INN HOTELS

Supervised Learning - Classification,
PGP - Data Science and Business Analytics

Arsalaan B. Saiyed

Date – 4th June 2022

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- Data Background and Contents
- EDA Results
- Data Preprocessing
- Model Performance Summary
  - Regression Model
  - Classification Model
- Appendices
  - Appendix-A: EDA Results
  - Appendix-B: Model Performance Summary

Executive Summary

- From exploratory analysis and relative importance we can ***conclude*** the likelihood of a booking cancellation depends primarily on the following parameters:
  - Lead time: There is a strong correlation between lead time and cancellation of booking. Higher the lead time the greater the likelihood of cancellation.
  - Average room price per day: Higher the room price greater the likelihood of booking cancellation.
  - Segment: From various market segments it was observed that the bookings made online were the most likely to get cancelled. Followed by offline bookings, corporate bookings and no cancellations for complementary bookings.
  - Speacial Requests: Greater the number of special requests the lesser is the likelihood of the booking being cancelled.
  - Arrival Month: We observed that the months April to September have the most likelihood of cancellation.
  - Week/Weekend Nights: It is also observed that if the numeber of weekend or weeknights increases the likelihood of cancellation increases. This means the longer the stay the more likely that it will be cancelled.
  - Arrival Month: The busiest months are usually June, July and August according and according to our data the most busiest month for 2017 and 2018 was October. October is also the month with the most number of cancellations and on average over the two years June has the highest cancellations followed by July and August.

Executive Summary

- ***Recommendations***: Based on the insights from EDA and building classification models to predict the booking cancellation we would like to make the following recommendations:

I. For bookings with long lead time:
   I. Offer some amount of discount for early payment and incase of last minute cancellation deduct amount for one day booking to compensate for the INNs loss of resources and to arrange another booking.
   II. Send monthly and then weekly confirmation/reminder emails close to the arrival date. If the customer is responsive then we can confirm that they are serious and will be staying with the INN. If the customer is unresponsive we can call to confirm or cancel their booking.
   III. Make sure you the hotels have a solid and reasonable cancellation policy in place.

II. Average room price: The INN Hotels must stratergize their prices and constantly maintains a competitive rate. A in depth price comparision should be conducted between INN hotels and other similar hotels to determine the best price per room depending on the following factors.
   I. Arrival month: Depending on the arrival month and its respective busyness the hotels must price their rooms appropriately and competitvely.
   II. Repeated guests: Target various coupons/offers for regular guests. Data also shows repeated guests are least likely to cancel bookings.
   III. New guests: Similary target various coupons/offers for first time guests to build popularity and gain trust through great service.
   IV. Families: Make your hotels to appear more family friendly by including ads and coupons or offers for family bookings. Data shows that guests with children i.e. families bring in more revenue so, it is good to have an image as a family friendly business.

III. Drive direct bookings rather than relying on online travel agents. As seen online bookings have the highest cancellation rate. With direct bookings there is a personalized connection and the hotel is better able to push rebooking or vouchers as an alternative to cancellation.

Executive Summary

- ***Recommendations***:

IV. Drive direct bookings rather than relying on online travel agents. As seen online bookings have the highest cancellation rate. With direct bookings there is a personalized connection and the hotel is better able to push rebooking or vouchers as an alternative to cancellation.

V. Upsell value-added packages or amenities during booking. Those who upgrade will have more invested with your hotel and therefore are more committed to their booking.

VI. Pre-arrival emails don't only have to be for reminders but can also be created to build excitement and anticipation for upcoming stays. Remind guests of their selected room type and ammenities they are sure to enjoy.

VII. Overbook. Overbooking is a tricky maneuver but it can work with contingency plan in place. The idea is to overbook more rooms then you have, in anticipation of cancellations or no shows. To this successfully we must analyze the reservation trends. Based on our data cancellation is most seen in the busiest months so it makes to overbook then. Also, you should have a backup plan, just in case. Shifting a guest to a similar property at the time of check-in could be one way to handle the situation.

VIII. Use Length of Stay Restrictions Checkout your competitors' policies on length of stay restrictions. Some hotels do not allow one night bookings on busy days, or when an event is happening nearby. Longer bookings mean less booking count, which reduces the number of cancellations. Our data also shows the same, the longer the stay the more the cancellations.

IX. Be Responsive and Proactive. Keep a track of cancellations. Your front desk needs to be proactive and responsive. When guests don't turn up, get in touch as soon as possible. Information is power and gives you the chance to resell the room.

Business Problem Overview and Solution Approach

- ***Problem Overview:***

  A significant number of hotel bookings are called off due to cancellations or no-shows. The cancellation of bookings impact a hotel on various fronts:

  - Loss of resources (revenue) when the hotel cannot resell the room.
  - Additional costs of distribution channels by increasing commissions or paying for publicity to help sell these rooms.
  - Lowering prices last minute, so the hotel can resell a room, resulting in reducing the profit margin.
  - Human resources to make arrangements for the guests.

  The objective is to analyze the data provided to find which factors have a high influence on booking cancellations, build a predictive model that can predict which booking is going to be canceled in advance, and help in formulating profitable policies for cancellations and refunds.

- ***Solution Approach:***

  - Data Overview & EDA
  - Data pre-processing
  - Regression Model Building & Testing
  - Decision Tree Model Building
  - Pruning & Model Comparison
  - Executive Summary - Actionable insights and summary

Data Background and Contents

- ***Observations***:

  I. Data was formatted appropriately.

  II. There is no missing or duplicated data

  III. This data is for the years 2017 & 2018

Observations	Variables
36,275	19

Table 1: Shape of the data

Variable	Description
Booking_ID	unique identifier of each booking
no_of_adults	Number of adults
no_of_children	Number of Children
no_of_weekend_nights	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
no_of_week_nights	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
type_of_meal_plan	Type of meal plan booked by the customer Not Selected – No meal plan selected Meal Plan 1 – Breakfast Meal Plan 2 – Half board (breakfast and one other meal) Meal Plan 3 – Full board (breakfast, lunch, and dinner)
required_car_parking_space	Does the customer require a car parking space? (0 - No, 1- Yes)
room_type_reserved	Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels.
no_of_special_requests	Total number of special requests made by the customer (e.g. high floor, view from the room, etc)
booking_status	Flag indicating if the booking was canceled or not.

Variable	Description
lead_time	Number of days between the date of booking and the arrival date
arrival_year	Year of arrival date
arrival_month	Month of arrival date
arrival_date	Date of the month
market_segment_type	Market segment designation.
repeated_guest	Is the customer a repeated guest? (0 - No, 1- Yes)
no_of_previous_cancellations	Number of previous bookings that were canceled by the customer prior to the current booking
no_of_previous_bookings_not_canceled	Number of previous bookings not canceled by the customer prior to the current booking
avg_price_per_room	Average price per day of the reservation; prices of the rooms are dynamic. (in euros)

Table 2: Details

EDA Results – Univariate Analysis

Upon conducting a univariate analysis of the raw data and computing the statistical summary we observed the following:

I. The average lead time is found to be 85 days and the median is 57 days. We also notice that longer the lead time the more likely the customer would cancel the booking.

II. Average price per room per day is found to be at a mean of 103 and a median of 99 euros.

III. Majority (72.1%) of the bookings are for two adults and very few of them (7-8%) having bookings with children.

IV. 77.5 % of the customers are given Room Type 1. 76.7% prefer meal type 1 only 3.1 % of customers require parking,

V. 64% bookings were online booking, 29% offline, 5.6% corporate, 1.1% complementary and 0.3% aviation.

VI. Almost half (54.5%) of the bookings have no special requests but the remaining do have 1 special request at least.

VII. 67.2% of all the bookings in the data were not cancelled while 32.8% were.

VIII. Only 2.6% of the guests are repeated customers

Note: The mean, min, max etc. for all columns are given in the statistical summary in a table in Appendix-A

Link to Appendix slide on data background check

EDA Results – Bivariate Analysis

Upon conducting a bivariate analysis of the raw data we observed the following:

I. From the heatmap we can see that lead time has a strong positive correlation with booking status. Average price also seems to have a slight positive correlation with booking status.

II. Across market segments average prices are highest for online bookings and lowest for complementary bookings. We notice a similar pattern for number of cancellations across market segments as it is highest for online bookings and lowest for complementary bookings.

III. The more special requests a customer has the lesser they are likely to cancel the booking but, we it is also seen that special requests come at a price as the average price per room increases with number of requests.

IV. The higher the room prices are the more likely is the customer to cancel the bookings.

V. Percentage of cancellation does not significantly differ depending on the number of family members per booking.

VI. Longer the booking is for the more likely is the customer to cancel the booking.

VII. Repeated guests are less likely to cancel bookings.

VIII. Based on total guests the busiest months seem to be August, September and October but, based on mean June, July and August seem to be the busiest month.

Link to Appendix slide on data background check

Data Preprocessing

- Duplicate value check
- Missing value treatment
- Outlier check (treatment if needed)
- Feature engineering
- Data preparation for modeling

***Note**: You can use more than one slide if needed*

Data Preprocessing

Let's take a quick look at a small sample of our data. This will help us get some idea of the attributes of the data and also help us understand the degree of cleaning needed before we can build a model.

Random sample of the data (5 rows, 19 columns):

	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
0	INN00001	2	0	1	2	Meal Plan 1	0	Room_Type 1
1	INN00002	2	0	2	3	Not Selected	0	Room_Type 1
2	INN00003	1	0	2	1	Meal Plan 1	0	Room_Type 1
3	INN00004	2	0	0	2	Meal Plan 1	0	Room_Type 1
4	INN00005	2	0	1	1	Not Selected	0	Room_Type 1

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

Data Preprocessing

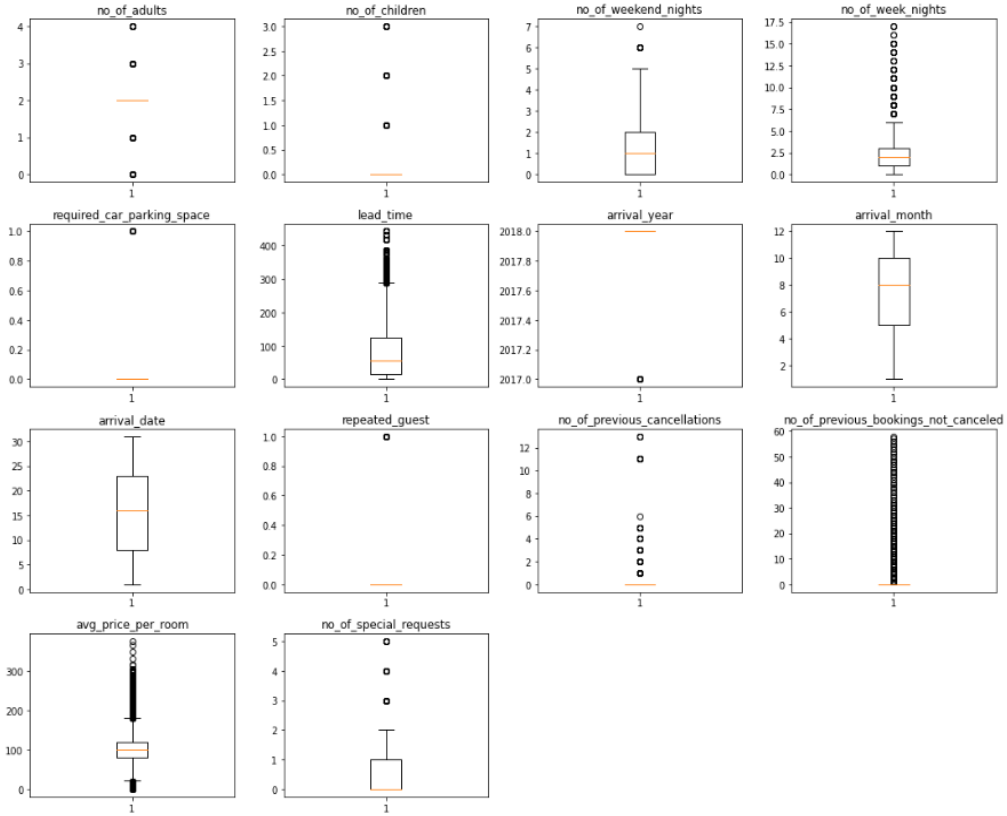
- Duplicate Value Check: No duplicate entries were found in the given data
- Missing Value Treatment:

  Our data has no missing value so no treatment was required.
- Outlier Check:

  We treated the outliers for average price per room for amounts that were greater than 500 euros because this was not a common price. We replaced all the values over 500 euros with the upper whisker value using the inter quartile range which came out to be 179.55 euros.

