DSBA
Machine Learning - 1
Coded Project
INN Hotels Business Report

By: Sagar R Naivaruni

Contents

Problem Statement

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.

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

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

Objective

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

Data Description

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

Data Dictionary:

- Booking_ID: the 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 weeknights (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 Group
- 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 Questions

1. What are the busiest months in the hotel?
2. Which market segment do most of the guests come from?
3. Hotel rates are dynamic and change according to demand and customer demographics. What are the differences in room prices in different market segments?
4. What percentage of bookings are canceled?
5. Repeating guests are the guests who stay in the hotel often and are important to brand equity. What percentage of repeating guests cancel?
6. Many guests have special requirements when booking a hotel room. Do these requirements affect booking cancellation?

Data Overview

Data Structure

Booking_ID	no_of_adults	no_of_children	no_of_weekenc	no_of_week_ni	type_of_meal_	required_car_p	room_type_res	lead_time	arrival_year	arrival_month	arrival_date	market_segmer	repeated_guest	no_of_previous	no_of_previous	avg_price_per_	no_of_special_	booking_statu
INN00001	2	0	1	2	Meal Plan 1	0	Room_Type 1	224	2017	10	2	Offline	0	0	0	65	0	Not_Canceled
INN00002	2	0	2	3	Not Selected	0	Room_Type 1	5	2018	11	6	Online	0	0	0	106.68	1	Not_Canceled
INN00003	1	0	2	1	Meal Plan 1	0	Room_Type 1	1	2018	2	28	Online	0	0	0	60	0	Canceled
INN00004	2	0	0	2	Meal Plan 1	0	Room_Type 1	211	2018	5	20	Online	0	0	0	100	0	Canceled
INN00005	2	0	1	1	Not Selected	0	Room_Type 1	48	2018	4	11	Online	0	0	0	94.5	0	Canceled

- There are 36275 rows and 19 columns.

