INN Hotels (Case Study)

Project 4: Classification

Date: December 10th, 2022.

By: Ijeoma Ejem

Contents / Agenda

EXECUTIVE SUMMARY

Executive Summary

	Conclusion	Recommendation
1	Returning guests had the least number of cancellations. Only 1.72% of returning guests canceled their reservation, while 34% of new guests canceled reservations in the same period.	There should be employees in charge of ensuring quality and unique experiences (customer experience manager) for returning guests. The hotel staff should know their names, what they like and have those things ready prior to their arrival. Reward programs should be very well designed.
2	Lead time appears to be the most important feature of all attributes. Exploratory data analysis showed that the average lead time is 85 days but most guest with lead times under 10 days retained their booking.	Advertising campaigns should cater more to demographics who frequently need last minute bookings, like; work travelers. This could mean providing services that reduce logistic planning for people in this segment like shuttle services.
3	While October appears to be the busiest month for INN Hotels, the cancellations around this month are still relatively higher than slower months. January has the least number of cancellations and reservations made in the year. July has the most cancellations.	Steps can be taken to reduce cancellations around busiest months; stringent penalties can be introduced so rooms that can be filled quickly aren’t held by frequent cancellers. More research should be done to understand traveler patterns in high cancellation months like July.

	Conclusion	Recommendation
4	Bookings with shorter total stays (below 10 days) generally tend to cancel reservations less frequently, with very few exceptions, than reservations made for long stays (10 and above).	People with short-stays, like; work, getaway and event travelers should be focused on. For instance, advertise to people traveling for local festivals or partner with local spas and grooming parlors to get guests the best deals.
5	Cancelled reservations tend to have higher median prices per day of the reservation than those that were not canceled.	Experiment with offers and price reduction tests to gauge how drastically price changes affect different demographics of guests across different seasons/months.
6	Coefficient of some levels of average price per room, no of previous cancellations, lead time, arrival year, no of guests per booking and no of nights booked are positive. An increase in these will lead to increase in chances of a guest canceling their reservation.	Pay attention to repeat cancellers and segment them. Enforce stricter penalties on these guests at different thresholds to prevent the pattern from continuing. Access yearly travel and hospitality trends to ensure INN Hotels is ahead of the curve. Socio-economic and political changes can and will influence consumer decisions.

	Conclusion	Recommendation
7	Coefficient of required car parking space, arrival month, repeated guest, no of special requests, market segment type, and room type are negative. An increase in these will lead to a decrease in chances of a guest canceling their reservation.	Consider more ways to add value to the guests experience by making every process as efficient and well designed as possible. Consider small things that will make big impacts for the guest. Convenience and added value should be designed into every experience.
8	There is enough evidence to suggest that market segments and no. of special requests are two of the most important feature next to lead time.	Create more segments by studying guests, even prior to arrival, and designing unique experiences that each segment will be most attracted to. This way your marketing techniques will be more diverse and more impactful at the same time. The occurrences of cancellations will also be reduced.
9	Our decision tree model can correctly predict 90% of booking cancellations and 85% of average recall and precision rates via f1 score. While, Our logistic regression model can correctly predict 73% of booking cancellations and 70% of average recall and precision rates via f1 score.	Use these models to gain a better understanding of customer behavior and enforce penalties when necessary to prevent total loss of revenue and resources.

PROBLEM AND SOLUTION OVERVIEW

Business Problem and Solution Overview

Problem:

- 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. made easier by the option to do so free of charge or preferably at a low cost.
- INN Hotels is impacted by cancellations through; loss of resources (revenue), additional costs of paying to advertise vacant rooms, last minute price reductions that reduce the profit margin, human resources.

Solution:

- Using exploratory data analysis (EDA) to find significant influences or patterns that affect booking cancellations.
- Building and testing a logistic regression model and decision tree to predict which booking is going to be canceled in advance, and help in formulating profitable policies for cancellations and refunds.
- Analyzing results from the data and providing business insights and recommendations to help take the company reduce its losses and maximize its gains.

DATA OVERVIEW

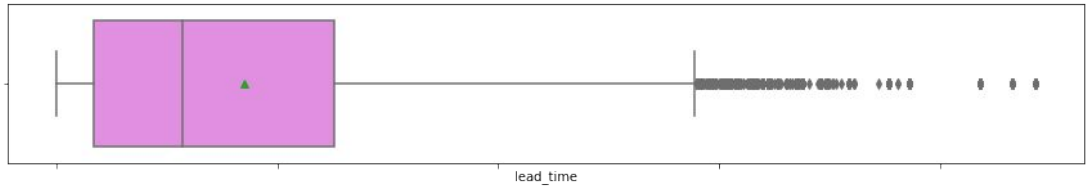
Data Overview

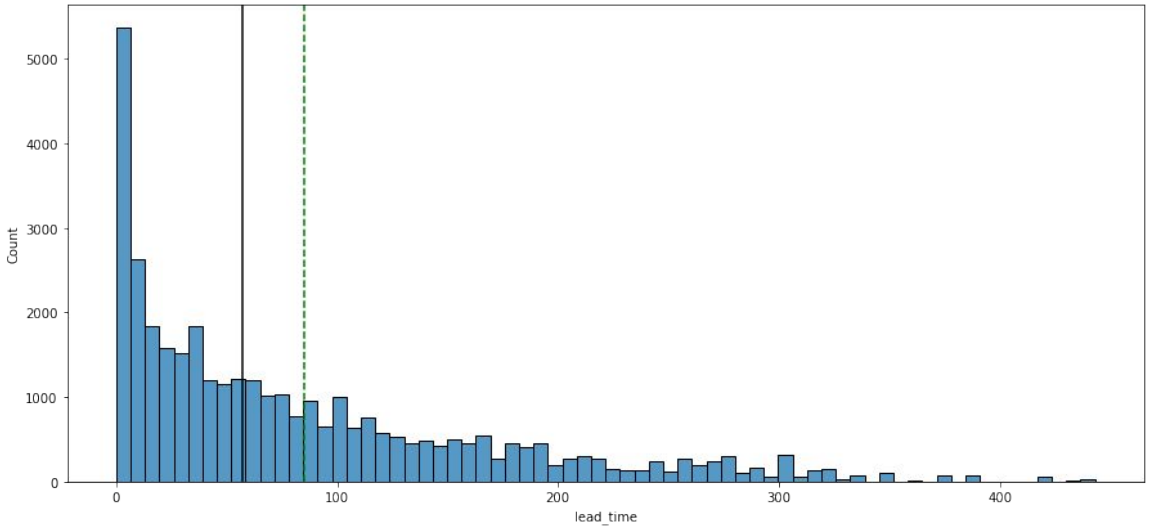
- There are 36,275 rows (observations) and 19 columns (attributes) in total.
- All datatypes are accurately represented; there are 5 objects, 13 integers and 1 float in the dataset. The attributes we are analyzing and building a model around have 5 categorical variables and 14 numerical variables.
- From the statistical summary of the combined numerical and categorical data, the following is deduced;
  - The average price of rooms per day is 103 euros.
  - The average arrival month is July.
  - The average number of days between the date of booking and the arrival date (lead time) is 85 days.
  - The maximum number of previous bookings not canceled by the customer prior to the current booking is 58.
  - The maximum number of previous bookings that were canceled by the customer prior to the current booking is 13.
- There are 0 missing values and duplicate values in the data.

EDA: UNIVARIATE ANALYSIS

EDA Results

UNIVARIATE ANALYSIS



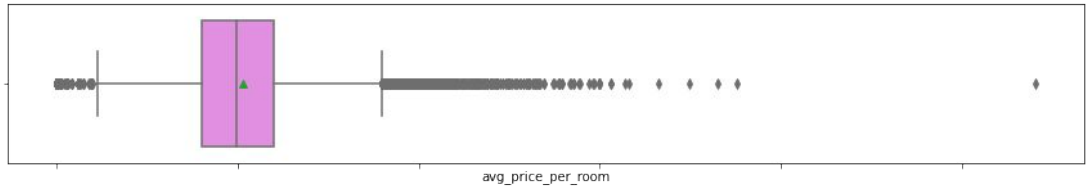


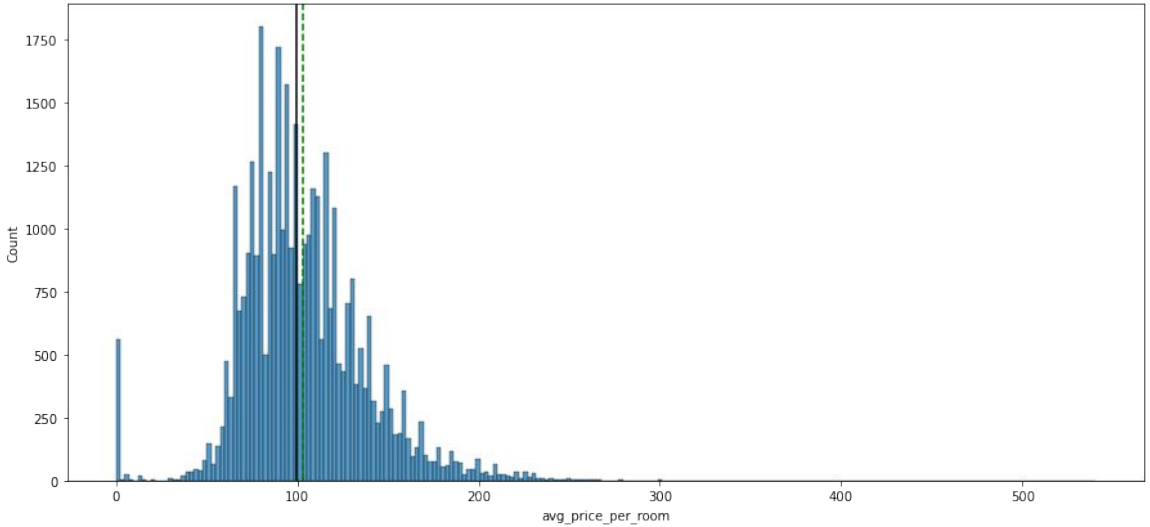
Lead Time

The median value for lead time is 57 days, while the mean value is 85 days. Lead time has numerous outliers with a maximum value of 443 days between the date of booking and the date of arrival. These plots show a right skewed distribution.

EDA Results

UNIVARIATE ANALYSIS





Average Price Per Room

The median value for average price rooms per day is 100 euros, while the mean value is 103 euros. Average price of rooms shows numerous outliers on both ends of the plot, with a minimum value of 0 euros and a maximum value of 540 euros per day. This distribution is relatively close to being normal or symmetrical.

EDA Results

UNIVARIATE ANALYSIS

Further Analysis of Average Price Per Room

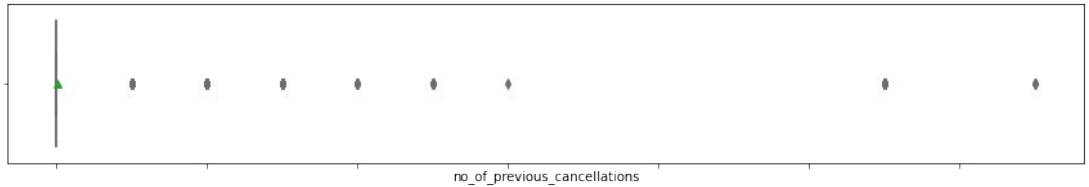
The number of bookings with room reservations averaging 0 euros per day is 545.

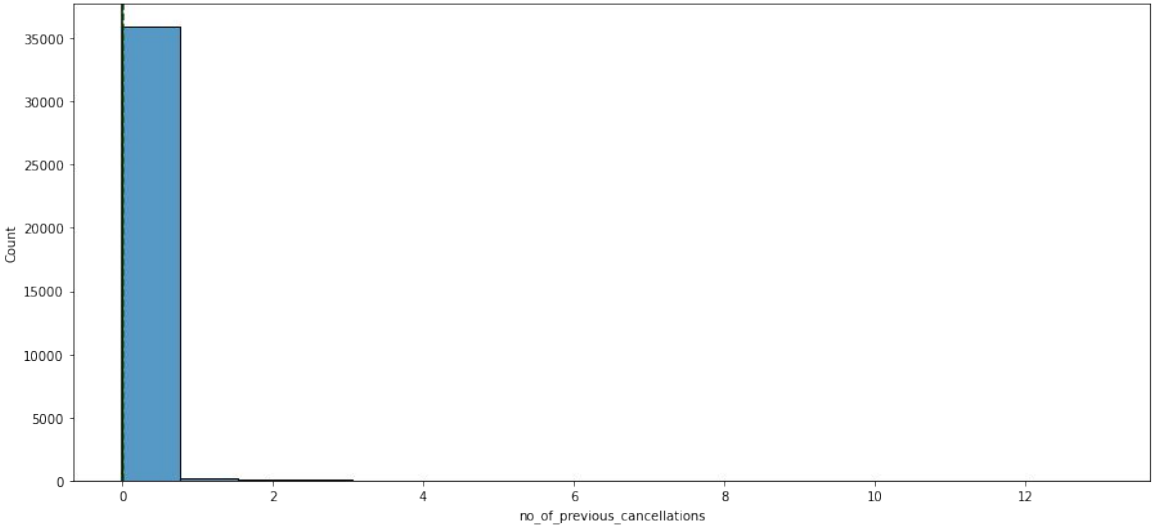
The number of bookings with room reservations averaging 0 euros per day when segmented into markets shows that 354 of those 0 euro bookings are complimentary bookings, while 191 are online customer bookings.

Extreme outlier values in the data can make accurate predictions harder to achieve, a maximum value of 540 euro per day of reservation is a highly infrequent and irregular event, so we calculated the upper whisker and replaced outlier values over or equal to 500 euros with the upper whisker value of 180 euros.

EDA Results

UNIVARIATE ANALYSIS



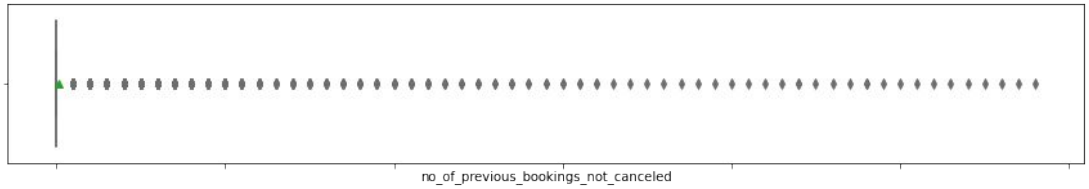


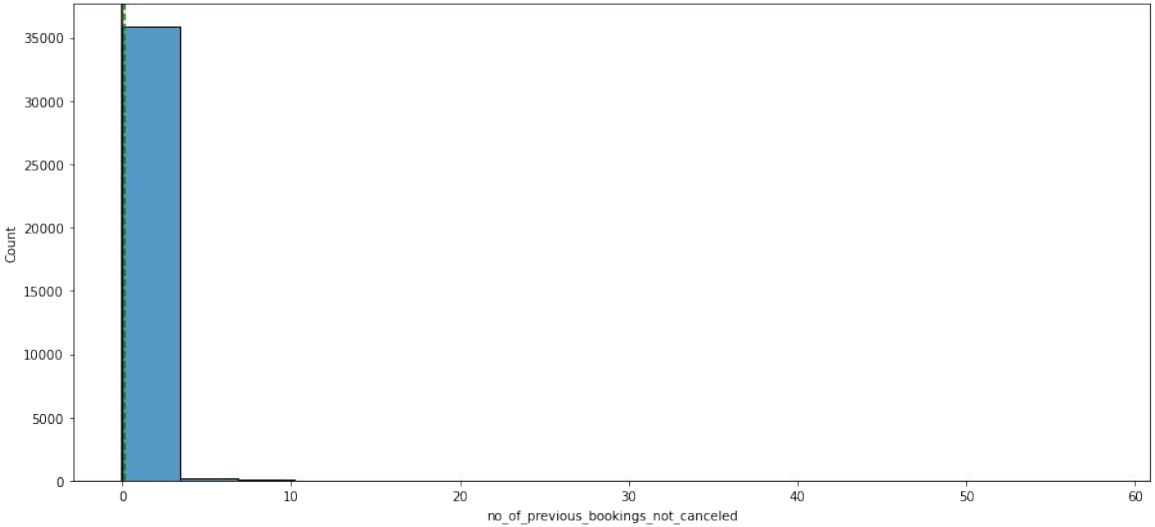
No. of Previous Cancellations

The mean and median value for number previous bookings that were canceled by the customer prior to the current booking is 0. There are some outliers present on the right end of these plots, the distribution is right-skewed.

