INN Hotels- Logistic Regression/Decision Tree Project

UT Data Science & Business Analytics

May 2023

Contents / Agenda

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

Executive Summary

- Lead time is a major factor in customers cancelling their reservations. Hotels could place a limit on the extent of how far out a customer can book their reservation to help cut down cancellations or hotels could implement a fee for cancellations to help deter customers from cancelling.

- Staff could offer more special requests available to customers since the data shows that customers who had special requests were less likely to cancel.

Business Problem Overview and Solution Approach

- A significant number of hotel bookings are called off due to cancellations or no-shows. The typical reasons for cancellations include change of plans, scheduling conflicts, etc. This is often made easier by the option to do so free of charge or preferably at a low cost which is beneficial to hotel guests but it is a less desirable and possibly revenue-diminishing factor for hotels to deal with. Such losses are particularly high on last-minute cancellations.

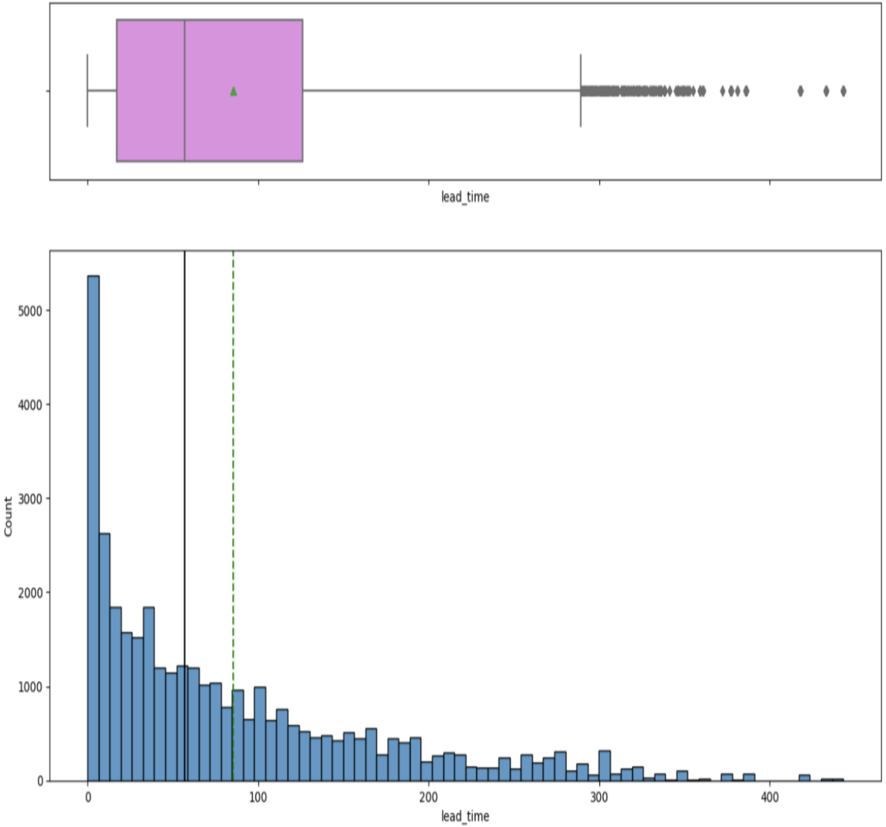
- The new technologies involving online booking channels have dramatically changed customers' booking possibilities and behavior. This adds a further dimension to the challenge of how hotels handle cancellations, which are no longer limited to traditional booking and guest characteristics.
- The cancellation of bookings impact a hotel on various fronts:
- Loss of resources (revenue) when the hotel cannot resell the room.
- Additional costs of distribution channels by increasing commissions or paying for publicity to help sell these rooms.
- Lowering prices last minute, so the hotel can resell a room, resulting in reducing the profit margin.
- Human resources to make arrangements for the guests.

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

- There are 36,275 rows and 19 columns in the dataset.
- No duplicate values are present.
- Booking ID column has been dropped from the data frame.

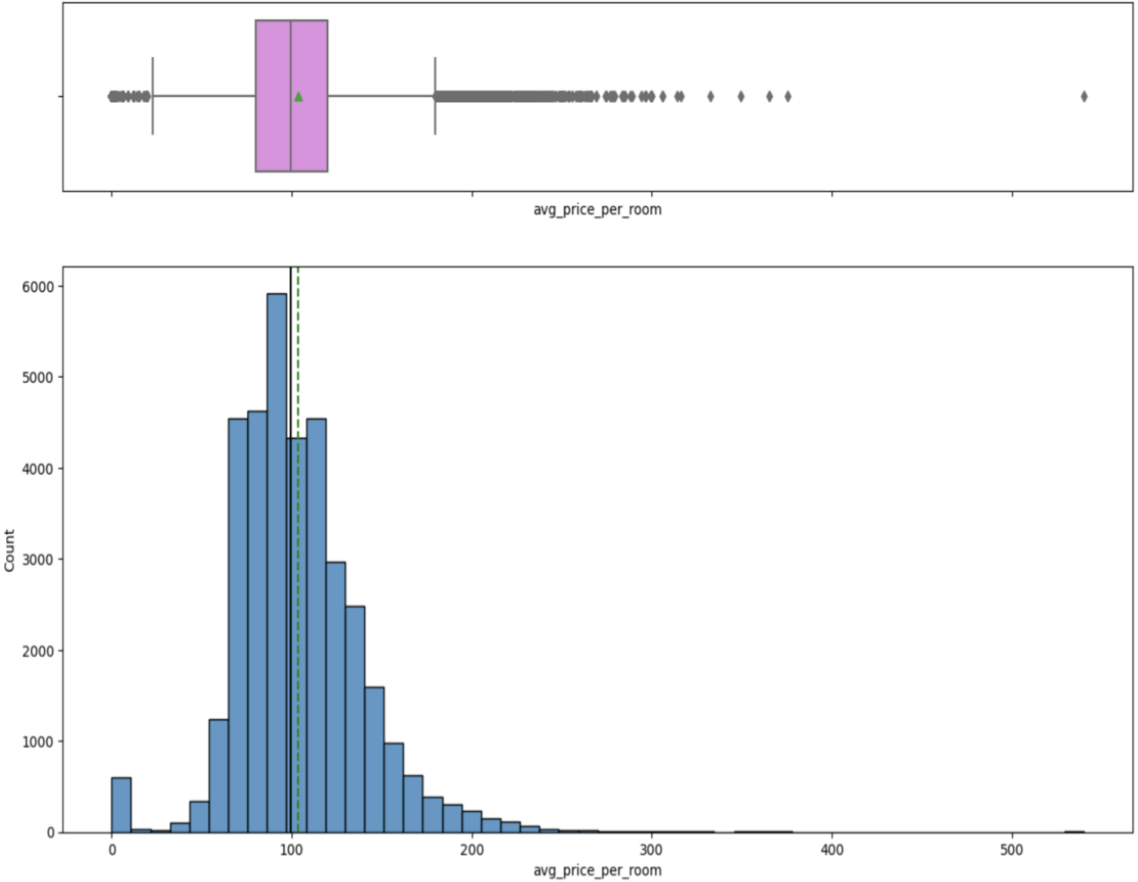
EDA- Lead Time



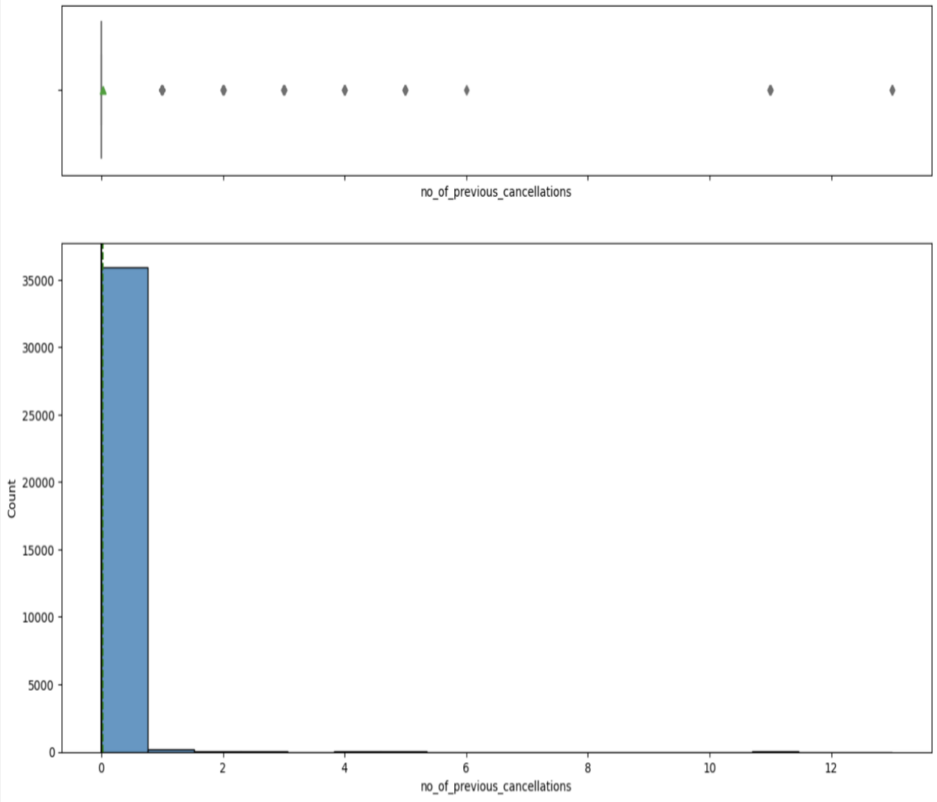
- Lead_time : Number of days between the date of booking and the arrival date.
- Mean lead time is ~85 days
- 25% are below 17 days, 50% are below 57 days, and 75% are below 126 days.
- The distribution is right-skewed.
- Outliers are present.

EDA- Average Price Per Room

- The mean average price per room is ~ 103 euros
- 25% are < 80 euros
- 50% are < 99 euros
- 75% are < 120 euros



EDA- Number of previous cancellations



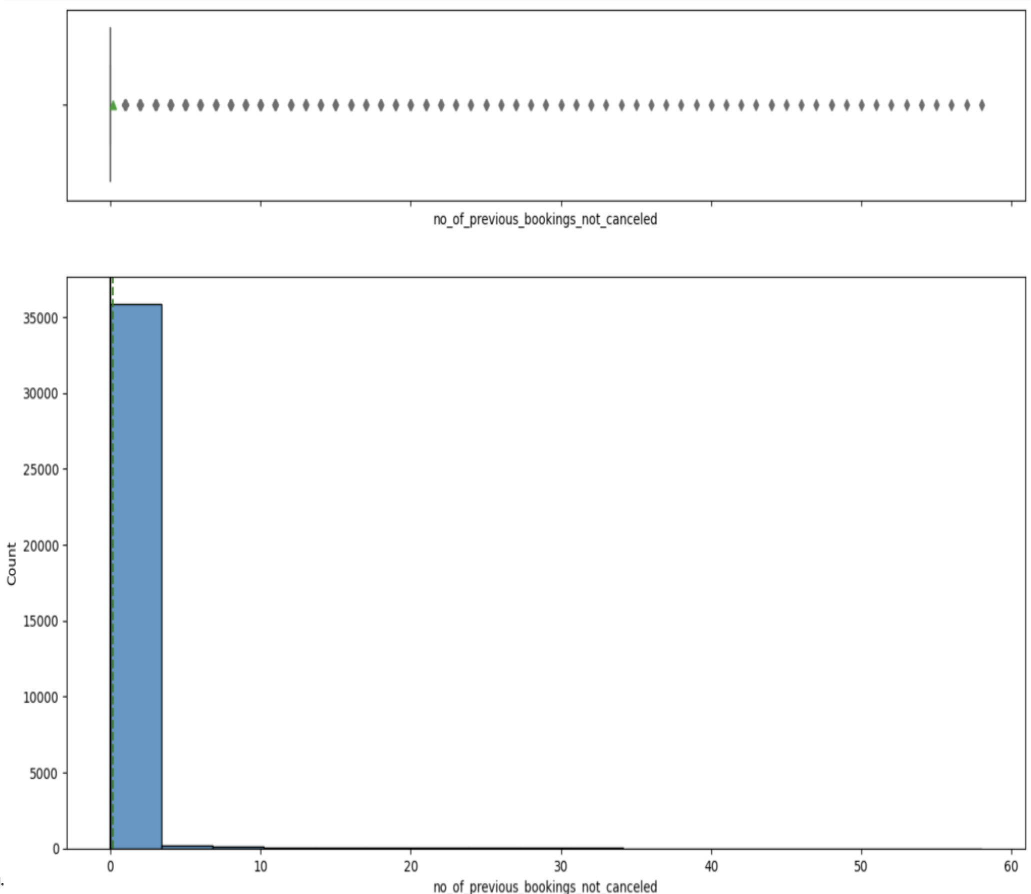
- 75% of the customers had never had a previous cancellation.
- The distribution is right-skewed.

EDA- Number of previous bookings not cancelled

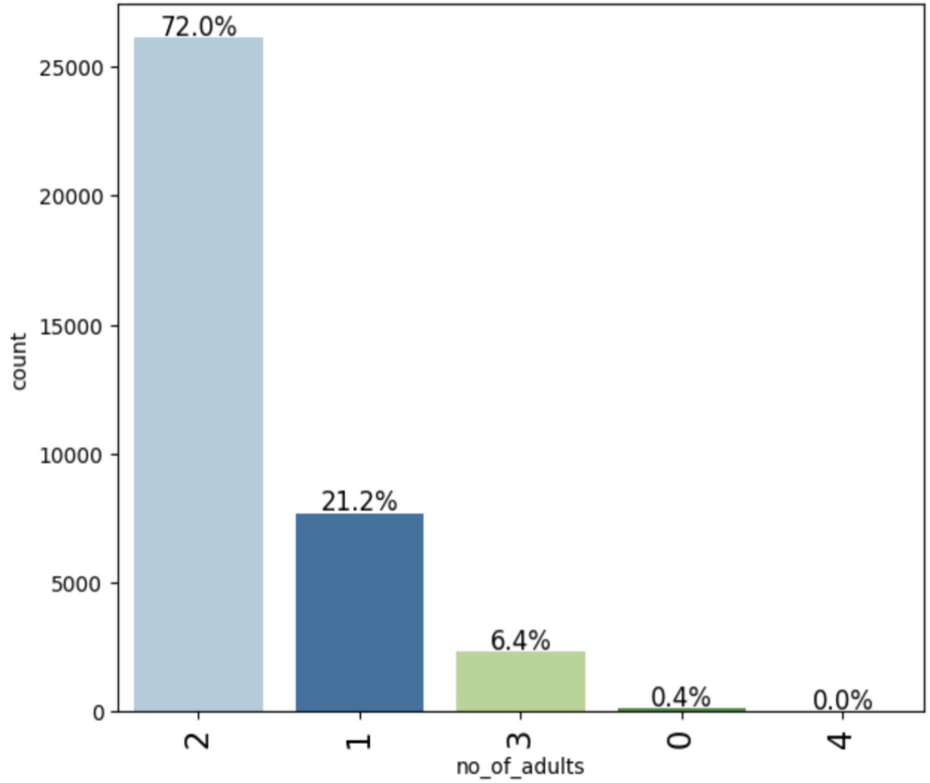
- Mean ~ .15

- 75% of customers have never had
A previous booking cancelled

- The distribution is right-skewed.