Data Types

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36275 entries, 0 to 36274
Data columns (total 19 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Booking_ID                            36275 non-null  object
 1   no_of_adults                          36275 non-null  int64
 2   no_of_children                        36275 non-null  int64
 3   no_of_weekend_nights                  36275 non-null  int64
 4   no_of_week_nights                     36275 non-null  int64
 5   type_of_meal_plan                     36275 non-null  object
 6   required_car_parking_space            36275 non-null  int64
 7   room_type_reserved                    36275 non-null  object
 8   lead_time                             36275 non-null  int64
 9   arrival_year                          36275 non-null  int64
 10  arrival_month                         36275 non-null  int64
 11  arrival_date                          36275 non-null  int64
 12  market_segment_type                   36275 non-null  object
 13  repeated_guest                        36275 non-null  int64
 14  no_of_previous_cancellations          36275 non-null  int64
 15  no_of_previous_bookings_not_canceled  36275 non-null  int64
 16  avg_price_per_room                    36275 non-null  float64
 17  no_of_special_requests                36275 non-null  int64
 18  booking_status                        36275 non-null  object
dtypes: float64(1), int64(13), object(5)
memory usage: 5.3+ MB
```

There are:

- 13 int data types.
- 5 object data types.
- 1 float data type.

Duplicate Values

	0
Booking_ID	0
no_of_adults	0
no_of_children	0
no_of_weekend_nights	0
no_of_week_nights	0
type_of_meal_plan	0
required_car_parking_space	0
room_type_reserved	0
lead_time	0
arrival_year	0
arrival_month	0
arrival_date	0
market_segment_type	0
repeated_guest	0
no_of_previous_cancellations	0
no_of_previous_bookings_not_canceled	0
avg_price_per_room	0
no_of_special_requests	0
booking_status	0

dtype: int64

- There are no duplicate values in the dataset.

Null Values

	0
Booking_ID	0
no_of_adults	0
no_of_children	0
no_of_weekend_nights	0
no_of_week_nights	0
type_of_meal_plan	0
required_car_parking_space	0
room_type_reserved	0
lead_time	0
arrival_year	0
arrival_month	0
arrival_date	0
market_segment_type	0
repeated_guest	0
no_of_previous_cancellations	0
no_of_previous_bookings_not_canceled	0
avg_price_per_room	0
no_of_special_requests	0
booking_status	0

dtype: int64

- There are no null values in the dataset.

Exploratory Data Analysis (EDA)

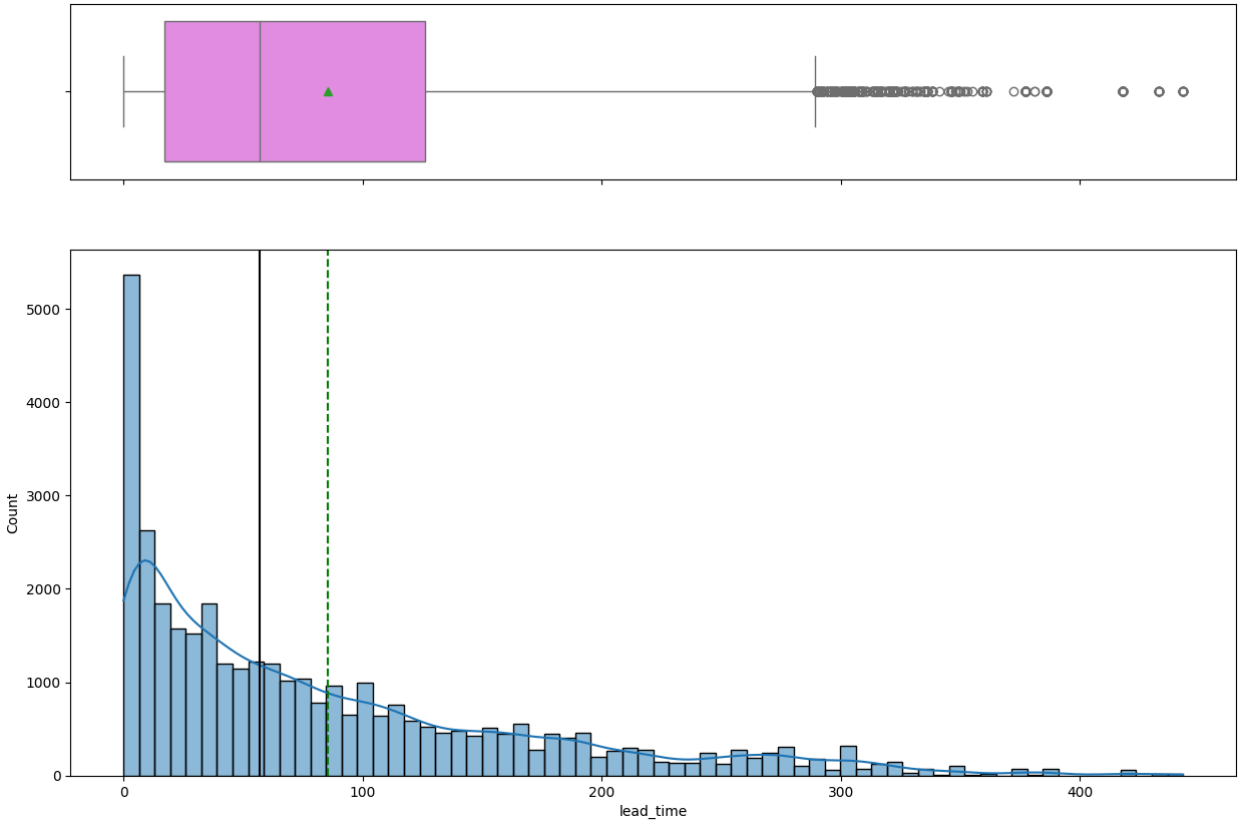
Statistical Summary

	count	mean	std	min	25%	50%	75%	max
no_of_adults	36275.0	1.844962	0.518715	0.0	2.0	2.00	2.0	4.0
no_of_children	36275.0	0.105279	0.402648	0.0	0.0	0.00	0.0	10.0
no_of_weekend_nights	36275.0	0.810724	0.870644	0.0	0.0	1.00	2.0	7.0
no_of_week_nights	36275.0	2.204300	1.410905	0.0	1.0	2.00	3.0	17.0
required_car_parking_space	36275.0	0.030986	0.173281	0.0	0.0	0.00	0.0	1.0
lead_time	36275.0	85.232557	85.930817	0.0	17.0	57.00	126.0	443.0
arrival_year	36275.0	2017.820427	0.383836	2017.0	2018.0	2018.00	2018.0	2018.0
arrival_month	36275.0	7.423653	3.069894	1.0	5.0	8.00	10.0	12.0
arrival_date	36275.0	15.596995	8.740447	1.0	8.0	16.00	23.0	31.0
