Star Hotel Project

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:

- Loss of resources (revenue) when the hotel cannot resell the room.
- 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 make arrangements for the guests.

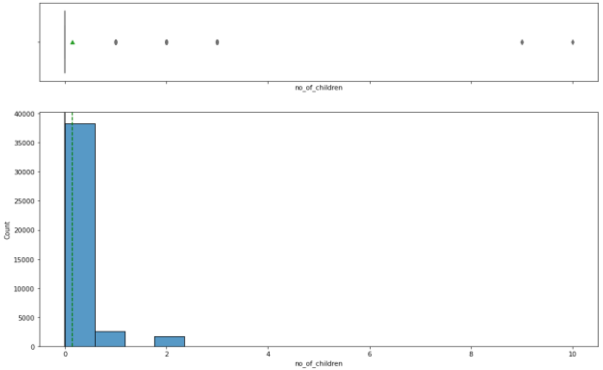
Objective

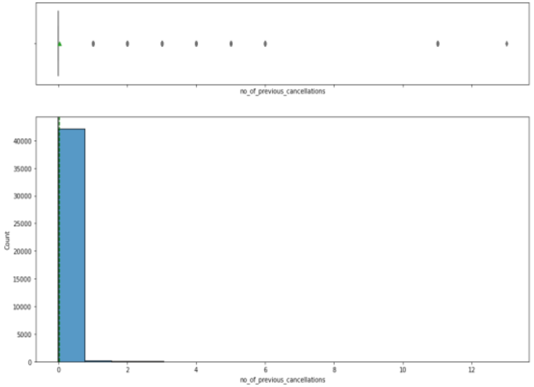
The increasing number of cancellations calls for a Machine Learning based solution that can help in predicting which booking is likely to be canceled. Star Hotels Group has a chain of hotels in Portugal, they are facing problems with the high number of booking cancellations and have reached out to your firm for data-driven solutions. You as a 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.

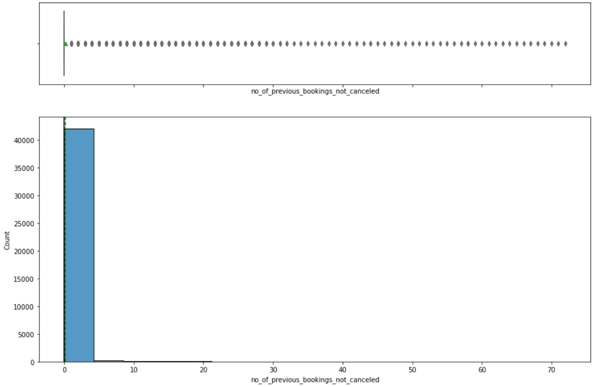
Data Information

Variable	Description
no_of_adults	Number of Adults
no_of_children	Number of Children
no_of_weekend_nights	Number of Weekend nights (Sat. and Sunday)
no_of_week_nights	Number of Weeknights (Mon. and Friday)
type_of_meal_plan	Type of meal plan booked by the customer
required_car_parking_space	Plan 2 – Half board (breakfast and one other meal)
room_type_reserved	Type of room reserved by the customer.
lead_time	Number of days between the date of booking/arrival
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
no_of_previous_bookings_not_canceled	Number of previous bookings not canceled
avg_price_per_room	Average price per day of the reservation
no_of_special_requests	Total number of special requests made by the cust.
booking_status	Flag indicating if the booking was canceled or not.

Exploratory Data Analysis



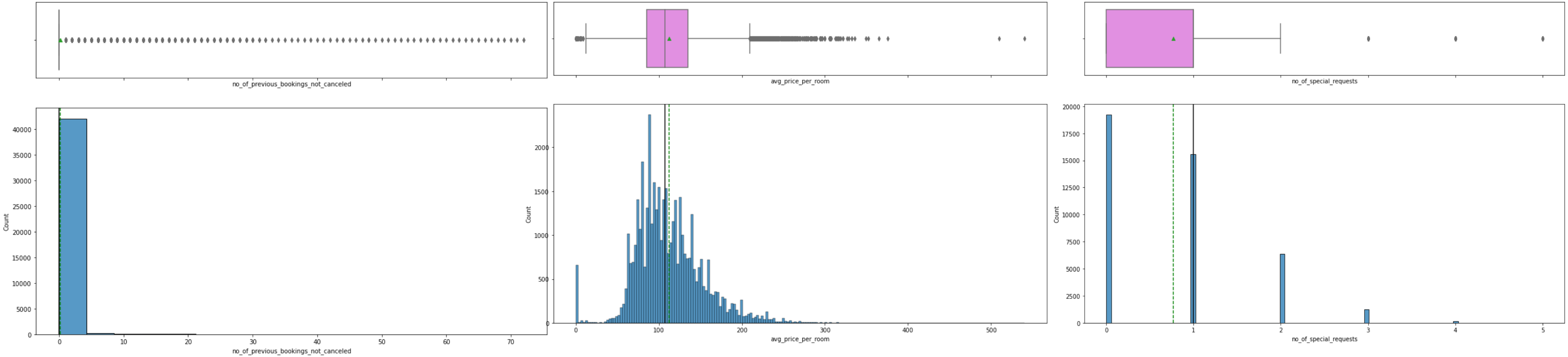




Observations;

- Above we can see that the average and median number of children are around zero. There do appear to be outliers going up all the way to 10 children.
- On average, most people have not cancelled their reservation previously. There are some outliers that go as high as 12 prior cancellations.
- Here we see that there on average there are zero previous bookings not cancelled, but we do see multiple outliers going above 70 of previous bookings not cancel