EDA- Number of Adults



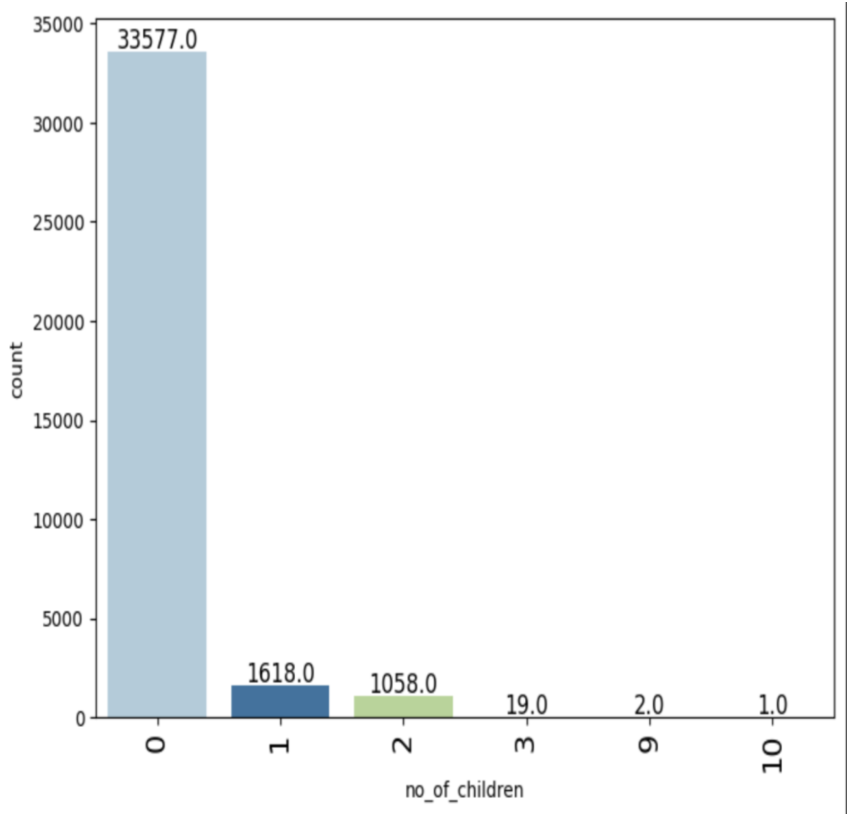
- 72% of guests have 2 adults staying

EDA- Number of Children

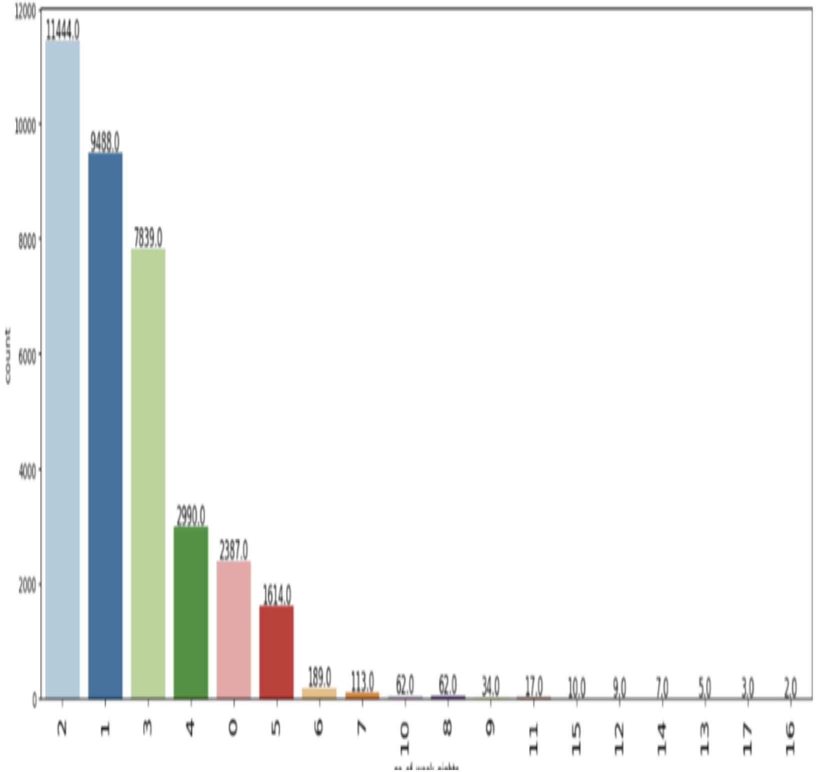
- Most of the guests staying had no children.

- 33,577 of the bookings had no children, and very few had children at all.

- Travelling with children is more expensive in general, which may contribute to the number of bookings including children.



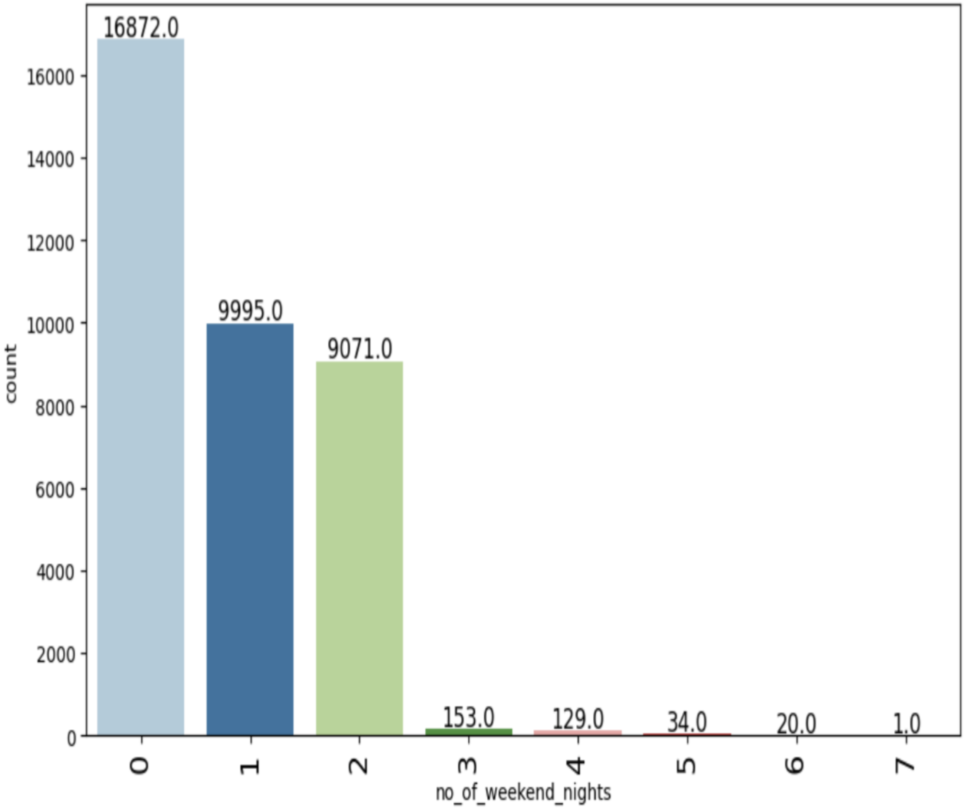
EDA- Number of Week nights



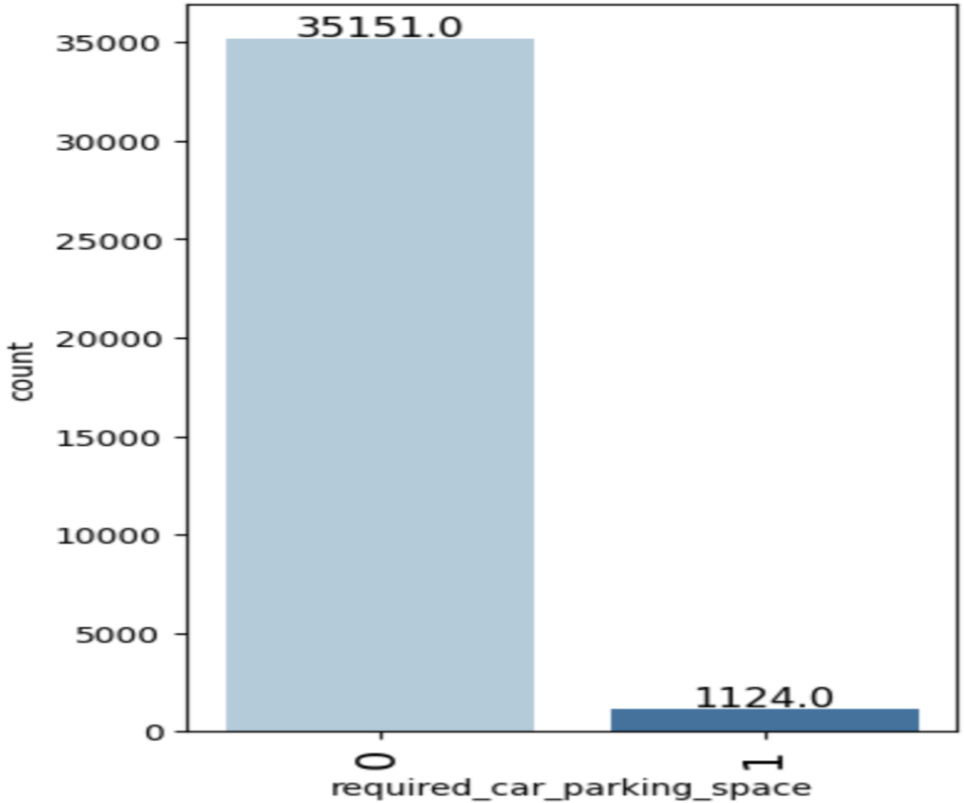
- Most guests only book for 2 week-nights in the hotel.
- Distribution is right-skewed

EDA- Number of Weekend Nights

- The distribution is right-skewed.
- Almost half of the guests didn't spend the weekend nights at the hotel.



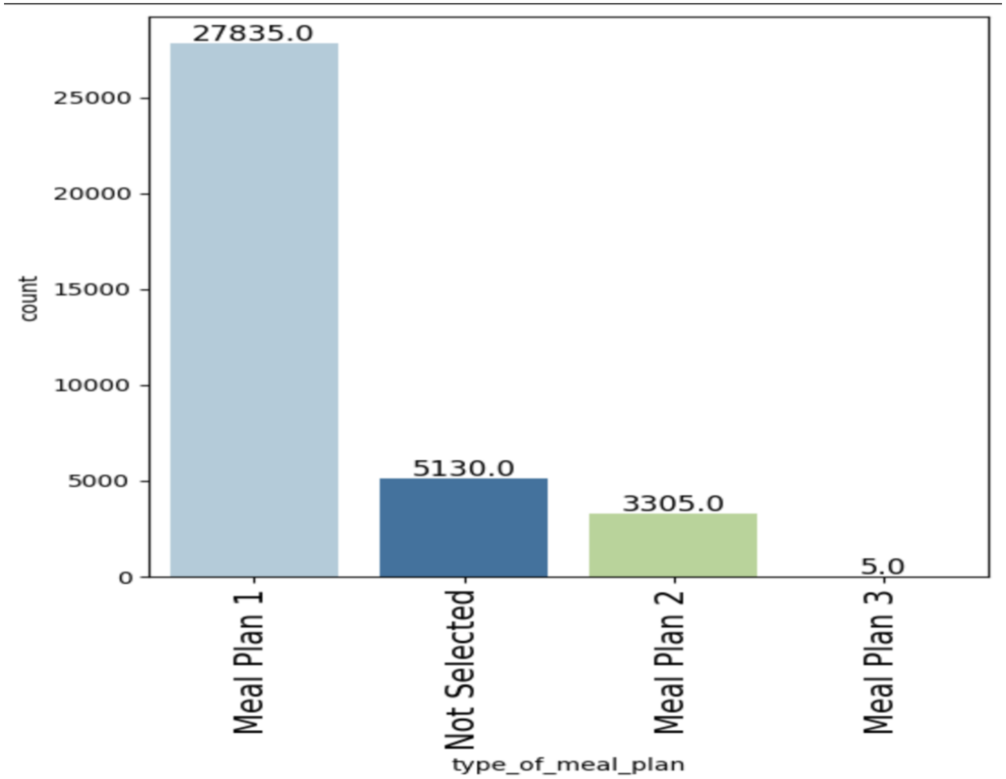
EDA- Required Car Parking Space



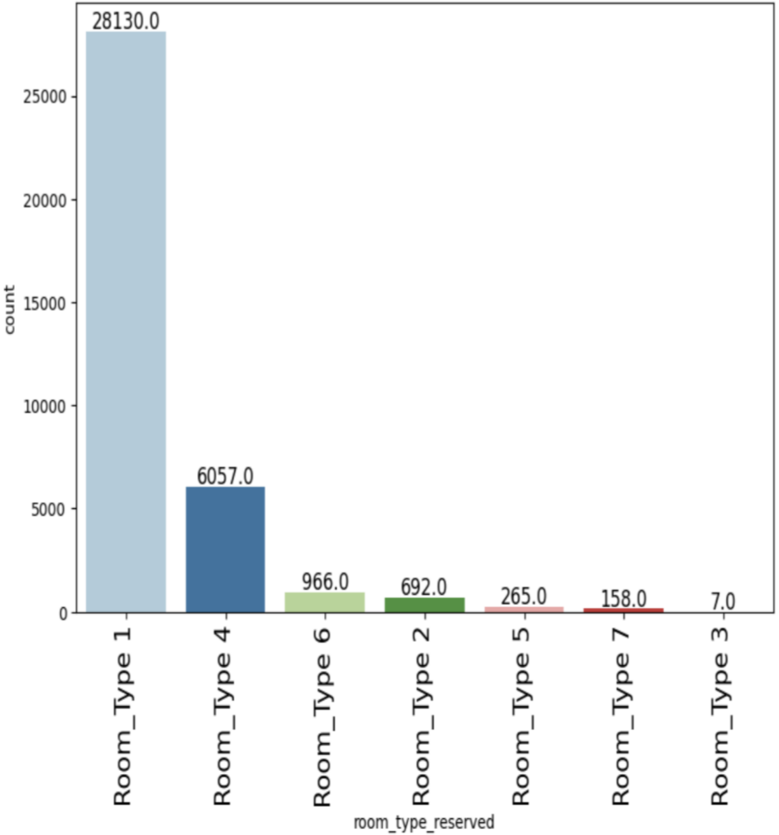
- Most guests require no parking space

EDA- Type of Meal Plan

- Meal Plan 1 is the most requested meal.



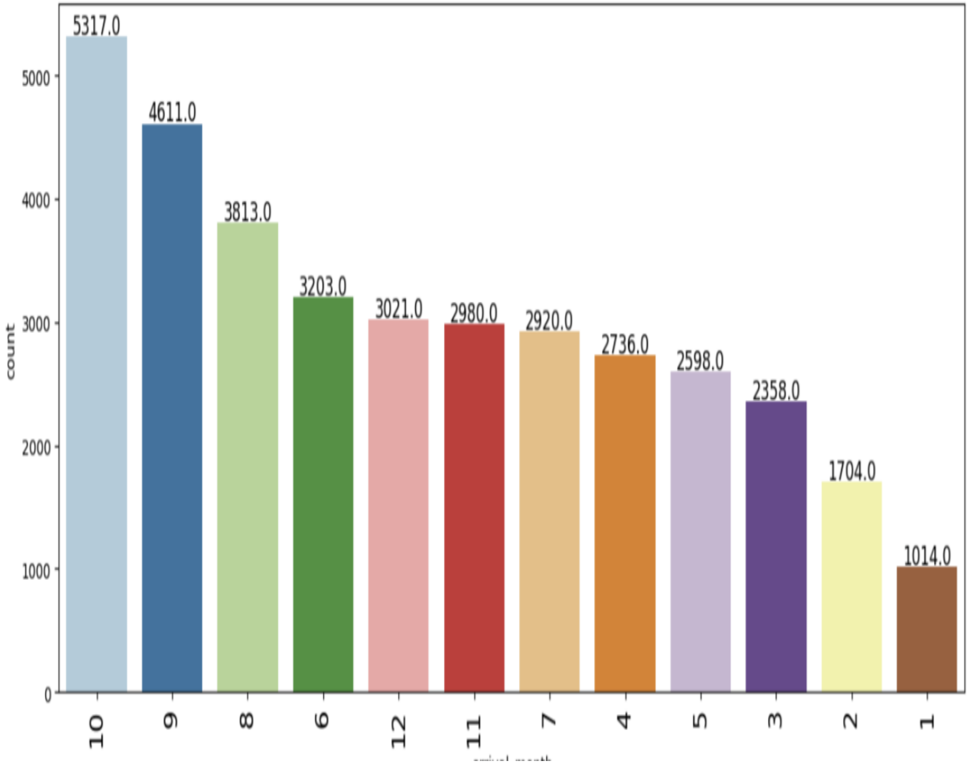
EDA- Room Type Reserved



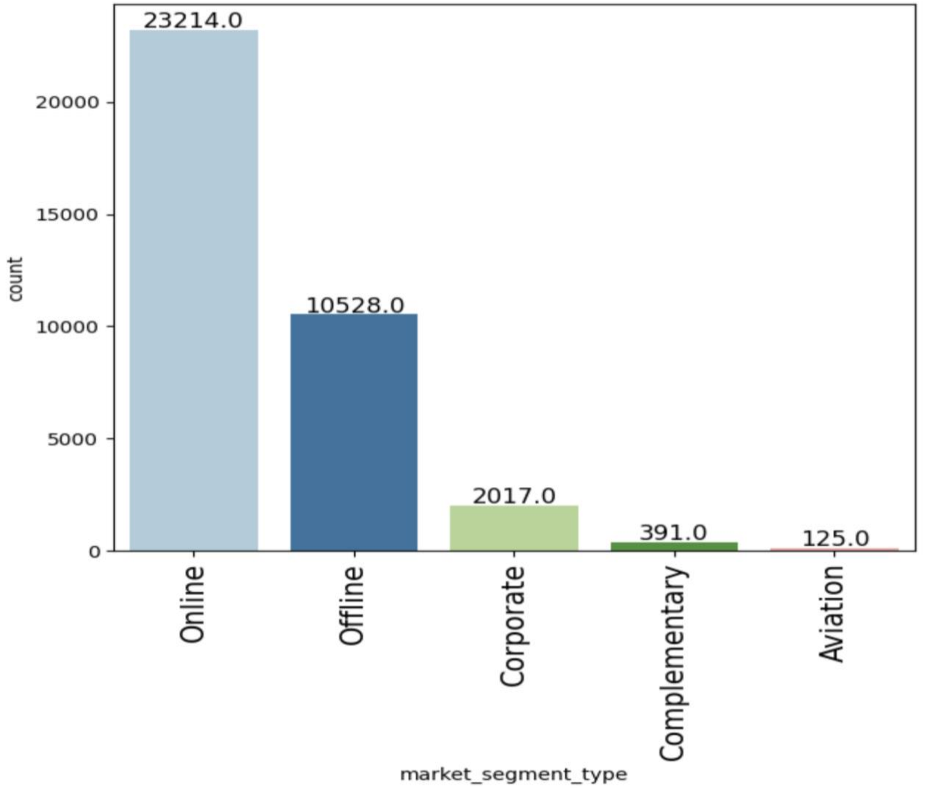
- Over 25,000 of the rooms reserved were Type 1
- Most of the rooms were Room Type 1, with nothing else even close

