INN HOTELS

Project 4 – Supervised Learning Classification

Joshua Willis, PGP-DSBA

CONTENTS

Business Problem:

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 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 must make arrangements for 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. INN Hotels Group has a chain of hotels in Portugal, they are facing problems with the high number of booking cancellations and is looking for a data-driven solution. This presentation analyzes the data provided to find which factors have a high influence on booking cancellations, builds 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 Summary

Data Dictionary

- Booking_ID: unique identifier of each booking
- 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 INN 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.

Data Summary

Dependent variable:

- booking_status = canceled

Object variables:

- type_of_meal_plan
- room_type_reserved
- market_segment_type

Integer variables:

- 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_date, repeated_guest, no_of_previous_cancellations,
- no_of_previous_bookings_not_canceled, no_of_special_requests

Float variables:

- avg_price_per_room

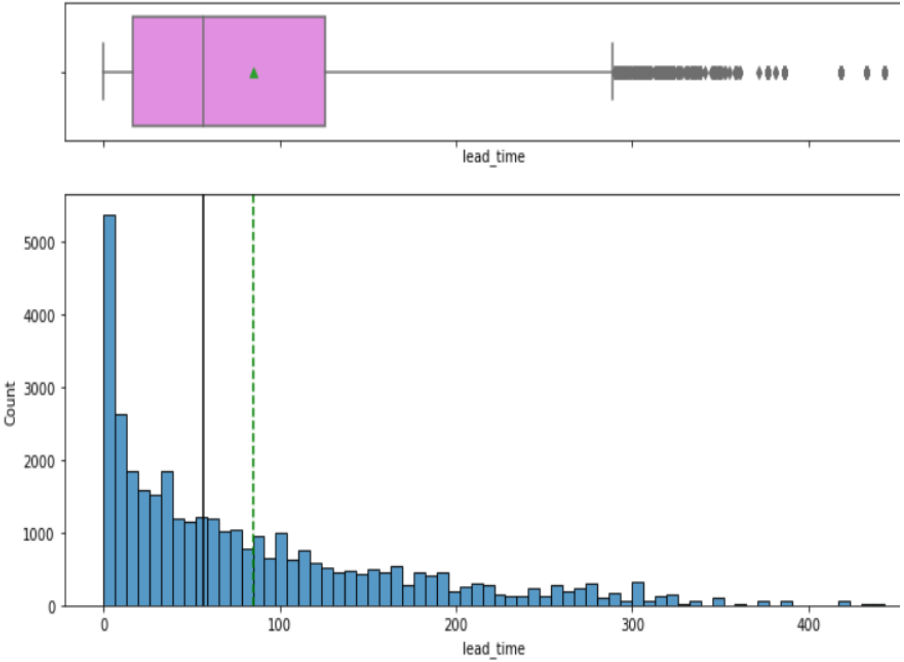
- ➢ Shape: 36,275 row and 19 columns
- ➢ There are no missing values in the dataset.
- ➢ There are no duplications in the dataset.

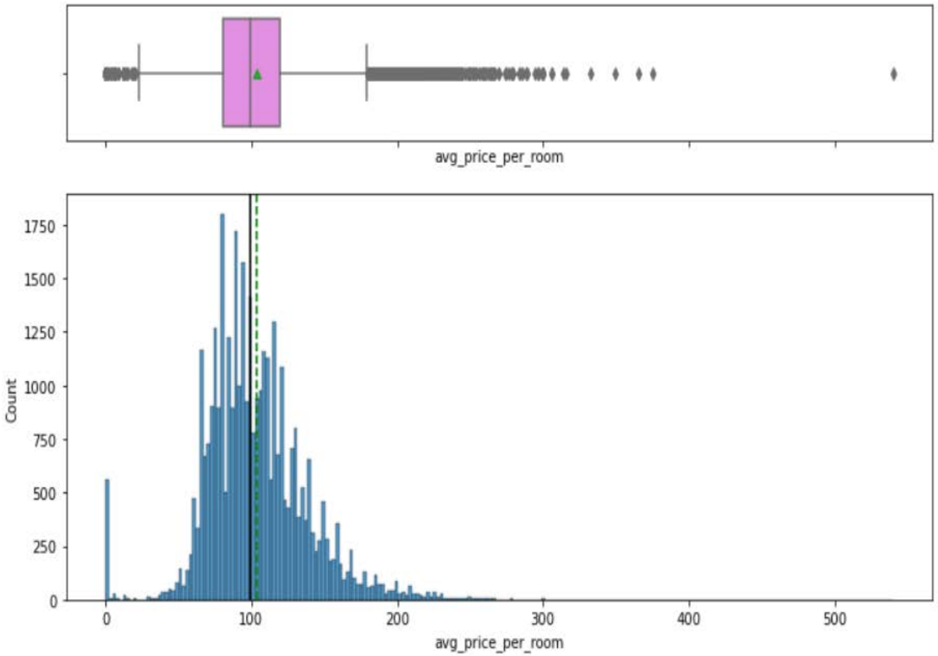
EDA Results

	count	mean	std	min	25%	50%	75%	max
no_of_adults	36275.00000	1.84496	0.51871	0.00000	2.00000	2.00000	2.00000	4.00000
no_of_children	36275.00000	0.10528	0.40265	0.00000	0.00000	0.00000	0.00000	10.00000
no_of_weekend_nights	36275.00000	0.81072	0.87064	0.00000	0.00000	1.00000	2.00000	7.00000
no_of_week_nights	36275.00000	2.20430	1.41090	0.00000	1.00000	2.00000	3.00000	17.00000
required_car_parking_space	36275.00000	0.03099	0.17328	0.00000	0.00000	0.00000	0.00000	1.00000
lead_time	36275.00000	85.23256	85.93082	0.00000	17.00000	57.00000	126.00000	443.00000
arrival_year	36275.00000	2017.82043	0.38384	2017.00000	2018.00000	2018.00000	2018.00000	2018.00000
arrival_month	36275.00000	7.42365	3.06989	1.00000	5.00000	8.00000	10.00000	12.00000
arrival_date	36275.00000	15.59700	8.74045	1.00000	8.00000	16.00000	23.00000	31.00000
repeated_guest	36275.00000	0.02564	0.15805	0.00000	0.00000	0.00000	0.00000	1.00000
no_of_previous_cancellations	36275.00000	0.02335	0.36833	0.00000	0.00000	0.00000	0.00000	13.00000
no_of_previous_bookings_not_canceled	36275.00000	0.15341	1.75417	0.00000	0.00000	0.00000	0.00000	58.00000
avg_price_per_room	36275.00000	103.42354	35.08942	0.00000	80.30000	99.45000	120.00000	540.00000
no_of_special_requests	36275.00000	0.61966	0.78624	0.00000	0.00000	0.00000	1.00000	5.00000

Number of children, number of weekend nights, lead time, previous cancellations, bookings not canceled, price per room appear to be right skewed.

EDA Results

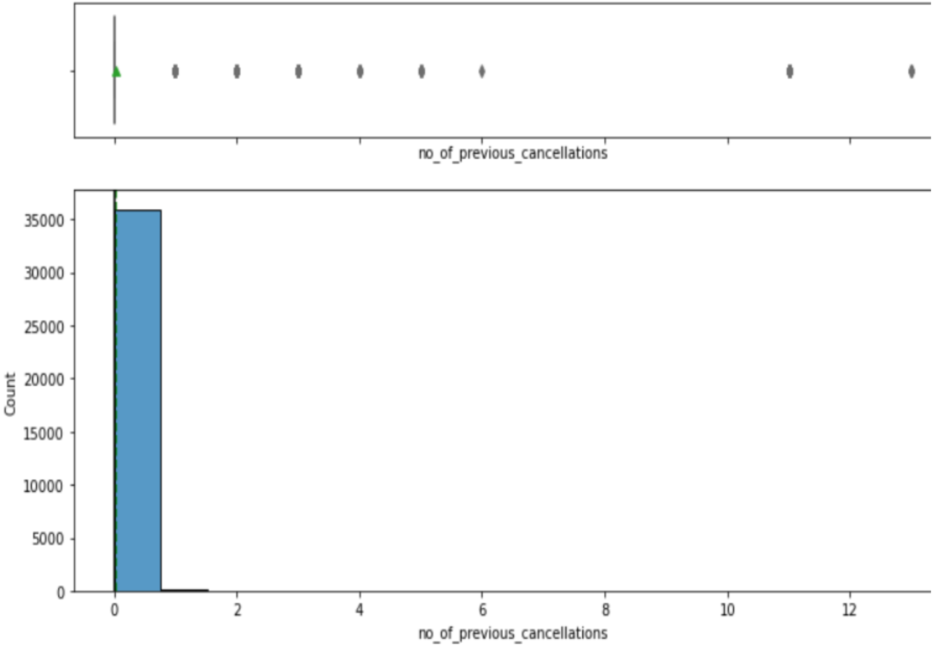


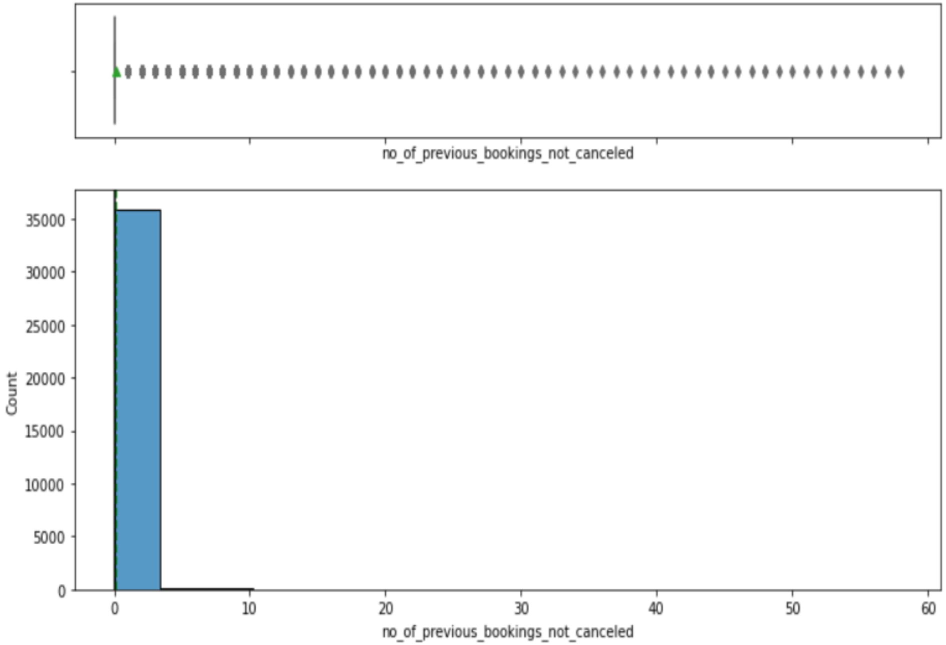


The distribution of lead_time is right-skewed. There are outliers in this variable. From the boxplot we can see that the third quartile(Q3) is equal to 126 which means 75% of the customer make reservations less than 126 days in advance.

There are outliers in this below and above lower/upper limits. The average price of the room is $103.42
Note that average price includes complimentary rooms as well.