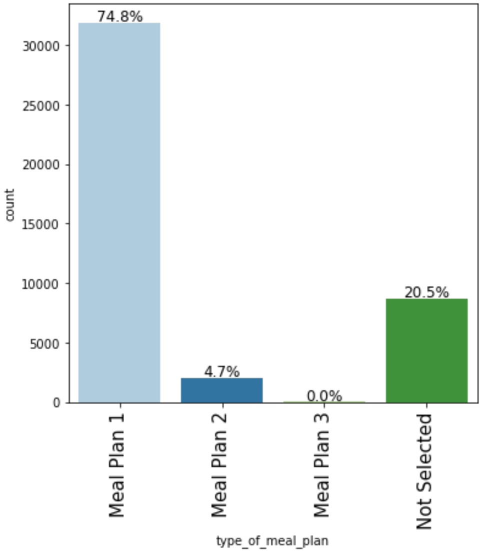
Exploratory Data Analysis

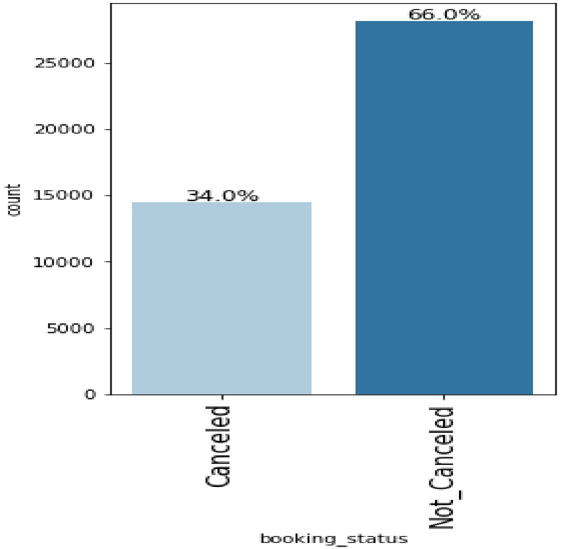


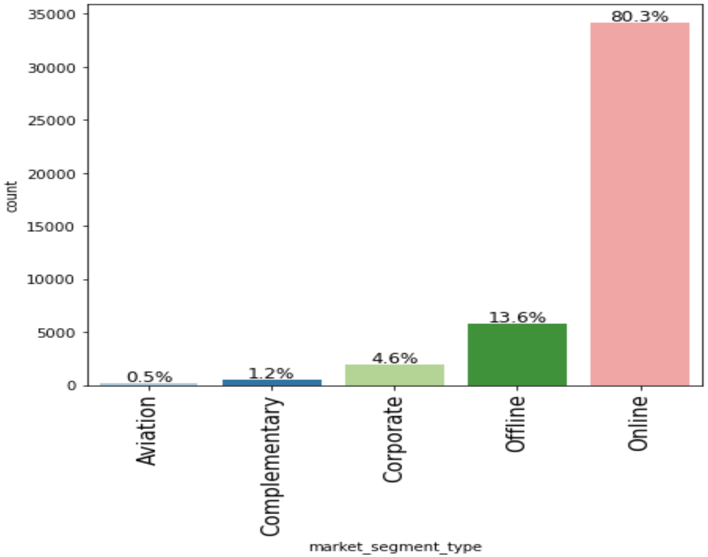
Observations;

- The average and median are close at about $100 per room. There are outliers on both sides.
- Here we see that there on average there are zero previous bookings not cancelled, but we do see multiple outliers going above 70 of previous bookings not cancel
- Most of the data lies between 0 and 1 number of special requests. There are outliers to the right with some people making multiple special requests.
- The average number of days a booking is made before the reservation is around 90 days or about 4 months. There are several outliers on the right which indicate that people are booking over a year in advance. This is a right skewed graph.

