Capstone: Airbnb DC

•••

Sarah Scolnik
DSI-US-06

Problem Statement

Problem:
Build a model to predict prices for Airbnb listings based on various features and identify the most influential features, with the goal of identifying strategies for Airbnb hosts to maximize profit.

Data

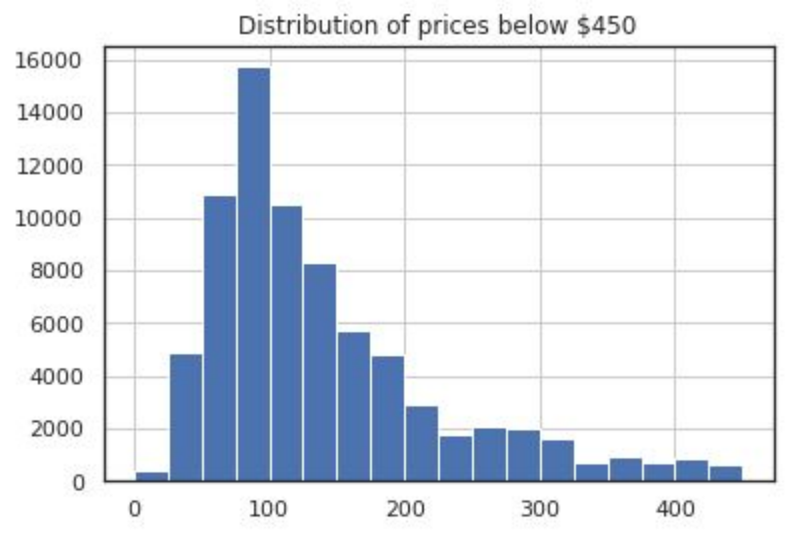
http://insideairbnb.com/get-the-data.html
Inside Airbnb scraped data on these dates from 2015 - 2018: 10/3/15, 3/10/17, 5/10/17, 4/15/18, 5/18/18, 7/20/18, 8/18/18, 9/14/18, 10/12/18, 11/15/18

Each scrape date has:

- listings.csv file: 1 record per listing, approx. 90 columns with data about the listing, including current price
- calendar.csv file: scraped from booking calendars. 365 records per listing with availability on each date for the next year, and price if available (if unavailable, price is null)

Target: Price

Price per night:
Median: $120
Mean: $217.67
Std dev: $364.99



Price Outliers

Low: shared room in townhouse in Fort Dupont, $10

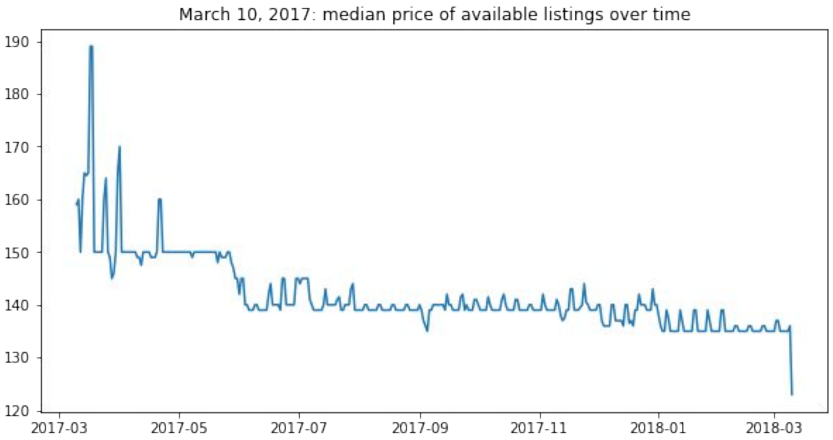
Price Outliers

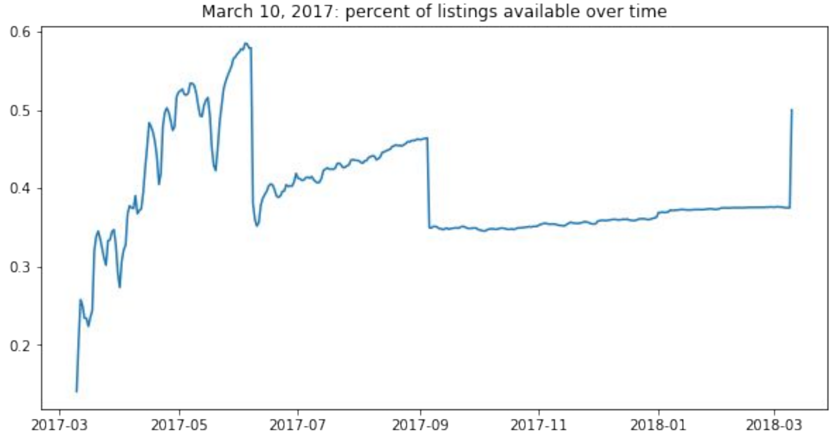
High: Historic Georgetown Residence, accommodates 8, 4 bedrooms, 5 beds, 6.5 baths

