Machine Learning

BUSINESS REPORT

PGP DSA 3 August 2025

ANJU SAINI

CONTENT

1) 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 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 a reduction of 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, and they are facing problems with the high number of booking cancellations and have reached out to your firm for data-driven solutions. 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 bookings are going to be canceled in advance, and help in formulating profitable policies for cancellations and refunds.

3) Data Description:

The data contains the different attributes of customers' booking details. The detailed data dictionary is given below.

4) 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.

5) Overview Of The Dataset

	Booking_ID	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	type_of_meal_plan	required_car_parking_space	room_type_reserved	lead_time	arrival_year
0	INN00001	2	0	1	2	Meal Plan 1	0	Room_Type 1	224	2017
1	INN00002	2	0	2	3	Not Selected	0	Room_Type 1	5	2018
2	INN00003	1	0	2	1	Meal Plan 1	0	Room_Type 1	1	2018
3	INN00004	2	0	0	2	Meal Plan 1	0	Room_Type 1	211	2018
4	INN00005	2	0	1	1	Not Selected	0	Room_Type 1	48	2018

In the dataset we have booking id , no of children , booking status and other

6) Checking The Shape Of The Dataset

```
(36275, 19)
```

There are 36275 rows and 19 columns in the dataset.

There are 36275 rows and 19 columns in the dataset.

7) Checking The Data Types

```
<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
```

There are total 19 columns and 36275 observations in the dataset

We have 13 continuous variables -

one variable is float type

All other variables are categorical : We can see that there are no missing values in the dataset.

8) Checking Duplicates

```
np.int64(0)
```

We have 10275 duplicate values after removing the booking ID so removing Duplicate Values

9) Checking Null Values

	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

There is no Null Value in the Data set . all values are Unique

10)Statistical Summary

Max No of Adults is 4.0

Max no of children is 10.0

	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

Avg price of pre room max is 540

11) Checking Unique Value

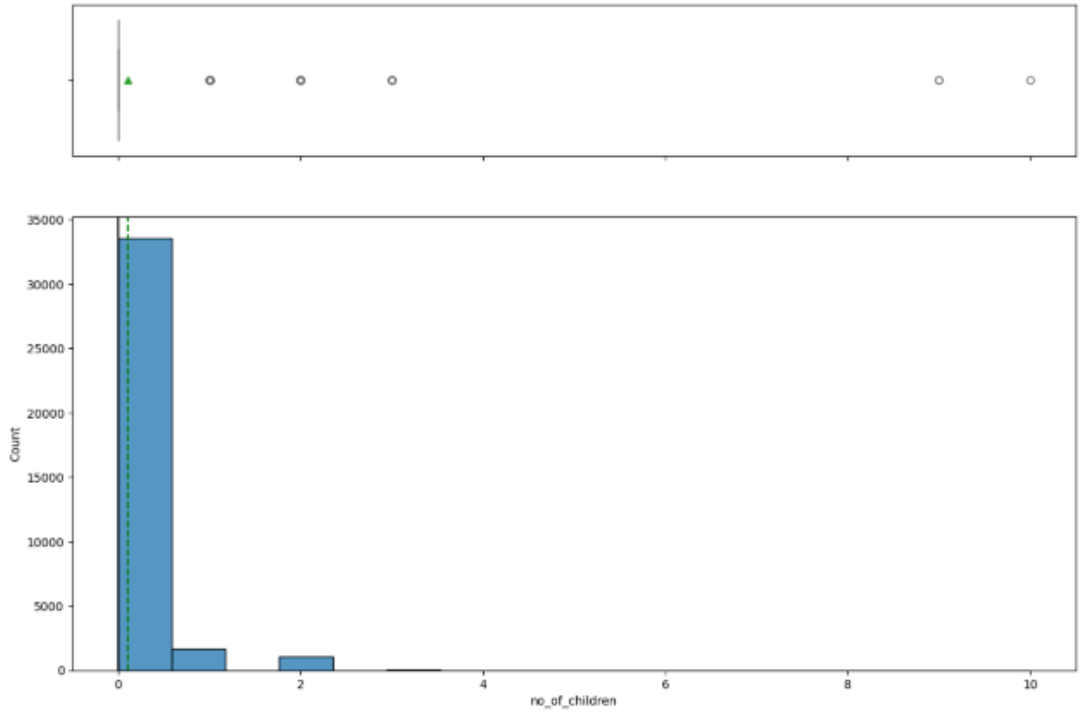
```
no_of_adults
2    26108
1     7695
3     2317
0      139
4       16
Name: count, dtype: int64
----------------------------------------
no_of_children
0     33577
1      1618
2      1058
3        19
9         2
10        1
Name: count, dtype: int64
----------------------------------------
no_of_weekend_nights
0    16872
1     9995
2     9071
3      153
4      129
5       34
6       20
7        1
Name: count, dtype: int64
----------------------------------------
```

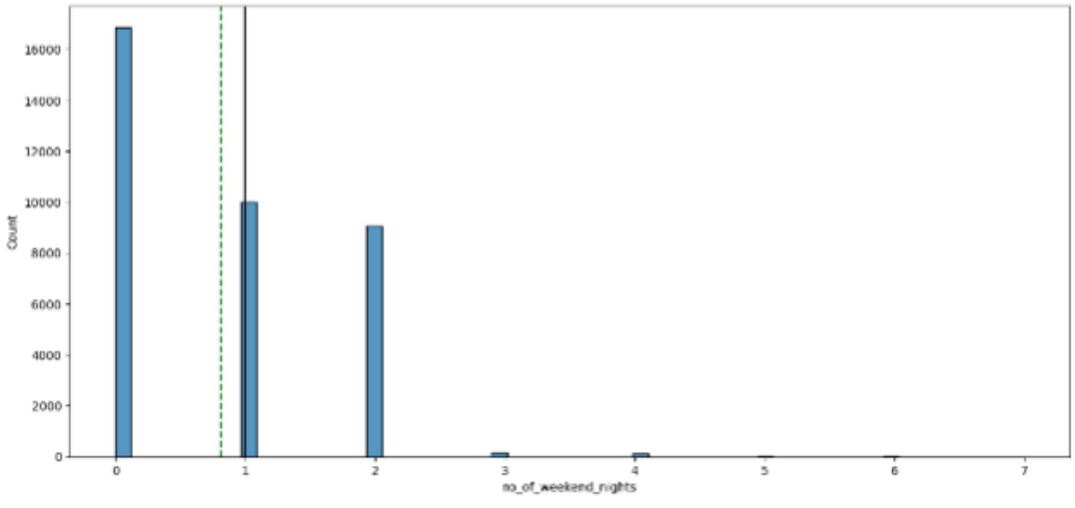
```
Name: count, Length: 3930, dtype: int64
----------------------------------------
no_of_special_requests
0    19777
1    11373
2     4364
3      675
4       78
5        8
Name: count, dtype: int64
----------------------------------------
booking_status
Not_Canceled    24390
Canceled        11885
Name: count, dtype: int64
----------------------------------------

Missing values per column:
```

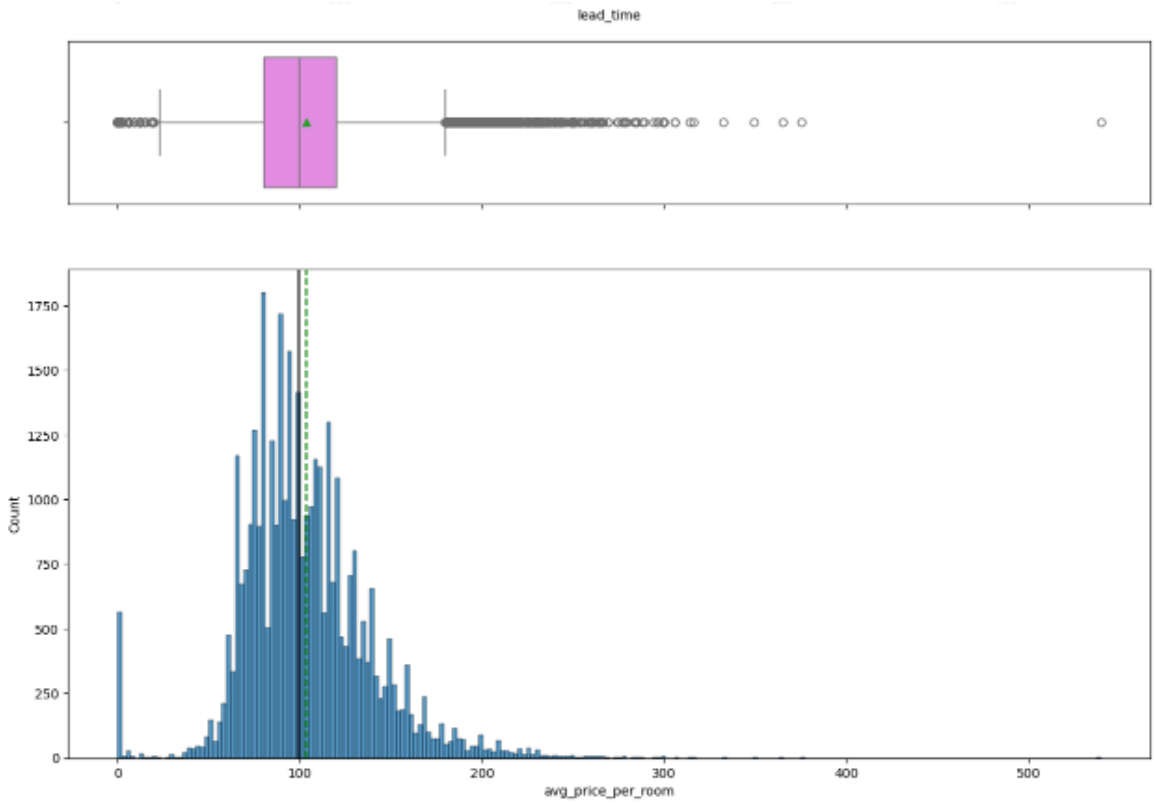
Not Canceled Count is More than Cancelled

There Are more People who would like to have some special request

12)Univariate analysis







Observations from univariate analysis of numerical variables:
no_of_adults:
The majority of bookings are for 2 adults, with a smaller number for 1 or 3 adults.
There are some outliers with 0 or 4 adults.
no_of_children:
Most bookings have no children.
There are outliers with a higher number of children.
no_of_weekend_nights:
The distribution is skewed towards fewer weekend nights.
The majority of bookings have 0, 1, or 2 weekend nights.
no_of_week_nights:
The distribution is skewed towards fewer weeknights.
The majority of bookings have between 1 and 3 weeknights.
lead_time:

- The distribution is highly skewed to the right, indicating that most bookings are made with a short lead time.