Exploratory Data Analysis



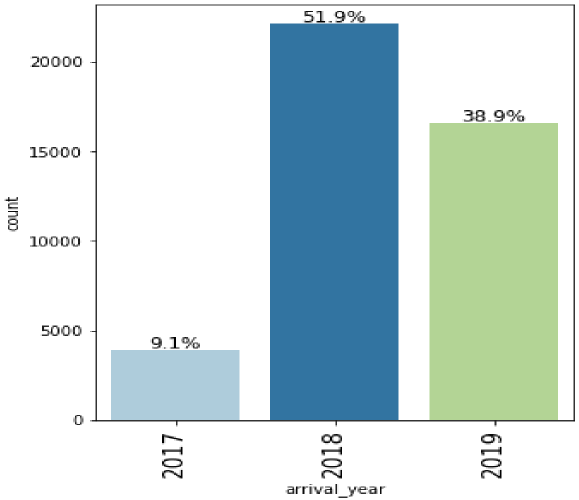


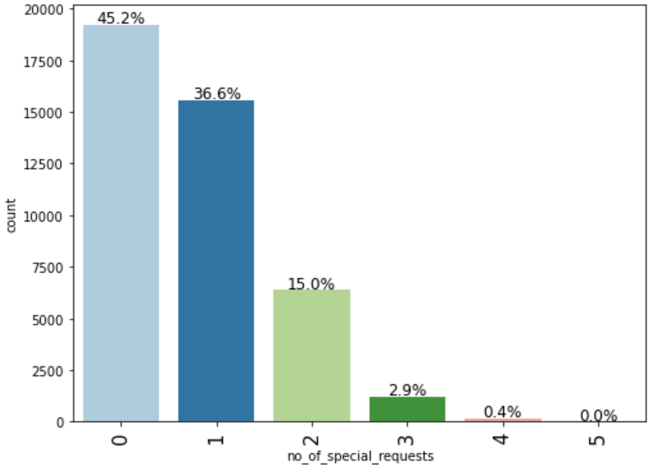


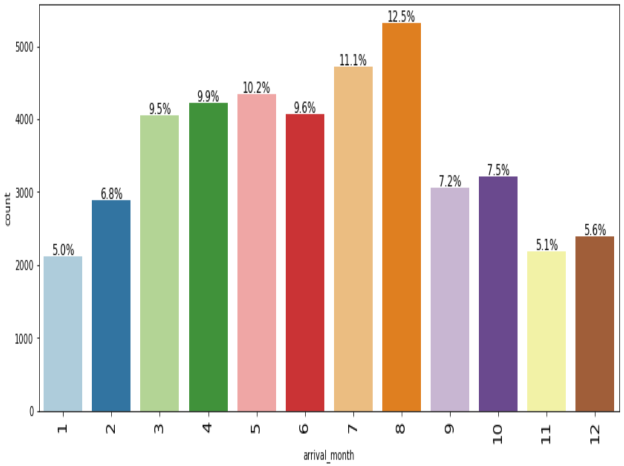
Observations;

- We see that there are more bookings that are not canceled vs canceled. There are 34.0% that are canceled and 66.0% that are NOT canceled.
- Meal plan 1 is the most order type of meal.
- For the market segment online bookings had the higher percentage with 80.3%

Exploratory Data Analysis



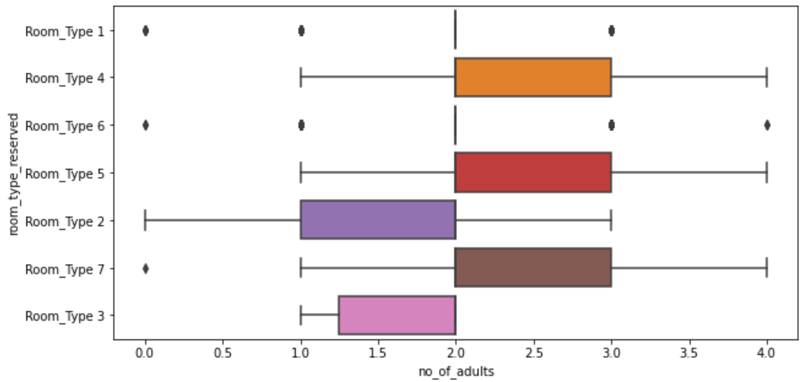


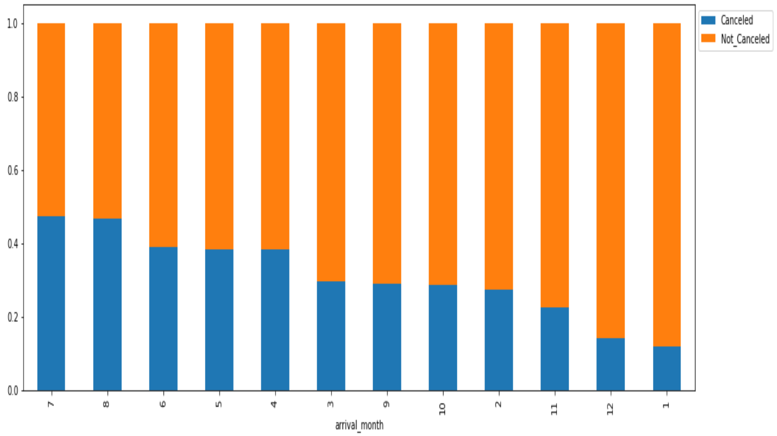


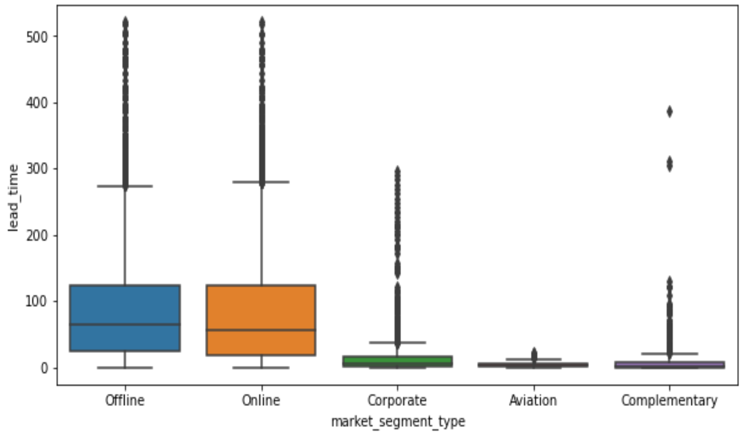
Observations;

- 2018 was the highest year of arrival
- 54.8%, nearly half, of all bookings require a special request.
- August was the highest arrival month which simply means the Hotel saw a higher rate of visitors visiting during that month.

