Business Presentation

Contents

- Business Problem Overview and Solution Approach
- Data Overview
- EDA
- Model Performance Summary
- Model Performance Comparison and Conclusions
- Business Insights and Recommendations

Business Problem Overview and Solution Approach

Core business idea

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. Idea is to reduce room cancel thereby improve the overall sales.

Problem to tackle

- Identify deficiencies in current target segmentation.
- Identifying potential cancelations

Financial implications

- Last minute cancellations that impact the hotel sales and resources

How to use ML model to solve the problem

- Logistic Regression and Decision Trees have been used to predict the model that can be implemented to predict the bookings that are likely to be canceled.

Data Overview

- The data contains information about 56926 customer bookings of Star Hotels and their characteristics

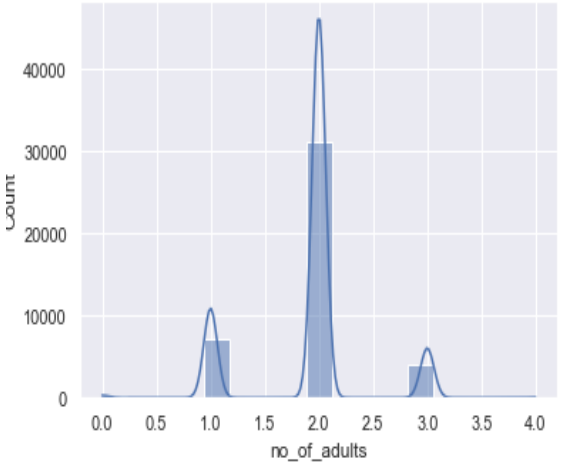
- The characteristics include no_of_children, number_of_adults, no_of_weekend_nights, no_of_week_nights, type_of_meal_plan, required_car_parking_space and many more. There are 18 such characteristics.

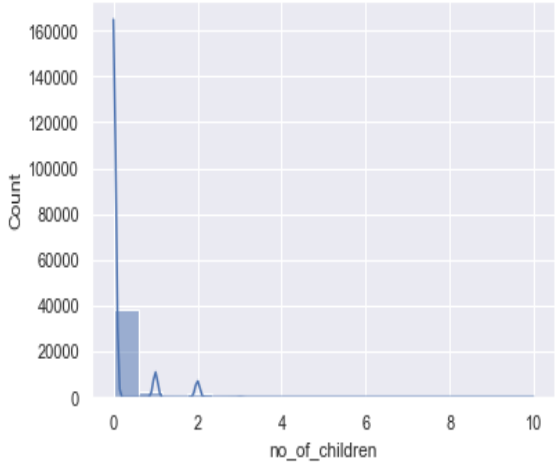
- Some rows are duplicates. Hence, the duplicate rows have been removed resulting in 42576 rows.

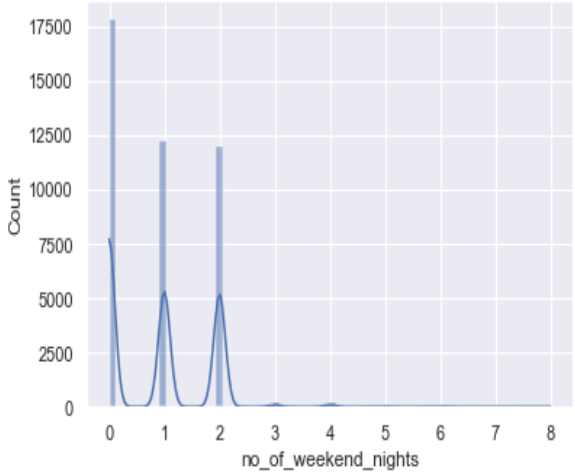
- Booking_status has been changed from object to numerical variable.
- object variables like type_of_meal_plan has been changed to categorical.
- There are no missing values.

- Factors like Lead time, average price per room, no of special requests, market segment type and many more can affect the cancellations.

