Hotel Bookings Classification Report

Presented by :
Sanjay Rajan J

TABLE OF CONTENTS

LIST OF FIGURES

➢ DATA OVERVIEW

CONTEXT

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 impact 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.

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.

DATA DICTIONARY:

- 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.

➢ Shape:

There are 1000 rows and 8 columns in this dataset.

➢ Duplicates:

There are no duplicate entries in this dataset.

- Basic Info:

Fig.1 Data Info

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36275 entries, 0 to 36274
Data columns (total 19 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Booking_ID                            36275 non-null  object
 1   no_of_adults                          36275 non-null  int64
 2   no_of_children                        36275 non-null  int64
 3   no_of_weekend_nights                  36275 non-null  int64
 4   no_of_week_nights                     36275 non-null  int64
 5   type_of_meal_plan                     36275 non-null  object
 6   required_car_parking_space            36275 non-null  int64
 7   room_type_reserved                    36275 non-null  object
 8   lead_time                             36275 non-null  int64
 9   arrival_year                          36275 non-null  int64
 10  arrival_month                         36275 non-null  int64
 11  arrival_date                          36275 non-null  int64
 12  market_segment_type                   36275 non-null  object
 13  repeated_guest                        36275 non-null  int64
 14  no_of_previous_cancellations          36275 non-null  int64
 15  no_of_previous_bookings_not_canceled  36275 non-null  int64
 16  avg_price_per_room                    36275 non-null  float64
 17  no_of_special_requests                36275 non-null  int64
 18  booking_status                        36275 non-null  object
dtypes: float64(1), int64(13), object(5)
memory usage: 5.3+ MB
```

- Null values Check:

Fig.2 Null values check

