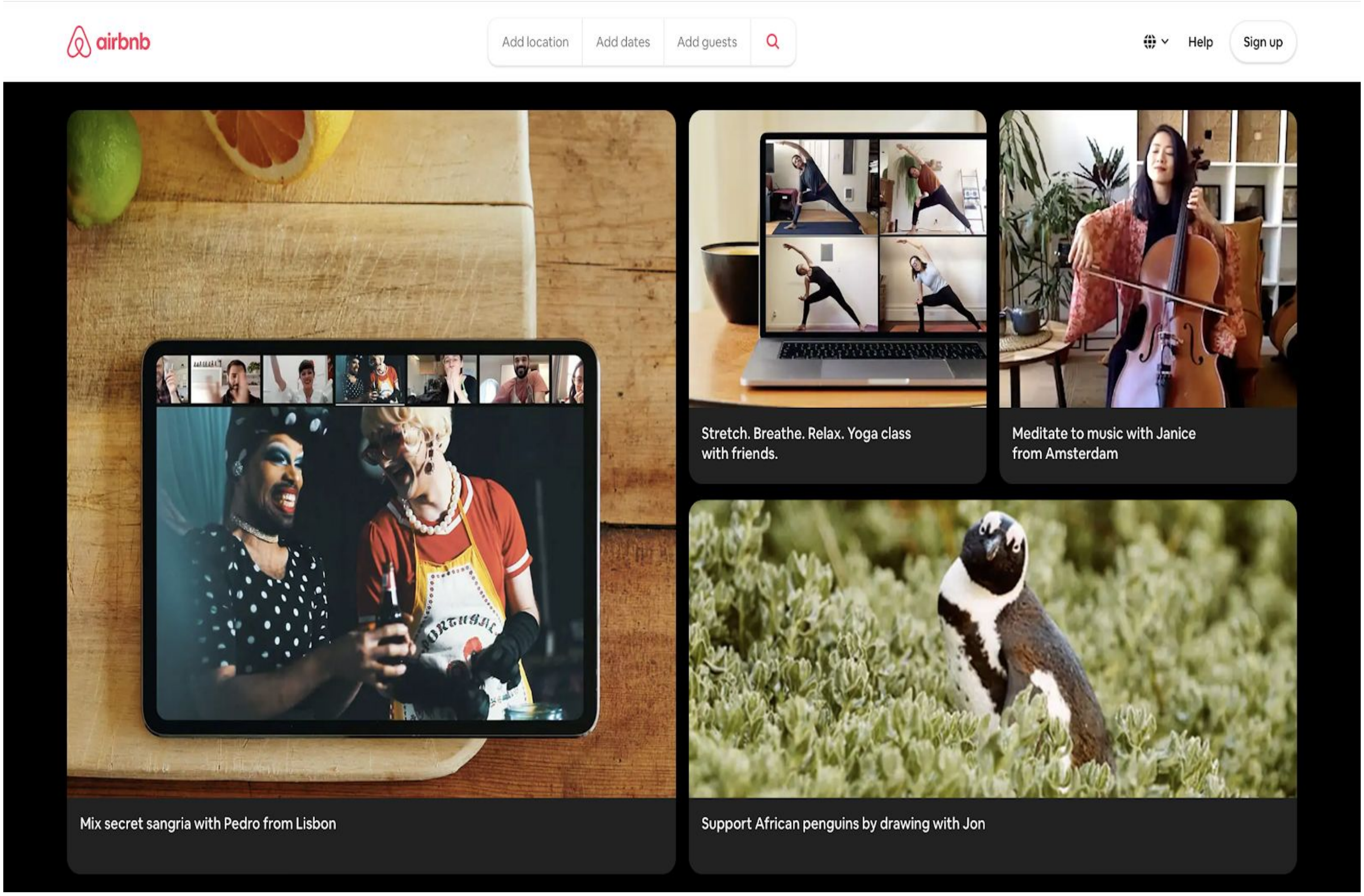
Airbnb Price Predictor


airbnb
Add location
Add dates
Add guests
Help
Sign up
Stretch. Breathe. Relax. Yoga class with friends.
Meditate to music with Janice from Amsterdam
Mix secret sangria with Pedro from Lisbon
Support African penguins by drawing with Jon

Introduction

"Airbnb happened because two guys could not pay their rent, but did have some space." That's how Air-bed and breakfast(Airbnb) started when Joe and Martin from San Francisco rented out their three airbeds to make some cash. Since then, guests and hosts have used Airbnb to travel in a more unique and personalized way. The story of Airbnb's foundation is fascinating.

