Capstone 3 Air bnb Report

By Thomas McMahon

When I was growing up if you wanted to go on away on a vacation for a week or even just a couple days your options for where to stay was limited to staying with friends, staying at a hotel, staying at a motel, buying a timeshare. Now with the boom of the internet, smartphones, and other technology new bossiness sectors have been created. One if the sectors that has been created is short term publicly owned rentals. Now with apps such as Air bnb people are able to rent out any property, they own to anybody that they want. With these apps customers no longer have to rely on finding a hotel room and often can do it at a much cheaper rate, because you are renting from an individual and not a corporation. This allows people to travel the world without having to worry about every hotel room being booked for an event because there will always be people willing to rent you a room if the price is right. In this project I will take a large city New York City and build a predictive model to predict how much the average price of an air bnb will be given how many people it can accommodate and where it is located. New York City is a perfect place to look because it has the five large boroughs as well as a very large dataset of rental histories. This project will be able to help both customers of air bnb and people looking to list their properties out to find a price range that will be able to complete in the open market while not being losing out on value by being under priced for the owners.

The datasets had all of the information that was needed to build the models, but the data did have quite a few missing datapoints. The two questions that came up were how to fill the null values and if the features with null values were even relevant to the problem at hand. When looking through the features it seemed that the most important features would be the location of the listing and the amount of people that the listing could accommodate. After reaching that conclusion all other features in the dataset were removed.

One feature that tuned into a problem was the location of the specific neighborhood the listing was in. The reason for this was when getting too specific with a feature that doesn’t have enough datapoint will cause a bias in the model. Each neighborhood have very few individual listings make it a bad feature to use in a predictive model. For this reason the name of the specific neighborhood was removed and for location just the name of the borough was used in the model.

Now with only the boroughs, number of people that the listing accommodates, and the price as the data points it was time to model the data. First a linear model was used as a baseline to see how well the model could predict the listing price using just two features. The mean absolute error of the linear regression model was 77.26. This was a good start, but needed to see how that would compare to a random forest model. The random forest model produced a mean absolute error of 74.95. Overall the mean absolute error was pretty good, but the question becomes what happens when a subset of the data without many datapoints is removed from the model.

In the dataset of almost 40 thousand listings only 970 listings could accommodate 8 or more people. To see how the model would change the data was split into two datasets, one of all the listing that could accommodate 8 or more people, and the other was the rest of the dataset. The results of the model created on the dataset that accommodates larger groups unsurprisingly preformed much worse than the dataset as a whole. The mean absolute error of the large dataset was 284.75. What isn’t known is the reason for the massive shift the lack of datapoints or was their more variance or something else going on that hurt the model. Also to be expected the smaller dataset had a significant drop in mean absolute error of about 20% to 62.72. Again, it isn’t known how much of that drop was from pulling datapoints out of the model vs actual improvements to the model. After running these models it was still needed to find the best hyperparameters to use them to build the best random forest model on the original dataset. When using the best parameters the random forest model did not see any improvement from the original random forest model.

In conclusion the model did work, but there were a goof amount of outliners that would really need to be looked at to see if there is any way to account for them. Having more relevant features could have helped the model, such as what the listing has in terms of a pool, grill, and other amenities. Even with other features there will still be outliners that need to be taken into account. From the business side the model still will be helpful be give a person a ballpark estimate that should make sure the listing price is in the right range of prices.