INN Hotel Booking Data Analysis (ML-1 Coded)

Apr B Sunday Onkar 10:30 AM Batch

Arindam Saha

Table of Contents

List of figures

List of tables

1. Background:

- A significant number of hotel bookings are called off due to cancellations or no-shows. The typical reasons for cancellations include change of plans, scheduling conflicts, etc. This is often made easier by the option to do so free of charge or preferably at a low cost which is beneficial to hotel guests but it is a less desirable and possibly revenue-diminishing factor for hotels to deal with. Such losses are particularly high on last-minute cancellations.
- The new technologies involving online booking channels have dramatically changed customers' booking possibilities and behavior. This adds a further dimension to the challenge of how hotels handle cancellations, which are no longer limited to traditional booking and guest characteristics.

The cancellation of bookings impacts a hotel on various fronts:

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

2. Objective

The increasing number of cancellations calls for a Machine Learning based solution that can help in predicting which booking is likely to be canceled. INN Hotels Group has a chain of hotels in Portugal, they are facing problems with the high number of booking cancellations and have reached out to your firm for data-driven solutions. You as a data scientist have 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.

3. Data Dictionary:

Variables	Description
Booking_ID	The 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 weeknights (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 Group
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.

Table 1: Data Dictionary

4. Data Information:

Note: -

- There are in total 36275 rows and 19 columns present.
- There are 14 numeric (float and int type) and 5 string (object type) columns in the data.
- The target variable is the booking_status (cancelled / not cancelled), which is categorical type.
- In total 3 meal plans observed ('Meal Plan 1', 'Meal Plan 2', 'Meal Plan 3')
- There are 7 different types of rooms available in the data.
- Market segments are ('Offline', 'Online', 'Corporate', 'Aviation', 'Complementary')
- Views trailer & ad impressions seem to have outliers.
- The mean views content has mean of 0.47 with standard deviation of 0.1
- Mean is almost similar or very low differnce to median for views content.
- There are no duplicates.
- There are no missing values in all the columns.
- For the column major_sports_event, we have replaced the 1 value with '***yes***' and 0 values with '***no***'.
- no_of_week_nights: Average no of week nights are 2.2. A vast difference in minimum value and 25th percentile, as well as 75th percentile and the maximum value, indicates that there might be outliers present in the variable.
- lead_time: The mean number of days between the date of booking and the arrival date is 85.23. A vast difference in minimum value and 25th percentile, as well as 75th percentile and the maximum value, indicates that there might be outliers present in the variable.
- avg_price_per_room: The average price per room is 103.42 in euros. A vast difference in minimum value and 25th percentile, as well as 75th percentile and the maximum value, indicates that there might be outliers present in the variable.

- No_of_previous_booking_not_cancelled: There's a huge difference in the 75th percentile and maximum value of No_of_previous_booking_not_cancelled indicating the presence of outliers. Also, 75% of the observations are 0.
- no_of_previous_cancelled: Same as no_of_previous_cancelled there's a huge difference in the 75th percentile and maximum value indicating the presence of outliers. Also, 75% of the observations are 0.
- repeated_guest: Same as repeated_guest there's a huge difference in the 75th percentile and maximum value indicating the presence of outliers. Also, 75% of the observations are 0.
- car_parking_space: Same as car_parking_space there's a huge difference in the 75th percentile and maximum value indicating the presence of outliers. Also, 75% of the observations are 0.
- no_of_children: Same as no_of_children there's a huge difference in the 75th percentile and maximum value indicating the presence of outliers. Also, 75% of the observations are 0.
- We labeled no_of_booking_not_cancelled in four different groups
    - Very Low (count < 1)
    - Low (Count 1 to 5)
    - Medium (count 6 to 15)
    - High (count 15 to 30)
- Also we have labelled no_of_previous_cancellations in three different groups
    - No Cancellations (Count = 0)
    - Few Cancellations (Count = 1 to 5)
    - Moderate Cancellations (Count = 6 to 10)

5. Exploratory Data Analysis (EDA)

5.1. Univariate Analysis

5.1.1. Observations on no of adults

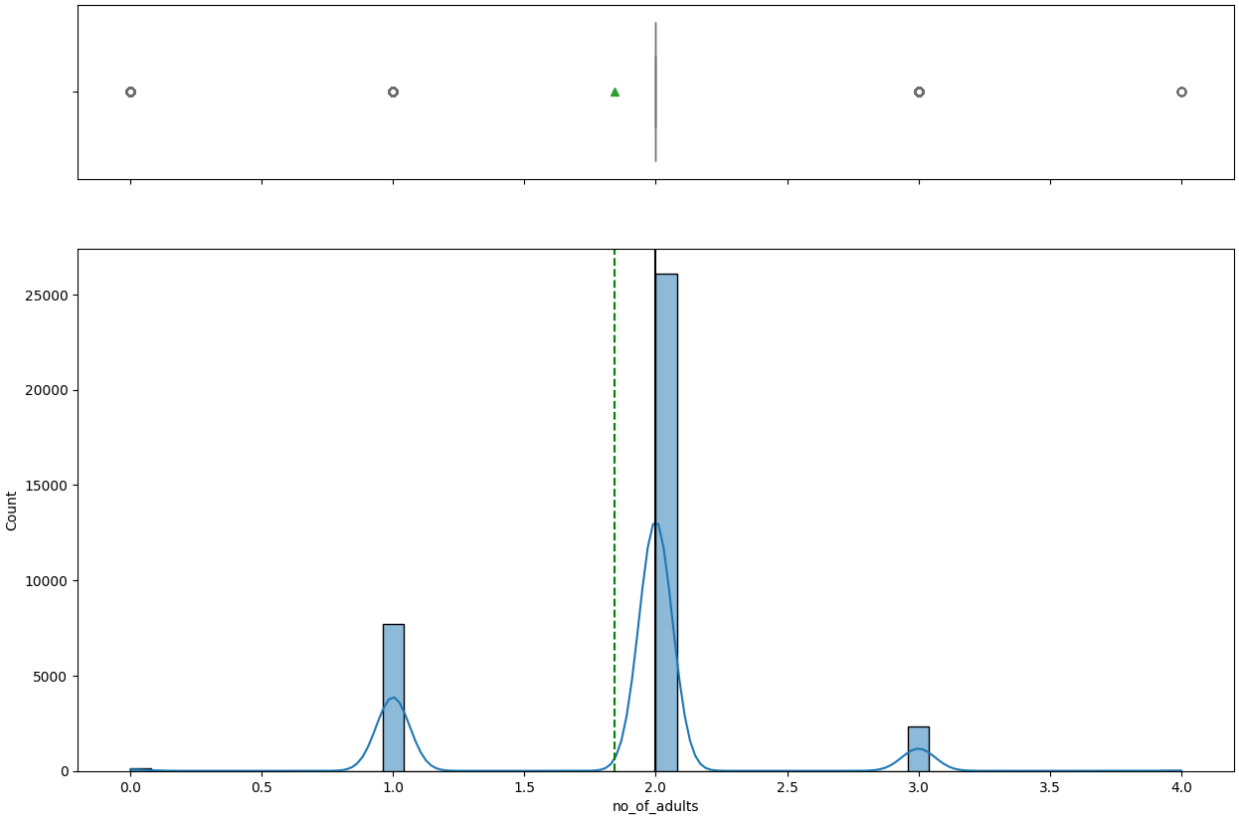

Figure 1: Univariate on number of adults

- The majority of bookings were for two adults, totaling over 25,000.

- Bookings for three adults were minimal (around 2,500), and bookings for one adult were about 7,500.
- The boxplot indicates that bookings with zero, one, and three adults are outliers.
- It's important to investigate the reasons behind bookings with zero adults.
- After investigation we came to know that around 31 percentile of the booking with zero adults got cancelled.
-

5.1.2. Observations on no of children

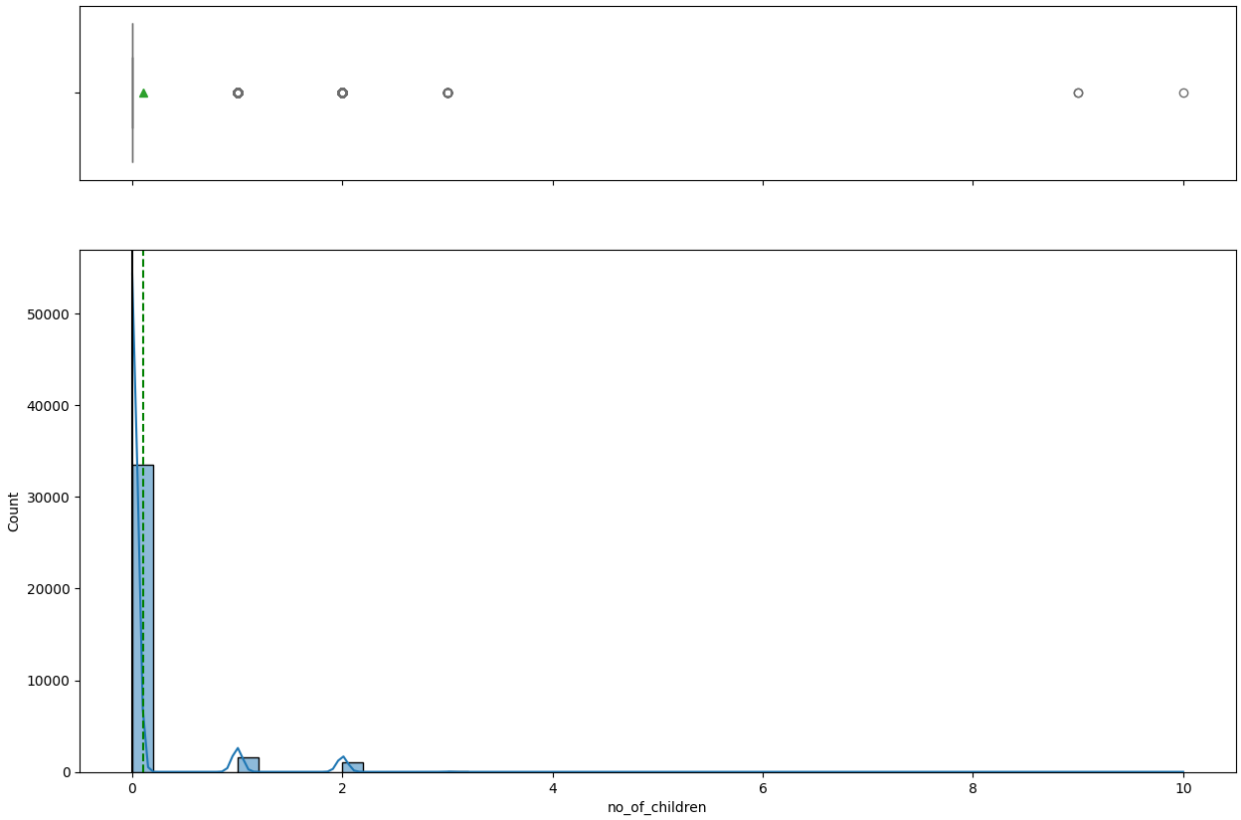

Figure 2: Univariate on Number of Children

- Most of the booking were made with zero no of child. But also seems to have outliers present in it.
- We need to check if there is any relation with multiple number of children.
- One or two adults with 9 to 10 children is a rare phenomenon.
- We must get more understanding on this from client.

5.1.3. Observation on no of weekend nights

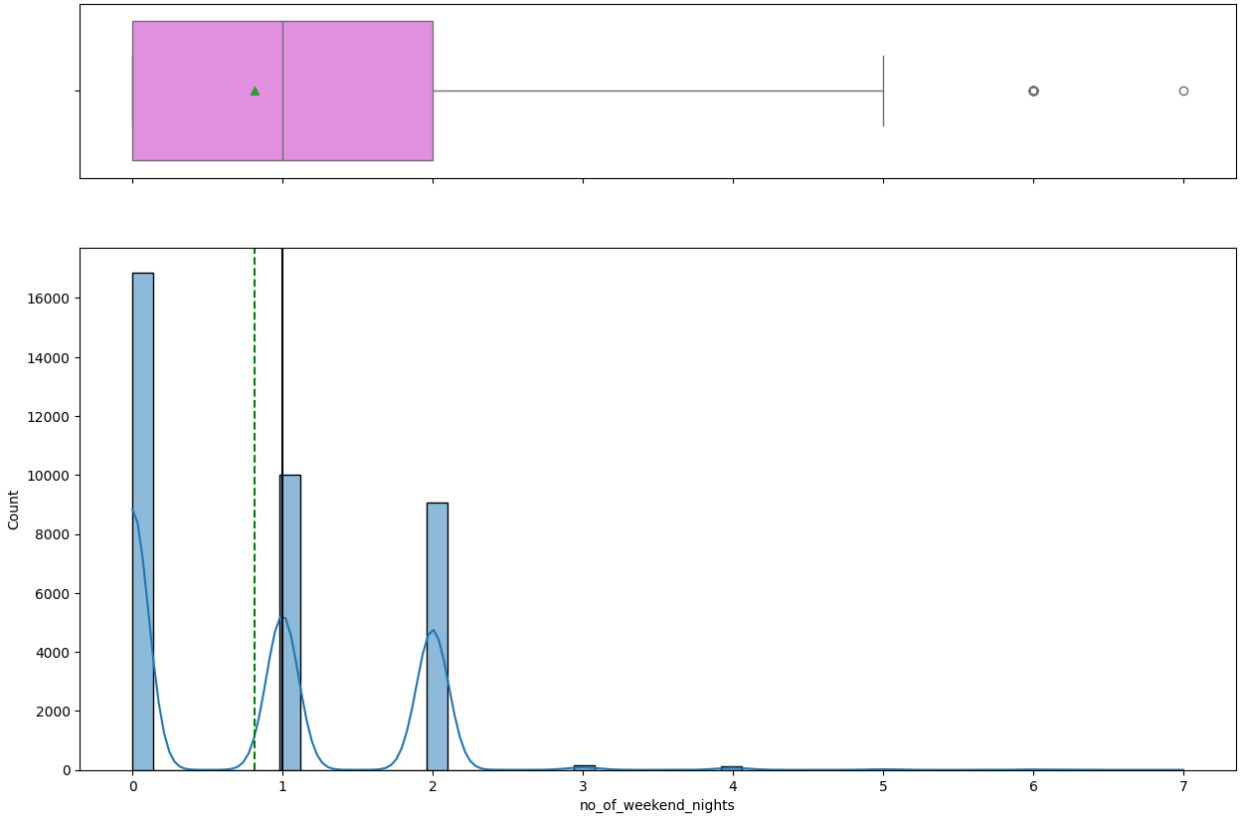

Figure 3: Univariate on Number of Weekends

- It is right skewed. As booking more than 2 weekend nights means long stay in the hotel.
- We shall check in depth to understand these bookings.
- We can clearly observe that booking with more than 7 days has more probability (~ 60%) of getting cancelled.
- We will try to analyze further if we can understand of getting cancelled for these bookings.

5.1.4. Observations on number of week nights

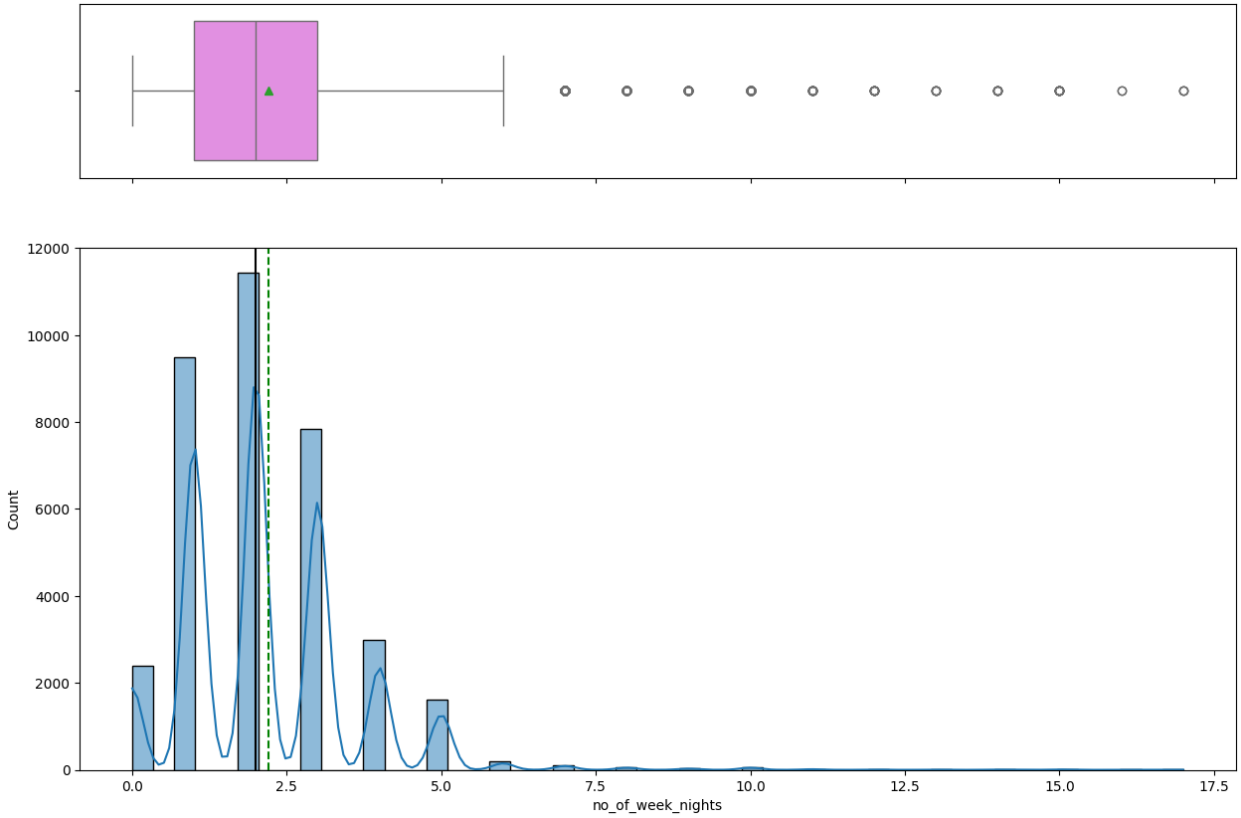

Figure 4: Univariate on Number of Week Nights

- Seems to have outliers for the number of week day nights more than ~5.5 days.
- We can also check those outliers.
- We can also observe here booking with more than 5.5 week nights has ~10% higher probability of getting cancelled.
- Price of rooms are likely to be the main reason for cancellation.

5.1.5. Observations on lead time

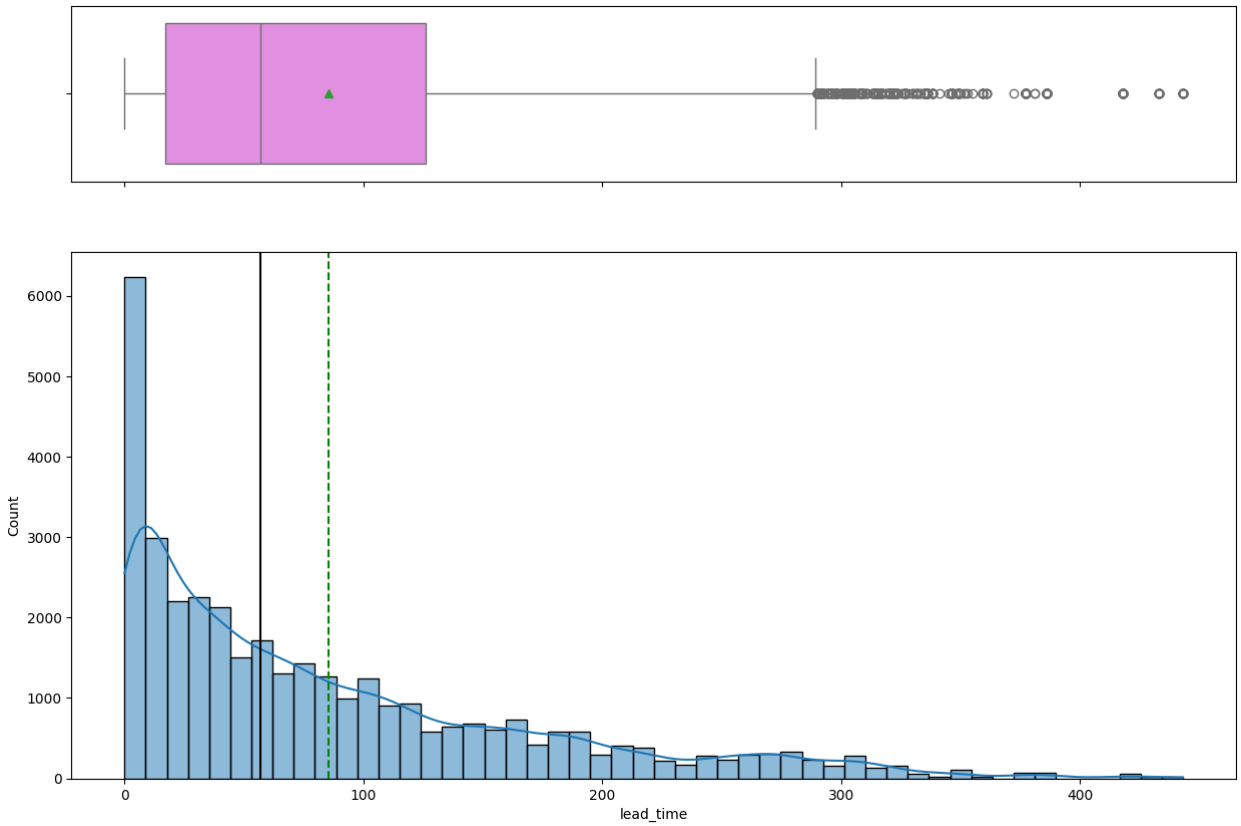

Figure 5: Univariate on Lead Time

- Lead time seems to be right skewed. It also have some outliers towards the right tail.
- It is very clear that if the booking date is very far / lead time more are likely to get cancelled.