EDA- Arrival Month

- Most rooms were booked for the month of October, and the least was January



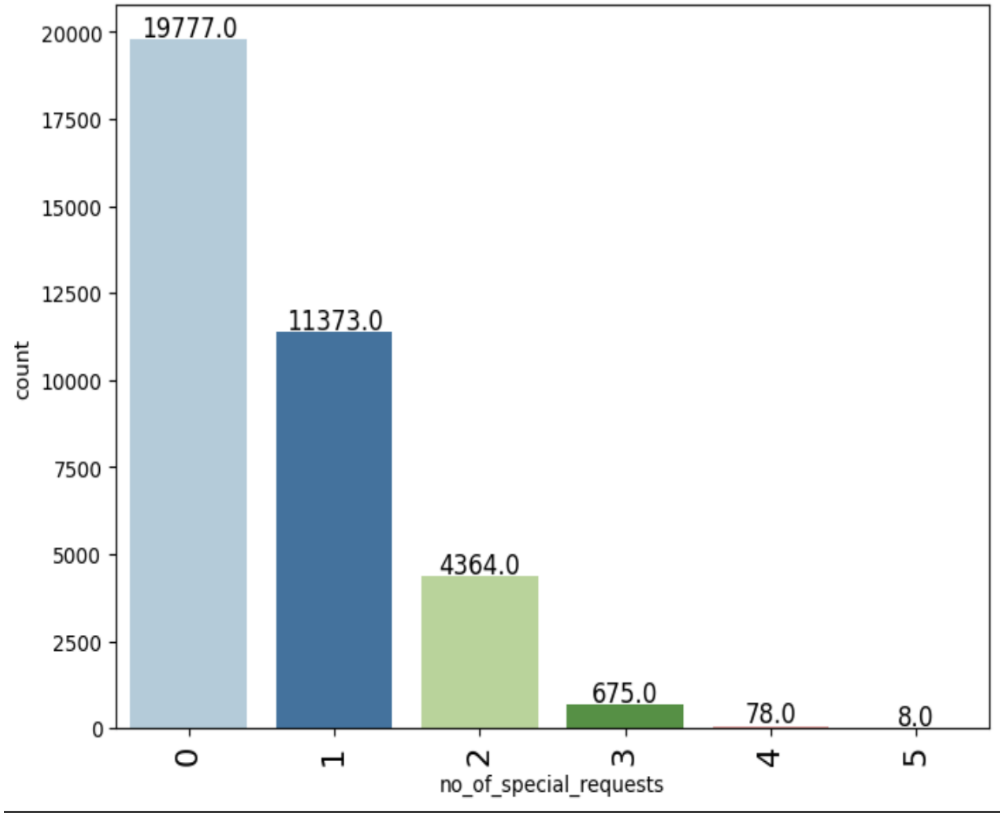
EDA- Market Segment Type



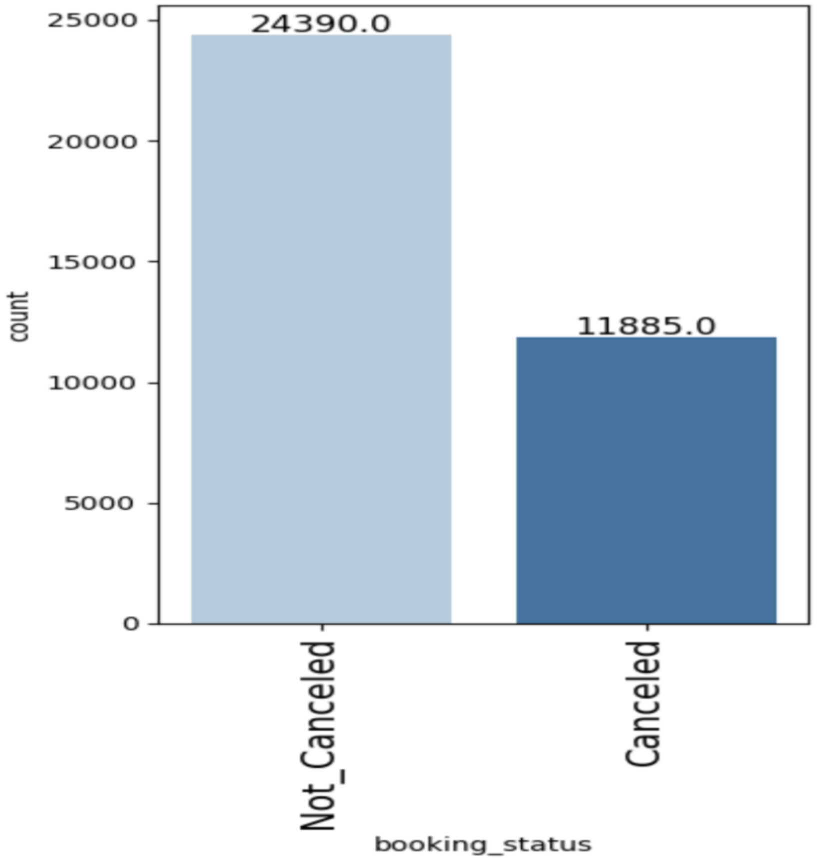
- The online market segment had the most bookings
- Complimentary bookings were the least.

EDA- Number of Special Requests

- Most bookings didn't have any special requests

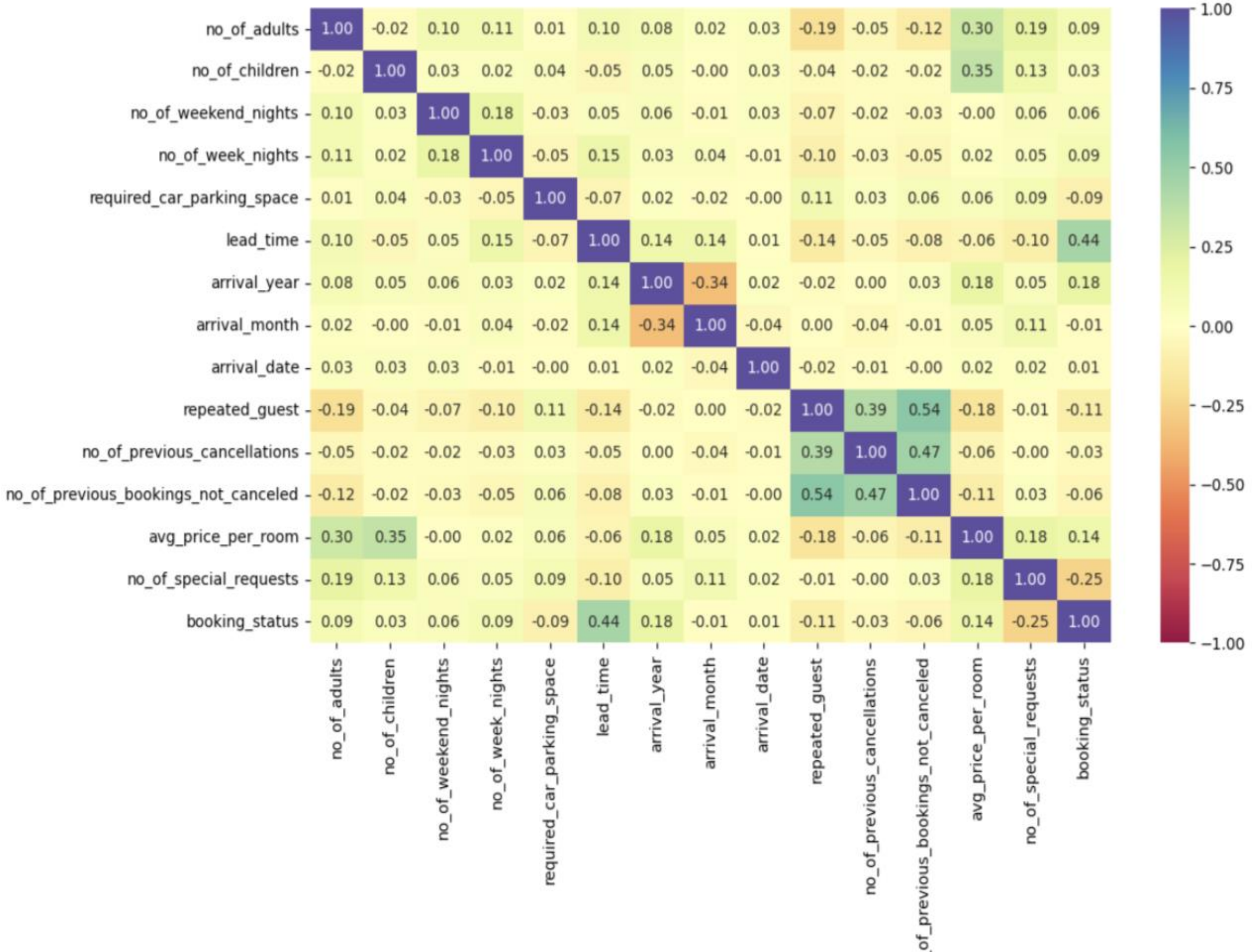


EDA- Booking Status



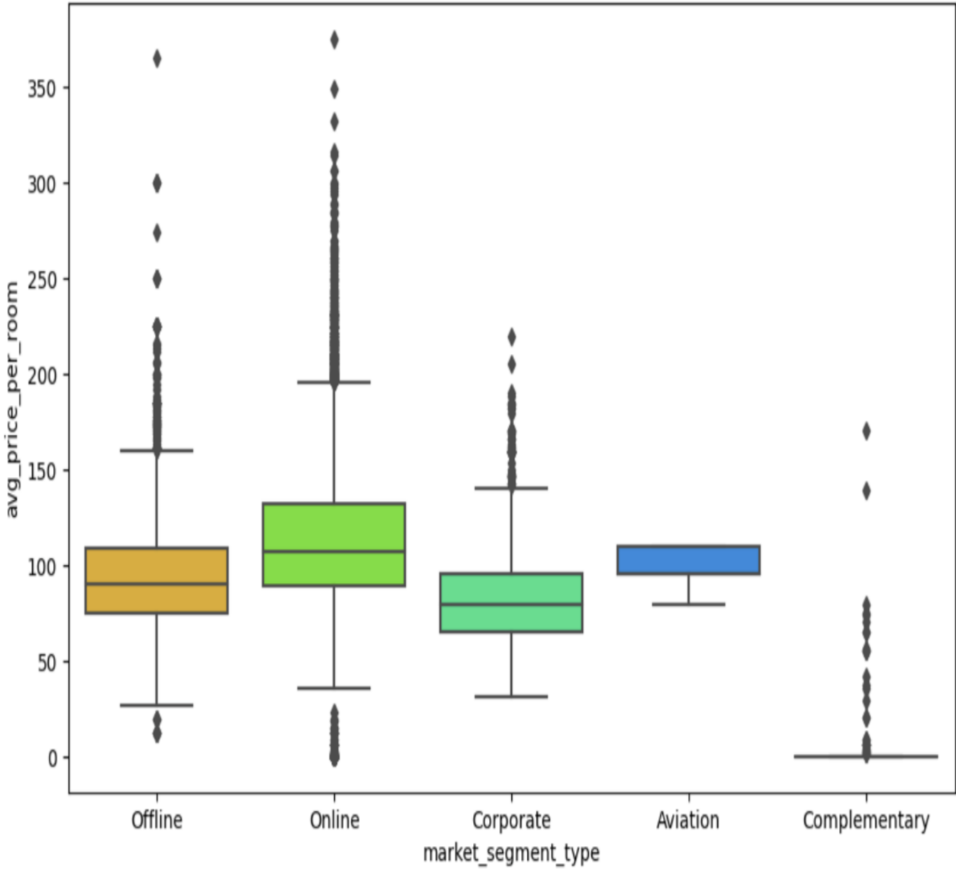
- 24, 390 reservations were not cancelled, and 11,885 were.

- There wasn't a strong correlation between any of the variables



	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	required_car_parking_space	lead_time	arrival_year	arrival_month	arrival_date	repeated_guest	no_of_previous_cancellations	no_of_previous_bookings_not_canceled	avg_price_per_room	no_of_special_requests	booking_status
no_of_adults	1.00	-0.02	0.10	0.11	0.01	0.10	0.08	0.02	0.03	-0.19	-0.05	-0.12	0.30	0.19	0.09
no_of_children	-0.02	1.00	0.03	0.02	0.04	-0.05	0.05	-0.00	0.03	-0.04	-0.02	-0.02	0.35	0.13	0.03
no_of_weekend_nights	0.10	0.03	1.00	0.18	-0.03	0.05	0.06	-0.01	0.03	-0.07	-0.02	-0.03	-0.00	0.06	0.06
no_of_week_nights	0.11	0.02	0.18	1.00	-0.05	0.15	0.03	0.04	-0.01	-0.10	-0.03	-0.05	0.02	0.05	0.09
required_car_parking_space	0.01	0.04	-0.03	-0.05	1.00	-0.07	0.02	-0.02	-0.00	0.11	0.03	0.06	0.06	0.09	-0.09
lead_time	0.10	-0.05	0.05	0.15	-0.07	1.00	0.14	0.14	0.01	-0.14	-0.05	-0.08	-0.06	-0.10	0.44
arrival_year	0.08	0.05	0.06	0.03	0.02	0.14	1.00	-0.34	0.02	-0.02	0.00	0.03	0.18	0.05	0.18
arrival_month	0.02	-0.00	-0.01	0.04	-0.02	0.14	-0.34	1.00	-0.04	0.00	-0.04	-0.01	0.05	0.11	-0.01
arrival_date	0.03	0.03	0.03	-0.01	-0.00	0.01	0.02	-0.04	1.00	-0.02	-0.01	-0.00	0.02	0.02	0.01
repeated_guest	-0.19	-0.04	-0.07	-0.10	0.11	-0.14	-0.02	0.00	-0.02	1.00	0.39	0.54	-0.18	-0.01	-0.11
no_of_previous_cancellations	-0.05	-0.02	-0.02	-0.03	0.03	-0.05	0.00	-0.04	-0.01	0.39	1.00	0.47	-0.06	-0.00	-0.03
no_of_previous_bookings_not_canceled	-0.12	-0.02	-0.03	-0.05	0.06	-0.08	0.03	-0.01	-0.00	0.54	0.47	1.00	-0.11	0.03	-0.06
avg_price_per_room	0.30	0.35	-0.00	0.02	0.06	-0.06	0.18	0.05	0.02	-0.18	-0.06	-0.11	1.00	0.18	0.14
no_of_special_requests	0.19	0.13	0.06	0.05	0.09	-0.10	0.05	0.11	0.02	-0.01	-0.00	0.03	0.18	1.00	-0.25
booking_status	0.09	0.03	0.06	0.09	-0.09	0.44	0.18	-0.01	0.01	-0.11	-0.03	-0.06	0.14	-0.25	1.00

