Star Hotels – ML model for predicting hotel cancelations

Problem – Significant number of hotel bookings are called off due to no-show or cancelations.

Business Impact due to Cancelations –

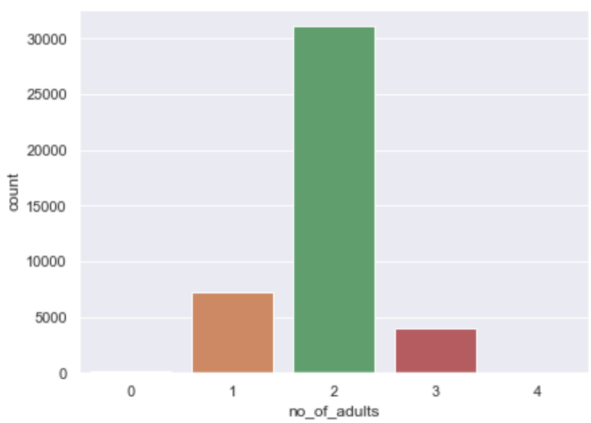
- **Loss of Revenue** when hotel cannot re-sell the room
- **Additional cost of distribution channels** (publicity to help sell these rooms)
- **Reduced Profit Margin**- lowered prices to help sell these rooms
- **Human Resources** to make arrangements for guest (who canceled)

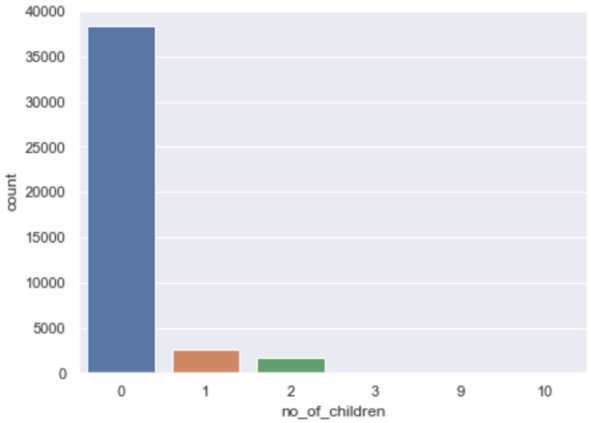
Attributes –

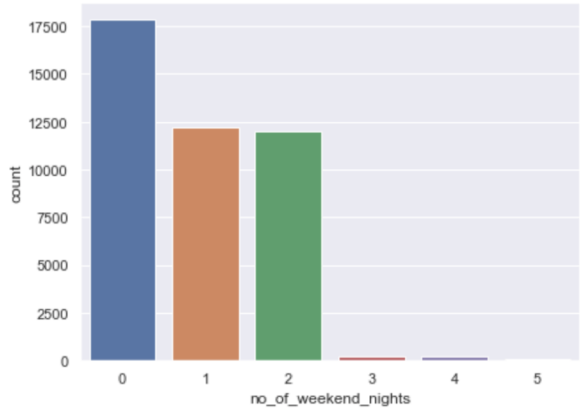
- No of adults
- No of children
- No of weekend nights in booking
- No of weekday nights in booking
- Type of Meal plan option selected
- If a car parking is required?
- Type of room reserved in booking
- Lead time (between booking & check-in)
- Arrival year, month & date
- Market Segment
- Is guest a repeat guest?
- No of previous cancelations by guest
- No of previous bookings not canceled by guest
- Average price per room
- No. of special requests

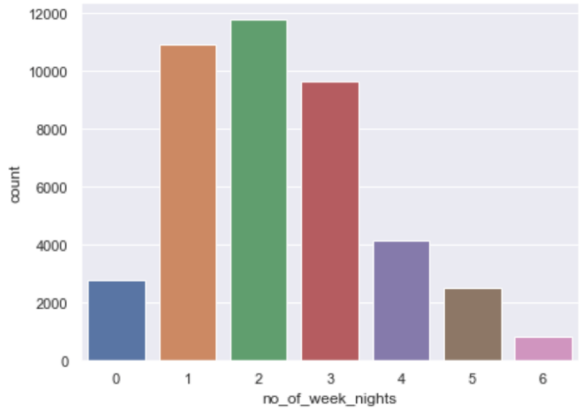
Predictor → Booking Status: Canceled or Not Canceled

