StarHotels Presentation

Business Problem Overview and Solution Approach

A significant number of hotel bookings are called off due to cancellations or no-shows. The typical reasons for cancellations include change of plans, scheduling conflicts, etc. This is often made easier by the option to do so free of charge or preferably at a low cost which is beneficial to hotel guests but it is a less desirable and possibly revenue-diminishing factor for hotels to deal with. Such losses are particularly high on last-minute cancellations.

The new technologies involving online booking channels have dramatically changed customers' booking possibilities and behavior. This adds a further dimension to the challenge of how hotels handle cancellations, which are no longer limited to traditional booking and guest characteristics.

The cancellation of bookings impact a hotel on various fronts:

1. Loss of resources (revenue) when the hotel cannot resell the room.

2. Additional costs of distribution channels by increasing commissions or paying for publicity to help sell these rooms.

3. Lowering prices last minute, so the hotel can resell a room, resulting in reducing the profit margin.

4. Human resources to make arrangements for the guests.

Business Objective

Objective

The increasing number of cancellations calls for a Machine Learning based solution that can help in predicting which booking is likely to be canceled. Star Hotels Group has a chain of hotels in Portugal, they are facing problems with the high number of booking cancellations and have reached out to your firm for data-driven solutions. You as a data scientist have to analyze the data provided to find which factors have a high influence on booking cancellations, build a predictive model that can predict which booking is going to be canceled in advance, and help in formulating profitable policies for cancellations and refunds

Data Description

Variable	Description
no_of_adults	Number of Adults
no_of_children	Number of Children
no_of_weekend_nights	Number of weekend nights booked or stayed
no_of_week_nights	Number of week nights booked or stayed
type_of_meal_plan	Type of meal plan booked
required_car_parking_space	0-No, 1-Yes
room_type_reserved	Room type reserved (encoded by Hotel)
lead_time	Number of days between booking and arrival
arrival_year	Year of arrival
arrival_month	Month of arrival
arrival_date	Date of the month
market_segment_type	Market segment designation
repeated_guest	0-No, 1-Yes

Data Description Ctd.

Variable	Description
no_of_previous_cancellations	Number of previous bookings cancelled by the customer prior to current booking
no_of_previous_bookings_not_cancelled	Number of previous bookings not cancelled by the customer prior to current booking
avg_price_per_room	Price per room per day. Prices are dynamic (in Euros)
no_of_special_requests	Total number of special requests made by the customer
booking_status	Flag indicating if booking was cancelled or not.

- Data from hotel contains 56,926 entries
- Data has no missing values
- Data is mostly in integer/float format with a couple factors in object format

Data Overview Ctd:

- Data contains object strings, floats and integer data types.
- Data has 14,350 duplicated values that will be dropped which will leave 42,576 values remaining.

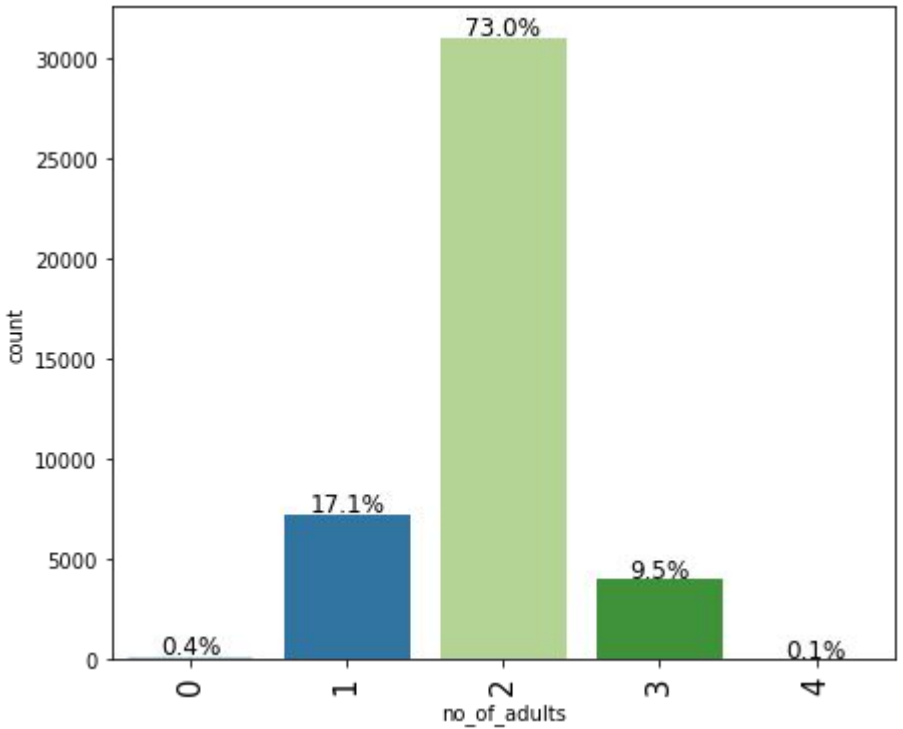
EDA: Univariate Analysis: Categorical Data

```
Meal Plan 1     31863
Not Selected     8716
Meal Plan 2      1989
Meal Plan 3         8
Name: type_of_meal_plan, dtype: int64
****************************************
Room_Type 1    29730
Room_Type 4     9369
Room_Type 6     1540
Room_Type 5      906
Room_Type 2      718
Room_Type 7      307
Room_Type 3        6
Name: room_type_reserved, dtype: int64
****************************************
Online          34169
Offline          5777
Corporate        1939
Complementary     496
Aviation          195
Name: market_segment_type, dtype: int64
****************************************
Not_Canceled    28089
Canceled        14487
Name: booking_status, dtype: int64
****************************************
```

Observations:

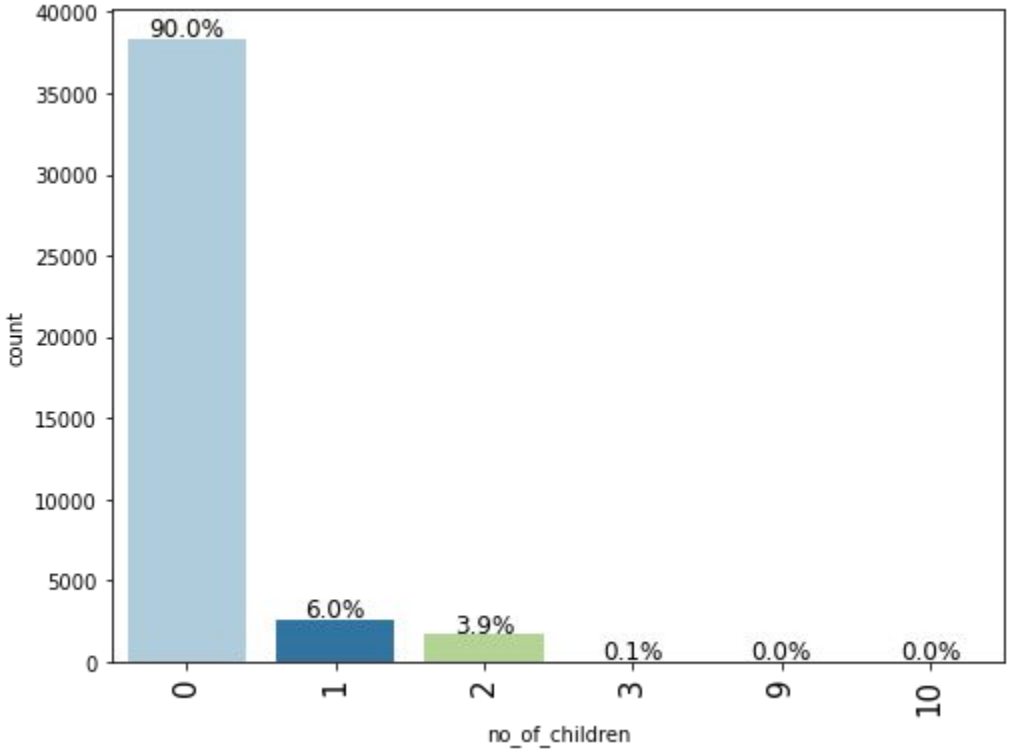
- Majority of customers purchased meal plan 1
- Room type 1 is most booked
- Online bookings make up most of the reservations
- 66% of the bookings are not cancelled.

EDA: Univariate Analysis: Number of Adults



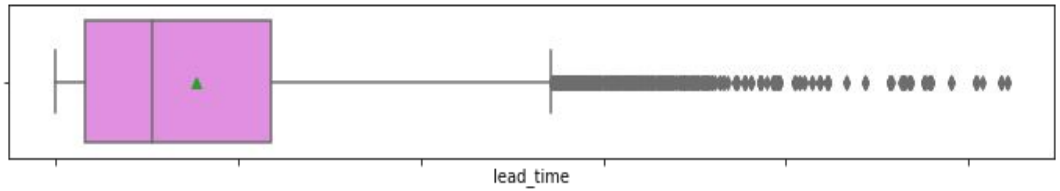
- 73% of bookings are for 2 adult guests,
- Next highest percentage is bookings of 1 adult guest at 17.1%.
- Max is 4 guests.

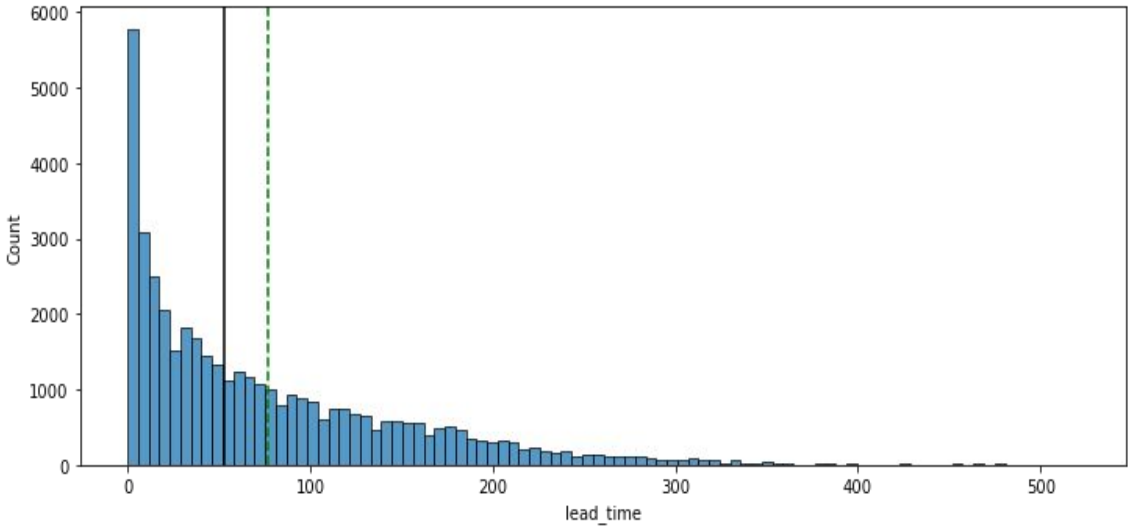
EDA: Univariate Analysis: Number of Children



