Analysis of Singapore HDB Resale Price

IBM Data Science Capstone Project

Sun QiFang

28 December 2019

Intro

The Housing & Development Board (HDB) is Singapore's public housing authority and a statutory board under the Ministry of National Development. They develop public housing to provide Singaporeans with affordable, quality homes, and a better living environment. Focusing on nurturing a strong family and community spirit, public housing policies and schemes are formulated to meet changing needs and aspirations. Foreigners who gained Permanent Residency are allowed to purchase HDB housing through resale. In this project, we are targeting to develop a machine learning model to predict the resale price of HDB housing.

The project is designed for Singaporean/Permanent Residents (PR) to predict the resale price of their prospect HDB housing. The result serve as guidance for pricing.

Data Acquisition

1. Government Data API: https://data.gov.sg

   - HDB property info: https://data.gov.sg/dataset/hdb-property-information
   - HDB median rental by town by flat type: https://data.gov.sg/dataset/median-rent-by-town-and-flat-type
   - Median Resale price by town by flat type: https://data.gov.sg/dataset/median-resale-prices-for-registered-applications-by-town-and-flat-type

2. Singapore official geocoding data through OneMap API: http://developers.onemap.sg

Total 34554 units of HDB with resale price scrapped and 9303 units to predict.

Data Processing

- Transform of year of completion to age of HDB housing
- Map all town code to town name for consistency
- Concatenate block number with street name
- Change Flat type to categorical
- Transform Yes or No feature to Boolean variables.

```
Int64Index: 34554 entries, 0 to 43855
Data columns (total 28 columns):
max_floor_lvl            34554 non-null int64
residential              34554 non-null int64
commercial               34554 non-null int64
market_hawker            34554 non-null int64
miscellaneous            34554 non-null int64
multistorey_carpark      34554 non-null int64
precinct_pavilion        34554 non-null int64
total_dwelling_units     34554 non-null int64
1room_sold               34554 non-null int64
2room_sold               34554 non-null int64
3room_sold               34554 non-null int64
4room_sold               34554 non-null int64
5room_sold               34554 non-null int64
exec_sold                34554 non-null int64
multigen_sold            34554 non-null int64
studio_apartment_sold    34554 non-null int64
1room_rental             34554 non-null int64
2room_rental             34554 non-null int64
3room_rental             34554 non-null int64
other_room_rental        34554 non-null int64
age                      34554 non-null int64
address                  34554 non-null object
town                     34554 non-null object
flat_type                34554 non-null float64
median_rent              34554 non-null float64
price                    34554 non-null float64
Latitude                 33891 non-null float64
Longitude                33891 non-null float64
```

Methodology

This is a regression model with features of HDB housing by type, age, commercial, facilities and location data, with target as resale price. We will use Lasso regression of L1 penalty to select the features, and build regression on it to target resale price.

Analysis

We have evaluated the correlations of all variables by heat map, as shown below:

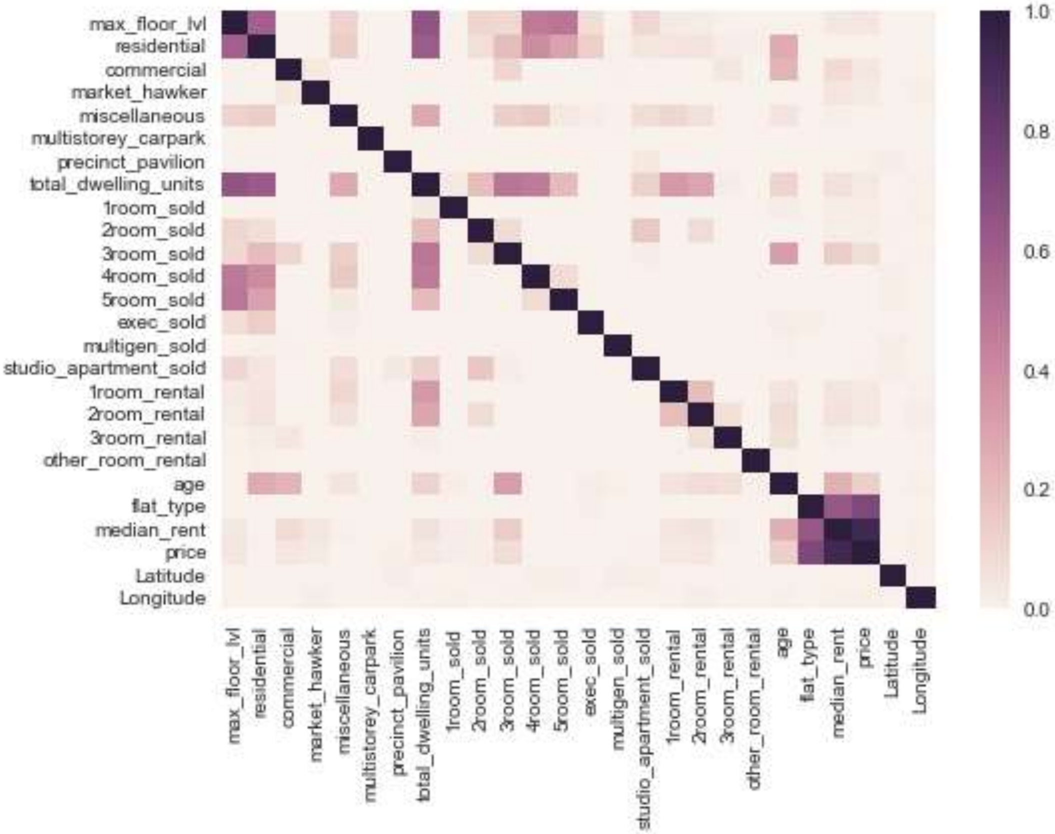


Figure 1. Heat Map of Correlations

Clearly Price is correlated mostly to flat type and median rent only. Even coordinates by Latitude and Longitude are not affecting the price too much.

The target of prediction, resale price ranges mostly from 300K SGD to 500K SGD, with discrete distributed 500+K to 800K SGD. Distribution as shown below.

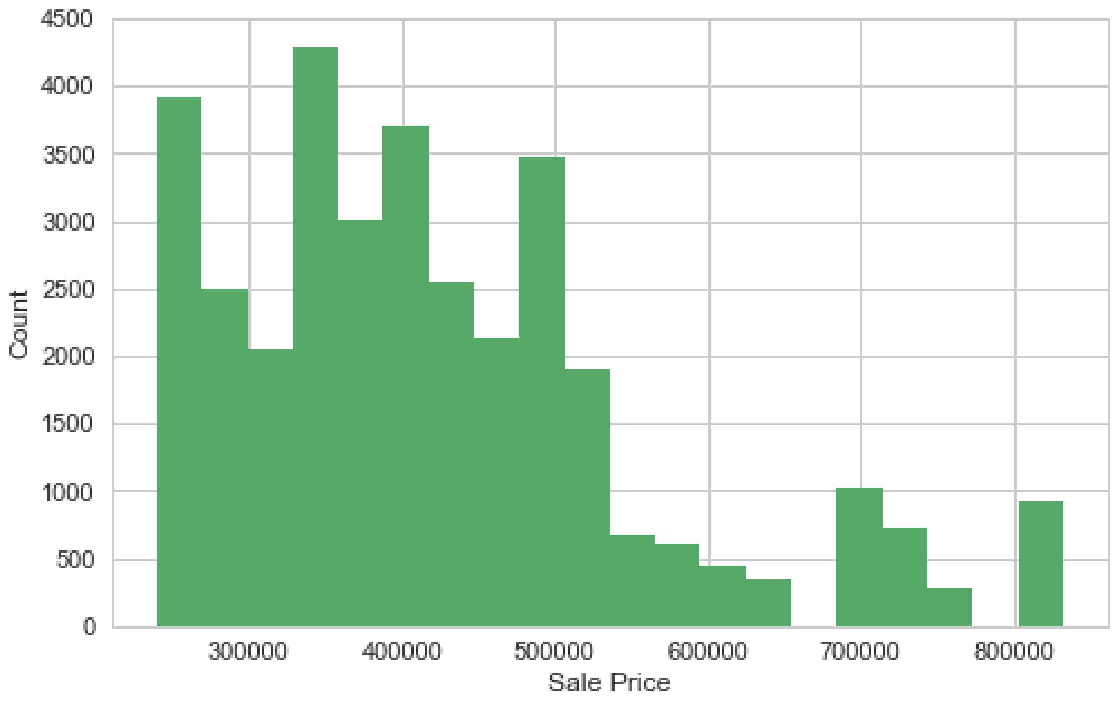


Figure 2. Resale Price Distribution.

