Introduction

For most people, purchasing a property is their biggest investment and so it is important for buyers to know if they can afford the property. In Singapore, the government gives generous subsidies (grants) for first-time buyers of public housing flats and most people use their mandatory saving fund (known as CPF) to pay for the properties. These flats were built by a government agency known as Housing and Development Board (HDB) and so locals refer to these properties as HDB flats.

Singaporeans that meet the criteria take advantage of this policy to own their homes, resulted in 80% of the residents living in HDB flats. They either buy newly built flats directly from the government, which require them to wait for a few years, or pick a ready-built flat from the resale market.

Problem Statement

Buying a property is a big investment. Buyers want to know the best estimate for a property for 2 reasons:

1. Can they afford the downpayment and mortgage?
2. Can they sell with a profit later?

Sellers want to ensure that they are not making a loss, they can meet the profit target they have set, or the sale proceed is enough to finance their next property.

The goal of this project is to predict/estimate the price of HDB resale flat using a regression model. The price of a property is determined by many factors, such as its condition, floor level, size and location. It is also affected by the current state of the economy, government policies and supply and demand.

This requires the necessary data to be collected, which are described in the next section.

The metric for measuring the performance is RMSE. The goal is to improve the baseline RMSE score of the model by at least 10%.

Data Collection

1. HDB Flat Resale transactions (from 2015 to May 2020)

- flat attributes and prices (Source: data.gov.sg)

2. Supply and Demand Factors

- Sales of new HDB flats (source: data.gov.sg)
- Sales of new private homes (source: data.gov.sg)
- Number of married people (source: singstat)
- Number of Residents (source: singstat)

3. Macroeconomic factors
   - Consumer Price Index (source: Singstat)
   - Purchasing Manager Index (source: SIPMM)
   - Composite Leading Index (source: Singstat)

- GDP Growth (source: Singstat)
- CPF interest rates (source: Singstat)
- Singapore Interbank Offered Rate (source: Singstat)
- umemployment rate (source: Singstat)
- median income of residents (source: Singstat)
- HDB flat price index (source: data.gov.sg)
- Private property price index (source: data.gov.sg)

4. Points of Interest

- Shopping Malls (source: wikipedia and other websites)
- Nature Parks (source: data.gov.sg)
- Schools (source: data.gov.sg)
- Sports Facitilites (source: data.gov.sg)
- MRT/LRT stations (source: kaggle)
- Hawker centres and markets (source: data.gov.sg)
- Librarues (source: data.gov.sg)

Most of the data came in the form of csv files, some are geojson and kml files. Shopping malls data were scraped from websites and locations' latitude and longitude were fetched through OneMap API.

Feature Engineering

Here I add macroeconomic data and Points of Interest to the HDB dataset.

Macroeconomic data are time based, meaning the values change over time, e.g. GDP growth is calculated quarterly, PMI is monthly, SIBOR interest rate is daily. I match macro date based on flats' sale dates.

Each HDB flat and Point of interest has a pair of longitude and latitude to represent its location, which I used for calculating the following distances between flats and points of interests:

dist_mrt - distance between flat and the nearest MRT station
dist_market - distance between flat and the nearest market/hawker centre
dist_mall - distance between flat and the nearest shopping mall
dist_library - distance between flat and the nearest library
dist_school - distance between flat and the nearest primary school
dist_sport - distance between flat and the nearest sport facilities
dist_park - distance between flat and the nearest nature park
dist_core – distance between flat and the downtown core

The Downtown Core Planning Area is the economic and cultural heart of Singapore.
https://www.ura.gov.sg/Corporate/Guidelines/Urban-Design/Downtown-Core

Modeling

I started with an Ordinary Least Square Linear Regression Model to establish the baseline, and then explored Keras and XGBoost, tuning their hyperparameters using RandomSearchCV.

