INN Hotels Project

Supervised Learning - Classification

Yair Brama – September 2024

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Logistic Regression Performance Summary
- Decision Tree Performance Summary
- Appendix

Executive Summary - Business Context

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.

By using machine learning we will build classification models that will help us to predict, when a hotel room is likely to be cancelled, so we can perform the steps that will help us to lower the cost of having un-planned empty rooms.

Supervised Machine Learning – Linear Regression

Performing a logistic regression on the data set (training and test) to find how much a unit change in each of the parameters will increase the odds of having a room cancelled last minute.

Supervised Machine Learning – Decision Tree

Performing a decision tree model on the same data set to find which parameters are the most indicative in classification an order, whether it will be cancelled last minute or not.

Executive Summary – Logistic Regression Model Results

Best results – Using threshold of 0.42

Training performance:

	Accuracy	Recall	Precision	F1
0	0.80128	0.69927	0.69794	0.69860

Test performance:

	Accuracy	Recall	Precision	F1
0	0.80364	0.70386	0.69381	0.69880

Summary – The best results of the logistic regression were achieved by using the threshold of **0.42**

This gives us a F1 score of almost 70% success in predicting if an order will be cancelled last minute, given the order details.

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25370
Method:	MLE	Df Model:	21
Date:	Sun, 22 Sep 2024	Pseudo R-squ.:	0.3282
Time:	09:41:25	Log-Likelihood:	-10810.
converged:	True	LL-Null:	-16091.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-916.8647	120.456	-7.612	0.000	-1152.953	-680.776
no_of_adults	0.1087	0.037	2.916	0.004	0.036	0.182
no_of_children	0.1520	0.057	2.656	0.008	0.040	0.264
no_of_weekend_nights	0.1086	0.020	5.497	0.000	0.070	0.147
no_of_week_nights	0.0417	0.012	3.400	0.001	0.018	0.066
required_car_parking_space	-1.5947	0.138	-11.564	0.000	-1.865	-1.324
lead_time	0.0157	0.000	59.216	0.000	0.015	0.016
arrival_year	0.4529	0.060	7.587	0.000	0.336	0.570
arrival_month	-0.0425	0.006	-6.586	0.000	-0.055	-0.030
repeated_guest	-2.7358	0.557	-4.914	0.000	-3.827	-1.645
no_of_previous_cancellations	0.2288	0.077	2.982	0.003	0.078	0.379
avg_price_per_room	0.0192	0.001	26.382	0.000	0.018	0.021
no_of_special_requests	-1.4700	0.030	-48.893	0.000	-1.529	-1.411
type_of_meal_plan_Meal Plan 2	0.1643	0.067	2.470	0.014	0.034	0.295
type_of_meal_plan_Not Selected	0.2860	0.053	5.407	0.000	0.182	0.390
room_type_reserved_Room_Type 2	-0.3557	0.131	-2.722	0.006	-0.612	-0.100
room_type_reserved_Room_Type 4	-0.2833	0.053	-5.344	0.000	-0.387	-0.179
room_type_reserved_Room_Type 5	-0.7362	0.208	-3.534	0.000	-1.144	-0.328
room_type_reserved_Room_Type 6	-0.9667	0.147	-6.584	0.000	-1.254	-0.679
room_type_reserved_Room_Type 7	-1.4338	0.293	-4.901	0.000	-2.007	-0.860
market_segment_type_Corporate	-0.7927	0.103	-7.710	0.000	-0.994	-0.591
market_segment_type_Offline	-1.7855	0.052	-34.377	0.000	-1.887	-1.684

Executive Summary – Decision Tree Model Results

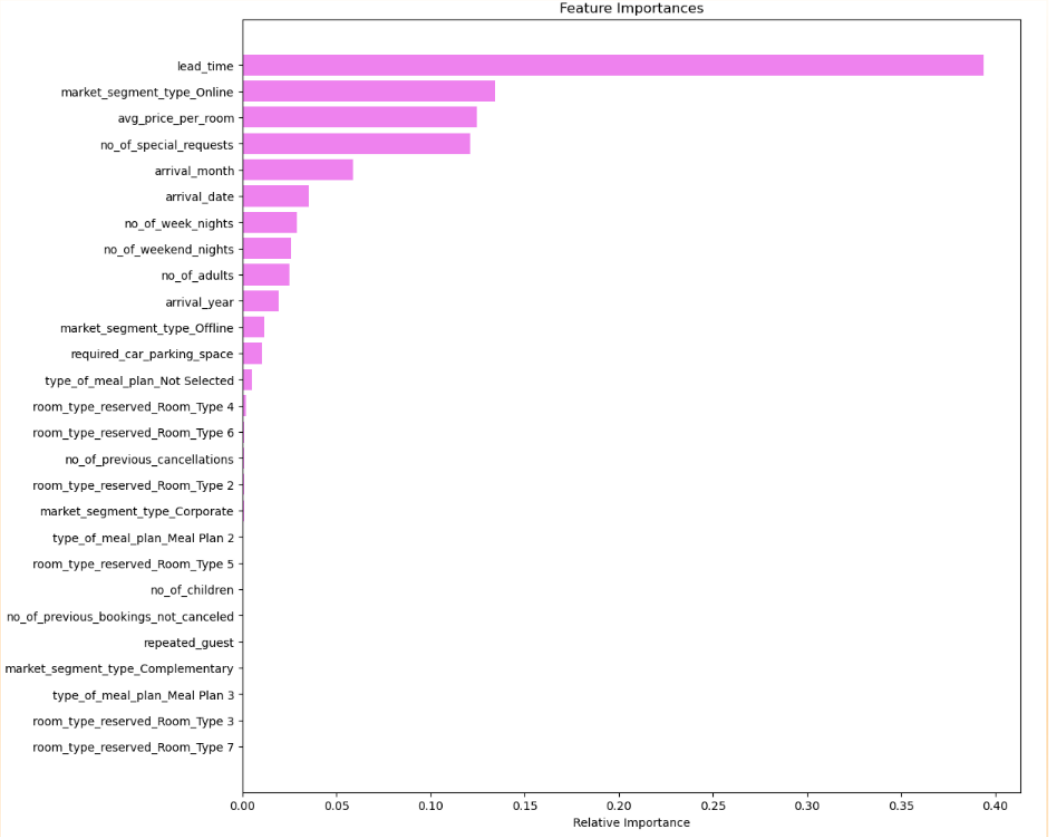
Best results – Using
ccp_alpha=0.00012267633155167048

Training set results

	Accuracy	Recall	Precision	F1
0	0.89993	0.90291	0.81369	0.85598

Test set results

	Accuracy	Recall	Precision	F1
0	0.86943	0.85662	0.76710	0.80939



Summary – The best results of the decision tree are received by post-pruning and re-running it with ccp_alpha - 0.00012267633155167048
This gives us a F1 score of almost 81% success in predicting if an order will be cancelled last minute given the order details.

Executive Summary – Final Recommendations

Model	Pros	Cons
Logistical Regression	Easier to understand how each parameter influences the prediction	The results show only (70%-80%) success in predicting if an order will be cancelled at the last minute
Decision Tree	• Excellent results in predicting cancelled orders (80%-86%) • No need to deal with model assumptions (multicollinearity, linearity, normality, homoscedasticity)	Less transparent about how a unit change in one feature (when all the other feature remain the same) will increase the odds of a cancellation

Based on the results of our models, we recommend to use the **decision tree model, with post-pruning** (cost-complexity pruning).

Solution Approach – Logistic Regression

Data initial review	EDA	Data Preprocessing	Model Building - Linear Regression	Check assumptions	Check different thresholds	Final Model Summary
Check for null values Check for duplicates Check data types	Univariate Analysis – Box plots, Histograms Multivariate Analysis - Check for correlation (Heatmap), visualize the correlated variables	Feature Engineering – Add field for number of years Create Dummy variables – For the categories' fields – market type, meal types, room types	Split the dataset between training and test data Run the model and check the performance	Remove Multicollinearity Remove high P-values	Based on ROC-AOC curve and Precision-Recall curve find the optimal thresholds Calculate the confusion matrix for each threshold	Perform the final run Compare the training and the test sets

Solution Approach – Decision Tree

Data initial review	EDA	Data Preprocessing	Model Building – Decision Tree	Model after pre-pruning	Model after post-pruning	Final Model Summary
Check for null values Check for duplicates Check data types	Univariate Analysis – Box plots, Histograms Multivariate Analysis - Check for correlation (Heatmap), visualize the correlated variables	Feature Engineering – Add field for number of years Create Dummy variables – For the categories' fields – market type, meal types, room types	Split the dataset between training and test data Run the model and check the performance	Perform GridSearch to find the optimal values of the hyperparameters Run the model with the new values Compare training and test results	Based on optimal ccp_alpha values, find the nodes that can be pruned Compare the confusion matrix of the training and test sets	Compare the results of all iterations and recommend the best option

Data Description

- 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.

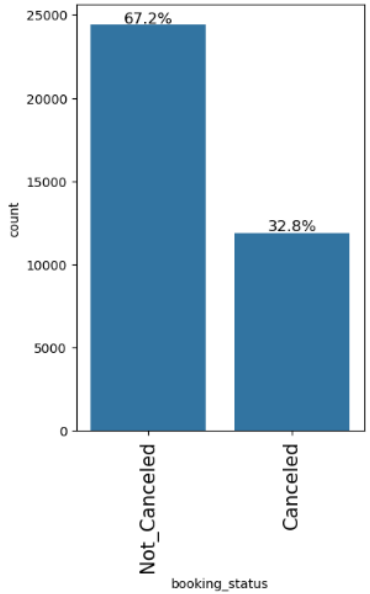
Data Description – Cont.

- 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.

EDA Results - Data Overview

The data includes 36275 rows and 19 columns. There are no missing values and no duplicates.

The data shows us that 1/3 of the orders are cancelled.

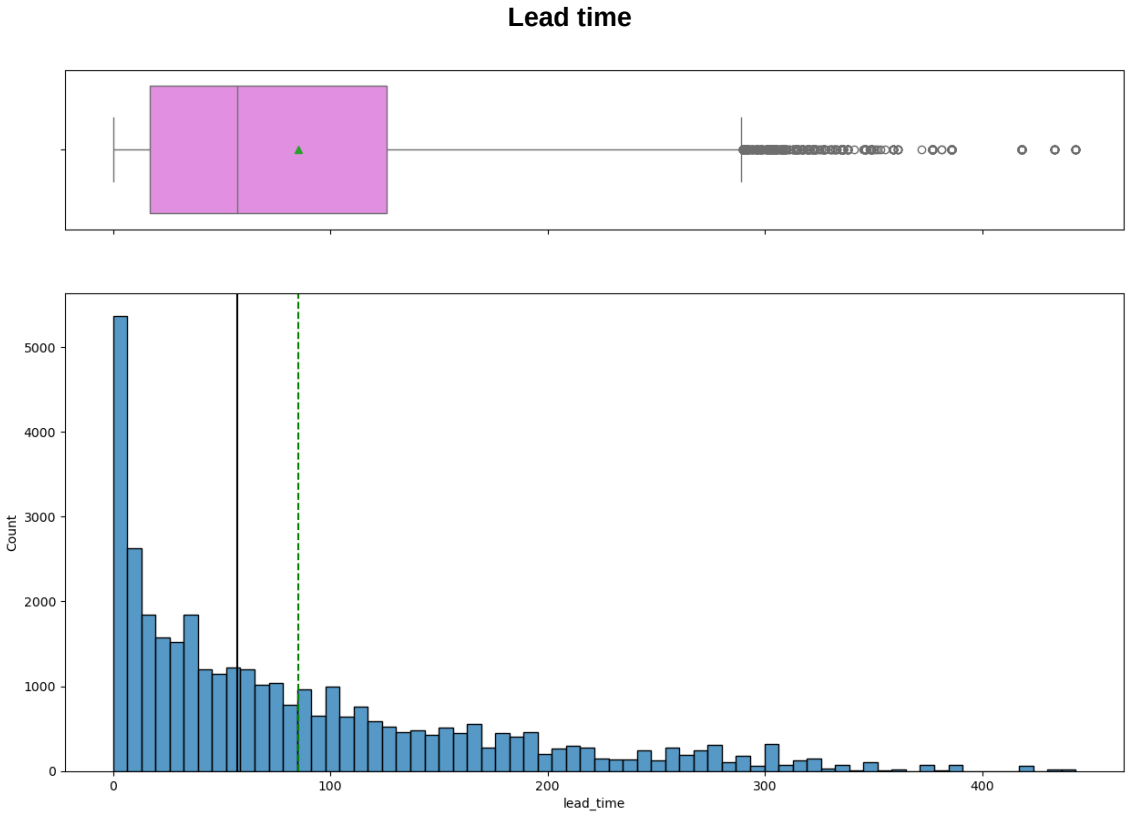


Statistical summary of the numeric fields:

	count	mean	std	min	25%	50%	75%	max
no_of_adults	36275.00000	1.84496	0.51871	0.00000	2.00000	2.00000	2.00000	4.00000
no_of_children	36275.00000	0.10528	0.40265	0.00000	0.00000	0.00000	0.00000	10.00000
no_of_weekend_nights	36275.00000	0.81072	0.87064	0.00000	0.00000	1.00000	2.00000	7.00000
no_of_week_nights	36275.00000	2.20430	1.41090	0.00000	1.00000	2.00000	3.00000	17.00000
required_car_parking_space	36275.00000	0.03099	0.17328	0.00000	0.00000	0.00000	0.00000	1.00000
lead_time	36275.00000	85.23256	85.93082	0.00000	17.00000	57.00000	126.00000	443.00000
arrival_year	36275.00000	2017.82043	0.38384	2017.00000	2018.00000	2018.00000	2018.00000	2018.00000
arrival_month	36275.00000	7.42365	3.06989	1.00000	5.00000	8.00000	10.00000	12.00000
arrival_date	36275.00000	15.59700	8.74045	1.00000	8.00000	16.00000	23.00000	31.00000
repeated_guest	36275.00000	0.02564	0.15805	0.00000	0.00000	0.00000	0.00000	1.00000
no_of_previous_cancellations	36275.00000	0.02335	0.36833	0.00000	0.00000	0.00000	0.00000	13.00000
no_of_previous_bookings_not_canceled	36275.00000	0.15341	1.75417	0.00000	0.00000	0.00000	0.00000	58.00000
avg_price_per_room	36275.00000	103.42354	35.08942	0.00000	80.30000	99.45000	120.00000	540.00000
no_of_special_requests	36275.00000	0.61966	0.78624	0.00000	0.00000	0.00000	1.00000	5.00000

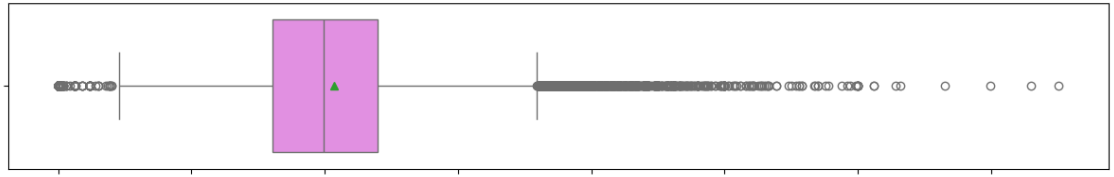
EDA Results - Univariate Analysis

Observation – The time between reserving the room and the actual stay is right-skewed, meaning most of the orders are done closer than 3 months. Point to check – Are cancellations happen more often when lead time is shorter?

