




# HOTEL BOOKING PREDICTION

DSC-550 Case Study Documentation

Vinay Nagaraj  
Bellevue University



## Abstract:

Due to recent events in the world, the hotel industry has taken a major setback on its business. Hotel room booking and subsequent cancellations are causing Hotel owners the constant headache of losing business.

This case study looks at a dataset comprised of data pertaining to Hotel booking and their attributes. Through this study we will try to look at what factors would be available to the Hotel owners which they could use and predict the outcome of a booking. Utilizing the insights gained by this analysis the Hotel owners could better plan their business and see how they can profit in this tough times. Using the features from each hotel booking, my model will predict whether the hotel booking will be cancelled or not.

For my analysis I have used the dataset downloaded from Kaggle.

Kaggle link: <https://www.kaggle.com/jessemotipak/hotel-booking-demand>

This data set contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things.

## Approach:

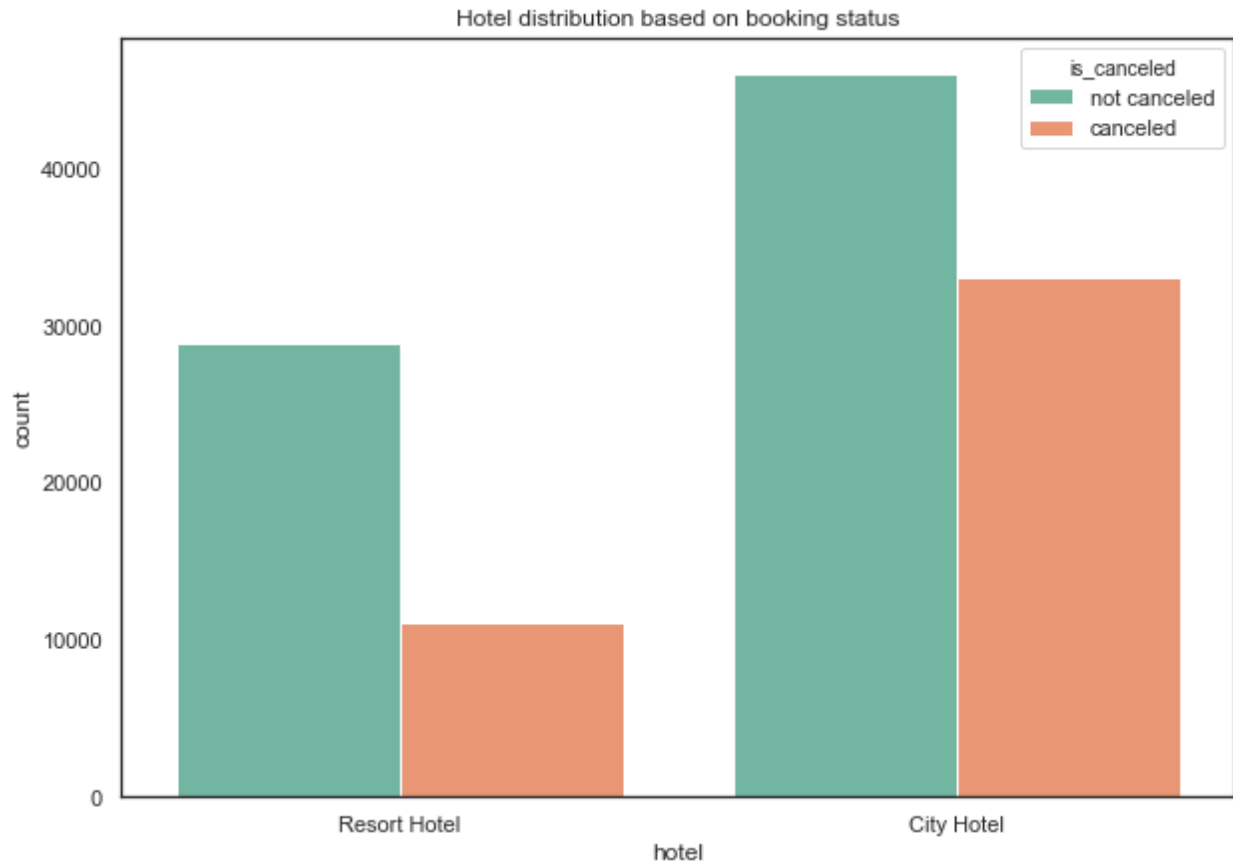
The aim here is to create meaningful estimators from the data set we have and to select the model that predicts the cancellation best by comparing them with the accuracy scores of different ML models and ROC Curves. Below are the steps:

1. Graph Analysis: Here we will understand a lot of aspects around this dataset through graphs. Hotel distribution based on booking status, Number of reservations per year along with its status, Average number of guests per month and some histograms & barchart to understand the dataset. We will also create heat maps to understand the relationship between the attributes.
2. Feature Reduction: As part of Feature reduction steps, I have handled missing values and features. I have tried with VarianceThreshold and SelectKBest feature selection to extract the important features for the model.
3. Model Evaluation & Selection: In this part, I have run summary scores on a few models to see which is the better one. I selected Logistic Regression and Random Forest Classifier models and compared the accuracy of each model for Train dataset and Test dataset and produces comparison results.
4. Conclusion: Based on the results from the models, I will lay out my observations/findings.

