ORIE 4171

Learning with Big Messy Data

Project Midterm Report

The Prediction of Airbnb Price in NYC: the Effect of COVID-19 in the Airbnb Market

December 2020

Linna Yue (ly389)

Sophie Zhao (sbz24)

Nicolás Oriol Guerra (no227)

1. Introduction

The Covid-19 pandemic gave the global economy a huge shock. The travel industry was clobbered heavily. Many trips were cancelled or held due to Covid-19 which has a devastating effect on the hotels, Airbnb and other short-term rental properties. Especially after March 22, 2020, the NYC government closed the non-essential economy and international travel restrictions enforced on February 2 and March 11. According to Forbes[1],47% of Airbnb hosts don't feel safe renting to guests and 70% of guests are fearful to stay at Airbnb apartments. The revenue from June to August was expected a 44% decrease for the Airbnb apartments. 47% of the hosts provided month-long rental, and 29% reduced the listing prices.

This project aims to explore the Airbnb data before and after Covid-19, and test different price prediction machine learning models to determine the influence of pandemic on the Airbnb listings. The assumptions are that the Airbnb prices, number of rentals might decrease after Covid-19. However, because of social distancing rules, the changes of listings might differ in distinct neighbourhoods and room types. Proximal gradient loss function optimization, random forest, XGBoost and SVR models will be trained. The most accurate model is chosen according to MSE. Then the best model will be compared before and after the Covid-19. To minimize the seasonal influence and the short-term to long-term pandemic impact, the listing data used are rentals from April to September (2019 and 2020).

2. Initial data treatment[2]

The datasets used in this project are from Inside Airbnb, 2019 (48,602 obs) and 2020 (44,651 obs) October listings. The original datasets contain all the listings with the last review date from 2011 to the date they were released. In order to compare the effect of Covid-19 on listings rented, we chose those listings with review records from April to September in 2019 and 2020 as before and after Covid-19. The variables in the datasets are *id, name, host_id, host_name, neighbourhood_group, neighbourhood, latitude, longitude, room_type, price, minimum_night, number_of_reviews, last_review (date), reviews_per_month, calculated_host_listings_count, availability_365*. The missing rate is pretty low here (less than 1%), thus missing value was deleted. In order to avoid overfitting, outliers (above 95th quantile) of prices were excluded and some models like random forest and XGBoost will be trained. When checking the correlation matrix, number_of_reviews and reviews_per_month were highly correlated. Then reviews_per_month was removed. The number of observations left is 16,617 and 7,400 respectively before and after Covid-19.

---

[1] LANE, L. *How Bad Are Covid-19 Pandemic On Airbnb Guests, Hosts?.* Forbes (Link)
[2] Code and process detailed in Data Mining and Cleaning.ipynb

Another step in the data treatment process was to encode some of the categorical variables. One hot encoding was conducted. The review_month was generated from the last_review month and was encoded as well (last_review was removed). At last, training and test datasets were split randomly by the rate of 8:2.

3. Airbnb listings on the map

In order to make a clear visualization of the Airbnb listings on NYC map, only those being reviewed in September were chosen. There were 12,086 listings in Sep 2019, however, there were only 5,721 listings being reviewed in Sep 2020. As shown in the two maps. The listings with the price under 200 dollars, 200 - 500 dollars, 500 - 1,000 dollars, and above 1,000 dollars are represented by grey, steelblue, blue and red respectively. It's obvious that rental in Sep 2020 decreased dramatically compared with that in 2019, especially the proportion of lower price listings (under 500 dollars). One interesting phenomenon is that in Long Island, the number of rentals with 500 - 1,000 dollars price increased after the lock down. Because the social distancing makes non-crowded places more preferable than those in commercial areas.

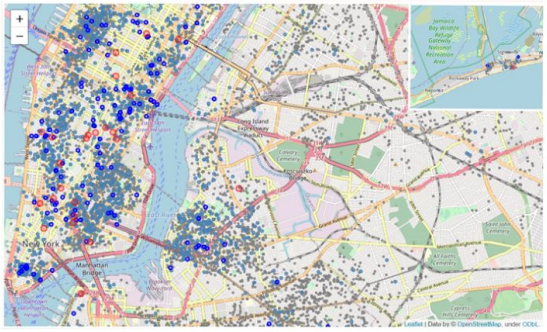

Figure 1. Listings in Sep. 2019.