EDA – no_of_adults, no_of_children, no_of_weekend_nights



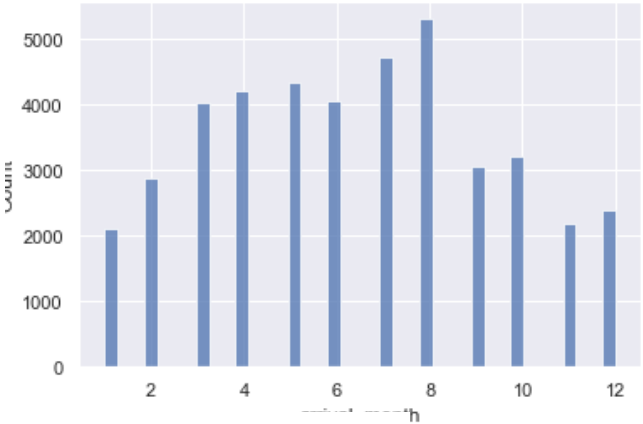


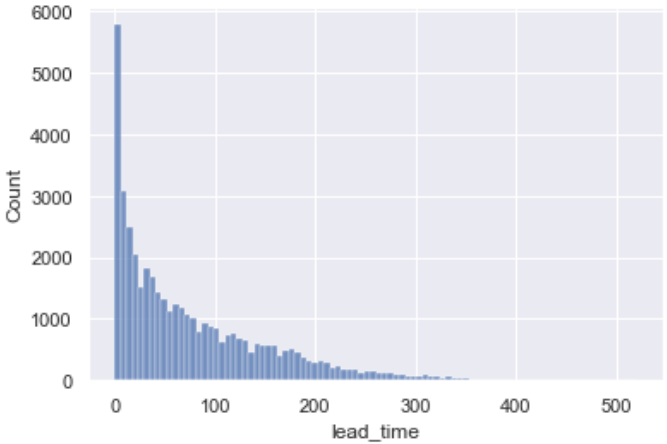


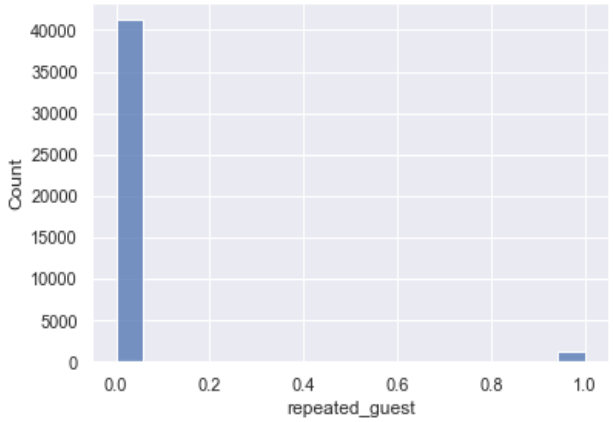
- Number of adults is 2 in most of the cases.
- In most of the cases Number of children = 0.The distribution is positively skewed. There are 5 outliers.
- Most of the booking are made for 1-2 weekend nights. There are 3 outliers.

EDA - lead_time, arrival_month, repeated_guest

- Lead_time → The distribution is positively skewed.
  - There are many outliers.
- Most of the booking are in August month.
  - The least number of bookings are in January and November
- Most of the customers are repeated guest.

