Modeling Price and Popularity of AirBnB listings in New-York

Melody Jiang and **Raphael Morsomme** and **Ezinne Nwankwo**

Abstract

In this paper, we seek to understand factors about a typical Airbnb listing that influence its price and popularity. We analyze a dataset with a total of 48,895 Airbnb listings in New York and 16 variables. We focus on feature engineering and fit two types of models: linear model using Bayesian Model Averaging and random forest with popularity and price as the outcome variables. Ultimately, our main finding in which all of our models agreed with was that room type and proximity to metro stations played a huge role in determining price and popularity.

1. Introduction

Airbnb has quickly grown into one of the largest online marketplace for arranging long or short term stays at homes and appartments owned by people. Airbnb does not own any of the property on its site; instead it acts as a broker between customers and hosts and earns a commission from each booking. Hosts have full control over aspects such as pricing, how to advertise their listings, and how often their listing is available throughout the year. For this project, we are interested in modelling price and popularity of listsings in order to best advise hosts on tactics that will maximize their profit and popularity on the platform. This includes (i.) identifying influential factors on the price and popularity of the listing (ii.) identifying heterogeniety across boroughs and neighborhoods, in particular which ones have the heaviest traffic and highest prices (iii.) identifying hetoergeneity across listing type and (iv.) providing recommendations on listing location and names.

2. Methods

2.1. Data Preparation

The quality of the data is questionable and the meaning of some variables is unclear. For instance, the data set contains listings with a price superior to USD$5,000$ per night or equal to 0, a minimum number of nights superior to $1,000$ and an average number of monthly review superior to 50. It was also unclear what the price variable represents (average price of booking, average listed price or current listed price?) or how the variable availability_365 was constructed (average number of available days per year, number of available days in the last year, current number of available days?).

For these reasons, we decide to simplify the dataset by removing listings that are not typical of the Airbnb platform. That is, we remove listing for long-term stays and listings that are owned by businesses (some owners possess several dozens of listings), since we believe that the set of factors determining their price and popularity strongly differs from that of a typical *short-term*, *privately owned* Airbnb listing. We also exclude listings that have not been reviewed in the last 12 months since factors that mattered several years ago may no longer be relevant. The resulting data set contains $24,255$ *short-term*, *active* and *privately owned* listings.

2.2. Feature Engineering

2.2.1. Spatial Variables

Figure 4 suggests that the most expensive listings are located close to metro stations. We therefore construct a variable indicating the proximity of the closest metro station. The proximity between two locations x and

y is computed with $\text{proximity}(x, y) = \dfrac{1}{\text{dist}(x, y)}$ where $\text{dist}(x, y)$ measures the distance between x and y. We decide to use the *manhattan* distance

$$\text{dist}_{\text{Manhattan}}(x, y) = |lat_x - lat_y| + |long_x - long_y|$$

which approximates the distance traveled by a pedestrian walking on the perpendicular streets of New-York.

Similarly, we compute the average proximity to the 36 attractions with

$$\text{proximity}_{\text{attraction}}(x) = \frac{1}{36} \sum_{i=1}^{36} \frac{1}{\text{dist}_{\text{Manhattan}}(x, \text{attraction}_i)}$$

2.2.2. Textual Variables

First, we conducted a sentiment analysis on the listing names in order to construct a variable indicating the sentiment of the listing name, that is, how positive the name sounds. The sentiment of a document $W = (w_1, w_2, \ldots, w_n)$ composed of n words is

$$\text{sentiment}(W) = \frac{1}{n} \sum_{i=1}^{n} \text{dictionary}(w_i)$$

where $\text{dictionary}(w_i)$ indicates the sentiment of the word w_i according to some sentiment dictionary. Since the listing names are relatively short, we decide to use the Affin dictionary which provides a gradual sentiment metric ranging from $-5, 5$; the other existing sentiment dictionaries only provide a binary metric (-1, 1) and would provide a sentiment that is too coarse for such short documents.

In addition, we attempted to model the origin of the owner's name. The rationale for this is that we expect renters to be less likely to book a listing whose owner has a name that does not look familiar. We obtained the relatively frequency of a someone's name in the data as a measurement of how common the name is. Since some owners own multiple listings, we filter by unique ID before computing the frequencies.