Exploratory Data Analysis





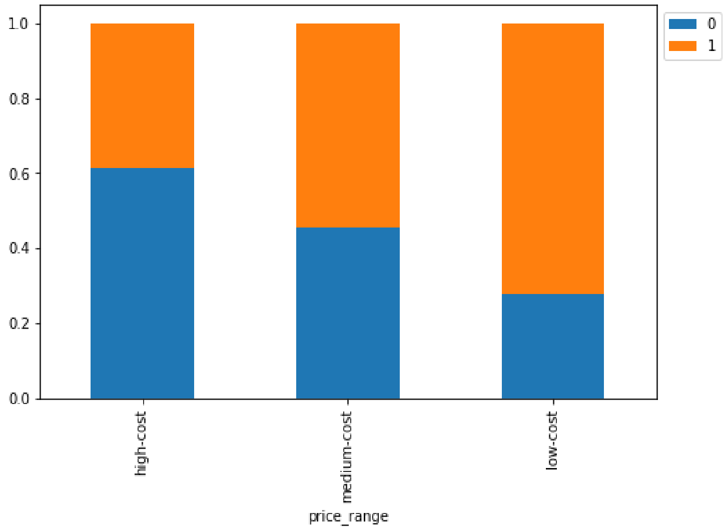


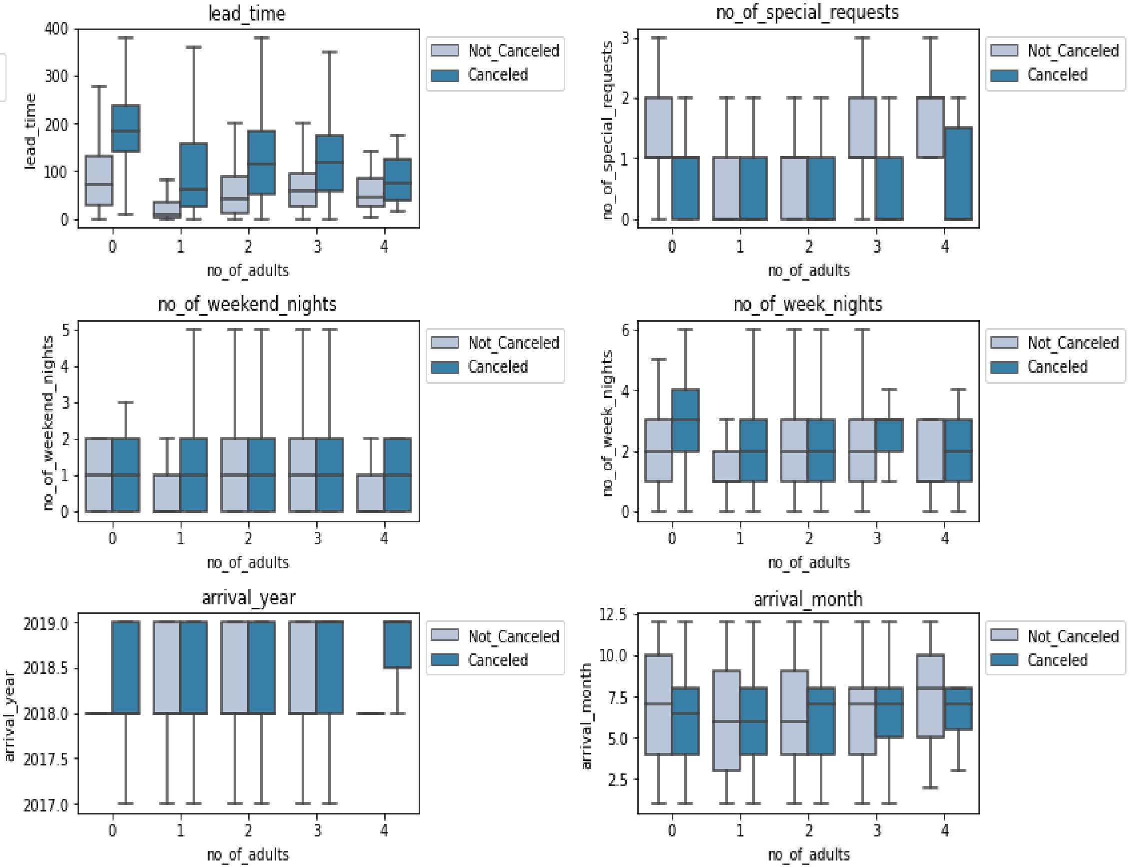
Observations;

- August is the busiest month of all, both Cancelled and not Cancelled bookings.

- We have outliers from both our market_segment_type and no_of_adult .