Key Findings – Exploratory Data Analysis



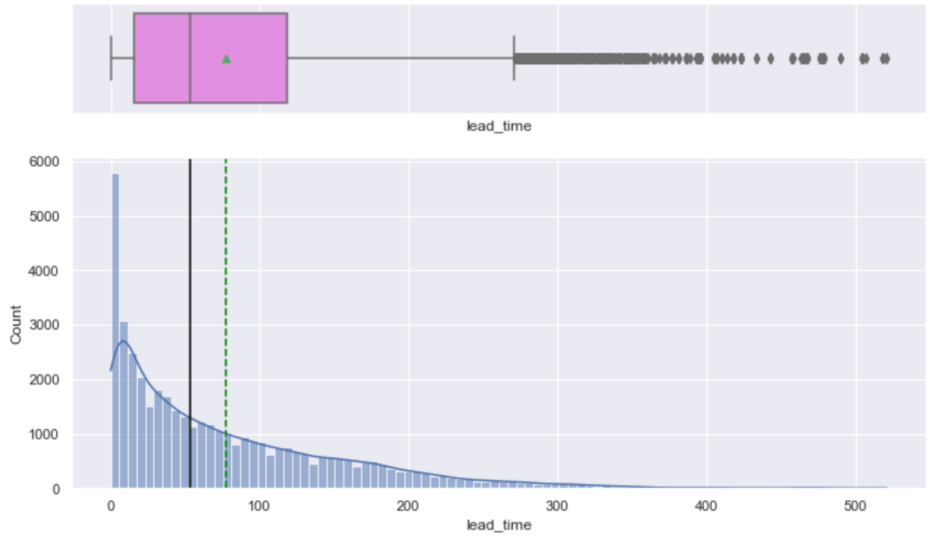


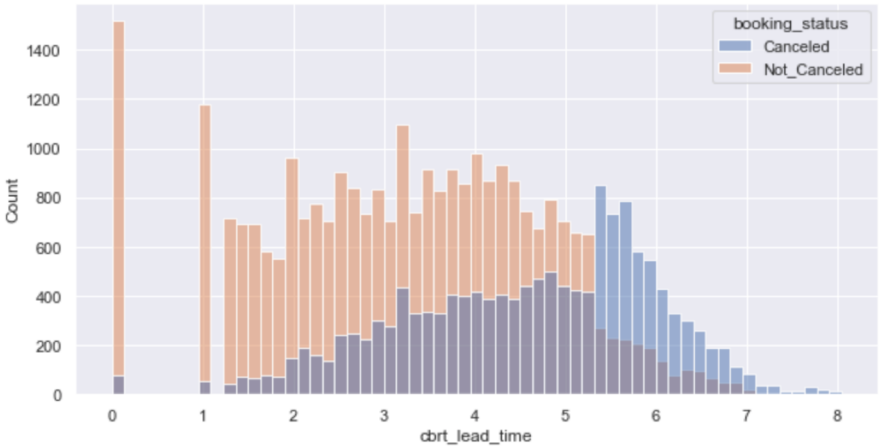




- Most bookings are made for 2 adults & 0 children

- Majority of the bookings are made over the weekdays (spread over 1-3 days) in comparison to weekends