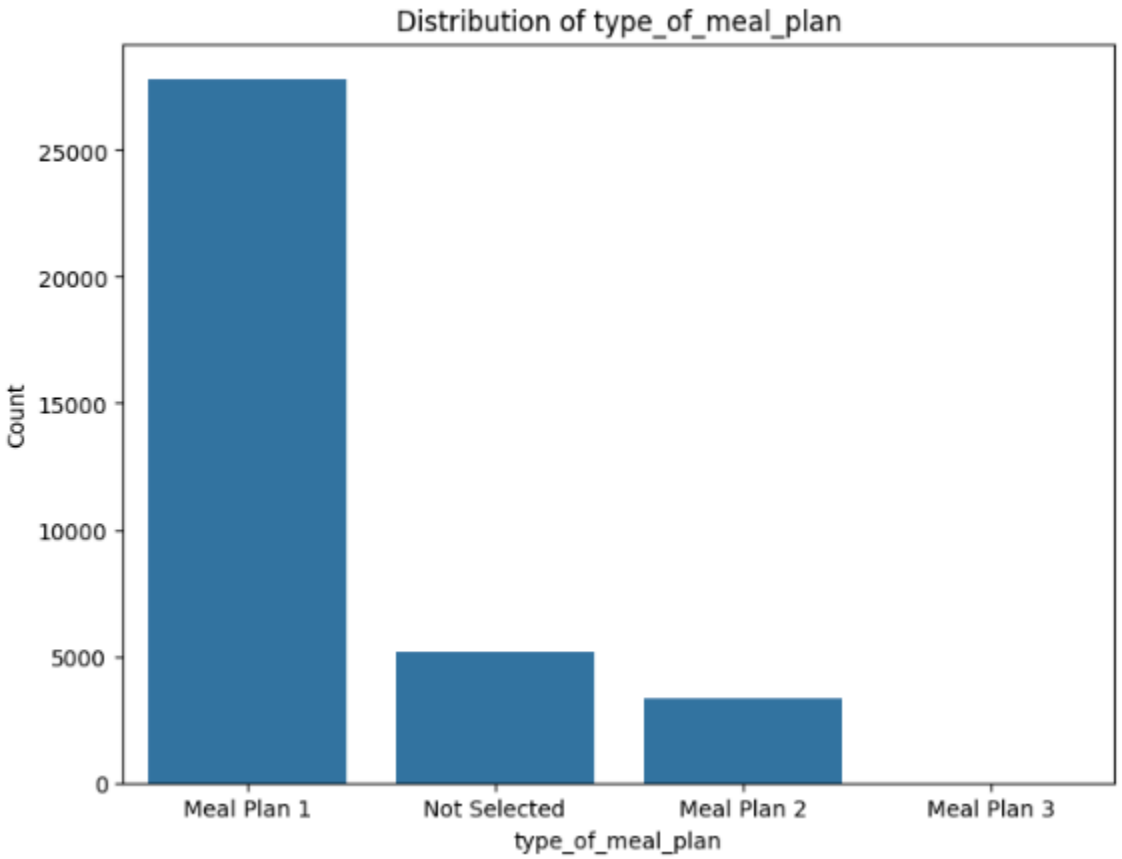
- There are many outliers with very long lead times.

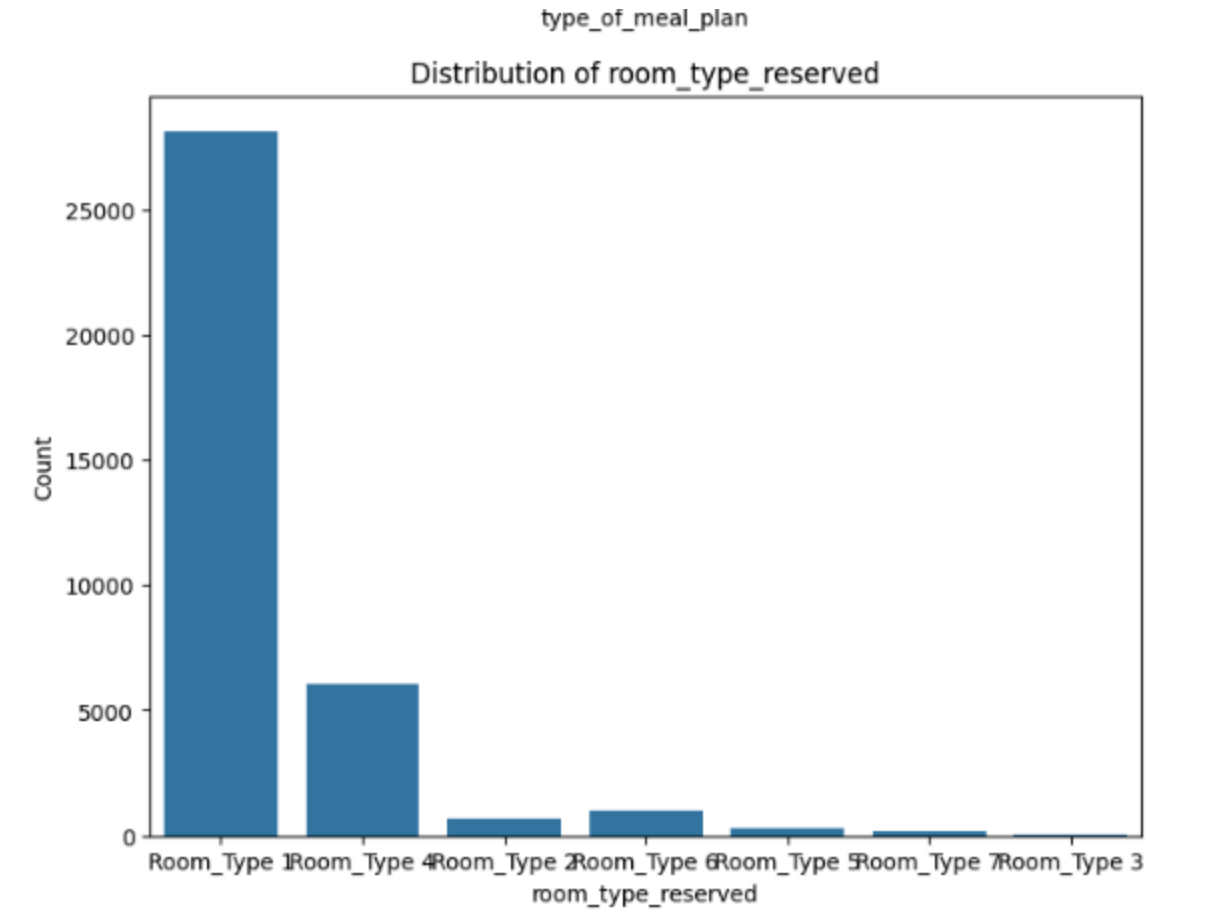
avg_price_per_room:

- The distribution is skewed to the right with a peak around the mean.
- There are outliers with very high average prices per room.

no_of_special_requests:

- The majority of bookings have no special requests.
- The distribution is skewed to the right, with fewer bookings having multiple special requests.





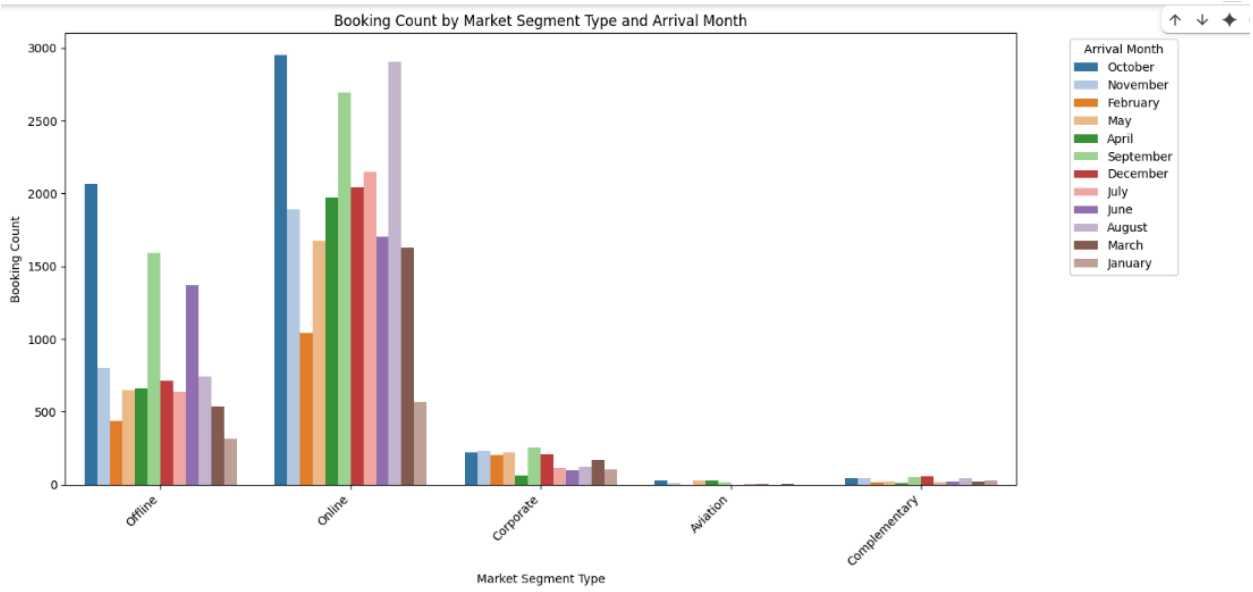
type_of_meal_plan:

- 'Meal Plan 1' is the most common meal plan booked by customers, followed by 'Not Selected'.
- 'Meal Plan 2' and 'Meal Plan 3' are significantly less frequent.

