Capstone Project 1: Final Report

Introduction

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. Building an appropriate model that predicts the likelihood of hotel cancelation helps revenue managers to better understand the situations and propose policies to accommodate or avoid hotel cancelation.

Data Wrangling

There are two datasets used in this project. The first dataset is directly downloaded from Kaggle (<https://www.kaggle.com/jessemostipak/hotel-booking-demand>) in a csv file. The dataset 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. It has 32 columns and 119390 entries.

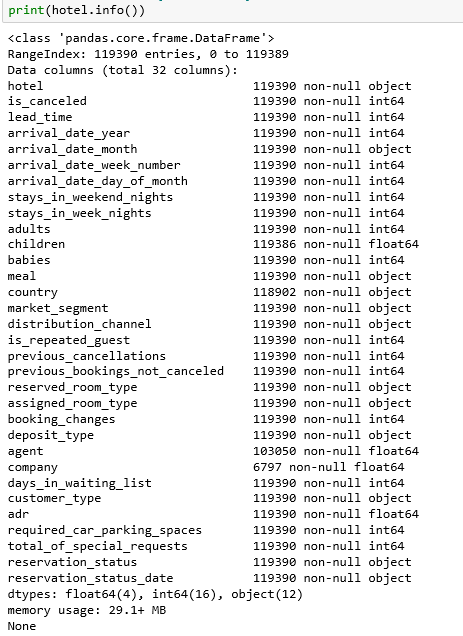
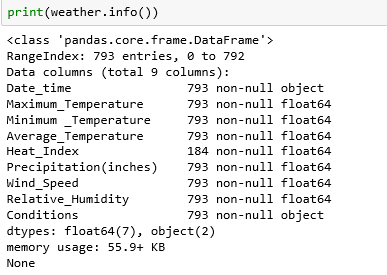


Figure 1. Kaggle Data

The other dataset collected from wunderground.com also comes in as a csv file. The dataset includes the average temperature, wind speed, precipitation and other things at Lisbon, Portugal from 7/1/2015 to 8/31/2017. It has 9 columns and 793 entries.

Figure 2. Weather Data

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.

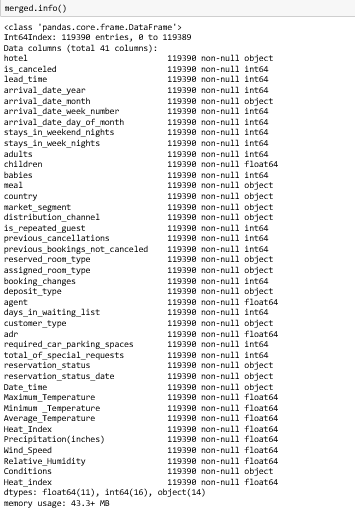


Figure 3.Merged Data*.*

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. I was interested in whether weather such as average temperature and wind speed have effects on the cancelation rate. The following graph shows the average temperature versus cancelation rate over time:

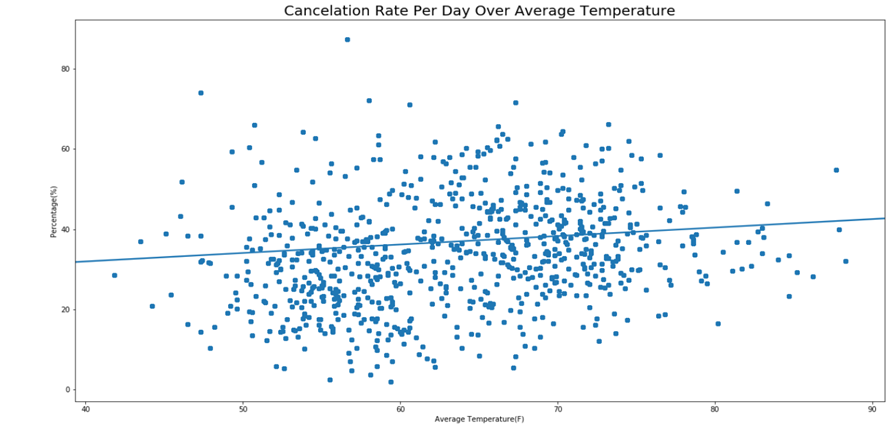


Figure 4. Cancelation Rate Per Day Over Average Temperature

From the graph, we can see there is a positive correlation between average temperature and cancellation rate over time. Now let’s take a look at if wind speed also have a positive influence on the cancelation rate:

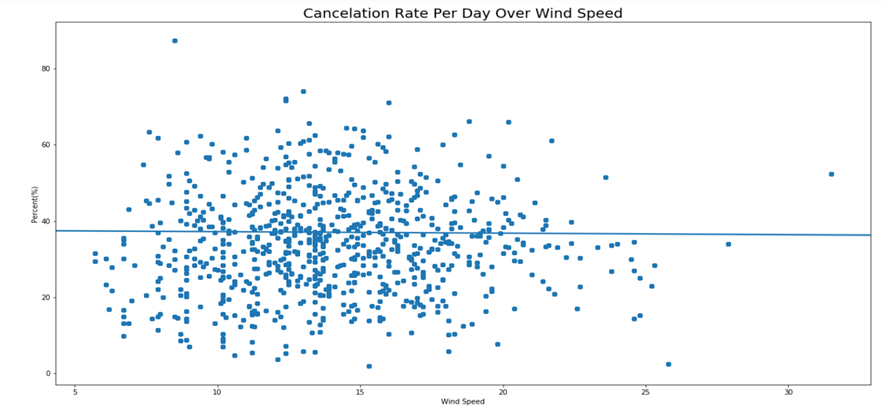


Figure 5. Cancelation Rate Per Day Over Wind Speed

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.

Another factor that might influence the cancelation rate can be the status of previous appointments. As a result, I plotted the graph below:

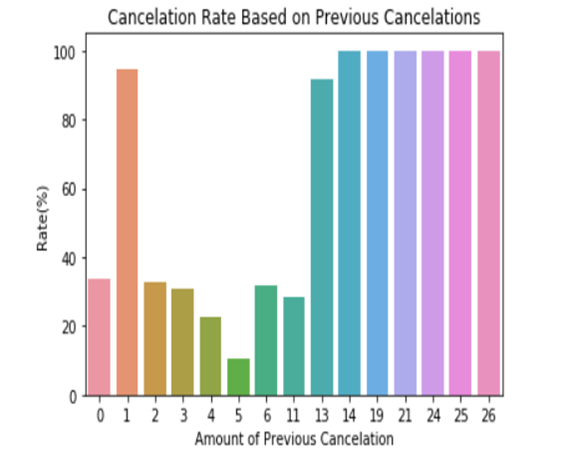
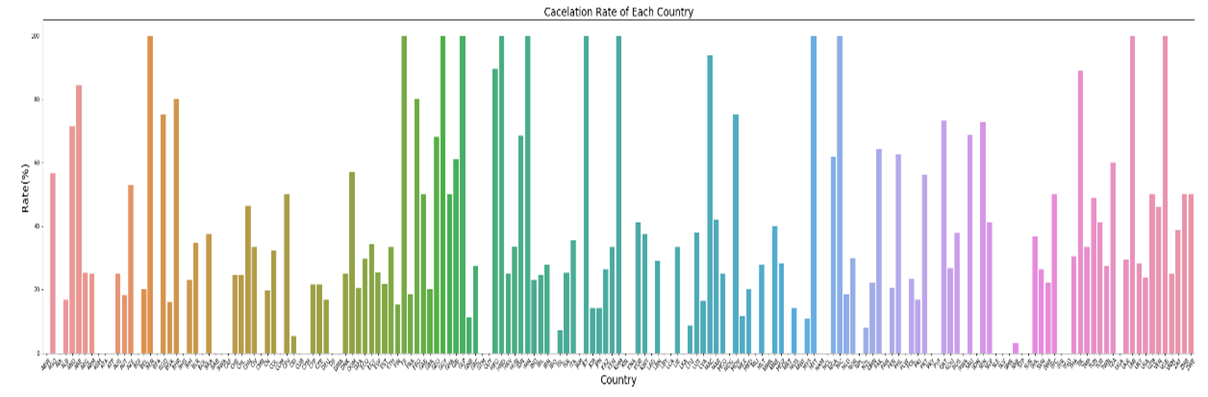


Figure 6. Cancelation Rate Over Previous Cancelations

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.

