INN HOTEL PROJECTS

Supervised Learning
Classification Project
by
Sania

11/05/2022

Contents / Agenda

Executive Summary

Business Problem Overview and Solution Approach

EDA Results

Data Preprocessing

Model Performance Summary

Appendix

EXECUTIVE SUMMARY

- INN hotel group has a chain of hotels in Portugal, who are facing increasing number of cancellations. They are looking for a machine learning based solutions to help them in predicting which booking is likely to be cancelled.

- Based on the data provided, we will be looking for factors that have a high influence on booking cancellations and build a MACHINE LEARNING model that can predict which booking is going to be cancelled in advance and help in formulating profitable policies for cancellations and refunds.

Business problem and solution

Business problem

INN hotel group in Portugal is facing increasing number of cancellations. This is affecting

- Their revenue and loss of resources.
- Additional costs of distribution channels by increasing commissions or paying for publicity to help sell these rooms.
- Lowering prices last minute, so the hotel can resell a room, resulting in reducing the profit margin.
- Human resources to decide for the guests.

Solution

- In order to avoid these negative factors and help the hotel group to formulate profitable cancellation policies.
- We are Building a Machine Learning based on **Logistic Regression and Decision Tree Model** that can help in predicting which booking is likely to be cancelled. And find out the factors which are influencing booking cancellations.

DATA OVERVIEW

The dataset has 36275 rows and 19 columns. There are no null values and Duplicate values in the dataset. 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 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)

DATA OVERVIEW CONTINUED.....

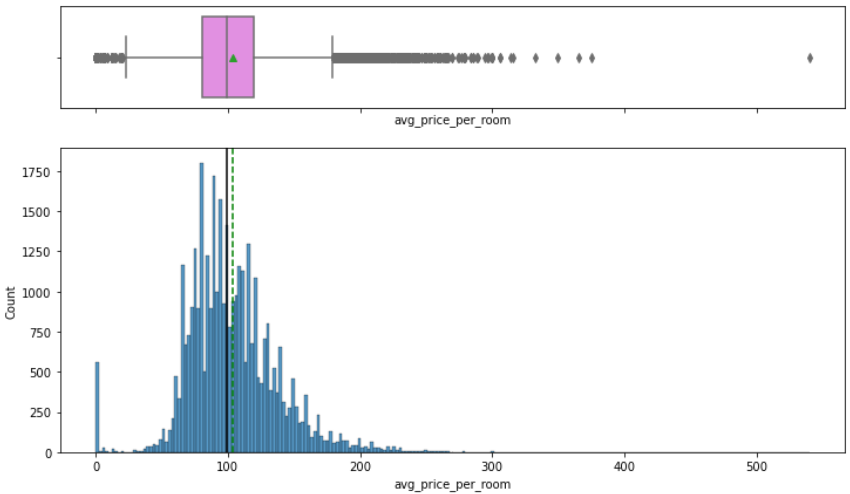
- required_car_parking_space: Does the customer require a car parking space? (0 - No, 1- Yes)
- room_type_reserved: Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels.
- lead_time: Number of days between the date of booking and the arrival date
- arrival_year: Year 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.

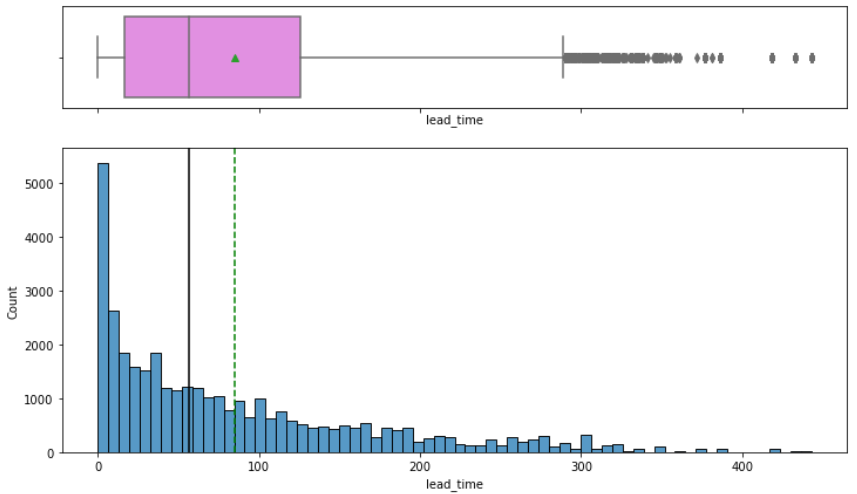
STATISTICAL SUMMARY

- We have two years of data, 2017 and 2018.
- At least 75% of the customers are not repeating customers.
- The average price per room is 103 euros. There's a huge difference between the 75th percentile and the maximum value which indicates there might be outliers present in this column.
- On average the customers book 85 days in advance.
- At least 75% of the customers do not require car parking space.
- The maximum value in the number of children column is 10.
- The number of adults ranges from 0 to 4.

EXPLORATORY DATA ANALYSIS

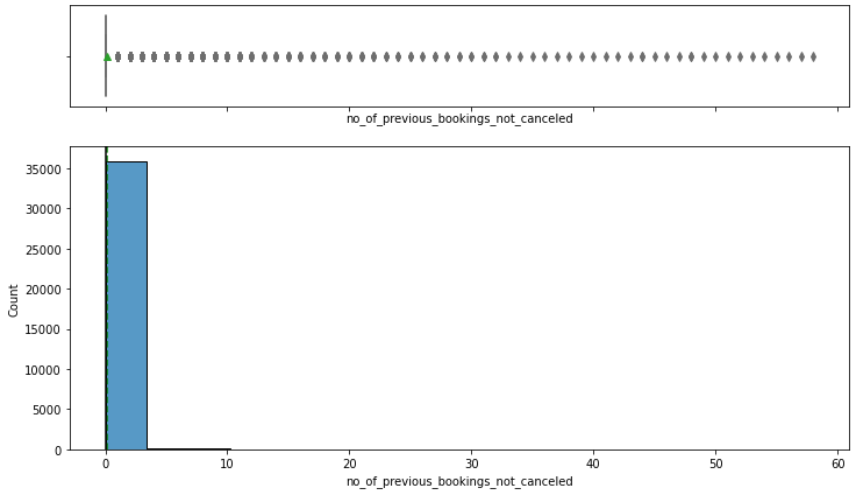
UNIVARIATE ANALYSIS ON LEAD_TIME and AVERAGE_PRICE_PER_ROOM

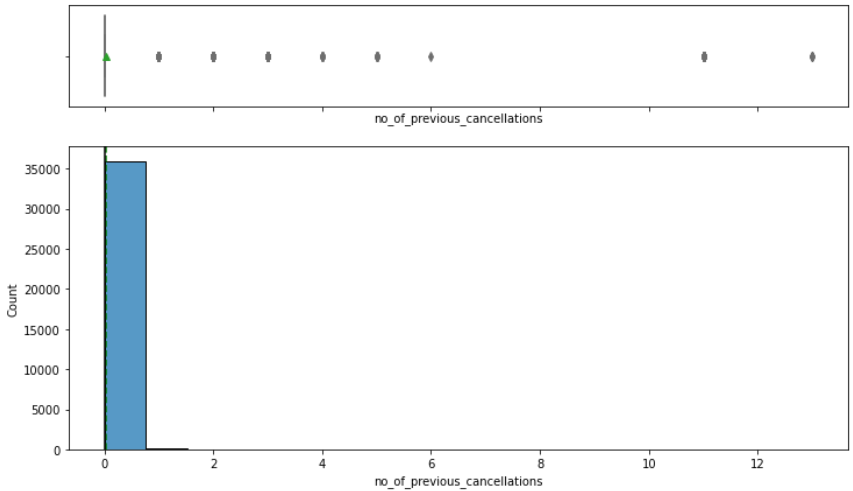




- The distribution of lead time is right-skewed, and there are many outliers.
- Some customers made booking around 500 days in advance.
- Many customers have made the booking on the same day
- The distribution of average price per room is skewed to right. There are outliers on both sides.
- The average price of a room is around ~100 euros.
- There is 1 observation where the average price of the room is more than 500 euros. This observation is quite far away from the rest of the values
- Interestingly some rooms have a price equal to 0.

