Predicting market rent prices of rental properties from Craigslist in the Bay Area

Melissa Han

November 2022

Table of Contents

1 The problem statement

For a given property, what is it’s market rent?

Goal of this capstone project: develop supervised machine learning models to predict the market rent price of a property based on relevant features found on a craigslist ad.

2 About the data

- Craigslist was chosen because it is one of the most common places to find a place to rent in the Bay Area.
- There is no craigslist API, listing information was scraped from Craigslist every 3 days over the period Aug - Oct 2022. The summer months are the most popular time to rent.
- Beautiful Soup and Playwright to scrape the search result index and listing pages
- HTML parsed to a dataframe with REGEX

3 Data Cleaning and Wrangling

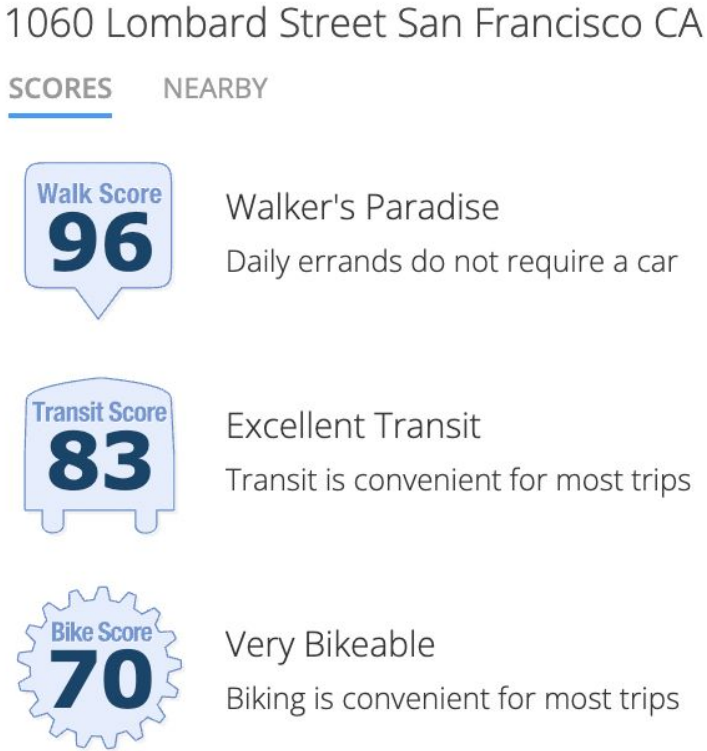
1. **Removing duplicate listings**

2. **Correct data type and filter to meet project assumptions:**
   - The property for rent is an entire property - that means not renting by room
   - The property is located in the Bay Area metros - San Francisco, Peninsula, East Bay
   - The rental price is quoted on a monthly basis

3 Data Cleaning and Wrangling

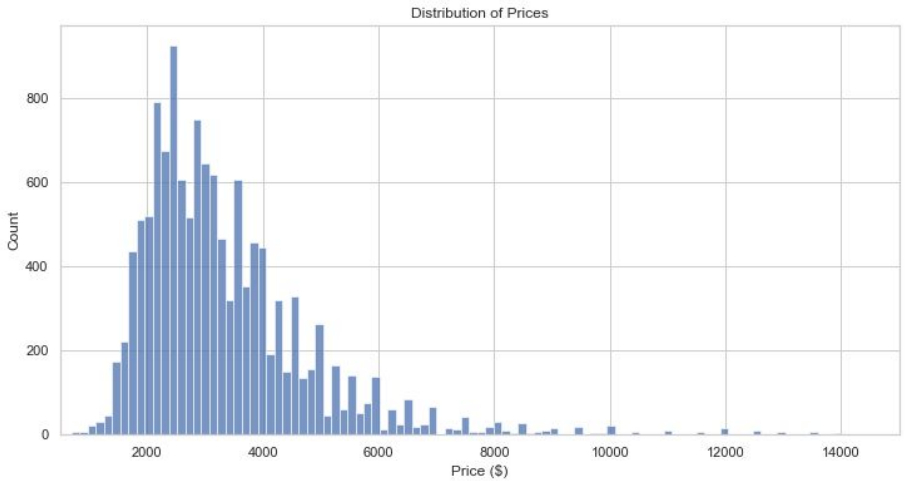
3. For each address, supplement the dataset with walk score, transit score, bike score from the walkscore.com website