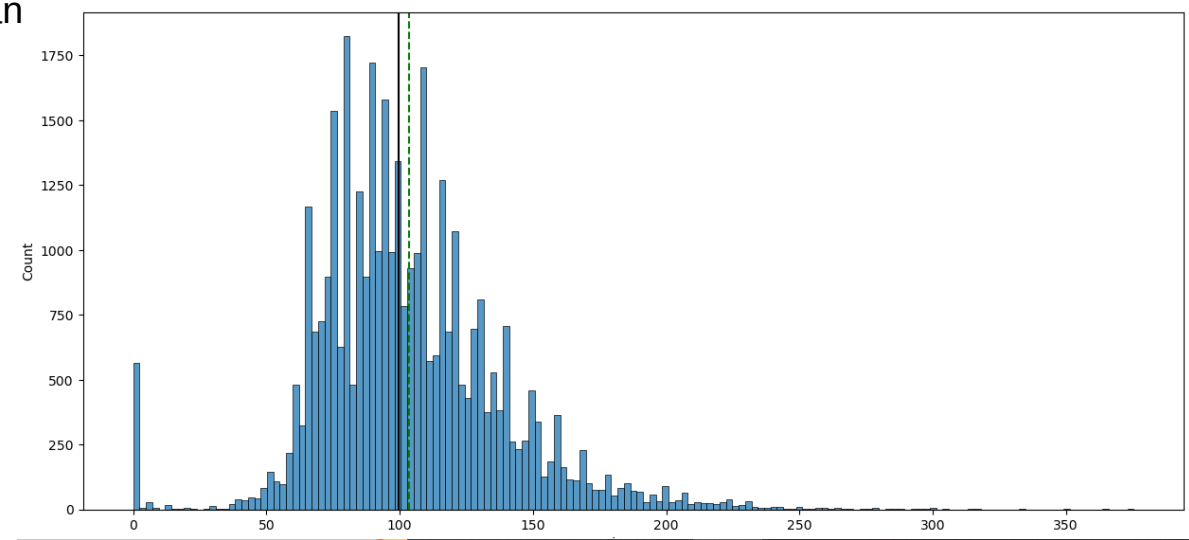


EDA Results - Univariate Analysis

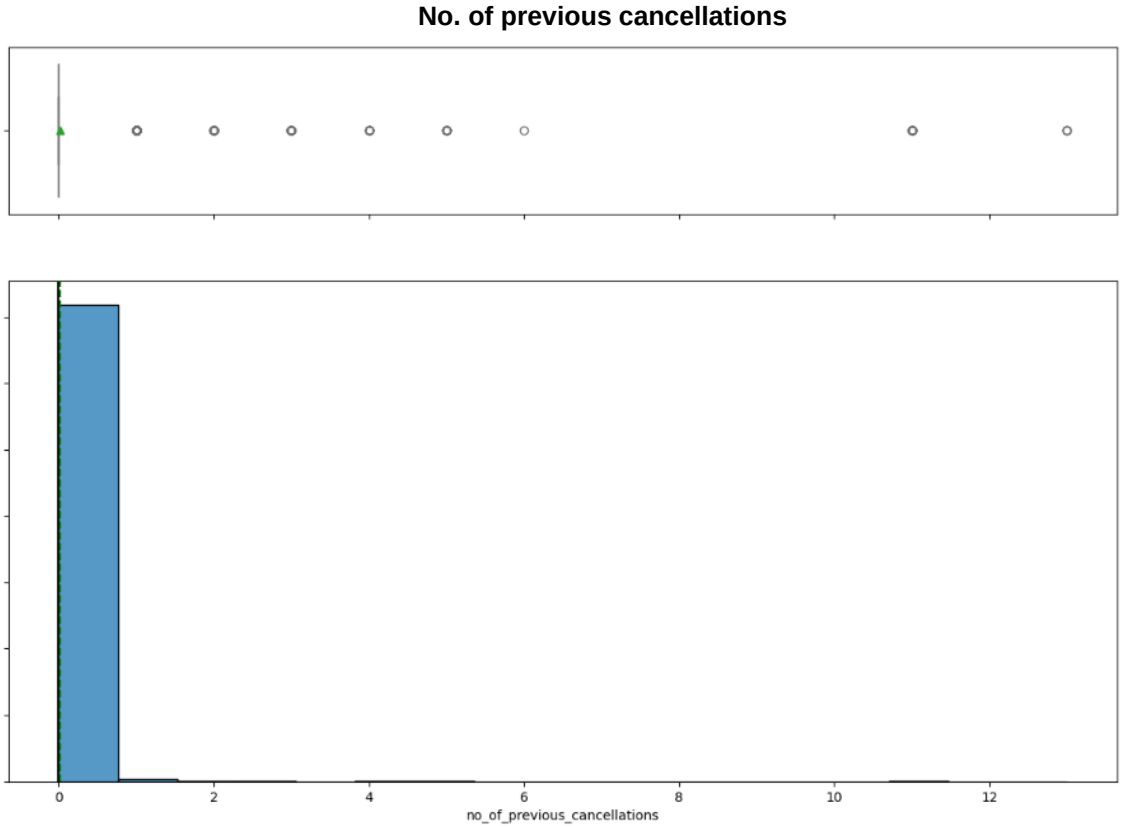
Avg price per night, after removing the >$500 prices



Observation – The average price per night is slightly right-skewed. The mean price is around $100.
Outliers – Prices of hotel rooms vary, and we need to see if it plays a part in cancellation considerations. However, we curbed the highest prices to be $500, since the highest end of the prices might impact the model negatively.



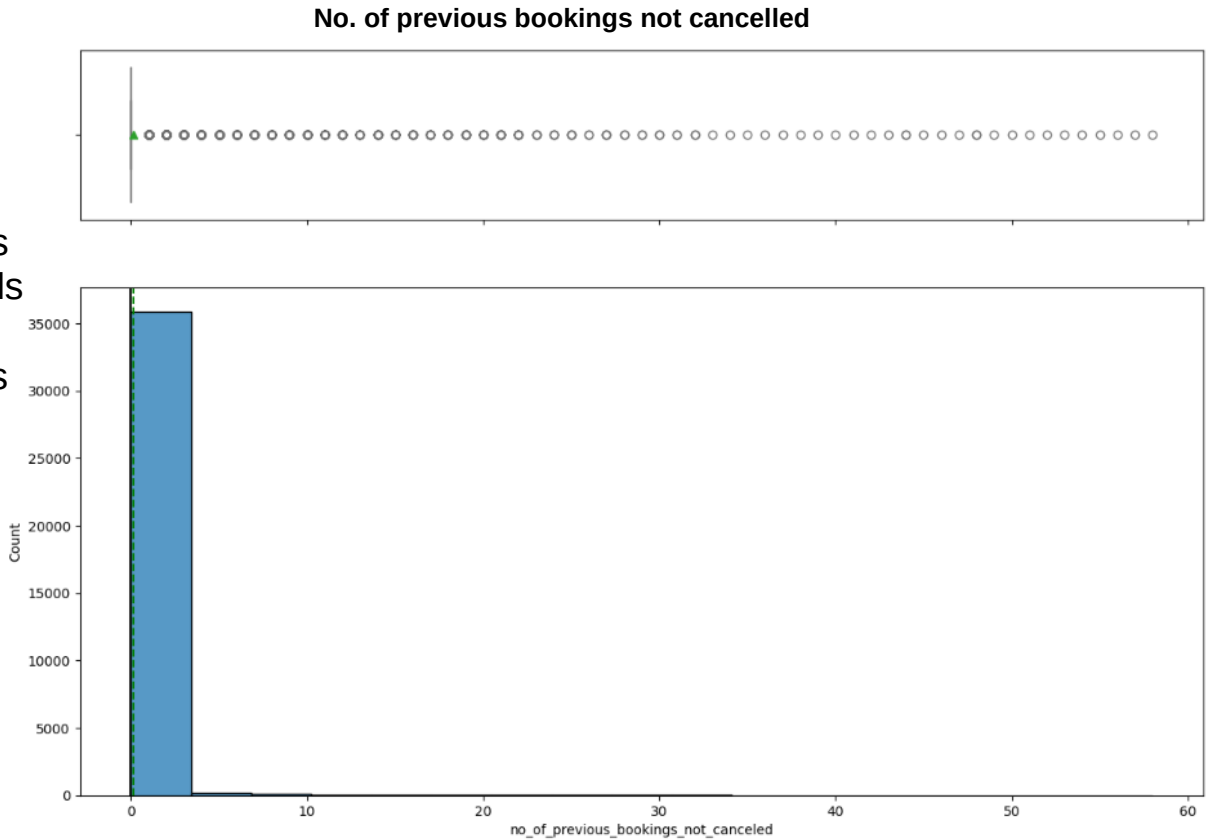
EDA Results - Univariate Analysis



Observation – The number of previous cancellations is between 0 and 12, which tells us that people are probably not booking and cancelling at the same hotel chain over and over.
It might indicate that people cancel due to a reason that is related to their stay, and therefore they will not try again in the same chain.

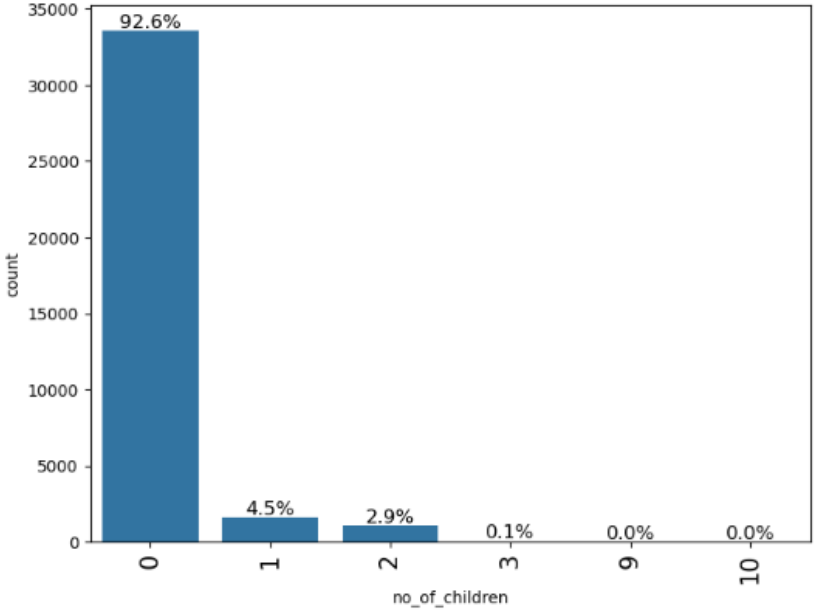
EDA Results - Univariate Analysis

Observation – The number of previous booking is between 0 and 60, which tells us that people who come back often to the same hotel, do that a lot, and that is a good sign of their satisfaction.
Further analysis is required to check if cancellations in these cases are more frequent, the more often the visits are.

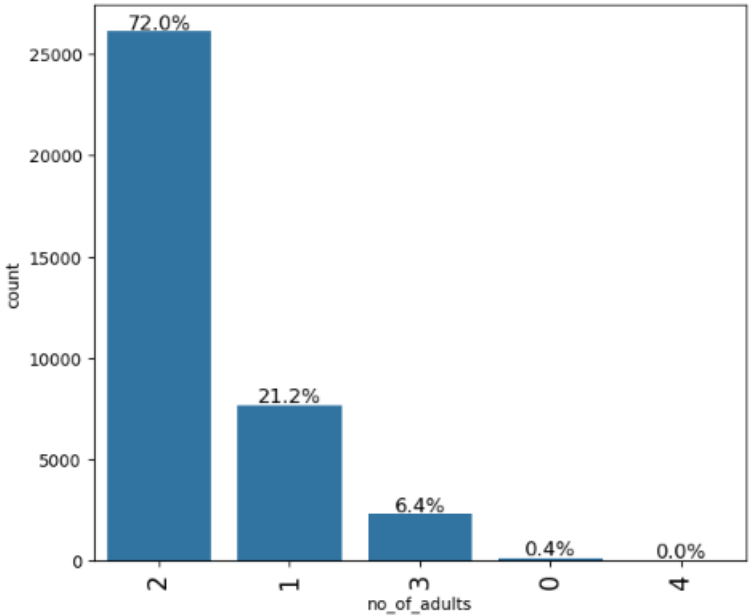


EDA Results - Univariate Analysis

No. of children



No. of adults

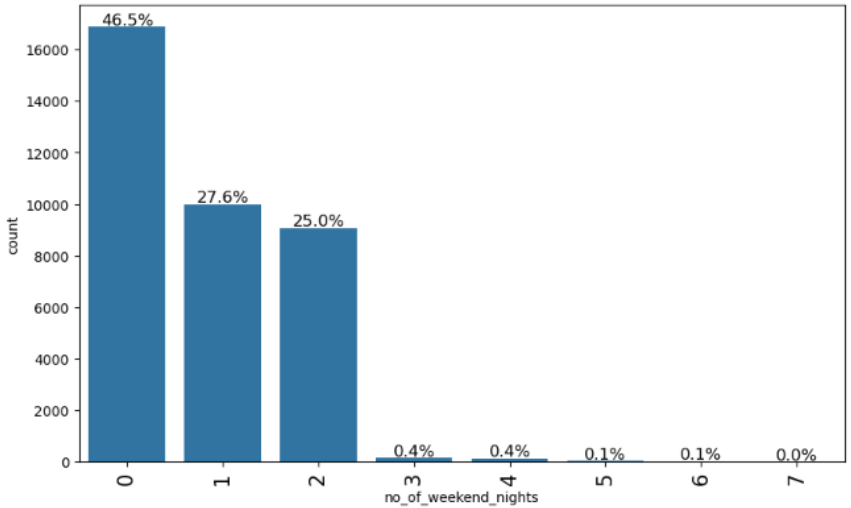


Observation – Not surprisingly, over 90% of the reservations are made by 1 or 2 adults, without childrens.

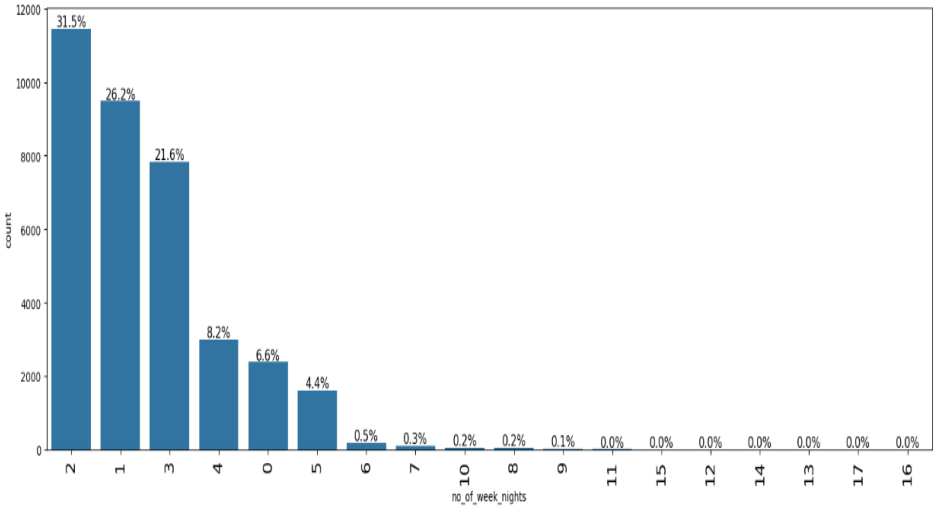
EDA Results - Univariate Analysis

Observation – Over 50% of the reservations are for 1-2 nights. Most of the weekend nights orders are for 1 or 2 nights (one weekend). Further analysis is needed to check if cancellations are more likely for a shorter duration of the stay or are there more likely for the short weekends stays.

No. of weekend nights



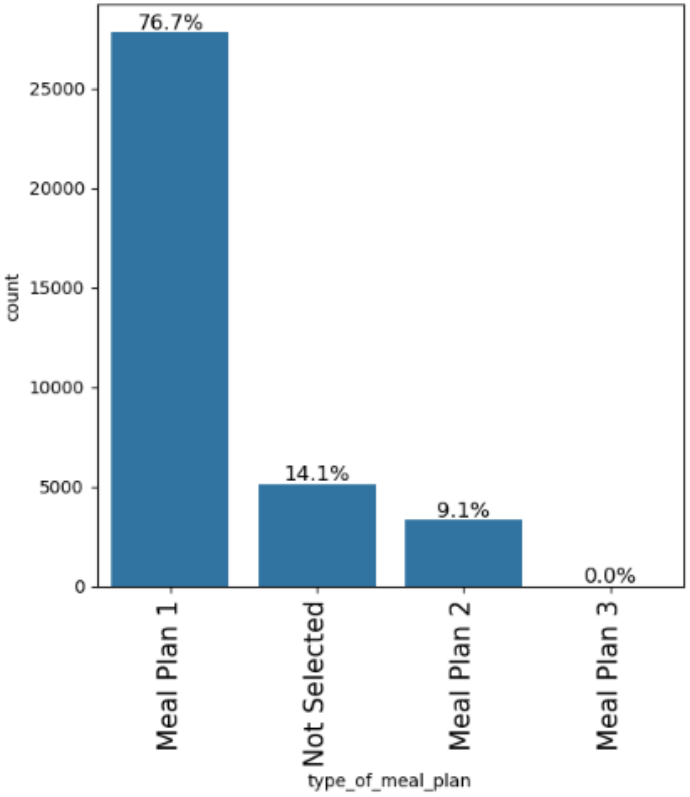
No. of weeknights



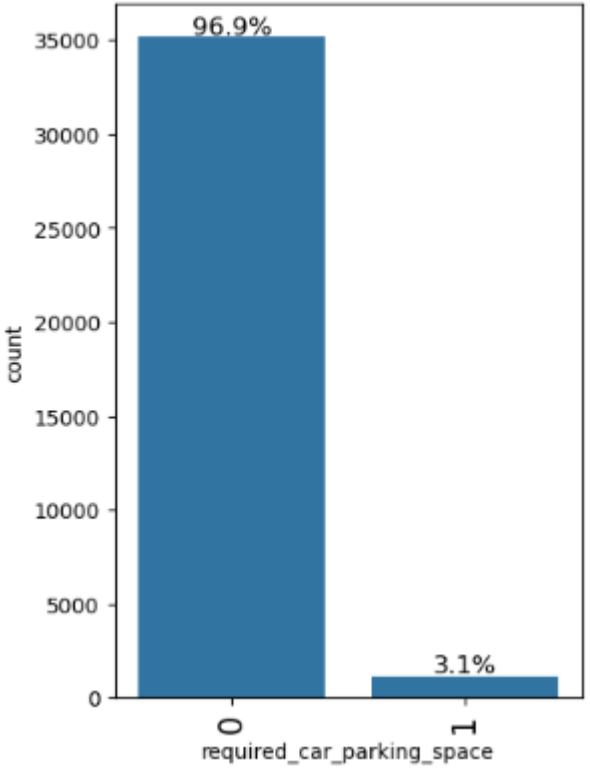
EDA Results - Univariate Analysis

Observation – 97% of reservations do not require car space, and 77% chose the basic meal plan. It seems that these attributes are not relevant to the cancellation rate.