Exploratory Data Analysis

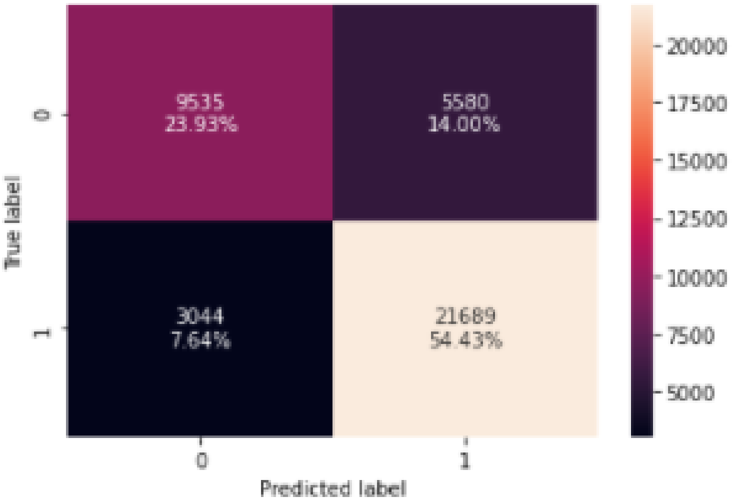


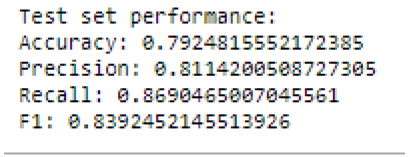


Observation

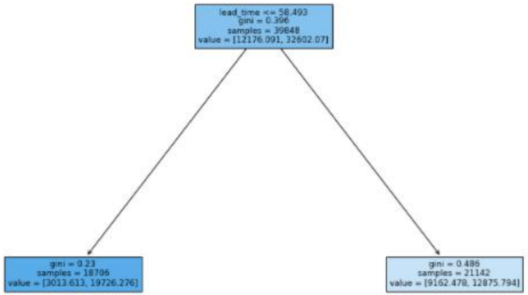
- It appears that high-cost rooms are more likely to cancel as their cancellation rate is higher than those who don't. Medium cost and low-cost rooms have more people that do not cancel than those who do cancel.

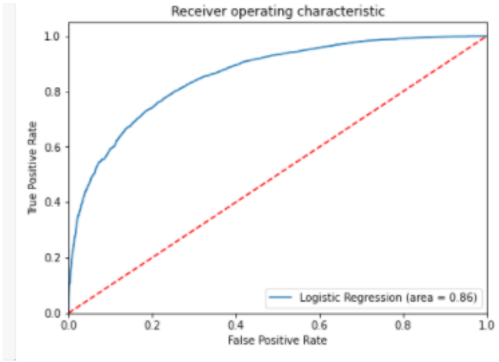
Model Prediction





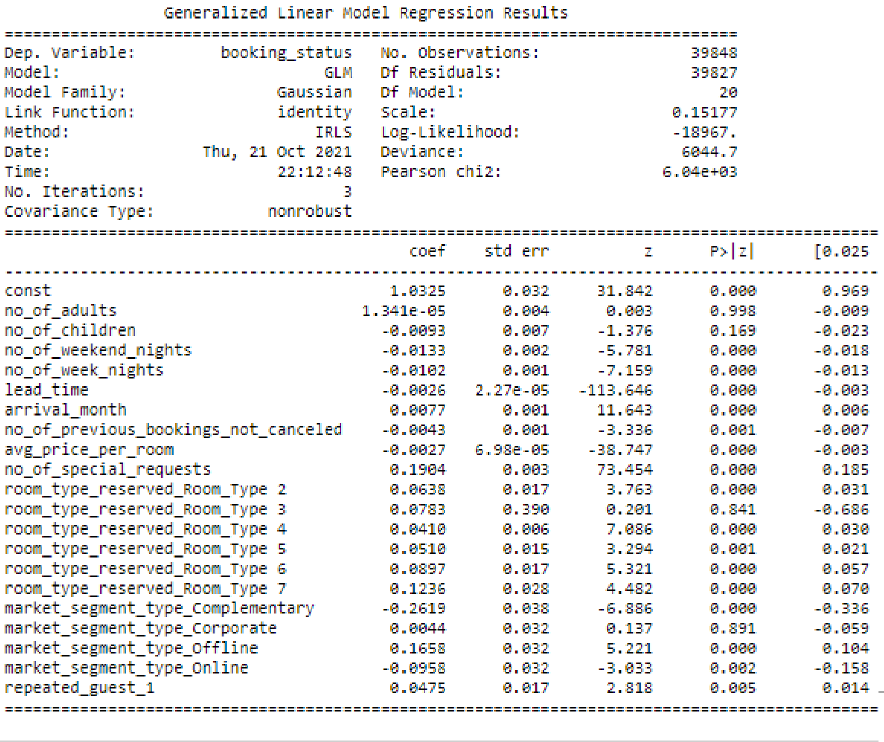
```
Test set performance:
Accuracy: 0.7924815552172385
Precision: 0.8114200508727305
Recall: 0.8690465007045561
F1: 0.8392452145513926
```



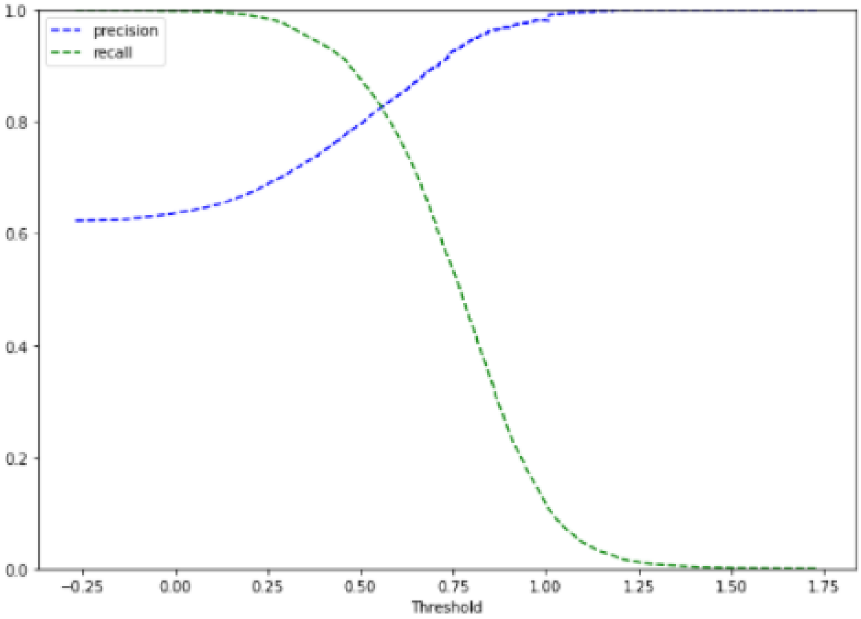


- Logistic Regression model is giving a good performance on training set.

