Business Report

ML 1 Coded Project

PGPDSBA

Chithira Raj

Table of Contents

List of Tables

List of Figures

1. Context

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

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

The cancellation of bookings impacts a hotel on various fronts:

1. Loss of resources (revenue) when the hotel cannot resell the room.

2. Additional costs of distribution channels by increasing commissions or paying for publicity to help sell these rooms.

3. Lowering prices last minute, so the hotel can resell a room, resulting in reducing the profit margin.

4. Human resources to make arrangements for the guests.

2. Objective

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

3. Data Dictionary

S.No.	Variables	Description
1	Booking_ID	the unique identifier of each booking
2	no_of_adults	Number of adults
3	no_of_children	Number of Children
4	no_of_weekend_nights	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
5	no_of_week_nights	Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel
6	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)
7	required_car_parking_space	Does the customer require a car parking space? (0 - No, 1-Yes)
8	room_type_reserved	Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels Group
9	lead_time	Number of days between the date of booking and the arrival date
10	arrival_year	Year of arrival date
11	arrival_month	Month of arrival date
12	arrival_date	Date of the month
13	market_segment_type	Market segment designation
14	repeated_guest	Is the customer a repeated guest? (0 - No, 1- Yes)

15	no_of_previous_cancellations	Number of previous bookings that were canceled by the customer prior to the current booking
16	no_of_previous_bookings_not_canceled	Number of previous bookings not canceled by the customer prior to the current booking
17	avg_price_per_room	Average price per day of the reservation; prices of the rooms are dynamic. (in euros)
18	no_of_special_requests	Total number of special requests made by the customer (e.g. high floor, view from the room, etc)
19	booking_status	Flag indicating if the booking was canceled or not.

Table 1: Data Dictionary

4. Data Overview

4.1. Import libraries and load the data

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

Figure 1: Data Overview

4.2. Check the structure of data

Shape of the dataset: 36275 rows and 19 columns

4.3. Check the types of the 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: Datatypes

4.4. Check for and treat (if needed) missing values

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

Figure 3: Missing values check

4.5. Data Duplicates

There are no duplicate rows.

4.6. Statistical Summary

	count	mean	std	min	25%	50%	75%	max
no_of_adults	36275.00000	1.84496	0.51871	0.00000	2.00000	2.00000	2.00000	4.00000
no_of_children	36275.00000	0.10528	0.40265	0.00000	0.00000	0.00000	0.00000	10.00000
no_of_weekend_nights	36275.00000	0.81072	0.87064	0.00000	0.00000	1.00000	2.00000	7.00000
no_of_week_nights	36275.00000	2.20430	1.41090	0.00000	1.00000	2.00000	3.00000	17.00000
required_car_parking_space	36275.00000	0.03099	0.17328	0.00000	0.00000	0.00000	0.00000	1.00000
lead_time	36275.00000	85.23256	85.93082	0.00000	17.00000	57.00000	126.00000	443.00000
arrival_year	36275.00000	2017.82043	0.38384	2017.00000	2018.00000	2018.00000	2018.00000	2018.00000
arrival_month	36275.00000	7.42365	3.06989	1.00000	5.00000	8.00000	10.00000	12.00000
arrival_date	36275.00000	15.59700	8.74045	1.00000	8.00000	16.00000	23.00000	31.00000
repeated_guest	36275.00000	0.02564	0.15805	0.00000	0.00000	0.00000	0.00000	1.00000
no_of_previous_cancellations	36275.00000	0.02335	0.36833	0.00000	0.00000	0.00000	0.00000	13.00000
no_of_previous_bookings_not_canceled	36275.00000	0.15341	1.75417	0.00000	0.00000	0.00000	0.00000	58.00000
avg_price_per_room	36275.00000	103.42354	35.08942	0.00000	80.30000	99.45000	120.00000	540.00000
no_of_special_requests	36275.00000	0.61966	0.78624	0.00000	0.00000	0.00000	1.00000	5.00000

Figure 4: Statistical Summary

4.7. Insights

- The majority of bookings are made for 2 adults, which is evident from the mean (1.84) and the median (2.00) of the no_of_adults column. This suggests that most guests are couples or pairs.
- The average stay includes slightly less than one weekend night (mean of 0.81) and about two weekday nights (mean of 2.20). This suggests that many guests might be staying for a long weekend or a short weekday trip.
- The average lead_time is around 85 days, showing that guests generally book well in advance, which can be beneficial for managing hotel occupancy and revenue strategies.
- The avg_price_per_room shows a mean of approximately 103.42, indicating the average price point for rooms. The large range (from 0 to 540) suggests varied pricing options, likely based on room types or seasons.
- The arrival_year and arrival_month suggest most data is from 2018, and bookings are spread across months with a peak around October. This indicates potential seasonality, with likely higher demand during summer months.