EDA Results

UNIVARIATE ANALYSIS



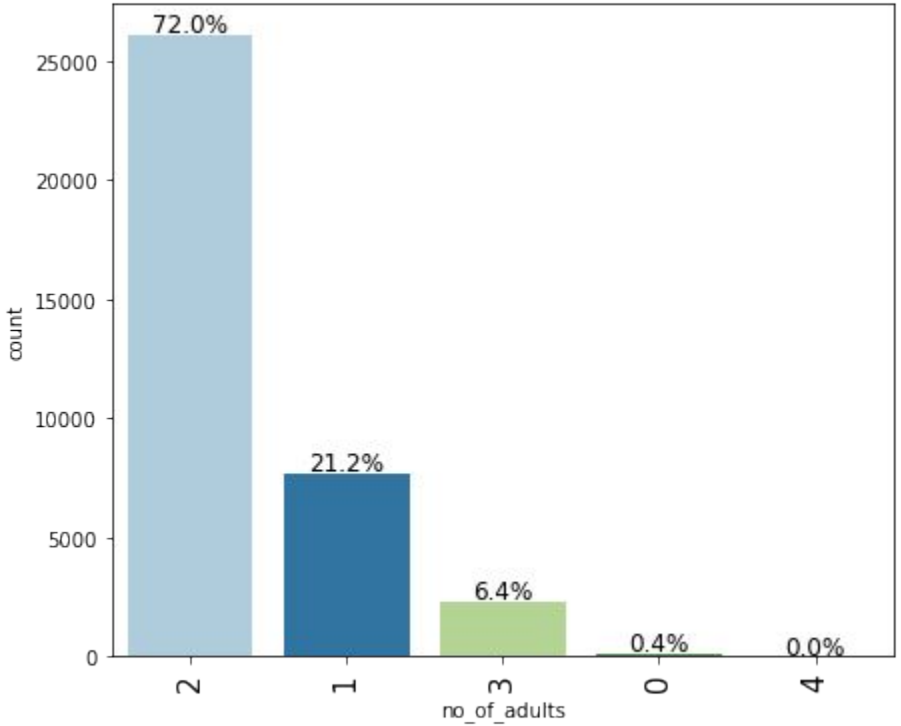


No. of Previous Booking not Cancelled

The mean and median value for number previous bookings that were not canceled by the customer prior to the current booking is 0. There are several outliers present on the right end of these plots, the distribution is right-skewed.

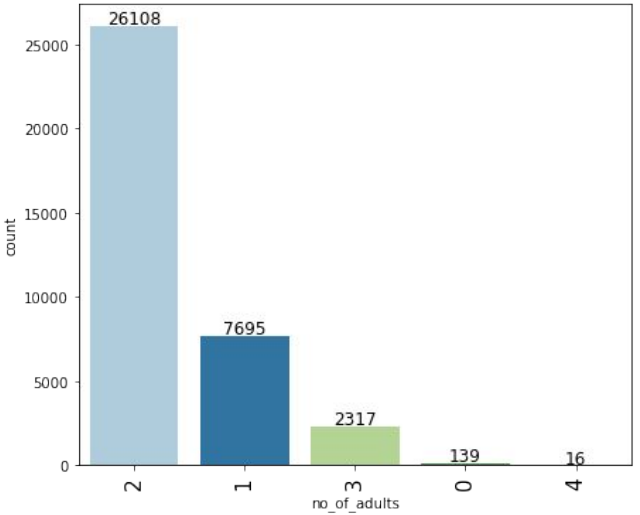
EDA Results

UNIVARIATE ANALYSIS



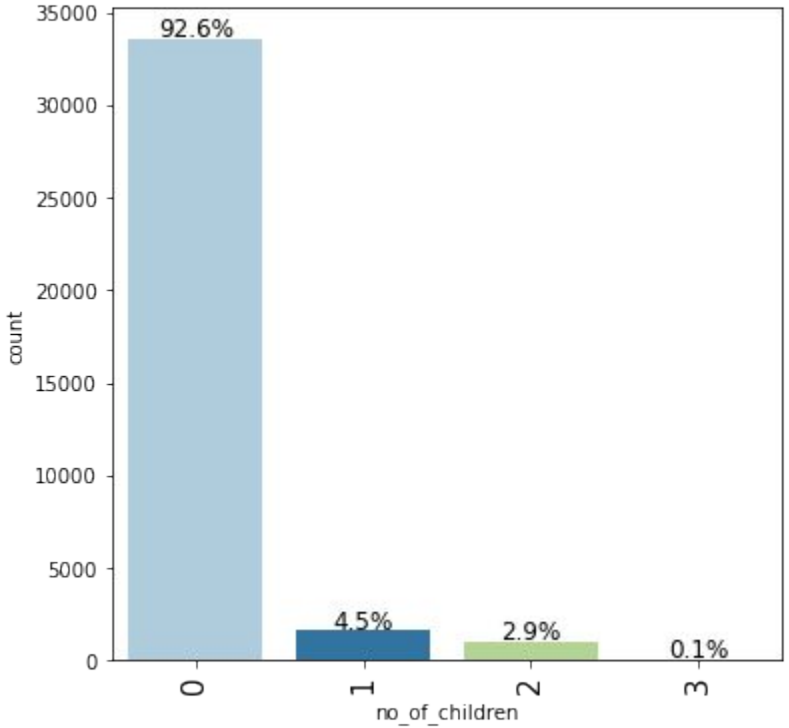
No. of Adults

There are no more than 4 adults per booking. The number of adults per reservation by percentage and count is;
0 adults at 0.4% (139),
1 adult at 21.2% (7,695),
2 adults at 72% (26,108),
3 adults at 6.4% (2,317), and
4 adults at 0% (16).



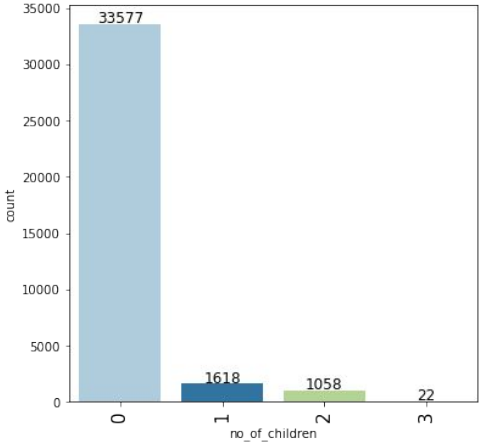
EDA Results

UNIVARIATE ANALYSIS



No. of Children

There are 3 bookings with more than 3 children per reservation. The number of children per reservation by percentage and count is;
0 children adults at 92.6% (33,577),
1 children at 4.5% (1,618),
2 children at 2.9% (1,058),
3 children at 0.1% (19),
9 children at 0% (2), and
10 children at 0% (1).

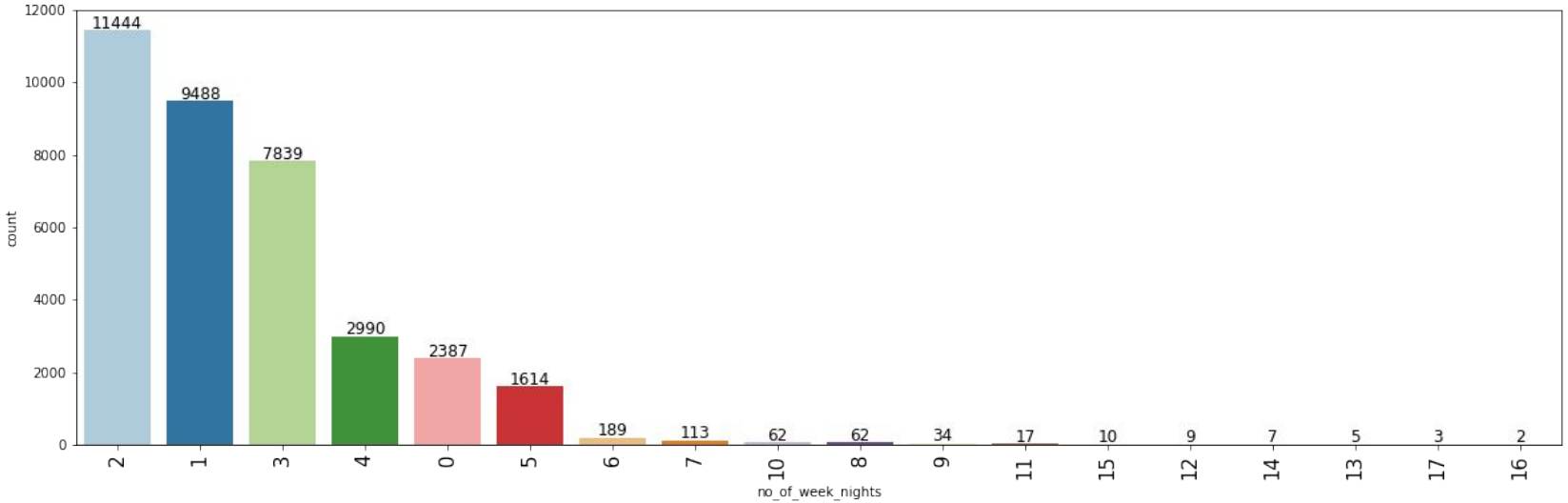


EDA Results

UNIVARIATE ANALYSIS

Number of Week Nights

The highest number of week nights booked per reservation is between 1 and 3, the counts for these 3 values are higher than the count of bookings with more or less week days in total and separately. The counts are as follows; 1 week night (11,444 bookings), 2 week nights (9,488 bookings), 3 week nights (7,839 bookings).

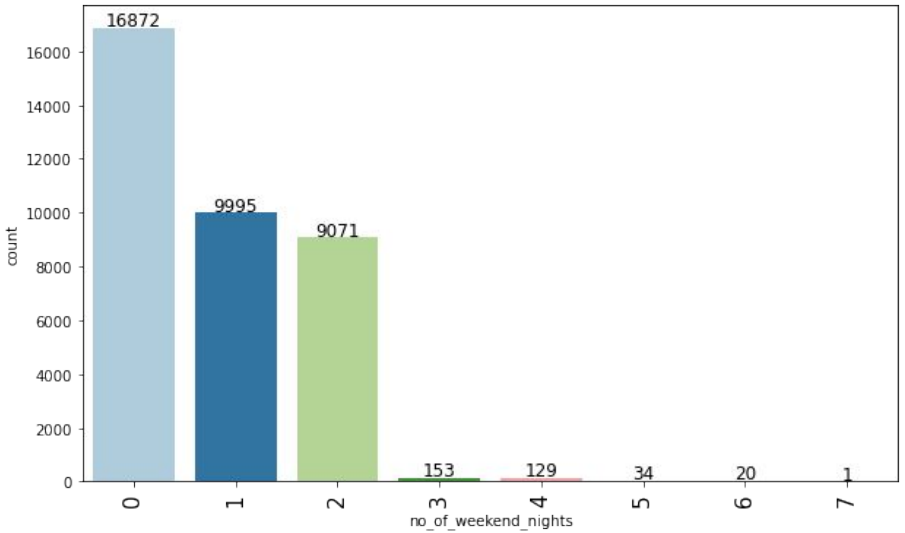


EDA Results

UNIVARIATE ANALYSIS

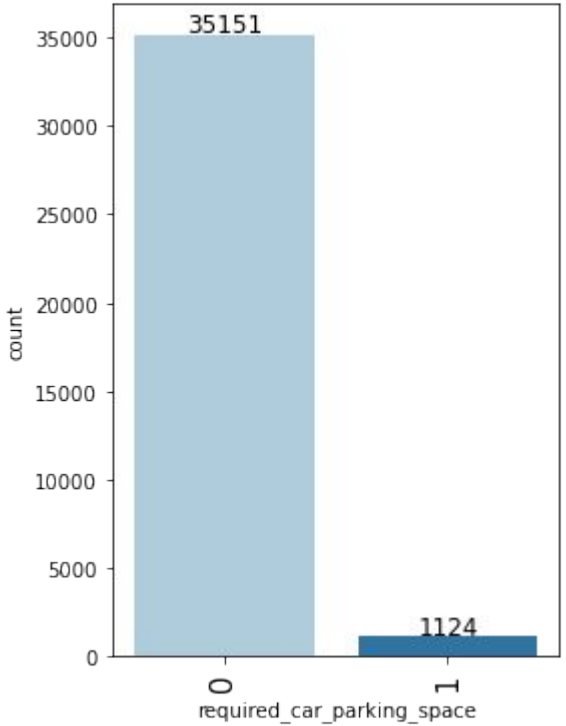
Number of Weekend Nights

The highest number of weekend nights booked per reservation is between 0 and 2, all other values in the data have counts far less significant counts. The counts are as follows; 0 weekend nights (16,872 bookings), 1 weekend night (9,995 bookings), 2 weekend nights (9,071 bookings).



EDA Results

UNIVARIATE ANALYSIS

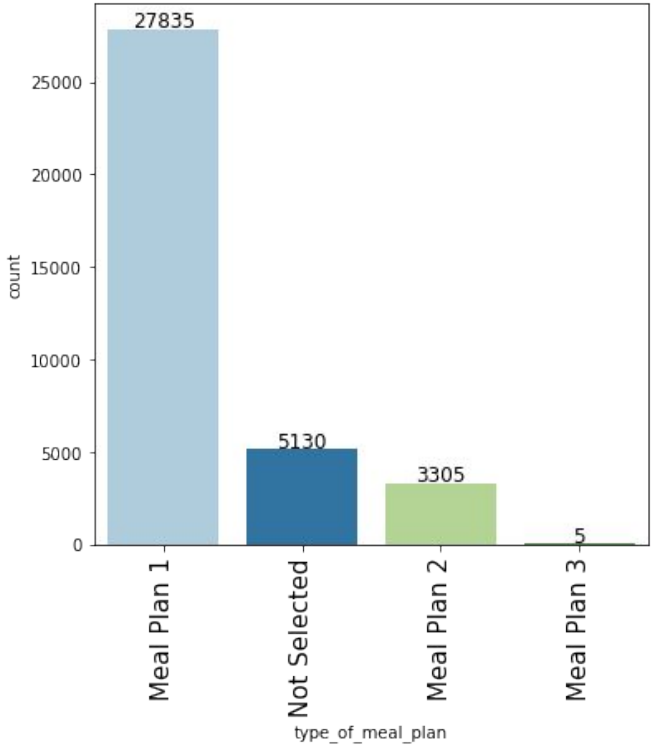


Required Car Parking Space

There are two unique values in the required car parking space dataset, 0 and 1. The counts of these values are as follows; 0 car parking space (35,151 bookings), 1 car parking space (1,124 bookings).

EDA Results

UNIVARIATE ANALYSIS

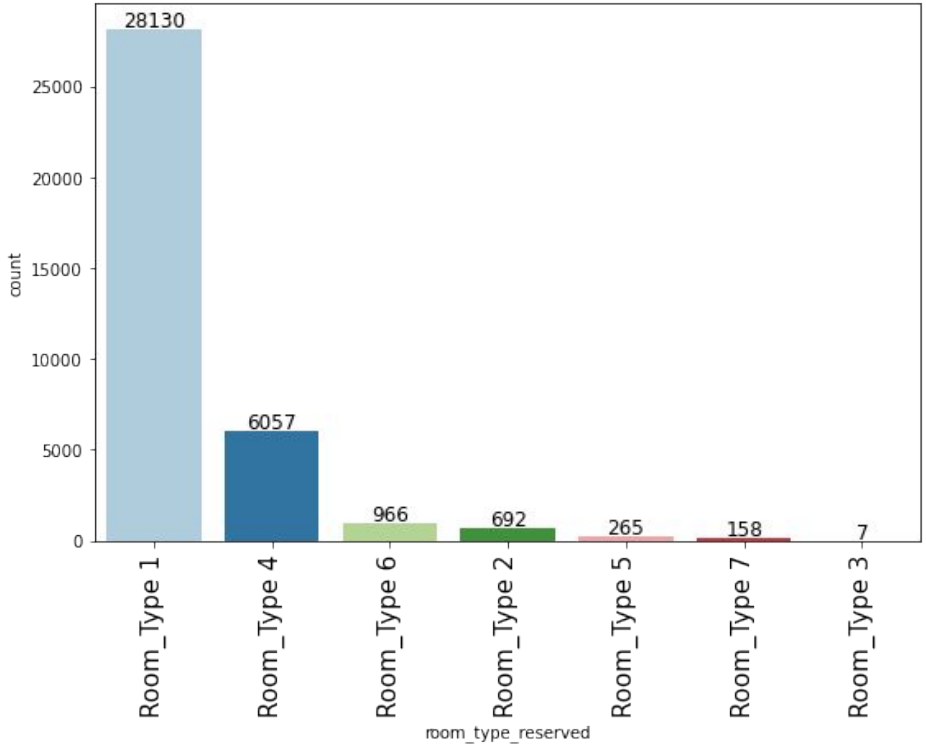


Type of Meal Plan

There are four unique values in the required car parking space dataset; not selected, meal plan 1, meal plan 2 and meal plan 3. The counts of these values are as follows; not selected (5,130 bookings), meal plan 1 (27,835 bookings), meal plan 2 (3,305 bookings), meal plan 3 (5 bookings).

