Key features determining HDB sale prices & data informed model for prediction of prices

Benny Teo

January 25, 2021

1. Introduction

1.1 Business Problem

In Singapore, public housing are named as the ubiquitious HDB flats. Due to the growing economy and land scarce nature of the island nature, property prices are on a general uptrend. Despite it being public housing, lease holders have the right to buy and sell their HDB flats allowing for a chance for profit.

Despite public data on transaction prices being available, it is still a challenge to determine the right price to sell at. This data science project aims to identify the factors that influence real estate prices and hopes to develop a predictive model considering those factors to provide future sellers with a data informed pricing guide to sell at.

The problem statement is: What is/are the key feature(s) that influences HDB transaction prices?

The major benefit of this business problem and its solution to clients or stakeholders is improved decision making in evaluating the impact of different factors to the price of flat for buying/selling evaluation.

2. Data

2.1 Data Sources

Transaction prices

Transaction prices are the end outcome of what we are trying to predict.
Source: https://data.gov.sg has a wealth of such publicly available data. For the purposes of this project, we will take the latest available year of transaction prices.

Details of the transacted flat

The details of the transacted flat will be needed to evaluate the influencing factors of the sale. Source: https://data.gov.sg has a database that includes Town, Flat type, Block, Street name, Storey range, Floor area sqm, Flat model, Lease commence date, Remaining lease. The 9 factors is presents a sufficiently wide range of factors for the purposes of the study.

Foursquare location data

Foursquare location data can be used to ascertain nearby amenities and provide a supporting view of the analysis. Source: https://api.foursquare.com

The venue data will be grouped by category types in order to give a perspective of the concentration of amenities in the towns
The focus of the study will be on a regression model and thus will not apply the clustering methodology used in the course. It is the author's intention to explore methods beyond what was taught in the course rather than replicating the course steps.

3. Methodology

3.1 Data preprocessing

Since I'm using 2 different API sources, the data had to be pulled, input into a data frame and combined.

I started first with the Data.gov.sg API as it contained the majority of the information required. Having read the dataset into a pandas dataframe, I sought to understand the data set with the describe() function. I checked for null values and found that the data set had none.

The data included 11 columns below:

Column	Data type
town	String
flat_type	String
flat_model	String
floor_area_sqm	Float
street_name	String
resale_price	Float
month	String
remaining_lease	String
lease_commence_date	String
storey_range	String
_id	Integer
block	Integer

Looking at the data columns, I determined the following features could be used:

1. Town/Location
   Rather than looking specifically into the geospatial coordinates, it is known in the industry that the town that the flat is in usually has some influence on the transaction price. I decided to look only at the Town rather than the exact streets or block where the transaction took place as the transaction data is not significantly wide enough to differentiate at such a level and might lead to false positives.

2. Resale price
   This field does not require any preprocessing as it is already in the correct data type.

3. Remaining lease
   Given that HDB flats have a across the board 99 year lease, this was an important feature to factor in. The field on remaining_lease however was not current as it was measured off the date when the data was ran. The remaining lease was instead obtained from the lease_commence_date and transposing the result to a float.

4. Storey
   It is also another industry axiom that the higher the selling unit is, the greater the value. Thus storey range was an important feature to obtain. Since the data indicated a range of levels rather than the actual unit level, I decided to use the median of the storey range that was indicated. This was done by extracted the lower & upper storey levels and obtaining the median of the 2. The result was stored into a new column as an integer.

Venues information from Foursquare API

To obtain the venues information in each town, I used the Foursquare API. I first created a new dataframe with which the unique town names were populated. The geospatial data was obtained from a csv file and the information added into the dataframe. Using the function learnt from the course, I called the Foursquare API to obtain the list of nearby venues in the town. The venues were then grouped by category type to enable the analysis of the number of surrounding venues.

3.2 Modelling

Regression Machine learning techniques were utilized for this project.
I utilized 2 of the most commonly used Regression modelling techniques (Linear Regression & Gradient Boost Regression).