After cleaning and filtering for duplicates, the data frame has 18338 rows.
The feature ‘address’ can be dropped from the dataset.

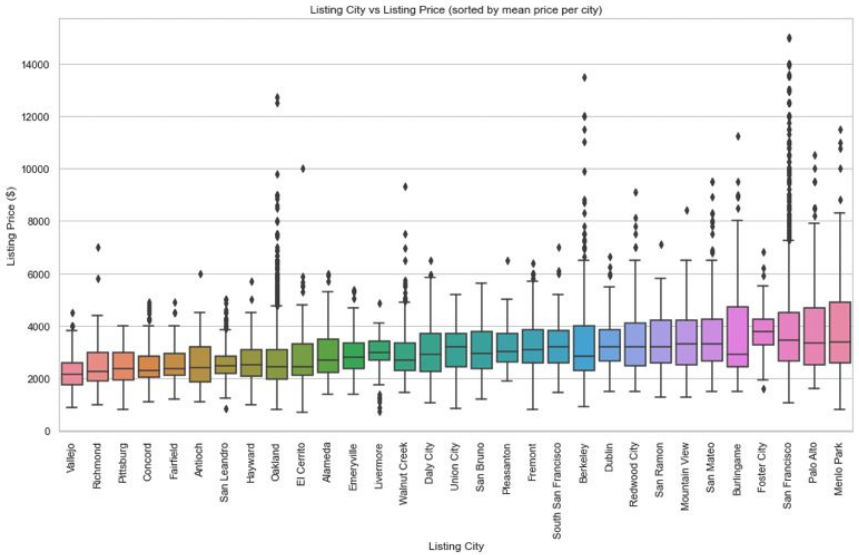


4 Data Visualization and Analysis

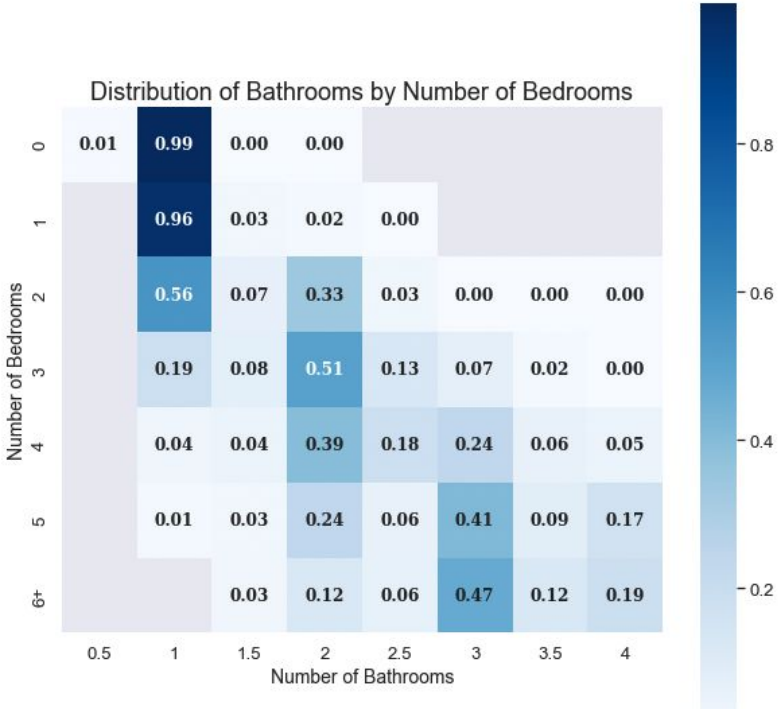
Look at the distribution of rent price



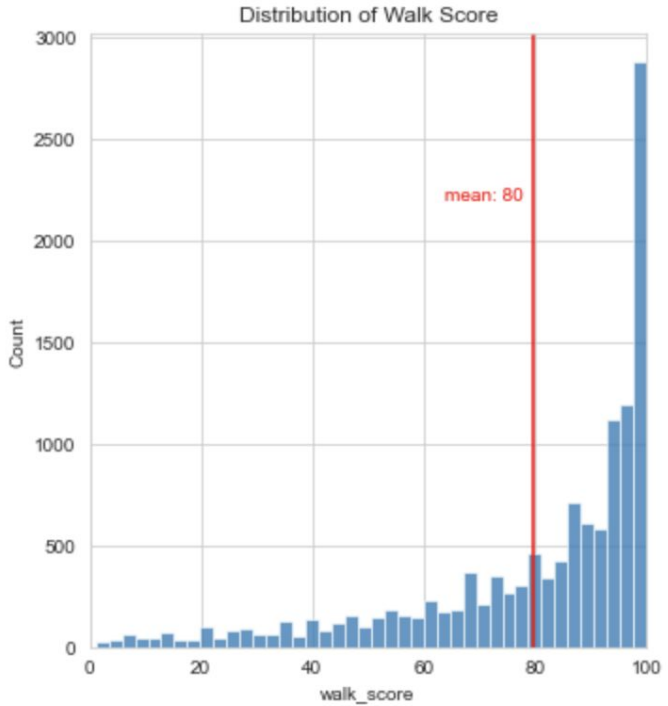
Distribution of prices for each listing city