Observations on number of previous booking cancellations and previous booking not cancelled

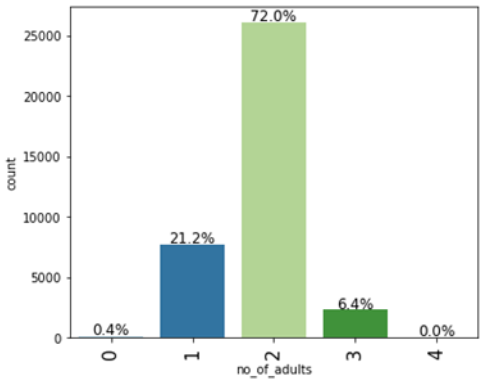


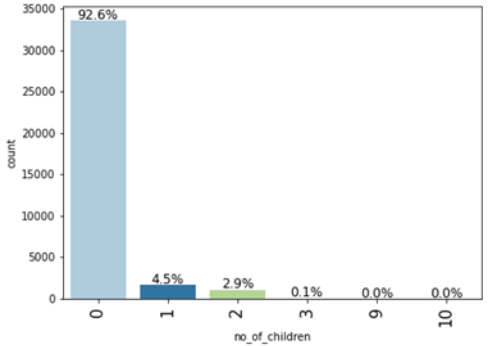


- Very few customers have more than one cancellation.
- Some customers canceled more than 12 times.
- Very few customers have more than 1 booking..
- Some customers have not canceled their bookings around 60 times.

Observations on number of adults and number of children

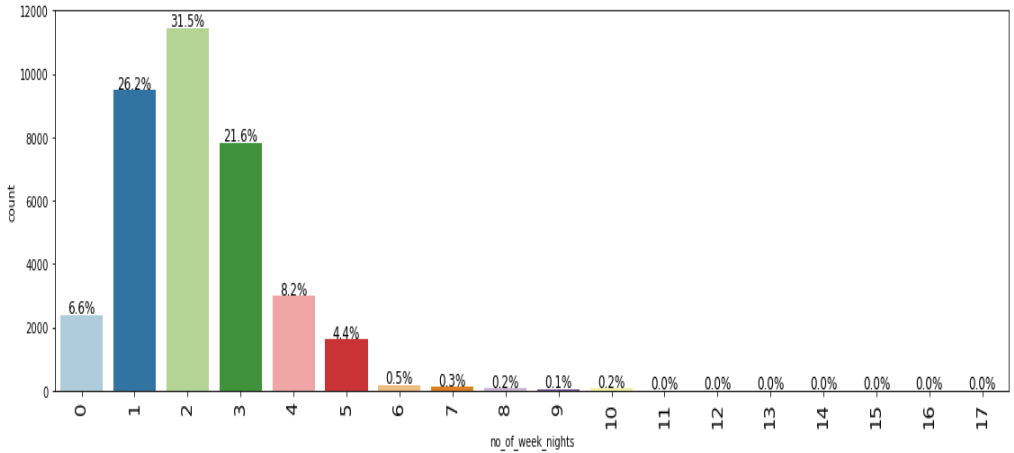
- 72% of the bookings were made for 2 adults.
- 93% of the customers didn't make reservations for children.
- There are some values in the data where the number of children is 9 or 10.





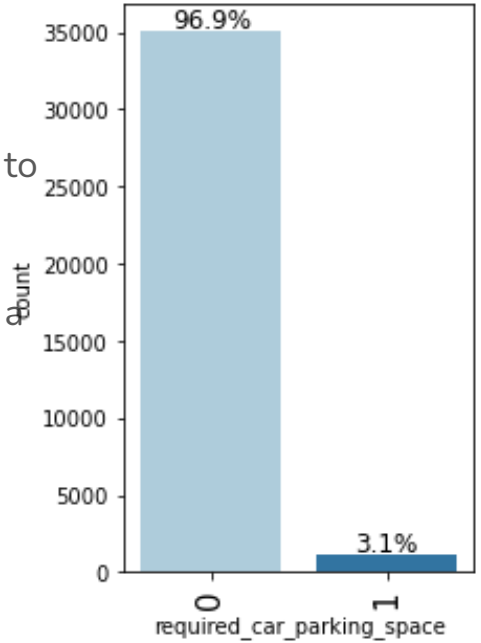
Labelled barplot on number_of_week_nights

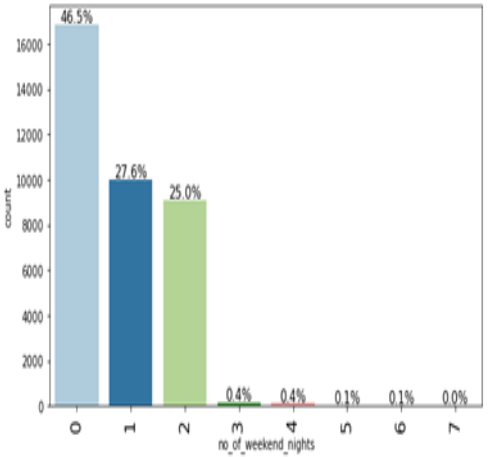
- Most bookings are made for 2 nights (31.5%) followed by 1 night (26.2%).
- A very less proportion of customers made the booking for more than 10 days.



Observations on number_of_weekend_nights and car parking space

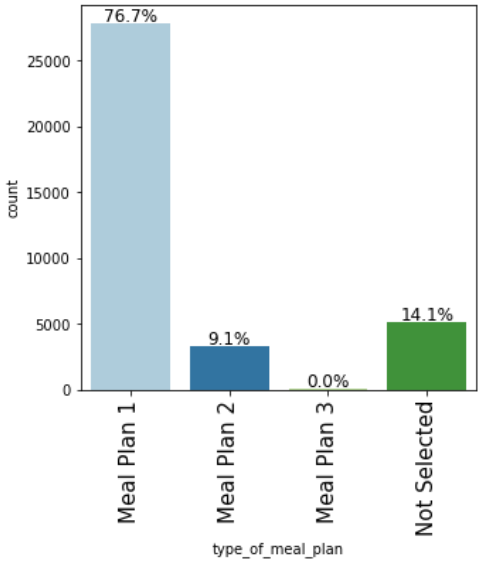
- 46.5% of the customers do not plan to spend the weekend in the hotel.
- The percentage of customers planning to spend 1 or 2 weekends in the hotel is almost the same.
- 96.9% of the customers do not require a car parking space.

