IEOR 4523 Data Analytics

Final Project

Airbnb Price Prediction

Group: Sales Machine

Group Member: Ellen Wang, Lacey Long,
Olivia Guo, Shenyi Lu, Yousa Song

I. Introduction

The dataset we are going to use for the final project is the New York City Airbnb data in the year 2021. Since Airbnb was founded in 2008, accommodating with Airbnb has become popular for all kinds of travel purposes, especially in NYC where hotel rates are expensive and room sizes vary a lot. We retrieved data from the website, Inside Airbnb, which contains detailed historical data of listings and reviews of airbnb hosts around the world. In this project, we used the detailed listing data, detailed review data, and neighborhood geojson data of New York City.

In this paper, we would like to figure out the pricing methodologies of the Airbnb hosts so that we could provide pricing suggestions for future hosts. First, we will explore the data in order to find insights into features influencing the listing price. Then, we will perform several machine learning models to predict the listing price based on our analysis.

II. Data Cleaning and Preprocessing

Our dataset consists of two sections. The first section is the major attributes of the housings while the second section is the comments from the customers for the housing they have selected. After brief browsing of both datasets, we keep the 18 attributes out of the complete 70+ ones. Then, we conduct further data cleaning for the columns we retain as follows.

1. Listings data

   In order to make the data type meaningful, we first convert the type of housing id into the string and the types of host response rate and price into the number.

   Next, we only keep the maximum minimum night as 365 and drop the rest, since the number of minimum nights would rarely exceed 365.

   At last, we convert the categorical variables including the neighborhood groups and the room types into dummy variables, so that it is easier for us to conduct the exploratory data analysis and the modeling of machine learning.

2. Reviews data

   For the review data, because the original dataset contains all the review data since 2009, we clean the dataset by conditioning on the review date from November 2020 to November 2021 and drop rows that contain missing values. All the following analysis is performed with the new dataset, ‘review_ltm.csv’.

Finally, we combine the two datasets into a comprehensive one to conduct the following analysis (Table 1).

III. Exploratory Data Analysis

1. Neighborhood

   In order to get a thorough knowledge about the neighborhood information, we use GeoJson to find the price of the housing. We start exploring the neighborhood information in two different approaches.

The first is to see how many housings are in each neighborhood (Figure 1). Here, we clearly find out that there are more housing listings in the Brooklyn area.

The second is to see how the price is distributed in each neighborhood (Figure 2). Here, we find that the pricing in Manhattan is slightly higher than in the other areas. In order to further confirm it, we draw the boxplot to compare the difference (Figure 3). Here, we clearly see that the pricing in Manhattan is a lot higher than in the other areas.

2. Room Type

When learning about the room types, we also use two dimensions to explore it, which is the distribution in each neighborhood (Figure 4) and the price difference among the room types (Figure 5).

We find that there are most rooms in the type of entire home/apartment. And most of them are located in Brooklyn and Manhattan.

Secondly, the most expensive listing belongs to the room type of hotel. But on average, the most expensive room type is the entire home/apartment.

3. Availability

For the availability, there is not an apparent pattern for the distribution (Figure 6). But, we can find that after removing the extreme values (availability > 365), the top-three room types with the highest availability are the shared rooms in Queens, the hotel rooms in Queens, and the hotel rooms in Manhattan.

4. Room Characteristics

In order to grab information from the name of the housing, we divide the housing into two groups, which are the housing with prices below average and above average.

For the listings with prices lower than average (Figure 7), the most common keywords are “private room”, “Brooklyn”, and “furnished”. Those features are not attractive enough for customers and thus result in a lower price.

For the listings with prices greater than average (Figure 8), the most common keywords are “garden”, “modern”, and “duplex”, which indicate the high value of the listing and thus lead to a higher price.

