Hotel Inn Data Analysis

Python Foundations : PGP-DSBA

March 16,2023

Contents / Agenda

- Business Problem
- Objective
- Data Overview
- EDA Results
- Data Preprocessing – Outlier check
- Model Performance Summary
- Decision Tree and Comparing Decision Tree models
- Conclusion and recommendations

Business Problem

- A significant number of hotel bookings are called off due to cancellation or no-shows. This has been made easier due to low cost / no charge cancellation option which benefit hotel guest but are less desirable and have negative impact on the hotel and have the potential of diminishing hotel revenue.
- The cancellation of bookings have a significant impact on the hotel:
    - Lost of revenue when the hotel cannot resell the room.
    - Additional cost 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 arrangement for the guests.

Objective

- Machine Learning based solution can help in predicting which booking system is likely to cancel.
- The INN Hotel Group have reached out our firm for a data driven solution.
    - Analyze the provided data to find which factors have a high influence on booking cancellations.
    - Build a predictive model that can predict which bookings are going to be cancelled.
    - Help in formulating profitable policies for cancellations and refunds.

Data Overview

```
In [15]: data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36275 entries, 0 to 36274
Data columns (total 19 columns):
 #   Column                                Non-Null Count  Dtype  
---  ------                                --------------  -----  
 0   Booking_ID                            36275 non-null  object 
 1   no_of_adults                          36275 non-null  int64  
 2   no_of_children                        36275 non-null  int64  
 3   no_of_weekend_nights                  36275 non-null  int64  
 4   no_of_week_nights                     36275 non-null  int64  
 5   type_of_meal_plan                     36275 non-null  object 
 6   required_car_parking_space            36275 non-null  int64  
 7   room_type_reserved                    36275 non-null  object 
 8   lead_time                             36275 non-null  int64  
 9   arrival_year                          36275 non-null  int64  
 10  arrival_month                         36275 non-null  int64  
 11  arrival_date                          36275 non-null  int64  
 12  market_segment_type                   36275 non-null  object 
 13  repeated_guest                        36275 non-null  int64  
 14  no_of_previous_cancellations          36275 non-null  int64  
 15  no_of_previous_bookings_not_canceled  36275 non-null  int64  
 16  avg_price_per_room                    36275 non-null  float64
 17  no_of_special_requests                36275 non-null  int64  
 18  booking_status                        36275 non-null  object 
dtypes: float64(1), int64(13), object(5)
memory usage: 5.3+ MB
```

Observations

- There are object, integer, and float data types.
- There are no missing values

Data Overview

```
In [14]: data.shape ##
Out[14]: (36275, 19)
```

```
In [16]: # checking for duplicate
         data.duplicated().sum()
Out[16]: 0
```

Observations

- There are 36,725 rows and 19 columns
- There are no duplicate entries or Data

EDA Results (Statistical Summary)

```
In [19]: data.describe()##  Complete the code to print the statistical summary of the data
```

Out[19]:

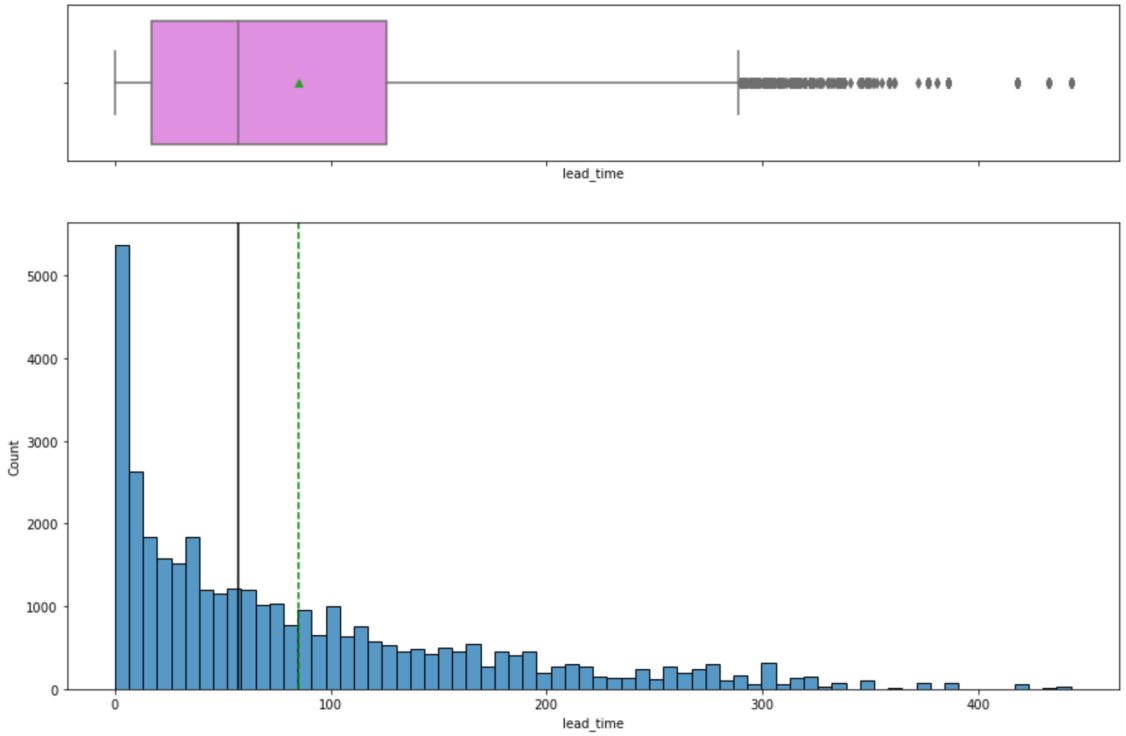
	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	required_car_parking_space	lead_time	arrival_year	arrival_month	arrival_d
count	36275.00000	36275.00000	36275.00000	36275.00000	36275.00000	36275.00000	36275.00000	36275.00000	36275.00(
mean	1.84496	0.10528	0.81072	2.20430	0.03099	85.23256	2017.82043	7.42365	15.597
std	0.51871	0.40265	0.87064	1.41090	0.17328	85.93082	0.38384	3.06989	8.74(
min	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	2017.00000	1.00000	1.00(
25%	2.00000	0.00000	0.00000	1.00000	0.00000	17.00000	2018.00000	5.00000	8.00(
50%	2.00000	0.00000	1.00000	2.00000	0.00000	57.00000	2018.00000	8.00000	16.00(
75%	2.00000	0.00000	2.00000	3.00000	0.00000	126.00000	2018.00000	10.00000	23.00(
max	4.00000	10.00000	7.00000	17.00000	1.00000	443.00000	2018.00000	12.00000	31.00(

