Charlie Denhart
10/23/18
DS4400

Project Proposal

Title: Predicting Airbnb host performance

Description:

My goal in this project is to attempt to predict how hosts and their properties will perform on Airbnb. In order to evaluate performance, several attributes will be predicted for each host including occupancy rate, number of stars in rating, length of presence in market, and revenue.

Dataset:

Link: http://insideairbnb.com/get-the-data.html

If I am short on time, I will just use the data from Boston, which consists of 11 datasets (from different scrapes) between 2015 – 2018. The aggregated data from all sets has 41,364 observations. Each observation has 96 different features. I will hopefully, however, pull data from several different cities, which would result in significantly more observations. The use of many cities, though, would force me to consider larger scale data processing techniques and therefore the effort is contingent upon how much time I have available for the project.

Approach:

Several features will be engineered in order both to provide labeling and to support models. These features include the occupancy rate, average distance between a host and their units, number of apartment amenities, description length, number of yearly price changes, added revenue from price changes. Other features could also be added from other data sources to supplement inputs for any models. For example, the number of google searches for a city or neighborhood could be used as a proxy for demand for an area.

Once I am satisfied with the features available, I will perform feature selection. Due to amount of data I am likely to have, I will probably stick to quicker methods like filter and regularization. As for models, I will use different types depending on what I am trying to predict. I will try first to predict occupancy rate and revenues as continuous variables using linear and/or non-linear regressions. I might also try binning the labels for those two variables and then trying to classify them. For star rating and length in market (and potentially the binned variables), I will create ensemble models that combine LDA, logistic regression, and SVM for classification. I will also try using neural networks for classification. Models will be evaluated by their accuracy in predicting the metrics named above. Cross-validation will be used to ensure thorough validation.

I will complete this project using python. Some of the packages I plan on using include Numpy, Pandas, Seaborn, and Scikit-learn. If I am using a significant amount of data, I will try to use PySpark or some other parallel computation tool.