## Part 1: Graph Analysis

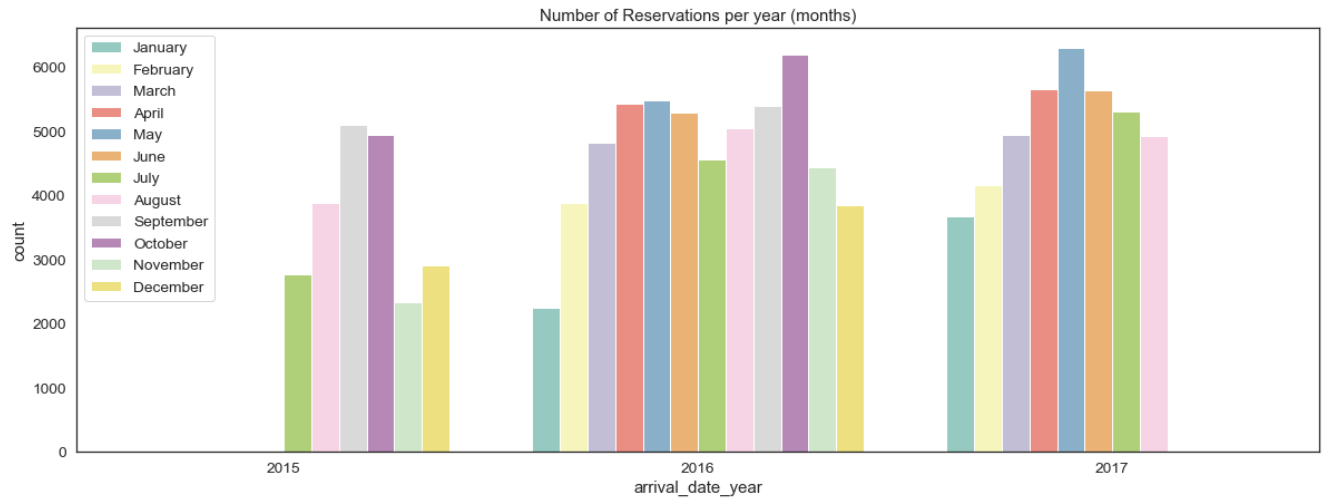
### Hotel Distribution Based on Booking Status:

Here we can see that the bookings are more in City Hotels but the cancellation percentage is less in Resort Hotels.



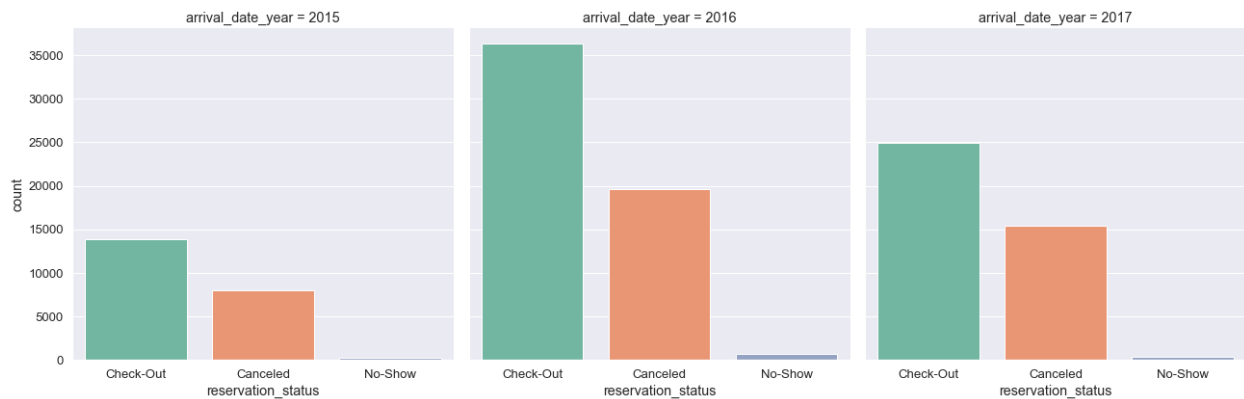
### Number of reservations per year (Month wise):

Summers are usually the time that families tend to go out for holidays and we can notice that below as the months in-between have more bookings than from Nov-Feb.