Observations

- Data consist of only arrival year 2017 and 2018.
- The mean number of adults is 1.85, min is 2 and max is 4
- The mean number of children is 0.15, min is 0 and max is 10
- The mean number of weekend nights is 0.81, min is 0 and max is 7
- The mean number of weeknights is 2.24, min is 0 and max is 17

EDA Results (Univariable Analysis)

```
In [21]: histogram_boxplot(data, "lead_time")
```

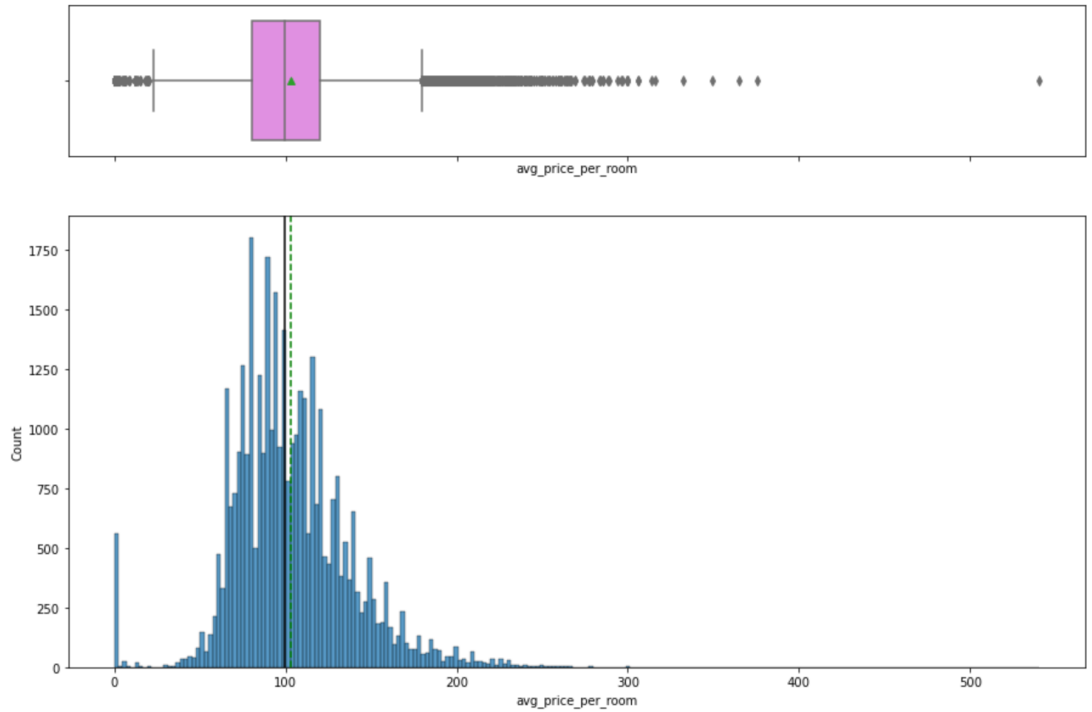


Observations

- There are outliers in this data more to the right indicating that less visibility in bookings some days.
- Both the mean and median are quite high so on most day we have good visibility on bookings.

EDA Results (Univariable Analysis)

```
In [22]: histogram_boxplot(data,"avg_price_per_room")  ## Complete the code to create histogram_boxplot for average p
```

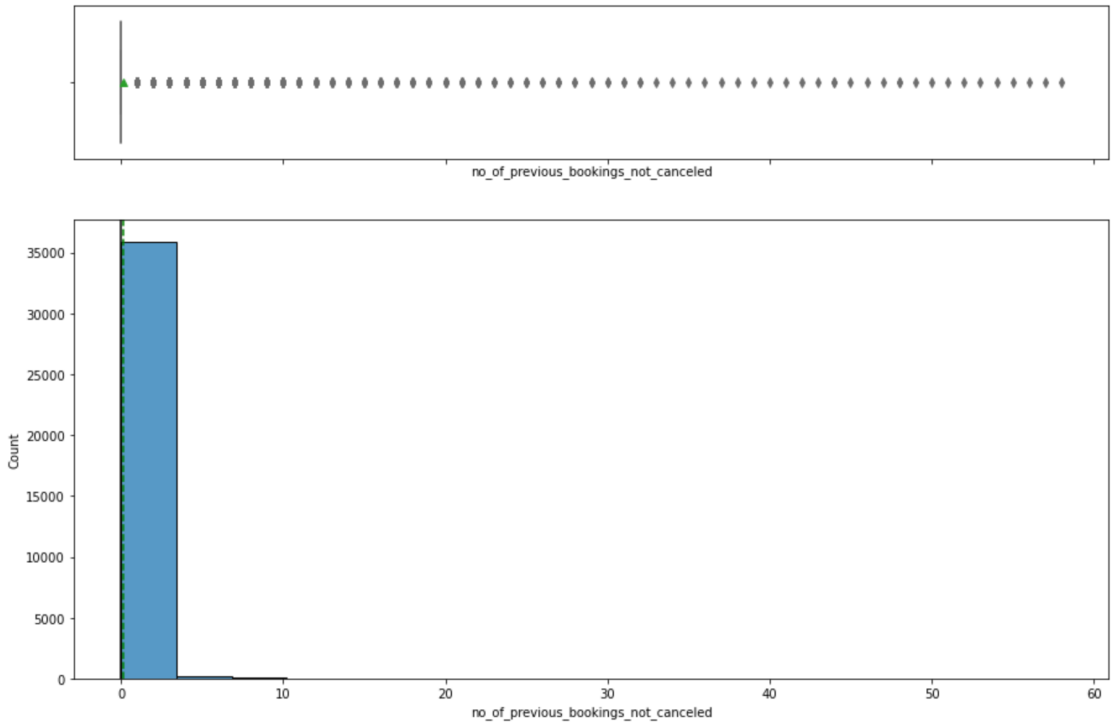


Observations

- There are outliers in this data more to the right.
- We observe that the price per room gradually decreases on the right (higher price
point).
- The mean and median are almost at the same point. ~ $100 per night.