Distribution of bedrooms and bathrooms



Distribution of walk scores



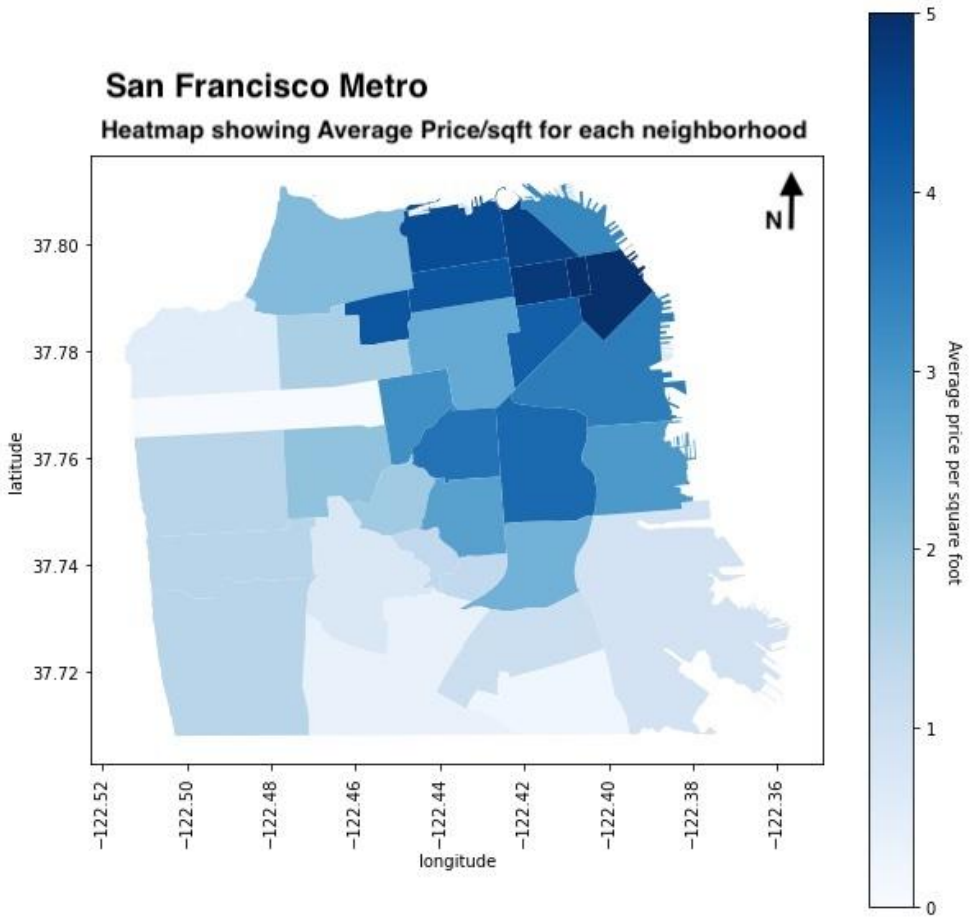
Correlation between numeric features

The numeric features that seem most correlated with price are:

- Listing_sqft
- Number of bedrooms
- Number of bathrooms
- Laundry in unit
- Parking in an attached garage
- A house, apartment or condo

There are also the categorical features neighborhood and city and how to handle them in the model.

Feature engineering: create a new feature: price per sqft for each neighborhood



5 Machine Learning Models

Train/test split is 70/30
Before preprocessing, the data frame has 13020 rows of listings.

