INN Hotels
(a project on identifying factors that influence booking cancellations)

INN Hotels – Data analysis of factors influencing booking cancellations
The University of Texas at Austin
McCombs School of Business

Itu Mukherjee
Date : 09.12.2022

Executive Summary

A significant number of hotel bookings were cancelled for INN Hotels. Some of the reasons for cancellations include change of plans, scheduling conflicts, etc.

We analyzed the data provided to identify factors that influence booking cancellations and constructed a model to predict which booking is going to be canceled in advance, to aid in formulating profitable policies for cancellations and refunds.

Our results indicate that the three most important variables in terms of cancellations are the lead time, which is a measure of how far in advance the rooms were booked; special request for the stay; and average price for the room. Rooms booked in advance of 151 days (5 months) or less were much less likely to be cancelled. Those who made a special request on top of that were very unlikely to cancel. This I believe is an opportunity. Rooms booked over 151 days in advance were more likely to cancel. Price was the determining factor for those cancellations.

Business Problem Overview and Solution Approach

- Our objective is to identify the factors that have a high impact on booking cancellations for INN Hotels and build a model that can predict which booking is going to be canceled in advance.

- Solution approach and methodology
  - EDA (bivariate and univariate analysis), duplicate value check, missing value treatment, outlier check (treatment if needed)
  - Logistic Regression model building
  - Train, test data split, model performance check
  - Checking multicollinearity, ROC curve analysis, model performance check with threshold 0.37 and 0.42
  - Decision Tree model building, Pre-Pruning, Cost Complexity Pruning, Comparing Decision Tree models

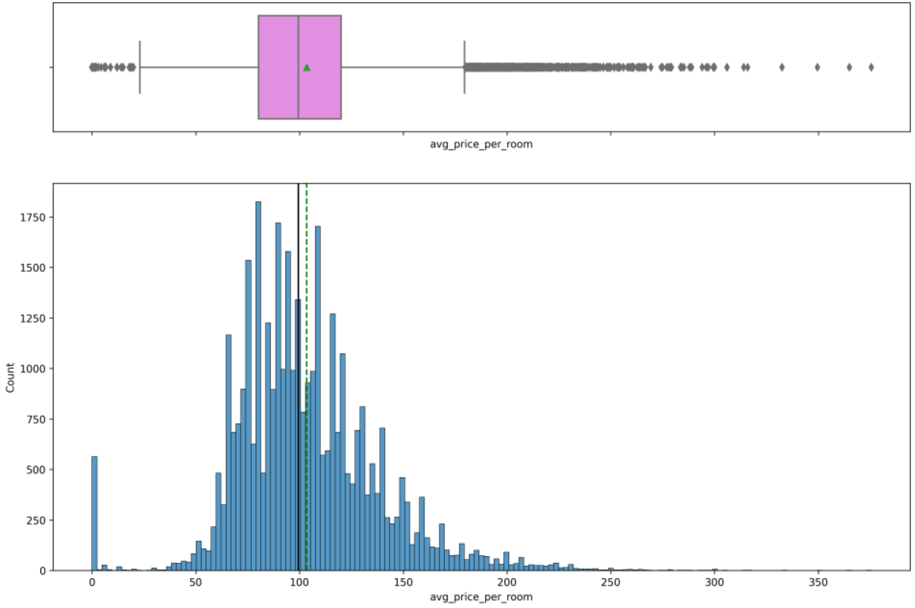
Data Overview

	Booking_ID	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	type_of_meal_plan	required_car_parking_space
0	INN00001	2	0	1	2	Meal Plan 1	0
1	INN00002	2	0	2	3	Not Selected	0
2	INN00003	1	0	2	1	Meal Plan 1	0
3	INN00004	2	0	0	2	Meal Plan 1	0
4	INN00005	2	0	1	1	Not Selected	0

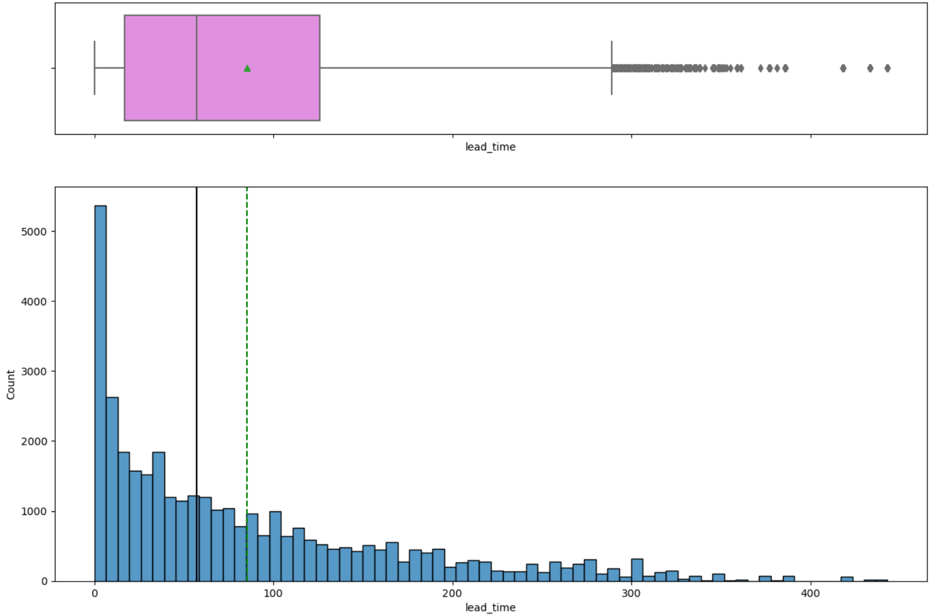
room_type_reserved	lead_time	arrival_year	arrival_month	arrival_date	market_segment_type	repeated_guest	no_of_previous_cancellations	no_of_previous_bookings_not_canceled	avg_price_per_room	no_of_special_requests	booking_status
Room_Type 1	224	2017	10	2	Offline	0	0	0	65.00000	0	Not_Canceled
Room_Type 1	5	2018	11	6	Online	0	0	0	106.68000	1	Not_Canceled
Room_Type 1	1	2018	2	28	Online	0	0	0	60.00000	0	Canceled
Room_Type 1	211	2018	5	20	Online	0	0	0	100.00000	0	Canceled
Room_Type 1	48	2018	4	11	Online	0	0	0	94.50000	0	Canceled

- 36275 entries (rows) of 19 data points (columns) with no missing or duplicated data.
- Booking ID, type of meal plan, room type reserved, market segment type and booking status are categorical while all others are numerical data types. However, one is the Booking ID.
- Booking status is the dependent variable.

Exploratory Data Analysis

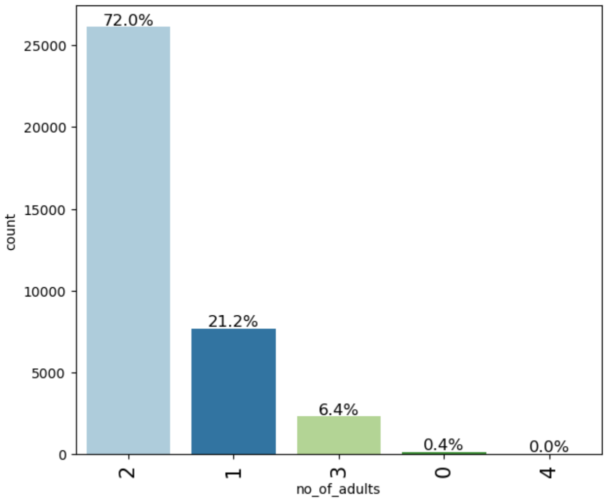


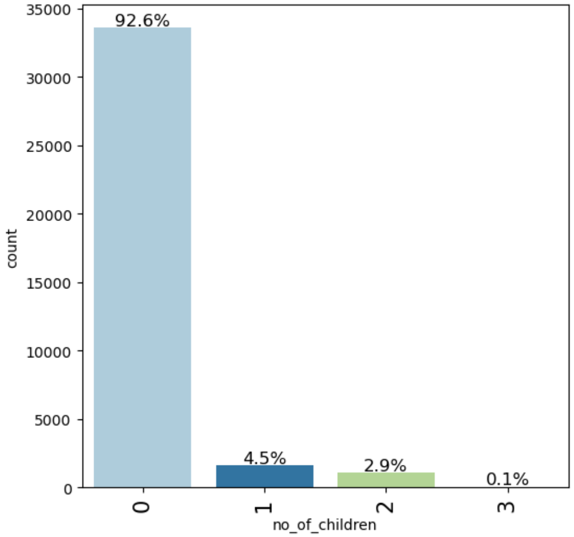
- ➢ Average price per room is normally distributed with a median price of EUR 100.

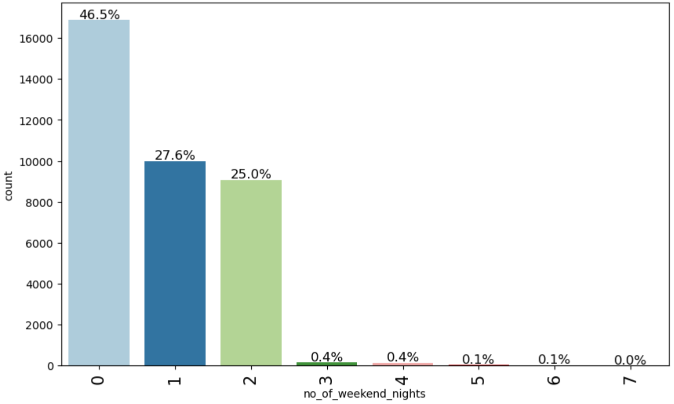


- ➢ The distribution of lead time is right skewed with a median of 90 days.

