



INN Hotels Group

An approach to understanding cancellation rates and potential remediation.



Data Analysis Observations & Recommendation

John D Noble
November 2021

Business Problem Overview & Solution Approach

Inn Hotels sees almost 1/3rd of all of there reservations canceled. While some level is expected, is there a way to a smart ML-based point classification system to dynamically price based, build customer loyalty and ultimately reduce the unforeseen cancelations and drive lost revenue.

Key Questions We Will Answer!

1. *What are the key predictors +/- of cancelation rates?*
2. *Can we build a model (s) to predict used phone price? If yes, how accurate will the model be?*
3. *What are the predictive variables actually affecting the rate?*
4. *How can we compare models, and chose the best one for solving this problem?*



What are we trying to solve for?

How best to take advantage of the historical data we have collected on each segment and build a “realtime” decision model to determine dynamic pricing and/or incent users to not cancel.

Financial Implications

A reservation booked and then canceled is lost opportunity/revenue as that room likely can not be resold for the same price. Building customer loyalty and attractive pricing remains first, however there are few barriers to cancelation, so balancing our bottom line financial interest with our customer –first staretgy needs to be balanced..

Exploratory Data Analysis (EDA): General Observations

AFTER ALL data preprocessing this is what we end up with:

- Number of variables 18
- Number of observations 36238
- Missing cells 0
- Missing cells (%) 0.0%
- Duplicate rows 0
- Duplicate rows (%) 0.0%
- Total size in memory 5.0 MiB
- Average record size in memory 144.0 B
- Numeric 11
- Categorical 7

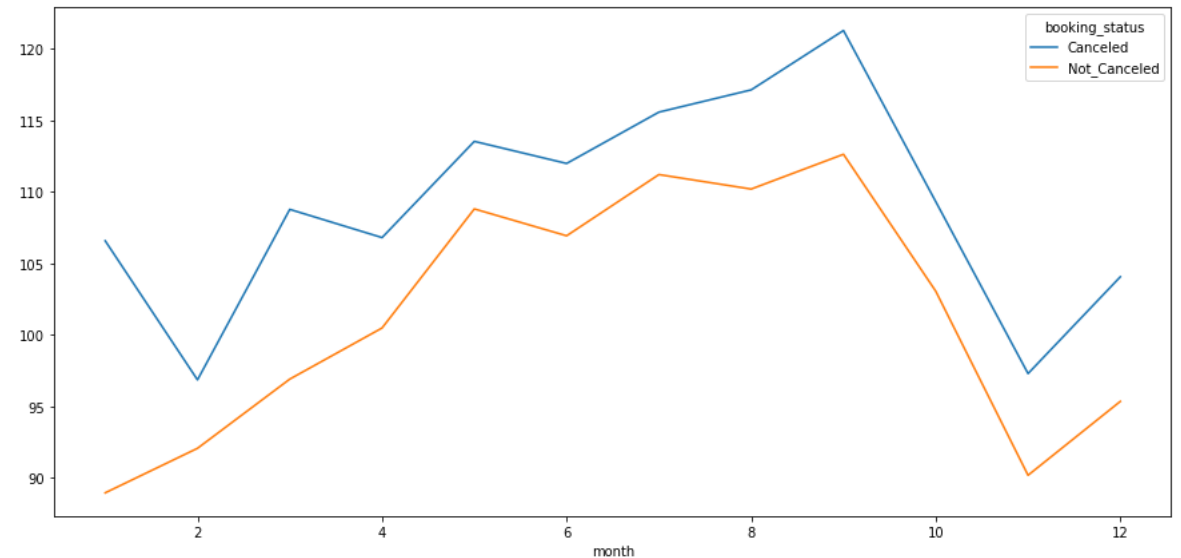
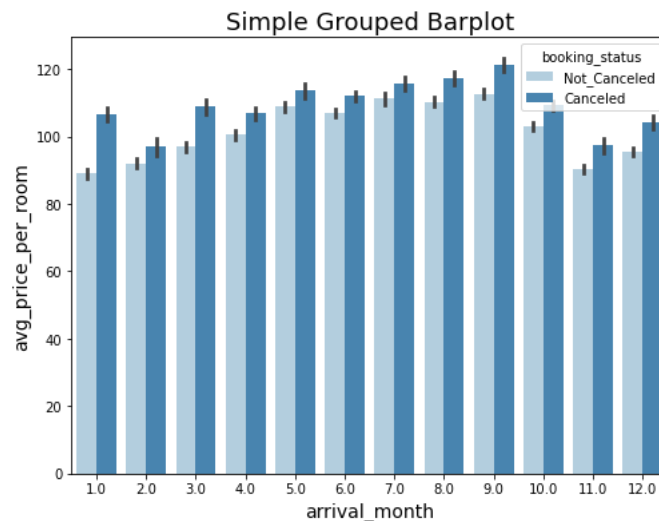
However ... the data suggest there are some interesting relationships that immediately jump out as areas to further explore.



Exploratory Data Analysis (EDA)

Number of canceled reservations: 11878 (32.78%)

Number of reservations not canceled: 24360 (67.22%)

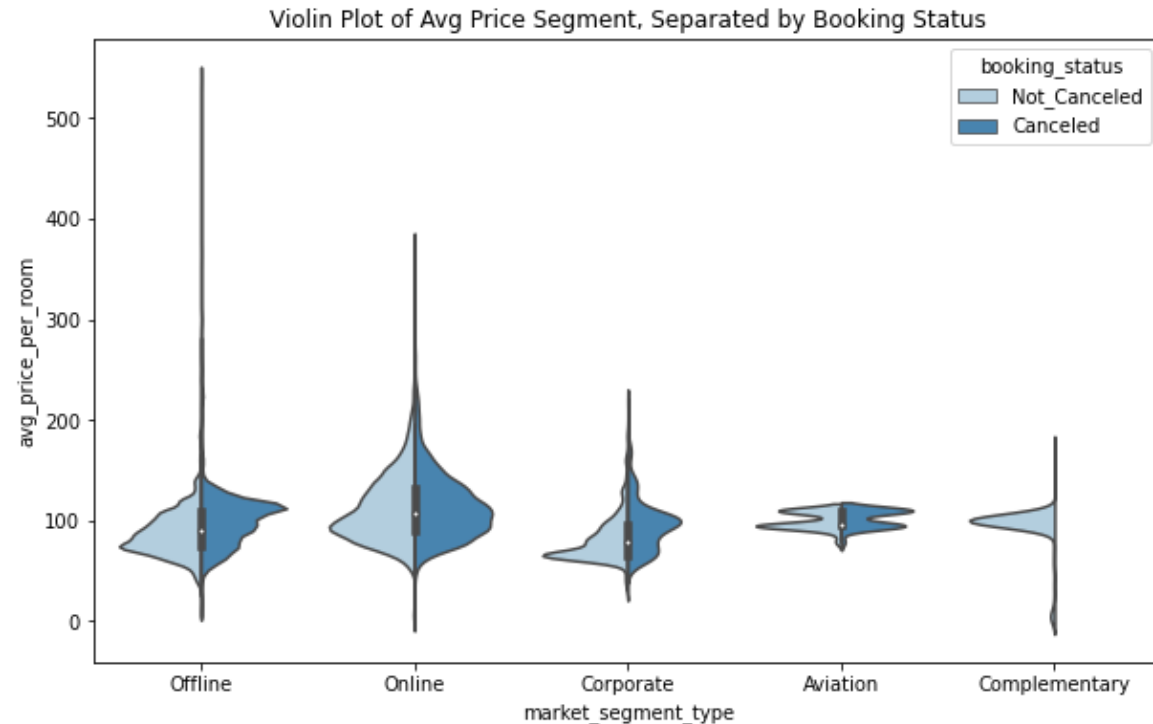


Cancelations vs avg price, goes up throughout the year especially in summer and fall.