Model Predictions

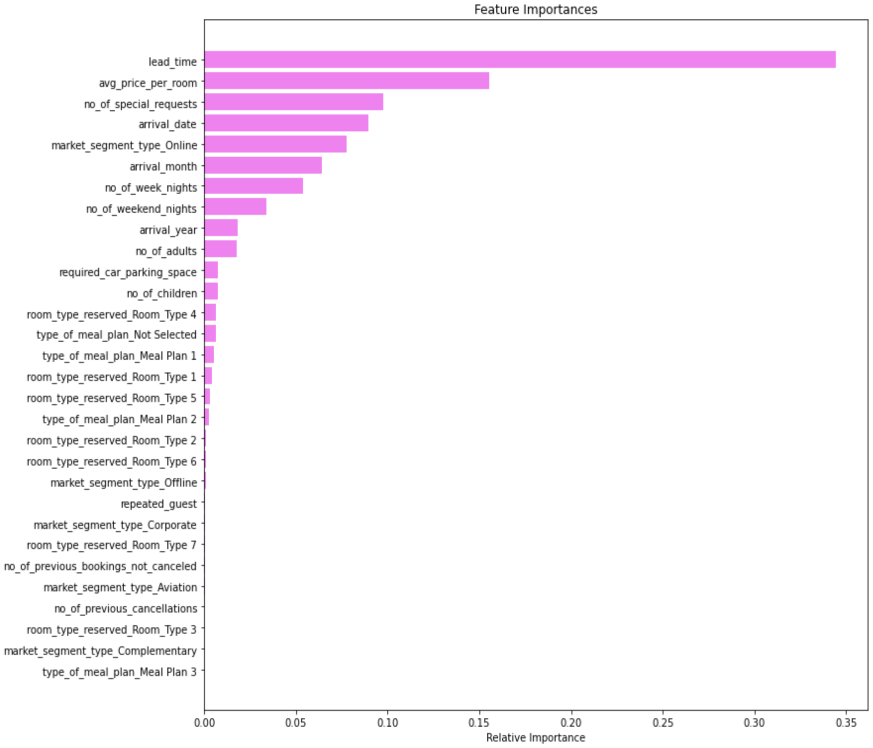


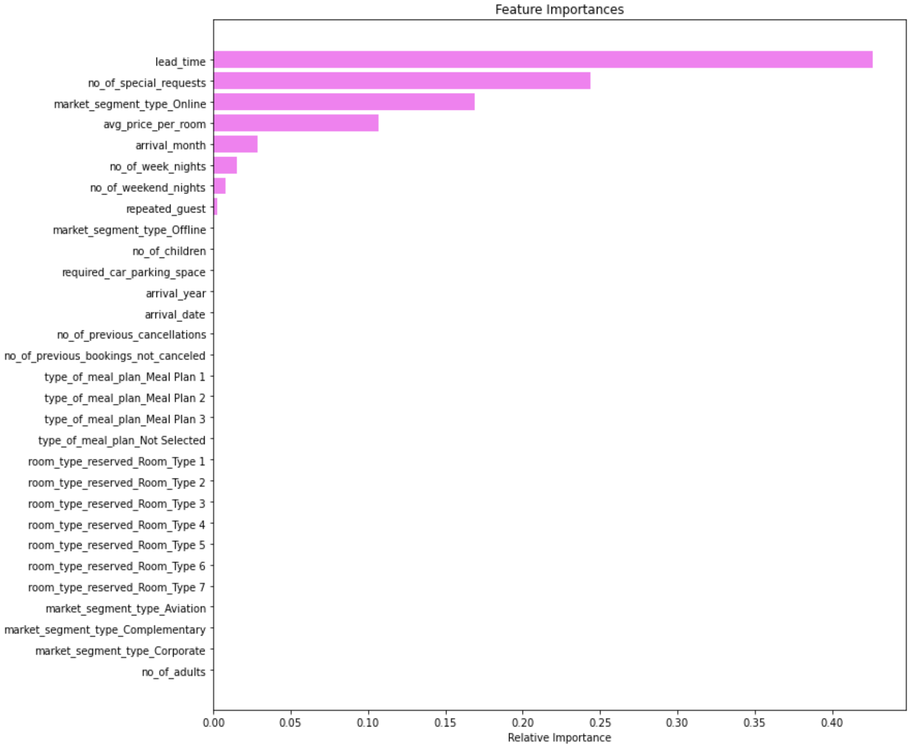
```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:          booking_status   No. Observations:                39848
Model:                             GLM   Df Residuals:                    39827
Model Family:                 Gaussian   Df Model:                           20
Link Function:                identity   Scale:                         0.15177
Method:                           IRLS   Log-Likelihood:                -18967.
Date:                 Thu, 21 Oct 2021   Deviance:                       6044.7
Time:                         22:12:48   Pearson chi2:                 6.04e+03
No. Iterations:                      3
Covariance Type:             nonrobust
===============================================================================================
                                        coef    std err          z      P>|z|      [0.025
-----------------------------------------------------------------------------------------------
const                                 1.0325      0.032     31.842      0.000       0.969
no_of_adults                       1.341e-05      0.004      0.003      0.998      -0.009
no_of_children                       -0.0093      0.007     -1.376      0.169      -0.023
no_of_weekend_nights                 -0.0133      0.002     -5.781      0.000      -0.018
no_of_week_nights                    -0.0102      0.001     -7.159      0.000      -0.013
lead_time                            -0.0026   2.27e-05   -113.646      0.000      -0.003
arrival_month                         0.0077      0.001     11.643      0.000       0.006
no_of_previous_bookings_not_canceled -0.0043      0.001     -3.336      0.001      -0.007
avg_price_per_room                   -0.0027   6.98e-05    -38.747      0.000      -0.003
no_of_special_requests                0.1904      0.003     73.454      0.000       0.185
room_type_reserved_Room_Type 2        0.0638      0.017      3.763      0.000       0.031
room_type_reserved_Room_Type 3        0.0783      0.390      0.201      0.841      -0.686
room_type_reserved_Room_Type 4        0.0410      0.006      7.086      0.000       0.030
room_type_reserved_Room_Type 5        0.0510      0.015      3.294      0.001       0.021
room_type_reserved_Room_Type 6        0.0897      0.017      5.321      0.000       0.057
room_type_reserved_Room_Type 7        0.1236      0.028      4.482      0.000       0.070
market_segment_type_Complementary    -0.2619      0.038     -6.886      0.000      -0.336
market_segment_type_Corporate         0.0044      0.032      0.137      0.891      -0.059
market_segment_type_Offline           0.1658      0.032      5.221      0.000       0.104
market_segment_type_Online           -0.0958      0.032     -3.033      0.002      -0.158
repeated_guest_1                      0.0475      0.017      2.818      0.005       0.014
===============================================================================================
```