5. Exploratory Data Analysis

5.1. Univariate Analysis

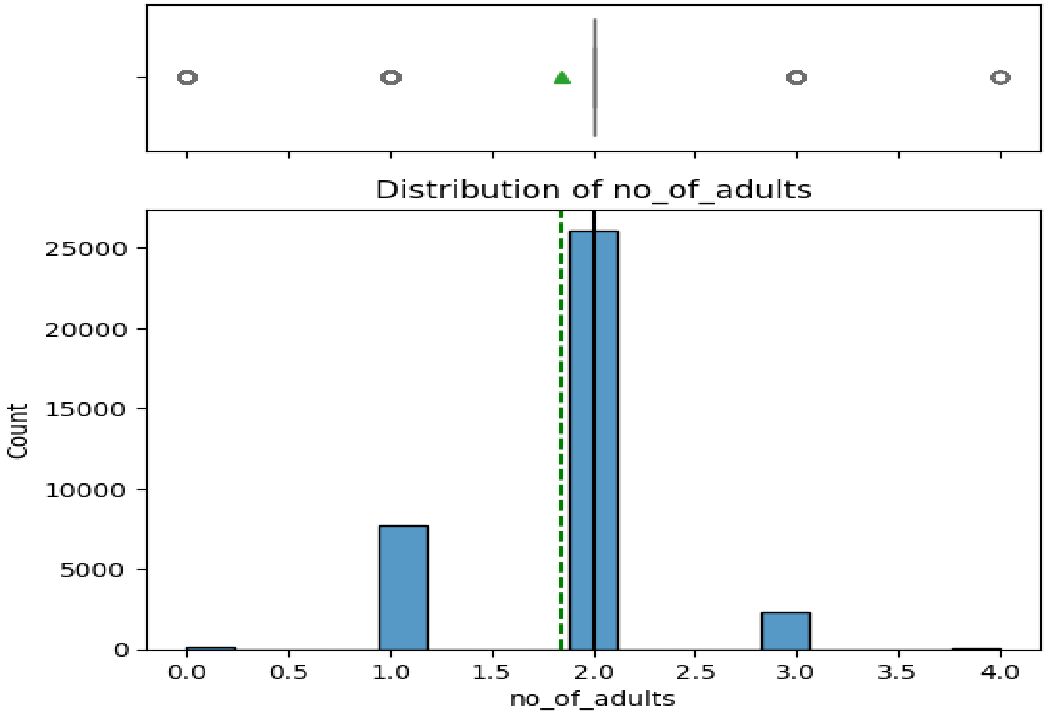

Figure 5: no_of_adults

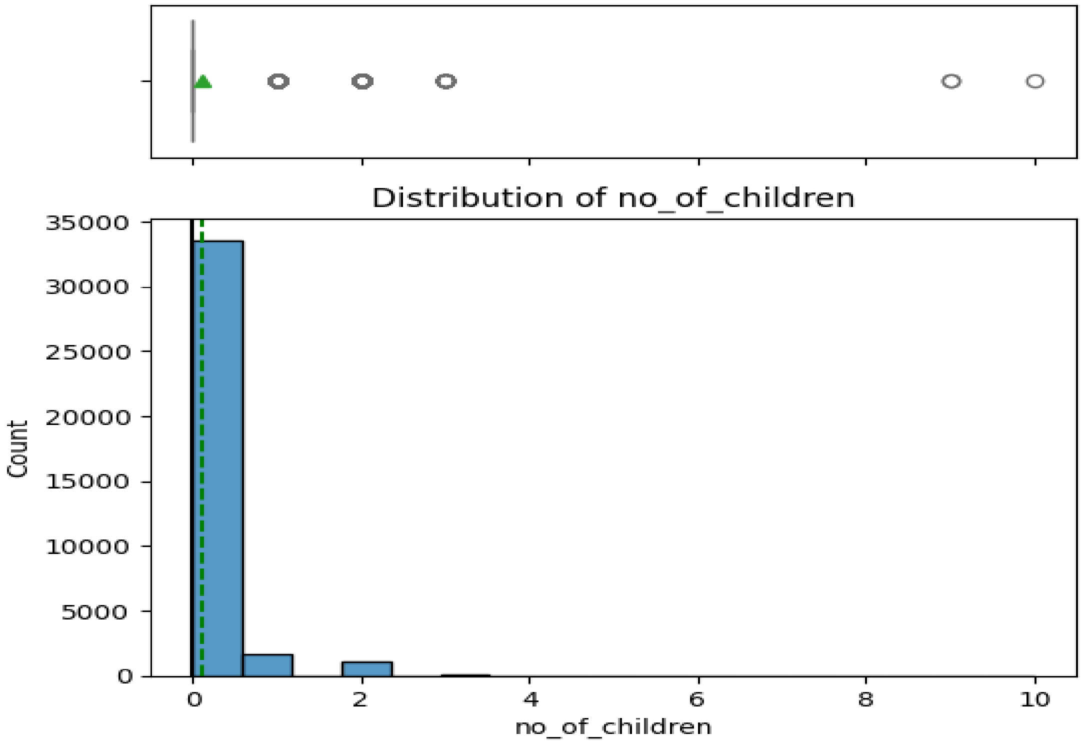

Figure 6: no_of_children

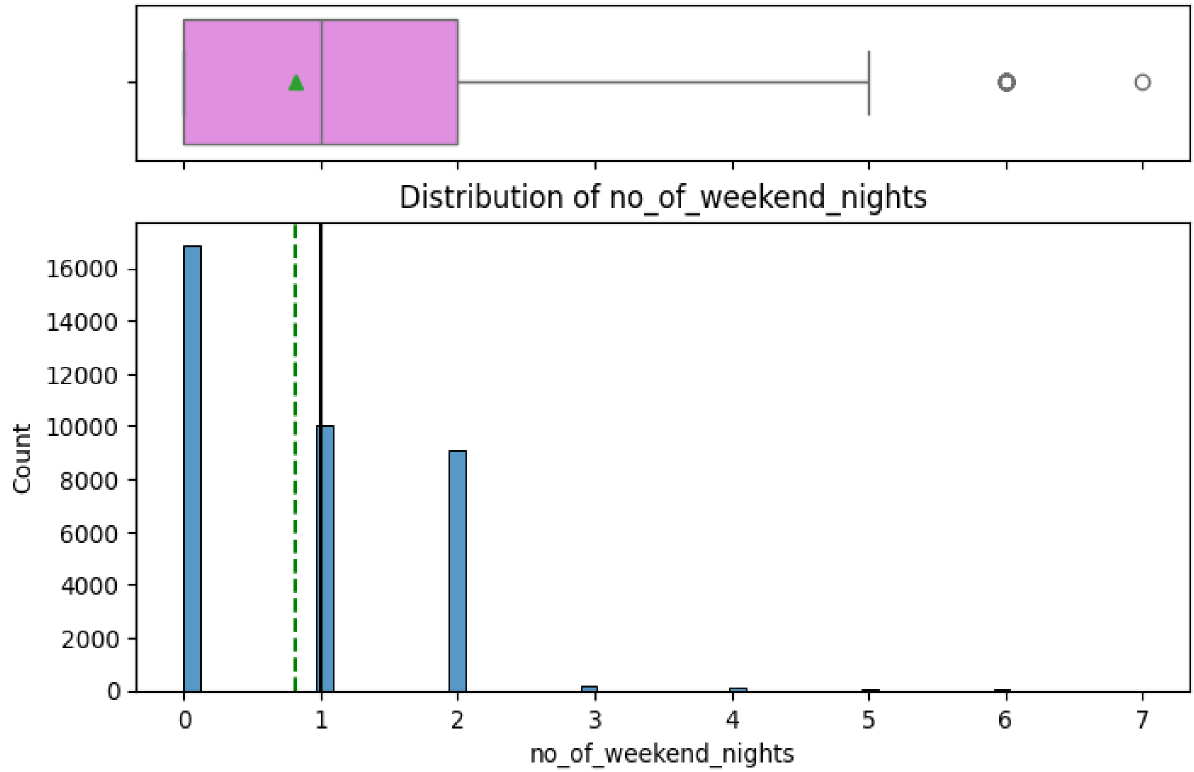

Figure 7: no_of_weekend_nights

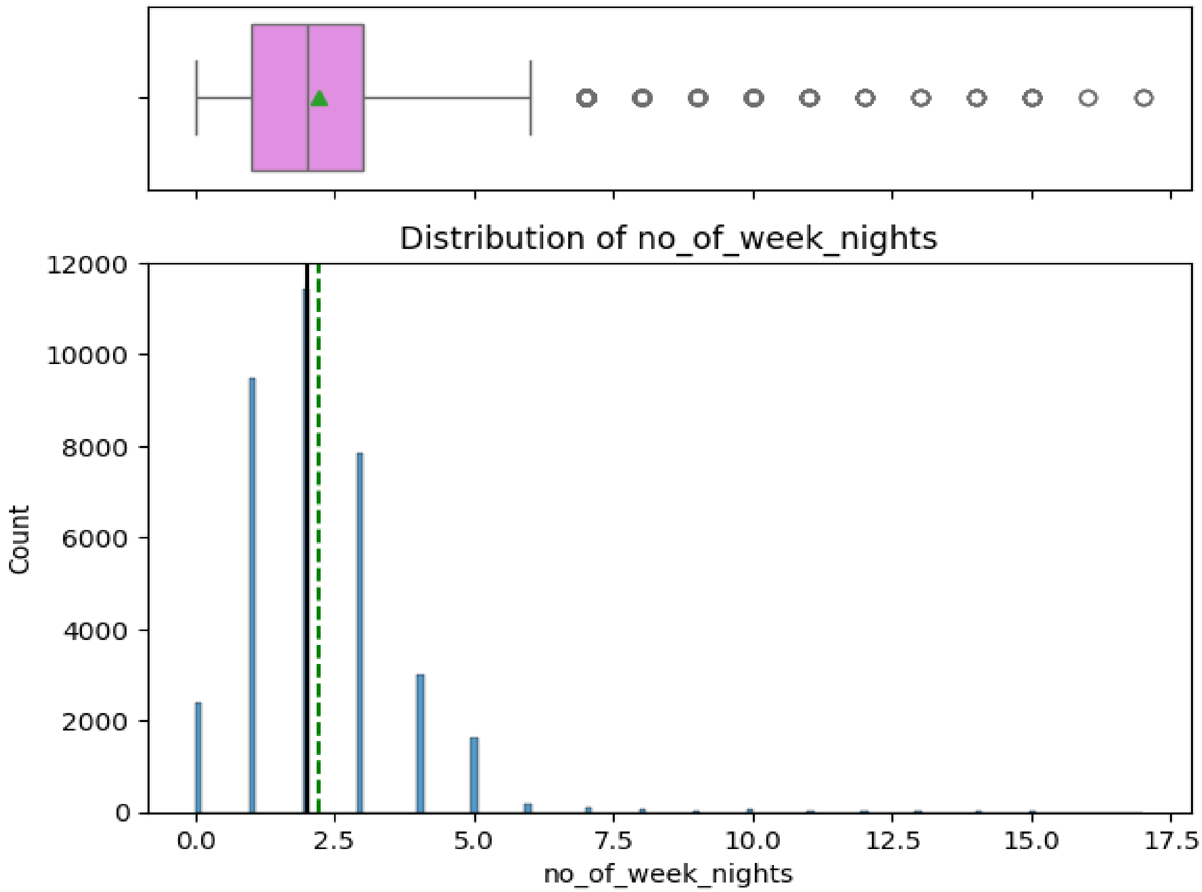

Figure 8: no_of_week_nights

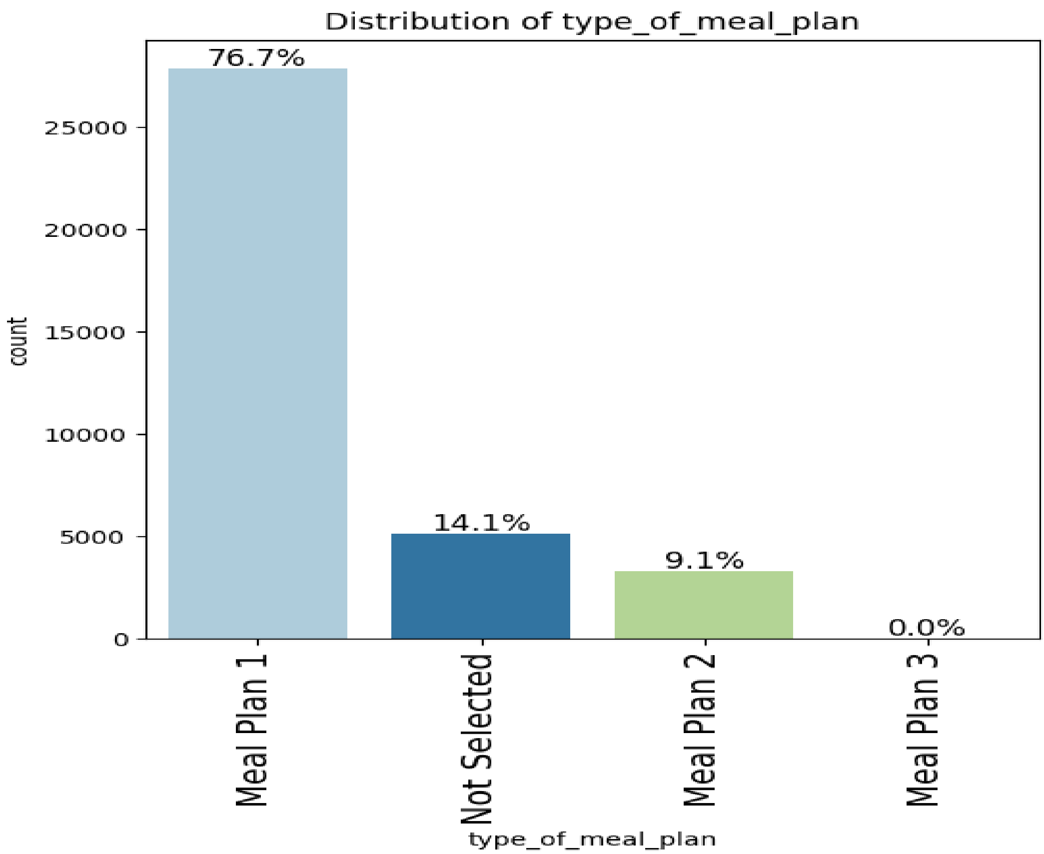

Figure 9: type_of_meal_plan

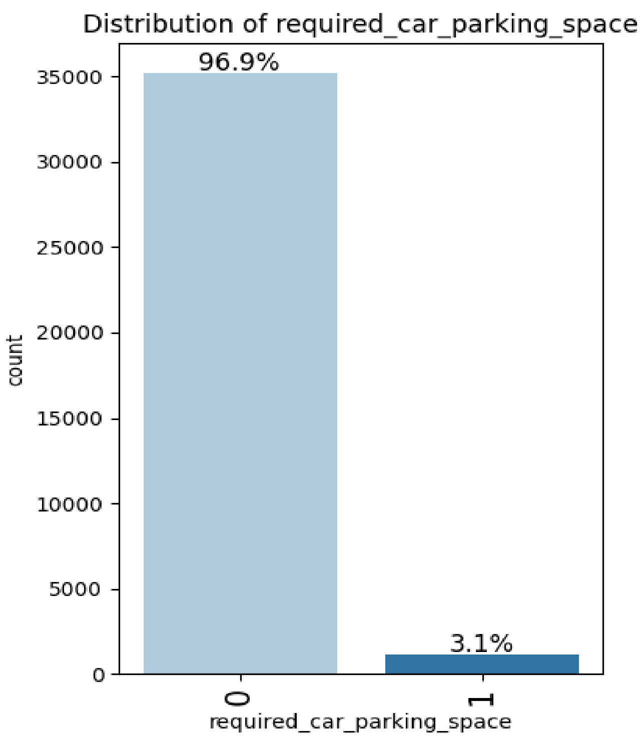

Figure 10: required_car_parking_space

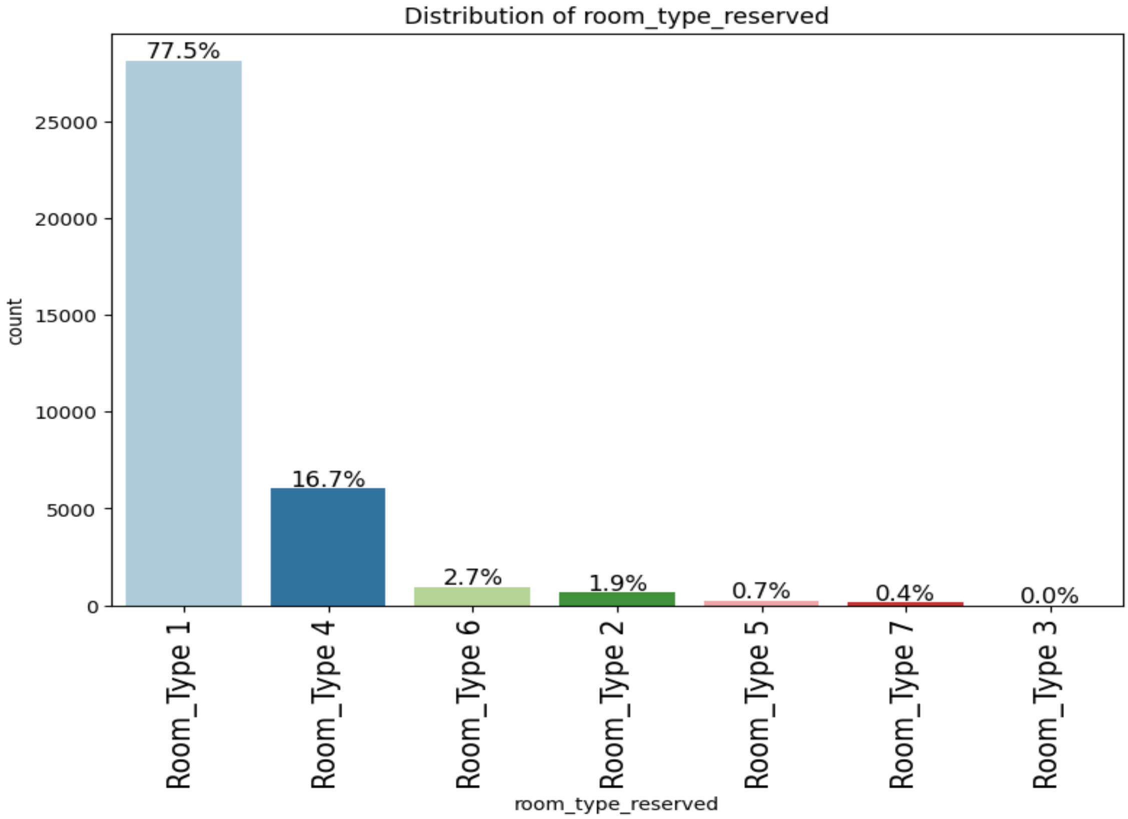

Figure 11: room_type_reserved

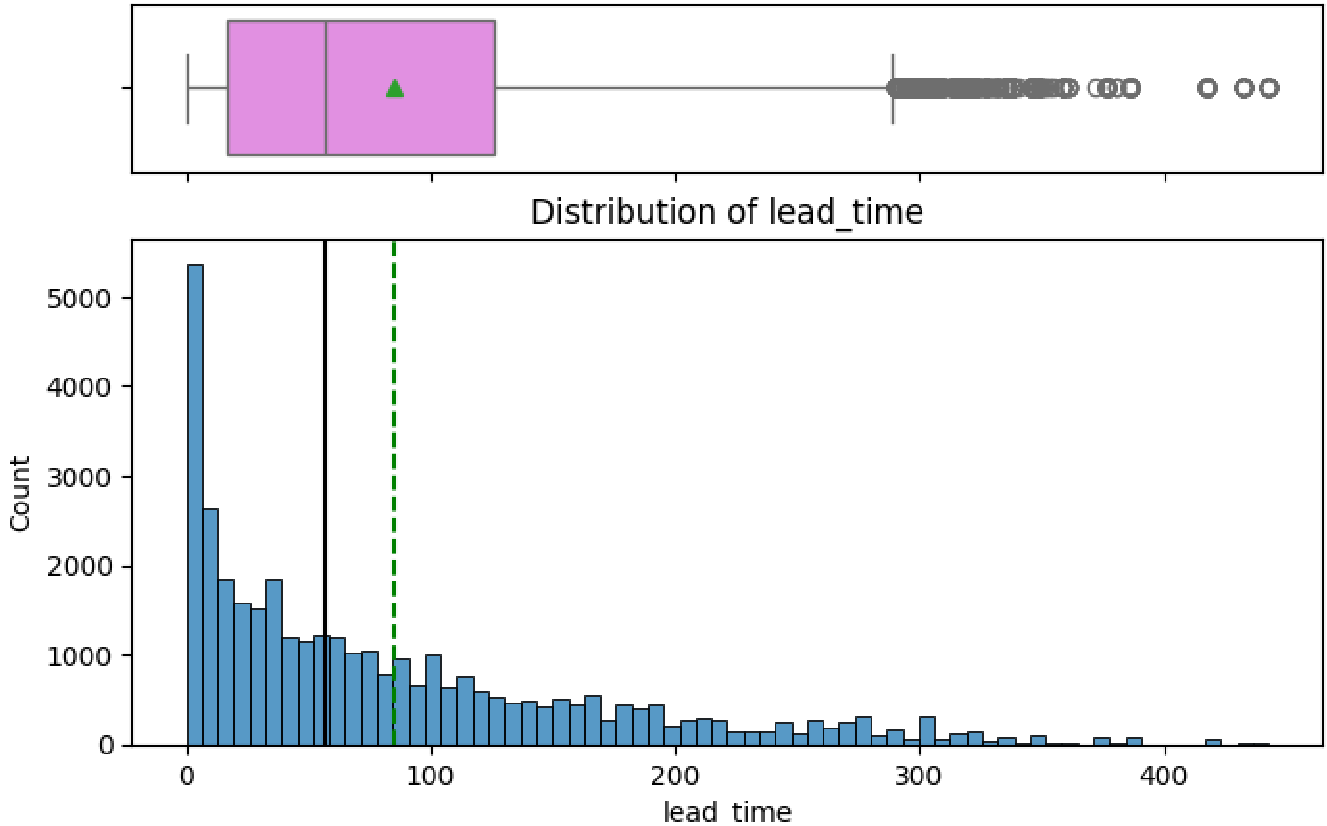

Figure 12: lead_time

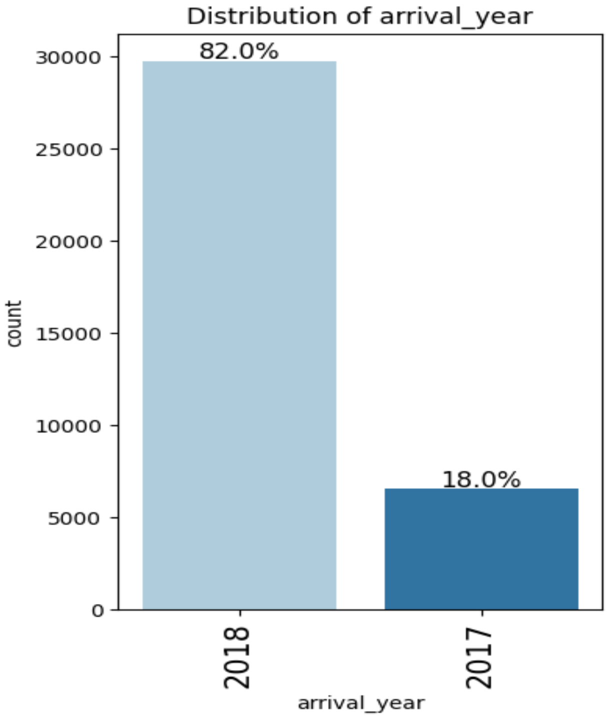

Figure 13: arrival_year

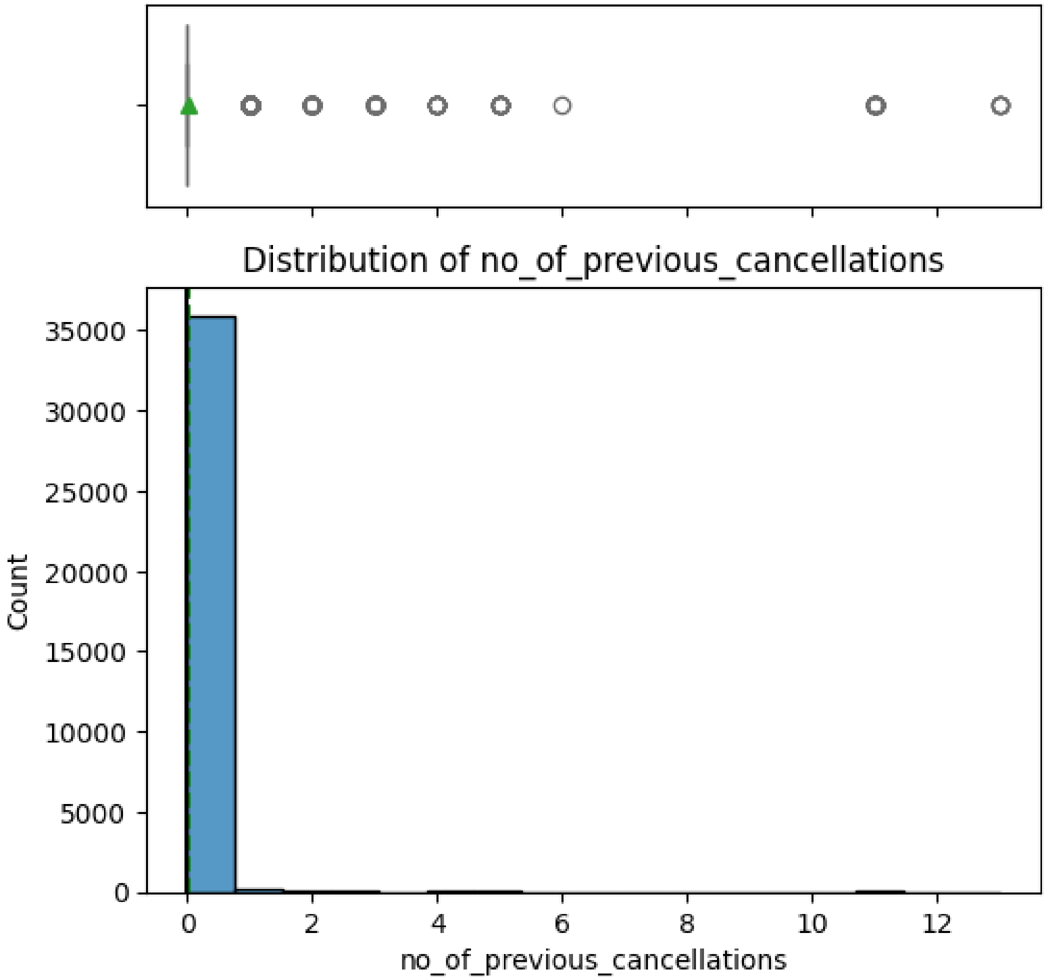

Figure 14: no_of_previous_cancellations

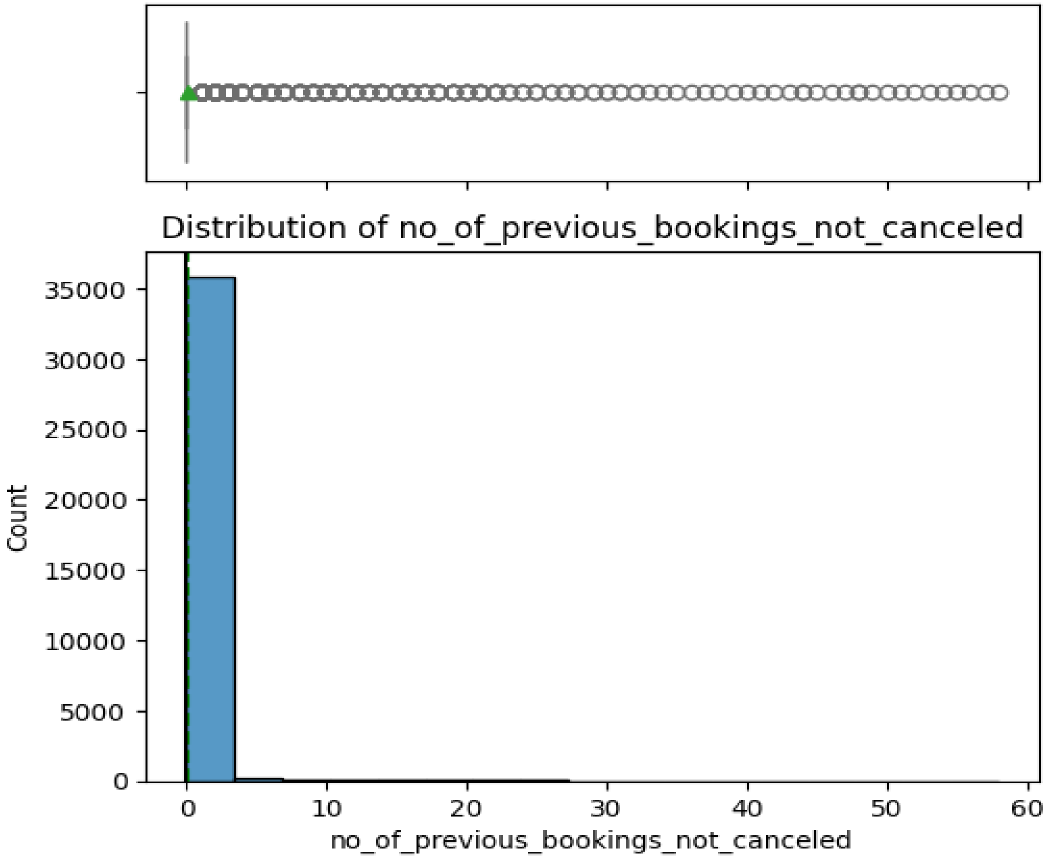

Figure 15: no_of_previous_bookings_ not_canceled

