Hotel Reservation

By Hantao Lin

Problem

According http://blog.experience-hotel.com/, the average rate of **hotel cancellation from all sources is about 24%**. Consequently, it has always been the hotel revenue manager's biggest concern to counteract hotel cancelation. The mission here is to **build an appropriate model that predicts the likelihood of hotel cancelation** that helps revenue managers to better understand the situations and propose policies to accommodate or avoid hotel cancelation.

Context of the Data

There are two datasets used in this project:

1. Kaggle Dataset (https://www.kaggle.com/jessemostipak/hotel-booking-demand)
    - The dataset contains booking information for a city hotel and a resort hotel.
    - 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 32 features.
    - It has 32 columns and 119390 entries.

```
print(hotel.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 32 columns):
hotel                             119390 non-null object
is_canceled                       119390 non-null int64
lead_time                         119390 non-null int64
arrival_date_year                 119390 non-null int64
arrival_date_month                119390 non-null object
arrival_date_week_number          119390 non-null int64
arrival_date_day_of_month         119390 non-null int64
stays_in_weekend_nights           119390 non-null int64
stays_in_week_nights              119390 non-null int64
adults                            119390 non-null int64
children                          119386 non-null float64
babies                            119390 non-null int64
meal                              119390 non-null object
country                           118902 non-null object
market_segment                    119390 non-null object
distribution_channel              119390 non-null object
is_repeated_guest                 119390 non-null int64
previous_cancellations            119390 non-null int64
previous_bookings_not_canceled    119390 non-null int64
reserved_room_type                119390 non-null object
assigned_room_type                119390 non-null object
booking_changes                   119390 non-null int64
deposit_type                      119390 non-null object
agent                             103050 non-null float64
company                           6797 non-null float64
days_in_waiting_list              119390 non-null int64
customer_type                     119390 non-null object
adr                               119390 non-null float64
required_car_parking_spaces       119390 non-null int64
total_of_special_requests         119390 non-null int64
reservation_status                119390 non-null object
reservation_status_date           119390 non-null object
dtypes: float64(4), int64(16), object(12)
memory usage: 29.1+ MB
None
```

Figure 1. Kaggle Data

2. Wunderground data (wunderground.com)

- The dataset includes the average temperature, wind speed, precipitation and other things at Lisbon,
- It is collected from 7/1/2015 to 8/31/2017.
- It has 9 columns and 793 entries.

```
print(weather.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 793 entries, 0 to 792
Data columns (total 9 columns):
Date_time                793 non-null object
Maximum_Temperature      793 non-null float64
Minimum _Temperature     793 non-null float64
Average_Temperature      793 non-null float64
Heat_Index               184 non-null float64
Precipitation(inches)    793 non-null float64
Wind_Speed               793 non-null float64
Relative_Humidity        793 non-null float64
Conditions               793 non-null object
dtypes: float64(7), object(2)
memory usage: 55.9+ KB
None
```

Figure 2. Weather Data

Data Wrangling

The datasets were joined for the analysis. In order to do so, a new column(“Date_time”) is created in Kaggle dataset based on the “arrival_year”, “arrival_month”, and “arrival_day” in the dataset. After joining the two datasets with “Date_time” as the common column, missing values were filled differently in the new dataset. Missing value in “Heat_Index” is calculated based on a formula provided online. Other missing values in other columns are filled with the mode of the column. Column “company” is dropped from the dataset because it is missing 80% of its values. After the cleaning, the final dataset has 40 columns in total and 119390 entries.

```
merged.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 119390 entries, 0 to 119389
Data columns (total 41 columns):
hotel                             119390 non-null object
is_canceled                       119390 non-null int64
lead_time                         119390 non-null int64
arrival_date_year                 119390 non-null int64
arrival_date_month                119390 non-null object
arrival_date_week_number          119390 non-null int64
arrival_date_day_of_month         119390 non-null int64
stays_in_weekend_nights           119390 non-null int64
stays_in_week_nights              119390 non-null int64
adults                            119390 non-null int64
children                          119390 non-null float64
babies                            119390 non-null int64
meal                              119390 non-null object
country                           119390 non-null object
market_segment                    119390 non-null object
distribution_channel              119390 non-null object
is_repeated_guest                 119390 non-null int64
previous_cancellations            119390 non-null int64
previous_bookings_not_canceled    119390 non-null int64
reserved_room_type                119390 non-null object
assigned_room_type                119390 non-null object
booking_changes                   119390 non-null int64
deposit_type                      119390 non-null object
agent                             119390 non-null float64
days_in_waiting_list              119390 non-null int64
customer_type                     119390 non-null object
adr                               119390 non-null float64
required_car_parking_spaces       119390 non-null int64
total_of_special_requests         119390 non-null int64
reservation_status                119390 non-null object
reservation_status_date           119390 non-null object
Date_time                         119390 non-null object
Maximum_Temperature               119390 non-null float64
Minimum _Temperature              119390 non-null float64
Average_Temperature               119390 non-null float64
Heat_Index                        119390 non-null float64
Precipitation(inches)             119390 non-null float64
Wind_Speed                        119390 non-null float64
Relative_Humidity                 119390 non-null float64
Conditions                        119390 non-null object
Heat_index                        119390 non-null float64
dtypes: float64(11), int64(16), object(14)
memory usage: 43.3+ MB
```