Type of meal plan

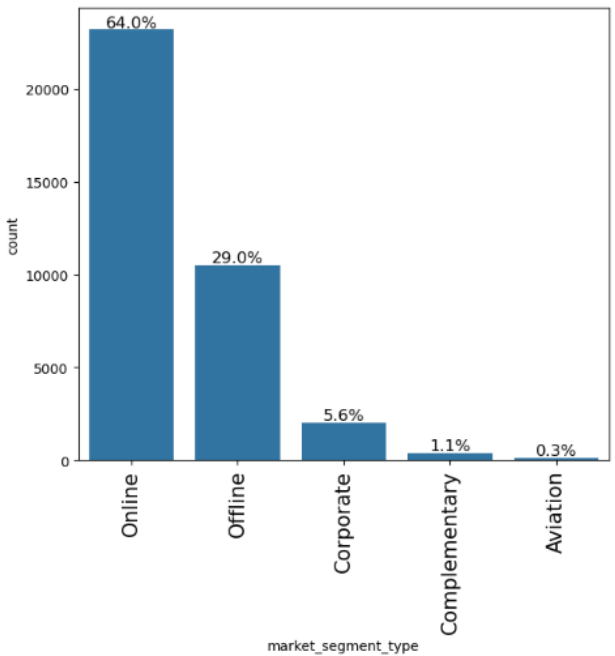


Reservations that required car space

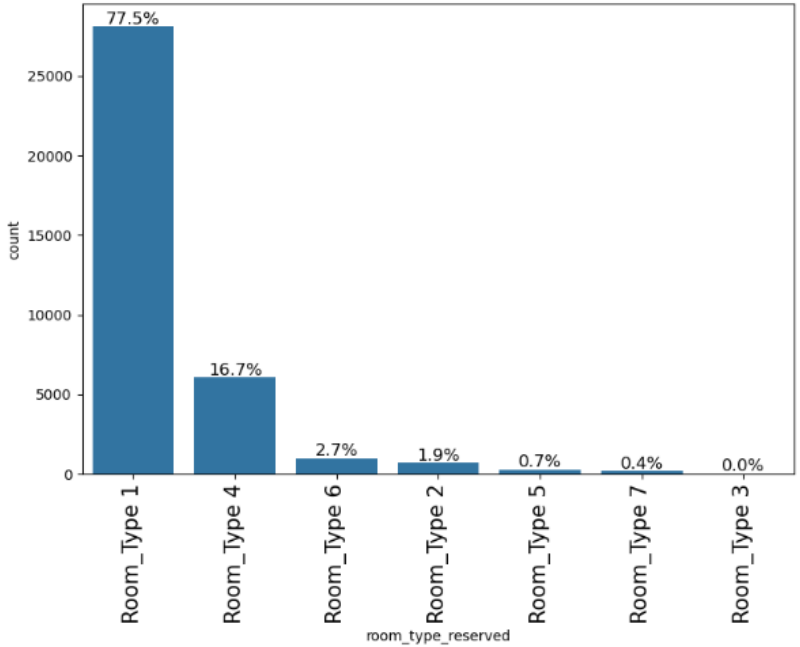


EDA Results - Univariate Analysis

Market Segment



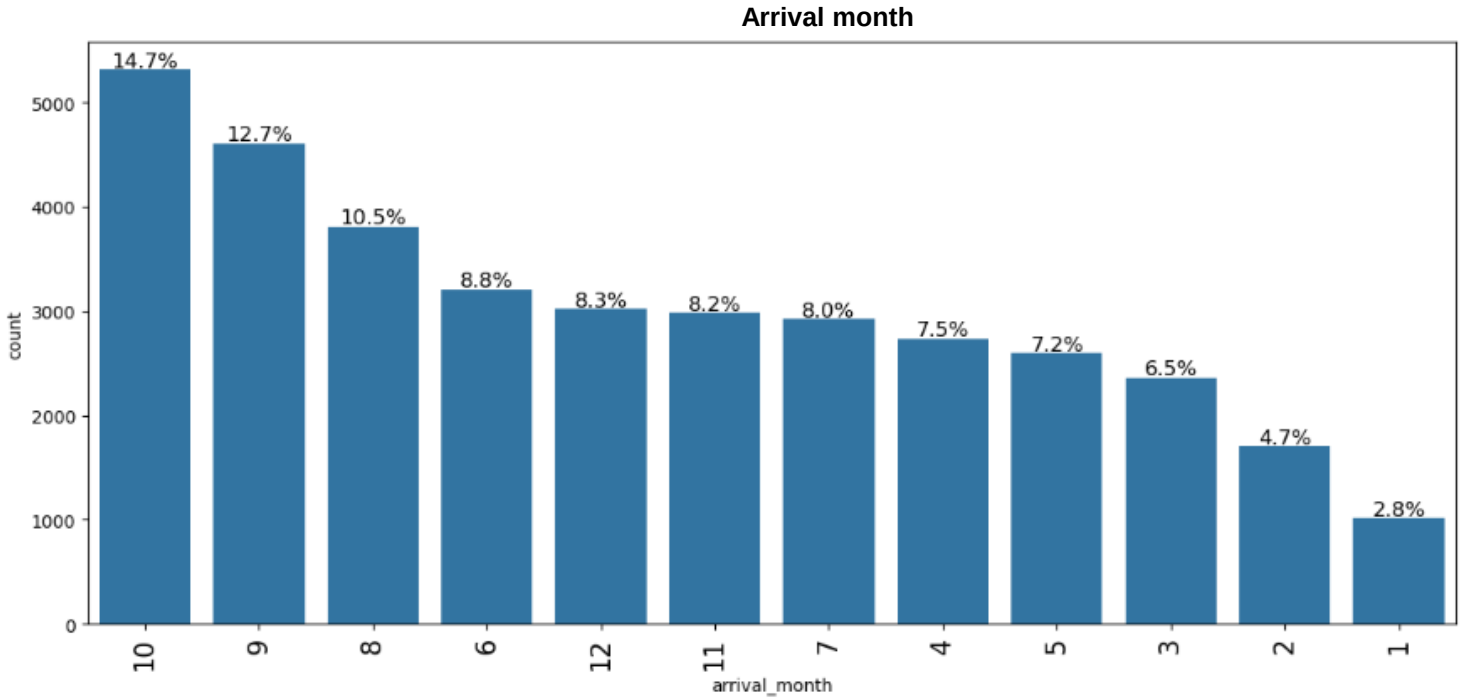
Room Type Reserved



Observation – 78% of the rooms are of one type, which doesn't seem indicative of cancellation in either direction. The same conclusion for the market segment, as 65% of the rooms are ordered online.

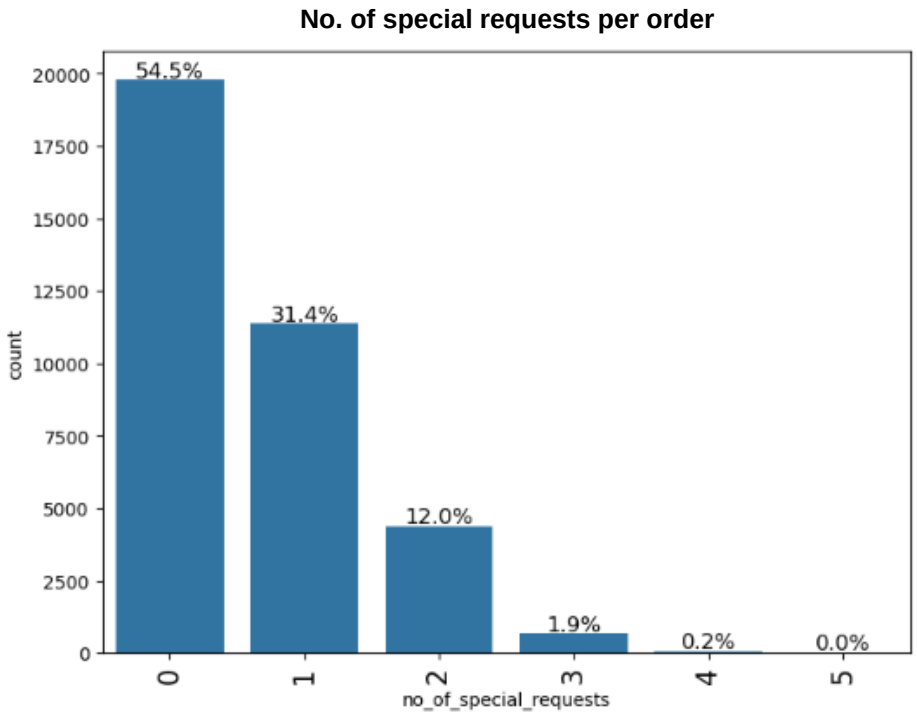
EDA Results - Univariate Analysis

Observation – Interestingly, the last months of the summer and early fall are the most popular for reservations. A further analysis is required to see if there are months that have more cancellations than others.



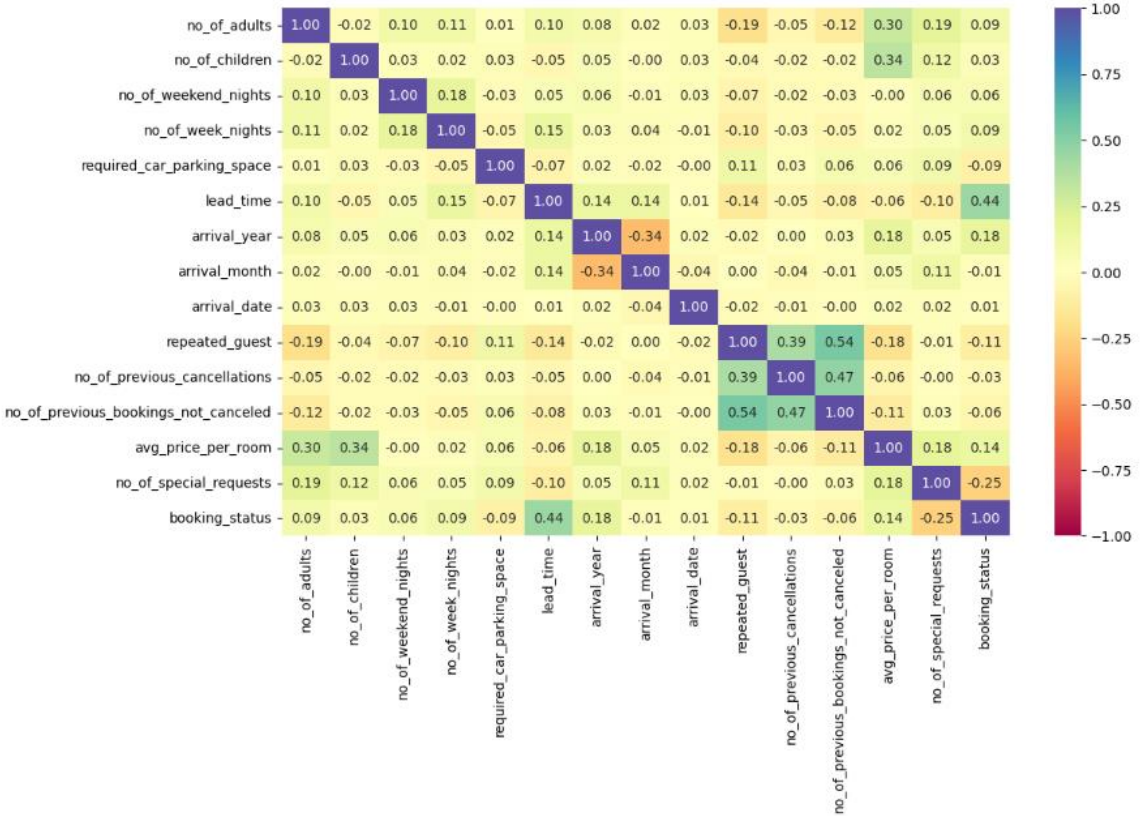
EDA Results - Univariate Analysis

Observation – Around 45% of orders have special requests.
Further analysis is required to see if there is a connection between cancellations and the existence of, or lack of special requests.



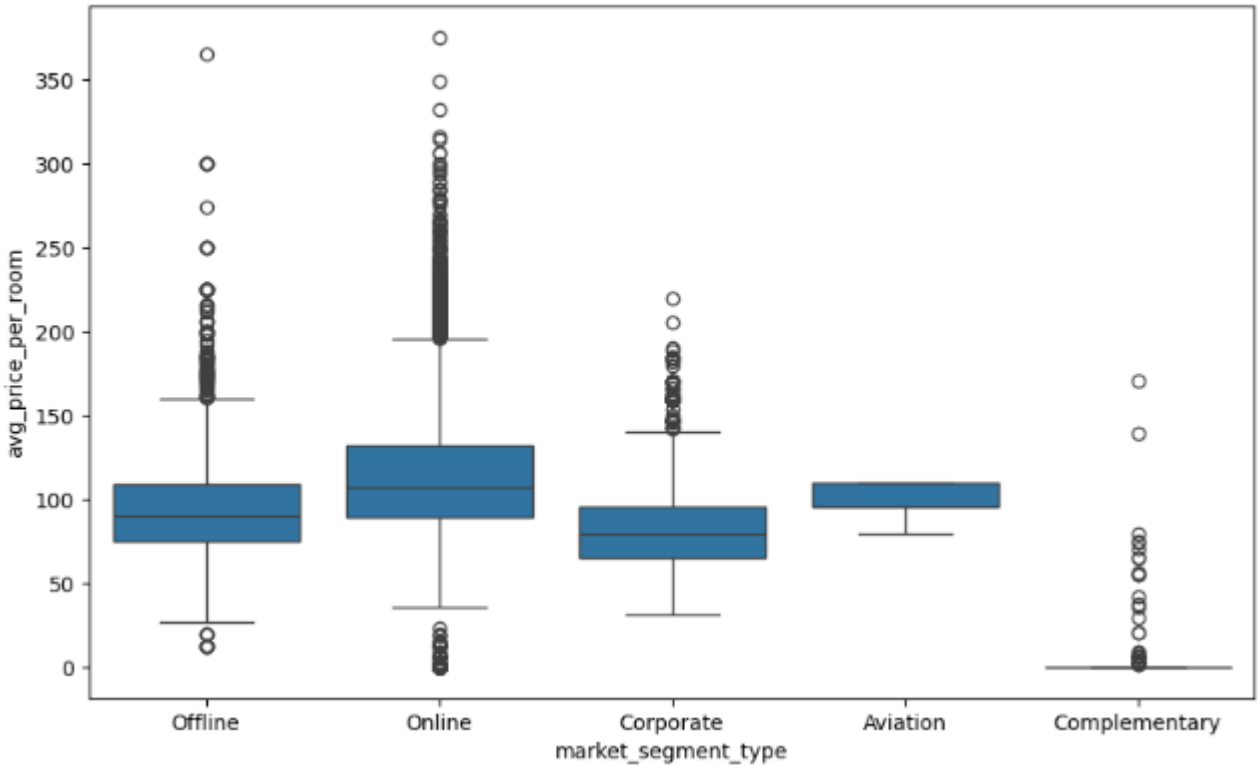
Multivariate Analysis – Heatmap/Correlation

Observation – We see some obvious correlation between repeated guests and the no. of previous reservations, cancelled or not.
There is also 0.44 correlation between lead time and cancellation, which indicates that longer lead time can predict more chance for cancellation.
There is a weak adverse correlation between no. of special requests and cancellation, showing that the more requests that guests have, it is less likely that their order will be cancelled.



Multivariate Analysis – Room Price vs. Market Segment

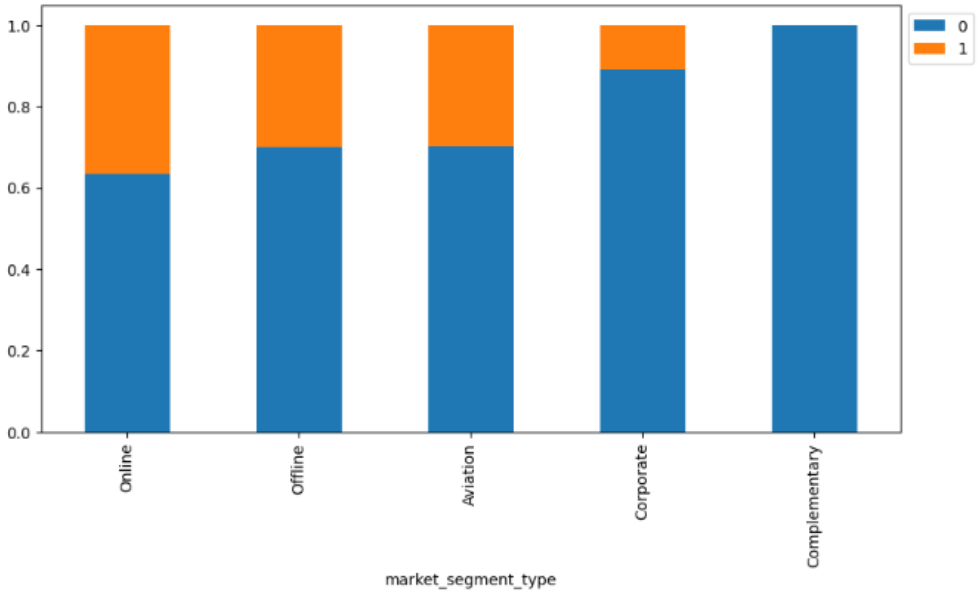
Observation – We can see that Online orders are also the more expensive, which is not surprising in today's world.



Multivariate Analysis – Cancellations vs. Market Segment

Observation – We see that there is not a big difference between the major segments when it comes to cancellations. Corporate orders are less likely to be cancelled, but this is a small segment.

