INN Hotels – Project 4

Supervised Learning Classification

January 6, 2023

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Appendix
  - Data Background and Contents
  - Model Building
  - Model Performance Evaluation and Improvement

Executive Summary

- The model built can be used to predict the cancelations to within ~**80%**.
- The most important variables in predicting a cancelation are **(1) lead time**, **(2) online bookings** market segment, **(3) average price per room**, the **(4) number of special requests**. Lead time showed twice the importance.
- All these being equal, a cancellation is lest likely when (1) the lead time is shorter, (2) the booking is made other than online, (3) the price is lower, and (4) at least one special request is asked by the customer.
- As an attempt to offset last minute cancelations, ***INN*** hotels could ask for a small deposit for all online bookings that balloons closer to arrival time. Then at the point of cancelation, offer a discount as cancelations tend to happen at a higher average room rate.
- Regarding special requests, ***INN*** could market common special requests that could be available to guests. This would display higher value, quality, and hospitality. Also look to which customer segments ask for special requests and market to them (i.e. possibly corporate or special events).
- Look for opportunities to: increase repeat customers as only 4% of the repeat guests cancel; increase corporate customers as they cancel less; meet higher demand with higher prices in more months, October in particular.

Business Problem Overview and Solution Approach

Context of Business Problem

- **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 **online booking channels** have dramatically changed **customers'** booking possibilities and **behavior**.
- The **cancellation of bookings impact** a hotel on various fronts:
    - 1. Loss of resources (**revenue**) when the hotel cannot resell the room.
    - 2. **Additional costs** of distribution channels by increasing commissions or paying for publicity to help sell these rooms.
    - 3. Lowering prices last minute, so the hotel can resell a room, resulting in **reducing the profit margin**.
    - 4. **Human resources** to make arrangements for the guests.

Business Problem Overview and Solution Approach

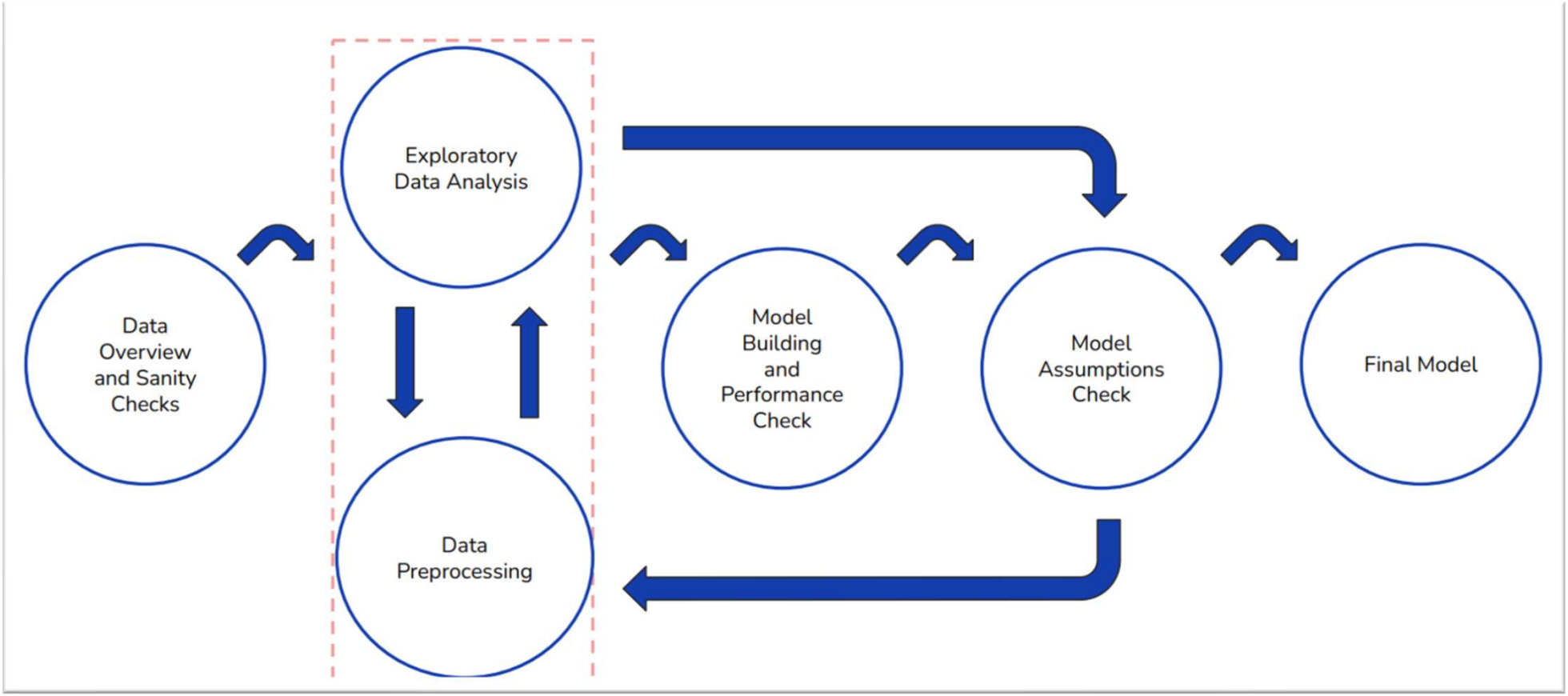
Objective

- **INN Hotels Group** has a chain of hotels in Portugal, they are facing problems with the high number of booking cancellations and are looking for a data-driven solution to reduce cancelations. **INN** need an ML-based solution that can help **in predicting which booking is likely to be canceled.**
- The objective is to analyze the data provided to:
  - **(1) find which factors have a high influence on booking cancellations**,
  - **(2) build a predictive model that can predict which booking is going to be canceled in advance, and**
  - **(3) help in formulating profitable policies for cancellations and refunds.**

Solution Approach

- Build models of **Logistic Regression** and **Decision Tree.** Try and improve each model performance. Compare the two model types: **decision tree** and **logistic regression.**

Business Problem Overview and Solution Approach



EDA Results – Overview

- Please mention the key results from EDA
- Please mention the answers to all the insight based questions

***Note**: You can use more than one slide if needed*

Link to Appendix slide on data background check

EDA Results – Overview & Questions

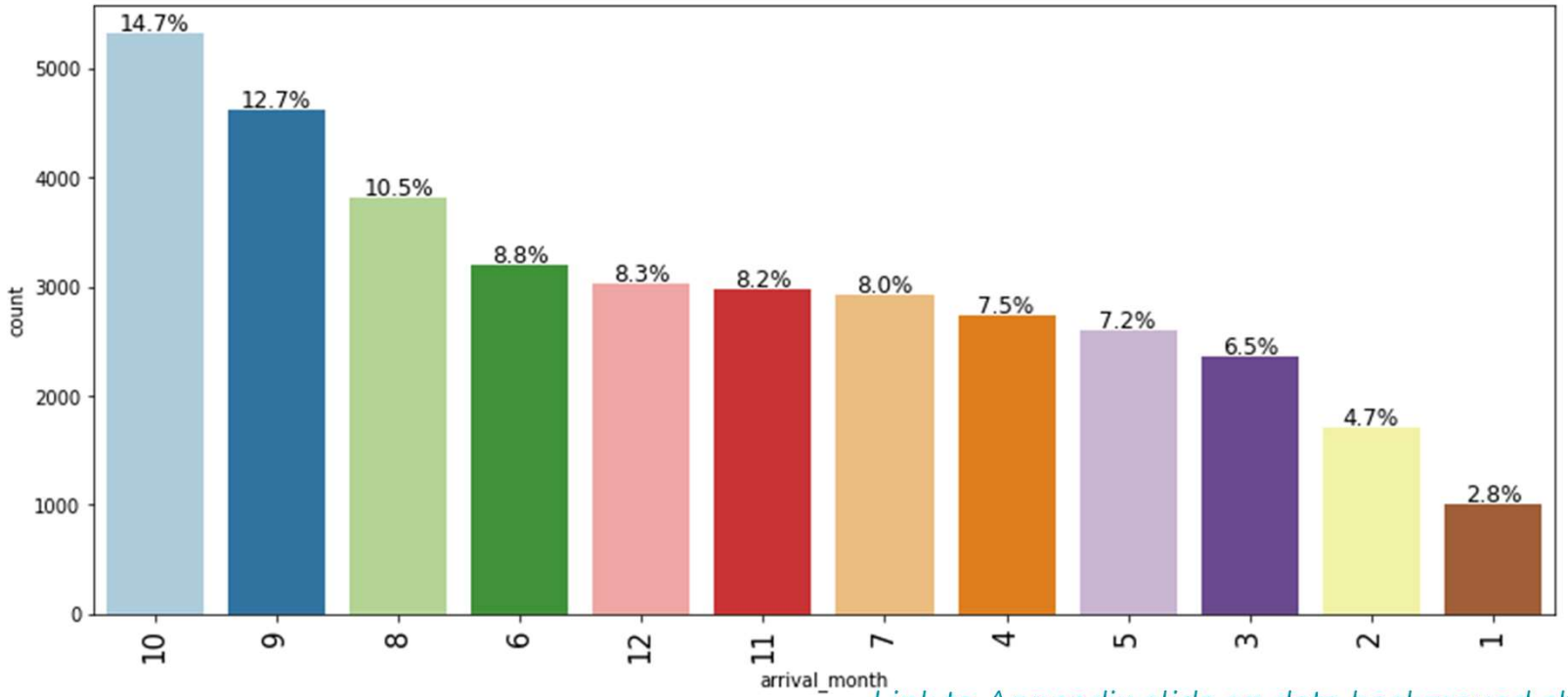
Leading Questions:

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

Link to Appendix slide on data background check

EDA Results – Q1

- Q1 --- What are the busiest months in the hotel?
- October, September, August with respective percent of bookings being ~15%, ~13%, ~11%



Link to Appendix slide on data background check

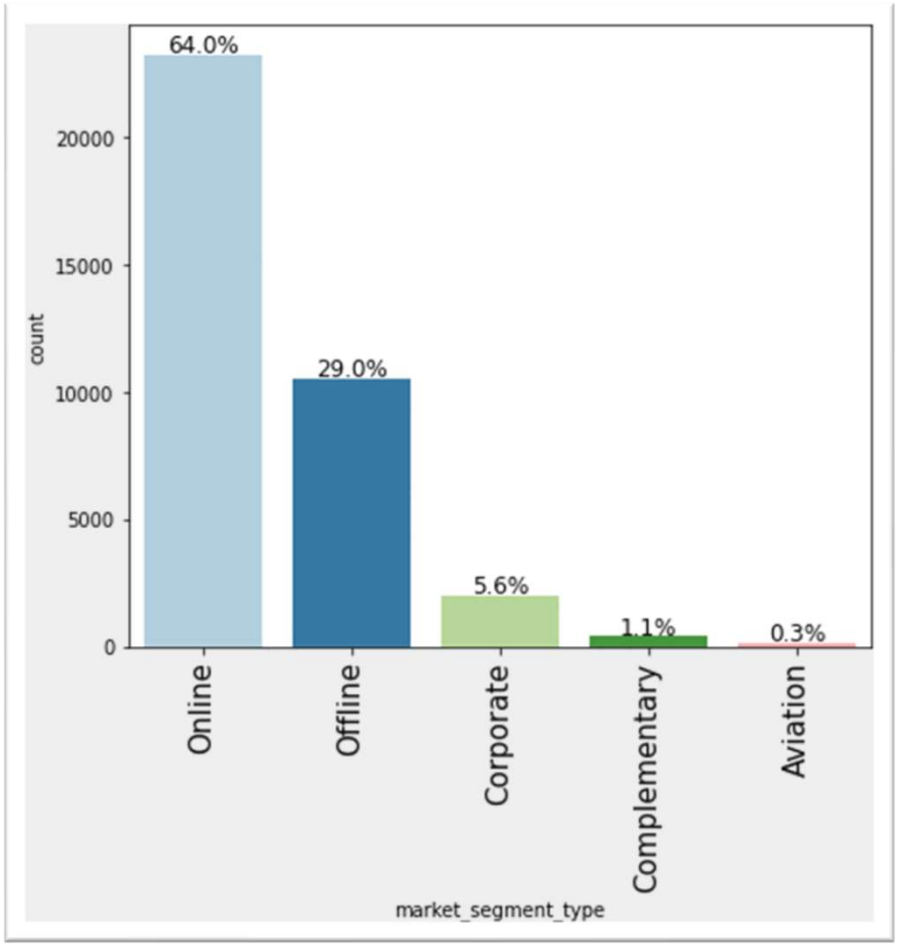
EDA Results – Q2

- Q2 --- Which market segment do most of the guests come from?
- Online ~64%

Link to Appendix slide on data background check

EDA Results – Univariate, market_segment_type

- Nearly two thirds of the bookings are Online ~64%, followed by ~29% offline, only ~6% Corporate



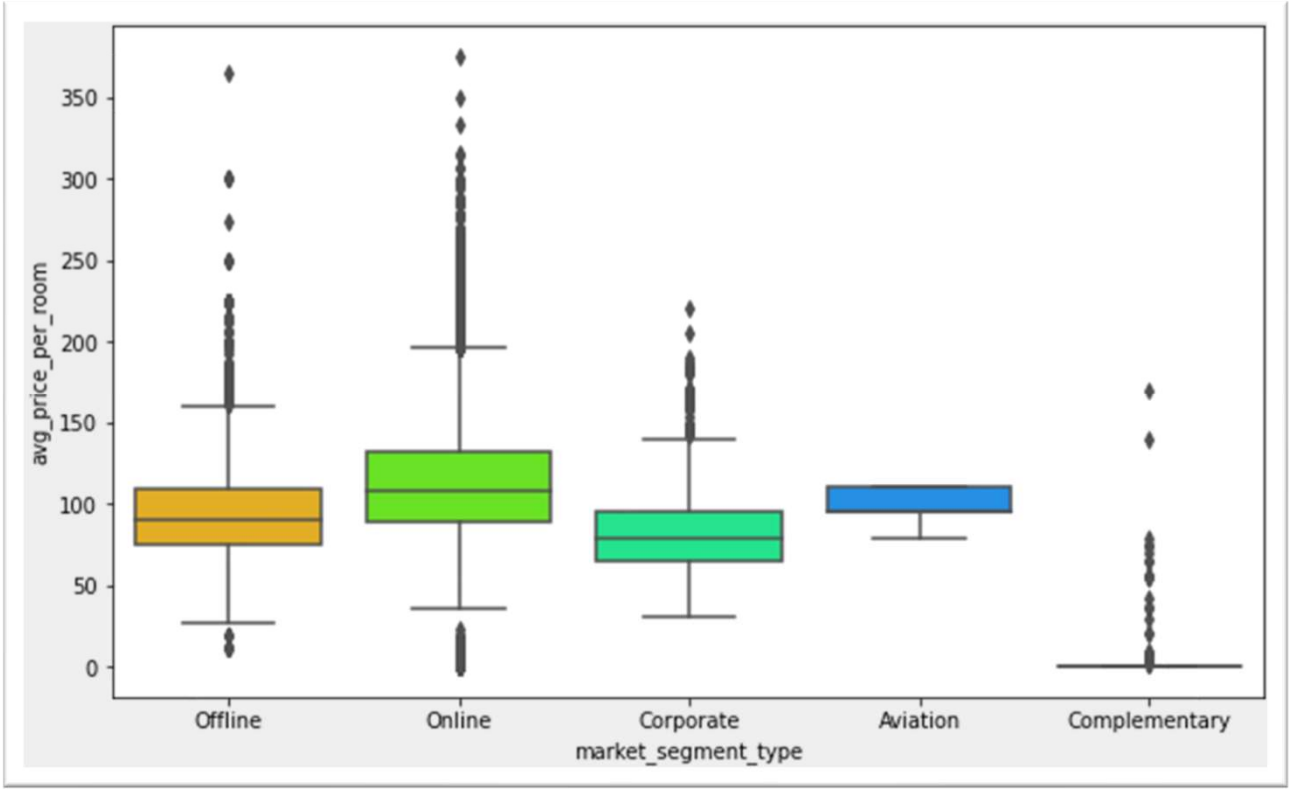
EDA Results – Q3

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

Link to Appendix slide on data background check

EDA Results – Bivariate, price variation per market segments

- Online bookings tend to have higher average price per room. ~15% (~$10 to $40 est.) compared to Offline.



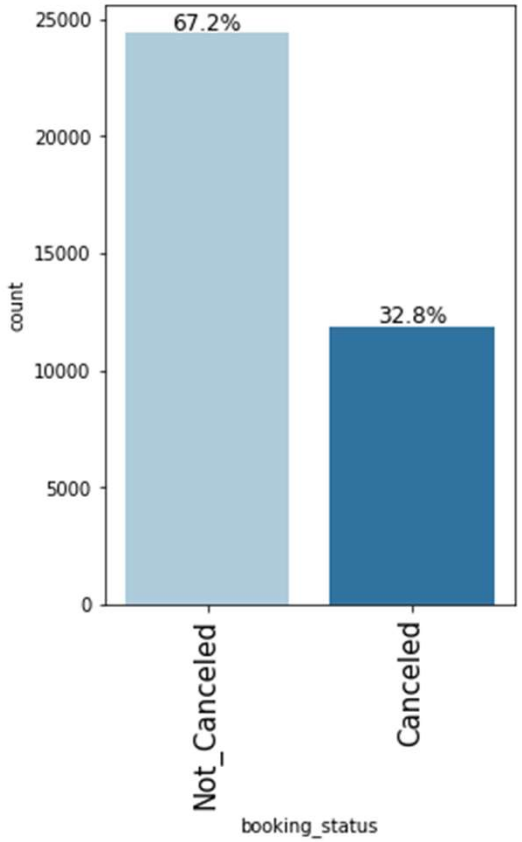
EDA Results – Q4

- Q4 --- What percentage of bookings are canceled?

Link to Appendix slide on data background check

EDA Results – Univariate, ...booking_status

- Two thirds (~67%) of the bookings are **not** canceled
- However, one third (~33%) of the bookings are canceled



EDA Results – Q5

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

Link to Appendix slide on data background check

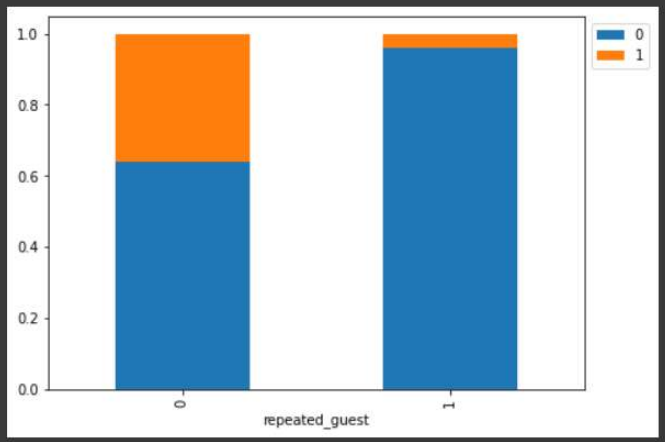
EDA Results – Bivariate,

- Repeat guest cancel, 4% of the repeat guests cancel

Repeating guests are the guests who stay in the hotel often and are important to brand equity. Let's see what percentage of repeating guests cancel?

```
[ ]  stacked_barplot(stay_data, "repeated_guest", "booking_status") ## Complete the code to plot stacked barplot for repeated guests and booking stat
```

```
booking_status      0     1    All
repeated_guest
All             10979  6115  17094
0               10812  6108  16920
1                 167     7    174
```



EDA Results – Q6

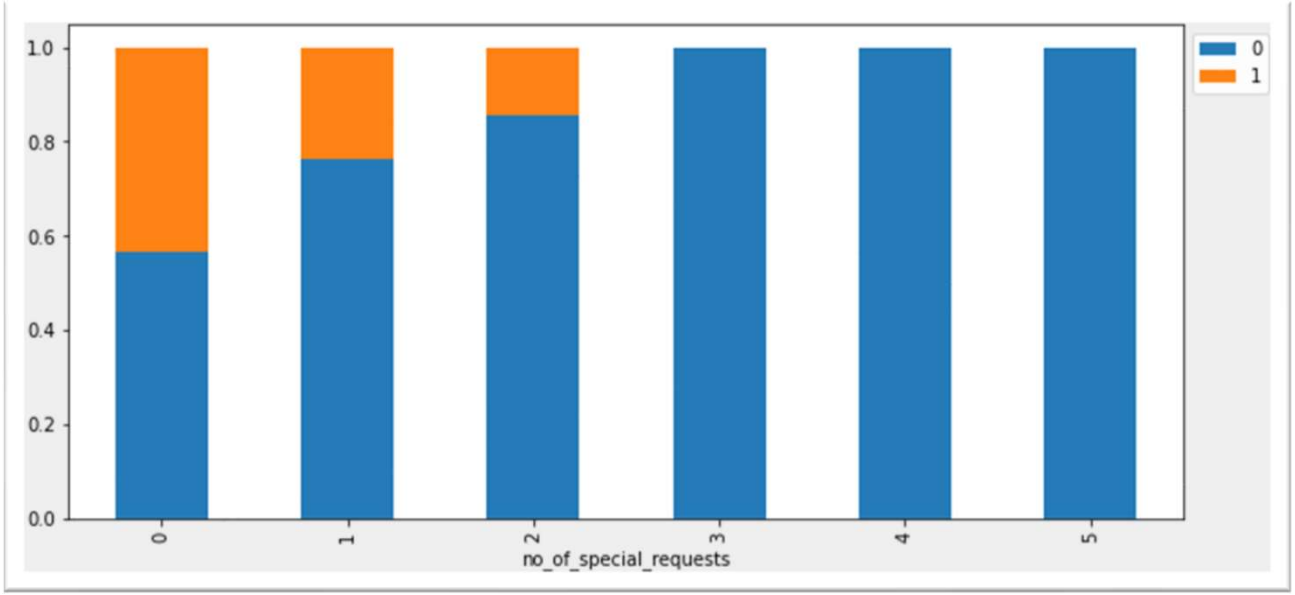
- Q6 --- Many guests have special requirements when booking a hotel room. Do these requirements affect booking cancellation? Yes, next slide.

Link to Appendix slide on data background check

EDA Results – Bivariate, booking status per special requests

- The higher the number of special requests the booking has, the less likely the booking will be canceled.

booking_status	0	1	All
no_of_special_requests			
All	24390	11885	36275
0	11232	8545	19777
1	8670	2703	11373
2	3727	637	4364
3	675	0	675
4	78	0	78
5	8	0	8

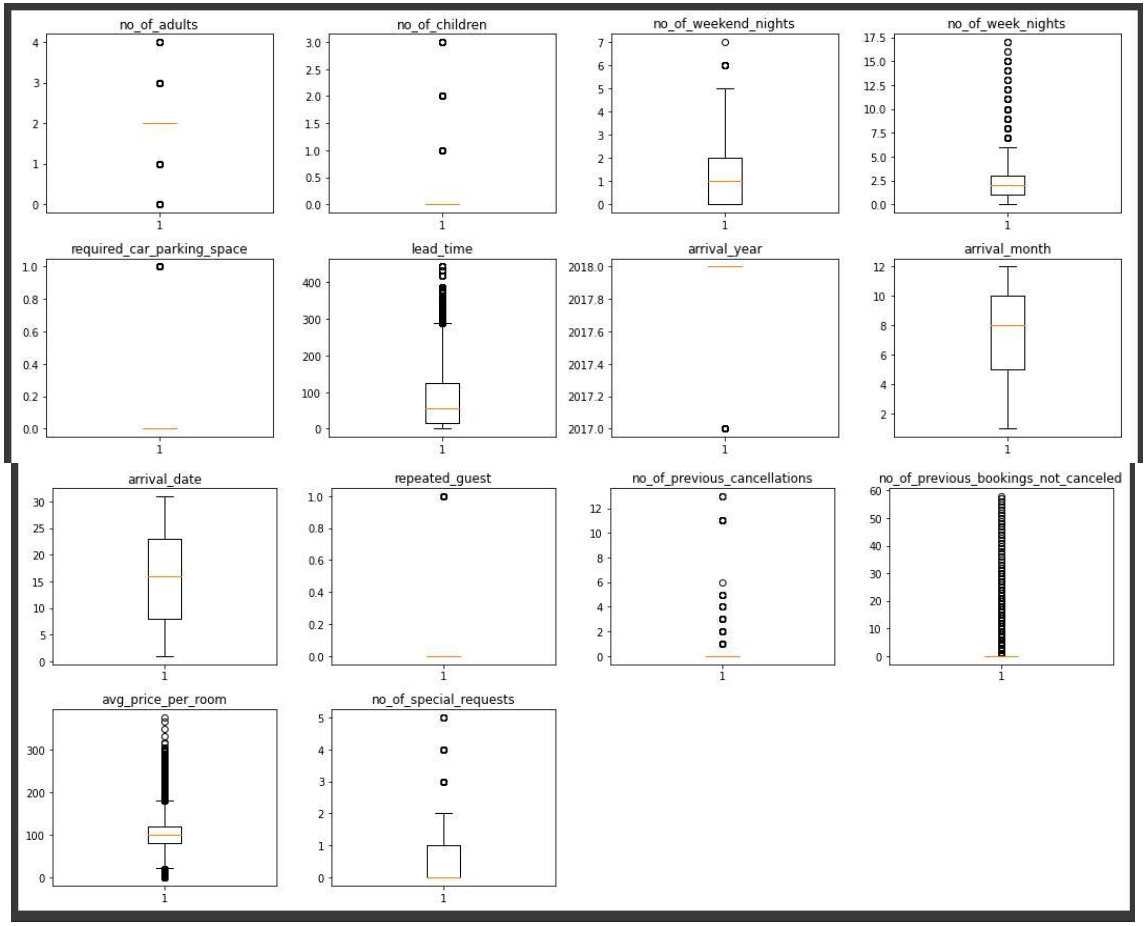


Data Preprocessing – Overview

- Duplicate value check --- NO DUPLICATES
- Missing value treatment --- NO MISSING VALUES
- Outlier check (treatment if needed) --- Data visualized with and without outliers during EDA,
  - Replaced no_of_children outliers 9, and 10 children with 3 children
- Feature engineering --- columns, combined
- Data preparation for modeling
- Any other preprocessing steps

Data Preprocessing – OUTLIER CHECK

- Outlier Check, addressed in overview previous slide.