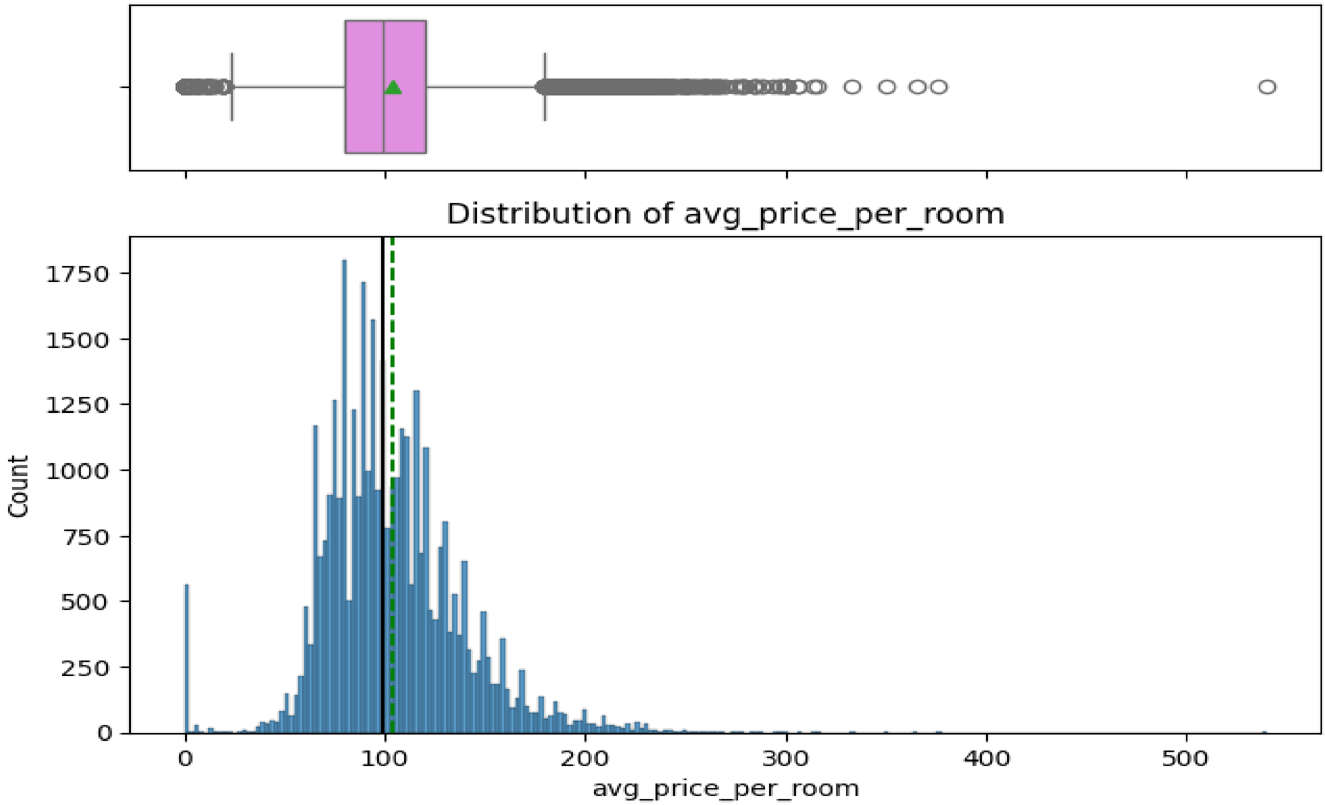

Figure 16: avg_price_per_room

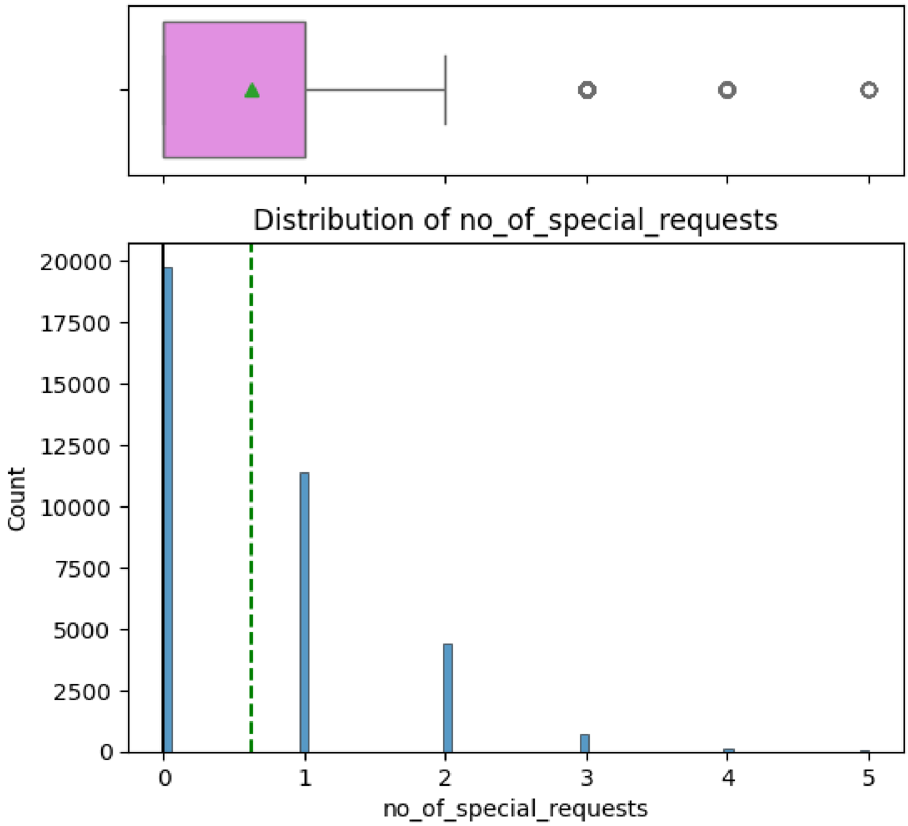

Figure 17: no_of_special_requests

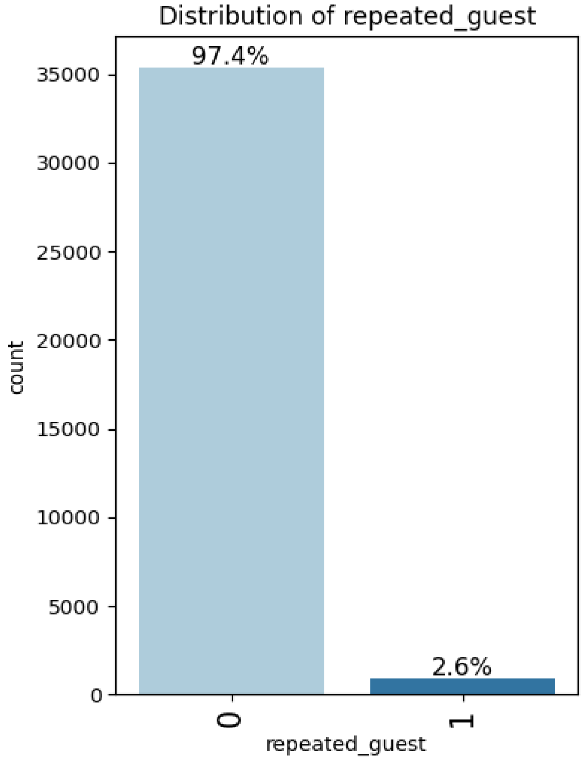

Figure 18: repeated_guest

5.2. Bivariate Analysis

Correlation Check

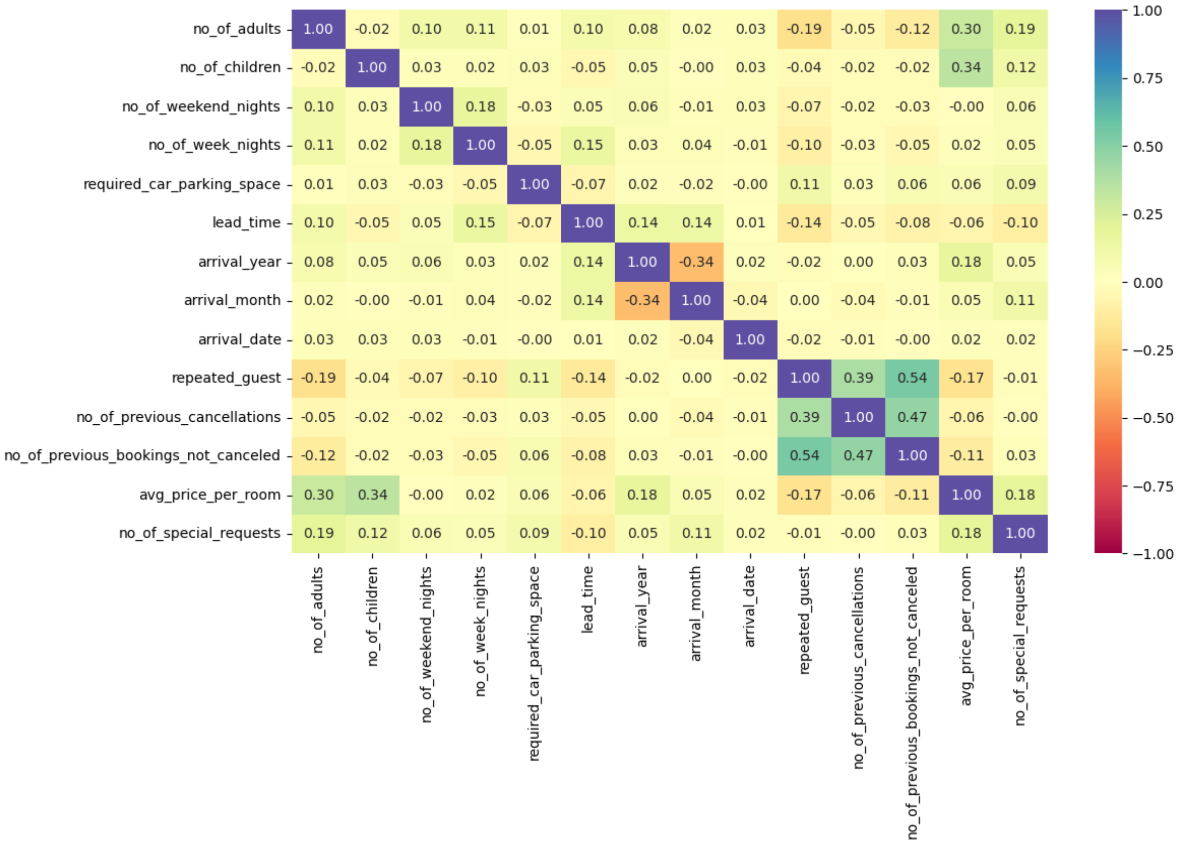

Figure 19: Heatmap

Insights

- Repeat Guest and No. of Previous Cancellations: A correlation of 0.39 suggests that guests who have previously canceled are moderately more likely to be repeat guests.
- Repeat Guest and No. of Previous Bookings Not Canceled: A correlation of 0.54 indicates a stronger likelihood of repeat bookings from guests who have not canceled before.
- Average Price Per Room and No. of Children: The correlation of 0.34 is quite notable, suggesting that rooms booked with children tend to be priced higher, possibly due to larger room types or additional amenities being booked.