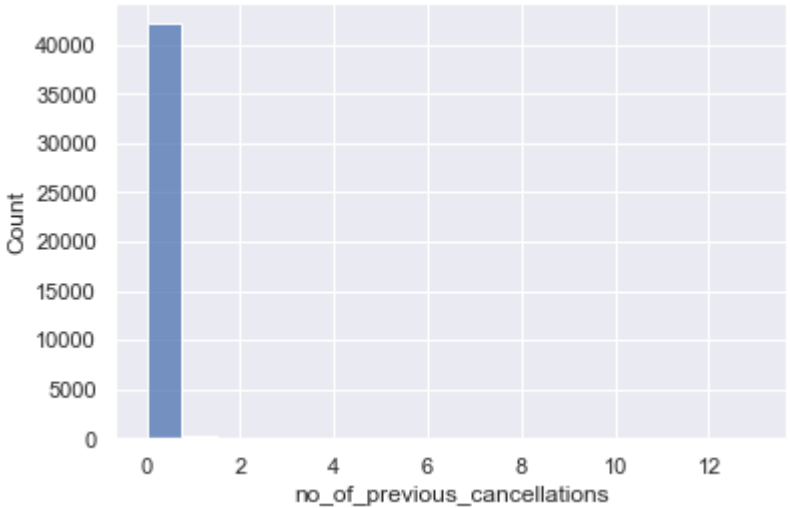


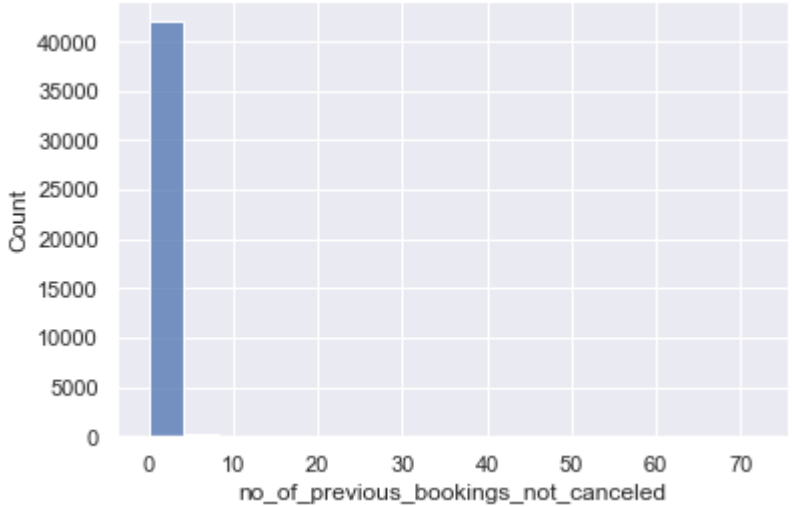




EDA - no_of_previous_bookings_not_cancelled, no_of_previous_cancellations

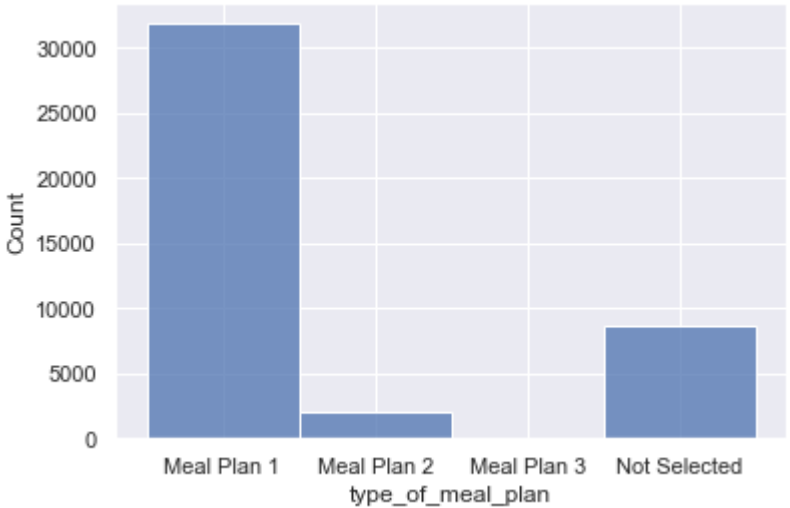
- Most of the customers have zero previous cancelations.
- Number of previous bookings not canceled is 0 in most of the cases

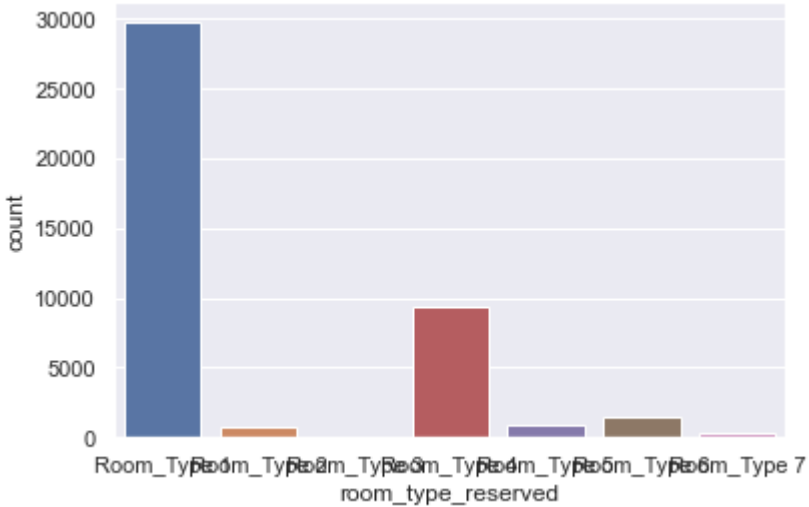




EDA – Categorical Variables --Meal_plan and room_type

1. Maximum number of the customers prefer Meal Plan 1. There are some customers who dont have any preference.
2. Most of the customers prefer Room Type 1. Least preferred room is Room Type 3 followed by Room Type 7.

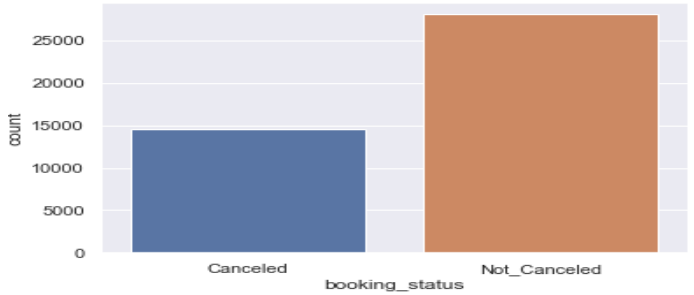




EDA–market segmentation type and booking status

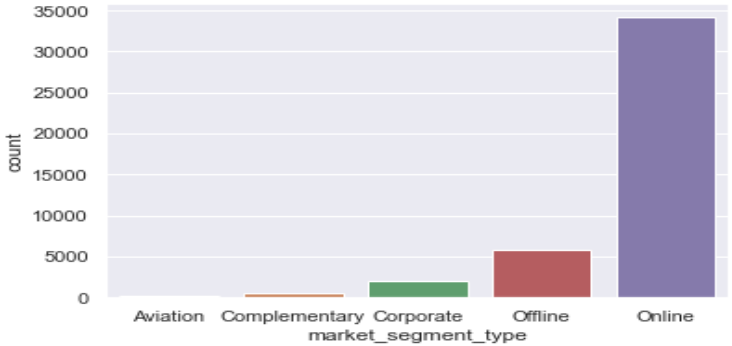
1. Most of the customers prefer 'online' market Segmentation type. Least preferred is Aviation.