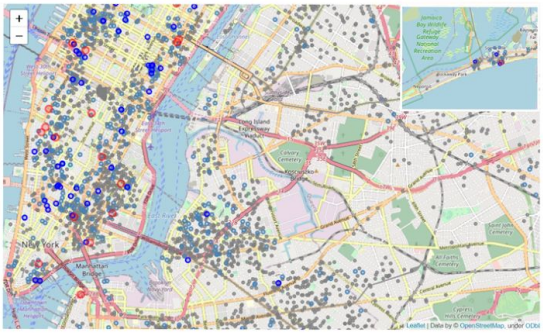

Figure 2. Listings in Sep. 2020.

4. Exploratory data analysis

Using the NLP analysis, we analyzed the name of the Airbnb listings being reviewed from April to September 2019 and 2020. The high frequency words in the listing names are shown on the word cloud. "Brooklyn" appeared more than "Manhattan", "Spacious", "private" and "clean" etc. appeared more than the names of popular tourist spots compared with that of 2019.

Figure 3. Frequency Words in the Listing Names 2019

Figure 4. Frequency Words in the Listing Names 2020

The total number of listings rented after April to September in 2020 in the five neighbourhood group dropped dramatically. However, the percentage of rentals in different neighbourhoods did not change much, except that in Brooklyn and Manhattan. Before Covid-19, the percentages of the listings rented in Brooklyn and Manhattan were both 41.5%. After Covid-19 lockdown, the listings rented in Manhattan were 2.9% less than that in Brooklyn. The proportions of rentings in Queens, Staten Island and Bronx all increased. This finding supports our hypothesis that people might avoid city centers in which the social distance is harder.

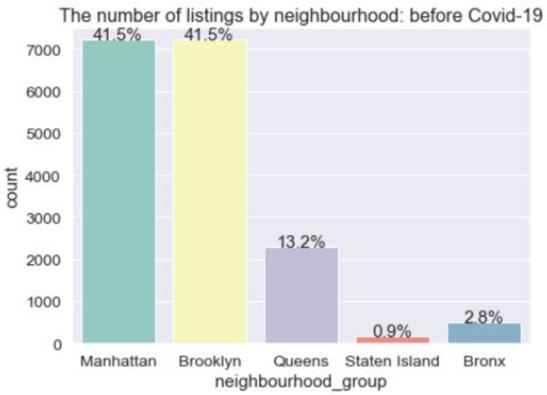

Figure 5. Neighborhood Listings before Covid-19

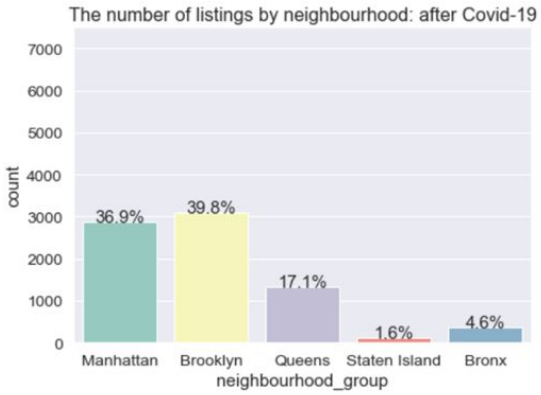

Figure 6. Neighborhood Listings after Covid-19

Nevertheless the density and distribution of prices (most of the listings prices were lower than $600) changed a little before and after the lockdown (outliers were removed). As the violin plots below show, the density and distribution of prices for listings in Brooklyn did not change much. However, for listings in Manhattan, the median price moved from 150 to around 103. The frequency of lower prices increased after lockdown. Nevertheless, the distribution of listing prices in other neighbourhoods did not change much. One interesting finding is that the average price for listings in

Manhattan dropped from 150 to 112 after Covid-19. The average price for Brooklyn did not change. However, the average price in Queens, Staten Island and Bronx increased by 2.0, 0.5, 4.5 dollars respectively.