EDA- Market Segment Type v. Average Price/Room

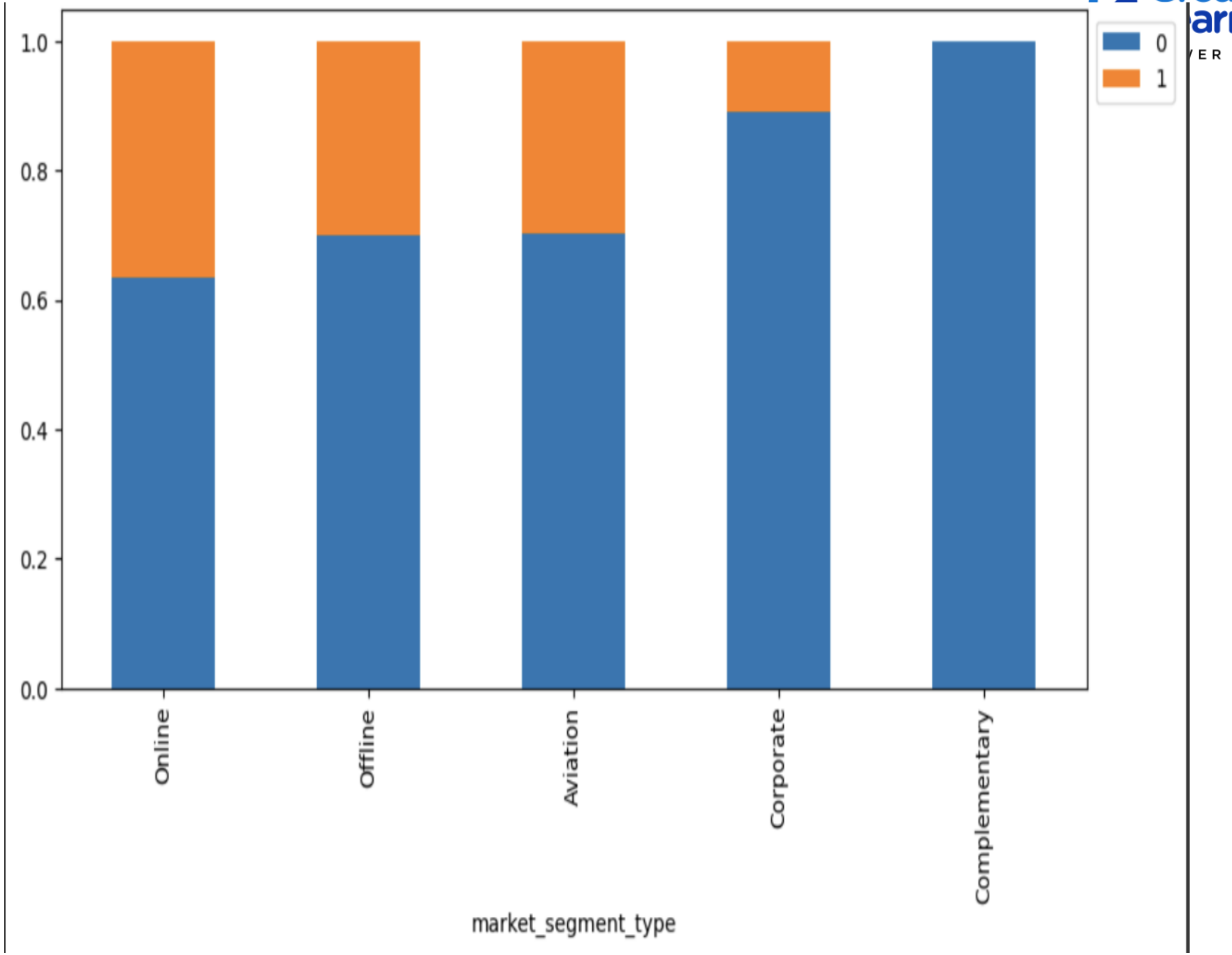


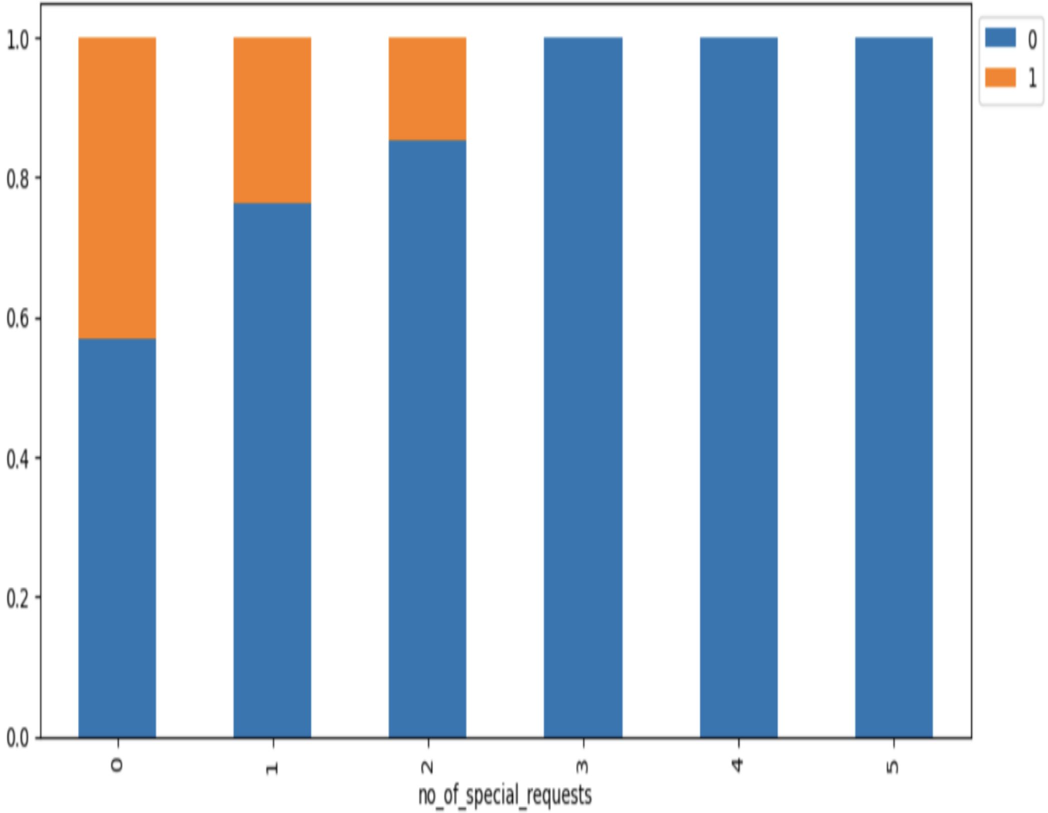
- Mean Average price per room is highest for online reservations

- There are many outliers with the online reservation segment.

EDA-

.

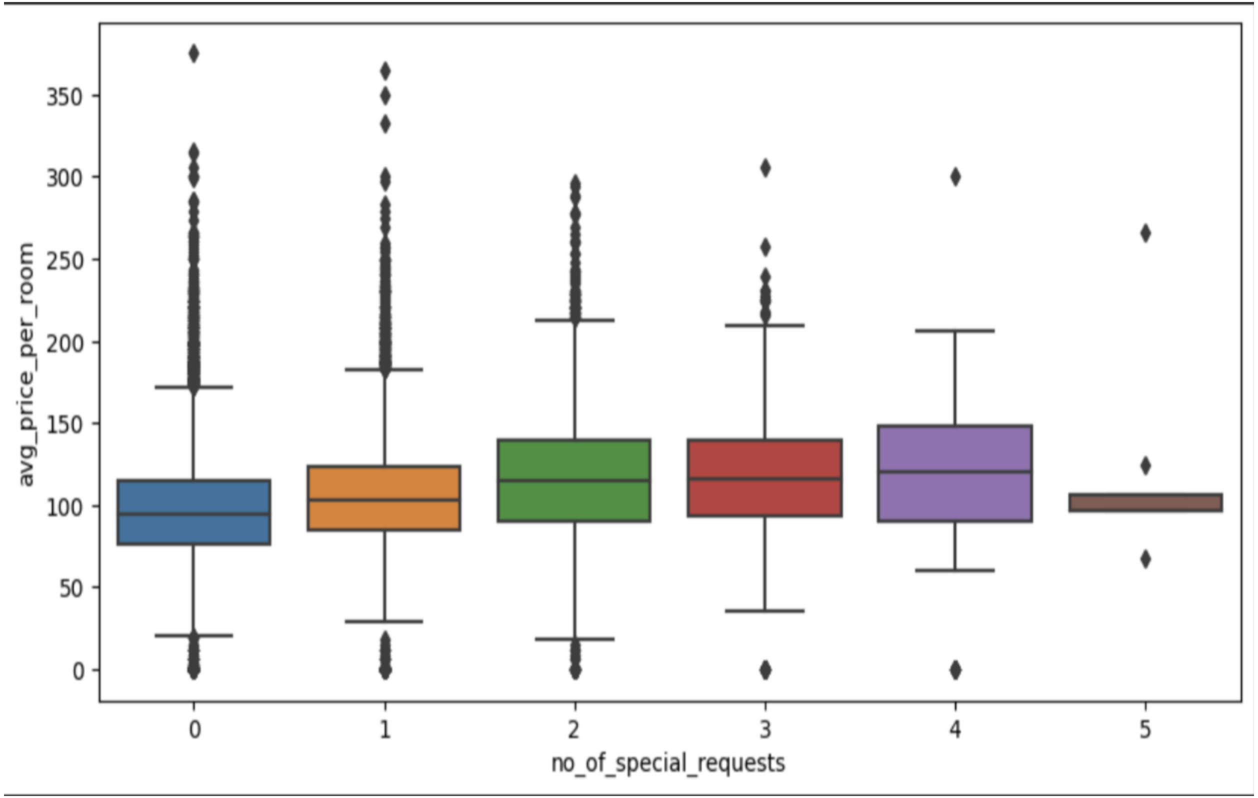

1.0
0.8
0.6
0.4
0.2
0.0
Online
Offline
Aviation
Corporate
Complementary
market_segment_type
0
1



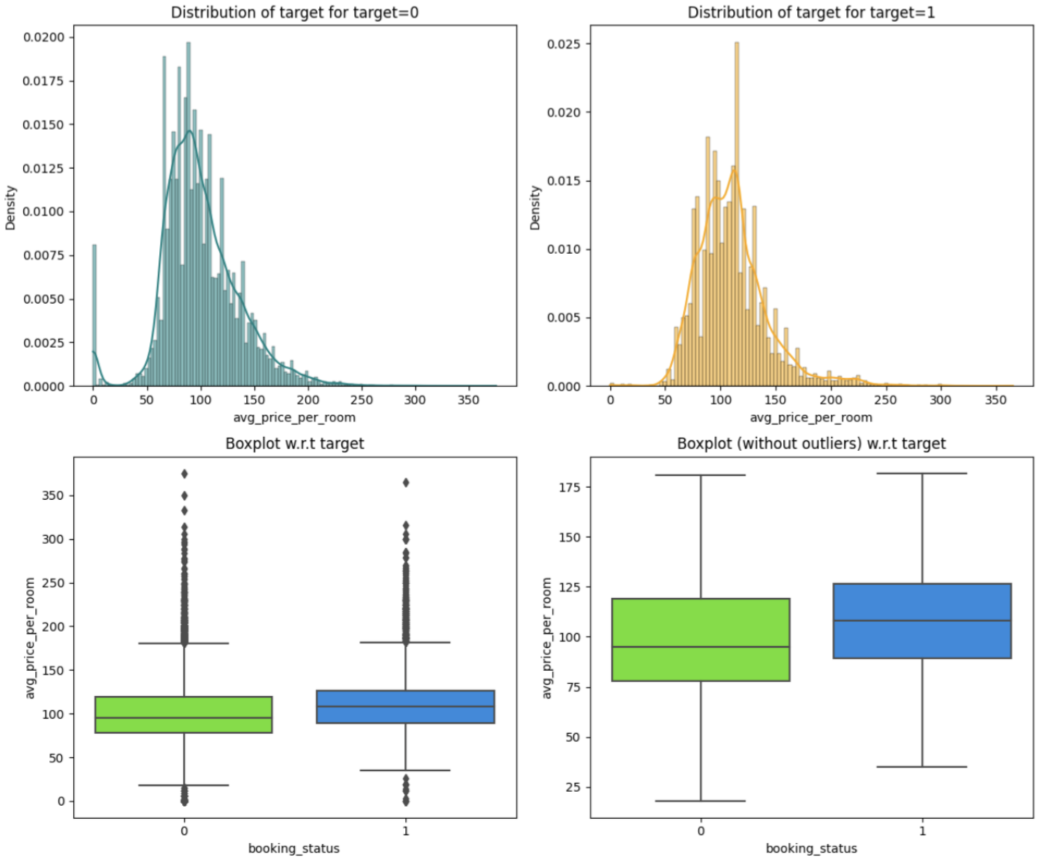
- The more specific a guest is with their special requests, the less likely they are to cancel.

EDA- Number of Special Requests v. Average Price/Room

- There is nothing distinct about this comparison.
- Most reservations h ad no special requests.
- There is no correlation between average room price and number of special requests.



EDA- Avg. Price/Room v. Booking Status



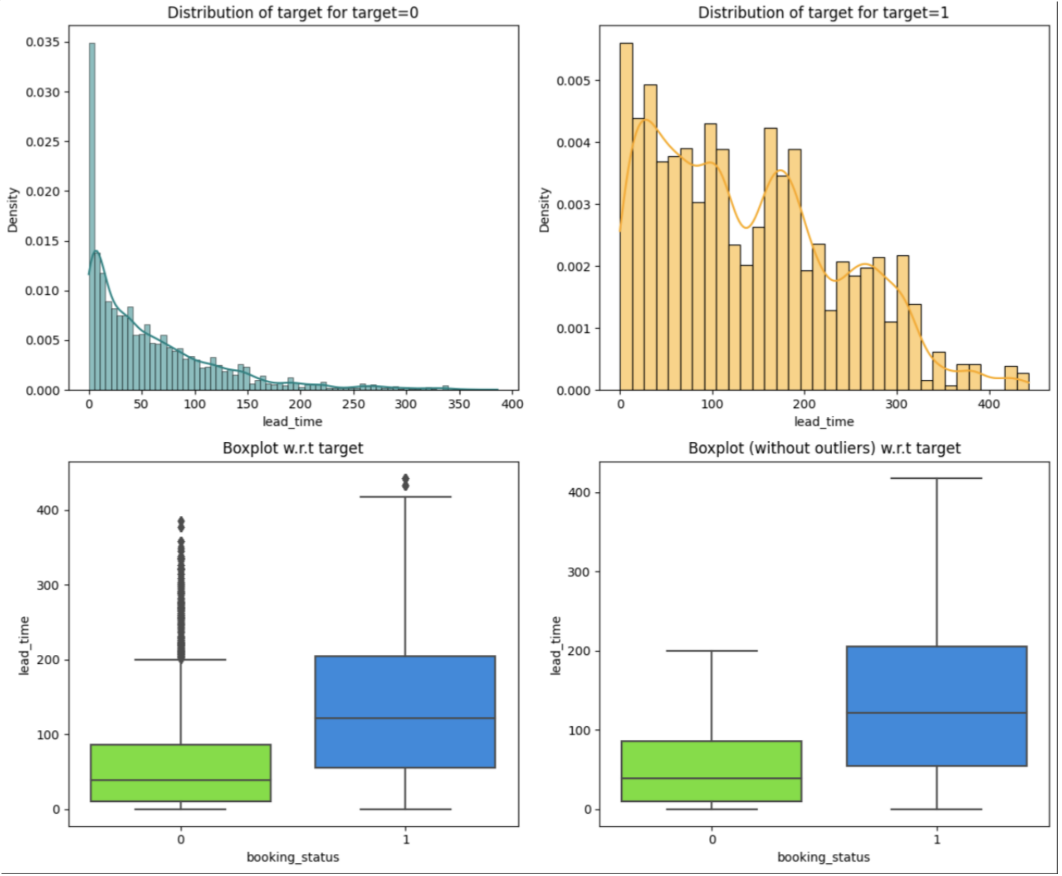
- Booking status and average price per room is slightly right-skewed. The rooms that weren't cancelled are cheaper than the average price for cancelled rooms.

EDA- Lead Time v. Booking Status

- The lead time distribution is right-skewed.

