Star Hotels Bookings Case Study

CONTENTS

BUSINESS OVERVIEW

PROBLEM STATEMENT

DATA OVERVIEW

EXPLORATORY DATA ANALYSIS

MODEL PERFORMANCE

BUSINESS RECOMMENDATIONS

Business Overview

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.

Problem Statement

The increasing number of cancellations calls for a Machine Learning based solution that can help in predicting which booking is likely to be canceled.

Star 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.

The data scientist must 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 Overview – Features Used

- 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 Star 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.

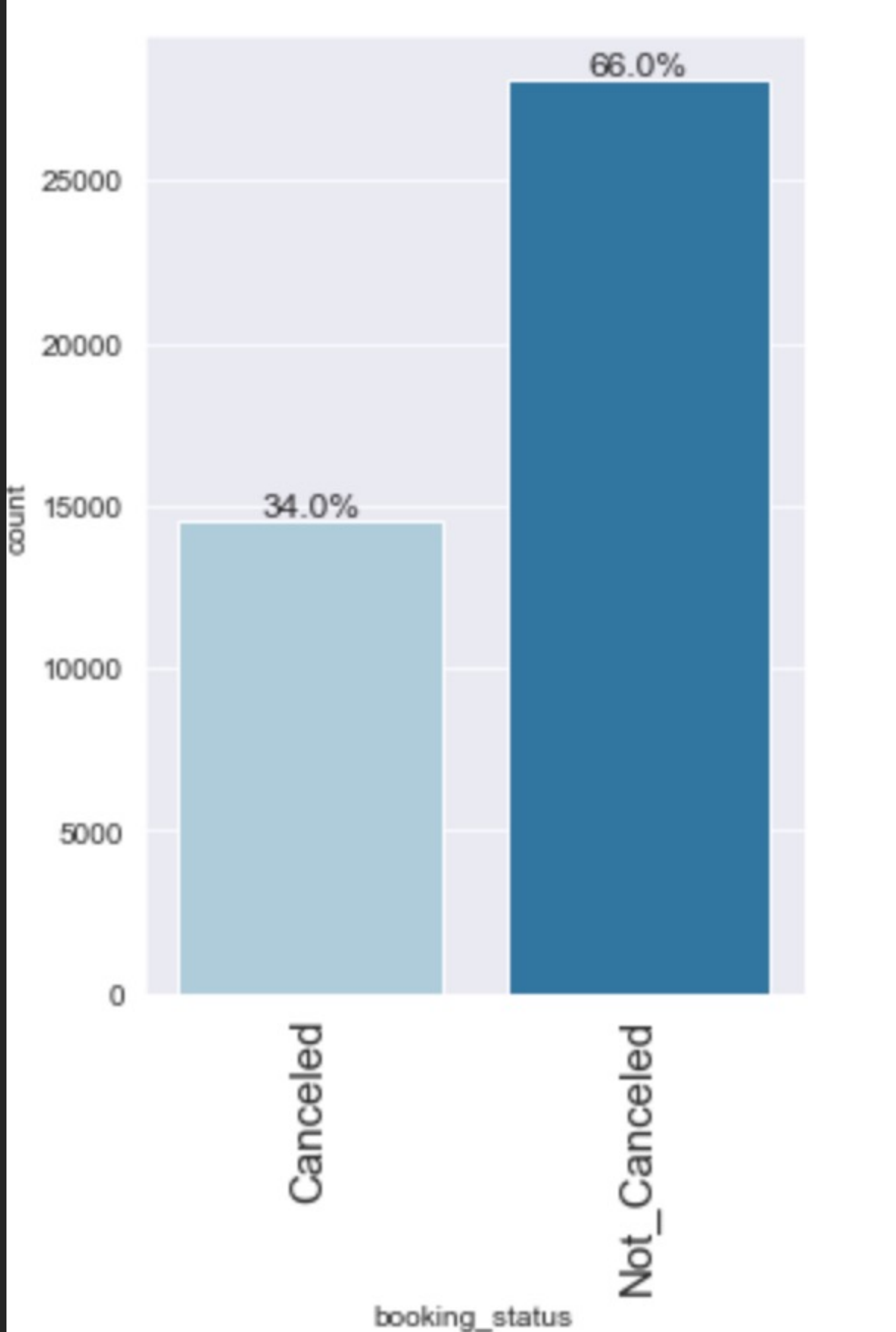
Exploratory Data Analysis

- The data has 56,926 observations with 17 features
- There are 14,350 duplicate records
- The dataset doesn’t have any missing values

