MACHINE LEARNING 1
BUSINESS REPORT

Krishnan CS

GREAT LEARNING DSBA

Contents

1. Problem Statement

1.1 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 behaviour. 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.

1.2 Objective

The increasing number of cancellations calls for a Machine Learning based solution that can help in predicting which booking is likely to be cancelled. 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 analyse 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 cancelled in advance, and help in formulating profitable policies for cancellations and refunds.

1.3 Data Description

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

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 cancelled by the customer prior to the current booking
- no_of_previous_bookings_not_canceled: Number of previous bookings not cancelled 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 cancelled or not.

1.4 EDA 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 cancelled?
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?

2 Exploratory Data Analysis

2.1 Problem Definition

INN Hotels Group has a chain of hotels in Portugal, they are facing problems with the high number of booking cancellations. The increasing number of cancellations calls for a Machine Learning based solution that can help in predicting which booking is likely to be cancelled.

2.2 Univariate Analysis

A description of all the columns is shown in the figure.

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Booking_ID	36275	36275	INN00001	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
no_of_adults	36275.0	NaN	NaN	NaN	1.84	0.52	0.0	2.0	2.0	2.0	4.0
no_of_children	36275.0	NaN	NaN	NaN	0.11	0.4	0.0	0.0	0.0	0.0	10.0
no_of_weekend_nights	36275.0	NaN	NaN	NaN	0.81	0.87	0.0	0.0	1.0	2.0	7.0
no_of_week_nights	36275.0	NaN	NaN	NaN	2.2	1.41	0.0	1.0	2.0	3.0	17.0
type_of_meal_plan	36275	4	Meal Plan 1	27835	NaN	NaN	NaN	NaN	NaN	NaN	NaN
required_car_parking_space	36275.0	NaN	NaN	NaN	0.03	0.17	0.0	0.0	0.0	0.0	1.0
room_type_reserved	36275	7	Room_Type 1	28130	NaN	NaN	NaN	NaN	NaN	NaN	NaN
lead_time	36275.0	NaN	NaN	NaN	85.23	85.93	0.0	17.0	57.0	126.0	443.0
arrival_year	36275.0	NaN	NaN	NaN	2017.82	0.38	2017.0	2018.0	2018.0	2018.0	2018.0
arrival_month	36275.0	NaN	NaN	NaN	7.42	3.07	1.0	5.0	8.0	10.0	12.0
arrival_date	36275.0	NaN	NaN	NaN	15.6	8.74	1.0	8.0	16.0	23.0	31.0
market_segment_type	36275	5	Online	23214	NaN	NaN	NaN	NaN	NaN	NaN	NaN
repeated_guest	36275.0	NaN	NaN	NaN	0.03	0.16	0.0	0.0	0.0	0.0	1.0
no_of_previous_cancellations	36275.0	NaN	NaN	NaN	0.02	0.37	0.0	0.0	0.0	0.0	13.0
no_of_previous_bookings_not_canceled	36275.0	NaN	NaN	NaN	0.15	1.75	0.0	0.0	0.0	0.0	58.0
avg_price_per_room	36275.0	NaN	NaN	NaN	103.42	35.09	0.0	80.3	99.45	120.0	540.0
no_of_special_requests	36275.0	NaN	NaN	NaN	0.62	0.79	0.0	0.0	0.0	1.0	5.0
booking_status	36275	2	Not_Canceled	24390	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figure 1 Summary of data

Booking_ID: This column only serves to give an identity to the booking. This will be dropped during analysis.

No_of_adults: Roughly 72% of the total bookings have 2 adults and 21% is for a single adult. There are some bookings with 0 adults. In such bookings, the *no_of_children* is greater than zero. Those rows of data are also valid. All values other than 2, are potential outliers here. However, they will not be treated.

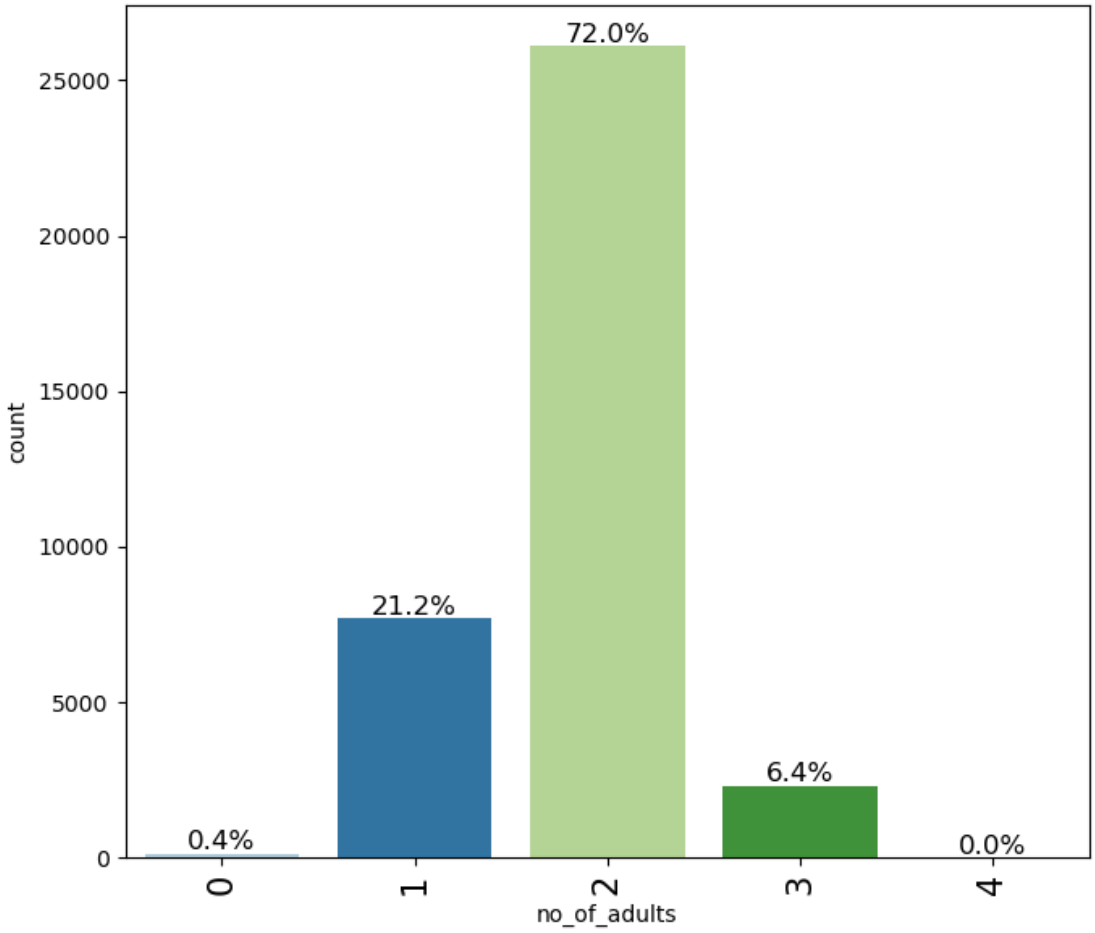

Figure 2 Distribution for number of adults

No_of_children: In 92.5% or more of the bookings, there are no children. In a few cases, these values even go to 10. All values other than zero, are potential outliers. They will not be treated, unless it is necessary.

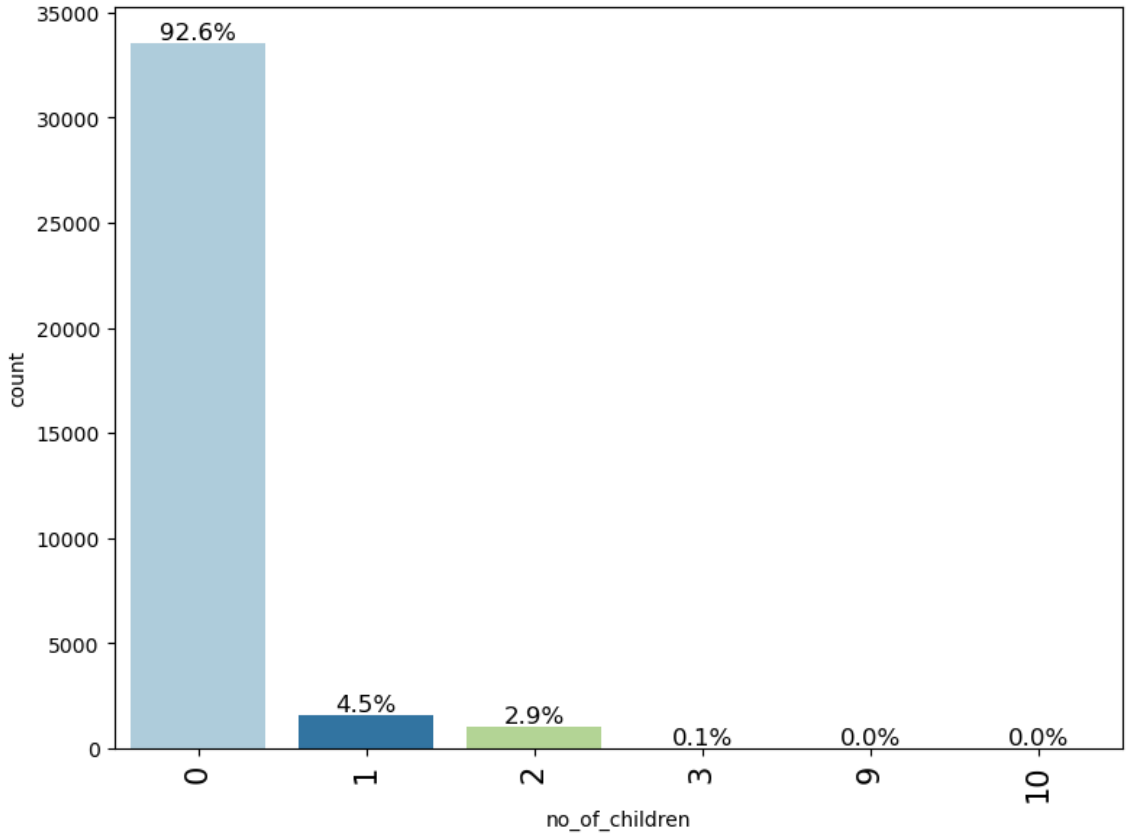

Figure 3 Distribution for number of children

No_of_weekend_nights:

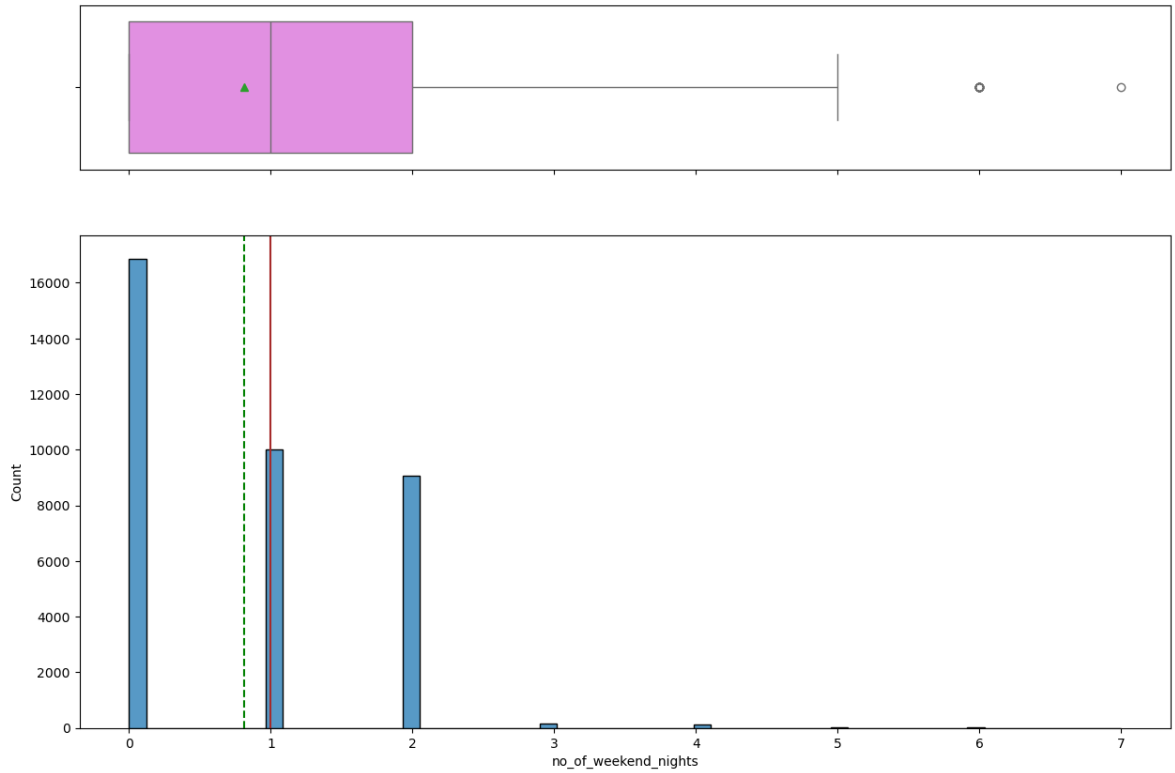

Figure 4 Distribution for number of weekend nights

In most of the cases, the number of weekend nights is 0, 1 or 2. Values greater than 5 are considered outliers. They are still valid values and will not be treated.

No_of_week_nights:

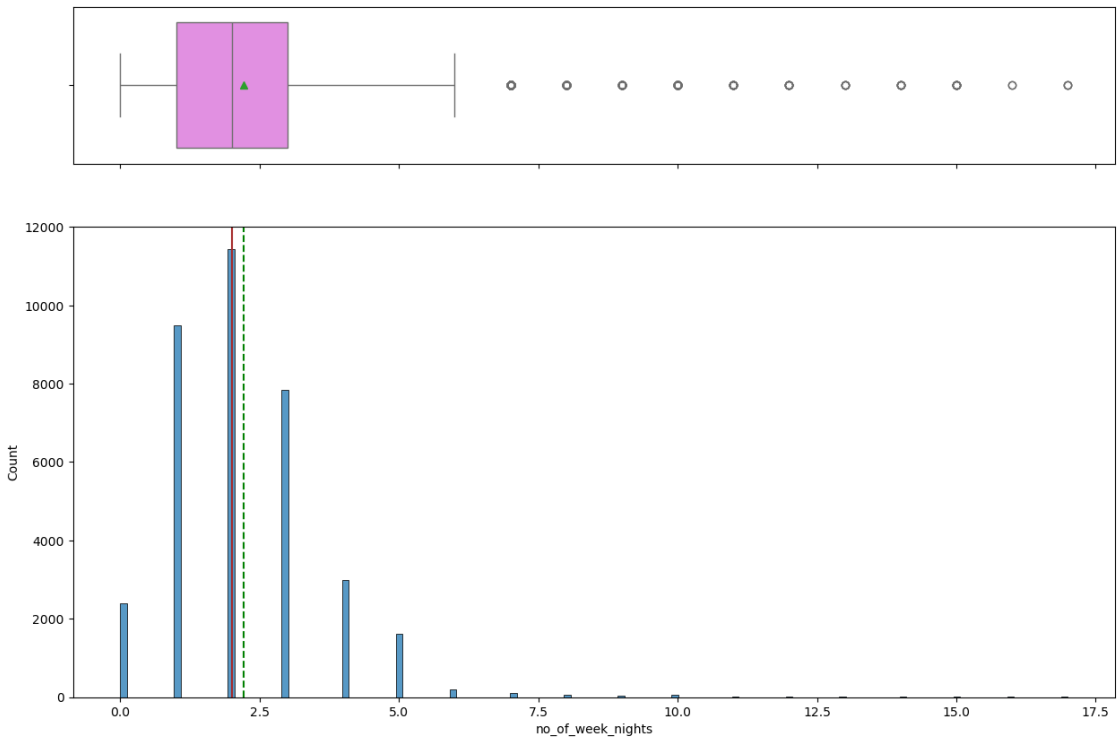

Figure 5 Distribution for number of week nights

For week_nights, the most common number of week_nights is 2. The values from 0 to 5 are not uncommon. Values beyond 6 are considered as outliers. They are still valid values and will not be treated.

Type_of_meal_plan: ~77% of the guests opted for Mean plan 1 and ~14% did not select any. Very few chose Mean plan 3. More analysis is needed to study the effect of meal plan on cancellations.

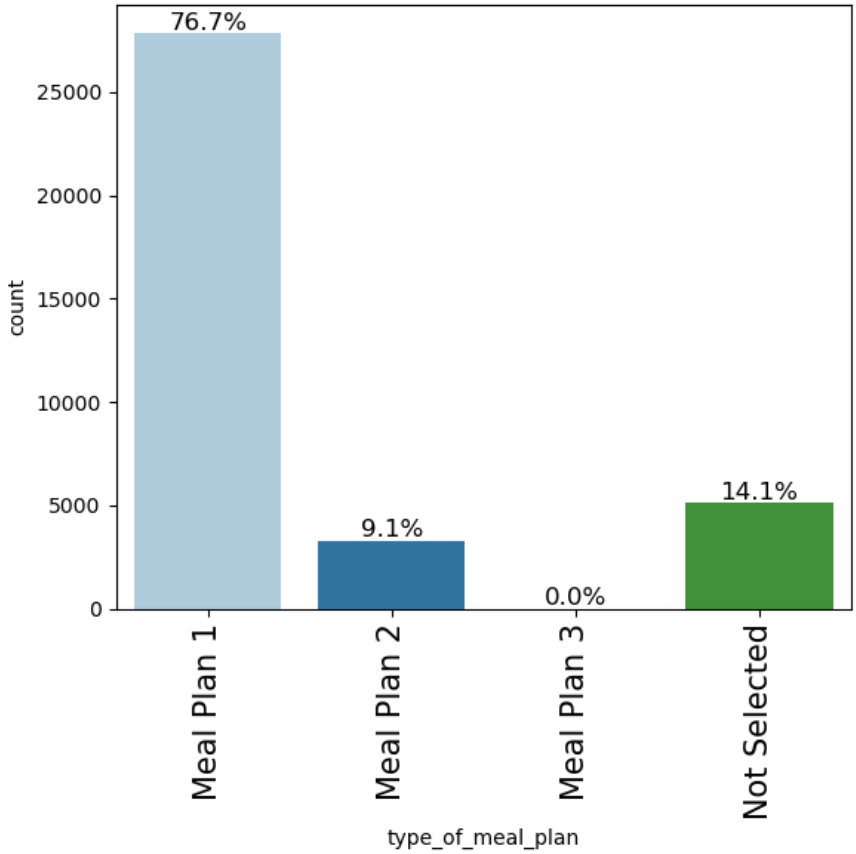

Figure 6 Distribution for type of meal plan

required_car_parking_space: Around 97% of the guests do not require for car parking space. The effect of this factor needs to be considered for cancellations.