- The greater the lead time, the greater likelihood of cancellation

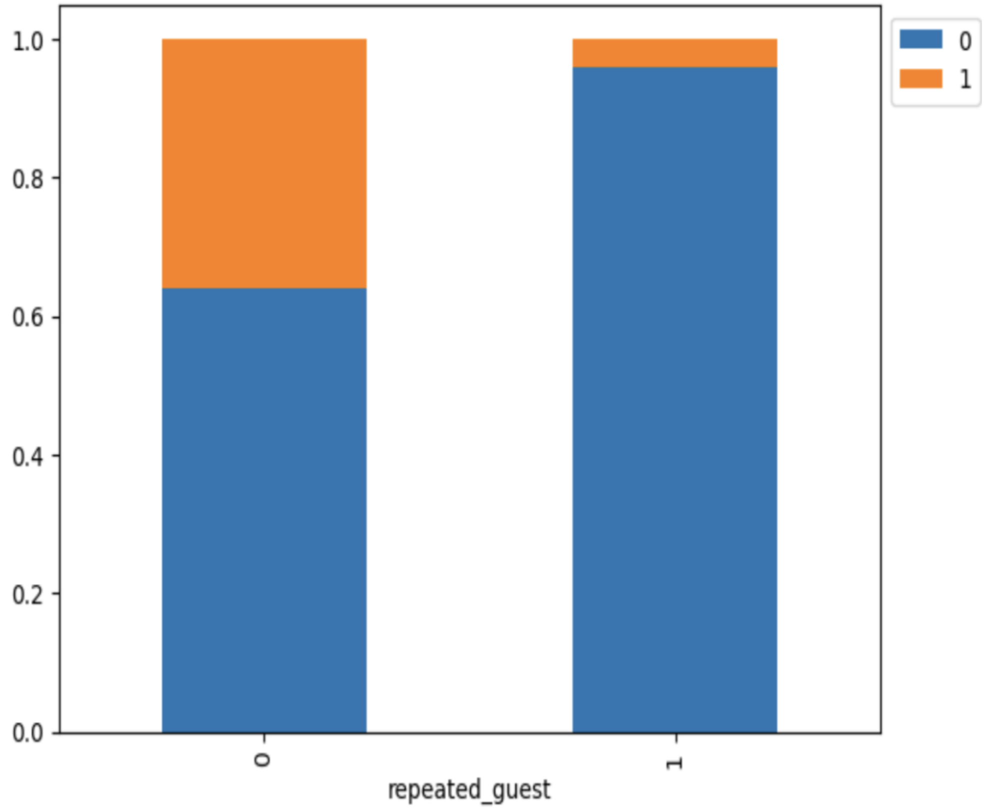
.



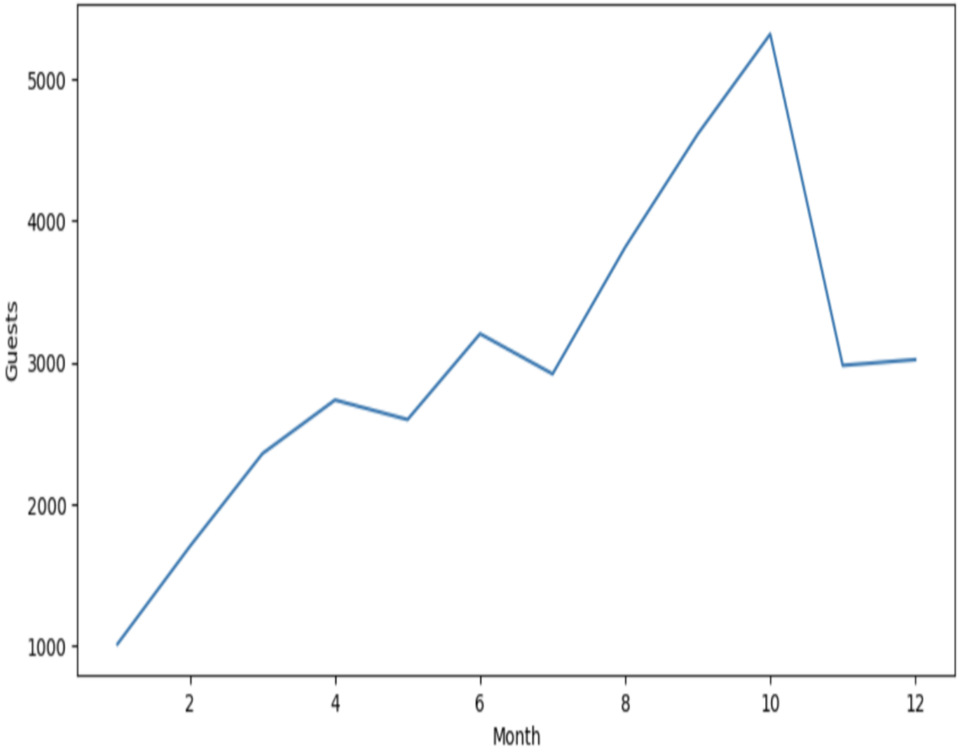
EDA- Repeated Guest

- Repeat guests have a very low number of cancellations.

- Most cancellations come from new customers.

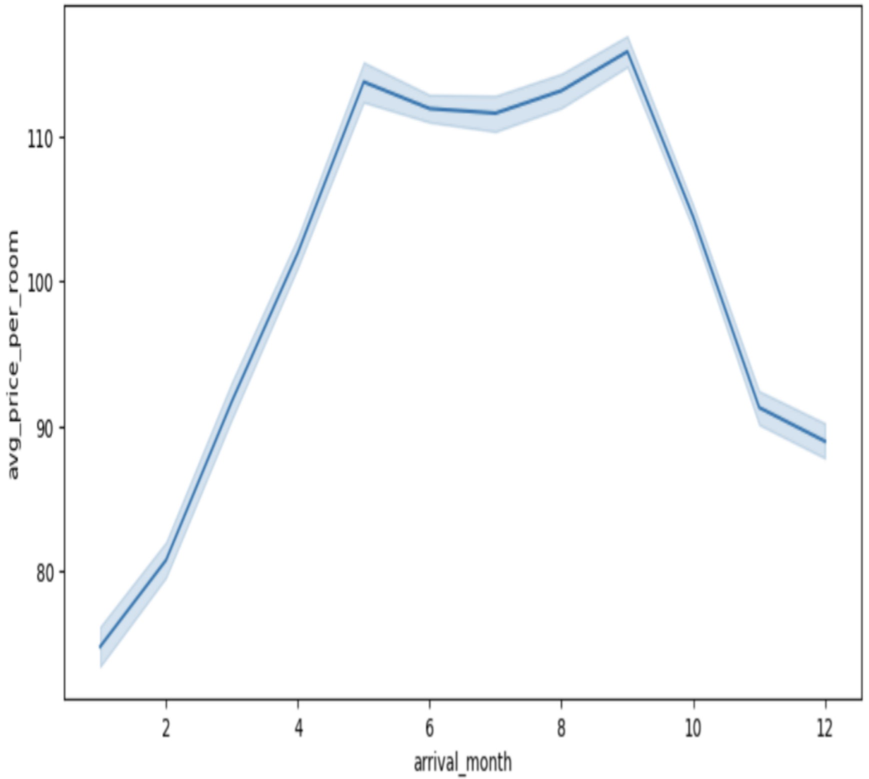


EDA- Month v. Guests

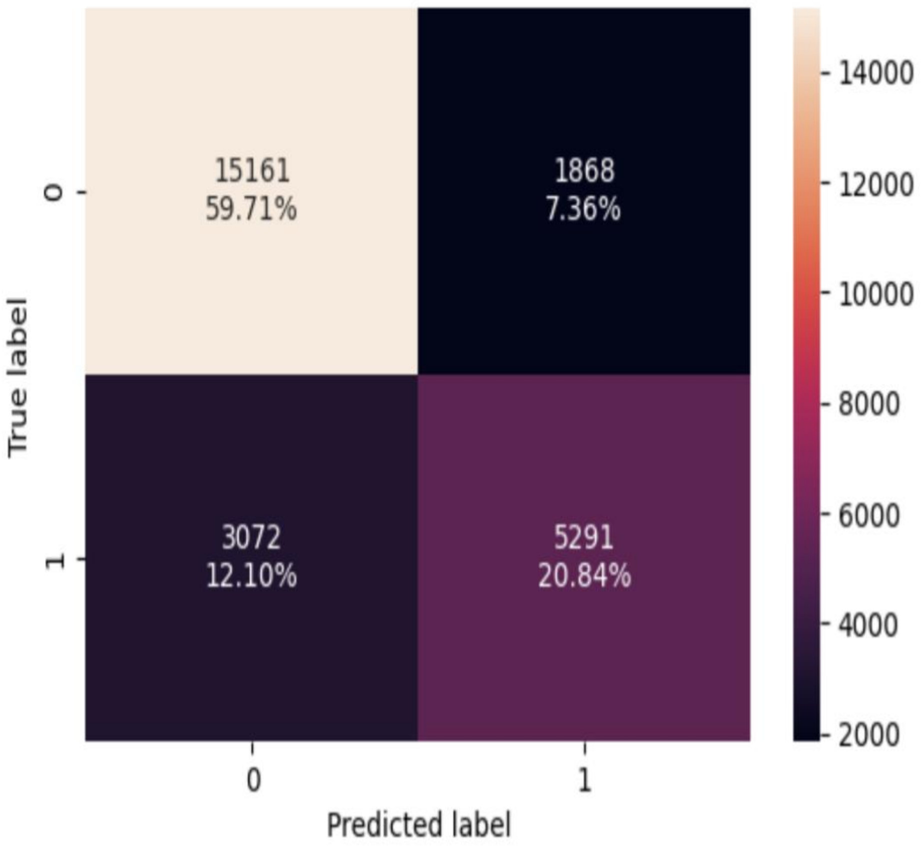


- Most bookings occur between August
- And October.

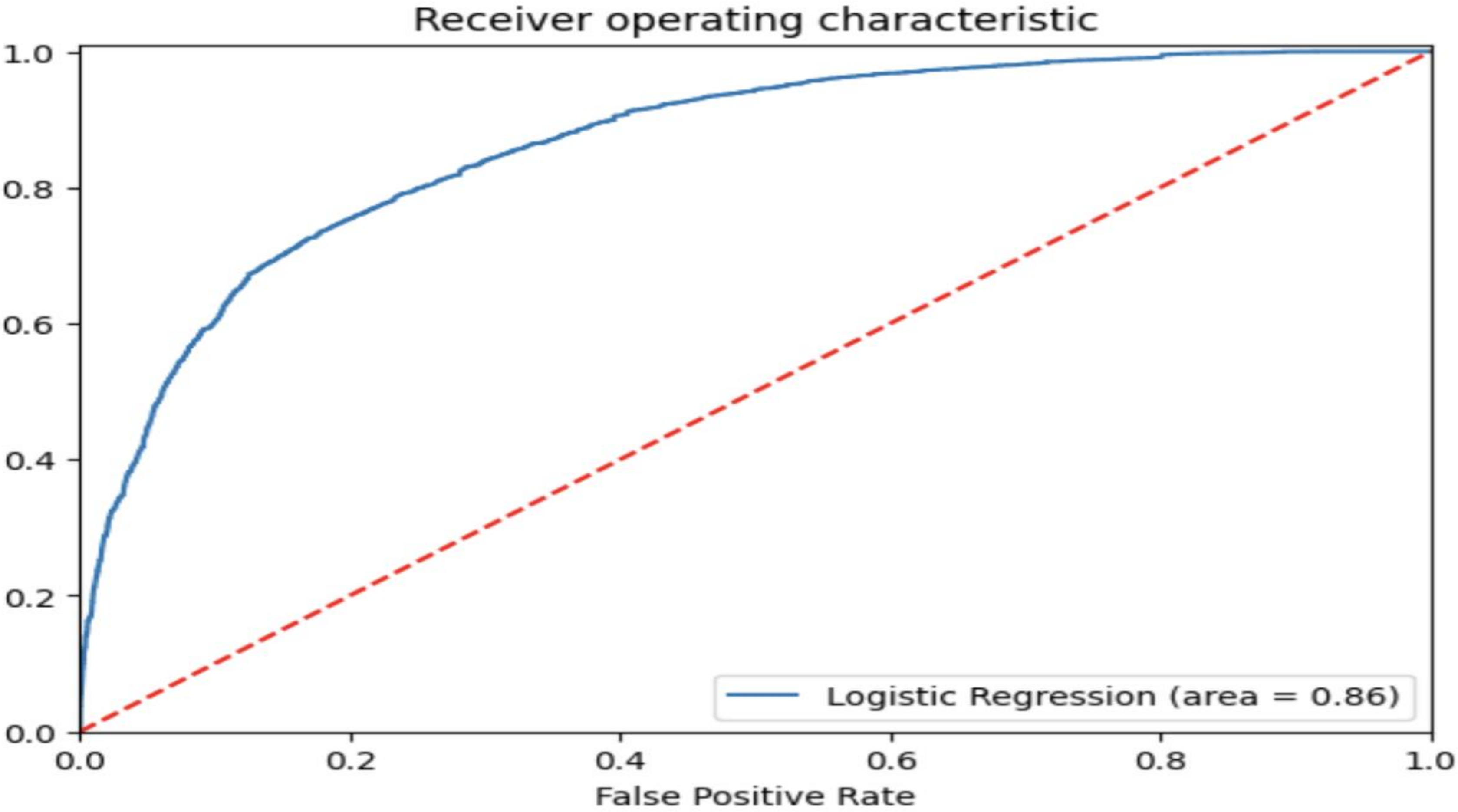
EDA- Arrival Month v. Avg. Price/Room

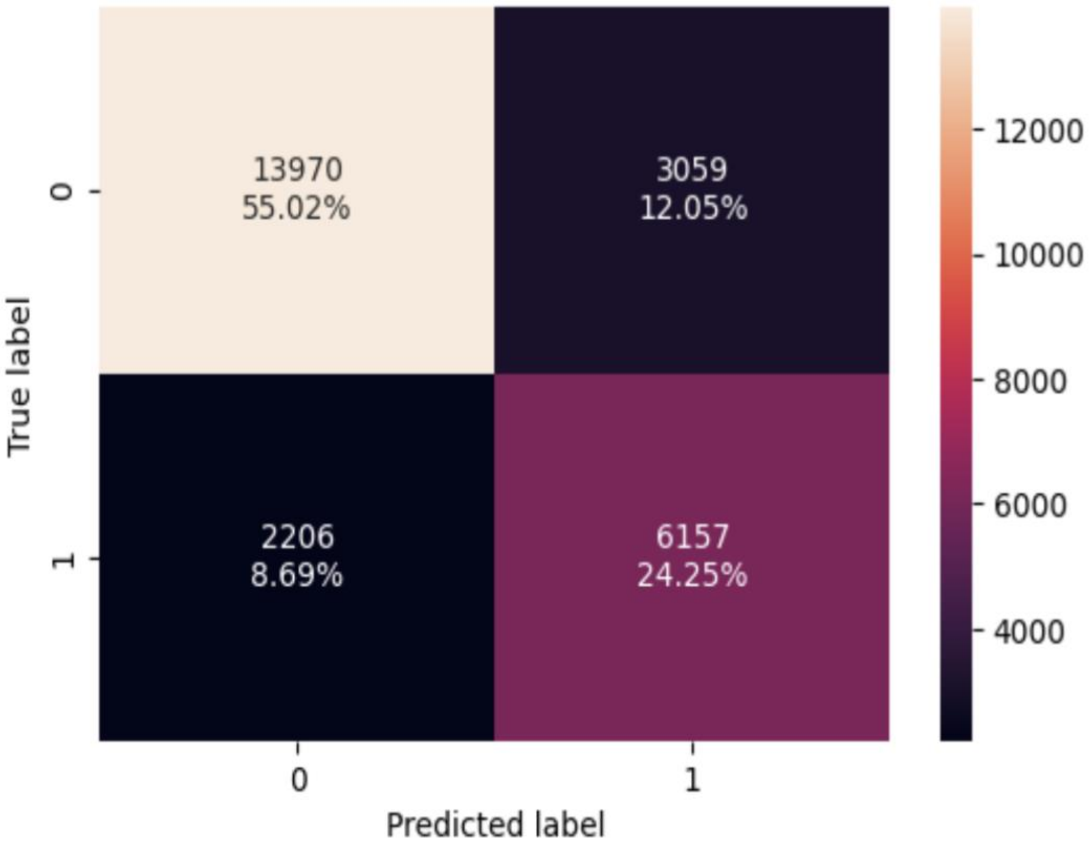


- Most families travel in the summer for summer vacations. Many hotels increase prices during this time to maximize profits.



