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

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

March 23, 2017

1 Business Problem

Founded in 2008, Airbnb keeps changing the way people live and travel. Not only do individuals share accommodations, but they also start to connect on unique stories and lifestyles through Airbnb, which has become a symbol of the sharing economy. With its rapid expansion in the whole world, Airbnb has already advertised over 3,000,000 lodging listings in more than 65,000 cities and 191 countries up till 2017. As students at New York University (NYU), we always love to explore the city we live in. In this project, we focus on Airbnb listings in New York City (NYC), a cosmopolitan and worldwide travel destination. We believe Airbnb could give us a distinct insight into characteristics of NYC neighborhoods.

We would like to propose a model to predict appropriate prices for Airbnb listings in NYC. Particularly, we are interested in studying how these features would help Airbnb provide an internal tool to assist hosts in determining reasonable prices. If time is allowed, we would also like to apply natural language processing (NLP) techniques on housing reviews, and integrate them into our predictive model.

2 Data

We will use datasets provided by Inside Airbnb where the data is sourced from publicly available information from the Airbnb site. For this project, we will concentrate on NYC data, which is compiled on December 3rd, 2016.

The data is separated into two datasets. We will mainly focus on *listing* dataset, which records detailed listing information in NYC. It contains 40,227 unique listings and 95 features in total. The features are roughly divided into the following groups.

- listing: name, space, description, accommodates, etc.

- host: about host, response rate, verification, etc.
- location: latitude, longitude, zip code, borough, etc.
- availability: maximum nights, security deposit, etc.
- review: review scores, reviews per month, etc.
- `price`

Specifically, our target variable `price` is continuous. From the summary table and the truncated histogram graph below, the variable has a heavy right tail, which indicates lots of outliers.

Mean	SD	Min	25%	50%	75%	Max
147.48	210.22	10	70	109	175	9999

Table 1: `price` summary

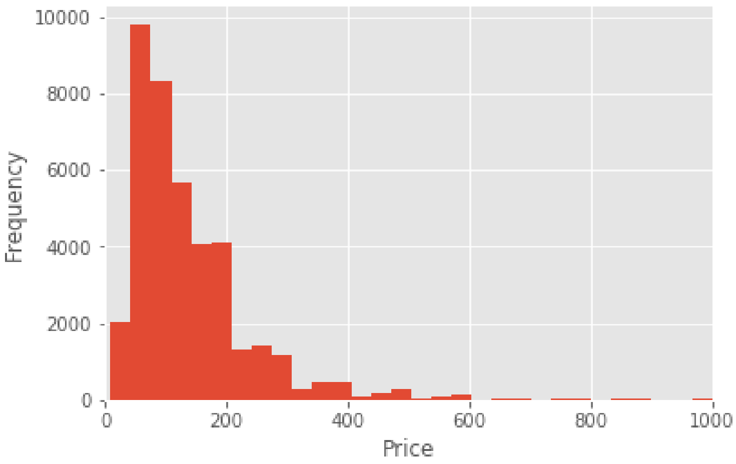

Figure 1: `price` distribution (truncated)

Optionally, the other *review* dataset collects detailed reviews for each listing. It contains a total of 614,127 reviews for 31,150 distinct listings.

Upon our initial data exploration, 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 out a reasonable way to clean null columns and impute null values.

- Lots of columns (63) are natural language, which requires NLP techniques to extract features. In addition, there are entries recorded in non-English language.
- Some features are categorical with a large number of unique values (for example, neighborhoods). We need to one-hot such variable, and thus are also likely to introduce too many dimensions.
- Not all the features are useful to predict the house price. Feature selection and feature engineering are required to select best subset to maximize the information gain.

3 Machine Learning Formulation

Some baseline algorithms to predict listing price include but not limited to linear regression (with regularization), linear decision tree and random forest based on a basic feature set. Since there are several options, we plan to accomplish some preliminary model performance evaluation before our second meeting with the adviser.

4 Timeline

- **March 23: Project Proposal Due**
- March 24 - March 31: Data inspection
  - Data cleaning and missing value imputation
  - Data Visualization
  - Descriptive statistical analysis
- April 1 - April 18: Feature engineering; Baseline evaluation
- **April 19: Second meeting with advisers**
- April 20 - May 2: Model selection and evaluation; Performance and Error Analysis
- **May 3: Third meeting with advisers**
- **May 9: Project Poster Session**
- **May 12: Final Project Report Due**