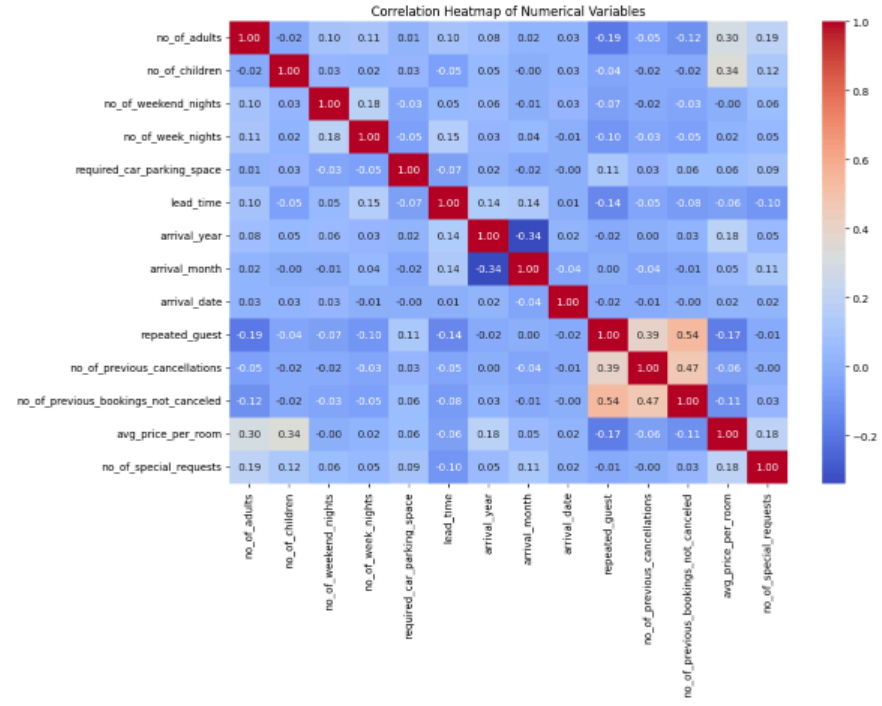
room_type_reserved:

- 'Room_Type 1' is by far the most frequently reserved room type.
- Other room types ('Room_Type 4', 'Room_Type 6', 'Room_Type 2', 'Room_Type 5', 'Room_Type 7', 'Room_Type 3') are reserved much less often.

13) Bivariate Analysis



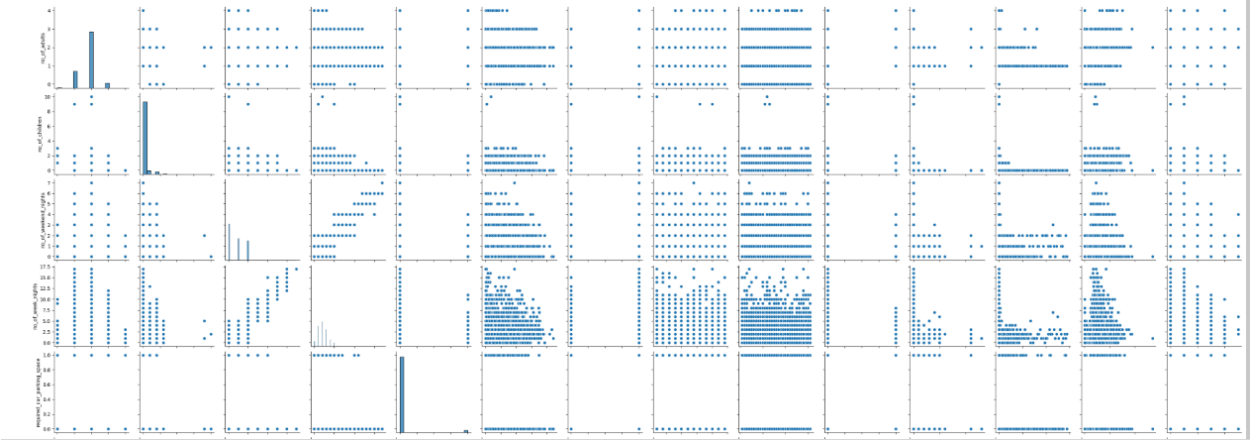
Most of booking online done on August Month while when see ofline most booking is on October Month



Lead Time and Average Price per Room: There appears to be a weak positive correlation between lead_time and avg_price_per_room. This suggests that bookings made further in advance might have slightly higher average prices, but the relationship is not very strong.

Number of Previous Bookings Not Canceled and Repeated Guest: There is a strong positive correlation between no_of_previous_bookings_not_canceled and repeated_guest. This is expected, as a repeated guest is likely to have previous bookings that were not canceled.

14) Pairplot of Numerical Variables

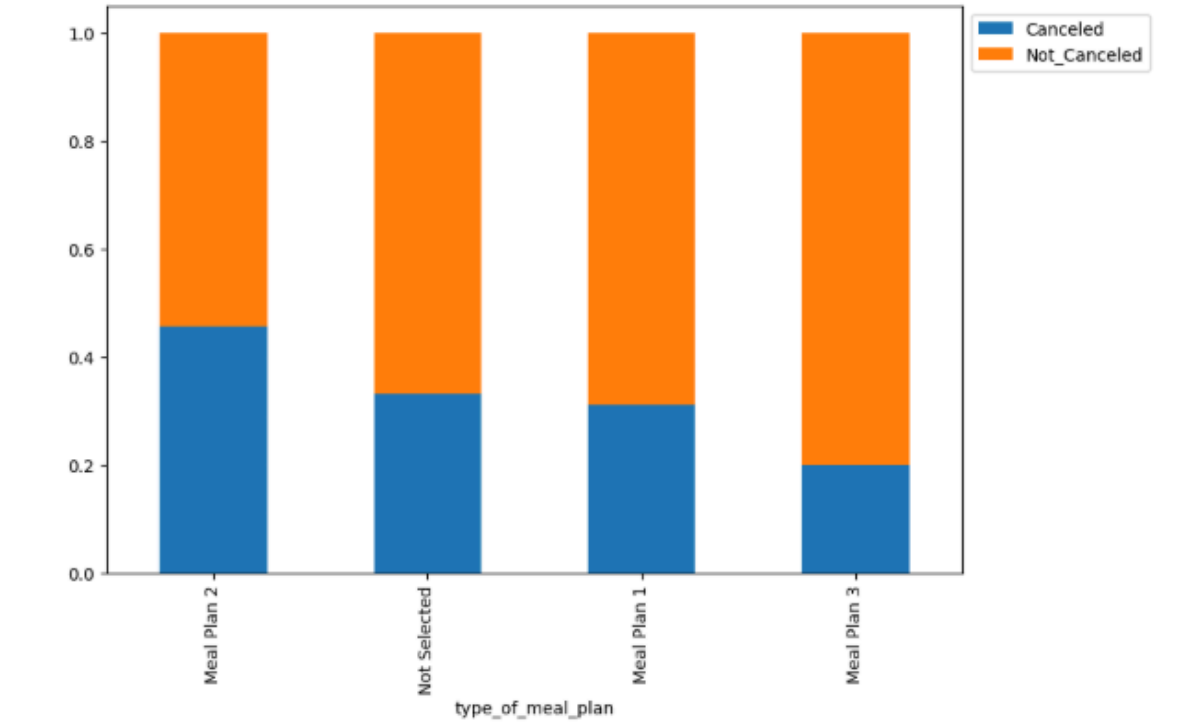


Number of Weekend Nights and Number of Week Nights: There is a weak positive correlation between no_of_weekend_nights and no_of_week_nights, suggesting that longer stays (more weeknights) might also include more weekend nights, but they are not strongly dependent on each other.

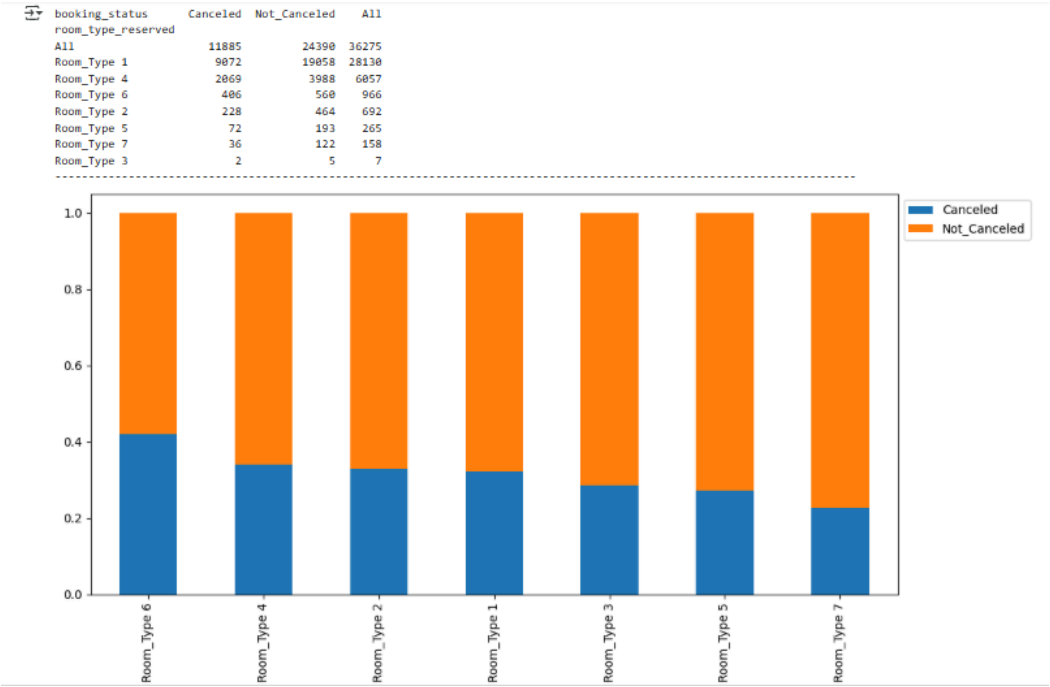
No Strong Linear Correlations Among Most Features: The heatmap generally shows relatively weak linear correlations between most pairs of numerical variables (values closer to 0). This indicates that many of the numerical features are not strongly linearly related to each other.