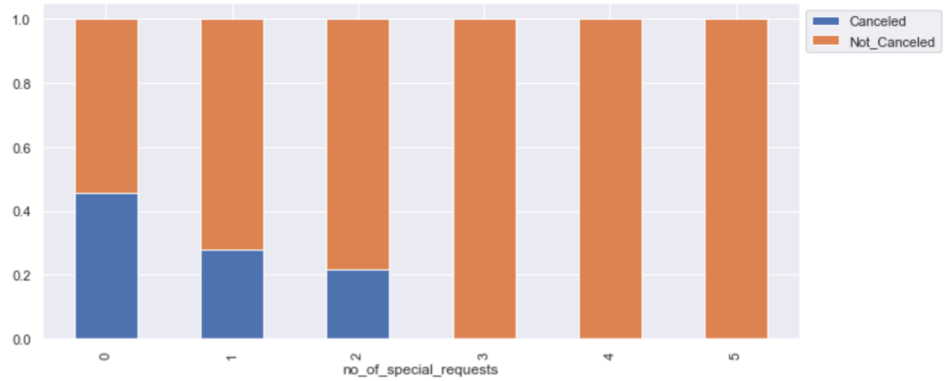


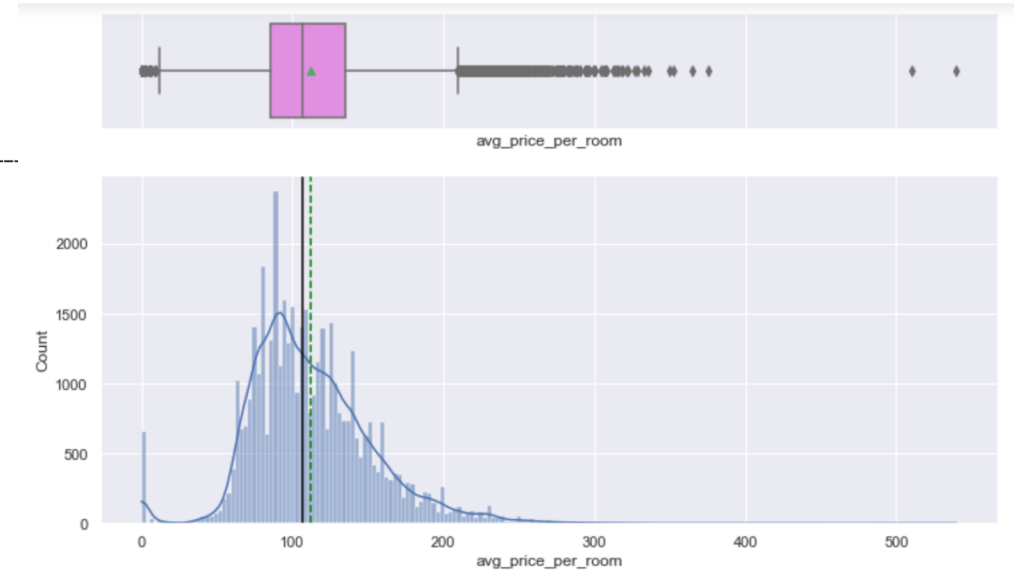


- Majority of guests book closer to check in-date with both average & median falling under 3 months (90 days)
- Lead time (time between booking & check-in) is right skewed with several outliers booking more than 6 months (240 days) in advance
- The lead time was transformed via a cube-root transformation (to treat skewness). As the lead time increases, it was observed that the odds of (Cancelation : No Cancelation) increases as well.

    - Hotel should introduce policies to restrict how far in advance a booking can be made to decrease the odds for cancelations

booking_status	Canceled	Not_Canceled	All
no_of_special_requests			
All	14480	28061	42541
0	8747	10457	19204
1	4344	11217	15561
2	1389	4991	6380
3	0	1230	1230
4	0	150	150
5	0	16	16



