12/5/2025

Meachine Learning

Project-4

Tushar Choudhury

Contents

Table Of Figures

Exploratory Data Analysis

Data Overview

	Booking_ID	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	type_of_meal_plan	required_car_parking_spa
0	INN00001	2	0	1	2	Meal Plan 1	
1	INN00002	2	0	2	3	Not Selected	
2	INN00003	1	0	2	1	Meal Plan 1	
3	INN00004	2	0	0	2	Meal Plan 1	
4	INN00005	2	0	1	1	Not Selected	
...	...	...	...	...	...	...	
36270	INN36271	3	0	2	6	Meal Plan 1	
36271	INN36272	2	0	1	3	Meal Plan 1	
36272	INN36273	2	0	2	6	Meal Plan 1	
36273	INN36274	2	0	0	3	Not Selected	
36274	INN36275	2	0	1	2	Meal Plan 1	

36275 rows × 19 columns

Figure 1 - Data

The data consists of 36275 rows and 19 columns.

Information about Data

```
<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
```

Figure 2 - Information about Data

- *The majority columns in the dataset are **numerical ie('int64','float64' dtype)**.*
- *The dataset contains **13 int64 dtype, 1 float64 dtype, 5 objec dtype**.*

Description Of Data

	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

Figure 3 - Description of Data

- *The data consists of numerical columns like **number of adults, children, nights, lead time, and price**.*
- *Data consists categorical columns like **type of meal plan, room type reserved, and booking status**.*
- *Numerical columns like **required_car_parking_space, repeated_guest, no_of_previous_cancellations, and no_of_special_requests** represent categorical information as int64 type.*

Univariate Analysis

no_of_adults

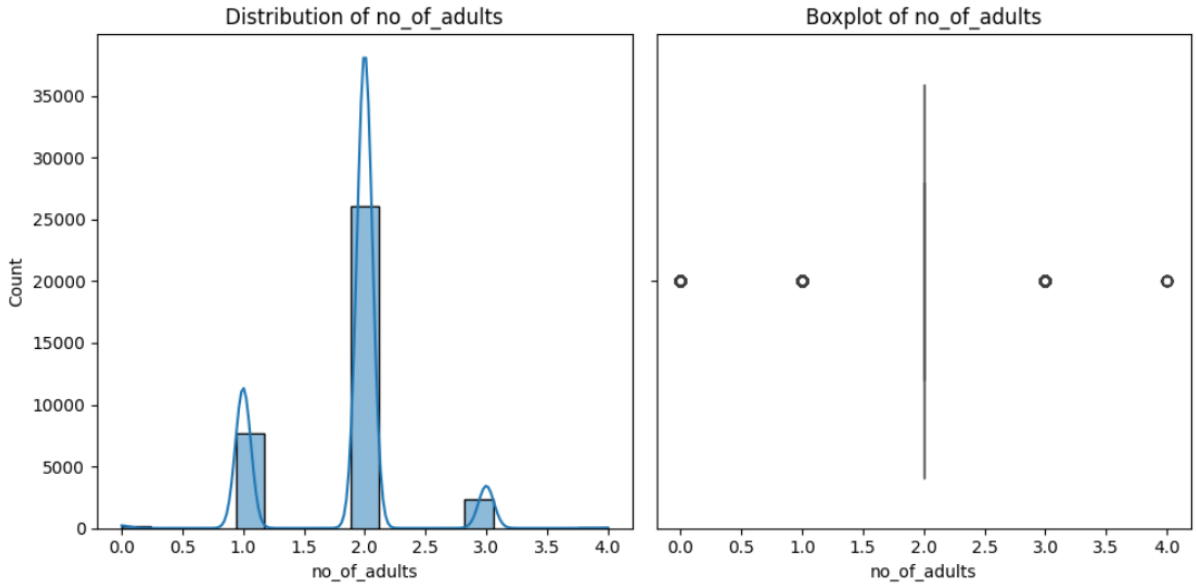


Figure 4 - Histogram and Boxplot of no_of_adults column

- *The number of visitors shows **right_skewed**.*

no_of_children

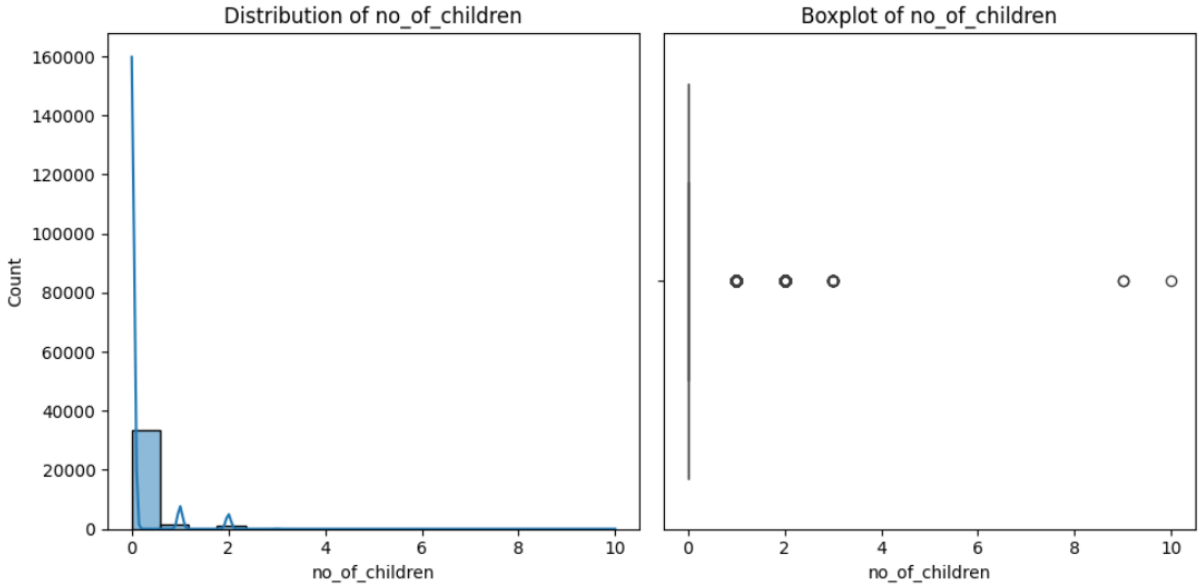


Figure 5 - Histogram and Boxplot of no_of_children column

- *The distribution of No_of_weekend_nights shows right skewed.*
- *The median number of weekedn nights guests stayed is **1 night**.*

no_of_week_nights

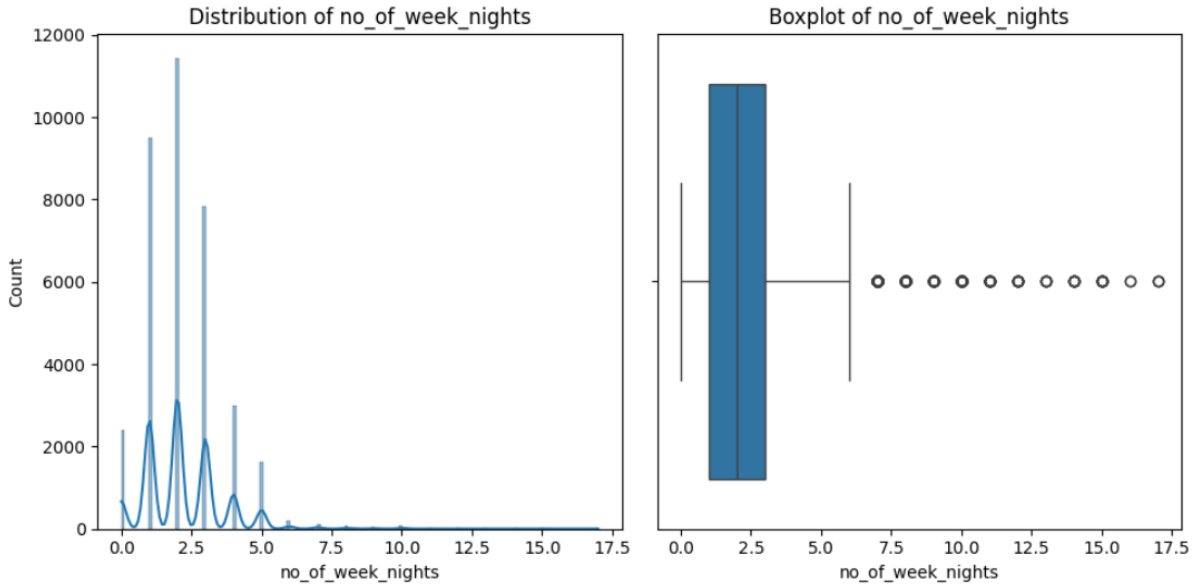


Figure 6 - Histogram and Boxplot of no_of_week_nights column

- *The distribution of **no_of_week_nights** is **right-skewed**.*
- *The median number of weeknights guests stayed is approxilately **2 nights**.*

lead_time

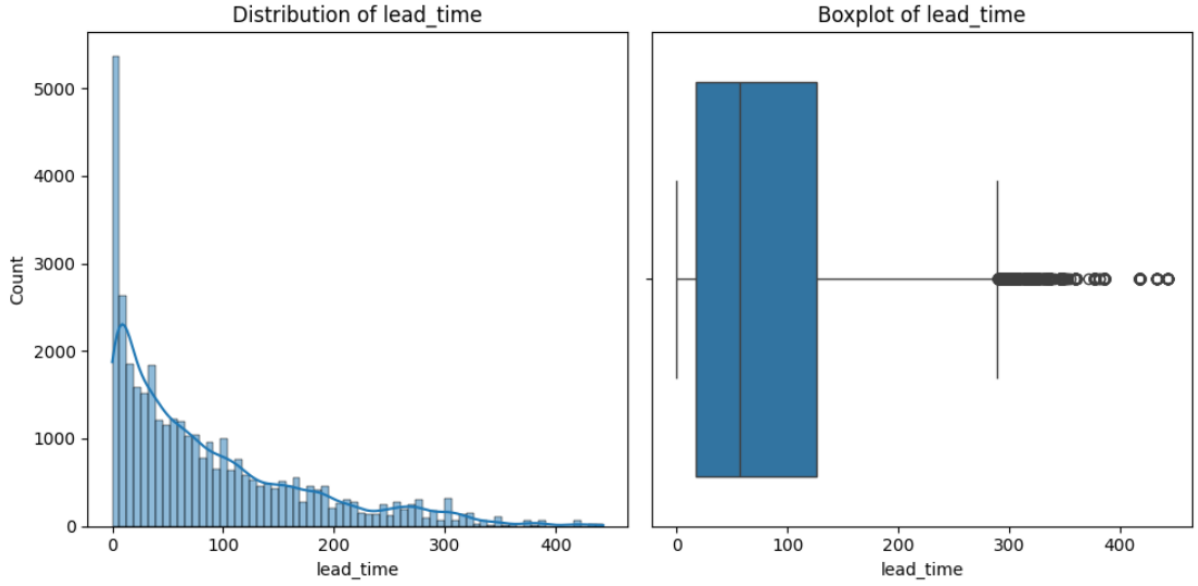


Figure 7 - Histogram and Boxplot of lead_time column

- The distribution of **lead_time** is **right-skewed**.
- The median number of days(lead_time) spent is approximately **57 days**.

avg_price_per_room

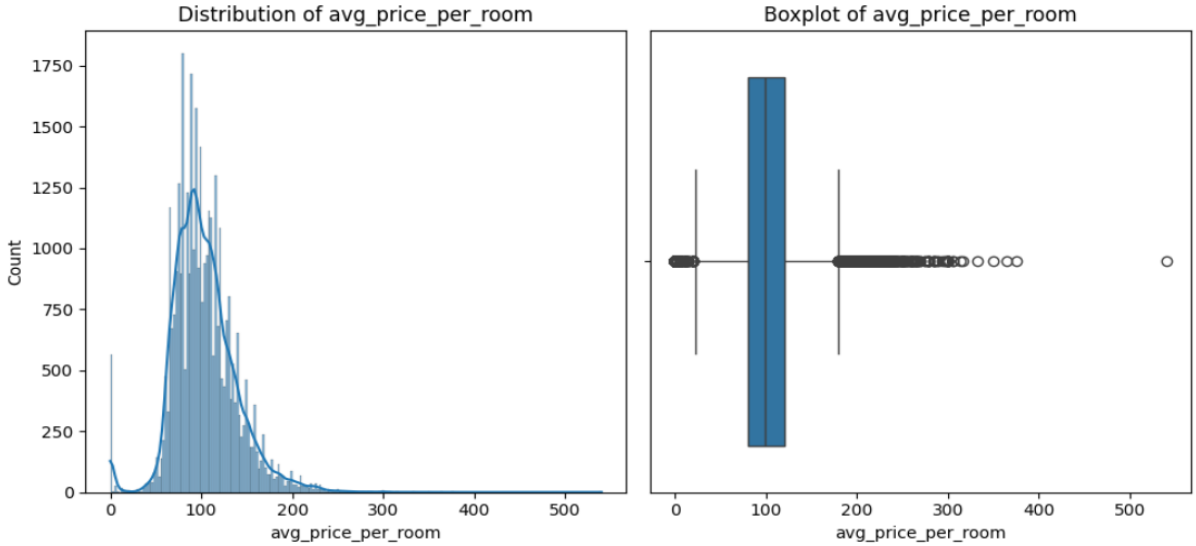


Figure 8 - Histogram and Boxplot of avg_price_per_room column

- The distribution of **avg_prive_per_room** is **right-skewed**.
- The median price per room is approximately **100 euros**.

type_of_meal_plan

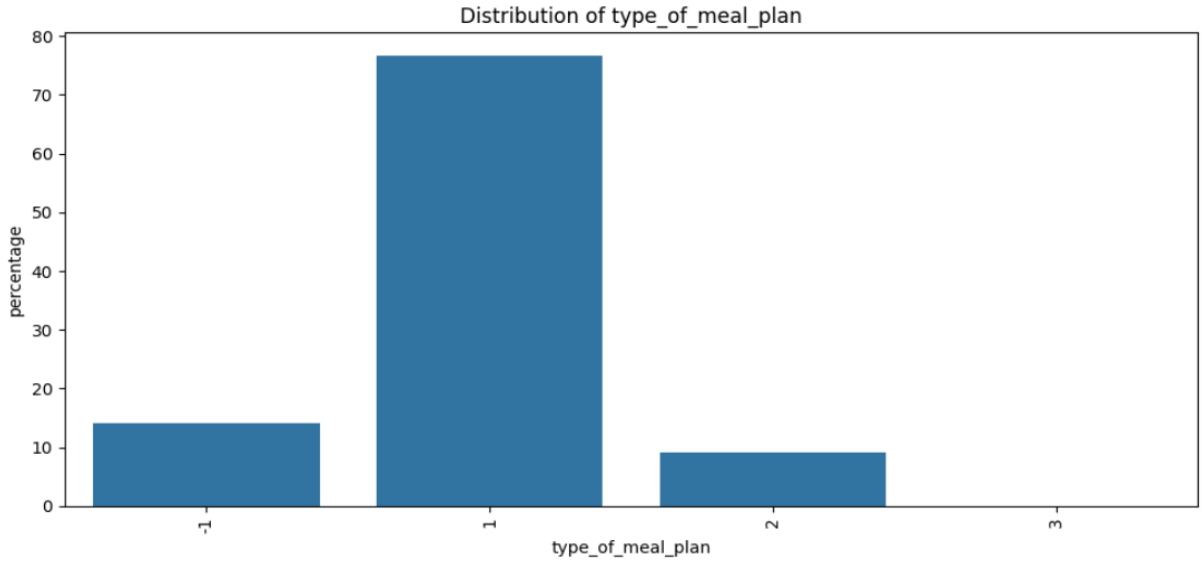


Figure 9 - Barplot of type_of_meal_plan column

type_of_meal_plan	proportion
Meal Plan 1	76.733287
Not Selected	14.141971
Meal Plan 2	9.110958
Meal Plan 3	0.013784

Figure 10 - Proportion of meal plans

- Approximately **76%** of the customers booked **Meal Plan 1**.
- Approximately **14%** of the customers have not booked any meal plan.
- Approximately **9%** of customers booked **Meal Plan 2**.

room_type_reserved

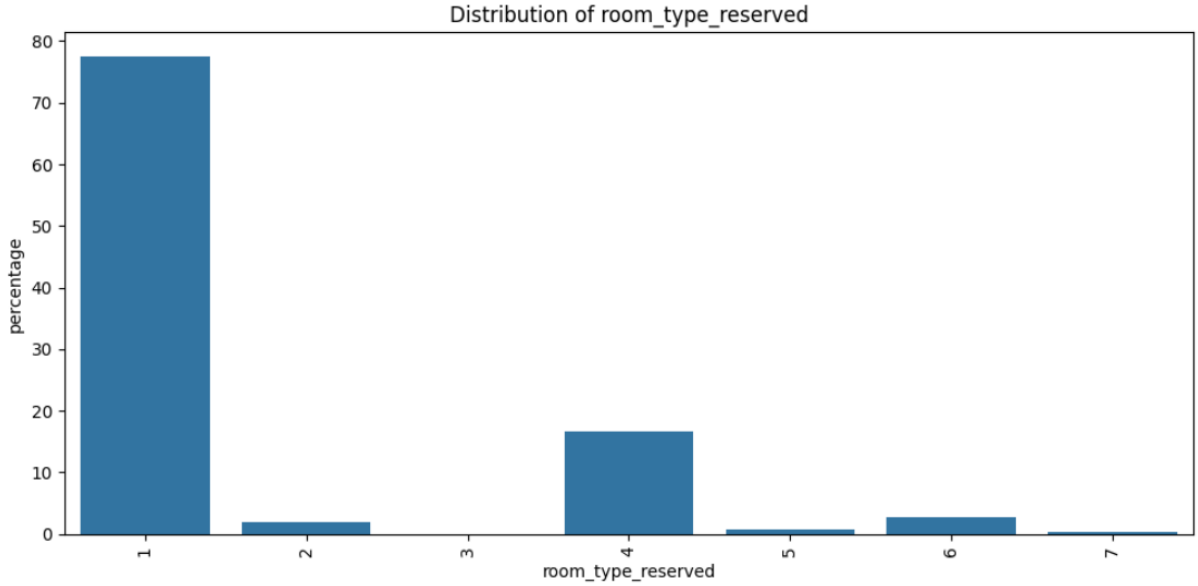


Figure 11 - Barplot of room_type_reserved column

room_type_reserved	proportion
Room_Type 1	77.546520
Room_Type 4	16.697450
Room_Type 6	2.662991
Room_Type 2	1.907650
Room_Type 5	0.730531
Room_Type 7	0.435562
Room_Type 3	0.019297

Figure 12 - Proportion of room_type_reseved

- Approximately **77.5%** of the customers booked **Room Type 1**.
- Approximately **17%** of the customers booked **Room Type 4**.
- Approximately **3%** of the customers booked **Room Type 6**.

booking_status

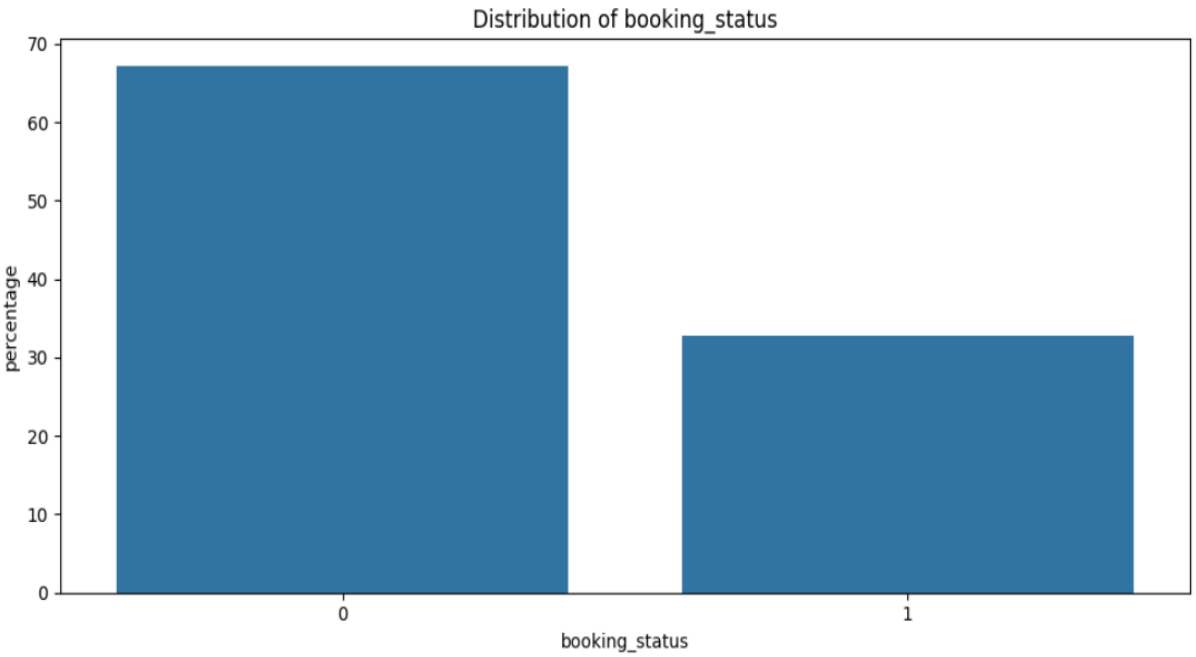


Figure 13 - Barplot of booking_status column

booking_status	proportion
Not_Canceled	67.236389
Canceled	32.763611

Figure 14 - Proportion of booking_status

- Approximately **67%** of the customers have **not cancelled** their booking.
- Approximately **33%** of the customers have **cancelled** their booking.