I also plotted the cancellation rate of each country just to see if there are countries that have an extremely higher cancelation rate compared to the others. The resulting graph is following:

Figure 7. Cancelation Rate Over Country*.*

Statistical Analysis

To further investigate the correlation between wind speed, average temperature, and cancelation rate, I set up hypothesis testings on average temperature vs cancelation and wind speed vs cancellation rate.

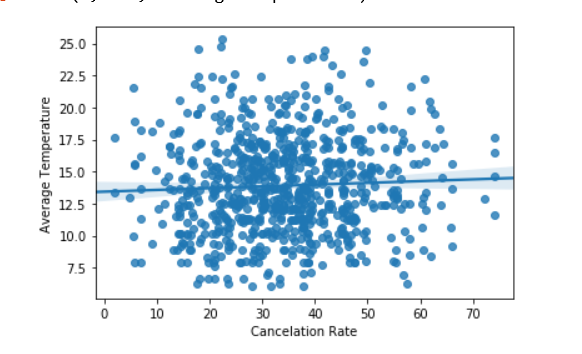
For average temperature vs cancelation rate, the null hypothesis is that the two variables are not correlated, else the Pearson’s coefficient will not equal to 0.  After obtaining the average temperature and cancelation rate, I used bootstrap inference to loop over 10000 cycles and calculated the Pearson’s coefficient and the 95% confidence interval that is equal to [-0.07093038,  0.07070765]. Because the 95% confidence interval contains 0, it is inconclusive whether the two variables have positive, negative, or no correlation. Moreover, as the graph shown below is the bootstrap points representing cancelation rate vs average temperature: 

Figure 8. Bootstraps Sample of Cancelation Rate Over 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.

For wind speed vs cancelation rate, the null hypothesis is that there is no correlation between wind speed and cancelation rate. The alternative hypothesis is that the Pearson’s coefficient is not equal to 0.After obtaining the wind speed and cancellation rate after performing bootstrap inference 10000 times, the calculated confidence interval is equal to [-0.06833989  0.06928068]. 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. In addition, the graph below is very similar to temperature vs cancelation rate:

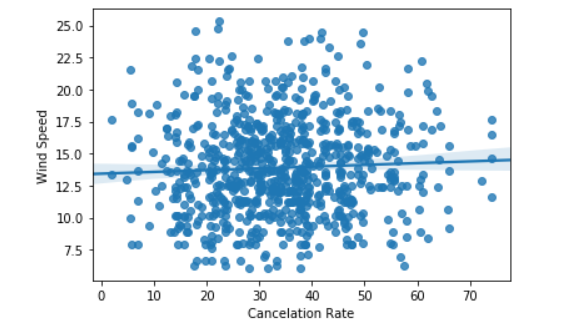
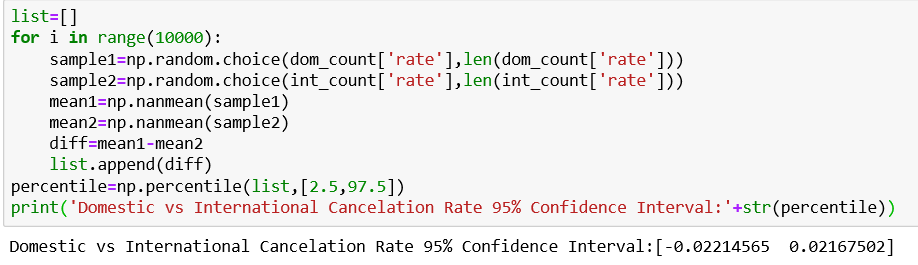


Figure 9. Bootstraps Sample of Cancelation Rate Over Wind Speed

This graph also illustrates that there is not a significant correlation between wind speed and cancellation rate.

