Hotel Cancellation Prediction

Thip Rattanavilay
Bellevue University
DSC550-T302 - Data Mining

Abstract

The purpose of this study is to provide new insights into the factors that influence cancellation behavior with respect to hotel bookings. The data are based on individual bookings drawn from a hotel reservation system database comprising of hotels provided by Kaggle.

Finding out what impact it has on hotels when a cancellation is made. Would a cancellation limit the production of accurate forecasts, overbooking, and strategies, which can also have a negative influence on revenue and reputation? This discovery will help the hotel owners and provide a prediction model that will unlock this information.

I have used the dataset downloaded from Kaggle.

Kaggle link: https://www.kaggle.com/jessemostipak/hotel-booking-demand

Introduction

A significant number of hotel bookings are not realized due to cancellations or no-shows. The typical reasons include sudden illnesses, accidents, schedule conflicts, unexpected (family) obligations, and natural catastrophes. Although the option to cancel a hotel booking (preferably at a low cost) is beneficial to presumptive hotel guests, it is a less desirable and possibly revenue-diminishing factor for hotel managers to deal with. Such losses are particularly high on last-minute cancellations, and hotel owners can not clearly predict the outcome.

Knowledge of cancellation behavior is relevant not only to hotel managers' predictions of future revenues and capacity utilization but also to their cancellation and pricing policies. However, obtaining such knowledge places high demands on the data source in question, which needs to be rich and (preferably) also correspond to the flow format of a hotel booking system or related databases drawn from enterprise resource planning (ERP) systems. Booking data drawn from hotel booking systems can be an essential pillar of destination management information systems that collect data from different sources. Utilizing the insights gained by this analysis, the Hotel owners could better plan their business and see how they can profit in these challenging times. Using the features from each hotel booking, my model will predict whether the hotel booking will be canceled or not.

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

Approach

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

- *1. Graph Analysis: Here, we will understand many aspects of this dataset through graphs. Hotel distribution based on booking status, number of reservations per year, status, Average number of guests per month, and some histograms & bar charts to understand the dataset. We will also create heat maps to understand the relationship between the attributes.*

- *2. Feature Reduction: As part of the Feature reduction steps, I have handled missing values and features. I have tried with VarianceThreshold and SelectKBest feature selection to extract the essential 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, compared each model's accuracy for the Train dataset and Test dataset, and produced comparable results.*

- *4. Conclusion: Based on the results from the models, I will lay out my observations/findings.*

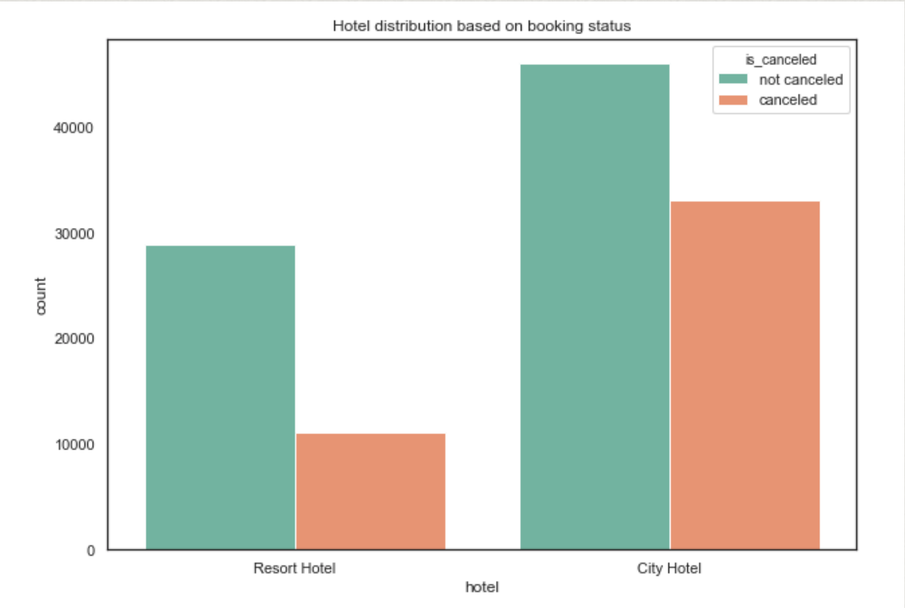
Part 1 - Graph Analysis

Here, we will understand many aspects of this dataset through graphs. Hotel distribution based on booking status, number of reservations per year, status, Average number of guests per month, and some histograms & bar charts to understand the dataset. We will also create heat maps to understand the relationship between the attributes.

