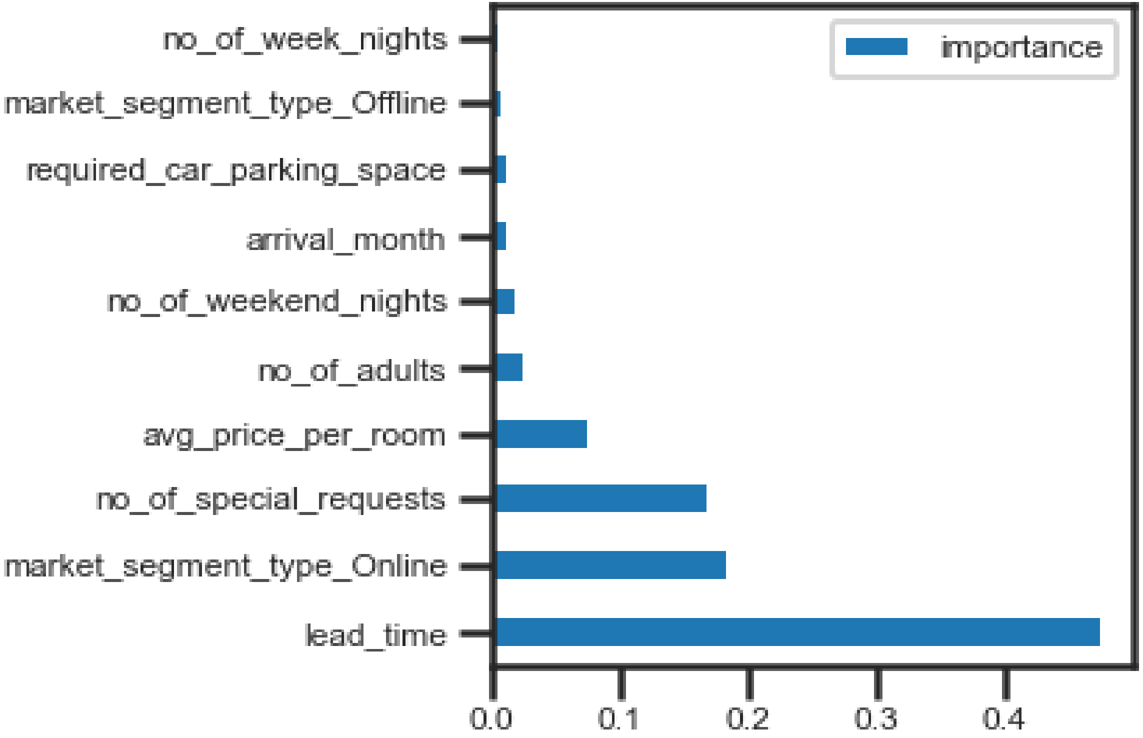


Model Performance Summary

- The final ML model (Decision Tree)and its parameters
  - The model building steps of Decision Tree
    - Split data into test and train sets
    - Hot encoding the catecorical features
    - Decision Tree model
    - DT parameters : (class_weight='balanced', max_depth=6, max_leaf_nodes=50, min_samples_split=10, random_state=1)

Model Performance Summary

- The feature ir



Model Performance Summary
Logistic Regression

Training	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80545	0.79265	0.80132
Recall	0.63267	0.73622	0.69939
Precision	0.73907	0.66808	0.69797
F1	0.68174	0.70049	0.69868

Testing	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80465	0.79555	0.80345
Recall	0.63089	0.73964	0.70358
Precision	0.72900	0.66573	0.69353
F1	0.67641	0.70074	0.69852

Model Performance Summary
Decision Tree

Yes!

Training	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.99421	0.83097	0.89954
Recall	0.98661	0.78608	0.90303
Precision	0.99578	0.72425	0.81274
F1	0.99117	0.75390	0.85551

Testing	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.87118	0.83497	0.86879
Recall	0.81175	0.78336	0.85576
Precision	0.79461	0.72758	0.76614
F1	0.80309	0.75444	0.80848

APPENDIX

Data Background and Contents

* Booking_ID: unique identifier of each booking
* no_of_adults: Number of adults
* no_of_children: Number of Children
* no_of_weekend_nights: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
* no_of_week_nights: Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
* type_of_meal_plan: 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: Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels.
* lead_time: Number of days between the date of booking and the arrival date
* arrival_year: Year of arrival date
* arrival_month: Month of arrival date
* arrival_date: 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: Number of previous bookings that were canceled by the customer prior to the current booking
* no_of_previous_bookings_not_canceled: Number of previous bookings not canceled by the customer prior to the current booking
* avg_price_per_room: Average price per day of the reservation; prices of the rooms are dynamic. (in euros)
* no_of_special_requests: 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.

