Airbnb Popularity Prediction Project

The Objective

Most people have used Airbnb before. I know when I want to go somewhere, Airbnb is the first platform I search. How does your search process work when you are looking for an Airbnb? Do you value location over everything, trust a few amazing reviews, or go for the cheapest option? **The motivation of this project was to find insight into what factors influence the popularity of Airbnb listings, and use that to predict whether or not a listing will be popular.** The data feature to determine popularity in our project is the number of reviews per month.

About the Data

Understanding the Dataset

Our journey to understanding the popularity of Airbnb listings starts with a public dataset containing data on 50 000 Airbnb listings in New York City in 2019. For each listing, the dataset contains data on:

- **Basic listing information**: listing name, host's name, the type of room
- **Location data**: the neighbourhood name, longitude, and latitude
- **Price and Availability**: the price per night, and number of days per year the property is available
- **Reviews**: Perhaps the most important feature for our analysis, this includes the number of reviews, the date of the last review, and the number of reviews per month.

Initial Dataset Insights

In our initial exploration, several intriguing patterns emerged:

1. As seen below, listings in certain neighbourhoods are on average more popular than in others. Evident in the figure below, the listings in Queens are often much more popular than listings in Manhattan and Brooklyn.

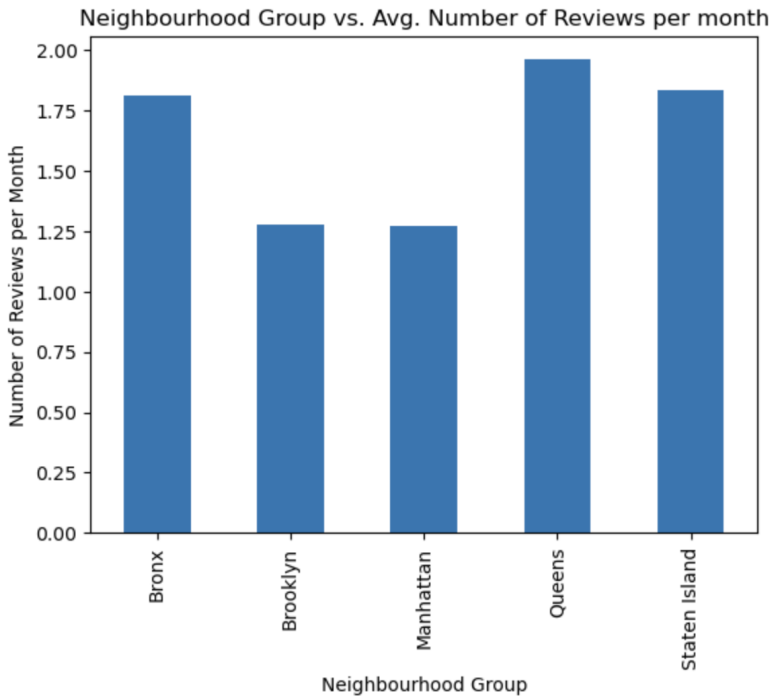

Figure 1: Neighbourhood vs. Average Number of Reviews per month in the Dataset

2. The dataset contains a wide variety of nightly prices. As you can see in the diagram below, it also appears upon initial look from the figure below that the lower-priced homes have more monthly reviews.

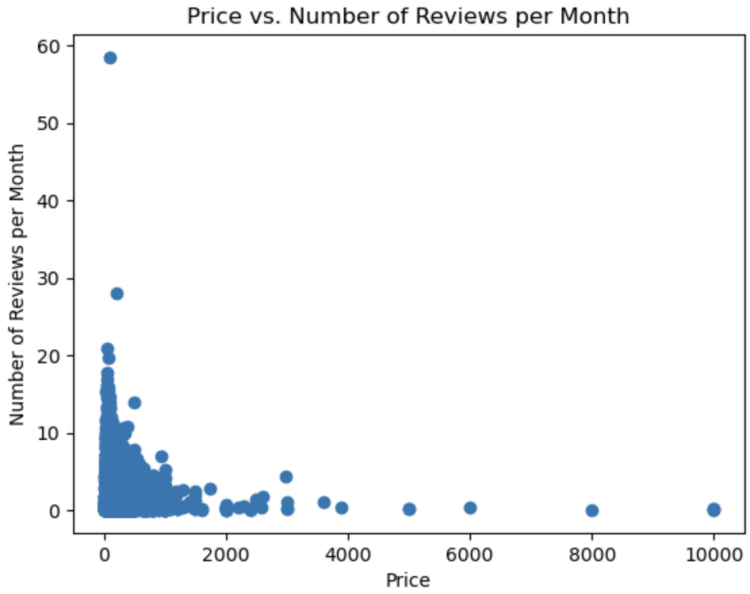

Figure 2: Price vs. Number of Reviews per month in the Dataset

About the Model

Machine Learning Models

For our project we tested many different machine learning models. For context, a machine learning model is like a defined methodology that obtains predictions from an input data set. These models range from quite simple and relatively easy to understand, to very complex and computationally expensive.

