Capstone 1 Project Proposal - AirBnB Price Prediction

Alexandra Michel

Background

AirBnb is a website that allows people to search and reserve short-term and apartment rentals, as well as become a host on the platform by opening up their own space to potential guests. All prices are set by the hosts, and guests/renters can narrow their search by different criteria to find a place to stay within their budget while still meeting their needs. As a host looking to earn supplemental income through AirBnB, it is important to know what the market price for your space is so that you can meet the demand of more consumers and ultimately have more guests book a stay with you.

Motivation

There are third-party companies (Like this one, and this one) out there who charge people (a lot) to price their AirBnB for them; probably implementing similar techniques on data that is free to access like the dataset I will use for this project. Also, AirBnB provides hosts with some advice when it comes to pricing, but who knows what attributes they are using to determine those prices, or how successful they are to attract guests (as hosts don't have to go with the suggested price, this would be hard to test!). AirBnB could be interested in this kind of exploration as a way to generate more revenue through their hosts, but more importantly, the average consumer with a spare room or living space is the ultimate consumer of the information this model would provide. Within the real estate industry, this would have a big impact for real estate agents and their clients (homebuyers and investors) looking to use a measure like this as part of the bigger picture in describing and analyzing the potential return on investment for a home purchase.

Data Sources and Methods

Source of data is Inside AirBnB, a group who scrape and compile data from active listings on www.AirBnB.com. They seem to have a mission along the lines of wanting to bring to light how AirBnB's mission may be mis-aligned with the social and economic impact it actually has.

> *Sidenote:* If you check out the InsideAirBnB site, they note "Airbnb claims to be part of the "sharing economy" and disrupting the hotel industry. However, data shows that the majority of Airbnb listings in most cities are entire homes, many of which are rented all year round - disrupting housing and communities. Browse the data for your city below, and see for yourself". I think what InsideAirBnB is doing is really rad because they are utilizing data to easily show the impact of capitalism on the community as a whole. It would be awesome if AirBnB took this information to heart and used it to realign with their mission. Who knows, maybe they do!

I chose to start out with the Los Angeles dataset, mainly because there is enough data in there to get some (hopefully) good machine learning results on, and it is still just one city. I'd hope to expand the techniques I use throughout this project to be applicable to any city, and to be able to compare results between cities to see if there are any differences.

The Data set

Each row in this data set is an available listing on the AirBnB site in Los Angeles, CA. The data set has features related to listing criteria (bedrooms, bathrooms, beds, number of guests), as well as reviews, location, amenities, min/max number of nights. It also has many free-text features which I will drop for the sake of this initial attempt, but I would include them in the future as I learn to implement NLP techniques.

Data Cleaning

I will drop a handful of columns which have mostly NaN values. There are also columns that are closely related to each other which I will condense, and categories that are sparse which I will group together (for example property_type). The dollar-value columns need to be changed to int values as well, and NaN values converted to zeros.

Statistical/ML Investigation

My goal for this project is to be able to predict, with a good level of accuracy, the price of a listing based on the other meaningful attributes in the dataset. I will investigate the relationship between price and a few other features I predict to play a role in this (e.g. bedrooms, number of guests, number of reviews, location, etc.). I will also look at the distribution and descriptive statistics related to the nightly price of the listings to get a sense for what the market looks like. For the actual prediction model, it will be a prediction of prices, so I'll start out with a linear regression model and use feature selection to improve the model. This is not a classification problem, so I will proceed with using linear regression and feature selection to create my model. In the future I could also see what a neural network / deep-learning approach would produce. For further investigation it may also be interesting to do some unsupervised learning on neighborhoods and price (reverse to predict neighborhood based on price and other criteria).

Deliverables

Code and a paper/slide deck with the results