Hotel distribution based on booking status

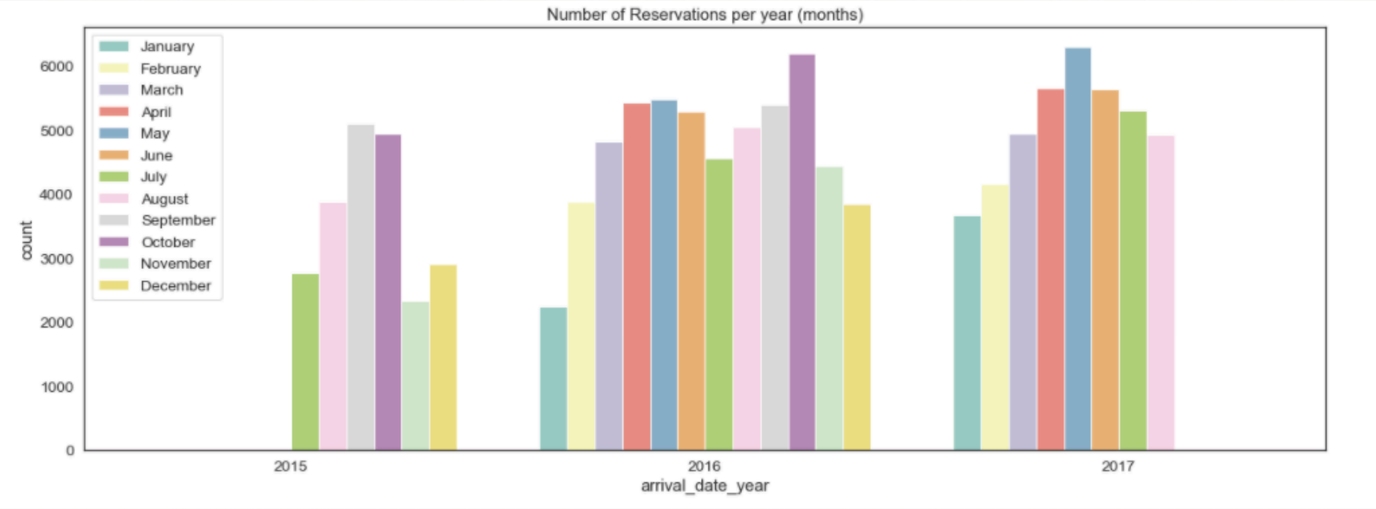
We can see that the bookings are more in City Hotels, but the cancelation percentage is much less at the Resort Hotels.

Green shows "not canceled"
Orange shows "canceled"



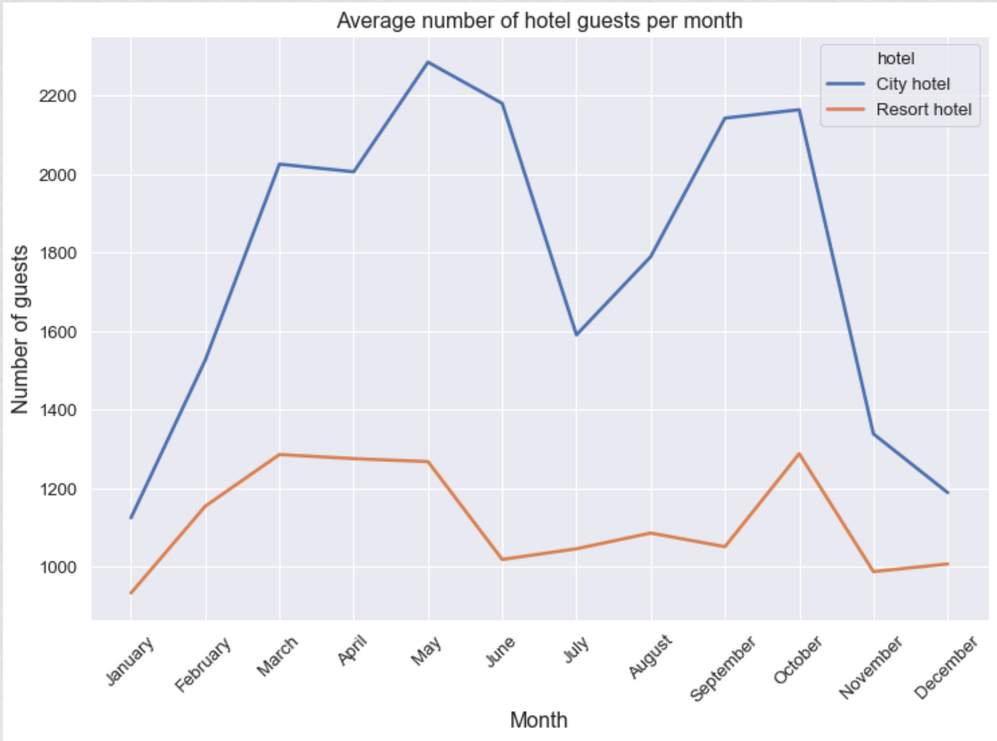
The number of reservations per year by month

Summers are usually when families tend to go out for holidays and notice that below as the months in-between has more bookings than Nov-Feb.