2.3. Model

We decided to implement two models, Bayesian Model Averaging (BMA) in linear models and random forests (RF). We build two models separately for log price and log popularity using all the variables as a linear combination of predictors. For the BMA models, we exclude latitude and longitude because there is no inpretable relationship between the magnitiude of, say, latitude and outcome variables. After sensitivity analysis, we decided to use a Cauchy prior for the predictors and a uniform prior over the model space, which assigns equal probability to all the models. We use an MCMC algorithm with 10^{16} iterations to sample from the model space of 2^{15} models. Lastly, we approximate the marginal inclusion probabilities of the predictors by taking the p-values from p simple linear regressions. The posterior inclusion probabilities (PIP) of each predictor was primarily what we used to determine influential factors in price and popularity of a listing.

For the RF models, we used all the predictors including longitude and latitude since this class of models does not make any linearity assumptions between predictors and outcome variable. We build these models using 1900 trees each built on subsamples of 19000 data points. Due to computational challenges, we could not go beyond this number of trees and subsamples. For the split criteria, we randomly sampled $m = \frac{p}{3} = 4$ variables as candidates for the split (Note that this is the default value for regression) and $n_{leaf} \geq 5$. We used a variable importance measure that is based on an increase in node purity, or Gini-based importance. It is is calculated based on the reduction in sum of squared errors whenever a variable is chosen to split.

3. Results

3.1. EDA

Figure 1 shows the distribution of required minimum number of nights to book a listing. We observe that minimum number of nights is concetnrated at below 14 days and around 30 days. Figure 2 displays the

distribution of days that listings are available for booking in a year. We see that there are data concentrated at 0, which means these listings are not open for booking. Such observation would inform our data cleaning.

To address the quesstion of whether the type of listing (shared room, private room, entire home) vary across neighbourhoods, we performed a chi-square test and plotted the results in Figure 3. The p-value of the chi-square test was less than 10^{-16}. In Figure 3, the size of dots represents the absolute standardized residuals. The color represents the value of standardized residuals. We see that difference in room types is most pronounced in Manhattan and Queens. Manhattan has more entire home than expected and Queens has more private room than expected.

Figure 4 shows a spatial map of of listings, metro stations, and attractions. Black dots represent metro stations, gree dots attractions, blue dots listings priced at bottom 80%, and red dots are listings priced at top 20%. We observe that listings priced at top 20% distribute close to metro stations. This observation motivates us to include spatial in formation of metro stations as explanatory variables.

3.2. Main Findings

Main findings.

3.3. Model Checking and Sensitivity Analysis

Our model was very sensitive to outliers. When we fit our models, we noticed some outliers in the residuals plot, and thus decided to remove points that were three times the standard deviation away from the mean of the outcome. After refitting, the model residual plots looked much better and had better R^2 values. Unfortunately, for the popularity model, the residual plots indicate that our models are not accounting for some relationship in the data very accurately , as shown in Figure 5 and Figure 6. For the BMA models, we also varied the choice in prior by testing out the g-prior for g = 1, 5, 8, 50, 100, 500, 1000. The results did not change for the different priors. Lastly, for the RF models, we varied m and the minimum n_{leaf}. The final values that we used for those parameters did not result in a change in MSE.

4. Conclusions and further discussion

(i) As shwon in our main findings, the most influential factors on price that BMA and RF agree on are room type and borough. For popularity, influential factors that both models agree on are minimum nights, listing sentiment, and borough.

(iii) As shown in the EDA section, the type of listing does vary across neighbourhoods, since the p-value indicates significance. We have also seen the most pronounced difference in Mahhattan and Queens, where Manhattan has more entire home than expected and Queens has more private room than expected.

(ii)(iv) Yes, there is heterogeneity across boroughs, and Manhattan is the most pricy. To make our listing the most pricy, We would locate an entire home/apartment in Manhattan and use a common name as host. Whether to use positive words in the name of listing does not affect the price.

We have seen in Results section that variable importance of BMA and random forest do not agree completely. This could be due to nonlinear trend that is hard to pick up in EDA. We could investigate this issue further in future work.

