# *Using Machine Learning Algorithms to Predict Airbnb Prices and Rental Probabilities in New York City*

**PRINCIPAL INVESTIGATORS:** BRADEN POE (BPOE@WISC.EDU) & KEVIN VAN LIESHOUT (KVANLIESHOUT@WISC.EDU)

## Purpose:

Airbnb has quickly become a leading tourism driver across the world. The company's ability to connect visitors with quick and easy places to stay in over 81,000 cities and 191 countries has simplified renting. In this project, we utilized a 2019 New York City Airbnb dataset in order to determine indicators of price and rental probability of individual properties. By applying machine learning techniques to a compiled dataset, we highlight features that drive price and rental probability and fold those features into accurate predictions using machine learning algorithms.

In our analysis we run a plethora of simple and complex algorithms. For price, we find that an ensemble method of bagging, using a base model of Random Forest predicted the best out of sample for Airbnb prices with a RMSE of about 74 dollars. In terms of rental probability, we find a weak, but useful predicative algorithm using a one-degree logistic regression, which accurately predicts the probability of Airbnb rentals 65% of the time. Neither algorithm is quite specific enough to be useful on a consumer-basis, however they uncover several trends that are of interest to potential Airbnb owners.

## Analysis:

### 1. Data:

Our main dataset comes from Kaggle, which provided us with a 2019 New York City Airbnb dataset that includes data from every Airbnb rented over the year. The dataset has 45000 observations with variables such as "price", "reviews per month", "days listed per year", "minimum nights", "number of reviews", "neighborhood", and "borough". Based off the listing description, we were able to generate a new variable using sentiment analysis on this description to see whether how an owner described their dataset impacted the prices. We generated heat maps to analyze the locations of properties with the highest rental prices and probabilities (Appendix A).

Some of the data that we introduced into the model start with Walk Scores which are described as, "the only international measure of walkability and the leading provider of neighborhood data to the real estate industry." Within Walk Score, we will also derive similar metrics for Biking and Transit. These metrics will allow us to determine the relationship that accessibility of an Airbnb location has on its prices and how predictive walk score is of price. This data comes out of the Walk Score database using their API to access it.

Other datasets we utilize are average housing price by neighborhood, number of restaurant inspections by neighborhood to represent food quality, crime data by neighborhood, recognized healthy stores by neighborhood, unemployment by neighborhood, and an indicator variable for borough. These data sets all range from 2016 – 2019 and come from Kaggle and or the NYC public dataset site. We merged these new datasets with the main database by using borough as the index. Lastly, we generated variables to look at rentals and help analyze rental probability. The final csv file is the culmination of all the data and the new variables we created within our final dataset.

There are some acknowledgements that we want to make with respect to our dataset. First, there is no time stamp so there is no way to control for seasonality or cyclicality of typical rental seasons. There are many trends that we would like to pursue if we had this timestamp in the dataset. Analyzing dense rentals around times of big events/holidays like the US Open, NYC marathon, New Years and more would add detail to our report however our dataset limits this ability. Additionally, since the dataset includes yearly totals for reviews, number of stays, etc., we assume all variables are uniformly distributed across the year. This assumption plays a crucial role in rental probability and was kept in mind when interpreting the validity of results.

# Methodology:

## 1. Price Prediction

We employed multiple machine learning techniques ranging from simple linear regressions to neural networks along with deploying feature section and a variety of ensemble methods. In this section, we will focus on the method that was more accurate in predicting NYC Airbnb prices. The histogram of prices after we cleaned the data of outliers is below. (Appendix A).

The definition of a bagging regressor from scikit learn[1] is: "a Bagging regressor is an ensemble meta-estimator that fits base regressors each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction." Bagging stands for bootstrap aggregating which compiles random subsets of training data to use to train a model with the best score it can find. We chose a random forest as the best estimator to be a part of this ensemble technique. A random forest model is a type of decision tree that creates different sets of decision trees leading to better predictions than individual trees themselves. Tony Yiu[2] describes the concept in a sense that "a large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models."