- 90% of the customers have 0 children with them

EDA: Univariate Analysis: Lead Time

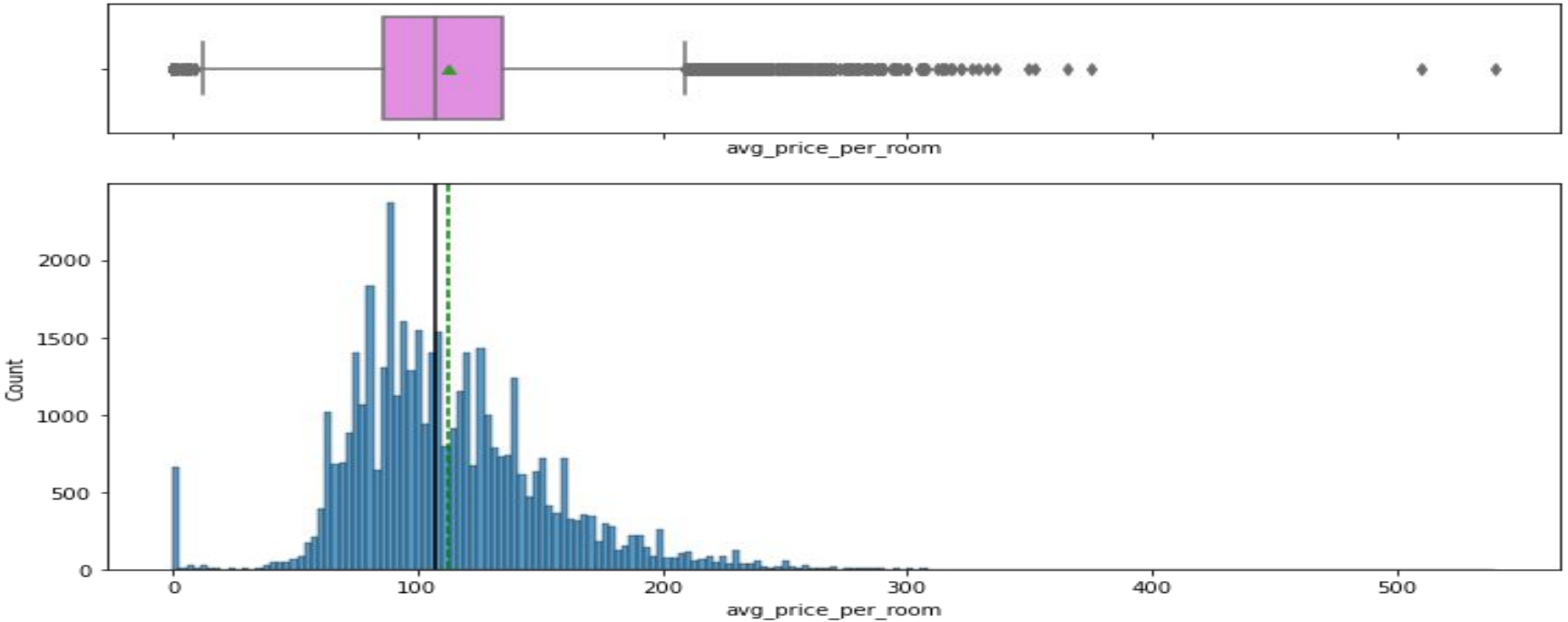




- Heavily right skewed

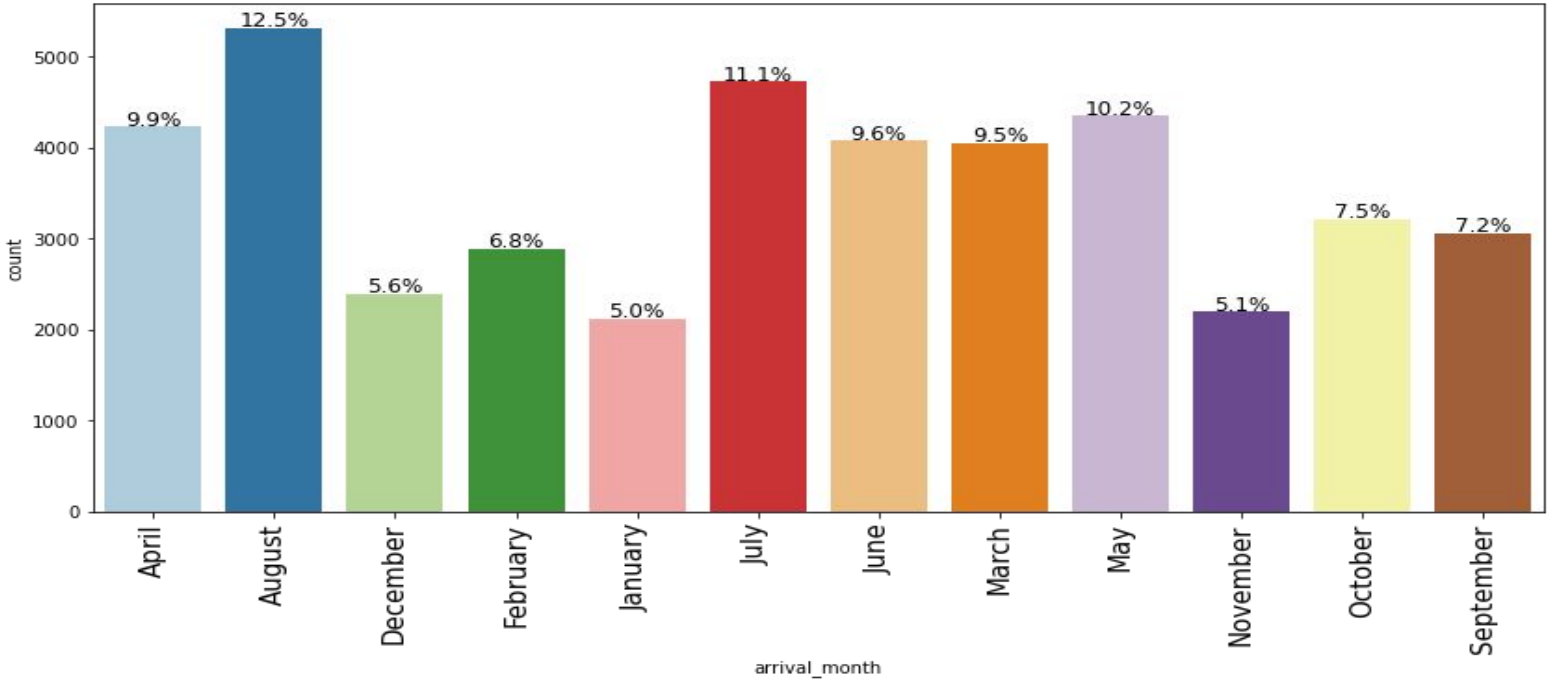
- Lead time has a very large range, from an average of 94 days to the upper limit of 500 days

EDA: Univariate Analysis: Price Per Room



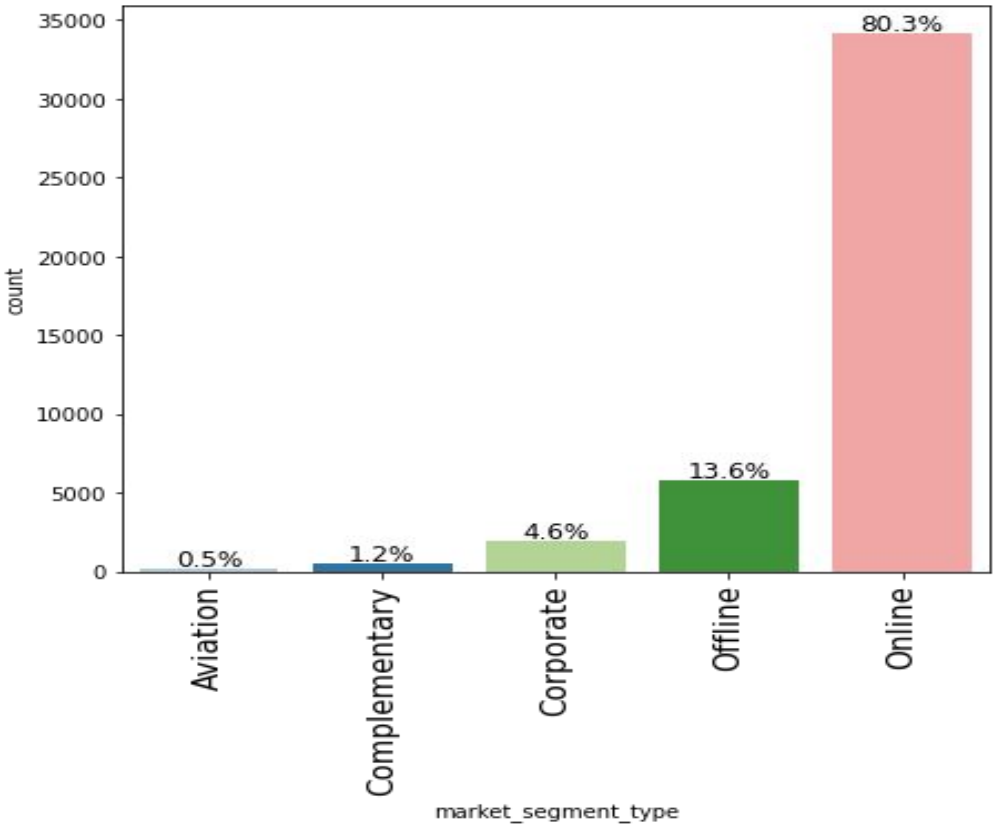
Average price per rom varies greatly and is heavily right skewed with outliers 5x the average of ~100 Euros.

EDA: Univariate Analysis: Arrival Month



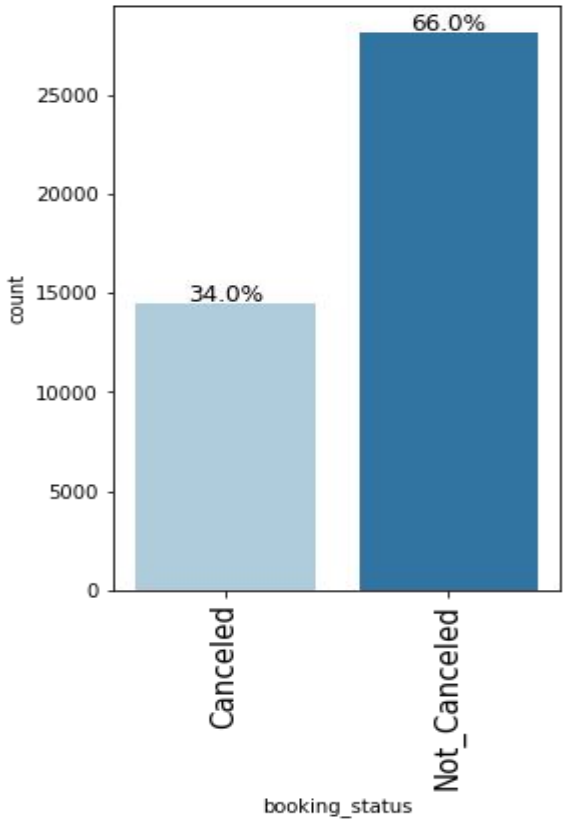
- The busiest month for bookings is August with 12.5%. Summer months (May/June/July/August) make up 43% of the reservations.

EDA: Univariate Analysis: Market Segment Type



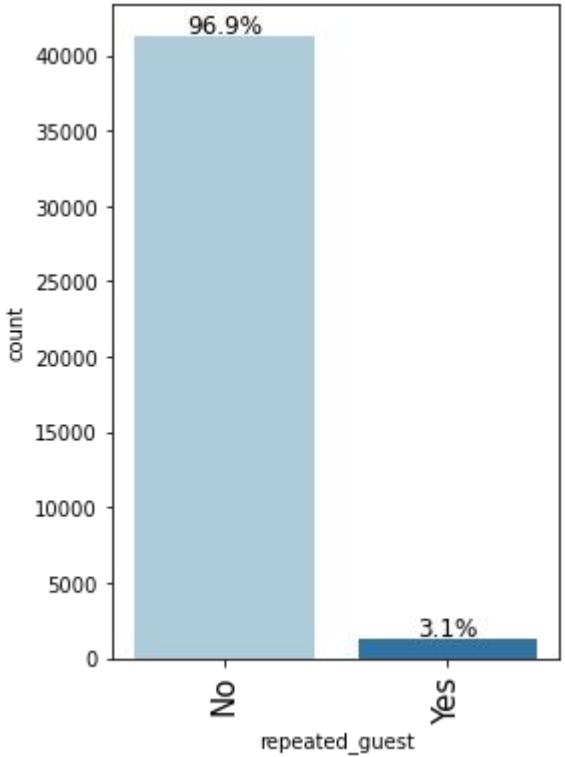
- Online bookings make up 80.3% of the reservations.
- Offline reservations are the next highest at 13.6%.

EDA: Univariate Analysis: Booking Status



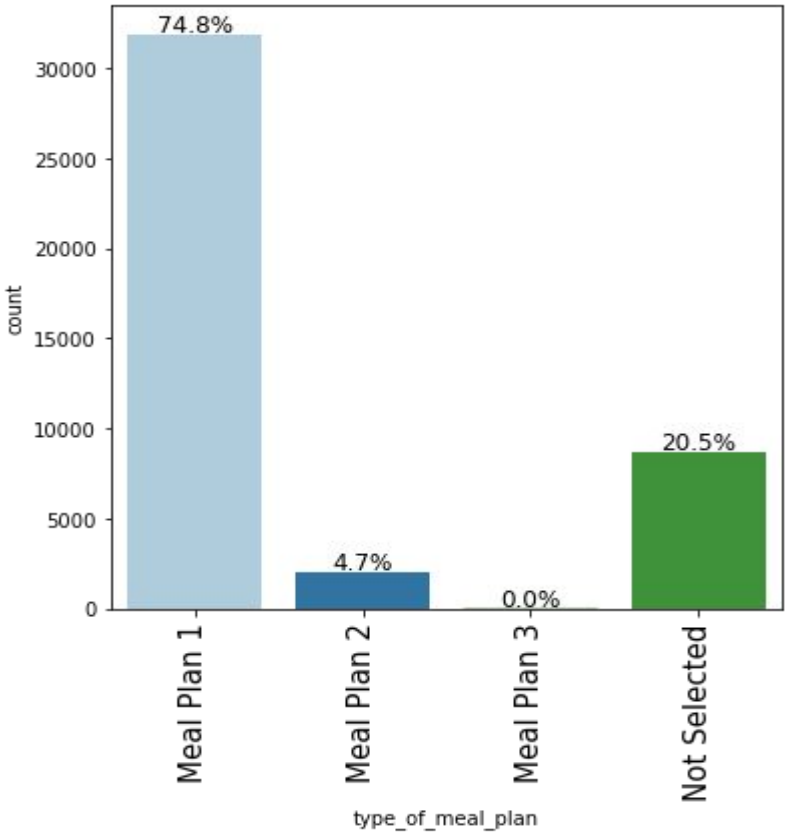
- 34.0% of the bookings are cancelled.
- 66.0% of the bookings are not cancelled.

