CS5228 Final Project
Team XYZ

Xingxuan Li ZhiHao Sun Ming Yuan
School of Computing, National University of Singapore
{e0404025,e0025739,e0524793}@u.nus.edu

1. Motivation

As Singapore's public housing authority, the Housing & Development Board (HDB) has provided quality and affordable public housing for generations of Singaporeans. More than 1 million HDB flats have been developed in 26 towns and 3 estates across the island, being home to over 80% of Singapore's resident population [2]. The main objective is to predict the housing resale price with higher accuracy, enabling both buyers and sellers to make a more informed decision with a better understanding of resale market movement. Second, the importance of each feature in predicting resale price is studied. Besides, from literature review, we found that there was a 60% increase in housing price after the 2008 Global Financial Crisis, followed by a series of strict housing policies to drive the housing price down [3]. The impacts of the economic trend and government policies is studied.

The following questions are structured and to be addressed:

- With all necessary information, what is a reasonable resale HDB price for a buyer and seller?
- What information would be of the most importance and usefulness for sellers and buyers in the HDB resale market?
- In alternative for a higher level, how much more should a buyer pay?
- What are the impacts of the economic trend and governmental policies?

2. Data Pre-processing

Understanding and pre-processing data is essential before deciding and training machine learning models.

Originally, there are 16 features given, most of which are geometric and flat intrinsic information. To avoid duplicate information in these features, features are modified and selected with solid assumptions and quality assessment in Section A and features from auxiliary data and external data are added in Section B with justification. Furthermore, feature selection is emphasized with different dimension reduction methods in Section C and different scaling techniques are explained and validated in Section D.

2.1. Feature Modification

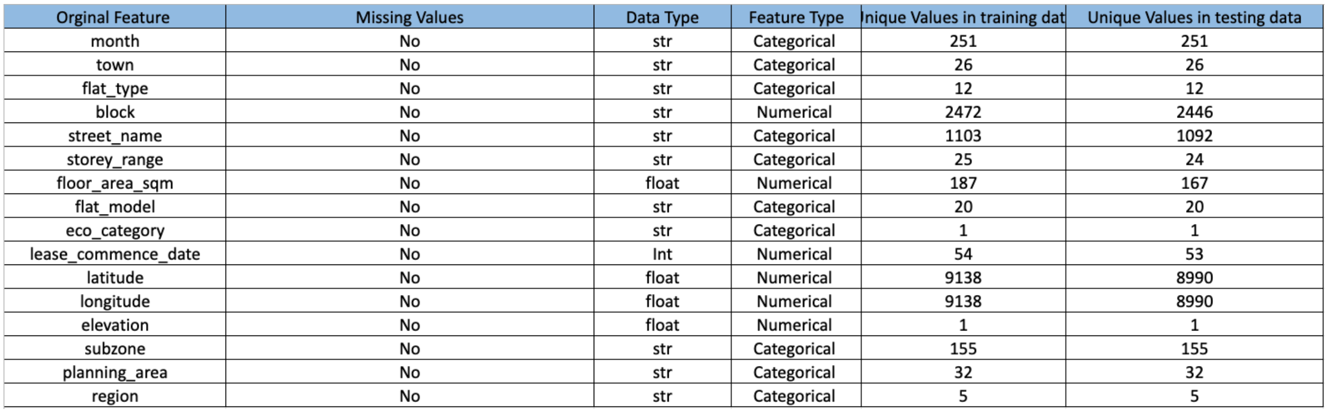
First of all, the quality of the data is assessed as shown in Fig. 1, based on the data type, feature type, and unique values, 7 assumptions are made to modify the features.

- **month**: it is a string format and it will be replaced with two new numerical features, namely 'saleyear' and 'salemonth'.
- **flat_type**: There are 12 distinct values in the **falt_type** while there are in fact only 7 types of flat. Two additional features are created namely, **room_number** and **toilet_number**.The mapping can be found in Tab. 1.