Initial Testing

Initially we tried a type of model that is relatively easy to understand. It works similar to a choose your own adventure book, asking a series of yes or no questions about the features of each data point and assigning a prediction based on the path the answers took. This model is called a decision tree. It performed surprisingly well for a relatively basic model, but we still wanted to see better results.

The Winning Model!

Due to the surprising performance of the decision tree, we looked for ways to use similar models with some additional insight. What we settled on was a model that essentially tests decision trees one by one, with each decision tree trying to correct mistakes and improve upon the previous decision tree that was tested. This model, known as a Light Gradient Boosting Machine, performed the best out of any of the models we tested. We decided to implement it as the machine learning model to help with our prediction problem.

Results

Accuracy

To find out how well your model predicts data, you should train it on data set aside for training purposes, where you can fit the model, and then test its performance on data it hasn't seen before to test it. The metric we measured our results with for this project was R-squared. It measures how well a model fits the data and explains variability in the data. 0 is the worst R-squared possible and 1 is the best possible R-squared. When we tested the model on the data set aside for the final test we obtained an R-squared of 0.61 which we were satisfied with given the complexity of the prediction problem. This was also an encouraging score because the test score was essentially the same as the scores obtained from training, which shows that we didn't create a model that only worked for our training data. As you can see in the diagram below, there is definitely a correlation between our predictions and the actual number of reviews, as our predicted number of reviews increases, the actual number of reviews tends to increase as well.

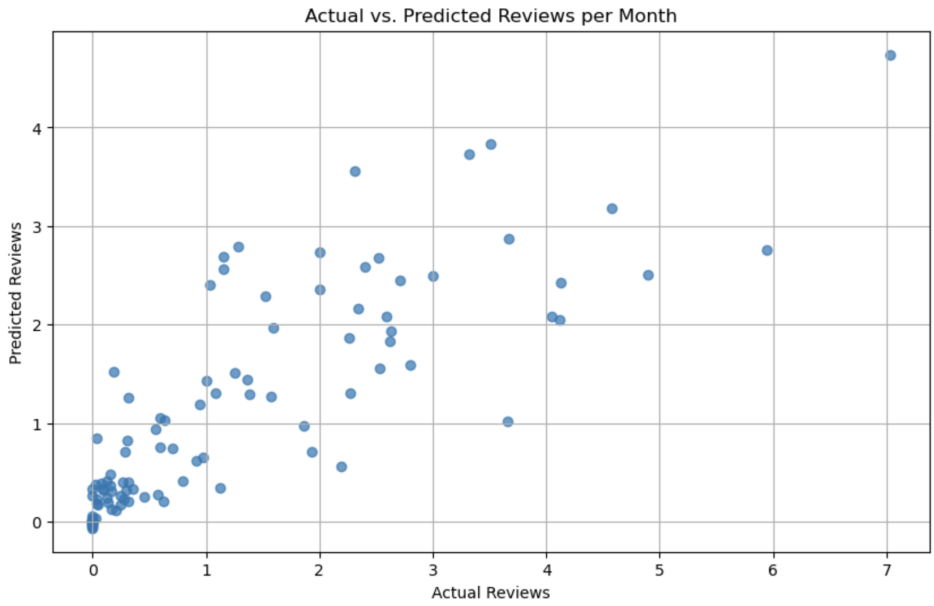

Figure 3: Actual vs. Predicted Reviews per Month of our Model for 100 Randomly Sampled Unseen Listings

Feature Importance

From our trained model, we can obtain insight on how it makes predictions. A really useful indication of this is feature importances. This is essentially a weight the model finds that determines how important a feature is to the prediction. For example, if you have a model predicting your score on a final exam, the feature importances of features like hours spent studying and midterm exam marks would be high, but feature importances of features like height and favourite color would be low. For our project, the most important feature by far is the number of days since the last review. Other important features were how many days in a year the Airbnb was available, price, and the minimum number of nights you have to stay at the Airbnb.

Caveats

1. **Proxy for popularity:** The number of reviews per month for a listing doesn't necessarily accurately represent a listing's popularity. Had the dataset contained information about each listing's average rating or vacancy rate, we could have used other data as proxies for a listing's popularity.
2. **Temporal relevance:** The dataset we used was based on data in 2019. However, Airbnb is a dynamic marketplace. Changes in travel trends, local events, or even global occurrences (like a pandemic) can significantly alter what makes an Airbnb listing popular. Therefore, our model might not fully capture these evolving trends, and its predictions could become less accurate over time without regular updates with new data.
3. **Use of R^2:** To prevent any misleading conclusions from our analysis, it should also be stated that R^2 measures correlation, not causation. For instance, just because a feature like a listing's minimum nights of stay required is strongly correlated with popularity, it

doesn't necessarily mean that reducing the minimum nights of stay will automatically increase a listing's popularity