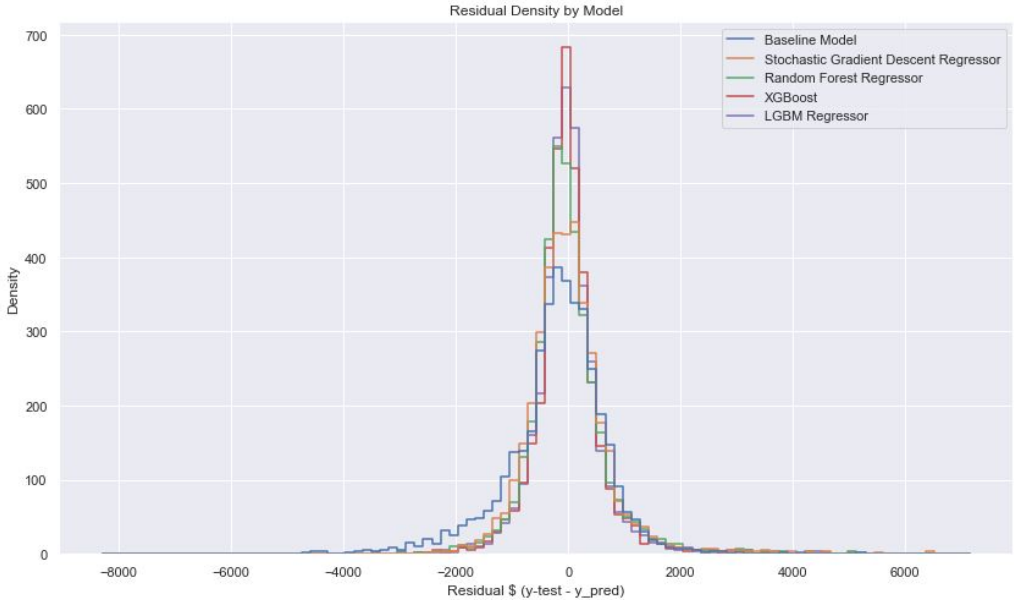
Many features are created in the preprocessing step after the train/test split to prevent data leakage:

- Replace neighborhood (81 in total) with the average price per square foot of that neighborhood
- Around 30% of listings did not provide a square footage value. Impute missing square footage values based on the median square footage grouped by number of bedrooms and number of bathrooms
- Remove price outliers for one bedroom places - these had the most number of mislabeled units and the biggest range of prices. Cap the top 1% percentile of prices for one bedroom listings.