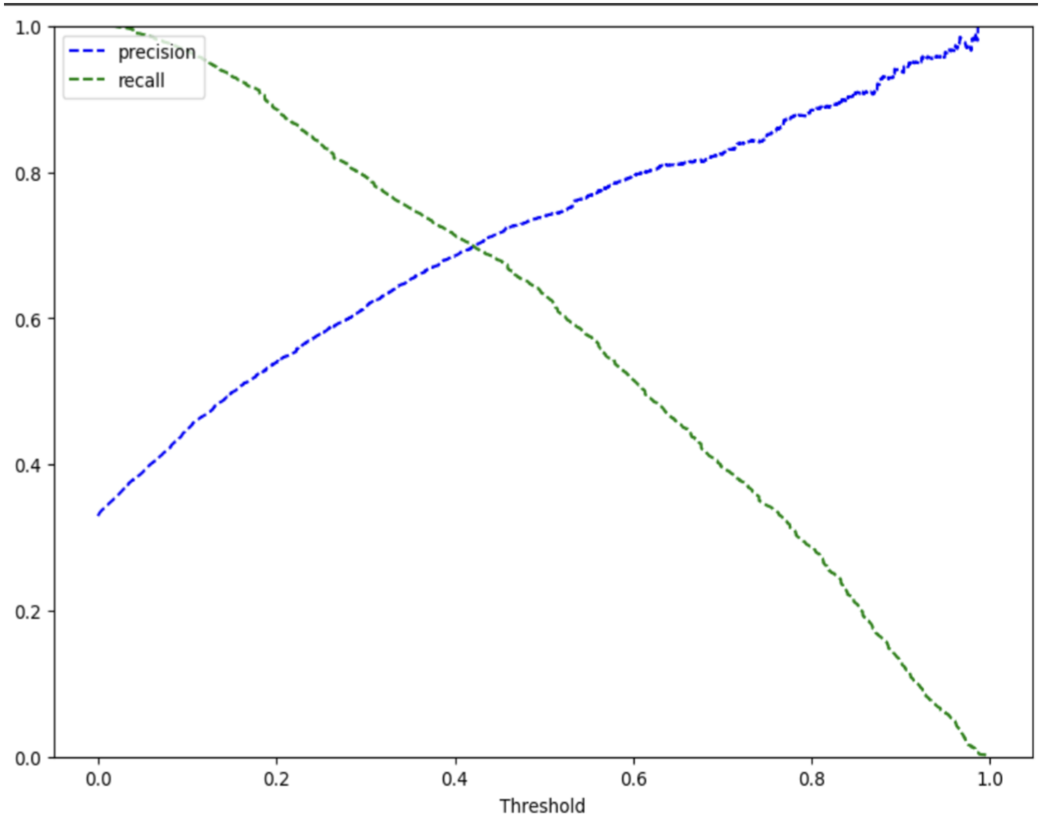
- The training data predicts a true positive ~ 60%
- Training data predicts True negative ~21%
- Training data predicts false positive at ~7%
- Training data predicts false negative at 12%

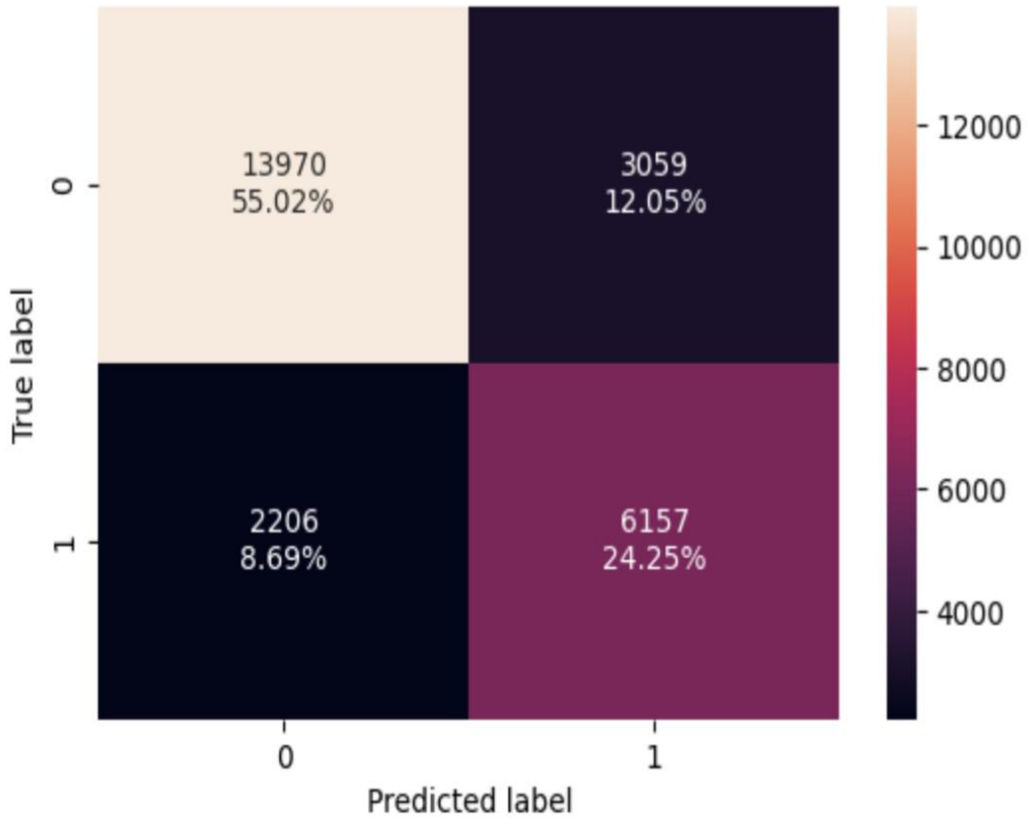

Receiver operating characteristic
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
Logistic Regression (area = 0.86)



- Confusion matrix on the test model shows:
- True Positive: ~55%
- False Positive: ~12%
- False Negative: ~ 24%
- True Negative: ~9%

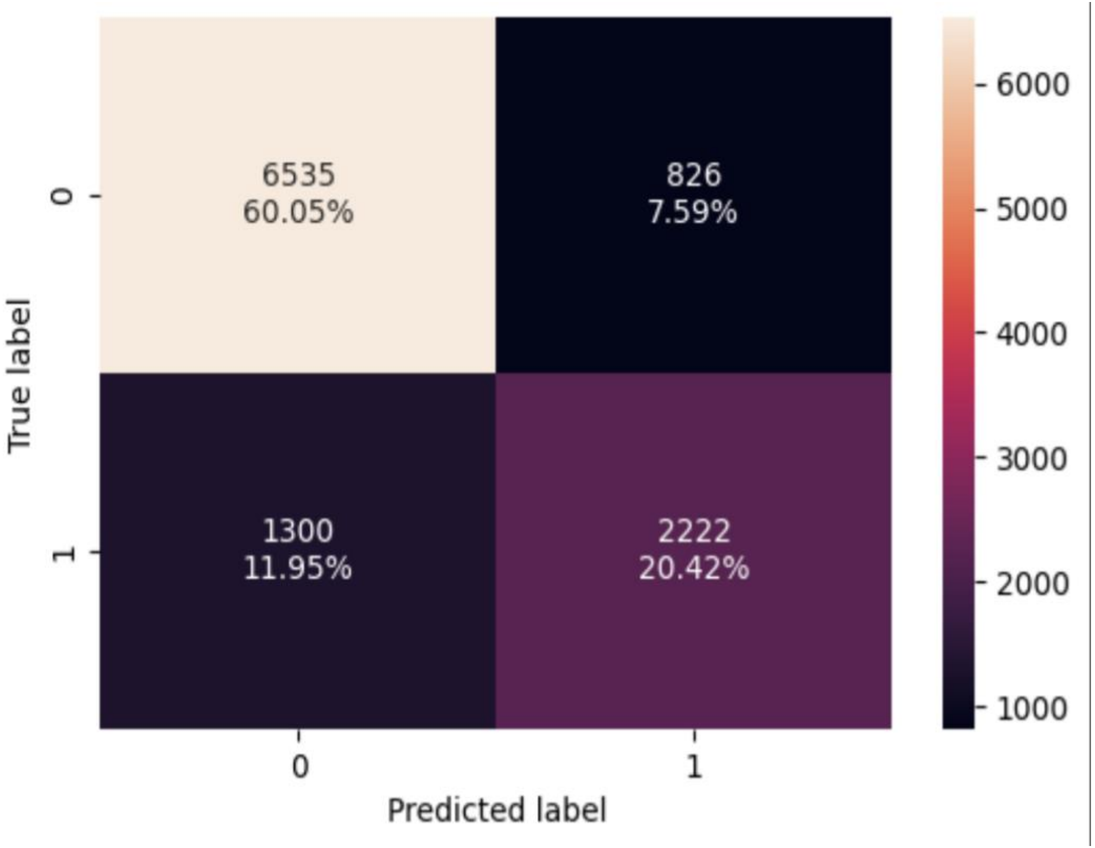
- Treshold of 0.42 gives equal precision and recall.
- The treshold will be set at 0.42

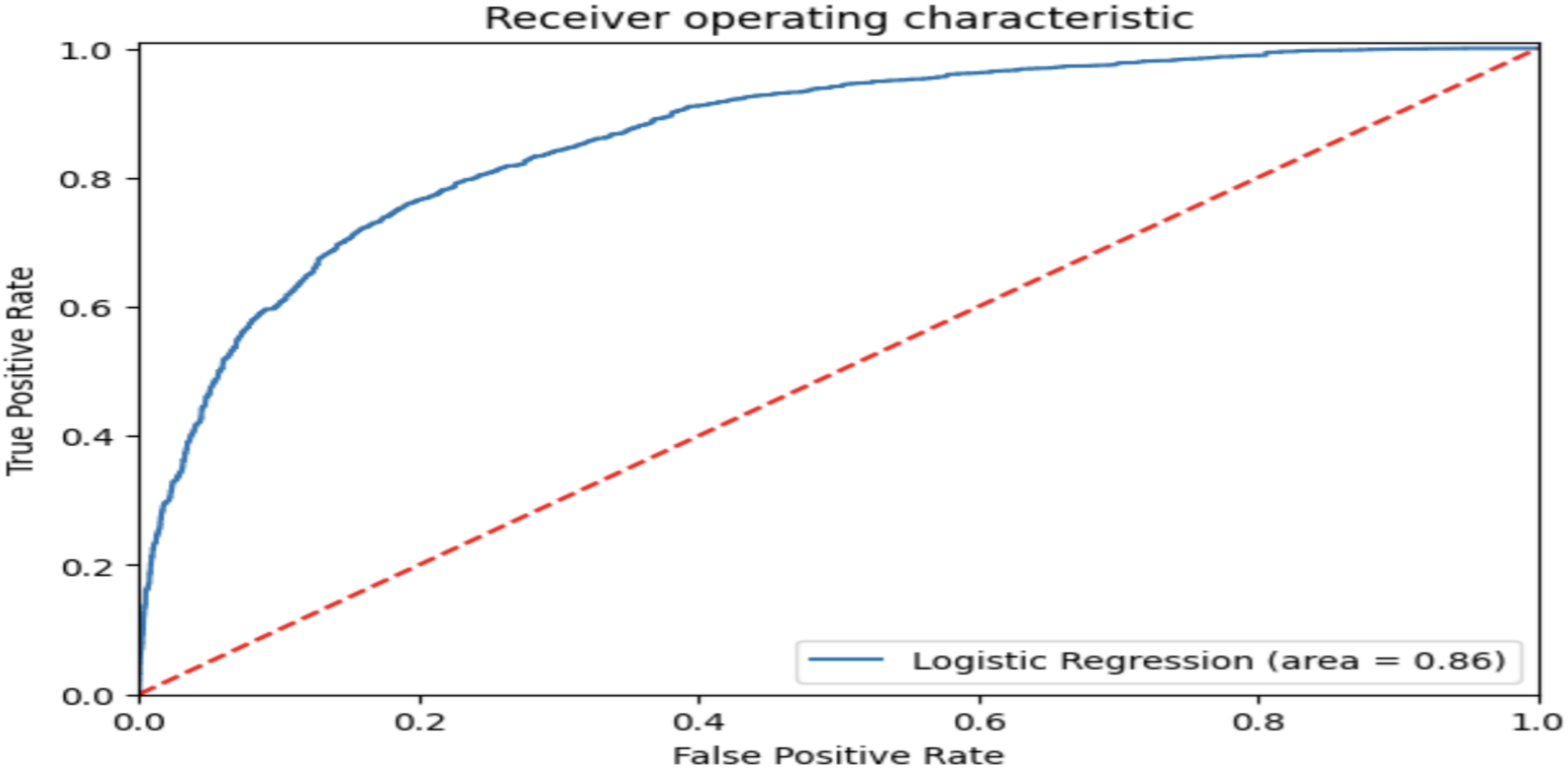


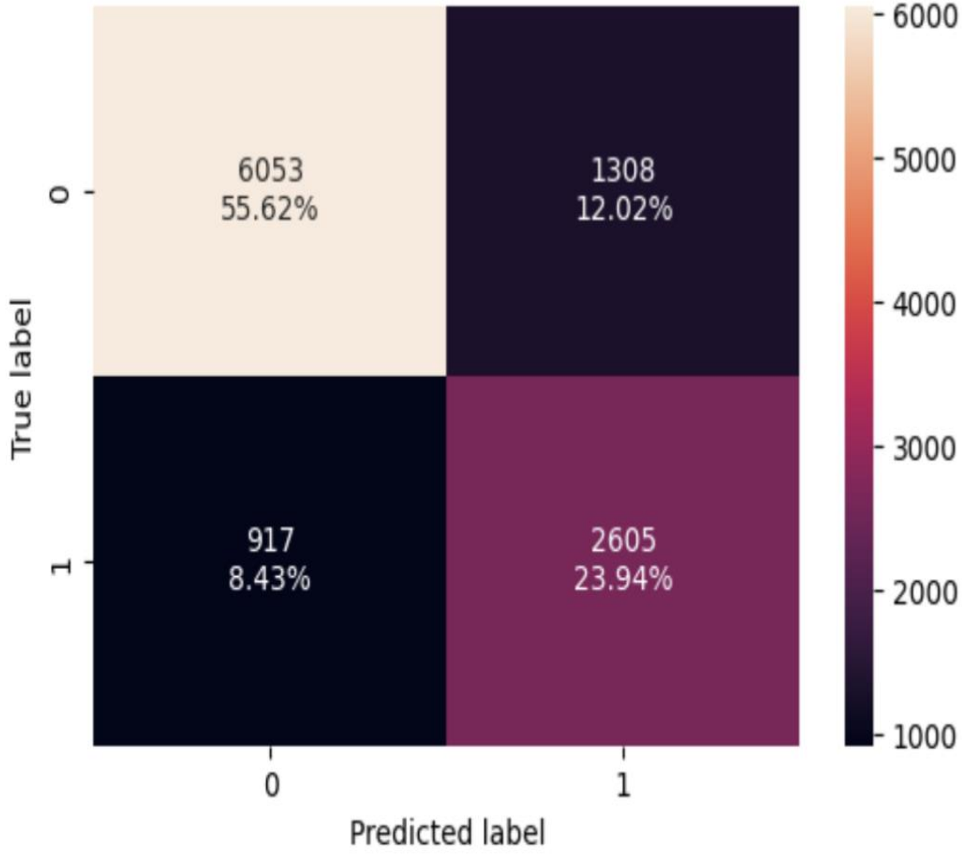


- Confusion Matrix shows:
- True Positive: 55%
- True Negative: ~ 24%
- False Positive: 12%
- False Negative: ~9%

- Confusion Matrix shows:
- True Positive: 60%
- True Negative: 20.42%
- False Positive: ~8%
- False Negative: ~12%