5.1.6. Observation on arrival month

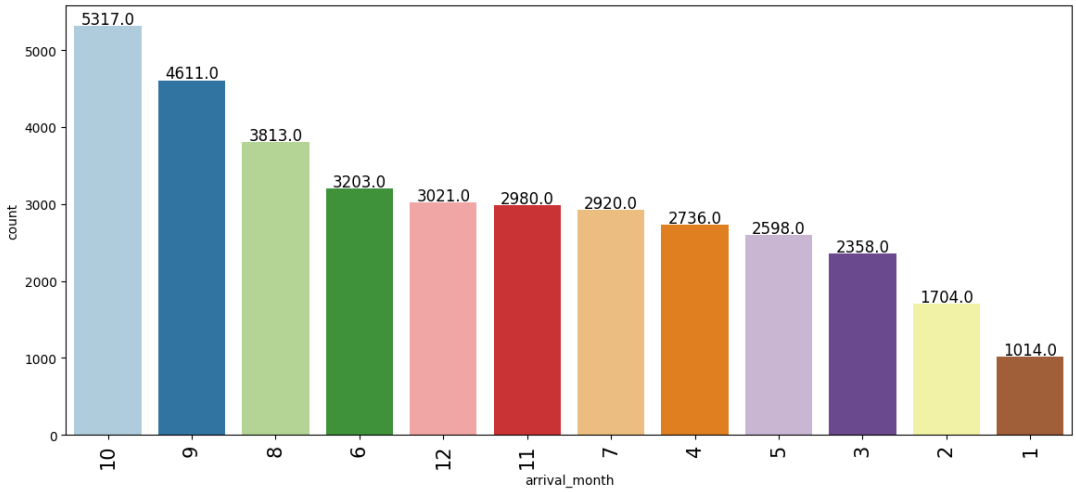

Figure 6: Univariate on Arrival Month

- October has the highest number of arrivals, followed by September and August. January and February have the lowest number of arrivals.

5.1.7. Observations on avg Price per room

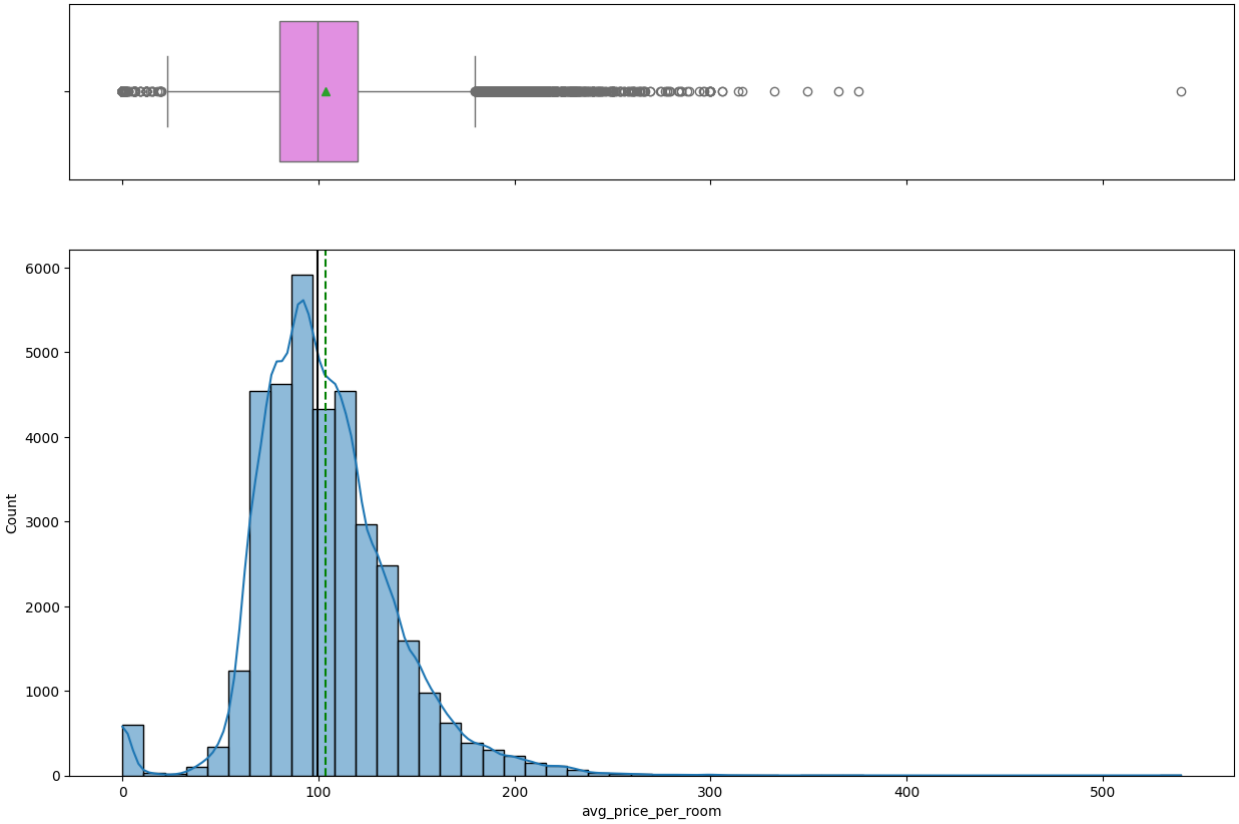

Figure 7: Univariate on AVG Price Per Rom

- Distribution looks normal and slightly right skewed with outliers.
- Mean and median are almost existing at same point ~ 100-105 Euros.
- Outliers seems to have no influence on booking cancellation.

5.1.8. Observation on market segment

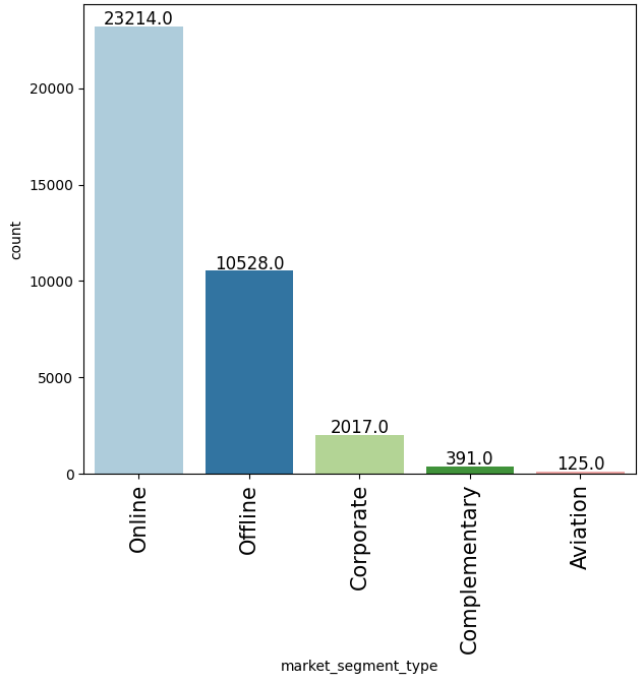

Figure 8: Univariate on Market Segments

- The online mode is the most dominant market segment (~64%), followed by the offline mode (29%).
- Bookings in the complimentary and aviation market segments are very few.

5.1.9. Observation on booking status

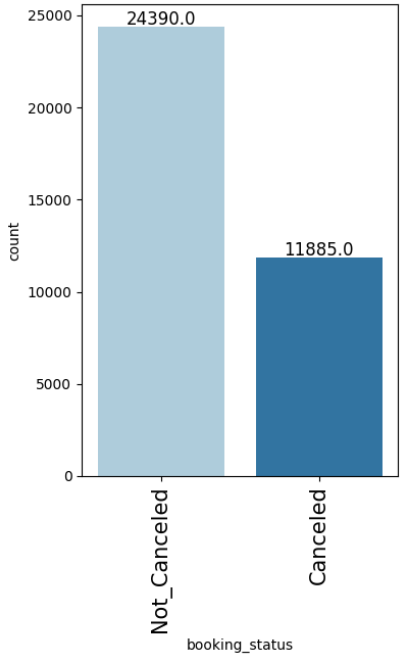

Figure 9: Univariate on Booking Status

- Almost one third of the total bookings got cancelled.

5.1.10. Observations on room type reserved

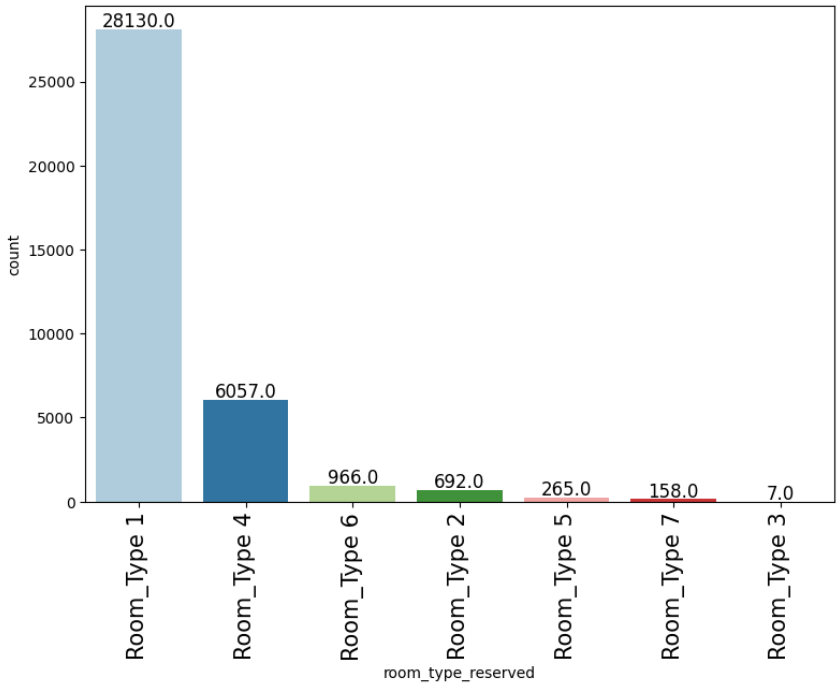

Figure 10: Univariate on Room Type Reserved

- Type one rooms were booked the most (77%), followed by type four rooms (~17%). Other types of rooms had significantly fewer bookings.

5.2. Bivariate Analysis

5.2.1. Pair Plot and Heat Map

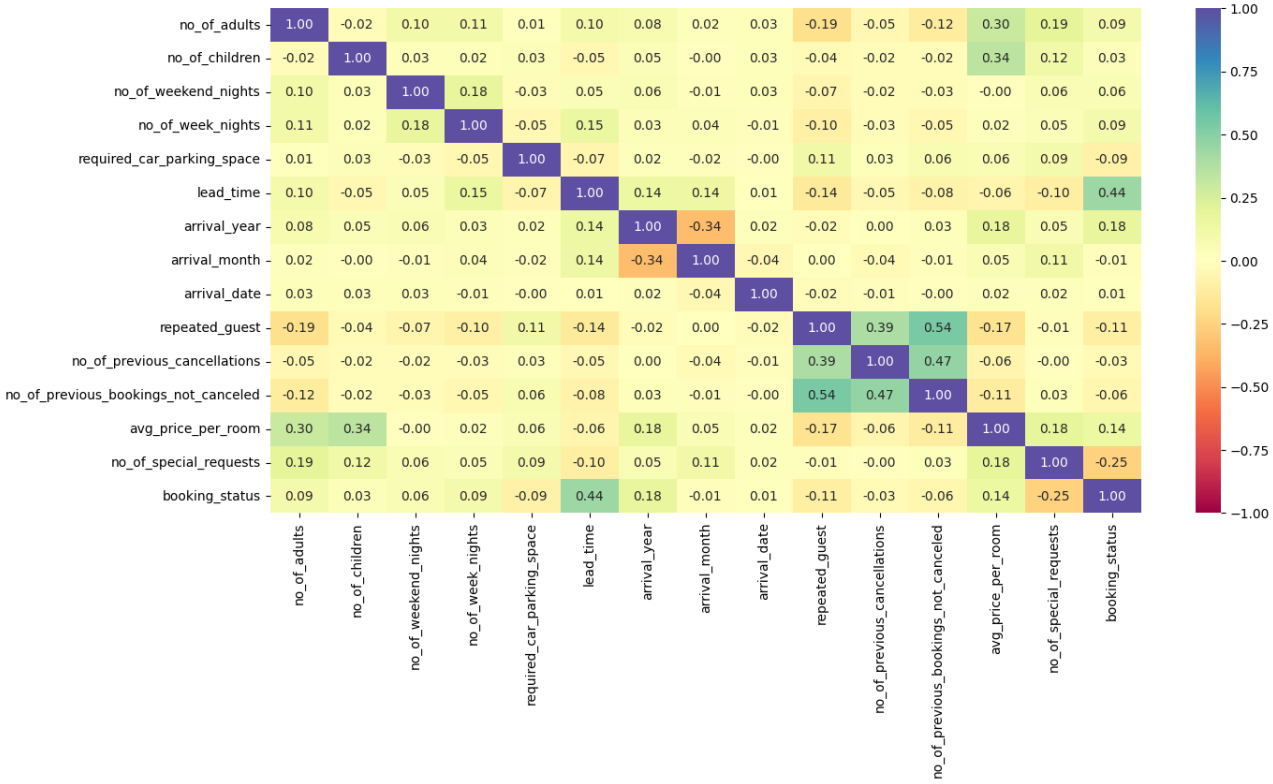

Figure 11: Heat map

Figure 12: Pair Plot

- Booking Status cancelled has got a good positive correlation with the Lead Time. And also, it is negatively correlated with the special requests and repeated guests.
- No of previous bookings not cancelled has got positive correlation with repeated guests.
- Relation can be also observed with respect to previous bookings not cancelled and previous cancellation.

5.2.2. Bivariate market segment vs price

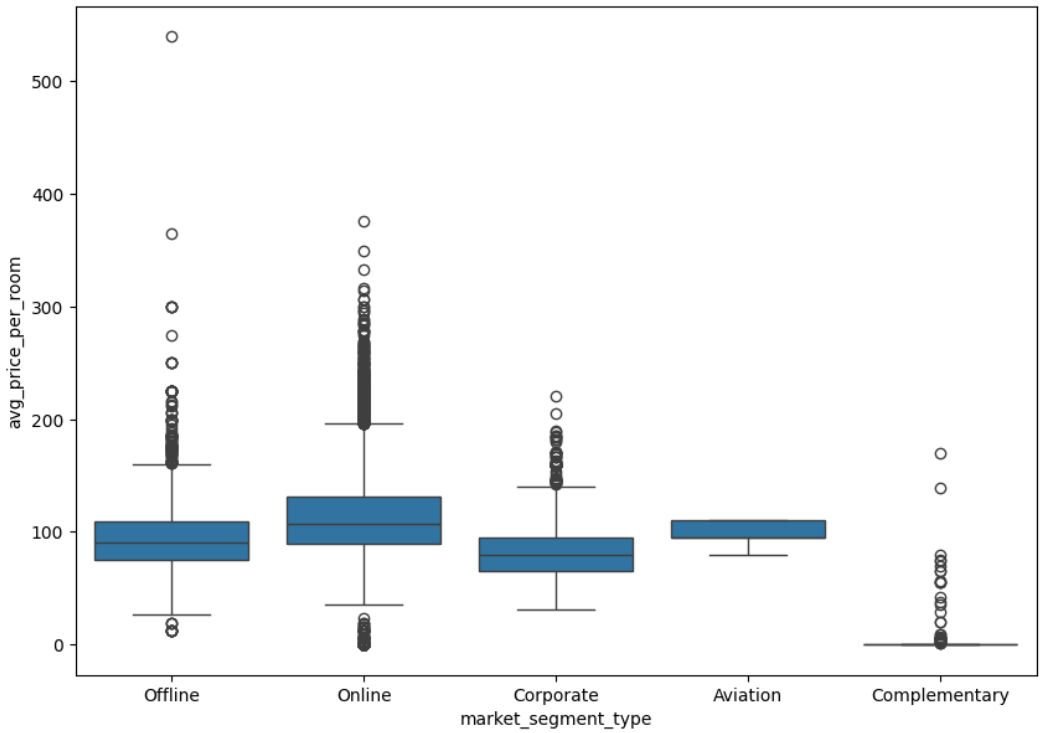

Figure 13: Bivariate market segment w.r.t. avg price of room with outliers

- Outliers can be observed in all the market segments.

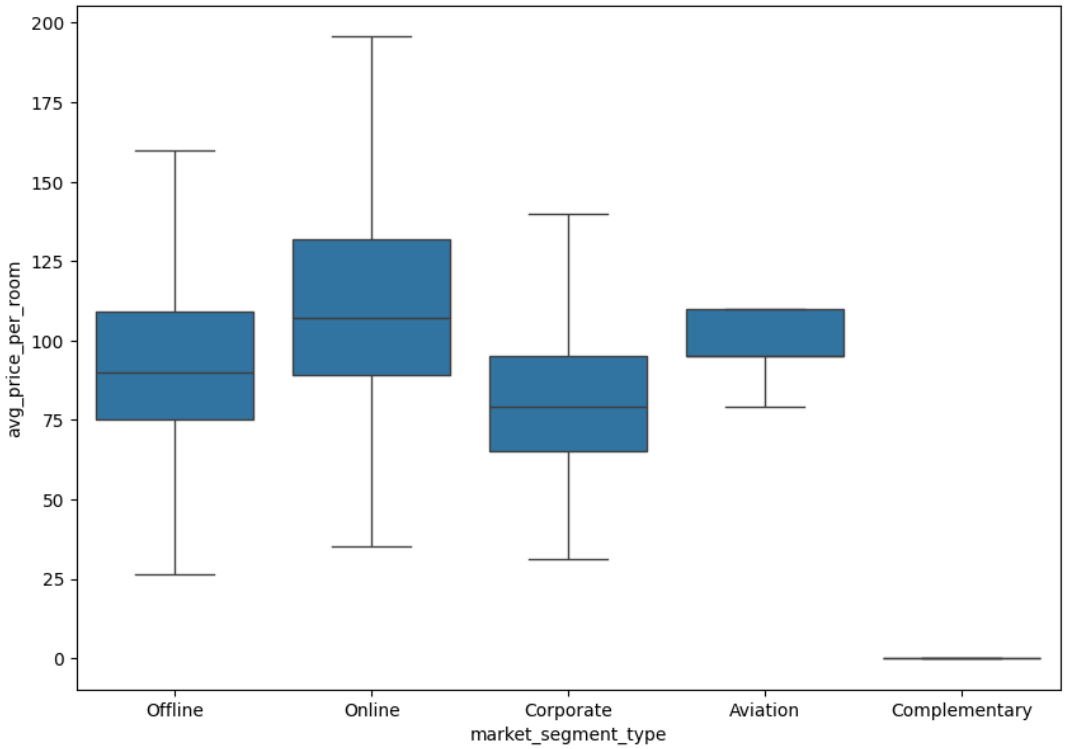

Figure 14: Bivariate market segment w.r.t. avg price of room without outliers

- The mean price per room for Online market segments are at the highest level.