Regarding BMA for popularity, we have seen the anti-intuitive result that popularity declines as sentiment of listing name becomes positive. This calls for a more detailed analysis of listing names.

For future improvements, we would like to include spatial modeling. For example, we could model spatial relationship between houses, as James et al. (2005) has shown that OLS including just boroughs as spatial information does not work well . In addition, we could utilize domain-specific knowledge to improve our modeling of price. We could potentially use a finite mixture model to identify submarkets as in the study by Belasco et al. (2012). We could also draw from knowledge in marketing to improve our feature engineering, for example consider pricing as an explanatory variable for popularity.

Appendix A. Figures

Figure 1: Distribution of minimum number of nights.

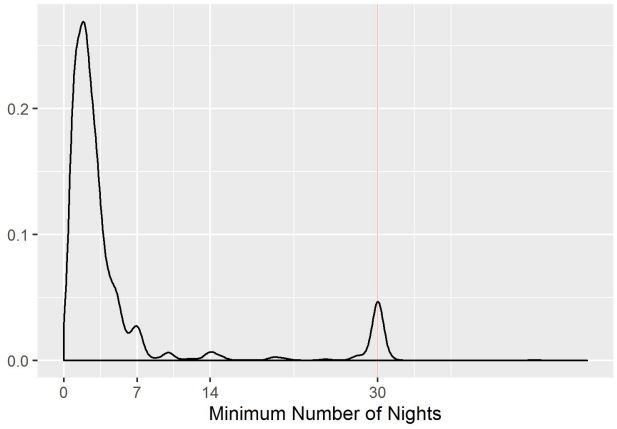


Figure 2: Distribution of number of days available for booking.

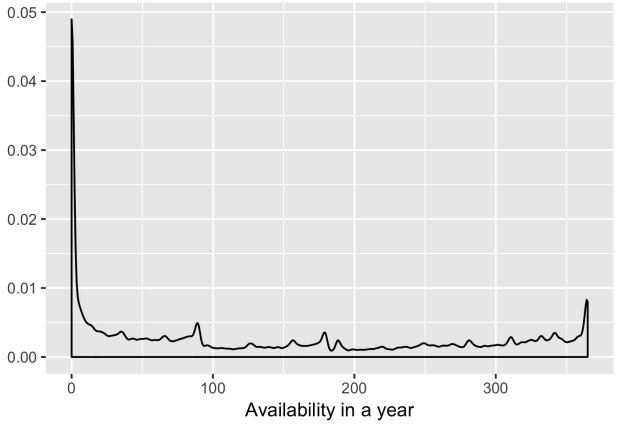


Figure 3: Output from chi-squared test.

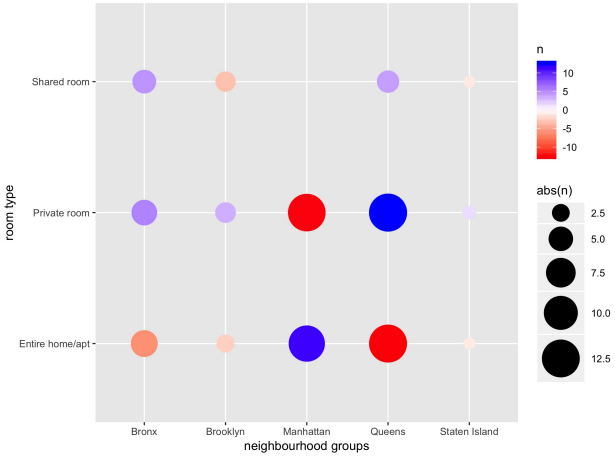


Figure 4: Map of listings, metro stations, and attractions. Black dots are metro stations, gree dots are attractions, blue dots are listings priced at bottom 80%, and red dots are listings priced at top 20%