EDA: Univariate Analysis: Repeat Guests



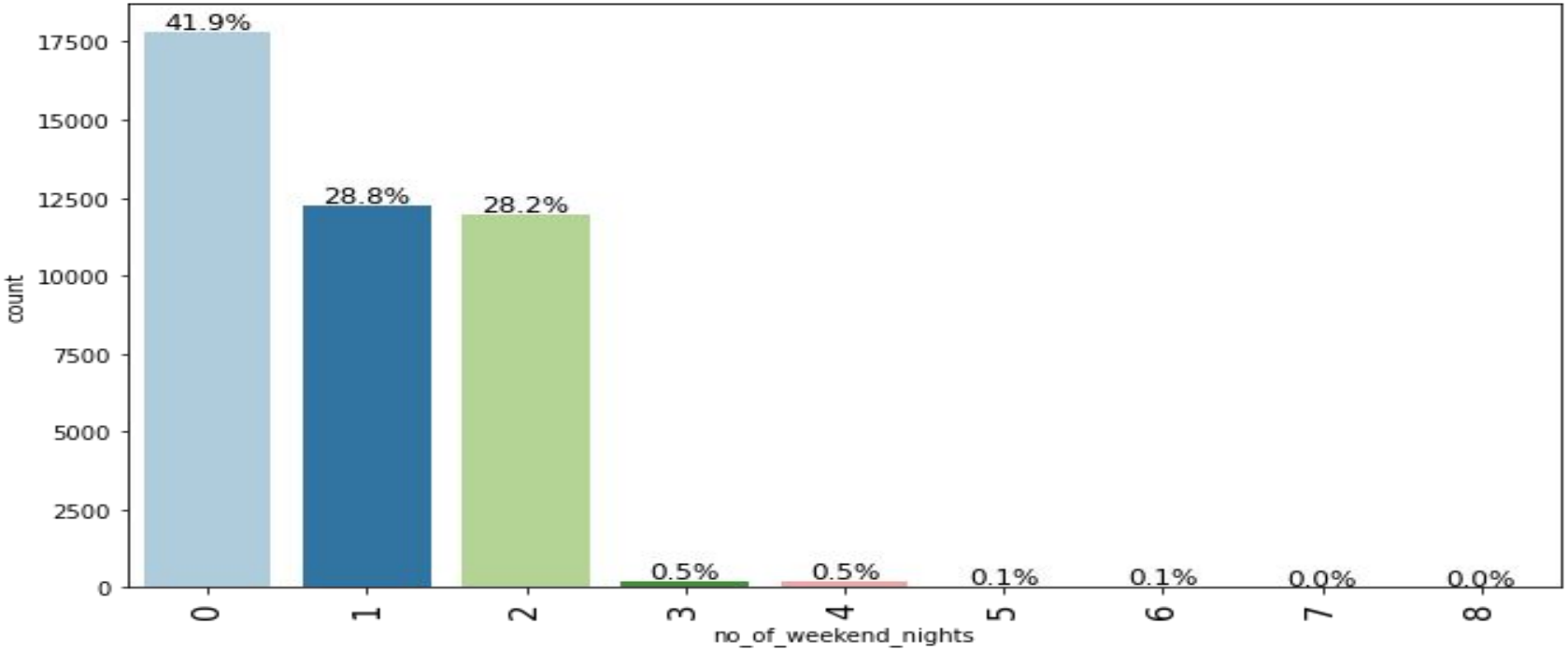
- 96.9% of the bookings are from new customers.
- Area of improvement here to increase repeat customers,

EDA: Univariate Analysis: Meal Plan



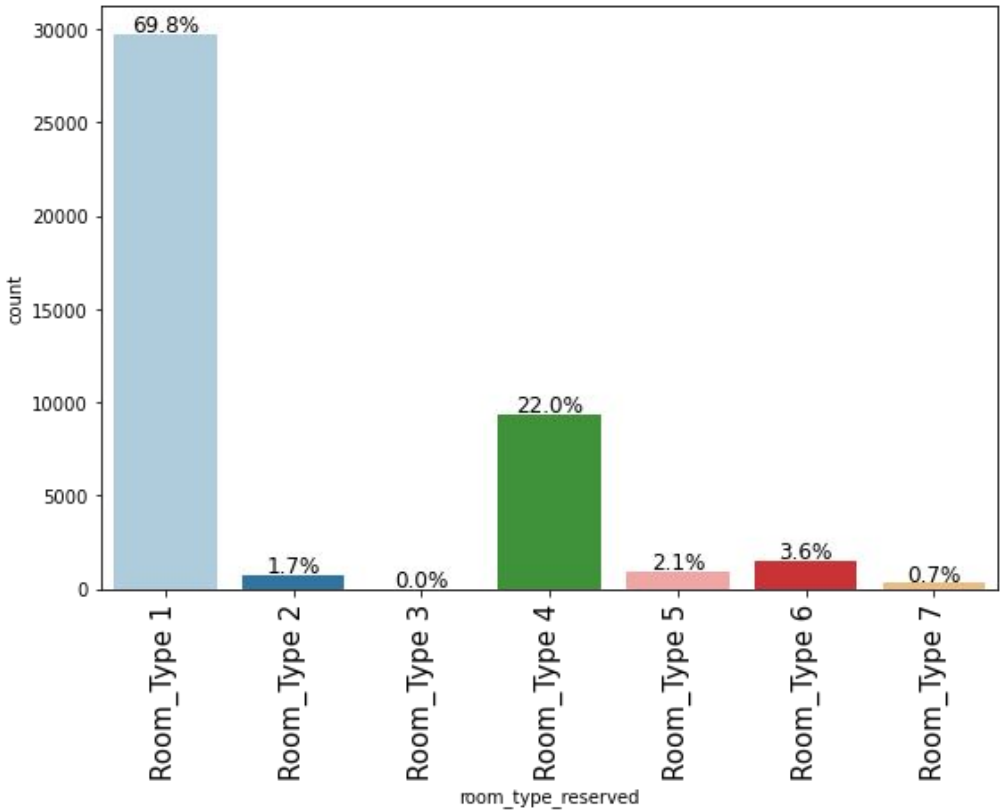
- Meal plan 1 is the most popular.
- 20.5% of customers opt for no meal plan.

EDA: Univariate Analysis: Number of Weekend Nights



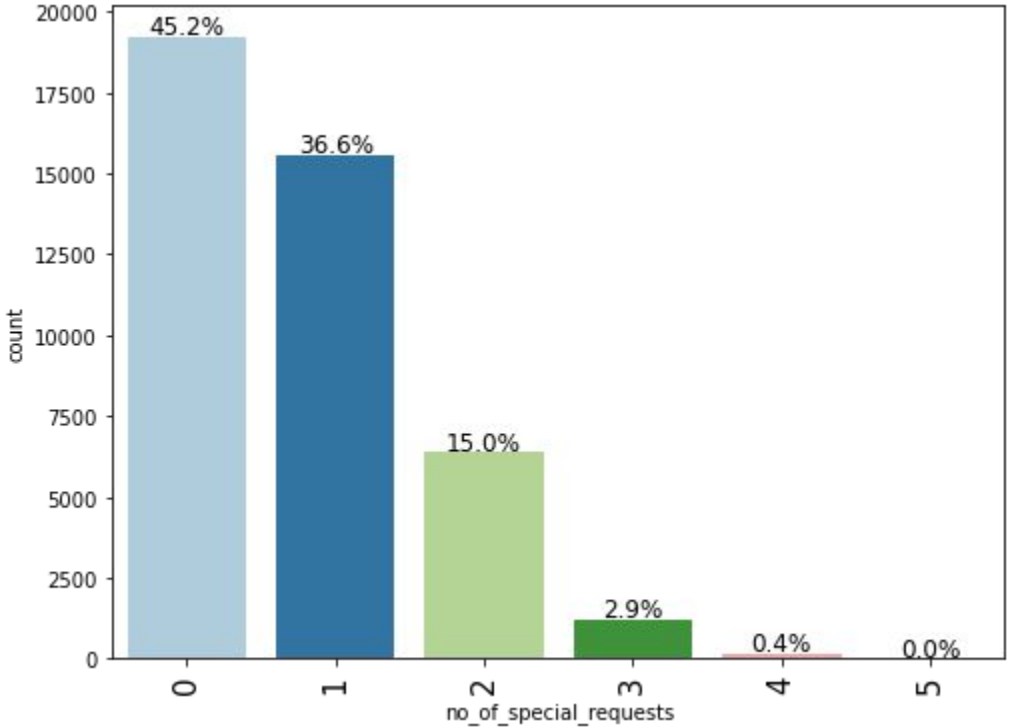
- 41.9% if customers do not book weekend nights.
- 57% of customers book weekend nights

EDA: Univariate Analysis: Room Types Reserved



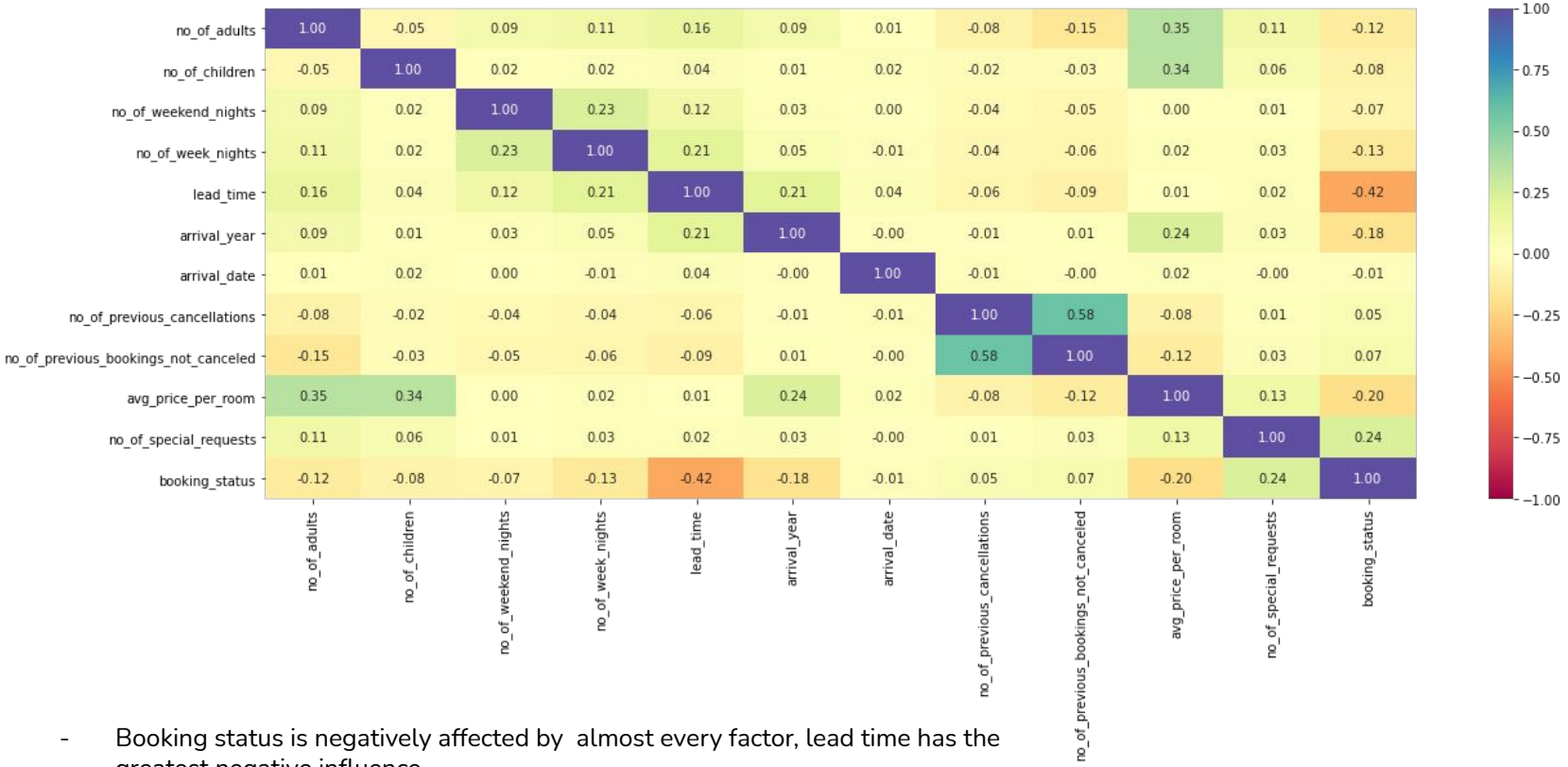
- 69.8% of customers reserve room type 1.
- 22.0% reserve room type 4