  All the other outliers are correct data so we don't treat them.
- Feature Engineering:

  There was no need for feature engineering.



Model Performance Summary

- Overview of ML model and its parameters
- Summary of most important factors used by the logistic regression model for classification (prediction)
- Summary of key performance metrics for training and test data in tabular format for comparison of logistic regression model
- Summary of most important factors used by the decision tree model for classification (prediction)
- Summary of key performance metrics for training and test data in tabular format for comparison decision tree model

Link to Appendix slide on model assumptions

Model Performance Summary

- Overview of ML model:

  Our was to provide INN Hotels insights as to what are the key indicators that show us that if a customer would cancel the booking or not, based on which they will organize their resources and optimize their profits by having maximum occupancy in their hotels and minimize their losses on cancelled bookings and lost resources. For this we used past data to build classification models using logistic regression and a decision tree to predict that if a new customer is likely to cancel their booking or not**.**

- ML model parameters:

  - **Model Selection & Fitting**: We build two models for classification and compare the results.
    a. Logistic Regression Model: Statsmodel was used in python to build a model that fit the train data set and tested against the test data set. Then, the model was treated to meet all the linear regression assumptions.
    b. Decision Tree: We will build another model using the DecisionTreeClassifier function. Using default 'gini' criteria to split.
  - **Model Evaluation Criterion**: Model can make wrong predictions as:
    a. Predicting a customer will not cancel their booking but in reality, the customer will cancel their booking.
    b. 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? Hotels would want F1 Score to be maximized, greater the F1 score higher are the chances of minimizing False Negatives and False Positives.

Model Performance Summary

- Summary of most important factors used by the **LINEAR REGRESSION** model for prediction:
  I. Pre-processed data was split into training and testing data with a 70:30 ratio.
  II. Using logit from statsmodel library we fit a logistic model to the training data set. F1 was found to be 0.68213 for our *first model*.
  III. No multicollinearity was detected when VIF was calculated.
  IV. Predictor variables having a p-value greater than the selected significance level (> 0.05) were dropped.
  V. We then fit the logit function to the new training set and arrived at a F1 score of 0.68191 for our *final model*.
  VI. Converting coefficients to odds: The coefficients of the logistic regression model are in terms of log(odd), to find the odds we have to take the exponential of the coefficients.
    I. Therefore, odds = exp(b)
    II. The percentage change in odds is given as odds = (exp(b) - 1) * 100

	const	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
Odds	0.00000	1.16153	1.03682	0.19886	1.01601	1.53653	0.95191	0.04623	1.33523	1.01925	0.22638	1.18026	1.23366	0.68845
Change_odd%	-100.00000	16.15325	3.68173	-80.11371	1.60134	53.65324	-4.80895	-95.37690	33.52334	1.92531	-77.36171	18.02564	23.36580	-31.15506

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
0.76451	0.50213	0.47521	0.26776	0.41854	0.17012
-23.54886	-49.78672	-52.47870	-73.22410	-58.14571	-82.98809

  VII. We then calculate the optimal threshold from the *ROC-AUC curve* and using that build a new logistic *model*
  VIII. We also tried using the *precision-recall curve* to see if we can find a better thresholds.
  IX. We then compare all the models and select the best performing one (Follow hyperlinks for comments on different models)

Model Performance Summary

- Summary of key performance metrics for training and test data in tabular format for comparison of logistic regression model:

Training performance comparison:

	Logistic Regression-default Threshold (0.5)	Logistic Regression-optimal-auc-roc Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80663	0.78430	0.79218	0.80265
Recall	0.63265	0.76608	0.73470	0.69876
Precision	0.73950	0.64348	0.66565	0.69885
F1	0.68191	0.69945	0.69847	0.69880

Test set performance comparison:

	Logistic Regression-default Threshold (0.5)	Logistic Regression-optimal-auc-roc Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80272	0.77846	0.78489	0.79656
Recall	0.62619	0.76612	0.72799	0.69405
Precision	0.73285	0.63402	0.65440	0.68788
F1	0.67534	0.69384	0.68923	0.69095

Conclusion:

We can conclude that the threshold 0.42 has the highest accuracy and F1 score for the training and testing data set are slightly lower than the optimal threshold model but we need least errors so we pick the model with the best F1 score. Hence, it would be our best logistic regression model.

Model Performance Summary

- Summary of most important factors used by the **DECISION TREE** model for prediction:
  I. Pre-processed data was split into training and testing data with a 70:30 ratio.
  II. Using DecisionTreeClassifier library we fit a decision tree model to the training data set. F1 was found to be 0.99117 for our *first model*. Our model was overfitting so we resolved this.
  III. Using GridSearchCV we apply pre-pruning before building the decision tree. Using the *grid search parameters* we get *model* with a F1 score of 0.77326.
  IV. Then using the *cost complexity method* we calculate various alphas and their impurities. Using different alphas we plot F1 score against the alpha values for training and testing set and determine the most optimal alpha for the highest F1 scores for training and testing data set.
  V. Although the decision tree might look complex and hard to visualize, we leave it as is because this is the most optimal machine learning model.
  VI. We then compare all the models and select the best performing one. (Follow hyperlinks for comments on different models)

Model Performance Summary

- Summary of key performance metrics for training and test data in tabular format for comparison of decision tree model:

Training performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.99421	0.84905	0.89989
Recall	0.98661	0.78154	0.90303
Precision	0.99578	0.76516	0.81353
F1	0.99117	0.77326	0.85594

Test set performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.87118	0.84765	0.86888
Recall	0.81175	0.77541	0.85576
Precision	0.79461	0.75903	0.76634
F1	0.80309	0.76713	0.80858

Conclusion:

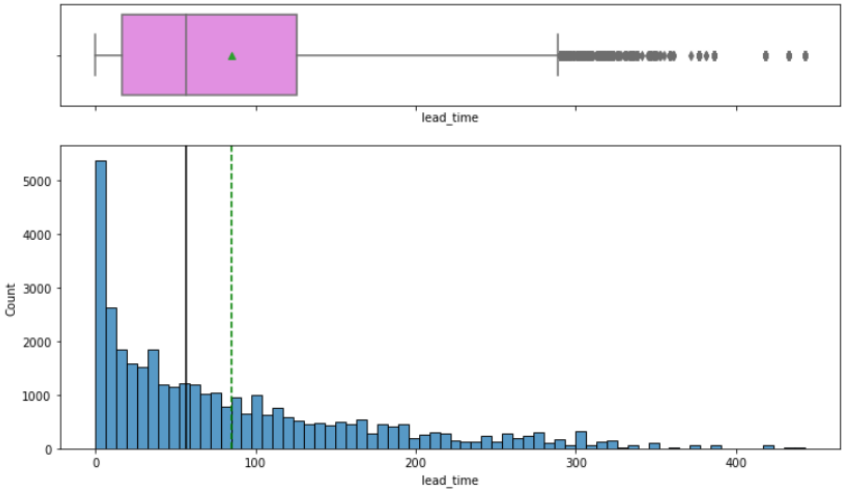
- Post-pruning model has a higher F1 and accuracy score for both training and testing data as compared to the pre-pruning model.
- The post pruning model seems to be more reliable as the performance for the training and testing data are relatively similar.
- The objective was minimize false positives and false negatives therefore, we will pick the model with the highest F1 score which is the post pruning model.

APPENDIX A – EDA Results (Univariate Analysis)

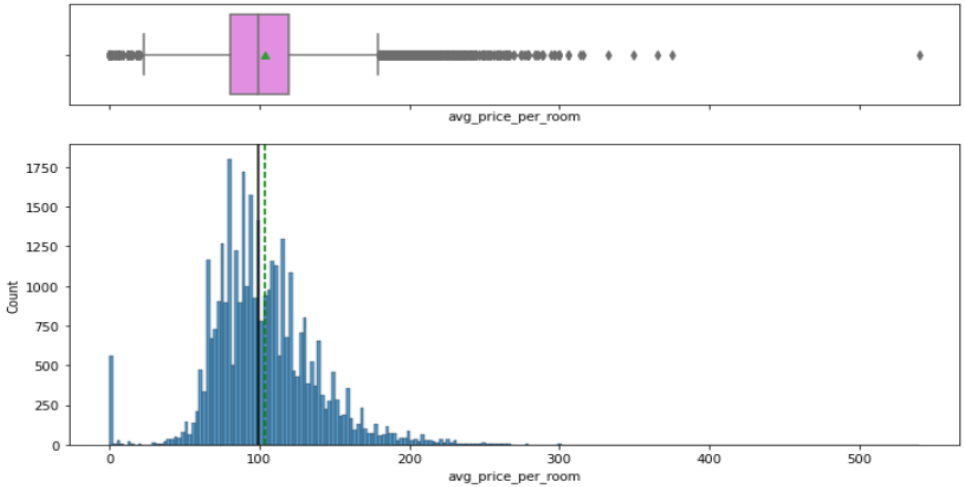
APPENDIX A: Statistical Summary

	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	0.10476	0.39466	0.00000	0.00000	0.00000	0.00000	3.00000
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	103.41360	35.01675	0.00000	80.30000	99.45000	120.00000	375.50000
no_of_special_requests	36275.00000	0.61966	0.78624	0.00000	0.00000	0.00000	1.00000	5.00000
booking_status	36275.00000	0.32764	0.46936	0.00000	0.00000	0.00000	1.00000	1.00000

APPENDIX A: Univariate Analysis

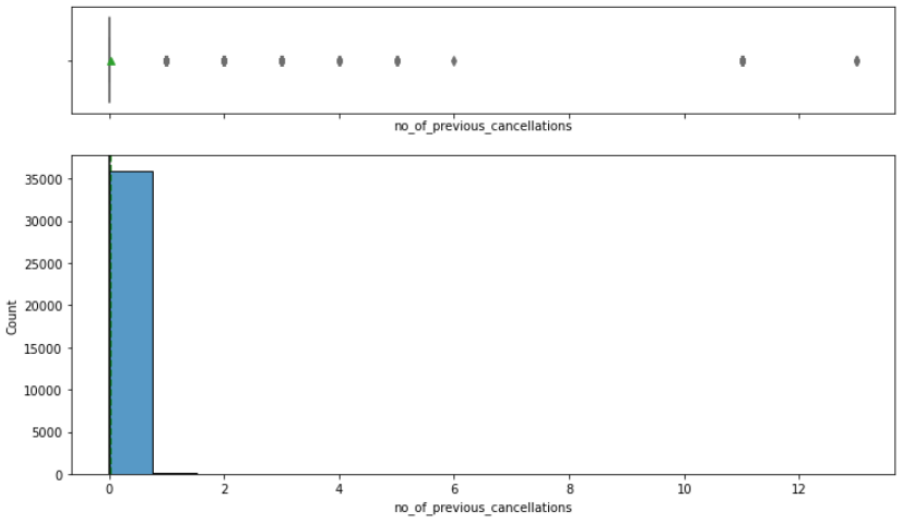


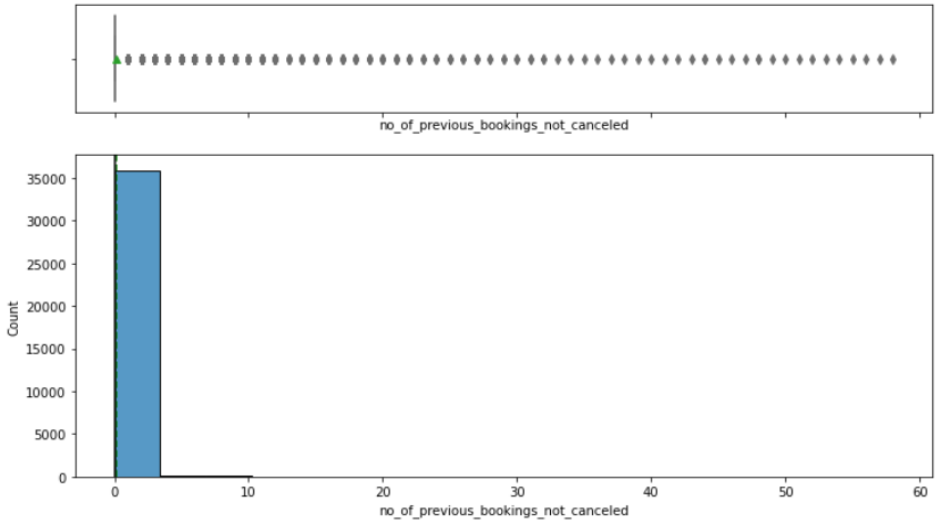
Mean: 85.23256 days
Median: 57 days
Strong Right-skew