5. Sentiment Analysis

To better measure the review rate for each listing, we decided to create our own measure based on the comments provided by the customer. We used the NRC emotion text file to evaluate the sentiment of each comment in the review table. After generating the sentiment score for each listing’s comments, we combined all positive and all negative sentiment scores and created ‘gd_review’ as one parameter to measure each listing’s review. The ‘gd_review’ is calculated by the number of positive comments (if the sum of all positive sentiment scores is higher than the sum of all negative sentiment scores) divided by the total number of reviews for the listing.

IV. Modeling and Result Analysis

We apply four modeling methodologies as our analysis techniques, which are multiple regression analysis, KNN regressor analysis, regression tree analysis, classification tree analysis, and random forest analysis. After thoughtful model establishment, we get the modeling results and the measurement metrics (Table 2).

1. Multiple Regression

   The multiple regression model is an extension of linear regression models that allows for multiple features. We used the variables in Table 1 (except Price Interval and Price ($)) as our independent variables and use Price ($) as our dependent variable. We applied the multiple regression in scikit-learn to the training set and obtained the intercept and coefficients. Then we used the trained model to predict the Y value of the new data point in the testing set.

2. K-NN Regression

   The K-Nearest Neighbors Algorithm uses feature similarities (usually a distance measure) of K closest points to the new data point and predicts the Y value of this new data point. We applied the KNN regression with the same training set, testing set, and features as in the multiple regression.

   To find the best K for the model, we tried K values ranging from 5 to 45 with a step size of 5. Similar to multiple regression, we evaluate our model on the square root of mean square error and R-square value of the testing data. The result of the KNN model shows that when k = 25, we have the lowest RMSE and highest testing R-square (Table 3).

3. Regression Tree

   The regression tree model is trained through binary recursive partitioning. The Y value here is the continuous valuable Price ($) in Table 1. For this type of model, the depth of the tree is the hyperparameter. We tried different values of the depth ranging from 1 to 10 and trained corresponding 10 models. Then we predicted the Y values for the data in the testing set and calculated the testing R-square and testing RMSE. The result shows that the model with depth 4 has the best testing R-square and RMSE.

4. Classification Tree

   Similar to the regression tree model, the classification tree model is also trained through binary recursive partitioning. The difference is that instead of using a continuous valuable as the Y value, this type of model uses a discrete variable. To use this type of model, we clustered the continuous valuable Price ($) into 21 intervals $[0, 100), [100, 200), \ldots, [1900, 2000), [2000, \infty)$. If a value of Price ($) belongs to the interval with index k, then we set the corresponding value of Price Interval to be k. Then the Y value here is the discrete valuable Price Interval. For this type of model, the depth of the tree is the hyperparameter. We tried different values of the depth ranging from 1 to 10 and trained corresponding 10 models. Then we predicted the Y values for the data in the testing set and calculated the testing R-square and testing RMSE. The result shows that the model with depth 8 has the best testing R-square, with an accuracy of 0.627 on the test data.

5. Random Forest

   The random forest model builds several decision trees from the training set and classifies the data in the testing set based on the "vote" from the built decision trees. In this type of model, we set the number of trees in the forest to be 10 to 100 with a step size of 10, the maximum depth of the tree to be 2 to 8, the minimum number of samples that split an internal node to be (2, 4, 8), and the minimum number of samples that are at a leaf node to be (4, 8, 12, 16). After training, the best model has an R-square of 0.6383 and an accuracy of 0.623. The parameters for the best model have a maximum depth of 8.

6. Model Comparison

   We divide our models into two categories, the regression models and the classification models. Among the regression models, the best model is the regression tree as it has the lowest RMSE and highest R-square. The performance of the classification models is similar, however, the Random Forest model has a higher R-square with slightly lower accuracy on the test data.

V. Improvements

1. Model Improvements

   We can perform the Principal Component Analysis on the original dataset containing all features of the listing to see if we've missed out on any important features and reduce the dimensions of total features of our models. By doing so, we can also see what are the influencing features for Airbnb listing price in practical terms. In addition to improving model performances, because of our limited study on machine learning, we only performed five models in this project. With additional time and resources, we can include more types of algorithms such as XGBoost and neural networks.

