Machine Learning
Business report

Context:

A significant number of hotel bookings are called off due to cancellations or no-shows. The typical reasons for cancellations include change of plans, scheduling conflicts, etc. This is often made easier by the option to do so free of charge or preferably at a low cost which is beneficial to hotel guests but it is a less desirable and possibly revenue-diminishing factor for hotels to deal with. Such losses are particularly high on last-minute cancellations.

The new technologies involving online booking channels have dramatically changed customers' booking possibilities and behavior. This adds a further dimension to the challenge of how hotels handle cancellations, which are no longer limited to traditional booking and guest characteristics.

The cancellation of bookings impact a hotel on various fronts:

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

Contents

Objective:

The increasing number of cancellations calls for a Machine Learning based solution that can help in predicting which booking is likely to be canceled. INN Hotels Group has a chain of hotels in Portugal, they are facing problems with the high number of booking cancellations and have reached out to your firm for data-driven solutions. You as a data scientist have to analyze the data provided to find which factors have a high influence on booking cancellations, build a predictive model that can predict which booking is going to be canceled in advance, and help in formulating profitable policies for cancellations and refunds.

Data Dictionary:

Booking_ID: the unique identifier of each booking

no_of_adults: Number of adults

no_of_children: Number of Children

no_of_weekend_nights: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel

no_of_week_nights: Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel

type_of_meal_plan: Type of meal plan booked by the customer:

Not Selected – No meal plan selected

Meal Plan 1 – Breakfast

Meal Plan 2 – Half board (breakfast and one other meal)

Meal Plan 3 – Full board (breakfast, lunch, and dinner)

required_car_parking_space: Does the customer require a car parking space? (0 - No, 1- Yes)

room_type_reserved: Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels Group

lead_time: Number of days between the date of booking and the arrival date

arrival_year: Year of arrival date

arrival_month: Month of arrival date

arrival_date: Date of the month

market_segment_type: Market segment designation.

repeated_guest: Is the customer a repeated guest? (0 - No, 1- Yes)

no_of_previous_cancellations: Number of previous bookings that were canceled by the customer prior to the current booking

no_of_previous_bookings_not_canceled: Number of previous bookings not canceled by the customer prior to the current booking

avg_price_per_room: Average price per day of the reservation; prices of the rooms are dynamic. (in euros)

no_of_special_requests: Total number of special requests made by the customer (e.g. high floor, view from the room, etc)

booking_status: Flag indicating if the booking was canceled or not.

Important questions:

1. What are the busiest months in the hotel?
2. Which market segment do most of the guests come from?
3. Hotel rates are dynamic and change according to demand and customer demographics. What are the differences in room prices in different market segments?
4. What percentage of bookings are canceled?
5. Repeating guests are the guests who stay in the hotel often and are important to brand equity. What percentage of repeating guests cancel?
6. Many guests have special requirements when booking a hotel room. Do these requirements affect booking cancellation?

Rubric:

1. Exploratory Data Analysis

- Problem definition - Univariate analysis - Bivariate analysis - Use appropriate visualizations to identify the patterns and insights - Answers to EDA questions provided - Key meaningful observations on individual variables and the relationship between variables.

2. Data preprocessing

- Missing Value Treatment (with rationale if needed) - Outlier Detection and Treatment (with rationale if needed) - Feature Engineering (with rationale if needed) - Data Scaling (with rationale if needed) - Train-test split.

3. Model building

- Choose the metric to optimize for the problem - Build the following models - Logistic Regression (statsmodels) - KNN Classifier (sklearn) - Naive-Bayes Classifier (sklearn) - Decision Tree Classifier (sklearn) - Check and comment on model performance across different metrics.

4. Model Performance Improvement

- Tune the following models to improve performance - Logistic Regression (deal with multicollinearity, remove high p-value variables, determine optimal threshold using ROC curve) - KNN Classifier - Decision Tree Classifier (pre-pruning or post-pruning) - Check and comment on tuned model performance across different metrics.

5. Model Performance Comparison and Final Model Selection

- Compare all the models and choose the best model - Comment on all the model performance and provide rationale for selecting the best model.

6. Actionable Insights & Recommendations

- Actionable insights and recommendations

Exploratory Data Analysis:

– The Name of the dataset is “INNHotelsGroup.csv”.

– The Dataset contain a total of 36275 rows and 19 columns.

– There are no Duplicates present in the Data.

– There are no missing values present in the dataset.

– There some irregularities in outliers present in the dataset. In no_of_children and avg_price_per_room.

– There are 4 Categorical Columns: type_of_meal_plan, market_segement_type booking_status room_type_reserved, season.

– And there are 5 Numerical column: no_of_adults, no_of_children, no_of_weekend_nights, lead_time, no_of_previous_cancellations, no_of_previous_bookings_not_canceled, avg_price_per_room, no_of_special_requests.

– The Datatypes present are :

- Int64.
- Object.
- Float64.

Univariate Analysis**:**

Categorical Variables:

1. **No_of_adults:**

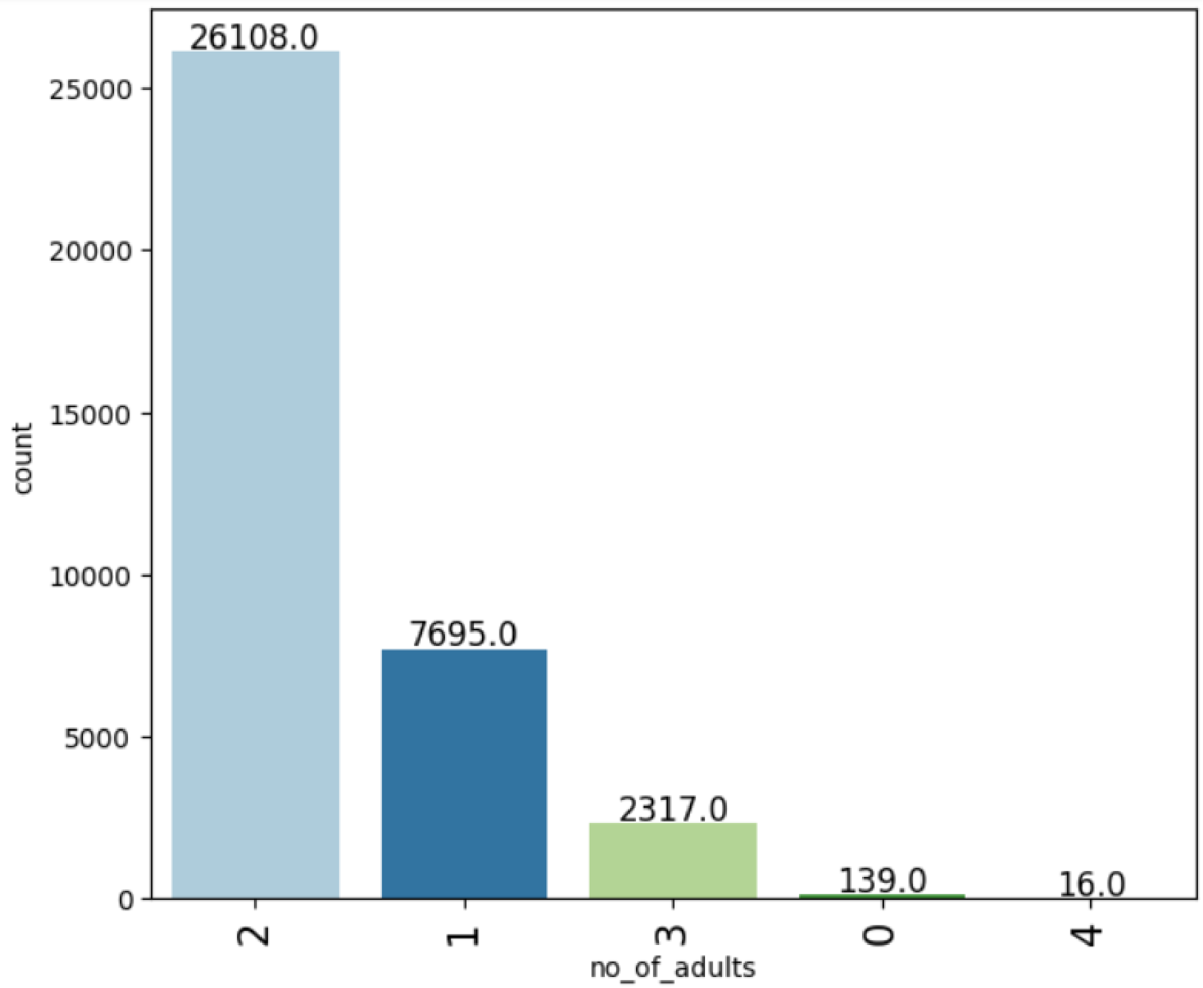


Figure 1: no_of_adults

- Most of the bookings are done for 2 adults and the no_of_adults value 2 is more in number.

- After that most bookings is done for 1 adult.

2. **no_of_children:**

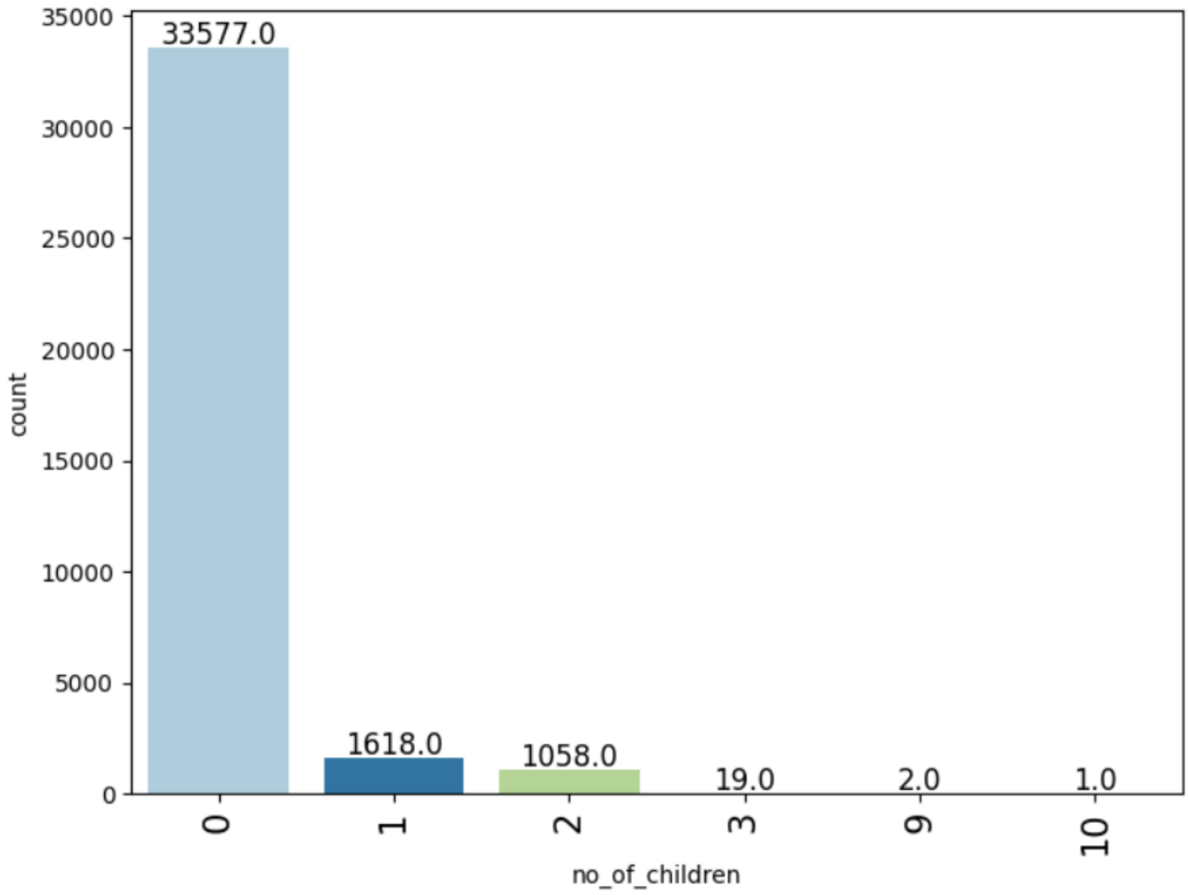


Figure 2: no_of_children

- The mean value of Ad impressions is somewhere around 1400 to 1500 million `ad_impressions`.

- Most of the data lies betwee 1200 to 1700 million `ad_impressions`.

- There is a slight right skew present in the plot.

3. type_of_meal_plan:

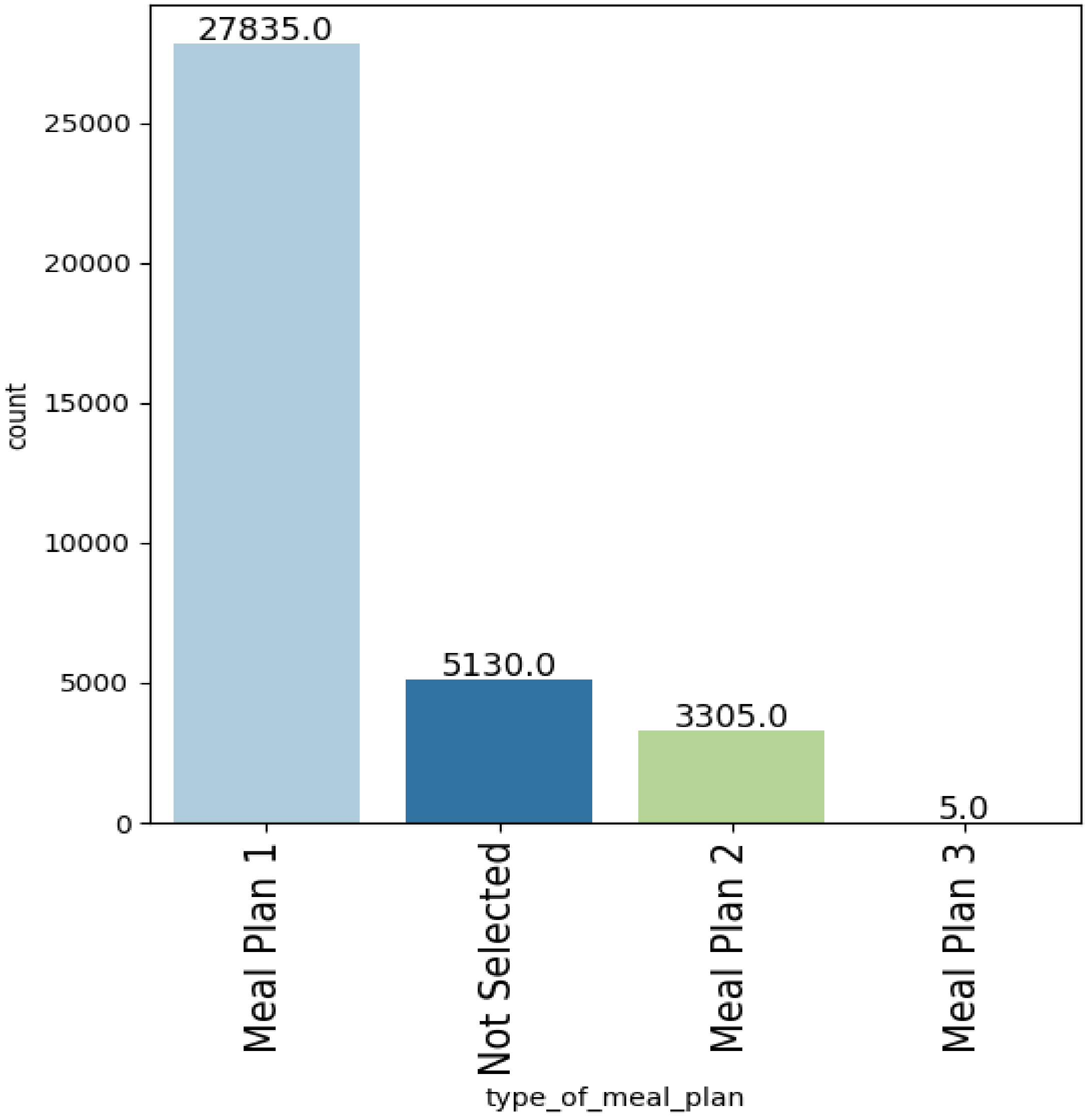


Figure 3: type_of_meal_plan

- Most preferred `type_of_meal_plan` is Meal Plan 1 with 27835 as its count.

- Most people haven't selected any meal type as we have 'Not Selected' type in the second place with 5130 as its count.

- Most people haven't selected any meal type as we have 'Meal plan 2' type in the second place with a count of 3305.

- Most least preferred meal plan seems to be meal plan 3 with count of only 5.

4. Room_type_reserved:

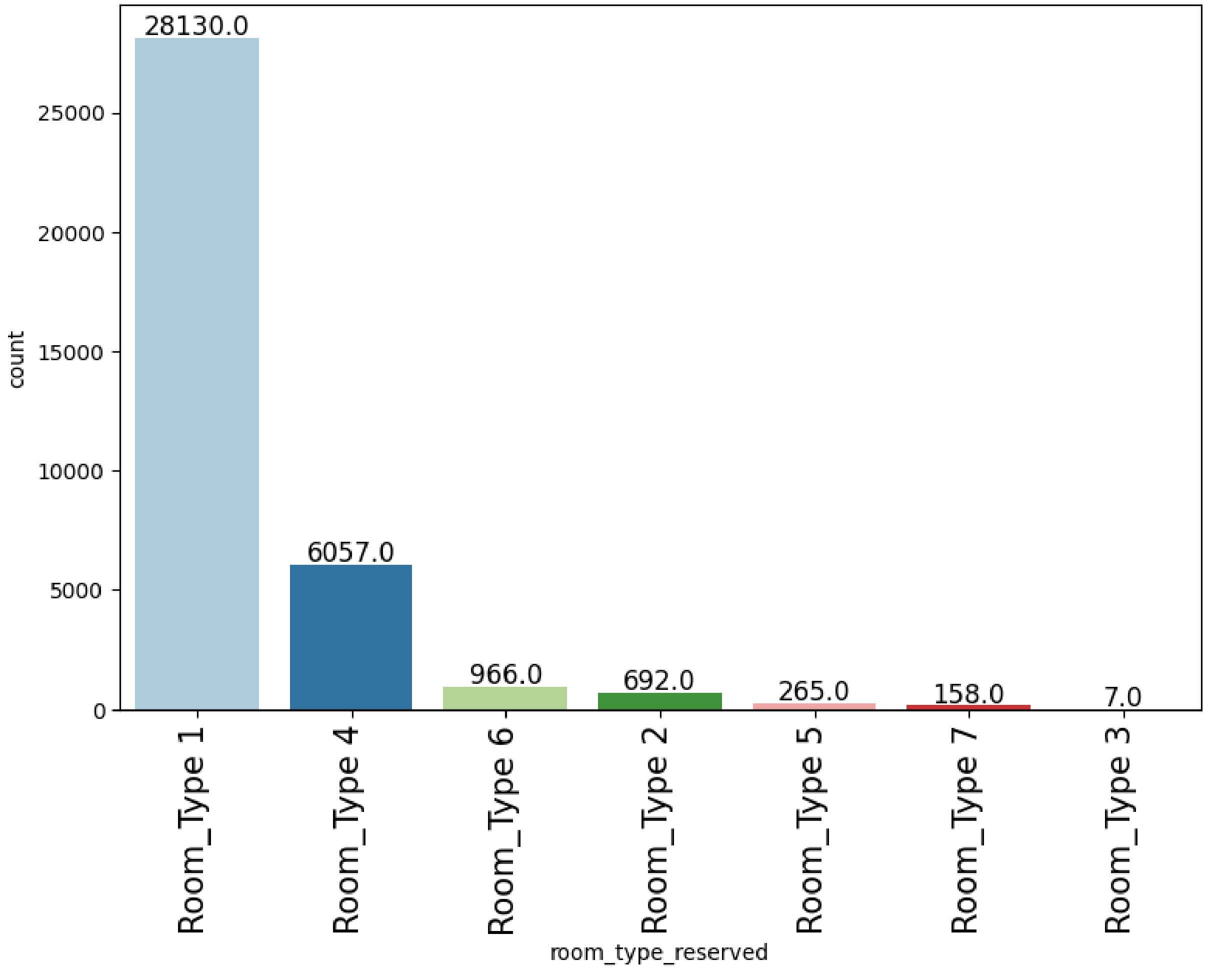


Figure 4: room_type_reserved

- From the above plot we can see that the room_type 1 is highly preferred with a count of 28130.

- 6057 people preferred Room_Type 2 which is second most preferred.

- All the other room_types are least preferred with the values of Room_Type 6 - 966, Room_Type 2 - 692, Room_Type 5 - 265, Room_Type 7 - 158, Room_Type 3 - 7

5. Market_segement_type:

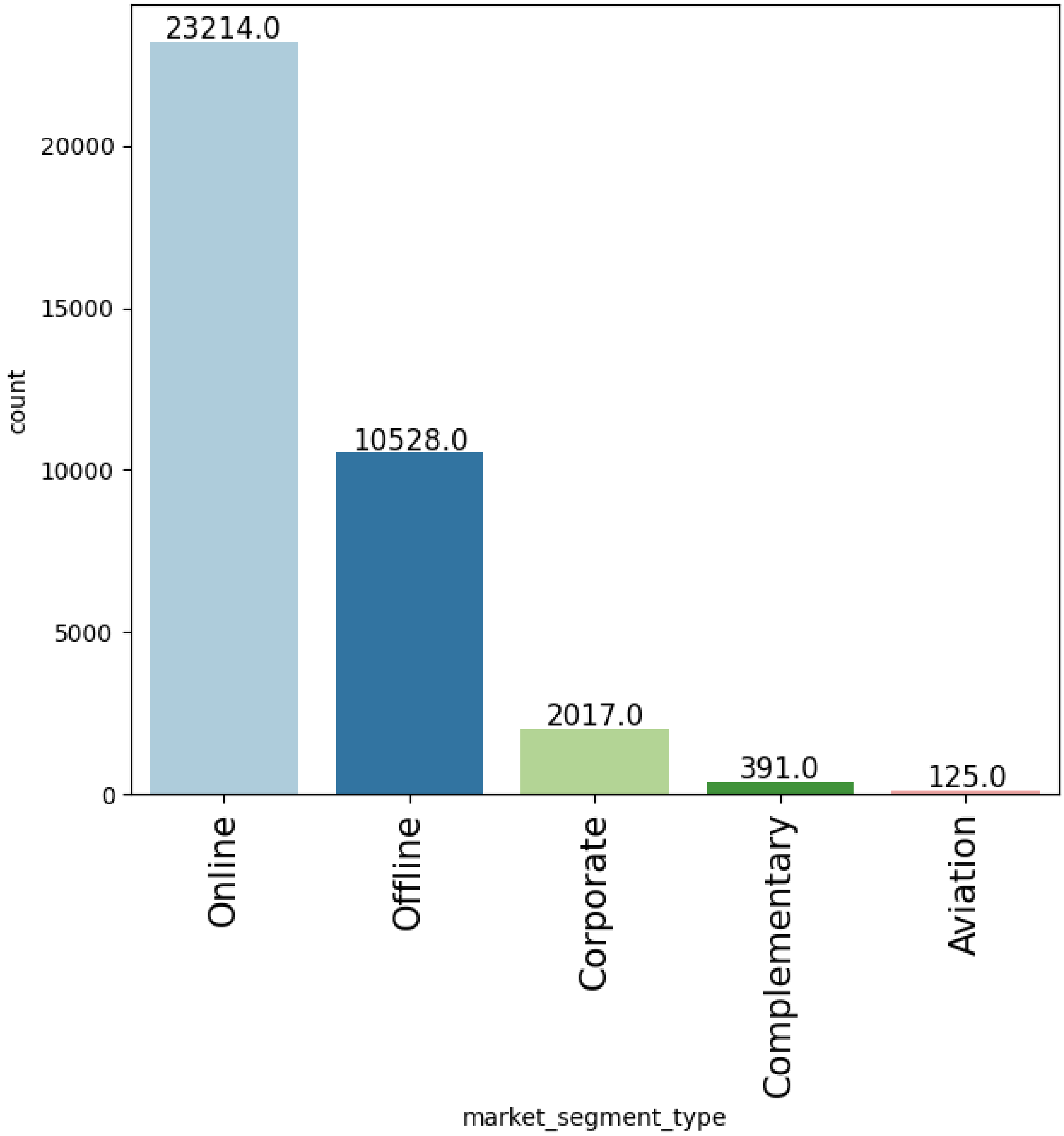


Figure 5:market_segement_type

- Most of the bookings are done through online booking with the count of 23214.

