INN Hotels Project

Supervised Learning - Classifications

July 6, 2023

Contents / Agenda

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

Executive Summary

- Summary of observations and conclusions

  - Log regression with 0.42 threshold is giving the best value (Accuracy: 80.51%, Recall: 69.96%, Precision: 69.86% and F1 score: 69.91%) compared with default threshold and 0.37 threshold. It also giving a comparable data between both training and test performance.

  - Log regression with 0.42 threshold also giving the balanced Precision-Recall Curve.

  - Post-pruning decision tree is the giving the best value (Accuracy: 88.28%, Recall: 79.05%, Precision: 83.81% and F1 score: 81.36%) compared with the pre-pruning value. It also giving the best F1 value compared with pre-pruning. Both data on training and test performance is comparable, an indication of model is not overfitting.

Business Problem Overview and Solution Approach

- Business problem overview:
  - Top 4 importance features were identified through decisions tree; which is lead_time, avg_price_per_room, market_segment_type_online and no_of_special_request. The company should look into these 4 top reasons to improve their business.
  - Price of room is very competitive during peak season
  - The highest customer who book the hotel is 2 adult and a couple with no children.
  - The breakfast meal plan becomes the highly sought after on the package of hotel booking.
  - Summer and end year holiday is the peak season, business owner should improve/comes out a plan to cater the customer booking hotel requirements.

Business Problem Overview and Solution Approach

- Solution approach/business improvement/recommendation
  - Suggested business owners to create a loyalty program for repeated customer.
  - During online booking process, the website/business owner should offering additional customizations/special request for customer, to prevent the likelihood of booking being cancelled.
  - Since most customer did not require car parking space, business owners should utilize the space to create hotel event/special activities to attract more customers.
  - Business owners may need to give vouchers/promotion/discount for meal plan 2 and 3 to attract customer to have a lunch and dinner at the hotel. Also, business owners may need to vary the lunch and dinner recipe for additional attraction. Food festival/food event also would help to increase the sales and hotel rating review.

EDA Results – Univariate Analysis

- Checking the first 5 data

	Booking_ID	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	type_of_meal_plan	required_car_parking_space	room_type_reserved	lead_time	arrival_year	arrival_month	arrival_date	market_segment_type	repeated_guest	no_of_previous_cancellations	no_of_previous_bookings_not_canceled	avg_price_per_room	no_of_special_requests	booking_status
0	INN00001	2	0	1	2	Meal Plan 1	0	Room_Type 1	224	2017	10	2	Offline	0	0	0	65.00000	0	Not_Canceled
1	INN00002	2	0	2	3	Not Selected	0	Room_Type 1	5	2018	11	6	Online	0	0	0	106.68000	1	Not_Canceled
2	INN00003	1	0	2	1	Meal Plan 1	0	Room_Type 1	1	2018	2	28	Online	0	0	0	60.00000	0	Canceled
3	INN00004	2	0	0	2	Meal Plan 1	0	Room_Type 1	211	2018	5	20	Online	0	0	0	100.00000	0	Canceled
4	INN00005	2	0	1	1	Not Selected	0	Room_Type 1	48	2018	4	11	Online	0	0	0	94.50000	0	Canceled

- Data shape – contains 36,275 rows and 19 columns

EDA Results – Univariate Analysis

```
<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
```

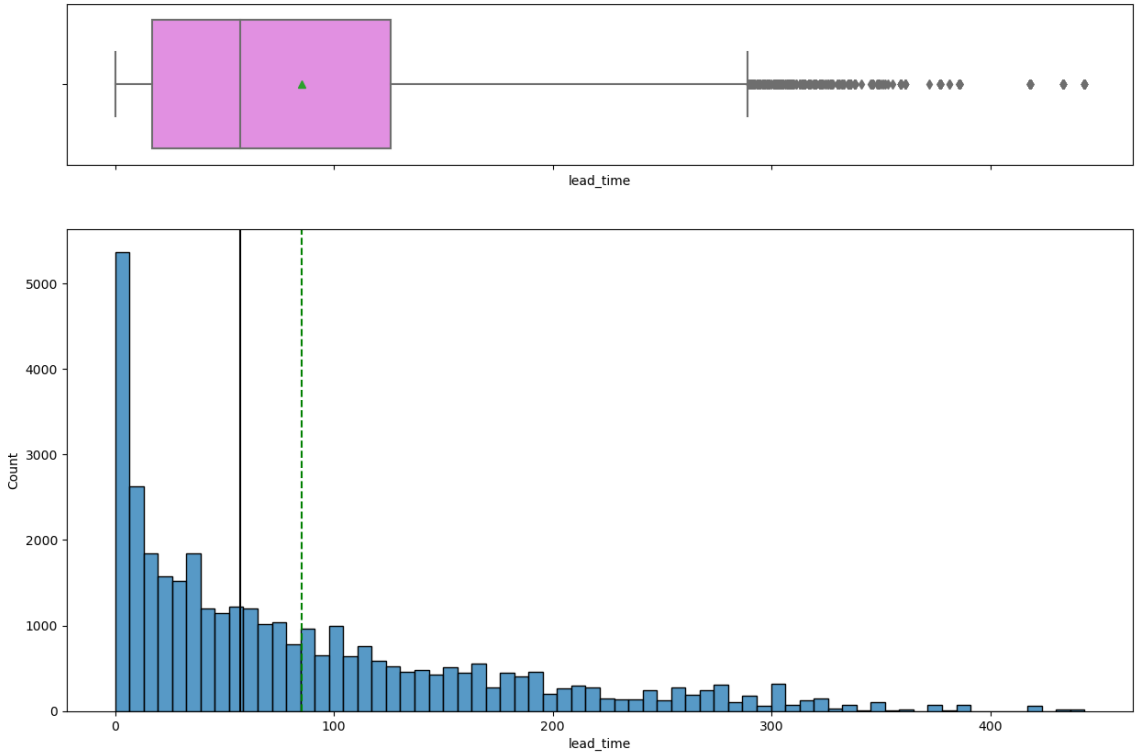
- Data info as above, with no duplicated values were found.
- Before further analysis, we dropped booking ID as it does not giving any information for data modelling.

EDA Results – Univariate Analysis

	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

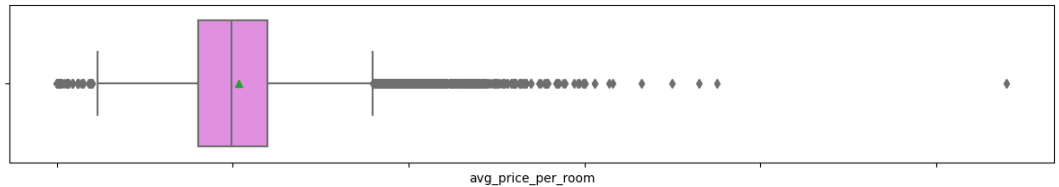
- No_of_adults are vary from one person to maximum 4 person, while no_of_children from none to maximum 10 person.
- Arrival_year can be seen from January until December with the average month is at July
- Avg_price_per_room is ranging between zero to 540 euro with average around 103.4 euro

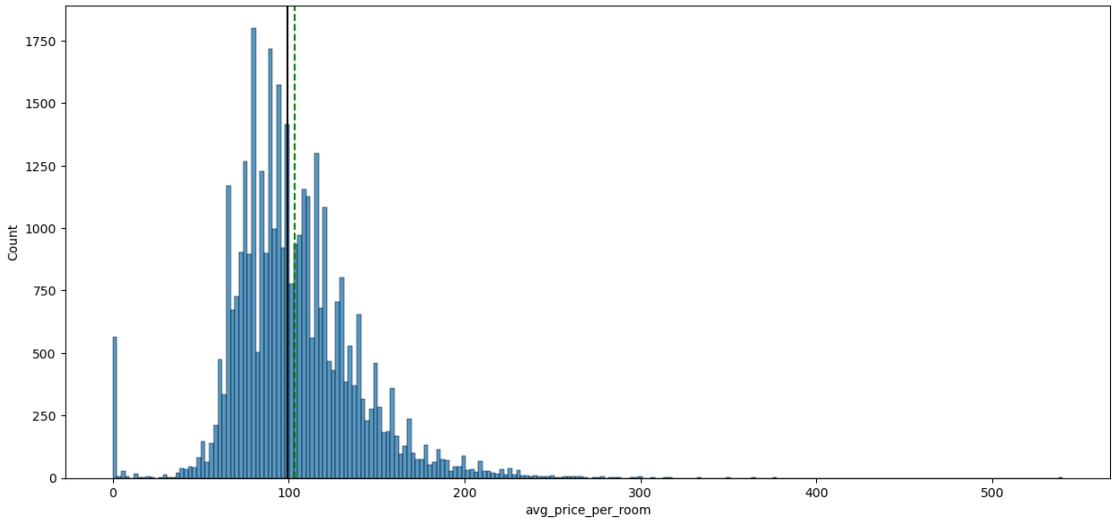
EDA Results – Univariate Analysis: 1. Lead Time



- Number of lead_time can be seen skewed to the right with some outlier.
- Mostly guests will book their hotel at last minute until 85 days.
- However, some can be seen to book early as long as long 300-400 days

EDA Results – Univariate Analysis: 2. Average Price per Room





- Avg_price_per_room can be seen quite competitive with average around 100 euros, and outlier can be seen at both sides

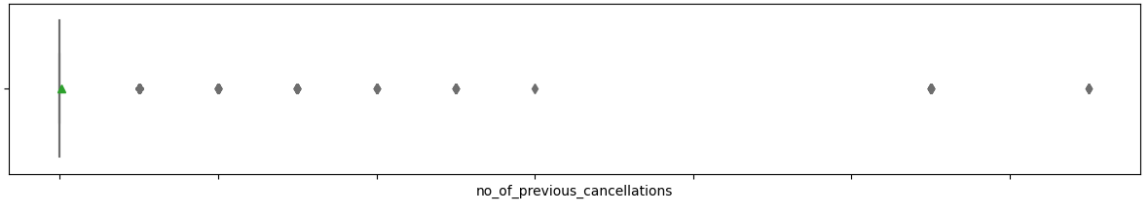
EDA Results – Univariate Analysis: 2. Average Price per Room

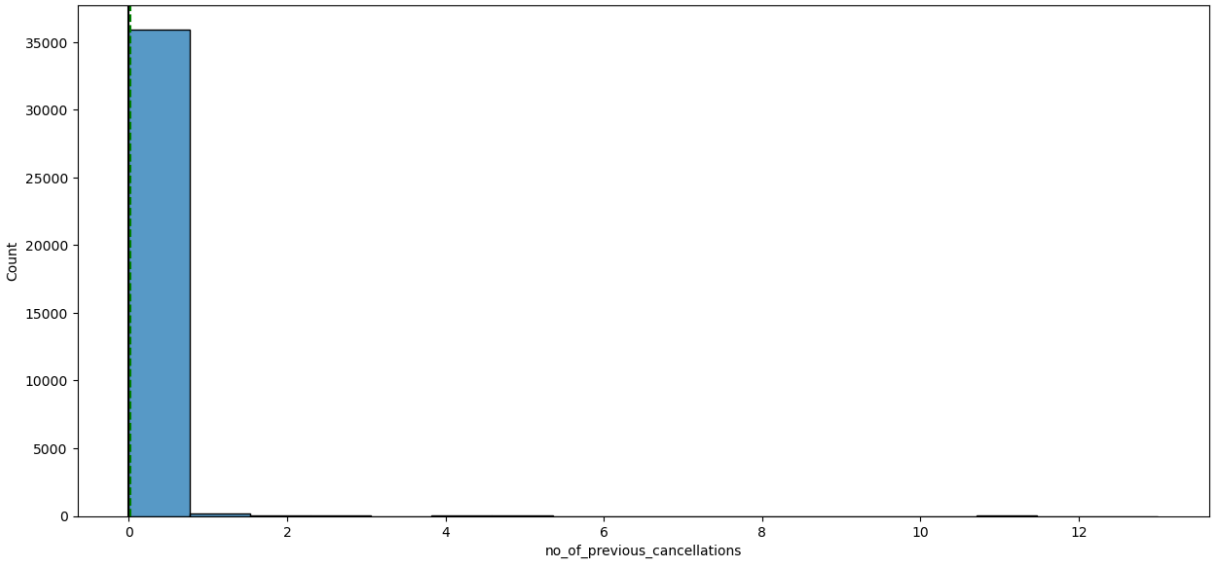
	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	type_of_meal_plan	required_car_parking_space	room_type_reserved	lead_time	arrival_year	arrival_month	arrival_date	market_segment_type	repeated_guest	no_of_previous_cancellations	no_of_previous_bookings_not_canceled	avg_price_per_room	no_of_special_requests	booking_status
63	1	0	0	1	Meal Plan 1	0	Room_Type 1	2	2017	9	10	Complementary	0	0	0	0.00000	1	Not_Canceled
145	1	0	0	2	Meal Plan 1	0	Room_Type 1	13	2018	6	1	Complementary	1	3	5	0.00000	1	Not_Canceled
209	1	0	0	0	Meal Plan 1	0	Room_Type 1	4	2018	2	27	Complementary	0	0	0	0.00000	1	Not_Canceled
266	1	0	0	2	Meal Plan 1	0	Room_Type 1	1	2017	8	12	Complementary	1	0	1	0.00000	1	Not_Canceled
267	1	0	2	1	Meal Plan 1	0	Room_Type 1	4	2017	8	23	Complementary	0	0	0	0.00000	1	Not_Canceled

```
Complementary    354
Online           191
Name: market_segment_type, dtype: int64
```

- Avg_price_per_room can be seen from sides, which is from online and complementary

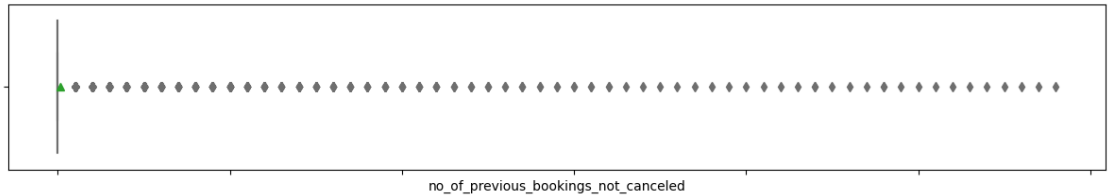
EDA Results – Univariate Analysis: 3. Number of Previous Booking Cancellations

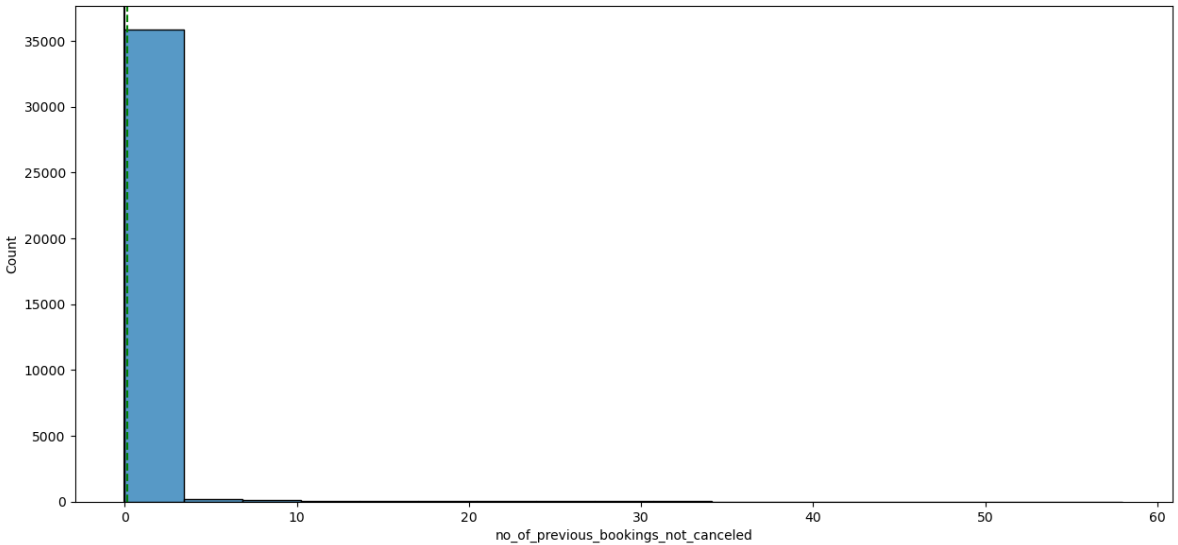




- It can be seen most customer did not cancelled their bookings.

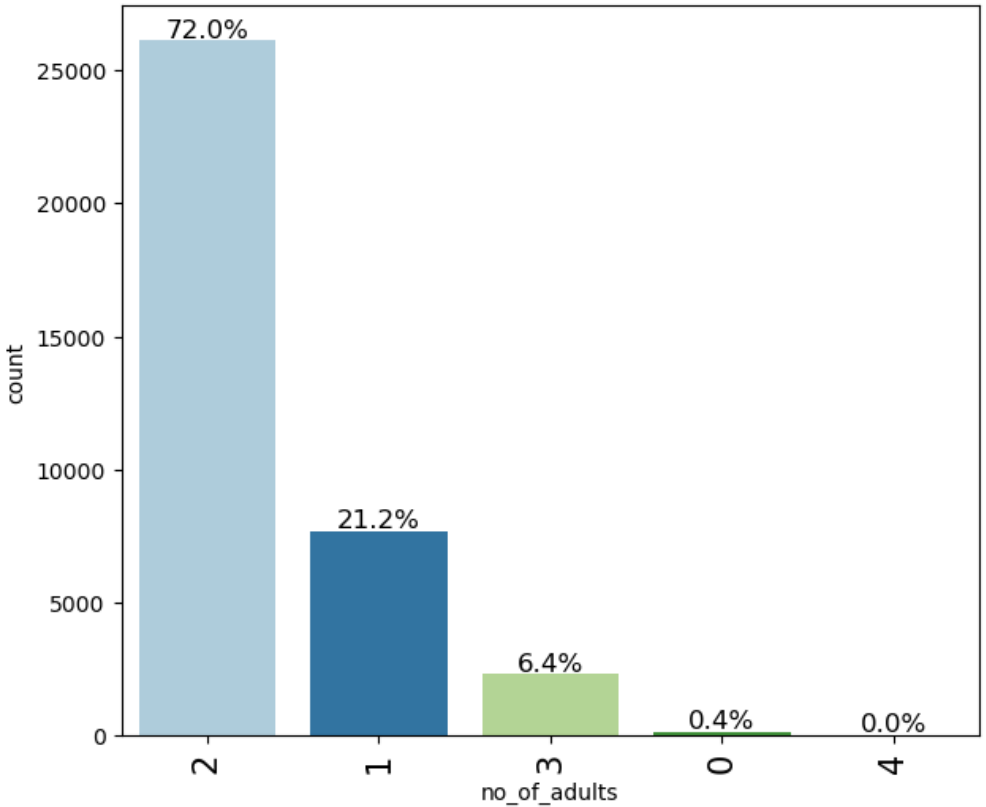
EDA Results – Univariate Analysis: 4. Number of previous booking not cancelled





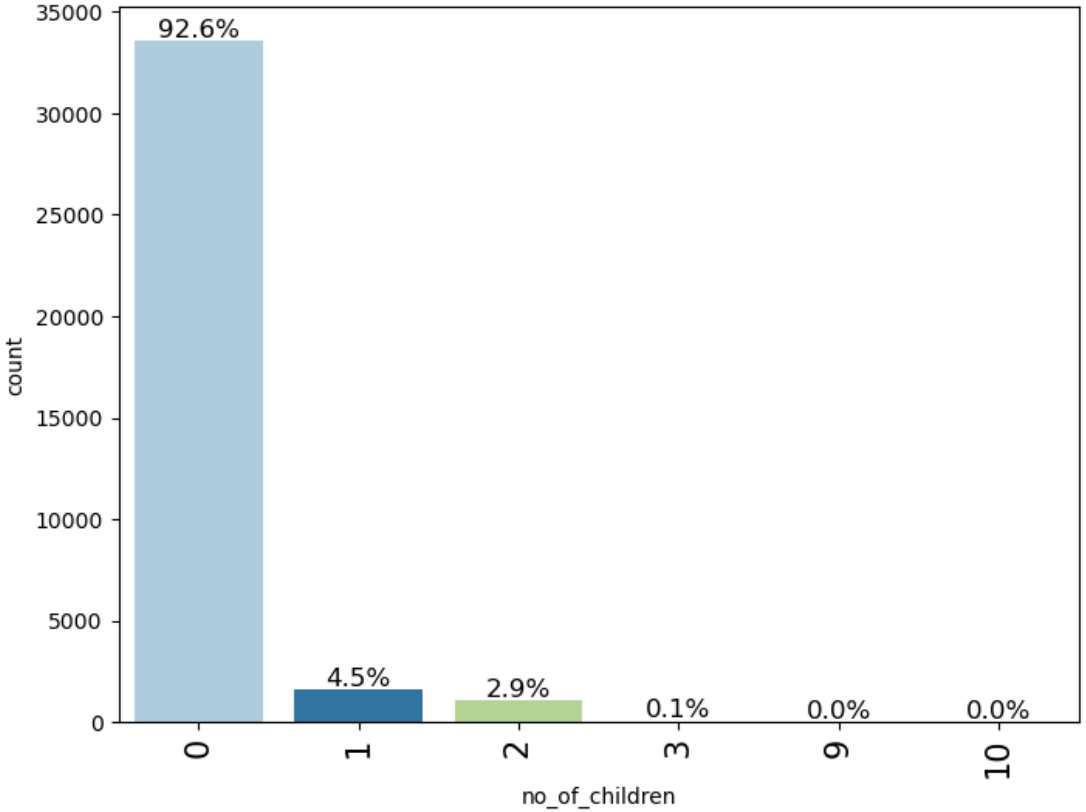
- This result is supported with the same finding on previous slide (slide#11).
- Most customer did not cancelled their bookings.