Cancellations

34% of bookings are cancelled

The loses are high for these cancellations

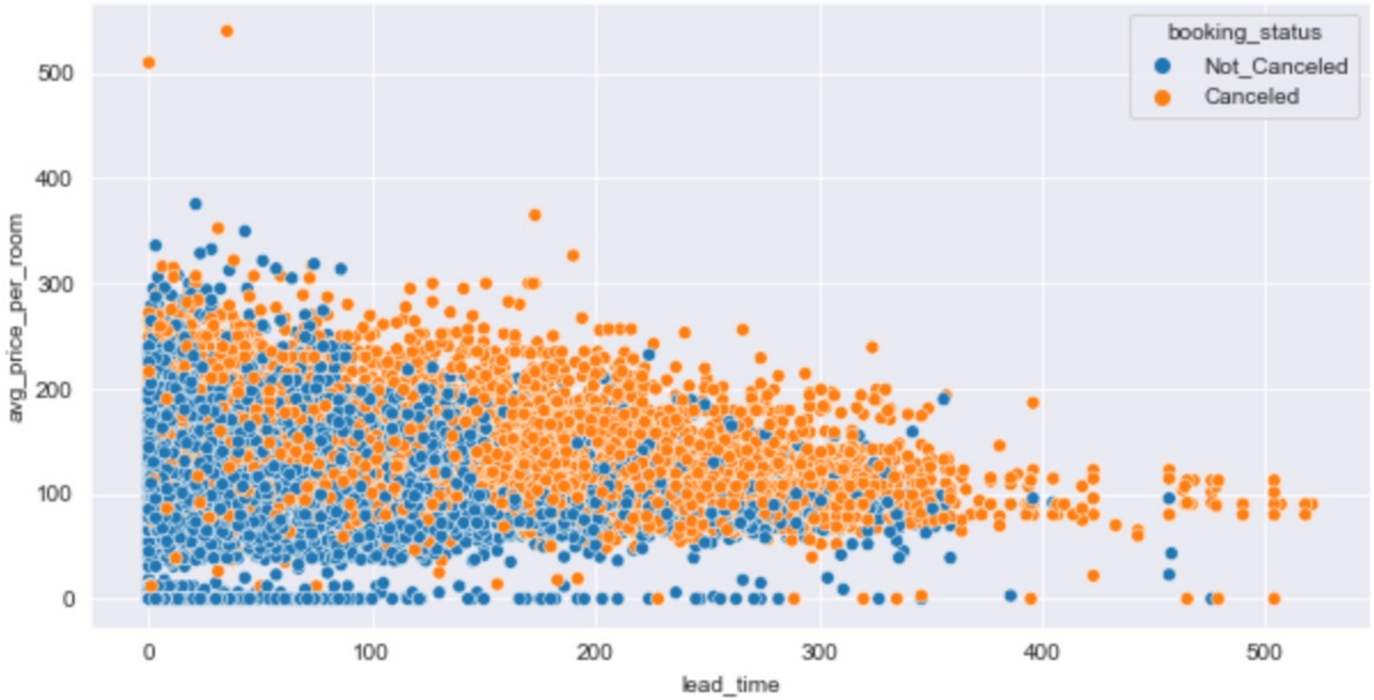


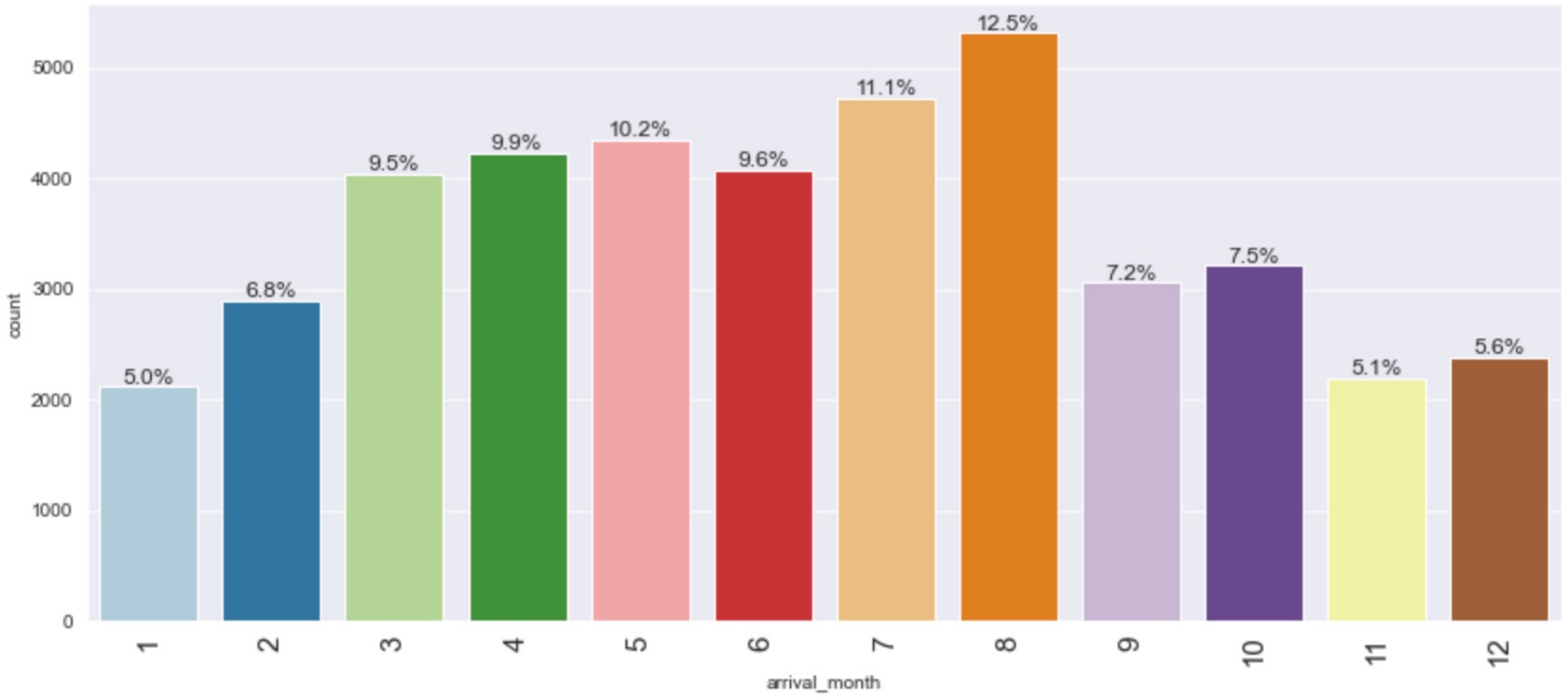
Lead Time and Average Price

Higher lead time has higher cancellations

Average price increases as the lead time decreases

Higher average price tends to have higher cancellations





Arrival Months

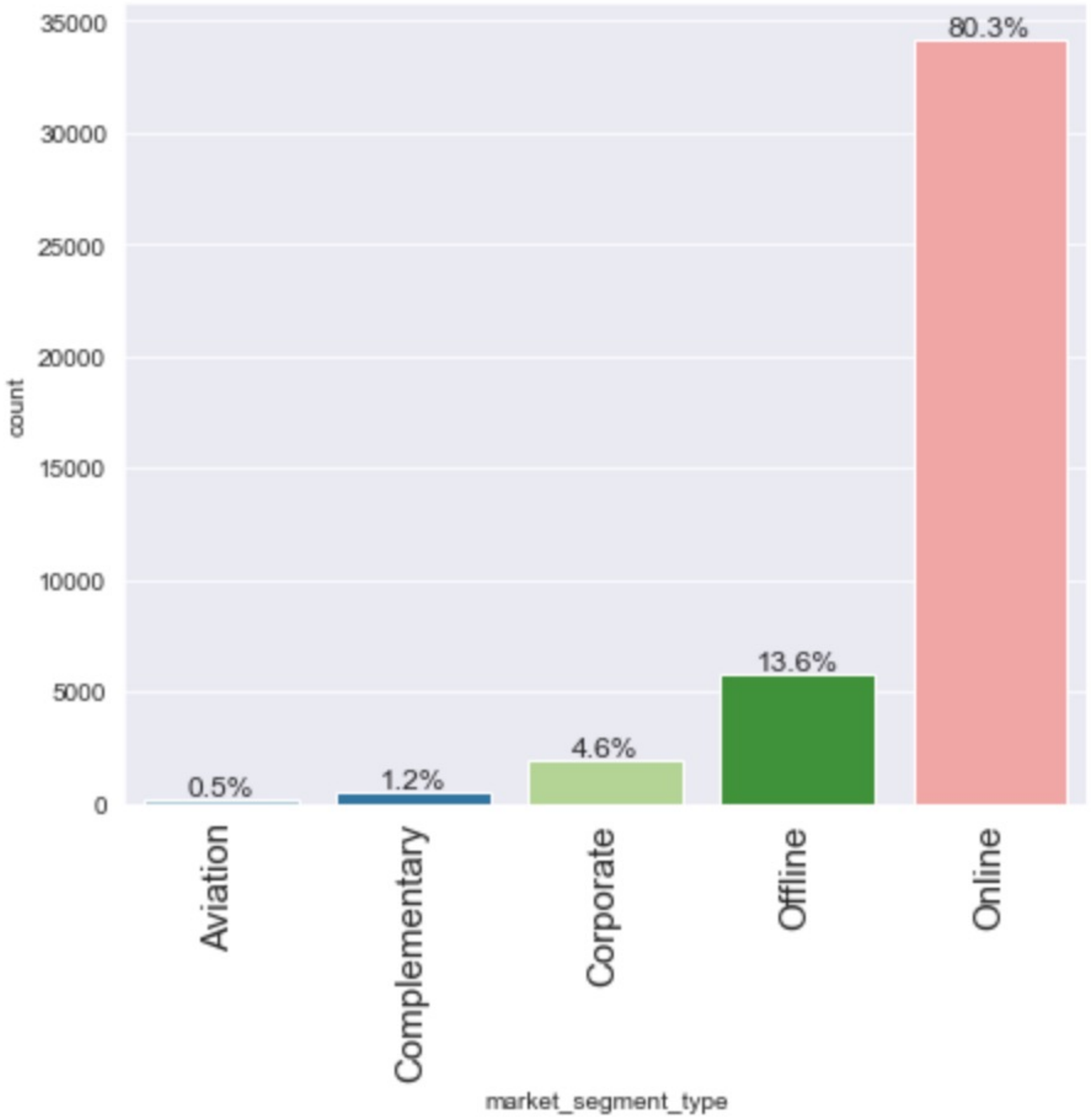
The busiest month is August