Cancellations vs. Lead Time

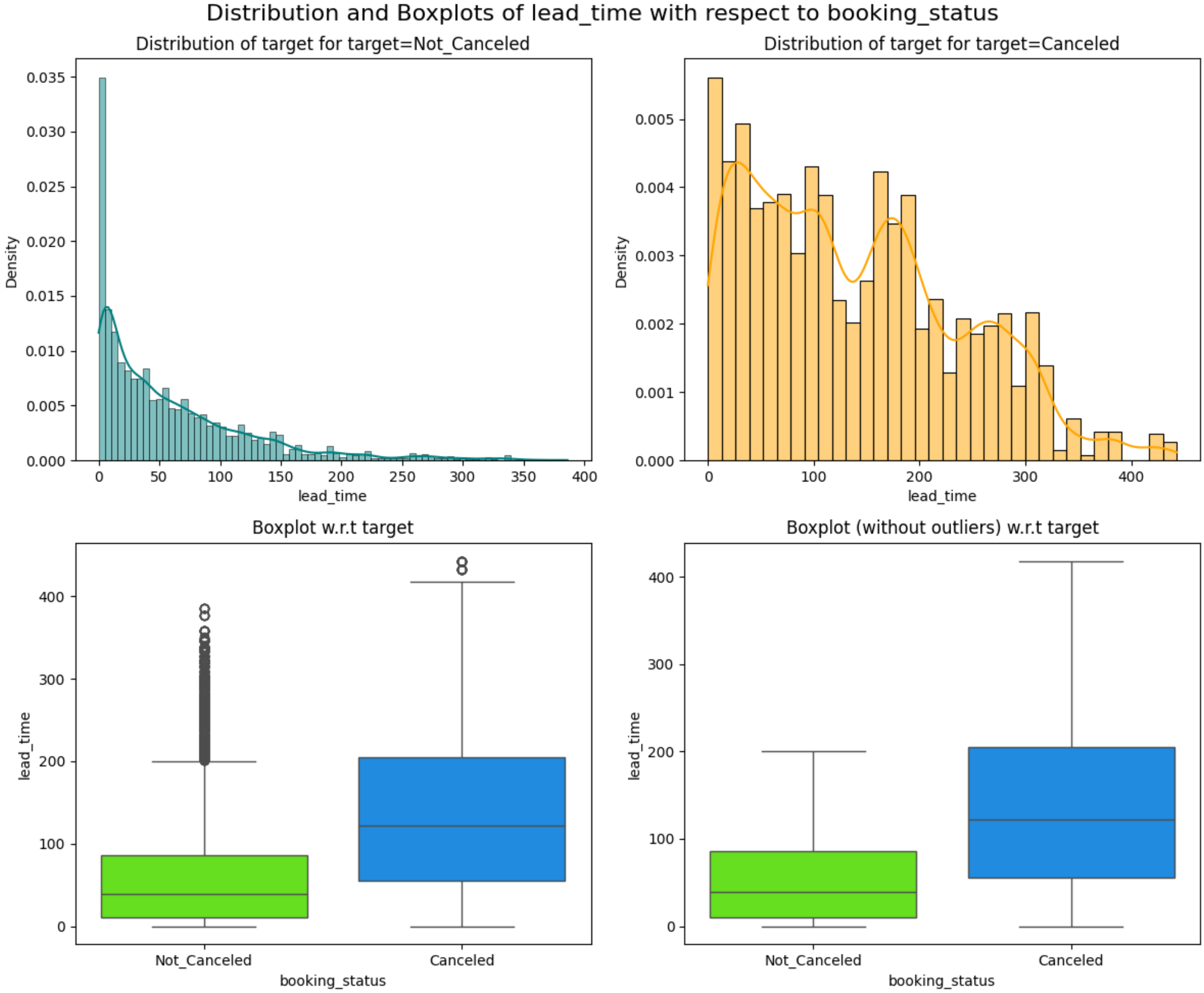

Figure 20: Cancellations vs. Lead Time

Room Type vs. Booking Status

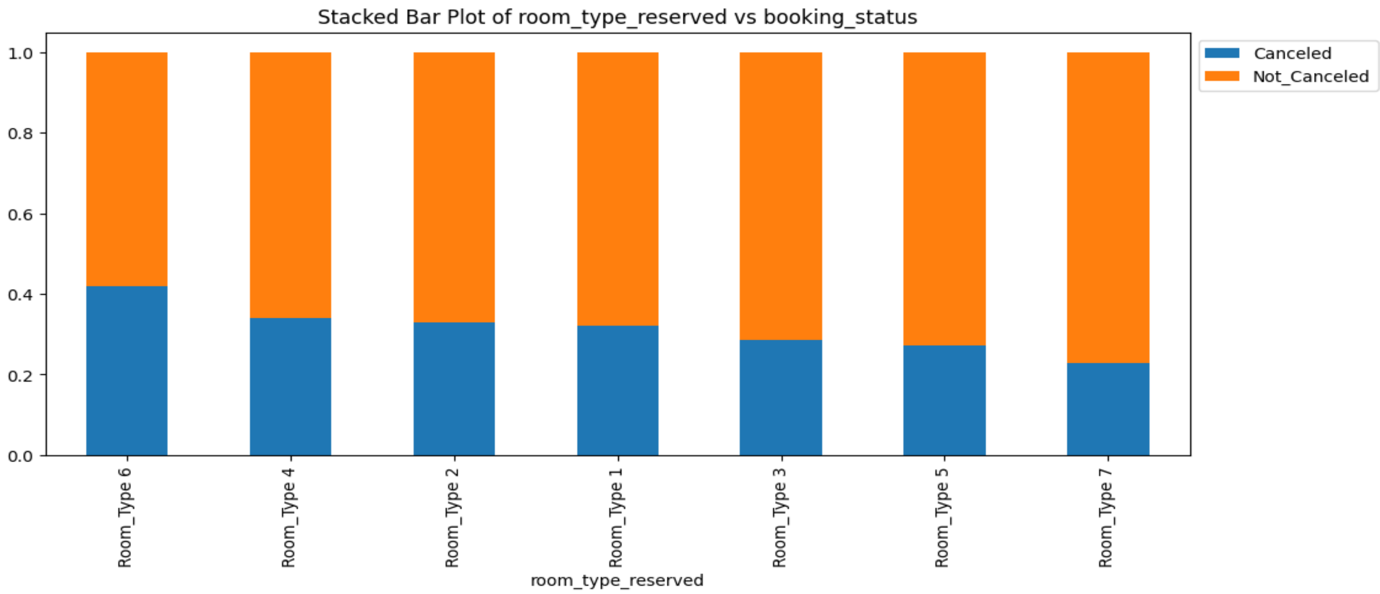

Figure 21: Room Type vs. Booking Status

Market Segment vs. Booking Trends

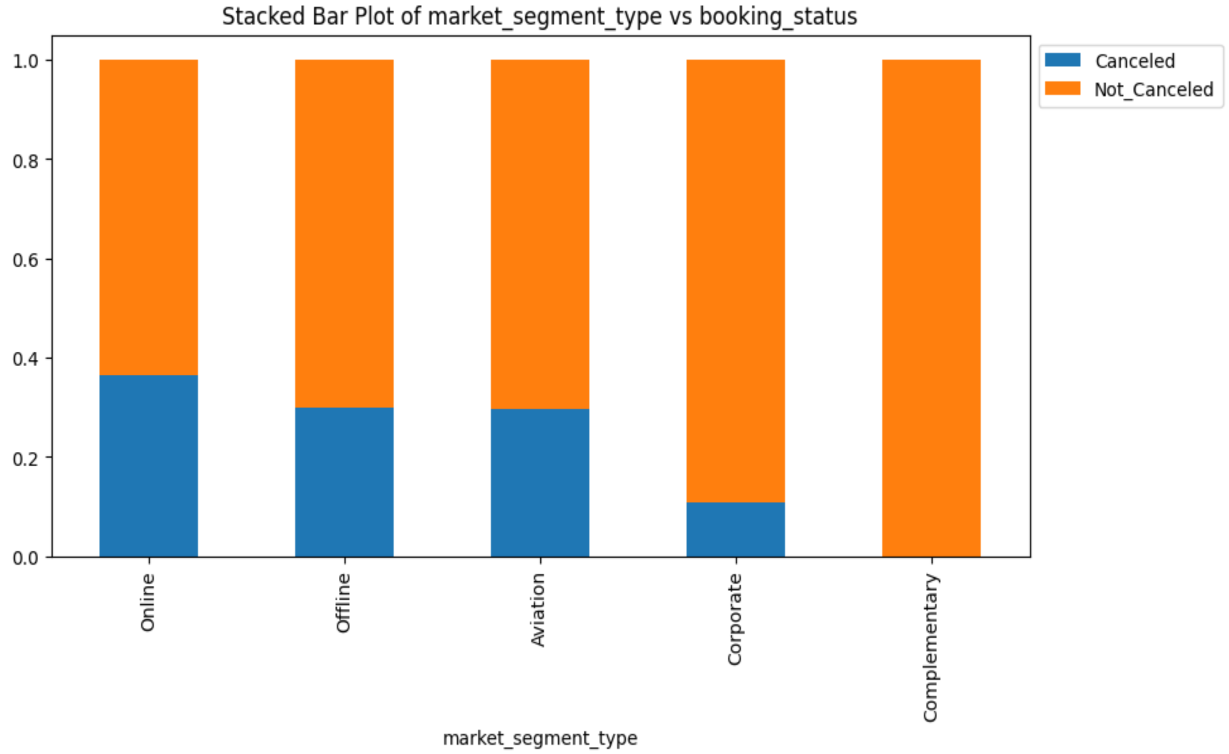

Figure 22: Market Segment vs. Booking Trends

Parking vs. Guest Type

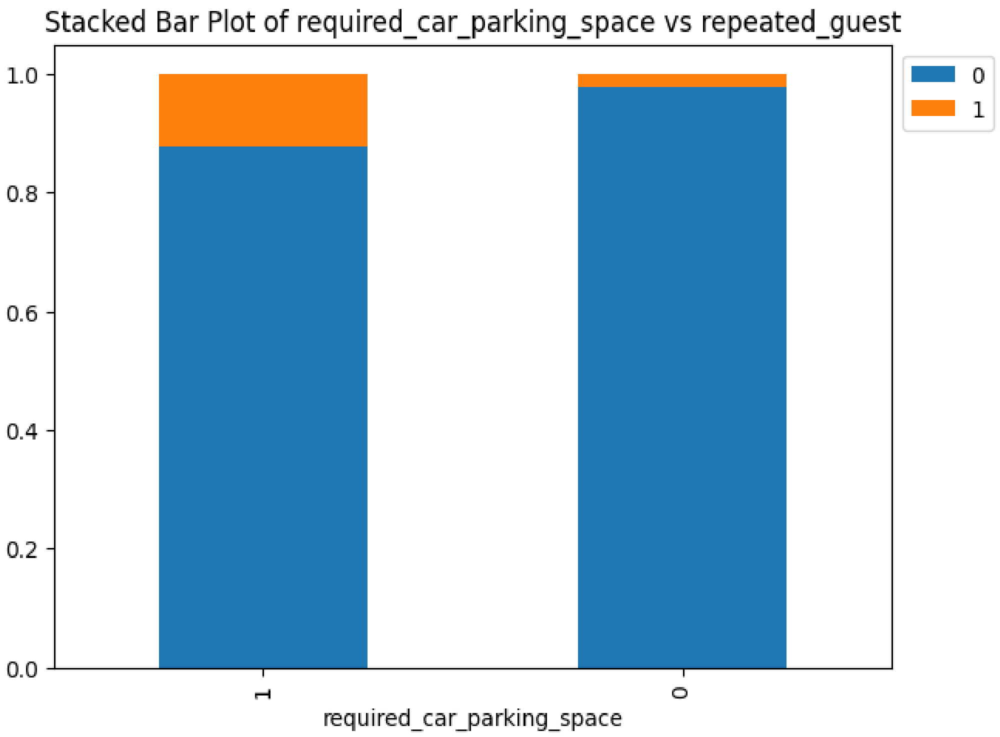

Figure 23: Parking vs. Guest Type

5.3. EDA Questions

Q1: What are the busiest months in the hotel?

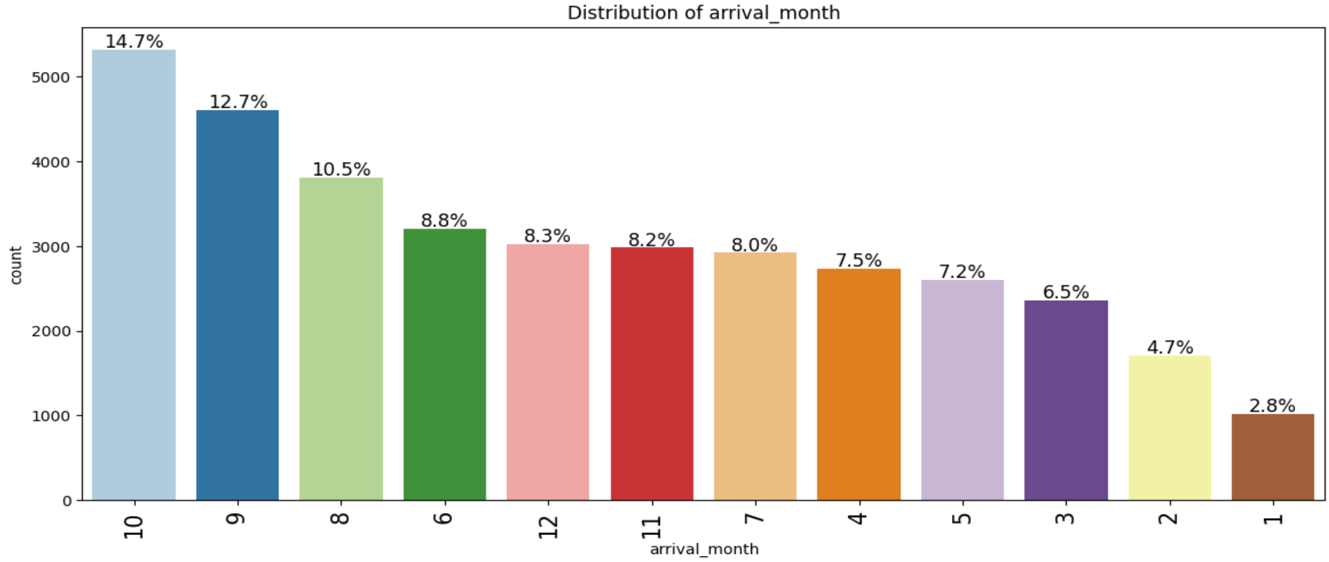

Figure 24: Arrival Month

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

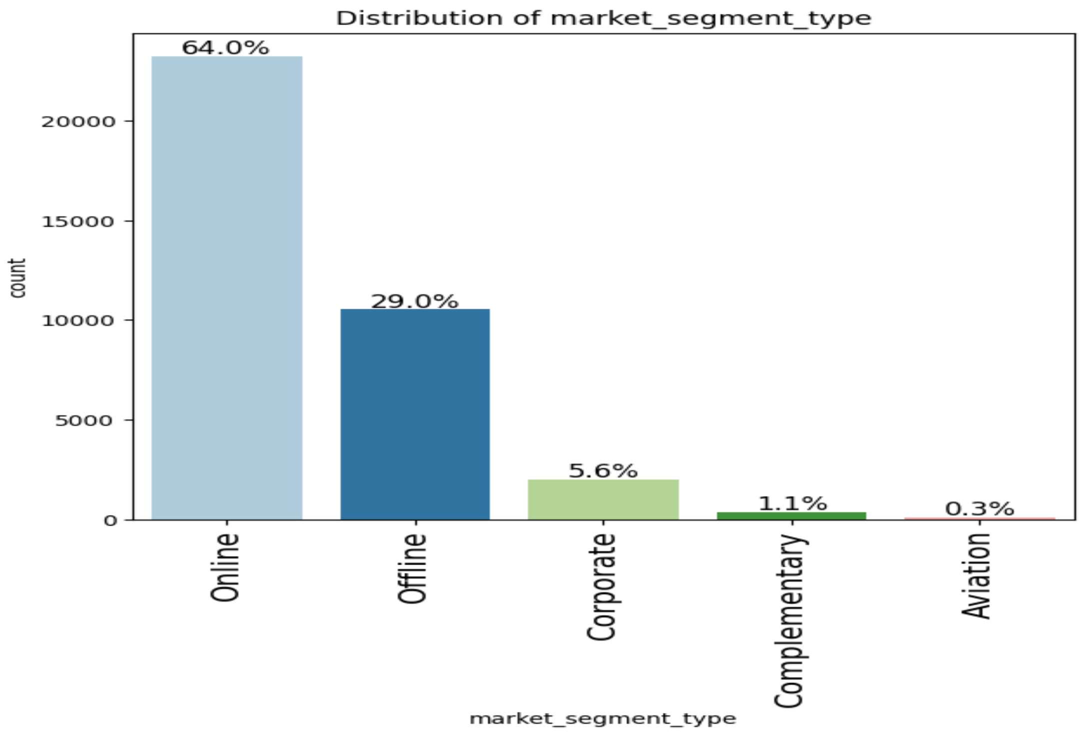

Figure 25: Market Segment

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

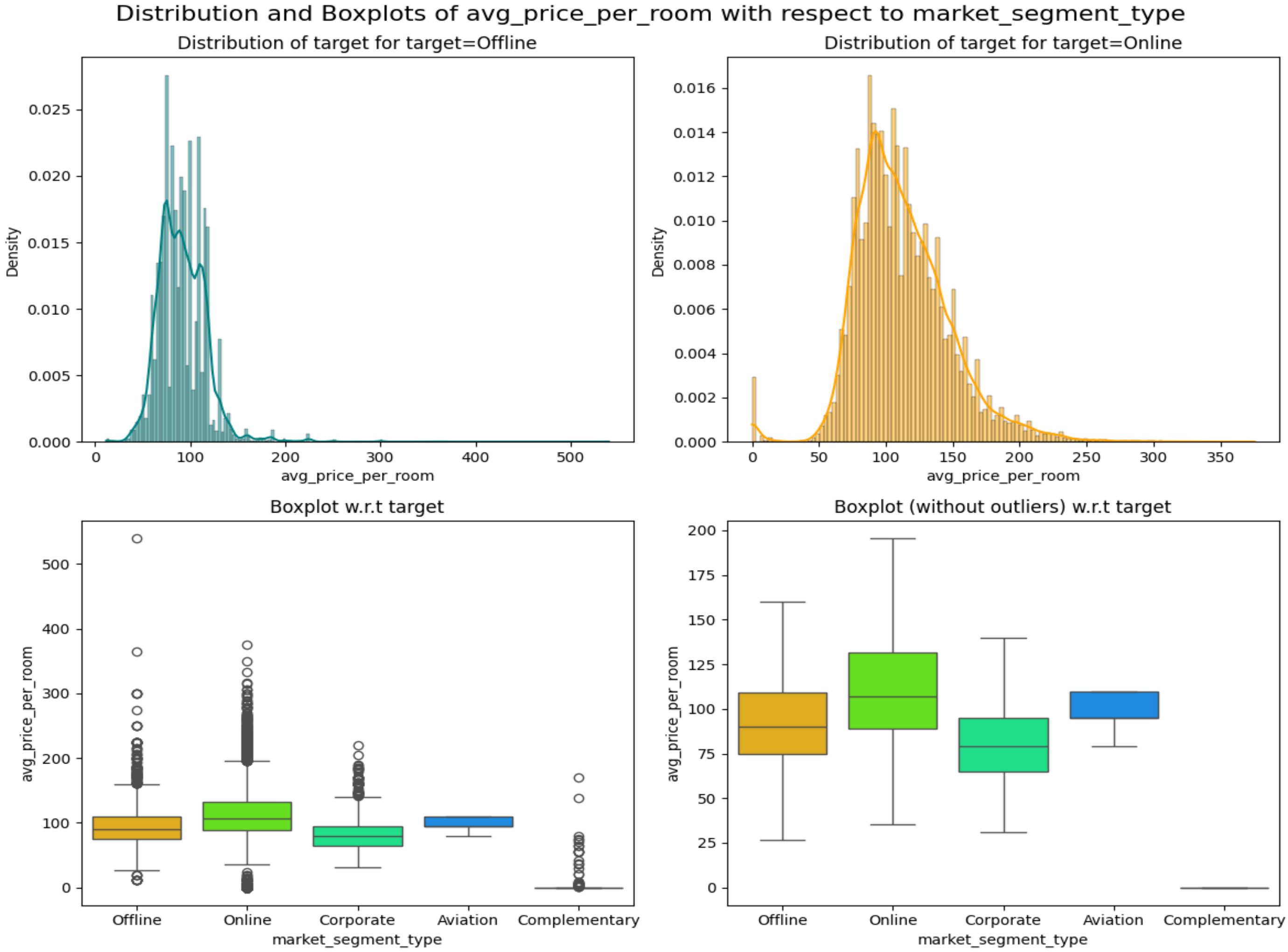

Figure 26: avg_price_per_room vs market_segment_type

Q4: What percentage of bookings are cancelled?