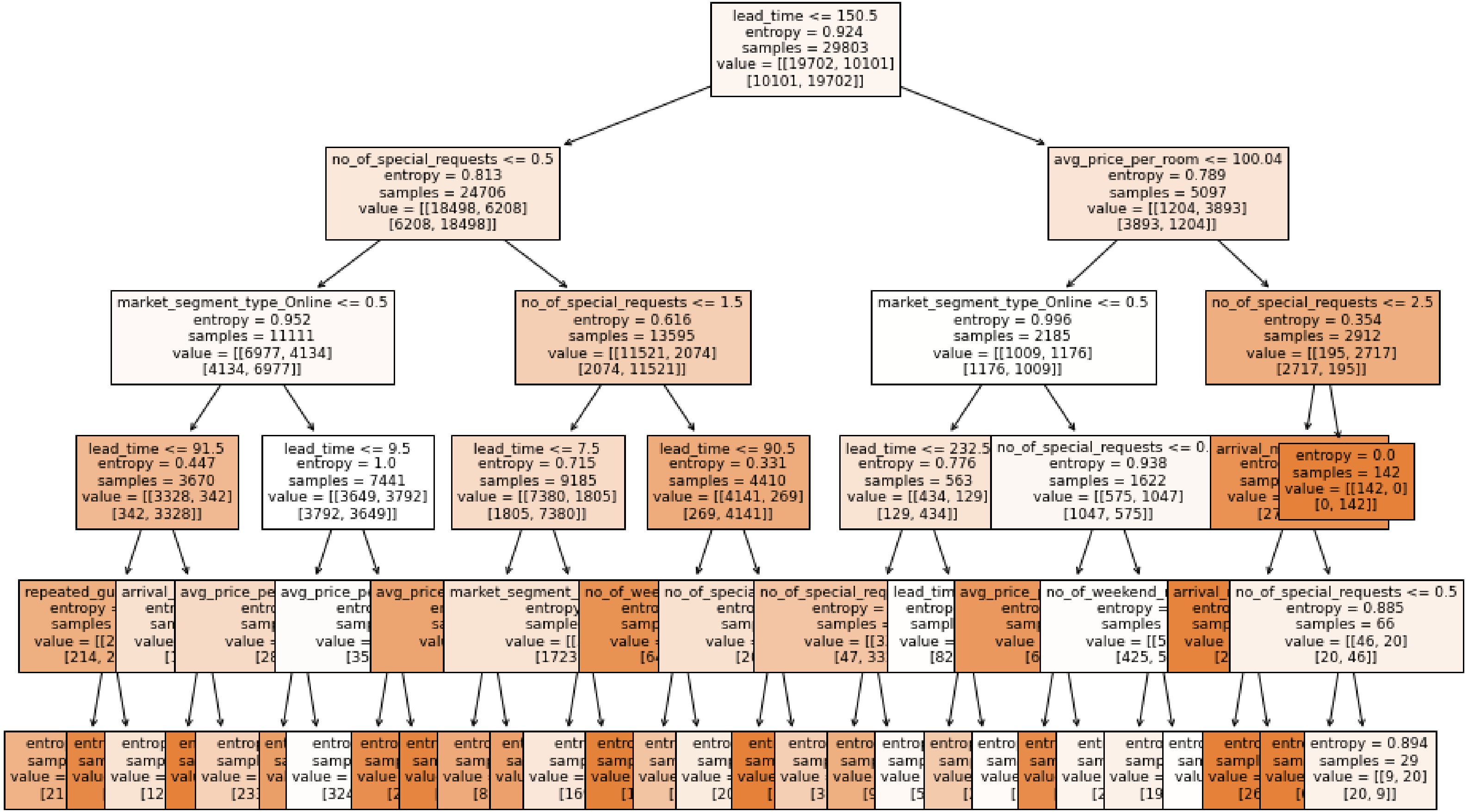
- At the threshold of 0.6, we get balanced recall and precision.