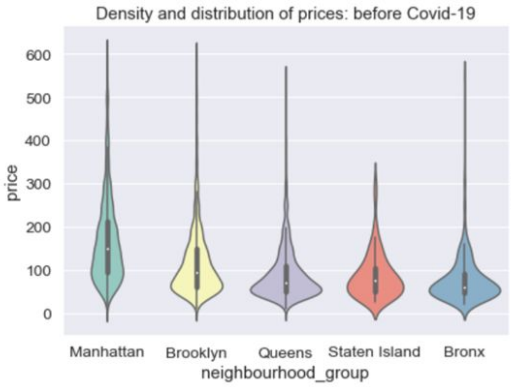

Figure 7. Price distribution before Covid-19

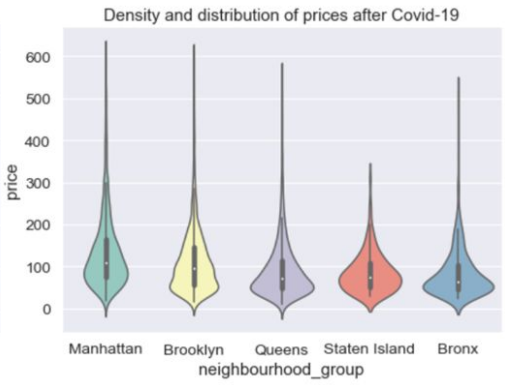

Figure 8. Price distribution after Covid-19

However, after analyzing the prices per room type (entire home, private room, shared rooms, and hotel rooms) as shown in Figures 5 and 6, prices for private rooms increased overall from April 2020 to September 2020 compared to April 2019 to September 2019. But there was not a significant difference in price change for the other room types. The reason for this trend is because of the increased demand for renting private rooms since more people have been practicing social distancing since the beginning of Covid-19.

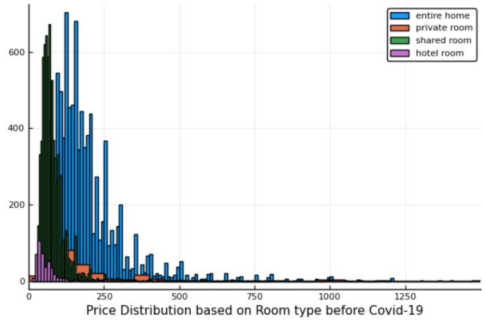

Figure 9: Price distribution per room type before Covid

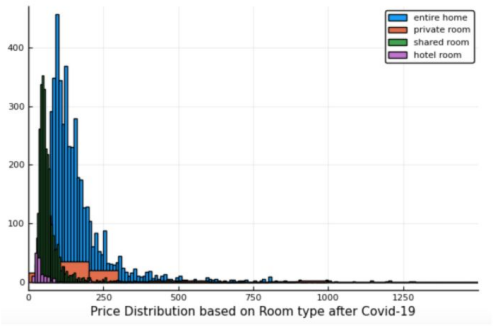

Figure 10: Price distribution per room type after Covid

5. Methodology and steps

After the initial analysis and variable exploration, a three-step-process was followed. Firstly, the dataset was randomly split into training and testing sets. Secondly, three different prediction models were developed separately with before and after Covid-19 datasets. Finally, the best model would be chosen by the accuracy rate, and a comparison would be made for the best model before and after Covid-19.

	Data processing	Data analysis	Mapping	Development	Model selection
Input	Raw data from Airbnb dataset (data/initial data)	Files with cleaned and encoded data	Files with cleaned and encoded data	Training and testing datasets obtained from cleaned datasets	Best model from previous step is selected
Process	Clean dataset and perform feature transformations	Study data and provide insights regarding features and pricing distributions	Plot listings in NY map comparing different areas and time periods	Build three parallel models with different procedures	Adjust and change parameters if needed to improve final model
Output	Files with encoded data (data/treated data)	Visualizations, graphs and plots added to final report	Visualizations, graphs and plots added to final report	Performance of the three models for comparison	Final model for drawing conclusions
File	1. Data Mining and Cleaning	2. EDA	3. Airbnb data visualization	5.1-3 Model Development	6. Final model