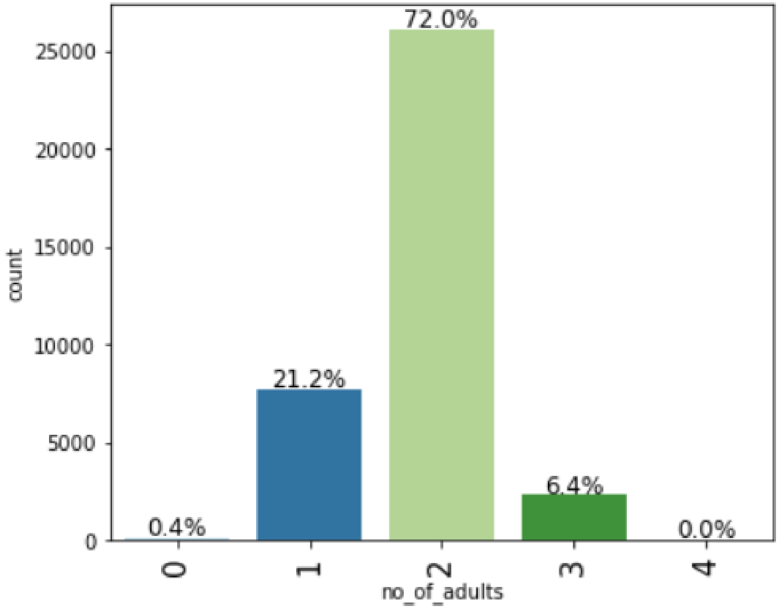
EDA Results

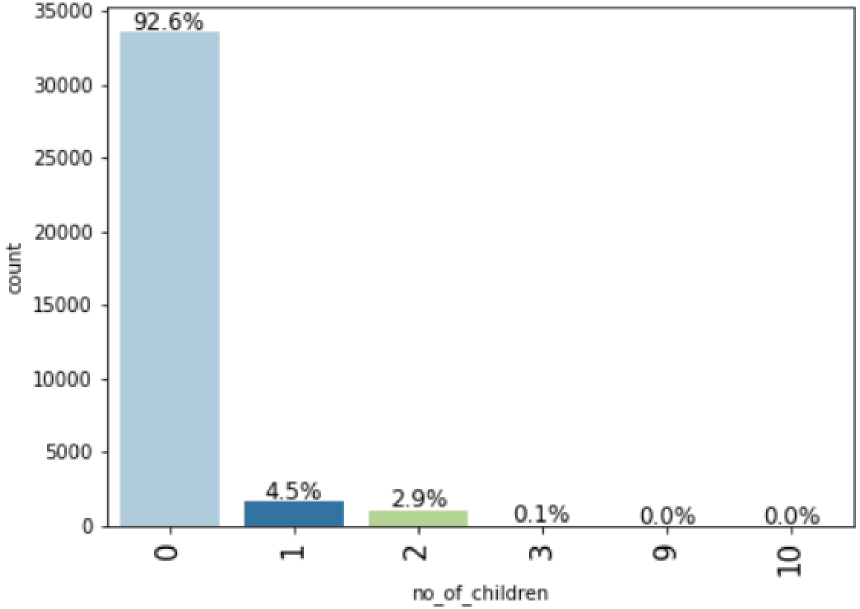




- Average number of previous cancellations is 0.02. There are some outliers.
- Average number of previous bookings not canceled is 0.15. There are some outliers.

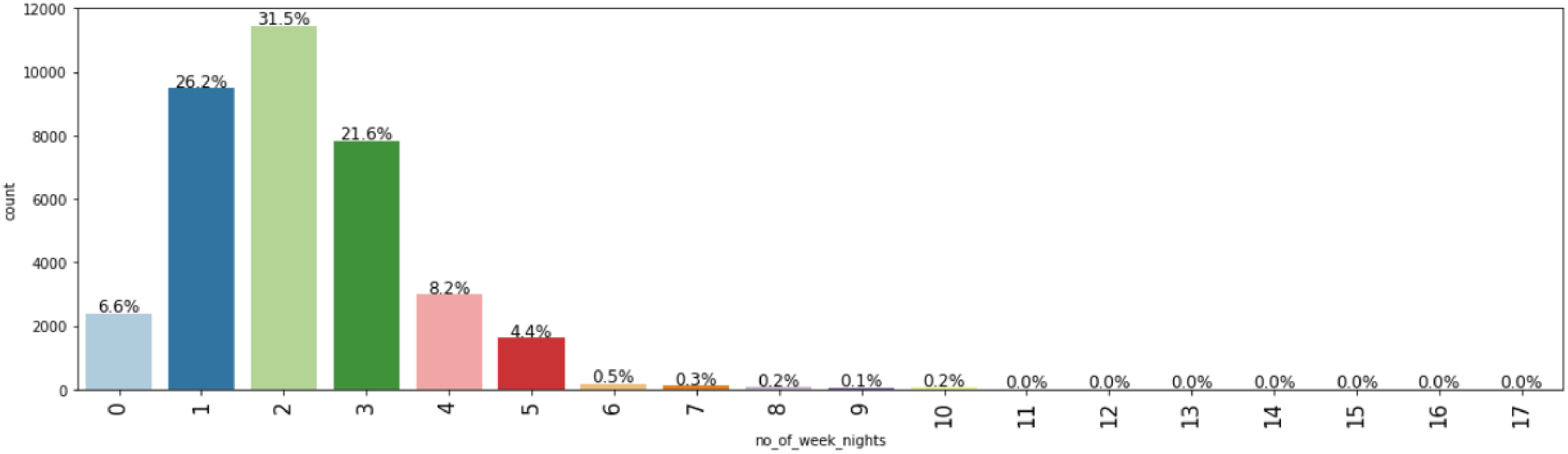
EDA Results





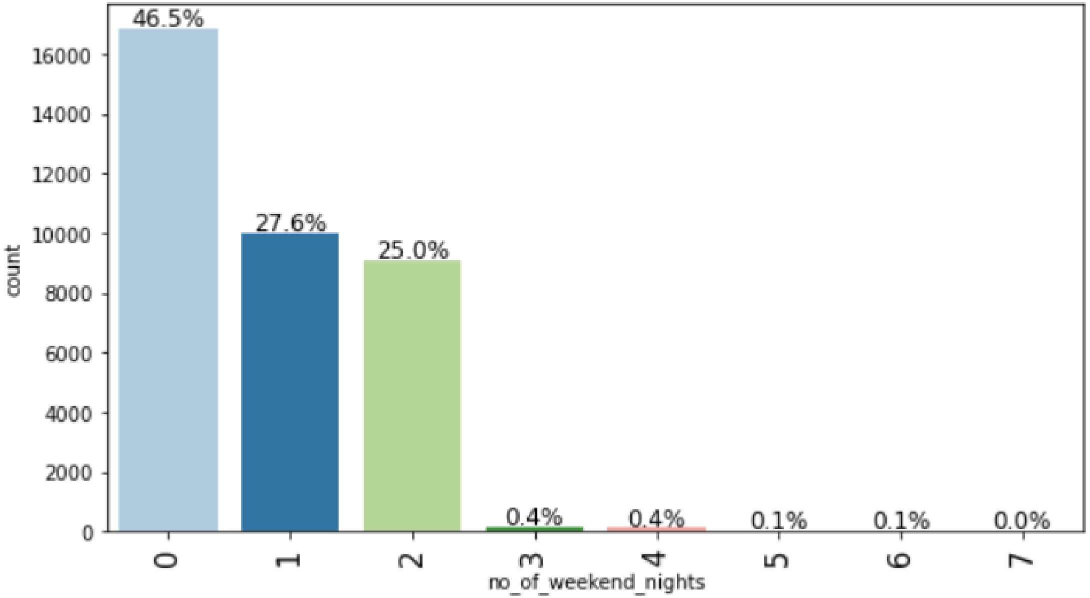
- Majority number of adults staying at the hotel is 2
- Majority of guests staying at hotel have no children with them (replaced those outliers with values of 9 or 10 children with 3 so data is not skewed)

EDA Results



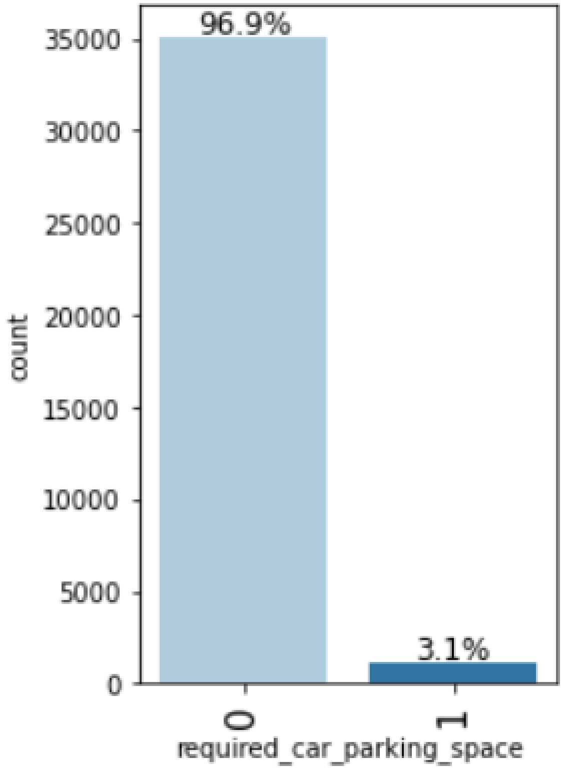
- Majority of guests stay for at least 2 weeknights