Bivariate Analysis

Heatmap

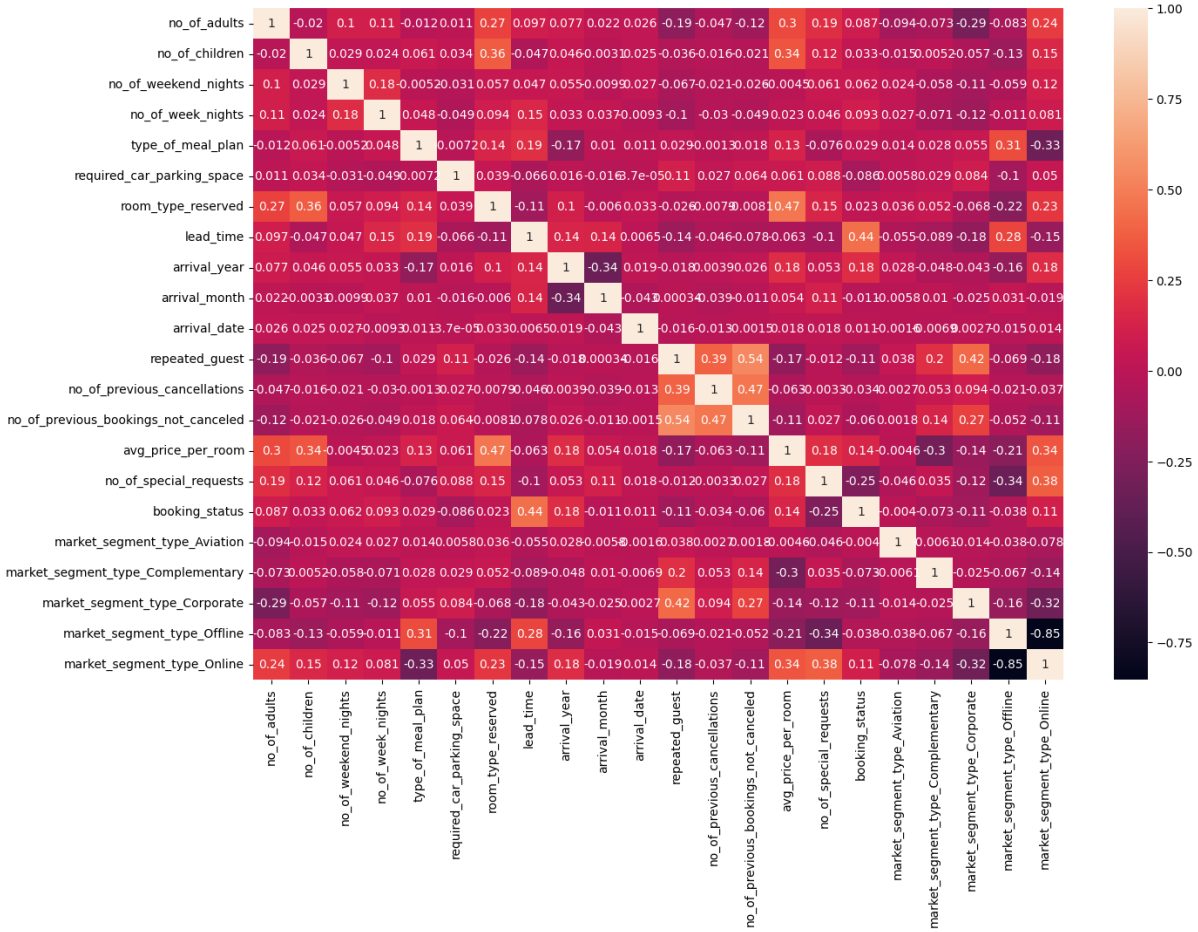


Figure 15 - Heatmap of Numerical Columns

- **Repeated_guest** is **highly positively correlated** with **no_of_previous_bookings_not_canceled** with correlation value of **0.54**.
- **Repeated_guest** is **postively correlated** with **no_of_previous_cancellations** with correlation value of **0.39**.

Pair plot

Figure 16 – Pairplot

- Bookings with longer lead times tend to have a higher proportion of cancellations.
- Very low prices tend to be cancelled more, while bookings with high prices are 'Not_Canceled'.
- No_of_special_requests and booking_status have some relation.

EDA Questions

What are the busiest months in the hotel?

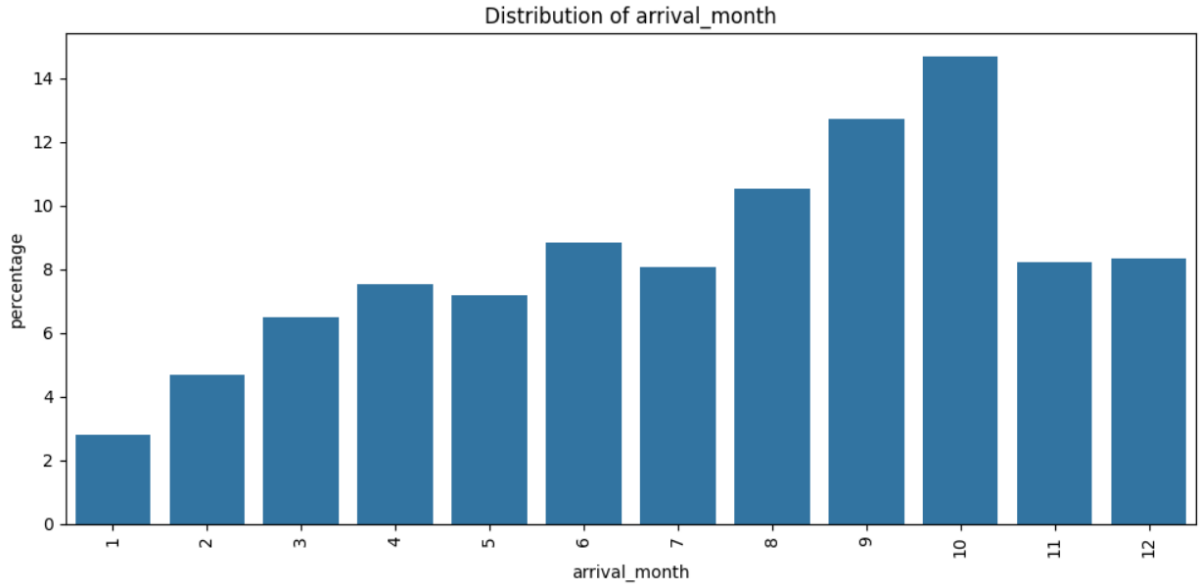


Figure 17 - Question 1

***10th month (October)** is the busiest month in the hotel approximately with **14% of bookings**.*

Which market segment do most of the guests come from?

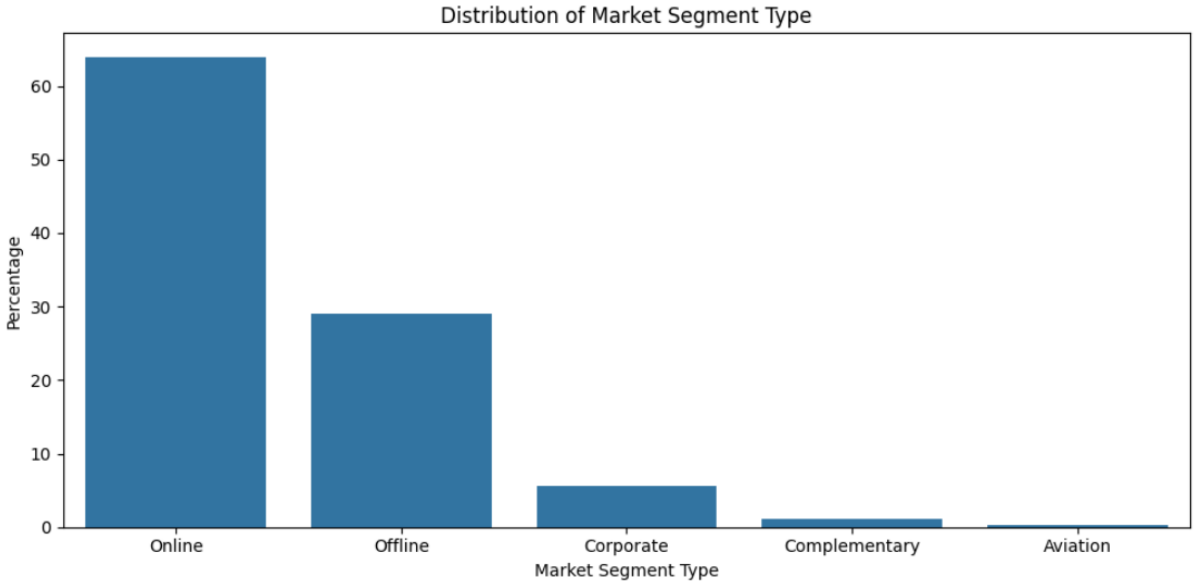


Figure 18 - Question 2

Most of the guests come from **Online** market_segment with **64%** of bookings.

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

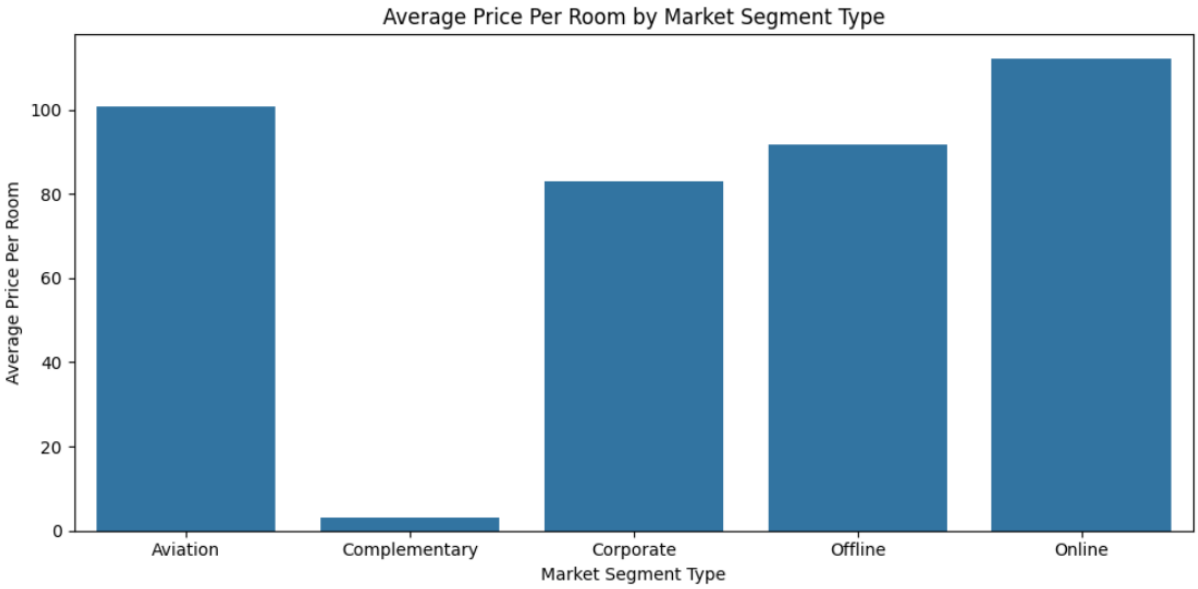


Figure 19 - Question 3

- *Bookings made through online channels generally command higher prices.*
- *The Aviation also has high average price but comparatively lower than 'Online'.*
- *Complementary has the low average prices.*
- *Bookings made through offline has 3rd highest average prices.*

What percentage of bookings are canceled?

```
Percentage of Bookings cancelled: 32.76361130254997
```

Figure 20 - Question 4

Approximately 32% of overall bookings are cancelled.

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

```
Percentage of repeating guests cancel: 1.7204301075268817
```

Figure 21 - Question 5

Approximately 1.7% of repeated guests cancelled their booking.

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

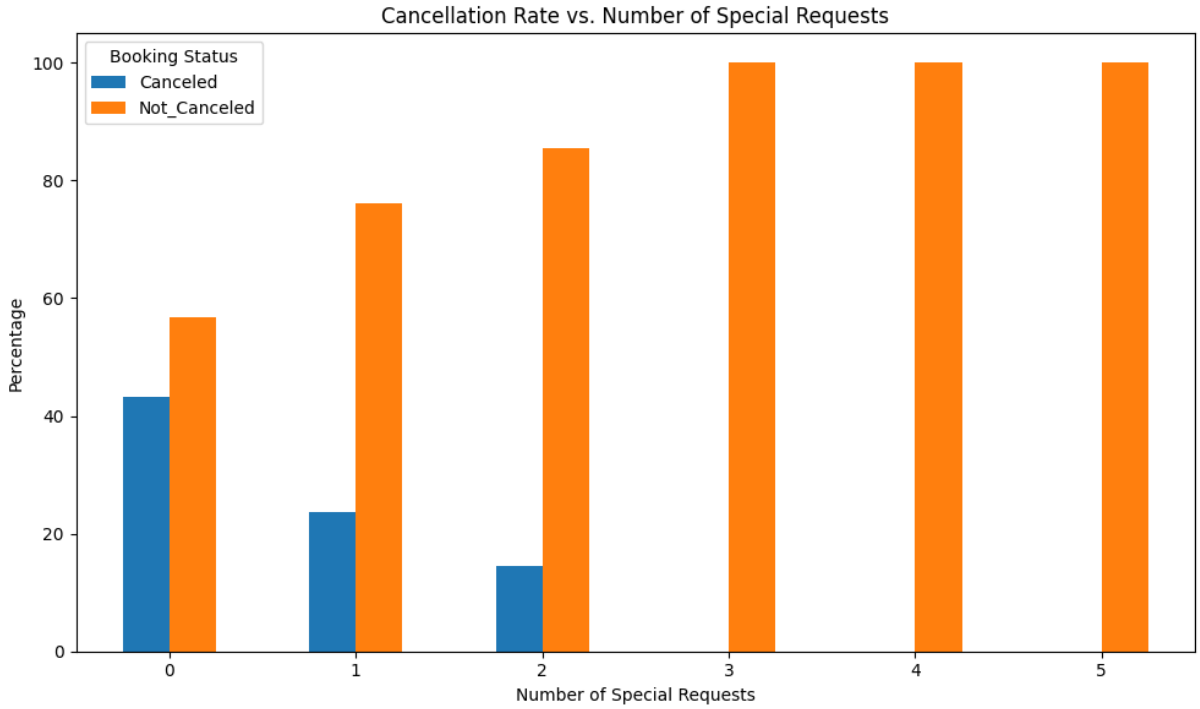


Figure 22-Question 6

- Customers with special requirements often do not cancel their booking.
- Lesser the special requirements higher the chances of cancelling the booking.

Data Preprocessing

- There are no duplicate values in this dataset.
- There are no missing values in this data.

Outlier Treatment

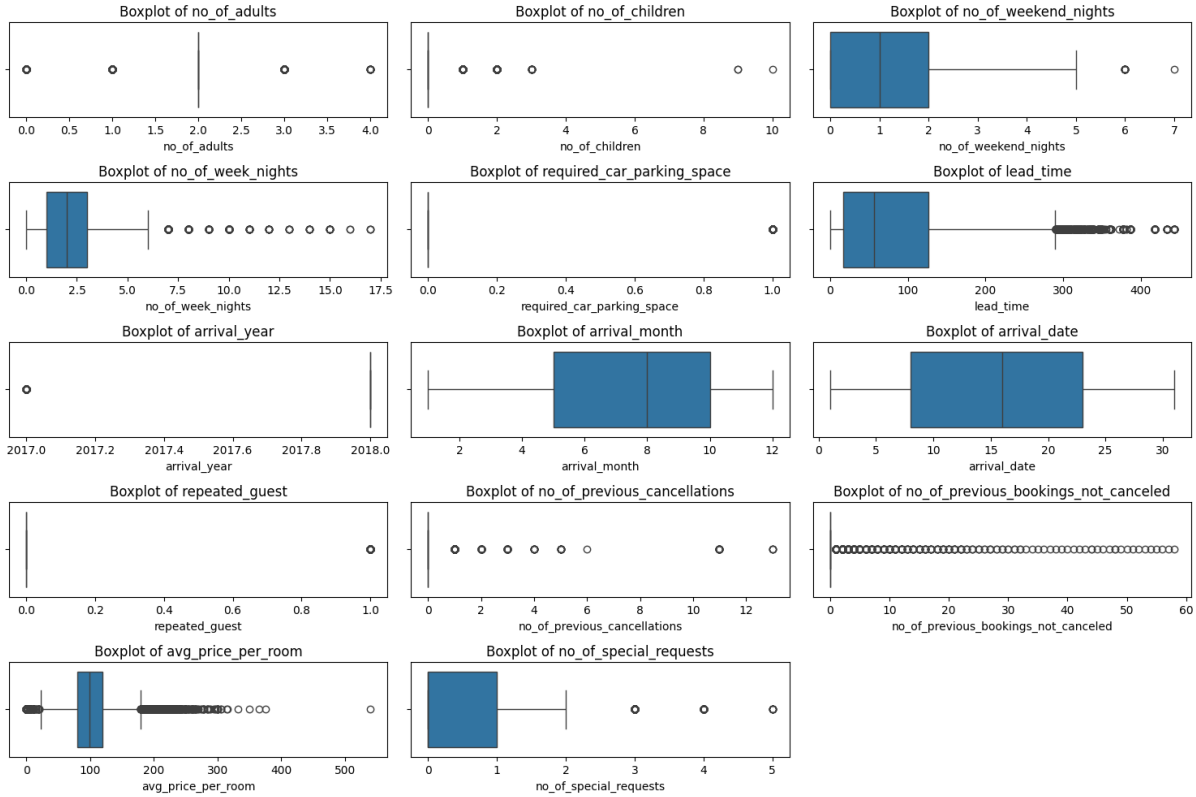


Figure 23 - Outliers of Numerical Columns

We do not treat the outliers because they are proper values that influence the target variable.