- Offline bookings are second most preferred with count of 10528.

- The remaining bookings are as follows: corparate - 2017, complementary - 391, Aviation - 125.

6. booking_status:

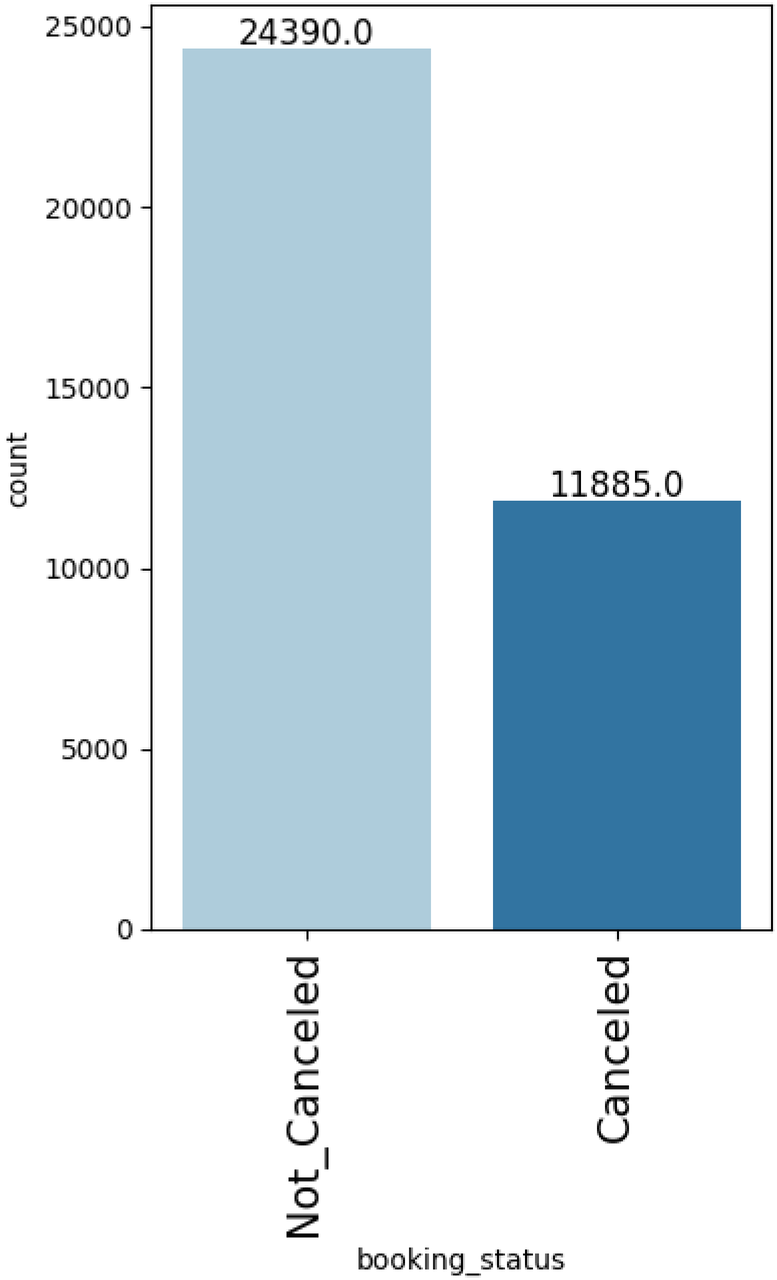


Figure 6: booking status

-As we inferred most of the bookings are not canceled with a count of 24390.

-The canceled bookings are less with a count of 11885.

7. No_of weekend_nights:

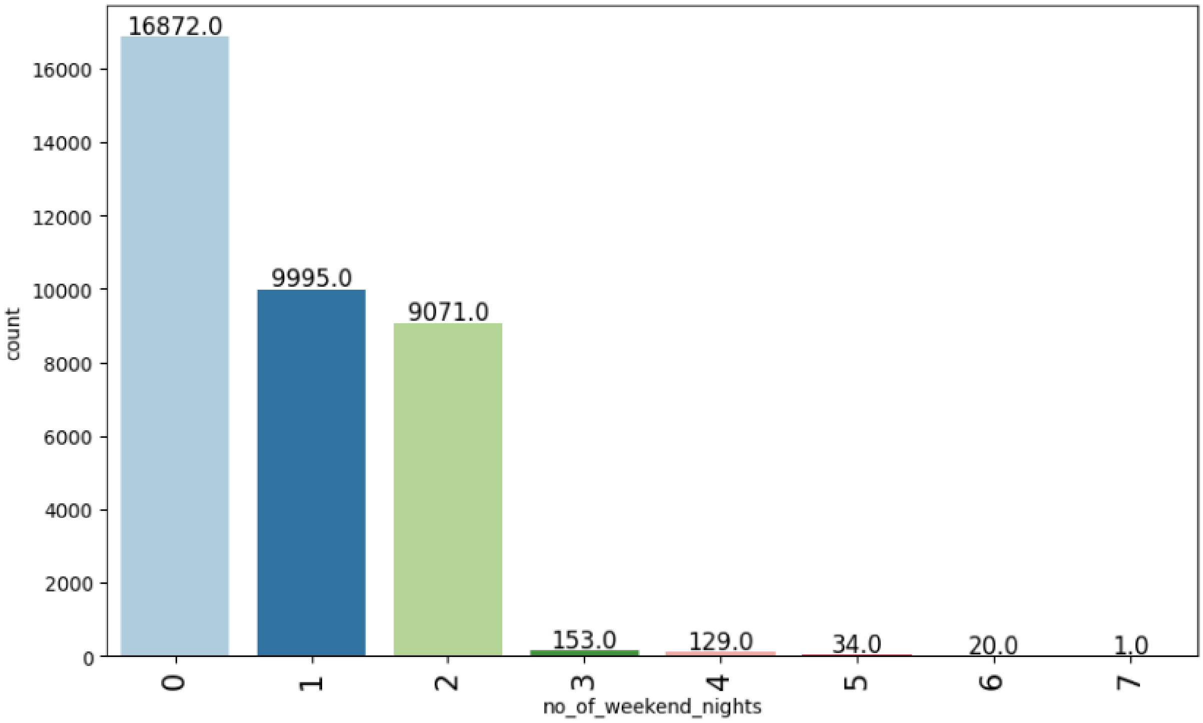


Figure 7: no_of_weekend_nights

- Most bookings are not done on weekend nights, `no_of_weekend_nights` value 0 has been repeated 16872 times in the dataset.

- Next we have 1 weekend night present at the second with 9995 values.

- There are 9071 people who have booked rooms for the whole weekend.

- Remaining values are very less in numbers.

8. No_of_week_nights:

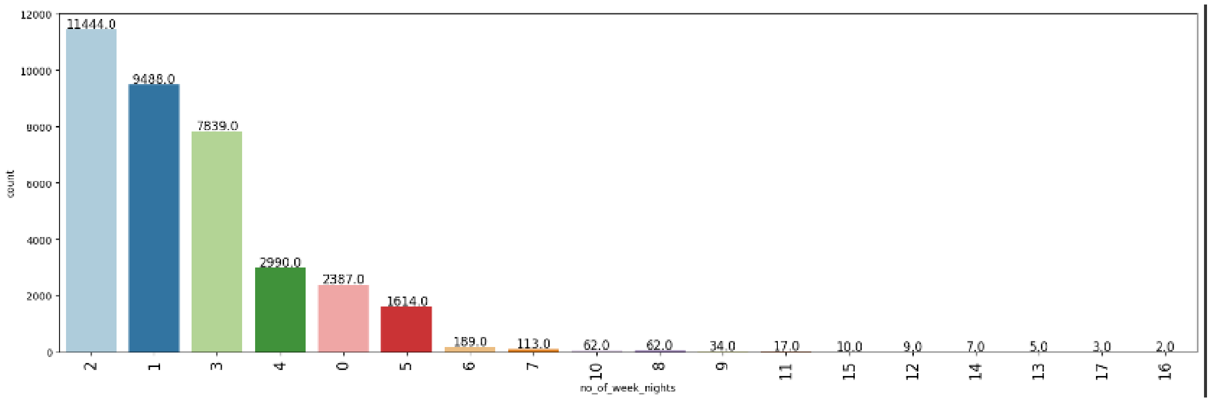


Figure 8: no_of_week_nights

- From the above graph we have most `no_of_week_nights` bookings done for 2 nights with a count of 11444.

- There are 9488 bookings done for 1 night bookings.

- Bookings that are done for 3 nights are done for 7839 in number.

- The remaining values are less in numbers, and there are some bookings done for 16 day as well which is interesting.

9. Required_car_parking_space:

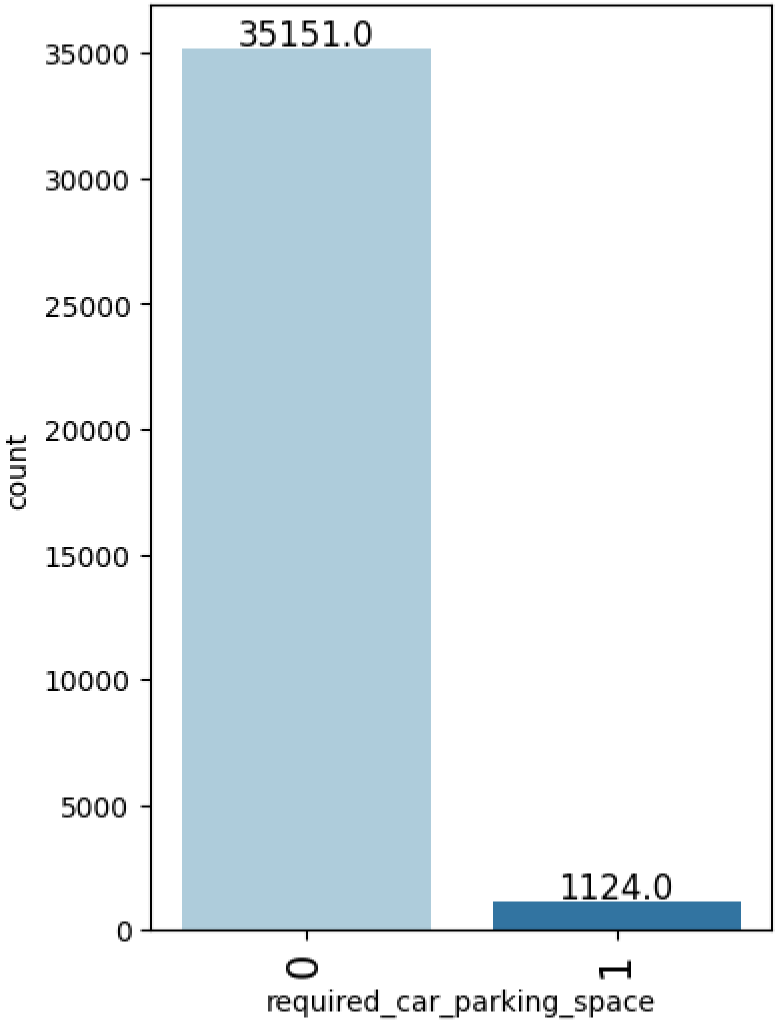


Figure 9: required_car_parking_space

- From the above graph we can conclude that most people doesn't require a car parking.

- The count of people who didn't choose car parking is 35151.

- Number of people who chose car parking is 1124.

10.arrival_month:

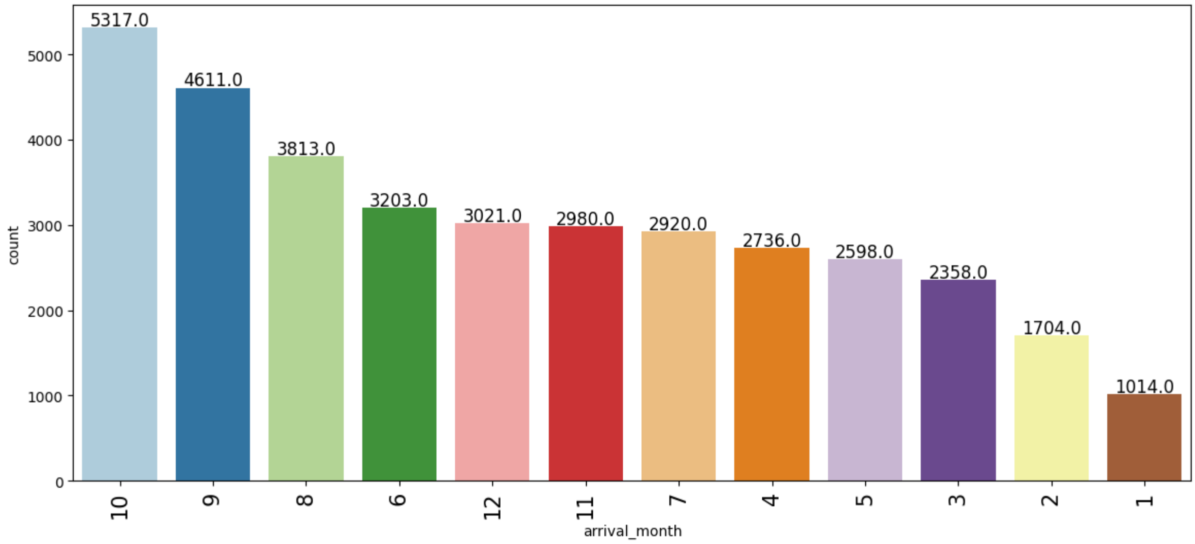


Figure 10: arrival_month

- From the above plot we can clearly see that most bookings are done in the second half of the year month of October, September, August.

11. no_of_special_requests:

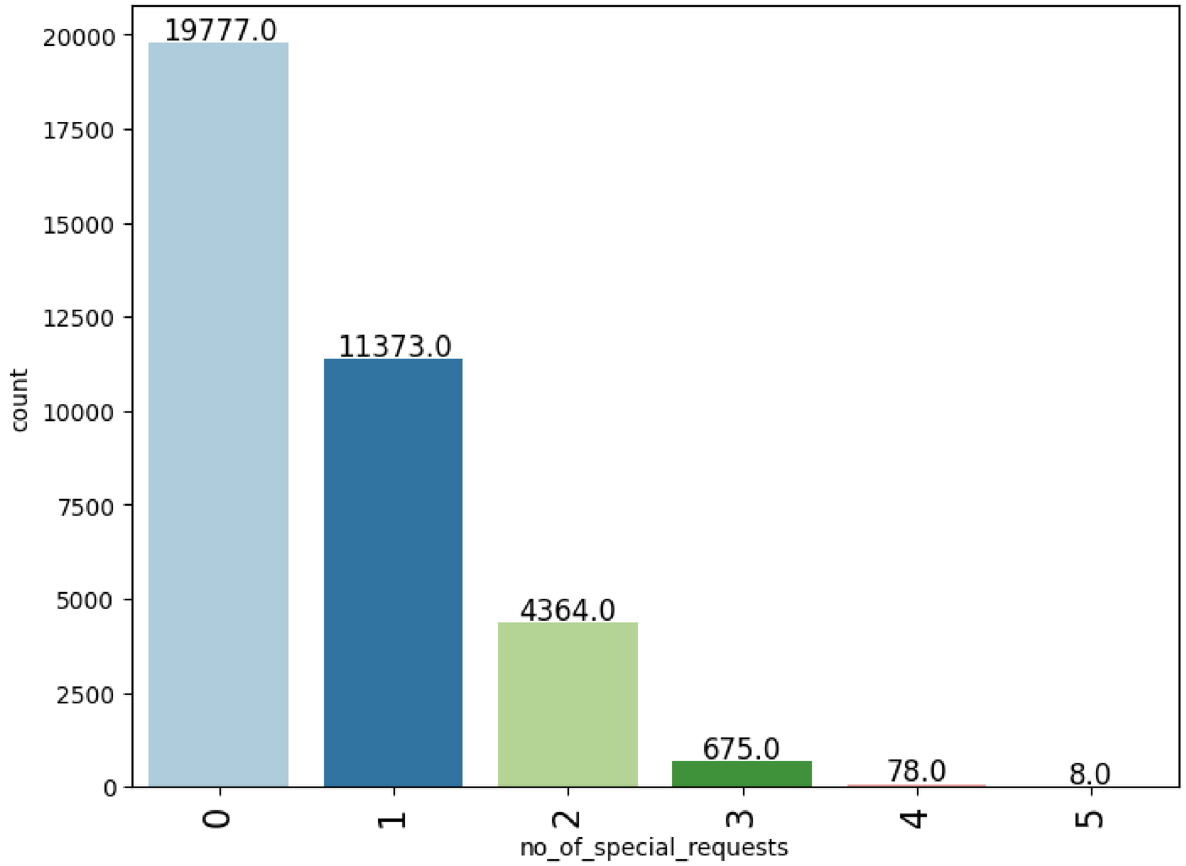


Figure 11: no_of_special_requests

- From the above graph we have people without special requests with a count of 19777.

- People who have had 1 special requests are 11373 in count.

Numerical Variables:

1. Lead_time:

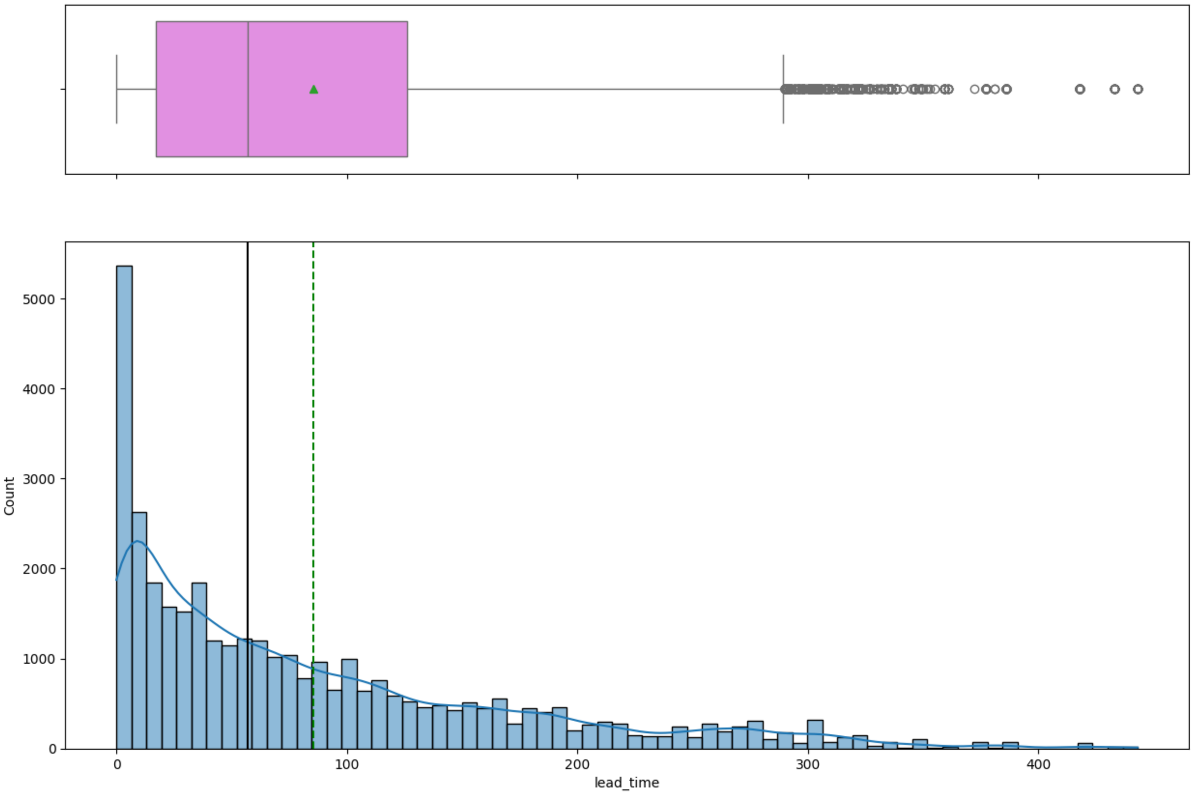


Figure 12: lead_time

- From the above graph we can conclude that the plot is right skewed.

- The lead time value gradually decreases in count.

- The mean value is somewhere near 60 days and the meadian value is somewhere aroud 80 to 90.

2. Avg_price_per_room:

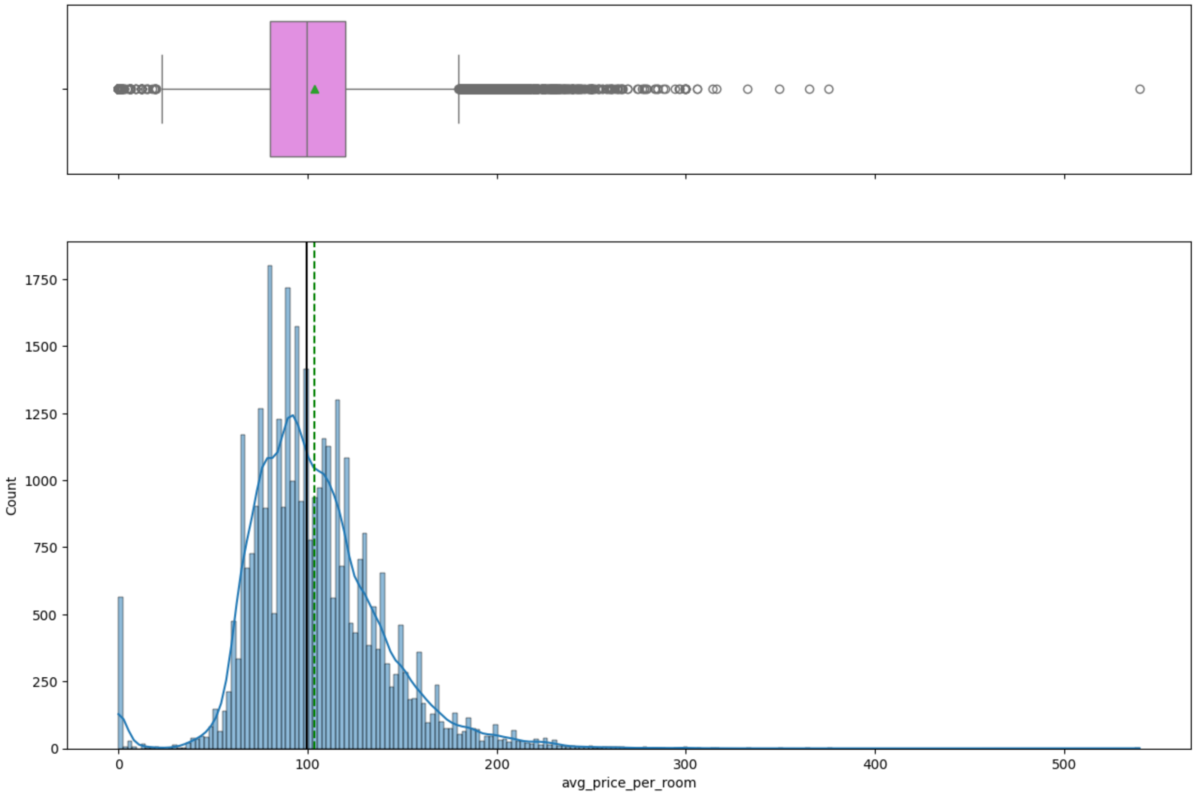


Figure 13: avg_price_per_room

- The above graph is right skewed and has the mean and median value close to 100.

- Hence, the average price per room is around 100 euros.