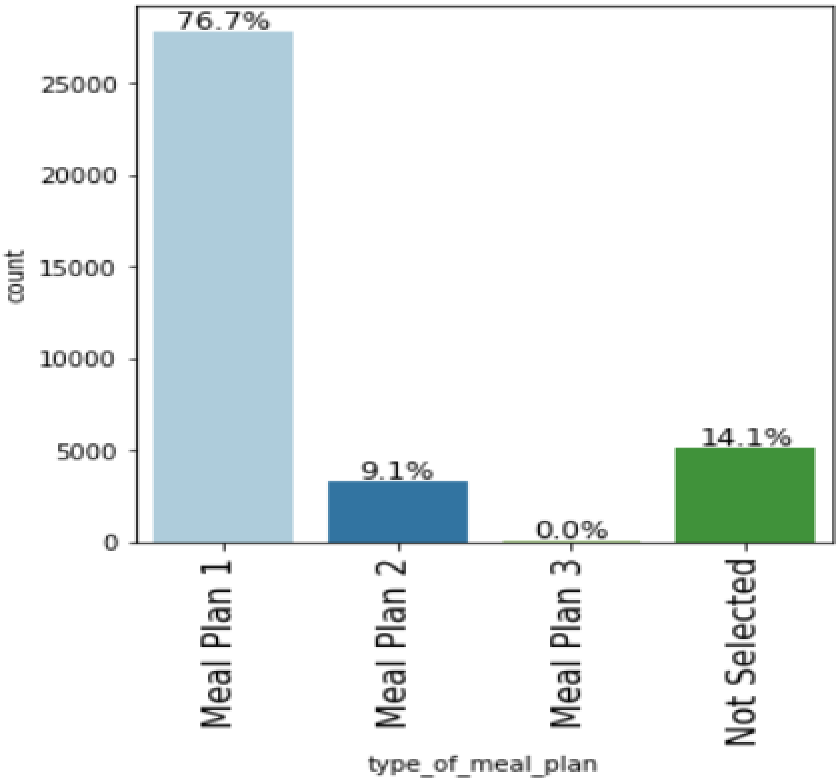
EDA Results



- Majority of guests only stayed for one weekend night

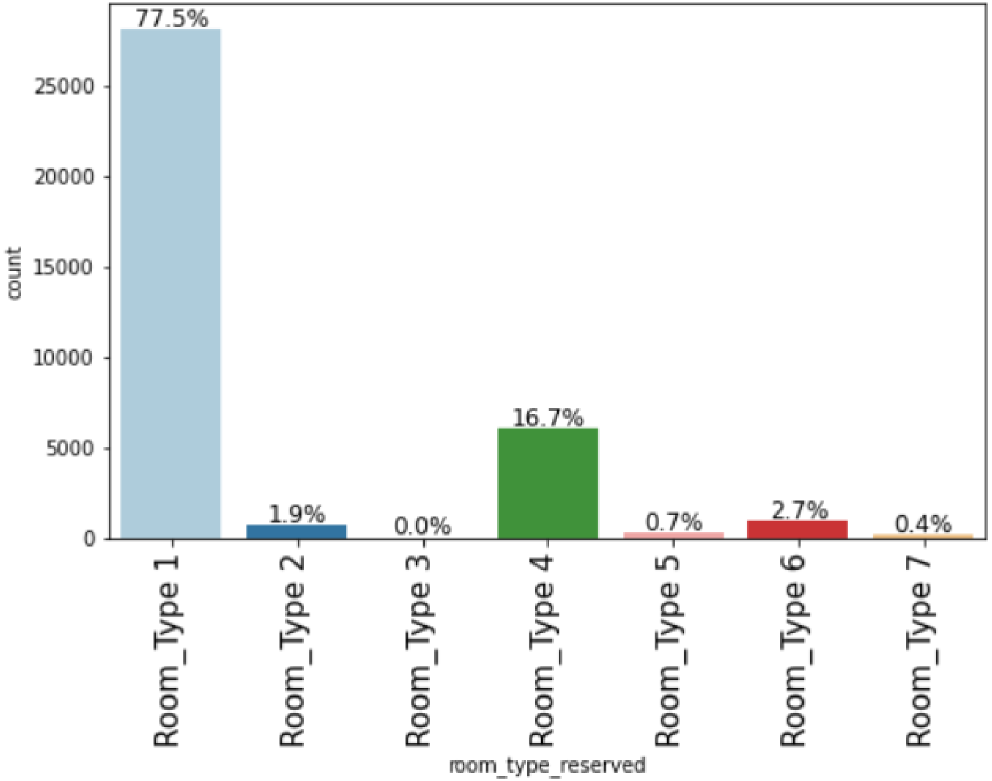
EDA Results





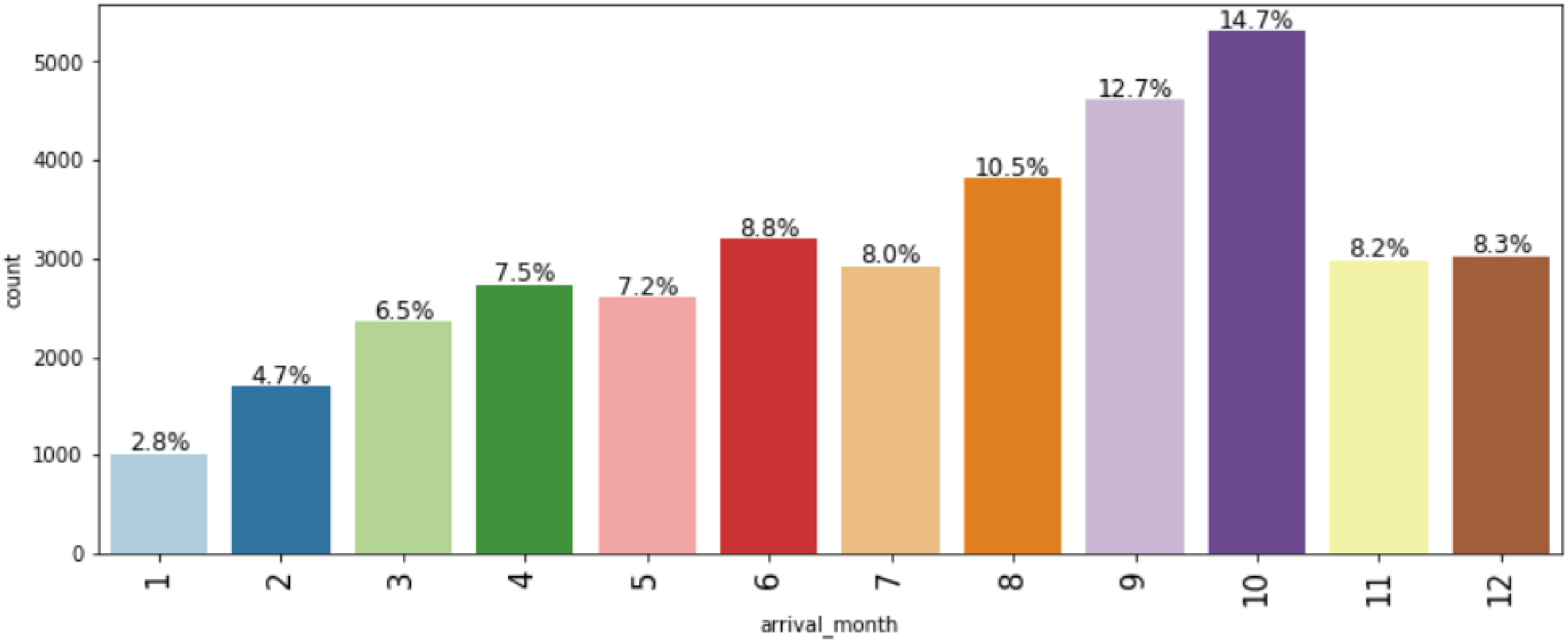
- Majority of guests don't require a parking space
- Majority of guest preferred breakfast meal plan

EDA Results



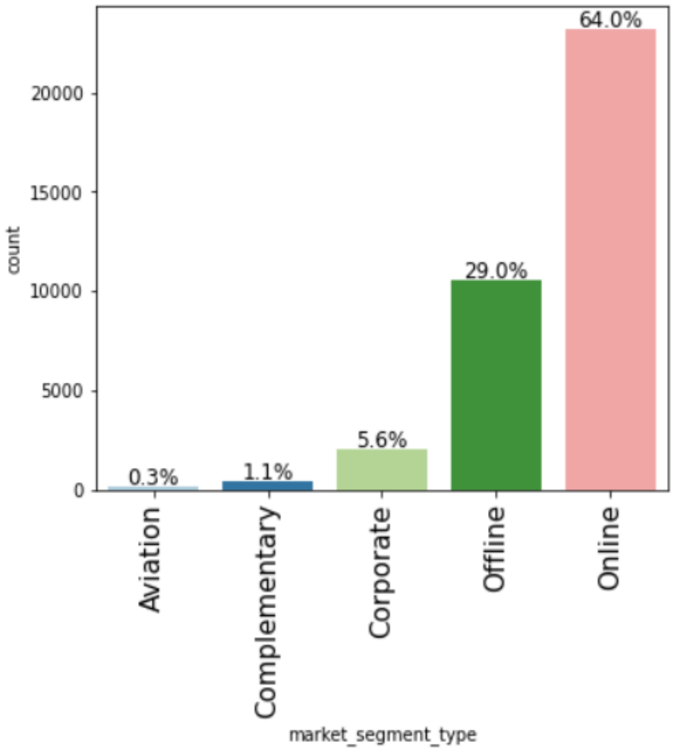
- Most of guests preferred Room Type 1 (encoded by INN Hotel)

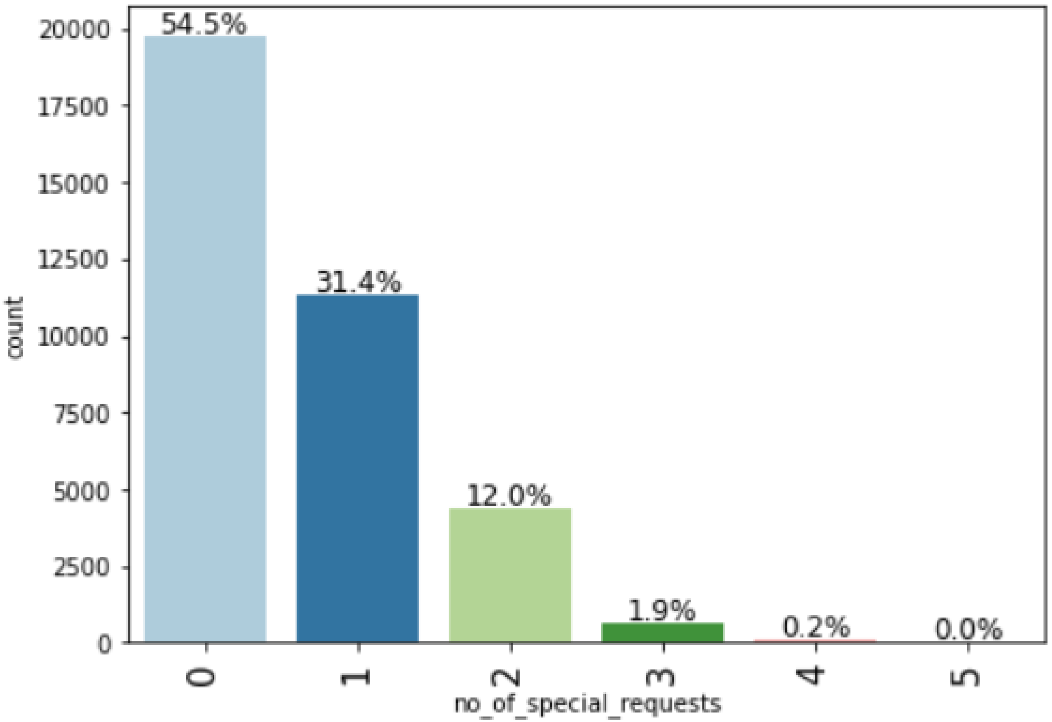
EDA Results



- Majority of guests booked for October; and other popular months were September and August