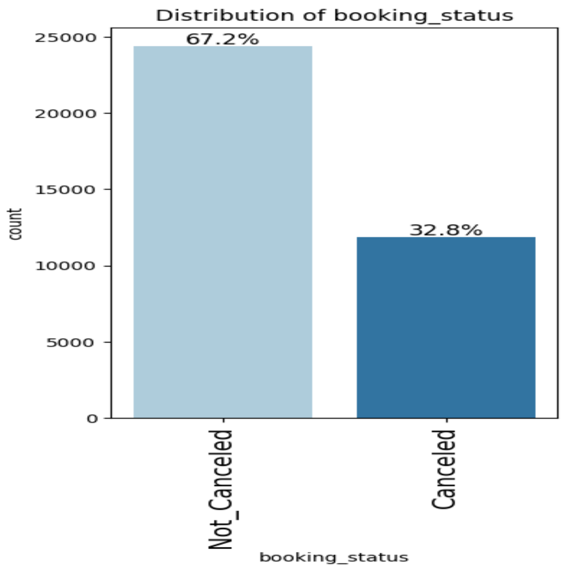

Figure 27: Booking Status

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

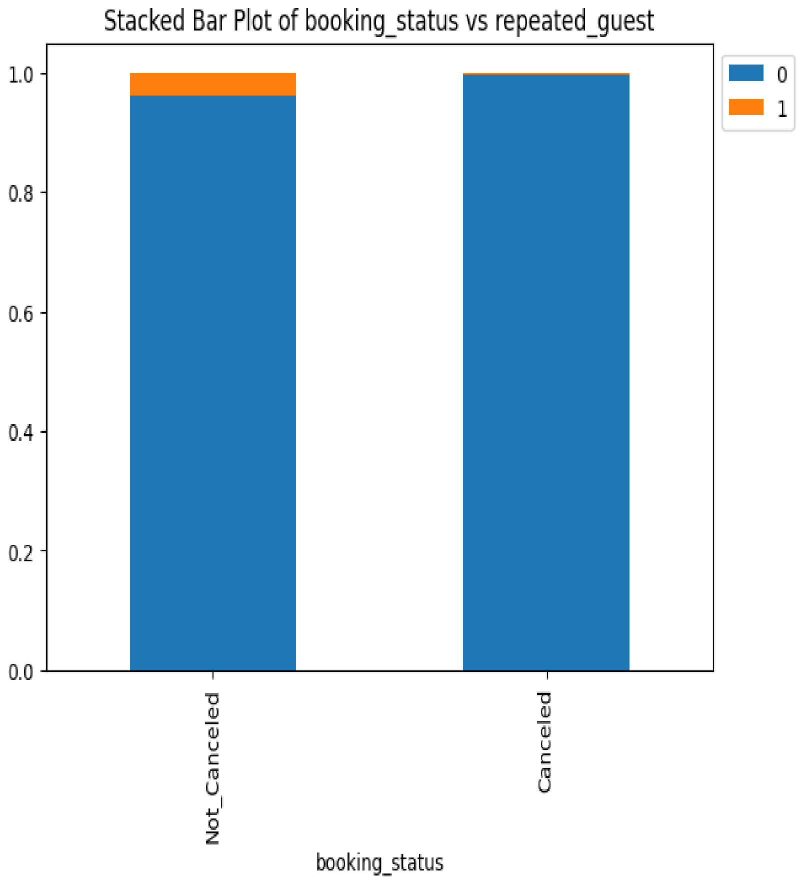

Figure 28: booking_status vs repeated_guest

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

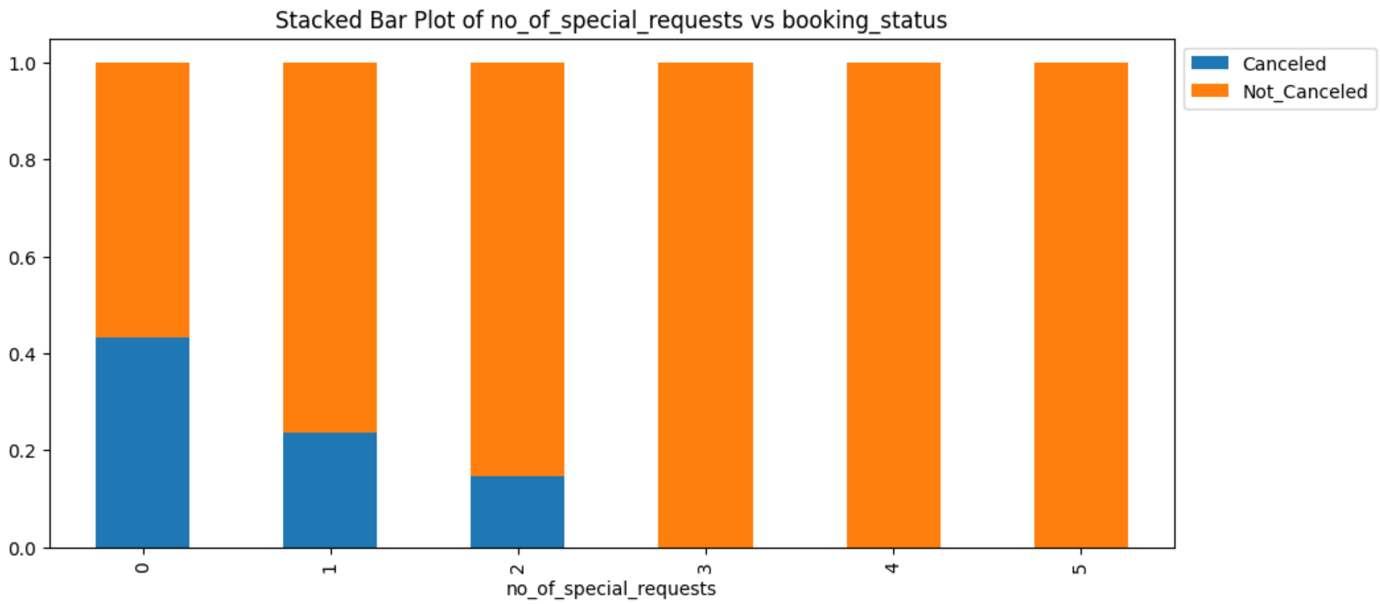

Figure 29: Special Requests vs. Cancellations

Insights

- Hotel bookings are high from August to October, with a peak in October.
- Customers prefer online bookings based on the distribution of market segment types.
- We observe that 32% of bookings are cancelled for various reasons.
- Around 28% of cancellations may be due to special requests.
- Prices are higher for rooms booked online, while offline and corporate bookings have similar room prices.
- Less than 1% of repeat guests cancel their bookings.

6. Data Preprocessing

6.1. Missing Value treatment

There are no missing values.

6.2. Duplicate value check

There are no duplicate rows.

6.3. Feature Engineering

Removing features from the dataset that have constant values and those that do not positively impact the prediction model.

Features removed: Booking ID, Arrival Year and Arrival date.

6.4. Outlier Detection

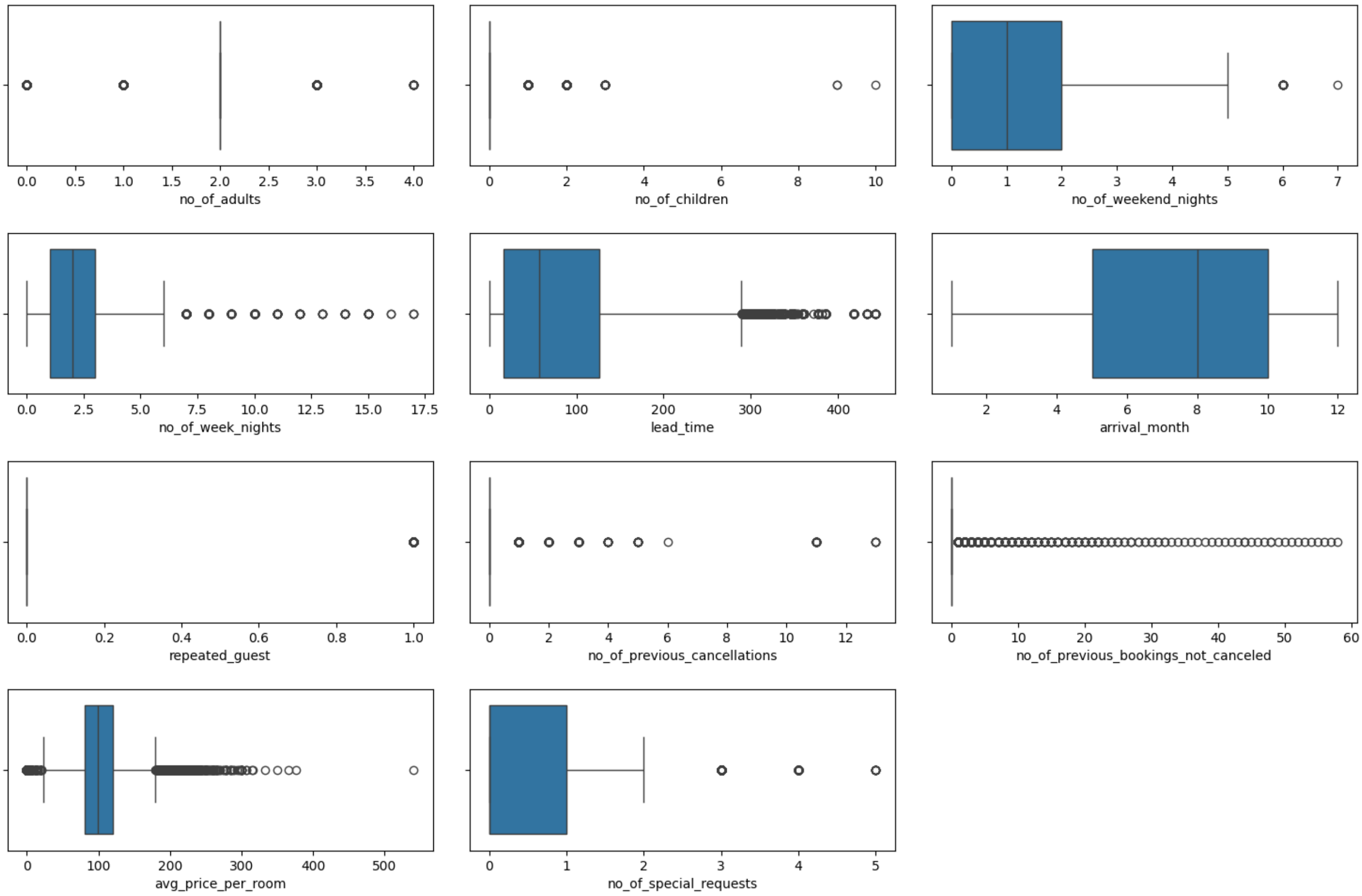

Figure 29: Outliers

There are outliers in the few columns like avg_price_per_room. We have a few options for handling these outliers:

- Use the IQR (Interquartile Range) to determine the lower and upper bounds of the column and either replace or remove the outliers.
- However, since we lack additional information from a subject matter expert, we may decide not to treat these outliers for now.
- The price varies with the seasons, so we can hold off on removing these outliers for the time being.

6.5. Data Preparation for Modeling

1. Our goal is to predict which bookings will be cancelled in advance, helping us develop profitable policies for cancellations and refunds.
2. Before building the model, we'll need to encode the categorical features.
3. We will split the data into training and testing sets to evaluate the model built on the training data.

6.5.1. Encoding Categorical Features

no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	required_car_parking_space	lead_time	arrival_month	repeated_guest	no_of_previous_cancellations	r
2.00000	0.00000	1.00000	3.00000	0.00000	200.00000	8.00000	0.00000	0.00000	
2.00000	0.00000	1.00000	3.00000	0.00000	79.00000	3.00000	0.00000	0.00000	
2.00000	0.00000	1.00000	4.00000	0.00000	78.00000	4.00000	0.00000	0.00000	
2.00000	0.00000	2.00000	0.00000	0.00000	61.00000	10.00000	0.00000	0.00000	
2.00000	1.00000	0.00000	4.00000	0.00000	201.00000	11.00000	0.00000	0.00000	
...	...	...	...	...	...	...	...	...	
2.00000	0.00000	2.00000	2.00000	0.00000	43.00000	12.00000	0.00000	0.00000	
1.00000	0.00000	0.00000	2.00000	0.00000	102.00000	10.00000	0.00000	0.00000	
2.00000	0.00000	0.00000	2.00000	0.00000	5.00000	5.00000	0.00000	0.00000	
2.00000	0.00000	0.00000	3.00000	0.00000	213.00000	6.00000	0.00000	0.00000	
2.00000	0.00000	0.00000	2.00000	0.00000	14.00000	9.00000	0.00000	0.00000	

Figure 30: Encoding

6.5.2. Train – Test Split

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

7. Model building

7.1. Logistic Regression

Logit Regression Results

Dep. Variable:	booking_status_Not_Canceled	No. Observations:	25392
Model:	Logit	Df Residuals:	25366
Method:	MLE	Df Model:	25
Date:	Fri, 09 Aug 2024	Pseudo R-squ.:	0.3258
Time:	19:32:02	Log-Likelihood:	-10808.
converged:	False	LL-Null:	-16030.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	2.7511	0.272	10.132	0.000	2.219	3.283
no_of_adults	-0.1119	0.038	-2.978	0.003	-0.186	-0.038
no_of_children	-0.1152	0.059	-1.951	0.051	-0.231	0.001
no_of_weekend_nights	-0.1229	0.020	-6.208	0.000	-0.162	-0.084
no_of_week_nights	-0.0333	0.012	-2.703	0.007	-0.057	-0.009
required_car_parking_space	1.7183	0.142	12.082	0.000	1.440	1.997
lead_time	-0.0160	0.000	-62.435	0.000	-0.016	-0.015
arrival_month	0.0613	0.006	10.193	0.000	0.050	0.073
repeated_guest	2.5769	0.656	3.930	0.000	1.292	3.862
no_of_previous_cancellations	-0.2653	0.078	-3.409	0.001	-0.418	-0.113
no_of_previous_bookings_not_canceled	0.0662	0.087	0.761	0.446	-0.104	0.237
avg_price_per_room	-0.0199	0.001	-27.746	0.000	-0.021	-0.018
no_of_special_requests	1.4736	0.030	48.973	0.000	1.415	1.533
type_of_meal_plan_Meal Plan 2	-0.0653	0.064	-1.024	0.306	-0.190	0.060
type_of_meal_plan_Meal Plan 3	-28.4637	1.27e+06	-2.24e-05	1.000	-2.49e+06	2.49e+06
type_of_meal_plan_Not Selected	-0.2533	0.052	-4.826	0.000	-0.356	-0.150
room_type_reserved_Room_Type 2	0.4408	0.130	3.389	0.001	0.186	0.696
room_type_reserved_Room_Type 3	-1.1876	2.013	-0.590	0.555	-5.133	2.758
room_type_reserved_Room_Type 4	0.2623	0.053	4.953	0.000	0.159	0.366
room_type_reserved_Room_Type 5	0.6550	0.208	3.152	0.002	0.248	1.062
room_type_reserved_Room_Type 6	1.0005	0.150	6.684	0.000	0.707	1.294
room_type_reserved_Room_Type 7	1.2899	0.305	4.227	0.000	0.692	1.888
market_segment_type_Complementary	42.1259	1.27e+06	3.32e-05	1.000	-2.49e+06	2.49e+06
market_segment_type_Corporate	0.9436	0.274	3.438	0.001	0.406	1.481
market_segment_type_Offline	1.9494	0.263	7.403	0.000	1.433	2.466
market_segment_type_Online	0.1275	0.260	0.490	0.624	-0.383	0.638

Figure 31: Model Statistics

Interpreting the Regression Results:

- Intercept: The constant (intercept) of 2.7511 suggests that when all other variables are held constant, there is a positive baseline log-odds of a booking not being canceled.
- Number of Adults and Children: For every additional adult, the log-odds of a booking not being canceled decrease by 0.1119, indicating a negative association. Similarly, for each additional child, the log-odds of a booking not being canceled decrease by 0.1152, with the result being marginally significant ($p = 0.051$).
- Number of Nights: An increase in the number of weekend nights booked decreases the log-odds of a booking not being canceled by 0.1229, suggesting that bookings with more weekend nights are more likely to be canceled. Likewise, each additional weeknight decreases the log-odds of a booking not being canceled by 0.0333.
- Required Car Parking Space: Bookings with required car parking space have significantly higher log-odds (1.7183) of not being canceled, indicating that this feature strongly predicts bookings that are likely to be fulfilled.
- Lead Time: A longer lead time significantly reduces the log-odds of a booking not being canceled by 0.0160 per day, implying that bookings made well in advance are more likely to be canceled.
- Arrival Month: The positive coefficient of 0.0613 for the arrival month suggests a seasonal effect, where bookings made closer to certain months (e.g., peak travel season) have higher chances of being fulfilled.
- Repeated Guests: Being a repeated guest increases the log-odds of a booking not being canceled by 2.5769, indicating that repeat customers are much more likely to follow through with their bookings.
- Previous Cancellations: A history of previous cancellations decreases the log-odds of a booking not being canceled by 0.2653, reflecting a negative impact on the likelihood of fulfilling the current booking.

- Average Price Per Room: An increase in the average price per room slightly decreases the log-odds of a booking not being canceled by 0.0199 per unit increase in price, suggesting that higher-priced rooms might be associated with a higher likelihood of cancellations.
- Number of Special Requests: The number of special requests significantly increases the log-odds of a booking not being canceled by 1.4736, suggesting that guests making special requests are more committed to their bookings.
- Type of Meal Plan: Meal Plan 3 has a very high negative coefficient, indicating an estimation issue possibly due to perfect separation, as the p-value suggests no significance. Not selecting a meal plan decreases the log-odds of a booking not being canceled by 0.2533.
- Room Type Reserved: Certain room types (e.g., Room Type 6 and 7) are associated with significantly higher log-odds of a booking not being canceled, indicating a preference or higher commitment to specific room types.
- Market Segment Type: The market_segment_type_Complementary has an unusually large coefficient and p-value suggesting estimation issues. Bookings made through the Corporate and Offline segments are significantly more likely to not be canceled, with coefficients of 0.9436 and 1.9494, respectively.

