Predicting HDB Resale Prices

ST4248 Project Group C4

This project aims to predict HDB resale prices and identify key factors that affect the resale price. This was done by engineering additional features for each HDB and selecting crucial features using an ensemble of feature selections. Six regressor models were fitted on the data to predict both price and price/sqm. The performance of each model was evaluated and important features were obtained. The result shows that XGBoost price/sqm model performs the best and MRTs, malls, remaining lease, and total resales in town are the top key features.

Clifton Felix, A0219735X

Kathy Fresilia Ijaya, A0200719M

Nicholas Russell Saerang, A0219718W

Teo Zay We, Simon, A0206249E

1 Problem Introduction

Singapore Housing and Development Board (HDB) flats are resold at various prices. The resale price can be affected by many factors, such as floor area, lease year, and many more. This project aims to perform prediction, which is to predict HDB resale prices, and inference, where we analyze how each feature contributes to the resale price. The model and findings can then be used to assist property agents and HDB owners in determining the resale price of HDB so as to attract buyers while still maximizing profits.

2 Dataset Description

The dataset is taken from government resale flat data at data.gov.sg/dataset/resale-flat-prices, managed by the Singapore Housing Development Board (HDB). Exactly 4410 resale transactions are taken from January to February 2023. Only these two months are taken for the dataset to eliminate the influence of time series on the data. The dataset has 11 variables: month, town, flat_type, block, street_name, storey_range, floor_area_sqm, flat_model, lease_commence_date, remaining_lease, and lastly resale_price as the response variable.

3 Exploratory Data Analysis

We generated four plots shown in Figure 1 as a preliminary analysis. The plot of resale prices against flat floor areas shows a noticeable upward trend, meaning that flats with higher floor areas tend to have higher resale prices. A similar conclusion is found when we plot the resale price against the storey, where flats with higher storeys tend to have higher prices. Another upward trend can be seen as we plot the resale price against the flat type, where flats with better types tend to have higher resale prices. Finally, plotting the resale price against the town gives us various distributions which imply that there is no clear town that has higher HDB prices than other towns and thus motivates us to perform further feature engineering.

