Cancellations of Hotel Reservation

Kendall Brown r0773111
Statistical Consulting
KU Leuven 2019-2020

May 12, 2020

Abstract

In this report an analysis of cancellations of hotel reservations will be explored using statistical methods. The primary objective is to build a predictive model which can accurately forecast potential cancellations. For this analysis a data set logging information regarding 119,390 hotel reservations made between 01-07-2015 and 31-08-2017 from 178 countries was used. Initial analysis revealed potential problems regarding this data set. For starters the hotel's average daily rate raised suspicion as there were a number of hotels with rates of zero, near-zero, or in one case below zero. Similarly, there were hotel reservations where the number of registered guests were zero. This cause for concern was remedied after measures were taken to account for these cases. It was determined that although these measurements appear to be mistakes, when corrected there was no significant increase or decrease in predictive performance. It was also discovered that reporting standards may have not been evenly enforced across all participating countries. This was discovered when analyzing potential time based trends as a number of countries were possibly not providing data regarding cancellations Unlike the previous situations, this problem did impact predictive performance quite substantially. As such dates prior to 01-01-2016 were redacted from the data set as well as instances where the hotel's country provided fewer than twenty instances of reservation data. The reason for redacting the lesser represented countries was to avoid making an incorrect analysis on a country that simply does not have enough information to substantiate a justifiable judgement regarding the behavior of that country.

Within the data set, there were thirty-two measurements of interest. For the sake of making the model both useful and interpretable, statistical methods were employed to determine which measurements were most important for the purposes of the model. After selecting these measurements, a model was drafted and optimized according to standard model building techniques. When performance was evaluated, this model out-performed a blind naive prediction by about forty-five percent, showing that it is indeed possible to build a predictive model for hotel reservation cancellations.

Initial Analysis

Initial analysis of the data gave useful insight into the data. To begin the base cancellation rate was calculated to be thirty-seven percent on a global scale. Seen here is a visualization of the daily cancellation rate over the analyzed period.

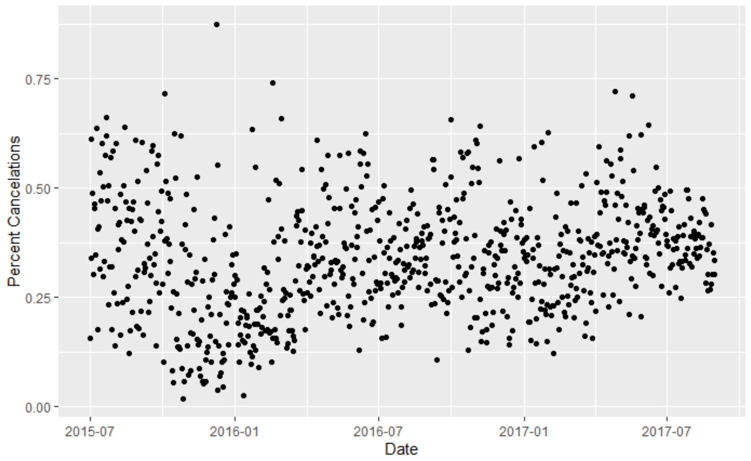


Figure 1: Global Daily Cancellation Rate over Time

As can be seen here, no obvious time based trends emerge when analyzed on the global scale. There is a period of high variance towards the beginning of 2016, however this pattern does not priciest through the same time period of 2017 and can be attributed to the following. When the country level visualizations are analyzed mixed results are apparent. To begin a number of countries seemingly did not report any cancellations for several months prior to 2016. Shown here are daily cancellation rates four countries, The USA, Belgium, China, and Germany. Cumulatively these nations, amongst others not listed, reported hundreds of reservations with a zero or near zero percent cancellation rate for the six months prior to 2016.