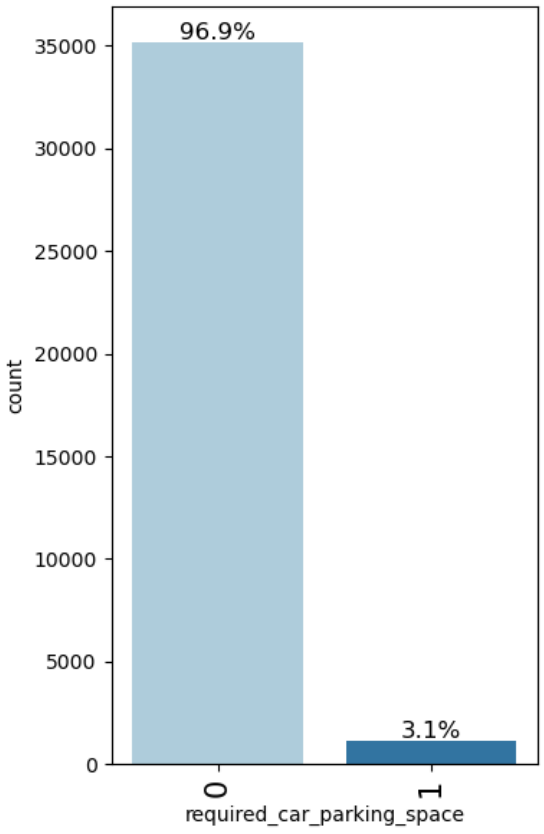

Figure 7 Distribution for required car space

room_type_reserved: Room type 1 and room type 4 are the most frequently chosen options. This data will also be considered for analysing cancellations.

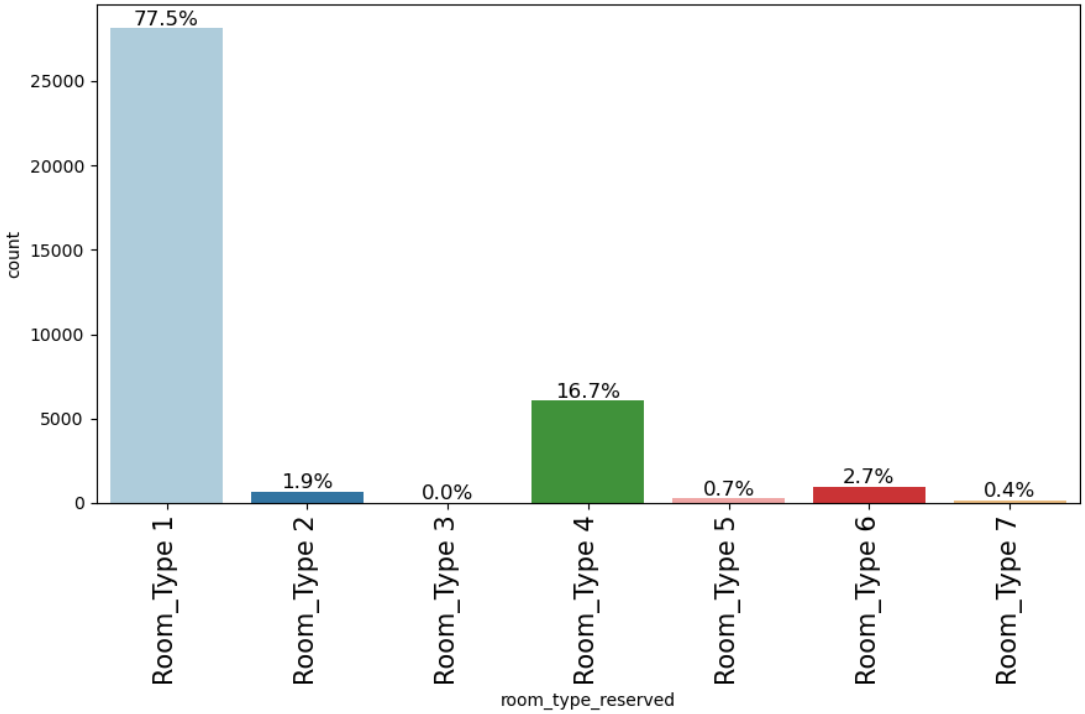

Figure 8 Distribution for room type reserved

lead_time: Each bar in the histogram represents ~20 days. The boxplot above, shows the 25th and 50th percentiles. From the combined plot, it is evident that more than 25% of the bookings have less than 20 days lead time and more than 50% have less than 60 days lead time.

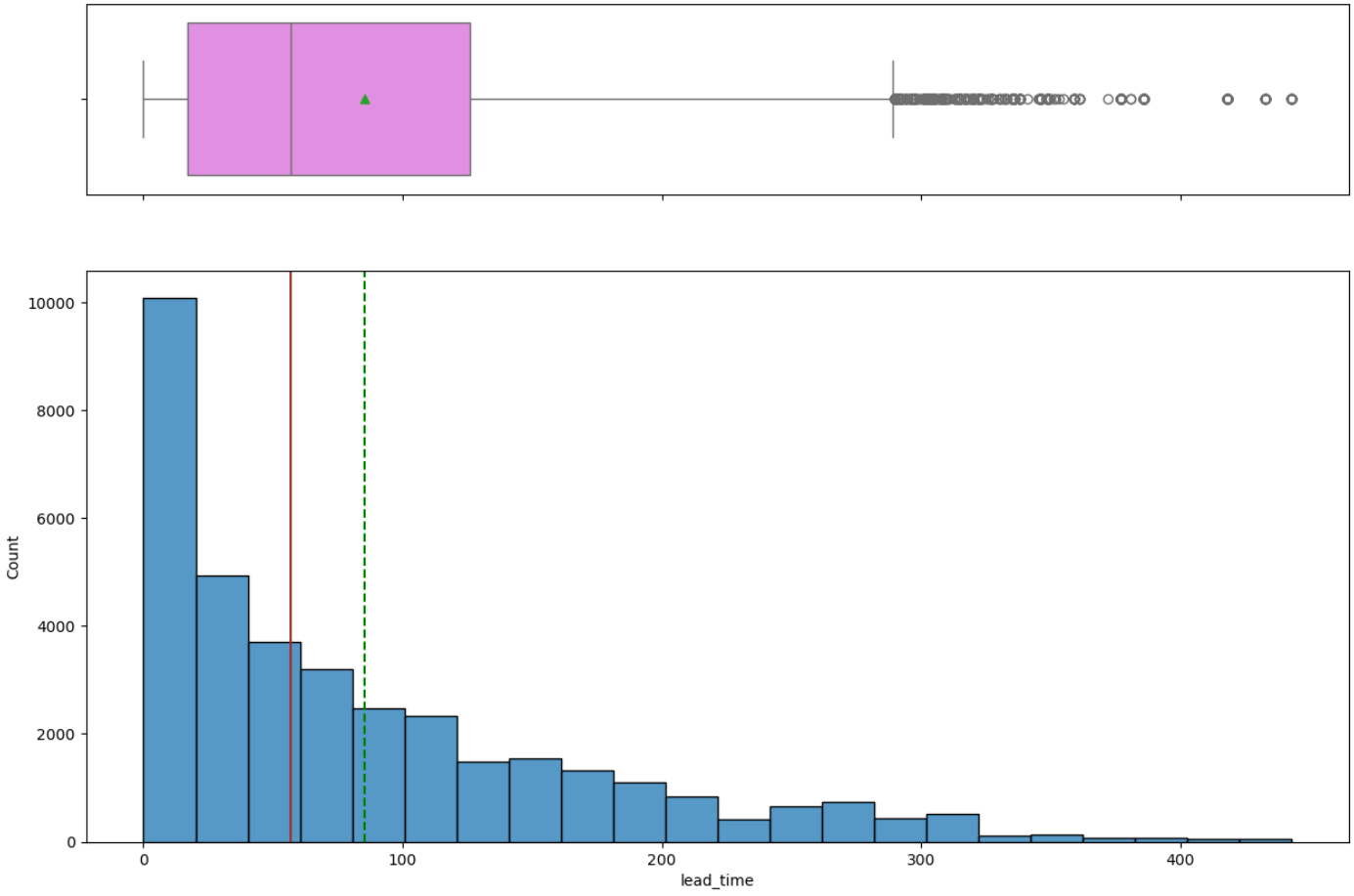

Figure 9 Distribution for lead time

arrival_year: 82% of the day is from 2018 and 18% of the day is from 2017. Since the data is available only from July 2017, this distribution is on expected lines.

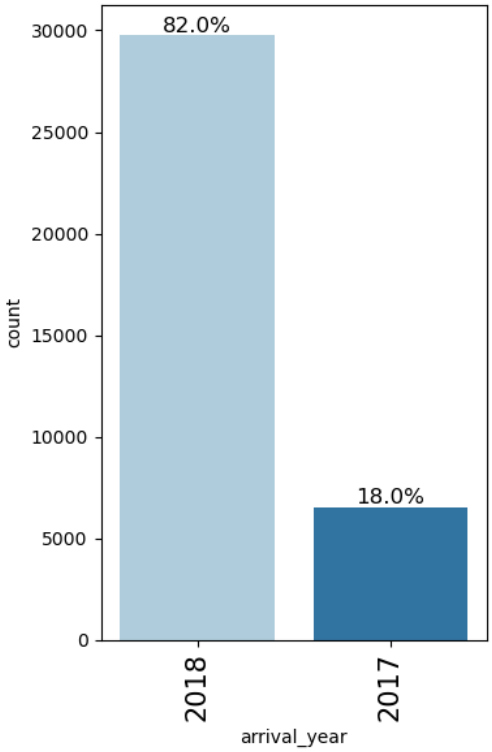

Figure 10 Distribution for arrival year

arrival_month: The data is available only from July 2017 to December 2018. From January to June, there is no data for the year 2017. October 2018 has been the busiest month with 3404 guests, closely followed by June 2018 at 3203 guests. The effect of month on the cancellations needs to be studied.

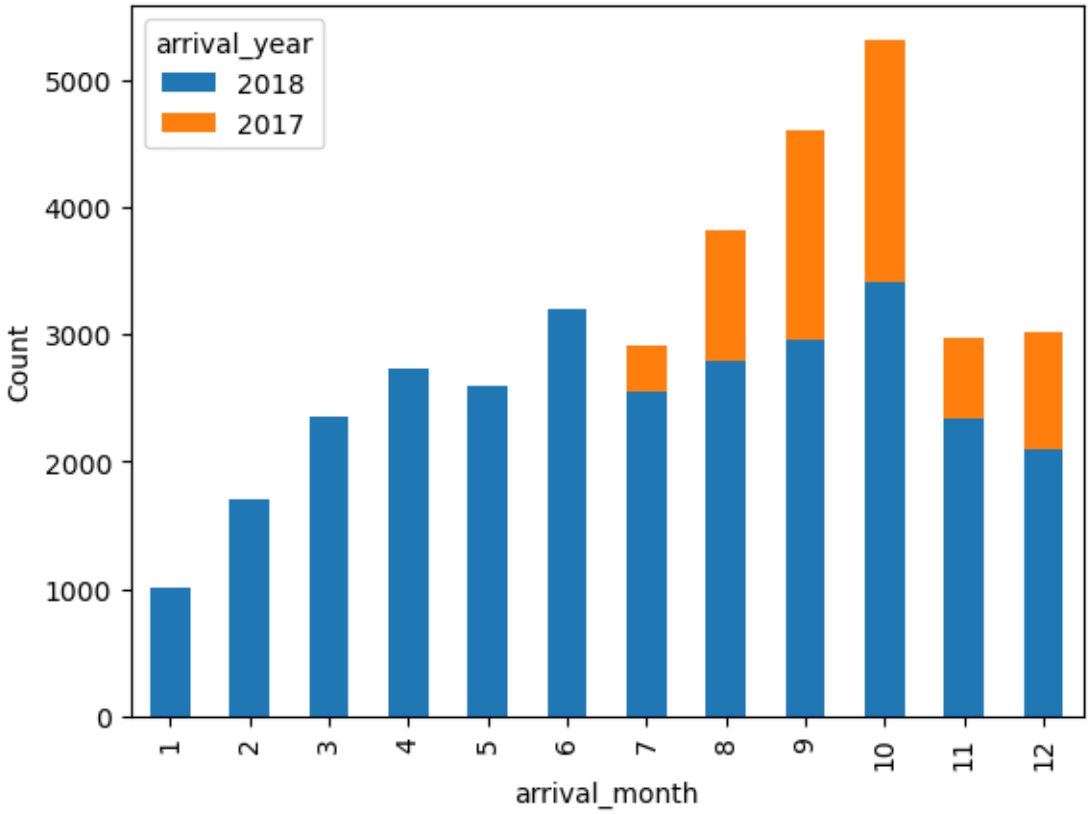

Figure 11 Distribution for arrival month

arrival_date: Date as a number by itself is insignificant and could be dropped.

market_segment_type: 64% of the guests book their stay online and 29% do it offline. The rest come from corporate, complementary or aviation. The cancellation trends for the different market segments would be studied.

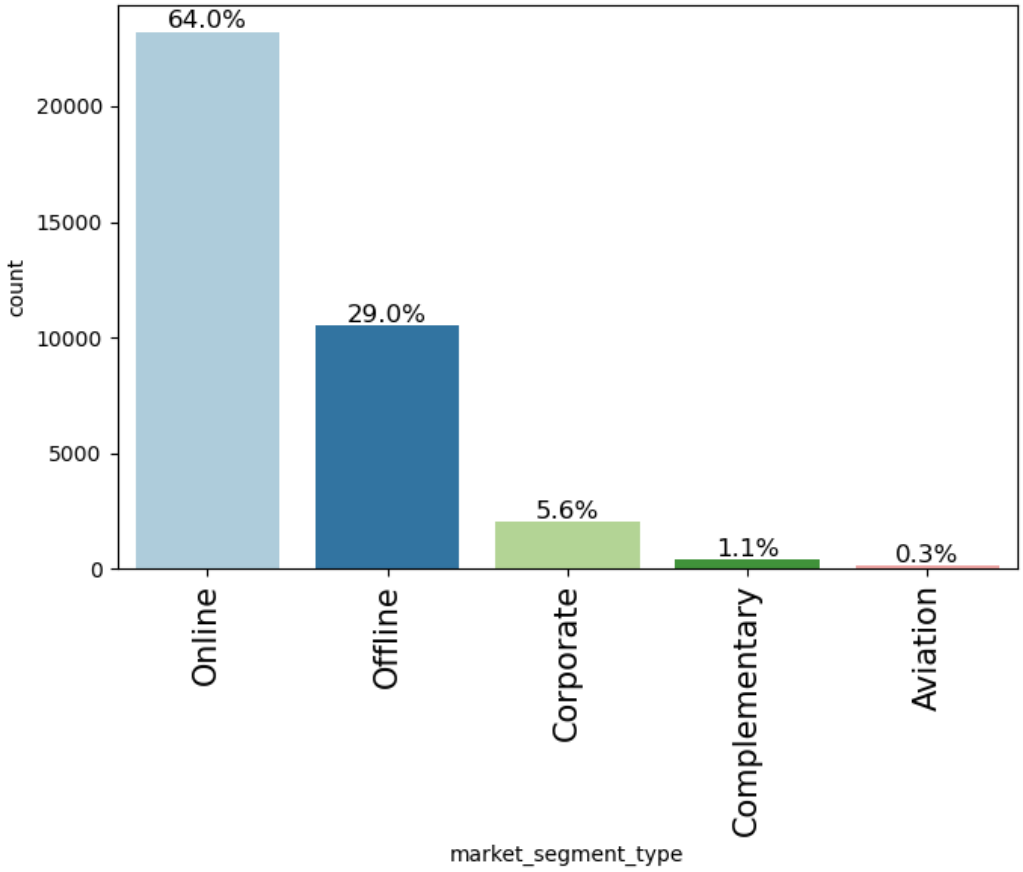

Figure 12 Distribution for market segment type

repeated_guest: 97.4% of the guests are first time visitors, while the remaining have experienced the hotel before. It would be important to check if the guests who visit repeatedly, cancel their stays.

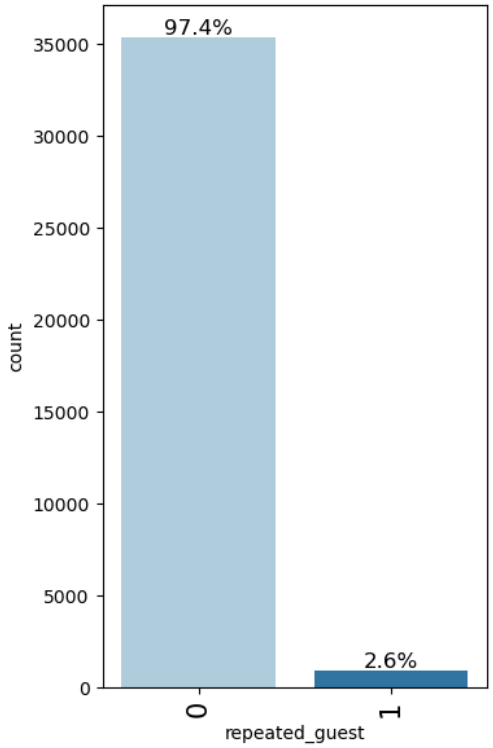

Figure 13 Distribution for repeated guest

no_of_previous_cancellations: 99% of the customers do not have a history of cancellations. The remaining 1% have cancelled their stays. A few have done it multiple times too. The outliers could be treated by converting all non-zero values to 1, making it essentially a Yes or No field.