EDA Results

UNIVARIATE ANALYSIS



Room Type Reserved

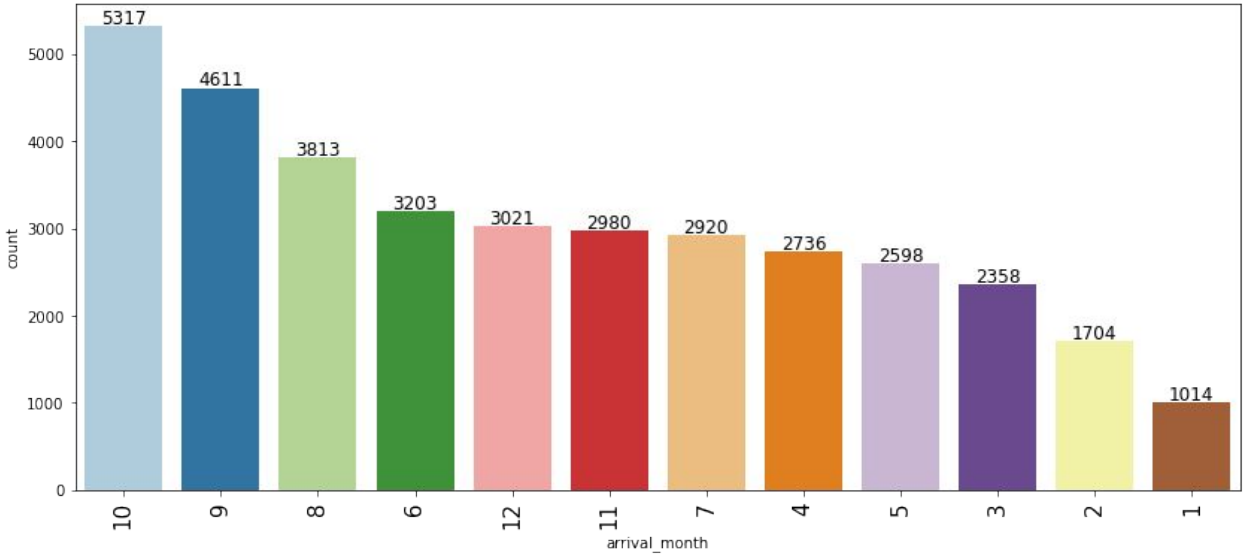
Room_type 1 has the highest reservation count at 28,130. Room_type 4 has the second highest reservation count at 6,057. All other room types have reservation counts below 1,000.

EDA Results

UNIVARIATE ANALYSIS

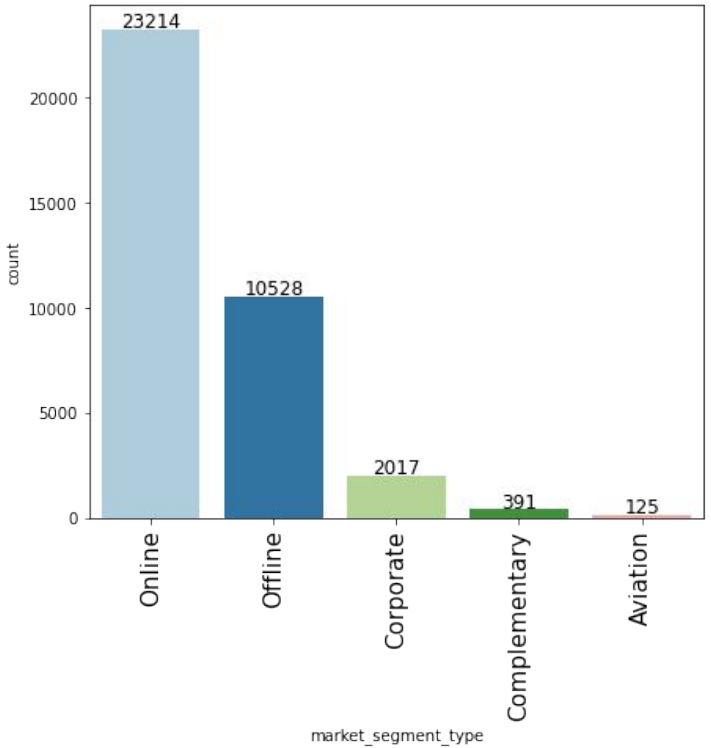
Arrival Month

The 10th month (October) has the highest number reservations recorded of all 12 months at 5,317. The data shows that the second half of the year has more reservations recorded within those months than within the first half of the year.



EDA Results

UNIVARIATE ANALYSIS

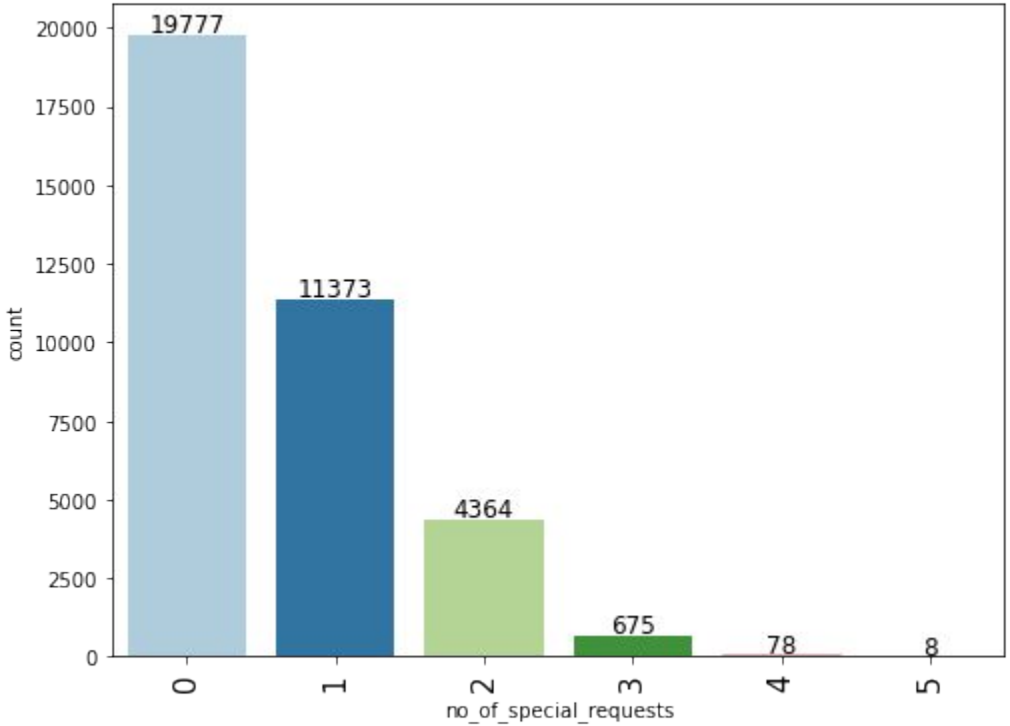


Market Segment Type

Most reservations made by customers are segmented into online and offline types. The online segment has a count of 23,214 and the offline segment has a count of 10,528. All other segments have made less than 2,500 reservations.

EDA Results

UNIVARIATE ANALYSIS

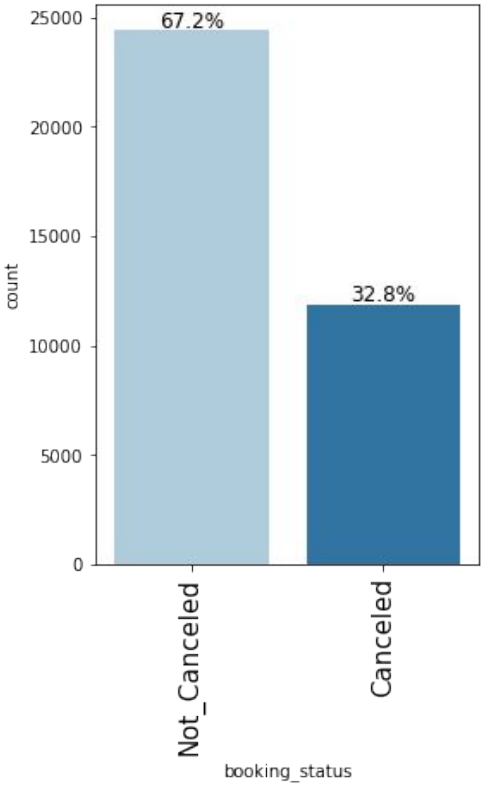


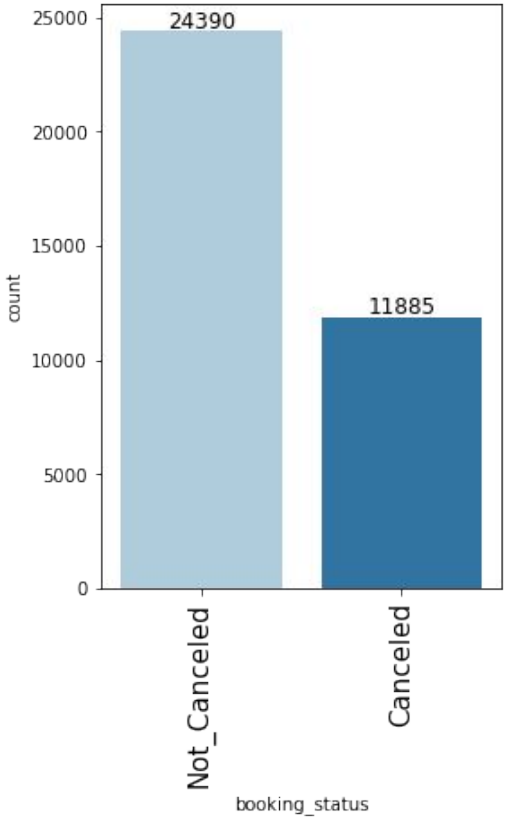
No. of Special Requests

Most reservations do not make any special requests. 19,777 reservations had 0 special requests. 11,373 reservations had 1 special request. 4,364 reservations had 2 special requests. 675 reservations had 3 special requests. Reservations with more than 3 special requests were less than 100.

EDA Results

UNIVARIATE ANALYSIS





Booking Status

24,390 reservations (67.2%) were not cancelled, while 11,885 reservations (32.8%) were cancelled.

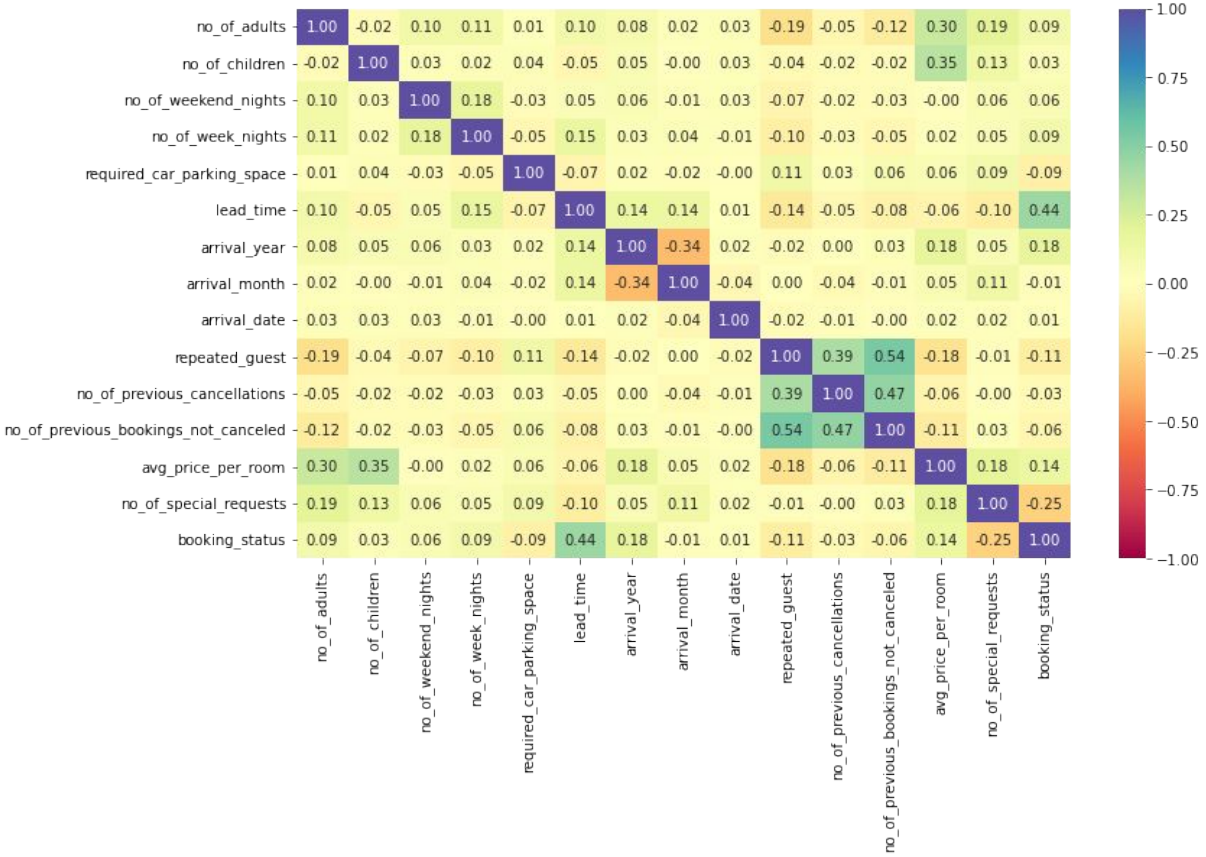
EDA: BIVARIATE ANALYSIS

EDA Results

BIVARIATE ANALYSIS

Heatmap

The plotted heatmap shows little to no correlation between columns in the dataset. The columns with the highest correlation ranging between 40% and 55% are; Number of previous bookings not canceled and repeated guest (54%), Number of previous bookings canceled and Number of previous bookings not canceled (47%), Booking status and lead time (44%).

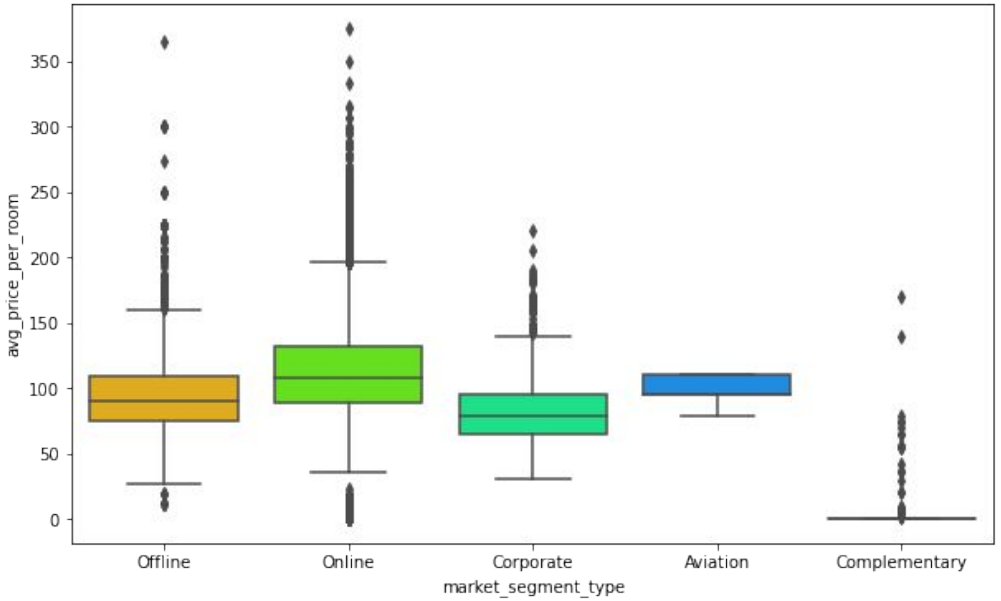


EDA Results

BIVARIATE ANALYSIS

Hotel rates are dynamic and change according to demand and customer demographics. We plotted a boxplot to see how prices vary across different market segments

Complementary segments on average pay less than other segments. Online segments on average pay higher prices per day of the reservation than other segments. All segments, with the exception of the complementary segment, have median values between 75 euros and 125 euros.

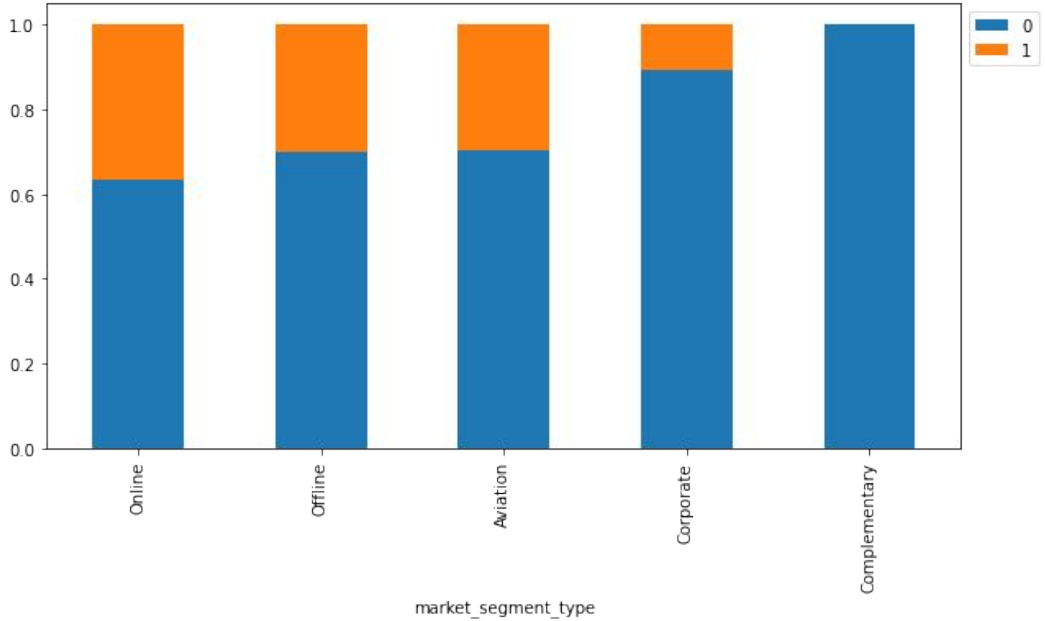


EDA Results

BIVARIATE ANALYSIS

We made a barplot to see how booking status varies across different market segments. Also, how average price per room impacts booking status

Bookings made by the online segment had the highest number of cancellations (8,475 cancellations) of all other segments. Complementary segment had no cancelation, and considering the complementary segment also pays the least average price per day of the reservation, this could be a contributing factor to the lack of cancellations in this market segment.