Figure 3. Merged Data

Data Storytelling

After getting the cleaned data, I had some initial guesses about the factors that influence the chance of people canceling their hotel appointment：

1. Whether weather such as average temperature and wind speed have effects on the cancelation rate?
2. How is status of previous appointments correlated with cancelation rate?
3. If there are countries that have an extremely higher cancelation rate compared to the others?

Cancelation Rate vs Average Temperature

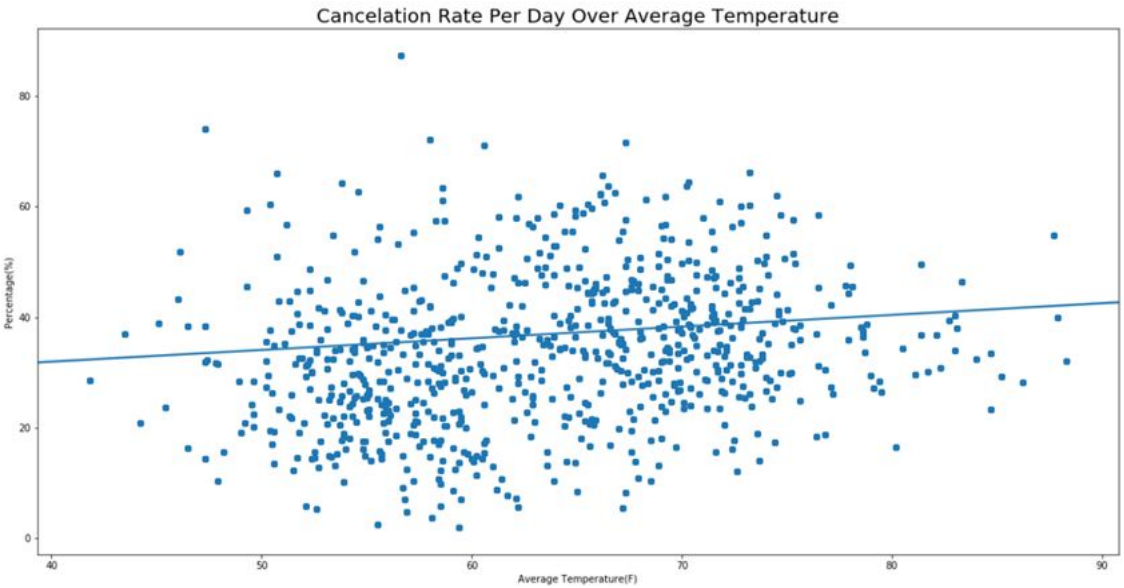

Figure 4. Cancelation Rate Per Day Over Average Temperature

From the graph, we can see there is a positive correlation between average temperature and cancellation rate over time