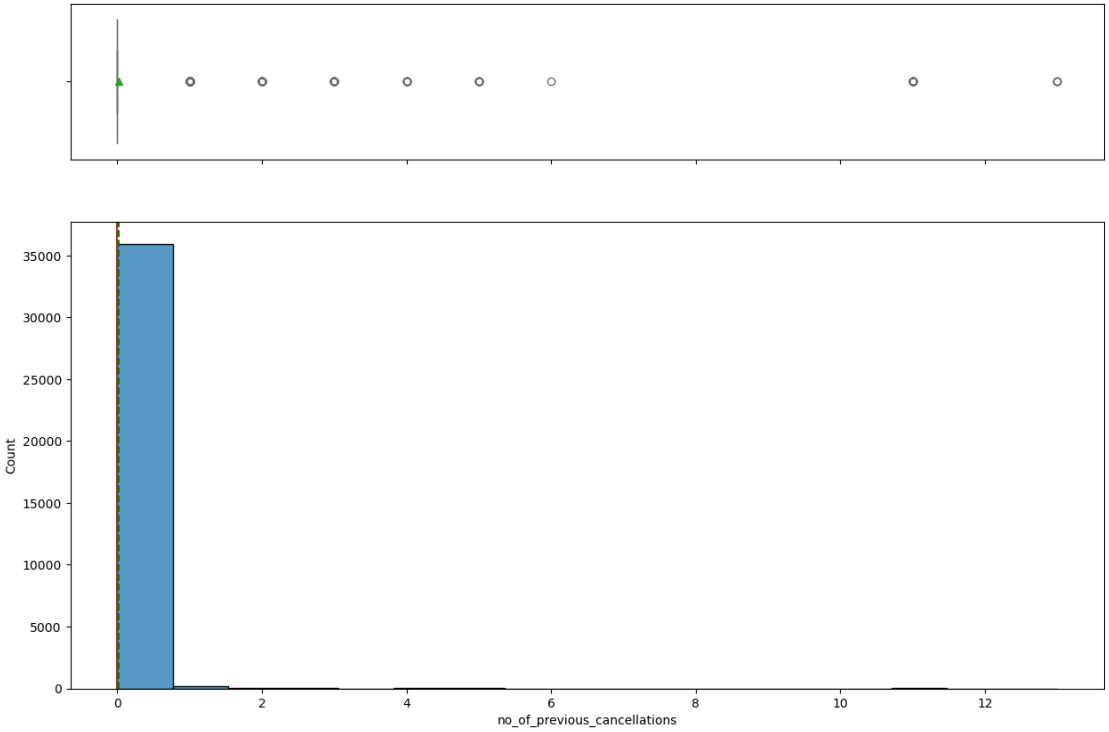

Figure 14 Distribution for number of previous cancellations

no_of_previous_bookings_not_canceled: This column shows a similar trend to *no_of_previous_cancellations* and will be treated in a similar fashion.

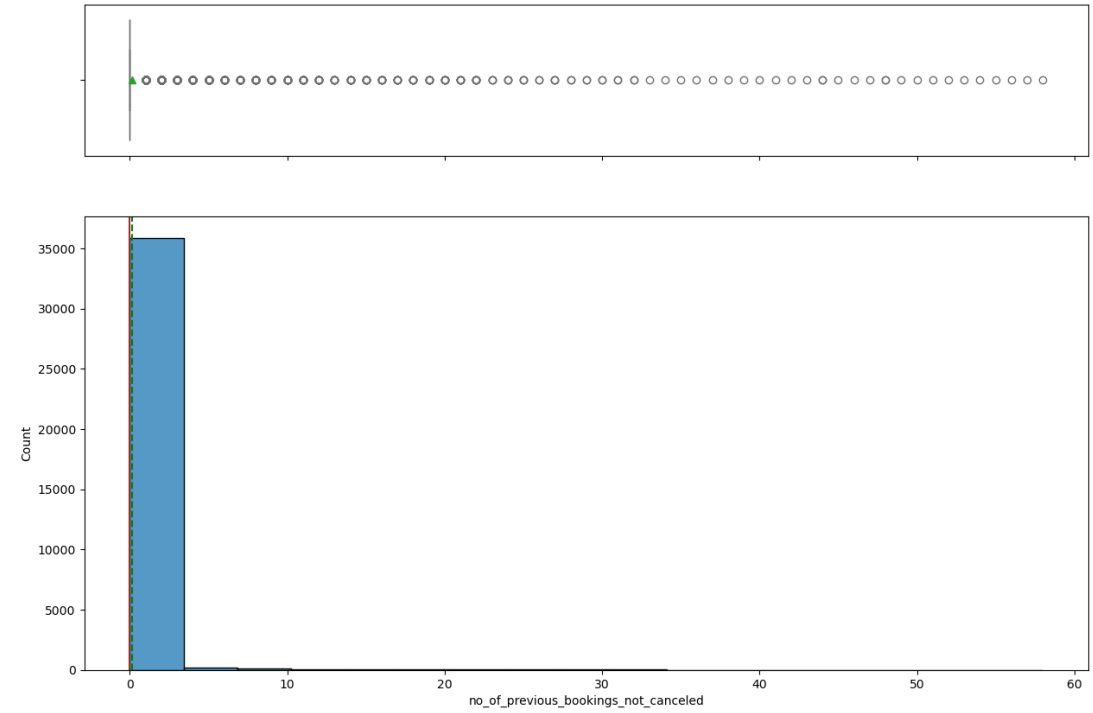

Figure 15 Distribution for number of previous bookings not cancelled

avg_price_per_room: The distribution of the average price per room is heavily clustered around the mean / median value of around 100 euros. There are some very expensive rooms at 540 euros and some complementary rooms as well (at 0 euros). None of these values are wrong. They all will be retained during analysis. However, if there are issues in the model, the outliers could be clipped.

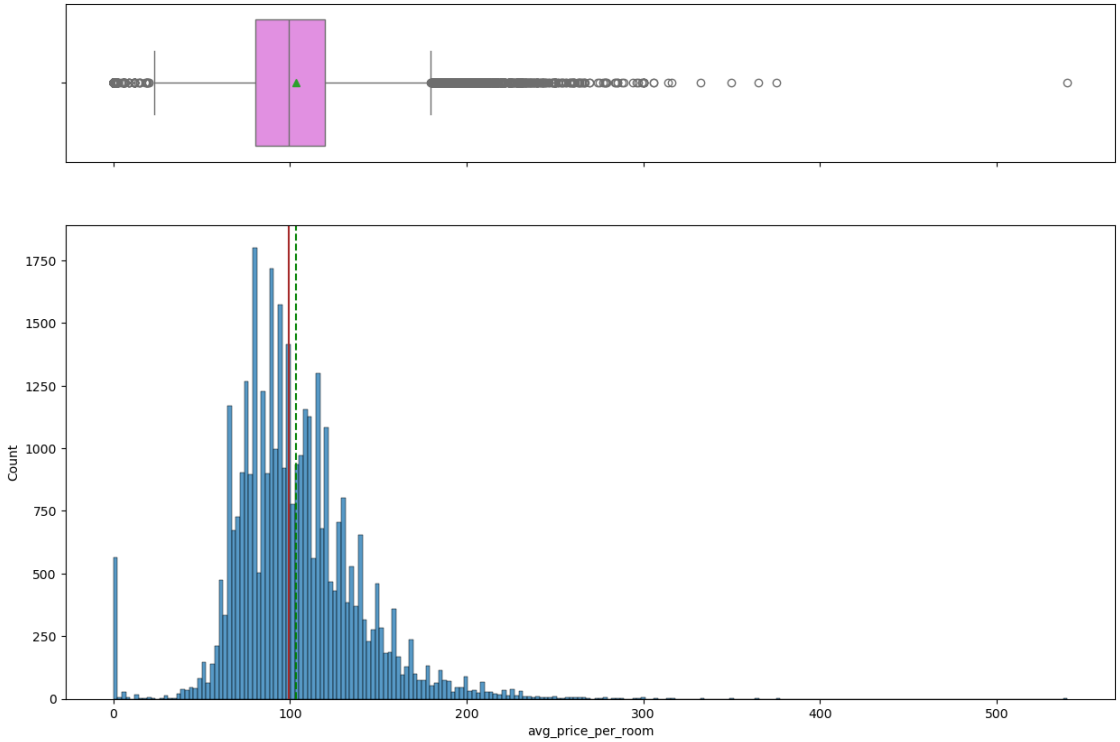

Figure 16 Distribution for average price per room

no_of_special_requests: In around 55% of the cases, there are no special requests. In another 43% there are 1-2 special requests. More than 2 requests are considered as outliers. They could be retained in the analysis and only treated if required.

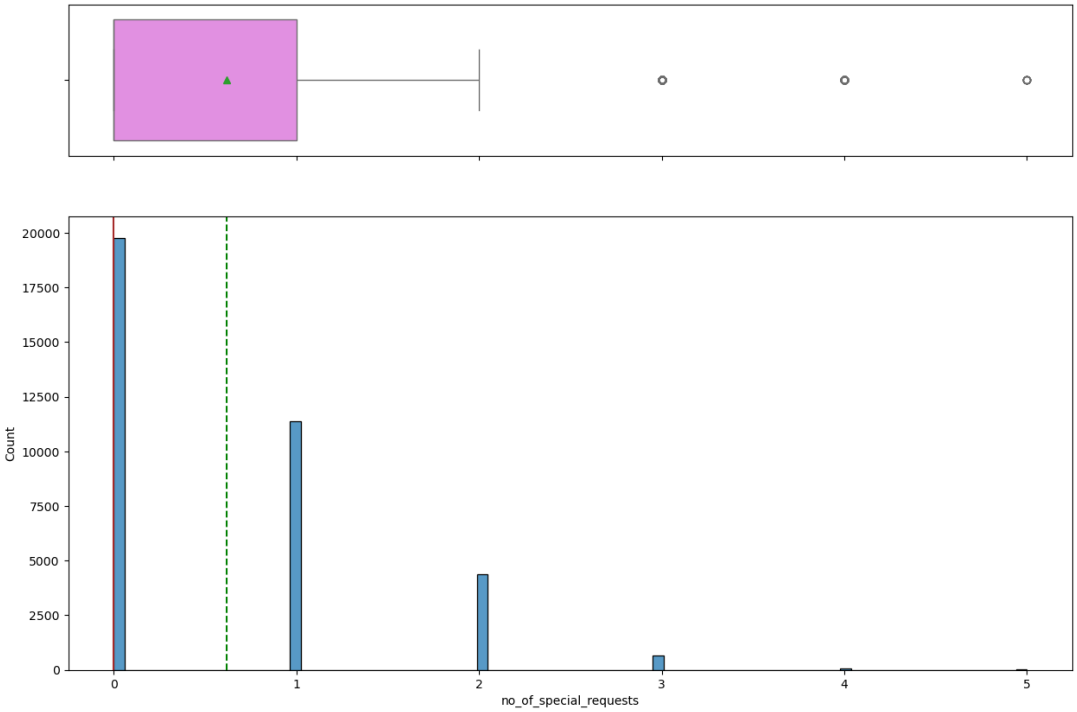

Figure 17 Distribution for number of special requests

booking_status: Around 67% of the data has the status of *not_canceled*, while the rest are *canceled*. This is the dependent variable for creating the model and all analysis will be centred around it. The value of *canceled* will be set to 1 and *non_canceled* will be set to 0 during analysis.

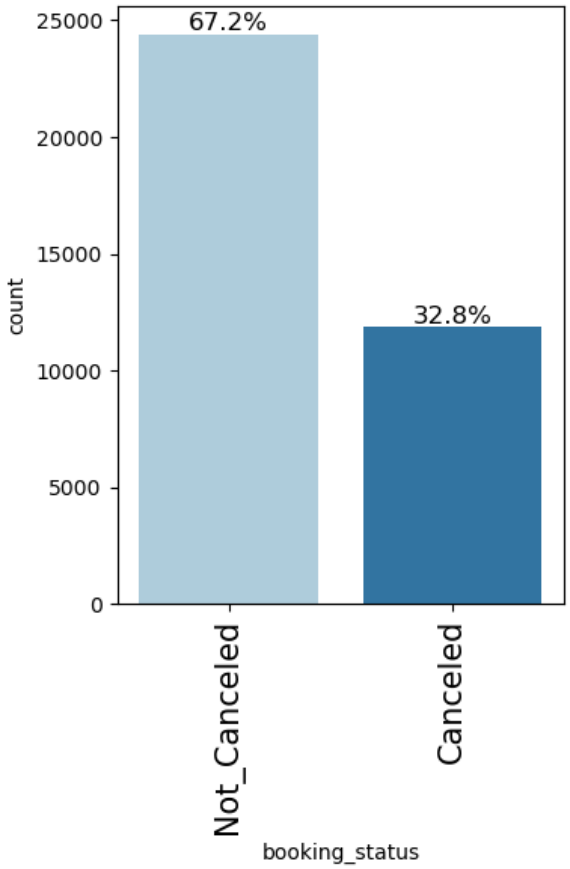

Figure 18 Distribution for booking status

2.3 Bivariate Analysis

A quick way to start the bivariate analysis is a heatmap, using which all the numerical columns could be studied.

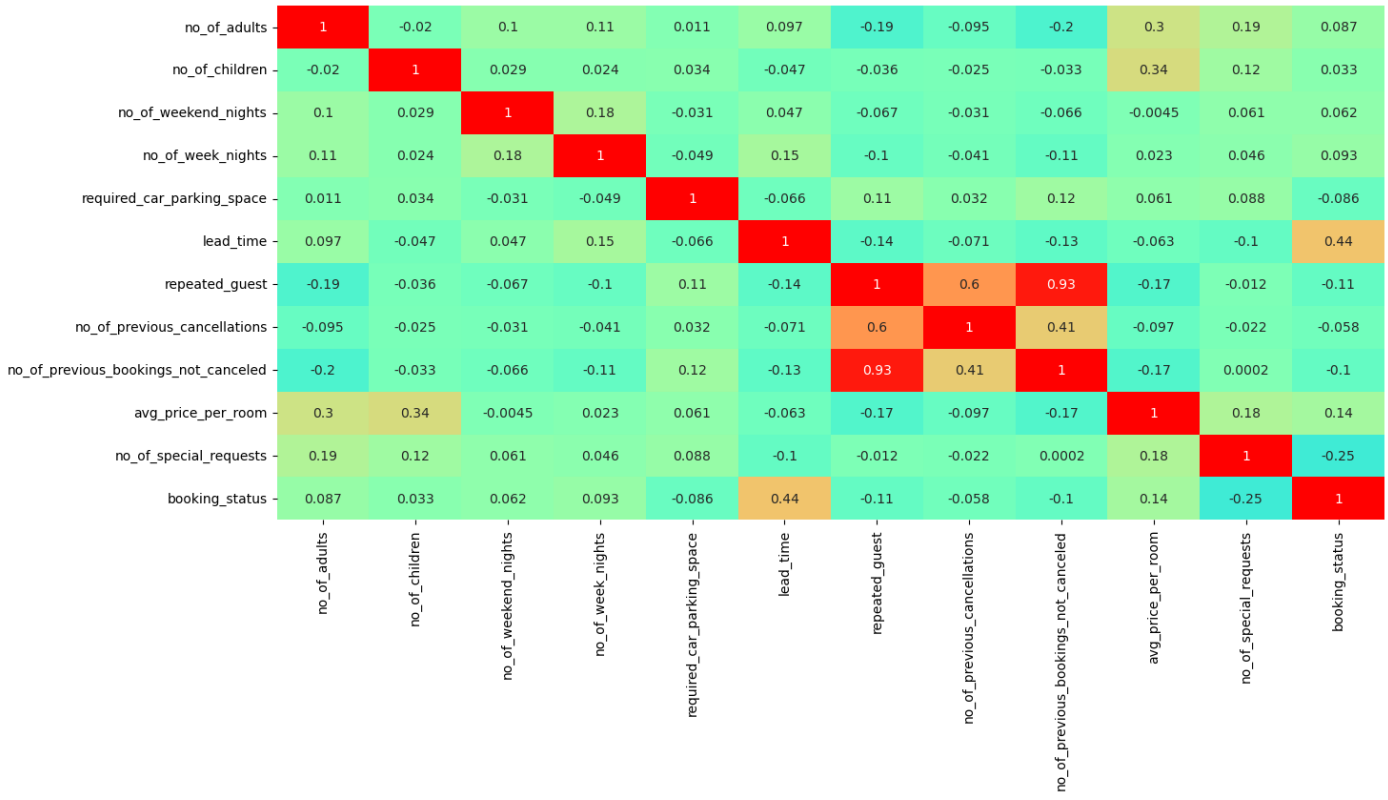

Figure 19 Correlation between the numeric columns

- *no_of_previous_bookings_not_canceled* displays a linear correlation with *repeated_guest* and *no_of_previous_cancellations*. It will be dropped to avoid multicollinearity.

- *Avg_price_per_room* shows a moderate positive correlation with *no_of_adults* and *no_of_children*.
- *Lead_time* has a moderate positive correlation with *booking_status*, which suggests that there are more cancellations for longer lead times. This would be an important parameter to analyze.
- *No_of_special_requests* has a weak negative correlation with *booking_status*. Guests who make special requests could be less likely to cancel rooms.

Since booking_status is our main field of interest, the effect of the categorical variables on this column needs to be checked.

Booking_status vs type_of_meal_plan: The ratio of cancelled to not cancelled is a little bit higher for Meal Plan 2. The effect of the different meal plans will be examined further.

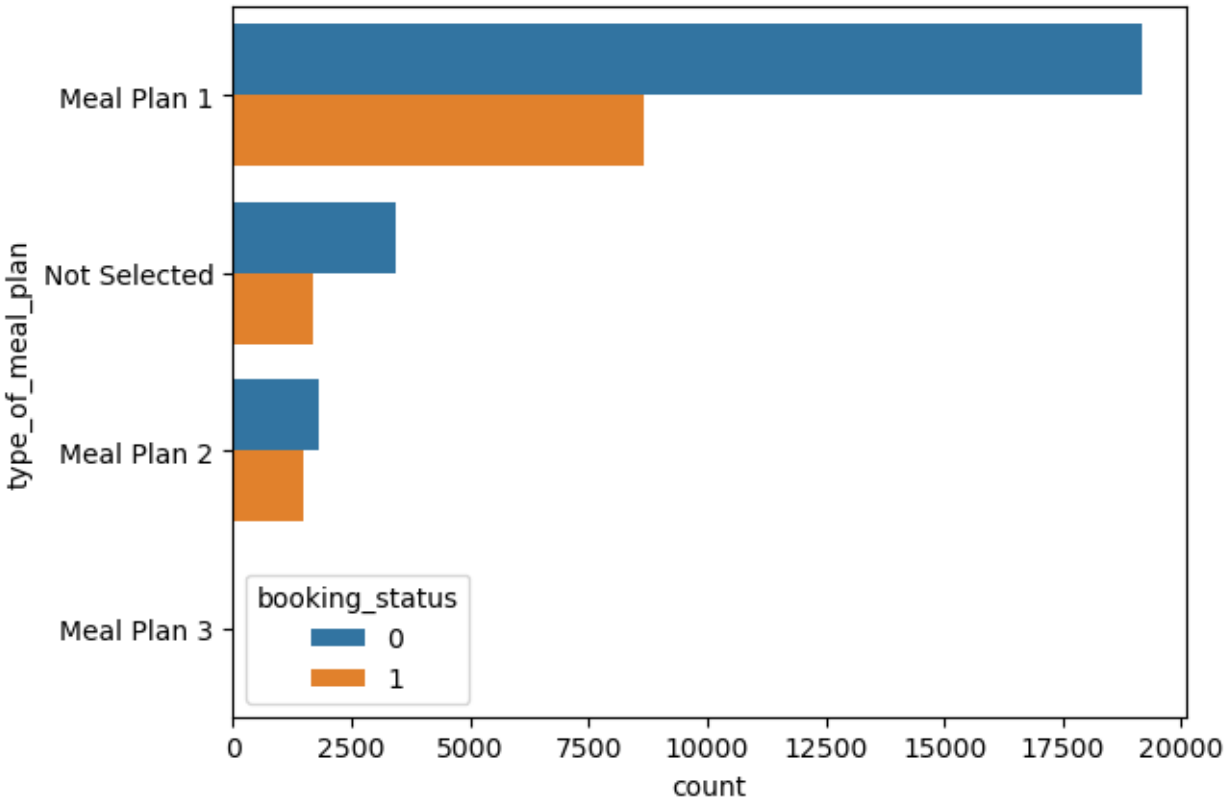

Figure 20 Booking status vs type of meal plan

Booking_status vs room_type: The bar chart does not reveal anything significant visually. A more detailed analysis is required to check if the room type affects cancellations.

