

AirBNB Project MVP

By: Belle Peng | 2018-08-30

Motivation

Travelling and experiencing the local lifestyle through staying at a local's home is a personal passion of mine, naturally I want to work on a project related to travels. My goal for this project is to showcase my diverse skillsets. I want to work on an aggregate project that demonstrates my versatility in working with problems including but not limited to linear modeling, classification, clustering, NLP, visualization, and deep learning. The AirBNB data is rich in depth and comes in many different forms (i.e. text, images, quantitative, categorical, ordinal, etc), which begs the question of how to manipulate several data sources and large datasets effectively. The questions that can be asked of the AirBNB business model open up enormous freedom for creative problem-solving and real world application.

Goals

1. *Host perspective*: Build a **Host Dashboard** that has a sentiment score over time
 - Review summaries over X period for hosts, so they know what to improve on and what they do well in
 - NLP and sentiment analysis on reviews for sentiment score time-series
 - Linear Models for pricing to let the seller know whether the pricing is above or below comparable homes
2. *Customer perspective*: Build **Recommendation Engine** to recommend similar homes in the same city when a user searches on one home
 - Clustering for different styles of stays (luxury, quaint, convenient, community feel, etc)
 - Use image recognition for photos classification (what type of home is it and how nice will it be? How quiet or urban will it be?) - Nice to have
 - Try to find a Classification problem - Nice to have
 - Recommend homes in a different city based on what user searched on / user history if I can find user history - Stretch goal
3. *Visualization*: Interactive Map and Host Dashboard

Data

My application for the AirBNB API has not been approved and may not ever be approved, however I found data updated monthly on many major cities on this site [InsideAirBNB](#). I plan to build the analysis and prototype of my product using just San Francisco data, which contains about 6600 listings as of August 2018 and 280,000 reviews since 2009 to August 2018. I will expand the next iteration of my project to the New York data which contains 51,000 listings and 1+ million reviews, and make recommendations on a NYC AirBNB given someone stayed at a SF AirBNB. The future of this project can be expanded to other cities abroad.

Known Unknowns / Limitations / Challenges / Concerns

With only 2.5 weeks, I may not get to all pieces of my proposal. I prioritized my projects so that it's completed in stages and each stage won't be dependent on another. The image recognition piece is left as the last piece to add on. Since I am getting the data from this static source online, the analysis will be as of August 2018, if I need a more refreshed dataset I am dependent on the refresh schedule of this site. The challenge to working with data abroad is that the reviews are written in different languages, and NLP techniques can be very different for different languages. Therefore, I won't be working on data outside of SF and possibly NYC in this projet, but it's a possible extention for the future.