Predicting Cancellations

Analysis of Hotel Reservations

Kendall Brown r0773111

KU LEUVEN

Semester 2, 2019-2020

Outline

- Problem Statement
- Task Objectives
- Initial Analysis
- Model Development
- Model Evaluation
- Results

Problem Statement

Can we use statistics to forecast potential cancellations of hotel reservations?

Task Objectives

- Establish a performance baseline from which to evaluate our model.
- Develop and optimize a predictive model which surpasses the baseline.
- Determine if the model is practically useful.

Initial Analysis

- Approximately 120,000 observations of 32 measurements.
- Significant number of unusual measurements.
  - Average Daily Rates of near or below 0.
  - Reservations without a guest count.
  - These proved to be inconsequential to the analysis once the country was taken into account.

Initial Analysis

- Cancellation rate does not appear to be influenced by time.
- Average Global Daily Cancellation Rate per day of .37

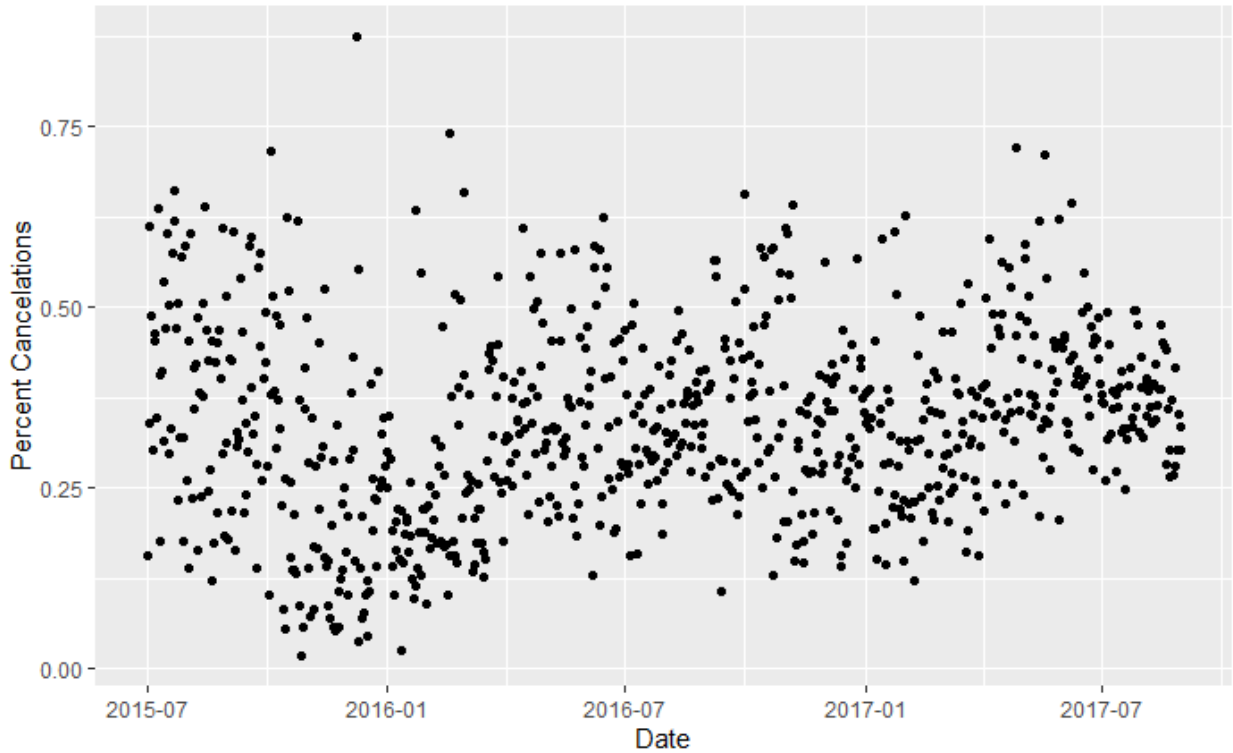
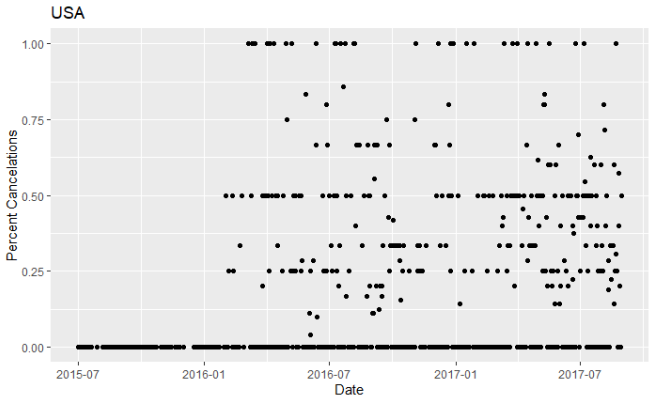