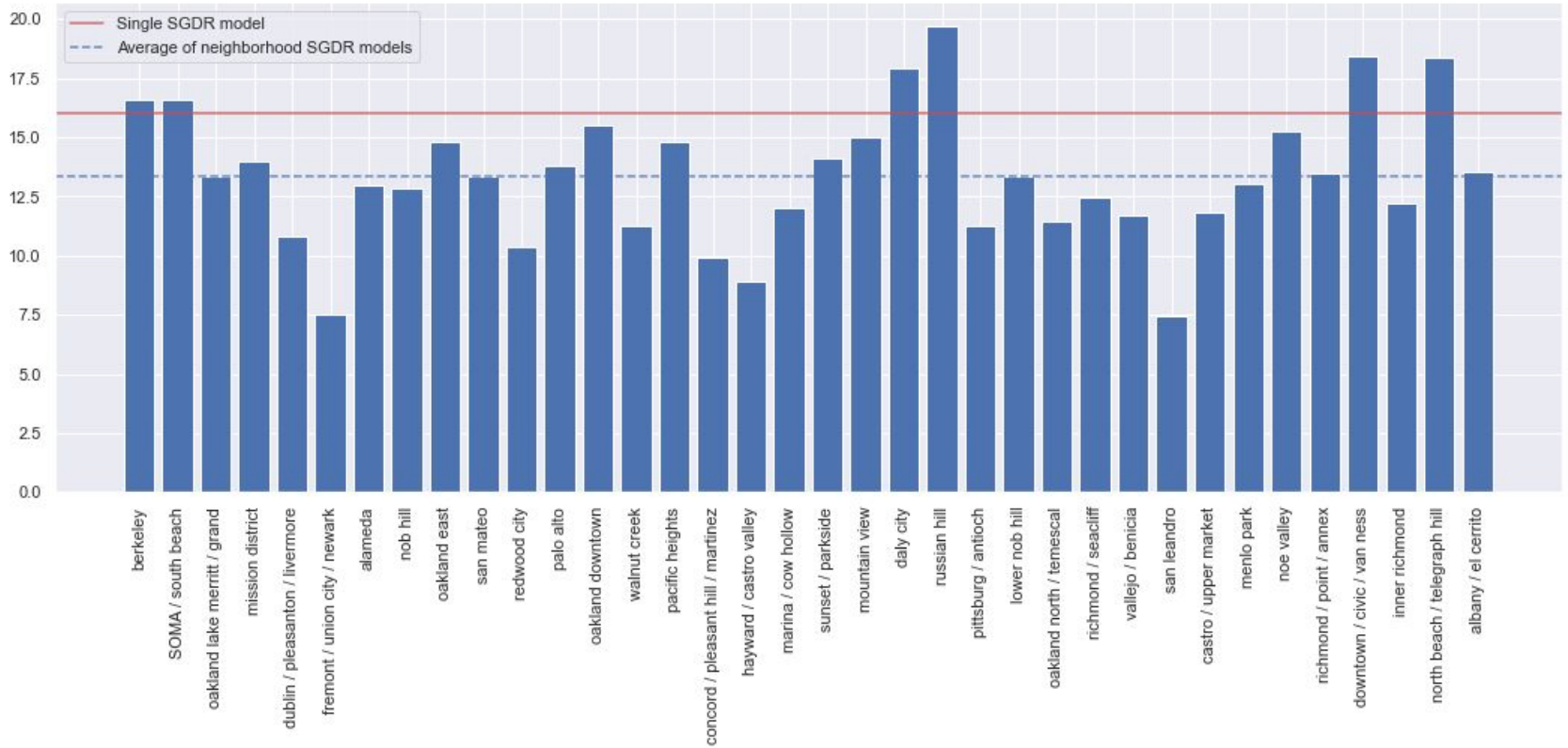
Performance Metrics: Mean Absolute Percentage Error (MAPE)

Tune hyperparameters using SKLearn's Gridsearch with 5 fold Cross validation.