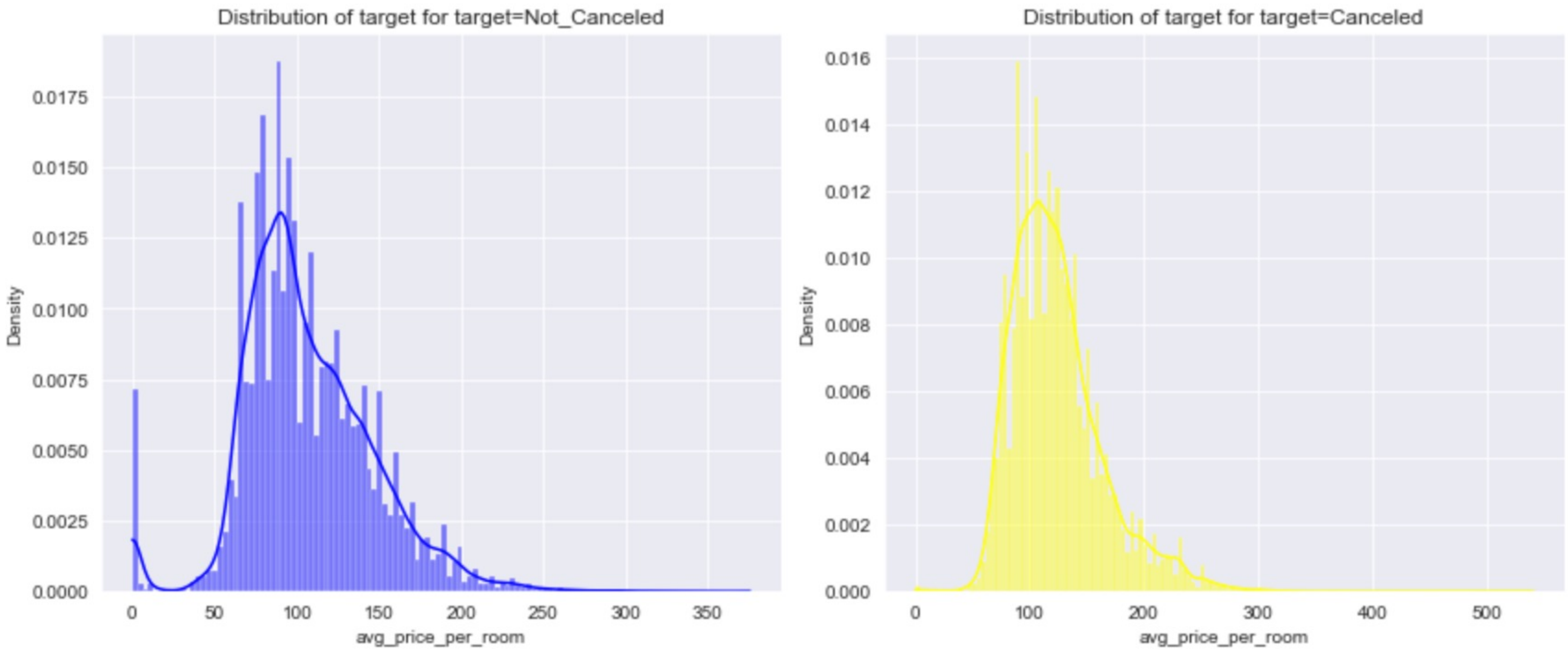
It's followed by July, May, April, March and June

October to January are the least busy months

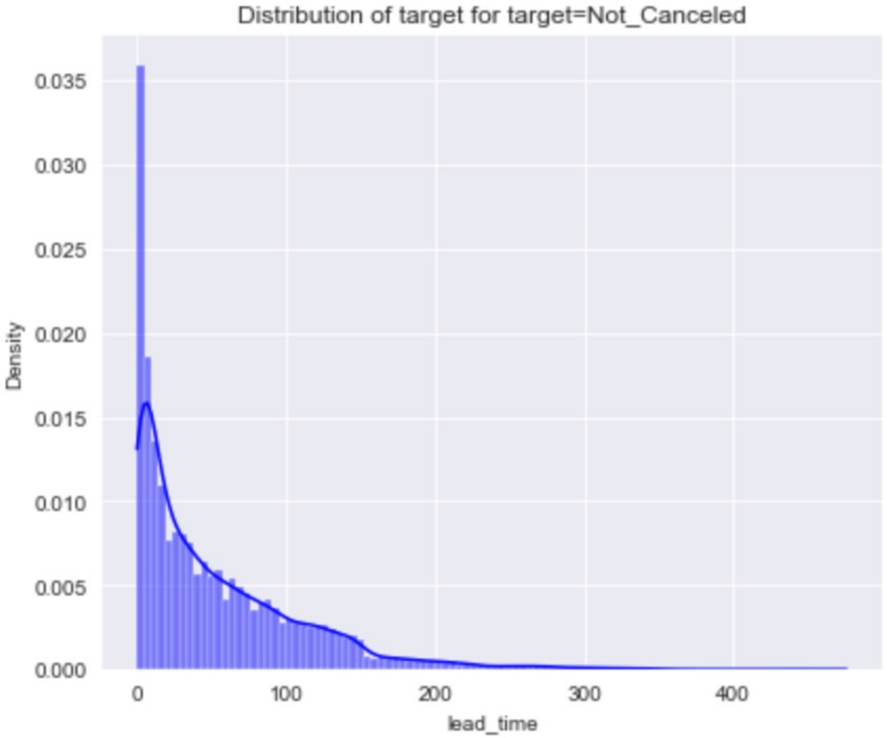
Market Segment

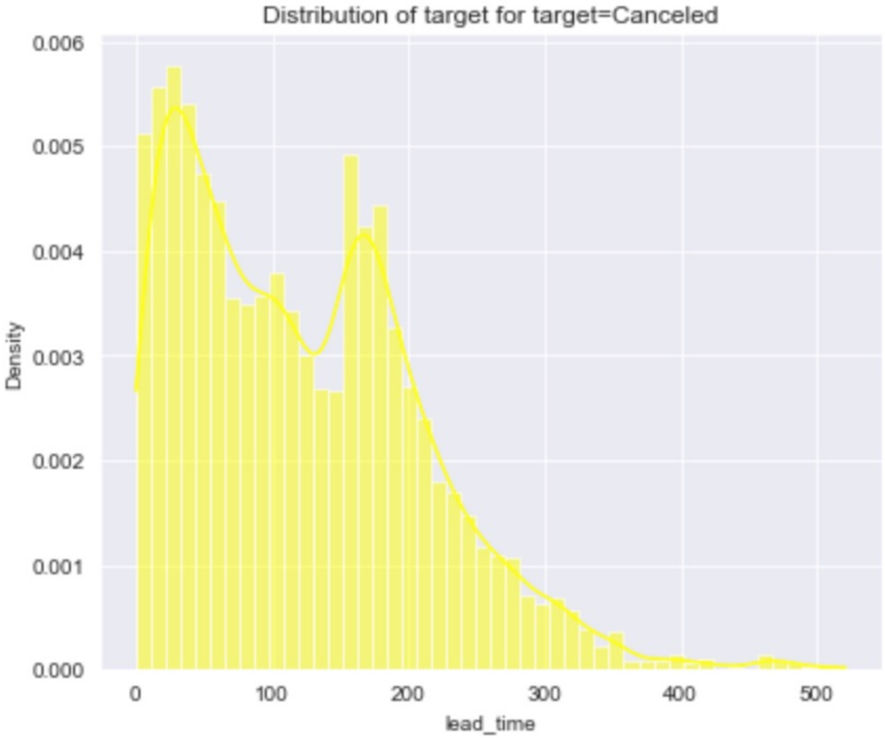
80% OF THE GUESTS COME FROM ONLINE MARKET SEGMENT





Average Price


Distribution of target for target=Not_Canceled
Density
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
100
200
300
400
lead_time