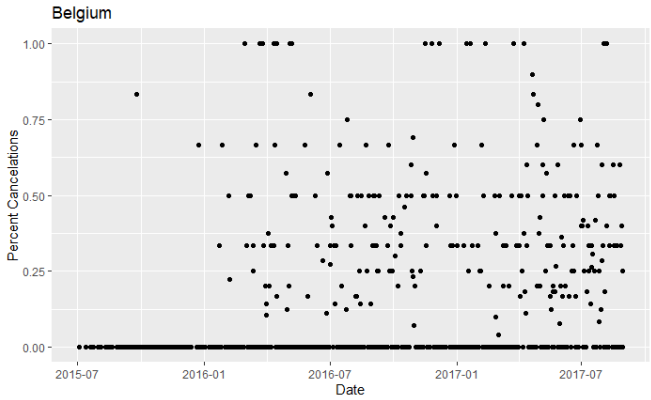
Figure 1: Global Cancellation Rate Per Day

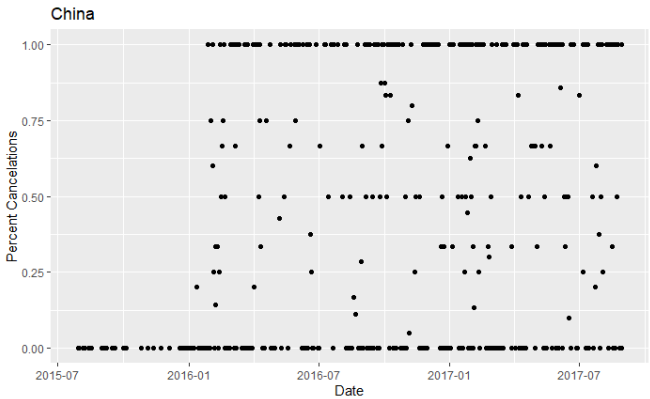
Initial Analysis

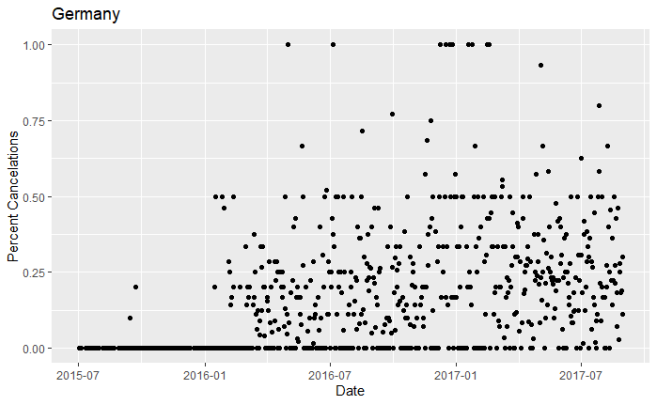
- Considering only certain countries reveals a possible error in the data for certain countries.
- For several months prior to January, 2016 a number of countries reported a zero or near zero cancellation rate despite exhibiting a more realistic metric immediately after for the next two years.
- Certain countries do not appear often enough to provide any meaningful insight.