repeated_guest	36275.0	0.025637	0.158053	0.0	0.0	0.00	0.0	1.0
no_of_previous_cancellations	36275.0	0.023349	0.368331	0.0	0.0	0.00	0.0	13.0
no_of_previous_bookings_not_canceled	36275.0	0.153411	1.754171	0.0	0.0	0.00	0.0	58.0
avg_price_per_room	36275.0	103.423539	35.089424	0.0	80.3	99.45	120.0	540.0
no_of_special_requests	36275.0	0.619655	0.786236	0.0	0.0	0.00	1.0	5.0

- Average no of adults is 2 and since the standard deviation is also 2 the dataset is normally distributed.
- There are outliers with respect to the number of children as parents don't bring their children with them to stay at the hotel.
- The avg price per room is €103 while the median is €99 which indicates there is a slight skewness.
- Most people have cancelled their booking.

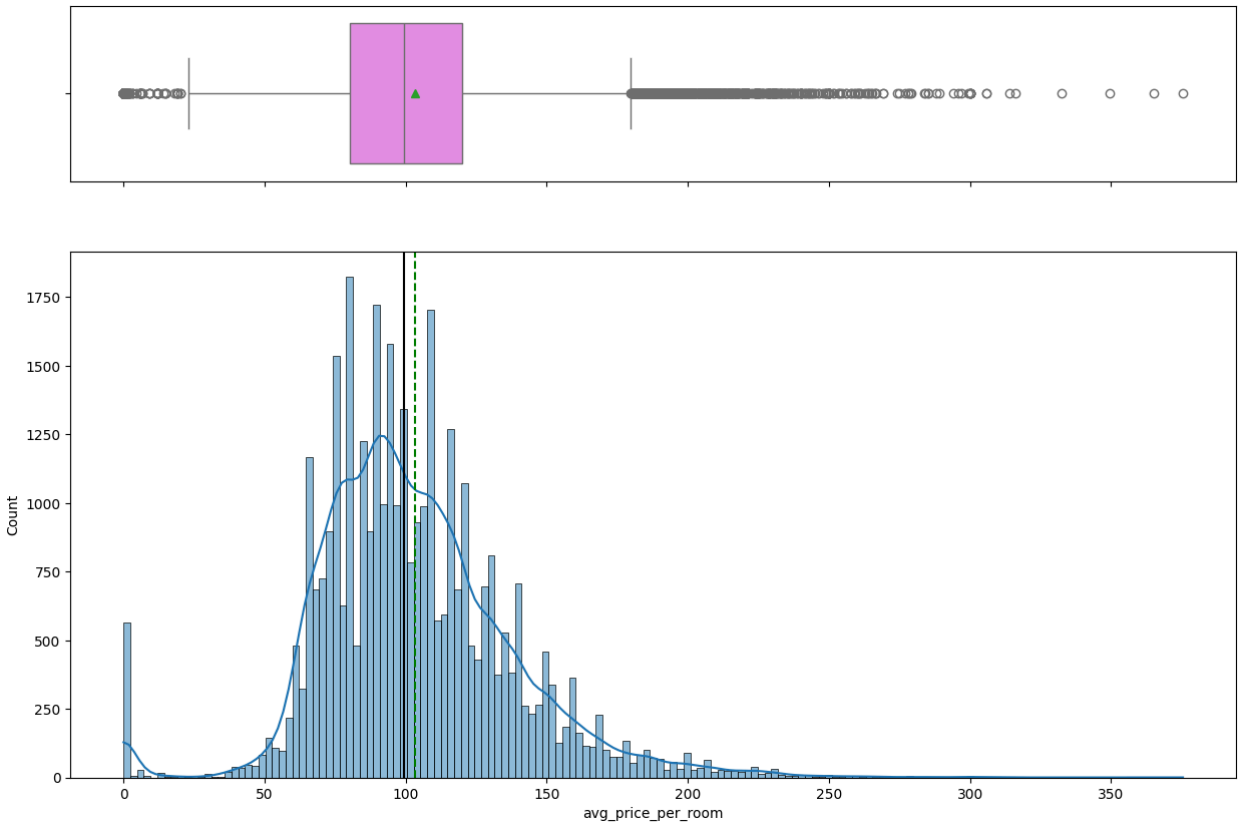
Univariate Analysis

Lead Time



- The distribution is skewed to the right and there are many outliers.
- Most customers made their booking on the day of arrival i.e day 0.
- There are some customers who made their booking 100 to 300 days in advance as well.

Average Price Per Room



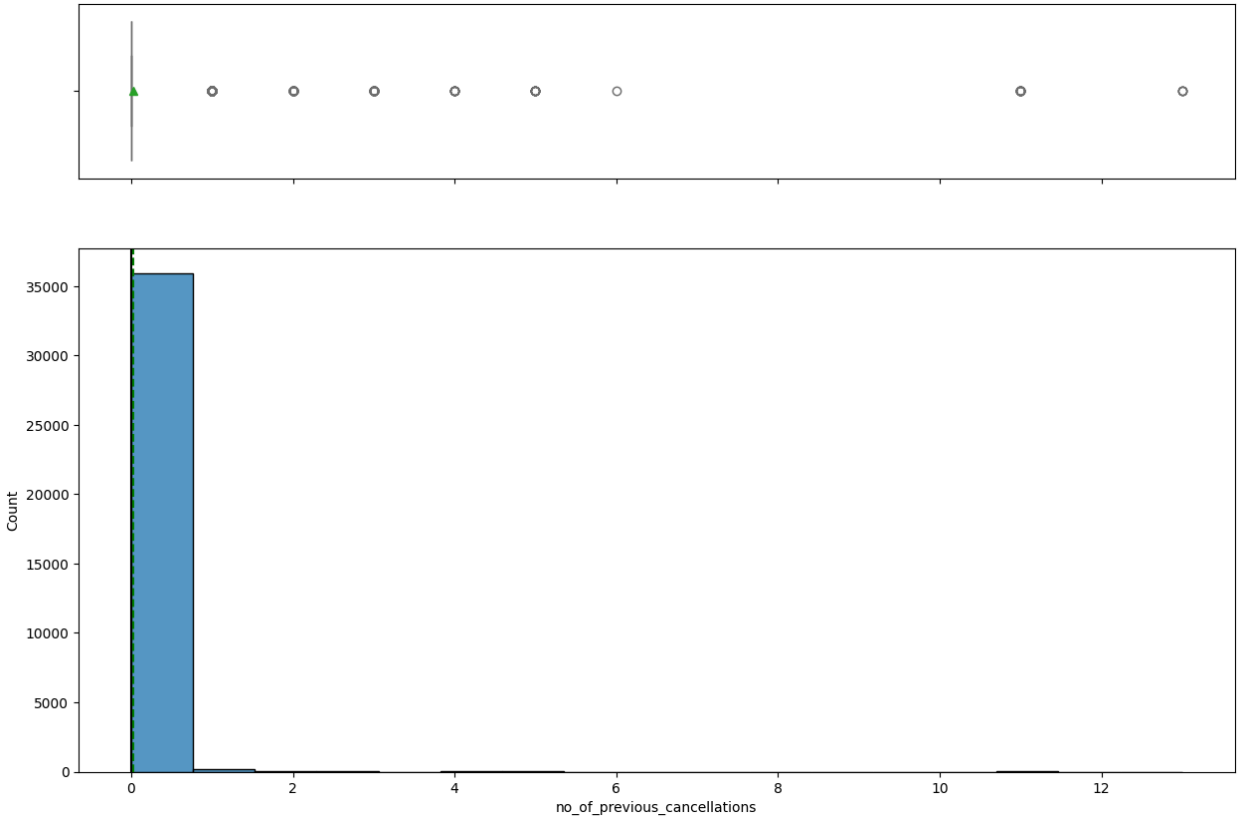
- There are outliers on both the sides and the curve is slightly skewed to the right.
- The avg price of the room per night is around €100.
- There are rooms with a price of €0 which has to be checked.

€0 Room

market_segment_type	count
Complementary	354
Online	191

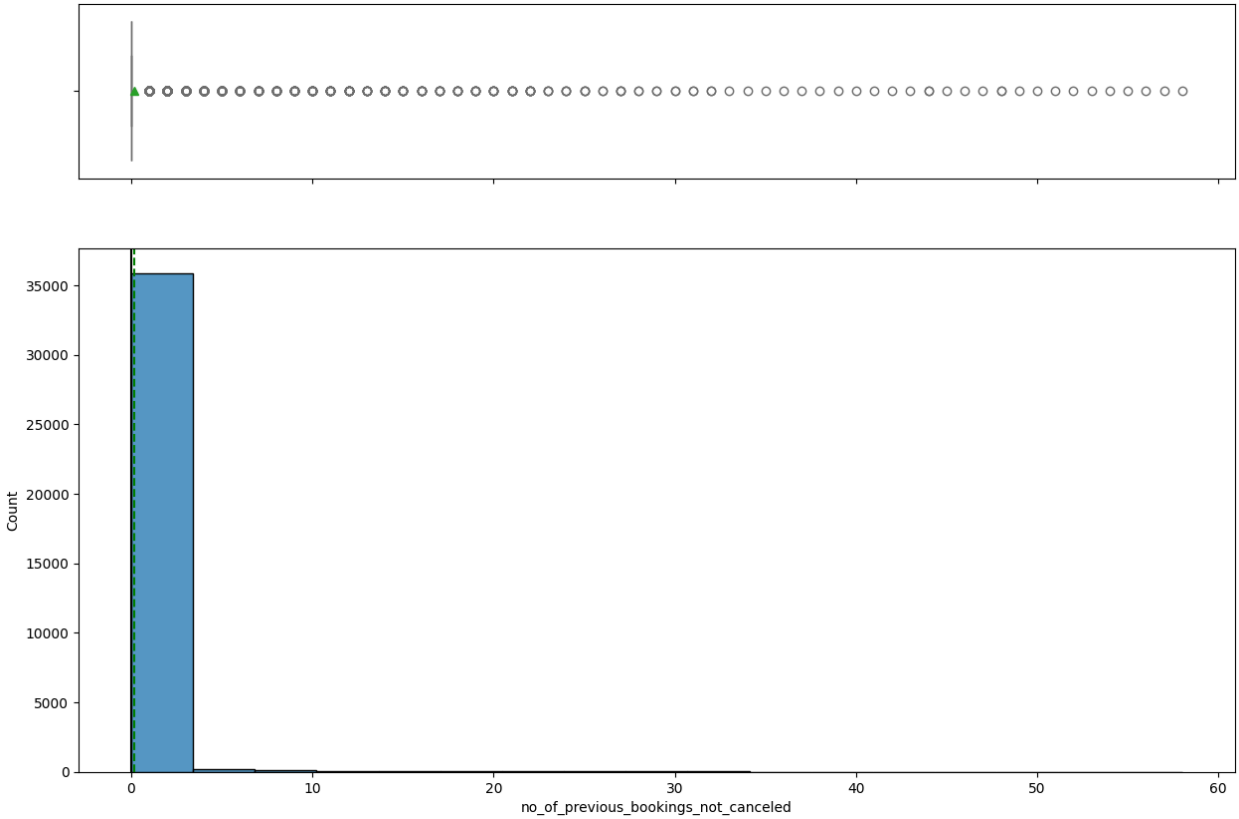
- There are some rooms which were given free at €0 as they were complimentary.
- The rooms that were booked online and were given free at €0 must be a part of some promotional offer conducted by the hotel.

Previous Booking Cancellations



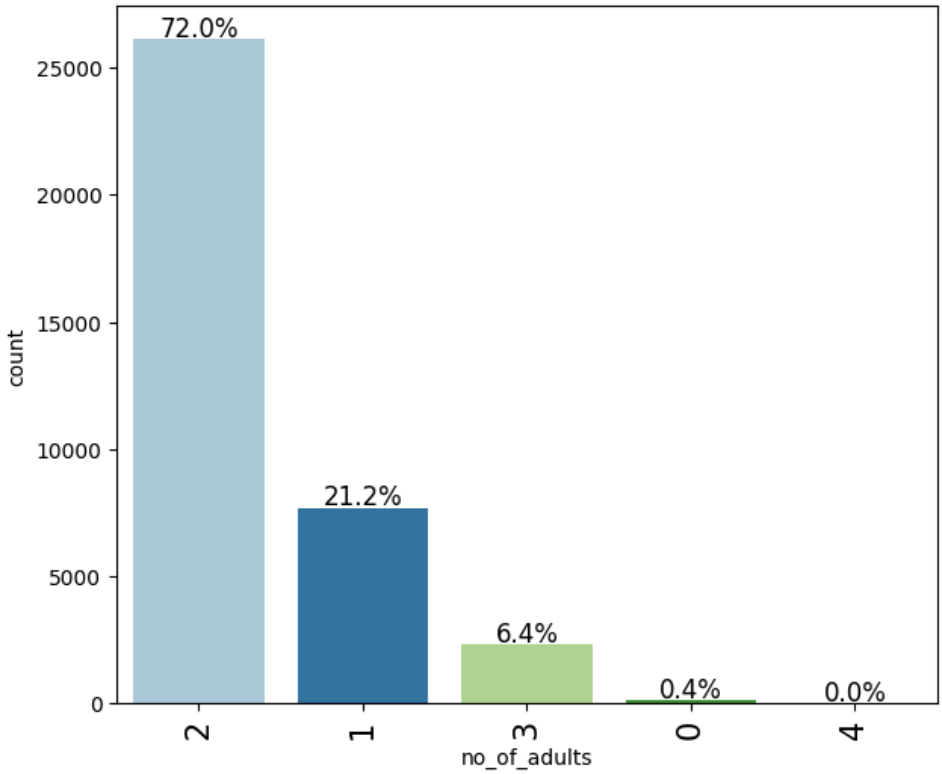
- Most customers who cancelled previously only cancelled 1 time while there are some outliers present.

Previous Bookings not Cancelled



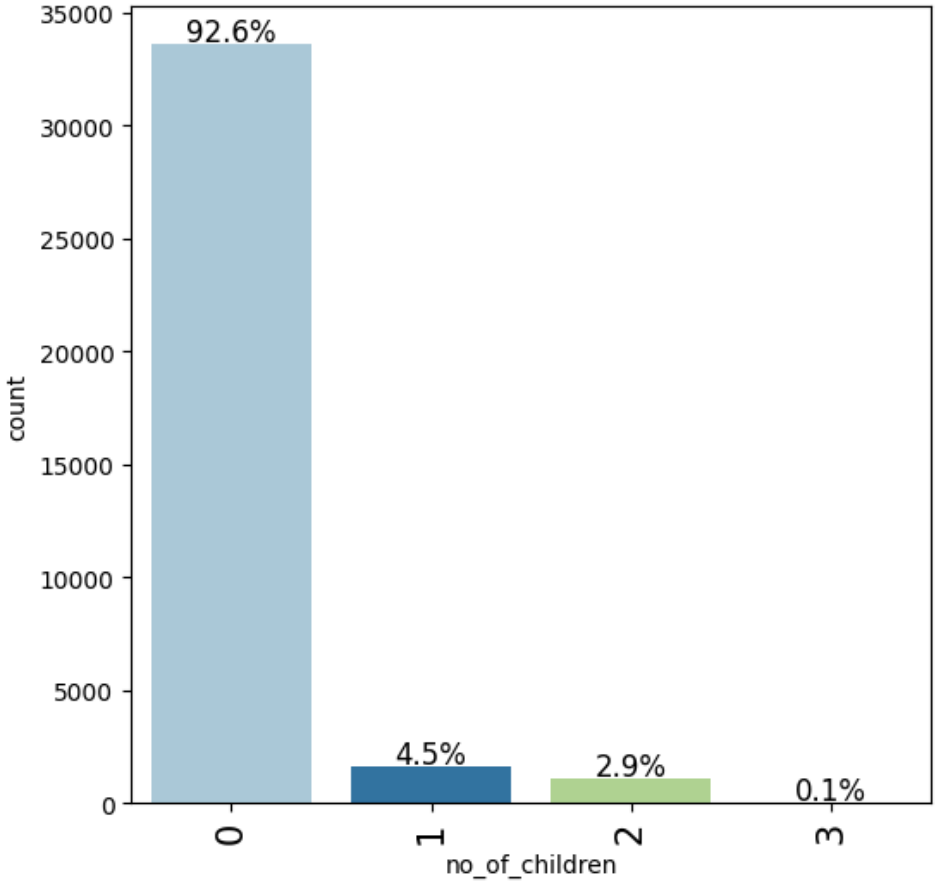
- Most of the customers did not cancel their bookings previously.

No of Adults



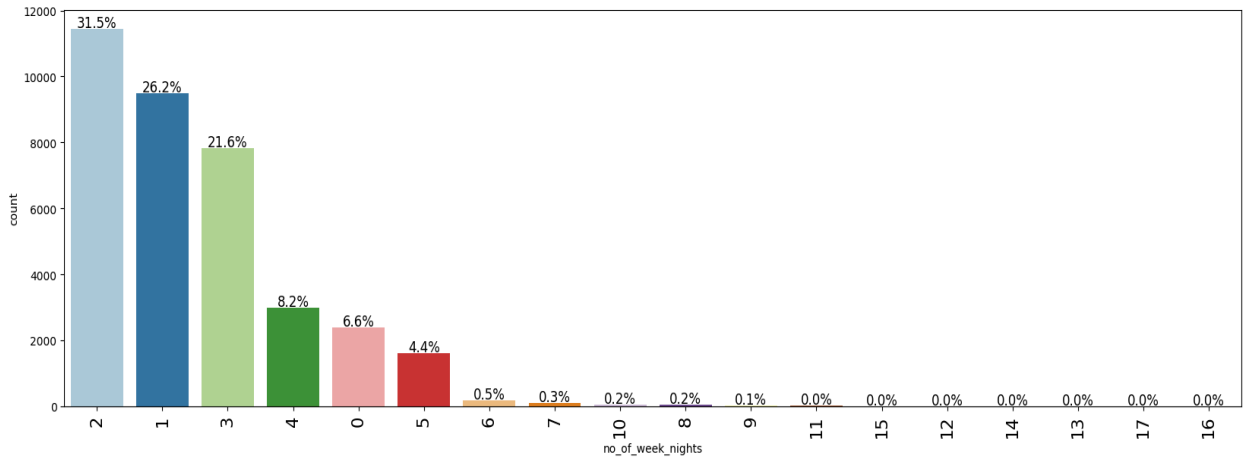
- 72% bookings were made for 2 adults.
- 21.2% bookings were made for 1 adult.
- The trend indicates that mostly couples and solo travellers preferred to stay at the hotel.

No of Children



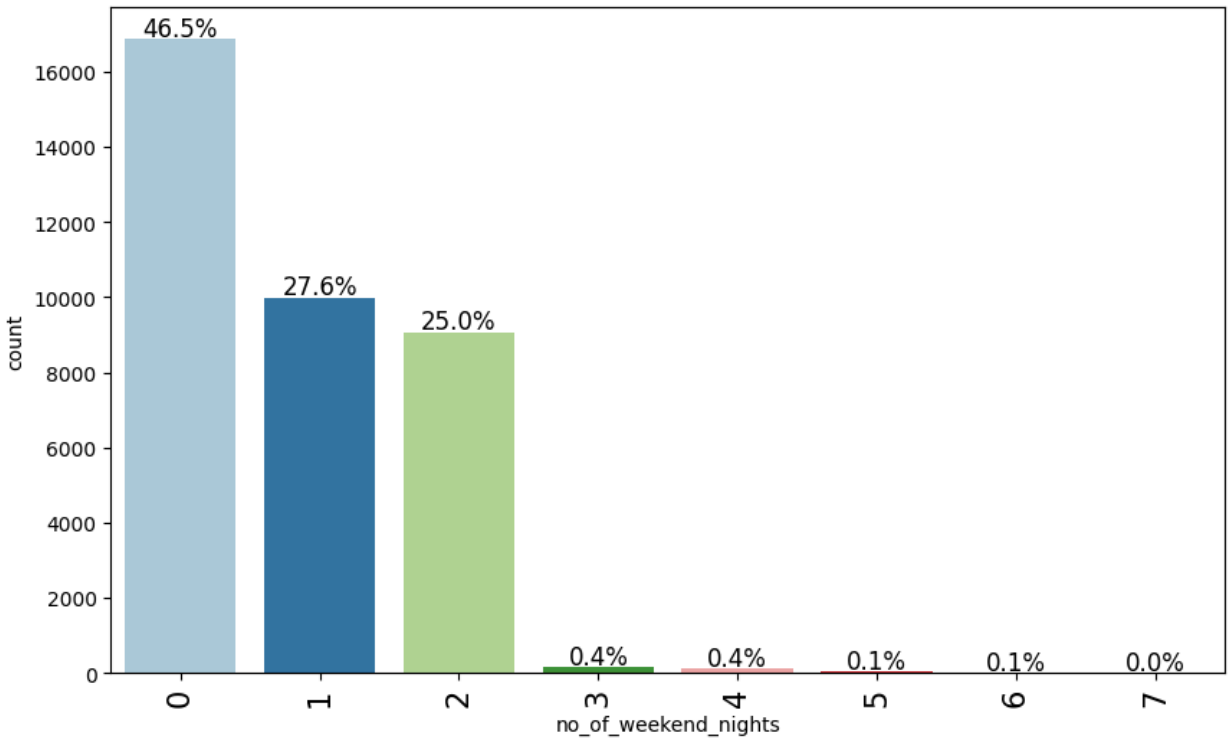
- 92.6% did not make any reservation for their children.
- There are some outliers of 9 & 10 children which are unlikely hence it is being replaced with the maximum value i.e 3.

No of Week Nights



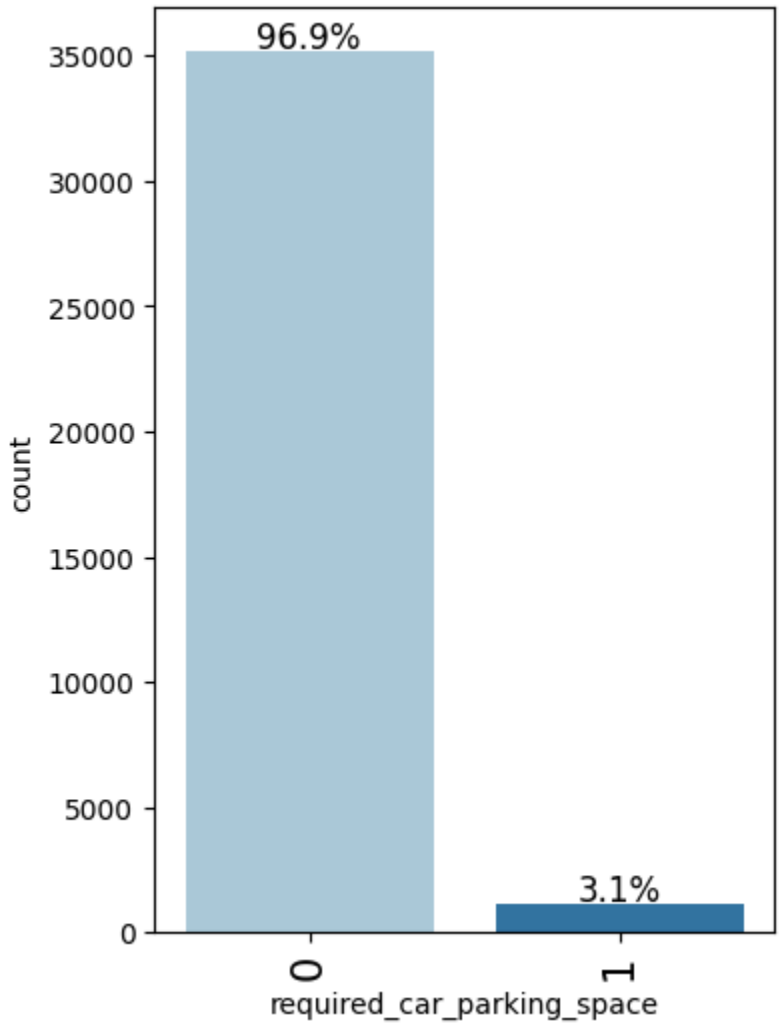
- 31.5% of the bookings were done for 2 nights.
- 26.2% of the bookings were done for 1 night.
- 21.6% bookings were done for 3 nights.

No of Weekend Nights



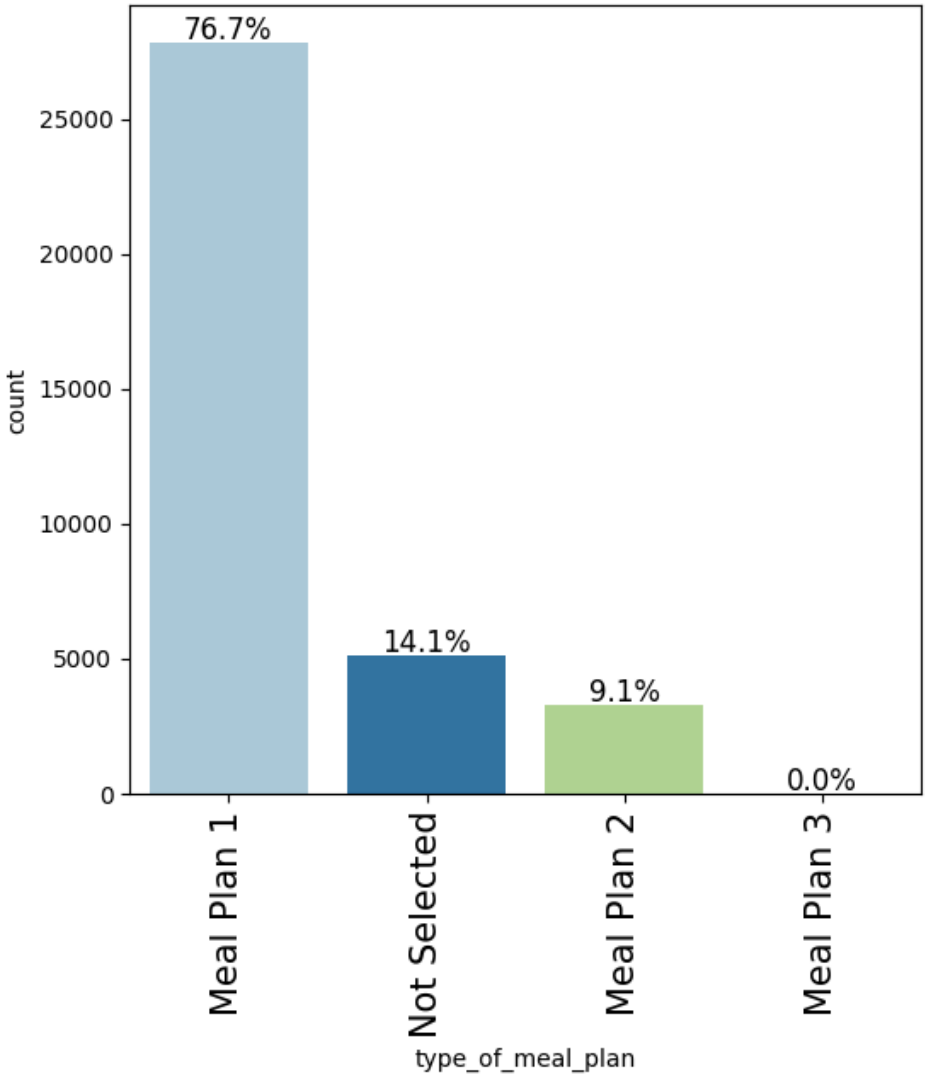
- 46.5% of the customers don't want to spend their weekend at the hotel.
- The percentage of customers who spent 1 & 2 nights at the hotel are almost the same.

Required Car Parking Space



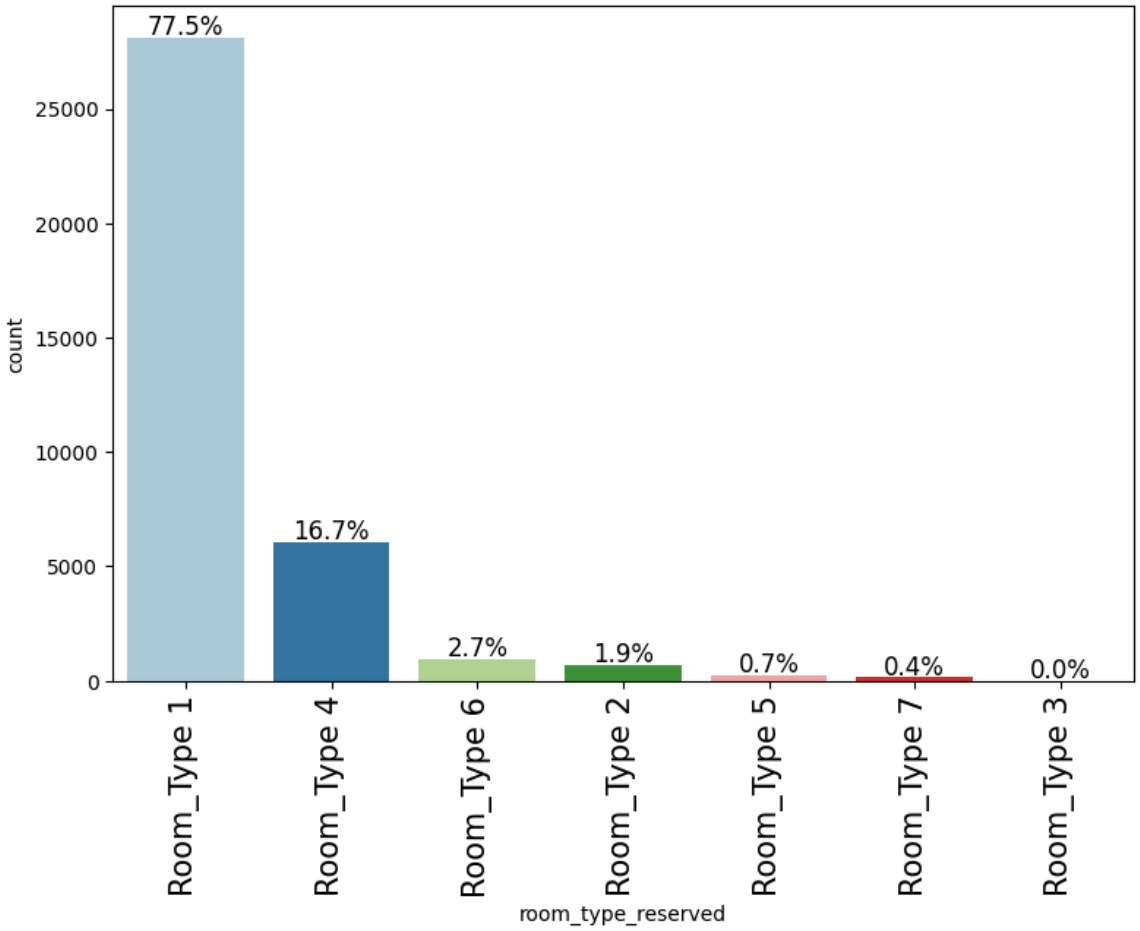
- 96.9% of the customers did not require a car parking space.
- While only 3% of the customers did require it.

Type of Meal Plan



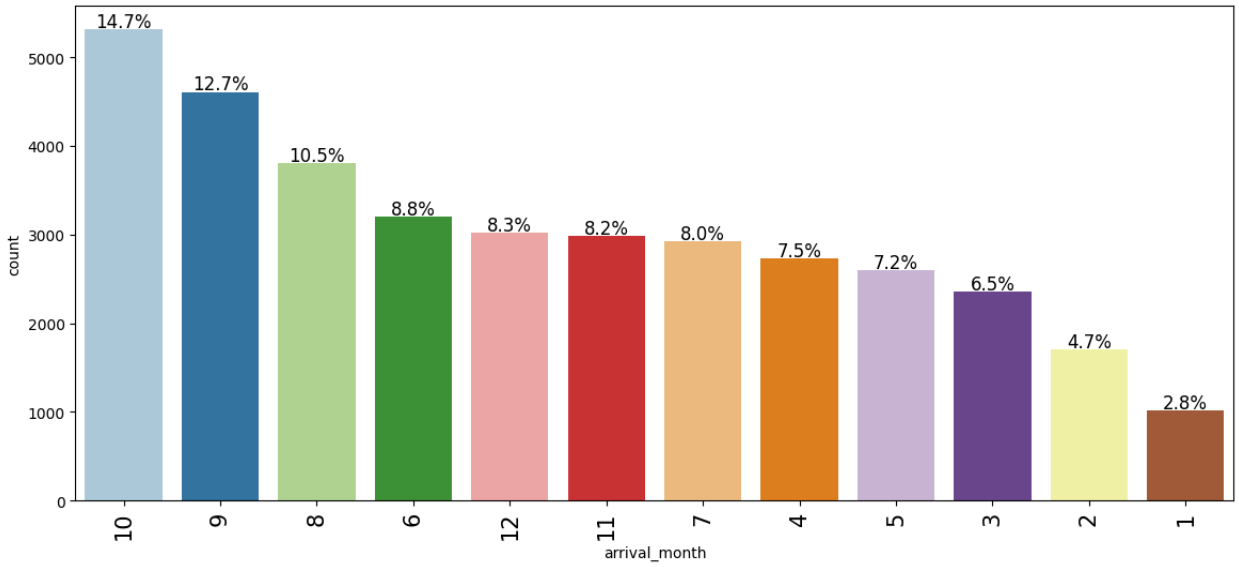
- Most of the customers (76.7%) preferred meal plan 1.
- While 14.1% of the customers did not select any meal plan.