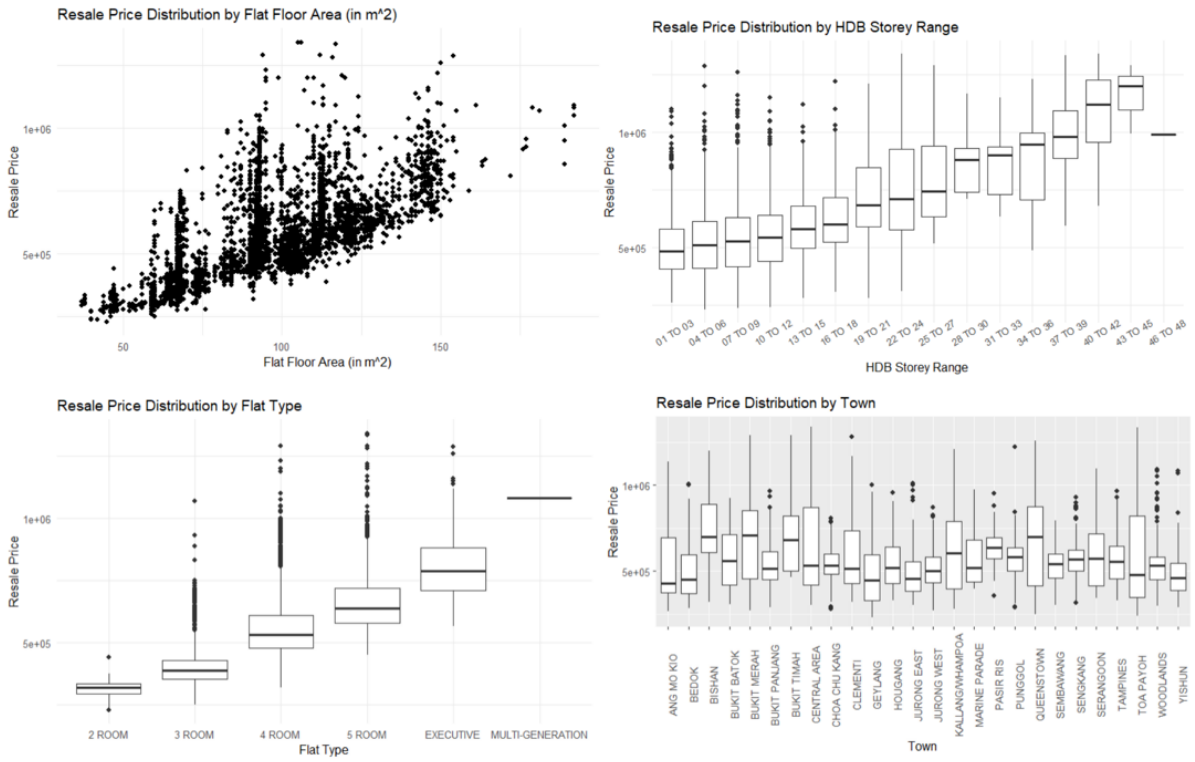

Figure 1: Exploratory Data Analysis Plots

4 Feature Engineering

After researching some property websites, we found out that HDB prices depend on some factors such as accessibility to transportation, education, and also entertainment. Hence, for each HDB, we engineered the nearest, distance to the nearest, and total nearby (within 1km) MRTs, bus stops, schools, primary schools, and malls.

MRTs, schools, and malls data were retrieved from data.gov.sg, while bus stops data was retrieved from data.busrouter.sg. To calculate their distances to the HDB unit, we used OneMap SG API to obtain the latitude and longitude for all MRTs, schools, malls, bus stops, and HDBs. Then, we split the data into 80% train and 20% test. After splitting, we added 3 more attributes that reflect the supply and demand, such as: total resales in town, block, and street. Subsequently, we one-hot encoded the categorical variables and standardized the predictors to mean 0 and variance 1. In the end, we have 4181 variables including the response variable.

5 Feature Selection

Since the data is of high dimension, we need to perform dimensionality reduction to remove less important features and improve model performance. However, since inference is one of our goals, we didn't do factor-based (PCA, Factor Analysis, etc) or projection-based (ISOMAP, t-SNE, etc) dimensionality reduction as it makes our models less interpretable for inference. Hence, we only performed feature selection methods on our data.

5.1 Filter Method

5.1.1 Variance Threshold

Firstly, we removed columns with variance 0 as it doesn't help us with prediction. In this stage, 233 predictors were removed.

5.2 Wrapper Method

5.2.1 Forward Selection

The first Wrapper method we applied was Forward Selection that selects 100 best predictors. Forward Selection begins with an empty model and adds in variables 1-by-1 based on RSS value until we get 100 predictors.

5.2.2 Recursive Feature Elimination

Another Wrapper method we used was Recursive Feature Elimination that fits a model and removes the weakest features sequentially based on the feature importance until we have 100 predictors. We used 2 models: Ridge Regression and Gradient Boosting Regressor as the estimators of the RFE. For Ridge Regression, we removed 1 feature at each step, while for Gradient Boosting Regressor, we removed 5% of the features at each step as it takes very long to remove 1 feature at each step.

5.3 Embedded Method

5.3.1 Select K Best

We also implemented Select K Best Embedded method to select 100 predictors with the highest scores based on the scoring function. We used 2 different scoring functions: F Regression and Mutual Info Regression, because F Regression detects the linear relationship between the predictor vs response, while Mutual Info Regression can capture the complex, non-linear relationship of each predictor vs response.

5.3.2 Lasso Regression

Lastly, we also used Lasso Regression to select 100 nonzero variables.