Distribution of Numerical Features: The pairplot's diagonal histograms confirm the observations from the univariate analysis regarding the skewed distributions of variables like lead_time, avg_price_per_room, no_of_children, no_of_previous_cancellations, no_of_previous_bookings_not_canceled, and no_of_special_requests. This skewness might need to be addressed during feature engineering for certain models.

booking_status	Canceled	Not_Canceled	All
type_of_meal_plan			
All	11885	24390	36275
Meal Plan 1	8679	19156	27835
Not Selected	1699	3431	5130
Meal Plan 2	1506	1799	3305
Meal Plan 3	1	4	5



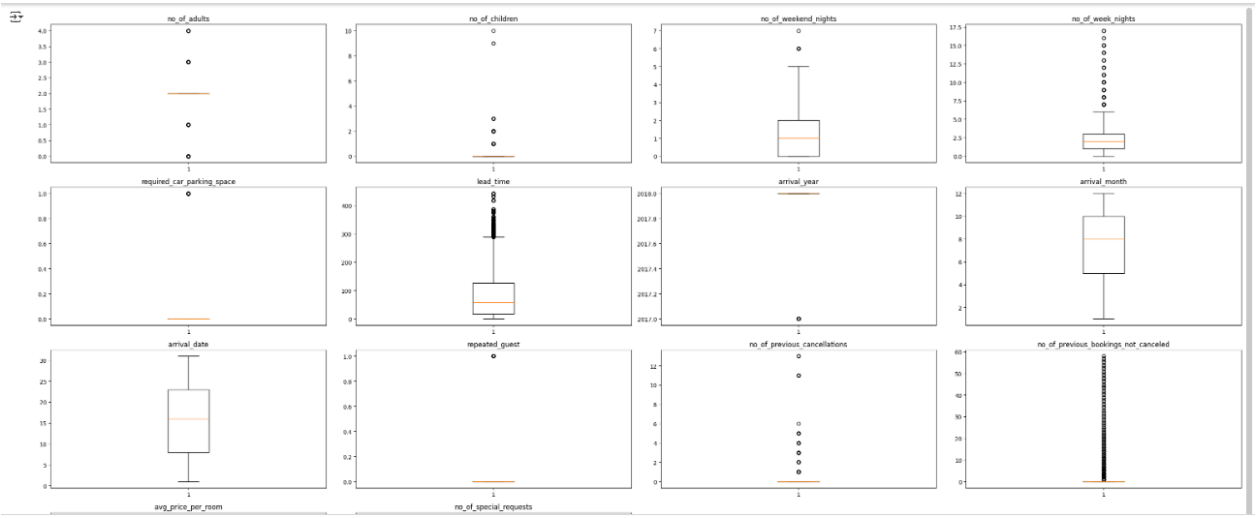
Not Cancelled is high with Meal dight plan 3
Cancelled ratio more with dight plane 2



booking_status	Canceled	Not_Canceled	All
room_type_reserved			
All	11885	24390	36275
Room_Type 1	9072	19058	28130
Room_Type 4	2069	3988	6057
Room_Type 6	406	560	966
Room_Type 2	228	464	692
Room_Type 5	72	193	265
Room_Type 7	36	122	158
Room_Type 3	2	5	7

Room Type Has Not More impact on cancel of a booking

15)Checking The Outliers



Although outliers Exist , we Will keep them as they may have a valuable input

16)Data Preparation

```
Shape of Training set :  (25392, 37)
Shape of test set :  (10883, 37)
Percentage of classes in training set:
booking_status
Not_Canceled    0.670644
Canceled        0.329356
Name: proportion, dtype: float64
Percentage of classes in test set:
booking_status
Not_Canceled    0.676376
Canceled        0.323624
Name: proportion, dtype: float64
```

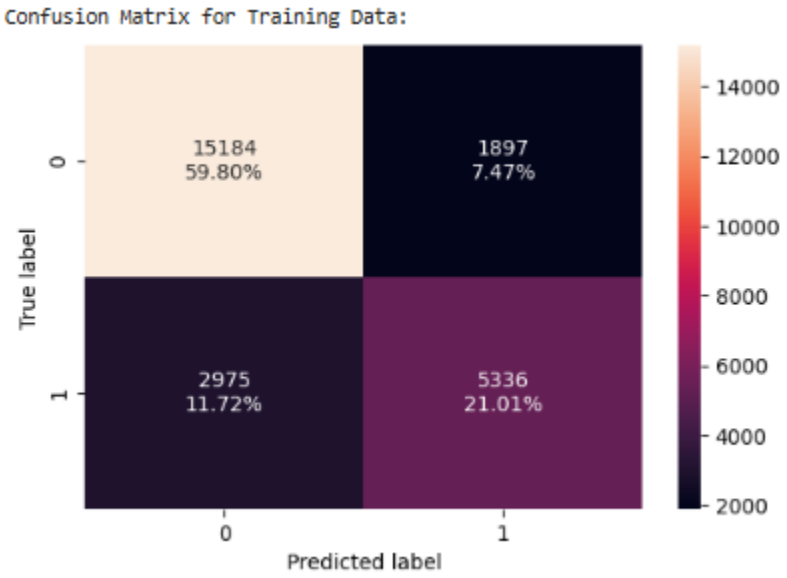
User 70,30 percentage to divide data into test and training , after divide data we have records like for training 25392 rows and 37 columns and for test 10883 and 37 columns

17)Building The Model

```
                           Logit Regression Results
==============================================================================
Dep. Variable:         booking_status   No. Observations:                25392
Model:                          Logit   Df Residuals:                    25354
Method:                           MLE   Df Model:                           37
Date:                Sun, 03 Aug 2025   Pseudo R-squ.:                 0.3511
Time:                        06:21:06   Log-Likelihood:                -10418.
converged:                      False   LL-Null:                       -16054.
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                         -901.1400        nan        nan        nan         nan         nan
no_of_adults                     0.1513      0.039      3.850      0.000       0.074       0.228
no_of_children                   0.2257      0.061      3.681      0.000       0.106       0.346
no_of_weekend_nights             0.1559      0.020      7.647      0.000       0.116       0.196
no_of_week_nights                0.0385      0.013      3.023      0.003       0.014       0.063
required_car_parking_space      -1.5290      0.143    -10.714      0.000      -1.809      -1.249
lead_time                        0.0161      0.000     55.922      0.000       0.016       0.017
arrival_year                     0.4778      0.064      7.506      0.000       0.353       0.603
arrival_month                  -16.5008        nan        nan        nan         nan         nan
arrival_date                     0.0037      0.002      1.867      0.062      -0.000       0.008
repeated_guest                  -1.9321      0.485     -3.982      0.000      -2.883      -0.981
no_of_previous_cancellations     0.3400      0.105      3.234      0.001       0.134       0.546
no_of_previous_bookings_not_canceled   -0.0930      0.080     -1.157      0.247      -0.251       0.065
avg_price_per_room               0.0190      0.001     22.602      0.000       0.017       0.021
no_of_special_requests          -1.5452      0.031    -49.488      0.000      -1.606      -1.484
type_of_meal_plan_Meal Plan 2    0.1475      0.069      2.142      0.032       0.013       0.282
type_of_meal_plan_Meal Plan 3   13.9916    711.371      0.020      0.984   -1380.269    1408.252
type_of_meal_plan_Not Selected   0.2064      0.054      3.806      0.000       0.100       0.313
room_type_reserved_Room_Type 2  -0.2776      0.137     -2.031      0.042      -0.546      -0.010
room_type_reserved_Room_Type 3  -0.4647      1.330     -0.349      0.727      -3.072       2.143
room_type_reserved_Room_Type 4  -0.2866      0.054     -5.300      0.000      -0.393      -0.181
room_type_reserved_Room_Type 5  -0.9994      0.213     -4.682      0.000      -1.418      -0.581
room_type_reserved_Room_Type 6  -1.0070      0.155     -6.485      0.000      -1.311      -0.703
room_type_reserved_Room_Type 7  -1.5729      0.323     -4.869      0.000      -2.206      -0.940
```

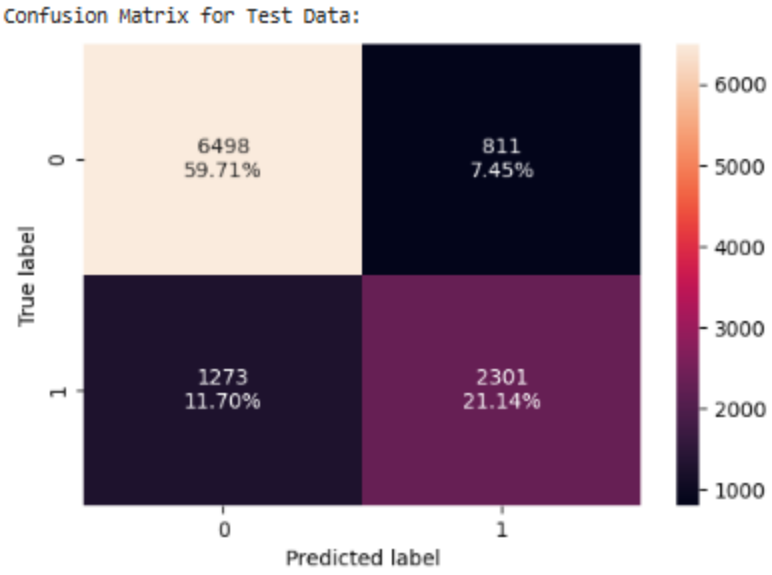
This is the first model with no column deleted in this we get sudo r square value around 0.3511