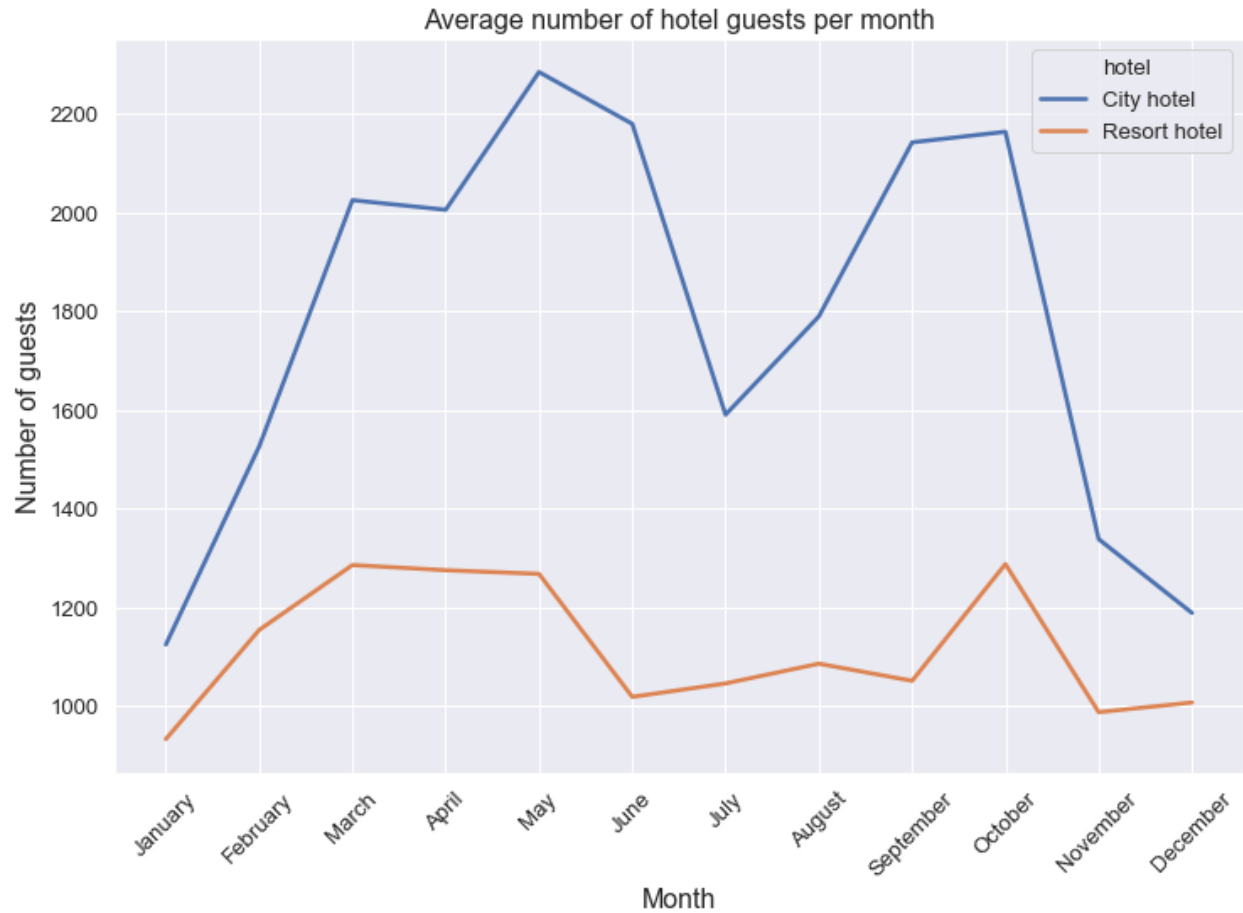
### Reservation status over the years:

No-Show has been very minimal over the years but surely the Hotels would need to look at the high number of cancelations.