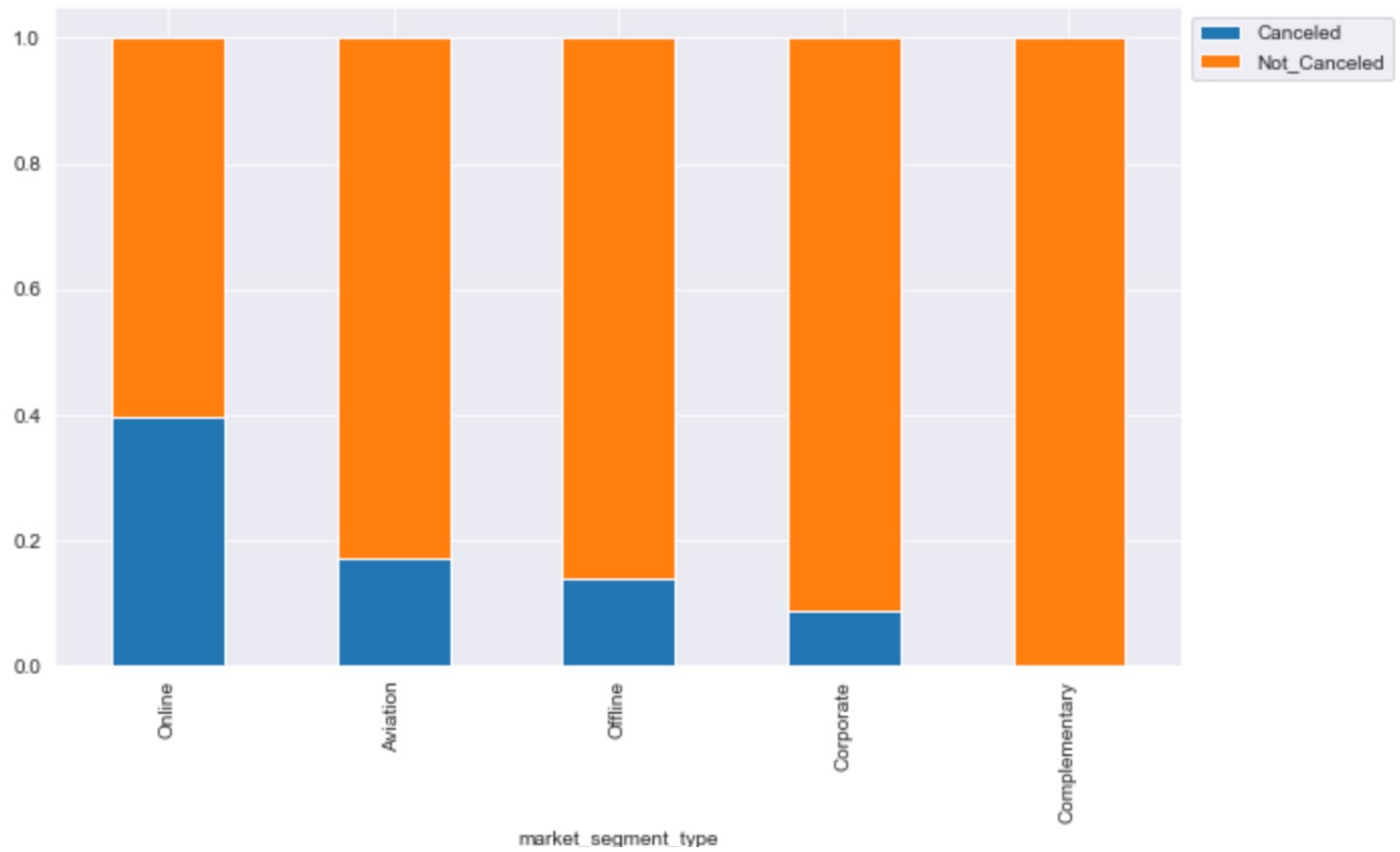

Distribution of target for target=Canceled
Density
0.006
0.005
0.004
0.003
0.002
0.001
0.000
0
100
200
300
400
500
lead_time

Lead Time

Market Segment - Booking Breakdown

No cancellations for complementary

Highest cancellations for Online



EDA Summary 1

Maximum of the bookings are for 2 adults

Bookings are made for 0-2 weekend nights are most popular

For the weeknights, the guests typically book for 2 weekdays

The average room price median is 110 and mean is 112

Meal Plan 1 is the most popular, 75% of the guests select this option

Room Type 1 has the most bookings 70%

Room Type 3 is is the least reserved type of room

80% of the guests make reservations online, followed by 13% of offline reservations

35% of overall booking gets cancelled incurring high amount loses for the hotel

EDA Summary 2

The average price per room is highly related with number of adults and children, which make sense as that’s how hotels determine pricing

Repeated guest, number of previous cancellations and number of previous bookings are highly correlated. That understandable as we only have previous information only available for repeated guests

The arrival date is uniform, there is slight drop in 31 reasonable as not all months have 31 date

The average price per room is higher as the lead time is reduces

The cancellations occurs only for weekend nights

Room type is not significant for cancellation

March August are the most popular months for the hotels

Chances of repeated guest cancelling is less than 1%

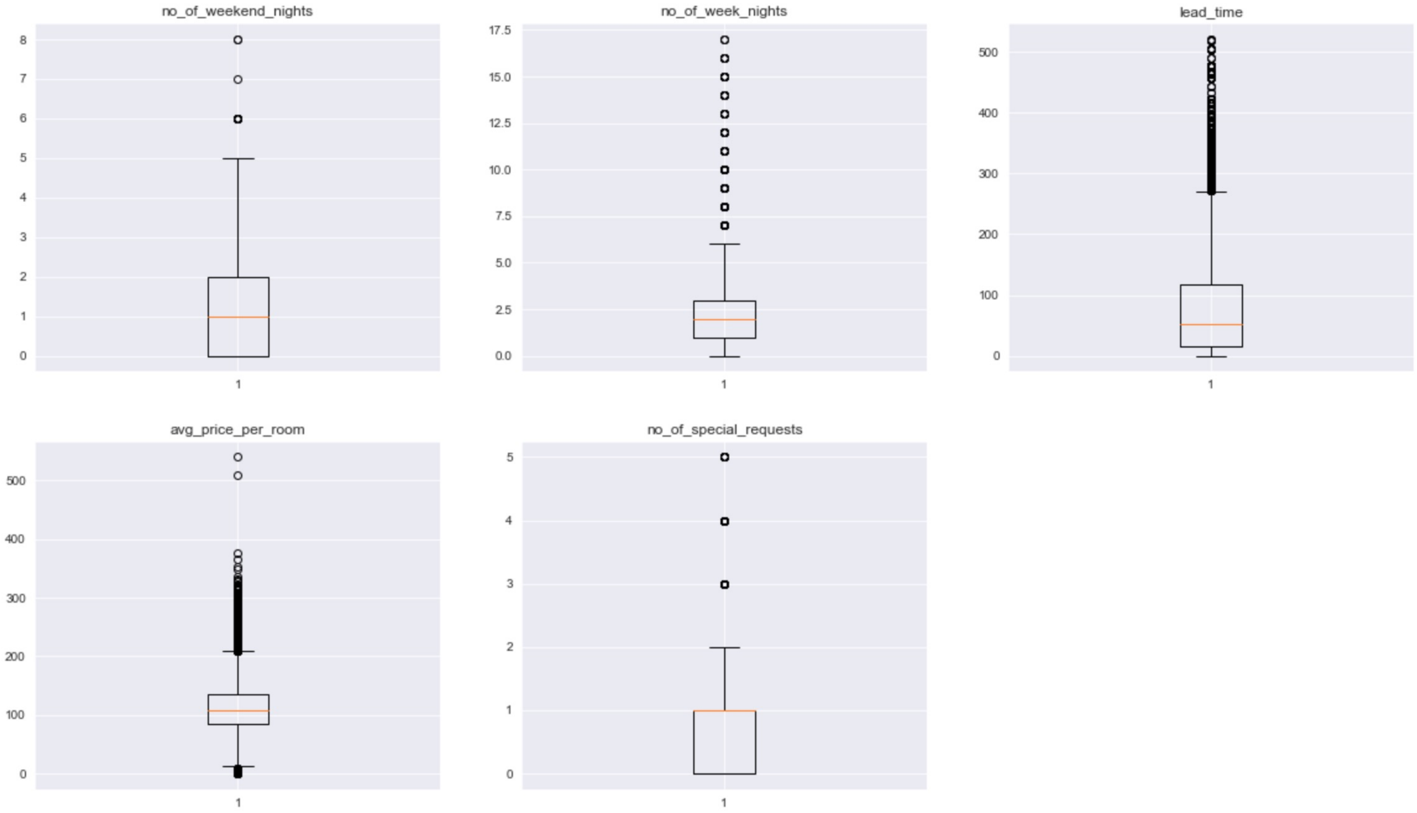
July and August are most popular months, however these months also have the maximum number of cancellations

