Exploring and Predicting Airbnb Price in NYC

Jiahui Bi, Ruoqi wang, Yicong Ma, Xin Chen
Instructor: Yifan Hu

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics
November 13th 30th, 2017

Introduction

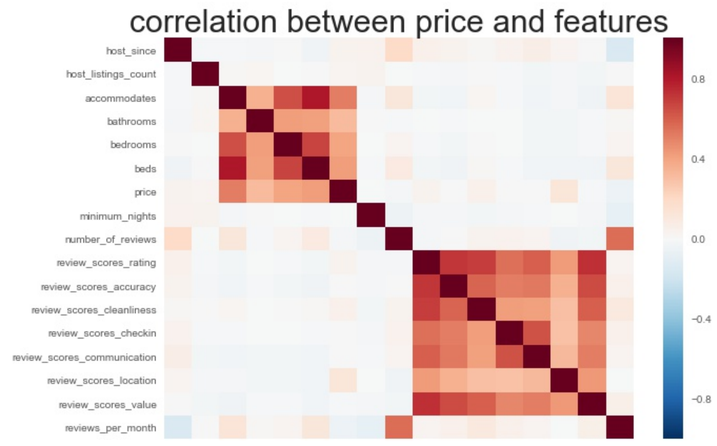
- Background: After showing up in the 2013, the sharing economy has been a more and more important role in everyone's life. Airbnb, as one of the representatives of sharing economy, is always focused by explorers and users.
- Purpose: exploring what factors will influence the price of the Airbnb houses/ apartments / rooms and using those factors to predict the price in New York City.
- Key Questions: What factors will influence the price in the New York and how will they influence?
- Technology: Python(sklearn, xgboost)

Data Preparation

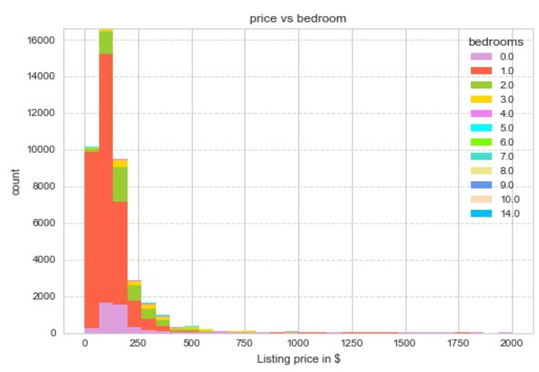
- Data Source : http://insideairbnb.com/get-the-data.html
  New York City, New York, United States
  October 2017
- Data cleaning : We get 44317 rows data which include 96 features
  - Drop entries that are missing (NaN) values for columns like "bedroom", "bed"
  - Substitute missing values with the mean of the columns like "review_scores_location"
  - Set the value of 'reviews_per_month' to 0 where there is currently a NaN
  - Transfer the string values into the integer or float type
  - Delete entries that are outliers (more than 2000$ or 0) for "price"
- Finally, we get 43148 rows data and 28 features

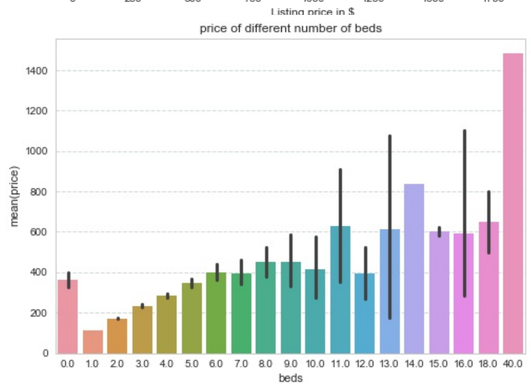
Exploratory Data Analysis

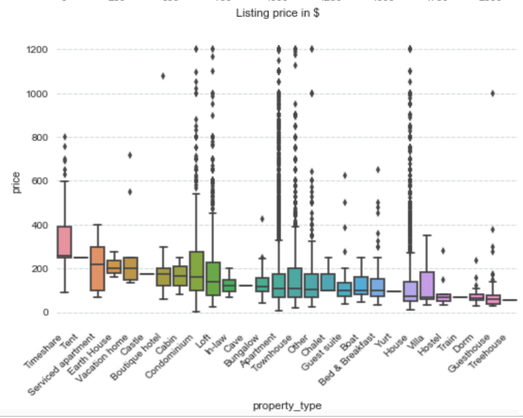
- Highly relevant features



- Price Distribution







Method 1

- Built Vanilla Linear Regression, Ridge regression, Lasso Regression, Bayesian Ridge on cross-validation split data
- Figure 1 shows when used 18 features
  Figure2 shows after One-Hot Encoding (create dummy variables)