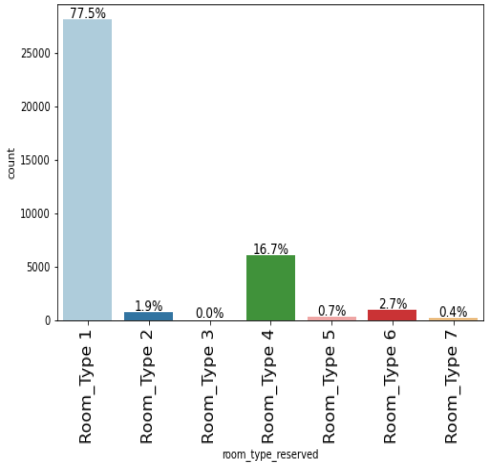




Observations on type of meal plan and ROOM TYPE RESERVED

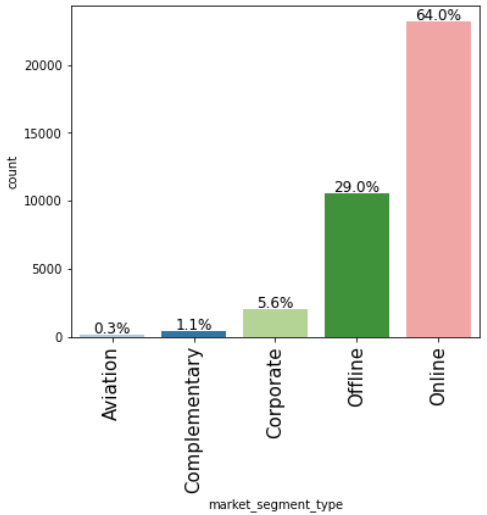
- Most of the customers prefer meal plan 1 that is only breakfast.
- 14.1% of the customers didn't select a meal plan.
- Around 77% of the customers booked Room_Type 1 followed by 17% of the customers booking Room_Type 4

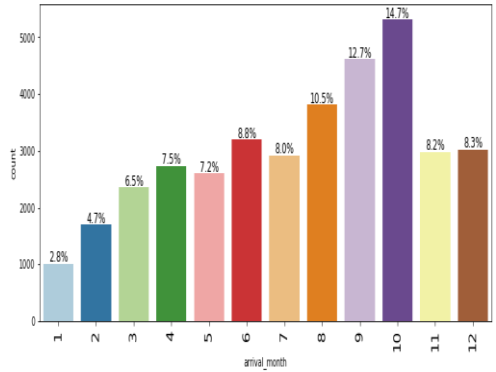




Observations on arrival month and market segment type

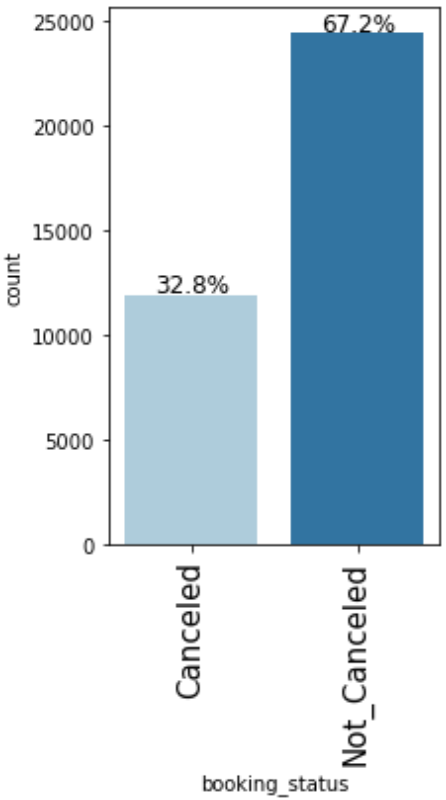
- October Month has the highest number of arrivals.
- 14.7% of the bookings were made in October.
- 64% of the hotel bookings were made online followed by 29% of the bookings which were made offline.

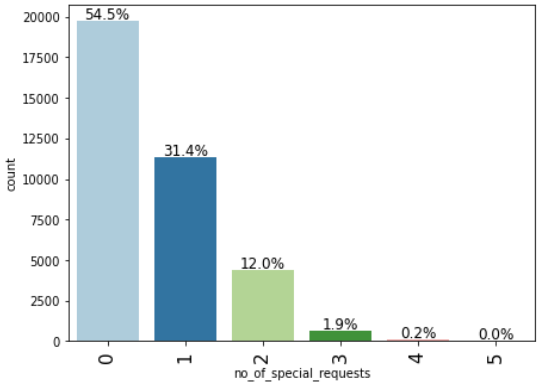




OBSERVATIONS ON BOOKING STATUS AND NUMBER OF SPECIAL REQUEST

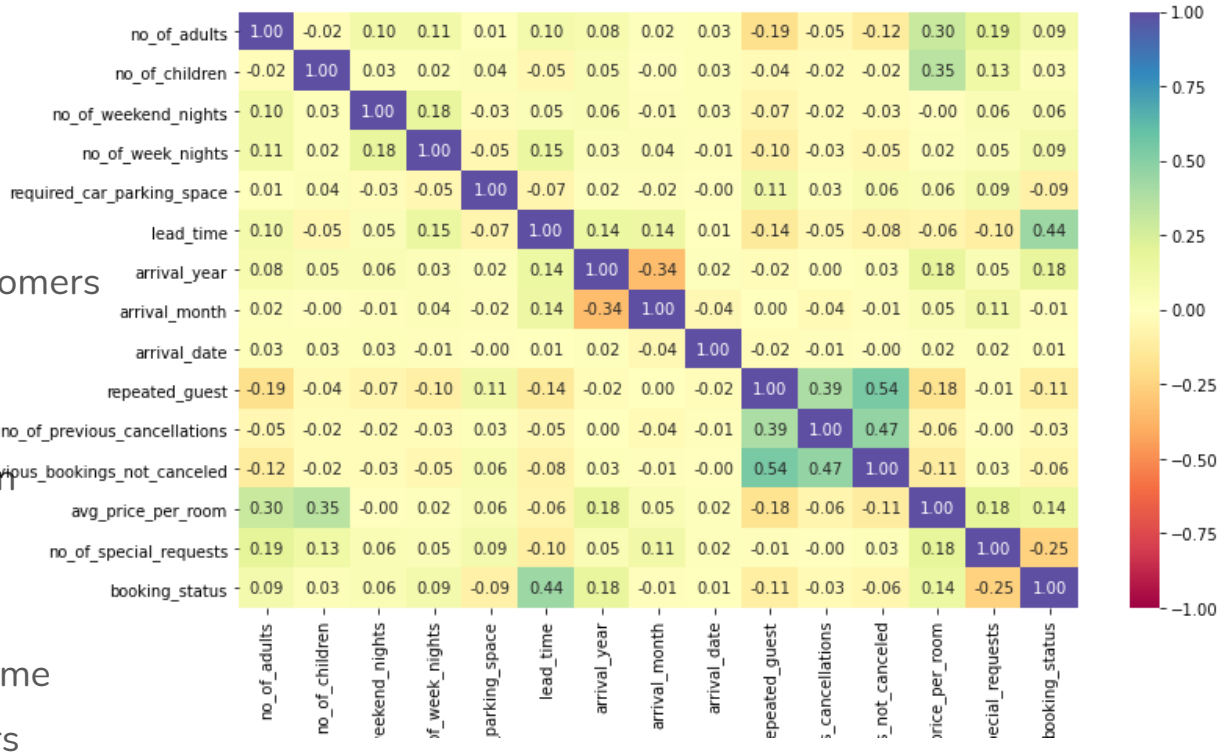
- 67.2% of the bookings were not cancelled by the customer.
- 54.5% of the customers generally do not make any requests while booking a hotel room.





