Regression: New York Airbnb Price Prediction

Zixin Huang

July 8, 2021

1 Objective and Data

In this project we use regressions to predict the Airbnb prices in New York city. The main objective of this project is to analyze the key attributes that determine the price of a listing. Therefore, we mainly focus on the interpretation aspect when choose the models.
The dataset[1] contains the information of 48895 Airbnb listings in New York in 2019. Attributes of listings include price, neighbourhood, room types, and host names. The names of all columns and their detailed explanations is shown in Table 1. There are 10052 null values in *last_review* column ans *reviews_per_month* columns respectively, 16 null values in *name*, and 21 null values in *host_name*.

Table 1: Columns and Explanations

Column	Explanation
id	listing ID
name	name of the listing
host_id	host ID
host_name	host name
neighbourhood_group	neighbourhood group name
neighbourhood	neighbourhood name
latitude	latitude of the listing
longitude	longitude of the listing
room_type	listing room type
price	price of the listing per night
minimum_nights	minimum nights for stay
number_of_reviews	total number of reviews
last_review	the date of the last review
reviews_per_month	number of reviews per month
calculated_host_listings_count	number of listings for the same host
availability_365	number of days a listing available for booking

2 Exploratory Analysis and Data Cleaning

We first replace all null values in *last_reviews* by 0, and replace all other null values by "None". The statistical summary of all numerical columns is shown in figure 1. The summary statistics indicates that most of the numerical features have an uneven distribution and possibly contain

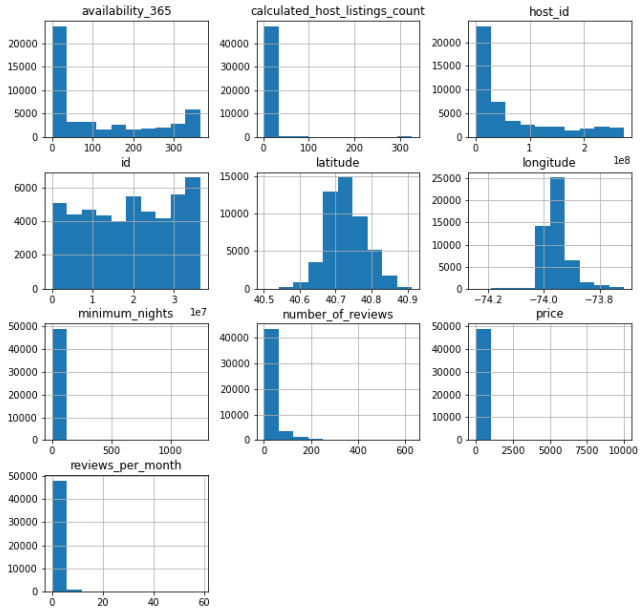
[1] https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

outliers. For example, 75% of the prices are lower than $175, but the maximum price is $10000. The standard deviations of price, minimum nights, number of reviews, and availability are very high. This can be seen more clear with histograms in Figure 2.

Figure 1: Statistical Summary of Numerical Columns

	count	mean	std	min	25%	50%	75%	max
id	48895.0	1.901714e+07	1.098311e+07	2539.00000	9.471945e+06	1.967728e+07	2.915218e+07	3.648724e+07
host_id	48895.0	6.762001e+07	7.861097e+07	2438.00000	7.822033e+06	3.079382e+07	1.074344e+08	2.743213e+08
latitude	48895.0	4.072895e+01	5.453008e-02	40.49979	4.069010e+01	4.072307e+01	4.076311e+01	4.091306e+01
longitude	48895.0	-7.395217e+01	4.615674e-02	-74.24442	-7.398307e+01	-7.395568e+01	-7.393627e+01	-7.371299e+01
price	48895.0	1.527207e+02	2.401542e+02	0.00000	6.900000e+01	1.060000e+02	1.750000e+02	1.000000e+04
minimum_nights	48895.0	7.029962e+00	2.051055e+01	1.00000	1.000000e+00	3.000000e+00	5.000000e+00	1.250000e+03
number_of_reviews	48895.0	2.327447e+01	4.455058e+01	0.00000	1.000000e+00	5.000000e+00	2.400000e+01	6.290000e+02
reviews_per_month	48895.0	1.090910e+00	1.597283e+00	0.00000	4.000000e-02	3.700000e-01	1.580000e+00	5.850000e+01
calculated_host_listings_count	48895.0	7.143982e+00	3.295252e+01	1.00000	1.000000e+00	1.000000e+00	2.000000e+00	3.270000e+02
availability_365	48895.0	1.127813e+02	1.316223e+02	0.00000	0.000000e+00	4.500000e+01	2.270000e+02	3.650000e+02

Figure 2: Distribution of Numerical Columns



2.1 Price

From Figure 2 we can see that the distribution of price, which is our target variable, is right skewed. To transform it to a normal-like distribution, we first perform log transformation to the

prices. Then we use boxplot to identify outliers. There are 48 observations exceed the lower bound and 589 observations exceed the upper bound. Next we remove the 48 lowest prices and the 589 highest prices. The distribution of price after transformation is shown in Figure 4, it is a more normal-like distribution with mean equals to 4.71 and standard deviation equals to 0.638.