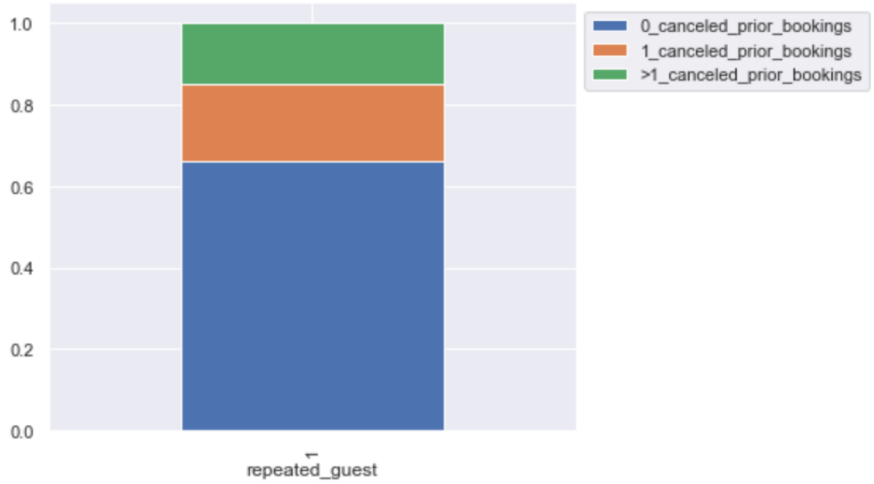


- Majority of guests have no special requests. Some have 1 or 2 requests and only a minority of guests have up-to 5 special requests

- More bookings are canceled when no special requests are made. Bookings with 3 or more special requests have 0 cancelations

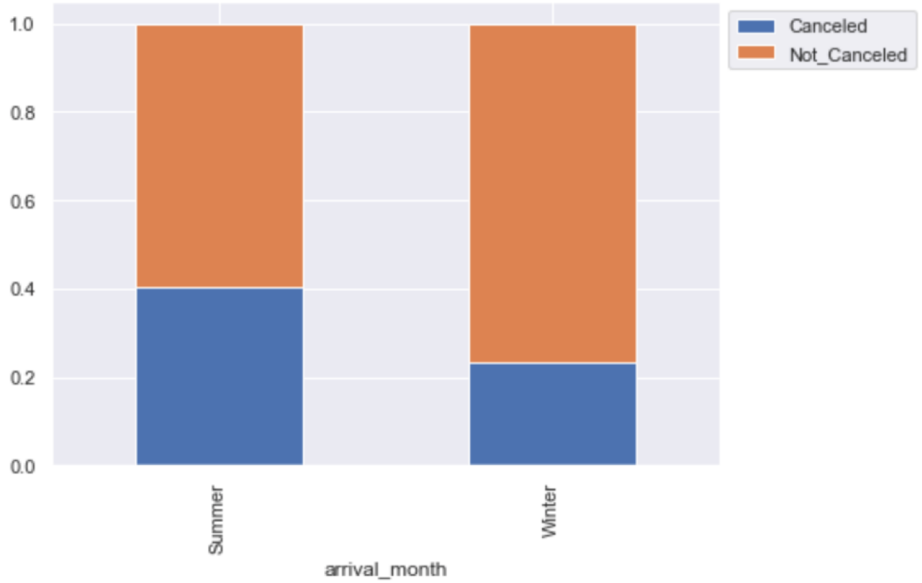
- Average price per room is skewed right with outliers in the range >£ 200

- Out of 40,000+ guests, less than 1500 guests indicated needing a parking spot
- Out of 40,000+ guests, less than 1500 guests were found to be repeat guests



- Out of the <1500 repeat guests, more than 60% have 0 prior canceled bookings and only less than 10% have more than 1 prior canceled bookings