Airbnb is a hospitality service which enables people to lease or rent short-term accommodation. The company does not own any lodging. It is merely an agent. It receives a percentage service fee (commission) from both guests and hosts for every booking.

A Host can create a listing by selecting the "Host" menu after logging in. Price for the listing is decided by the host. Hosts can charge different prices for nightly, weekly, and monthly stays. Host then adds a description of the residence, amenities, available dates, cancellation policies, and any house rules.

Potential guests are required to message the host directly through Airbnb to ask questions regarding the property. After the reservation, the hosts coordinate meeting times and contact information with the guests. Airbnb guests can leave a review and rate a listing after their stay. This rating can be used as an indicator of their experience during the stay.

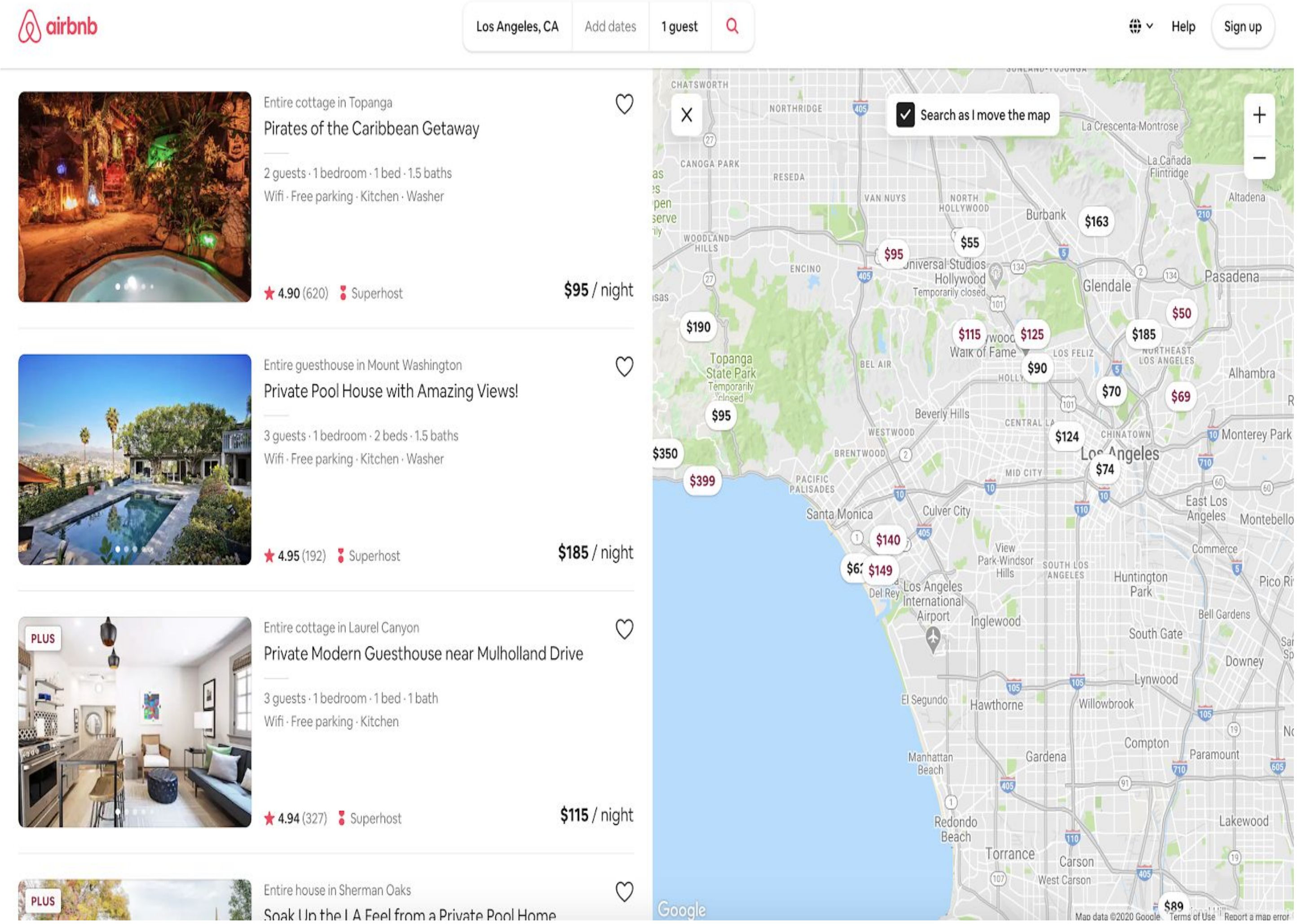
What is the problem I want to solve?