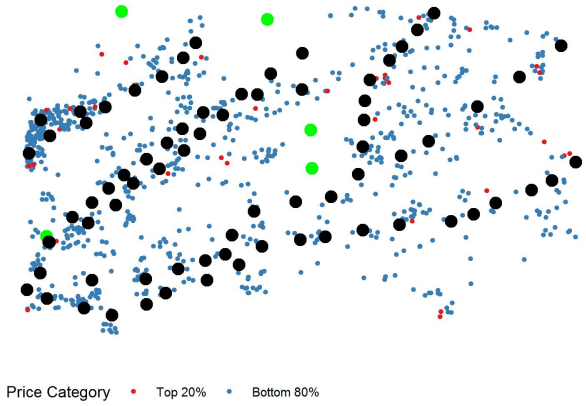


Figure 5: Model diagnostics for BMA for price

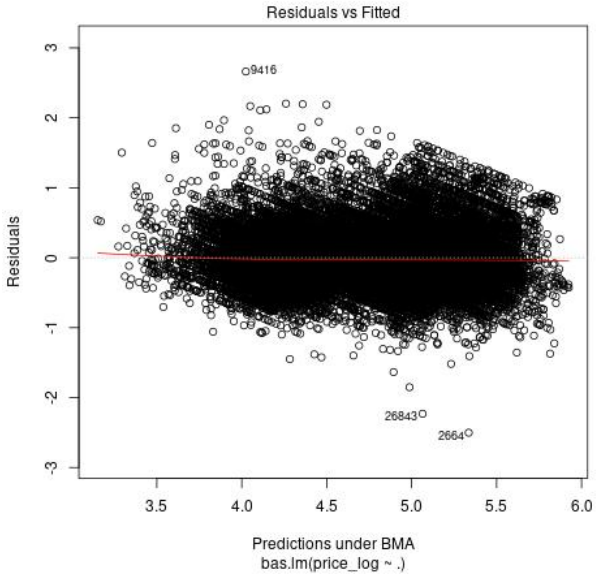


Figure 6: Model diagnostics for BMA for popularity

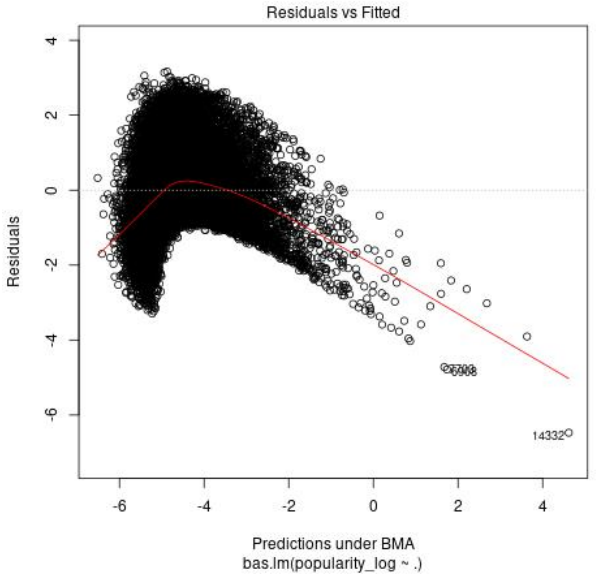
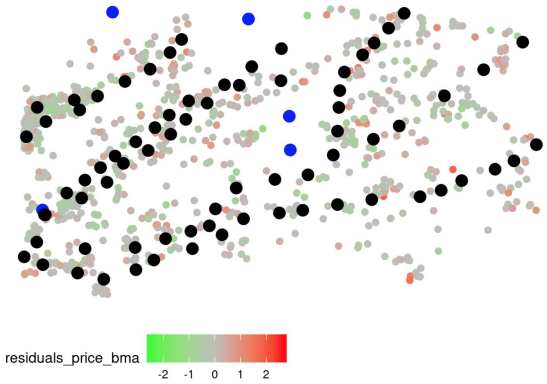


Figure 7: Residuals for price BMA model



Appendix B. Full Model Output

	Predictors	Variable.Importance
1	longitude	978.10
2	proximity_metro	231.64
3	reviews_per_month	107.88
4	listing_sentiment	101.15
5	number_of_reviews	100.73
6	neighbourhood_group	95.16
7	latitude	94.24
8	room_type	90.71
9	name_listing_length	89.26
10	minimum_nights	88.53
11	popularity_log	61.00
12	name_host_special	53.69
13	proximity_attraction	47.78
14	name_host_freq	39.16
15	listing_count	16.42