- Most of the customers have not canceled their booking.



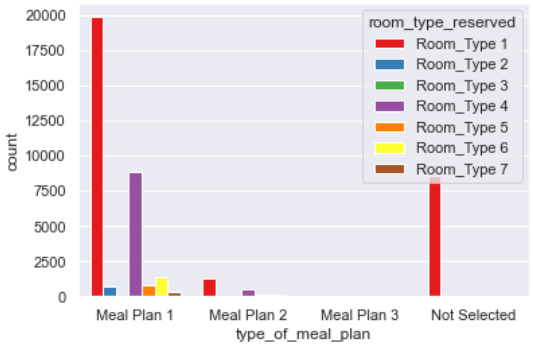
BiVariate Analysis – (Heat_map)

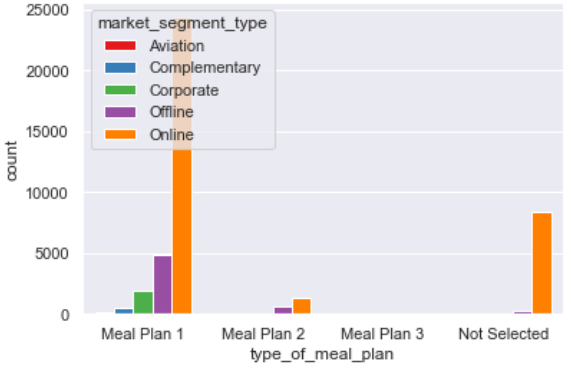
1. Number of previous cancelations show high correlation with number of previous bookings not canceled (0.58)
2. Repeated guest show high correlation with number of previous bookings not canceled(0.56)
3. It is important to note that correlation does not imply causation.
4. There are some negatively correlated.

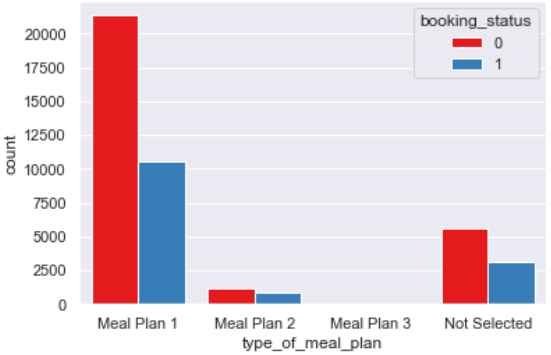


BiVariate Analysis(Heat_Map)

EDA – meal_plan vs room_type, market_segment and booking_status

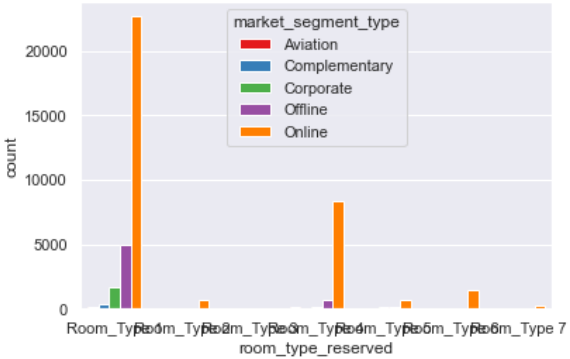


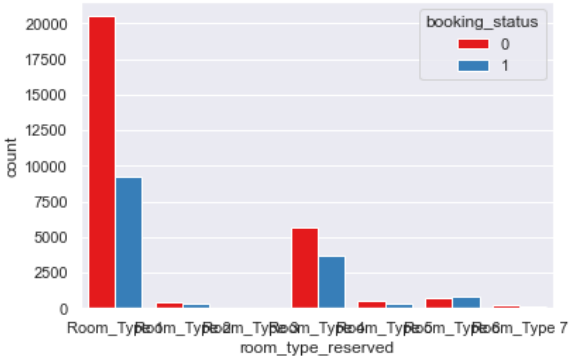




1. Room type 1 is mostly preferred followed by Room type 4 . Customers who prefer Meal Plan 1 mostly use Room type 1.
2. Most of the customers prefer Meal Plan 1. Mean plan 1 customers prefer online market segment type. Customers who have not selected any meal plan prefer online market segment type.
3. Customers who prefer Meal Plan 1 have more cancellations than other meal plan groups.