Cancelation Rate vs Wind Speed

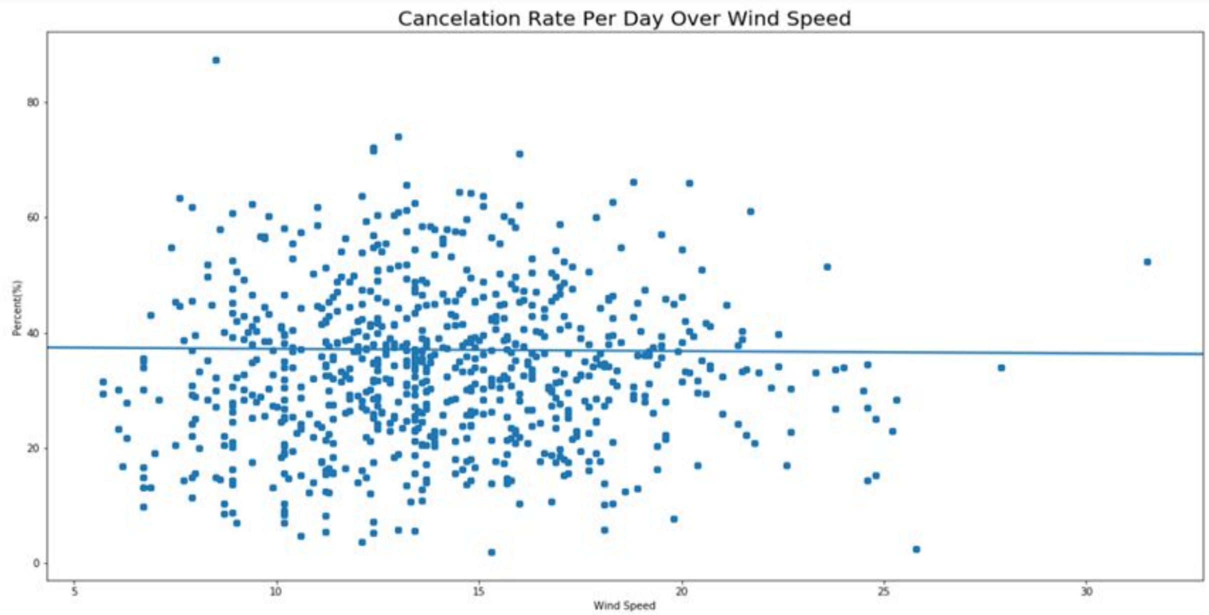

Figure 5. Cancelation Rate Per Day Over Wind Speed

From the graph above, wind speed and cancelation rate has a flat to negative regression line. This tells us that there is not much correlation between wind speed and rate of cancelation.

Cancelation Rate VS Previous Appointment Status

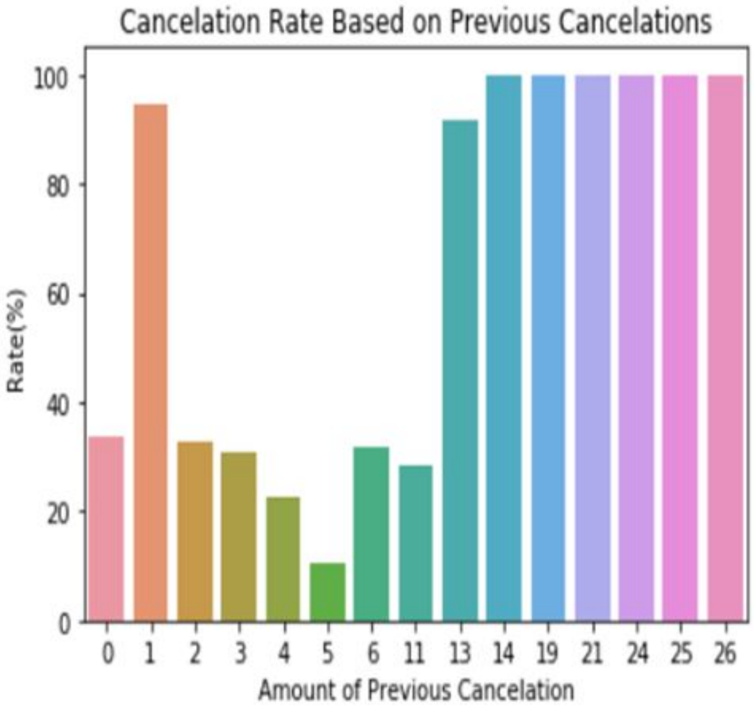

Figure 6. Cancelation Rate Over Previous Cancellations

From the graph we can conclude that people with 14 or more times of cancelation records have a 100% chance of cancelling again. Moreover, people who canceled 1 time and 13 times before are also extremely likely to cancel their reservation.