Table 1: Variance Importance for RF Price Model

	Predictors	Variable.Importance
1	number_of_reviews	199.74
2	latitude	62.77
3	listing_sentiment	59.50
4	minimum_nights	57.85
5	neighbourhood_group	54.93
6	room_type	50.03
7	reviews_per_month	49.69
8	longitude	45.87
9	name_listing_length	41.50
10	name_host_special	35.34
11	popularity_log	33.77
12	proximity_attraction	20.73
13	listing_count	8.76
14	name_host_freq	8.72
15	proximity_metro	0.42

Table 2: Variance Importance for RF Popularity Model

	Predictors	PIP	Estimate	Lower.Confint	Upper.Confint	Significance
1	Intercept	1.00	4.70	4.70	4.71	*
2	name_host_freq	1.00	0.07	0.07	0.07	*
3	Neighbourhood_group:Manhattan	1.00	0.17	0.15	0.18	*
4	number of reviews	1.00	-0.01	-0.02	-0.01	*
5	Room_type:Entire home/apt	1.00	0.74	0.73	0.75	*
6	availability_365	1.00	0.00	0.00	0.00	
7	name_listing_sentiment	1.00	0.00	0.00	0.00	
8	last_review	1.00	-0.00	-0.00	-0.00	
9	Neighbourhood_group:Queens	1.00	-0.11	-0.13	-0.10	*
10	name_host_special:True	1.00	10.42	7.16	13.66	*
11	Room_type:Shared room	1.00	-0.51	-0.54	-0.47	*
12	calculated_host_listings_count	1.00	-0.01	-0.02	-0.01	*
13	type_stay:Long	1.00	0.00	0.00	0.00	
14	Neighbourhood_group:Bronx	1.00	-0.17	-0.20	-0.14	*
15	name_listing_length	1.00	0.06	0.04	0.08	*
16	proximity_metro	0.89	-0.00	-0.00	0.00	
17	proximity_attraction	0.50	-0.01	-0.04	0.00	
18	reviews_per_month	0.07	-0.00	-0.00	0.00	
19	Neighbourhood_group:Staten Island	0.02	-0.00	0.00	0.00	

Table 3: Coeffients and 95% Confidence Intervals from Price BMA model

References

Eric Belasco, Michael Farmer, and Clifford Lipscomb. Using a finite mixture model of heterogeneous households to delineate housing submarkets. *Journal of Real Estate Research*, 34(4):577–594, 2012.

Valente James, ShanShan Wu, Alan Gelfand, and C Sirmans. Apartment rent prediction using spatial modeling. *Journal of Real Estate Research*, 27(1):105–136, 2005.

	Predictors	PIP	Estimate	Lower.Confint	Upper.Confint	Significance
1	Intercept	1.00	-4.47	-4.49	-4.45	*
2	listing_count	1.00	-0.10	-0.12	-0.09	*
3	minimum_nights	1.00	-0.07	-0.08	-0.07	*
4	name_listing_length	1.00	0.02	0.02	0.02	*
5	number_of_reviews	1.00	0.01	0.01	0.01	*
6	listing_sentiment	1.00	-0.16	-0.22	-0.09	*
7	proximity_attraction	0.99	0.02	0.01	0.03	*
8	Room_type:Shared room	0.85	-0.20	-0.34	0.00	
9	Neighbourhood_group:Staten Island	0.73	-0.22	-0.44	0.00	
10	proximity_metro	0.22	-0.00	-0.00	0.00	
11	Room_type:Entire home/apt	0.10	0.00	-0.00	0.04	
12	Neighbourhood_group:Manhattan	0.09	0.00	-0.00	0.04	
13	name_host_freq	0.04	0.16	0.00	0.00	
14	name_host_special:True	0.03	-0.00	0.00	0.00	
15	Neighbourhood_group:Queens	0.03	-0.00	0.00	0.00	
16	Neighbourhood_group:Bronx	0.03	-0.00	0.00	0.00	

Table 4: Coeffients and 95% Confidence Intervals from Popularity BMA model