Feature Engineering

booking_status	count
Not_Canceled	24390
Canceled	11885

Figure 24 - Booking status count

type_of_meal_plan	count
Meal Plan 1	27835
Not Selected	5130
Meal Plan 2	3305
Meal Plan 3	5

Figure 25 - Meal plan value count

room_type_reserved	count
Room_Type 1	28130
Room_Type 4	6057
Room_Type 6	966
Room_Type 2	692
Room_Type 5	265
Room_Type 7	158
Room_Type 3	7

Figure 26 - Room type value count

market_segment_type	count
Online	23214
Offline	10528
Corporate	2017
Complementary	391
Aviation	125

Figure 27 - Market segment values count

Dummies

	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	arrival_month
0	2	0	1	2	1	0	1	224	2017	10
1	2	0	2	3	-1	0	1	5	2018	11
2	1	0	2	1	1	0	1	1	2018	2
3	2	0	0	2	1	0	1	211	2018	5
4	2	0	1	1	-1	0	1	48	2018	4
5	2	0	0	2	2	0	1	346	2018	9
6	2	0	1	3	1	0	1	34	2017	10
7	2	0	1	3	1	0	4	83	2018	12
8	3	0	0	4	1	0	1	121	2018	7
9	2	0	0	5	1	0	4	44	2018	10

10 rows × 22 columns

Figure 28 - Data after Dummies

- Number of rows in train data = 25392
- Number of rows in test data = 10883

Model Building

Logistic Regression Model

```
                           Logit Regression Results
==============================================================================
Dep. Variable:         booking_status   No. Observations:                25392
Model:                          Logit   Df Residuals:                    25371
Method:                           MLE   Df Model:                           20
Date:                Fri, 05 Dec 2025   Pseudo R-squ.:                  0.3312
Time:                        11:51:57   Log-Likelihood:                -10741.
converged:                      False   LL-Null:                       -16060.
Covariance Type:            nonrobust   LLR p-value:                     0.000
=================================================================================================
                                    coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                          -834.4989    118.427     -7.047      0.000   -1066.612    -602.386
no_of_adults                      0.0497      0.037      1.328      0.184      -0.024       0.123
no_of_children                   -0.0341      0.048     -0.717      0.474      -0.127       0.059
no_of_weekend_nights              0.1449      0.020      7.313      0.000       0.106       0.184
no_of_week_nights                 0.0329      0.012      2.694      0.007       0.009       0.057
type_of_meal_plan                -0.0536      0.025     -2.180      0.029      -0.102      -0.005
required_car_parking_space       -1.6252      0.137    -11.858      0.000      -1.894      -1.357
room_type_reserved               -0.1260      0.016     -8.080      0.000      -0.157      -0.095
lead_time                         0.0158      0.000     59.630      0.000       0.015       0.016
arrival_year                      0.4123      0.059      7.026      0.000       0.297       0.527
arrival_month                    -0.0484      0.006     -7.478      0.000      -0.061      -0.036
arrival_date                      0.0032      0.002      1.644      0.100      -0.001       0.007
repeated_guest                   -1.9120      0.763     -2.506      0.012      -3.408      -0.416
no_of_previous_cancellations      0.3470      0.103      3.383      0.001       0.146       0.548
no_of_previous_bookings_not_canceled -1.3929  0.913     -1.526      0.127      -3.181       0.396
avg_price_per_room                0.0188      0.001     26.252      0.000       0.017       0.020
no_of_special_requests           -1.4855      0.030    -48.986      0.000      -1.545      -1.426
market_segment_type_Complementary -21.1663  1.27e+04     -0.002      0.999   -2.5e+04    2.49e+04
market_segment_type_Corporate    -0.9186      0.275     -3.336      0.001      -1.458      -0.379
market_segment_type_Offline      -1.7783      0.264     -6.743      0.000      -2.295      -1.261
market_segment_type_Online       -0.0196      0.261     -0.075      0.940      -0.531       0.492
=================================================================================================
```

Figure 29 - Logistic Regression Model Summary

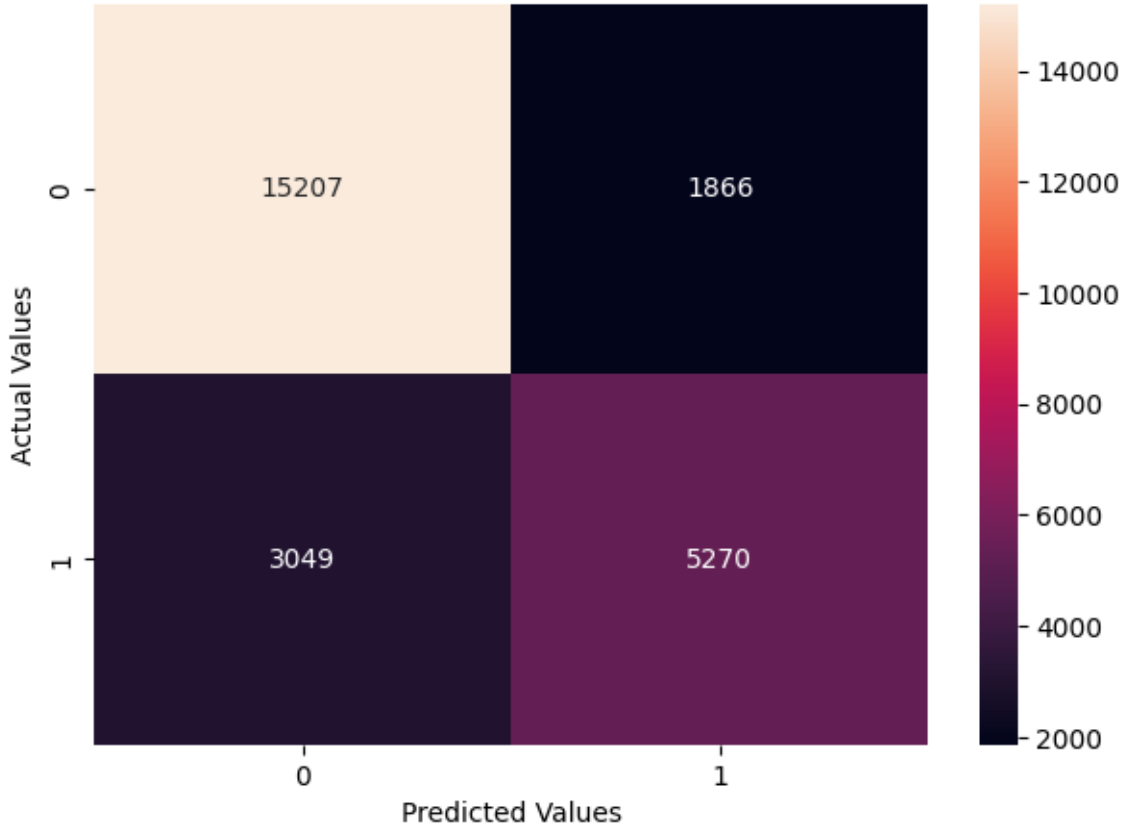
Model Performance

Training performance:

	Accuracy	Recall	Precision	F1
0	0.806435	0.63349	0.738509	0.68198

Figure 30 - Training Performance values

Confusion Matrix



- ***True Positives (TP):*** *The customer has not cancelled the booking and the model predicted customer has not canceled the booking.*
- ***True Negatives (TN):*** *The customer had cancelled the booking and the model predicted customer has cancelled the booking.*
- ***False Positives (FP):*** *The model predicted customer has not cancelled the booking but the customer has cancelled the booking.*
- ***False Negatives (FN):*** *The model predicted customer has cancelled the booking but the customer has not cancelled the booking.*

Decision Tree Model (Without Pruning)

Decision Tree Visualization

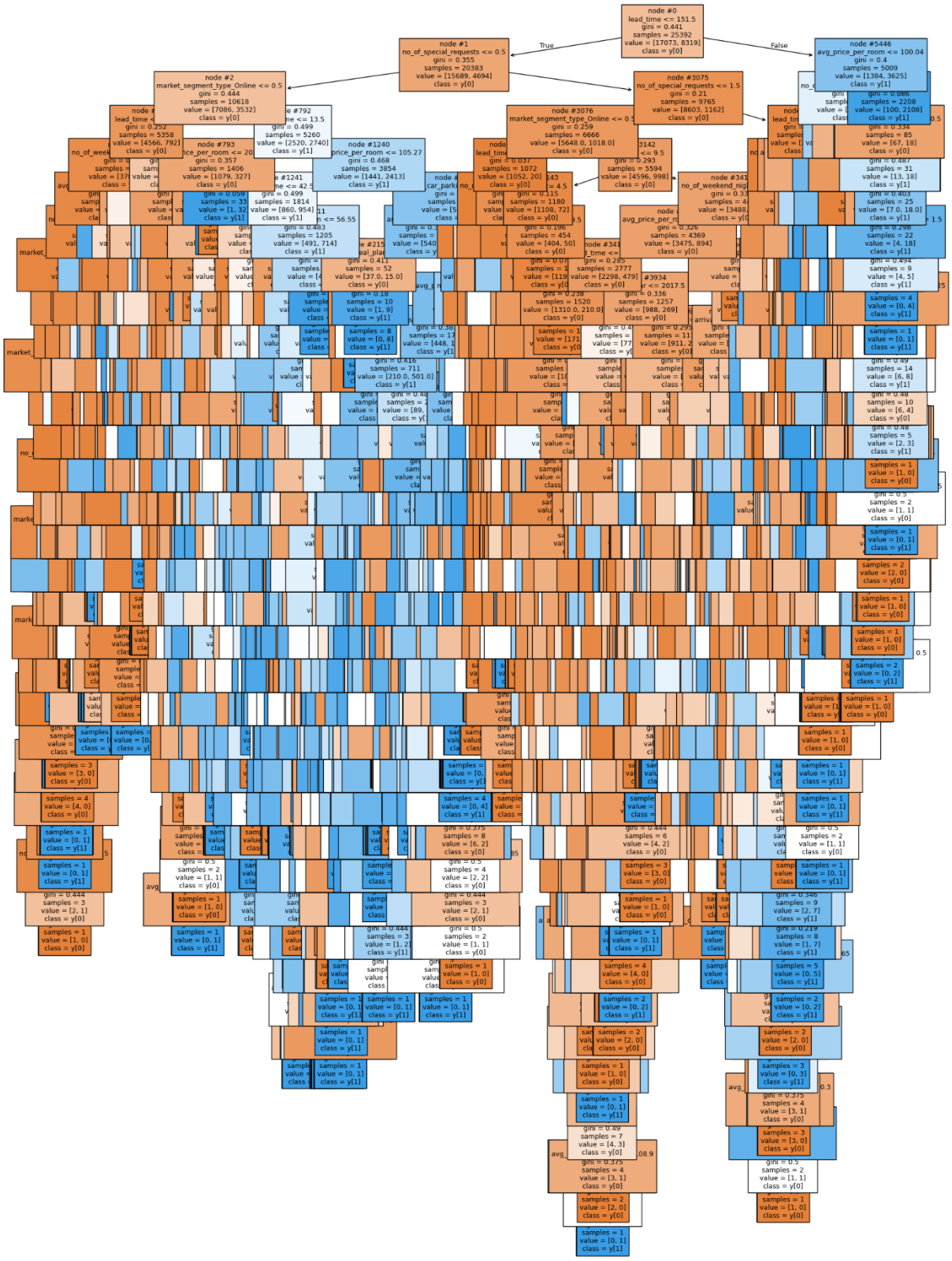


Figure 31 - Decision Tree (Without Pruning)

	Imp
lead_time	0.345855
avg_price_per_room	0.174797
market_segment_type_Online	0.086834
arrival_date	0.084449
no_of_special_requests	0.071219
arrival_month	0.070196
no_of_week_nights	0.046483
no_of_adults	0.034465
no_of_weekend_nights	0.032904
arrival_year	0.014476
type_of_meal_plan	0.012505
room_type_reserved	0.008796
required_car_parking_space	0.006685
no_of_children	0.005101
market_segment_type_Offline	0.002518
market_segment_type_Corporate	0.001808
no_of_previous_bookings_not_canceled	0.000654
repeated_guest	0.000254
no_of_previous_cancellations	0.000000
market_segment_type_Complementary	0.000000

Figure 32 - Feature Significance Values

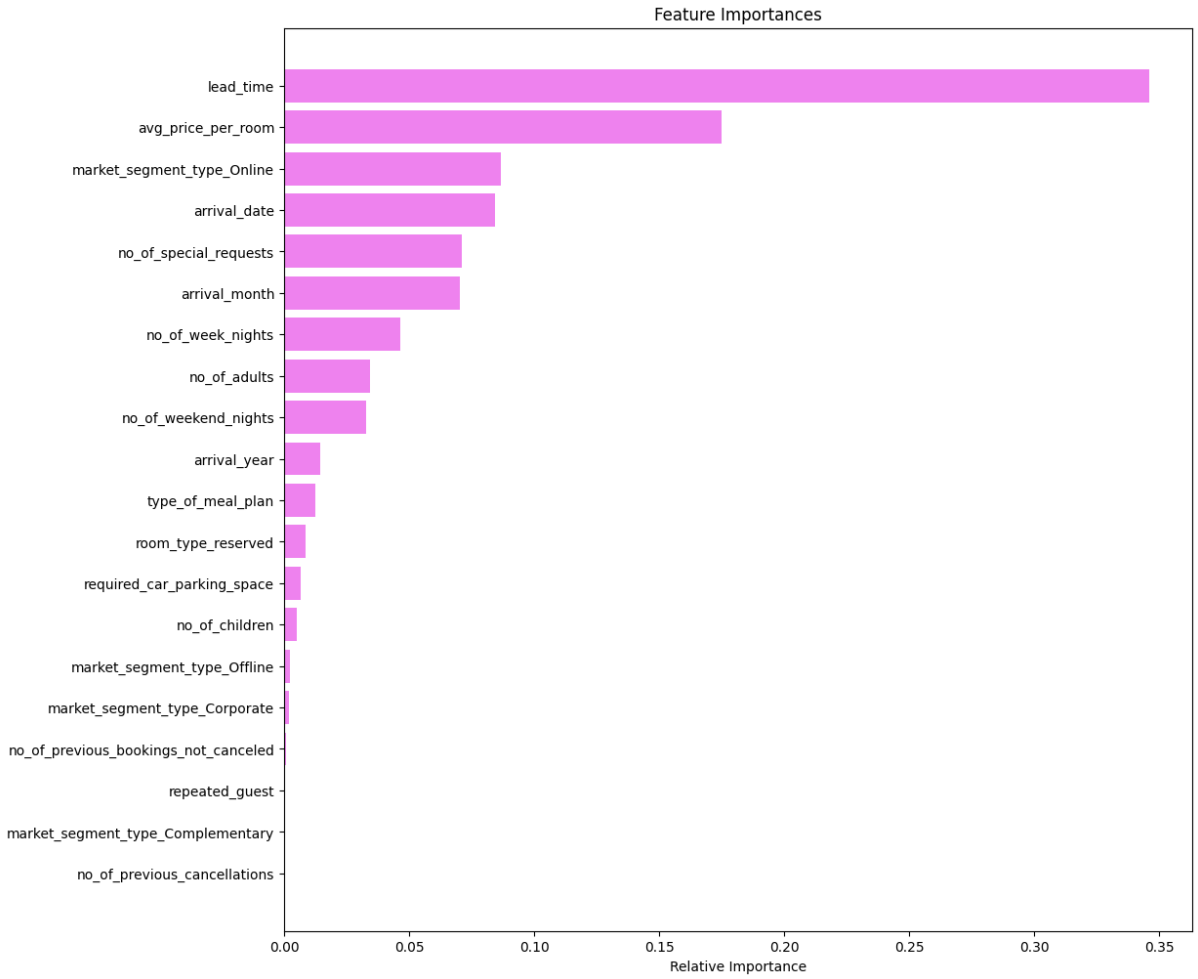


Figure 33 - Feature Importance Chart

Decision Tree Model Performance

Training set Performance

	Accuracy	Recall	Precision	F1
0	0.994368	0.985695	0.997082	0.991356

Figure 34 - Training set performance of without Pruning DTree mode

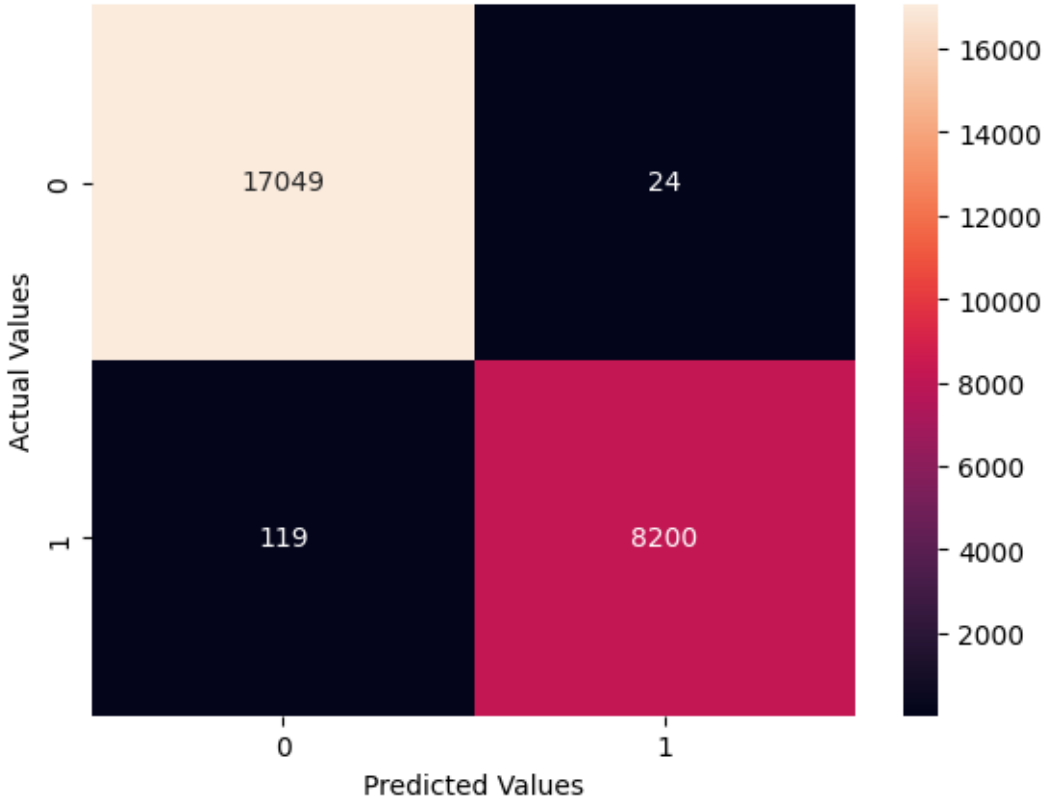


Figure 35 - Confusion Matrix of Training set without pruning model

Testing Set Performance

	Accuracy	Recall	Precision	F1
0	0.861068	0.790802	0.786392	0.788591

Figure 36 - Testing set performance of without Pruning DTree mode

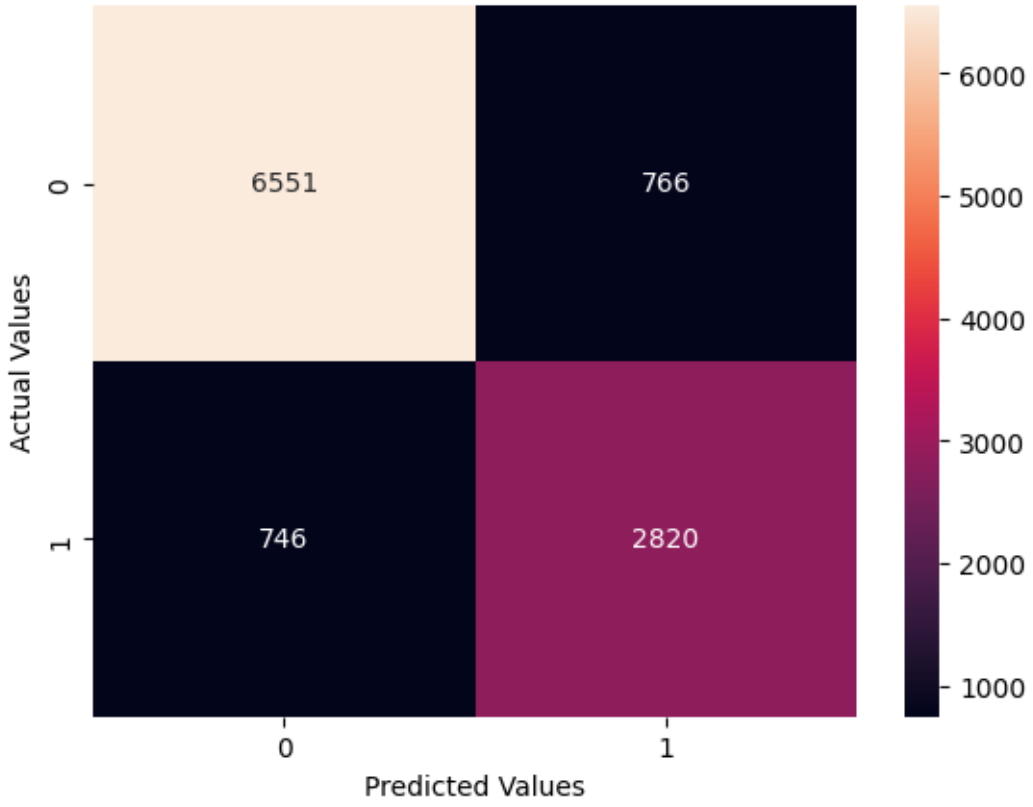


Figure 37 - Confusion Matrix of Testing set without pruning model

- The model is performing well on both Training set and Testing set.
- The difference between accuracy of Training set and Testing set is large.
- The difference between recall score of Training set and Testing set is very large.