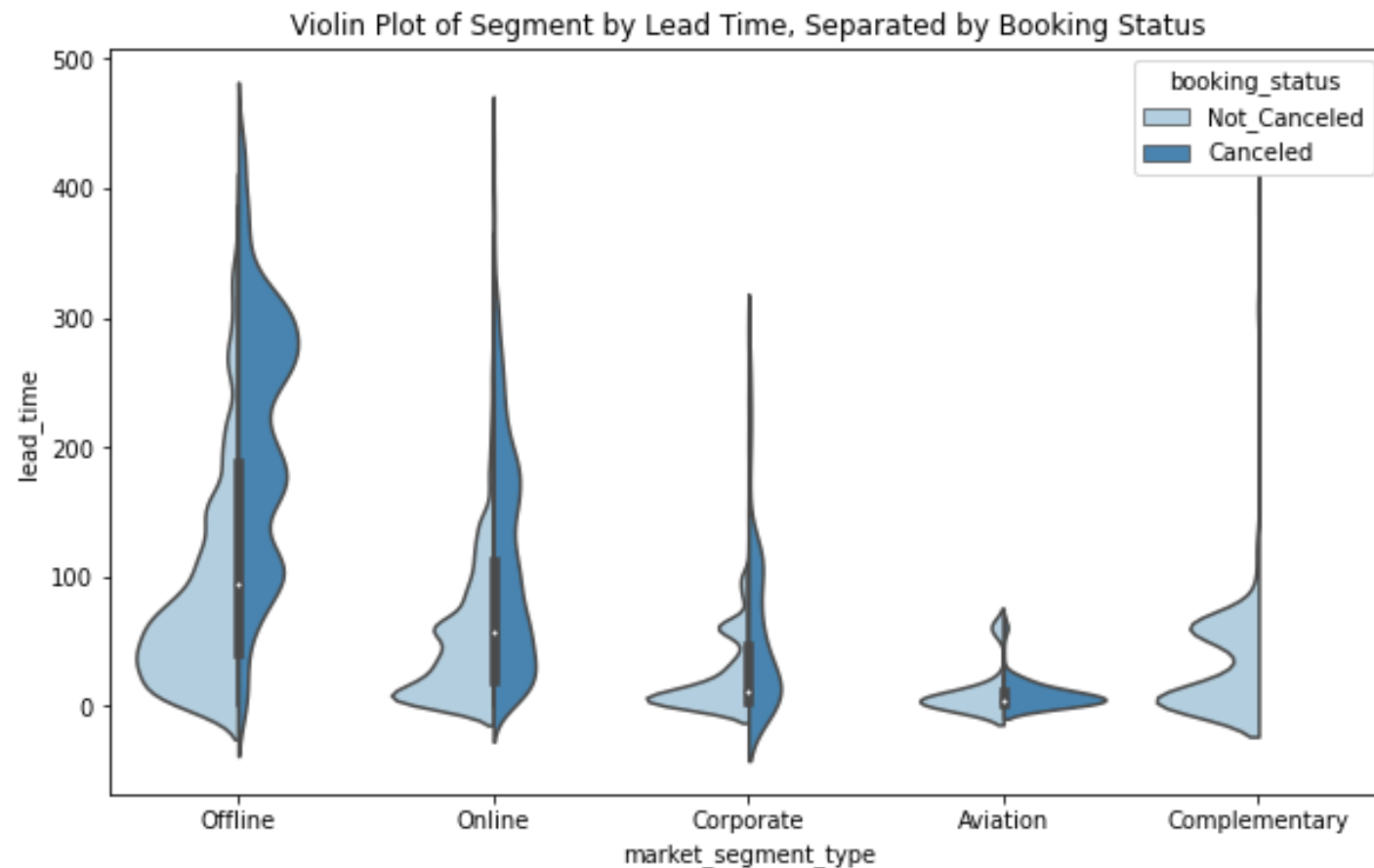
Exploratory Data Analysis (EDA)

There are differences in the avg price paid for each reservation by segment.

Online users tend to have the widest spread, compared to our Aviation segment clients who have renegotiated rates.



Exploratory Data Analysis (EDA)

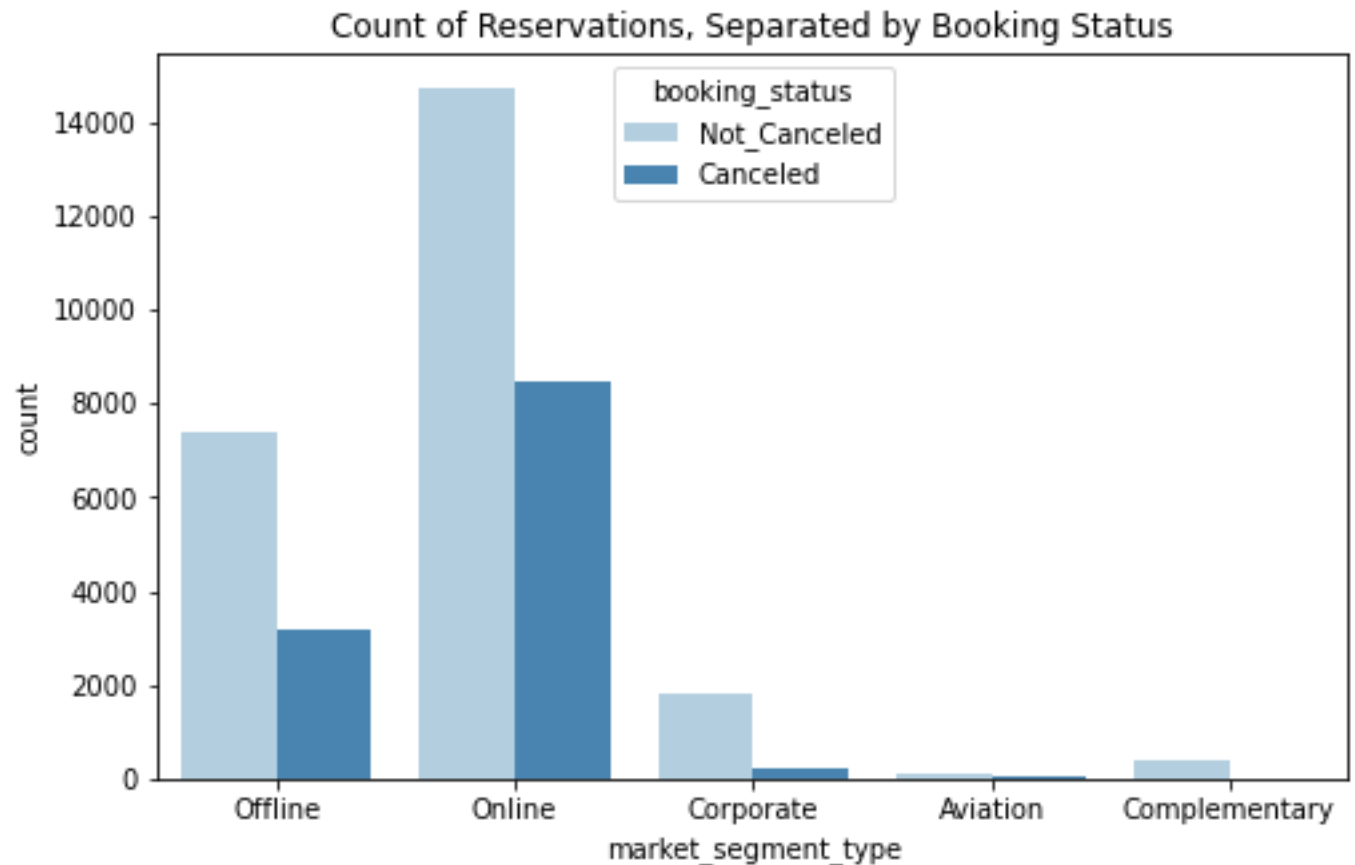


Lead time have very large differences where offline (old school) users tend to plan way in advance, followed by online users.

Our Aviation clients, have very dynamic schedules and hence the lead times we have are the shortest compared to all other segments.