EDA - room_type, market_segment and booking_status

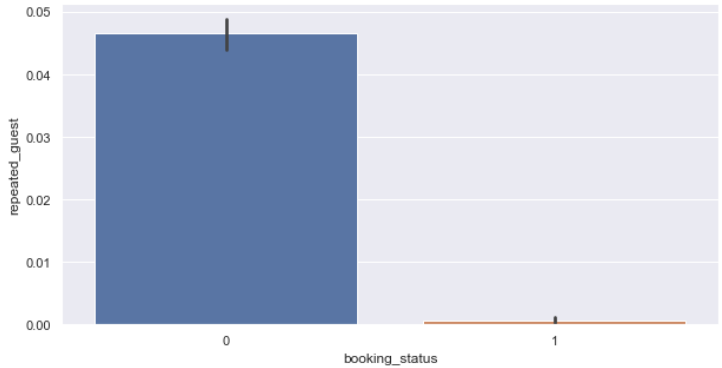


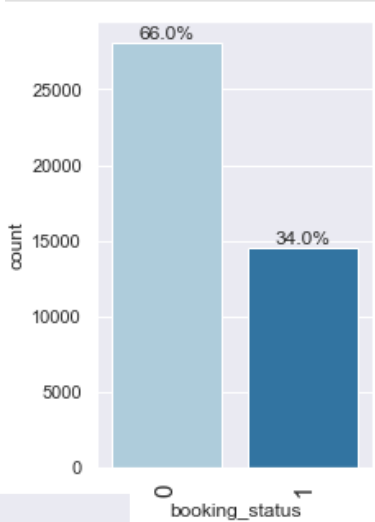


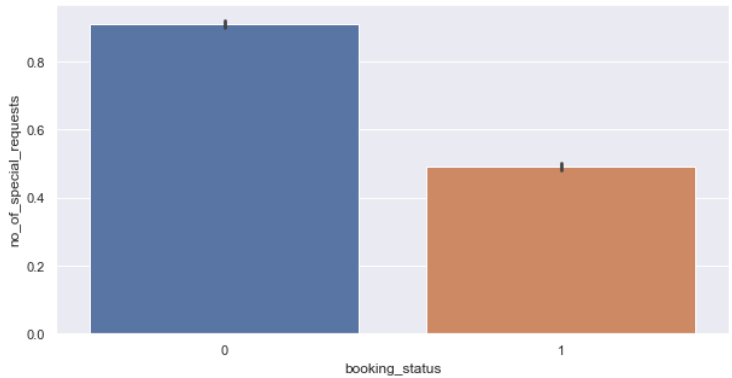
1. Customers prefer Online market segment type. Most of the customer choose Room Type 1 and online market segment type.
2. It is observed that customers who prefer Online market segmentation had more cancellations than others. There are more customers with booking status is not canceled.
3. Customers who prefer Room type 1 had more cancellations than other.

EDA (Contd)

1. 34% of the bookings are cancelled.
2. The rest of the 66% of the bookings are not canceled.
3. More that 97% of the repeating guests donot cancel the bookings.

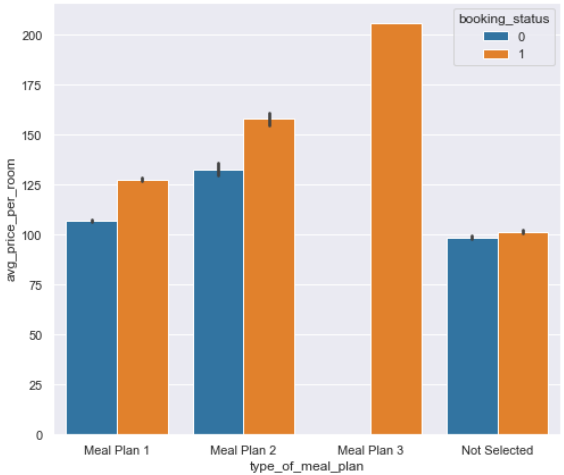


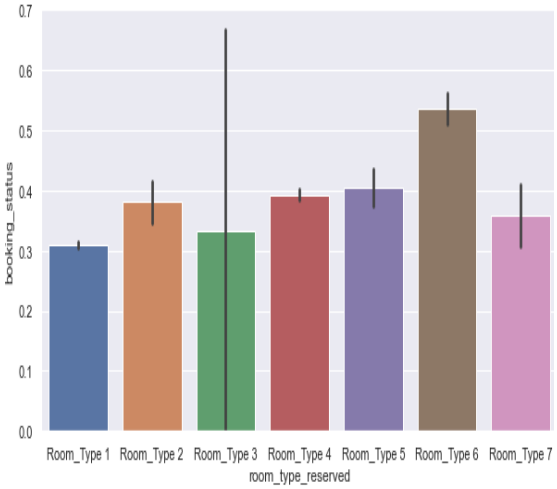


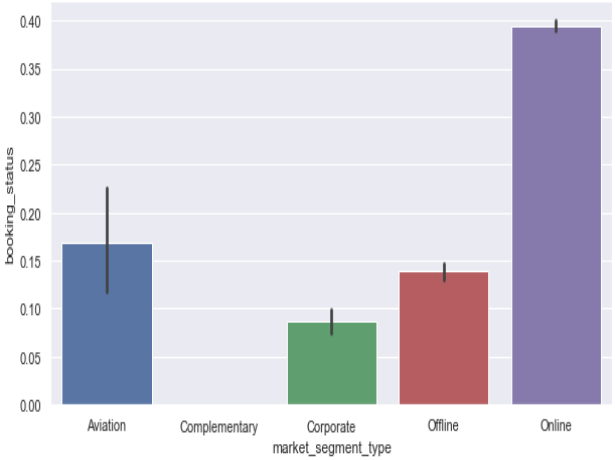


EDA(Contd)