EDA Results – Univariate Analysis: 5. Number of adults



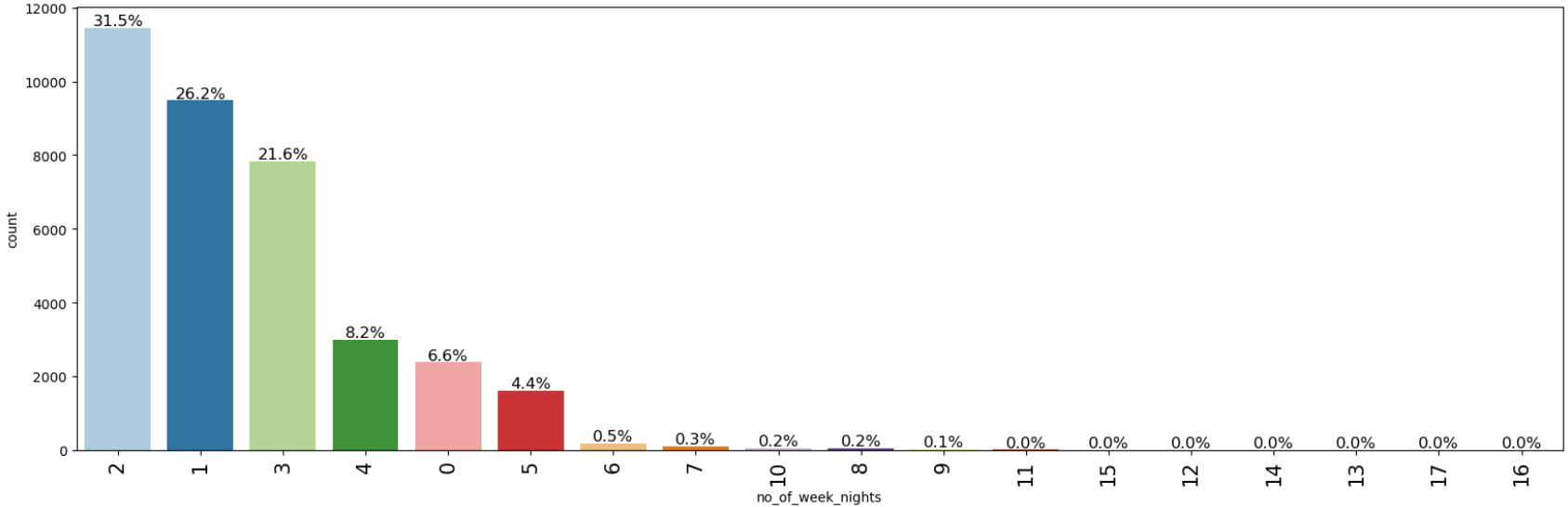
- Maximum number of adults to book a room is two person, followed by one person.

EDA Results – Univariate Analysis: 6. Number of children



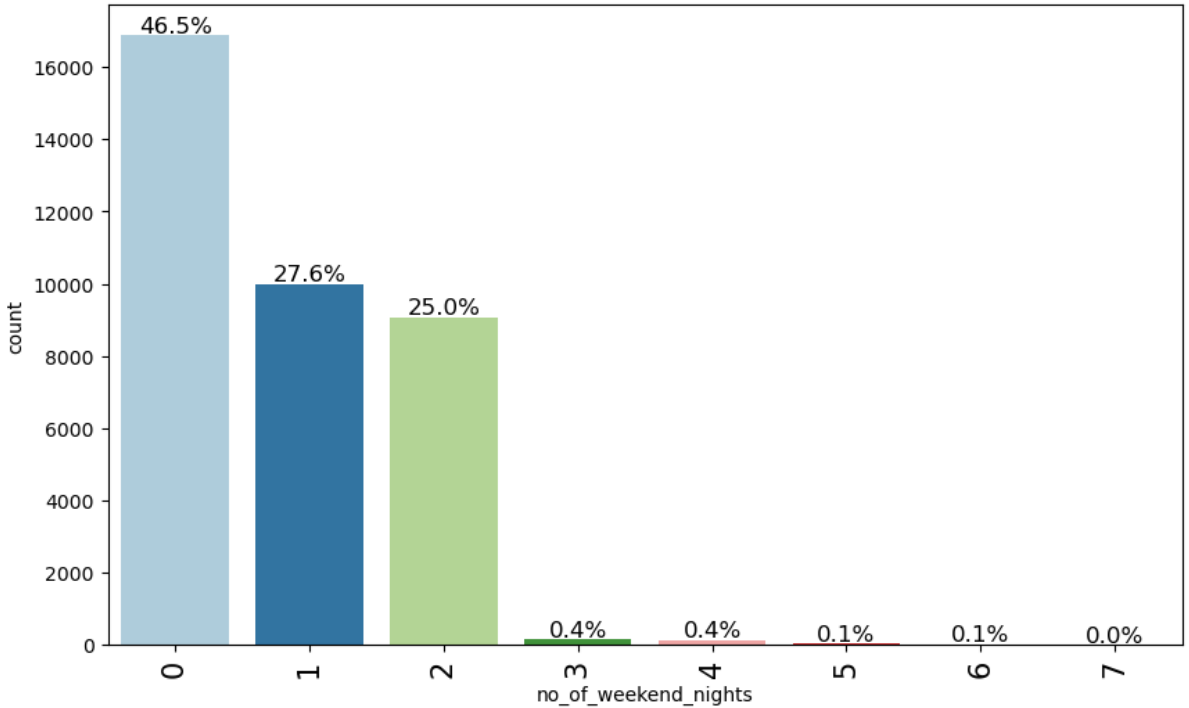
- A couple with no children becomes a top detractor, which is around 92.6%, followed by a couple with 1 child, around 4.5% and 2 children, around 2.9%

EDA Results – Univariate Analysis: 7. Number of week nights



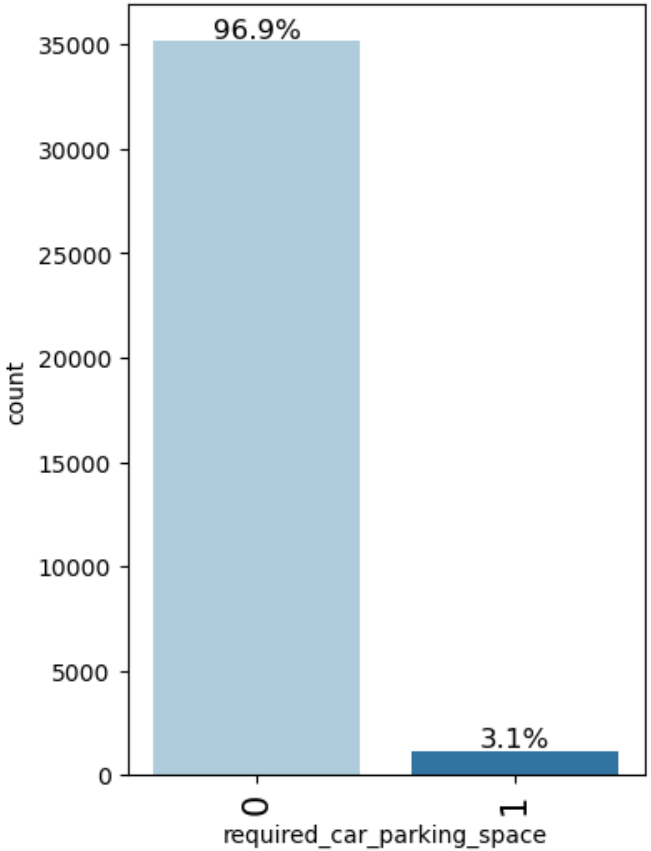
- Top night for holiday: 2 nights, followed with 1 night only
- Maximum night customer book the hotel: 7 nights

EDA Results – Univariate Analysis: 8. Number of weekend nights



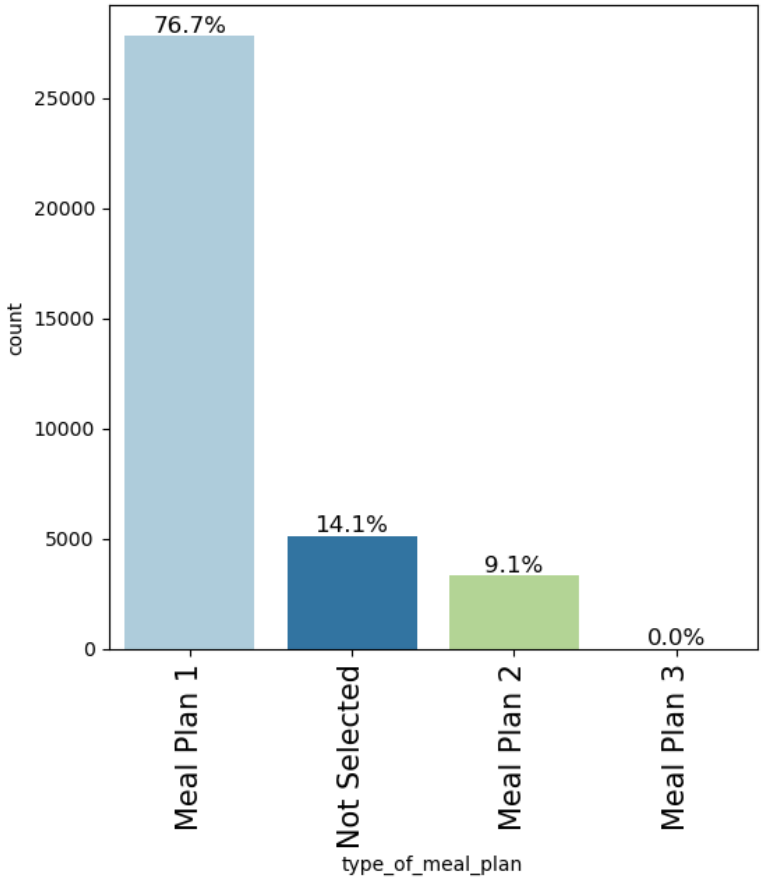
- Estimated around 46.5% of customer book the hotel not on weekend nights, while 27.6% on 1 night and 25% only on 2 nights

EDA Results – Univariate Analysis: 9. Required car parking space



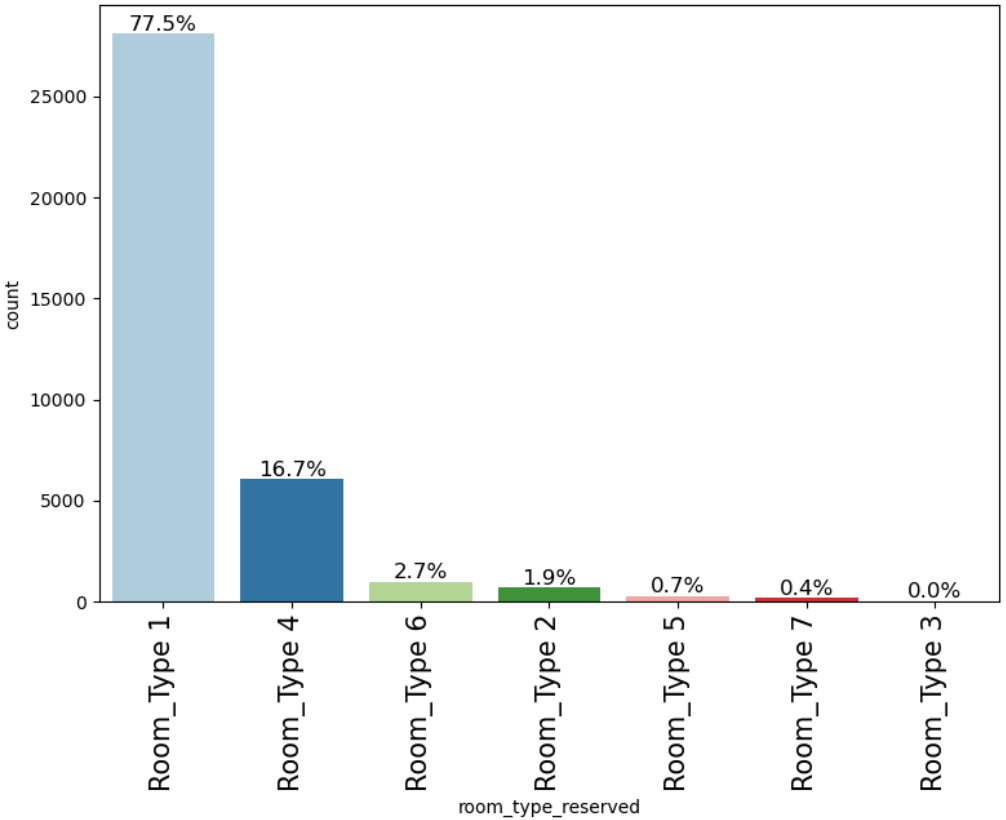
- Majority of customer (around 96.9%) did not require parking space when they book the hotel, probably coming by using public transport, while only 3.1% require car parking space at the hotel.

EDA Results – Univariate Analysis: 10. Type of meal plan



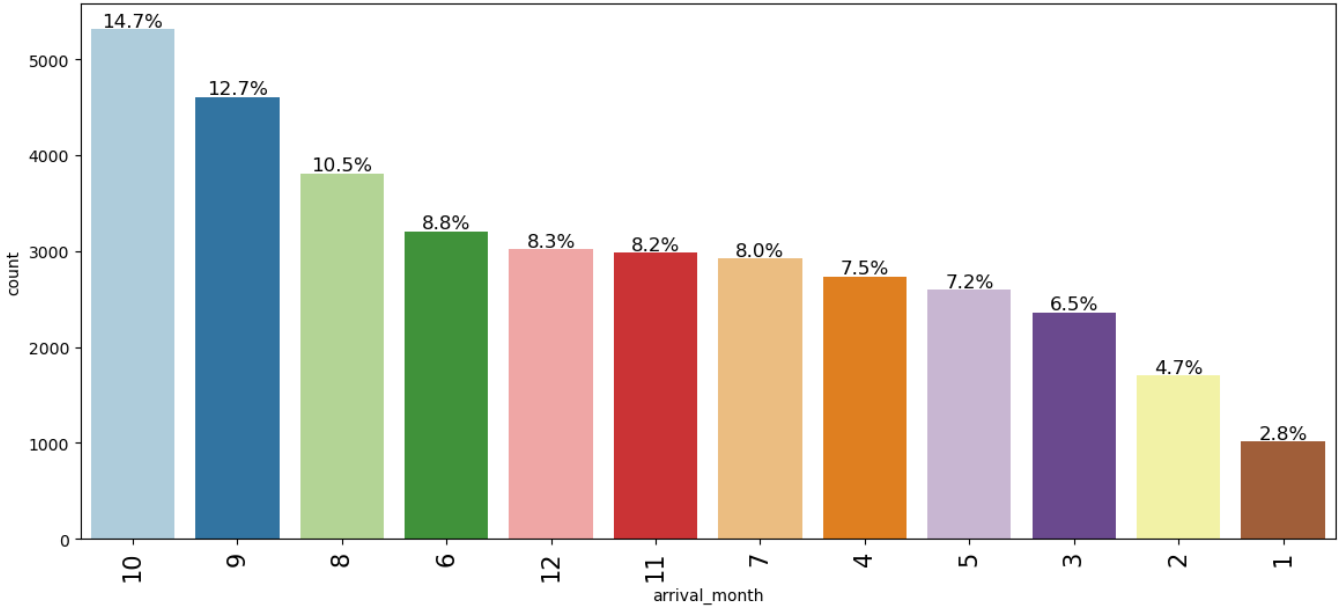
- Majority customer, estimated 76.7% book meal plan 1, which is breakfast meal, probably due to personal activities during daytime or checking out the hotel.
- Other 14.1% did not select any breakfast/meal plan, only 9.1% book meal plan 2.

EDA Results – Univariate Analysis: 11. Type of room observed



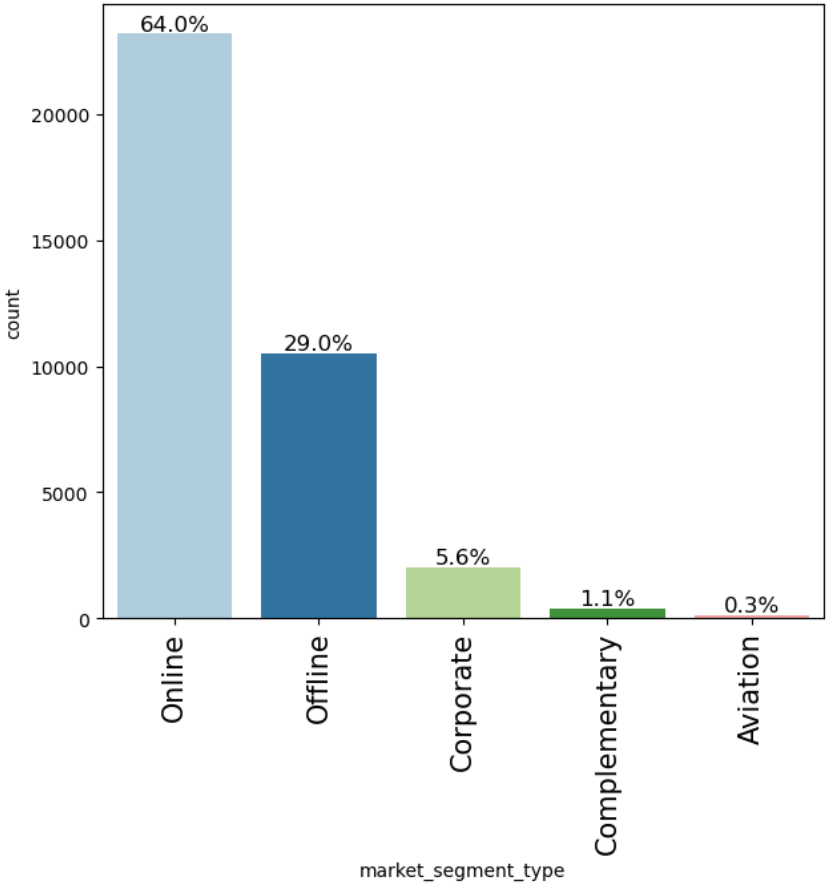
- Majority customer around 77.5% prefers room type 1, followed by room type 4 (16.7%) and other small contributions are coming from room type 6, type 2, type 5 and 7.