There are significant outliers in lead time and average price per room. These need to be handled in data processing

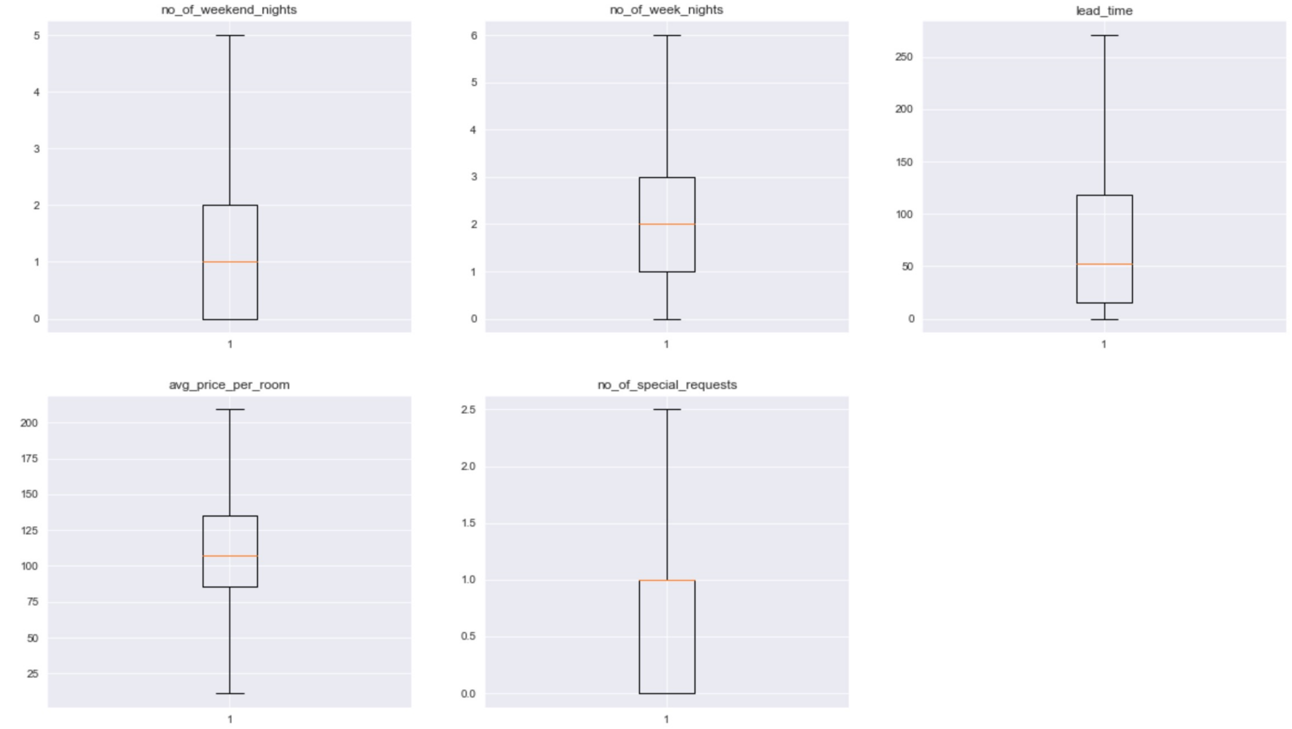
Data Preprocessing

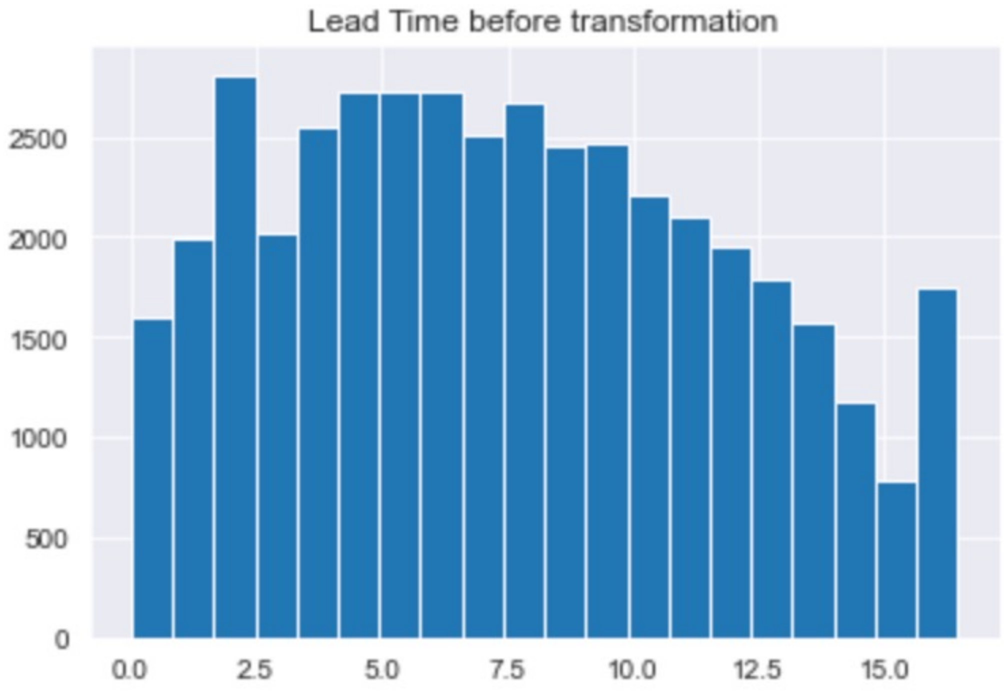
1. No missing values
2. Removed outliers for number of weekend nights, number of weeknights, lead time, average price per room and number of special requests (based on IQR)
3. SQRT transformation on Lead Time and Standardized Lead Time
4. Standardized Average Price Per Room