Table 1. Mapping for **room_number**, **toilet_number**

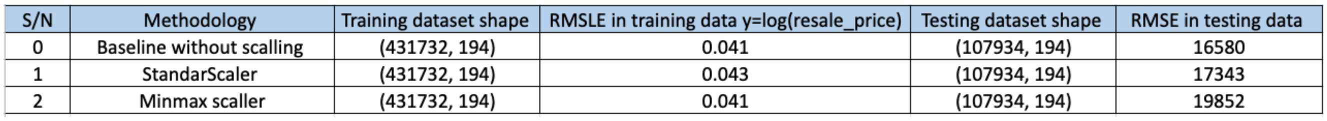
flat_type	room_number	toilet_number
1 room	1	1
2 room	2	1
3 room	3	2
4 room	4	2
5 room	5	2
executive	3	2
multi generation	4	3

- **eco_category**, **elevation**: For numerical features with only 1 unique value, namely **eco_category**, **elevation** will be removed with the advantage of reducing both the dimension and running time.
- **block**, **street_name**: For categorical features with more than 200 unique values, **block** and **street_name** are removed.



Orginal Feature	Missing Values	Data Type	Feature Type	Unique Values in training dat	Unique Values in testing data
month	No	str	Categorical	251	251
town	No	str	Categorical	26	26
flat_type	No	str	Categorical	12	12
block	No	str	Numerical	2472	2446
street_name	No	str	Categorical	1103	1092
storey_range	No	str	Categorical	25	24
floor_area_sqm	No	float	Numerical	187	167
flat_model	No	str	Categorical	20	20
eco_category	No	str	Categorical	1	1
lease_commence_date	No	Int	Numerical	54	53
latitude	No	float	Numerical	9138	8990
longitude	No	float	Numerical	9138	8990
elevation	No	float	Numerical	1	1
subzone	No	str	Categorical	155	155
planning_area	No	str	Categorical	32	32
region	No	str	Categorical	5	5

Figure 1. Data quality assessment.



S/N	Methodology	Training dataset shape	RMSLE in training data y=log(resale_price)	Testing dataset shape	RMSE in testing data
0	Baseline without scalling	(431732, 194)	0.041	(107934, 194)	16580
1	StandarScaler	(431732, 194)	0.043	(107934, 194)	17343
2	Minmax scaller	(431732, 194)	0.041	(107934, 194)	19852

Figure 2. Performance Comparison between different Scaling Methods

- **storey_range**: It is in sting format with 25 unique values in training data and 24 unique values in testing data. Two numerical features 'story_low' and 'story_high' are created to replace the original type.

- **lease_commerce_date** For **lease_commerce_date**, a more reasonable feature **lease_commerce_year** indicating the years the HDB has been in use since the commencement date is created.

- **geometric information** For geometric information, a feature **distance_to_orchard** is created to replace the original **latitude** and **longitude**. In terms of the dimension, **subzone** $>>$ **planning_area** $>$ **town** $>$ **region** and thus **subzone** is chosen to indicate the geometric zone with the maximum information. This assumption is further validated by comparison between keeping **subzone** vs keeping **planning_area**.

2.2. Feature Addition

Besides the original data, 10 more features are added from auxiliary files and external data.

- **mrt_0_5**, **mrt_1_0**, **mrt_1_5**:Three features based on the **longitude** and **latitude** of the flat and available MRT stations. Each of them represents the number of MRT stations within 0.5km, 1km and 1.5km. Intuitively, it is believed that more MRT stations in a closer range, the higher price of the flat will be with the convenience of transportation.

- **mrt_closest**: Apart from the above three features, another feature is created, indicating the distance of the nearest MRT station in km. It helps to distinguish two flats within the same zone but with a different distance to MRT stations.

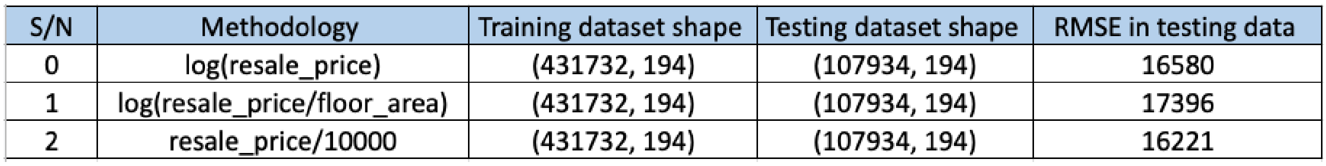
- **shopping_mall_1**, **commercial_1**, **hawker_1**: Similarly, features for shopping mall, commercial centre, and hawker centre within 1 km are created to indicate the accessibility to life supplies and recreation within walking distance.

