AirBnB Price Prediction in Los Angeles, CA

Springboard Capstone 1
Alexandra Michel
May, 2020

Background + Motivation

What is AirBnB?

- AirBnb is a website that allows people to search and reserve short-term and apartment rentals, as well as become a host on the platform by opening up their own space to potential guests.
- Prices are set by the hosts, and guests/renters can narrow their search by different criteria
- Important for hosts to know what the market price is for their space

AirBnB could be interested in enhancing their price prediction model as a way to generate more revenue through their hosts.

Problem Statement

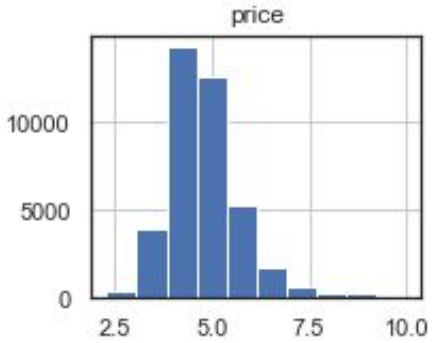
- Since hosts set their own nightly rate, it is important to get that price right in order to attract guests.
- Leave less money on the table by pricing according to the current market in the area
- Provide insights to hosts to help maximize their profit on renting out their space

The Dataset

- Using data from www.insideairbnb.com to obtain information about listings in Los Angeles, CA
- Raw dataset has 160 features and ~38,000 rows
- Each row is an available listing
- Features related to listing criteria (bedrooms, bathrooms, beds, number of guests), as well as reviews, location, amenities, min/max number of nights.
- The dataframe had many free-text features which I dropped

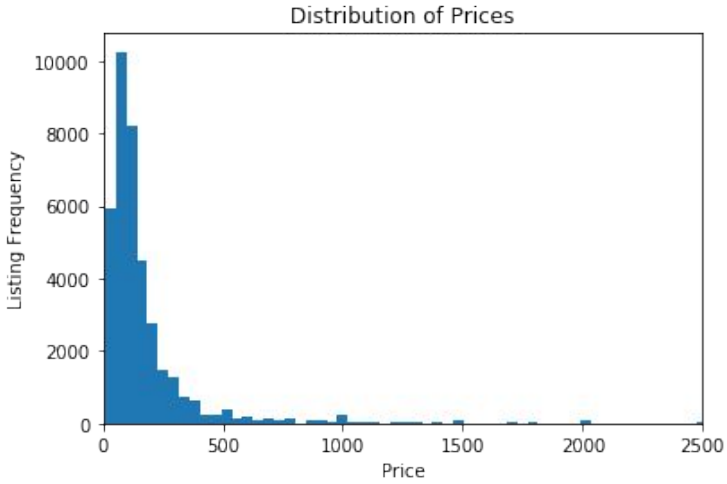
Cleaning the Dataset

- Dollar-value columns price, security_deposit, cleaning_fee and the column extra_people needed to be changed from string to int
- sparse categories within property_type which I grouped together into House, Apartment, and Other
- Dropped columns with high collinearity
- Log-transformed numerical columns



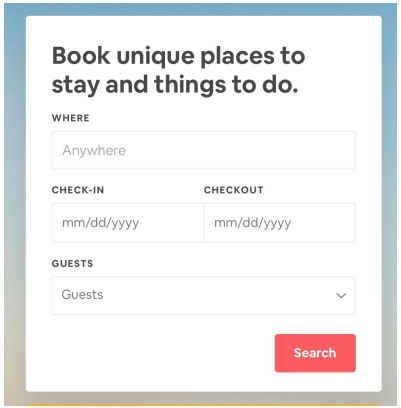
Exploratory Data Analysis

- 94.2% of the listings fall under $500/night
- Median price is $110.0/night
- Mean price is $226.88/night
- Prices range all the way up to $22,000/night
- The amount of listings above $2,500/night is less than 1% of the dataset



Exploratory Data Analysis

- Users input # of guests as first criteria aside from location
  → Should be important feature
- As # of guests increases, price also increases
- Outliers for luxury real estate, accommodating 18+ people

