CS221 Fall 2017 Project Proposal

Patrick Cho (), Sudarshan Seshadri () & Jianqing Yang (yangjq)

Task Definition

We are developing a predictor for housing prices in Singapore for an industry project. The industry concept is similar to Opendoor[1] where the industry client wants to be able to algorithmically predict the market value for a property in order to make an initial purchase offer (the output) solely based on the property's metadata (the input) without the need for physical inspection of the property or interviewing the seller. (TODO: is the output price used directly or will it be adjusted by a human first?)

Similar to anywhere else in the world, the housing prices in Singapore are affected by location, floor area and number of rooms which can be captured in the data. Other conditions such as the upkeep and interior design of the property are not represented in the data thus there would need to be some room for variance in the output price[2]. However, there are some characteristics particular to Singapore which our model will need to account for:

- Most residential areas in Singapore are densely packed, with the vast majority of housing as apartments[3]. However, within the same apartment block there can be significant variations in price depending on which floor the unit is on (higher floors tend to fetch higher prices) and the overall facing of the unit (units which do not receive direct sunlight in the afternoon tend to fetch higher prices). A floor range for each unit is available in our data while facing is not.

- The vast majority is also sold on a 99 year lease (the alternative is a 999 year lease), which can have a large impact on market value depending on the number of years remaining. Although units within the same apartment block will have roughly the same remaining lease, it is also common to have neighboring blocks built decades apart. (TODO: is this in the data?)

This gives us the challenge to avoid over-generalizing our output towards location without properly representing the above features, especially since the number of recent previous transactions for a property with very similar characteristics will generally be sparse.

We will evaluate our system based on the accuracy of the output market price (TODO: specific metrics). The memory or running time is not a major concern for our application as long as we are able to generate our output within minutes[4] using readily available commercial resources (e.g. AWS, Google Cloud).

[1] http://www.opendoor.com

[2] Again, similar to Opendoor which adjusts their offer price after an inspection

[3] https://data.gov.sg/dataset/resident-households-by-type-of-dwelling-annual

[4] Taking Opendoor's reply within 24 business hours as the benchmark

Infrastructure

We have raw data from the client (TODO: elaborate, give concrete examples of inputs and outputs)

Approach

(TODO: describe baseline algo and results)
    (TODO: we need an oracle algo)