Data Preprocessing – Feature engineering

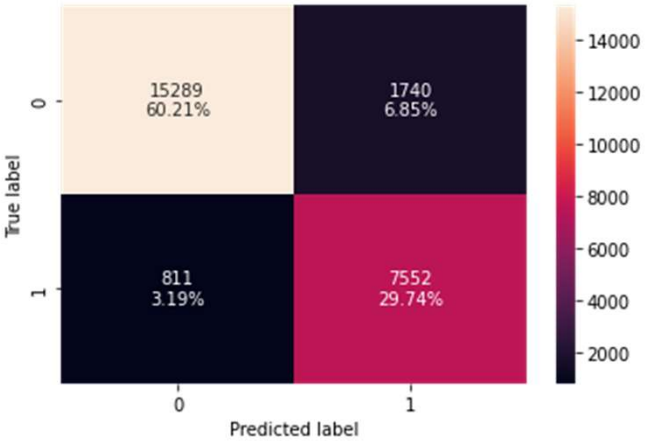
- Concider createing new column or changing **arrival_year** from the 2017 and 2018 to 0 and 1.
- Data processing post-model and assumption checks may require transforming variables. To be determined in next iteration.

Model Performance Summary

- Overview of the final ML model and its parameters
- Summary of most important features used by the ML model for prediction
- Summary of key performance metrics for training and test data of all the models in tabular format for comparison

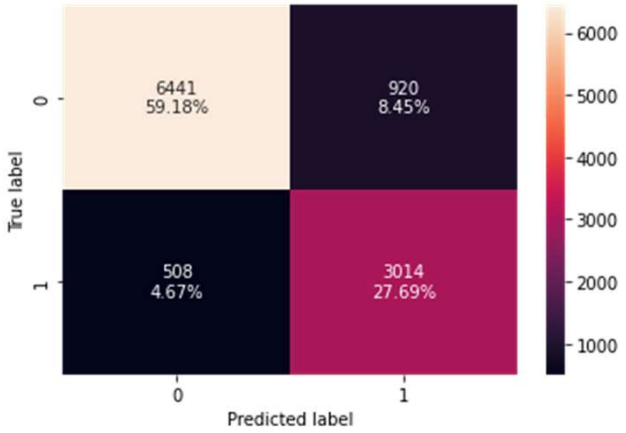
Model Building – Decision Tree

- Checking performances again for training and testing sets… **Not overfitting!**
- Since **F1 score** is within **five %** between **train ~0.8555** and **test ~0.8085** we can say that the model is not over fitting
- Training



	Accuracy	Recall	Precision	F1
0	0.89954	0.90303	0.81274	0.85551

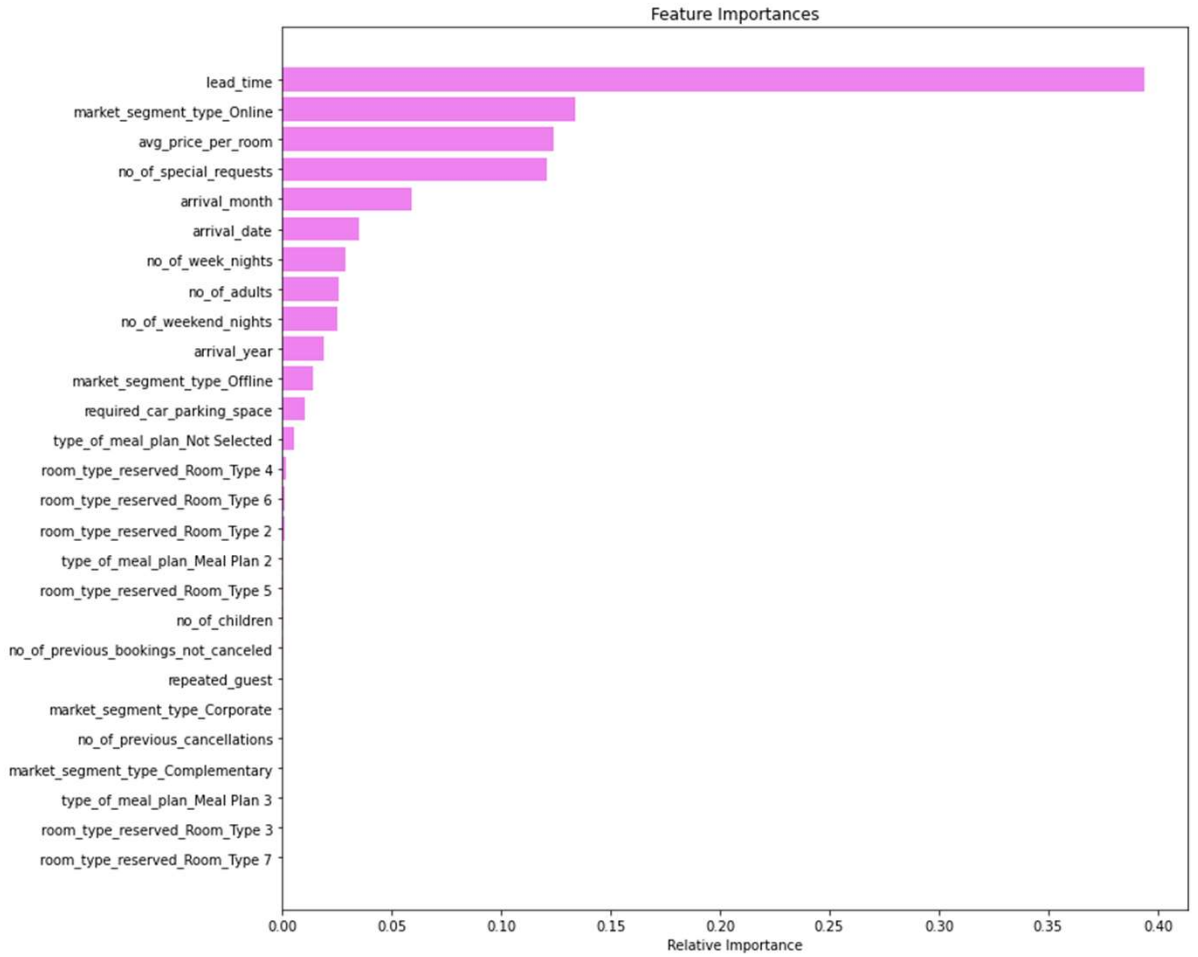
Testing



	Accuracy	Recall	Precision	F1
0	0.86879	0.85576	0.76614	0.80848

Model Building – Decision Tree – Feature Importances

- **lead_time ~0.40** first (1st) place
- **Market_segment type_online ~0.15** second (2nd)
- **average price per room ~0.14** to third (3th) most important.
- **Number of special requests ~0.13** moving up two spots to fourth (4th)
- And so on… lead_time, market_segment_type_online avg_price_per_room, …



APPENDIX

Data Background and Contents

- Please mention the data background and contents

Data Background and Contents – Business Context Given

- Please mention the data background and contents

Context

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

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

The cancellation of bookings impact a hotel on various fronts:

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

Data Background and Contents – Business Objective Given

- Please mention the data background and contents

Objective

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

Data Background and Contents – Description, Dictionary

Data Description

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

Data Dictionary

- Booking_ID: 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 week nights (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)

Data Background and Contents – Dictionary CONTINUED.

- 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.
- lead_time: Number of days between the date of booking and the arrival date
- arrival_year: Year of arrival date
- arrival_month: Month of arrival date
- arrival_date: Date of the month
- market_segment_type: Market segment designation.
- repeated_guest: Is the customer a repeated guest? (0 - No, 1- Yes)
- no_of_previous_cancellations: Number of previous bookings that were canceled by the customer prior to the current booking
- no_of_previous_bookings_not_canceled: Number of previous bookings not canceled by the customer prior to the current booking
- avg_price_per_room: Average price per day of the reservation; prices of the rooms are dynamic. (in euros)
- no_of_special_requests: Total number of special requests made by the customer (e.g. high floor, view from the room, etc)
- booking_status: Flag indicating if the booking was canceled or not.

Da

-

Data Dictionary

- Booking_ID: 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 week nights (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.
- lead_time: Number of days between the date of booking and the arrival date
- arrival_year: Year of arrival date
- arrival_month: Month of arrival date
- arrival_date: Date of the month
- market_segment_type: Market segment designation.
- repeated_guest: Is the customer a repeated guest? (0 - No, 1- Yes)
- no_of_previous_cancellations: Number of previous bookings that were canceled by the customer prior to the current booking
- no_of_previous_bookings_not_canceled: Number of previous bookings not canceled by the customer prior to the current booking
- avg_price_per_room: Average price per day of the reservation; prices of the rooms are dynamic. (in euros)
- no_of_special_requests: Total number of special requests made by the customer (e.g. high floor, view from the room, etc)
- booking_status: Flag indicating if the booking was canceled or not.

at
rning
AHEAD

Da

-

Data Dictionary

- Booking_ID: 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 week nights (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.
- lead_time: Number of days between the date of booking and the arrival date
- arrival_year: Year of arrival date
- arrival_month: Month of arrival date
- arrival_date: Date of the month
- market_segment_type: Market segment designation.
- repeated_guest: Is the customer a repeated guest? (0 - No, 1- Yes)
- no_of_previous_cancellations: Number of previous bookings that were canceled by the customer prior to the current booking
- no_of_previous_bookings_not_canceled: Number of previous bookings not canceled by the customer prior to the current booking
- avg_price_per_room: Average price per day of the reservation; prices of the rooms are dynamic. (in euros)
- no_of_special_requests: Total number of special requests made by the customer (e.g. high floor, view from the room, etc)
- booking_status: Flag indicating if the booking was canceled or not.

Data Background and Contents – TEXT

- no_of_adults: Number of adults
- no_of_children: Number of Children
- no_of_weekend_nights: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
- no_of_week_nights: Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel
- type_of_meal_plan: Type of meal plan booked by the customer:
- Not Selected – No meal plan selected
- Meal Plan 1 – Breakfast
- Meal Plan 2 – Half board (breakfast and one other meal)
- Meal Plan 3 – Full board (breakfast, lunch, and dinner)
- required_car_parking_space: Does the customer require a car parking space? (0 - No, 1- Yes)
- room_type_reserved: Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels Group
- lead_time: Number of days between the date of booking and the arrival date
- arrival_year: Year of arrival date
- arrival_month: Month of arrival date
- arrival_date: Date of the month
- market_segment_type: Market segment designation.
- repeated_guest: Is the customer a repeated guest? (0 - No, 1- Yes)
- no_of_previous_cancellations: Number of previous bookings that were canceled by the customer prior to the current booking
- no_of_previous_bookings_not_canceled: Number of previous bookings not canceled by the customer prior to the current booking
- avg_price_per_room: Average price per day of the reservation; prices of the rooms are dynamic. (in euros)
- no_of_special_requests: Total number of special requests made by the customer (e.g. high floor, view from the room, etc)
- booking_status: Flag indicating if the booking was canceled or not.

Data Background & Contents – Rows, Columns, Data types

Rows and Columns

- 36,275 rows,
- 19 columns

```
data.shape

(36275, 19)
```

Data types include

- Fourteen (14) **numeric**, consisting of
- One (1) float64 and
- Thriteen (13) int64.

- Five (5) **categorical**, consisting of
- Five (5) object type.

There are no missing values

There are no duplicates

```
# checking for duplicate values
data.duplicated().sum() ##  Comp

0
```

```
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
```

Let's check the statistical summary of the data.

```
data.describe().T ##  Complete the code to print the statistical summary of the data
data.describe().T
```

	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

Let's check the statistical summary of the data.

```
data.describe().T ##  Complete the code to print the statistical summary of the data
data.describe().T
```

	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

Data Background and Contents – STATS SUMMARY NOTES

- no_of_adults: Number of adults **---- no more than four? Why is min. zero? Canceled?**
- no_of_children: Number of Children **---- max is ten, that's a lot; other wise zero for all other quartiles 0,0,0,0**
- no_of_weekend_nights: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel **---- 75% (2) to max (10) indicates outlier**
- no_of_week_nights: Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel ---- 75% (3) to max (17) indicates outlier
- type_of_meal_plan:
- 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 **---- 75% (126) to max (443) indicates outlier**
- arrival_year: Year of arrival date **---- 2017 or 2018**
- arrival_month: Month of arrival date ---- travel later in the year?
- arrival_date: Date of the month ---- looked evenly distributed, uniform distribition
- market_segment_type: Market segment designation. **----**
- repeated_guest: Is the customer a repeated guest? (0 - No, 1- Yes)
- no_of_previous_cancellations: Number of previous bookings that were canceled by the customer prior to the current booking
- no_of_previous_bookings_not_canceled: Number of previous bookings not canceled by the customer prior to the current booking
- avg_price_per_room: Average price per day of the reservation; prices of the rooms are dynamic. (in euros)
- no_of_special_requests: Total number of special requests made by the customer (e.g. high floor, view from the room, etc)
- booking_status: Flag indicating if the booking was canceled or not.

```
[19] data.describe(include='all').T
```

	count	unique	top	freq	mean	std	min	25%	50%	75%
no_of_adults	36275.00000	NaN	NaN	NaN	1.84496	0.51871	0.00000	2.00000	2.00000	2.00000
no_of_children	36275.00000	NaN	NaN	NaN	0.10528	0.40265	0.00000	0.00000	0.00000	0.00000
no_of_weekend_nights	36275.00000	NaN	NaN	NaN	0.81072	0.87064	0.00000	0.00000	1.00000	2.00000
no_of_week_nights	36275.00000	NaN	NaN	NaN	2.20430	1.41090	0.00000	1.00000	2.00000	3.00000
type_of_meal_plan	36275	4	Meal Plan 1	27835	NaN	NaN	NaN	NaN	NaN	NaN
required_car_parking_space	36275.00000	NaN	NaN	NaN	0.03099	0.17328	0.00000	0.00000	0.00000	0.00000
room_type_reserved	36275	7	Room_Type 1	28130	NaN	NaN	NaN	NaN	NaN	NaN
lead_time	36275.00000	NaN	NaN	NaN	85.23256	85.93082	0.00000	17.00000	57.00000	126.00000
arrival_year	36275.00000	NaN	NaN	NaN	2017.82043	0.38384	2017.00000	2018.00000	2018.00000	2018.00000
arrival_month	36275.00000	NaN	NaN	NaN	7.42365	3.06989	1.00000	5.00000	8.00000	10.00000
arrival_date	36275.00000	NaN	NaN	NaN	15.59700	8.74045	1.00000	8.00000	16.00000	23.00000
market_segment_type	36275	5	Online	23214	NaN	NaN	NaN	NaN	NaN	NaN
repeated_guest	36275.00000	NaN	NaN	NaN	0.02564	0.15805	0.00000	0.00000	0.00000	0.00000
no_of_previous_cancellations	36275.00000	NaN	NaN	NaN	0.02335	0.36833	0.00000	0.00000	0.00000	0.00000
no_of_previous_bookings_not_canceled	36275.00000	NaN	NaN	NaN	0.15341	1.75417	0.00000	0.00000	0.00000	0.00000
avg_price_per_room	36275.00000	NaN	NaN	NaN	103.42354	35.08942	0.00000	80.30000	99.45000	120.00000
no_of_special_requests	36275.00000	NaN	NaN	NaN	0.61966	0.78624	0.00000	0.00000	0.00000	1.00000
booking_status	36275	2	Not_Canceled	24390	NaN	NaN	NaN	NaN	NaN	NaN

```
[19] data.describe(include='all').T
```

	count	unique	top	freq	mean	std	min	25%	50%	75%
no_of_adults	36275.00000	NaN	NaN	NaN	1.84496	0.51871	0.00000	2.00000	2.00000	2.00000
no_of_children	36275.00000	NaN	NaN	NaN	0.10528	0.40265	0.00000	0.00000	0.00000	0.00000
no_of_weekend_nights	36275.00000	NaN	NaN	NaN	0.81072	0.87064	0.00000	0.00000	1.00000	2.00000
no_of_week_nights	36275.00000	NaN	NaN	NaN	2.20430	1.41090	0.00000	1.00000	2.00000	3.00000
type_of_meal_plan	36275	4	Meal Plan 1	27835	NaN	NaN	NaN	NaN	NaN	NaN
required_car_parking_space	36275.00000	NaN	NaN	NaN	0.03099	0.17328	0.00000	0.00000	0.00000	0.00000
room_type_reserved	36275	7	Room_Type 1	28130	NaN	NaN	NaN	NaN	NaN	NaN
lead_time	36275.00000	NaN	NaN	NaN	85.23256	85.93082	0.00000	17.00000	57.00000	126.00000
arrival_year	36275.00000	NaN	NaN	NaN	2017.82043	0.38384	2017.00000	2018.00000	2018.00000	2018.00000
arrival_month	36275.00000	NaN	NaN	NaN	7.42365	3.06989	1.00000	5.00000	8.00000	10.00000
arrival_date	36275.00000	NaN	NaN	NaN	15.59700	8.74045	1.00000	8.00000	16.00000	23.00000
market_segment_type	36275	5	Online	23214	NaN	NaN	NaN	NaN	NaN	NaN
repeated_guest	36275.00000	NaN	NaN	NaN	0.02564	0.15805	0.00000	0.00000	0.00000	0.00000
no_of_previous_cancellations	36275.00000	NaN	NaN	NaN	0.02335	0.36833	0.00000	0.00000	0.00000	0.00000
no_of_previous_bookings_not_canceled	36275.00000	NaN	NaN	NaN	0.15341	1.75417	0.00000	0.00000	0.00000	0.00000
avg_price_per_room	36275.00000	NaN	NaN	NaN	103.42354	35.08942	0.00000	80.30000	99.45000	120.00000
no_of_special_requests	36275.00000	NaN	NaN	NaN	0.61966	0.78624	0.00000	0.00000	0.00000	1.00000
booking_status	36275	2	Not_Canceled	24390	NaN	NaN	NaN	NaN	NaN	NaN

Data Background and Contents – Categorical

- Statistical Summary, Categorical

catigorical column statistics

count and percentage of categorical levels in each column

```
# catigorical column statistics

# Making a list of all categorical variables
cat_cols_inn =  ['type_of_meal_plan', 'room_type_reserved', 'market_segment_type', 'booking_status']

# Printing the count (and the percentage) of unique categorical levels in each column
for column in cat_cols_inn:
    print(data[column].value_counts())
    print(data[column].value_counts(normalize=True))
    print("-" * 50)
```

Data Background and Contents – Categorical

- **type_of_meal_plan**, 76% Meal Plan 1, 14% not selected, 9% Meal Plan 2,...
- **room_type_reserved**, 78% Room_Type 1, 17% Room_Type 4,...

```
Meal Plan 1     27835
Not Selected     5130
Meal Plan 2      3305
Meal Plan 3         5
Name: type_of_meal_plan, dtype: int64
Meal Plan 1    0.76733
Not Selected   0.14142
Meal Plan 2    0.09111
Meal Plan 3    0.00014
Name: type_of_meal_plan, dtype: float64
```

```
----------------------------------------
Room_Type 1    28130
Room_Type 4     6057
Room_Type 6      966
Room_Type 2      692
Room_Type 5      265
Room_Type 7      158
Room_Type 3        7
Name: room_type_reserved, dtype: int64
Room_Type 1   0.77547
Room_Type 4   0.16697
Room_Type 6   0.02663
Room_Type 2   0.01908
Room_Type 5   0.00731
Room_Type 7   0.00436
Room_Type 3   0.00019
Name: room_type_reserved, dtype: float64
```

Data Background and Contents – Categorical

- **market_segment_type**, 64% Offline, 29% Offline, 5% Corporate,...
- **booking_status**, 67% **Not**_Canceled, 33% Canceled,...

```
Online           23214
Offline          10528
Corporate         2017
Complementary      391
Aviation           125
Name: market_segment_type, dtype: int64
Online          0.63994
Offline         0.29023
Corporate       0.05560
Complementary   0.01078
Aviation        0.00345
Name: market_segment_type, dtype: float64
```

```
Not_Canceled    24390
Canceled        11885
Name: booking_status, dtype: int64
Not_Canceled   0.67236
Canceled       0.32764
Name: booking_status, dtype: float64
```

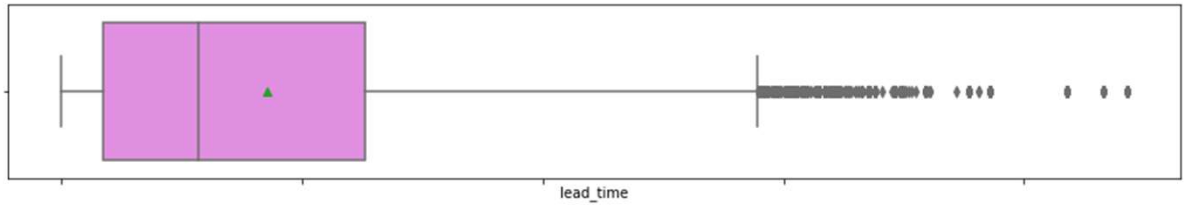
EDA – Nearly Full Supplementary INFO

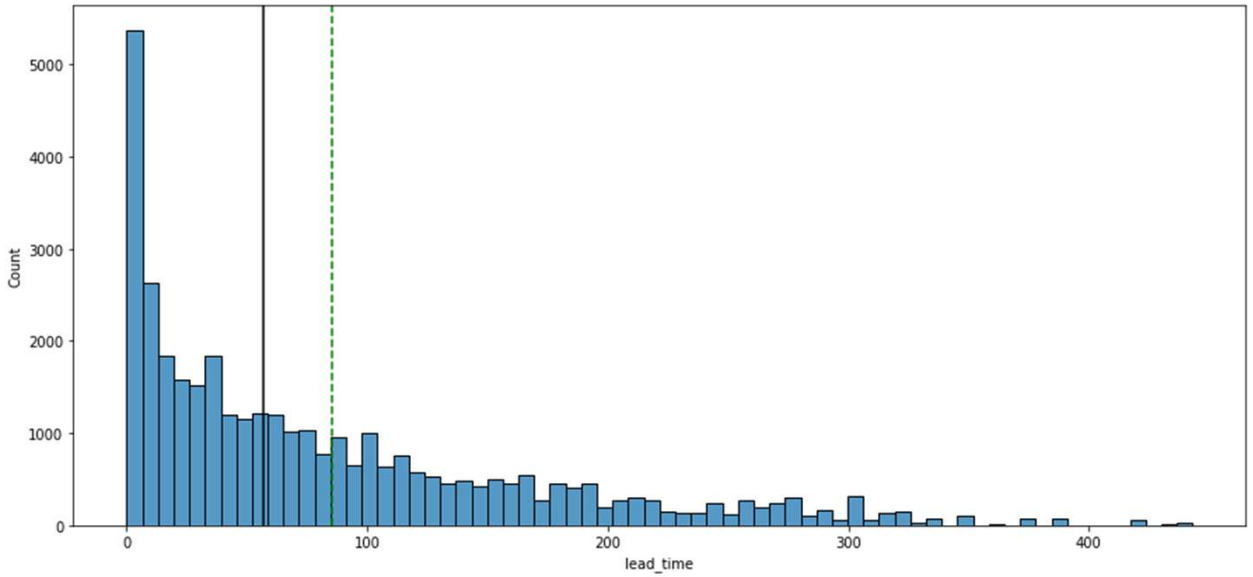
EDA Results – Univariate, Overview

- **Univariate Analysis**
- Observations on lead time
- Observations on average price per room
- Observations on number of previous booking cancellations
- Observations on number of previous booking not canceled
- Observations on number of adults
- Observations on number of children
- Observations on number of week nights
- Observations on number of weekend nights
- Observations on required car parking space
- Observations on type of meal plan
- Observations on room type reserved
- Observations on arrival month
- Observations on market segment type
- Observations on number of special requests
- Observations on booking status

