IS424 Data Mining and Business Analytics

Project Report - HDB Resale Price Prediction

Prepared for: Professor Steven Hoi

Date: 13 November 2017

Section	G1
Group	4
Members	Matthew Ang Wei Ming Nguyen Dang Thanh Ha Shilpa Suresh Ya Min Nyi Nyi

1. Executive Summary

In recent years, there has been increasing interest and literature in using machine learning techniques to analyse and forecast house sales prices. However, the field remains largely unexplored. Moreover, there exists no prior housing price study using machine learnings within the Singapore context. Therefore, we would like to perform a data-driven analysis on Housing Development Board (HDB) resale prices, to deliver accurate and clear explanations of Singapore housing prices. In this report, we will be focusing on the HDB resale prices of Singapore.

Our data sets contain the HDB resale prices from January 2007 to July 2017 with the explanatory variables that affect the HDB resale prices such as sales month, lease year, town, flat type, flat model, block, street name, storey range and floor area. Since the geographical location plays an important role in determining the resale prices, we used the Google Maps Geocoding API on the flat's block and street name to deliver latitude and longitude for each HDB block. We then incorporated such locational features into the model in the calculated Euclidean distance from amenities such as MRT stations and key locations such as Central Business District. Since the macroeconomic data also plays a crucial role in determining the HDB prices, we also used the variables such as monthly prime lending rates from MAS and monthly manufacturing Purchasing Manager's Index from Singapore Institute of Purchasing and Materials Management to deliver more precise result.

After processing the data, building three different models and performing data validation, we have concluded that linear regression is the best for our case. The reason for this conclusion is based on the outcome of graphical residual analysis and the assessment on how close the predicted and actual resale prices are. For graphical analysis, the regression model demonstrates that the relationship between the explanatory variables and response variable is a statistical relationship because of the random residual distribution. For predicted vs actual value assessment, the resulting regression slope of the linear regression model is the closest to 1 among all models, which suggests that linear regression is the best fit model for this real estate data.

The results of our model can be used by the Singapore government, financial institutions, property investors, or even real estate brokers—to analyse and forecast future housing resale prices with heightened accuracy and confidence.

2. Technical Summary

To have a better prediction of the HDB prices, we used 4 datasets to train the model and predict the HDB prices using the trained model. The datasets we used are as follows:

1. resale-flat-prices.xlsx – HDB resale records with variables affecting the prices from January 2007 to July 2017
2. station-mappings.xlsx - how the distances from each flat to MRT evolve over time
3. addresses.xlsx - geocoded information of the HDB addresses
4. macro data.xlsx - macroeconomic indicators that possibly affect resale prices

We will be applying Knowledge Discovery from Data (KDD) process to build a prediction model to accurately predict the HDB prices. We validated the accuracy of the model using the residual analysis and other accuracy metrics to assess the predictive power of the models, and to estimate their forecasting accuracy in a future context with developments of new MRT stations.

To predict the future HDB prices accurately, we first performed the data cleaning on the data sets. This is to ensure that there are no outliers and missing values in the data since they can greatly affect the accuracy of our model. After data cleaning, we performed data transformation and discretization of data where applicable. This stage was followed by running 3 different prediction models - Multiple Linear Regression, Random Forest, and XGBoost. Next, we performed model validation which provides the information necessary to choose the most accurate model which is used in the final predictor which takes in user defined values as the inputs and outputs a predicted resale price.

3. Introduction & Problem Definition

Traditionally, regression analysis of hedonic housing price models has been used to predict its sale price. Model results are then used to analyse real estate portfolios or mortgage-lending decisions by financial institutions. However, there is no prior housing price study using machine learning within the Singapore context. Although there are some online articles and domain knowledge on how factors such as MRT station proximity or district reputation can affect the housing prices, we would like to deliver more accurate and deeper insights on what factors affect Singapore housing prices and how these factors affect the prices using various machine learning techniques. Our goal is to identify the different factors affecting the HDB resale prices and develop the prediction model using those factors to forecast the HDB sales accurately.