```
Booking_ID                            0
no_of_adults                          0
no_of_children                        0
no_of_weekend_nights                  0
no_of_week_nights                     0
type_of_meal_plan                     0
required_car_parking_space            0
room_type_reserved                    0
lead_time                             0
arrival_year                          0
arrival_month                         0
arrival_date                          0
market_segment_type                   0
repeated_guest                        0
no_of_previous_cancellations          0
no_of_previous_bookings_not_canceled  0
avg_price_per_room                    0
no_of_special_requests                0
booking_status                        0
dtype: int64
```

- Numerical Statistics:

Fig.3 Numerical Statistics

	count	mean	std	min	25%	50%	75%	max
no_of_adults	36275.0	1.844962	0.518715	0.0	2.0	2.00	2.0	4.0
no_of_children	36275.0	0.105279	0.402648	0.0	0.0	0.00	0.0	10.0
no_of_weekend_nights	36275.0	0.810724	0.870644	0.0	0.0	1.00	2.0	7.0
no_of_week_nights	36275.0	2.204300	1.410905	0.0	1.0	2.00	3.0	17.0
required_car_parking_space	36275.0	0.030986	0.173281	0.0	0.0	0.00	0.0	1.0
lead_time	36275.0	85.232557	85.930817	0.0	17.0	57.00	126.0	443.0
arrival_year	36275.0	2017.820427	0.383836	2017.0	2018.0	2018.00	2018.0	2018.0
arrival_month	36275.0	7.423653	3.069894	1.0	5.0	8.00	10.0	12.0
arrival_date	36275.0	15.596995	8.740447	1.0	8.0	16.00	23.0	31.0
repeated_guest	36275.0	0.025637	0.158053	0.0	0.0	0.00	0.0	1.0
no_of_previous_cancellations	36275.0	0.023349	0.368331	0.0	0.0	0.00	0.0	13.0
no_of_previous_bookings_not_canceled	36275.0	0.153411	1.754171	0.0	0.0	0.00	0.0	58.0
avg_price_per_room	36275.0	103.423539	35.089424	0.0	80.3	99.45	120.0	540.0
no_of_special_requests	36275.0	0.619655	0.786236	0.0	0.0	0.00	1.0	5.0

- No. of Unique values:

Fig.4 Unique Values

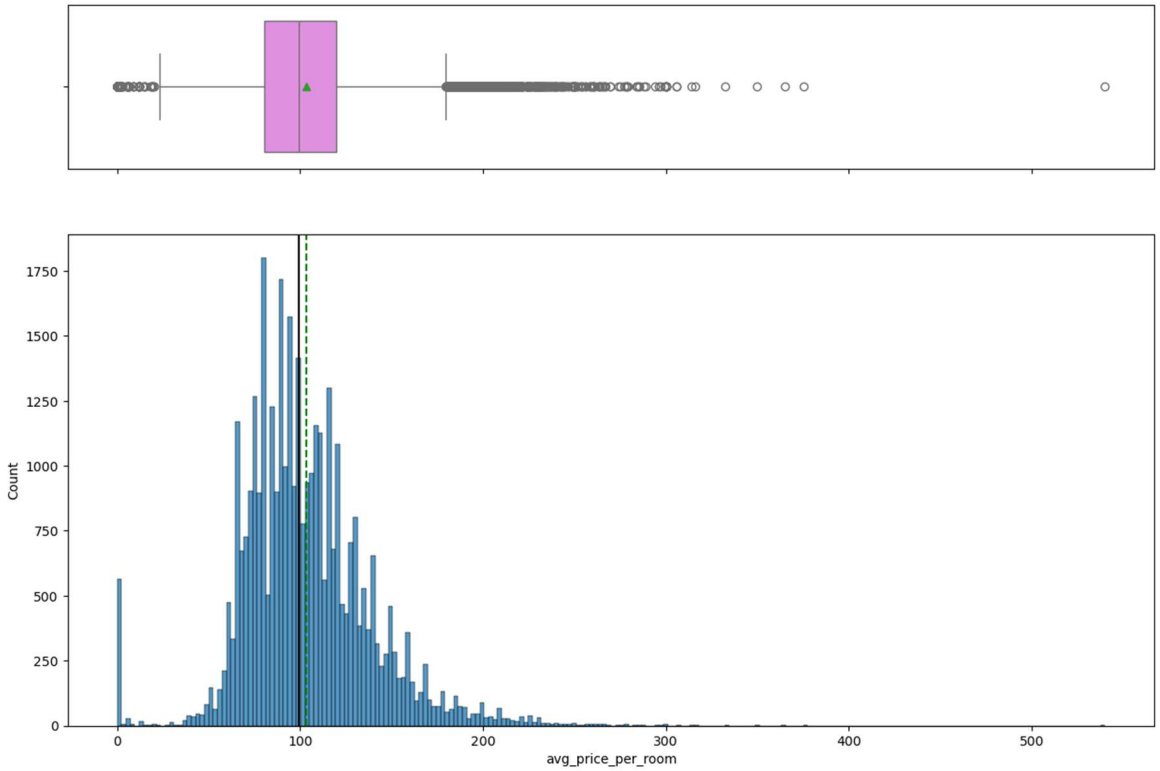
```
Booking_ID : 36275
no_of_adults : 5
no_of_children : 6
no_of_weekend_nights : 8
no_of_week_nights : 18
type_of_meal_plan : 4
required_car_parking_space : 2
room_type_reserved : 7
lead_time : 352
arrival_year : 2
arrival_month : 12
arrival_date : 31
market_segment_type : 5
repeated_guest : 2
no_of_previous_cancellations : 9
no_of_previous_bookings_not_canceled : 59
avg_price_per_room : 3930
no_of_special_requests : 6
booking_status : 2
```

2. EXPLORATORY DATA ANALYSIS

2.1 UNIVARIATE ANALYSIS:

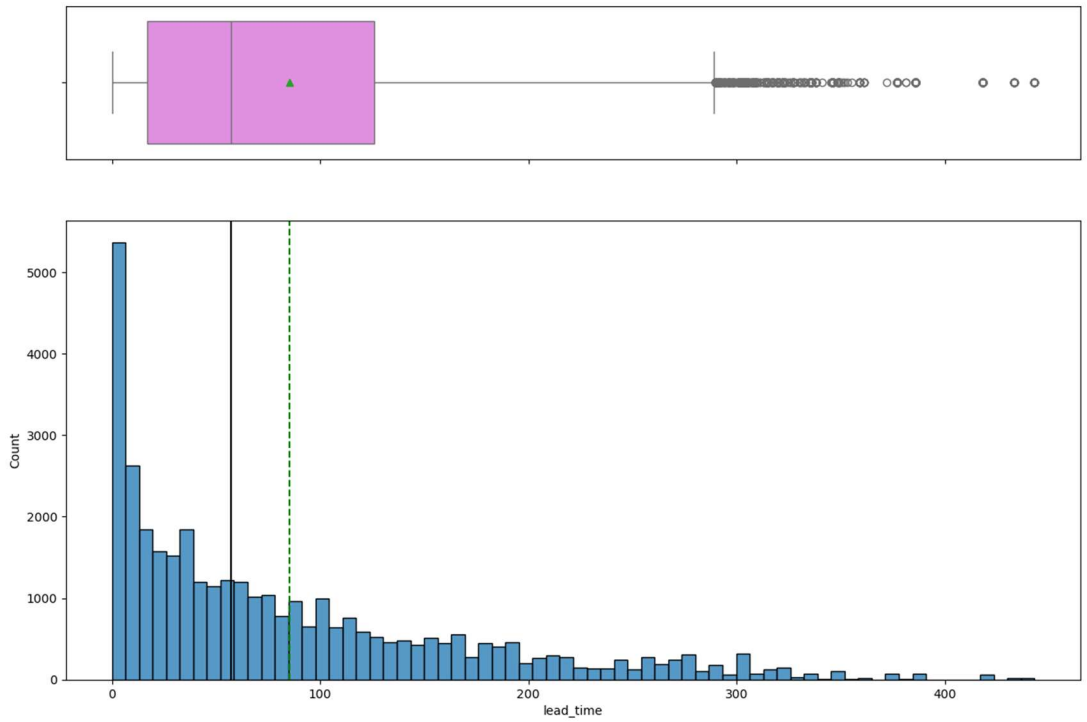
- Distribution of Average Price:

Fig.5 Average Price distribution



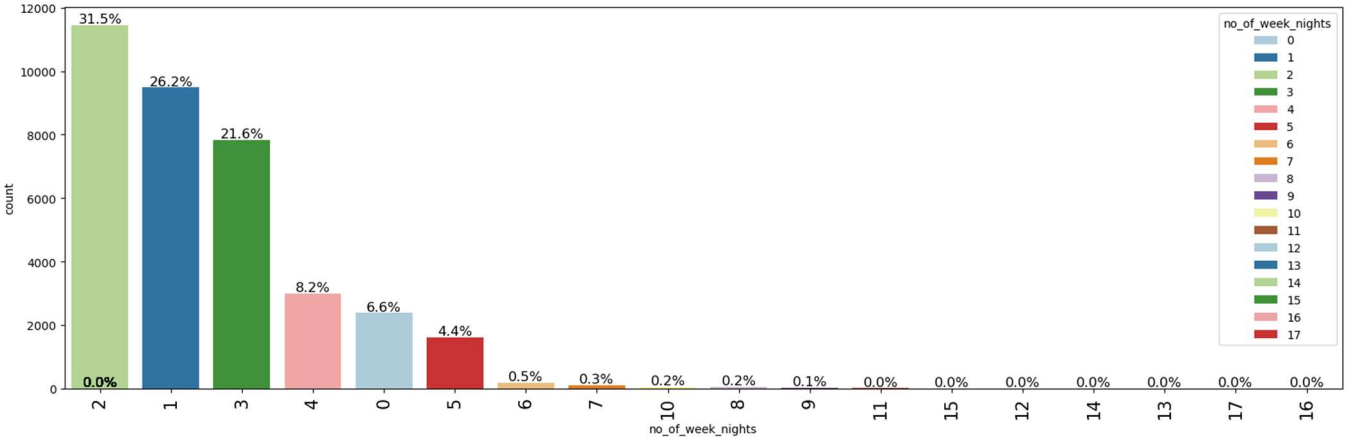
- Distribution of Lead Time:

Fig.6 Lead Time distribution



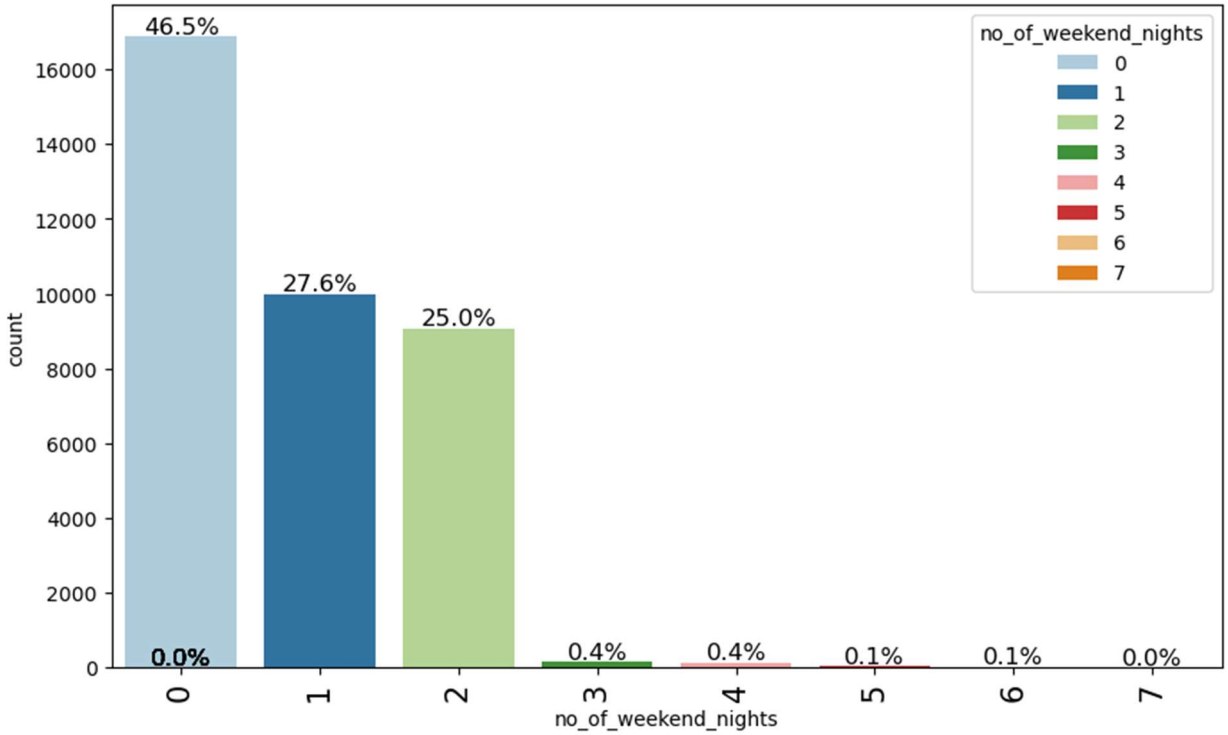
➢ Distribution of No.of week nights:

Fig.7 No.of week nights distribution



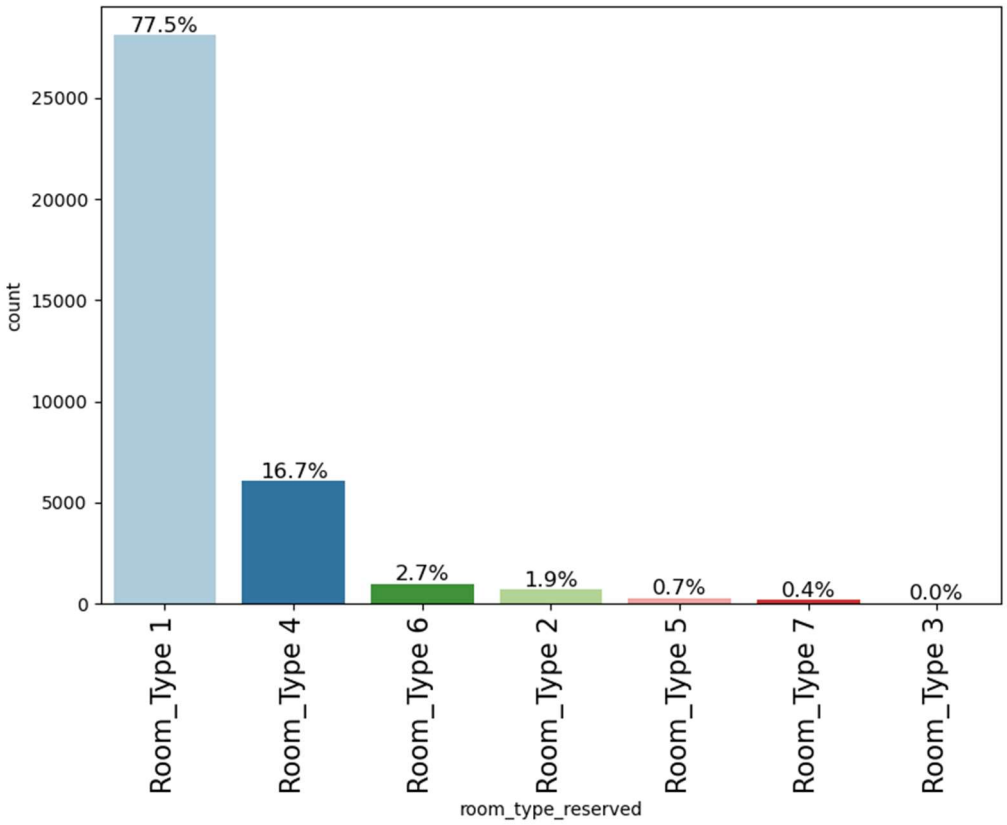
➢ Distribution of No.of weekend nights:

Fig.8 No.of weekend nights distribution



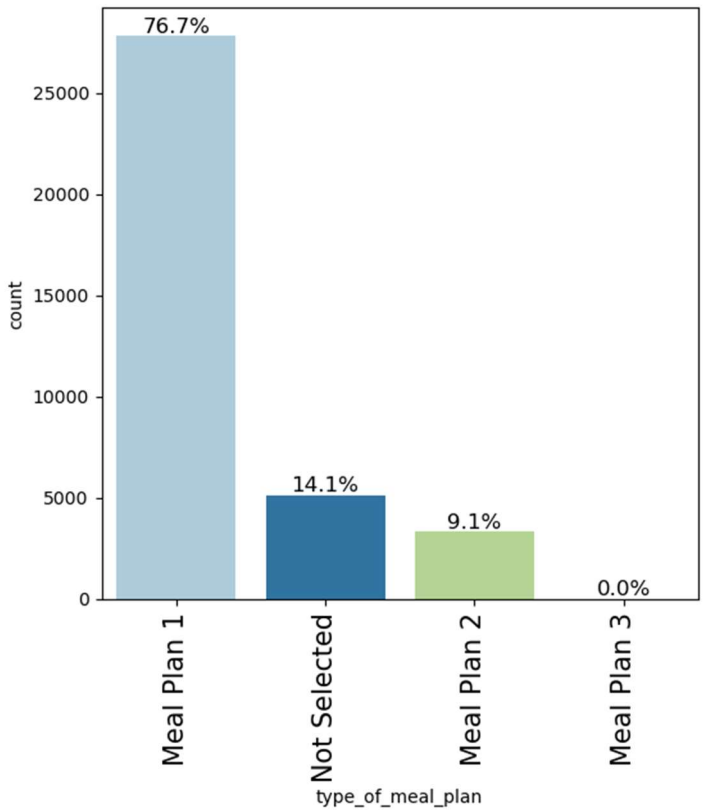
➢ Distribution of Room type reserved:

Fig.9 Room type distribution



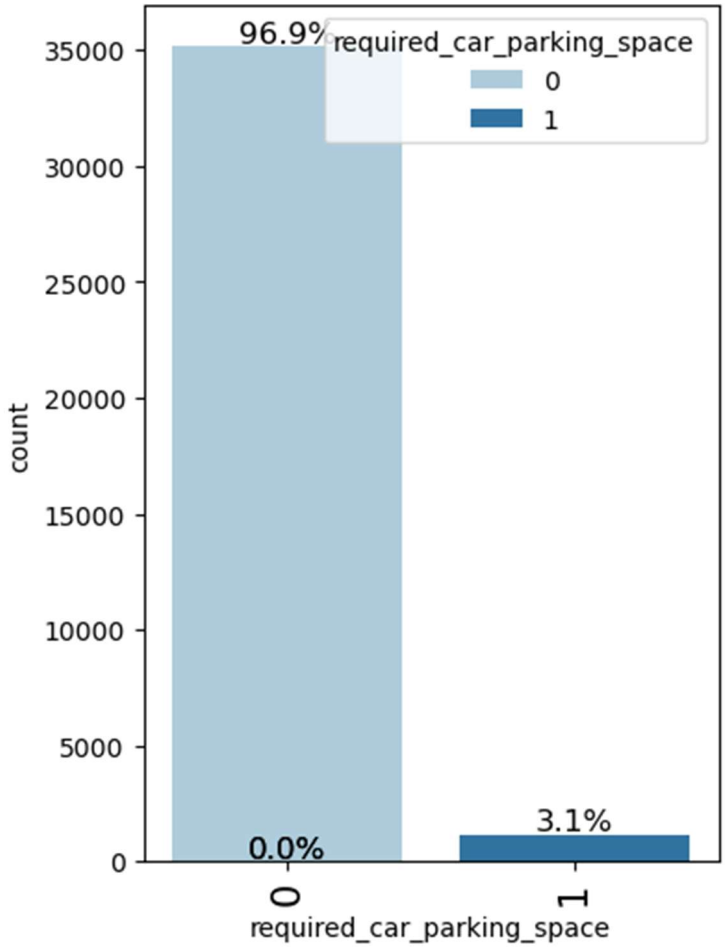
➢ Distribution of Meal plan:

Fig.10 Meal plan distribution



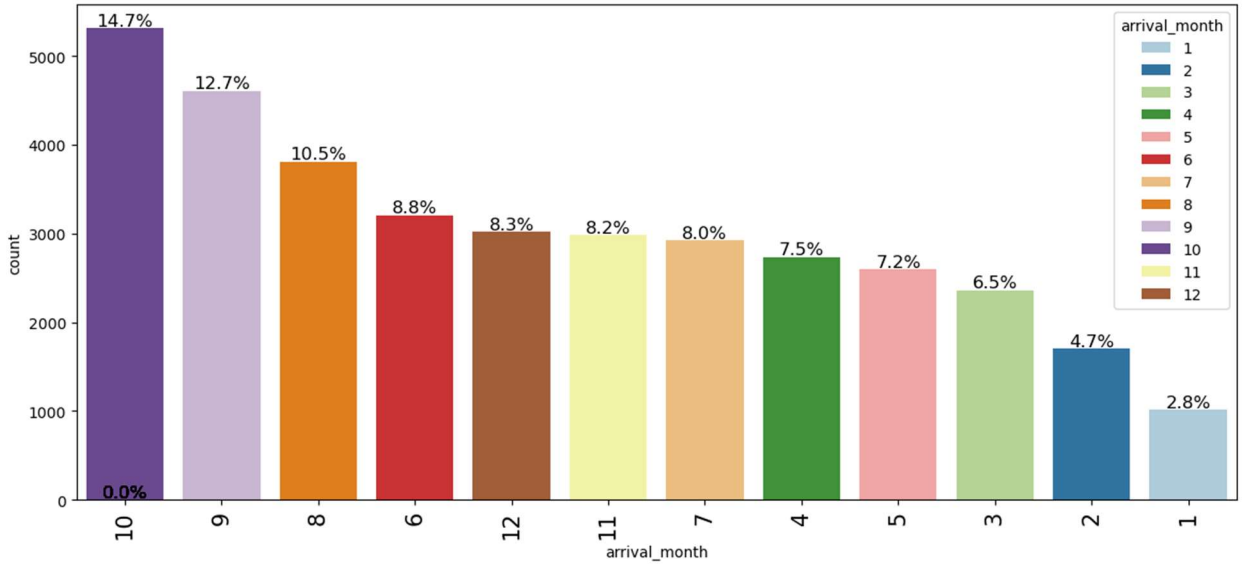
➢ Distribution of Required_car_parking:

Fig.11 Required_car_parking distribution



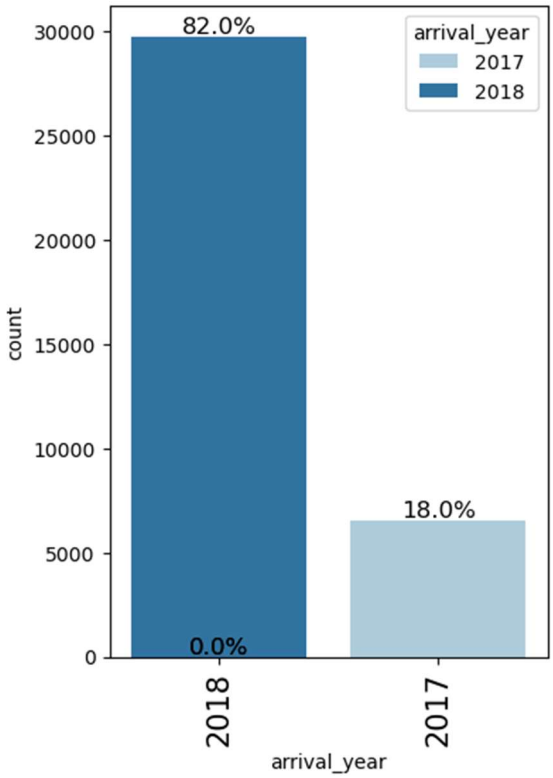
➢ Distribution of Arrival Month:

Fig.12 Arrival Month distribution



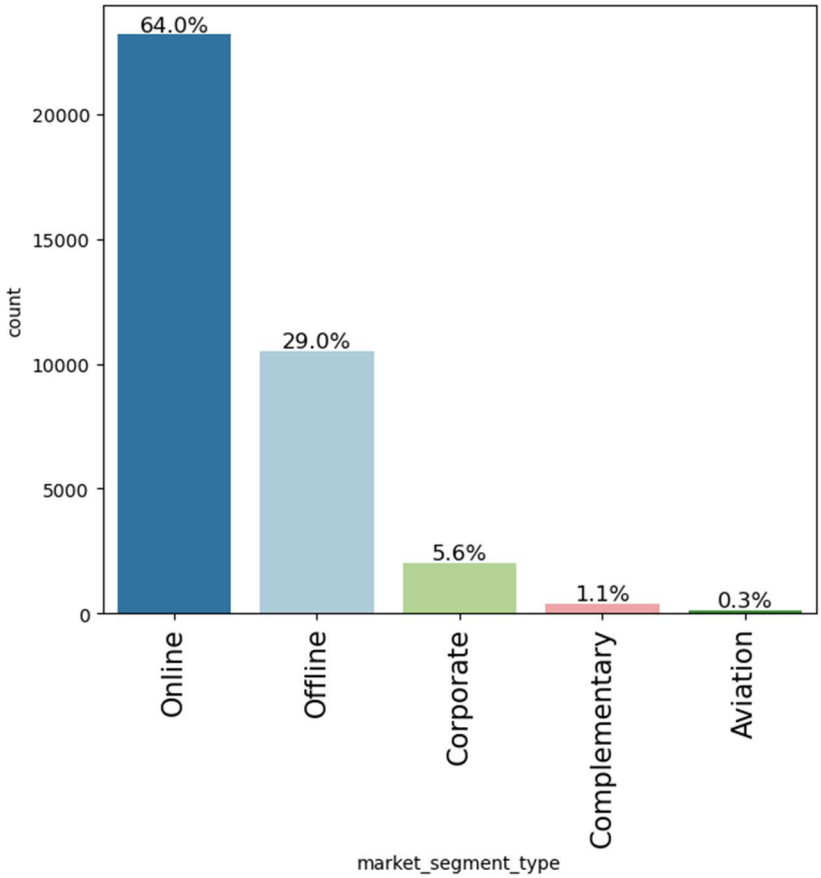
➢ Distribution of Arrival Year:

Fig.13 Arrival Year distribution



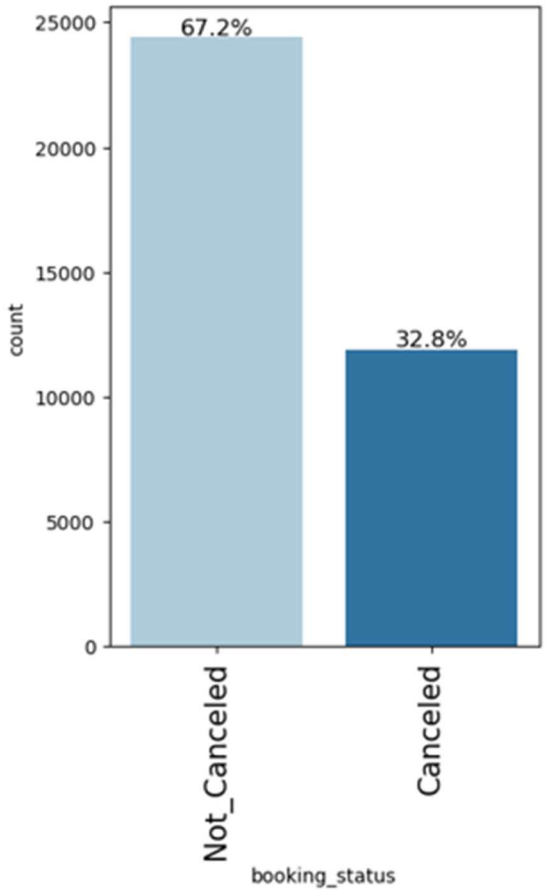
➢ Distribution of Market segment:

Fig.14 Market Segment distribution



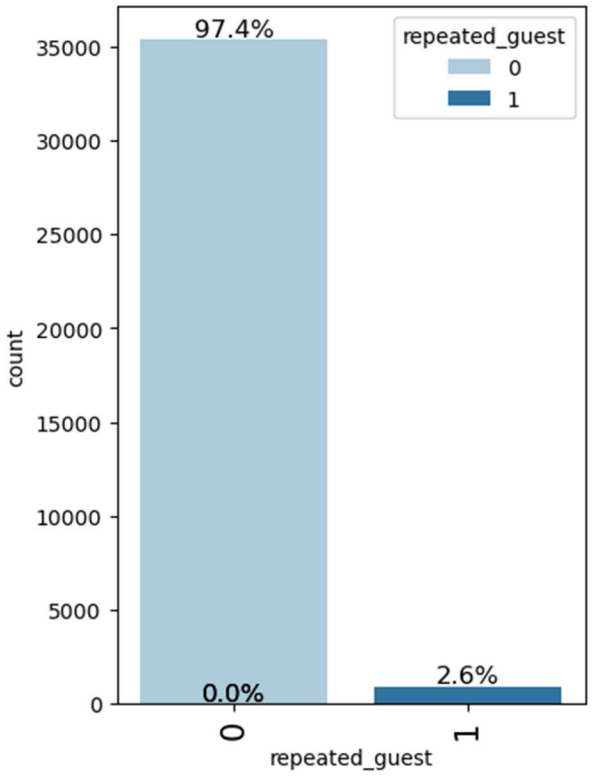
- Distribution of Booking status:

Fig.15 Booking status distribution



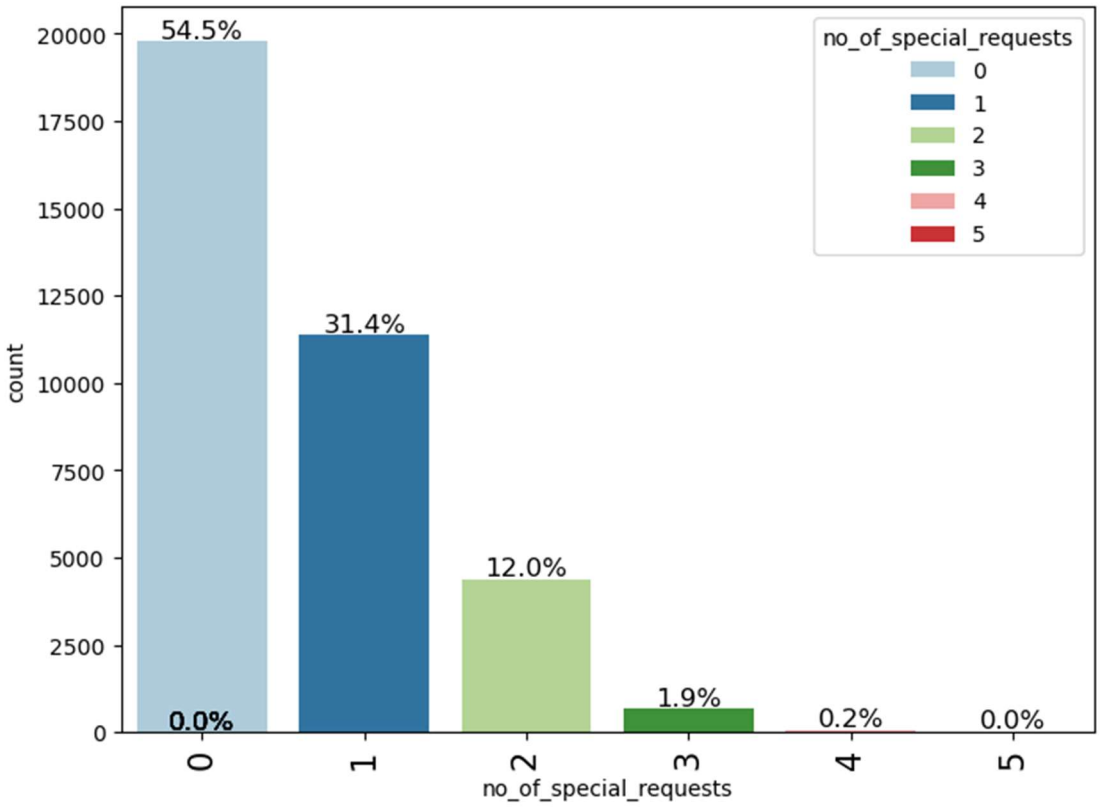
- Distribution of Repeated_guest:

Fig.16 Repeated guest distribution



- Distribution of No_of_special_requests:

Fig.17 Special Request distribution



COMMENTS:

- Lead_time and avg_price_per_room is skewed to the right.
- Majority of the people doesn't prefer special requests.
- The "Online" market segment has the highest count, with 23,214 entries.
- Repeated Guests constitutes only 2.6%.
- October was the busiest month (14.7%) followed by September and August.
- Majority of the people prefer Meal plan-1
- Majority of the people prefer Room type-1
- Over two-thirds of bookings are not cancelled, reflecting decent booking stability.

2.2 BIVARIATE ANALYSIS:

Fig.18 Heatmap

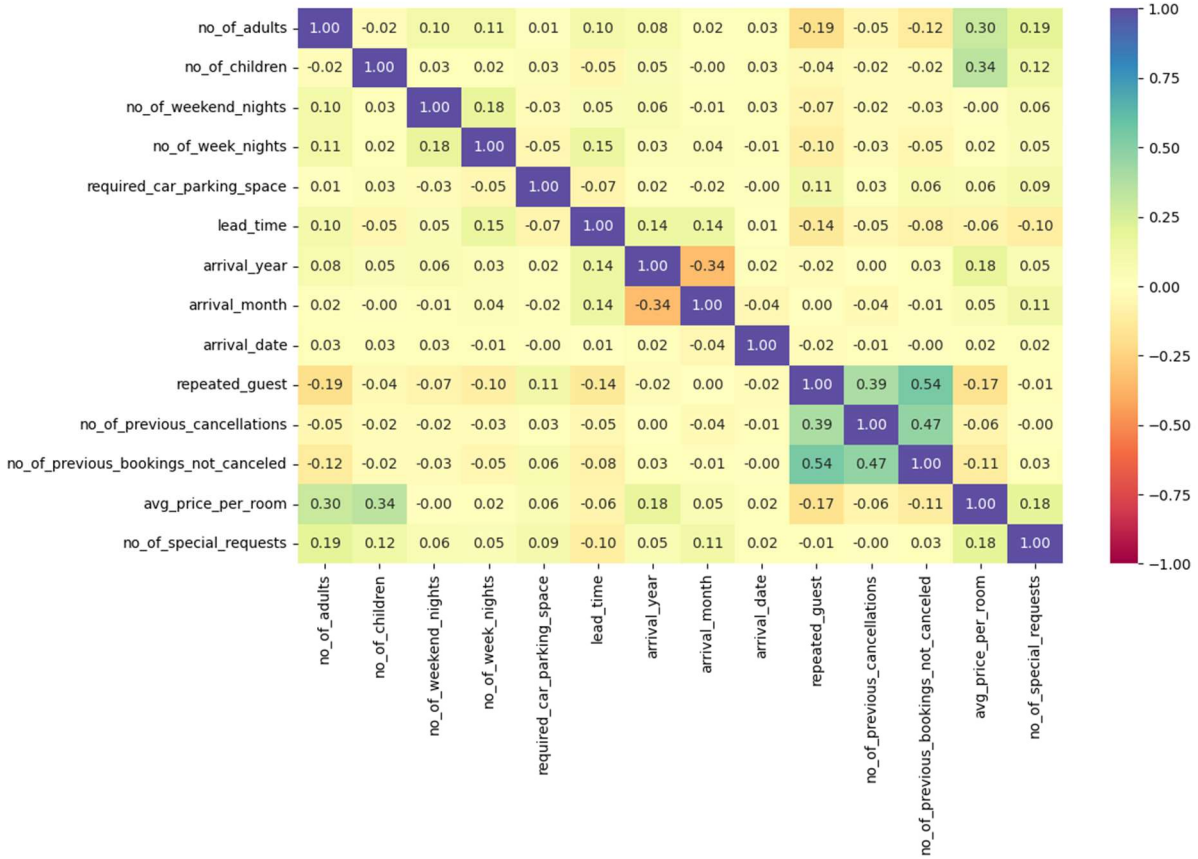


Fig.19 Avg_price vs Market segment

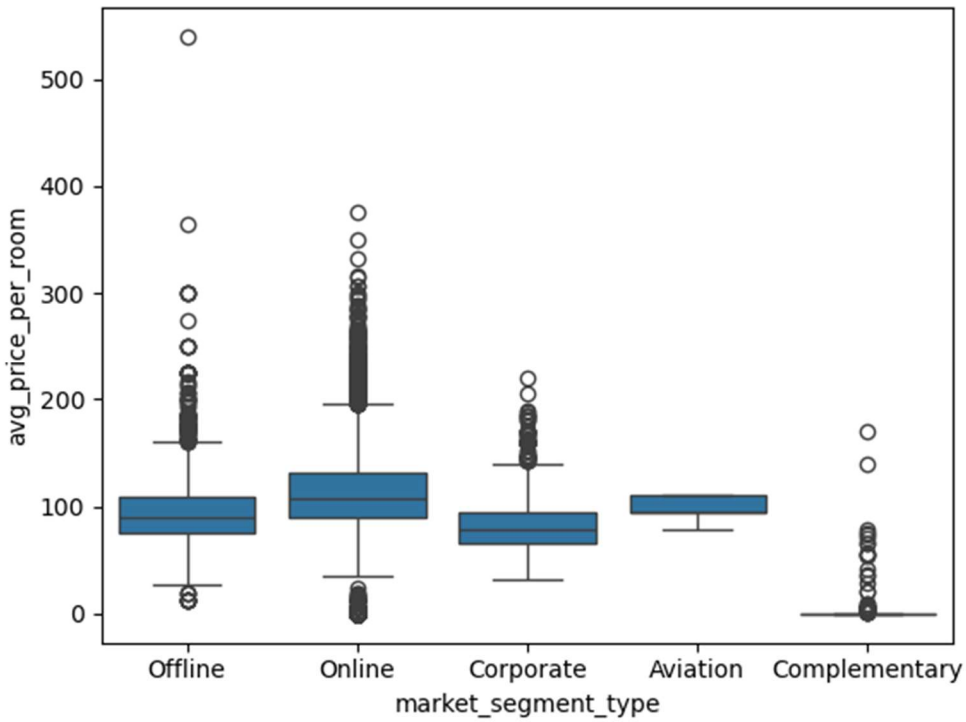


Fig.20 Avg_price vs special req

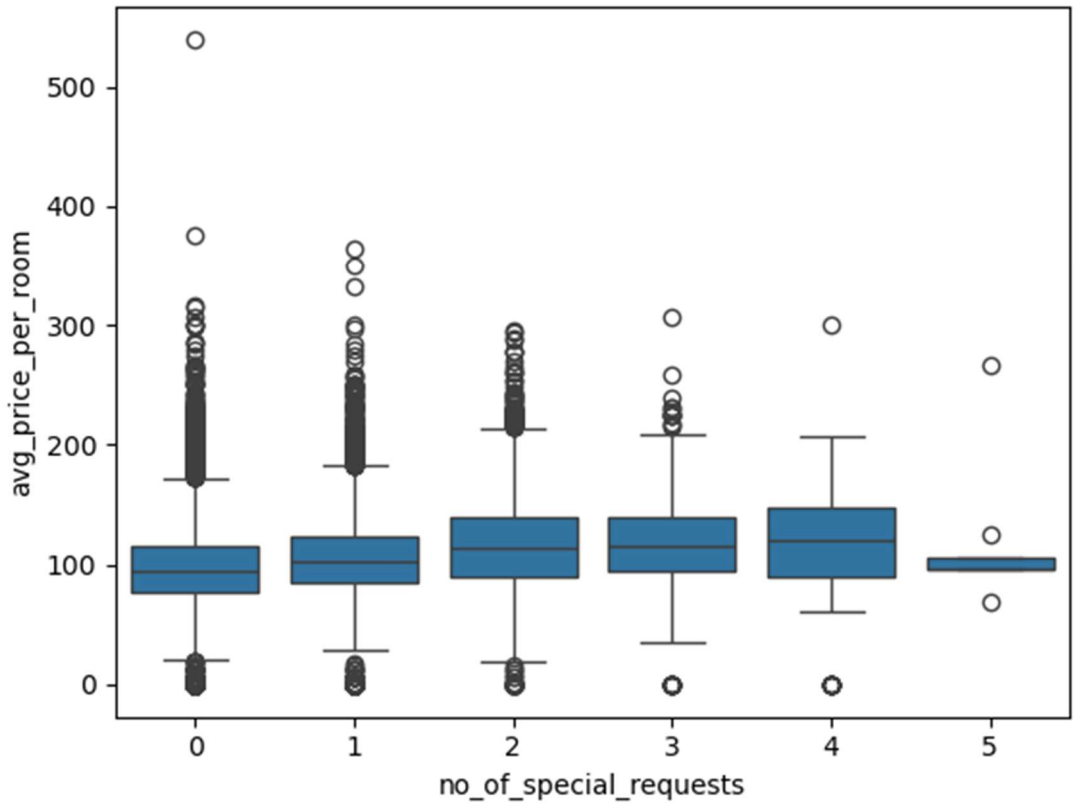


Fig.21 Booking status vs Meal plan

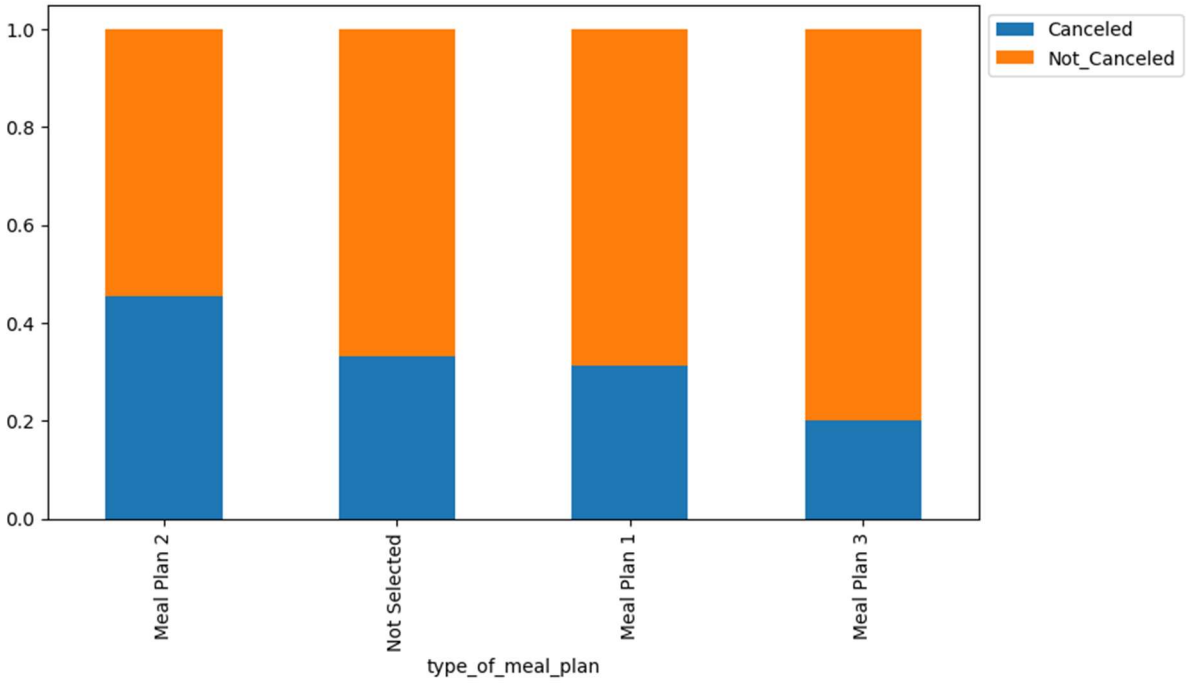


Fig.22 Booking status vs Month

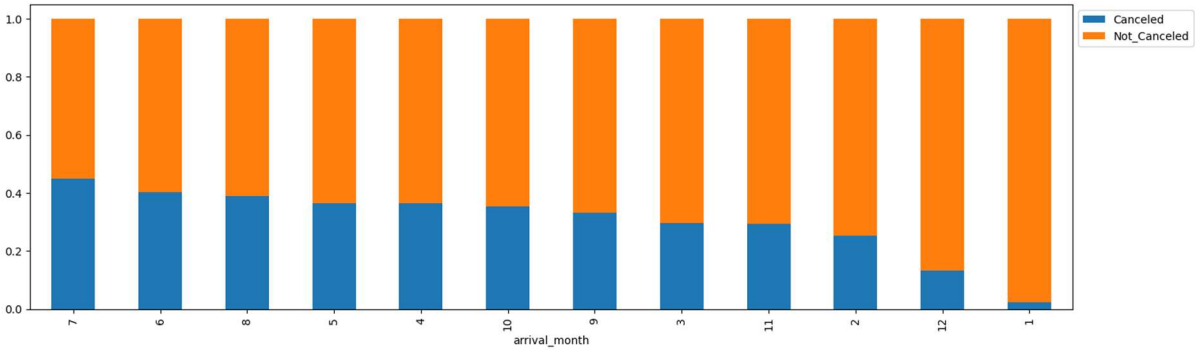


Fig.23 Booking status vs Special req

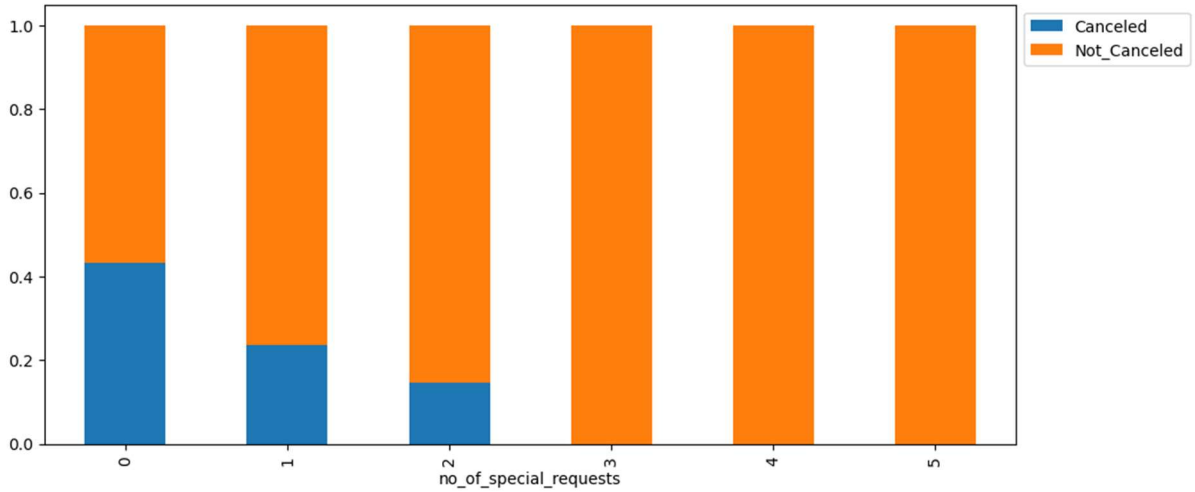


Fig.24 Booking status vs Required car parking

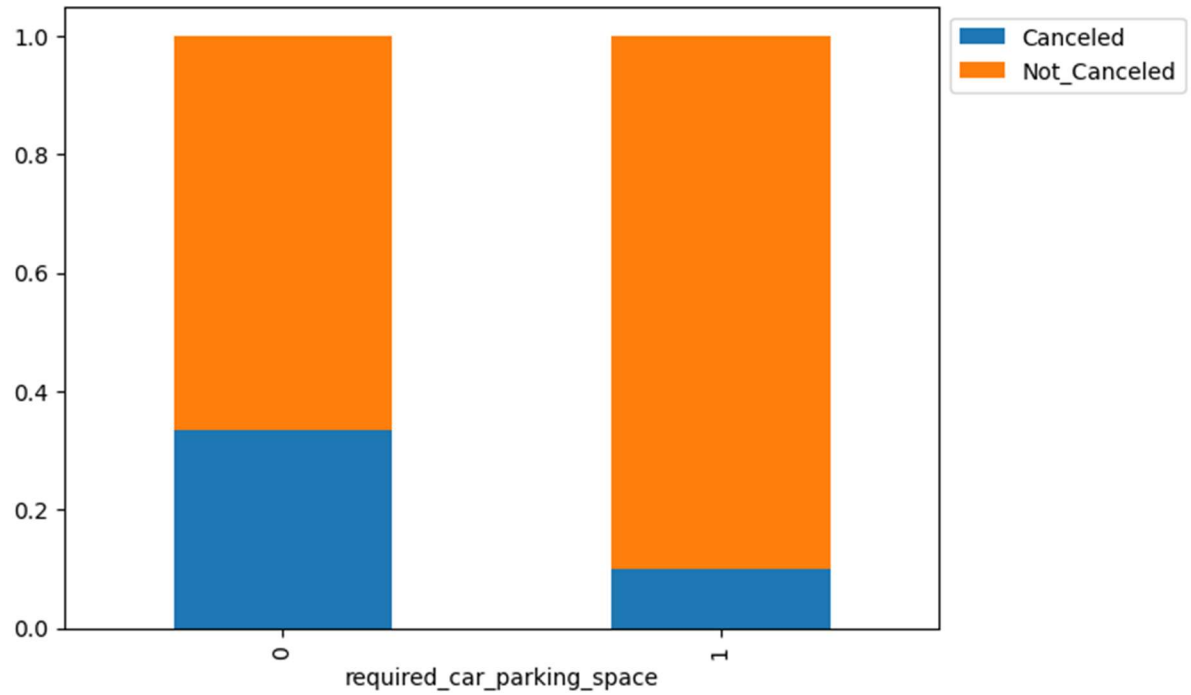


Fig.25 Booking status vs Market segment

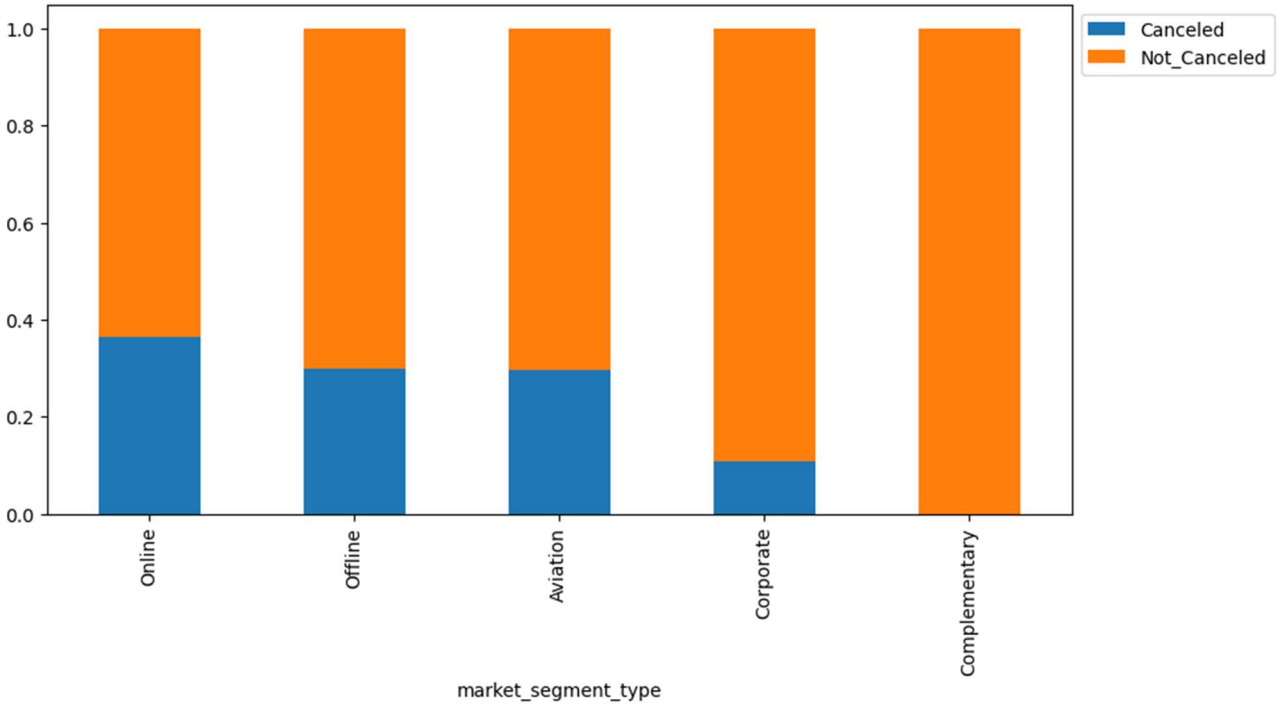


Fig.26 Booking status vs repeated guest

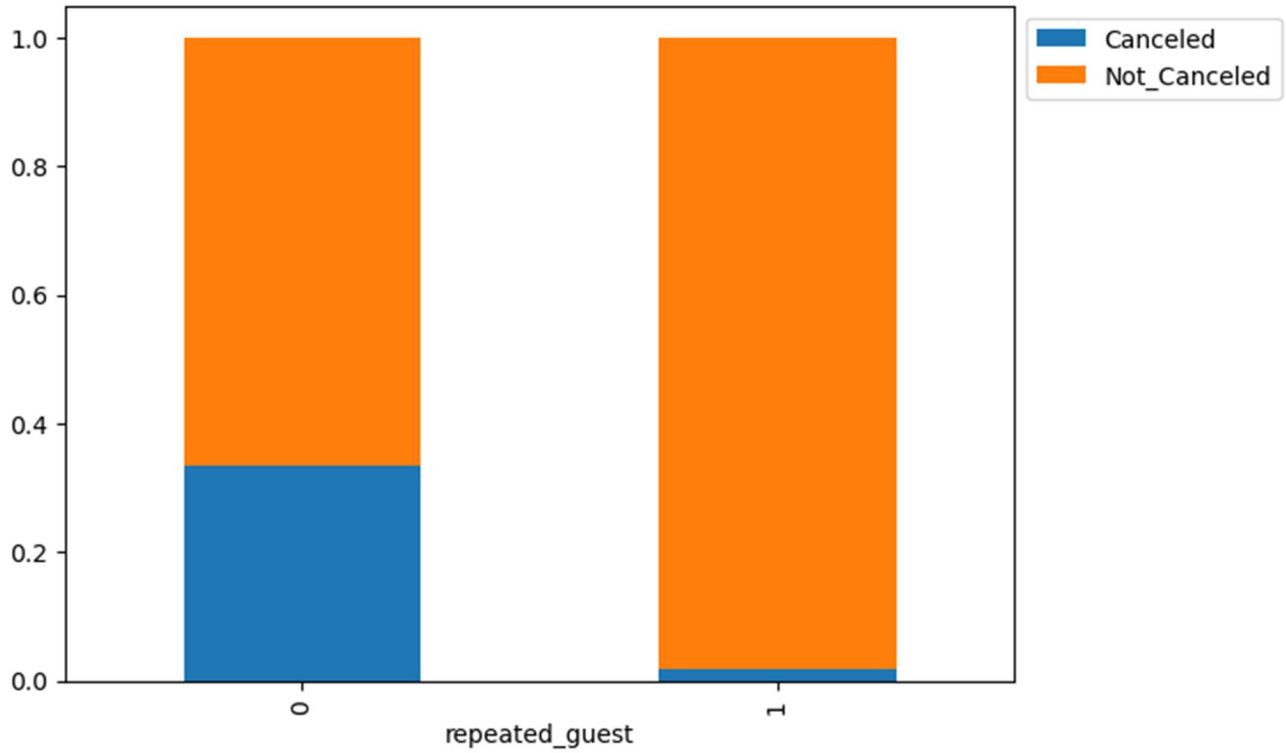


Fig.27 Distribution plot 1

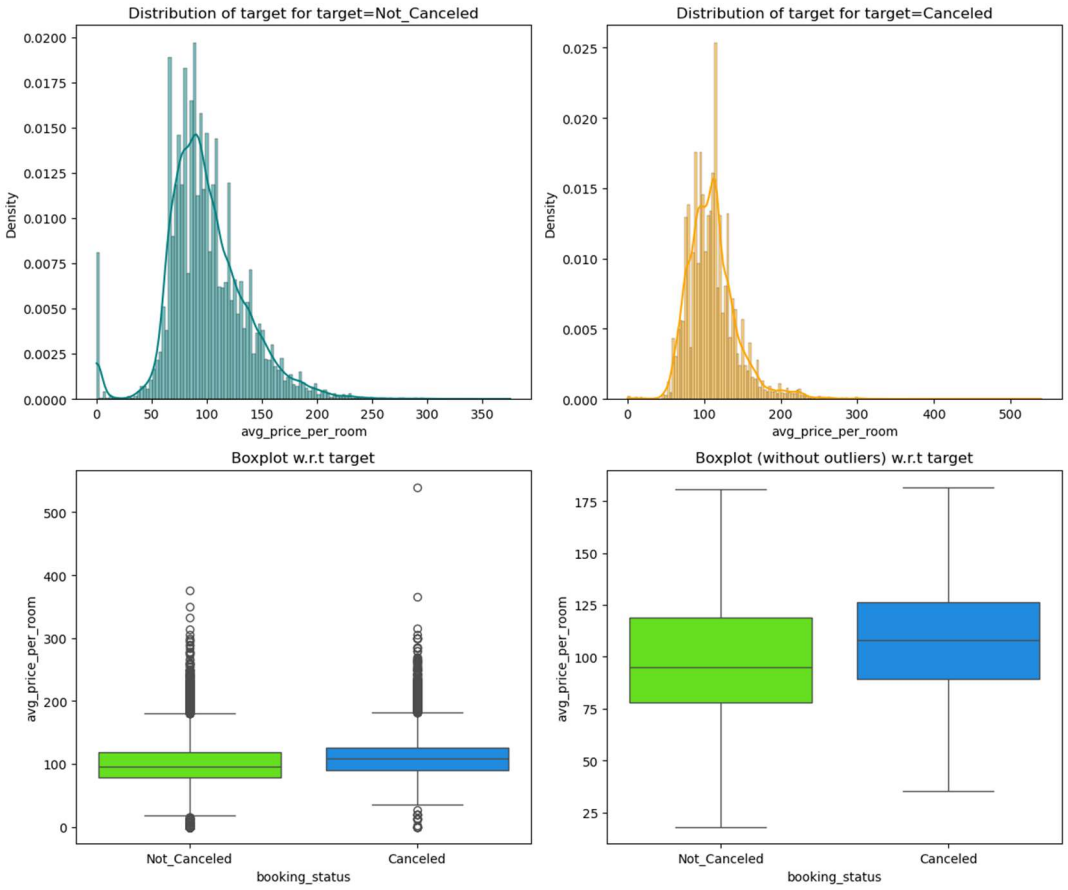
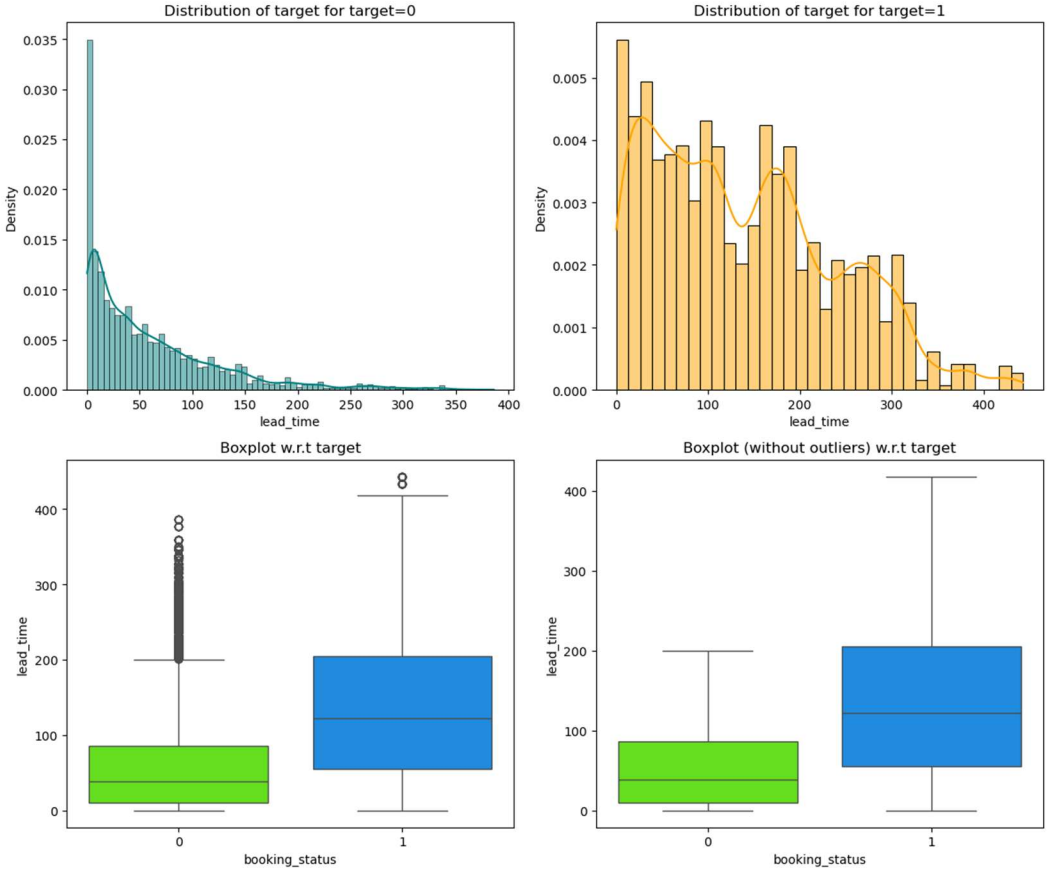