Exploratory Data Analysis

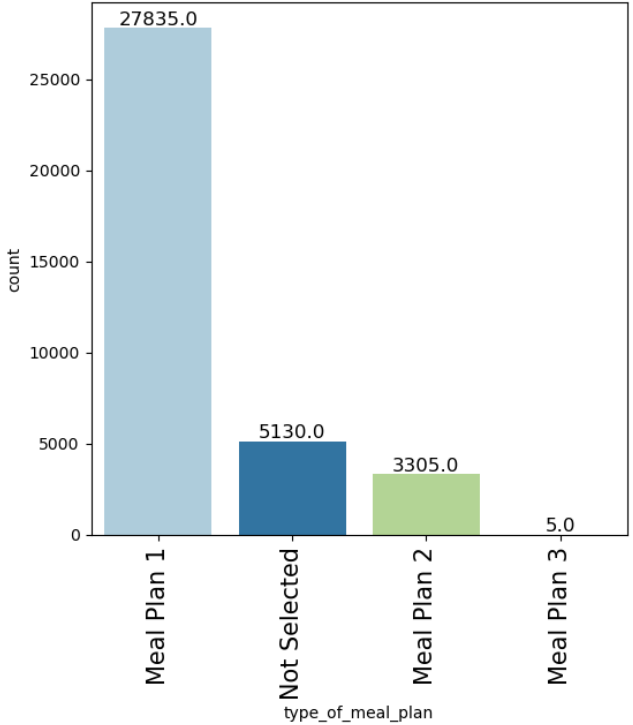


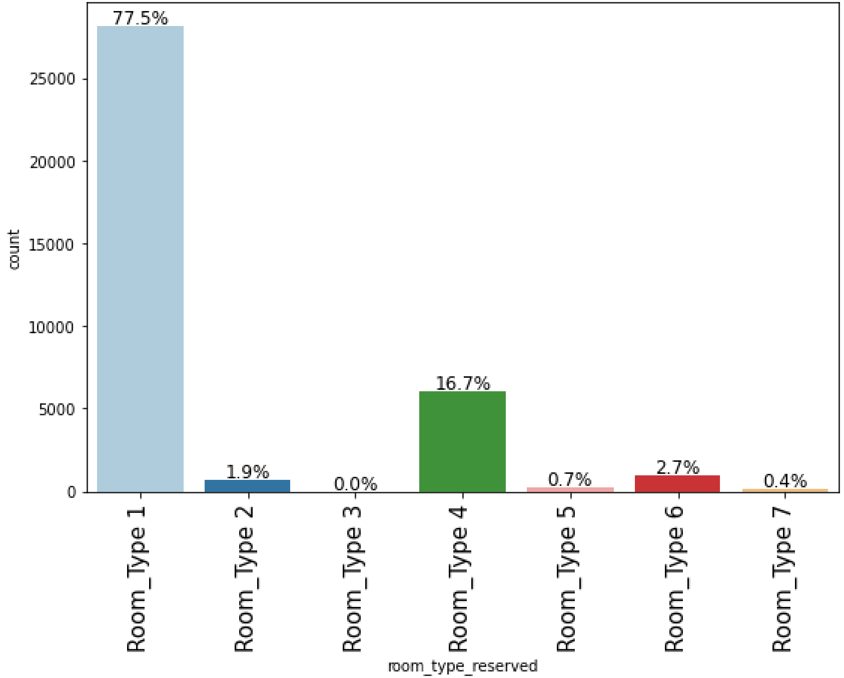




- 72% of adults who book rooms have another adult staying with them.
- Children are rare at the hotels, as 92.6% of booking don't include children in the rooms.
- 52.6% of bookings include at least one weekend night.

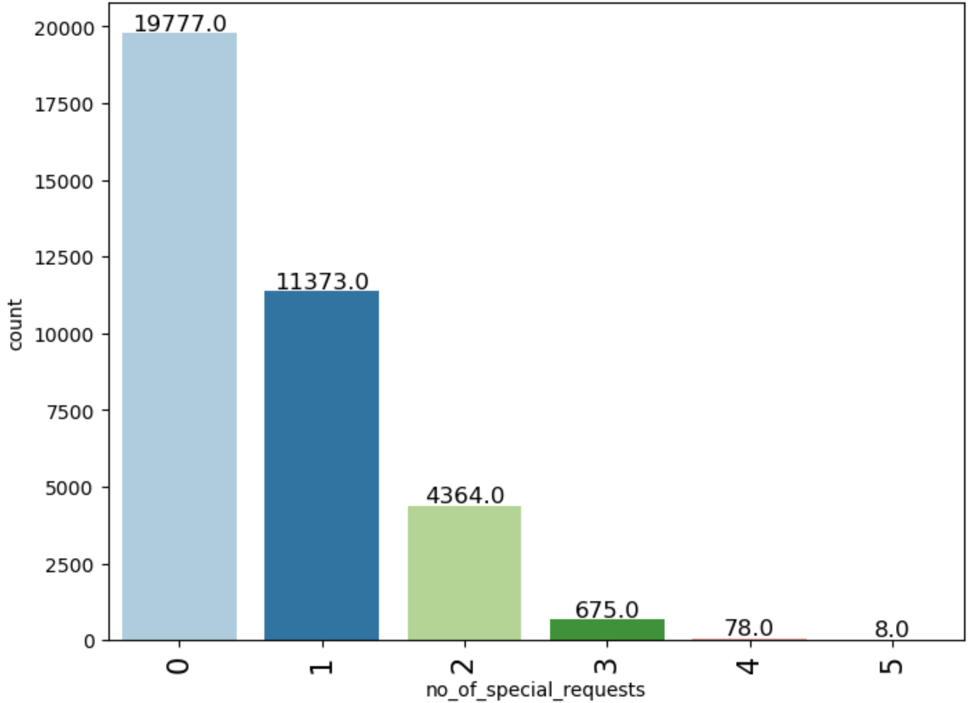
Exploratory Data Analysis

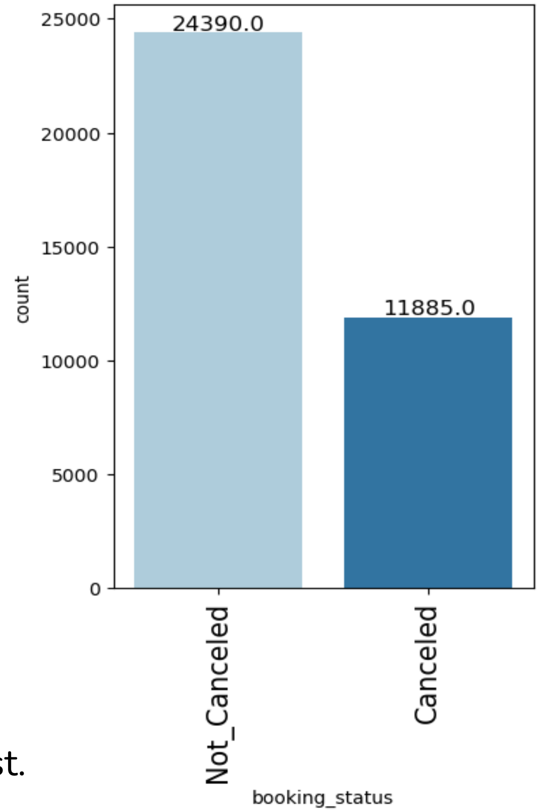




- 90% of room type booked are type1 & type4.
- Breakfast is the only plan which is the most popular one which is in this case is Meal Plan 1, and the ‘FULL BOARD’ plan is almost never booked.