EDA Results – Univariate, lead_time

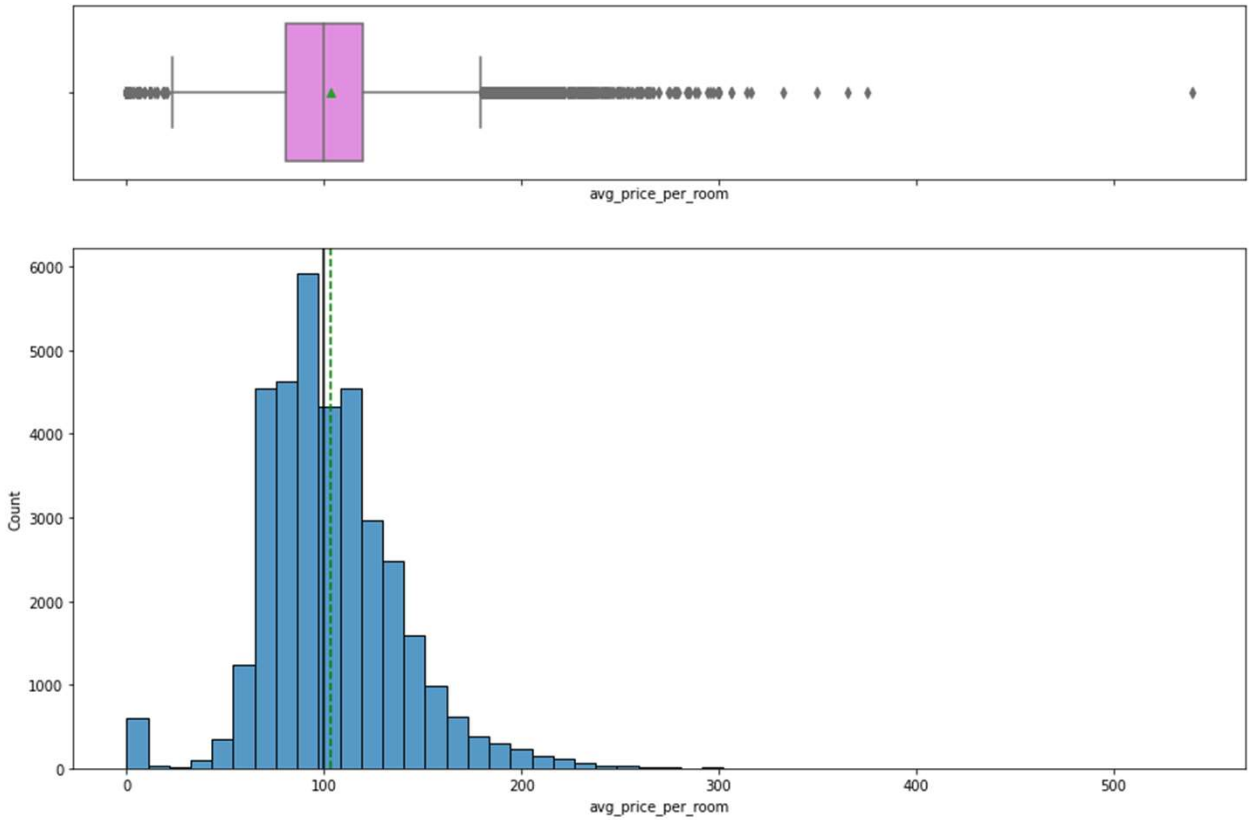
- Lead time, long right tail, many outliers concentrated to the right.
- Not normal distribution (expo)
- Most booking are done less than 50 days out.
- Spike of bookings within a week.





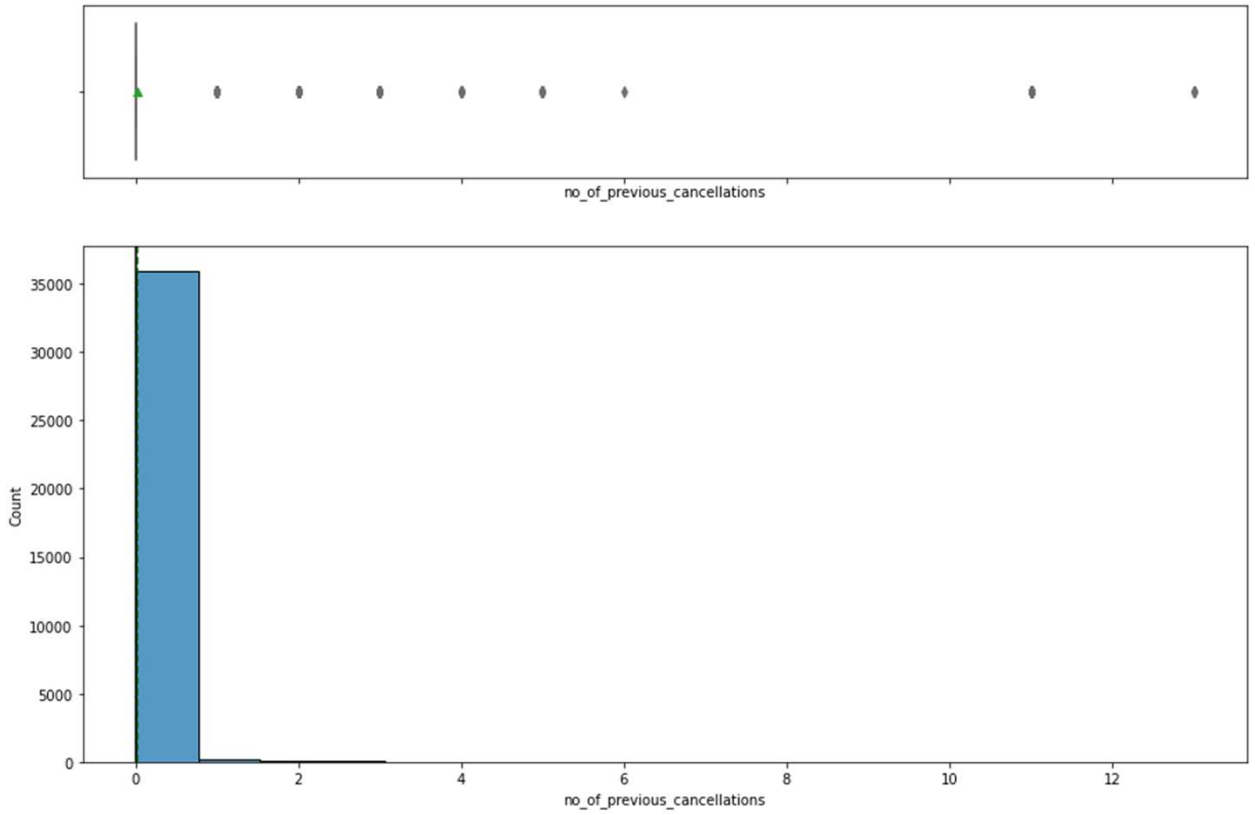
EDA Results – Univariate, avg_price_per_room

- Average price per room, long right tail, many outliers on both left and right, one extreme outlier to the right.



EDA Results – Univariate, no_of_previous_cancelations

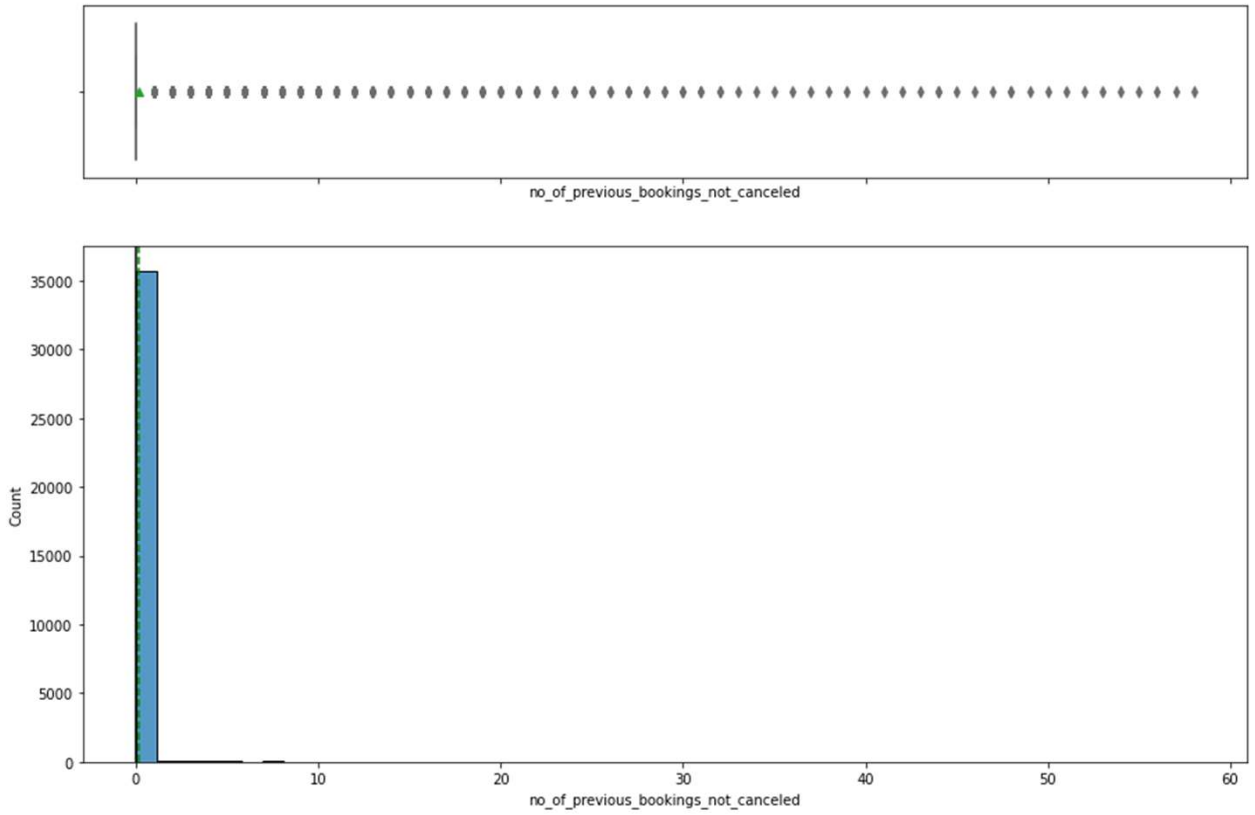
- Number of previous booking cancelations, box plot smashed to zero, 8 outliers, consider review without zero.



EDA Results – Univariate, no_of_previous_NOT_canceled

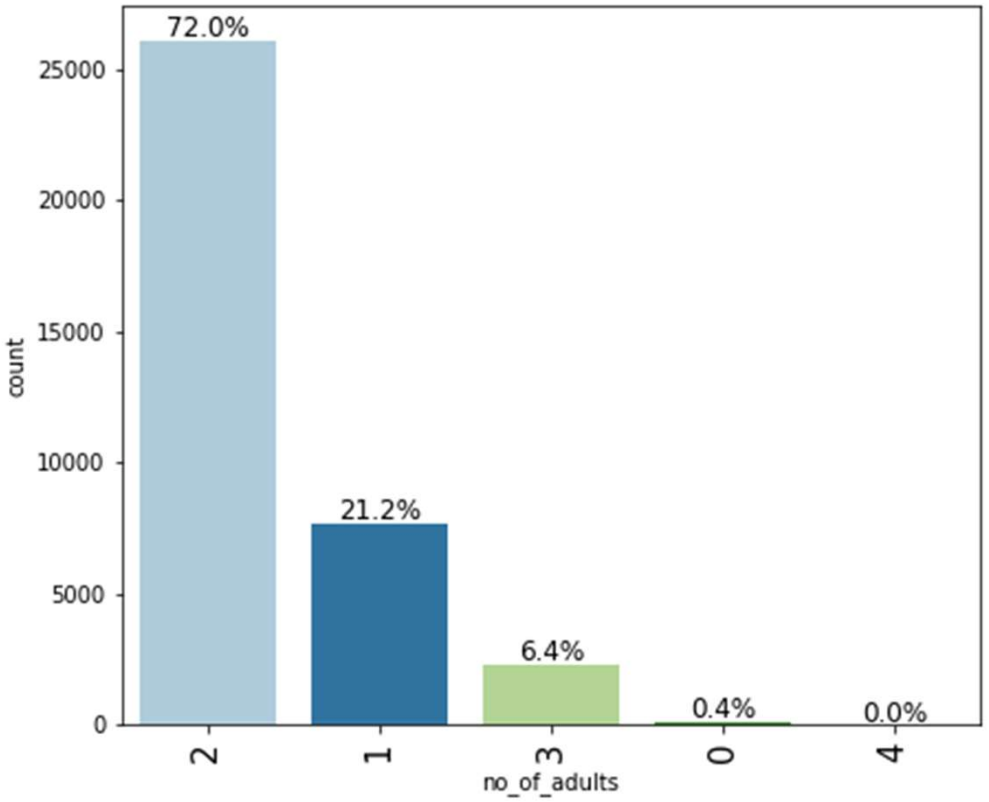
- Number of previous booking NOT canceled, box plot smashed to zero, many outliers, consider review without zero, one or two. Evaluate for treatment.

- Overwhelming majority 1 or 2 previous bookings



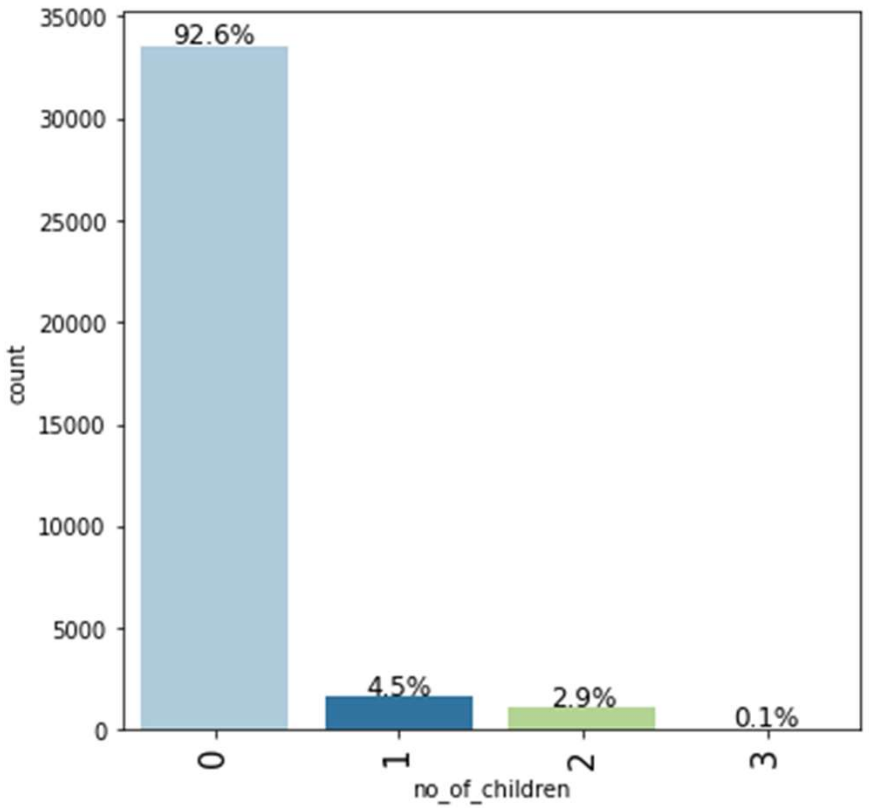
EDA Results – Univariate, no_of_adults

- Number of adults,
  - 72% two adults
  - 21% one adult,
  - 6.4 three adults
- Overwhelming majority of bookings have 1 or 2 adults



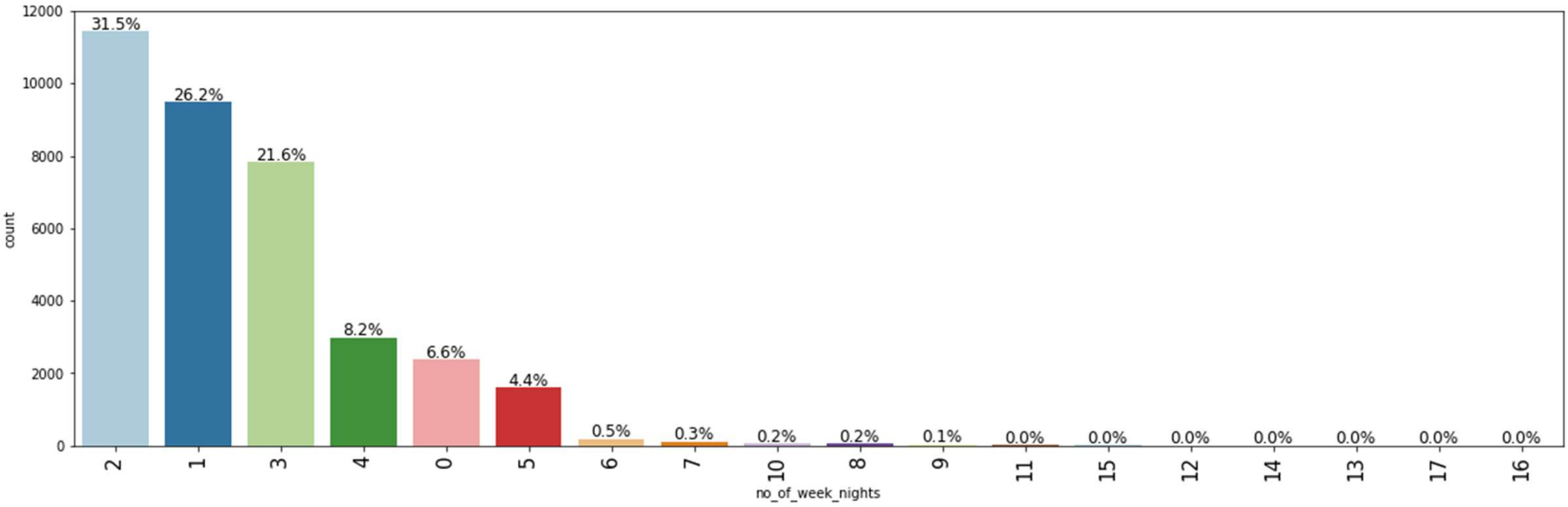
EDA Results – Univariate, no_of_children

- Number of children,
  - 93% zero children
- Overwhelming majority of bookings don't have children.



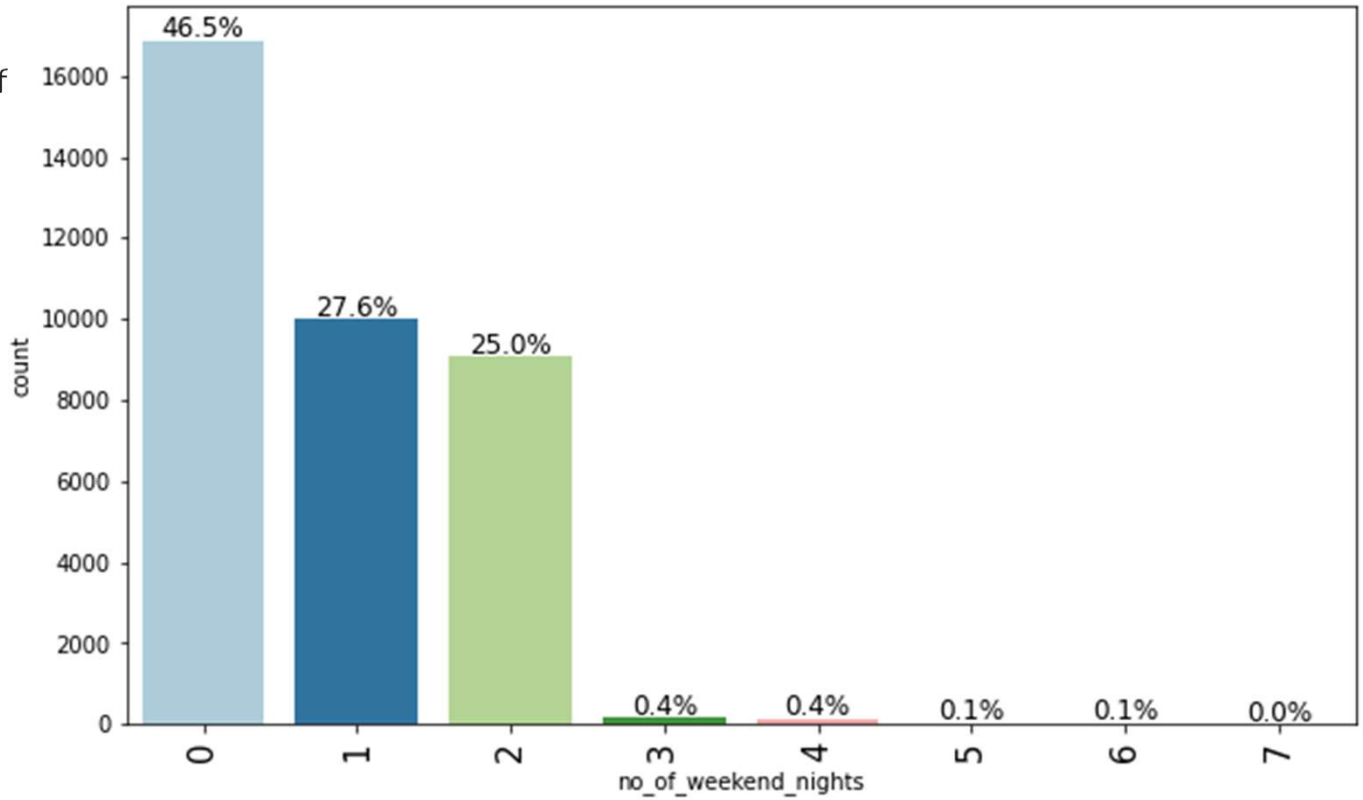
EDA Results – Univariate, no_of_week_nights

- no_of_week_nights, highest are 2, 1 ,3 at respectively ~32%, ~26%, ~22%



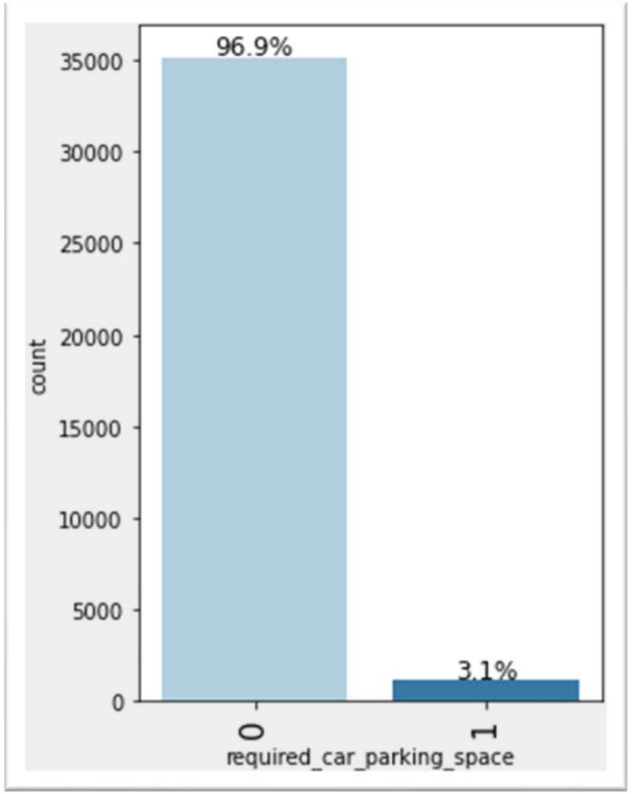
EDA Results – Univariate, ...no_of_weekend_nights

- no_of_week**end**_nights,
- Zero weekend nights is nearly half (~47%) of the weekend bookings
- 1 and 2 weekend nights are split nearly even at ~28% and ~25% respectively

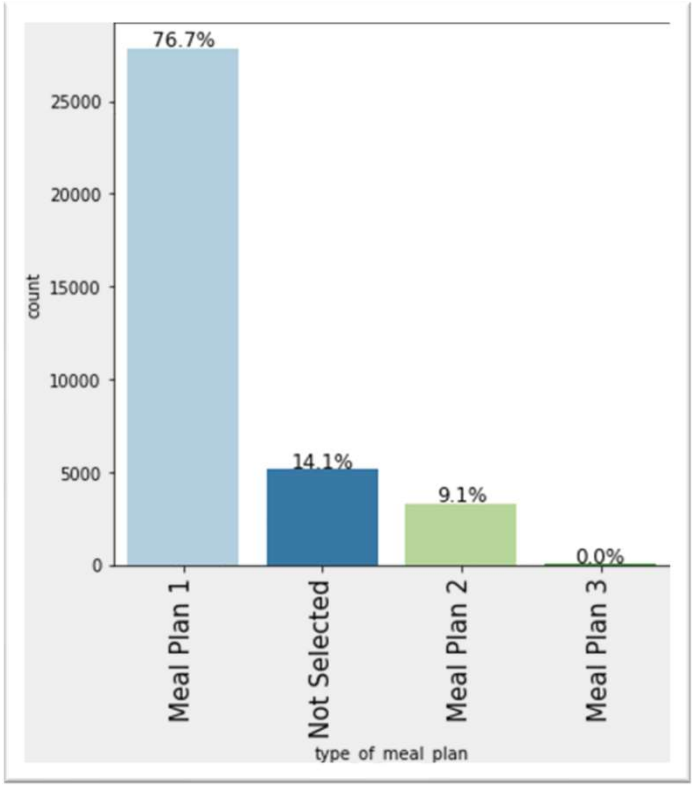


EDA Results – Univariate, ...