CORRELATION MATRIX

- There's a positive correlation between the number of customers (adults and children) and the average price per room.
- As more the number of customers more rooms they will require to increasing the cost.
- There's a negative correlation between average room price and repeated guests.
- The hotel might be giving some loyalty benefits to the customers

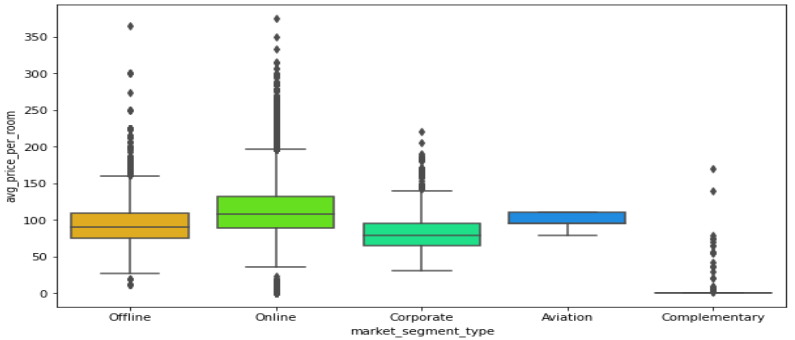


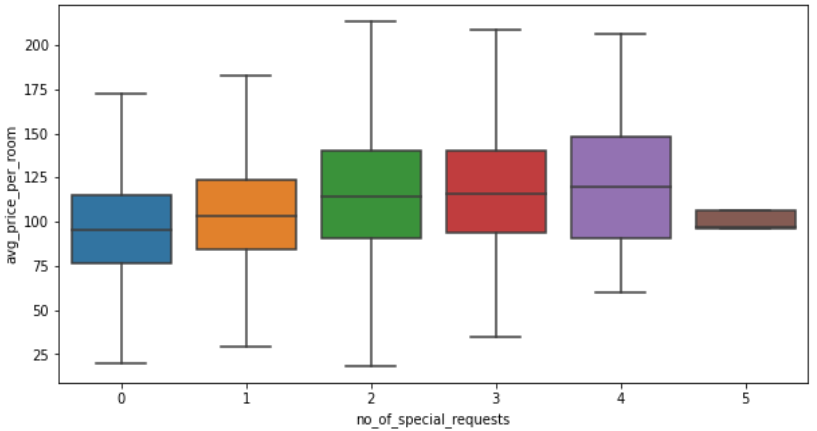
OBSERVATIONS ON BIVARIATE ANALYSIS continued....

- There's a positive correlation between the number of previous bookings canceled and previous bookings not canceled by a customer and repeated guest.
- There's a positive correlation between lead time and the number of weeknights a customer is planning to stay in the hotel.
- There's a positive correlation between booking status and lead time, indicating higher the lead time higher are the chances of cancellation. We will analyze it further.
- There's a negative correlation between the number of special requests from the customer and the booking status, indicating if a customer has made some special requests the chances of cancellation might decrease. We will analyze it further.

Observations on how prices vary across different market segments

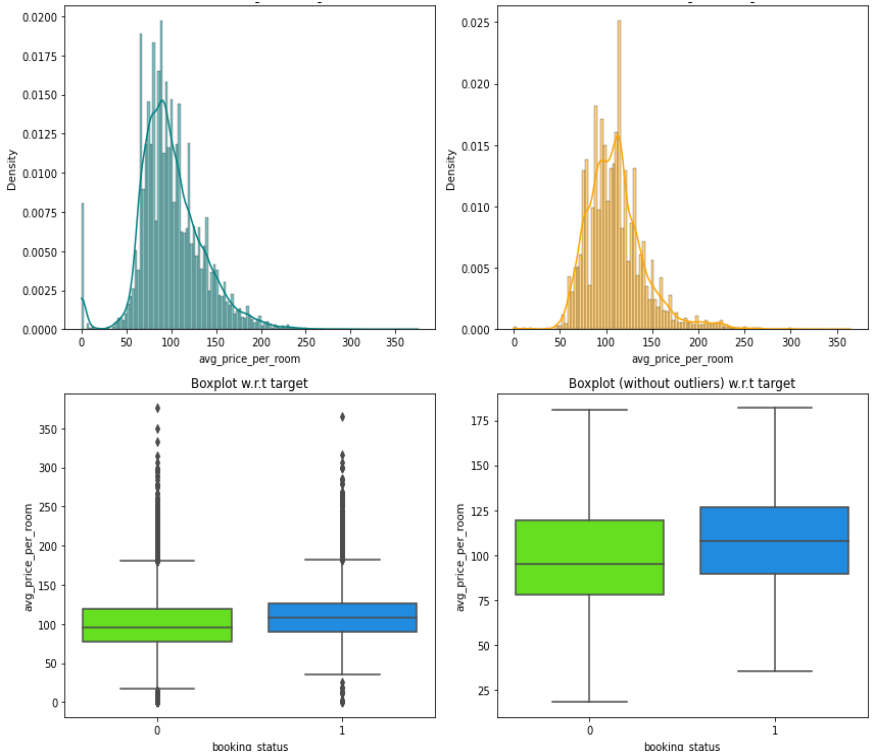
- Rooms booked online have high variations in prices.
- The offline and corporate room prices are almost similar.
- Complementary market segment gets the rooms at very low prices, which makes sense.
- The median prices of the rooms where some special requests were made by the customers are slightly higher than the rooms where customer didn't make any requests.





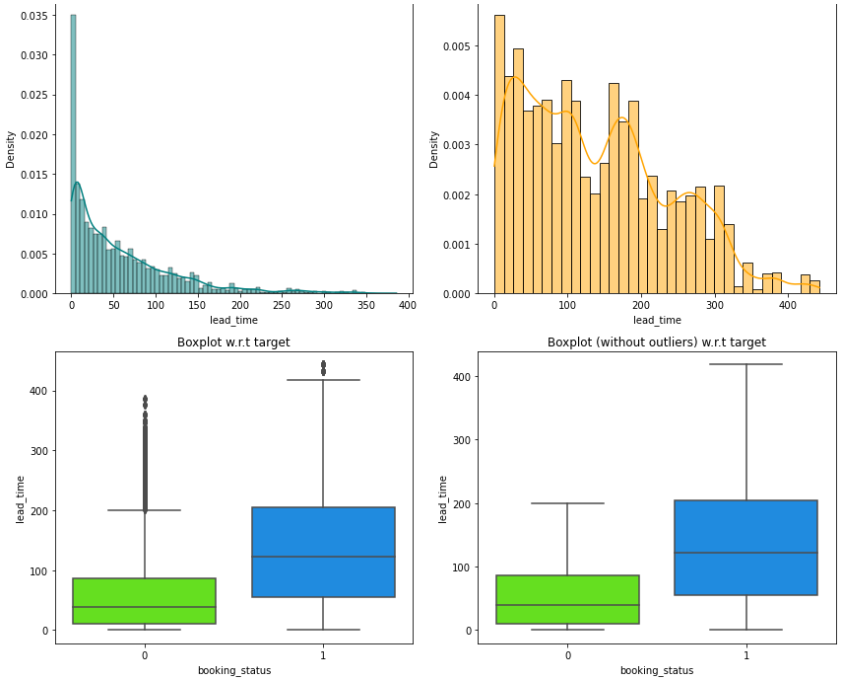
Observation of booking status and AVG price PER ROOM

- The distribution of price for canceled bookings and not canceled bookings is quite similar.
- The prices for the canceled bookings are slightly higher than the bookings which were not canceled.



