Project Description

Capstone Project U2

Team Name: Reveal Estate

Nora Barry	neb330@nyu.edu	github: neb330
Laura Buchanan	lcb402@nyu.edu	github: lcb402
Jacqueline Gutman	jg3862@nyu.edu	github: jgutman

Advisor: Luc Wilson lwilson@kpf.com

Project Description

For our MSDS capstone project, our goal is to first join city planning and zoning data with property sales data, along with other relevant NYC open datasets, to explore the factors most strongly influencing property sales prices in residential properties compared to commercial properties in New York City. In particular, we are interested in uncovering observations that may provide financially impactful recommendations for future development. For example, we plan to investigate the role of limited height districts in controlling property sales prices, or the importance of 311 complaints in deflated property values in a particular neighborhood.

After merging PLUTO and Department of Finance data, we will investigate the dataset both visually and via mathematical models. We anticipate building a visual representation that will allow off-the-cuff filtering of data to see spatial distributions of features of interest, and aid in hypothesis building. Our model, built with machine learning techniques, will enable us to gain insight into which features wield the greatest predictive influence. We intend to focus on the comparison between residential and commercial sales prices and contrast the most critical factors for each type of property to gain insights on what drives property value and how it might be maximized.

Introduction to the Datasets

PLUTO

The New York City Department of City Planning releases the PLUTO dataset as part of the NYC Open Data platform containing data pulled from multiple city agencies on every tax lot across the five city boroughs. The PLUTO data contains over seventy fields describing these lots, including which police precinct, fire company, sanitation district, school district, and other public services serve that lot. In addition, the PLUTO data can be joined to other public data sets through census tract and X-Y coordinate information. Census tract codes allow the lots to be joined together with American Community Survey (ACS) census data, and the X-Y coordinates can be translated to latitude and longitude coordinates to allow the incorporation of many additional sources

of geocoded spatial data. Additionally, zoning information on whether the lot is a historical landmark or part of a limited height district help account for restrictions on further development on a particular lot, which may artificially alter the property sales value in a predictable way.

Department of Finance Sales

The New York City Department of Finance releases aggregated data annually for each borough of New York City containing all properties, both residential and commercial, sold in that borough during the preceding year. Currently, these data are available for the years 2003-2015, with rolling data released monthly for the most recent sales taking place over the past 12 months. These property sales are sorted by borough and neighborhood, and include information about the class of the building being sold--for example, whether it is a two-family home, an office building, a commercial garage, or any of hundreds of other commercial and residential building codes.

The data includes the sale price and sale date, as well as the building size in square feet and the number of commercial and residential units in the building. Many properties include information about the year the property was built, but for a large proportion of buildings, this information is missing or unknown. From the sale price and property square footage information, we can normalize the price per square foot of each property to extract the desired target for prediction in all models. This data includes information about the size of the property in terms of both land square feet (the square footage of the property on the ground level) and the gross square feet (the square footage of the property across all levels of the building)--for our purposes, unless otherwise specified, square feet refers to the gross square footage of the property, as this relates more directly to the property sales price.

Because of the vast differences between the factors influencing sales price for commercial compared to residential buildings, we can separate these classes of properties for modeling by considering the number of commercial units and the number of residential units in a building, potentially excluding mixed-use buildings that include both types of units in the same property. From the tax block and tax lot number of each

property sold, we can join the Department of Finance sales data to the properties referenced in the PLUTO data set.

Supplementary Urban Data

We will incorporate a variety of location based data onto the datasets described above. These datasets will primarily come from NYC Open Data. We anticipate features from this data (such as crime rate, number of farmer's markets, etc.) will prove value in predicting property value. As we finalize our visualization tool and model framework, we may merge model data, allowing for a broad range of features to be considered.

Proposed Models

Baseline Model

Our baseline model will be a straightforward regularized linear regression approach where we see which features are most predictive of property value. We will incorporate machine learning methods. For example, in addition to optimizing via stochastic gradient descent, we can build a model to prefer a sparser solution using Lasso regularization, so that we focus on the most predictive features and avoid focusing on the relative weights of non-predictive features.

Extended Models

Once our visualization tool is built and our baseline model framework running, we will begin a thorough search for hypotheses concerning impacts on property value. From there, we will build additional models to test these hypotheses. In particular, we will consider nonlinear regressors, including Random Forests and Gradient Boosting Machines, to explore more complex, nonlinear relationships between the predictors and the sales price per square foot, allowing for separate, non-interacting models of the commercial and residential properties.