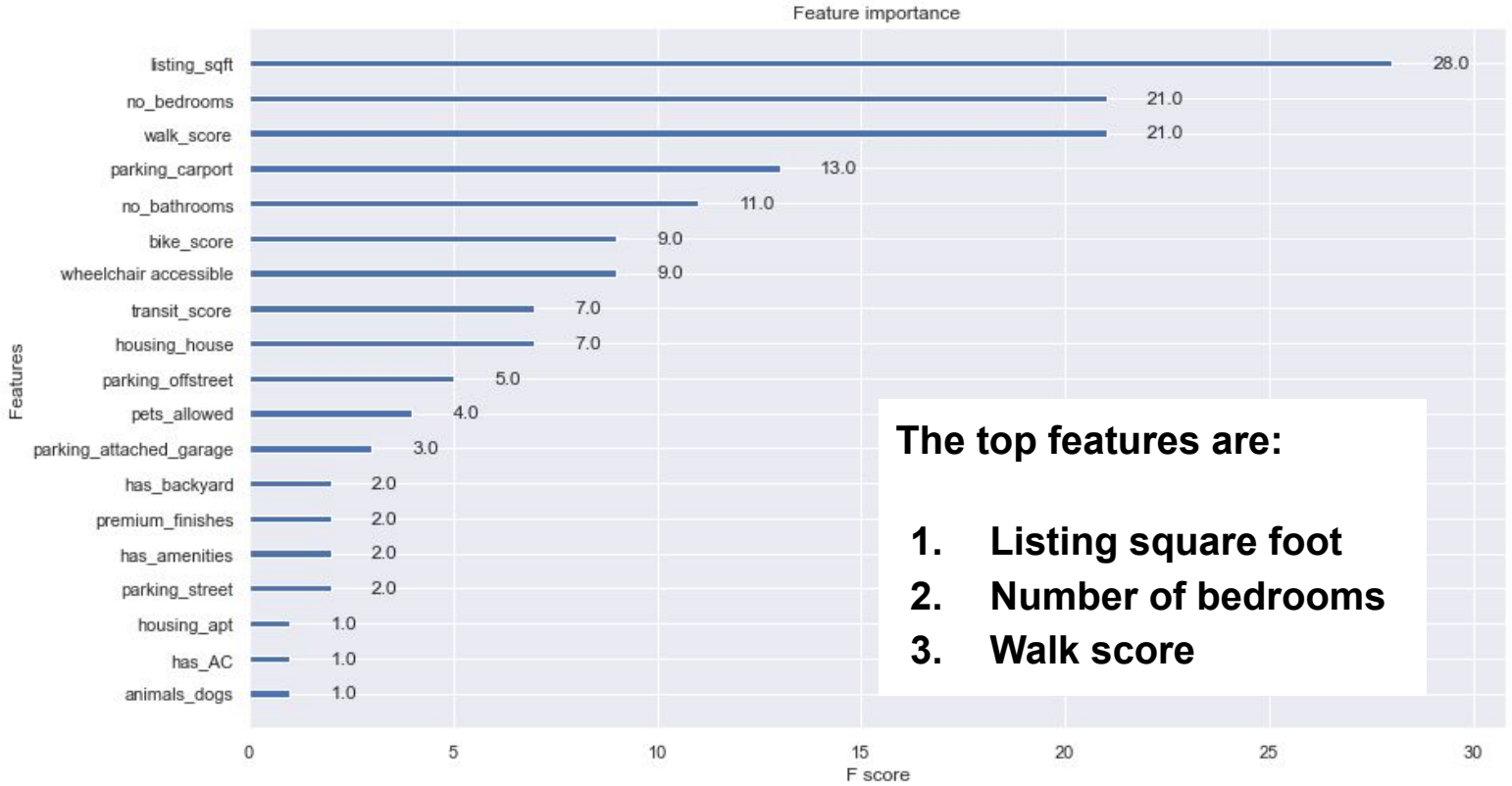
Machine Learning models:	**MAPE %**
1. Baseline Model	19.8%
2. Stochastic Gradient Descent Regressor	16.0%
3. Random Forest Regressor	14.6%
4. XGBoost Regressor	12.3%
5. LGBM Regressor	12.4%
6. Piecewise SGDR by number of bedrooms	14.1%
7. Piecewise SGDR by neighborhoods	13.3%