4. Approach

4.1. Data Cleaning and Compilation

The input data was compiled from multiple sources and cleaned and transformed to arrive at a suitable train and test dataset. The sources and accompanying data are listed below:

Data	**Source**	**Variables**	**Variable type**
HDB Internal Data	www.data.gov.sg	Date (YYYY-MM format) Town Flat type Block Street name Storey Floor area (sqm) Flat model Lease commence year Resale price Address Storey (normalized)	Ordinal (pre-parsed) Nominal Nominal Nominal Nominal Ratio Ratio Nominal Interval Ratio Nominal Ratio
MRT Station Data	www.lta.gov.sg	Address Date of change Closest 3 MRT stations Distance to closest 3 MRT stations No. of stations within 500m	Nominal Interval Nominal Ratio Ratio
Location Data	Google Geocoding API	Address Latitutde Longitude Distance to CBD	Nominal Interval Interval Ratio
Macroeconomic Data	www.mas.gov.sg SIPMM	Date Mortgage Rate Purchasing Manager's Index	Interval Ratio Interval

4.1.1 Data Transformation

Variables	Transformation
resale_price	Natural logarithm
flat_type	Binarization (one-of-k encoding)
block	Merged to become an address, and geocoded
street_name	
storey_range	Converted to “storey_normalized” by taking the average of range
flat_model	Binarization
town	Binarization
year	lease_age = year - lease_commence_date
lease_commence_date	
Stations within 500m	Generated using the guiding Stations Dataset

We did not convert the flat type to number of rooms, because although they are named as such, they do not actually truly represent the number of rooms in the apartment, as seen from the following table by HDB:

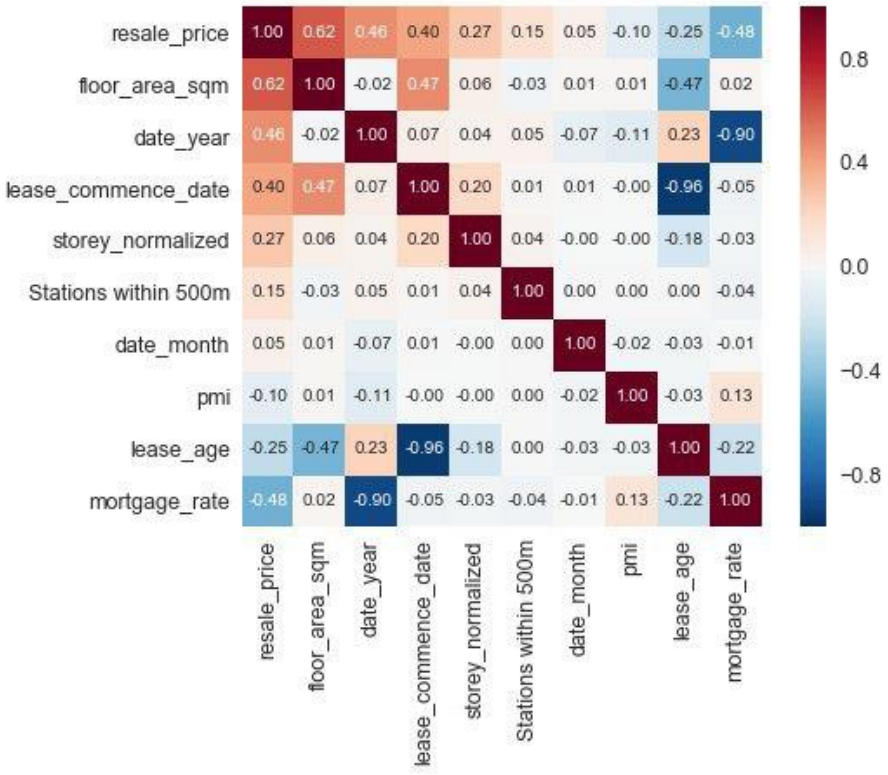
HDB Flat Types	2-Room Flexi	3-Room	4-Room	5-Room	3Gen	Executive Flat
Approx. floor area (square metres)	36 and 45	60 to 65	90	110	115	130
Total no. of bedrooms	1	2	3	3	4	3
Total no. of bathrooms	1	2	2	2	3	2

4.1.2 Stations Dataset

The stations dataset is prepared by making use of the MRT Station Data (Name, Establishment Date) and the Location Data obtained from the Google Geocoding API. Calculations are performed to come up with the number of MRT Stations that are located within 500m of each HDB. This dataset also contains information about the MRT Stations that are due to be constructed in the future and hence, the prediction model adjusts for future developments such as new MRT stations.