EDA: Univariate Analysis: Number of special requests



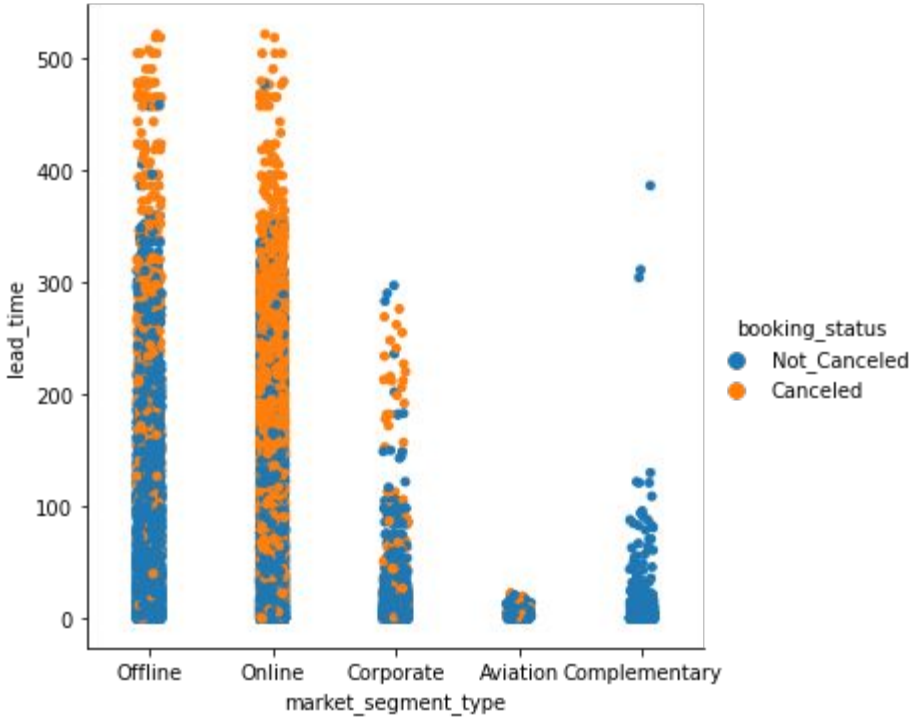
- 45.2% if customers do not need any special requests.
- 36.9% of customers ask for at least one special request.

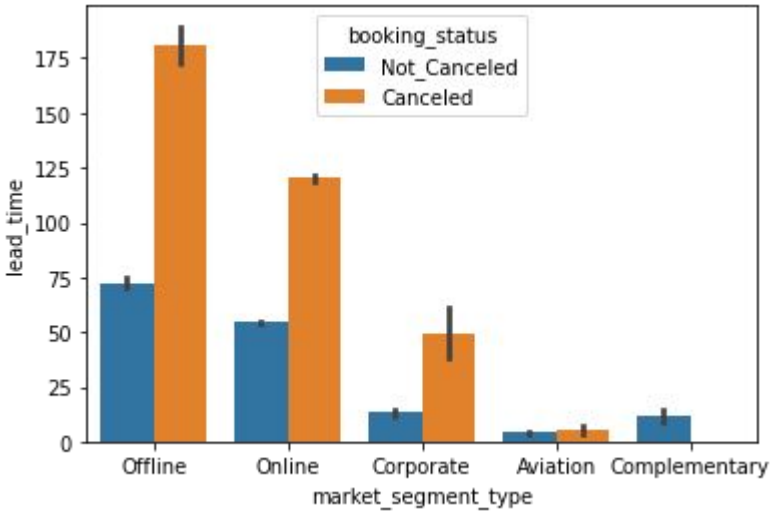
EDA: Factor Correlation Analysis



- Booking status is negatively affected by almost every factor, lead time has the greatest negative influence.

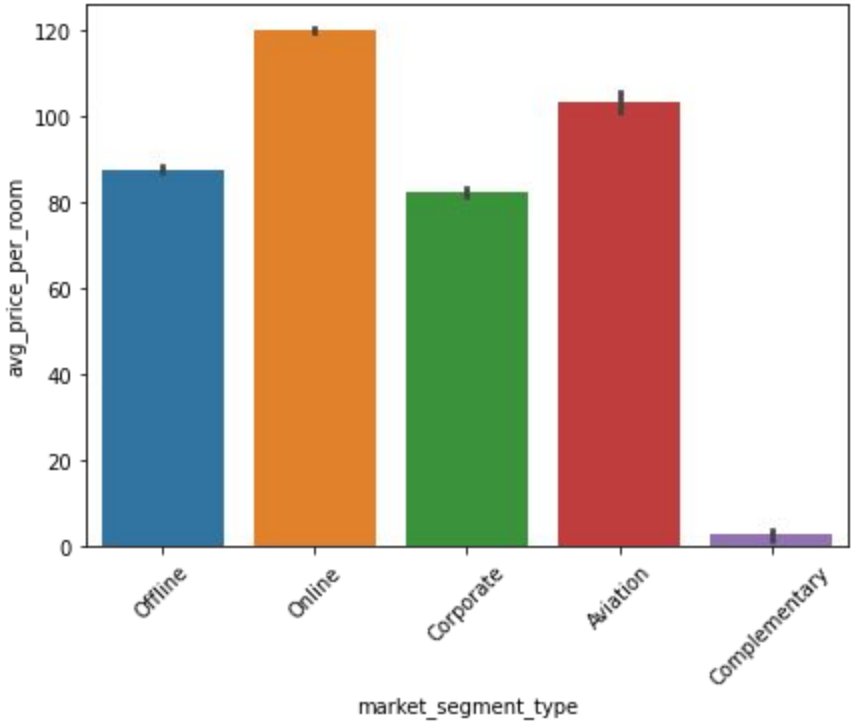
Bivariate Analysis: Lead Time vs Market Type Cancellations

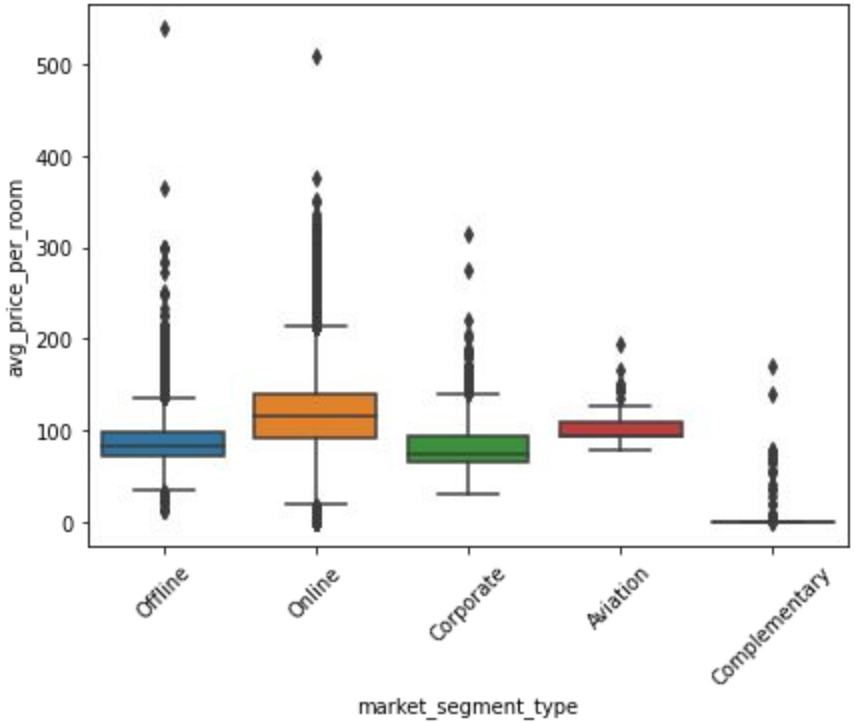




-Offline and online have by far the largest lead time of over 50 days on average.
-The longer the lead time the higher the cancellation rate.
-Complimentary has no cancellations, which makes sense as they are free to guests

Bivariate Analysis: Room Price vs Market Type





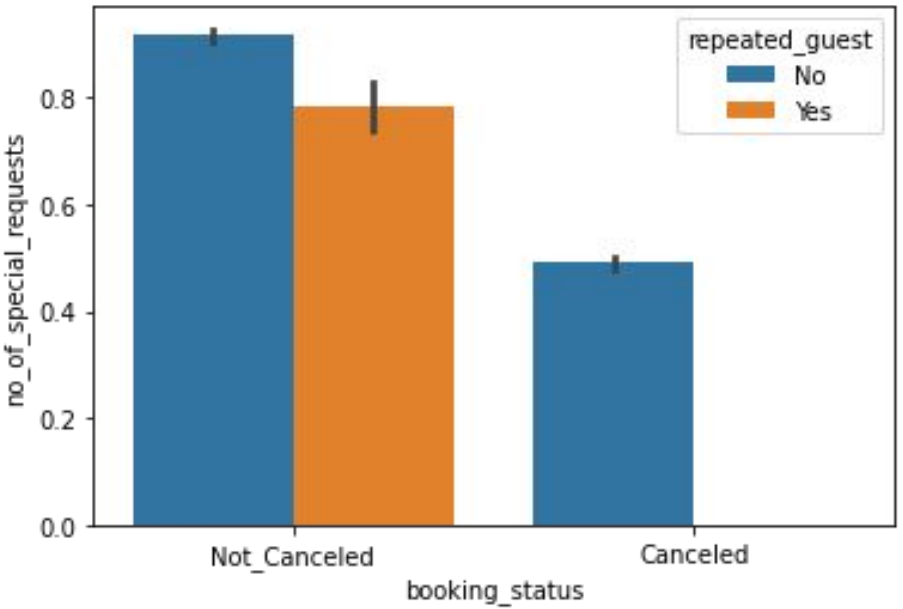
- Most market types have large ranges in room prices with many top heavy outliers.
- Online has the highest average room price: 120 Euros, with a max price of 510 Euros per room

Bivariate Analysis: Booking Status vs Repeat Guests

booking_status	repeated_guest	
Canceled	No	14477
	Yes	10
Not_Canceled	No	26784
	Yes	1305

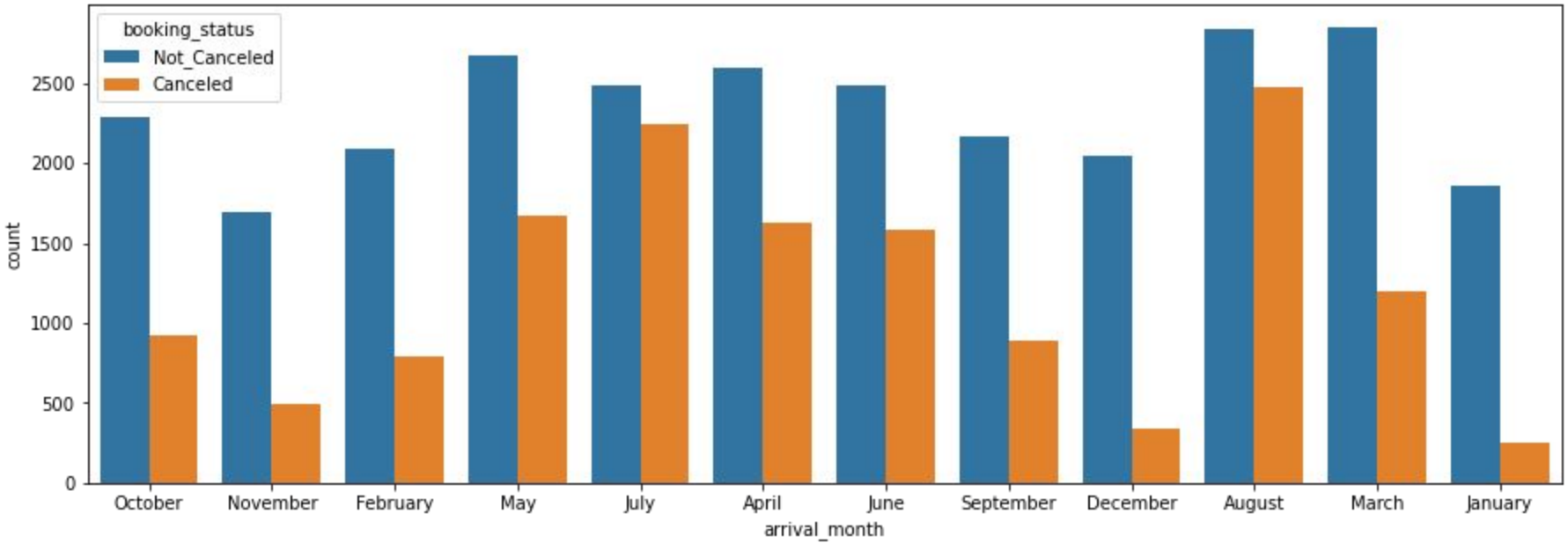
- The percentage of repeat guests that cancel is 0.76%

