INN Hotels Project

Dharitri Bollini

04-14-2023

Contents / Agenda

- Business problem Overview
- Data Overview
- EDA - Results
- Data Preprocessing
- Model Performance Summary
- Business Insights & Recommendations
- Appendix

Business Problem Overview

- This data is from a chain of hotels for INN hotel group in Portugal and they are facing increasing number of booking cancellations or no-shows.
- To keep customers satisfied, the hotel makes it easy for guests to cancel their bookings with no fee or small fees. However, the hotels are loosing revenue :
  - ○ Loss of revenue when the hotel cannot resell the room.
  - ○ Additional costs when the hotel pays commissions for publicity to help sell these rooms.
  - ○ Reducing revenue when lowering the pricing last minute to get those room re-booked.
  - ○ Utilizing human resources inefficiently: over-staff or under-staff because last minute cancellations or no-shows make it difficult to estimate the efficiency amount of human resources.
- The task is to analyze the data provided and develop a strategy to help in formulating profitable policies for cancellations and refunds using a logistic regression model and a decision tree model and identifying factors that highly influence which booking is going to get cancelled in advance.

Data Overview

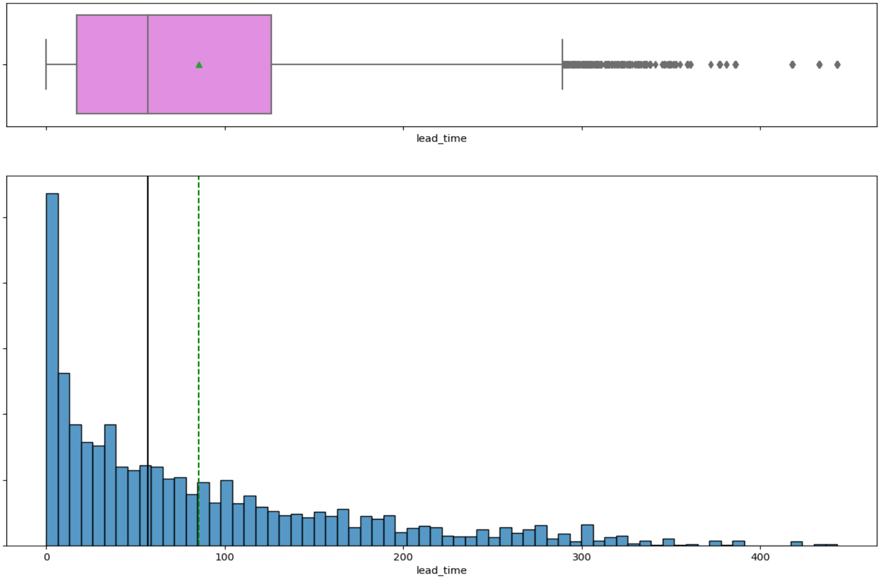
- The data contains 36,275 bookings and 19 customers’ booking details with no missing or duplicated data.
- Booking details include:

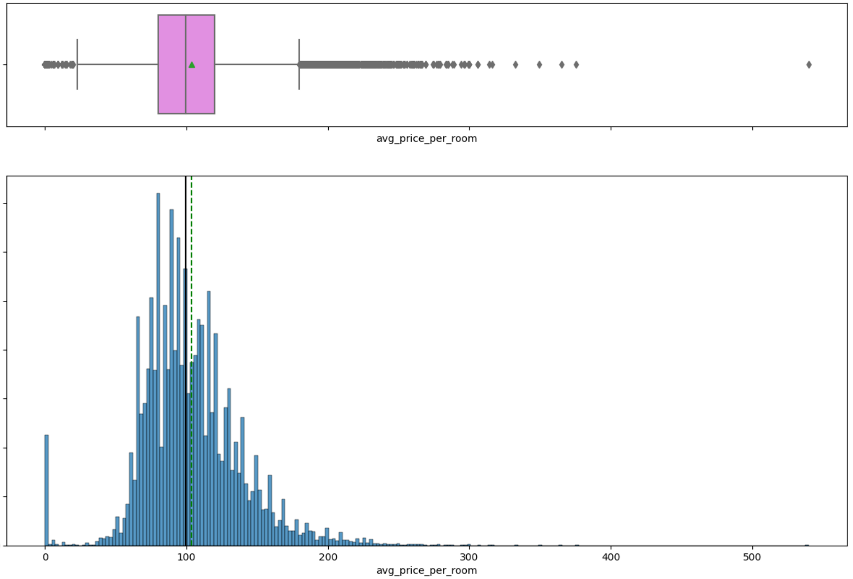
  booking status, average price per room, number of guests, no of weekday
  no of weekend, type of room, lead time from booking to arrival,
  arrival date/month/year, market segment type, repeated guests,
  no of special requests, and more.
- We will use booking status (cancelled and not cancelled) as target.

Data-Dictionary

- 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.
- 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)
- 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.

EDA-Lead Time & Average Price per Room

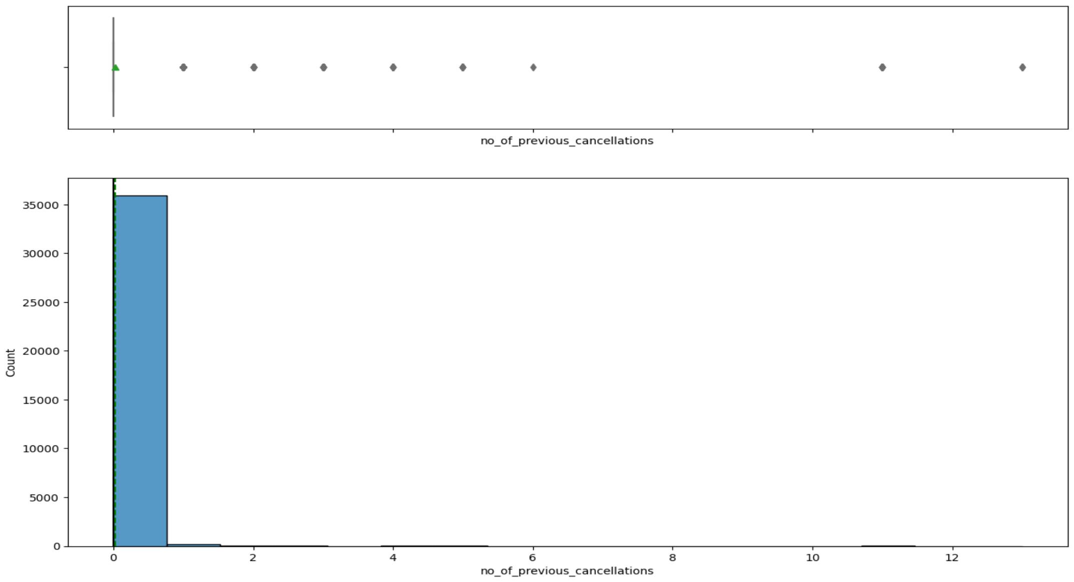


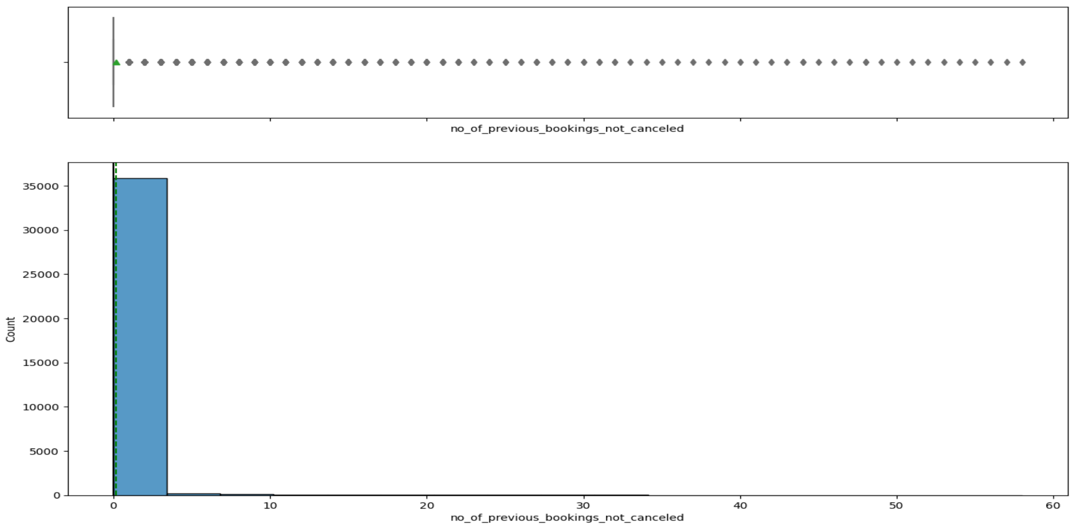


● ***Lead time***, number of days from booked to arrival date, has a heavily right-skewed with lot of outliers. More than 5,000 booking were made on the day of or a few days before an arrival date. Maximum lead time is 443 days and median is 57 days.

● ***Average price per room*** is a right-skewed distribution with lots of outliers. Mean and Median are around 100 Euros.

EDA-Bookings & Cancellations

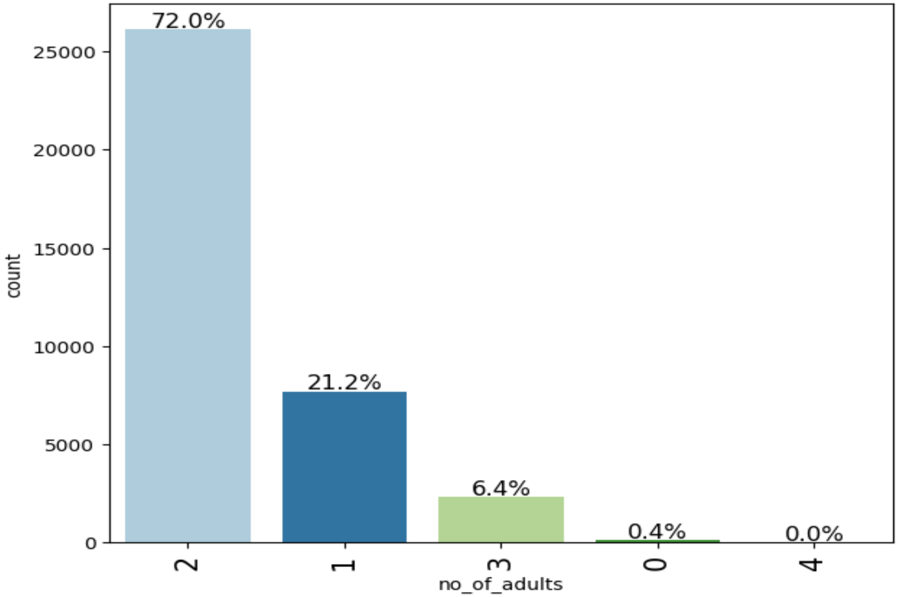


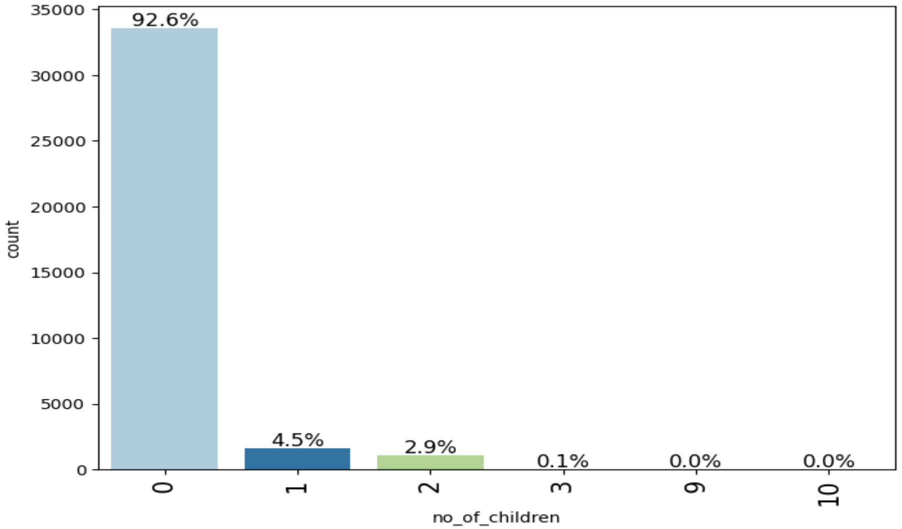


● **No of Previous Cancellations** boxplot shows the number of previous bookings that were cancelled by the same customer. Again they are mostly at zero.