Data Background and Contents

- 1- 32 % of booking was canceled.
- 2- online market segment has the higher percentage of orders ~63%
- 3- 71% of the booking are 2 adults and 92% without kids.
- 4- Online booking had higher median price while the complementary is the least median price.
- 5- Booking with lower price median ~90$ did not canceled comparing with median price 110$ which was canceled .
- 6- Complementary and Corporate have the least cancelation number comparing to other market segments.
- 7- Guests with more than 2 requests did not cancel their booking
- 8- The more lead time, the more likely the booking will be canceled. Lead time less than ~80 most likely will not be canceled
- 9- Heatmap show that There is no strong correlation between the features.
- 10- 10% repeating guests cancel while ~40% of non repeated guest was canceled
- 11- 7 night and less have 50% cancelation comparing to >7 nights. The less week-nights (1-3 nights) the more likely to not cancel the booking.
- 12- Cancelation is much less during the Winter months (Jan, Dec, February) comparing to summer months (July, June)

Model Building - Logistic Regression

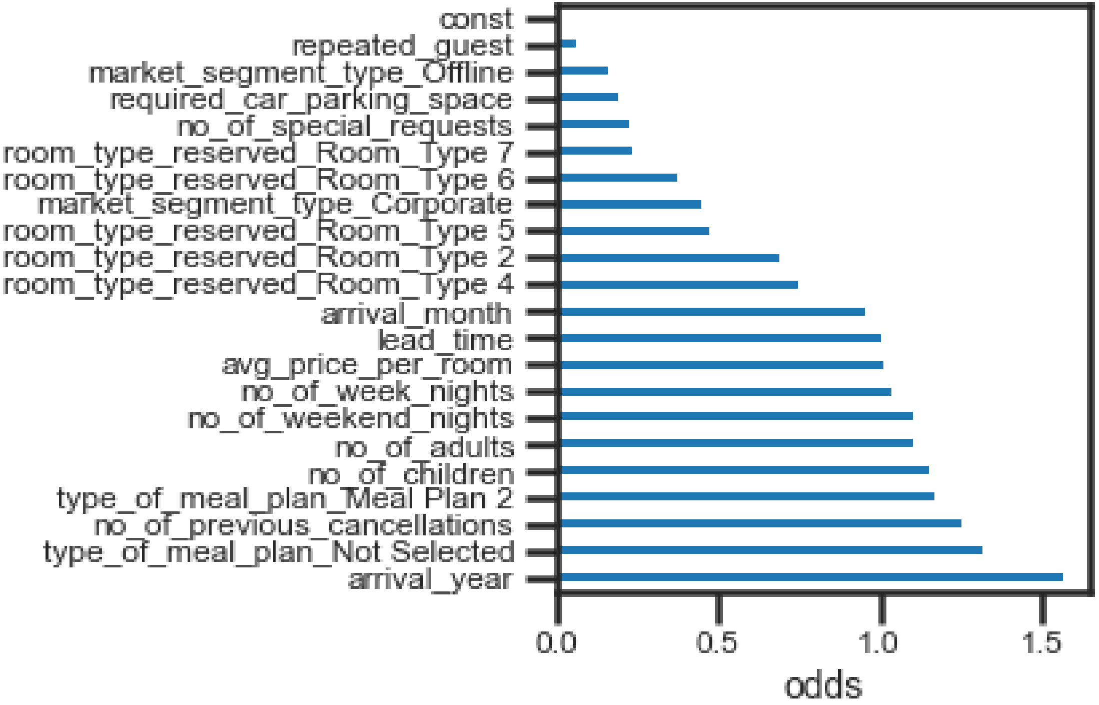
The tests conducted to check the assumptions of Logistic Regression

Checked for P values if P values >0.05 that means the feature does not impact the target. We check the multicollinearity by applying Variance inflation factor (VIF) and therefore we drop these features.