18) Confusion Matrix



Test Data Confusion Matrix

- **True Negatives (TN): 6,498 (59.71%)**
    - The model correctly predicted class `0` when the true label was `0`.
- **False Positives (FP): 811 (7.45%)**
    - The model predicted `1` incorrectly when the true label was `0`.
- **False Negatives (FN): 1,273 (11.70%)**
    - The model predicted `0` incorrectly when the true label was `1`.
- **True Positives (TP): 2,301 (21.14%)**



2. Training Data Confusion Matrix

- True Negatives (TN): 15,184 (59.80%)
- False Positives (FP): 1,897 (7.47%)
- False Negatives (FN): 2,975 (11.72%)
- True Positives (TP): 5,336 (21.01%)

Model Performance on Training Data with threshold = 0.4:

	Accuracy	Recall	Precision	F1
0	0.794423	0.70894	0.677787	0.693013

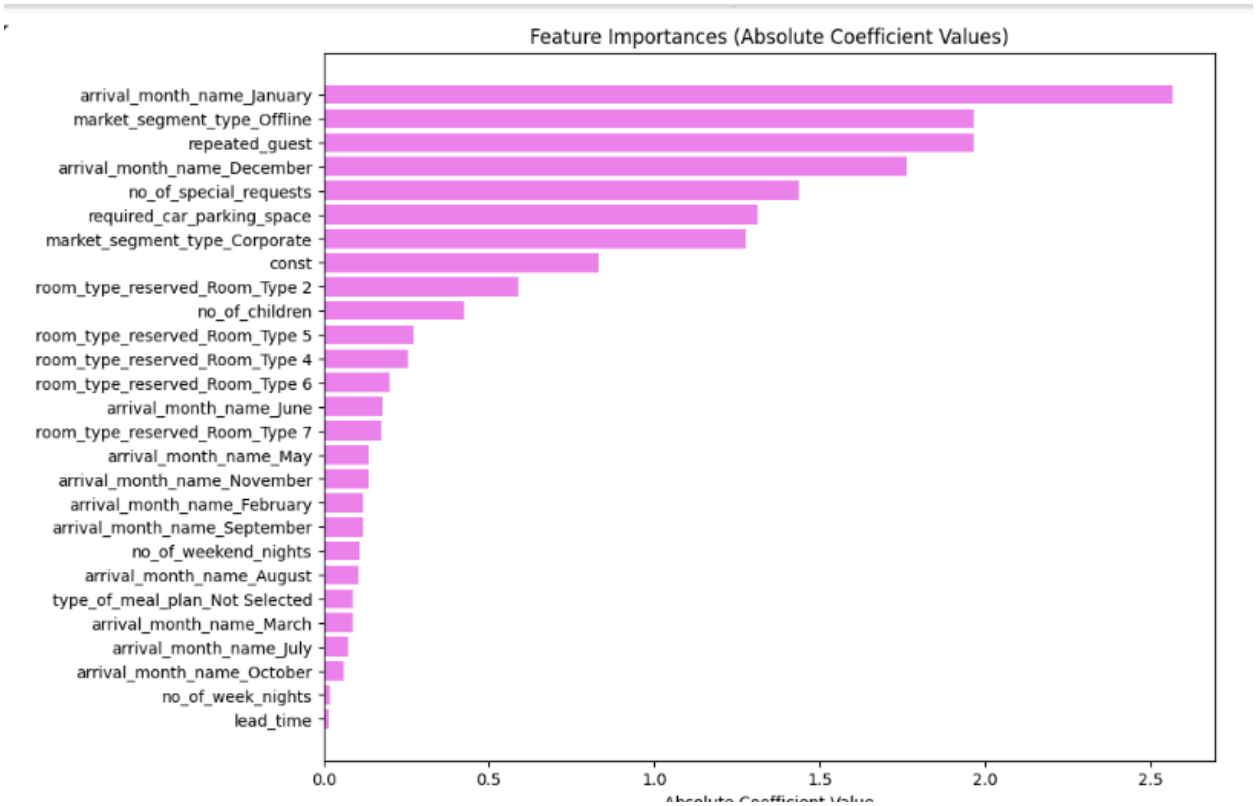
Model Performance on Test Data with threshold = 0.4:

	Accuracy	Recall	Precision	F1
0	0.797299	0.717124	0.682012	0.699127

Balanced Performance:

- Training and test scores are very close → **the model is not overfitting**.
  Threshold = 0.4:
- Lowering the threshold from the default 0.5 likely **improved recall** at the cost of some precision.
- This is useful if **missing positives (false negatives) is more costly**.

19)Feature Importances (Absolute Coefficient Value)



The model relies heavily on arrival month, market segment, and guest behaviors to make predictions, while some time-related or meal-type features contribute very little.

20)Final Logist Model

```
                           Logit Regression Results
==============================================================================
Dep. Variable:         booking_status   No. Observations:                25392
Model:                          Logit   Df Residuals:                    25369
Method:                           MLE   Df Model:                           22
Date:                Sun, 03 Aug 2025   Pseudo R-squ.:                 0.3189
Time:                        06:21:38   Log-Likelihood:                -10934.
converged:                       True   LL-Null:                       -16054.
Covariance Type:            nonrobust   LLR p-value:                     0.000
=================================================================================================
                                    coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                            -0.8114      0.046    -17.507      0.000      -0.902      -0.721
no_of_children                    0.4145      0.059      6.997      0.000       0.298       0.531
no_of_weekend_nights              0.1068      0.020      5.428      0.000       0.068       0.145
no_of_week_nights                 0.0197      0.012      1.604      0.109      -0.004       0.044
required_car_parking_space       -1.3118      0.141     -9.324      0.000      -1.588      -1.036
lead_time                         0.0153      0.000     61.486      0.000       0.015       0.016
repeated_guest                   -1.9751      0.320     -6.172      0.000      -2.602      -1.348
no_of_special_requests           -1.4431      0.030    -48.615      0.000      -1.501      -1.385
room_type_reserved_Room_Type 2   -0.5947      0.133     -4.475      0.000      -0.855      -0.334
room_type_reserved_Room_Type 4    0.2291      0.046      4.985      0.000       0.139       0.319
room_type_reserved_Room_Type 5   -0.2892      0.208     -1.391      0.164      -0.697       0.118
room_type_reserved_Room_Type 6    0.1922      0.142      1.356      0.175      -0.086       0.470
room_type_reserved_Room_Type 7    0.1473      0.285      0.517      0.605      -0.411       0.705
market_segment_type_Corporate    -1.3096      0.101    -13.014      0.000      -1.507      -1.112
market_segment_type_Offline      -1.9919      0.047    -42.188      0.000      -2.084      -1.899
arrival_month_name_August        -0.0816      0.058     -1.416      0.157      -0.195       0.031
arrival_month_name_December      -1.7347      0.087    -19.840      0.000      -1.906      -1.563
arrival_month_name_February       0.1358      0.079      1.724      0.085      -0.019       0.290
arrival_month_name_January       -2.5513      0.250    -10.210      0.000      -3.041      -2.062
arrival_month_name_June           0.2019      0.059      3.434      0.001       0.087       0.317
arrival_month_name_May           -0.1151      0.065     -1.777      0.075      -0.242       0.012
arrival_month_name_November       0.1597      0.066      2.430      0.015       0.031       0.288
arrival_month_name_September     -0.0959      0.056     -1.720      0.086      -0.205       0.013
=================================================================================================
```

Model Performance on Training Data (Reduced Model):

	Accuracy	Recall	Precision	F1
0	0.793163	0.705691	0.676393	0.690731

Model Performance on Test Data (Reduced Model):

	Accuracy	Recall	Precision	F1
0	0.796839	0.715445	0.681685	0.698157

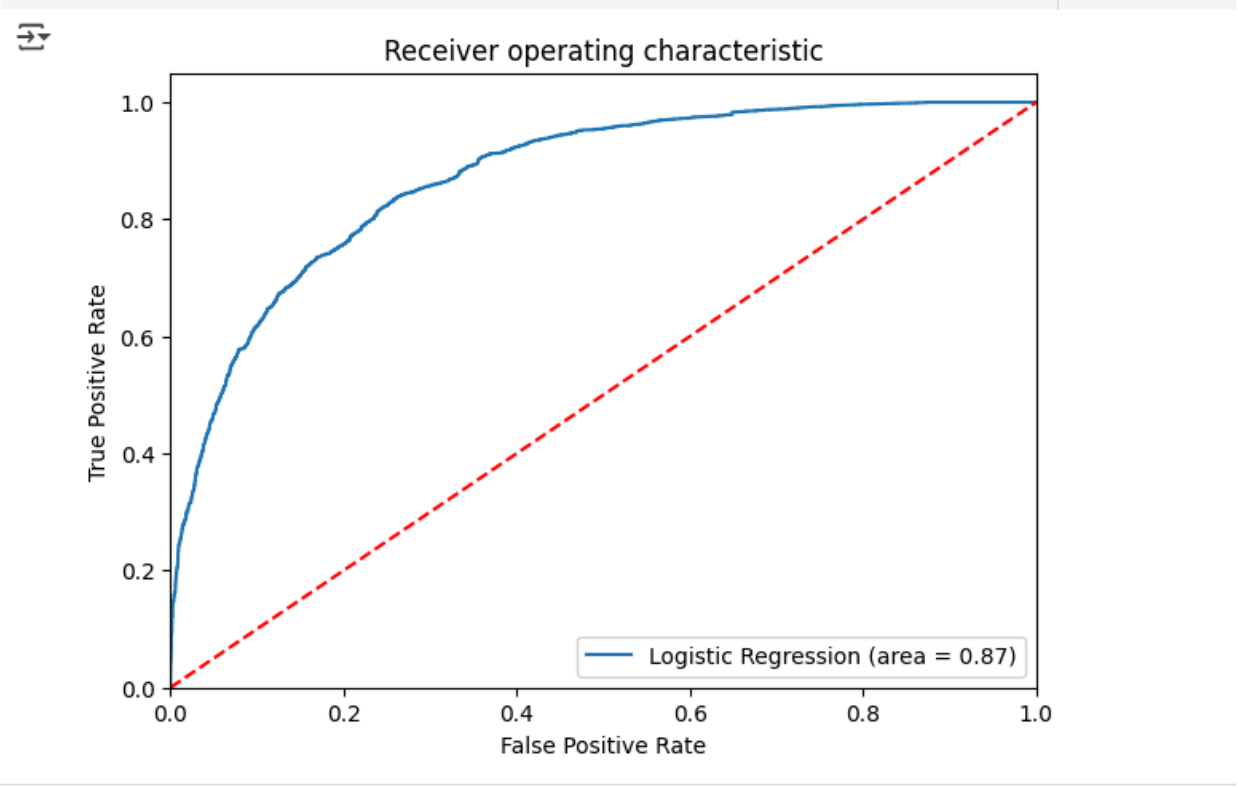
1. Training Data Performance (Reduced Model)