Fig.28 Distribution plot 2



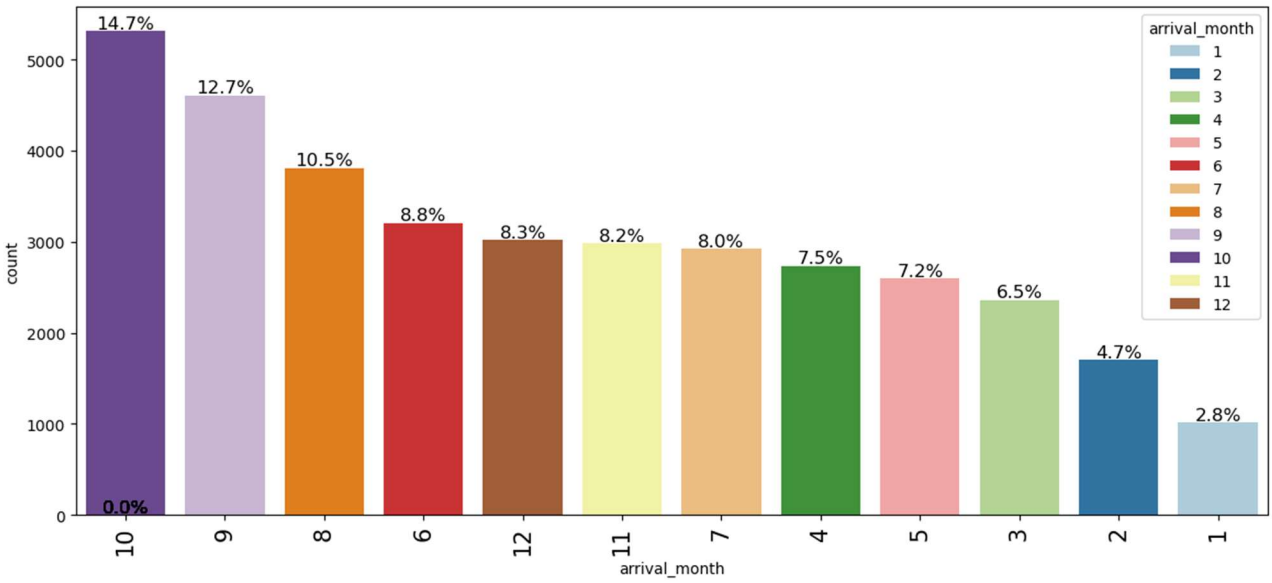
COMMENTS:

- We can observe that there are more cancellations when there are no specific requests.
- There is less chance of cancellation when there are more special requests.
- Online reservations result in the majority of cancellations
- Most bookings that were not canceled tend to have lower average price, with a peak around 50–100.
- The median price for both canceled and not canceled bookings looks similar.
- The distribution is highly skewed to the right, with most lead times below 50 days.
- Shorter lead times are associated with bookings that are less likely to be canceled.
- Customers not requiring a parking space are more likely to cancel, possibly reflecting less committed or more flexible travel plans.

2.3 EDA QUESTIONS:

2.3.1 What are the busiest months in the hotel?

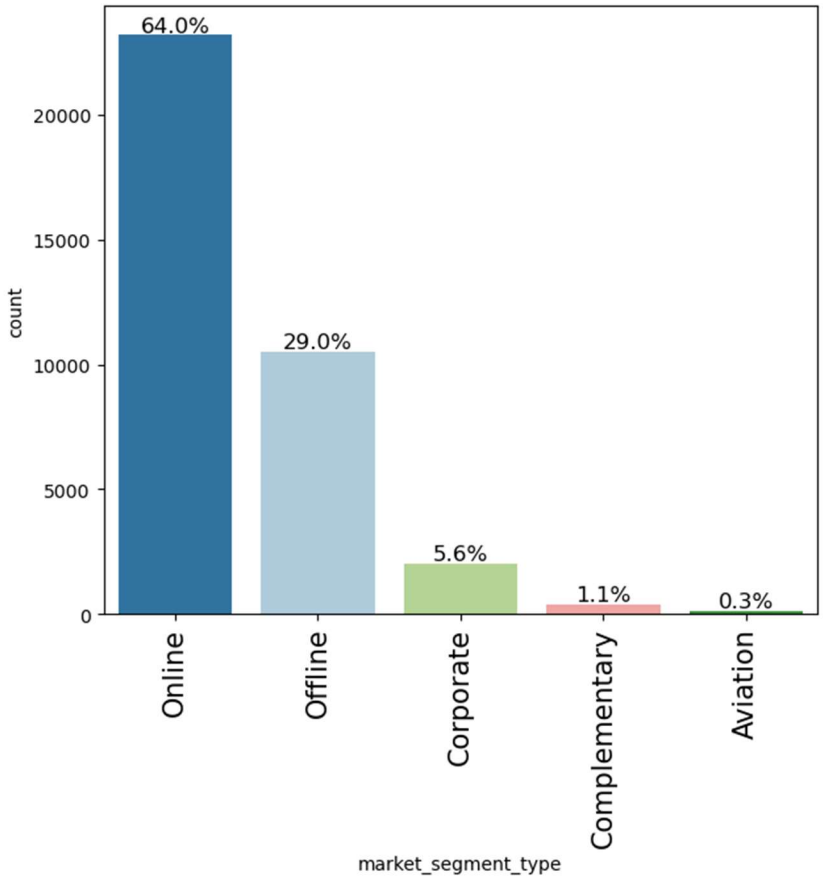
Fig.29 Arrival Month distribution



- October was the busiest month with 14.7% entries followed by September (12.7%), August (10.5%) and June (8.8%).
- The top 3 busiest months of the hotel are October, September and August.

2.3.2 Which market segment do most of the guests come from?

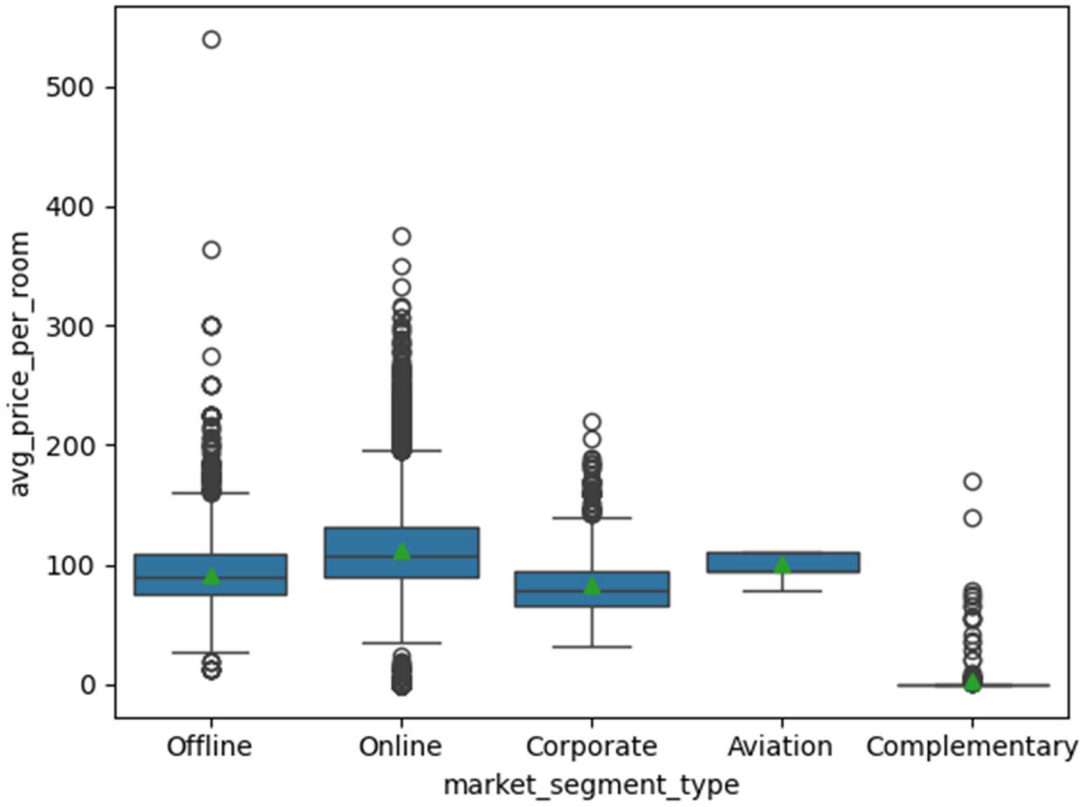
Fig.30 Market Segment distribution



- The "Online" segment accounts for the largest share at 64%.
- This indicates that the majority of transactions or interactions in this dataset occur through online channels.
- The online segment is followed by offline and Corporate.

2.3.3 Hotel rates are dynamic and change according to demand and customer demographics. What are the differences in room prices in different market segments?

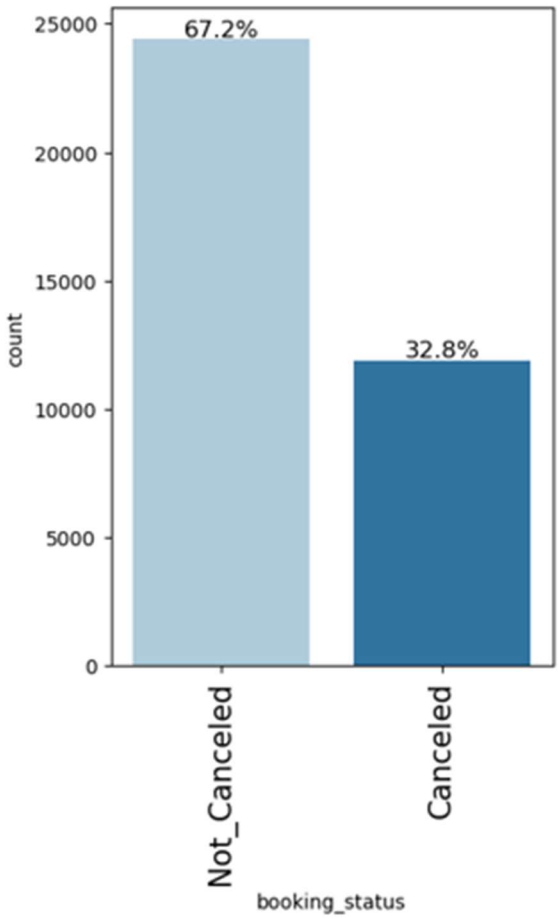
Fig.31 Market segment vs Avg price 2



- As hotel prices are very demand and dynamic as mentioned, Online market segment seems to have the highest median price, since online is where the most booking takes place.
- In general, the prices of Corporate, Offline, and Aviation are marginally lower, with Complementary having the lowest prices as they are free.

2.3.4 What percentage of bookings are canceled?

Fig.32 Booking status distribution

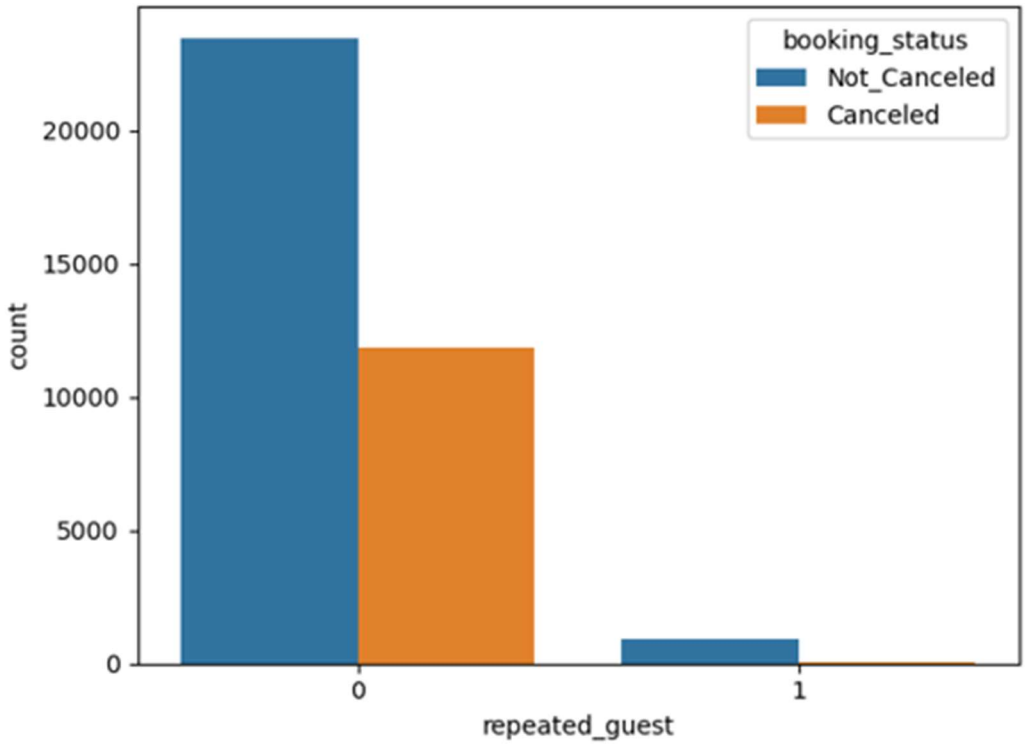


booking_status	Count	Perecentage
Not_Canceled	24390	67.236389
Canceled	11885	32.763611

- 32.7% of the entries are labelled as Canceeled.
- 1/3 of bookings are cancelled

2.3.5 Repeating guests are the guests who stay in the hotel often and are important to brand equity. What percentage of repeating guests cancel?

Fig.33 Repeated Guests vs Booking status

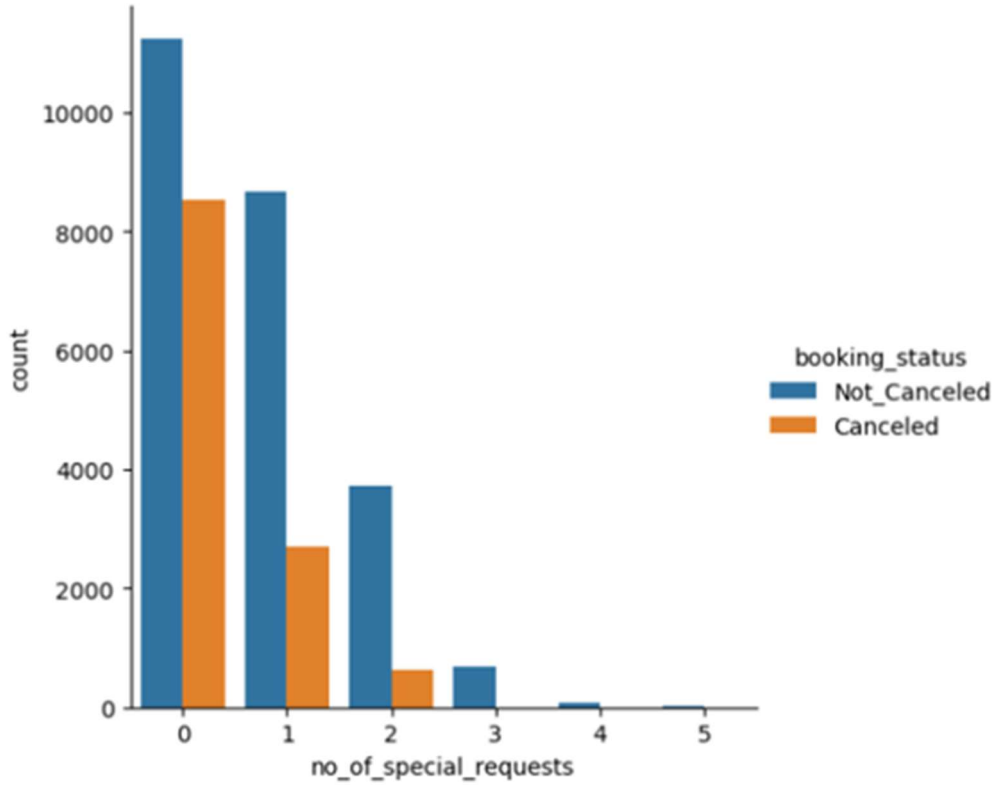