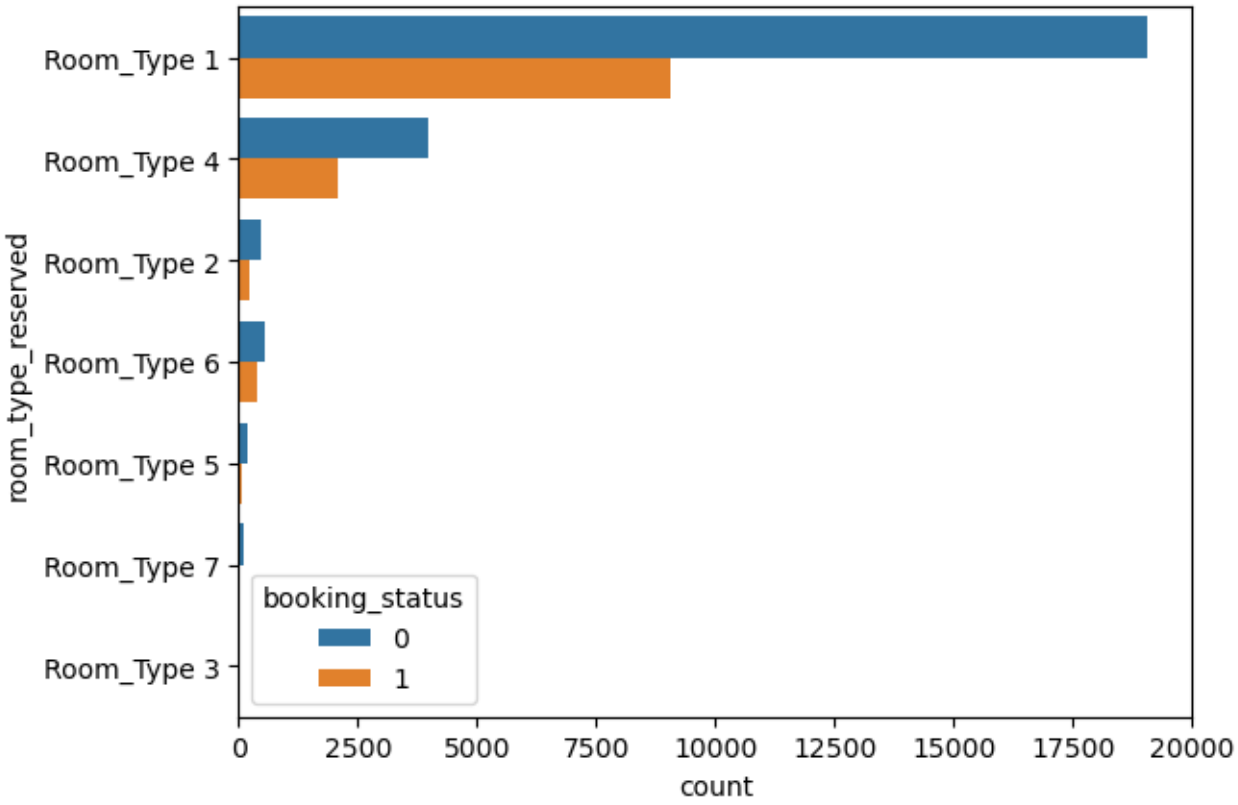

Figure 21 Booking status vs room type reserved

Booking_status vs arrival_year: The cancellations were much lower in 2017. However, the amount of data available for that year was also low. The change in the cancellation ratio between years should be included in the analysis.

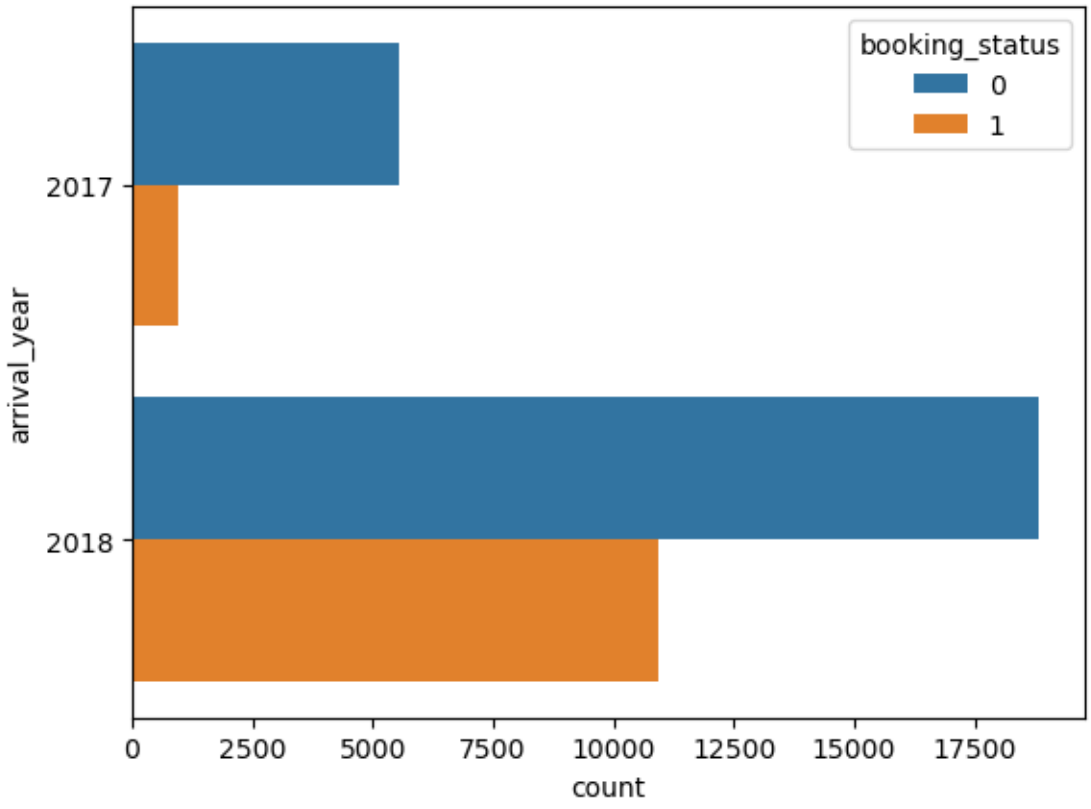

Figure 22 Booking status vs arrival year

Booking_status vs arrival_month: There are some months like January which show a low cancellation to non-cancellation ratio. The significance of month in terms of cancellations needs to be studied.

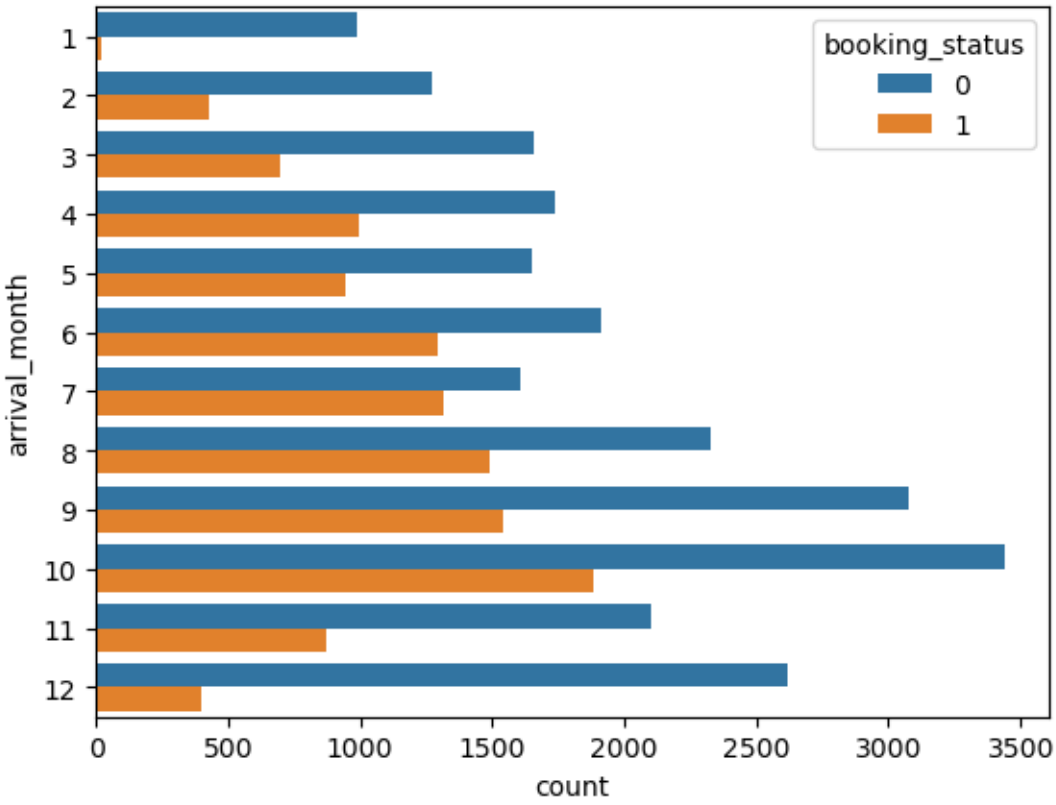

Figure 23 Booking status vs arrival month

Booking_status vs market_segment_type: The more common segments, Online and Offline, have a high cancellation to non-cancellation ratio. There are categories like complementary and aviation which have almost 0 cancellations. The effect of the market_segment_type will also be important in the analysis.

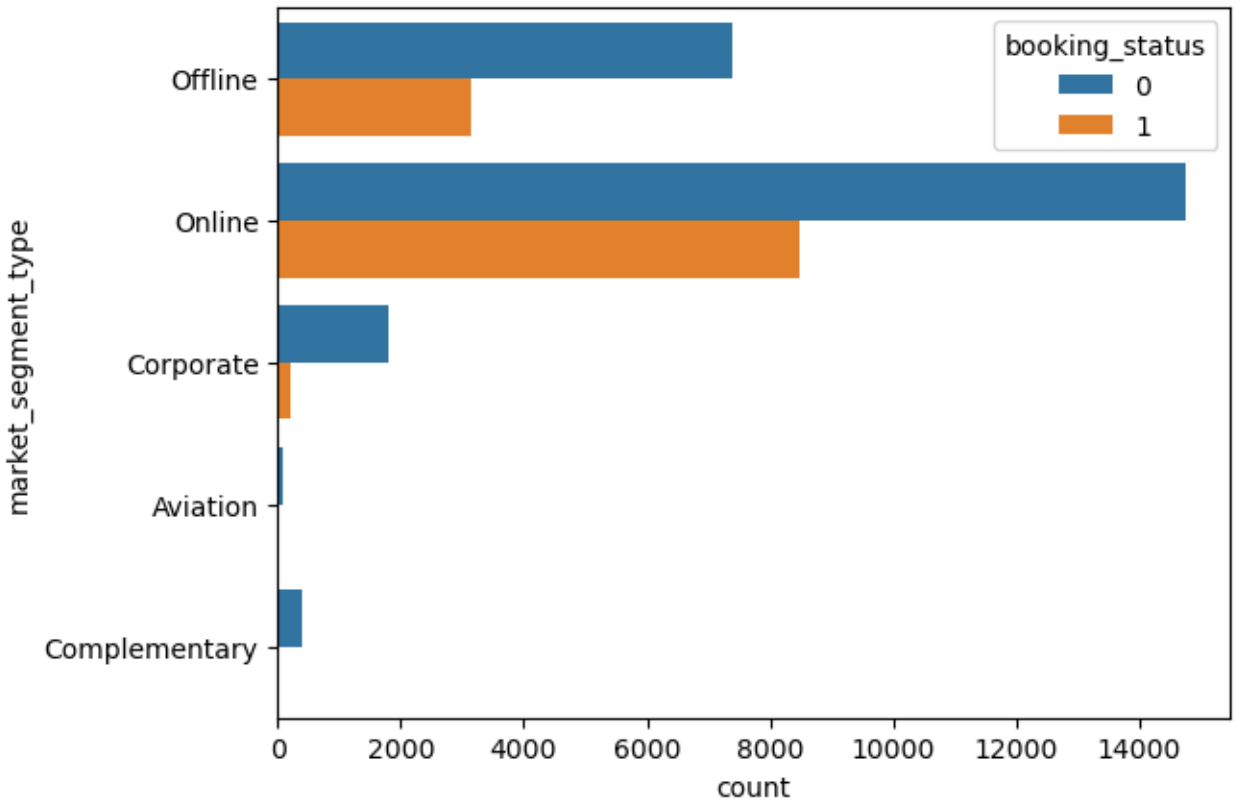

Figure 24 Booking status vs market segment type

2.4 Patterns and Insights

2.5 Answers to the EDA Questions

1. What are the busiest months in the hotel?

arrival_month	1	2	3	4	5	6	7	8	9	10	11	12
arrival_year												
2017	0	0	0	0	0	0	363	1014	1649	1913	647	928
2018	1014	1704	2358	2736	2598	3203	2557	2799	2962	3404	2333	2093

Figure 25 Number of guests visiting the hotel each month

Since there is a lot of missing data for the year 2017, the busiest months are decided based on the year 2018. October has the maximum number of guests, followed by June and September.

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

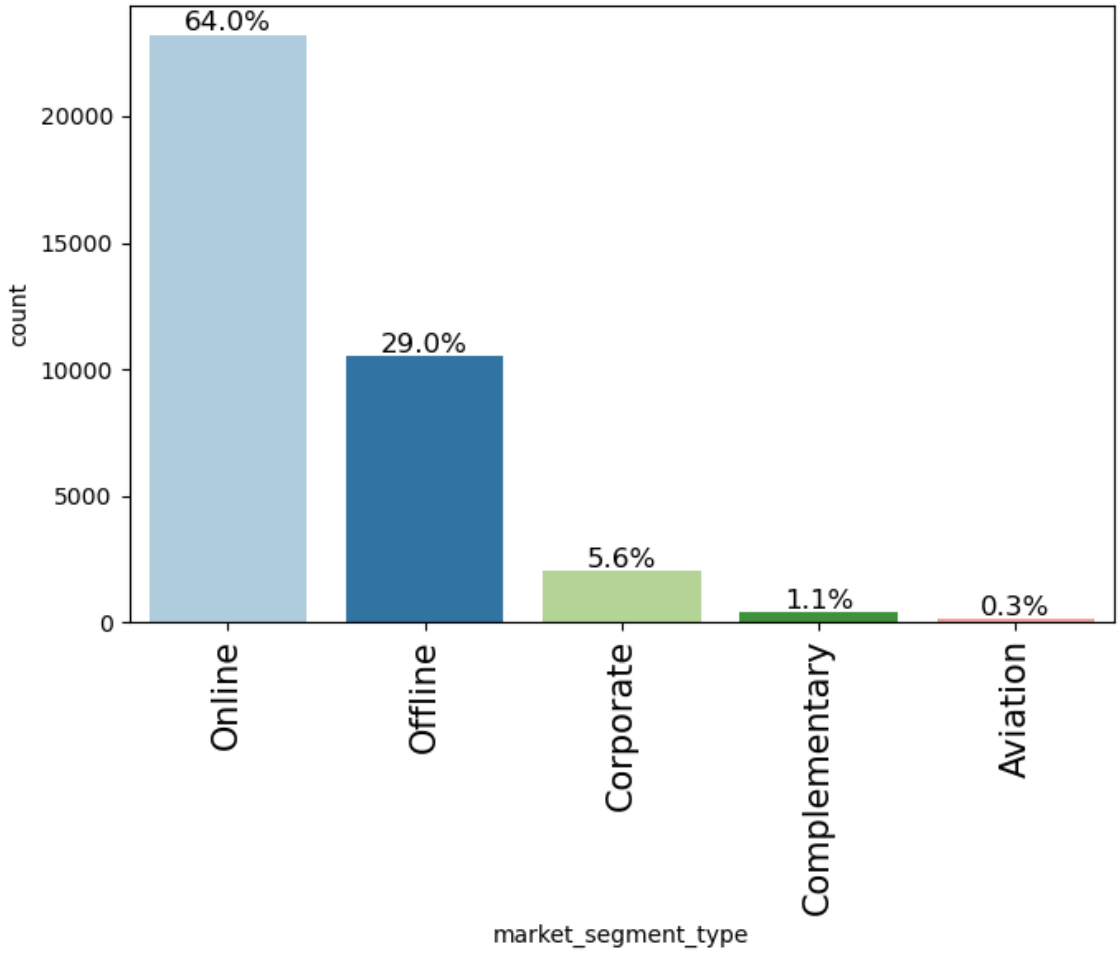

Figure 26 Market segment for guests

Most of the guests (64%) come from the **Online** market segment.

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

```
market_segment_type
Aviation          100.70
Complementary       3.14
Corporate          82.91
Offline            91.63
Online            112.26
Name: avg_price_per_room, dtype: float64
```

Figure 27 Average price per room for different market segment type

The price per room is maximum in the **Online** market segment at 112.26 euros. In the complementary category, it is close to 0 euros!

4. What percentage of bookings are cancelled?

   The percentage of bookings that were cancelled comes to **32.76%**

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

   The percentage of cancellations for repeated guests is **1.72%**

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

no_of_special_requests	0	1	2	3	4	5
booking_status						
0	0.57	0.76	0.85	1.0	1.0	1.0
1	0.43	0.24	0.15	0.0	0.0	0.0

Figure 28 Booking status for different number of special requests

Booking_status of 1, indicates that the booking is cancelled. The values have been normalized to sum to 1 along the columns. For example, when there are no special requests, the cancellation is 0.43, or 43%.

Based on this table, it is evident that the percentage of cancellation reduces with an increase in the number of special requests.

2.6 Observations on individual variables and relationship between variables

- In the data, 67% of the bookings were *not_canceled* and 33% were *canceled.*
- *Lead_time* has a moderate positive correlation with *booking_status*, which suggests that there are more cancellations for longer lead times.
- The percentage of cancellation reduces with an increase in the number of special requests.
- The percentage of cancellation for repeated guests is **1.72%** as against the overall cancellation rate of **32.76%.** Guests who visit repeatedly are less likely to cancel.