3. Arrival_date:

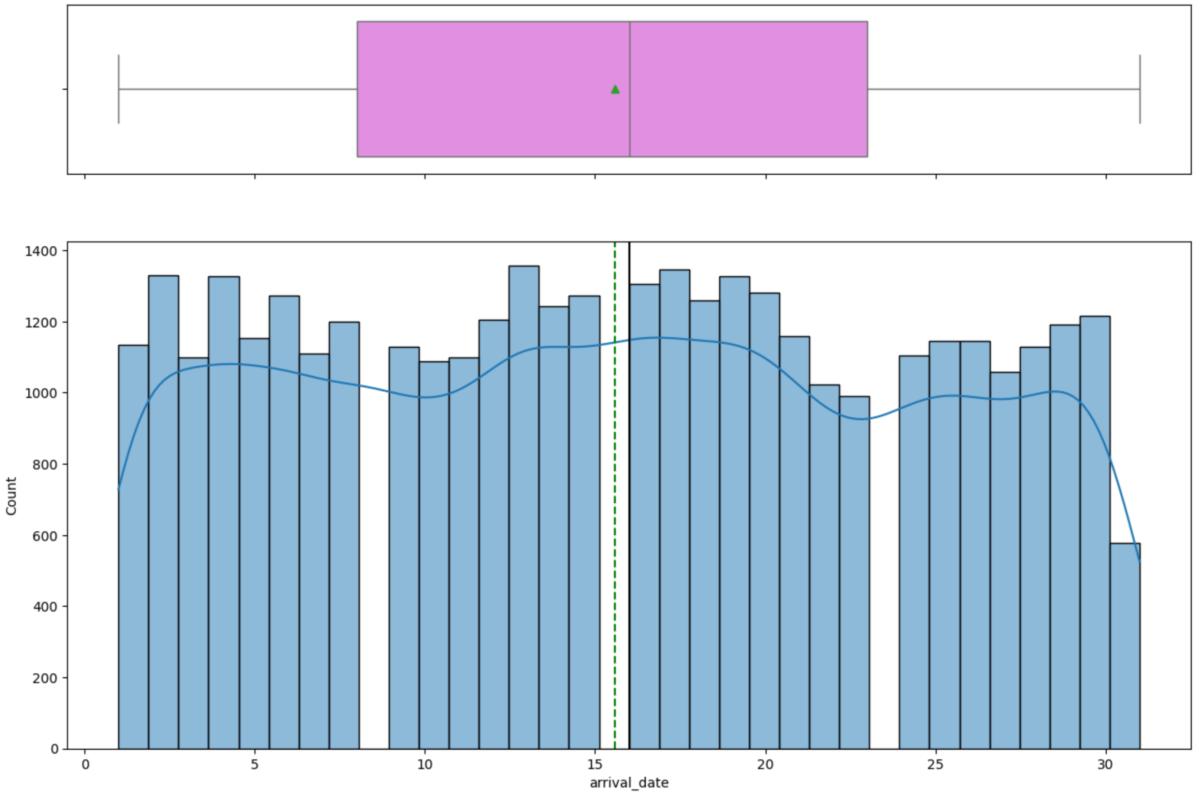
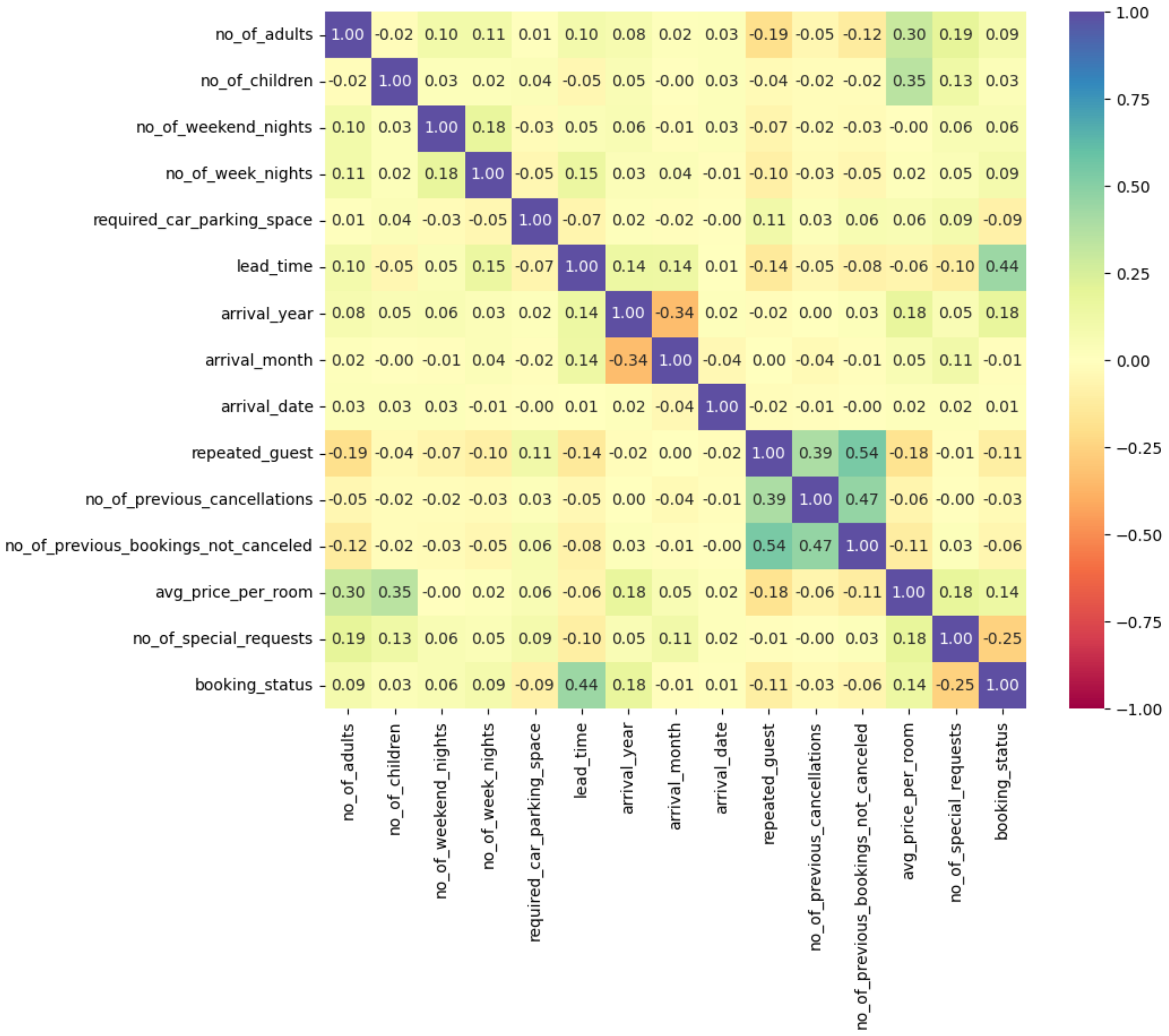


Figure 14: arrival_date

- The arrival date seems to be uniformly distributed.

Bivariate Analysis:

Heatmap:



	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	required_car_parking_space	lead_time	arrival_year	arrival_month	arrival_date	repeated_guest	no_of_previous_cancellations	no_of_previous_bookings_not_canceled	avg_price_per_room	no_of_special_requests	booking_status
no_of_adults	1.00	-0.02	0.10	0.11	0.01	0.10	0.08	0.02	0.03	-0.19	-0.05	-0.12	0.30	0.19	0.09
no_of_children	-0.02	1.00	0.03	0.02	0.04	-0.05	0.05	-0.00	0.03	-0.04	-0.02	-0.02	0.35	0.13	0.03
no_of_weekend_nights	0.10	0.03	1.00	0.18	-0.03	0.05	0.06	-0.01	0.03	-0.07	-0.02	-0.03	-0.00	0.06	0.06
no_of_week_nights	0.11	0.02	0.18	1.00	-0.05	0.15	0.03	0.04	-0.01	-0.10	-0.03	-0.05	0.02	0.05	0.09
required_car_parking_space	0.01	0.04	-0.03	-0.05	1.00	-0.07	0.02	-0.02	-0.00	0.11	0.03	0.06	0.06	0.09	-0.09
lead_time	0.10	-0.05	0.05	0.15	-0.07	1.00	0.14	0.14	0.01	-0.14	-0.05	-0.08	-0.06	-0.10	0.44
arrival_year	0.08	0.05	0.06	0.03	0.02	0.14	1.00	-0.34	0.02	-0.02	0.00	0.03	0.18	0.05	0.18
arrival_month	0.02	-0.00	-0.01	0.04	-0.02	0.14	-0.34	1.00	-0.04	0.00	-0.04	-0.01	0.05	0.11	-0.01
arrival_date	0.03	0.03	0.03	-0.01	-0.00	0.01	0.02	-0.04	1.00	-0.02	-0.01	-0.00	0.02	0.02	0.01
repeated_guest	-0.19	-0.04	-0.07	-0.10	0.11	-0.14	-0.02	0.00	-0.02	1.00	0.39	0.54	-0.18	-0.01	-0.11
no_of_previous_cancellations	-0.05	-0.02	-0.02	-0.03	0.03	-0.05	0.00	-0.04	-0.01	0.39	1.00	0.47	-0.06	-0.00	-0.03
no_of_previous_bookings_not_canceled	-0.12	-0.02	-0.03	-0.05	0.06	-0.08	0.03	-0.01	-0.00	0.54	0.47	1.00	-0.11	0.03	-0.06
avg_price_per_room	0.30	0.35	-0.00	0.02	0.06	-0.06	0.18	0.05	0.02	-0.18	-0.06	-0.11	1.00	0.18	0.14
no_of_special_requests	0.19	0.13	0.06	0.05	0.09	-0.10	0.05	0.11	0.02	-0.01	-0.00	0.03	0.18	1.00	-0.25
booking_status	0.09	0.03	0.06	0.09	-0.09	0.44	0.18	-0.01	0.01	-0.11	-0.03	-0.06	0.14	-0.25	1.00

Figure 15: Heatmap

- From the above heatmap we can see most of the columns don't really have correlation.

- The maximum correlation we have is between `no_of_previous_bookings_not_canceled` and `repeated_guest` with correlation of 0.54.

- `no_of_previous_bookings_not_canceled` and

`no_of_previous_cancellations` also have a positive correlation of 0.47 which is interesting.

- `no_of_previous_cancellations` and `repeated_guest` also have a positive correlation with value 0.39.

- We can see a negative correlation between `arrival_month` and `arrival_year` with a value of -0.34.

- `avg_price_per_room` and `no_of_children` have a positive correlation of 0.34 and `avg_price_per_room` and `no_of_adults` have a positive correlation of 0.30.

- Other than these we have a correlations whic are very minimal in number.

1. Booking_status with hue arrival_month:

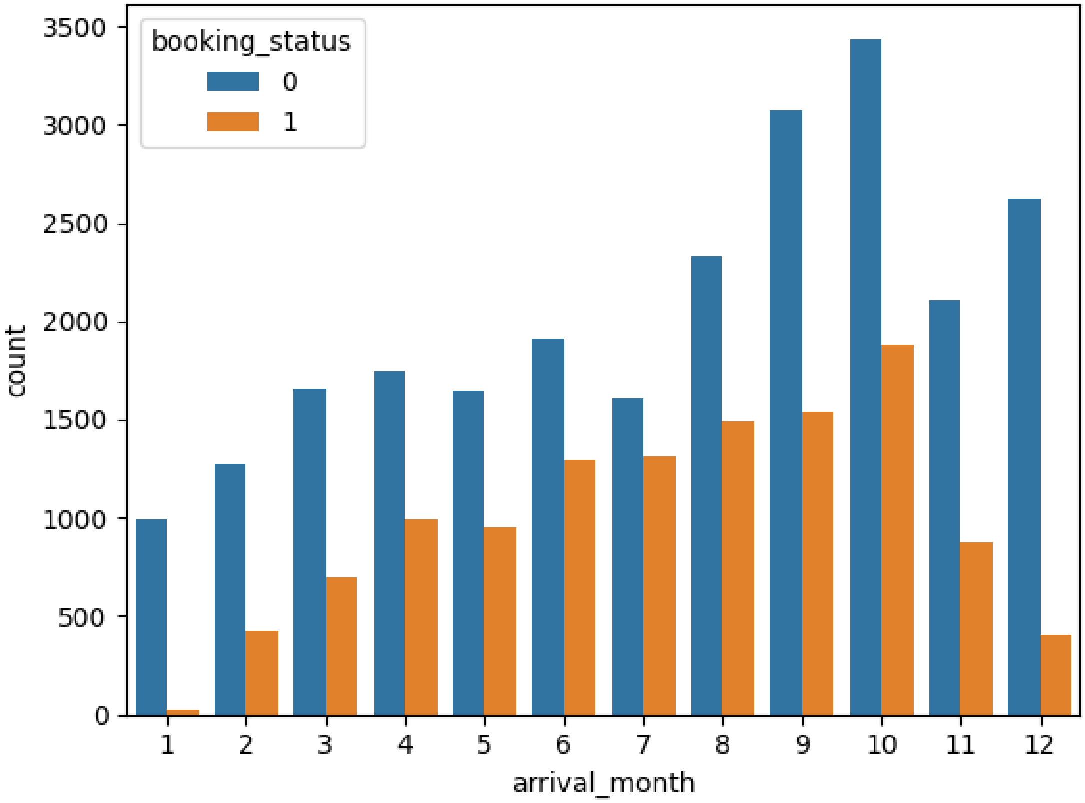


Figure 16: Booking_status with hue arrival_month:

- we can se that the month of October, September, August have highest number of bookings but in contrast the also have the higest number of bookings cancellation value.

- The month of January and February have least number of cancellations.

2. Market_segement_type vs avg_price_per_room:

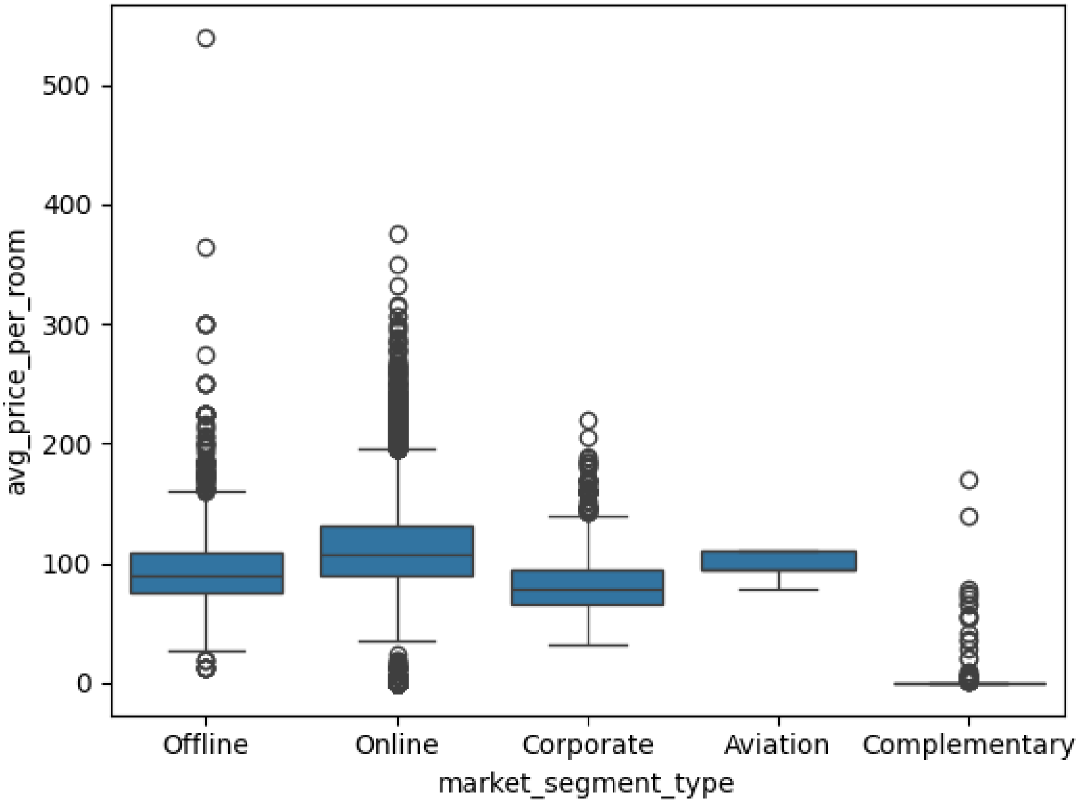


Figure 17: Market_segement_type vs avg_price_per_room

- We can clearly see the price of rooms are high in online and Offline mode of booking. W =e can also see the highest peak of more than 500 euros for Offline booking.

- The Corparate market segment is more in number but the median of aviation is higher than the corporate.

- Complementary market segment gets the rooms at very low prices, which makes sense.

- The distribution for offline and corporate room prices are almost similar except for some outliers.

3. Market_segement_type with hue booking_status:

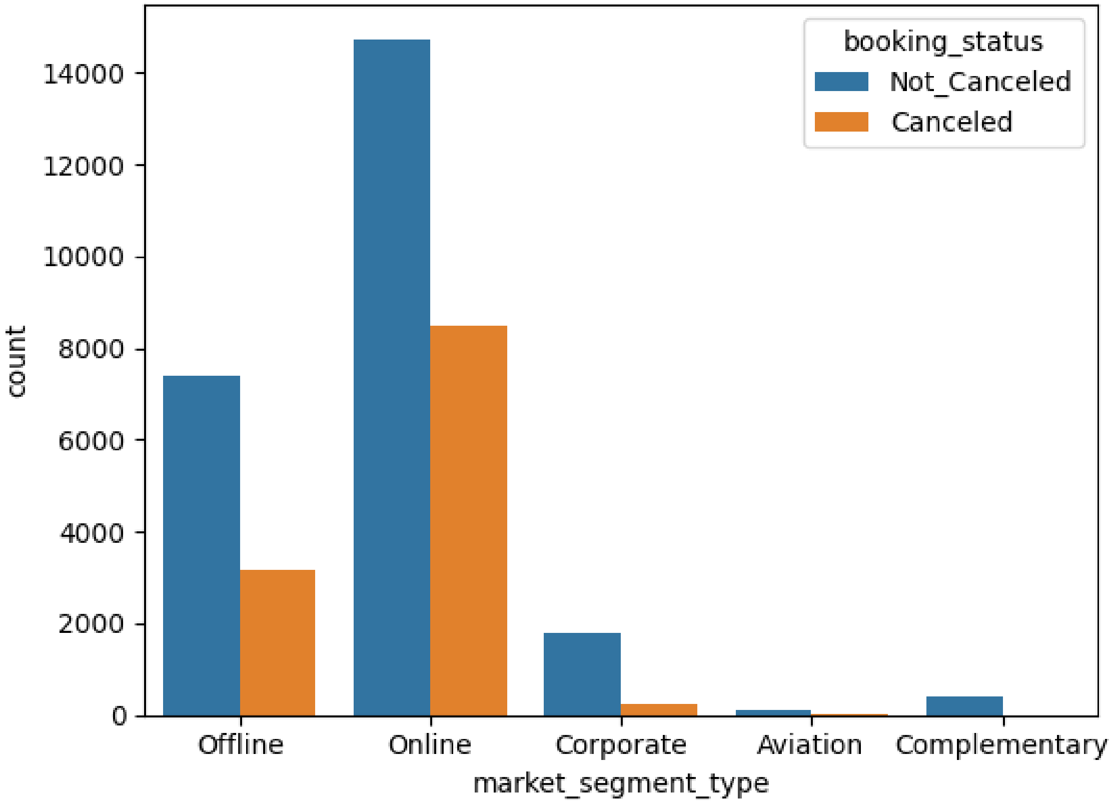


Figure 18: Market_segement_type with hue booking_status

- The Online mode of booking is more in count and also has the more number of cancellation of bookings.

- Whereas the other mode of booking has the least number of cancellations

Important Questions:

1. What are the busiest months in the hotel?

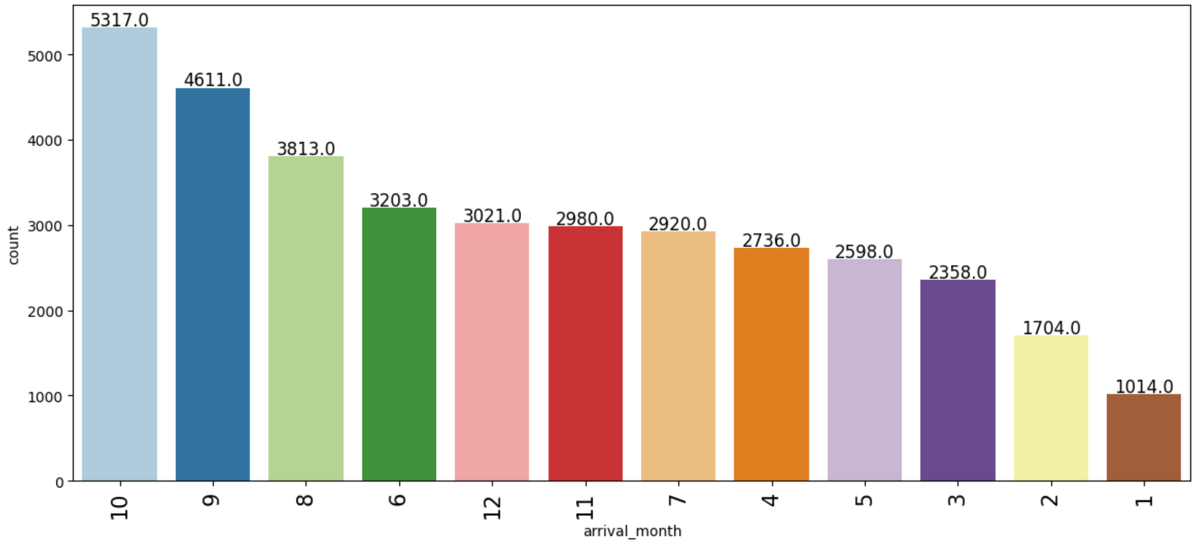


Figure 19: What are the busiest months in the hotel?

- From the above plot we can clearly see that most bookings are done in the second half of the year month of October, September, August.

- Seems like the busiest month is October in the hotel. And then September, August.

2. Which market segment do most of the guests come from?

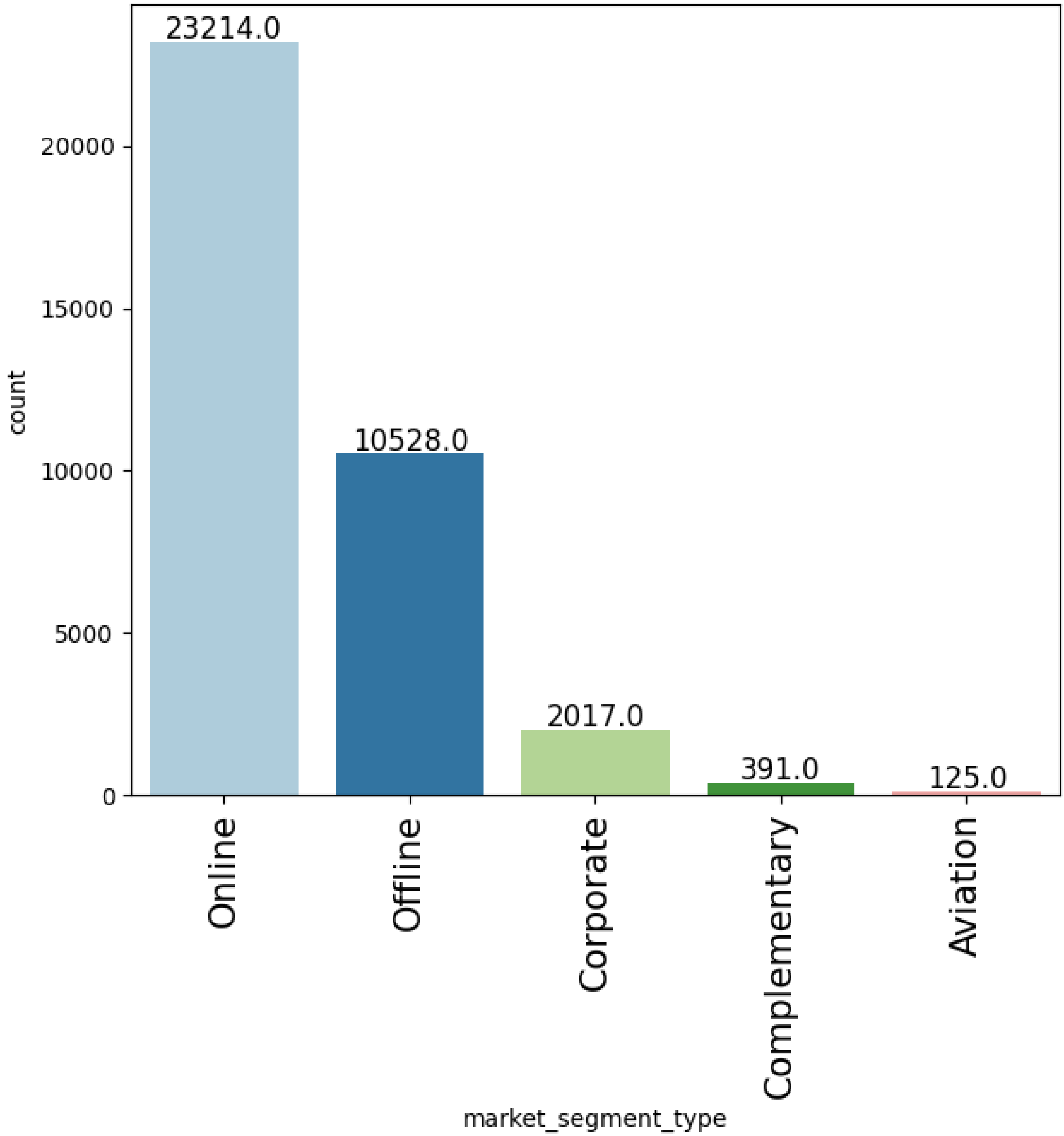


Figure 20: Which market segment do most of the guests come from?