Exploratory Data Analysis (EDA)

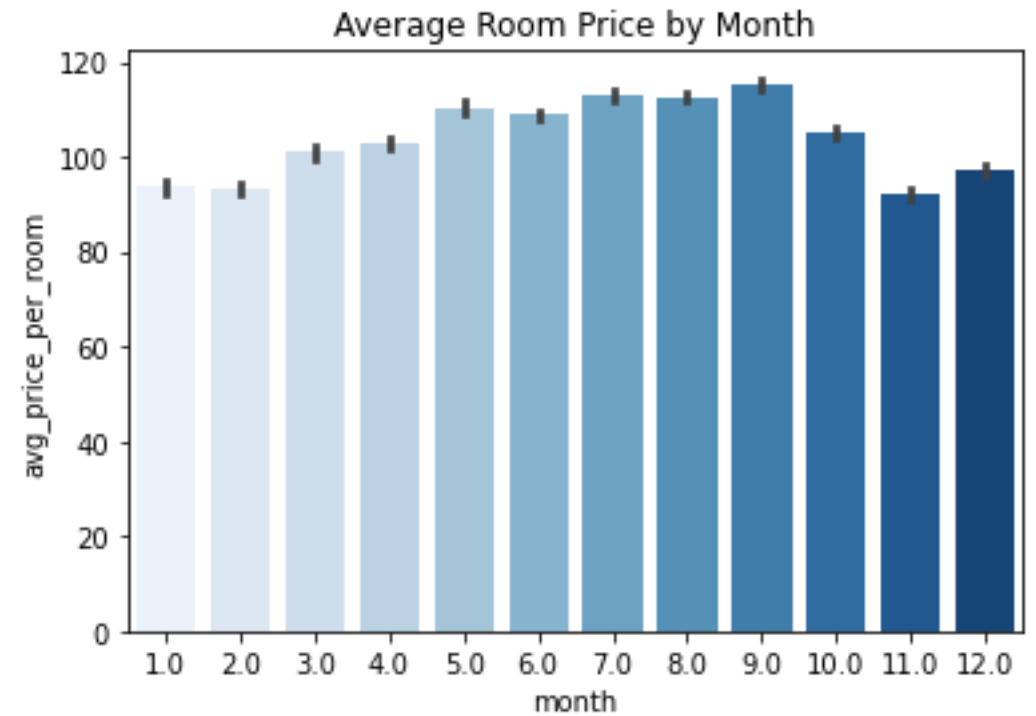
We believe the convenience of 'online booking' drives that channel usage, but the downside is it is an unemotional/quick ability to cancel which has no barriers and is almost 2x the rate of the next highest category.



Exploratory Data Analysis (EDA)

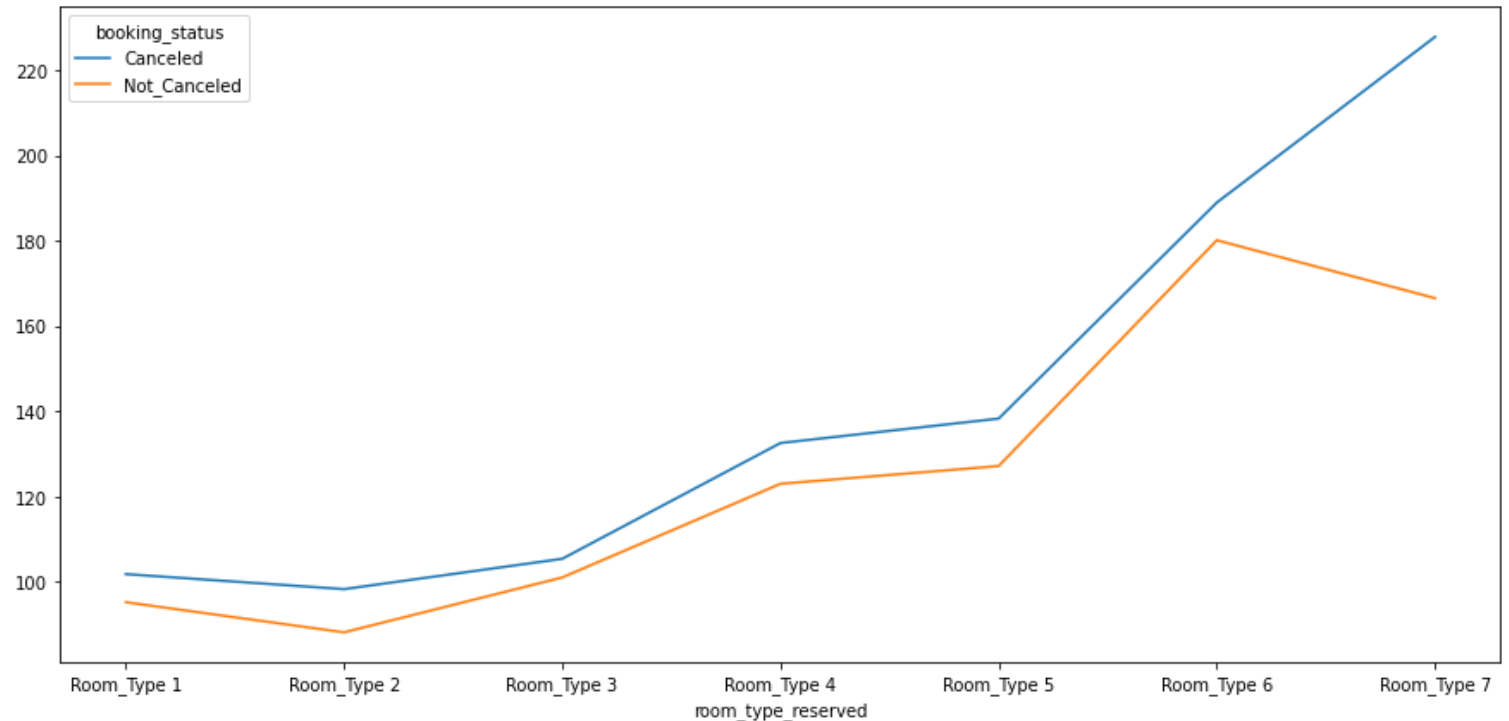
We typically see demand increase throughout the year, rising in the summer through September.

Traditionally we have raised rates based on the demand in those months.



Exploratory Data Analysis (EDA)

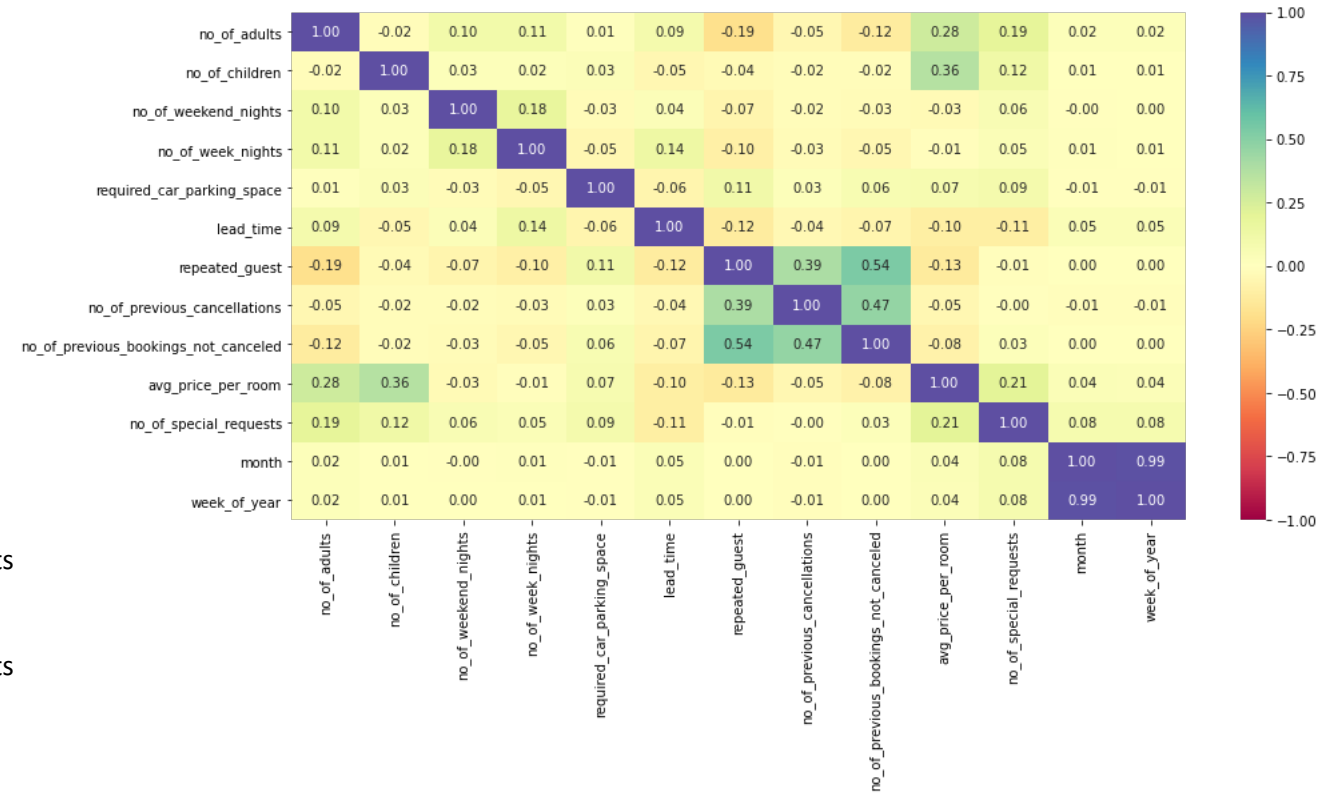
Insight: The ratio of cancelations between room rates are constant, except for some of the Cat4 rooms, and especially goes with room types beyond CAT 6.