Feature Importance

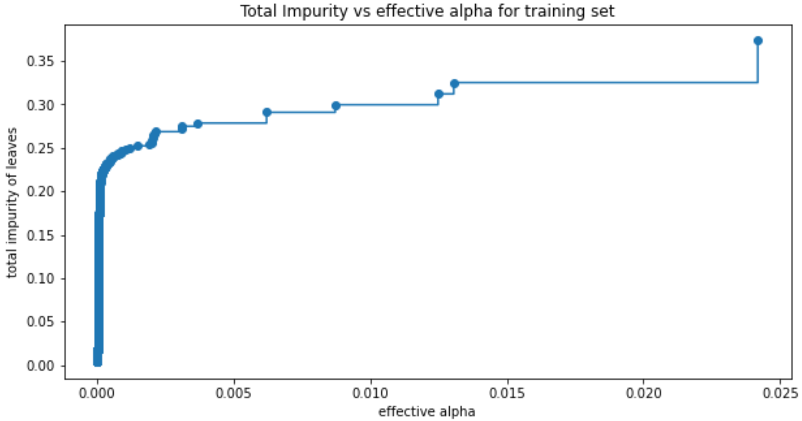


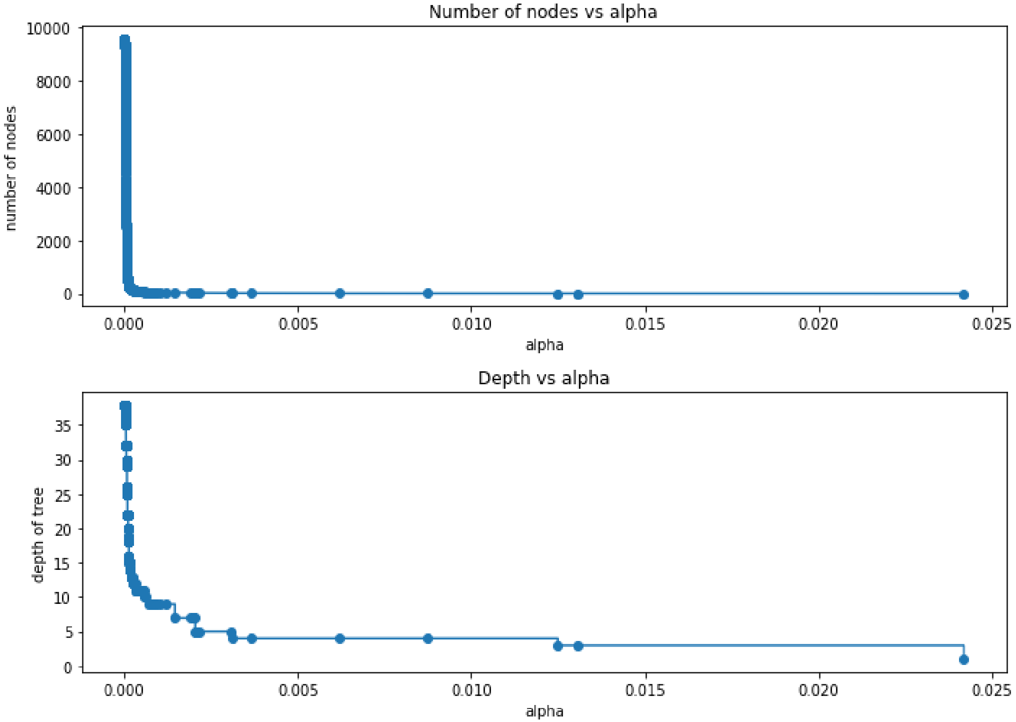


Visualizing the Decision Tree



Impurity vs effective alpha for training set





Observation

- We do not need to prune this tree.

Conclusion

- I analyzed "Star Hotels Group" and using different techniques and used Decision Tree Classifier to build a predictive model for the same
- The model can be used to predict if a hotel room booking is going to be canceled or not
- We visualized different trees and their confusion matrix to get a better understanding of the model.
- Lead_time, avg_price_per_room, and arrival month are the most important variable in predicting the customers that will cancel their reservation.
- We established the importance of hyper-parameters / pruning to reduce overfitting

Recommendations

- According to the decision tree model if a customer books their reservation in less than 58.49 days then the customer will not be cancelling their reservation.
- It is observed that booking status has a negative correlation with the market segment type as complementary
- It is also observed that when the price of a hotel room is lower, the booking is less likely to be canceled. Room Type 7, 6 and 5 have strong correlation to booking. Those who book those rooms are less likely to cancel. They are more expensive. If this is a destination hotel, people may be booking these for vacation and leisure.
- It is also observed that the higher number of special requests means that the customer will not cancel
- Most bookings occur during the months of June, July, and August. Summer months are the most popular to travel so you can assume that these will be less likely to be canceled.
- Weeknights have high correlation with booking and less likely to be canceled than weekend nights.
- When the reservation is made online, the booking is less likely to get canceled. This makes sense as most users have shifted to making reservations where it is online and user-friendly. People are less likely to cancel their reservation if the hotel has good reviews as well. So being able to see positive reviews from past guests is a plus on the website.

greatlearning
Power Ahead

Happy Learning !