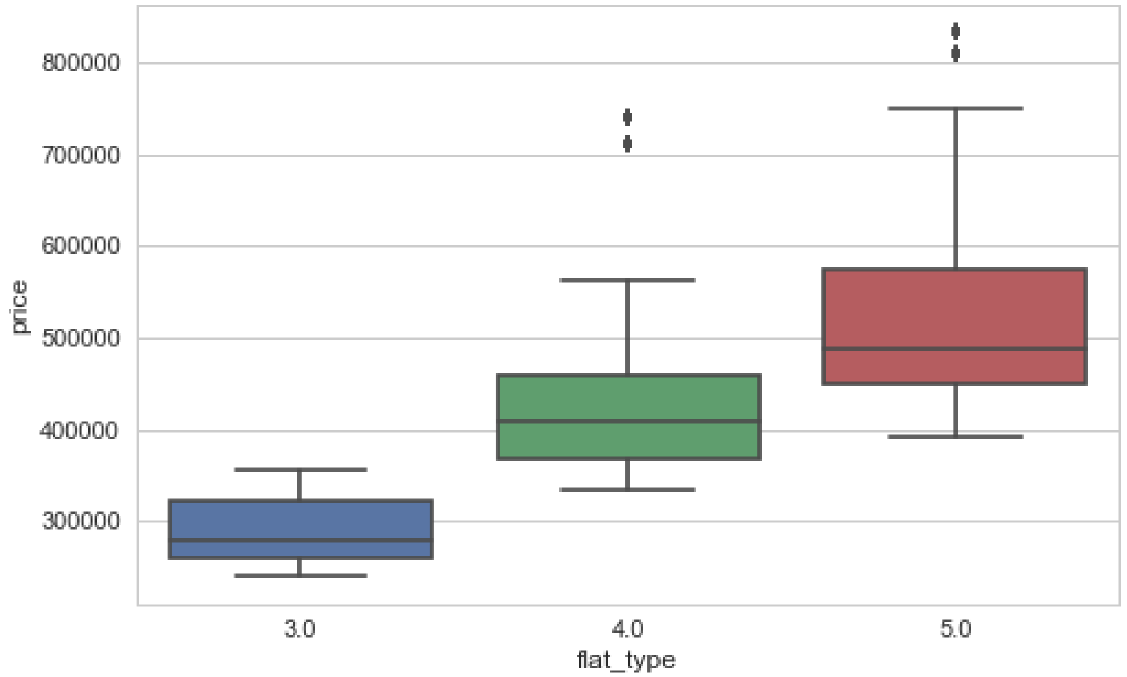


Figure 3. Flat Types box plot to resale price

Price is positively proportional to flat type by number of rooms. The relation is linear.