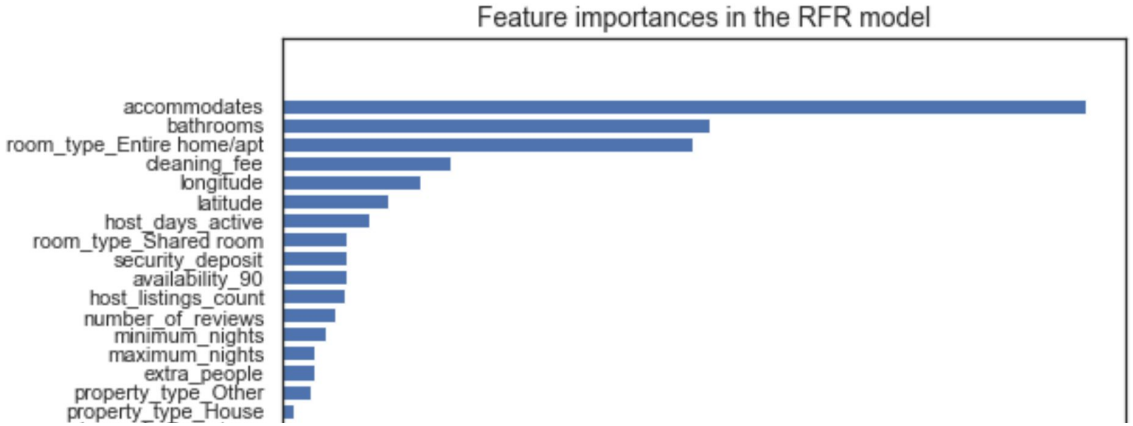


RandomForestRegressor Model

- Split data into 80% train / 20% test sets
- Returned 80.7% R^2 score on test data
- Found that neighborhood names didn't have hardly any importance, and when dropped, reduced complexity and increased to 80.87% R^2 score.

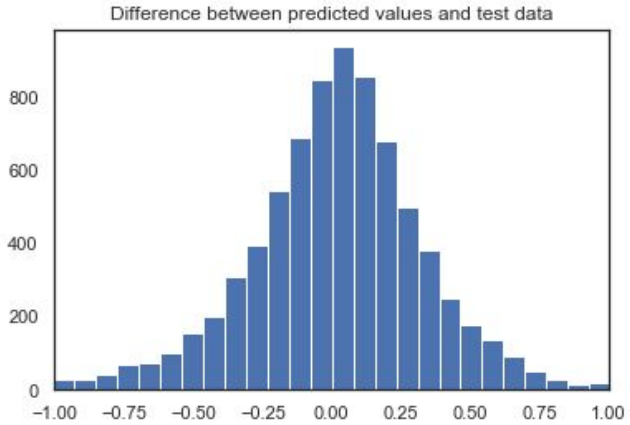
Feature Importance

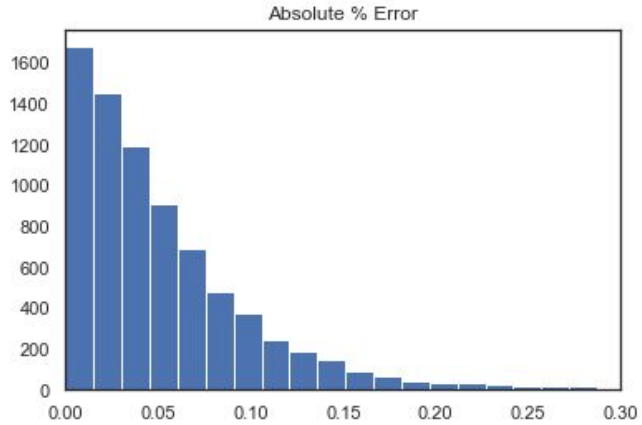
- Over 70% of the model is based on the ‘accommodates’, ‘bathrooms’, ‘room_type’ and ‘latitude’/‘longitude’ attributes.



Further Insights

- I was also interested to see how many predictions fell within 5% of the actual price, within 10%, and above 15%. Only 5% of the test data falls into more than 15% error.





Total # test data	Within 5% error	Within 10% error	More than 15% error
7770	4595	6663	414

Recommendations

- Maximizing guests optimizes revenue, and is important in determining market price for a listing.
- Develop a product for hosts to calculate ROI on home improvement projects looking to maximize accommodated guests.
- We have created a simple model with areas to expand on. It is recommended to look into text sentiment analysis for understanding guest reviews, and find patterns in the listings that are harder to predict.