Mean: 103.42354 Euros
Median: 99.45000 Euris
Light Right-skew

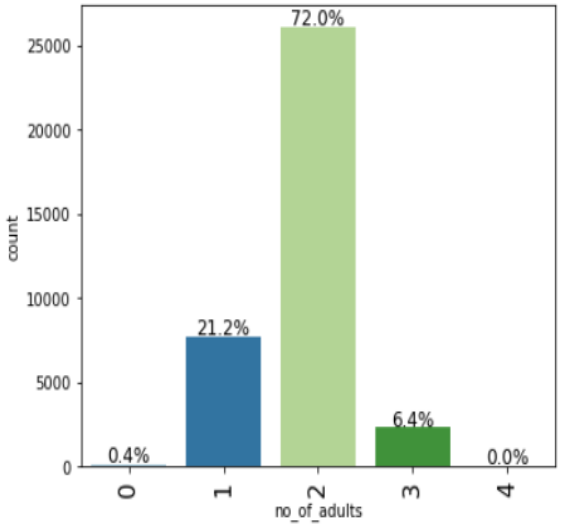
APPENDIX A: Univariate Analysis

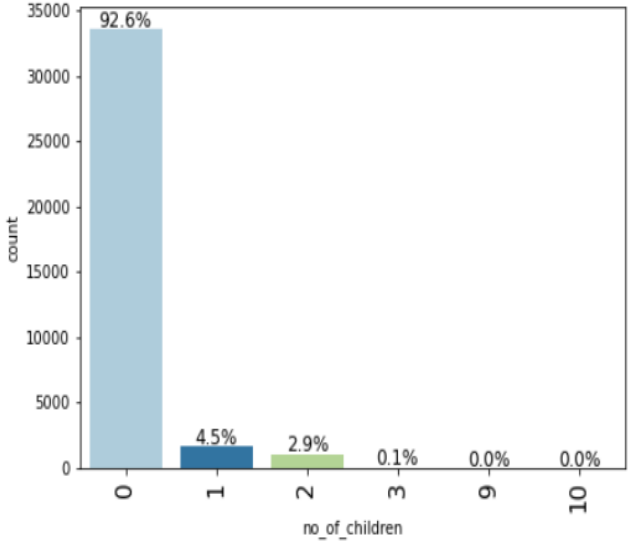


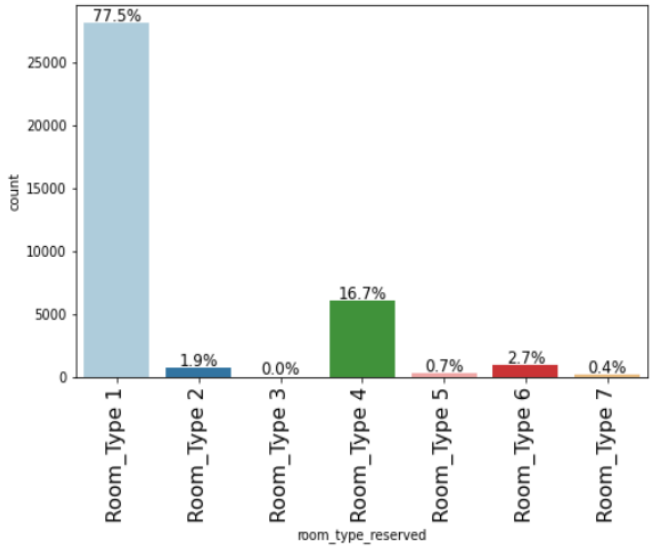


The two graphs show us that most customers have zero cancellations and a very small percentage has at least once cancellation. Most customers are not repeated customers because they don't have any value for both the parameters.