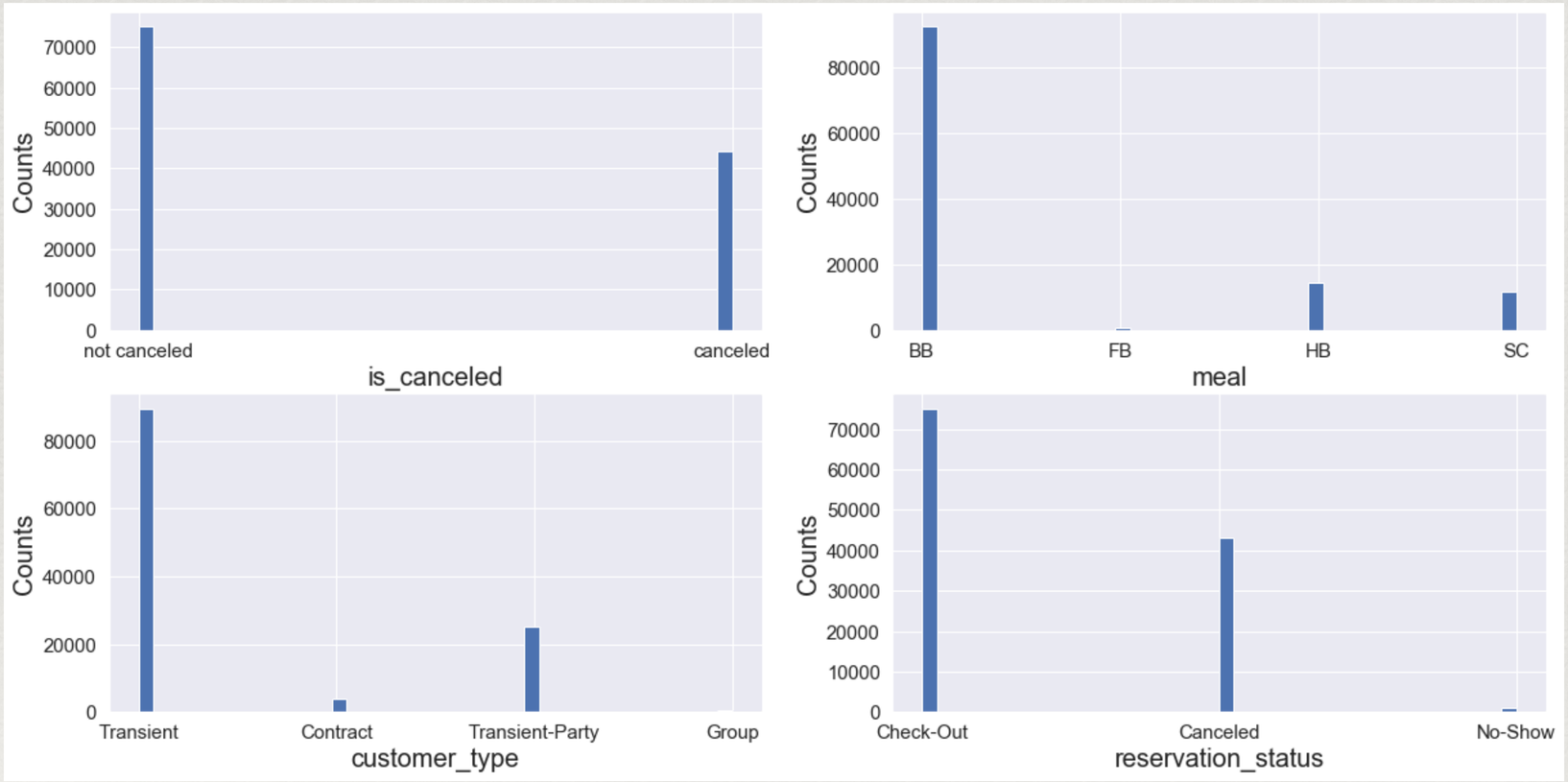


Average number of Hotel guest per month.