EDA Results – Univariate Analysis: 12. Arrival month



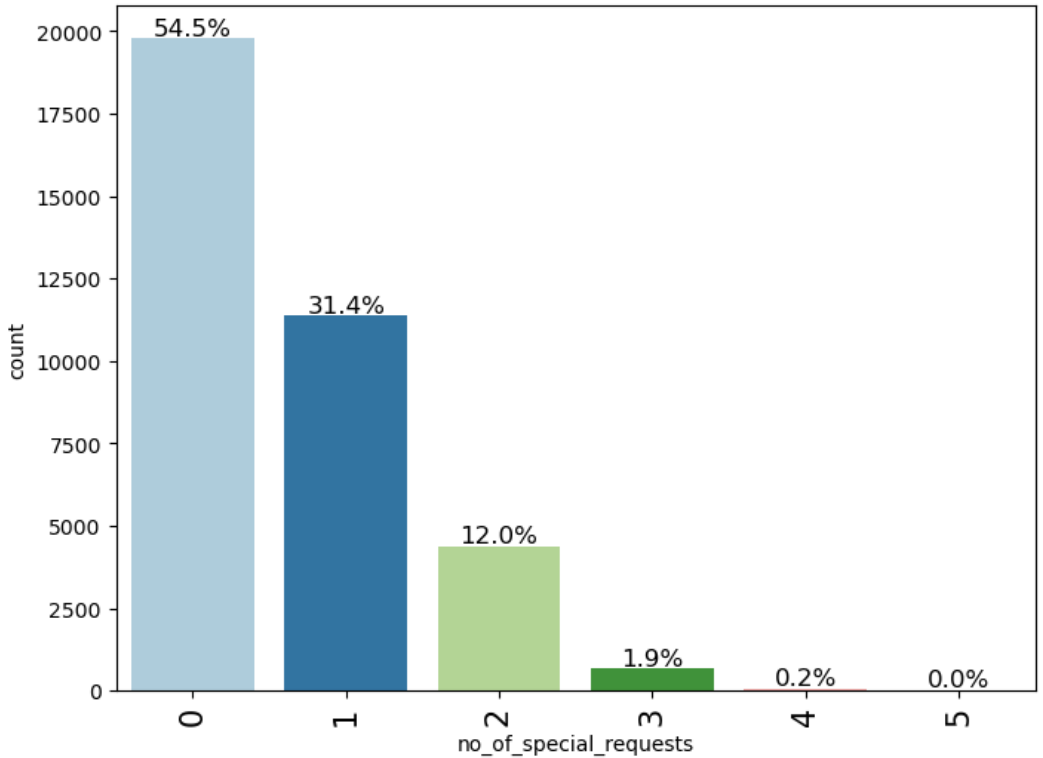
- The arrival month during hotel booking is vary. The highest contribution is from October (14/7%), followed by September (12.7%), August (10.5%) and June (8.8%). All the months is known for long summer holiday
- End year holiday is followed closely by December and November month, respectively contribute about 8.3% and 8.2%.

EDA Results – Univariate Analysis: 13. Market segment type



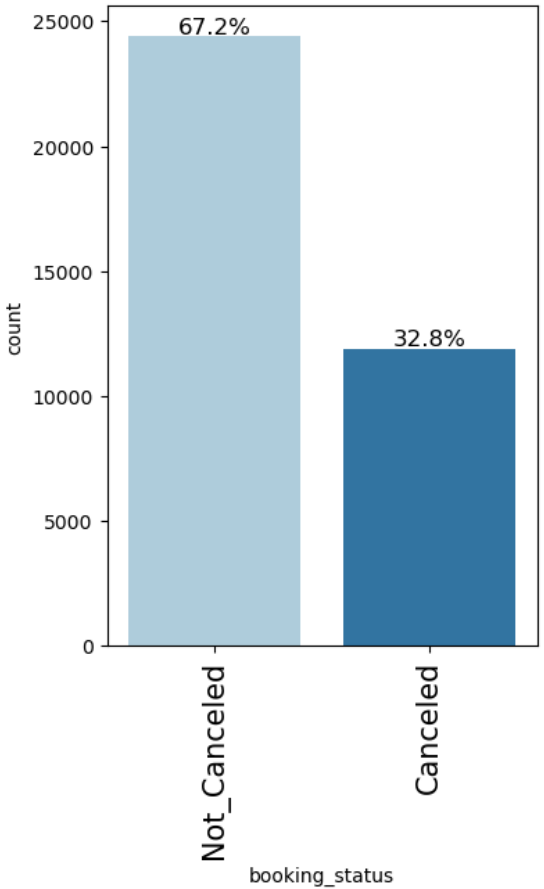
- Online booking become the highest contribution, estimated around 64.0% and offline estimated around 29.0%.
- This may be due to customer has made the holiday plan ahead and book the hotel early.

EDA Results – Univariate Analysis: 14. Number of special requests



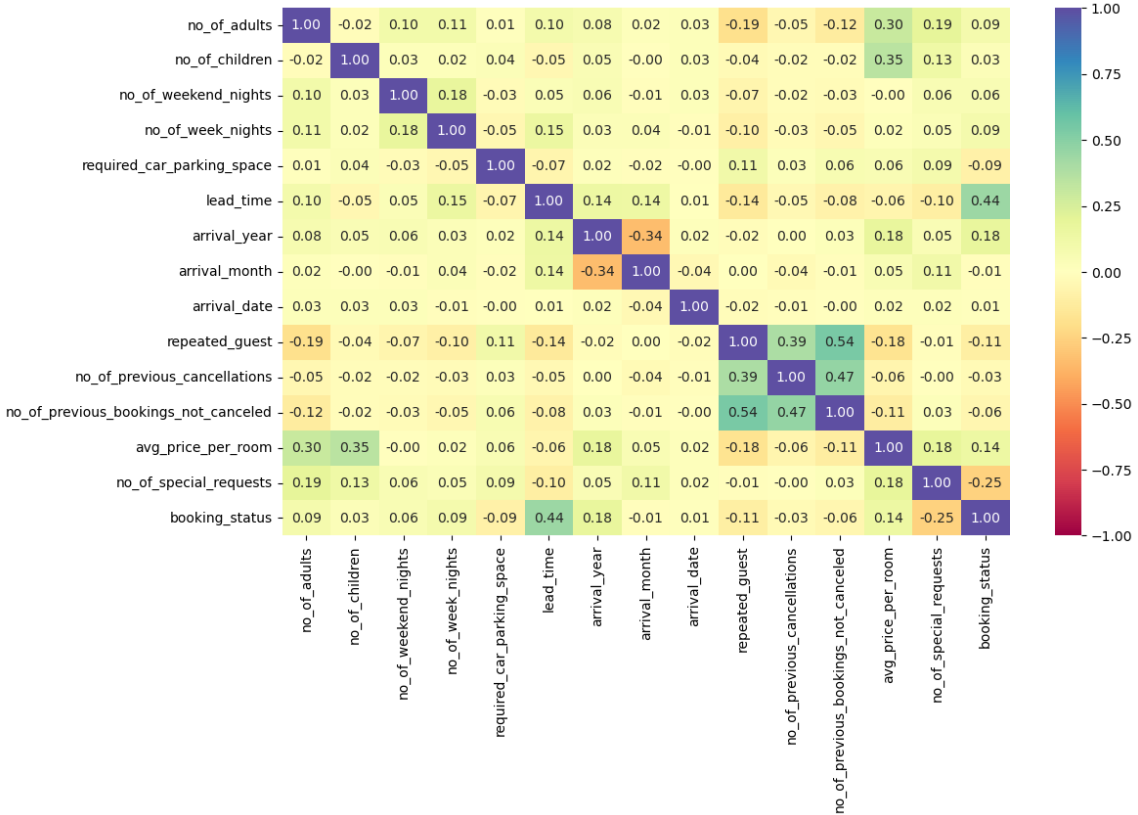
- Half percentage of customer, estimated around 54.5% did not have any special request during their booking, while there is 31.4% of 1 special request from customer to the booked hotel.
- The maximum request made by customer is only 3 request, which contribute to 1.9%.

EDA Results – Univariate Analysis: 15. Booking status



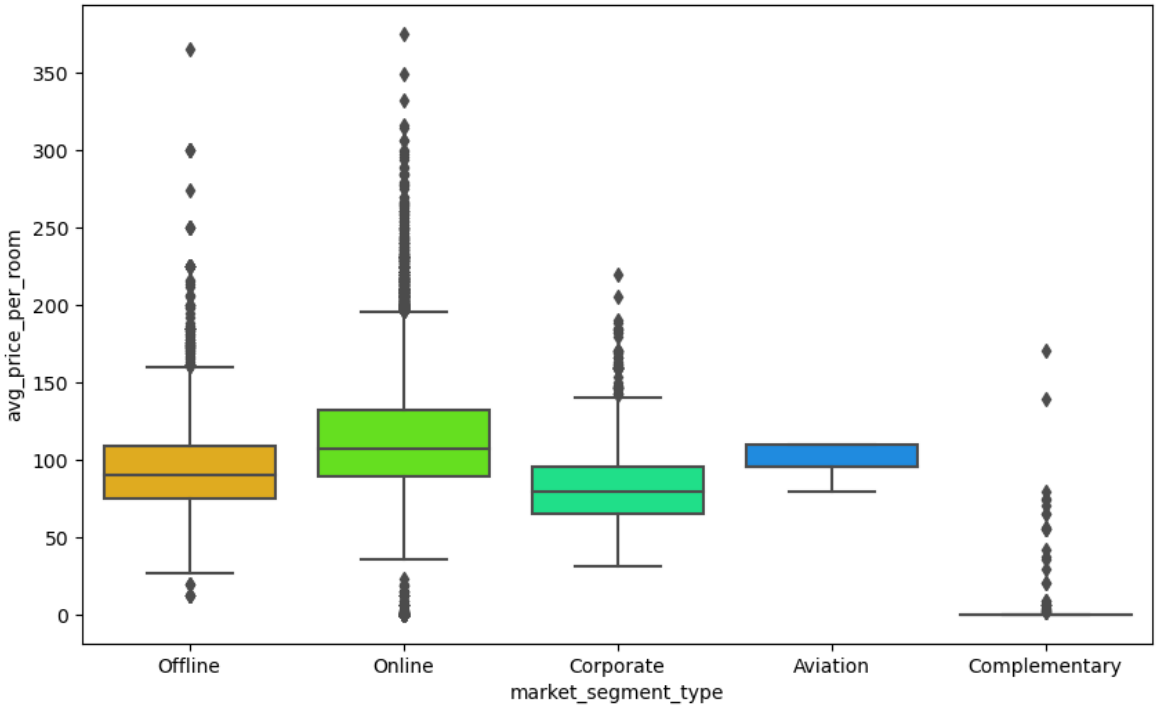
- Estimated 67.2% customer did not cancelled their hotel booking, while 32.8% has made cancellation for the book.

EDA Results – Bivariate Analysis



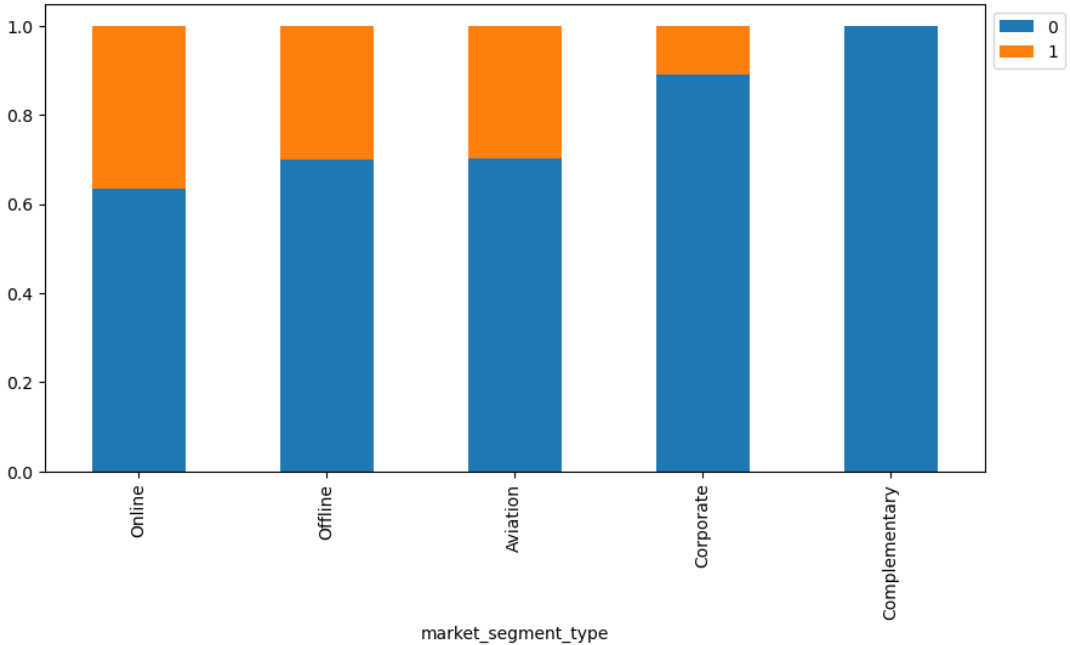
- Most of the parameter is having low to moderate relationship.
- However, moderate relationship is noticed is on repeated_guest, no_of_previous_cancellations and no_of_previous_bookings_not_canceled

EDA Results – Bivariate Analysis: 1. Prices vary across different market segment



- Outlier price was noticed on offline, online, corporate and complementary segment.
- Corporate and complementary segment is having right skewed.

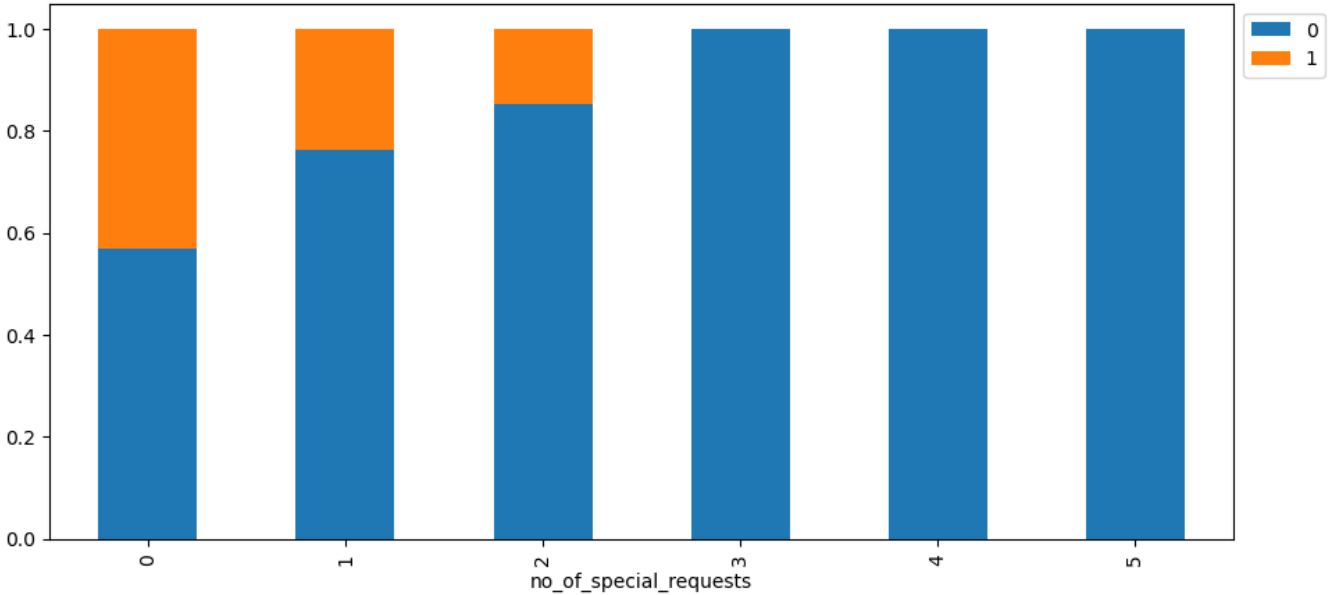
EDA Results – Bivariate Analysis: 2. Average price per room impacts booking status



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

- All booking segment is not affected by price per room, especially on complementary segment. The customer book the hotel mostly looking into other things besides room price.

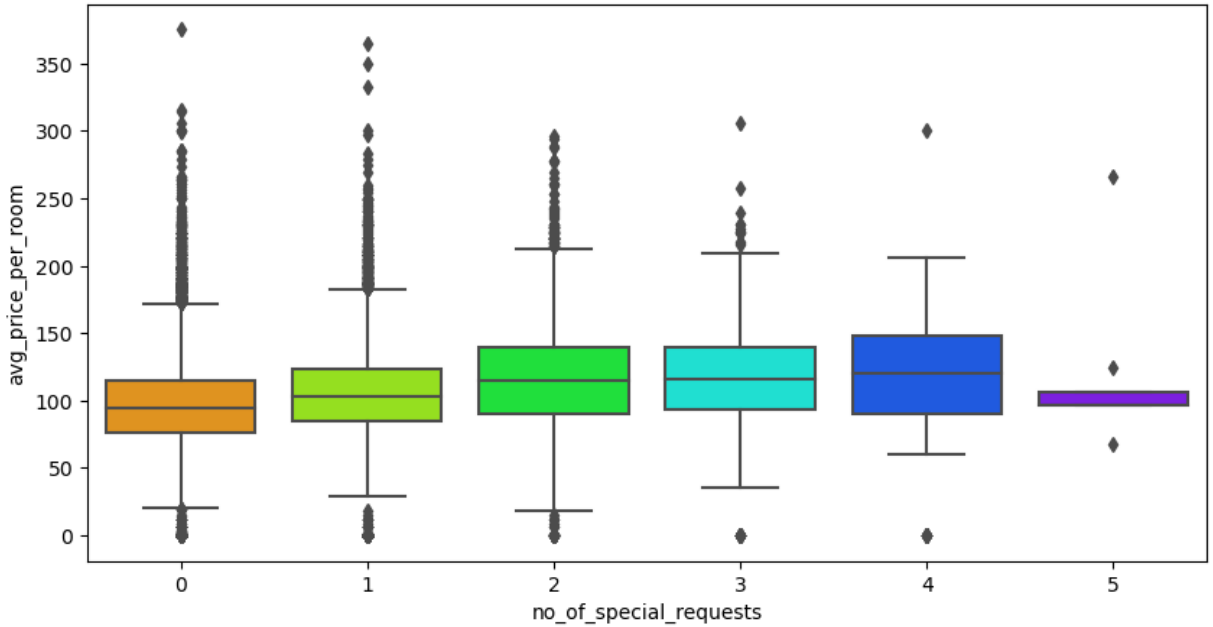
EDA Results – Bivariate Analysis: 3. Impact cancellations



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

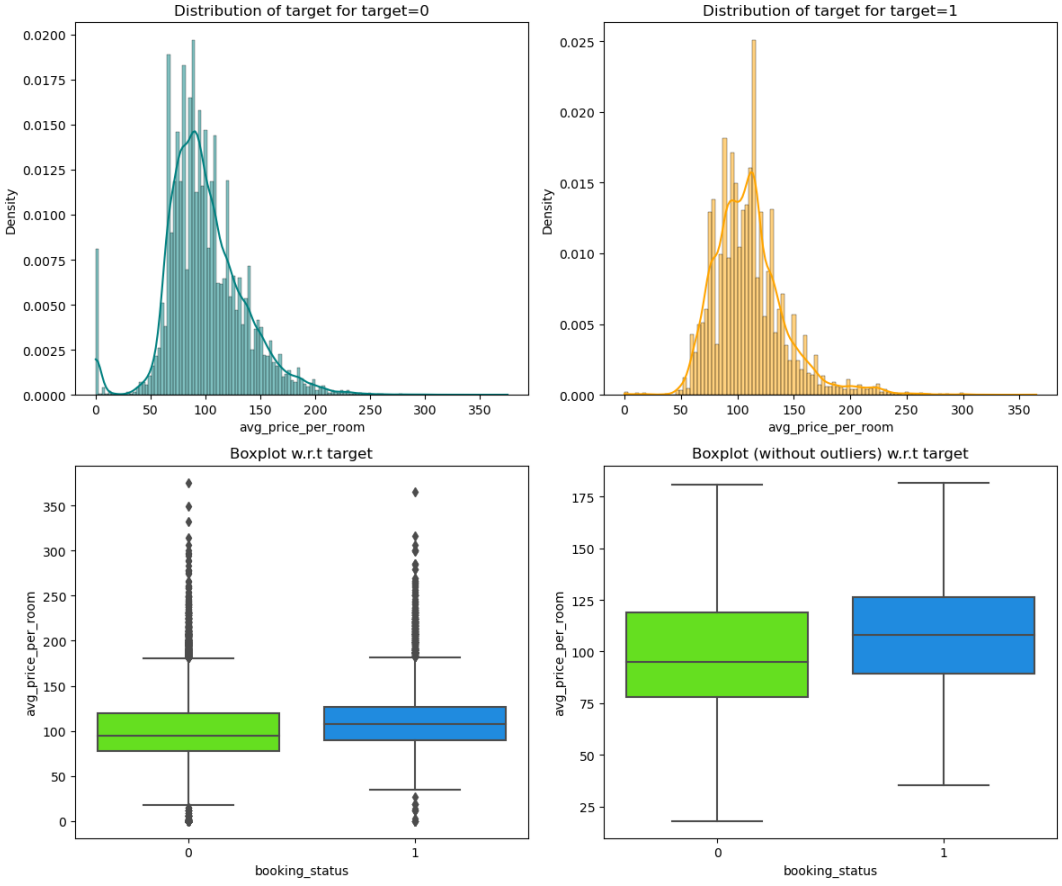
- No impact on booking cancellations were observed, as long as customer requests are being fulfilled.

