Predicting AirBnB Prices

Chan, Natural

14 December, 2021

Abstract

The goal of this project is to estimate Airbnb prices in the San Francisco - Bay Area by using information that we can obtain from querying Airbnb. We will primarily observe listings available on weekends of December, 2021 in order to focus on a set time period. From there, we will process our data and predict Airbnb prices by using a base OLS model. Once we establish a base model, we will continue to refine our data and adjust our model to improve on our predictions.

1 Design

New York City is an incredibly popular but expensive holiday destination for many people in the United States. Although barely comparable, San Francisco is also a great place to visit during the holidays and has a wide variety of affordable places to stay. Using a scraped list of zip codes within 20 miles of San Francisco, we will query Airbnb and create a dataset containing all listings at these locations. We will then use this data to predict Airbnb prices for weekends in the month of December.

2 Data

After web scraping from Airbnb, our dataset contains 20321 observations and 14 features. Unfortunately, Airbnb does not provide the exact location of each listing so we have to settle for the general district/area for each listing. Several of the features such as amenities, listing type, and area can be broken down into categorical features while others such as number of bedrooms or bathrooms can either be categorical or numerical for our purposes.

3 Algorithms

3.1 Feature Engineering

- Removal of duplicate listings (after running several models, we found that date did not have any influence on the pricing so we only keep one of each listing, regardless of date)

- Removal of features (such as rating, number of reviews, fees, etc.) an individual would not have when posting or searching for a listing
- Creating dummy variables for amenities, listing types and location
- Removal of dummy variables that had little impact on the pricing

3.2 Models

We began with Ordinary Least Squares on the entire cleaned dataset as a base model and used this as a starting point to find any features that have minimal impact on the pricing. Once we removed all the features, we used several models with regularization on 80% of the data (testing on the remaining 20%). Using an Elastic Net model with 5-fold cross validation and log transformation on the target variable, our results were as follows:

- Adj. R2: 0.734
- Test R2: 0.716

We also attempted to build use a decision tree regression model in an attempt to see if we can do better. Using XGBoost with minimal parameter tuning, we achieved the following results:

- R2 on Training Set: 0.940
- R2 on Test Set: 0.780

which is a significant improvement despite overfitting on the training set.

4 Tools

- Selenium and Beautiful Soup to scrape data on zip codes and Airbnb listings
- NumPy and Pandas to process data
- Scikit-learn, statsmodels, and XGBoost for modeling
- Seaborn and Matplotlib for visualizations

5 Communication

Presentation slides and plots are located in my github repository.