5.4 Ensemble of 6 Feature Selections

After removing 233 features using Variance Threshold, we built an ensemble of 6 feature selections using the Wrapper and Embedded methods specified above. Ensemble of 6 feature selections specified above is useful in determining the best set of features as it combines the strength of multiple methods, which can increase model robustness and performance as well as reduce multicollinearity (in this case is done by Lasso and Ridge) and overfitting.
With this ensemble, we selected variables that are chosen by at least 3 of the 6 methods. In the end, 74 final predictors were chosen. Furthermore, the top 5 variables selected by all 6 methods are floor_area_sqm, total_resales_in_town, nearest_mrt_dist, remaining_lease, town_BUKIT MERAH.

6 Models

For the models, we used the price variable as the response. In addition to price, we also wanted to verify our hypothesis that variable floor_area_sqm has multiplicative effects on price. Hence, we created other models with price/sqm as the response variable.

All models hyperparameters were tuned using GridSearchCV with the exception of linear regression and neural networks. Additionally, all models except linear regression utilized scaled data.

6.1 Linear Regression

After converting all categorical predictors to one-hot variables, we fit 2 linear regression models. One on standardized data and one on unstandardized data. There was no noticeable difference in performance which is reasonable since linear regression is scale invariant. Since the model fit on unstandardized is more interpretable (in terms of coefficients), we decided to use that instead.

6.2 ElasticNet

ElasticNet is the same as linear regression but the objective function we are trying to minimize has an addition of the L1 penalty and the L2 penalty. For both price and price/sqm models, the hyperparameter α is 0.001 and the mixing parameter, which determines the ratio between the L1 penalty and L2 penalty, is 0.5, meaning both penalties contribute equally to the objective function.

6.3 Neural Network

Neural Network used for this project contains one input layer, three hidden layers with ReLU activations, and one output layer where the 3 hidden layers have 64, 32, and 16 neurons.

6.4 Random Forest

Random Forest Regressor was implemented with 500 trees for both price and price/sqm models. For the price model, nodes are expanded until a maximum depth of 20, while for price/sqm model, they are expanded until there are not enough samples to split.

Metrics	Model	LR	EN	NN	RF	GBR	XGB
RMSE	Price	54316	52786	43526	46939	50644	38256
	Price/sqm	49426	49330	44579	42254	35003	34987
MAPE	Price	7.81%	7.53%	5.3%	5.53%	6.19%	4.53%
	Price/sqm	6.62%	6.65%	5.71%	5.09%	4.34%	4.40%
Adj R^2	Price	87.51%	89.77%	92.42%	91.91%	87.38%	94.14%
	Price/sqm	89.40%	91.07%	92.06%	93.45%	94.75%	95.00%

Table 1: Model Performances

6.5 Gradient Boosting Regressor

Gradient Boosting Regressor produces a predictive model from an ensemble of weak predictive models by fitting new models sequentially on residuals of the previous models. It uses gradient descent to minimize the residuals. Using GridSearchCV, the optimal hyperparameters we found were learning rate = 0.01, maximum tree depth = 10, number of boosting stages = 1500, fraction of samples to be used for fitting individual base learners = 0.2.

6.6 XGBoost

The last model we implemented was XGBoost, a gradient-boosted model for supervised learning. XGBoost is very similar to Gradient Boosting Regressor, except that it's optimized for parallel processing, faster training, and better performance. The hyperparameters used for XGBoost price model were colsample_bytree = 0.8, eta = 0.1, max_depth = 10, min_child_weight = 5, while for price/sqm model were colsample_bytree = 0.8, eta = 0.1, max_depth = 10, min_child_weight = 1.

7 Evaluation

To fairly compare the metrics of price and price/sqm models, we first converted the price/sqm prediction back to price before calculating the metrics.