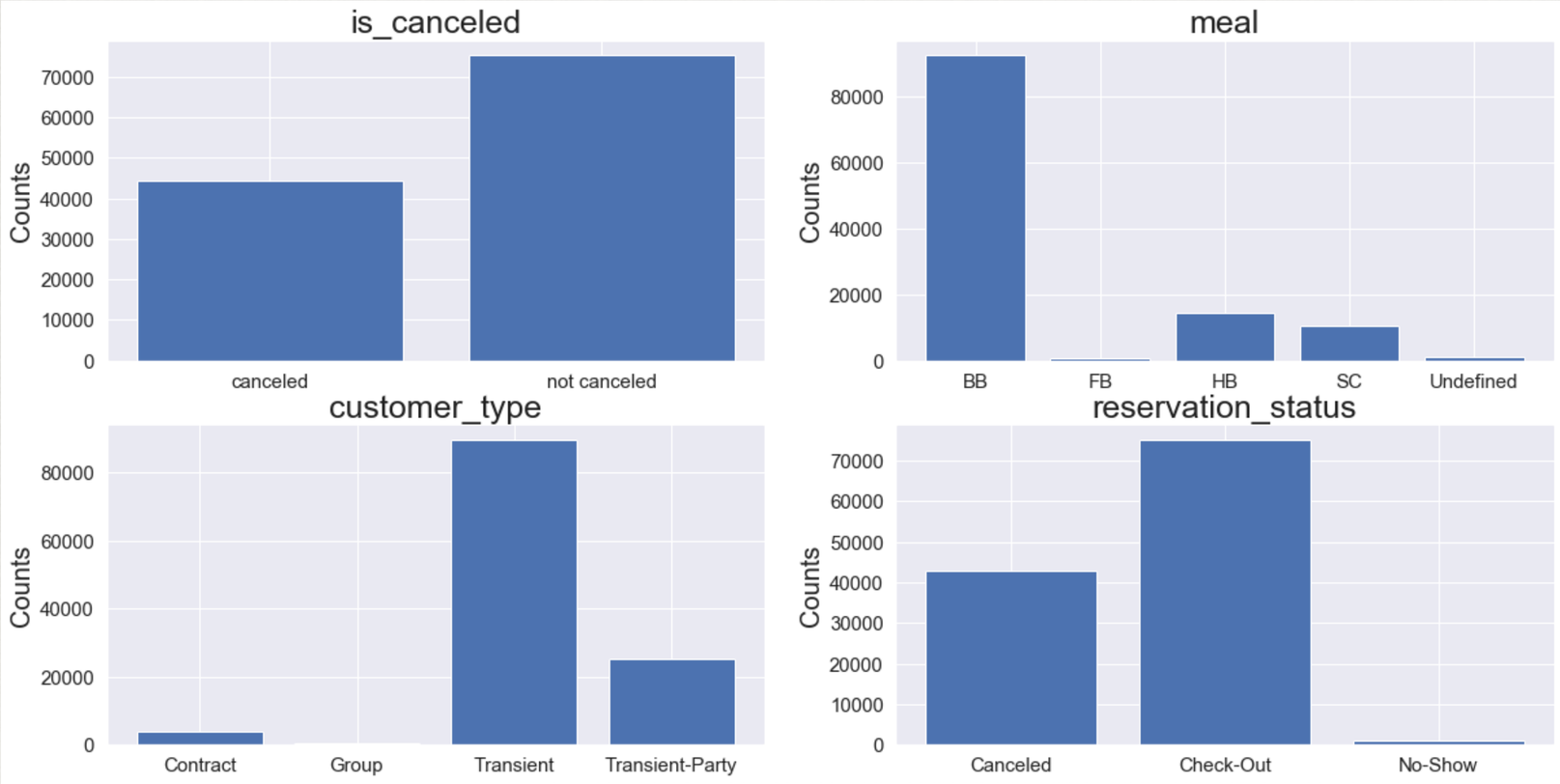
As you can see, May, June, September & October seem to be the busiest months for the Hotel Industry.



Histograms - *is_canceled, meal, customer type, reservation_status.*



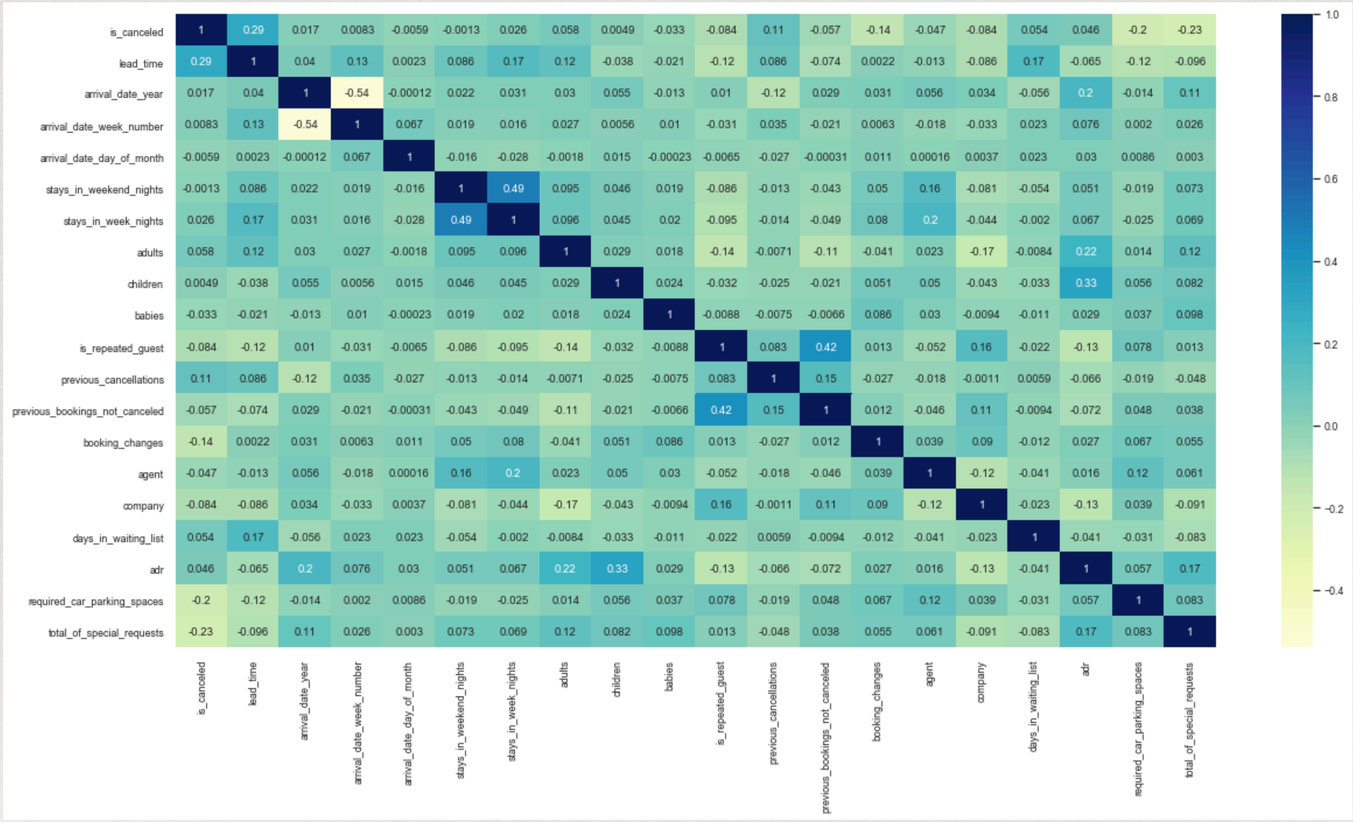
Barchart - *is_canceled, meal, customer type, reservation_status*