- 9.1% of the customers preferred meal plan 2.

Room Type Reserved



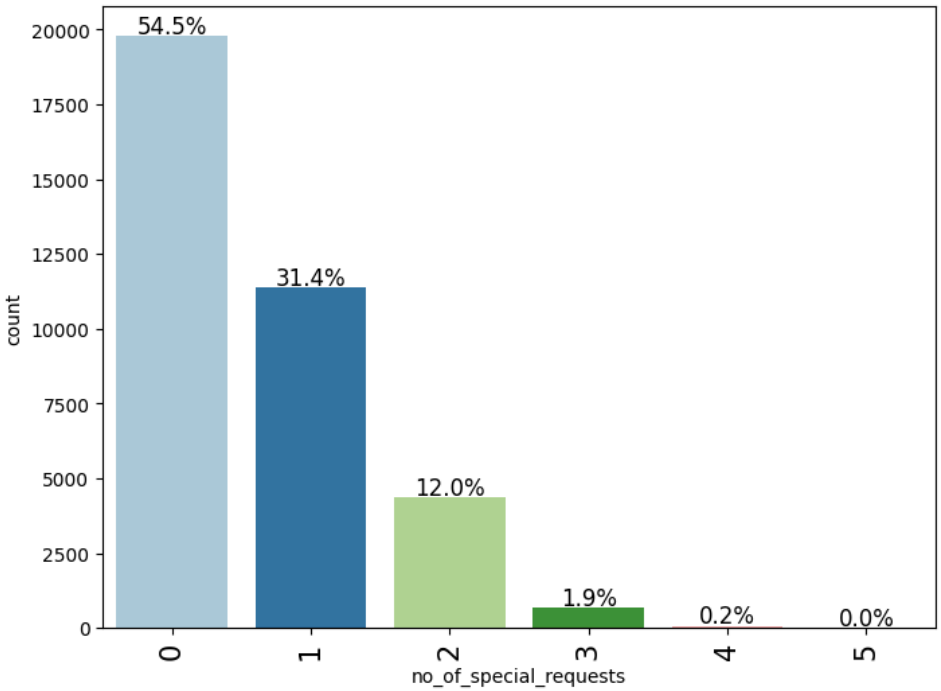
- 77.5% of the customers preferred room type 1.
- Followed by 16.7% preferred room type 4.

Arrival Month



- October is the busiest month with 14.7% bookings.
- Followed by September 12.7% bookings.
- And August with 10.5% of the bookings.

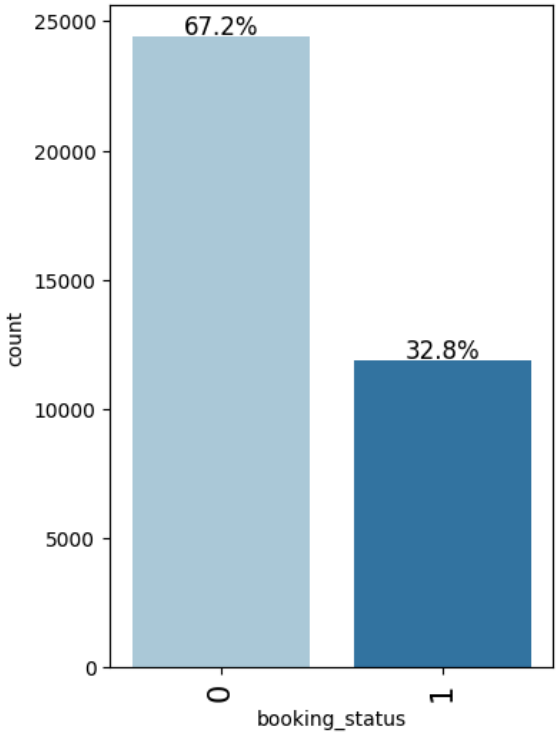
No of Special Requests



- Most customers do not make any special requests while booking the room.
- While some customers make 1 special request while booking the room.

4. What percentage of bookings are canceled?

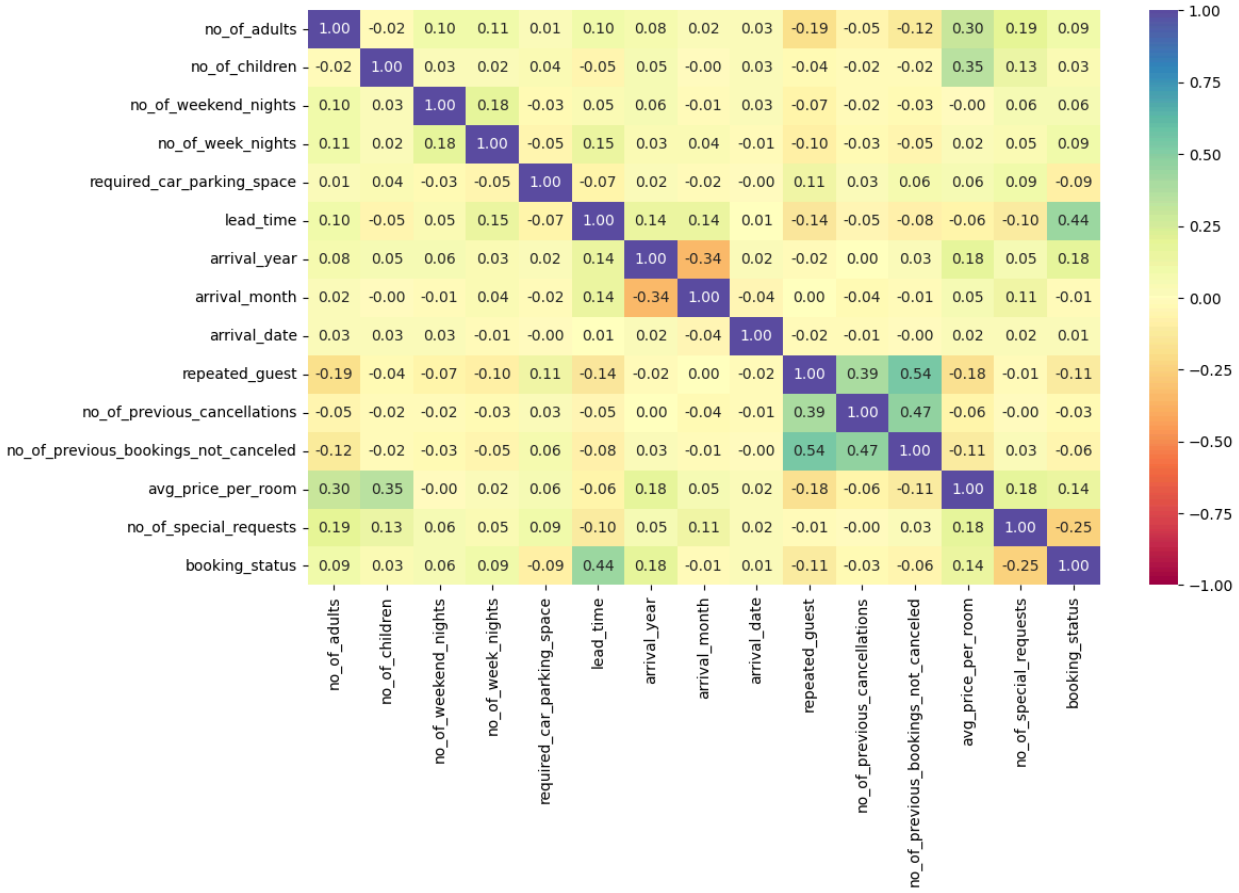
Booking Status



- 32.8% bookings were cancelled by the customers which is really high for a hotel.
- Booking status is coded as 1 for cancelled bookings and 0 for not cancelled bookings.

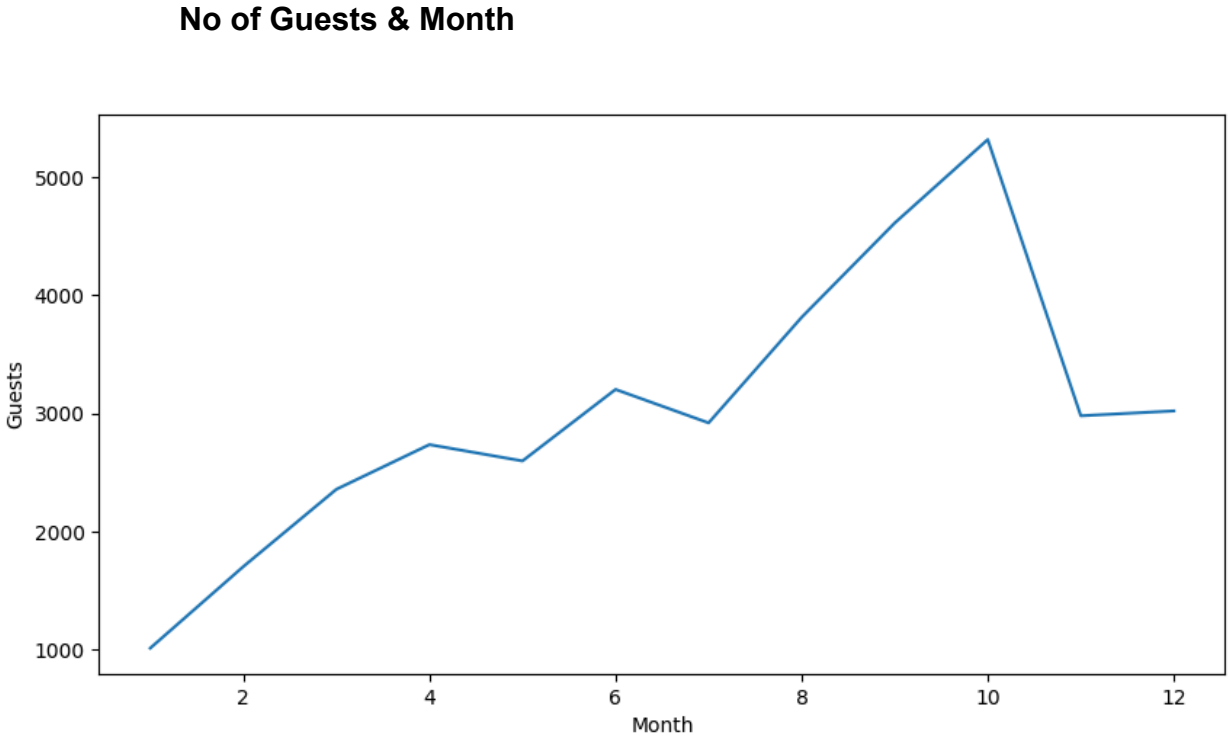
Bivariate Analysis

Correlation



- There is a positive correlation between the avg price of the room and the no of customers meaning more customers equals to more price.
- There is a positive correlation between the booking status and the lead time meaning higher the lead time higher the chances of cancellation.
- There is a positive correlation between the previous booking not cancelled and previous booking cancelled by the customer.
- There is a negative correlation between repeat guests and the room price meaning the hotel might be giving loyalty benefits to the customer.
- There is a negative correlation between the number of special requests and the booking status meaning if the customer has made special requests while booking then the chances of cancellation decreases.

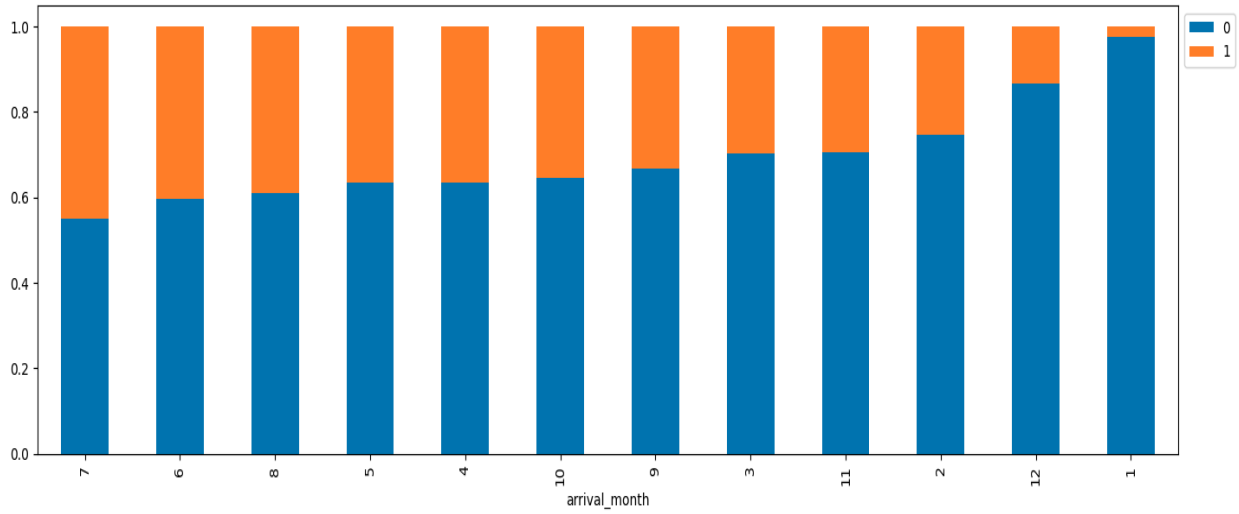
1. What are the busiest months in the hotel?



- Most no of bookings were made in the month of October which had a little over 5000 bookings.
- The least no of bookings were made in the month of January which had only 1000 bookings

Arrival Month & Booking Status

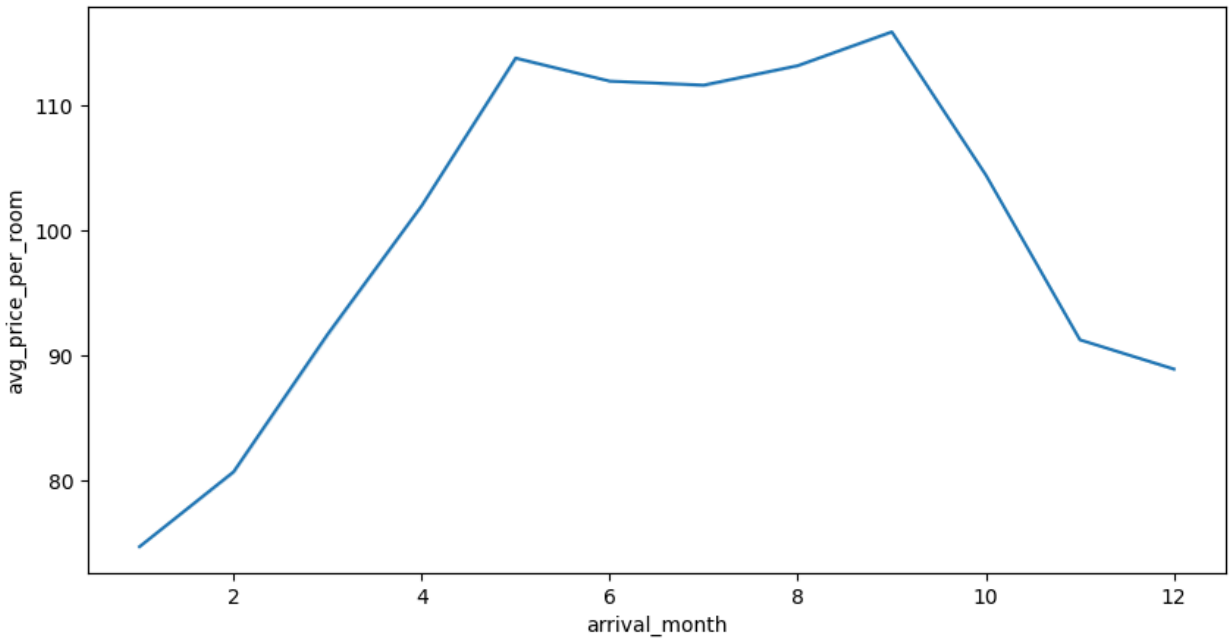
booking_status	0	1	All
arrival_month			
All	24390	11885	36275
10	3437	1880	5317
9	3073	1538	4611
8	2325	1488	3813
7	1606	1314	2920
6	1912	1291	3203
4	1741	995	2736
5	1650	948	2598
11	2105	875	2980
3	1658	700	2358
2	1274	430	1704
12	2619	402	3021
1	990	24	1014



0: Not Cancelled Bookings
1: Cancelled Bookings

- Though the highest no of bookings are made for the month of September and October there is a 40% cancellation rate while January had the least no of bookings it also had the least no of booking cancellations

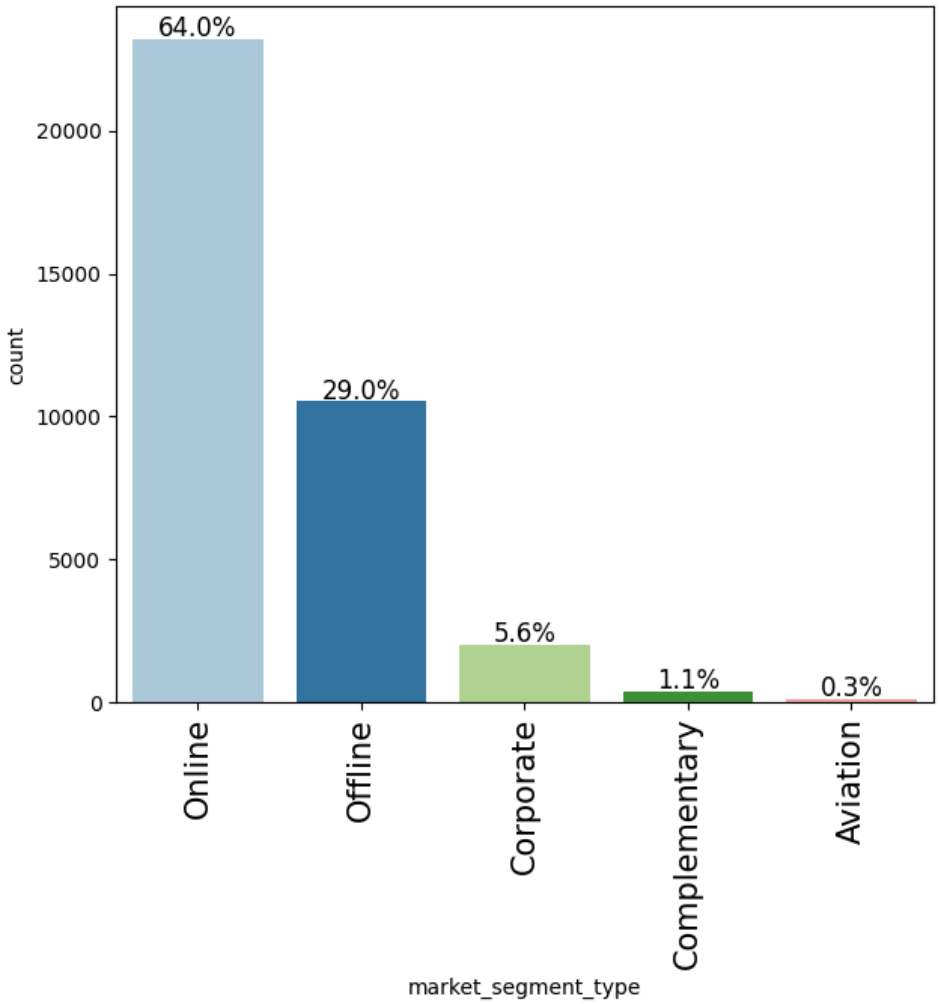
Arrival Month & Average Price



- The rooms cost the highest from May to October, it costs €115 on an average.

2. Which market segment do most of the guests come from?

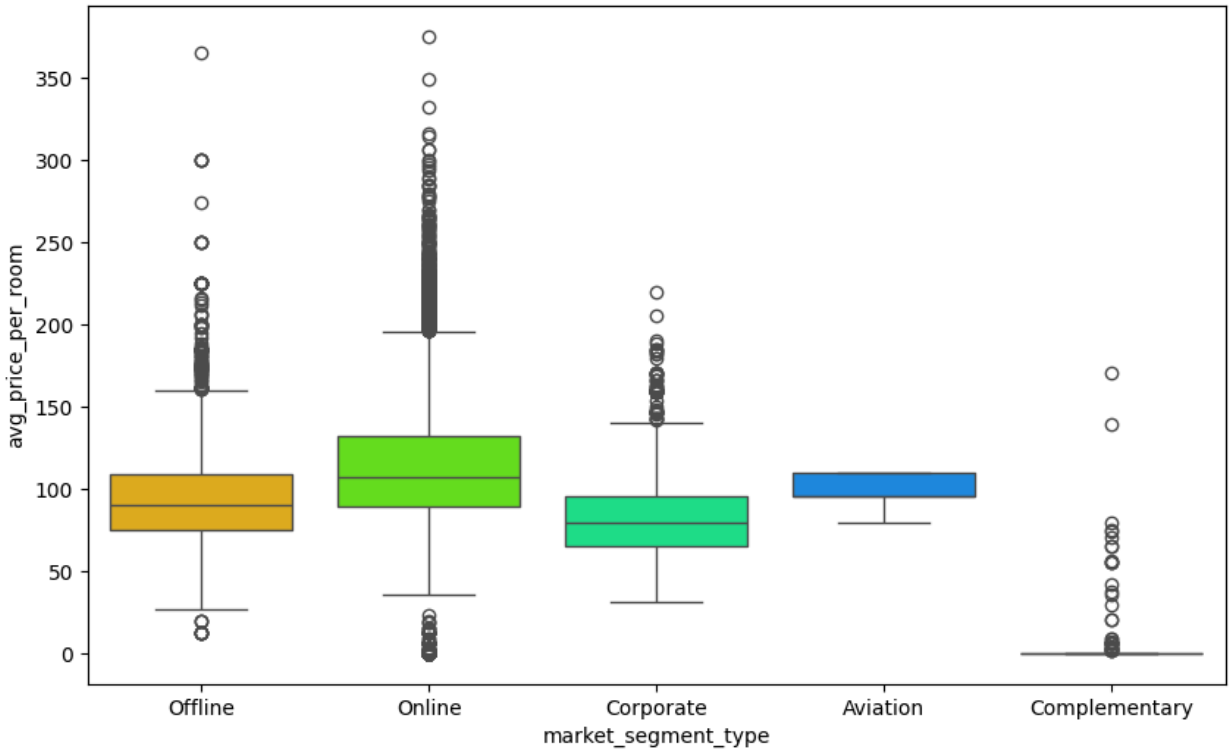
Market Segment Type



- Majority i.e 64% of the bookings were made online.
- And only 29% bookings were made offline.

3. Hotel rates are dynamic and change according to demand and customer demographics. What are the differences in room prices in different market segments?

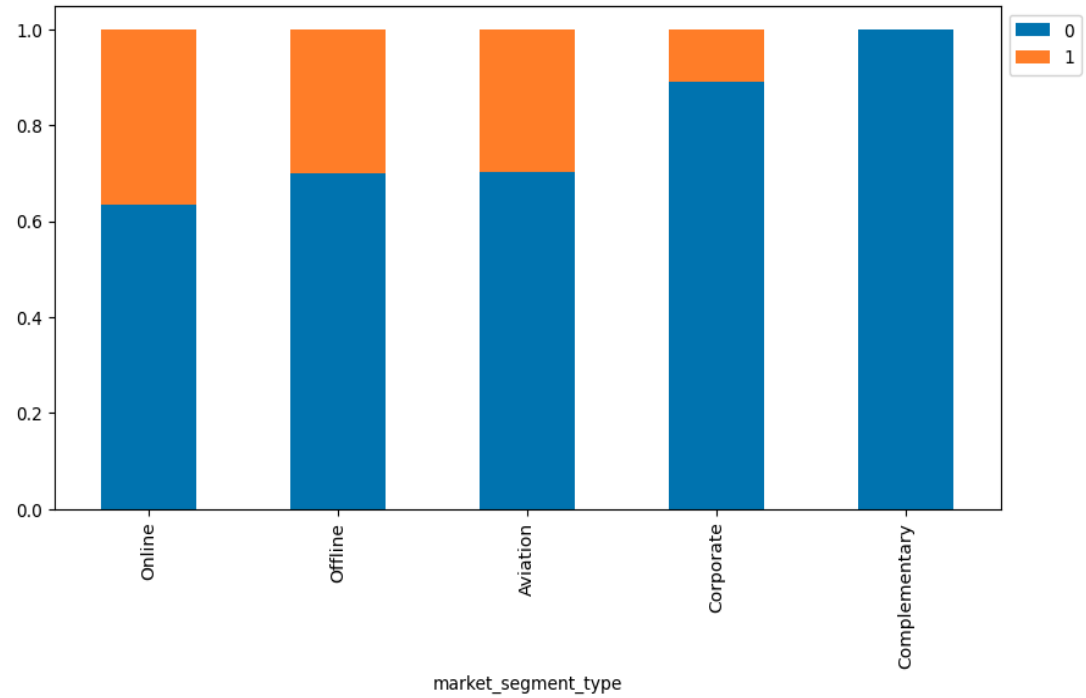
Average Price Per Room & Market Segment Type



- Offline and corporate prices are almost similar.