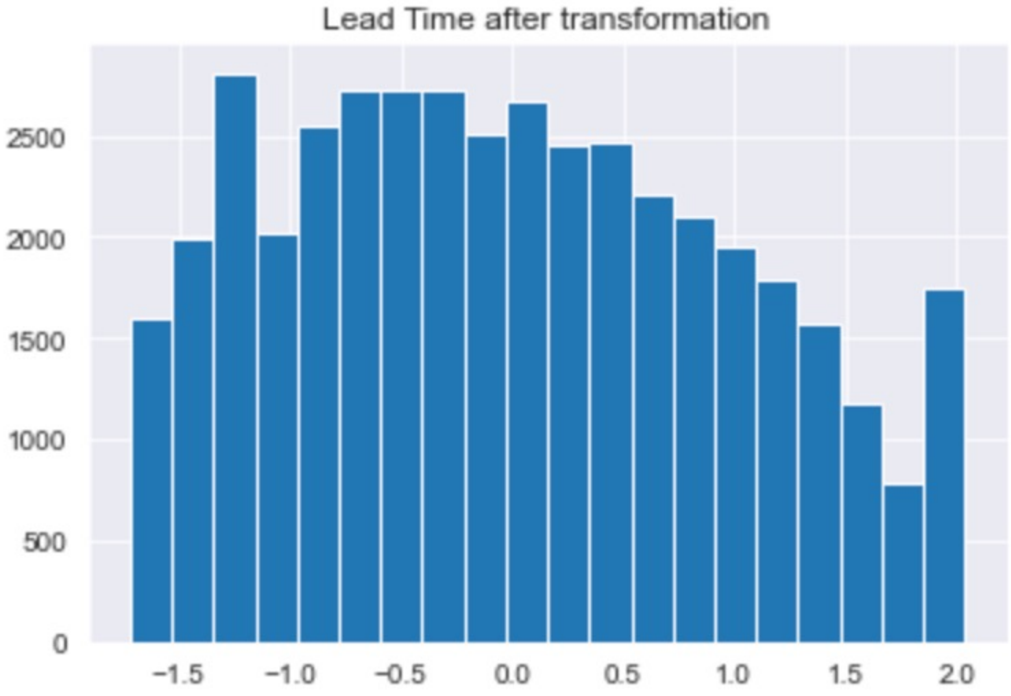
Outlier Detection



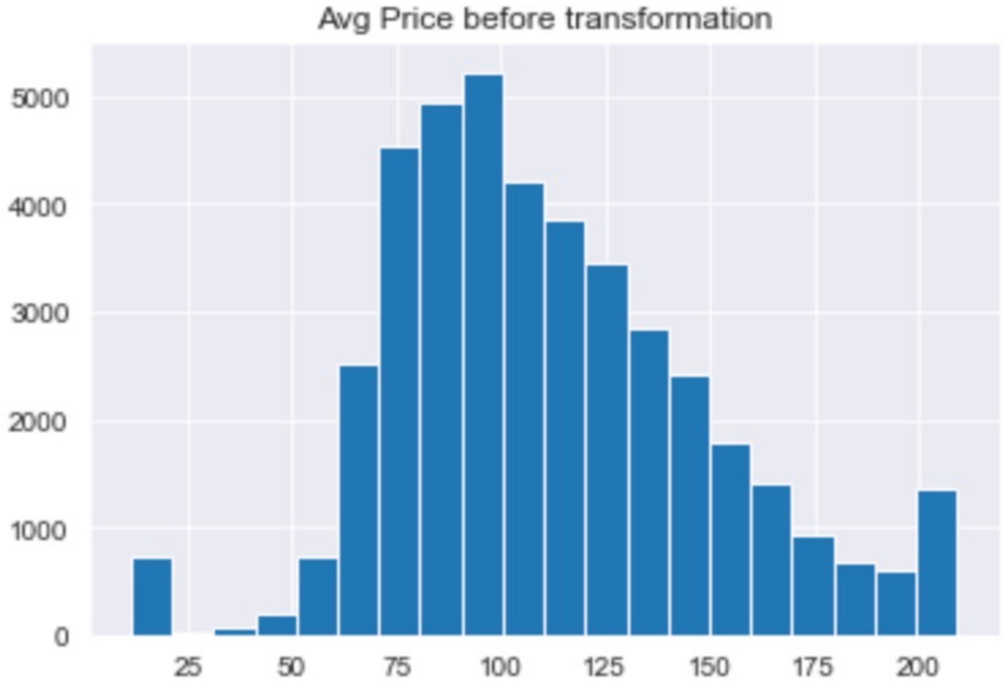
Outlier Treatment

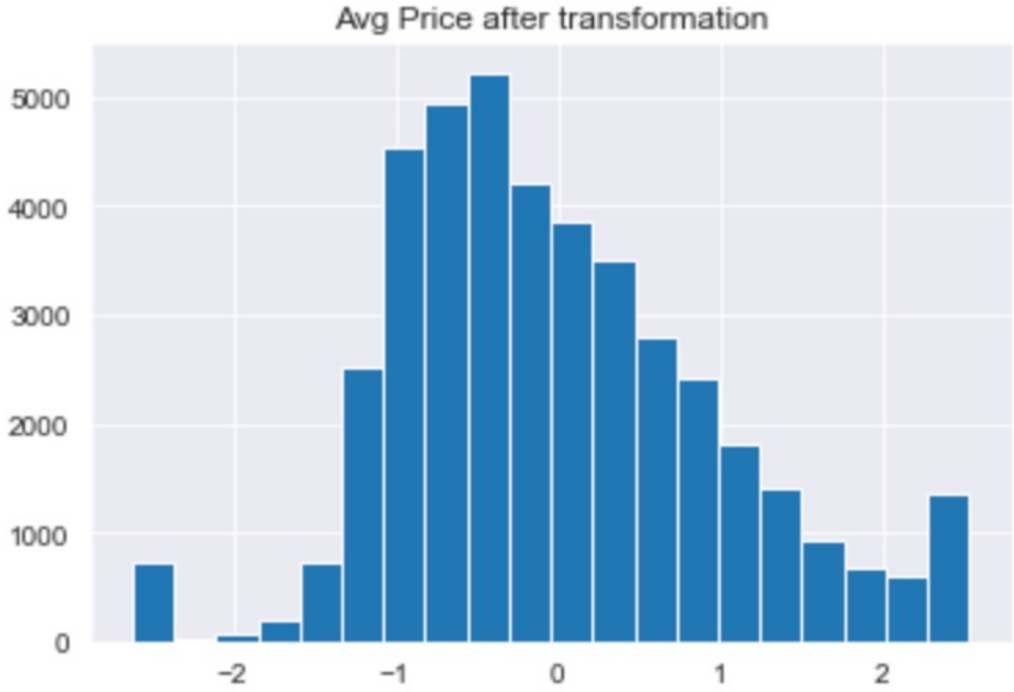






Transformations – Lead Time





Transformations – Average Price

Regressionm Model

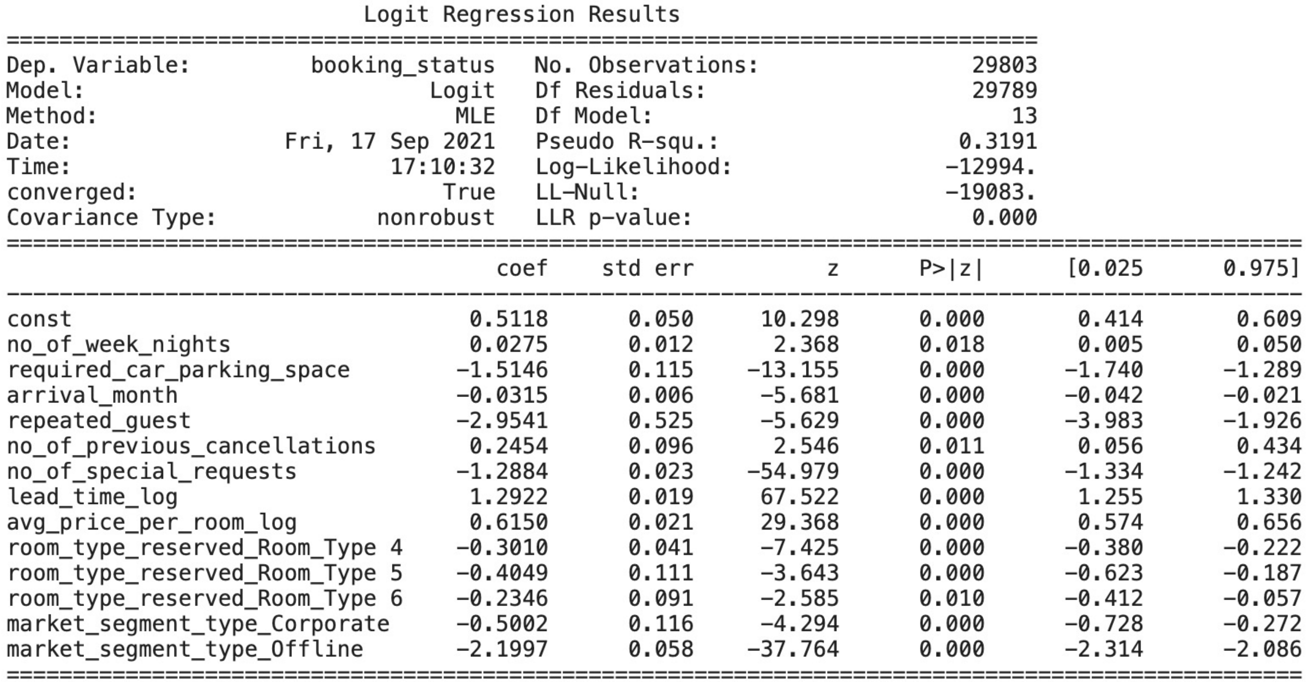
Dropped variables with high multicollinearity