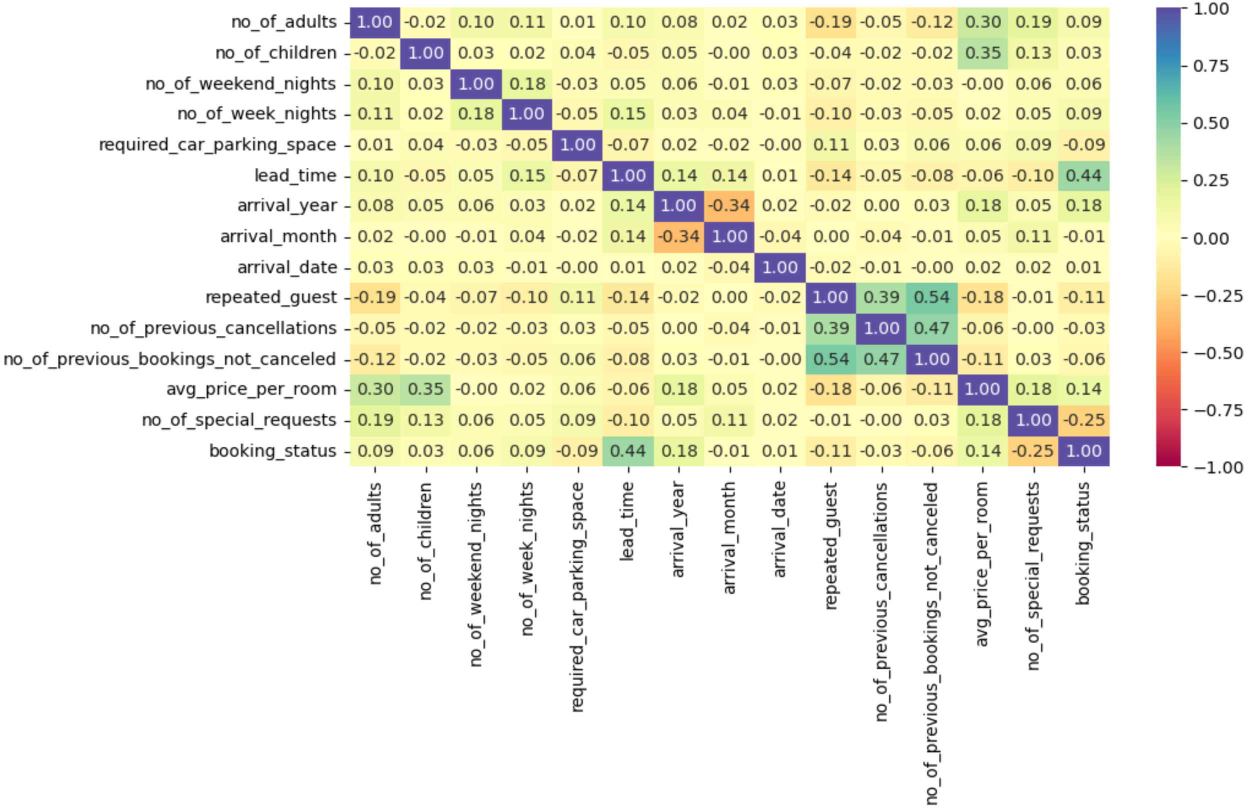
Exploratory Data Analysis





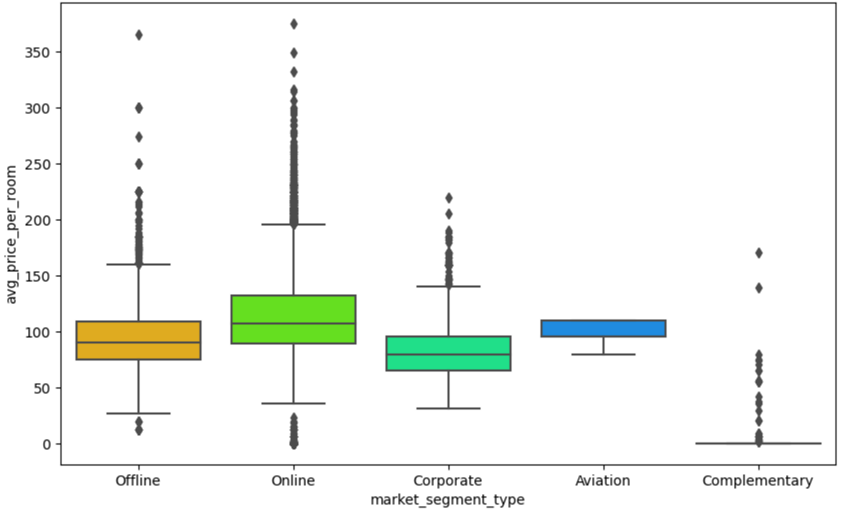
- The majority of bookings did not have a special request.
- No. of bookings not cancelled is more than the cancellations.

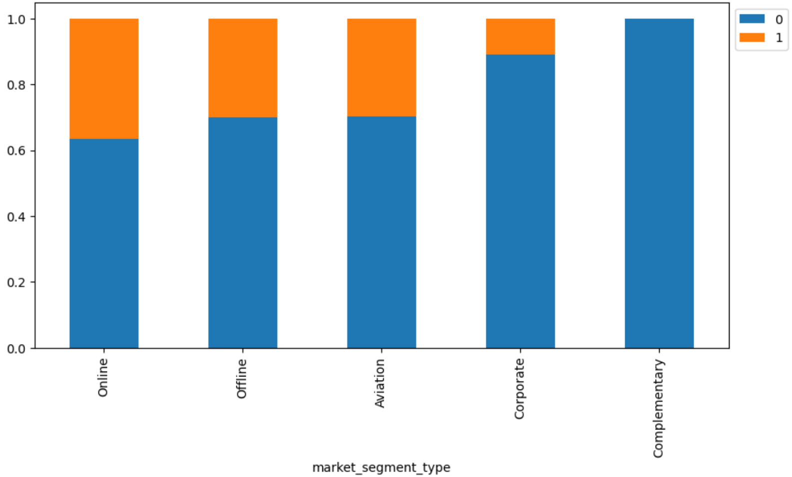
Exploratory Data Analysis

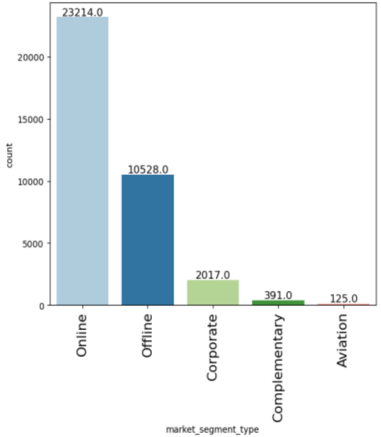


- There is a strong positive association between no. of previous bookings not cancelled and repeated guests
- Also, there is a positive correlation between no. of previous cancellations and guest repeated and lead time and booking status.
- It shows a negative correlation between arrival month and arrival year.

Exploratory Data Analysis

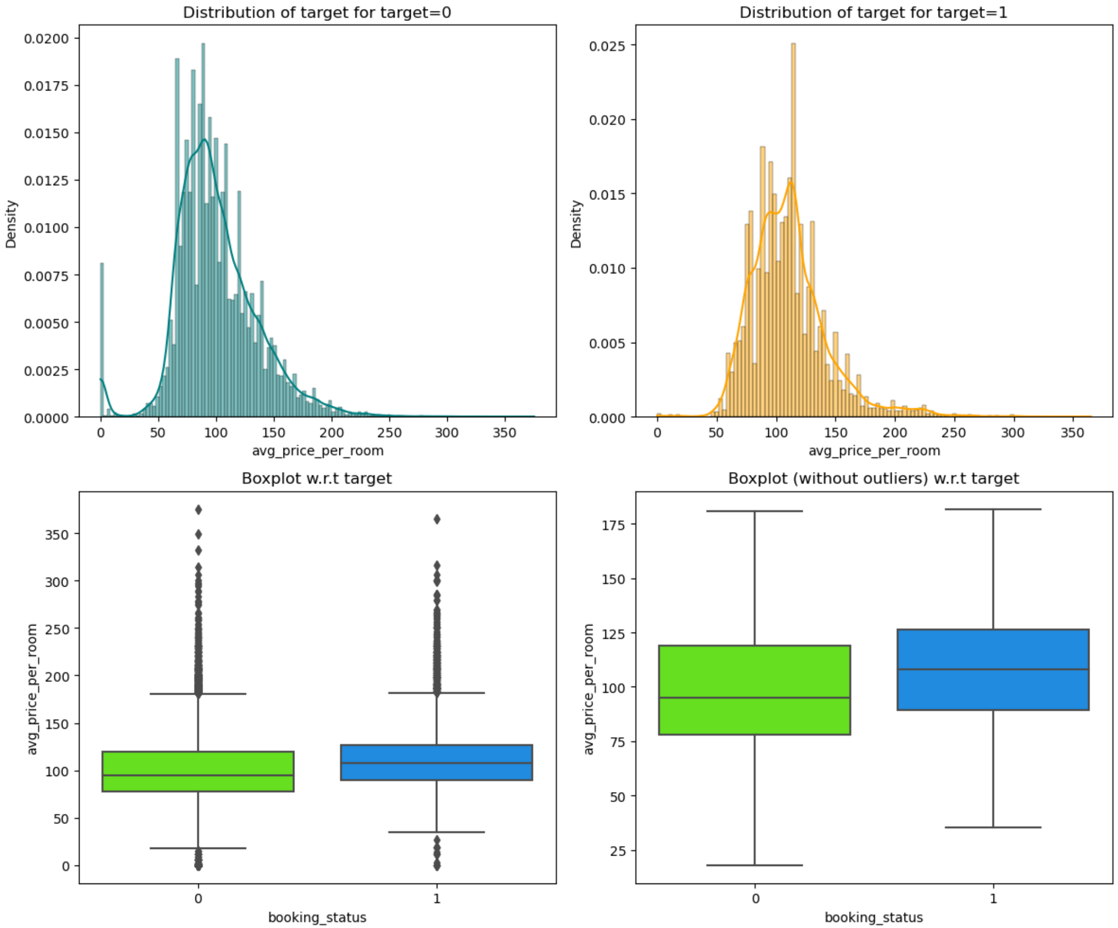






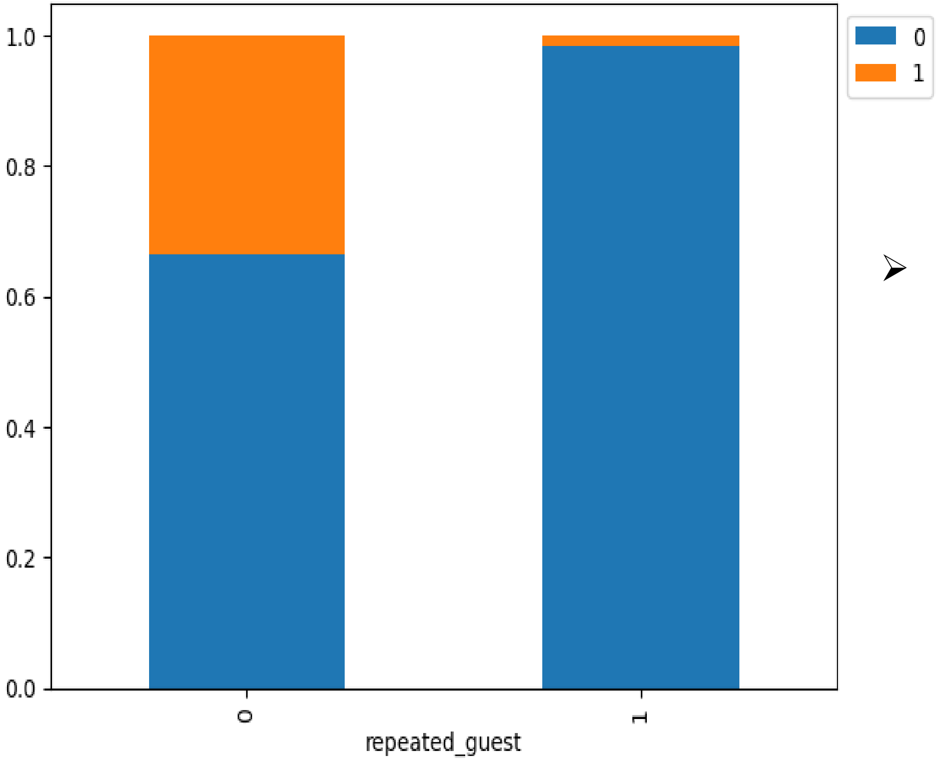
- Online booked rooms have the highest cost of booking. Aviation, Offline, and Corporate are generally slightly lower priced with Corporate edging out for the lowest.