Although Airbnb and other sites provide some general guidance, there are currently no free and accurate services which help hosts price their properties using a wide range of data points. Also there is no way for travellers to figure out the correct price range based on the features they are looking for.

Paid third party pricing software is available, but generally hosts are required to put in their own expected average nightly price ('base price'), and the algorithm will vary the daily price around that base price on each day depending on the day of the week, seasonality, how far away the date is, and some other factors.

It is very important to get the listings' price correctly, particularly in big cities like New York, San Francisco, Los Angeles, Boston where competition is so high and some small differences in prices can make a big difference.

This project aims to solve this problem, by using machine learning to predict the base price for properties in Los Angeles.



Entire cottage in Topanga
Pirates of the Caribbean Getaway
2 guests · 1 bedroom · 1 bed · 1.5 baths
Wifi · Free parking · Kitchen · Washer
★ 4.90 (620) Superhost
$95 / night

Entire guesthouse in Mount Washington
Private Pool House with Amazing Views!
3 guests · 1 bedroom · 2 beds · 1.5 baths
Wifi · Free parking · Kitchen · Washer
★ 4.95 (192) Superhost
$185 / night

Entire cottage in Laurel Canyon
Private Modern Guesthouse near Mulholland Drive
3 guests · 1 bedroom · 1 bed · 1 bath
Wifi · Free parking · Kitchen
★ 4.94 (327) Superhost
$115 / night

Entire house in Sherman Oaks
Soak Up the LA Feel from a Private Pool Home

Clients

I have two types of client clients, it can be used for both hosts and airbnb for better pricing optimization & maximize the revenue.

1. Airbnb hosts

If the listing price is high, then chances of listings getting booked will go down. Also if the price is low, then the host will be missing out on a lot of potential income. This project will help Airbnb hosts to price their listings competitively and thereby increase the bookings. The prices predicted by this model can also be used by existing hosts to increase bookings which will result in more earnings.

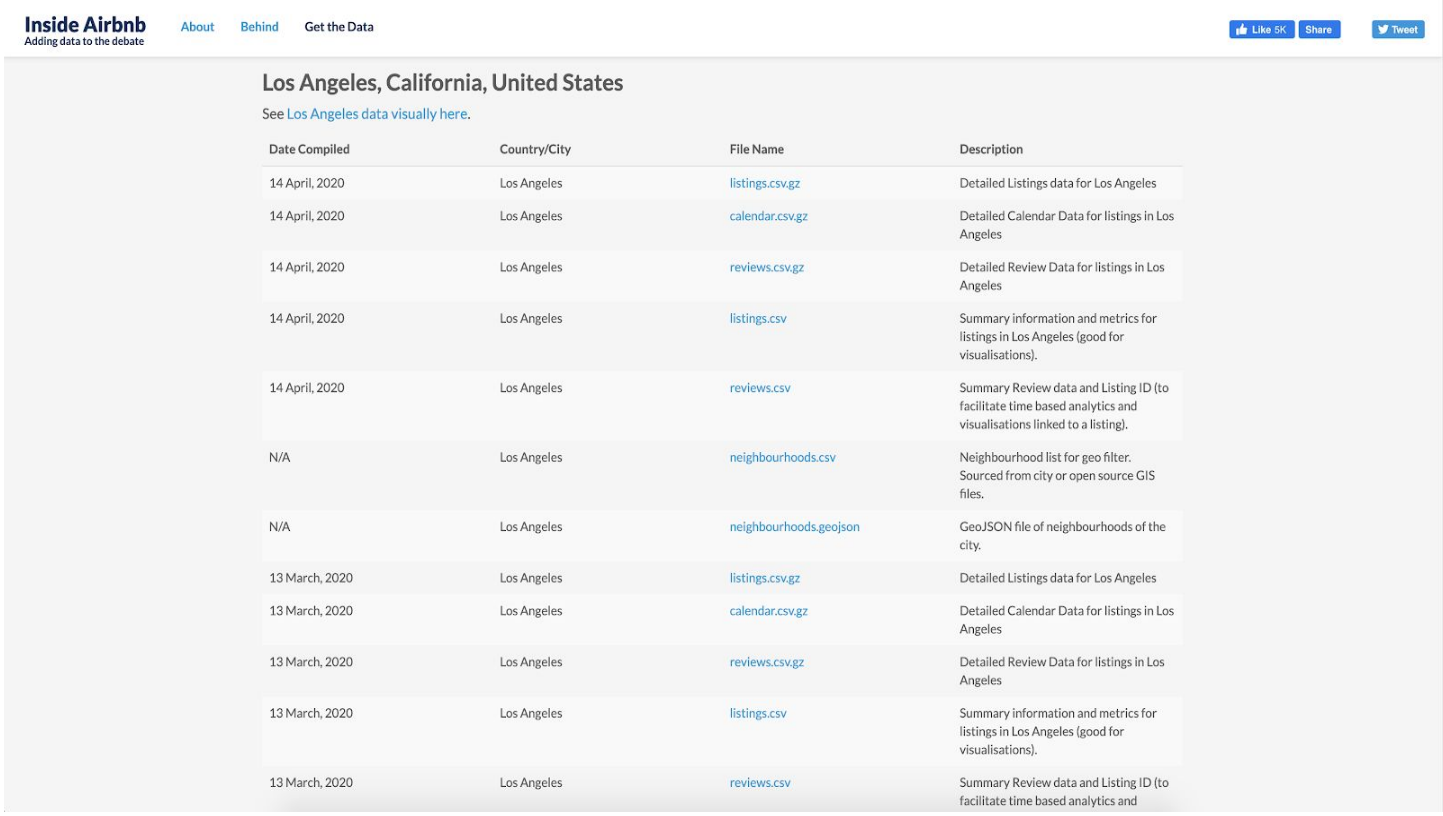
2. Guests looking to book Airbnb Listings