● **No of Previous bookings not cancelled** boxplot shows the number of previous bookings that were not cancelled by the same customer. As we can see, they are usually zero that means they are either a first time customer or did not cancel any bookings.

EDA: Number of Adults & Children

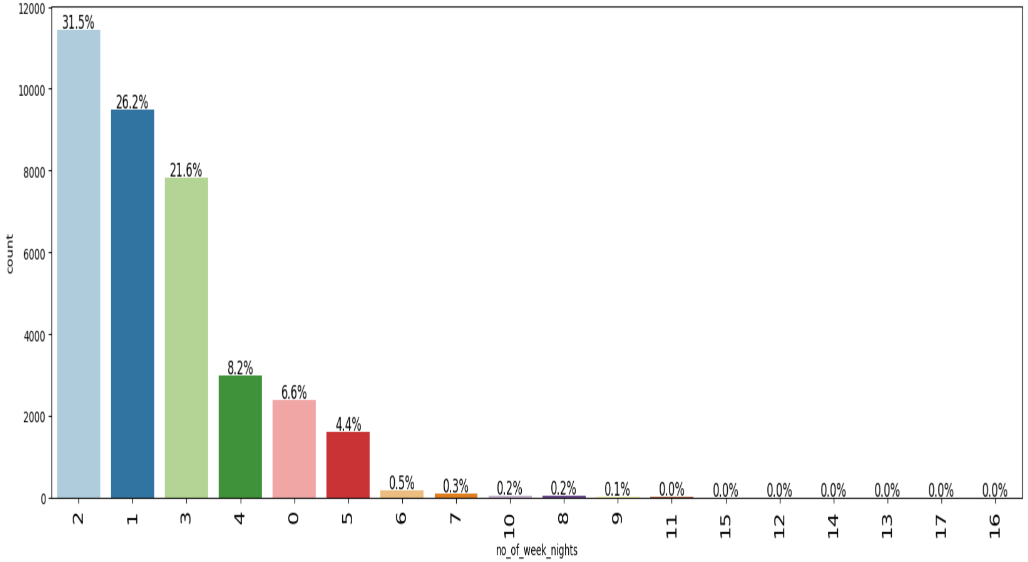


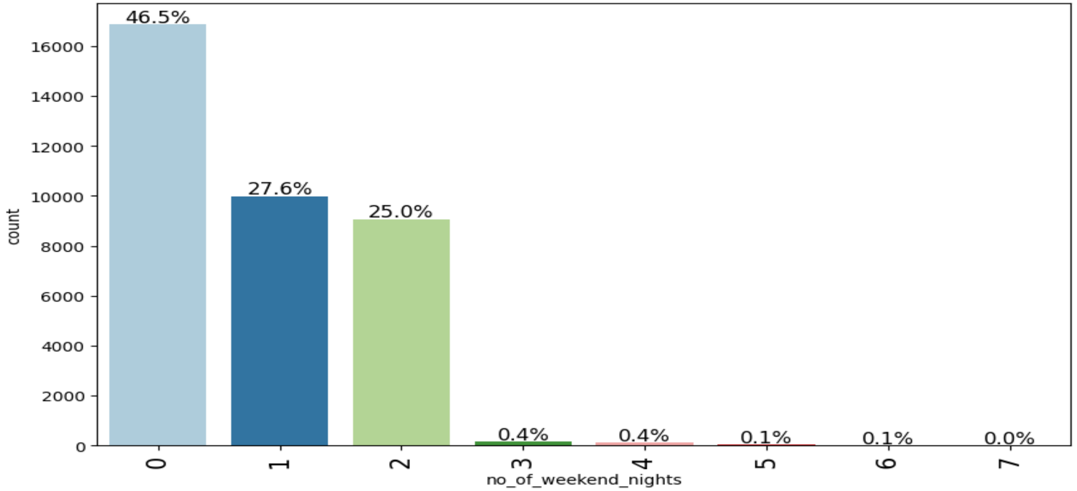


● **No of Adults** booking were72% which were made for 2 adults and 21% was for 1 adult.

● **No of Children** bookings constituted very small amount so accommodating the adults is important.

EDA: Number of Week & Weekend Nights



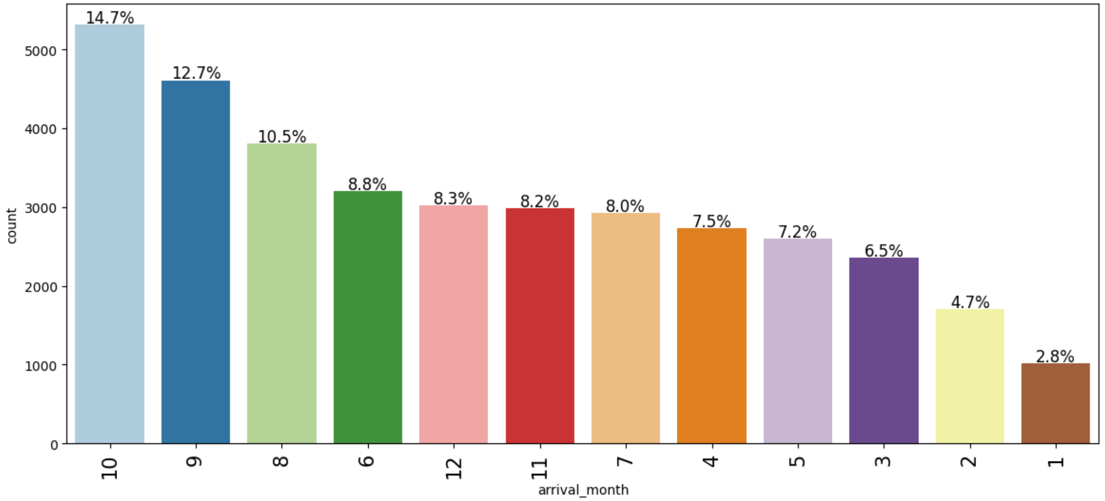


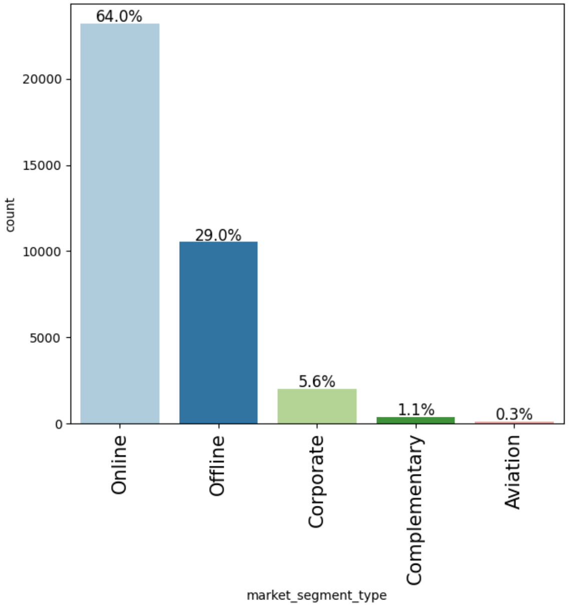
● ***No of weekday nights*** has a range from 0 to 17 days with 2nights as the most frequent bookings, following by 1 day. The bookings that have 0 weekday night, assuming that they are leisure travels.

● ***No of weekend nights*** has a range from 0 to 7 days. The most frequent bookings is 0 nights, assuming customers booked rooms for business. Following by 1 nights and 2 nights.

● A customer is most likely to book a room for business.

EDA: Arrival Months

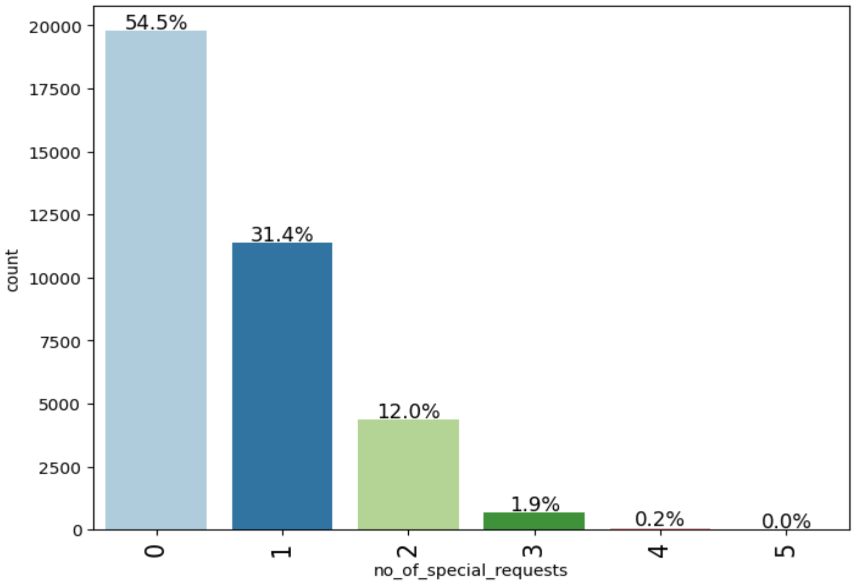


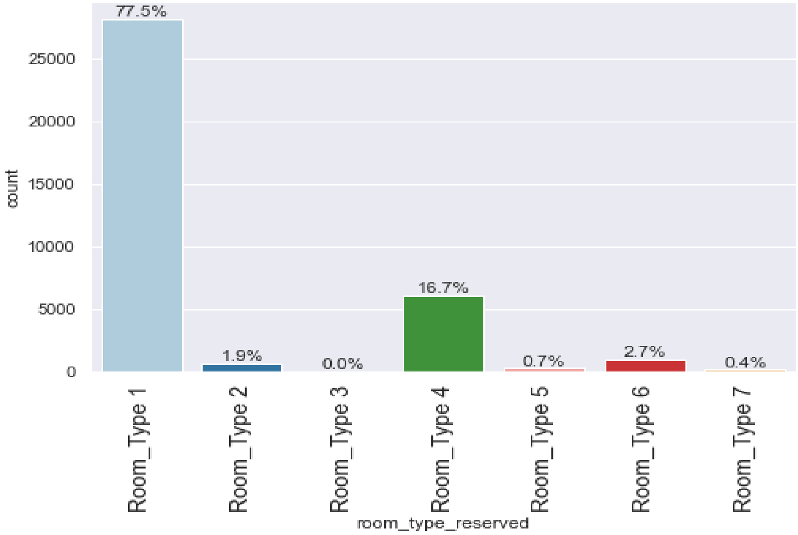


● The top 3 ***arrival months*** are October at 14.7% bookings, September at 12.7% bookings, and August 10.5% bookings. Holidays season in November and December have similar booking counts

● There are 5 type of ***market segments***. Customers made room reservations via online, most convenience. Followed by offline, corporate, complementary and aviation.