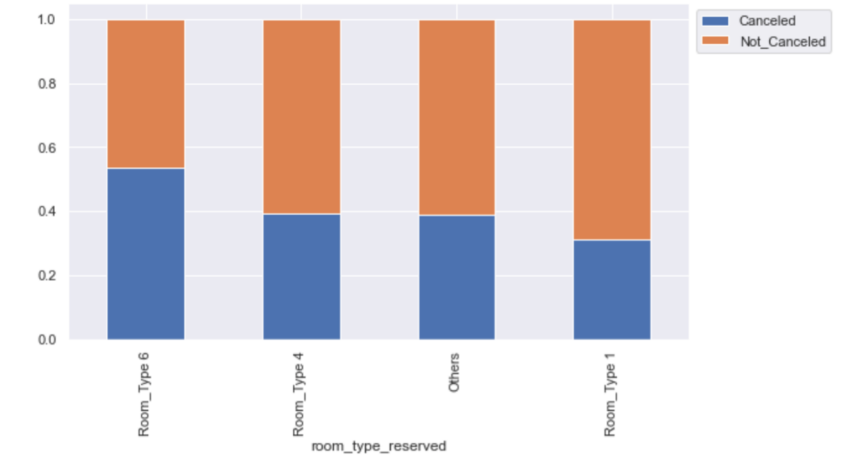
- Dataset has entries between July 2017- August 2019
- Summer (March-August) & Winter (September-February)



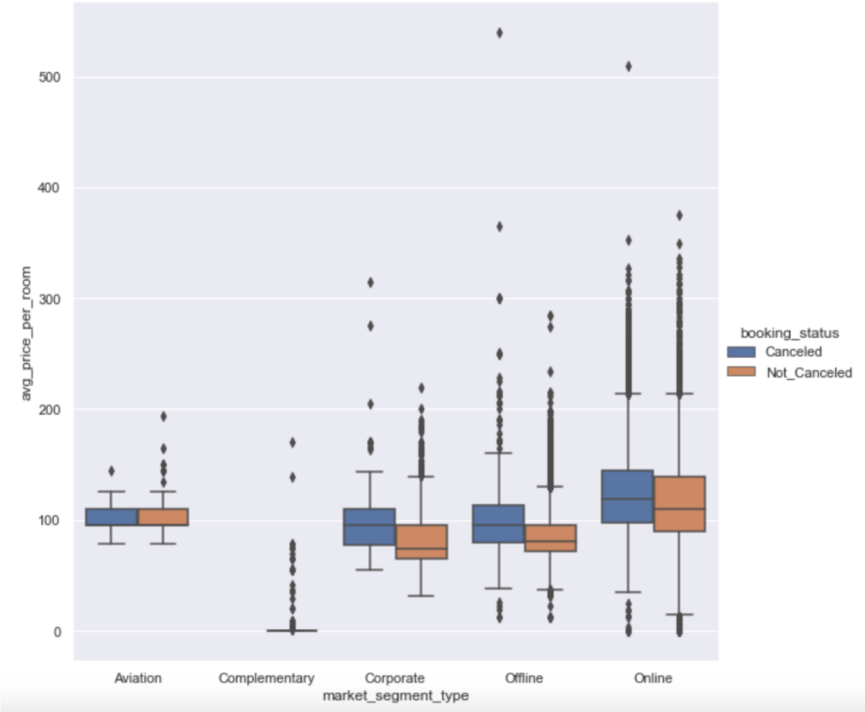
- More bookings are made over the summer months (26K+) over the winter months (15K+). About 40% and 20% of all bookings are canceled in summer and winter months respectively

- Majority of customers have the following room order preference: Room type 1 > type 4 > type 6
- Cancelation follows the following room order preference: Room type 1 < type 4 < type 6

All	14480	28061	42541
Room_Type 1	9219	20488	29707
Room_Type 4	3682	5679	9361
Room_Type 6	826	712	1538
Others	753	1182	1935



- Hence, a guest preferring a room type 1 is less likely to cancel. Hotel needs to communicate these findings to market each room appropriately



- Room prices are dynamic in nature. Prices are higher in online market segment, than other segments like aviation, corporate and offline
- Across all segments, bookings have been canceled in instances where prices are higher & not canceled when prices are lower
- There are no cancelations in Complimentary category