3 Data Preprocessing

There are 36275 rows of data with 19 columns. The columns have been described in the data description section.

3.1 Missing value treatment

There are no missing values in 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 29 Information on columns for the data

3.2 Outlier Detection and Treatment

Logistic regression model: Logistic regression is sensitive to outliers. Treating outliers could help in building a more accurate model.

no_of_previous_cancellations and *no_of_previous_bookings_not_canceled* have a lot of outliers. For both columns, all values above 1 could be set to 1, since only a small percentage of values lie in that zone. This would not affect the study of cancellations based on these features.

Decision Tree classifier: This model is robust to outliers. Outlier treatment is not necessary.

3.3 Feature Engineering

Logistic regression:

- Drop *Booking_ID* and *arrival_date*.

- Convert *arrival_month, arrival_date* to a category column.
- Drop *no_of_previous_bookings_not_canceled* since it is correlated to multiple other columns.
- Perform one-hot encoding for *arrival_month, arrival_date*, *type_of_meal_plan*, *room_type_reserved* and *market_segment_type*. One category from each of these column needs to be dropped. For improved interpretability, the following values have been dropped - **'type_of_meal_plan_Meal Plan 1', 'room_type_reserved_Room_Type 1', 'arrival_year_2018', 'arrival_month_10', 'market_segment_type_Online'** as they could be used as references.
- In *booking_status* column, set 'Canceled' to 1 and 'Not_Canceled' to 0. Since 'Canceled' is our class of interest, its value is set to 1.

Decision Tree Classifier:

- Drop *Booking_ID* and *arrival_date*.
- Perform one-hot encoding for *type_of_meal_plan*, *room_type_reserved* and *market_segment_type*.
- In *booking_status* column, set 'Canceled' to 1 and 'Not_Canceled' to 0. Since 'Canceled' is our class of interest, its value is set to 1.

3.4 Data Scaling

Data scaling has not been performed for any model.

3.5 Train-test split

In this step, the dependent variable column, *booking_status*, is separated out from the dataset. This will be called '**y**'. A second dataset is created with all the features except *booking_status*. This will be represented as '**X**'. In the X dataset, a new constant column is introduced and the whole dataset is converted to float. Both these datasets will be further split into training dataset and test dataset.

The training dataset will comprise of roughly 70% of the observations. This will be used for model building. The remaining 30% of the data will be classified as the test dataset. Test data is not seen by the model and it will be used to evaluate how good our model is.

4 Model Building

4.1 Metric to optimize for the problem

There are multiple metrics which can be used to optimize the model.

Accuracy: Based on how many cancellations and non-cancellations have been predicted correctly.

Recall: Based on correctly predicting cancellations, considering only the actual cancellations.

Precision: Based on correctly predicting cancellations, considering only the predicted cancellations.

F1 score: Computed using recall and precision as **2 * Precision * Recall / (Precision + Recall)**.

The goal of the case study is to correctly predict the cancellations and reduce the number of empty rooms. As a result, the focus should be on correctly identifying the actual cancellations and the **Recall** metric will be given higher priority. It must be noted that other metrics will not be compromised because predicting every booking as cancelled will give a 100% recall score! This approach would not be correct.

4.2 Logistic Regression

An initial logistic regression model will be built using the training dataset, that was explained in train-test split. Then, multicollinearity will be treated by studying the VIF followed by removal of insignificant features.

```
no_of_children                          2.096151
no_of_weekend_nights                    1.907268
no_of_week_nights                       3.502009
required_car_parking_space              1.068890
lead_time                               2.430130
repeated_guest                          1.982575
no_of_previous_cancellations            1.568832
no_of_special_requests                  1.781720
type_of_meal_plan_Meal Plan 2           1.310088
type_of_meal_plan_Meal Plan 3           1.008496
type_of_meal_plan_Not Selected          1.289154
room_type_reserved_Room_Type 2          1.109802
room_type_reserved_Room_Type 3          1.004282
room_type_reserved_Room_Type 4          1.359355
room_type_reserved_Room_Type 5          1.026708
room_type_reserved_Room_Type 6          1.886556
room_type_reserved_Room_Type 7          1.057261
arrival_year_2017                       1.495476
arrival_month_1                         1.101181
arrival_month_2                         1.175287
arrival_month_3                         1.249177
arrival_month_4                         1.246491
arrival_month_5                         1.259835
arrival_month_6                         1.321171
arrival_month_7                         1.360260
arrival_month_8                         1.455980
arrival_month_9                         1.524074
arrival_month_11                        1.324333
arrival_month_12                        1.359692
market_segment_type_Aviation            1.020015
market_segment_type_Complementary       1.114732
market_segment_type_Corporate           1.446544
market_segment_type_Offline             2.008442
dtype: float64
```

Figure 30 VIF values for different features in the model

The VIF values for all the features are shown in the figure. From this list, if there are any VIF values above 5, the column with the maximum VIF will be dropped. A new set of VIF will be calculated after dropping the high VIF column. This process is repeated until all the VIF values are below 5.

Removal of insignificant features is based on the p-value from the logistic regression summary. If there are any features with p-value greater than 0.05, the feature with the highest p-value will be dropped. A new model will be created with the updated features and the p-values will be checked again. An iterative process will be followed until all the columns have a p-value less than 0.05. The final model summary is shown.

Table 1 Logistic regression summary

Logit Regression Results

Dep. Variable:	booking_status	**No. Observations:**	25392
Model:	Logit	**Df Residuals:**	25369
Method:	MLE	**Df Model:**	22
Date:	Fri, 21 Mar 2025	**Pseudo R-squ.:**	0.3321
Time:	12:16:57	**Log-Likelihood:**	-10727.
converged:	True	**LL-Null:**	-16060.
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-0.5446	0.047	-11.522	0.000	-0.637	-0.452
no_of_children	0.3269	0.057	5.754	0.000	0.216	0.438
no_of_weekend_nights	0.1191	0.020	6.010	0.000	0.080	0.158
no_of_week_nights	0.0372	0.012	2.998	0.003	0.013	0.062
required_car_parking_space	-1.3752	0.138	-9.997	0.000	-1.645	-1.106
lead_time	0.0143	0.000	55.990	0.000	0.014	0.015
repeated_guest	-2.4249	0.389	-6.238	0.000	-3.187	-1.663
no_of_special_requests	-1.4939	0.030	-49.239	0.000	-1.553	-1.434
type_of_meal_plan_Meal Plan 2	0.6485	0.064	10.090	0.000	0.523	0.775

room_type_reserved_Room_Type 2	-0.5640	0.135	-4.179	0.000	-0.829	-0.299
room_type_reserved_Room_Type 4	0.1442	0.046	3.106	0.002	0.053	0.235
room_type_reserved_Room_Type 6	0.2790	0.141	1.979	0.048	0.003	0.555
arrival_year_2017	-0.9246	0.056	-16.389	0.000	-1.035	-0.814
arrival_month_1	-3.0751	0.275	-11.190	0.000	-3.614	-2.537
arrival_month_3	-0.1642	0.067	-2.460	0.014	-0.295	-0.033
arrival_month_4	-0.2499	0.061	-4.080	0.000	-0.370	-0.130
arrival_month_5	-0.3281	0.065	-5.064	0.000	-0.455	-0.201
arrival_month_7	-0.1360	0.062	-2.209	0.027	-0.257	-0.015
arrival_month_8	-0.2034	0.059	-3.470	0.001	-0.318	-0.089
arrival_month_12	-1.8107	0.088	-20.478	0.000	-1.984	-1.637
market_segment_type_Aviation	-0.5682	0.254	-2.237	0.025	-1.066	-0.070
market_segment_type_Corporate	-1.1682	0.101	-11.601	0.000	-1.366	-0.971
market_segment_type_Offline	-1.9856	0.050	-39.667	0.000	-2.084	-1.888

- Interpreting the coefficients is different for categorical and numerical columns.
- For numeric columns, a positive value represents a positive correlation and a negative value represents a negative correlation. The actual value represents the strength of the relationship.
- For example, the coefficient of -2.4249 for *repeated_guest* denotes that a repeated guest is very less likely to cancel the booking, while the value of 0.3269 for *no_of_children* shows that there is more likelihood of cancellation as the number of children increases.
- Interpreting categorical features is based on the reference category. The references chosen are **'type_of_meal_plan_Meal Plan 1', 'room_type_reserved_Room_Type 1', 'arrival_year_2018', 'arrival_month_10', 'market_segment_type_Online'.**
- For example, Room_Type_2 has a coefficient of -0.564. A customer who has booked Room_Type_2 is less likely to cancel than someone who has booked Room_Type_1. For Room_Type_6, the coefficient of 0.279 indicates that the probability of cancellation is higher for Room_Type_6 compared to Room_Type_1.
- Other coefficients can be interpreted in a similar way.
- Booking_status output will be a value ranging from 0 to 1. Depending on the threshold set, a prediction for cancellation will be obtained.

4.3 Decision Tree Classifier

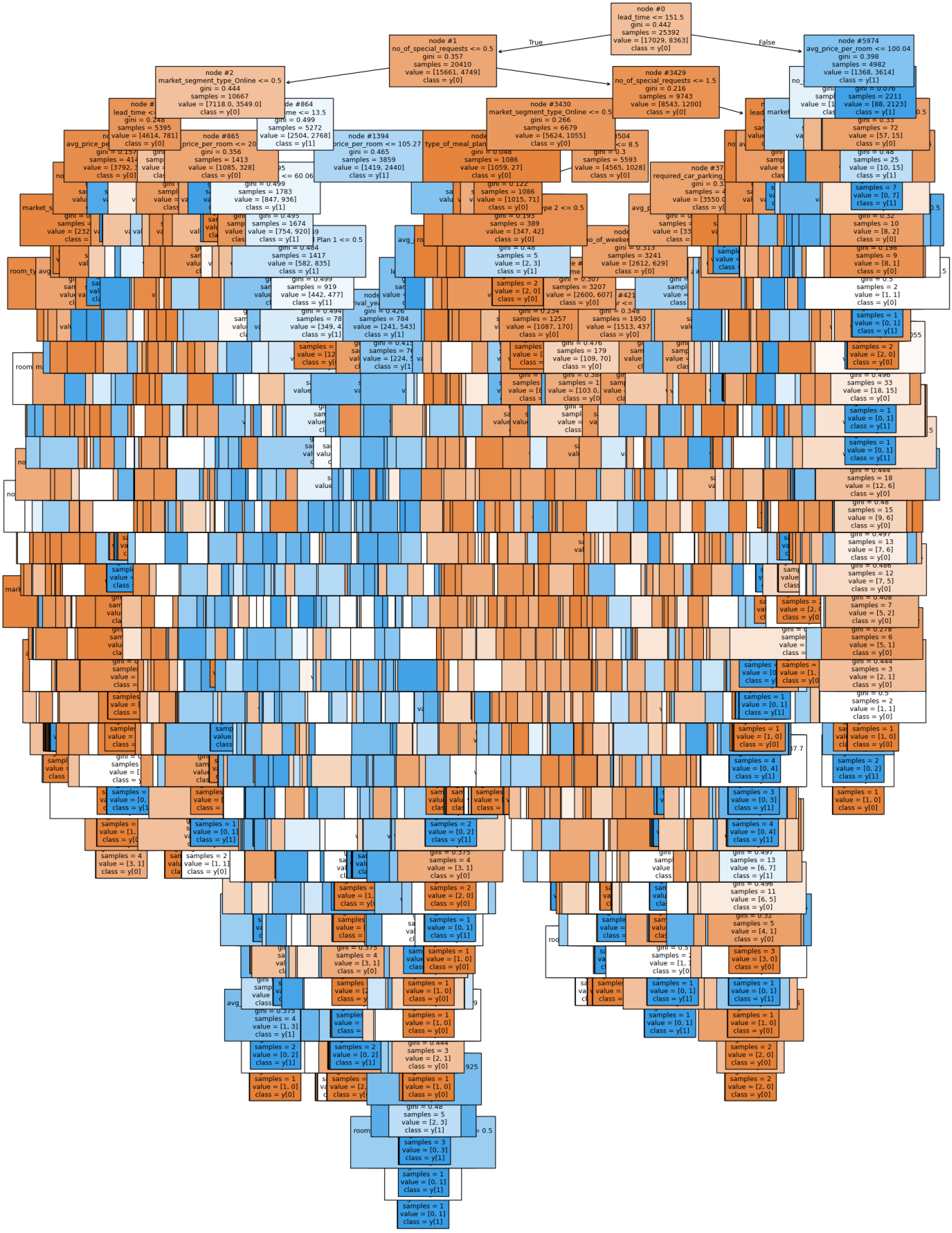

Figure 31 Decision Tree Max Depth

- The decision tree classifier constructs a sequence of ‘Yes’ or ‘No’ questions based on the different features to come up with predictions.
- The default tree continues the sequence until a very high accuracy is reached.
- A high accuracy on the training data generally results in a lower accuracy in the test data because the noise the data is also captured.
- Pre-pruning and post-pruning will be performed. The better model among the two will be chosen.
- The decision tree algorithm comes up with a table which highlights the importance of different features. The order of importance is shown in the figure.

	Importance
lead_time	0.377554
avg_price_per_room	0.191108
market_segment_type_Online	0.094015
arrival_month	0.080991
no_of_special_requests	0.068257
no_of_week_nights	0.057794
no_of_weekend_nights	0.038489
no_of_adults	0.029264
arrival_year	0.013218
type_of_meal_plan_Meal Plan 1	0.007499
required_car_parking_space	0.007060
room_type_reserved_Room_Type 1	0.006347
room_type_reserved_Room_Type 4	0.006305
no_of_children	0.005927
type_of_meal_plan_Not Selected	0.003781
market_segment_type_Offline	0.002400
type_of_meal_plan_Meal Plan 2	0.002319
room_type_reserved_Room_Type 2	0.001913
room_type_reserved_Room_Type 5	0.001452
market_segment_type_Corporate	0.001135
room_type_reserved_Room_Type 6	0.001029
no_of_previous_bookings_not_canceled	0.000783
repeated_guest	0.000633
room_type_reserved_Room_Type 7	0.000352
market_segment_type_Aviation	0.000279
no_of_previous_cancellations	0.000091
room_type_reserved_Room_Type 3	0.000000
type_of_meal_plan_Meal Plan 3	0.000000
market_segment_type_Complementary	0.000000

Figure 32 Importance of features in the decision tree model

4.4 Model performance across different metrics

Logistic Regression: For the default threshold value of 0.5, the performance on the training and testing data is very similar for the logistic regression model. The accuracy is 0.8, recall is 0.62 for training and 0.61 for testing. The recall is a little low for this model. The model could be improved.

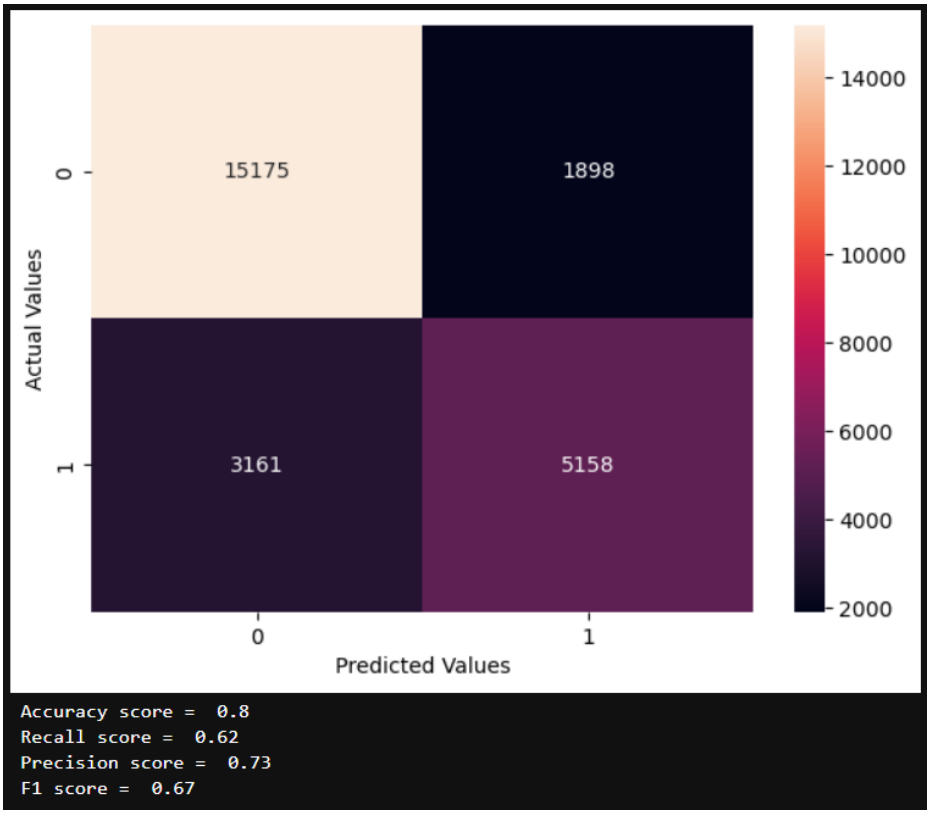

Figure 33 Performance on logistic regression training data

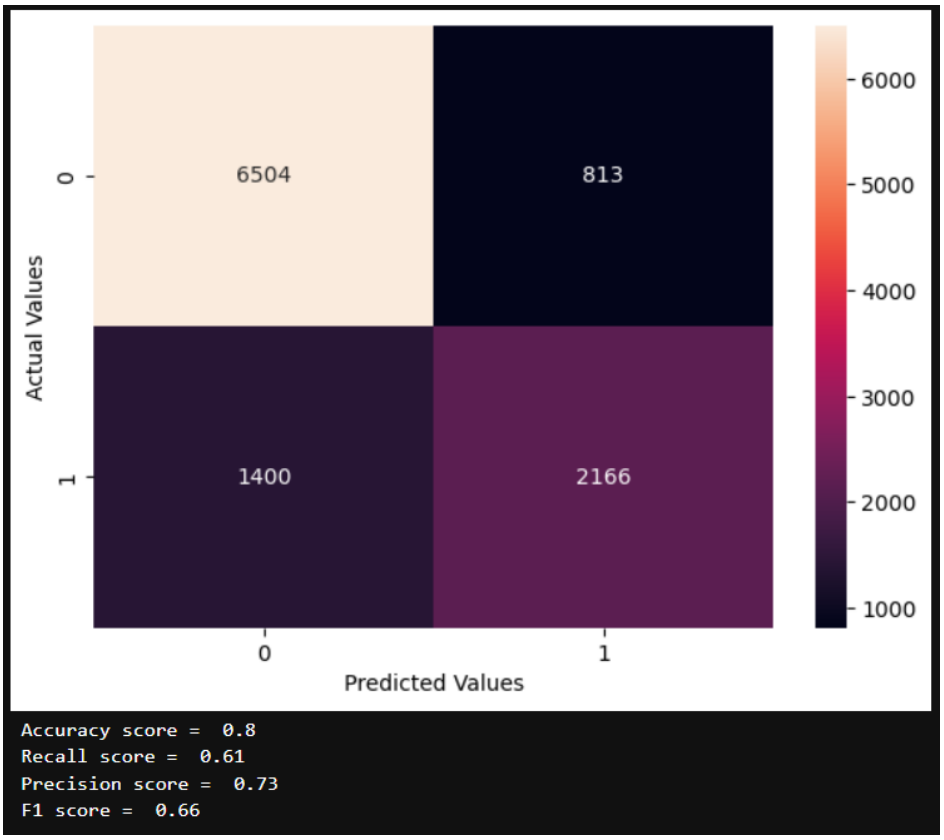

Figure 34 Performance on logistic regression testing data

Decision Tree Classifier: For the decision tree classifier model, there is a big difference between training and testing data. Accuracy drops from 0.99 to 0.87 and recall drops from 0.98 to 0.81 when switching from training to test data. The model captures a lot of noise.