Model Performance Improvement

Regression Model Performance Improvement

Checking for Multicollinearity

```
VIF values:

const                                 3.839020e+07
no_of_adults                          1.331019e+00
no_of_children                        1.271569e+00
no_of_weekend_nights                  1.066172e+00
no_of_week_nights                     1.088316e+00
type_of_meal_plan                     1.361876e+00
required_car_parking_space            1.033816e+00
room_type_reserved                    1.563090e+00
lead_time                             1.359257e+00
arrival_year                          1.389441e+00
arrival_month                         1.270822e+00
arrival_date                          1.006548e+00
repeated_guest                        1.747595e+00
no_of_previous_cancellations          1.321458e+00
no_of_previous_bookings_not_canceled  1.569723e+00
avg_price_per_room                    1.901873e+00
no_of_special_requests                1.243814e+00
market_segment_type_Complementary     4.296740e+00
market_segment_type_Corporate         1.657424e+01
market_segment_type_Offline           6.238500e+01
market_segment_type_Online            6.941084e+01
dtype: float64
```

Figure 38 - VIF values of logistic Regression Model

- There is no multicollinearity present in the data.

Removing high P_values

market_segment_type_Complementary	0.998674
market_segment_type_Online	0.940168
no_of_children	0.473506
no_of_adults	0.184317
no_of_previous_bookings_not_canceled	0.126908
arrival_date	0.100264

Figure 39 - Columns with P_values > 0.05

Dropping market_segment_type_Complementary as it has highest p_value

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25372
Method:	MLE	Df Model:	19
Date:	Sat, 06 Dec 2025	Pseudo R-squ.:	0.3308
Time:	03:38:21	Log-Likelihood:	-10748.
converged:	True	LL-Null:	-16060.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-838.3317	118.330	-7.085	0.000	-1070.254	-606.409
no_of_adults	0.0440	0.037	1.178	0.239	-0.029	0.117
no_of_children	-0.0401	0.048	-0.844	0.399	-0.133	0.053
no_of_weekend_nights	0.1470	0.020	7.422	0.000	0.108	0.186
no_of_week_nights	0.0345	0.012	2.820	0.005	0.011	0.058
type_of_meal_plan	-0.0552	0.025	-2.248	0.025	-0.103	-0.007
required_car_parking_space	-1.6257	0.137	-11.857	0.000	-1.894	-1.357
room_type_reserved	-0.1263	0.016	-8.099	0.000	-0.157	-0.096
lead_time	0.0158	0.000	59.682	0.000	0.015	0.016
arrival_year	0.4140	0.059	7.061	0.000	0.299	0.529
arrival_month	-0.0486	0.006	-7.518	0.000	-0.061	-0.036
arrival_date	0.0032	0.002	1.627	0.104	-0.001	0.007
repeated_guest	-1.9072	0.762	-2.503	0.012	-3.401	-0.414
no_of_previous_cancellations	0.3462	0.103	3.375	0.001	0.145	0.547
no_of_previous_bookings_not_canceled	-1.3785	0.905	-1.522	0.128	-3.153	0.396
avg_price_per_room	0.0190	0.001	26.717	0.000	0.018	0.020
no_of_special_requests	-1.4870	0.030	-49.038	0.000	-1.546	-1.428
market_segment_type_Corporate	-0.5244	0.258	-2.030	0.042	-1.031	-0.018
market_segment_type_Offline	-1.3831	0.245	-5.634	0.000	-1.864	-0.902
market_segment_type_Online	0.3741	0.243	1.540	0.123	-0.102	0.850

Figure 40 - Model Summary after removing market_segment_type_Complementary Column

```
Accuracy on training set :  0.8061988027725268
```

- There is no significant change in the model performance as compared to initial model.

Dropping no_of_children due to it's highest p_value

no_of_children	0.398900
no_of_adults	0.238629
no_of_previous_bookings_not_canceled	0.127893
market_segment_type_Online	0.123470
arrival_date	0.103781

Figure 41 - P-Values

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25373
Method:	MLE	Df Model:	18
Date:	Sat, 06 Dec 2025	Pseudo R-squ.:	0.3307
Time:	03:38:21	Log-Likelihood:	-10748.
converged:	True	LL-Null:	-16060.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-842.7455	118.183	-7.131	0.000	-1074.380	-611.111
no_of_adults	0.0509	0.036	1.394	0.163	-0.021	0.122
no_of_weekend_nights	0.1466	0.020	7.406	0.000	0.108	0.185
no_of_week_nights	0.0345	0.012	2.822	0.005	0.011	0.058
type_of_meal_plan	-0.0548	0.025	-2.231	0.026	-0.103	-0.007
required_car_parking_space	-1.6272	0.137	-11.866	0.000	-1.896	-1.358
room_type_reserved	-0.1295	0.015	-8.558	0.000	-0.159	-0.100
lead_time	0.0158	0.000	59.757	0.000	0.015	0.016
arrival_year	0.4162	0.059	7.107	0.000	0.301	0.531
arrival_month	-0.0484	0.006	-7.493	0.000	-0.061	-0.036
arrival_date	0.0031	0.002	1.615	0.106	-0.001	0.007
repeated_guest	-1.9063	0.761	-2.504	0.012	-3.399	-0.414
no_of_previous_cancellations	0.3458	0.103	3.371	0.001	0.145	0.547
no_of_previous_bookings_not_canceled	-1.3777	0.905	-1.522	0.128	-3.152	0.396
avg_price_per_room	0.0189	0.001	27.335	0.000	0.018	0.020
no_of_special_requests	-1.4883	0.030	-49.131	0.000	-1.548	-1.429
market_segment_type_Corporate	-0.5231	0.258	-2.027	0.043	-1.029	-0.017
market_segment_type_Offline	-1.3850	0.245	-5.646	0.000	-1.866	-0.904
market_segment_type_Online	0.3703	0.243	1.526	0.127	-0.105	0.846

Figure 42 - Model Summary after removing no_of_children column

```
Accuracy on training set :  0.8060806553245117
```

- There is no significant change in the model performance as compared to initial model.

Dropping no_of_adults as it has highest p_value

no_of_adults	0.163411
no_of_previous_bookings_not_canceled	0.127954
market_segment_type_Online	0.126969
arrival_date	0.106228

Figure 43 - p_value

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25374
Method:	MLE	Df Model:	17
Date:	Sat, 06 Dec 2025	Pseudo R-squ.:	0.3307
Time:	03:38:22	Log-Likelihood:	-10749.
converged:	True	LL-Null:	-16060.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-838.1585	118.149	-7.094	0.000	-1069.727	-606.590
no_of_weekend_nights	0.1477	0.020	7.468	0.000	0.109	0.187
no_of_week_nights	0.0347	0.012	2.840	0.005	0.011	0.059
type_of_meal_plan	-0.0564	0.025	-2.300	0.021	-0.104	-0.008
required_car_parking_space	-1.6238	0.137	-11.844	0.000	-1.892	-1.355
room_type_reserved	-0.1260	0.015	-8.448	0.000	-0.155	-0.097
lead_time	0.0159	0.000	60.100	0.000	0.015	0.016
arrival_year	0.4139	0.059	7.070	0.000	0.299	0.529
arrival_month	-0.0485	0.006	-7.507	0.000	-0.061	-0.036
arrival_date	0.0032	0.002	1.662	0.097	-0.001	0.007
repeated_guest	-1.9114	0.765	-2.497	0.013	-3.411	-0.411
no_of_previous_cancellations	0.3476	0.103	3.386	0.001	0.146	0.549
no_of_previous_bookings_not_canceled	-1.3823	0.906	-1.525	0.127	-3.158	0.394
avg_price_per_room	0.0190	0.001	27.641	0.000	0.018	0.020
no_of_special_requests	-1.4839	0.030	-49.282	0.000	-1.543	-1.425
market_segment_type_Corporate	-0.5143	0.258	-1.992	0.046	-1.020	-0.008
market_segment_type_Offline	-1.3569	0.245	-5.544	0.000	-1.836	-0.877
market_segment_type_Online	0.4030	0.242	1.667	0.095	-0.071	0.877

Figure 44 - Model Summary after removing no_of_adults

```
Accuracy on training set :  0.8054505356017644
```

- There is no significant change in the model performance as compared to initial model.

Dropping no_of_previous_bookings_not_canceled due to it's highest p_value

no_of_previous_bookings_not_canceled	0.127137
arrival_date	0.096613
market_segment_type_Online	0.095477

Figure 45 - P_values

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25375
Method:	MLE	Df Model:	16
Date:	Sat, 06 Dec 2025	Pseudo R-squ.:	0.3304
Time:	03:38:22	Log-Likelihood:	-10754.
converged:	True	LL-Null:	-16060.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-833.9312	118.159	-7.058	0.000	-1065.518	-602.345
no_of_weekend_nights	0.1478	0.020	7.468	0.000	0.109	0.187
no_of_week_nights	0.0348	0.012	2.850	0.004	0.011	0.059
type_of_meal_plan	-0.0572	0.025	-2.333	0.020	-0.105	-0.009
required_car_parking_space	-1.6203	0.137	-11.821	0.000	-1.889	-1.352
room_type_reserved	-0.1261	0.015	-8.457	0.000	-0.155	-0.097
lead_time	0.0159	0.000	60.183	0.000	0.015	0.016
arrival_year	0.4118	0.059	7.034	0.000	0.297	0.527
arrival_month	-0.0485	0.006	-7.502	0.000	-0.061	-0.036
arrival_date	0.0032	0.002	1.630	0.103	-0.001	0.007
repeated_guest	-3.0360	0.601	-5.055	0.000	-4.213	-1.859
no_of_previous_cancellations	0.2853	0.078	3.651	0.000	0.132	0.438
avg_price_per_room	0.0190	0.001	27.666	0.000	0.018	0.020
no_of_special_requests	-1.4852	0.030	-49.327	0.000	-1.544	-1.426
market_segment_type_Corporate	-0.5219	0.258	-2.023	0.043	-1.028	-0.016
market_segment_type_Offline	-1.3547	0.245	-5.539	0.000	-1.834	-0.875
market_segment_type_Online	0.4056	0.242	1.679	0.093	-0.068	0.879

Figure 46 - Model Summary after removing no_of_previous_bookings_not_cancelled

```
Accuracy on training set :  0.805489918084436
```

- There is no significant change in the model performance as compared to initial model.

Dropping arrival_date due to highest p_value

arrival_date	0.103148
market_segment_type_Online	0.093203

Figure 47 - P_values

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25376
Method:	MLE	Df Model:	15
Date:	Sat, 06 Dec 2025	Pseudo R-squ.:	0.3303
Time:	03:38:22	Log-Likelihood:	-10755.
converged:	True	LL-Null:	-16060.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-833.9237	118.210	-7.055	0.000	-1065.612	-602.236
no_of_weekend_nights	0.1486	0.020	7.511	0.000	0.110	0.187
no_of_week_nights	0.0344	0.012	2.814	0.005	0.010	0.058
type_of_meal_plan	-0.0569	0.025	-2.319	0.020	-0.105	-0.009
required_car_parking_space	-1.6221	0.137	-11.836	0.000	-1.891	-1.353
room_type_reserved	-0.1255	0.015	-8.421	0.000	-0.155	-0.096
lead_time	0.0159	0.000	60.197	0.000	0.015	0.016
arrival_year	0.4119	0.059	7.031	0.000	0.297	0.527
arrival_month	-0.0491	0.006	-7.612	0.000	-0.062	-0.036
repeated_guest	-3.0472	0.603	-5.053	0.000	-4.229	-1.865
no_of_previous_cancellations	0.2850	0.078	3.638	0.000	0.131	0.439
avg_price_per_room	0.0190	0.001	27.659	0.000	0.018	0.020
no_of_special_requests	-1.4835	0.030	-49.323	0.000	-1.542	-1.425
market_segment_type_Corporate	-0.5166	0.258	-2.002	0.045	-1.022	-0.011
market_segment_type_Offline	-1.3559	0.245	-5.541	0.000	-1.835	-0.876
market_segment_type_Online	0.4059	0.242	1.679	0.093	-0.068	0.880

Figure 48 - Model Summary after removing arrival_date column

```
Accuracy on training set :  0.8062381852551985
```

- There is no significant change in the model performance as compared to initial model.

Dropping market_segment_type_Online due to highest p_value

```
                          Logit Regression Results
==============================================================================
Dep. Variable:         booking_status   No. Observations:                25392
Model:                          Logit   Df Residuals:                    25377
Method:                           MLE   Df Model:                           14
Date:                Sat, 06 Dec 2025   Pseudo R-squ.:                  0.3302
Time:                        03:38:23   Log-Likelihood:                -10756.
converged:                       True   LL-Null:                       -16060.
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                         -822.8170    118.050     -6.970      0.000   -1054.190    -591.444
no_of_weekend_nights             0.1485      0.020      7.512      0.000       0.110       0.187
no_of_week_nights                0.0339      0.012      2.778      0.005       0.010       0.058
type_of_meal_plan               -0.0606      0.024     -2.479      0.013      -0.108      -0.013
required_car_parking_space      -1.6247      0.137    -11.857      0.000      -1.893      -1.356
room_type_reserved              -0.1266      0.015     -8.514      0.000      -0.156      -0.097
lead_time                        0.0159      0.000     60.637      0.000       0.015       0.016
arrival_year                     0.4065      0.059      6.949      0.000       0.292       0.521
arrival_month                   -0.0498      0.006     -7.728      0.000      -0.062      -0.037
repeated_guest                  -3.0722      0.600     -5.116      0.000      -4.249      -1.895
no_of_previous_cancellations     0.2875      0.078      3.675      0.000       0.134       0.441
avg_price_per_room               0.0192      0.001     28.192      0.000       0.018       0.020
no_of_special_requests          -1.4811      0.030    -49.310      0.000      -1.540      -1.422
market_segment_type_Corporate   -0.9136      0.102     -8.950      0.000      -1.114      -0.714
market_segment_type_Offline     -1.7574      0.051    -34.709      0.000      -1.857      -1.658
================================================================================================
```

Figure 49 - Final model summary after removing high p_values

```
Accuracy on training set :  0.8062775677378702
```

- *There is no significant change in the model performance as compared to initial model.*
- *All the insignificant predictor columns have been removed.*
- *Now all the columns left are significant predictors of regression model.*

Coefficient interpretations

- *Coefficient of columns named no_of_weekend_nights, no_of_week_nights, lead_time, arrival_year, no_of_previous_cancellations, avg_price_per_room are positive. An increase in these columns will lead to increase in chances of a customer not cancelling the booking.*

- *Coefficients of columns named type_of_meal_plan, required_car_parking_space, room_type_reserved, arrival_month, repeated_guest, no_of_special_requests, market_segment_type_Corporate, market_segment_type_Offline are negative. Increase in these columns lead to decrease in chances of a customer not cancelling the booking.*

Odds from coefficients

	const	no_of_weekend_nights	no_of_week_nights	type_of_meal_plan	required_car_parking_space	room_type_reserved	lead_time	arrival_year	arrival_month	repeated_guest
odds	0.0	1.160148	1.034506	0.941232	0.196969	0.881046	1.01605	1.501627	0.951464	0.046319

Figure 50 - Change in Odds

Percentage change in odds

	const	no_of_weekend_nights	no_of_week_nights	type_of_meal_plan	required_car_parking_space	room_type_reserved	lead_time	arrival_year	arrival_month
change_odds%	-100.0	16.014806	3.450587	-5.876846	-80.303095	-11.895437	1.60502	50.162716	-4.85361

Figure 51 - Percentage Change in Odds

- ***no_of_weekend_nights:*** *Holding all other features constant, a unit change in no_of_weekend_nights will increase the odds of a customer not cancelling the booking by 1.16 times or a 16.01% increase in the odds of not cancelling the booking.*
- ***no_of_week_nights:*** *Holding all other features constant, a unit change in no_of_week_nights will increase the odds of a customer not cancelling the booking by 1.03 times or a 3.45% increase in the odds of not cancelling the booking.*
- ***type_of_meal_plan:*** *Holding all other features constant, a unit change in type_of_meal_plan will increase the odds of a customer not cancelling the booking by 0.94 times or a 5.87% decrease in the odds of not cancelling the booking.*
- ***required_car_parking_space:*** *Holding all other features constant, a unit change in required_car_parking_space will increase the odds of a customer not cancelling the booking by 0.19 times or a 80.3% decrease in the odds of not cancelling the booking.*
- ***room_type_reserved****: Holding all other features constant, a unit change in room_type_reserved will increase the odds of a customer not cancelling the*

booking by 0.88 times or a 11.89% decrease in the odds of not cancelling the booking.

