DS-GA-1003
Machine learning and Computational Statistics
Project Proposal

Siyang Sun(ss9558), Xinsheng Zhang(xz1757), Zemin Yu(zy937)

March 22, 2017

1 Business Problem

Airbnb, founded in 2008, keeps changing the way people live and travel. Through Airbnb, people starts to share space, experience, stories and lifestyle. It becomes a symbol of the sharing economy. Airbnb expands rapidly in the whole world, up to 2017, it has already advertised over 3,000,000 lodging listings in 65000 cities and 191countries. As students at NYU, we always love to explore the city we live in. In this project, we focus on Airbnb listing in New York City, where a cosmopolitan and world-wide travel destination. We believe Airbnb could give us a unique insight into the characteristic and neighborhoods of NYC. The high-level problem that we are trying to solve is to predict Airbnb housing price in New York City based on features including booking period, geographic information, housing layouts and amenities, etc. If time is allowed, we also would like to apply NLP on the housing review data, and adding review as a feature to our predictive model.

2 Data

We will use the data provided by Inside Airbnb where the data is sourced from publicly available information from the Airbnb site. We will concentrate on NYC Airbnb data.

There are mainly two parts of the NYC Airbnb data. For one 'listing' dataset contains the basic information of host such as host location (latitude, longitude, neighborhood, zipcode, etc.), house facilities (number of bathrooms and bedrooms, square feet, etc.), availability (maximum nights, security deposit, extra guest cost, etc.) and reviews (number of reviews, review scores, etc.). For the other dataset contains 614127 distinct customer reviews of the 31150 hosts.
For our modeling, we will more focus on using 'listing' dataset where contains 95 features and 40227 observations. Out of 95 features, there are 53 features have null value and 9 features have more 50% null values. 63 features are text features.

3 Machine Learning Formulation

Some base line algorithms to predict housing price includes but not limited to linear regression, regressions with penalty, linear decision tree and random forest based on a basic feature set. Since there are some options, We plan to accomplish model performance evaluation before our second meeting with advisers.

We will face these following technical difficulties in our data mining problem:

- Most of the columns (53) contains null value, and 9 of them has more than 50% null. We need to figure a reasonable way to clean null columns and impute null values.

- Lots of columns (63) are in natural language which requires NLP techniques to extract features.
- Not all the features (95) are useful to predict the house price, feature engineering are required to select best subset of feature to maximize the information gain.

4 Methodology and Timeline

March 31st: Data inspection and descriptive statistical analysis:

- Data cleaning and dealing with missing data
- Data Visualization
- Descriptive statistical analysis

April 1st - 18th: Feature engineering and baseline model evaluation.
April 19th: Second meeting with advisers.
April 20th - May 2nd: Model tuning and improvement.
May 3rd: Third meeting with advisers.
May 4th - May 9th: Final adjustment.