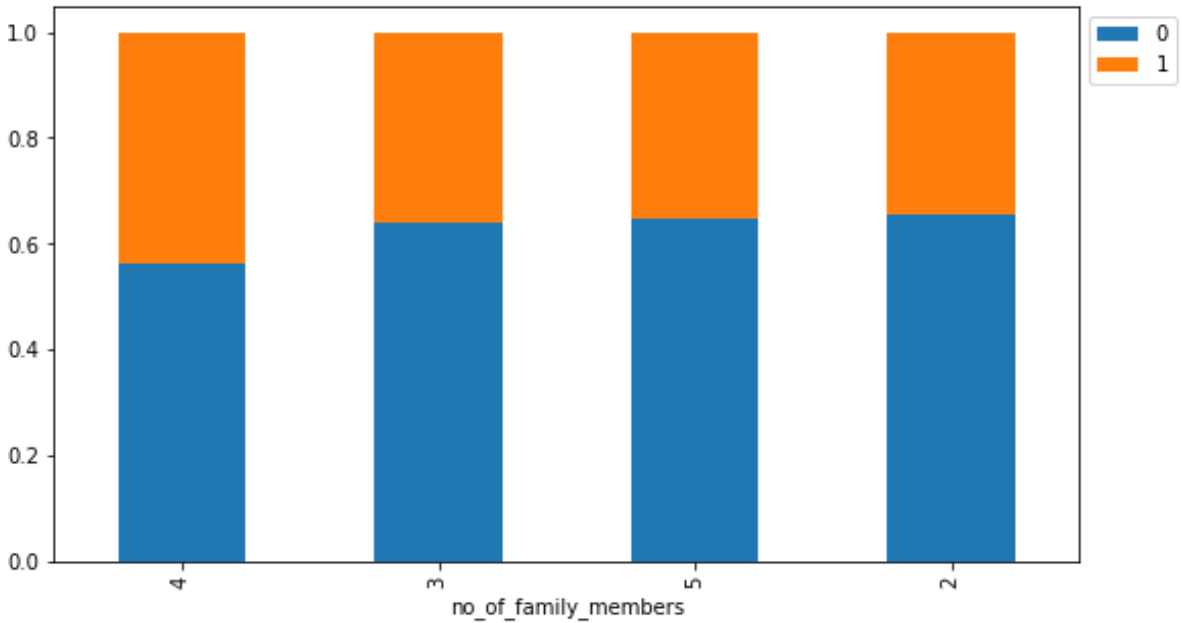
OBSERVATION OF BOOKING STATUS AND LEAD TIME

- There's a big difference in the median value of lead time for bookings that were canceled and bookings that were not canceled.
- Higher the lead time higher are the chances of a booking being canceled.

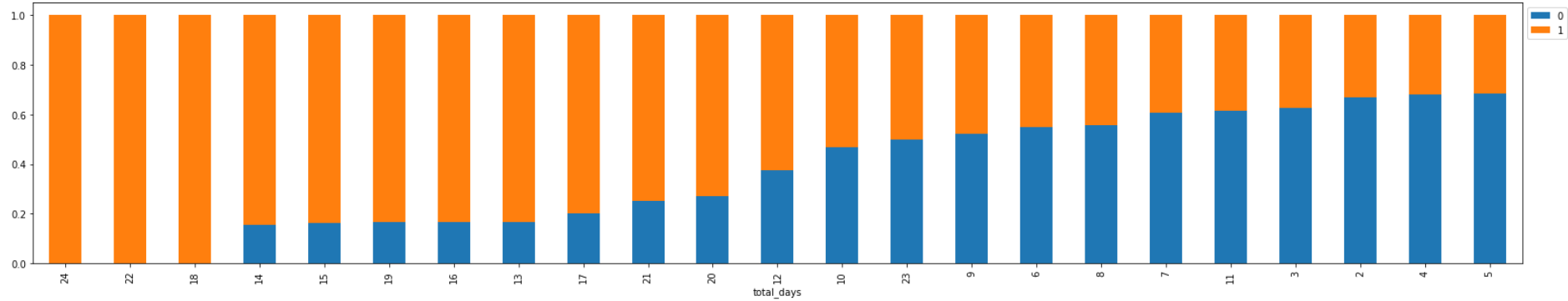


Observation of booking status with number of family members

- * There's a ~40% chance of a booking getting canceled if the booking is made for 4 family members.



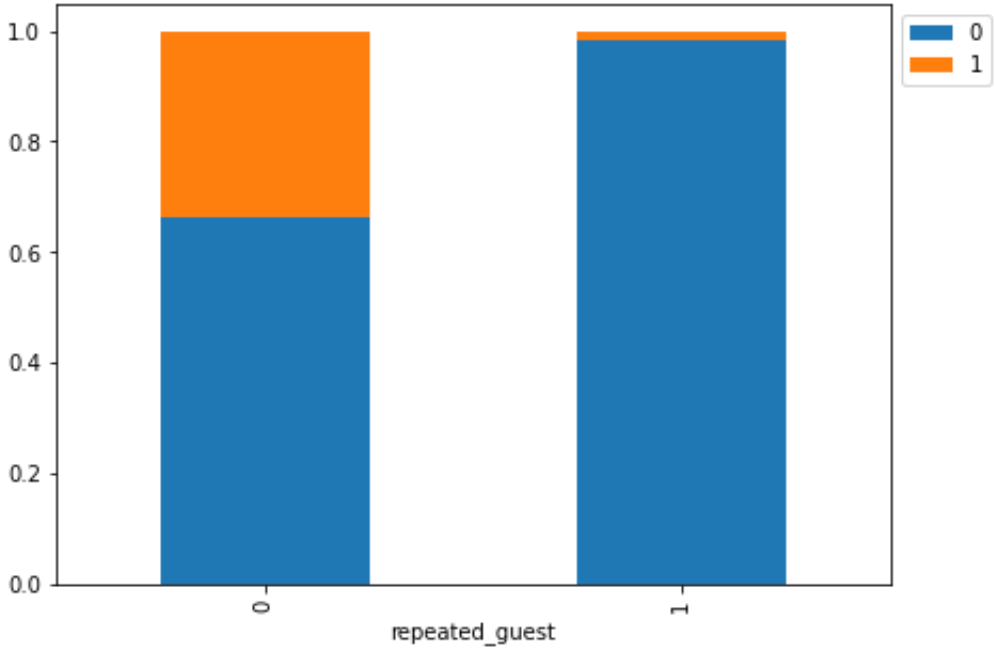
OBSERVATION OF THE CANCELLATION TREND WITH THE NUMBER OF STAY



- The general trend is that the chances of cancellation increase as the number of days the customer planned to stay at the hotel increases.