- ... required_car_parking_space ...only ~3% of guests require parking

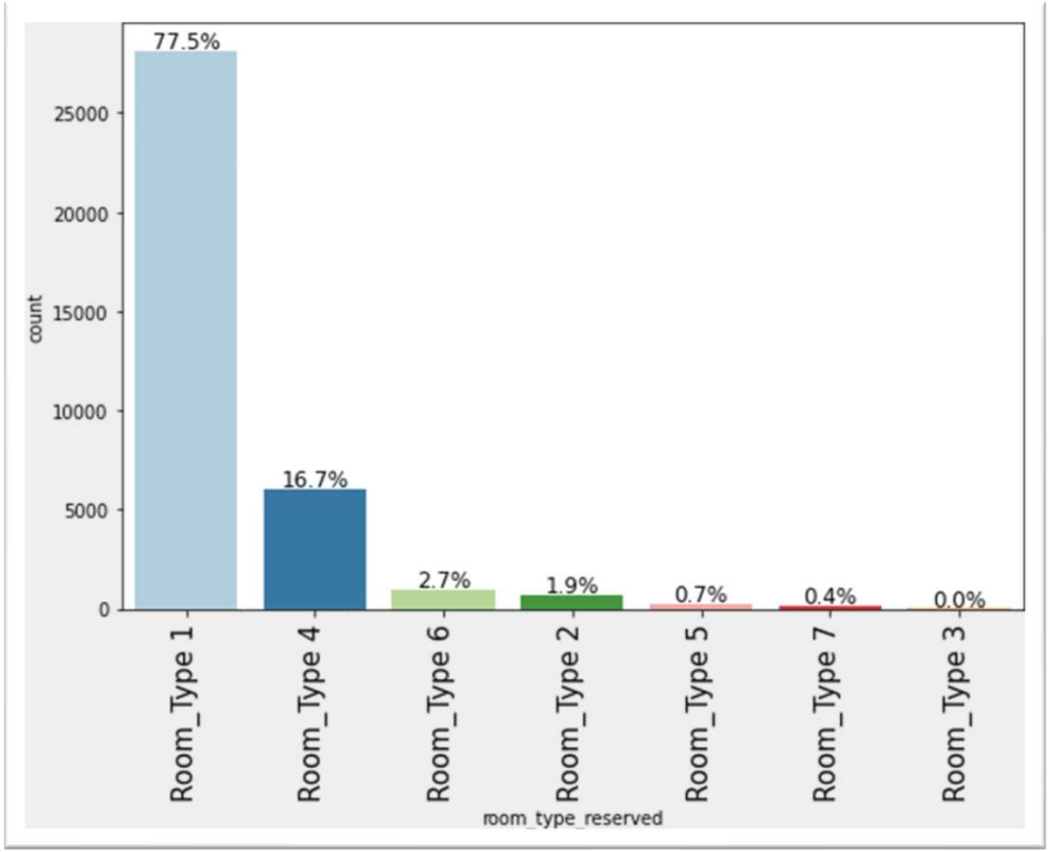


- ...type_of_meal_plan, Meal plan 1 most popular



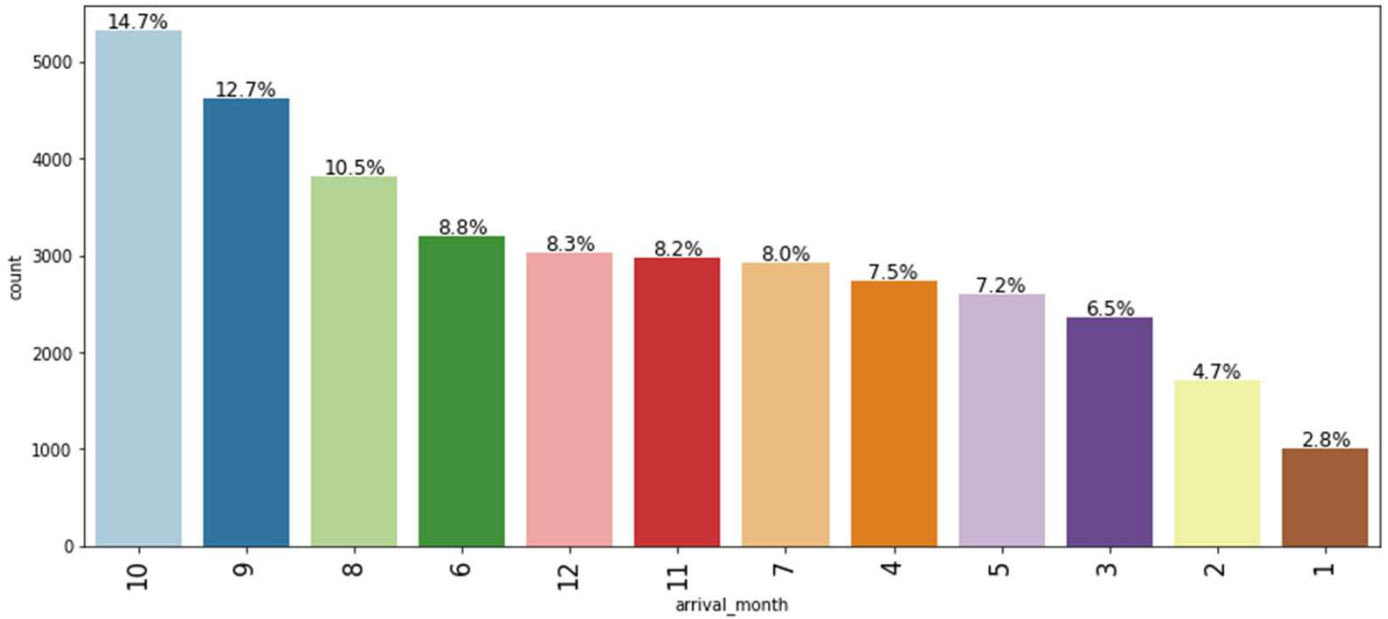
EDA Results – Univariate, … room_type_reserved

- room_type_reserved,
  Room_Type1 is most popular



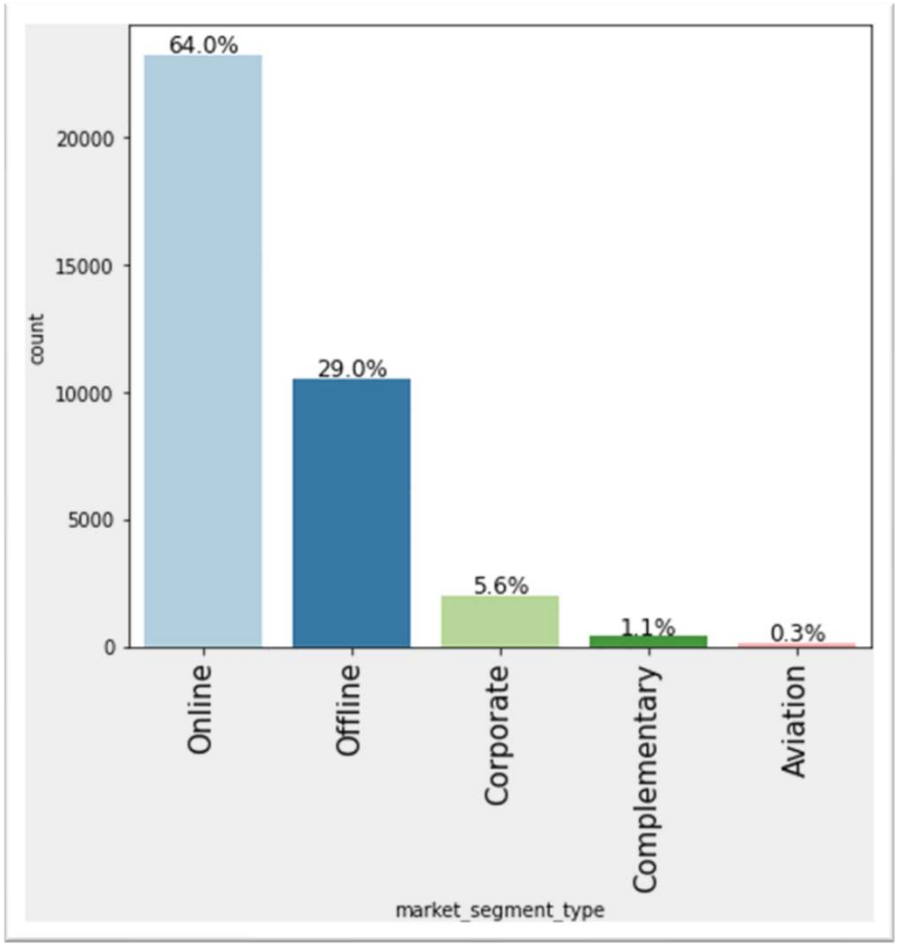
EDA Results – Univariate, ... arrival_month

- October, September, August with respective percent of bookings being ~15%, ~13%, ~11%
- Looks like a steady increase from January (~3%), Feb (~5%), March (~7%), then level off in April May, June, July then steady strong increase Aug, Sept, Oct, then drop off in Nov and Dec.
- Speculating, a fiscal year for most companies end in October?



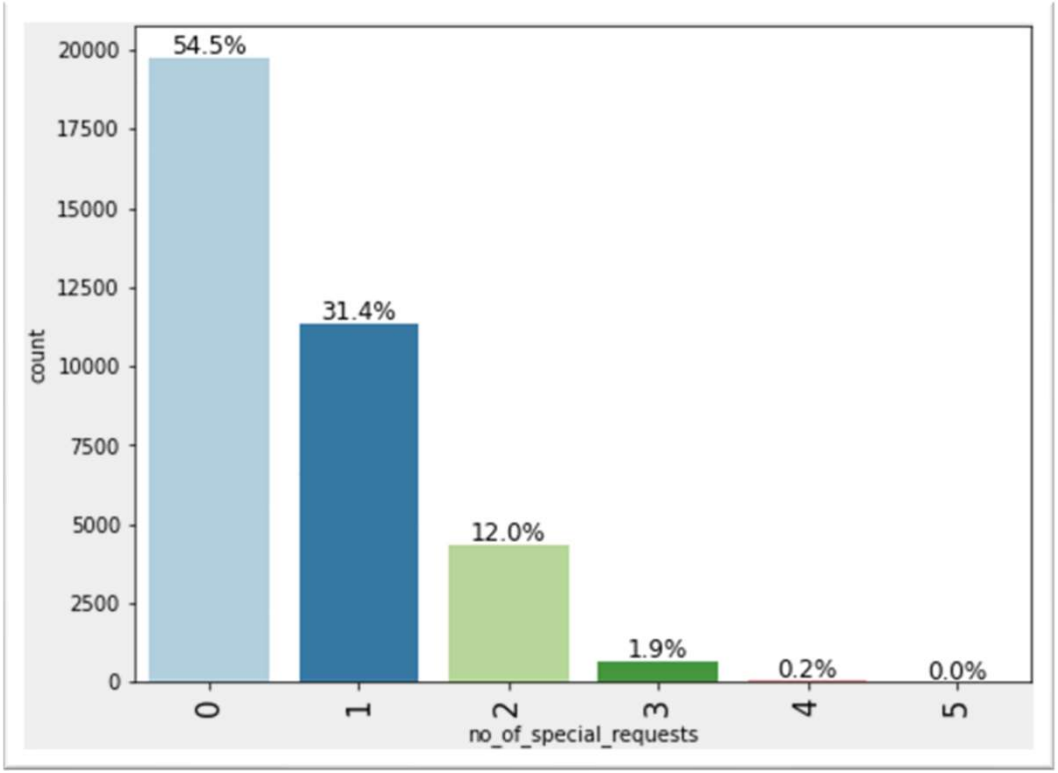
EDA Results – Univariate, ...market_segment_type

- Nearly two thirds of the bookings are Online ~64%, followed by ~29% offline, only ~6% Corporate



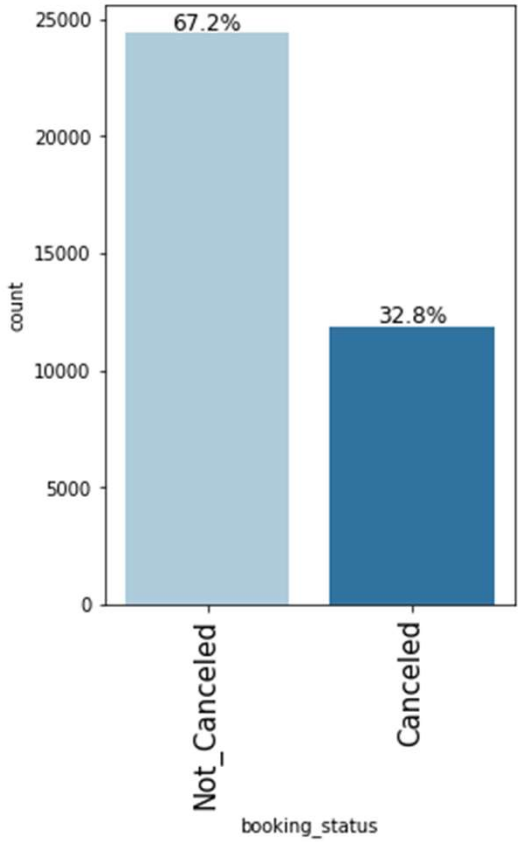
EDA Results – Univariate, ...no_of_special_requests

- Most guests ~55% don't have any special requests,
- However, all others have at least one special request:
  - a third ~31% have one spec
  - 12% have two spec req
- followed by ~29% offline, only ~6% Corporate



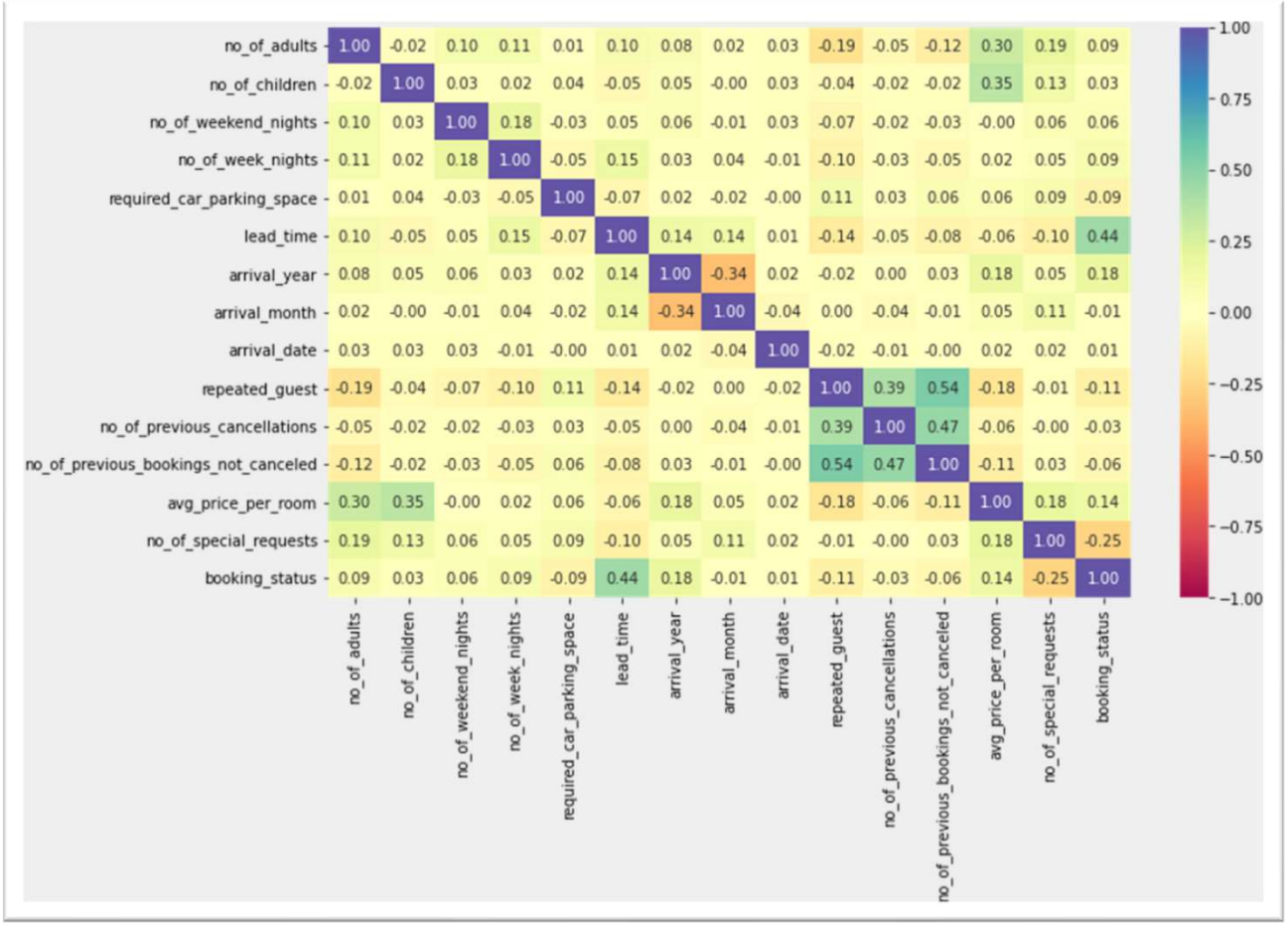
EDA Results – Univariate, ...booking_status

- Two thirds (~67%) of the bookings are **not** canceled
- However, one third (~33%) of the bookings are canceled



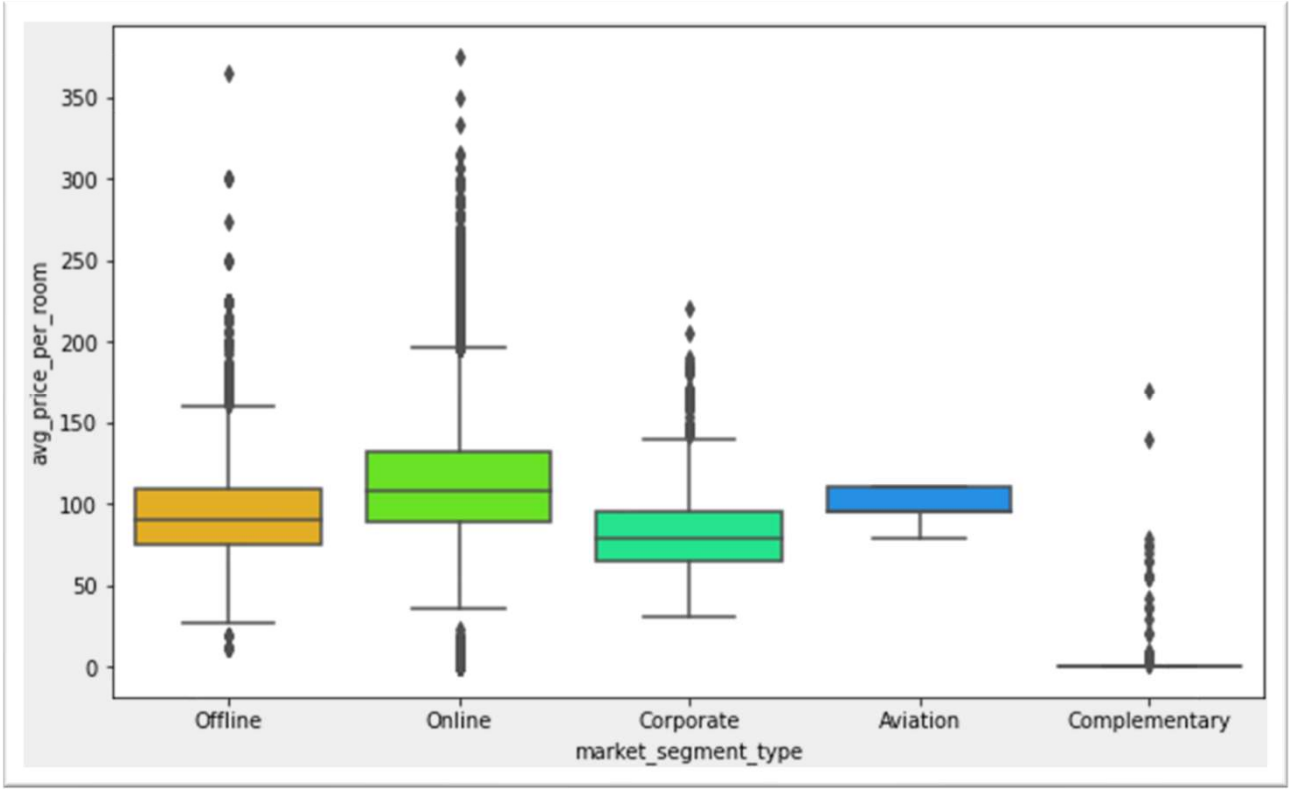
EDA Results – Bivariate – Correlation

- Highest correlations are moderate:
- 0.54, 0.47, 0.39, clustered around repeat guests, and number of previous bookings canceled and not cancelded
- Moderate correlation of 0.44 for lead_time and booking status
- Moderate correlation of average price per room to number of guests (adults 0.30 and children 0.35).
- Most others are weekly corelated
- A few inverse -0.34, only two years
- Addressed later -0.25



EDA Results – Bivariate, price variation per market segments

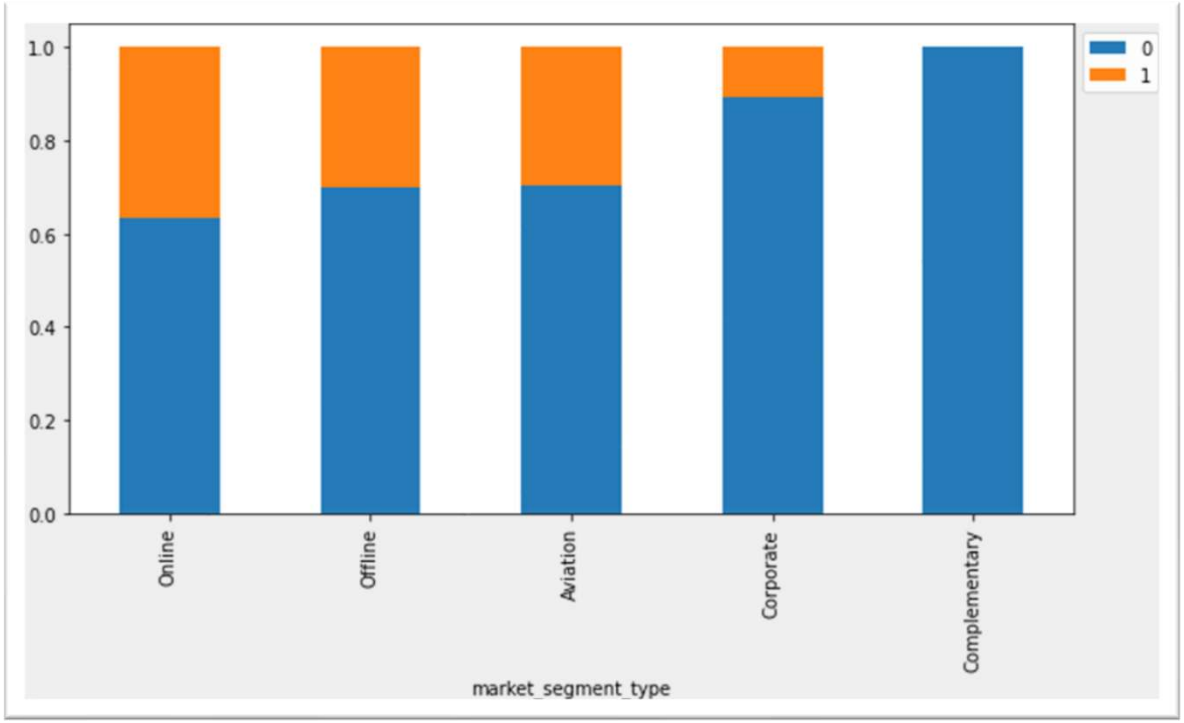
- Online bookings tend to have higher average price per room. ~15% (~$10 to $40 est.) compared to Offline.