['arrival_date', 'no_of_previous_bookings_not_canceled', 'type_of_meal_plan_Meal Plan 3', 'room_type_reserved_Room_Type 3', 'market_segment_type_Complementary', 'market_segment_type_Online’]

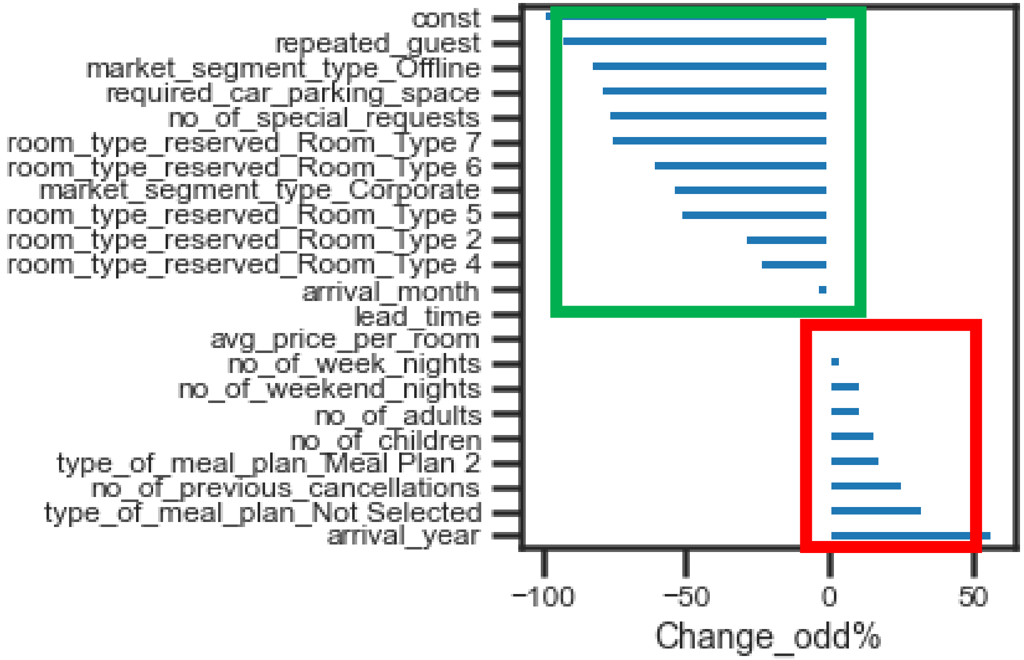
Model Building - Logistic Regression

- Summary of most important features used by the LR model for prediction after convert logit to Odds



Model Building - Logistic Regression

- Summary of most important features used by the ML model for prediction after convert logit to percentage change in odds



Increase the none-cancelation

Increase the cancelation

Model Building - Logistic Regression

Interpret the results based on coefficients and odds

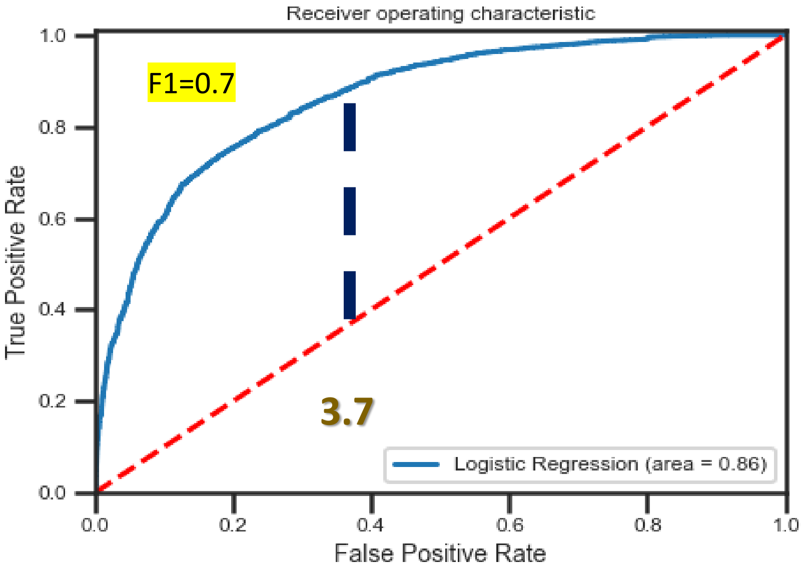
Let’s pick *arrival year* and see how it impacts the chances of cancelation. Increasing the arrival year by 1 unit (1 year) will increase the odd by 1.56, which is 57% increase in the odd of cancelation.