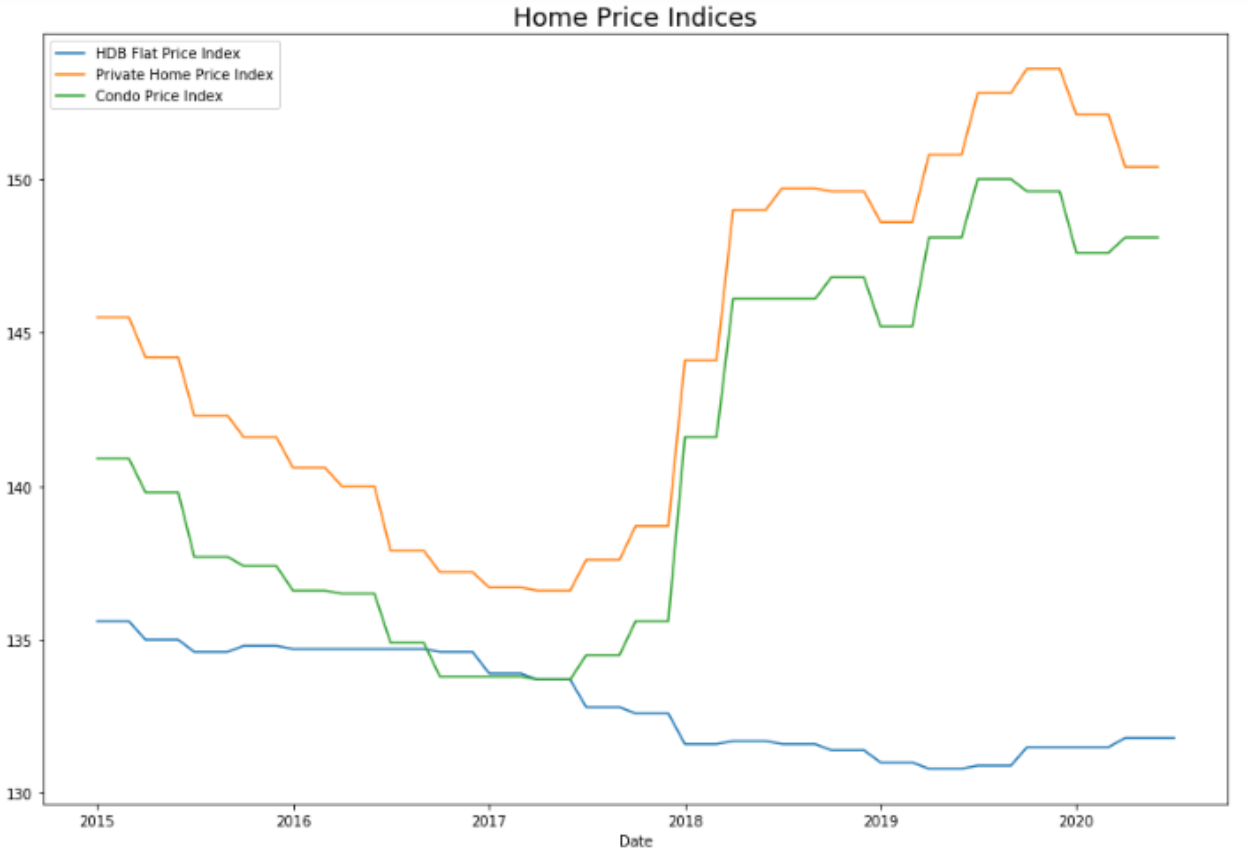
Guided by RMSE and R2 score, I compared Keras with the different configurations of XGBoost and finally settled on a XGBoost model.

Initially I used records from 2015 to 2019 as training and validation. After confirming the model, I used the data 2015 to 2019 as training set and the unseen data from 2020 as test set.