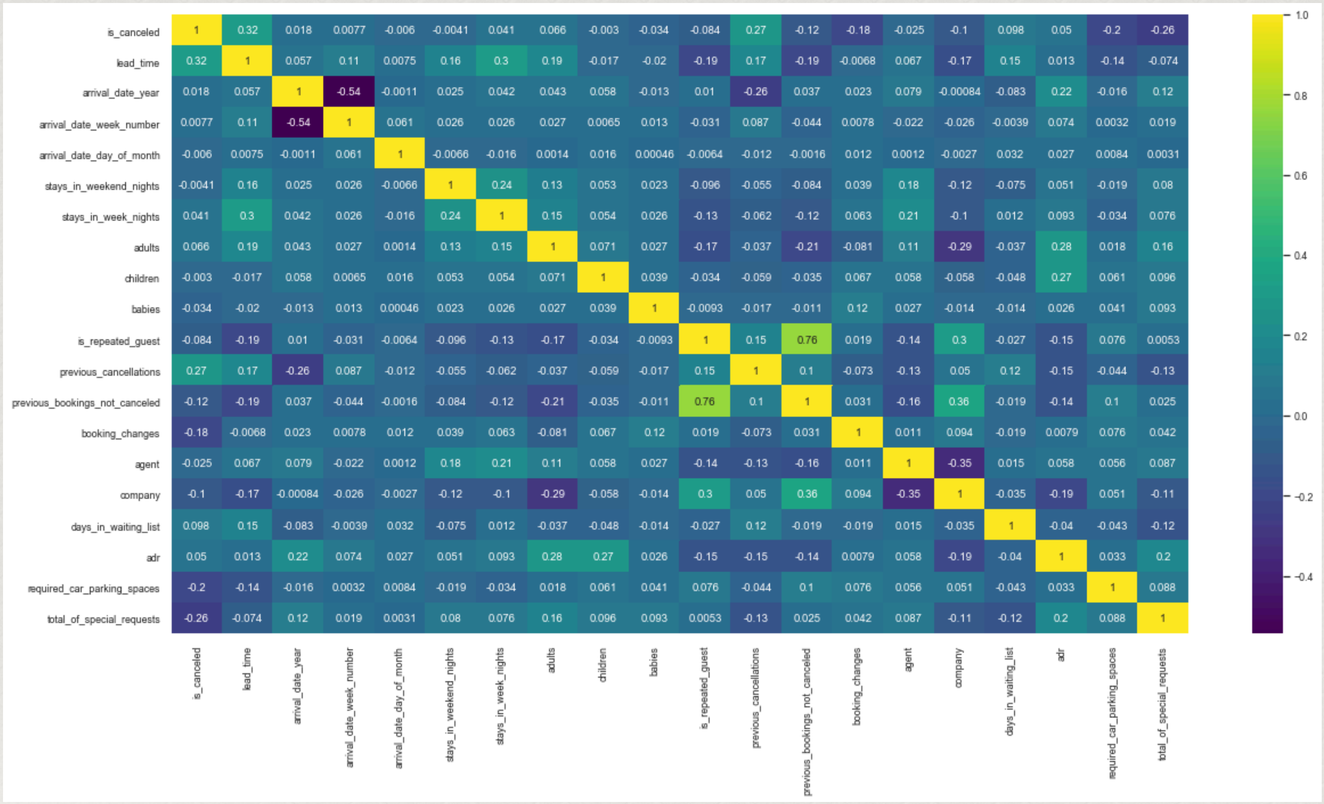
Part 2 - Feature Reduction

As part of the Feature reduction steps, I have handled missing values and features. I have tried with VarianceThreshold and SelectKBest feature selection to extract the essential features for the model.