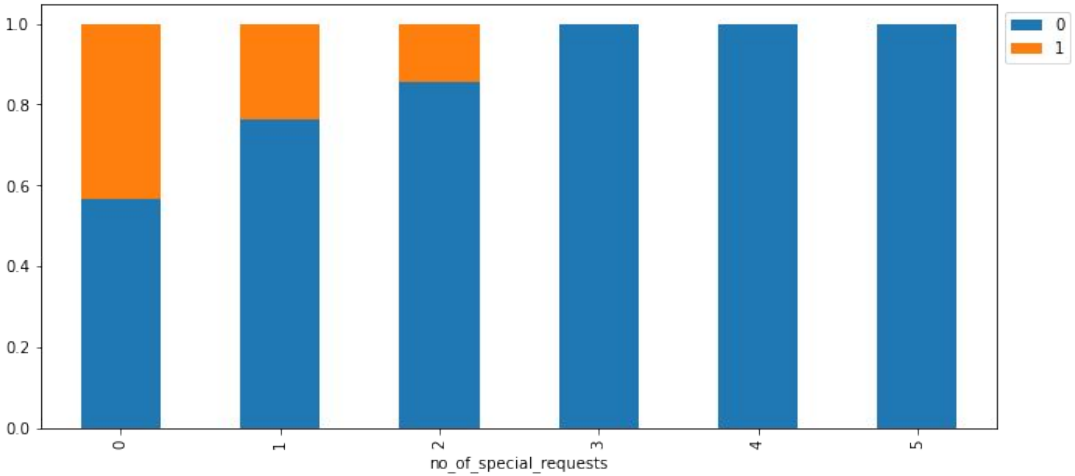


EDA Results

BIVARIATE ANALYSIS

Many guests have special requirements when booking a hotel room. We want to see how this impacts cancellations

Most cancellations made had no special requests. 2,703 reservations with 1 special request eventually canceled. 637 reservations with 2 special requests eventually canceled. No reservation with more than 2 special requests was canceled.

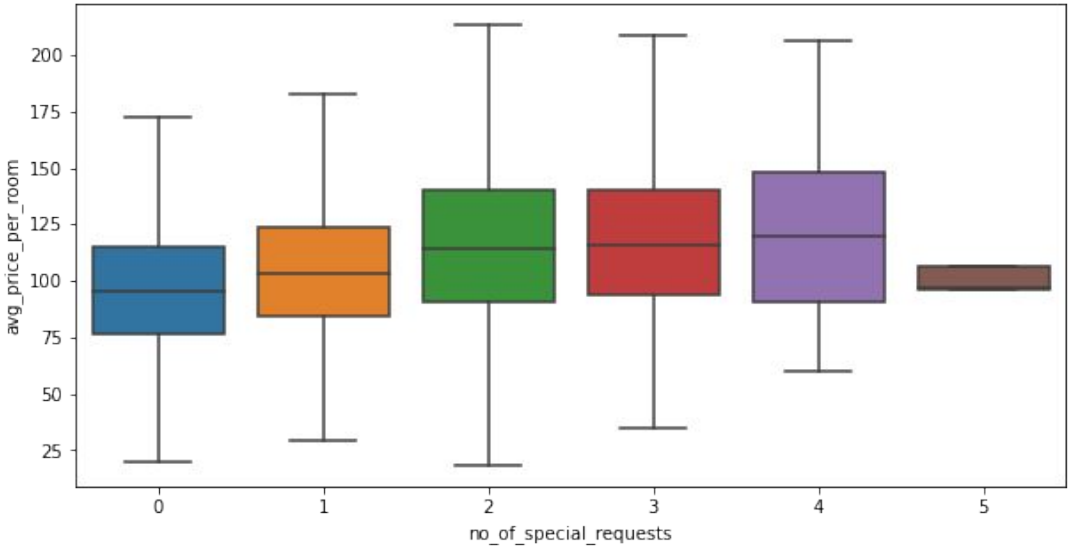


EDA Results

BIVARIATE ANALYSIS

We also checked if the special requests made by the customers impacts the prices of a room

In the absence of outliers, reservations made with one or more special requests tend to have higher median prices per day of the reservation. They also tend to have higher minimum and maximum prices than reservations without special requests.

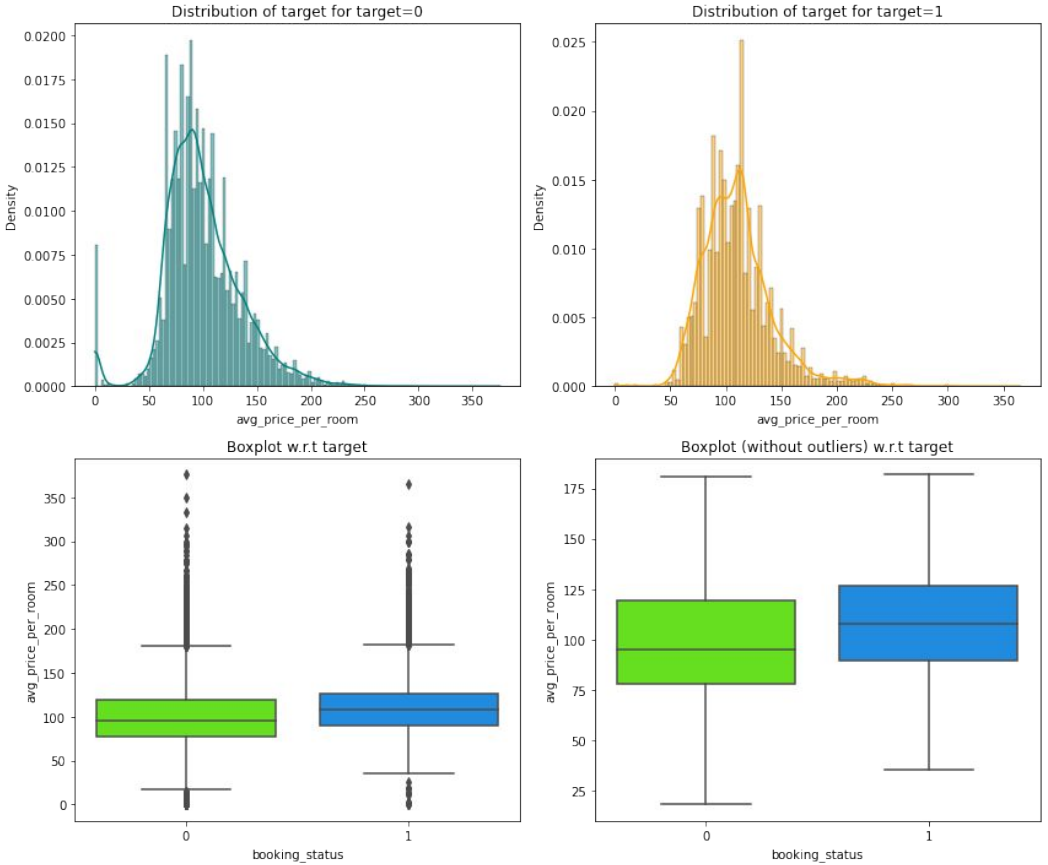


EDA Results

BIVARIATE ANALYSIS

We saw earlier that there is a positive correlation between booking status and average price per room. Upon analysis, we found that;

Reservations that were canceled tend to have higher median prices per day of the reservation than those that were not canceled. The box plot also shows higher lower whiskers in canceled reservations compared to bookings not canceled. The histogram shows the highest density of prices for canceled bookings to be somewhere around 115 euros, whereas, the highest density area(s) in average price per day for retained reservations were around 90 euros and below.



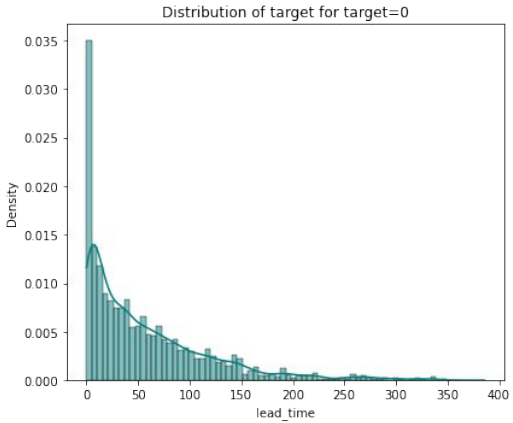
EDA Results

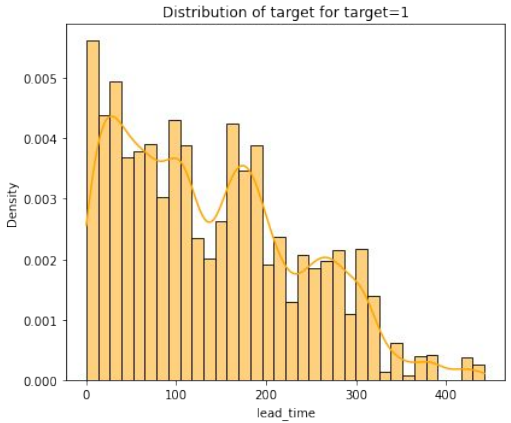
BIVARIATE ANALYSIS

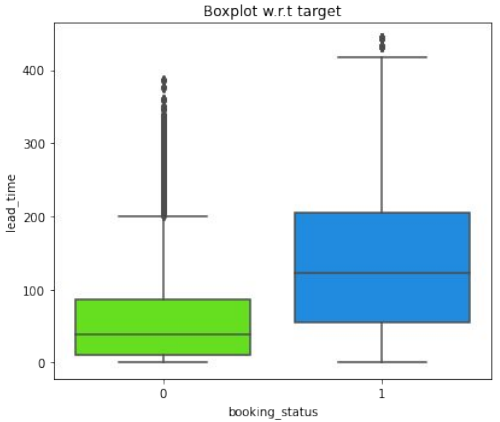
There is a positive correlation between booking status and lead time also. Upon analysis, we found that;

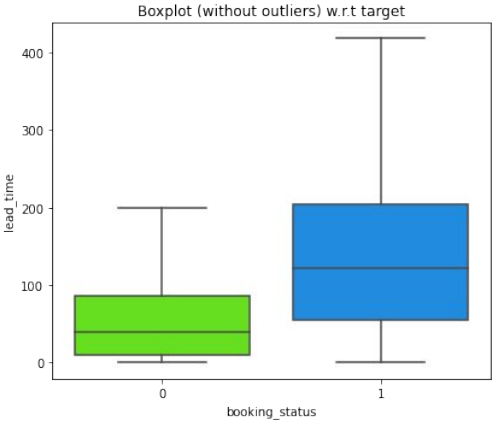
The distribution of both canceled and retained reservations are highly right skewed. The highest density area for retained reservations had a lead time of less than 10 days. Similarly, canceled reservations had the highest density area of lead times under than 15 days. However, unlike retained reservations the difference isn't so stark in comparison with other high density areas within the same plot.

The box plot clarifies even further and confirms that canceled reservations tend to have higher median lead times and maximum values than retained reservations.









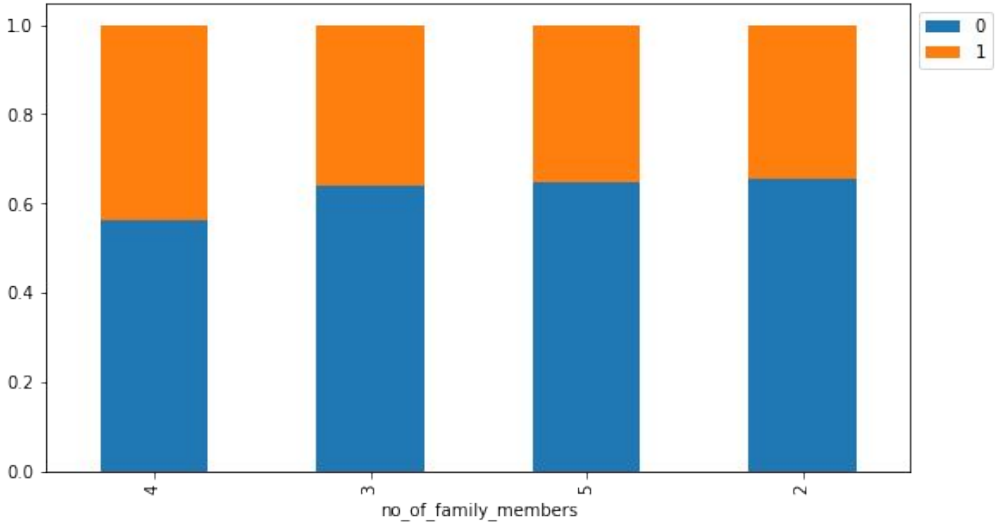
EDA Results

BIVARIATE ANALYSIS

Generally people travel with their spouse and children for vacations or other activities. We created a new dataframe of the customers who traveled with their families and analyzed the impact on booking status

We merged the number of children greater or equal to 0 with the number of adults greater than 1 to form a unique column for family data. Although, not every reservation for more than one adult constitutes a family, we will make the assumption for the sake of the observation and analysis.

Family members of 4 canceled on more counts than smaller groups. Together, all family groups canceled 9,985 reservations, this is 35.11% of the total reservations made by family groups. There is a uniform cancelation and retention rate across all family groups in the data.

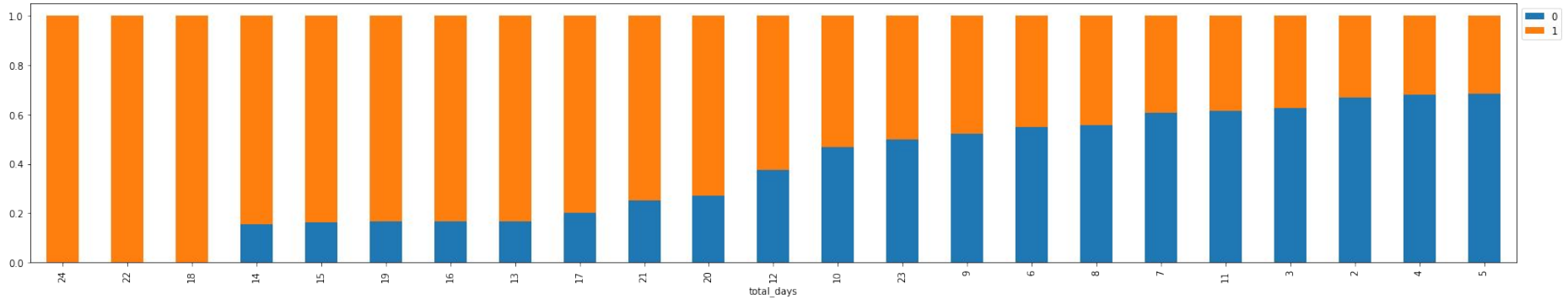


EDA Results

BIVARIATE ANALYSIS

We did a similar analysis for the customer who stay for at least a day at the hotel.

We merged the number of week and weekend nights greater than 0 into a unique column for total stay data. Every booking made for stays totaling 18, 22, and 24 nights were canceled. The bar plot shows visually how the cancelation rate increases as the total number of nights booked increases. Bookings with shorter total stays (below 10 days) generally tend to cancel reservations less frequently, with very few exceptions, than reservations made for long stays (10 and above).

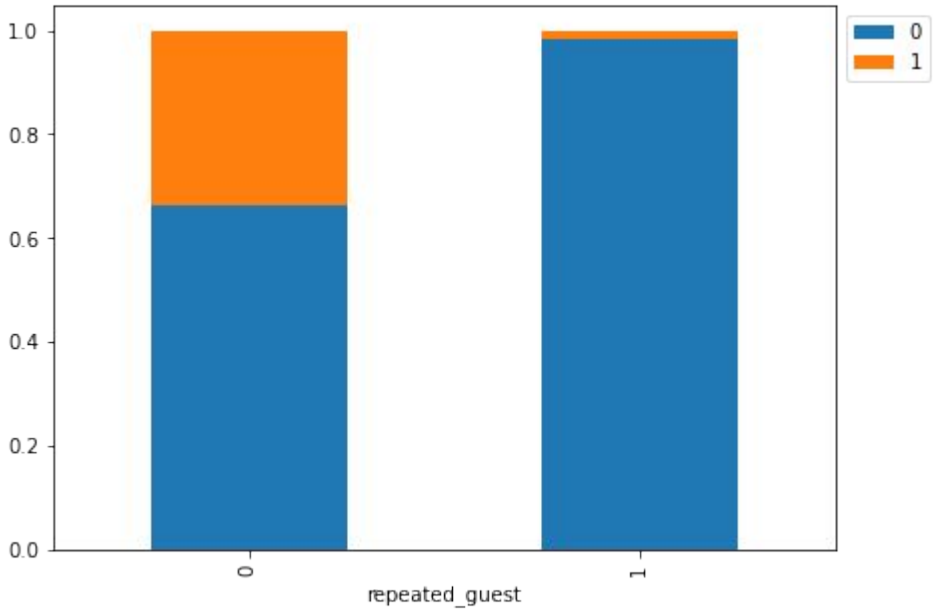


EDA Results

BIVARIATE ANALYSIS

Repeating guests are the guests who stay in the hotel often and are important to brand equity. We checked what percentage of repeating guests cancel.