- Further analysis will be needed to check if the cancellation probability is more or not.

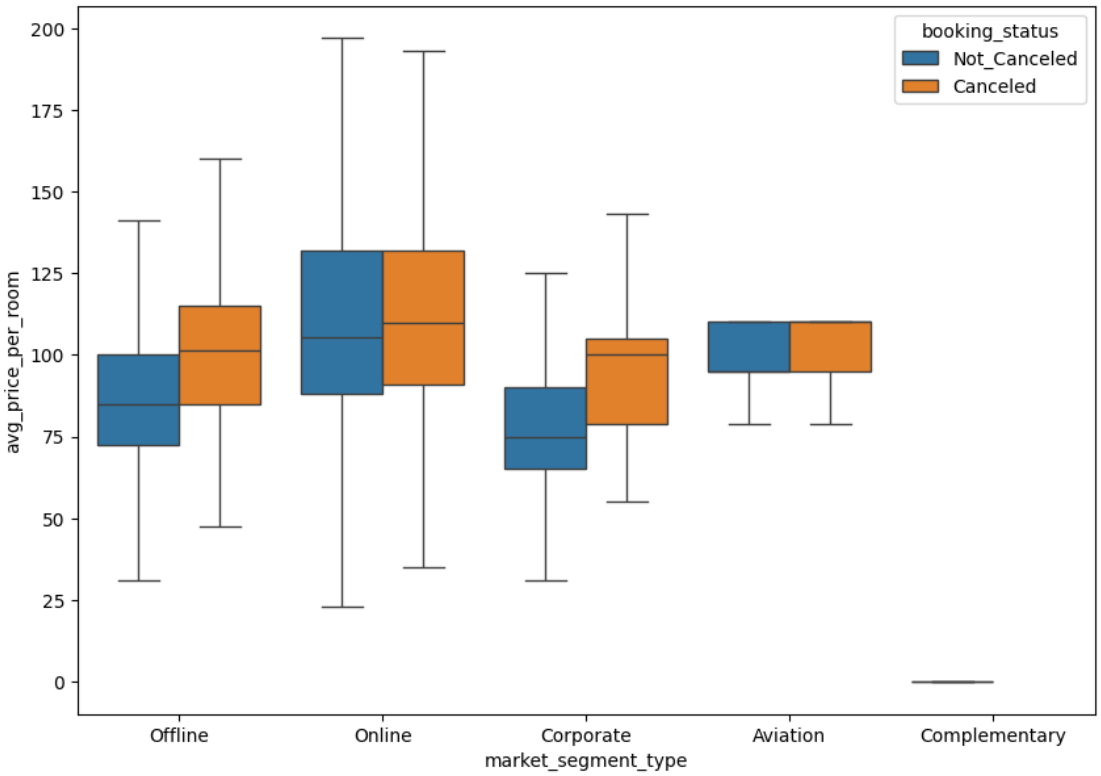

Figure 15: Avg room price vs market segment type w.r.t. booking status

- For corporate and offline bookings, the mean price per room for cancelled bookings was significantly higher compared to non-cancelled bookings.
- In contrast, online and aviation bookings exhibited nearly identical room prices regardless of whether the bookings were cancelled or not.
- Online bookings have a higher probability of cancellation at 36.5%, compared to offline and aviation bookings, which have a cancellation probability of around 30%.

5.2.3. Bivariate on repeating guest

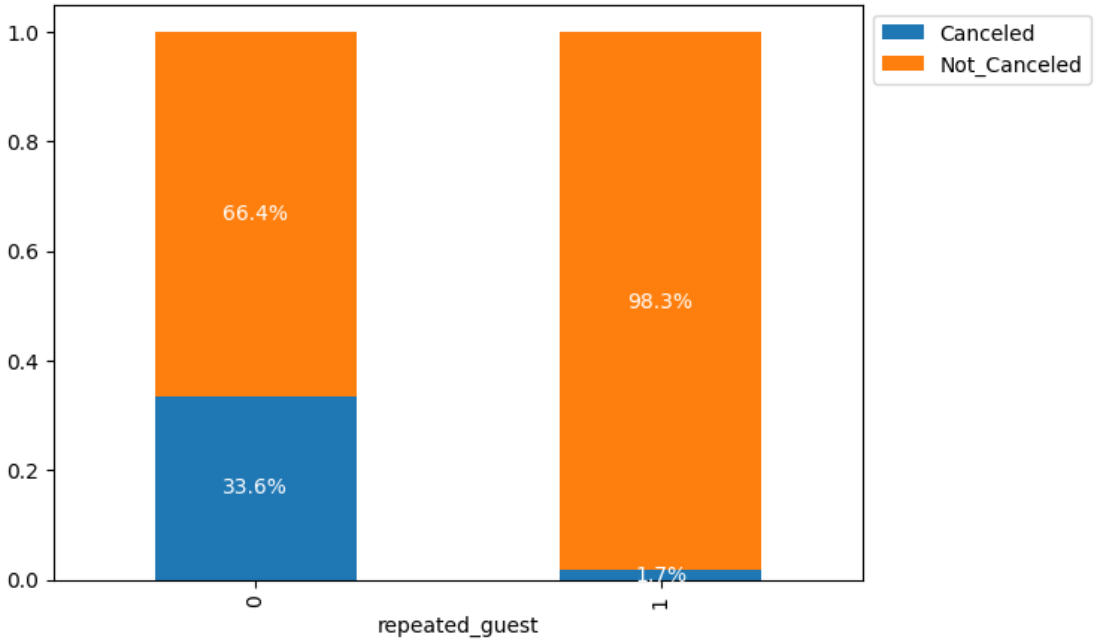

Figure 16: Repeating guest w.r.t. booking status

- Repeat guests are less likely to cancel (1.7%) their bookings compared to new guests.
- We can also check what are the no of previous booking not cancelled and how many of them are fall in the repeated guests.
- The guests who did not cancelled their bookings previously, seems to be the repeated guests who did not cancelled new bookings as well.

5.2.4. Observation on special requests

Guests who make special requests are less likely to cancel their bookings. Historically, only 69 guests with special requests have cancelled their bookings. Currently, the cancellation rate for bookings with special requests is just 20%.

5.2.5. Observation on lead time

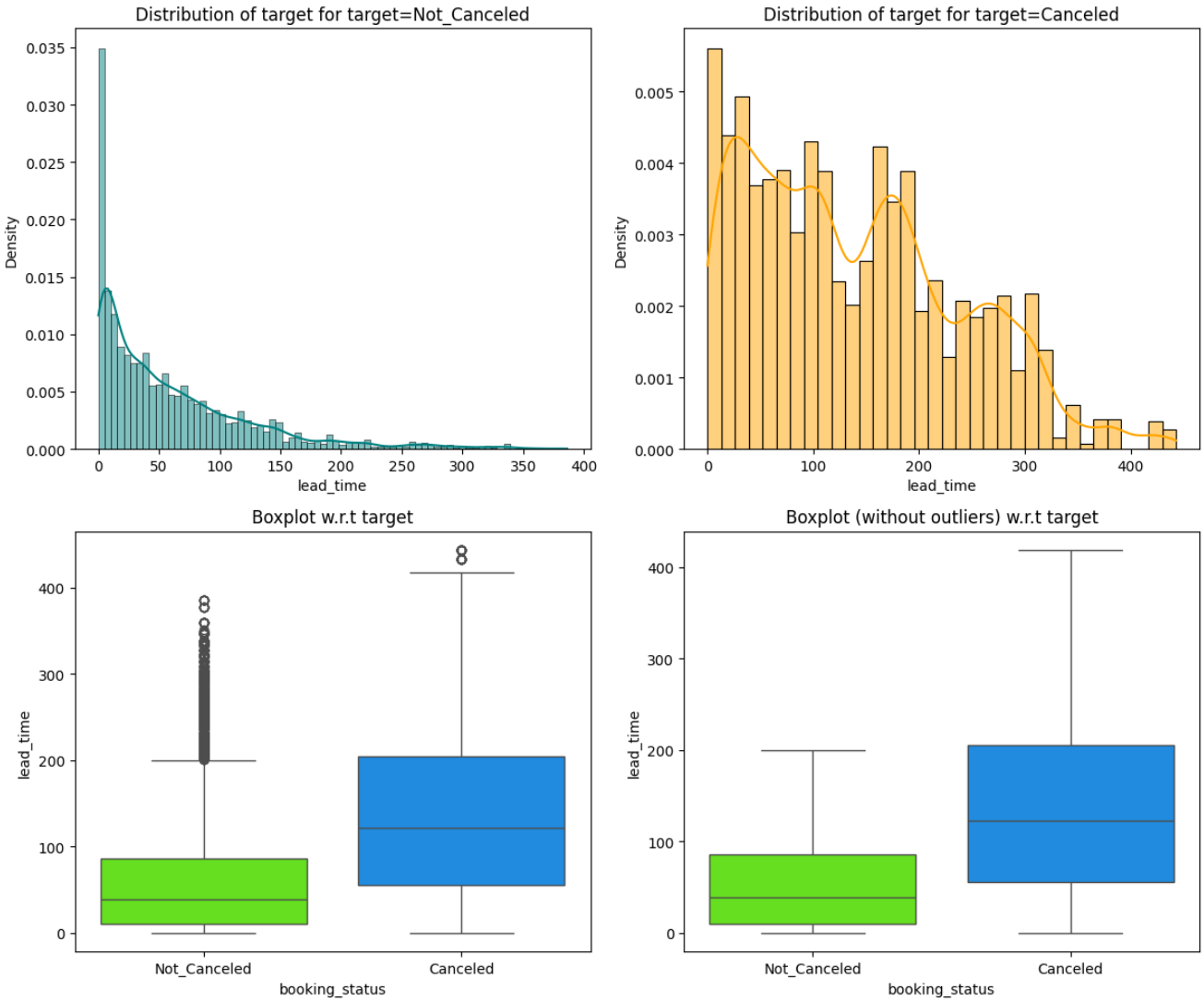

Figure 17: Observation on lead time w.r.t. booking status

- As the lead time increases probability of booking getting canceled increases.

5.2.6. Observation on avg room price

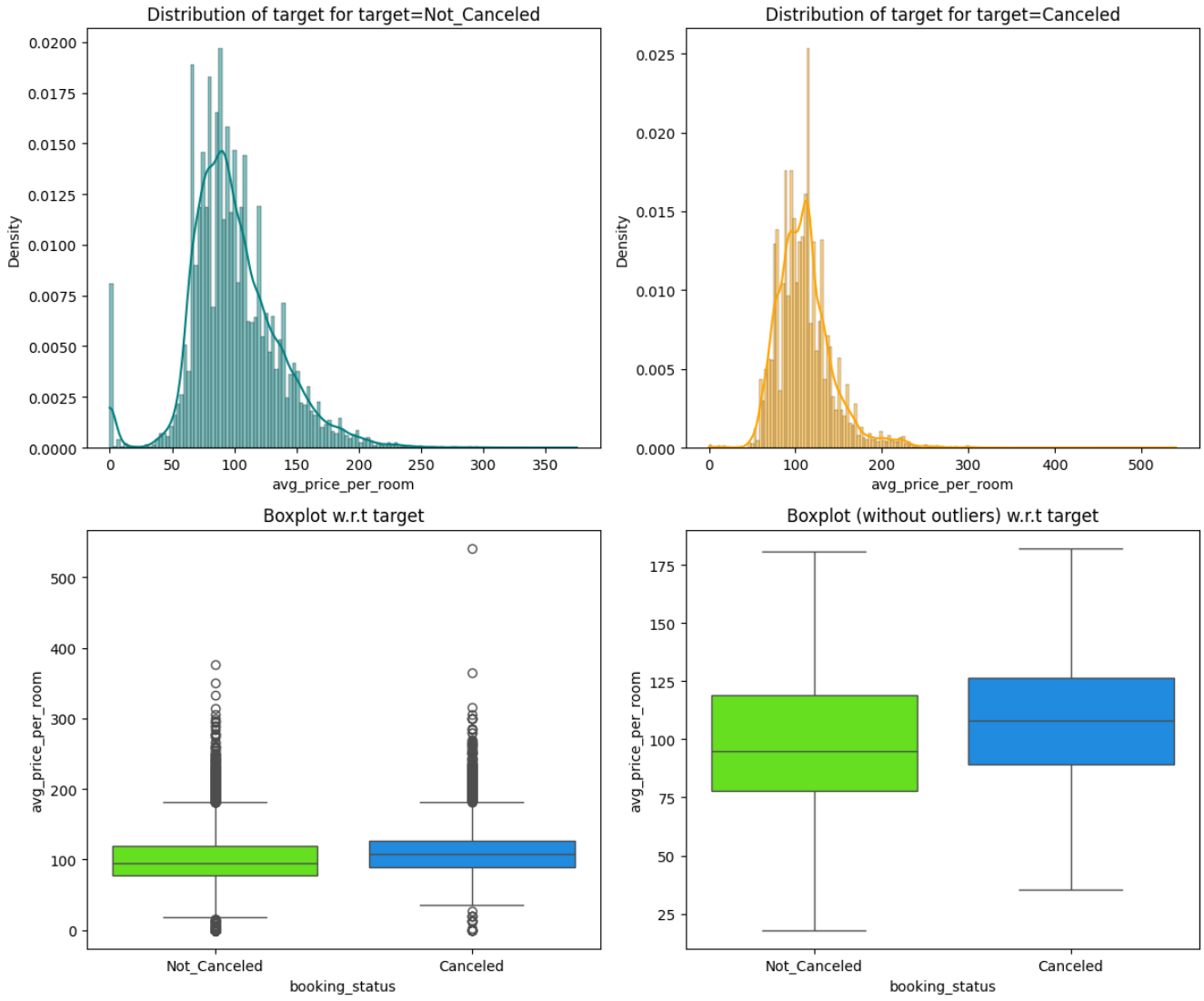

Figure 18: Observation on avg room price w.r.t. booking status

- The mean price per room is slightly higher for the booking getting canceled than the not canceled.

6. Logistic Regression

We can drop 'no_of_children', 'required_car_parking_space', 'no_of_previous_cancellations','no_of_previous_bookings_not_canceled', and 'repeated_guest' as most of the observations are 0 in these variables.

'Booking_ID' we can drop as it is a unique value no influence can be observed.

There are many outliers in the data which we will treat (perform capping of outliers)

All the values smaller than the lower whisker will be assigned the value of the lower whisker, and all the values above the upper whisker will be assigned the value of the upper whisker.

6.1. Data Preprocessing

- We labeled no_of_booking_not_cancelled in four different groups
  - Very Low (count < 1)

    - Low (Count 1 to 5)
    - Medium (count 6 to 15)
    - High (count 15 to 30)
- Also we have labelled no_of_previous_cancellations in three different groups
    - No Cancellations (Count = 0)
    - Few Cancellations (Count = 1 to 5)
    - Moderate Cancellations (Count = 6 to 10)
- We can drop 'no_of_children', 'required_car_parking_space', 'no_of_previous_cancellations', and 'no_of_previous_bookings_not_canceled' as most of the observations are 0 in these variables.
- 'Booking_ID' we can drop as it is a unique value no influence can be observed.
- There are many outliers in the data which we will treat (perform capping of outliers)
- All the values smaller than the lower whisker will be assigned the value of the lower whisker, and all the values above the upper whisker will be assigned the value of the upper whisker.

6.1.2. Outliers Detection

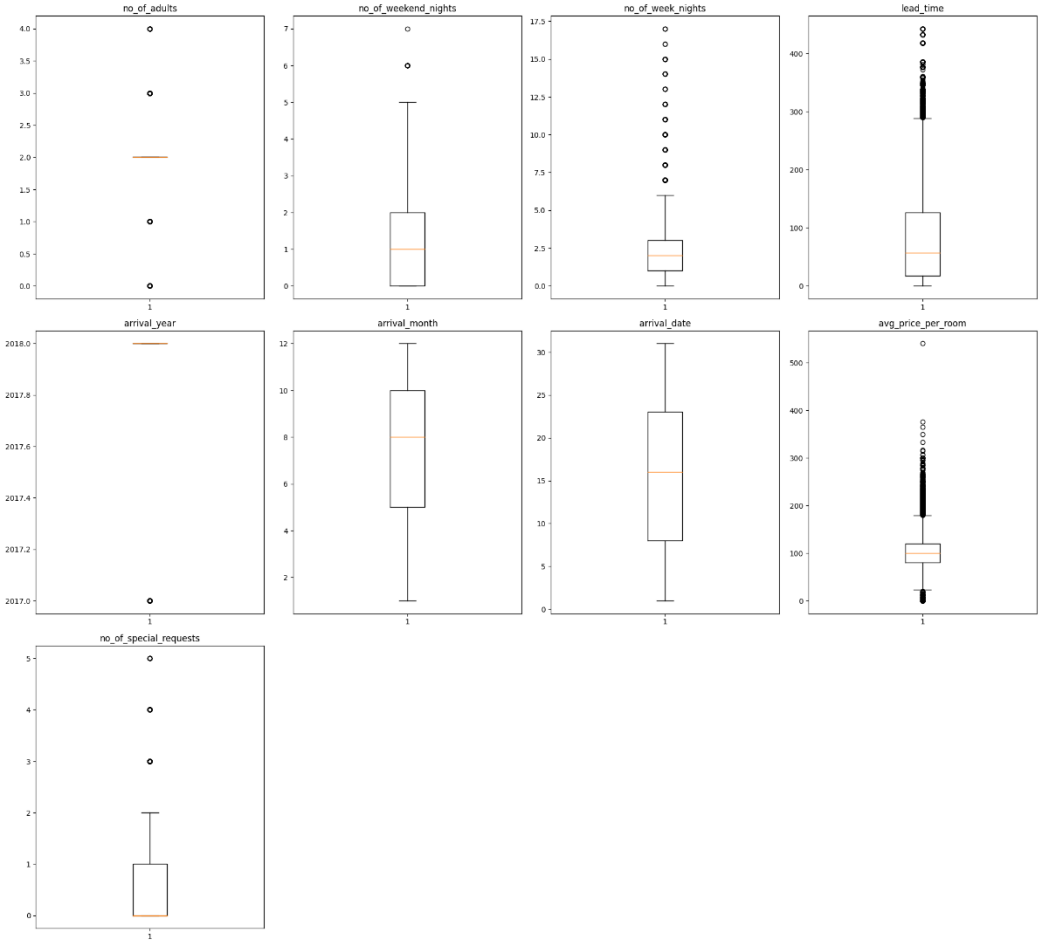

Figure 19: Outliers in the date

- No of week / weekend nights, no of special request, and lead time has upper outliers.
- Average price per room has outliers at both the tails.