```
repeated_guest  booking_status
0               Not_Canceled      66.419578
                Canceled          33.580422
1               Not_Canceled      98.279570
                Canceled           1.720430
Name: proportion, dtype: float64
```

- Repeating guest very rarely cancel their booking (1.72%) and constitutes very low.

2.3.6 Many guests have special requirements when booking a hotel room. Do these requirements affect booking cancellation?

Fig.34 Special req vs Booking status




```
no_of_special_requests  booking_status
0                       Not_Canceled       56.793245
                        Canceled           43.206755
1                       Not_Canceled       76.233184
                        Canceled           23.766816
2                       Not_Canceled       85.403300
                        Canceled           14.596700
3                       Not_Canceled      100.000000
4                       Not_Canceled      100.000000
5                       Not_Canceled      100.000000
Name: proportion, dtype: float64
```

- As we can clearly observe, that the number of special requests affects the booking cancellation.
- The addition of special request begins to reduce the cancellation rate at one and progressively reduces cancellation to zero from third request.

3. DATA PREPROCESSING

Missing Value Treatment:

Fig.35 Null value checks

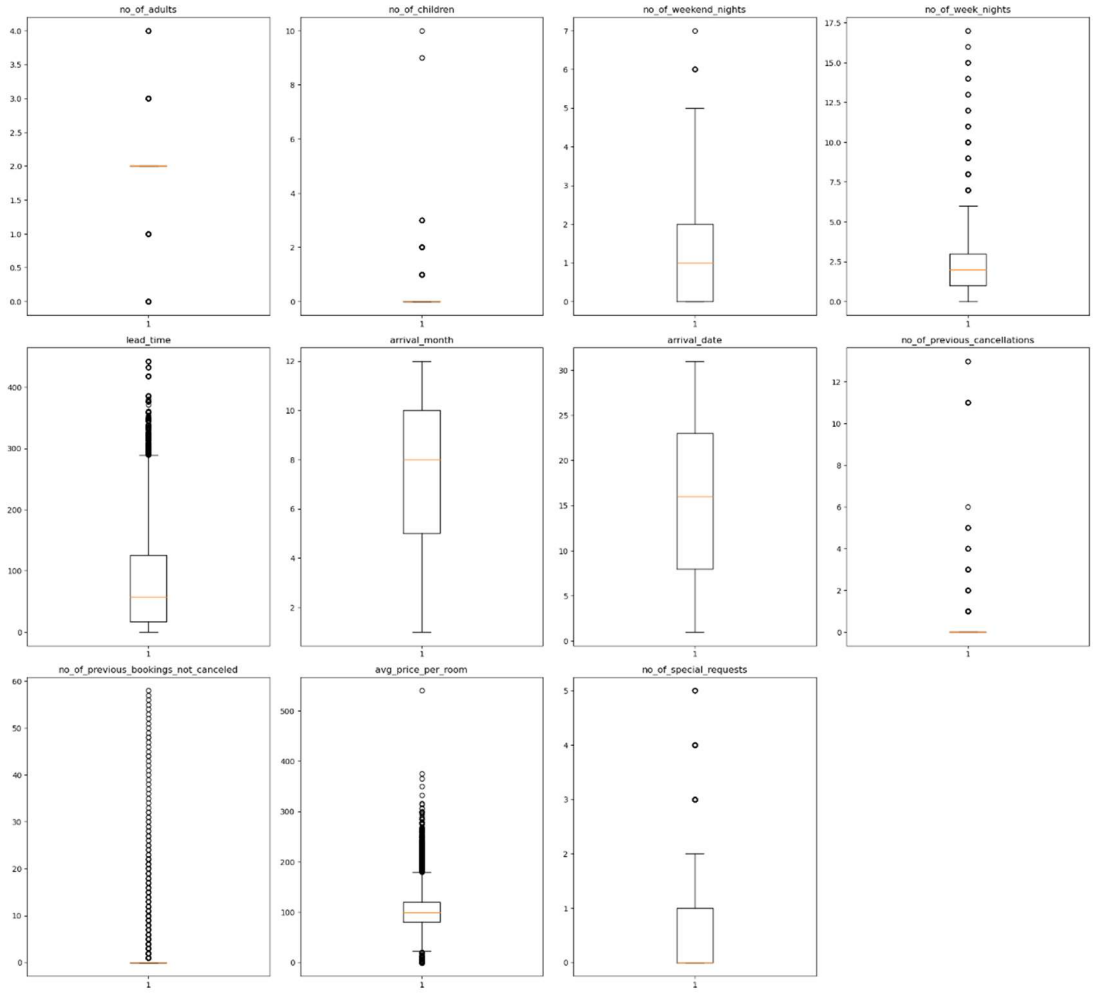
```
Booking_ID                            0
no_of_adults                          0
no_of_children                        0
no_of_weekend_nights                  0
no_of_week_nights                     0
type_of_meal_plan                     0
required_car_parking_space            0
room_type_reserved                    0
lead_time                             0
arrival_year                          0
arrival_month                         0
arrival_date                          0
market_segment_type                   0
repeated_guest                        0
no_of_previous_cancellations          0
no_of_previous_bookings_not_canceled  0
avg_price_per_room                    0
no_of_special_requests                0
booking_status                        0
dtype: int64
```

There are no null values or missing values in this dataset.

Outlier Treatment:

- As we can clearly see, there are many outliers in several columns in this dataset.
- These outliers are meaningful and we will not treat them.

Fig.36 Outlier checks



Feature Engineering:

- The "booking_id column" is dropped, as it is not useful.
- The "arrival_date" columns is dropped, because the distribution of data is approximately equal across all the dates except 31
- Dropping "arrival_year", as there is only 2 years.
- There are some entries with "no_of_children" as 9 and 10. They seems like misinterpreted values and they are replaced with the maximum value of 3.
- The target variables are labelled as 0 and 1 (Not cancelled: 0, Cancelled: 1)
- Data is splitted in the ratio of 7:3 for training and testing.
- One-hot Encoding is used for "room_type_reserved", "market_segment_type" and "type_of_meal_plan".

4. MODEL BUILDING – LOGISTIC REGRESSION

This is the initial model summary before treating multicollinearity and p-values.

Fig.37 Initial Model Summary

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25366
Method:	MLE	Df Model:	25
Date:	Sun, 17 Nov 2024	Pseudo R-squ.:	0.3304
Time:	01:46:25	Log-Likelihood:	-10753.
converged:	False	LL-Null:	-16060.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.6266	0.272	-9.649	0.000	-3.160	-2.093
no_of_adults	0.0239	0.038	0.633	0.526	-0.050	0.098
no_of_children	0.0815	0.063	1.289	0.197	-0.042	0.205
no_of_weekend_nights	0.1497	0.020	7.560	0.000	0.111	0.189
no_of_week_nights	0.0354	0.012	2.887	0.004	0.011	0.059
required_car_parking_space	-1.6401	0.137	-11.983	0.000	-1.908	-1.372
lead_time	0.0163	0.000	63.001	0.000	0.016	0.017
arrival_month	-0.0673	0.006	-11.211	0.000	-0.079	-0.056
repeated_guest	-2.0376	0.744	-2.738	0.006	-3.496	-0.579
no_of_previous_cancellations	0.3513	0.102	3.442	0.001	0.151	0.551
no_of_previous_bookings_not_canceled	-1.2961	0.878	-1.477	0.140	-3.016	0.424
avg_price_per_room	0.0198	0.001	27.457	0.000	0.018	0.021
no_of_special_requests	-1.4750	0.030	-48.805	0.000	-1.534	-1.416
room_type_reserved_Room_Type 2	-0.4158	0.133	-3.124	0.002	-0.677	-0.155
room_type_reserved_Room_Type 3	1.0909	1.797	0.607	0.544	-2.432	4.614
room_type_reserved_Room_Type 4	-0.2456	0.053	-4.614	0.000	-0.350	-0.141
room_type_reserved_Room_Type 5	-0.6404	0.215	-2.977	0.003	-1.062	-0.219
room_type_reserved_Room_Type 6	-0.9097	0.154	-5.888	0.000	-1.212	-0.607
room_type_reserved_Room_Type 7	-1.4045	0.299	-4.695	0.000	-1.991	-0.818
