New York City Airbnb Rental Price Prediction

Karl Wang
University of California, San Diego
kawang@ucsd.edu

Abstract

This project explores how location can help predict Airbnb listing price in NYC, one of the most expensive cities in the world. In this project, I compare 2 kinds of models to incorporate locality to predict rental price: K-Nearest Neighbour and Linear Regression.

1 Introduction

New York City (NYC) is one of the largest metropolitan in the world. With such a high population density, the cost of living is expensive. Airbnb is a vacation rental marketplace that allows customers to rent a house, a room, or a shared room. With iconic tourist attractions such as Statue of Liberty, Empire State building, Times Square and so on, Airbnb has a thriving market in the city. I am interested in predicting the Airbnb rental price given the information of the listing. I think this problem is interesting because if I am able to predict the rental price, then as a customer, I would know that if a listing is a good value or not. As a listing owner, I would be able to set a more appropriate price tag.

2 Exploratory Data Analysis

To explore this prediction problem, I acquired a dataset from Kaggle. It contains information about NYC 2019 Airbnb listings. The dataset has 48,895 number of listings. For each listing, the following information is given.

- *id*: the id of the listing
- *name*: the title name of the listing
- *host id*: the id of the host
- *host name*: the name of the host
- *neighbourhood group*: the neighbourhood group the listing belongs to.
- *neighbourhood*: the neighbourhood the listing belongs to.
- *latitude*: the latitude location of the listing.
- *longitude*: the longitude location of the listing.
- *room type*: the rental type of the listing.
- *price*: the rental price per night in dollars.
- *minimum nights*: the minimum nights the customers have to purchase.
- *number of reviews*: total number of reviews of the listing.
- *reviews per month*: number of reviews per month.
- *calculated host listings counts*: number of listings that the host has.
- *availability 365*: number of days the listing is available in 365 days.

In this section, I am only going to explore and process the features relevant to the predictive task.

2.1 Room Type

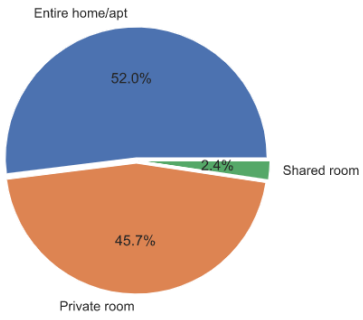

Figure 1: Percentage of different room types

Airbnb allows list owner to rent out the entire home/apartment, a private room, or a shared room. From common sense, one should expect the rental price of a entire house to be more expensive than that of a private room. From **Figure 1**, we see that most listings rent out either a private room or the entire

home/apartment. To make the prediction simpler, I am going to drop the 2.4% of shared room listings, so that the listings are either a private room or a entire house. After dropping shared room listings, private room has 47% and entire home has 53% of data.

2.2 Neighborhood Group

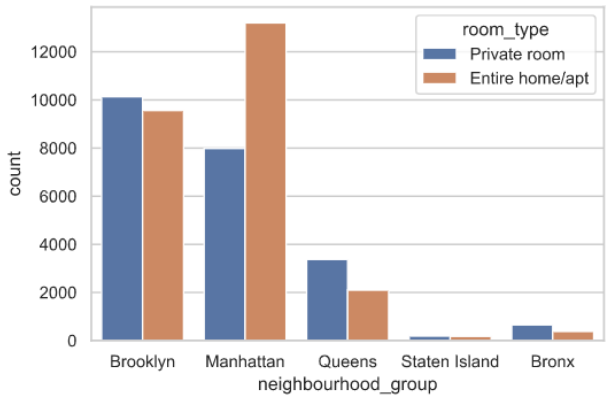

Figure 2: Counting of listings by neighborhood group

NYC is consists of 5 neighborhood groups: Brooklyn, Manhattan, Queens, Staten Island, and Bronx. With Manhattan being the downtown of NYC and having the most popular tourist attractions, it is unsurprising that most of the listings are located in Manhattan. However, it is to interesting note that Brooklyn has more private room listings than Manhattan. Since Staten Island only has 364 listings (out of 47,735), it is going to be difficult to make general predictions from such small number of samples. Hence, I am going to drop listings from Staten Island.

2.3 Rental Price

Statistics

From the dataset, the minimum of listings' price is 0, and the maximum of listings' price is 10,000. These two prices are clearly extreme outliers. To get rid of the outliers, I decide to take the listings with rental price between 5 percentile to 95 percentile from each neighborhood group. After getting rid of the outliers, here are the statistics for each neighborhood.

The sample mean is calculated using $\bar{P} = \frac{1}{n}\sum_{i=1}^{n} P_i$, where n is the count of the listing from the neighborhood group, P_i is the price of the listing. The sample standard deviation is calculated by taking the square root of unbiased variance estimator, which is $\sigma_P = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(P_i - \bar{P})^2}$.

