Airbnb Pricing in New York City

—

Jixi Dai * Zicheng Hu † Ziyu Liu ‡

November 5, 2019

Abstract

Home-stay is a rising alternative of hotels for many people around the world and home-stay is the cheaper option for most of the time. Therefore, the pricing of the home-stay property plays an important role in selecting properties. This project employed big data techniques to establish models that examines what price a rational property owner would probably want to rent to others, where the price is neither too high to scare people off nor too low to dissatisfy him or herself. The main goal of this project is to provide a reasonable price range of the Airbnb room listings in New York City area, as a reference for both the hosts and the tenants. Although the price of a listing is typically set by the host, there is always a need for good price estimation. On the one hand, the hosts need this information for strategic pricing in order to make their listings both attractive to tenants and profitable to themselves. On the other hand, the travelers seeking for home-stays want to know the potential price range of listings near their destination so that they can better plan their expenditure and find good bargains.

*Cornell University, *jd999@cornell.edu*
†Cornell University, *zh343@cornell.eduu*
‡Cornell University, *zl722@cornell.edu*

Contents

1 Explanatory Data Analysis

1.1 Data Description

Our project examines the *New York City Airbnb Open Data* from Kaggle.com. This dataset provides detailed features of the listed properties on Airbnb, such as listing price, neighborhood, room type, availability and minimum number of nights required for booking. It also includes additional information such as property descriptions and reviews from previous bookings. With this dataset, we would like to examine how various features influence the listing price of a property. We will try to capture such relationships by building different models (linear model, polynomial model, etc) after some appropriate data preprocessing.

1.2 Data Visualization

1.2.1 Variable Correlations

For numerical features, we plotted a correlation plot to visualize which aspects of the listed properties are relatively important in determining the price (unit: dollars/night).

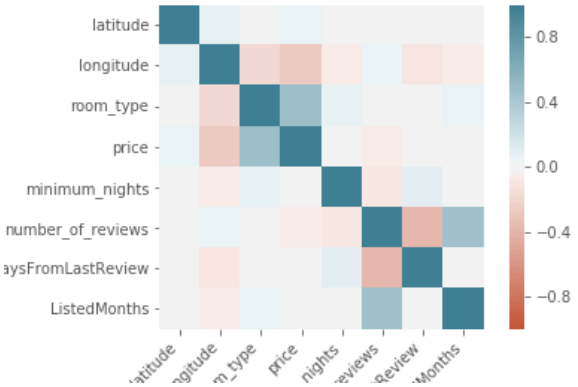

Figure 1: Correlation Plot

As we can see in Figure 1, price is highly correlated with latitude and room_type (property room type) and number_of_reviews (number of past reviews). This inference is in line with reality as the rent of properties located in Manhattan is generally more expensive than the rent of properties located in other counties, and an entire room normally costs more than a shared room or a private room. Also, some numerical features seem to have little correlation with the rent price so it's appropriate to use regularization or variable selection while training models.

1.2.2 Airbnb Prices

Figure 2 shows the distribution of listed Airbnb prices. Based on the plot, the listing price distribution is right-skewed and we fixed it by taking the log transformation (we will discuss the reasons for such transformation later in section 1.3.6).

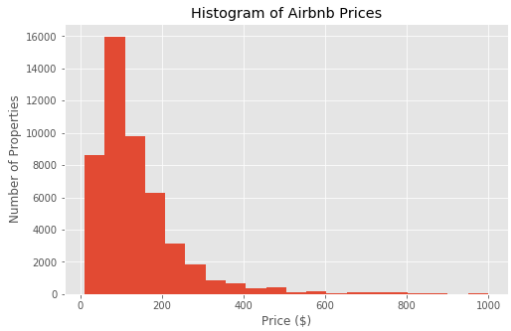

Figure 2: Airbnb Prices Histogram

As the rent price is correlated with latitude, we plotted the distributions of the rent prices in the five neighborhood groups around New York City (Manhattan, Bronx, Queens, Brooklyn, Staten Island). From Figure 3, the average price in Manhattan is the highest while the average price in Bronx is the lowest. Moreover, Manhattan also has the highest price variability compared to the other four neighborhood groups. These observations indicate that features containing property geographic information could be useful for predicting the listing prices.

Figure 3: Aribnb Prices in Different Counties

1.3 Data Preprocessing

1.3.1 Outliers

The conventional definition of an outlier is a data point that falls more than 1.5 times the interquartile range below the first quartile or above the third quartile. In this case, however, we do not stick with this definition due to the nature of our data. Instead, we only considered the price interval [10, 1000] based on common sense. (Prices below $10 are unlikely and barely profitable, and luxurious listings higher than $1000 are rare.) As a result, we dropped 250 outliers out of 48895 observations (0.5% of the data).