Bivariate Analysis: Special Requests Vs Booking Status



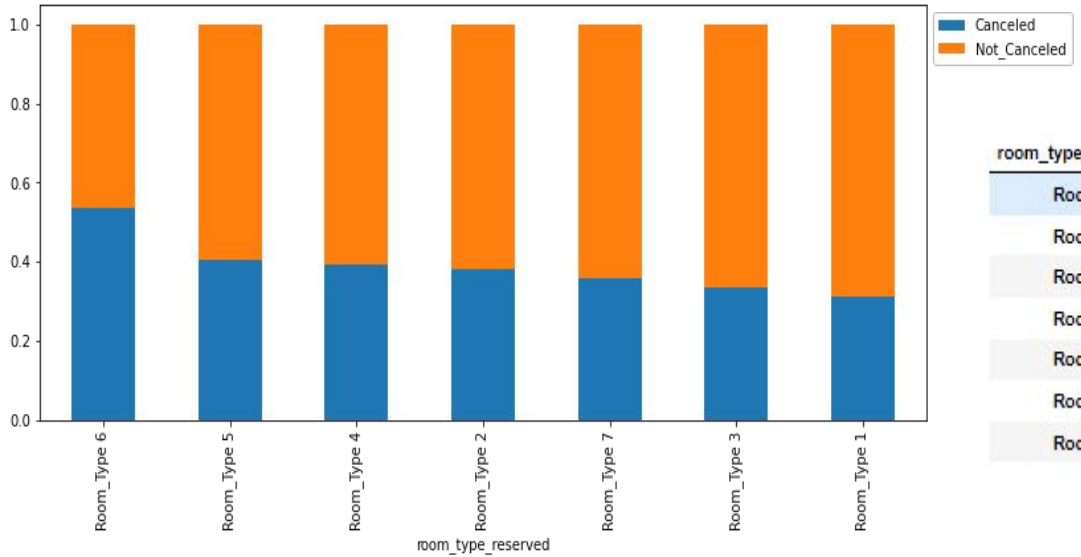
- Repeat customers do not appear to influence cancellation rates when comparing to number of special requests.
- As the number of special requests increases the rate of cancellation is lower .
- New customers do not cancel as much as repeat customers.

Bivariate Analysis: Arrival Month Vs Booking Status



- Summer months have higher cancellation rates than winter months and also an overall higher number of bookings.
- Winter months have much lower cancellation rate and lower guests reserving rooms.

Bivariate Analysis: Room Type Vs. Price per room



room_type_reserved	count	mean	std	min	25%	50%	75%	max
Room_Type 1	29730.0	100.092176	30.690012	0.0	80.0000	98.300	119.00	540.00
Room_Type 2	718.0	90.586657	35.885009	0.0	77.2500	88.630	103.05	284.10
Room_Type 3	6.0	85.958333	49.623688	0.0	68.9375	95.375	125.00	130.00
Room_Type 4	9369.0	133.247350	35.743346	0.0	110.0000	133.100	155.00	375.50
Room_Type 5	906.0	158.718366	50.939994	0.0	125.0000	162.000	198.00	269.00
Room_Type 6	1540.0	190.853740	45.094137	0.0	167.4500	190.000	220.00	349.63
Room_Type 7	307.0	186.015212	94.121170	0.0	170.9450	211.410	243.90	352.50

- Highest cancellation rate is with room type 6, which is over 50%.
- Room type 6 is also one of the most expensive rooms.
- Room type 7 is also expensive but has a much lower cancellation rate.
- All of the remaining rooms have over a 60% not cancellation rate.

Data Preprocessing

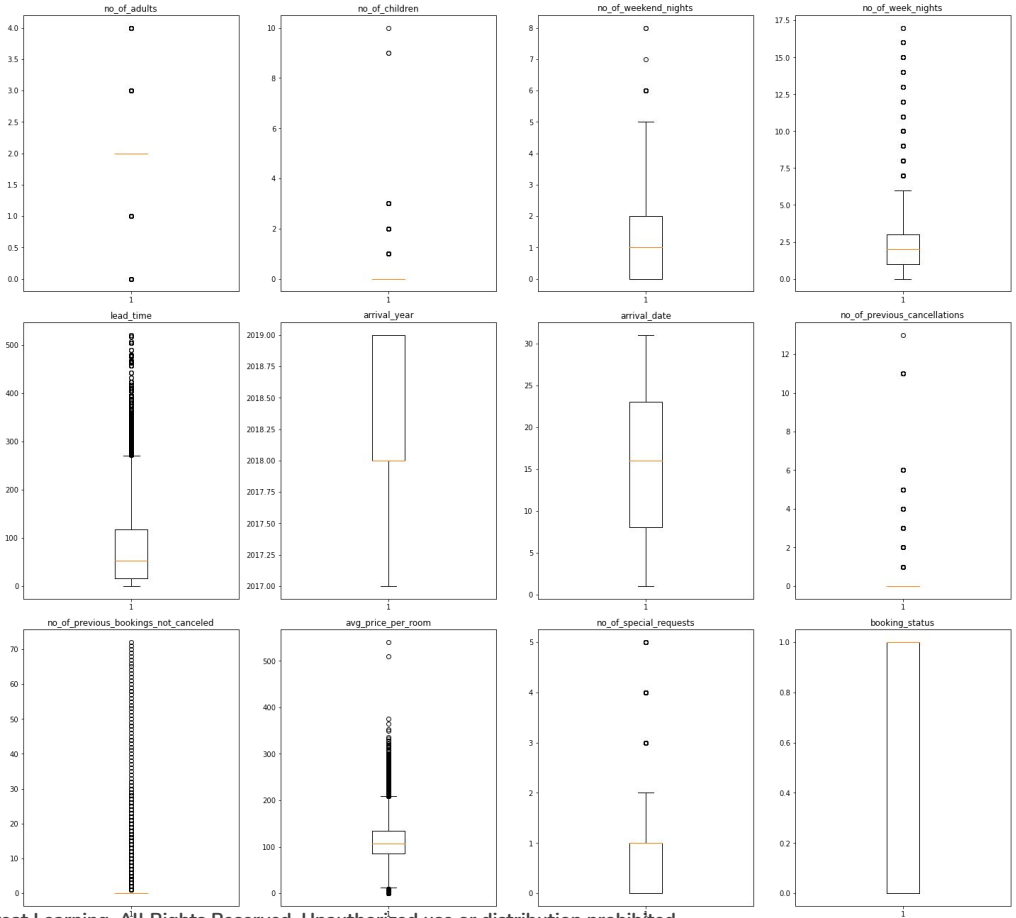
- No null values in the data
- Arrival_month was changed from numbers to string object to make reading visualizations easier.
- Duplicated values were dropped before EDA.
- Repeat_guest and required_car_parking_space were changed from 0-”No”, and 1-”Yes”
- Dummy variables made from non numeric columns: “Type_of_meal_plan”, “required_parking_space”, “room_type_reserved”, “arrival_month”, “market_segment_type”, “repeated_guest”, “booking_status”.
- Outlier treatment:
    - Due to high number of outliers skewing data to the right they were treated by assigning all values smaller than the lower whisker (25th quantile) the value of the lower whisker and all values greater than the upper whisker (75th quantil) the value of the upper whisker.

With null values

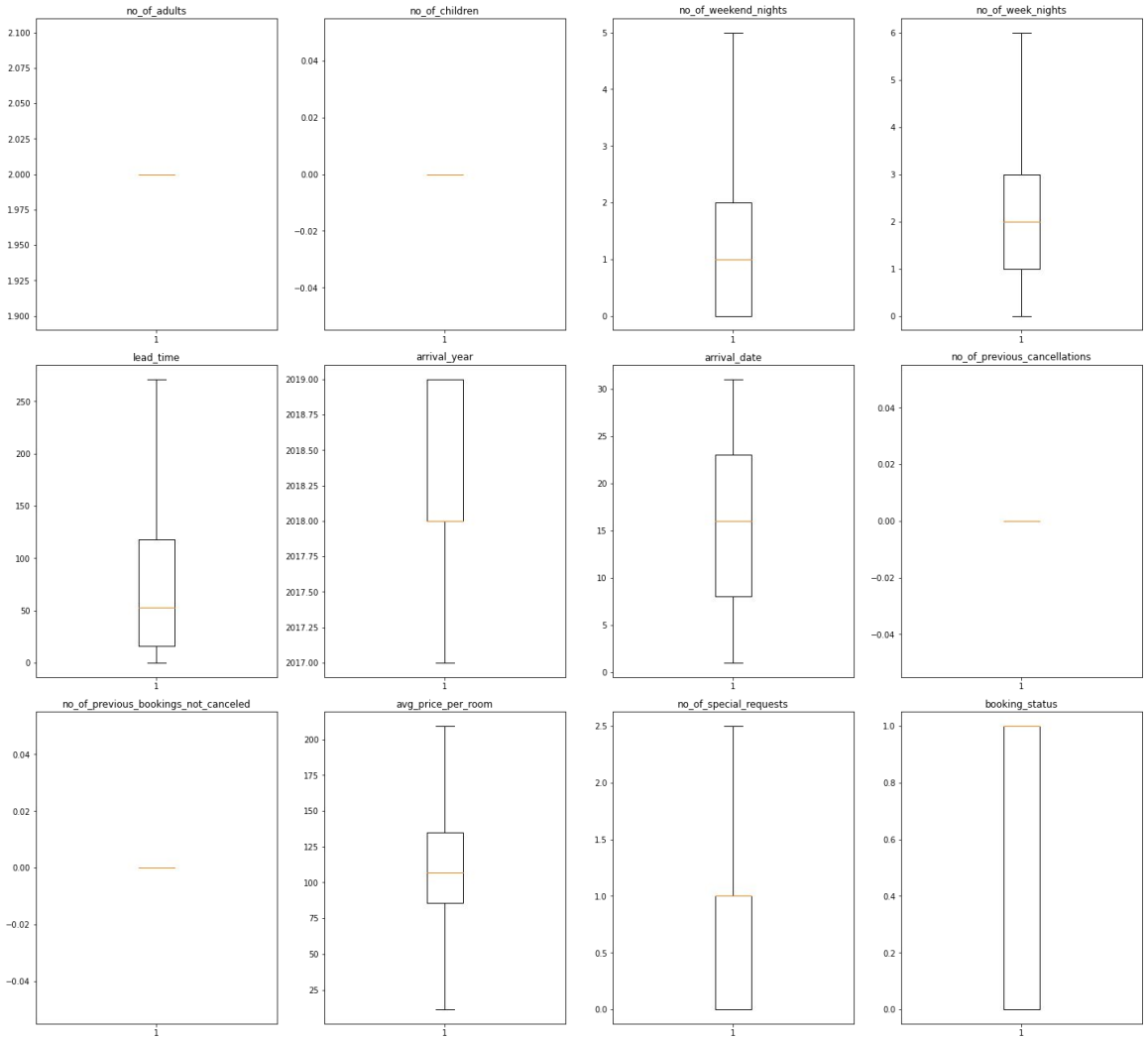
With column median replacing null

Data Preprocessing Continued

- Data is very heavy on outliers in all categories
- We will treat the outliers by assigning all values below lower whisker the value of lower whisker. Same formula will be assigned to upper whisker values.



- Outliers adjusted.



Testing for Multicollinearity in Data

- Multicollinearity will be tested using Variance Inflation Factor (VIF), which measures how much the variance of the estimated regression coefficient is “inflated” by the existence of correlation among the predicator variables in the model.
- If VIF is 1 then there is no correlation, whereas if the VIF exceeds 5 there is moderate correlation. If the VIF is greater than 10, it shows signs of high multicollinearity.
- None of the factors show a high relationship with others as all are sub VIF of 5.
- Some factors have VIF close to 5, but no generalizations can be drawn from that factor’s category.
- Columns removed from modeling due to having very low VIF:
    - Numer of children
    - Number of previous cancellations
    - Number of previous bookings not cancelled

Logistic Regression Model

Model evaluation criterion

Model can make wrong predictions as:

1. Predicting a customer will contribute to the revenue but in reality the customer would not have contribute to the revenue. - Loss of resources
2. Predicting a customer will not contribute to revenue but in reality the customer would have contributed to revenue. - Loss of opportunity

Which case is more important?

- If we predict a customer who was going to contribute to the revenue as a customer who will not contribute to the revenue.

How to reduce this loss i.e need to reduce False Negatives?

- `recall` should be maximized, the greater the recall higher the chances of minimizing the false negatives.