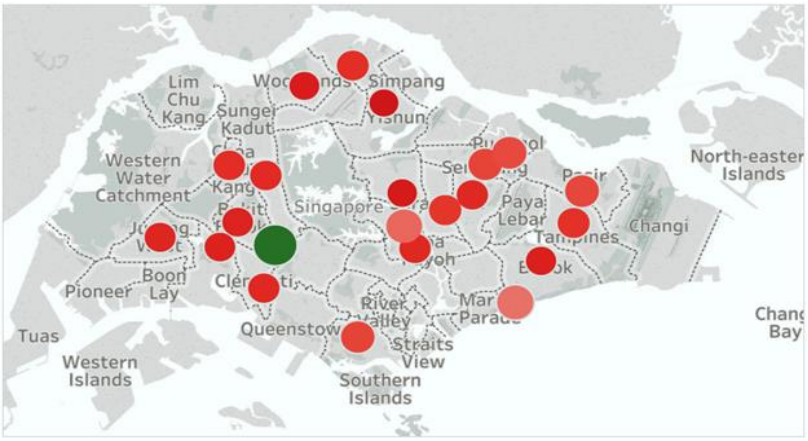
4.2 Exploratory Data Analysis

Prior to running any model on our data, it is imperative to gain a fundamental understanding of the resale price mechanics within the Singapore market. This is even more critical when our objective is not simply to generate accurate predictions, but also to offer intuitive interpretations to the model. We understood, through online articles and common knowledge, that resale price is related to factors such as number of rooms, proximity to amenities and transportation, housing condition, as well as location. However, we were neither able to quantify the effect of each factor, especially when considered among a variety of other housing factors.



The **target variable** resale_price is positively correlated with floor_area_sqm, lease_commence_date, and storey_normalized and negatively correlated with mortgage_rate, lease_age and pmi in descending order. This confirms our hypothesis—that larger-sized homes, with a later lease date, and at a higher level—have a higher price and that the mortgage rate and pmi are higher for cheaper flat types, but it is interesting to see their respective importance for a prospective buyer.

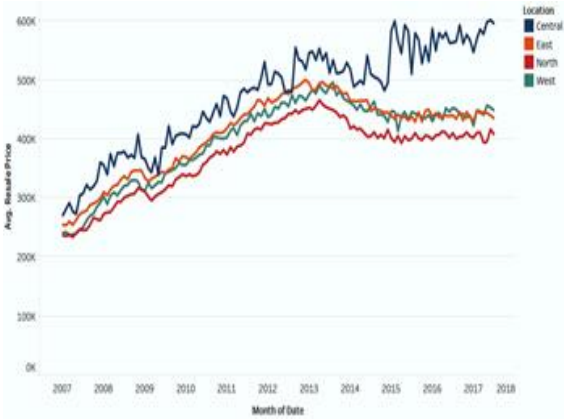
In terms of **explanatory variables**, we noticed that lease_age and lease_commence_date have a very high negative correlation and based on our additional research, we found out that the lease age is actually the period between the start date of the lease of the property and the resale date. Since the high correlation between the explanatory variables can affect the accuracy of our prediction model, we have decided to use only the lease_age for our model among the two variables.
This correlation chart provides us a better understanding on how resale price is affected by different internal and external factors and the insights from the chart helps us in developing a prediction model using those variables.

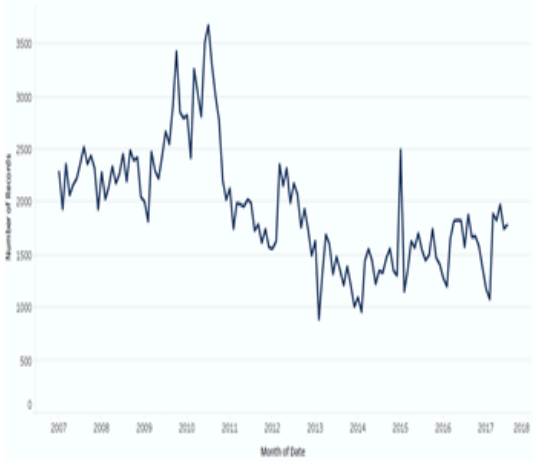


The geospatial plot above shows the distribution of resale prices across Singapore in July 2017. We deduce that the average resale price increases as they near the Central Business District of Singapore—but this is, however, not the only determining factor. For example, Bedok (avg: $373,956) is closer to the CBD, but is less expensive than Pasir Ris (avg: $467,445)—which lies at the most eastern of Singapore.