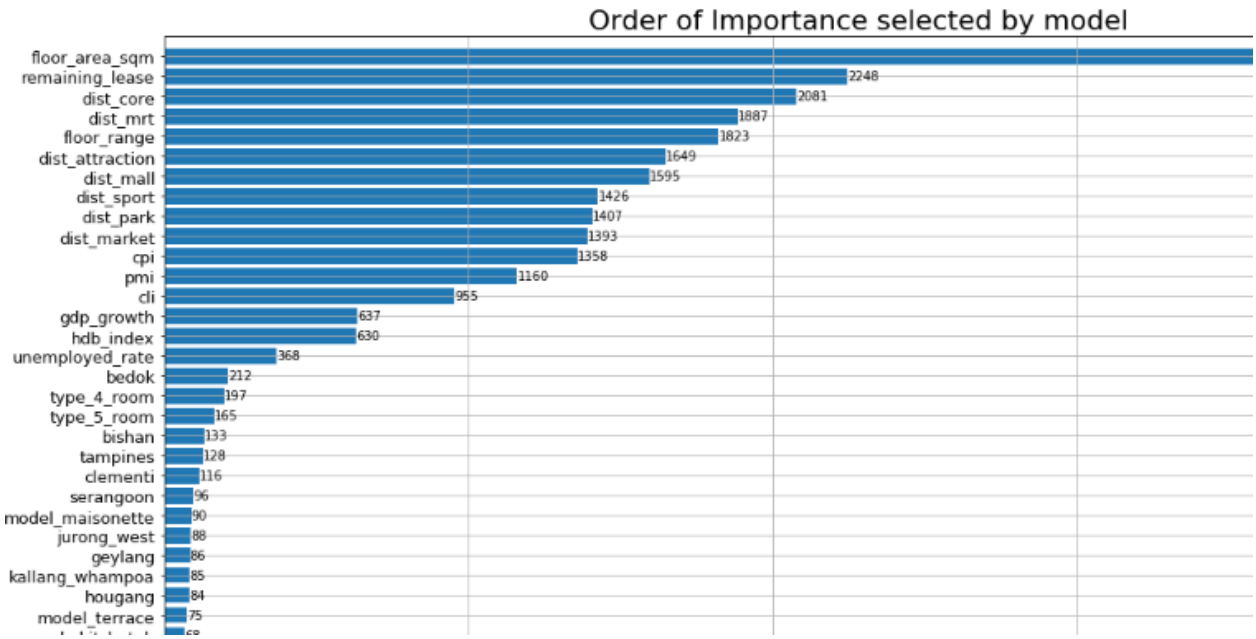
Findings

- Between Jan 2015 and May 2020, average price of HDB flats is SGD 438,000 with standard deviation of SGD 148,000
- Between Jan 2015 and May 2020, there were 107,000 resale transactions

  - HDB flat prices have started to go down in 2018 due to oversupply, and will continue to remain low due to poor economic condition in the next few quarters. This can be seen from the plot below.



- Most important feature that influences prices is floor area, followed by
  - Flat attributes (floor area, storey, remaining lease)
  - Location (mrt, market, attraction, mall, park)
  - Macroeconomic (cpi, pmi, cli, gdp)



The above plot shows which are the features that influenced the predictions, with the most important features at the top. It is not surprising that Floor Area is the most important feature, but it is also important to note that this is not the only feature that affects the price of a property.

Next is the Remaining Lease, which is important for HDB flats because all of them are 99 years leasehold properties, which means that at the end of 99 years the property will be returned to the state at zero value. That is why older flats fetch lower prices, assuming all other factors being equal.

The location of a flat means its proximities to points of interest, including the Downtown Core. Buyers are willing to pay more if the flats are near to Downtown Core, which is the economic and cultural heart of Singapore. In private residential market the government maintain indices for three different regions, and downtown core belongs to the region with the most expensive private properties.

Other Points of Interests also influence the price of a flat, such as nature parks, shopping malls, markets, sport facilities and attractions. But the one that most buyers look for is MRT station. With 60% of the population traveling to work/school in public transport, proximity to MRT station became an important price predictor.

Flats at higher floors are getting lesser noise from the street and offer better views, so it is shown here as an important factor.

Macroeconomic factors also play a part in determining flat prices, such as CPI (consumer price index), PMI (purchasing manager index), CLI (composite index) and GDP growth rate. Buying a home is a big investment. During a recession people are worried about their jobs and their ability to service the loan, and that would put downward pressure on price.

Conclusion

In the problem statement I mentioned that the goal was to improve the baseline RMSE score by at least 10%.

The RMSE score for baseline Ordinary Least Square model was 46987. With XGBoost the RMSE was reduced(improved) to 25204

More than 60% of people get to work or school by public transport, and 40% of those trips were made with MRT, which makes proximity to MRT station an important factor in home prices
https://www.budgetdirect.com.sg/car-insurance/research/public-transport-singapore

First-timers and young couples usually go for BTOs. Though they take at least two years to complete and are smaller, they are more affordable than resale flats and therefore the ideal purchase for young couples who have just started working. But with the enhanced grant introduced in 2019, more first-timers are turning to the resales market.

Advantages of resale flat are:

- Choose the location (new flats are mostly in new estate) including living near parents to get additional grant and childcare
- Ready to move in (in weeks rather than years)
- Mature estates (amenities ready)

Deployment
I used Flask micro web framework to turn the model into a web application and deploy it to Heroku - https://hdbprice.herokuapp.com/.

There is certainly improvements needed on the web application UI design, which I will continue to work on.

For production use it is also necessary to perform maintenance tasks to keep the model up-to-date:

- Periodically update time-sensitive data to Herokua such as cpi, gdp growth rate and pmi
- New Points of interest will change over time so it is necessary to get new lists, fetch their geocodes and calculate their distances from flats

Limitations

- Condition of the flat is not taken into consideration. Some buyers consider that to be important
- Other than large nature parks, each point of interest is currently represented by a single geocode, but the physical size of the location can be more than a hundred metre end to end

- Currently a user needs to provide the flat's model, type, floor area, year built and address to get a prediction. Ideally the user only needs to enter his address but flat attributes are only available for those that have changed hands before. I hope HDB can provide flat attributes for all flats