We chose this set of methods because with the belief that since the rentals are not normally distributed across New York City, the use of bagging allows the algorithm to grab different representative samples of data points across all boroughs and will penalize itself for misrepresented data. The use of the random forest helps combine variables that are then useful for determining price.

## 2. Rental Probability Prediction

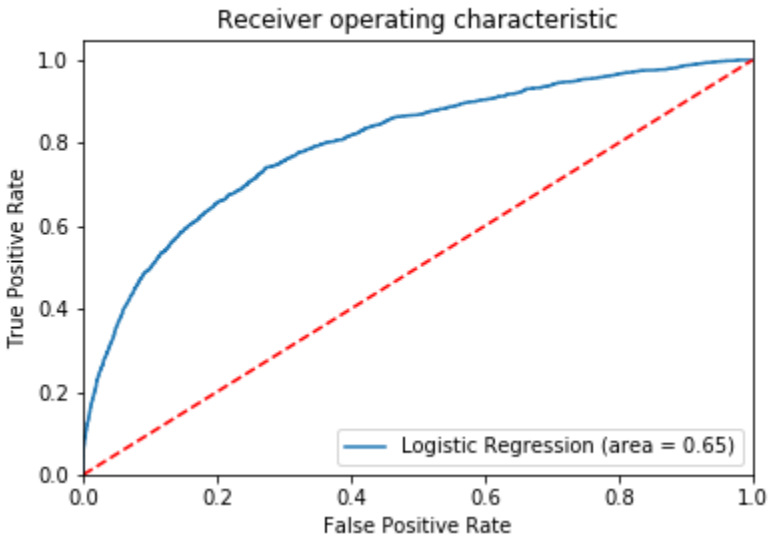
Predicting rental probability is a relatively simple task given proper data, but as stated, our dataset presented limitations that made predictive tasks difficult. Since our data is cross-sectional and lacks timestamps for individual rentals, we were forced to make a handful of strong assumptions during data manipulation. Without going too in depth, the total number of reviews over the year was used as a proxy for total stays at a specific property. Estimated total stays were then weighted based on the days each unit was available over the 2019 year and a "rentals per week" variable was generated.

Binary classification was applied to a new rental censor variable where "r_censor" equals 1 if "rentals per week" is greater than 0.5 and equals 0 if rentals per week is less than 0.5. By making this classification, we implemented a few additional assumptions: 1) all guests are equally likely to leave a review no matter where in the city they stay, 2) factors determining yearly property availability do not

[1] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingRegressor.html
[2] https://towardsdatascience.com/understanding-random-forest-58381e0602d2

determine rental probability, 3) misclassifications would occur for “rentals per week” between 0.6 and 0.4, but those misclassifications would not severely impact our model.

The limitations of our data set, and the strong assumptions required, led to a ROC curve with an unsatisfactory area. Our logistic model accurately predicts the probability of Airbnb rental 65% of the time and without additional micro-level data there’s not much room for model tweaks that could lead to greater accuracy. Despite predictive inaccuracy, the probability model still provides many helpful insights for potential Airbnb owners and renters.