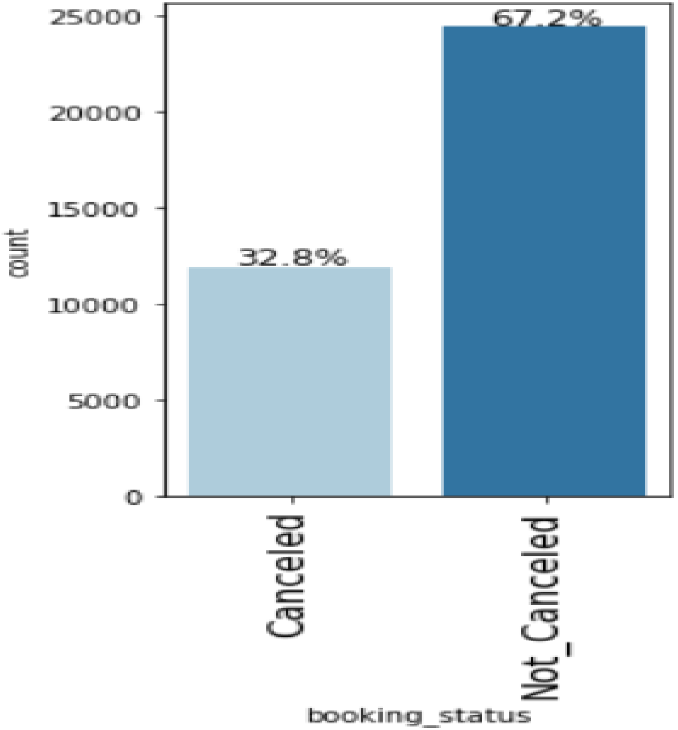
EDA Results





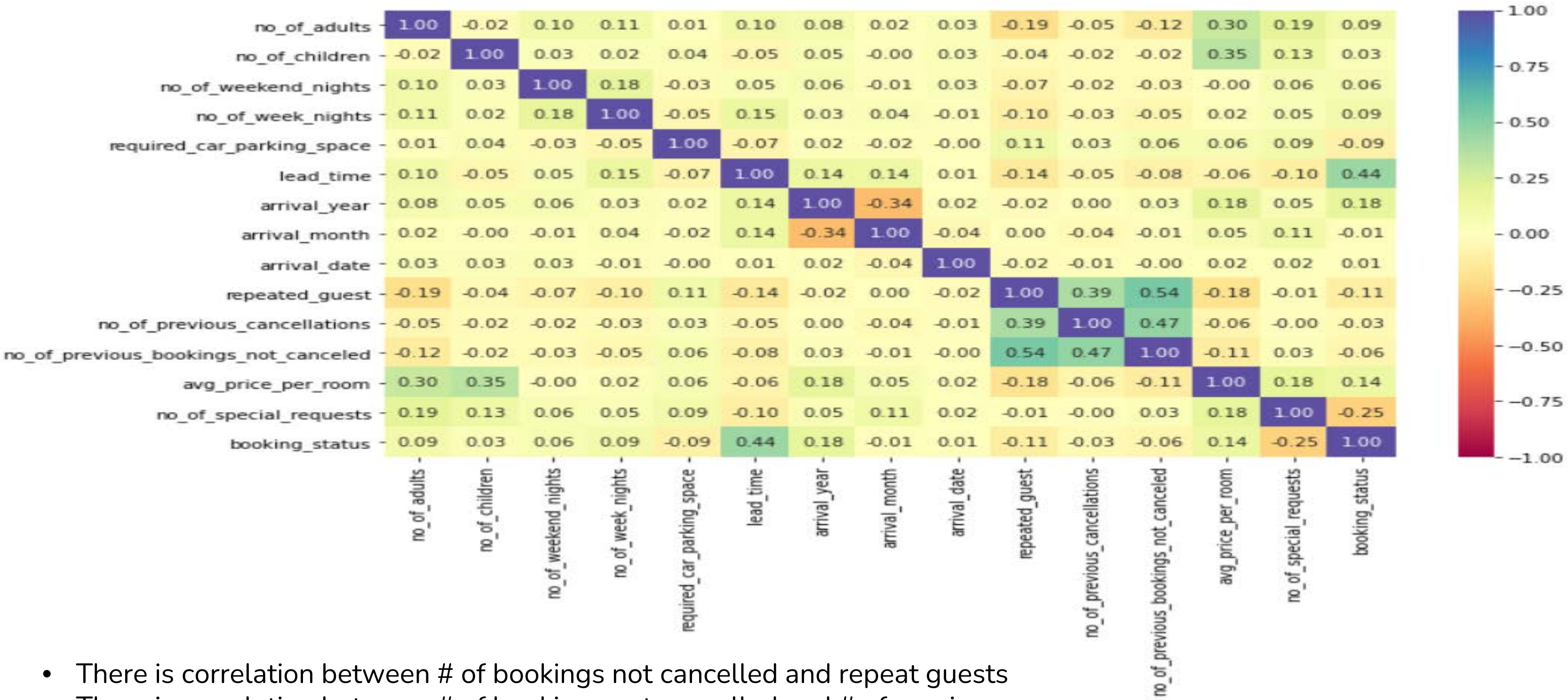
- Majority of reservations are made online.
- Majority of guests don't have a special request.

EDA Results



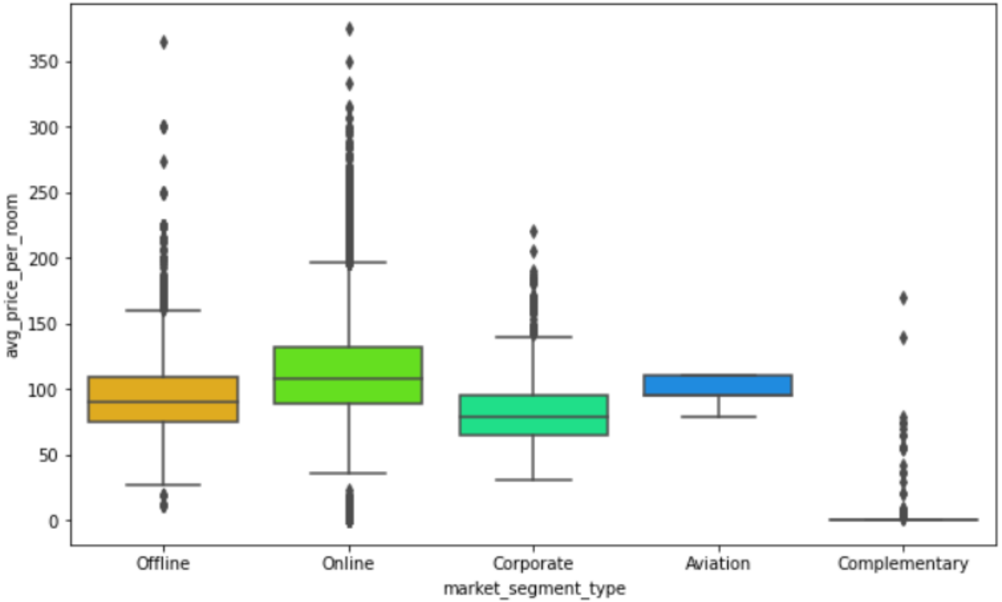
- Majority of reservations don't cancel. However, 32.8% of reservations are canceled which is too high.

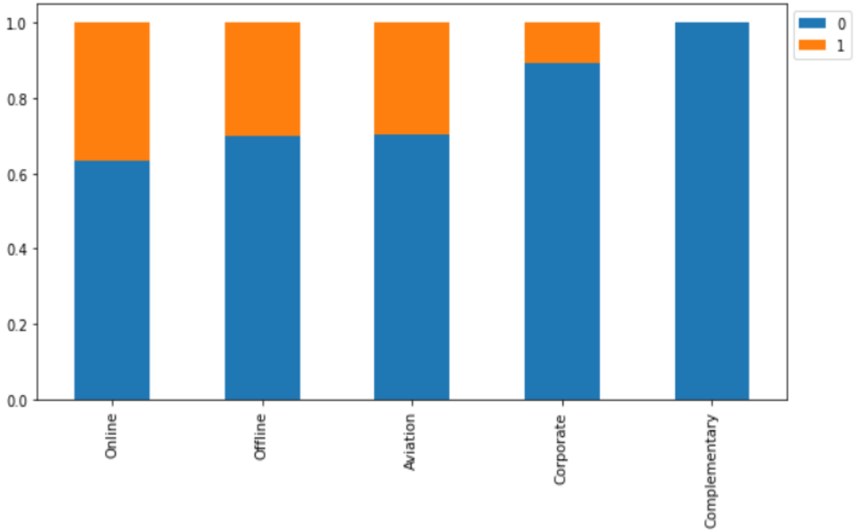
EDA Results



- There is correlation between # of bookings not cancelled and repeat guests
- There is correlation between # of bookings not cancelled and # of previous cancellations
- There is correlation between average price of room and # adults / children
- There is correlation between lead time and booking status

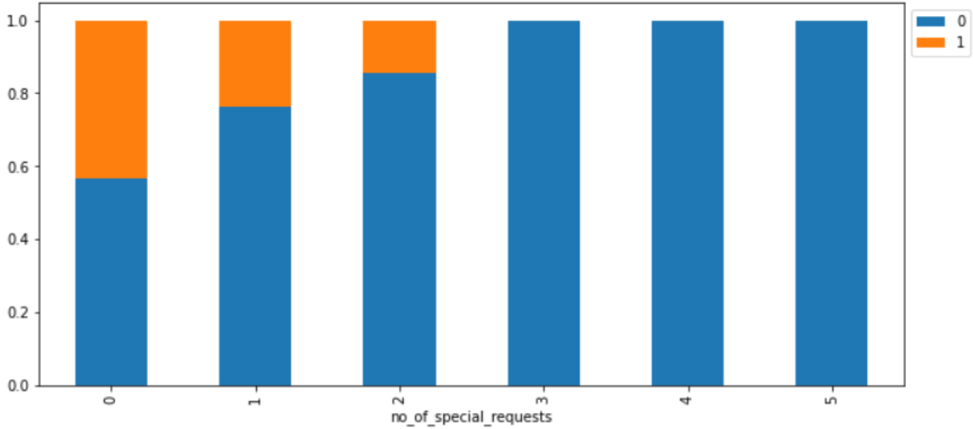
EDA Results



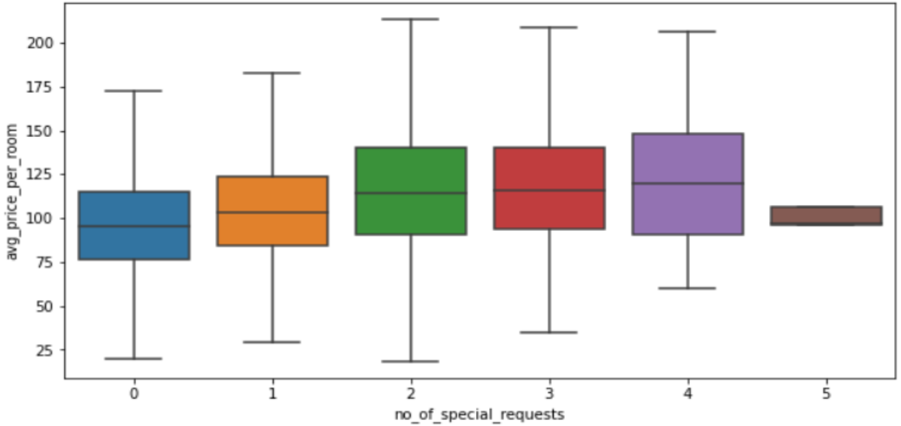


- Prices seems to be higher when rooms are booked online
- Majority of cancellations are those that are complimentary followed behind corporate