Cancelation Rate VS Countries

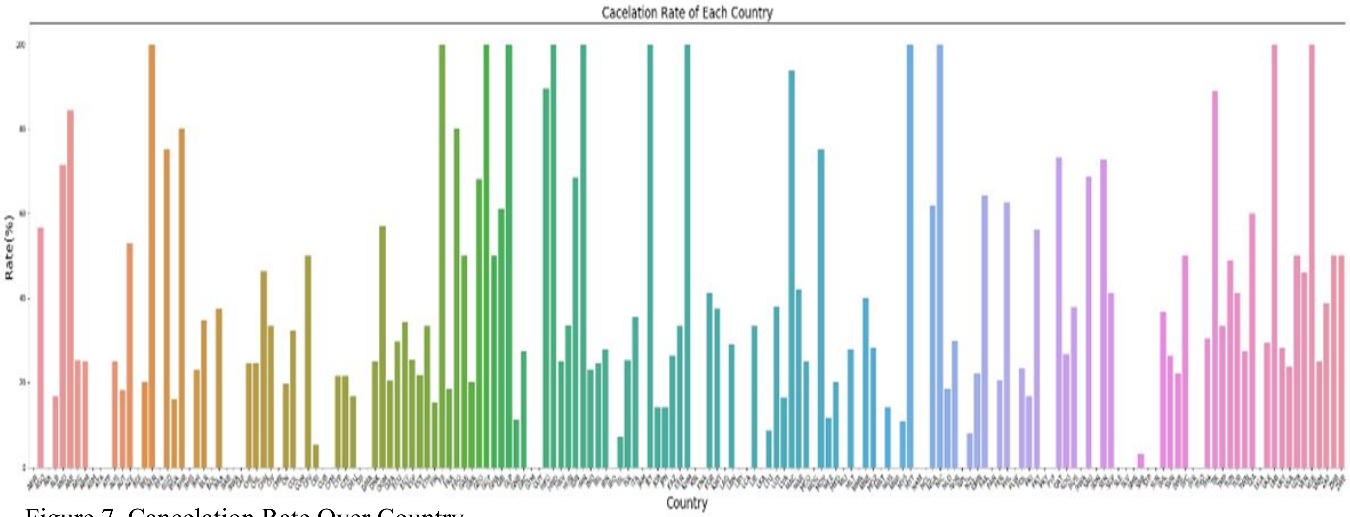

Figure 7. Cancelation Rate Over Country.

The are indeed some countries have higher cancelation rate compare to other countries.

Statistical Analysis

In this step, I will further investigate the correlation between wind speed, average temperature, and cancelation rate. Moreover, I will expand on the graph of cancelation rate of each country by examining if the domestic (Portugal) cancelation rate is equal to the international cancelation rate

Hypothesis Test for Average Temperature VS Cancelation Rate

Null Hypothesis: Pearson's coefficient = 0

Alternative Hypothesis: Pearson's coefficient != 0

```
cor_list=[]
for i in range(10000):
    sample1=np.random.choice(d['percent'],len(d['percent']))
    sample2=np.random.choice(d['temperature'],len(d['temperature']))
    cor=stats.pearsonr(sample1,sample2)
    cor_list.append(cor)
```

```
correlation=[]
for i in cor_list:
    correlation.append(i[0])
```

```
print('Correlation of Average Temperature and Cancelation Rate 95% Confidence Interval:'
      +str(np.percentile(correlation,[2.5,97.5])))
```

```
Correlation of Average Temperature and Cancelation Rate 95% Confidence Interval:[-0.07026314  0.07109968]
```

Figure 8. Confidence Interval of Correlation of Average Temperature and Cancellations Rate

Because the 95% confidence interval contains 0, it is inconclusive whether the two variables have positive, negative, or no correlation.