Polarity
North: Sembawang $414,490
South: Bukit Merah $459,289
West: Jurong West $394,392
East: Pasir Ris $467,445

Most expensive	Least expensive
1. Bishan $721,390	1. Yishun $340,099
2. Bukit Timah $613,520	2. Woodlands $384,608
3. Central Area $593,182	3. Bukit Batok $386,222
4. Queenstown $573,834	4. Choa Chu Kang $388,811

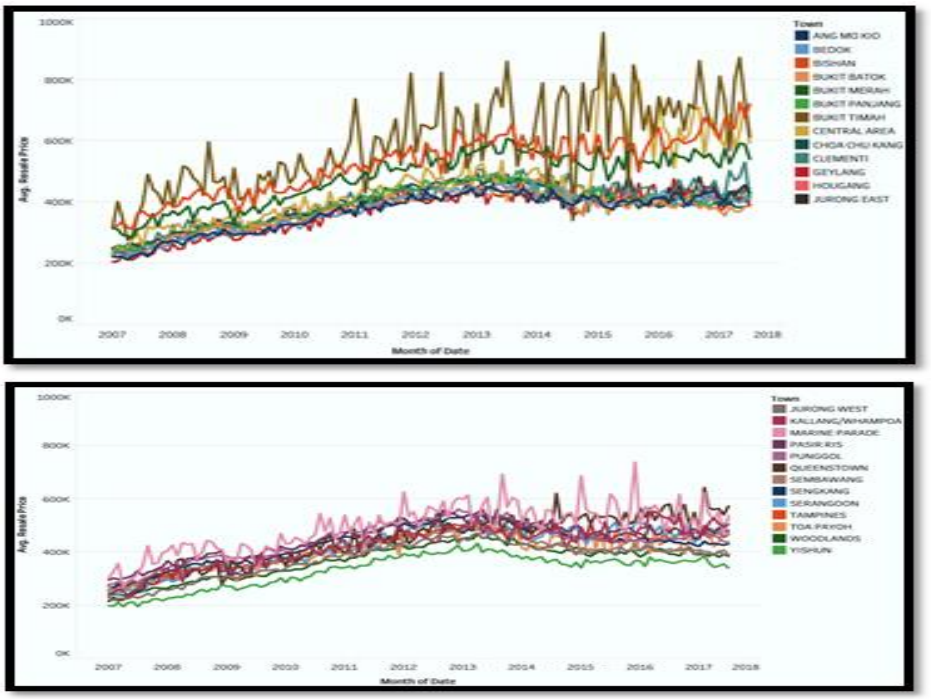




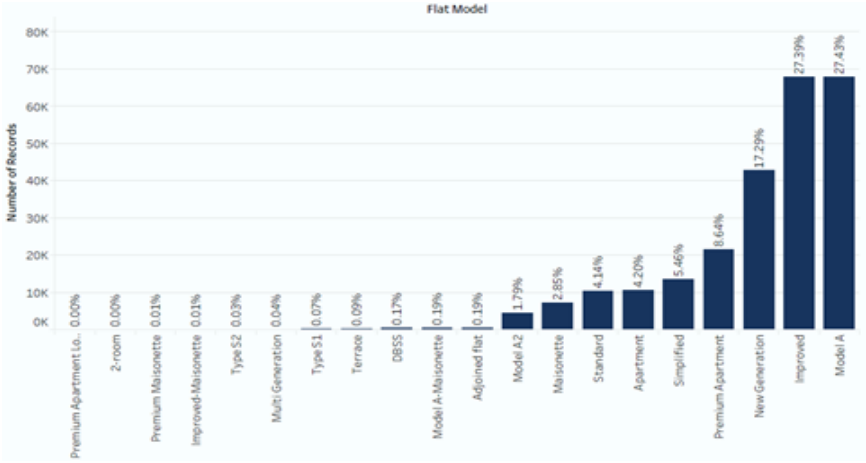
The above graph shows how resale prices have evolved over the years. Price changes have generally been smooth, but Central-located flats are most volatile.

From the above line graph, it can be seen that 2010 saw the highest number of HDB resale flats sold, while the lowest number of HDB resale flats were transacted in 2013.

We also use line graphs to understand the fluctuations of average resale prices in different areas over the years. From the below graphs, we can see that Bukit Timah has the largest fluctuations of average resale prices. It is surprising to see that the areas with large fluctuations of average resale prices also hold higher average resale prices than those areas with small fluctuations of average resale prices.