Exploratory Data Analysis (EDA)

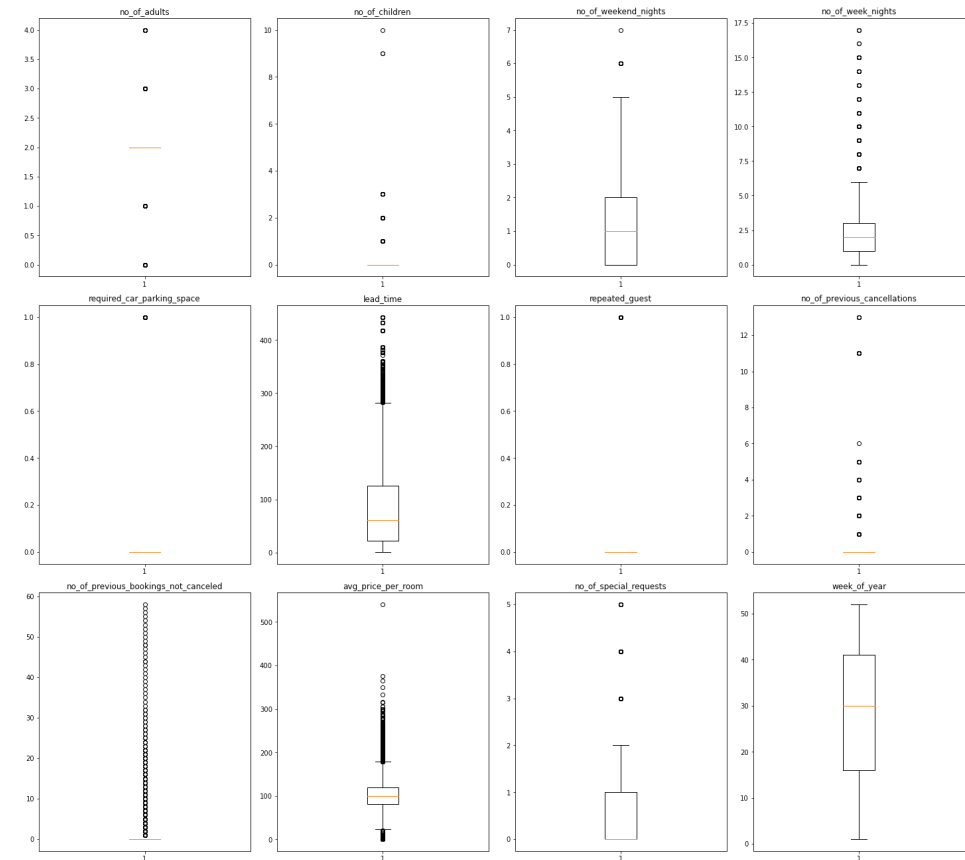
Insight:

1. month is highly correlated with week_of_year
2. week_of_year is highly correlated with month
3. month is highly correlated with week_of_year
4. week_of_year is highly correlated with month
5. month is highly correlated with week_of_year
6. week_of_year is highly correlated with month
7. booking_status is highly correlated with lead_time
8. no_of_previous_bookings_not_canceled is highly correlated with repeated_guest and 1 other fields
9. no_of_week_nights is highly correlated with no_of_weekend_nights
10. lead_time is highly correlated with booking_status
11. no_of_children is highly correlated with room_type_reserved
12. no_of_weekend_nights is highly correlated with no_of_week_nights
13. repeated_guest is highly correlated with no_of_previous_bookings_not_canceled
14. room_type_reserved is highly correlated with no_of_children
15. month is highly correlated with week_of_year
16. no_of_previous_cancellations is highly correlated with no_of_previous_bookings_not_canceled
17. week_of_year is highly correlated with month
18. no_of_previous_cancellations is highly skewed



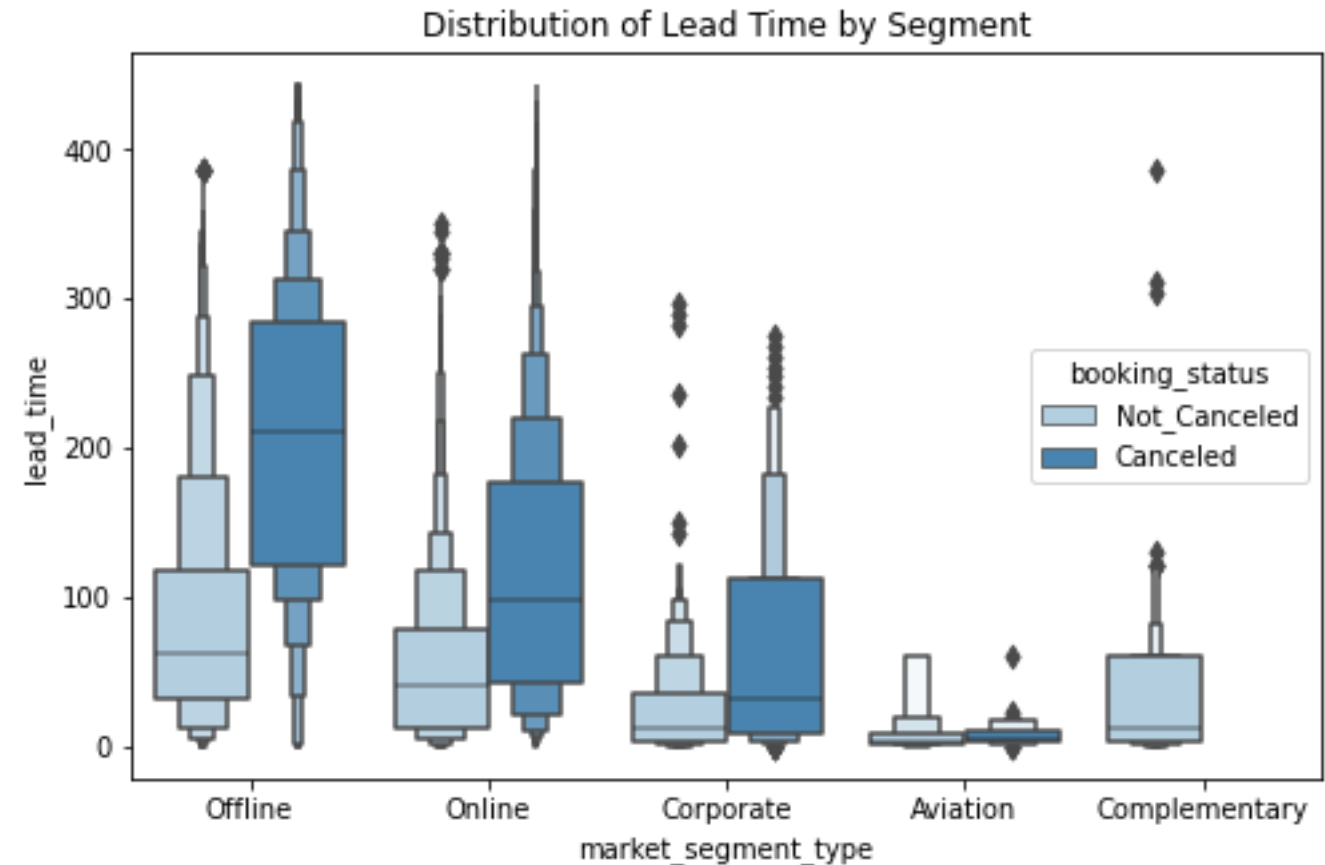
Exploratory Data Analysis (EDA)

Insight: although it is typical that most analysis is sensitive to extremes in the data, for this analysis we have shown that there is important information captured in those data points.