- Accuracy: 0.7931 (~79.31%)
- Recall: 0.7057 (~70.57%)
- Precision: 0.6764 (~67.64%)
- F1-score: 0.6907 (~69.07%)

2. Test Data Performance (Reduced Model)

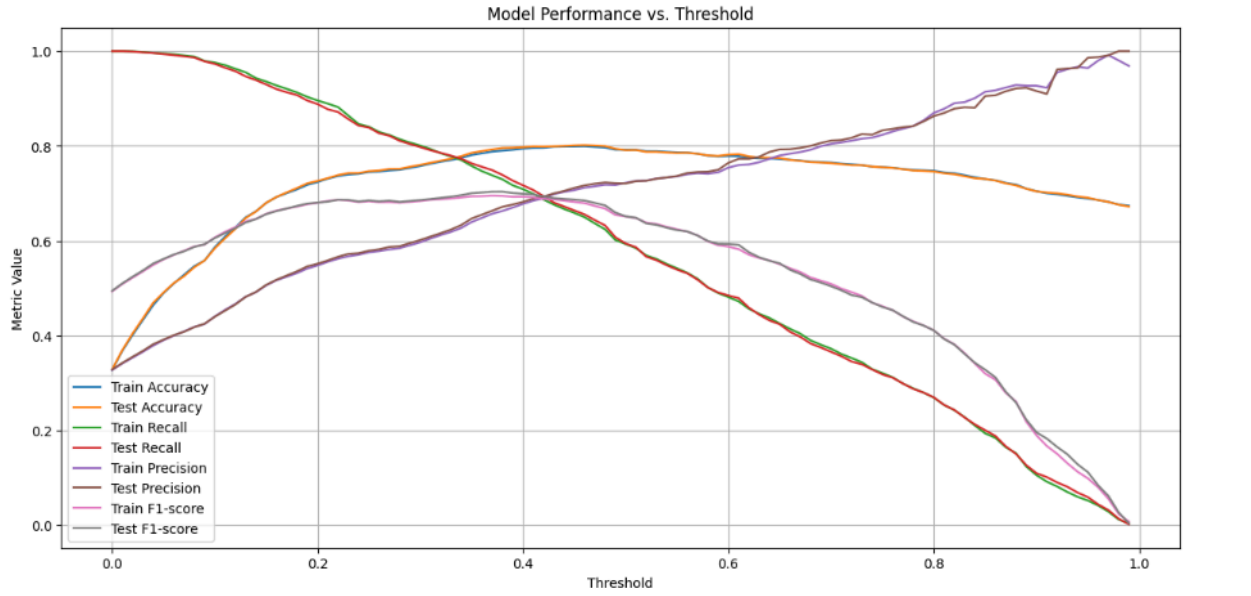
- Accuracy: 0.7968 (~79.68%)
- Recall: 0.7154 (~71.54%)
- Precision: 0.6817 (~68.17%)
- F1-score: 0.6982 (~69.82%)

21)Receiver operating characteristic



This ROC (Receiver Operating Characteristic) curve shows the performance of a logistic regression model for binary classification. The area under the curve (AUC) is 0.87, which indicates that the model has good discriminative ability to distinguish between the two classes. The closer the AUC is to 1, the better the model is at predicting 0s as 0 and 1s as 1. An AUC of 0.87 means the model is performing well and is much better than random guessing (the red dashed line, AUC = 0.5).

22)Model Performance Vs Threshold



Now, based on the plot above, we can choose a threshold that balances the trade-off between different metrics (Accuracy, Recall, Precision, F1-score) according to the specific goals of the problem. A lower threshold will increase recall but decrease precision, and vice-versa. To reduce both False Positives (0,1) and False Negatives (1,0) in the confusion matrix, we should look for a threshold where precision and recall are reasonably balanced, or prioritize the metric that is most important for the business problem (e.g., if minimizing lost revenue from cancellations is critical, we might prioritize recall).

23) Model Building - Decision Tree Model

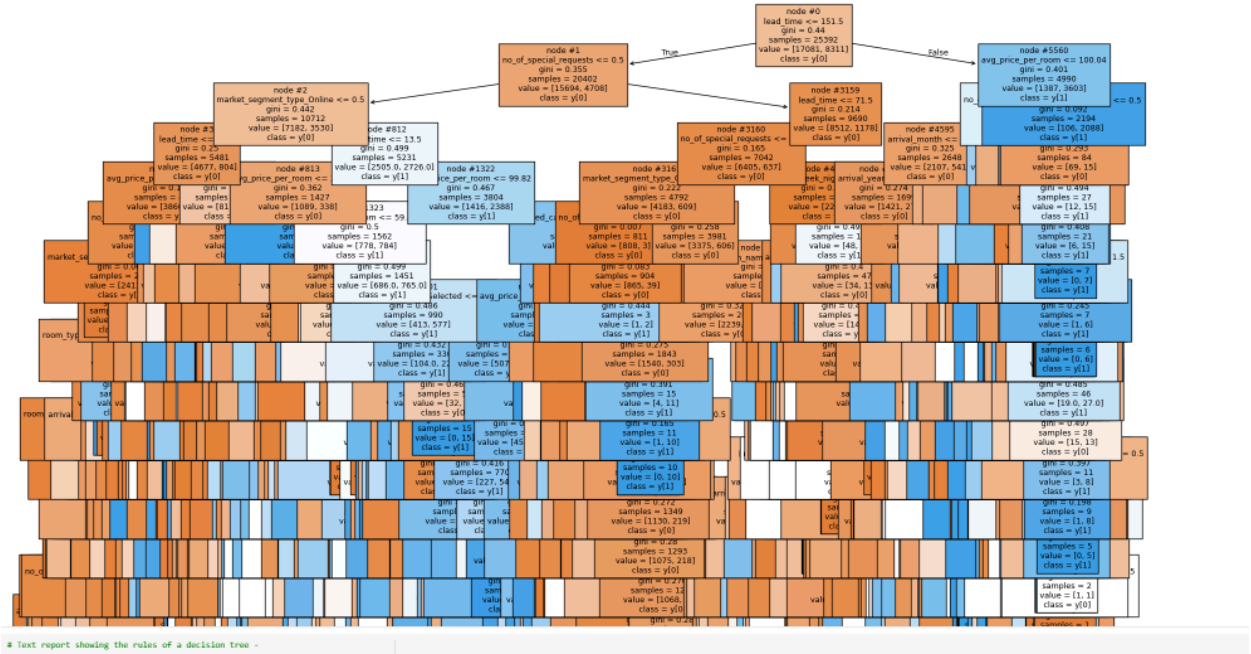
```
DecisionTreeClassifier
DecisionTreeClassifier(random_state=1)
```

24)Checking model performance on test set

	Accuracy	Recall	Precision	F1
0	0.868143	0.797426	0.800337	0.798879

This table shows the performance metrics of the model on the test set. The model achieved an accuracy of 0.87, recall of 0.80, precision of 0.80, and an F1 score of 0.80. These values indicate that the model performs well, with a good balance between precision and recall.

25)Visualizing the Decision Tree



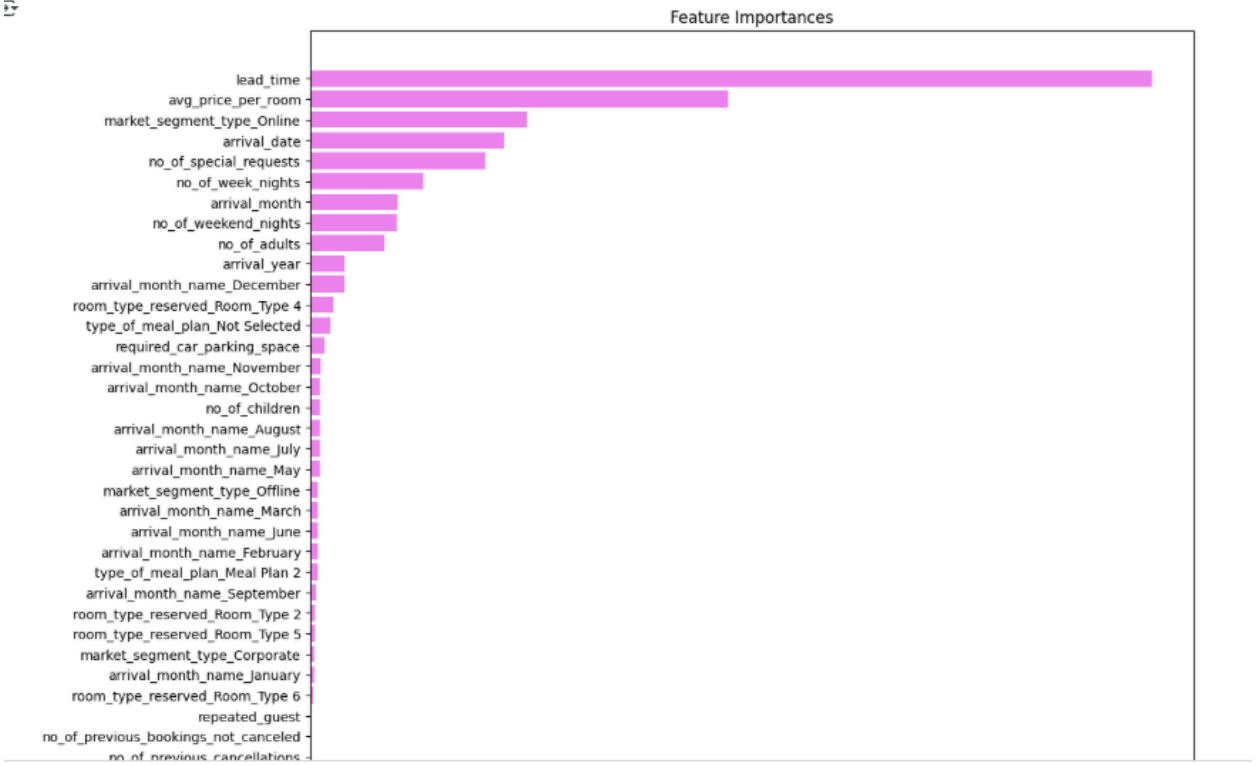
This image displays a graphical visualization of the trained decision tree model. Each node represents a decision based on a feature and threshold, and the branches show the possible outcomes. The colors indicate different classes, helping to understand how the model makes predictions.

