Introduction

Correct pricing has always been one of the most important factors in for-profit organizations. An inappropriate price either above or below the intrinsic product value will result in a loss. Unlike traditional hotels, which have professional pricing strategies to adjust their price depending on factors such as seasonality, location, etc. Individual Airbnb hosts might not be able to set an appropriate price due to lacking market information.

In this analysis, we will develop a pricing model by taking Airbnb market information in New York City into account and predicting an appropriate price to assist local hosts in better setting their listing prices. This analysis would be a regression task since our target is to predict the Airbnb price per night in USD.

The data we used in this analysis is New York City Airbnb Open Data. This dataset contains information about all Airbnb Listings in New York City in 2019. The original data was sourced from publicly available information from the Airbnb site and collected by Inside Airbnb site. This dataset has been previously used for many Kaggle Data Challenge projects. Among them, the most common projects are Exploratory Data Analysis (EDA) and predict the availability of a given Airbnb listing. For the EDA-related projects, their potential goal is to have a descriptive analysis of the New York City Airbnb market in 2019. In their study, they found that the Airbnb price per night is highly skewed and the downtown area tends to have a higher price. Besides, among all types of rooms, the entire room is the most popular. For the availability prediction-related projects, the author has rearranged the numerical variable availability into a binary categorical variable which represents whether a given Airbnb has a whole year availability. The author eventually received an Accuracy of 0.83 and a ROC AUC of 0.77. However, I think this result is not appropriate since the target variable is highly imbalanced. I would recommend using precision and recall instead of ROC AUC.

This dataset contains 48,895 rows and 16 columns. Among them, 4 variables contain missing values, they are Last_review (20% are missing), reviews_per_month(20% are missing), host_name (0.04% are missing), and name (0.03% are missing).

Exploratory Data Analysis

In our dataset, there are originally 15 predictor variables and 1 target variable. The target variable y is the price per night in USD for a given Airbnb. The distribution of this variable is highly right-skewed. This makes sense because prices can be very expensive but cannot be below zero. It is also a unimodal distribution with its mode at $100. By calculation, about 2.7% of data can be considered extreme outliers. Among them, the most extreme outlier or the maximum price is $10,000. The minimum price in this dataset is $0, we should drop those 11 observations with a price equal to $0 since it is impossible to book an Airbnb for free. In another scenario where $0 might indicate a missing value, we should also drop them because missing values are not allowed in the target variable. After dropping those rows we will have 48,884 remaining.

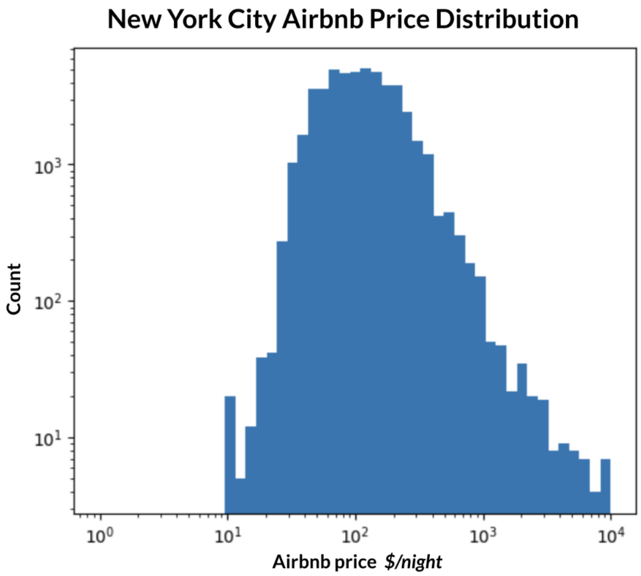

Fig. 1, This plot shows the distribution of Airbnb price per night in USD

For the predictor variables, we first dropped identifier variables id (Airbnb id) and host_id. The reason is that they just function as an identifier and will not give us any information related to the target price. We will also drop unstructured text variables such as name (Airbnb name) and host_name. They might influence the price in such a way that certain words may be more attractive to the customer. However, it requires more advanced NLP strategies which are not in the scoop of this analysis. Next, we convert a date type variable last_review which shows the date of the last review

Airbnb received into a numerical variable gap_between_last_review_and_end_of_2019 which shows the number of days between the last review date and 2019/12/31. After feature engineering, we have 11 predictor variables remaining.

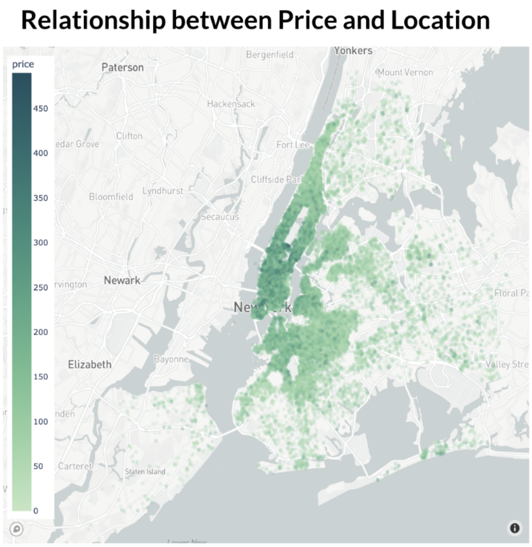

Fig. 2, This plot shows the relationship between Airbnbs price and their geological distribution.