- Most of the bookings are done through online booking with the count of 23214.

- Offline bookings are second most preferred with count of 10528.

- The remaining bookings are as follows: corparate - 2017, complementary - 391, Aviation - 125.

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

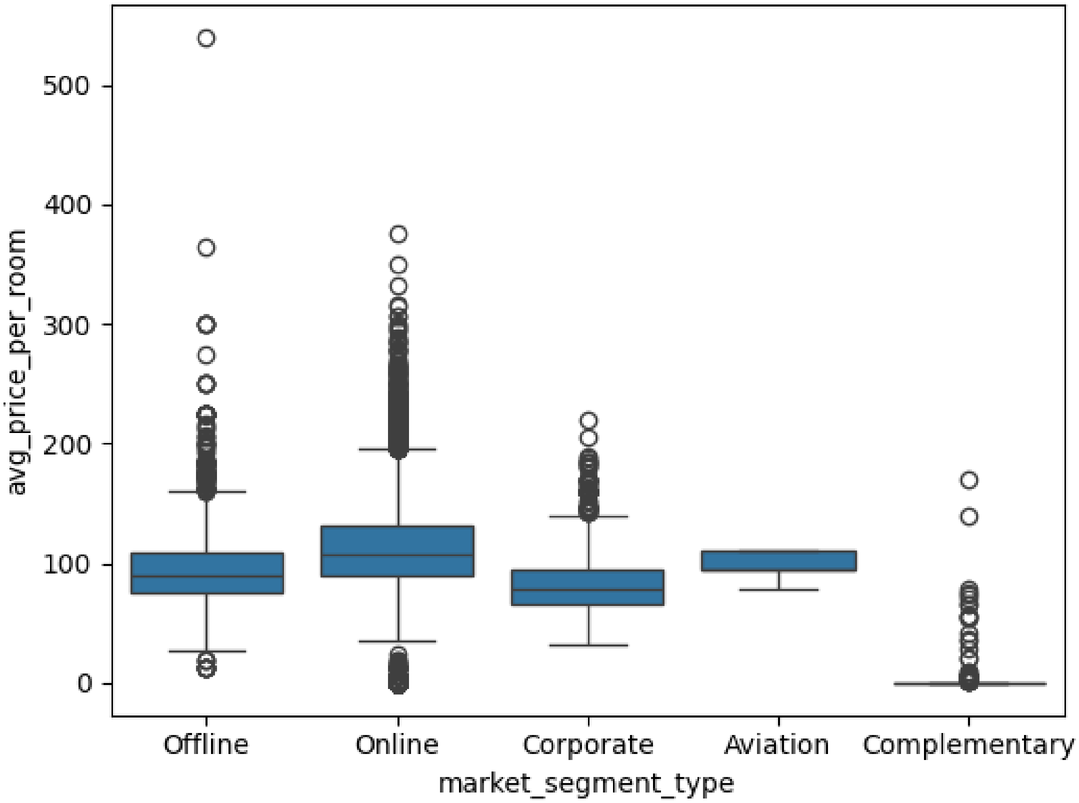


Figure 21: Hotel rates are dynamic and change according to demand and customer demographics

- We can clearly see the price of rooms are high in online and Offline mode of booking. W =e can also see the highest peak of more than 500 euros for Offline booking.

- The Corparate market segment is more in number but the median of aviation is higher than the corporate.

- Complementary market segment gets the rooms at very low prices, which makes sense.

- The distribution for offline and corporate room prices are almost similar except for some outliers.

4. What percentage of bookings are canceled?

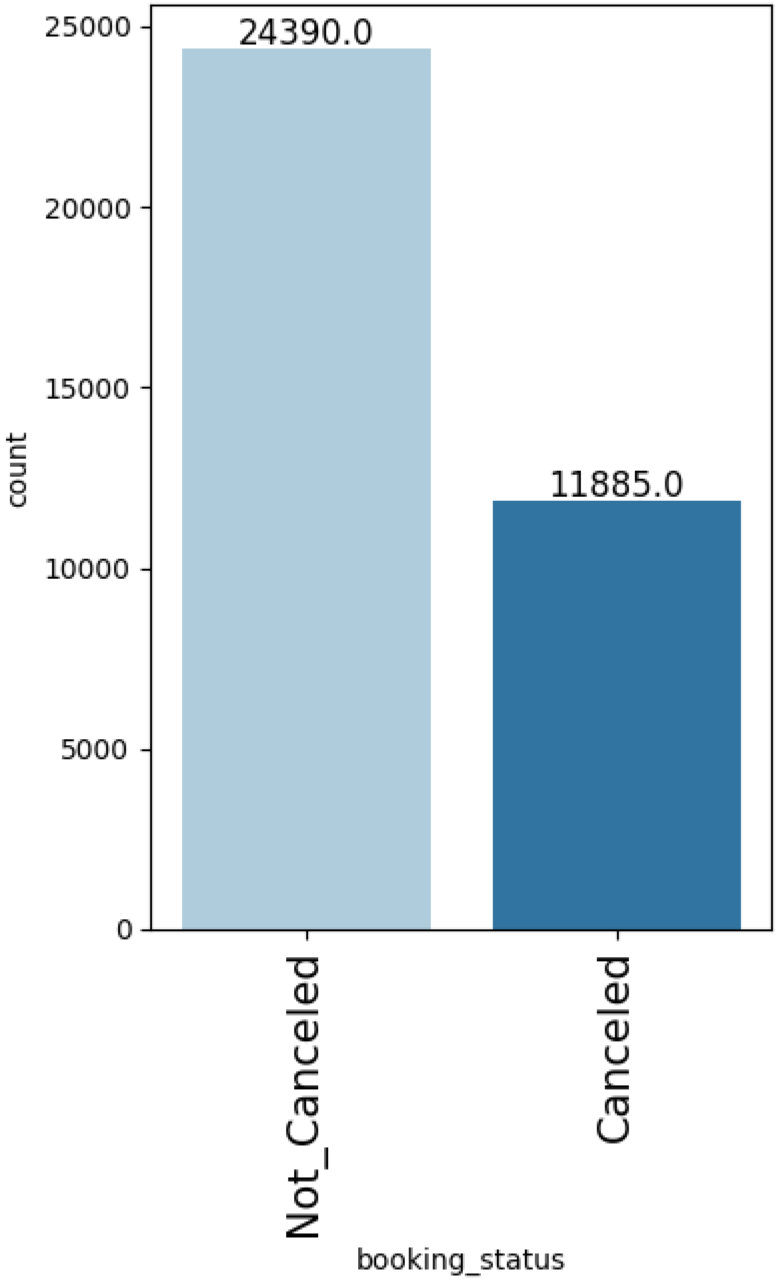
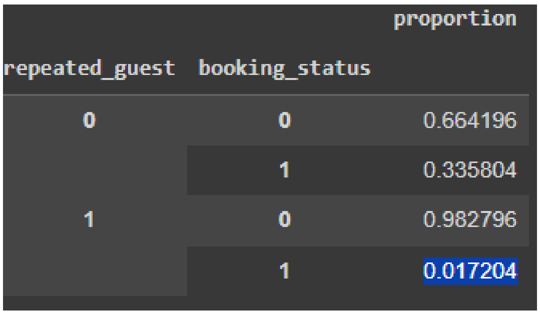


Figure 22: What percentage of bookings are canceled?

-The canceled bookings are with a count of 11885 and the percentage of cancelled bookings is 32.7 percentage.

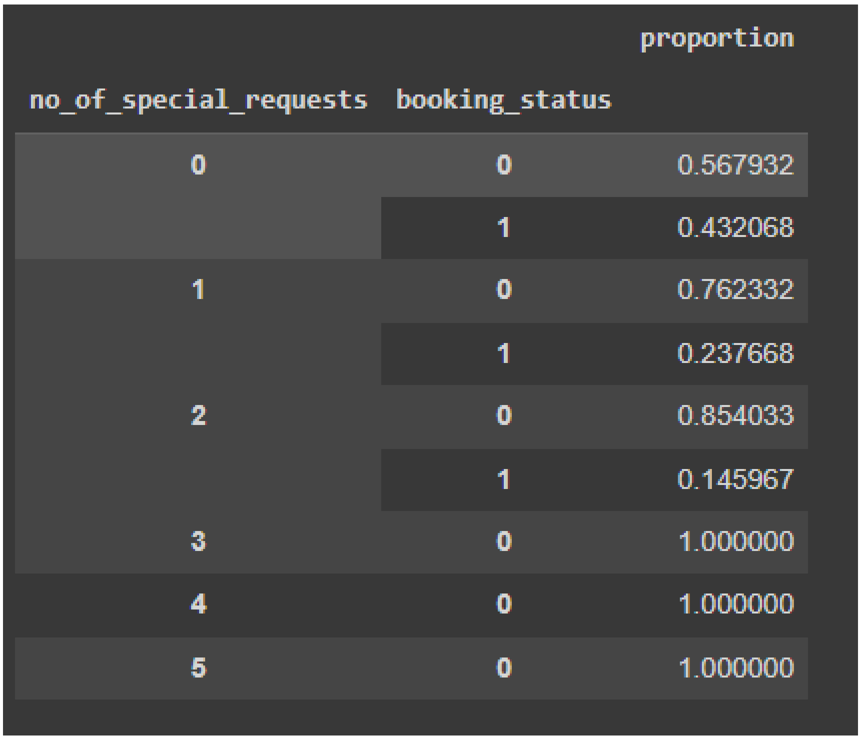
5. Repeating guests are the guests who stay in the hotel often and are important to brand equity. What percentage of repeating guests cancel?



		proportion
repeated_guest	booking_status	
0	0	0.664196
	1	0.335804
1	0	0.982796
	1	0.017204

-The percentage of repeating guests cancel is 0.017%.

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



		proportion
no_of_special_requests	booking_status	
0	0	0.567932
	1	0.432068
1	0	0.762332
	1	0.237668
2	0	0.854033
	1	0.145967
3	0	1.000000
4	0	1.000000
5	0	1.000000

-Seems like people who have more than 2 special request have less percentage of booking cancellation.

-The number of booking cancellation also reduces on increasing special requests.

Data preprocessing:

Outlier Treatment:

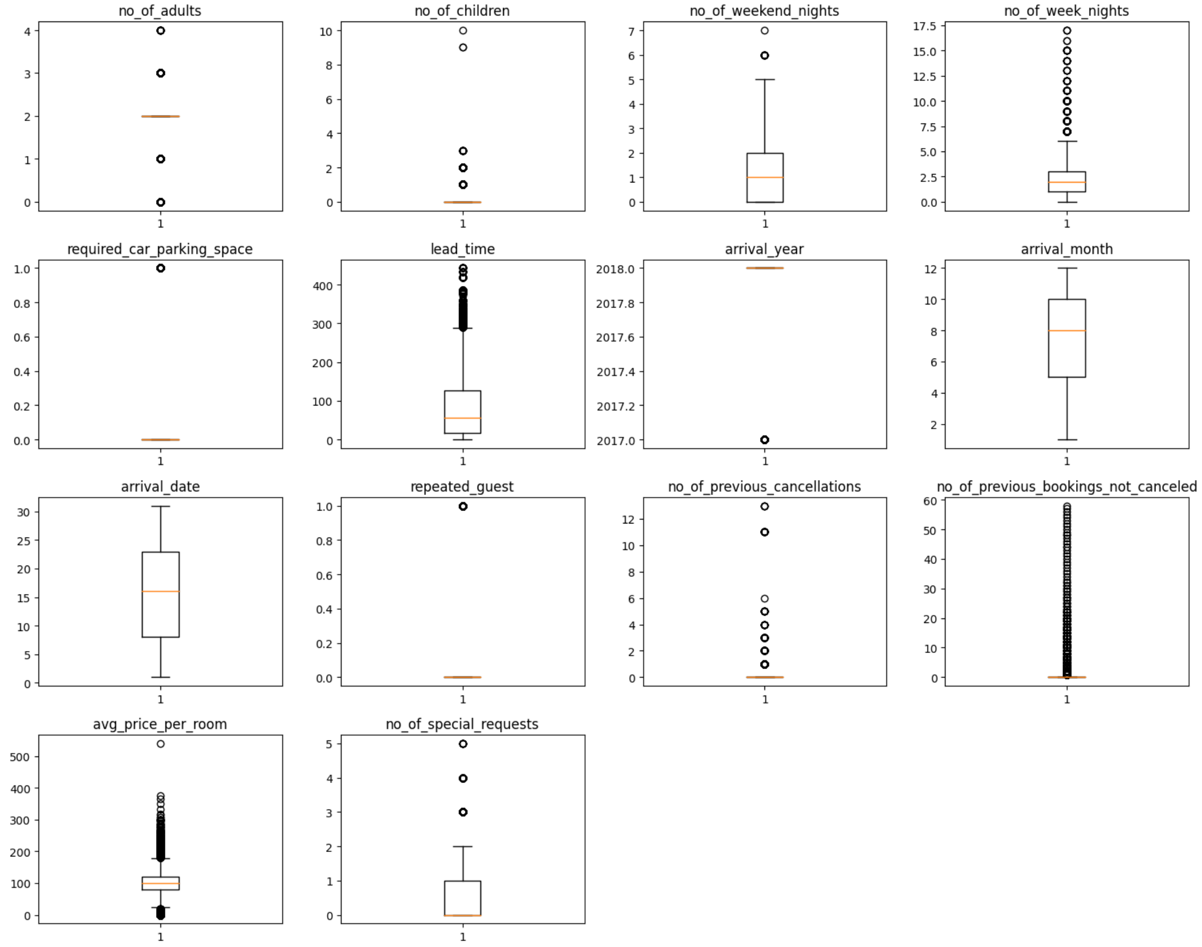
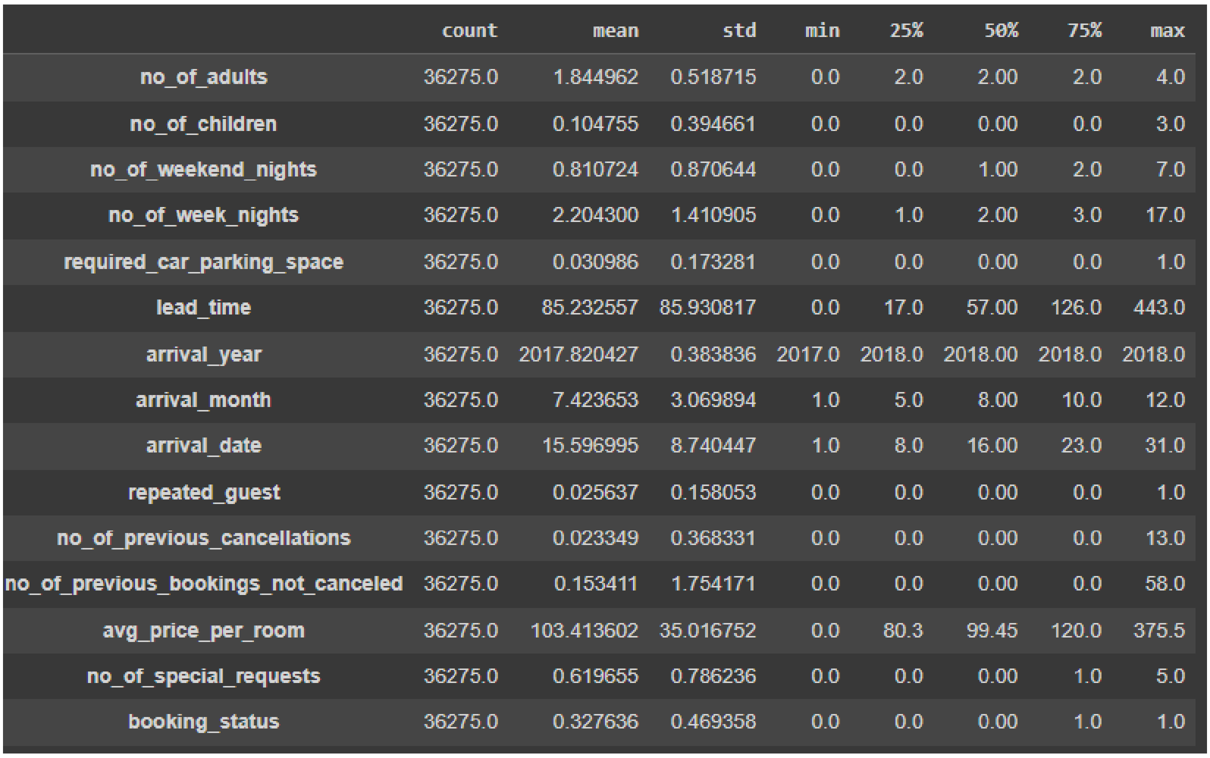


Figure 23: Outlier Treatment

- From, the above plot we can see that the no_of_children value have extreme outliers of 9 and 10 which should be treated.

- The avg_price_per_room also have an outlier of more 500 euros which should be treated.

- All other values seems to be good.



	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	0.104755	0.394661	0.0	0.0	0.00	0.0	3.0
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	103.413602	35.016752	0.0	80.3	99.45	120.0	375.5
no_of_special_requests	36275.0	0.619655	0.786236	0.0	0.0	0.00	1.0	5.0
booking_status	36275.0	0.327636	0.469358	0.0	0.0	0.00	1.0	1.0

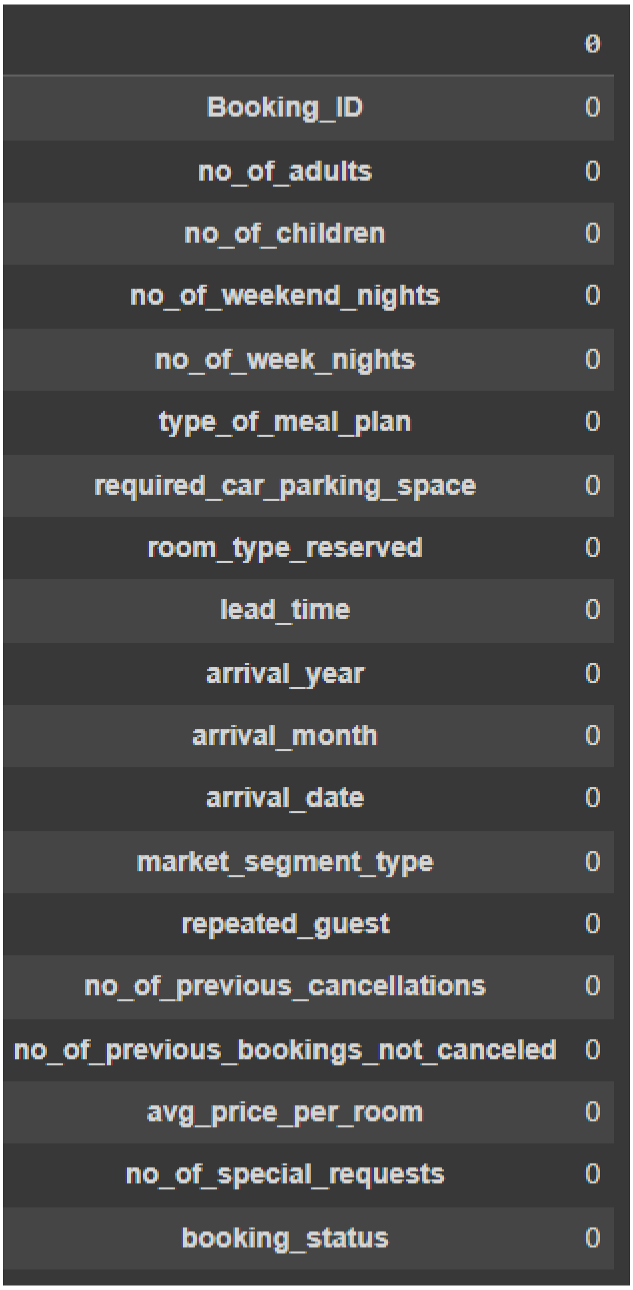
- As we have some outliers in number of children we are treating the values of 9 and 10 to the upper whiskers which is 3.

- The avg_price_per_room max value also got changed to 443

Feature engineering:

- We are not doing any feature engineering here since all the columns seems to have independent effect on the target variable.

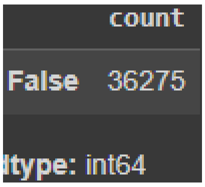
Missing value treatment:



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

- We don’t have any missing values in the dataset.

Duplicate value:

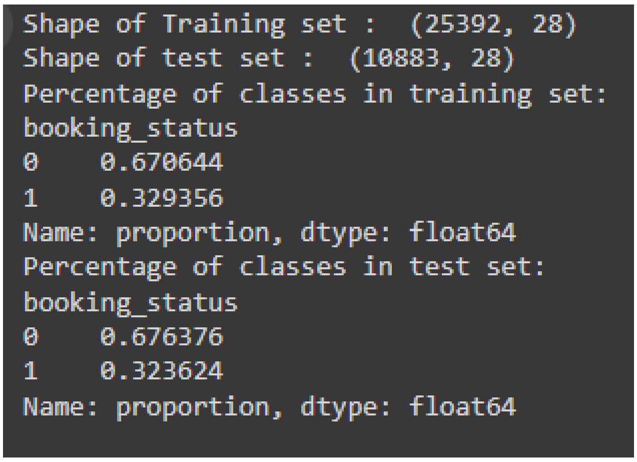


```
       count
False  36275

ltype: int64
```

- Since, we don’t see any duplicate values we are good to go without any treatment for the duplicate.