When guests are on Airbnb they do not have any means of knowing if the price that the hosts are asking for is reasonable for given features or not. This project will help guests to find real value for their money in the Airbnb marketplace.

Dataset

The dataset I am using is available on Inside Airbnb for Los Angeles location. "Inside Airbnb" is an independent, non-commercial set of tools and data that allows us to explore how Airbnb is really being used in cities around the world.

As per the information available on Inside Airbnb the dataset was last scraped on July 8, 2019 and contains information of all 44,504 Los Angeles Airbnb Listings on Airbnb site on July 8, 2019.



Los Angeles, California, United States

See Los Angeles data visually here.

Date Compiled	Country/City	File Name	Description
14 April, 2020	Los Angeles	listings.csv.gz	Detailed Listings data for Los Angeles
14 April, 2020	Los Angeles	calendar.csv.gz	Detailed Calendar Data for listings in Los Angeles
14 April, 2020	Los Angeles	reviews.csv.gz	Detailed Review Data for listings in Los Angeles
14 April, 2020	Los Angeles	listings.csv	Summary information and metrics for listings in Los Angeles (good for visualisations).
14 April, 2020	Los Angeles	reviews.csv	Summary Review data and Listing ID (to facilitate time based analytics and visualisations linked to a listing).
N/A	Los Angeles	neighbourhoods.csv	Neighbourhood list for geo filter. Sourced from city or open source GIS files.
N/A	Los Angeles	neighbourhoods.geojson	GeoJSON file of neighbourhoods of the city.
13 March, 2020	Los Angeles	listings.csv.gz	Detailed Listings data for Los Angeles
13 March, 2020	Los Angeles	calendar.csv.gz	Detailed Calendar Data for listings in Los Angeles
13 March, 2020	Los Angeles	reviews.csv.gz	Detailed Review Data for listings in Los Angeles
13 March, 2020	Los Angeles	listings.csv	Summary information and metrics for listings in Los Angeles (good for visualisations).
13 March, 2020	Los Angeles	reviews.csv	Summary Review data and Listing ID (to facilitate time based analytics and

Data Wrangling

I started this project because as of today Airbnb doesn't have a system where a host can go and find out what is the competitive price they can charge for their listing on Airbnb and at the same time guests don't know if they are getting real value for their money or not.

I took Airbnb LA data and I am working towards predicting Airbnb listing prices based on various conditions. This model will help hosts and guests both. First step after loading the data is Data Wrangling. It involves looking at the data, removing outliers, taking care of missing data and null values, converting data into a format that can be analyzed and studied.

The original dataset contained 44620 Airbnb listings and 106 features but I dropped many of those.

1. Some of the features are free text variables, like the host description of the property and all the written reviews. To include that data, I would have to perform Natural Language processing operations and at present it is out of the scope of this project. So those features were dropped.
2. Several features only contained one category, so they were dropped too.
3. Most listings don't offer the experiences feature, so it was also dropped.
4. Some columns contained mostly 'None' or 'NaN' values and therefore were dropped.
5. There were some features which were duplicates, so those were also dropped.
6. Data in money columns was stored as string, so I fixed those data types and converted those to floats.
7. Categorical data was converted to numeric type.
8. Biggest challenge was the amenities column. It had a lot of useful information but extracting that info was not easy. I used various python functions to extract the data from this column and then find out the most important items and then create separate columns for these items. Data in these columns was filled as amenities present or absent.
9. I did draw various plots to see the relationships between various features and price of the listing. I also draw graphs to visualize price distribution for these listings.