EDA: Market segment type and preferences

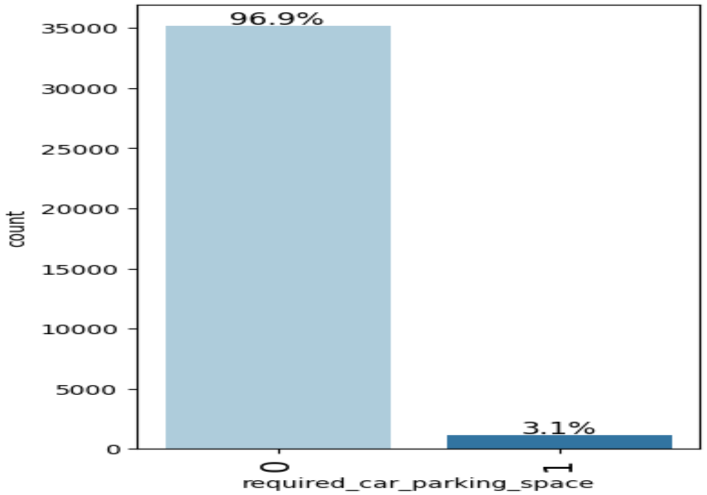


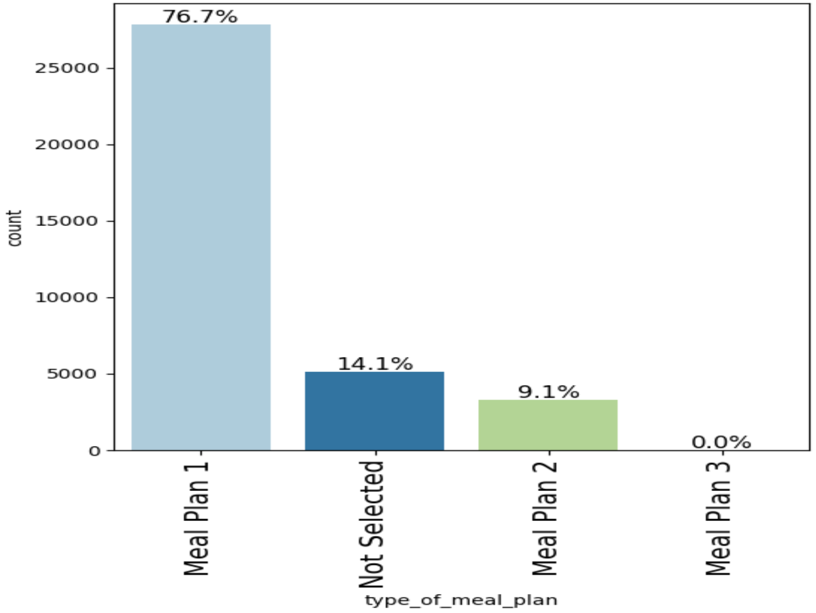


- ***No of special requests*** has a range from 0-5 requests. Most bookings 54.5% bookings, have no special request, followed by 1 request, 2 requests, and the rest.

- ***Room type reserved*** has 7 types of rooms. Most of customers choose room type 1 at 77.5% bookings.

EDA: Meal & Car parking

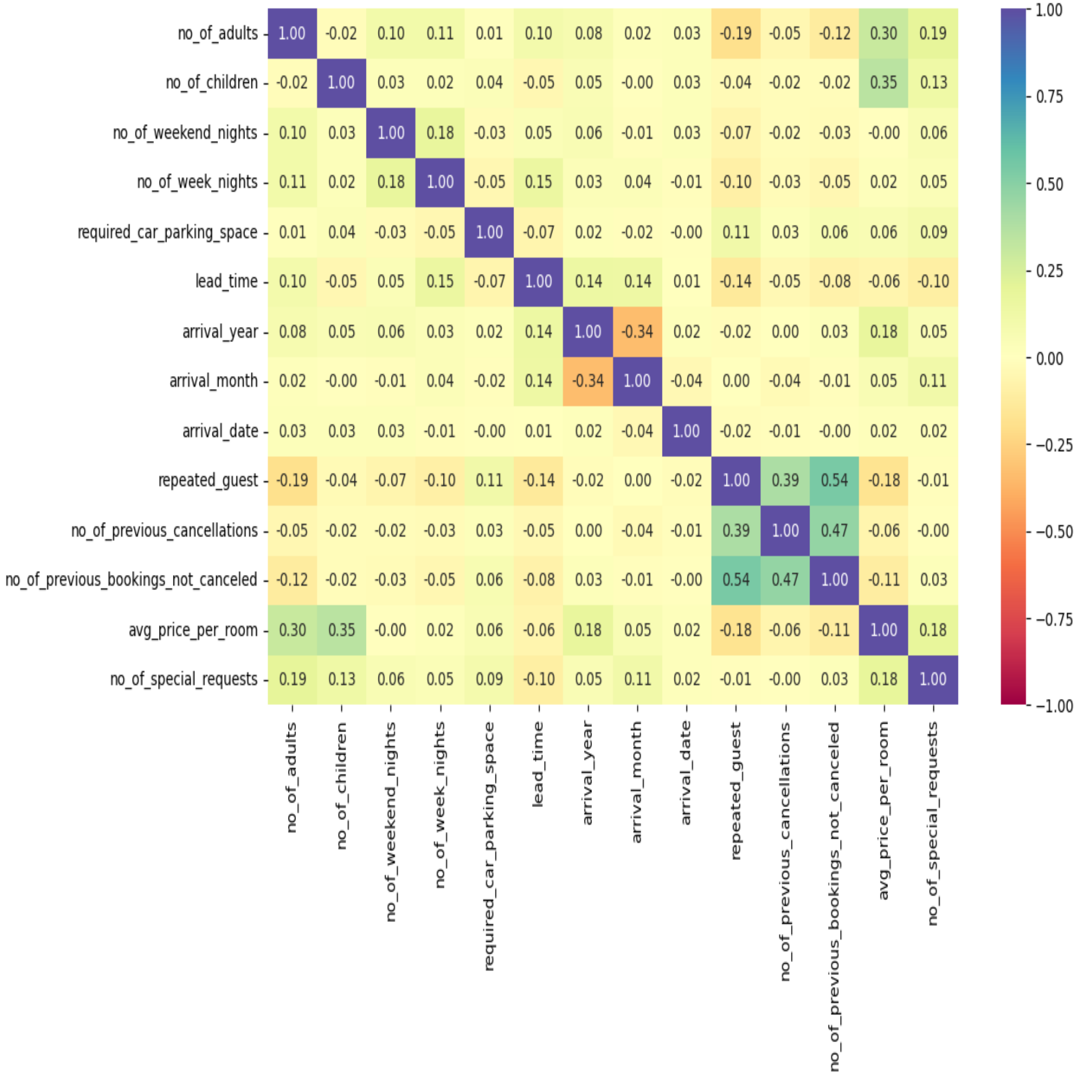




- ***Required Car Parking space** is requested by most guests when they book.*

- ***Type of meal plan*** has 3 types of plans. Most of customers choose meal plan type 1 at 76.7% bookings.

EDA: Bivariate Analysis

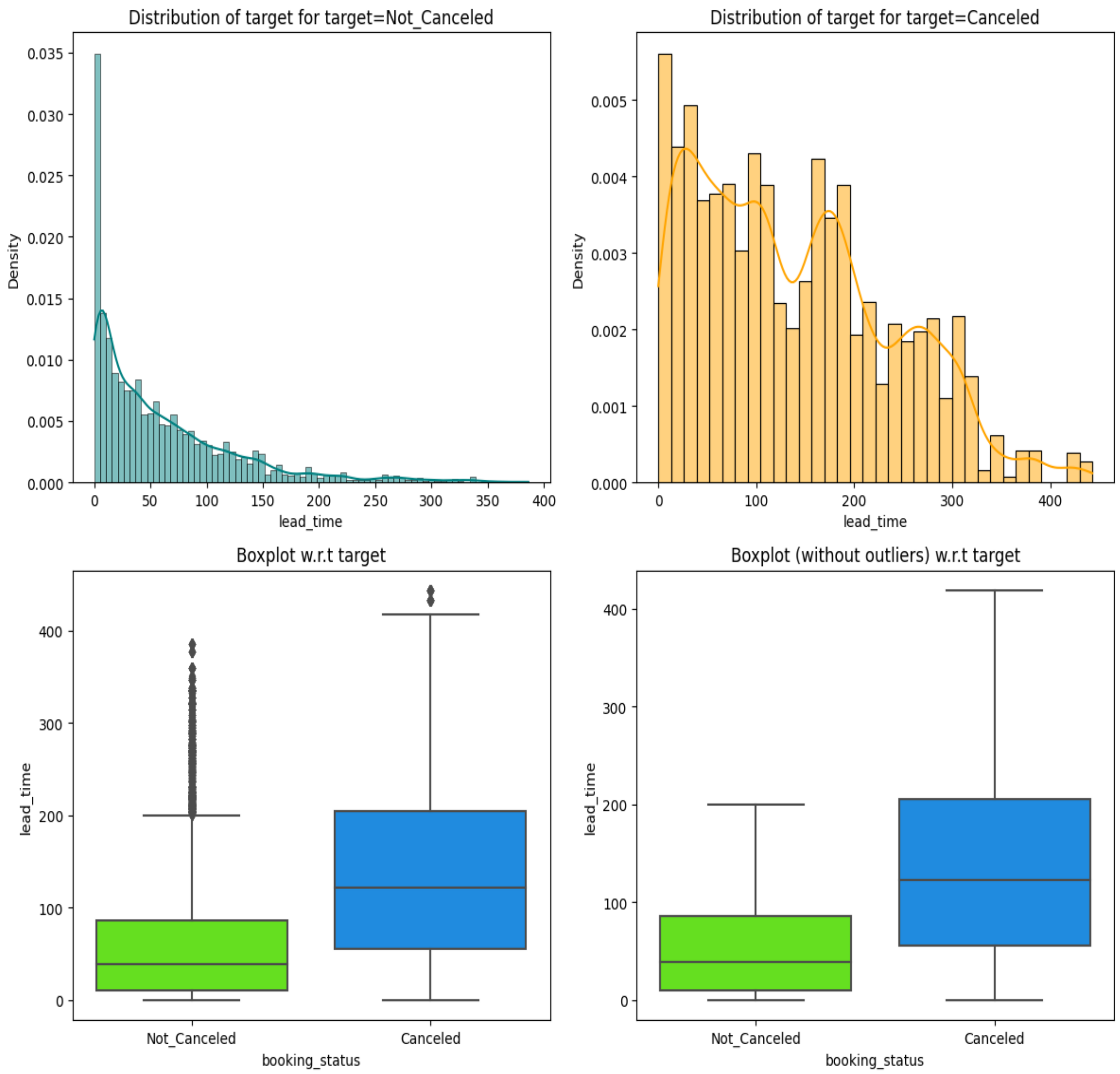


	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	required_car_parking_space	lead_time	arrival_year	arrival_month	arrival_date	repeated_guest	no_of_previous_cancellations	no_of_previous_bookings_not_canceled	avg_price_per_room	no_of_special_requests
no_of_adults	1.00	-0.02	0.10	0.11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

Correlation

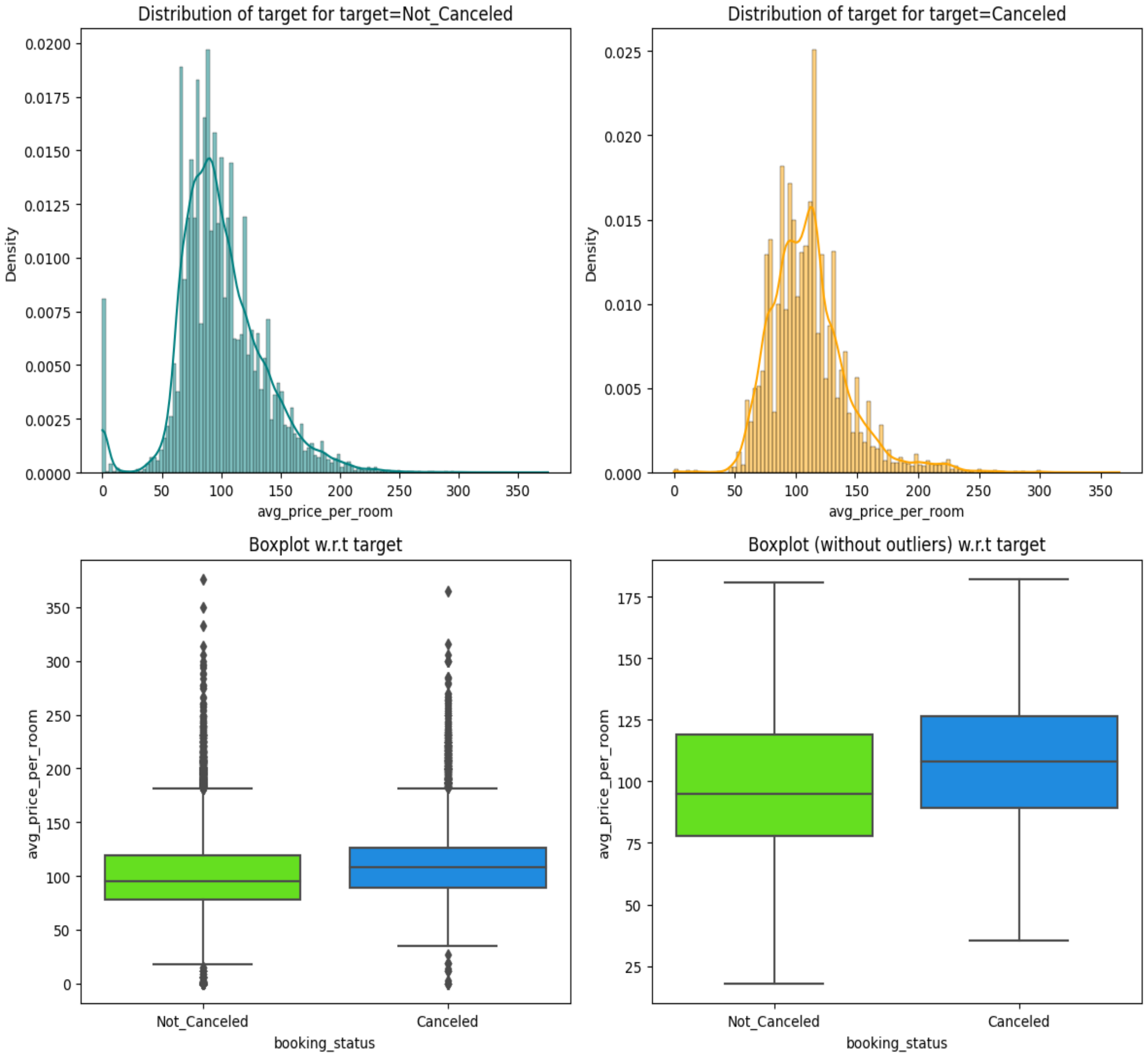
- ***Lead_time*** has the highest correlation with booking_status at 0.44. Following by no_of_special_request at -0.25.
- ***repeated_guest*** has 0.54 correlation with no_of_previous_bookings_not_canceled.
- ***avg_price_per_room*** has 0.30 correlation with no_of_adults and 0.35 correlation with no_of_children.
- ***booking_status*** has 0.14 correlation with avg_price_per_room and -0.11 with Repeated_guest
- We will have to investigate more on the relationship between lead_time and booking_status.

