Benjamin Jakubowski (buj201)
Haonan Zhou (hz1004)

DS-GA 1006 Project Proposal

Project Name:

Machine Learning for Real Estate Comps (U8)

Project Description:

The value of real estate units are often understood by examining sales of real estate comps (comparables). This is used by property owners to assess the market values of their existing units and property developers to decide whether it is profitable to construct new units. Very often, much of the process of determining real estate comps are conducted in a qualitative and haphazard method that only uses basic information such as price per square foot of recent sales in similar neighborhoods. Moreover, often this information is gather informally (through professional networks) as a limited sample of known recent sales. Our goal for this project is to use a machine learning approach that allows us to incorporate more factors in the process. This will hopefully produce an automated, systematic, and more accurate method of valuing real estate properties by controlling for more complex features of real estate properties.

Project Data:

The primary source of data we will use is sales data from real estate websites such as StreetEasy.com, Trulia.com, and Zillow.com. These sites do not provide databases of sales or listings records (and their APIs are throttled), so we will need to write a suite of scripts to scrape data from their websites and extract records of current listings and sales that were closed in the past few year. This will give us some basic features about real estate listings and transactions that we will use as the basis of our analysis.

To obtain more features for the real estate units, we will also use Primary Land Use Tax Lot Output (PLUTO) data, which contains detailed information of land use and geographic data on a tax lot level. This is a public dataset maintained by NYC's Department of City Planning, which is available for download on their website. Our goal is to map each sales or listing that we scrape to a tax lot in order to obtain features that are not directed associated with the listings and sales on the real estate websites. In

addition, we will also consider using other geographic data for certain landmarks such such as schools, subway stations, and grocery stores (data sources TBD). This data will allow us to examine the proximity of real estate listings to these landmarks, which can also be very useful predictors of their value. These datasets will provide us with a wealth of features for each sales or listing that we can then proceed to prune and select during the analysis phase of our project.

Analysis:

Currently we anticipate using the following approaches to learning:

- We will likely start by clustering neighborhoods based on a limited set of residential unit features, in order to see where neighborhoods with similar residential properties diverge in sale prices. This will guide hypothesis generation as we consider adding additional features to our model.
- Similarly, we will likely try various clustering algorithms to determine comparable neighborhoods and properties. Note this clustering could be used to generate features for our supervised regression problem.
- Perform supervised learning with listing/sales price or price per square foot as target variable. We can try both linear models and more advanced methods (for example, Gaussian Process regression as suggested by Aaron Ng[1]). Regardless of modeling approach, one concern is the number of sales; we anticipate the number of residential property sales in NYC per year will be relatively small (since a large proportion of residential properties are rentals).
- There will be a large number of binary/categorical features in our data (e.g. condo vs co-op, borough of property). We may consider splitting the data into a few distinct groups and modeling each separately (i.e. learning a hierarchical model).
- We need to think about how to factor time of sale into our model because real estate prices are clearly changing over the years. One simple approach is to bucket each sale by year or month of sale, though more advanced approaches may be worth considering.

Challenges:

Additional considerations include the project objective. Framing the problem as either objective

[1] http://www.doc.ic.ac.uk/~mpd37/theses/2015_beng_aaron-ng.pdf

- **Price prediction:** minimize a cost function for sale price prediction
- **Comps identification:** identify comparable properties

will likely lead to very different solutions (especially given the constraints implied by either approach). For example, if our objective is to identify comparable properties in the same neighborhood sold in the recent past, we are likely losing information that could improve the predicted sale price implied by our comps-identifying model (under which predictions could be made using KNN-regression). Conversely, an optimized price prediction model may not immediately support comps identification.