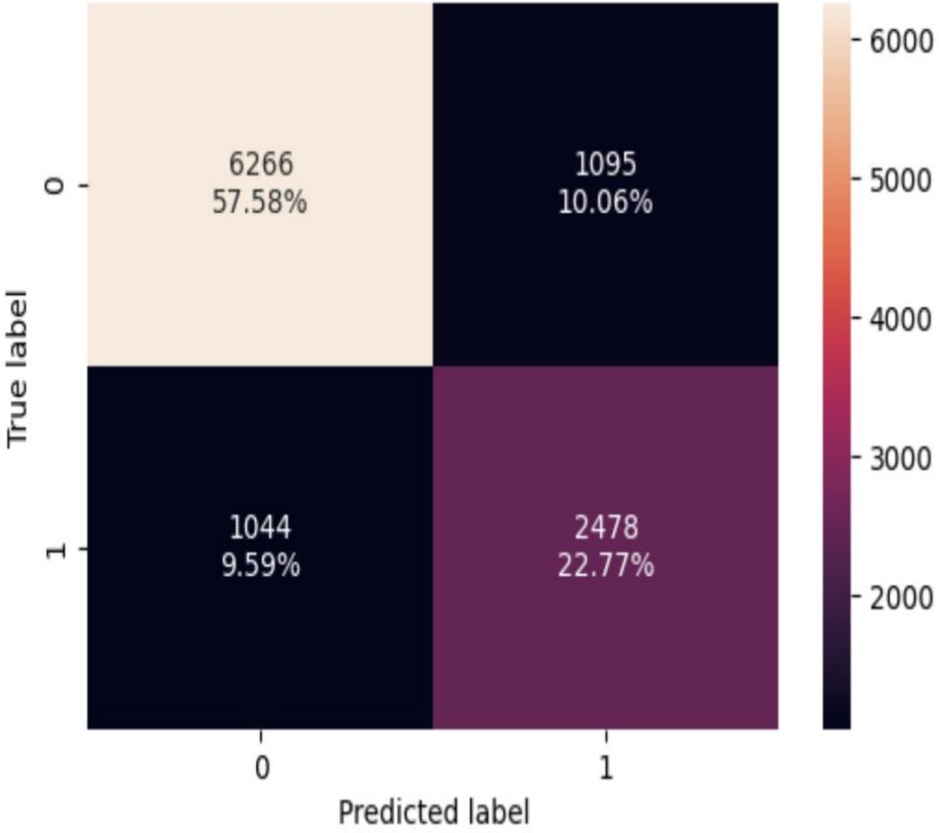


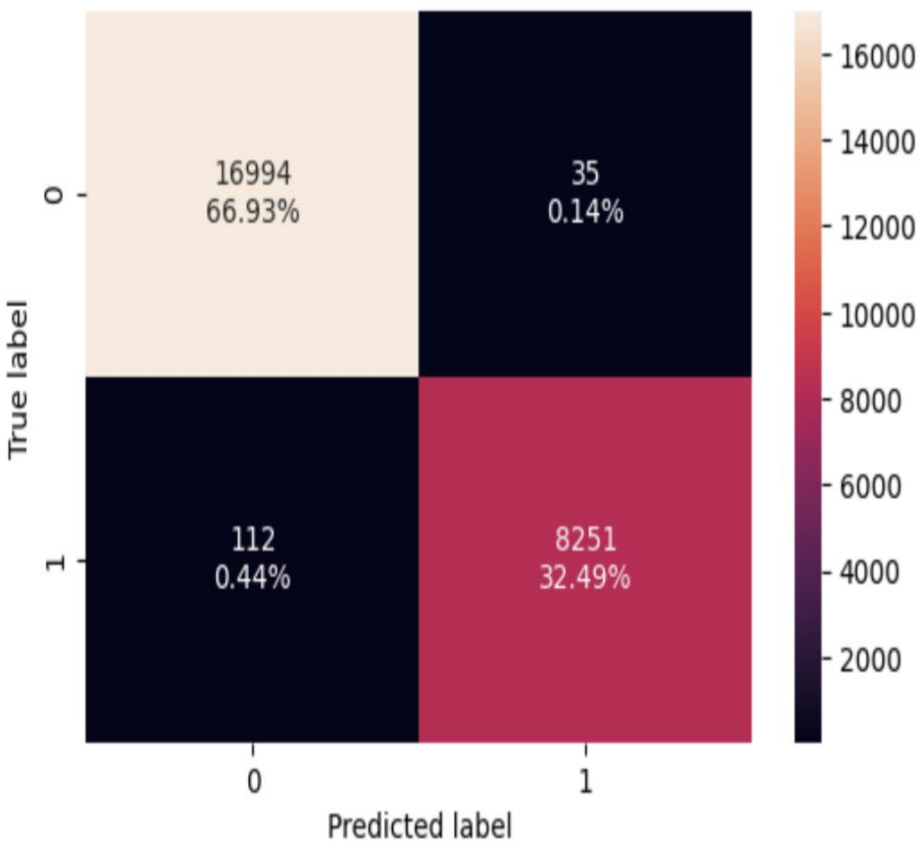

Receiver operating characteristic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression (area = 0.86)



- Confusion Matrix shows:
- True Positive: ~56%
- True negative: ~24%
- False Positive: 12%
- False Negative: ~8%

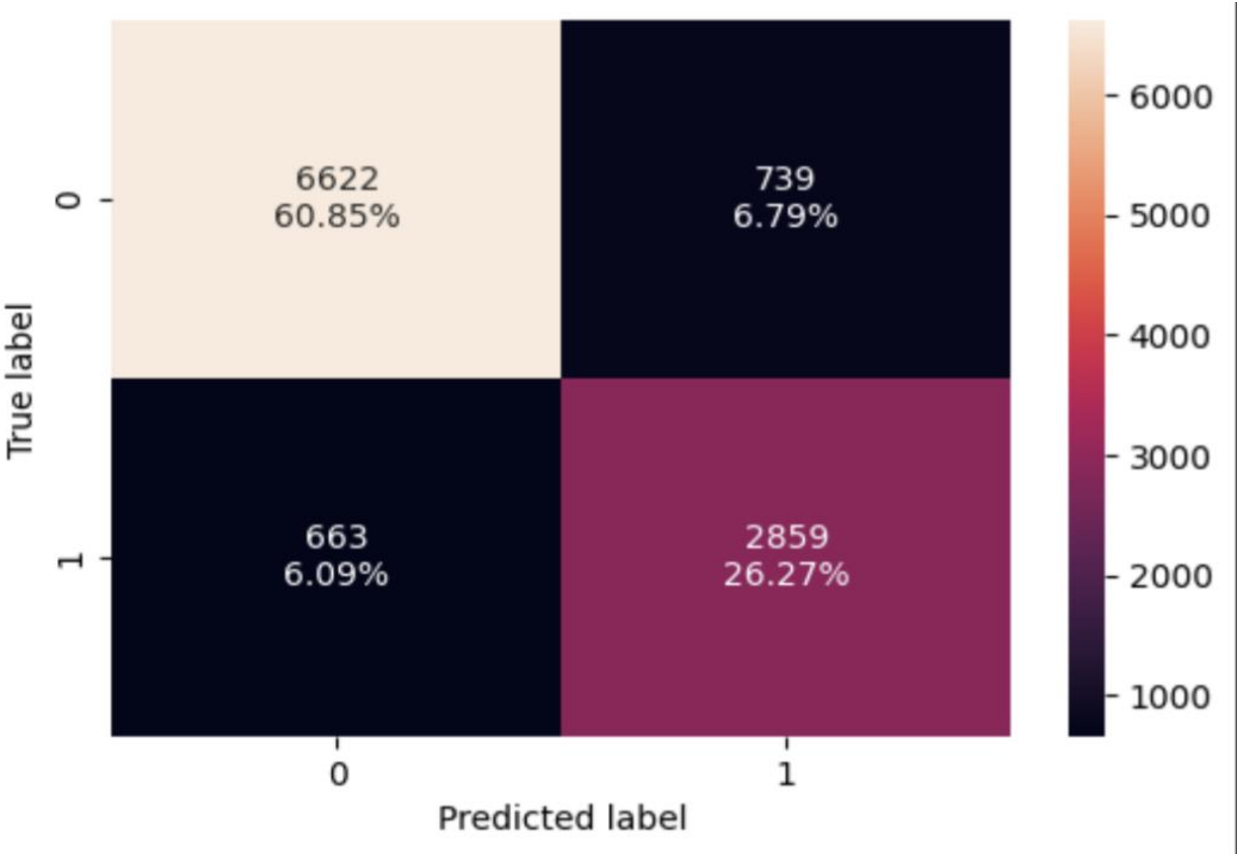
- Confusion Matrix shows:
- True Positive: ~57%
- True negative: ~23%
- False Positive: 10%
- False Negative: ~10%

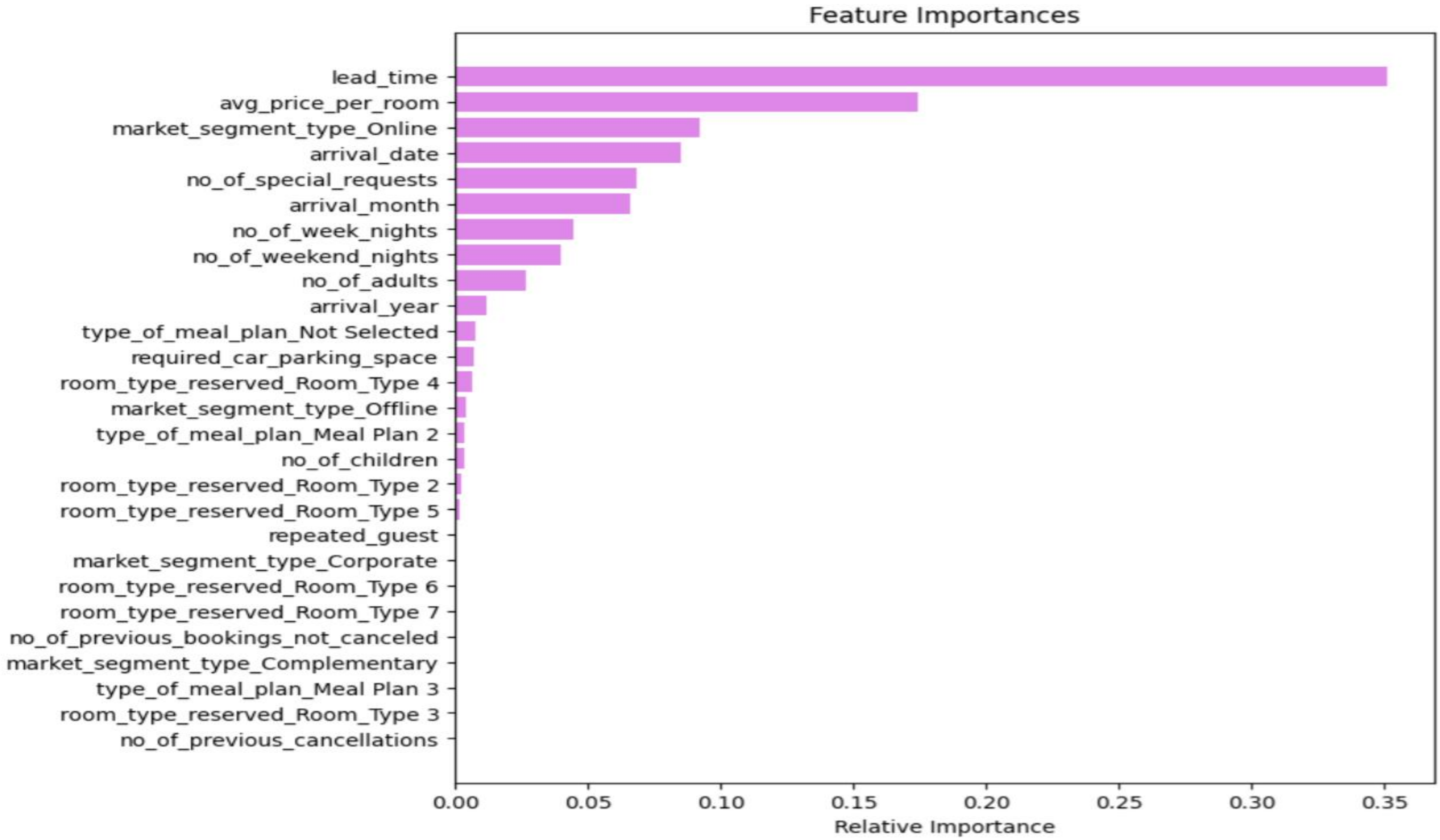




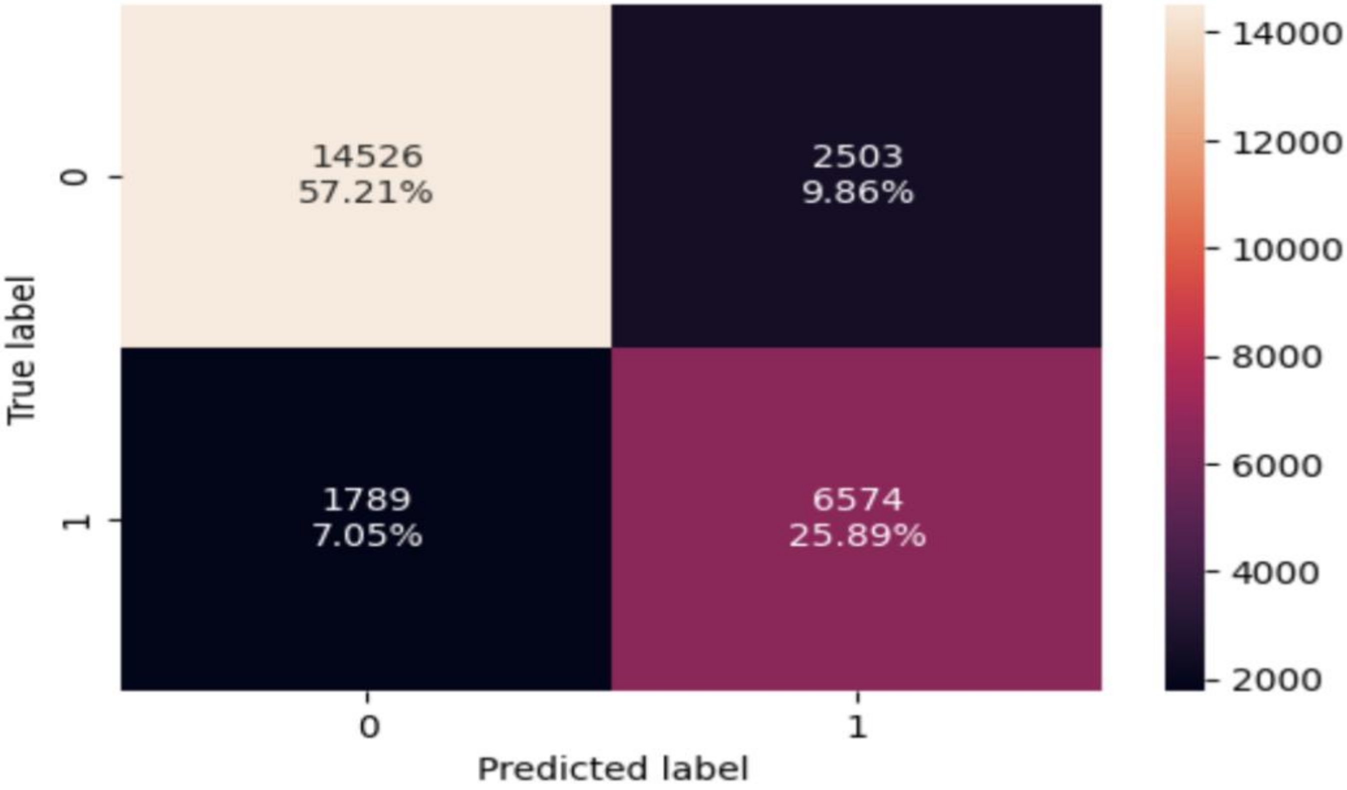
- Confusion Matrix shows:
- True Positive: ~67%
- True negative: ~33%
- False Positive: 0.14%
- False Negative: ~0.44%