- Online prices have higher prices.
- Complementary rooms are priced at €0 which is a part of promotional activity run by the hotel.

Market Segment Type & Booking Status

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



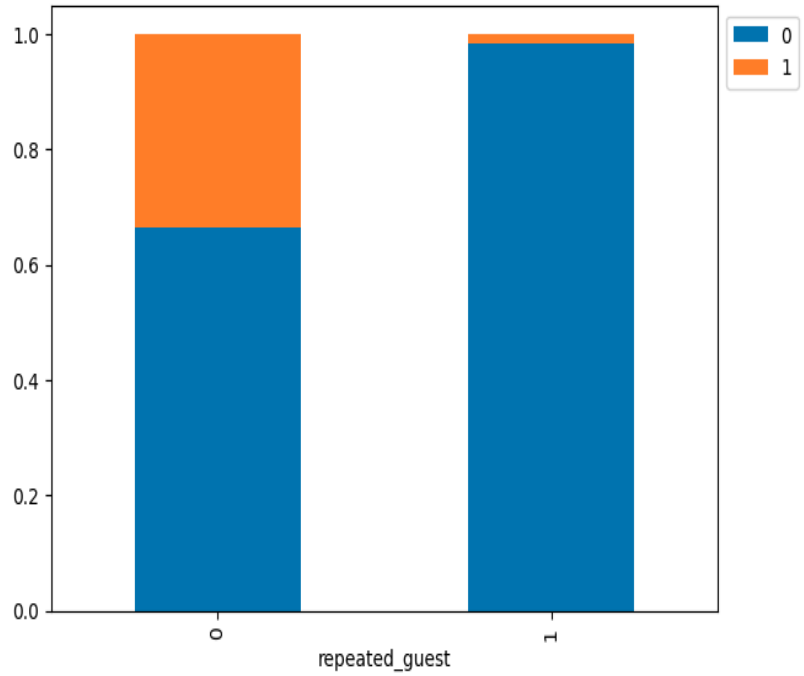
0: Not Cancelled Bookings
1: Cancelled Bookings

- Complimentary rooms were never cancelled.
- Corporate bookings had the least cancellation.
- Online bookings about 40% had cancellation and is the highest.
- While offline and aviation had the same number of cancellations.

5. Repeating guests are the guests who stay in the hotel often and are important to brand equity. What percentage of repeating guests cancel?

Repeated Guests & Booking Status

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



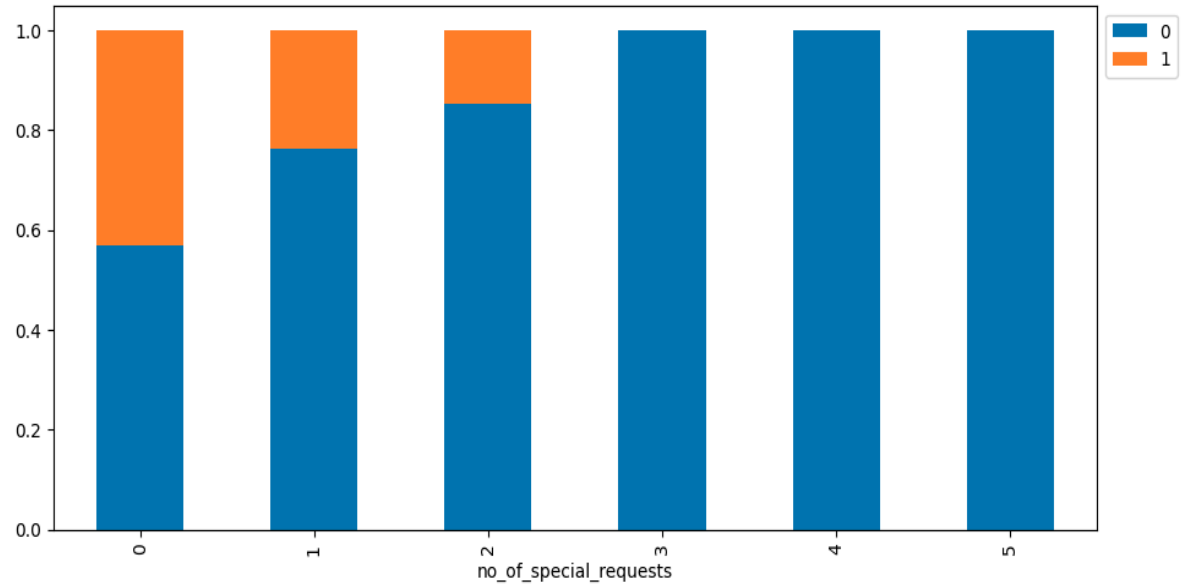
0: Not Cancelled Bookings
1: Cancelled Bookings

- There are a very few repeat guests but the cancellation among the repeat guests are also the least.
- It indicates that attracting a new customer is costly and a tedious task while loyal guests are usually more profitable.

6. Many guests have special requirements when booking a hotel room. Do these requirements affect booking cancellation?

No of Special Requests & Booking Status

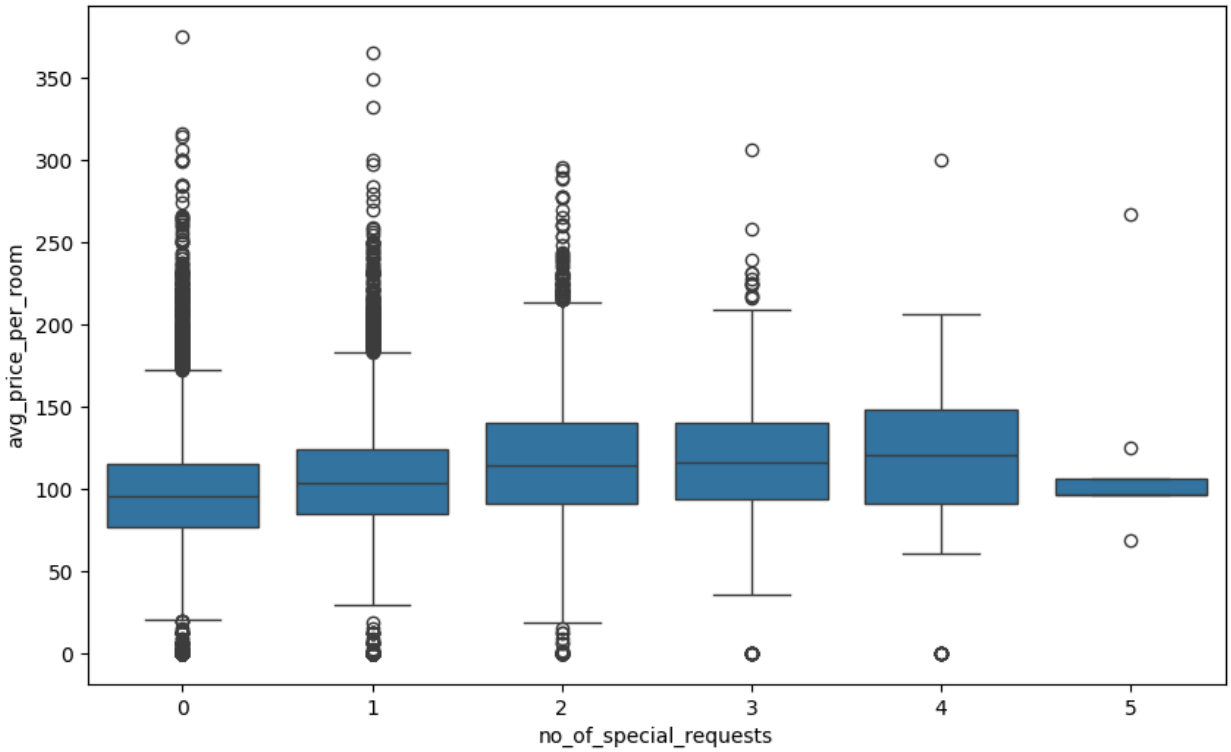
booking_status	0	1	All
no_of_special_requests			
All	24390	11885	36275
0	11232	8545	19777
1	8670	2703	11373
2	3727	637	4364
3	675	0	675
4	78	0	78
5	8	0	8



0: Not Cancelled Bookings
1: Cancelled Bookings

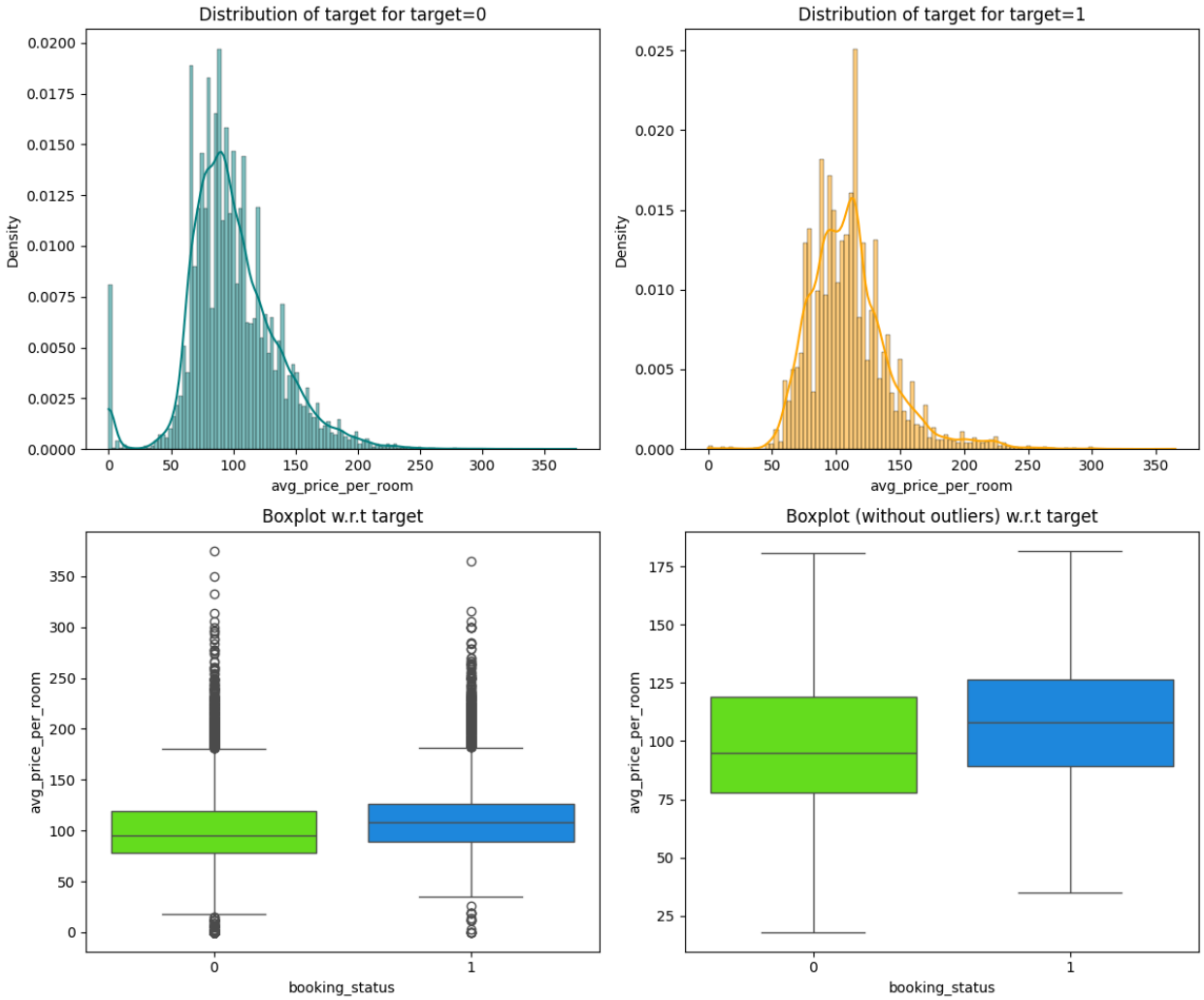
- There are high chances of customers not cancelling the bookings if they had special requests.

No of Special Requests & Average Price Per Room



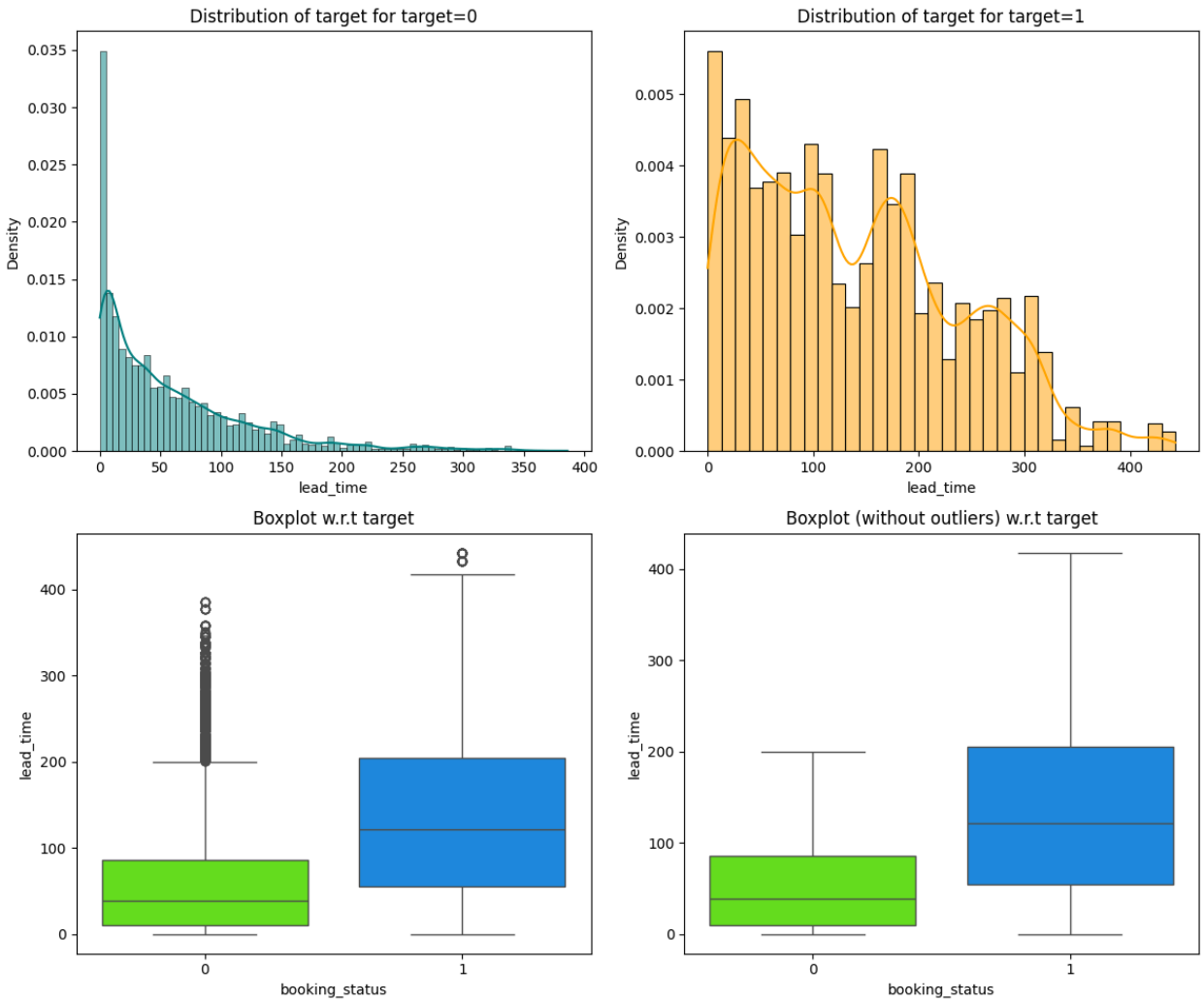
- Rooms with special requests had a higher price compared to the rooms they did not have any special requests for.

Average Price Per Room & Booking Status



- Prices of cancelled bookings are higher than the bookings that were not cancelled.
- Distribution for cancelled and not cancelled bookings is very similar.

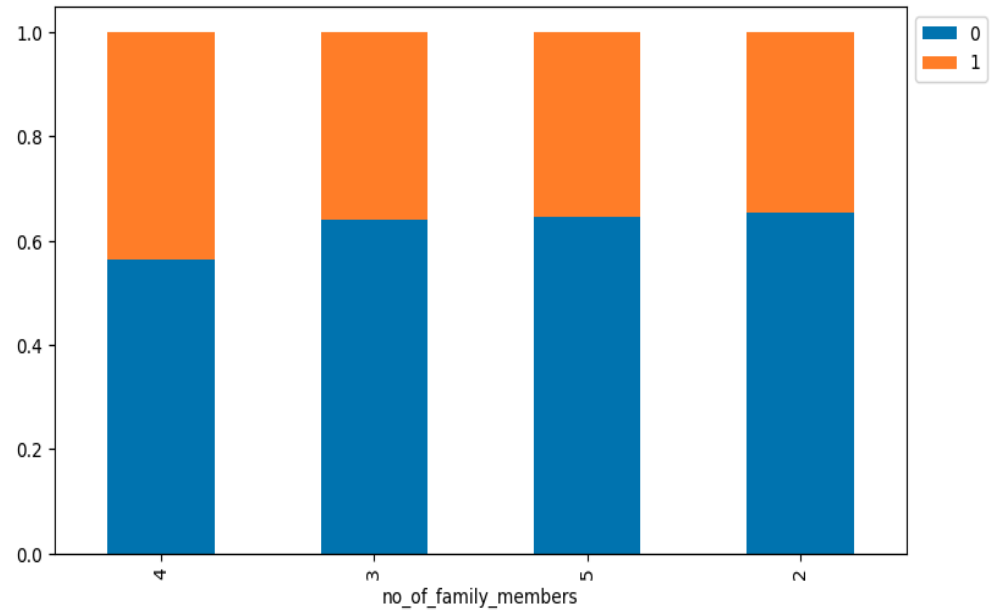
Lead Time & Booking Status



- There is a big difference in the lead time for the bookings that were cancelled and the bookings that were not cancelled.
- Higher lead time results in higher chances of booking cancellation.

No of Family Members & Booking Status

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	11	6	17

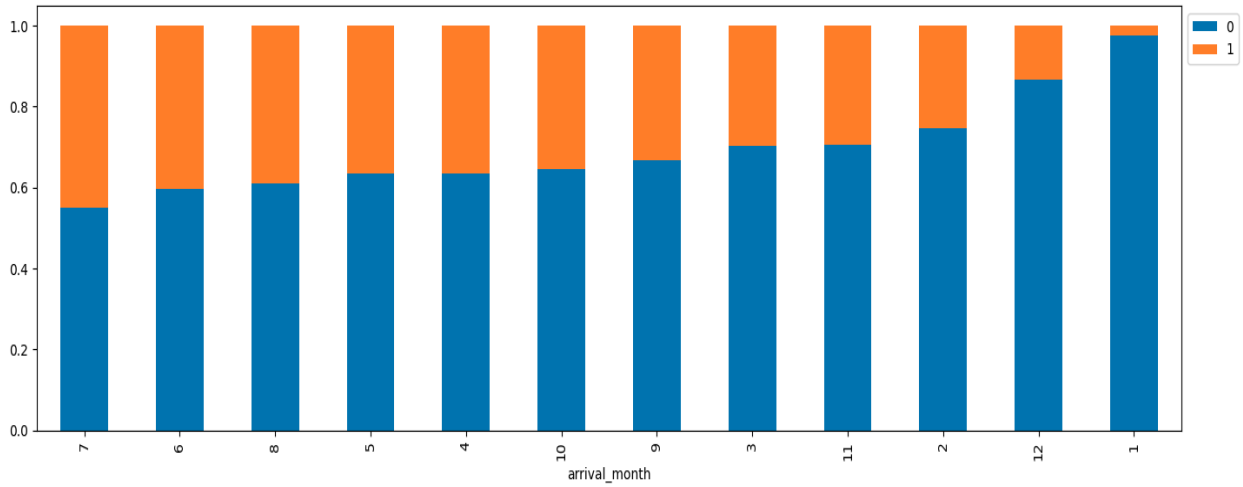


0: Not Cancelled Bookings
1: Cancelled Bookings

- Higher the no of family members the higher the chances of booking cancellation.

Arrival Month & Booking Status

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

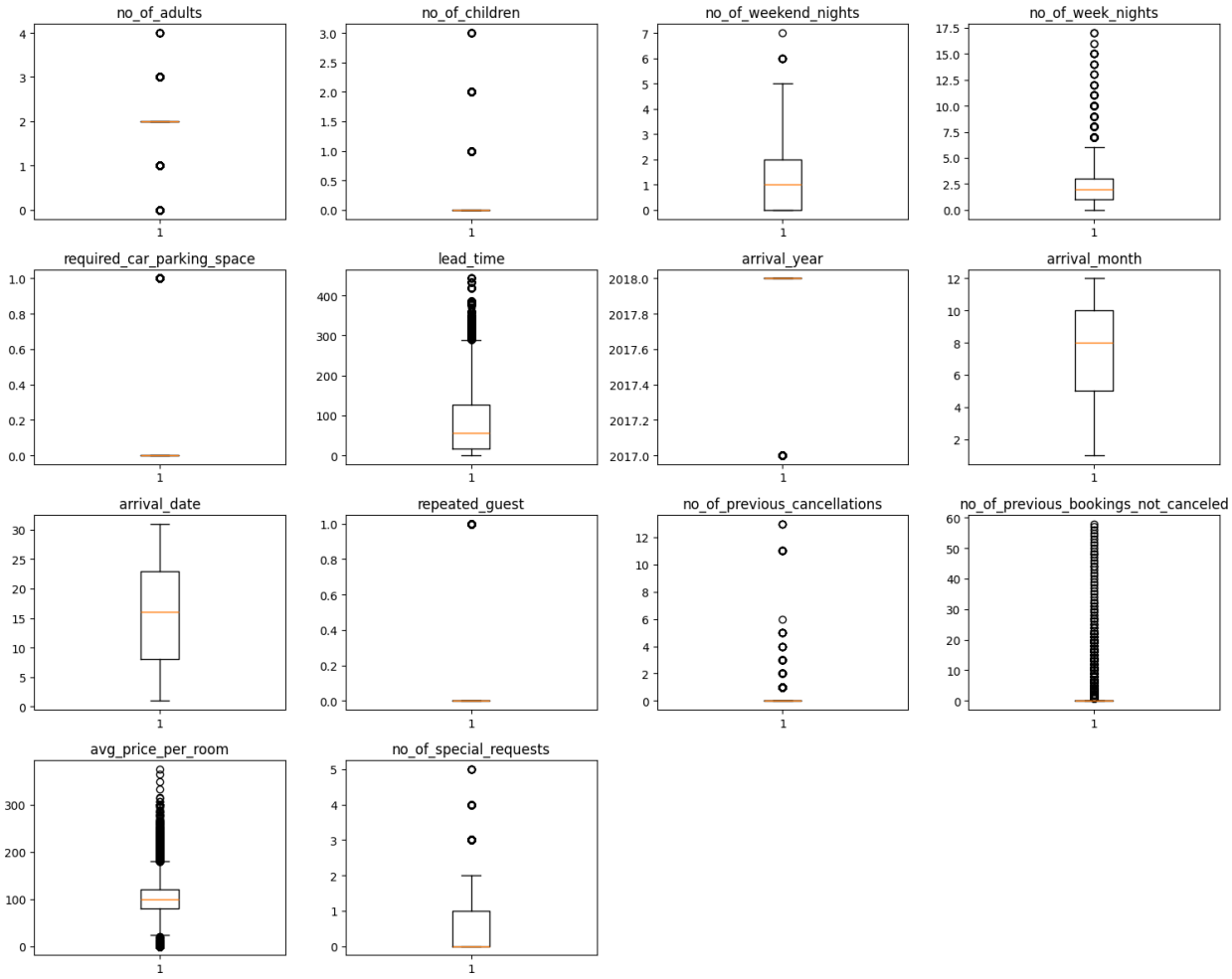


0: Not Cancelled Bookings
1: Cancelled Bookings

- The highest no of booking cancellation was in the month of July.
- The lowest no of booking cancellation was in the month of January.
- While the highest bookings were made in the month of September and October there was a 40% cancellation rate.

Data Preprocessing

Outliers Check



- There are outliers present and we are keeping it as it is important in analyzing the data further.

Model Building

Data Preparation for Modelling