## Results:

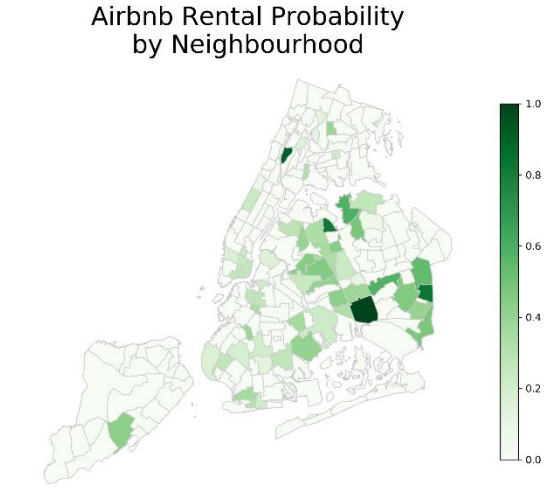
### 1. Price Prediction

The bagging ensemble method using a random forest as the baseline model was the best performer in our prediction analysis. The training score of fit was a .8 and the test score accuracy was a .48, not the best result. It yielded a root mean squared error of about 74 dollars. Although it did not perform great out of sample, this was the closest RMSE we could find. Even after cleaning the data of price outliers, we still struggled with an abnormal distribution in our variables which lead to high penalties in a root mean squared error framework. We are most likely overfitting the data with the number of variables we created, and we need to utilize the use of greedy algorithms better, but these can be computationally expensive. We are happy with this result but not completely satisfied; future research, model simplification, and a more robust dataset would significantly increase accuracy.

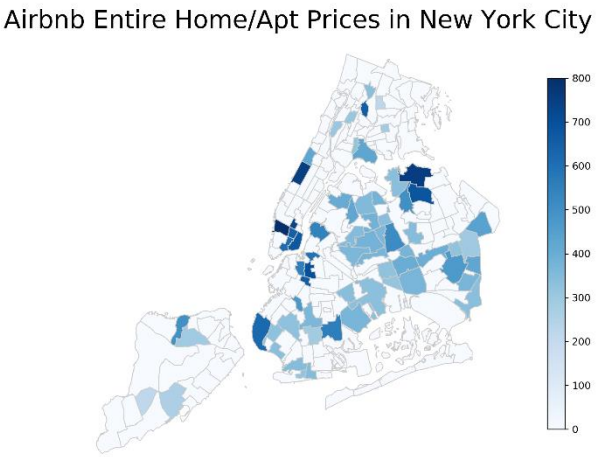
### 2. Rental Probability Prediction

While our logistic model is not ideal for the specific predictive tasks, it still provides plenty of broad takeaways and allows us to tease out rental heterogeneity between neighborhoods. When we compare average neighborhood rental prices to average neighborhood rental probabilities, some stark differences arise. Higher priced Airbnb’s in neighborhoods such as Manhattan’s Upper West Side, East Village, and Greenwich Village also have some of the lowest rental probabilities among all neighborhoods. It is natural to think that the opposite must be true for lower priced properties, but we do not see that

trend in the heat maps below (prices in USD). Properties with lower prices in Central Brooklyn neighborhoods have nearly equivalent rental probabilities to high-priced Manhattan locations.




Source: Kaggle NYC Open Data, 2019




Source: Kaggle NYC Open Data, 2019

The discrepancy between price and probability signals a few important trends with Airbnb rentals in New York City. One being, prime location does not necessarily maximize profits for an Airbnb owner. A potential property investor would be much wiser to consider property in Queens, which holds several neighborhoods than contain average-to-above average prices and rental probabilities—a rare combination within the city. In addition, potential Airbnb guests looking to stay in some of New York City's upscale areas should not book far in advance. Low rental probabilities indicate that supply outweighs consumer demand, which is further supported by the upward effect our model imposed on rental probability.[3] An optimal search strategy for consumers would be to not book far in advance, but rather wait until you are within the month of travelling. Most units in upscale areas have a high probability of being unrented in any given week, so it is unnecessary to pay a premium for security of stay.