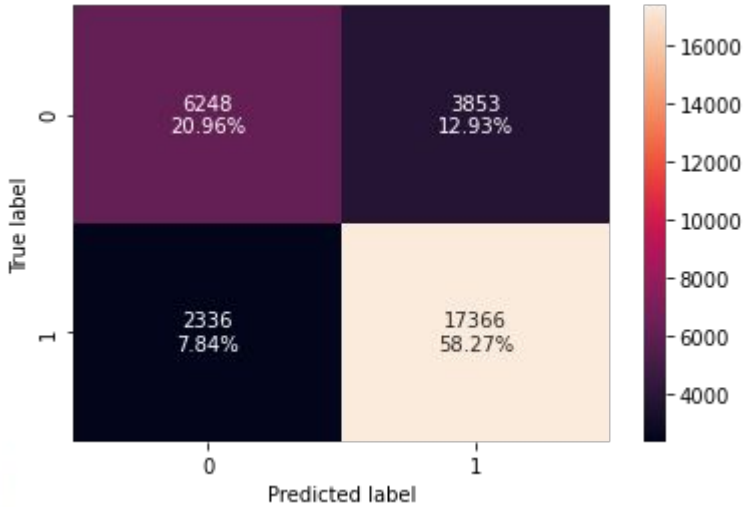
Logistic Modeling

- Model prepared on a 70:30 split of data into training and test sets.
-

```
Number of rows in train data = 29803
Number of rows in test data = 12773
```

Training performance:

	Accuracy	Recall	Precision	F1
0	0.792806	0.881484	0.818928	0.849055



The Confusion Matrix

- True Positives (TP): Correctly predicted that 6248, or 20.96% will cancel
- True Negatives (TN): We predicted that 17,366 or 58.27% will not cancel their booking
- False Positives (FP): We incorrectly predicted that 12.93%, or 3853 would cancel
- False Negatives (FN): We incorrectly predicted that 7.84% or 2336 would not cancel their bookings
- Overall our model accounted for 79.23% of the true values

Logistic Modeling Coefficient Analysis

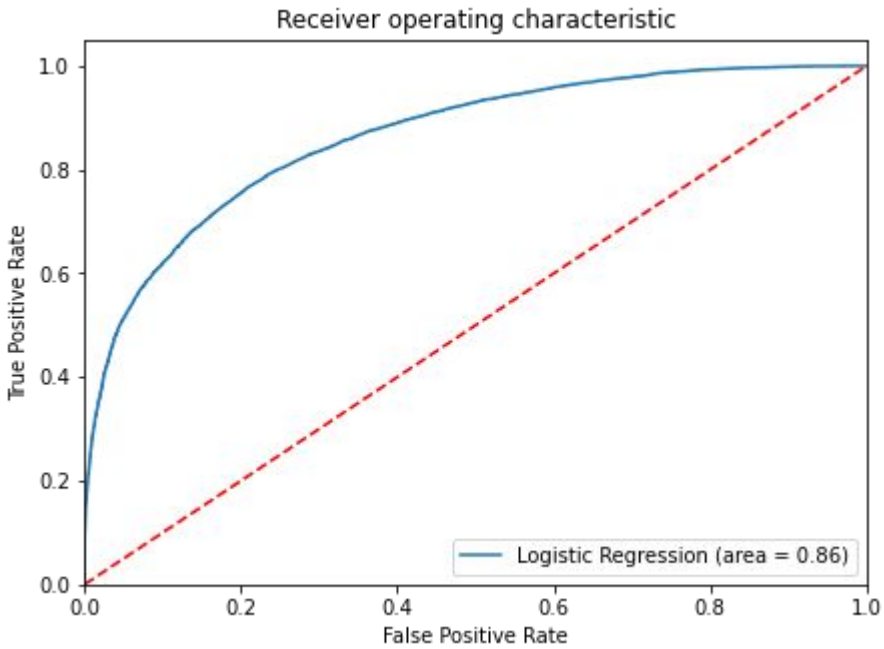
Coefficient Analysis

- The largest negative contribution to canceling is number of adults.
- The largest positive contribution to not canceling is from complementary rooms, which would make sense as they would be free and so a guest would be extremely unlikely to cancel.
- The second largest positive contribution comes from being a repeated guest, which makes sense as if a guest enjoyed their stay they are more likely to return and not cancel.
- Other major positive contributors to not cancelling are bookings done offline, in the month of December, and requesting a parking space.
- The impact of special requests also contributes to not cancelling a booking.

- Curve is giving good performance on training set.

Try to improve the model

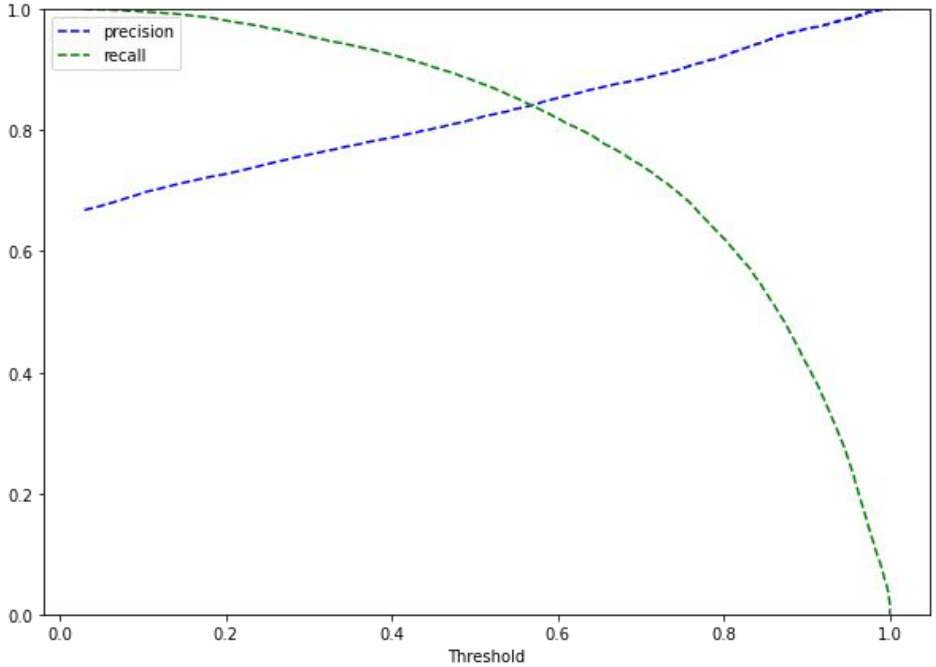
- Find optimum threshold and ajust log model and compare to see if before or after is better



New Model with adjusted threshold from AUC

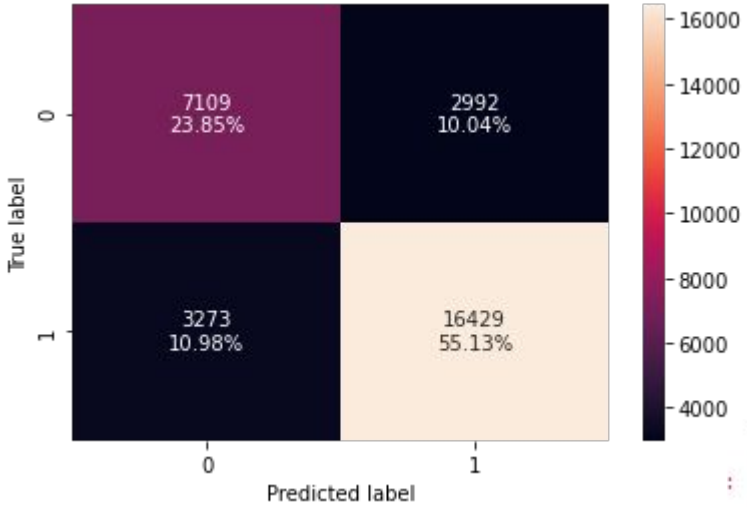
	Accuracy	Recall	Precision	F1
0	0.774956	0.767435	0.876776	0.81847

- Accuracy and Recall have reduced
- Check Precision-Recall curve to see if another threshold becomes viable



- Threshold of ~0.58 gives a balances precision/recall score.
- new model with threshold changed to prevision/recall forecast.

New Model from Precision/Recall Curve



	Accuracy	Recall	Precision	F1
0	0.789786	0.833875	0.84594	0.839864

	Logistic Regression sklearn	Logistic Regression-0.33 Threshold	Logistic Regression-0.58 Threshold
Accuracy	0.792336	0.774956	0.789786
Recall	0.881433	0.767435	0.833875
Precision	0.818417	0.876776	0.845940
F1	0.848757	0.818470	0.839864

Test set performance comparison:

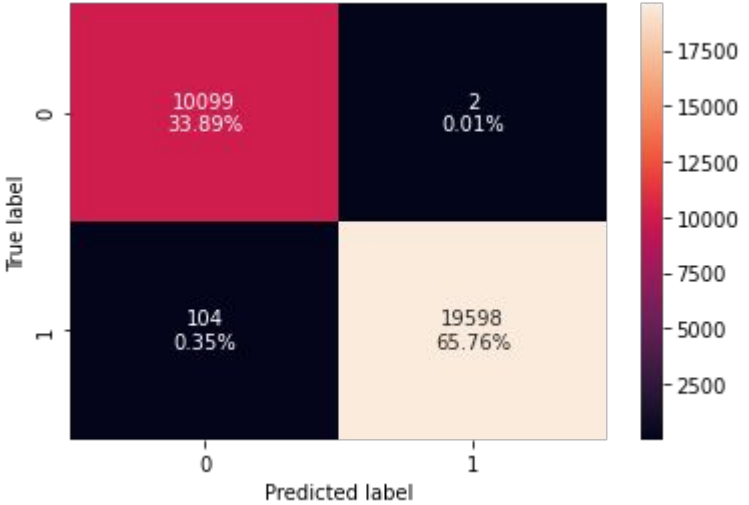
	Logistic Regression sklearn	Logistic Regression-0.76 Threshold	Logistic Regression-0.58 Threshold
Accuracy	0.793471	0.772959	0.789948
Recall	0.881602	0.762609	0.835460
Precision	0.818011	0.875565	0.843201
F1	0.848617	0.815192	0.839312

Logistic Model Conclusions:

Final Observations on Logistic Model

- A predictive model has been built that provides insight into the hotel predicting whether or not a guest will cancel their booking with a good recall and an F1 score of 0.85 on training and test data.
- Coefficients showing a positive affect on keeping guests include requiring parking spaces, number of special requests,bookings in December, offline reservations, corporate reservations and repeated guests.
- The most impactful factor in increasing a guests cancellations are the number of guests, so as the number of guests for a reservation increase so to does the rate of cancellation.
- Number of guests is followed by bookings in February, and rooms reserving room type 3.

Decision Tree Modeling: Default Gini Criteria Training Set

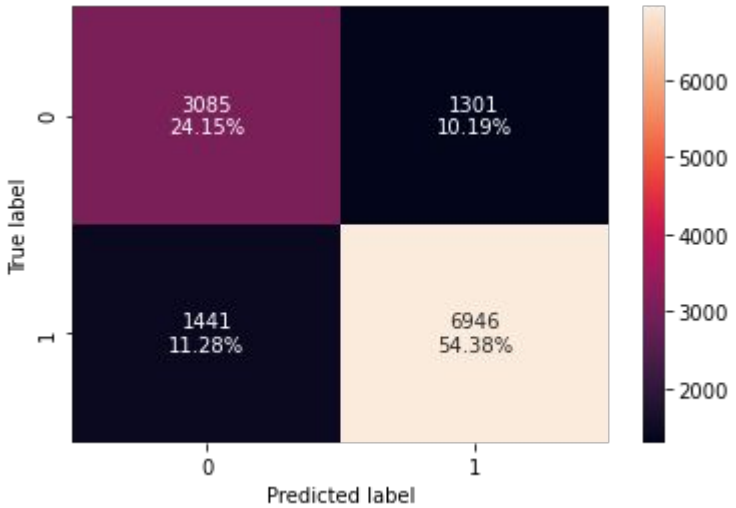



```
Recall Score: 0.9947213480864887
```

Decision Tree Model Observations

- Model predicts very well the true positives and the true negatives, predicting 99.65% correct
- This model has no restrictions on the tree and is un-pruned, creating potential overfitting of the data as the model aims for perfect predicting.