```
Shape of the training set: (25392, 28)
Shape of test set: (10883, 28)
Percentage of Classes in Training Set: booking_status
0    0.670644
1    0.329356
Name: proportion, dtype: float64
Percentage of Classes in Test Set: booking_status
0    0.676376
1    0.323624
Name: proportion, dtype: float64
```

Train Size: 70%
Test Size: 30%

Building the Logistic Regression Model

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25364
Method:	MLE	Df Model:	27
Date:	Sat, 01 Feb 2025	Pseudo R-squ.:	0.3292
Time:	16:53:16	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	-922.8266	120.832	-7.637	0.000	-1159.653	-686.000
no_of_adults	0.1137	0.038	3.019	0.003	0.040	0.188
no_of_children	0.1580	0.062	2.544	0.011	0.036	0.280
no_of_weekend_nights	0.1067	0.020	5.395	0.000	0.068	0.145
no_of_week_nights	0.0397	0.012	3.235	0.001	0.016	0.064
required_car_parking_space	-1.5943	0.138	-11.565	0.000	-1.865	-1.324
lead_time	0.0157	0.000	58.863	0.000	0.015	0.016
arrival_year	0.4561	0.060	7.617	0.000	0.339	0.573
arrival_month	-0.0417	0.006	-6.441	0.000	-0.054	-0.029
arrival_date	0.0005	0.002	0.259	0.796	-0.003	0.004
repeated_guest	-2.3472	0.617	-3.806	0.000	-3.556	-1.139
no_of_previous_cancellations	0.2664	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	25.396	0.000	0.017	0.020
no_of_special_requests	-1.4689	0.030	-48.782	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	17.3584	3987.836	0.004	0.997	-7798.656	7833.373
type_of_meal_plan_Not Selected	0.2784	0.053	5.247	0.000	0.174	0.382
room_type_reserved_Room_Type 2	-0.3605	0.131	-2.748	0.006	-0.618	-0.103
room_type_reserved_Room_Type 3	-0.0012	1.310	-0.001	0.999	-2.568	2.566
room_type_reserved_Room_Type 4	-0.2823	0.053	-5.304	0.000	-0.387	-0.178
room_type_reserved_Room_Type 5	-0.7189	0.209	-3.438	0.001	-1.129	-0.309
room_type_reserved_Room_Type 6	-0.9501	0.151	-6.274	0.000	-1.247	-0.653
room_type_reserved_Room_Type 7	-1.4003	0.294	-4.770	0.000	-1.976	-0.825
market_segment_type_Complementary	-40.5975	5.65e+05	-7.19e-05	1.000	-1.11e+06	1.11e+06
market_segment_type_Corporate	-1.1924	0.266	-4.483	0.000	-1.714	-0.671
market_segment_type_Offline	-2.1946	0.255	-8.621	0.000	-2.694	-1.696
market_segment_type_Online	-0.3995	0.251	-1.590	0.112	-0.892	0.093

- Negative values of the coefficient show that the probability of customers canceling the booking decreases with the increase of the corresponding attribute value.
- Positive values of the coefficient show that the probability of customer canceling increases with the increase of corresponding attribute value.
- p-value of a variable indicates if the variable is significant or not. If we consider the significance level to be 0.05 (5%), then any variable with a p-value less than 0.05 would be considered significant.
- But these variables might contain multicollinearity, which will affect the p-values.
- Multicollinearity from the data needs to be removed to get reliable coefficients and p-values. coefficients and p-values.
- There are different ways of detecting (or testing) multicollinearity, one such way is the Variation Inflation Factor.

Criteria for Model Evaluation

Wrong predictions can be made by the model

1. Model can predict that a customer will not cancel their booking, but in reality the custom will cancel their booking.

➔ In this case the hotel will lose the resources and will have to bear additional cost for filling the room.

2. Model can predict that the custom will cancel their booking, but in reality the custom will not cancel their booking

➔ In this case the reputation of the hotel would get damaged and will make the customer mad.

- To reduce the loss and damage to the hotel's reputation the hotel the F1 score should be maximized meaning greater the F1 score the lower the chances of getting False Negatives and False Positives.

Training Performance

	Accuracy	Recall	Precision	F1
0	0.806002	0.634103	0.739713	0.682848

Multicollinearity

	Features	VIF
0	const	3.949769e+07
1	no_of_adults	1.351135e+00
2	no_of_children	2.093583e+00
3	no_of_weekend_nights	1.069484e+00
4	no_of_week_nights	1.095711e+00
5	required_car_parking_space	1.039972e+00
6	lead_time	1.395175e+00
7	arrival_year	1.431904e+00
8	arrival_month	1.276334e+00
9	arrival_date	1.006795e+00
10	repeated_guest	1.783576e+00
11	no_of_previous_cancellations	1.395693e+00
12	no_of_previous_bookings_not_canceled	1.652000e+00
13	avg_price_per_room	2.068603e+00
14	no_of_special_requests	1.247981e+00
15	type_of_meal_plan_Meal Plan 2	1.273283e+00
16	type_of_meal_plan_Meal Plan 3	1.025258e+00
17	type_of_meal_plan_Not Selected	1.273060e+00
18	room_type_reserved_Room_Type 2	1.105954e+00
19	room_type_reserved_Room_Type 3	1.003303e+00
20	room_type_reserved_Room_Type 4	1.363606e+00
21	room_type_reserved_Room_Type 5	1.028000e+00
22	room_type_reserved_Room_Type 6	2.056136e+00
23	room_type_reserved_Room_Type 7	1.118156e+00
24	market_segment_type_Complementary	4.502756e+00
25	market_segment_type_Corporate	1.692829e+01
26	market_segment_type_Offline	6.411564e+01
27	market_segment_type_Online	7.118026e+01

- VIF for dummy variables is ignored.
- Numerical variables does not show high or moderate multicollinearity.
- Predictor variables with a p-value of greater than 0.05 will be dropped as they do not significantly impact the target variable.
- Not all the p-values are dropped as once as sometimes p-values change after dropping a variable.
- A new model is then created without the dropped feature and the steps are repeated until there are no p-values greater than 0.05.

New Logistic Regression Model After Treatment of Multicollinearity

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25370
Method:	MLE	Df Model:	21
Date:	Sat, 01 Feb 2025	Pseudo R-squ.:	0.3282
Time:	16:53:34	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	-915.6391	120.471	-7.600	0.000	-1151.758	-679.520
no_of_adults	0.1088	0.037	2.914	0.004	0.036	0.182
no_of_children	0.1531	0.062	2.470	0.014	0.032	0.275
no_of_weekend_nights	0.1086	0.020	5.498	0.000	0.070	0.147
no_of_week_nights	0.0417	0.012	3.399	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	59.213	0.000	0.015	0.016
arrival_year	0.4523	0.060	7.576	0.000	0.335	0.569
arrival_month	-0.0425	0.006	-6.591	0.000	-0.055	-0.030
repeated_guest	-2.7367	0.557	-4.916	0.000	-3.828	-1.646
no_of_previous_cancellations	0.2288	0.077	2.983	0.003	0.078	0.379
avg_price_per_room	0.0192	0.001	26.336	0.000	0.018	0.021
no_of_special_requests	-1.4698	0.030	-48.884	0.000	-1.529	-1.411
type_of_meal_plan_Meal Plan 2	0.1642	0.067	2.469	0.014	0.034	0.295
type_of_meal_plan_Not Selected	0.2860	0.053	5.406	0.000	0.182	0.390
room_type_reserved_Room_Type 2	-0.3552	0.131	-2.709	0.007	-0.612	-0.098
room_type_reserved_Room_Type 4	-0.2828	0.053	-5.330	0.000	-0.387	-0.179
room_type_reserved_Room_Type 5	-0.7364	0.208	-3.535	0.000	-1.145	-0.328
room_type_reserved_Room_Type 6	-0.9682	0.151	-6.403	0.000	-1.265	-0.672
room_type_reserved_Room_Type 7	-1.4343	0.293	-4.892	0.000	-2.009	-0.860
market_segment_type_Corporate	-0.7913	0.103	-7.692	0.000	-0.993	-0.590
market_segment_type_Offline	-1.7854	0.052	-34.363	0.000	-1.887	-1.684

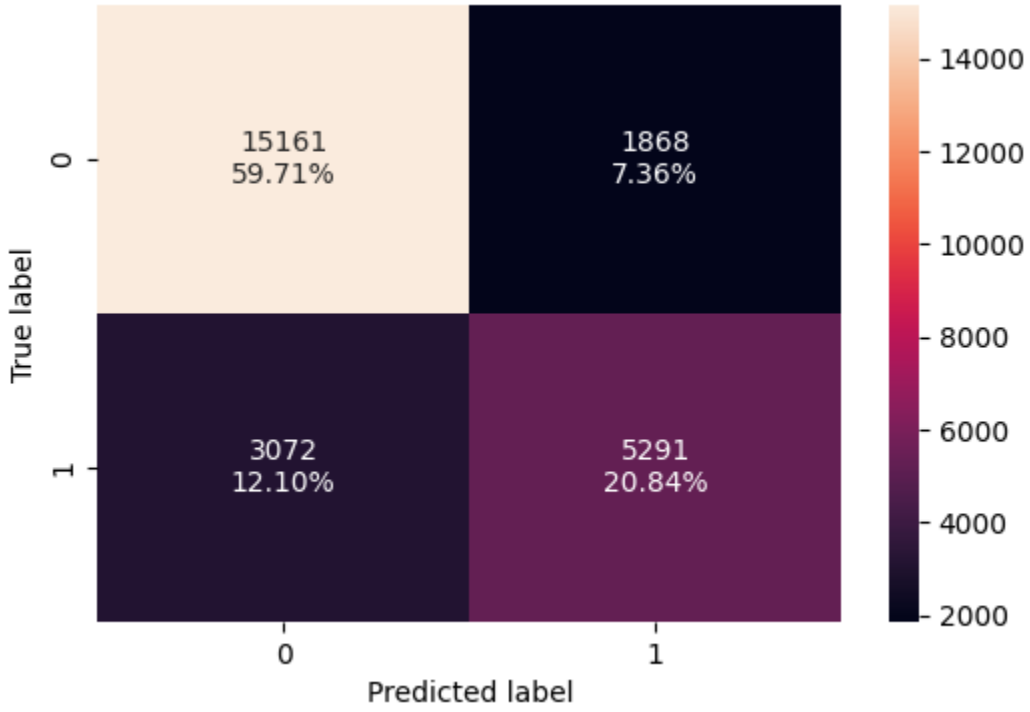
- All the variables left has a p-value less than 0.05 hence this is the best model for making any inference.
- Also the performance on the training data is the same as before dropping the variables with p-value greater than 0.05.

Interpretation of Coefficient

- The likelihood of a customer cancelling their reservation will reduce if the coefficients for required_car_parking_space, arrival_month, repeated_guest, no_of_special_requests, and a few others grow from negative values.
- There are positive coefficients for the number of adults, children, weekend evenings, weeknights, lead time, average price per room, kind of meal plan not selected, and a few more; a higher number of these will enhance the likelihood that a customer will cancel their reservation.

Checking the Model Performance on the Train Set

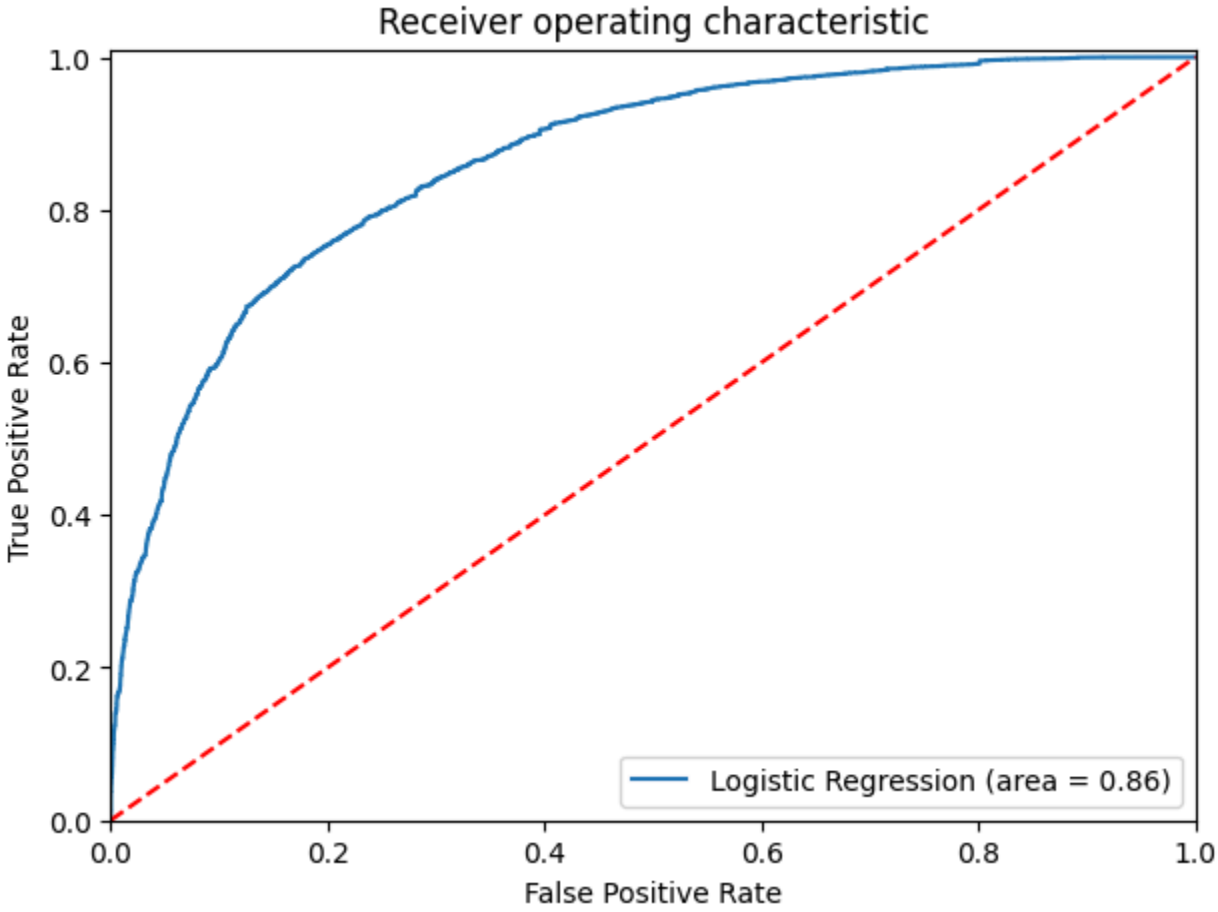
Confusion Matrix Train Set



Training Performance

	Accuracy	Recall	Precision	F1
0	0.805451	0.632668	0.73907	0.681742

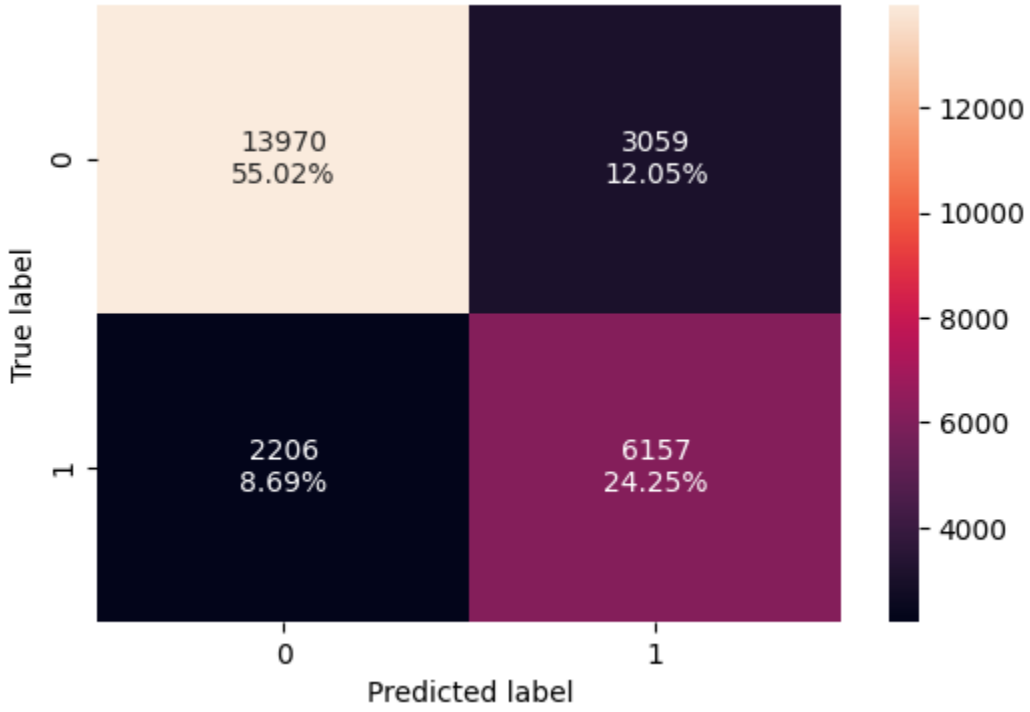
ROC-AUC Curve



- ROC-AUC curve score of 0.86 on training is good.
- The Logistic Regression model is giving a generalized performance on both the training and the testing set.

Model Performance Improvement

Checking if the recall score can be improved by changing the model threshold using the AUC-ROC Curve.



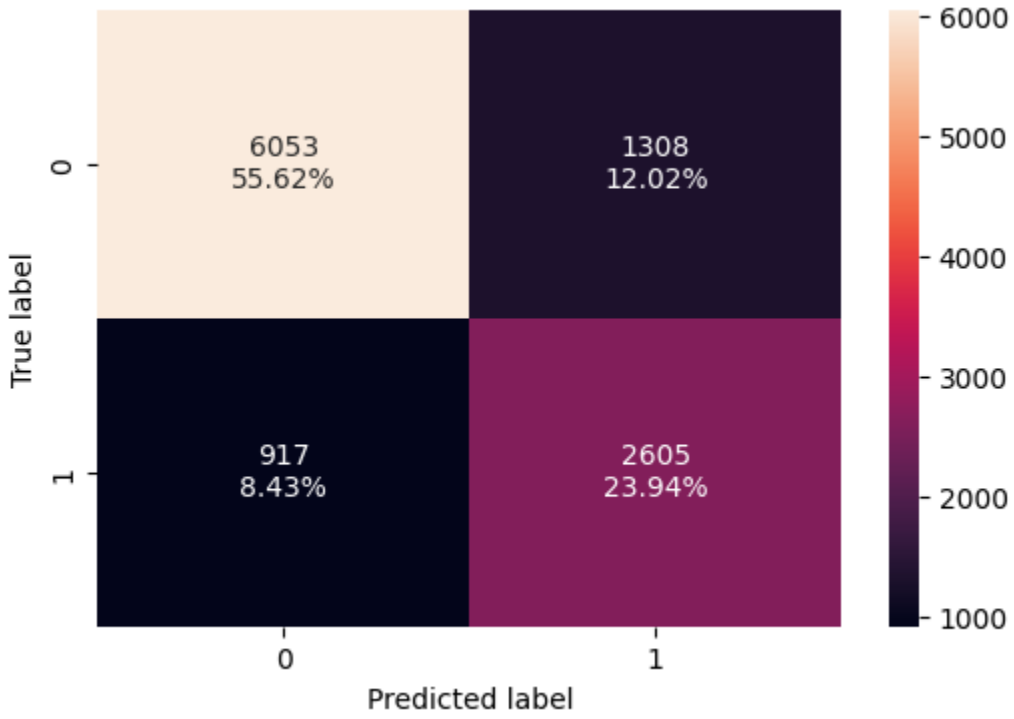
Training performance:

	Accuracy	Recall	Precision	F1
0	0.792651	0.736219	0.668077	0.700495

- Recall has increased compared to the previous model.
- As the threshold keeps decreasing the recall will keep on increasing and the precision will decrease hence an optimal balance between the both need to be reached.

Checking the Model Performance on the Test Set

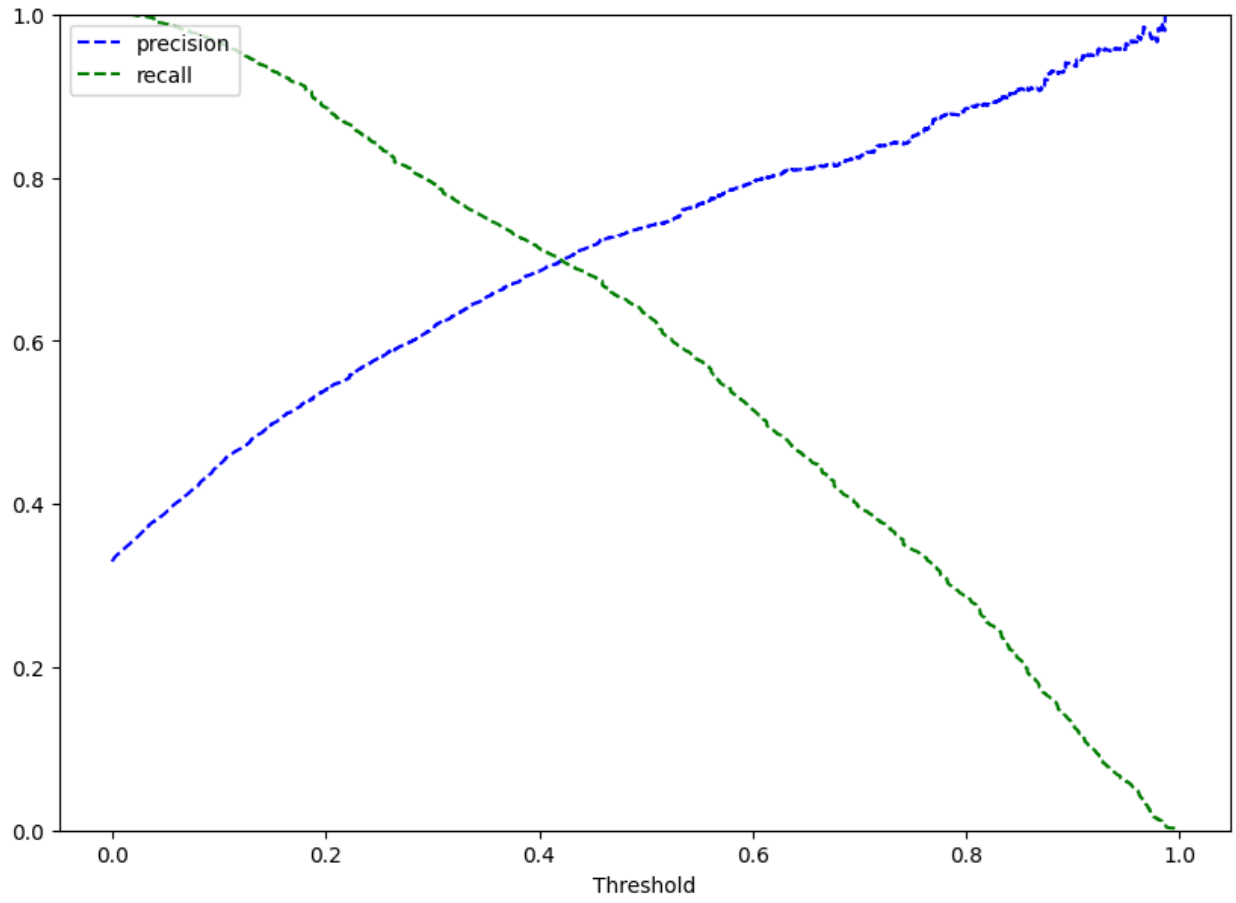
Confusion Matrix Test Set



Test performance:

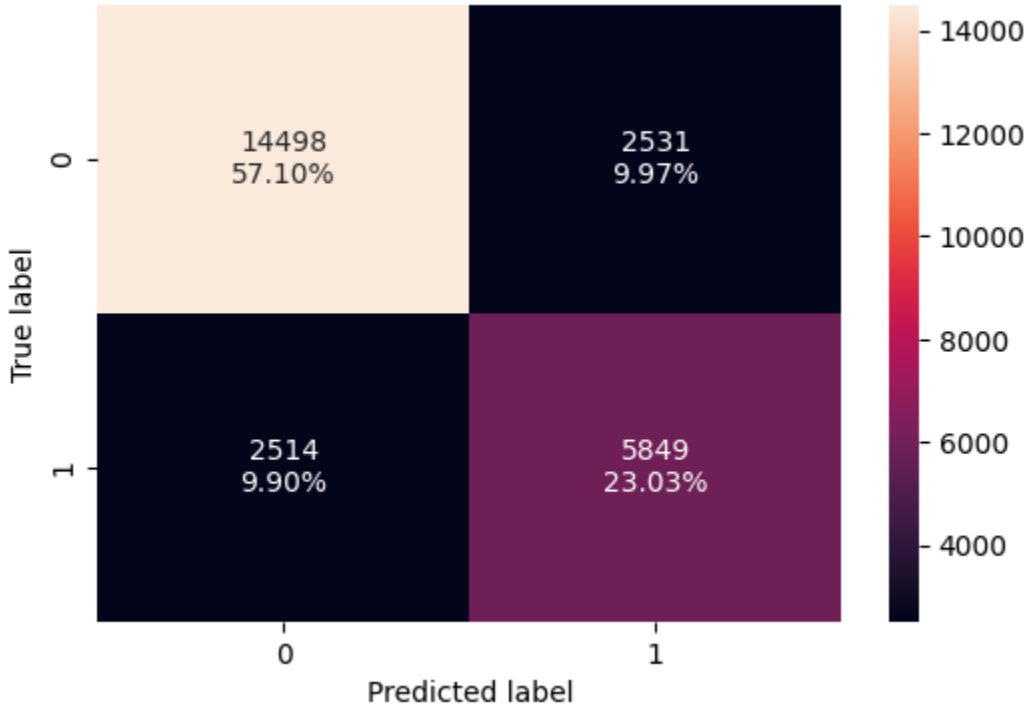
	Accuracy	Recall	Precision	F1
0	0.795553	0.739637	0.66573	0.70074

Precision Recall Curve



- At a threshold of 0.42 there is a balance between precision and recall.

Checking model performance on training set with 0.42 as threshold

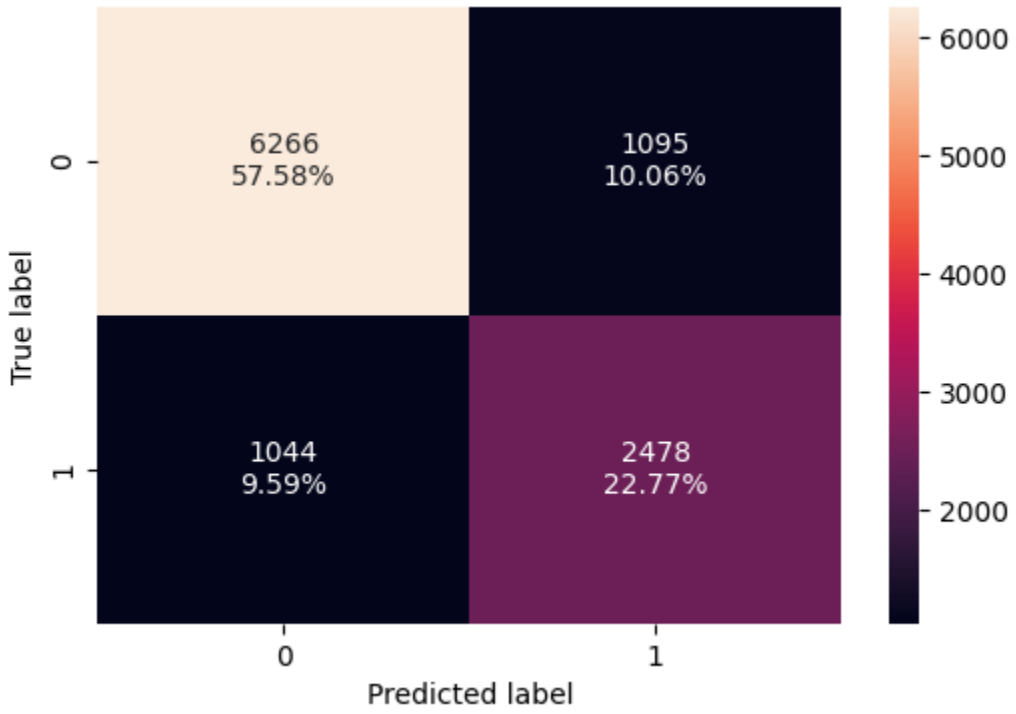


Training performance:

	Accuracy	Recall	Precision	F1
0	0.801315	0.69939	0.697971	0.69868

- The performance of the model has improved compared to the initial model.
- There is a balanced performance in terms of precision and recall.

Checking Model Performance on Test Set



Test performance:

	Accuracy	Recall	Precision	F1
0	0.803455	0.703578	0.693535	0.69852

Model Performance Summary

Training Performance Comparison

	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.805451	0.792651	0.801315
Recall	0.632668	0.736219	0.699390
Precision	0.739070	0.668077	0.697971
F1	0.681742	0.700495	0.698680

Testing Performance Comparison

	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.804649	0.795553	0.803455
Recall	0.630892	0.739637	0.703578
Precision	0.729003	0.665730	0.693535
F1	0.676408	0.700740	0.698520

- With an F1 score of 0.69 on the training set, we have developed a predictive model that the hotel can use to identify which reservations are most likely to be cancelled.
- On both training and test sets, the logistic regression models provide a generalised performance.
- The hotel can predict which reservations won't be cancelled and will be able to offer those customers satisfactory services, which helps to maintain brand equity but will cost money. This is achieved by using a model with a default threshold, which will yield a low recall but good precision score.
- The hotel can save money by accurately forecasting the bookings that are likely to be cancelled, but doing so could harm its brand equity. The model with a 0.37 threshold will provide a high recall but poor precision score.
- The hotel will be able to preserve a balance between resources and brand equity by using the model with a 0.42 threshold, which will provide a balanced recall and precision score.
- Required_car_parking_space, arrival_month, repeated_guest, no_of_special_requests, and a few other coefficients are negative; if they grow, the likelihood of a consumer cancelling their reservation will decrease.

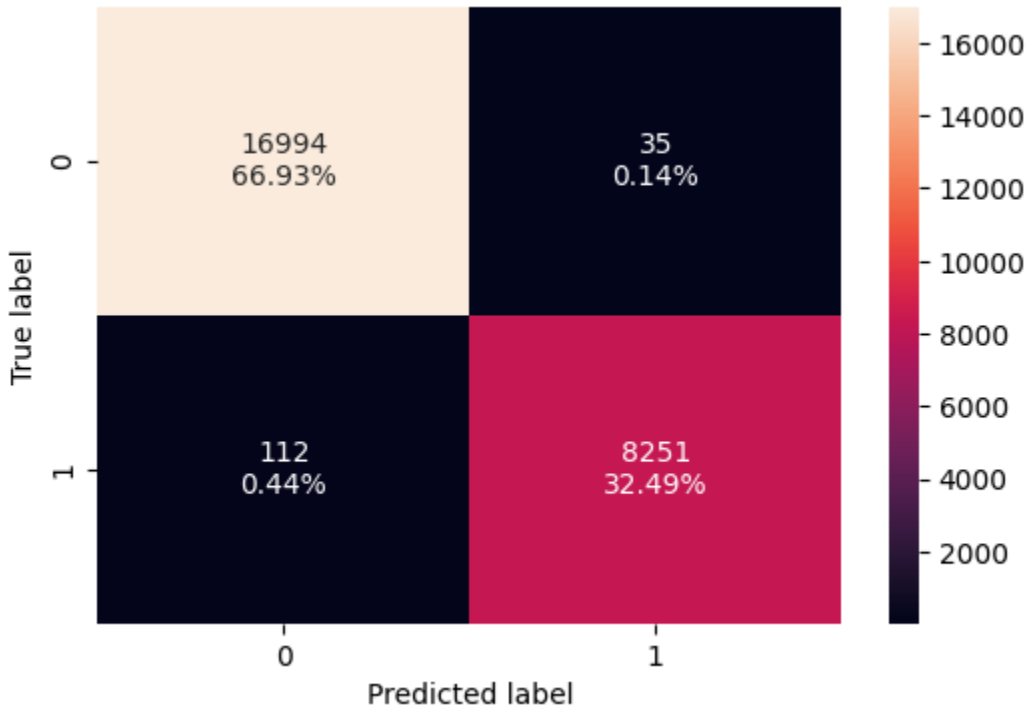
Decision Tree

Shape of Training and Testing Set