EDA Results (Univariable Analysis)

```
In [28]: histogram_boxplot(data,'no_of_previous_bookings_not_canceled')  ## Complete the code to create histogram_box
```



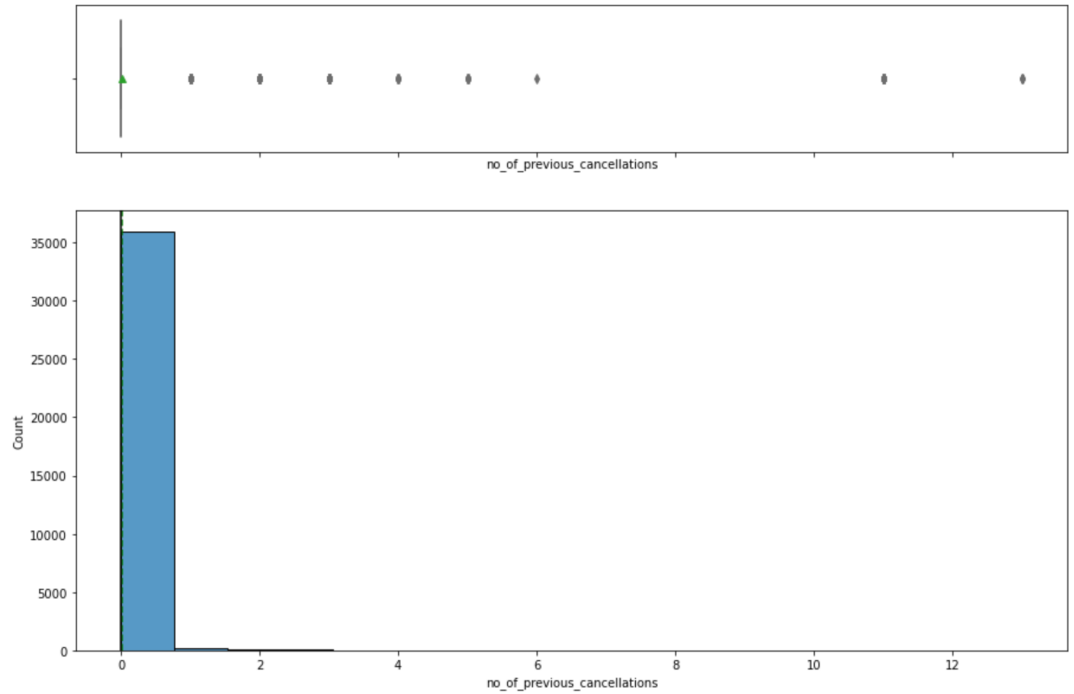
Observations

- This boxplot shows Outliers which means there is an impact.
- Both the mean and median are quite high so on most day we have good visibility on bookings.

EDA Results (Univariable Analysis)

Observations on number of previous booking cancellations

```
In [27]: histogram_boxplot(data,'no_of_previous_cancellations')  ## Complete the code to create histogram_boxplot for
```



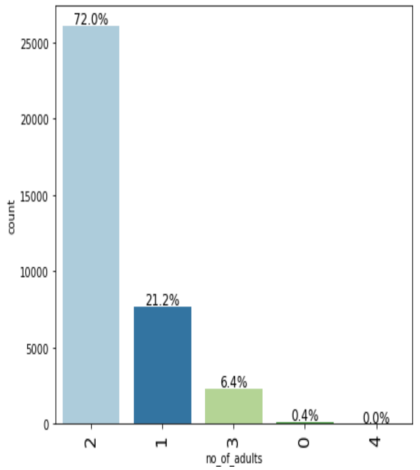
Observations

- This boxplot shows Outliers which means there is an impact.
- There is no mean and median shown, indicating the high impact cancellation have .

EDA Results (Univariable Analysis)

Observations on number of adults

```
labeled_barplot(data, "no_of_adults", perc=True)
```

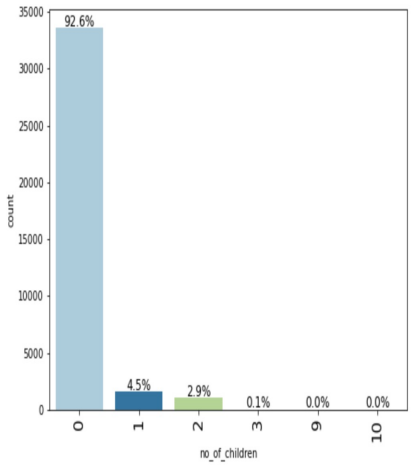


Observations

The large group of booking are for two adults. They likely have a large impact on booking retention.

Observations on number of children

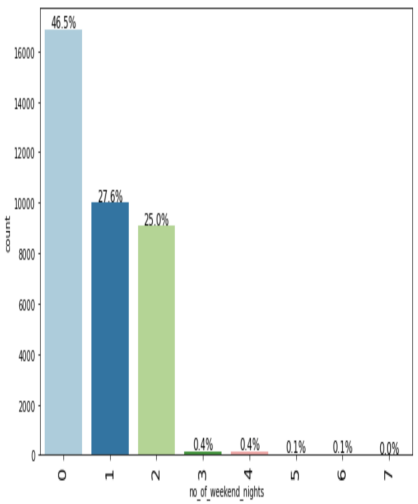
```
labeled_barplot(data,'no_of_children',perc=True)  ## Complete
```



Observations

Most bookings do not include children therefore there have net to no impact on booking retention.

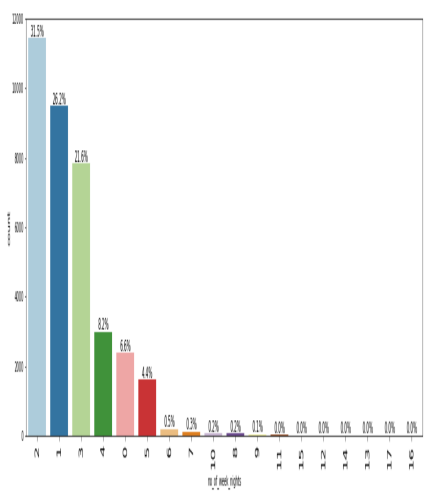
```
labeled_barplot(data,'no_of_weekend_nights',perc=True)  ## Complete the co
```



Observations

Nearly 50% of bookings do not include a weekend nights and less then 1% include a long weekend.

```
labeled_barplot(data,'no_of_week_nights',perc=True)  ## Complete the code to create labeled_barplot for number of we
```



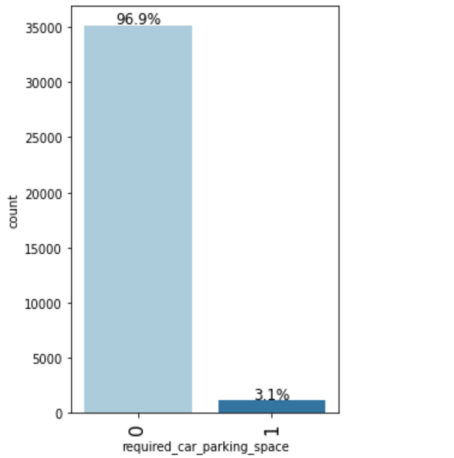
Observations

Most booking during week nights are for 2 days.

