Won't You Be My Neighbor?

And For How Long?

By: Ben Chapdelaine, Adam Chappell, Michael Friedman, Tausha Morris

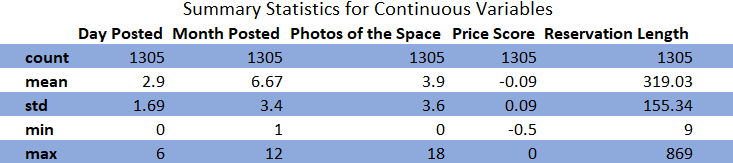
**Introduction:**

The explosive growth of online service companies ranging from real estate sharing (AirBnB) to pet sitting (Rover) has spurred the necessity for predicting how customers will utilize these services. Neighbor markets itself as the “AirBnB of storage” and has garnered millions in seed investment[[1]](#footnote-0). The company connects individuals who have extra space on their home/property (such as in the basement, attic, or garage) with people who are looking for low cost, reliable storage in the area. Neighbor’s website and app display listings posted by “hosts” who are trying to rent out their storage space. These listings include information including area, size, ease of access, surveillance, discounted price, etc. helping customers to make decisions about the space they would like to reserve. Both Neighbor and the storage hosts only make money when a space is under reservation, so it is in both parties’ best interest to understand how long that space is likely to stay reserved. A host may want to know a prediction of how long a reservation will last for a specific listing based off the traits of the listing. Neighbor may wish to offer predicted projections of reservation length based on the information provided by the host. Our model identifies the most important traits in determining reservation length and uses these traits to make predictions about how long a specific space will be reserved given these features. In addition to the financial benefit to both Neighbor and its clients, this prediction data could be used in future economic models as an instrumental variable used to better understand consumer decisions in the ever growing peer-to-peer services market.

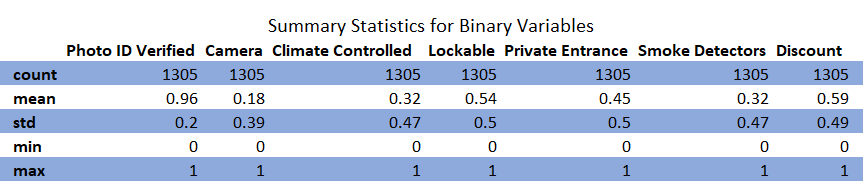
**Data:**

All of the data for our project came from Neighbor's private database of company data. One of our group members was able to get access to this data because he had previously worked as an intern for the company. This data was well structured when we received it but it was spread across many different tables with different types of information. One table that we thought would be especially useful was the table that had listing information on it. It contains several variables that we thought would be useful, including: the type of storage that was being offered, what time of year the posting was listed, and price per square foot. Another useful table has information on the hosts that were offering the space. Relevant variables include: the host’s level (an indicator of how trusted the host was and how many verification steps they had taken), whether they had a bio on their profile, and whether they had a profile photo or not. Figures 1 and 2 give summary statistics for the binary and non-binary variables.

**Summary Statistics for Continuous Variables (Figure 1):**



**Summary Statistics for Binary Variables (Figure 2):**



These tables also show data that we do not have including what the renter was storing, how many people had viewed the listing, and demographic data on the renter. These variables either weren't collected or would require natural language processing to be useful which we were unable to implement.

Our response variable was more difficult to manage. We had observations where the reservation was still active. We discarded these which reduced the number of observations from roughly 3,000 to just 1,300. We believe this caused major issues for our model and is further explained later on. For future implementation, a possible model could be the Cox Proportional Hazard for this type of data which would accurately include the data points that we were forced to throw out.

**Methods:**

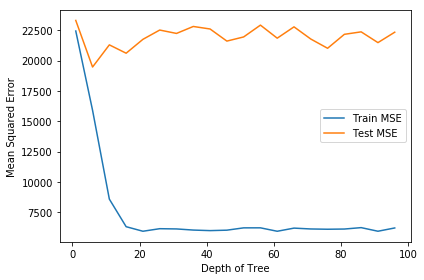
We decided to use Random Forest because it does not make any assumptions about the functional form. We’re not sure about how our predictors are related to the outcome, so random forest accommodates here nicely. Since random forest is a bit of a “black box” approach, it doesn’t take away from our goal of prediction. As long as it’s yielding accurate predictions, we’re not concerned with the interpretability of the model or with insights on causality.

To identify the best parameter for our Random Forest Model we started off with a Gridsearch with Cross Fold Validation. We started off with values for max\_depth ranging from 1 to 9 and it returned the best parameter as 9. We then reset our grid to range from 9 to 19. Again it returns the largest number in our search and we had to increase the range. After several instances of this we landed on the best max\_depth parameter being 50. The exact same process was used to identify the best number of estimators which was 400.

As an alternative method we tried lasso. We selected lasso for a lot of the same reasons as random forest. We have a relatively large amount of variables, so lasso helps guard against overfitting by setting many variables to zero. Thus we were able to take all of our variables (of which we weren’t sure which ones would turn out to be most predictive) and make a prediction out of it. One common pitfall in using Lasso misinterpreting the variables that it chooses (doesn’t set to zero) as the true causal or “most important” predictors. If we remember that the model is for prediction purposes only and not to explore possible causality, we don’t need to worry about Lasso hindering our results. Therefore, if Lasso throws out some potentially “truly causal” variables, it is of no concern.

**Results:**

In our results our Random Forest MSE training and MSE test were extremely high. This was due to overfitting and the same was for our Lasso results. Our MSE training was 2763.09 and MSE for testing was 12105.76. This difference is indicative of high overfitting. One cause of overfitting may be due to the amount of observation that the reservation had ended. In the graph below we ran a simulation to see what has happened to our training and testing MSE as the tree depth increases. The result from the graph shows that our training mse drops quickly and remains around the 6,000 MSE. The test MSE begins to drop around 10 trees, but begins to increase and is starting to decrease.



Our test set continues to have a high amount of overfitting, but as more trees are added the test set looks like it will decrease. The top five variables in our random forest method were the month posted, superhost level, price of rental, security camera and manual score. Our lasso method found security camera, manual, storage type, discount and enterprise to be the most helpful. There is a slight difference between the two top variable lists, but security camera and manual score appear in both.

**Conclusion:**

We set out to predict the timing of demand of the peer-to-peer storage market at Neighbor so that hosts can plan ahead and know how long their space is likely to be reserved. This is also helpful for Neighbor to know in order to better predict demand in the future.

Unfortunately our model does not have the predictive power we hope for, and we’re not really sure why. It could be due to a number of reasons. It might be due to data issues. One issue discussed to improve our results was to change our reservation time to months instead of days. This would help improve our result and probably decrease the depth of tree in our random forest. If there was a way to include current ongoing reservations, this could increase our dataset and possibly yield more predictive results. The explanatory variables we chose just might not be very indicative of reservation length. In the future it might be valuable to gather relevant customer data such as demographics, what items being stored, etc. These we guess might hold more predictive relationships with storage reservation length. A natural language processor could be used to parse customer’s input descriptions to make sense of it. This procedure can be rerun periodically to adjust for new customers/areas as well as trends that arise over time. If a more predictive model is attained, the outcome could be instrumented in future economic models involving the peer-to-peer services market.

1. <https://techcrunch.com/2018/03/28/neighbor-a-p2p-self-storage-marketplace-bags-2-5m-seed/> [↑](#footnote-ref-0)