2. Dataset Improvements

   Since our price prediction model only focuses on the data of NYC listings in the last twelve months, we are interested in expanding our data from 2020 to the present to see how COVID-19 has influenced the pricing of Airbnb. To better understand the trend, we can also include multiple cities' listing data to see if the trend is consistent with other cities across the world.

3. Project Improvements

   Because our dataset provides limited features for the listings, by incorporating new information such as distance to subway stations, taxi availability, or the number of restaurants near the listing we could build a more comprehensive dataset. These additional features can be obtained by web scraping using the longitude and latitude data. After having all these features, we can create a recommendation model to recommend listings to consumers based on their preferences on transportation, price, neighborhood location, the purpose of the trip (business, shopping, sightseeing, or food hunting) to provide the top listings.

VI. Appendix

Table 1: Full description of the dataset after cleaning and combination

Variable Name	**Description**
ID	The unique ID for each housing
Name	The name of the housing
Instant Bookable	Whether the housing could be immediately booked
Host is Superhost	Whether the host is a superhost
Accommodates	How many people the housing can occupy
Minimum Nights Type	How many nights the customers should at least book (categorical)
Minimum Nights	How many nights the customers should at least book (numerical)
Availability 365	How many nights the housing is available in the next 365 days
Neighbourhood Group Manhattan	Whether the housing is located in Manhattan
Neighbourhood Group Brooklyn	Whether the housing is located in Brooklyn
Neighbourhood Group Queens	Whether the housing is located in Queens
Neighbourhood Group StatenIsland	Whether the housing is located in StatenIsland
Longitude	The longitude of the exact location of the housing
Latitude	The latitude of the exact location of the housing
Room Type Entire	Whether the room is a entire room
Room Type Private	Whether the room is a private room
Room Type Hotel	Whether the room is a hotel room
Numbers of Review LTM	How many reviews there are in the past 365 days
Host Response Rate (%)	What percentage of the reviews the host response
Review Scores Rating	The average rating of the housing from the customers
GD Review Rate (%)	The percentage of the positive reviews for the housing
Price Interval	How much is the housing for one night (categorical)
Price ($)	How much is the housing for one night (numerical)