Table 1

Additionally, some measures were taken to prevent overfitting: excluding outliers from the training set, adding regularization parameters and removing some redundant and overly specific variables. Exclusion of outliers is justified in the plots below where it can be seen that removing entries with prices higher than the 95th quantile ($350 for 2019, $300 for 2020) reduces the magnitude of errors and greatly increases the generalizability of our dataset. The addition of regularization parameters is done in the training process of the model to ensure that the final selection does not overfit the training set. This penalizes large weights when calculating the optimal model. Variable removal for simplification purposes includes the elimination of host_id, neighbourhood (but not neighbourhood_group) and listing_id.

Figure 11. Quantile distribution of prices

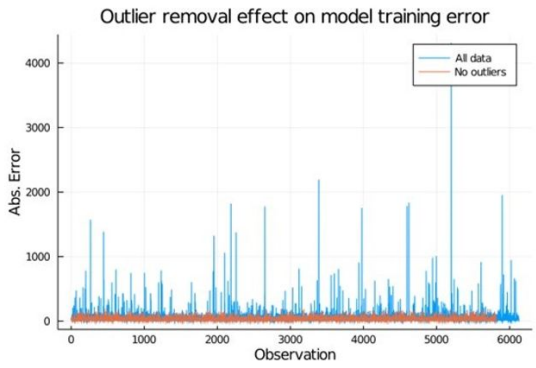

Figure 12. Model improvements with data treatment

6. Model demonstration

6.1 Minimizing loss functions with proximal gradient method

The proximal operator $R^d \rightarrow R^d$ for a function $r : R \rightarrow R^d$ is defined as:

$$\text{prox}_r(z) = \text{argmin}_w(r(w) + \frac{1}{2}||w - z||_2^2)$$

This operator will be used in an algorithm to minimize the objective function where l(w) and r(w) are loss and regularizer functions.

$$l(\text{w}) + r(\text{w})$$

Given a loss function, a regularizer, and stepsizes (α) the following steps are done for *maxiters* iterations:

1. The subgradient (*g)* of the loss function is calculated
2. w is updated in the following manner:

$$\text{w} \leftarrow \text{prox}_{\alpha r}(w - \alpha g)$$

This approach is taken to build three different models with quadratic, huber and l1 as target loss functions. Their performances are calculated for different regularizations functions and varying parameter values. The top performing models for each target function are displayed in the results table below.

6.2 Random Forest

Random forest is an ensemble learning method that creates multiple decision trees and merges those decision trees together to give a prediction. Being able to combine multiple decision trees reduces variances and will thus produce more accurate results. To prevent overfitting, random forest also uses a bagging technique, by randomly selecting a subset of the features at each split. The objective function of random forest is:

$$\hat{y} = \frac{1}{K}\sum_{k=1}^{K}(f_k(x_i)), f_k \in F$$

We decided to fit our random forest model with 3000 trees, because the more trees, the more unlikely it is for our model to overfit. As for the number of features selected, we decided to run the random forest model with 10 features selected and with 150 features selected at each split. The reasoning behind this is because random forest had the lowest training MSE on the pre-covid Airbnb dataset when we randomly selected 150 features. For the post-covid Airbnb dataset, random forest had the lowest training MSE when we randomly selected 10 features. We can visualize the result in figures 13 and 14 below. Additionally, there is a tradeoff between selecting less features and selecting more features. The lower the number of features selected, the higher the decorrelation effect. On the other hand, selecting more features increases the strength of the individual decision trees in the forest.