Repeat guests cancel reservations far less often than new guests. Of all repeat guests, only 1.72% canceled their reservation, while 34% of new guests canceled reservations in the same period.

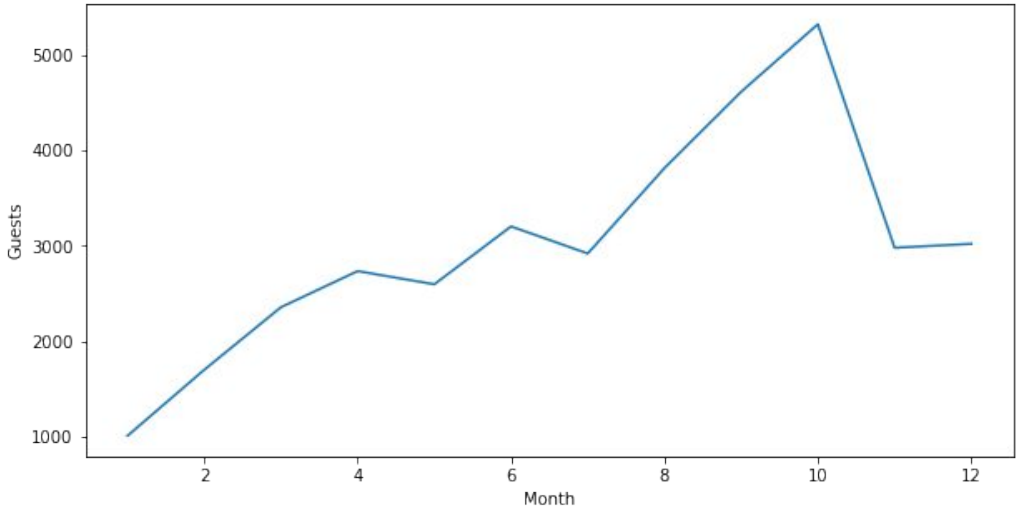


EDA Results

BIVARIATE ANALYSIS

We checked the data for the busiest months in the hotel.

As confirmed previously in our univariate analysis of arrival month data, the 10th month (October) has the highest number reservations recorded of all 12 months. October is the busiest month for INN Hotels, followed by September and August.

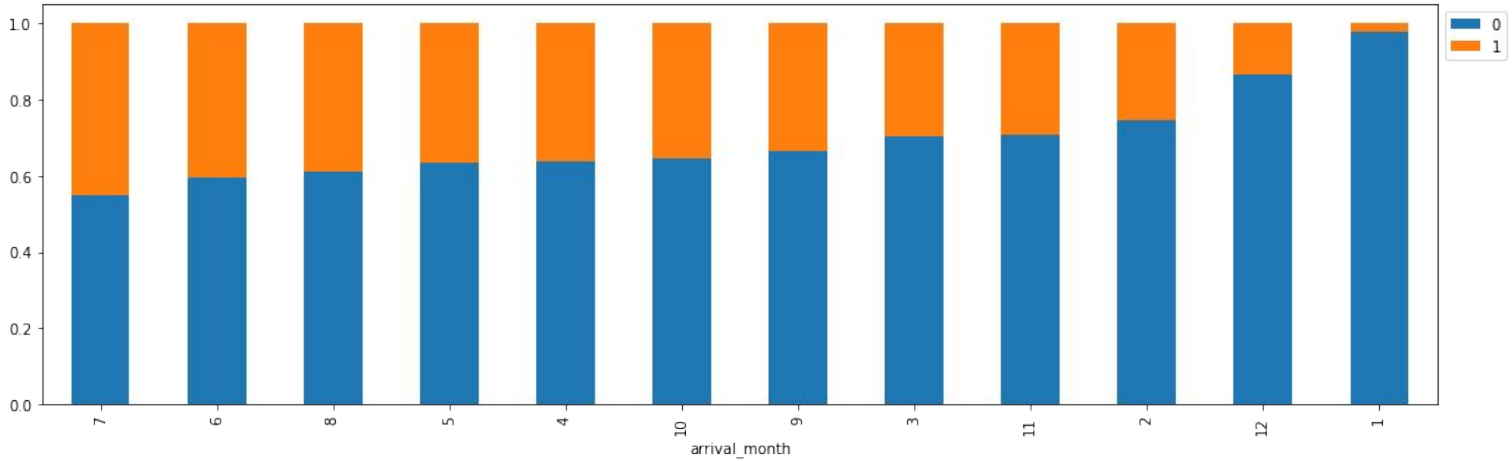


EDA Results

BIVARIATE ANALYSIS

We checked the percentage of bookings canceled in each month.

January has the lowest percentage of cancellations, while July has the highest percentage of canceled bookings. These are the percentages of canceled bookings in each month; 1 January (2.37%), 2 February (25.23%), 3 March (29.69%), 4 April (36.37%), 5 May (36.49%), 6 June (40.31%), 7 July (45%), 8 August (39.02%), 9 September (33.36%), 10 October (35.36%), 11 November (29.36%), 12 December (13.31%).

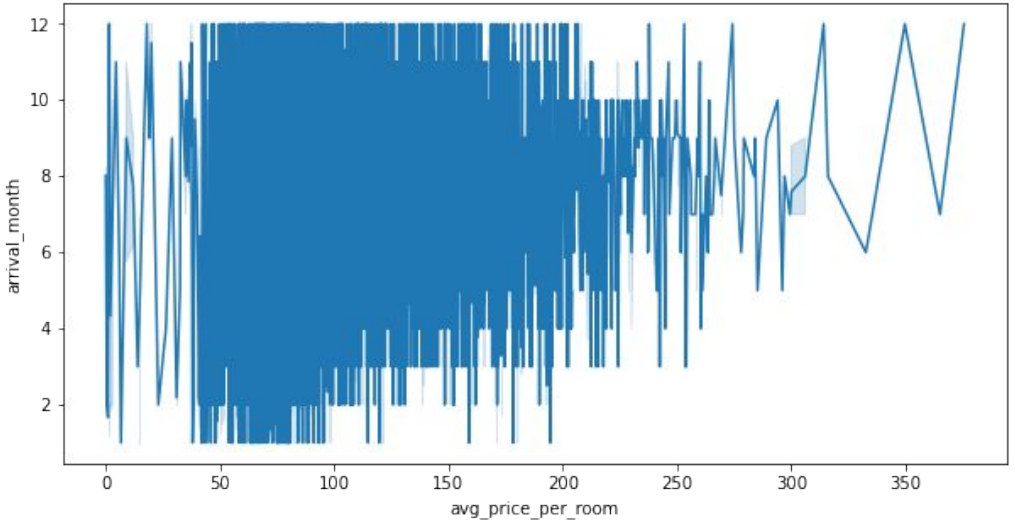


EDA Results

BIVARIATE ANALYSIS

As hotel room prices are dynamic, we checked how prices vary across different months.

The average price per day of the reservation is highest in the second half of the year. From March, prices appear to steadily increase and after july prices shoot up even higher.

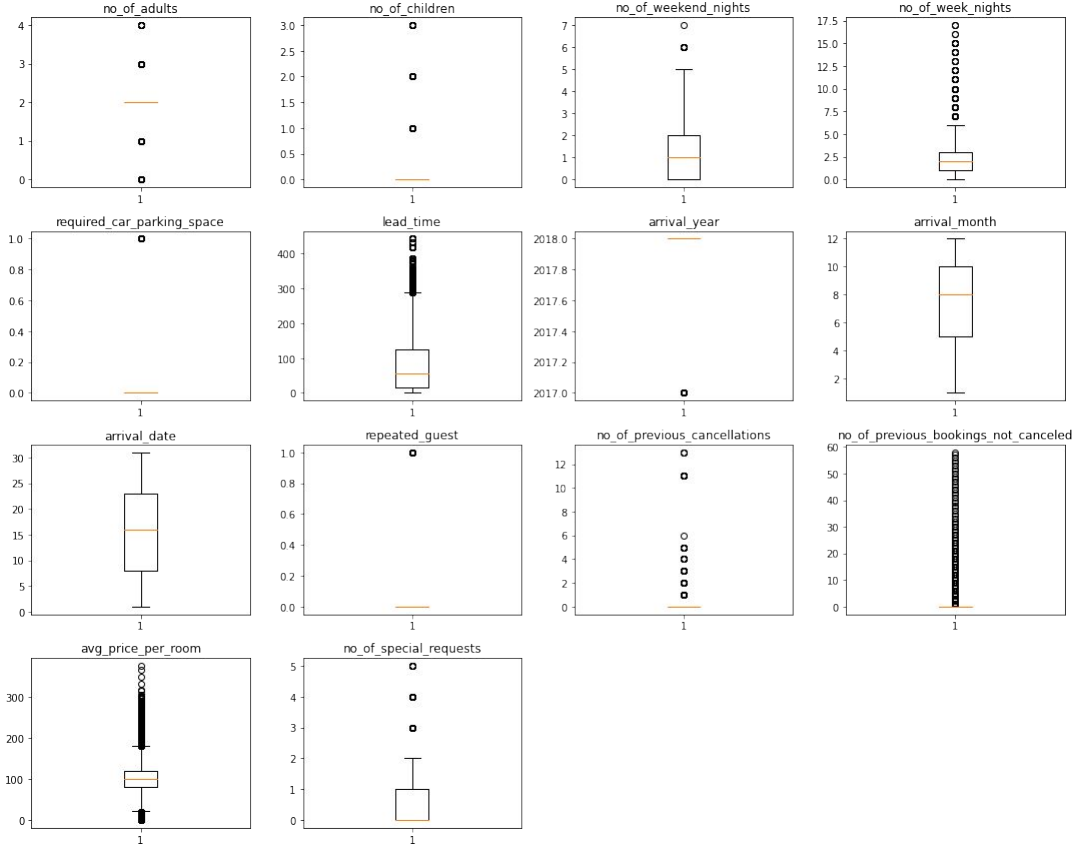


DATA PROCESSING & MODELING CRITERIA

Data Processing: Outlier Detection and Treatment

Outlier Check:

There are still numerous outliers in the data but we will not treat them as they are proper values. Outliers that we need to be corrected have already been in the univariate analysis section of average price per room.



Model Building Criterion

Model can make wrong predictions as:

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

Which case is more important?

Both the cases are important as:

- If we predict that a booking will not be canceled and the booking gets canceled then the hotel will lose resources and will have to bear additional costs of distribution channels.
- If we predict that a booking will get canceled and the booking doesn't get canceled the hotel might not be able to provide satisfactory services to the customer by assuming that this booking will be canceled. This might damage the brand equity.

How to reduce the losses?

- Hotel would want F1 score to be maximized, greater the F1 score higher are the chances of minimizing False Negatives and False Positives.

LOGISTIC REGRESSION MODELING

Data Processing Overview

Data Preprocessing:

Since there are no missing values, we do not have to treat the data for this.

Data Preparation for Modeling:

We encoded the categorical columns/features, then split the data into ‘training’ and ‘testing’ data in order to train and test the built model. We split the data into 70:30; 67% of observations belongs to class 0 and 32.9% observations belongs to class 1. The training set has 25,392 rows and 28 columns, while the test set has 10,883 rows and 28 columns.

Initial 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, 10 Dec 2022	Pseudo R-squ.:	0.3292
Time:	12:04:49	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

Training performance:

	Accuracy	Recall	Precision	F1
0	0.80600	0.63410	0.73971	0.68285

The f1 score of the model is 0.68 and we will try to maximize it further.

The variables used to build the model contain multicollinearity, which will affect the p-values.

We'll have to remove multicollinearity from the data to get reliable coefficients and p-values.

Initial Model Building and Performance Evaluation

Initial Model Performance Evaluation:

Negative values of the coefficient show that the probability that a person will cancel their reservation decreases with the increase of the corresponding attribute value.

Positive values of the coefficient show that the probability that a person will cancel their reservation increases with the increase of the 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.

market_segment_type_Complementary, market_segment_type_Online, room_type_reserved_Room_Type 3, type_of_meal_plan_Meal Plan 3, no_of_previous_bookings_not_canceled, and arrival_date appear to be the most insignificant variables in this model.

We'll be dropping these values using a loop statement, building separate models, and checking their performances to see if we have successfully fixed any multicollinearity issues and gauge the impact on the model's performance.

Checking Multicollinearity

Checking VIF (Multicollinearity)

Three market segment types exhibit high multicollinearity, these variables have VIF values of greater than 5. Such high VIF scores indicate perfect correlation between variables and since these variable capture similar information, it is understandable that they have high VIF values.

	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

Final Logistic Regression Model

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25370
Method:	MLE	Df Model:	21
Date:	Sat, 10 Dec 2022	Pseudo R-squ.:	0.3282
Time:	12:04:51	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

	Accuracy	Recall	Precision	F1
0	0.80545	0.63267	0.73907	0.68174

We dropped variables with high p-values and checked the performance of the new model. There is no significant change in the model performance as compared to initial model.
Now no categorical feature has p-value greater than 0.05, so we'll consider the features in X_train1 as the final ones and lg1 as final model.

Converting Coefficients to Odds

Results:

	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
Odds	0.00000	1.11491	1.16546	1.11470	1.04258	0.20296	1.01583	1.57195	0.95839	0.06478
Change_odd%	-100.00000	11.49096	16.54593	11.46966	4.25841	-79.70395	1.58331	57.19508	-4.16120	-93.52180

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
1.25712	1.01937	0.22996	1.17846	1.33109	0.70104	0.75364
25.71181	1.93684	-77.00374	17.84641	33.10947	-29.89588	-24.63551

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
0.47885	0.37977	0.23827	0.45326	0.16773
-52.11548	-62.02290	-76.17294	-54.67373	-83.22724

Coefficient Interpretations

Lead Time: Holding all other features constant, a 1 unit change in lead time will increase the odds of a guest canceling their reservation by 1.01 times or a 1.58% increase in odds of a guest canceling their reservation.

Repeated Guest: Holding all other features constant, a 1 unit change in repeated guest will decrease the odds of a guest canceling their reservation by 0.06 times or a 93.52% decrease in odds of a guest canceling their reservation.

Average Price Per Room: Holding all other features constant, a 1 unit change in average price per room will increase the odds of a guest canceling their reservation by 1.01 times or a 1.93% increase in odds of a guest canceling their reservation.

Type of Meal Plan: Holding all other features constant, the odds of a guest whose meal plan type is type 2 canceling their reservation is 1.178 more than the guest whose meal plan type is not selected, or 17.8% greater odds of canceling their reservation than the guest whose meal plan type is not selected.

Similarly, the odds of guest whose meal plan type is not selected canceling their reservation is 1.33 times more than a guest whose meal plan type is type 2, or 33.1% greater odds of canceling their reservation than the guest whose meal plan type is not selected.

Other attributes in the data can be accessed and interpreted similarly.

Confusion Matrix Overview

True Positives (TP): A guest cancels their reservation and the model predicted guest will cancel their reservation.

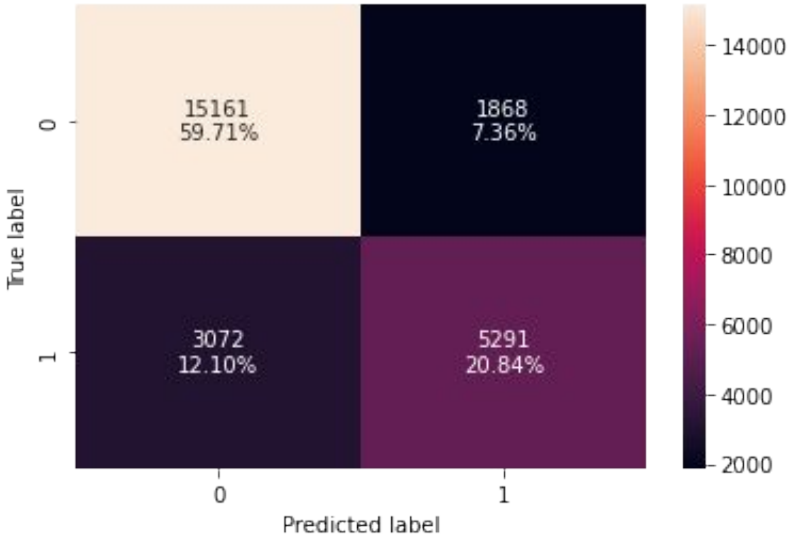
True Negatives (TN): A guest doesn't cancel their reservation and the model predicted guest won't cancel their reservation.

False Positives (FP): The model predicted guest will cancel their reservation but the guest doesn't cancel their reservation.

False Negatives (FN): The model predicted guest won't cancel their reservation but the guest cancels their reservation.