Exploratory Data Analysis



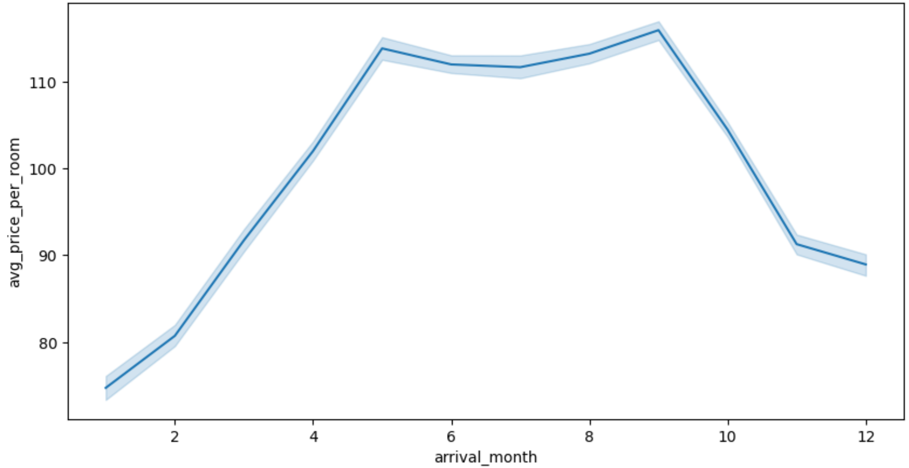
- Canceled bookings appear to be slightly more expensive.

Exploratory Data Analysis

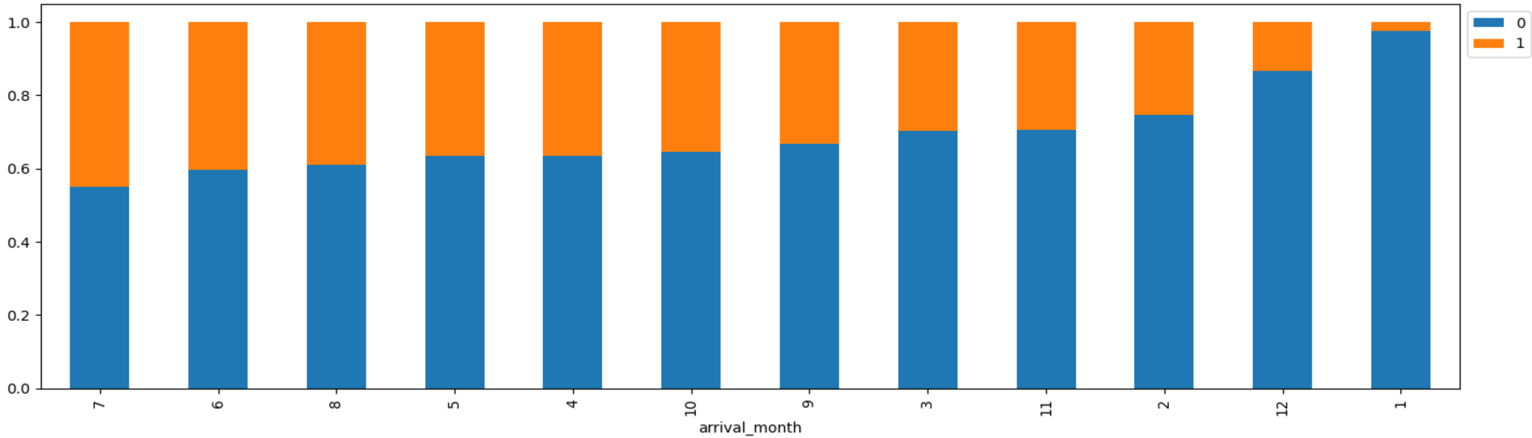


- There is very less cancellation from regular guests, meaning the level of satisfaction for the regular guests are likely very high.

Exploratory Data Analysis



- ➢ Both the graphs shows that late summer / early fall (AUG – OCT) is the busiest time of the year for the hotel industry and price is also high at those times.



Exploratory Data Analysis Results

- Late Summer / Early Fall (AUG – OCT) is the busiest time of the year for the hotel industry.
- Nearly 2/3 of bookings come from online sources.
- Room rent ranges between EUR100 plus or minus 25.
- Of 36275 room rentals 545 were free of charge, 354 are complementary and 191 were online over the course of the survey.
- Online booked rooms have the highest cost of booking.
- Repeated guest rarely cancel, meaning that these customers are less likely to cancel booking.
- Guest who make special request for their stay, are significantly less likely to cancel the reservation.
- There is a strong correlation between no. of previous bookings not cancelled and repeated guests

Exploratory Data Analysis Results

- 72% of adults who book rooms have another adult staying with them.
- Children are rare at the hotels, as 92.6% of booking don't include children in the rooms.
- 52.6% of bookings include at least one weekend night.
- The hotel rarely has long stay guests.
- Parking is not a factor for almost all the guests, I wouldn't bother promoting it.
- 90% of room type booked are type1 & type4.
- Breakfast is the only plan which is the most popular one, and the 'FULL BOARD' plan is almost never booked.
- The further out in terms of day that rooms are booked the more likely they are to be canceled.
- The client has a robust pricing structure.

Data Preprocessing

Great Learning
POWER AHEAD

- Duplicate value check - There are no duplicate values in the Data Frame
- Missing value treatment - There are no missing values
- Outlier check (treatment if needed) - No insignificant outliers found
- Data preparation for modeling
  - To evaluate the model, we split the data into train and test in the ratio of 70-30.
  - We built a Logistic Regression model using the train data and then checked its performance. Before model building, we encoded the categorical features.
  - Booking Status is the dependent variable.

Model Building - Logistic Regression

```
                           Logit Regression Results
==============================================================================
Dep. Variable:         booking_status   No. Observations:                25392
Model:                          Logit   Df Residuals:                    25364
Method:                           MLE   Df Model:                           27
Date:                Fri, 09 Dec 2022   Pseudo R-squ.:                  0.3292
Time:                        00:23:01   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

- Our logistic regression model has a high accuracy.
- We will remove predictor variables with high p-values and rerun our model to check if it improves performance

Model Building - Logistic Regression

Great Learning
POWER AHEAD

Multicollinearity

