Project Proposal

Description of the problem of interest: predict how popular a rental listing is based on the listing content such as descriptions of the listing, photos, number of bedrooms, number of bathrooms, listing address, features of the apartment, longitude and latitude, and price. The interest level, our response variable, is a categorical variable that has three different levels: low, medium and high.

The description of the dataset includes data source, data size and variable information: The data is provided by Renthop, an apartment listing website. All of the apartments are located in New York City which includes not just Manhattan but also the outer boroughs such as Brooklyn and Queens. The data comes in the form of a json file that is 70 megabytes. We used pandas to convert this json file to a csv file. The csv file is 1.69 megabytes which consists of 50,000 rows and 15 columns of data. Some of the features are numerical such as price, number of bedrooms and bathrooms, others are textual. The description of the listing, for example, is largely unmined and contains various descriptions of apartments that agents choose to provide.

Explanation on the type of machine learning involved and the techniques that will be used

**classification methods**

logistic regression for more than 2 responses

lda for p>1

QDA

KNN

**Best subset selection**

Pca

Pls

Ridge and lasso

**Beyond linear**

Polynomial

Regression splines

Additive mixed models

Svm

**NLP**

Spacy for nlp

Discussions on the potential challenges and a brief plan on how to handle the difficulties

Classification of mixed categorical and numerical variables

Some very long descriptions were translated correctly, while some shorter ones generated new lines

The descriptions of the listings are in no inherent order so we may have to mine the text using natural language techniques before we can make use of the descriptions data