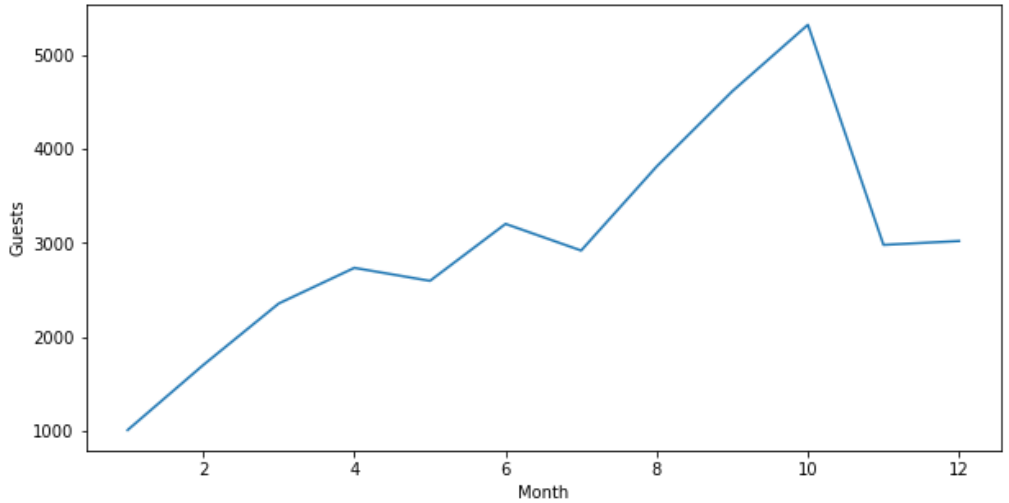
OBSERVATION WITH THE PERCENTAGE OF REPEATING GUESTS

- There are very few repeat customers but the cancellation among them is very less.
- Repeating Customers who stay in the hotel often are more important for business equity.



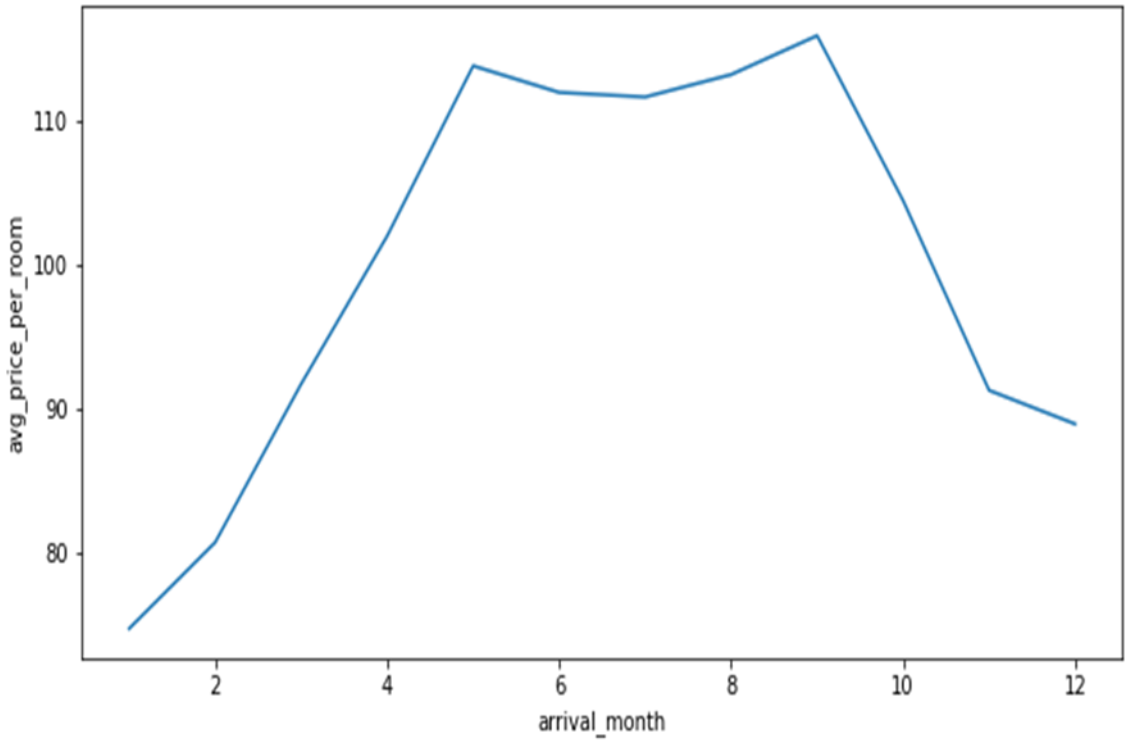
OBSERVATION OF BUSIEST MONTHS IN THE HOTEL

- The trend shows the number of bookings remains consistent from April to July and the hotel sees around 3000 to 3500 guests.
- Most bookings were made in October - more than 5000 bookings.
- Least bookings were made in January - around 1000 bookings.