```
                           Logit Regression Results
==============================================================================
Dep. Variable:         booking_status   No. Observations:                25392
Model:                          Logit   Df Residuals:                    25370
Method:                           MLE   Df Model:                           21
Date:                Fri, 09 Dec 2022   Pseudo R-squ.:                  0.3282
Time:                        00:23:09   Log-Likelihood:                -10810.
converged:                       True   LL-Null:                       -16091.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==================================================================================================
                                     coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
const                           -915.6391    120.471     -7.600      0.000   -1151.758    -679.520
no_of_adults                       0.1088      0.037      2.914      0.004       0.036       0.182
no_of_children                     0.1531      0.062      2.470      0.014       0.032       0.275
no_of_weekend_nights               0.1086      0.020      5.498      0.000       0.070       0.147
no_of_week_nights                  0.0417      0.012      3.399      0.001       0.018       0.066
required_car_parking_space        -1.5947      0.138    -11.564      0.000      -1.865      -1.324
lead_time                          0.0157      0.000     59.213      0.000       0.015       0.016
arrival_year                       0.4523      0.060      7.576      0.000       0.335       0.569
arrival_month                     -0.0425      0.006     -6.591      0.000      -0.055      -0.030
repeated_guest                    -2.7367      0.557     -4.916      0.000      -3.828      -1.646
no_of_previous_cancellations       0.2288      0.077      2.983      0.003       0.078       0.379
avg_price_per_room                 0.0192      0.001     26.336      0.000       0.018       0.021
no_of_special_requests            -1.4698      0.030    -48.884      0.000      -1.529      -1.411
type_of_meal_plan_Meal Plan 2      0.1642      0.067      2.469      0.014       0.034       0.295
type_of_meal_plan_Not Selected     0.2860      0.053      5.406      0.000       0.182       0.390