- **primary_1**, **primary_2**: Another important element for families or even couples who intend to have children is the distance to primary schools. According to Ministry of Education, priority admission is applicable to primary schools and not applicable to secondary schools [1]. Home-to-school distance of 1km and 2km are determining factors in priority admission besides the citizenship. Therefore, the counts of primary schools within 1 km and 2km to the house are included.

- **gold_price**: Considering the importance of housing in the economy to hedge against inflation and provide premises for income-generating activities, monthly gold price data in USD is collected as an external economic indicator from the public source [8].

2.3. Feature Scaling

Feature scaling is essential for machine learning algorithms, especially those that calculate distances between data [4]. In addition, many estimators are designed with the assumption that each feature takes values close to zero or



S/N	Methodology	Training dataset shape	Testing dataset shape	RMSE in testing data
0	log(resale_price)	(431732, 194)	(107934, 194)	16580
1	log(resale_price/floor_area)	(431732, 194)	(107934, 194)	17396
2	resale_price/10000	(431732, 194)	(107934, 194)	16221

Figure 3. Performance of Different Scales of Y

more importantly that all features vary on comparable scales [5]. Particularly, estimators of metric-based and gradient-based usually assume centralized features with unit variances. Notably, decision tree-based estimators are robust and insensitive to arbitrary scaling of features.

Different types of scaling techniques are explored with their pros and cons together with the prediction performance before deciding the best scaling method.

- **StandarScaler** As a very common scaling technique, StandarScaler removes the mean and scales each feature/variable to unit variance with operation performed feature-wise independently. It can be impacted by outliers because it utilizes the estimation of the empirical mean and standard deviation of each feature.

- **MinmaxScaler** Minmax scaler scales features between a given minimum and maximum value, often between zero and one, or so that the maximum absolute value of each feature is scaled to unit size. The advantage is the robustness to small standard deviations of features and preserving zero entries in sparse data.

The results of 2 different scaling and the baseline is shown in Fig. 2.

In addition, different scales in y are explored and compared shown in Fig. 3.

To sum up, scaling on y instead of features produces the lowest RMSE in testing data. This can be explained because most features in the dataset are categorical features and RMSE is chosen as the evaluation criteria. Scaling may improve other performance beyond RMSE.

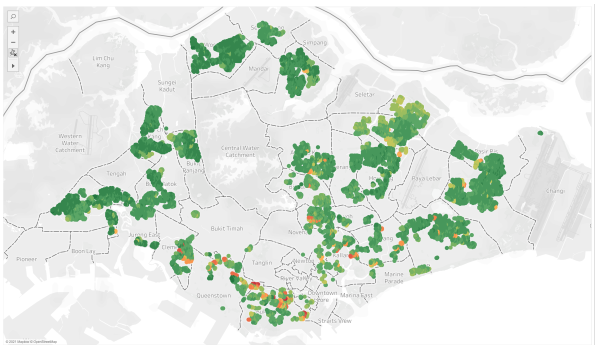


Figure 4. Average housing price distributed in Singapore Map

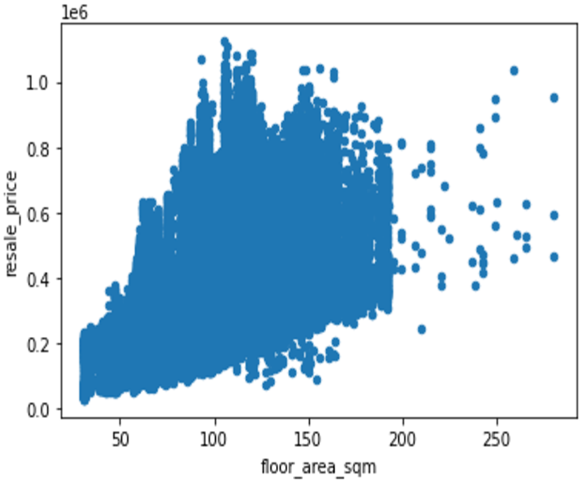


Figure 5. resale price vs floor area

2.4. Data Visualizations

From Singapore's map and the distribution of average housing price per sqm in Fig. 4, red colour indicates higher prices with green indicating lower prices. Geometric location is important because more hot spots are observed in