```
Shape of Training set :  (25392, 27)
Shape of test set :  (10883, 27)
Percentage of classes in training set: booking_status
0    0.670644
1    0.329356
Name: proportion, dtype: float64
Percentage of classes in test set: booking_status
0    0.676376
1    0.323624
Name: proportion, dtype: float64
```

Building the Decision Tree Model

Checking Model Performance on Training Set

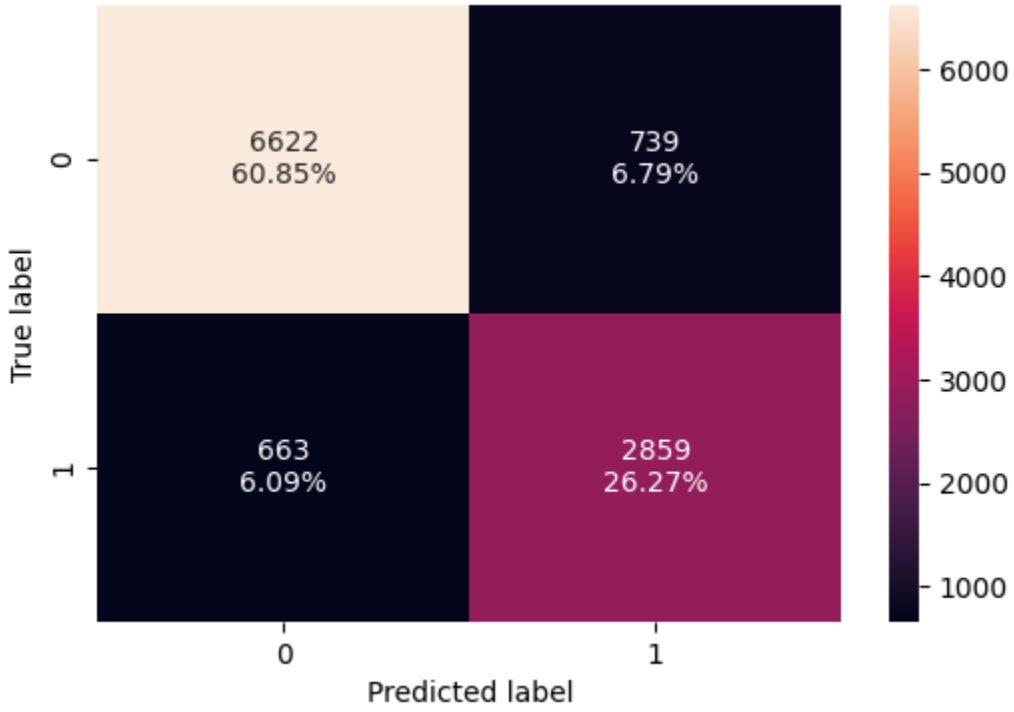


Training Performance

	Accuracy	Recall	Precision	F1
0	0.994211	0.986608	0.995776	0.991171

- There are 0 errors in the training set as each sample has been classified correctly.
- Model performance is good on the training set.

Checking Model Performance on Testing Set

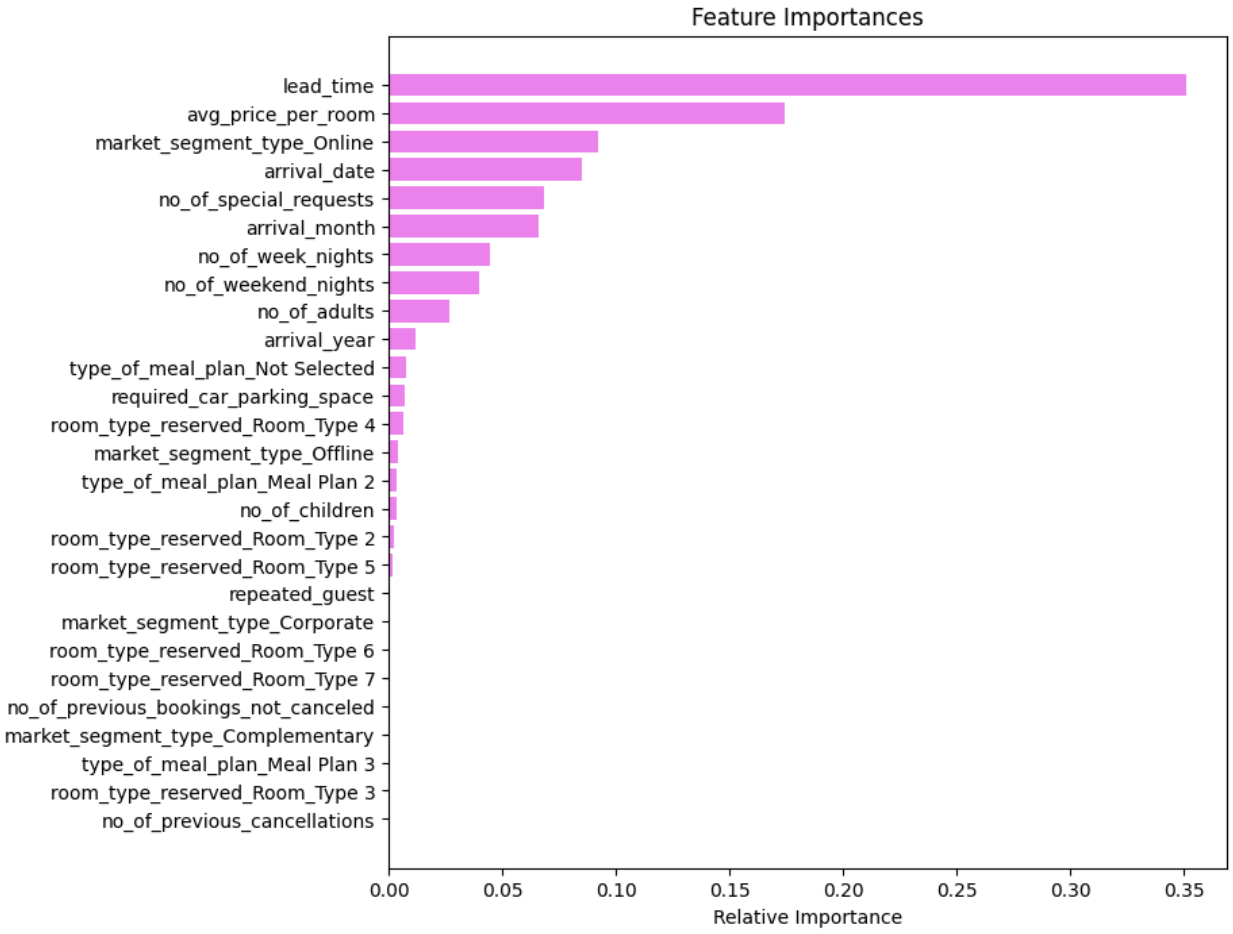


Testing Performance

	Accuracy	Recall	Precision	F1
0	0.871175	0.811755	0.794608	0.80309

- The decision tree model is overfitting the data and is not able to generalize well on the test set.
- Pruning of the decision tree is required.

Important Features



- Lead time is the most important factor followed by average price per room.

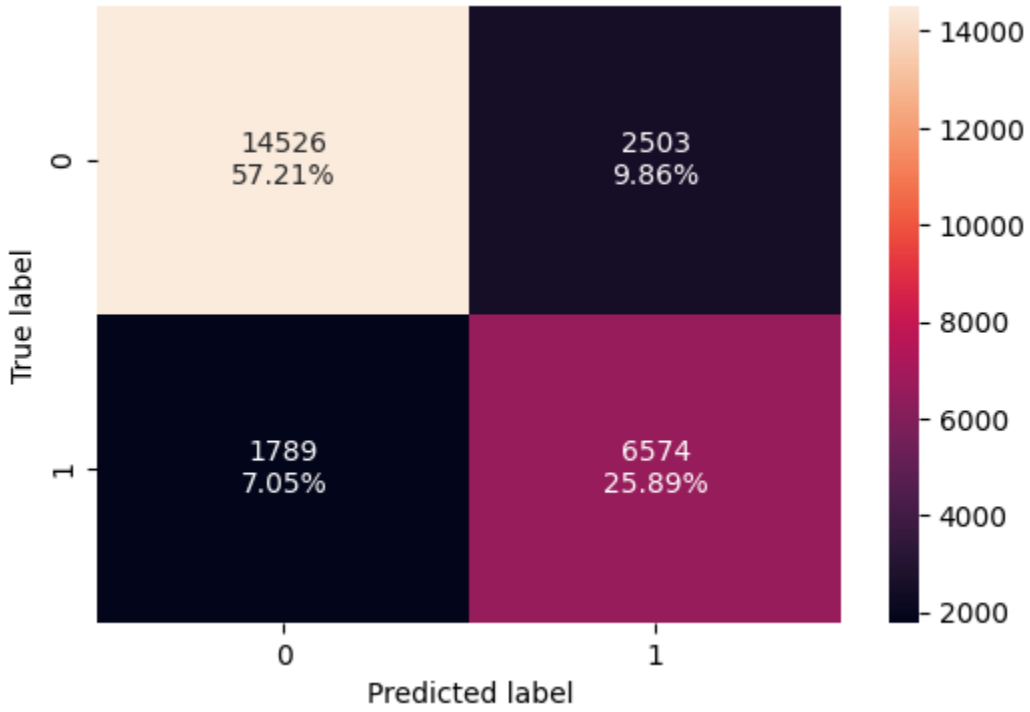
Pruning the Tree

Pre-Pruning