$10,000

Challenges: calendar price data

Missing data for unavailable days:
Approximately 2/3 of listings have null price for > 50% of days of the year





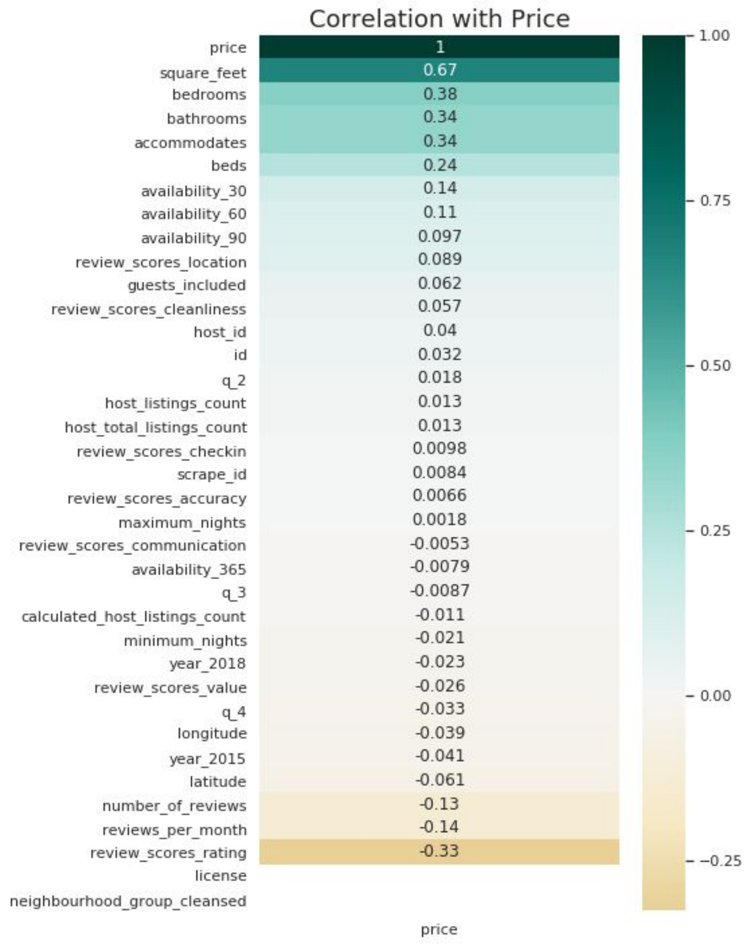
Methods & Models

Model: Linear regression with regularization (Lasso). Linear regression was chosen as main analysis tool for interpretability of results.

Feature Engineering: Data was cleaned, a subset of relevant features was selected, and interaction features were created.

Methods & Models

Correlation



Results

Baseline (predicting mean of target):

- baseline R^2: 0
- baseline RMSE: 363.71
- baseline MAE: 170.73

Better than the baseline!

Lasso model:

- train/test R^2: 0.47 / 0.48
- train/test RMSE: 266.69 / 261.00
- train/test MAE: 118.47 / 121.60 (vs mean of predicted prices: $216.92)

Results

Features/ coefficients:

Of 3322 features, lasso zeroed out coefficients for 1904 features.

Features with largest coefficients:

- accommodates (120.585354)
- accommodates * review_scores_rating (-114.632033)
- bathrooms * zipcode_20007 (84.229379)
- review_scores_rating * host_is_superhost_t (83.072543)
- host_is_superhost_t (-75.466774)

Conclusions & Recommendations

Features with highest positive coefficients in the linear regression models were indicators of size/number of guests. This result makes sense but is not very helpful for prospective Airbnb hosts looking for factors that could help them get a higher price.

Other features that influenced model included: review_scores_rating (negative), and neighborhoods:

highest prices: Downtown, Capitol Hill, Shaw, Union Station, Southwest

lowest prices: Ivy City, Historic Anacostia, Fort Totten

Future Improvements

- NLP on listing descriptions and reviews
- Add geographic features (distance to Mall, Metro, etc)
- Analysis of photo quality
- Add any additional features to model that may be useful for hosts to estimate effect of changes
- Time series modeling for calendar price data: impute missing values, or set up a daily scrape

Questions?