Exploration and Prediction of Airbnb Housing Prices in London

Xiaozhou Yu, Xiao Wan, Xiaoyu Zhang
Instructor: Yifan Hu

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics
April 30th, 2015

Introduction

- Airbnb is an online marketplace and homestay network that enables people to list or rent short-term house, with the cost of such accommodation set by the house owner. The company receives percentage service fees from both guests and hosts in conjunction with every booking.
- Airbnb was founded in August 2008. Now It has over 2,000,000 listings in 34,000 cities and 191 countries. London, the capital and most popular city of Britain, has more than 42,000 listings in 2016.
- Our goal is to predict housing prices for new listings. Using this model we can help travelers choose best-fitting house within budget and help house owners make a better price. From the perspective of Airbnb, the result can help them detect fraudulent listings.

Experiment

Data preparation

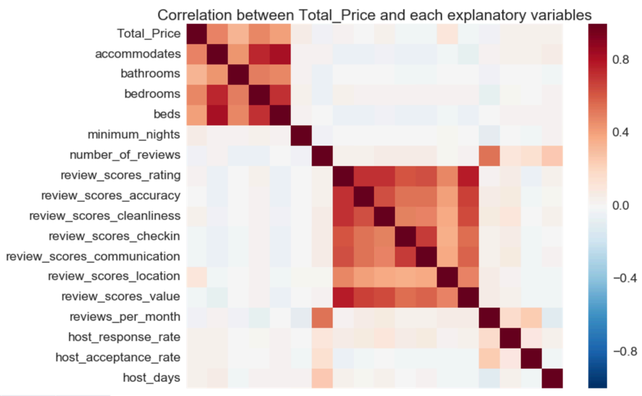
- Data resource: Inside Airbnb provides scraped data from publicly available information from the Airbnb site. http://insideairbnb.com/get-the-data.html
- This dataset contains 42,646 records and 94 features.
- After data cleaning, our new dataset has 41772 records and 74 features.
- Training data 80%, testing data 20%

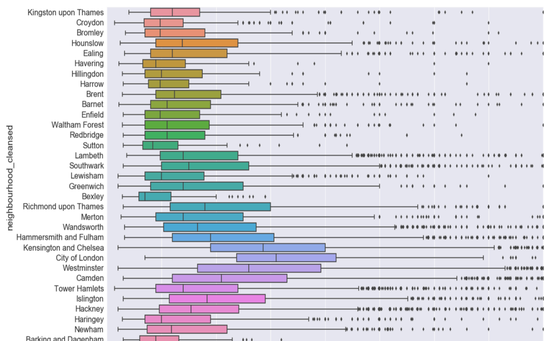
Exploratory Analysis

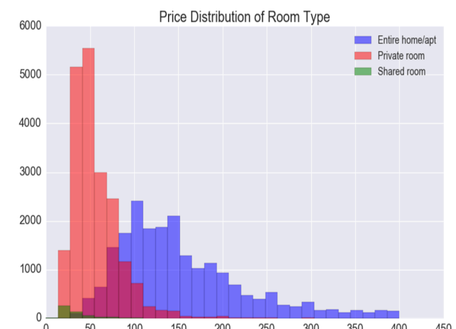
Highly relevant features:

neighborhood, room_type

- Prices differs a lot due to the location and room type, so add dummy variables for each categorical feature







Model Evaluation Metrics:

- To avoid influence of different price level, use relative error instead of absolute error
- RMSPE -- Root Mean Square Percentage Error
- $$RMSPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \widehat{y_i}}{y_i}\right)^2}$$

Models & Methodology

Baseline Model:

- Use only 2 features: neighborhood, room_type

Linear Model:

- Apply simple linear regression on all features.

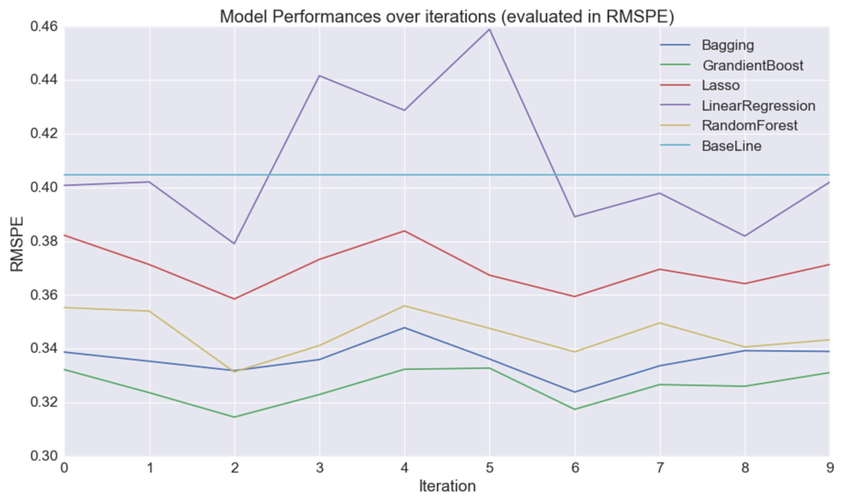
Regularized Linear Model:

- Feature selection method: LASSO
- Tuning parameter: Lowest RMSPE ←→ penalty = 0.7

Ensemble methods:

- Bagging, Random Forest, Gradient Boosting

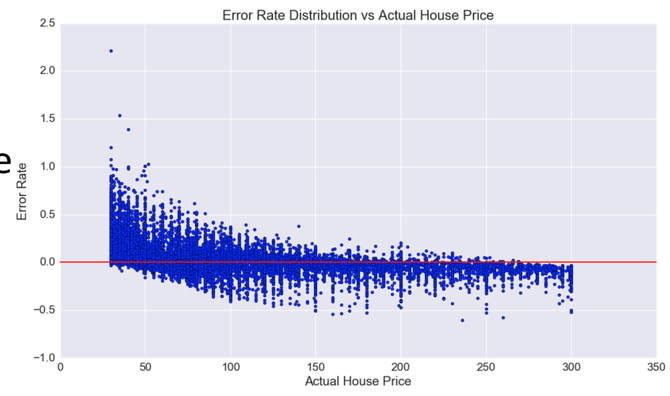
Choose Gradient Boosting model for further analysis

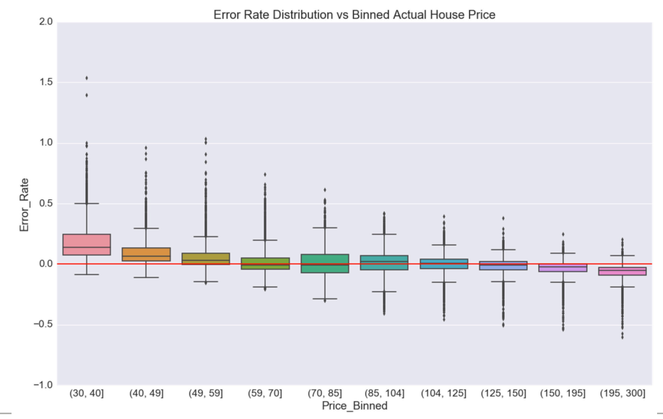


Evaluation & Deviation Analysis

- Analyze the distribution of error vs price
- Analyze the distribution of price and decile the data
- Use error_rate to describe how prediction deviates form actual price

$$Error\ Rate = \frac{price\_predicted}{price\_actual} - 1$$





- Over-predict at low price level;
- Under-predict at high price level.

Next Step

- Parameter tuning on ensemble models
- Analyze the distribution of deviation of prediction for further exploration
- To improve model performance at both ends of price range: consider building different models on different price level