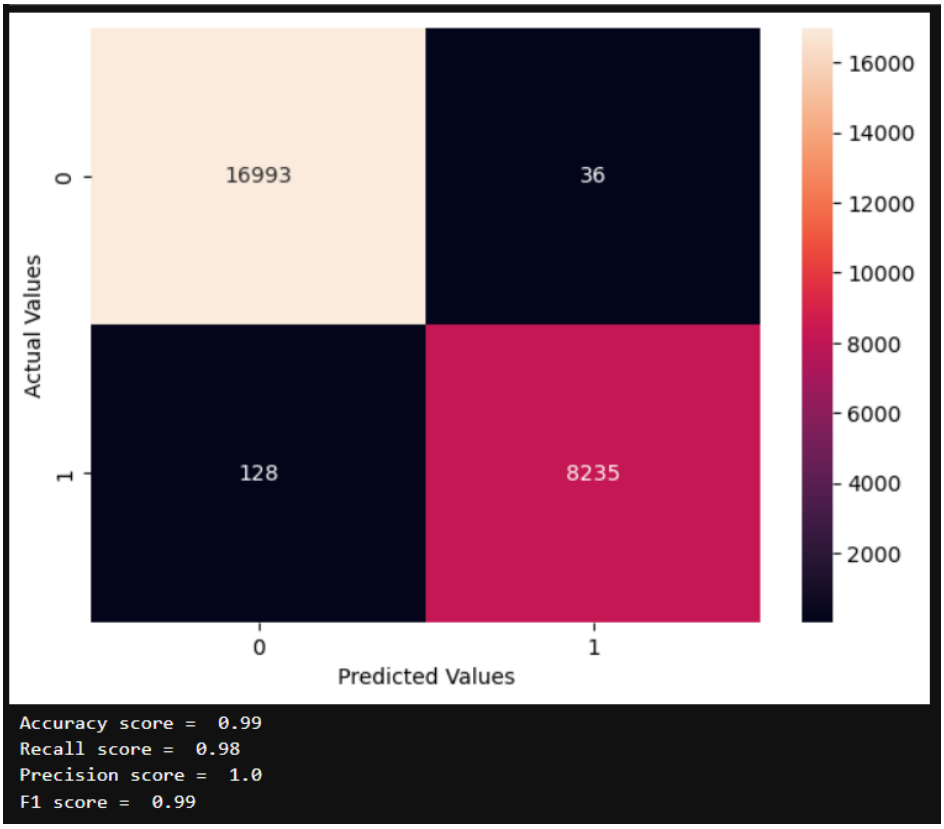

Figure 35 Performance on decision tree classifier training data

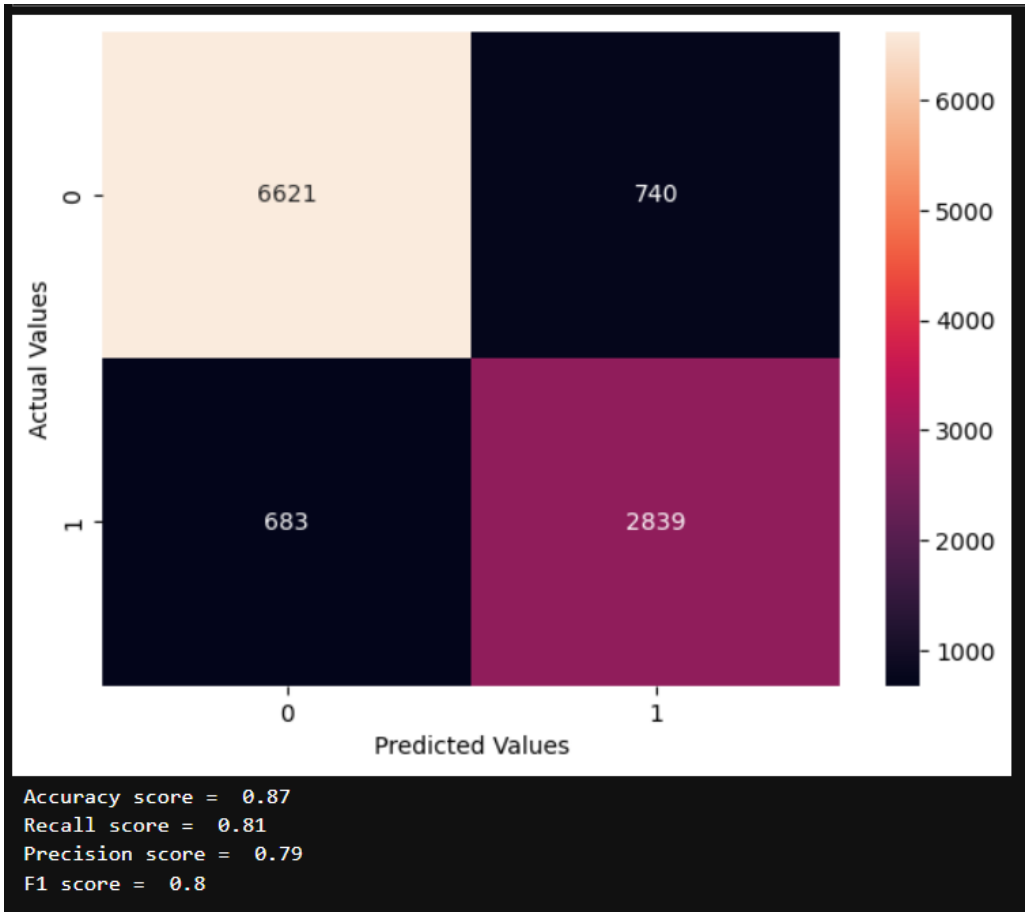

Figure 36 Performance on decision tree classifier testing data

5 Model Performance Improvement

For both logistic regression and decision tree classifier models, the performance can be improved by a couple of methods. They will be discussion in the sections below.

5.1 Logistic Regression performance tuning

The default model uses a threshold value of 0.5 to classify the bookings as cancelled or not cancelled. A different threshold value can be set using the values from the ROC-AUC curve or the Precision-Recall curve.

5.1.1 Optimization using ROC-AUC curve

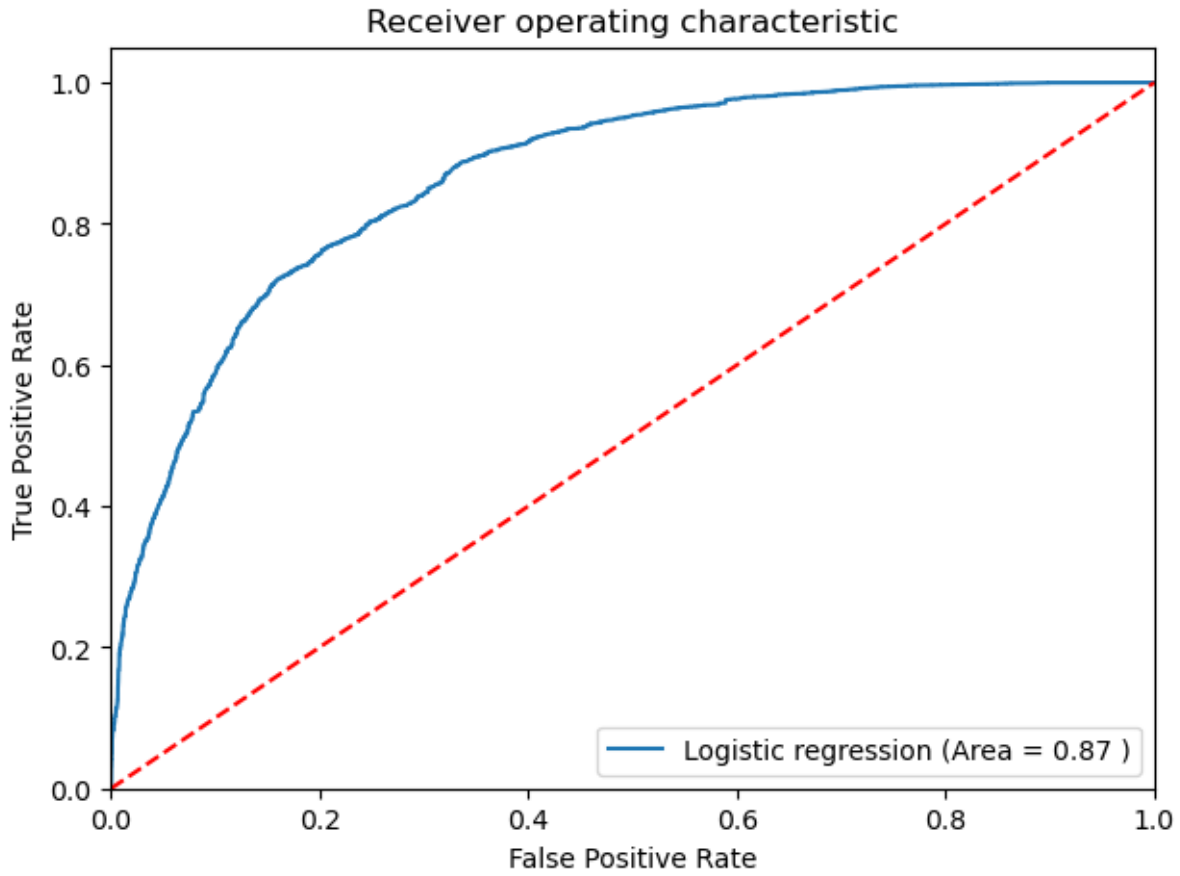

Figure 37 ROC AUC curve

- From the ROC-AUC curve, the optimal threshold value is the point of maximum difference between the true positive rate and the false positive rate.
- The threshold value turns out to be 0.406
- This threshold is used when predicting the cancellations on the training and testing data.
- On both training and testing data, the accuracy is 0.8 and recall is 0.72. This is a significant improvement over the model with default threshold of 0.5.
- Since the metrics are similar for train and test data, there is no underfitting or overfitting.

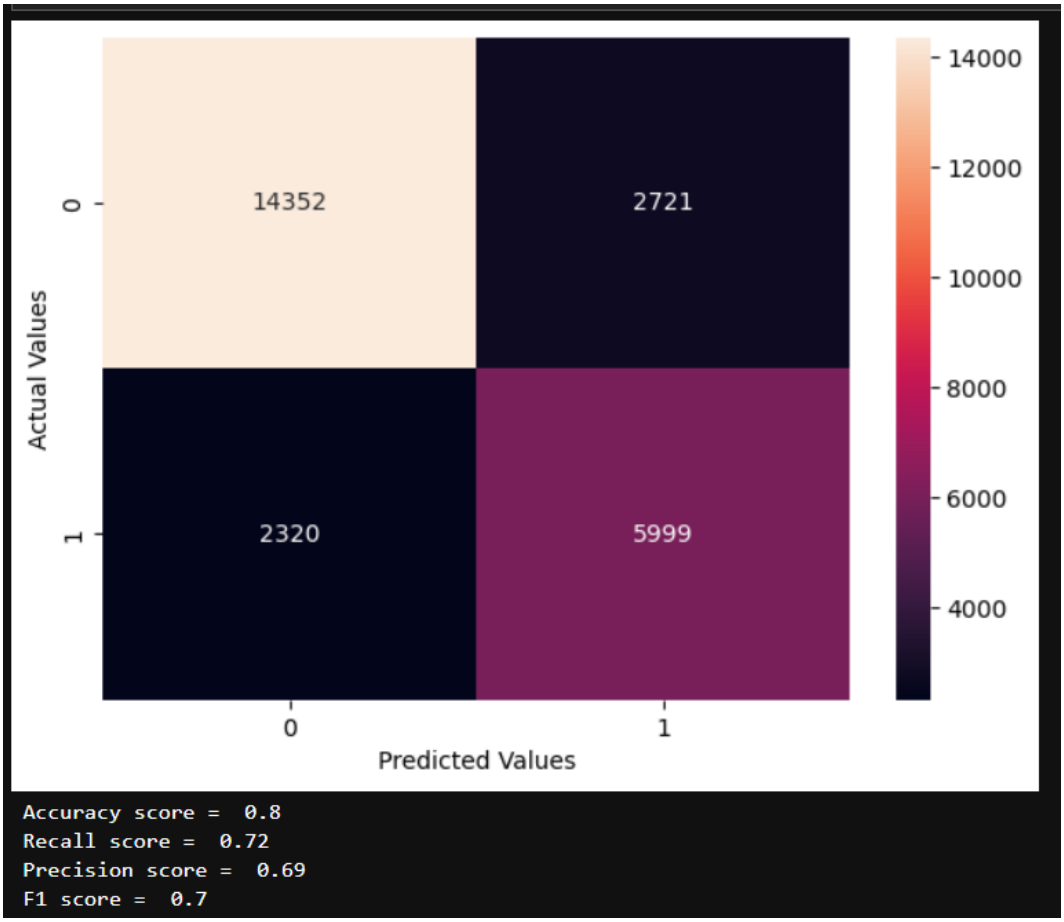

Figure 38 Performance metrics for optimal ROC-AUC threshold on training data

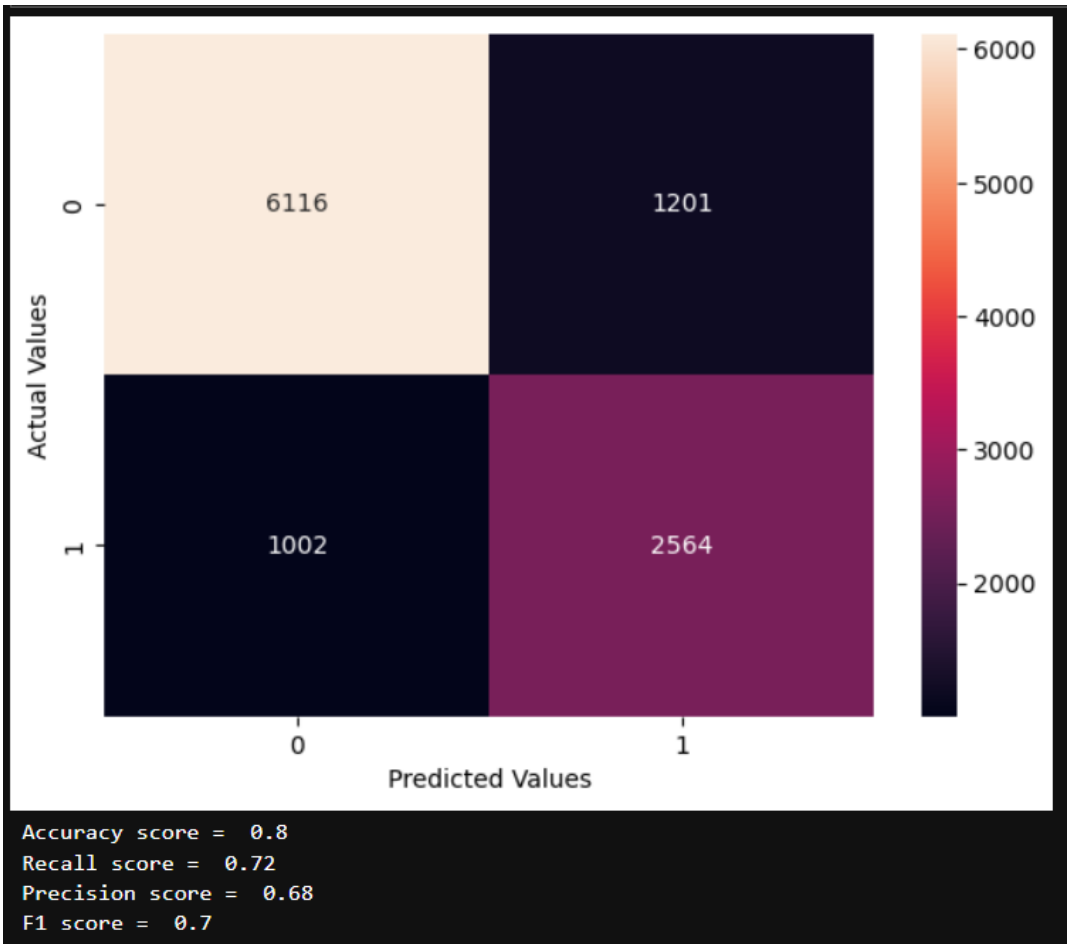

Figure 39 Performance metrics for optimal ROC-AUC threshold on testing data

5.1.2 Optimization using Precision-Recall curve

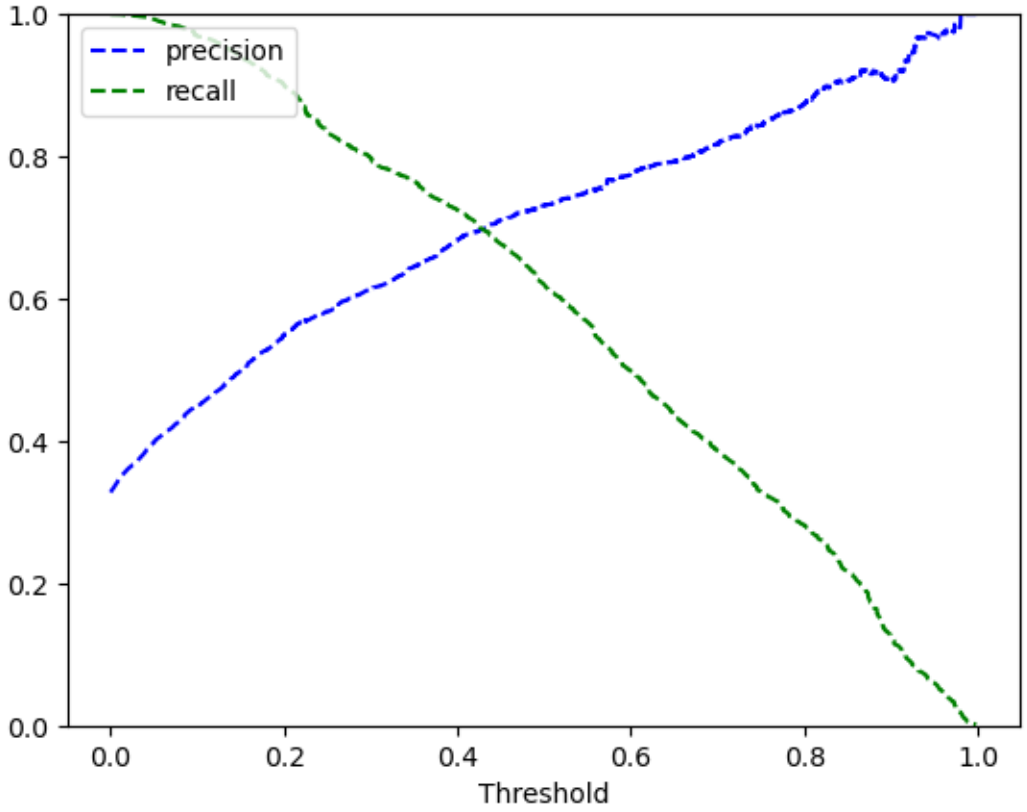

Figure 40 Precision Recall curve

- Another method to tune the performance is to use the intersection of the precision and recall curve as the threshold value.
- The threshold value turns out to be 0.4294.
- This threshold is used when predicting the cancellations on the training and testing data.
- On both training and testing data, the accuracy is 0.8 and while the recall is 0.7 on training data and 0.69 on testing data.
- This model also performs better than the one with default threshold of 0.5.
- Since the metrics are similar for train and test data, there is no underfitting or overfitting.