Confusion Matrix: Default Threshold (0.50)

TRAINING SET:

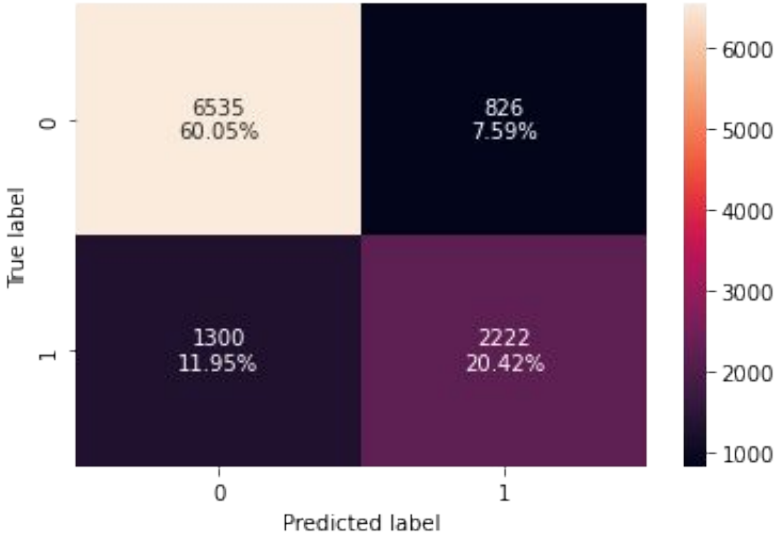


	Accuracy	Recall	Precision	F1
0	0.80545	0.63267	0.73907	0.68174

The final model is giving an f1 score of 0.681 on the tuned training set but we want to improve the performance of this model.

Confusion Matrix: Default Threshold (0.50)

TEST SET:



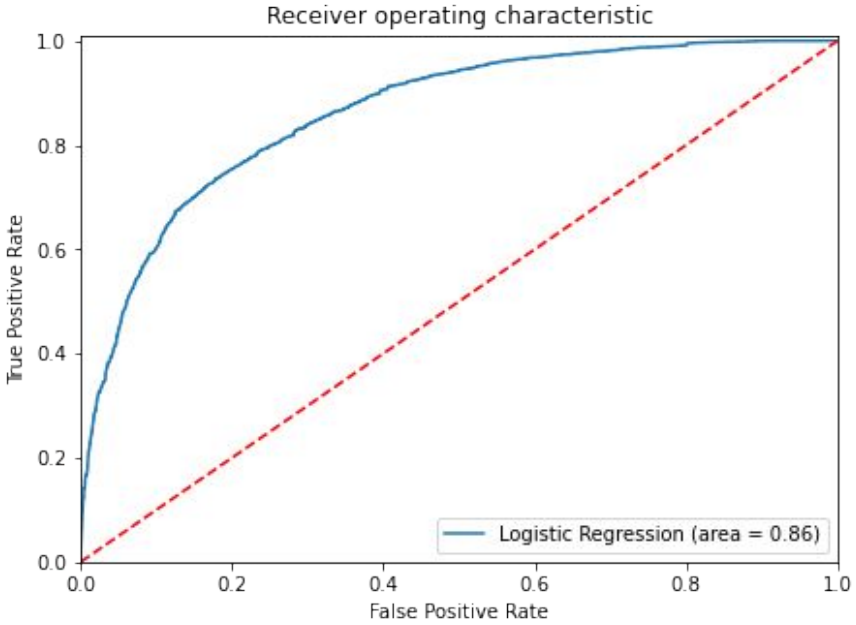
	Accuracy	Recall	Precision	F1
0	0.80465	0.63089	0.72900	0.67641

We checked the performance of this model on the test set and got an f1 score of 0.67. All other resulting values are comparable to the same performance check values gotten from the training set.

ROC-AUC

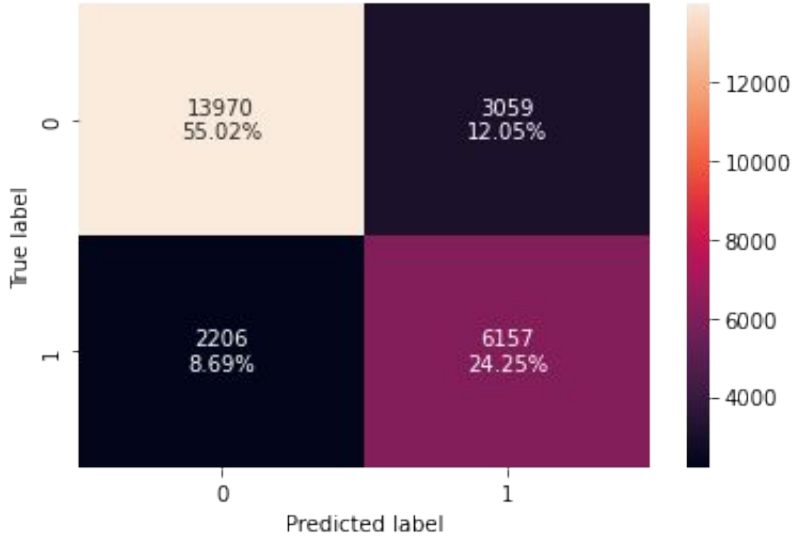
Model Performance Improvement:
We want to see if the recall score can be improved further, by changing the model threshold using AUC-ROC Curve.

Our optimal threshold using AUC-ROC curve is found to be 0.37.



Confusion Matrix: Optimal threshold using AUC-ROC curve (0.37)

TRAINING SET:

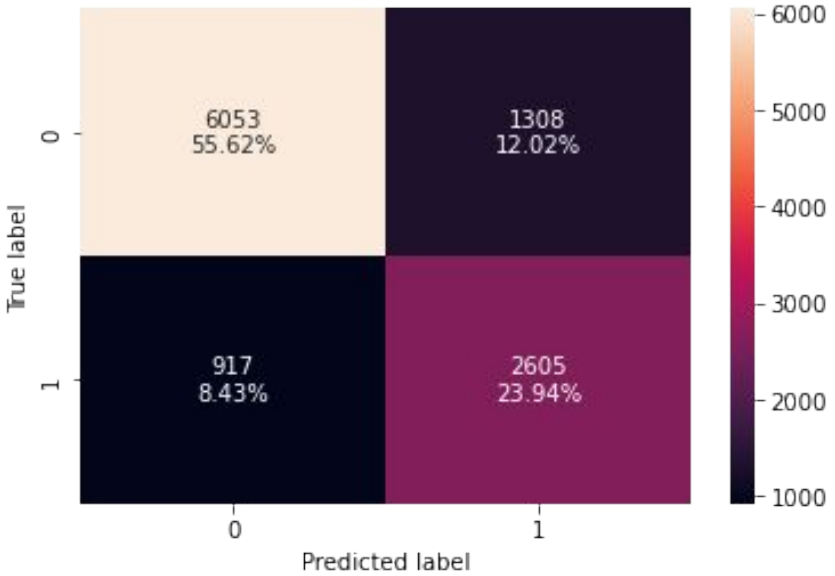


	Accuracy	Recall	Precision	F1
0	0.79265	0.73622	0.66808	0.70049

We changed the model threshold using AUC-ROC Curve value of 0.37. We checked the performance of this model on the training set and found that we were able to improve our recall and precision scores.

Confusion Matrix: Optimal threshold using AUC-ROC curve (0.37)

TEST SET:



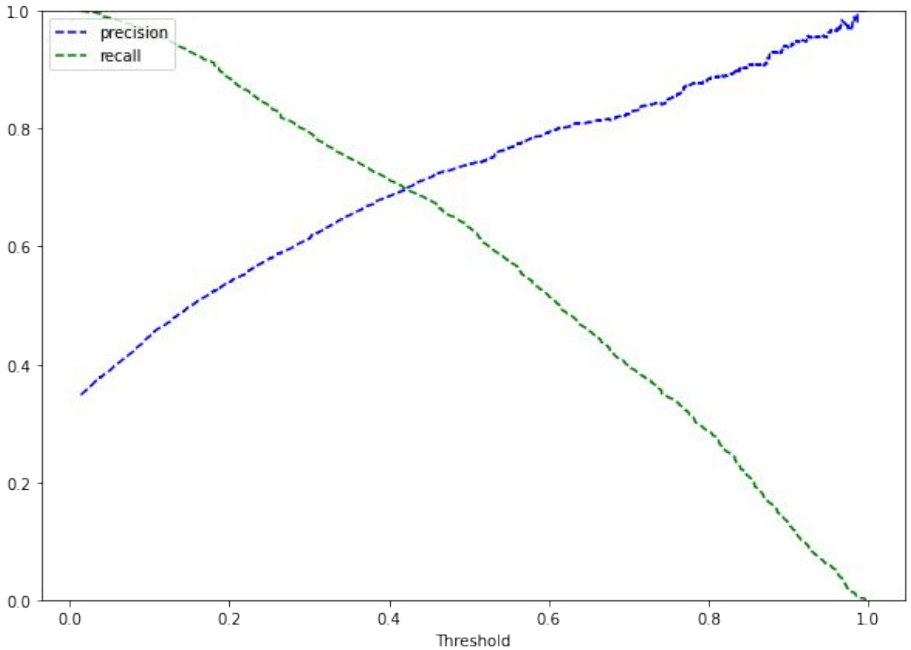
	Accuracy	Recall	Precision	F1
0	0.79555	0.73964	0.66573	0.70074

We changed the model threshold for the test set, as we did with the training set, using AUC-ROC Curve value of 0.37. We checked the performance of this model and found the training and test set results to be comparable and satisfactory.

Precision-Recall Curve

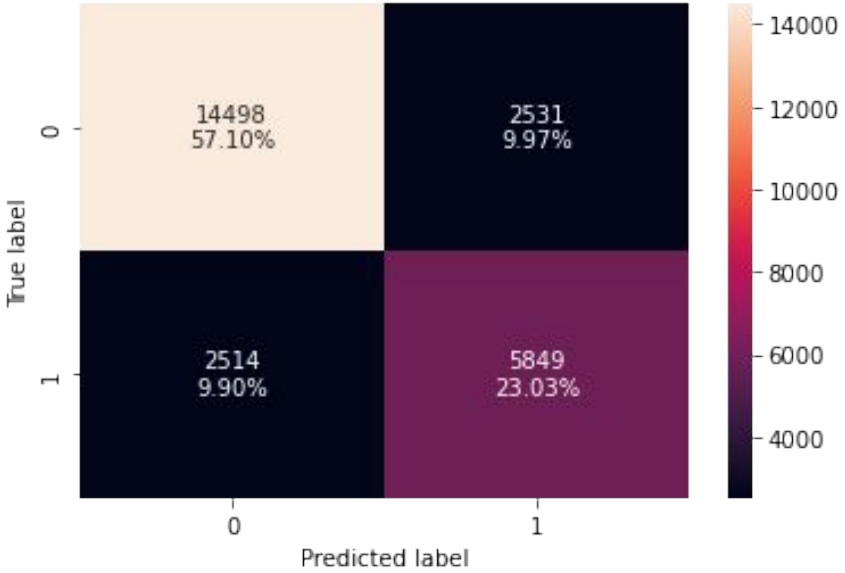
Model Performance Improvement:

We want to see if our f1 score can be improved further, by changing the model threshold using Precision-Recall curve. Our optimal threshold using Precision-Recall curve is found to be 0.42.



Confusion Matrix: Optimal threshold, Precision-Recall curve (0.42)

TRAINING SET:

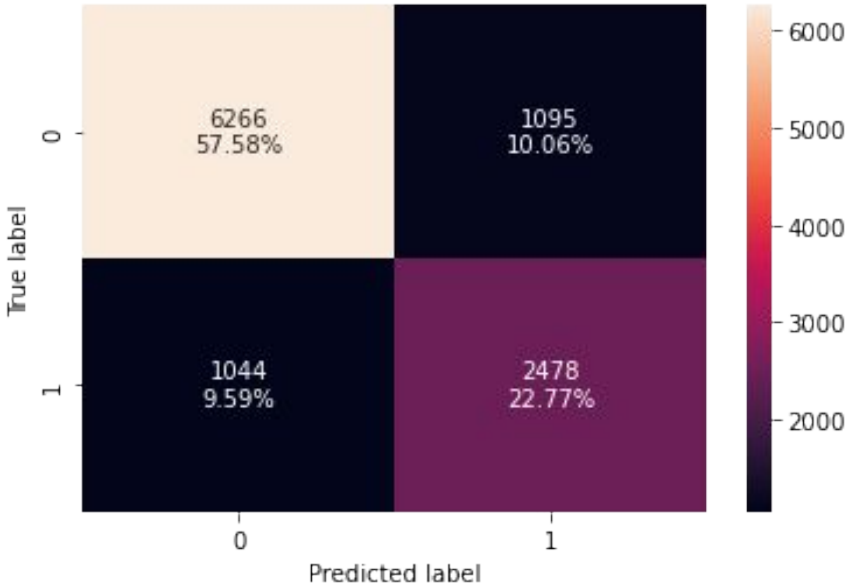


	Accuracy	Recall	Precision	F1
0	0.80132	0.69939	0.69797	0.69868

To find a better threshold, we changed the model threshold using Precision-Recall curve value of 0.42. We checked the performance of this model on the training set and found that we were able to improve our accuracy and precision from the previous AUC-ROC curve threshold, but experienced a fall in our recall and f1 scores.

Confusion Matrix: Optimal threshold, Precision-Recall curve (0.42)

TEST SET:



	Accuracy	Recall	Precision	F1
0	0.80345	0.70358	0.69353	0.69852

We changed the model threshold on the test set, as we did with the training set, using Precision-Recall curve value of 0.42. We checked the performance of this model and found that the results were comparable to the training set.

Model Performance Summary: Logistic Regression

TRAINING SET:

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

TEST SET:

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

Conclusion: Of the three models, some performed better than others based on different criteria in both training and test data without overfitting the data. The model with a AUC-ROC curve threshold (0.37) is giving the best F1 score. Therefore, we will choose this as the final model.

DECISION TREE MODELING

Data Processing Overview

Data Preprocessing:

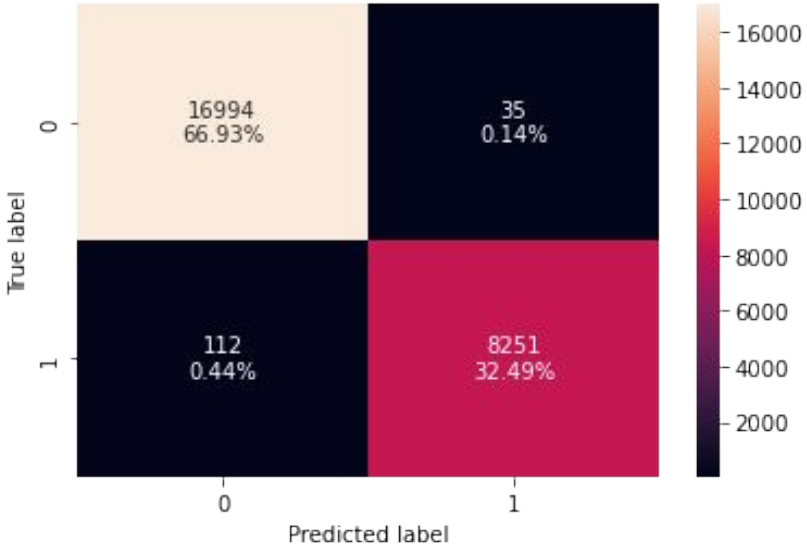
Since there are no missing values, we do not have to treat the data for this.

Data Preparation for Modeling:

We encoded the categorical columns/features, then split the data into ‘training’ and ‘testing’ data in order to train and test the built model. We split the data into 70:30; 67% of observations belongs to class 0 and 32.9% observations belongs to class 1. The training set has 25,392 rows and 27 columns, while the test set has 10,883 rows and 27 columns.