room_type_reserved_Room_Type 2    -0.3552      0.131     -2.709      0.007      -0.612      -0.098
room_type_reserved_Room_Type 4    -0.2828      0.053     -5.330      0.000      -0.387      -0.179
room_type_reserved_Room_Type 5    -0.7364      0.208     -3.535      0.000      -1.145      -0.328
room_type_reserved_Room_Type 6    -0.9682      0.151     -6.403      0.000      -1.265      -0.672
room_type_reserved_Room_Type 7    -1.4343      0.293     -4.892      0.000      -2.009      -0.860
market_segment_type_Corporate     -0.7913      0.103     -7.692      0.000      -0.993      -0.590
market_segment_type_Offline       -1.7854      0.052    -34.363      0.000      -1.887      -1.684
==================================================================================================
```

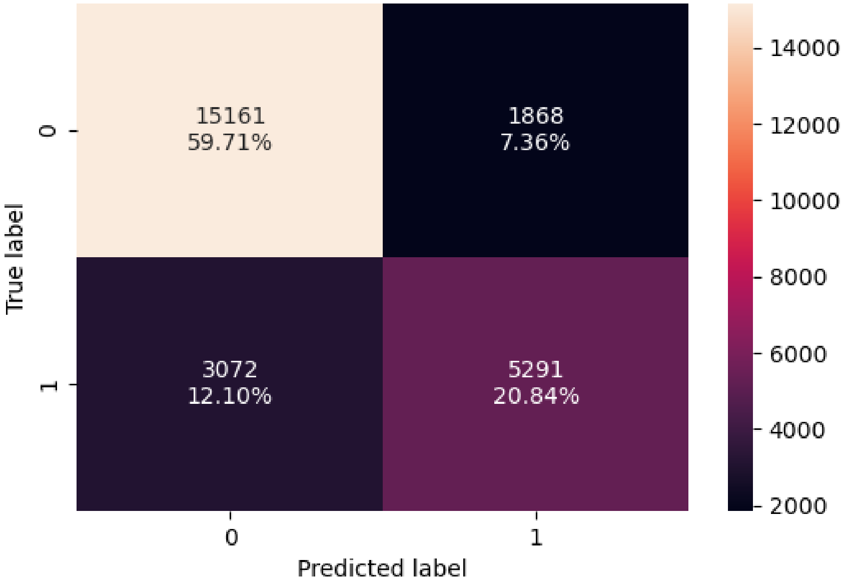
Training performance:

	Accuracy	Recall	Precision	F1
0	0.80545	0.63267	0.73907	0.68174

- Removal of predictor variables with high p-values did not affect model accuracy.

Model Building - Logistic Regression

Confusion Matrix after adding coefficients and odds



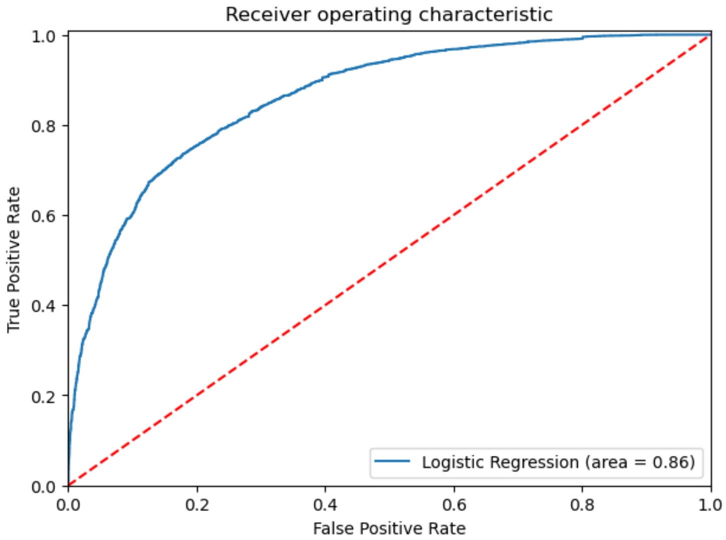
Performance check on training set

	Accuracy	Recall	Precision	F1
0	0.80545	0.63267	0.73907	0.68174

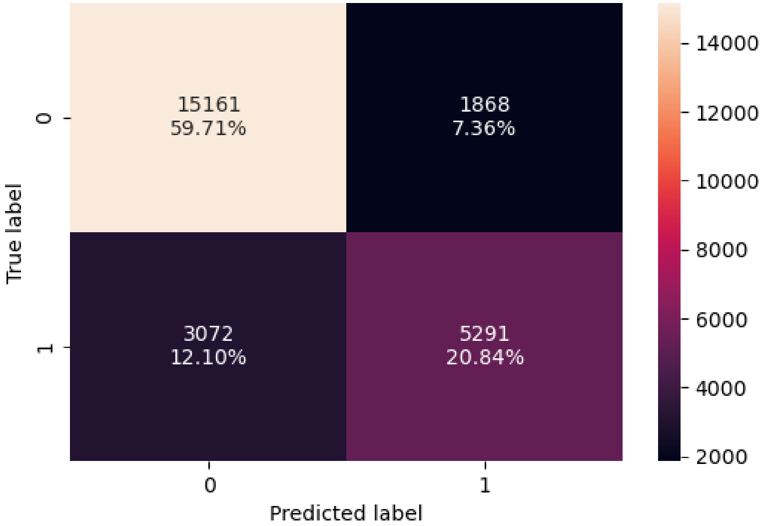
Model Building - Logistic Regression

Performance on training set

ROC-AUC curve



Confusion matrix using 0.37 threshold

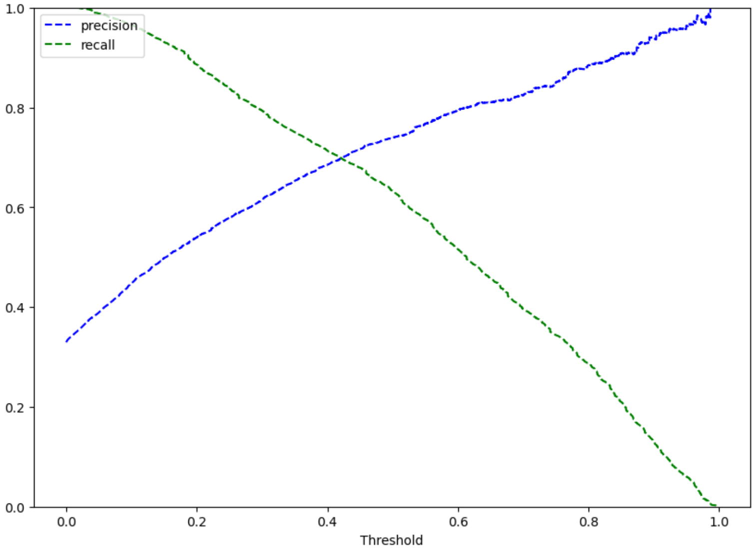


	Accuracy	Recall	Precision	F1
0	0.79265	0.73622	0.66808	0.70049

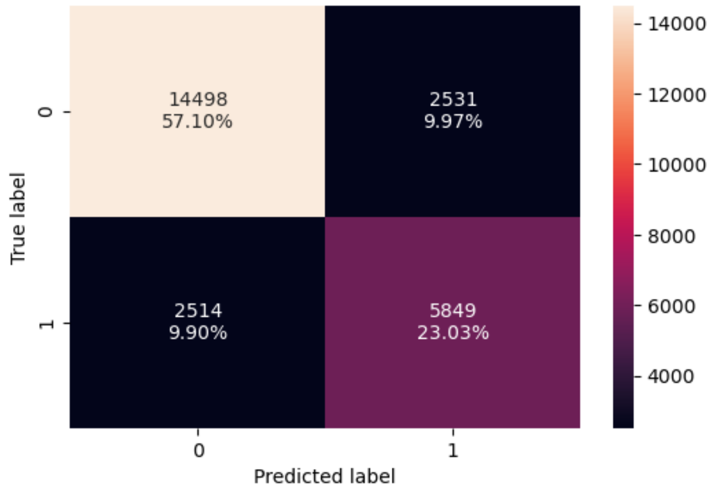
Model Building - Logistic Regression

Performance on training set

Precision-Recall curve



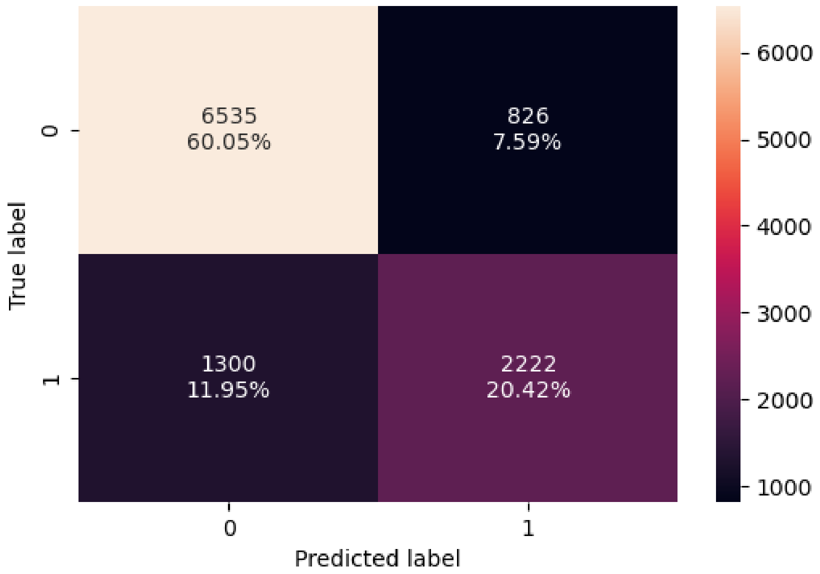
Confusion matrix at 0.42 threshold



	Accuracy	Recall	Precision	F1
0	0.80132	0.69939	0.69797	0.69868

Model Building - Logistic Regression

Performance check on the test set

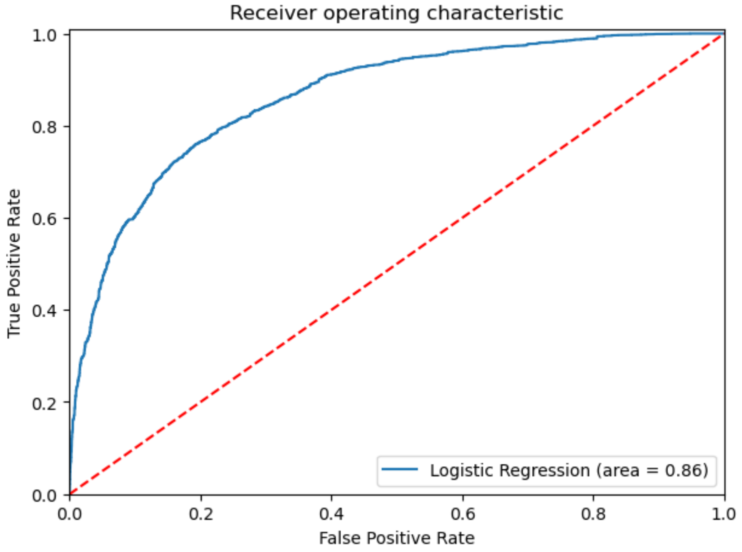


	Accuracy	Recall	Precision	F1
0	0.80465	0.63089	0.72900	0.67641

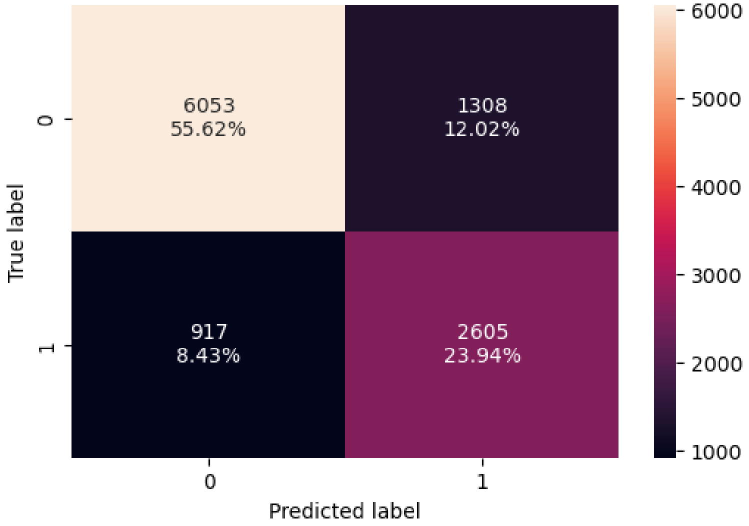
Model Building - Logistic Regression

Performance on test set

ROC-AUC curve



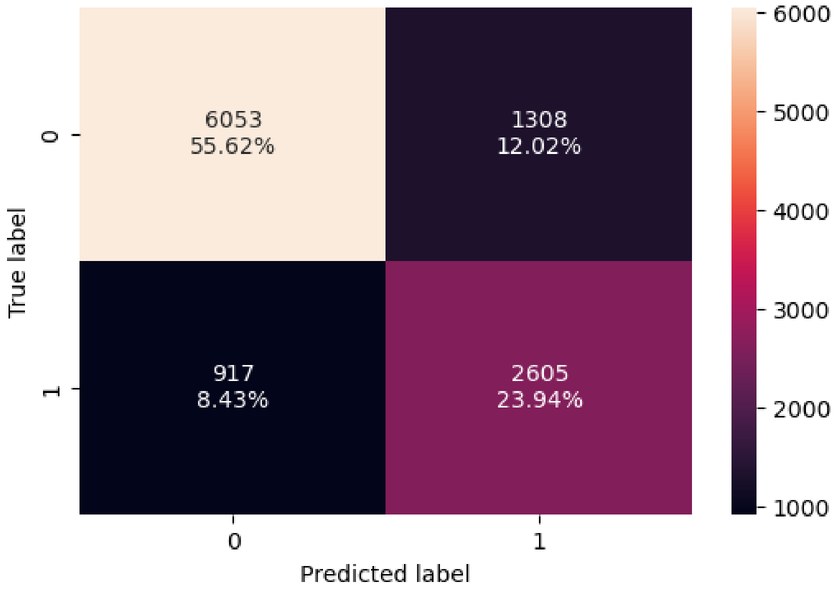
Confusion matrix at 0.37 threshold



	Accuracy	Recall	Precision	F1
0	0.79555	0.73964	0.66573	0.70074

Model Building - Logistic Regression

Confusion matrix at 0.42 threshold