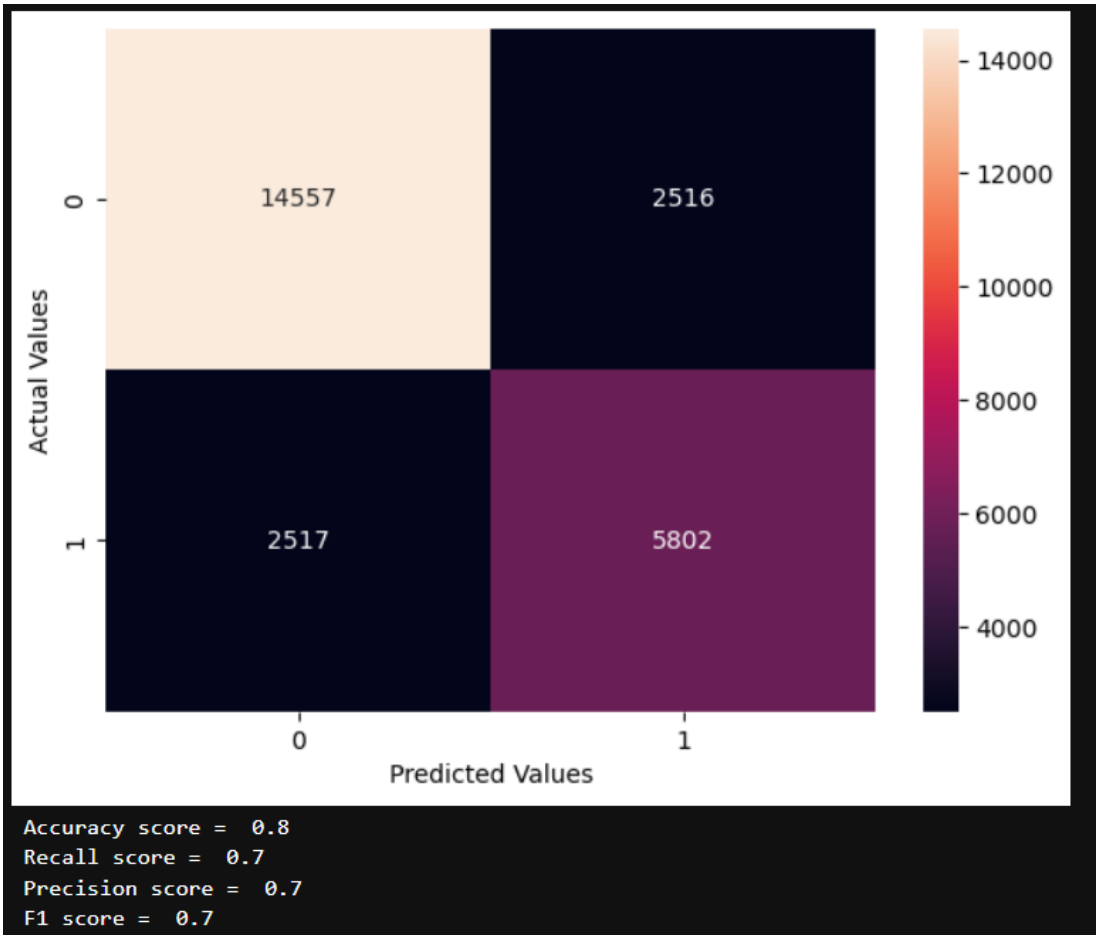

Figure 41 Performance metrics for optimal precision-recall threshold on training data

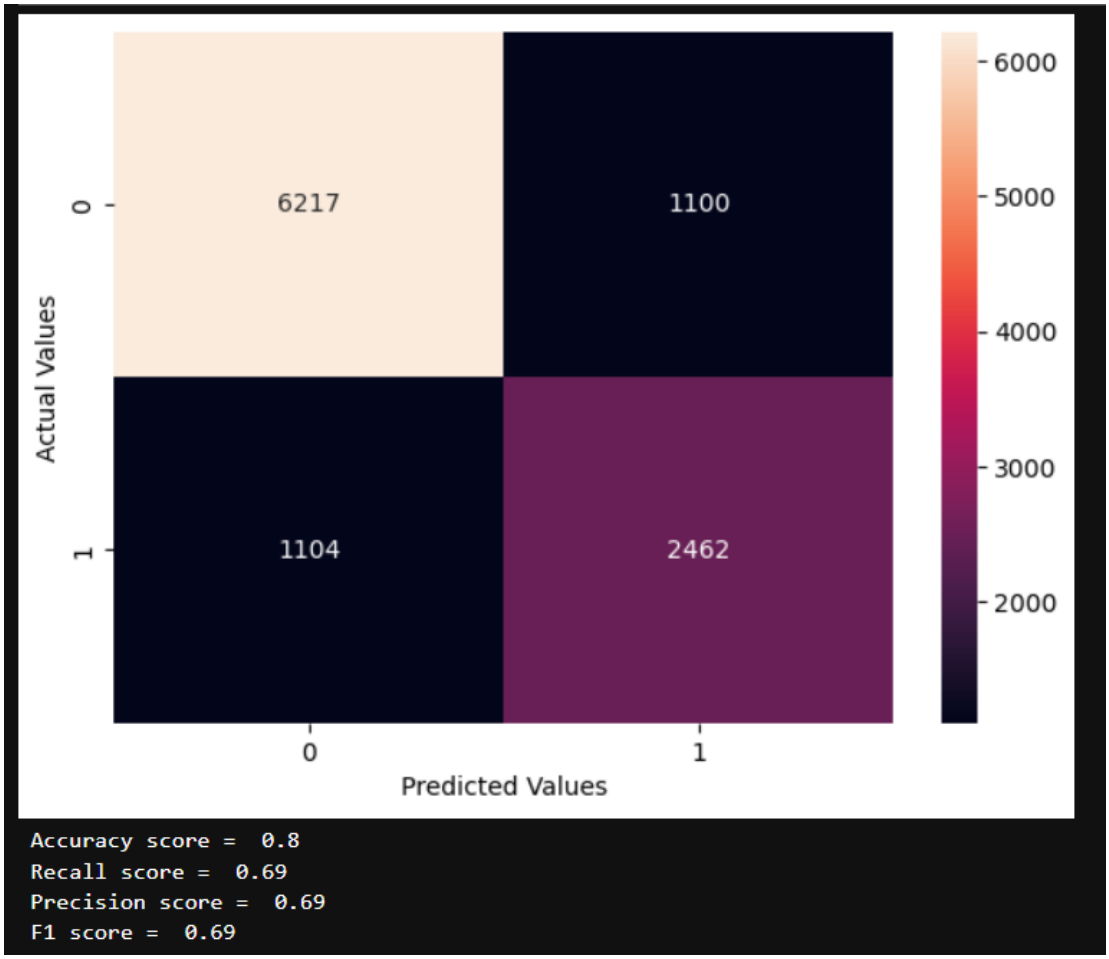

Figure 42 Performance metrics for optimal precision-recall threshold on testing data

5.2 Decision Tree Classifier performance tuning

5.2.1 Pre-pruning

- In pre-pruning, the model is not allowed to grow to the maximum depth.
- A bunch of hyperparameters are set to reduce overfitting.
- Max_depth is used to stop the growth of the tree after a certain point.
- With min_samples_leaf and max_leaf_nodes, the minimum number of samples required at a leaf node and the maximum number of leaf nodes in the tree are set.
- Further min_impurity_decrease is also tuned.
- The result is a much smaller tree.
- The pre-pruned tree has an accuracy of 0.77 on both the training and testing data.
- The recall score is 0.76 on the training data and 0.77 on the testing data.
- The model does not overfit or underfit the data.

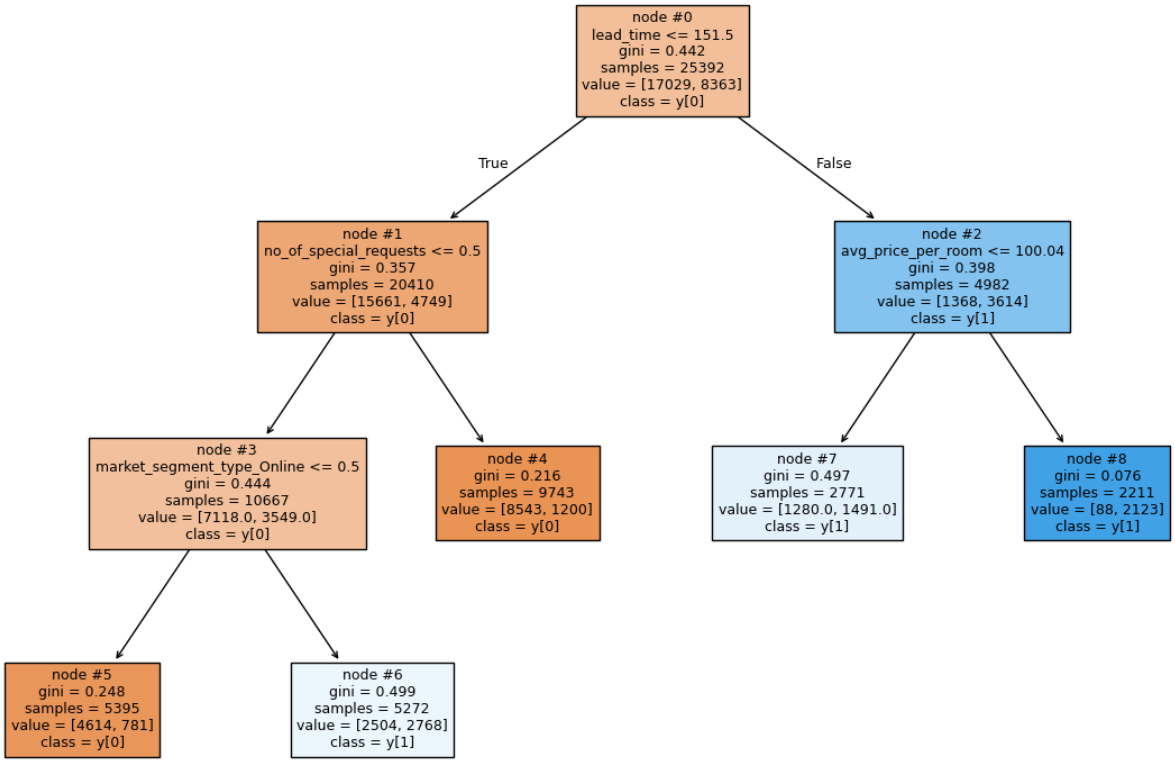

Figure 43 Decision tree classifier with pre pruning

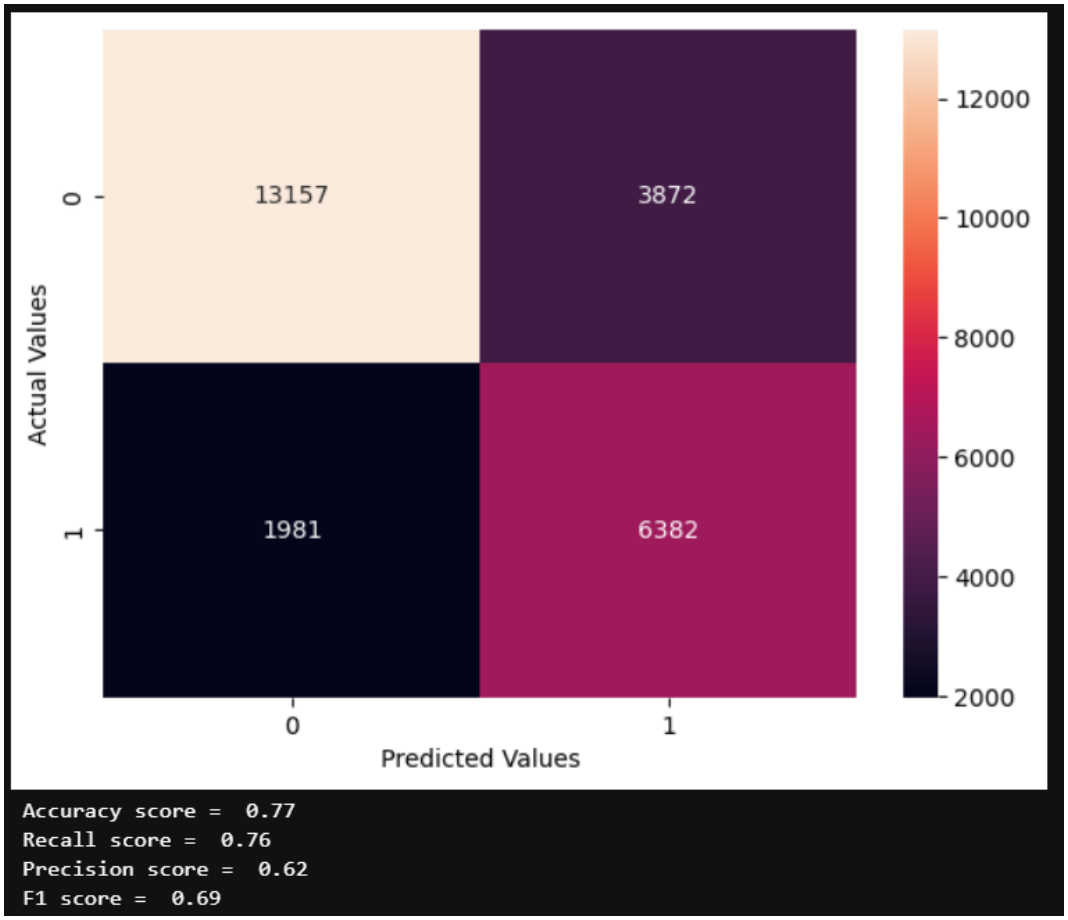

Figure 44 Performance metrics for pre-pruned decision tree on training data

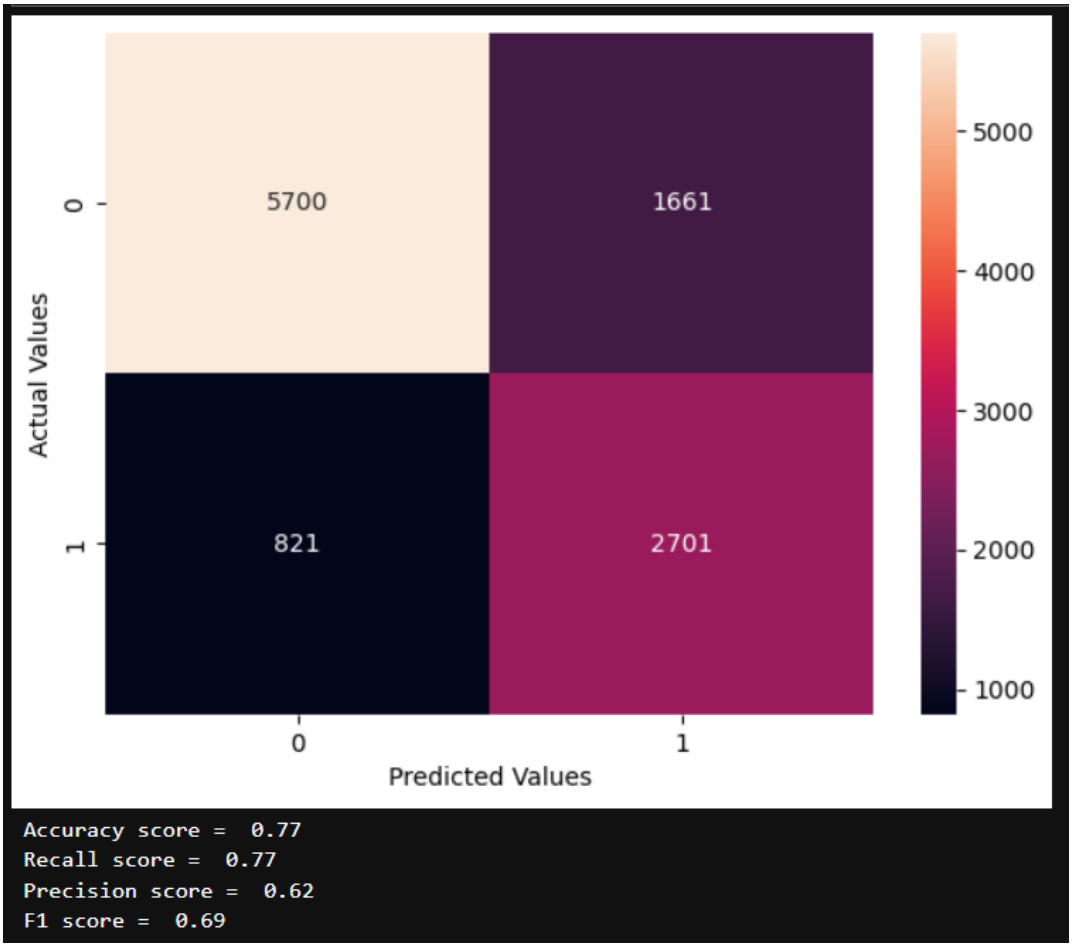

Figure 45 Performance metrics for pre-pruned decision tree on testing data

5.2.2 Post-pruning

- In pre-pruning, the model is allowed to grow to the maximum depth.
- A cost complexity path is created and the parameter alpha is calculated. Based on this path, insignificant branches are removed.
- The alpha value is plotted against one of the performance metrics (recall in this case) on the training and testing data.
- The model with the best recall score on the test data can be selected.
- The result is a much smaller tree.
- However, in this particular dataset, the decision tree from post-pruning is almost as complex as the max depth tree.