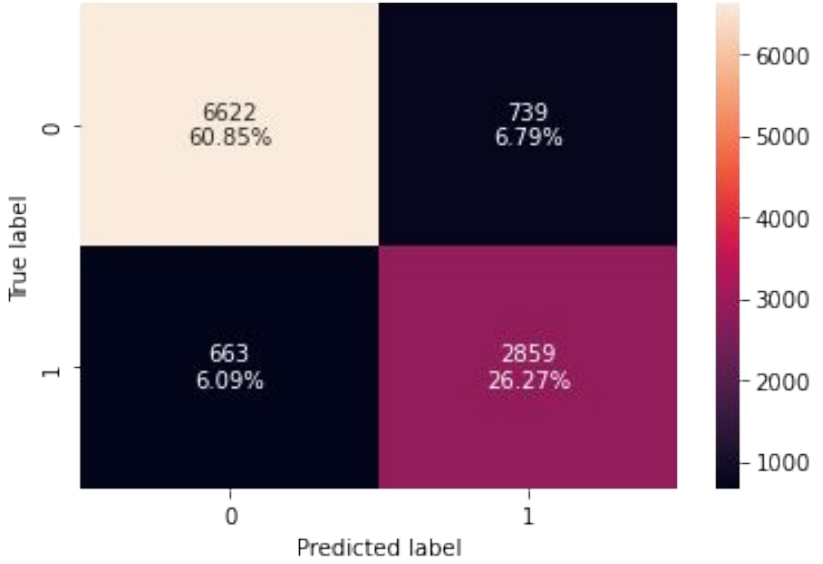
Confusion Matrix: No Pruning

TRAINING SET:



	Accuracy	Recall	Precision	F1
0	0.99421	0.98661	0.99578	0.99117

TEST SET:



	Accuracy	Recall	Precision	F1
0	0.87118	0.81175	0.79461	0.80309

Model Performance Evaluation

UN-PRUNED DECISION TREE:

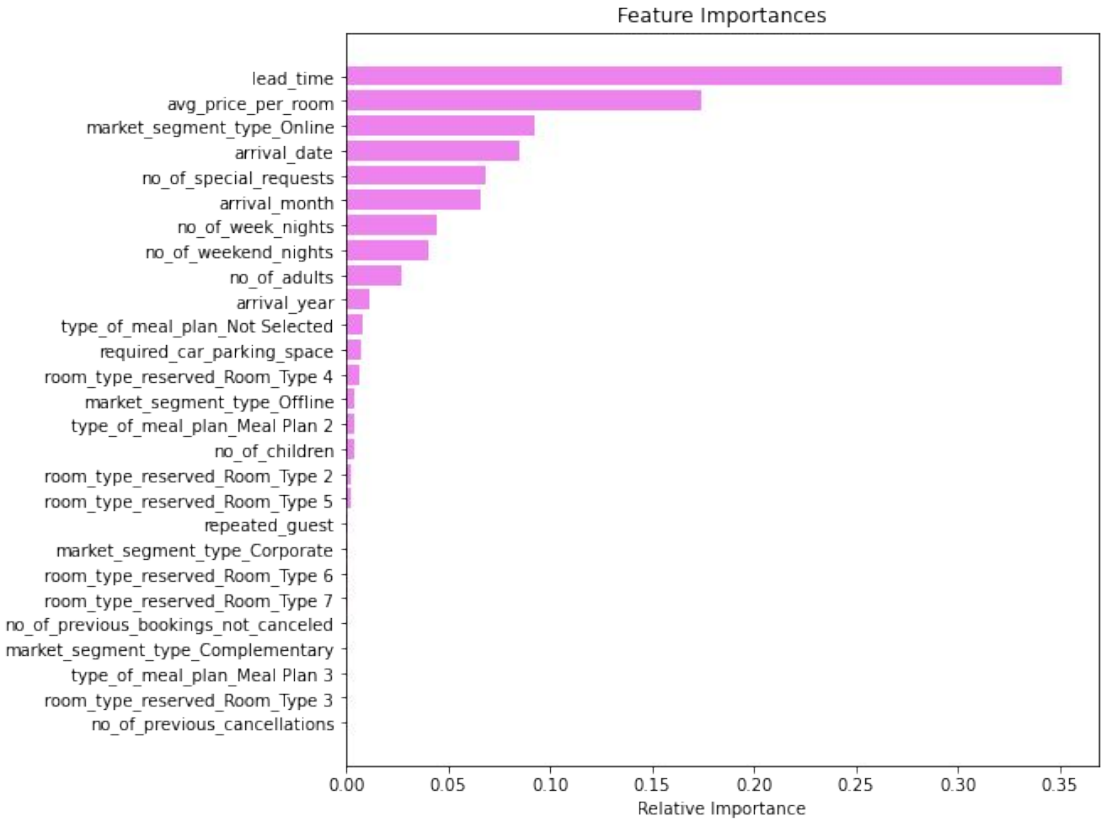
This model is able to almost perfectly classify all the data points on the training set. Less than 1% errors on the training set means that more than 99% of samples have been classified correctly.

Decision trees, without restrictions, will continue to grow until all data points are correctly classified and the trees will learn all the patterns in the training set.

The training set has performed excellently, while the test set has performed poorly when compared to the former. This model performance check shows us that there is overfitting in this model.

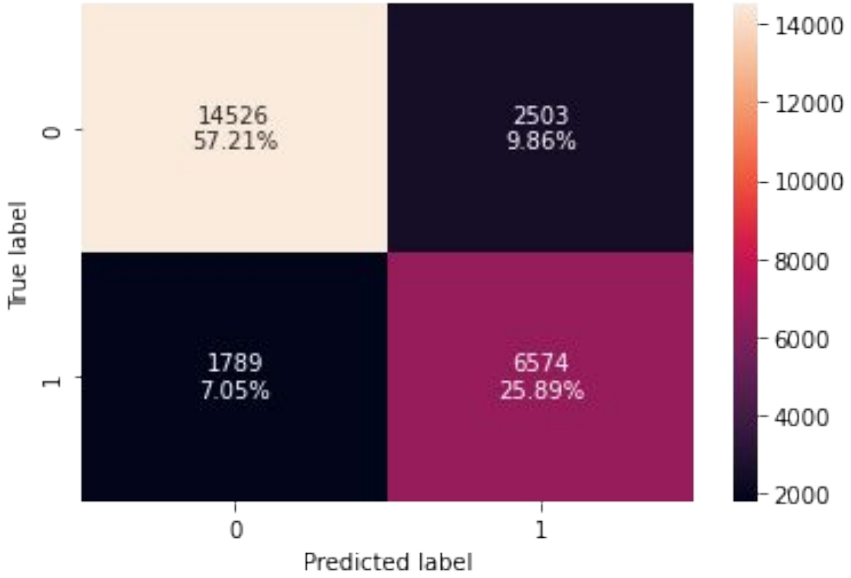
Feature Importance Evaluation

The most important features in our pre-tuned decision tree are; lead time and average price per room.



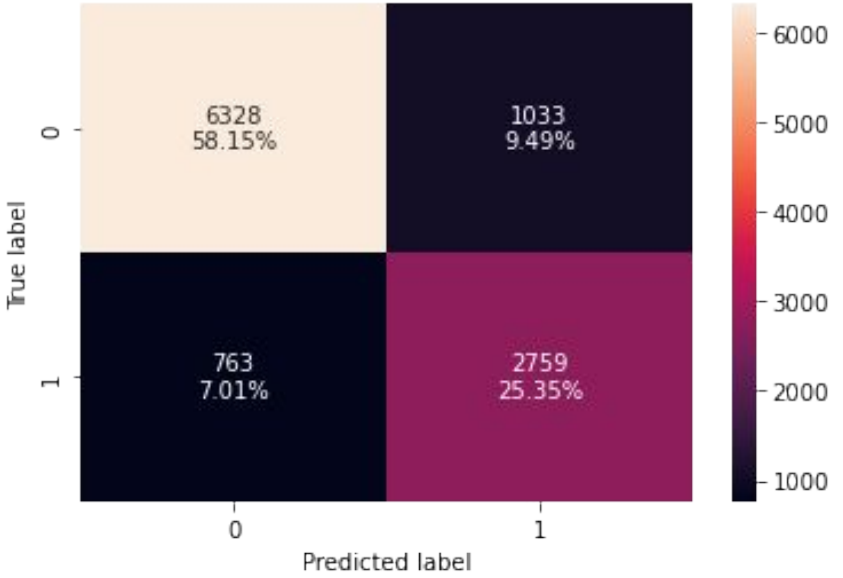
Confusion Matrix: Pre-Pruned

TRAINING SET:



	Accuracy	Recall	Precision	F1
0	0.83097	0.78608	0.72425	0.75390

TEST SET:



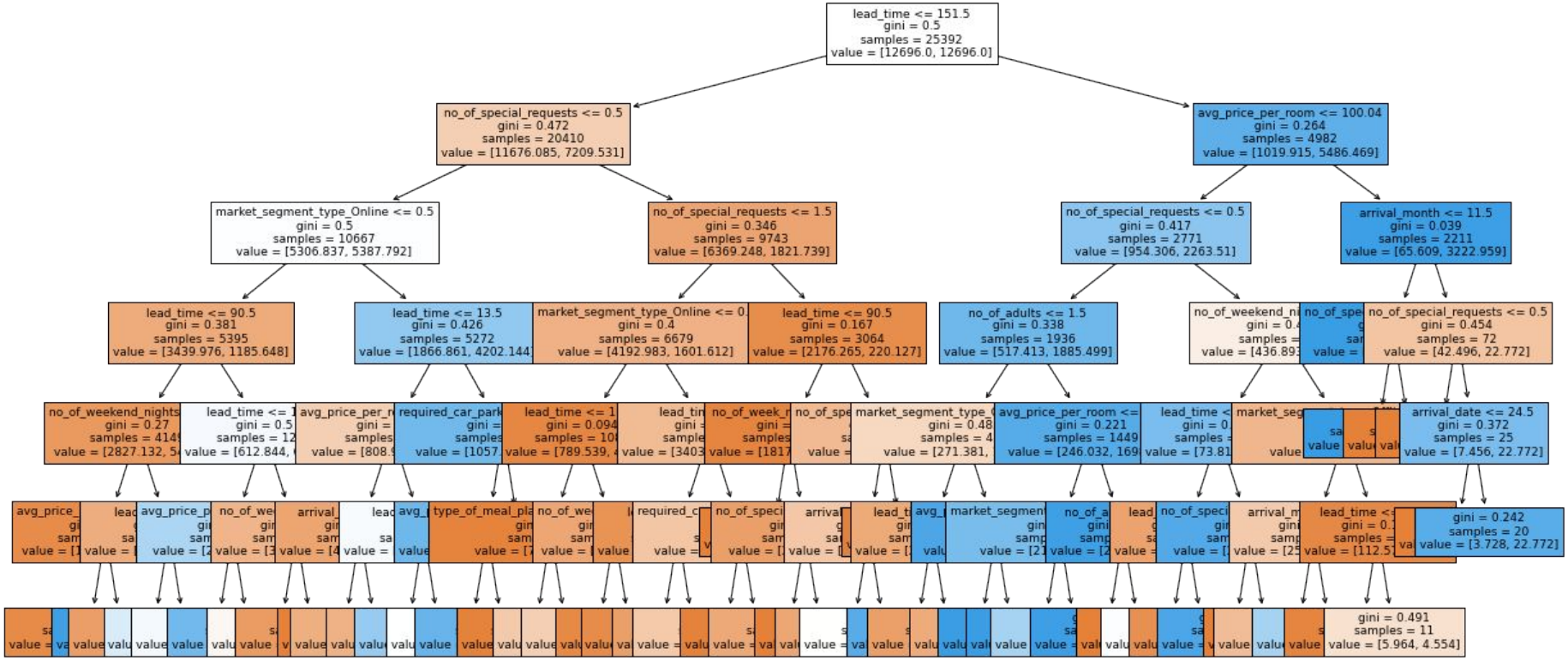
	Accuracy	Recall	Precision	F1
0	0.83497	0.78336	0.72758	0.75444

Model Performance Evaluation

PRE-PRUNED DECISION TREE:

We pre-pruned the decision tree using all hyperparameters. The model is now giving a generalized result and shows that the current model is able to generalize better on unseen data than the un-tuned model. Recall scores on both the train and test data are around 0.78, while f1 score on both are around 0.75.

Decision Tree Visualization: Pre-Pruned



Text Report: Rules of the Decision Tree

Using these extracted decision rules, we can make interpretations from the decision tree model like:

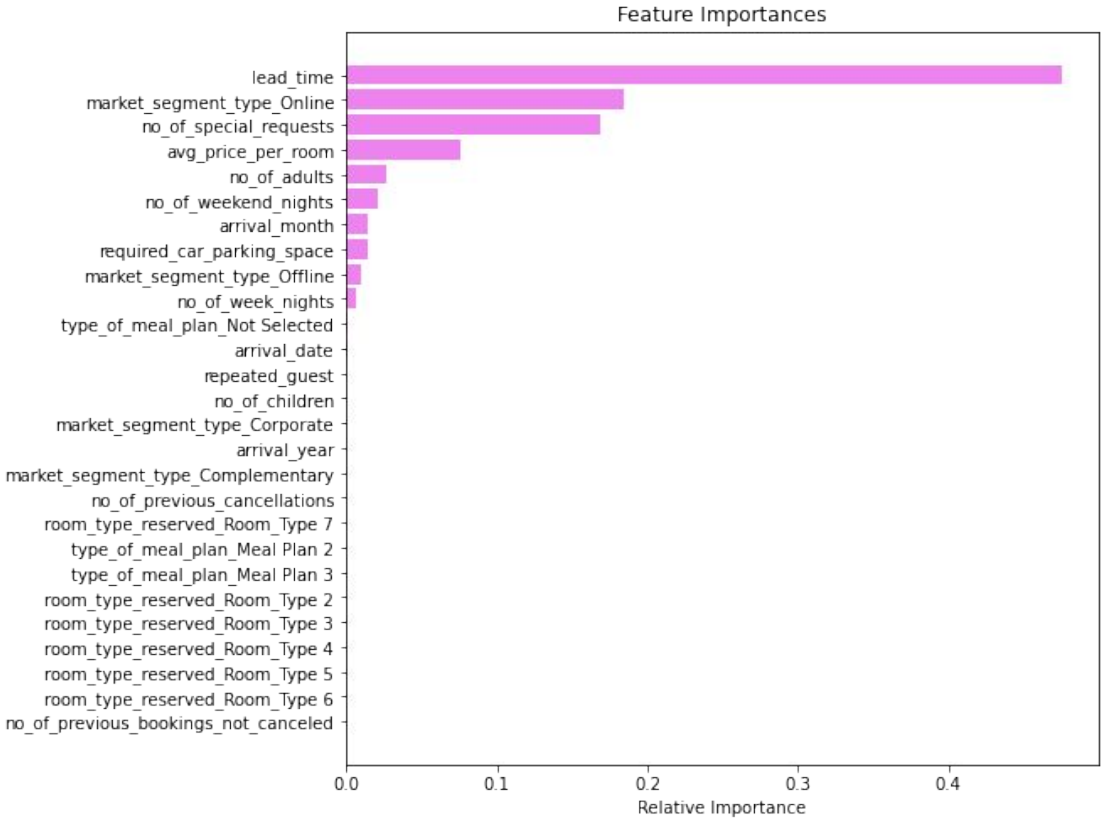
If the lead time is less than or equal to 151.50, the no. of special requests is less than or equal to 0.50, the market segment type (online) is less than or equal to 0.50, the lead time is less than or equal to 90.50, the no. of weekend nights is less than or equal to 0.50, and the average price per room is less than or equal to 196.50, then the guest will be classed 0 (not likely to cancel their reservation).

Interpretations from other decision rules can be made similarly.

```
|--- lead_time <= 151.50
|   |--- no_of_special_requests <= 0.50
|   |   |--- market_segment_type_Online <= 0.50
|   |   |   |--- lead_time <= 90.50
|   |   |   |   |--- no_of_weekend_nights <= 0.50
|   |   |   |   |   |--- avg_price_per_room <= 196.50
|   |   |   |   |   |   |--- weights: [1736.39, 133.59] class: 0
|   |   |   |   |   |--- avg_price_per_room >  196.50
|   |   |   |   |   |   |--- weights: [0.75, 24.29] class: 1
|   |   |   |   |--- no_of_weekend_nights >  0.50
|   |   |   |   |   |--- lead_time <= 68.50
|   |   |   |   |   |   |--- weights: [960.27, 223.16] class: 0
|   |   |   |   |   |--- lead_time >  68.50
|   |   |   |   |   |   |--- weights: [129.73, 160.92] class: 1
|   |   |   |--- lead_time >  90.50
|   |   |   |   |--- lead_time <= 117.50
|   |   |   |   |   |--- avg_price_per_room <= 93.58
|   |   |   |   |   |   |--- weights: [214.72, 227.72] class: 1
|   |   |   |   |   |--- avg_price_per_room >  93.58
|   |   |   |   |   |   |--- weights: [82.76, 285.41] class: 1
|   |   |   |   |--- lead_time >  117.50
|   |   |   |   |   |--- no_of_week_nights <= 1.50
|   |   |   |   |   |   |--- weights: [87.23, 81.98] class: 0
|   |   |   |   |   |--- no_of_week_nights >  1.50
|   |   |   |   |   |   |--- weights: [228.14, 48.58] class: 0
```

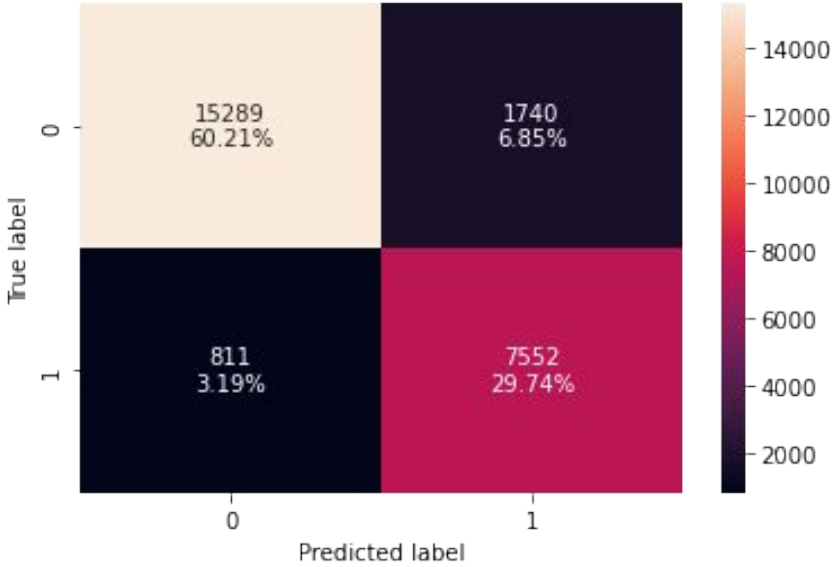
Feature Importance Evaluation

The most important feature in our pre-pruned decision tree is lead time.