Pearson Correlation Heat map



Spearman Correlation Heat map



Based on the Pearson Correlation Heat map & Spearman Correlation Heat map, we can clearly see that 'lead_time' has a stronger connection with the 'is_canceled' column.

As part of the 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 essential features for the model.

The variance threshold is calculated based on the probability density function of a particular distribution. If a feature has 95% or more variability, then it's very close to zero, and the feature may not help in the model prediction and can be removed. The values with True are the features selected using the 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 the SelectKBest technique. The most relevant 10 features are selected. The chosen features 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 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, compared each model's accuracy for the Train dataset and Test dataset, and produced comparable results.

In the Feature Selection part, the 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.

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 specific booking would be canceled or not.

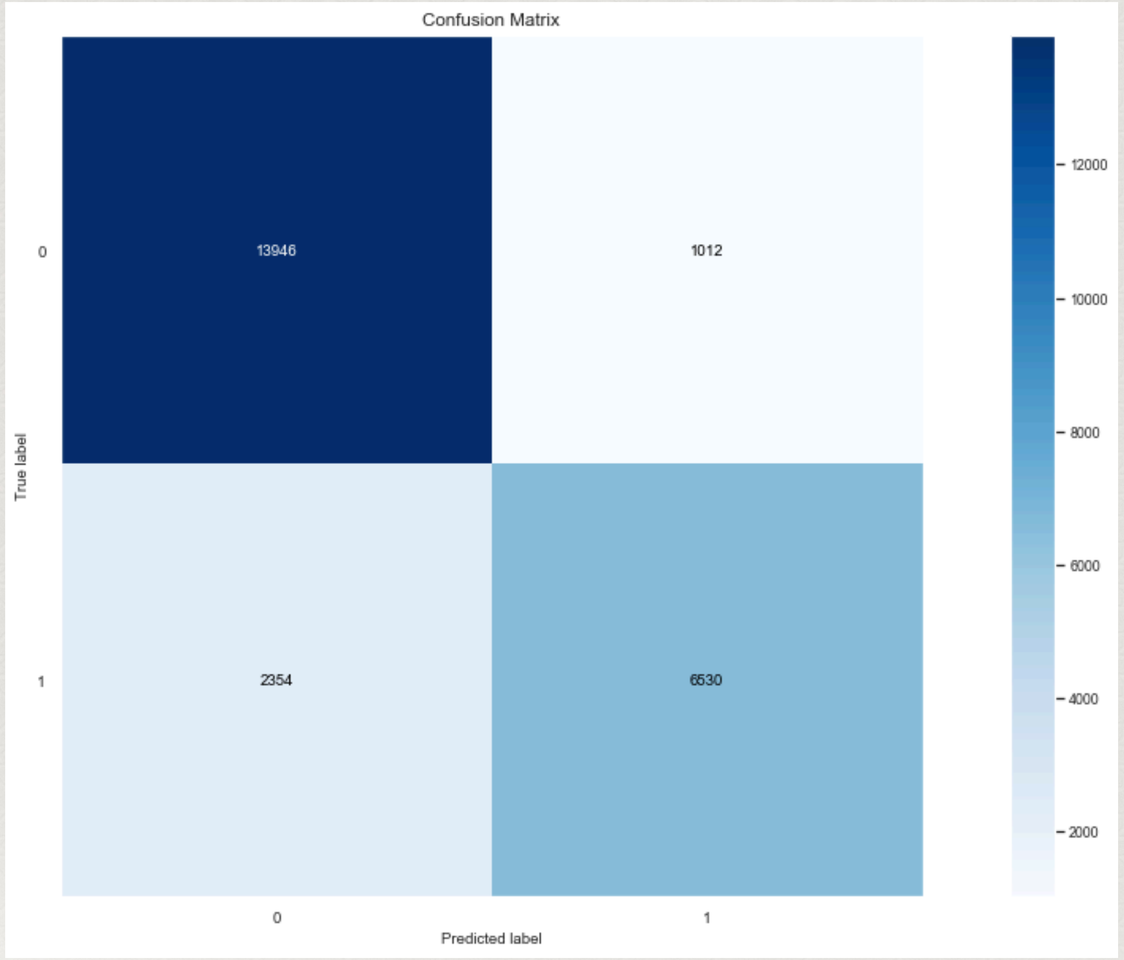
1 - Canceled
0 - Not Canceled

	roc_auc
RandomForestClassifier	0.913100
DecisionTreeClassifier	0.793645
SGDClassifier	0.750435
LogisticRegression	0.759874

