Isabelle Fitzpatrick, Erina Fukuda, Subashini Sridhar
SDS 293 - Machine Learning Final Project

Predicting Price for Airbnb Listings in New York City

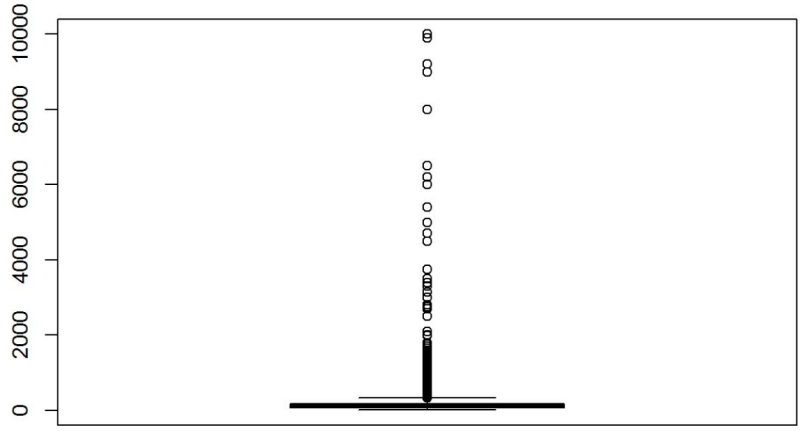
Introduction

Since 2008, Airbnb started a radical change in how people travel, creating an alternative to staying in hotels and hostels. Not only are there now more options than ever for the choosy vacationer, but thousands of people around the world have been empowered to make an income renting out their homes or spare rooms. Today, there are Airbnb listings in 34,000 cities in 190 countries, which is every country except Cuba, North Korea, Iran and Syria; and there are hundreds of thousands of people in the world renting out Airbnb listings on a peak night. The popularity of this business concept has created a sharing economy which allows anyone with property to become a microentrepreneur.

Our research delves into what influences predicting the price of an Airbnb listing which can range from around ten dollars to thousands within the same city. We looked specifically at New York City, an international hub ripe with Airbnb listings for every kind of traveler, varying from a shared room for the low-budget student to luxury penthouses in Manhattan fit for a celebrity.

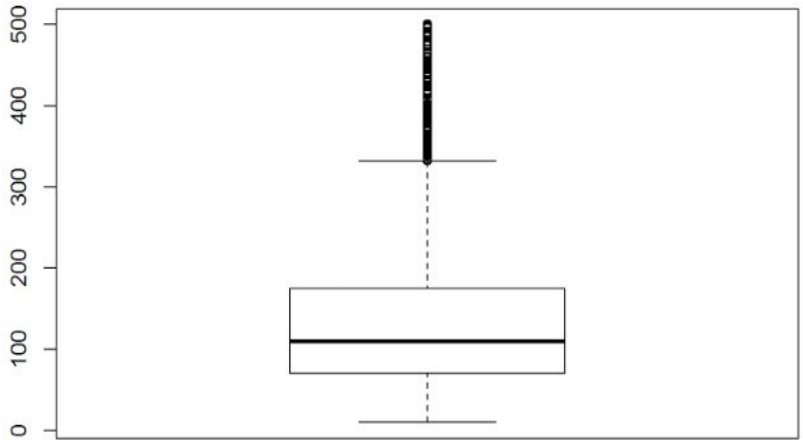
Data Analysis

We chose to perform our analysis on the 2015 Airbnb data set for the listings in New York City. In preparing the data, we filtered unusual observations such as listings that charged more than 100 dollars for extra people, and any observations where the minimum number of nights was greater than 90. For maximum number of nights, the third quartile and the median were equal which means that a value greater than 1,125 was a strange occurrence; therefore, we removed all of those observations. Because Airbnb appeals to such a wide audience, there were 693 listings out of 39,553 which charged more than $500.00 a night. As you can see in the boxplot below, this led to a very skewed distribution.



Price

Therefore, we removed those data points since the unusually high prices were skewing the analysis. The boxplot below reflects the improved distribution for price according to the average person's budget.



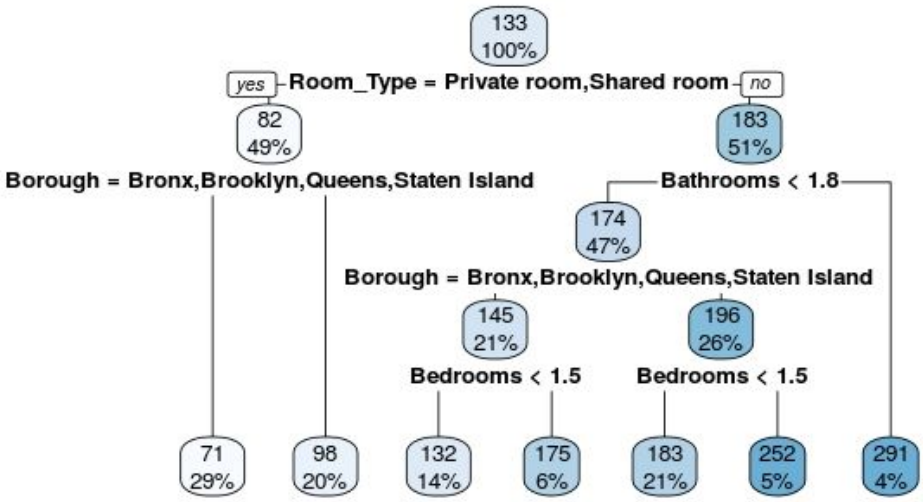
Price

We proceeded to split the data into test and training. For the training and test data we fit the same regression decision tree model and compared the MSEs as measure of performance of the model.