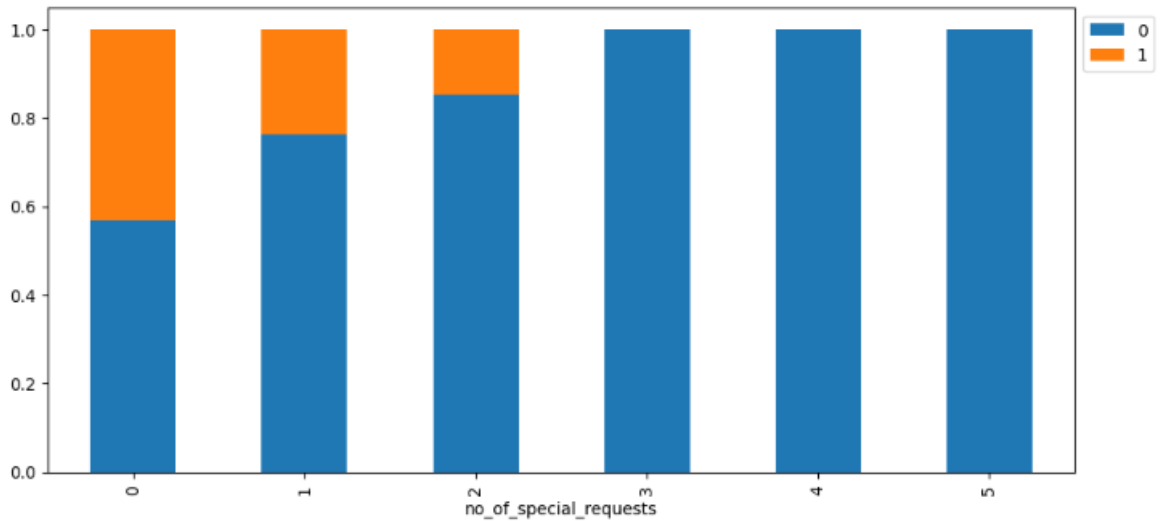
booking_status	0	1	All
market_segment_type			
All	24390	11885	36275
Online	14739	8475	23214
Offline	7375	3153	10528
Corporate	1797	220	2017
Aviation	88	37	125
Complementary	391	0	391



Multivariate Analysis – Cancellations vs. Special Requests

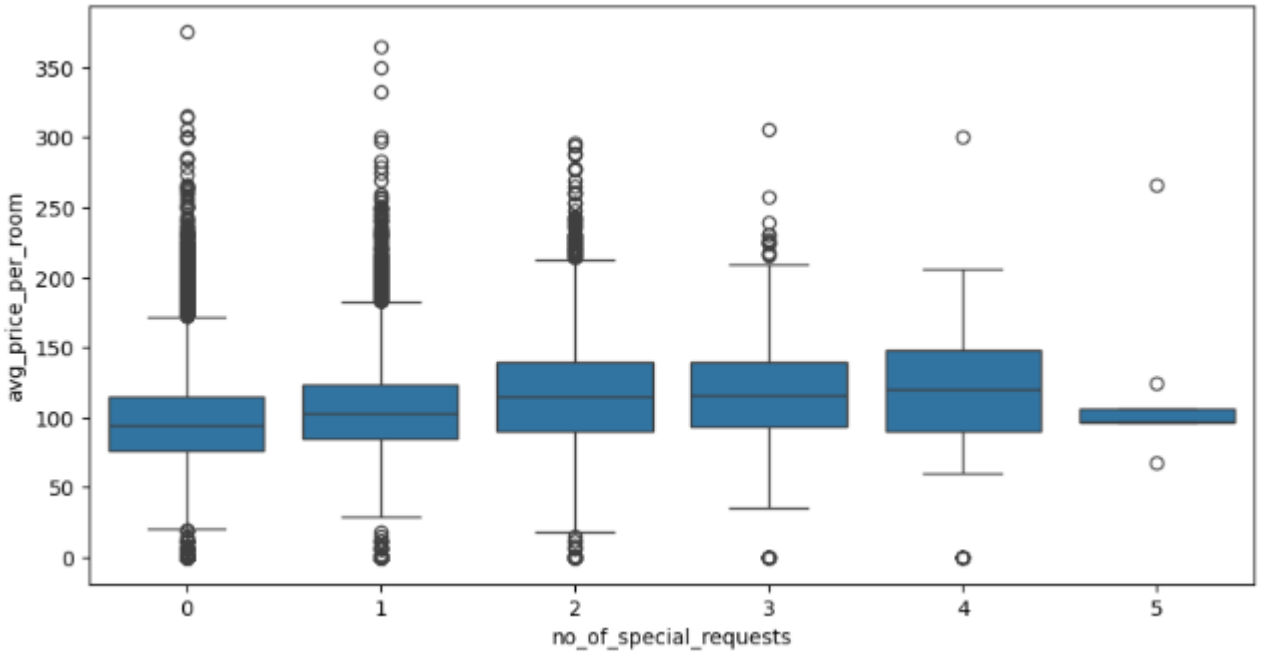
Observation – As we suggested earlier, the more requests people make, the less likely their order will be cancelled.

booking_status	0	1	All
no_of_special_requests			
All	24390	11885	36275
0	11232	8545	19777
1	8670	2703	11373
2	3727	637	4364
3	675	0	675
4	78	0	78
5	8	0	8



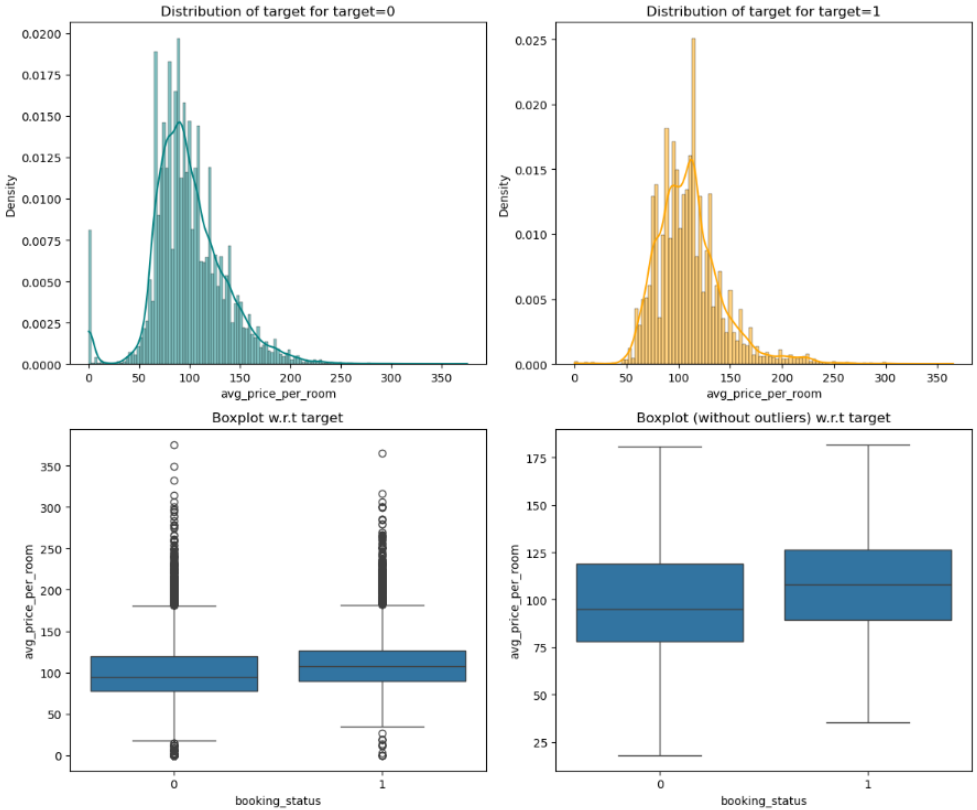
Multivariate Analysis – Room Prices vs. Special Requests

Observation – The higher the room price, the more requests people make. That might be explained by the fact that people expect more, when they pay more.



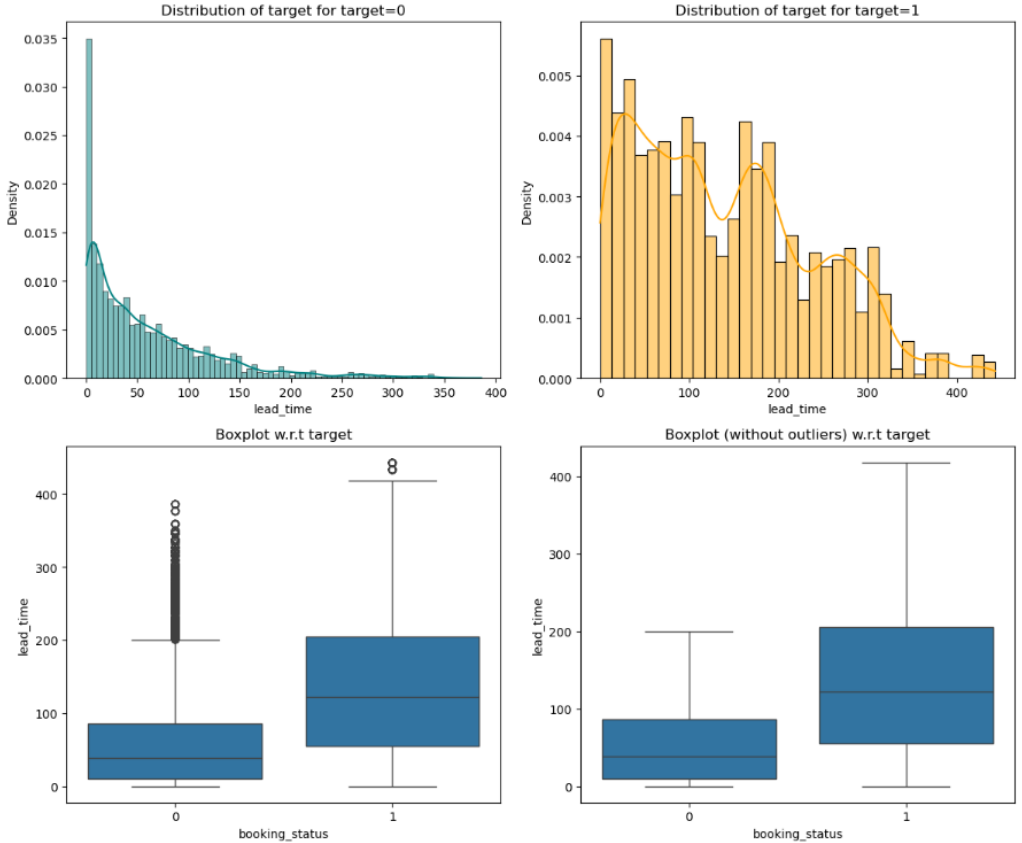
Multivariate Analysis – Cancellations and Room Prices

Observation – The cancellations are distributed normally (for the most part) between the room prices.



Multivariate Analysis – Cancellations and Lead Time

Observation – As we saw in the heat map, lead time plays a bigger part in cancellation rates. The longer the lead time is, it is more likely to see last minute cancellations.

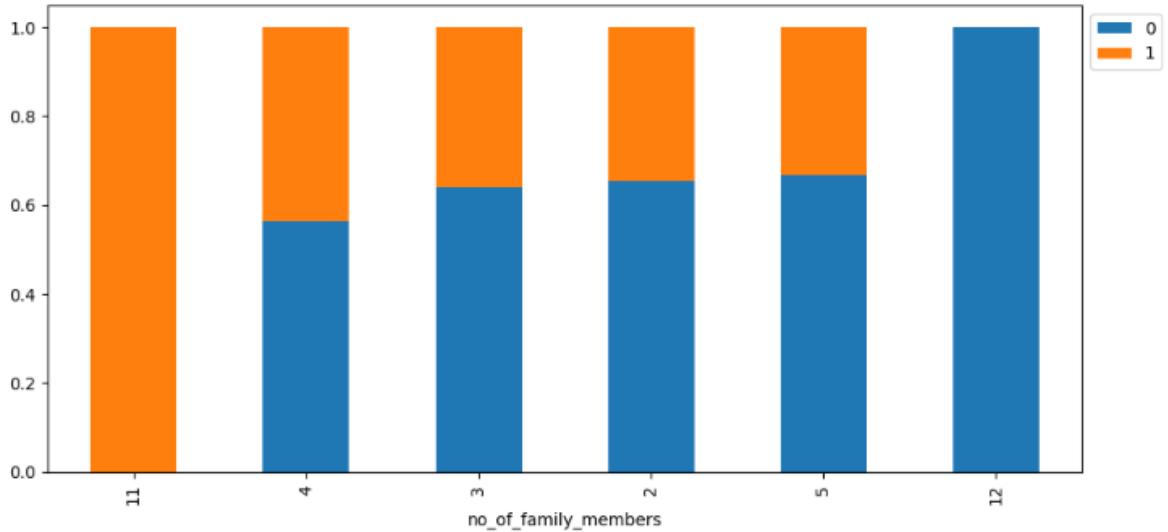


Multivariate Analysis – Cancellations and Party Size

Observation – For most cases, there are the same cancellations regardless of the no. of family members.

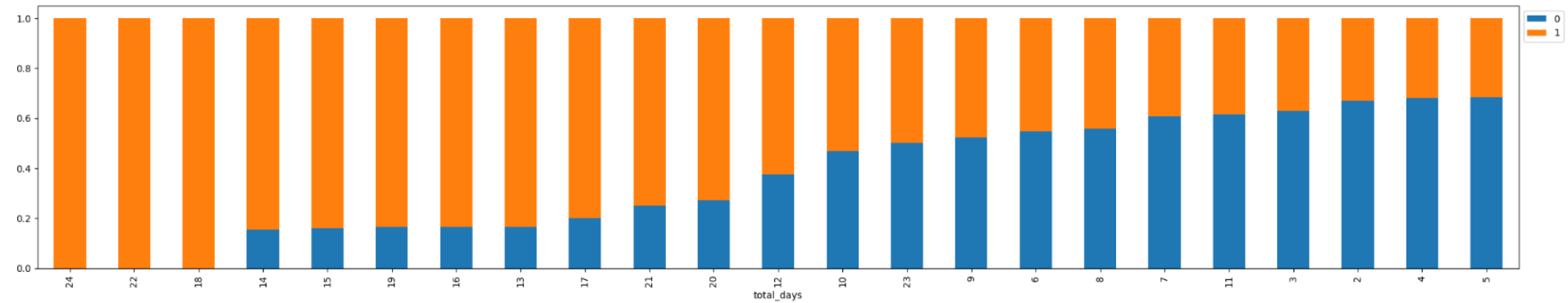
There are 2 cases (out of 36275) where there are 11 and 12 family members, so the fact that the orders was cancelled or not is not significant.

booking_status	0	1	All
no_of_family_members			
All	18456	9985	28441
2	15506	8213	23719
3	2425	1368	3793
4	514	398	912
5	10	5	15
11	0	1	1
12	1	0	1



Multivariate Analysis – Cancellations and Party Size

booking_status	0	1	All
total_days			
All	10979	6115	17094
3	3689	2183	5872
4	2977	1387	4364
5	1593	738	2331
2	1301	639	1940
6	566	465	1031
7	590	383	973
8	100	79	179
10	51	58	109
9	58	53	111
14	5	27	32
15	5	26	31
13	3	15	18
12	9	15	24
11	24	15	39
20	3	8	11
19	1	5	6
16	1	5	6
17	1	4	5
18	0	3	3
21	1	3	4
22	0	2	2
23	1	1	2
24	0	1	1

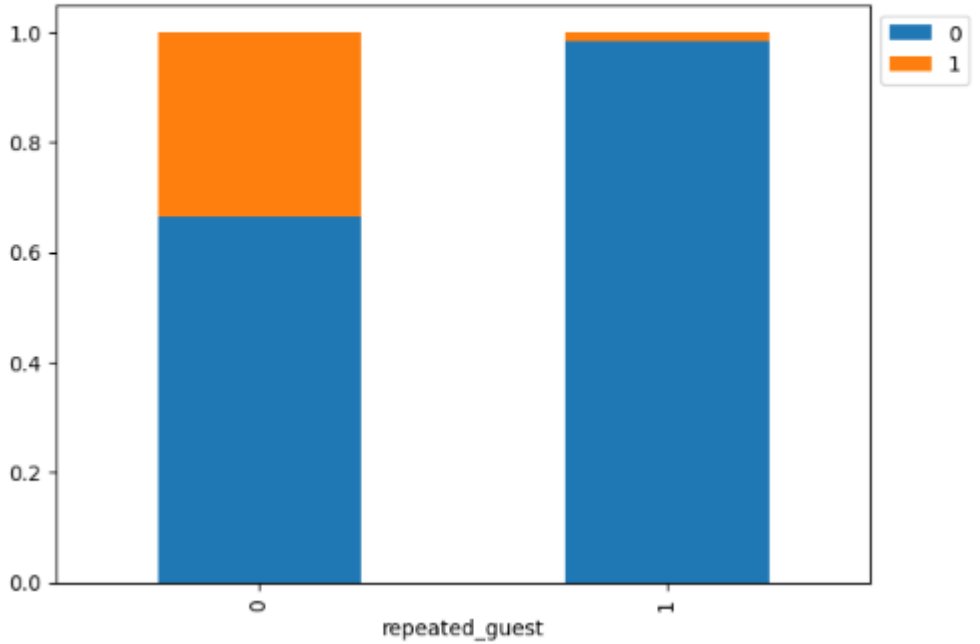


Observation – We can see a very gradual rise of cancellations the longer stays. However, there are much fewer orders with more than 10 days, so it is not relevant for cancellation prediction

Multivariate Analysis – Cancellations and Repeated Guests

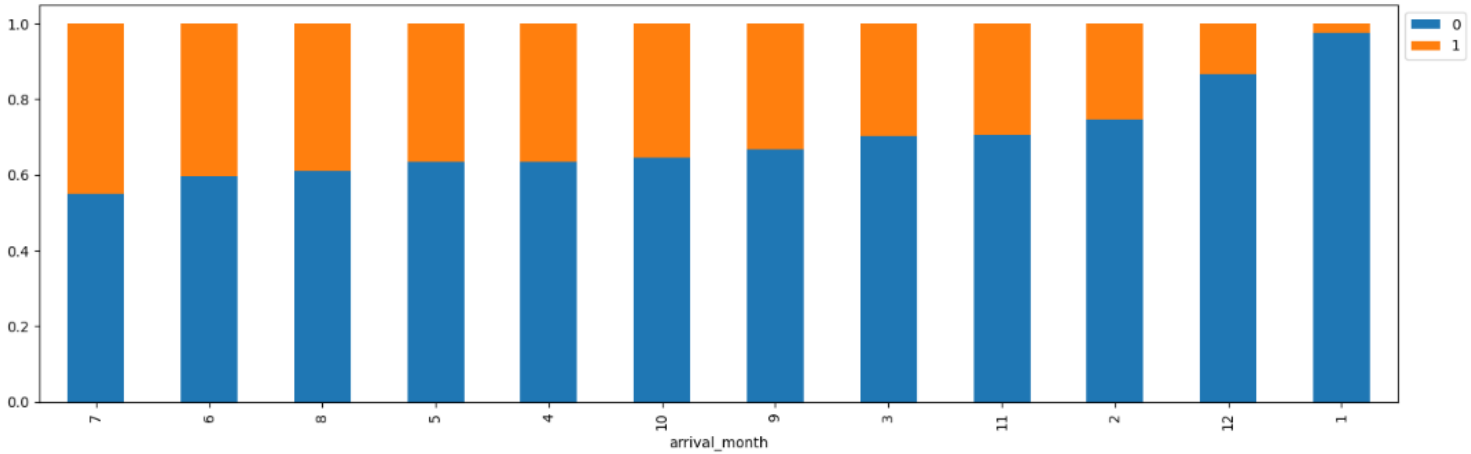
booking_status	0	1	All
repeated_guest			
All	24390	11885	36275
0	23476	11869	35345
1	914	16	930

Observation – We can see that repeated guests tend to have less cancellations, as we saw in the heat map.