7.1.1. Model Performance

Train Data

Accuracy	Recall	Precision	F1
0.80289	0.88910	0.83043	0.85876

Figure 32: Model Performance

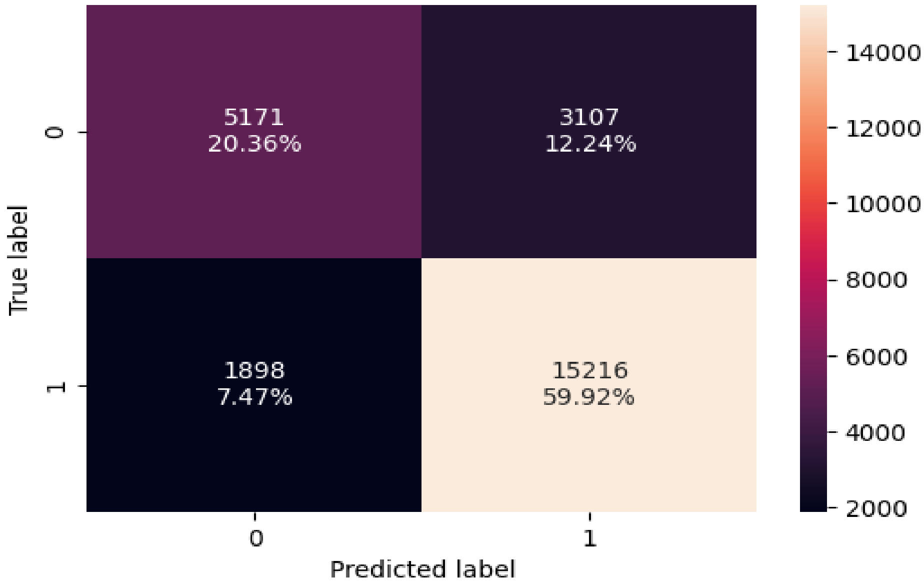

Figure 33: Confusion Matrix

Test Data

Accuracy	Recall	Precision	F1
0.80474	0.89720	0.82581	0.86002

Figure 34: Model Performance

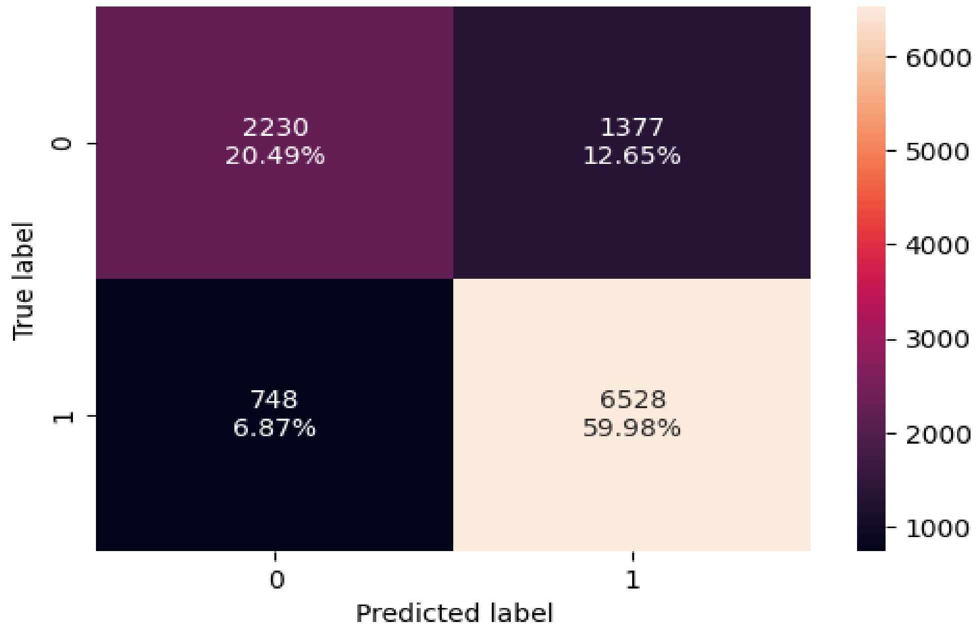

Figure 35: Confusion Matrix

7.2. KNN Classifier

In this study, a k-Nearest Neighbors (k-NN) algorithm was implemented with k = 3 to classify bookings.

7.2.1. Model Performance

Train Data

Accuracy	Recall	Precision	F1
0.91671	0.94992	0.92818	0.93892

Figure 36: Model Performance

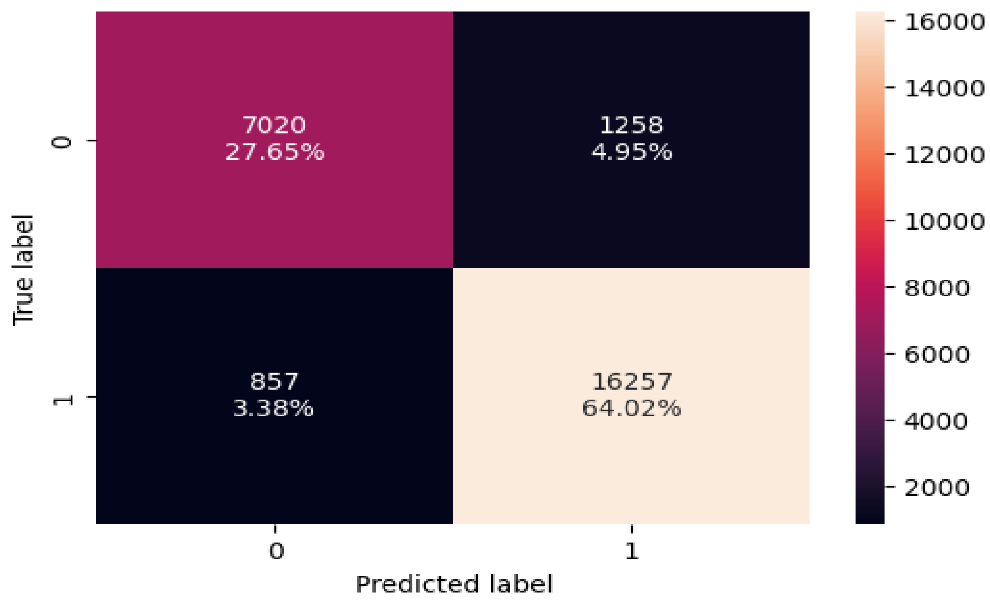

Figure 37: Confusion Matrix

Test Data

Accuracy	Recall	Precision	F1
0.85234	0.90104	0.88083	0.89082

Figure 38: Model Performance

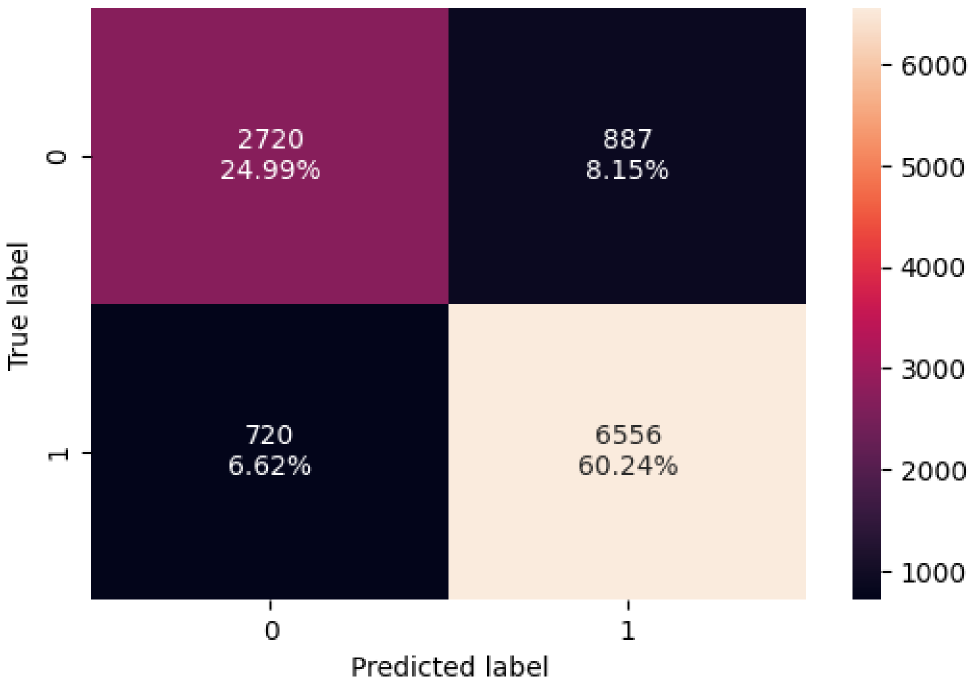

Figure 39: Confusion Matrix

7.3. Naive- Bayes Classifier

7.3.1. Model Performance

Train Data

Accuracy	Recall	Precision	F1
0.40855	0.14029	0.88729	0.24228

Figure 40: Model Performance

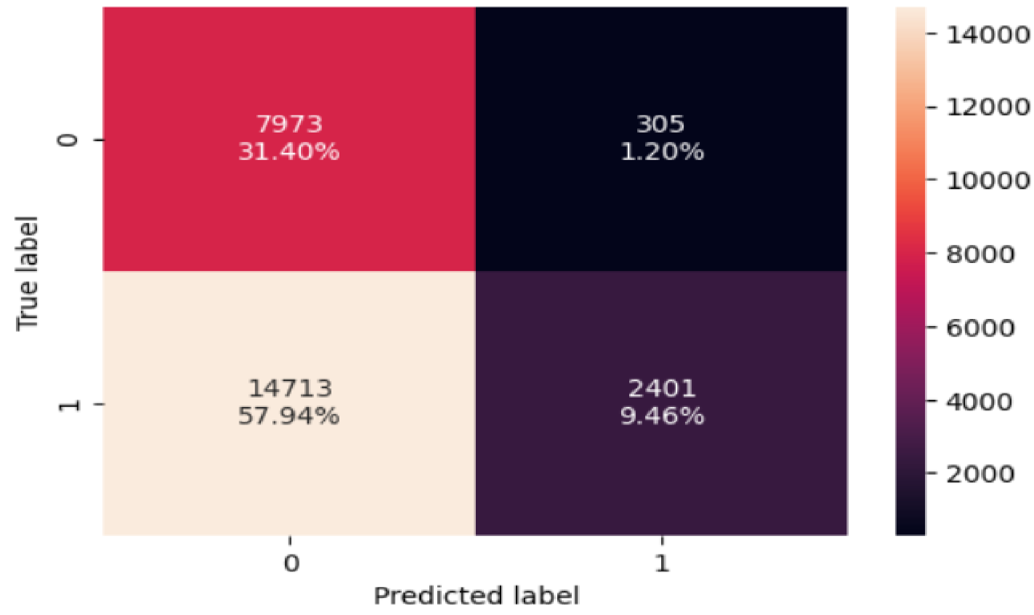

Figure 41: Confusion Matrix

Test Data

Accuracy	Recall	Precision	F1
0.41459	0.14184	0.89042	0.24469

Figure 42: Model Performance

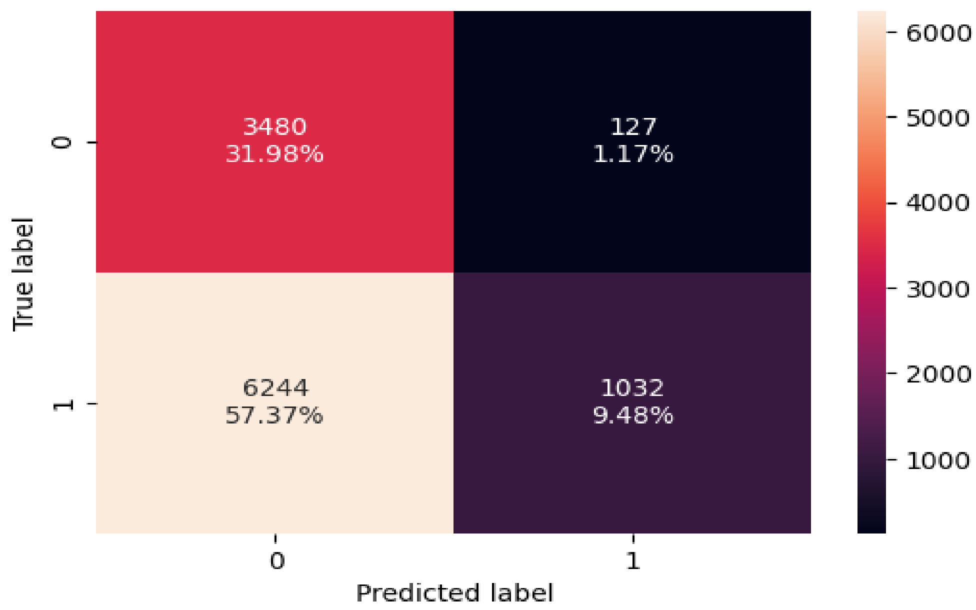

Figure 43: Confusion Matrix

7.4. Decision Tree Classifier

7.4.1. Model Performance

Train Data

Accuracy	Recall	Precision	F1
0.99374	0.99509	0.99562	0.99535

Figure 44: Model Performance

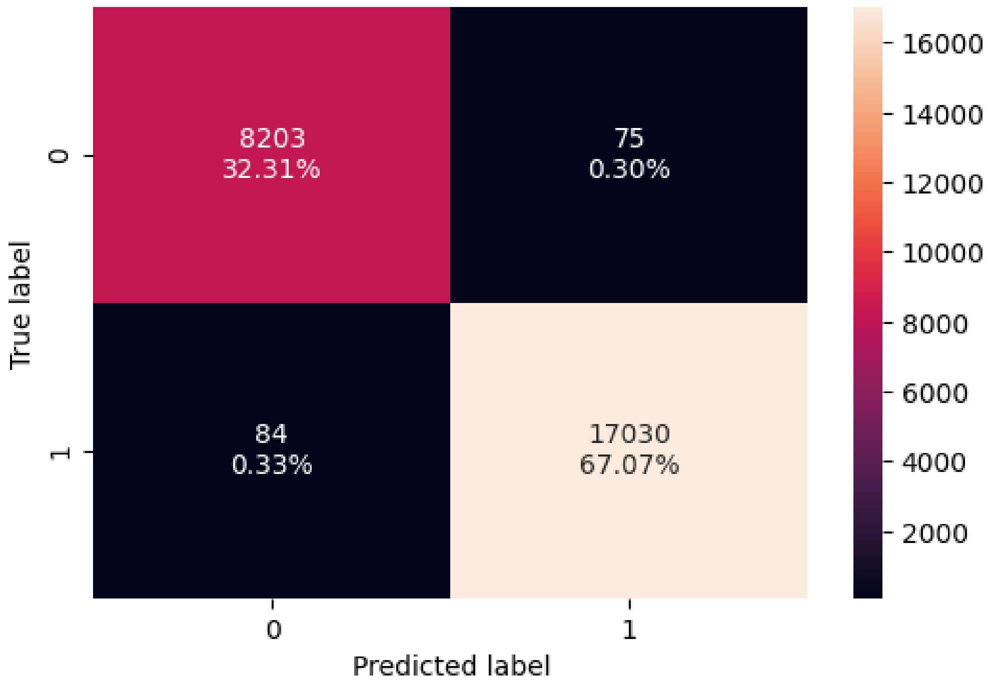

Figure 45: Confusion Matrix

Test Data

Accuracy	Recall	Precision	F1
0.86364	0.89321	0.90189	0.89753

Figure 46: Model Performance

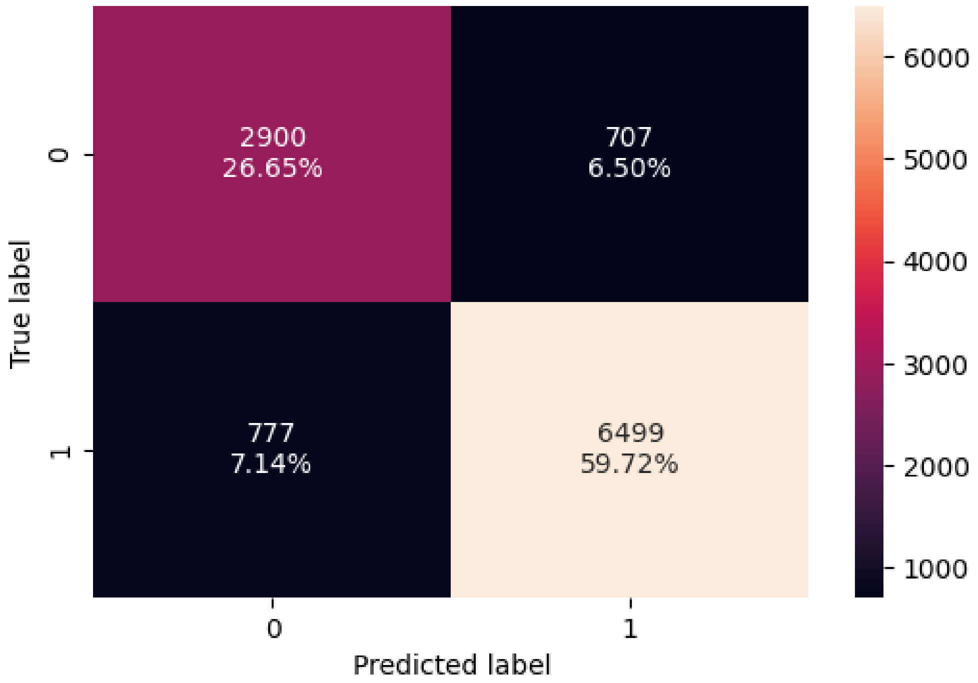

Figure 47: Confusion Matrix

8. Model Performance Improvement

8.1. Logistic Regression

8.1.1. Dealing with Multicollinearity

VIF is used to measure how much the variance of an estimated regression coefficient increases when your predictors are correlated.

