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 scraped
  - 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

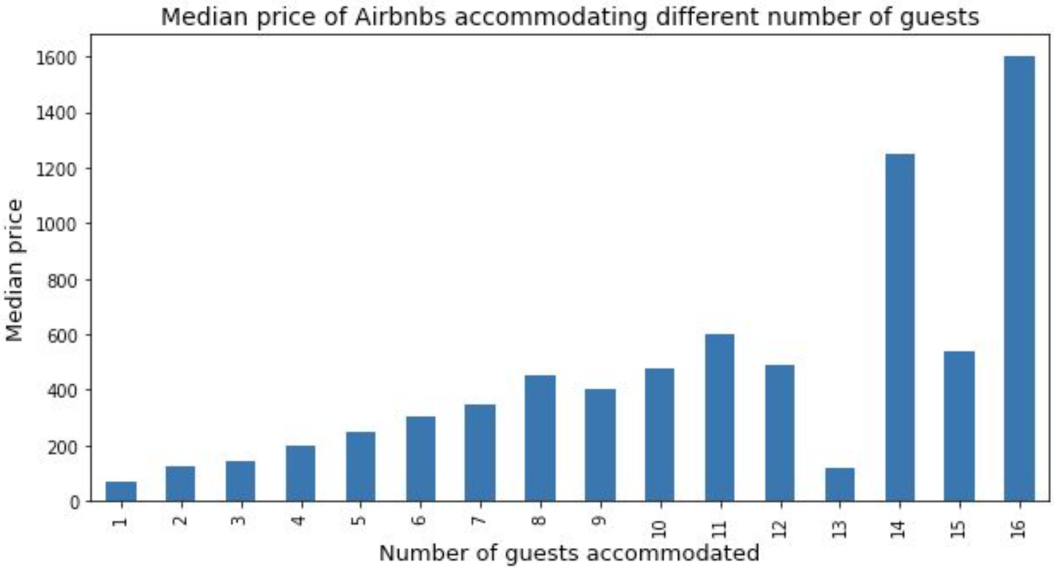
  Removing $ from Price and making it into int

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 (some examples of the amenities are wifi, microwave, free parking, gym,...)
  - 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
  - To further evaluate the amenities , a binary feature is created for each amenity, with 1 and 0 indicating the presence or absence of the amenity in the listing(84 new features were added here)
  - Went through them and if for example one was happening in less than 10, I removed that column
  - After extracting the info from amenities I 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

EDA

How number of guests a place accommodates related to the price?



We can see there is a clear linear relationship between price and the number of guest a place can accommodate

Machine Learning Model

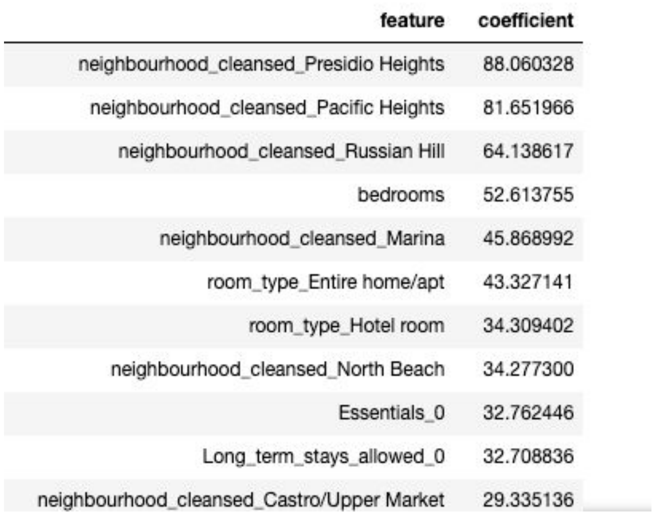
- Pipeline
- Ridge Regression & GridSearch
- Cleaning Null Values on Bedroom, Bathroom and Beds
    - Use Median for missing values as part of the pipeline
- Split the training and test data
- Using hyperparameter tuning for the ML model

Conclusion and future work

R^2 Score : 0.46

RMSE : 21287

The features that had the highest coefficient are :



feature	coefficient
neighbourhood_cleansed_Presidio Heights	88.060328
neighbourhood_cleansed_Pacific Heights	81.651966
neighbourhood_cleansed_Russian Hill	64.138617
bedrooms	52.613755
neighbourhood_cleansed_Marina	45.868992
room_type_Entire home/apt	43.327141
room_type_Hotel room	34.309402
neighbourhood_cleansed_North Beach	34.277300
Essentials_0	32.762446
Long_term_stays_allowed_0	32.708836
neighbourhood_cleansed_Castro/Upper Market	29.335136

Conclusion and future work

- Some features that are most important are neighborhood ,bedrooms and entire apartment/house.
- Another interesting feature was if the long term stay is allowed the price will be higher. I guess that is because some people may stay in Airbnb for a long time instead of leasing a place.
- From the accommodate column that we extracted other features it seems only the “essential”is in the top 10 features.

In this model we eliminated the prices that were over 2000 a night and we consider them luxury. So our model will work for the non luxury rentals.

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.For the hosts that are already in the platform we can use those information to give them a suggestion to adjust their price.Those reviews and the information about the host can be used to see if the result will change.

So for a future work , I recommend using features related to host and reviews and see the result based on that for the hosts that have their property already listed so that they can get a better estimate of the price they should list.

Conclusion and Future work