market_segment_type_Complementary	-15.9555	294.223	-0.054	0.957	-592.621	560.710
market_segment_type_Corporate	-0.9505	0.276	-3.438	0.001	-1.492	-0.409
market_segment_type_Offline	-1.8885	0.264	-7.142	0.000	-2.407	-1.370
market_segment_type_Online	-0.0853	0.262	-0.326	0.744	-0.598	0.428
type_of_meal_plan_Meal Plan 2	0.0554	0.065	0.858	0.391	-0.071	0.182
type_of_meal_plan_Meal Plan 3	10.5231	155.309	0.068	0.946	-293.876	314.922
type_of_meal_plan_Not Selected	0.2636	0.053	5.009	0.000	0.160	0.367

4.1 TREATING MULTICOLLINEARITY AND P-VALUE VARIABLES:

Fig.38 VIF values before and after treating

VIF values:

const	318.673648
no_of_adults	1.344685
no_of_children	2.085915
no_of_weekend_nights	1.064976
no_of_week_nights	1.093798
required_car_parking_space	1.034669
lead_time	1.251766
arrival_month	1.050871
repeated_guest	1.748572
no_of_previous_cancellations	1.321104
no_of_previous_bookings_not_canceled	1.565753
avg_price_per_room	1.911579
no_of_special_requests	1.245456
room_type_reserved_Room_Type 2	1.096033
room_type_reserved_Room_Type 3	1.003719
room_type_reserved_Room_Type 4	1.352339
room_type_reserved_Room_Type 5	1.030425
room_type_reserved_Room_Type 6	2.035961
room_type_reserved_Room_Type 7	1.093409
market_segment_type_Complementary	4.350270
market_segment_type_Corporate	16.609607
market_segment_type_Offline	62.343603
market_segment_type_Online	69.339302
type_of_meal_plan_Meal Plan 2	1.196447
type_of_meal_plan_Meal Plan 3	1.007906
type_of_meal_plan_Not Selected	1.241857

dtype: float64

VIF values:

const	38.332470
no_of_adults	1.327850
no_of_children	2.084994
no_of_weekend_nights	1.064533
no_of_week_nights	1.093213
required_car_parking_space	1.034596
lead_time	1.249320
arrival_month	1.050718
repeated_guest	1.745278
no_of_previous_cancellations	1.320982
no_of_previous_bookings_not_canceled	1.565522
avg_price_per_room	1.911422
no_of_special_requests	1.240928
room_type_reserved_Room_Type 2	1.095840
room_type_reserved_Room_Type 3	1.003718
room_type_reserved_Room_Type 4	1.347772
room_type_reserved_Room_Type 5	1.030422
room_type_reserved_Room_Type 6	2.035726
room_type_reserved_Room_Type 7	1.093279
market_segment_type_Complementary	1.319317
market_segment_type_Corporate	1.529895
market_segment_type_Offline	1.597484
type_of_meal_plan_Meal Plan 2	1.196395
type_of_meal_plan_Meal Plan 3	1.007906
type_of_meal_plan_Not Selected	1.240450

dtype: float64

- If VIF is between 1 and 5, then there is low multicollinearity.
- As we can clearly observe from the above table, some of the variables has greater VIF values indicating high multicollinearity.
- So, the necessary columns are removed and treated.

Dealing with high p-value variables:

- After treating multicollinearity in the data, we can deal the variables having high p-values.
- Some variables have p-value greater than 0.05 and can be dropped one by one, as they are not significant in predicting "booking_status".

Model Summary after treating multicollinearity and dealing p-values:

Fig.39 Summary 2

Logit Regression Results

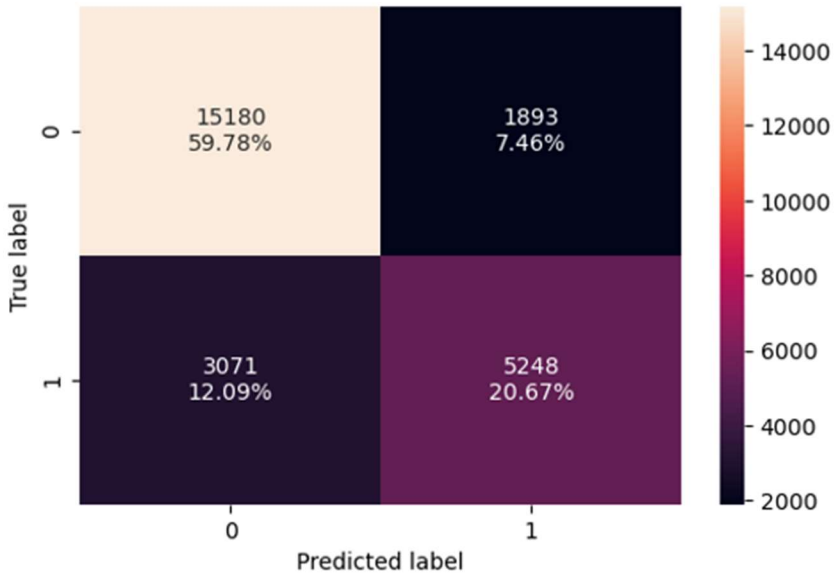
Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25374
Method:	MLE	Df Model:	17
Date:	Sun, 17 Nov 2024	Pseudo R-squ.:	0.3296
Time:	01:46:38	Log-Likelihood:	-10767.
converged:	True	LL-Null:	-16060.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.7358	0.092	-29.603	0.000	-2.917	-2.555
no_of_weekend_nights	0.1525	0.020	7.715	0.000	0.114	0.191
no_of_week_nights	0.0364	0.012	2.972	0.003	0.012	0.060
required_car_parking_space	-1.6394	0.137	-11.986	0.000	-1.907	-1.371
lead_time	0.0164	0.000	64.203	0.000	0.016	0.017
arrival_month	-0.0680	0.006	-11.338	0.000	-0.080	-0.056
repeated_guest	-3.0374	0.599	-5.067	0.000	-4.212	-1.863
no_of_previous_cancellations	0.2867	0.078	3.690	0.000	0.134	0.439
avg_price_per_room	0.0204	0.001	29.982	0.000	0.019	0.022
no_of_special_requests	-1.4726	0.030	-49.164	0.000	-1.531	-1.414
room_type_reserved_Room_Type 2	-0.3763	0.129	-2.919	0.004	-0.629	-0.124
room_type_reserved_Room_Type 4	-0.2532	0.051	-4.932	0.000	-0.354	-0.153
room_type_reserved_Room_Type 5	-0.6513	0.214	-3.038	0.002	-1.071	-0.231
room_type_reserved_Room_Type 6	-0.8136	0.119	-6.860	0.000	-1.046	-0.581
room_type_reserved_Room_Type 7	-1.3781	0.292	-4.713	0.000	-1.951	-0.805
market_segment_type_Corporate	-0.8794	0.103	-8.529	0.000	-1.082	-0.677
market_segment_type_Offline	-1.7928	0.050	-35.894	0.000	-1.891	-1.695
type_of_meal_plan_Not Selected	0.2689	0.052	5.153	0.000	0.167	0.371

Training set Performance:

Fig.40 Metrics train lg1

	Accuracy	Recall	Precision	F1
0	0.804505	0.630845	0.734911	0.678913



Test set Performance:

Fig.41 Metrics test lg1

	Accuracy	Recall	Precision	F1
0	0.799504	0.621985	0.726737	0.670293

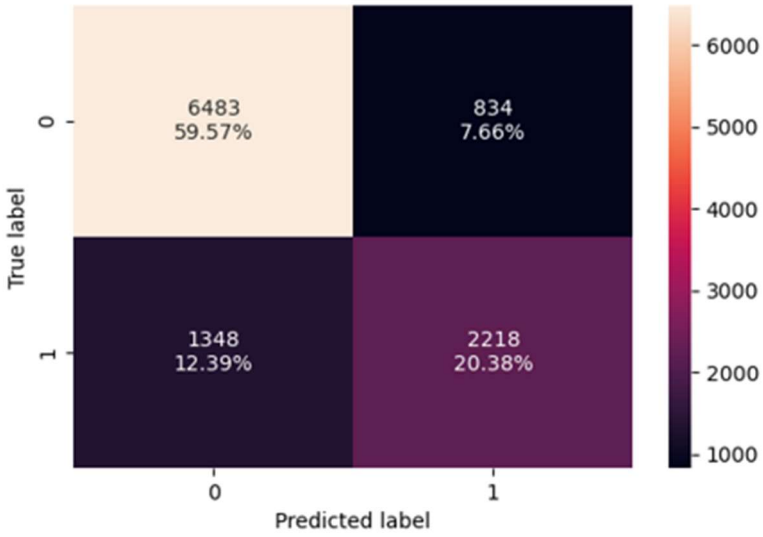