26)Text Report Showing of A decision Tree

```
|--- lead_time <= 151.50
|   |--- no_of_special_requests <= 0.50
|   |   |--- market_segment_type_Online <= 0.50
|   |   |   |--- lead_time <= 90.50
|   |   |   |   |--- avg_price_per_room <= 201.50
|   |   |   |   |   |--- no_of_weekend_nights <= 0.50
|   |   |   |   |   |   |--- market_segment_type_Offline <= 0.50
|   |   |   |   |   |   |   |--- no_of_adults <= 1.50
|   |   |   |   |   |   |   |   |--- avg_price_per_room <= 162.53
|   |   |   |   |   |   |   |   |   |--- lead_time <= 19.50
|   |   |   |   |   |   |   |   |   |   |--- room_type_reserved_Room_Type 4 <= 0.50
|   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 13
|   |   |   |   |   |   |   |   |   |   |--- room_type_reserved_Room_Type 4 >  0.50
|   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 7
|   |   |   |   |   |   |   |   |   |--- lead_time >  19.50
|   |   |   |   |   |   |   |   |   |   |--- arrival_date <= 18.50
|   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 9
|   |   |   |   |   |   |   |   |   |   |--- arrival_date >  18.50
|   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 5
|   |   |   |   |   |   |   |   |--- avg_price_per_room >  162.53
|   |   |   |   |   |   |   |   |   |--- lead_time <= 2.50
|   |   |   |   |   |   |   |   |   |   |--- weights: [0.00, 2.00] class: 1
|   |   |   |   |   |   |   |   |   |--- lead_time >  2.50
|   |   |   |   |   |   |   |   |   |   |--- arrival_month_name_October <= 0.50
|   |   |   |   |   |   |   |   |   |   |   |--- weights: [3.00, 0.00] class: 0
|   |   |   |   |   |   |   |   |   |   |--- arrival_month_name_October >  0.50
|   |   |   |   |   |   |   |   |   |   |   |--- weights: [0.00, 1.00] class: 1
```

This is a text-based representation of the decision tree structure. It shows the sequence of splits based on different features and their thresholds, leading to the final classification at each leaf node. This format helps to interpret the logic and rules used by the decision tree for making predictions.

27)Feature Importance



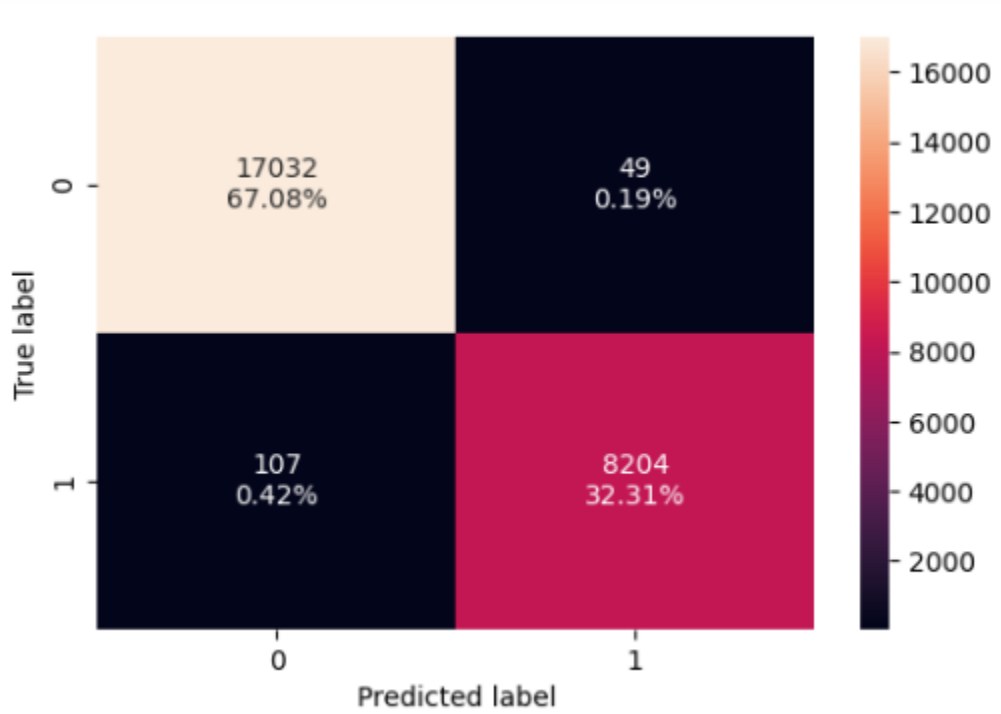
lead time , avg price room , market segment type online , arrival data, no of specific request ,no of week night , no of weekend nights are important factors

28)Decision Tree (Pre-pruning)

```
DecisionTreeClassifier(criterion='entropy', min_impurity_decrease=1e-05,
                       random_state=1)
```

This shows the initialization of a Decision Tree Classifier with pre-pruning parameters. The model uses the 'entropy' criterion and a minimum impurity decrease to control tree growth, helping to prevent overfitting by stopping splits that do not improve purity enough.

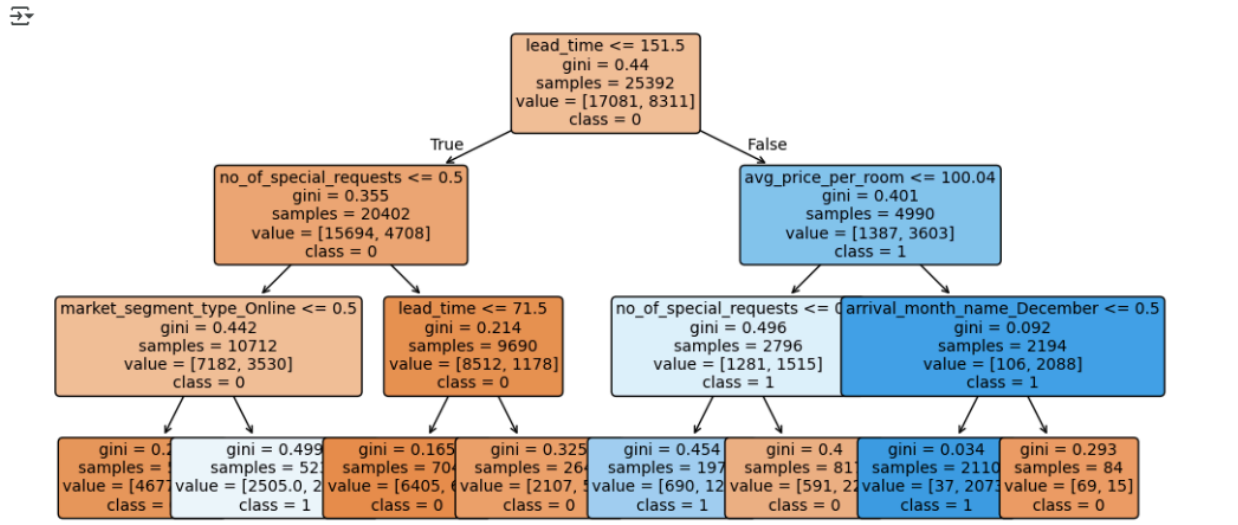
29)Checking model performance on training set



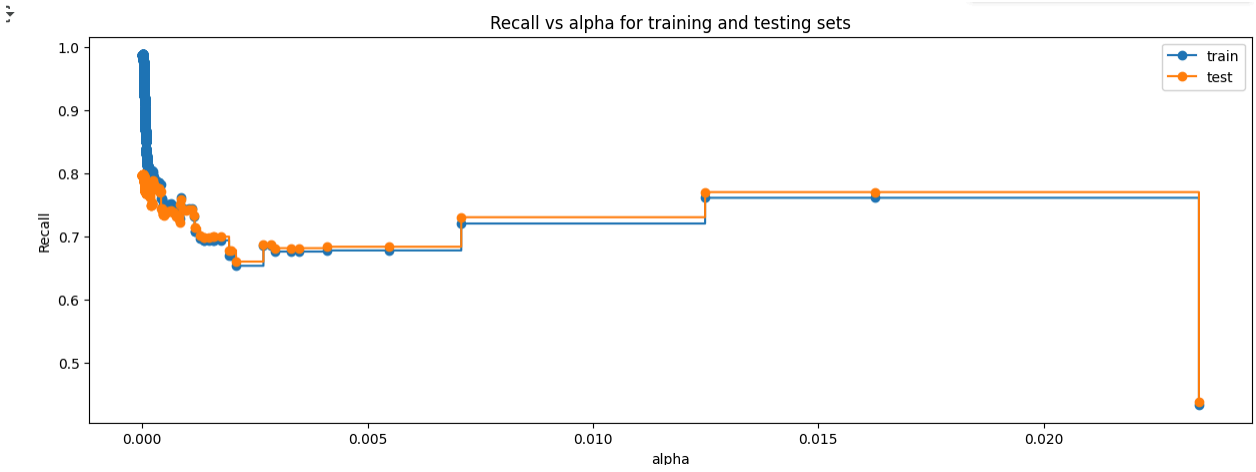
This confusion matrix visualizes the performance of a classification model. The majority of predictions are correct, with high true positives and true negatives, and very few false positives and false negatives, indicating strong model performance.