Initial Analysis

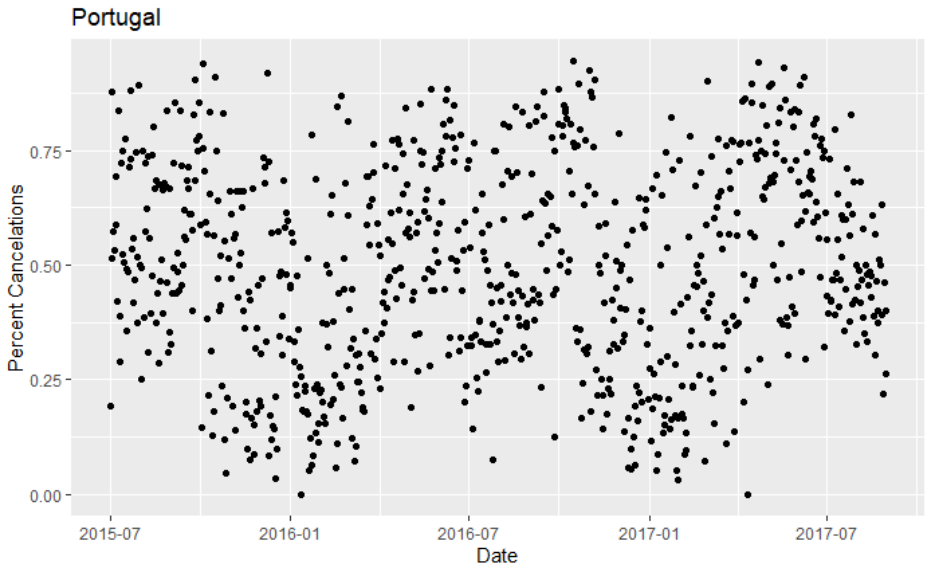


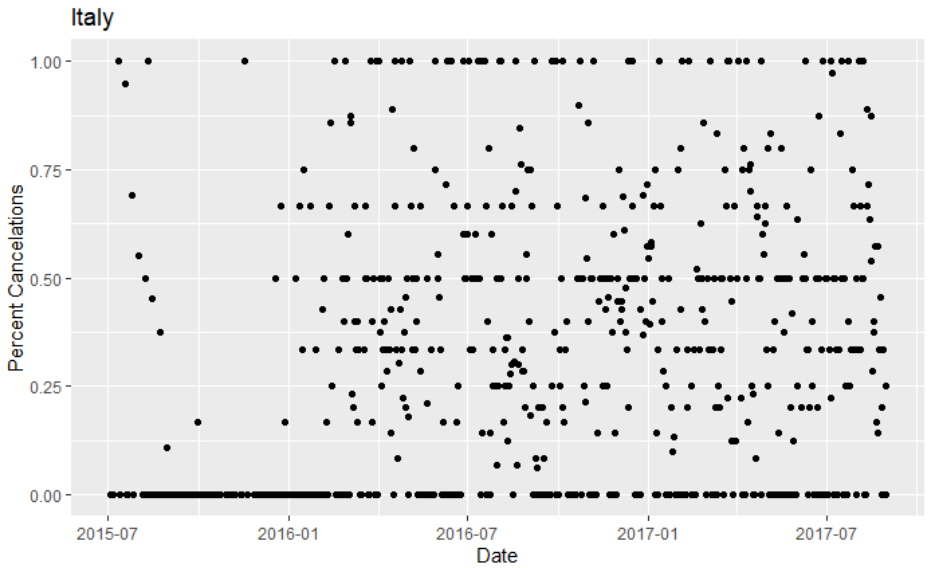






Initial Analysis





- This pattern may be the result of varying enforcement standards regarding data collection.
- As the problem appears to have been remedied post December, 2015 we shall only consider data points taken after the first of January 2016.
- Similarly several countries did not provide enough data points to be considered useful for the analysis. As such, countries which provided fewer than 20 data points were excluded from the analysis as well.
  - Ex. Aruba(2), Honduras(1), Senegal(11), Macao(16),...

Model Building

- When building a model we can start with a naive approach where we consider all gathered measurements.
  - Prone to errors, may lack performance if too small, and may lack interpretability if it grows too large.
- We can use statistical techniques to avoid the problems presented in naive models.
  - Doing so allows us to focus in on the most important measurements whilst reducing the number of possible errors within the analysis. Additionally, the model may become more interpretable and can provide stronger analytical power.

Model Building

- The goal of our model is to calculate a probability of cancellation given a certain set of measurements that we deem important.
- Through the use of statistical software, the following measurements were determined to be the most important for our model.
  - Average-Daily Rate
  - Lead-Time(Time between reservation date and arrival date).
  - Type of hotel (City or Resort).
  - Total number of guests.
  - If the guests are bringing children.
  - If they are a repeated guest.
  - The type of deposit they are asked of.
  - Whether or not they received the room they requested.
  - How they booked the hotel room (online, through an agent, etc.)
  - The time of year they are arriving.
  - The country for which the hotel resides in.

Model Evaluation

- True Positive/Negative: A prediction that matches the truth.
- False Positive/Negative: A prediction that does not match the truth.

Model Evaluation

- Raw Accuracy: *CorrectPredictions* / *TotalPredictions*
  - Not the best metric for evaluating predictive models as it does not account for false positives/negatives.
  - As a reminder 37% of reservations result in cancellation. Meaning if we were to blindly assign each reservation as a non-cancellation we would achieve an accuracy of 63%.
- We must re-balance our accuracy measurement to account for the fact that cancellations are rarer than non-cancellations. Doing so would result in a balanced accuracy rating of 50% when a blind prediction technique is employed regardless of the rarity of cancellations. This is done by penalizing false positives and negatives and then calculating the accuracy score.

Model Evaluation

- When considering raw accuracy, the model that was built correctly identified approximately 72% of reservations set aside for model testing.
  - Considering a blind prediction calculated with raw accuracy would be 63%, we can say that our model outperforms a blind prediction by about 15%.
  - As stated earlier this raw accuracy metric is misleading as the relative rarity of cancellations raises the accuracy metric to higher values than it deserves.
- When considered a balanced accuracy rating, the model achieves an approximate rating of 73% as well.
  - Considering a blind prediction should always yield a balanced accuracy of 50%, we discover that the model drafted here outperforms a blind prediction by about 46%.

Results

- We can predict a reservation cancellation using statistical methods quite well.
- If you wish to avoid cancellations, do not require a non-refundable deposit as this was shown to be a very significant in deterring cancellations.
- Guest with children are about as likely to cancel than those that do not, but this measurement changes with group size.
- Higher ADRs lead to more cancellation's.
- Repeated guests cancel less frequently.
- Cancellations seemingly occur less frequently as the year passes.

Questions

I will now be taking questions.