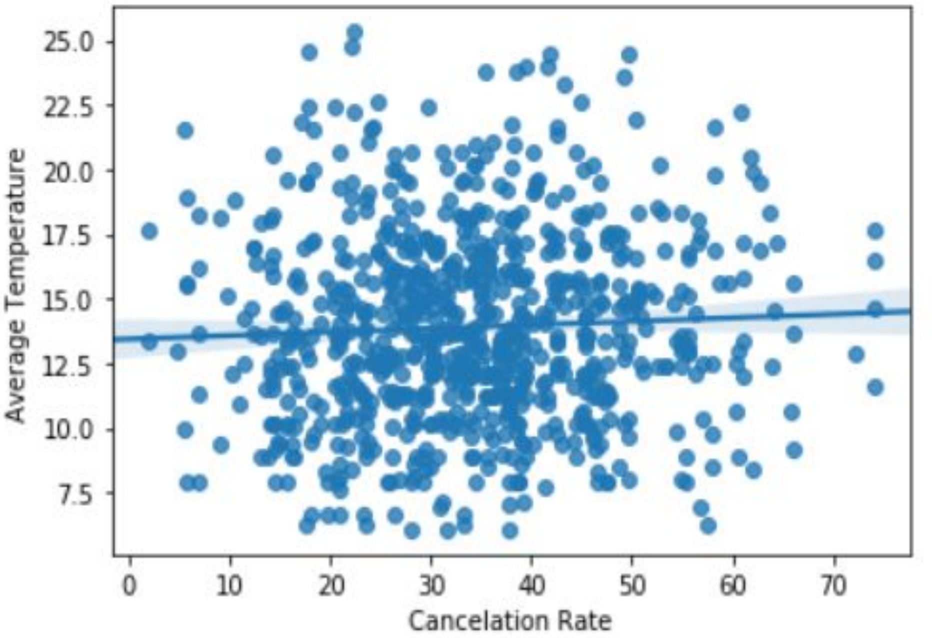

Figure 9. Bootstraps Sample of Cancelation Rate Over Average Temperature

The graph is the bootstrap points representing cancelation rate vs average temperature.It is very hard to conclude that there exists a significant correlation between the two variables since the regression line is not very steep.

Hypothesis Test for Wind Speed VS Cancelation Rate

Null Hypothesis: Pearson's coefficient = 0

Alternative Hypothesis: Pearson's coefficient != 0

```
cor_list2=[]
for i in range(10000):
    sample1=np.random.choice(d['percent'],len(d['percent']))
    sample2=np.random.choice(d['wind_speed'],len(d['wind_speed']))
    cor=stats.pearsonr(sample1,sample2)
    cor_list2.append(cor)
```

```
correlation2=[]
for i in cor_list2:
    correlation2.append(i[0])
```

```
p=[]
for i in cor_list2:
    p.append(i[1])
print('Correlation of Wind Speed and Cancelation Rate 95% Confidence Interval:'+
      str(np.percentile(correlation2,[2.5,97.5])))
print('p-value:'+str(np.mean(p)))
```

```
Correlation of Wind Speed and Cancelation Rate 95% Confidence Interval:[-0.06967139  0.06925268]
p-value:0.5026847741192151
```

Figure 10. Confidence Interval of Cancellation Rate and Wind Speed

The result made the test inconclusive because the confidence interval contains both negative and positive numbers. Similar to the average temperature vs cancelation rate, the correlation between wind speed and cancelation rate could be positive, negative, or not correlated at all.