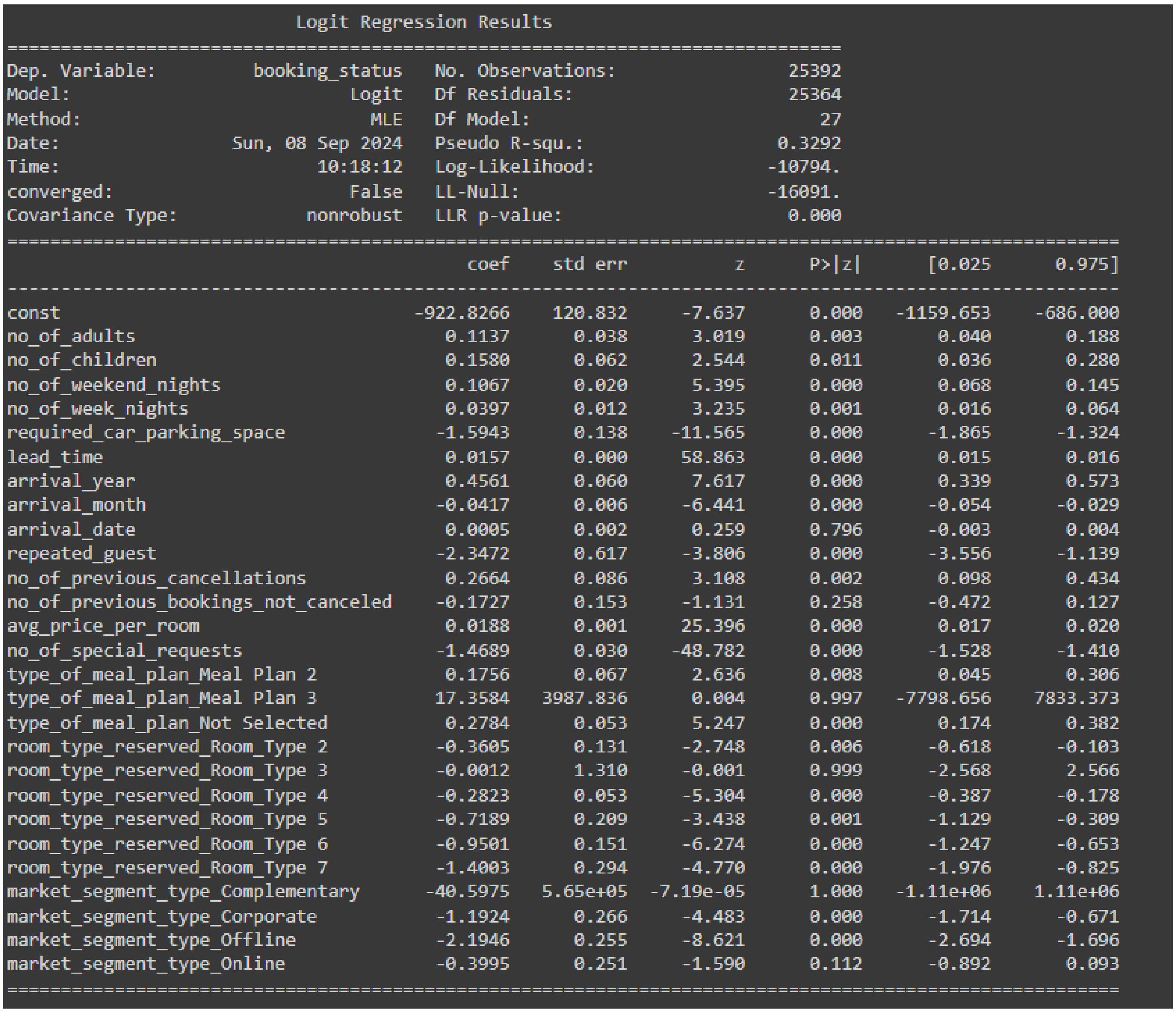
Splitting dataset to Training and Testing dataset:



```
Shape of Training set :  (25392, 28)
Shape of test set :  (10883, 28)
Percentage of classes in training set:
booking_status
0    0.670644
1    0.329356
Name: proportion, dtype: float64
Percentage of classes in test set:
booking_status
0    0.676376
1    0.323624
Name: proportion, dtype: float64
```

- 67% of the data is used for training.
- 32% of the data is used for testing the efficiency of the model.

Model building - Logistic Regression:



```
                           Logit Regression Results                           
==============================================================================
Dep. Variable:         booking_status   No. Observations:                25392
Model:                          Logit   Df Residuals:                    25364
Method:                           MLE   Df Model:                           27
Date:                Sun, 08 Sep 2024   Pseudo R-squ.:                  0.3292
Time:                        10:18:12   Log-Likelihood:                -10794.
converged:                      False   LL-Null:                       -16091.
Covariance Type:            nonrobust   LLR p-value:                     0.000
========================================================================================================
                                           coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------
const                                 -922.8266    120.832     -7.637      0.000   -1159.653    -686.000
no_of_adults                             0.1137      0.038      3.019      0.003       0.040       0.188
no_of_children                           0.1580      0.062      2.544      0.011       0.036       0.280
no_of_weekend_nights                     0.1067      0.020      5.395      0.000       0.068       0.145
no_of_week_nights                        0.0397      0.012      3.235      0.001       0.016       0.064
required_car_parking_space              -1.5943      0.138    -11.565      0.000      -1.865      -1.324
lead_time                                0.0157      0.000     58.863      0.000       0.015       0.016
arrival_year                             0.4561      0.060      7.617      0.000       0.339       0.573
arrival_month                           -0.0417      0.006     -6.441      0.000      -0.054      -0.029
arrival_date                             0.0005      0.002      0.259      0.796      -0.003       0.004
repeated_guest                          -2.3472      0.617     -3.806      0.000      -3.556      -1.139
no_of_previous_cancellations             0.2664      0.086      3.108      0.002       0.098       0.434
no_of_previous_bookings_not_canceled    -0.1727      0.153     -1.131      0.258      -0.472       0.127
avg_price_per_room                       0.0188      0.001     25.396      0.000       0.017       0.020
no_of_special_requests                  -1.4689      0.030    -48.782      0.000      -1.528      -1.410
type_of_meal_plan_Meal Plan 2            0.1756      0.067      2.636      0.008       0.045       0.306
type_of_meal_plan_Meal Plan 3           17.3584   3987.836      0.004      0.997   -7798.656    7833.373
type_of_meal_plan_Not Selected           0.2784      0.053      5.247      0.000       0.174       0.382
room_type_reserved_Room_Type 2          -0.3605      0.131     -2.748      0.006      -0.618      -0.103
room_type_reserved_Room_Type 3          -0.0012      1.310     -0.001      0.999      -2.568       2.566
room_type_reserved_Room_Type 4          -0.2823      0.053     -5.304      0.000      -0.387      -0.178
room_type_reserved_Room_Type 5          -0.7189      0.209     -3.438      0.001      -1.129      -0.309
room_type_reserved_Room_Type 6          -0.9501      0.151     -6.274      0.000      -1.247      -0.653
room_type_reserved_Room_Type 7          -1.4003      0.294     -4.770      0.000      -1.976      -0.825
market_segment_type_Complementary      -40.5975   5.65e+05  -7.19e-05      1.000   -1.11e+06    1.11e+06
market_segment_type_Corporate           -1.1924      0.266     -4.483      0.000      -1.714      -0.671
market_segment_type_Offline             -2.1946      0.255     -8.621      0.000      -2.694      -1.696
market_segment_type_Online              -0.3995      0.251     -1.590      0.112      -0.892       0.093
========================================================================================================
```

Figure 24: Logistic regression model

- The coef values in the above logistic regression results we have

- From the above result we can see the coefficient values range from positive to negative.

- Positive coefficient values are no_of_adults, no_of_children, nno_of_weekend_nights, no_of_week_nights, lead_time, arrival_year, arrival_date, no_of_previous_cancellations, avg_of_previous_per_room, type_of_meal_plan_Meal plan 2, type_meal_plan_Meal plan 3, type_of_meal_plan_not_selected

coefficient:

- Negative coefficient values are required_car_parking_space, arrival_month, repeated_guest, no_of_previous_bookings_not_canceled, no_of_special_requests, room_type_reserved_Room_Type 2, room_type_reserved_Room_Type 3, room_type_reserved_Room_Type 4, room_type_reserved_Room_Type 5, room_type_reserved_Room_Type 6, room_type_reserved_Room_Type 7, market_segment_type_Complementary, market_segment_type_Corporate, market_segment_type_Offline, market_segment_type_Online.

P-Value:

- The some P values also have values greater than 0.05 we might need to see in multicolinearity.

Confusion matrix on Training dataset:

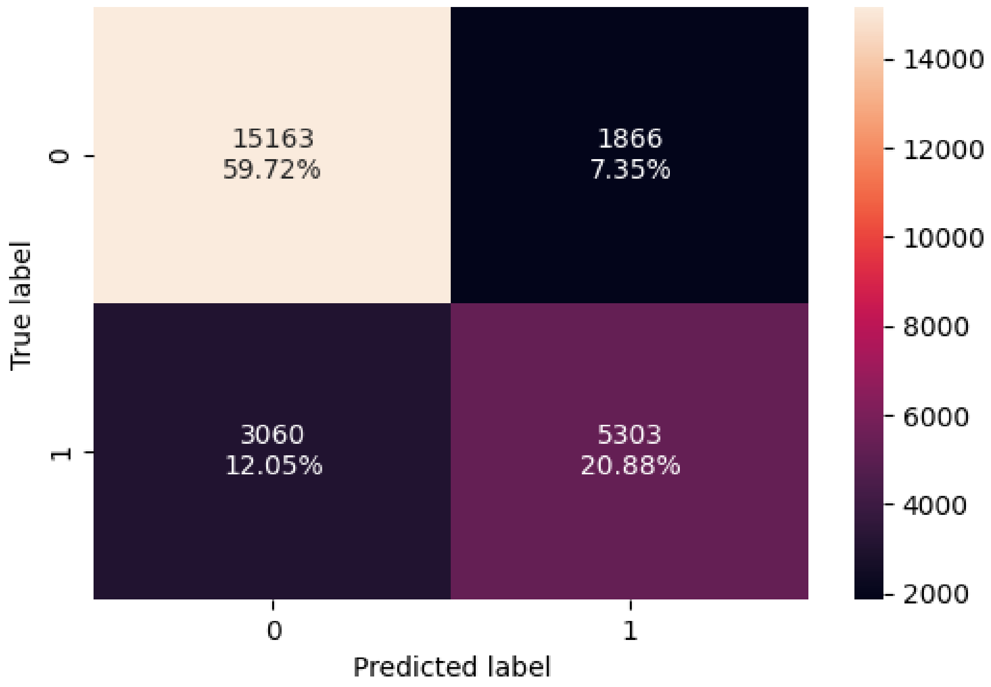


Figure 25: confusion matrix

So we have,

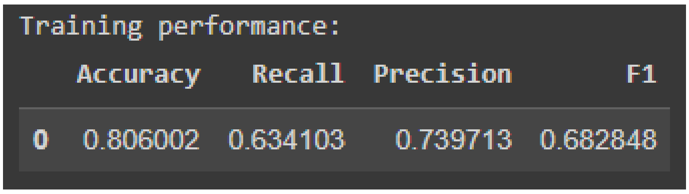
True positive – 5303 values (20.88%)

False Positive – 1866 values (7.35%)

False Negative – 3060 values (12.05%)

True Negative – 15163 values (59.72%)

Performance the model in Training dataset:



```
Training performance:
   Accuracy    Recall  Precision        F1
0  0.806002  0.634103   0.739713  0.682848
```

Figure 26: Performance of the Model

- The recall value for training dataset for logistic regression before fine tuning is 0.634.
-

Confusion matrix for Testing dataset:

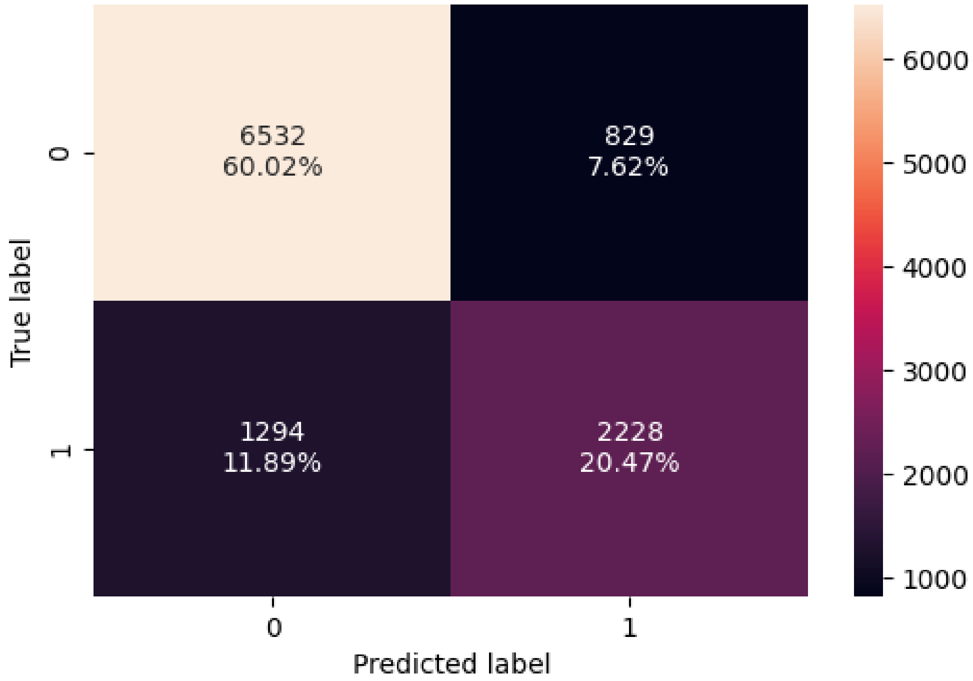


Figure 27: Confusion Matrix

So we have,

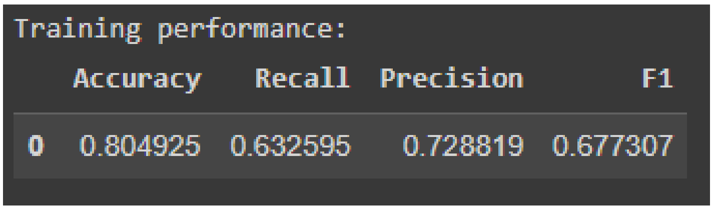
True positive – 2228 values (20.47%)

False Positive – 829 values (7.62%)

False Negative – 1294 values (11.89%)

True Negative – 6532 values (60.02%)

Performance the model in Testing dataset:



```
Training performance:
   Accuracy    Recall  Precision        F1
0  0.804925  0.632595   0.728819  0.677307
```

Figure 28: Performance of the model

- The recall value for training dataset for logistic regression before fine tuning is 0.632.

****

Naïve Bayes:

Model Score for Training dataset:

- We have a Model score for our dataset in naïve bayes algorithm for training dataset is 0.409.

Confusion matrix on Training dataset:

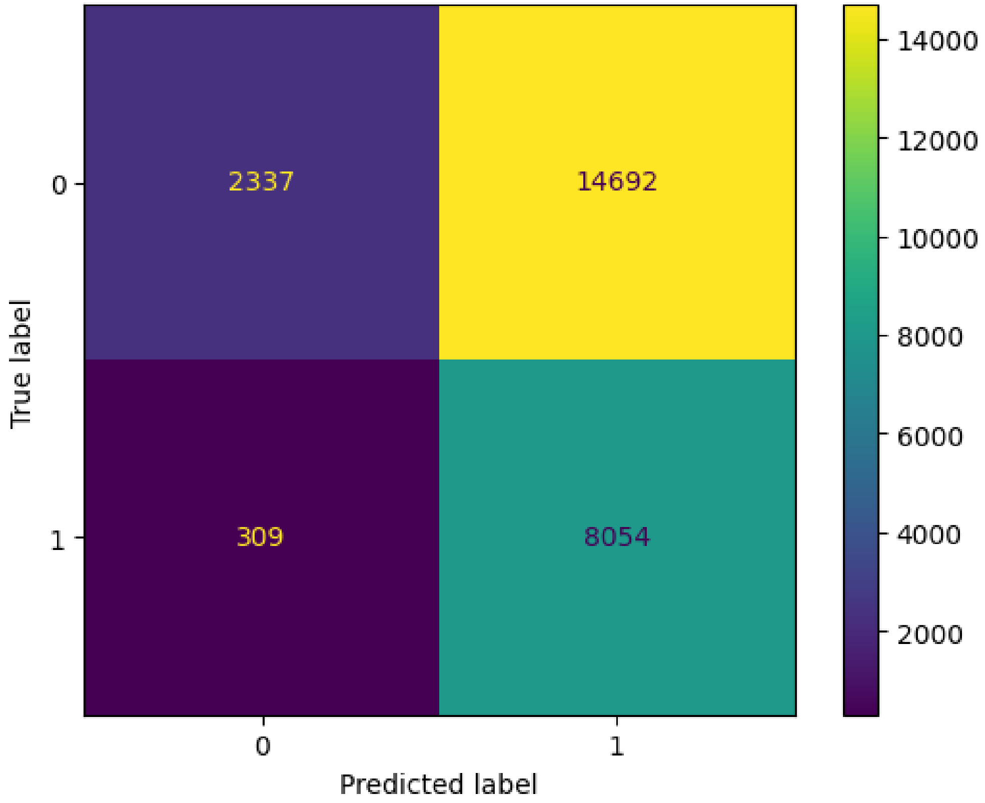


Figure 29: Confusion Matrix

So we have,

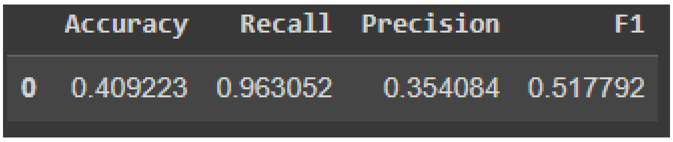
True positive – 8054 values.

False Positive – 14692 values.

False Negative – 309 values.

True Negative – 2337 values.

Performance the model in Training dataset:



	Accuracy	Recall	Precision	F1
0	0.409223	0.963052	0.354084	0.517792

Figure 30: Performance of the model

- The recall value for training dataset for logistic regression before fine tuning is 0.9630.

Model Score Testing dataset:

- We have a Model score for our dataset in naïve bayes algorithm for testing dataset is 0.406.

Confusion matrix for Testing dataset:

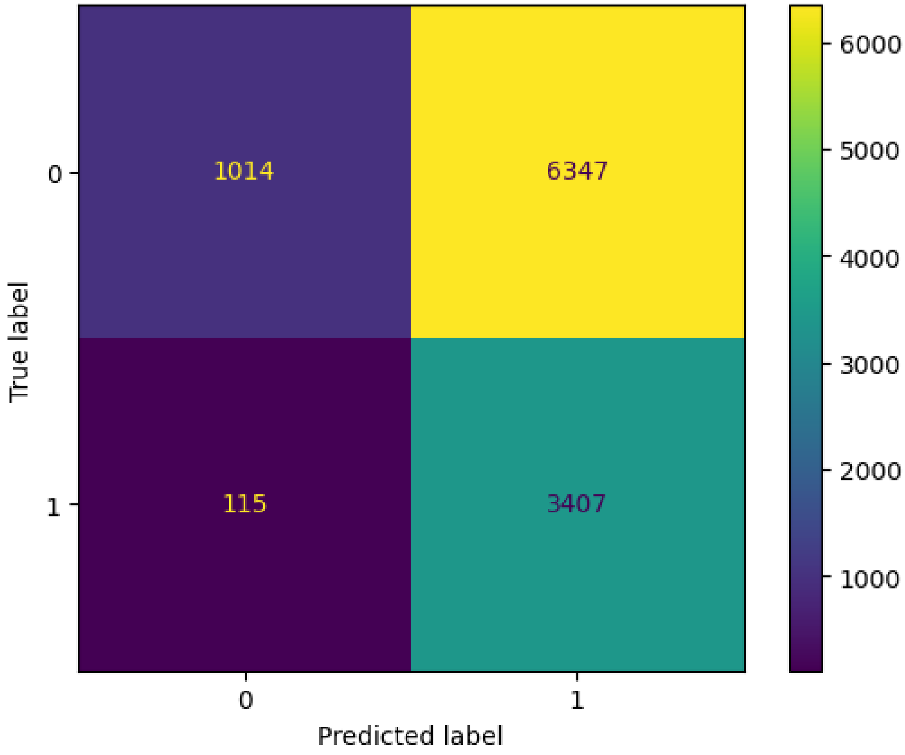


Figure 31: Confusion Matrix

So we have,

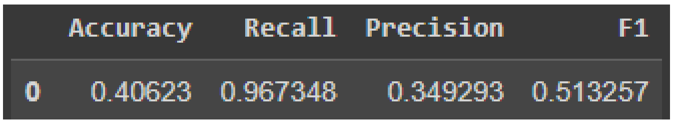
True positive – 3407 values

False Positive – 6347 values

False Negative – 115 values

True Negative – 1014 values

Performance the model in Testing dataset:



	Accuracy	Recall	Precision	F1
0	0.40623	0.967348	0.349293	0.513257

Figure 32: Performance of the model

- The recall value for training dataset for logistic regression before fine tuning is 0.967.

****

KNN for K = 3:

Model Score for Training dataset:

- We have a Model score for our dataset in KNN algorithm for training dataset is 0.914.

Confusion matrix on Training dataset:

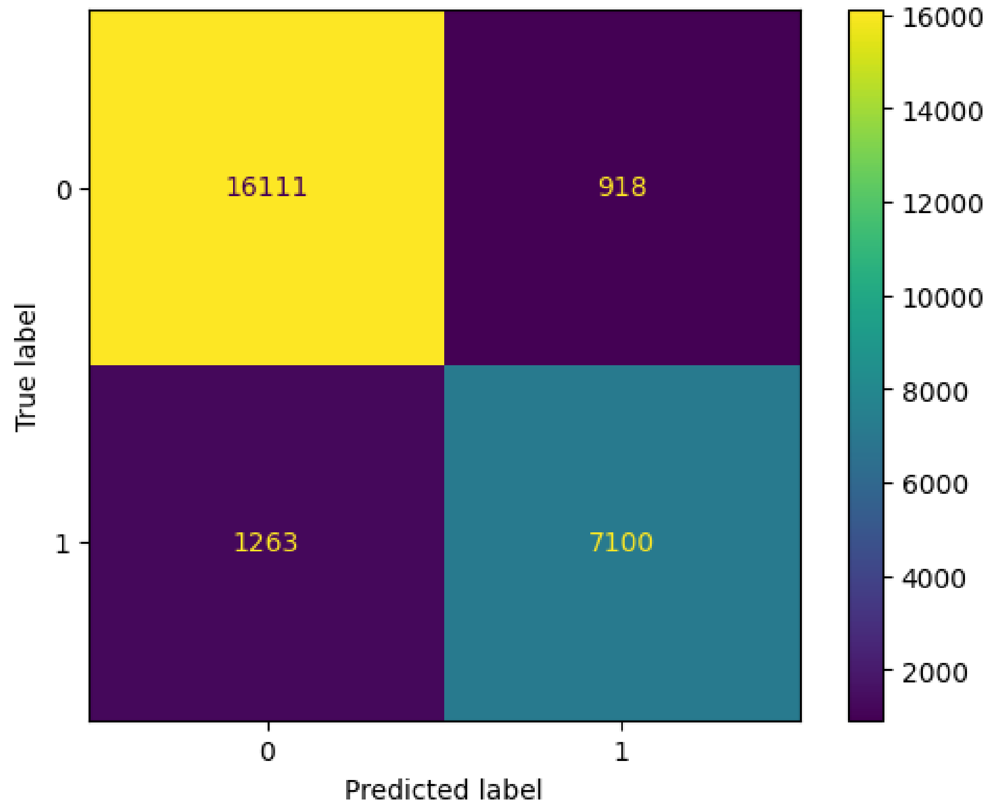


Figure 33: Confusion Matrix

So we have,

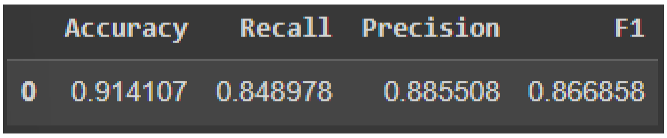
True positive – 7100 values.

False Positive – 918 values.

False Negative – 1263 values.

True Negative – 16111values.

Performance the model in Training dataset:



	Accuracy	Recall	Precision	F1
0	0.914107	0.848978	0.885508	0.866858

Figure 34: Performance of the model

- The recall value for training dataset for logistic regression before fine tuning is 0.848.

Model Score for Testing dataset:

- We have a Model score for our dataset in KNN algorithm for Testing dataset is 0.849.