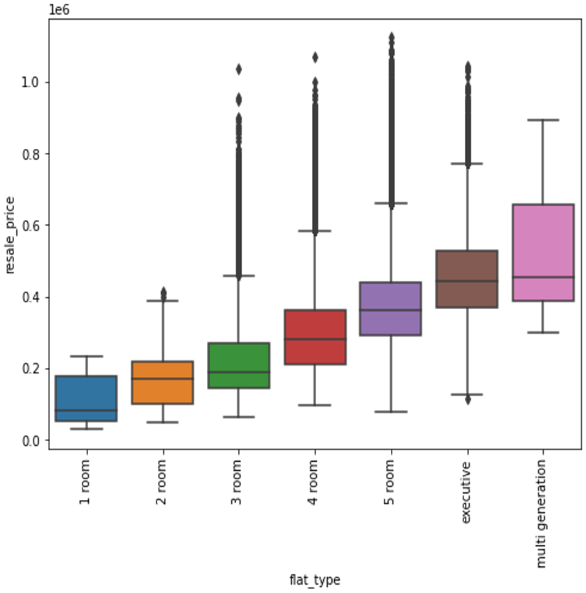


Figure 6. Flat Type

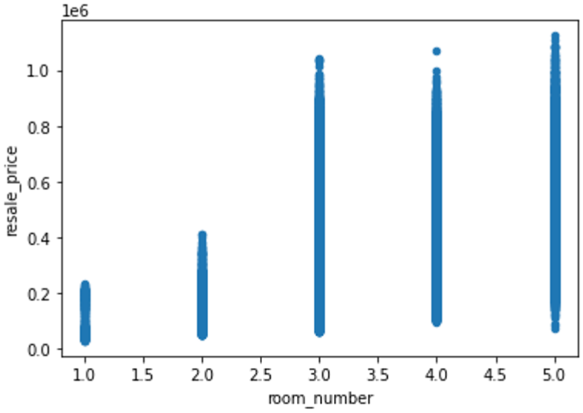


Figure 7. Number of Rooms

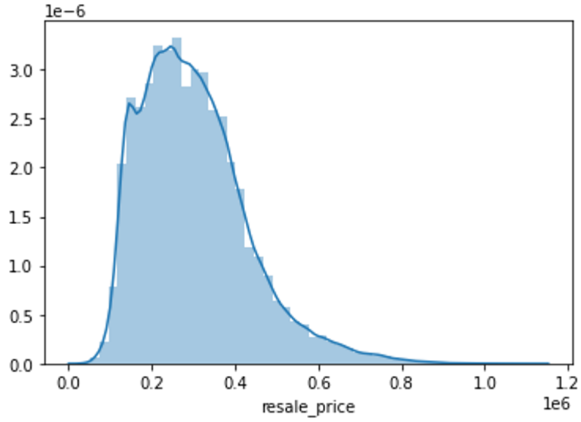


Figure 8. Distribution of Price

specific towns like Clementi, Queenstown and Marina Bay.

From Fig. 5, we can infer that the resale price increases with the floor area. And we can spot if there are any outliers.

From Fig. 6 and Fig. 7, the distribution of the flat price for different flat types is shown. We can infer that the resale price increases with the number of rooms.

From Fig. 8, the distribution of resale price is plotted. This is to understand if imbalance of data happens and sampling techniques can be used to avoid data imbalance.

Fig. 10 is the correlation matrix to study the relationship of the features, indicating if data reduction is required.

3. Models

With feature engineering with validation of different preprocessing methods, different machine learning models are explored for prediction. The evaluation criteria is mainly RMSE and meanwhile, we also compare the accuracy, F1 score to avoid over-fitting and under-fitting.

We have attempted several models for prediction. Below are all our attempts:

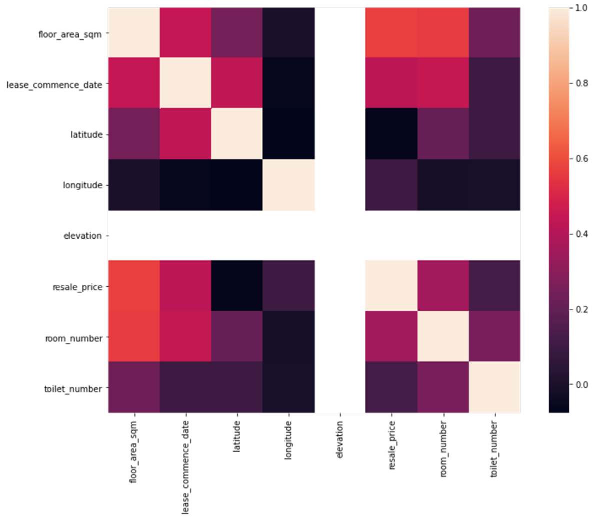


Figure 9. Features Correlations

- **Lasso**

  Lasso model is a Linear model and is an advanced variation of Logistic Regression. The major control factor is alpha value, but however value it is, validation RMSE cannot get better than 0.1, so a single model using Lasso is not considered. Kernel Ridge Regression It is a variation of a Linear model which can learn non-linear relationships, however this model trains very slowly, and cannot get a comparable RMSE as other models, so we gave up this model [9].

- **ElasticNet**

  This is a flexible version of a neural network model. Training on this model is very slow, and it is highly prone to overfitting. We get a training RMSE at 0.04, but validation RMSE at 0.15. We fine-tuned this model for a long time, but failed to overcome the overfitting issue, so we gave up this model [6].

- **Random Forest Regressor**

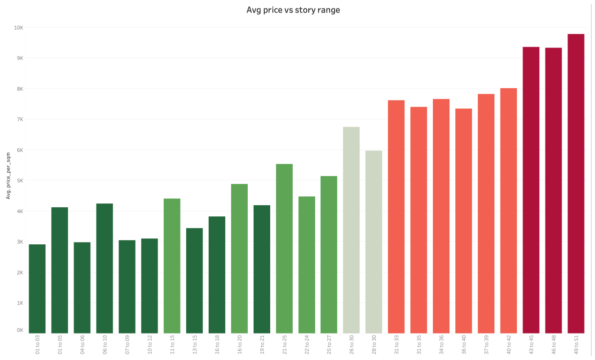


Figure 10. Avg housing price vs floor level

This a tree-based model. This model performs very well with minor parameter tuning. Just adjusted n_estimators a bit, and we managed to get a validation RMSE at around 0.04. It is robust against overfitting, as validation performance is near to training performance. This model alone can achieve a submission score of 17243. We decide to adopt this model, to be assembled with other models.

- **Extreme Tree Regressor**

  This is also a tree-based model. It also performs well compared with Random Forest Regressor. Similar to RF regressor, n_estimator is the major parameter to fine-tune, and min_sample_split and max_depth to fight against overfitting. The single model can achieve comparable performance as an RF regressor, so we decided to keep this model [5].

- **GBDT Regressor**

  This is another tree-based model. Training time is much longer compared with the Random Forest model with equal parameter setting. Its performance is slightly worse than Random Forest regressor. As we already have two tree-based models, to keep the variety of models, we decided to drop this model.

- **Multilayer perceptron , or neural network**

  Multi layer perceptron is a type of neural network model. It can fit a non-linear relationship, and is not a tree-based model. The most important parameter to fine-tune is the shape of the network. After several rounds of trial-and-error, we found that the layers of the network shouldn't be too much, as two layers perform best. And as our number of features is not too high, only around 80, each layer has a neuron size of 100 is proper. In this way, we found that the MLPClassifier with layer size (100, 100) performs best, and a result of 17152 is obtained. This is good enough to be a candidate for ensemble.

4. Model Ensembling

After we obtained three models that can perform comparably well, we decide to ensemble to obtain a better result. As more wisdom gives suggestion, final result will be more prone to overfitting and should perform better. Our candidate models are Random ForestRegressor, Extra Tree Regressor and neural network. This is a good combination as both tree-based model and neural network types of models are considered. A variety of models are kept. We have attempted three different ways of model ensembling:

- **Simple average**

  Simply taking the average of three models gives a slightly improvement of performance. Result submission achieves 16945 on the leaderboard.

- **Weighted Sum**

  We adjusted the weight of each model, and found out a weight of 0.3, 0.3, and 0.4 is given for RF, Extra tree, and neural network. As the neural network model is a different model, higher weight is given to it. The final performance on the leaderboard is 16221, which is our best result so far.

- **Linear Regression learning of weights**

  We trained a simple Linear regression with three weights, based on the validation score of every model. The result gives two tree-based models higher weight, each around 0.4, as they perform a bit better than neural network models. However, this way reduces the variety of models, and thus ensemble result is not so well as weighted sum, only 17015 is obtained for the final result.