Methods & Results

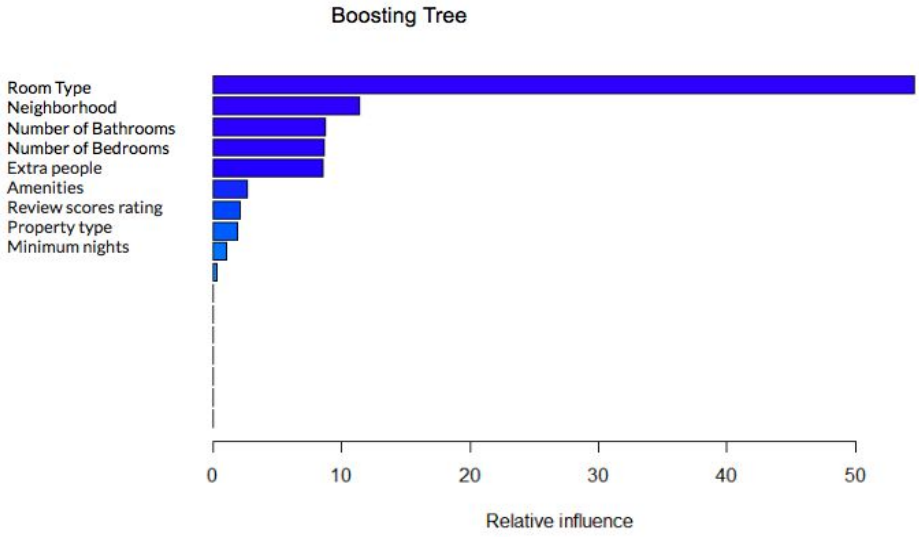
We chose to model price based on a regression decision tree as this would be an easily interpretable model for someone who is interested in determining how much to charge for their New York City Airbnb listing. At the top of the tree, the deciding predictor is room type. If the tree cut off at the first level, 100% of our listings will be priced $133.00. At the second level, the room type is divided into either private/shared room or entire home/apartment. When the listing is a private/shared room, we expect the price to be $82.00 on average. Then if the listing is located in the boroughs of Bronx, Brooklyn, Queens, or Staten Island, we expect the price of a listing to be $70.00 on average. If the listing is located Manhattan borough, we expect the price to be $99.00. If the room type is an entire home, we expect the price of a listing to be $183.00 on average. If there were less than 1.2 bathrooms, we would expect the price to be $172.00 on average. Then if the borough is Bronx, Brooklyn, Queens, or Staten Island, we expect the price of a listing to be $144.00 on average. If the listings in these neighborhoods have less than 1.5 bedrooms the average price is $131.00; whereas, if they have more than 1.5 bedrooms it would be $173.00. If the borough is Manhattan, we would expect the average price to be $194.00. If there were less than 1.5 bedrooms, a listing would be $181.00 on average. If there were more than 1.5 bedrooms, we would expect the price on average to be $251.00. In a listing with more than 1.2 bathrooms in Manhattan, we would expect the price on average to be $319.00.

Modeling AirBnB Listing Prices Based on a Decision Tree



.

In order to determine whether we could improve the regression decision tree, we fit a boosting tree on the testing and training data. Then, we calculated the test MSE of our new model to compare with the test MSE of our original regression tree. Our boosting test MSE was 3042.09 in comparison with our regression test MSE of 3613.839; the smaller test MSE indicates that we have improved our model. The boosting tree makes its decisions based on relative influence of predictors. It showed that room type and borough were the most important deciding factors for price. Below, we can see from the boosting tree which predictors have the strongest influence relative to one another. Just as in our regression decision tree, the four top predictors are *room type, neighborhood, number of bathrooms*, and *number of bedrooms*. *Extra people* comes in right after as having about the same amount of influence in predicting price as the number of bedrooms and bathrooms.



Conclusions & Future Work

From our analysis, we can infer that when a host is trying to determine how much to charge, the most important deciding factors are room type, number of bedrooms/bathrooms, and neighborhood. In the future, we would be interested in expanding the variables that we consider for our model, and looking into variables that are not only quantitative or categorical, but also textual so that we could do a textual analysis on aspects such as each listing's description. We could also take into consideration the listings' popularity to find the optimal price. We hope that our research can be useful to anyone interested in joining the sharing economy and listing their home or property with Airbnb.