Multivariate Analysis – Cancellations and Arrival Months

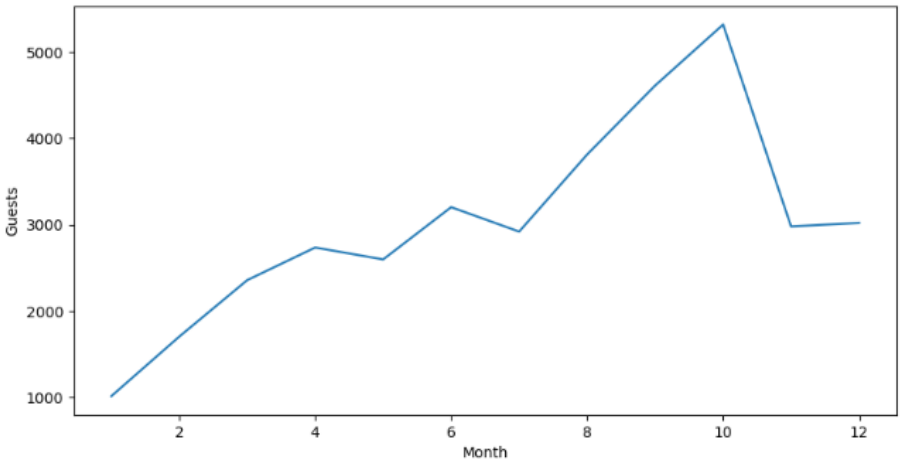
booking_status	0	1	All
arrival_month			
All	24390	11885	36275
10	3437	1880	5317
9	3073	1538	4611
8	2325	1488	3813
7	1606	1314	2920
6	1912	1291	3203
4	1741	995	2736
5	1650	948	2598
11	2105	875	2980
3	1658	700	2358
2	1274	430	1704
12	2619	402	3021
1	990	24	1014



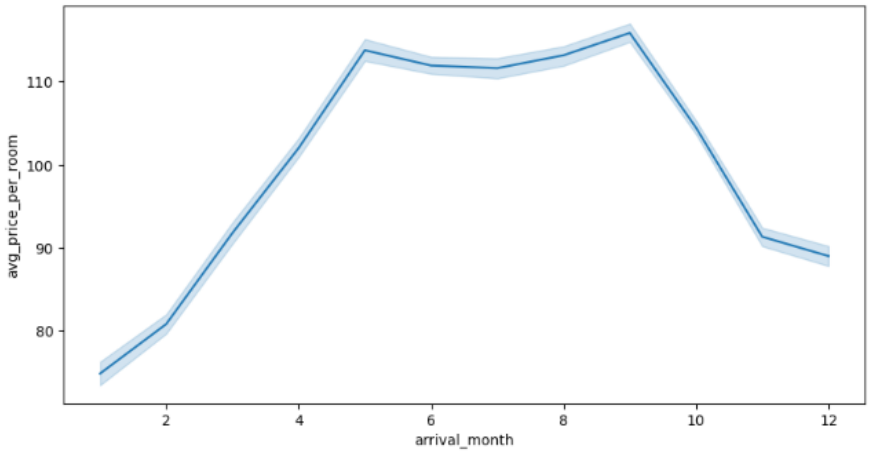
Observation – December and January are the months with the least number of cancellations, and the months with the least amounts of orders. The rest of the months have pretty much the same amounts of cancellations.

Multivariate Analysis – Prices and Arrival Months

No. of guests in each month



Room prices in each month



Observation – The prices are higher in the spring, when the occupancy starts to pick up, all the way to the October, with a little reduction in July, where there are less reservations, compared to spring and fall.

Data Pre-processing – Feature Engineering and Outlier Check

- **Outliers Check** - Most of the variables have outliers, however we will move on with considering them as part of the model

- **Data Types** – All fields are set to be numeric, including Booking Status (Cancelled = 1, Non-Cancelled = 0)

Logistic Regression Model

Model Performance Summary – Logistic Regression

Data Preperation

Create train set (70%) and test set (30%)

Dummy Variables

Create dummy variables

Initial Model Run

Run the model on the training set and check the results

Check Assumptions

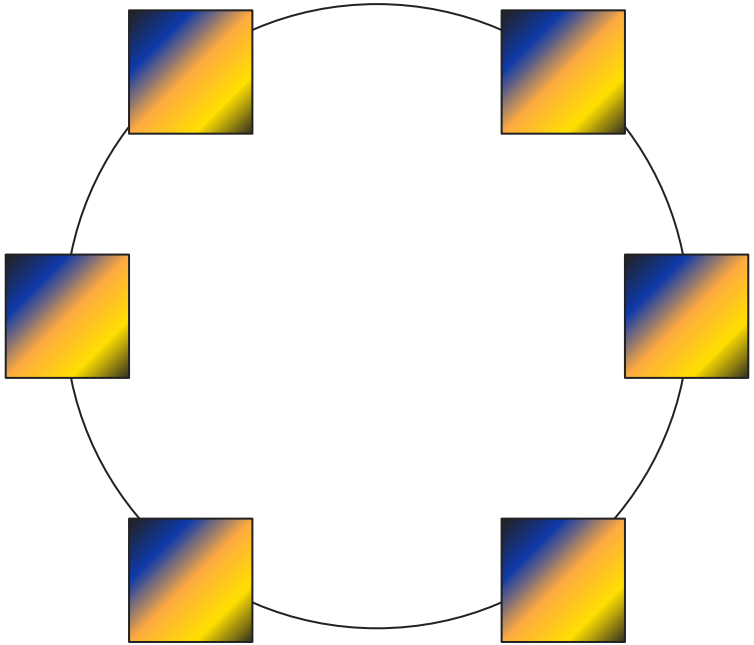
Check VIF > 5 and P values > 0.05 of the variables, adjust thresholds

Compare Thresholds

Find the optimal threshold that should be used to predict the outcome

Final Model Run

Run the final model and compare the results of all variations



Logistic Regression Performance Summary – Data Prep

Train and test sets

```
Shape of Training set :  (25392, 28)
Shape of test set :  (10883, 28)
Percentage of classes in training set:
booking_status
0   0.67064
1   0.32936
Name: proportion, dtype: float64
Percentage of classes in test set:
booking_status
0   0.67638
1   0.32362
```

Creating dummy variables

```
type_of_meal_plan_Meal Plan 2
type_of_meal_plan_Meal Plan 3
type_of_meal_plan_Not Selected
room_type_reserved_Room_Type 2
room_type_reserved_Room_Type 3
room_type_reserved_Room_Type 4
room_type_reserved_Room_Type 5
room_type_reserved_Room_Type 6
room_type_reserved_Room_Type 7
market_segment_type_Complementary
market_segment_type_Corporate
market_segment_type_Offline
market_segment_type_Online
```

Logistic Regression Performance Summary – Initial Run

	feature	VIF
0	const	39491186.47744
1	no_of_adults	1.34849
2	no_of_children	1.97862
3	no_of_weekend_nights	1.06949
4	no_of_week_nights	1.09567
5	required_car_parking_space	1.03998
6	lead_time	1.39518
7	arrival_year	1.43167
8	arrival_month	1.27637
9	arrival_date	1.00674
10	repeated_guest	1.78361
11	no_of_previous_cancellations	1.39569
12	no_of_previous_bookings_not_canceled	1.65200
13	avg_price_per_room	2.06421
14	no_of_special_requests	1.24730
15	type_of_meal_plan_Meal Plan 2	1.27325
16	type_of_meal_plan_Meal Plan 3	1.02522
17	type_of_meal_plan_Not Selected	1.27252
18	room_type_reserved_Room_Type 2	1.10151
19	room_type_reserved_Room_Type 3	1.00330
20	room_type_reserved_Room_Type 4	1.36261
21	room_type_reserved_Room_Type 5	1.02797
22	room_type_reserved_Room_Type 6	1.97490
23	room_type_reserved_Room_Type 7	1.11559
24	market_segment_type_Complementary	4.50229
25	market_segment_type_Corporate	16.92846
26	market_segment_type_Offline	64.11425
27	market_segment_type_Online	71.17686

- **Checking VIF for Multicollinearity**

Results – Only dummy variables have VIF > 5, therefore we will keep all existing fields in the X set

- **Removing High P-Values**

The following fields will be removed from the X set due to high P-values:

arrival_date, no_of_previous_booking_not_cancelled, type_of_meal_plan_Meal_Plan_3, Room_type_reserved_Room_Type_3, market_segment_type_Complementary, market_segment_type_Online

Logistic Regression Second Run – Odds from Coefficiency

Column	Odds	Change_odd%
no_of_adults	1.11487	11.48701
no_of_children	1.16411	16.41072
no_of_weekend_nights	1.11466	11.46616
no_of_week_nights	1.0426	4.25996
required_car_parking_space	0.20298	-79.7024
lead_time	1.01583	1.58342
arrival_year	1.57291	57.29054
arrival_month	0.95841	-4.15858
repeated_guest	0.06484	-93.5161
no_of_previous_cancellations	1.25705	25.7046
avg_price_per_room	1.01937	1.93739
no_of_special_requests	0.22993	-77.0067
type_of_meal_plan_Meal Plan 2	1.17855	17.85489
type_of_meal_plan_Not Selected	1.33103	33.10347
room_type_reserved_Room_Type 2	0.70069	-29.9311
room_type_reserved_Room_Type 4	0.75327	-24.6735
room_type_reserved_Room_Type 5	0.47893	-52.1065
room_type_reserved_Room_Type 6	0.38034	-61.9663
room_type_reserved_Room_Type 7	0.2384	-76.16
market_segment_type_Corporate	0.45262	-54.738
market_segment_type_Offline	0.16771	-83.229

Converting coefficients to odds:

The coefficients of the logistic regression model are in terms of log(odd), to find the odds we must take the exponential of the coefficients. Therefore:

odds = exp(b)

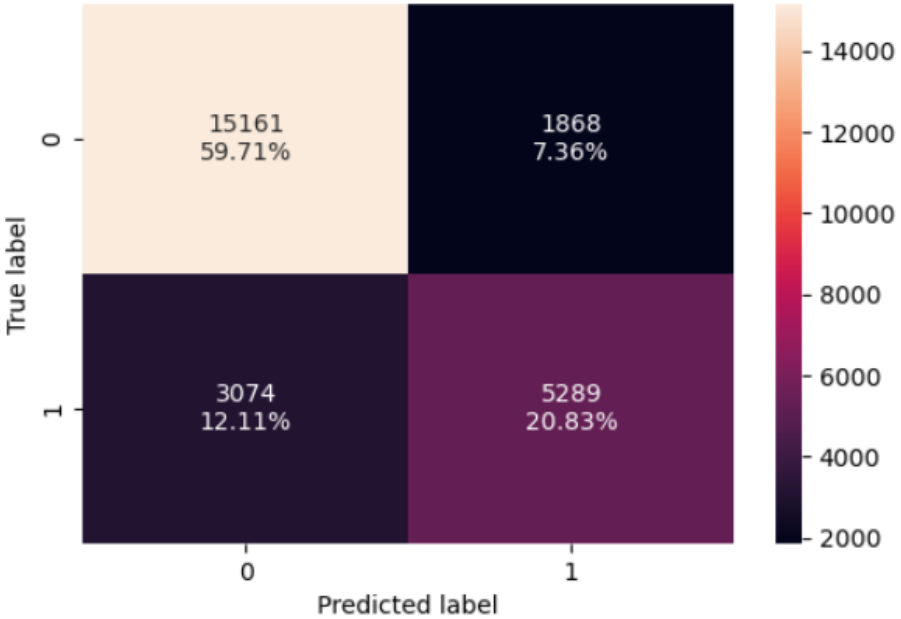
The percentage change in odds is given as:

odds = (exp(b) - 1) * 100

Explanation:
Holding all other features constant, a unit change in a feature will increase the odds of a cancellation by 'Odd' value times or a 'change_odd%' value increase in the odds of having a cancellation.

Logistic Regression Performance Summary – Training Set

Confusion Matrix with threshold of 0.5



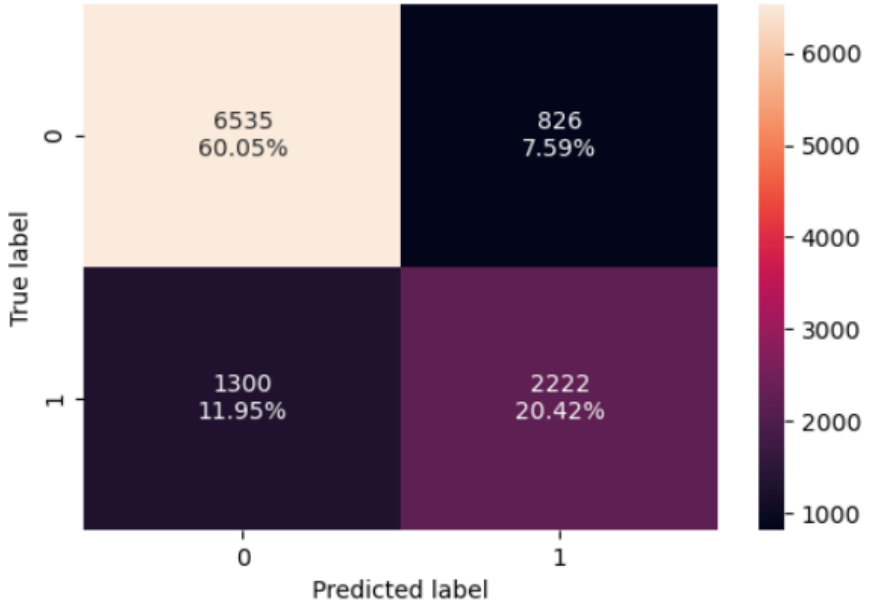
Training performance:

	Accuracy	Recall	Precision	F1
0	0.80537	0.63243	0.73900	0.68157

Results – The evaluations methods for performance give us F1 score of 68%, which means that in 68% of the time, the model predictions matched the actual result, if a room is cancelled or not

Logistic Regression Performance Summary – Test Set

Confusion Matrix with threshold of 0.5



Test performance:

	Accuracy	Recall	Precision	F1
0	0.80465	0.63089	0.72900	0.67641

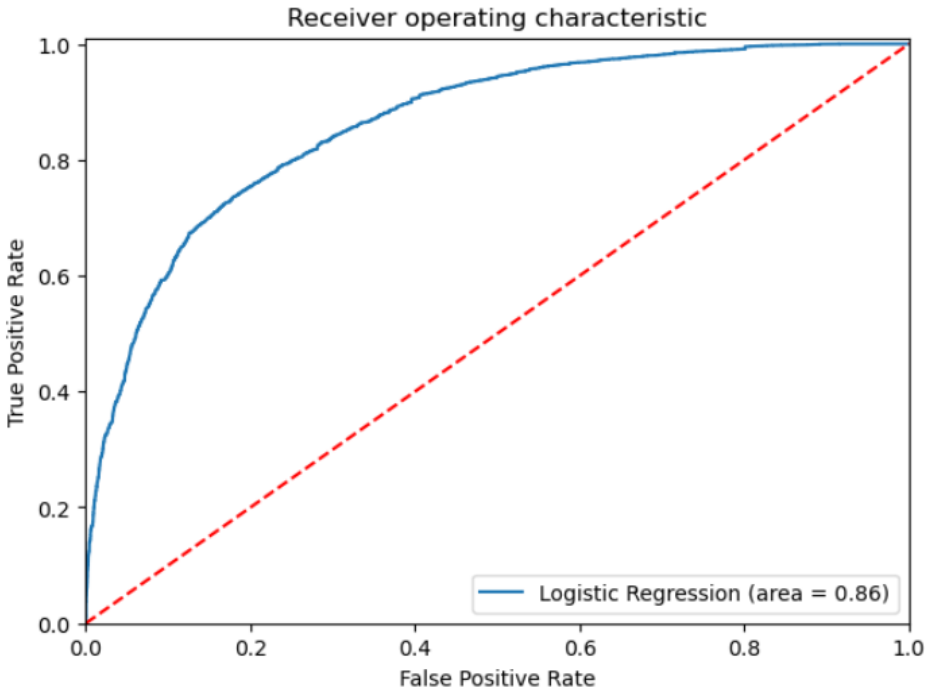
Results – The evaluations methods for performance give us F1 score of 67%, which is not far from the training performance. This means that there is no overfitting in this model