Confusion matrix for Testing dataset:

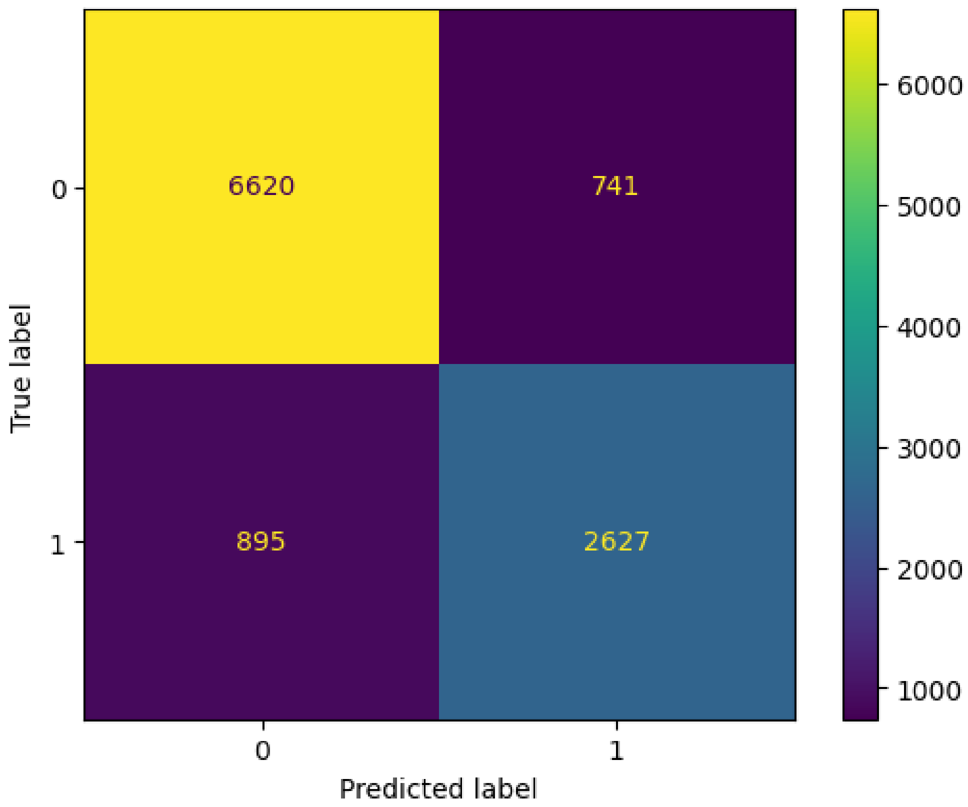


Figure 35: Confusion Matrix

So we have,

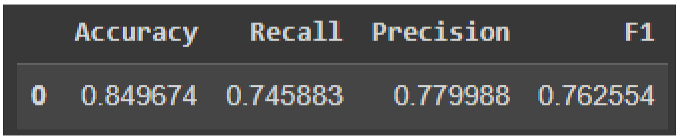
True positive – 2627 values

False Positive – 741 values

False Negative – 895 values

True Negative – 6620 values

Performance the model in Testing dataset:



	Accuracy	Recall	Precision	F1
0	0.849674	0.745883	0.779988	0.762554

Figure 36: Performance of the model

- The recall value for training dataset for logistic regression before fine tuning is 0.745.

****

Decision Tree:

Modal Score for Training dataset:

- We have a Model score for our dataset in KNN algorithm for training dataset is 0.849.

Confusion matrix on Training dataset:

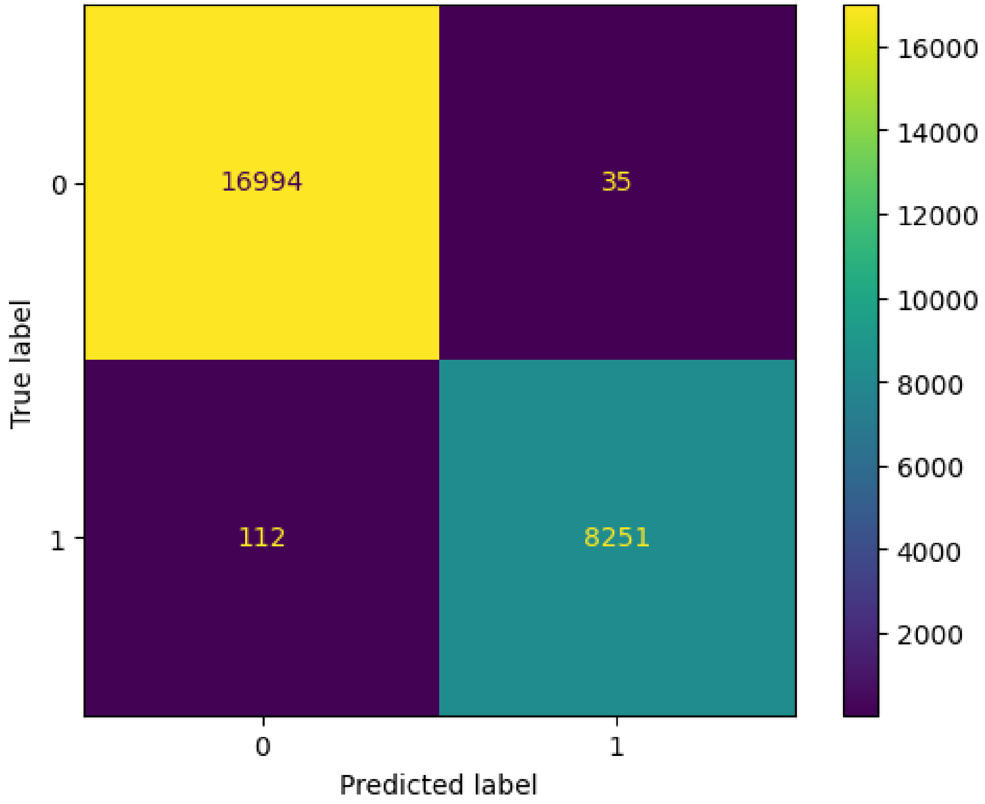


Figure 37: Confusion Matrix

So we have,

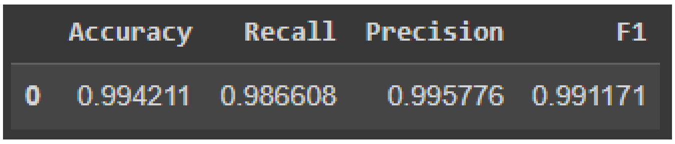
True positive – 8251 values

False Positive – 35 values

False Negative – 112 values

True Negative – 16994 values

Performance of Decision tree model for Training dataset:



	Accuracy	Recall	Precision	F1
0	0.994211	0.986608	0.995776	0.991171

Figure 38: Performance of the model

- The recall for Decision tree modal for Training dataset in 0.986 which is good.

Model Score for Testing datasets:

- We have a Model score for our dataset in KNN algorithm for training dataset is 0.873.

Confusion matrix for Testing dataset:

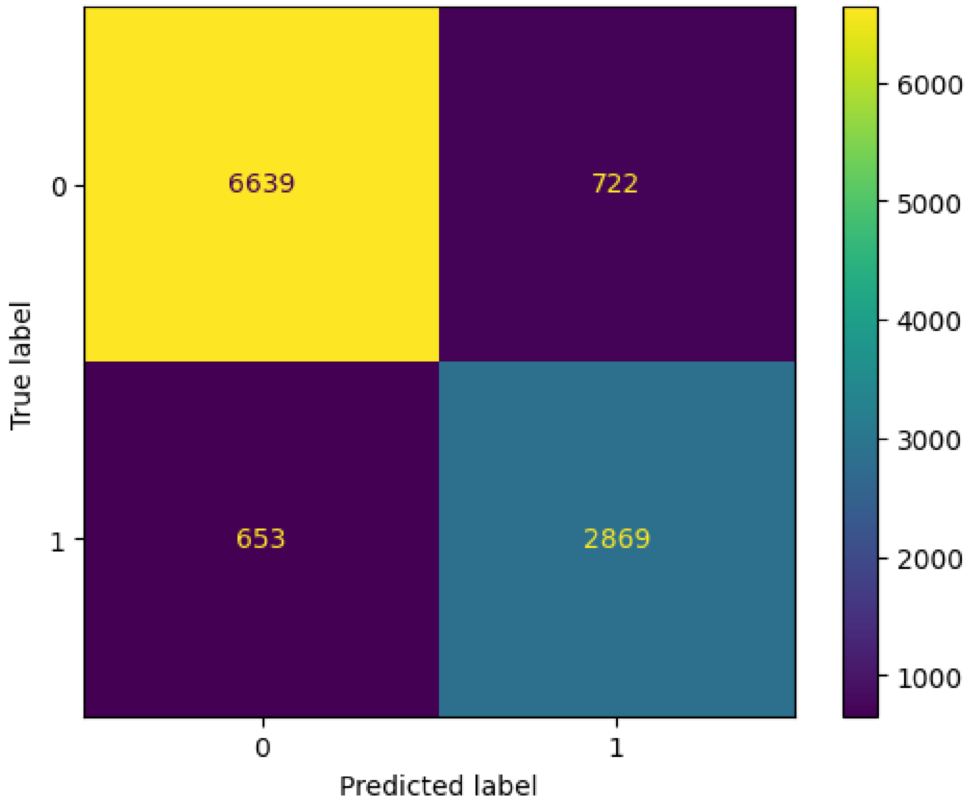


Figure 39: Confusion Matrix

So we have,

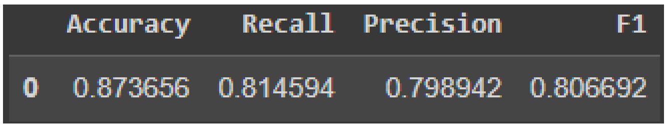
True positive – 8251 values

False Positive – 35 values

False Negative – 112 values

True Negative – 16994 values

Performance of the model for Testing data:



```
   Accuracy    Recall  Precision        F1
0  0.873656  0.814594   0.798942  0.806692
```

Figure 40: Performance of the model

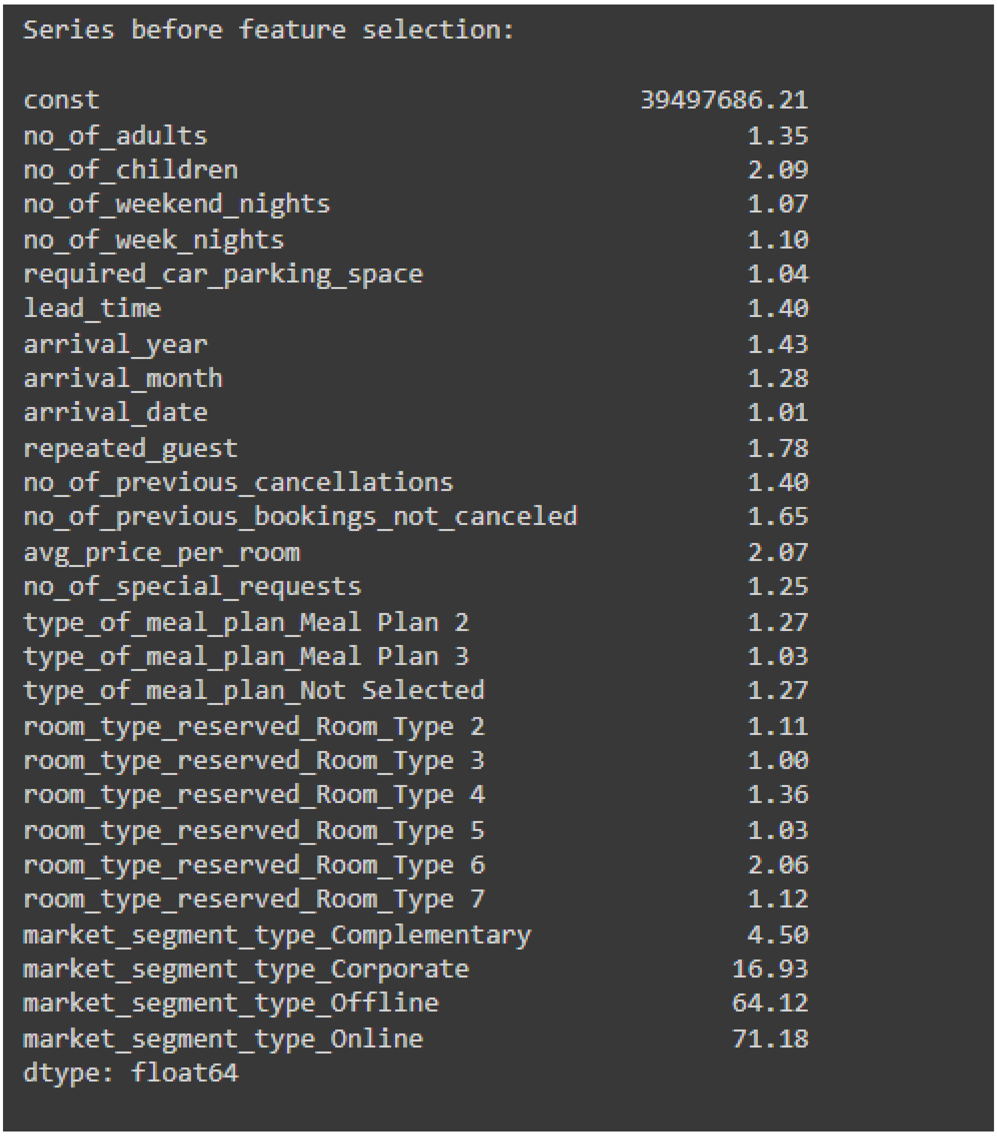
- The model has a recall value of 0.814 which seems to be good.

****

Fine Tuning the Model:

Logistic Regression:

1. **Checking Multicollenearity:**

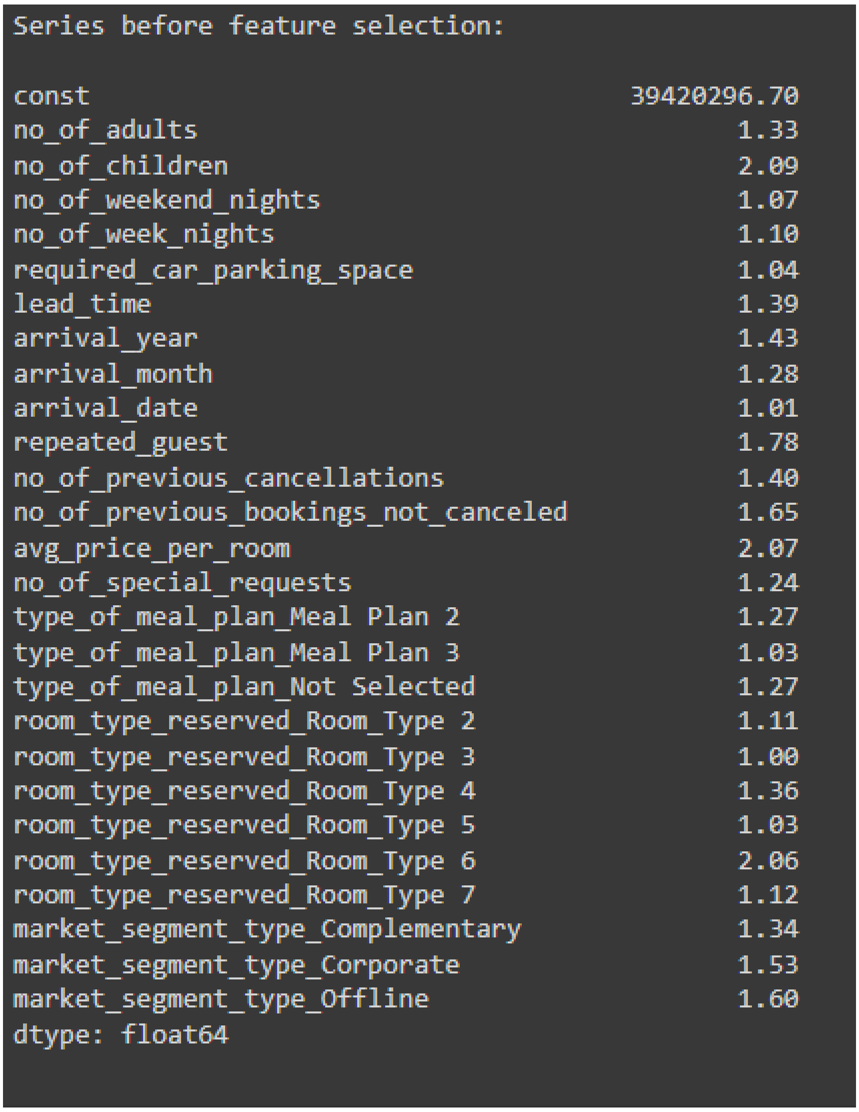


```
Series before feature selection:

const                                  39497686.21
no_of_adults                                  1.35
no_of_children                                2.09
no_of_weekend_nights                          1.07
no_of_week_nights                             1.10
required_car_parking_space                    1.04
lead_time                                     1.40
arrival_year                                  1.43
arrival_month                                 1.28
arrival_date                                  1.01
repeated_guest                                1.78
no_of_previous_cancellations                  1.40
no_of_previous_bookings_not_canceled          1.65
avg_price_per_room                            2.07
no_of_special_requests                        1.25
type_of_meal_plan_Meal Plan 2                 1.27
type_of_meal_plan_Meal Plan 3                 1.03
type_of_meal_plan_Not Selected                1.27
room_type_reserved_Room_Type 2                1.11
room_type_reserved_Room_Type 3                1.00
room_type_reserved_Room_Type 4                1.36
room_type_reserved_Room_Type 5                1.03
room_type_reserved_Room_Type 6                2.06
room_type_reserved_Room_Type 7                1.12
market_segment_type_Complementary             4.50
market_segment_type_Corporate                16.93
market_segment_type_Offline                  64.12
market_segment_type_Online                   71.18
dtype: float64
```

Figure 41: VIF

- We are having `market_segment_type_Corporate`, `market_segment_type_Offline` and `market_segment_type_Online` having VIF> 5 which exhibits high Multicollinearity so we are removing it one by one.

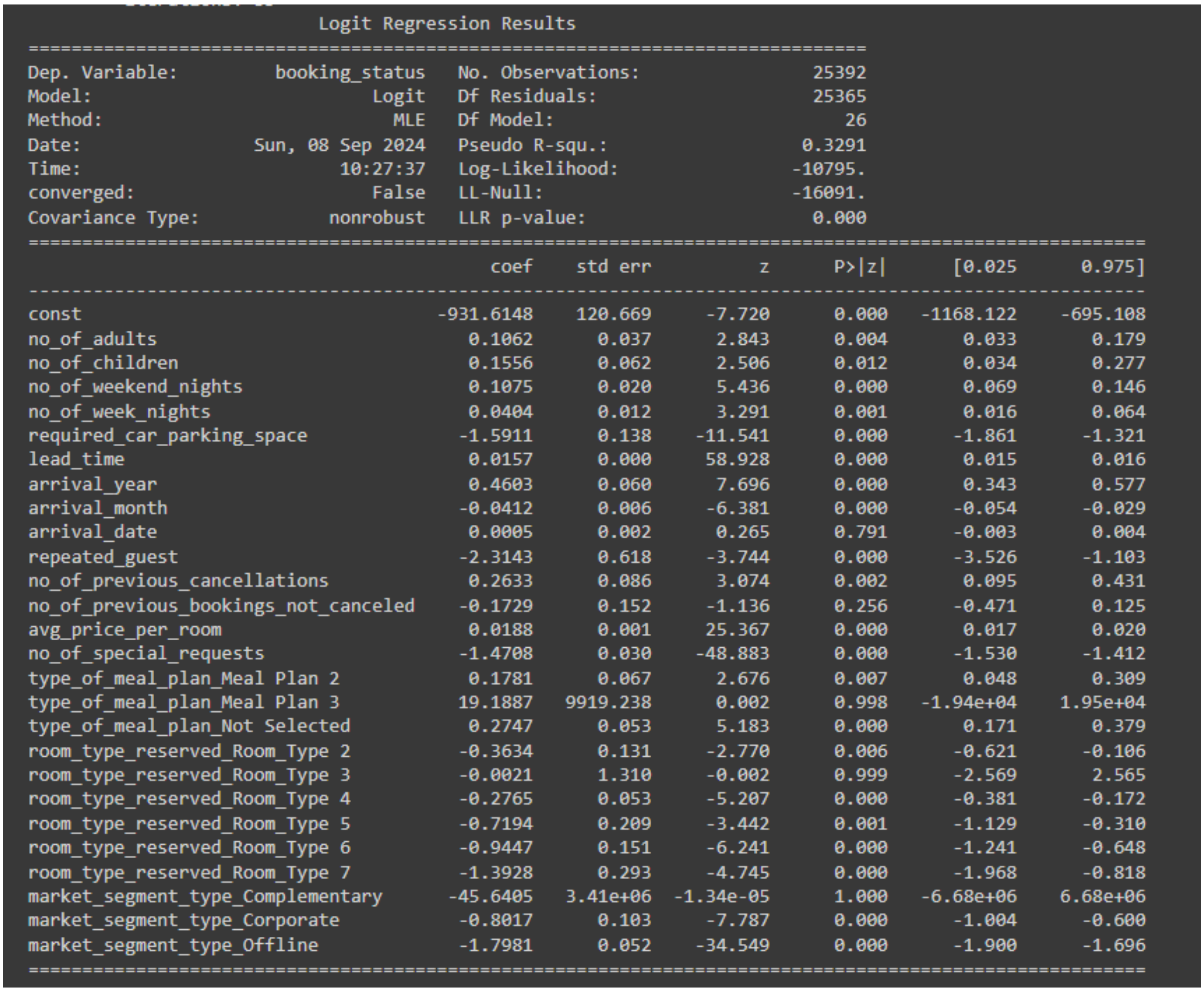


```
Series before feature selection:

const                                      39420296.70
no_of_adults                                      1.33
no_of_children                                    2.09
no_of_weekend_nights                              1.07
no_of_week_nights                                 1.10
required_car_parking_space                        1.04
lead_time                                         1.39
arrival_year                                      1.43
arrival_month                                     1.28
arrival_date                                      1.01
repeated_guest                                    1.78
no_of_previous_cancellations                      1.40
no_of_previous_bookings_not_canceled              1.65
avg_price_per_room                                2.07
no_of_special_requests                            1.24
type_of_meal_plan_Meal Plan 2                     1.27
type_of_meal_plan_Meal Plan 3                     1.03
type_of_meal_plan_Not Selected                    1.27
room_type_reserved_Room_Type 2                    1.11
room_type_reserved_Room_Type 3                    1.00
room_type_reserved_Room_Type 4                    1.36
room_type_reserved_Room_Type 5                    1.03
room_type_reserved_Room_Type 6                    2.06
room_type_reserved_Room_Type 7                    1.12
market_segment_type_Complementary                 1.34
market_segment_type_Corporate                     1.53
market_segment_type_Offline                       1.60
dtype: float64
```

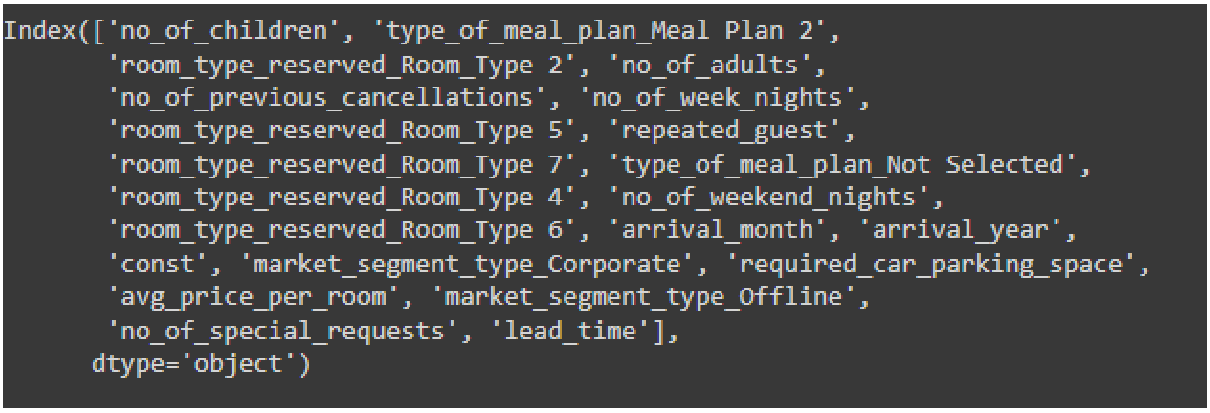
Figure 42: VIF values

- Now we don't have any VIF value > 5.