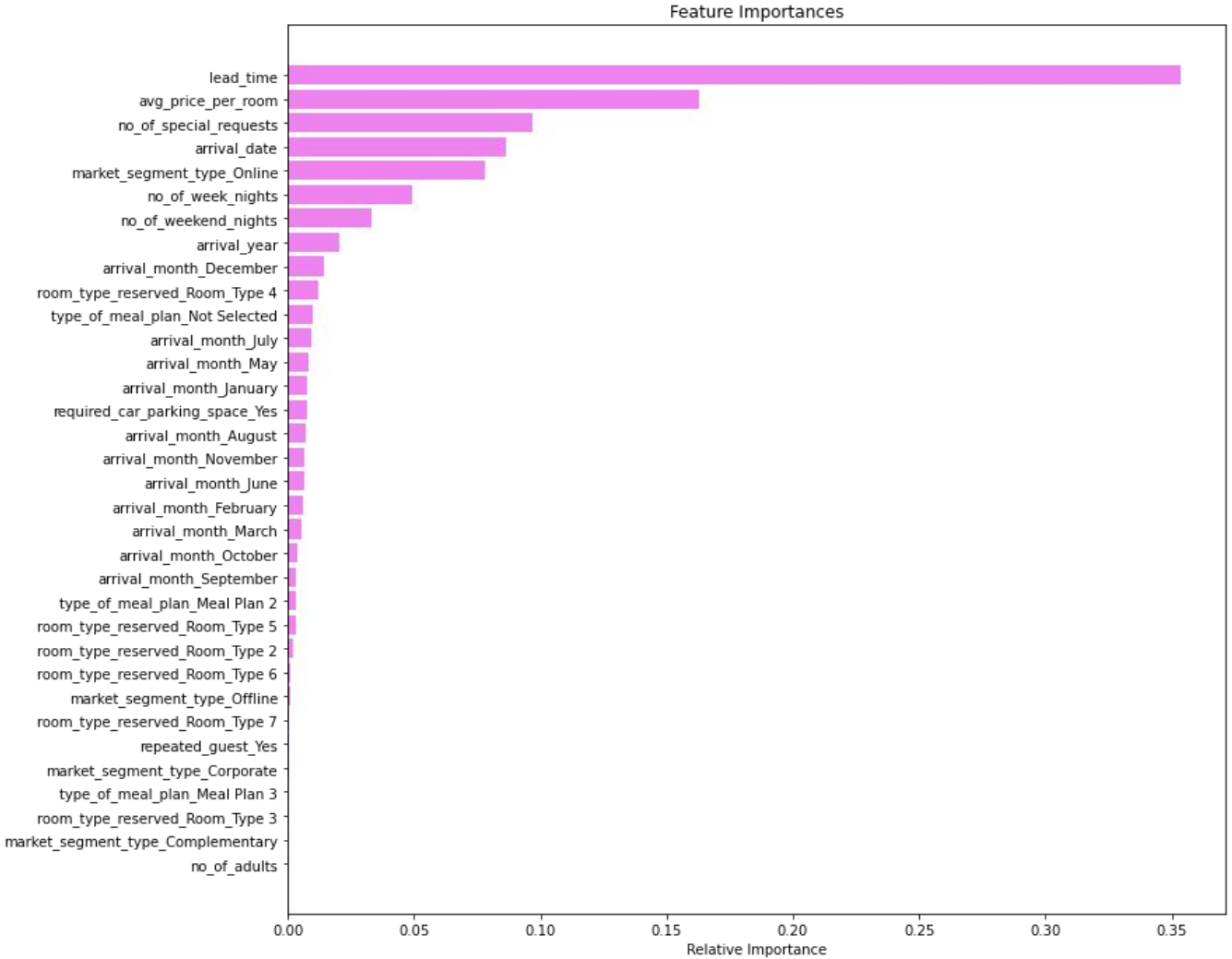
Decision Tree Modeling: Default Gini Criteria Test Set




```
Recall Score: 0.8281864790747585
```

- Recall score between training and test is very far off, showing training model is overfitting.
- We will prune the tree to bring the training and test results closer.

Decision tree feature importance showing lead time as most influential in cancellations.



Decision Tree: Reducing Overfit using Hyperparameters

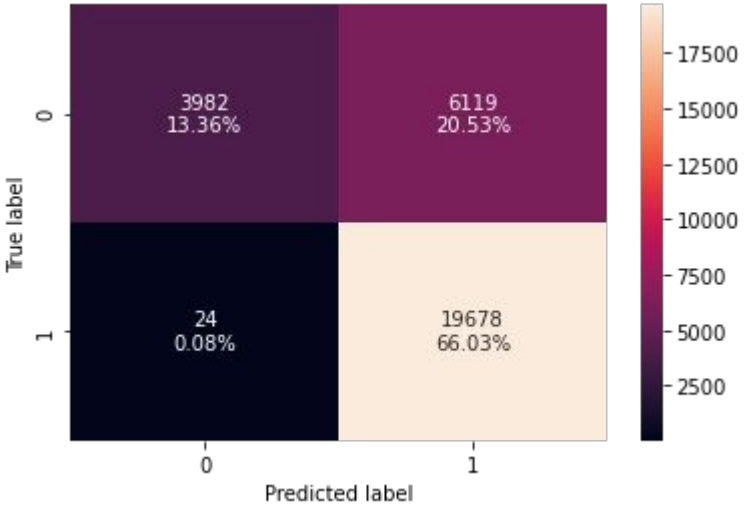
Great Learning
POWER AHEAD

Reducing the overfitting model

Using GridSearch for Hyperparameter tuning of our tree model

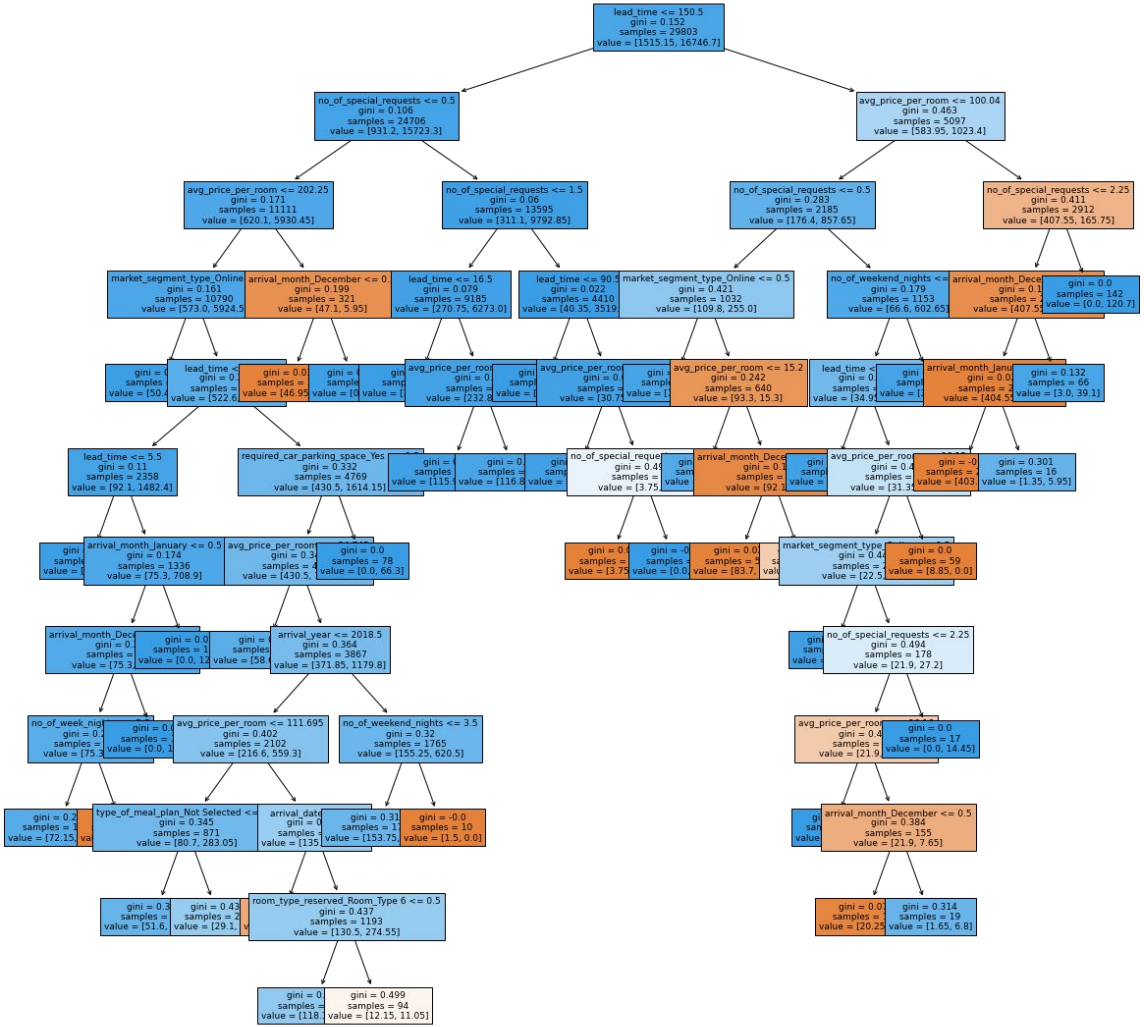
- Hyperparameter tuning is also tricky in the sense that there is no direct way to calculate how a change in the hyperparameter value will reduce the loss of your model, so we usually resort to experimentation. i.e we'll use Grid search
- Grid search is a tuning technique that attempts to compute the optimum values of hyperparameters.
- It is an exhaustive search that is performed on a the specific parameter values of a model.
- The parameters of the estimator/model used to apply these methods are optimized by cross-validated grid-search over a parameter grid.

Decision Tree: Optimial Hyperparameter training set

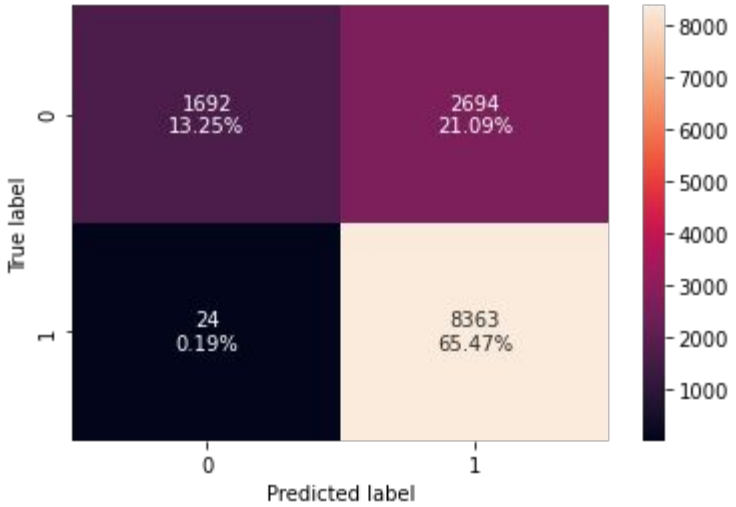


	Accuracy	Recall	Precision	F1
0	0.79388	0.998782	0.762802	0.864986

Decision Tree from optimized hyperparameters with first split at lead time of 150 days.



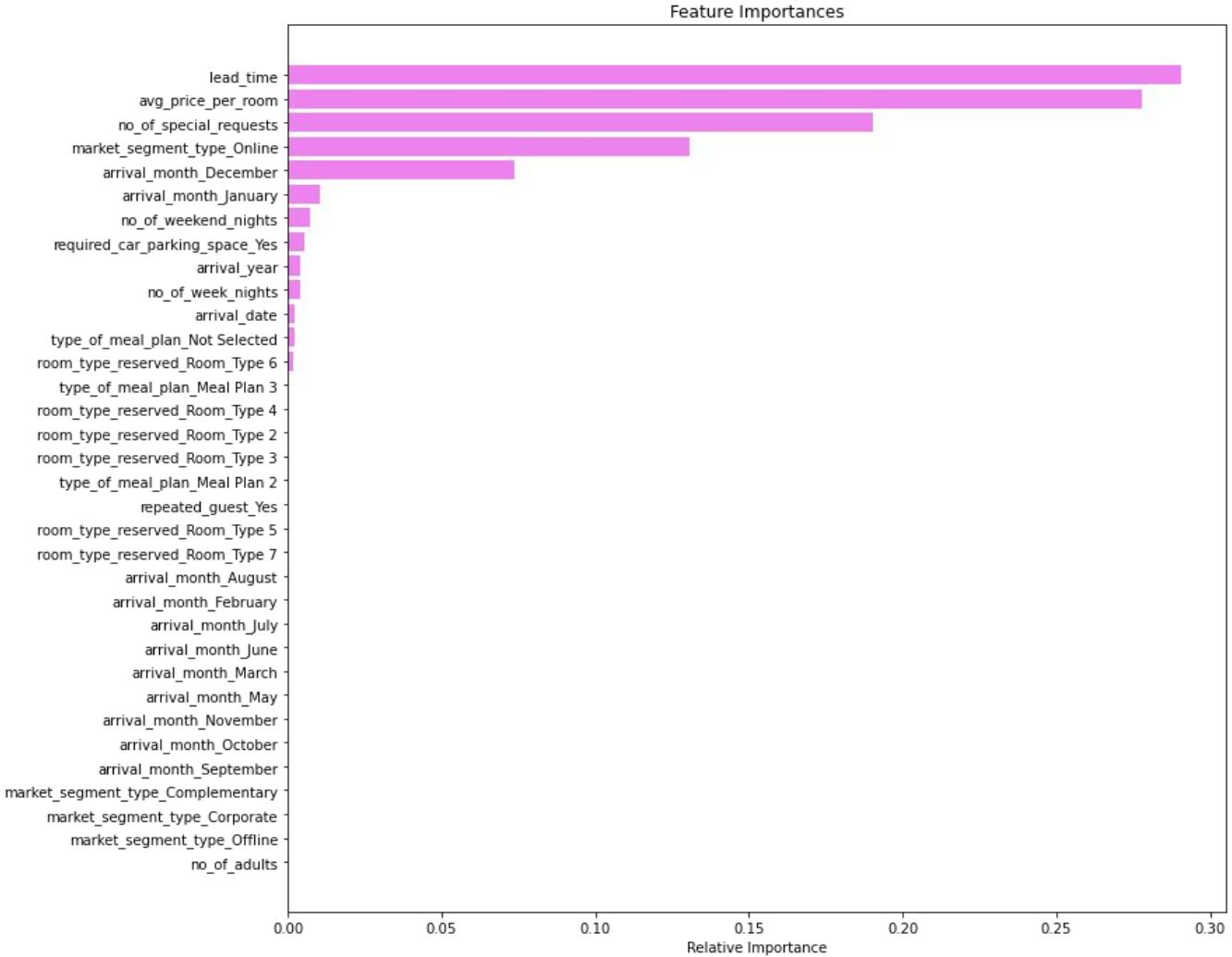
Decision Tree: Optimal Hyperparameters test set