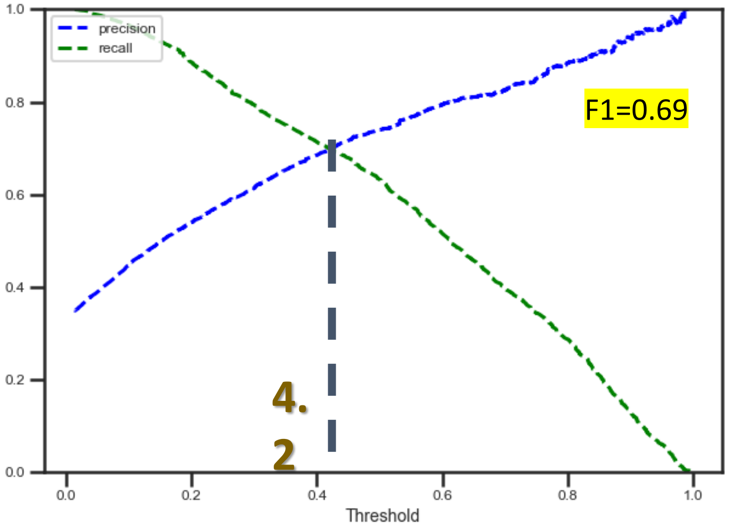
***type_of_meal_plan_Not Selected:* the odd is 1.33 . The odds for cancelation is 33% higher for type_of_meal_plan_Not Selected .**

***Repeated_guest*: the odd is 0.06 . The odds for cancelation is 93% lower for repeated guest.**

Model Performance Evaluation and Improvement - Logistic Regression

- Improvement in the model performance by changing the classification threshold





Model Building - Decision Tree

- **The model building steps of Decision Tree**
  - **Split data into test and train sets**
  - **Hot encoding the catecorical features**
  - **Decision Tree model.**
  - **Pre pruning (Grid search with cross validation)**
  - **Post pruning (cost complexity pruning)**

Model Performance Evaluation and Improvement - Decision Tree

Training	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.99421	0.83097	0.89954
Recall	0.98661	0.78608	0.90303
Precision	0.99578	0.72425	0.81274
F1	0.99117	0.75390	0.85551

Test	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.87118	0.83497	0.86879
Recall	0.81175	0.78336	0.85576
Precision	0.79461	0.72758	0.76614
F1	0.80309	0.75444	0.80848

Model Performance Evaluation and Improvement - Decision Tree

- Decision rules

1. Predicting a customer will not cancel their booking but in reality, the customer will cancel their booking.

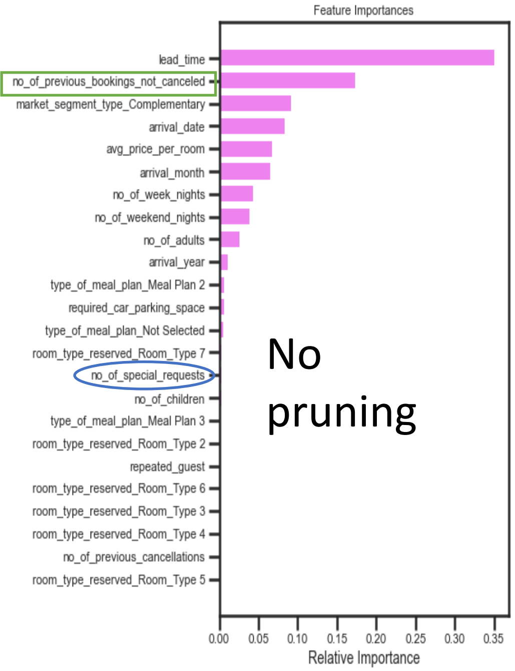
If we predict that a booking will not be canceled and the booking gets canceled, then the hotel will lose resources and will have to bear additional costs of distribution channels.

