DS-GA-1003: Machine learning and Computational Statistics, Spring 2017

Prediction on NYC Airbnb Listing Price

NEW YORK UNIVERSITY

Siyang Sun, Xinsheng Zhang, Zemin Yu

Research Question

We would like to propose a model to predict appropriate prices for Airbnb listings in NYC. Particularly, we were interested in studying how these features would help Airbnb provide an internal tool to assist hosts in determining reasonable prices.

Data

The Inside Airbnb project included detailed information about all listings available in New York City as of December 3, 2016. Each listing includes:

- listing: `name`, `description`, `bedrooms`, etc.
- host: `host_about`, `host_verification`, etc.
- location: `latitude`, `longitude`, `zipcode`, etc.
- availability: `minimum_nights`, `availability_30`, etc.
- review: `review_score`, `reviews_per_month`, etc.
- price: `price`, `cleaning_fee`, etc.

Figure 1 visualized total number and average price of Airbnb listings in NYC by neighborhood.

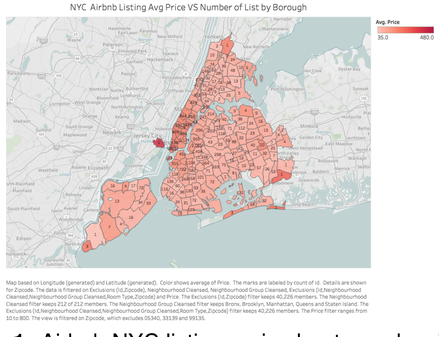
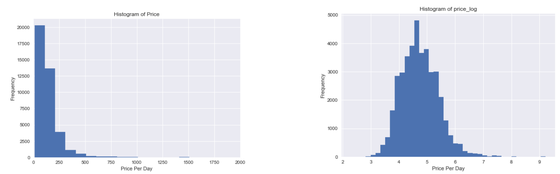


Figure 1: Airbnb NYC listings price heat map by zipcode

Target Variable Transformation

The original label data was skewed heavily. Therefore we did a log transformation and deleted outliers which lies outside 2 standard deviations.
Figure 2 shows the distribution change.



(a) First sub-figure **(b)** First sub-figure

Figure 2: Airbnb NYC listings price and log price distribution

Features

Missing Imputation

KNN and mode imputation for different categories of variables.

Feature Extraction

Categorical Variables

One-hot encoding.

Text variables

- Sentiment analysis: included both `_polarity` and `_subjectivity`.
- Topics: used word counts to vectorize text values and categorized those vectors into 10 topics using Latent Dirichlet Allocation technique. Finally, represented `_topic` features into boolean features.

Feature Engineering

Created five new features:

- `count_near_subway`: the number of subway stations within a small distance to the listing.
- `dist_to_nearest_subway`: the distance from the listing to the nearest subway station.
- `count_near_park`: the number of parks within a small distance to the listing.
- `dist_to_nearest_park`: the distance from the listing to the nearest park.
- `dist_to_famous_attraction`: the distance from the listing to the nearest tourist attraction.

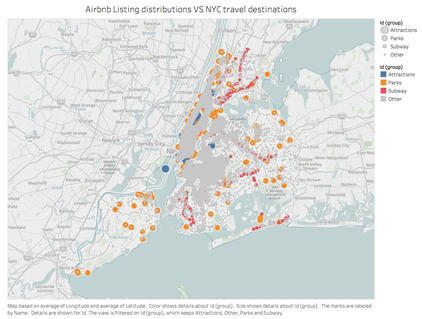


Figure 3: Listing distributions VS NYC travel destinations

Feature Importance

We used a number of randomized decision trees on various sub-samples of the dataset to calculate our feature importance. Figure 4 illustrated top 25 important features among around 400 features in total. New features `dist_to_famous_attraction`, `dist_to_nearest_park`, `count_near_subway` and `dist_to_neareast_subway` were among the list.

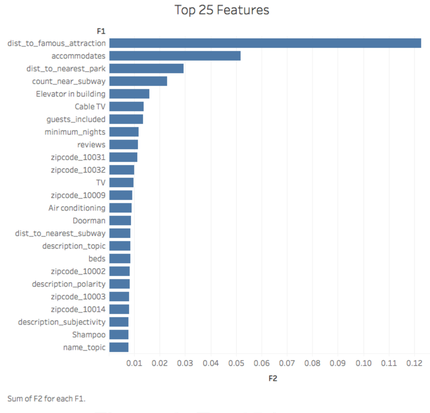


Figure 4: Top 25 features

Models

We divided the data into entire home and private room based on room types, and built models on both datasets. We tried three types of machine learning algorithms: regression models, regression trees/random forest and gradient boosting model. We used cross validation method to find our best parameters and chose the final prediction model from candidates based on the validation error. We used RMSE as our metrics.
Table 1 illustrates the RMSE for different models.

Results

	Regression		Tree/Forest		Boosting	
Party	Entire Home	Private Room	Entire Home	Private Room	Entire Home	Private Room
RMSE	0.308	0.326	0.302	0.319	0.286	0.305

Table 1: Model Selection

Conclusion & Error Analysis

We chose extreme gradient boosting model (xgboost) as our final prediction model, which minimized RMSE score as 0.286 for entire home and 0.305 for private room. The heat map visualized how our prediction is biased from the true price, shown in Figure 5. Cold colors refer to neighborhoods that we overestimated while warm colors stand for underestimation.
From the map, we could see that our model mostly overestimated NYC neighbourhoods, especially in the west side. Hamilton heights is the one mostly overestimated while East village is the one mostly underestimated.

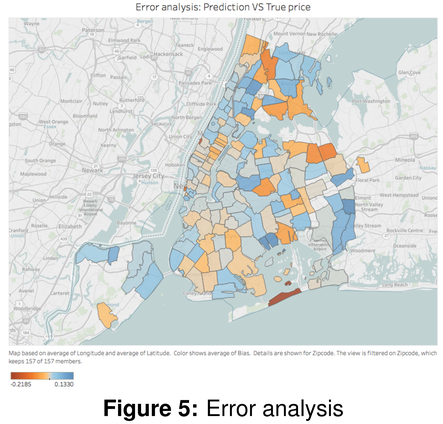


Figure 5: Error analysis

Advisors: Bonnie Ray , David Rosenberg