EDA Results

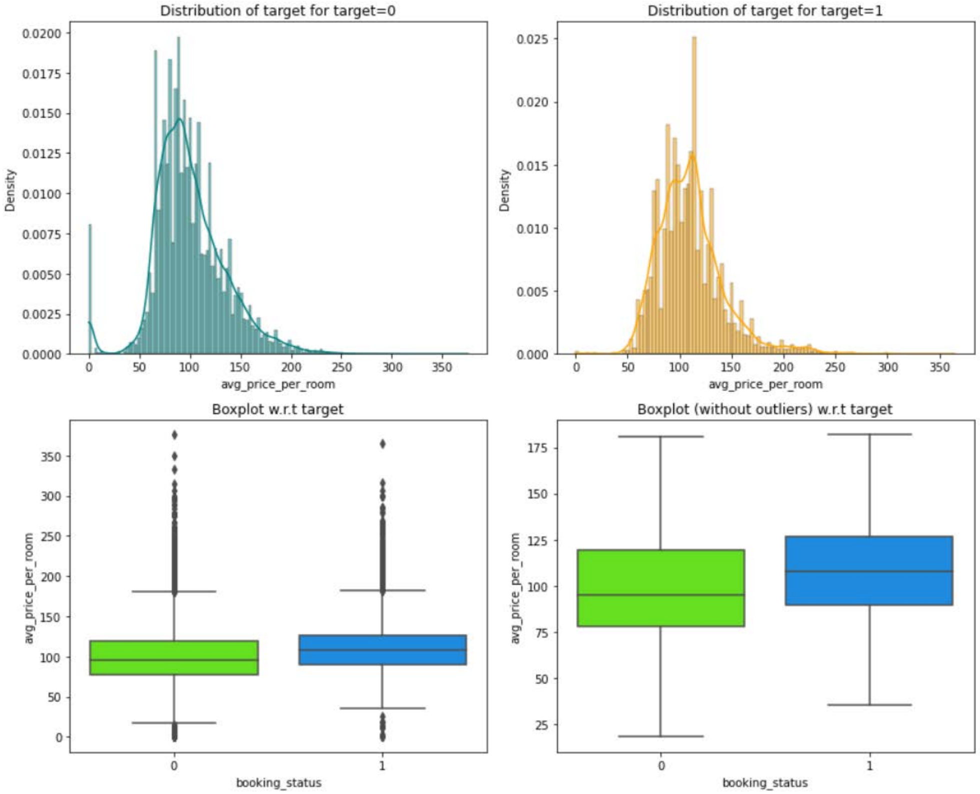


- Those who canceled reservation had the most requests



- Average price per room doesn't seem to have much impact on the number of special requests.
- However, it appears that those who paid more did have more special requests.

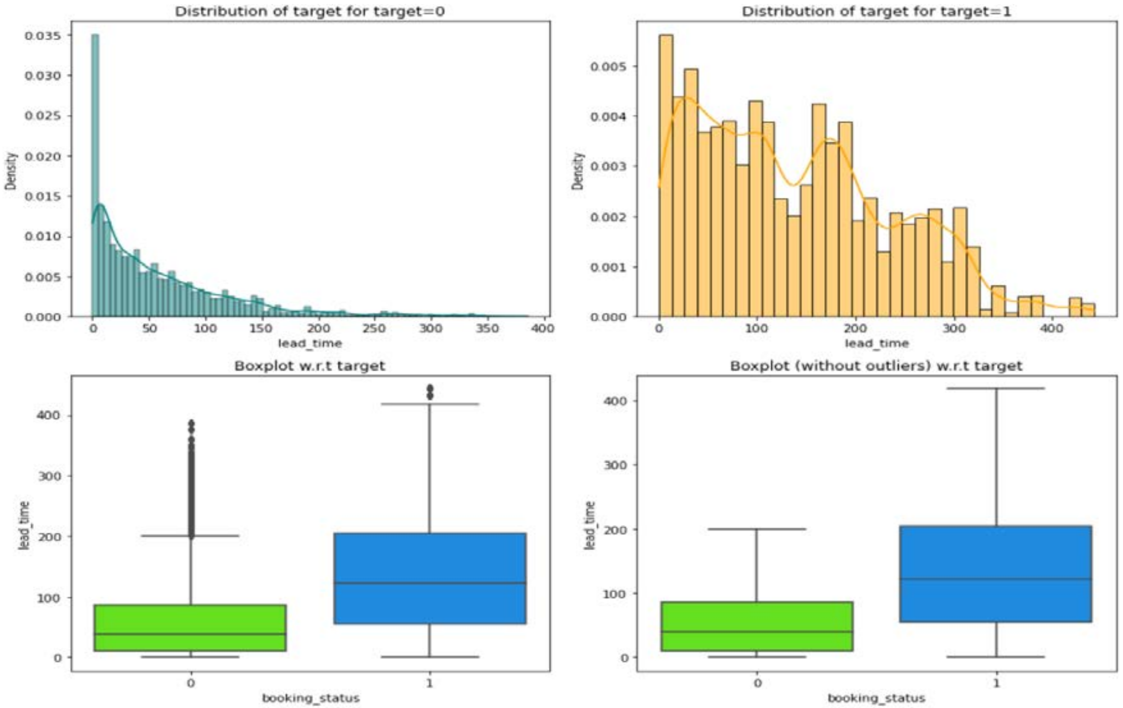
EDA Results



- We saw earlier that there is a positive correlation between booking status and average price per room.

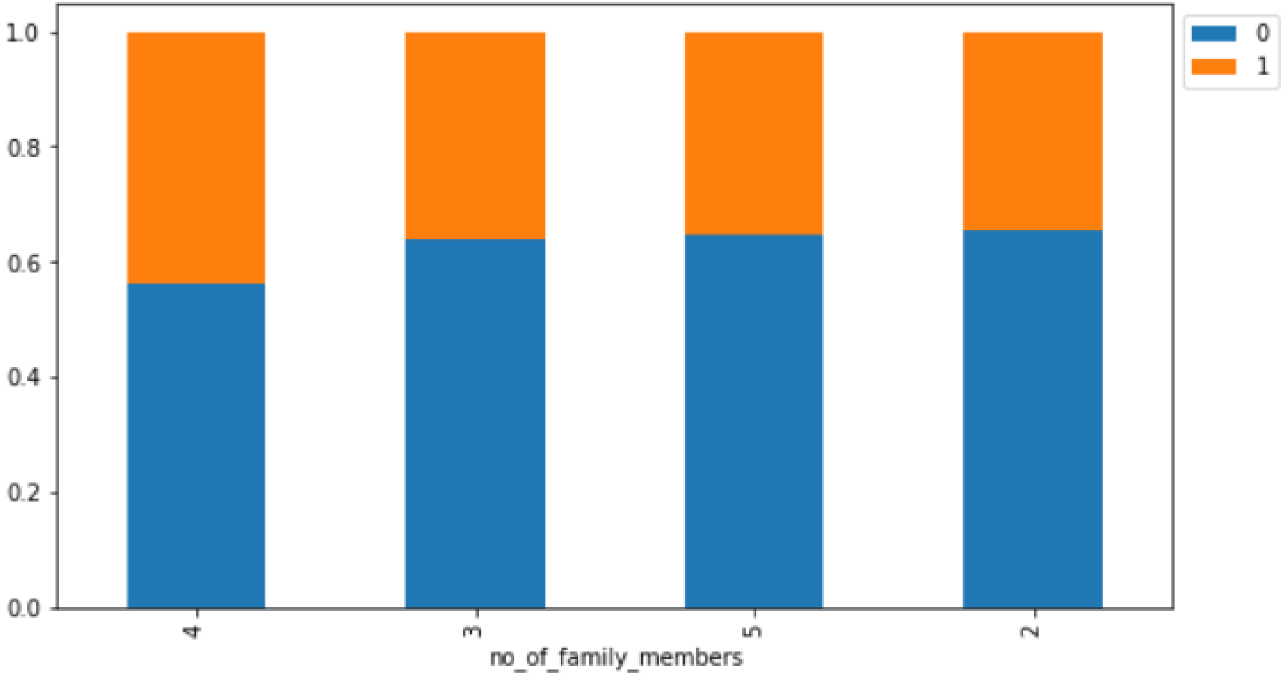
- Again, appears that those who paid less or not at all were more inclined to cancel booking

EDA Results



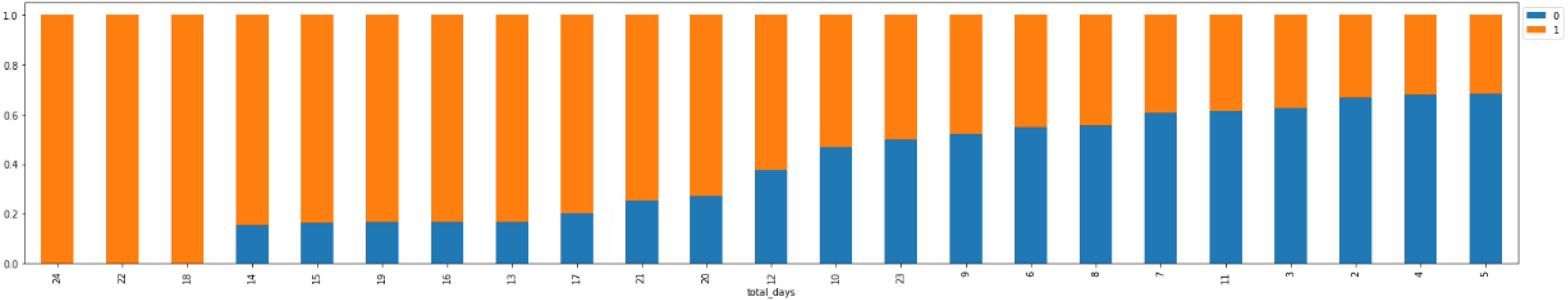
- There is a positive correlation between booking status and lead time
- People who canceled had less lead time

EDA Results

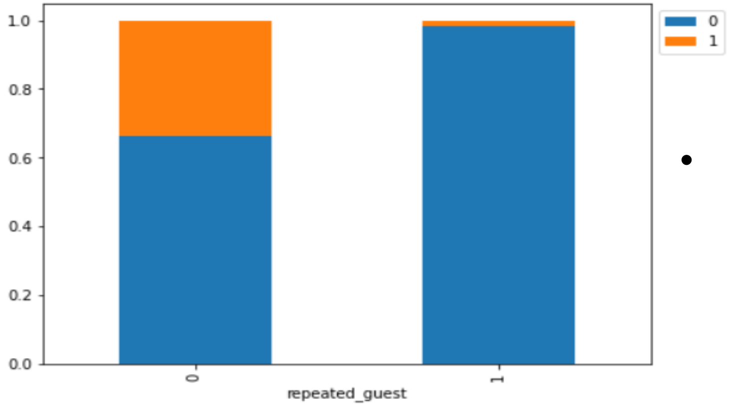


- There isn't much difference in the number of family members who canceled booking