- Correlation observed b/w price per room & no of adults & children which makes intuitive sense
- Correlation observed b/w no of week nights and weekend nights (as longer stays will cover more of both)
- Linear correlation observed b/w lead time & no of week nights indicating longer trips are booked in advance
- Strong relationship observed between previous bookings not canceled & no of previous cancelations (verified by statistical tests)
- Weak correlation observed b/w lead time & price with odd of cancelations being high for both high lead time and high price

***multicollinearity needs to be treated for solving logistic regression models. Decision trees are however immune to linearly correlated attributes**

Model Evaluation Criteria

Model can make a wrong prediction as:

- Predicting a person will cancel a booking, when a person will not cancel the booking (False Negatives).

This will result in loss of potential revenue & business for the hotel chain

- Predicting a person will not cancel a booking, when a person will cancel the booking (False Positives)

This will result in last minute cancellations -loss of revenue due to hiring of human resources for guests who will no longer come, as well as profit-margin loss in case of trying to price the room cheap to get last minute bookings

- **F1_score should be maximized. The greater the F1_score higher the chances of identifying both the classes correctly**

The dataset was split 70%-30% to be used for training & testing purposes.

- Logistic Regression model was fit using Sklearn & Stats libraries (Model was further trained & refined)
- Decision trees was used for classification (Model was further pre/post pruned for comparison & refinement)

- Accuracy, Precision, Recall & F1 score metrics were evaluated & important attributes for predictions were identified

Logistic Regression Model Comparison

Training performance comparison:

	Accuracy	Recall	Precision	F1
Sk-learn	0.787058	0.867981	0.819581	0.843087
Logistic Regression-0.50 Threshold	0.772080	0.865841	0.803566	0.833542
Logistic Regression-0.65 Threshold	0.753509	0.749363	0.858652	0.800294
Logistic Regression-0.56 Threshold	0.768386	0.822939	0.825167	0.824052

Test set performance comparison:

	Accuracy	Recall	Precision	F1
Sk-learn	0.788843	0.873859	0.818818	0.845444
Logistic Regression-0.50 Threshold	0.773408	0.864019	0.806819	0.834440
Logistic Regression-0.65 Threshold	0.757345	0.755305	0.860481	0.804470
Logistic Regression-0.56 Threshold	0.769882	0.824659	0.826717	0.825687

***default threshold is 0.5, AUC-ROC optimization gives a threshold of 0.65 (with high precision) & Precision-Recall curve optimization gives a threshold of 0.56 (balanced high of both recall & precision)**

- The models are able to give generalized performance on both the training as well as the testing datasets
- Model using sklearn library gives F1 values of 0.843 & 0.845 on training & testing datasets, i.e. able to explain over 84.5% of the information
- Optimized threshold for stats library was found to be ~0.56 with a balanced high precision & recall. This model gives F1 values of 0.824 & 0.825 on training & testing datasets, i.e. able to explain over 82.4% of the information

Attributes contributing to No_Cancelations

- No_of_adults, No_of_special_requests, type_meal_plan_selected_MealPlan1, required_car_parking_space_1, & repeated_guests

- Repeated_guests contribute the maximum to No_Cancelations (intuitive as they have likely had prior good experience staying at the hotel). Guests that are specific in their requests or their requirements like meal plan & parking or planning to stay with larger group are also less likely to cancel intuitively

- Attributes contributing to Cancelations

- No_of_weekend_nights, no_of_week_nights, avg_price_per_room, cbrt_lead_time, binned_no_of_previous_cancellations_1_canceled_prior_bookings

- Guests who have previously canceled a booking (intuitive given past behavior) or bookings with more expensive rates are the most contributing to Cancelations. Guests who plan longer stays covering many weekdays, weekends, or planning way in advance (months before checking) are likely to contribute to Cancelations