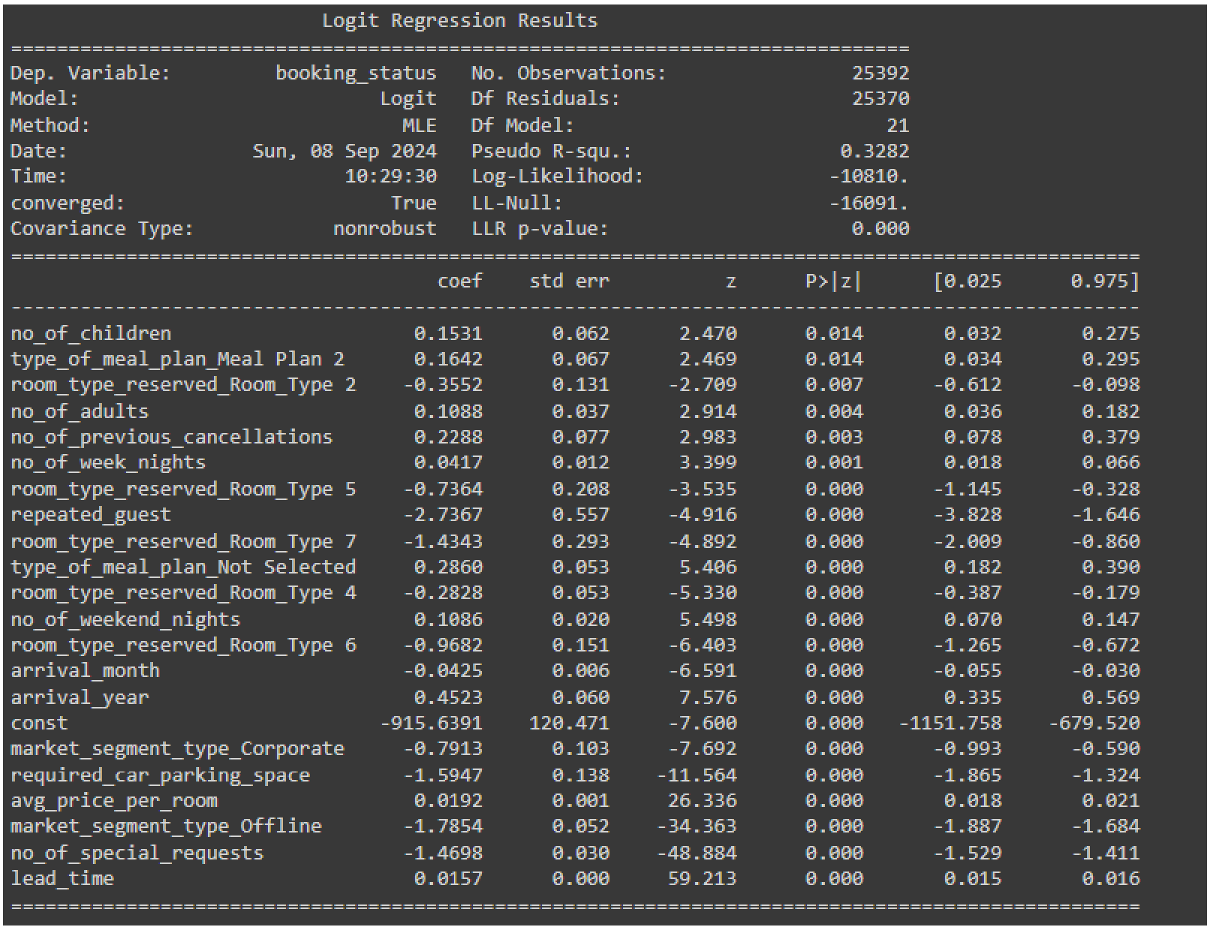
```
                           Logit Regression Results
==============================================================================
Dep. Variable:         booking_status   No. Observations:                25392
Model:                          Logit   Df Residuals:                    25365
Method:                           MLE   Df Model:                           26
Date:                Sun, 08 Sep 2024   Pseudo R-squ.:                  0.3291
Time:                        10:27:37   Log-Likelihood:                -10795.
converged:                      False   LL-Null:                       -16091.
Covariance Type:            nonrobust   LLR p-value:                     0.000
========================================================================================================
                                           coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------
const                                 -931.6148    120.669     -7.720      0.000   -1168.122    -695.108
no_of_adults                             0.1062      0.037      2.843      0.004       0.033       0.179
no_of_children                           0.1556      0.062      2.506      0.012       0.034       0.277
no_of_weekend_nights                     0.1075      0.020      5.436      0.000       0.069       0.146
no_of_week_nights                        0.0404      0.012      3.291      0.001       0.016       0.064
required_car_parking_space              -1.5911      0.138    -11.541      0.000      -1.861      -1.321
lead_time                                0.0157      0.000     58.928      0.000       0.015       0.016
arrival_year                             0.4603      0.060      7.696      0.000       0.343       0.577
arrival_month                           -0.0412      0.006     -6.381      0.000      -0.054      -0.029
arrival_date                             0.0005      0.002      0.265      0.791      -0.003       0.004
repeated_guest                          -2.3143      0.618     -3.744      0.000      -3.526      -1.103
no_of_previous_cancellations             0.2633      0.086      3.074      0.002       0.095       0.431
no_of_previous_bookings_not_canceled    -0.1729      0.152     -1.136      0.256      -0.471       0.125
avg_price_per_room                       0.0188      0.001     25.367      0.000       0.017       0.020
no_of_special_requests                  -1.4708      0.030    -48.883      0.000      -1.530      -1.412
type_of_meal_plan_Meal Plan 2            0.1781      0.067      2.676      0.007       0.048       0.309
type_of_meal_plan_Meal Plan 3           19.1887   9919.238      0.002      0.998   -1.94e+04    1.95e+04
type_of_meal_plan_Not Selected           0.2747      0.053      5.183      0.000       0.171       0.379
room_type_reserved_Room_Type 2          -0.3634      0.131     -2.770      0.006      -0.621      -0.106
room_type_reserved_Room_Type 3          -0.0021      1.310     -0.002      0.999      -2.569       2.565
room_type_reserved_Room_Type 4          -0.2765      0.053     -5.207      0.000      -0.381      -0.172
room_type_reserved_Room_Type 5          -0.7194      0.209     -3.442      0.001      -1.129      -0.310
room_type_reserved_Room_Type 6          -0.9447      0.151     -6.241      0.000      -1.241      -0.648
room_type_reserved_Room_Type 7          -1.3928      0.293     -4.745      0.000      -1.968      -0.818
market_segment_type_Complementary      -45.6405   3.41e+06  -1.34e-05      1.000   -6.68e+06    6.68e+06
market_segment_type_Corporate           -0.8017      0.103     -7.787      0.000      -1.004      -0.600
market_segment_type_Offline             -1.7981      0.052    -34.549      0.000      -1.900      -1.696
========================================================================================================
```

- Checking whether the P-value is greater than 0.05.
- There are some values > 0.05 which we should be removing.



```
Index(['no_of_children', 'type_of_meal_plan_Meal Plan 2',
       'room_type_reserved_Room_Type 2', 'no_of_adults',
       'no_of_previous_cancellations', 'no_of_week_nights',
       'room_type_reserved_Room_Type 5', 'repeated_guest',
       'room_type_reserved_Room_Type 7', 'type_of_meal_plan_Not Selected',
       'room_type_reserved_Room_Type 4', 'no_of_weekend_nights',
       'room_type_reserved_Room_Type 6', 'arrival_month', 'arrival_year',
       'const', 'market_segment_type_Corporate', 'required_car_parking_space',
       'avg_price_per_room', 'market_segment_type_Offline',
       'no_of_special_requests', 'lead_time'],
      dtype='object')
```

- The above columns are having the values <= 0.05
- We are going to add these columns in the new variable.



```
                           Logit Regression Results
==============================================================================
Dep. Variable:         booking_status   No. Observations:                25392
Model:                          Logit   Df Residuals:                    25370
Method:                           MLE   Df Model:                           21
Date:                Sun, 08 Sep 2024   Pseudo R-squ.:                  0.3282
Time:                        10:29:30   Log-Likelihood:                -10810.
converged:                       True   LL-Null:                       -16091.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==================================================================================================
                                     coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
no_of_children                     0.1531      0.062      2.470      0.014       0.032       0.275
type_of_meal_plan_Meal Plan 2      0.1642      0.067      2.469      0.014       0.034       0.295
room_type_reserved_Room_Type 2    -0.3552      0.131     -2.709      0.007      -0.612      -0.098
no_of_adults                       0.1088      0.037      2.914      0.004       0.036       0.182
no_of_previous_cancellations       0.2288      0.077      2.983      0.003       0.078       0.379
no_of_week_nights                  0.0417      0.012      3.399      0.001       0.018       0.066
room_type_reserved_Room_Type 5    -0.7364      0.208     -3.535      0.000      -1.145      -0.328
repeated_guest                    -2.7367      0.557     -4.916      0.000      -3.828      -1.646
room_type_reserved_Room_Type 7    -1.4343      0.293     -4.892      0.000      -2.009      -0.860
type_of_meal_plan_Not Selected     0.2860      0.053      5.406      0.000       0.182       0.390
room_type_reserved_Room_Type 4    -0.2828      0.053     -5.330      0.000      -0.387      -0.179
no_of_weekend_nights               0.1086      0.020      5.498      0.000       0.070       0.147
room_type_reserved_Room_Type 6    -0.9682      0.151     -6.403      0.000      -1.265      -0.672
arrival_month                     -0.0425      0.006     -6.591      0.000      -0.055      -0.030
arrival_year                       0.4523      0.060      7.576      0.000       0.335       0.569
const                           -915.6391    120.471     -7.600      0.000   -1151.758    -679.520
market_segment_type_Corporate     -0.7913      0.103     -7.692      0.000      -0.993      -0.590
required_car_parking_space        -1.5947      0.138    -11.564      0.000      -1.865      -1.324
avg_price_per_room                 0.0192      0.001     26.336      0.000       0.018       0.021
market_segment_type_Offline       -1.7854      0.052    -34.363      0.000      -1.887      -1.684
no_of_special_requests            -1.4698      0.030    -48.884      0.000      -1.529      -1.411
lead_time                          0.0157      0.000     59.213      0.000       0.015       0.016
==================================================================================================
```

- Now we are clean with the P-value all the values are <= 0.05.

ROC-AUC:

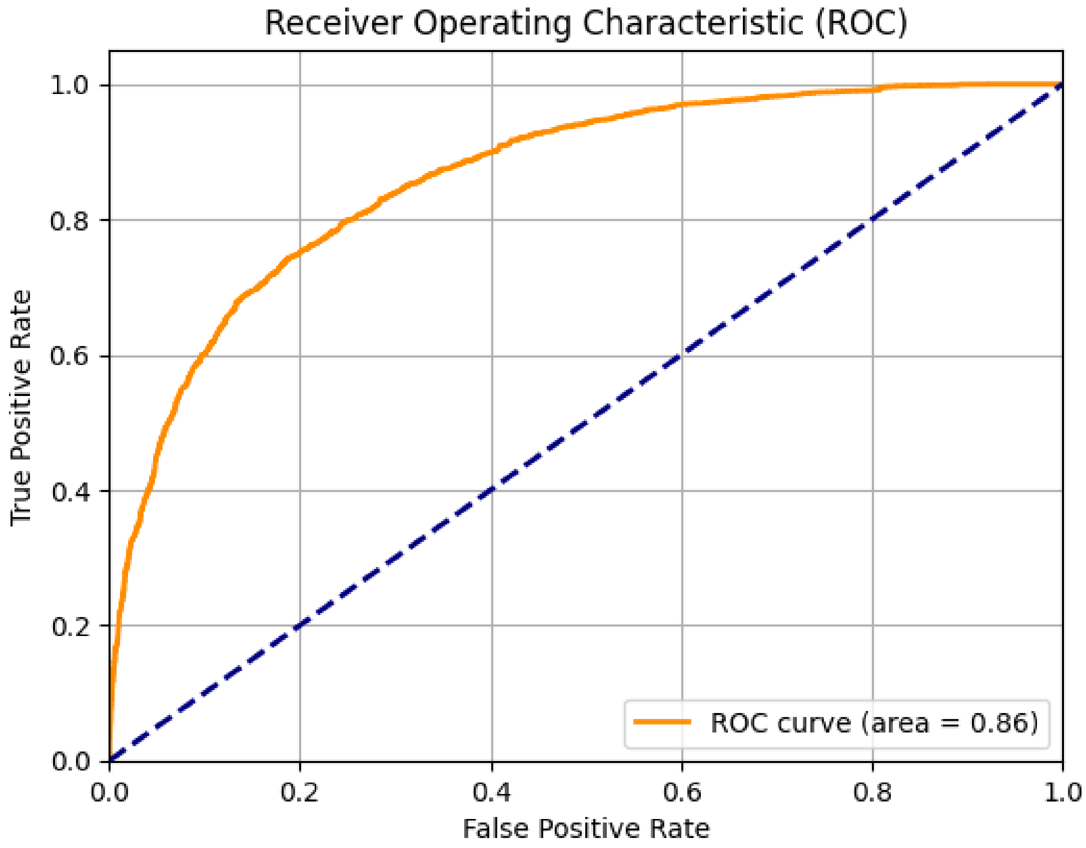


Figure 43: ROC-AUC curve

- The curve seems to be above the Threshold value which is 0.362.

Confusion Matrix for Training data after Fine Tuning:

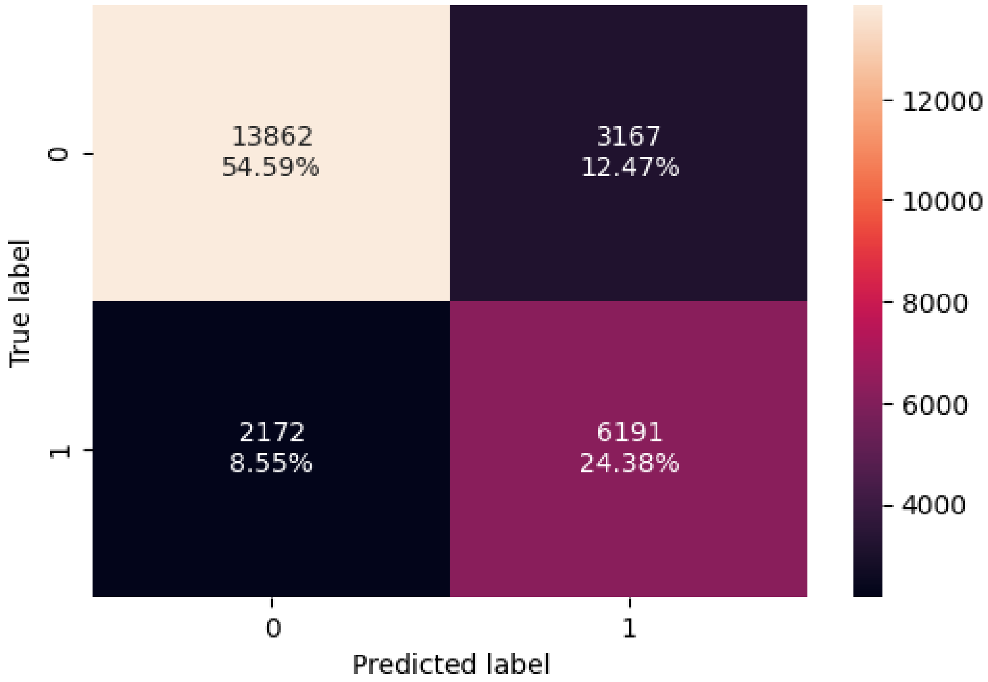


Figure 44: Confusion Matrix

So we have,

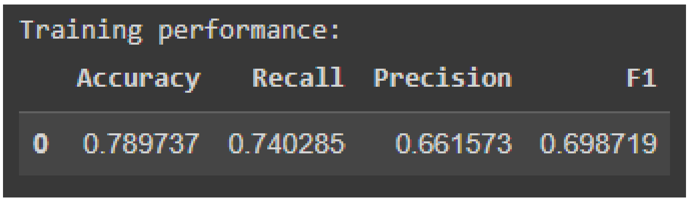
True positive – 6191 values

False Positive – 3167 values

False Negative – 2172 values

True Negative – 13862 values

Performance of the model for Training after fine tuning:



```
Training performance:
   Accuracy    Recall  Precision        F1
0  0.789737  0.740285   0.661573  0.698719
```

Figure 45: Performance of the model

- The recall value of logistic regression model is 0.740 which seems to be good.

Confusion Matrix for Testing data after Fine Tuning:

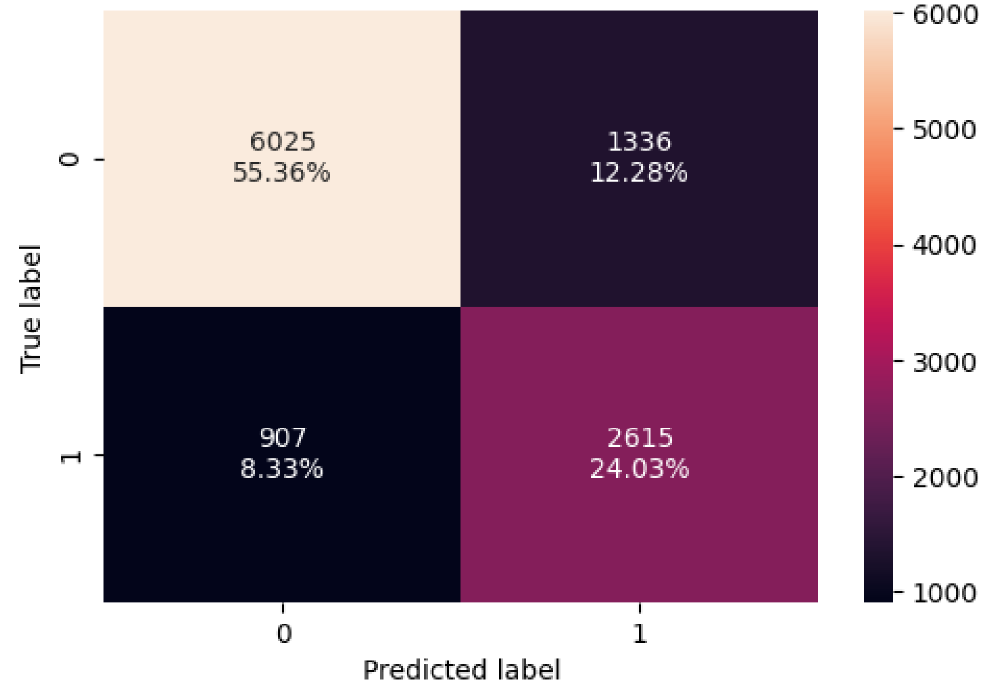


Figure 46: Confusion Matrix

So we have,

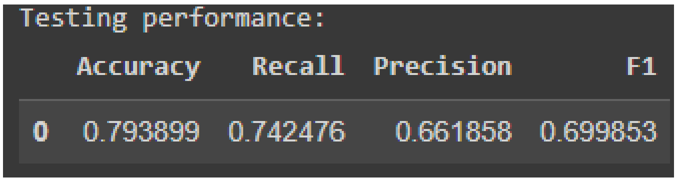
True positive – 2615 values

False Positive – 1336 values

False Negative – 907 values

True Negative – 6025 values

Perfomance of the model on testing data after fine tuning:



```
Testing performance:
    Accuracy    Recall  Precision        F1
0   0.793899  0.742476   0.661858  0.699853
```

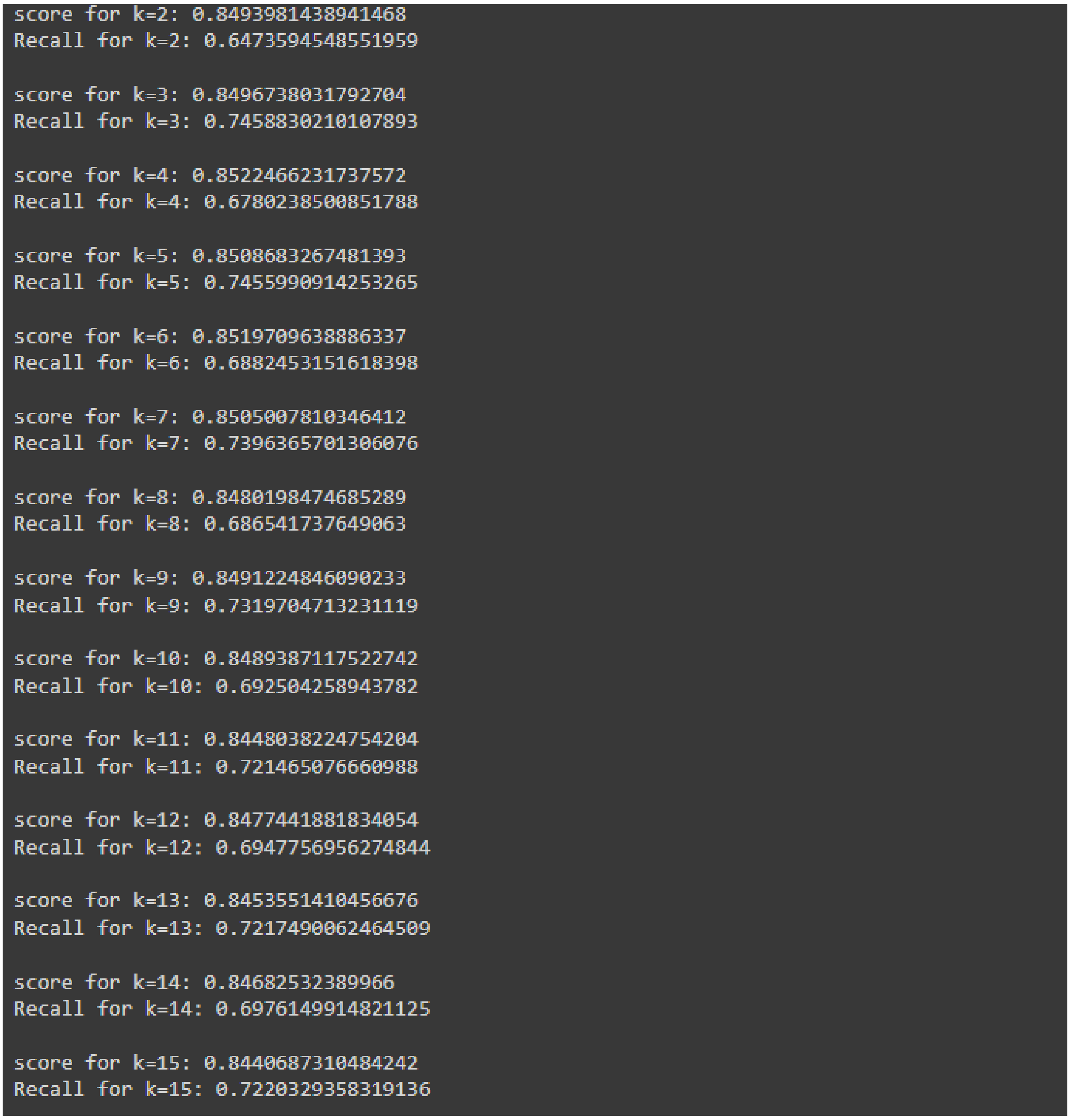
Figure 47: Performance of the model

- The recall value of logistic regression model is 0.742 which seems to be good.

****

KNN:

Checking for multiple values from 2 to 20 for K:



```
score for k=2: 0.8493981438941468
Recall for k=2: 0.6473594548551959

score for k=3: 0.8496738031792704
Recall for k=3: 0.7458830210107893

score for k=4: 0.8522466231737572
Recall for k=4: 0.6780238500851788

score for k=5: 0.8508683267481393
Recall for k=5: 0.7455990914253265

score for k=6: 0.8519709638886337
Recall for k=6: 0.6882453151618398

score for k=7: 0.8505007810346412
Recall for k=7: 0.7396365701306076

score for k=8: 0.8480198474685289
Recall for k=8: 0.686541737649063

score for k=9: 0.8491224846090233
Recall for k=9: 0.7319704713231119

score for k=10: 0.8489387117522742
Recall for k=10: 0.692504258943782

score for k=11: 0.8448038224754204
Recall for k=11: 0.721465076660988

score for k=12: 0.8477441881834054
Recall for k=12: 0.6947756956274844

score for k=13: 0.8453551410456676
Recall for k=13: 0.7217490062464509

score for k=14: 0.84682532389966
Recall for k=14: 0.6976149914821125

score for k=15: 0.8440687310484242
Recall for k=15: 0.7220329358319136
```

Figure 48: KNN with multiple K values

- Seems like K = 3 value is good for our dataset base on the recall K: 3 is 0.745 and the score is 0.849.