From the table above, we see that listings from Manhattan has the highest sample median and sample mean among all neighborhood groups. This is to be expected

	Brooklyn	Manhattan	Queens	Bronx
count	17878.0	19116.0	4923.0	938.0
mean	107.08	166.54	87.15	76.88
std	54.94	85.03	41.79	35.81
min	40.0	55.0	37.0	34.0
25%	65.0	100.0	55.0	50.0
50%	93.0	150.0	75.0	69.0
75%	140.0	206.0	105.0	98.0
max	289.0	450.0	225.0	200.0

as Manhattan is the most popular tourist places in NYC. On the other hand, listings from Manhattan also has the highest sample standard deviation.

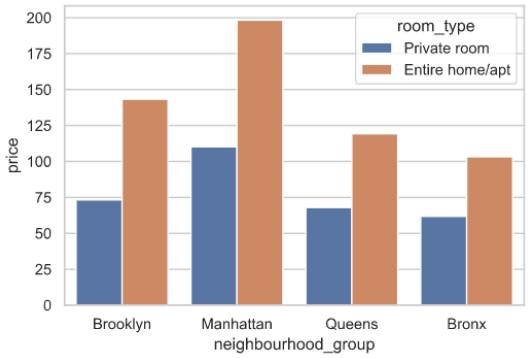

Figure 3: Mean price by neighborhood and room type

Figure 3 shows that rental prices from different room type and different neighborhood group are quite distinctive. Although it is expected that the price of a private room is cheaper than the price a entire house, it is interesting to observe that the average price of a private room in Manhattan is more expensive than the average price of a entire house in Bronx. This suggests that both neighborhood group and room type are important features that distinguish listing price.

Gaussian Approximation

The first figure of the **Figure 4** is the density of the listing price of entire home/apt. With a mean of 170 and a median of 150, the distribution is positively skewed. To rectify the skew, I take the natural log of the listing price, and the result is shown in the second figure of **Figure 4**. The distribution of the log price is approximately normal. The red curve is a Gaussian distribution with $\mu_N = 5.06$ and $\sigma_N = 0.4$. Using the Gaussian distribution, the 95% confidence interval is between (4.28, 5.85). Transforming the log interval back to standard price, the log interval corresponds to (71.9, 345.6). After checking the actual percentage of listings between the interval, the percentage comes out to be 92.7%. Therefore, the Gaussian model is quite accurate.

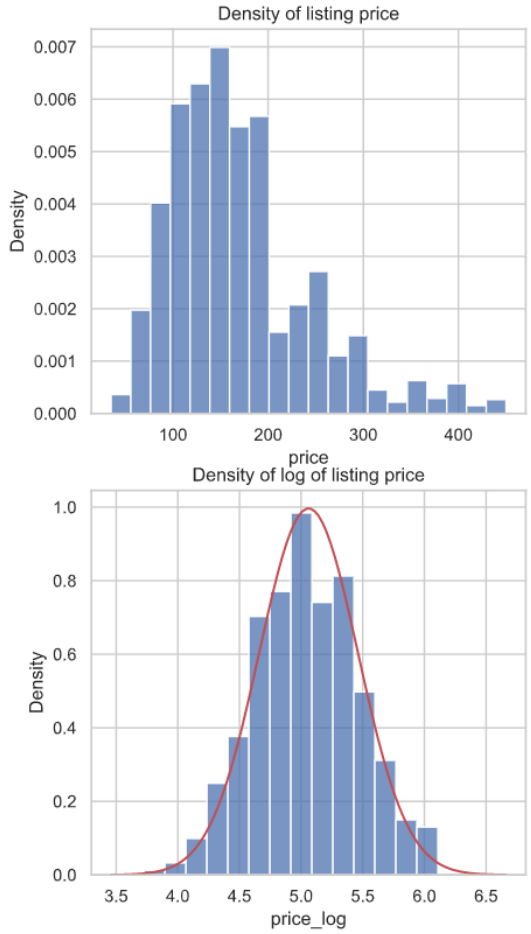

Figure 4: Fitting a Gaussian model to the listing price

2.4 Longitude and Latitude

Longitude and latitude plays a key role in the rental price of the listing. One should expect listings located by popular tourist attractions to be more expensive than listings that are not.

In **Figure 5**, the center rectangle with no data points is the central park. From the figure, we can see that the listings to the southwest of central park is generally more expensive than listings to the northeast. This is mainly because southwest of Manhattan is the downtown, and has popular attractions such as Empire State building, Times Square, Broadway and etc.

Longitude and Latitude can be useful in the prediction of rental price. For a new listing X, the rental price should be similar to the listings near X longitude and latitude.

