The Cost of Staying in the Big Apple

July 2020
Group 7

Executive Summary

This report will provide an overview of the methods applied to develop a better price recommender for Airbnb listings in New York City. It will also highlight the chosen model and overall insights.

Project Goals

The goal of this project was to determine the feasibility of developing an accurate price recommender tool for Airbnb. If successful, this tool would replace the existing algorithm, which has received complaints from users in the past about being inaccurate and unhelpful. A price recommender that is more precise would help increase Airbnb's ease of use for hosts, and could potentially increase bookings by more accurately reflecting the supply and demand of a certain offering in the price of a listing. Since Airbnb makes a percentage of all booking rates, an increase in bookings would positively impact the revenue of the company.

Method

The overall approach was to analyze the data and evaluate the application of various machine learning algorithms to determine if a model could be constructed that would achieve the desired accuracy. The dataset provided by Airbnb included information on listings in New York City from 2011-2019, with close to fifty thousand observations. The dataset was used to train various models on the pricing patterns within the data. The algorithms included in model testing were KNN, Random Forest, Gradient Boosting Machine, and Linear Regression. Root mean squared error (RMSE) was used to evaluate and compare the various models. Ultimately, the model chosen was a GBM model, which had a best RMSE of 0.439.

KNN

Using 10-fold cross validation, different k values from 1 to 100 were tried for each combination of variables to find the optimal model and corresponding k value. Distinct combinations of variables were also run, to see which subset of variables would result in the most accurate model. Ultimately, the most successful model was the one that used all of the independent variables (days since last review, availability, listings count, number of reviews, reviews per month, minimum nights, bronx, queens, staten island, brooklyn, shared room, private room). This model had an optimal k value of 17, and yielded the lowest RMSE value of 0.443.

Tree-Based Methods

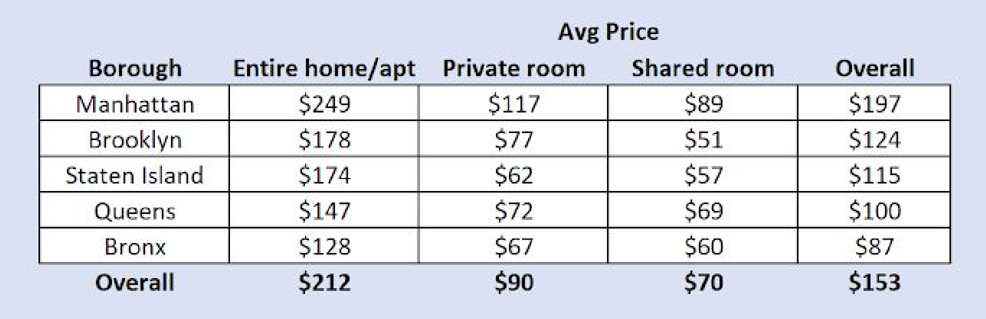
Several tree-based models were evaluated, including a classic decision tree algorithm, random forest, and gradient boosting machine (GBM). The parameters for the final random forest model were: mtry = 3, number of trees = 500, and max nodes = 15. For the gradient boosting model, the parameters used were: distribution = gaussian, max tree depth = 4, learning rate = .2, number of trees = 5000. The best random forest model had an RMSE of 0.491, compared to the best RMSE for GBM of 0.439. The variable importance ranking from both models had room type at the top of the list by a large margin, suggesting that whether or not a listing is for a private space is the biggest indicator of ultimate pricing. Longitude and latitude also ranked highly for both models, which is not surprising given the different cultural offerings available to Airbnb guests in each of the five boroughs.

Stepwise Regression

A stepwise regression model was also evaluated. For this model, an additional set of dummy variables was created from the neighborhood feature in order to get a better sense of variable importance. Ultimately, the model with the lowest AIC utilized all of the variables. The optimal stepwise selection model returned an AIC of 45662.78, RSS of 429818, and MSE of 2.964. Lasso and Ridge regression models were also evaluated, with optimal lambda of 0.01 and 0.08, each resulting in an RMSE of just under 0.45.