	Odds	Change_odd%
const	310.662714	30966.271435
no_of_adults	1.203417	20.341695
no_of_weekend_nights	0.958035	-4.196452
no_of_week_nights	0.941352	-5.864831
avg_price_per_room	0.978584	-2.141565
no_of_special_requests	2.956435	195.643491
cbrt_lead_time	0.450457	-54.954336
type_of_meal_plan_Not Selected	0.500418	-49.958161
type_of_meal_plan_Not_Meal_Plan 1	1.636813	63.681316
required_car_parking_space_1	3.872028	287.202758
room_type_reserved_Room_Type 1	0.735929	-26.407056
room_type_reserved_Room_Type 4	0.734002	-26.599751
repeated_guest_1	47.239698	4623.969818
binned_no_of_previous_cancellations_1_canceled_prior_bookings	0.058671	-94.132905

Decision Tree Comparison

```
Training performance comparison:
                                  Accuracy    Recall  Precision        F1
Grid Search Hyperparameter Pruning  0.813251  0.864567   0.853908  0.859204
Cost Complexity Pruning             0.810860  0.859751   0.854886  0.857312


Test set performance comparison:
                                  Accuracy    Recall  Precision        F1
Grid Search Hyperparameter Pruning  0.786420  0.811576   0.856805  0.833578
Cost Complexity Pruning             0.785865  0.809247   0.858616  0.833201
```

*** The dataset had ~65% no-cancelations and 35% cancelations. Weights 0.35 & 0.65 were assigned to booking status 1 (no cancelations) & 0 (cancelations) so the decision tree maybe balanced. In cost complexity (post-) pruning, an alpha value of 0.01 was chose to maximize F1 score on both training & testing datasets without loss of information**

- Unlike a normal decision tree which is prone to overfitting, both grid search hyperparameter i.e., pre pruning & cost complexity i.e., post pruning are giving generalized results on both training & testing datasets
- In comparison to all models sklearn, statsmodel (default, AUC-ROC optimization, recall-precision), cost complexity/ post pruned decision tree, – **the grid search hyperparameter / pre tuned decision tree has the highest F1 score and is able to explain 86% of information contained in the dataset**