- ***lead_time:*** *Holding all other features constant, a unit change in lead_time will increase the odds of a customer not cancelling the booking by 1.01 times or a 1.60% increase in the odds of not cancelling the booking.*
- ***arrival_year:*** *Holding all other features constant, a unit change in arrival_year will increase the odds of a customer not cancelling the booking by 1.5 times or a 50.16% increase in the odds of not cancelling the booking.*
- ***arrival_month:*** *Holding all other features constant, a unit change in arrival_month will increase the odds of a customer not cancelling the booking by 0.85 times or a 4.85% decrease in the odds of not cancelling the booking.*
- ***repeated_guest:*** *Holding all other features constant, a unit change in repeated_guest will increase the odds of a customer not cancelling the booking by 0.046 times or a 95.36% decrease in the odds of not cancelling the booking.*
- ***no_of_previous_cancellations****: Holding all other features constant, a unit change in no_of_previous_cancellations will increase the odds of a customer not cancelling the booking by 1.33 times or a 33.3% increase in the odds of not cancelling the booking.*
- ***avg_price_per_room :*** *Holding all other features constant, a unit change in avg_price_per_room will increase the odds of a customer not cancelling the booking by 1.01 times or a 1.93% increase in the odds of not cancelling the booking.*
- ***no_of_special_requests****: Holding all other features constant, a unit change in no_of_special_requests will increase the odds of a customer not cancelling the booking by 0.22 times or a 77.26% decrease in the odds of not cancelling the booking.*
- ***market_segment_type_Corporate:*** *Holding all other features constant, a unit change in market_segment_type_Corporate will increase the odds of a customer not cancelling the booking by 0.40 times or a 59.89% decrease in the odds of not cancelling the booking.*
- ***market_segment_type_Offline****: Holding all other features constant, a unit change in market_segment_type_Offline will increase the odds of a customer not cancelling the booking by 0.17 times or a 82.75% decrease in the odds of not cancelling the booking.*

Performance Metrics of the final Regression model - 'lg6'

Training set Performance

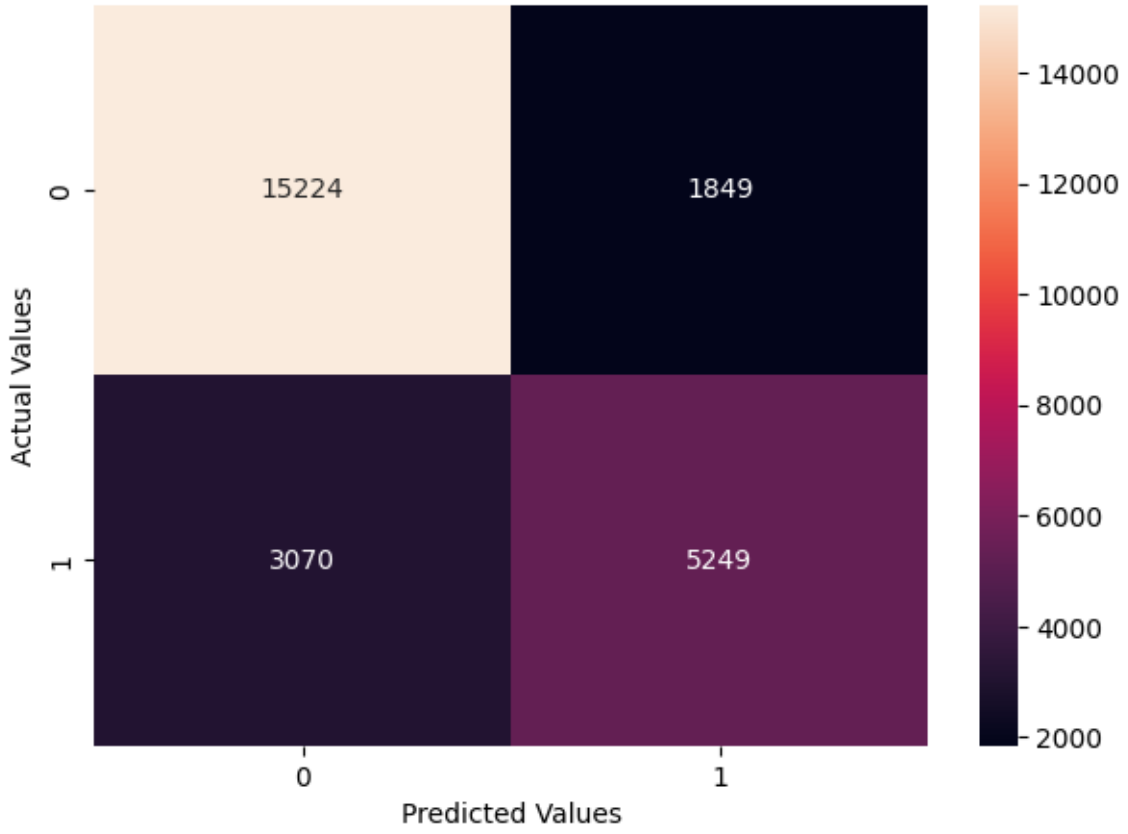


Figure 52 - Confusion Matrix of final model training set

Training performance:

	Accuracy	Recall	Precision	F1
0	0.806278	0.630965	0.739504	0.680937

Figure 53 - Performance Metrics of final model training set

Test set Performance

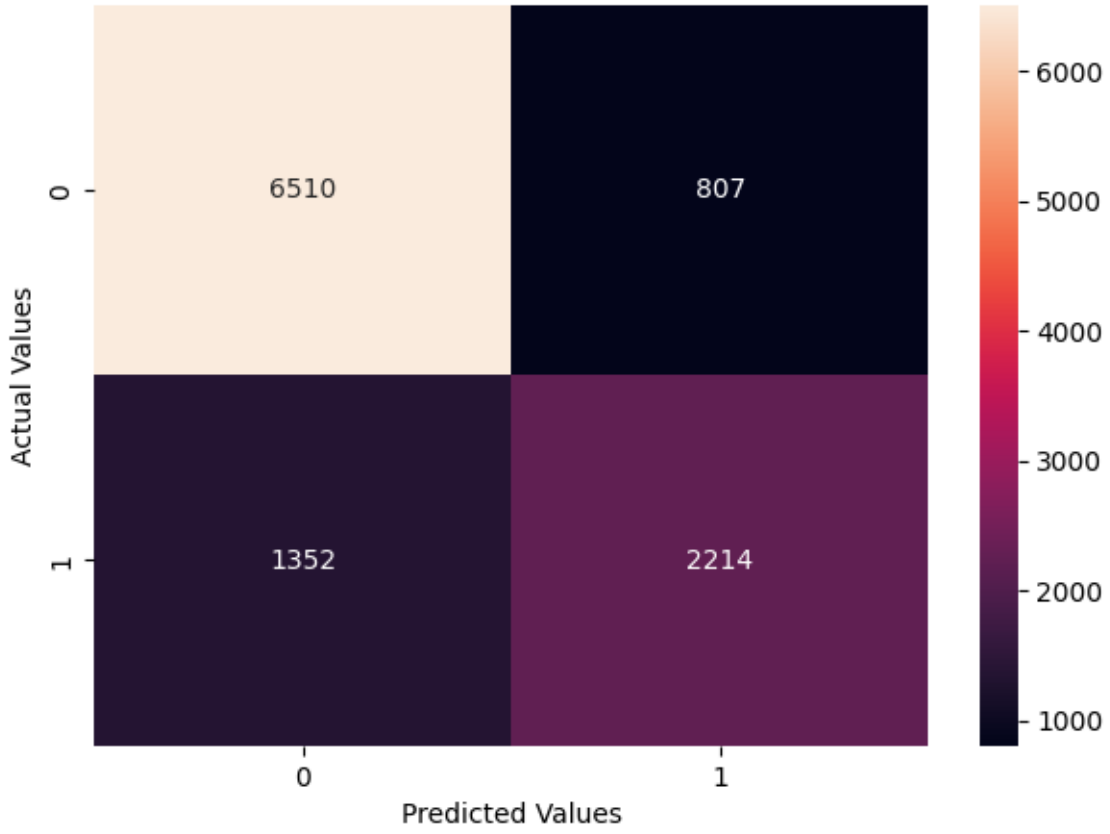


Figure 54 - Confusion Matrix of final model testing set

Test performance:

	Accuracy	Recall	Precision	F1
0	0.801617	0.620864	0.73287	0.672233

Figure 55 - Performance Metrics of final model testing set

- *The model is giving an average f1_score of ~0.680 and ~0.672 on the train and test sets respectively.*
- *As the train and test performances are comparable, the model is not overfitting.*
- *Let's see if the f1_score can be improved further by changing the model threshold.*

ROC Curve and ROC-AUC

Model performance on Training set with Optimal Threshold

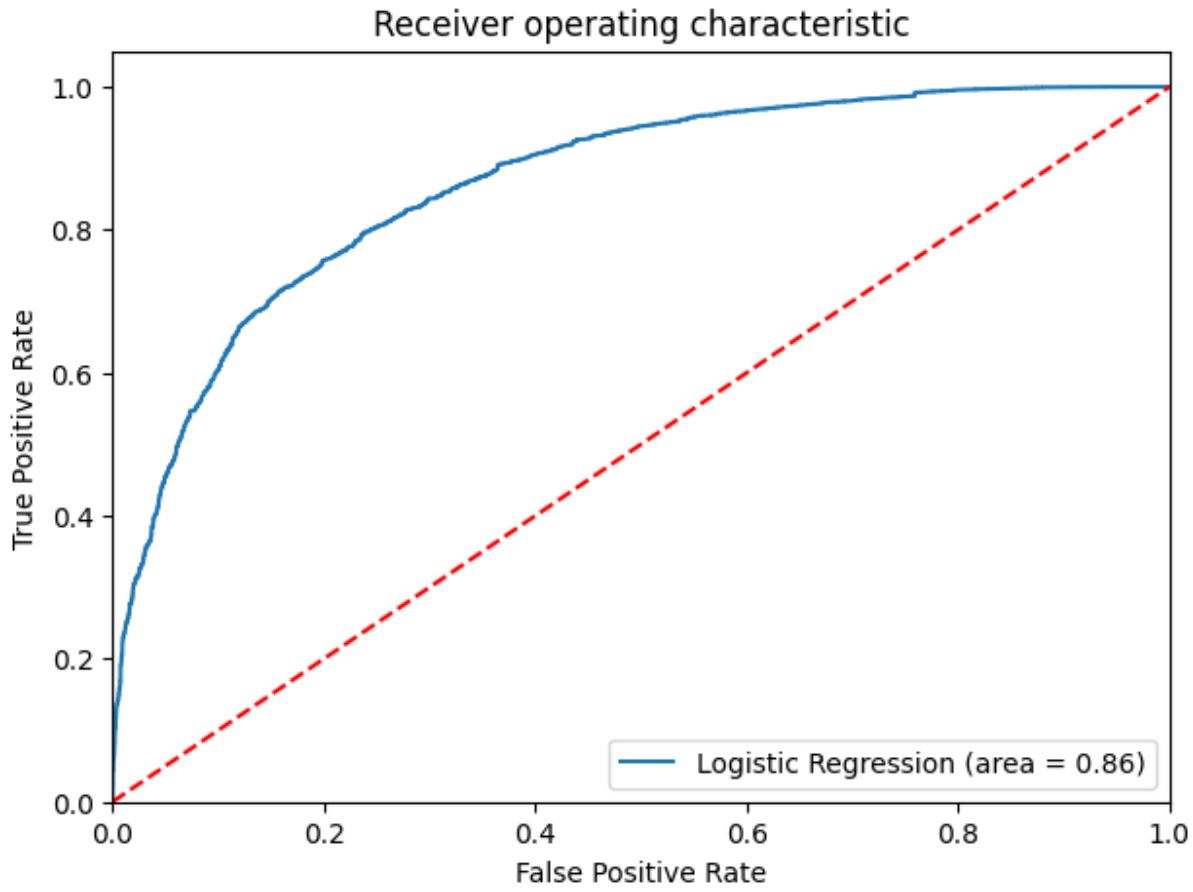


Figure 56 - Roc Curve with Optimal Threshold

- *Logistic Regression model is giving a good performance on training set.*

Optimal threshold using AUC-ROC curve

- ***Optimal Threshold Value: 0.30381746229941814***

Model performance on Training set with Optimal Threshold

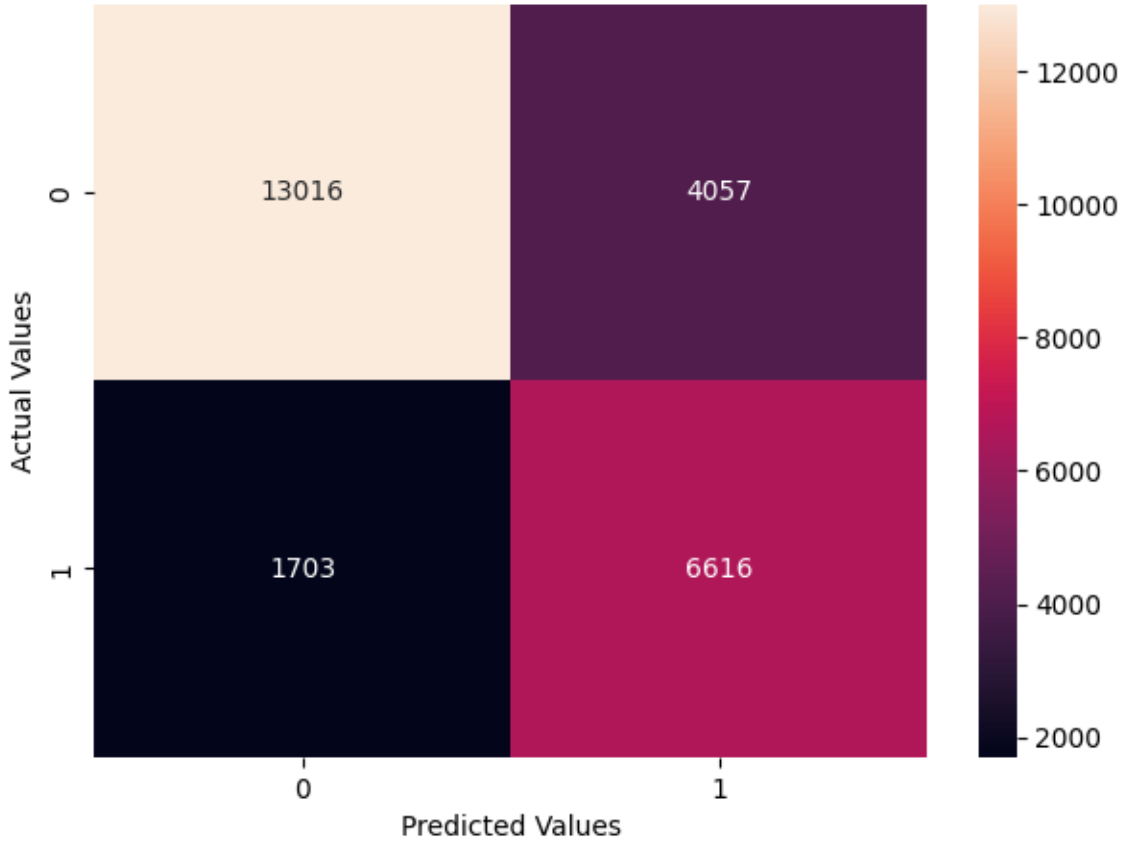


Figure 57 - Confusion matrix training set of final logistic model with threshold > 0.30

Training performance:

	Accuracy	Recall	Precision	F1
0	0.773157	0.795288	0.619882	0.696714

Figure 58 - Performance Matrix training set of final logistic model with threshold > 0.30

- *Recall and F1 of model has increased but the other metrics have reduced.*
- *The model is still giving a good performance.*

Model performance on Testing set with Optimal Threshold

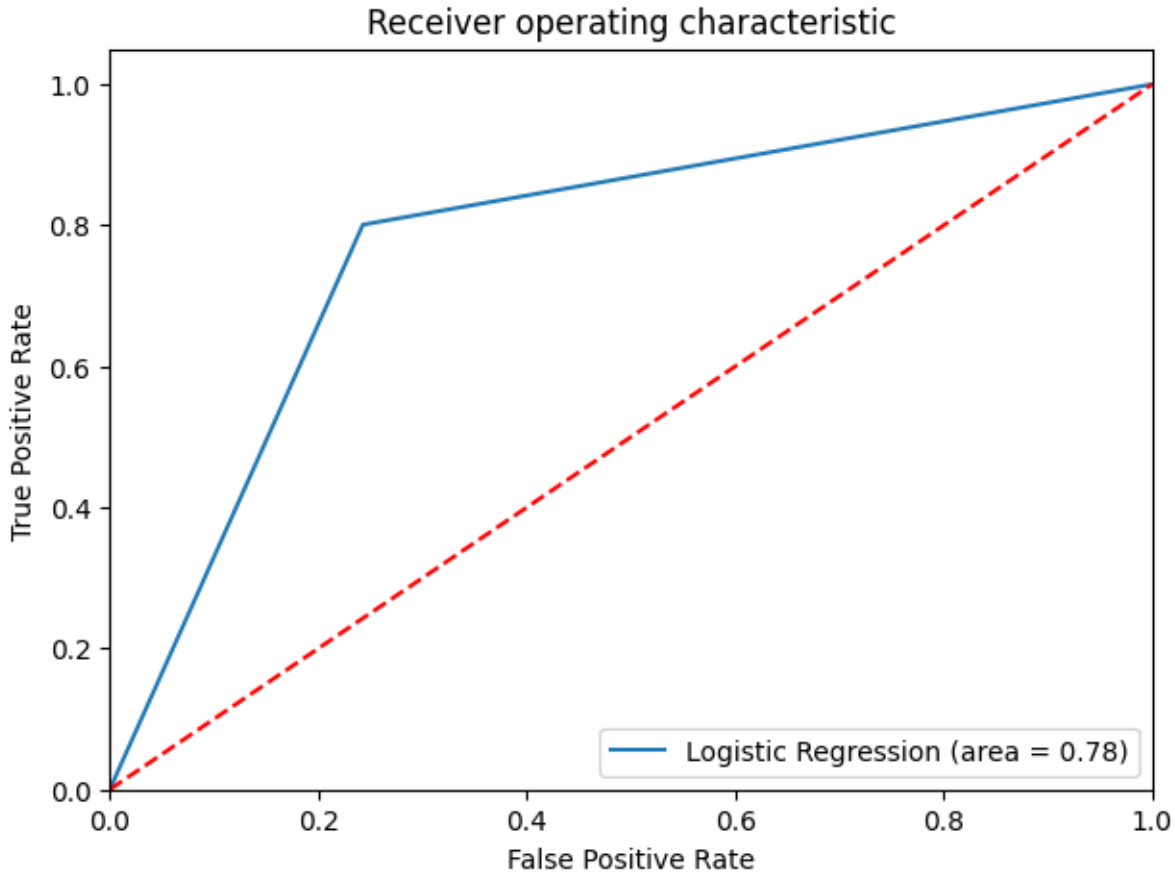


Figure 59 - Confusion matrix testing set of final logistic model with threshold > 0.30

Test performance:

	Accuracy	Recall	Precision	F1
0	0.77157	0.795289	0.617596	0.695268

Figure 60 - Performance Matrix testing set of final logistic model with threshold > 0.30

- *Recall and F1 of model has increased but the other metrics have reduced.*
- *The model is still giving a good performance.*

Precision-Recall Curve

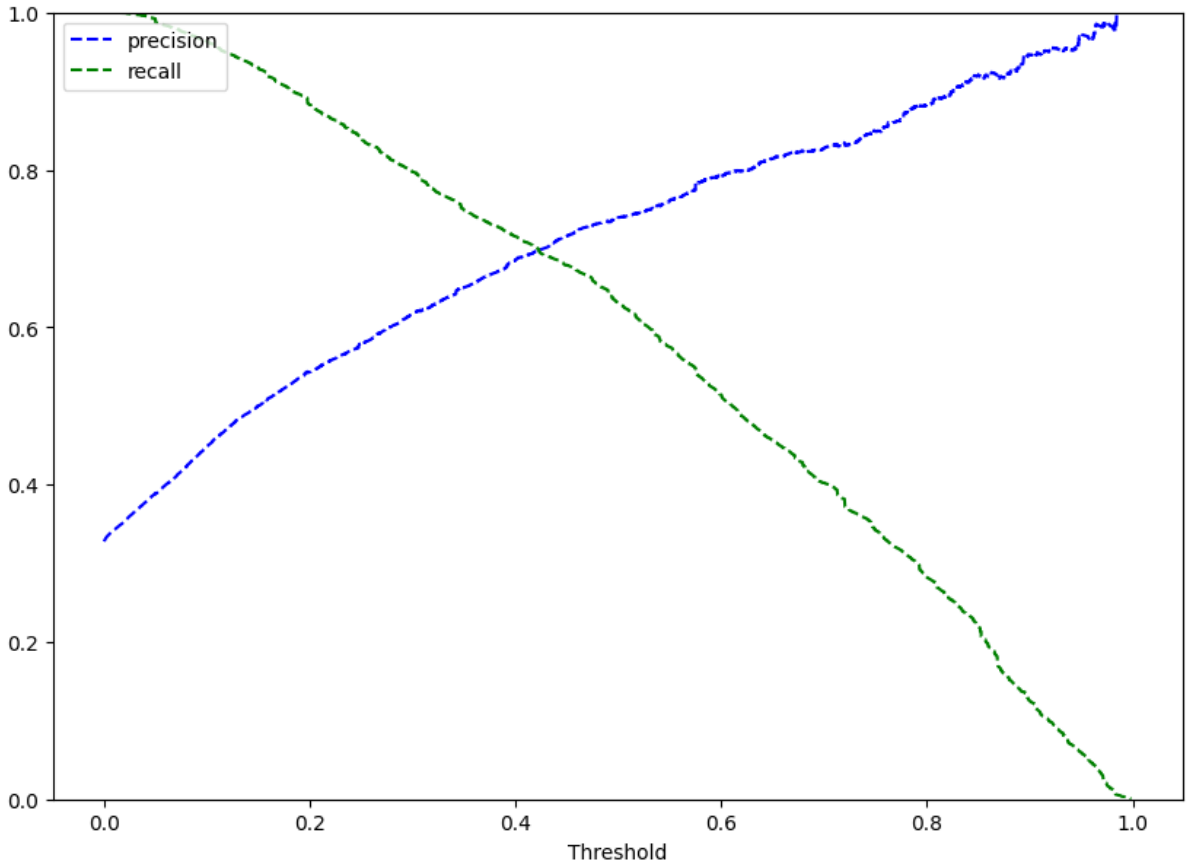


Figure 61 - precision-Recall Curve

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

Model performance on training set with Threshold > 0.42

Training performance:

	Accuracy	Recall	Precision	F1
0	0.801827	0.701046	0.696192	0.69861

Figure 62 - Performance Matrix training set of final logistic model with threshold > 0.42

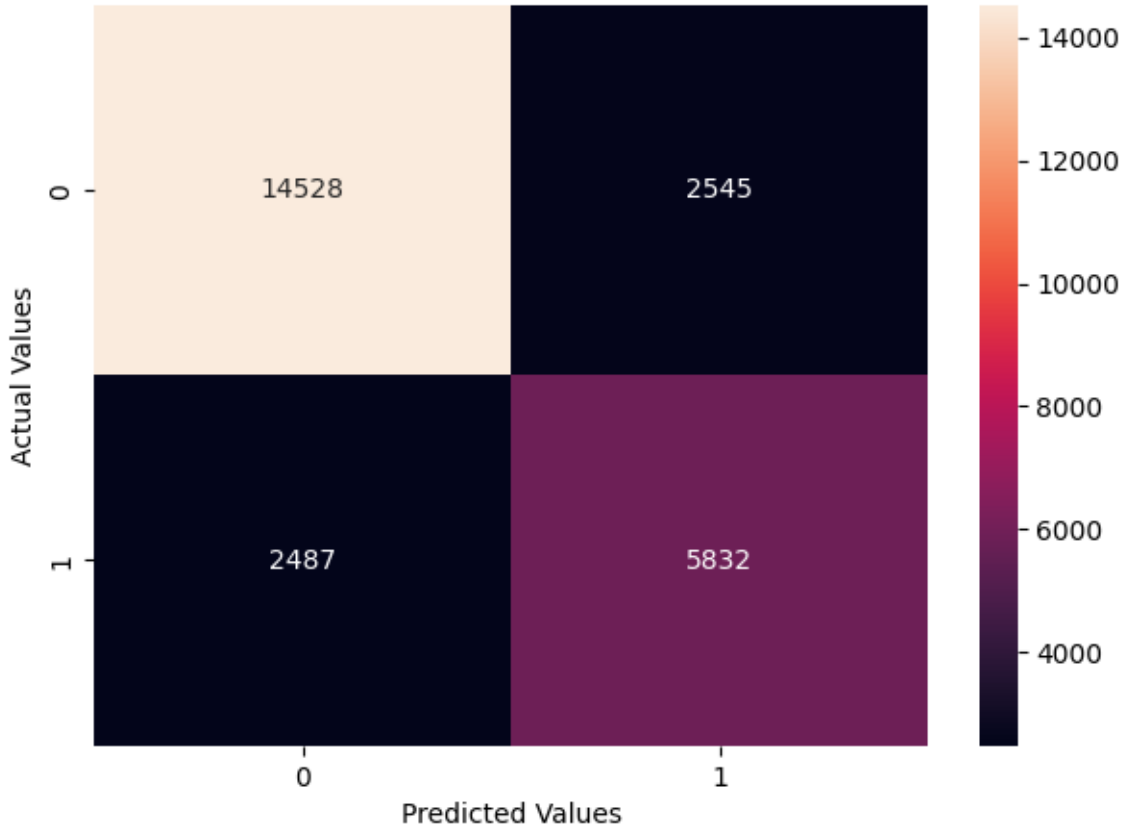


Figure 63 - Confusion matrix training set of final logistic model with threshold > 0.42

- *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.*

Model performance on test set with Threshold > 0.42

Test performance:

	Accuracy	Recall	Precision	F1
0	0.795001	0.695457	0.684138	0.689751

Figure 64 - Performance Matrix testing set of final logistic model with threshold > 0.42

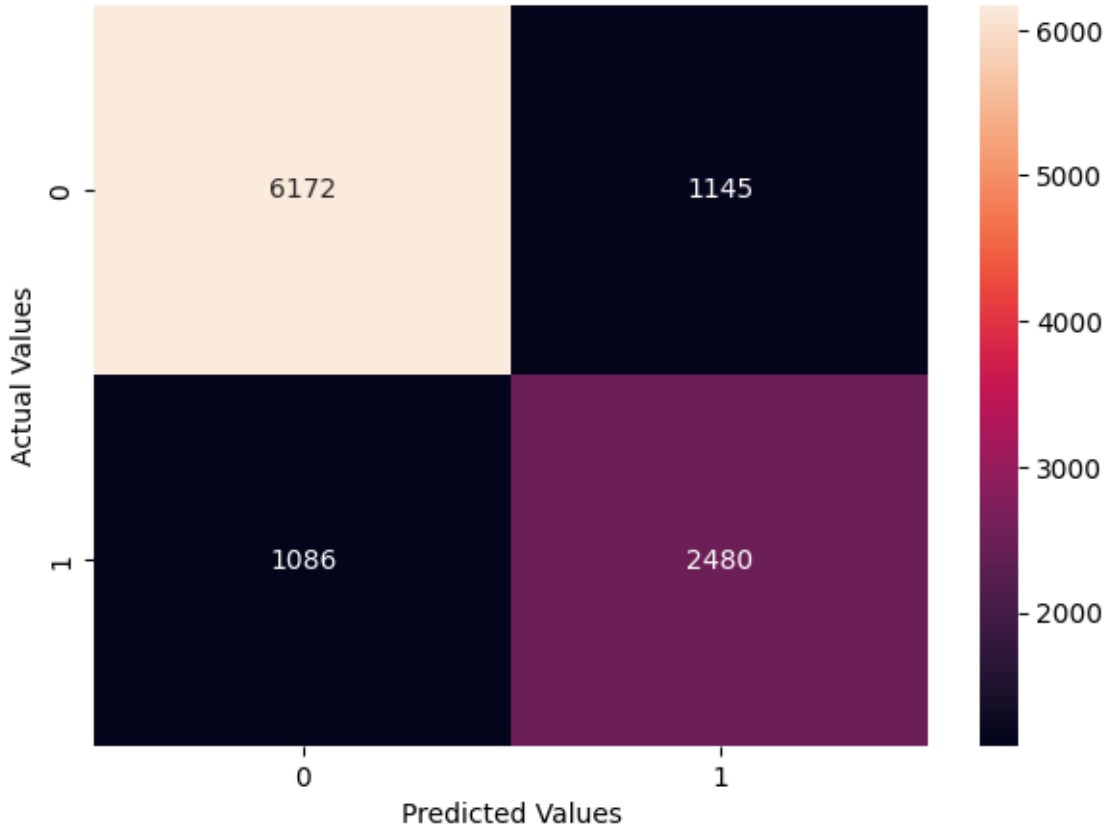


Figure 65 - Confusion matrix testing set of final logistic model with threshold > 0.42

- *Recall and F1 of model has increased but the other metrics have reduced.*
- *Model is performing well on testing 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.*

Decision Tree Model Performance Improvement

Decision Tree with restricted Depth

The Depth is restricted to 3.

Training Set performance for restricted depth model

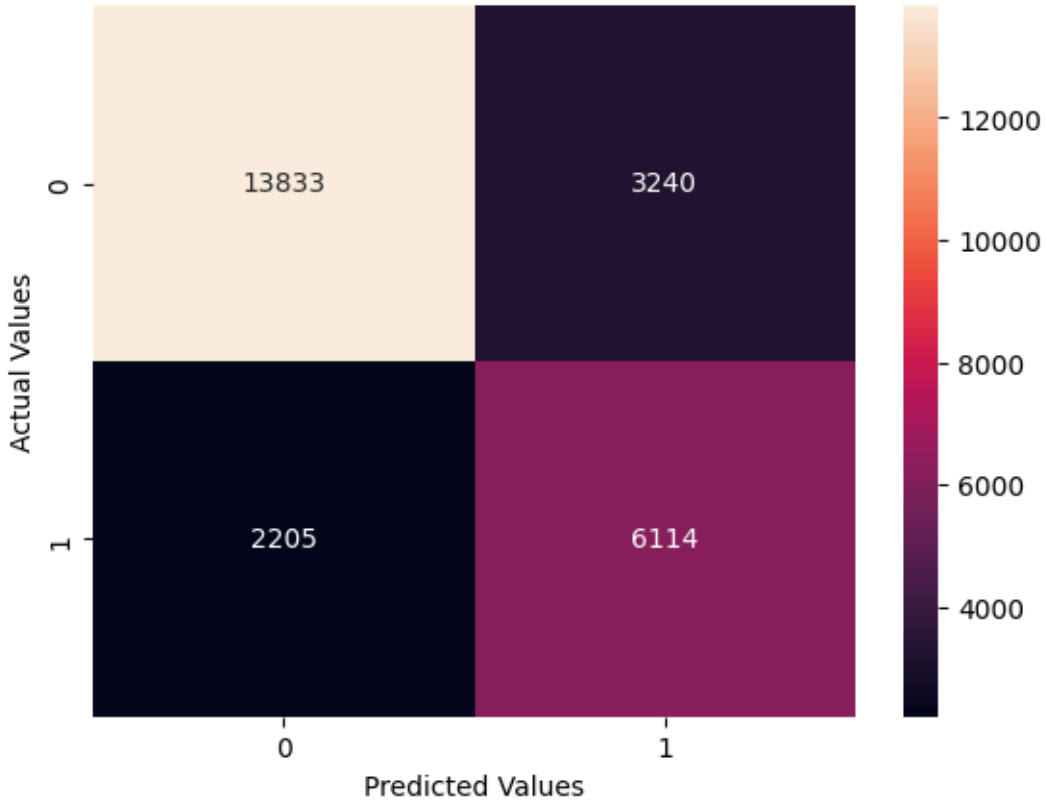


Figure 66 - Confusion Matrix of training set of restricted depth DTree

	Accuracy	Recall	Precision	F1
0	0.785562	0.734944	0.653624	0.691903

Figure 67 - Performance Matrix of Training set of restricted Depth Dtree

- Recall on training set on default model is 0.985.
- Recall on training set on restricted depth model is 0.734.
- Recall on training sets has reduced from 0.985 to 0.734 but this is an improvement because now the restricted depth model is not overfitting and we have a generalized model.

Testing Set performance for restricted depth model

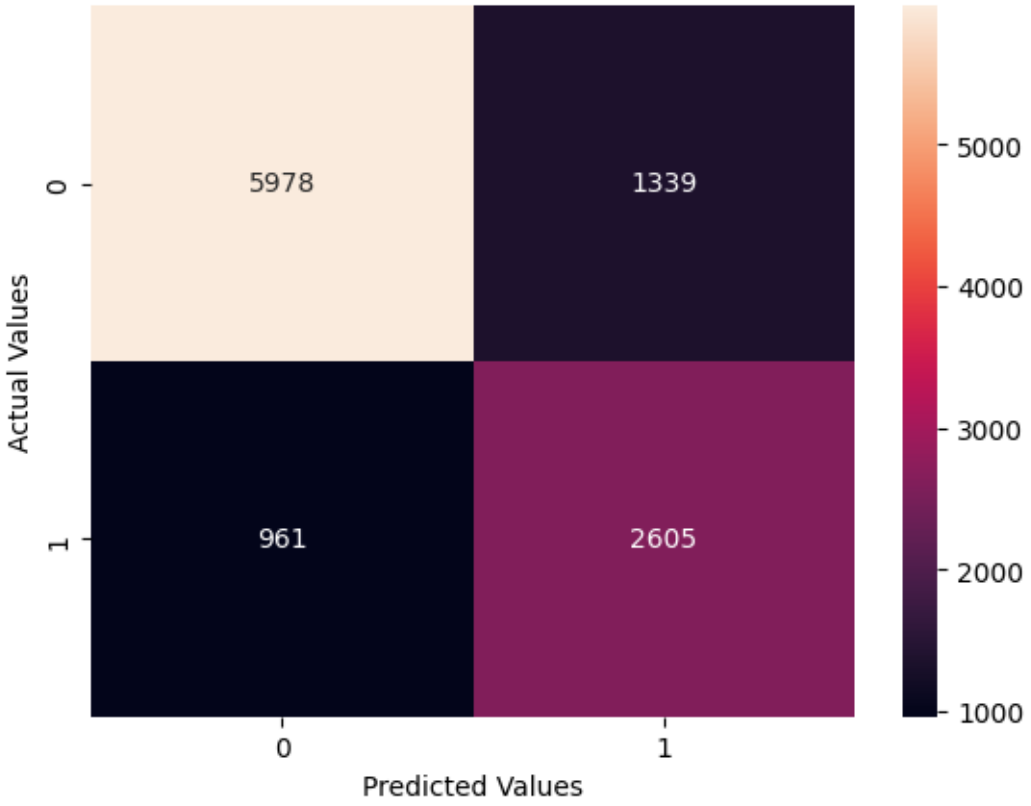


Figure 68 - Confusion Matrix of testing set of restricted depth DTree

	Accuracy	Recall	Precision	F1
0	0.788661	0.73051	0.660497	0.693742

Figure 69 - Performance Matrix of Testing set of restricted Depth Dtree

- *Recall on testing set on default model is 0.785.*
- *Recall on testing set on restricted depth model is 0.730.*
- *Recall on testing sets has reduced from 0.785 to 0.730 but this is an improvement because now the restricted depth model is not overfitting and we have a generalized model.*

Visualizing the Restricted model Decision Tree

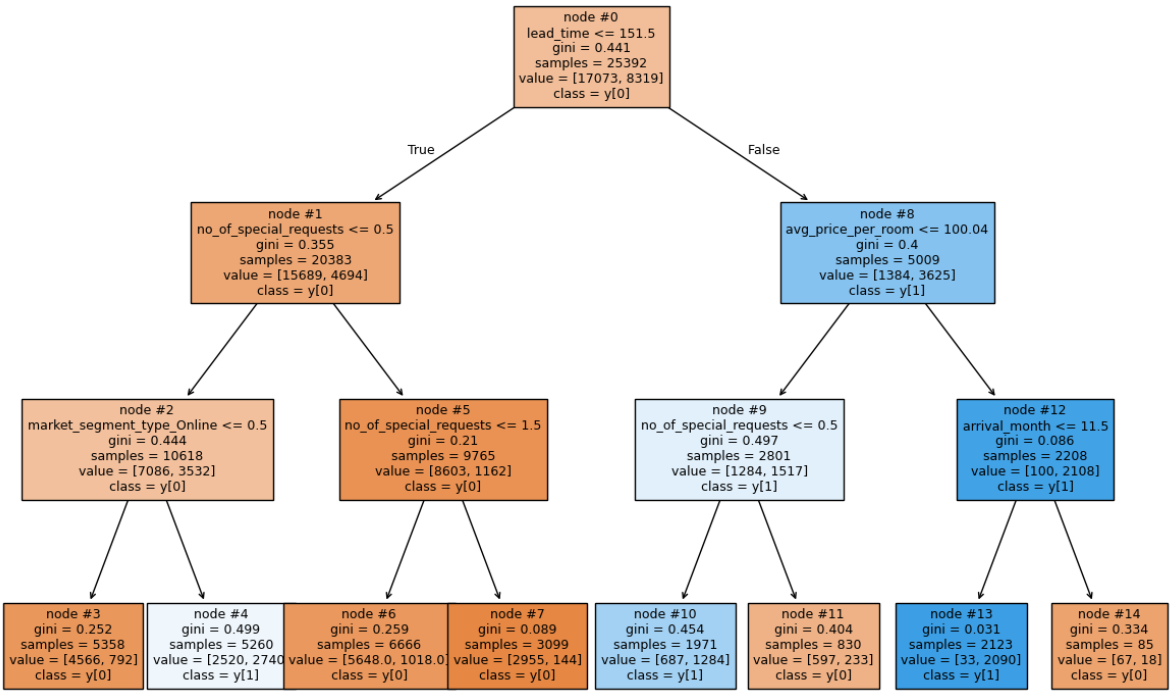


Figure 70 - Decision Tree of Restricted Depth

	Imp
lead_time	0.503486
market_segment_type_Online	0.190039
no_of_special_requests	0.172994
avg_price_per_room	0.108384
arrival_month	0.025098
no_of_weekend_nights	0.000000
no_of_children	0.000000
no_of_adults	0.000000
no_of_week_nights	0.000000
room_type_reserved	0.000000
required_car_parking_space	0.000000
type_of_meal_plan	0.000000
repeated_guest	0.000000
arrival_date	0.000000
arrival_year	0.000000
no_of_previous_bookings_not_canceled	0.000000
no_of_previous_cancellations	0.000000
market_segment_type_Complementary	0.000000
market_segment_type_Corporate	0.000000
market_segment_type_Offline	0.000000

Figure 71 - Feature Significance of Restricted Depth Model

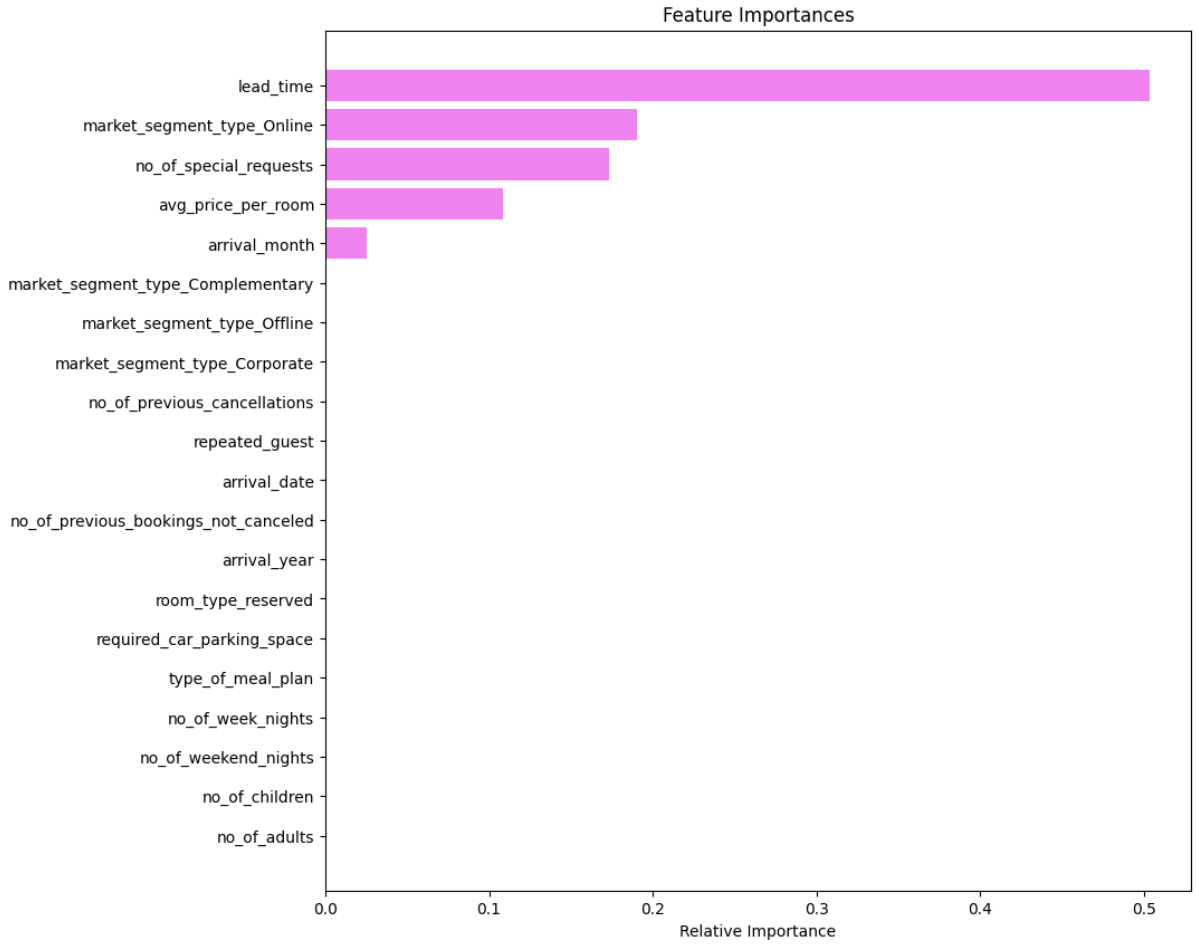


Figure 72 - Feature Importance Chart of Restricted Depth Model

- *Important features of arrival_date, no_of_week_nights was on top in default model, but here importance of arrival_date and no_of_week_nights variable is zero this is the shortcoming of pre pruning, we just limit it even before knowing the importance of features and split.*
- *That's why we will go for pre pruning using grid search, maybe setting max_depth to 3 is not good enough.*
- *It is bad to have a very low depth because your model will underfit.*

Decision Tree Model (Pre-Pruning)

we use GridSearchCV for Hyper parameter tuning for our Decision Tree Model.