3 Methodology

In this section, I will describe the general procedure to model the predictive task and test the correctness of the model.

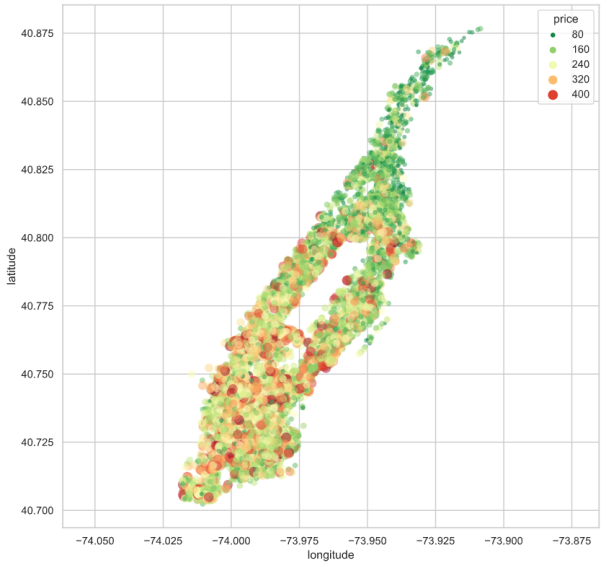

Figure 5: Manhattan listings price for entire house

3.1 Data Preparation

In order to test the robustness of the prediction model, I split 60% of data into training set, 20% of data into validation set, and 20% of data into testing set. The validation set is mainly use for tuning parameters of the model and testing set is for the final evaluation of the model.

3.2 Features

From the last section, I concluded that locality of the listing is a major factor to the listing price. Therefore I decide to use longitude and latitude as the features. In addition to longitude and latitude, the listing's neighborhood group and room type are also key features to distinguish the type of the listing.

3.3 Proposed Model

For a new listing with longitude and latitude X, we want to predict the the listing's rental price from the neighbors according to X. Therefore, K-Nearest Neighbor (KNN) is the suited model for this predictive task. More specifically, I use a KNN regressor to predict as the prediction target, rental price, is a continuous value.

The KNN regressor finds k nearest neighbors of X using Euclidean distance, or the L2 norm. Below is the formula for finding one closest neighbor. X_i is a 2 dimension feature point, (longitude, latitude), from the training set, and X is the target features we are trying to predict.

$$\arg\min_{X_i} \|X_i - X\|_2$$

After finding the k closest points, the KNN regressor interpolates between the k points to output the prediction $\hat{y}$.

3.4 Alternative Model

Beside KNN, I am also going to implement a linear regression model for comparison. The linear regression model is the following, where X is the $m \times n$ data matrix with m as number of samples, n as number of features, θ is the weights for each feature that we are trying to fit. θ is a $n \times 1$ vector. Finally $\hat{y}$ would be the predicted price.

$$\hat{y} = X\theta$$

The optimal θ would come from minimizing the loss function.

$$\theta^* = \arg\min_{\theta} \frac{1}{m}\|y - X\theta\|_2^2 + \lambda\|\theta\|_2^2$$

I decide to use the linear regression with L2 regularization, the ridge regression. With regularization parameter λ, we can prevent the model from overfitting by reducing the variance and increasing the bias.

3.5 Metrics

I use root mean square error (RMSE) as my evaluation metrics. In the following formula for RMSE, y_i is the actual price of the listing i, $\hat{y}_i$ is the predicted price by the model for the listing i, and n is the number of samples.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

The RMSE is essentially the average square deviation square between the predicted price and actual price, taking the square root so that the unit will be in dollars instead of dollars square.

3.6 Algorithm

From section 2.3, I discover that rental price for different room types and different neighborhood groups vary drastically. To incorporate this difference, I separate the data by its neighborhood group and room type. Since there are 4 neighborhood groups and 2 room types, the data is split into 8 parts according to the combination of neighborhood groups and room types.

After that, for each (neighborhood group, room type) in the training set, I trained a regressor using all the samples with the same combination of neighborhood group and room type. For each sample in the validation and testing set, I use the regressor with the corresponding neighborhood group and room type to predict.

3.7 Baseline

To compare the effectiveness of the models, I also implement a simple baseline model that uses the same algorithm as above. However, instead of training a regressor, I am simply using the sample mean of the samples with the same neighborhood group and room type from training data as prediction, $\hat{y} = \bar{y}$. Essentially, only working with features of neighborhood group and room type.

4 Prediction

This section goes through the process of model parameters tuning, the prediction result, and conclusion of the predictive task.

4.1 Proposed Model Parameters

The question arises with the KNN model is that how many neighbor points to consider? Or, what is the optimal k value? To find this out, I run the KNN model with 1 neighbor to 59 neighbors on the validation set.