1.3.2 Ordinal Values

To facilitate our regression modeling, we transformed the "room_type" column into ordinal values. The average prices for the 3 room types are respectively $66 for shared rooms, $85 for private rooms, and $193 for entire homes/apts. This demonstrates an ordinal relationship among room types so we labeled 1, 2, and 3 for each type in ascending price order.

1.3.3 Nominal Values

The columns "neighborhood_group" and "neighborhood" contains nominal values. Since "neighborhood" specifies the district within each "neighborhood_group", the two columns contain overlapping information. To avoid correlation-related problems while keeping the geographic information intact, we concatenate these two columns into a single column of nominal values (named as "neighborhoods"). In addition, since there is no apparent ordinal relationship between different neighborhood values, we applied one-hot encoding on the new column to facilitate our model building and predictions.

1.3.4 Additional Features

The column "last_review" contains hidden significance indicating the freshness of a home listing, but its timestamp values are not intuitive enough for interpretation and modeling. Therefore, we reshape this column by calculating the difference between the current date and the original stamps, which is just the number of days since the last review. The new column ("days_since_last_review") gives a simple and reader-friendly integer value to measure how recent the home has been reviewed. Besides, we also replaced "reviews_per_month" with "month_listed" (i.e. "number_of_reviews" / "reviews_per_month"). This new column shows how long a listing has existed, which is more relevant to the price.

1.3.5 NA Values

Due to high data integrity, the NA values only appeared in new columns from 1.3.4, which were inherited from the original "last_review" and "reviews_per_month". For "days_since_last_review", it is hard to determine whether these missing values were caused by misrecording or simply no user reviews (we believe both causes are reasonable and likely), so we replaced these missing values by the overall median. The mean replacement is not suitable in this situation because our data has an apparent right-skewed pattern. For "month_listed", missing values from "reviews_per_month" generally indicate freshly listed properties, so we filled in 0 as replacement.

1.3.6 Response Transformation

According to the histogram of our price range (between $10 and $1000), the response variable is highly right-skewed with the majority of points lying between $10 and $250, and a long tail to the right. This is consistent with our intuition that most people are going for fair-priced listings rather than premium or luxurious listings. In order to reduce the effect of the minority expensive listings and stabilize the variance in our data, we can transform the distribution of prices by taking the logarithm of their values. This technique will help us improve the accuracy of regression models in the later phase.

2 Initial Modeling

2.1 Model Evaluation

We used Root Mean Squared Error (RMSE) between the logarithm of the predicted price and the logarithm of the actual sales price. As we mentioned in the previous section, by taking the logarithm on the prices we can make sure that the errors in predicting high prices and low prices will affect the result equally.

2.2 Linear Model

We first fitted a simple linear model by solving the least squares problem and used it as a baseline for our project. In this case, there is no unique solution to the problem due to some linear dependence between features. Therefore, we need to include a regularization term to ensure the validity of our linear model. Firstly, we added a quadratic regularizer to the following optimization problem to ensure a unique solution.

$$minimize \quad \sum_{i=1}^{n}(y_i - w^\intercal x_i)^2 + \lambda \sum_{i=1}^{n} w^2 \qquad (1)$$

Then we implemented a 5-fold cross validation to find the best λ that minimizes the RMSE. As shown in Figure 4, the optimal λ is 3.9 and the resulting RMSE is 0.45.

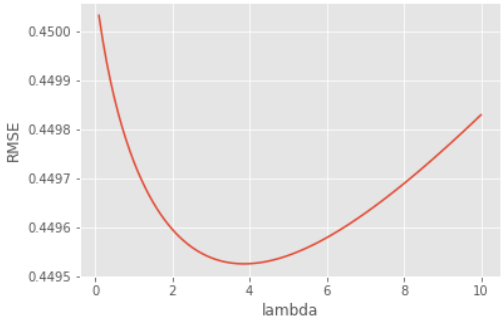

Figure 4: 5-fold Cross-Validation for Ridge Regression

3 Potential Improvement

We have performed some EDA and simple modeling but there are still space for improvement over the rest of the semester. The best RMSE available we have so far is 0.45, which is relatively high. We will try to build more complex models to reduce the RMSE and prevent underfitting. Potential choices could be tree-based models and model aggregation. Another idea we plan to work on is to implement word embedding that analyzes how the name of the listing(for example, "Clean & quiet apt home by the park") might influence price. In the next stage of the project, we aim to predict a fair price of a property, and possibly also predict the desired room type a tenant can find given his/her pricing budget.