6.1.3. Outliers Treatment

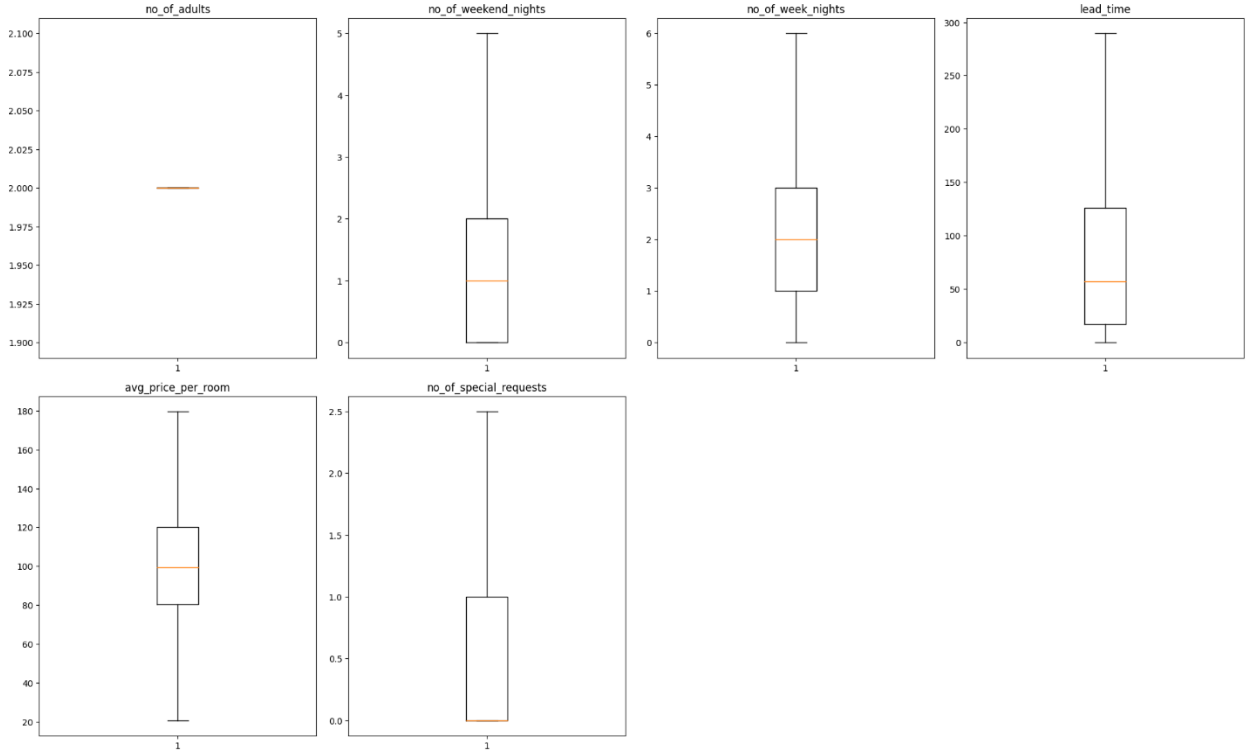

Figure 20: Outliers Treatment

- We have treated all the values smaller than Lower Whisker will be assigned the value of Lower Whisker and all the values greater than Upper Whisker will be assigned the value of Upper Whisker.
- We should not remove the outliers which are present in the date/age column. Hence, we did not remove outliers from arrival year, arrival month, arrival date columns.

6.1.4. Data Preprocessing and splitting training and testing data

Encoding not cancelled as 0 and cancelled as 1 as the INN hotel authority wants the reason behind booking cancellation.

- Shape of Training set : (25392, 30)
- Shape of test set : (10883, 30)
- Percentage of classes in training set:

  booking_status

  0 0.670644

  1 0.329356

- Percentage of classes in test set:

  booking_status

  0 0.676376

  1 0.323624

6.2. Model Building - Logistic Regression

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25362
Method:	MLE	Df Model:	29
Date:	Sun, 04 Aug 2024	Pseudo R-squ.:	0.3180
Time:	15:16:22	Log-Likelihood:	-10974.
converged:	False	LL-Null:	-16091.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
no_of_adults	-0.0019	57.611	-3.23e-05	1.000	-112.917	112.913
no_of_weekend_nights	0.0958	0.019	4.922	0.000	0.058	0.134
no_of_week_nights	0.0324	0.013	2.416	0.016	0.006	0.059
lead_time	0.0167	0.000	61.160	0.000	0.016	0.017
arrival_year	-0.0020	0.057	-0.036	0.972	-0.114	0.110
arrival_month	-0.0583	0.007	-8.965	0.000	-0.071	-0.046
arrival_date	0.0017	0.002	0.904	0.366	-0.002	0.006
avg_price_per_room	0.0203	0.001	26.613	0.000	0.019	0.022
no_of_special_requests	-1.4687	0.030	-49.718	0.000	-1.527	-1.411
type_of_meal_plan_Meal Plan 2	0.2457	0.064	3.842	0.000	0.120	0.371
type_of_meal_plan_Meal Plan 3	0.0076	1.985	0.004	0.997	-3.883	3.898
type_of_meal_plan_Not Selected	0.3720	0.053	7.048	0.000	0.269	0.475
room_type_reserved_Room_Type 2	-0.1833	0.124	-1.476	0.140	-0.427	0.060
room_type_reserved_Room_Type 3	6.327e-05	1.239	5.11e-05	1.000	-2.428	2.428
room_type_reserved_Room_Type 4	-0.2164	0.051	-4.208	0.000	-0.317	-0.116
room_type_reserved_Room_Type 5	-0.1655	0.198	-0.835	0.404	-0.554	0.223
room_type_reserved_Room_Type 6	-0.3681	0.111	-3.314	0.001	-0.586	-0.150
room_type_reserved_Room_Type 7	-0.0701	0.264	-0.266	0.790	-0.587	0.447
market_segment_type_Complementary	-0.0729	0.600	-0.122	0.903	-1.249	1.103
market_segment_type_Corporate	-0.0891	0.346	-0.258	0.797	-0.767	0.589
market_segment_type_Offline	-0.9137	0.336	-2.716	0.007	-1.573	-0.254
market_segment_type_Online	0.8830	0.334	2.647	0.008	0.229	1.537
repeated_guest_Yes	-0.3244	0.842	-0.385	0.700	-1.975	1.326
booking_label_Low	-0.1993	1.069	-0.187	0.852	-2.294	1.895
booking_label_Medium	-0.0398	1.160	-0.034	0.973	-2.314	2.235
booking_label_Very High	-0.0075	1.928	-0.004	0.997	-3.786	3.771
booking_label_Very Low	0.2695	1.106	0.244	0.808	-1.899	2.438
cancellation_label_Frequent Cancellations	-0.0005	0.944	-0.001	1.000	-1.850	1.849
cancellation_label_Moderate Cancellations	-6.172e-05	27.103	-2.28e-06	1.000	-53.121	53.121
cancellation_label_No Cancellations	0.1058	0.695	0.152	0.879	-1.257	1.469

Figure 21: Logistic Regression model before treating multicollinearity

Observations

- Negative values of the coefficient show that the probability of a booking to get canceled decreases with the increase of the corresponding attribute value.
- Positive values of the coefficient show that the probability of a booking to get canceled increases with the increase of the corresponding attribute value.
- p-value of a variable indicates if the variable is significant or not. If we consider the significance level to be 0.05 (5%), then any variable with a p-value less than 0.05 would be considered significant.

6.2.1. Model Performance Evaluation

Model can make wrong predictions as:

1. Predicting a booking status as cancelled but in reality the booking status is not cancelled.
2. Predicting a booking status as not cancelled but in reality the booking status is cancelled.

Which case is more important?

- If we predict that a booking will not get cancelled but in reality, the guest cancels the booking, then the hotel company has to face losses due to cancellation of booking in following ways
    - 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.
- If we predict that a booking will get cancelled but in reality, the guest does not cancels the booking, then the company will have to bear the cost of inspection
    - Due to wrong prediction the booking will not take place and then hotel take care cost will increase.

How to reduce this loss?

- We need to reduce both False Negatives and False Positives
- f1_score should be maximized as the greater the f1_score, the higher the chances of reducing both False Negatives and False Positives and identifying both the classes correctly

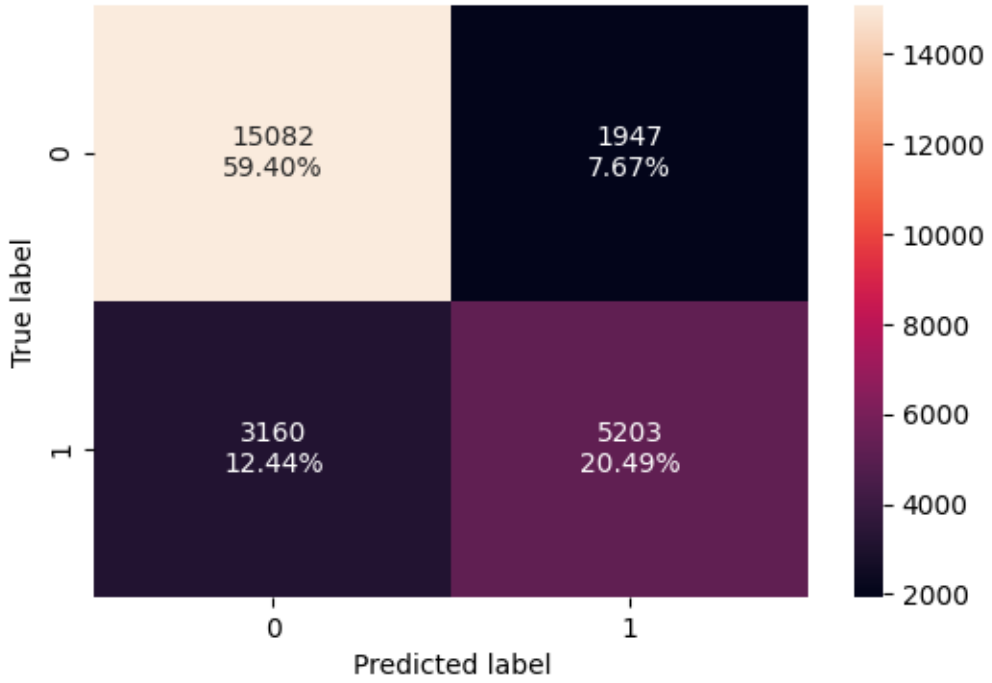

Figure 22: Confusion Matrix before treating multicollinearity

- Here 12.44% of our data has been misclassified under false negative and 7.67% data has been misclassified under false positive.

Training performance:

	Accuracy	Recall	Precision	F1
0	0.798874	0.622145	0.727692	0.670792

Table 2: Training set performance before treating multicollinearity

Observations

- The f1_score of the model is ~0.67 and we will try to maximize it further
- The variables used to build the model might contain multicollinearity, which will affect the p-values
- We will have to remove multicollinearity from the data to get reliable coefficients and p-values

6.2.2. Detecting and Dealing with Multicollinearity

There are different ways of detecting (or testing for) multicollinearity. One such way is using the Variation Inflation Factor (VIF).

- **Variance Inflation factor**: Variance inflation factors measure the inflation in the variances of the regression coefficients estimates due to collinearities that exist among the predictors. It is a measure of how much the variance of the estimated regression coefficient βk is "inflated" by the existence of correlation among the predictor variables in the model.
- **General Rule of thumb**:
    - If VIF is 1 then there is no correlation among the kth predictor and the remaining predictor variables, and hence the variance of βk is not inflated at all
    - If VIF exceeds 5, we say there is moderate multicollinearity
    - If VIF is equal or exceeding 10, it shows signs of high multi-collinearity
- The purpose of the analysis should dictate which threshold to use

```
no_of_adults                              3.961279e+07
no_of_weekend_nights                      1.050307e+00
no_of_week_nights                         1.091266e+00
lead_time                                 1.372143e+00
arrival_year                              1.435178e+00
arrival_month                             1.273729e+00
arrival_date                              1.006523e+00
avg_price_per_room                        1.881548e+00
no_of_special_requests                    1.235526e+00
type_of_meal_plan_Meal Plan 2             1.263667e+00
type_of_meal_plan_Meal Plan 3             1.025023e+00
type_of_meal_plan_Not Selected            1.267547e+00
room_type_reserved_Room_Type 2            1.032519e+00
room_type_reserved_Room_Type 3            1.003285e+00
room_type_reserved_Room_Type 4            1.300120e+00
room_type_reserved_Room_Type 5            1.028277e+00
room_type_reserved_Room_Type 6            1.226133e+00
room_type_reserved_Room_Type 7            1.067304e+00
market_segment_type_Complementary         4.420189e+00
market_segment_type_Corporate             1.687804e+01
market_segment_type_Offline               6.342970e+01
market_segment_type_Online                7.040308e+01
repeated_guest_Yes                        1.535815e+01
booking_label_Low                         1.000514e+01
booking_label_Medium                      3.729003e+00
booking_label_Very High                   1.565878e+00
booking_label_Very Low                    2.121260e+01
cancellation_label_Frequent Cancellations 1.192417e+00
cancellation_label_Moderate Cancellations 1.050292e+00
cancellation_label_No Cancellations       2.823621e+00
dtype: float64
```

Figure 23: VIF Score before treating multicollinearity

- Sme categorical levels of market_segment_type, and booking_label exhibit high multicollinearity

Removing market_segment_type_Online:

```
no_of_adults                              3.952654e+07
no_of_weekend_nights                      1.050077e+00
no_of_week_nights                         1.091236e+00
lead_time                                 1.365546e+00
arrival_year                              1.432406e+00
arrival_month                             1.272339e+00
arrival_date                              1.006521e+00
avg_price_per_room                        1.878098e+00
no_of_special_requests                    1.228129e+00
type_of_meal_plan_Meal Plan 2             1.263307e+00
type_of_meal_plan_Meal Plan 3             1.025023e+00
type_of_meal_plan_Not Selected            1.264818e+00
room_type_reserved_Room_Type 2            1.032414e+00
room_type_reserved_Room_Type 3            1.003285e+00
room_type_reserved_Room_Type 4            1.297708e+00
room_type_reserved_Room_Type 5            1.028277e+00
room_type_reserved_Room_Type 6            1.226132e+00
room_type_reserved_Room_Type 7            1.067304e+00
market_segment_type_Complementary         1.297087e+00
market_segment_type_Corporate             1.515237e+00
market_segment_type_Offline               1.618406e+00
repeated_guest_Yes                        1.535797e+01
booking_label_Low                         1.000354e+01
booking_label_Medium                      3.728993e+00
booking_label_Very High                   1.565877e+00
booking_label_Very Low                    2.121229e+01
cancellation_label_Frequent Cancellations 1.192304e+00
cancellation_label_Moderate Cancellations 1.050292e+00
cancellation_label_No Cancellations       2.823511e+00
dtype: float64
```

Figure 24: VIF Score after removing market segment type online

- Now our model has very less multicollinearity. And we can proceed for removing variables having p value more than 0.05.

Training performance:

	Accuracy	Recall	Precision	F1
0	0.800725	0.623221	0.73192	0.673211

Table 3: Training set performance after treating multicollinearity

- No significant change in the model performance

Observations:

1. Dropping market_segment_type_Online doesn't have a significant impact on the model performance.
2. We can choose any model to proceed to the next steps.
3. Some of the categorical levels of a variable have VIF< 5 which can simply be ignored.

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25363
Method:	MLE	Df Model:	28
Date:	Sun, 04 Aug 2024	Pseudo R-squ.:	0.3183
Time:	16:00:55	Log-Likelihood:	-10969.
converged:	False	LL-Null:	-16091.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
no_of_adults	-0.0018	58.224	-3.09e-05	1.000	-114.118	114.114
no_of_weekend_nights	0.1009	0.019	5.185	0.000	0.063	0.139
no_of_week_nights	0.0294	0.013	2.186	0.029	0.003	0.056
lead_time	0.0169	0.000	61.521	0.000	0.016	0.017
arrival_year	-0.0017	0.058	-0.029	0.977	-0.115	0.111
arrival_month	-0.0590	0.007	-9.045	0.000	-0.072	-0.046
arrival_date	0.0011	0.002	0.584	0.559	-0.003	0.005
avg_price_per_room	0.0205	0.001	26.849	0.000	0.019	0.022
no_of_special_requests	-1.4679	0.030	-49.649	0.000	-1.526	-1.410
type_of_meal_plan_Meal Plan 2	0.2669	0.064	4.158	0.000	0.141	0.393
type_of_meal_plan_Meal Plan 3	0.0064	1.829	0.003	0.997	-3.578	3.591
type_of_meal_plan_Not Selected	0.3734	0.053	7.062	0.000	0.270	0.477
room_type_reserved_Room_Type 2	-0.1137	0.124	-0.917	0.359	-0.357	0.129
room_type_reserved_Room_Type 3	-0.0002	1.195	-0.000	1.000	-2.343	2.342
room_type_reserved_Room_Type 4	-0.2111	0.051	-4.110	0.000	-0.312	-0.110
room_type_reserved_Room_Type 5	-0.1961	0.198	-0.988	0.323	-0.585	0.193
room_type_reserved_Room_Type 6	-0.3409	0.111	-3.067	0.002	-0.559	-0.123
room_type_reserved_Room_Type 7	-0.0741	0.263	-0.282	0.778	-0.589	0.441
market_segment_type_Complementary	-0.1036	0.372	-0.278	0.781	-0.832	0.625
market_segment_type_Corporate	-0.7938	0.100	-7.936	0.000	-0.990	-0.598
market_segment_type_Offline	-1.8287	0.052	-35.030	0.000	-1.931	-1.726
repeated_guest_Yes	-0.4104	0.862	-0.476	0.634	-2.099	1.279
booking_label_Low	-0.2600	1.106	-0.235	0.814	-2.427	1.907
booking_label_Medium	-0.0564	1.194	-0.047	0.962	-2.396	2.283
booking_label_Very High	-0.0102	1.979	-0.005	0.996	-3.889	3.868
booking_label_Very Low	0.3535	1.139	0.310	0.756	-1.878	2.585
cancellation_label_Frequent Cancellations	0.0021	1.005	0.002	0.998	-1.968	1.972
cancellation_label_Moderate Cancellations	-7.919e-05	28.034	-2.82e-06	1.000	-54.946	54.946
cancellation_label_No Cancellations	0.1308	0.708	0.185	0.853	-1.257	1.518

Figure 25: Logistic Regression model after treating multicollinearity

6.2.3. Removing insignificant variables with High P value

- For other attributes present in the data, the p-values are high only for few dummy variables and since only one (or some) of the categorical levels have a high p-value we will drop them iteratively as sometimes p-values change after dropping a variable. So, we'll not drop all variables at once.
- Instead, we will do the following repeatedly using a loop:
    - Build a model, check the p-values of the variables, and drop the column with the highest p-value.
    - Create a new model without the dropped feature, check the p-values of the variables, and drop the column with the highest p-value.
    - Repeat the above two steps till there are no columns with p-value > 0.05.
- Note: The above process can also be done manually by picking one variable at a time that has a high p-value, dropping it, and building a model again. But that might be a little tedious and using a loop will be more efficient.
- After removing the higher p values below are the variables which we have considered in our logistic regression model.

'no_of_weekend_nights', 'lead_time', 'arrival_year', 'arrival_month', '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 4',

'room_type_reserved_Room_Type 6', 'market_segment_type_Corporate', 'market_segment_type_Offline', 'repeated_guest_Yes'

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25379
Method:	MLE	Df Model:	12
Date:	Sun, 04 Aug 2024	Pseudo R-squ.:	0.3177
Time:	16:05:06	Log-Likelihood:	-10979.
converged:	False	LL-Null:	-16091.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
no_of_weekend_nights	0.1283	0.019	6.604	0.000	0.090	0.166
lead_time	0.0171	0.000	65.549	0.000	0.017	0.018
arrival_year	-0.0014	4.55e-05	-31.287	0.000	-0.002	-0.001
arrival_month	-0.0600	0.006	-10.048	0.000	-0.072	-0.048
avg_price_per_room	0.0210	0.001	29.259	0.000	0.020	0.022
no_of_special_requests	-1.5013	0.030	-50.300	0.000	-1.560	-1.443
type_of_meal_plan_Meal Plan 2	0.2450	0.062	3.930	0.000	0.123	0.367
type_of_meal_plan_Not Selected	0.3572	0.052	6.882	0.000	0.255	0.459
room_type_reserved_Room_Type 4	-0.2022	0.051	-3.989	0.000	-0.301	-0.103
room_type_reserved_Room_Type 6	-0.4207	0.111	-3.799	0.000	-0.638	-0.204
market_segment_type_Corporate	-1.0105	0.104	-9.763	0.000	-1.213	-0.808
market_segment_type_Offline	-1.8570	0.052	-35.881	0.000	-1.958	-1.756
repeated_guest_Yes	-0.6097	0.225	-2.708	0.007	-1.051	-0.168

Figure 26: Logistic Regression model after treating variables with p value more than 0.05

Now no categorical feature has p-value greater than 0.05, so we'll consider the features in *X_train2* as the final ones and *lg3* as final model.

6.2.4. Coefficient Interpretations

- Coefficient of some levels of meal plan, weekend night, lead time and avg price per room are positive an increase in these will lead to increase in chances of booking get cancelled.
- Coefficient of some levels of market segment types, room types, special request, arrival date, and repeated guests are negative increase in these will lead to decrease in chances of booking get cancelled.