30) Visualizing the Decision Tree

This is a graphical representation of the trained decision tree after pre-pruning. Each node shows the feature, threshold, gini impurity, sample count, and class distribution, making it easy to interpret the model's decision-making process.

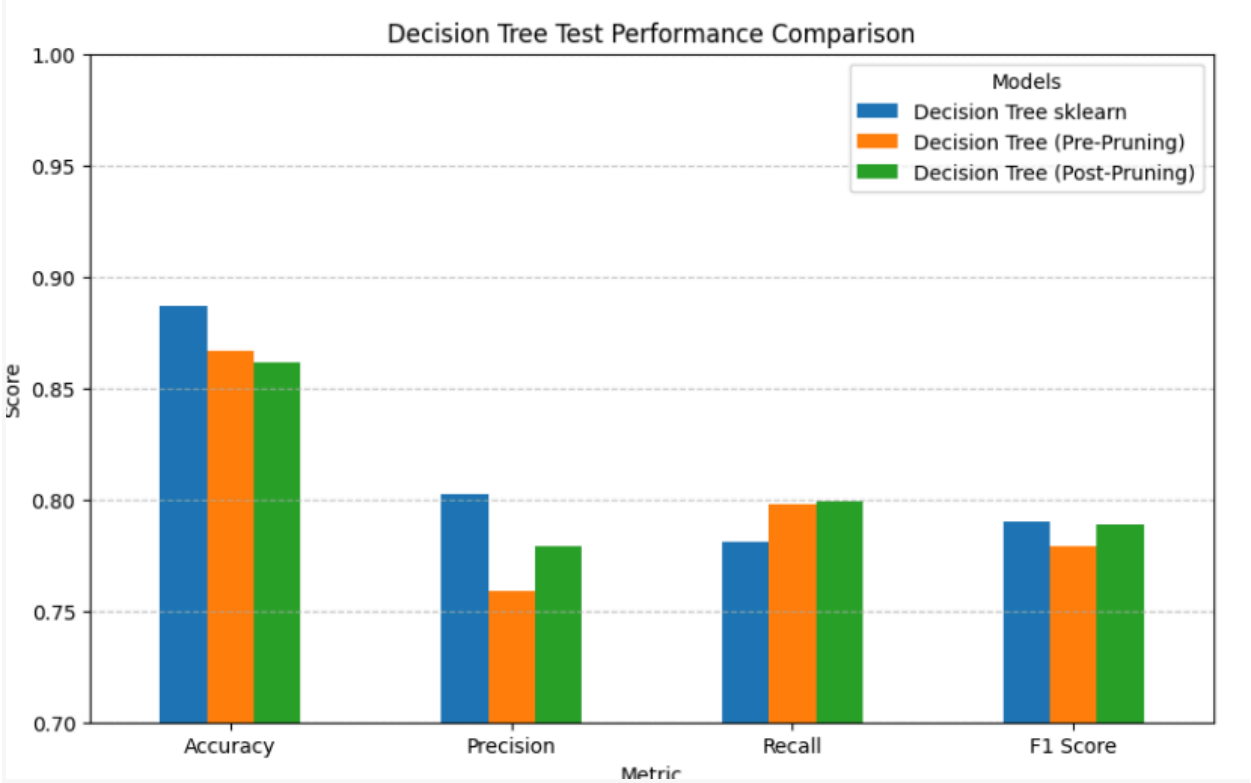


31)Recall vs alpha for training and testing sets



This line plot shows how recall changes for both training and testing sets as the alpha parameter (regularization strength) varies. It helps identify the optimal alpha value that balances recall between train and test, reducing overfitting.

30)Decision Tree Test Performance Comparison



This bar chart compares the test performance (accuracy, precision, recall, and F1 score) of three decision tree models: the default sklearn tree, a pre-pruned tree, and a post-pruned tree. It helps visualize the impact of pruning on model performance

30)EDA Questions:

Q1) What are the busiest months in the hotel?

```
Busiest Months:
 arrival_month_name
October      5317
September    4611
August       3813
June         3203
December     3021
November     2980
July         2920
April        2736
May          2598
March        2358
February     1704
January      1014
Name: count, dtype: int64
```

October months have the highest number of bookings.

Q2) Which market segment do most of the guests come from?

```
Guest Distribution by Market Segment (%):
 market_segment_type
Online           63.994487
Offline          29.022743
Corporate         5.560303
Complementary     1.077877
Aviation          0.344590
Name: proportion, dtype: float64
```

This shows which market segment Online brings most guests.

Q3) Hotel rates are dynamic and change according to demand and customer demographics. What are the differences in room prices in different market segments?

```
Average Room Price by Market Segment:
 market_segment_type
Online           112.256855
Aviation         100.704000
Offline           91.632679
Corporate         82.911740
Complementary      3.141765
Name: avg_price_per_room, dtype: float64
```

Online bookings pay the highest on average.

Complementary guests generate almost no revenue.

Q4)What percentage of bookings are canceled?

```
Booking Status (%):
 booking_status
Not_Canceled    67.236389
Canceled        32.763611
Name: proportion, dtype: float64
```

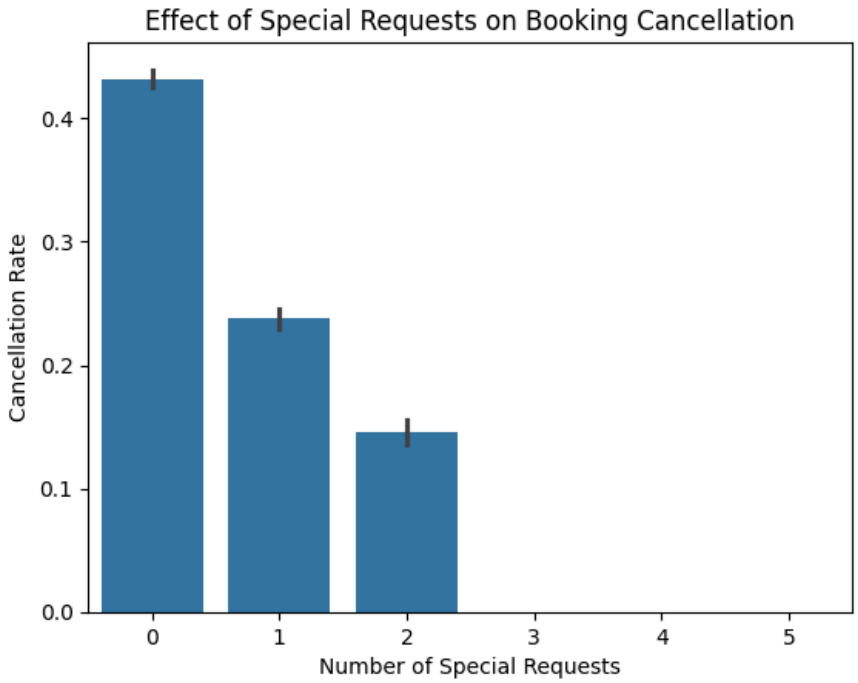
Almost **1 in 3 bookings gets canceled**, which is significant for revenue planning.

Q5)Repeating guests are the guests who stay in the hotel often and are important to brand equity. What percentage of repeating guests cancel?

```
Repeating Guests Cancellation Rate (%):
 booking_status
Not_Canceled    98.27957
Canceled         1.72043
Name: proportion, dtype: float64
```

Repeating guests rarely cancel, which means they are **loyal customers** and critical for the hotel's revenue stability.

Q6)Many guests have special requirements when booking a hotel room. Do these requirements affect booking cancellation?



Most of the customers has no or 0 special Requests

31) Actionable Insights & Recommendations

Actionable Insights & Recommendations

1. Busiest Months

Increase room prices during busy months to maximize earnings.
Hire extra staff and stock up on supplies in advance.

Run special marketing campaigns to attract guests in slower months.

2. Market Segment Contribution

Spend more on **online marketing and travel websites**, where most customers are.
Offer **loyalty programs or special packages** for corporate guests to boost weekday occupancy.
Encourage online guests to buy **premium services** like spa or dining to increase revenue.

3. Room Price Differences

Offer **premium online-only deals** to capture higher-paying guests.
Reduce the number of **free (complementary) bookings**, or combine them with paid services like meals or spa packages.

Use **different prices for different segments** to attract more guests without lowering overall revenue.

4. Booking Cancellations

Introduce **non-refundable discounted options** for budget travelers.
Offer **credits or rebooking options** instead of full cancellations.
Track **which segments cancel most** and adjust policies accordingly.

5. Repeating Guests

Create a **loyalty or membership program** to encourage first-time guests to return.
Provide **personalized offers and special recognition** to repeat guests to build long-term loyalty.

6. Special Requests

Encourage guests to add special requests during booking.
Identify these guests as **high-value customers** and offer **personalized upselling opportunities** like premium services.

32)Model Comparison

When compare logit model and desission tree we get better answer with decision tree because it will predict more accurate value with more confidence

33) Conclusions

Unpruned Decision Tree shows overfitting: The unpruned model performs extremely well on the training data but loses some accuracy on the test set, which indicates it does not generalize as well to new data.

Pruning improves model generalization: Both pre-pruning and post-pruning help simplify the tree, reduce overfitting, and make the model's performance more stable on unseen data.

Post-pruned tree gives the best balance: While its accuracy is slightly lower than the unpruned model, its F1 score is almost the same. More importantly, the tree is simpler, easier to interpret, and better suited for real-world deployment.

34)Recommendations

Deploy the Post-Pruned Decision Tree

It provides a better balance between accuracy and simplicity.

The model is easier to explain to non-technical stakeholders.

It carries a lower risk of overfitting compared to the fully grown tree.

When recall is the top priority

If the main goal is to catch as many cancellations as possible, the pre-pruned tree is a better choice since it achieved the highest recall (~0.80).

Future Improvements

Experiment with ensemble methods like Random Forest or XGBoost to improve accuracy and robustness.

Use cross-validation to fine-tune pruning parameters and select the most reliable model.

Thankyou