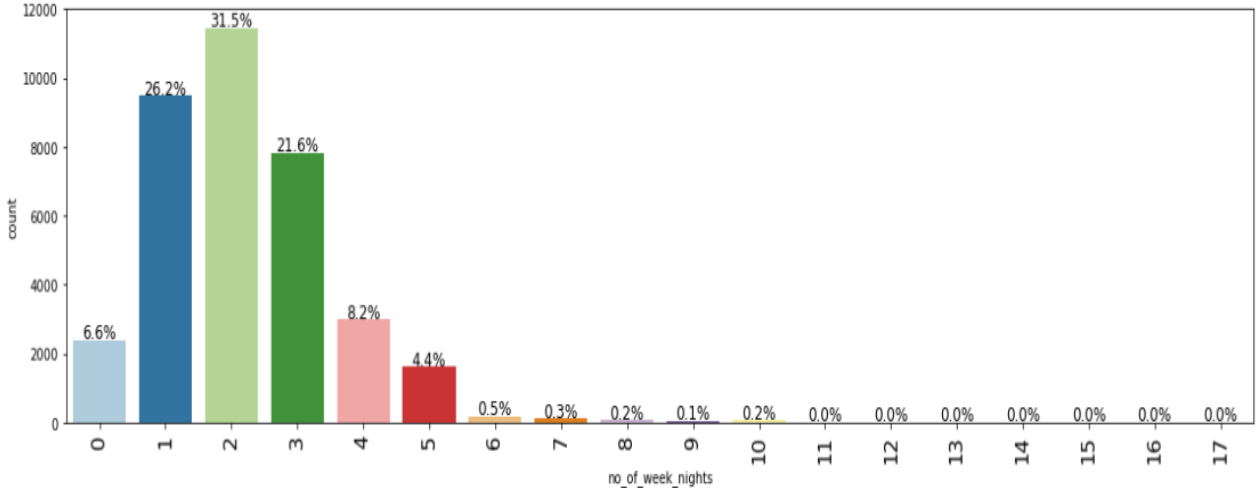
APPENDIX A: Univariate Analysis

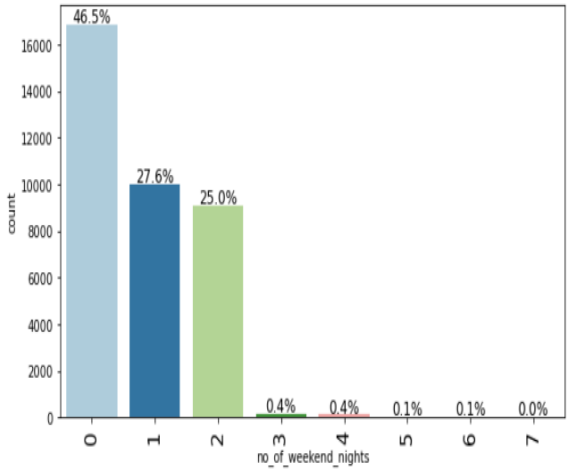






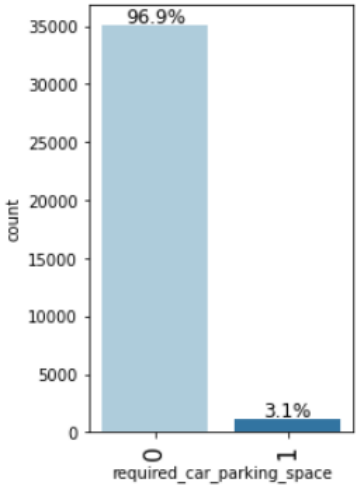
APPENDIX A: Univariate Analysis

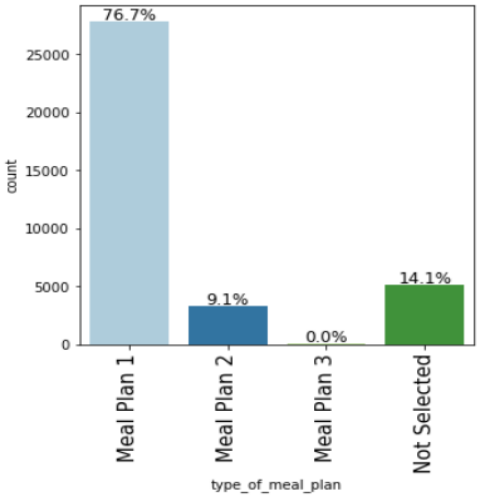


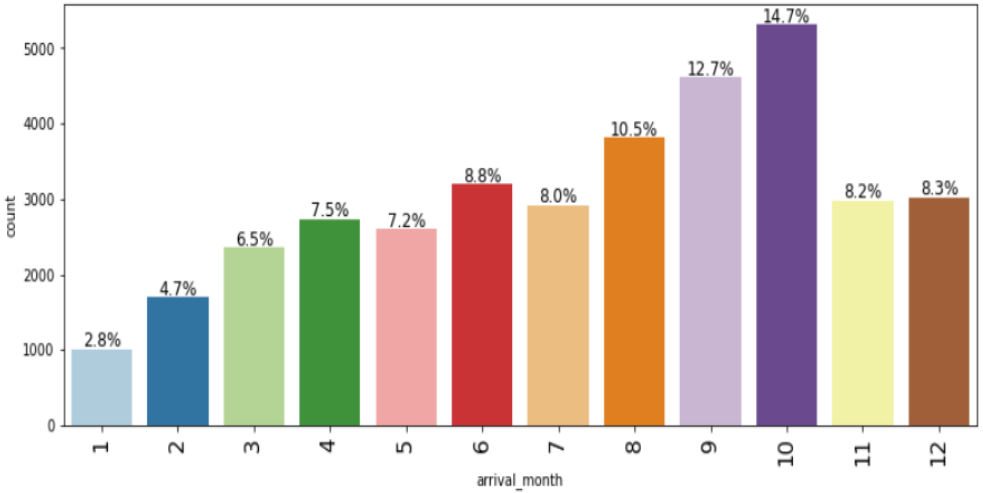


Most bookings are for short stays and most of them are over the weekend

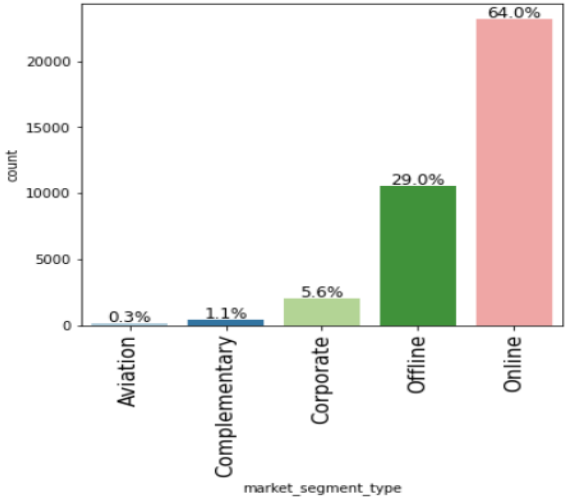
APPENDIX A: Univariate Analysis

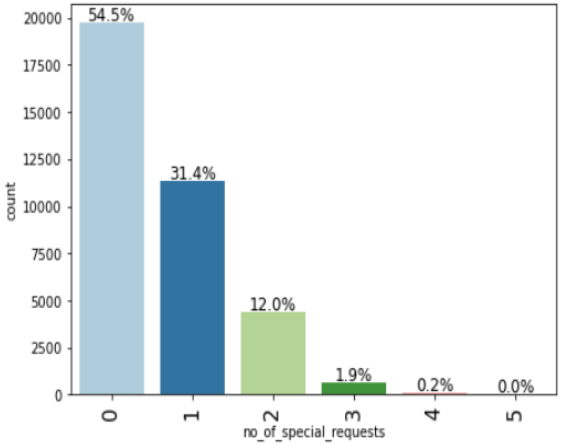


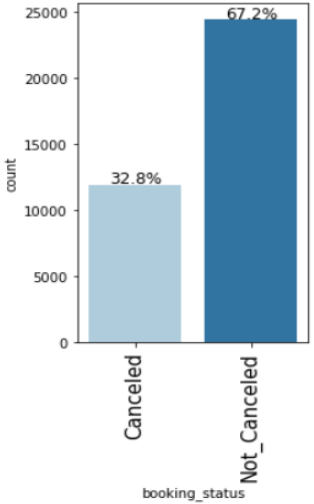


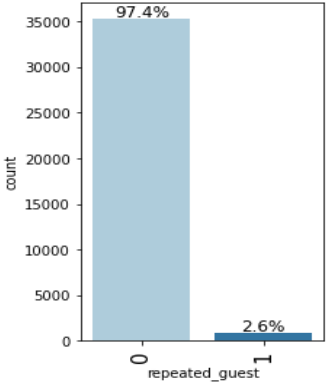


APPENDIX A: Univariate Analysis





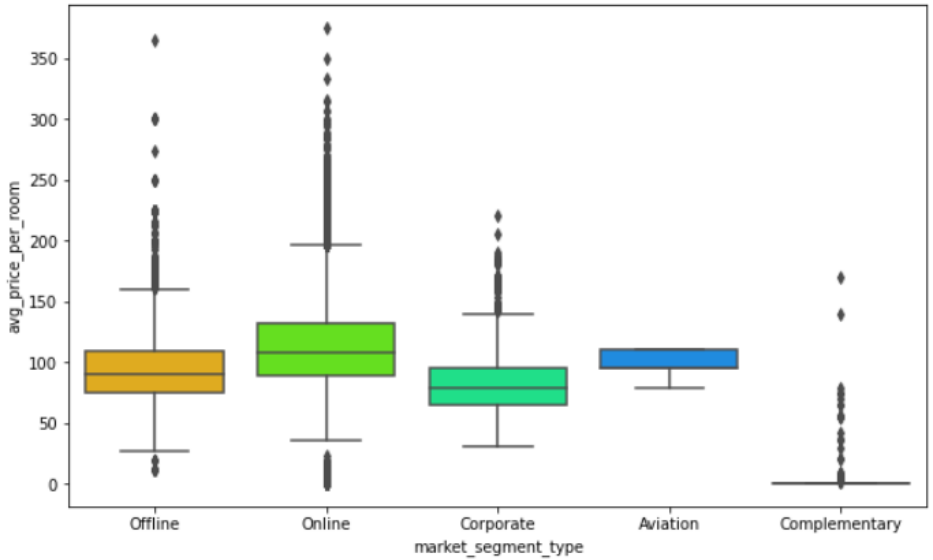




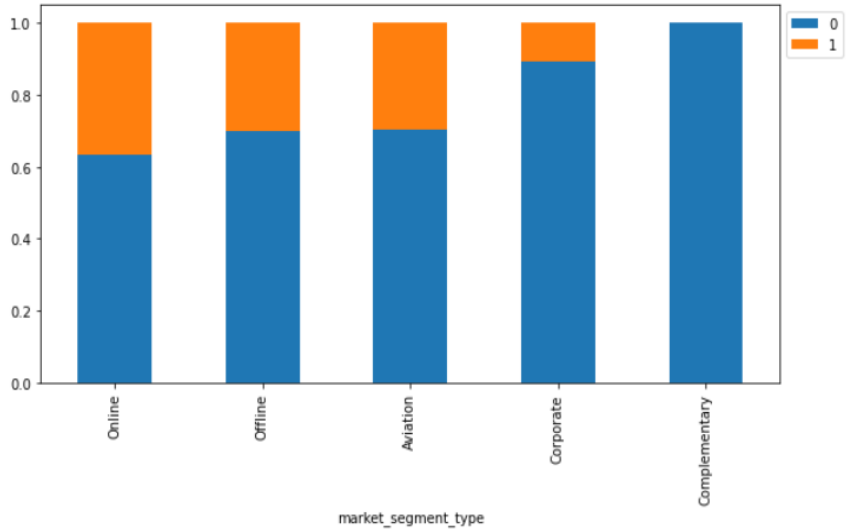
APPENDIX A – EDA Results (Bivariate Analysis)

APPENDIX A: Bivariate Analysis

APPENDIX A: Bivariate Analysis

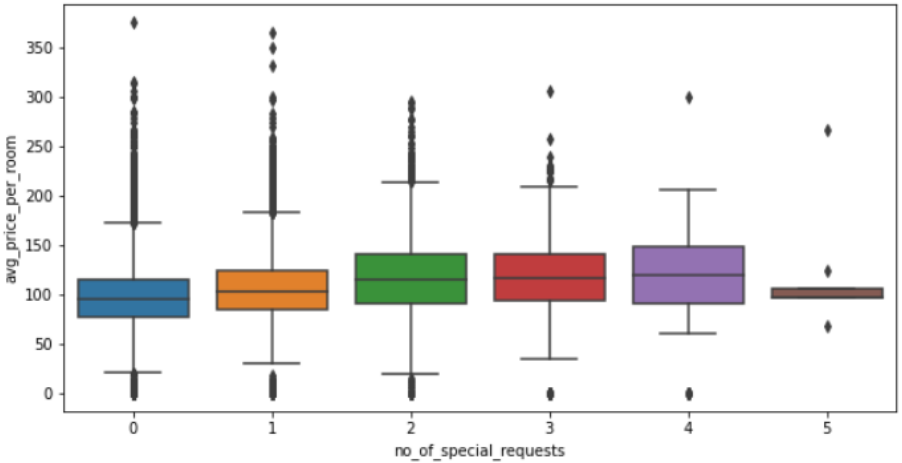


booking_status	0	1	All
market_segment_type			
All	24390	11885	36275
Online	14739	8475	23214
Offline	7375	3153	10528
Corporate	1797	220	2017
Aviation	88	37	125
Complementary	391	0	391

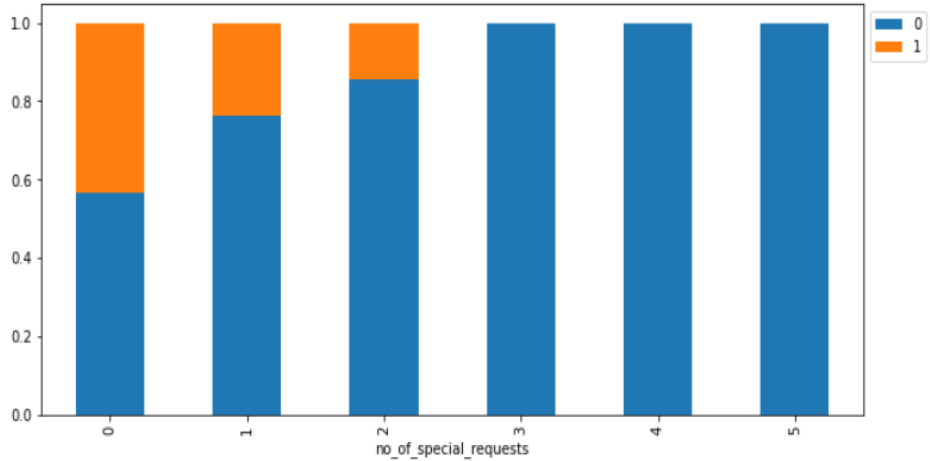


Most bookings are made online and prices for online booking are highest.
Lowest cancellation and prices are for complementary booking.
Aviation bookings also have a high proportion or cancellations.

APPENDIX A: Bivariate Analysis

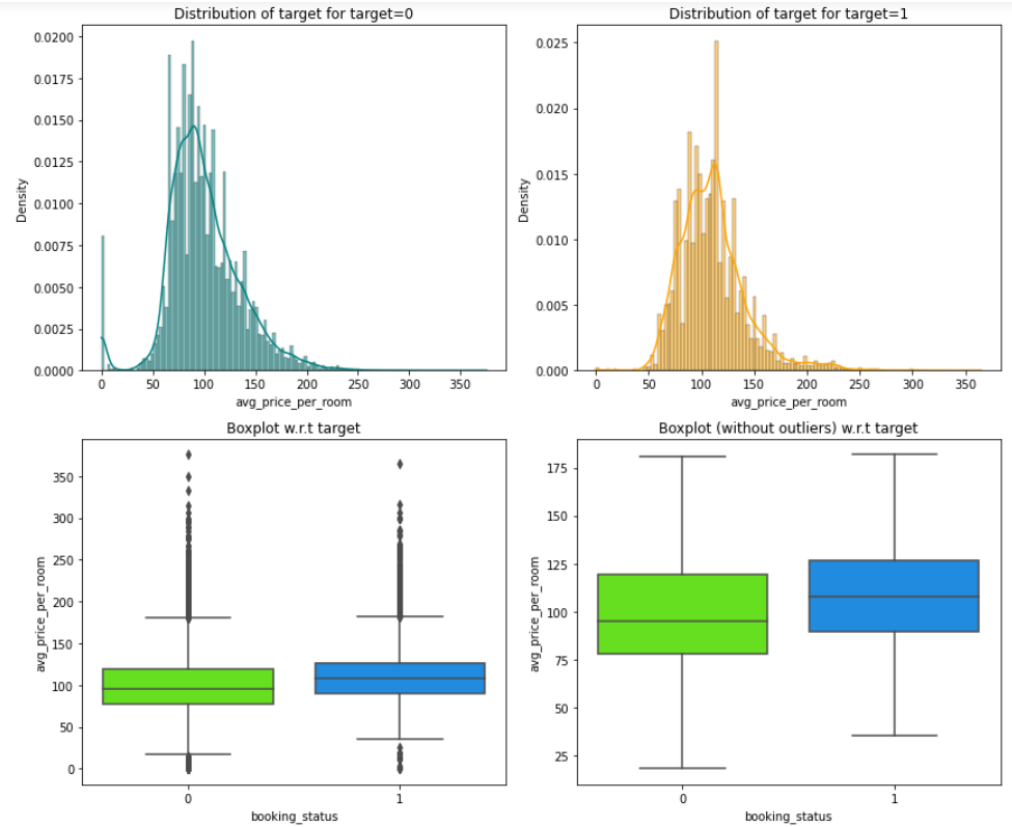


booking_status	0	1	All
no_of_special_requests			
All	24390	11885	36275
0	11232	8545	19777
1	8670	2703	11373
2	3727	637	4364
3	675	0	675
4	78	0	78
5	8	0	8