```
DecisionTreeClassifier(class_weight='balanced', max_depth=6, max_leaf_nodes=50,
                       min_samples_split=10, random_state=1)
```

Checking Performance on Training Set

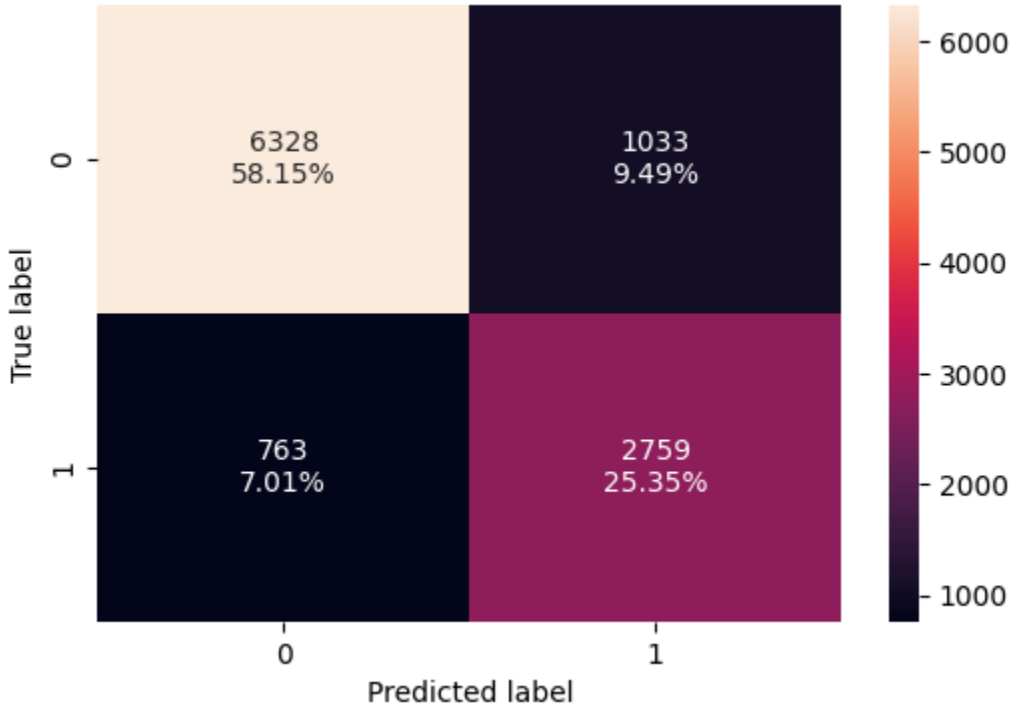
Training Set Confusion Matrix



Training Performance

	Accuracy	Recall	Precision	F1
0	0.83097	0.786082	0.724248	0.753899

Checking Performance on Testing Set

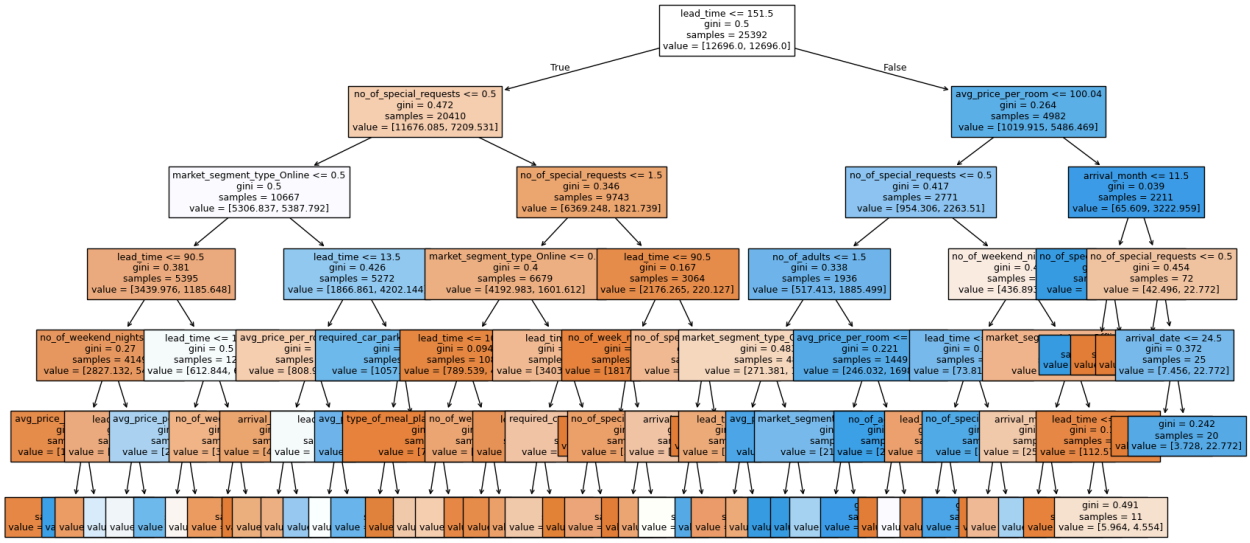


Testing Performance

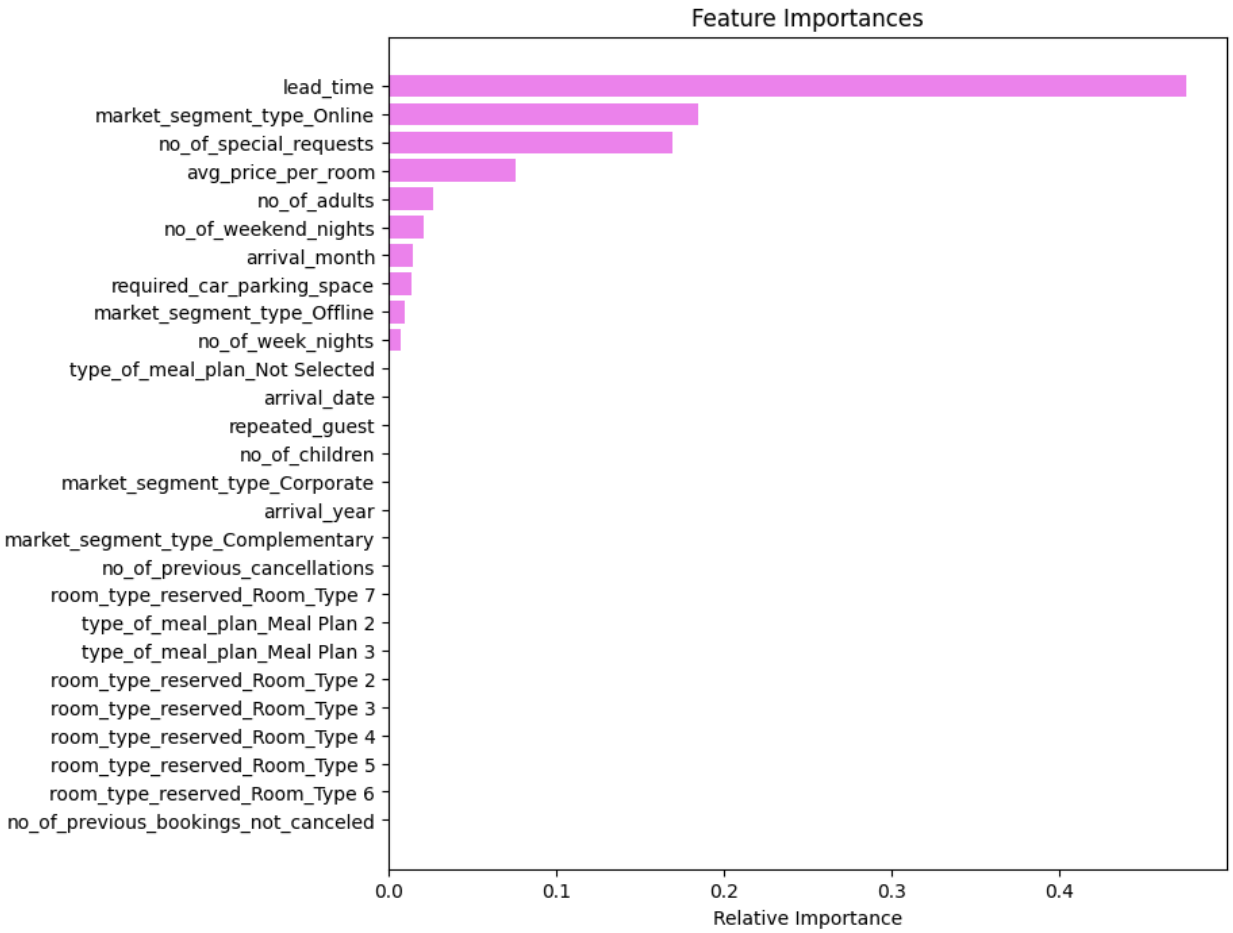
	Accuracy	Recall	Precision	F1
0	0.834972	0.783362	0.727584	0.754444

Visualizing the Decision Tree

Decision Tree



Importance of Features in the Decision Tree



- The tree has become simpler and the rules of the trees are readable.
- Model performance of the model is generalized.
- The most important factors are:
    - Lead time
    - Market segment - Online
    - No of special request
    - Avg price per room
- The rules show that lead time plays a key role in identifying if a booking will be cancelled or not. 151 days has been considered as a threshold value by the model to make the first split.

Bookings made more than 151 days before the date of arrival:

- If the average price per room is greater than 100 euros and the arrival month is December, then the booking is less likely to be cancelled.

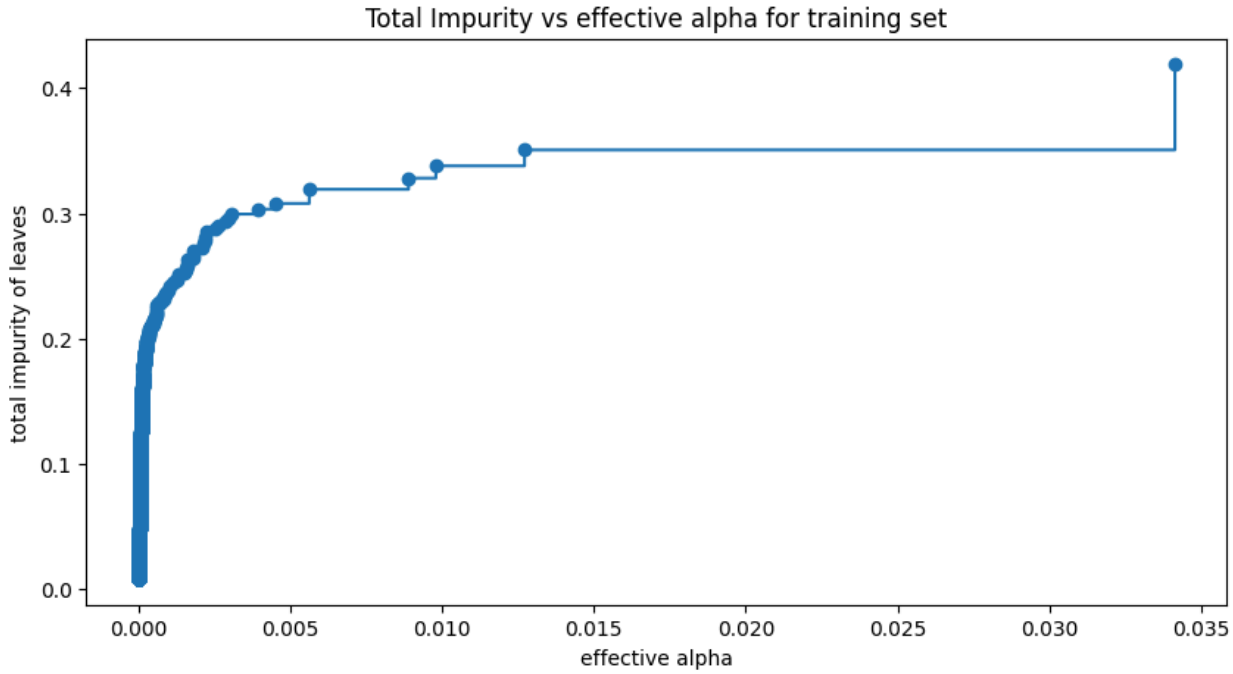
- If the average price per room is less than or equal to 100 euros and the number of special requests is 0, then the booking is likely to get canceled.

Bookings made under 151 days before the date of arrival:

- If a customer has at least 1 special request the booking is less likely to be cancelled.
- If the customer didn't make any special requests and the booking was done Online it is more likely to get cancelled, if the booking was not done online, it is less likely to be cancelled.

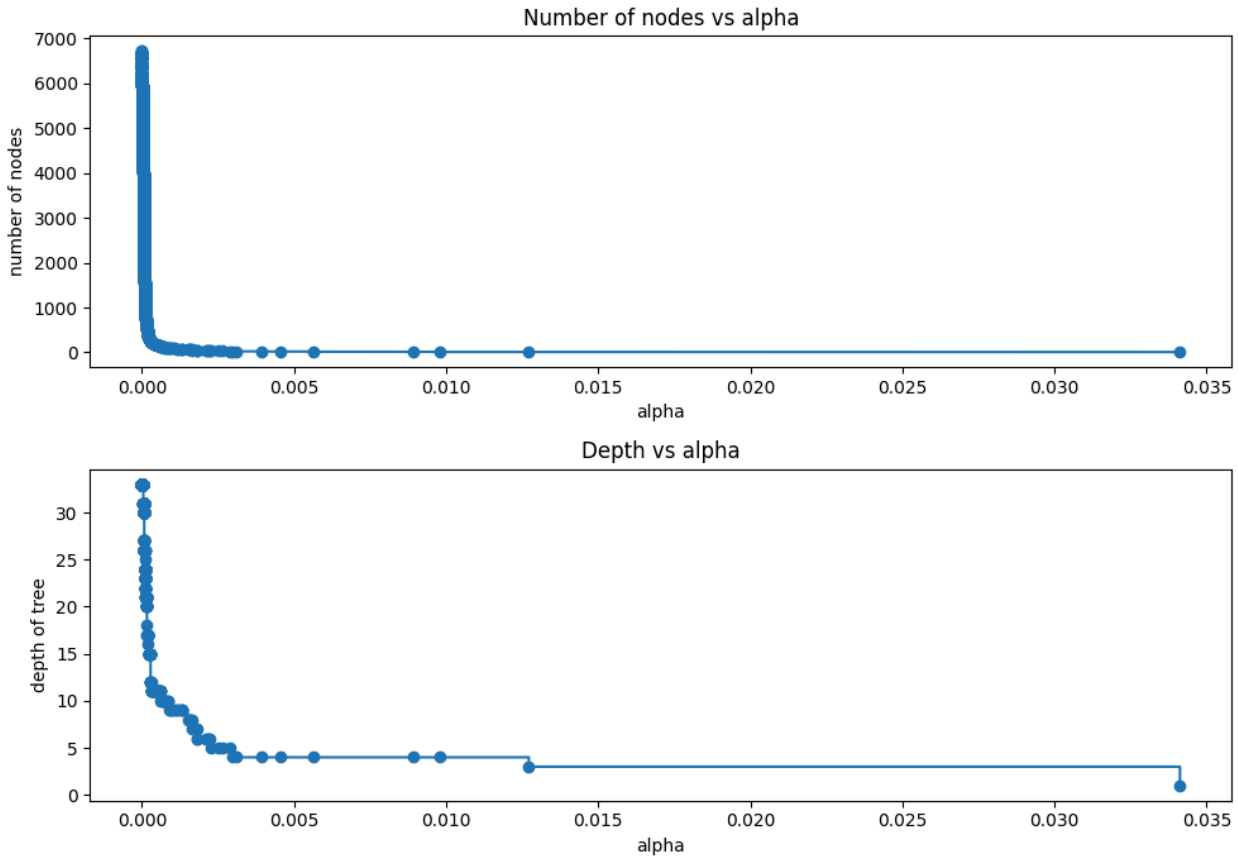
Cost Complexity Pruning

Total Impurity vs Effective Alpha(Training Set)

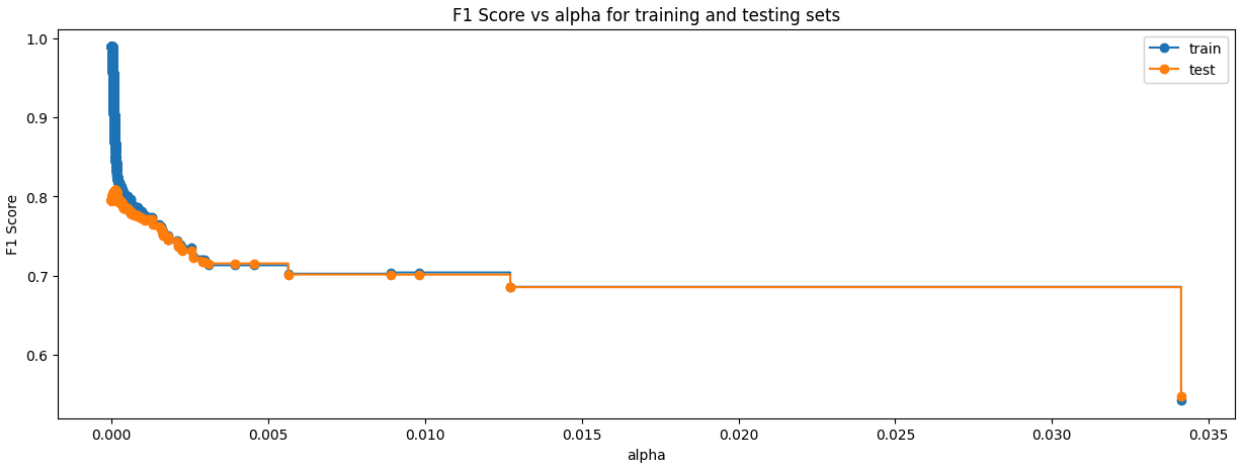


- Training the decision tree using effective alphas. The last value in ccp_alphas is the alpha value that prunes the whole tree, leaving the tree, clfs[-1], with one node.

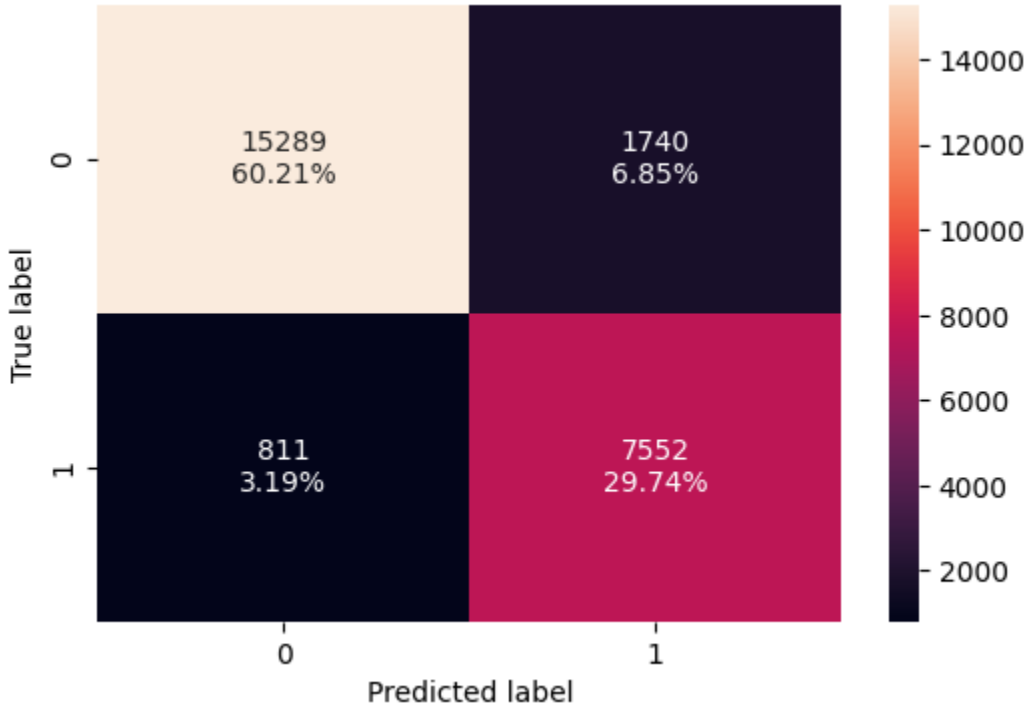
No of Nodes vs Alpha & Depth vs Alpha



F1 Score vs Alpha for Training & Testing Sets



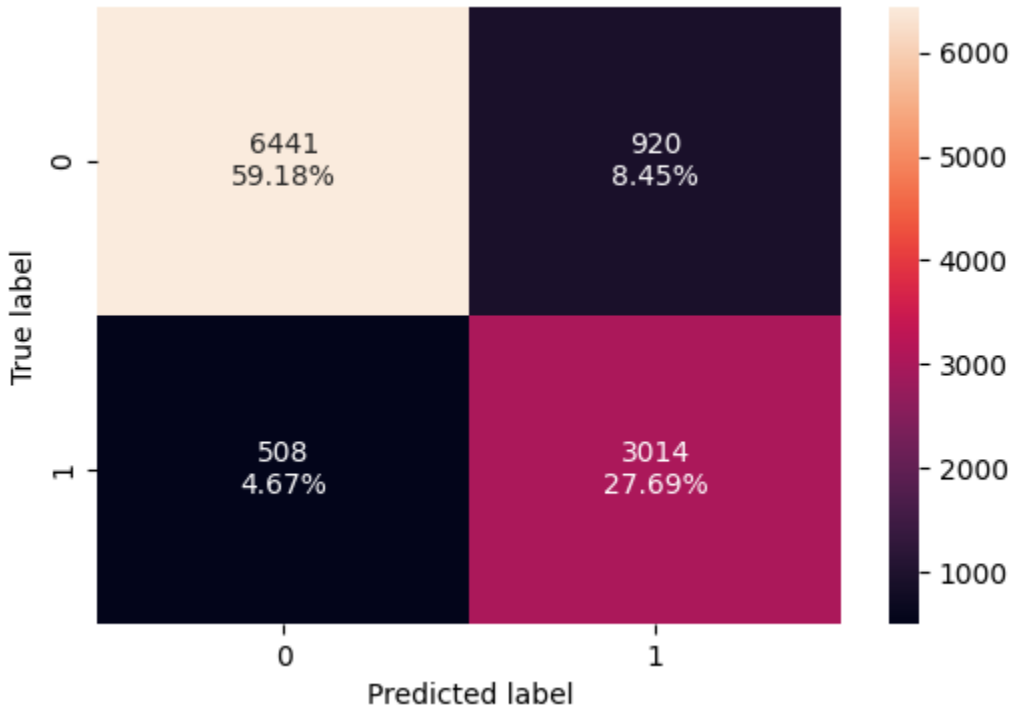
Checking Performance on Training Set



Training Performance

	Accuracy	Recall	Precision	F1
0	0.899535	0.903025	0.812742	0.855508

Checking Importance on Testing Set

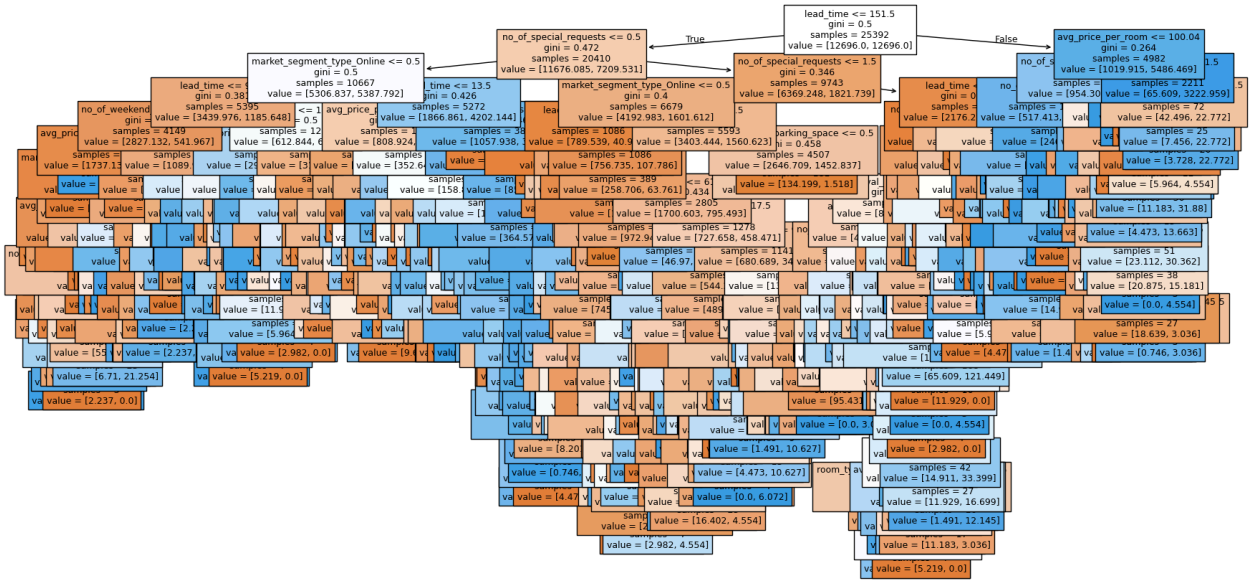


Testing Performance

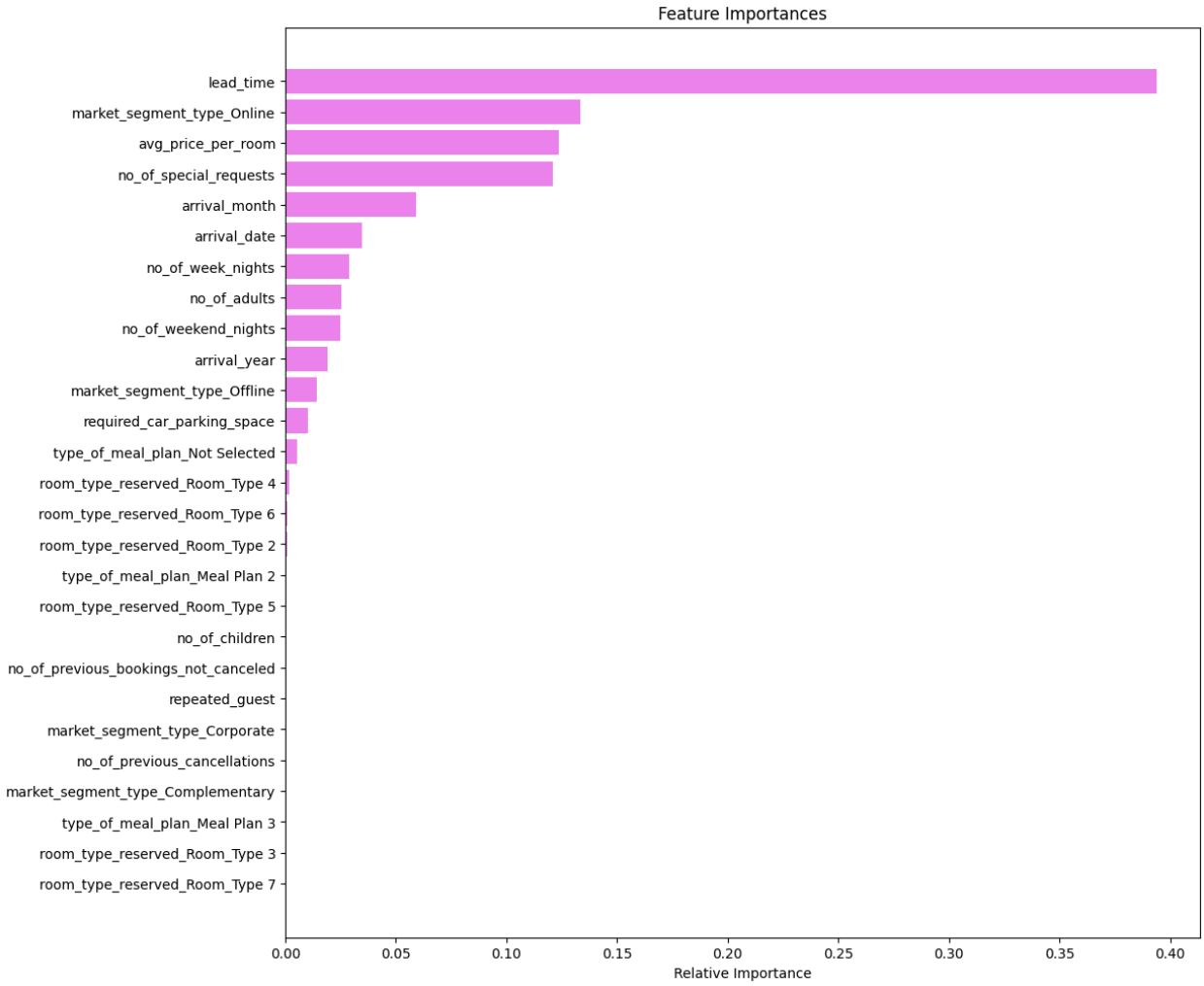
	Accuracy	Recall	Precision	F1
0	0.899535	0.903025	0.812742	0.855508

- Post pruning the decision tree the performance has been generalized on both training and testing sets.
- The recall is high in this model and also the difference between recall and precision has increased.

Decision Tree Post Pruning



Feature Importance



- Tree is very complex compared to the pre-pruned tree.
- Feature importance is the same as the pre-pruned tree.

Comparing the Decision Tree Models

Training Performance Comparison

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.994211	0.830970	0.899535
Recall	0.986608	0.786082	0.903025
Precision	0.995776	0.724248	0.812742
F1	0.991171	0.753899	0.855508

Testing Performance Comparison

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.871175	0.834972	0.899535
Recall	0.811755	0.783362	0.903025
Precision	0.794608	0.727584	0.812742
F1	0.803090	0.754444	0.855508

- The decision tree model's default settings cause it to overfit the training set and perform poorly in general.
- With pre-pruned trees, there is a generalised performance with balanced precision and recall values.
- When compared to other models, the post-pruned tree has a high F1 score; nonetheless, the precision and recall differences are significant.
- The hotel will be able to balance resources and brand equity by using the pre-pruned decision tree model.

Actionable Insights

- Considering key variables based on p-values in the Decision Tree model and p-values in the Logistic regression model.
- Lead time, the quantity of special requests, and the average cost per room are significant in both models.
- According to the Logistic Regression model, the average price per room and lead time are positively correlated with cancellations of reservations. Additionally, there is a negative correlation between the quantity of specific requests and cancelled reservations.

Business Recommendations

- The hotel should implement stricter cancellation policies: Reservations with specific demands and high average room prices shouldn't receive a complete refund because there will be a significant loss in resource. Although it would be ideal for cancellation procedures to be uniform across all market groups, the data shows that a significant portion of online reservations are cancelled. Customers that cancel their reservations online should receive a less amount of refund.
- When determining whether or not to cancel a reservation, the lead time and the quantity of specific requests made by the client are crucial factors. Reservations made with a special request from a customer and made less than 151 days prior to the arrival date are less likely to be cancelled.
- Before the customers' arrival date, set up a system that can automatically send them an email asking for a confirmation of their reservation and any adjustments they would like to make.
- December and January have a low cancellation to noncancellation ratio. Consumers may take trips to enjoy the holidays and the new year. The hotel should make sure that there are adequate staff members on hand to meet the needs of the visitors.
- The busiest months for bookings and cancellations were October and September. The hotel needs to look into this more thoroughly.
- According to our analysis, there aren't many repeat clients, and those that do exist have very low cancellation rates, which is encouraging because repeat business is crucial to the hospitality sector because it can help spread the word. Compared to returning consumers, attracting new ones is time-consuming and more expensive.
- These consumers' experience can be enhanced by a loyalty program that provides them with exclusive discounts, hotel amenities, etc.