Midterm Report: Home Purchase Assistant

Shibo Zang (sz428), Yangwen Wan (yw762)

Dataset Description

Feature Selection & Engineering

We would like to select those features that have close relationship with the price of real estate properties and hopefully can be independent with each other. There are several factors that we felt would have an impact on the price: location / neighborhood, building type (apartment / house / condo), year built, building class (one-bedroom / two-bedroom / studio), etc.. Basing on the dataset we have on hand, we selected or generated the following features for our model:

- Building Class At Time of Sale: The building classification follows the “NYC Building Classifications”, which is used to describe a property’s constructive use. For example, “A” signifies on-family homes, “O” signifies office building, “R” signifies condominiums. More specifically, the number on the second position adds more specific information about the property’s use or construction style, e.g. “O4” is a tower type office buliding.
- Tax Class at Present: Every property in the city is assigned to one of four tax classes based on the use of the property. Total Units: Sum of residential units (number of residential units at the listed property) and commercial units (number of commercial units at the listed property).
- Land Square Feet: The land area of the property listed in square feet.
- Gross Square Feet: The total area of all the floors of a building. The reason we would like to keep both Land Square Feet and Gross Square Feet as features is that we figured they are not totally positive correlated - Gross Square Feet contains some information that Land Square Feet doesn’t possess, like number of floors, other land area and space within any building or structure on the property.
- Year Built: Year the structure on the property was built.
- Neighborhood Average Price: Each real estate property unit is marked uniquely by a Borough-Block-Lot identifier. This gave us a natural advantage to group close real properties together by “Block” number, whereas using street address, which is hard to analyze and organize, and zipcode is too broad to define a neighborhood. Therefore, the neighborhood average price would be a strong indicator signifying the price of the property you are looking at.

As we discussed above, you might find that some features are numerical, and some features are categorical

like ‘Building Class At Time Of Sale’ and ‘Tax Class At Present’. Since most regression models or ML algorithms cannot fit categorical variables, we are going to use dummy coding to convert a categorical feature into continuous variable. For example, there are four tax classes, we can have four features ‘Tax Class A’, ‘Tax Class B’, ‘Tax Class C’, and ‘Tax Class D’. Presence of a tax class is represented by 1 and absence is represented by 0.

The other thing we did when we cleaned and prepared the data is to standardize the range of those features. As you can see the range of the year when the property was built is from 1900 to 2009, whereas the neighborhood average price ranges from 10,000 to 5,000,000. We normalize features by calculating their standard scores $x' = \frac{x-\bar{x}}{\sigma}$, where $\bar{x}$ is mean and σ is the standard deviation.

Since we have more than 10,000 records per borough each year, we are expecting to have the underfitting issue in our model. From feature engineering’s perspective, the way to avoid underfitting is to add more features, either from external resources or apply transformation on existing features. There are some other financial factors that may affect price of real estate properties such as people’s average annual income, GDP, inflation rate, etc. We are trying to gather these information and append them as new features to our model. On the other hand, we can generate new features using existing features by computing the production or exponentiation of feature values.