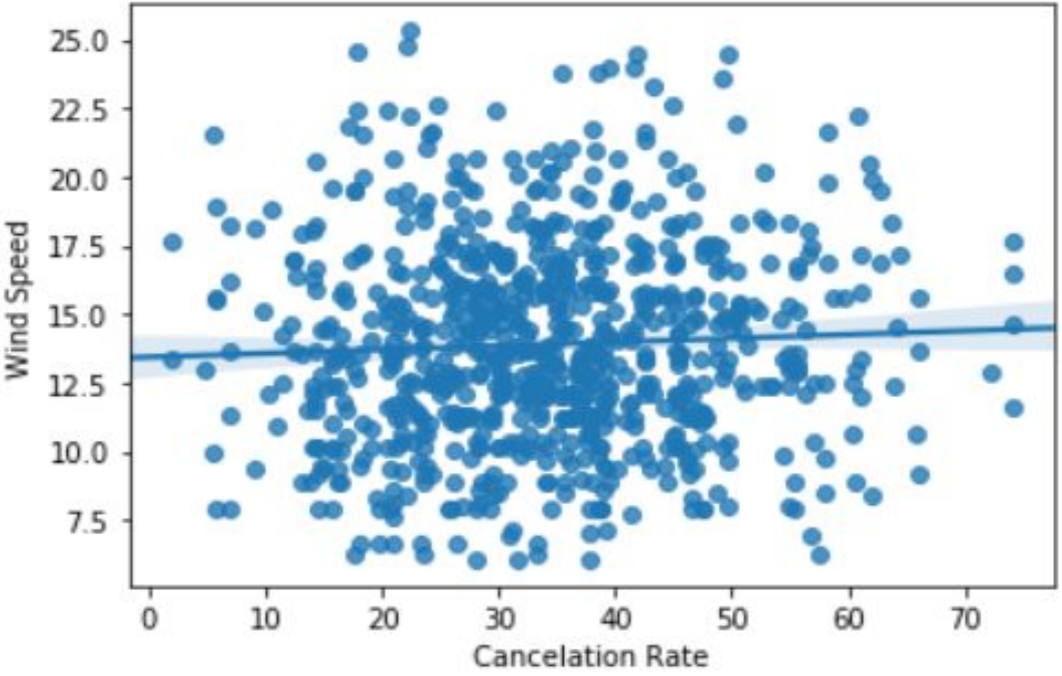

Figure 11. Bootstraps Sample of Cancelation Rate Over Wind Speed

The graph is very similar to temperature vs cancelation rate. This graph also illustrates that there is not a significant correlation between wind speed and cancellation rate.

Hypothesis Test for Domestic Cancelation Rate VS International Cancelation Rate

Null Hypothesis: mean domestic cancelation rate - mean international cancelation rate = 0

Alternative Hypothesis: Pearson's coefficient: mean domestic cancelation rate - mean international cancelation rate != 0

```
list=[]
for i in range(10000):
    sample1=np.random.choice(dom_count['rate'],len(dom_count['rate']))
    sample2=np.random.choice(int_count['rate'],len(int_count['rate']))
    mean1=np.nanmean(sample1)
    mean2=np.nanmean(sample2)
    diff=mean1-mean2
    list.append(diff)
percentile=np.percentile(list,[2.5,97.5])
print('Domestic vs International Cancelation Rate 95% Confidence Interval:'+str(percentile))
```

```
Domestic vs International Cancelation Rate 95% Confidence Interval:[-0.02214565  0.02167502]
```

Figure 12. Confidence Interval of Domestic vs International Cancelation Rate

The 95% confidence interval is (-0.02134154, 0.02174621) which includes 0. This made the hypothesis test inconclusive.

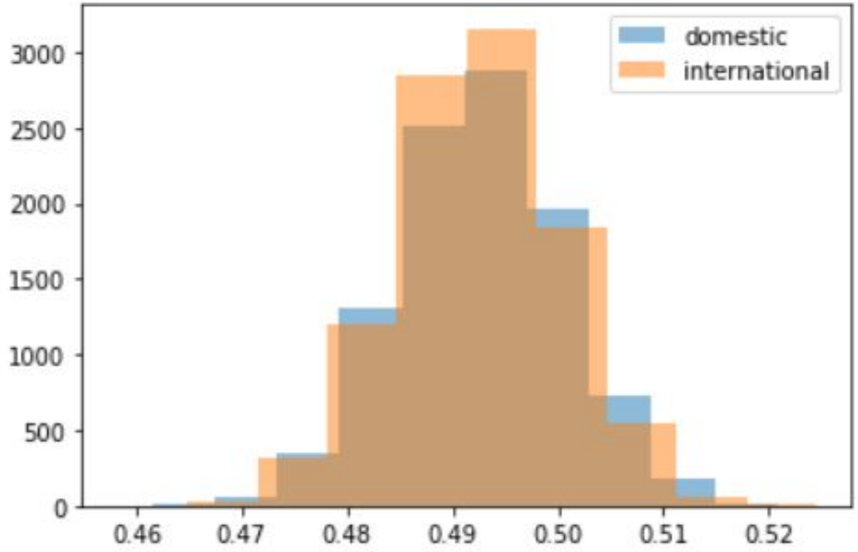

Figure 13. Distribution of Domestic and International Cancelation Rate

The distribution plot vividly demonstrates that it is not possible to conclude as the two distributions overlap each other heavily.

Machine Learning

In this part of the project, I compared 3 machine learning algorithms:

1. DecisionTreeClassifier

2. LogisticRegression

3. RandomForestClassifier.

Performance of 3 Algorithms

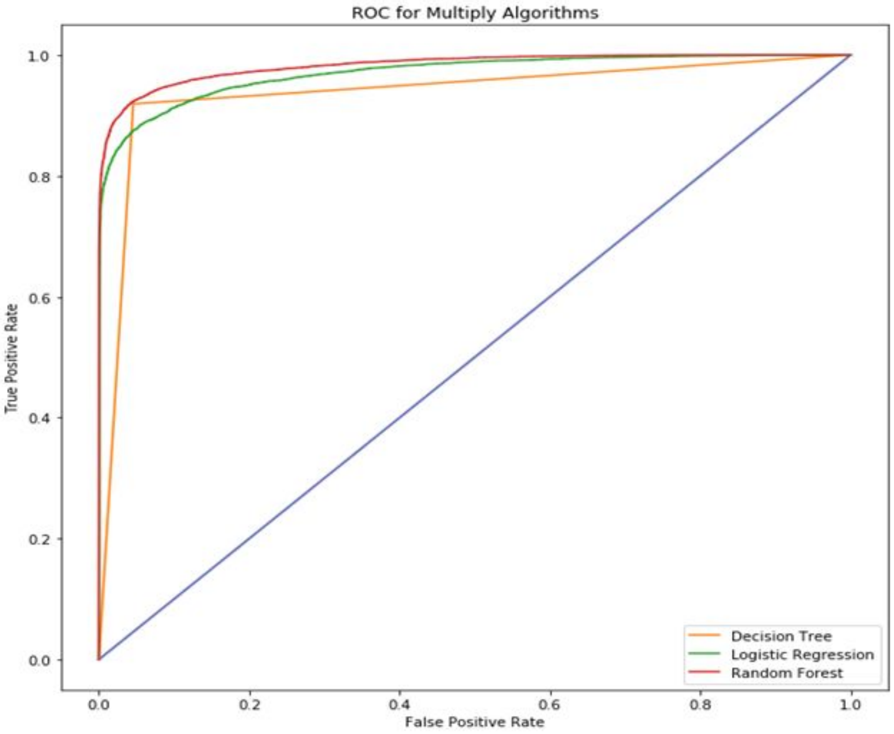

Figure 14. ROC for 3 Machine Learning Algorithms

Most Important Features

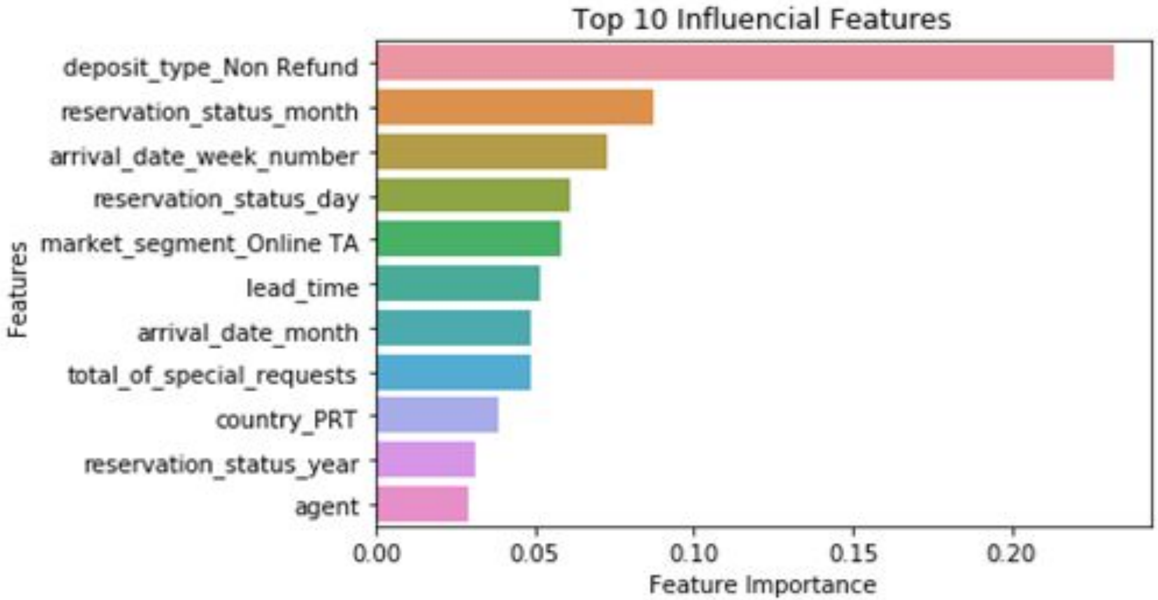

Figure 15. Top 10 Influential Features

Conclusion

In conclusion, to make the model better I think I should add holiday labels to each day because we see some influences from reservation status date. We can also maybe look at the agent-client pool and compare it to people with a different agent or no agent to better understand why that is influencing the cancelation of a reservation. For recommendation, because non-refund type room contribute to almost one-fourth of the feature importance, the hotel manager should increase their non-refund type rooms to reduce reservation cancelation.