EDA Results (Univariable Analysis)

Observations on required car parking space

```
labeled_barplot(data,'required_car_parking_sp
```

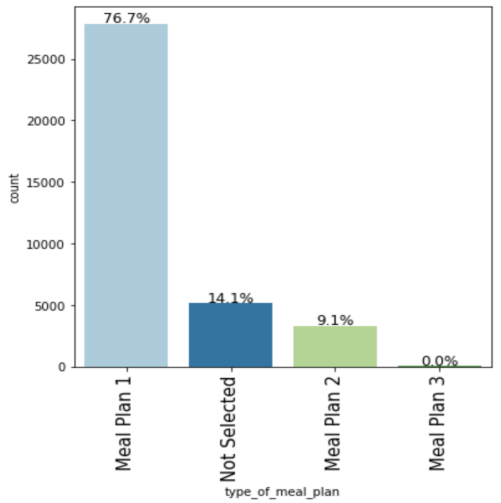


Observations

Most bookings required a space to park their car. Likely an important factor in there booking.

Observations on type of meal plan

```
labeled_barplot(data,'type_of_meal_plan',perc=Tr
```

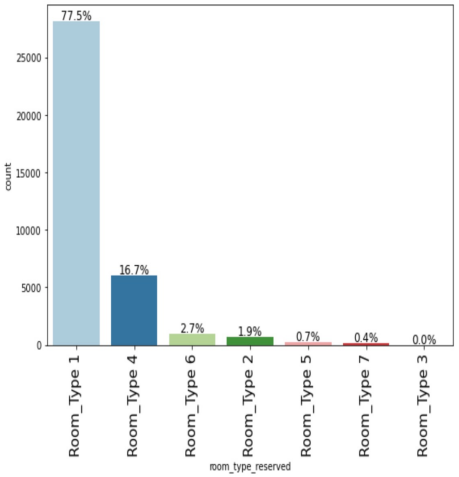


Observations

- Most bookings included a breakfast options. Guess which make other meal plan option likely have little impact on booking retention.
- This could have a significant impact on cost if meals need to be planned out for.

Observations on room type reserved

```
labeled_barplot(data,'room_type_reserved',perc=True)  ## Complete th
```

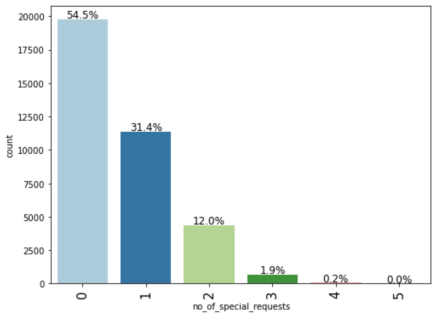


Observations

- Most bookings are for Room_Type 1. They will likely have the greater impact on booking retention.

Observations on number of special requests

```
labeled_barplot(data,'no_of_special_requests',perc=True)  ##
```



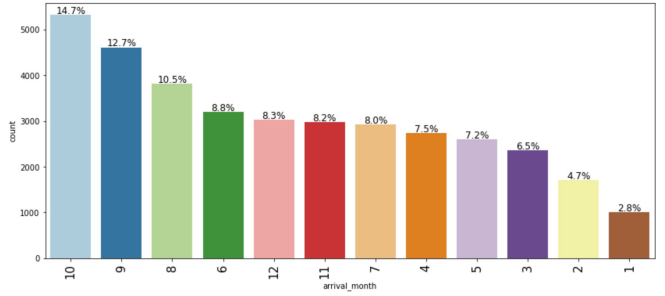
Observations

- Most booking have NO special request. With approximately 86% of booking required 1 or lest special request. Special request likely have any impact on booking retention.

EDA Results (Univariable Analysis)

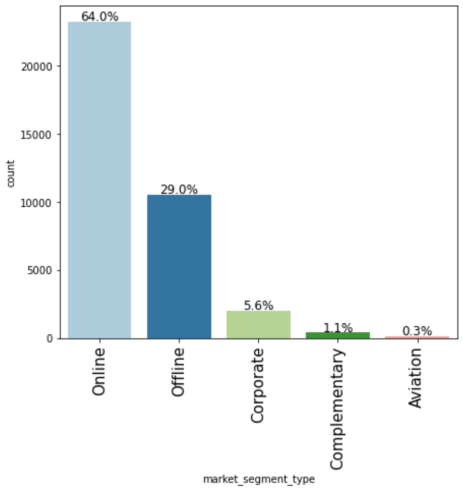
Observations on arrival month

```
labeled_barplot(data,'arrival_month',perc=True)  ## Complete the code to create labeled_barplot for a
```



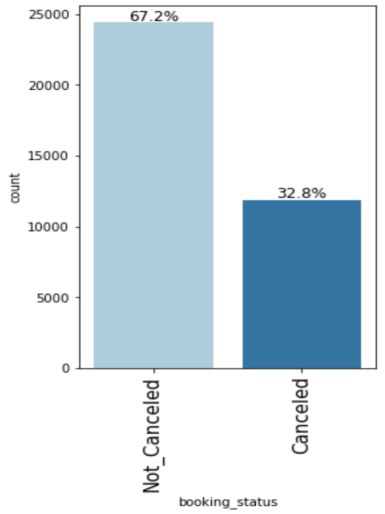
Observations on market segment type

```
: labeled_barplot(data,'market_segment_type',perc=True)
```



Observations on booking status

```
labeled_barplot(data,'booking_statu
```



Observations

Last summer and fall are the months when the most bookings are received and therefore will likely have a higher impact on booking retention.