1. Average price of the room if the customer chooses Meal Plan 3 and booking status is cancelled is most expensive. The customer whose booking status is not cancelled and if he chooses Meal Plan 1, the average price of the room will be lesser.
2. Max number of cancelations are seen when Room Type 6 is chosen.
3. Max number of cancelations are seen form market segment type is Online and least for market segment type is Complementary.

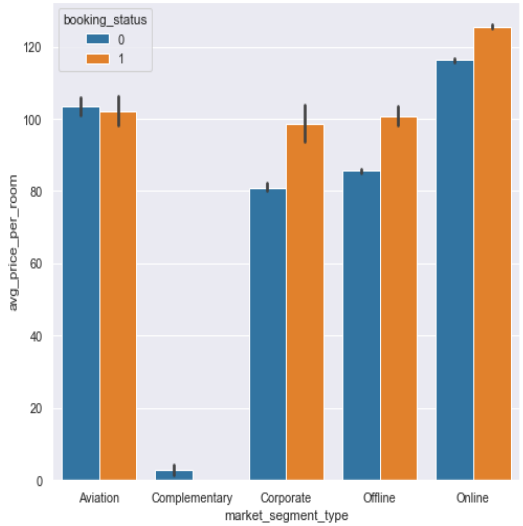


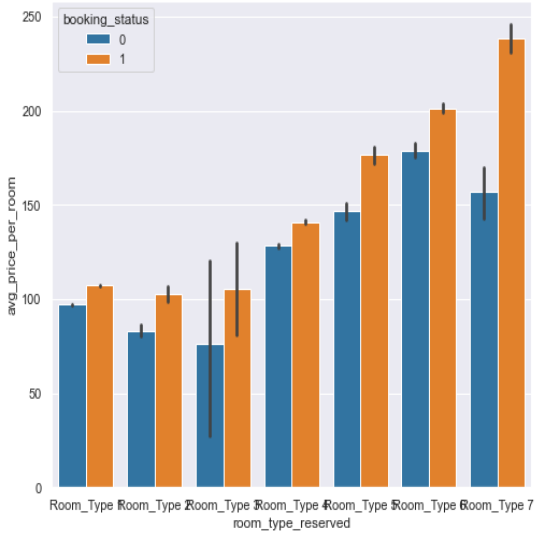


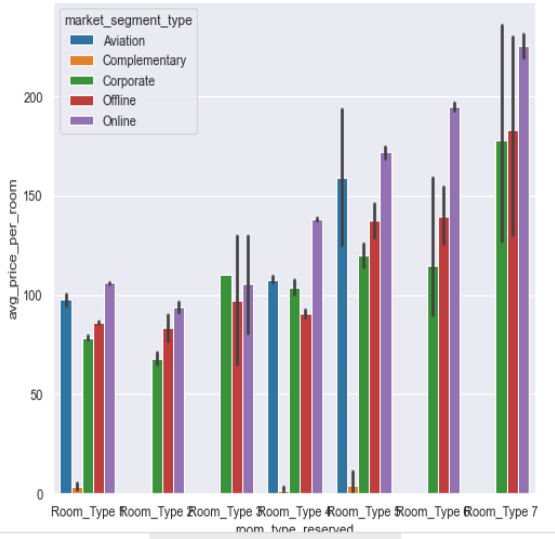


EDA(Contd.)

1. For all kinds of room types, there are more cancelations than non Canceled rooms.
2. Room Type 7 is most expensive for all types of market segmentation
3. Room Type 2 is cheaper for all types of market segmentation