Converting coefficients to odds

- The coefficients (βs) of the logistic regression model are in terms of log(odds) and to find the odds, we have to take the exponential of the coefficients
- Therefore, odds=exp(β)
- The percentage change in odds is given as (exp(β)−1)∗100

Coefficient interpretations

- arrival year: Holding all other features constant a 1 unit change in the arrival year will decrease the odds of a booking gets canceled by ~1 times or a decrease of ~0.1% decrease in odds of a booking gets canceled.
- arrival month: Holding all other features constant a 1 unit change in the arrival year will decrease the odds of a booking gets canceled by ~0.94 times or a decrease of ~6% decrease in odds of a booking gets canceled.
- Special requests: Holding all other features constant a 1 unit change in the arrival year will decrease the odds of a booking gets canceled by ~0.22 times or a decrease of ~78% decrease in odds of a booking gets canceled.

- Repeated guests: Holding all other features constant a 1 unit change in the case of repeated guests will decrease the odds of a booking gets canceled by ~0.54 times or a decrease of ~46% decrease in odds of a booking gets canceled.
- room type reserved: The odds of a booking reserved for type 4 room having a cancellation is ~0.82 less than the booking which were done for other types (1-3,5) of rooms or ~18.3% fewer odds of cancellation than the booking which were done for other types of rooms. Similarly, a booking reserved for type 6 room having a cancellation is is ~0.65 times less than the booking which were done for other types of rooms or ~34.3% fewer odds of cancellation than the booking which were done for other types of rooms. [the dropped category type 1-3 and 5 is taken as a reference level]
- market segment type: The odds of a booking done at corporate market segment having cancellation of booking is ~0.36 times less than the booking which was done at online or complementary market segment having cancellation of booking or ~63.5% fewer odds of cancellation than the booking which were done at online or complementary market segment. Similarly, the odds of a booking done at offline market segment having cancellation of booking is ~0.15 times less than the booking which was done at online or complementary market segment having cancellation of booking or ~84.3% fewer odds of cancellation than the booking which were done at online or complementary market segment. [the dropped categories online and complementary are taken as a reference level]

6.2.5. Checking performance of the new model

- Test data confusion matrix:

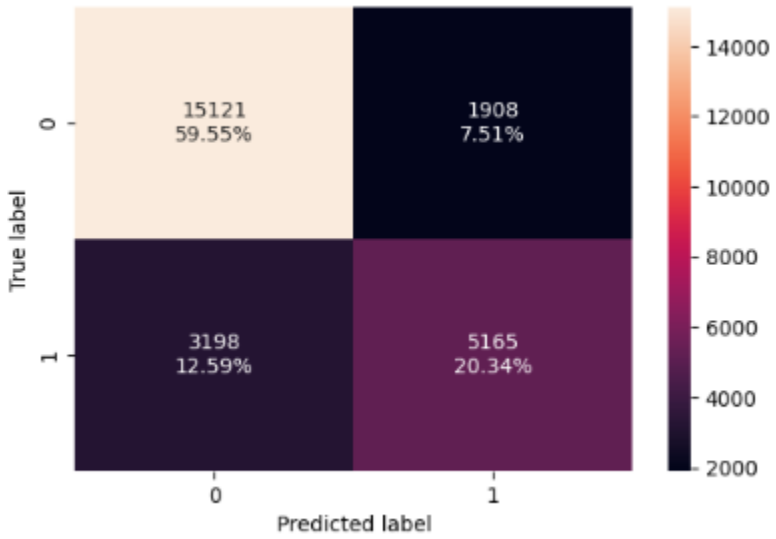

Figure 27:Confusion matrix of X_train2 and y_train

- Training performance:

	Accuracy	**Recall**	**Precision**	**F1**
0	0.798913	0.617601	0.730242	0.669215

Table 4: Training set performance of our final training data

- Test data confusion matrix:

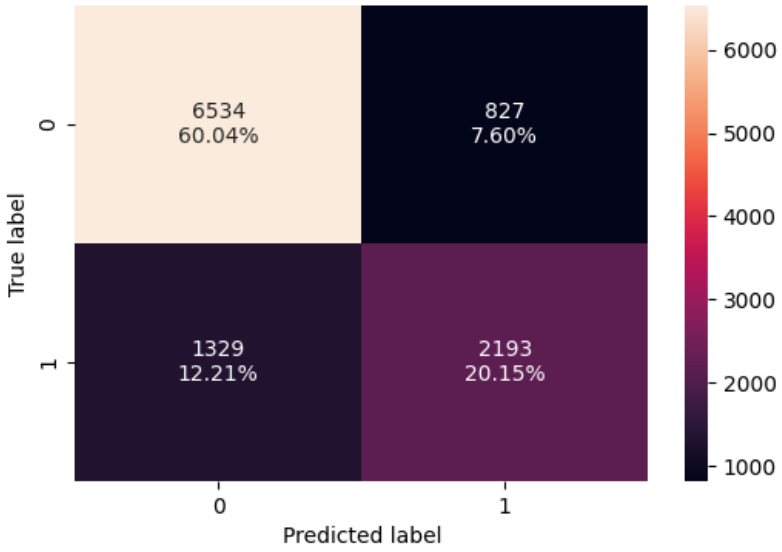

Figure 28: Confusion matrix of X_test2 and y_test

Test performance:

	Accuracy	Recall	Precision	F1
0	0.801893	0.622658	0.726159	0.670437

Figure 29: Testing set performance of our final testing data

- The model is giving a good f1_score of ~0.67 on the train and test sets respectively.
- As the train and test performances are comparable, the model is not overfitting.
- Moving forward we will try to improve the performance of the model.

6.3. Model Performance Improvement

- Let's see if the f1_score can be improved further by changing the model threshold
- First, we will check the ROC curve, compute the area under the ROC curve (ROC-AUC), and then use it to find the optimal threshold
- Next, we will check the Precision-Recall curve to find the right balance between precision and recall as our metric of choice is f1_score

6.3.1. ROC Curve and ROC-AUC

Checking model performance on training set

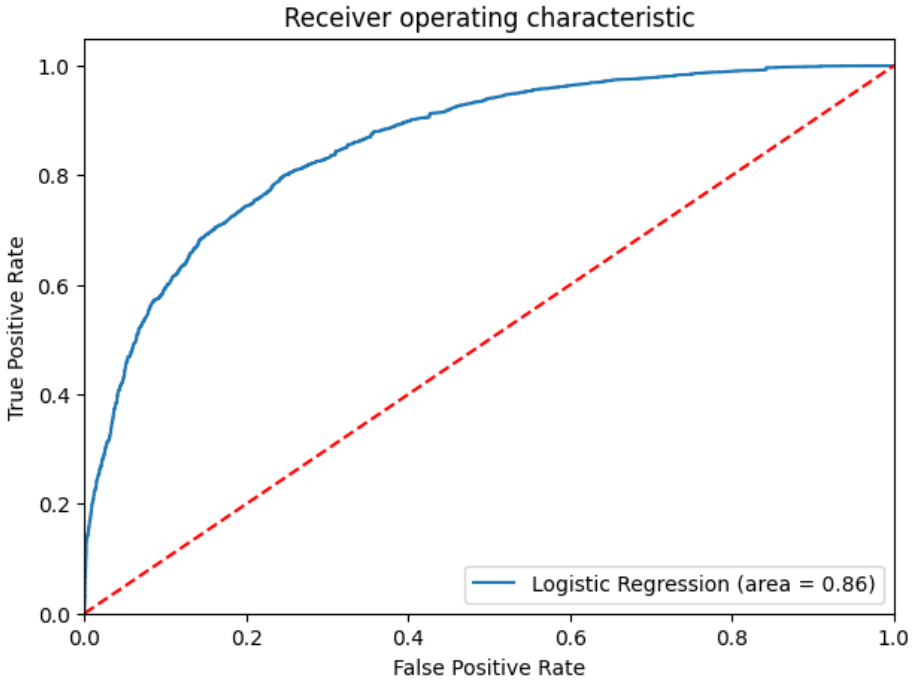

Figure 30: ROC-AUC on training set

- Logistic Regression model is giving a good performance on training set.
- Optimal threshold for training data set is 0.2955693634213808.

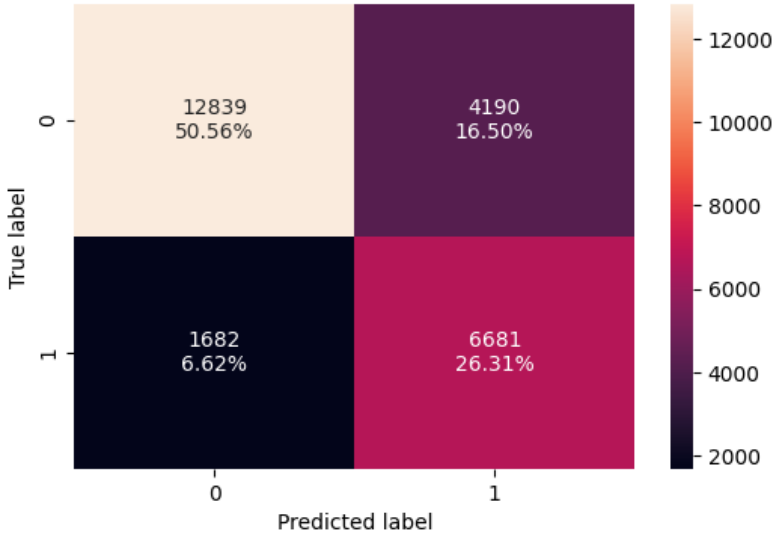

Figure 31: Confusion matrix of training data set at optimal threshold value of 0.29

Training performance:

	Accuracy	Recall	Precision	F1
0	0.768746	0.798876	0.614571	0.694707

Table 5: Training data set performance at optimal threshold value of 0.29

- Recall has increased, accuracy and precision has reduced.
- The model is still giving a good performance.

Checking model performance on test set

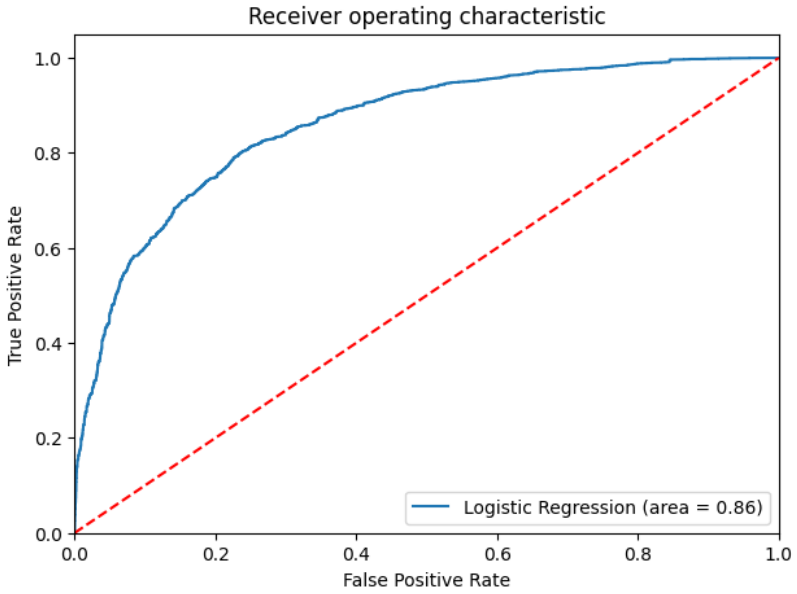

Figure 32: ROC-AUC on testing set

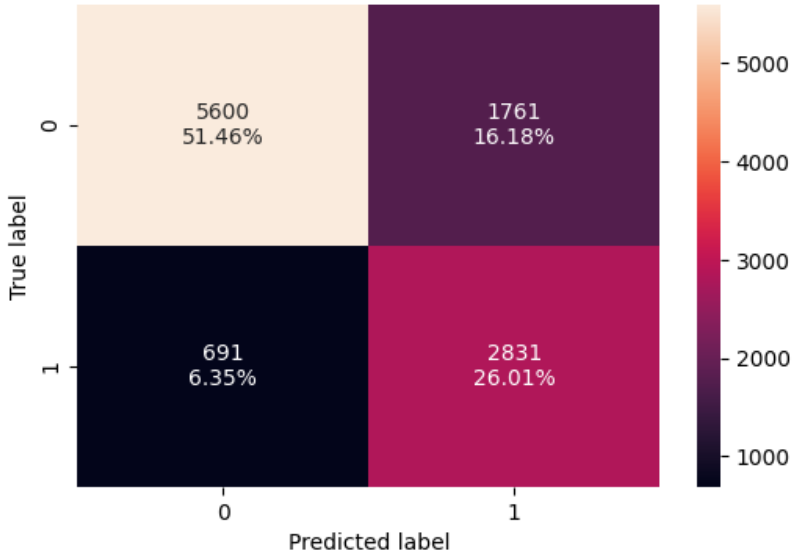

Figure 33: Confusion matrix of testing data set at optimal threshold value of 0.29

Test performance:

	Accuracy	Recall	Precision	F1
0	0.774694	0.803805	0.616507	0.697806

Table 6: Testing data set performance at optimal threshold value of 0.29

6.2.2. Precision-Recall Curve

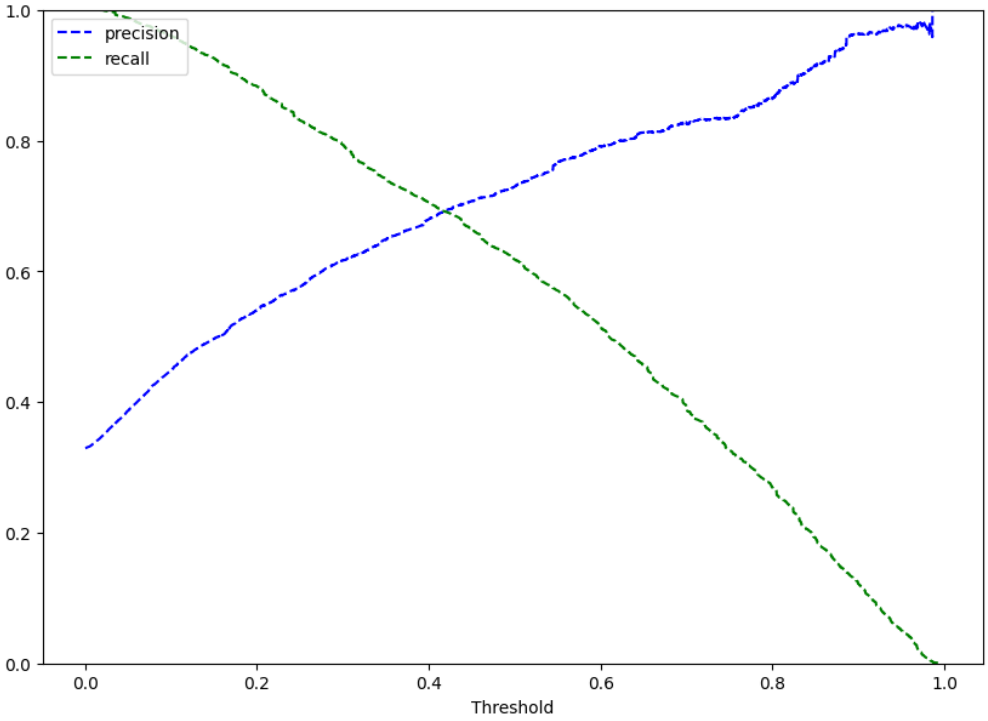

Figure 34: Precision-Recall Curve

- At the threshold of 0.42, we get balanced recall and precision.

Checking model performance on training set

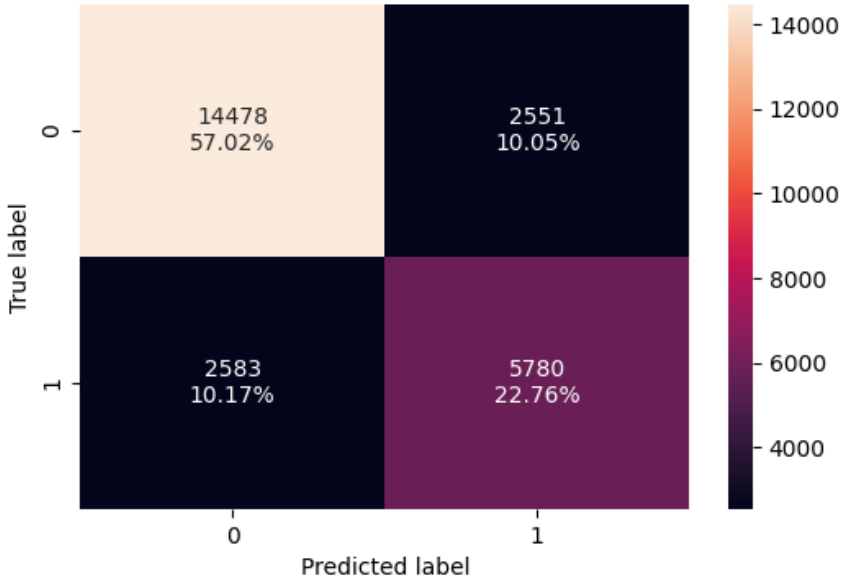

Figure 35: Confusion Matrix of training data at threshold of 0.42

Training performance:

	Accuracy	Recall	Precision	F1
0	0.79781	0.69114	0.693794	0.692464

Table 7: Training data set performance at optimal threshold value of 0.42

- Model is performing well on training set.
- There's not much improvement in the model performance as the default threshold is 0.50 and here we get 0.42 as the optimal threshold.

Checking model performance on test set

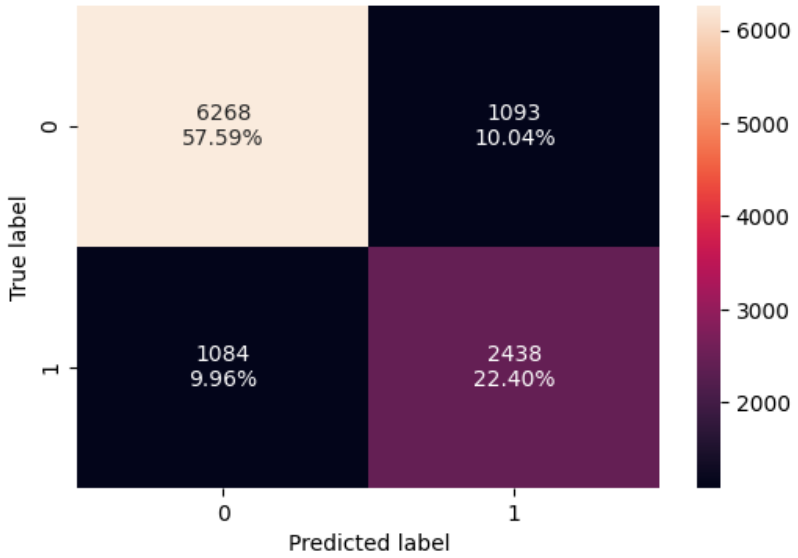

Figure 34: Confusion Matrix of testing data at threshold of 0.42

Test performance:

	Accuracy	Recall	Precision	F1
0	0.799963	0.69222	0.690456	0.691337

Table 8: Testing data set performance at optimal threshold value of 0.42

6.4. Model Performance Comparison and Final Model Selection

- Training performance comparison:

	Logistic Regression-default Threshold (0.5)	Logistic Regression-0.81 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.798913	0.768746	0.797810
Recall	0.617601	0.798876	0.691140
Precision	0.730242	0.614571	0.693794

	Logistic Regression-default Threshold (0.5)	Logistic Regression-0.81 Threshold	Logistic Regression-0.42 Threshold
F1	0.669215	0.694707	0.69246

Table 9: Training set performance at different thresholds

- Testing performance comparison:

	Logistic Regression-default Threshold (0.5)	Logistic Regression-0.81 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.801893	0.774694	0.799963
Recall	0.622658	0.803805	0.692220
Precision	0.726159	0.616507	0.690456
F1	0.670437	0.697806	0.691337

Table 10: Testing set performance at different thresholds

- Almost all the three models are performing well on both training and test data without the problem of overfitting.
- The model with a threshold (0.81) is giving the best F1 score (~0.68). Also, recall and precision are optimized. Therefore, it can be selected as the final model.

7. KNN Classifier and Naïve Bayes

7.1. Normalizing the numerical variables

The 'lead_time', 'avg_price_per_room' we have kept as a numerical variable.