Observations

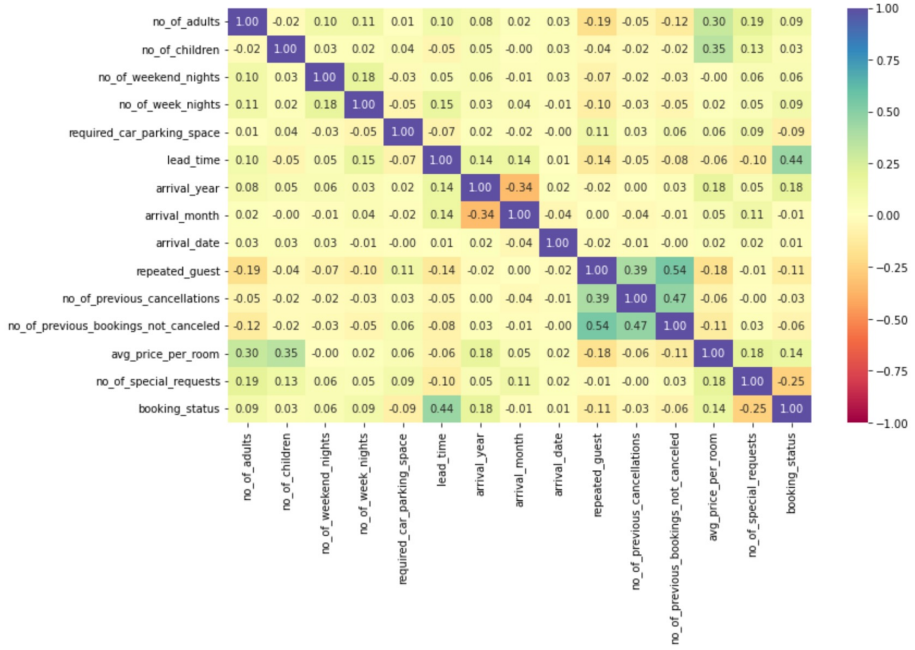
Most bookings are received online, followed by offline and the corporate. Online bookings are likely to have the higher impact on booking retention due to the favorable cancellation policy online booking sites have.

Observations

- 32.8% of bookings are cancelled and 67.2% are retained.

- Focus is one creating policies that will help with retaining a high portion of bookings.

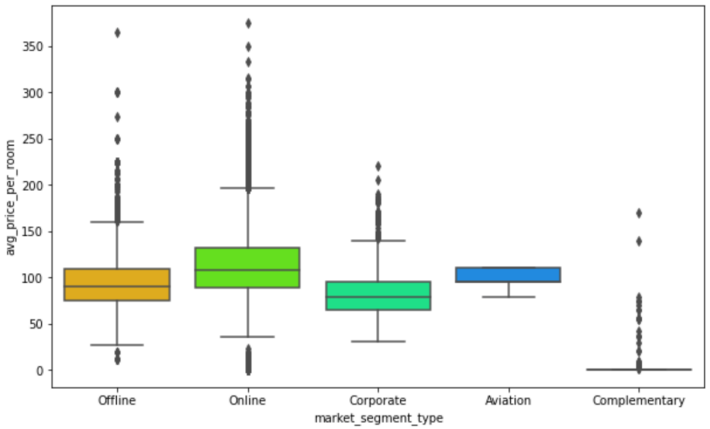
EDA Results (Bivariate Analysis)

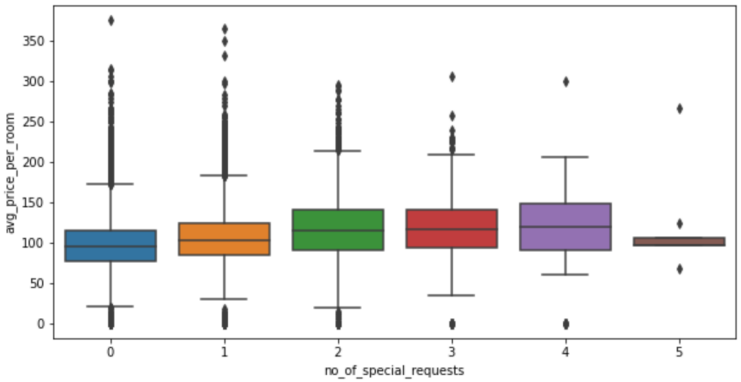


Observations

- There is limited correlation between most of the criteria in the data set.
- The highest correlation is between “# of previous booking not cancelled” and “repeated guest”.
- The next highest is between “# of previous booking not cancelled” and “# of previous cancellation”
- Lead Time and Booking status have a noticeable correlation that should be noted. As well as # of adults and avg price per room.

EDA Results (Bivariate Analysis)





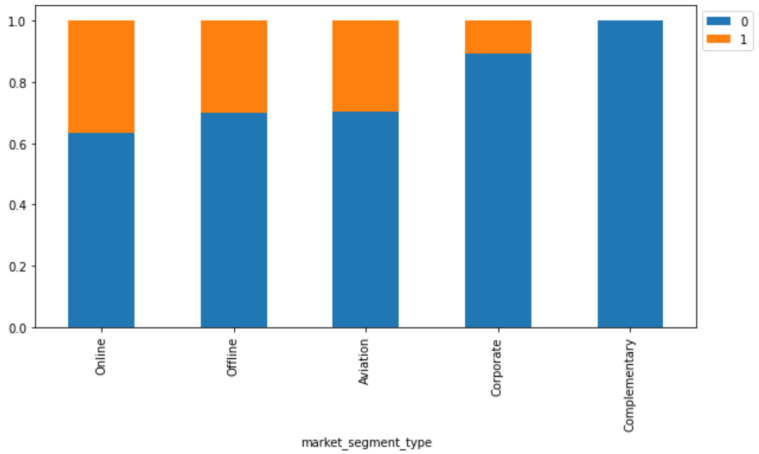
Observations

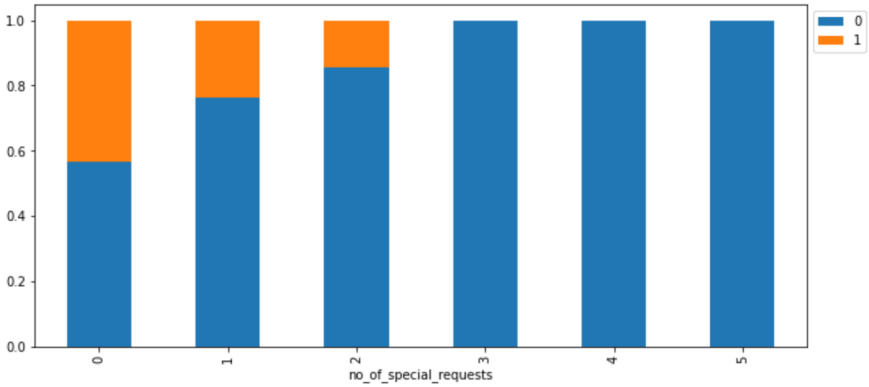
- Hotel room rate vary by market segment. As you can see the “Online” channel bring in the highest “Avg Price per Room”, followed by Offline, Corporate and Aviation.
- There are also significant outliers by market segments with the exception of Aviation.

Observations

- There are less outlier in the Average price per Room as the number of Specialist request increases.
- Also, as the number of Specialist request increase the Average price per Room increases.