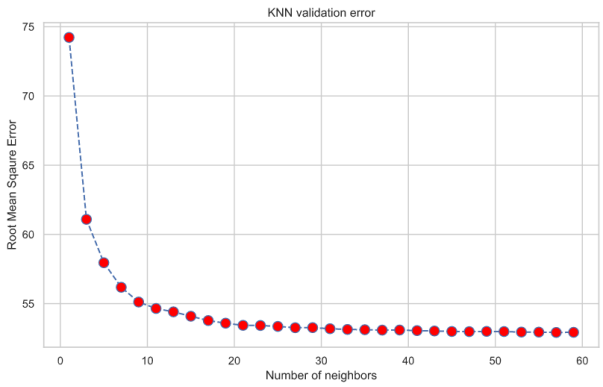

Figure 6: KNN with different number of neighbors

From **Figure 6**, I notice that the improvement from the first few neighbors are drastic. Using 3 neighbors is 17.7% better than using just 1 neighbor in terms of RMSE. However, the improvement beyond 20 neighbors is minimal. Nonetheless, concluding from the KNN validation error, the optimal number of neighbors is 57, with a RMSE of 52.9.

4.2 Alternative Model Parameters

From **Figure 7**, the result shows that ridge regression has the lowest validation error with $\lambda = 0$, which means the loss function has no regularization at all. The RMSE of ridge regression with no regularization is 54.

4.3 Result

After finding the optimal model parameters using validation error, the models are used to predict the test set. From **Figure 8**, we see that KNN has the lowest overall RMSE compare to the other two models. KNN also has the lowest RMSE for all neighborhood groups except for Bronx. Surprisingly, the baseline model, which is simply using the sample mean as the prediction, has the lowest testing error for Bronx. I suspect that this is due to that the sample size of Bronx is too small that the model does

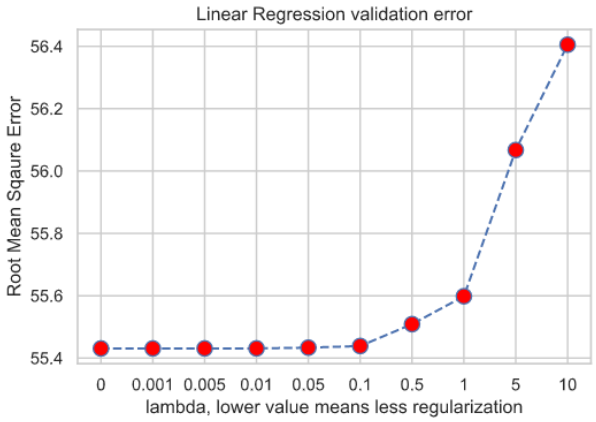

Figure 7: Ridge Regression with various regularization

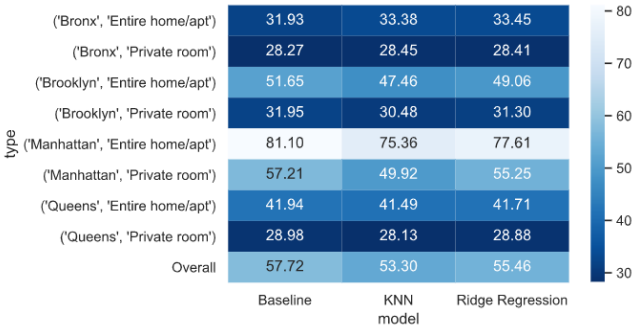

Figure 8: Testing errors of the 3 models

not have enough samples to train on. Compare to other neighborhood groups, Manhattan with entire house has the highest RMSE across all three models. This is to be expected as in section 2.3 we discover that Manhattan neighborhood has the highest sample standard deviation.

4.4 Conclusion

From the result above, I conclude that KNN performed the best compare to linear regression and the baseline model. The testing errors of KNN and linear regression is slightly higher than the corresponding validation errors. This might be due to that the model parameters overfitted the validation set.

I suspect the reason KNN performed better than linear regression is because that linear regression assumes that there is a linear relationship between X, longitude and latitude, and y, the rental price. This assumption might be true for specific case, as demonstrated in **Figure 5**. However, it is hard to say this holds true in all cases.On the other hand, KNN has no assumption on X and y. However, they both performed better than the baseline.

I believe the predictive task can be improved further with better algorithm and more useful features. For example, if there is the date of listing, perhaps we could observe some temporal pattern, such as higher price during holiday season. Maybe the ratings of the listing can also help judge the rental price.

From this project, I conclude that locality of a listing is a key determining factor to help predict the rental price.

References

[New York City Airbnb Open Data] https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data