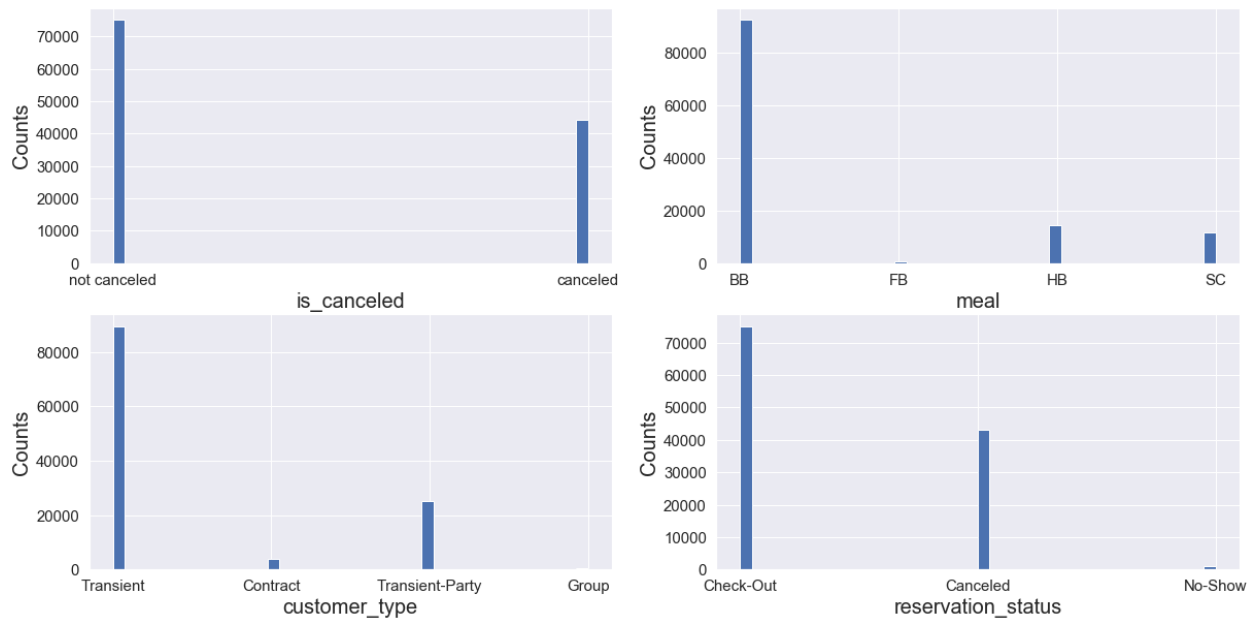


### Avg number of Hotel guests per month:

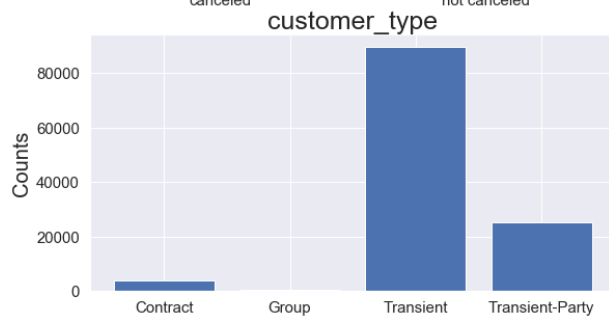
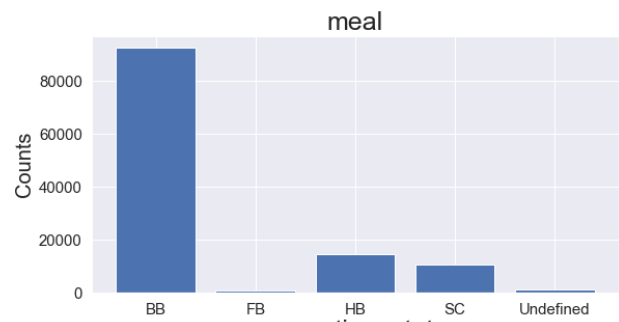
May, June, September & October seems to be the busiest months for the Hotel Industry.