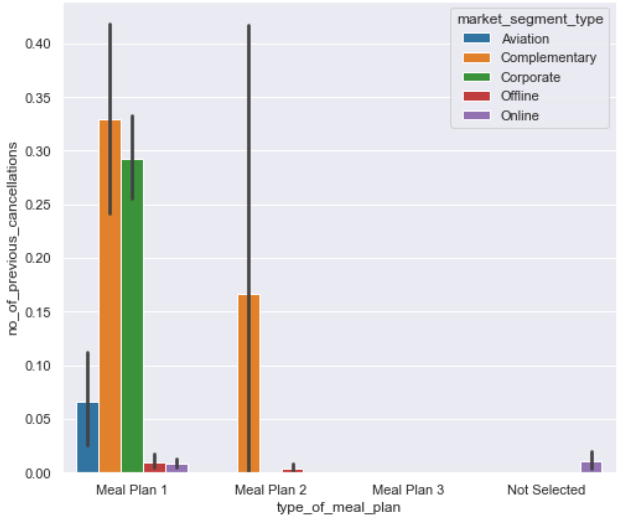


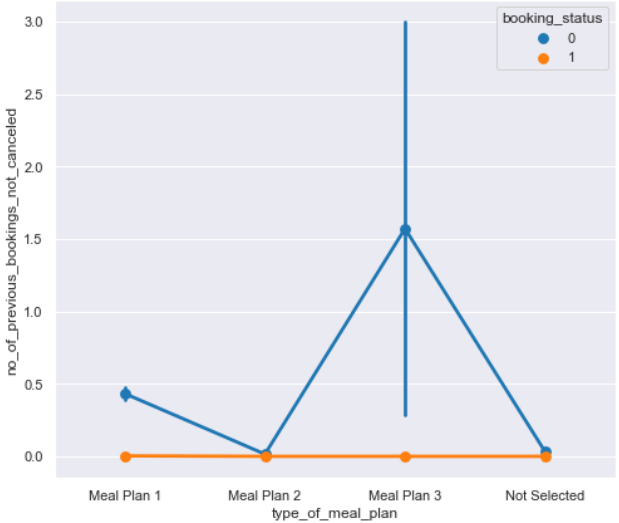


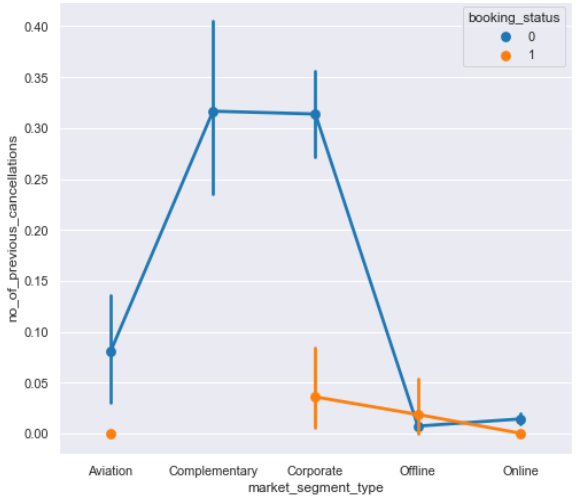


EDA(Correlation)

1. Number of previous cancelations is least for all the Market Segments if the booking status is canceled.
2. Number of previous bookings not canceled is most for Meal Plan 3 and booking status is not canceled.
3. Number of previous bookings not canceled is least for booking status is canceled for all the meal plans.
4. Number of previous cancelations is least for market segment type is online and offline for all the meal plans.
5. Number of previous cancelations is highest for market segment type is Complementart and Corporate







EDA - Summary

1. Most of the customers prefer Meal Plan 1, followed by Mean plan 2. Rest of customers who have not selected a meal preference.
2. There are least cancellations in Meal plan 3. Cancellations are highest in Meal plan 2.
3. Most preferred room type is 1. followed by type 4, type 6 and 5. Least preferred room type is type3 followed by 7.
4. Percentage of cancellations is high in type 6. There are no cancellations in type 7 and 3.
5. Most of customers are online bookers followed by offline and coroprate. Least is aviation.
6. There are no cancellations in market segment type - Aviation, complementary and corporate.
7. Very few cancellations in offline segment. However, close to 65% cancellations in online segment.

Model Performance Summary

- Logistic Regression and Decision trees were used to implement Machine Learning techniques
- Factors Accuracy, Recall, Precision and F1 values are calculated and compared to find the best model.

Logistic Regression

Test set performance comparison:

	Logistic Regression sklearn	Logistic Regression-0.306 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.790730	0.771549	0.791044
Recall	0.610123	0.808254	0.721158
Precision	0.735367	0.630559	0.686266
F1	0.666916	0.708433	0.703280

Training performance comparison:

	Logistic Regression sklearn	Logistic Regression-0.306 Threshold	Logistic Regression-0.40 Threshold
Accuracy	0.793410	0.772506	0.788109
Recall	0.613306	0.806752	0.714088
Precision	0.733483	0.627957	0.677914
F1	0.668033	0.706214	0.695531

Decision Tree - StatsModel

Training

	Accuracy	Recall	Precision	F1
0	0.760863	0.760816	0.619958	0.683202

Testing

	Accuracy	Recall	Precision	F1
0	0.76098	0.770178	0.622902	0.688755

Decision Tree -with depth limited to 3

```
Accuracy value of training set :  0.7710968694426735
Accuracy value of test set :  0.7707664605026228
```

```
get_recall_value(d_Tree)
```

```
Recall on training set :  0.7480447480447481
Recall on test set :  0.7549019607843137
```

```
get_f1_value(d_Tree)
```

```
f1 on training set :  0.6889760189659889
f1 on test set :  0.6934031413612565
```