Figure 10. Confidence Interval of Domestic vs International Cancelation Rate

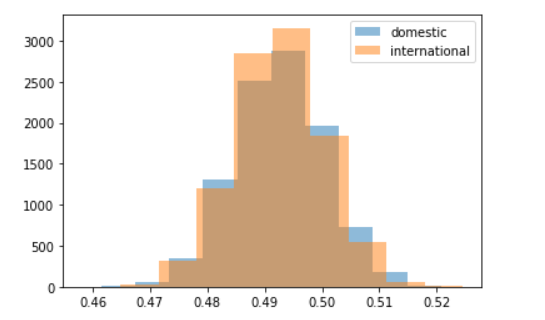
Lastly, I want to expand on the graph of cancelation rate of each country by examining if the domestic (Portugal) cancelation rate is equal to the international cancelation rate. After extracting the domestic and international cancelation rate, I used bootstrap inference to loop over the data for 10000 times and calculated the difference between domestic and international cancellation rates. The 95% confidence interval is (-0.02134154, 0.02174621) which includes 0. This made the hypothesis test inconclusive. The graph below also vividly demonstrates that it is not possible to conclude as the two distributions overlap each other heavily. 

Figure 11. Distribution of Domestic and International Cancelation Rate

Machine Learning

In this part of the project, I am comparing three types of machine learning models to find the one that has the best performance in predicting or classifying a client cancel their hotel reservation.

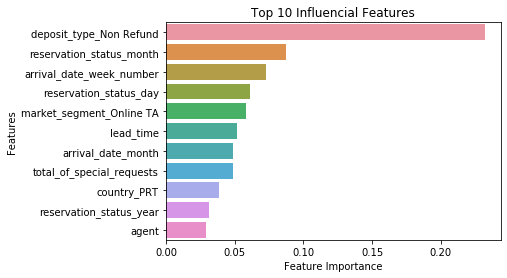
To feed the data into models, I need to preprocess the data. First, I use the function get\_dummy from pandas to create categories for string inputs. The next step I have done is feature selection to reduce model dimensions and have a better result. I separated the data into a training set and test set for evaluation later and avoid overfitting. Then I feed the training set into DecisionTreeClassifier and look for important features using the attribute ‘feature\_importance’. The result I got back is that only ‘reservation\_status’ have 100% contribution and the rest have 0% contribution to the feature importance. This represents that the feature ‘reservatio\_status’ is leaking the information about whether a client is canceling their reservation. This makes sense and I must remove this feature to build my model. After removing the feature and rerun the codes, I got a list of feature importance. I dropped all the features that have 0 contributions to the model and keep the rest. Here are the top ten most influential features:

Figure 12. Top 10 Influential Features

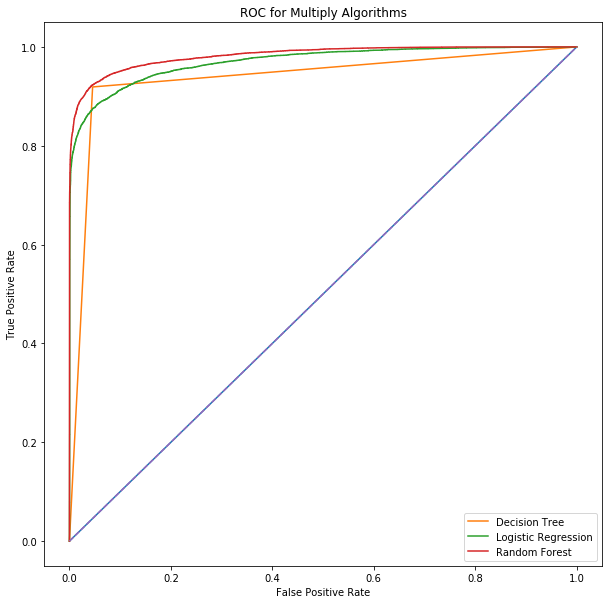
I feed the necessary features into DecisionTreeClassifier and calculate the AUC score; the result is about 0.93665. For logistic regression, the default setting gave an AUC score of 0.96977 and with tuning the C value using grid search, the AUC score improved a little at a value of 0.97776. Lastly, for RandomForestClassifier, the default setting returns an AUC score of 0.97534. After tuning the parameter using a randomized grid search, the AUC score is 0.983285. As a result, the RandomForestClassifier has the best performance in predicting the probability of a client canceling their reservation. Here is a ROC curve plot of the three algorithms:

Figure 13. ROC Plot for Multiple Machine Learning Algorithms

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.