Figure 1: Number of rooms across various neighbourhoods

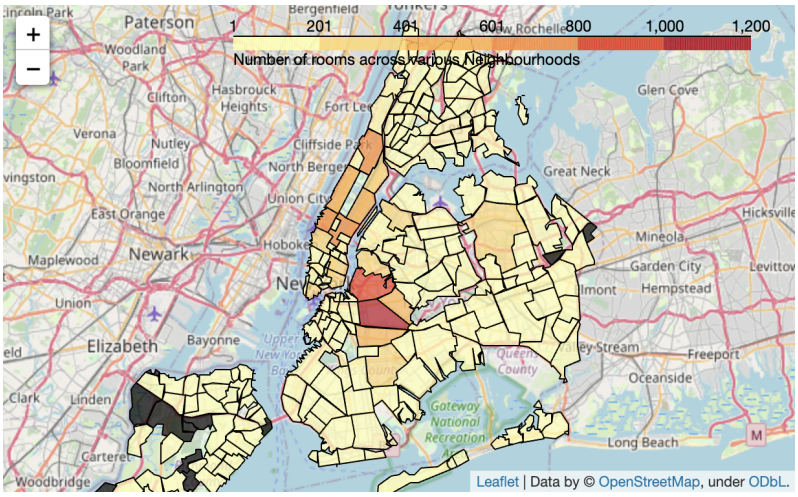

Figure 2: Average price across various neighbourhoods

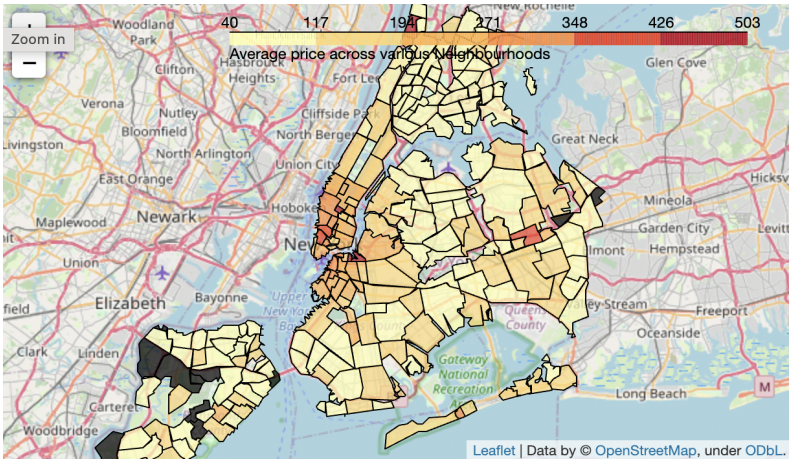

Figure 3: Neighbourhood group price distribution

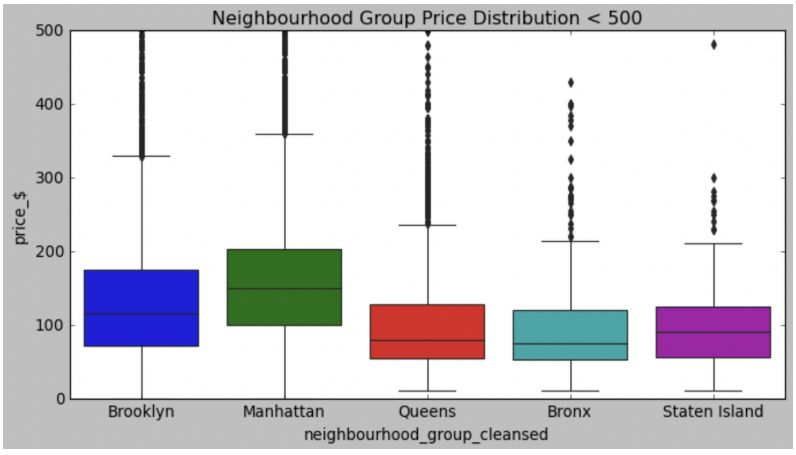

Figure 4: Number of different room types in each neighbourhood

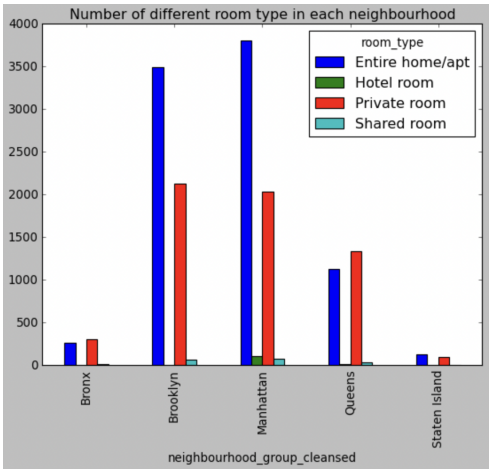

Figure 5: Average price for different room types in each neighbourhood

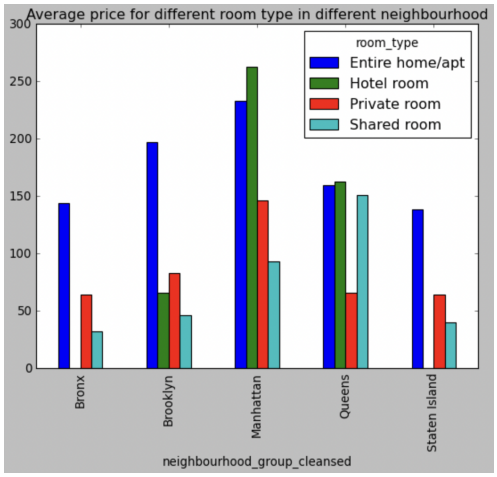

Figure 6: Average availability for different room types in each neighbourhood

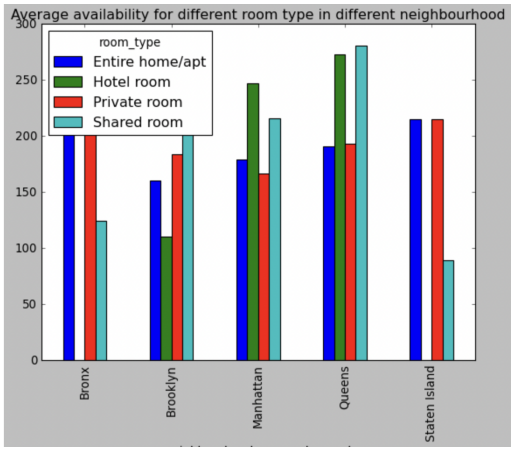

Figure 7: Listings whose price is lower than average

Figure 8: Listings whose price is greater than average

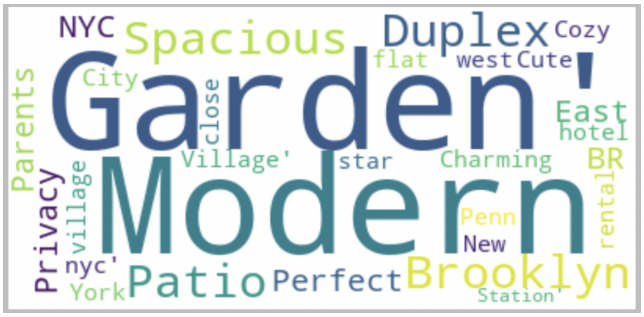

