Capstone Project

Airbnb price prediction in SanFrancisco

Shabnam.attaran@gmail.com

Outline

- Overview
- Business Objective
- Data
- Data Cleaning & Exploratory data analysis(EDA)
- Machine Learning Model
- Conclusion and future work

Overview

Airbnb(Air bed and breakfast)

Airbnb is an online marketplace for arranging or offering lodging, primarily homestays, or tourism experiences. The company does not own any of the real estate listings, nor does it host events; it acts as a broker, receiving commissions from each booking.

Since Airbnb has started more people are using the platform to list their house/apartment/room… for short term rent. The host can rent the place from one night to months. Sometimes the host will rent their house when they are away for a week.

Business Objective

Giving a tool to hosts to be able to estimate the price for the property that they want to list in Airbnb

The challenge that the host face is finding the right price for their listing. One way they can do it is by searching similar places and see what are those listed as and then estimate a price for their listing but this is not accurate and it does not include all the information. (for example the user may think only number of bedrooms is important whereas there are many factors in predicting the price)

So having a tool for predicting the listing price of an Airbnb that takes into consideration several factor (location, number of bedrooms, time of the year for listing,..) would help the host estimate their listing price.

Data

I am using the data from http://insideairbnb.com/get-the-data.html and it is free. They data is as of November 30 ,2019. It contains about 8533 record for San Francisco.

Data Cleaning

Dropping features

- Features related to host were dropped
- Features that had many text were dropped since I was not doing NLP
- Information about when data was scrapped
- Columns that had all null values
- Different columns for the same feature
- Columns with only one unique value
- Dropped “other” column since it was not clear what it is

- Cleaning price, security_deposit , cleaning_fee and extra_people

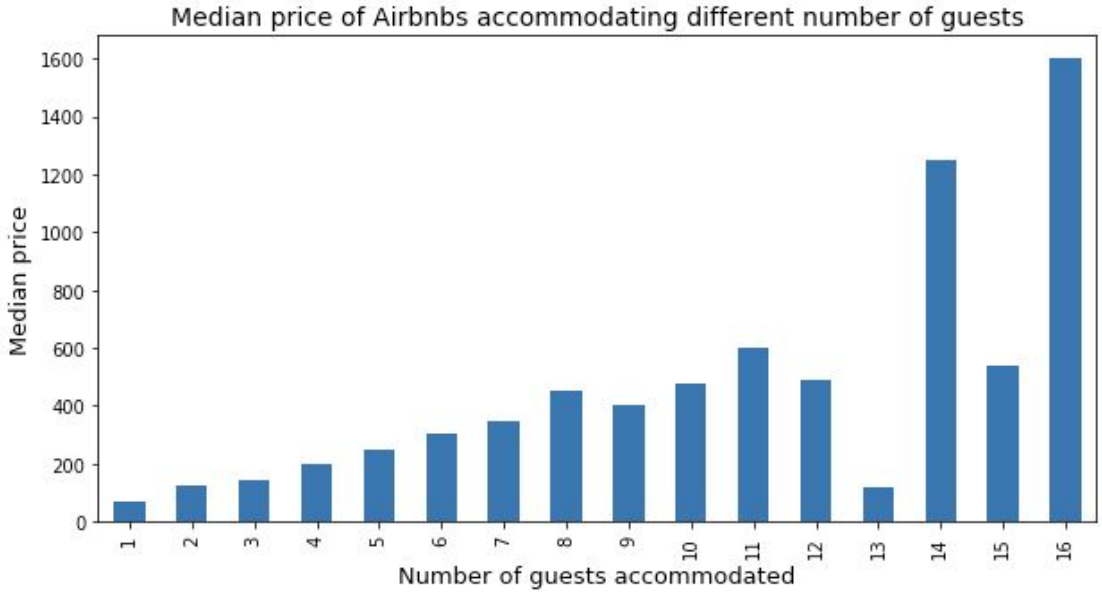
  Removing $ from Price and making it into int

Data Cleaning

- Extracting info from amenities feature
  - amenities gave information about different amenities that are provided in the listing
  - This had 186 different values in different listings
  - I parsed this and grouped some of them together and checked for each entry if they had that amenity and added new columns for those.
  - Added 84 columns .
  - Went through them and if for example one was happening in less than 10, I removed that column.
  - Dropped amenities feature
- Outliers
  - There were places where market was “D.C” . This was kept since other columns were correctly indicating the location is SF. so this must have been an error in the data
  - city= San Jose . This was dropped since the focus was in SF area
  - 54 rows were all null so they were dropped

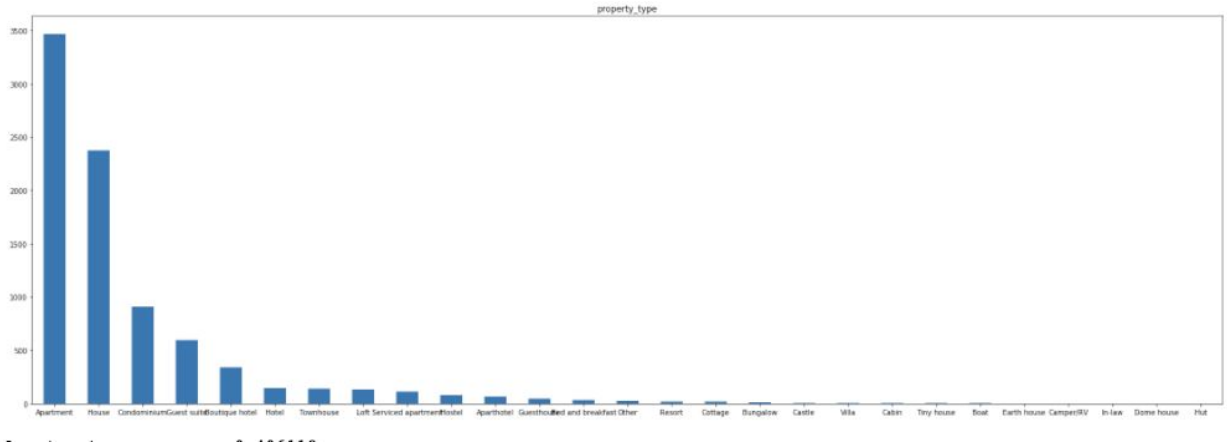
EDA

How number of guests a place accommodates related to the price



EDA

What are the most common properties that are listed in SF?



Machine Learning Model

- Pipeline
- Ridge Regression & GridSearch
- Cleaning Null Values on Bedroom, Bathroom and Beds
    - Use Median for missing values as part of the pipeline
- Split the training and test data
- Using hyperparameter for the ML model

Conclusion and future work

The features that had the highest coefficient are :

Accommodates	22.870323
Cleaning_fee	0.3148890
security_deposit	0.0423644

Surprisingly features related to location (zipcode,neigborhood,lattitude/longtitude) didn’t really make any differences in the result and they did not appear in the top features with high coefficient.

Conclusion and future work

Based on the model and the output I got this is not giving the most accurate result.
One of the thing I did not use in my model is anything related to the host since I was assuming if someone wants to put their property in Airbnb for the first time there wouldn't be any previous knowledge of the host but it may be related for the hosts that are already in the platform and the reviews that have received on other properties. Those reviews and the information about the host can be used to see if the result will change.
This dataset may have not been enough to give the first time host the best possible model.

Also I was highly expecting the location to be an important feature in predicting the price which didn't seem so. That can be because of multiple reasons. We may have different prices in the same neighborhood depending on whether if it is the whole apartment or if it is only one bedroom that is being rented.

So for a future work , I recommend using features related to host and reviews and see the result based on that.