**Histograms:** is\_canceled, meal, customer\_type, reservation\_status

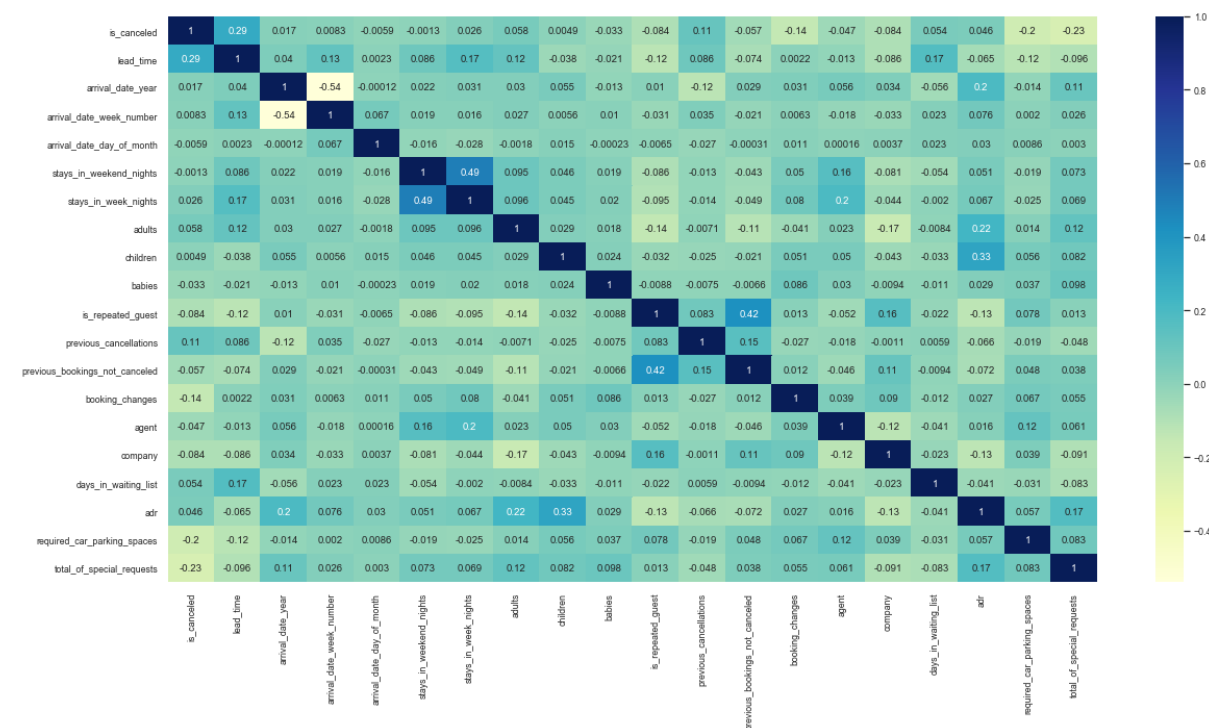


**Barchart:** is\_canceled, meal, customer\_type, reservation\_status

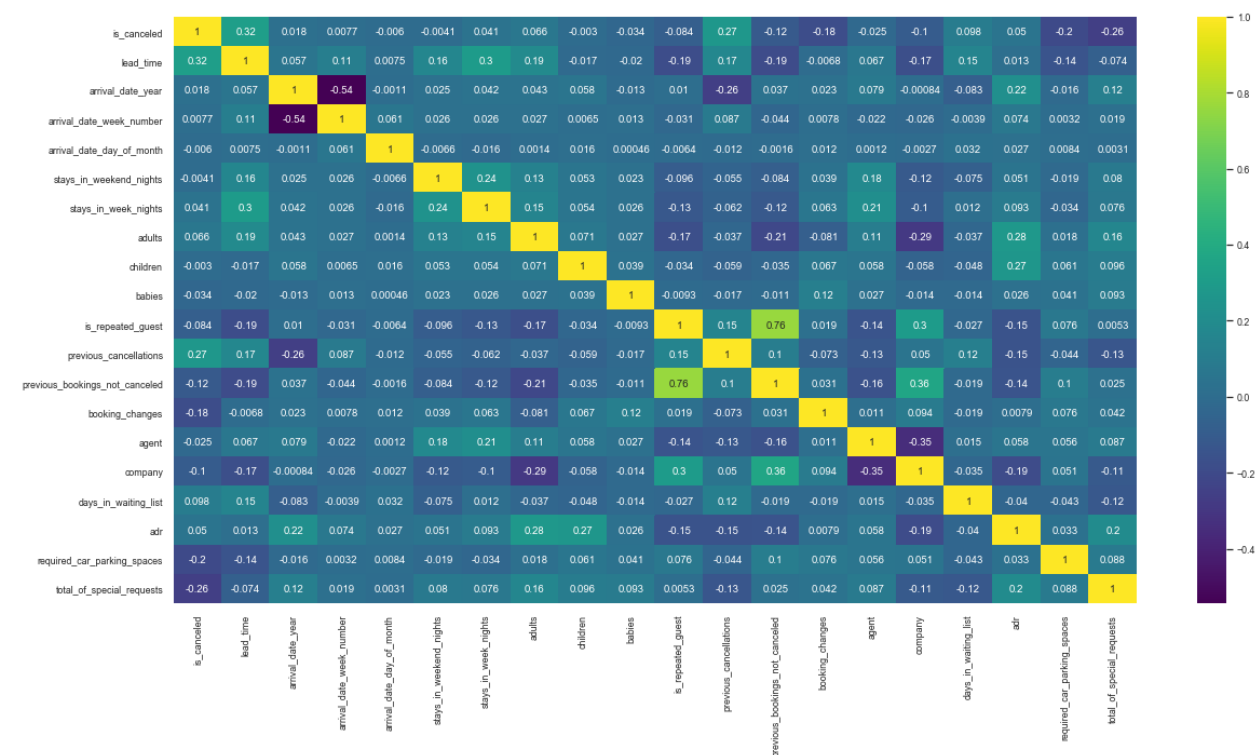


## Part 2: Feature Reduction

### Pearson Correlation Heatmap:



### Spearman Correlaton Heatmap:



Based on the Pearson Correlation Heatmap & Spearman Correlation Heatmap, we can clearly see that 'lead\_time' has a stronger connection with the 'is\_canceled' column.

As part of Feature reduction steps, I decided to create a new column with Total Guests instead of having 3 columns for adults, children and babies. After that I tried with VarianceThreshold and SelectKBest feature selection to extract the important features for the model.

Variance threshold is calculated based on probability density function of a particular distribution. If a feature has 95% or more variability then is very close to zero and the feature may not help in the model prediction and it can be removed. The values with True are the features selected using Variance threshold technique. The columns hotel, arrival\_date\_year, is\_repeated\_guest, booking\_changes, deposit\_type & required\_car\_parking\_spaces are removed.

The values with True are the features selected using SelectKBest technique. Most relevant 10 features are selected. The features selected can be tested by running through the model.

	Feature_Name	Score