Table 2: The measurement metrics of the machine learning methodologies

Model Group	Model	R-Square	RMSE (Accuracy)
Group 1	Multiple Regression	0.346	138.3605
	KNN Regressor	0.135	159.2085
	Regression Tree	0.381	134.6834
Group 2	Classification Tree	0.627	0.627
	Random Forest	0.638	0.624

Table 3: KNN regression metric output

	K	RMSE	Training R-square	Testing R-square
0	5	165.072673	0.365978	0.070268
1	10	159.919347	0.267264	0.127412
2	15	159.501065	0.222980	0.131971
3	20	159.374499	0.195932	0.133348
4	25	159.208562	0.174848	0.135151
5	30	159.777871	0.157739	0.128955
6	35	160.057934	0.146053	0.125899
7	40	160.359197	0.135113	0.122605
8	45	160.564622	0.126781	0.120356

Figure 9: R-square over models

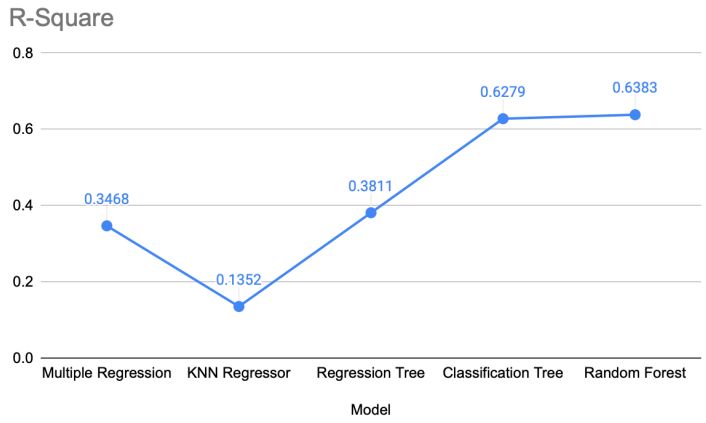

Figure 10: RMSE over models