EDA Results – Bivariate Analysis: 4. Customers impacts the prices of a room



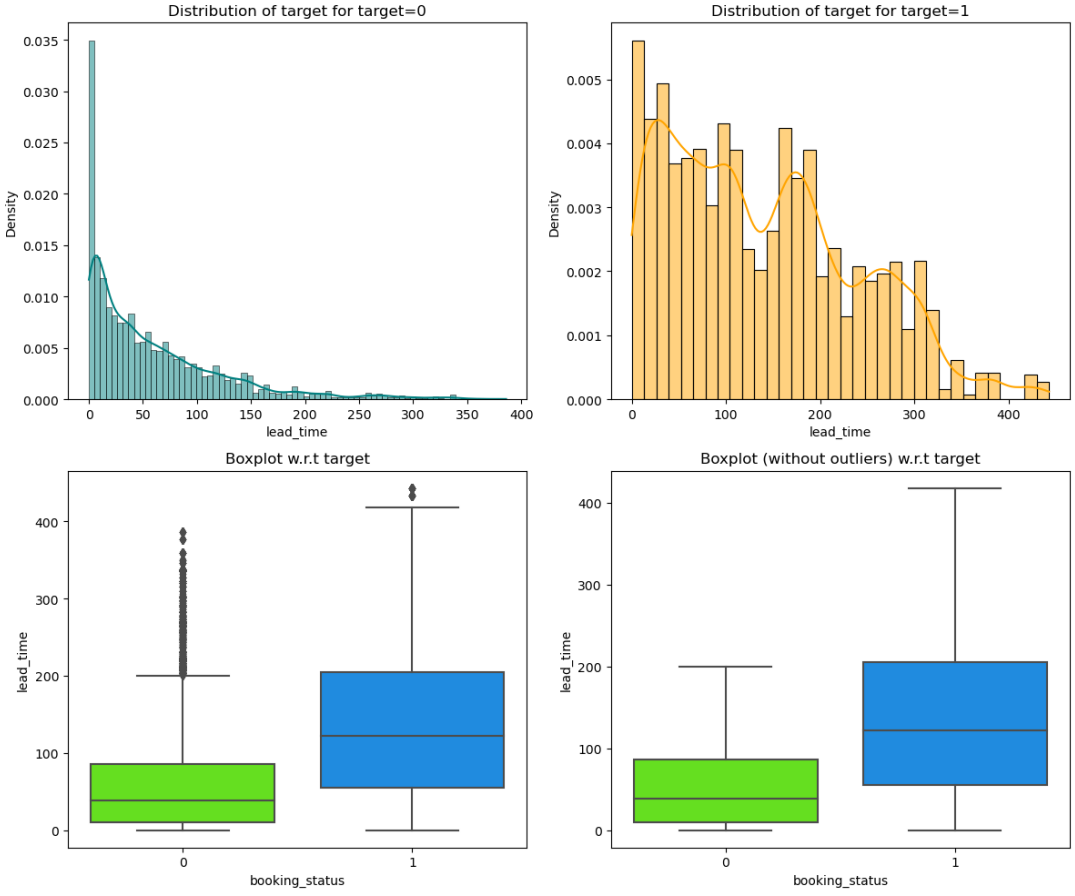
- All requests are observed having quite comparable average price per room, despite all of the data is right skewed.

EDA Results – Bivariate Analysis: 5. Positive correlation between booking status and average price per room



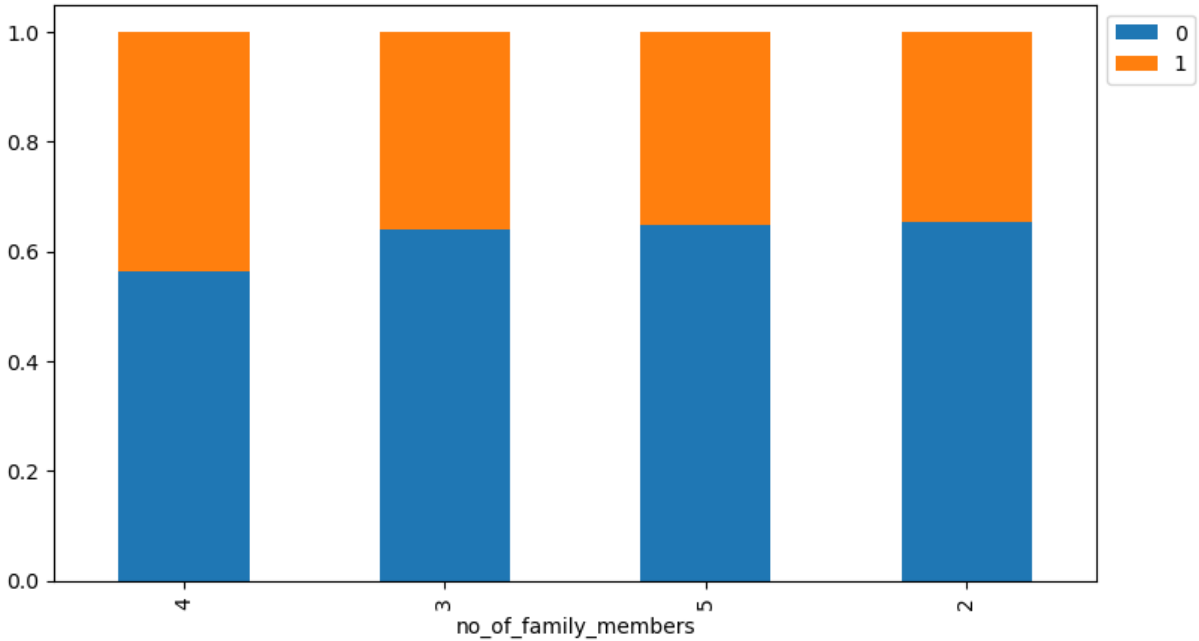
- Both booking status are having normal distribution (normal bell-curved) even after outlier removed. Both booking status is having comparable average price per room.

EDA Results – Bivariate Analysis: 6. Positive correlation between booking status and lead time



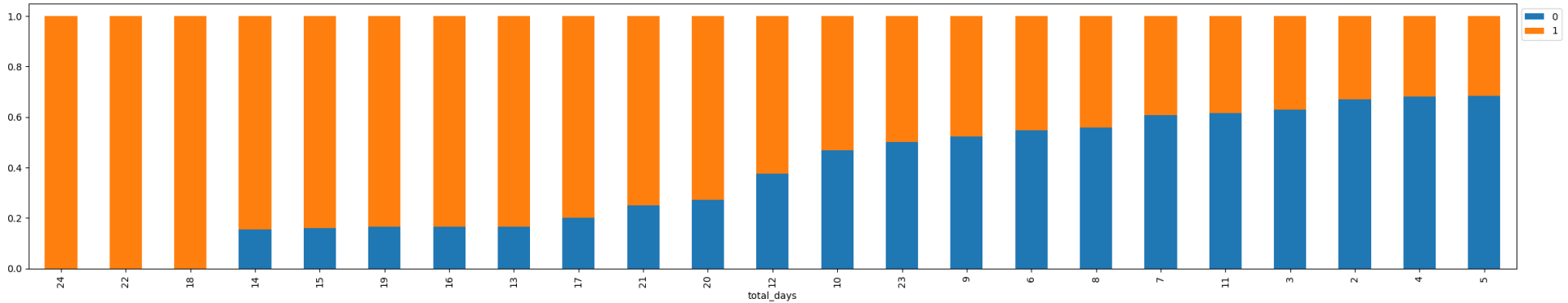
- Both booking type (cancelled and not cancelled having some outliers, and after outlier removed, found cancelled booking status is having wider range compared with customer not cancelling their booking.

EDA Results – Bivariate Analysis: 7. Travelling with families



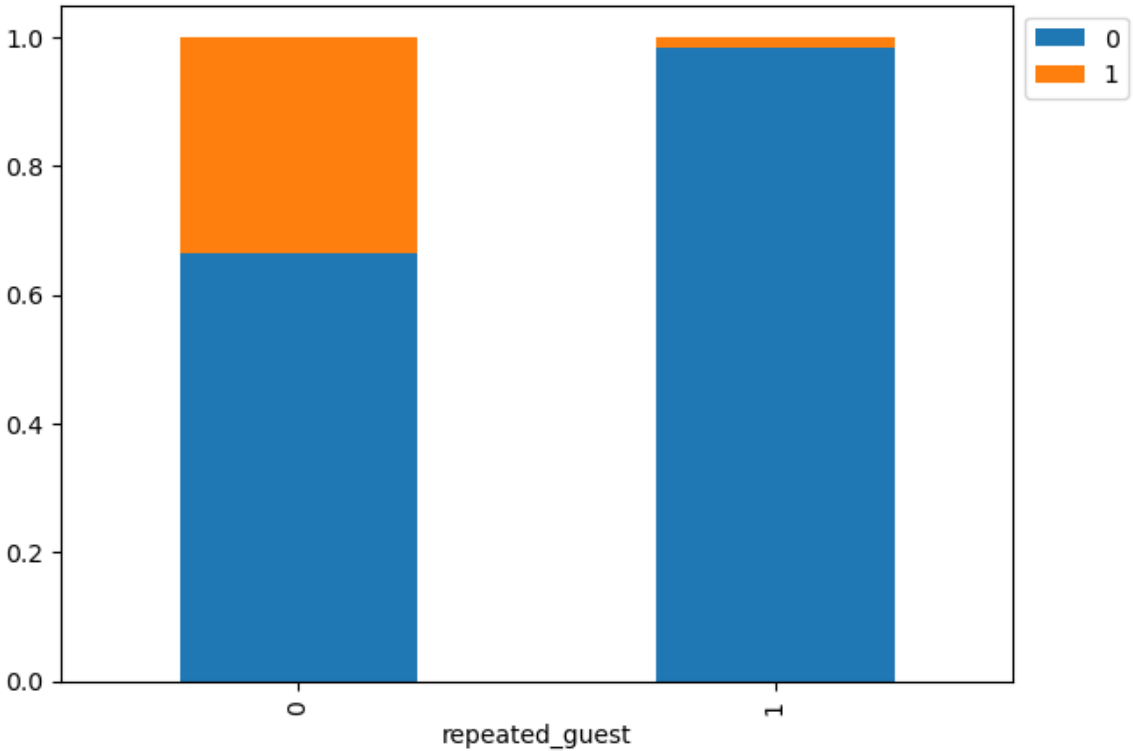
- All family members regardless the family quantity did not affect any booking status.

EDA Results – Bivariate Analysis: 8. Stay at least one day at the hotel



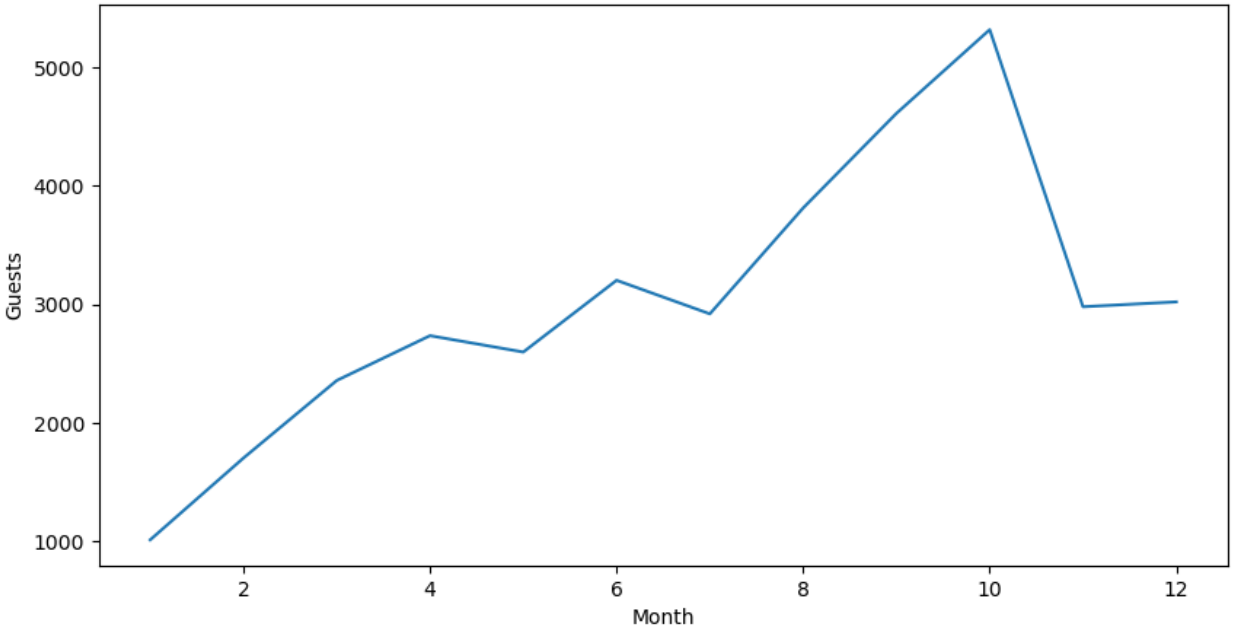
- The most customer is booked and stay at the hotel is 24days, while the minimum days is 5.
- Most customer is spending a long holiday at the hotel.

EDA Results – Bivariate Analysis: 9. Repeated guests cancel booking



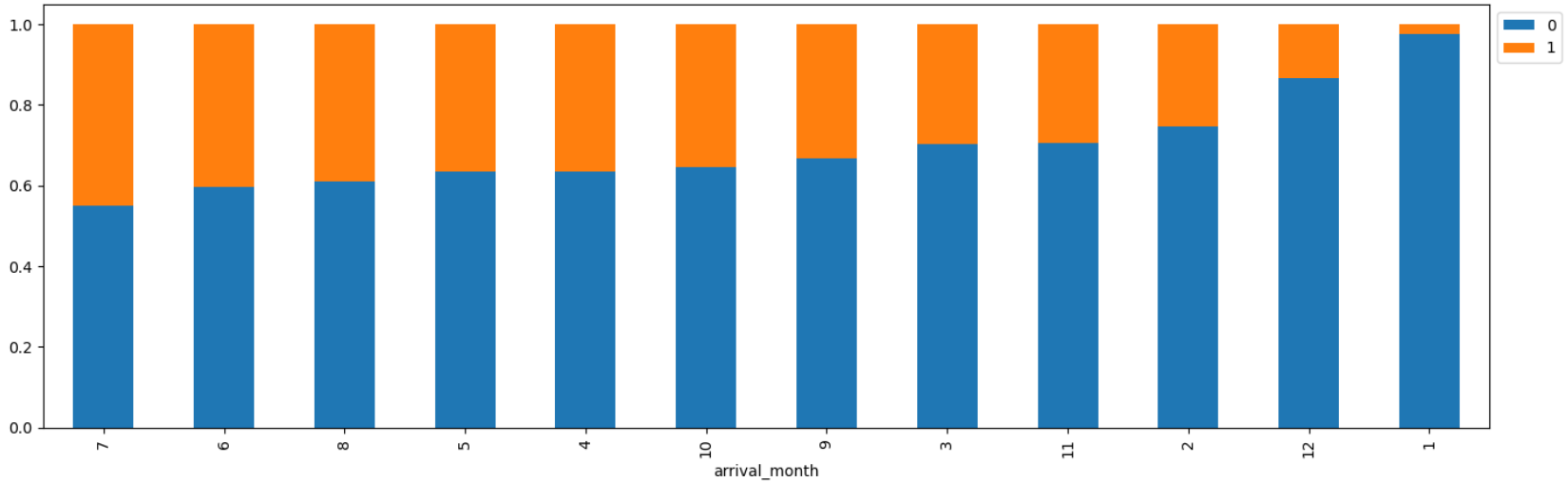
- Less customer was observed to cancelled their hotel booking.

EDA Results – Bivariate Analysis: 10. The busiest month in the hotel



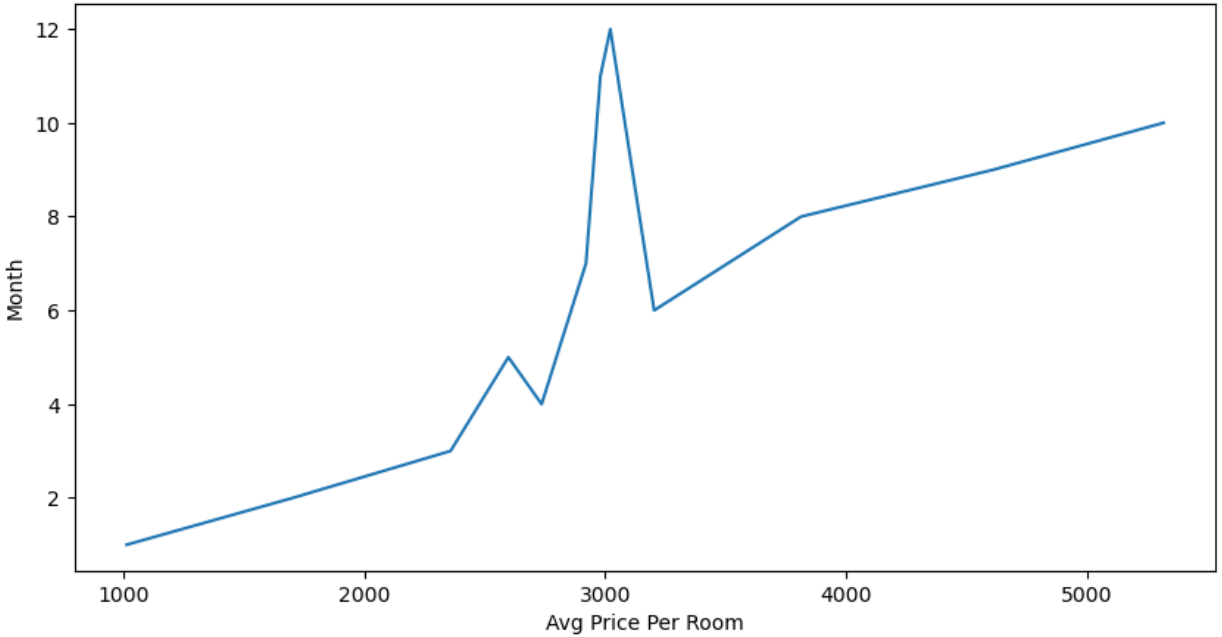
- The busiest month was observed during summer holiday, with the peak month is at October.
- End holiday is observed to pick up the similar trend with Jun-July guests.