EDA Results (Bivariate Analysis)





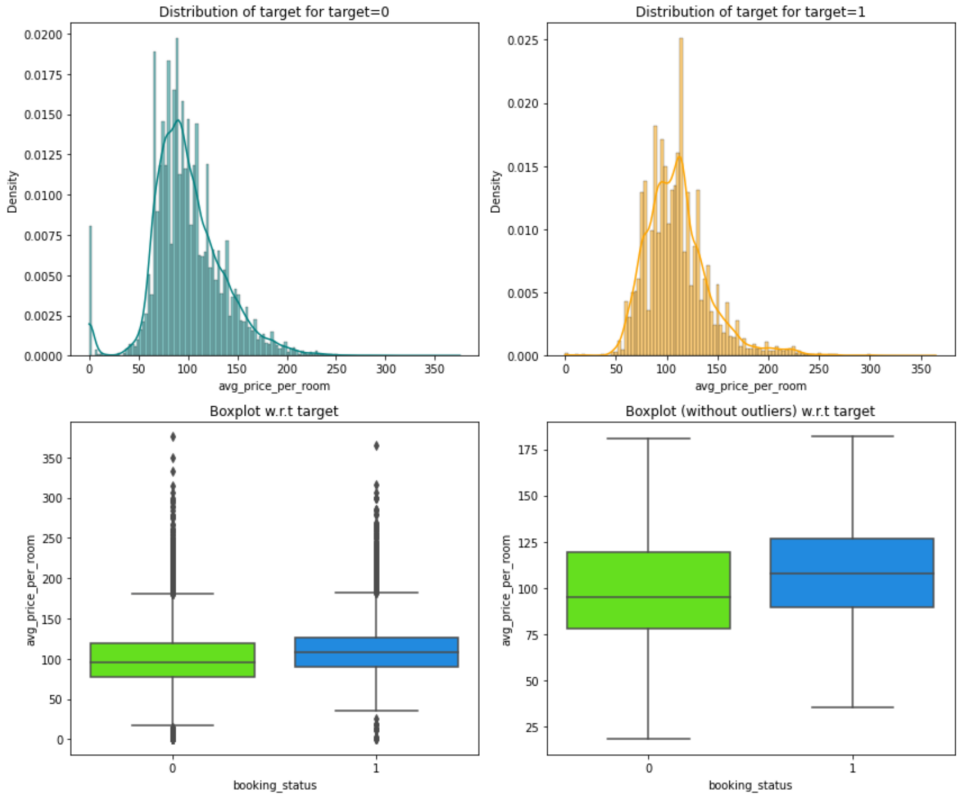
Observations

- Each Market Segment is impacted by booking status except for “Complementary” which makes sense since they are NOT paying for the room.
- Online room are the most negatively impacted following by Offline and Aviation. The least impacted or Corporate bookings.

Observations

- Bookings with the least number of special requests are negatively impacted. A guest with a largest number of specialist request is least likely to cancel there booking.

EDA Results (Bivariable Analysis)

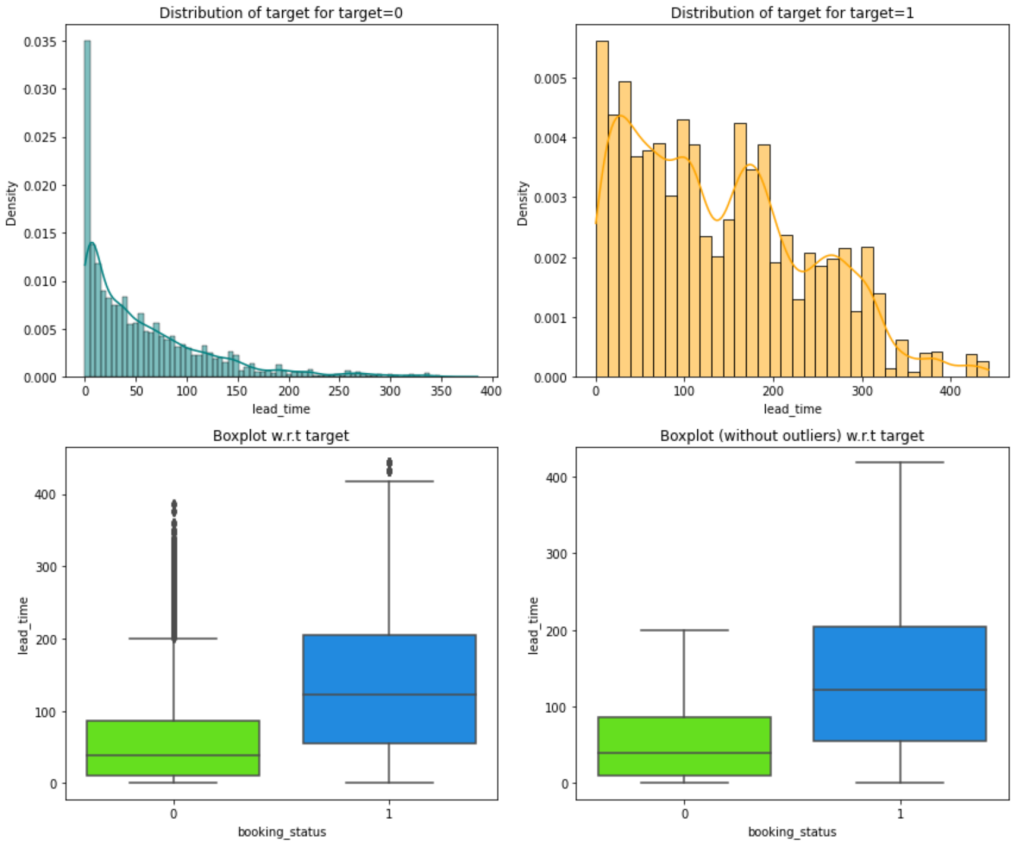


Observations

- There is positive impact between booking status and average price per room.

- Bookings that have been cancelled have a higher average price per room.