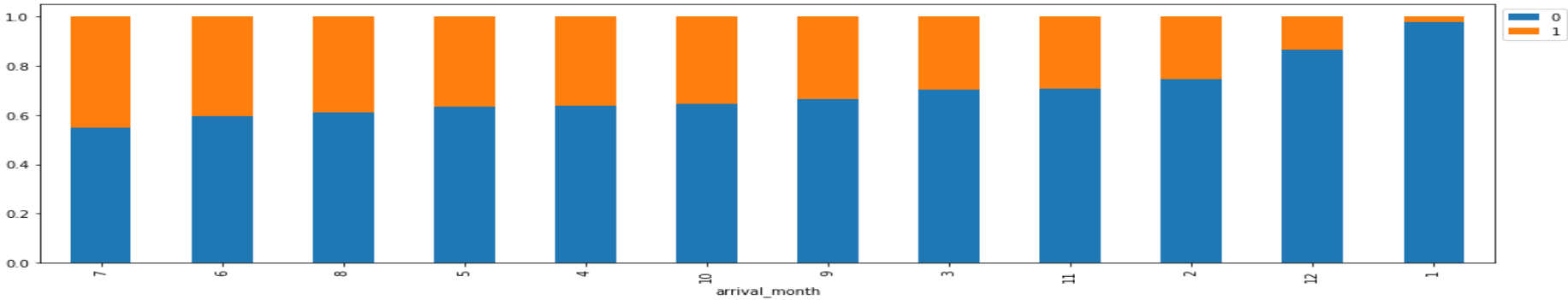


Observations on average_price and arrival month

- The price of the rooms are high from MAY-SEPT i.e ~115 euros



OBSERVATION OF PECENTAGE OF BOOKINGS CANCELLED IN EACH MONTH

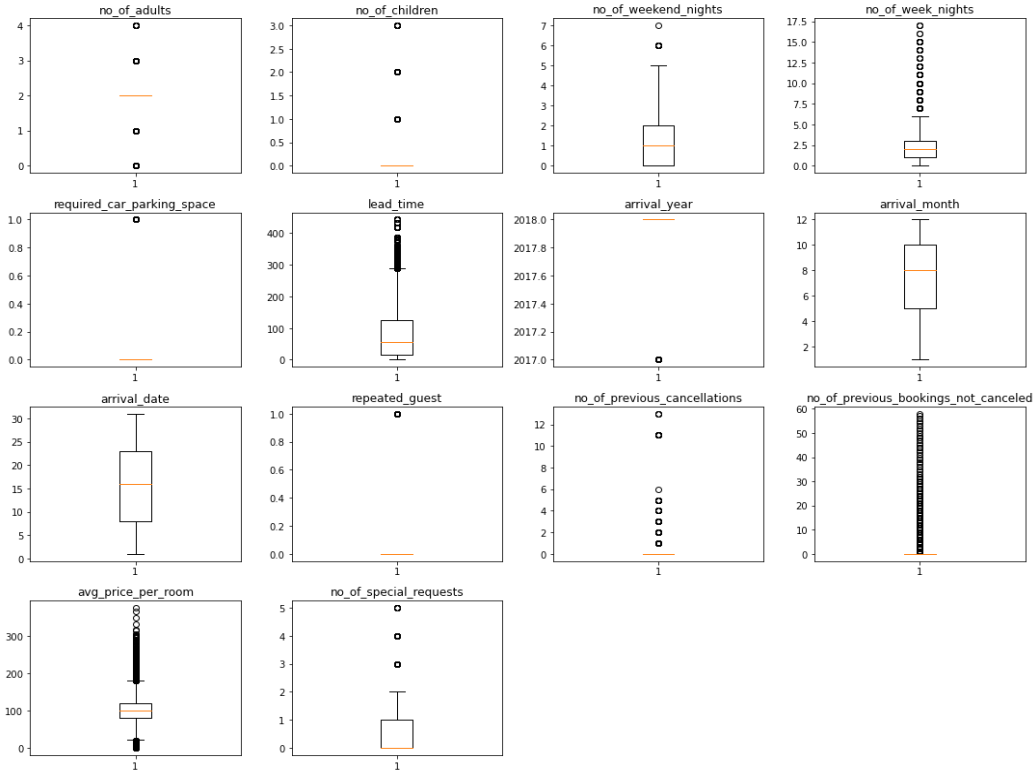


- We see that even though the highest number of bookings were made in September and October - around 40% of these bookings got canceled.
- Least bookings were canceled in December and January - customers might have traveled to celebrate Christmas and New Year.

OUTLIERS CHECK

OBSERVATIONS

- There are quite a few outliers in the data.
- However, we will not treat them as they are proper values.



INSIGHTS FROM EDA

- Data Trend Shows OCTOBER is the busiest month of the year.
- Collected Data shows that ONLINE MARKET SEGMENT through which majority of bookings are made.
- Hotel prices are dynamic and change according the demands and Customer Demographics.
- 32.8% of total hotel bookings are cancelled.
- Data trend shows that returning guests have very low cancellation rate.

LOGISTIC REGRESSION

Data preparation for modelling

- We want to predict which bookings will be canceled.
- Before we proceed to build a model, we'll have to encode categorical features.
- We'll split the data into train and test to be able to evaluate the model that we build on the train data.

Model can make wrong predictions as:

- Predicting a customer will not cancel their booking but in reality, the customer will cancel their booking.
- Predicting a customer will cancel their booking but in reality, the customer will not cancel their booking.
- Hotel would want F1 Score to be maximized, greater the F1 score higher are the chances of minimizing False Negatives and False Positives.

Training performance:(Default Threshold)

Accuracy	Recall	Precision	F1
0.80600	0.63410	0.73971	0.68285

❑ OBSERVATIONS

- ❑ Negative values of the coefficient show that the probability of customers canceling the booking decreases with the increase of the corresponding attribute value.
- ❑ Positive values of the coefficient show that the probability of customer canceling increases with the increase of 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.
- ❑ But these variables might contain multicollinearity, which will affect the p-values.
- ❑ We will have to remove multicollinearity from the data to get reliable coefficients and p-values.
- ❑ There are different ways of detecting (or testing) multi-collinearity, one such way is the Variation Inflation Factor.

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25364
Method:	MLE	Df Model:	27
Date:	Tue, 14 Dec 2021	Pseudo R-squ.:	0.3292
Time:	18:36:08	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	3987.822	0.004	0.997	-7798.629	7833.346
type_of_meal_plan_Not Selected	0.2784	0.053	5.247	0.000	0.174	0.382
room_type_reserved_Room_Type 2	-0.3605	0.131	-2.748	0.006	-0.618	-0.103
room_type_reserved_Room_Type 3	-0.0012	1.310	-0.001	0.999	-2.568	2.566
room_type_reserved_Room_Type 4	-0.2823	0.053	-5.304	0.000	-0.387	-0.178
room_type_reserved_Room_Type 5	-0.7189	0.209	-3.438	0.001	-1.129	-0.309
room_type_reserved_Room_Type 6	-0.9501	0.151	-6.274	0.000	-1.247	-0.653
room_type_reserved_Room_Type 7	-1.4003	0.294	-4.770	0.000	-1.976	-0.825
market_segment_type_Complementary	-40.5975	5.65e+05	-7.19e-05	1.000	-1.11e+06	1.11e+06
market_segment_type_Corporate	-1.1924	0.266	-4.483	0.000	-1.714	-0.671
market_segment_type_Offline	-2.1946	0.255	-8.621	0.000	-2.694	-1.696
market_segment_type_Online	-0.3995	0.251	-1.590	0.112	-0.892	0.093

Coefficient interpretations

Coefficients of required_car_parking_space, arrival_month, repeated_guest, no_of_special_requests and some others are negative, an increase in these will lead to a decrease in chances of a customer canceling their booking.

Coefficients of no_of_adults, no_of_children, no_of_weekend_nights, no_of_week_nights, lead_time, avg_price_per_room, type_of_meal_plan_Not Selected and some others are positive, an increase in these will lead to a increase in the chances of a customer canceling their booking.

Converting coefficients to odds

- The coefficients of the logistic regression model are in terms of log(odd), to find the odds we have to take the exponential of the coefficients.
- Therefore, odds = exp(b)
- The percentage change in odds is given as odds = (exp(b) - 1) * 100

COEFFICIENT INTERPRETATIONS

no_of_adults: Holding all other features constant a 1 unit change in the number of children will increase the odds of a booking getting cancelled by 1.11 times or a 11.49% increase in the odds of a booking getting cancelled.

no_of_children: Holding all other features constant a 1 unit change in the number of children will increase the odds of a booking getting cancelled by 1.16 times or a 16.54% increase in the odds of a booking getting cancelled.

no_of_weekend_nights: Holding all other features constant a 1 unit change in the number of weeknights a customer stays at the hotel will increase the odds of a booking getting cancelled by 1.11 times or a 11.46% increase in the odds of a booking getting cancelled.

no_of_week_nights: Holding all other features constant a 1 unit change in the number of weeknights a customer stays at the hotel will increase the odds of a booking getting cancelled by 1.04 times or a 4.25% increase in the odds of a booking getting cancelled.

required_car_parking_space: The odds of a customer who requires a car parking space are 0.2 times less than a customer who doesn't require a car parking space or a 79.70% fewer odds of a customer canceling their booking.

lead_time: Holding all other features constant a 1 unit change in the lead time will increase the odds of a booking getting cancelled by 1.01 times or a 1.58% increase in the odds of a booking getting cancelled.

no_of_special_requests: Holding all other features constant a 1 unit change in the number of special requests made by the customer will decrease the odds of a booking getting cancelled by 0.22 times or a 77% decrease in the odds of a booking getting cancelled.

avg_price_per_room: Holding all other features constant a 1 unit change in the lead time will increase the odds of a booking getting cancelled by 1.01 times or a 1.93% increase in the odds of a booking getting cancelled.