EDA Results – Bivariate Analysis: 11. Bookings cancelled in each month



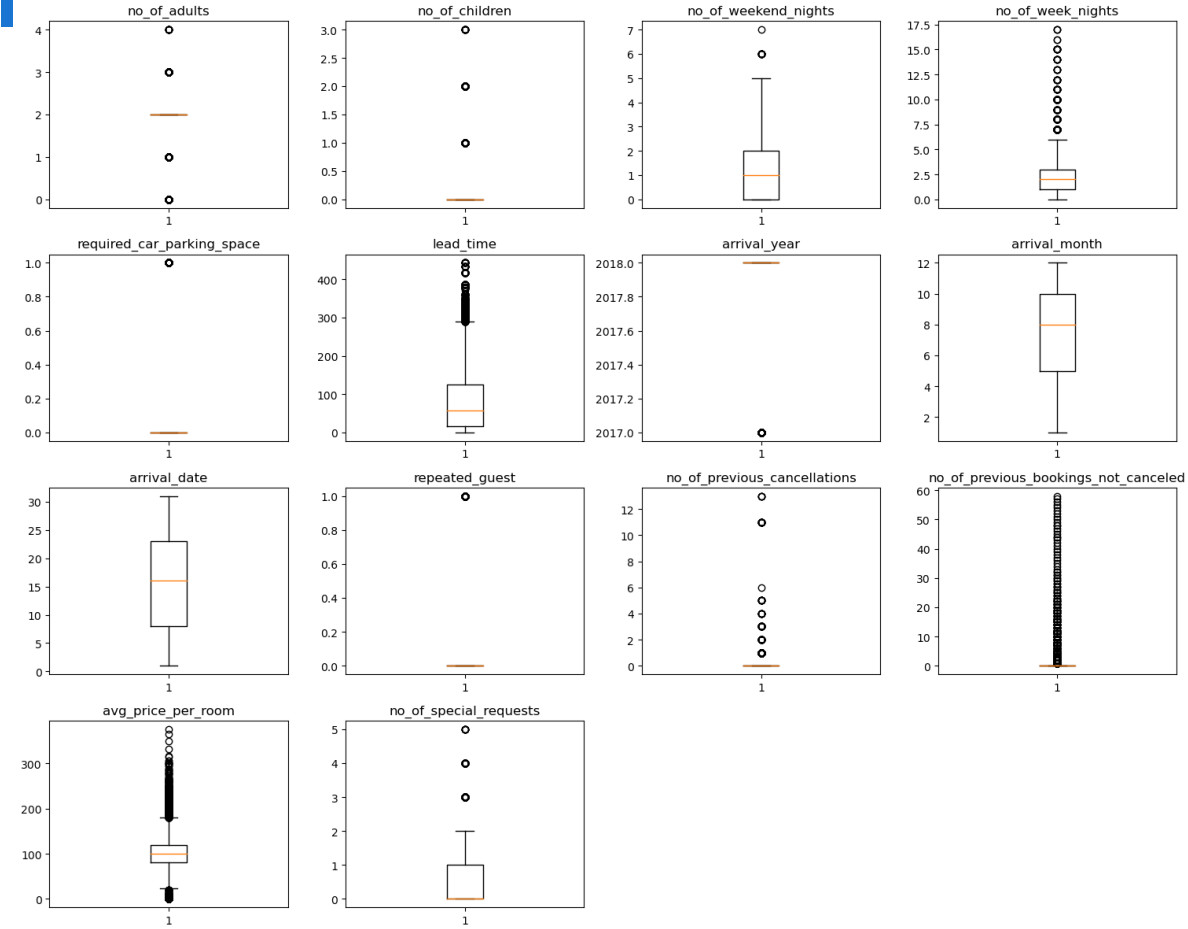
- The highest quantity bookings were cancelled are from summer holiday (top 3 detractor is from July, June and Aug)
- The lowest was seen on Dec and Jan.

EDA Results – Bivariate Analysis: 12. Hotel room prices vary across different months



- End holiday is at the peak with average price 3000 euros and 5000 euros.

Data Preprocessing – 1. Outlier Check



- Few outliers found, but we are not going to remove the data, as we need this outlier data for our modelling later

Data Preprocessing – 2. Building Logistic Regression and Model Performance Check

- Model result

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25364
Method:	MLE	Df Model:	27
Date:	Wed, 28 Jun 2023	Pseudo R-squ.:	0.3292
Time:	13:01:42	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

- Training Performance

	Accuracy	Recall	Precision	F1
0	0.80600	0.63410	0.73971	0.68285

- First log model is having moderate percentage of prediction.

Checking Logistic Regression Assumption – 3. Checking multicollinearity

- Checking for VIF – total 28 of VIF factor, 3 of them is higher than 5 (market_segment_type_Corporate, market_segment_type_Offline and market_segment_type_Online)

	feature	VIF
0	const	39497686.20788
1	no_of_adults	1.35113
2	no_of_children	2.09358
3	no_of_weekend_nights	1.06948
4	no_of_week_nights	1.09571
5	required_car_parking_space	1.03997
6	lead_time	1.39517
7	arrival_year	1.43190
8	arrival_month	1.27633
9	arrival_date	1.00679
10	repeated_guest	1.78358
11	no_of_previous_cancellations	1.39569
12	no_of_previous_bookings_not_canceled	1.65200
13	avg_price_per_room	2.06860
14	no_of_special_requests	1.24798
15	type_of_meal_plan_Meal Plan 2	1.27328
16	type_of_meal_plan_Meal Plan 3	1.02526
17	type_of_meal_plan_Not Selected	1.27306
18	room_type_reserved_Room_Type 2	1.10595
19	room_type_reserved_Room_Type 3	1.00330
20	room_type_reserved_Room_Type 4	1.36361
21	room_type_reserved_Room_Type 5	1.02800
22	room_type_reserved_Room_Type 6	2.05614
23	room_type_reserved_Room_Type 7	1.11816
24	market_segment_type_Complementary	4.50276
25	market_segment_type_Corporate	16.92829
26	market_segment_type_Offline	64.11564
27	market_segment_type_Online	71.18026

Checking Logistic Regression Assumption – 4. Removing multicollinearity

- 3 with high VIF – remove market_segment_type_Corporate

```
Series before feature selection:

const                                 39427863.33934
no_of_adults                                1.34710
no_of_children                              2.09347
no_of_weekend_nights                        1.06743
no_of_week_nights                           1.09387
required_car_parking_space                  1.03997
lead_time                                   1.39443
arrival_year                                1.42968
arrival_month                               1.27585
arrival_date                                1.00678
repeated_guest                              1.77844
no_of_previous_cancellations                1.39515
no_of_previous_bookings_not_canceled        1.64964
avg_price_per_room                          2.06807
no_of_special_requests                      1.24772
type_of_meal_plan_Meal Plan 2               1.27308
type_of_meal_plan_Meal Plan 3               1.02526
type_of_meal_plan_Not Selected              1.27260
room_type_reserved_Room_Type 2              1.10591
room_type_reserved_Room_Type 3              1.00330
room_type_reserved_Room_Type 4              1.35633
room_type_reserved_Room_Type 5              1.02736
room_type_reserved_Room_Type 6              2.05605
room_type_reserved_Room_Type 7              1.11815
market_segment_type_Complementary           1.37923
market_segment_type_Offline                 5.78960
market_segment_type_Online                  6.42810
dtype: float64
```

Checking Logistic Regression Assumption – 4. Removing multicollinearity

- 3 with high VIF – remove market_segment_type_Offline. After removing, another high VIF, market_segment_type_Online has dropped becomes low VIF value, therefore this factor will not be removed and proceed with another model building

```
Series before feature selection:

const                                 39354327.92909
no_of_adults                                1.30235
no_of_children                              2.09223
no_of_weekend_nights                        1.06588
no_of_week_nights                           1.09270
required_car_parking_space                  1.03605
lead_time                                   1.35291
arrival_year                                1.42709
arrival_month                               1.27539
arrival_date                                1.00585
repeated_guest                              1.61102
no_of_previous_cancellations                1.38180
no_of_previous_bookings_not_canceled        1.64082
avg_price_per_room                          2.06799
no_of_special_requests                      1.24751
type_of_meal_plan_Meal Plan 2               1.26051
type_of_meal_plan_Meal Plan 3               1.02526
type_of_meal_plan_Not Selected              1.27082
room_type_reserved_Room_Type 2              1.10581
room_type_reserved_Room_Type 3              1.00330
room_type_reserved_Room_Type 4              1.35408
room_type_reserved_Room_Type 5              1.02368
room_type_reserved_Room_Type 6              2.05544
room_type_reserved_Room_Type 7              1.11798
market_segment_type_Complementary           1.26980
market_segment_type_Online                  1.66252
dtype: float64
```

Data Preprocessing – 5. Re-check Modeling and its performance

- Model result

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25366
Method:	MLE	Df Model:	25
Date:	Wed, 28 Jun 2023	Pseudo R-squ.:	0.3252
Time:	13:02:16	Log-Likelihood:	-10858.
converged:	False	LL-Null:	-16091.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-981.0147	120.357	-8.151	0.000	-1216.910	-745.119
no_of_adults	0.0527	0.037	1.419	0.156	-0.020	0.125
no_of_children	0.1421	0.062	2.304	0.021	0.021	0.263
no_of_weekend_nights	0.1035	0.020	5.260	0.000	0.065	0.142
no_of_week_nights	0.0426	0.012	3.478	0.001	0.019	0.067
required_car_parking_space	-1.5523	0.137	-11.305	0.000	-1.821	-1.283
lead_time	0.0151	0.000	58.608	0.000	0.015	0.016
arrival_year	0.4840	0.060	8.114	0.000	0.367	0.601
arrival_month	-0.0404	0.006	-6.274	0.000	-0.053	-0.028
arrival_date	0.0013	0.002	0.679	0.497	-0.002	0.005
repeated_guest	-1.9114	0.618	-3.093	0.002	-3.123	-0.700
no_of_previous_cancellations	0.2156	0.086	2.497	0.013	0.046	0.385
no_of_previous_bookings_not_canceled	-0.1450	0.140	-1.039	0.299	-0.419	0.129
avg_price_per_room	0.0186	0.001	25.291	0.000	0.017	0.020
no_of_special_requests	-1.4509	0.030	-48.699	0.000	-1.509	-1.393
type_of_meal_plan_Meal Plan 2	0.1149	0.066	1.752	0.080	-0.014	0.243
type_of_meal_plan_Meal Plan 3	18.5374	7576.711	0.002	0.998	-1.48e+04	1.49e+04
type_of_meal_plan_Not Selected	0.2668	0.053	5.065	0.000	0.164	0.370
room_type_reserved_Room_Type 2	-0.3533	0.130	-2.715	0.007	-0.608	-0.098
room_type_reserved_Room_Type 3	-0.0364	1.283	-0.028	0.977	-2.552	2.479
room_type_reserved_Room_Type 4	-0.2440	0.053	-4.609	0.000	-0.348	-0.140
room_type_reserved_Room_Type 5	-0.5666	0.205	-2.767	0.006	-0.968	-0.165
room_type_reserved_Room_Type 6	-0.9137	0.150	-6.074	0.000	-1.209	-0.619
room_type_reserved_Room_Type 7	-1.3597	0.293	-4.642	0.000	-1.934	-0.786

- Training Performance

	Accuracy	Recall	Precision	F1
0	0.80498	0.63028	0.73917	0.68039

- All data is comparable and not affected after removing high VIF value.
- Next, we will proceed to remove high P-value.

Data Preprocessing – 6. Dropping high P-value

- High P-value after dropping, as below:
  - ‘const’, ‘no_of_adults’, ‘no_of_children’,’no_of_weekend_nights’,’no_of_week_nights’, ‘required_car_parking_space’, ‘lead_time’, ‘arrival_year’, ‘arrival_month’, ‘repeated_guest’, ‘no_of_previous_cancellations’, ‘avg_price_per_room’, ‘no_of_special_requests’, ‘type_of_meal_plan_Meal Plan 2’, 'type_of_meal_plan_Not Selected', 'room_type_reserved_Room_Type 2', '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_Corporate', 'market_segment_type_Offline'

Data Preprocessing – 7. Re-check Modeling and its performance

- Model result

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25370
Method:	MLE	Df Model:	21
Date:	Wed, 28 Jun 2023	Pseudo R-squ.:	0.3282
Time:	13:02:26	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

- Training Performance

	Accuracy	Recall	Precision	F1
0	0.80545	0.63267	0.73907	0.68174

- Test Performance

	Accuracy	Recall	Precision	F1
0	0.80594	0.62834	0.73375	0.67697

- All data is comparable and not affected after removing high P-value.
- The performance for both training and test is comparable, and it seems the data is not overfitting.

Data Preprocessing – 8. Converting coefficients to odds

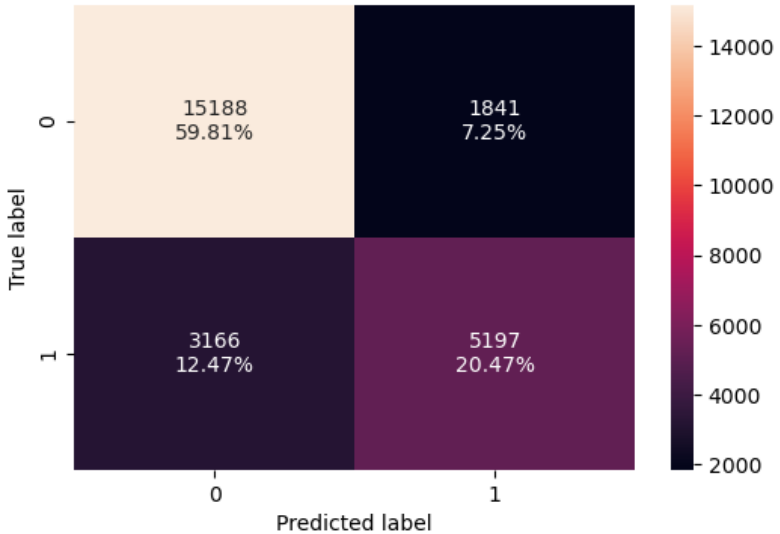
- After conversion

	const	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	required_car_parking_space	lead_time	arrival_year	arrival_month	repeated_guest	no_of_previous_cancellations
Odds	0.00000	0.99302	1.08217	1.23292	1.03744	0.21588	1.01584	1.48900	0.95653	0.11991	1.20525
Change_odd%	-100.00000	-0.69770	8.21664	23.29156	3.74360	-78.41195	1.58373	48.90013	-4.34714	-88.00906	20.52512

	avg_price_per_room	no_of_special_requests	type_of_meal_plan_Meal Plan 2	type_of_meal_plan_Not Selected	room_type_reserved_Room_Type 2	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_Corporate	market_segment_type_Offline
Odds	1.01883	0.22108	1.13020	1.17693	0.60571	0.80656	0.40807	0.40762	0.17829	0.37098	0.17860
Change_odd%	1.88297	-77.89242	13.01959	17.69350	-39.42881	-19.34361	-59.19278	-59.23830	-82.17075	-62.90248	-82.14031

Data Preprocessing – 9. Checking performance on the training set

- Confusion matrix

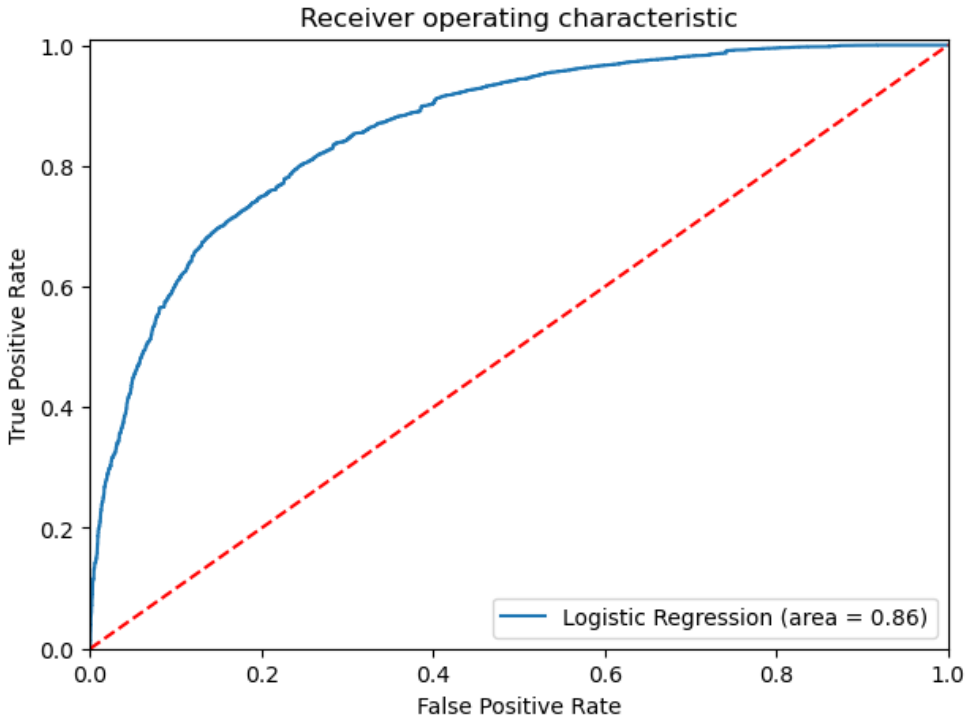


- Training Performance

	Accuracy	Recall	Precision	F1
0	0.80281	0.62143	0.73842	0.67489

- All data is still comparable from previous performance after conversion.

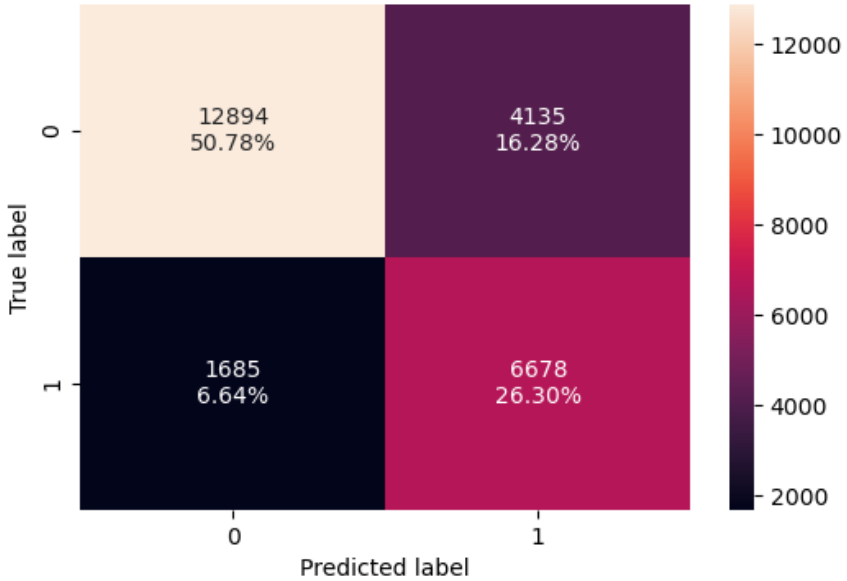
Model Performance Evaluation and Improvement – 1. Checking ROC-AUC on the training set



- Based on ROC-AUC curve, the model is performing well on the training set.