EDA Results – Bivariate, booking status per market segments

- Online bookings (~0.35) tend to have more cancelations than Offline (~0.30).

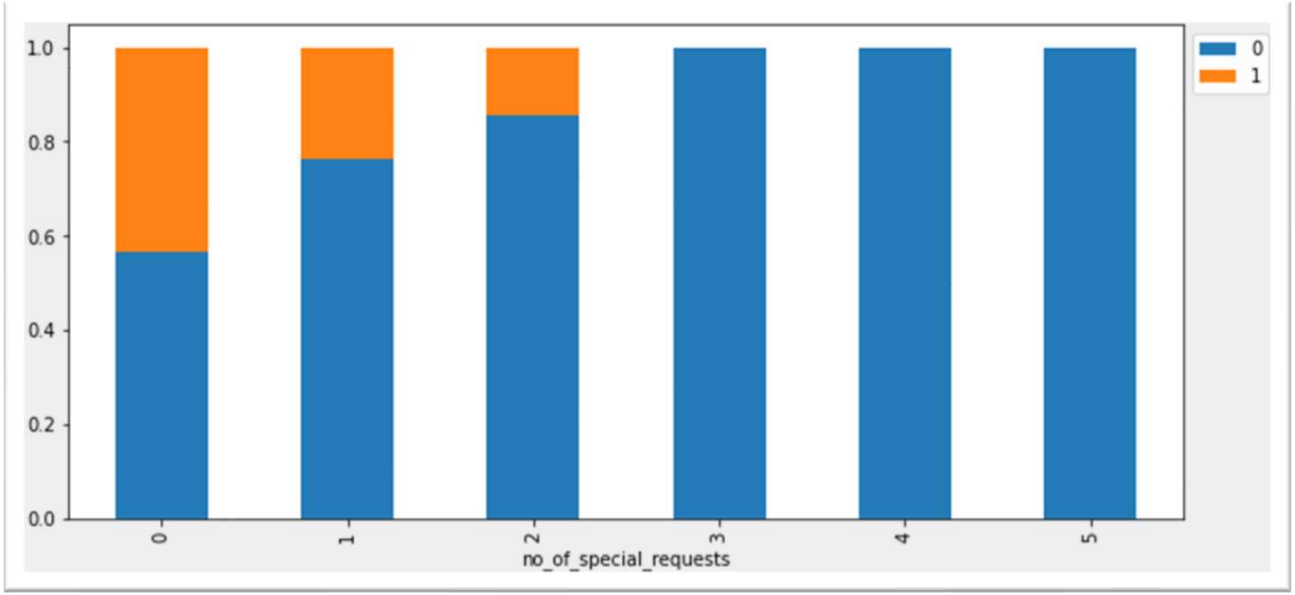
booking_status	0	1	All
market_segment_type			
All	24390	11885	36275
Online	14739	8475	23214
Offline	7375	3153	10528
Corporate	1797	220	2017
Aviation	88	37	125
Complementary	391	0	391



EDA Results – Bivariate, booking status per special requests

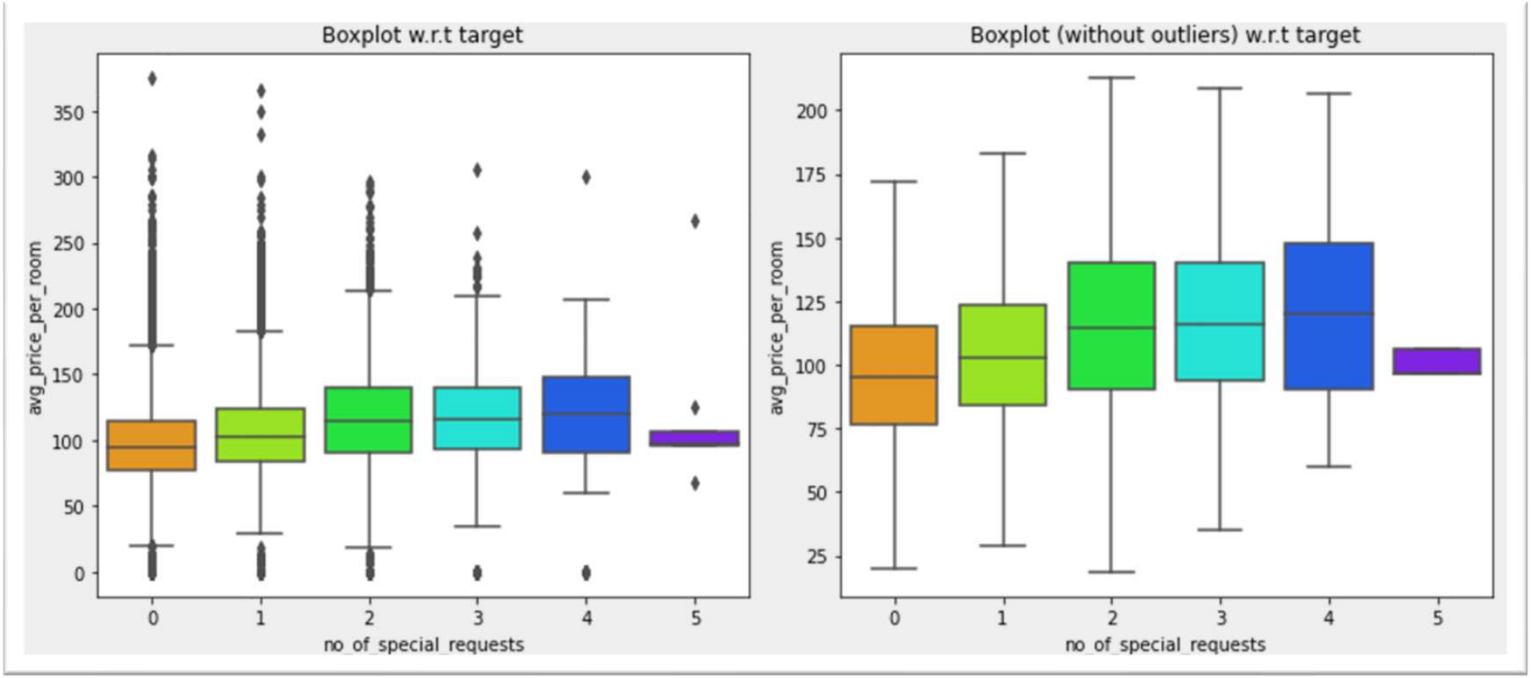
- The higher the number of special requests the booking has, the less likely the booking will be canceled.

booking_status	0	1	All
no_of_special_requests			
All	24390	11885	36275
0	11232	8545	19777
1	8670	2703	11373
2	3727	637	4364
3	675	0	675
4	78	0	78
5	8	0	8



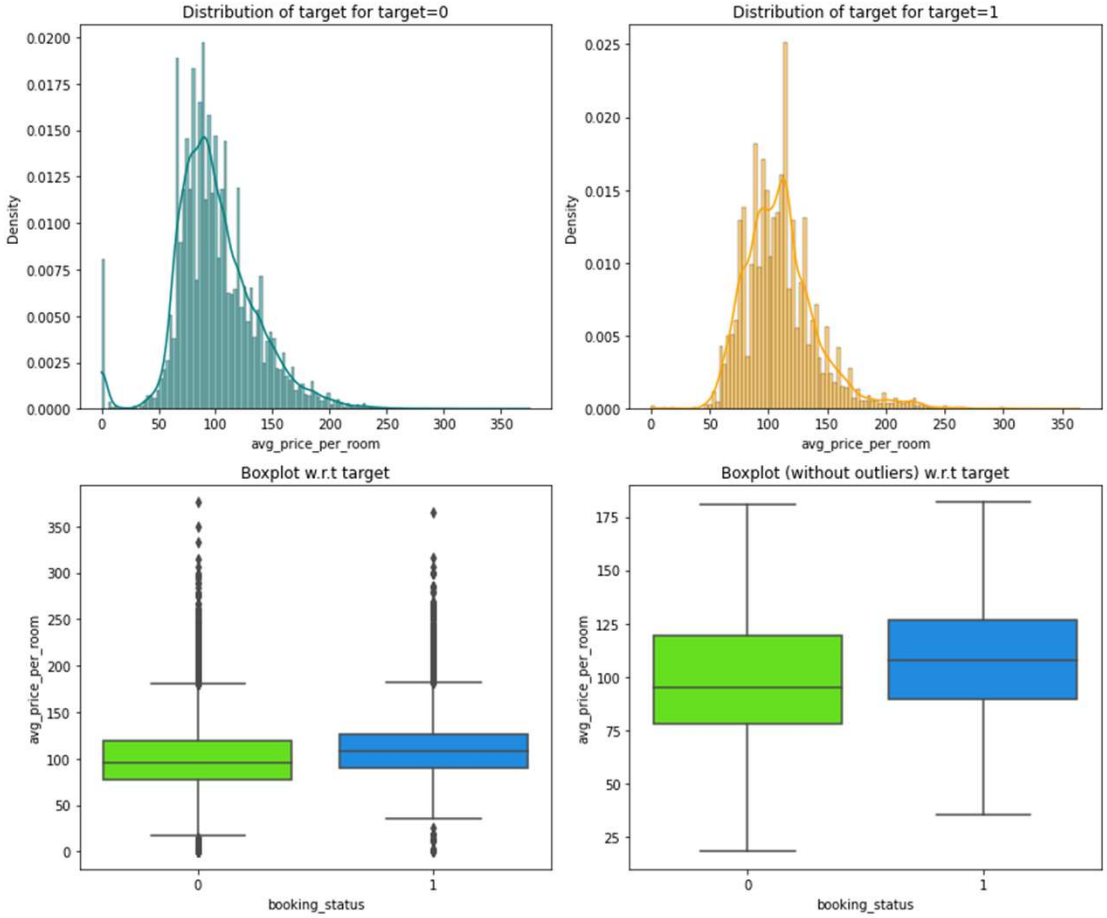
EDA Results – Bivariate, room price vs special requests

- The higher the number of special requests the booking has, the higher the room price is likely to be.



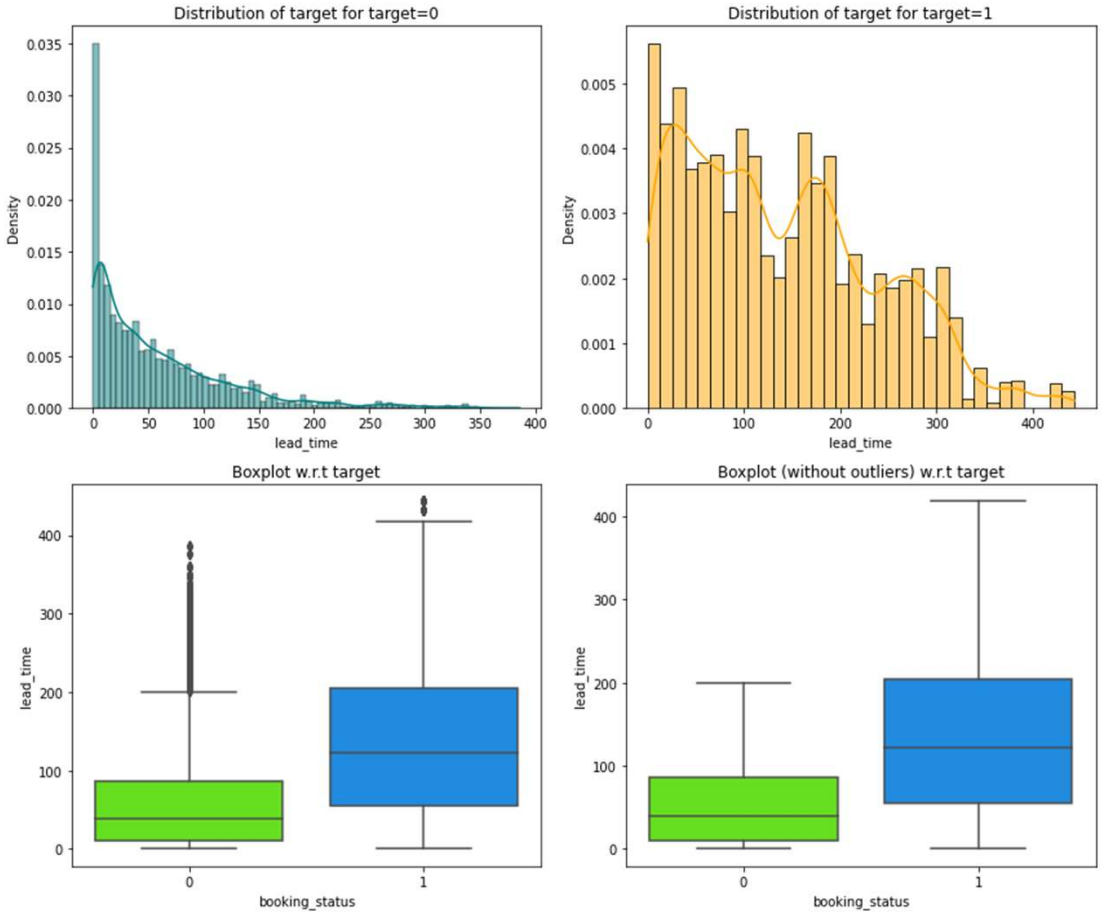
EDA Results – Bivariate, room price vs booking status

- Cancelations tend to happen at a higher average room rate.



EDA Results – Bivariate, lead time vs booking status

- Cancelations tend to happen at a larger lead time.
- Maybe people make tentative plans further out and less committed than those making the booking closer to the actual time of arrival.



EDA Results – Bivariate,

Generally people travel with their spouse and children for vacations or other activities. Let's create a new dataframe of the customers who traveled with their families and analyze the impact on booking status.

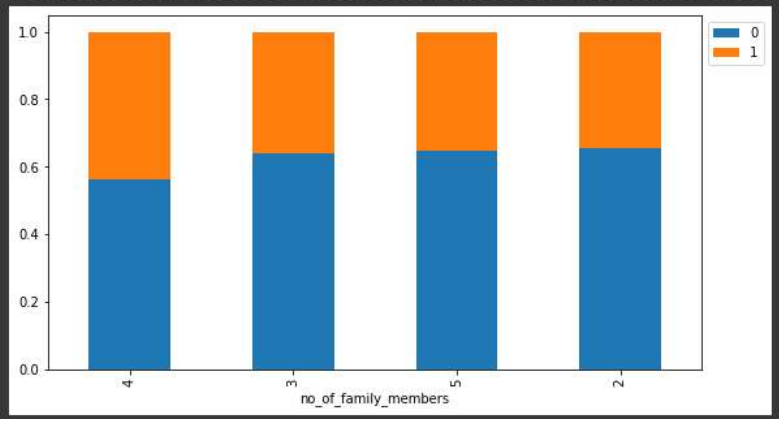
```
[57] family_data = data[(data["no_of_children"] >= 0) & (data["no_of_adults"] > 1)]
     family_data.shape
```

```
(28441, 18)
```

```
[58] family_data["no_of_family_members"] = (
         family_data["no_of_adults"] + family_data["no_of_children"]
     )
```

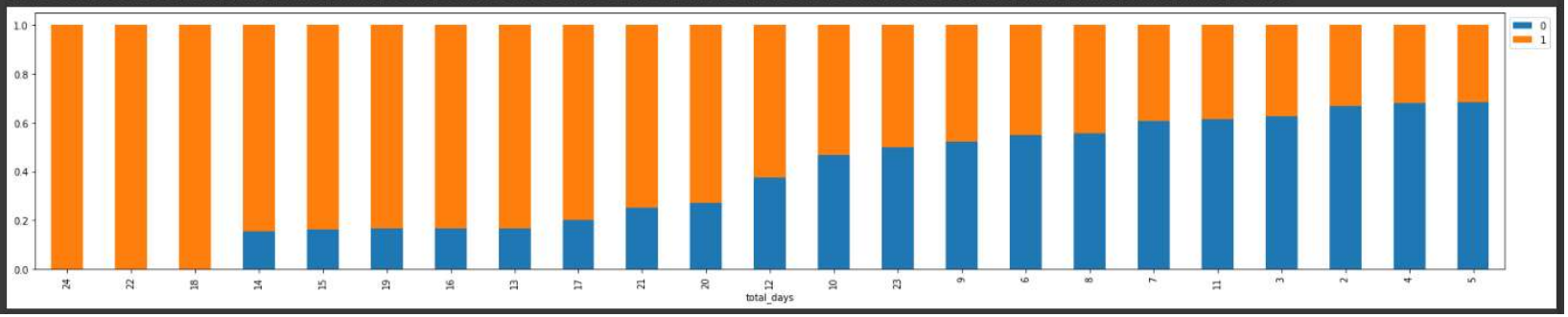
```
[59] stacked_barplot(family_data, "no_of_family_members", "booking_status") ## Complete the code to plot stacked barplot for no of
```

```
booking_status            0     1    All
no_of_family_members
All                   18456  9985  28441
2                     15506  8213  23719
3                      2425  1368   3793
4                       514   398    912
5                        11     6     17
```



EDA Results – Bivariate,

booking_status	0	1	All
total_days			
All	10979	6115	17094
3	3689	2183	5872
4	2977	1387	4364
5	1593	738	2331
2	1301	639	1940
6	566	465	1031
7	590	383	973
8	100	79	179
10	51	58	109
9	58	53	111
14	5	27	32
15	5	26	31
13	3	15	18
12	9	15	24
11	24	15	39
20	3	8	11
19	1	5	6
16	1	5	6
17	1	4	5
18	0	3	3
21	1	3	4
22	0	2	2
23	1	1	2
24	0	1	1

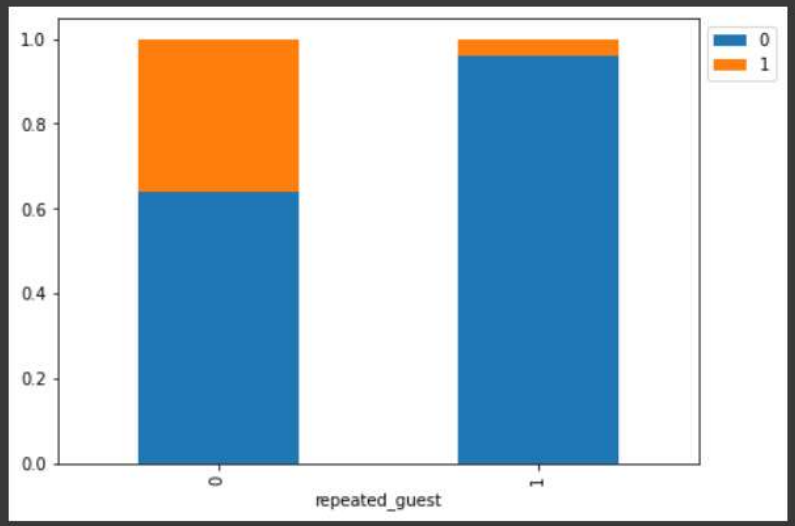


EDA Results – Bivariate,

Repeating guests are the guests who stay in the hotel often and are important to brand equity. Let's see what percentage of repeating guests cancel?

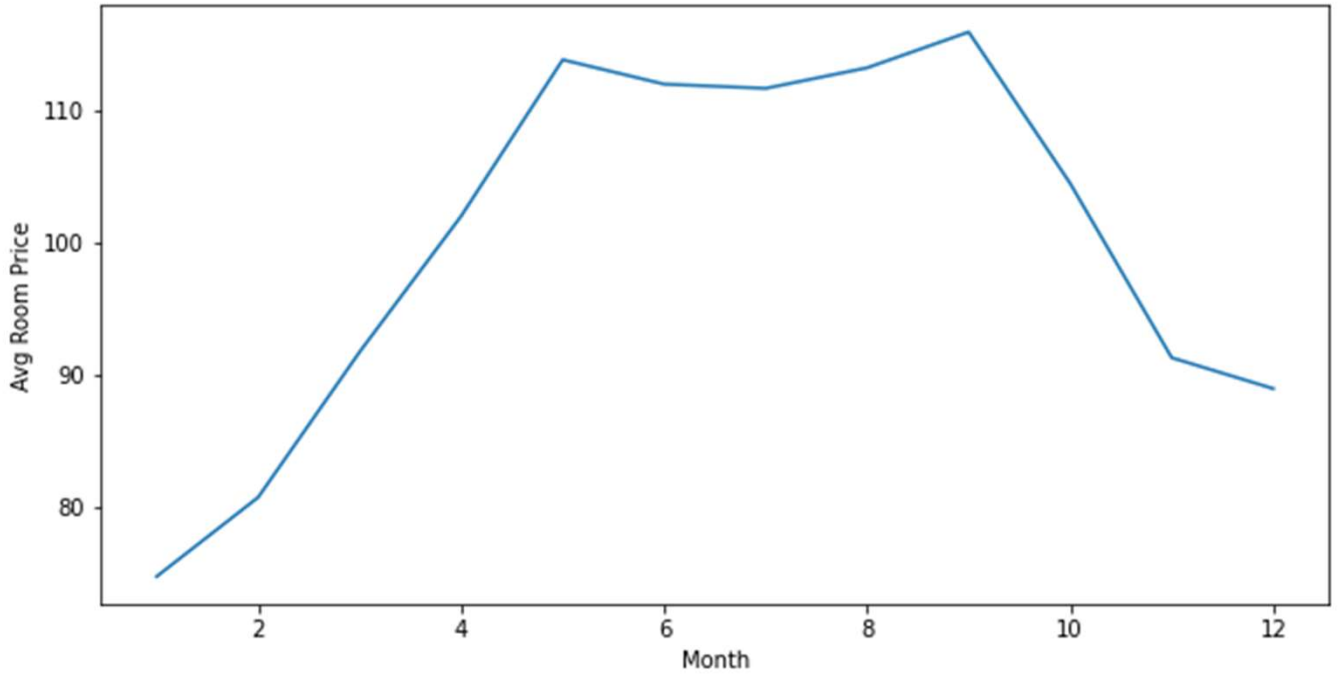
```
[ ] stacked_barplot(stay_data, "repeated_guest", "booking_status") ## Complete the code to plot stacked barplot for repeated guests and booking stat
```

repeated_guest	booking_status 0	1	All
All	10979	6115	17094
0	10812	6108	16920
1	167	7	174



EDA Results – Bivariate,

- Average price is up between May and September. This seems to correspond to higher number of bookings.

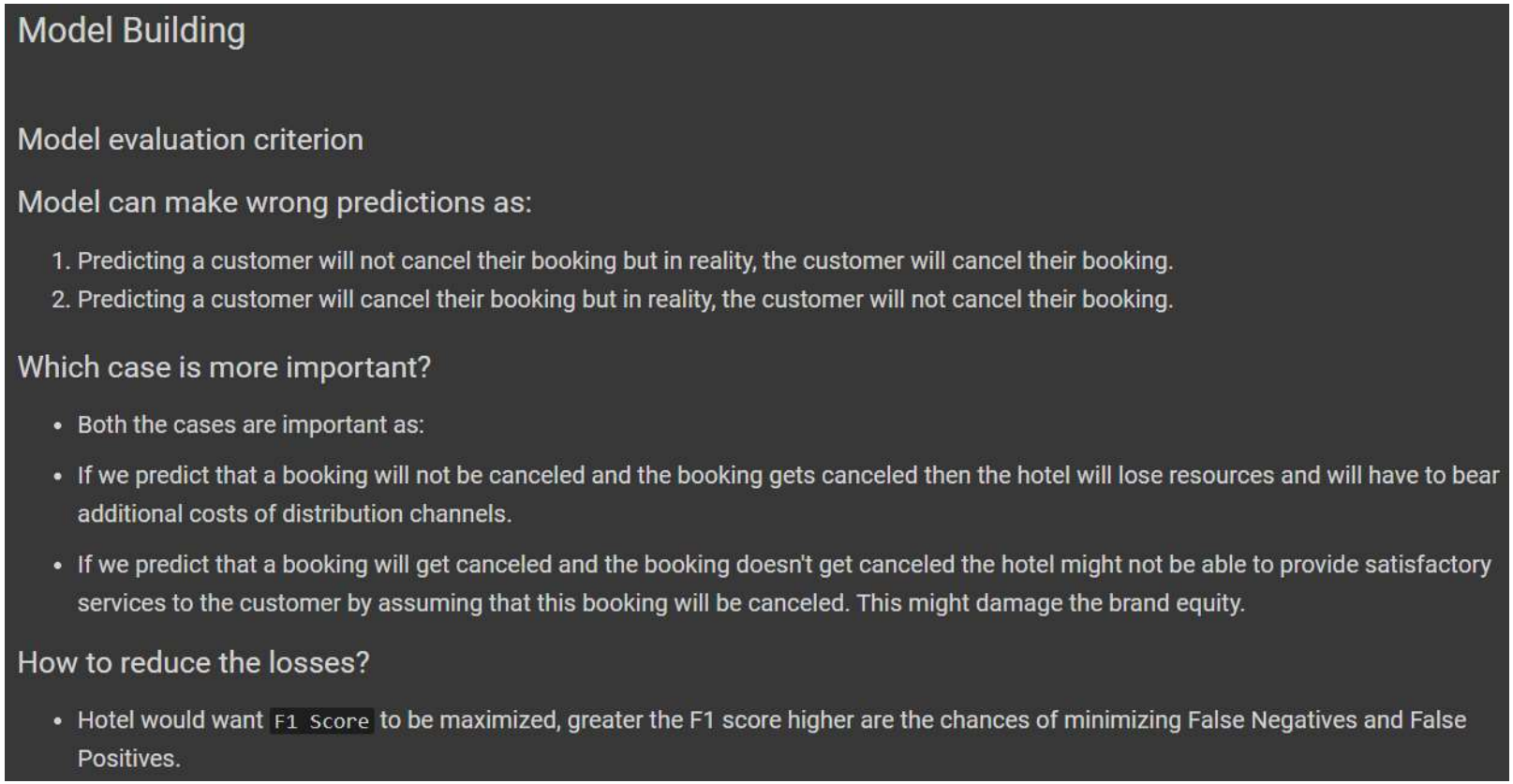


Model Building - Logistic Regression

- Please mention regarding the tests conducted to check the assumptions of Logistic Regression
- Interpret the results based on coefficients and odds
- Comment on the model performance

Model Building - Logistic Regression

- Plan to maximize **F1 Score**



Model Building

Model evaluation criterion

Model can make wrong predictions as:

1. Predicting a customer will not cancel their booking but in reality, the customer will cancel their booking.
2. Predicting a customer will cancel their booking but in reality, the customer will not cancel their booking.