Dropped variables with high p-values

Build model based on various thresholds

Finalized the model with best F1 score

Regression Model Summary



```
                           Logit Regression Results
==============================================================================
Dep. Variable:         booking_status   No. Observations:                29803
Model:                          Logit   Df Residuals:                    29789
Method:                           MLE   Df Model:                           13
Date:                Fri, 17 Sep 2021   Pseudo R-squ.:                  0.3191
Time:                        17:10:32   Log-Likelihood:                -12994.
converged:                       True   LL-Null:                       -19083.
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                            0.5118      0.050     10.298      0.000       0.414       0.609
no_of_week_nights                0.0275      0.012      2.368      0.018       0.005       0.050
required_car_parking_space      -1.5146      0.115    -13.155      0.000      -1.740      -1.289
arrival_month                   -0.0315      0.006     -5.681      0.000      -0.042      -0.021
repeated_guest                  -2.9541      0.525     -5.629      0.000      -3.983      -1.926
no_of_previous_cancellations     0.2454      0.096      2.546      0.011       0.056       0.434
no_of_special_requests          -1.2884      0.023    -54.979      0.000      -1.334      -1.242
lead_time_log                    1.2922      0.019     67.522      0.000       1.255       1.330
avg_price_per_room_log           0.6150      0.021     29.368      0.000       0.574       0.656
room_type_reserved_Room_Type 4  -0.3010      0.041     -7.425      0.000      -0.380      -0.222
room_type_reserved_Room_Type 5  -0.4049      0.111     -3.643      0.000      -0.623      -0.187
room_type_reserved_Room_Type 6  -0.2346      0.091     -2.585      0.010      -0.412      -0.057
market_segment_type_Corporate   -0.5002      0.116     -4.294      0.000      -0.728      -0.272
market_segment_type_Offline     -2.1997      0.058    -37.764      0.000      -2.314      -2.086
================================================================================================
```

Regression Model Coefficients

Number of weeknights, number of previous cancellations, lead time, and average price per room have positive impact, will increase chances of cancellations.

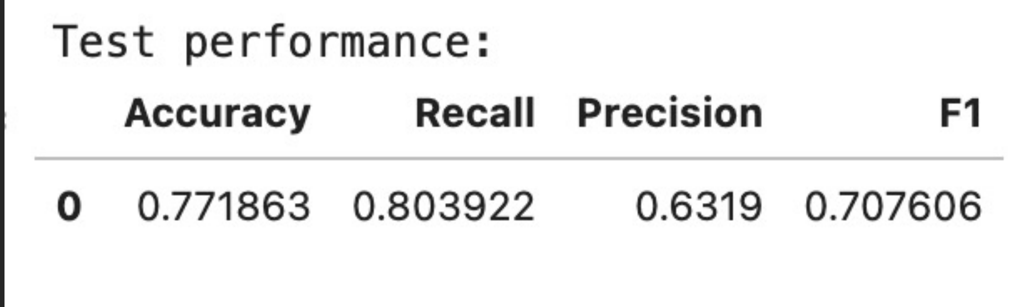
Required car parking space, arrival month, repeated guest, number of special requests, room type 4, room type 5, room type 6, market segment corporate and market segment offline have negative impact, will reduce chances of cancellations.

Regression Model Performance

Using the threshold value 0.31 gives the best model

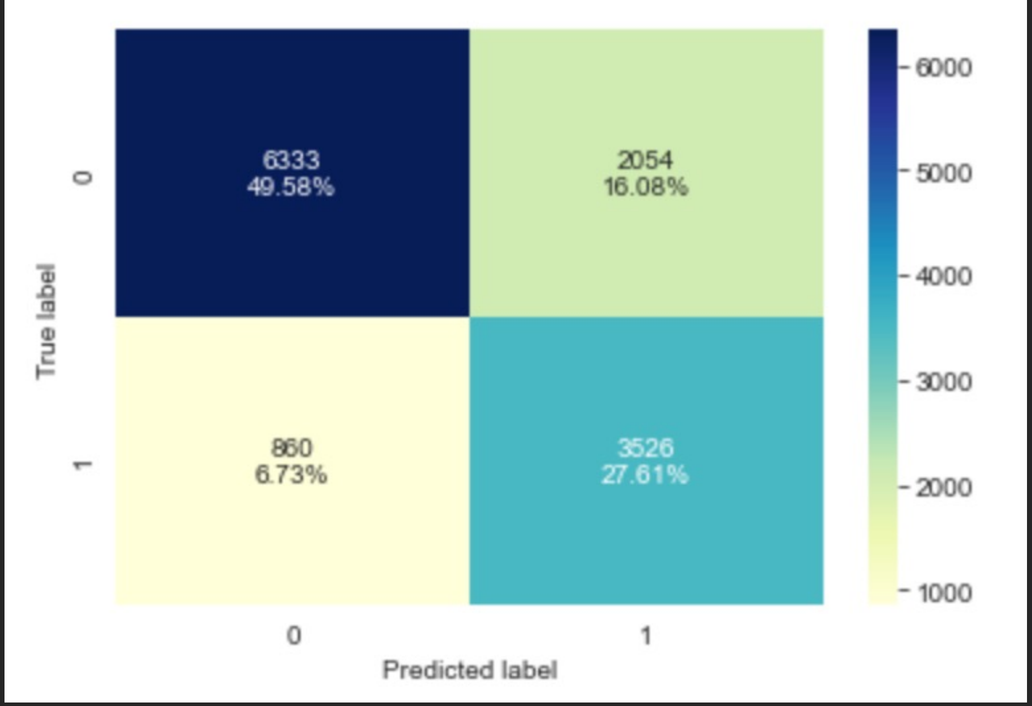
We get best Recall and F1 scores using this threshold value

Showing confusion matrix and performance metrics on the test data on the right



Test performance:

	Accuracy	Recall	Precision	F1
0	0.771863	0.803922	0.6319	0.707606



Regression Model Summary

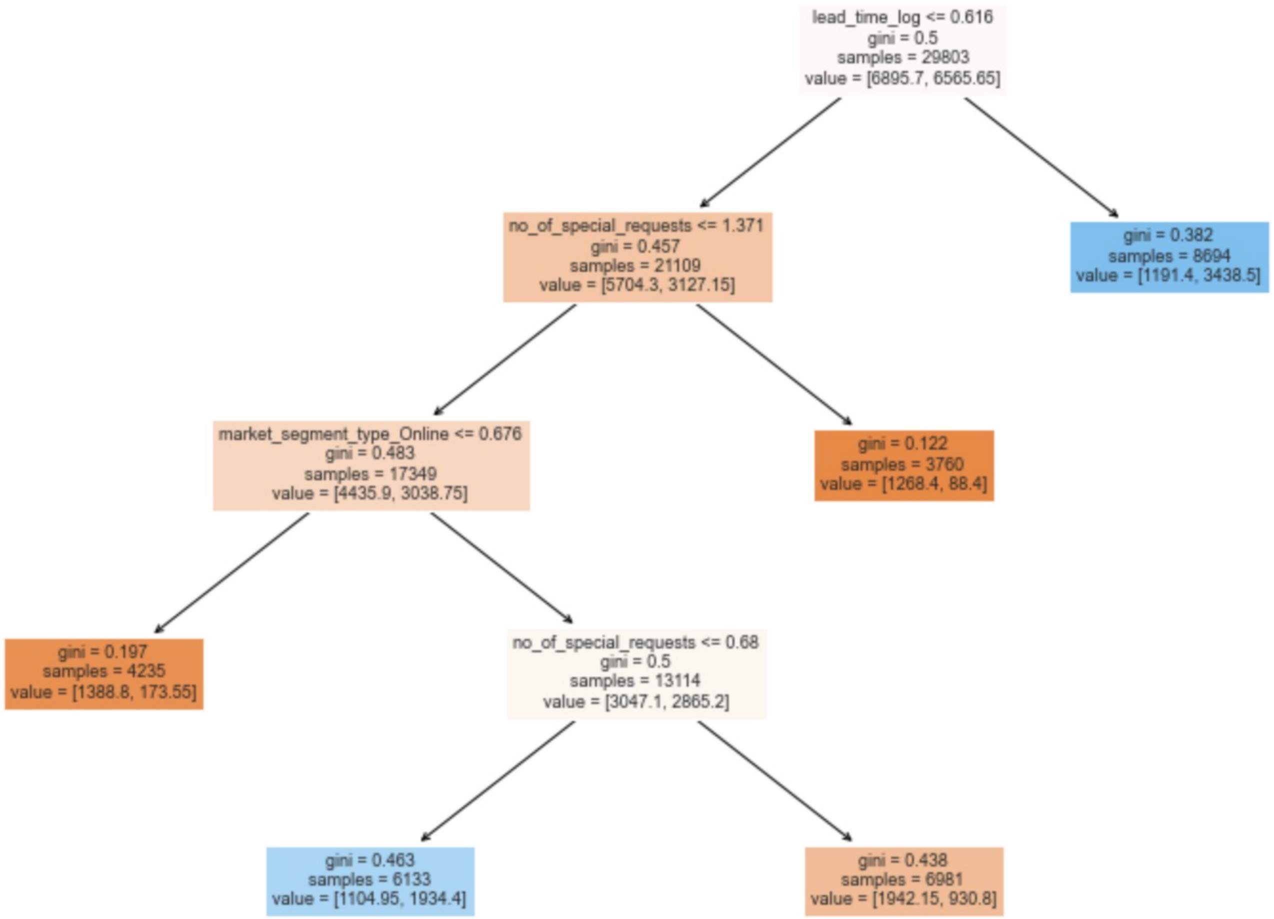
We have been able to build a predictive model that can be used by the star hotels to find the bookings that can may result into cancellation with an f1_score of 0.70 on the training set and the testing set.

All the logistic regression models have given a generalized performance on the training and test set.

Coefficient of number of weeknights, having previous cancellation, large lead time and higher average price per room are positive and an increase in these will lead to increase in chances of a cancellation.

Coefficient of required car parking, arrival month, repeated guest, number of special request, room type 4, 5 and 6 and market segment corporate and offline are negative increase in these will lead to decrease in chances of a cancellation.

Decision Tree Model



Decision Tree Performance

The estimator provided the best tree

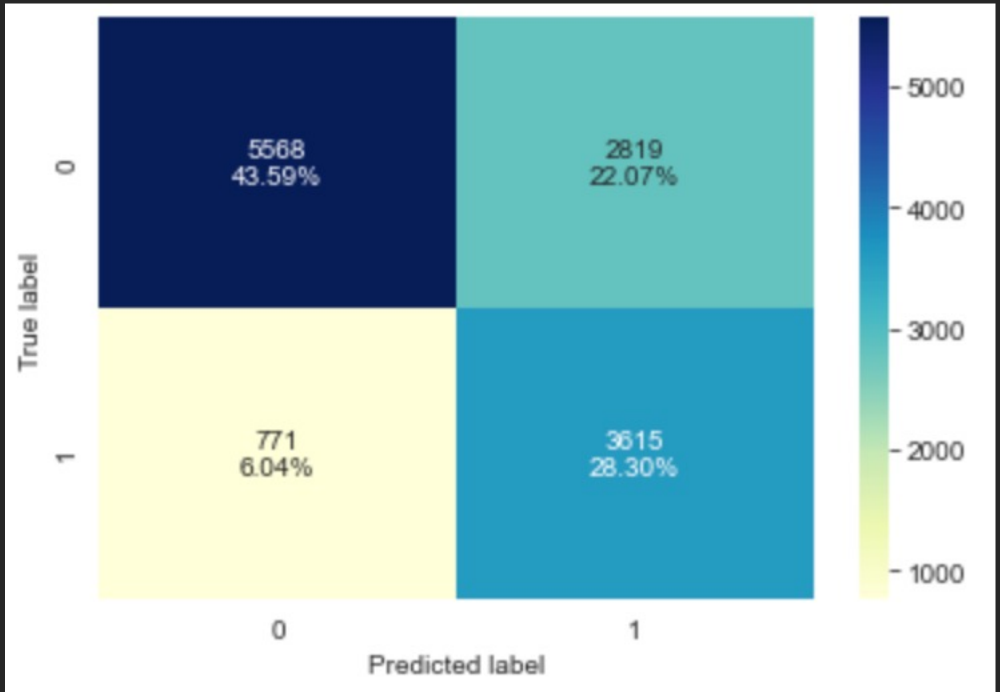
The decision tree max depth is 5

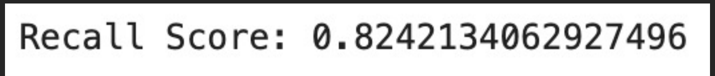
Random splitter is used for the decision tree

Showing confusion matrix and performance metrics on the test data on the right

The most important features are:

1. Lead Time
2. Number of Special Requests
3. Market Segment Online





```
Recall Score: 0.8242134062927496
```

Decision Tree Summary

Decision tree model with pre-pruning has given the best recall score on training data.

The pre-pruned and the post-pruned models have reduced overfitting and the model is giving a generalized performance.

The post-pruned model gives good result, but we'll select the tree from pre-pruning as it's the simple and efficient model.

The lead time is significant feature that drives cancellations the most.

The longer the lead time the higher possibility of cancellations.

Special requests affect the cancellations. If there are more special requests, it's cancellations is unlikely

Market segment online is important feature that determines cancellations. Market segments other than online more likely not to cancel.

Business Recommendations

The **lead time** is a critical metric for cancellations. Longer the lead time higher are the chances of cancellations. The hotel can introduce policy for no refunds for longer bookings. The hotel can also add conditions like no cancellations allowed within 30 days of arrival date.

The next important feature is special requests. The guests who make 3 or more special requests do no cancel their bookings. The hotel should take special requests into consideration. Offering more options for the guests and providing guest service accordingly.

The number of repeated guests are on the lower side. The hotel should provide incentives for customers for revisit. The repeated guests rarely cancel their booking. By creating loyalty, they can increase LTV of repeated guests.

The guests who book online have highest number of cancellations. Policies can be implemented to avoid cancellations by securing deposit, partial refunds, etc.

The percentage of repeated guests is low. The hotel should gather feedbacks for new guests and do improvements to encourage repeat customers