EDA: Bivariate Analysis



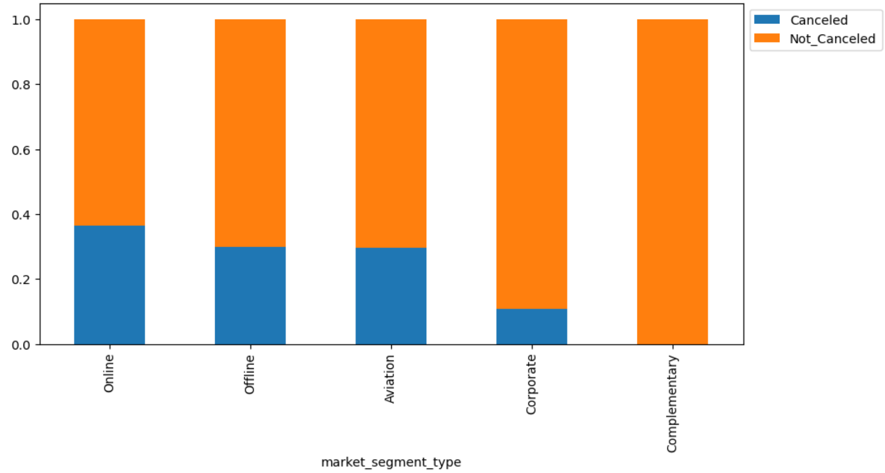
- The median of ***lead_time*** of not Cancelling the booking is around 50 days.
- The median of lead_time of cancelled booking is around 120 days.
- Without outliers, the range for lead_time of not cancelling the booking is 0-200 days and lead_time of cancelled booking is 0-400 days.
- From this observation, the longer the lead_time of booking, the higher chance for cancellations.

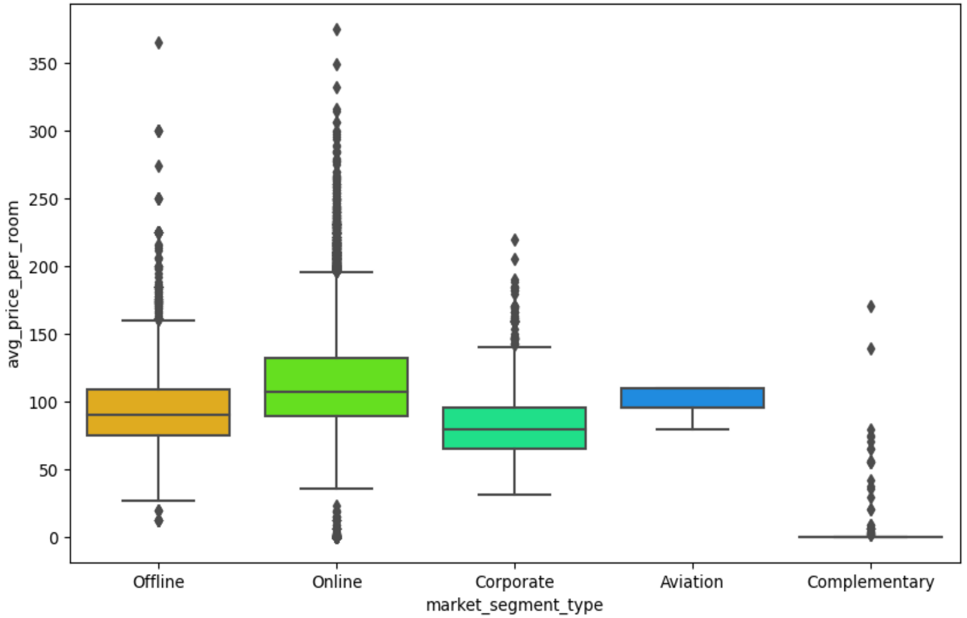
EDA: Bivariate Analysis



- With and without outliers, the ***average price per room*** for cancelled booking is ~110 dollars and for not cancelled bookings is ~95 dollars.
- Customers who cancelled their booking may find a more affordable room from different hotels.
- Customers who did not cancel their booking are satisfied with their room prices and see it as affordable price.

EDA: Bivariate Analysis

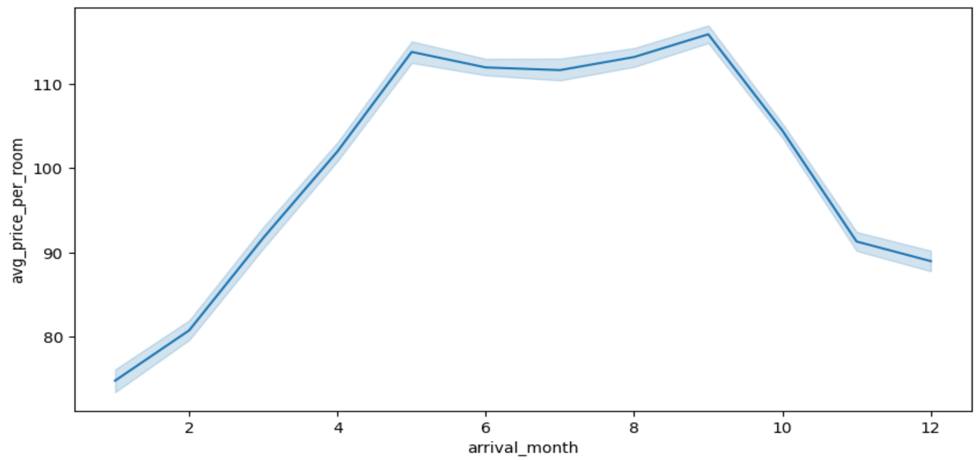


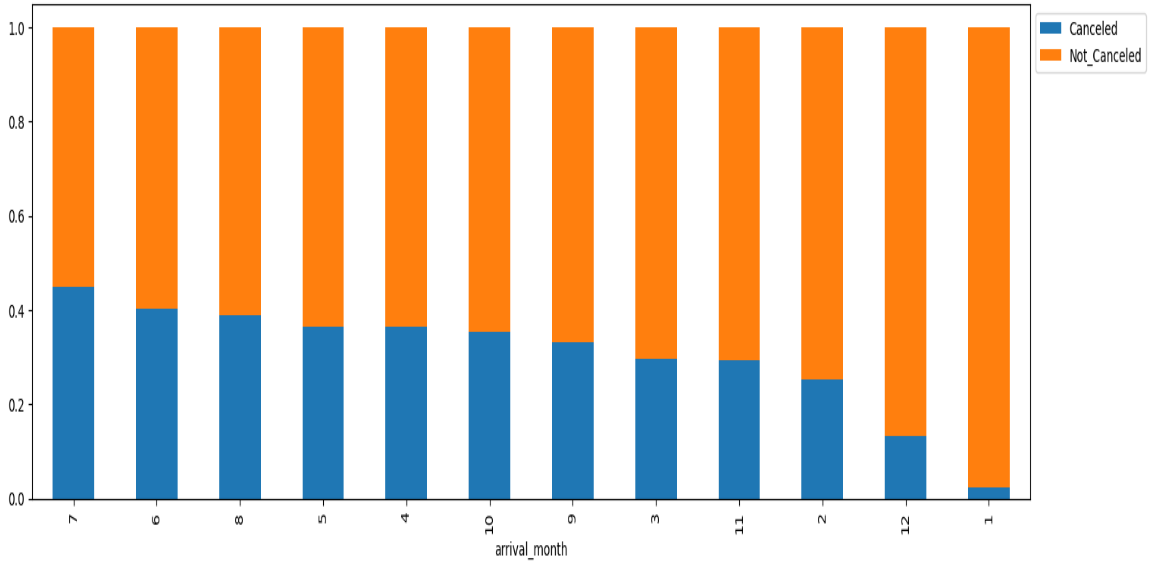


- Observation on ***market segment type***, Online booking has the highest cancellations (~40%), Followed by Offline, Aviation, Corporate and Complementary .Aviation cancellations might be due to flight delays or flights getting cancelled.

- ***Average price per room*** for online bookings has the highest price (over 100 dollars), followed by aviation bookings, offline bookings and complementary (data shown that a large number of bookings are free).

EDA: Bivariate Analysis

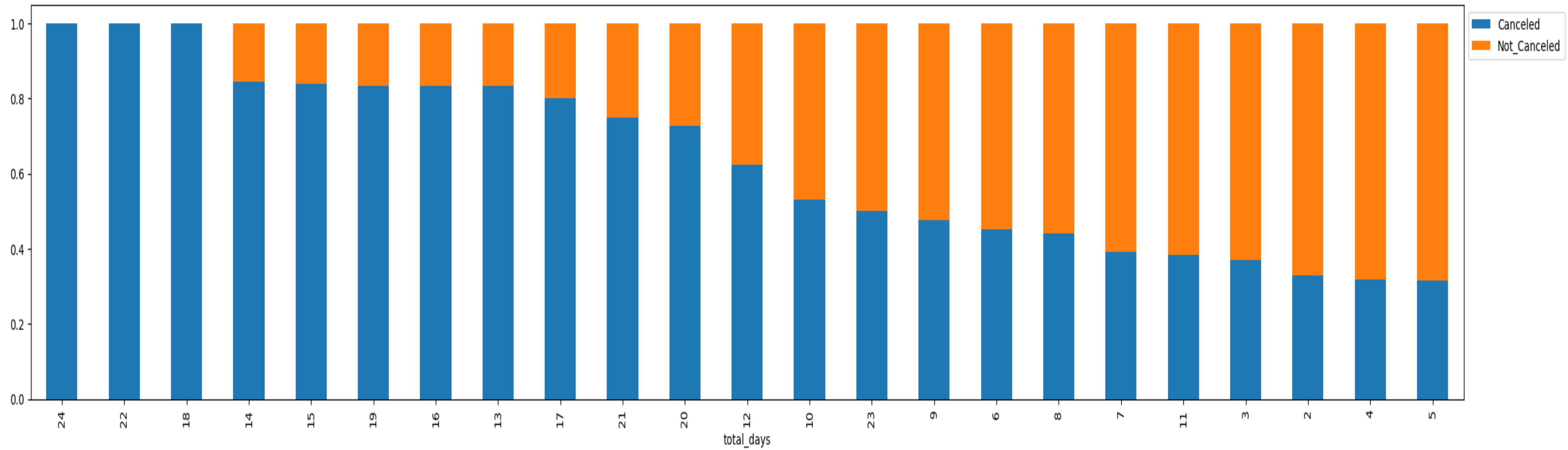




- The most expensive **arrival months** are September, May, August, June and July . The average price in these months are around 110 dollars or higher. Correlating to the most cancelled reservations those months.