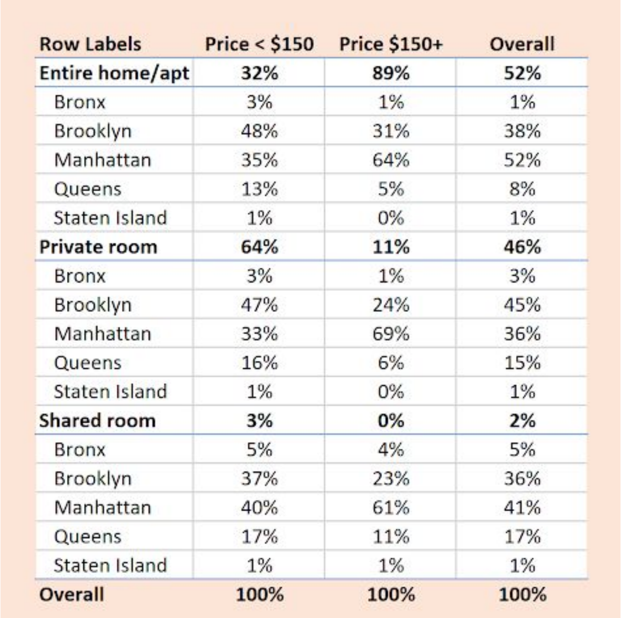
Insights and Recommendations

After evaluating the various models and their variable importance rankings, it is clear that the biggest predictor of an Airbnb listing price in New York City is the room type. Specifically, whether or not a listing is for a private room or an entire apartment has the largest impact on the listing price. In fact, the average price for renting an entire unit in NYC is $212 versus $89 for a private room and just $70 for a shared room. This makes sense, as most people enjoy their privacy and prefer to have their own space when traveling to come back to at the end of the day. Another factor that plays a big role in setting prices is location, which is also understandable. There are significant differences between the five boroughs with regards to cultural and culinary offerings. Manhattan, for example, is home to some of the world's most visited attractions, including Central Park, the Empire State Building, the Metropolitan Museum of Art, and Broadway. As one might expect, Manhattan also has the highest average Airbnb listing price of the five boroughs at $197 a night (vs the overall NYC average of $153). In fact, 64% of higher-priced listings (those over $150 a night) can be found in Manhattan.



		Avg Price		
Borough	**Entire home/apt**	**Private room**	**Shared room**	**Overall**
Manhattan	$249	$117	$89	$197
Brooklyn	$178	$77	$51	$124
Staten Island	$174	$62	$57	$115
Queens	$147	$72	$69	$100
Bronx	$128	$67	$60	$87
Overall	**$212**	**$90**	**$70**	**$153**

A deeper dive into the similarities of listings at various prices also reveals some interesting insights. The majority of listings under $150 are private rooms (64%) followed by entire units (32%), whereas the overwhelming majority (89%) of higher-priced listings are for entire units.



Row Labels	Price < $150	Price $150+	Overall
Entire home/apt	**32%**	**89%**	**52%**
Bronx	3%	1%	1%
Brooklyn	48%	31%	38%
Manhattan	35%	64%	52%
Queens	13%	5%	8%
Staten Island	1%	0%	1%
Private room	**64%**	**11%**	**46%**
Bronx	3%	1%	3%
Brooklyn	47%	24%	45%
Manhattan	33%	69%	36%
Queens	16%	6%	15%
Staten Island	1%	0%	1%
Shared room	**3%**	**0%**	**2%**
Bronx	5%	4%	5%
Brooklyn	37%	23%	36%
Manhattan	40%	61%	41%
Queens	17%	11%	17%
Staten Island	1%	1%	1%
Overall	**100%**	**100%**	**100%**