	5	10	25	50	100	150	200	225
MSE	273.6228	272.4503	272.9807	272.7047	272.7972	273.8041	272.7733	273.8379

Figure 13: MSE with 3000 decision trees on pre-covid dataset per number of random features selected

	5	10	25	50	100	150	200	225
MSE	314.4819	314.3826	314.0159	314.4656	314.7267	313.9879	314.4071	314.8065

Figure 14: MSE with 3000 decision trees on post-covid dataset per number of random features selected

6.3 XGBoost

XGBoost is a decision-tree-based ensemble machine learning algorithm which uses a gradient boosting framework. It optimized the gradient boosting algorithm through parallel processing, tree-pruning, and regularization to avoid overfitting. XGBoost builts one tree at a time which is different from random forests building trees independently. The Objective function is:

$$\mathcal{L}^t = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i^{(t-1)} + f_t(X_i)) + \Omega(f_t)$$

6.4 Support Vector Regression (SVR)

SVR uses the same algorithm as SVM but on regression tasks. It constructs a hyperplane in a high dimensional space. The Objective function is:

$$\frac{1}{2}||w||^2 + C\sum_{i=1}^{N}(\xi + \xi^*)$$

7. Model training and selection

The results of all the models were listed in Table 1. Among the predictive models trained on the dataset *Before Covid-19,* XGBoost had the lowest test MSE, 2264.07, with the highest test R-square 0.57. However, when training all the same models on the dataset *After Covid-19*, the Random Forest model with 3000 trees and 150 features outperformed. The test MSE is 1952.01, and R-square is 0.48. Therefore, the Covid-19 had some effects on the predictive models on the listing prices.

	Before Covid-19			**After Covid-19**		
Model	**Training MSE**	**Test MSE**	**Test R^2**	**Training MSE**	**Test MSE**	**Test R^2**
Proxgrad (huber loss)	4925.40	28498.89	-0.07	3574.41	23894.62	-0.05
Proxgrad (quadratic loss)	4568.82	27184.58	-0.01	3439.75	23186.54	-0.02
Proxgrad (l1 loss)	4922.83	28493.25	-0.06	3579.69	23906.77	-0.05
Random Forest 3000 trees, 10 features	272.45	2350.93	0.56	272.45	1953.51	0.48

Random Forest 3000 trees, 150 features	273.80	2354.61	0.55	273.80	1952.01	0.48
XGBoost	1534.12	2264.07	0.57	1069.27	1954.89	0.48
SVR	5231.16	5527.91	-0.05	3728.17	3962.89	-0.05

Table 2

8. Weapons of math destruction

Our project is unlikely to produce a weapon of math destruction. First off, the outcomes of our models are easy to measure. The outcomes will be the estimated Airbnb price before and after the Covid-19 pandemic based on previous Airbnb listings. Since the intention of our model was to only gain insight on the Airbnb price based on the property information, it is also hard to foresee any negative consequences that our predictions. Lastly, there does not seem to be a self-fulfilling feedback loop. Our price predictions are for informational purposes and the features used to predict the price such as reviews, room type, and neighborhood would not be impacted by a high or low price of an Airbnb.

9. Fairness

Fairness is not an important criterion to consider while selecting a model for our application. While the dataset does contain information about the reviews for each of the Airbnbs, the algorithms that were used to run the model did not bias the positive versus negative reviews. Additionally, the dataset we used for this study contained no information about the property owners such as race, gender, etc., or anything that would potentially pose ethnic issues. The other features such as neighborhood and room type would by nature, differentiate the price of an Airbnb. For example, the cost of a shared room in uptown Manhattan would be significantly cheaper than the cost of a house in the financial district of Manhattan.

10. Conclusions and insights regarding the Airbnb market

After Covid-19, the average price for listings in Manhattan dropped from $150 to $112 and that for Brooklyn did not change. However, the average price in Queens, Staten Island and Bronx increased by 2.0, 0.5, 4.5 dollars respectively. One possible reason is that the social distancing made non-crowded places more preferable than those in center areas. The frequency names of those listings being rented recently also proved this point. Manhattan and the names of tourist attractions appeared less than before. Proximal gradient regression with Huber, quadratic, and L1 loss functions, Random Forest, XGBoost and SVR models were tested on the dataset before and after Covid-19 to predict the listing prices. Before covid-19, XGBoost is the best model with the lowest test MSE. However after Covid-19, the best model is the Random Forest with 300 trees and 150 features. Therefore, Covid-19 did influence the Airbnb listing prices.