5. Results and Analysis

5.1. Feature Importance

Since we are using an ensemble model of EtraTreeRegressor, RandomForestRegressor and MLPRegressor, we try to analyse the feature importance by looking at both ExtraTreeRegressor and RandomForestRegressor. Fig. 11 plots the feature importances of each feature of the model. And Tab. 2 and Tab. 3 list the top 10 most important features for both models.

From Fig. 11, we can observe that around 10 features have importance significantly higher than all the other 70 features which are nearly equal to 0. And in terms of model, RandomForestRegressor amplifies the importance of most of the top 10 features compared with ExtraTreeRegressor.

From Tab. 2 and Tab. 3, we observe that both ExtraTreeRegressor and RandomForestRegressor have similar top 10 features. And some of the features we created from auxiliary data as well as external data are also listed. For example, **mrt_closest**, **distance_to_orchard** and **gold_price**. This proves some of our hypotheses: transportation convenience, distance to the central area, and global economic environment are critical to the housing price.

5.2. Study Impact of Economic Conditions and Policies

From Tab. 2 and Tab. 3, it has been observed that **year** is an important feature to the resale price. And this is because the global economy is a key factor to the housing price. To further prove this point, we plot the average housing price

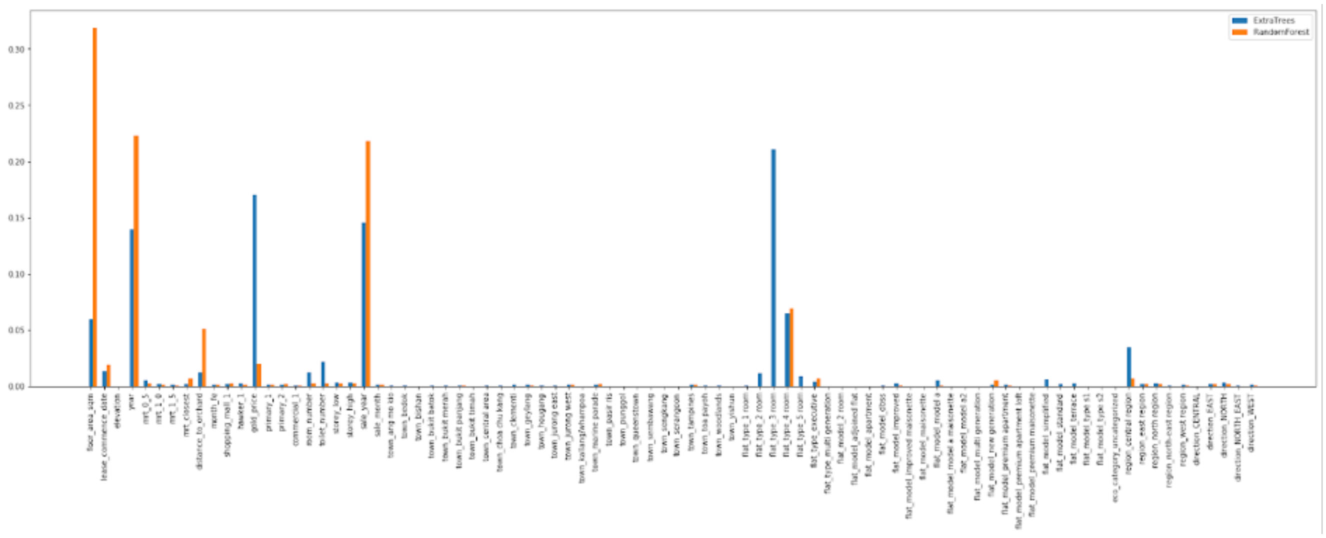


Figure 11. All Feature Importance for ExtraTreeRegressor and RandomForestRegressor

Table 2. Top 10 Important Features ExtraTreeRegressor

Rank	**ET_feat**	**ET_score**
1	flat_type_3_room	0.211
2	gold_price	0.171
3	sale_year	0.146
4	year	0.14
5	flat_type_4_room	0.065
6	floor_area_sqm	0.059
7	region_central_region	0.035
8	toilet_number	0.022
9	lease_commence_date	0.014
10	distance_to_orchard	0.012

Table 3. Top 10 Important Features RandomForestRegressor

Rank	**RF_feat**	**RF_score**
1	floor_area_sqm	0.319
2	year	0.223
3	sale_year	0.218
4	flat_type_4_room	0.069
5	distance_to_orchard	0.051
6	gold_price	0.02
7	lease_commence_date	0.019
8	mrt_closest	0.007
9	region]_central_region	0.007
10	flat_type_executive	0.007

and gold price in one graph. We observe that the changing trends are almost identical for the two. As the gold price can be seen as an indicator of the global economic environment, the graph proves our hypothesis.