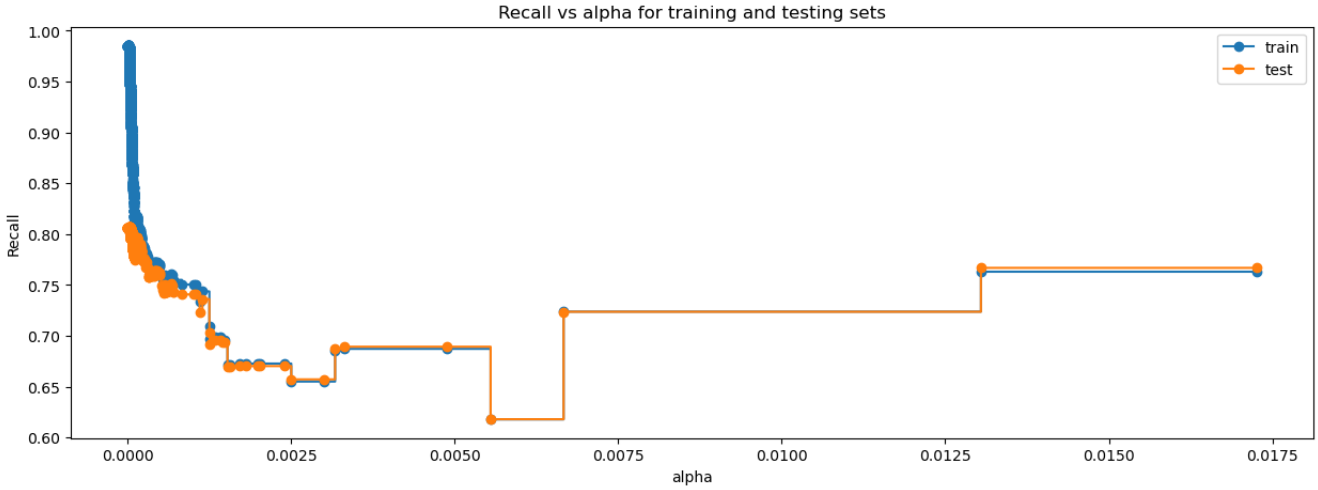

Figure 46 Recall vs alpha for training and testing data

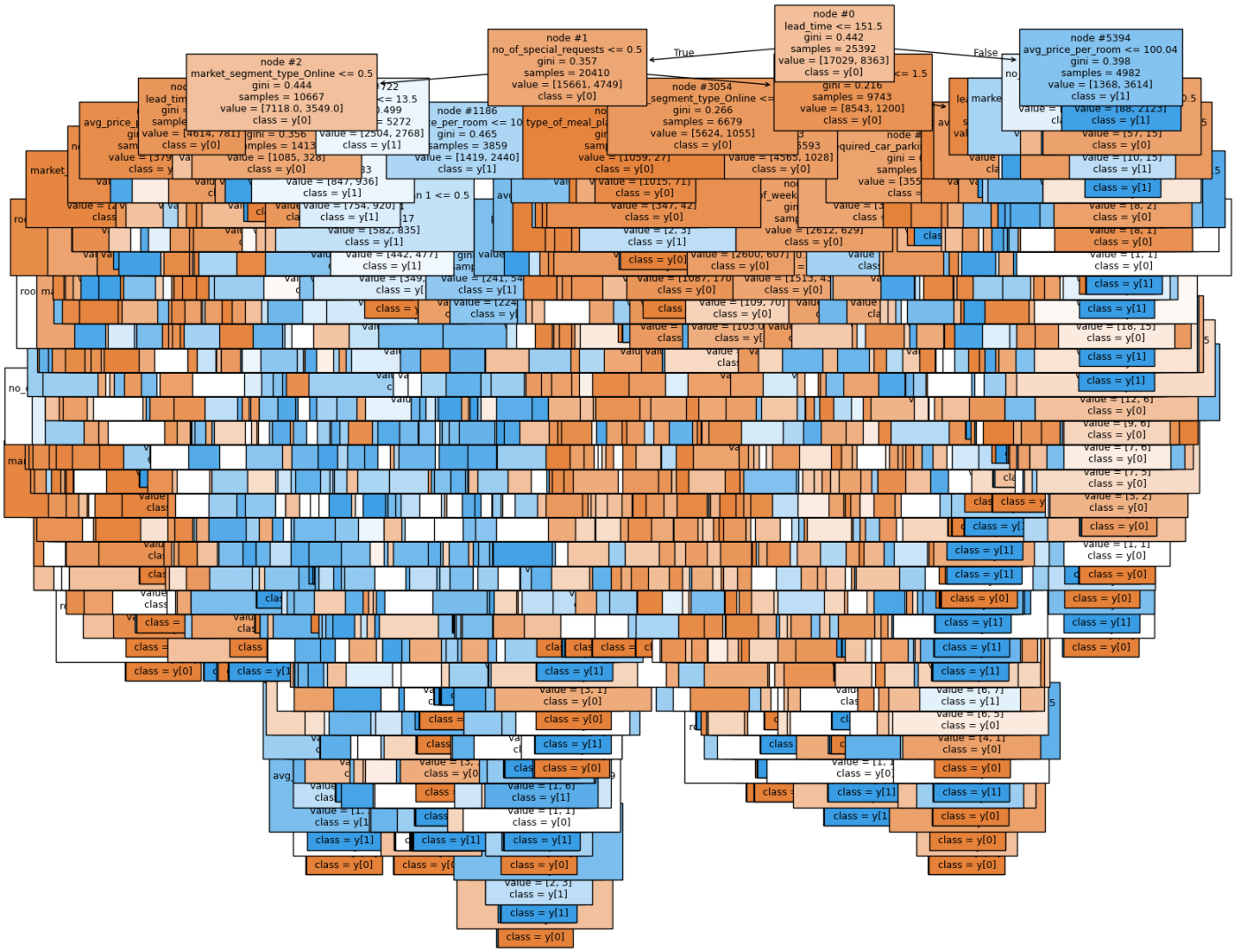

Figure 47 Decision tree classifier with post pruning

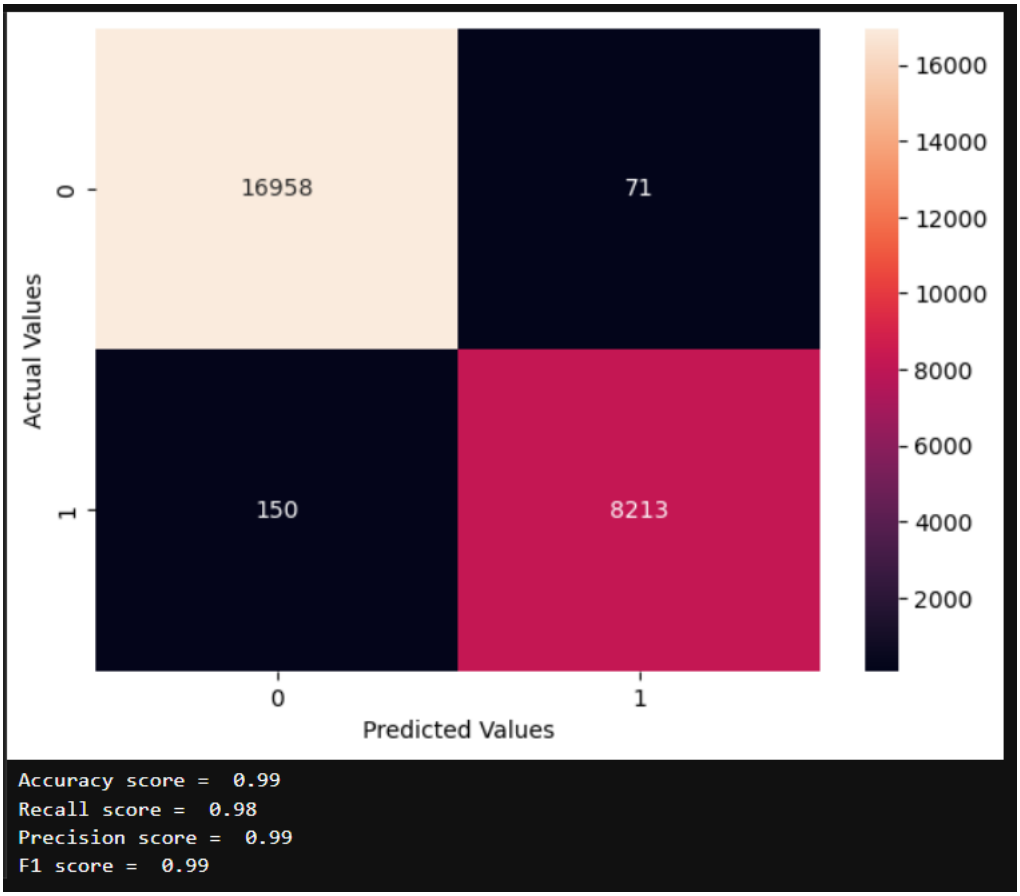

Figure 48 Performance metrics for post-pruned decision tree on training data

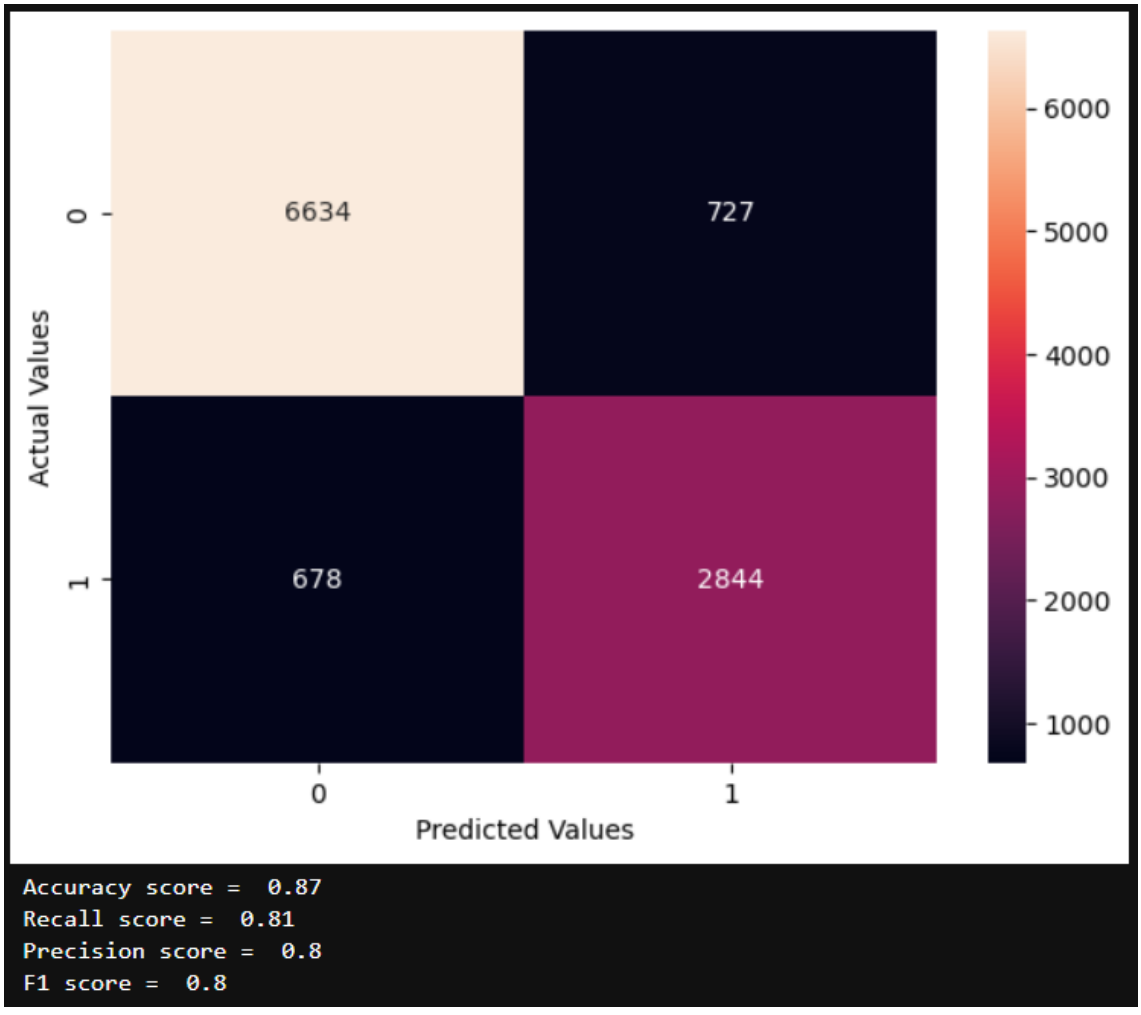

Figure 49 Performance metrics for post-pruned decision tree on testing data

6 Model Performance Comparison and Final Model Selection

- 6 models were created in total. Three using logistic regression and three using decision tree classifier.
- All the models were evaluated on the training and testing datasets.
- Performance metrics for logistic regression models:

	Accuracy	Recall	Precision	F1 score
Train_Default	0.8	0.62	0.73	0.67
Test_Default	0.8	0.61	0.73	0.66
Train_ROC	0.8	0.72	0.69	0.70
Test_ROC	0.8	0.72	0.68	0.70
Train_PR	0.8	0.70	0.70	0.70
Test_PR	0.8	0.69	0.69	0.69

Figure 50 Performance metrics for logistic regression models

- Performance metrics for decision tree models:

	Accuracy	Recall	Precision	F1 score
Train_Default	0.99	0.98	1.00	0.99
Test_Default	0.87	0.81	0.79	0.80
Train_Pre_Pruning	0.77	0.76	0.62	0.69
Test_Pre_Pruning	0.77	0.77	0.62	0.69
Train_Post_Pruning	0.99	0.98	0.99	0.99
Test_Post_Pruning	0.87	0.81	0.80	0.80

Figure 51 Performance metrics for decision tree classifier models

- Max depth decision tree performs best in training but not so well in testing. This model overfits the data and it will not be chosen.
- Post-pruned decision tree ended up almost as complex as the max depth decision tree and suffers from overfitting. This model will also be rejected.
- Pre-pruned decision tree has the best recall score of 0.77.
- The logistic regression models from the ROC-AUC method and the precision-recall method have a higher accuracy of 0.8.
- The precision-recall logistic regression model has a slightly better precision than the ROC-AUC model and a slightly lower recall.
- The F1 score for all the models are very closely matched.
- The decision tree model has a much lower precision compared to the logistic regression models.
- Since our performance metric of choice is recall, the pre-pruned decision tree classifier will be used as the final model.

Final model

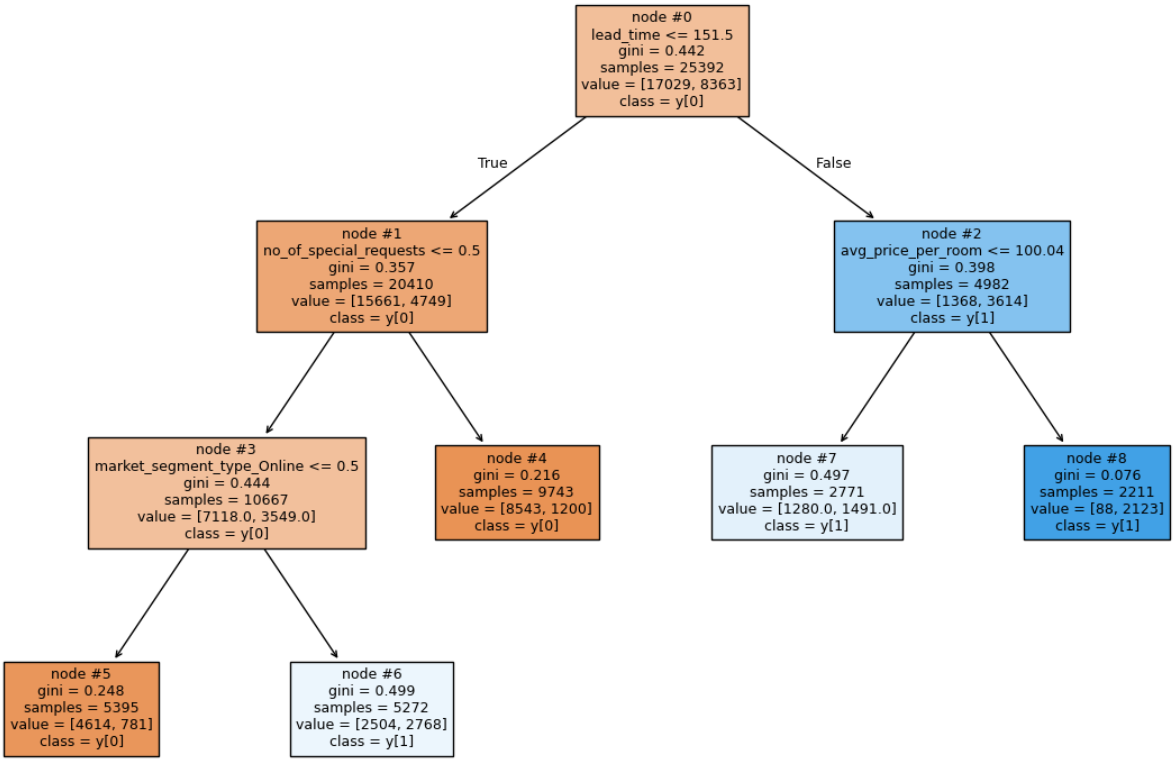

Figure 52 Final model

7 Actionable Insights & Recommendations

- A simple model based on decision tree has been created for predicting cancellations.
- The most important features are lead_time, market_segment_type_Online, no_of_special_requests and avg_price_per_room.
- A combination of lead_time > 151.5 days and avg_price_per_room > 100.04 euros, has 2123 cancellations in a total of 2211 records, which is ~96%. Early bird booking discounts or offering additional upgrades or perks can be considered for such guests.
- When lead_time <= 151.5 days and no_of_special_requests > 1, there are only 1200 cancellations in 9743 samples, which is ~12%. Low lead times and special requests have lower cancellations.
- In a scenario with lead_time <= 151.5 days and no special requests, market_segment_Online plays an important role. When the segment is not online, the cancellation is around 14.5%, but it shoots up to 52.5% when the segment is online. Establishing a connection with an online customer could help.
- Another strategy to improve profits is overbooking. For this hotel, the overall cancellation proportion of 33%. For this cancellation rate, when there are 130 bookings for 100 rooms, there is only a 1% probability of more than 100 customers turning up! A more precise model for optimal overbooking could be created, considering all factors like season, market segment of customer, room price, lead time etc.