	Accuracy	Recall	Precision	F1
0	0.787207	0.997138	0.756353	0.860214

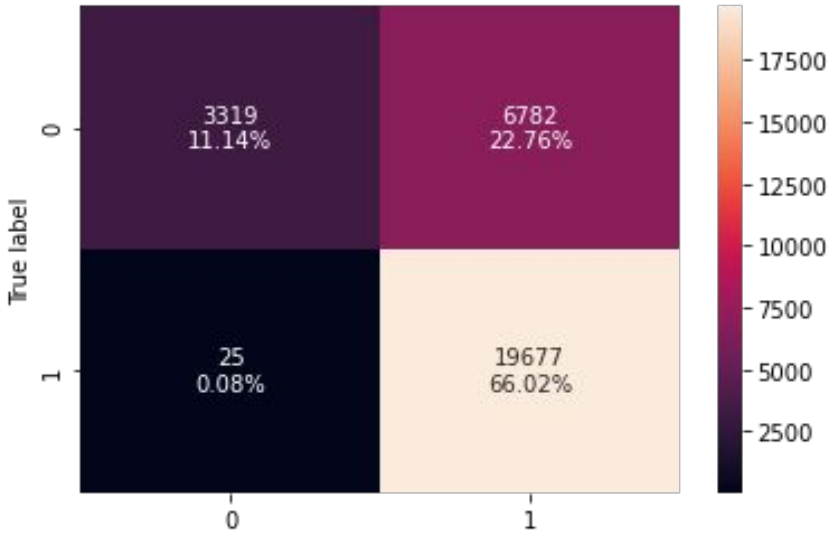
- Training and Test data results are very close.
- No need to further prune the Decision Tree.

Feature importance using optimal hyperparameters. Lead time is still most influential, however now other factors have increased their influence.



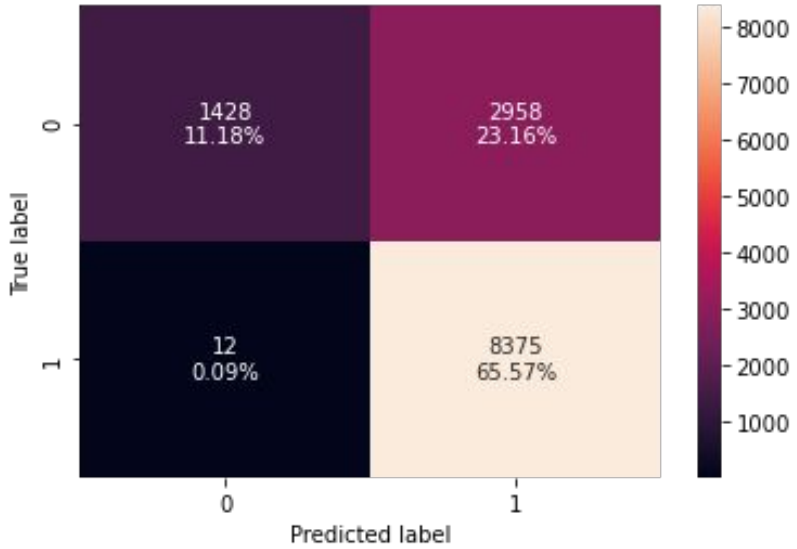
Post Prune Decision Tree with Alpha

Training



	Accuracy	Recall	Precision	F1
0	0.7716	0.998731	0.743679	0.852538

Test



	Accuracy	Recall	Precision	F1
0	0.767478	0.998569	0.738992	0.849391

Model has very good recall and high F1

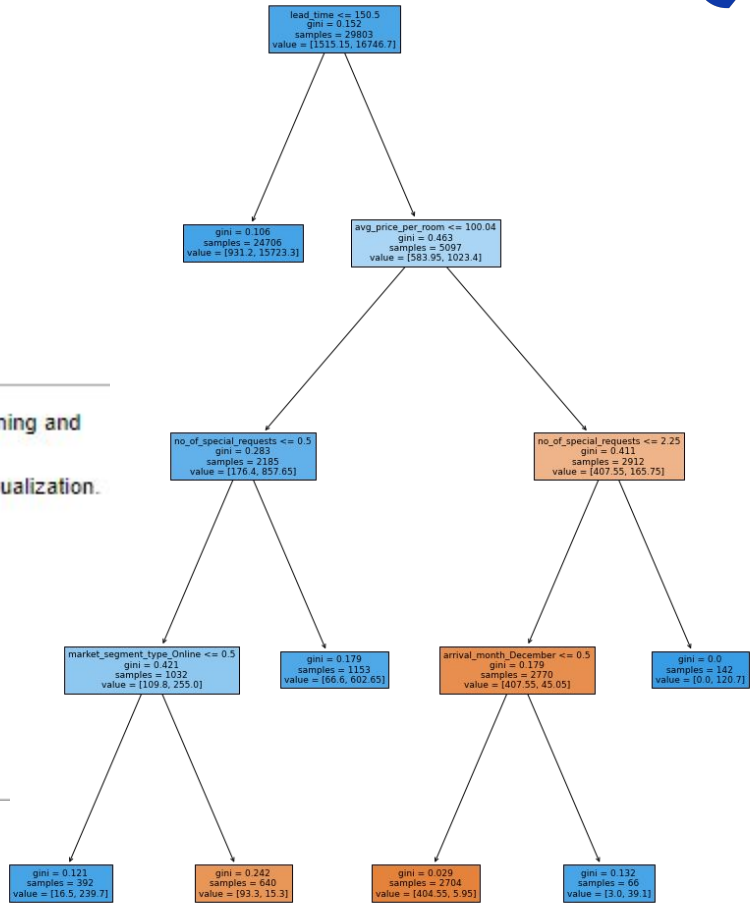
Post Pruned Decision Tree

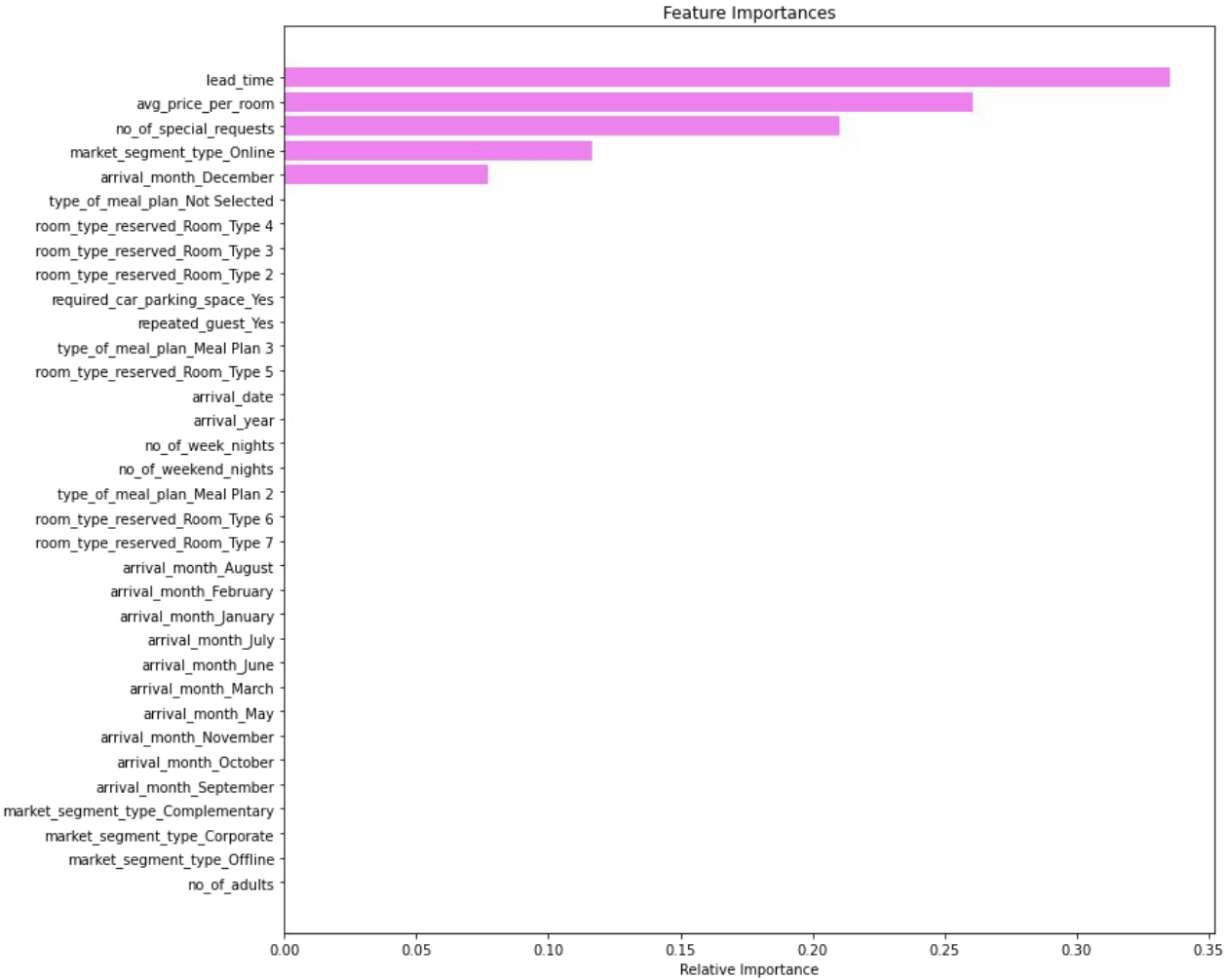
- Model is much simpler once we set the alpha level.
- Very easy to read and understand.

- We will go with the post-pruning model as it offers very good statistics on both training and test data.
- The last decision tree model is also very easy to understand with the simplified visualization.

Model Performance Comparison and Conclusions

- The post-pruned decision tree model best fits the data.
- The decision tree better parallels the EDA of the data, following the trends that affect cancellations the most: lead time, price per room, online bookings,Winter arrival months, and special requests.





Post pruned decision tree feature importance still has lead time as most influential in cancellation decisions followed by price per room and special requests

Actionable Insights from models:

Actionable Insights and Recommendations

- What profitable policies for cancellations and refunds can the hotel adopt?
  - According to the final decision tree:
    - A customer with a booking lead time of over 150 days has a higher chance of cancelling.
    - A customer with a booking lead time of under 150 days and purchasing a room below 100 Euros has a high chance of not cancelling.

Business Recommendations:

Recommendations based upon current information:
-Attempt to cater to more families and especially focus on summer months as currently winter months have the lowest cancellations rates. Why is the Hotel losing out on Summer bookings? Being in Portugal the summer months should be prime opportunity for more guests. Does the site have a pool? Close to the beach? Summer festivals/specials to entice more families to come.
-The Hotel needs to look at the current prices of rooms and determine if room type 6 and 7 are truly worth the extra expense. Type 6 has the highest cancellation rate and is also the most expensive.
-Work on retaining guests that book out over 150 days. Guests wanting a summer getaway often shop well in advance for lower room rates. Incentivize customers that purchase that far out with complimentary popular special request items, or perhaps reduced/free meal plans or room upgrades. And if the price changes, notify the customer and offer them the difference.
-Online purchases have the highest cancellation rates. Is this due to the purchases on the Hotel website or third party websites?
-The hotel does a good job in catering to guests with the special requests. We recommend the hotel consider offering the most popular requests as complimentary services for returning guests.

Further Recommendations.

-More data is needed in order to build a more robust model for the Hotel.

-The room types needs to be further broken down to truly analyze why certain rooms have higher rates.

-More information on online purchases is necessary. Are they from third party sites, or the hotel site? If the hotel site then perhaps a remodel of the site to make things easier and help retain customers.

-More information is needed for offline purchases. Are they through a travel agency? On-site sales?

-The Hotel needs to research ways to bring in families in the summer months. Festivals? Local attractions? Better summer activities on-site?

-Incentives need to be given to current guests to help them come back for more stays. Perhaps complimentary parking, meal plans, room upgrades, or certain special requests.

-What are the current cancellation and refund policies? Do they differ depending on whether the booking is online vs offline? Are they full refunds? Partial refunds?

greatlearning
Power Ahead

Happy Learning !