SUMMARY OF CORRELATIONS & OTHER FACTS. (EDA)

1. month is highly correlated with week_of_year
2. week_of_year is highly correlated with month
3. month is highly correlated with week_of_year
4. week_of_year is highly correlated with month
5. month is highly correlated with week_of_year
6. week_of_year is highly correlated with month
7. booking_status is highly correlated with lead_time
8. no_of_previous_bookings_not_canceled is highly correlated with repeated_guest and 1 other fields
9. no_of_week_nights is highly correlated with no_of_weekend_nights
10. lead_time is highly correlated with booking_status
11. no_of_children is highly correlated with room_type_reserved
12. no_of_weekend_nights is highly correlated with no_of_week_nights
13. repeated_guest is highly correlated with no_of_previous_bookings_not_canceled
14. room_type_reserved is highly correlated with no_of_children
15. month is highly correlated with week_of_year
16. no_of_previous_cancellations is highly correlated with no_of_previous_bookings_not_canceled
17. week_of_year is highly correlated with month
18. no_of_previous_cancellations is highly skewed ($\gamma_1 = 25.19401422$) ** We will drop either Month or Week
19. These columns have valid '0s'...these should not be treated!!!
20. no_of_children has 33544 (92.6%) zeros
21. no_of_weekend_nights has 16872 (46.6%) zeros
22. no_of_week_nights has 2383 (6.6%) zeros
23. no_of_previous_cancellations has 35901 (99.1%) zeros
24. no_of_previous_bookings_not_canceled has 35429 (97.8%) zeros
25. no_of_special_requests has 19751 (54.5%) zeros



Data Transformations

As noted initially, a number of clues were given up front that we were going to need to do some data preprocessing and transformations:

1. Missing data
2. outliers in the data do exist
3. skewness exists in a number of variables preprocessing/feature creation.

TRANSFORMATIONS

1. Created a new combined date feature; and then used that to create the Week #
2. Filled missing data w/ the median for avg price and lead times given their importance to the model.
3. DID NOT Floor & cap all other variables to address skewness by eliminating the outliers.
 - Used a new Zscore method to confirm no adjustment is needed.
4. Change categorical & object values to 1 or zero using pd.Dummies to make it easier to interpret their impact on price.
5. Dropped 'month' bit it gave us exactly the same information as the "week"
6. Encoding and replacing the words ' Not_Canceled and Canceled' with 0 and 1

Step 1: We want to predict the the likelihood of cancelation.

- Before we proceed to build a model, we'll have to encode categorical features.
- We'll split the data into train and test to be able to evaluate the model that we build on the train data.
- We will build a model using the train data and then check it's performance.

Shape of Training set : (25366, 25)

Shape of test set : (10872, 25)

Training set performance:

Accuracy: 0.80

Precision: 0.73

Recall: 0.61

F1: .66

Test set performance:

Accuracy: 0.80

Precision:0.72

Recall: 0.63

F1: 0.67

Logit Regression Results						
Dep. Variable:	booking_status	No. Observations:	25366			
Model:	Logit	Df Residuals:	25340			
Method:	MLE	Df Model:	25			
Date:	Fri, 19 Nov 2021	Pseudo R-squ.:	0.3165			
Time:	12:56:42	Log-Likelihood:	-10986.			
converged:	False	LL-Null:	-16073.			
Covariance Type:	nonrobust	LLR p-value:	0.000			
	coef	std err	z	P> z	[0.025	0.975]
const	-2.7129	0.253	-10.721	0.000	-3.209	-2.217
no_of_adults	0.1062	0.037	2.863	0.004	0.033	0.179
no_of_children	0.2652	0.059	4.522	0.000	0.150	0.380
no_of_weekend_nights	0.1595	0.020	8.142	0.000	0.121	0.198
no_of_week_nights	0.0311	0.012	2.583	0.010	0.007	0.055
required_car_parking_space	-1.5030	0.132	-11.383	0.000	-1.762	-1.244
lead_time	0.0156	0.000	62.075	0.000	0.015	0.016
repeated_guest	-3.0909	0.620	-4.985	0.000	-4.306	-1.876
no_of_previous_cancellations	0.3082	0.075	4.089	0.000	0.160	0.456
no_of_previous_bookings_not_canceled	-0.0074	0.061	-0.122	0.903	-0.126	0.111
avg_price_per_room	0.0179	0.001	24.881	0.000	0.017	0.019
no_of_special_requests	-1.4660	0.030	-49.364	0.000	-1.524	-1.408
week_of_year	-0.0037	0.001	-3.076	0.002	-0.006	-0.001
type_of_meal_plan_Meal Plan 2	0.0992	0.063	1.567	0.117	-0.025	0.223
type_of_meal_plan_Meal Plan 3	-11.4274	383.310	-0.030	0.976	-762.701	739.846
type_of_meal_plan_Not Selected	0.2756	0.052	5.288	0.000	0.173	0.378
room_type_reserved_Room_Type 2	-0.5085	0.132	-3.855	0.000	-0.767	-0.250
room_type_reserved_Room_Type 3	-0.0495	1.211	-0.041	0.967	-2.423	2.324
room_type_reserved_Room_Type 4	-0.1623	0.053	-3.084	0.002	-0.265	-0.059
room_type_reserved_Room_Type 5	-0.5010	0.206	-2.429	0.015	-0.905	-0.097
room_type_reserved_Room_Type 6	-1.0058	0.149	-6.743	0.000	-1.298	-0.713
room_type_reserved_Room_Type 7	-1.2870	0.308	-4.178	0.000	-1.891	-0.683
market_segment_type_Complementary	-54.6268	6.39e+10	-8.55e-10	1.000	-1.25e+11	1.25e+11
market_segment_type_Corporate	-1.1388	0.255	-4.467	0.000	-1.639	-0.639
market_segment_type_Offline	-2.0814	0.243	-8.554	0.000	-2.558	-1.604
market_segment_type_Online	-0.3463	0.240	-1.441	0.149	-0.817	0.125

Model Building – *Most Important Features That Need to BE ADDRESSED*

1. no_of_adults 1.33
2. no_of_children 2.00
3. no_of_weekend_nights 1.07
4. no_of_week_nights 1.10
5. required_car_parking_space 1.03
6. lead_time 1.20
7. repeated_guest 1.80
8. no_of_previous_cancellations 1.28
9. no_of_previous_bookings_not_canceled 1.56 avg_price_per_room 1.83
10. no_of_special_requests 1.24
11. week_of_year 1.02
12. type_of_meal_plan_Meal Plan 2 1.22
13. type_of_meal_plan_Meal Plan 3 1.03
14. type_of_meal_plan_Not Selected 1.24
15. room_type_reserved_Room_Type 2 1.10
16. room_type_reserved_Room_Type 3 1.00
17. room_type_reserved_Room_Type 4 1.35
18. room_type_reserved_Room_Type 5 1.03
19. room_type_reserved_Room_Type 6 2.01
20. room_type_reserved_Room_Type 7 1.12
21. market_segment_type_Complementary 4.02
22. market_segment_type_Corporate 15.62
23. market_segment_type_Offline 58.96
24. market_segment_type_Online 65.37

Variance Inflation factor:
Variance inflation factors measure the inflation in the variances of the regression coefficients estimates due to collinearity that exist among the predictors. It is a measure of how much the variance of the estimated regression coefficient β_k is "inflated" by the existence of correlation among the predictor variables in the model.

