Airbnb Listing Price Prediction

Benny Kouk

Introduction

Background

Airbnb, Inc. started in 2008 when two friends hosted a space to three travelers who needed a space to stay. Today, Airbnb, Inc.is one of the largest online marketplaces worldwide, for connecting people wanting to rent out their properties and people looking for accommodation in the same area. On average at least two million people rest their heads in an Airbnb property each night. Prices of these rental properties are dependent on myriad of factors, from the type of place offered to the level of amenities provided. This project explores how the location of a rental properties, specifically the venues surrounding it affects its listing price.

Problem

Predicting the listing price of a rental property or referred to as an “Airbnb”, given its nearby venues.

Interest

Knowing the price of an Airbnb can be valuable information for property owners new to the service or those who cannot decide on a price for their Airbnb. Having this information gives property owners some insight on the value of their Airbnb and ensures that a property owner does not undercharge and end up being shortchanged – or overcharge and not get any customers. On the economic level, this brings value by effectively maximizing the profits of said property owners. One interesting venture could be for Airbnb, Inc. to adopt this as a feature on their website or application, suggesting prices when property owners create a listing. This orientates new users in the creation of their listing, who might not know how much to charge.

Data

Data source

The data used in this project was from the Kaggle dataset U.S. Airbnb Open Data. The dataset is csv file which has features such as host id, hostname, listing id, listing name, latitude and longitude of listing, the neighbourhood, price, room type, minimum number of nights, number of reviews, last review date, reviews per month, availability, host listings and city.

Methodology

Outliers

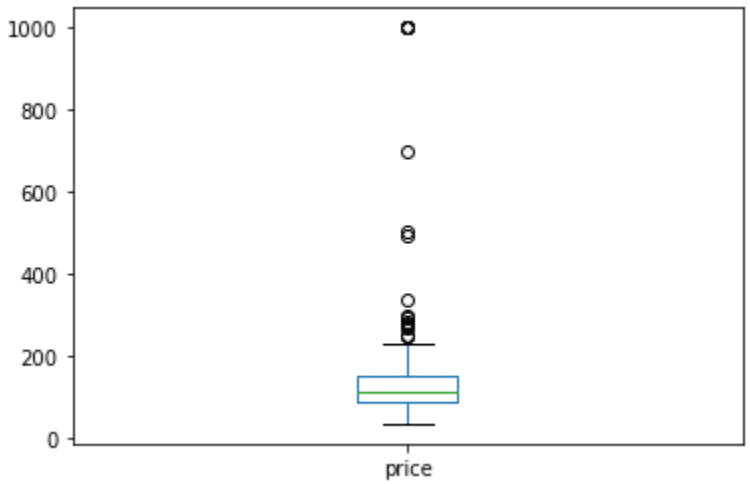

Fig.1 Box plot of dataframe

Some evidence of outliers is visualized with an initial box plot of the dataframe. However, they were kept in the dataframe they seem to be mainly concentrated in the Midtown, Manhattan locale, signaling some possible correlation between the location and the unusually high prices of Airbnbs there.

Foursquare API

The data of venues surrounding an Airbnb is obtained using the Foursquare API. A sandbox account entitles 950 Regular Calls/Day, 50 Premium Calls/Day, 1 Photo per Venue and 1 Tip per Venue. The data obtained using the API is then parsed into a new dataframe containing the nearby venues for all Airbnbs in the dataset, and its corresponding venue categories.

Scope

Variations in the price of an Airbnb could be accounted for by a multitude of variables, and hence, it is a highly sensitive variable. In order test the correlation of Airbnb prices and with its profile of nearby venues, a subgroup of Airbnbs was chosen. These were Airbnbs that were full studio apartments in New York City, not requiring a minimum lease of 30 days, having more than 3 reviews. Not only does selecting such a specific subset of Airbnbs increase the credibility of results, but it also allows us to fit the dataset into the Foursquare API's daily limit. The dataframe of this subgroup consisted of 528 Airbnbs.

One-hot encoding

Since venues categories are categorical values, one-hot encoding is performed on the data before fitting it to a machine learning algorithm.

Choice of Models

Since the problem is price prediction, and price being a continuous value. The choice of models can be narrowed down to a choice of regression models. Experimenting with various regression models, using Mean Squared Error and R^2 Score as model evaluation metrics. The models that evaluated as the best were GradientBoostingRegressor and DecisionTreeRegressor.

Results and Discussion

GradientBoostingRegressor

GradientBoostingRegressor predicted Airbnb prices with a Mean Squared Error of 1619.56, and a R^2 Score of 0.85.

DecisionTreeRegressor

DecisionTreeRegressor predicted Airbnb prices with a Mean Squared Error of 2135.13, and a R^2 Score of 0.86.

Performance of Models

Model	MSE	R^2 Score
GBR	1619.56	0.85
DTR	2135.13	0.86

Fig.2 Performance of Models

The performance of the regression models used are limited by the size of the data. This is largely a limitation due to the Foursquare API limit of 950 Regular Calls/Day for the sandbox account associated to this project. It is likely that the performance of these models will improve given a larger dataset.

Conclusions

GradientBoostingRegressor and DecisionTreeRegressor performed the best on the dataset, based on the evaluation metrics used, achieving the smallest Mean Squared Errors (Mean Squared Error = 1619.56 and 2135.13) and R^2 Scores closest to 1 (R^2 Score = 0.85 and 0.86). Creating a model achieving a R^2 Score of 0.86 is by no means conclusive of any correlation, and the data used is on a specific subgroup of Airbnbs. It does, however, beg the question of how accurate price prediction would have been given a different subgroup of Airbnbs. Another area of interest could be to examine any tradeoffs when the scope of this project is expanded. Widening the scope, and by extension, size of the dataset could introduce more variables to the dataset; as such, data cleaning will be necessary to preserve a level of accuracy in the results. Undoubtedly, a plethora of insights could be uncovered given changes in methodology, this project being one with a well-defined focus.