Figure 3: Boxplot of the Price

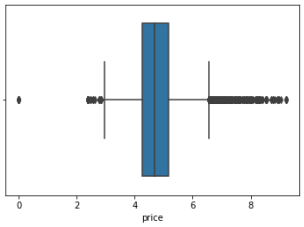
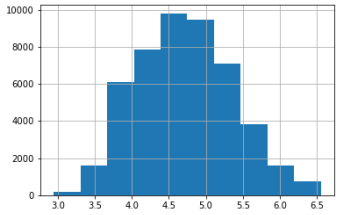


Figure 4: Distribution of the Price After Transformation



2.2 Neighbourhoods

There are 5 neighbourhood groups in this dataset, which are Manhattan, Brooklyn, Queens, Bronx, and Staten Island. As shown in Figure 5, most of listings are in Manhattan and Brooklyn, and only around 14% of the listings are in other neighbourhood groups. Figure 6 shows the relationship between price and neighbourhood groups. Listings in Manhattan have the highest average log price, and listings in Brooklyn have the second highest price. Average log price of listings in other neighbourhood groups are not very different.

Figure 5: Distribution of Neighbourhood Groups

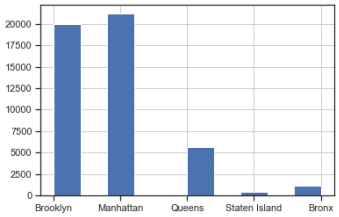
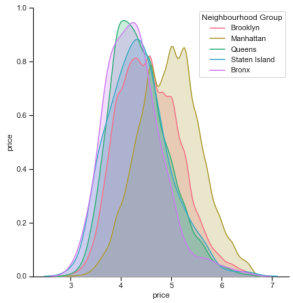


Figure 6: Relationship between Neighbourhood Groups and Log Price

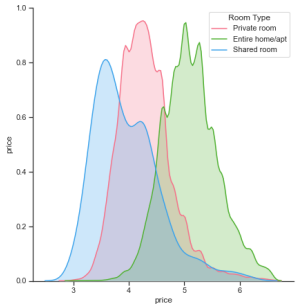


The total number of neighbourhood in this dataset is 220, with about 71% of the listings in the top 23 neighbourhoods. Therefore, we map those less represented neighbourhoods, which have listings less than 500, into "Others".

2.3 Room Type

The relationship between room type and price can be seen in Figure 7. Listings with room type as entire room or apartment have the highest average log price, whereas listings with shared rooms have the lowest average log price.