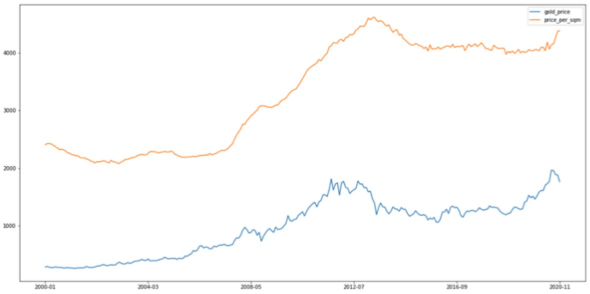


Figure 12. Housing and gold price

To study the impact of housing policies and economic conditions on the housing price, housing price per sqm and gold price in USD is plotted in Fig.12 Noticeably, a similar pattern is observed between the two trends. In 2008, both housing prices went down due to the Global Financial Crisis and observed a dramatic increase up to 60% due to the housing bubble and increasingly unaffordable housing. The government has intervened with a series of policy controls to cool down the market since 2009 and in 2011, more effective policies were introduced, after which the housing prices started to decrease and remained stable. These policies were relaxed in 2017 March by the government to stimulate economic growth, after which the housing prices started to increase.

5.3. Recommendation on Price per sqm Increment in Alternative For a Higher Level

In the resale market, interestingly story is one of the important factors for buyers to consider because story level is related to coolness, convenience, noises [7]. In alternative for a higher level, the price increase is studied extensively with 40k transaction data. As shown in Fig. 10, there is an obvious increment in the story level with the average price per sqm. Based on this figure, recommendations can be made for buyers, for example, when they compare two options, one-story level in 46-48 and the other in range 49-51, a 4.9% increase in the unit price per sqm is reasonable.

6. Conclusion

With an ensemble model, the RMSE achieved is 16221in the testing data. The selected model will be integrated into an application for deployment to users, so that users are able to get an estimate of the HDB Flat resale price, quickly and cost-free. Limitations are analysed for future work and improvement.

Besides, the impact of economic conditions and government policies on housing prices empathizes with interesting insights and strong correlation found. Furthermore, with the significance of story range when buyers deciding the house to buy, a comprehensive study is carried out to recommend to buyers how much they should increase in the alternative for a higher floor.

7. Contribution

For this project, the efforts are equally contributed by each of the team members. Yuan Ming and Xingxuan are mainly responsible of EDA, feature engineering, and report writing. Zhihao is responsible of model selection and training.

References

[1] Ministry of Education Singapore. How distance affects priority admission. Available at https://www.hdb.gov.sg/about-us (2021/04/17).

[2] Government of Singapore. About us - housing development board (hdb). Available at https://www.hdb.gov.sg/about-us (2021/04/17).

[3] Sock Yong PHANG. Singapore's housing policies: 1960-2013. Available at https://ink.library.smu.edu.sg/cgi/viewcontent.cgi?article=2543&context=soe_research (2013/11/01).

[4] ScikitLearn. Compare the effect of different scalers on data with outliers. Available at https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html (2021/04/17).

[5] ScikitLearn. sklearn.ensemble.extratreesregressor. Available at https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesRegressor.html (2021/04/17).

[6] ScikitLearn. sklearn.linear_model.elasticnet. Available at https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html (2021/04/17).

[7] The Stratistimes. Many factors considered before price of hdb flat is set. Available at https://www.straitstimes.com/forum/letters-in-print/many-factors-considered-before-price-of-hdb-flat-is-set (2017/04/27).

[8] theDIG95. Gold-prices-ann. Available at https://github.com/theDIG95/Gold-prices-ANN/blob/master/data_importer.py (2021/04/17).

[9] Wikipedia. Lasso(statistics). Available at https://en.wikipedia.org/wiki/Lasso_(statistics) (2021/04/17).