Training Set performance for Decision Tree model (Pre-Pruning)

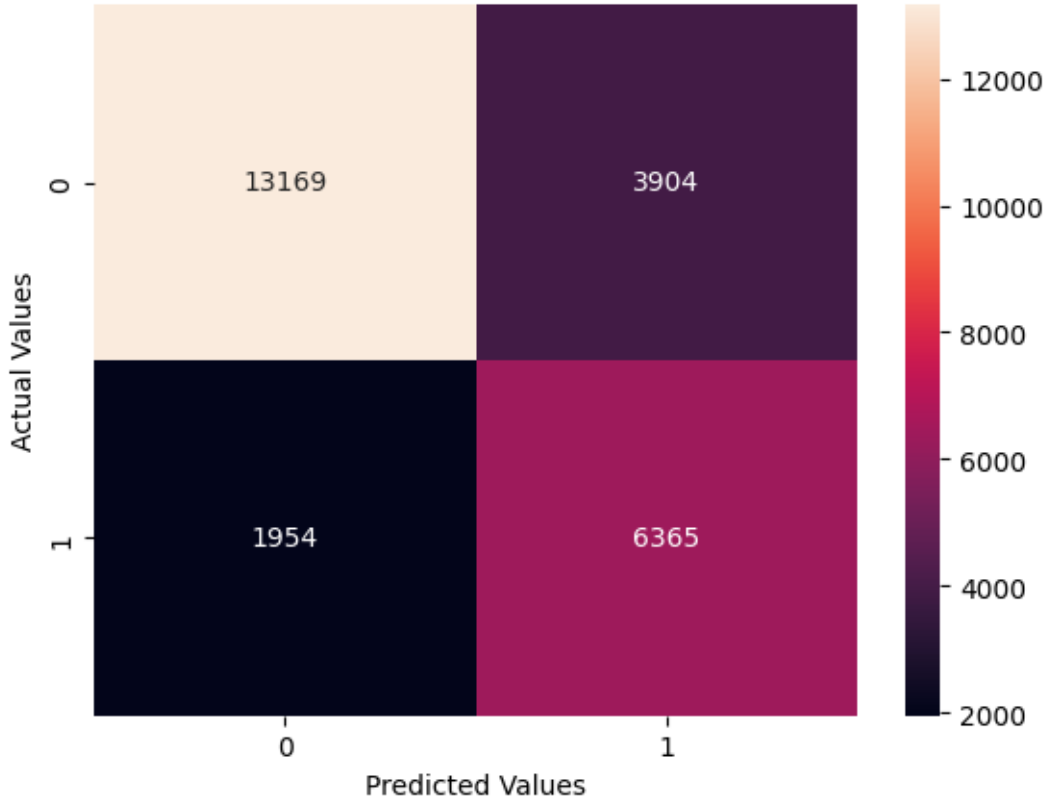


Figure 73 - Confusion Matrix of training set DTree model (Pre-Pruning)

	Accuracy	Recall	Precision	F1
0	0.769297	0.765116	0.619827	0.68485

Figure 74 - Performance Metrics of Training set Dtree model (Pre-Pruning)

- Recall on training set on default model is 0.985.
- Recall on training set on restricted depth model is 0.765.
- Recall on training sets has reduced from 0.985 to 0.765 but this is an improvement because now the Pre_pruning model is not overfitting and we have more generalized model.

Testing Set performance for Decision Tree model (Pre-Pruning)

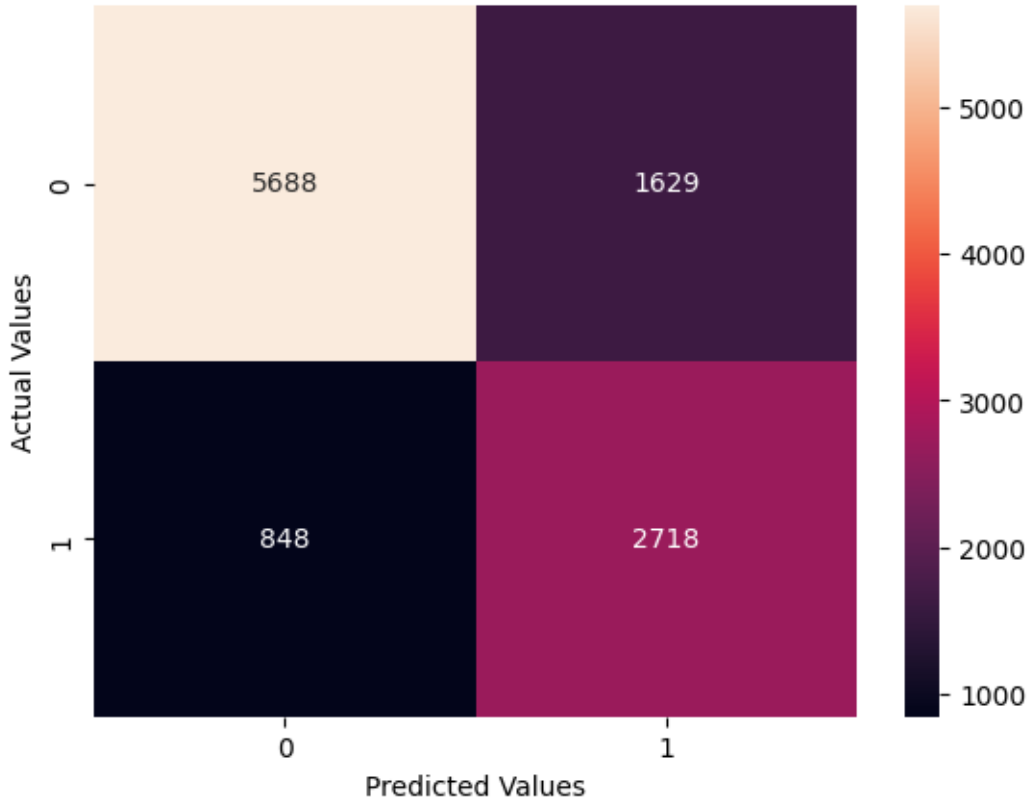


Figure 75 - Confusion Matrix of testing set DTree model (Pre-Pruning)

	Accuracy	Recall	Precision	F1
0	0.772397	0.762199	0.625259	0.686971

Figure 76 - Performance Metrics of Testing set Dtree model (Pre-Pruning)

- Recall on testing set on default model is 0.785.
- Recall on testing set on restricted depth model is 0.762.
- Recall on testing sets has reduced from 0.785 to 0.762 but this is an improvement because now the Pre_Pruning model is not overfitting and we have more generalized model.

Visualizing the Decision Tree model (Pre-Pruning)

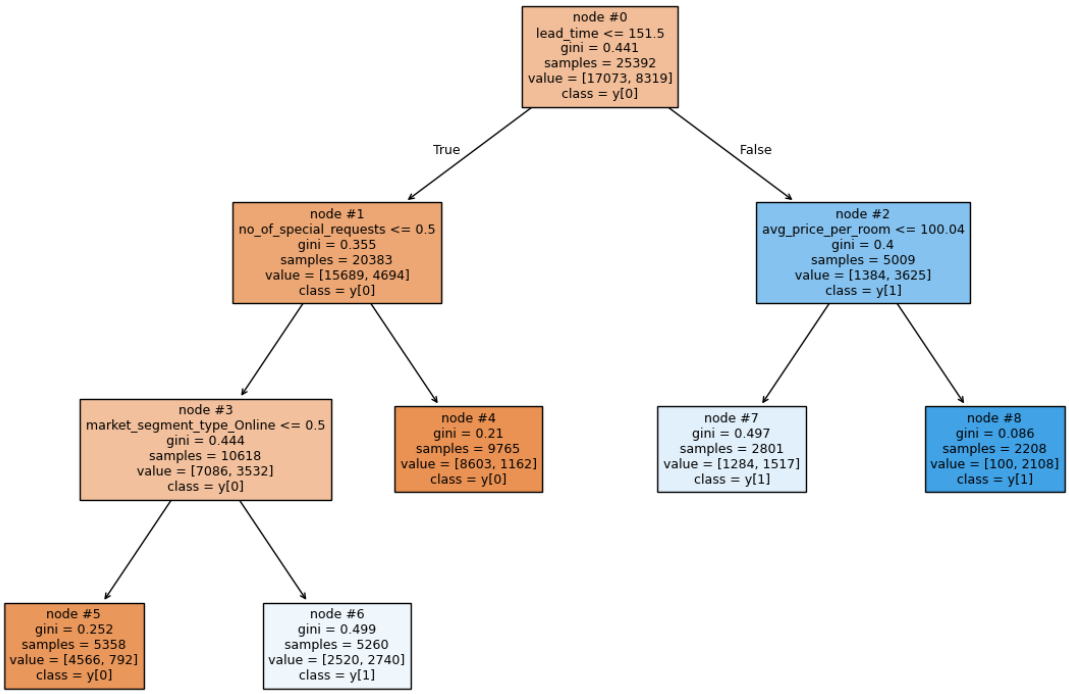


Figure 77 - Decision Tree After Pre-Pruning

	Imp
lead_time	0.546476
market_segment_type_Online	0.206265
no_of_special_requests	0.129621
avg_price_per_room	0.117638
no_of_week_nights	0.000000
no_of_weekend_nights	0.000000
no_of_children	0.000000
no_of_adults	0.000000
arrival_year	0.000000
type_of_meal_plan	0.000000
required_car_parking_space	0.000000
room_type_reserved	0.000000
repeated_guest	0.000000
arrival_date	0.000000
arrival_month	0.000000
no_of_previous_bookings_not_canceled	0.000000
no_of_previous_cancellations	0.000000
market_segment_type_Complementary	0.000000
market_segment_type_Corporate	0.000000
market_segment_type_Offline	0.000000

Figure 78 - Feature Importance Values after Pre-Pruning

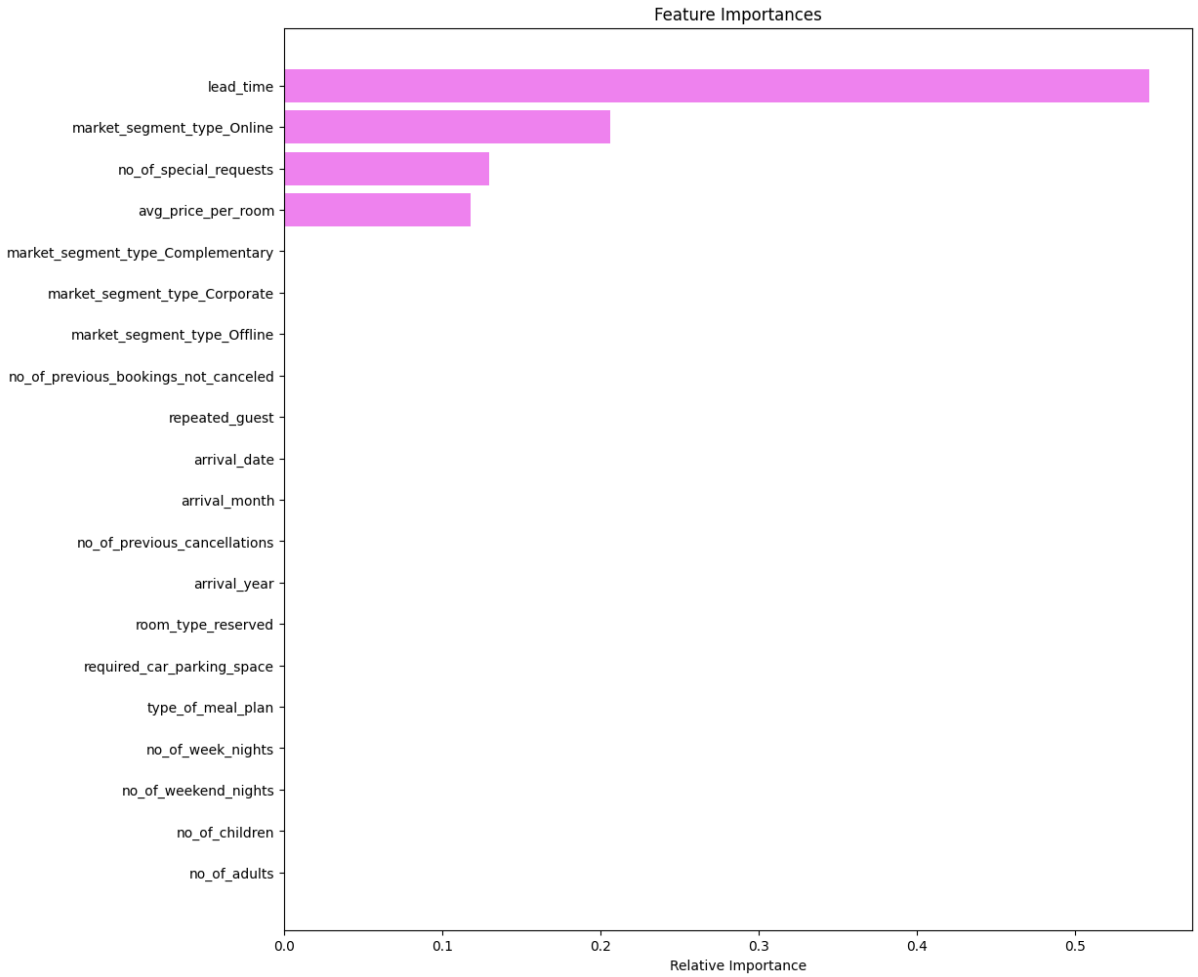


Figure 79 - Feature Importance Chart after Pre-Pruning

- You can see in important features of previous model is same but the Decision Tree of pre-pruned is the best than Depth limiting model.
- This shows that hyperparameter tuning using Grid Search is better than randomly limiting a Hyperparameter.

Decision Tree Model (Post-Pruning)

we use Cost Complexity Pruning method for Post-Pruning.

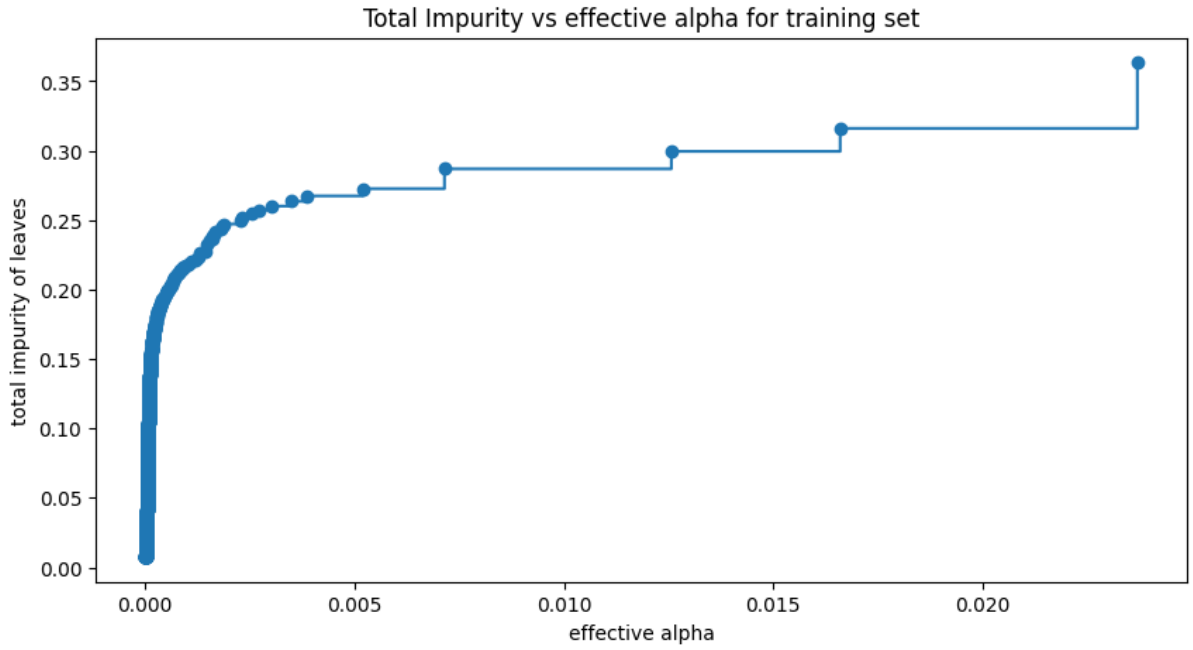
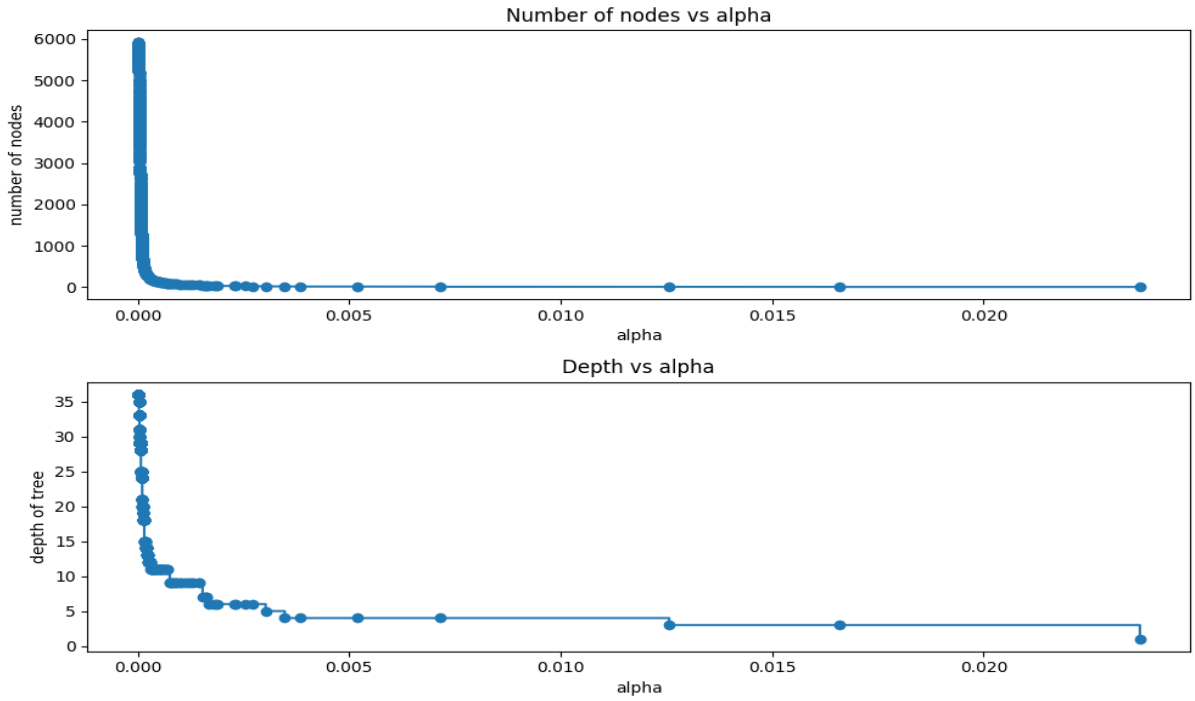


Figure 80 - Total Impurity vs effective alpha for training set

Number of nodes in the last tree is: 1 with ccp_alpha: 0.07710217431082483



- *We can see a decreasing trend when Alpha value Increases.*

Accuracy vs alpha for training and testing sets

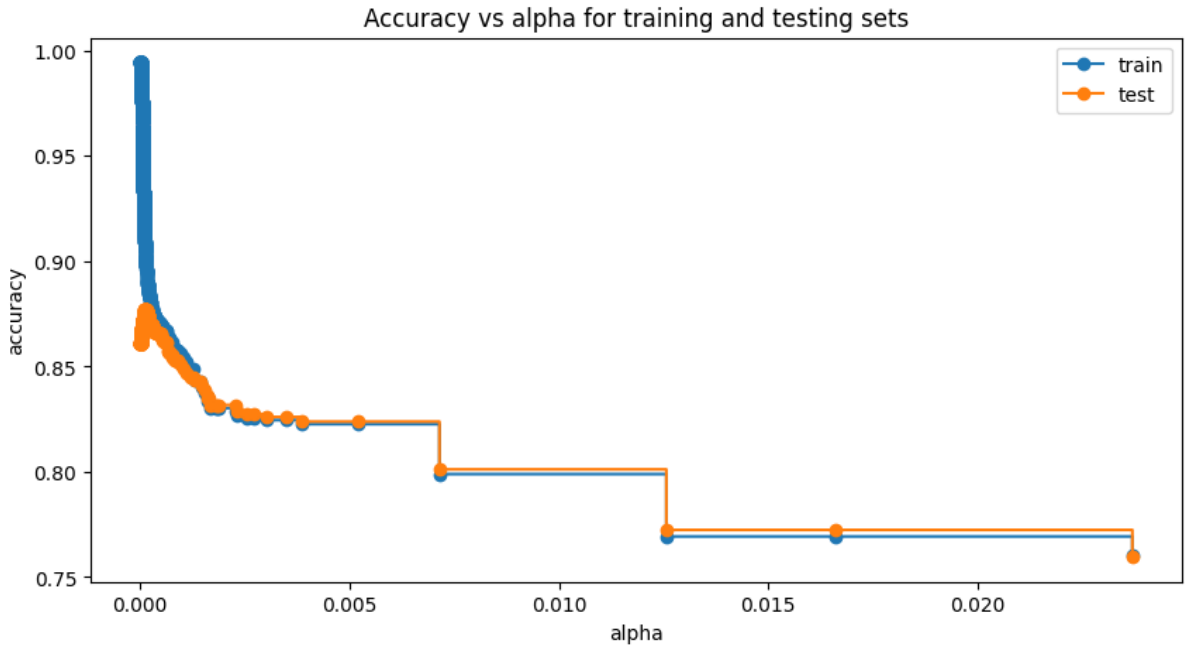


Figure 81 - Accuracy VS Alpha plot

- Training accuracy of best model: 0.902882797731569
- Test accuracy of best model: 0.8774235045483783

Since accuracy isn't the right metric for our data we would want high recall.