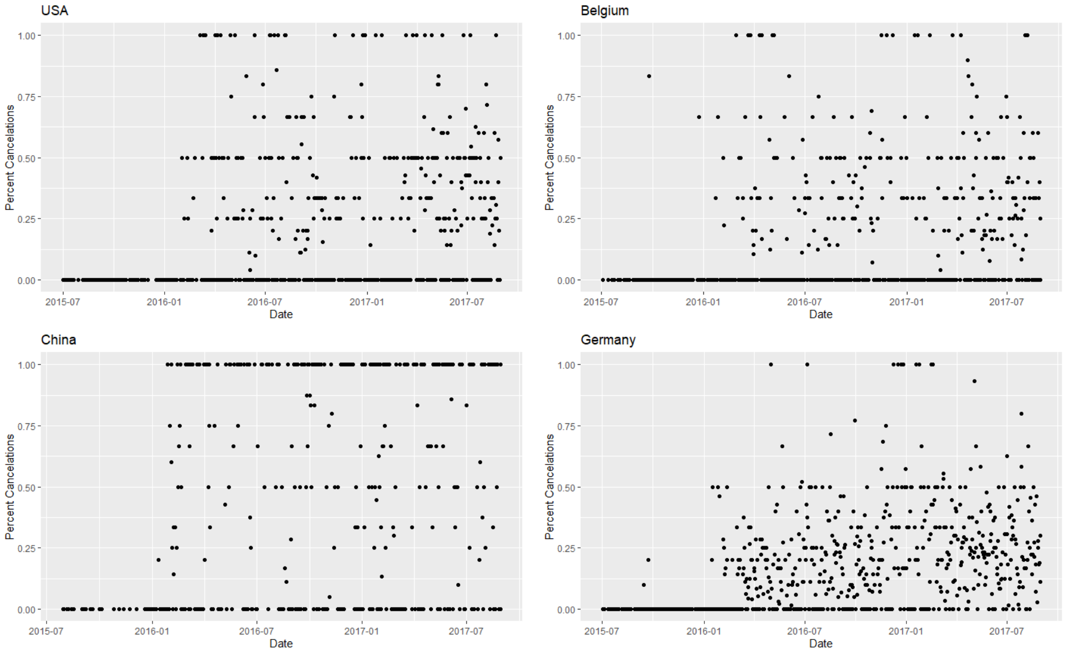
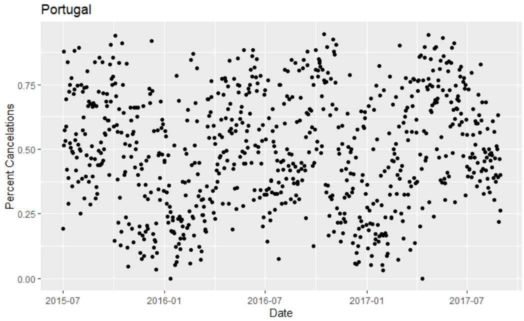
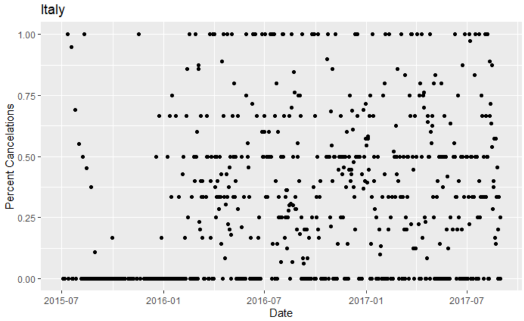


Figure 2: Cancellation Rates of USA, Belgium, China, and Germany

Other countries such as Portugal and to a lesser extent Italy did not exhibit this same pattern of cancellation prior to 2016. Initial reactions say that it may be incorrect to outright remove these data points from the analysis, however As the data from this period is in some cases quite obviously contradictory to the data provided after 01-01-2016, and because this time period is relatively small in comparison to the full size of the data, it was decided that all observations made before this date were to be considered invalid. This decision was made after a cross-comparison of models and will be further explained in the model building section of this report. Found here are plots detailing the daily cancellation rates of Portugal and Italy.





A number of potential errors in the data collection were discovered during this step and were handled accordingly. To begin, a number of hotels were offering daily rates of zero, near zero, or below zero. In total these observations represented about three percent of all observations. With these reservations making up a small minority of the total data set it is tempting to redact them from the data set and perform the analysis. This may be a mistake as the other measurements that describe these reservations may still be important. Similarly there were a number of reservations which did not log the expected number of guests. For the reasons stated earlier in regards to the measurement of average daily rate, it may be unwise to redact these measurements. This assumption was tested and is discussed in the model building section of this report.

Model Building

Before building a predictive model it is important to understand what the purpose of the model is to be. Here, a model is to be built with the goal of calculation of the probability of cancellation. There are three possible outcomes considered for a hotel reservation. Either the customer arrives and completes their reservation, they cancel beforehand, or they do not arrive at the hotel. For the purposes of this model, non-arrivals will be considered cancellations. Another model building procedure to determine is which measurements actually provide value to the model. To begin this process, all non-redundant measurements were considered as they were reported. Doing this revealed a number of problems, as discussed earlier there are countries with simply too few data points to provided meaningful information for the model and would often make the model nearly impossible to evaluate fairly with the data set given. As such the decision to redact all of the countries with least representation in the data set was made.

With the underrepresented countries no longer present in the data set, the model building procedure can continue. To ensure the model remains as interpretable as possible, only a few measurement were considered a time and the model was built gradually using statistically based optimization techniques. During this process the possible problems regarding average daily rate and total guest count were explored. A model built using the original measurements showed that these particular measurements were informative. To confirm this, these data measurements were divided into groups. For the average daily rate, these measurements were categorized into budget categories. Economy representing the cheapest rooms, standard, representing a typical room, expensive representing rooms as such, and luxury representing the most expensive rooms provided. For total guests, these groups were divided into unknown, single, double, and triple+. In regards to average daily rate, after testing multiple price cut-off points for each group and re-evaluating the model at each step it was determined that the original data was functionally equivalent for the purposes of the model. As such the original measurements were used. In dissimilar fashion, the re categorization of group sizes yielded fruitful results. When attempts were made to account for the unknown group size model performance slightly improved when the unknown group size was merged completely with the single guests. As such, the group size was restructured and the model was built using this updated measurement.