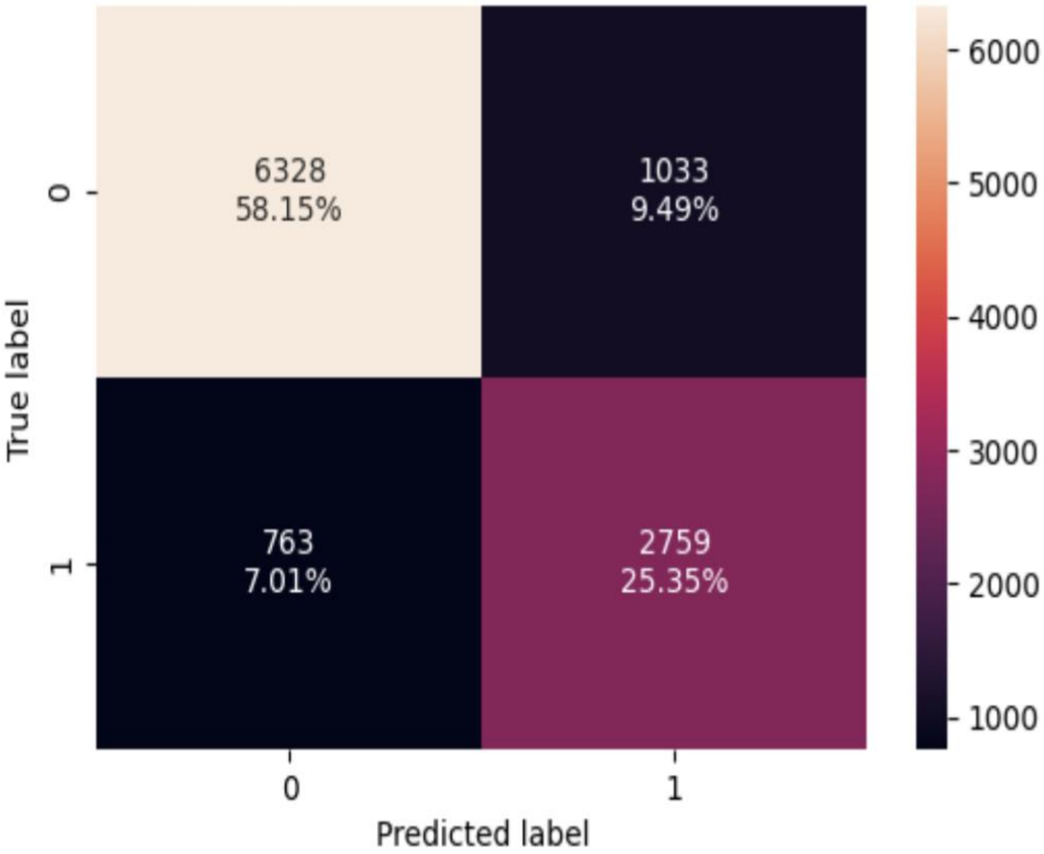
- Confusion Matrix shows:
- True Positive: ~61%
- True negative: ~26%
- False Positive: ~7%
- False Negative: ~6%



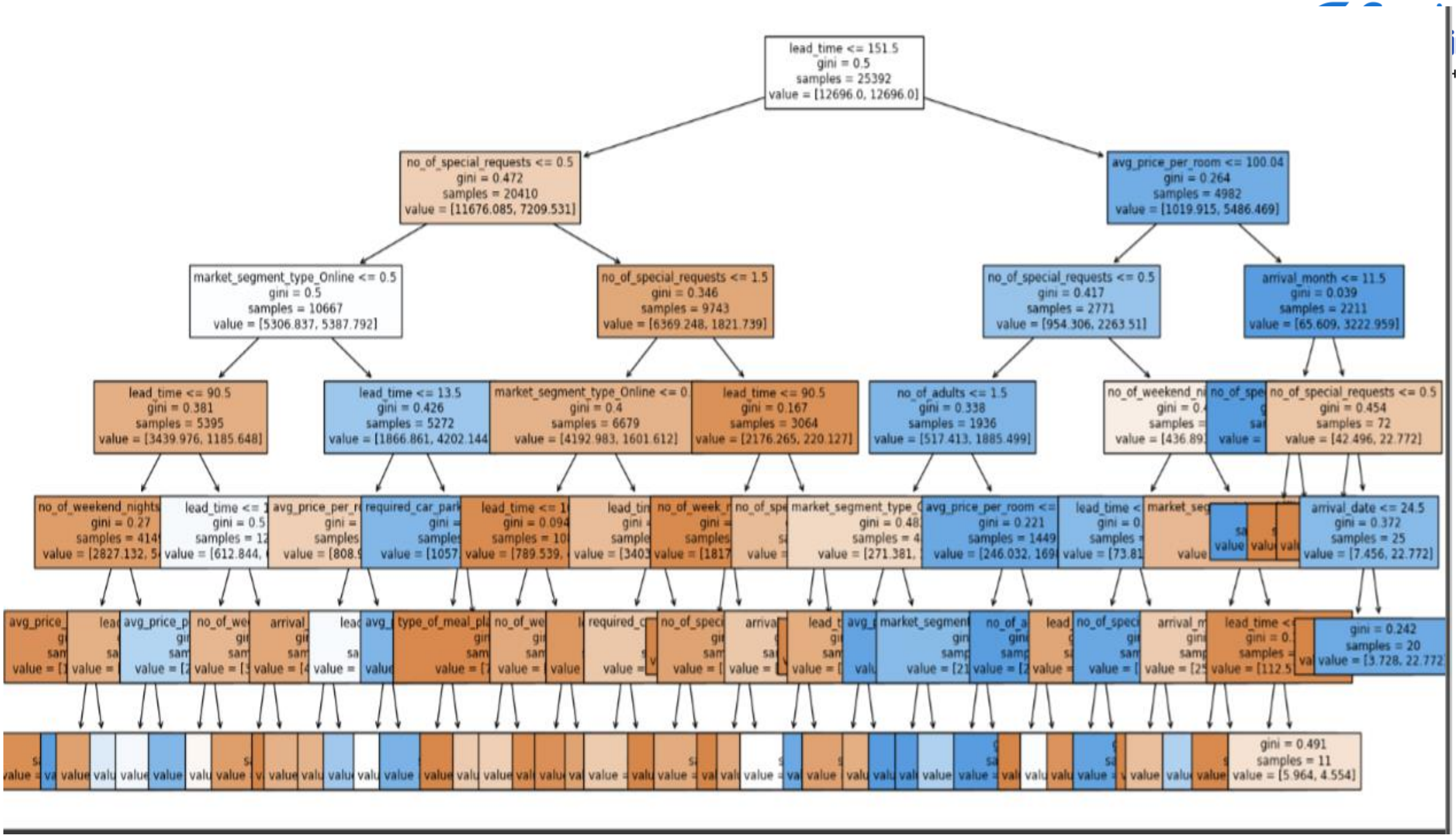

Feature Importances
lead_time
avg_price_per_room
market_segment_type_Online
arrival_date
no_of_special_requests
arrival_month
no_of_week_nights
no_of_weekend_nights
no_of_adults
arrival_year
type_of_meal_plan_Not Selected
required_car_parking_space
room_type_reserved_Room_Type 4
market_segment_type_Offline
type_of_meal_plan_Meal Plan 2
no_of_children
room_type_reserved_Room_Type 2
room_type_reserved_Room_Type 5
repeated_guest
market_segment_type_Corporate
room_type_reserved_Room_Type 6
room_type_reserved_Room_Type 7
no_of_previous_bookings_not_canceled
market_segment_type_Complementary
type_of_meal_plan_Meal Plan 3
room_type_reserved_Room_Type 3
no_of_previous_cancellations
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Relative Importance

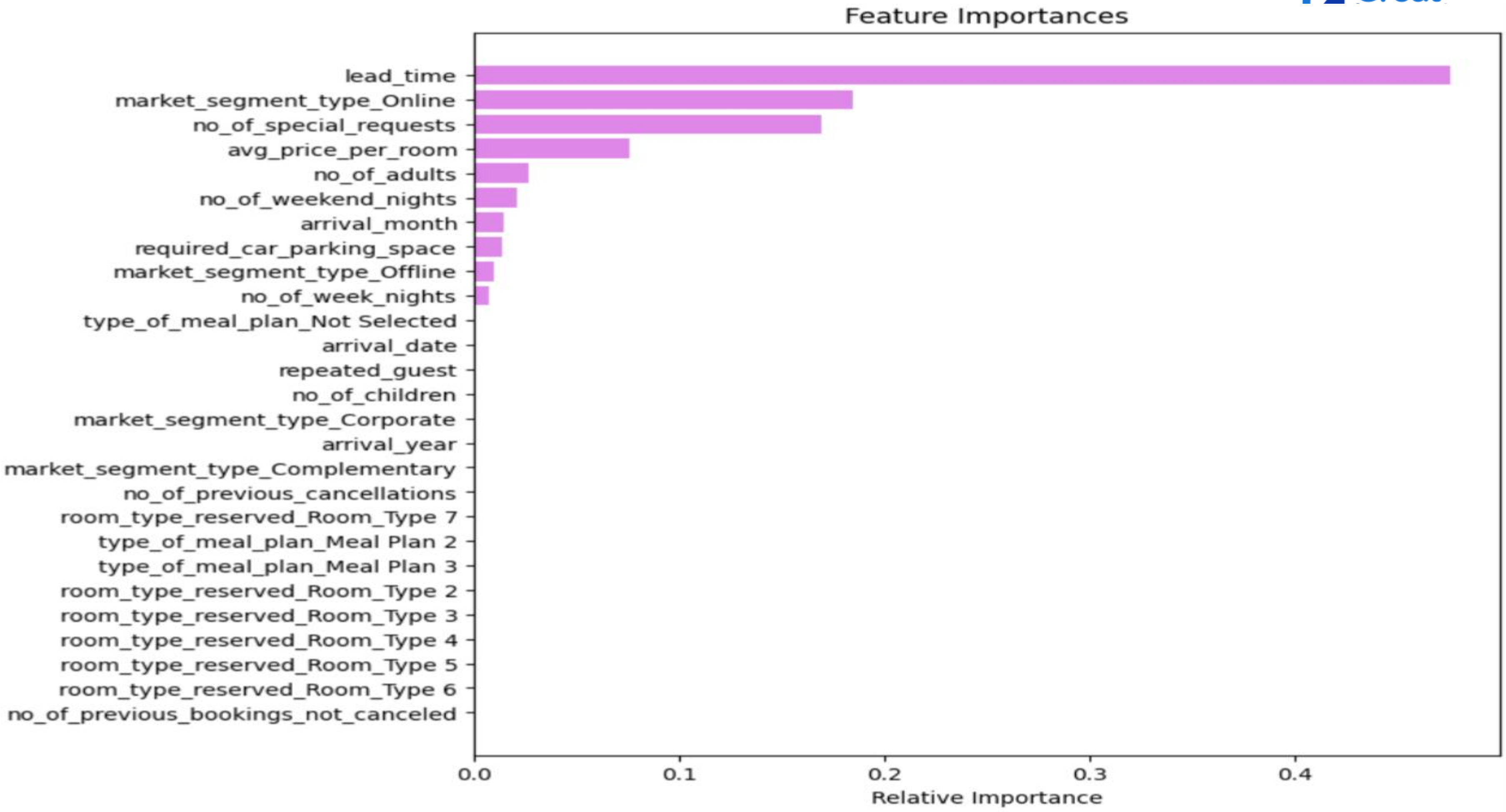
- Confusion Matrix shows:
- True Positive: ~57%
- True negative: ~26%
- False Positive: ~10%
- False Negative: ~7%

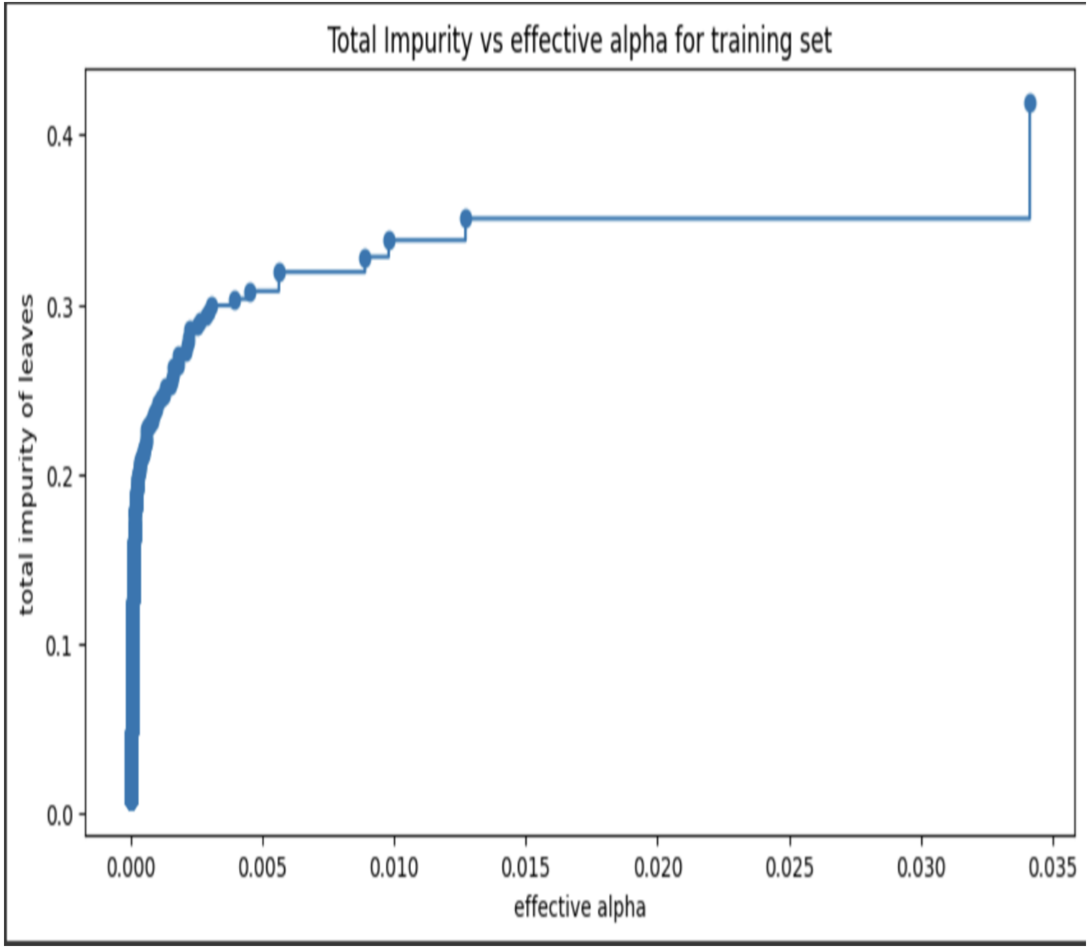




- Confusion Matrix shows:
- True Positive: ~58%
- True negative: ~25%
- False Positive: ~10%
- False Negative: ~7%


lead_time <= 151.5
gini = 0.5
samples = 25392
value = [12696.0, 12696.0]
no_of_special_requests <= 0.5
gini = 0.472
samples = 20410
value = [11676.085, 7209.531]
avg_price_per_room <= 100.04
gini = 0.264
samples = 4982
value = [1019.915, 5486.469]
market_segment_type_Online <= 0.5
gini = 0.5
samples = 10667
value = [5306.837, 5387.792]
no_of_special_requests <= 1.5
gini = 0.346
samples = 9743
value = [6369.248, 1821.739]
no_of_special_requests <= 0.5
gini = 0.417
samples = 2771
value = [954.306, 2263.51]
arrival_month <= 11.5
gini = 0.039
samples = 2211
value = [65.609, 3222.959]
lead_time <= 90.5
gini = 0.381
samples = 5395
value = [3439.976, 1185.648]
lead_time <= 13.5
gini = 0.426
samples = 5272
value = [1866.861, 4202.144
market_segment_type_Online <= 0
gini = 0.4
samples = 6679
value = [4192.983, 1601.612]
lead_time <= 90.5
gini = 0.167
samples = 3064
value = [2176.265, 220.127]
no_of_adults <= 1.5
gini = 0.338
samples = 1936
value = [517.413, 1885.499]
no_of_special_requests <= 0.5
gini = 0.454
samples = 72
value = [42.496, 22.772]
arrival_date <= 24.5
gini = 0.372
samples = 25
value = [7.456, 22.772]
gini = 0.242
samples = 20
value = [3.728, 22.772]
gini = 0.491
samples = 11
value = [5.964, 4.554]


Feature Importances
lead_time
market_segment_type_Online
no_of_special_requests
avg_price_per_room
no_of_adults
no_of_weekend_nights
arrival_month
required_car_parking_space
market_segment_type_Offline
no_of_week_nights
type_of_meal_plan_Not Selected
arrival_date
repeated_guest
no_of_children
market_segment_type_Corporate
arrival_year
market_segment_type_Complementary
no_of_previous_cancellations
room_type_reserved_Room_Type 7
type_of_meal_plan_Meal Plan 2
type_of_meal_plan_Meal Plan 3
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
no_of_previous_bookings_not_canceled
0.0
0.1
0.2
0.3
0.4
Relative Importance


Total Impurity vs effective alpha for training set
total impurity of leaves
0.0
0.1
0.2
0.3
0.4
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
effective alpha

Happy Learning !