11	deposit_type	26849.743593
1	lead_time	8917.683060
17	total_of_special_requests	5593.493295
16	required_car_parking_spaces	3730.453374
10	booking_changes	2034.273401
0	hotel	1810.499652
8	previous_cancellations	1184.390357
7	is_repeated_guest	668.337403
13	company	651.244405
9	previous_bookings_not_canceled	312.704213



## Part 3: Model Evaluation and Selection

In the Feature Selection part, Variance threshold had returned 14 features. I have used this training and test data for further process. I have run model scores across RandomForestClassifier, DecisionTreeClassifier, SGDClassifier and LogisticRegression. Below is a snapshot of the scores.

```
Out[46]:
```

	roc_auc
RandomForestClassifier	0.913729
DecisionTreeClassifier	0.793289
SGDClassifier	0.749051
LogisticRegression	0.759874

From the above scores, RandomForestClassifier is the best algorithm for this dataset.

The dependent variable for my models is "is\_canceled" and I am trying to predict the possibility of a booking. Whether a certain booking would be canceled or not.

1 - Canceled

0 - Not Canceled

### Random Forest Model Evaluation:

Confusion Matrix

```
[[13946  1012]
 [ 2354  6530]]
```

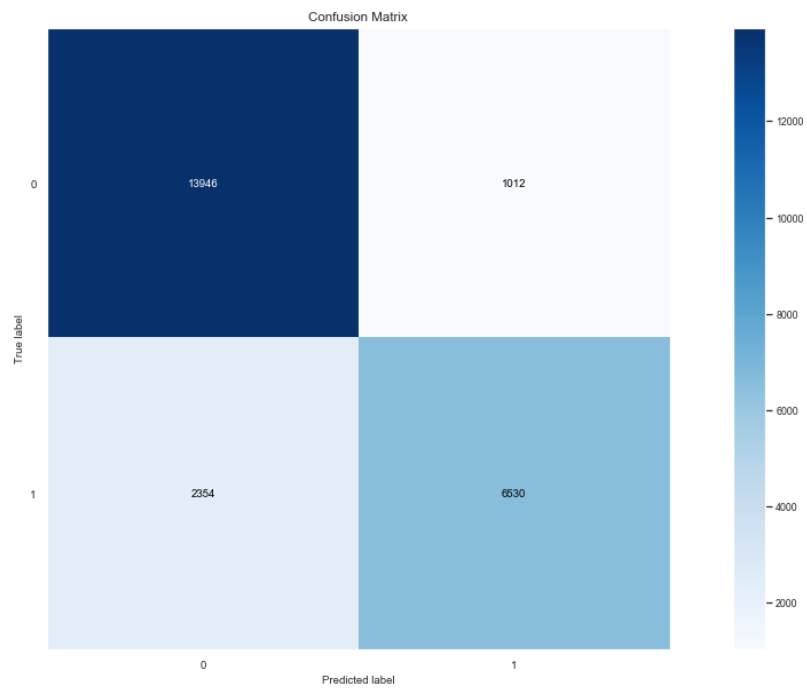
Classification report

	precision	recall	f1-score	support
0	0.85558	0.93234	0.89232	14958
1	0.86582	0.73503	0.79508	8884
accuracy			0.85882	23842
macro avg	0.86070	0.83369	0.84370	23842
weighted avg	0.85940	0.85882	0.85608	23842