EDA Results

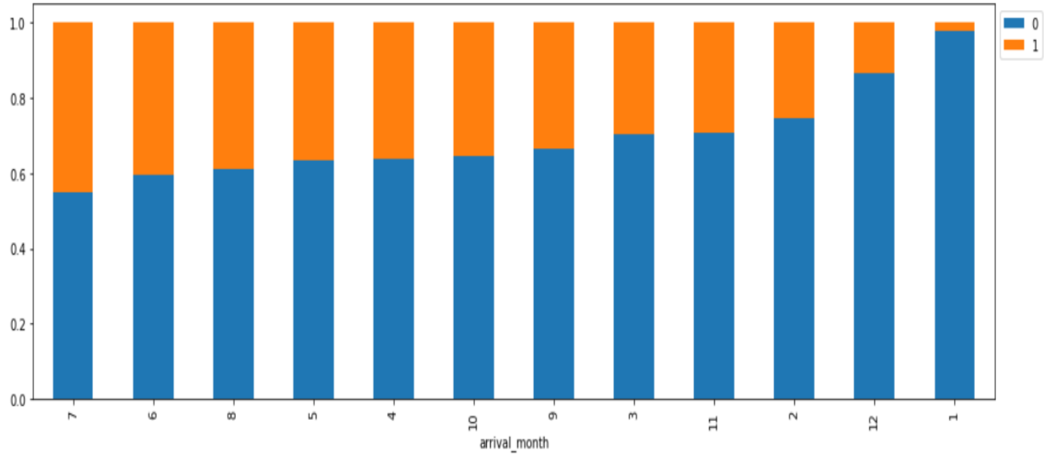


- Majority of those who cancelled booked for less amount of days

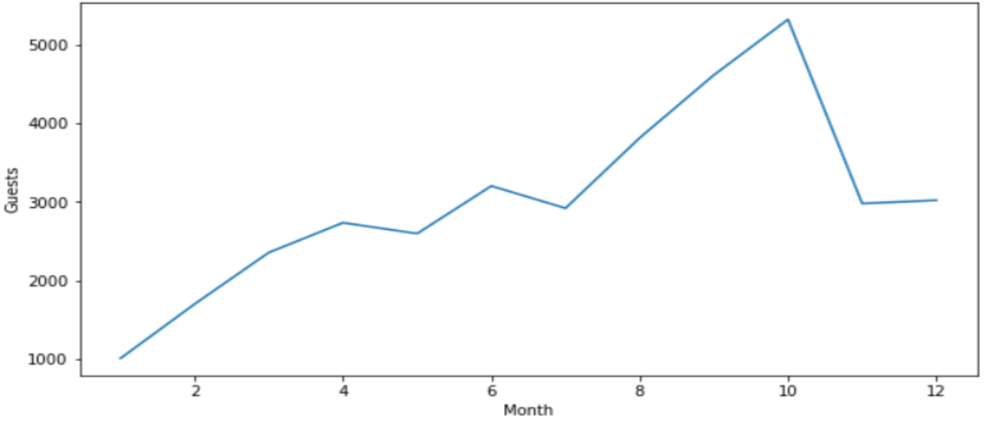


- Majority of those who canceled were repeat guests

EDA Results

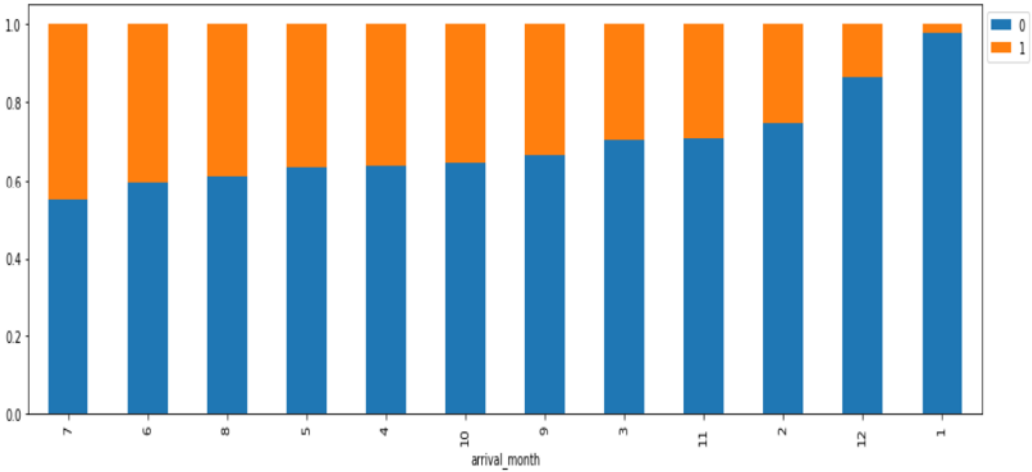


- Majority of those who canceled booked for less amount of days

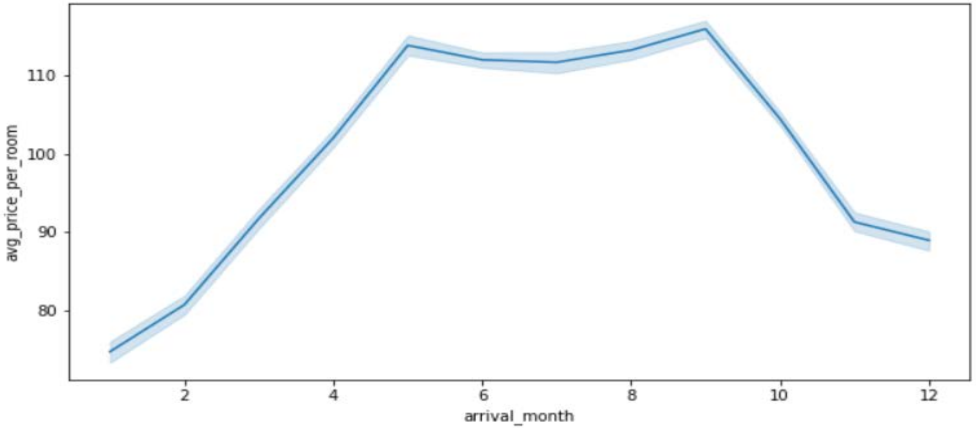


- Majority of those who canceled were repeat guests

EDA Results



- Majority of cancellations were in December and January.



- The prices are higher in the summer and lower in spring and fall.

EDA Observations:

- Average prices are higher for online guests
- Majority of cancellations come from complimentary booking and corporations is second
- The reservations with the most special requests are those that are canceled the most
- The people with the greatest number of special requests paid a little more for the room.
- Guest who paid less / complimentary were more inclined to cancel reservation
- Those who canceled had less lead time ~ 75% quartile was less than 100 days
- People with children were still more likely to cancel except for those with 4 children
- People who cancelled stayed for shorter period
- A repeat guest cancelled more than first time guests
- Busiest time of the year was October.
- Prices were more expensive in the summer than rest of the year
- Majority of free rooms were complimentary and came from online.

Data Preprocessing:

- Dropped booking ID
- Converted object to category data types
- Created a data frame that include number of family members (adults + children)
- Created a data frame to include total days including weekdays and weekends
- Number of weeknights, lead time, number of previous cancellations, number of previous bookings not canceled, and avg. price per room has the most outliers. Will not treat them to preserve integrity of the data.
- Increased upper whisker for avg. price for room to account for outlier over $500

Model Building (Logistics Regression):

- Performed logistics regression using statmodels; will identify significant predictors based upon p-values.
- The following predictors were dropped based upon p-values:

- ❑ arrival_date
- ❑ no_of_previous_bookings_not_canceled
- ❑ type_of_meal_plan_Meal Plan 3
- ❑ room_type_reserved_Room_Type 3
- ❑ market_segment_type_Complementary
- ❑ market_segment_type_Online

For collinearity when attempted to drop variables based upon VIF, didn't make a significant impact on model and causes other values to inflate. As result, just dropped predictors based on p-values greater than 0.05

Model Building (Logistics Regression):

```
                           Logit Regression Results
==============================================================================
Dep. Variable:         booking_status   No. Observations:                25392
Model:                          Logit   Df Residuals:                    25370
Method:                           MLE   Df Model:                           21
Date:                Tue, 26 Apr 2022   Pseudo R-squ.:                  0.3282
Time:                        17:08:36   Log-Likelihood:                -10810.
converged:                       True   LL-Null:                       -16091.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==================================================================================================
                                     coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
const                           -915.6391    120.471     -7.600      0.000   -1151.758    -679.520
no_of_adults                       0.1088      0.037      2.914      0.004       0.036       0.182
no_of_children                     0.1531      0.062      2.470      0.014       0.032       0.275
no_of_weekend_nights               0.1086      0.020      5.498      0.000       0.070       0.147
no_of_week_nights                  0.0417      0.012      3.399      0.001       0.018       0.066
required_car_parking_space        -1.5947      0.138    -11.564      0.000      -1.865      -1.324
lead_time                          0.0157      0.000     59.213      0.000       0.015       0.016
arrival_year                       0.4523      0.060      7.576      0.000       0.335       0.569
arrival_month                     -0.0425      0.006     -6.591      0.000      -0.055      -0.030
repeated_guest                    -2.7367      0.557     -4.916      0.000      -3.828      -1.646
no_of_previous_cancellations       0.2288      0.077      2.983      0.003       0.078       0.379
avg_price_per_room                 0.0192      0.001     26.336      0.000       0.018       0.021
no_of_special_requests            -1.4698      0.030    -48.884      0.000      -1.529      -1.411
type_of_meal_plan_Meal Plan 2      0.1642      0.067      2.469      0.014       0.034       0.295
type_of_meal_plan_Not Selected     0.2860      0.053      5.406      0.000       0.182       0.390
room_type_reserved_Room_Type 2    -0.3552      0.131     -2.709      0.007      -0.612      -0.098
room_type_reserved_Room_Type 4    -0.2828      0.053     -5.330      0.000      -0.387      -0.179
room_type_reserved_Room_Type 5    -0.7364      0.208     -3.535      0.000      -1.145      -0.328
room_type_reserved_Room_Type 6    -0.9682      0.151     -6.403      0.000      -1.265      -0.672
room_type_reserved_Room_Type 7    -1.4343      0.293     -4.892      0.000      -2.009      -0.860
market_segment_type_Corporate     -0.7913      0.103     -7.692      0.000      -0.993      -0.590
market_segment_type_Offline       -1.7854      0.052    -34.363      0.000      -1.887      -1.684
==================================================================================================
```

	Accuracy	Recall	Precision	F1
0	0.80545	0.63267	0.73907	0.68174

Model Building (Logistics Regression):

Interpreting Coefficients:

-Coefficient of **no_of_adults, no_of_children, no_of_weekend_nights, no_of_week_nights , lead_time, arrival_year, no_of_previous_cancellations, avg_price_per_room, type_of_meal_plan_Meal Plan 2, type_of_meal_plan_Not Selected,** are positive and an increase in these will lead to an increase in chances of a person cancelling booking.

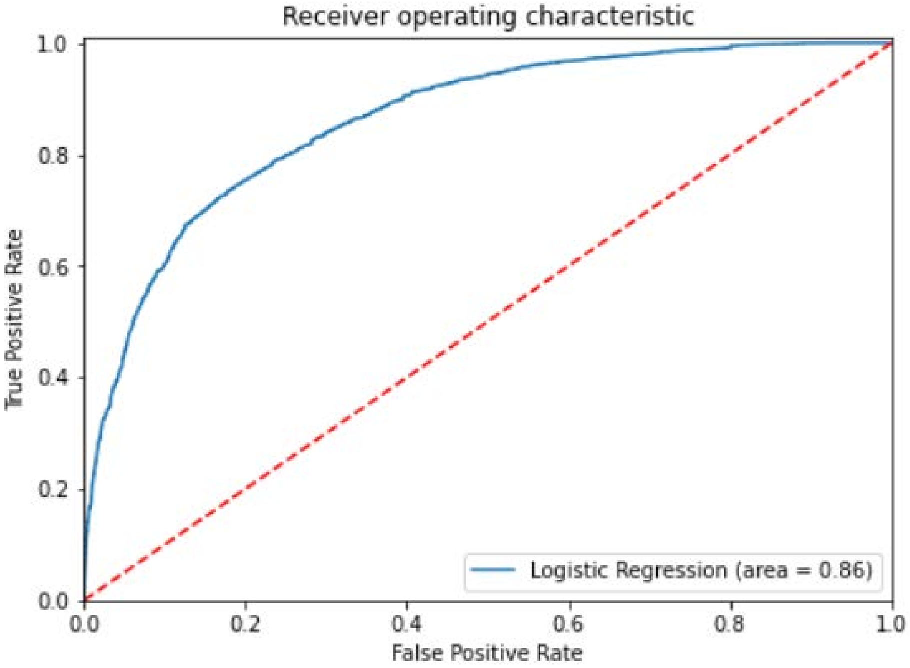
-Coefficient of **required_car_parking_space, arrival_month, repeated_guest, no_of_special_requests, room_type_reserved_Room_Type 2, 4, 5, 6, 7, market_segment_type_Corporate, market_segment_type_Offline** are a negative increase in these will lead to a decrease in chances of a person cancelling booking

When converting coefficients to odds:

- **arrival_year:** Holding all other features constant a 1 unit change in no_of_adults will increase the odds of a booking cancellation by 0.97 times or a 11.49% decrease in odds of a booking cancellation.
- **no_of_children:** Holding all other features constant a 1 unit change in the no_of_children will increase the odds of a booking being cancelled by 1.16 times or an increase of 16.5% decrease in odds of cancellation.
- **type_of_meal_plan_Not Selected:** Holding all other features constant a 1 unit change in the type_of_meal_plan_Not Selected will increase the odds of a booking being cancelled by 1.33 times or an increase of 33.1% in odds of a cancellation.
- **market_segment_type_Offline:** Holding all other features constant a 1 unit change in market_segment_type_Offline will decrease the odds of a booking cancellation by 0.16 times or 83.22% decrease in odds of a booking cancellation. Interpretation for other attributes can be done similarly.