7.2. Train Test Split

- Shape of Training set: (25392, 30)
- Shape of test set: (10883, 30)
- Percentage of classes in training set:

booking_status

0 0.670644

1 0.329356

- Percentage of classes in test set:

booking_status

0 0.676376

1 0.323624

7.3. Model Evaluation

Model evaluation criterion

Model can make wrong predictions as:

- Predicting a booking will not get cancelled but in reality, the booking gets cancelled (FN)

- Predicting a booking will get cancelled but in reality, the booking does not cancel (FP)

Which case is more important?

- If we predict that a booking will not get cancelled but in reality, the guest cancels the booking, then the hotel company has to face losses due to cancellation of booking in following ways
    - 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.
- If we predict that a booking will get cancelled but in reality, the guest does not cancels the booking, then the company will have to bear the cost of inspection
    - Due to wrong prediction the booking will not take place and then hotel take care cost will increase.

How to reduce the losses?

The company would want the recall to be maximized, greater the recall score higher are the chances of minimizing the False Negatives.

7.4. Model Building – KNN (K- Nearest Neighbor) Classifier

In order to optimize our model, it's essential to experiment with different values of k to find the most suitable fit for our data. We can commence this process by setting k equal to 3 and gradually exploring other values to assess their impact on the model's performance.

- We'll only consider odd values of K as the classification will be done based on majority voting.

7.4.1. K=3

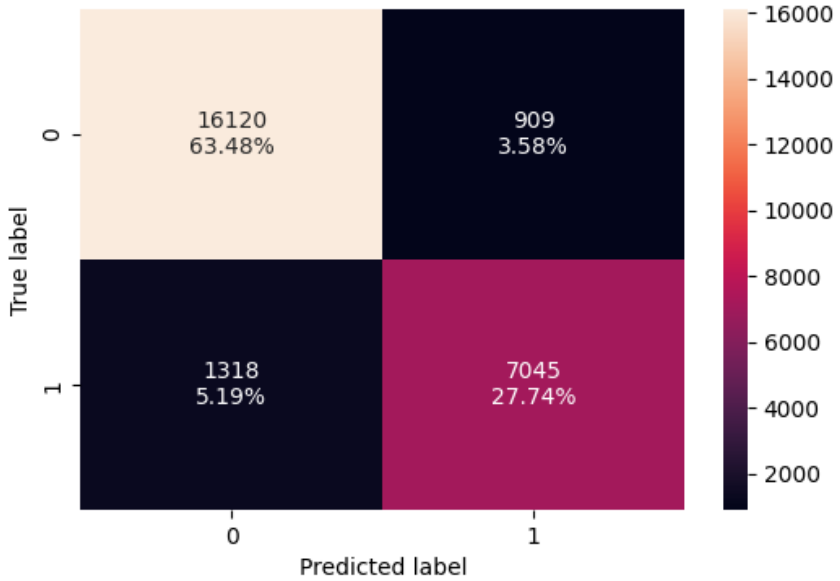

Figure 36: Confusion matrix of training data for KNN Classification where k=3

	Accuracy	Recall	Precision	F1
0	0.912295	0.842401	0.885718	0.863517

Table 11: Performance of the training data at k=3

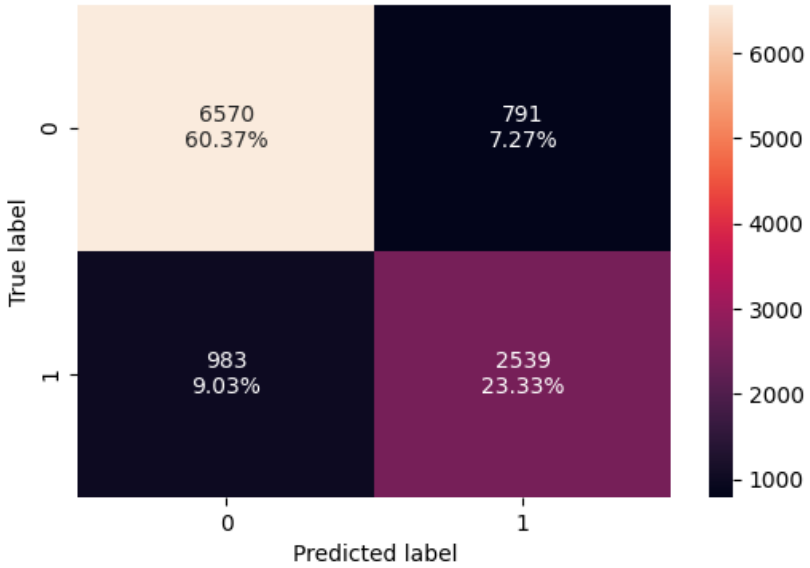

Figure 37: Performance of testing data for KNN Classification (where k=3)

	Accuracy	Recall	Precision	F1
0	0.836993	0.720897	0.762462	0.741097

Table 12: Performance of the testing data at k=3

7.4.2. K with different values

Let's run the KNN with no of neighbors to be 1,3,5..19 and find the optimal number of neighbors from the above list using the recall score.

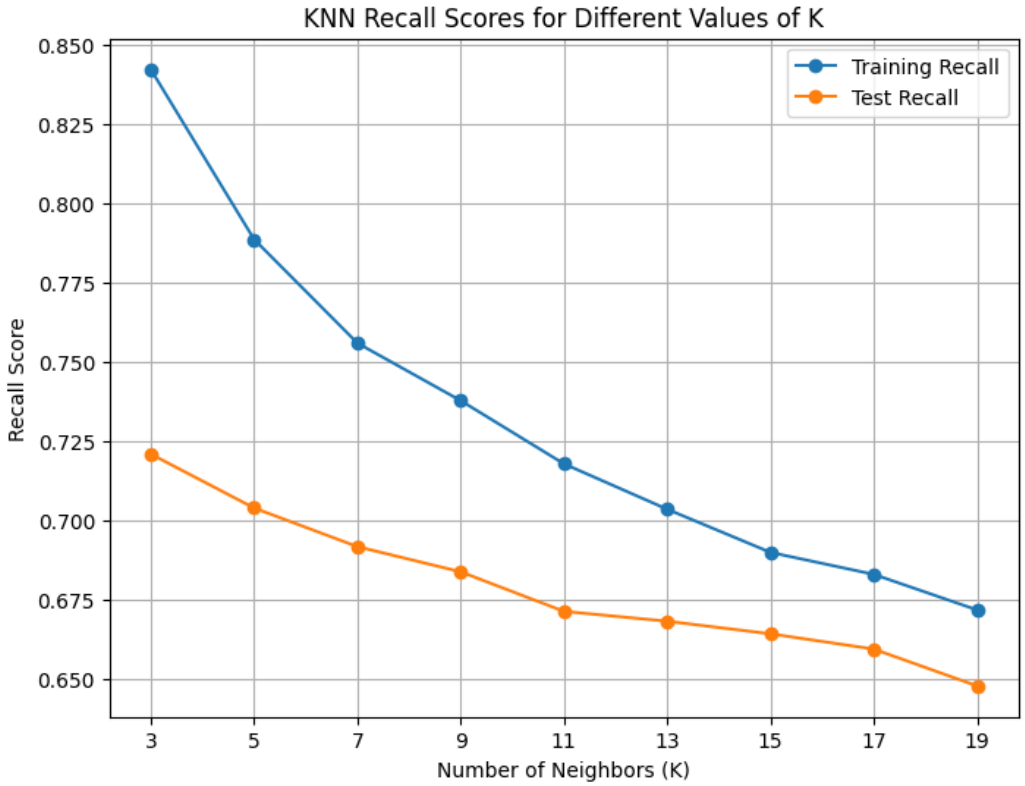

Figure 38: KNN Recall Score for Different Values of K

- The recall scores for both training and test sets are highest when k=3. This suggests that with k=3, the model is better at identifying positive instances in both the training and test data compared to other values of k.
- As the value of k increases beyond 3, the recall scores tend to decrease for both training and test sets. This indicates a potential risk of the model not being able to identify the underlying patterns in the data.
- Therefore, based on the provided recall scores, k=3 appears to be the most suitable choice for balancing model performance between capturing positive instances effectively and generalizing well to new data.

7.5. Model building- Naïve Bayes Classification

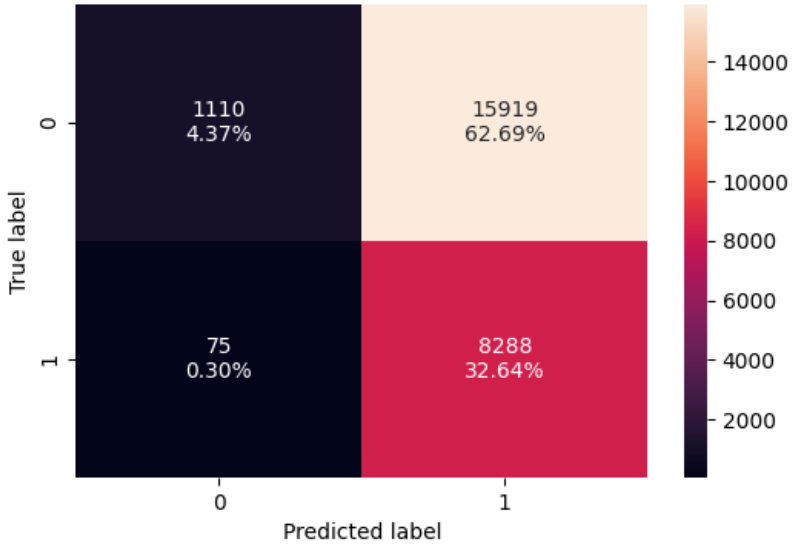

Figure 39: Confusion Matrix of training data in Naive Bayes Classification

	Accuracy	Recall	Precision	F1
0	0.370117	0.991032	0.34238	0.508935

Table 13: Performance of training data in Naïve Bayes Classification

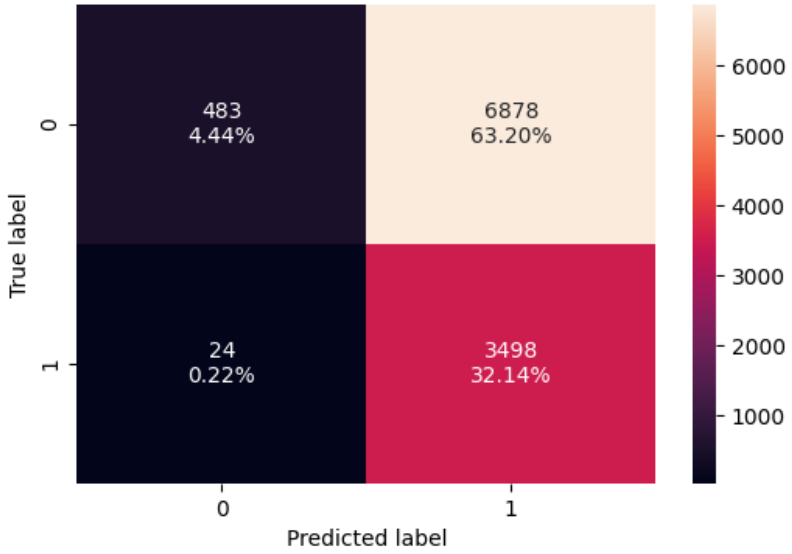

Figure 40: Confusion Matrix of testing data in Naive Bayes Classification

	Accuracy	Recall	Precision	F1
0	0.3658	0.993186	0.337124	0.503382

Table 14: Performance of testing data in Naïve Bayes Classification

7.6. Performance comparison between KNN (k=3) and Naïve Bayes model

Training performance comparison:

	K Nearest Neighbor k=3	Naive Bayes
Accuracy	0.912295	0.370117
Recall	0.842401	0.991032
Precision	0.885718	0.342380
F1	0.863517	0.508935

Table 15: Training set performance comparison between KNN and Naïve Bayes Classification

Test set performance comparison:

	K Nearest Neighbor k=3	Naive Bayes
Accuracy	0.835799	0.358449
Recall	0.716922	0.994889
Precision	0.761689	0.334734
F1	0.738628	0.500929

Table 16: Testing set performance comparison between KNN and Naïve Bayes Classification

- In both the training and test sets, the K Nearest Neighbor model with k=3 demonstrates the highest recall among all compared models. This indicates that the model with k=3 is better at correctly identifying positive instances compared to the models with different k values and Naive Bayes.
- Naive Bayes consistently shows lower recall values compared to K Nearest Neighbor models with different k values. This suggests that Naive Bayes may struggle to capture positive instances as effectively as K Nearest Neighbor models in both training and test datasets, highlighting potential limitations in its performance for this specific task.

8. Decision Tree

8.1. Data Preprocessing

- We labeled no_of_booking_not_cancelled in four different groups
  - Very Low (count< 1)
  - Low (Count 1 to 5)
  - Medium (count 6 to 15)
  - High (count 15 to 30)
- Also, we have labelled no_of_previous_cancellations in three different groups
  - No Cancellations (Count = 0)
  - Few Cancellations (Count = 1 to 5)
  - Moderate Cancellations (Count = 6 to 10)
- We can drop 'no_of_children', 'required_car_parking_space', 'no_of_previous_cancellations', and 'no_of_previous_bookings_not_canceled' as most of the observations are 0 in these variables.
- 'Booking_ID' we can drop as it is a unique value no influence can be observed.
- There are many outliers in the data which we will treat (perform capping of outliers)
  - All the values smaller than the lower whisker will be assigned the value of the lower whisker, and all the values above the upper whisker will be assigned the value of the upper whisker.

8.2. Outliers Detection

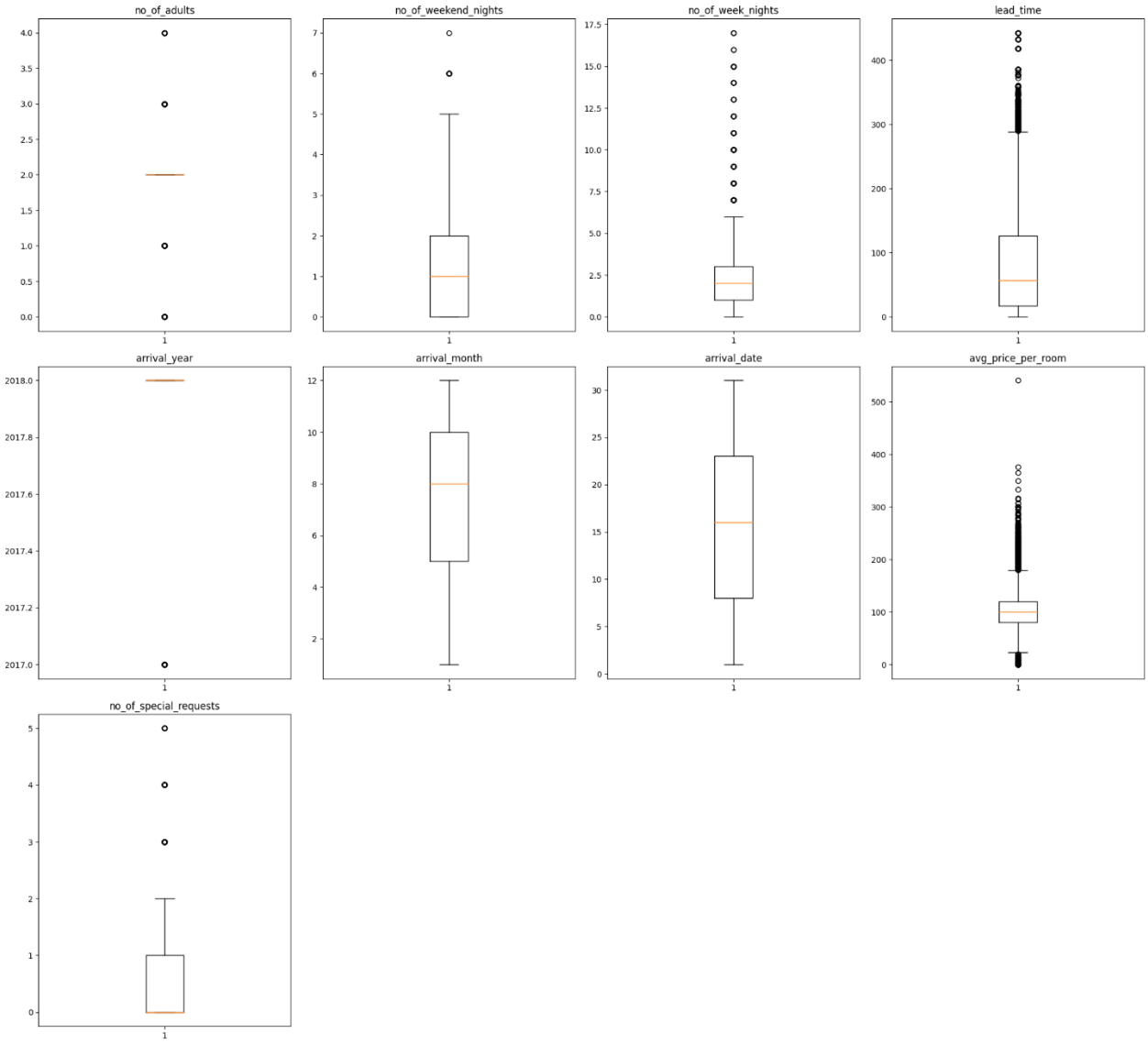

Figure 41: Outliers in the dataset

Observations

- There are quite a few outliers in the data.
- However, we will not treat them as they are proper values

8.3. Data Preparation for Modeling

- We have changed the booking status from objects to integer. Assigned 1 for booking getting canceled and 0 for not canceled.
- Shape of Training set : (25392, 31)
- Shape of test set : (10883, 31)
- Percentage of classes in training set:

  booking_status
  0 0.670644
  1 0.329356
- Percentage of classes in test set:

  booking_status

0 0.676376
1 0.323624

- Around ~67% of observations belong to not canceled and ~33% observations belong to canceled, and this is preserved in the train and test sets.

8.4. Model Building

- We fit our model in decision tree classifier.

8.5. Model Evaluation

Model evaluation criterion

Model can make wrong predictions as:

- Predicting a booking will not get cancelled but in reality, the booking gets cancelled (FN)
- Predicting a booking will get cancelled but in reality, the booking does not cancel (FP)

Which case is more important?

- If we predict that a booking will not get cancelled but in reality, the guest cancels the booking, then the hotel company has to face losses due to cancellation of booking in following ways
    - 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.
- If we predict that a booking will get cancelled but in reality, the guest does not cancels the booking, then the company will have to bear the cost of inspection
    - Due to wrong prediction the booking will not take place and then hotel take care cost will increase.

How to reduce the losses?

The company would want the recall to be maximized, greater the recall score higher are the chances of minimizing the False Negatives.