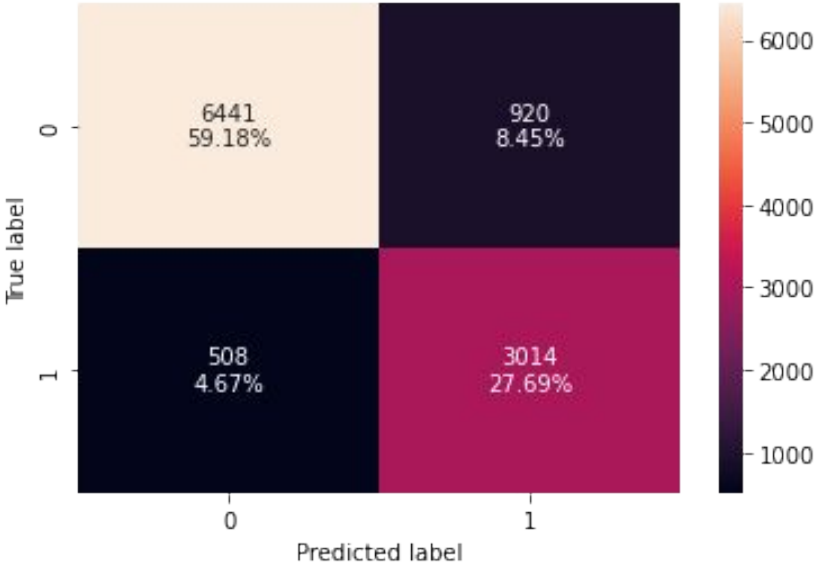
Confusion Matrix: Post-Pruned

TRAINING SET:



	Accuracy	Recall	Precision	F1
0	0.89954	0.90303	0.81274	0.85551

TEST SET:



	Accuracy	Recall	Precision	F1
0	0.86879	0.85576	0.76614	0.80848

Model Performance Evaluation

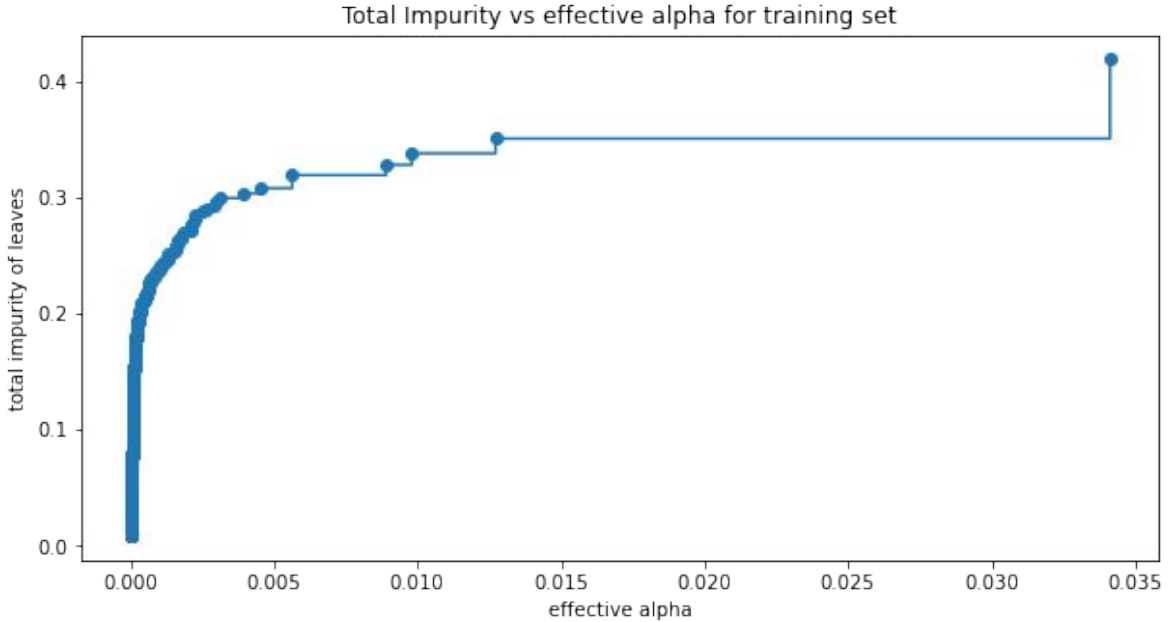
POST-PRUNED DECISION TREE:

There are more discrepancies between the training and test data sets in the post-pruned decision tree than was evident in the pre-pruned trees. In the pre-pruned decision tree, we saw that the test set performed better than the training set. Whereas, in the post-pruned tree, the training set performed better than the test set. However, in the post-pruned trees the overall results and scores were greater than in the pre-pruned tree.

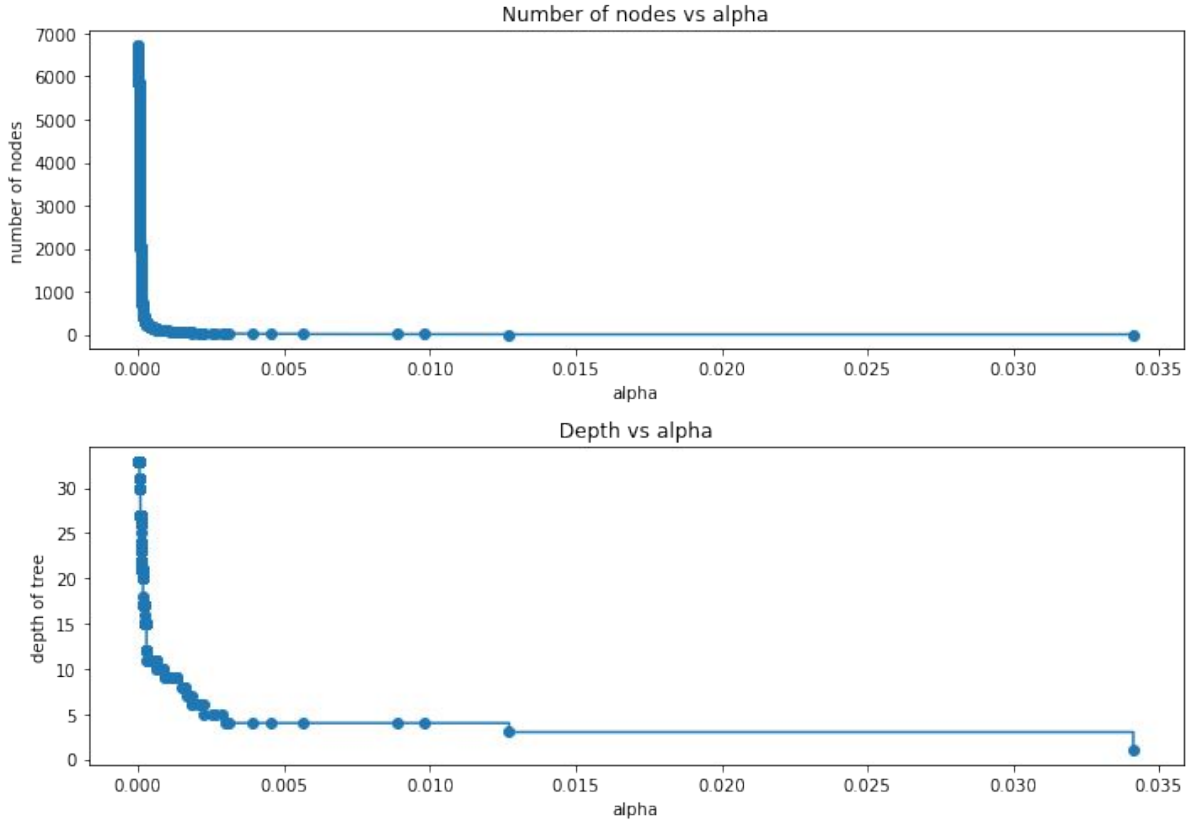
Training the Decision Tree using Effective Alphas

Next, we train a 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. The number of nodes and tree depth decreases as alpha increases, the total impurity of leaves also increases as alpha increases.

The next page shows visualizations for number of nodes and depth of tree.

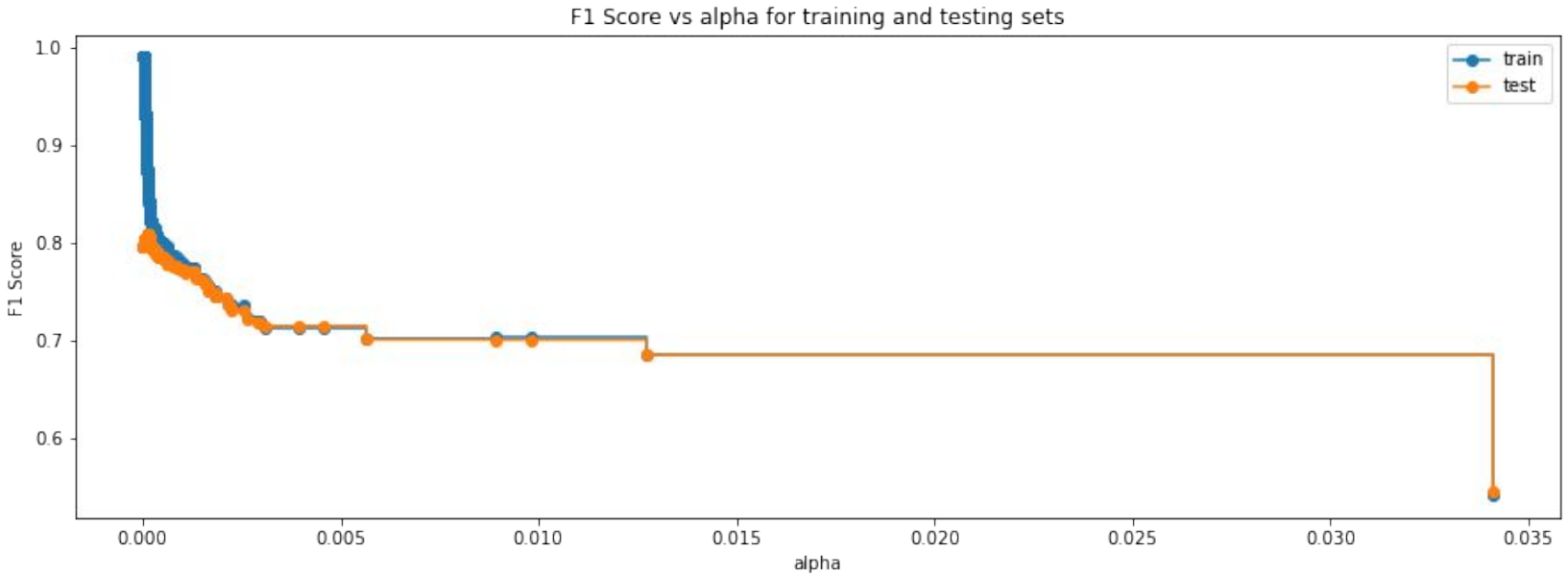


Visualization: Alphas, Nodes and Tree Depth

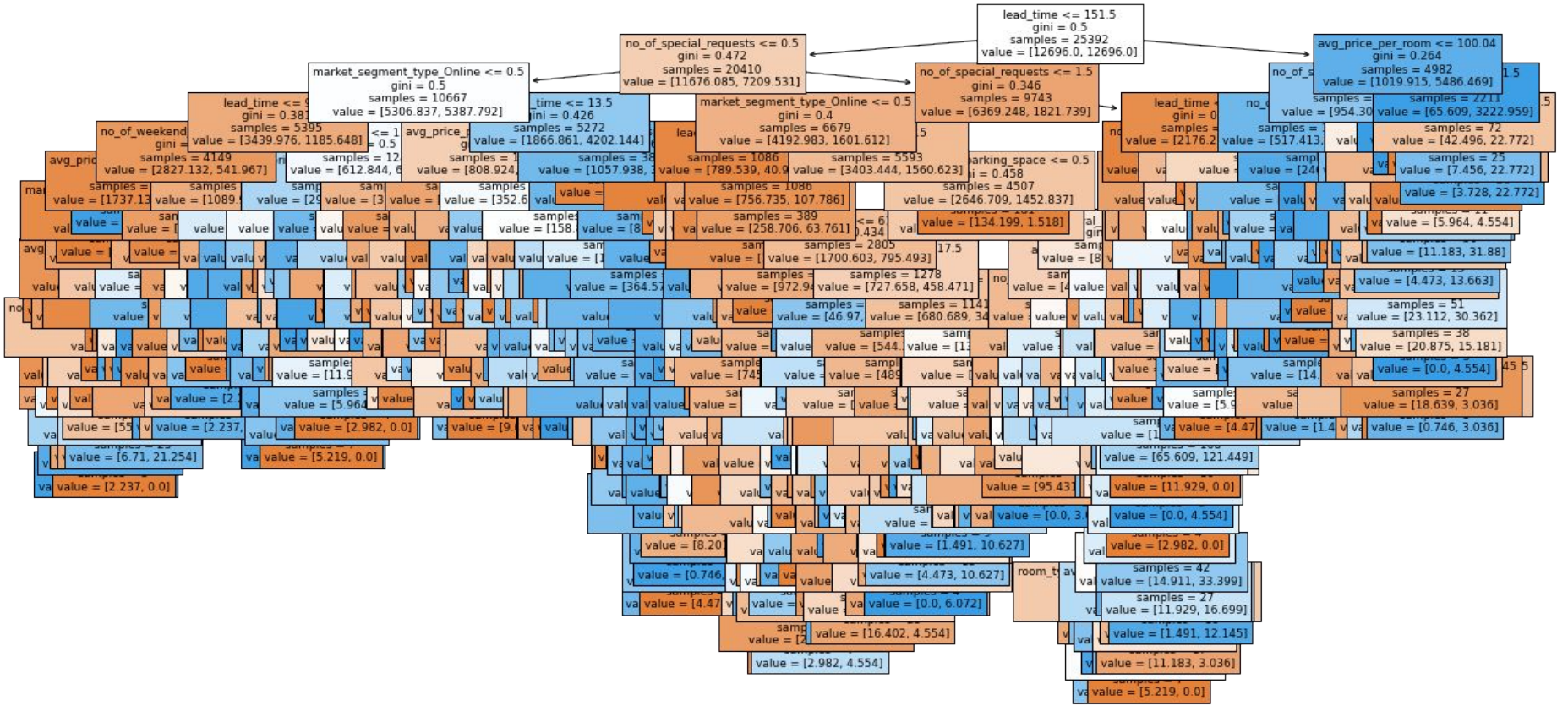


F1 Score vs Alpha for Training and Testing Sets

Both F1 score and alpha follow a similar pattern in both training and test sets. As alpha increases, the f1 score gradually declines.

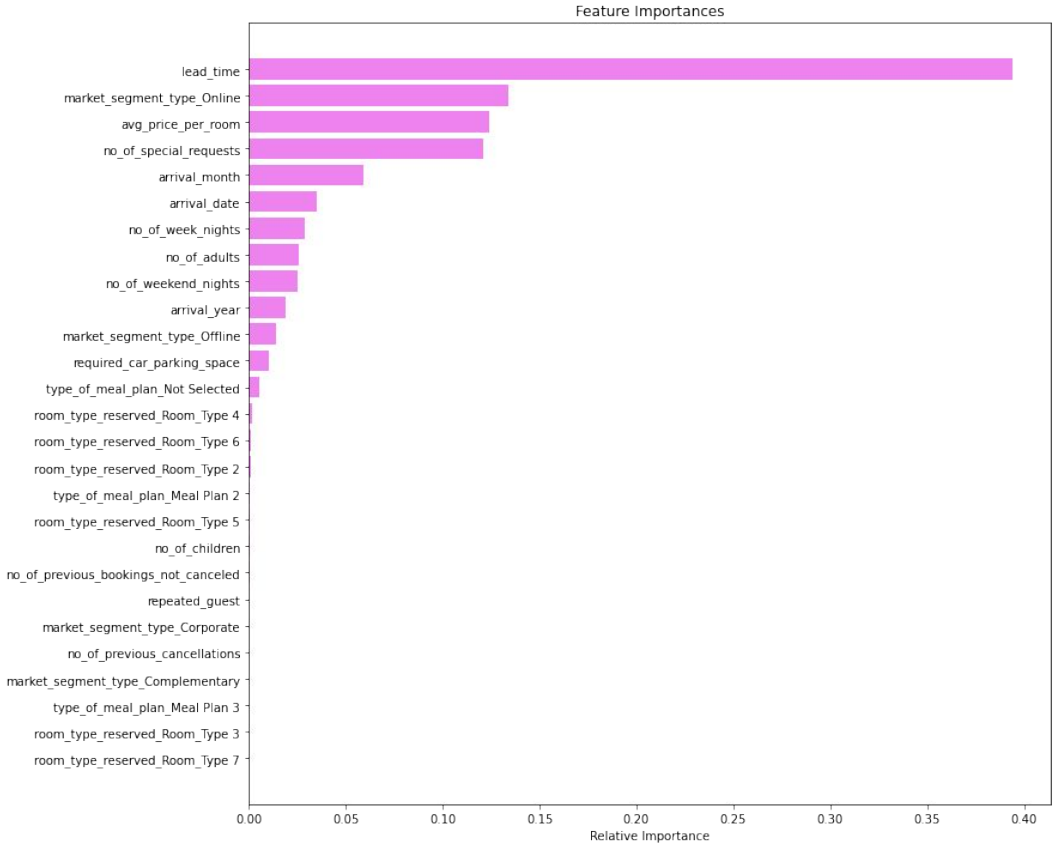


Decision Tree Visualization: Post-Pruned



Feature Importance Evaluation

The most important features in our post-pruned tree match that of our pre-pruned decision tree. Lead time is still the most important feature to consider.



Model Performance Summary: Decision Tree

TRAINING SET:

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

TEST SET:

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

Conclusion: Since the post-pruned decision tree has overall better results and a greater f1 score than the pre-pruned tree, we are satisfied with our post-pruned model and will select this model.

THE END :)