2. Predicting a customer will cancel their booking but in reality, the customer will not cancel their booking.

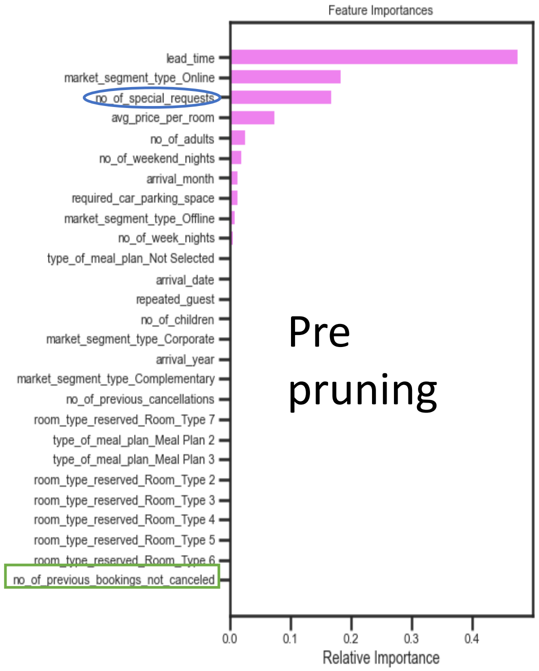
If we predict that a booking will not be canceled and the booking gets canceled, then the hotel will lose resources and will have to bear additional