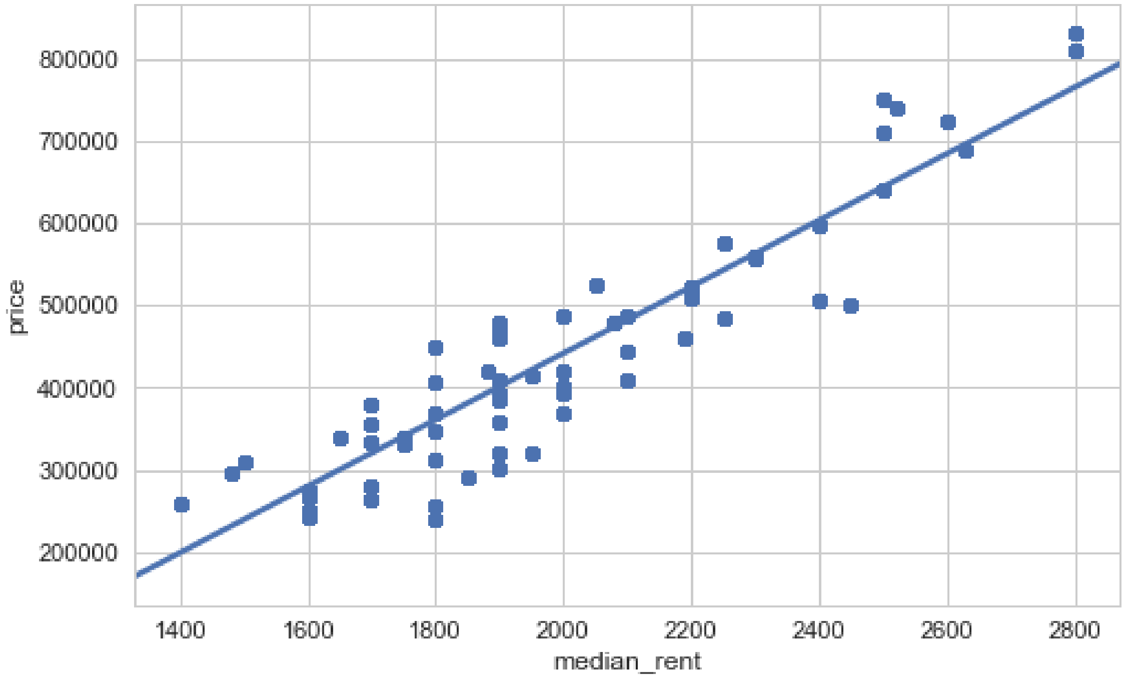


Figure 4. Median Rent to Resale Price

Resale price is also positively proportional to Median Rent in a linear relation as well.

Meanwhile, the next correlated feature of building age is not proportional to resale price.

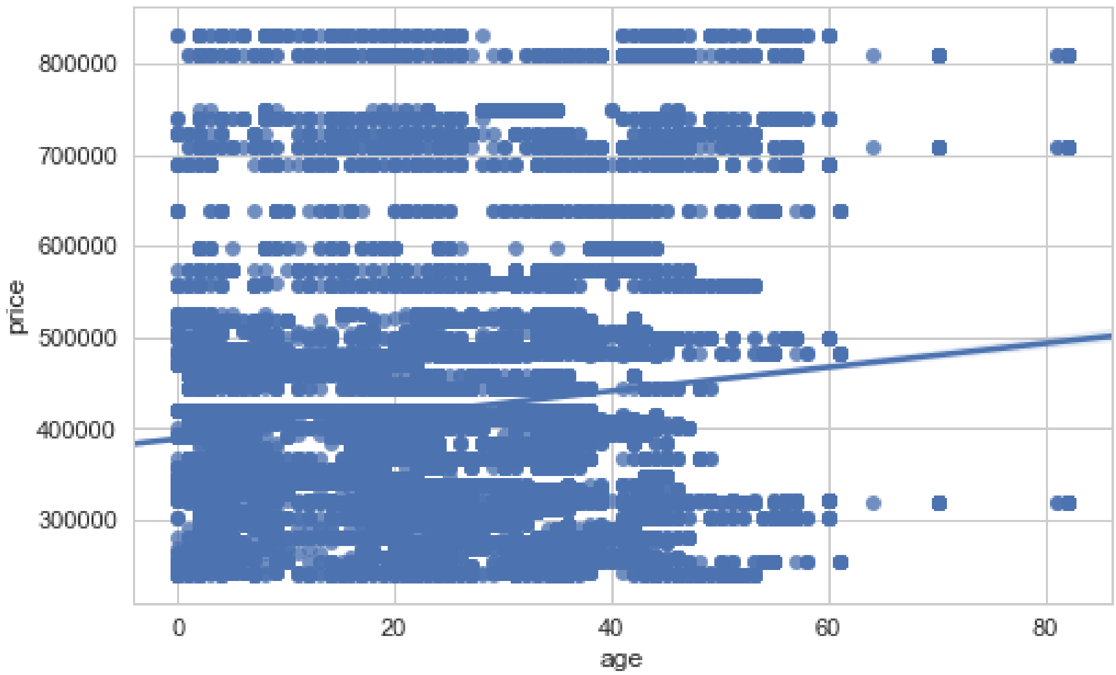


Figure 5. Age to Resale Price

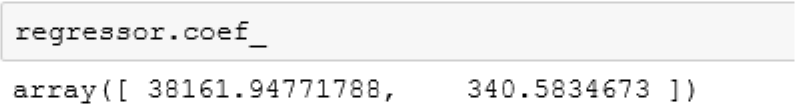
We have split the data set to Train and test set in a ratio of 7:3, which resulting 24187 units and 10367 units.

Result and Discussion

We have run Lasso regression which uses L1 penalty to shrink unnecessary features with coefficients to almost 0. Then with remain features, we ran linear regression to fit on train data set and validate with test data set. The hyper parameter of Lasso ranges widely from 0.1 to 10000 to increase the discrimination power of feature selection. Then we evaluate the model fitted with Mean Absolute Error, Root Mean Squared Error and R2 score of determination.