With the suspected problems addressed the optimization of the model can now take focus. Using statistical techniques each measurement's influence on the model was evaluated and quantified. After discerning the the most important measurements, the final model was built. These measurements include, the average daily rate, the lead time between reservation and arrival, the type of hotel, the total number of guests, whether or not children or babies are present, if the guest is a repeated guest, the type of deposit, the type of meal provided by the hotel, if the guest received the same type of room they tired to reserve, the time of year for their arrival, and the country of the hotel.

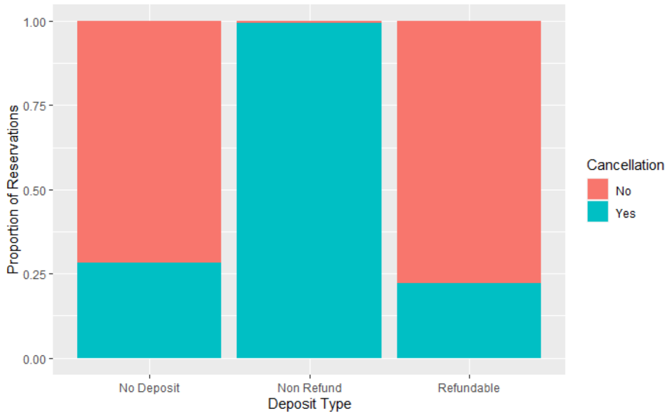
Model Evaluation

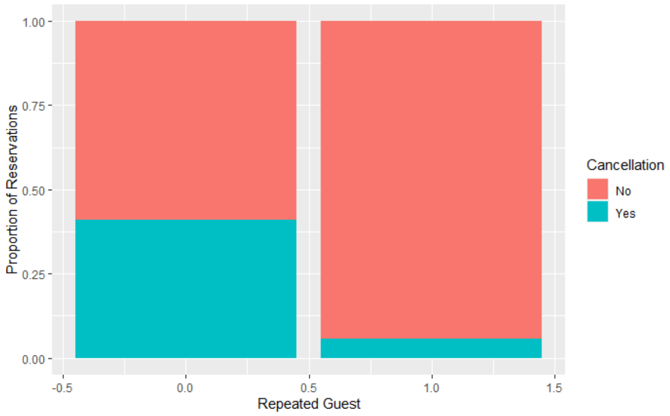
To begin the model evaluation, it must first be discussed how predictive models are evaluated. What is possibly the most popular form of model evaluation is the metric of accuracy. We define accuracy to be the number of successes divided by the number of attempts. This would be fine for the purposes of evaluating the model drafted here had it not been for the fact that cancellations are quite a bit rarer than non-cancellations. In the initial analysis it was shown that only about thirty-seven percent of reservations result in cancellation. This implies that if a person were to assume that there would be no cancellations, they would achieve an accuracy rating of sixty-three percent. This is accuracy metric is misleading as it does not account for what are referred to as false positives and false negatives. False positives and negatives are, simply put, incorrect predictions. When evaluating models it is important to consider these metrics as indications of strength. Strong models minimize the rate at which these errors occur, and will ideally grant a higher level of raw accuracy over a blind prediction. To minimize the rate of false positives and negatives, a balanced accuracy rating will be considered. This balanced accuracy rating will penalize high levels of false positives and negatives such that in the event of a blind prediction, the balanced accuracy rating will remain close to fifty-percent.

In consideration of the balanced accuracy rating the model was evaluated to have a raw accuracy rating of approximately seventy-two percent with a balanced accuracy of seventy-three percent as well. In consideration of raw accuracy it can be said that the model drafted here out-performs a blind prediction by about fourteen to sixteen percent, whereas the balanced accuracy rating shows that the model outperforms a blind prediction by over forty-five percent.

Results

This exercise in predictive modeling demonstrates that it indeed possible to use statistical techniques to build a predictive model for hotel reservation cancellations. When interpreting the model it can be said that one of the most important metric to consider when trying to reduce cancellations is to not require a non-refundable deposit. As can be seen in the plot below the rate of cancellations amongst hotels which require this type of deposit is quite high in comparison to hotels that do not. Similarly cancellations occur less frequently as the year passes, the hotel country does play a significant role in determining the probability of cancellation, as well as the hotels average daily rate. Single+Unknown groups cancel less frequently than doubles or triple+. Repeated guests cancel much less frequently then those that are not. Two plots visualizing some of the more extreme cases of cancellation rate differences are shown here. To conclude, it can be stated that it is indeed possible to build a predictive model utilizing adequate data and statistical techniques. When attempted here, the model drafted out-performed blind prediction by quite a substantial margin and can likely be improved upon with more data. This is especially true for groups which may have been underrepresented in the data set used here.


Proportion of Reservations
No Deposit
Non Refund
Refundable
Deposit Type
Cancellation
No
Yes


Proportion of Reservations
Repeated Guest
Cancellation
No
Yes