Model Performance Evaluation and Improvement – 2. Optimize threshold using ROC-AUC curve

- Optimal threshold is on value 0.2993111390299003
- Therefore, the confusion matrix as below:

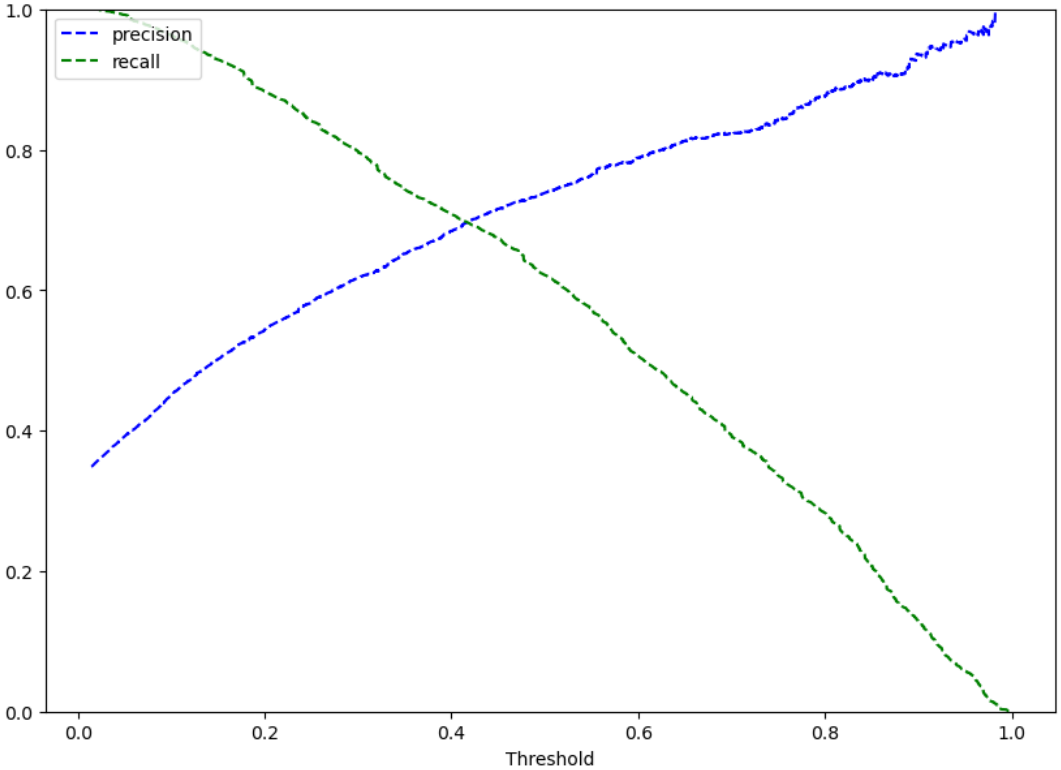


- Training Performance

	Accuracy	Recall	Precision	F1
0	0.77079	0.79852	0.61759	0.69650

- After we set the optimal threshold value, we observed the training performance slightly decrease from previous performance.
- However, we will check the precision-recall curve first before performing any assumption/summary.

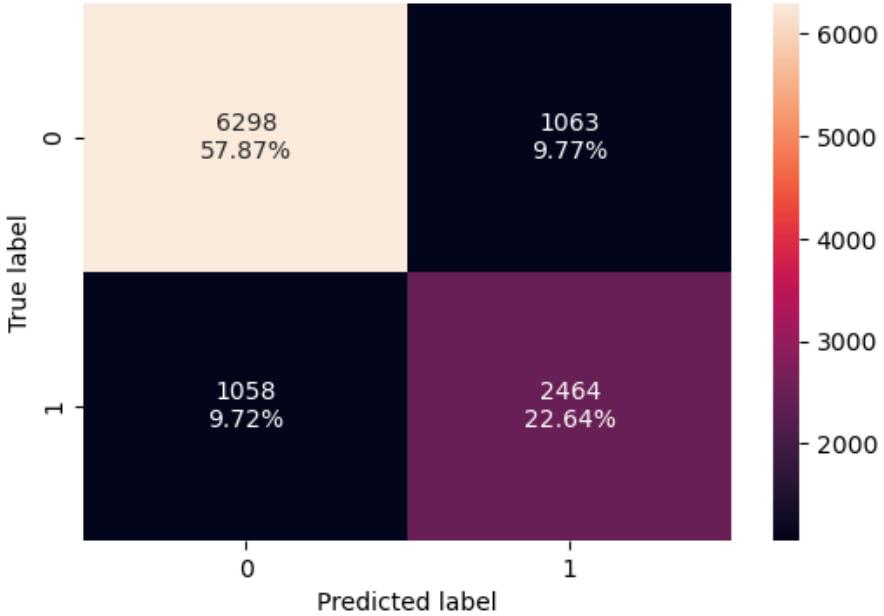
Model Performance Evaluation and Improvement – 2. Checking the Precision-Recall Curve



- A threshold around 0.42, we may get a balanced precision and recall.

Model Performance Evaluation and Improvement – 3. Checking model performance on the test set

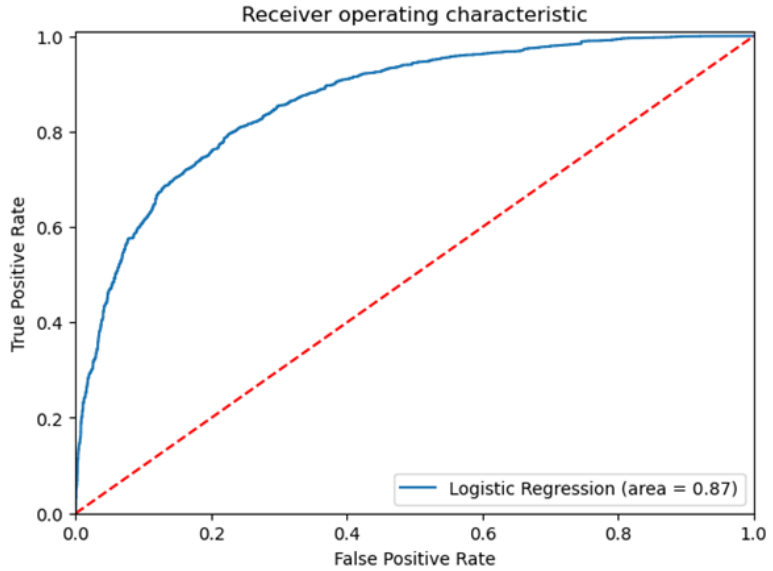
- First optimal threshold was set at default.
- The confusion matrix as below:



- Training Performance

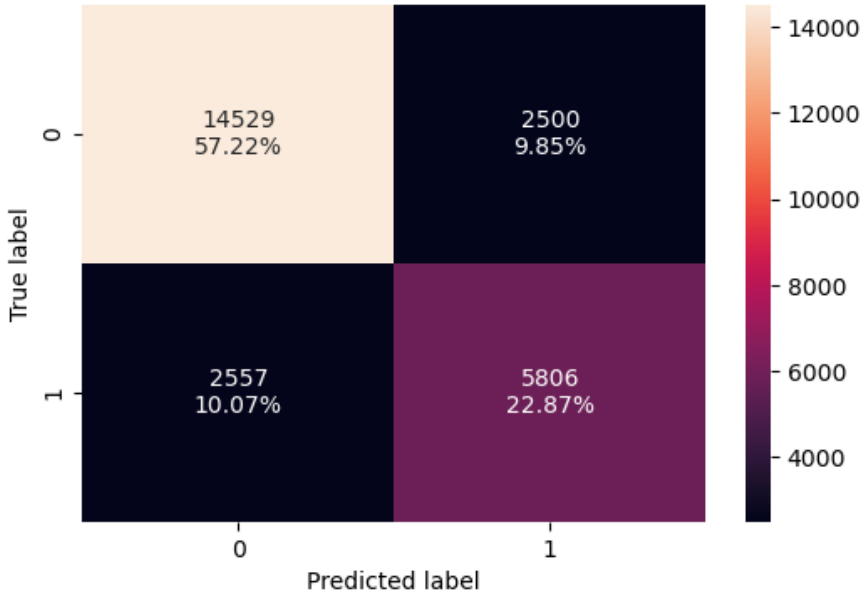
	Accuracy	Recall	Precision	F1
0	0.80511	0.69960	0.69861	0.69911

- ROC Curve



Model Performance Evaluation and Improvement – 4. Checking model performance on the training set

- Second optimal threshold was set at: 0.42
- The confusion matrix as below:

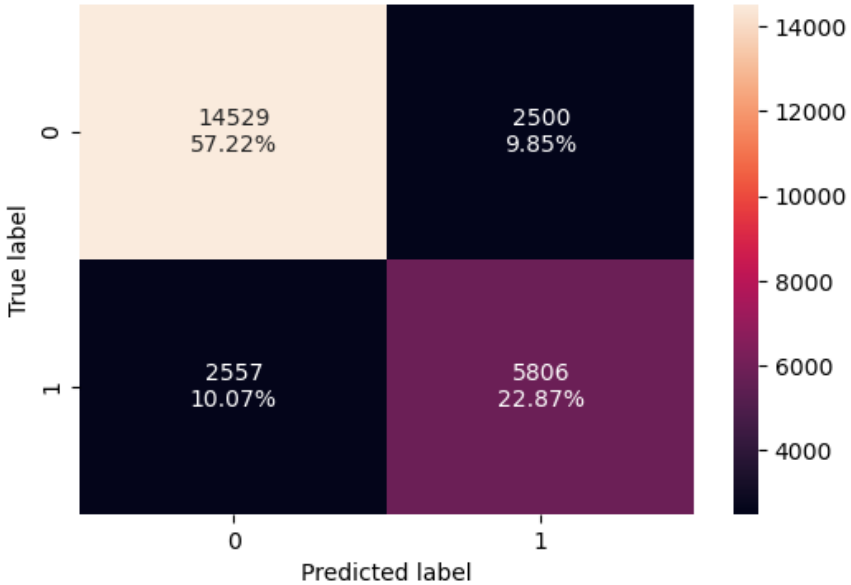


- Training Performance

	Accuracy	Recall	Precision	F1
0	0.80084	0.69425	0.69901	0.69662

Model Performance Evaluation and Improvement – 5. Checking model performance on the test set

- Testing the test set with optimal threshold was set at: 0.42
- The confusion matrix as below:

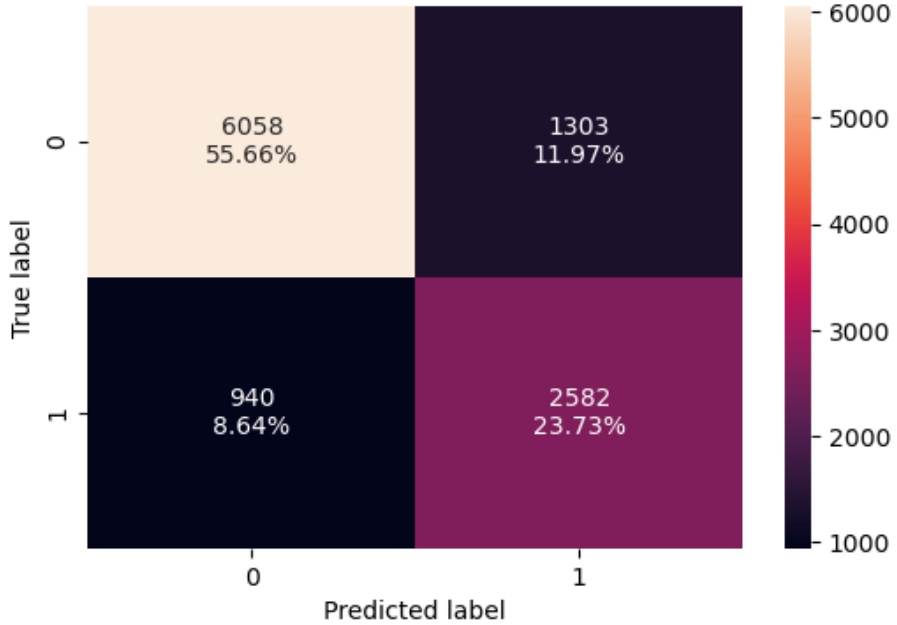


- Test Performance

	Accuracy	Recall	Precision	F1
0	0.80084	0.69425	0.69901	0.69662

Model Performance Evaluation and Improvement – 6. Checking model performance on the test set

- Testing the test set with optimal threshold was set at: 0.37
- The confusion matrix as below:



- Test Performance

	Accuracy	Recall	Precision	F1
0	0.79390	0.73311	0.66461	0.69718

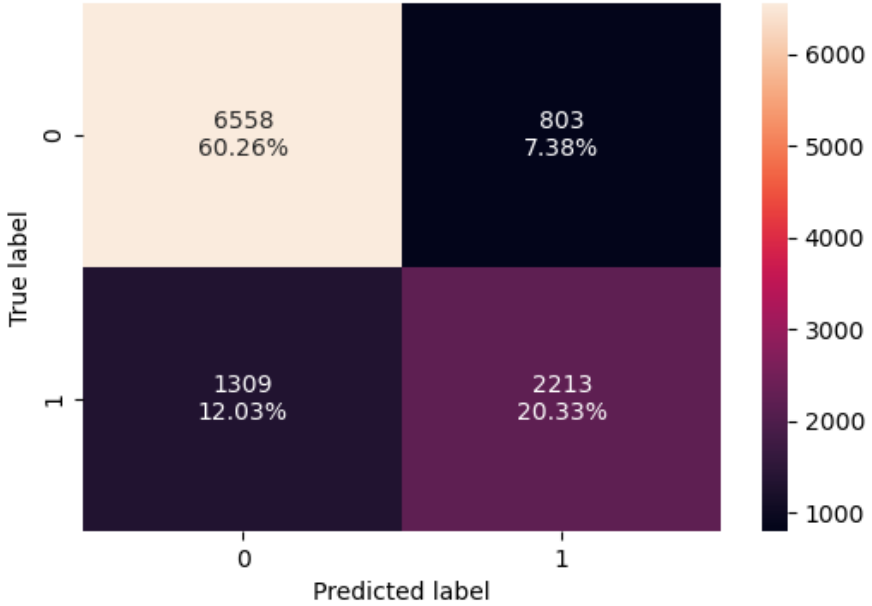
Model Performance Summary – 7. Training Performance Comparison

	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80281	0.77079	0.80084
Recall	0.62143	0.79852	0.69425
Precision	0.73842	0.61759	0.69901
F1	0.67489	0.69650	0.69662

- All three thresholds are giving the reliable and good performance without overfitting.
- However, among the three performance, the 0.42 threshold is giving the best value of F1. Therefore, it will be selected as the final model.
- All three thresholds will be tested on the test data.

Model Performance Summary – 8. Checking the Test Performance

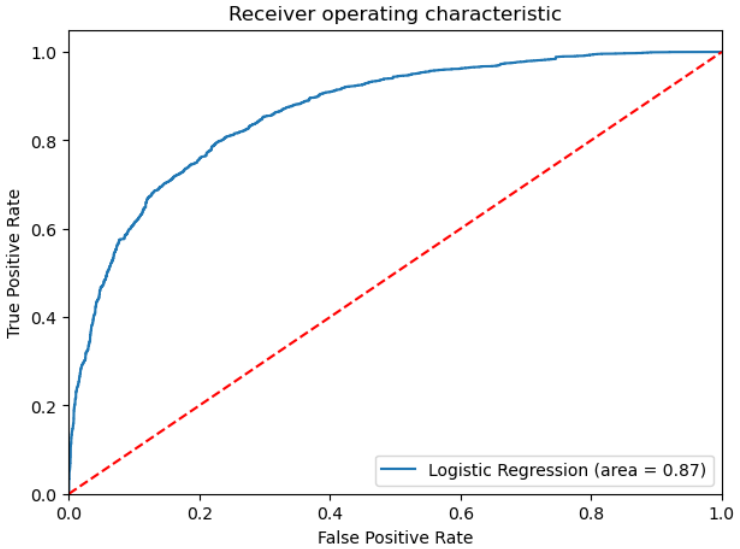
- First optimal threshold was set at default.
- The confusion matrix as below:



- Test Performance

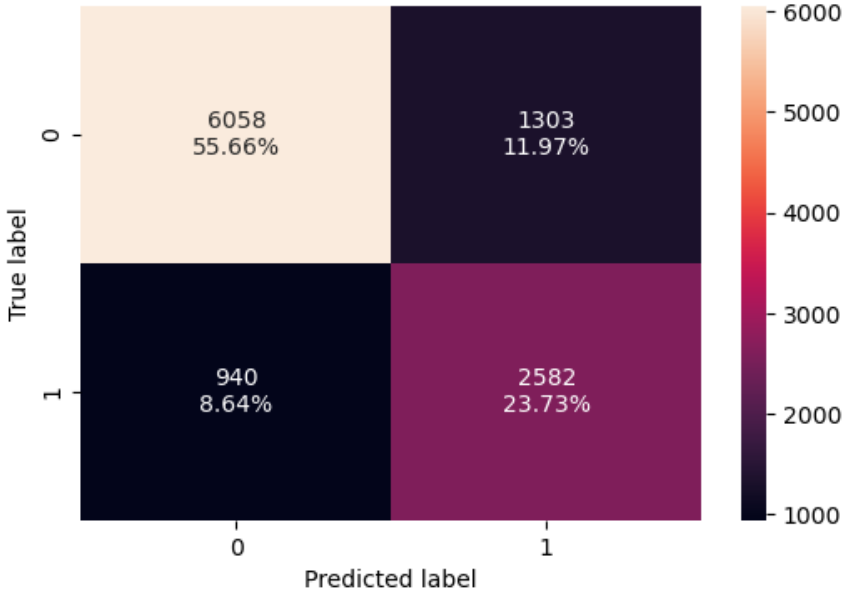
	Accuracy	Recall	Precision	F1
0	0.80594	0.62834	0.73375	0.67697

- ROC Curve



Model Performance Summary – 9. Checking the Test Performance

- Testing with optimal threshold was set at: 0.37
- The confusion matrix as below:

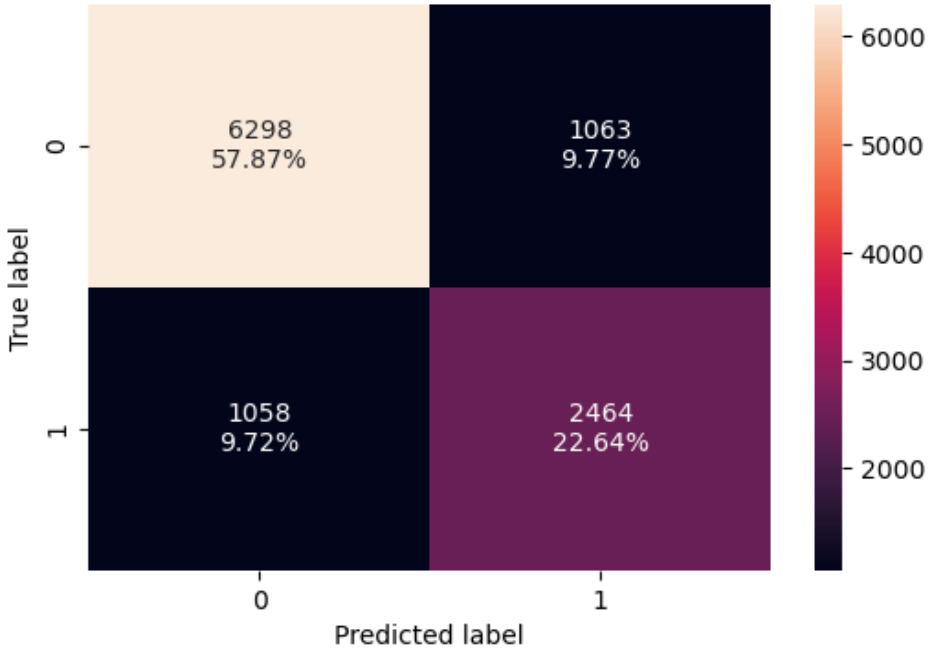


- Test Performance

	Accuracy	Recall	Precision	F1
0	0.79390	0.73311	0.66461	0.69718

Model Performance Summary – 10. Checking the Test Performance

- Testing with optimal threshold was set at: 0.42
- The confusion matrix as below:



- Test Performance

	Accuracy	Recall	Precision	F1
0	0.80511	0.69960	0.69861	0.69911

Model Performance Summary – 7. Test Performance Comparison

	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80594	0.79390	0.80511
Recall	0.62834	0.73311	0.69960
Precision	0.73375	0.66461	0.69861
F1	0.67697	0.69718	0.69911

- All three thresholds are giving the reliable and good performance without overfitting on a test data.
- As predicted on the training performance data, threshold 0.42 is giving the best F1 result. It also can be seen above on the test data by giving the best F1 value compared on default and 0.37 threshold.
- Therefore, we will take threshold 0.42 as the best threshold.