Which case is more important?

- Both the cases are important as:
- If we predict that a booking will not be canceled and the booking gets canceled then the hotel will lose resources and will have to bear additional costs of distribution channels.
- If we predict that a booking will get canceled and the booking doesn't get canceled the hotel might not be able to provide satisfactory services to the customer by assuming that this booking will be canceled. This might damage the brand equity.

How to reduce the losses?

- Hotel would want `F1 Score` to be maximized, greater the F1 score higher are the chances of minimizing False Negatives and False Positives.

Model Building - Logistic Regression

- Objective is to predict which bookings will be canceled.
- So, we define ***booking_status*** as the **y, dependent variable, target variable**
- All other columns as **X, independent variables**.
- Added intercept to data.
- Used one hot encoding on categorical features to create dummy variables
- Split the data into train (70% of data) and test (30% of data).
- Fit the logistic regression model

```
Shape of Training set :  (25392, 28)
Shape of test set :  (10883, 28)
Percentage of classes in training set:
0   0.67064
1   0.32936
Name: booking_status, dtype: float64
Percentage of classes in test set:
0   0.67638
1   0.32362
Name: booking_status, dtype: float64
```

Model Building – Logistic Regression Results – Baseline, lg

- Built model for logistic regression using Logit and logit.fit to fit the logistic regression. ...lg
- Performance checked

Training performance:

	Accuracy	Recall	Precision	F1
0	0.80600	0.63410	0.73971	0.68285

- Multicolloinearity checked with VIF
- Dropped high p-values,
- Retreaved **selected_features**

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25364
Method:	MLE	Df Model:	27
Date:	Sun, 08 Jan 2023	Pseudo R-squ.:	0.3292
Time:	05:32:24	Log-Likelihood:	-10794.
converged:	False	LL-Null:	-16091.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-922.8266	120.832	-7.637	0.000	-1159.653	-686.000
no_of_adults	0.1137	0.038	3.019	0.003	0.040	0.188
no_of_children	0.1580	0.062	2.544	0.011	0.036	0.280
no_of_weekend_nights	0.1067	0.020	5.395	0.000	0.068	0.145
no_of_week_nights	0.0397	0.012	3.235	0.001	0.016	0.064
required_car_parking_space	-1.5943	0.138	-11.565	0.000	-1.865	-1.324
lead_time	0.0157	0.000	58.863	0.000	0.015	0.016
arrival_year	0.4561	0.060	7.617	0.000	0.339	0.573
arrival_month	-0.0417	0.006	-6.441	0.000	-0.054	-0.029
arrival_date	0.0005	0.002	0.259	0.796	-0.003	0.004
repeated_guest	-2.3472	0.617	-3.806	0.000	-3.556	-1.139
no_of_previous_cancellations	0.2664	0.086	3.108	0.002	0.098	0.434
no_of_previous_bookings_not_canceled	-0.1727	0.153	-1.131	0.258	-0.472	0.127
avg_price_per_room	0.0188	0.001	25.396	0.000	0.017	0.020
no_of_special_requests	-1.4689	0.030	-48.782	0.000	-1.528	-1.410
type_of_meal_plan_Meal Plan 2	0.1756	0.067	2.636	0.008	0.045	0.306
type_of_meal_plan_Meal Plan 3	17.3584	3987.873	0.004	0.997	-7798.729	7833.446
type_of_meal_plan_Not Selected	0.2784	0.053	5.247	0.000	0.174	0.382
room_type_reserved_Room_Type 2	-0.3605	0.131	-2.748	0.006	-0.618	-0.103
room_type_reserved_Room_Type 3	-0.0012	1.310	-0.001	0.999	-2.568	2.566
room_type_reserved_Room_Type 4	-0.2823	0.053	-5.304	0.000	-0.387	-0.178
room_type_reserved_Room_Type 5	-0.7189	0.209	-3.438	0.001	-1.129	-0.309
room_type_reserved_Room_Type 6	-0.9501	0.151	-6.274	0.000	-1.247	-0.653
room_type_reserved_Room_Type 7	-1.4003	0.294	-4.770	0.000	-1.976	-0.825
market_segment_type_Complementary	-40.5976	5.65e+05	-7.19e-05	1.000	-1.11e+06	1.11e+06
market_segment_type_Corporate	-1.1924	0.266	-4.483	0.000	-1.714	-0.671
market_segment_type_Offline	-2.1946	0.255	-8.621	0.000	-2.694	-1.696
market_segment_type_Online	-0.3995	0.251	-1.590	0.112	-0.892	0.093

Model Building – Logistic Regression Results – lg1

- Re-ran model using selected features, … **lg1**
- Performance checked

Training performance:

	Accuracy	Recall	Precision	F1
0	0.80545	0.63267	0.73907	0.68174

- The **f1_score** of the model is ~0.68174 … we will try to maximize **f1_score** further

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25370
Method:	MLE	Df Model:	21
Date:	Sun, 08 Jan 2023	Pseudo R-squ.:	0.3282
Time:	05:32:28	Log-Likelihood:	-10810.
converged:	True	LL-Null:	-16091.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-915.6391	120.471	-7.600	0.000	-1151.758	-679.520
no_of_adults	0.1088	0.037	2.914	0.004	0.036	0.182
no_of_children	0.1531	0.062	2.470	0.014	0.032	0.275
no_of_weekend_nights	0.1086	0.020	5.498	0.000	0.070	0.147
no_of_week_nights	0.0417	0.012	3.399	0.001	0.018	0.066
required_car_parking_space	-1.5947	0.138	-11.564	0.000	-1.865	-1.324
lead_time	0.0157	0.000	59.213	0.000	0.015	0.016
arrival_year	0.4523	0.060	7.576	0.000	0.335	0.569
arrival_month	-0.0425	0.006	-6.591	0.000	-0.055	-0.030
repeated_guest	-2.7367	0.557	-4.916	0.000	-3.828	-1.646
no_of_previous_cancellations	0.2288	0.077	2.983	0.003	0.078	0.379
avg_price_per_room	0.0192	0.001	26.336	0.000	0.018	0.021
no_of_special_requests	-1.4698	0.030	-48.884	0.000	-1.529	-1.411
type_of_meal_plan_Meal Plan 2	0.1642	0.067	2.469	0.014	0.034	0.295
type_of_meal_plan_Not Selected	0.2860	0.053	5.406	0.000	0.182	0.390
room_type_reserved_Room_Type 2	-0.3552	0.131	-2.709	0.007	-0.612	-0.098
room_type_reserved_Room_Type 4	-0.2828	0.053	-5.330	0.000	-0.387	-0.179
room_type_reserved_Room_Type 5	-0.7364	0.208	-3.535	0.000	-1.145	-0.328
room_type_reserved_Room_Type 6	-0.9682	0.151	-6.403	0.000	-1.265	-0.672
room_type_reserved_Room_Type 7	-1.4343	0.293	-4.892	0.000	-2.009	-0.860
market_segment_type_Corporate	-0.7913	0.103	-7.692	0.000	-0.993	-0.590
market_segment_type_Offline	-1.7854	0.052	-34.363	0.000	-1.887	-1.684

Model Building – Logistic Regression – coeff. to odds

- Converted coefficients to odds
- **Negative** values of the **coefficient** show that the **probability of a cancelation happening** will **decrease** with the **increase** of the corresponding attribute value.
- **Positive** values of the **coefficient** show that the **probability of a cancelation happening** will **increase** with the **increase** of the corresponding attribute value.
- p-value of a variable indicates if the variable is significant or not. If we consider the significance level to be 0.05 (5%), then any variable with a p-value less than 0.05 would be considered significant.

Model Building – Logistic Regression – Odds

- Odds

	const	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	required_car_parking_space	lead_time	arrival_year	arrival_month	repeated_guest
Odds	0.00000	1.11491	1.16546	1.11470	1.04258	0.20296	1.01583	1.57195	0.95839	0.06478
Change_odd%	-100.00000	11.49096	16.54593	11.46966	4.25841	-79.70395	1.58331	57.19508	-4.16120	-93.52180

no_of_previous_cancellations	avg_price_per_room	no_of_special_requests	type_of_meal_plan_Meal Plan 2	type_of_meal_plan_Not Selected	room_type_reserved_Room_Type 2	room_type_reserved_Room_Type 4
1.25712	1.01937	0.22996	1.17846	1.33109	0.70104	0.75364
25.71181	1.93684	-77.00374	17.84641	33.10947	-29.89588	-24.63551

room_type_reserved_Room_Type 5	room_type_reserved_Room_Type 6	room_type_reserved_Room_Type 7	market_segment_type_Corporate	market_segment_type_Offline
0.47885	0.37977	0.23827	0.45326	0.16773
-52.11548	-62.02290	-76.17294	-54.67373	-83.22724

Model Building – Logistic Regression – Odds

- Odds comments

Coefficient interpretations

- `lead_time`: Holding all other features constant a 1 unit change in lead_time will increase the odds of a booking cancelation by ~1.02 times or an increase of ~58%.
- `avg_price_per_room`: Holding all other features constant a 1 unit change in avg_price_per_room will increase the odds of a booking cancelation by ~1.02 times or an increase of ~94%.
- `no_of_special_requests`: Holding all other features constant a 1 unit change in no_of_special_requests will decrease the odds of a booking cancelation by ~0.22 times or an increase of ~77%.

Interpretation for other attributes can be done similarly.

Model Performance Evaluation and Improvement - Logistic Regression

- Please comment on the improvement in the model performance by changing the classification threshold

Model Performance Evaluation and Improvement - Logistic Regression

- Checked model performance on **training** set
- ROC-AUC on training set
- Model Performance Improvement

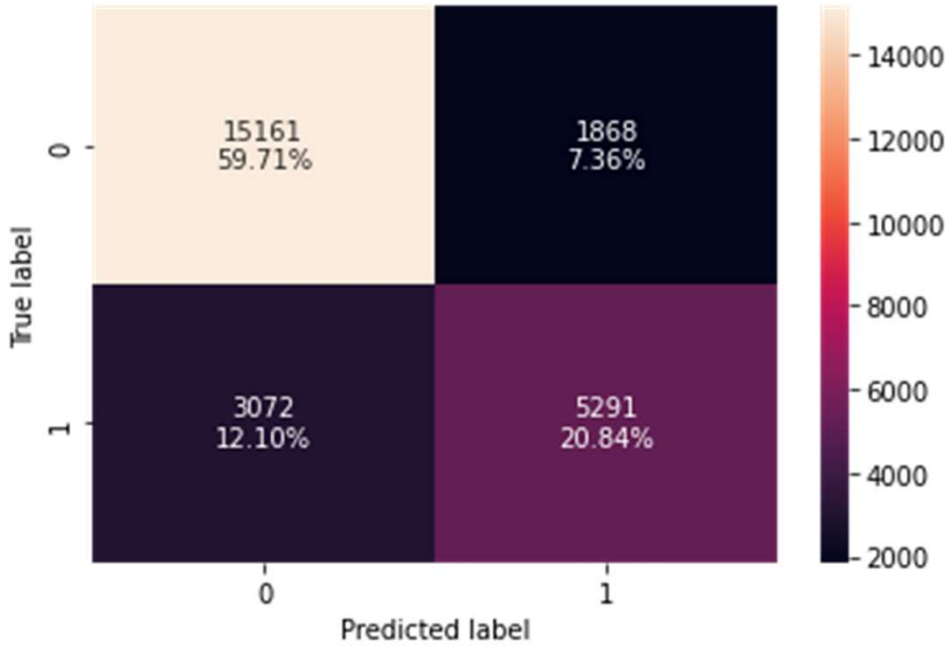
- Optimal threshold using AUC-ROC curve
- Used Precision-Recall curve and see if we can find a better threshold

- Checked model performance on **training** set
- Checked model performance on **test** set
- Model performance summary

Model Performance Evaluation and Improvement - Logistic Regression

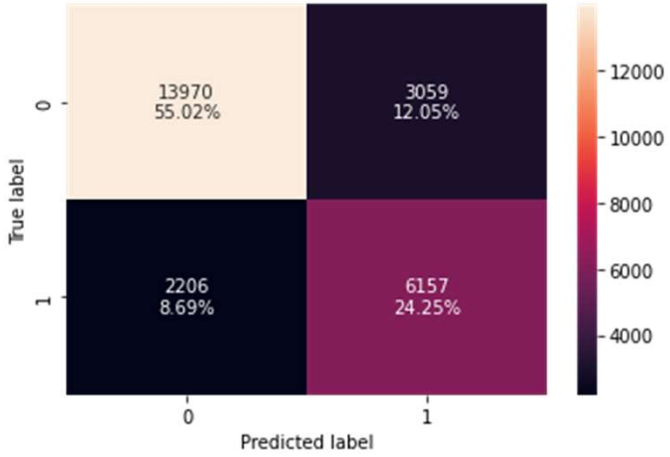
- Checked model performance on **training** set

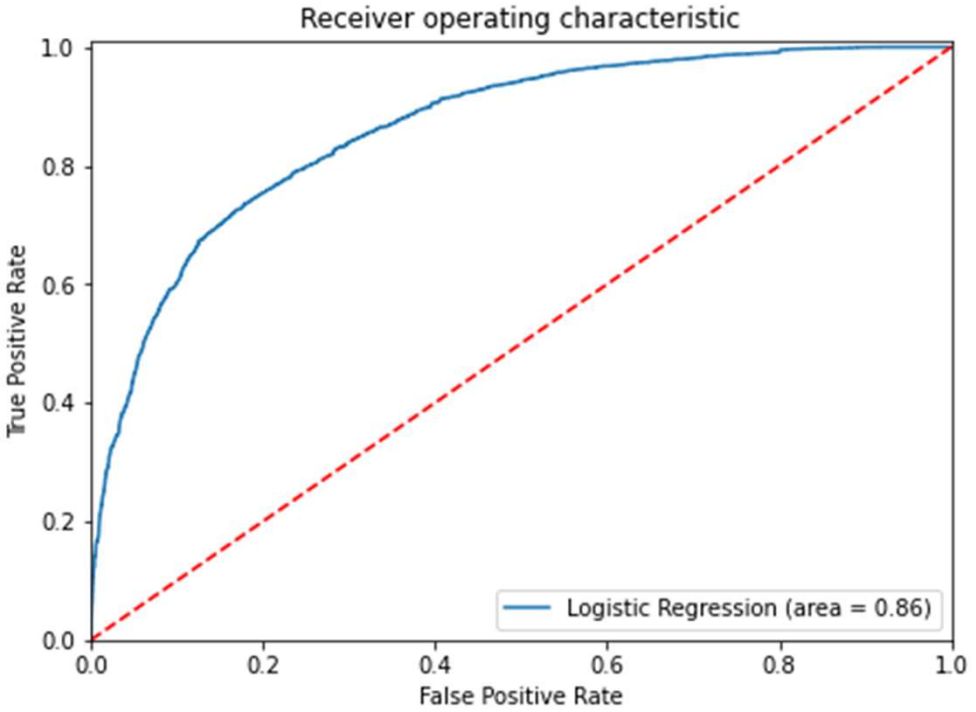
- TP = ~60%
- FP = ~ 7%
- FN = ~12%
- TN = ~21%



Model Performance Evaluation and Improvement - Logistic Regression

- ROC-AUC on training set
- Receiver Operating Characteristics
- Area Under the Curve = 0.86
- optimal_threshold_**auc_roc =~ 0.37**
  0.3700522558707859
- **F1 score better = 0.70049**



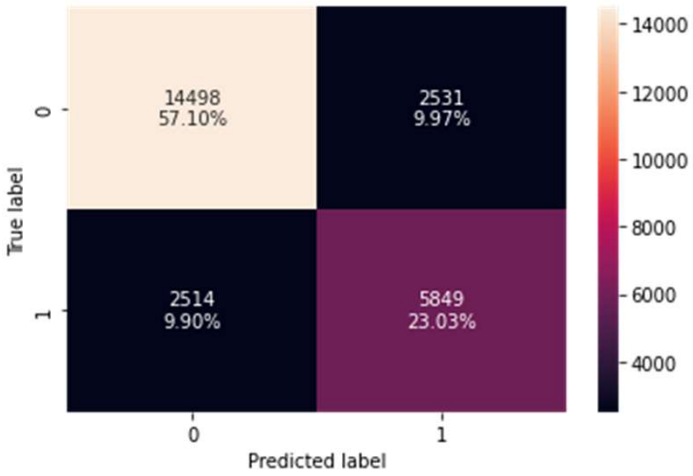


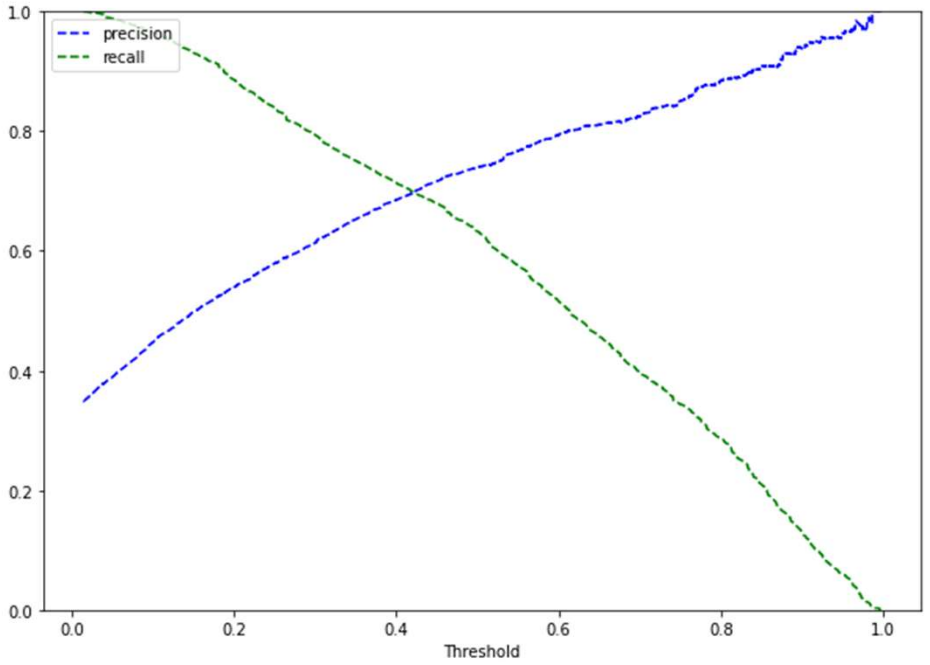
Training performance:

	Accuracy	Recall	Precision	F1
0	0.79265	0.73622	0.66808	0.70049

Model Performance Evaluation and Improvement - Logistic Regression

- Precision-Recall curve
- optimal_threshold_**curve = 0.42**
- **F1 score = 0.69868**
- Accuracy is better, recall and precision is more balanced , but F1 is not as high.



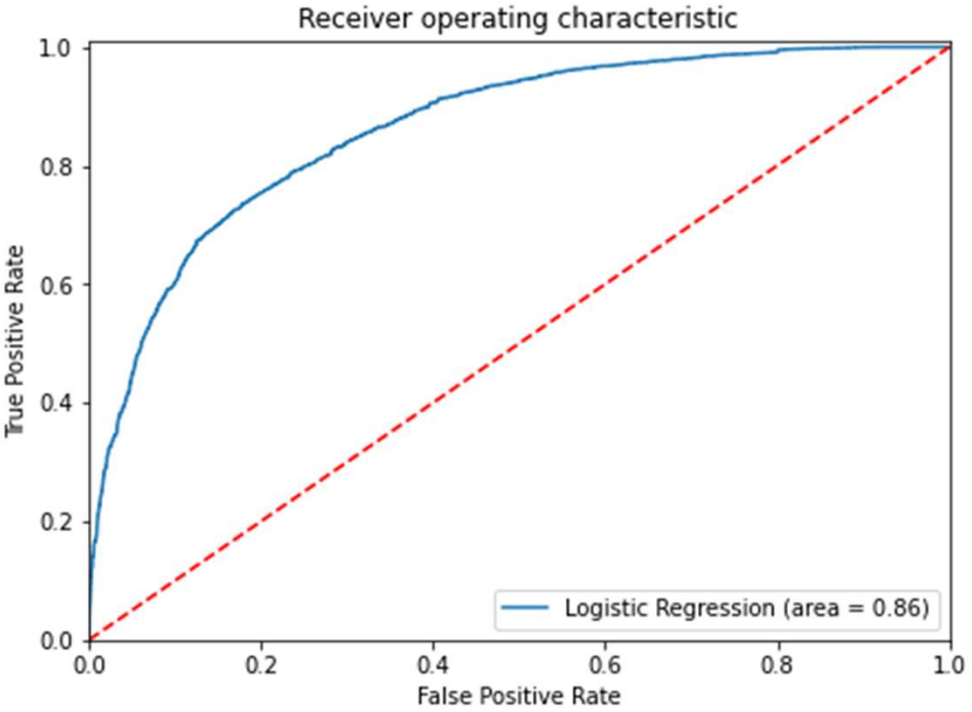


Training performance:

	Accuracy	Recall	Precision	F1
0	0.80132	0.69939	0.69797	0.69868

Model Performance Evaluation and Improvement - Logistic Regression

- ROC-AUC on **test** set
- Receiver Operating Characteristics
- Area Under the Curve = 0.86
- optimal_threshold_**auc_roc** **=~ 0.37**
  0.3700522558707859
- **F1 score for test = 0.70074**

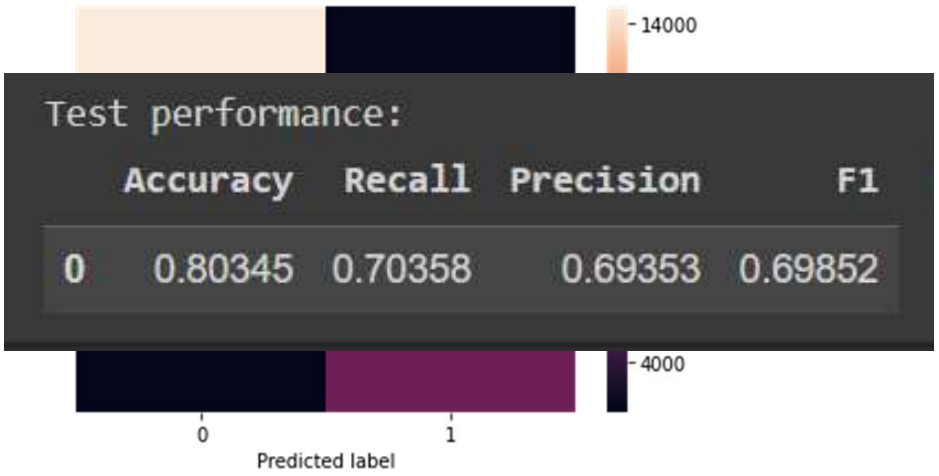


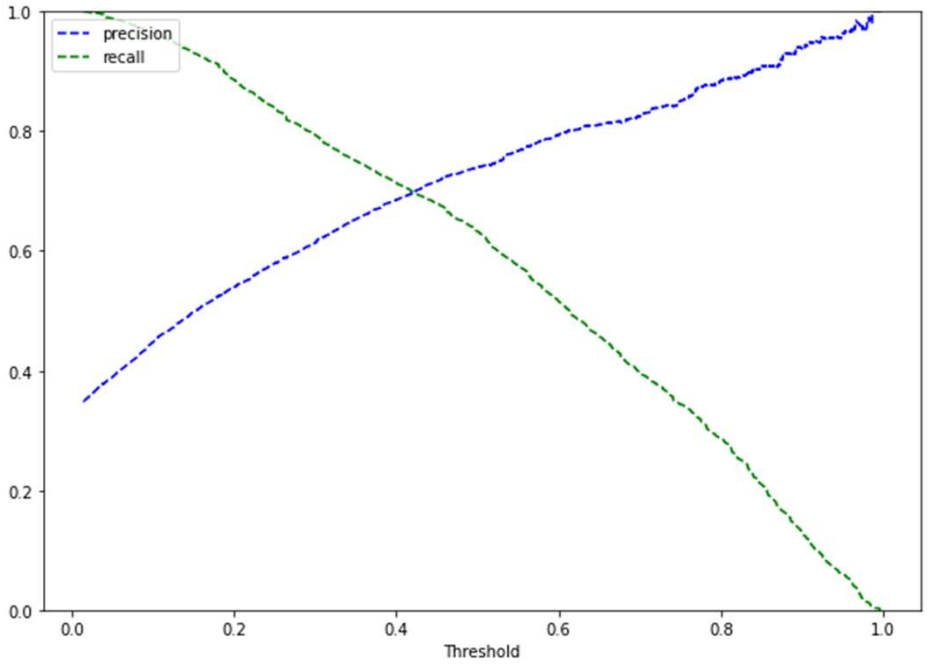
Test performance:

	Accuracy	Recall	Precision	F1
0	0.79555	0.73964	0.66573	0.70074

Model Performance Evaluation and Improvement - Logistic Regression

- **test**
- optimal_threshold_**curve = 0.42**
- **F1 score = 0.69852**
- So not better





Model Performance Evaluation and Improvement - Logistic Regression

- Model performance summary

Training performance comparison:

	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80545	0.79265	0.80132
Recall	0.63267	0.73622	0.69939
Precision	0.73907	0.66808	0.69797
F1	0.68174	0.70049	0.69868

Model Performance Evaluation and Improvement - Logistic Regression

- Model performance summary
- **Looking to optimize F1 so 0.70049 is best here using a threshold of 0.37**

Training performance comparison:

	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80545	0.79265	0.80132
Recall	0.63267	0.73622	0.69939
Precision	0.73907	0.66808	0.69797
F1	0.68174	0.70049	0.69868

Model Building - Decision Tree

- Please mention the model building steps of Decision Tree
- Comment on the model performance

Model Building – Decision Tree – Steps

- Decision Tree
- Data Preparation for modeling (Decision Tree)
- The model_performance_classification_sklearn function is used to check the model performance of models.
- Building Decision Tree Model
- Checking model performance on training set
- Checking model performance on test set
- Pruning the tree
- Checking performance on training set
- Checking performance on test set
- Visualizing the Decision Tree
- F1 Score vs alpha for training and testing sets
- Checking performance on training set
- Checking performance on test set
- Comparing Decision Tree models

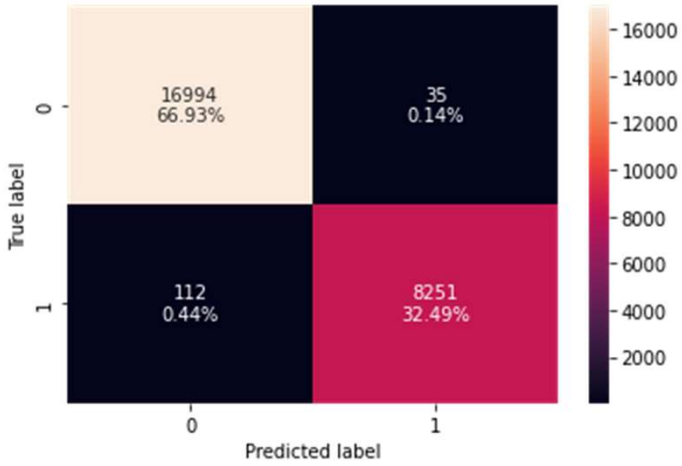
Model Building – Decision Tree – Steps

- Objective is to predict which bookings will be canceled.
- So, we define ***booking_status*** as the **y, dependent variable, target variable**
- All other columns as **X, independent variables**.
- Used one hot encoding on categorical features to create dummy variables
- Split the data into train (70% of data) and test (30% of data).
- Built a model named ***model*** using ***DecisionTreeClassifier***
- Then fit the model on the decision tree

```
Shape of Training set :  (25392, 27)
Shape of test set :  (10883, 27)
Percentage of classes in training set:
0   0.67064
1   0.32936
Name: booking_status, dtype: float64
Percentage of classes in test set:
0   0.67638
1   0.32362
Name: booking_status, dtype: float64
```

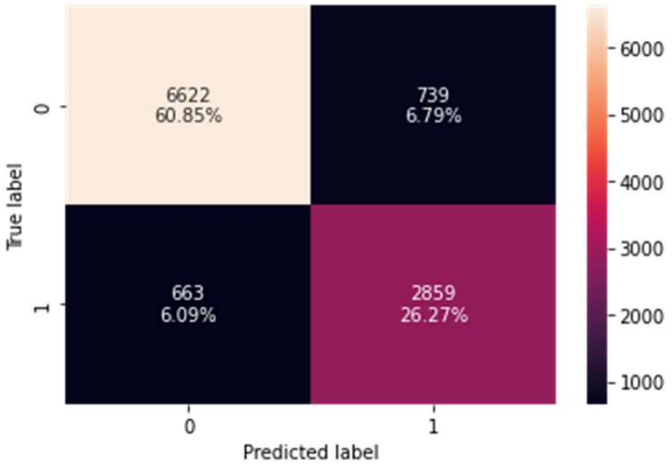
Model Building – Decision Tree

- Checked the model performance on the training set, which looks to be overfit.
- Train



	Accuracy	Recall	Precision	F1
0	0.99421	0.98661	0.99578	0.99117

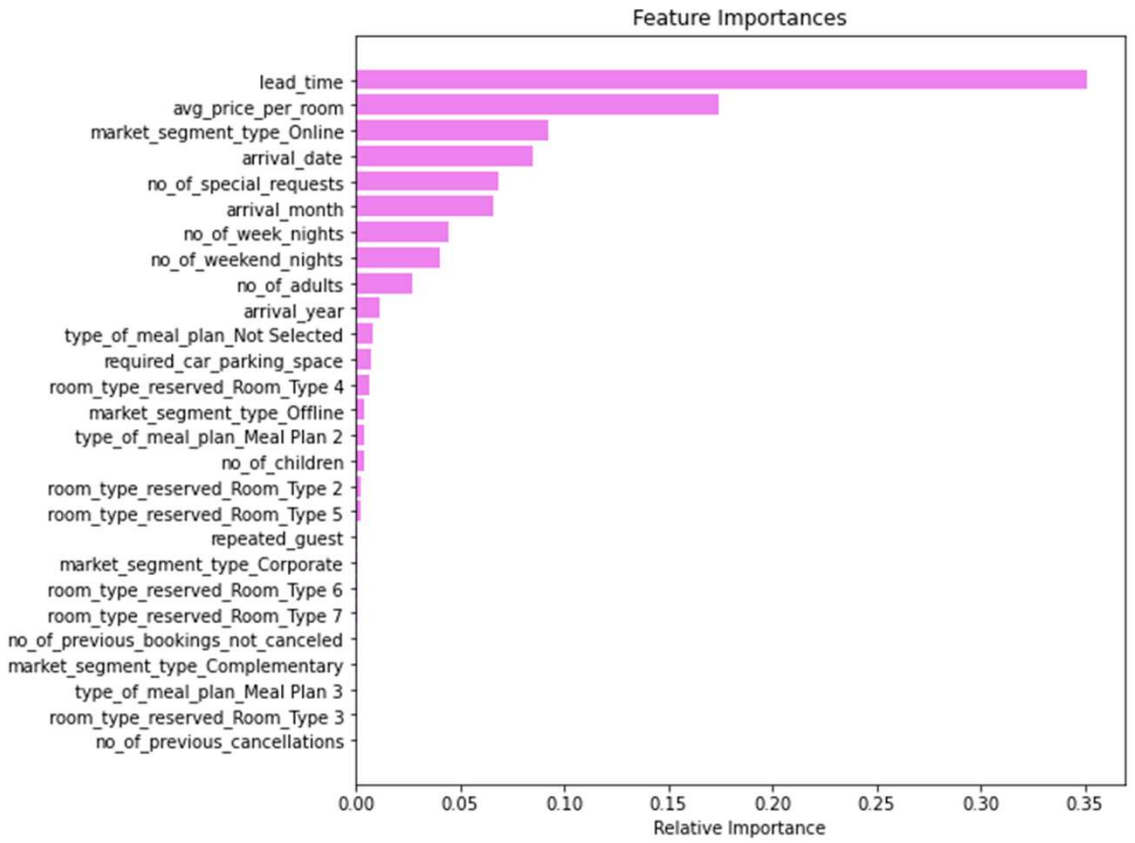
Test



	Accuracy	Recall	Precision	F1
0	0.87118	0.81175	0.79461	0.80309

Model Building – Decision Tree

- Feature Importance **before pruning**
- The most important feature seams to be **lead_time** ~**0.35** that is almost twice as important as the second most important, **average price per room**. ~**0.18**
- Market_segment type_online ~**0.10**
- Arival_date ~ **0.09**
- Number of special requests
- And so on...



Model Performance Evaluation and Improvement - Decision Tree

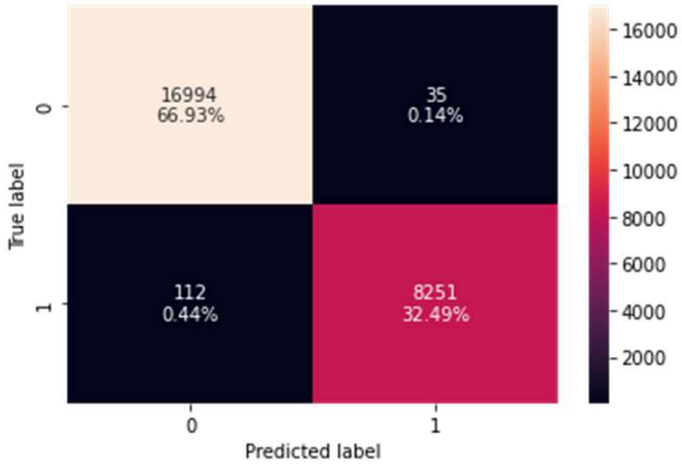
- Please comment on the improvement in the model performance by trying the different pruning techniques
- Please mention the decision rules and check the feature importance

Model Performance Evaluation and Improvement - Decision Tree

- Performed some pre-pruning by limiting the max depth, max leaf nodes, and minimum samples split
- Used **GridSearchCV** to find the best hyper parameters and giving the model a scorer of **f1_score** to use to optimize.
- Results:
    - DecisionTreeClassifier(class_weight='**balanced**', **max_depth=6, max_leaf_nodes=50, min_samples_split=10**, random_state=1)

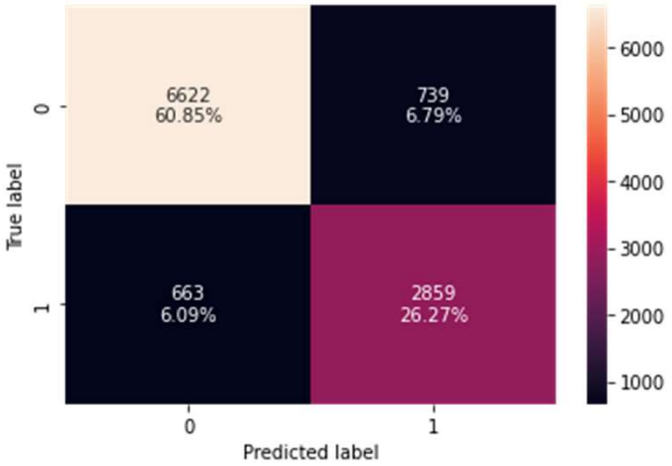
Model Building – Decision Tree

- Checked the model performance on the training set, which looks to be same?
- Train



	Accuracy	Recall	Precision	F1
0	0.99421	0.98661	0.99578	0.99117

Test



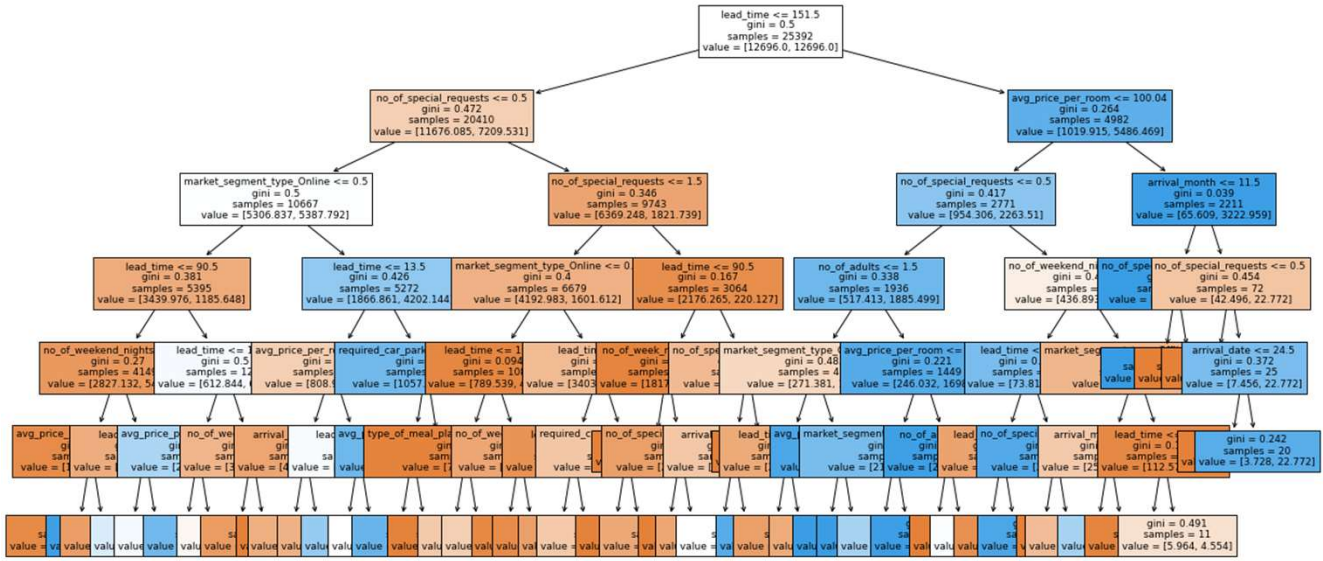
	Accuracy	Recall	Precision	F1
0	0.87118	0.81175	0.79461	0.80309

Model Building – Decision Tree

- Visualized tree and may need to be re-ran or change parameters?

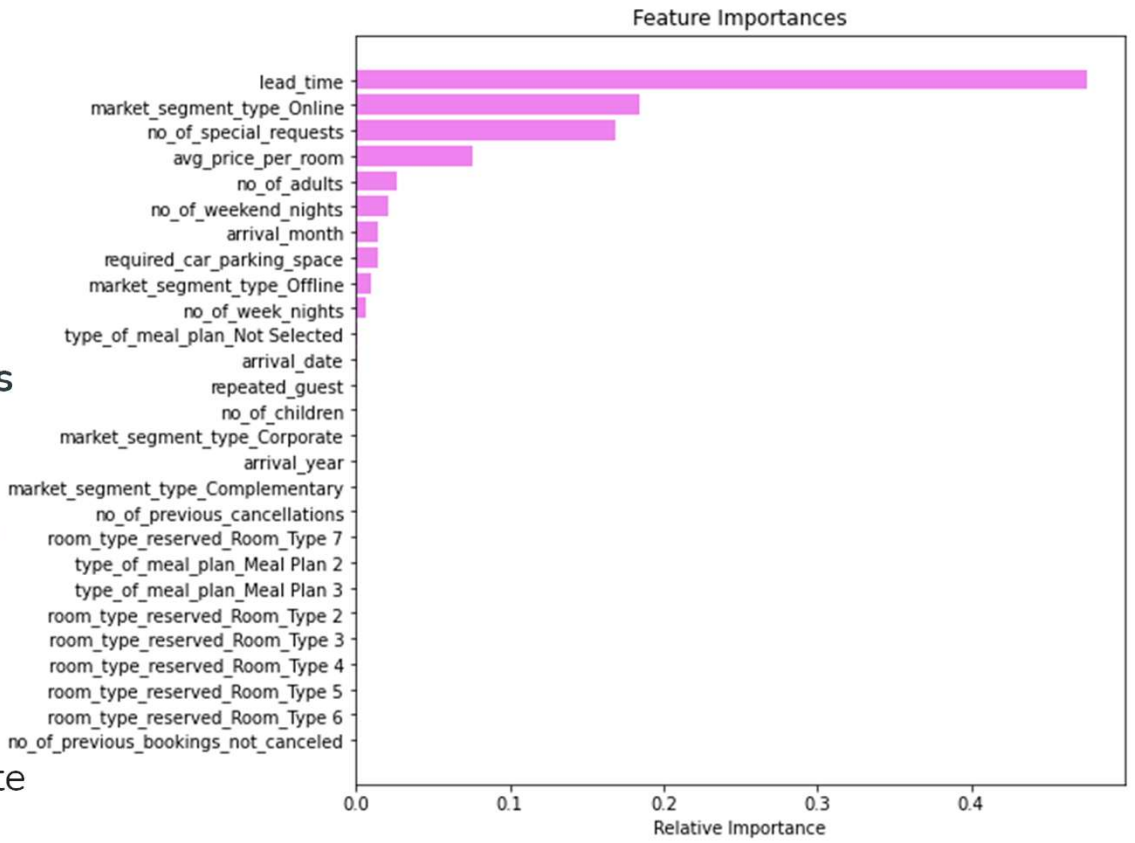
```
# Text report showing the rules of a decision tree -
print(tree.export_text(estimator, feature_names=feature_names, show_weights=True))
```

```
|--- lead_time <= 151.50
|   |--- no_of_special_requests <= 0.50
|   |   |--- market_segment_type_Online <= 0.50
|   |   |   |--- lead_time <= 90.50
|   |   |   |   |--- no_of_weekend_nights <= 0.50
|   |   |   |   |   |--- avg_price_per_room <= 196.50
|   |   |   |   |   |   |--- weights: [1736.39, 133.59] class: 0
|   |   |   |   |   |--- avg_price_per_room >  196.50
|   |   |   |   |   |   |--- weights: [0.75, 24.29] class: 1
|   |   |   |   |--- no_of_weekend_nights >  0.50
|   |   |   |   |   |--- lead_time <= 68.50
|   |   |   |   |   |   |--- weights: [960.27, 223.16] class: 0
|   |   |   |   |   |--- lead_time >  68.50
|   |   |   |   |   |   |--- weights: [129.73, 160.92] class: 1
|   |   |   |--- lead_time >  90.50
|   |   |   |   |--- lead_time <= 117.50
|   |   |   |   |   |--- avg_price_per_room <= 93.58
|   |   |   |   |   |   |--- weights: [214.72, 227.72] class: 1
|   |   |   |   |   |--- avg_price_per_room >  93.58
|   |   |   |   |   |   |--- weights: [82.76, 285.41] class: 1
|   |   |   |   |--- lead_time >  117.50
|   |   |   |   |   |--- no_of_week_nights <= 1.50
|   |   |   |   |   |   |--- weights: [87.23, 81.98] class: 0
|   |   |   |   |   |--- no_of_week_nights >  1.50
|   |   |   |   |   |   |--- weights: [228.14, 48.58] class: 0
|   |   |--- market_segment_type_Online >  0.50
|   |   |   |--- lead_time <= 13.50
|   |   |   |   |--- avg_price_per_room <= 99.44
|   |   |   |   |   |--- arrival_month <= 1.50
|   |   |   |   |   |   |--- weights: [92.45, 0.00] class: 0
|   |   |   |   |   |--- arrival_month >  1.50
|   |   |   |   |   |   |--- weights: [363.83, 132.08] class: 0
|   |   |   |   |--- avg_price_per_room >  99.44
|   |   |   |   |   |--- lead_time <= 3.50
|   |   |   |   |   |   |--- weights: [219.94, 85.01] class: 0
|   |   |   |   |   |--- lead_time >  3.50
|   |   |   |   |   |   |--- weights: [132.71, 280.85] class: 1
|   |   |   |--- lead_time >  13.50
```



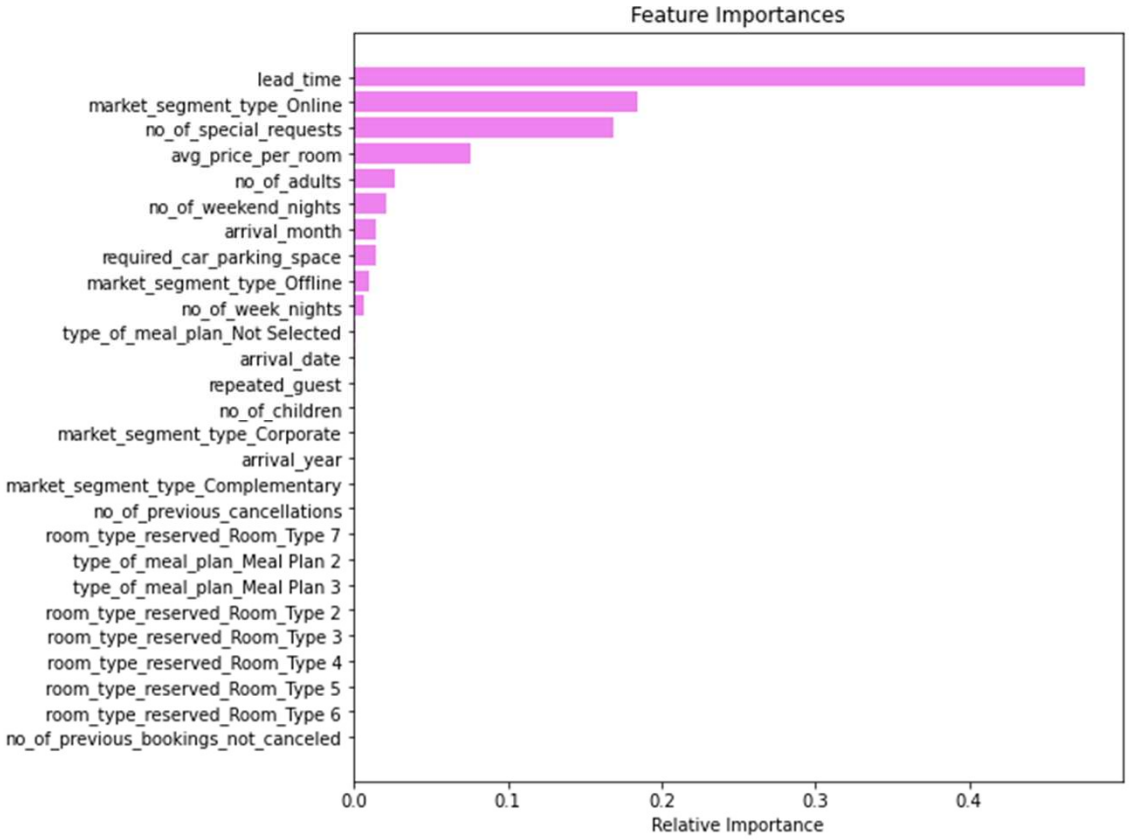
Model Building – Decision Tree

- Feature Importance **after pruning**
- **lead_time ~0.50** becomes even more pronounced at first (1st) place
- **Market_segment type_online ~0.20** moves up in rank to second (2nd)
- Followed by **Number of special requests ~0.20** moving up two spots to third (3rd)
- **average price per room ~0.08** decreases to fourth (4th) most important
- And so on… lead_time, avg_price_per_room, market_segment_type_online, arrival_date