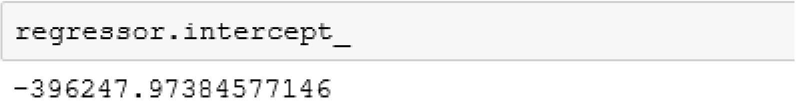
Alpha	feature_selected	feature_names	MAE	RMSE	R2_score
0.1	25	Index(['max_floor_lvl', 'residential', 'commer...	36532.51	45919.30	0.87
1.0	25	Index(['max_floor_lvl', 'residential', 'commer...	36532.51	45919.30	0.87
10.0	24	Index(['max_floor_lvl', 'residential', 'commer...	36532.51	45919.30	0.87
100.0	21	Index(['max_floor_lvl', 'commercial', 'market_...	36539.40	45914.75	0.87
1000.0	8	Index(['max_floor_lvl', 'commercial', 'market_...	36604.45	45952.70	0.87
10000.0	2	Index(['flat_type', 'median_rent'], dtype='obj...	37577.78	46732.99	0.87

The performance of metrics is close with different number of features. However, the least feature used is best, as it is more general and less sensitive which will not cause over-fitting. Also, as suggested by the analysis section, the selected two features: flat_type and median_rent is linear variable which is highly proportional to resale price. This is further proved by the coefficient of determination R2 score, which is 0.87 in range of 0 to 1.



```
regressor.coef_
```

```
array([ 38161.94771788,    340.5834673 ])
```



```
regressor.intercept_
```

```
-396247.97384577146
```

The regression coefficient and intercept are shown above.

Prediction on unknown 9303 units of HDB are available based on above formula. The prediction result is distributed as below:

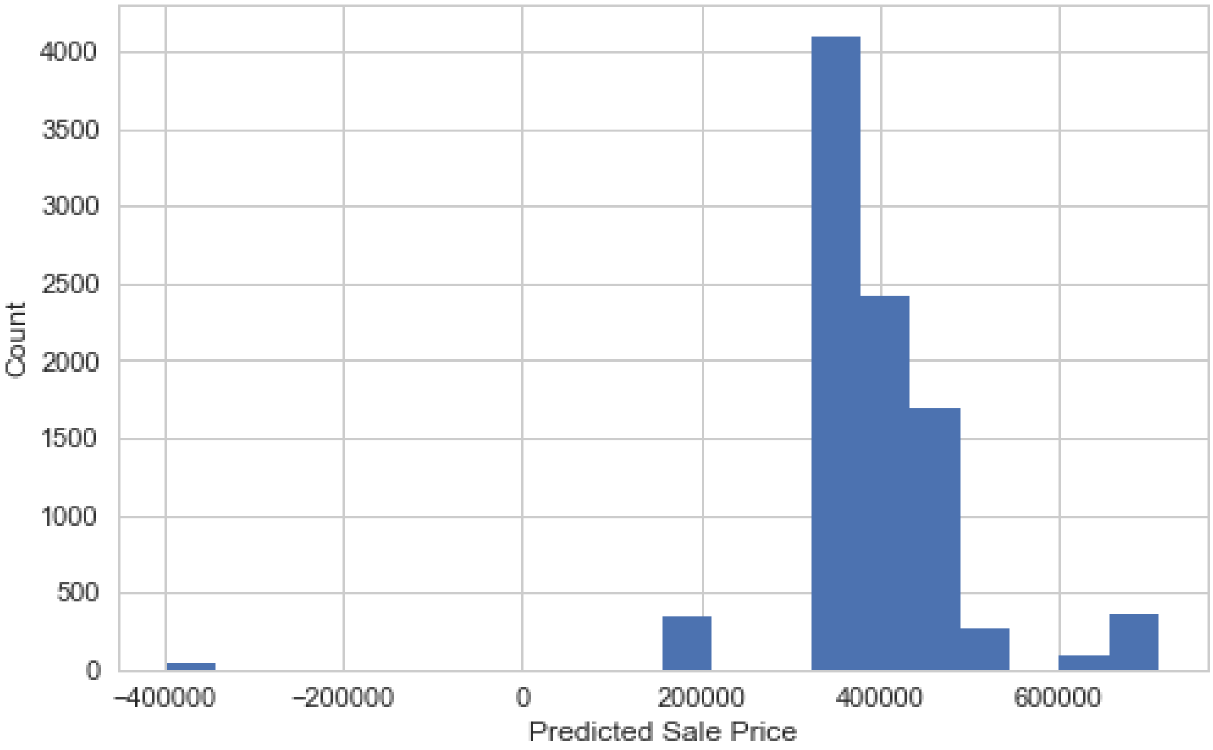


Figure 6. Predicted Sale Price