Data Preparation for Decision Tree – 1. Encode Categorical Value

	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
0	2	0	1	2	0	224	2017	10	2	0	0	0	65.00000
1	2	0	2	3	0	5	2018	11	6	0	0	0	106.68000
2	1	0	2	1	0	1	2018	2	28	0	0	0	60.00000
3	2	0	0	2	0	211	2018	5	20	0	0	0	100.00000
4	2	0	1	1	0	48	2018	4	11	0	0	0	94.50000

no_of_special_requests	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
0	0	0	0	0	0	0	0	0	0	0	0	1	0
1	0	0	1	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	1	0	0	0	0	0	0	0	0	0	1

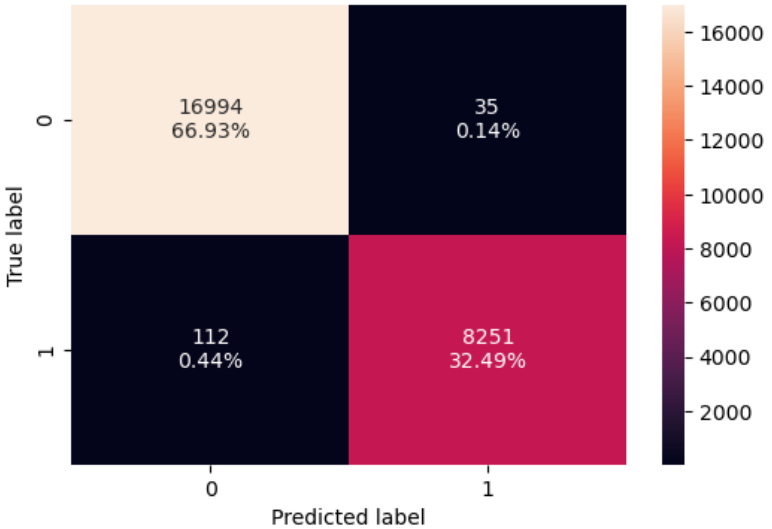
Data Preparation for Decision Tree – 2. Split data

- Split data into 70:30

```
Shape of Training set :  (25392, 27)
Shape of test set :  (10883, 27)
Percentage of classes in training set:
0   0.67064
1   0.32936
Name: booking_status, dtype: float64
Percentage of classes in test set:
0   0.67638
1   0.32362
Name: booking_status, dtype: float64
```

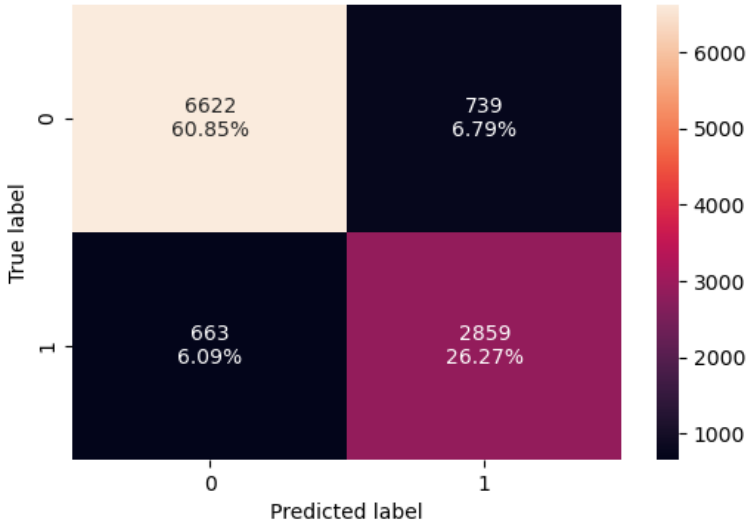
Data Preparation for Decision Tree – 3. Building Decision Tree and Model Performance Check

- Training Performance



	Accuracy	Recall	Precision	F1
0	0.99421	0.98661	0.99578	0.99117

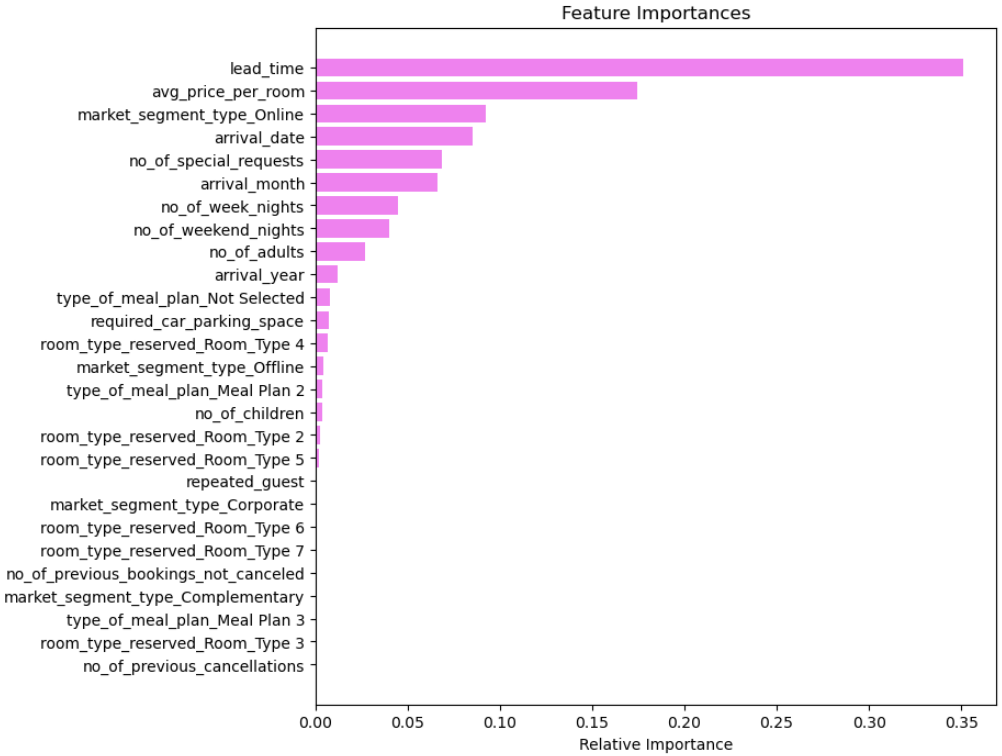
- Test Performance



	Accuracy	Recall	Precision	F1
0	0.87118	0.81175	0.79461	0.80309

- Training and test performance is far different from each other, an indication of overfitting.
- Pre-pruning is necessary.

Data Preparation for Decision Tree – 4. Check Important Features



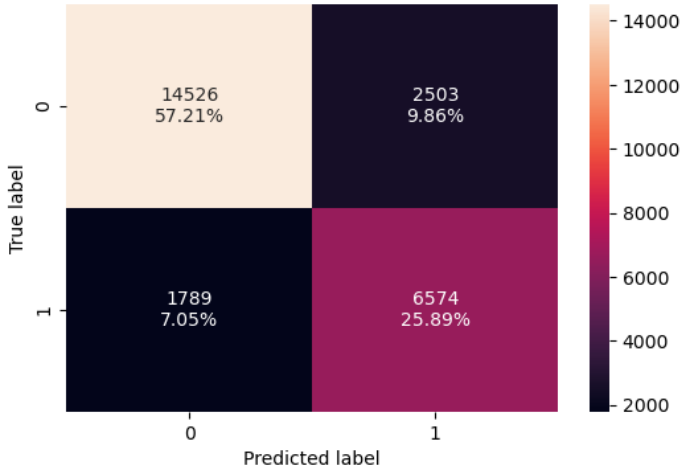
- Among importance feature is lead_time, avg_price_per_room and market_segment_type_online (top 3)
- The parameter need to observe again after pre-pruning.

Data Preparation for Decision Tree – 5. Pre-pruning

- Pre-pruning

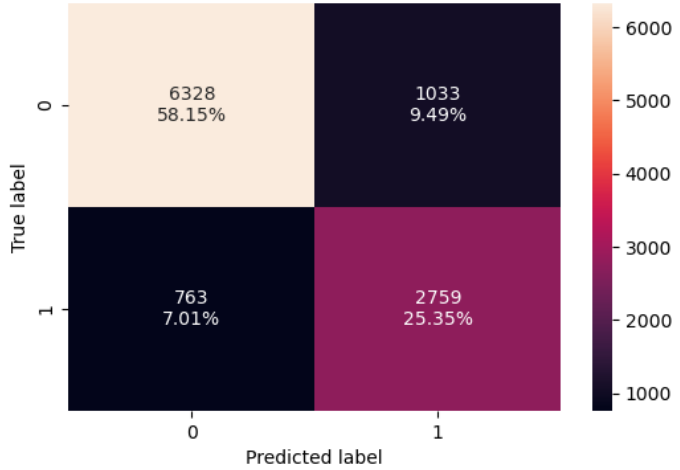
```
DecisionTreeClassifier(class_weight='balanced', max_depth=6, max_leaf_nodes=50,
                       min_samples_split=10, random_state=1)
```

- Training Performance



	Accuracy	Recall	Precision	F1
0	0.83097	0.78608	0.72425	0.75390

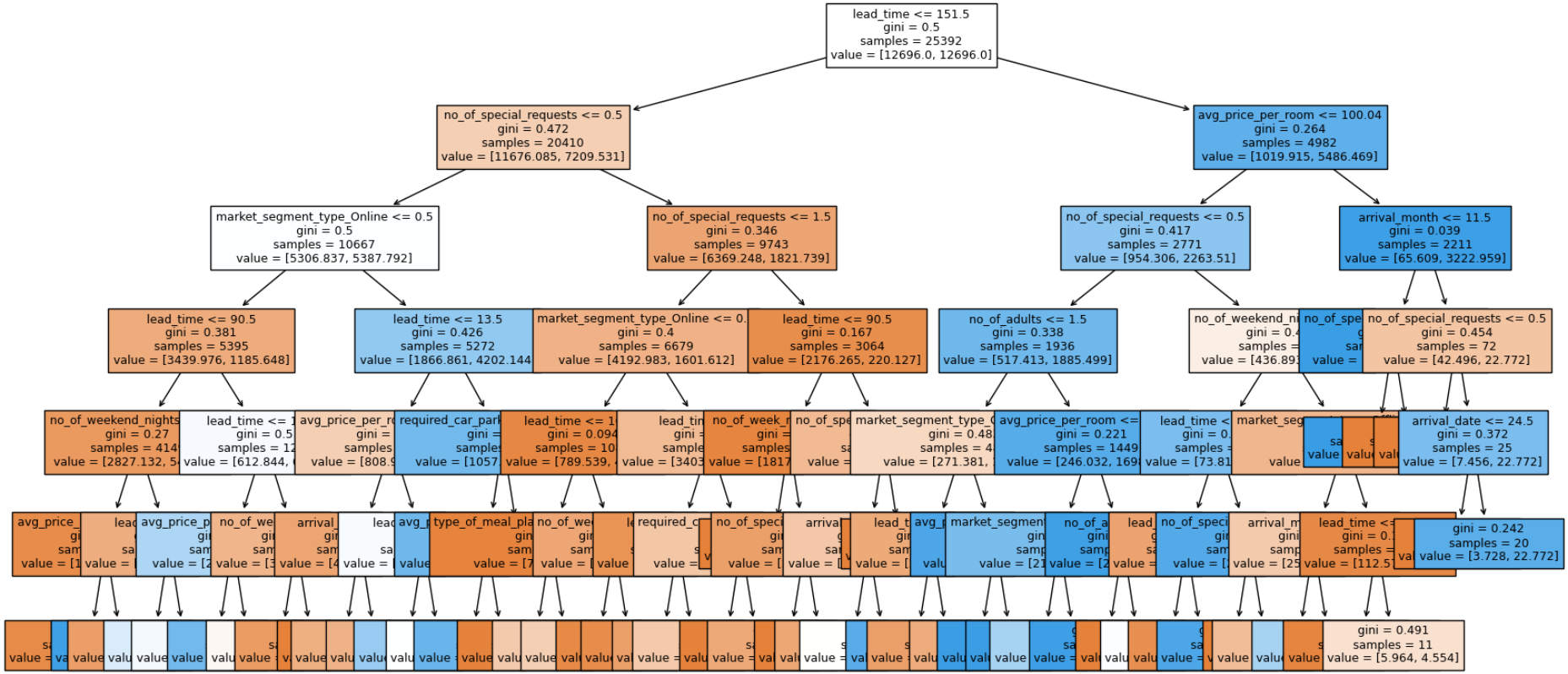
- Test Performance



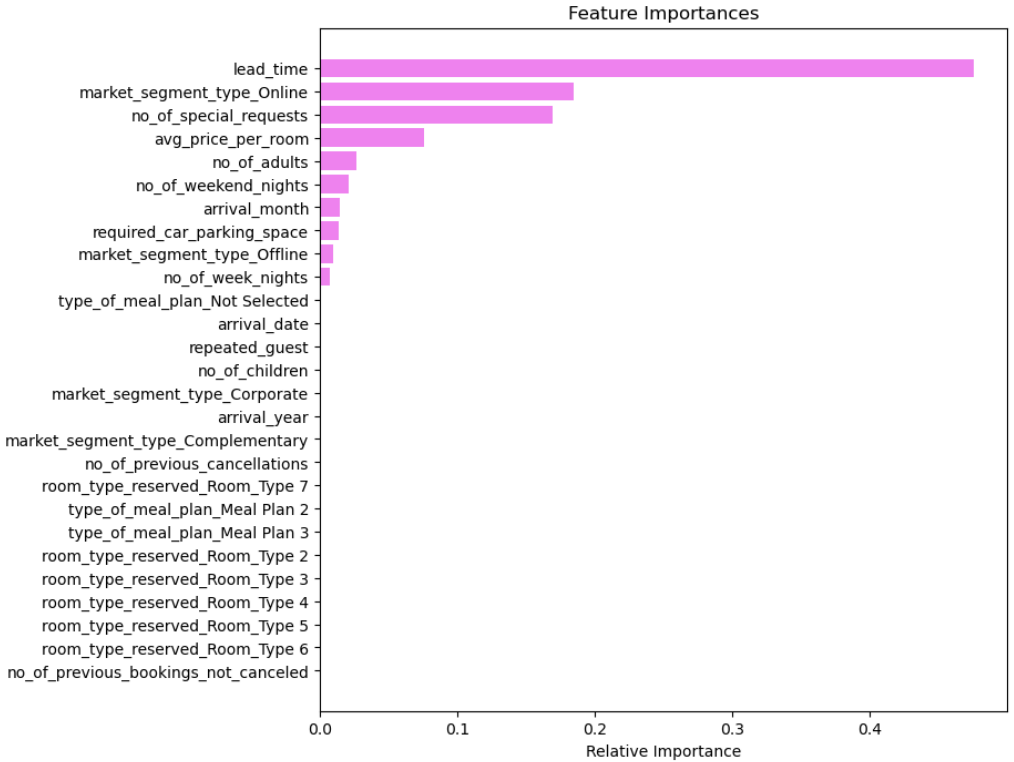
	Accuracy	Recall	Precision	F1
0	0.83497	0.78336	0.72758	0.75444

- The performance for both training and test is now comparable after pre-pruning.

Data Preparation for Decision Tree – 6. Visualising the Decision Tree



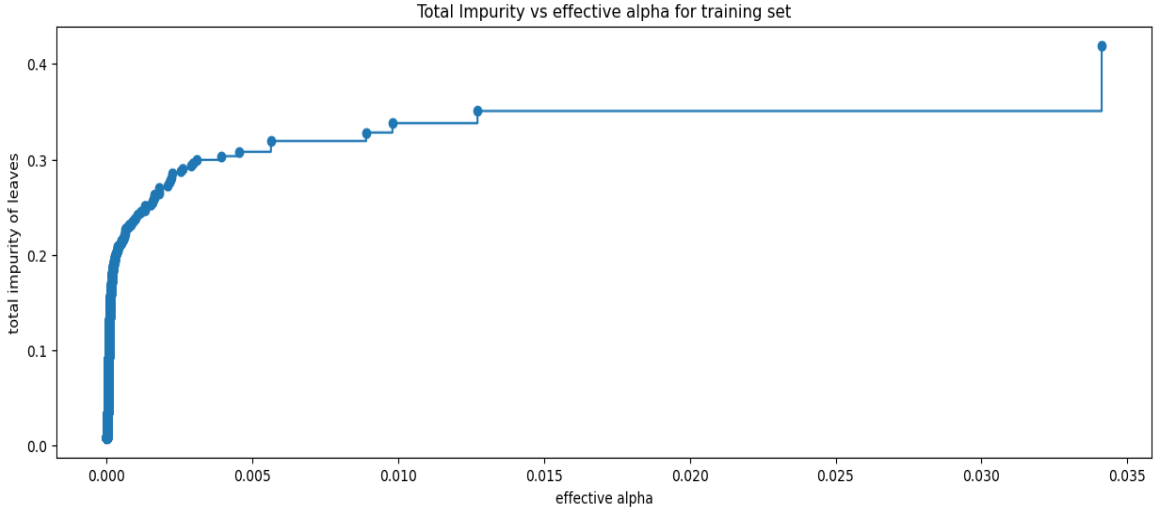
Data Preparation for Decision Tree – 7. Re-check Important Features



- After pre-pruning, some important features has changed.
- Now the top three after pre-pruning is lead_time, market_segment_type_online and no_of_special_request.