Performance check on the test set

	Accuracy	Recall	Precision	F1
0	0.80345	0.70358	0.69353	0.69852

Model Performance Summary

- A logistic regression model performs well in classifying bookings status.
- No. of adults, no. of children, lead time, repeated guests, average price per room, no. of weekend nights are the most important contributing factors distinguishing canceled bookings from the retained ones.
- Removal of non-significant predictor variables did not improve model performance.
- Varying the threshold for the classifier also did not impact performance

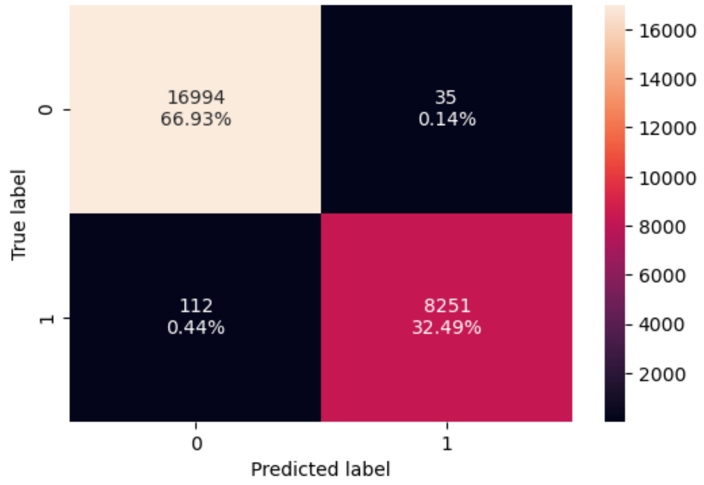
Model Performance Summary

	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

	Logistic Regression statsmodel	Logistic Regression-0.27 Threshold	Logistic Regression-0.36 Threshold
Accuracy	0.80465	0.79555	0.80345
Recall	0.63089	0.73964	0.70358
Precision	0.72900	0.66573	0.69353
F1	0.67641	0.70074	0.69852

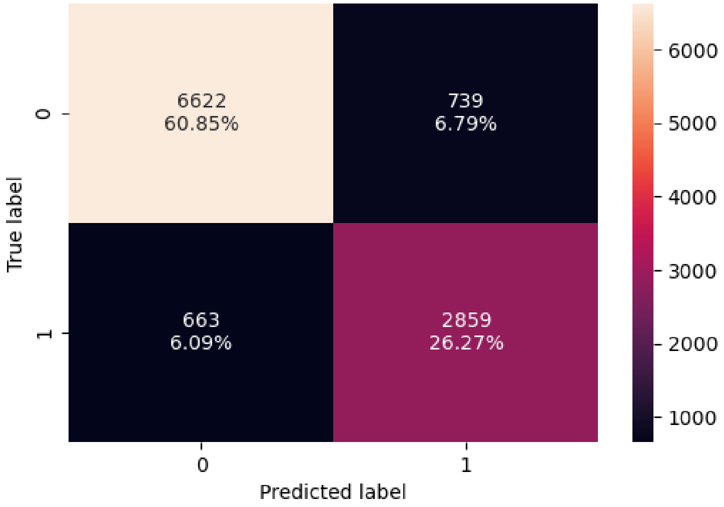
Decision Tree

Performance check on the train set



	Accuracy	Recall	Precision	F1
0	0.99421	0.98661	0.99578	0.99117

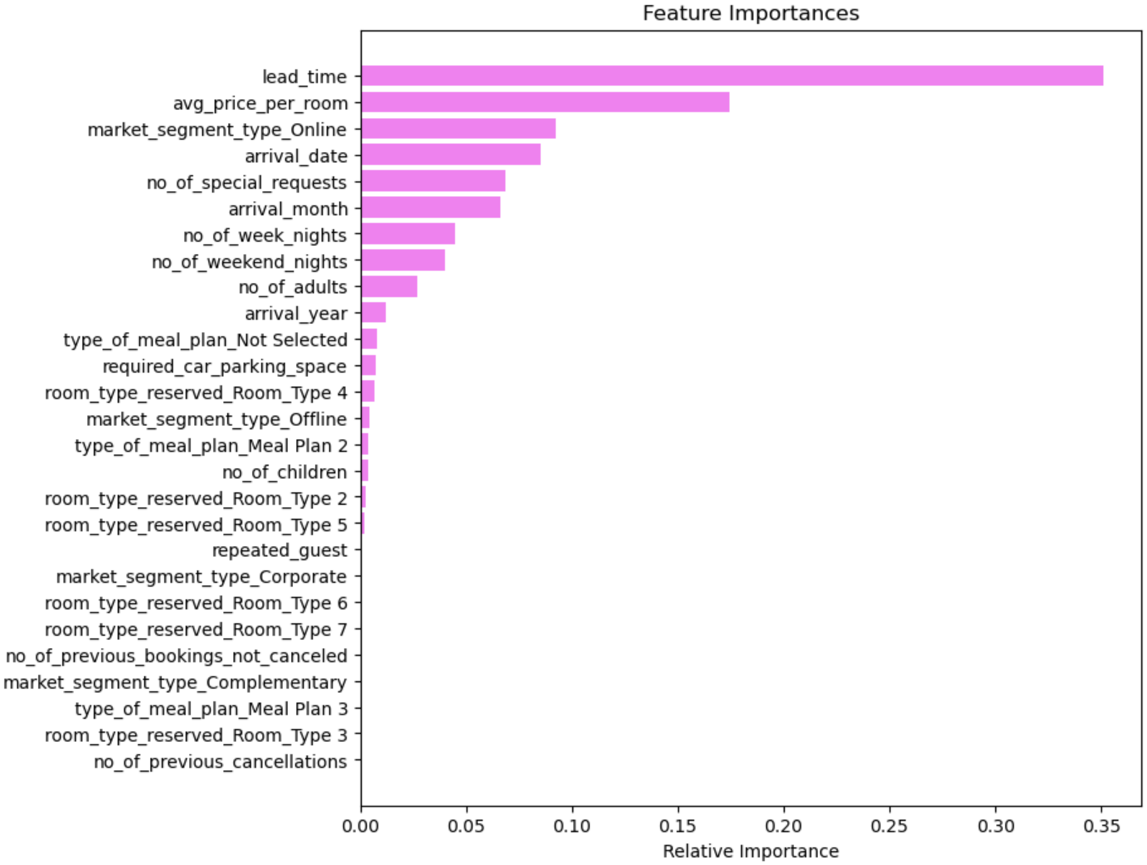
Performance check on the test set



	Accuracy	Recall	Precision	F1
0	0.87118	0.81175	0.79461	0.80309

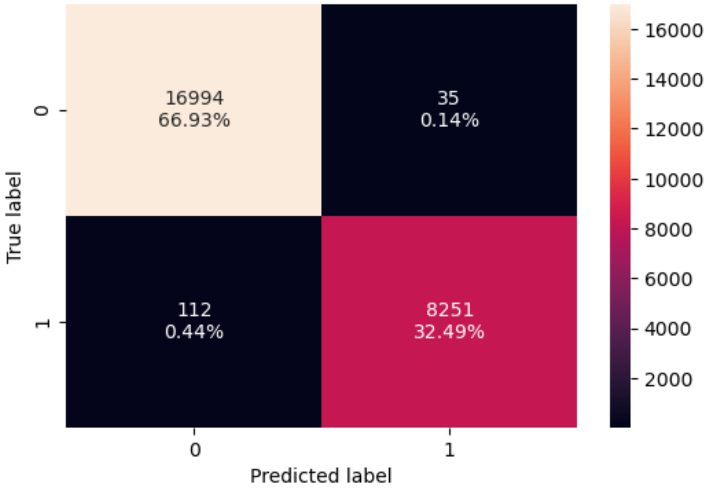
Decision Tree

Important features before pruning of decision tree



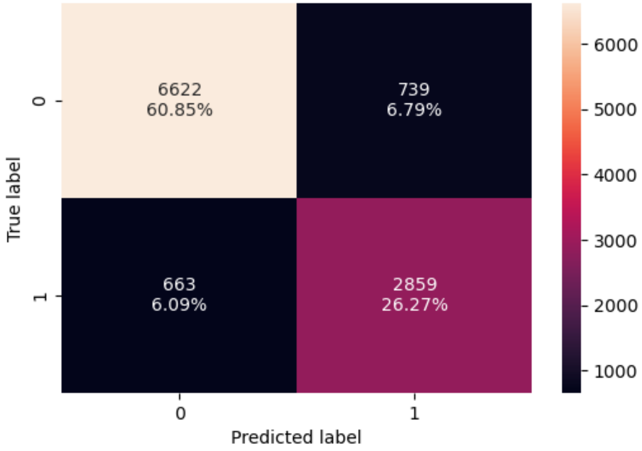
Decision Tree

Performance check on the train set after pruning



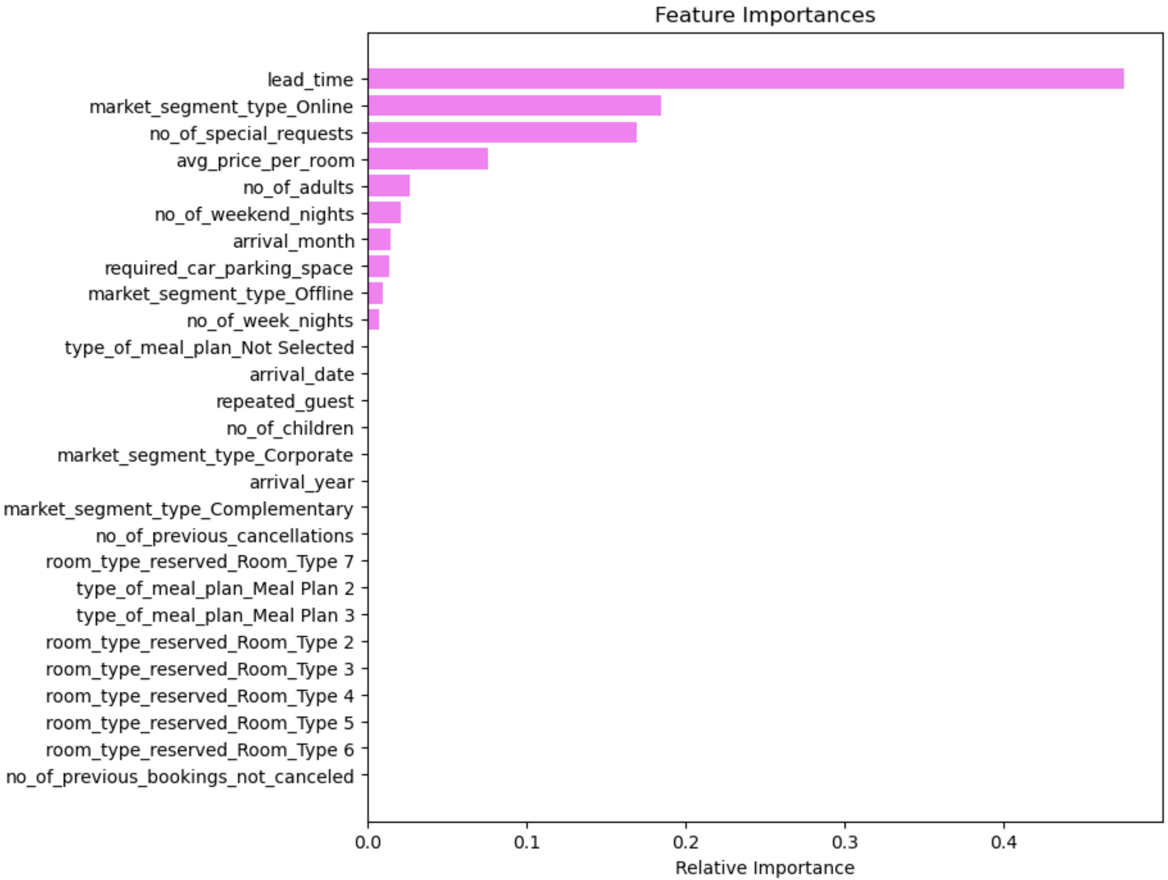
	Accuracy	Recall	Precision	F1
0	0.99421	0.98661	0.99578	0.99117

Performance check on the test set after pruning



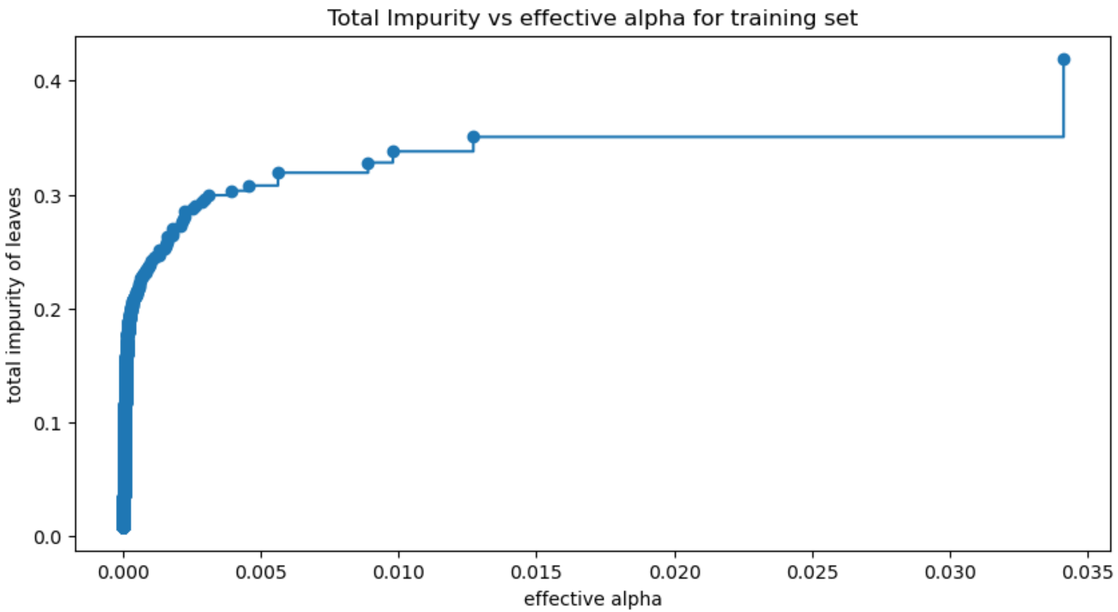
	Accuracy	Recall	Precision	F1
0	0.87118	0.81175	0.79461	0.80309

Decision Tree

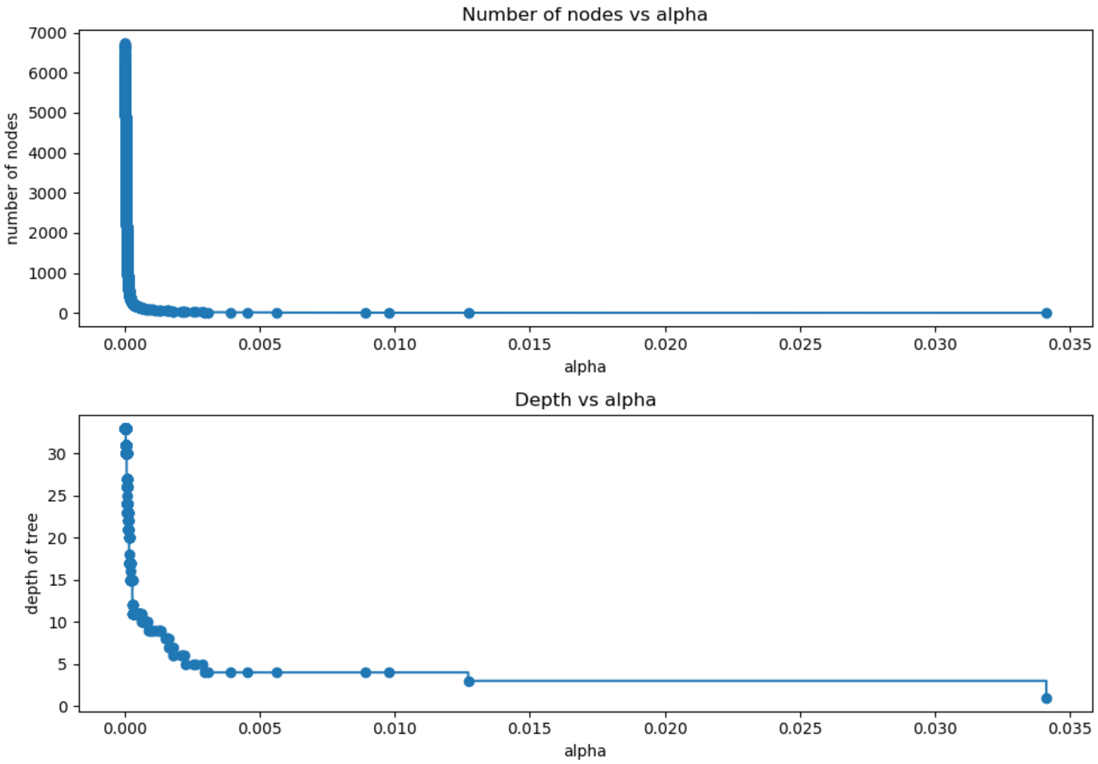


Decision Tree

Cost Complexity Pruning



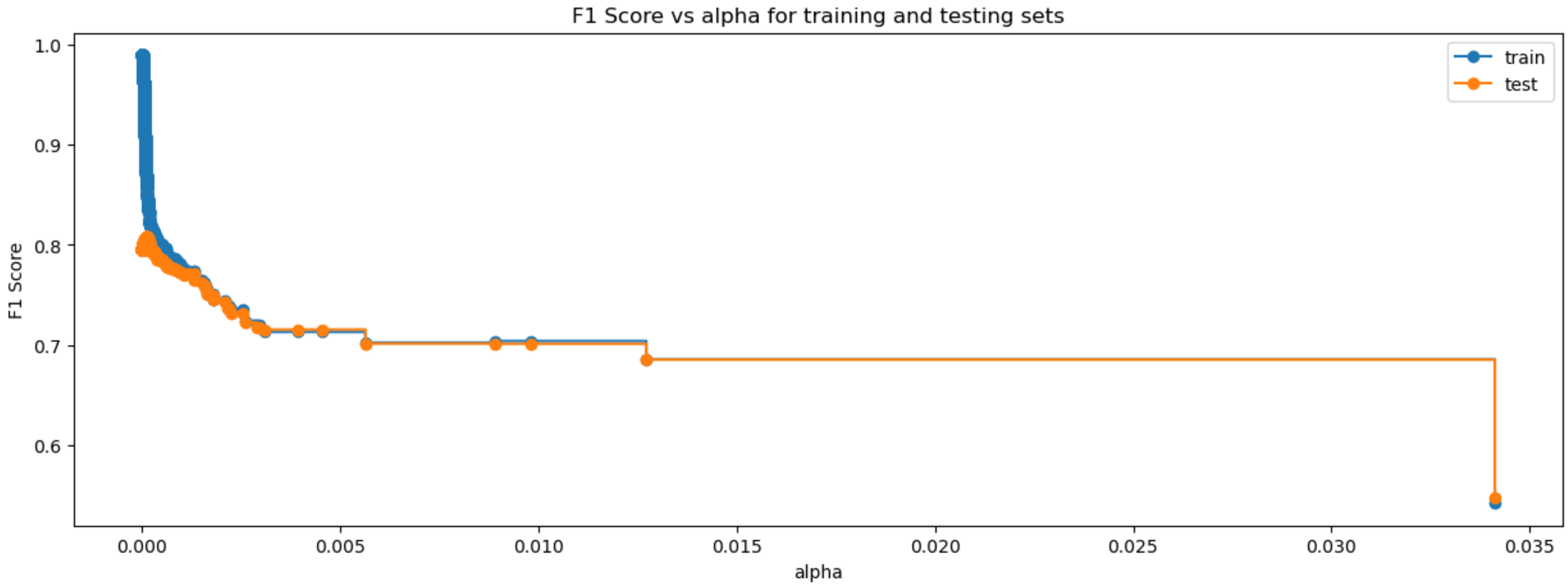
Decision Tree



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

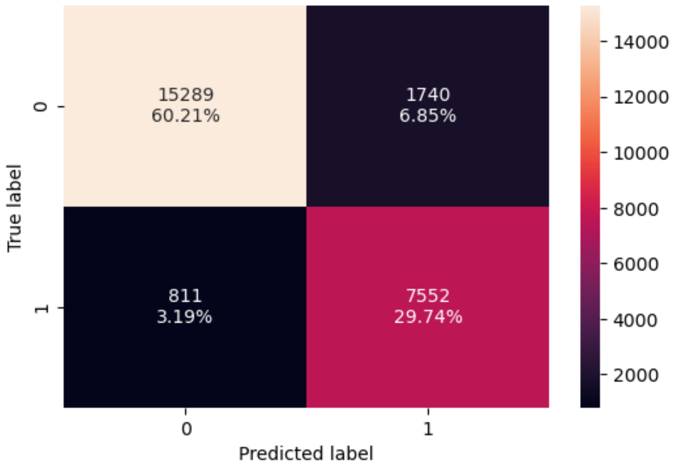
Decision Tree

F1 Score vs alpha for training and testing sets



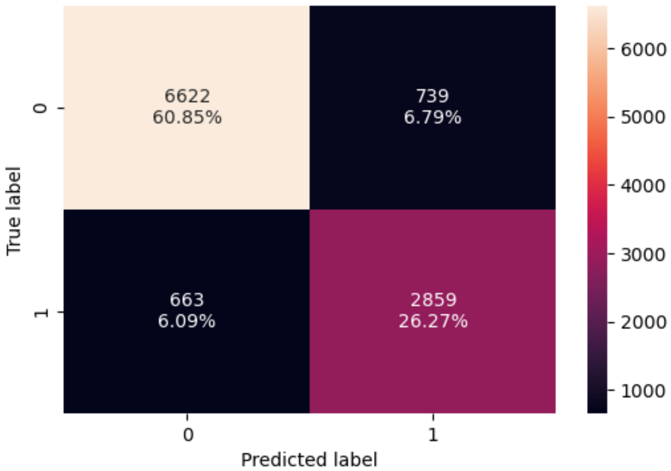
Decision Tree

Performance check on the train set



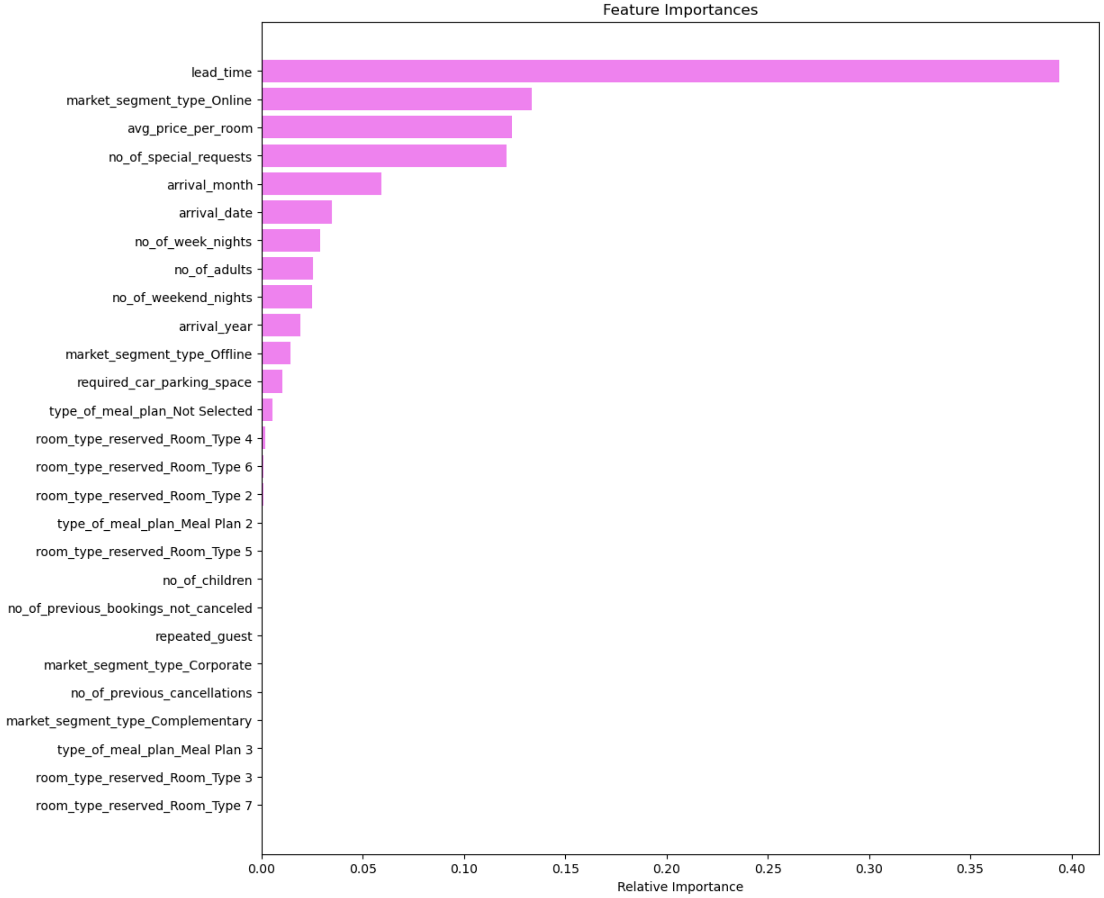
	Accuracy	Recall	Precision	F1
0	0.89954	0.90303	0.81274	0.85551

Performance check on the test set



	Accuracy	Recall	Precision	F1
0	0.87118	0.81175	0.79461	0.80309