MODEL PERFORMANCE SUMMARY

Training performance:

	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

Test performance:

	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80465	0.79555	0.80345
Recall	0.63089	0.73964	0.70358
Precision	0.72900	0.66573	0.69353
F1	0.67641	0.70074	0.69852

Model performance observations

- We have been able to build a predictive model that can be used by the hotel to predict which bookings are likely to be cancelled with an F1 score of 0.69 on the training set accordingly.
- The logistic regression models are giving a generalized performance on training and test set.
- Using the model with default threshold the model will give a low recall but good precision score - The hotel will be able to predict which bookings will not be cancelled and will be able to provide satisfactory services to those customers which help in maintaining the brand equity but will lose on resources.
- Using the model with a 0.37 threshold the model will give a high recall but low precision score - The hotel will be able to save resources by correctly predicting the bookings which are likely to be cancelled but might damage the brand equity.
- Using the model with a 0.42 threshold the model will give a balance recall and precision score - The hotel will be able to maintain a balance between resources and brand equity.
- Coefficients of required_car_parking_space, arrival_month, repeated_guest, no_of_special_requests and some others are negative, an increase in these will lead to a decrease in chances of a customer canceling their booking.
- Coefficients of no_of_adults, no_of_children, no_of_weekend_nights, no_of_week_nights, lead_time, avg_price_per_room, type_of_meal_plan_Not Selected and some others are positive, an increase in these will lead to a increase in the chances of a customer canceling their booking.

DECISION TREE

Checking Model performance on Trainingset and Testset

TRAINING SET

	Accuracy	Recall	Precision	F1
0	0.99421	0.98661	0.99578	0.99117

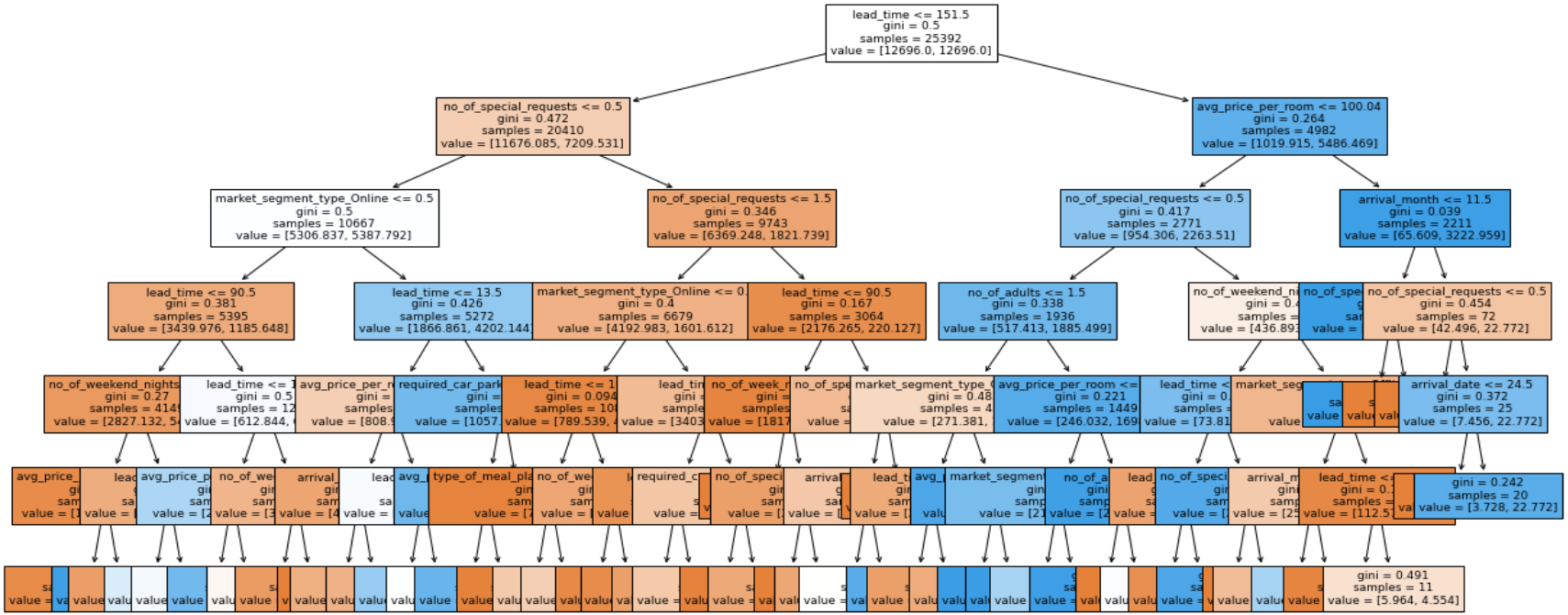
- There are no errors on the training dataset.
- Model has performed very well on the training set.
- As we know a decision tree will continue to grow and classify each data point correctly if no restrictions are applied, as the trees will learn all the patterns in the training set.

TEST SET

	Accuracy	Recall	Precision	F1
0	0.87118	0.81175	0.79461	0.80309

- The decision tree model is overfitting the data as expected and not able to generalize well on the test set.
- We will have to prune the decision tree to avoid overfitting.
- Due to the overfitting, the decision tree branches are very complex. Should look at limiting depth and post pruning.

Visualizing the decision tree after pre-pruning



Observation on decision tree

- We observe that the most important features are:

1) **Lead Time**
2) **Market Segment - Online**
3) **Number of special requests**
4) **Average price per room**

<u>According to the Decision tree model</u>

1.We should look at booking with lead time greater the 151 as they have a high probability of being cancelled.

2. If the booking has a lead time of less than 151 we should focus on the number of the special request . The higher the number the higher probability they will not cancel there booking.

3. Lastly, we should also focus on booking receive via Online as the have second highest level of cancellation

Continued.....

- If the average price per room is greater than 100 euros and the arrival month is December, then the booking is less likely to be cancelled.
- If the average price per room is less than or equal to 100 euros and the number of special request is 0, then the booking is likely to get canceled.
- If a customer has at least 1 special request the booking is less likely to be cancelled.
- If the customer didn't make any special requests and the booking was done Online it is more likely to get canceled, if the booking was not done online, it is less likely to be canceled.

Comparison of decision tree

- **Training performance comparison**

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.99421	0.83097	0.89954
Recall	0.98661	0.78608	0.90303
Precision	0.99578	0.72425	0.81274
F1	0.99117	0.75390	0.85551

- **Test set performance comparison**

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.87118	0.83497	0.86879
Recall	0.81175	0.78336	0.85576
Precision	0.79461	0.72758	0.76614
F1	0.80309	0.75444	0.80848

OBSERVATIONS

- Decision tree model with default parameters is overfitting the training data and is not able to generalize well.
- Reducing the number of branches and generalizing the testing data allowed for a higher f1 score.
- Decision tree model with pre-pruning has given the best recall score on training data.
- Pre-pruned tree has given a generalized performance with balanced values of precision and recall.
- Post-pruned tree is giving a high F1 score as compared to other models but the difference between precision and recall is high.
- The hotel will be able to maintain a balance between resources and brand equity using the pre-pruned decision tree model.

RECOMMENDATIONS

Knowing that lead time, price, and online bookings have the highest influence on cancellations, we can infer that having a cancellation clause during the online booking process would impact booking commitment.

Repeating guests have a very low cancellation rate . Creating a loyalty program for those guests would reward your customers with personalized incentives.

During the online booking process, offering additional customizations or special requests would help reduce the likelihood of a cancelled booking.

Keep price of rooms near competitive pricing as it seems like guest will be booking a room with the expectation of continued searches.

Great Learning

Happy Learning !