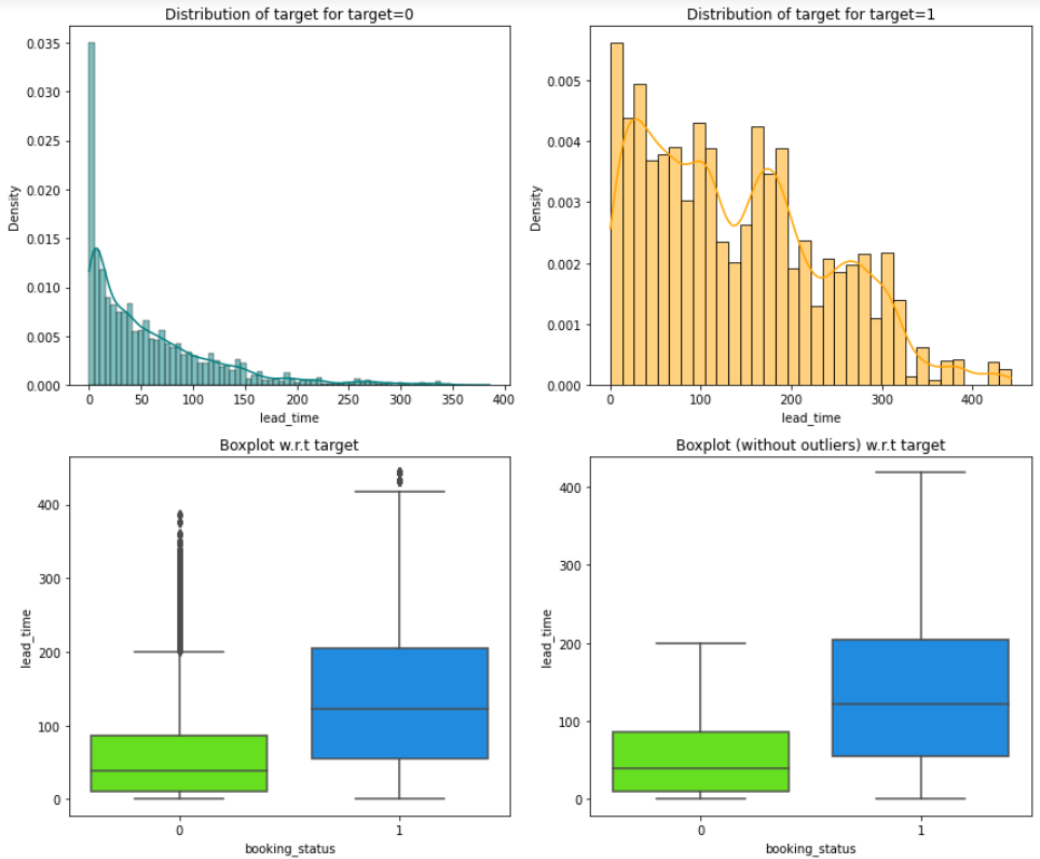
Price increases with number of special requests.
Higher the special requests the more the customer is invested in the booking the lesser the likelihood of cancellation.

APPENDIX A: Bivariate Analysis



The average room price follows a normal distribution curve.
The rooms with higher prices are more likely to be cancelled.

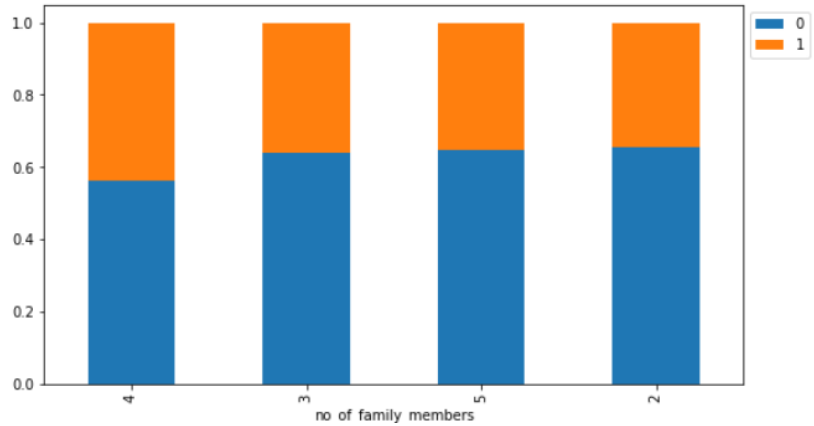
APPENDIX A: Bivariate Analysis



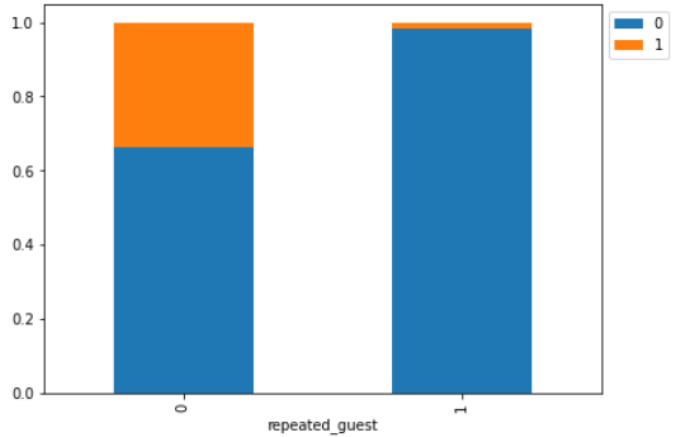
The bookings with higher lead times are more likely to be cancelled.

APPENDIX A: Bivariate Analysis

booking_status	0	1	All
no_of_family_members			
All	18456	9985	28441
2	15506	8213	23719
3	2425	1368	3793
4	514	398	912
5	11	6	17



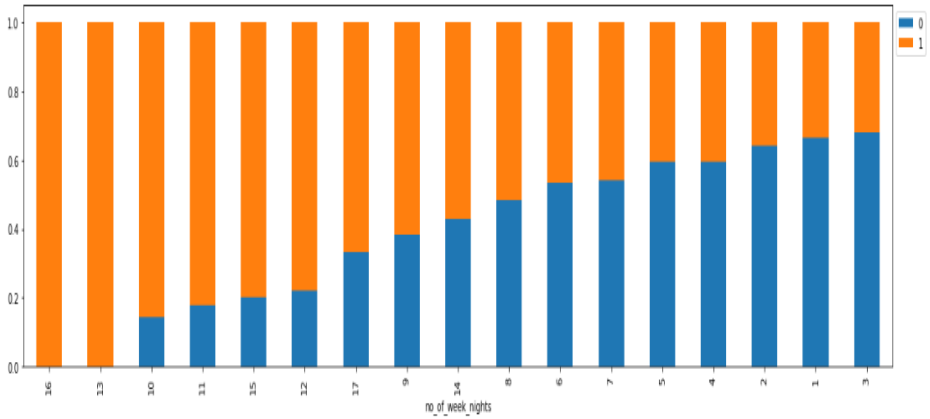
booking_status	0	1	All
repeated_guest			
All	24390	11885	36275
0	23476	11869	35345
1	914	16	930

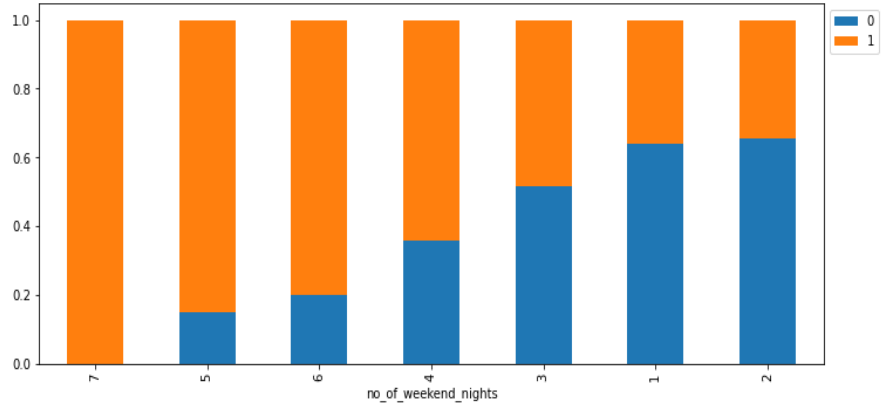


APPENDIX A: Bivariate Analysis

booking_status	0	1	All
no_of_week_nights			
All	10979	6115	17094
2	3663	2036	5699
1	2922	1484	4406
3	2488	1171	3659
4	872	589	1461
5	809	547	1356
6	101	88	189
10	9	53	62
7	61	52	113
8	30	32	62
9	13	21	34
11	3	14	17
15	2	8	10
12	2	7	9
13	0	5	5
14	3	4	7
16	0	2	2
17	1	2	3

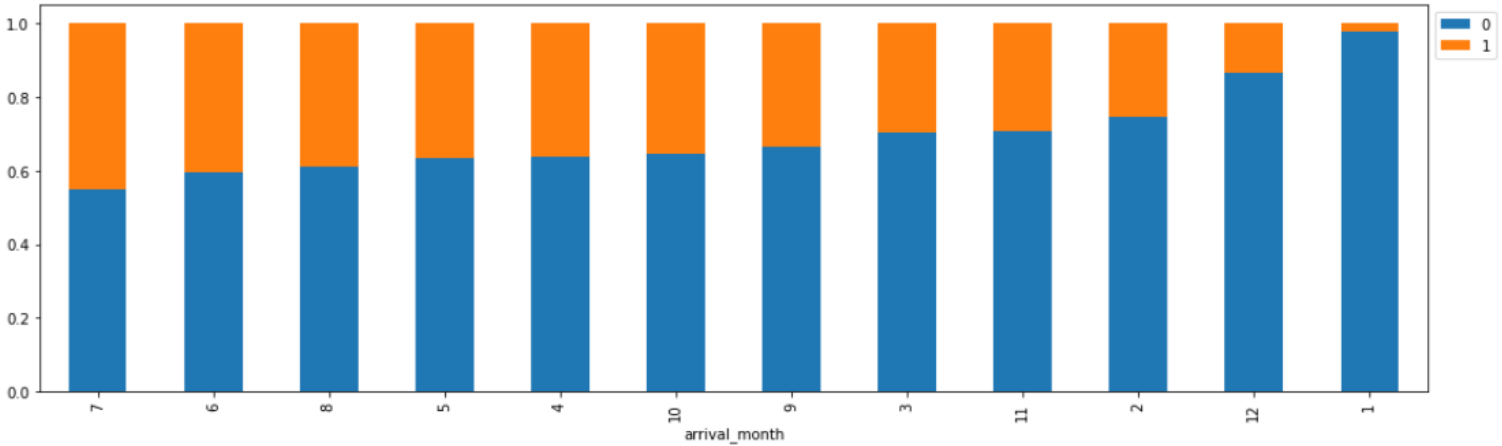
booking_status	0	1	All
no_of_weekend_nights			
All	10979	6115	17094
1	5419	3054	8473
2	5426	2858	8284
4	46	83	129
3	79	74	153
5	5	29	34
6	4	16	20
7	0	1	1



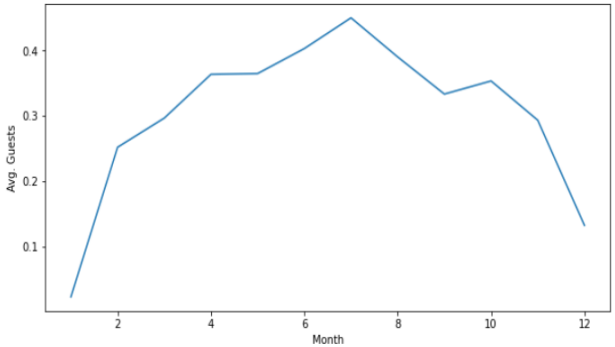


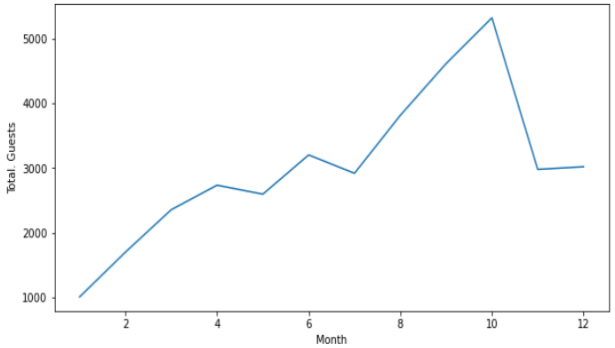
APPENDIX A: Bivariate Analysis

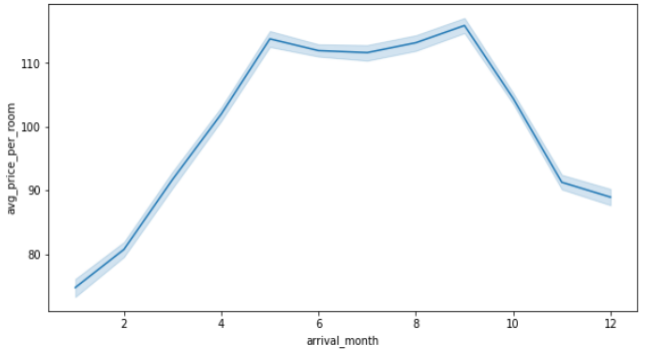
booking_status	0	1	All
arrival_month			
All	24390	11885	36275
10	3437	1880	5317
9	3073	1538	4611
8	2325	1488	3813
7	1606	1314	2920
6	1912	1291	3203
4	1741	995	2736
5	1650	948	2598
11	2105	875	2980
3	1658	700	2358
2	1274	430	1704
12	2619	402	3021
1	990	24	1014