Figure 7: Relationship between Room Type and Log Price



2.4 Minimum Nights and Reviews

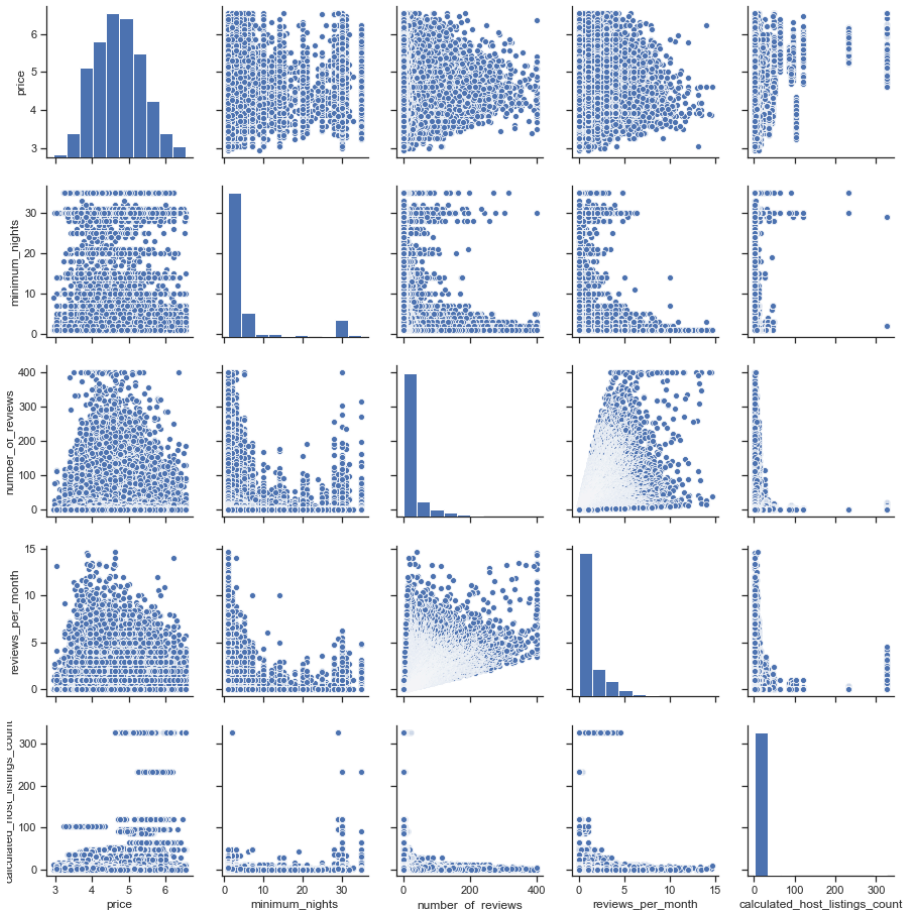
The summary statistics indicates that the average and 75% quantile of minimum nights is 6.97 and 5, but the maximum value of minimum nights is 1250. Since there are only 495 listings that have minimum nights requirement longer than 35 days, we replace the minimum nights of those listings with 35.
For the same reason, we replace the number of reviews of listings that have reviews more than 400 with 400, and replace reviews per month of listings that have reviews per month more than 15 with 15.

2.5 Other Features

We drop listing ID, name, host ID, host name, last review date, availability days, longitude, and latitude since we believe they are not helpful in predicting the listing price. Therefore, the dataset now contains 48258 observations and 8 columns, with 3 categorical features and 4 numerical features. The pair plot of all numerical features and price can be seen in Figure 8. We do not observe a clear relationship between minimum nights and price, though they may be positively associated.

Number of reviews and reviews per month seem negatively associated with price, but their relationships are not clear as well. The calculated host listing number appears to be positively associated with price.

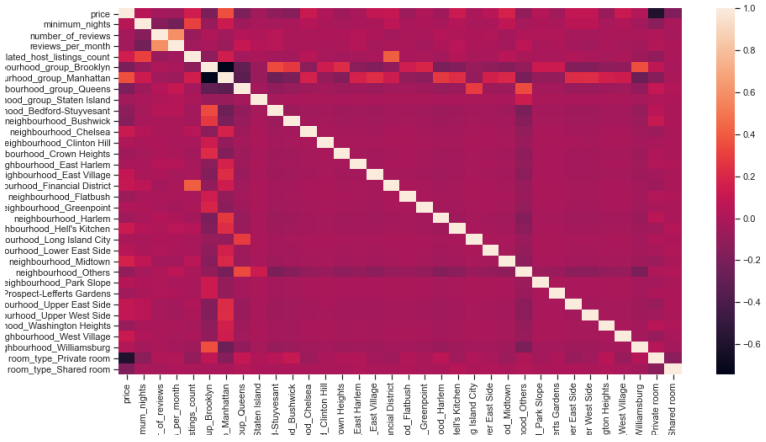
Figure 8: Pair Plot of Numerical Features and Price



3 Feature Engineering

We then one-hot encode the categorical features in the dataset. For the purpose of interpretation we drop the first category for every categorical feature. This results in a total number of 34 columns in our dataset. The heatmap of all the columns is shown in Figure 9. From the heatmap we can see that features that mostly correlated with price are whether the listing is a private room or entire apartment and whether the listing is in Manhattan.

Figure 9: Heatmap of All Columns



4 Regression

To avoid the impact of the magnitude of features, we use scale the features by subtracting their means then dividing by their standard deviation. Next we split our dataset randomly so that 70% of the data is in the training set and 30% of the data is in the test set. Then in order to choose a model, we randomly separate our training set into 5 folds to perform cross validation. We first use a simple linear regression as base line regression

$$Y = \boldsymbol{\beta}\boldsymbol{X} + \varepsilon$$

where Y is the log price of listings and X is the matrix of features.
Next we add polynomial features into our regression. We set the polynomial degree equals to 2 and let the algorithm to find the best polynomial feature for the model.
Finally, we add regularization term to our model. Since we focus on the interpretation of the model, we use Lasso regression which minimize the objective

$$\sum_i (y_i - \sum_j \beta_j x_{ij})^2 + \lambda \sum_j |\beta_j|$$

We use cross validation to tune hyper-parameter λ[2], and record the R^2 score and the mean square error of each regression to compare their performance.