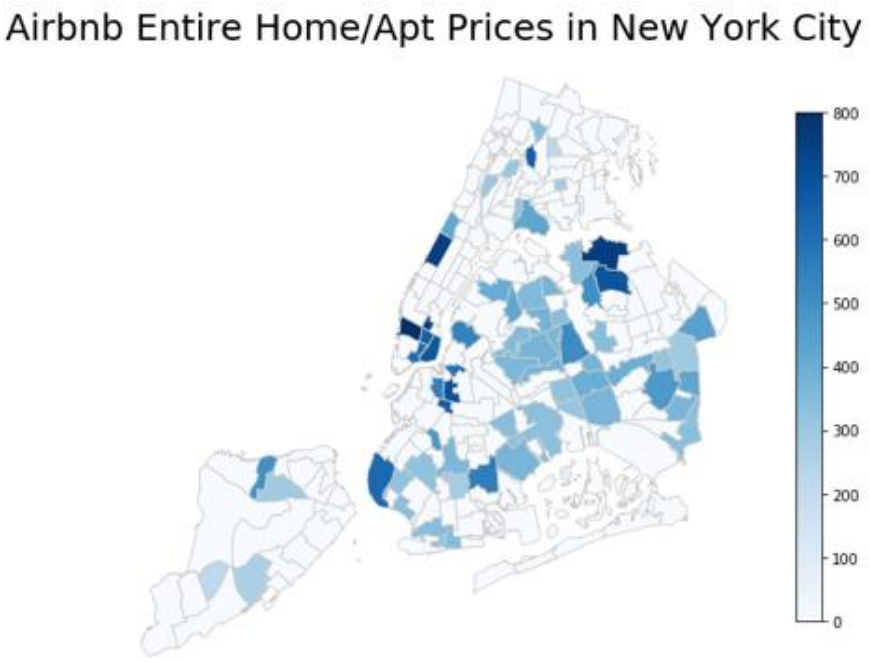
## Conclusions and areas for future research:

In summary, the machine learning algorithms we implemented ended up aiding us more in understanding the composition of Airbnb properties in New York City rather than predicting individual rental prices and probabilities. The results highlighted large discrepancies between price and rental probability within many of the cities neighborhoods, which suggests that there are variables preventing Airbnb markets from adjusting properly. This conclusion opens up several areas for future research including rental market frictions, Airbnb demand estimation, and optimal search strategy in rental

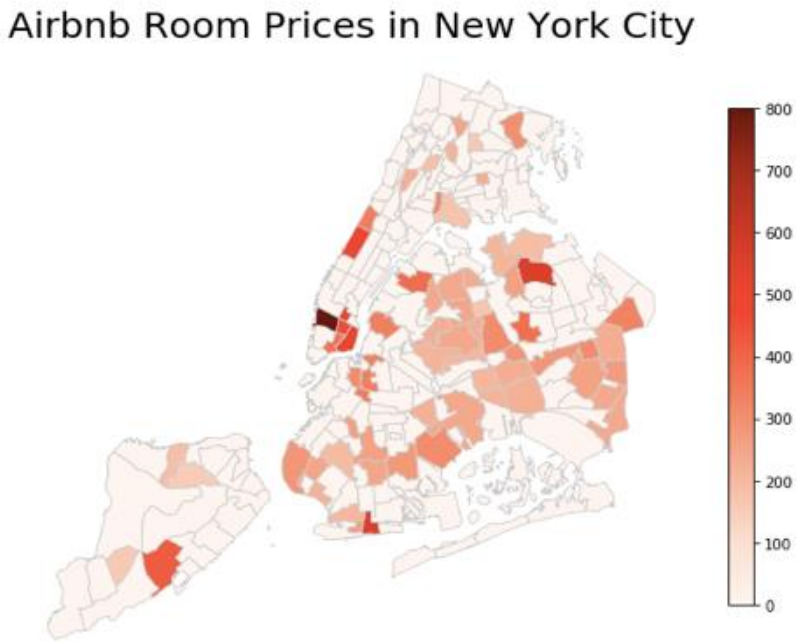
[3] By censoring at 0.5 rentals per week (rpw) and not at 1 rpw, we attempted to include fringe properties (between 0.5 and 1.0 rpw) as ones likely to be rented. We took this measure as a precaution, since censoring at one risks misclassifying many observations. By essentially boosted the "r_censor" variable with a few additional 1's, we may slightly bias rental probability upwards across the board. Thus, low probability properties would see an even lower probability with censoring greater than 0.5.

markets. Further, if we manage to acquire timestamped, time-series data for Airbnb rentals, we would be able to predict prices and rental probabilities with the accuracy and precision we set out to obtain.

## Appendix: Figures and tables




Source: Kaggle NYC Open Data, 2019




Source: Kaggle NYC Open Data, 2019