Recall vs alpha for training and testing sets

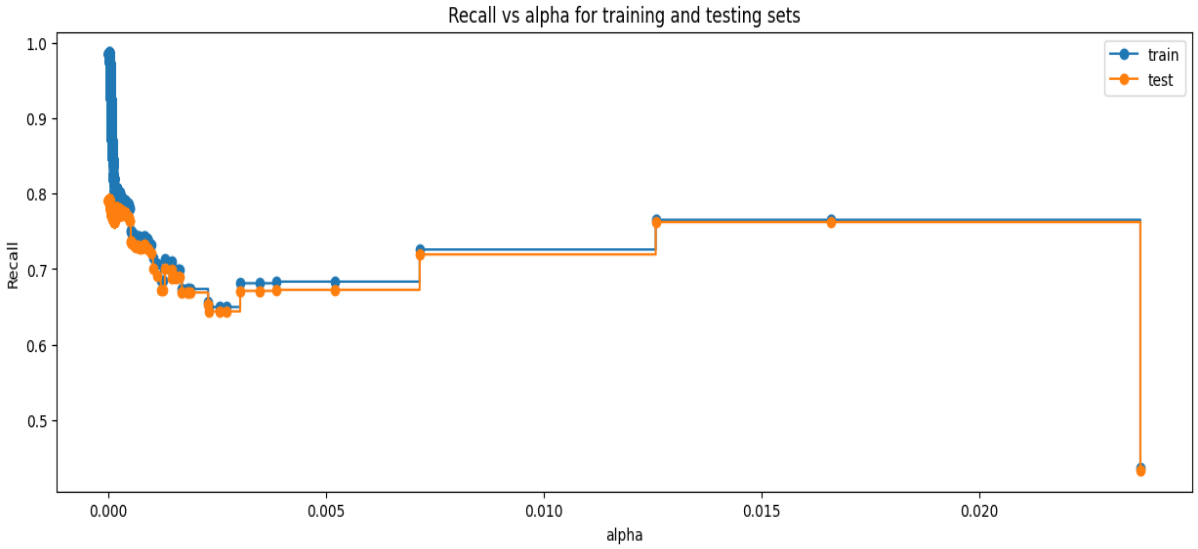


Figure 82 - Recall Vs Alpha Plot

- Training accuracy of best model: 0.9842529150138237
- Testing accuracy of best model: 0.7916432978126753

Testing set performance of Decision Tree (Post-Pruning)

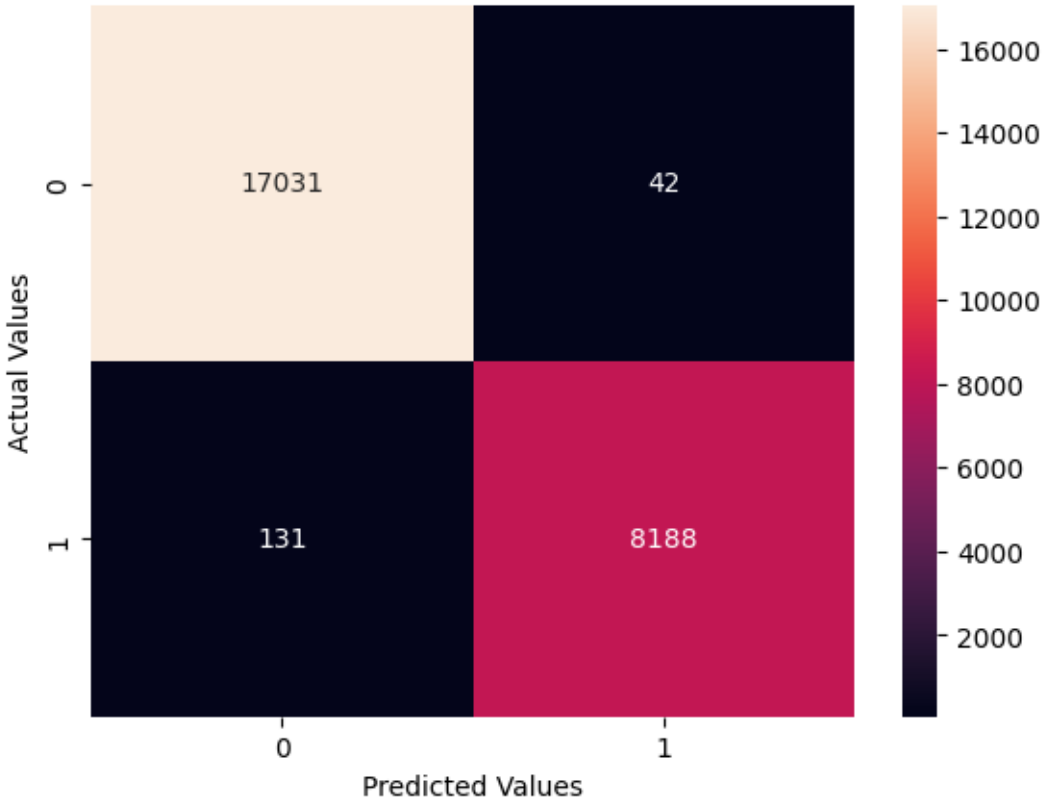


Figure 83 - Confusion Matrix of training set DTree model (Post-Pruning)

Accuracy	Recall	Precision	F1
0.993187	0.984253	0.994897	0.989546

Figure 84 - Performance Metrics of Training set Dtree model (post-pruning)

Testing set performance of Decision Tree (Post-Pruning)

Accuracy	Recall	Precision	F1
0.861619	0.791643	0.787228	0.78943

Figure 85 - Performance Metrics of Testing set Dtree model (post-pruning)

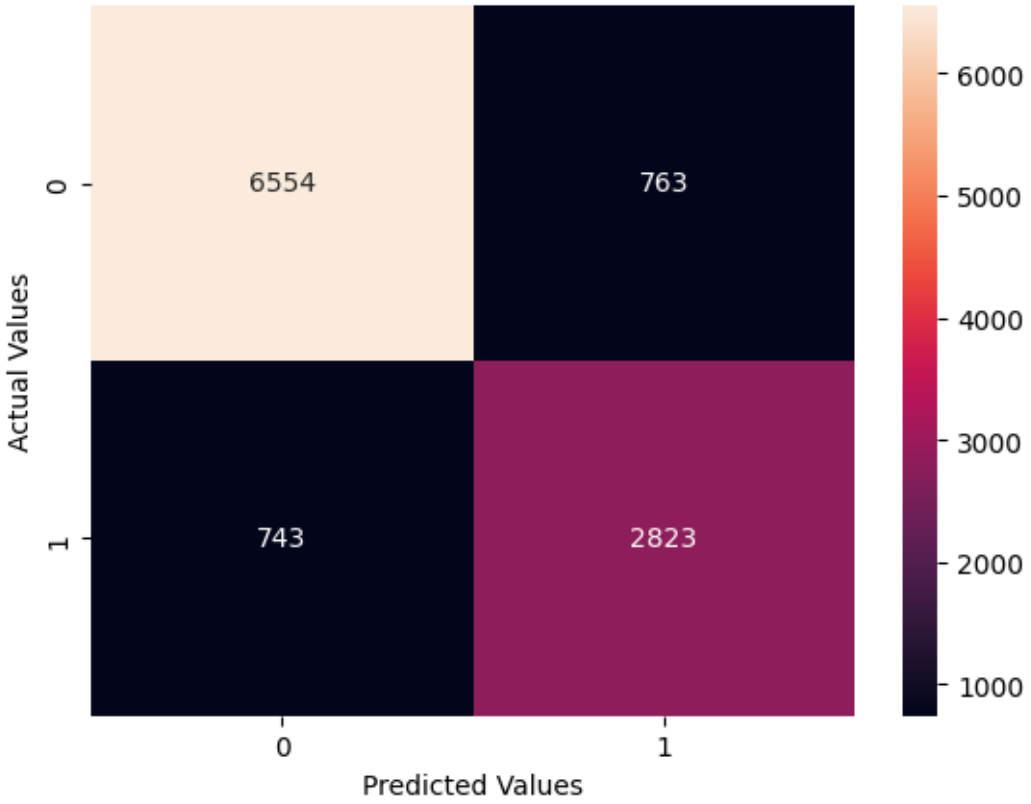


Figure 86 - Confusion Matrix of Testing set DTree model (Post-Pruning)

Visualization of Decision Tree (Post-Pruning)

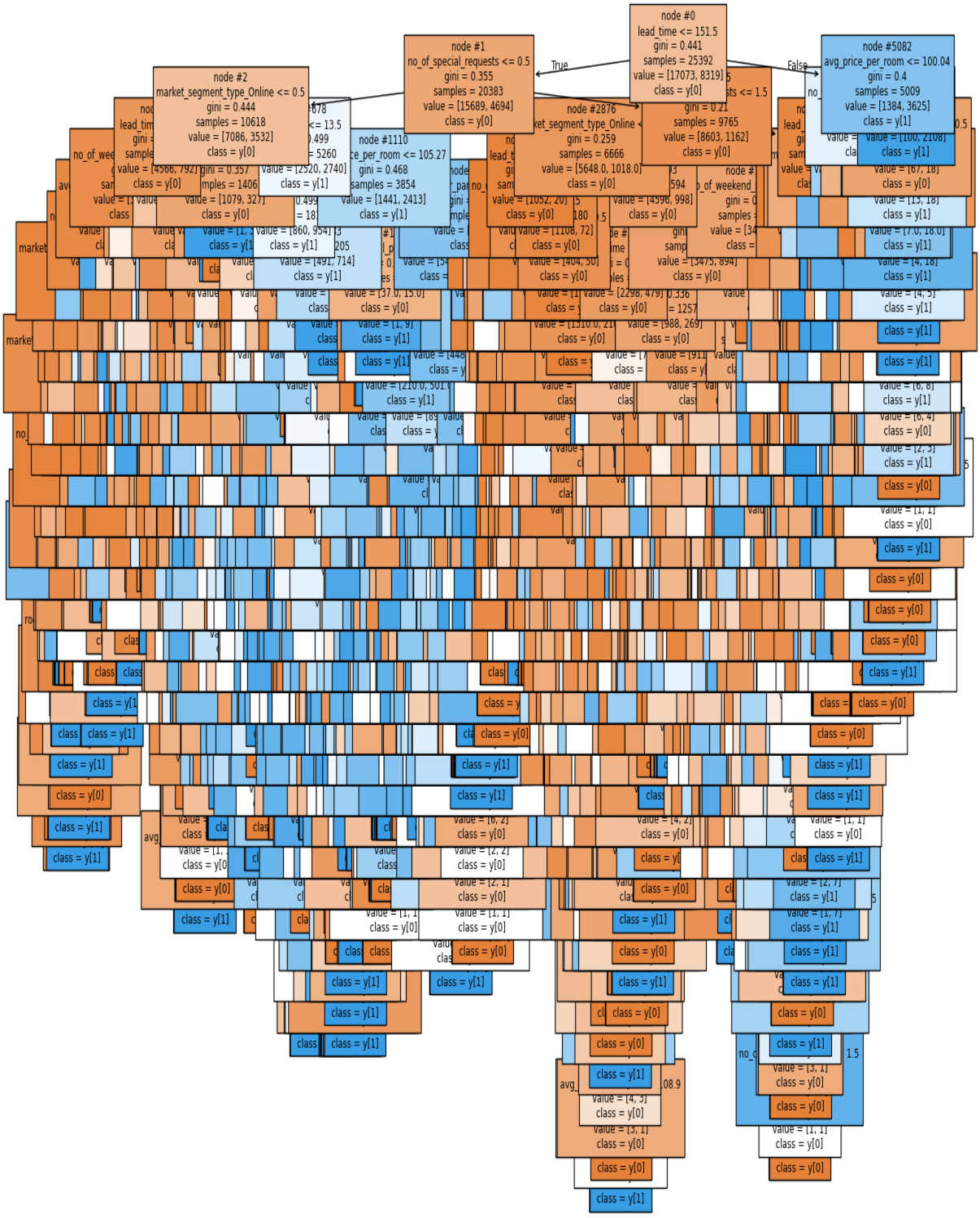


Figure 87 - Decision Tree after Post-Pruning

	Imp
lead_time	0.346710
avg_price_per_room	0.175029
market_segment_type_Online	0.087476
arrival_date	0.083405
no_of_special_requests	0.071745
arrival_month	0.069930
no_of_week_nights	0.045850
no_of_adults	0.034507
no_of_weekend_nights	0.032606
arrival_year	0.014552
type_of_meal_plan	0.012460
room_type_reserved	0.008720
required_car_parking_space	0.006734
no_of_children	0.005017
market_segment_type_Offline	0.002537
market_segment_type_Corporate	0.001806
no_of_previous_bookings_not_canceled	0.000659
repeated_guest	0.000256
no_of_previous_cancellations	0.000000
market_segment_type_Complementary	0.000000

Figure 88 - Feature Importances Values

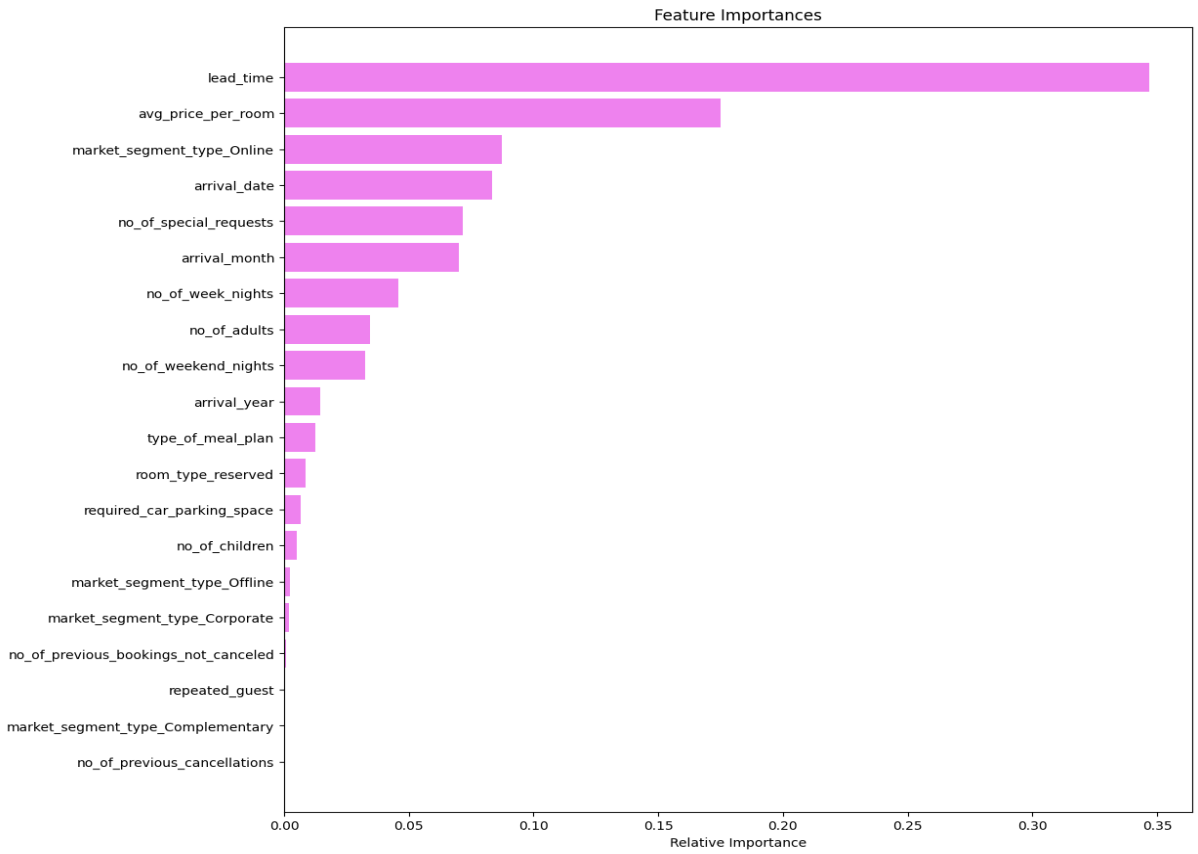


Figure 89 - Post-Pruning Feature Importance chart

- Most the features are captured in the Post-Pruning model showing the significance of features in predicting the dependent variable.

Model Performance Comparison and Final Model Selection

Training set performance comparison

Training performance comparison:

	Logistic Regression-default Threshold (0.5)	Logistic Regression-0.30 Threshold	Logistic Regression-0.42 Threshold	Decision Tree-without_Pruning	Decision Tree with Restricted Depth to 3	Decision Tree with Pre-Pruning	Decision Tree with Post-Pruing
Accuracy	0.806278	0.773157	0.801827	0.994368	0.785562	0.769297	0.993187
Recall	0.630965	0.795288	0.701046	0.985695	0.734944	0.765116	0.984253
Precision	0.739504	0.619882	0.696192	0.997082	0.653624	0.619827	0.994897
F1	0.680937	0.696714	0.698610	0.991356	0.691903	0.684850	0.989546

Figure 90 - Comparison of All models Training Set

Testing set performance comparison

Testing performance comparison:

	Logistic Regression-default Threshold (0.5)	Logistic Regression-0.30 Threshold	Logistic Regression-0.42 Threshold	Decision Tree-without_Pruning	Decision Tree with Restricted Depth to 3	Decision Tree with Pre-Pruning	Decision Tree with Post-Pruing
Accuracy	0.801617	0.771570	0.795001	0.861068	0.788661	0.772397	0.861619
Recall	0.620864	0.795289	0.695457	0.790802	0.730510	0.762199	0.791643
Precision	0.732870	0.617596	0.684138	0.786392	0.660497	0.625259	0.787228
F1	0.672233	0.695268	0.689751	0.788591	0.693742	0.686971	0.789430

Figure 91 - Comparison of All models Testing Set

- *Almost all the models are performing well on both training and test data without the problem of overfitting.*
- ***The Decision Tree model with Post-Pruning is giving the best F1 score. Therefore it can be selected as the final model.***