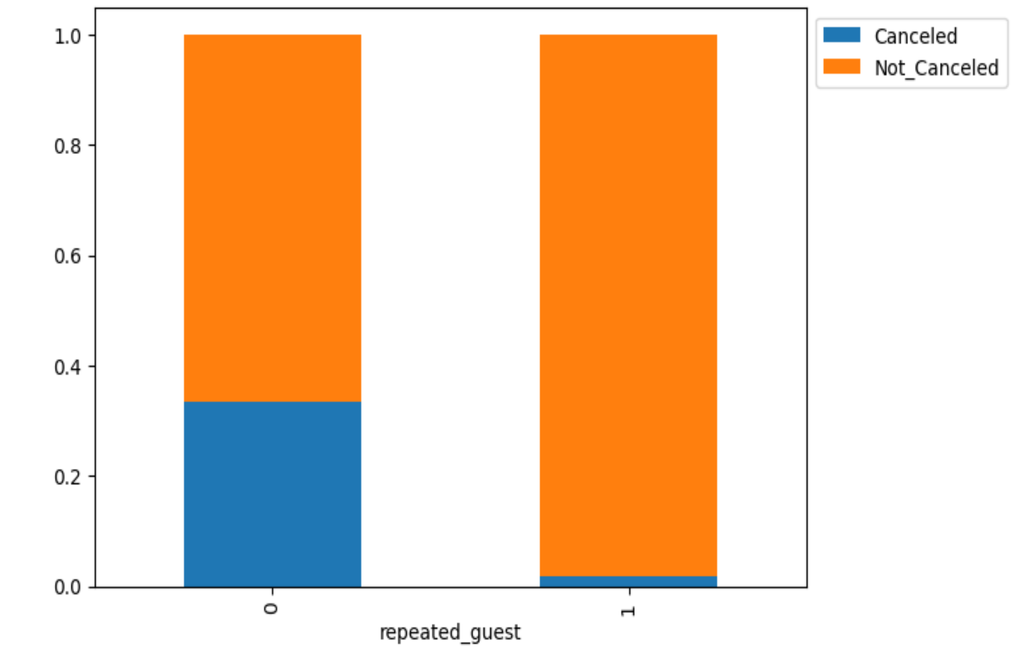
- The least expensive **arrival months** are January, February, March, December, and November. Those prices are around 90 dollars or less. Correlating to the least cancelled reservations in those months.

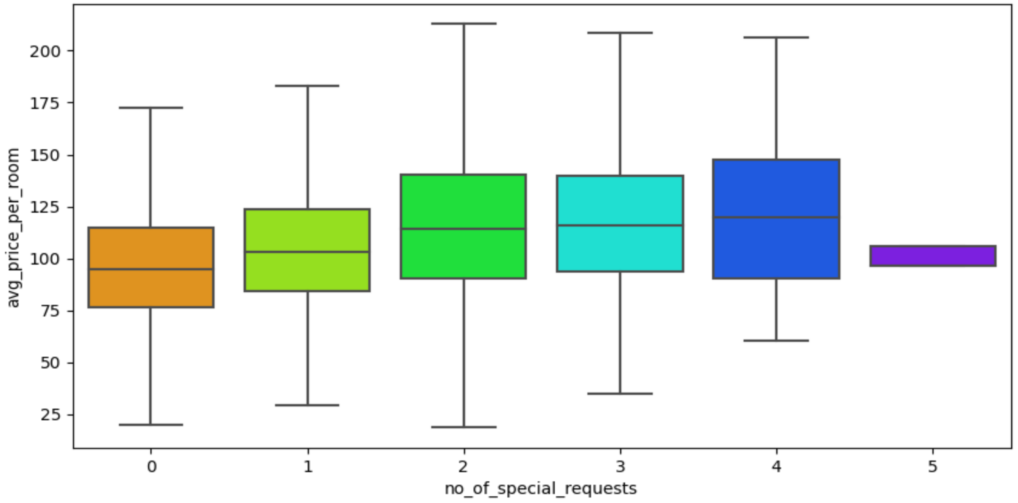
EDA: Bivariate Analysis



- There are 17094 bookings that stay longer than 2 days.
- The range is from 2 days to 24 days.
- The longer time the guests stay, the higher percentage of cancellations occur.
- Bookings for 22-24 days of stay have almost 100 percent chance that the bookings will be canceled.
- Bookings for 2-5 days of stay have the least chance of bookings to be cancelled(~30%).

EDA: Bivariate Analysis

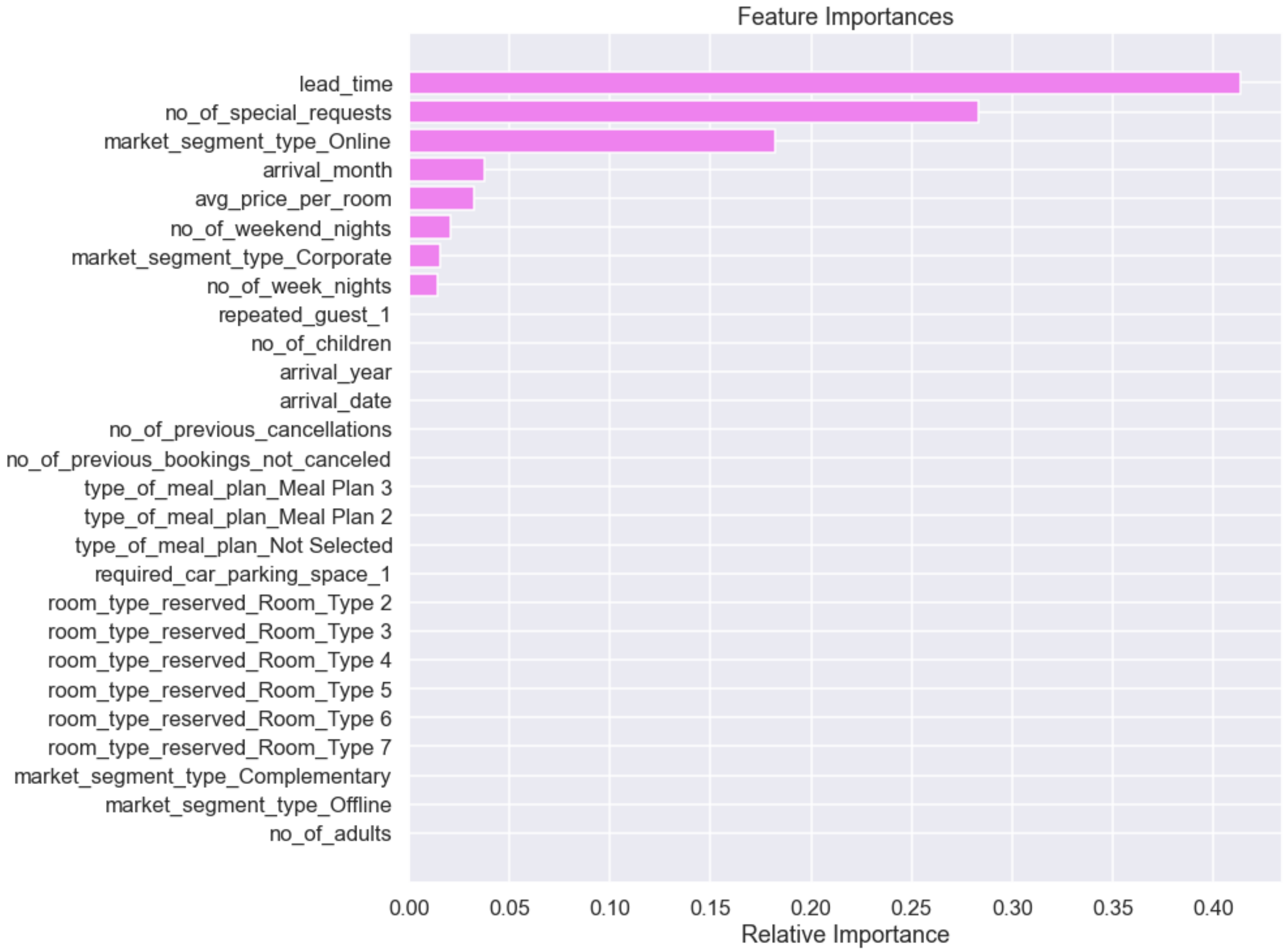




● Bookings of ***repeated customers*** is almost 100% sure that bookings did not get cancelled and for not repeated customers accounts to ~65% cancellations.

● Top 3 of median price of Bookings with ***No of special requests*** is 2, 3, and 4, which is around 120 dollars. For 0 request bookings price is around 90 dollars. All of them have outliers.

Data Preprocessing



● There are no duplicated or missing values.

●The outliers also do not require treating.

Data Preparation

Before we proceed to build a model, we will:

- Drop Booking_ID column. We decided to group booking ID because they are unique numbers. We cannot use it for pattern recognitions.
- Treat outliers:
  - Avg_price_per_room: There are only outliers above upper whisker.
  - Calculated upper whisker which is 179.55 dollars
  - Assigned 179.55 dollars to outliers greater or equal to 500 dollars.
  - No_of_children: We used 3 children to replace the bookings with 9 or 10 children.
- Encode categorical features: Type_of_meal_plan, Room_type_reserved,

  Market_segment_type, and Booking_status
- Split the data into train (70%) and test (30%) to evaluate the model that we build on the train data.

Model Performance Summary

- We want to predict which bookings will be cancelled.
- Model can make wrong prediction as ***false negative*** (predicting a booking to be cancelled when it does not) and ***false positive*** (predicting a booking to not get cancelled when it does).
- We decided that both false negative and false positive are important.

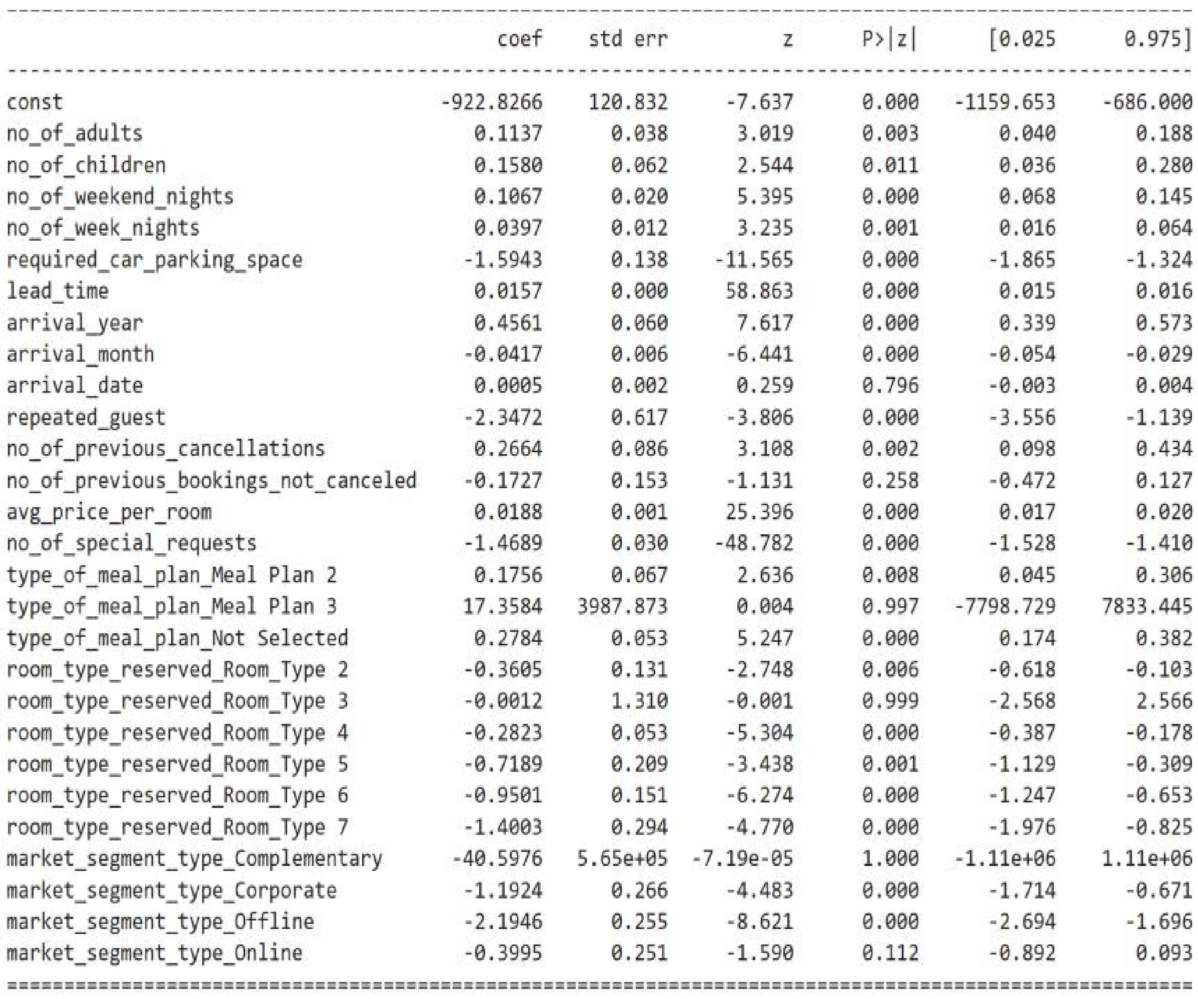
○ ***False negative***: The hotel might not be able to provide satisfactory services to the customer by assuming that this booking will be cancelled. This might damage the brand equity.