From Table 1, it can be seen that the RMSE (Root Mean Square Error) scores are not too bad, considering the price range for the units is between S$200,000 to S$1,000,000. This is further

Variables	Coefficients
Intercept	4246.0972
Nearest MRT distance	-392.0935
Total nearby MRTs	87.7080
Nearest mall distance	-159.0245
Remaining lease	64.8747
Total resales in town	-5.1221

Table 2: Linear Regression Top Features

supported by the MAPE (Mean Absolute Percentage Error) values, where all are below 10% and some even reach below 5%, which indicate very good performances of the models employed. Moreover, adjusted R^2 was used as a metric instead of the standard R^2 because price/sqm models have 1 less predictor, which is floor_area_sqm.

The results for the models are better when the response variable is price/sqm as compared to price, except for the neural network model. However, it can be noted that the differences between the metrics of the two response variables using Neural Network are the least compared to other models.

The best result was obtained when using XGBoost with Price/sqm as the response variable when scored using RMSE and Adjusted R^2. When assessed using MAPE, the best value is obtained when using Gradient Boosting Regressor price/sqm model. Overall, it can be concluded that tree methods with gradient boosting give the best results compared to linear or other non-linear models.

8 Learnings

We then chose our top-performing and most interpretable models to infer what factors really contribute to HDB resale prices. Since all price/sqm models performed better in general, we will be using that to evaluate.

8.1 Linear Regression

Table 2 shows the intercept and top 5 significant features for linear regression (identified using p-values). From the intercept coefficient, we expect an HDB flat to have a starting price/sqm of $S\$4246$ when all other features are 0. The next few features have the following impacts on the price:

1. **Nearest MRT distance**: If the nearest MRT distance increases by 1km, the price/sqm decreases by $S\$392$. Having to walk 2km to the nearest MRT instead of 1km can be much more inconvenient, hence the stark drop in price.

2. **Total nearby MRTs**: If the total nearby MRTs (within 1km) increases by 1, the price/sqm increases by $S\$87$. Choosing from more MRT stops and possible different MRT lines makes travel much more convenient.

3. **Nearest mall Distance**: If the nearest mall distance increases by 1km, the price/sqm decreases by $S\$159$. Having to walk 2km to the nearest mall instead of 1km can be much more inconvenient, hence the stark drop in price.

4. **Remaining lease**: If the remaining lease increases by 1 year, the price/sqm increases by $S\$64$. This is because a flat with a higher remaining lease is much more valuable.

5. **Total resales in town**: If the total resales in town increase by 1, the price/sqm decreases by $S\$5$. This might be because every new resale increases the supply, driving down the price.

8.2 Top 5 Features by Importance

Then, according to our top 3 performing models (Random Forest, Gradient Boosted Regression, XGBoost), the common top 5 features (in no particular order) are: remaining lease, nearest MRT distance, total resales in town, nearest mall distance, and total resales in street.
This is in line with the top 5 features selected by feature selection. The only major difference is that feature selection included town_Bukit_Merah as one of the top few features. However,