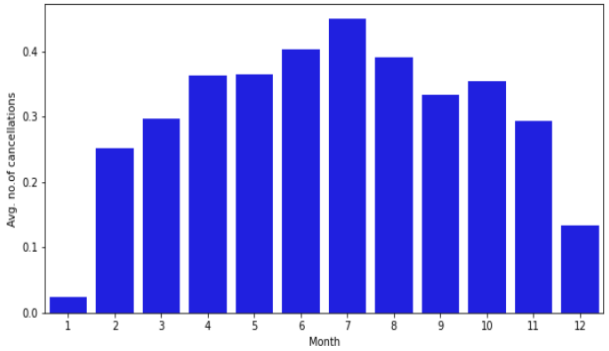


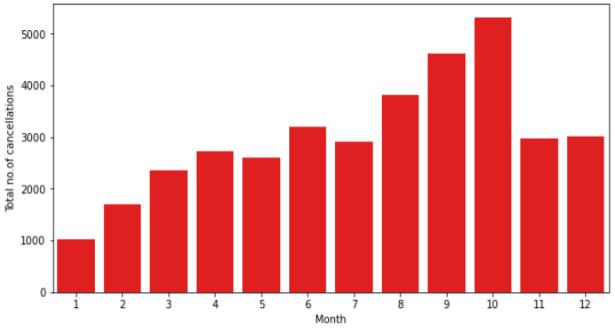
APPENDIX A: Bivariate Analysis











Average guests and average price is highest in the same in the busiest middle months of the year.
Average guests and number of cancellations is the same in the busiest middle months of the year.
Total guests and total cancellations also follow the same trends as each other.

APPENDIX B – Model Performance Summary

First Logistic Regression Model

```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.422404
         Iterations: 35
                           Logit Regression Results
==============================================================================
Dep. Variable:         booking_status   No. Observations:                25392
Model:                         Logit   Df Residuals:                    25364
Method:                          MLE   Df Model:                           27
Date:                Thu, 02 Jun 2022   Pseudo R-squ.:                0.3321
Time:                        11:26:51   Log-Likelihood:                -10726.
converged:                      False   LL-Null:                       -16060.
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                         -886.9564    121.335     -7.310      0.000   -1124.769    -649.144
no_of_adults                     0.0323      0.038      0.856      0.392      -0.042       0.106
no_of_children                   0.0680      0.063      1.075      0.282      -0.056       0.192
no_of_weekend_nights             0.1461      0.020      7.367      0.000       0.107       0.185
no_of_week_nights                0.0354      0.012      2.881      0.004       0.011       0.059
required_car_parking_space      -1.6149      0.137    -11.772      0.000      -1.884      -1.346
lead_time                        0.0158      0.000     58.942      0.000       0.015       0.016
arrival_year                     0.4382      0.060      7.288      0.000       0.320       0.556
arrival_month                   -0.0476      0.006     -7.336      0.000      -0.060      -0.035
arrival_date                     0.0030      0.002      1.543      0.123      -0.001       0.007
repeated_guest                  -1.9184      0.767     -2.502      0.012      -3.421      -0.416
no_of_previous_cancellations     0.3476      0.102      3.413      0.001       0.148       0.547
no_of_previous_bookings_not_canceled -1.3824  0.906     -1.527      0.127      -3.157       0.393
avg_price_per_room               0.0185      0.001     24.946      0.000       0.017       0.020
no_of_special_requests          -1.4899      0.030    -48.954      0.000      -1.550      -1.430
type_of_meal_plan_Meal Plan 2    0.1732      0.067      2.586      0.010       0.042       0.305
type_of_meal_plan_Meal Plan 3   14.0463    758.641      0.019      0.985   -1472.863    1500.956
type_of_meal_plan_Not Selected   0.1986      0.053      3.724      0.000       0.094       0.303
room_type_reserved_Room_Type 2  -0.4087      0.134     -3.061      0.002      -0.670      -0.147
room_type_reserved_Room_Type 3   1.1880      1.891      0.628      0.530      -2.519       4.895
room_type_reserved_Room_Type 4  -0.2697      0.053     -5.050      0.000      -0.374      -0.165
room_type_reserved_Room_Type 5  -0.6814      0.215     -3.170      0.002      -1.103      -0.260
room_type_reserved_Room_Type 6  -0.8243      0.155     -5.317      0.000      -1.128      -0.520
room_type_reserved_Room_Type 7  -1.3507      0.298     -4.536      0.000      -1.934      -0.767
market_segment_type_Complementary -22.7477 1563.580     -0.015      0.988   -3087.308    3041.813
market_segment_type_Corporate   -0.8518      0.276     -3.088      0.002      -1.392      -0.311
market_segment_type_Offline     -1.7631      0.264     -6.683      0.000      -2.280      -1.246
market_segment_type_Online       0.0082      0.261      0.031      0.975      -0.503       0.520
================================================================================================
```

Training performance:

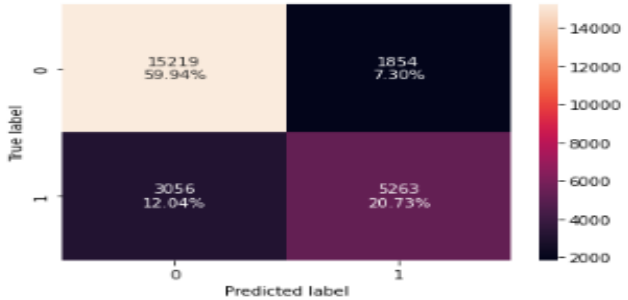
	Accuracy	Recall	Precision	F1
0	0.80679	0.63277	0.73985	0.68213

Testing performance:

	Accuracy	Recall	Precision	F1
0	0.80391	0.62675	0.73568	0.67686

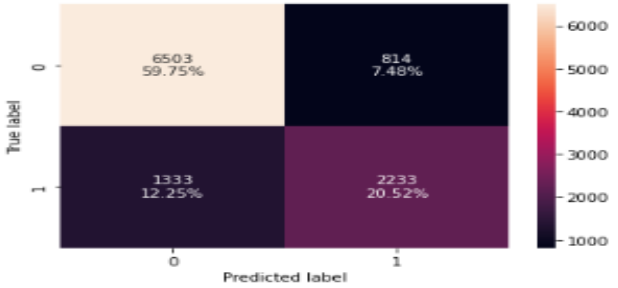
Final Logistic Regression Model (Default Threshold = 0.5)

```
Optimization terminated successfully.
         Current function value: 0.423028
         Iterations 11
                          Logit Regression Results
==============================================================================
Dep. Variable:         booking_status   No. Observations:                25392
Model:                          Logit   Df Residuals:                    25372
Method:                           MLE   Df Model:                           19
Date:                Thu, 02 Jun 2022   Pseudo R-squ.:                 0.3312
Time:                        11:26:54   Log-Likelihood:                -10742.
converged:                       True   LL-Null:                       -16060.
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                         -869.4077    120.911     -7.191      0.000   -1106.388    -632.427
no_of_weekend_nights             0.1497      0.020      7.564      0.000       0.111       0.189
no_of_week_nights                0.0362      0.012      2.950      0.003       0.012       0.060
required_car_parking_space      -1.6151      0.137    -11.783      0.000      -1.884      -1.346
lead_time                        0.0159      0.000     59.906      0.000       0.015       0.016
arrival_year                     0.4295      0.060      7.168      0.000       0.312       0.547
arrival_month                   -0.0493      0.006     -7.614      0.000      -0.062      -0.037
repeated_guest                  -3.0741      0.597     -5.151      0.000      -4.244      -1.904
no_of_previous_cancellations     0.2891      0.078      3.721      0.000       0.137       0.441
avg_price_per_room               0.0191      0.001     26.685      0.000       0.018       0.020
no_of_special_requests          -1.4855      0.030    -49.274      0.000      -1.545      -1.426
type_of_meal_plan_Meal Plan 2    0.1657      0.067      2.475      0.013       0.035       0.297
type_of_meal_plan_Not Selected   0.2100      0.053      3.977      0.000       0.106       0.313
room_type_reserved_Room_Type 2  -0.3733      0.129     -2.890      0.004      -0.627      -0.120
room_type_reserved_Room_Type 4  -0.2685      0.052     -5.203      0.000      -0.370      -0.167
room_type_reserved_Room_Type 5  -0.6889      0.214     -3.216      0.001      -1.109      -0.269
room_type_reserved_Room_Type 6  -0.7440      0.120     -6.203      0.000      -0.979      -0.509
room_type_reserved_Room_Type 7  -1.3177      0.292     -4.512      0.000      -1.890      -0.745
market_segment_type_Corporate   -0.8710      0.103     -8.444      0.000      -1.073      -0.669
market_segment_type_Offline     -1.7713      0.052    -34.275      0.000      -1.873      -1.670
================================================================================================
```



Training performance:

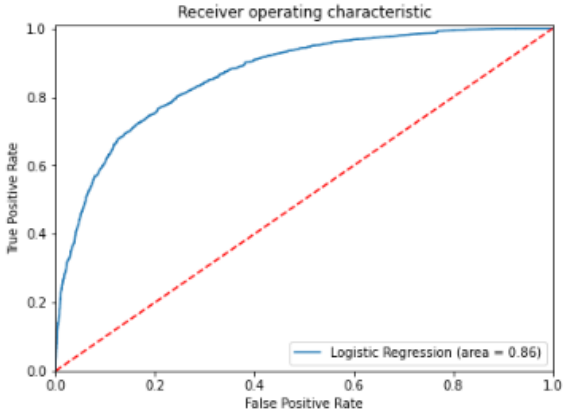
	Accuracy	Recall	Precision	F1
0	0.80663	0.63265	0.73950	0.68191

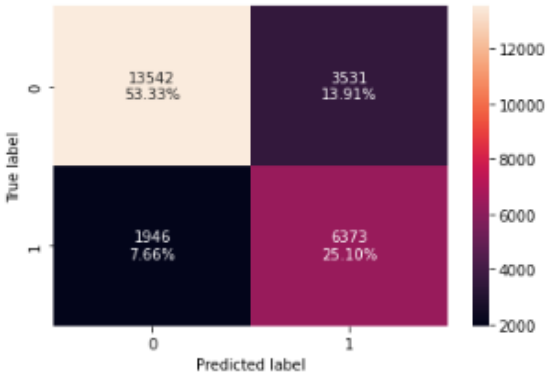


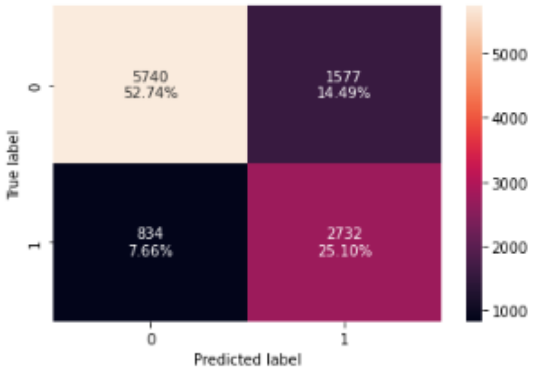
Training performance:

	Accuracy	Recall	Precision	F1
0	0.80272	0.62619	0.73285	0.67534

Model using ROC-AUC Curve







Optimal threshold value from the ROC-AUC curve was found to be:

0.33300753336836625

Training performance:

	Accuracy	Recall	Precision	F1
0	0.78430	0.76608	0.64348	0.69945

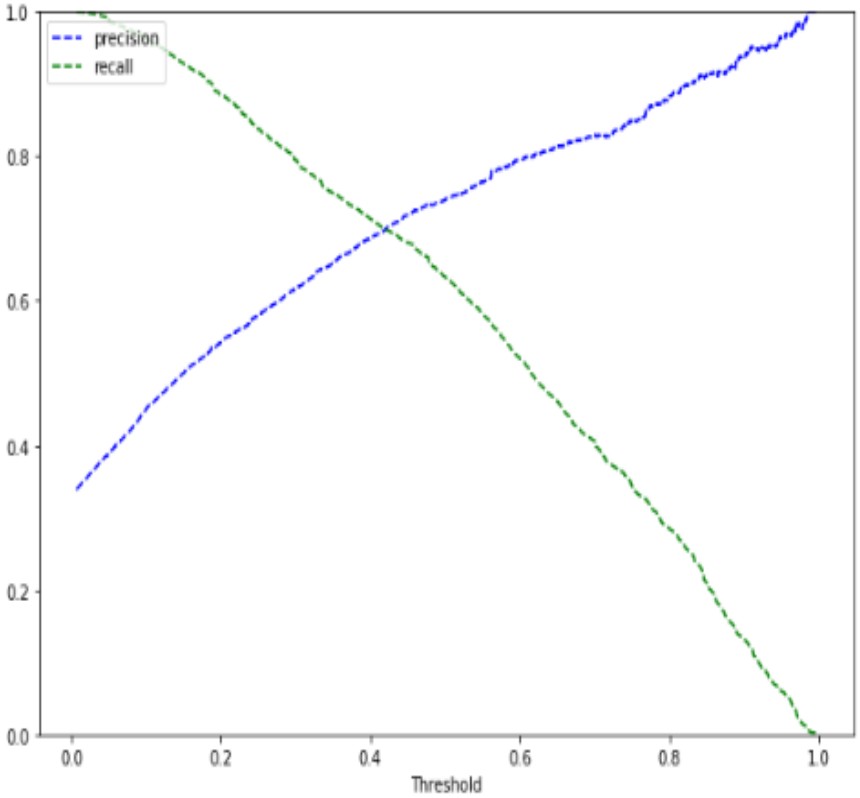
Testing Performance:

	Accuracy	Recall	Precision	F1
0	0.77846	0.76612	0.63402	0.69384

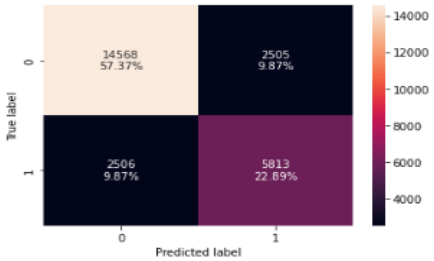
Observation:

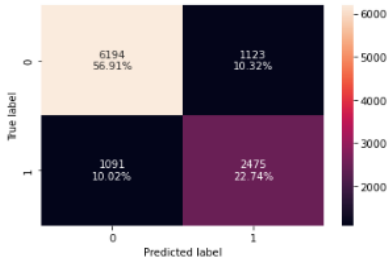
ROC-AUC model performance slightly better than the default threshold model.

Model using Precision-Recall Curve



Threshold = 0.42





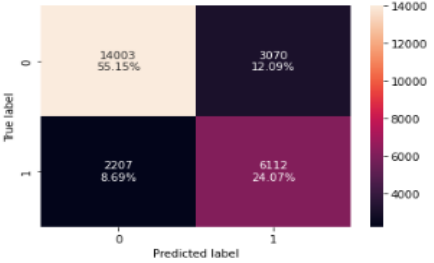
Training performance:

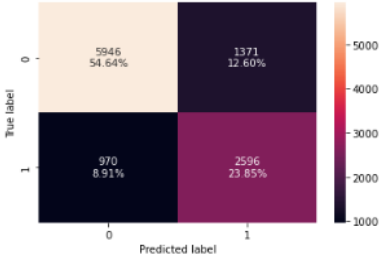
	Accuracy	Recall	Precision	F1
0	0.80265	0.69876	0.69885	0.69880

Test performance:

	Accuracy	Recall	Precision	F1
0	0.79656	0.69405	0.68788	0.69095

Threshold = 0.37





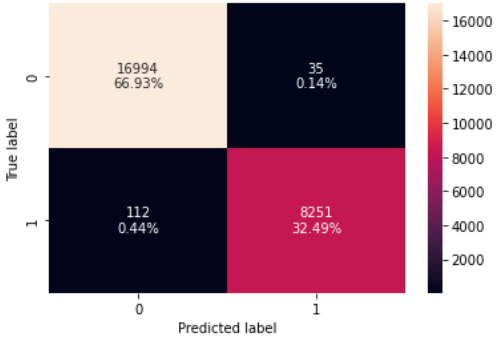
Training performance:

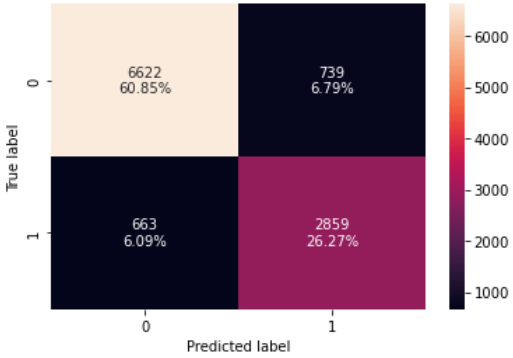
	Accuracy	Recall	Precision	F1
0	0.79218	0.73470	0.66565	0.69847

Test performance:

	Accuracy	Recall	Precision	F1
0	0.78489	0.72799	0.65440	0.68923

First Decision Tree Model



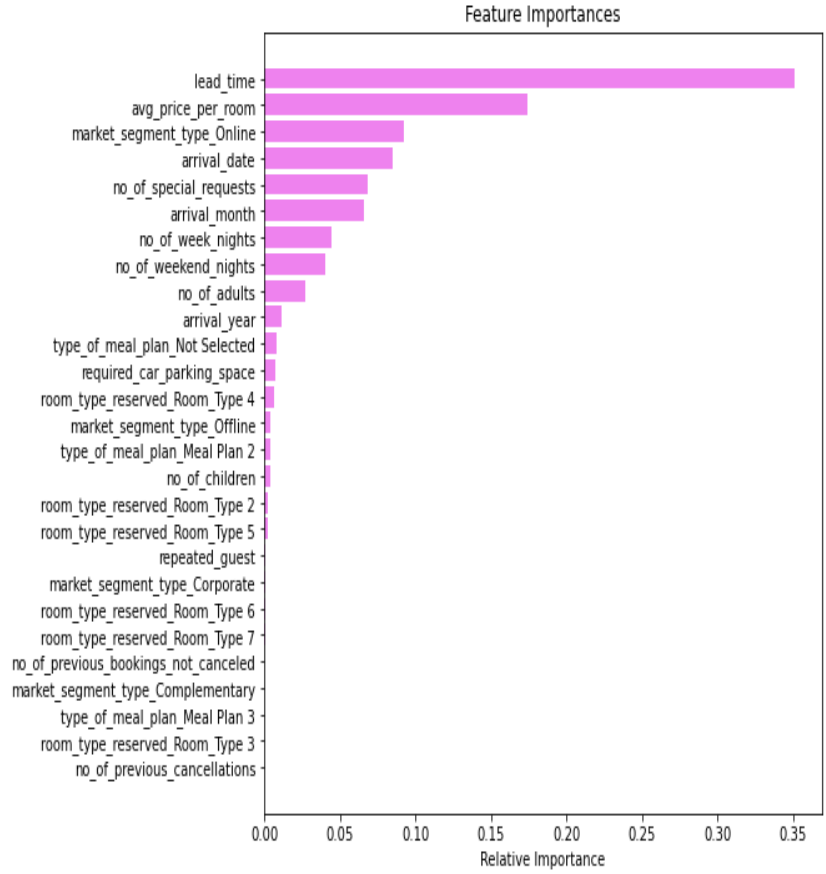


Training performance:

	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

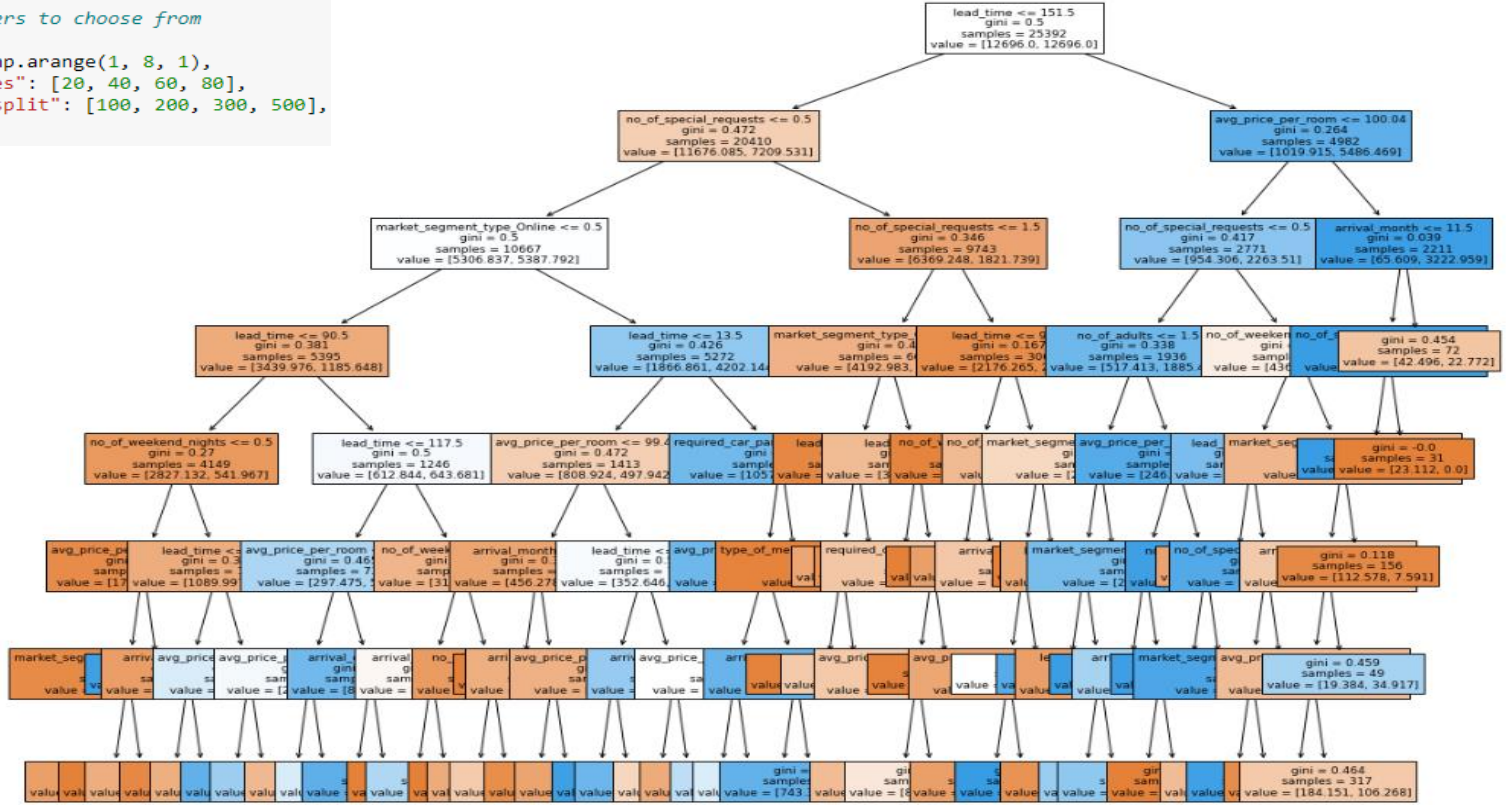


Observations:

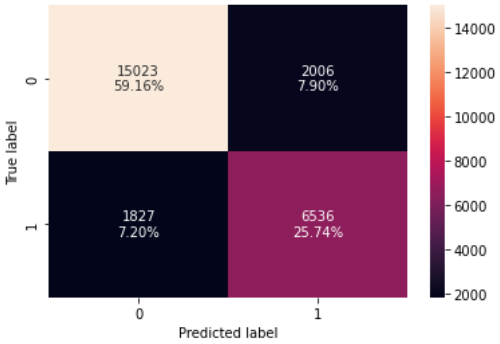
- The training model is overfitted as our performance on testing data is relatively weaker.
- Lead time, average price per room per day and market segment are the top three important feature in this model tree

Pre-pruning using Grid Search - Decision Tree Model

```
# Grid of parameters to choose from
parameters = {
    "max_depth": np.arange(1, 8, 1),
    "max_leaf_nodes": [20, 40, 60, 80],
    "min_samples_split": [100, 200, 300, 500],
}
```

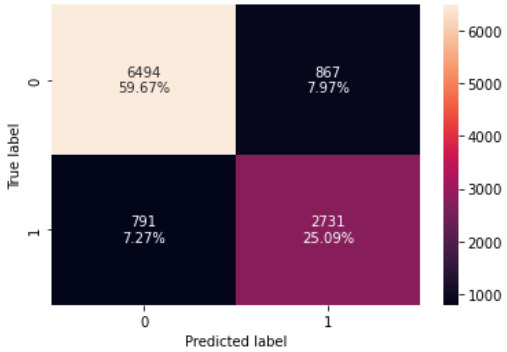


Pre-pruning using Grid Search - Decision Tree Model



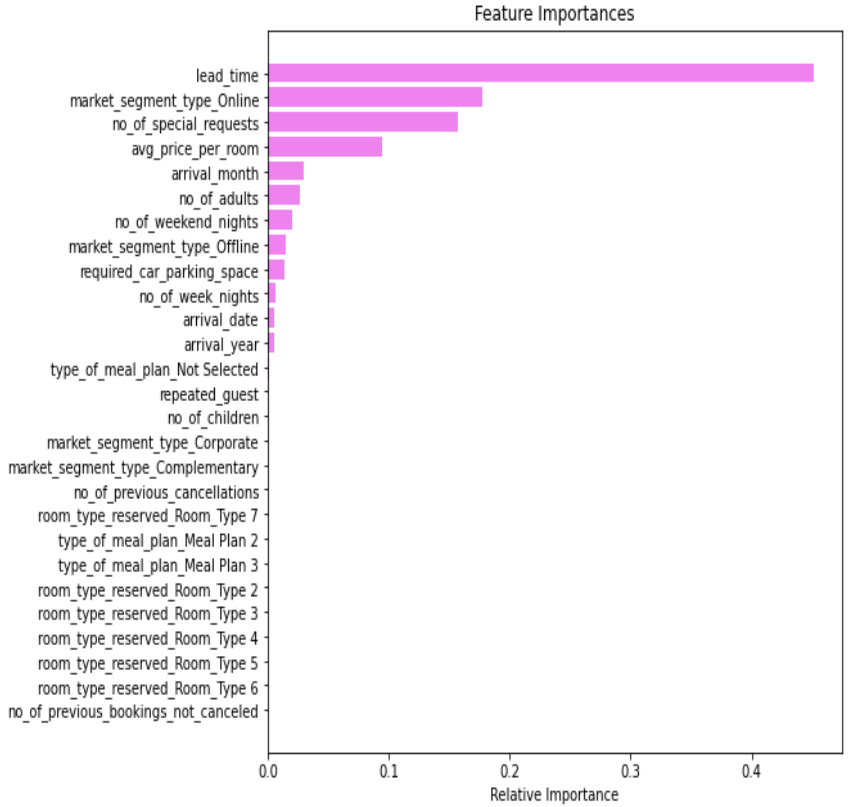
Training performance:

	Accuracy	Recall	Precision	F1
0	0.84905	0.78154	0.76516	0.77326



Test performance:

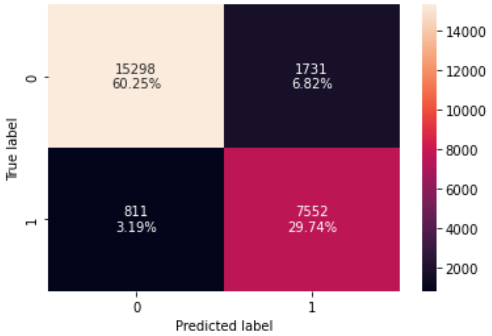
	Accuracy	Recall	Precision	F1
0	0.84765	0.77541	0.75903	0.76713

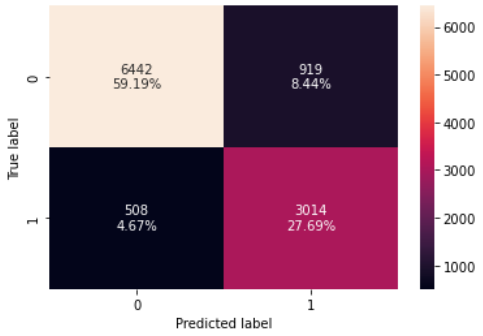


Observations:

- The training and test performance parameters are relatively similar.
- The decision tree has:
  - Max depth: 7
  - Max leaf nodes: 60
  - Min samples split: 100

Cost Complexity Method- Decision Tree Model



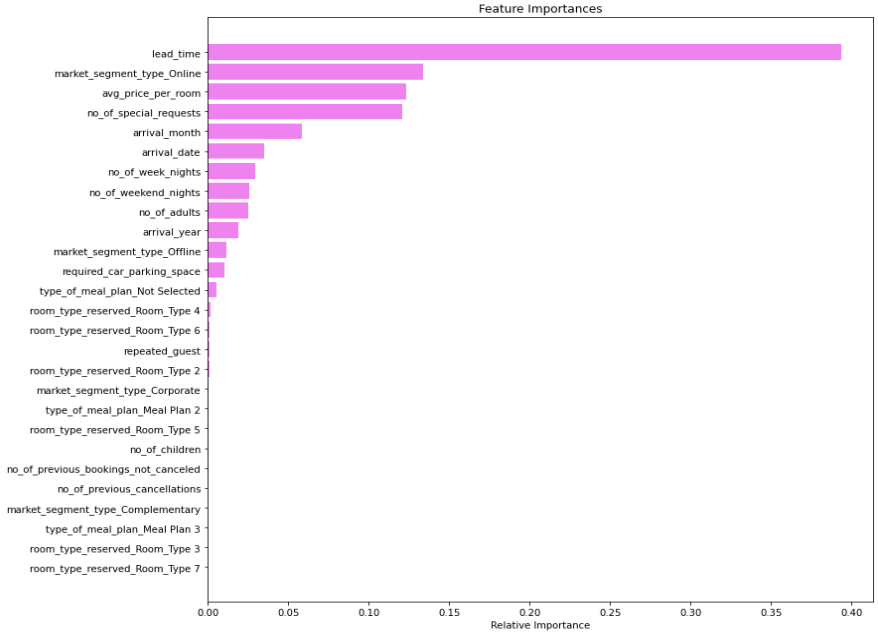


Training performance:

	Accuracy	Recall	Precision	F1
0	0.89989	0.90303	0.81353	0.85594

Test performance:

	Accuracy	Recall	Precision	F1
0	0.86888	0.85576	0.76634	0.80858

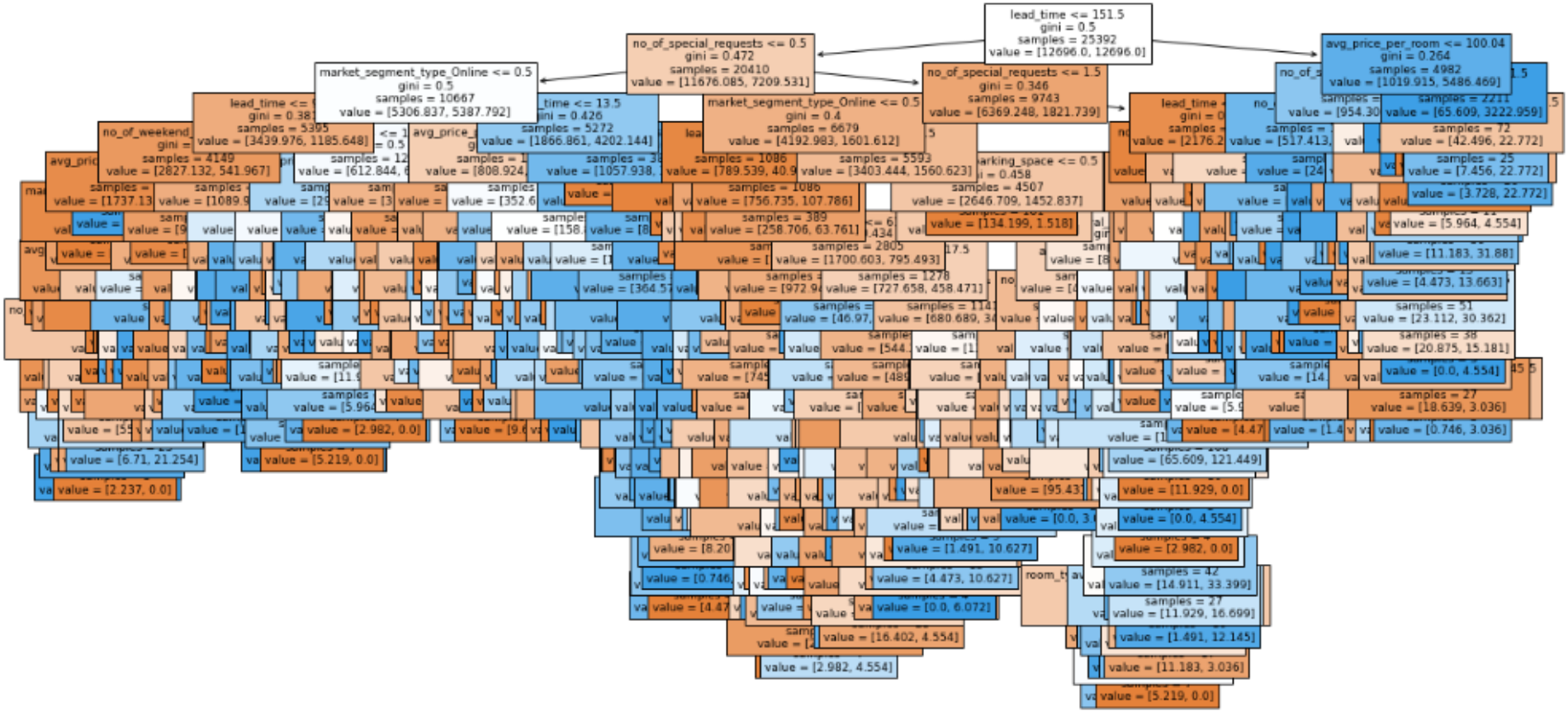


Observations:

- The best alpha value was found to be:

  0.0001226763315516706

Cost Complexity Method- Decision Tree Model



Great Learning

Happy Learning !