10. After performing all these operations I saved the clean data into a separate .csv file. This file will be loaded in the next phase (EDA phase) and data will be studied further.

Exploratory Data Analysis

During exploratory data analysis, following questions were asked:

1. What are the most important features affecting listing price?
2. Does location/neighborhood affect listing price?
3. If location/neighborhood affects listing price, what made these neighborhoods special? and ask follow-up questions if needed.

Initial Findings

1. As expected, listing price is correlated to accommodates, bedrooms, bathroom, neighborhood, and property type.
    - Interestingly, some neighborhoods (like Hollywood Hills West, Beverly Hills, Bel-Air, Hollywood Hills) seem to have higher listing prices.
    - However there might be some exceptions:
        - There might be extremely high listing prices although there's only one bedroom, but rarely happened.
        - If the property has a lot more bathrooms, ex. 8, but it's a hostel, the listing price would not increase proportionally.
        - If the property has a lot more bathrooms, ex. 8, but can only accommodate 2 people, the listing price would not increase proportionally, either.

        - If the property is a villa which can accommodate 14 or more people but its accommodates vs. bedrooms ratio and/or accommodates vs. bathrooms ratio are higher, that means more travelers need to share the bedrooms / bathrooms, the listing price would not increase proportionally, either.

2. Property-type-wise, most of the property type of LA listings are house and apartment.
3. In terms of bedrooms and bathrooms, 1 bedroom 1 bathroom Airbnbs which can accommodate 1 - 4 people are most common in LA.
    - About 52% of LA Airbnb listings are of this type.
    - For the rest of Airbnbs, not many of them can accommodate more than 10 people. In fact, only 2.75% of listings can accommodate more than 10 people.

Exploratory Data Analysis

Correlation between an independent and a dependent variable

- accommodates, bedrooms, beds, bathrooms do correlate with price.
- property_type & neighbourhood_cleansed are also correlated with price.

Correlation between pairs of independent variables

- For numerical variables, I checked correlation by pearson correlation coefficient.
- For categorical variable vs. numerical data, I used a logistic test.

Modeling

With insights based on exploratory data analysis (EDA), I started to train predictive models.

I checked the distribution of the target, which is price & confirmed that taking log (using logPrice column) can make it distribute more normally & skew is much improved.

I identified five models to try first:

1. **K-Nearest Neighbors**
    - Predict the label of a data point by
        - Looking at the 'k' closest labeled data points
        - Taking a majority vote
2. **Linear Regression**
    - Predict the label of a data point by a linear function
    - Loss function = Ordinary least squares (OLS), which is sum of squares of residuals
3. **Ridge Regression (L2 regularization)**
    - Same as Linear Regression but penalize large coefficients by L2 regularization:

$$Loss\ Function = OLS\ Loss\ Function + \alpha * \sum_{i=1}^{n} {a_i}^2$$

4. **Lasso Regression (L1 regularization)**
    - Same as Linear Regression but penalize large coefficients by L1 regularization:

$$Loss\ Function = OLS\ Loss\ Function + \alpha * \sum_{i=1}^{n} |a_i|$$

5. Random Forest

   - Predict the label of a data point by ensembling decision trees, which correct for decision trees' habit of overfitting.

And train the above models with three features:

1. Accommodates
2. Bathrooms
3. Bedrooms

Test size of Train-test-split is set to 33%. We chose Root-Mean-Squared-Error (RMSE) as our scoring metric.

Summary

1. I tried 5 different models on Airbnb listing price prediction
   - K-Nearest Neighbors
   - Linear Regression
   - Ridge Regression
   - Lasso Regression
   - Random Forest
2. KNN & random forest outperforms linear regression models. After adding more features, random forest performs better than KNN.
   - If looking at KNN results, after adding more features, the best parameter for n_neighbors becomes less. It might be because KNN depends on similar neighbor data points to get better prediction results. When adding more features, it increased dimensionality. With the curse of dimensionality.

    Number of similar data points seems to be becoming less, so RMSE started to get higher.

    - On the other hand, random forests by nature automatically select useful features when splitting so did not have this issue.

3. Linear regression models did not perform well because there are less data points with price above 500 and looks like they have a different linear relationship. For this piecewise linear regression is sometimes used.

Future

1. Apply NLP on Airbnb reviews for better listing price prediction.
2. Apply models on other cities or training models on other cities.
3. Apply piecewise linear regression.
4. Apply ensemble models like Gradient Boosting