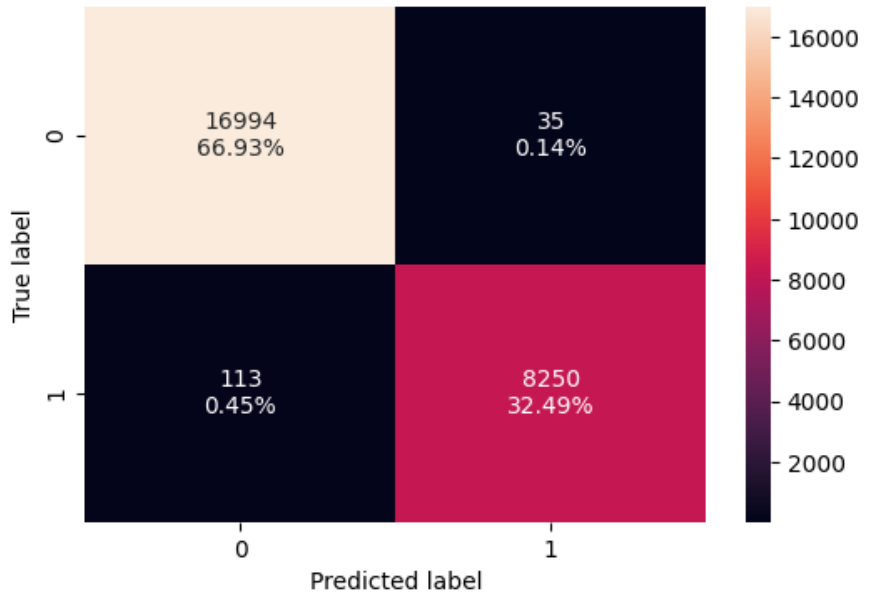

Figure 42: Confusion matrix for Train data (CART)

	Accuracy	Recall	Precision	F1
0	0.994171	0.986488	0.995775	0.99111

Table 17: Model performance for Train data (CART)

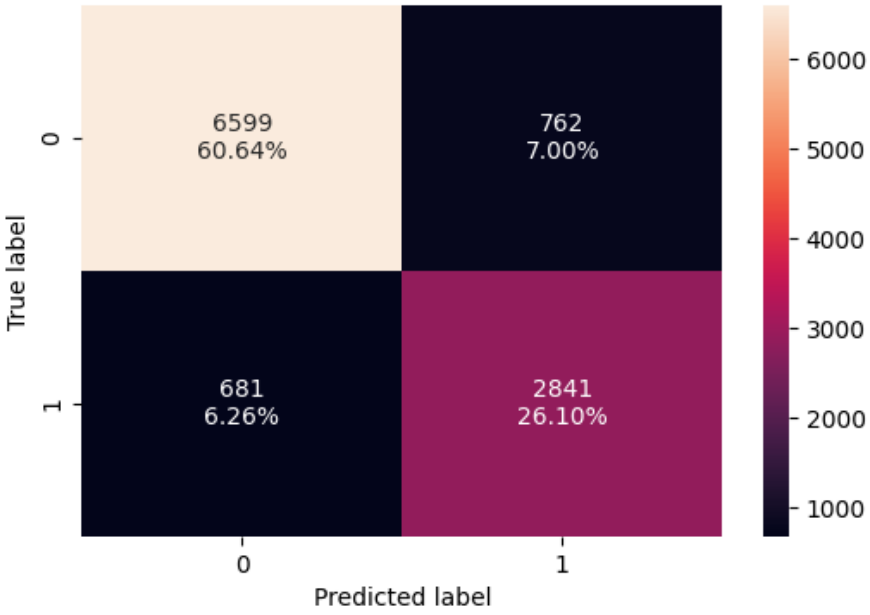

Figure 43: Confusion matrix for Test data (CART)

	Accuracy	Recall	Precision	F1
0	0.867408	0.806644	0.78851	0.797474

Table 18: Model performance for Test data (CART)

- The training set is performing well but the testing set does not show the same accuracy level.

8.5.1. Decision tree with class weight balanced

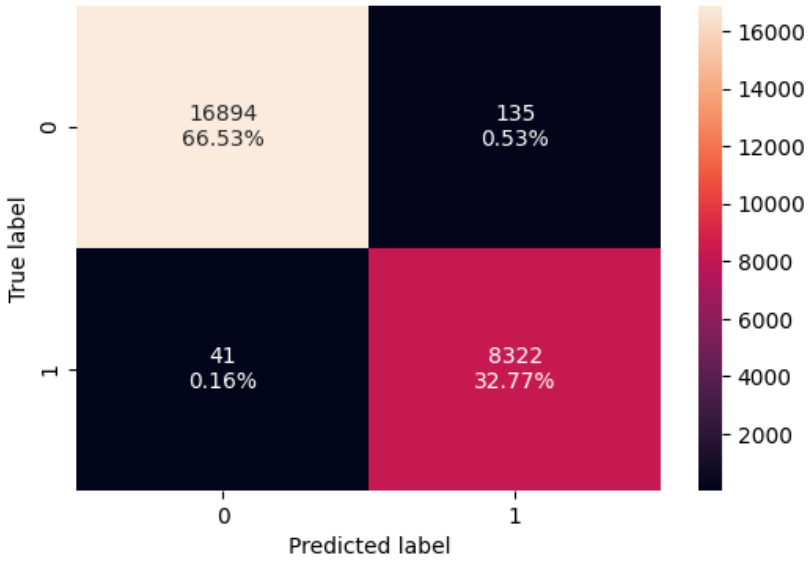

Figure 44: Confusion Matrix for Train data (balanced weight)

	Accuracy	Recall	Precision	F1
0	0.993069	0.995097	0.984037	0.989536

Table 19: Model Performance for Train data (balanced weight)

- Model is able to perfectly classify all the data points on the training set.
- As we know a decision tree will continue to grow and classify each data point correctly if no restrictions are applied as the trees will learn all the patterns in the training set.
- This generally leads to overfitting of the model as Decision Tree will perform well on the training set but will fail to replicate the performance on the test set.

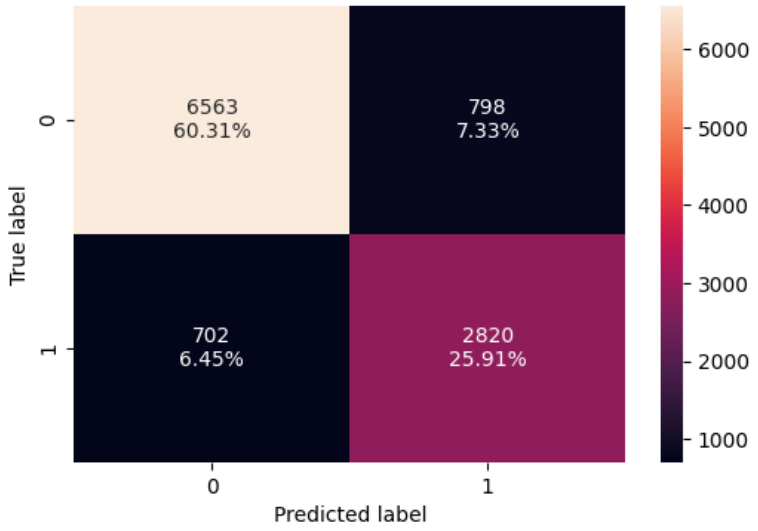

Figure 45: Confusion Matrix for Test data (balanced weight)

	Accuracy	Recall	Precision	F1
0	0.86217	0.800681	0.779436	0.789916

Table 20: Model Performance for Test data (balanced weight)

- There is a huge disparity in performance of model on training set and test set, which suggests that the model is overfitting.
- We need to use pruning techniques to reduce overfitting.

8.5.2. Decision Tree (Pre-pruning)

Using Grid Search for Hyperparameter tuning of our tree model

- Hyperparameter tuning is also tricky in the sense that there is no direct way to calculate how a change in the hyperparameter value will reduce the loss of your model, so we usually resort to experimentation. i.e we'll use Grid search
- Grid search is a tuning technique that attempts to compute the optimum values of hyperparameters.
- It is an exhaustive search that is performed on the specific parameter values of a model.
- The parameters of the estimator/model used to apply these methods are optimized by cross-validated grid-search over a parameter grid.

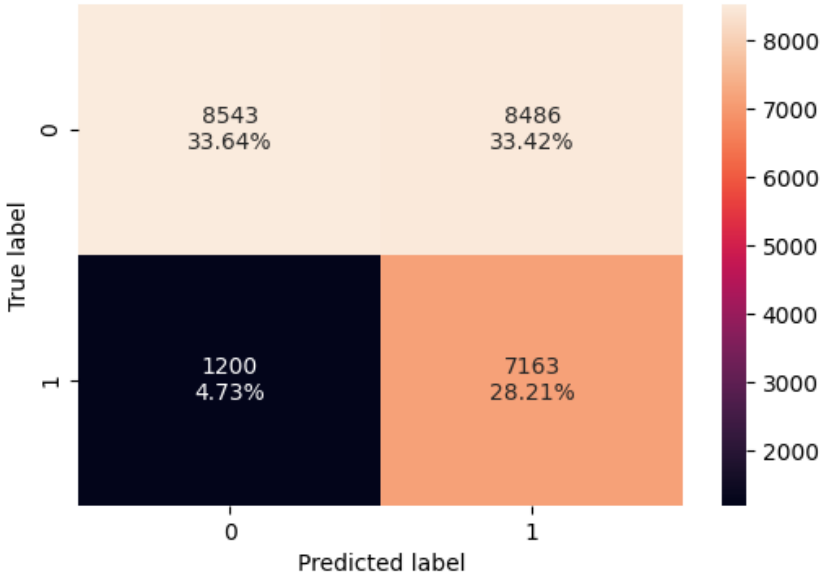

Figure 46: Confusion Matrix for train data (Pre-pruning)

	Accuracy	Recall	Precision	F1
0	0.618541	0.856511	0.457729	0.596618

Table 21: Model Performance for Train data (Pre-pruning)

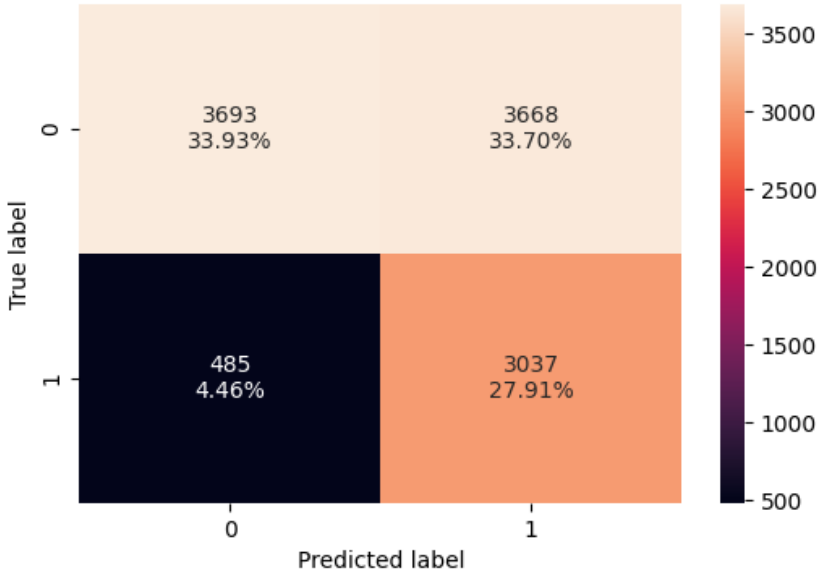

Figure 47: Confusion Matrix for Test data (Pre-pruning)

	Accuracy	Recall	Precision	F1
0	0.618396	0.862294	0.452946	0.593918

Table 22: Model Performance for Test data (Pre-pruning)

- The model is giving a generalized result now since the recall scores on both the train and test data are coming to be around 0.86 which shows that the model is able to generalize well on unseen data.

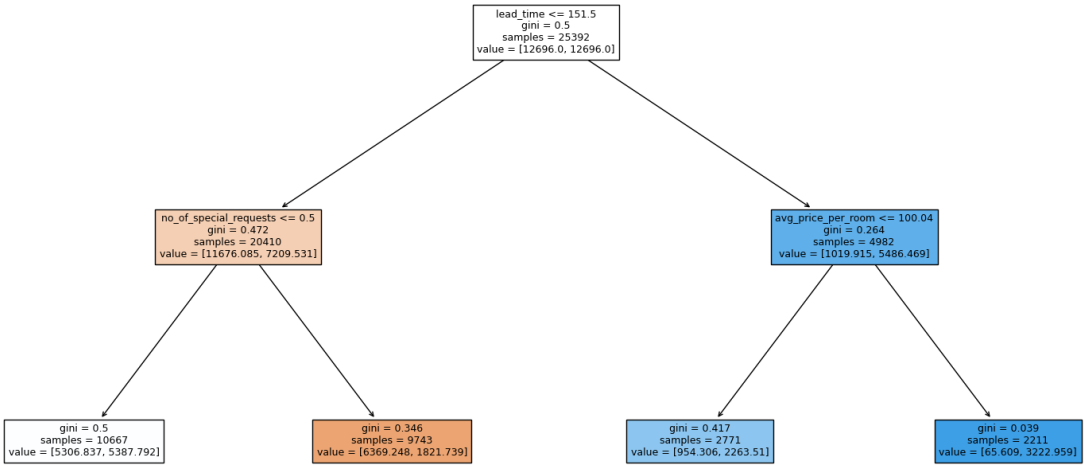

Figure 48: Pre-pruned Decision Tree

Observations from the pre-pruned tree:

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

- If the lead time is less than or equal to 151.5, the number of special request is less than or equal to 0.50, and avg price per room is less than or equal to 100.04 then the booking is likely to get canceled.
- Interpretations from other decision rules can be made similarly.

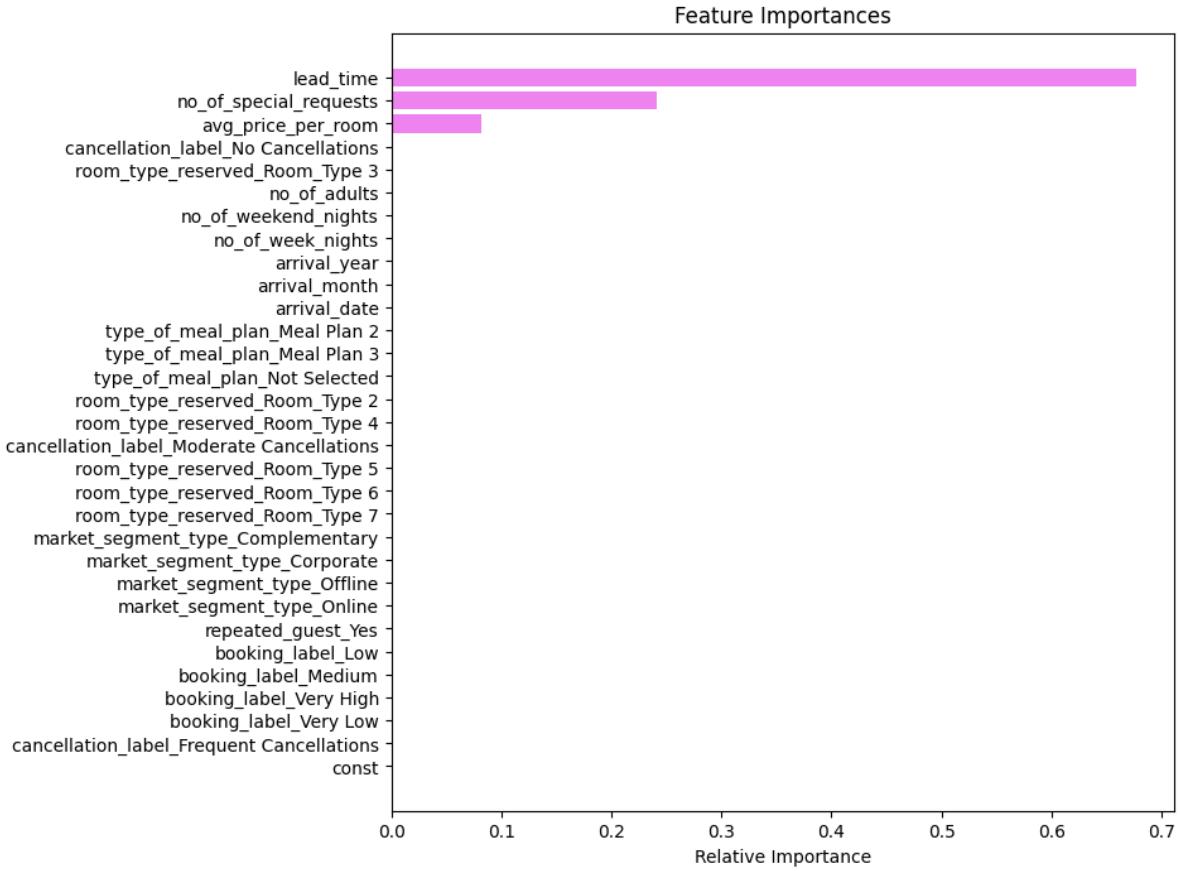

Figure 49: Importance of features in the tree building (Pre-Pruning)

- In the pre tuned decision tree also, lead time special request and room average price are the most important features.

8.5.3. Decision Tree (Post-pruning)

- Next, we train a decision tree using the effective alphas. The last value in ccp_alphas is the alpha value that prunes the whole tree, leaving the tree, clfs[-1], with one node.
- Number of nodes in the last tree is: 1 with ccp_alpha: 0.08117914389137032

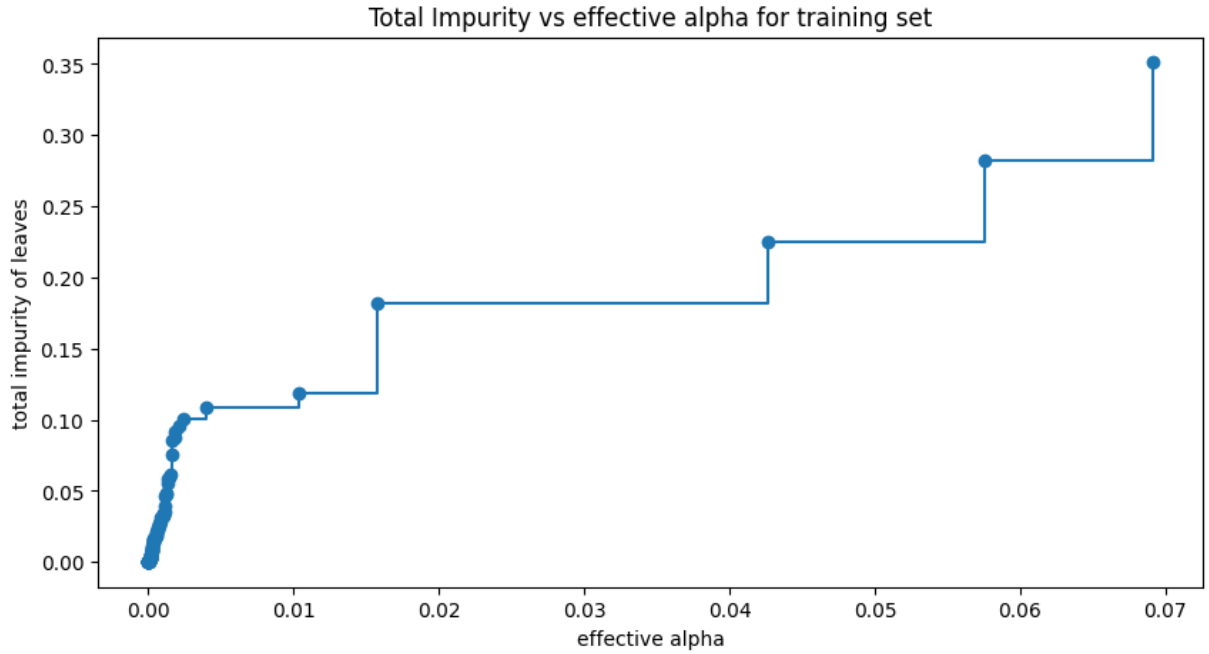

Figure 50: Impurity vs Alpha (Training Set)