Piecewise SGDR by neighborhoods



Feature Importance - XGBoost Model

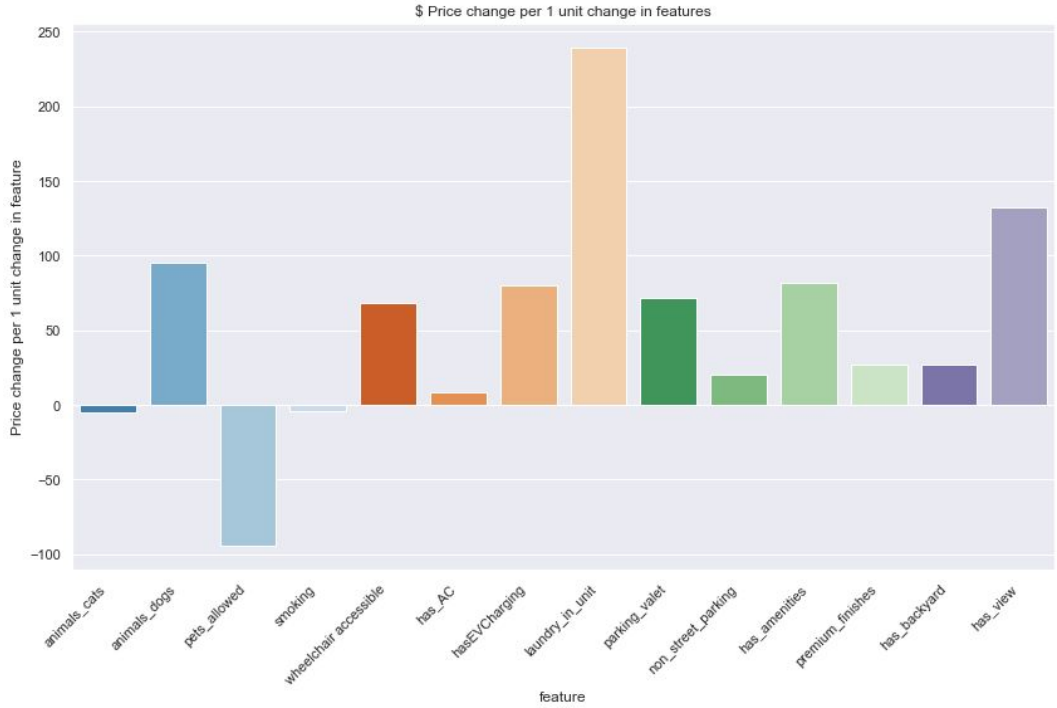


The top features are:

1. **Listing square foot**
2. **Number of bedrooms**
3. **Walk score**

Applications of the price prediction model

- Renters can optimize for below market rent properties and get a good deal on rent.
- Property owners can determine the market price for their property, which they can then adjust to increase revenue or increase competitiveness
- Cost Benefit analysis of craigslist listing features: knowing how much features are worth on a unit basis can help people make informed decisions.



Applications of the price prediction model

- Flag craigslist scams priced below market rent

- Counterfactual analysis

Using Bayes Optimization and work backwards to predict the combinations of features to change the listing price of a property by a chosen amount.

Thanks!

Github portfolio:
https://melissavhan.github.io/CapstoneTwoProject/