Model Building (Logistics Regression):

Next, we use ROC-AUC on training set in attempt to improve F1 score:

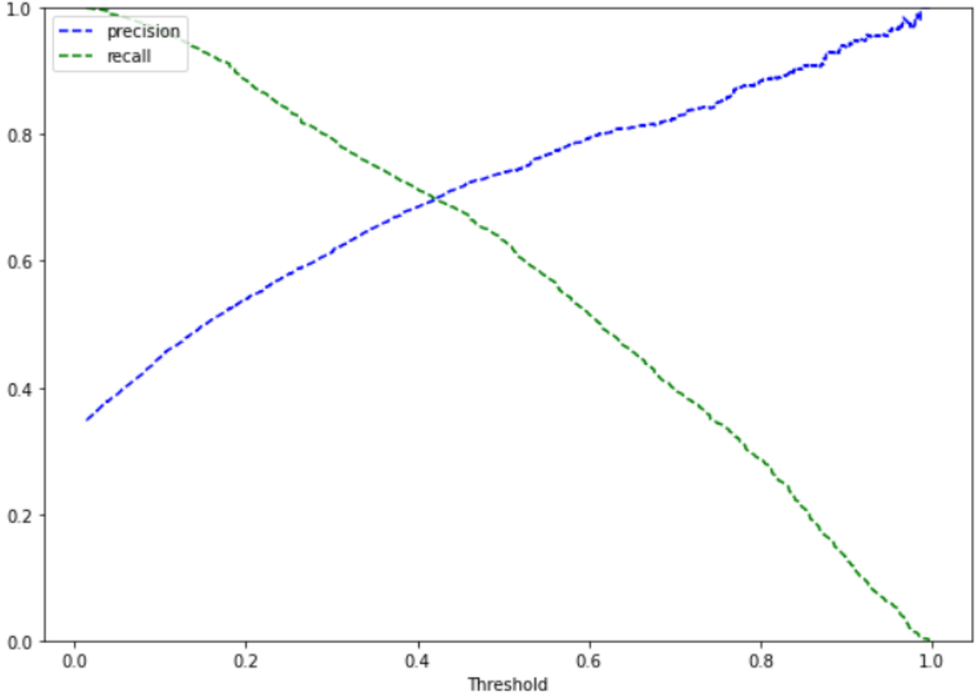


- Optimal threshold using AUC-ROC curve is 0.37
- This increases F1 score from 0.68 to 0.70

	Accuracy	Recall	Precision	F1
0	0.79265	0.73622	0.66808	0.70049

Model Building (Logistics Regression):

Next, we use Precision-Recall curve and see if we can find a better threshold:

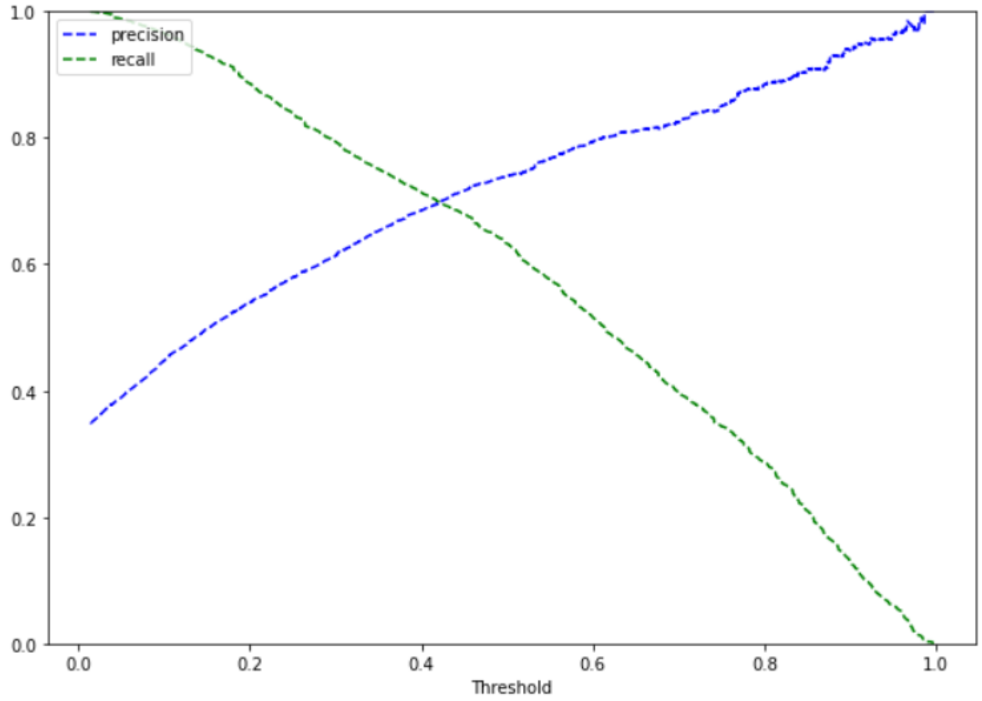


- Optimal threshold using AUC-ROC curve is 0.42
- This decreases F1 score from 0.70 to 0.69

	Accuracy	Recall	Precision	F1
0	0.80132	0.69939	0.69797	0.69868

Model Building (Logistics Regression):

Next, we use ROC curve on test set at threshold 0.37 & 0.42



0.37 Threshold

	Accuracy	Recall	Precision	F1
0	0.79555	0.73964	0.66573	0.70074

0.42 Threshold

	Accuracy	Recall	Precision	F1
0	0.80345	0.70358	0.69353	0.69852

Model Building (Logistics Regression):

Training & Testing Set Performance Comparison

Training performance comparison:

	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 comparison:

	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

- We have been able to build a predictive model that can be used by the hotel to find what factors contribute to cancelations with an F1_score of 0.70 on the training set and can formulate policies accordingly.

Model Building (Decision Tree):

- Performed model using Decision Tree Classifier
- Initial Training performance is 0.99% which is overfitting
- Initial Testing performance is 0.80%
- Will need to prune model

	Accuracy	Recall	Precision	F1
0	0.99421	0.98661	0.99578	0.99117

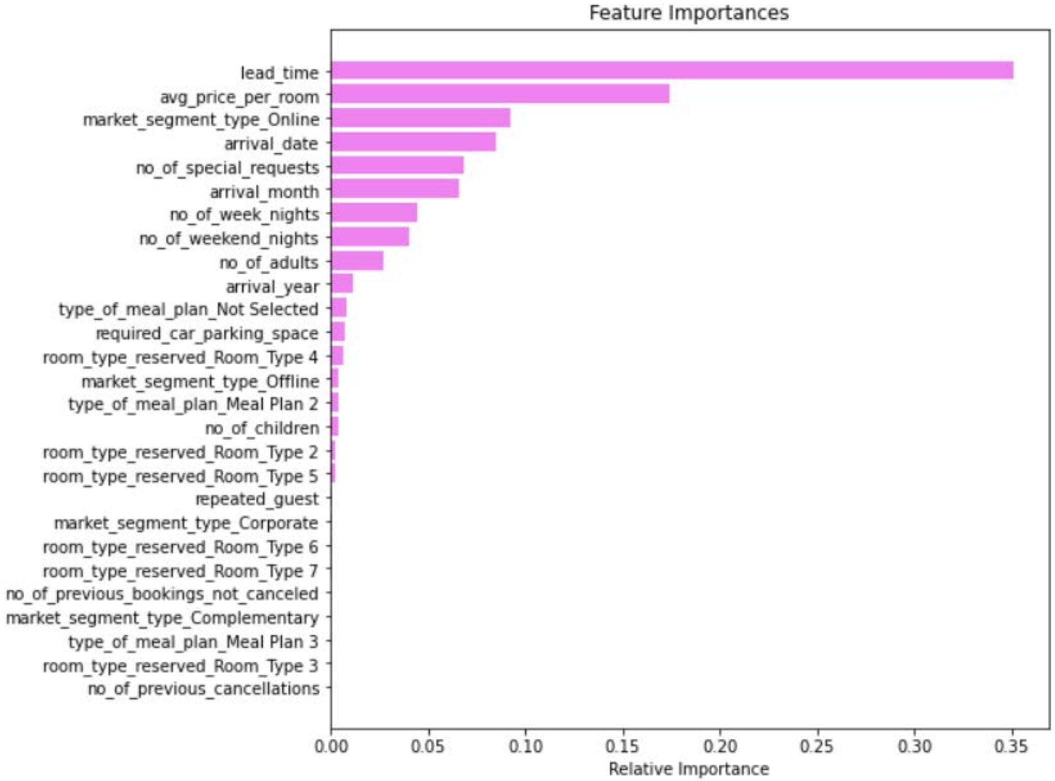
Training

	Accuracy	Recall	Precision	F1
0	0.87118	0.81175	0.79461	0.80309

Testing

Model Building (Decision Tree):

- Prior to pruning we checked out the important features:



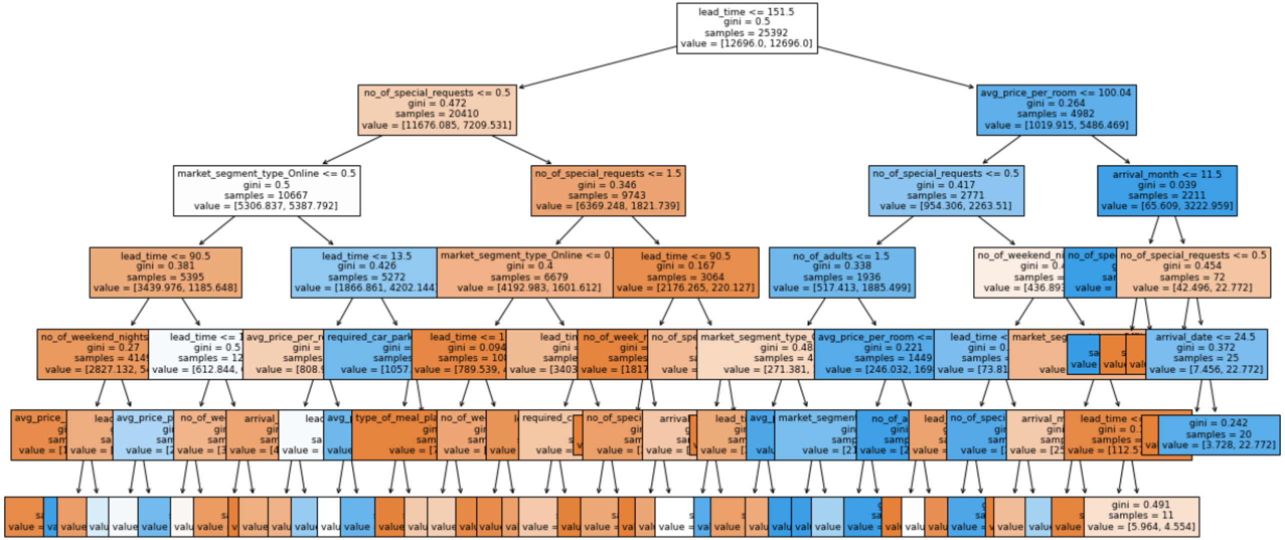
Important Features are:

- lead_time
- avg_price_per_room
- market_segment_type_Online
- arrival_date
- no_of_special_requests

Model Building (Decision Tree):

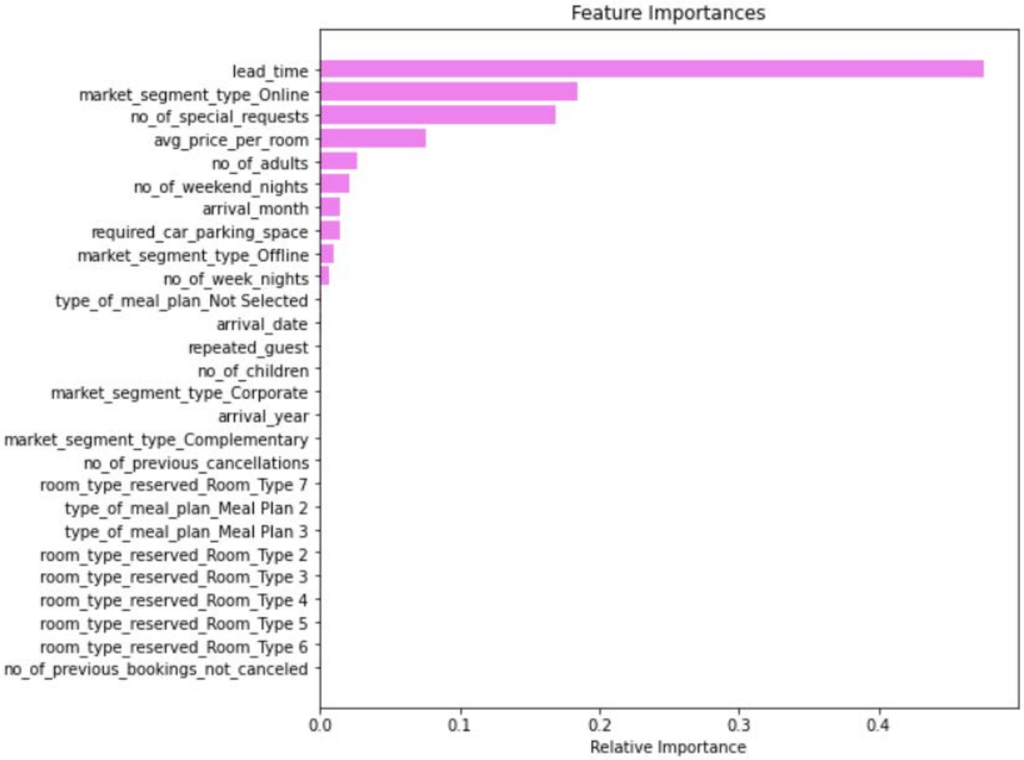
- Pre-Pruning using the below parameters:

```
DecisionTreeClassifier(class_weight='balanced', max_depth=6, max_leaf_nodes=50,
                       min_samples_split=10, random_state=1)
```



Model Building (Decision Tree):

- Important features after pre-pruning:

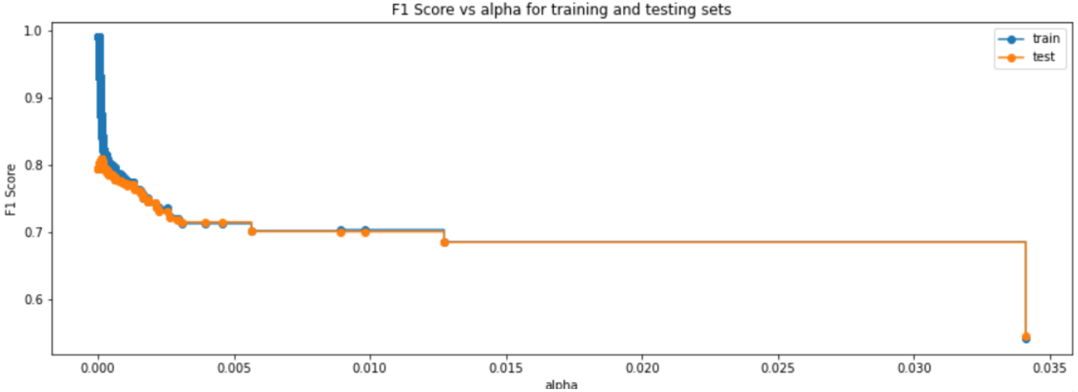


Important Features are:

- lead_time
- market_segment_type_Online
- no_of_special_requests
- avg_price_per_room
- no_of_weekend_nights

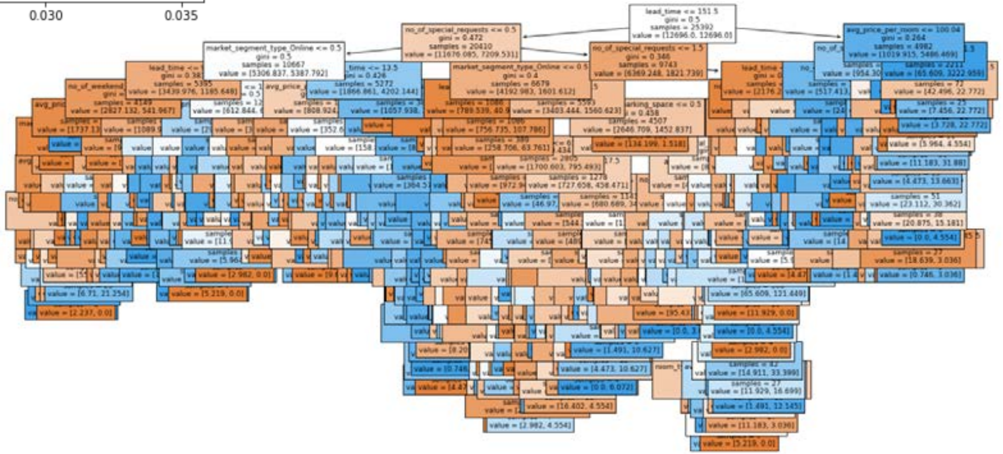
Model Building (Decision Tree):

- Post-Pruning using ccp_alphas:



Based upon pruned decision tree if the lead time is less than 16 days and the average price per room is less than $68.50 then there is very good chance that guest will cancel booking

```
|--- lead_time <= 151.50
|   |--- no_of_special_requests <= 0.50
|   |   |--- market_segment_type_Online <= 0.50
|   |   |   |--- lead_time <= 90.50
|   |   |   |   |--- no_of_weekend_nights <= 0.50
|   |   |   |   |   |--- avg_price_per_room <= 196.50
|   |   |   |   |   |   |--- market_segment_type_Offline <= 0.50
|   |   |   |   |   |   |   |--- lead_time <= 16.50
|   |   |   |   |   |   |   |   |--- avg_price_per_room <= 68.50
|   |   |   |   |   |   |   |   |   |--- weights: [207.26, 10.63] class: 0
|   |   |   |   |   |   |   |   |--- avg_price_per_room >  68.50
|   |   |   |   |   |   |   |   |   |--- arrival_date <= 29.50
|   |   |   |   |   |   |   |   |   |   |--- no_of_adults <= 1.50
|   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 2
|   |   |   |   |   |   |   |   |   |   |--- no_of_adults >  1.50
|   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 5
|   |   |   |   |   |   |   |   |   |--- arrival_date >  29.50
|   |   |   |   |   |   |   |   |   |   |--- weights: [2.24, 7.59] class: 1
|   |   |   |   |   |   |   |--- lead_time >  16.50
```



Model Building (Decision Tree):

- Training & Testing Model Summary

```
Training performance comparison:
```

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.99421	0.99421	0.89989
Recall	0.98661	0.98661	0.90303
Precision	0.99578	0.99578	0.81353
F1	0.99117	0.99117	0.85594

```
Test performance comparison:
```

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.87118	0.87118	0.86888
Recall	0.81175	0.81175	0.85576
Precision	0.79461	0.79461	0.76634
F1	0.80309	0.80309	0.80858

Training model is not overfitting now it has been pruned. F1 score is 0.85 which is a pretty good model.

Testing model F1 score is 0.80 which is pretty good as well.

CONCLUSIONS

- We have been able to build a predictive model that can be used by the hotel to find the guests who will cancel with an F1 score of 0.70 on the training set and can formulate policies accordingly.
- Coefficient of no_of_adults, no_of_children, no_of_weekend_nights, no_of_week_nights, lead_time, arrival_year, no_of_previous_cancellations, avg_price_per_room, type_of_meal_plan_Meal Plan 2, type_of_meal_plan_Not Selected, are positive and an increase in these will lead to increase in the chances of a guest cancelling booking.
- Coefficient of required_car_parking_space, arrival_month, repeated_guest, no_of_special_requests, room_type_reserved_Room_Type 2, room_type_reserved_Room_Type 4, room_type_reserved_Room_Type 5, room_type_reserved_Room_Type 6, room_type_reserved_Room_Type 7, market_segment_type_Corporate, market_segment_type_Offline are negative and an increase in these will lead to a decrease in chances of a guest cancelling booking.
- Important features are lead time, online market segment, number of special requests, average price per room, and number of weekend nights.
- Decision tree model showed us that if lead time is less than 16 days and the average price per room is less than $68.50 than there is a very good chance that guest will cancel booking.

RECOMMENDATIONS

- For complimentary and corporate reservations require a deposit that is only refundable once the guests checks in.
- Revise policy on complimentary rooms as this is where a lot of cancellations are coming from. If the guests cancel more than one time in a year, then guest is no longer entitled to a complimentary room for a period.
- Reduce options for meal plans to breakfast only. Most guests choose breakfast or no meal plan. This can reduce food costs.
- Majority of guests are purchasing Room Type 1, offer discounts and/or incentives to book other room types.
- Most guests don’t use parking. There is an opportunity to evaluate parking to assess why guests are not using it and to repurpose the space.
- Majority of guests are booking through the week. Offer promotions for weekend get-a-ways.
- Most guest aren't bringing children. The hotel could incorporate more family-oriented activities to motivate parents to bring their children.
- Offer spring specials to encourage guests to attend hotel in the spring during slowest period.
- Offer loyalty points for repeats guests and additional bonus points if they don’t cancel.
- Send survey to guests asking about hotel experience and follow-up asking about cancellation reasons to find patterns.
- Allow more than 2 special requests for loyalty members and those who haven't canceled within in the last year.

THANKS!