The features used were ‘ ‘floor_area_sqm’ (from the original data set) and the ‘remaining_lease_int’, storey’ (both engineered features mentioned in the “Data preprocessing” section

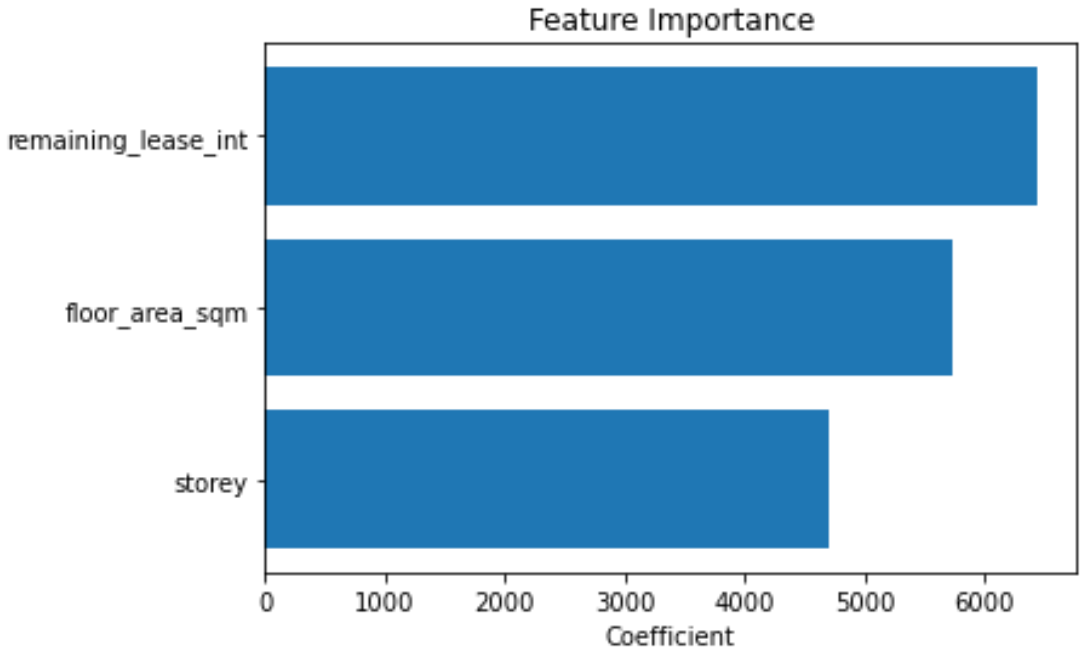
Although the Linear regression model was already giving a very high Rsq score, the Gradient Boost Regression was also conducted to validate if there is a better model that can more accurately predict the features influence on the transaction prices.

4. Results

4.1 Linear regression model

Rsq = 0.9609234411537675 (High accuracy in predicting transaction prices)

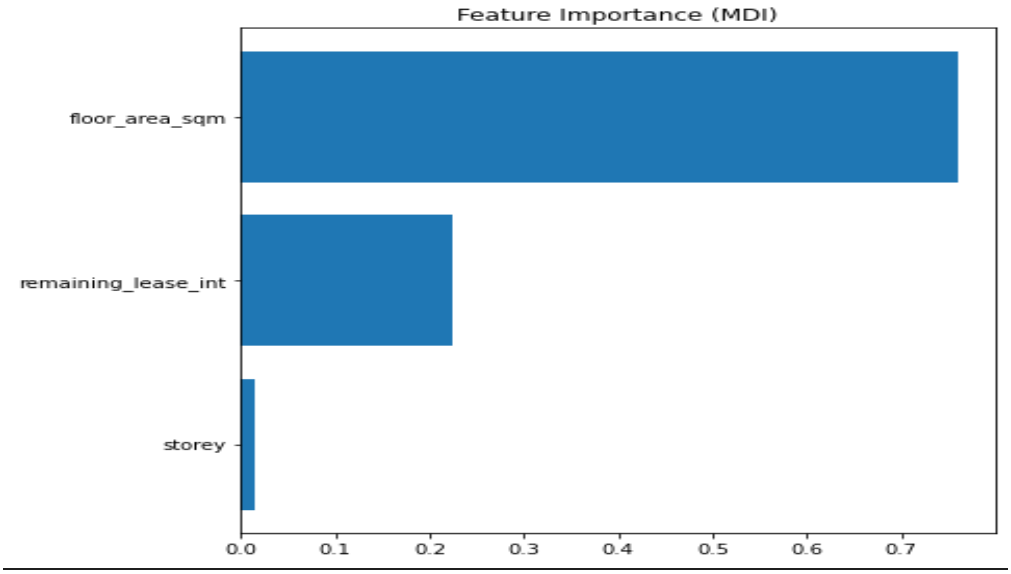
Linear Regression Feature Importance Chart