○ ***False positive***: The hotel will loose resources and will have to bear additional costs of distribution channels trying to resell the room.

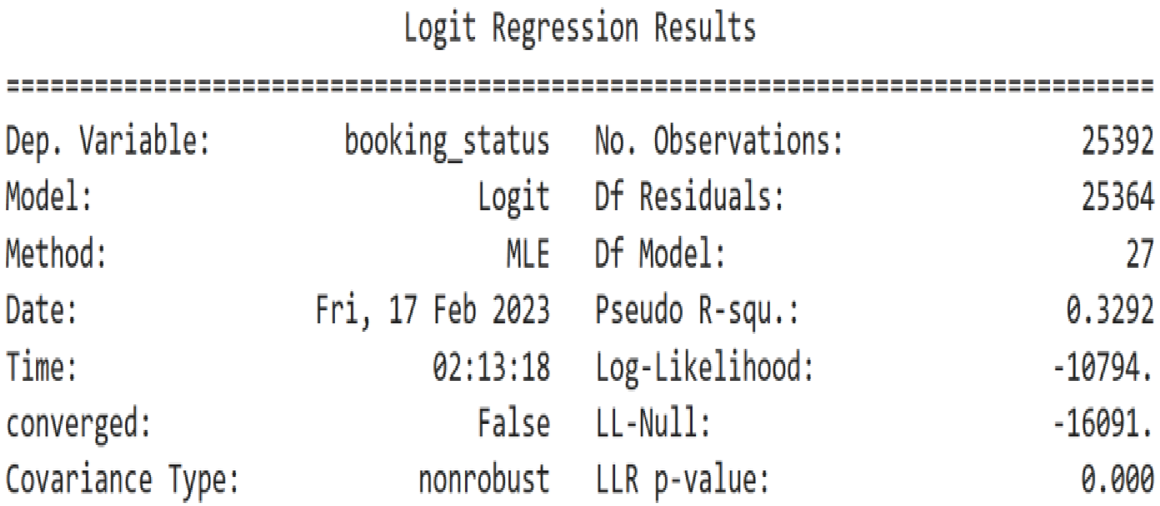
- We want to reduce the losses by which ***F1 Score*** *has* to be maximized for higher chances of minimizing False Negatives and False Positives.
- We will use ***Logistic Regression Model*** and ***Decision Tree Model*** for prediction.

Model Performance Summary

- Data preparation for modeling: we built a logistic regression model

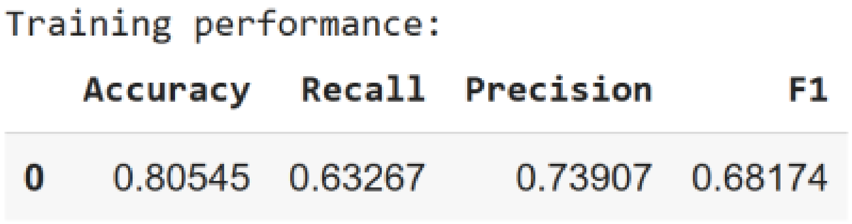


```
==============================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                         -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   3987.873      0.004      0.997   -7798.729    7833.445
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 -40.5976  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
==============================================================================================
```



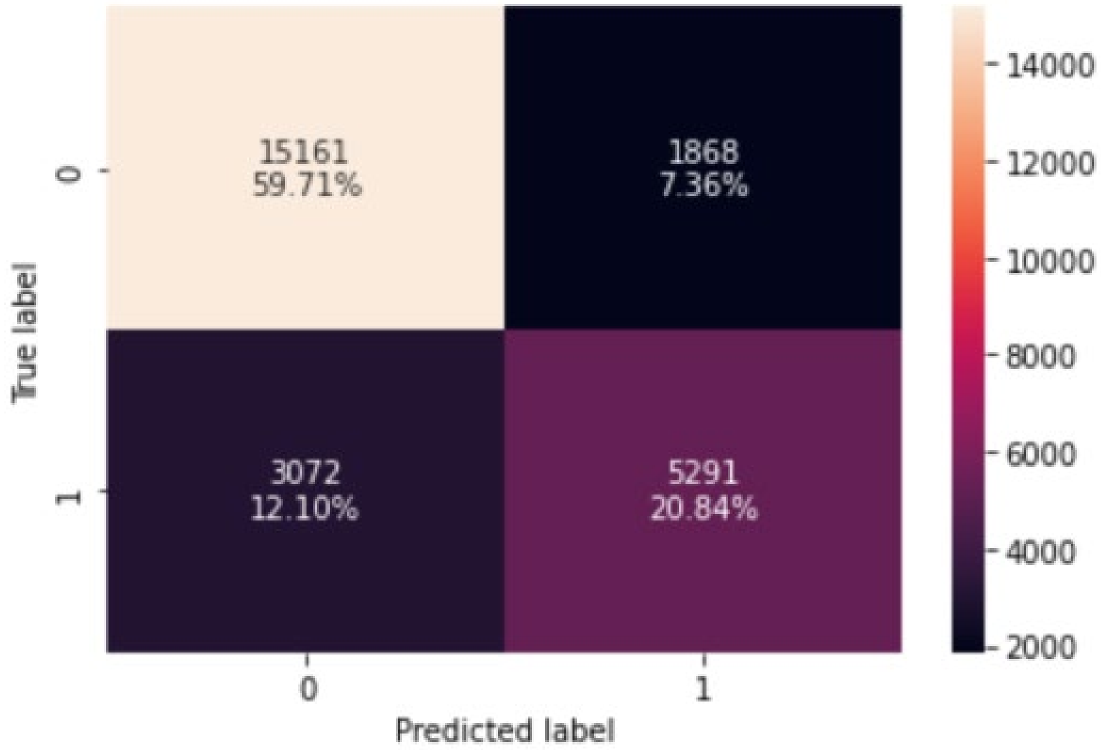
```
                           Logit Regression Results
==============================================================================
Dep. Variable:         booking_status   No. Observations:                25392
Model:                          Logit   Df Residuals:                    25364
Method:                           MLE   Df Model:                           27
Date:                Fri, 17 Feb 2023   Pseudo R-squ.:                  0.3292
Time:                        02:13:18   Log-Likelihood:                -10794.
converged:                      False   LL-Null:                       -16091.
Covariance Type:            nonrobust   LLR p-value:                     0.000
```

Model Performance Summary



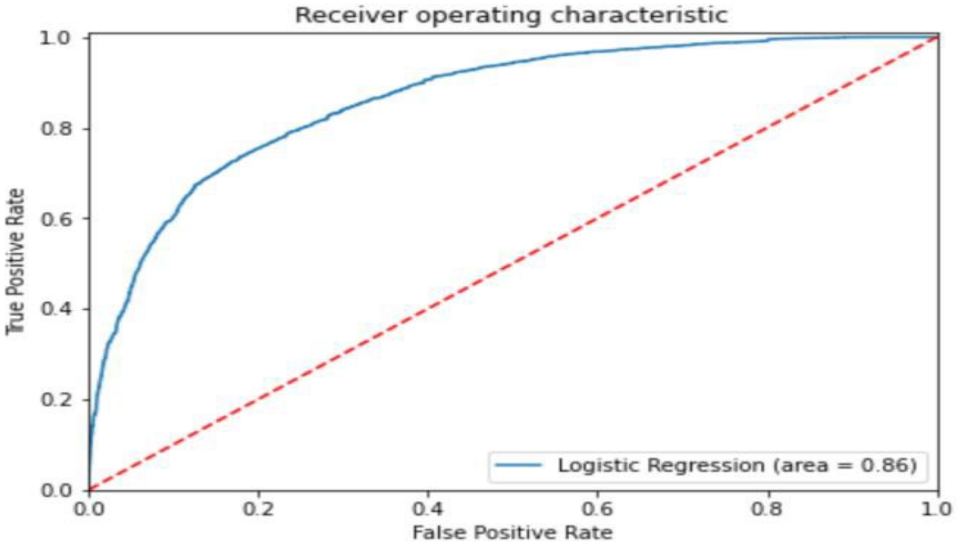
Training performance:

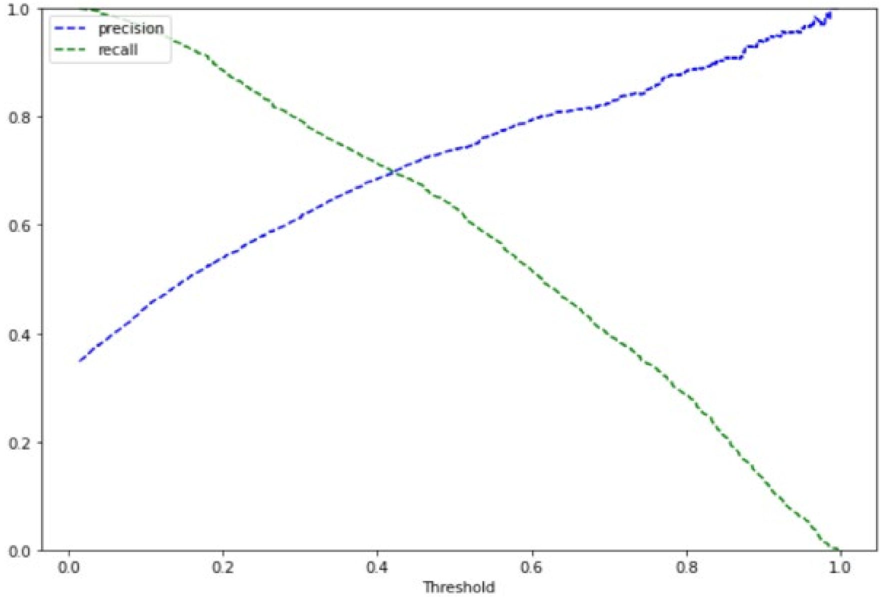
	Accuracy	Recall	Precision	F1
0	0.80545	0.63267	0.73907	0.68174



- We checked VIFs for multicollinearity and dropped the high p-value variables and executed a new regression as well as a confusion matrix. The Results are :

Model Performance Evaluation & Improvement – Logistic Regression





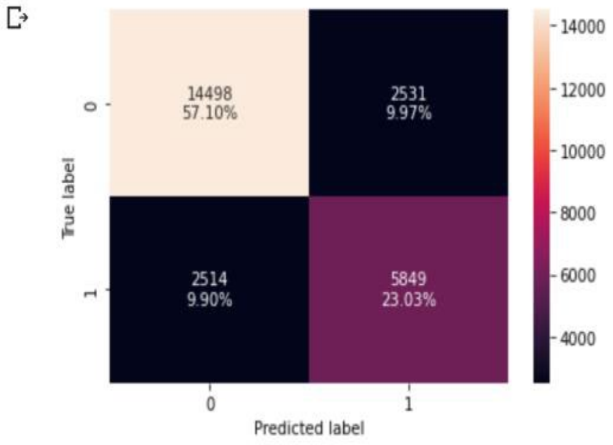
The AUC-ROC curve of our model is shown to the left:

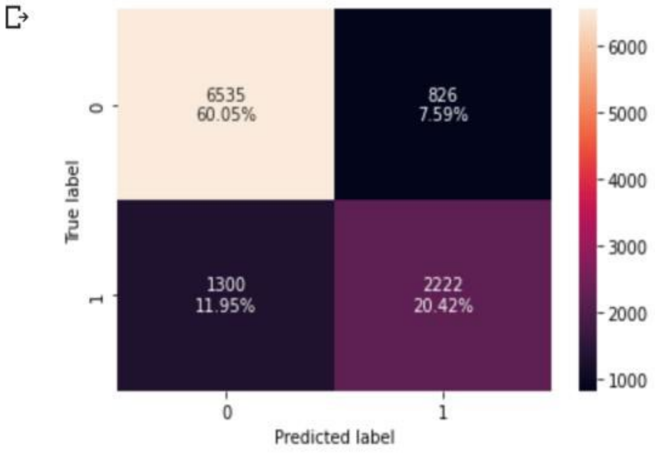
● We wanted to improve the recall score by changing the model threshold with this curve. The optimal threshold cutoff is where tpr(Total Positive rate) is high and fpr(False Positive rate) is low

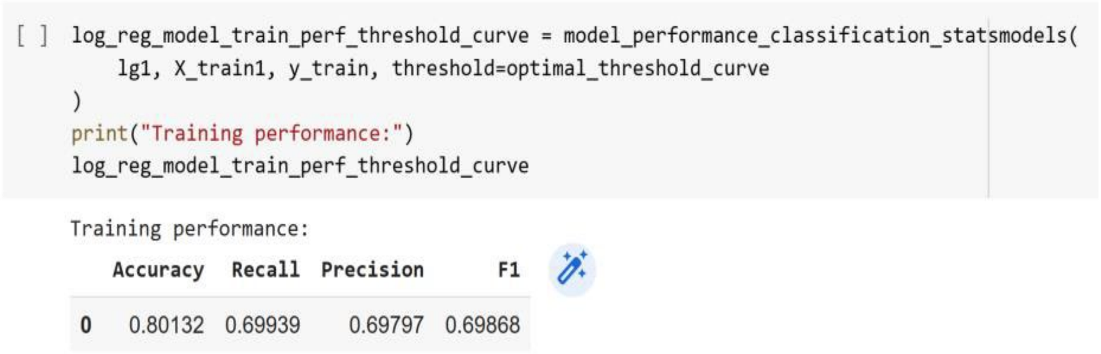
● Using the Precision-Recall curve to the left, we found the optimal threshold for our model to be 0.42.

Model Performance Evaluation & Improvement – Logistic Regression

- With the new threshold, we tested both our test and train data. Here are the results for both:



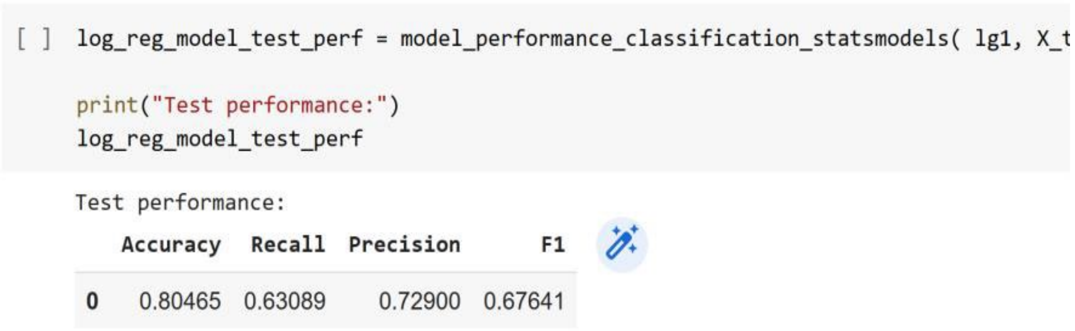




```
[ ] log_reg_model_train_perf_threshold_curve = model_performance_classification_statsmodels(
        lg1, X_train1, y_train, threshold=optimal_threshold_curve
    )
    print("Training performance:")
    log_reg_model_train_perf_threshold_curve
```

Training performance:

	Accuracy	Recall	Precision	F1
0	0.80132	0.69939	0.69797	0.69868



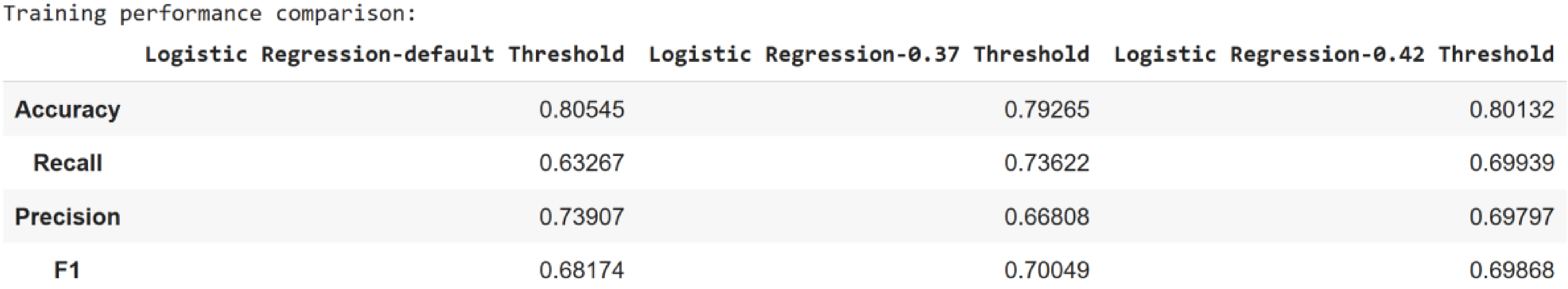
```
[ ] log_reg_model_test_perf = model_performance_classification_statsmodels( lg1, X_t

    print("Test performance:")
    log_reg_model_test_perf
```

Test performance:

	Accuracy	Recall	Precision	F1
0	0.80465	0.63089	0.72900	0.67641

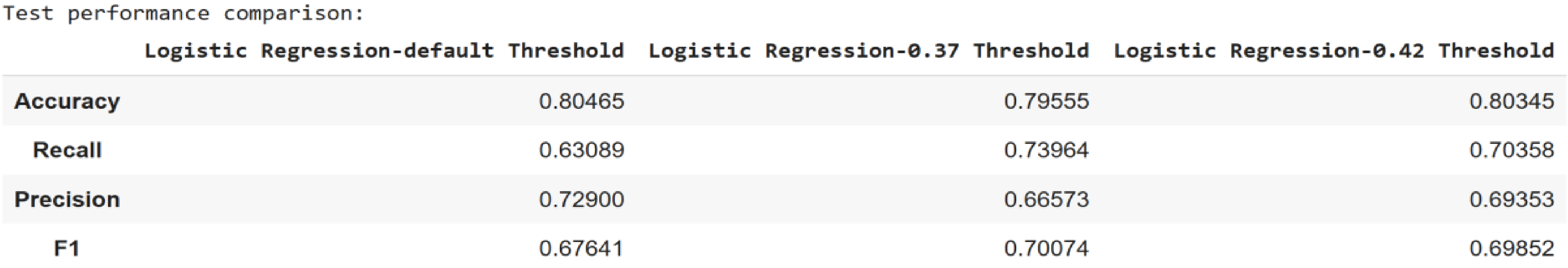
Model Performance Evaluation & Improvement – Logistic Regression

- We then compared the performance of testing and training data on both of the thresholds:



Training performance comparison:

	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



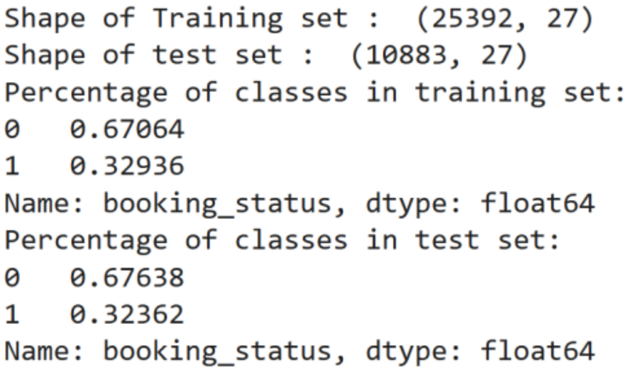
Test performance comparison:

	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 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

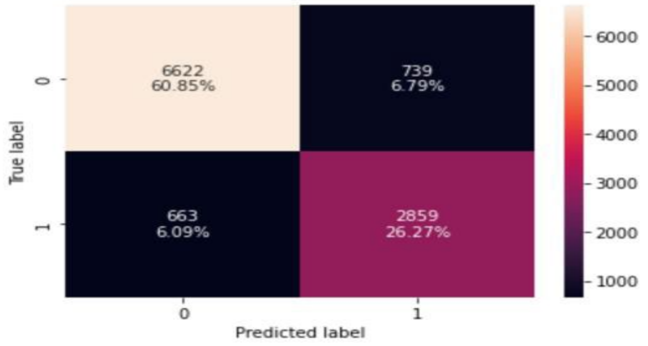
- We concluded that the threshold of 0.42 is preferred.

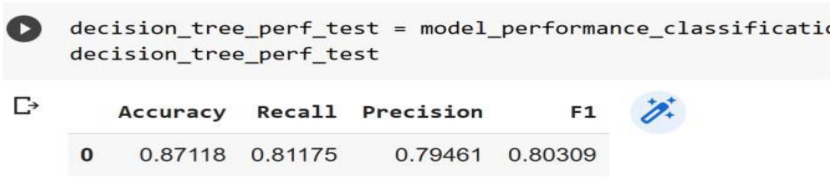
Model Building-Decision Tree

- First ,we split the data:



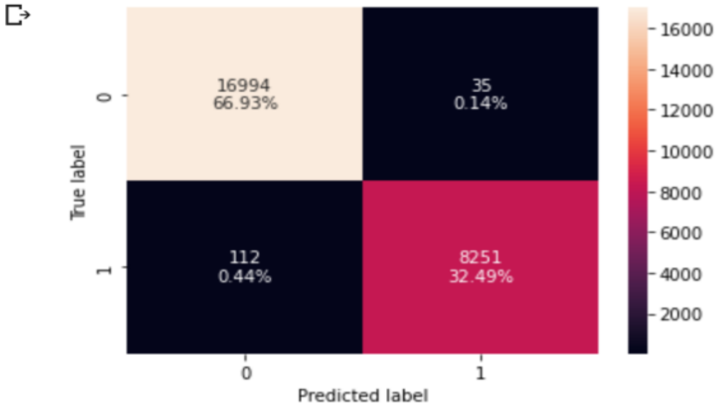
```
Shape of Training set :  (25392, 27)
Shape of test set :  (10883, 27)
Percentage of classes in training set:
0    0.67064
1    0.32936
Name: booking_status, dtype: float64
Percentage of classes in test set:
0    0.67638
1    0.32362
Name: booking_status, dtype: float64
```

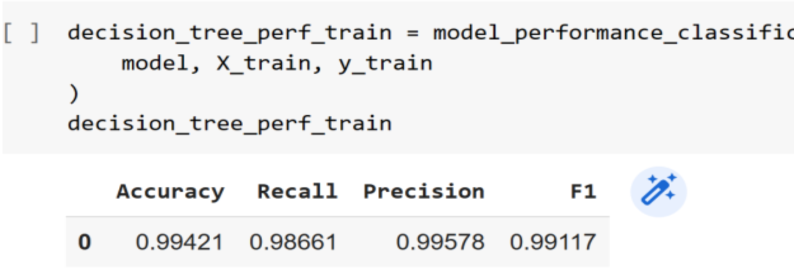




```
decision_tree_perf_test = model_performance_classificati
decision_tree_perf_test
```