EDA Results (Bivariate Analysis)



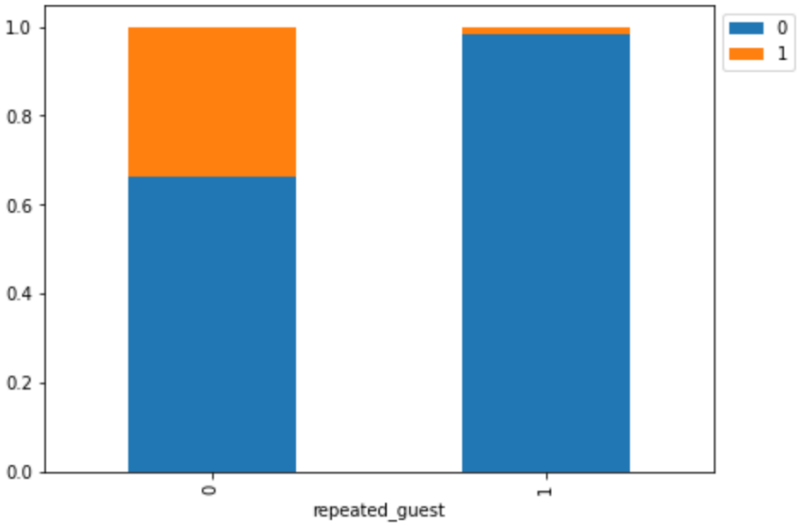
Observations

- There is positive impact between booking status and Lead time.

- Bookings that are cancelled then to give you a longer lead time. This should allow enough time to develop an algorithm to determine the likely hood of cancelled.

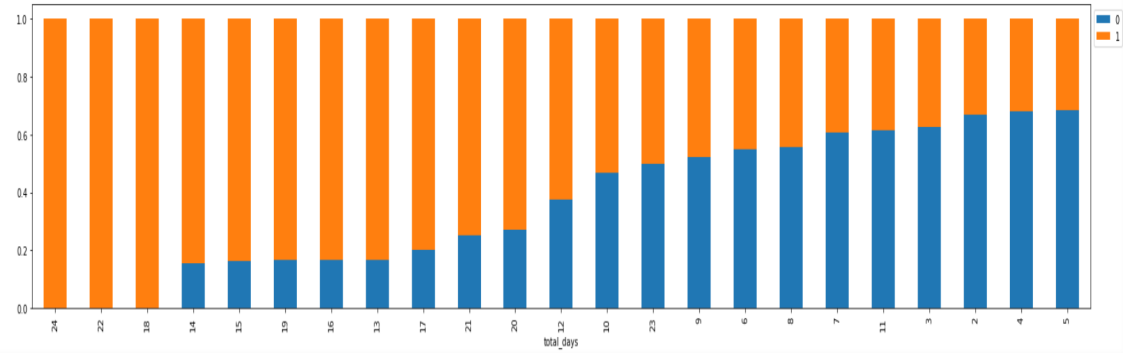
- Guest with bookings with shorter lead times tend not to cancel.

EDA Results (Bivariate Analysis)



Observations

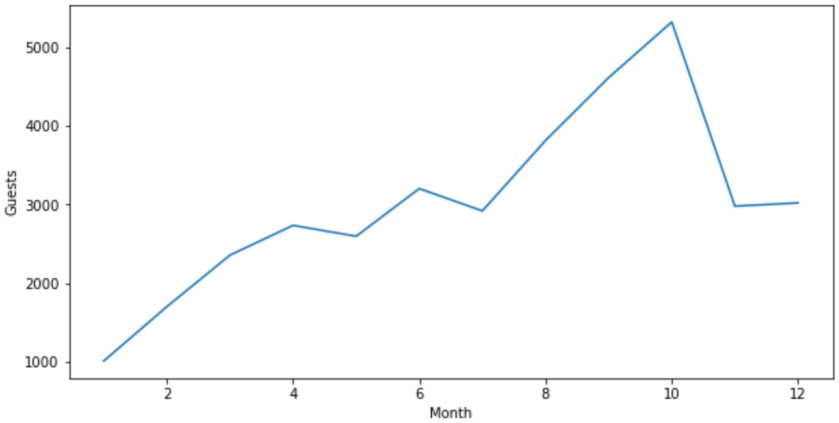
- Repeat guest tend to cancel their booking less often.

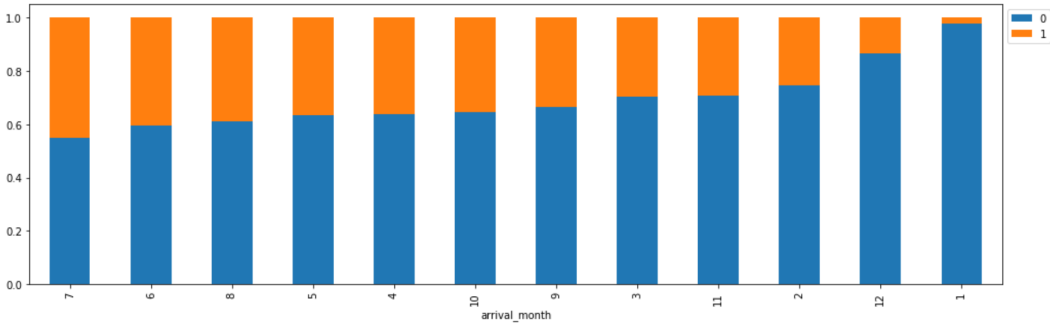


Observations

- Bookings that booked farther in advance have a higher probability of being cancelled than booked more recently.

EDA Results (Bivariate Analysis)