Logistic Regression Performance - ROC-AUC

Optimal threshold as per AUC-ROC curve:

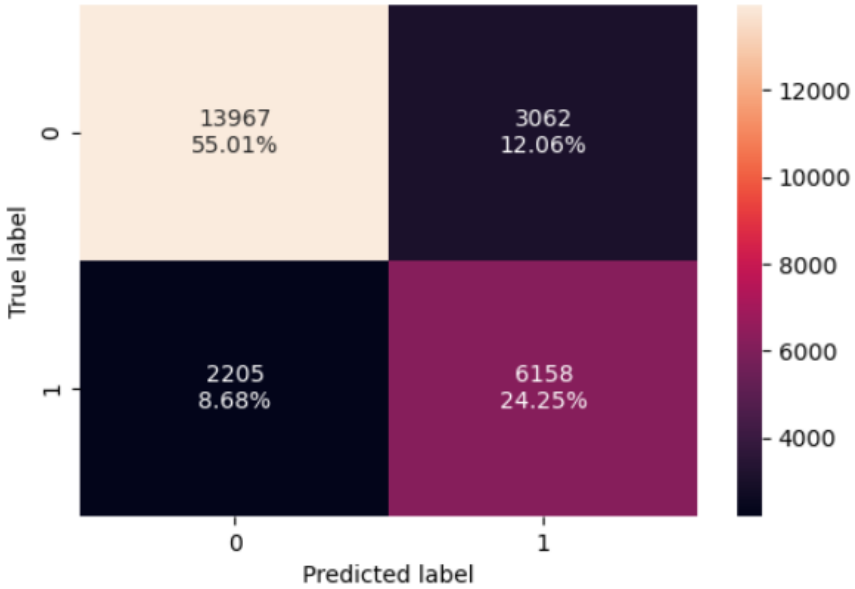
The optimal Threshold would be where TPR (True Positive Rate) is high and FPR (False Positive Rate) is low

The optimal threshold is - **0.3696037915893037**



Logistic Regression Performance Summary – Training Set

Confusion Matrix with threshold of 0.3696037915893037



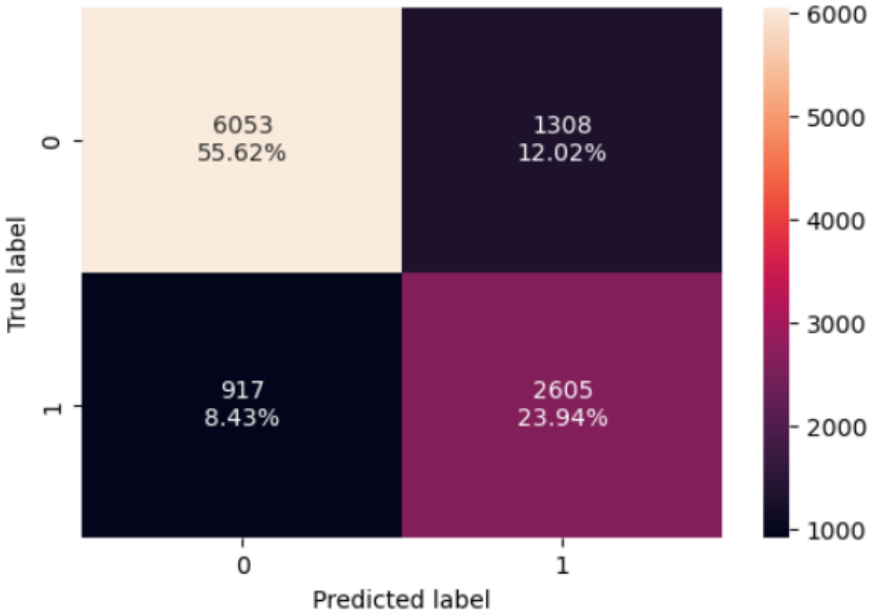
Training performance:

	Accuracy	Recall	Precision	F1
0	0.79257	0.73634	0.66790	0.70045

Results – The evaluations methods for performance give us F1 score of 70%, which is better than the initial threshold (68%)

Logistic Regression Performance Summary – Test Set

Confusion Matrix with threshold of 0.37



Test performance:

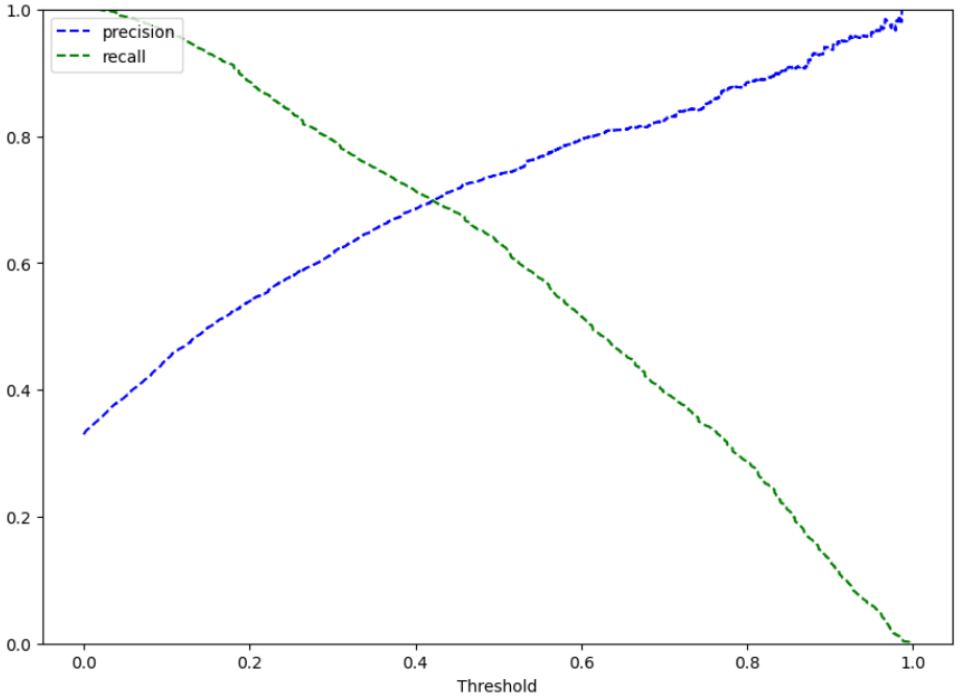
	Accuracy	Recall	Precision	F1
0	0.79555	0.73964	0.66573	0.70074

Results – The evaluations methods for performance give us F1 score of 70%, which is almost the same as in the training set

Logistic Regression Performance - Precision-Recall curve

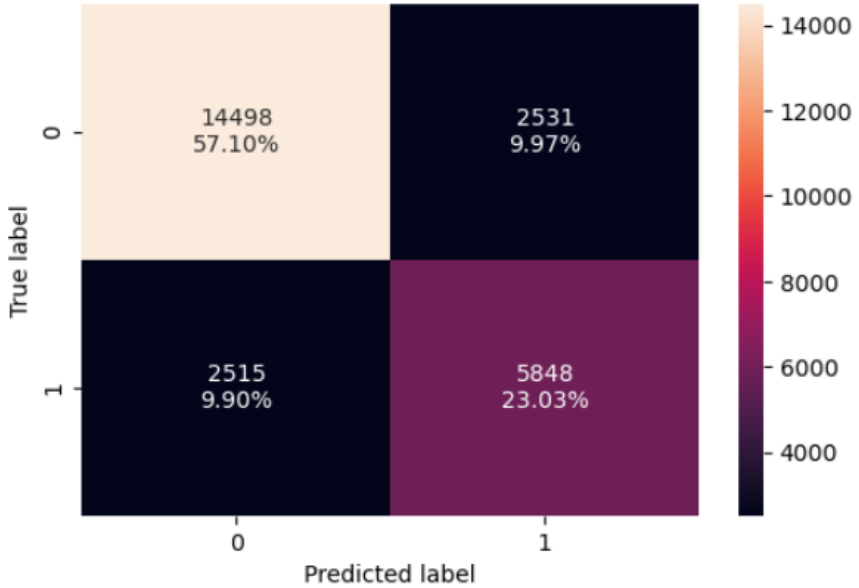
Optimal threshold as per Precision-Recall curve:

The optimal threshold is - **0.42**



Logistic Regression Performance Summary – Training Set

Confusion Matrix with threshold of 0.42



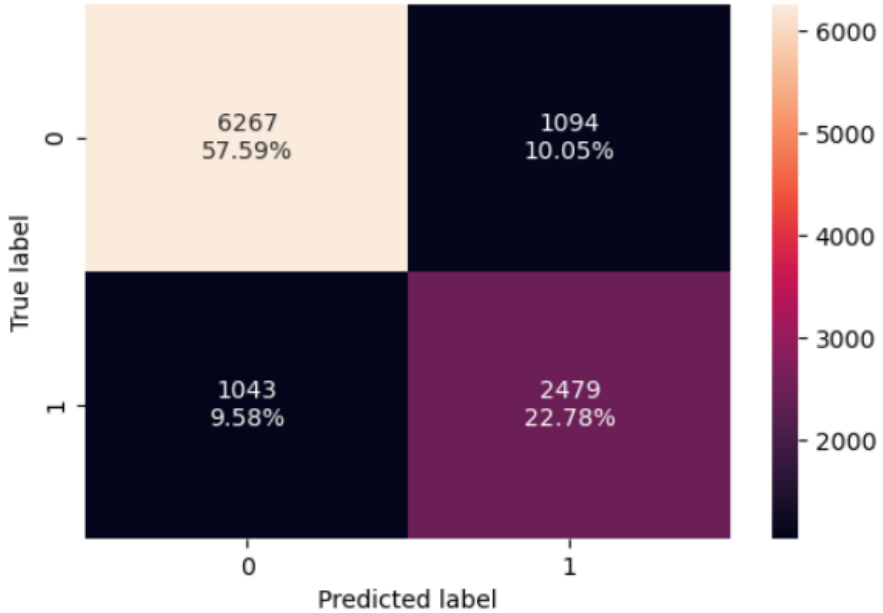
Training performance:

	Accuracy	Recall	Precision	F1
0	0.80128	0.69927	0.69794	0.69860

Results – The evaluations methods for performance give us F1 score of 69%, which is still better than the initial threshold (68%)

Logistic Regression Performance Summary – Test Set

Confusion Matrix with threshold of 0.42



Test performance:

	Accuracy	Recall	Precision	F1
0	0.80364	0.70386	0.69381	0.69880

Results – The evaluations methods for performance give us F1 score of 69.88%, which is slightly better than the training set

Logistic Regression Performance Summary

Training performance comparison:

	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80537	0.79257	0.80128
Recall	0.63243	0.73634	0.69927
Precision	0.73900	0.66790	0.69794
F1	0.68157	0.70045	0.69860

Testing performance comparison:

	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80465	0.79555	0.80364
Recall	0.63089	0.73964	0.70386
Precision	0.72900	0.66573	0.69381
F1	0.67641	0.70074	0.69880

Results – This model gives us slightly different results with thresholds between 0.37 – 0.5
In all thresholds, it is very similar between training and test sets, which means it's not overfitting.
Based on the F1 results, we will use the 0.42 threshold to predict the cancellation of a hotel room.

Decision Tree Model

Model Performance Summary – Decision Tree

Final Model Run

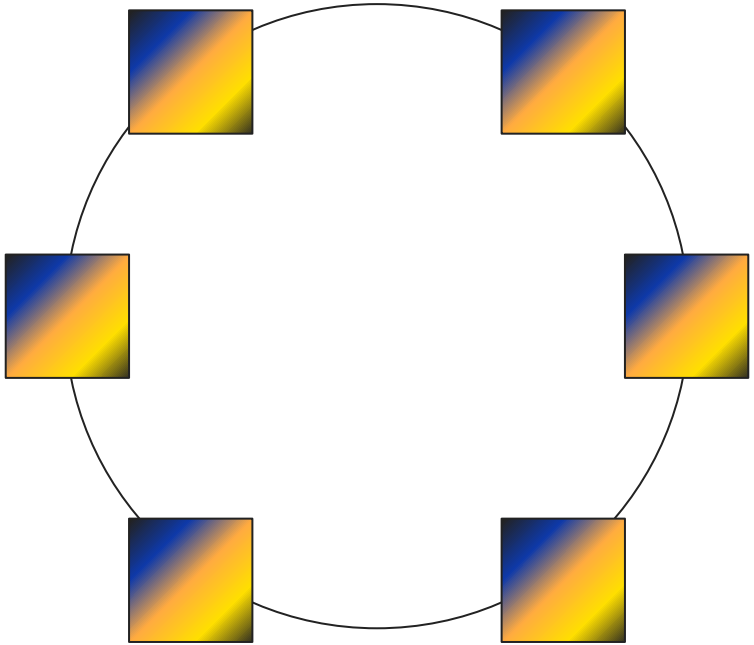
Run the final model and compare the results of all variations

Post-Pruning

Perform post pruning optimization and re-run the model

Pre-Pruning

Perform post-pruning and re-run the model



Data Preperation

Create train set (70%) and test set (30%)

Dummy Variables

Create dummy variables

Initial Model Run

Run the model on the training set and check the results

Decision Tree Performance Summary – Data Prep

Train and test sets

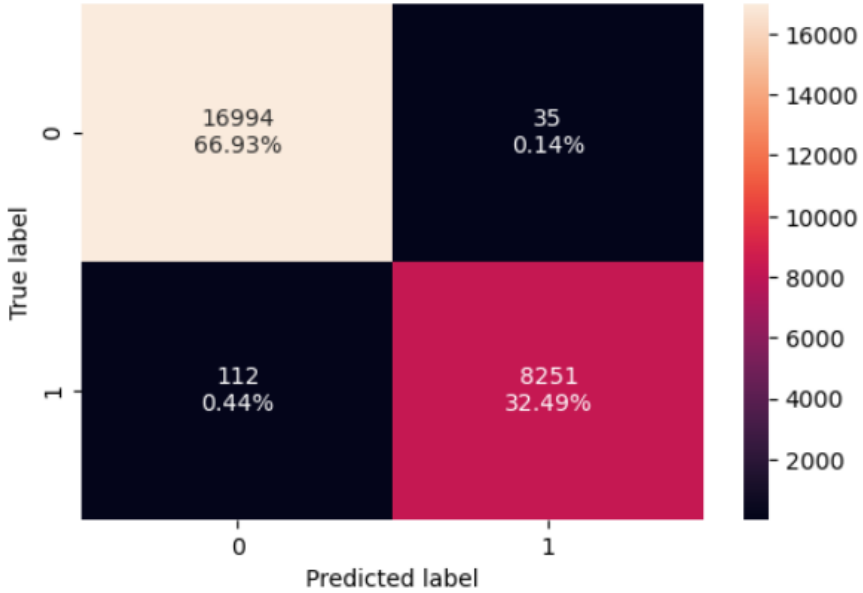
```
Shape of Training set :  (25392, 28)
Shape of test set :  (10883, 28)
Percentage of classes in training set:
booking_status
0   0.67064
1   0.32936
Name: proportion, dtype: float64
Percentage of classes in test set:
booking_status
0   0.67638
1   0.32362
```

Creating dummy variables

```
type_of_meal_plan_Meal Plan 2
type_of_meal_plan_Meal Plan 3
type_of_meal_plan_Not Selected
room_type_reserved_Room_Type 2
room_type_reserved_Room_Type 3
room_type_reserved_Room_Type 4
room_type_reserved_Room_Type 5
room_type_reserved_Room_Type 6
room_type_reserved_Room_Type 7
market_segment_type_Complementary
market_segment_type_Corporate
market_segment_type_Offline
market_segment_type_Online
```

Decision Tree Performance Summary – Training Set

Default Setting (random_state=1)

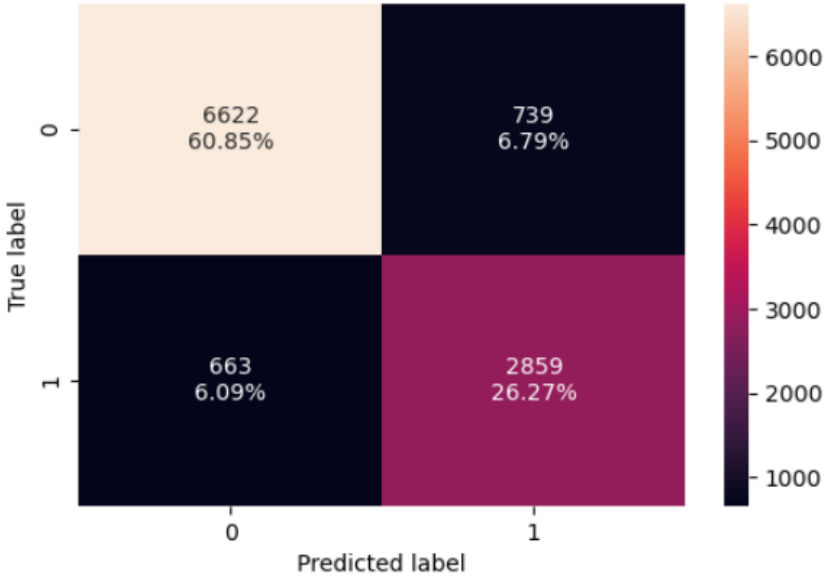


	Accuracy	Recall	Precision	F1
0	0.99421	0.98661	0.99578	0.99117

Results – As expected from a decision tree model, the results are very good for the training set for all types of tests, as we didn't limit (prune) the tree in any way.