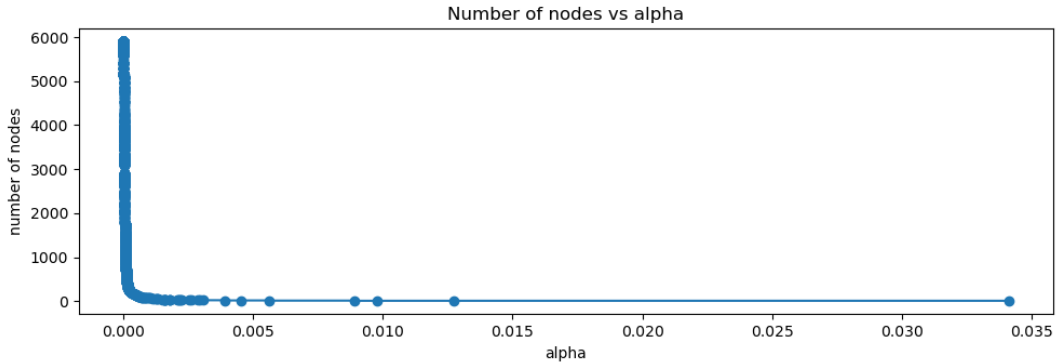
Data Preparation for Decision Tree – 8. Cost Complexity Pruning

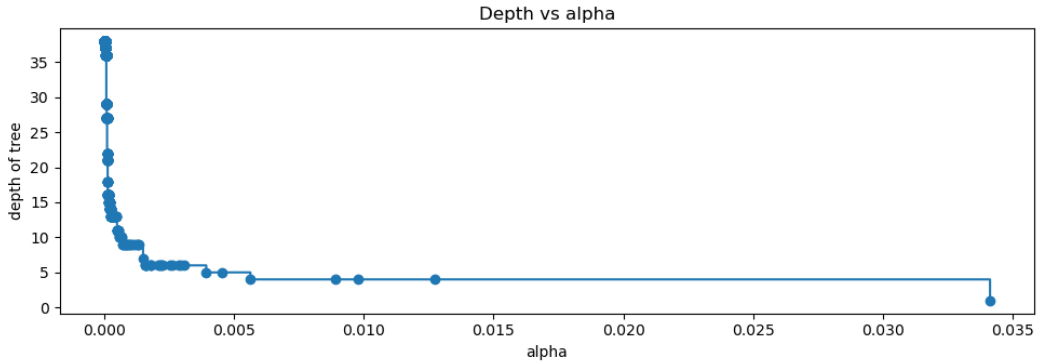
	ccp_alphas	impurities
0	0.00000	0.00838
1	0.00000	0.00838
2	0.00000	0.00838
3	0.00000	0.00838
4	0.00000	0.00838
...	...	...
1839	0.00890	0.32806
1840	0.00980	0.33786
1841	0.01272	0.35058
1842	0.03412	0.41882
1843	0.08118	0.50000



Data Preparation for Decision Tree – 9. Train using effective alphas

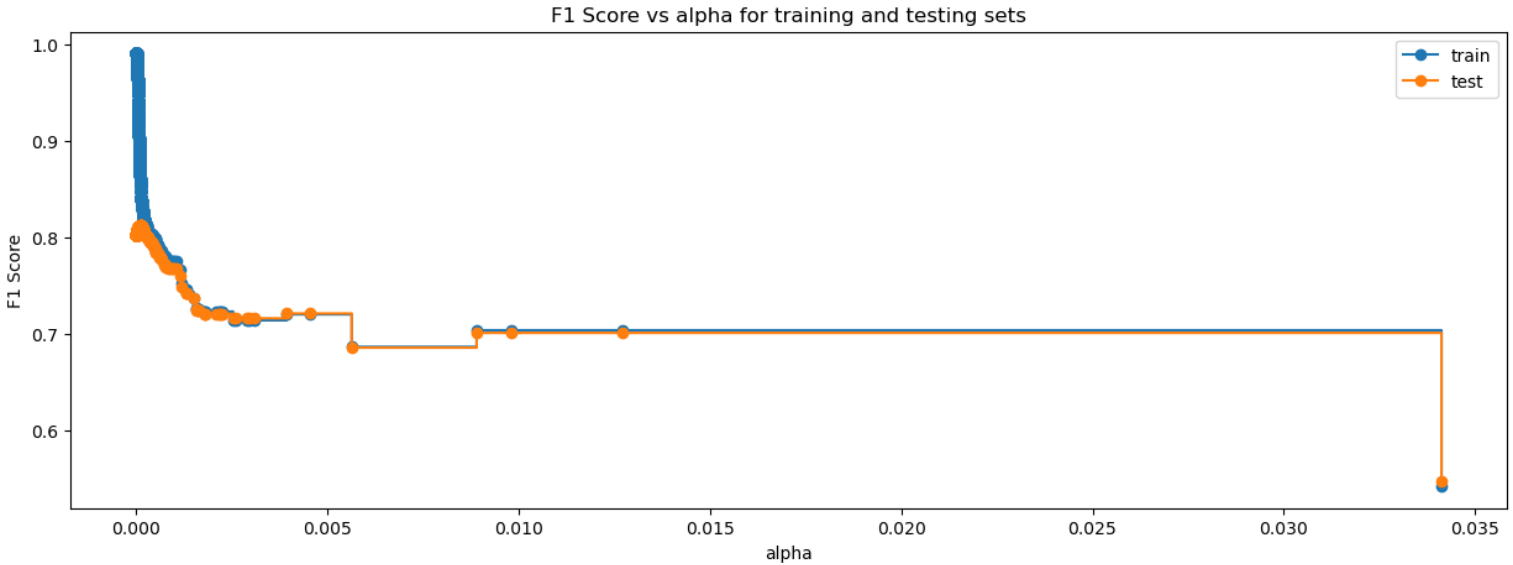
- Number of nodes in the last tree is: 1 with ccp_alpha: 0.0811791438913696





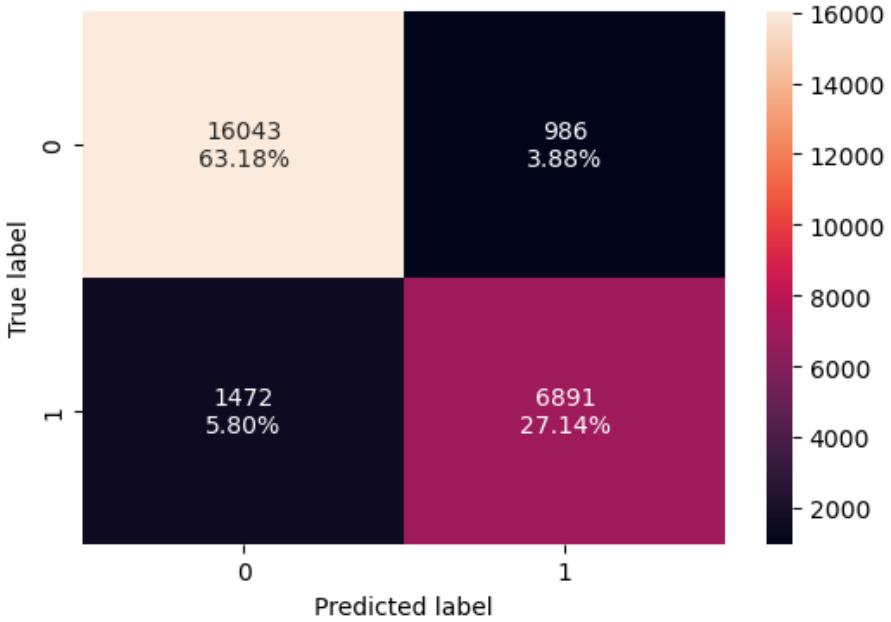
Data Preparation for Decision Tree – 10. F1 Score for alpha

- Best model: decision tree classifier (ccp_alpha=0.00011676062470523539, random_state=1)
- Both value is comparable.



Data Preparation for Decision Tree – 11. Model Performance Check

- Training Performance



	Accuracy	Recall	Precision	F1
0	0.90320	0.82399	0.87483	0.84865

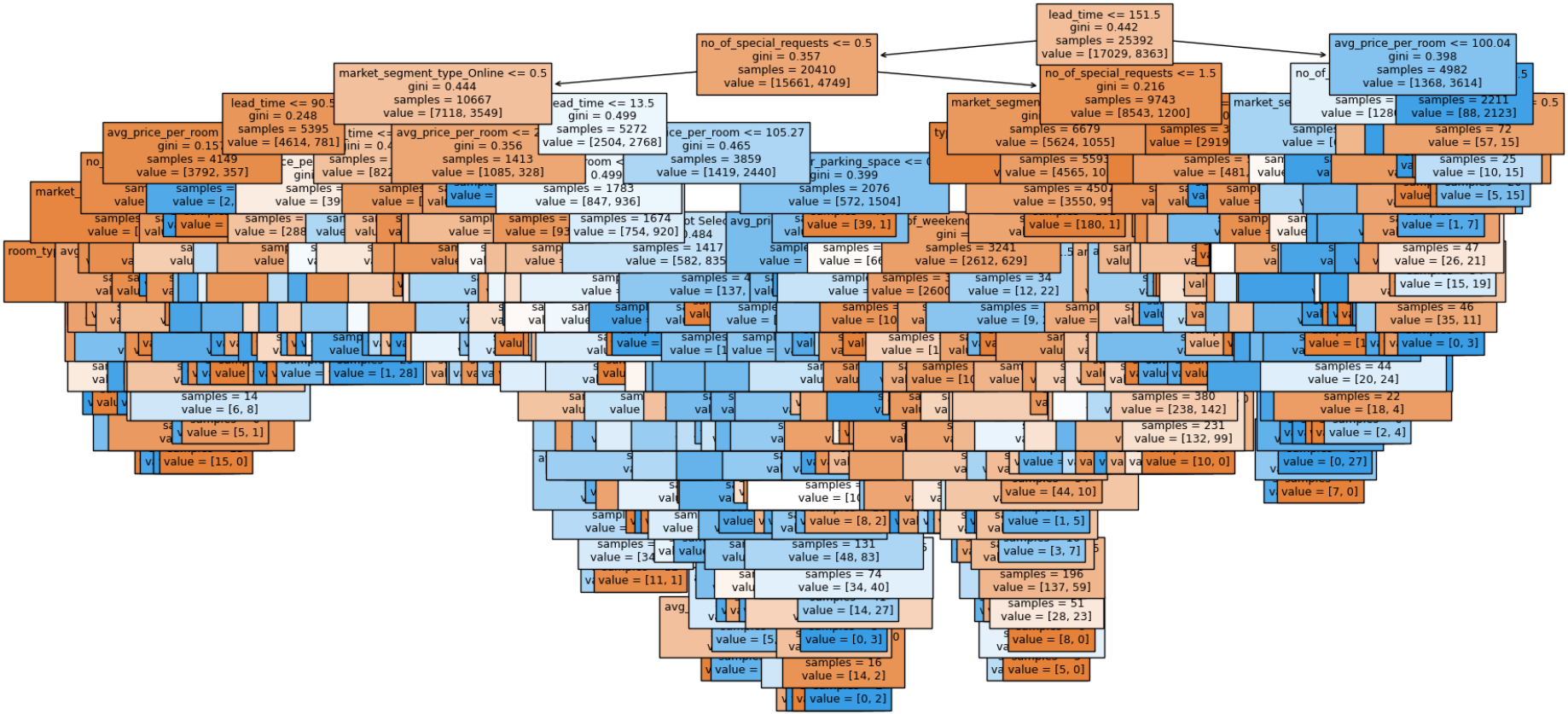
- Test Performance

	Accuracy	Recall	Precision	F1
0	0.88275	0.79046	0.83805	0.81356

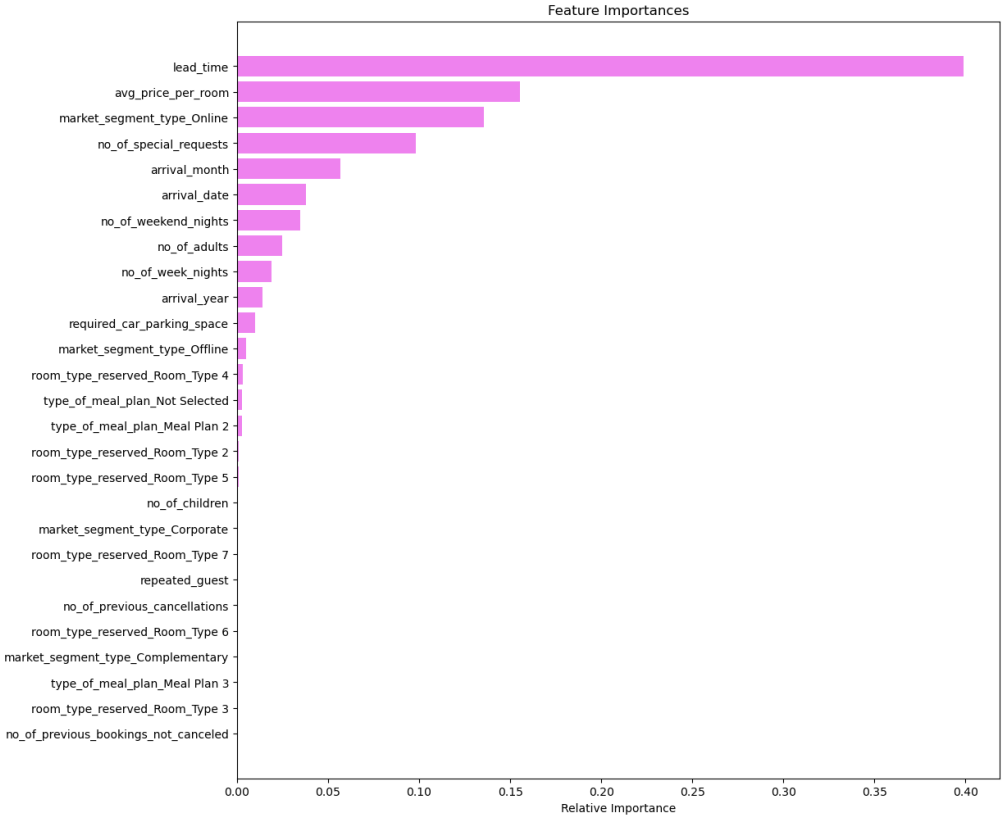
- Post Test Performance

	Accuracy	Recall	Precision	F1
0	0.83497	0.78336	0.72758	0.75444

Model Building – 12. Decision Tree



Model Building – 14. Important Features



- Some important features change again post pruning
- Now the top three after pre-pruning is lead_time, avg_price_per_room and market_segment_type_online

Model Building – 15. Final Model Performance Check

- Training Performance

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.99421	0.83097	0.90320
Recall	0.98661	0.78608	0.82399
Precision	0.99578	0.72425	0.87483
F1	0.99117	0.75390	0.84865

- Test Performance

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.87118	0.83497	0.88275
Recall	0.81175	0.78336	0.79046
Precision	0.79461	0.72758	0.83805
F1	0.80309	0.75444	0.81356

Model Performance Evaluation and Improvement - Decision Tree

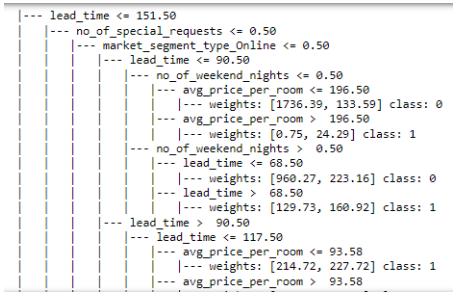
- Decision tree before pre-pruning giving the best training performance but the difference on the test performance is high, an indication of overfitting.
- Decision tree post pruning is giving us the best F1 score and both training and test performance is comparable, an indication of data is not overfitting.
- It is giving us the prediction of 79% of why the hotel bookings are being cancelled, based on the listed reasons on the importance features.
- We conclude that lead_time, avg_price_per_room, market_segment_type_online and no_of_special_request is the most top 4 importance features when we predict why the significant number of hotel bookings are being called off/cancelled.
- The company should focus to improve on those 4 features to prevent for another cancellations by customers in the near future.

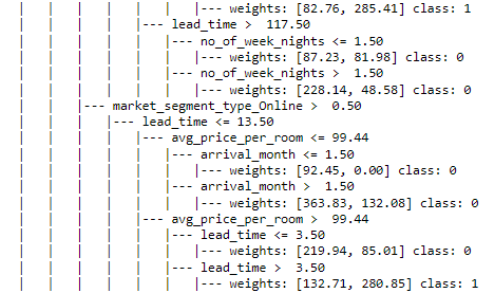
APPENDIX

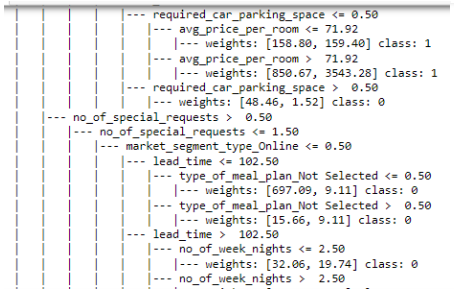
Data Background and Contents

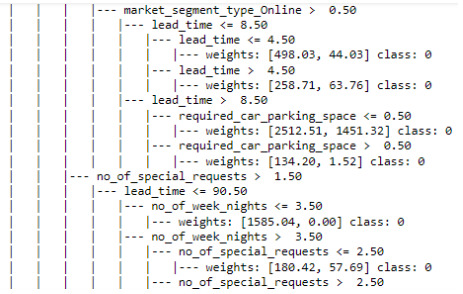
- Data.shape() – 36275 columns, 19 rows
- No duplicated data found.
- No missing data found.

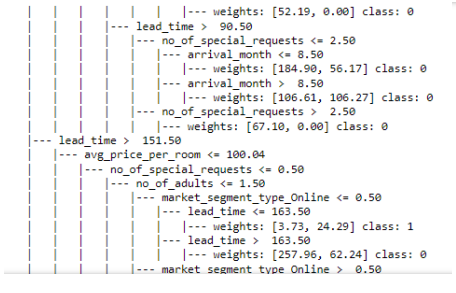
Data Preparation for Decision Tree – Pre-pruning

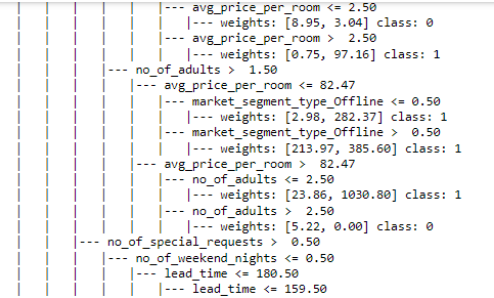


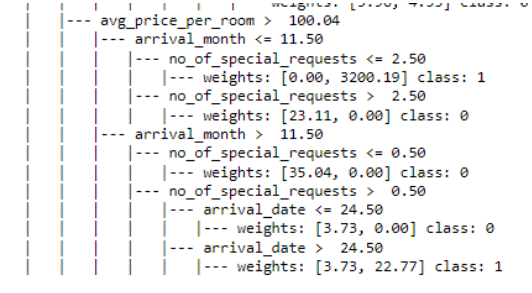




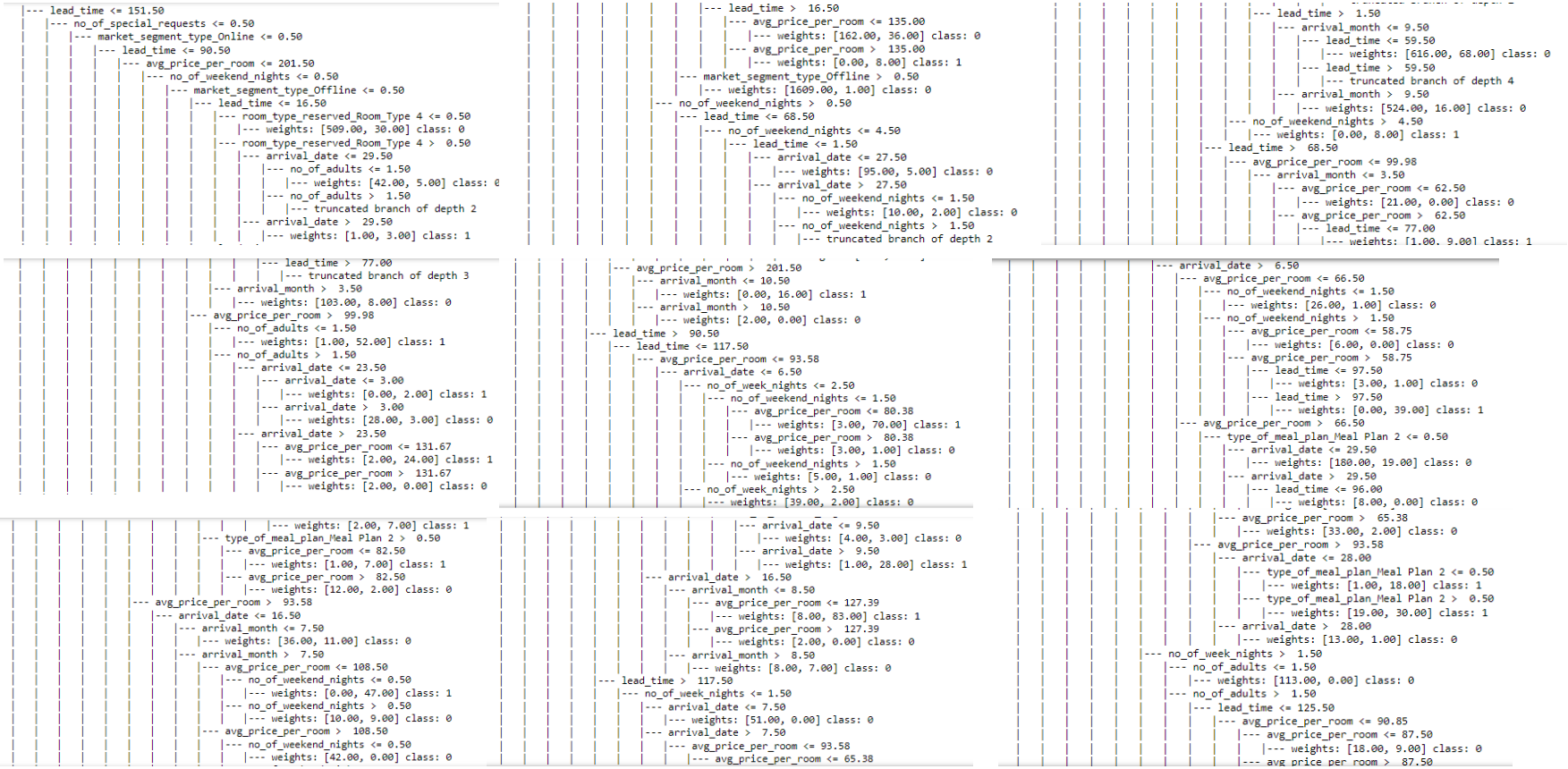








Model Building – Decision Tree



Great Learning

Happy Learning !