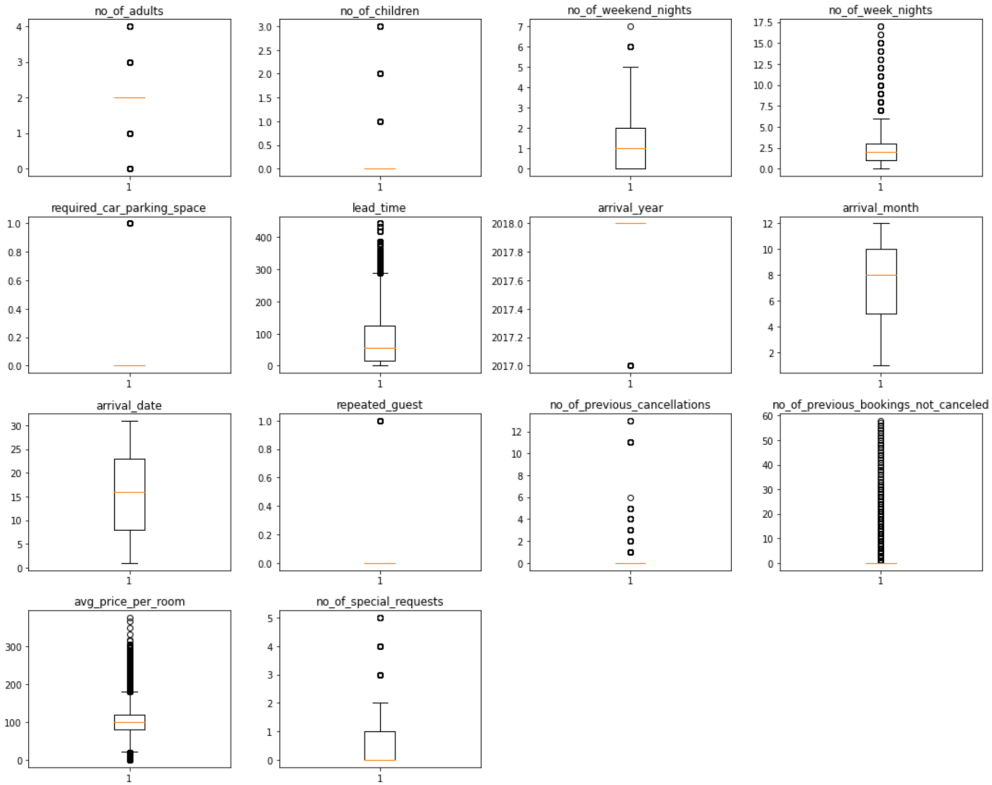
Observations

- Guest numbers peek in the 10th month (October).

Observations

- In summer months people tend to cancelling book as we can see summer months have most booking cancellation.

Data Preprocessing – Outlier check



Observations

No of Weeknights, Lead Time, # of previous cancellation, # of previous booking not cancelled and avg price per room have the most outliers.

Model Performance Summary

Default Threshold:

Out[68]:

	Accuracy	Recall	Precision	F1
0	0.80600	0.63410	0.73971	0.68285

```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.425090
         Iterations: 35
                         Logit Regression Results
==============================================================================
Dep. Variable:         booking_status   No. Observations:                25392
Model:                          Logit   Df Residuals:                    25364
Method:                           MLE   Df Model:                           27
Date:                Wed, 15 Mar 2023   Pseudo R-squ.:                 0.3292
Time:                        18:57:50   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   3987.836      0.004      0.997   -7798.656    7833.373
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
================================================================================================
```

Observations

- Negative value in the coefficient shows the probability of the guest book cancellation decrease with the increase of corresponding attribute value.
- Positive value in the coefficient shows the probability of the guest book cancellation increases with the increase of corresponding attribute value.
- The p-value of a variable indicates if the variable is significant or not. If we consider the significant level to be 0.05%, then any variable with a p-value less then 0.05 would be consider 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.

Model Performance Summary

Training performance :

Out[93]:

	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

Testing performance :

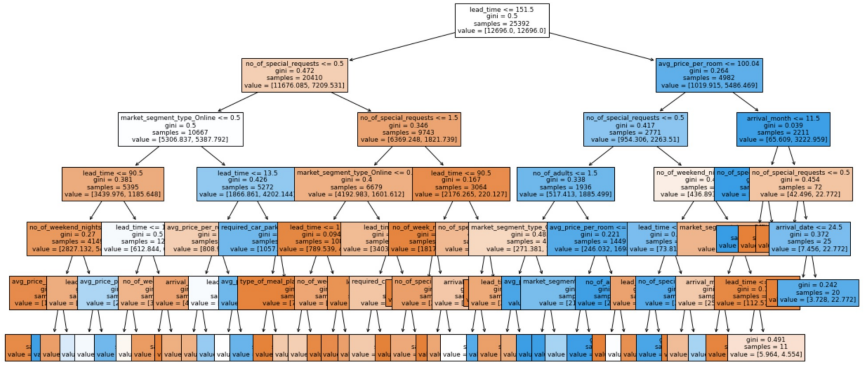
Out[94]:

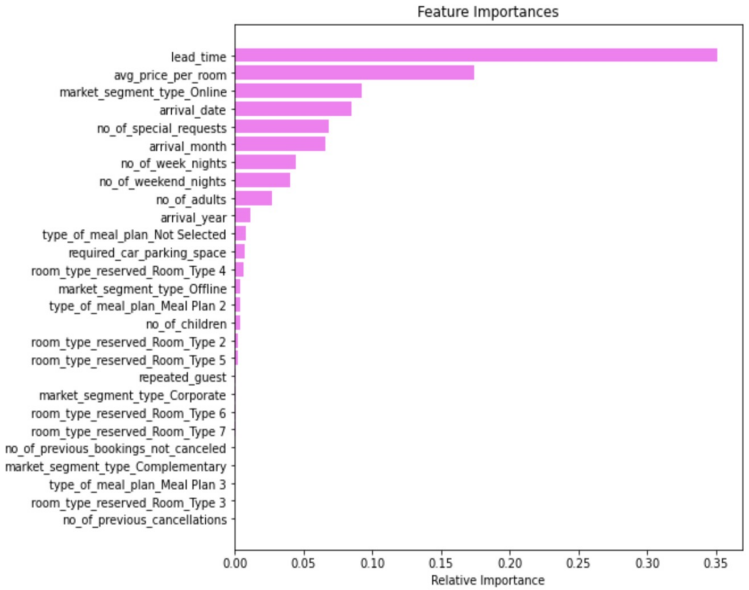
	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80465	0.79555	0.80345
Recall	0.63089	0.73964	0.70358
Precision	0.72900	0.66573	0.69353
F1	0.67641	0.70074	0.69852

Observations

- We built a predictive model that can be used by the Hotel to find out which bookings will be cancelled with a F1 score of 0.68 on the training set and formulate policies accordingly.
- All the logic regression model have given a generalized performance on the train and test set.
- The coefficient of some attributes will increase the chances of a booking being cancelled.

Decision Tree





Observations

- Using the above extracted decision rules, we can make interpretation from the decision tree model.
- In tuned decision tree “Lead Time” is the most important feature following by Market Segment Type Online and # of Special Request.

Comparing Decision Tree Models

Training performance :

	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 performance :

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

Observations

- 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.

Conclusion and Recommendation

Conclusion:

- We analyzed the INN Hotel booking data using a variety of techniques and built a predictive model using the Decision Tree.
- The model can be used to predict whether a guest will result in a cancellation, resulting in lost revenue and additional costs for the hotel.
- We determined that the Lead Time, Market Segment type online, and the number of Special Requests are the most important variables in predicting whether or not a guest will cancel the booking.
- We've established the significance of pruning in reducing overfitting.

Recommendation:

According to Decision tree model,

1. we should avoid bookings with lead times greater than 151 because they are more likely to be canceled.
2. If the booking has a lead time of less than 151 days, we should concentrate on how many special requests the guest has. The higher the number, the less likely they are to cancel their reservation.
3. Finally, we should focus on online bookings because they have the second highest level of cancellation.

Great Learning

Happy Learning !