Decision Tree



Comparing Decision Tree models

Performance check on the train set

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

Performance check on the test set

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.87118	0.87118	0.87118
Recall	0.81175	0.81175	0.81175
Precision	0.79461	0.79461	0.79461
F1	0.80309	0.80309	0.80309

Decision Tree

- Training a decision tree is able to predict room cancelations based on the observed variables
- The initial decision tree performed very well on the training set as given by the accuracy score of 0.99 but seems to overfit the data as observed by a classification accuracy score of 0.87.
- To take care of the issue of overfitting we pruned the tree based on the effective alpha value and comparing classification performance between training and test data.
- In our final model we were able to predict cancelation status with close to 90% accuracy in both training and test data sets.
- A plot of feature importance clearly demonstrated lead time as the single most important contributor towards cancelation status.

Business Recommendations

Great Learning
POWER AHEAD

It is likely that cancelation will increase if the room was priced over 100 Euros. This suggests that early booking customers are more likely to cancel booking if a better deal is available at a later date.

- Offer your best room rates before 5 months ahead. After that you may increase your prices slightly and increase profit.
- Require a nonrefundable deposit on all rooms in advance of over 5 months.
- Replace the ‘Full Board’ option on your booking with a menu of special requests available.
    - Wi-Fi
    - VIP a champagne toast at sunset your first night.
    - Laundry Bag
    - Slippers
    - Room upgrades

Business Recommendations

- I believe that seasonal high prices may peak from Aug to early in October.
- Online booking are the highest priced despite also having the highest number of free rooms. Aviation, Offline, and Corporate are generally slightly lower priced with Corporate edging out for the lowest.
- The absence of special request increases the likelihood of cancellation, the addition of special request begins to reduce the likelihood of cancellation at one and progressively reduces cancellation to Zero on the instance of a third request.

Great Learning

Happy Learning !