	Accuracy	Recall	Precision	F1
0	0.87118	0.81175	0.79461	0.80309





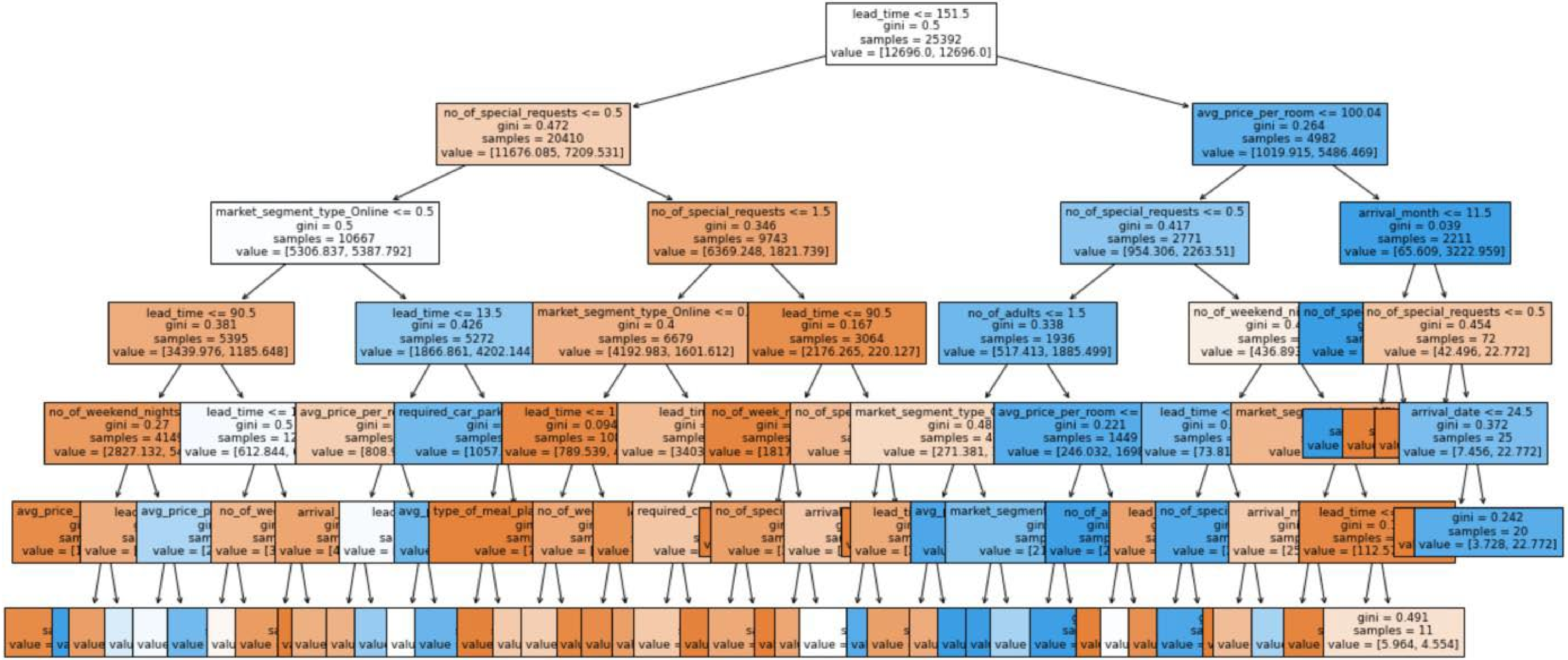
```
[ ]  decision_tree_perf_train = model_performance_classific
         model, X_train, y_train
     )
     decision_tree_perf_train
```

	Accuracy	Recall	Precision	F1
0	0.99421	0.98661	0.99578	0.99117

- Then, we tested the model on these sets of data:

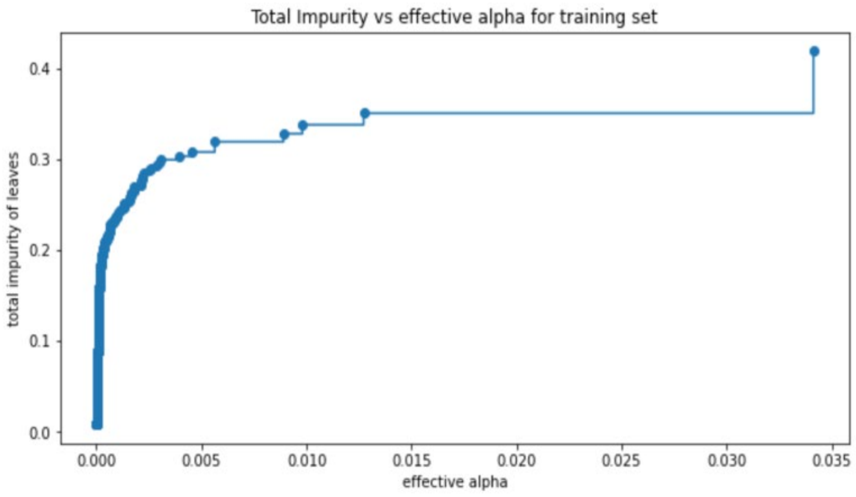
Model Building-Decision Tree

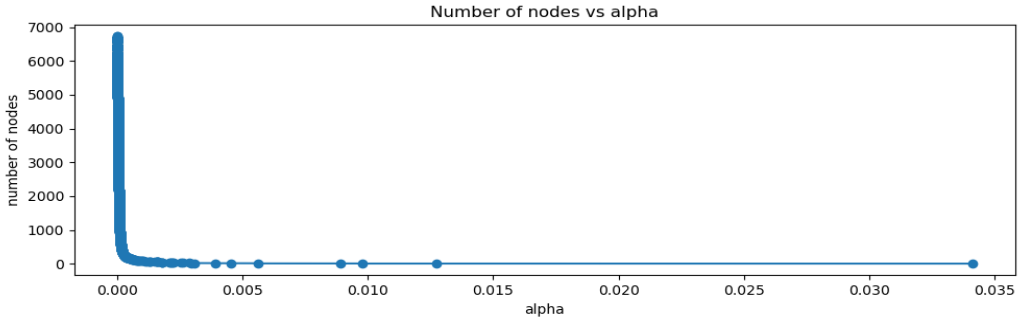
- Next, We began to visualize the tree

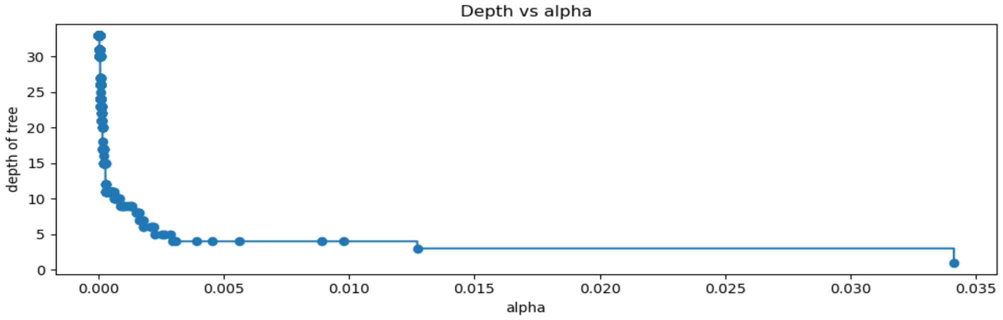


Model Performance Evaluation & Improvement

Cost Complexity Pruning







● Next, we train a decision tree using effective alphas

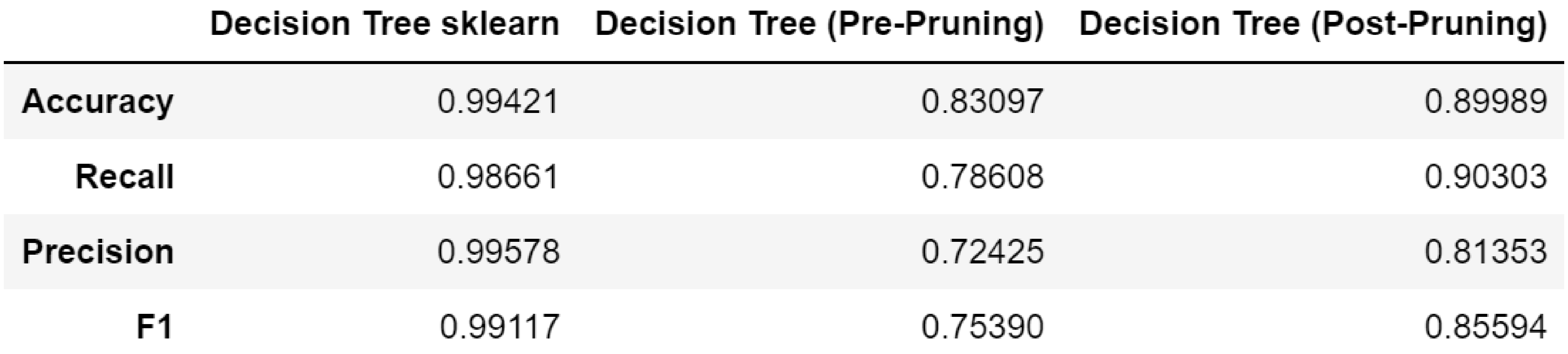
●The last value in ccp_alphas is the alpha value that prunes the whole tree, leaving the tree with one node.

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

Model Performance Evaluation –Decision Tree

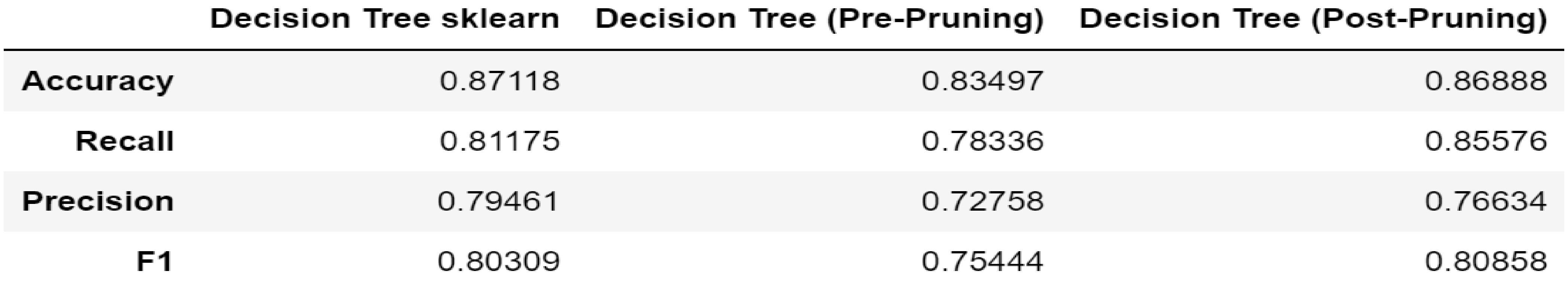
- We compared the performance of testing and training data on Decision Tree Models:

Performance check on the train set



	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.99421	0.83097	0.89989
Recall	0.98661	0.78608	0.90303
Precision	0.99578	0.72425	0.81353
F1	0.99117	0.75390	0.85594

Performance check on the test set



	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.87118	0.83497	0.86888
Recall	0.81175	0.78336	0.85576
Precision	0.79461	0.72758	0.76634
F1	0.80309	0.75444	0.80858

Model Performance Summary

- We want to pick the best models to predict if a booking will be cancelled.
- We want to avoid false negatives and false positives as they are harmful to our predictive models, so we want the highest F1 score in our model for both the logistic regression and decision tree.
- The decision tree post-pruning is the model with the highest F1(0.80858) and accuracy of 0.86888, recall of 0.85576, and precision of 0.76634. This is our best model.
- As we found, our top features for prediction were lead time, market segment type, avg. price per room, no. of special requests, and arrival month.

Executive Summary

INSIGHTS:

- Bookings with a lead time of 120 days or more are most likely to be cancelled
- Online bookings are most likely to be cancelled & complimentary ones are least likely to be cancelled.
- The more special requests, the less likely the booking will be cancelled
- Bookings with 3 or more guests are more likely to be cancelled
- More bookings and cancellations were found to occur over months (March-August) compared to (September-February)
- Observing market segments, the avg price per room has been higher in instances where bookings have been cancelled than in cases in which bookings have not been cancelled. More competition information is required to ensure that our pricing is competitive to retain guests.

Executive Summary

Business Recommendations:

- Increasing available room inventory by converting the unpopular floor plans such as room type 3, 5, and 7 to the most popular plan, room type 1. If the hotels have a renovation budget and if the building floor plans are permitted.
- Reducing cost and other resources by removing Meal Plan 3 which is the least popular with only 5 orders.
- Giving guests rewards for more stays/less cancellations such as discounts, free parking, etc would promote business.
- The hotels should implement a cancellation fee for within 24 hours of the booking and for no-shows as well.
- Guests should not be allowed to make bookings too far in advance(1 year or more).
- Promotion of booking a room for Aug-Oct (highest number of bookings and highest

  cancellations) and get a one night stay complimentary to be redeemed in the next 6 months for free.
- The lead time was identified as the most important feature; a longer lead time increases the odds of cancellations. Policies need to be introduced to restrict how far in advance bookings can be made before the check-in date.