Here's a quick overview of what the VIF values indicate:

- VIF = 1: No correlation between the variable and other variables.
- 1 < VIF < 5: Moderate correlation; generally considered acceptable.
- VIF >= 5: High correlation; may indicate problematic multicollinearity.
- VIF > 10: Very high correlation; suggests significant multicollinearity issues.

```
Variance Inflation Factors:
                                   Variable        VIF
0                                     const  326.53376
1                              no_of_adults    1.34103
2                            no_of_children    2.00531
3                      no_of_weekend_nights    1.06289
4                         no_of_week_nights    1.09128
5                required_car_parking_space    1.03706
6                                 lead_time    1.24328
7                             arrival_month    1.04936
8                            repeated_guest    1.76556
9              no_of_previous_cancellations    1.36938
10     no_of_previous_bookings_not_canceled    1.61311
11                         avg_price_per_room    1.92811
12                     no_of_special_requests    1.24369
13             type_of_meal_plan_Meal Plan 2    1.19892
14             type_of_meal_plan_Meal Plan 3    1.00600
15            type_of_meal_plan_Not Selected    1.23946
16            room_type_reserved_Room_Type 2    1.09085
17            room_type_reserved_Room_Type 3    1.00484
18            room_type_reserved_Room_Type 4    1.35215
19            room_type_reserved_Room_Type 5    1.03285
20            room_type_reserved_Room_Type 6    1.97980
21            room_type_reserved_Room_Type 7    1.10427
22         market_segment_type_Complementary    4.49241
23             market_segment_type_Corporate   17.19096
24               market_segment_type_Offline   64.26317
25                market_segment_type_Online   71.31989
```

Figure 48: VIF

Removing some dummy variables of market_segment_type to remove multicollinearity.

	Variable	VIF
0	const	38.14690
1	no_of_adults	1.32534
2	no_of_children	2.00368
3	no_of_weekend_nights	1.05815
4	no_of_week_nights	1.08644
5	required_car_parking_space	1.03699
6	lead_time	1.24154
7	arrival_month	1.04792
8	repeated_guest	1.75775
9	no_of_previous_cancellations	1.36878
10	no_of_previous_bookings_not_canceled	1.61142
11	avg_price_per_room	1.70757
12	no_of_special_requests	1.23411
13	type_of_meal_plan_Meal Plan 2	1.19094
14	type_of_meal_plan_Meal Plan 3	1.00503
15	type_of_meal_plan_Not Selected	1.23786
16	room_type_reserved_Room_Type 2	1.09056
17	room_type_reserved_Room_Type 3	1.00223
18	room_type_reserved_Room_Type 4	1.33320
19	room_type_reserved_Room_Type 5	1.02815
20	room_type_reserved_Room_Type 6	1.95493
21	room_type_reserved_Room_Type 7	1.06401
22	market_segment_type_Offline	5.28275
23	market_segment_type_Online	5.95241

Figure 49: VIF after removing dummy variables

8.1.2. Dealing with high p-value variables

```
Optimization terminated successfully.
         Current function value: 0.426183
         Iterations 10
Dropping column room_type_reserved_Room_Type 3 with p-value: 0.6257719482026117
Optimization terminated successfully.
         Current function value: 0.426188
         Iterations 10
Dropping column no_of_previous_bookings_not_canceled with p-value: 0.4368540065144816
Optimization terminated successfully.
         Current function value: 0.426205
         Iterations 9
Dropping column type_of_meal_plan_Meal Plan 2 with p-value: 0.34777192684515545
Optimization terminated successfully.
         Current function value: 0.426222
         Iterations 9
Dropping column type_of_meal_plan_Meal Plan 3 with p-value: 0.32806390350922154
Optimization terminated successfully.
         Current function value: 0.426253
         Iterations 9
Dropping column no_of_children with p-value: 0.06007277921533484
Optimization terminated successfully.
         Current function value: 0.426321
         Iterations 9
Dropping column no_of_adults with p-value: 0.009257527645957633
```

Figure 50: Dropped Columns

Selected Features:

['const', 'no_of_adults', 'no_of_weekend_nights', 'no_of_week_nights', 'required_car_parking_space', 'lead_time', 'arrival_month', 'repeated_guest', 'no_of_previous_cancellations', 'avg_price_per_room', 'no_of_special_requests', 'type_of_meal_plan_Not Selected', 'room_type_reserved_Room_Type 2', '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_Offline', 'market_segment_type_Online']

8.1.3. Determining optimal threshold using ROC Curve

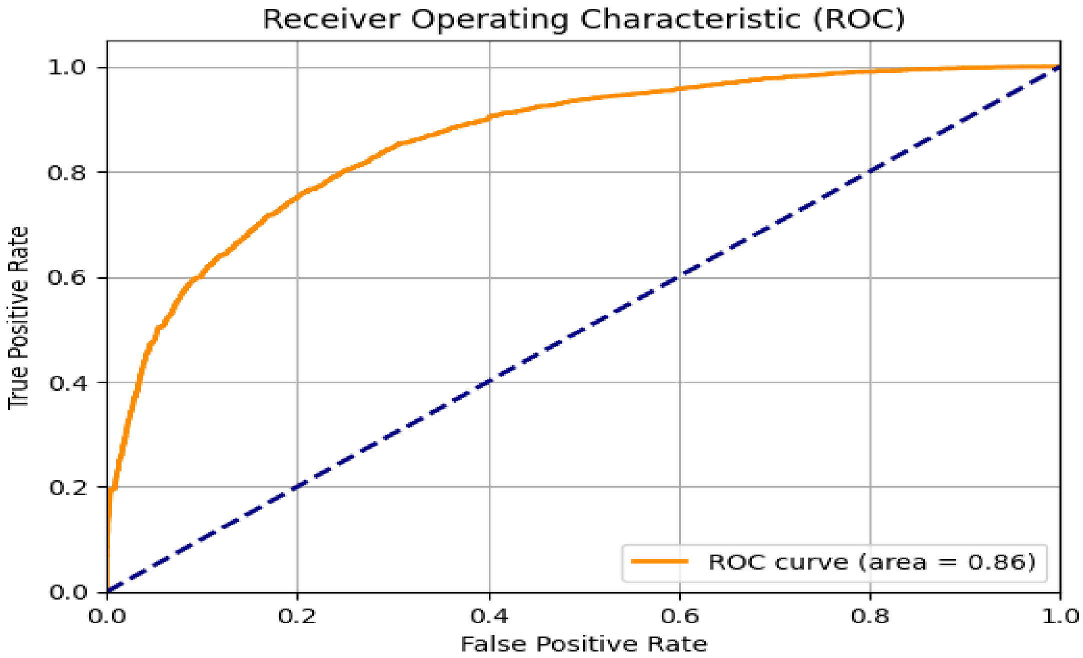

Figure 51: ROC curve

8.1.4. Tuning Logistic Regression model with significant features

```
                           Logit Regression Results
==============================================================================
Dep. Variable:     booking_status_Not_Canceled   No. Observations:        25392
Model:                                   Logit   Df Residuals:            25373
Method:                                    MLE   Df Model:                   18
Date:                         Fri, 09 Aug 2024   Pseudo R-squ.:          0.3247
Time:                                 19:37:33   Log-Likelihood:         -10825.
converged:                                True   LL-Null:                -16030.
Covariance Type:                     nonrobust   LLR p-value:             0.000
===========================================================================================
                                coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------
const                         3.6751      0.122     30.107      0.000       3.436       3.914
no_of_adults                 -0.0967      0.037     -2.602      0.009      -0.170      -0.024
no_of_weekend_nights         -0.1275      0.020     -6.449      0.000      -0.166      -0.089
no_of_week_nights            -0.0363      0.012     -2.959      0.003      -0.060      -0.012
required_car_parking_space    1.7169      0.142     12.081      0.000       1.438       1.995
lead_time                    -0.0160      0.000    -62.895      0.000      -0.016      -0.015
arrival_month                 0.0619      0.006     10.304      0.000       0.050       0.074
repeated_guest                2.8076      0.616      4.562      0.000       1.601       4.014
no_of_previous_cancellations -0.2551      0.075     -3.391      0.001      -0.403      -0.108
avg_price_per_room           -0.0204      0.001    -29.602      0.000      -0.022      -0.019
no_of_special_requests        1.4707      0.030     49.021      0.000       1.412       1.529
type_of_meal_plan_Not Selected -0.2504    0.052     -4.787      0.000      -0.353      -0.148
room_type_reserved_Room_Type 2 0.3841     0.127      3.031      0.002       0.136       0.632
room_type_reserved_Room_Type 4 0.2696     0.053      5.132      0.000       0.167       0.373
room_type_reserved_Room_Type 5 0.6873     0.207      3.315      0.001       0.281       1.094
room_type_reserved_Room_Type 6 0.8482     0.117      7.228      0.000       0.618       1.078
room_type_reserved_Room_Type 7 1.2391     0.299      4.143      0.000       0.653       1.825
market_segment_type_Offline   1.0370      0.100     10.382      0.000       0.841       1.233
market_segment_type_Online   -0.7701      0.097     -7.978      0.000      -0.959      -0.581
```

Figure 52: Model Summary

Interpretations:

- Intercept (const): Coefficient: 3.6751, which indicates a strong positive log-odds of a booking not being canceled when all other predictors are at their reference levels.
- Number of Adults (no_of_adults): Coefficient: -0.0967, slightly negative, suggesting that an increase in the number of adults slightly decreases the likelihood of the booking being completed.
- Number of Weekend Nights (no_of_weekend_nights): Coefficient: -0.1275, indicating that bookings over the weekend are slightly more likely to be canceled compared to weekday bookings.
- Required Car Parking Space (required_car_parking_space): Coefficient: 1.7169, showing a strong positive association with the likelihood of a booking not being canceled, similar to previous interpretations.
- Lead Time (lead_time): Coefficient: -0.0160, confirms that longer lead times are associated with higher chances of cancellation.
- Repeated Guest (repeated_guest): Coefficient: 2.8076, a very strong positive effect, indicating that bookings made by repeat guests are much more likely to be completed.
- Average Price Per Room (avg_price_per_room): Coefficient: -0.0204, implying that higher room prices slightly deter bookings from being completed.
- Type of Meal Plan (type_of_meal_plan_Not Selected): Coefficient: -0.2504, suggesting that not selecting a meal plan is associated with a higher chance of cancellation.
- Market Segment (market_segment_type_Online): Coefficient: -0.7701, indicating that bookings made through online market segments are more likely to be canceled compared to other segments.

8.1.5. Tuned Model Performance

Train Data

Accuracy	Recall	Precision	F1
0.80305	0.88886	0.83076	0.85883

Figure 53: Model Performance

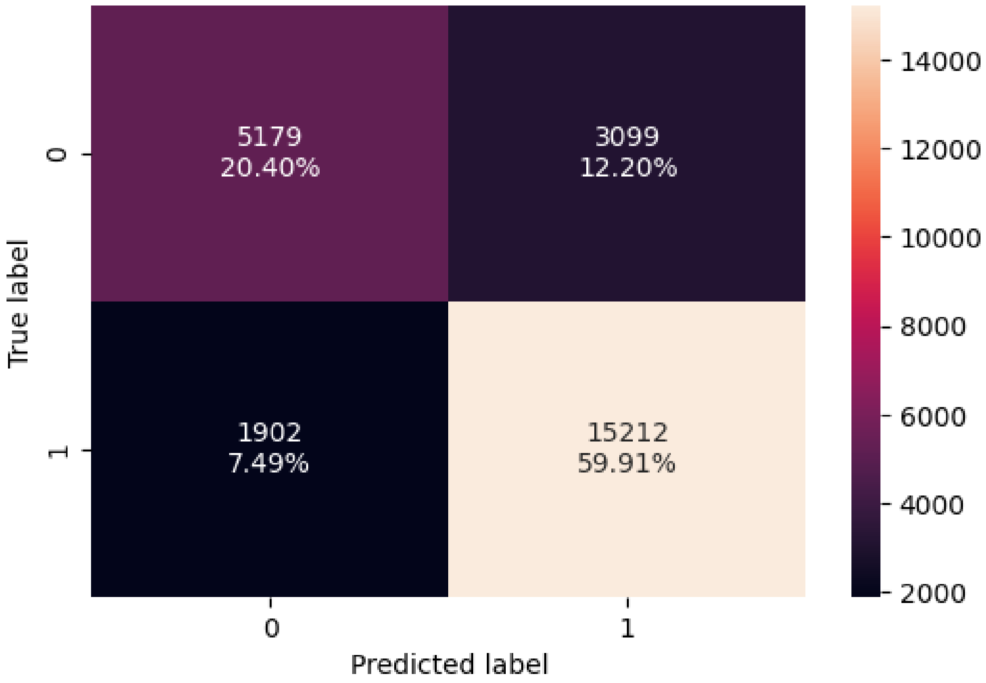

Figure 54: Confusion Matrix

Test Data

Accuracy	Recall	Precision	F1
0.80584	0.89761	0.82681	0.86076

Figure 55: Model Performance

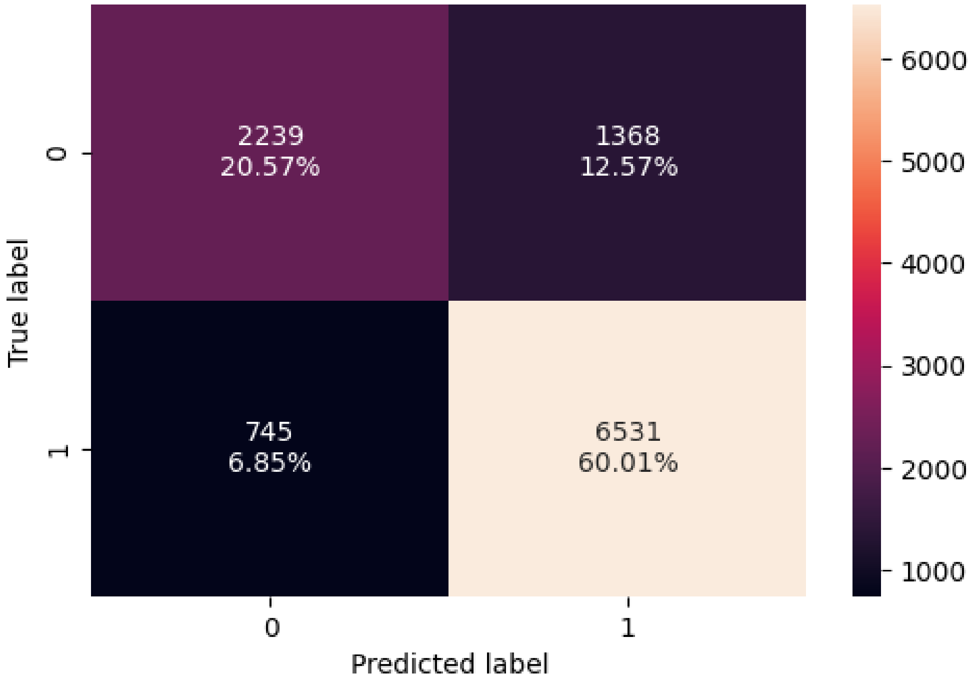

Figure 56: Confusion Matrix

8.2. KNN Classifier

With different k values,

```
Recall for k=2: 0.802913688840022
Recall for k=3: 0.9010445299615173
Recall for k=4: 0.8545904343045629
Recall for k=5: 0.909978009895547
Recall for k=6: 0.8767179769103903
Recall for k=7: 0.9092908191313909
Recall for k=8: 0.8840021990104453
Recall for k=9: 0.9097031335898845
Recall for k=10: 0.8919736118746564
Recall for k=11: 0.9125893347993403
Recall for k=12: 0.8919736118746564
Recall for k=13: 0.9131390874106652
Recall for k=14: 0.8993952721275426
Recall for k=15: 0.9171247938427708
Recall for k=16: 0.902693787795492
Recall for k=17: 0.9175371083012644
Recall for k=18: 0.903243540406817
Recall for k=19: 0.9153380978559648

The best value of k is: 17 with a recall of: 0.9175371083012644
```