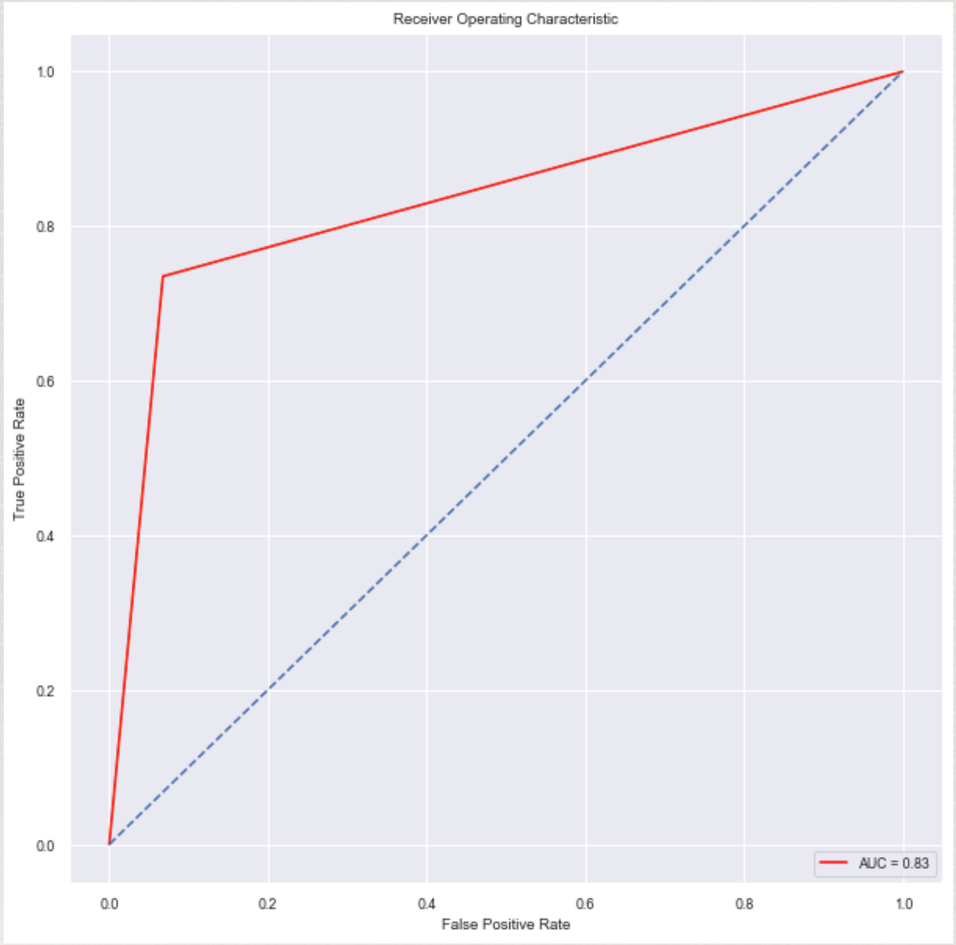
Random Forest Model Evaluation

```
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



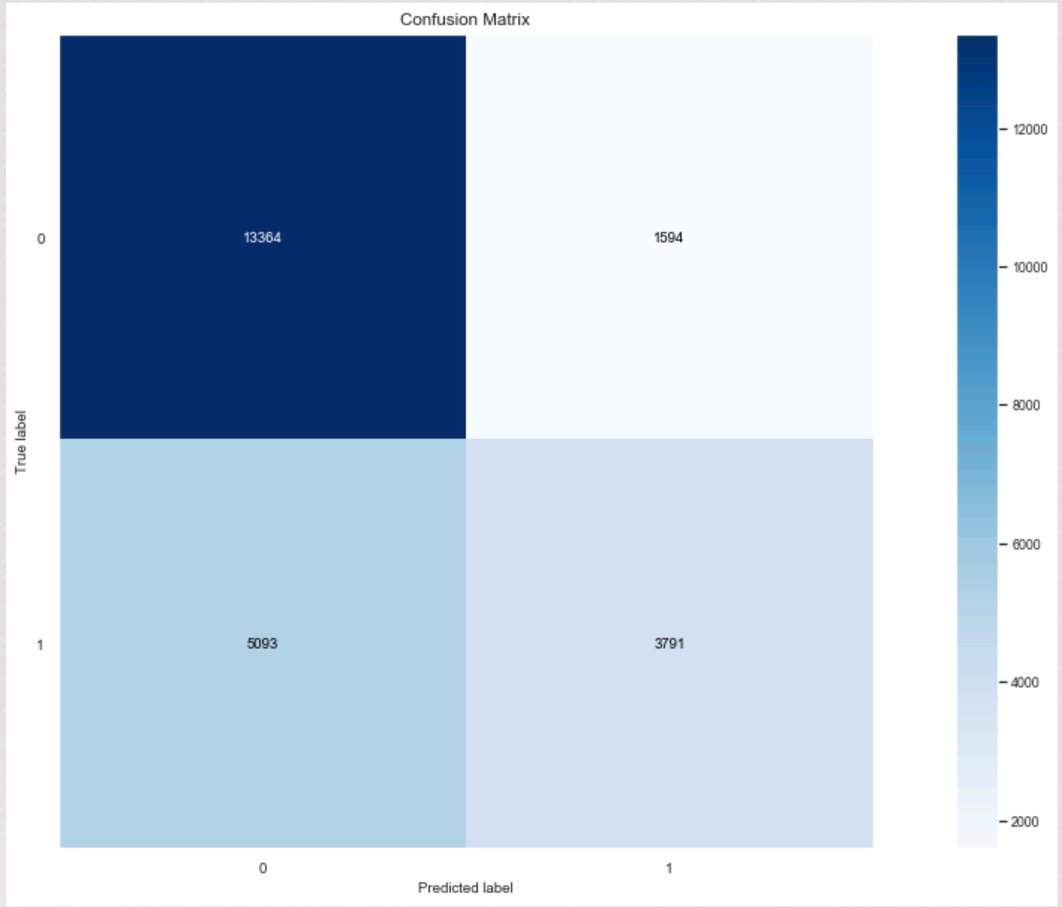
Random Forest ROC AUC



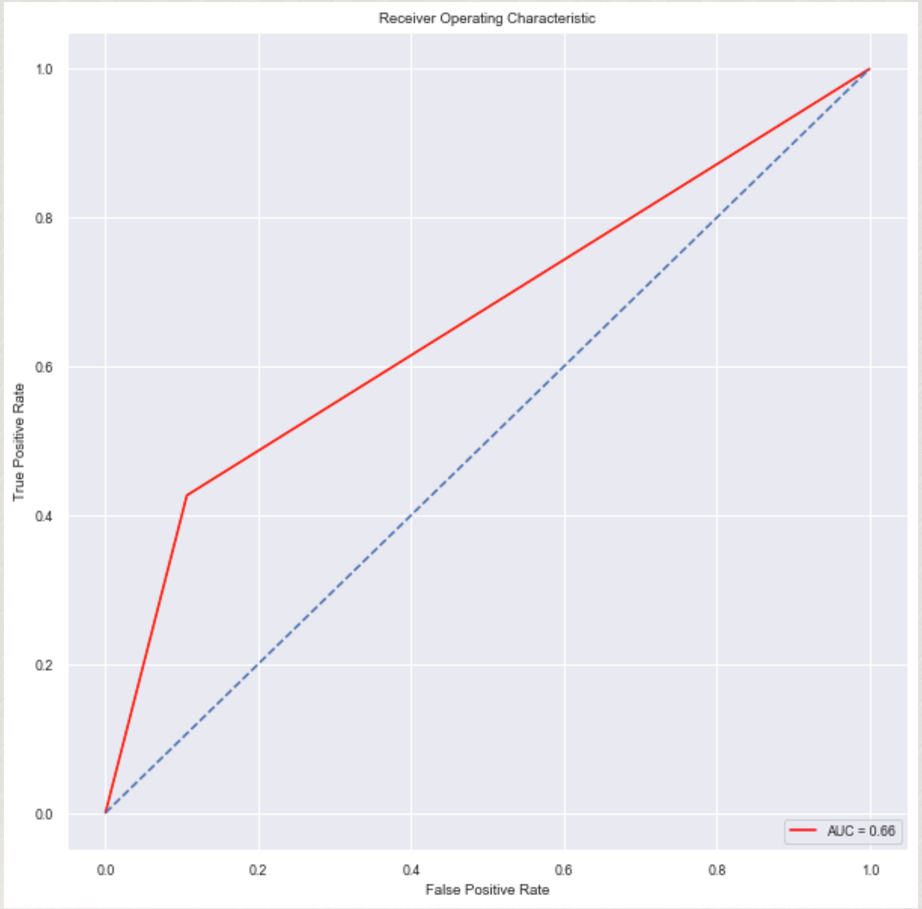
Logistic Regression Model Evaluation

```
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



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 then 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.*

Conclusion

By applying graph analysis, feature reduction, model evaluation & selection, and machine learning algorithms to each property, the dynamics between features per property can be highlighted. These dynamics will help the hotel owners to understand whenever the cancellations act in a similar way for different properties. The importance tells which features are used in the model and which are not. Each feature will get a score, and ranked scores between hotels per model are presented in tables in all 3 sections. This importance can be extracted for logistic regression and random forest. It is impossible to make a ranking between the features of the model. Overall the Random Forest model out preforms the logistic regression model on every point.

The approach that is suggested in this paper can be applied in different industries such as the airline industry and car rental industry, as well as the models that are applied. Inevitably, issues will be faced regarding the data selection and feature engineering because the data structure is in a different format, however, the same type of steps can be taken.

Hotel owners would benefit from the implementation of this algorithm into a dashboard, because these predictions should create more control in certain situations. The ability to see which reservations are likely to cancel in an overbooking situation for example. Besides their experience, they can act with more knowledge in crucial situations with the goal of generating extra revenue.

References

- https://builtin.com/data-science/random-forest-algorithm
- https://www.kaggle.com/jessemostipak/hotel-booking-demand
- https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8
- https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/