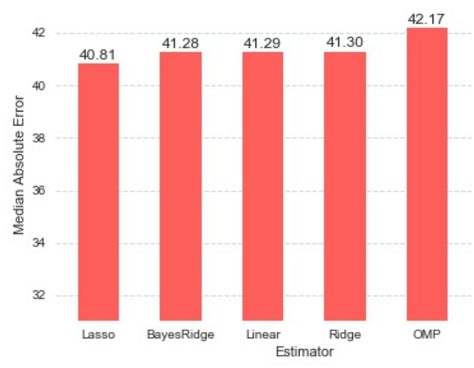


Figure 1

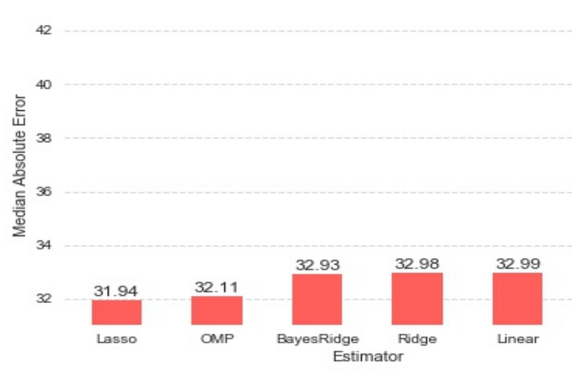
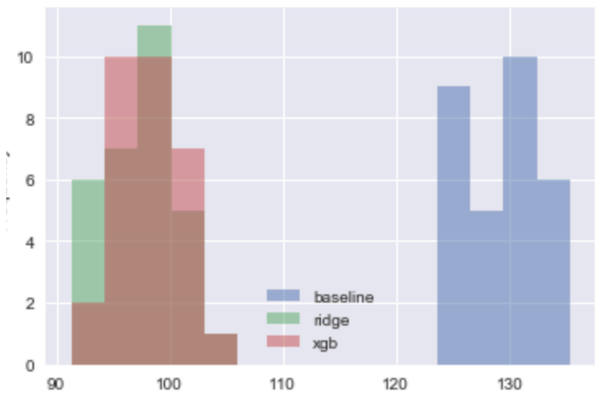


Figure 2

- So, we can see Lasso comes out on top
- Ensemble model: Gradient Boosting
- We're doing better with Gradient Boosting Regressor with MAE=25.52, almost 20% less than previous method

Method 2

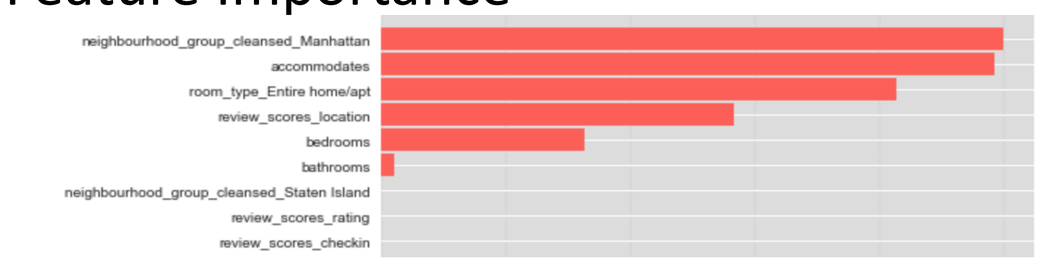
- Ridge, Baseline & Xgboost
- To test our different models in depth, we repeated train-validation split and then see how our errors are distributed. We also added a baseline model to compare.



	Mean(RMSE)
baseline	129.69
ridge	97.24
xgboost	98.31

- Both ridge and xgboost beat the baseline model
- Ridge Regression is performing slightly better than xgboost model

Feature Importance



- The price is highly positively related to location, in our case, listings in Manhattan is highly correlated with the pricing.
- Accommodates, room type, the number of bathrooms, bedrooms, beds also have strong correlations with the price.

Conclusions

- Conducted 5 models and ensemble the model
- Tested xgboost model with baseline model and Ridge Regression
- The most important features that influence the prices in NYC are location, the type of room and the number of bedrooms

Future Work

- Abstract new features, such as amenities.