Decision Tree Performance Summary – Test Set

Default Setting



	Accuracy	Recall	Precision	F1
0	0.87118	0.81175	0.79461	0.80309

Results – The default setting give us pretty good results even for the test set. However, there is a 20% difference that suggest overfitting of this version of the model.

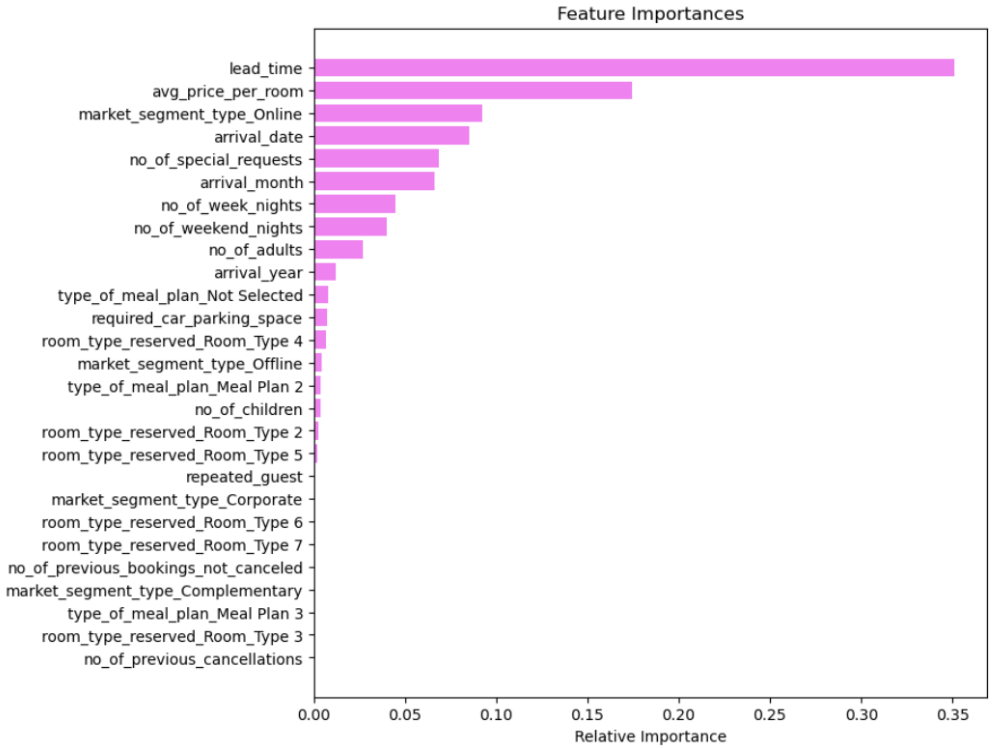
Decision Tree Model – Feature Importance

Observation:

The most important feature according to this model is the lead time.

It means that the longer the reservation is made in advance, the more likely it will be cancelled.

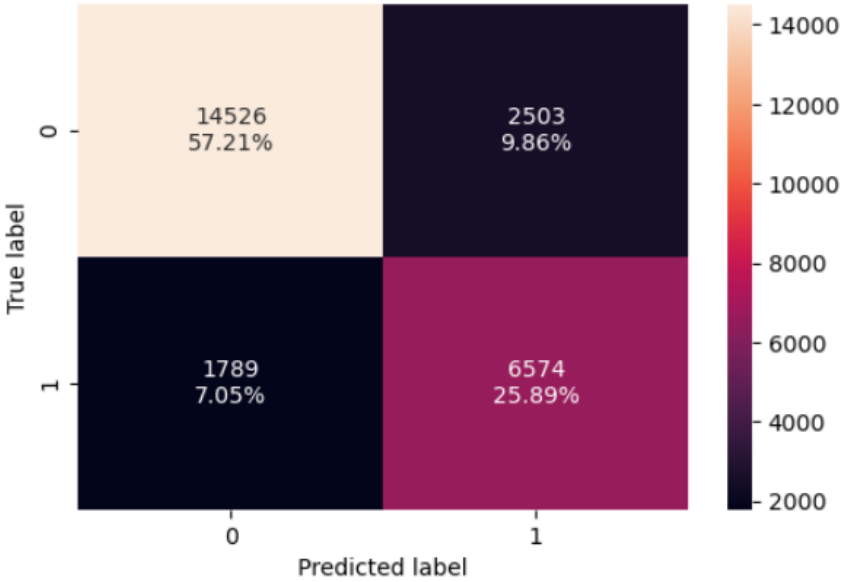
Other important features are the room price and whether it is an online order.



Decision Tree – Pre-Pruning

Decision Tree Performance Summary – Training Set – Pre-Pruning

(max_depth=6, max_leaf_nodes=50, min_samples_split=10, random_state=1)



	Accuracy	Recall	Precision	F1
0	0.83097	0.78608	0.72425	0.75390

Using GridSearch for Hyperparameter tuning:

Grid search is a tuning technique that attempts to compute the optimum values of hyperparameters.

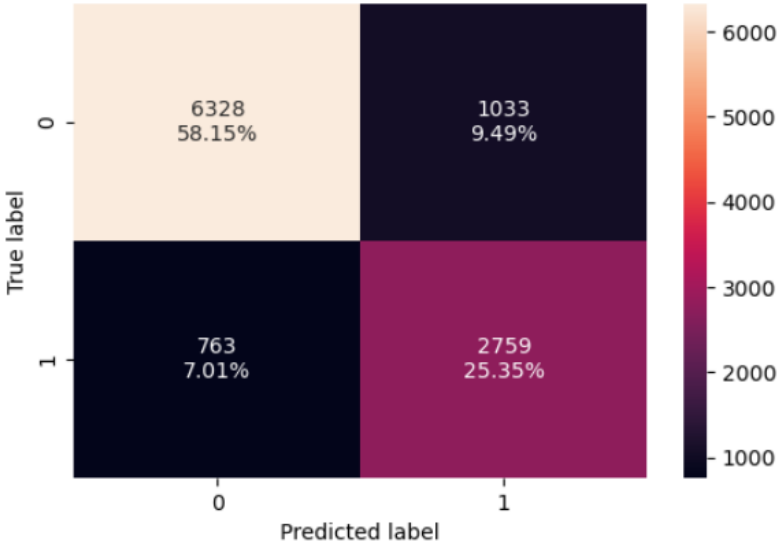
The options we give to the grid search are:

"max_depth": [2, 4, 6],
"max_leaf_nodes": [50, 75, 150, 250],
"min_samples_split": [10, 30, 50, 70]

Results – After limiting the growth of the tree before running the model (pre-pruning) the results are not as good as before, but 75% for F1 is not a bad performance

Decision Tree Performance Summary – Test Set – Pre-Pruning

(max_depth=6, max_leaf_nodes=50, min_samples_split=10, random_state=1)



	Accuracy	Recall	Precision	F1
0	0.83497	0.78336	0.72758	0.75444

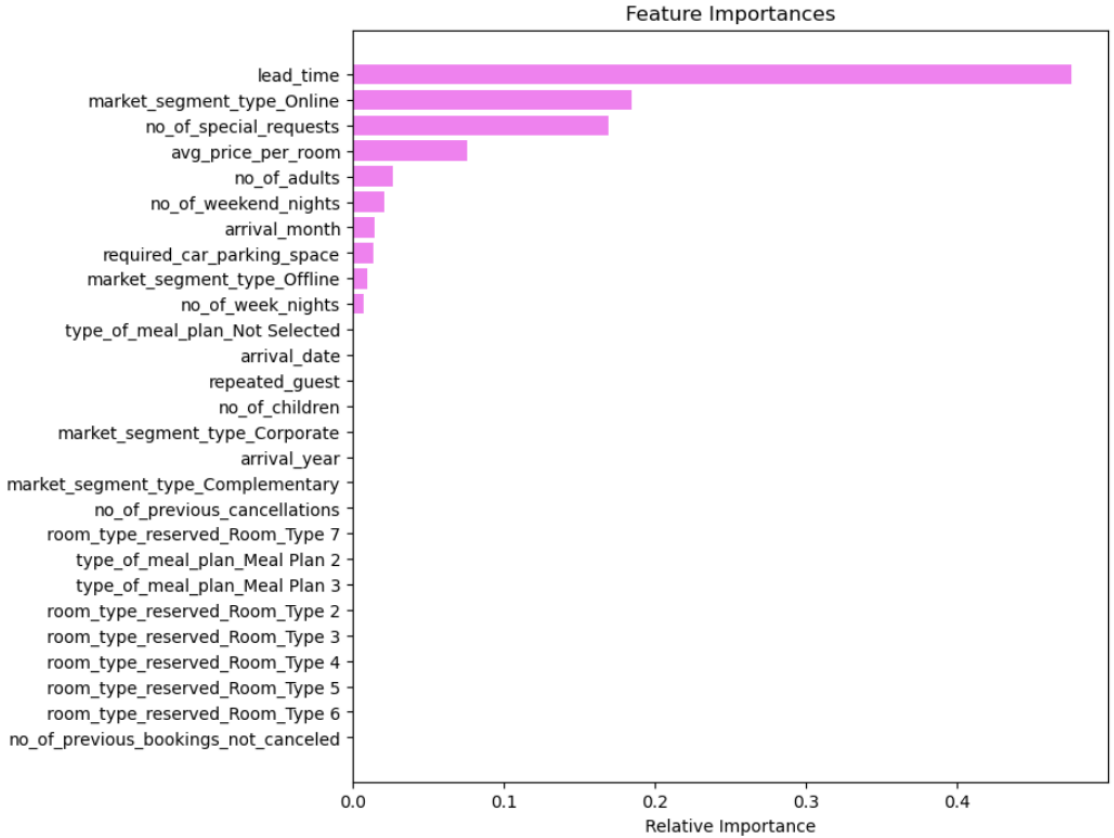
Results – When running the model on the test set, we see that F1 score is like the training results, around 75%, which indicates that the model is not overfitting.

Decision Tree Model – Feature Importance – Pre-Pruning

Observation:

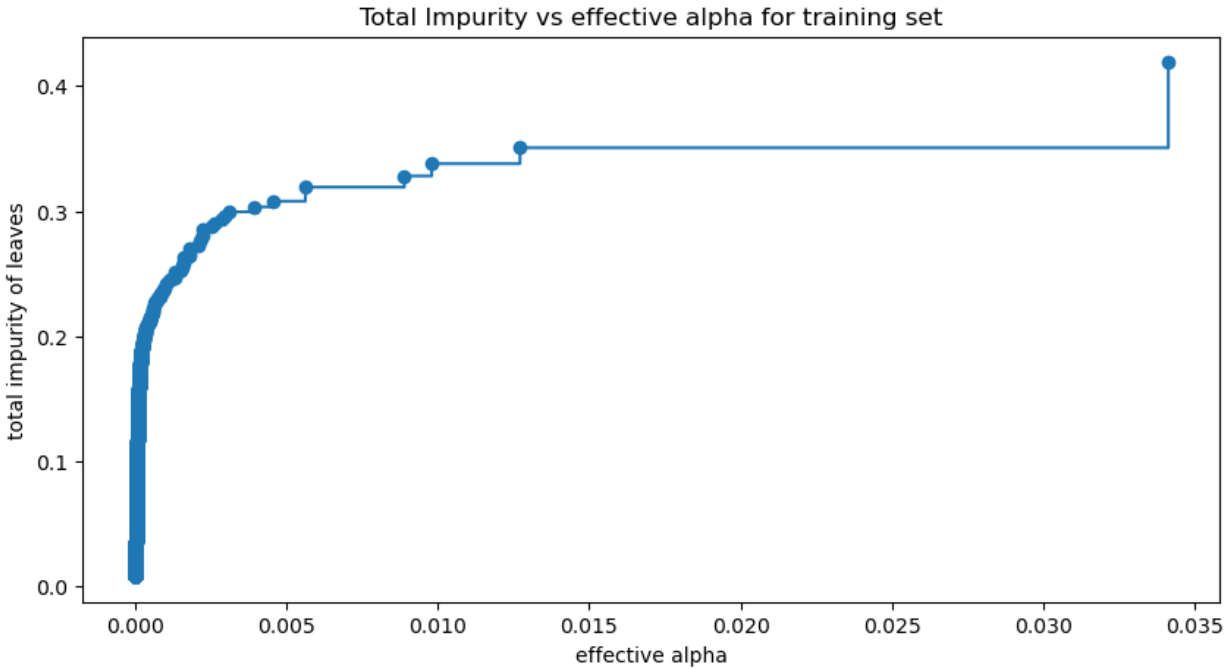
By pre-pruning we narrow the list of parameters that play a big role in the decision tree.

The main features are lead time, online orders, special requests and room price.



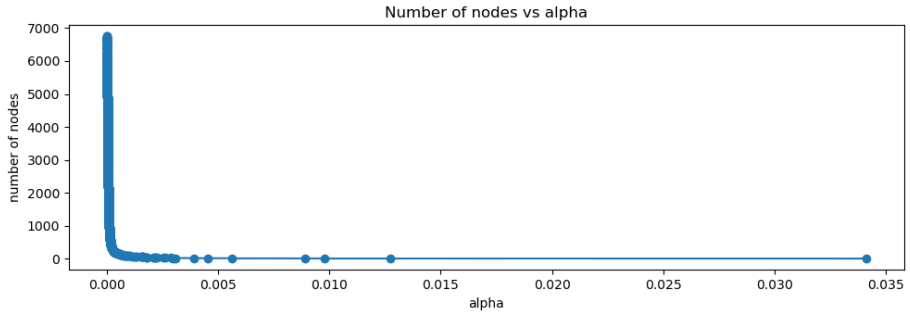
Decision Tree – Post-Pruning

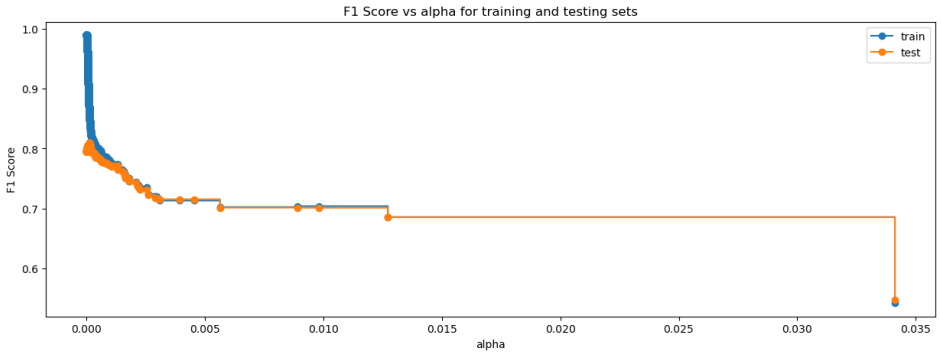
Decision Tree Performance – Post-Pruning Preparation

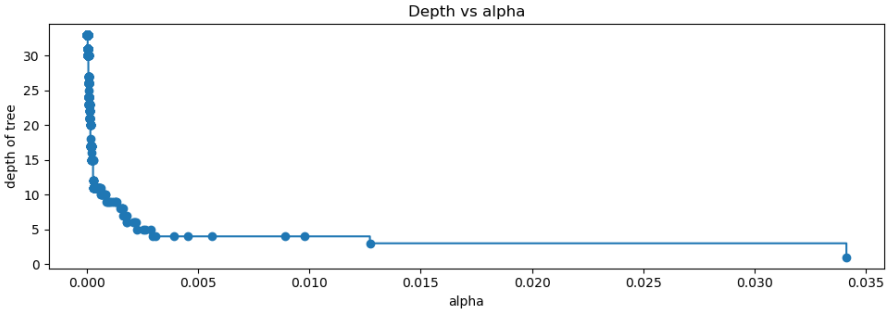


Minimal cost complexity pruning recursively finds the node with the "weakest link". The weakest link is characterized by an effective alpha, where the nodes with the smallest effective alpha are pruned first. We found the effective alphas and the corresponding total leaf impurities at each step of the pruning process. As alpha increases, more of the tree is pruned, which increases the total impurity of its leaves.