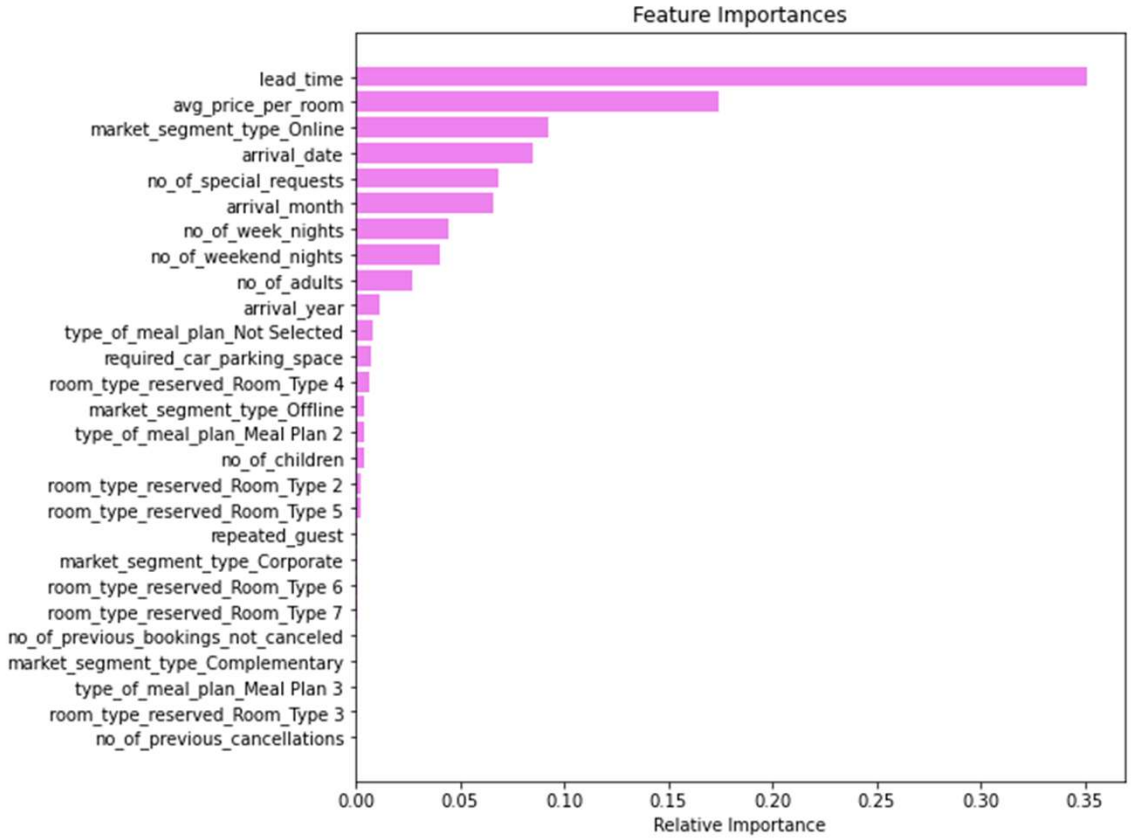
Model Building – Decision Tree

- Toggle after



Model Building – Decision Tree

- Toggle before

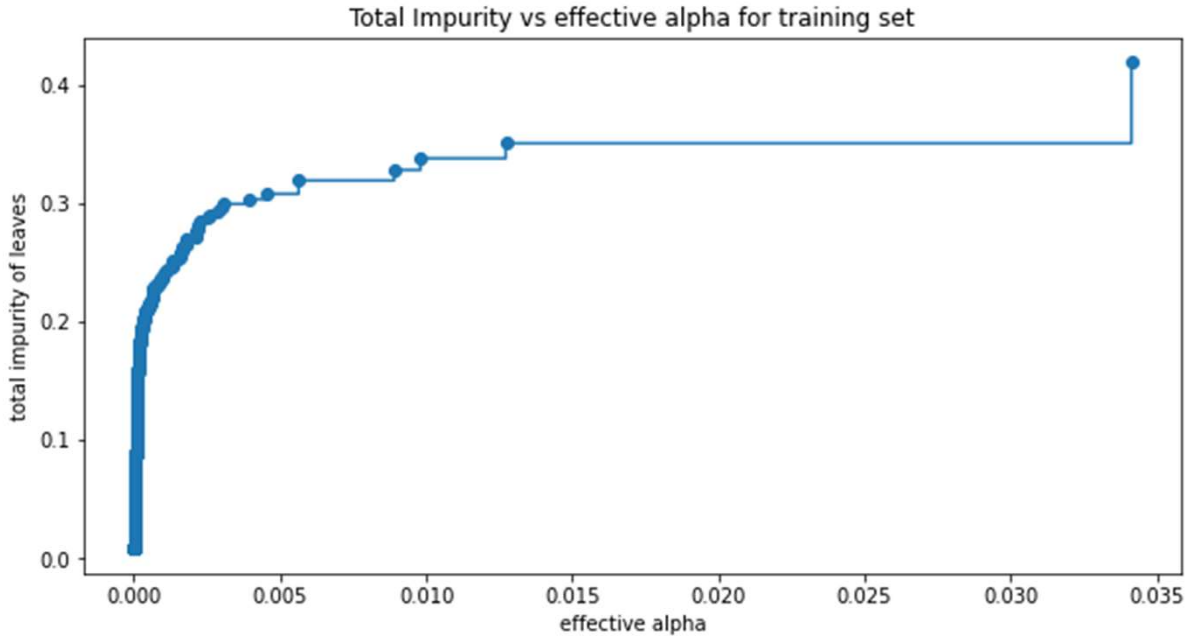


Model Building – Decision Tree

- Cost Complexity Pruning
- Checked ccp alphas

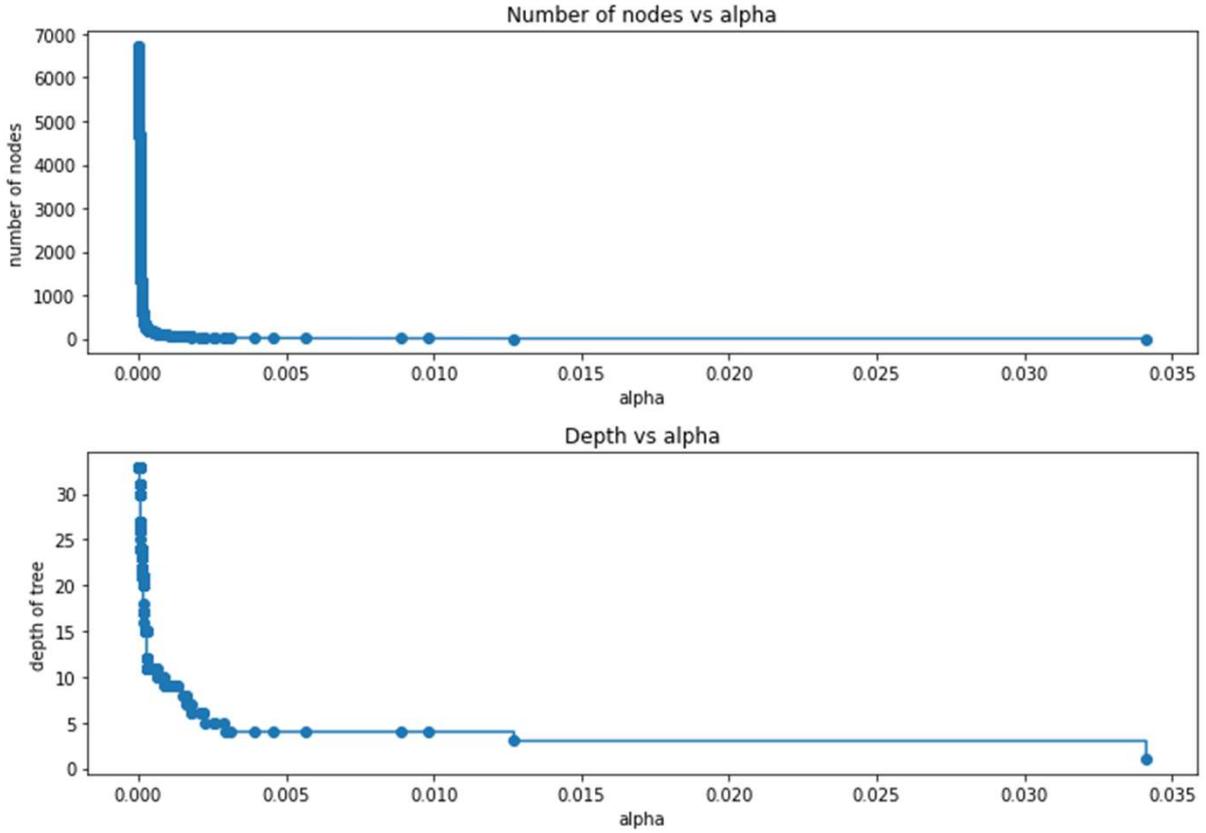
	ccp_alphas	impurities
0	0.00000	0.00838
1	0.00000	0.00838
2	0.00000	0.00838
3	0.00000	0.00838
4	0.00000	0.00838
...	...	...
1839	0.00890	0.32806
1840	0.00980	0.33786
1841	0.01272	0.35058
1842	0.03412	0.41882
1843	0.08118	0.50000

1844 rows × 2 columns



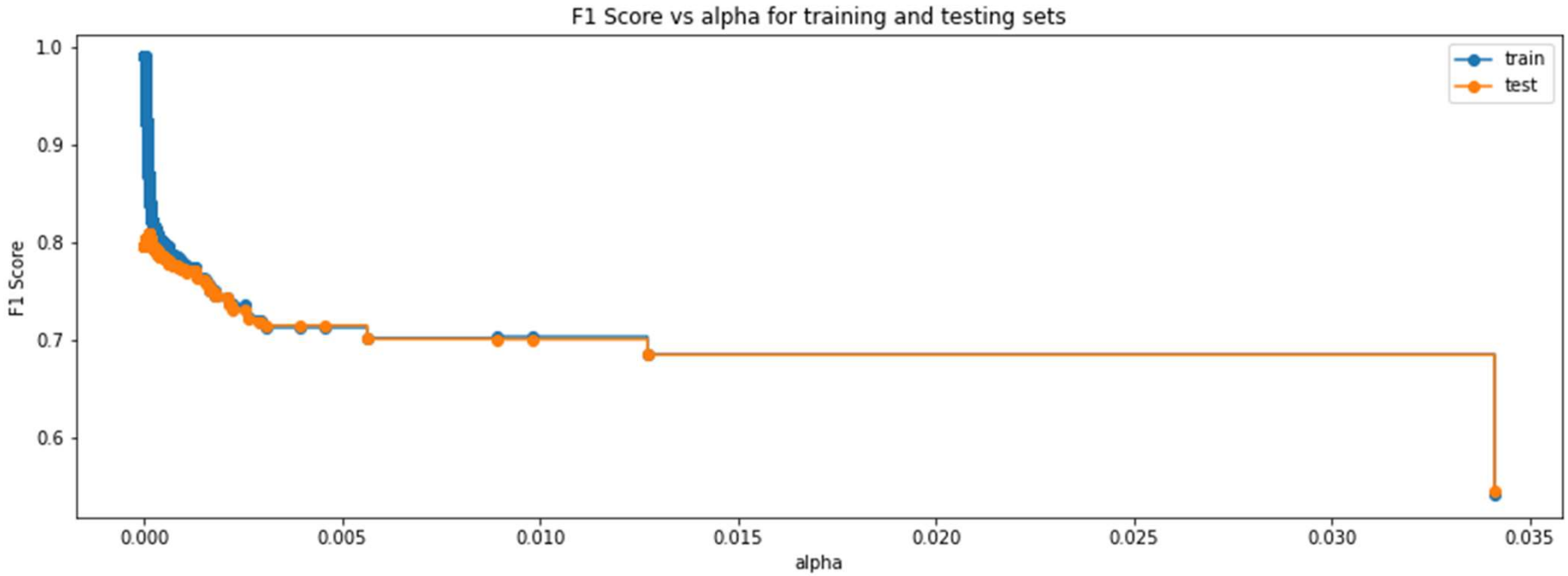
Model Building – Decision Tree

- Cost Complexity Pruning
- Trained a decision tree using effective alphas
- Number of nodes in the last tree is: 1 with ccp_alpha: 0.0811791438913696
- F1 Score vs alpha for training and testing sets



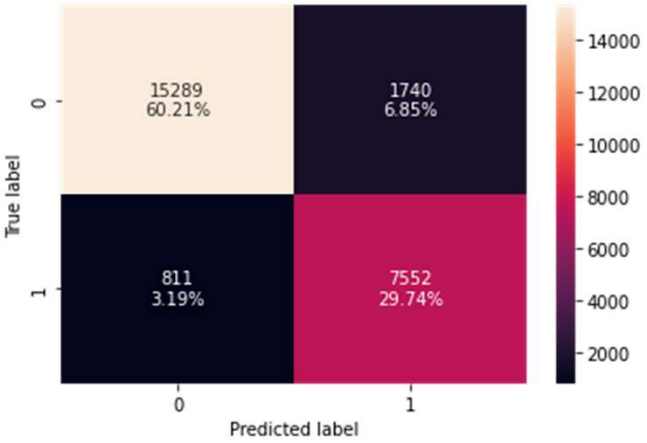
Model Building – Decision Tree

- F1 Score vs alpha for training and testing sets
- DecisionTreeClassifier(ccp_alpha=0.00012267633155167043, class_weight='balanced', random_state=1)



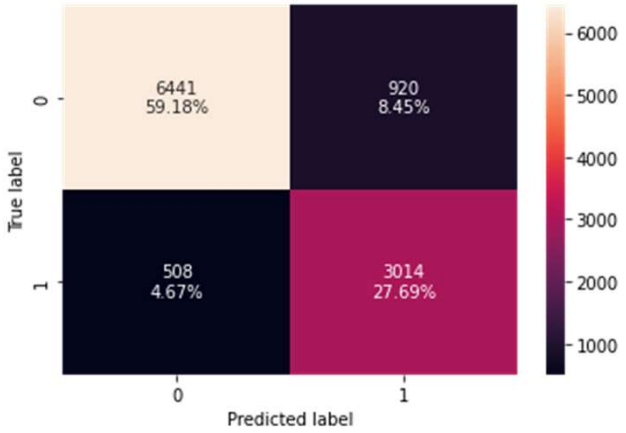
Model Building – Decision Tree

- Checking performances again for training and testing sets… **Not overfitting!**
- Since **F1 score** is within **five %** between **train ~0.8555** and **test ~0.8085** we can say that the model is not over fitting
- Training



	Accuracy	Recall	Precision	F1
0	0.89954	0.90303	0.81274	0.85551

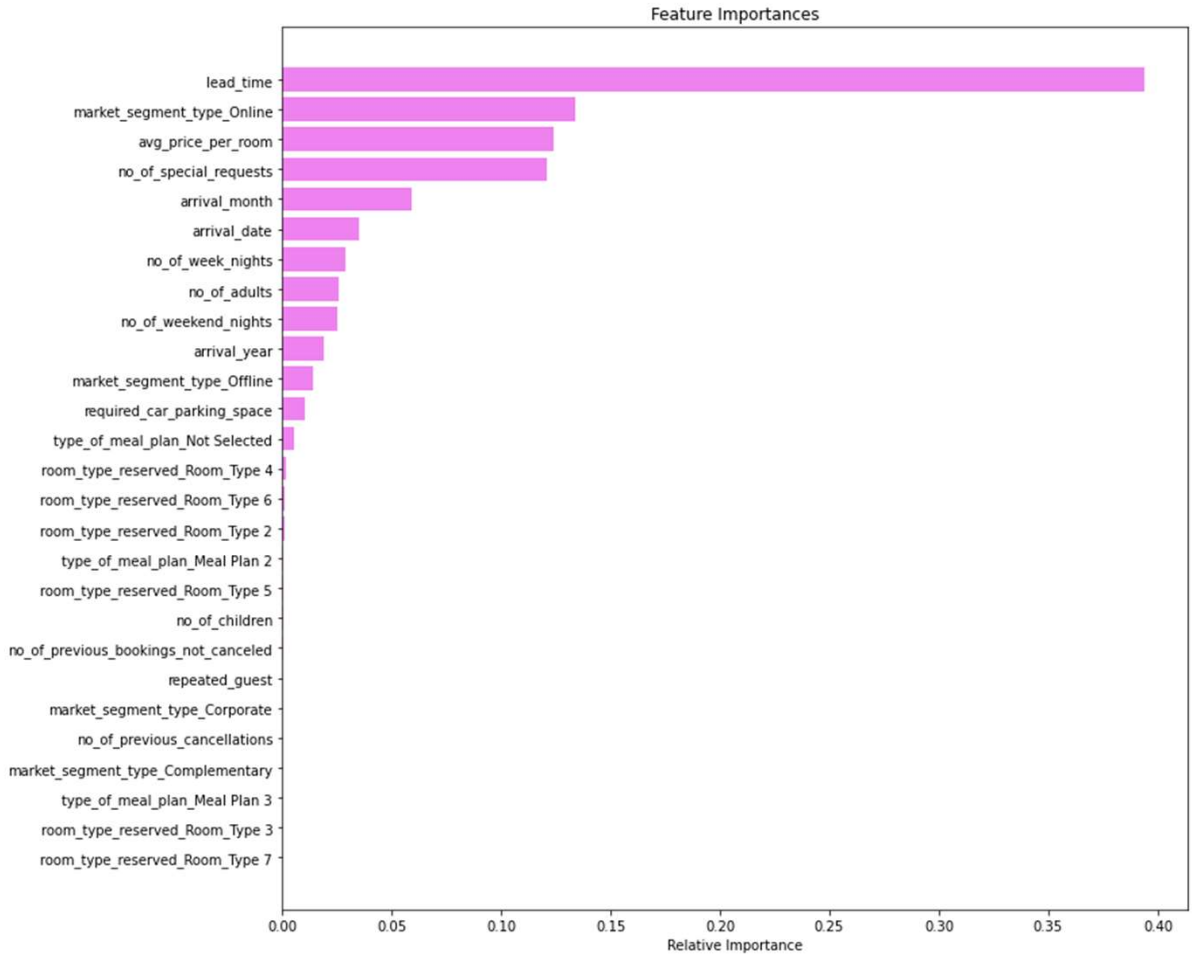
Testing



	Accuracy	Recall	Precision	F1
0	0.86879	0.85576	0.76614	0.80848

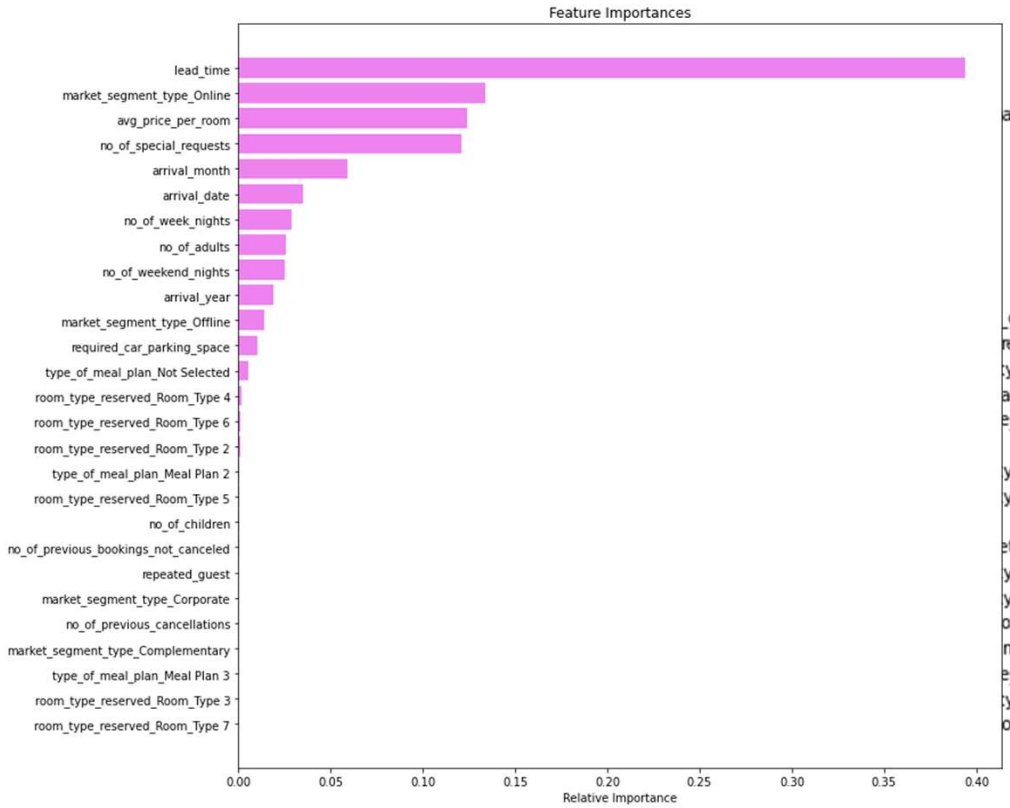
Model Building – Decision Tree – Feature Importances

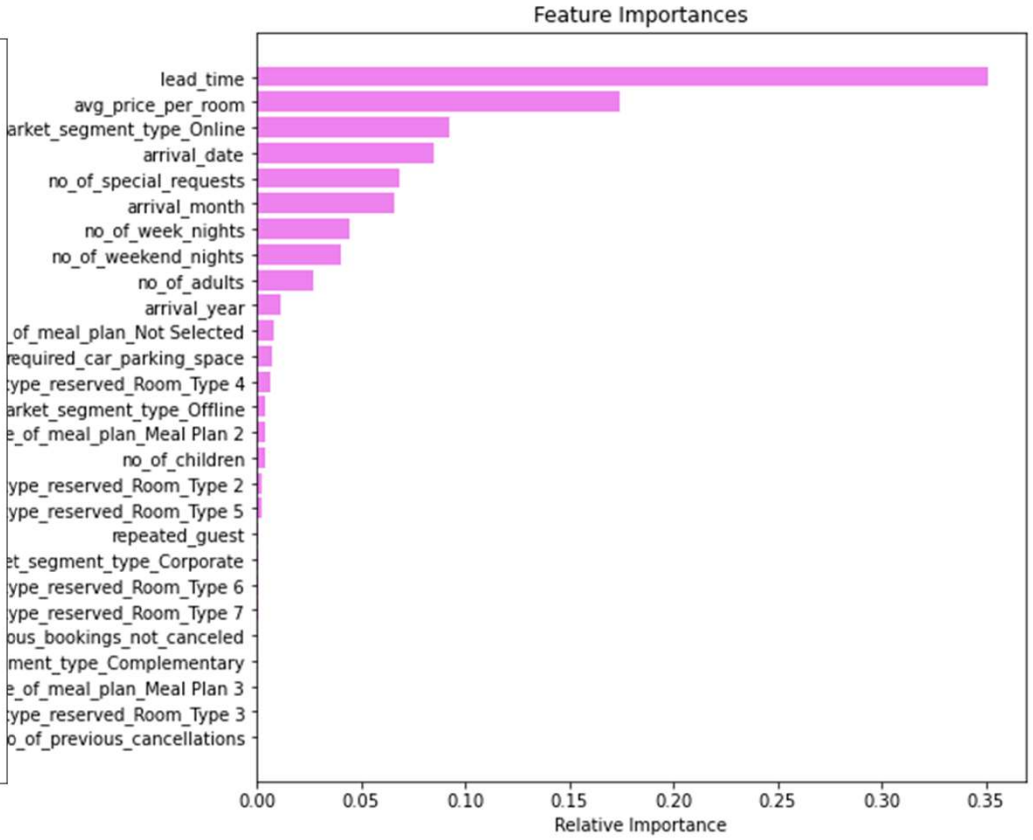
- **lead_time ~0.40** first (1st) place
- **Market_segment type_online ~0.15** second (2nd)
- **average price per room ~0.14** to third (3th) most important.
- **Number of special requests ~0.13** moving up two spots to fourth (4th)
- And so on… lead_time, market_segment_type_online avg_price_per_room, …



Model Building – Decision Tree

- Toggle before





Model Building – Decision Tree

- Comparing Decision Tree models
- Training

Training performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.99421	0.99421	0.89954
Recall	0.98661	0.98661	0.90303
Precision	0.99578	0.99578	0.81274
F1	0.99117	0.99117	0.85551

- Testing

Training performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.99421	0.99421	0.89954
Recall	0.98661	0.98661	0.90303
Precision	0.99578	0.99578	0.81274
F1	0.99117	0.99117	0.85551

	Accuracy	Recall	Precision	F1
0	0.86879	0.85576	0.76614	0.80848

Model Building – Decision Tree – Scratch Notes

- Scratch notes
- The model build can be used to predict the cancelations to with ~80%.
- The most important variables in predicting a cancelation are **(1) lead time**, **(2)online** bookings market segment, **(3) average price per room**, the **(4)number of special requests**.
- Inn hotels should focus on efforts to decrease the lead time,....
- Twice as important...The most importain features Leadtime is by far...
- lead_time ~0.40 first (1st) place ...Market_segment type_online ~0.15 second (2nd)
- average price per room ~0.14 to third (3th) most important. ...Number of special requests ~0.13 moving up two spots to fourth (4th)...

Model Building – Logistic Regression – Assumptions

- Checking the following Logistic Regression assumptions:
  - **No Multicollinearity** among independent variables
  - **Linearity of variables**. There should be a linear relationship between dependent and independent variables.
  - **Independence of error terms**. The residuals should be independent of each other.
  - **Normality of error terms**. The residuals must be normally distributed.
  - **No Heteroscedasticity**. The residuals must have constant variance.
- Tests conducted for checking model assumptions and the Results obtained

Great Learning

Happy Learning !