Confusion Matrix for KNN:

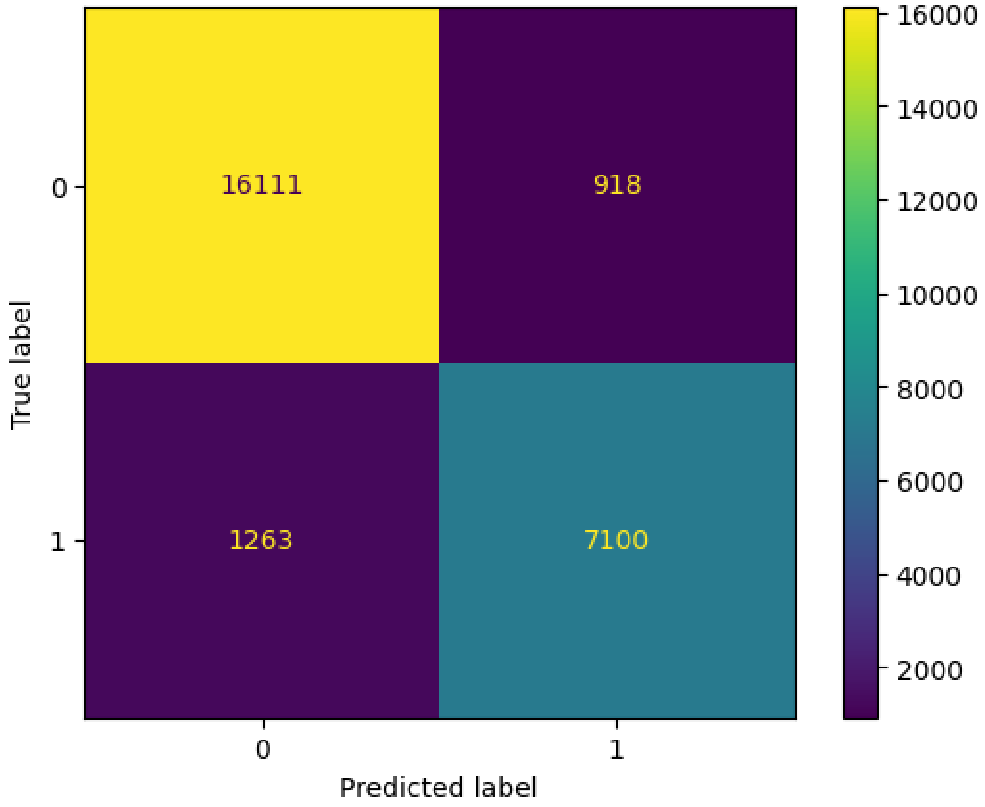


Figure 49: Confusion Matrix

So we have,

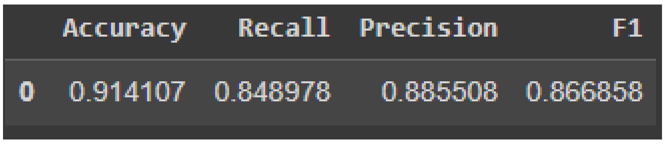
True positive – 7100 values

False Positive – 918 values

False Negative – 1263 values

True Negative – 16111 values

Performance of KNN model for K = 3:



	Accuracy	Recall	Precision	F1
0	0.914107	0.848978	0.885508	0.866858

Figure 50: Performance of the model

- The recall value for the Training data for K = 3 is 0.754 which is good

Confusion Matrix for KNN:

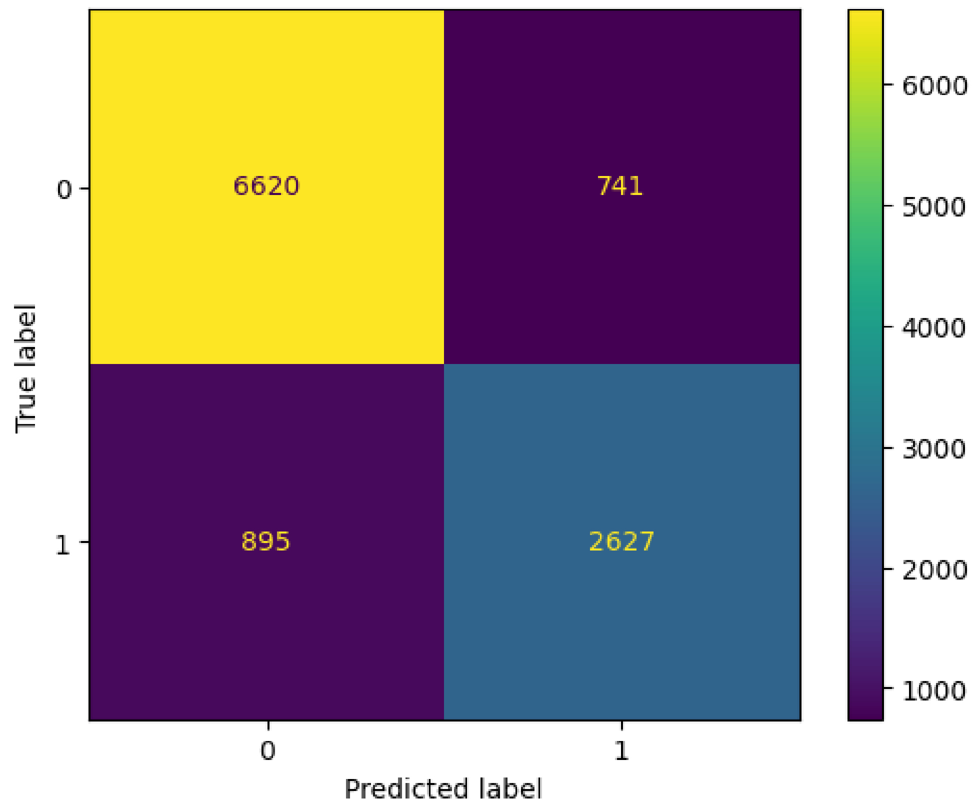


Figure 51: Confusion Matrix

So we have,

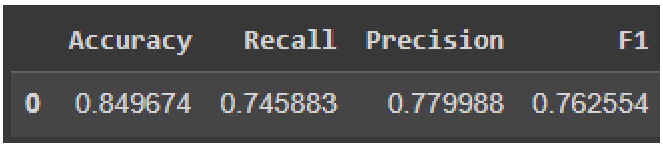
True positive – 2627 values

False Positive – 741 values

False Negative – 895 values

True Negative – 6620 values

Performance of KNN model for K = 3:



	Accuracy	Recall	Precision	F1
0	0.849674	0.745883	0.779988	0.762554

Figure 52: Performance of the model

- The recall value for the testing data for K = 3 is 0.754 which is good after fine tuning.
- The value of k is the same (k = 3) as the base model. So the model performance remains unaltered on both training and test sets.

****

Decision Tree:

Pruning:

Confusion Matrix for training data after pre-pruning:

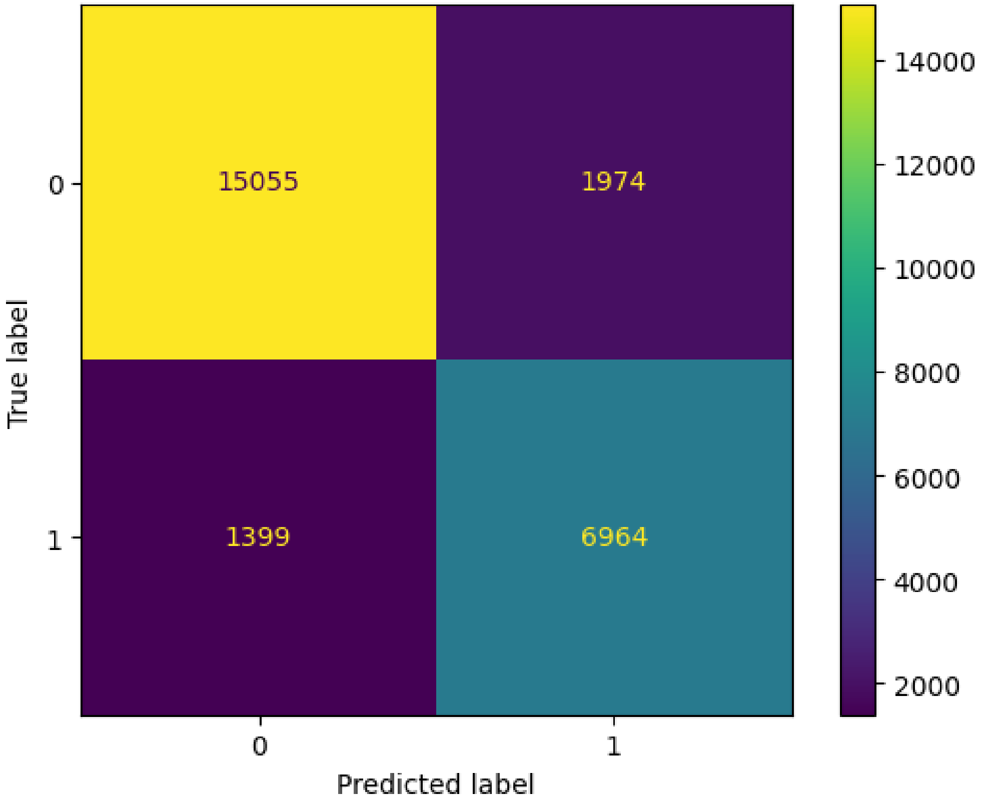


Figure 53: Confusion Matrix

So we have,

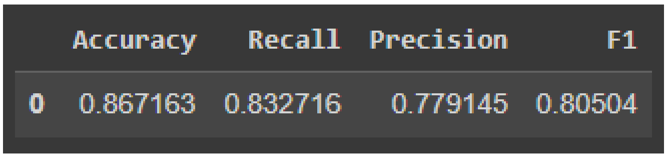
True positive – 6964 values

False Positive – 1974 values

False Negative – 1399 values

True Negative – 15055 values

Performance of Decision Tree after pre-pruning:



	Accuracy	Recall	Precision	F1
0	0.867163	0.832716	0.779145	0.80504

Figure 54: Performance of the model

- The recall value of the decision tree for training data after pruning is 0.832 which good

Confusion Matrix for testing data after pre-pruning:

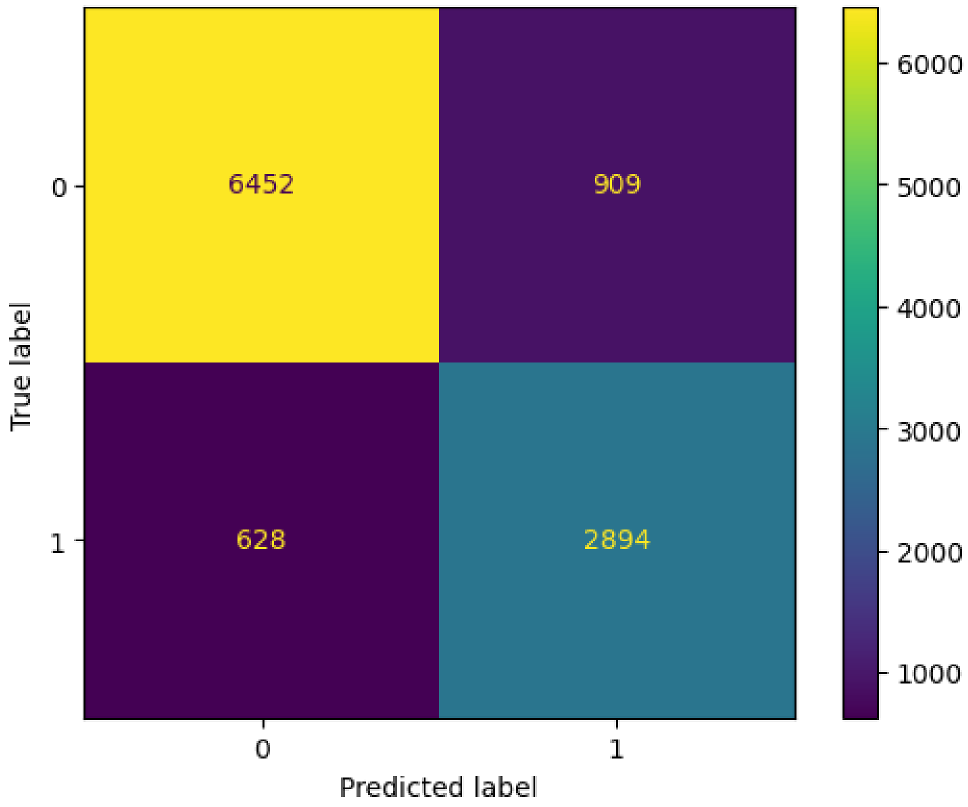


Figure 55: Confusion Matrix

So we have,

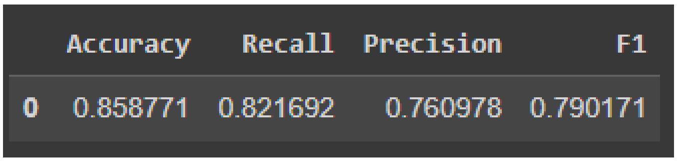
True positive – 6452 values

False Positive – 909 values

False Negative – 628 values

True Negative – 6452 values

Performance of Decision Tree after pre-pruning:



	Accuracy	Recall	Precision	F1
0	0.858771	0.821692	0.760978	0.790171

Figure 56: Performance of the model

- The recall value of the decision tree for training data after pruning is 0.821 which good and better compared to other models.

Visualizing the decision tree:

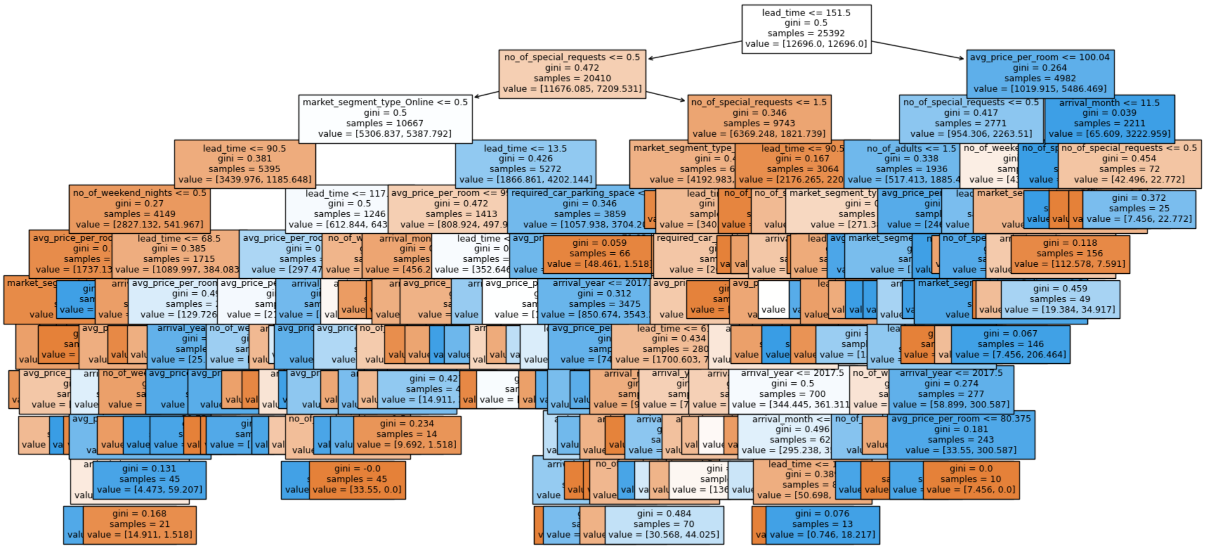


Figure 57: Decision Tree visual

- This is the visualization of the decision tree after pruning.

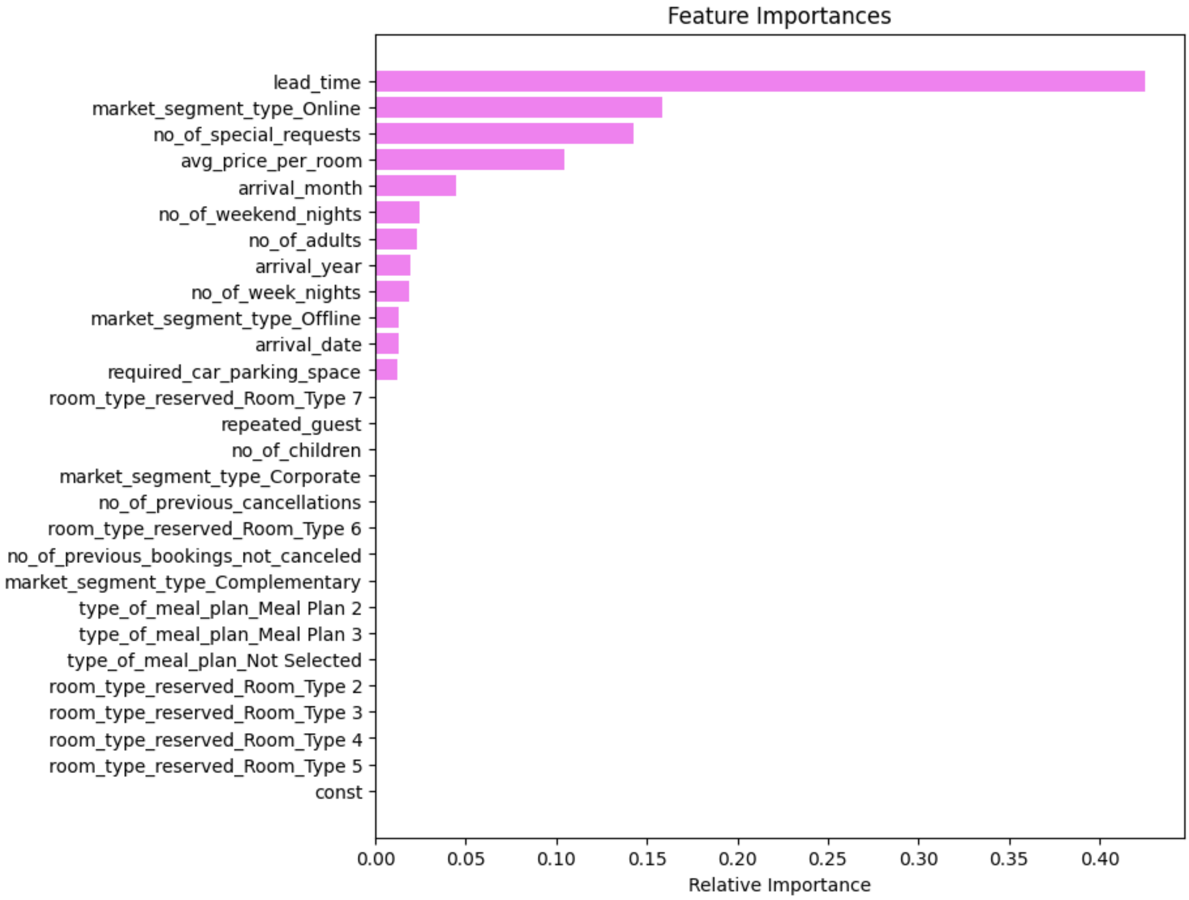
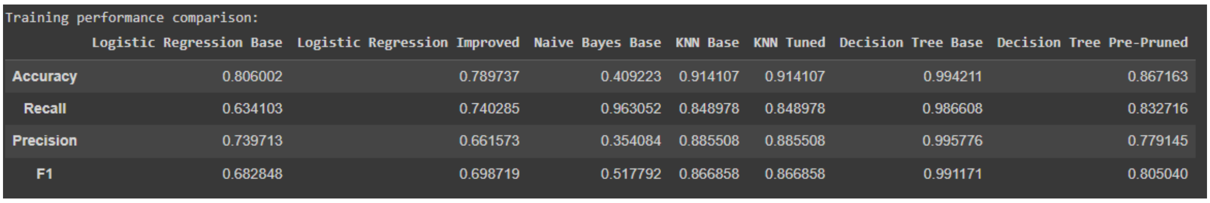


Figure 58: Feature importance

- We can conclude that the lead time feature seems to be much more important than others.

The performance comparision of models for training dataset:

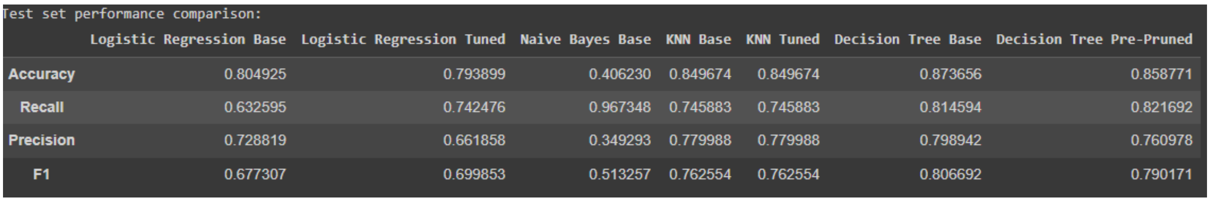


Training performance comparison:

	Logistic Regression Base	Logistic Regression Improved	Naive Bayes Base	KNN Base	KNN Tuned	Decision Tree Base	Decision Tree Pre-Pruned
Accuracy	0.806002	0.789737	0.409223	0.914107	0.914107	0.994211	0.867163
Recall	0.634103	0.740285	0.963052	0.848978	0.848978	0.986608	0.832716
Precision	0.739713	0.661573	0.354084	0.885508	0.885508	0.995776	0.779145
F1	0.682848	0.698719	0.517792	0.866858	0.866858	0.991171	0.805040

Figure 59: Performance comparision on Training dataset

The performance comparision of models for testing dataset:



Test set performance comparison:

	Logistic Regression Base	Logistic Regression Tuned	Naive Bayes Base	KNN Base	KNN Tuned	Decision Tree Base	Decision Tree Pre-Pruned
Accuracy	0.804925	0.793899	0.406230	0.849674	0.849674	0.873656	0.858771
Recall	0.632595	0.742476	0.967348	0.745883	0.745883	0.814594	0.821692
Precision	0.728819	0.661858	0.349293	0.779988	0.779988	0.798942	0.760978
F1	0.677307	0.699853	0.513257	0.762554	0.762554	0.806692	0.790171

Figure 60: Performance comparision on testing dataset

- For testing data we can clearly conclude that the Decision tree is much more suitable for this dataset and is the correct to choose

---

Actionable Insights and Business Recommendations

Based on the analysis of hotel booking data and the machine learning model, here are key insights and strategic recommendations for improving business performance:

1. Customer Segmentation and Personalization

- **Insight**: Distinct customer profiles (e.g., family size, meal plans, special requests) allow for segmentation.
- **Recommendation**: Develop targeted marketing campaigns and personalized offers, such as family packages or meal plan upgrades, to increase conversion rates and enhance customer experience.

2. Cancellation Prediction and Prevention

- **Insight**: The model predicts high-risk bookings based on factors like previous cancellations and booking behavior.
- **Recommendation**: Proactively engage customers with high cancellation risk through personalized offers or reminders, and adjust cancellation policies based on customer segmentation to reduce cancellations.

3. Revenue Management and Dynamic Pricing

- **Insight**: Room pricing and booking trends reveal price sensitivity among customer groups.
- **Recommendation**: Implement dynamic pricing strategies to adjust room rates in real-time based on demand and customer behaviour. Optimize cross-selling of services like parking or meal plans to boost revenue.

4. Service Enhancements

- **Insight**: Analysis of special requests highlights customer preferences for specific services.
- **Recommendation**: Tailor offerings (e.g., parking availability, meal plan options) to meet common customer needs, improving satisfaction and operational efficiency.

5. Customer Loyalty and Retention

- **Insight**: Repeat bookings provide an opportunity to boost long-term value.
- **Recommendation**: Create a loyalty program with personalized benefits, encouraging repeat bookings and fostering customer loyalty through exclusive offers or room upgrades.

6. Operational Efficiency

- **Insight**: Booking patterns provide insights for better resource allocation.
- **Recommendation**: Use data to optimize staffing and inventory management, particularly during peak times, reducing operational costs while ensuring customer needs are met.

Strategic Outcomes

- **Increased Revenue** through dynamic pricing and personalized offers.
- **Reduced Cancellations** via proactive engagement with high-risk customers.
- **Improved Satisfaction** through tailored services and enhanced operational efficiency.
- **Stronger Customer Loyalty** by implementing a data-driven loyalty program.

These recommendations will help improve revenue, customer retention, and operational performance.