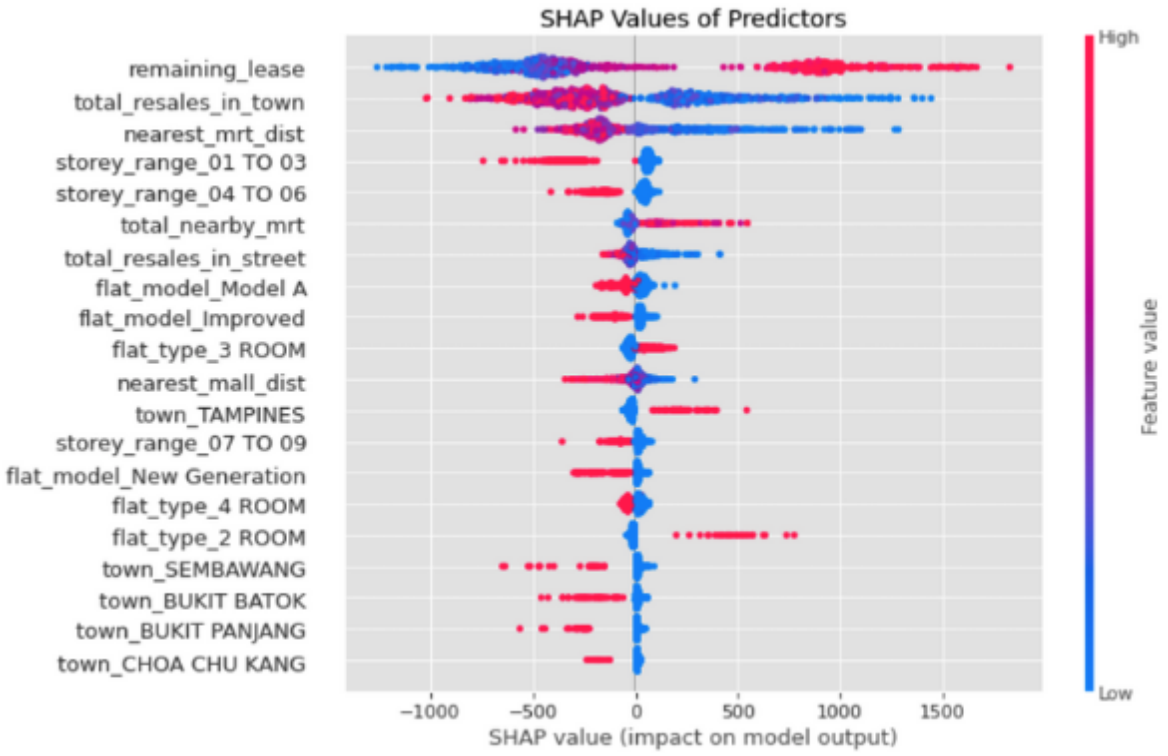

Figure 2: XGBoost SHAP Values

this was not selected by our top 3 models most likely because the resale prices in every town are varied making this feature less important than others.

8.3 XGBoost SHAP values

Lastly, we would like to know how each feature affects the resale price using Shapley values. SHAP, which stands for SHapley Additive exPlanations, measures how much a certain feature X contributes to the overall prediction. This is measured by taking random subsets of features (that do not include feature X) and then adding feature X, and measuring the difference in predicted values.

Each dot seen in Figure 2 is the Shapley value for each data point from training. When the red dots are on the right, the value of the predictor is directly proportional to the resale price/sqm. On the other hand, when the red dots are on the left, the value of the predictor is inversely proportional to the resale price/sqm.

Add to overall price/sqm	Subtract from overall price/sqm
Remaining lease	Nearest MRT distance
Total nearby MRTs	Nearest mall distance
Floor number > 20	Total resales in town

Table 3: Top 3 Positive and Negative Features

From Figure 2, we can infer the following:

1. A higher remaining lease leads to a higher price/sqm.
2. The lower the total number of resales in town, the higher the price/sqm.
3. The nearer the MRT, the higher the price.
4. HDBs located on floors 1 - 3, 4 - 6, and 7 - 9 tend to have a lower price/sqm than flats on higher floors.

9 Conclusion

To sum up, we have collated the top 3 features (by importance) that add to price/sqm and subtract from price/sqm which can be seen from Table 3.

Throughout this project, we have shown that we can quite accurately predict the price/sqm for HDB resale flats using the original and engineered features. We have also inferred the most important features that add to and subtract from the HDB resale price. With this model, we hope that property agents or sellers can have a better benchmark on the price they should set when reselling an HDB unit.

10 Future Work

Future work for this project includes:

1. Building models with price/remaining lease as response variable to study whether remaining lease has a multiplicative effect on the resale price or not.
2. Engineer more features that may affect HDB resale prices (e.g. distance to CBD area)