Decision Tree Performance – Post-Pruning Preparation



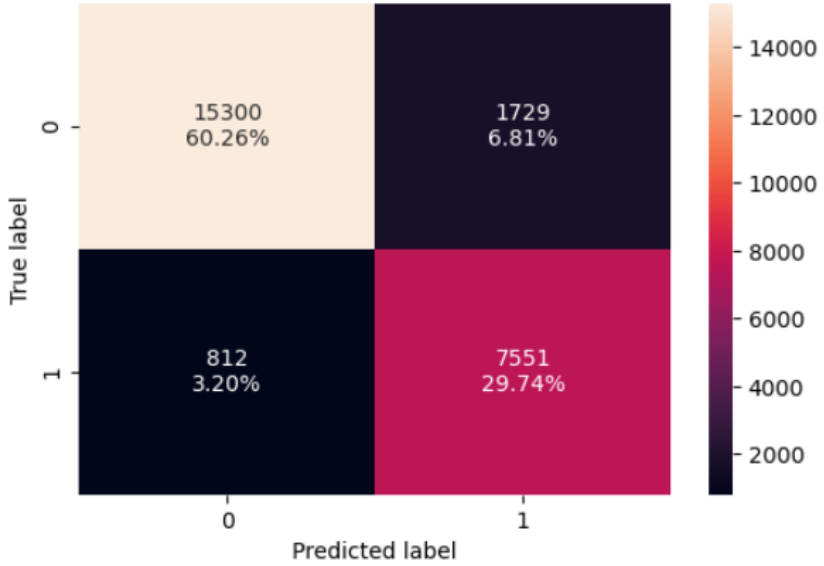




We are finding the balance between pruning the nodes and limiting the depth of the tree, according to the alpha score.
Then, by comparing the F1 scores of the training and test sets, we find the alpha value that will make the model not overfitting (F1 score around 80%, alpha= 0.00012267633155167048)

Decision Tree Performance Summary – Training Set – Post-Pruning

(ccp_alpha=0.00012267633155167048, class_weight='balanced', random_state=1)

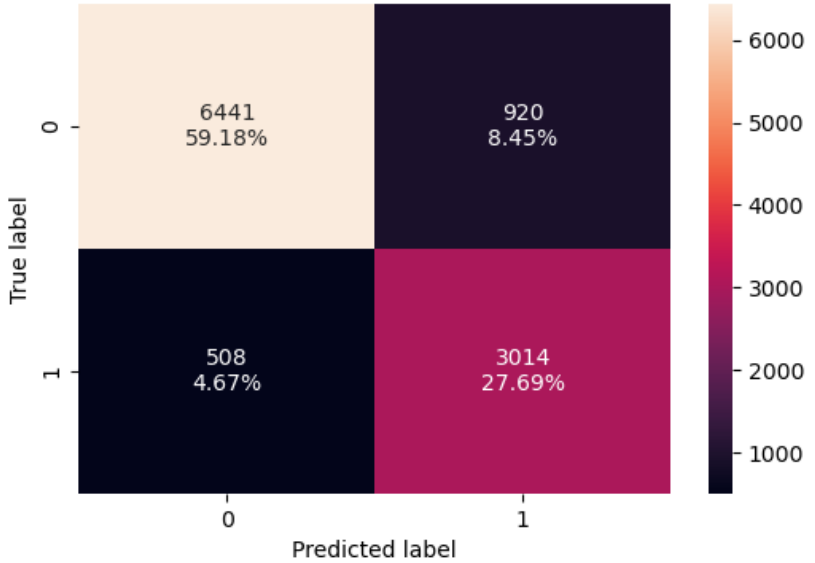


	Accuracy	Recall	Precision	F1
0	0.89993	0.90291	0.81369	0.85598

Results – When using the ccp_alpha to prune the original model, we see better results than before. F1 score of 86% is very good, compared to 70% in the previous run.

Decision Tree Performance Summary – Test Set – Post-Pruning

(ccp_alpha=0.00012267633155167048, class_weight='balanced', random_state=1)



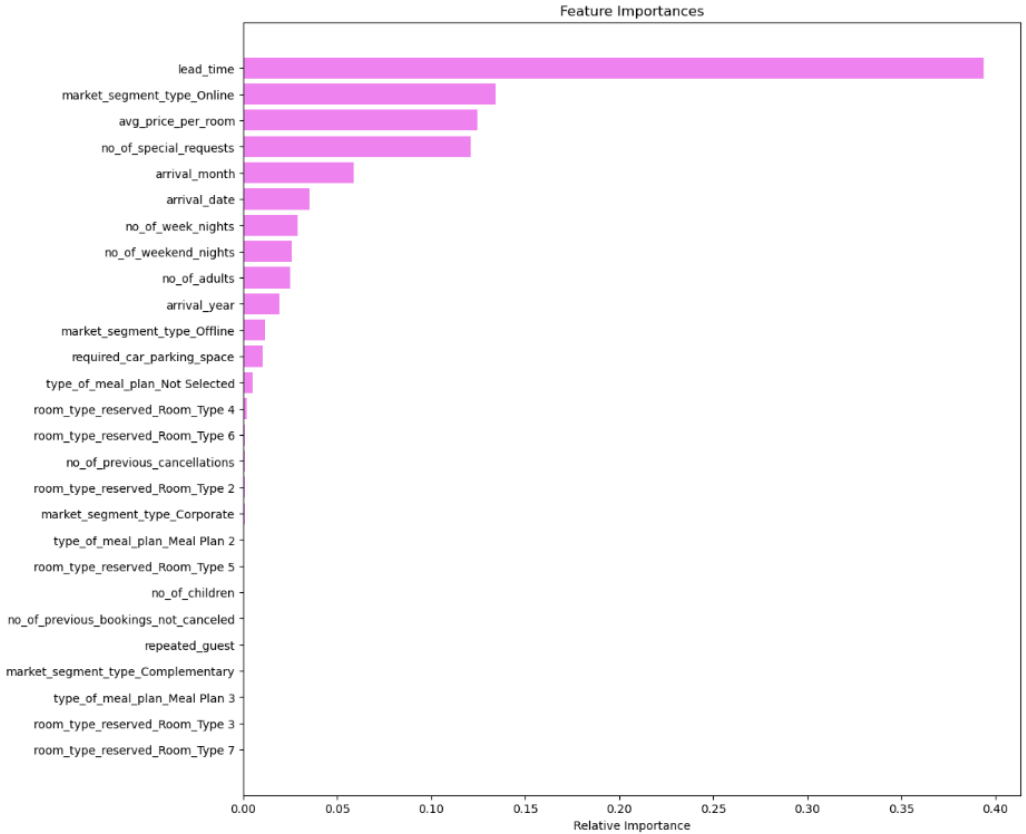
	Accuracy	Recall	Precision	F1
0	0.86943	0.85662	0.76710	0.80939

Results – Since we used the optimal ccp_alpha, we can maximize the F1 score while avoiding overfitting

Decision Tree Model – Feature Importance – Post-Pruning

Observation:

By post-pruning we added some to the list of parameters that play a role in the decision tree.
Also, we can see that the weights of the following 3 features (**online orders, special requests and room price**) are similar to each other, in the previous variations their weights were not as similar.



Decision Tree Model – Performance Summary

Train set results

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.99421	0.83097	0.89993
Recall	0.98661	0.78608	0.90291
Precision	0.99578	0.72425	0.81369
F1	0.99117	0.75390	0.85598

Test set results

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.87118	0.83497	0.86943
Recall	0.81175	0.78336	0.85662
Precision	0.79461	0.72758	0.76710
F1	0.80309	0.75444	0.80939

Final Results:

The initial run, where we allowed the tree to go as deep as possible, we can see the overfitting of the model, and how the test set is 20% worse than the training set.
Between the 2 pruning options, we can see that pre-pruning brings a result that is almost identical between the training and testing sets, but the F1 score is only 75%.

Our best result comes from the post-pruning run, that despite a slight overfitting of the training set, gives us a better result on the test set than all the others.

APPENDIX

Model Performance Summary – Initial Run Full Results

Great Learning
POWER AHEAD

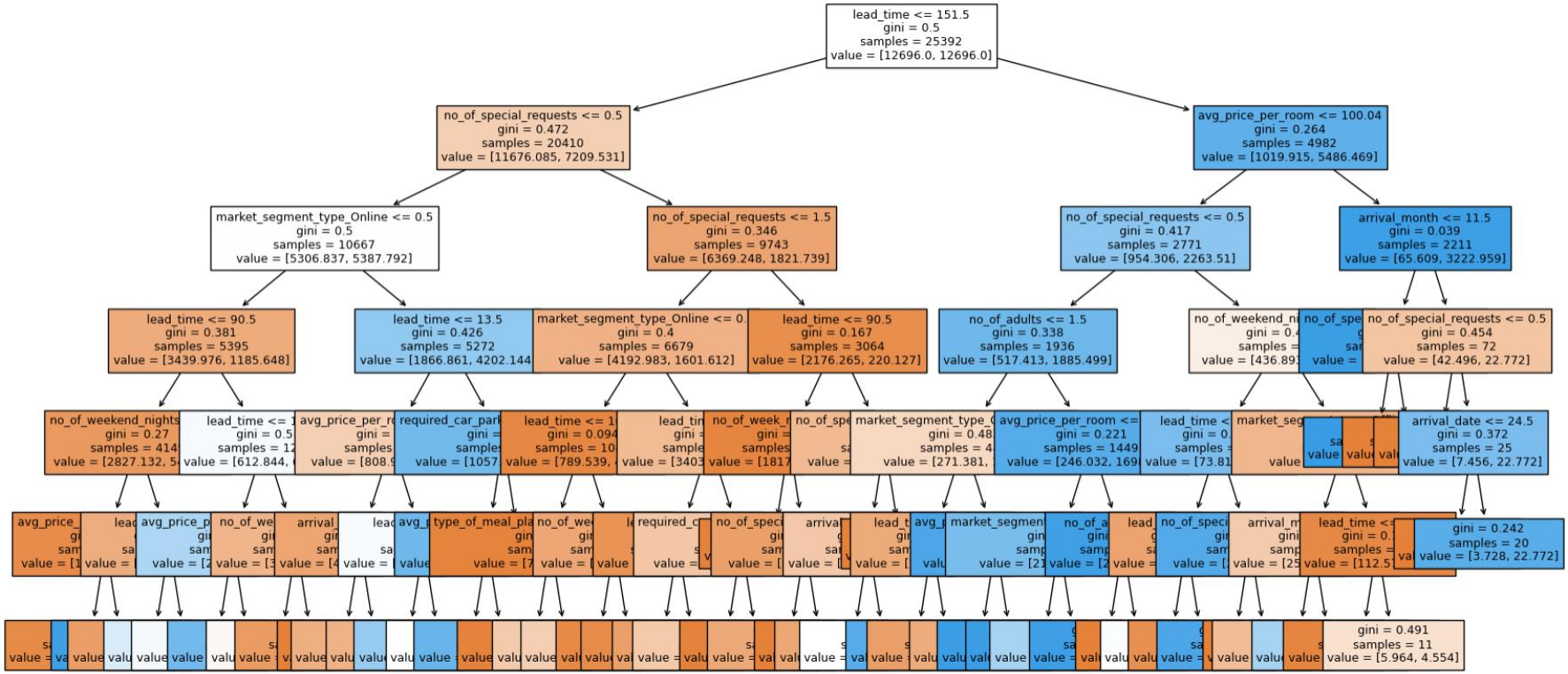
Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25364
Method:	MLE	Df Model:	27
Date:	Sun, 22 Sep 2024	Pseudo R-squ.:	0.3292
Time:	09:37:06	Log-Likelihood:	-10794.
converged:	False	LL-Null:	-16091.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-924.1330	120.816	-7.649	0.000	-1160.928	-687.338
no_of_adults	0.1136	0.038	3.020	0.003	0.040	0.187
no_of_children	0.1561	0.057	2.727	0.006	0.044	0.268
no_of_weekend_nights	0.1067	0.020	5.394	0.000	0.068	0.145
no_of_week_nights	0.0398	0.012	3.236	0.001	0.016	0.064
required_car_parking_space	-1.5942	0.138	-11.564	0.000	-1.864	-1.324
lead_time	0.0157	0.000	58.866	0.000	0.015	0.016
arrival_year	0.4567	0.060	7.629	0.000	0.339	0.574
arrival_month	-0.0416	0.006	-6.436	0.000	-0.054	-0.029
arrival_date	0.0005	0.002	0.264	0.792	-0.003	0.004
repeated_guest	-2.3465	0.617	-3.805	0.000	-3.555	-1.138
no_of_previous_cancellations	0.2663	0.086	3.108	0.002	0.098	0.434
no_of_previous_bookings_not_canceled	-0.1727	0.153	-1.131	0.258	-0.472	0.127
avg_price_per_room	0.0188	0.001	25.444	0.000	0.017	0.020
no_of_special_requests	-1.4690	0.030	-48.791	0.000	-1.528	-1.410
type_of_meal_plan_Meal Plan 2	0.1756	0.067	2.636	0.008	0.045	0.306
type_of_meal_plan_Meal Plan 3	17.4845	4247.878	0.004	0.997	-8308.204	8343.173
type_of_meal_plan_Not Selected	0.2783	0.053	5.247	0.000	0.174	0.382
room_type_reserved_Room_Type 2	-0.3607	0.131	-2.758	0.006	-0.617	-0.104
room_type_reserved_Room_Type 3	-0.0014	1.310	-0.001	0.999	-2.569	2.566
room_type_reserved_Room_Type 4	-0.2829	0.053	-5.320	0.000	-0.387	-0.179
room_type_reserved_Room_Type 5	-0.7187	0.209	-3.437	0.001	-1.128	-0.309
room_type_reserved_Room_Type 6	-0.9474	0.147	-6.447	0.000	-1.235	-0.659
room_type_reserved_Room_Type 7	-1.3992	0.293	-4.776	0.000	-1.973	-0.825
market_segment_type_Complementary	-40.9223	6.23e+05	-6.56e-05	1.000	-1.22e+06	1.22e+06
market_segment_type_Corporate	-1.1940	0.266	-4.489	0.000	-1.715	-0.673
market_segment_type_Offline	-2.1948	0.255	-8.622	0.000	-2.694	-1.696
market_segment_type_Online	-0.3996	0.251	-1.590	0.112	-0.892	0.093

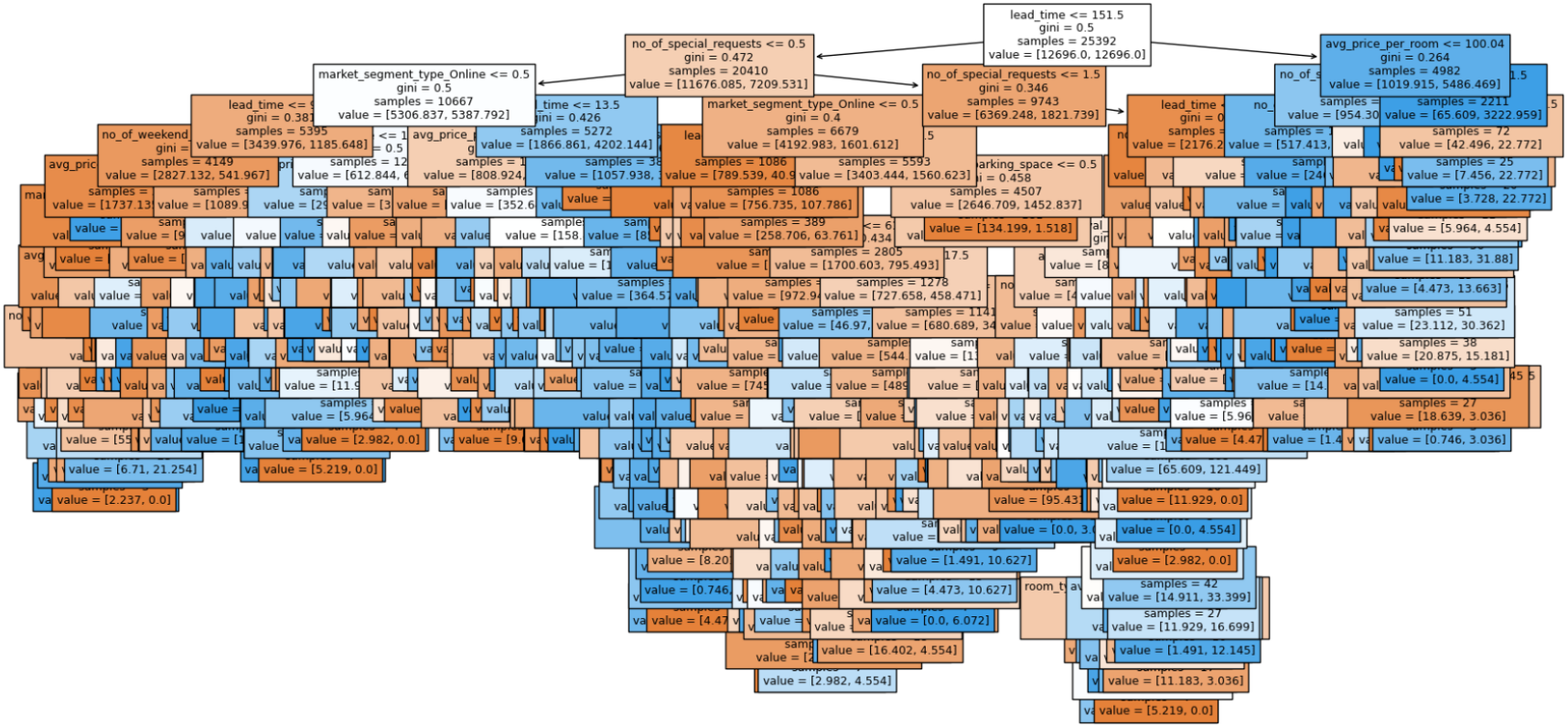
Decision Tree Visualization – Pre-Pruning Settings

(max_depth=6, max_leaf_nodes=50, min_samples_split=10, random_state=1)



Decision Tree Visualization – Post-Pruning Settings

(ccp_alpha=0.00012267633155167048, class_weight='balanced', random_state=1)



Great Learning

Happy Learning !