Important Attributes

- bookings with higher lead times had a pattern of higher cancelations as well

- bookings with more no. of special requests had a pattern of lower cancelations

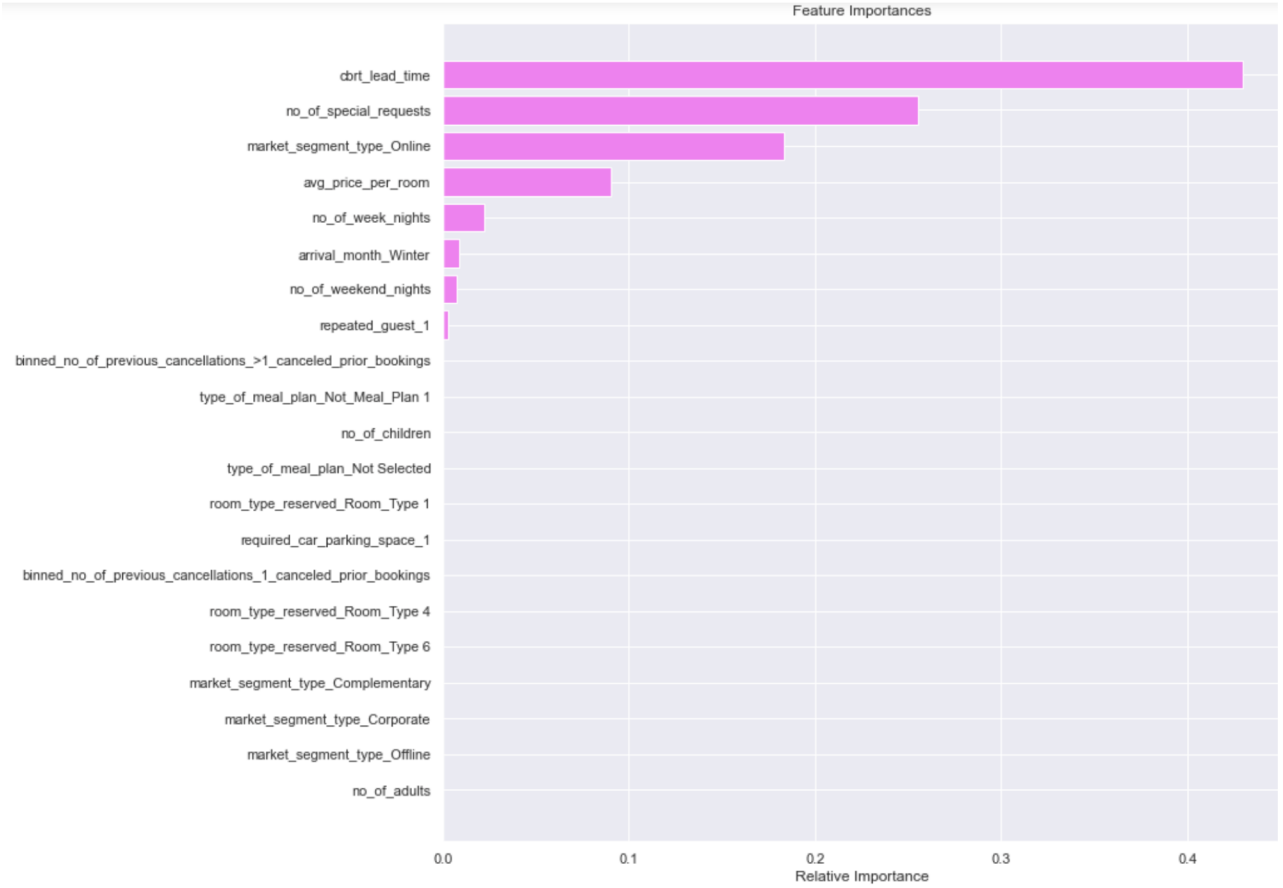
- market segment online was excluded from regression fit, as it had collinear dependencies on other attributes, however included in the decision tree model, as trees are immune to multicollinearity. Online market segment was found to have some of the higher min, median, max avg_price_per_room values compared to any other market segment like offline, aviation, complimentary etc.,

- avg price per room was found to be higher for canceled bookings than for not canceled bookings (across the board for all market segments) during the EDA process

- higher no of week nights indicates longer booking duration, which as was observed from the regression fit, has a higher odds of booking cancelations (similar for no of weekend nights)

- arrival_month winter (Sep-Feb) was found to have a lesser percentage of booking cancelations than arrival_month summer (March-Aug). Although, also noticed number of bookings in summer months are higher than number of bookings in winter months

- a repeated guest is less likely to cancel a booking



Recommendations

- ML model is able to predict cancelations or no cancelations for bookings with a confidence of ~86%. Hotel policies for staffing, publicity and dynamic room pricing need to take into consideration the odds for cancelations & have contingency plans in place
- Lead time was identified as the most important feature with a longer lead time increasing the odds for cancelations. Policies need to be introduced to restrict how far in advance bookings can be made before the check in date
- Similarly, hotel policies need to restrict the length of hotel stay as bookings made for longer stay periods were also found to have increased odds of cancelations
- The repeat guests (although few) were identified to have lower odds of cancelations. Hotel policies need to incentivize current & previous guests to increase conversion as repeated guests
- More bookings (as well as more cancelations) were found to occur over months (March - August) than months (September - February). Broadly policies and plans can be formulated estimating business on this biannual basis
- Majority of customers preferred Room Type 1. As well, this room has a pattern of not having as many bookings cancelled. The room has to be adequately marketed, and priced in order to capitalize on its strengths
- Across all market segments, avg price per room has been higher in instances where bookings have been canceled than in instances where bookings have not been canceled. More competition information is required to ensure that our pricing is competitive to retain guests