Figure 57: KNN Classifier Performance Improvement using different k values

8.2.1. Tuned Model Performance

Train Data

Accuracy	Recall	Precision	F1
0.85944	0.91989	0.87749	0.89819

Figure 58: Model Performance

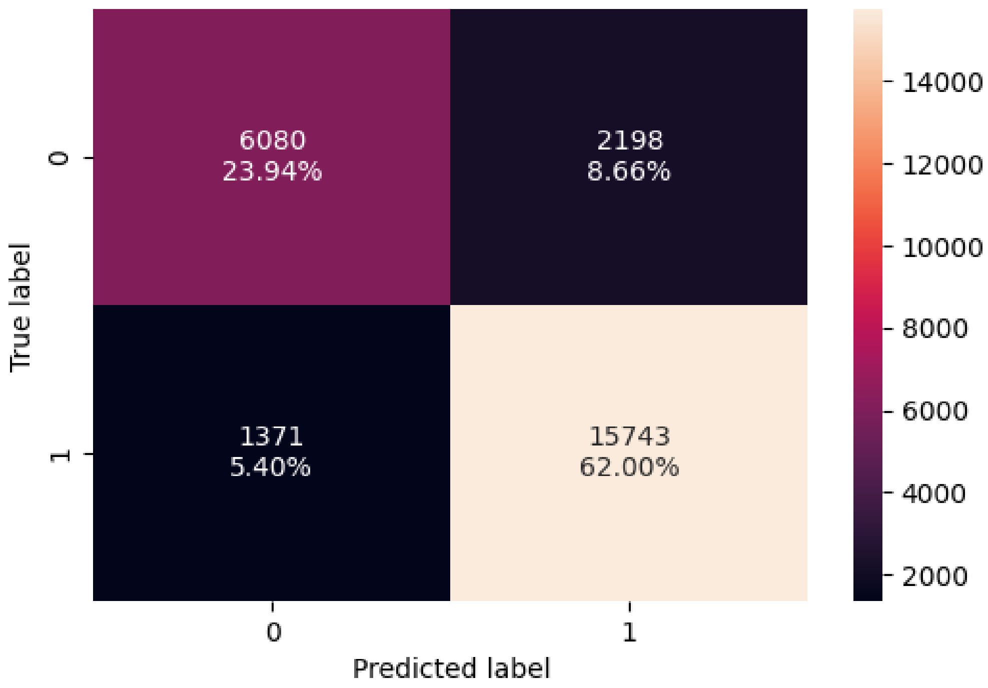

Figure 59: Confusion Matrix

Test Data

Accuracy	Recall	Precision	F1
0.85087	0.91754	0.86713	0.89162

Figure 60: Model Performance

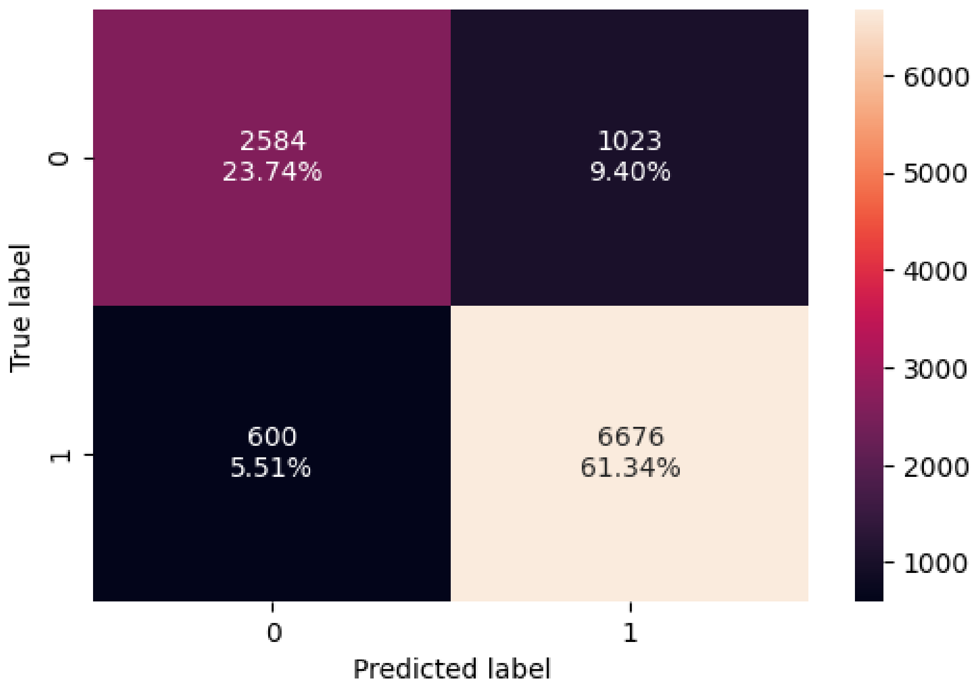

Figure 61: Confusion Matrix

8.3. Decision Tree Classifier

8.3.1. Pre-Pruning the tree

```
DecisionTreeClassifier
DecisionTreeClassifier(max_depth=11, max_leaf_nodes=100, random_state=42)
```

Figure 62: Best Estimators

8.3.2. Tuned Model Performance

Train Data

Accuracy	Recall	Precision	F1
0.87287	0.92924	0.88744	0.90786

Figure 63: Model Performance

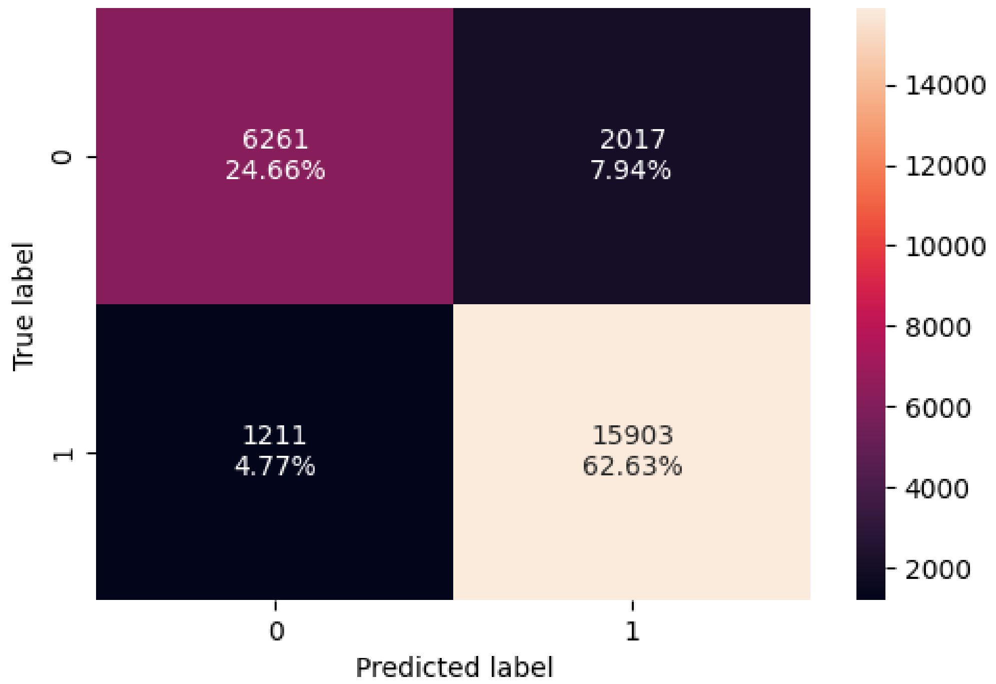

Figure 64: Confusion Matrix

Test Data

Accuracy	Recall	Precision	F1
0.87108	0.93101	0.88261	0.90616

Figure 65: Model Performance

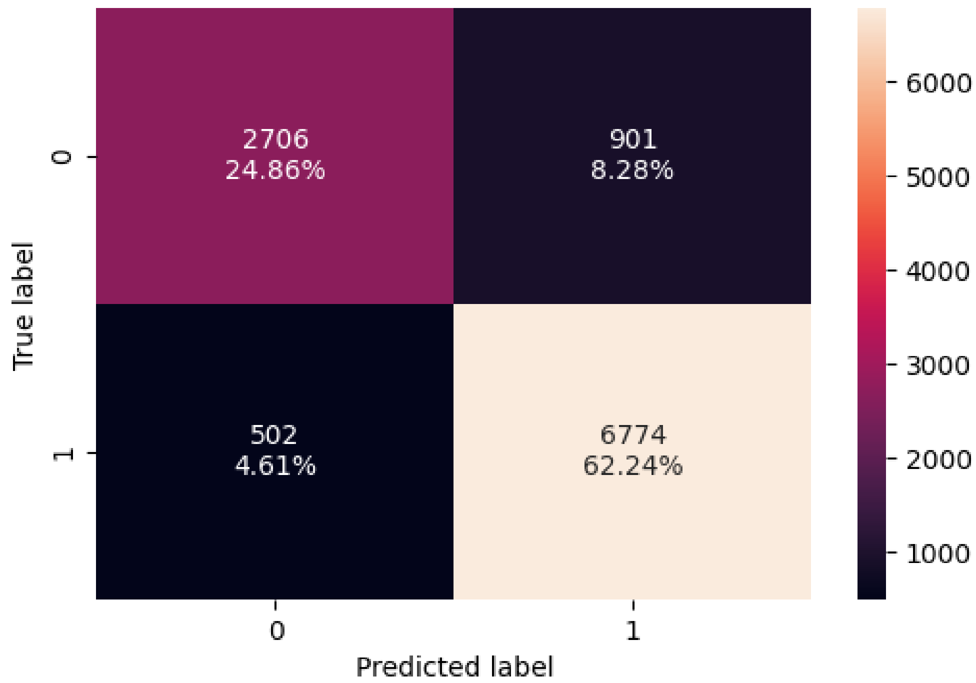

Figure 66: Confusion Matrix

8.3.3. Visualizing decision tree

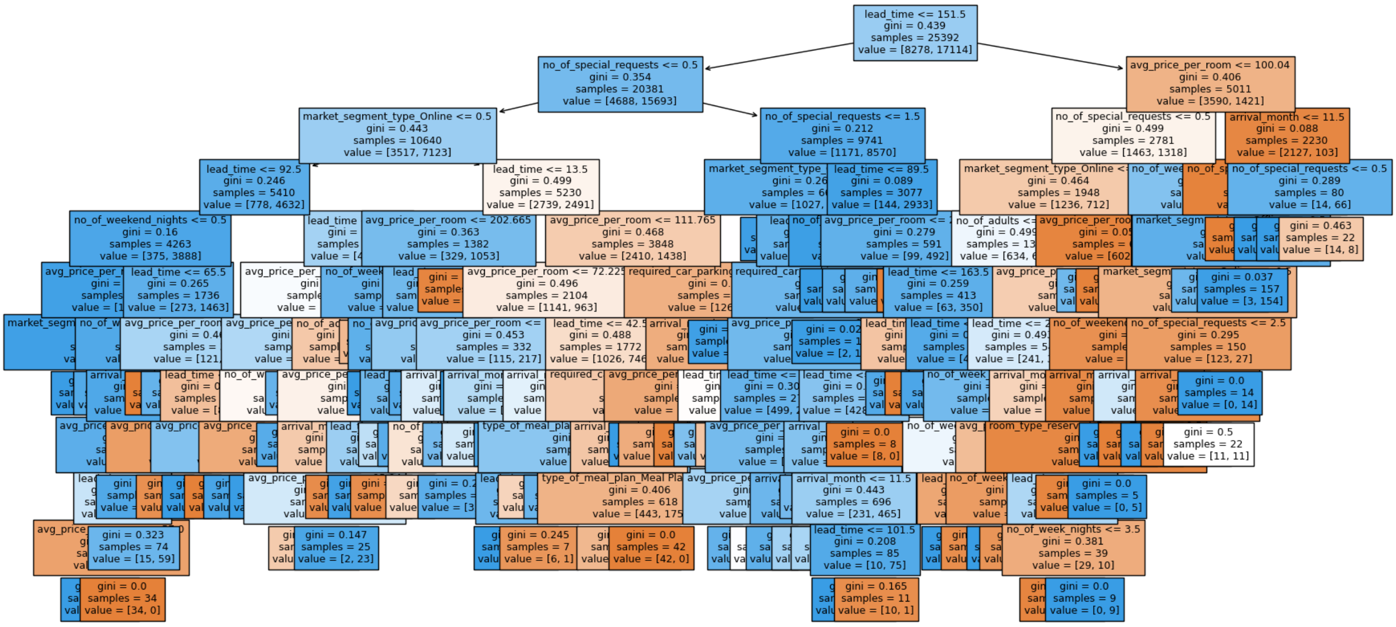

Figure 67: Tree

8.3.4. Feature Importance

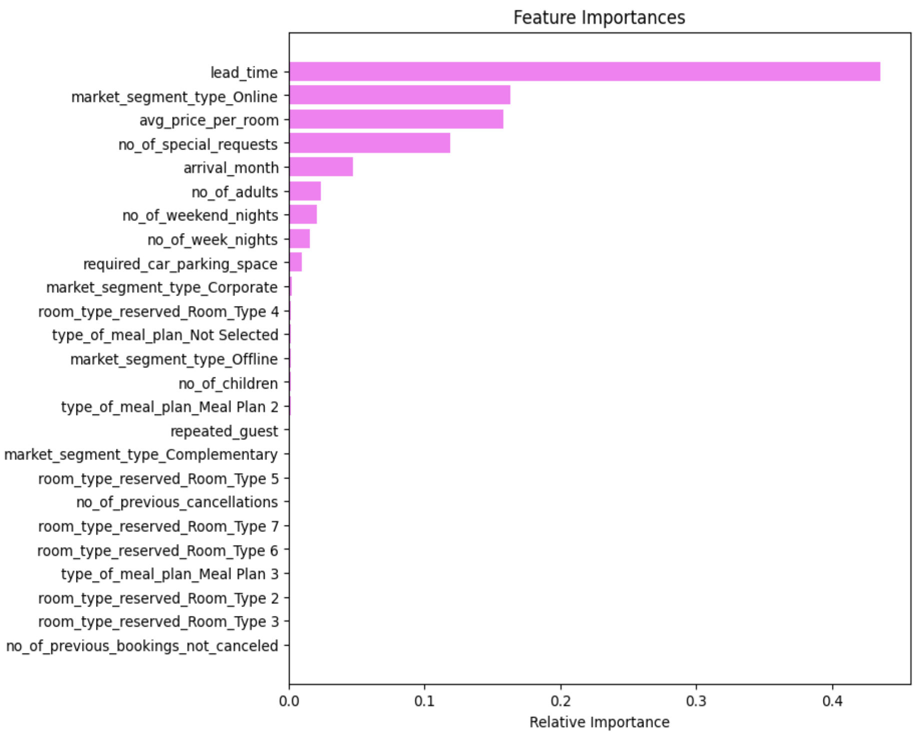

Figure 68: Feature Importance

9. Model Performance Comparison and Final Model Selection

Train:

	Logistic Regression Base	Logistic Regression Tuned	Naive Bayes Base	KNN Base	KNN Tuned	Decision Tree Base	Decision Tree Tuned
Accuracy	0.80289	0.80305	0.40855	0.91671	0.85944	0.99374	0.87287
Recall	0.88910	0.88886	0.14029	0.94992	0.91989	0.99509	0.92924
Precision	0.83043	0.83076	0.88729	0.92818	0.87749	0.99562	0.88744
F1	0.85876	0.85883	0.24228	0.93892	0.89819	0.99535	0.90786

Figure 69: Train data Model Performance Comparison

Test:

	Logistic Regression Base	Logistic Regression Tuned	Naive Bayes Base	KNN Base	KNN Tuned	Decision Tree Base	Decision Tree Tuned
Accuracy	0.80289	0.80305	0.41459	0.85234	0.85087	0.86364	0.87108
Recall	0.88910	0.88886	0.14184	0.90104	0.91754	0.89321	0.93101
Precision	0.83043	0.83076	0.89042	0.88083	0.86713	0.90189	0.88261
F1	0.85876	0.85883	0.24469	0.89082	0.89162	0.89753	0.90616

Figure 70: Test data Model Performance Comparison

9.1. Insights and Final Model Selection - Decision Trees

- Logistic Regression: Since tuning only slightly improves the model, logistic regression may not be the best choice. However, its high recall indicates it's good for scenarios where missing a positive instance is costly. Further improvements might not yield substantial gains.
- Naive Bayes: Naive Bayes is not suitable for this problem due to poor performance in recall and accuracy. Consider removing it from the model set.
- KNN: The base KNN model performs exceptionally well, making it a strong candidate. The tuned model's decline suggests tuning may have negatively impacted its performance. Focus on optimizing hyperparameters or feature selection for potential gains.
- Decision Tree: The base decision tree model performs the best overall. It should be considered the primary model due to its superior balance of metrics. Further tuning seems to degrade its performance, so focus on maintaining its current state.

Given the high accuracy, recall, precision, and F1 score, decision trees are the most promising.

10. Actionable Insights and Recommendations

- Focus on Decision Trees: Given the high accuracy, recall, precision, and F1 score, decision trees (base version) are the most promising.
- Consider Ensemble Methods: Use ensemble methods like Random Forests or Gradient Boosting to potentially improve model performance further.
- Model Interpretability: Decision trees provide easily interpretable models, which can be an advantage in understanding the factors influencing predictions.
- Feature Engineering: Explore additional feature engineering to enhance model inputs, which might improve the performance of logistic regression or KNN models.
- Based on the Feature importance plot:
- Lead Time: This is the most important feature, suggesting that the amount of time between booking and arrival significantly influences the outcome.
- Market Segment Type (Online and Offline): These features also hold substantial importance, indicating that the channel through which the booking is made affects the results.
- Average Price Per Room: Another critical feature that impacts the model, likely reflecting the customer's budget or the quality of the room.
- Number of Special Requests: This feature's importance suggests that more personalized service requests may correlate with specific outcomes (like customer satisfaction or repeat bookings).
- Investigate creating new features that might capture the underlying processes better, such as categorizing 'lead time' into different time frames or deriving features from 'avg_price_per_room' that relate to service or amenities offered.