5 Results

The R^2 score and mean square error of linear regression, polynomial regression and Lasso regression with tuned λ can be seen in Table 2. The table shows that the Lasso regression with $\lambda = 0.0001$ has the highest performance since it has the highest R^2 score and the lowest mean square error (MSE). Polynomial regression is better than simple linear regression, which has the lowest R^2 score and highest MSE. Therefore, we use the Lasso regression with $\lambda = 2$ in our whole training set, then use the fitted model to make prediction on the test set.

[2]Due to limited computing power, we only try λ equals to 0.0001, 0.001, 0.01, and 0.1.

Table 2: Performance of Models

Regression	R^2 Score	MSE
Linear Regression	0.5544	0.1814
Polynomial Regression	0.5689	0.1755
Lasso Regression ($\lambda = 0.0001$)	0.5694	0.1753

The test result using Lasso regression is shown in Table 3. This suggests that our model has over 82% accuracy and about 57% interpretability.

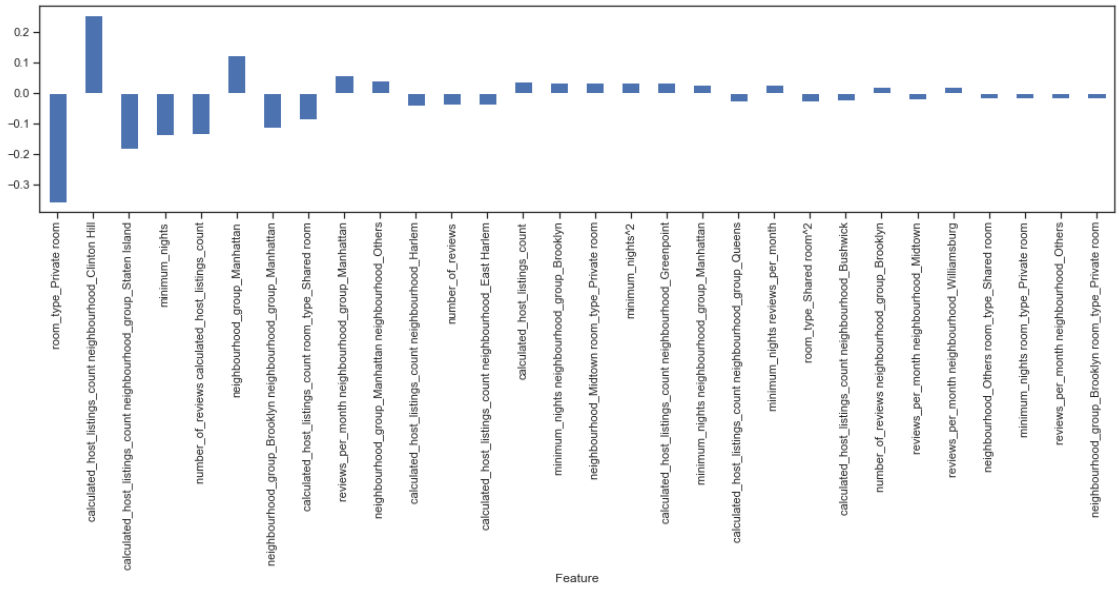
Table 3: Test Result using Lasso Regression ($\lambda = 2$)

R^2 Score	MSE
0.5731	0.1749

6 Insights

The results indicate that our model can only explain about 57% of the price of listings in the New York City. To gain a deeper insight of attributes that have the greatest influence on price, we plot the top 30 features having largest absolute coefficient in Figure 10.

Figure 10: Top 30 Features with Largest Absolute Coefficient



The plot shows that our model determine the price of listings primarily use the first 6 features in Figure 10. These key features are whether the room is private room instead of entire apartment, calculated host listing number combined with whether the neighbourhood is Clinton Hill or Staten Island, minimum nights stay, number of reviews combined with calculated host listing number, and whether the listing is in Manhattan.

In general, if the listing is in Manhattan, the price of this listing would be higher than listings in other area. For listing belongs to hosts who have many other listings, its prices would be higher

if it is in Clinton Hill, but lower if it is in Staten Island or has many guests before. In addition, number of minimum night stay has a negative impact on price.

7 Next Steps

To further improve our model we may need to try a wider range of λ for the Lasso regression. We may also try other machine learning models or revisit linear regression model but improving feature engineering by adding some key polynomial features, such as calculated host listing count times number of reviews and calculated host listing count times neighbourhood Staten Island.