4.2 Gradient Boost Regression Model

Rsq = 0.9514292231963996(High accuracy in predicting transaction prices)

Gradient Booth Feature importance Chart



Discussion of results

Both models give a R Sq of over 95% which is a good level for the purposes of this study.

The Linear regression model has a higher Rsq score at 0.9609 than the Gradient Boost regression model's Rsq at 0.9514 suggesting that the linear regression model though simpler is a better predictor of the sale prices of the flat.

The Feature importance chart (which identifies the impact of the features to the sales prices) differ for the Top 2 features. The Linear regression model identifies the remaining lease as the top feature influencing transaction prices with the floor area of the flat following closely behind. The Gradient Boost model however prioritizes the floor area of the flat over the remaining lease and determines the impact of the floor area to be more than 2 times that of the remaining lease.

Both models are consistent in identifying that the storey level at which the unit is located has the smallest impact to the sale price of the unit.

Analysis of venues

----ANG MO KIO----

Category	Count
Asian Restaurant	1
Burger Joint	1
Japanese Restaurant	1
Noodle House	1
Park	1
Ramen Restaurant	1
Snack Place	1
Spa	1
Supermarket	2

----BEDOK----

Category	Count
Asian Restaurant	1
Bakery	1
Breakfast Spot	1
Café	1
Chinese Restaurant	1
Coffee Shop	1
Food Court	1
Supermarket	1
Thai Restaurant	1
Vegetarian / Vegan Restaurant	1

----BISHAN----

Category	Count
Bookstore	1
Chinese Restaurant	1
Coffee Shop	1
Dumpling Restaurant	1
Food Court	1
Gym	1
Japanese Restaurant	1
Pool	1
Shaanxi Restaurant	1
Stadium	1

Looking at the venues in each town, although it seems like there is a wide spread of categories, most of them can be classified as Food & Beverages (i.e restaurants, burger, food court, coffee shop, cafes, noodle house). Other than Food & Beverages, the unique venues for the towns are as follows:

- Ang Mo Kio: Park, Spa, Supermarket
- Bedok: Supermarket
- Bishan: Bookstore, Gym, Pool, Stadium

5. Discussion

Observations noted
From the venues data, all 3 towns have a wide variety of places to eat. Buyers that are looking for more recreational facilities can look to Bishan where there are sports facilities like gyms, pools and stadiums. A neighborhood covering more needs would be Ang Mo Kio where there are recreational (Park & Spa) and amenities(supermarket) as well. Bedok has a supermarket but the venues do not reveal amenities in the neighborhood.

Recommendation:

From the regression models feature importance chart, all else being equal, when deciding on the premium to sell a HDB flat at, sellers should focus on their units' floor area and remaining lease rather than placing a premium on the level at which their unit was located. Although it is commonly regarded that every storey level brings a premium, the model reveals that whilst that may be true in certain cases, the floor area and years of remaining lease are the key features in determining a selling price.

Given that it has a high predictive accuracy, potential flat sellers can utilize the linear regression model developed to get a data backed estimate of a reasonable selling price base on the features of the flat.

6. Conclusion

This project has provided a data informed pricing model to guide sellers on HDB prices to sell at. It has answered the problem statement and identified the key features influencing HDB transaction prices namely (floor area & years of remaining lease). It has also provided a categorisation of venues type to provide a perspective on the venues in each neighborhood to better inform HDB flat purchases. Clients now have an improved decision making in evaluating the impact of different factors to the price of flat for buying/selling evaluation.

7. Opportunities for further development:

1. This work can be further extended to include more HDB flat features to make the model more robust.
2. URA API source (https://www.ura.gov.sg/maps/api/#introduction) includes only Private residential/landed prices. The dataset may be relevant if there is an extension of the work to private housing.