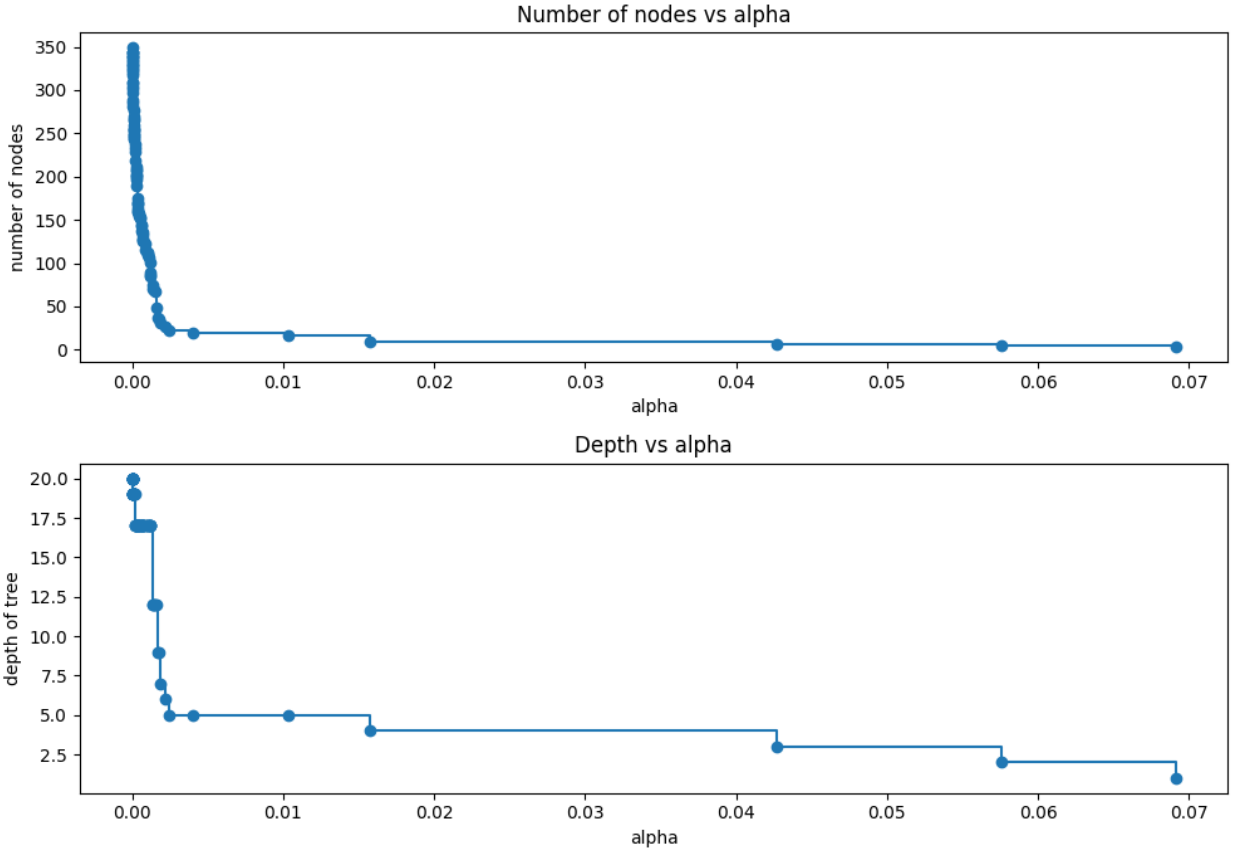

Figure 51: Nodes and Depth vs Alpha

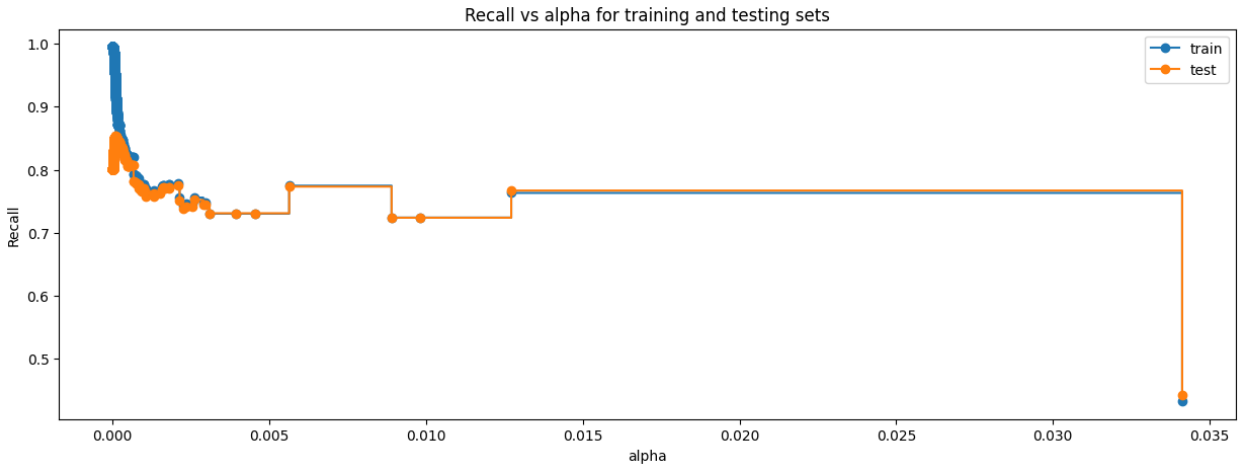

Figure 52: Recall vs alpha for training and testing sets

- Creating the model where we get highest train and test recall (ccp_alpha= 0.00009648231815018171)

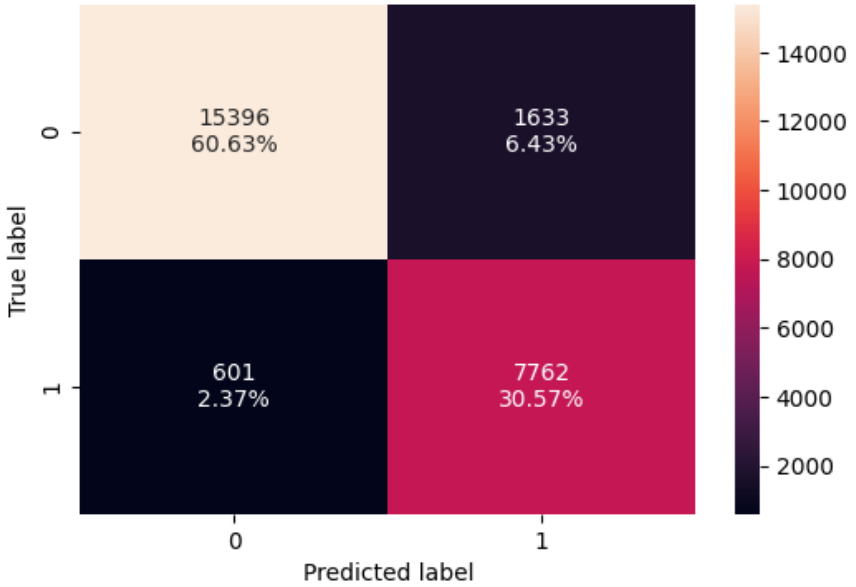

Figure 53: Confusion Matrix for Train Data (best model)

	Accuracy	Recall	Precision	F1
0	0.91202	0.928136	0.826184	0.87419

Table 23: Model Performance for Train Data (best model)

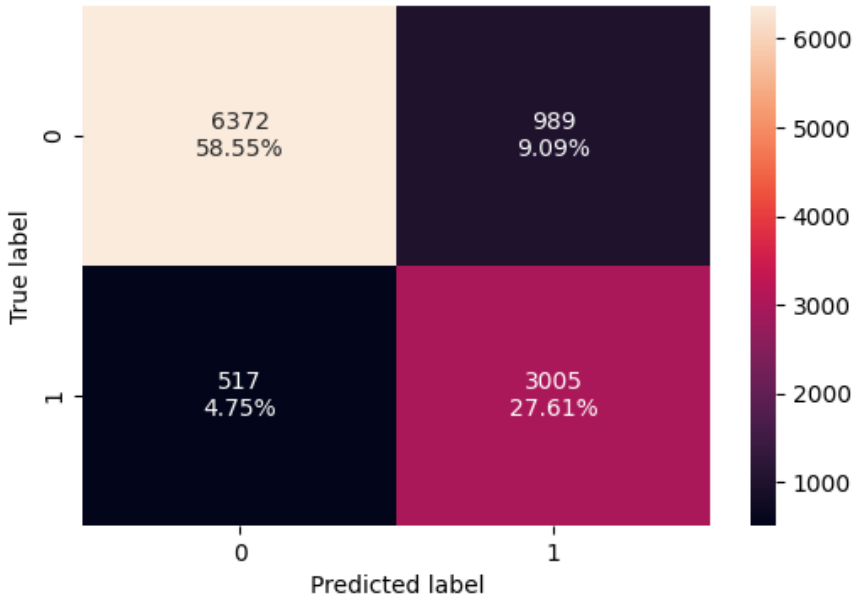

Figure 54: Confusion Matrix for Test Data (best model)

	Accuracy	Recall	Precision	F1
0	0.861619	0.853208	0.752379	0.799627

Table 24: Model Performance for Test Data (best model)

- Training data has recall of 0.92 and testing data has recall of 0.85. Precision and recall have significantly increased from pre-pruning method. This shows this model is explaining sufficiently good. As our alpha value is very low, we are getting more branches.

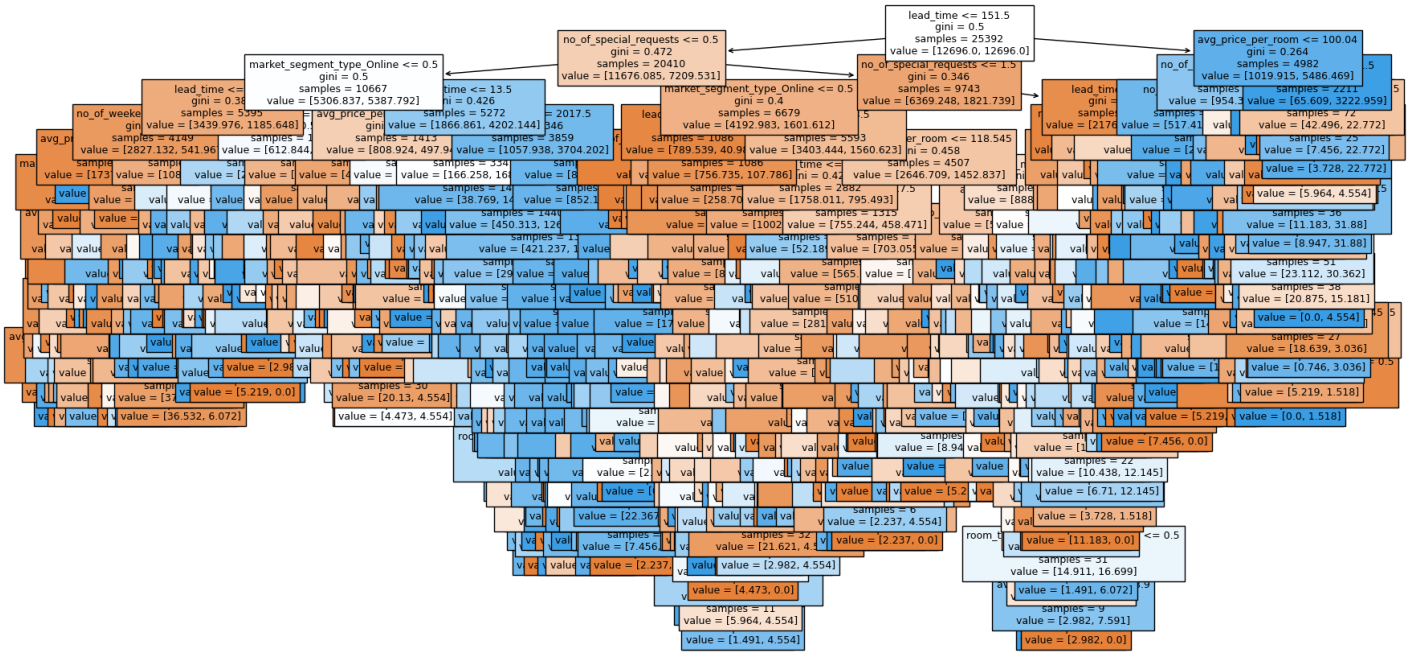

Table 25: Decision tree post-pruning

Observations from the post-pruned tree:

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

- If the lead time is less than or equal to 151.5, the number of special request is less than or equal to 0.50, and market segment type online is less than or equal to 0.50 then the booking is likely to get canceled.

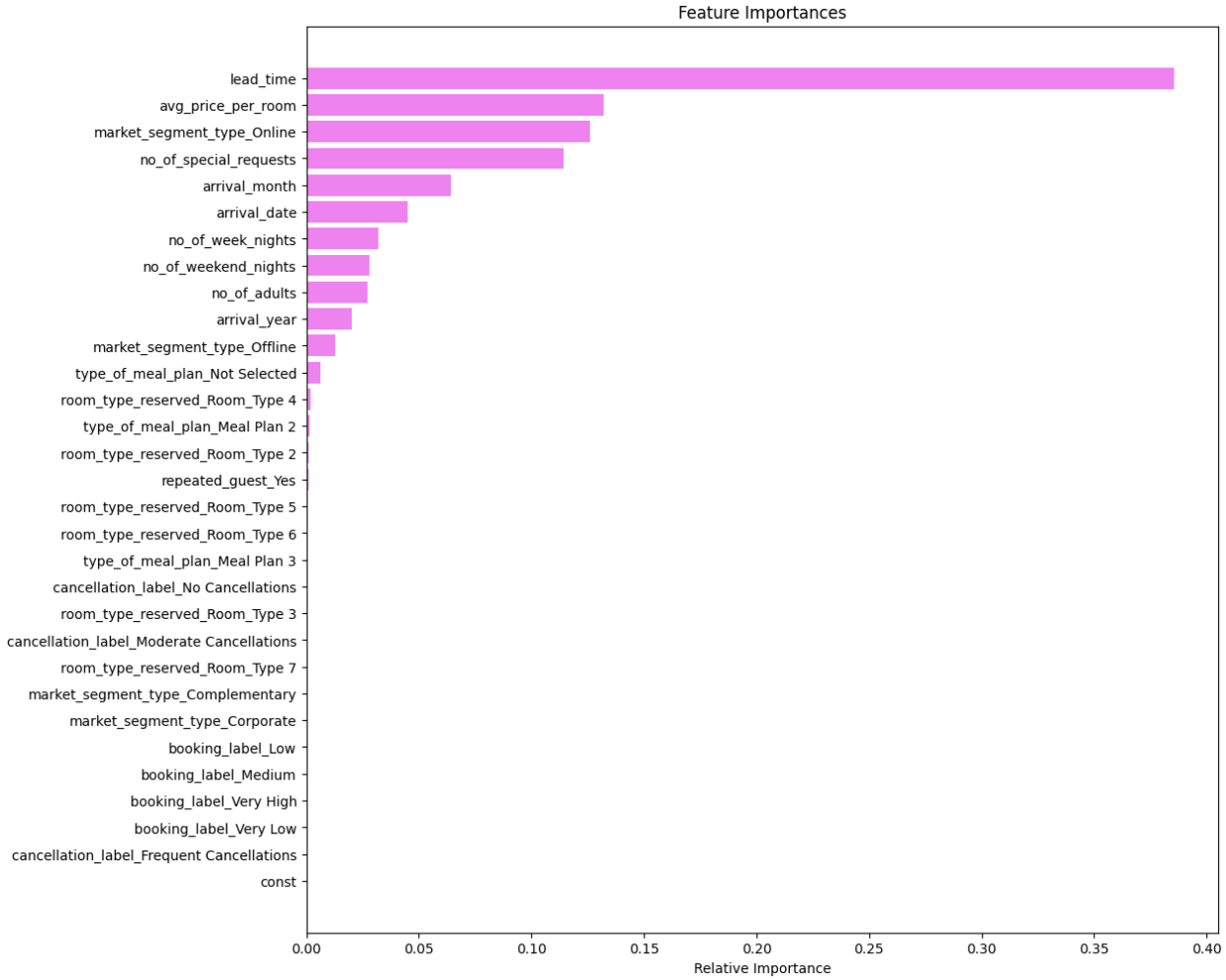

Figure 55: Important features post-pruning

- Lead time, avg price per room, online market segment type, and special requests are the main features of the post-pruned tree.

8.5.4. Model Comparison

Training performance comparison:

	Decision Tree without class_weight	Decision Tree with class_weight	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.994171	0.993069	0.618541	0.912020
Recall	0.986488	0.995097	0.856511	0.928136
Precision	0.995775	0.984037	0.457729	0.826184
F1	0.991110	0.989536	0.596618	0.874198

Table 26: Decision Tree All models for training data set

Test set performance comparison:

	Decision Tree without class_weight	Decision Tree with class_weight	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.867408	0.862170	0.618396	0.861619
Recall	0.806644	0.800681	0.862294	0.853208
Precision	0.788510	0.779436	0.452946	0.752379
F1	0.797474	0.789916	0.593918	0.799627

Table 27: Decision Tree All models for testing data set

- The most effective decision tree was achieved using the post-pruning method. It demonstrated significant recall and precision, with differences between the train and test data within a 10% margin.

9. All Models Comparison

	Logistic Regression-default Threshold (0.5)	Logistic Regression-0.81 Threshold	Logistic Regression-0.42 Threshold	K Nearest Neighbor k=3	Naive Bayes	Decision Tree without class_weight	Decision Tree with class_weight	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Training set performance comparison									
Accuracy	0.798913	0.768746	0.79781	0.912295	0.370117	0.994171	0.993069	0.618541	0.91202
Recall	0.617601	0.798876	0.69114	0.842401	0.991032	0.986488	0.995097	0.856511	0.928136
Precision	0.730242	0.614571	0.693794	0.885718	0.34238	0.995775	0.984037	0.457729	0.826184
F1	0.669215	0.694707	0.692464	0.863517	0.508935	0.99111	0.989536	0.596618	0.874198
Testing set performance comparison									
Accuracy	0.801893	0.774694	0.799963	0.836993	0.3658	0.867408	0.86217	0.618396	0.861619
Recall	0.622658	0.803805	0.69222	0.720897	0.993186	0.806644	0.800681	0.862294	0.853208
Precision	0.726159	0.616507	0.690456	0.762462	0.337124	0.78851	0.779436	0.452946	0.752379
F1	0.670437	0.697806	0.691337	0.741097	0.503382	0.797474	0.789916	0.593918	0.799627

Table 28: Performance comparison between all the models used

- The best performance was achieved using the decision tree post-pruning method, resulting in the highest F1 score for both training and testing data, along with the highest precision and recall scores.

10. Actionable Insights & Recommendations

1. Lead Time Management

- **Encourage Shorter Lead Times:** Since it is evident that a longer lead time might enhance the chances of cancellations, consider offering a discount or incentive for those making bookings closer to the check-in date, which includes last-minute deals and promotions targeting spontaneous travelers.
- **Flexible Booking Policies:** Offer flexible cancellation policies for long lead time bookings to minimize cancellations. For example, offer free cancellations up until a particular point closer to the check-in date.

2. Optimize Your Booking Channels

- **Promotion of Offline Channels:** Since offline bookings generally have a lower cancellation rate, you can encourage more bookings through this channel. You accomplish this by offering special offline discounts or by tying up with some local travel agencies.
- **Improve Online Booking Experience**: Improve the online booking experience to reduce cancellations. This can be done by providing more customer support, clearly stating the cancelation policies, and providing some additional offers on holding bookings.

3. Repeat Guest Retention

- **Loyalty Programs:** Refine or develop loyalty programs that would increase repeat bookings with benefits like discounts, room upgrades, or additional services for loyal customers—lowering their chances of canceling.
- **Personalized Marketing:** Based on the data collected regarding prior bookings, more emphasis should be given to repeat guests by designing personalized marketing campaigns and emphasizing the special benefits given to loyal customers.

4. Special Requests Handling

- **Encourage Special Requests:** Since bookings with special requests are much less often canceled, it is important to motivate guests to make special requests. Make this an option at time of booking and accommodation of such requests very good.
- **Better Customer Service:** Train staff to effectively handle special requests and follow up with guests to ensure all their requirements are being met, which would help in reaffirming the guest's commitment to the booking.

5. Pricing Strategies

- **Competitive Pricing**: Monitoring and changing the cost of the rooms to be competitive, especially for online bookings where price is more sensitive. Then there is dynamic pricing to optimize on occupancy and minimize cancellations.
- **Price Guarantees:** Provide price guarantees to reduce cancellations triggered by price change. Give confidence that the guests are receiving the optimum price.

6. Seasonal Promotion

- **Low Season Bookings:** Try to have more promotion during January and February months, which contribute the most to the lowest number of arrivals. This could include special winter packages or themed events.
- **Manage High-Season Demand:** Especially in peak months, such as October, make sure policies are in place to minimize cancellations through more stringent cancellation policies or nonrefundable rates.

7. Data-Driven Decision Making

- **Regular Analysis:** Keep a constant view of the booking data, looking for trends, and adjust strategies accordingly. Run prediction models that look out for cancellations so proactive measures may be taken against them.

- **Feedback Loop:** Design a system to take feedback from people who are not going to arrive, especially from canceling guests, in order to understand the reasons behind it and make any necessary changes to the booking process for future guests.

8. Enhance Communication

- **Pre-Arrival Engagement:** Engage more before the guests' arrival to confirm their stay and raise/redress concerns that may lead to cancellations. These may involve emails, texts, or even calls.
- **Transparency of Policies**: The cancellation policies ought to be clearly stated at the point of reservation so as to have as minimal a level of misunderstanding as possible, which could in turn reduce last-minute cancellations.

If these suggestions are followed by the INN hotels, then the cancellations would significantly drop and eventually work on more efficient booking procedures while satisfying guests.