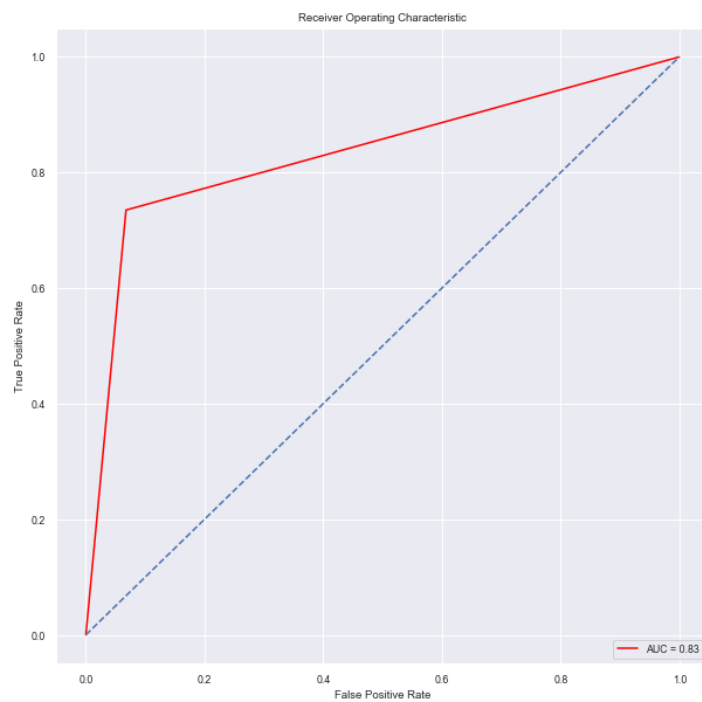
Scalar Metrics

AUROC = 0.91734

## Random Forest Confusion Matrix:



## Random Forest ROC AUC:



**Logistic Regression Model Evaluation:**

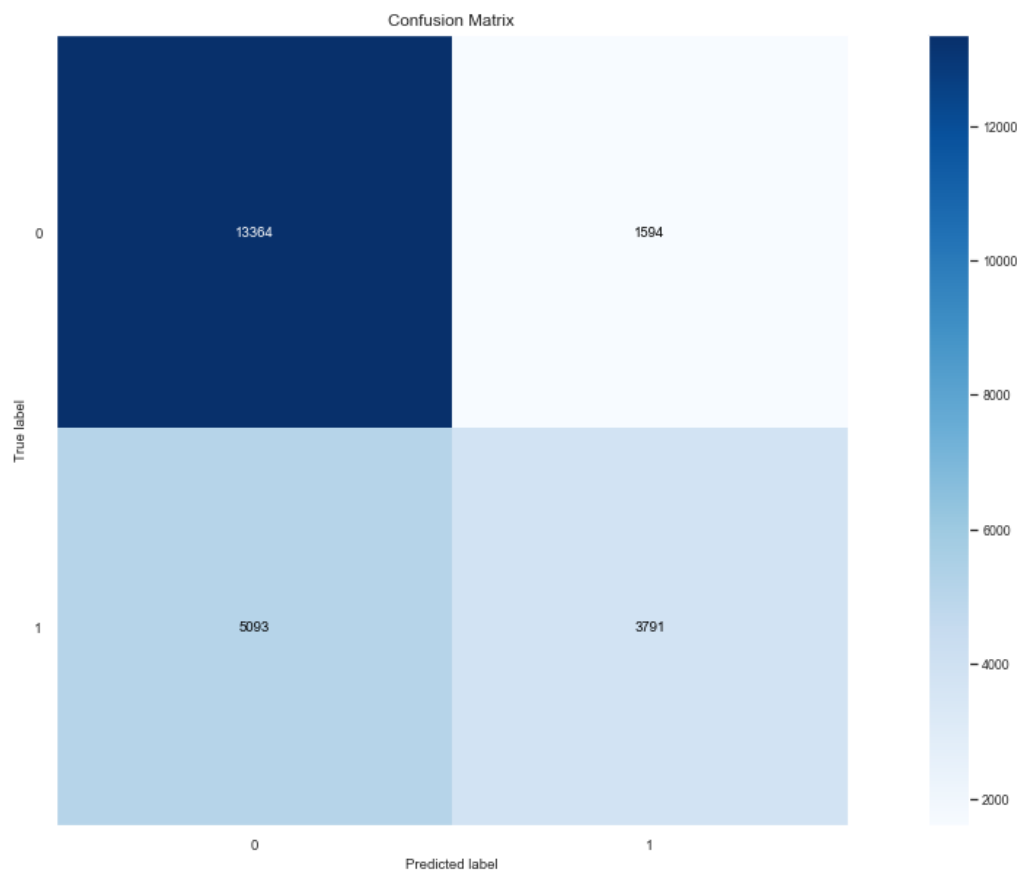
Confusion Matrix  
[[13364 1594]  
[ 5093 3791]]