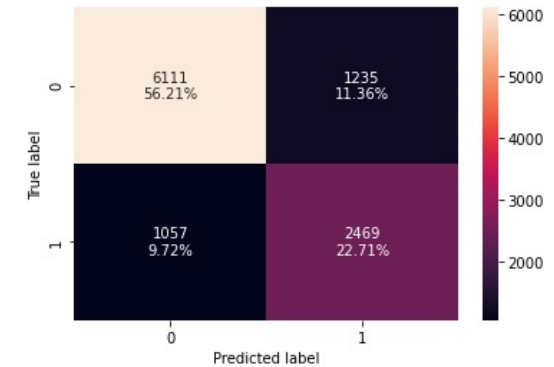
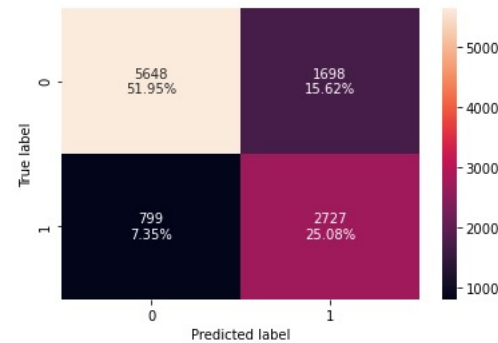
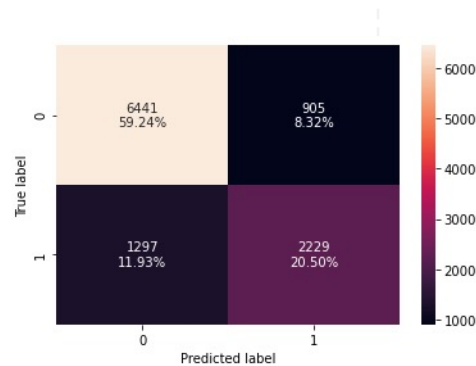
Model Building – *Variance Inflation factor Adjusted Results*

Adjusted Results should that all variables are significant and have P Value $s < .05$!!

Logit Regression Results						
Dep. Variable:	booking_status	No. Observations:	25366			
Model:	Logit	Df Residuals:	25346			
Method:	MLE	Df Model:	19			
Date:	Fri, 19 Nov 2021	Pseudo R-squ.:	0.3148			
Time:	12:56:43	Log-Likelihood:	-11014.			
converged:	True	LL-Null:	-16073.			
Covariance Type:	nonrobust	LLR p-value:	0.000			
	coef	std err	z	P> z	[0.025	0.975]
const	-3.9163	0.121	-32.379	0.000	-4.153	-3.679
no_of_adults	0.0919	0.037	2.491	0.013	0.020	0.164
no_of_children	0.2598	0.058	4.446	0.000	0.145	0.374
no_of_weekend_nights	0.1653	0.020	8.456	0.000	0.127	0.204
no_of_week_nights	0.0357	0.012	2.980	0.003	0.012	0.059
required_car_parking_space	-1.4982	0.132	-11.347	0.000	-1.757	-1.239
lead_time	0.0156	0.000	62.486	0.000	0.015	0.016
repeated_guest	-3.0949	0.577	-5.362	0.000	-4.226	-1.964
no_of_previous_cancellations	0.3061	0.075	4.062	0.000	0.158	0.454
avg_price_per_room	0.0182	0.001	26.020	0.000	0.017	0.020
no_of_special_requests	-1.4674	0.030	-49.504	0.000	-1.525	-1.409
week_of_year	-0.0036	0.001	-2.968	0.003	-0.006	-0.001
type_of_meal_plan_Not Selected	0.2779	0.052	5.342	0.000	0.176	0.380
room_type_reserved_Room_Type 2	-0.5137	0.132	-3.903	0.000	-0.772	-0.256
room_type_reserved_Room_Type 4	-0.1522	0.052	-2.914	0.004	-0.255	-0.050
room_type_reserved_Room_Type 5	-0.5289	0.204	-2.591	0.010	-0.929	-0.129
room_type_reserved_Room_Type 6	-1.0189	0.148	-6.863	0.000	-1.310	-0.728
room_type_reserved_Room_Type 7	-1.3194	0.307	-4.303	0.000	-1.920	-0.718
market_segment_type_Offline	-0.8737	0.096	-9.139	0.000	-1.061	-0.686
market_segment_type_Online	0.8321	0.093	8.986	0.000	0.651	1.014

Model Building – *Confusion Matrix & Train Test Comparison*

	Logistic Regression sklearn	Logistic Regression-0.33 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80	0.77	0.79
Recall	0.63	0.77	0.70
Precision	0.71	0.62	0.67
F1	0.67	0.69	0.68



- All the models are giving a generalized performance on training and test set.
- The highest recall is 77% on the training & had the same result on the testing data!!
- Using the model with default threshold the model will give a low recall but good precision scores - This model will help the hotel resell rooms but lose on potential customers.
- Using the model with 0.33 threshold the model will give a high recall but low precision scores - This model will help the hotel identify potential customers that may cancel but tmay not be precise and charge higher cancelation fees than they should.
- Using the model with 0.42 threshold the model will give a balance recall and precision score - This model will help the hotel maintain a balance in identifying potential customer that will cancel vs those that will not and charge the cancelation fee to those that shouldn't get it.

Step 1: We want to predict the the likelihood of cancelation.

- Before we proceed to build a model, we'll have to encode categorical features.
- We'll split the data into train and test to be able to evaluate the model that we build on the train data.
- We will build a model using the train data and then check it's performance.

Shape of Training set : (25366, 25)

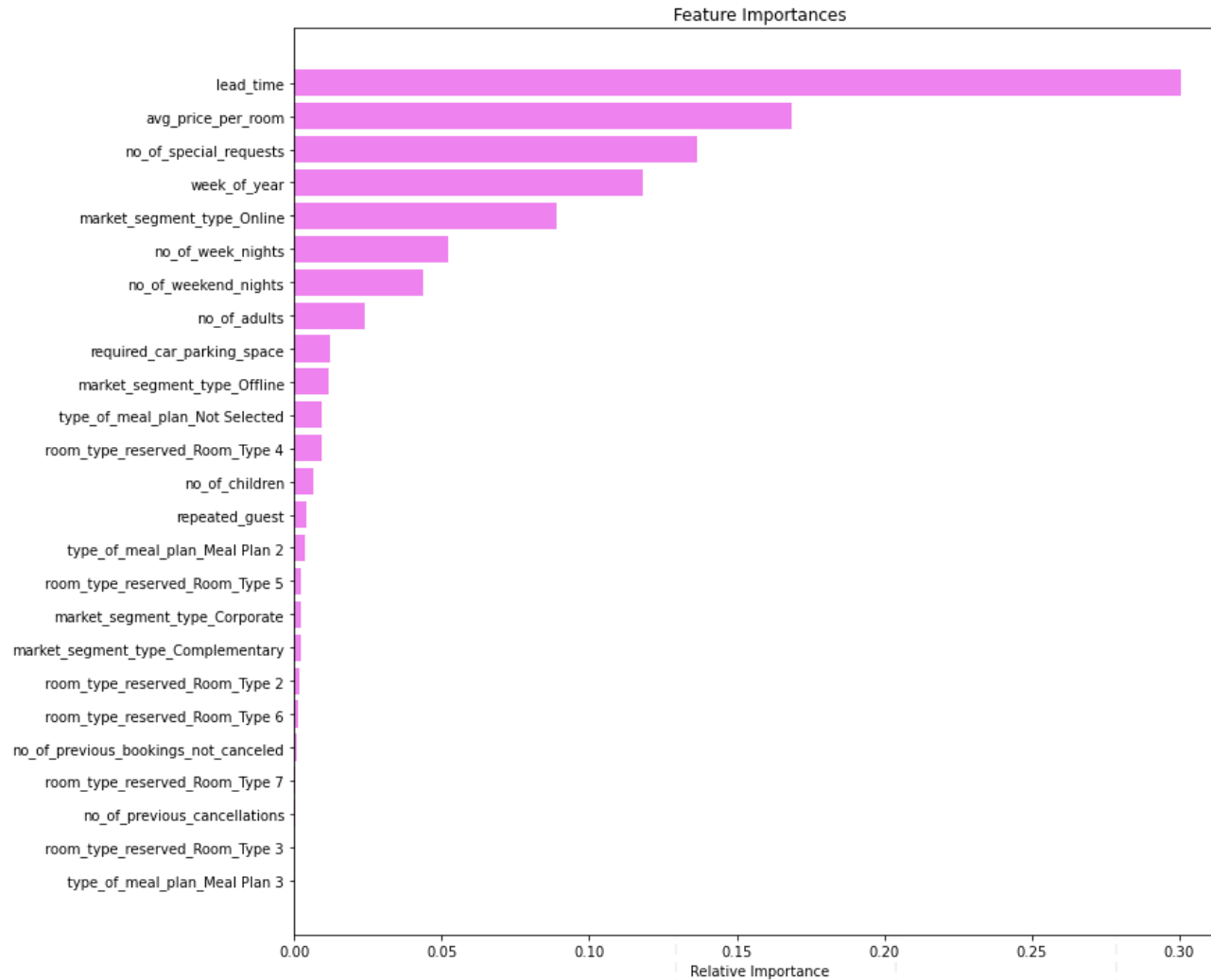
Shape of test set : (10872, 25)