Model Performance Evaluation and Improvement - Decision Tree

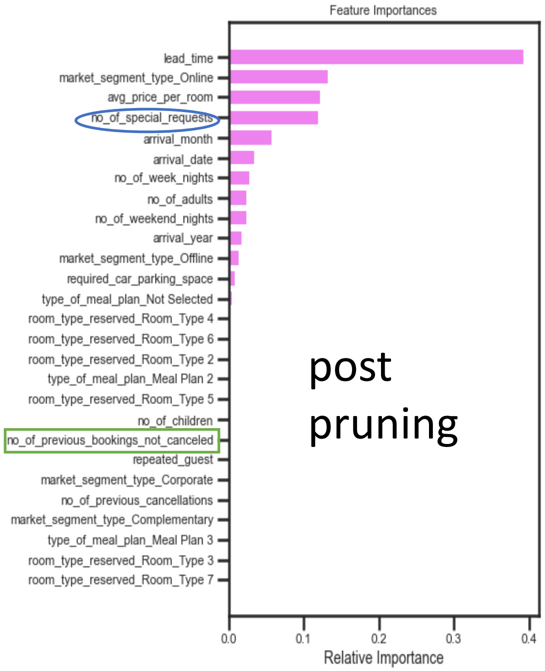
- the feature importance



No pruning

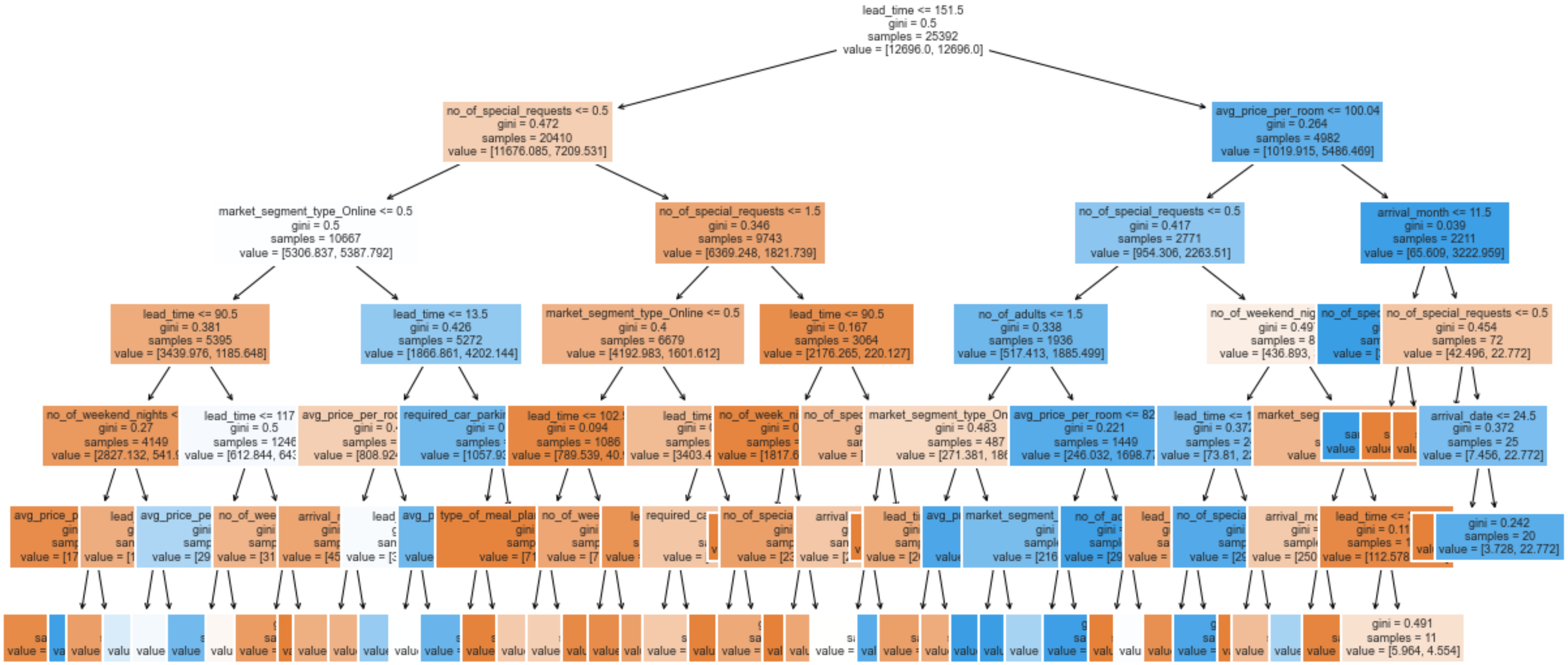


Pre pruning



post pruning

Pre Pruning DT



lead_time <= 151.5
gini = 0.5
samples = 25392
value = [12696.0, 12696.0]
no_of_special_requests <= 0.5
gini = 0.472
samples = 20410
value = [11676.085, 7209.531]
avg_price_per_room <= 100.04
gini = 0.264
samples = 4982
value = [1019.915, 5486.469]
market_segment_type_Online <= 0.5
gini = 0.5
samples = 10667
value = [5306.837, 5387.792]
no_of_special_requests <= 1.5
gini = 0.346
samples = 9743
value = [6369.248, 1821.739]
no_of_special_requests <= 0.5
gini = 0.417
samples = 2771
value = [954.306, 2263.51]
arrival_month <= 11.5
gini = 0.039
samples = 2211
value = [65.609, 3222.959]
lead_time <= 90.5
gini = 0.381
samples = 5395
value = [3439.976, 1185.648]
lead_time <= 13.5
gini = 0.426
samples = 5272
value = [1866.861, 4202.144]
market_segment_type_Online <= 0.5
gini = 0.4
samples = 6679
value = [4192.983, 1601.612]
lead_time <= 90.5
gini = 0.167
samples = 3064
value = [2176.265, 220.127]
no_of_adults <= 1.5
gini = 0.338
samples = 1936
value = [517.413, 1885.499]
no_of_weekend_nig
gini = 0.49
samples = 8
value = [436.893,
no_of_spec
value = [
no_of_special_requests <= 0.5
gini = 0.454
samples = 72
value = [42.496, 22.772]
no_of_weekend_nights <
gini = 0.27
samples = 4149
value = [2827.132, 541.9
lead_time <= 117
gini = 0.5
samples = 1246
value = [612.844, 643
avg_price_per_roo
gini = 0.
samples =
value = [808.92
required_car_parkir
gini = 0
samples
value = [1057.93
lead_time <= 102.5
gini = 0.094
samples = 1086
value = [789.539, 40.
lead_time
gini =
samples
value = [3403.4
no_of_week_ni
gini = 0
samples =
value = [1817.6
no_of_spec
gi
sar
value = [
market_segment_type_On
gini = 0.483
samples = 487
value = [271.381, 186
avg_price_per_room <= 82
gini = 0.221
samples = 1449
value = [246.032, 1698.7
lead_time <= 1
gini = 0.372
samples = 2
value = [73.81, 2
market_seg
value
arrival_date <= 24.5
gini = 0.372
samples = 25
value = [7.456, 22.772]
avg_price_p
no_of_wee
arrival_
type_of_meal_plai
required_ca
no_of_specia
arrival
lead_ti
avg_p
market_segment_
no_of_ac
lead_time <=
gini = 0.11
value = [112.578
gini = 0.242
samples = 20
value = [3.728, 22.772]
gini = 0.491
samples = 11
value = [5.964, 4.554]