Classification report

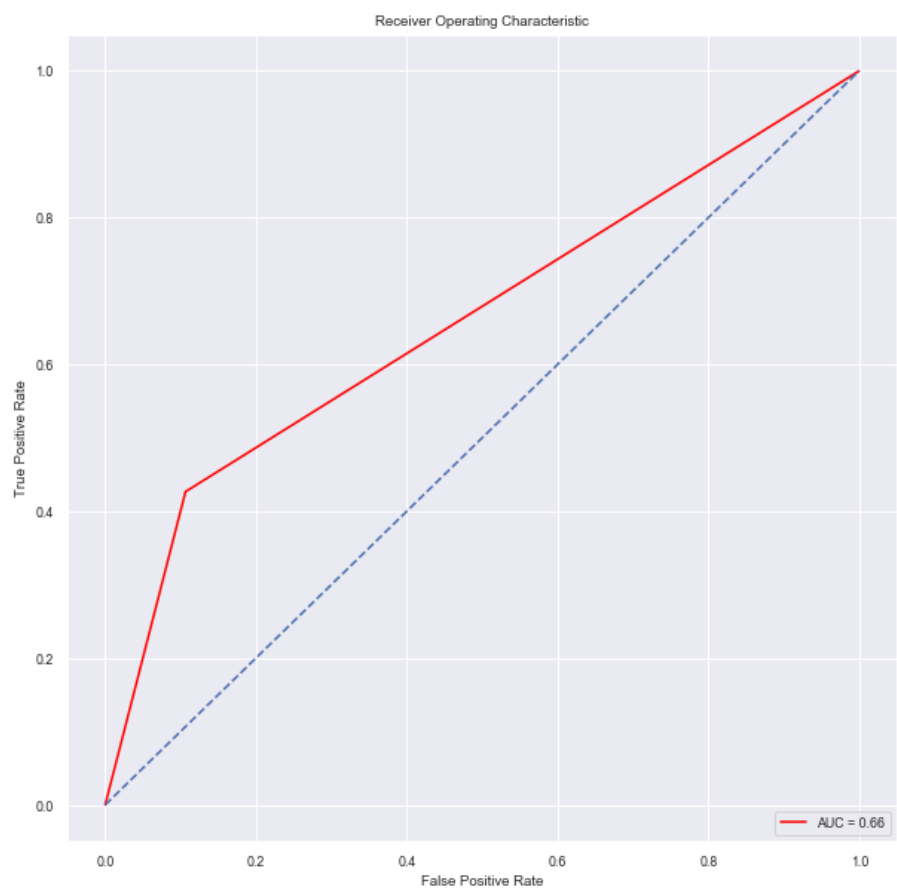
	precision	recall	f1-score	support
0	0.72406	0.89343	0.79988	14958
1	0.70399	0.42672	0.53136	8884
accuracy			0.71953	23842
macro avg	0.71403	0.66008	0.66562	23842
weighted avg	0.71658	0.71953	0.69982	23842

Scalar Metrics  
AUROC = 0.76038

Logistic Regression Confusion Matrix:



Logistic Regression ROC AUC:



## Conclusion

Based on the above observations:

- lead\_time is one of the most importance feature. This means that the sooner the reservation is made compared to arrival time, the more likely it will be cancelled.
- Comparing the Random Forest model and Logistic Regression model I think Random Forest model is much better with this dataset than Logistic Regression mode.
- In the confusion matrix the random forest predicts the values for both classes more accurately than the logistic regression confusion matrix.
- The LogisticRegression model predicted the 1,594 Not Canceled bookings as Canceled bookings and 5,093 canceled bookings as Not Canceled.
- The RandomForestClassifier model predicted the 1,012 Not Canceled bookings as Canceled bookings and 2,354 canceled bookings as Not Canceled.
- In the precision recall and F1 score the Random forest model scored better on all three metrics for both classes. In the ROC AUC curve the random forest preforms much better.
- Overall, the Random Forest model out preforms the logistic regression model on every point.

Hotel Owners can surely use the Random Forest model to predict about a booking.