Training set performance:
Recall: 0.99

Test set performance:
Recall: .79

- * Model is able to almost perfectly classify all the data points on the training set @ 99.7%.
- * 0 errors on the training set, each sample has been classified correctly.
- * As we know a decision tree will continue to grow and classify each data point correctly if no restrictions are applied as the trees will learn all the patterns in the training set.
- * This generally leads to overfitting of the model as Decision Tree will perform well on the training set but will fail to replicate the performance on the test set.

Model Building – *Most Important Features That Need to BE ADDRESSED*

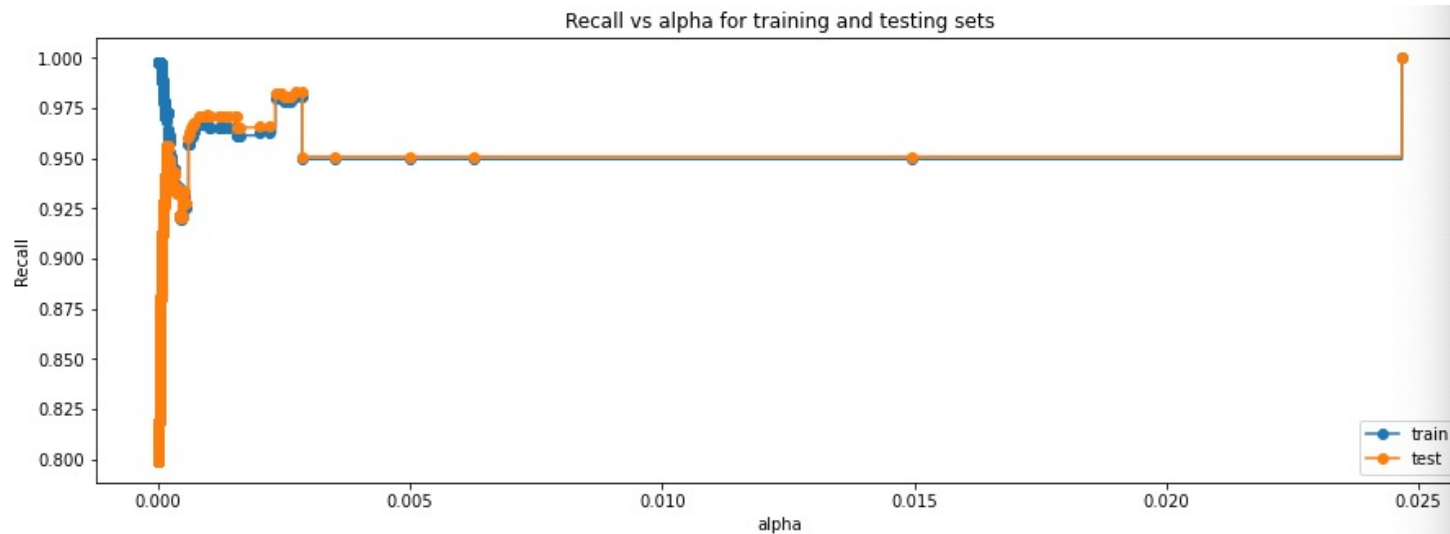


The first tree nearly fits every possible combination so is not insightful.

However, looking at the sense of what variables are important starts to give us a sense of what is, but further analysis is needed.

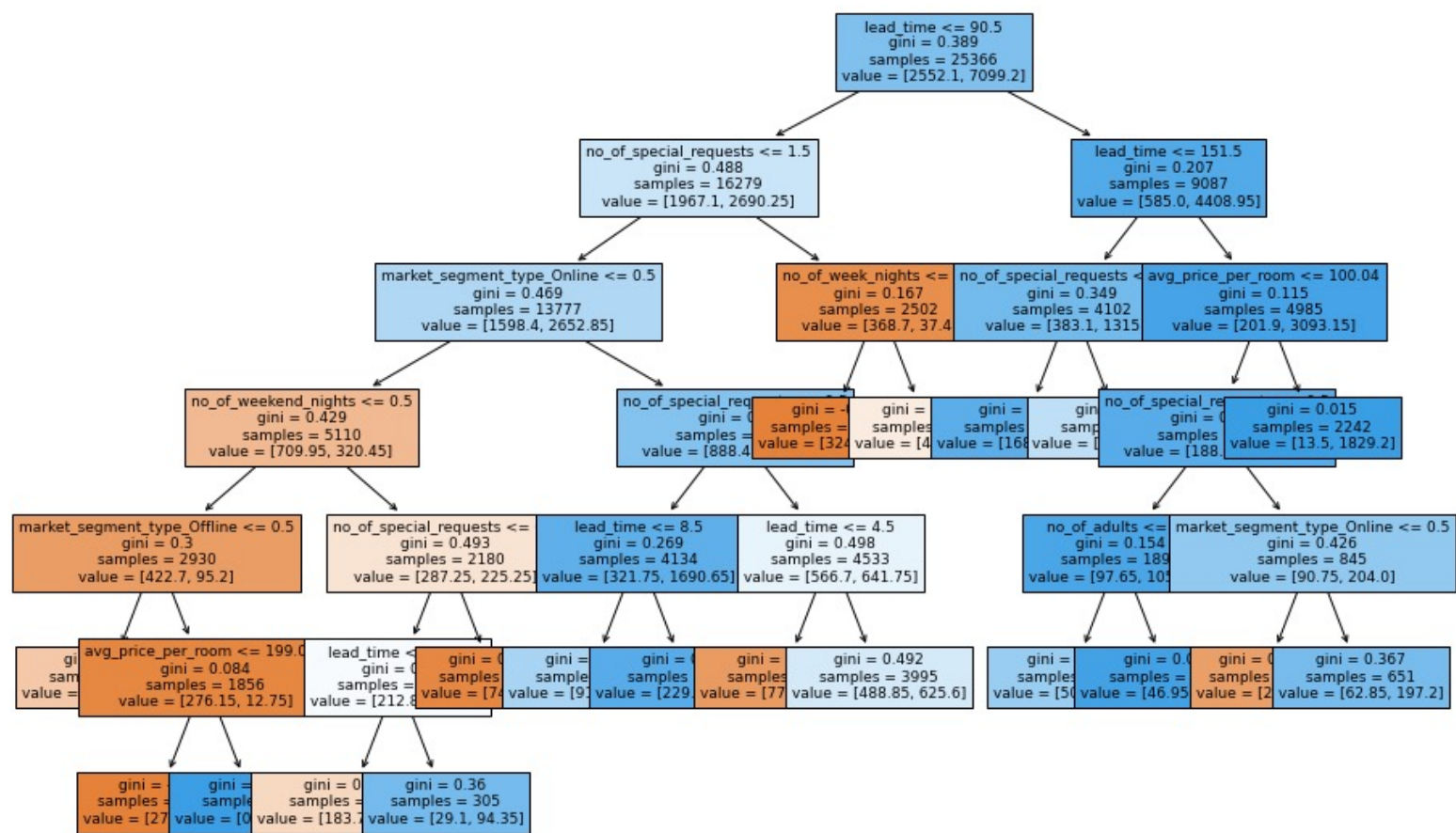
Model Building – OVER FITTING & Using GridSearch for Hyperparameter tuning to address it.