1. Recall on training data has reduced from 0.99 to 0.74 which is an improvement as the model is not overfitting.

Decision Tree -Hyperparameters

```
Accuracy on training set :  0.76086300036909O4
Accuracy on test set :  0.7609801925937525
Recall on training set :  0.7608157608157609
Recall on test set :  0.7701778385772914
f1 on training set :  0.6832022047384095
f1 on test set :  0.6887552247935569
```

Decision Tree -Best Model

```
DecisionTreeClassifier(ccp_alpha=0.00013438608720037193, random_state=1)
Training accuracy of best model:  0.8453511391470657
Test accuracy of best model:  0.838409144288734
```

```
# Recall of best model on train and test
get_recall_value(best_model_head1)
```

```
Recall on training set :  0.7608157608157609
Recall on test set :  0.7701778385772914
```

```
get_f1_value(best_model_head1)
```

```
f1 on training set :  0.6789169132912232
f1 on test set :  0.68408262454435
```

Model Performance Comparison and Conclusions

	Model	Train_Recall	Test_Recall
0	Logistic Regression sklearn	0.610	0.610
1	Logistic Regression-0.306	0.810	0.810
2	Logistic Regression - 0.40	0.710	0.710
3	Decision Tree stats Model	0.760	0.770
4	Decision tree with restricted maximum depth	0.748	0.745
5	Decision treee with hyperparameter tuning	0.760	0.770
6	Decision tree with post-pruning	0.760	0.770

	Model	Train_accuracy	Test_accuracy
0	Logistic Regression sklearn	0.79	0.79
1	Logistic Regression-0.306	0.77	0.77
2	Logistic Regression - 0.40	0.79	0.79
3	Decision Tree stats Model	0.76	0.76
4	Decision tree with restricted maximum depth	0.77	0.77
5	Decision treee with hyperparameter tuning	0.76	0.76
6	Decision tree with post-pruning	0.85	0.84

	Model	Train_f1	Test_f1
0	Logistic Regression sklearn	0.67	0.67
1	Logistic Regression-0.306	0.71	0.71
2	Logistic Regression - 0.40	0.70	0.70
3	Decision Tree stats Model	0.68	0.69
4	Decision tree with restricted maximum depth	0.69	0.69
5	Decision treee with hyperparameter tuning	0.68	0.69
6	Decision tree with post-pruning	0.68	0.68

Model Performance Comparison and Conclusions

1. According to the Recall definition, high recall means less false negatives. Lower chances of predicting the cancelation as non-cancelation. Recall should be maximized, the greater the Recall higher the chances of identifying.
2. Here, the highest recall value is obtained for Logistic Regression Stats Model - 0.306 Threshold.
3. The highest accuracy value is obtained for Decision tree with post-pruning
4. The difference between recall values for Logistic Regression Stats Model - 0.306 Threshold and Decision tree with post-pruning is 4%. The value of recall for Logistic Regression Stats Model - 0.306 Threshold is only slightly more.
5. The Decision tree with post-pruning has highest accuracy 0.84 and recall value is 0.77. This tree with post pruning is not complex and easy to interpret.
6. The Decision tree with post-pruning is preferred.

Business Insights and Recommendations

- The following attributes play major role in cancellations.
    1. Lead time,
    2. average price per room,
    3. no of special requests,
    4. market segment type online
- Bookings with high lead time have high cancellations. Lead time > 150 are likely to be cancelled.
- Bookings with high Average price per room are likely to be cancelled. Average price per room > 100 are likely to be cancelled.
- If Lead time is < 150, # special requests <= 0.5 and booking is online then it is likely to be cancelled.

Business Insights and Recommendations

- Special attention should be paid to online bookings. Analyze further to see if these online bookings being cancelled are from same customer or agency.
- These could be double bookings. If such a pattern is identified, we could mark them for pre cancellation and make them available for other customers.
- On Bookings with high lead time, try adding a higher cancellation fee. Alternately, disallow bookings more than 150 days in advance.
- Make cancellation fee a percentage of the Average price per room so that higher priced rooms are not easily cancellable.
- Alter refund policy such that cancellations closer to stay date are disallowed or have a high penalty.
- Since percentage of cancellations is high in type 6. Discontinue type 6. Alternately, model type 6 rooms on similar lines as types 7 and 3 as they have least cancellations.

Business Insights and Recommendations

- Try to allocate more rooms via corporate tie-ups for - Aviation, complementary and corporate customers as these customers seem more reliable. Increasing sales of these segments will mitigate risk of cancellations in online segment.

- Encourage longer weekend nights for hotel bookings as such bookings seem to have least cancellations. Provide special discounts on such packages to entice customers.

greatlearning
Power Ahead

Happy Learning !