Section C Team 39
Paul Kim, Hillman Han, Ashley Teo, Yilun Li, Pallavi Rao

Team 39 Project Proposal

Assumptions:

Airbnb hosts already have the house and therefore don't have any additional costs.

1. What is the exact business problem?
   Before becoming an Airbnb host, the most common question people have will be how attractive my Airbnb list will be. Therefore, having a prediction about the popularity of the list will be important for those who want to become Airbnb hosts. Since there is no direct measurement of the popularity, our team decides to use the rating as a proxy for the popularity.
2. What is the use scenario?
   People who consider becoming an Airbnb host can determine how to design the cancelation policy, which neighborhood to choose, what room type is, what price and service fee to set, …, in order to design satisfying airbnb properties.
3. Which is the data source?
   NYC Airbnb Dataset which includes listings, full descriptions, average review score, reviews, unique id for each reviewer, detailed comments, calendar, listing id, price and availability for that day.
4. What is a data instance/unit?
   Data unit will be an Airbnb property's rating (1 through 5).
5. What might be the target variable?
   Rating of the Airbnb properties in New York city.
6. What features would be useful?
   Some of the features that would be useful : Neighbourhood_group, Neighborhood, instant_bookable, country, cancellation_policy, service_fee, number of reviews, availability 365, minimum nights.
7. What precisely is the data mining problem?
   a. Business and domain understanding: How does opening an Airbnb business work

b. Data preparation:
    - Cleaning the data. (Null values, empty data sets, etc.)
    - Identifying the numeric and categorical variables.
    - Setting the dummy variables.

c. Modeling
    - Regressions (linear, lasso, XGBoost, etc.)
    - Class classification (Random Forest)

d. Evaluation
    - Subsetting the main data into training data and test data.

e. Deployments
    - Might be unrealistic to deploy as trends, purchasing power, rent in the city, etc. change over time.

8. Supervised or unsupervised?

    This should be a supervised model because all the data is properly labeled. We will focus on using regressions and classifications to predict the dependent variable (rating) with the independent variables(each column in this dataset except for rating).

9. How exactly would it add business value?

    This will add business value to the Airbnb Hosts, as this will give them a better idea of the qualities of highly-rated Airbnb properties. This is very important to Airbnb hosts, as highly-rated Airbnb properties can lead to higher occupancy rate.