1. ***Start by the most generalized results ... on node. BUT WE CANT USE IT AS ITS TOO SIMPLE!!***
2. ***Lets change the number of nodes & alpha to find the optimal spot.***



3. **Maximum value of Recall is at 0.025 alpha, but if we choose decision tree will only have a root node and we would lose the buisness rules, instead we can choose alpha 0.002~3 retaining information and getting higher recall.**

Model Building – The optimal tree looks like this!



Model Building – The “tree” is hard to read, this is an easier way to follow the path of each segment.

```
--- lead time <= 90.50
  --- no_of_special_requests <= 1.50
    --- market segment type Online <= 0.50
      --- no_of_weekend_nights <= 0.50
        --- market_segment_type_Offline <= 0.50
          --- weights: [146.55, 82.45] class: 0
          --- market segment type Offline > 0.50
            --- avg_price_per_room <= 199.01
              --- weights: [276.00, 0.00] class: 0
              --- avg_price_per_room > 199.01
                --- weights: [0.15, 12.75] class: 1
            --- no_of_weekend_nights > 0.50
              --- no_of_special_requests <= 0.50
                --- lead_time <= 65.50
                  --- weights: [183.75, 126.65] class: 0
                  --- lead_time > 65.50
                    --- weights: [29.10, 94.35] class: 1
              --- no_of_special_requests > 0.50
                --- weights: [74.40, 4.25] class: 0
```

<<< This segment is
likely to cancel

Model Building – *Decision Tree Confusion Matrix & Train Test Comparison*



- All the models are giving a generalized performance on training and test set.
- The highest recall is 77% on the training & had the same result on the testing data!!
- Using the model with default threshold the model will give a low recall but good precision scores - This model will help the hotel resell rooms but lose on potential customers.
- Using the model with 0.33 threshold the model will give a high recall but low precision scores - This model will help the hotel identify potential customers that may cancel but tmay not be precise and charge higher cancelation fees than they should.
- Using the model with 0.42 threshold the model will give a balance recall and precision score - This model will help the hotel maintain a balance in identifying potential customer that will cancel vs those that will not and charge the cancelation fee to those that shouldn't get it.

Model Recommendation

While both models produce results, if you consider comparing the Logistic Regression (Statsmodel & Sklearn) the following observations can be made:

Logistic

1. Logistic requires a lot more effort in terms of data cleaning and observation, which can be time consuming and error prone.
2. Logistic regression is much faster in terms of how quickly the analysis can take place based on current environment
3. Logistic regression is easier to understand in terms of the coefficient meaning +/- and it's impact on the dependent variable
4. There performance was OK ... changing the threshold balanced the recall and precision resulting in a F1 score of .69
5. The main gap I see is there is still quite a bit of explainability that can not be discovered with the data we have.

Decision Tree

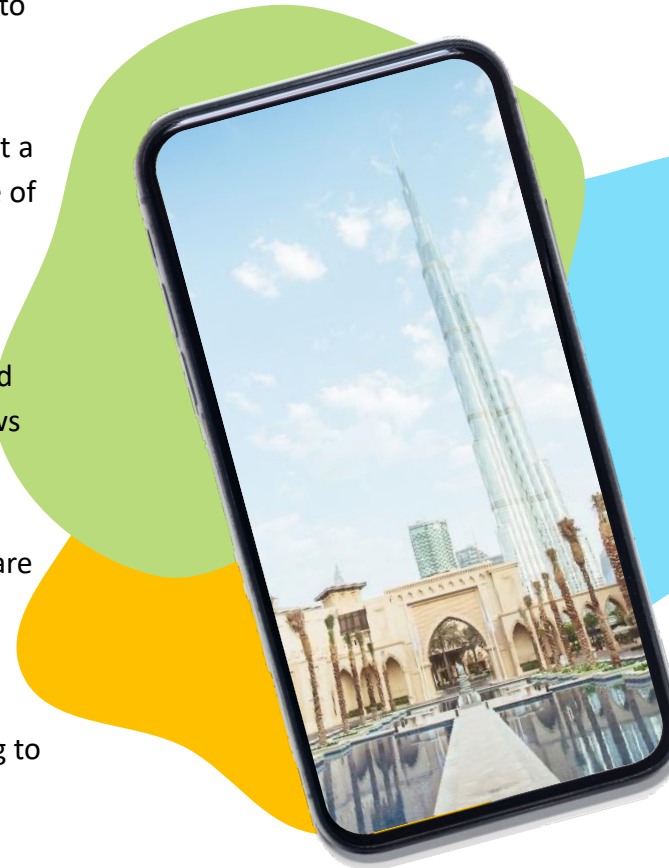
1. The results would "appear" initially to be better, but that is a red herring b/c the model is overfitting on the training data.
2. Hyperparameter tuning does appear to have very good results on both test and train, but may still be overfitting
3. The **post-tuned tree is the model I would select**, as it has performance better than that of the logistic model in terms of closing the recall gap, and amongst the other options of the decision tree.
4. The only draw back is that it is computationally intensive and the results graphically are difficult to understand (vs coefficients) which may be a personal level of comfort with the latter.

1. What also needs to be considered is the production environment. If stable, meaning the data comes in with known defects and remains stable possibly the Logistic model can be used as a second check where the delta between the two becomes key metric or warning signal if something changes.
2. As for the decision tree, computationally the environment would need to be sized accordingly to ensure performance isn't an issue. If it is for some reason not feasible to run the DT in real-time ... use the Logistic model to screen the "easy" decisions and the DT to run a second pass against those that are not clearly candidates for approval.

Business Insights & Recommendations

Recap Observations

1. We analyzed the "INNHOTEL" using different techniques and used Decision Tree Classifier to build a predictive model for the same.
2. The model built can be used to predict if a customer is going to cancel their reservation or not.
3. We visualized different trees and their confusion matrix to get a better understanding of the model. Easy interpretation is one of the key benefits of Decision Trees.
4. We verified the fact that how much less data preparation is needed for Decision Trees and such a simple model gave good results even with outliers and imbalanced classes which shows the robustness of Decision Trees.
5. Lead Time, 2 Special Request, Online, and 1 Special Request are the most important variables in predicting customer cancellation or not.
6. We established the importance of hyper-parameters/ pruning to reduce overfitting.




Recommendations

While the decision Trees allows us to make even more discrete A/B tests compared to the logistic model, the basic principles are the same:

1. To drive the likelihood of decreasing cancellations build pricing and programs around:
2. Build incentive programs/price reductions to become a "repeated guest" to build loyalty
3. room customization / special requests / pillows to other types beyond the basic categories we have
4. upgrade @ a minimal cost of "free" to categories 6 & 7
5. discounted parking in advance required_car_parking_space
6. Online booking is barrier free, and most of the cancellations come from that segment:
7. consider limiting the number of cancellations allowed w/in a 6month period
8. consider lower cost, but non-refundable reservations
9. consider implementing % of fee based for canceling an online reservation
10. consider allowing reservations up-to > 90 to be changed or canceled, with an increasing cancelation fee as the date is < 90days. Within 7days consider "no refund"

Business Insights & Recommendations

Examples - each of these likely buckets of folks that will cancel could be target through a/b to test various levels of fees (incentives)

1. Implement Short Term Cancellation Fees For Any Reservation Made ≤ 90 Days.
 2. For All Corporate Reservations That Are On A Weekend, Implement A Cancellation Fees
 3. For Online Reservations Where The Lead Time Is Lead-time ≤ 8.50 , Implement A Non Refundable "Short Term" Premium Fee Equal To 40% Of The Total Stay And/Or Refund Only 60%.
 4. For Example, For Rooms $> \$200$ Consider Increasing The Cancellation Fee.
- 

• • •
• • •
• • •
• • •
• • •
• • •
• • •
• • •
• • •
• • •
• • •

Thank You

John Noble

617.519.9065