Fig.42 Coefficient Interpretations

	Odds	Change_odd%
const	0.064844	-93.515634
no_of_weekend_nights	1.164793	16.479261
no_of_week_nights	1.037052	3.705179
required_car_parking_space	0.194103	-80.589658
lead_time	1.016553	1.655293
arrival_month	0.934300	-6.570017
repeated_guest	0.047958	-95.204213
no_of_previous_cancellations	1.332056	33.205602
avg_price_per_room	1.020612	2.061228
no_of_special_requests	0.229324	-77.067647
room_type_reserved_Room_Type 2	0.686378	-31.362175
room_type_reserved_Room_Type 4	0.776334	-22.366635
room_type_reserved_Room_Type 5	0.521365	-47.863549
room_type_reserved_Room_Type 6	0.443251	-55.674941
room_type_reserved_Room_Type 7	0.252069	-74.793125
market_segment_type_Corporate	0.415016	-58.498434
market_segment_type_Offline	0.166501	-83.349933
type_of_meal_plan_Not Selected	1.308544	30.854446

4.2 MODEL PERFORMANCE IMPROVEMENT:

- The optimal threshold is obtained from AUC-ROC curve.
- Optimal threshold = 0.3104

Training set Performance:

Fig.43 ROC-AUC curve train

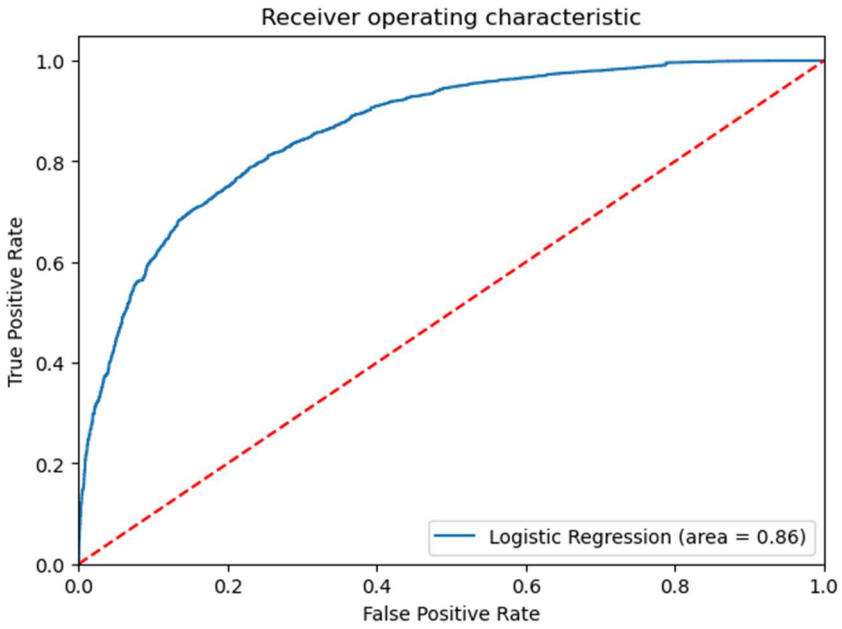
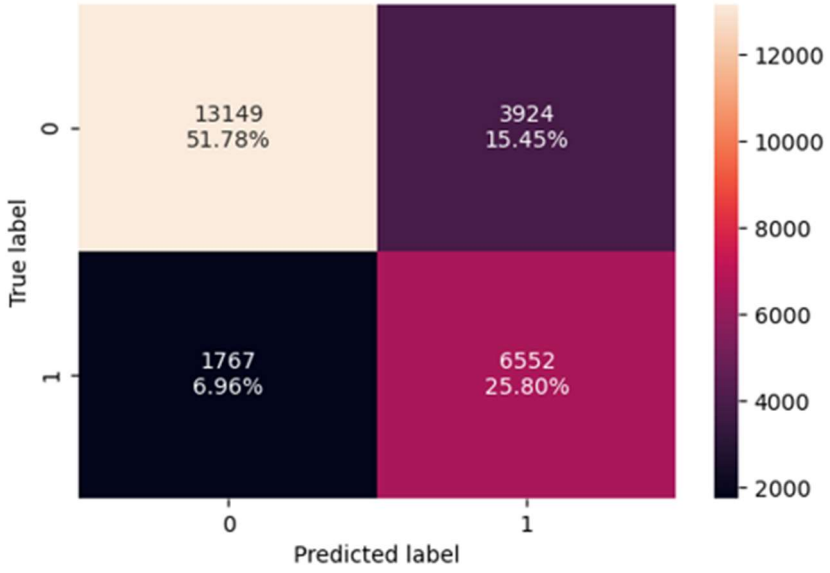


Fig.44 Model metrics lg train

	Accuracy	Recall	Precision	F1
0	0.775874	0.787595	0.62543	0.697207



Test set Performance:

Fig.45 ROC on Test set

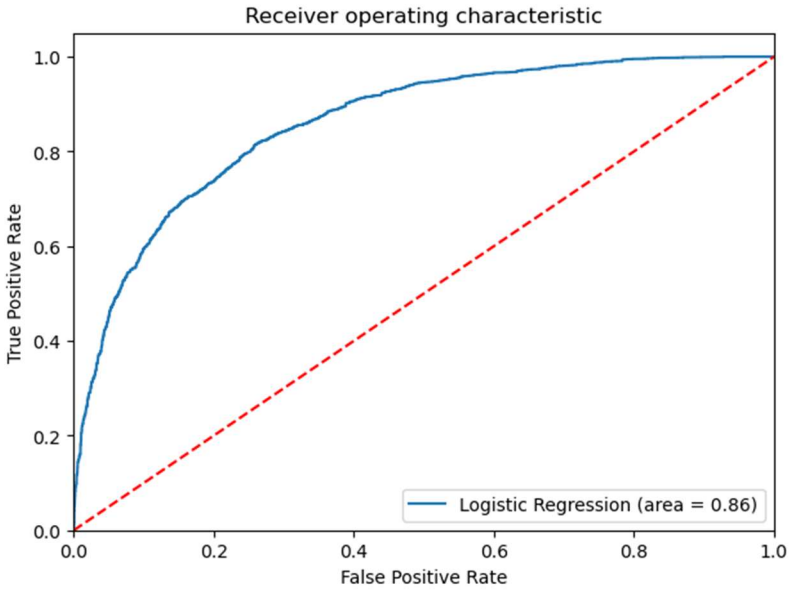
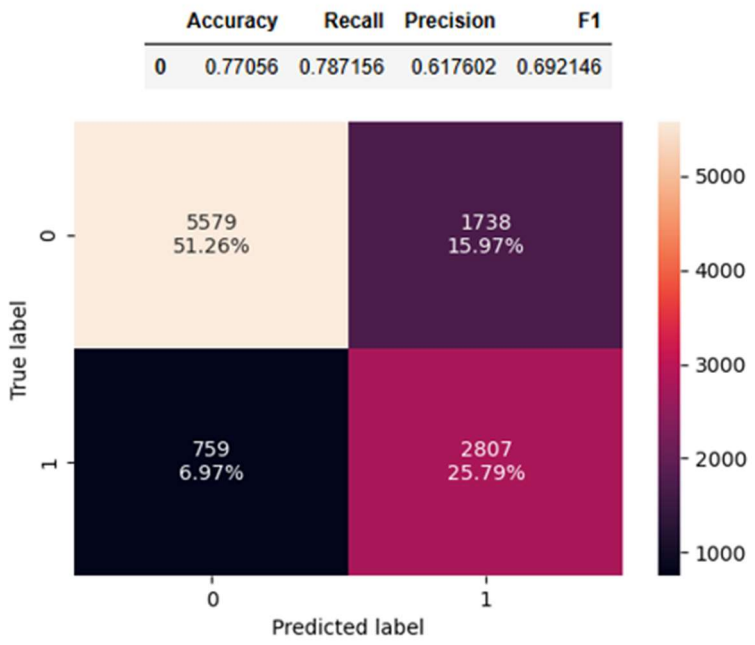


Fig.46 Model metrics lg test



	Accuracy	Recall	Precision	F1
0	0.77056	0.787156	0.617602	0.692146

COMMENTS:

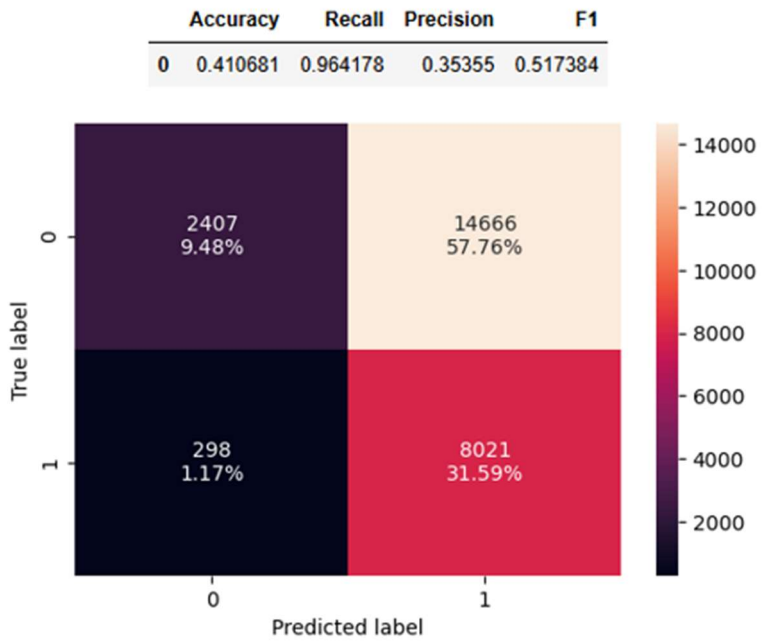
- Accuracy of default threshold model indicates that model performs well.
- Accuracy is decreased by 2% in ROC-AUC model compared to the default threshold model.
- The F1 score is increased.

5. MODEL BUILDING – NAIVE BAYES MODEL

- The Gausian Navie Bayes Model from sklearn library is used here.

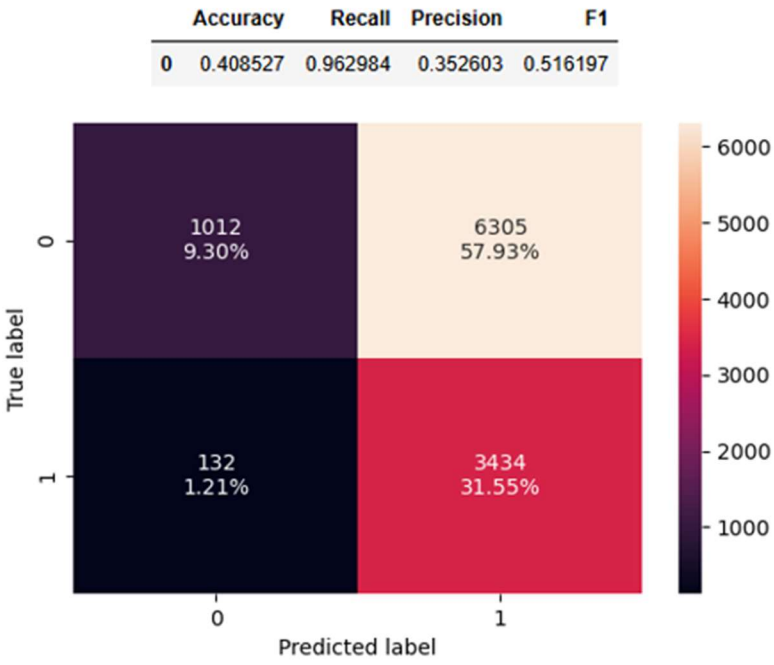
Training set Performance:

Fig.47 Train Perf NB



Test set Performance:

Fig.48 Train Perf NB



COMMENTS:

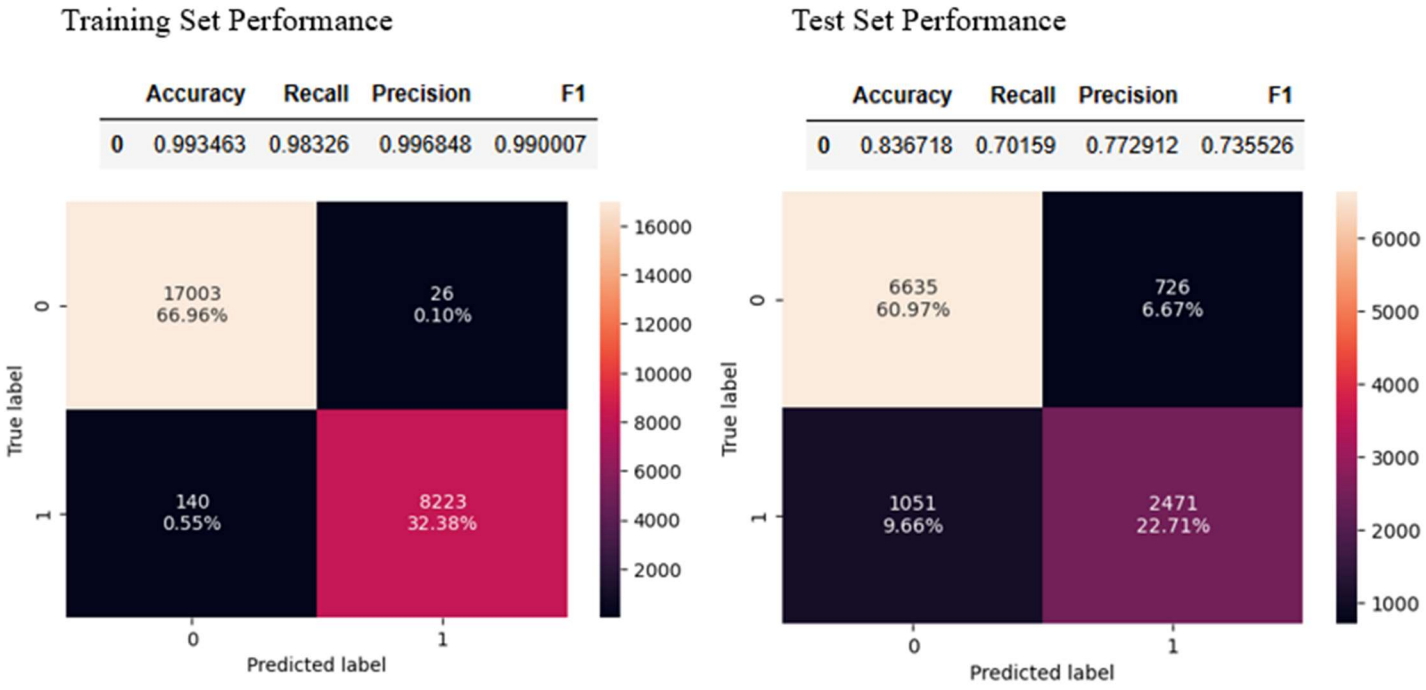
- The Naive- Bayes model poorly performs in both training and testing set.
- Comparing to other models, the accuracy score and F1 score is very low.

6. MODEL BUILDING –

KNN MODEL

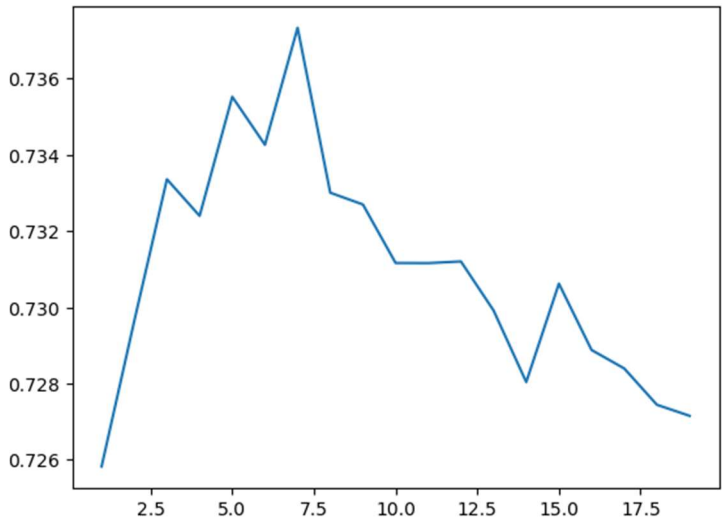
- Performance of the initial KNN model when the k value is set as 5.

Fig.49 Initail knn model metrics



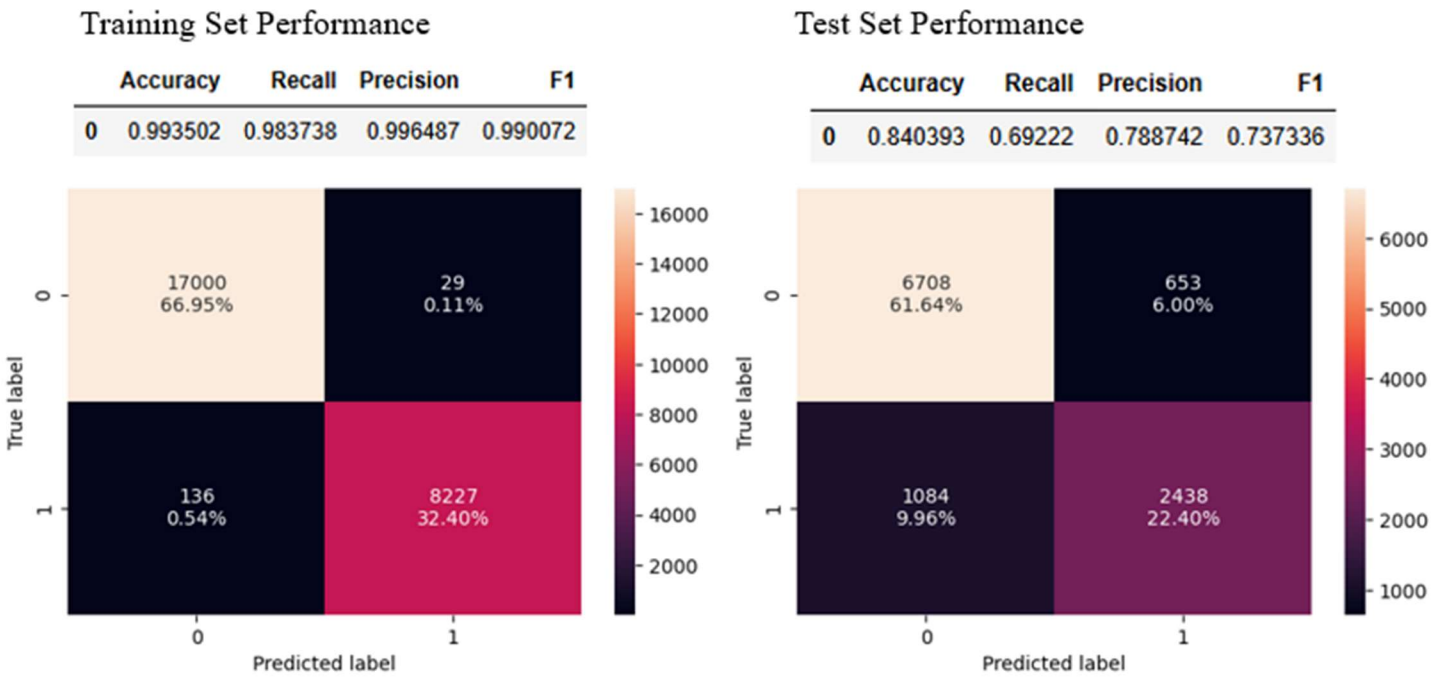
- The best K-value is chosen on the basis of F1 score.
- In the context of predicting booking cancellations for INN Hotels Group, using the F1 score is recommended as it provides a balanced evaluation of both precision and recall.

Fig.50 k-values vs f1-score



➢ Performance of the KNN model after choosing the best k-value (7).

Fig.51 Final knn model metrics



COMMENTS:

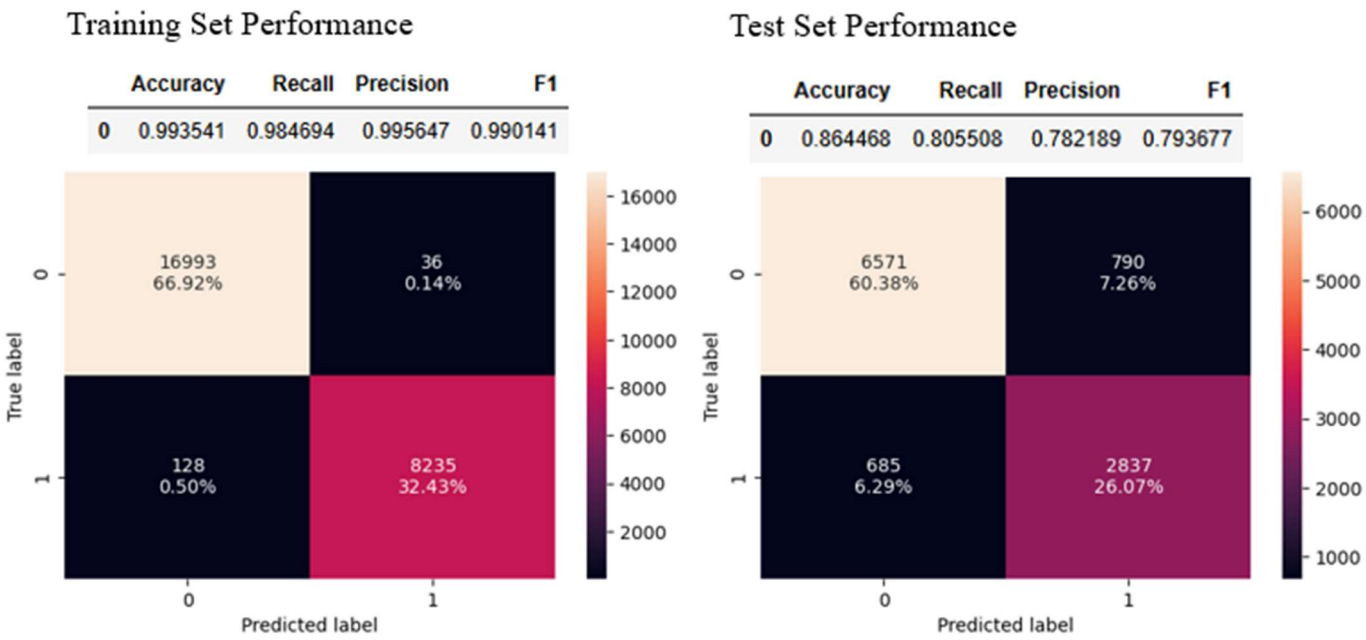
- A significant improvement is visible after selecting the best k values as 7.
- Accuracy and F1 scores are increased.
- Even though the KNN model performs very well in training set, it still underperforms in test set.

7. MODEL BUILDING – DECISION TREE MODEL

7.1 DECISION TREE (default)

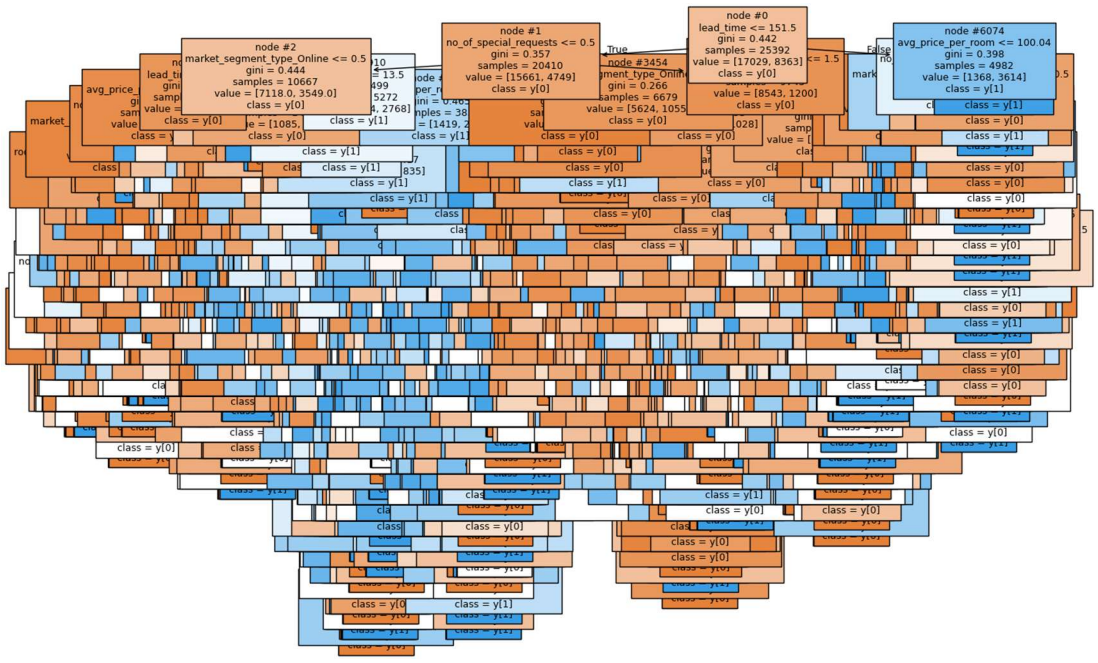
Initial Model Perfomance:

Fig.52 Initial DT perf



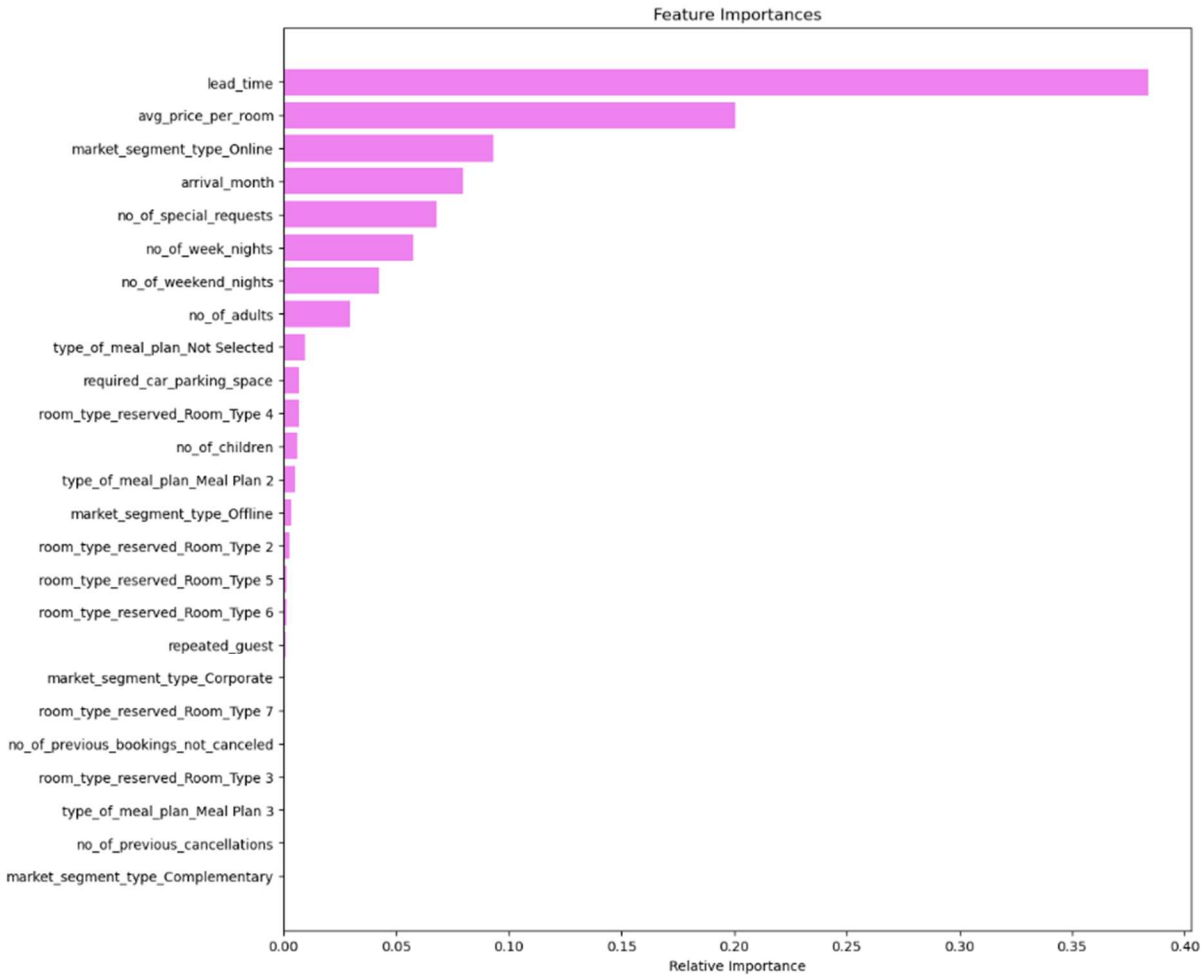
Decision Tree:

Fig.53 Visual of initial DT



Feature Importances:

Fig.54 Feature importances Initial DT



COMMENTS:

- The model tends to overfit with an accuracy of 99% in training set.
- As a result of overfitting, the model poorly performs in the test dataset.
- According to the decision tree, lead_time is the most important variable for predicting the booking status.
- The tree above is very complex and such a tree often overfits.

7.2 DECISION TREE (max_depth = 4)

- The max_depth is restricted to 4 to reduce the complexity of the tree.
- Model Performance:

Fig.55 DT-4 perf

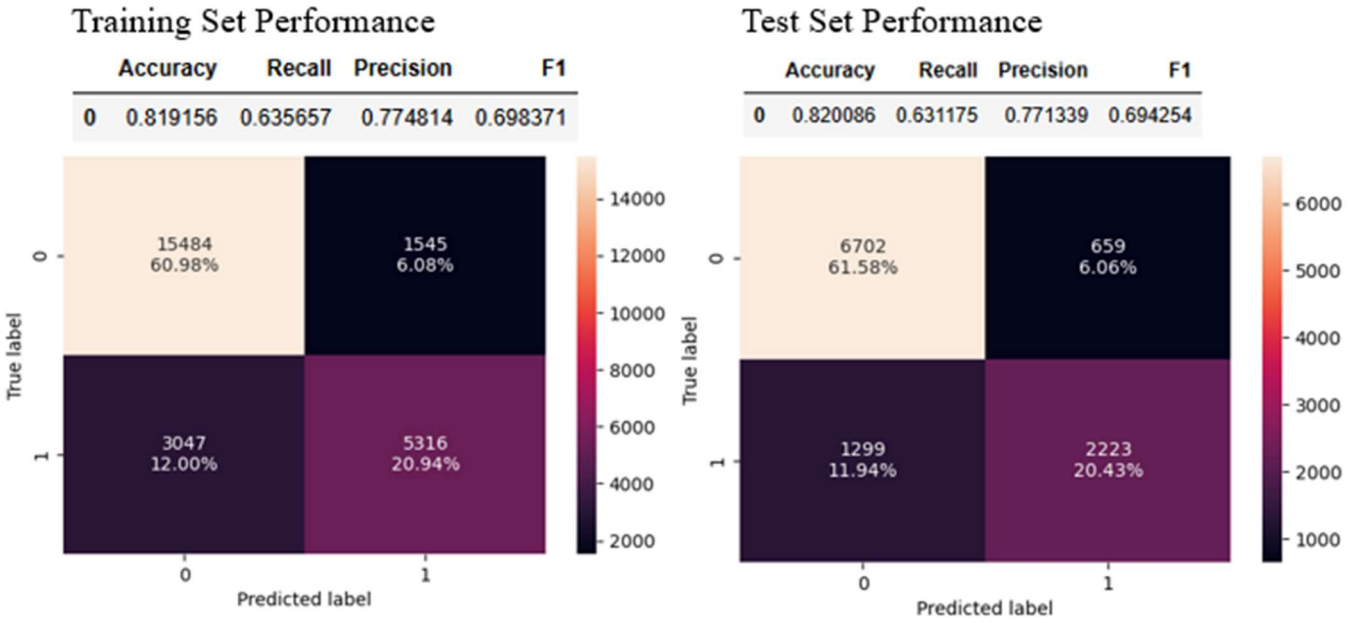


Fig.56 Visual of DT-4

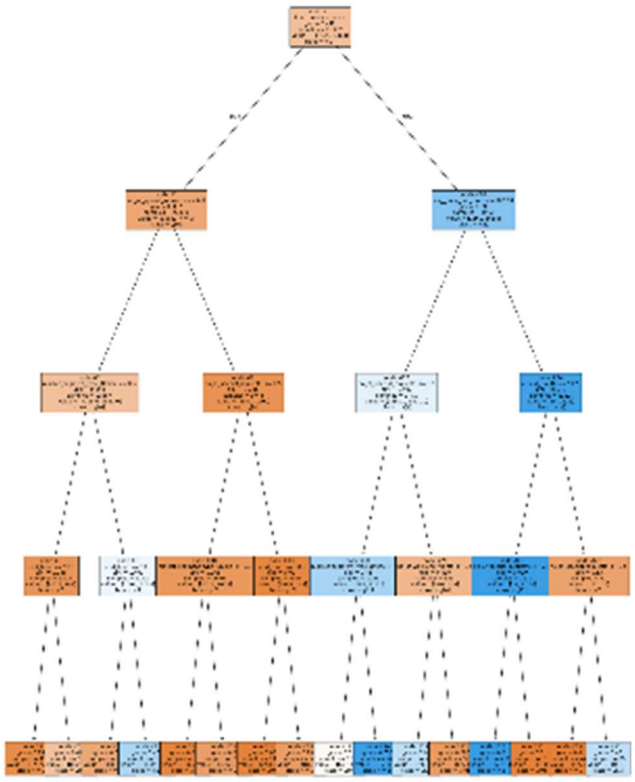
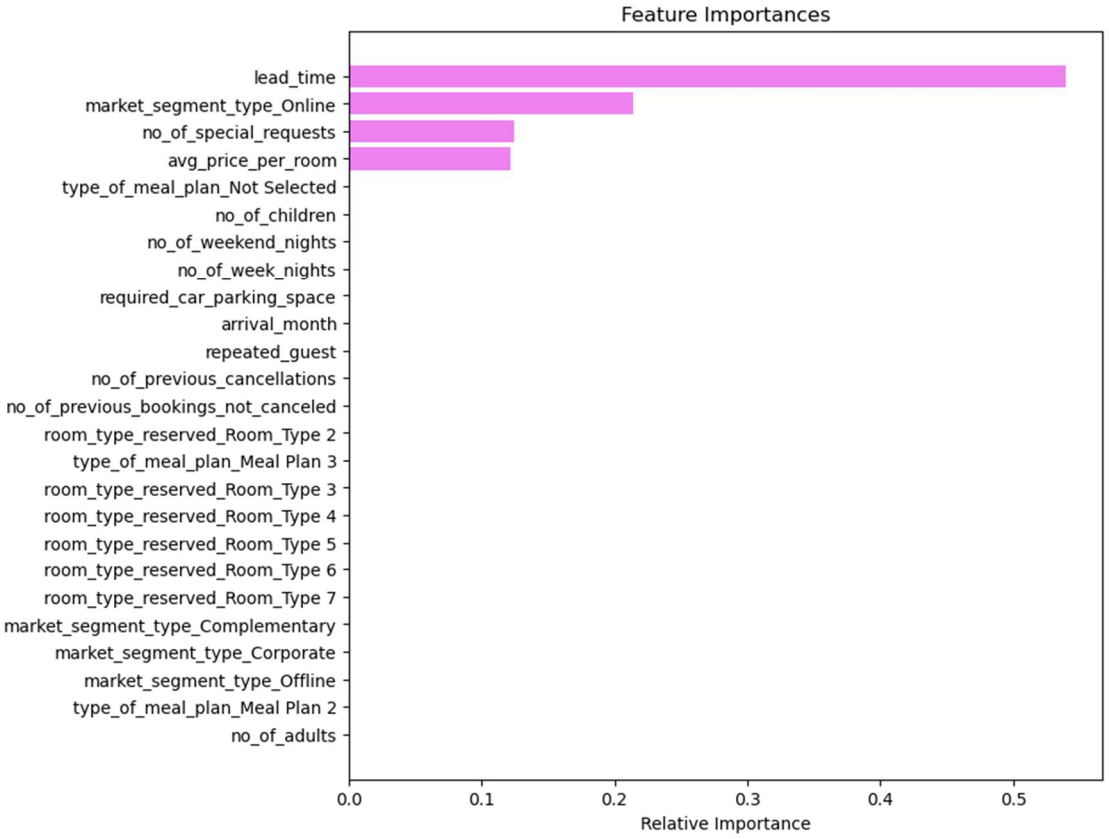


Fig.57 Feature Importances of DT-4



COMMENTS:

- According to this decision tree, the “lead_time”, “market_segment”, “no_of_special_requests” and “avg_price_per_room” are the important variables for predicting the booking status.
- Training accuracy and Test accuracy are pretty much same.
- The F1_score is reduced in this decision tree and we will proceed with pre-pruning.

7.3 DECISION TREE (Pre-pruned)

- The GridSearchCV is used to find the best parameters for pre-pruning.
- The Decision tree with parameters max_depth=3, max_leaf_nodes=5, min_impurity_decrease=0.001 are used.

Fig.58 DT pre-pruned perf

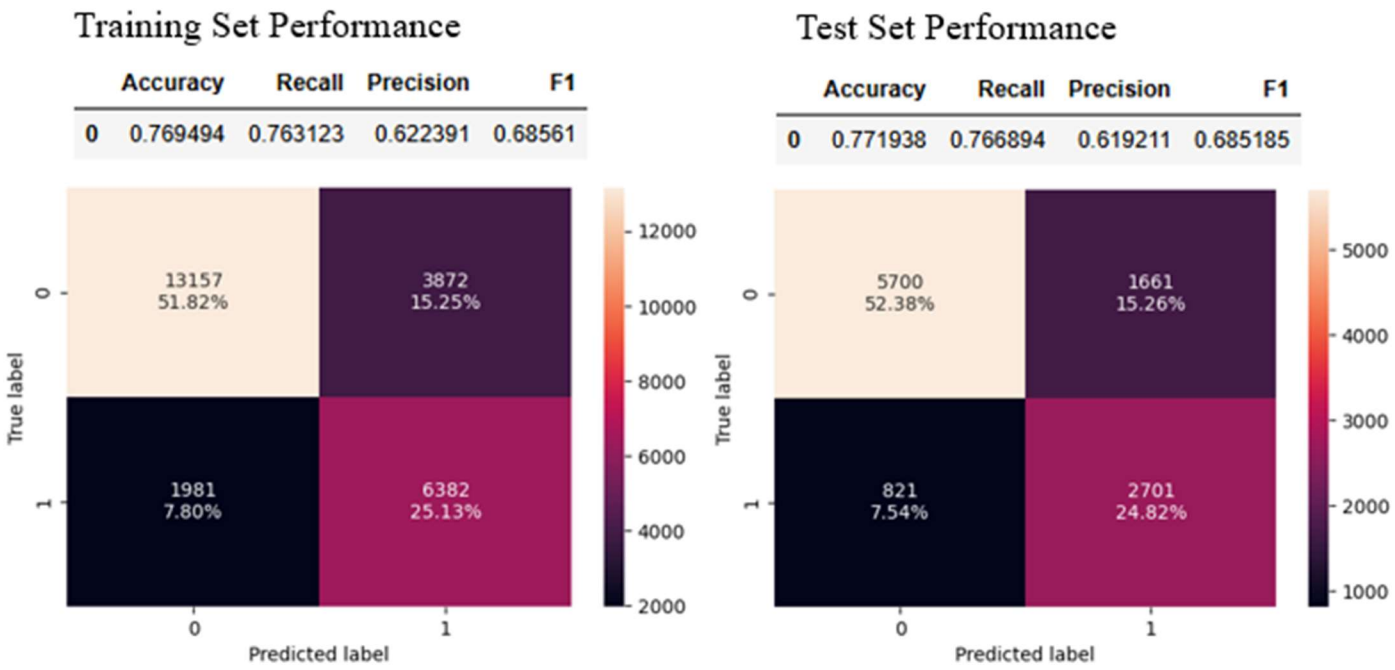


Fig.59 Visual of DT pre-pruned

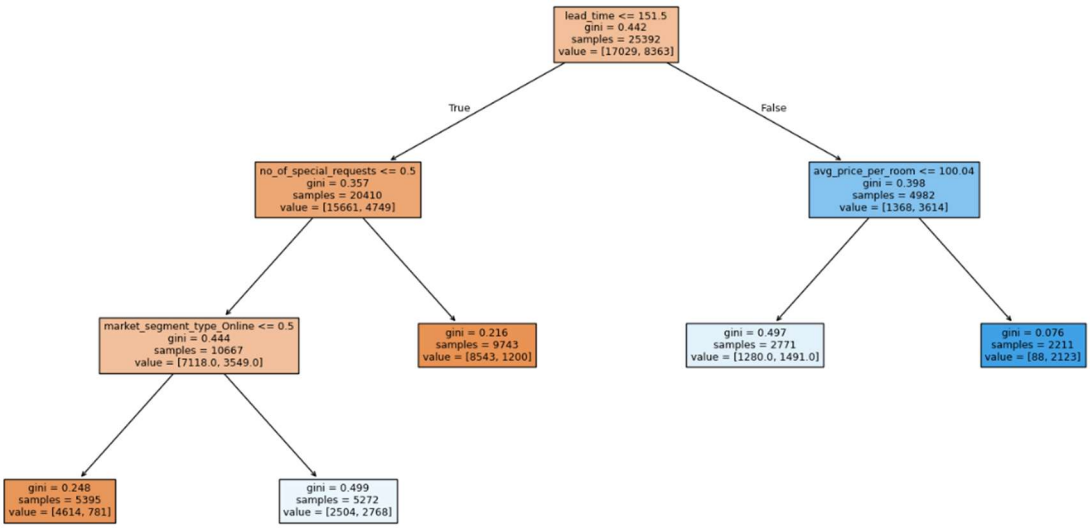
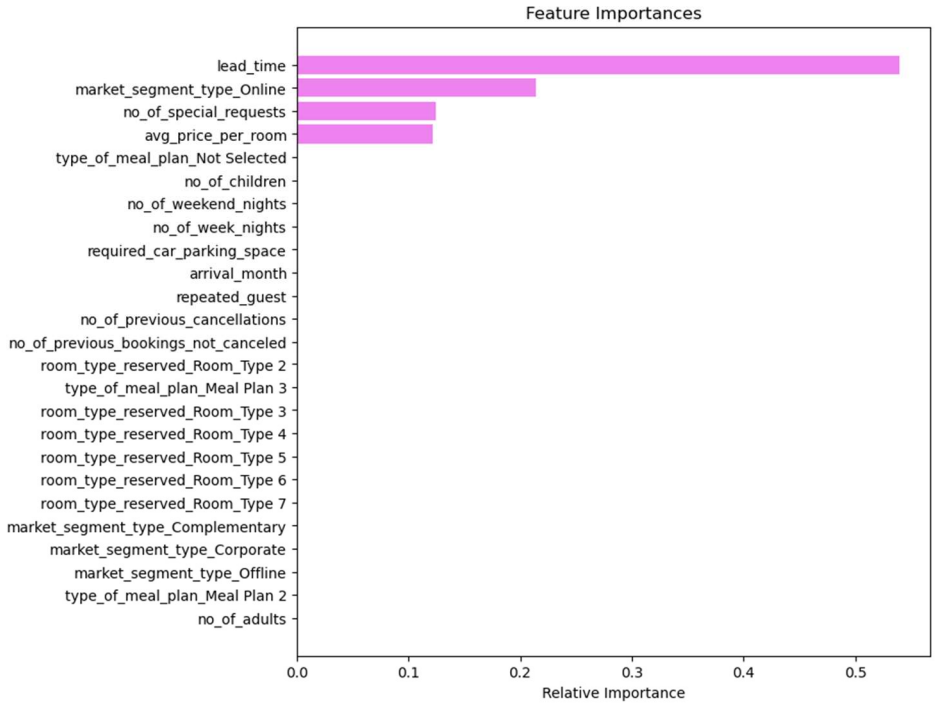


Fig.60 Features of DT pre-pruned



COMMENTS:

- The important features are same with compared to the decision tree with max_depth = 4
- In pre-pruning, there is no improvement in accuracy scores and F1 scores. We will proceed with Post pruning.

7.4 DECISION TREE (Post-pruned)

- This pruning technique is parameterized by the cost complexity parameter, ccp_alpha.
- Greater values of ccp_alpha increase the number of nodes pruned.
- In the context of predicting booking cancellations for INN Hotels Group, using the F1 score is recommended as it provides a balanced evaluation of both precision and recall.
- So, the best model is selected in the basis of F1 score.

Fig.61 Total Impurity vs Effective alpha

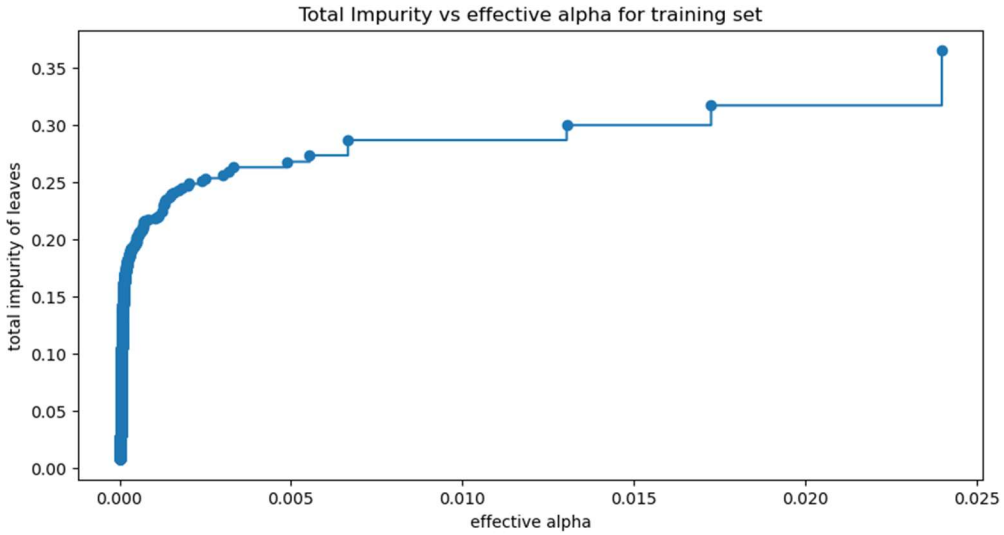


Fig.62 Nodes and Depth vs alpha

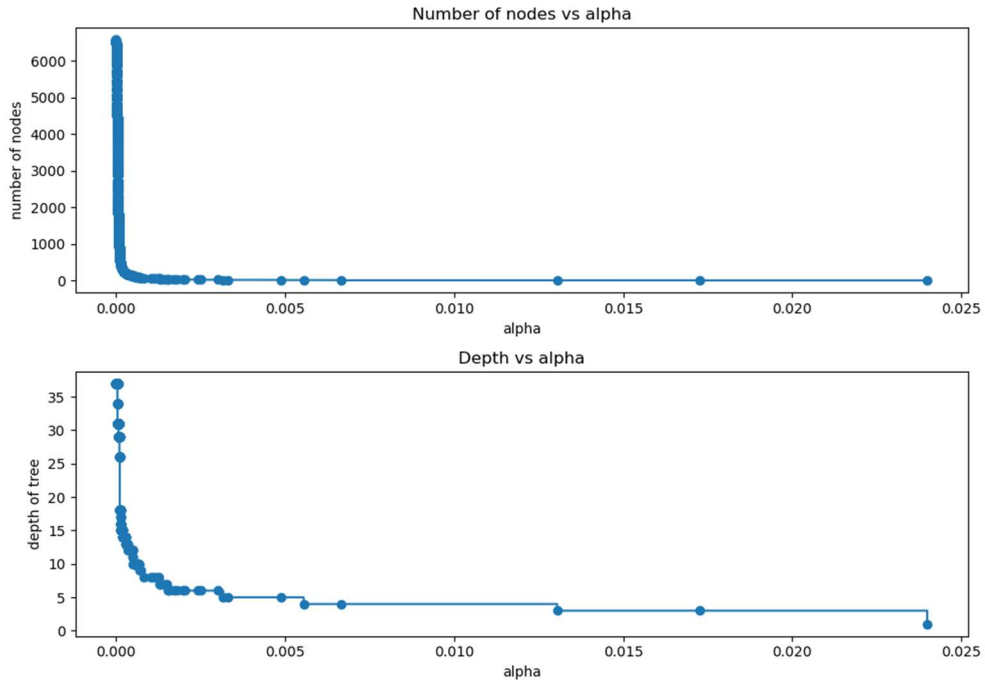


Fig.63 Accuracy vs alpha

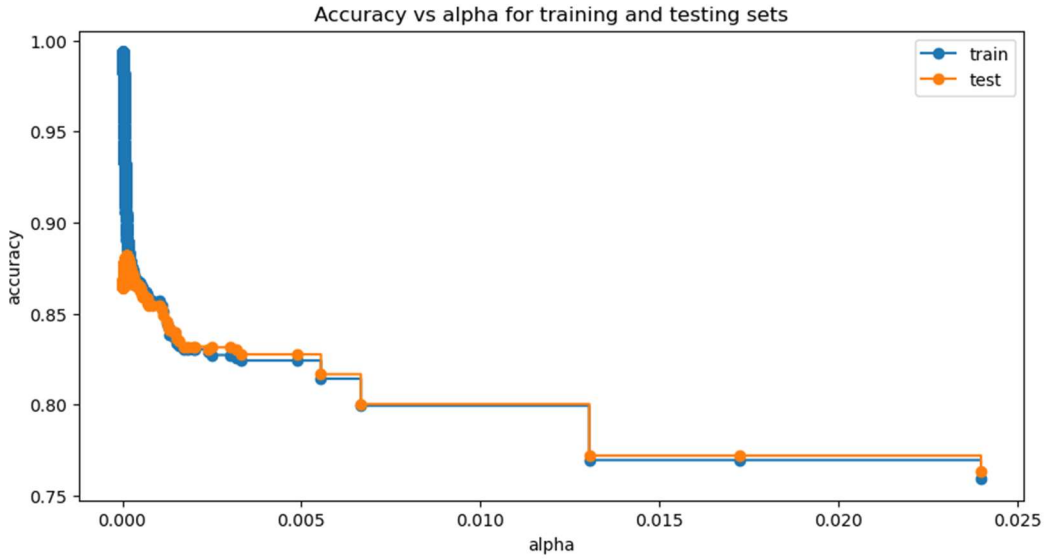


Fig.64 F1 score vs alpha

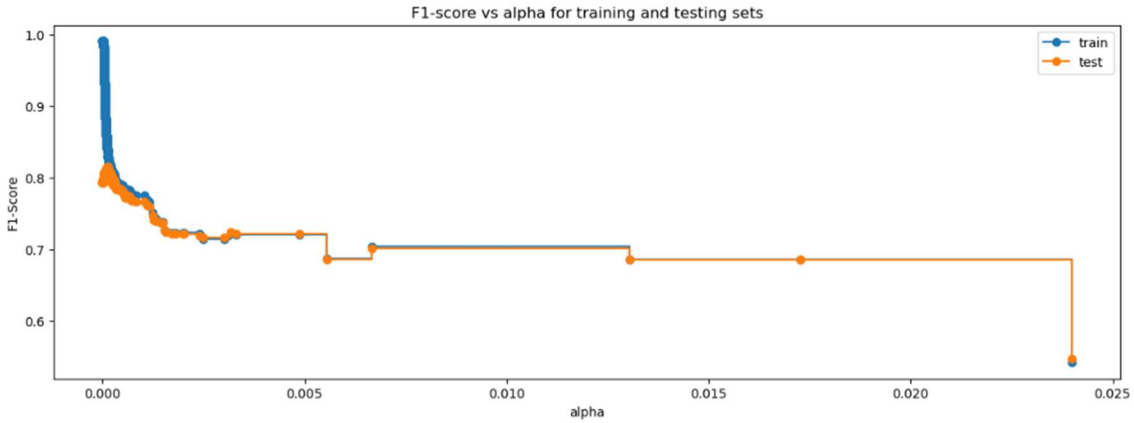
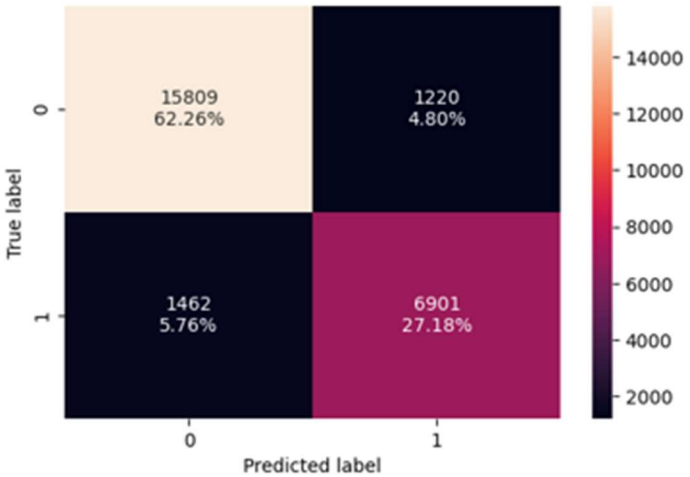


Fig.65 Performance of DT post-pruned

Training Set Performance

	Accuracy	Recall	Precision	F1
0	0.894376	0.825182	0.849772	0.837297



Test Set Performance

	Accuracy	Recall	Precision	F1
0	0.881834	0.801533	0.827859	0.814484

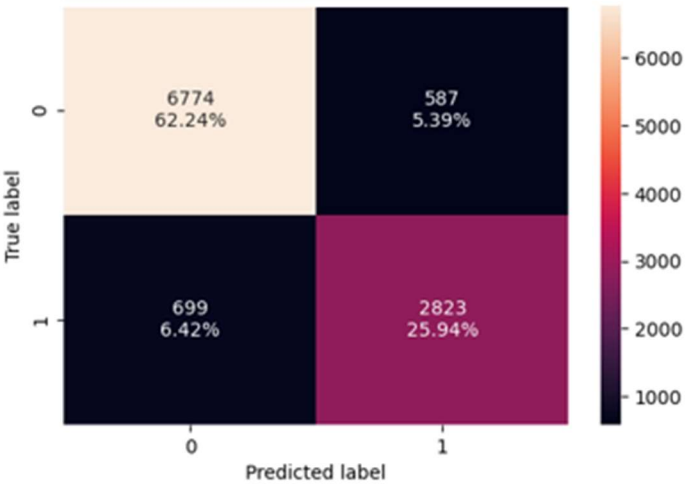


Fig.66 Visual of DT post-pruned

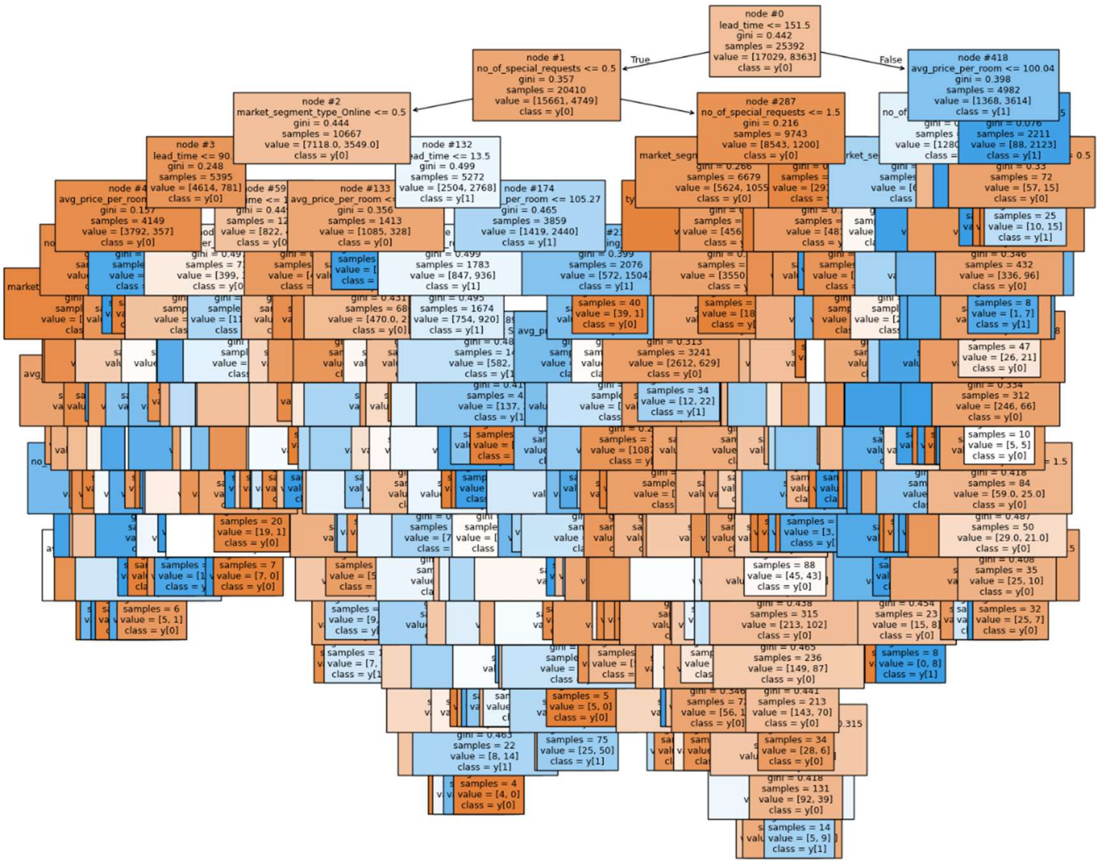
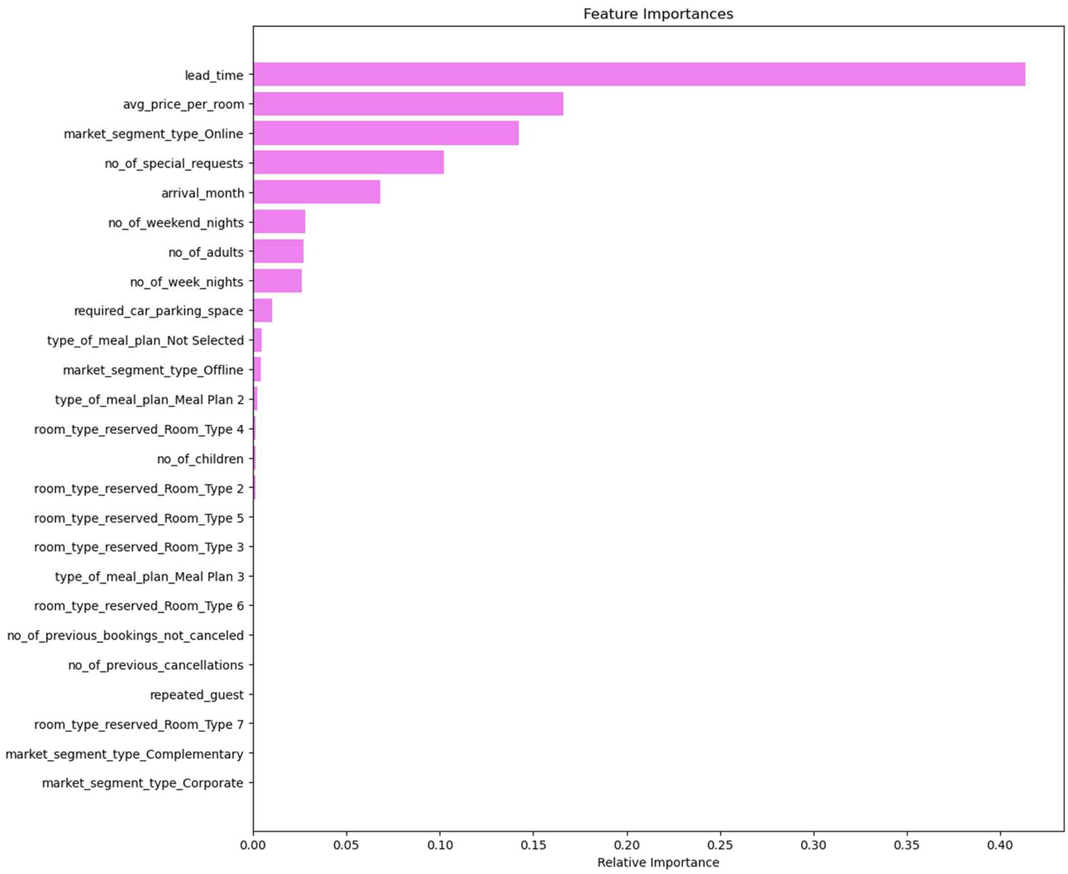


Fig.67 Features of DT post-pruned



COMMENTS:

- The Accuracy and F1 scores are very much improved in the training set and test test.
- This indicates that the model performs well in both sets.
- The important variables for predicting the booking status are also increased in this decision tree.

8. MODEL PERFORMANCE COMPARISON AND FINAL MODEL SELECTION

Training Set Performance:

Fig.68 All model Metrics train

	Accuracy	Recall	Precision	F1
Logis Reg Model 0.5 default Thresold	0.804505	0.630845	0.734911	0.678913
Logis Reg Model -0.31 Threshold ROC-AUC curve	0.775874	0.787595	0.625430	0.697207
Logis Reg Model - 0.42 Threshold Pression-Recall curve	0.801197	0.698521	0.695845	0.697181
KNN Model	0.993463	0.983260	0.996848	0.990007
Naive-bayes model	0.410681	0.964178	0.353550	0.517384
Decision Tree with no parameters	0.993541	0.984694	0.995647	0.990141
Decision Tree with max-depth=4	0.819156	0.635657	0.774814	0.698371
Decision Tree (Pre-Pruning)	0.769494	0.763123	0.622391	0.685610
Decision Tree (Post-Pruning)	0.894376	0.825182	0.849772	0.837297

Training Set Performance:

Fig.69 All model Metrics test

	Accuracy	Recall	Precision	F1
Logis Reg Model 0.5 default Thresold	0.799504	0.621985	0.726737	0.670293
Logis Reg Model -0.31 Threshold ROC-AUC curve	0.770560	0.787156	0.617602	0.692146
Logis Reg Model - 0.42 Threshold Pression-Recall curve	0.796288	0.691812	0.688145	0.689973
KNN Model	0.836718	0.701590	0.772912	0.735526
Naive-bayes model	0.408527	0.962984	0.352603	0.516197
Decision Tree with no parameters	0.864468	0.805508	0.782189	0.793677
Decision Tree with max-depth=4	0.820086	0.631175	0.771339	0.694254
Decision Tree (Pre-Pruning)	0.771938	0.766894	0.619211	0.685185
Decision Tree (Post-Pruning)	0.881834	0.801533	0.827859	0.814484

FINAL MODEL SELECTION:

- Even though the KNN model performs very well in training set, it still underperforms in test set.
- Naive-Bayes model poorly performs in this dataset (both training and testing).
- Decision Trees with no parameters tends to overfit.
- Post-Pruned Decision Tree performs very well in both training and testing sets.
- Post-Pruned Decision Tree has the best accuracy and F1 score on testing set compared to other models.
- Thus, the best Model for Predicting booking status in this dataset is **"Post-Pruned Decision Tree Model"**.

9. ACTIONABLE INSIGHTS & RECOMMENDATIONS

- Providing incentives for parking reservations could potentially reduce cancellations.
- Identifying trends in parking-related bookings may help businesses predict and manage cancellations effectively.
- Businesses may focus on shorter lead-time bookings to reduce cancellation rates.
- The lead time—that is, how far in advance they booked the hotel or rooms—special requests for the stay, and the average room price were the three most significant factors in terms of cancellations.
- Hotel Group should mainly focus on Lead time, Special requests and Average room price.
- Hotel group may provide discounts on higher stay and also provide special treatments to those who ask request.
- Reservation cancellation rates were significantly lower for rooms reserved 151 days (5 months) or less in advance.
- It was quite improbable that anyone who added a specific request would cancel.
- Hotel should ask for a non-refundable deposit paid at least 2-3 months in advance. This will reduce the chances of cancellation.