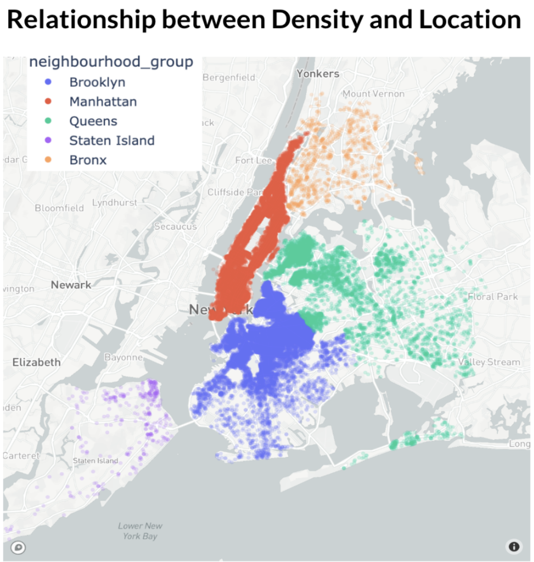

Fig. 3, This plot shows the relationship between Airbnbs neighbourhood and their geological distribution.

By setting aside extreme outliers, we can see there is a clear relationship between the pair of coordinate variables and the target variable price. In Fig. 2, the darker shade represents a more expensive price. As we can see, high price Airbnbs are mostly located in the downtown area (Manhattan and North Brooklyn region). And the price gradually decreased as we moved away from the downtown area. Similarly, from Fig. 3 you might notice that dots are more close to each other in downtown areas. This means that there are more Airbnb in North Brooklyn, West Queen, and Manhattan region. This makes sense because most tourist spots are located in the downtown area, thus the high demand results in an increase in Airbnb prices. Meanwhile, there will be more supply of Airbnbs to meet the demand.

Another interesting relationship I found is between price and reviews received by Airbnb per month. From Fig. 4 we can see that most dots are gathered in the bottom left corner. And there is a clear negative correlation between price and the review rate. This is reasonable because as the price increases, fewer people can afford it, which results in lower review rates.

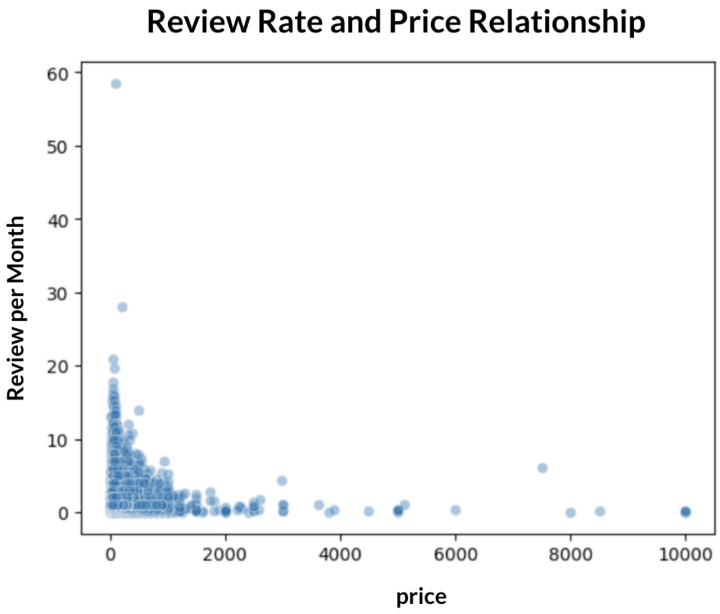

Fig. 4, This plot shows the correlation between monthly review rate and the daily price for Airbnb

Data Preprocessing

Since our dataset does not have any group structures or time-series patterns, we will keep the iid assumption. However, since the target variable price is highly skewed, I have used the package scsplit to perform the basic regression stratified splitting. The reason I used basic splitting instead of K-fold is that we have a reasonably large dataset with 48,884 observations. The randomness due to splitting for a large dataset is not a big issue. After splitting, we have 70%

for training, 20% for validation, and 10% for tests. The reason I used 7:2:1 instead of 6:2:2 is that the later proportion will result in an automatic row dropping due to regression stratification. The 7:2:1 proportion can help us to preserve the original number of observations.

In terms of preprocessing. I have used 4 types of preprocessors.

- For variable room_type I used the ordinal encoder because there is an ordinal relationship among shared rooms, private rooms, and entire rooms.
- For variable neighborhood groups and neighborhoods, I used the one hot encoder because there is not an ordinal relationship among different neighborhoods.
- For variable availability_365, I used MinMaxScaler since the available day in a year is bounded between 0 and 365.
- For the rest of the variables, I used StandardScaler because all of them are numerical variables and are not bounded.

After preprocessing, the number of predictor variables has been increased from 11 to 231. This is because the one hot encoder has converted each element in categorical variables into a new feature.

In the original dataset, 4 out of 16 variables contain missing values. They are last_review (20% missing), review_per_month (20% missing), host_name (0.04% missing), and name (0.03% missing). We will ignore the last two variables since we have already dropped them. And we will keep those missing values in the first two variables until I learn more advanced methods to handle missing values.

Reference

1. *Dgomonov. (2019, August 12). New York City airbnb open data. Kaggle. Retrieved October 18, 2022, from https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data*
2. *How is Airbnb really being used in and affecting the neighbourhoods of your city?* Inside Airbnb. (n.d.). Retrieved October 18, 2022, from http://insideairbnb.com/
3. Cao, J., Choi, T., & Khare, R. (n.d.). *NYC Airbnb Listings in 2019: Determining Factors that Affect a Listing's Price*. Retrieved from https://www.stat.cmu.edu/capstoneresearch/spring2021/315files/team5.html
4. Gcdatkin. (2020, October 30). *NYC Airbnb availability prediction*. Kaggle. Retrieved October 18, 2022, from https://www.kaggle.com/code/gcdatkin/nyc-airbnb-availability-prediction

GitHub Repository