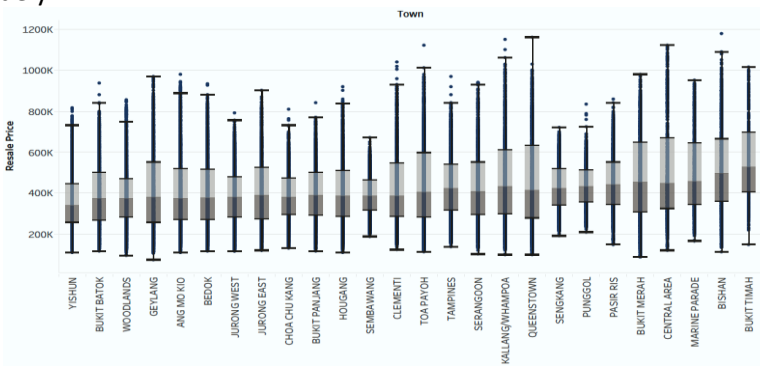


As flat_model is likely to be used as a categorical variable, it is critical that the number of records are sufficient for each category; else this would increase variance and instability within the model. From the chart below, we observe that 11 flat models contain less than 1% of the total number of records—which poses a critical data sparsity issue for us. How we treat and model the data will be explained later under 4.3 Model Building.



Box plots help us understand the relationships between categorical variables and resale prices, and the relationship among categorical variables. The box plot on the following page (sorted by average price, ascending) shows us that Yishun is the cheapest on average; and Bukit Timah is the most expensive.

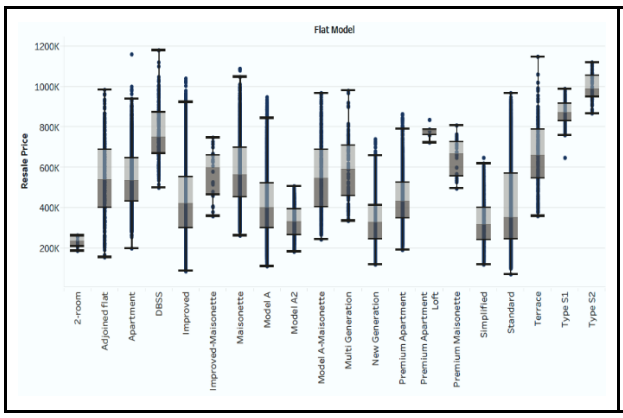
It also shows that towns share approximately the same price for cheaper flats, but otherwise the upper tails vary widely.

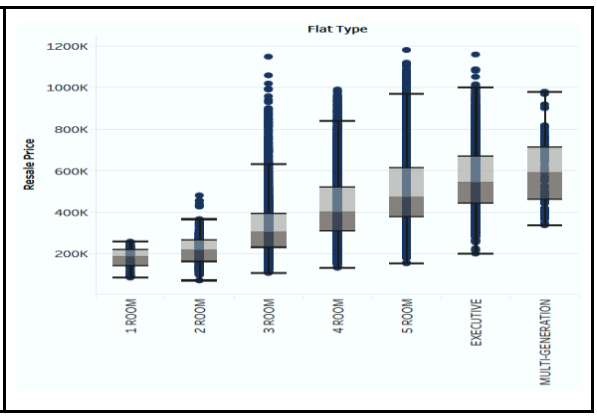


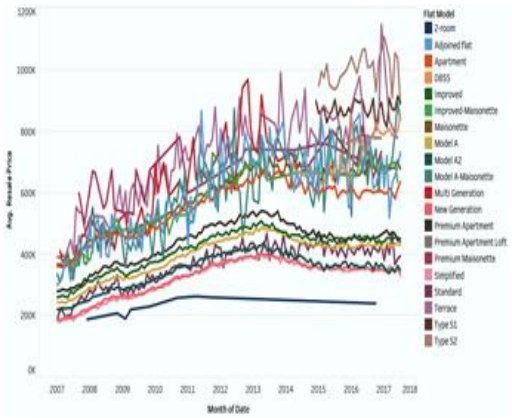
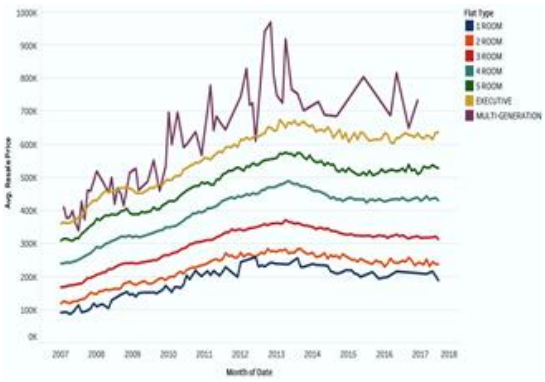
As for flat model and flat type, their distributions are not as uniform as those in towns. Some interesting things to note: the variance of Premium Apartment Loft and 2-room prices are significantly low. A possible explanation would be:

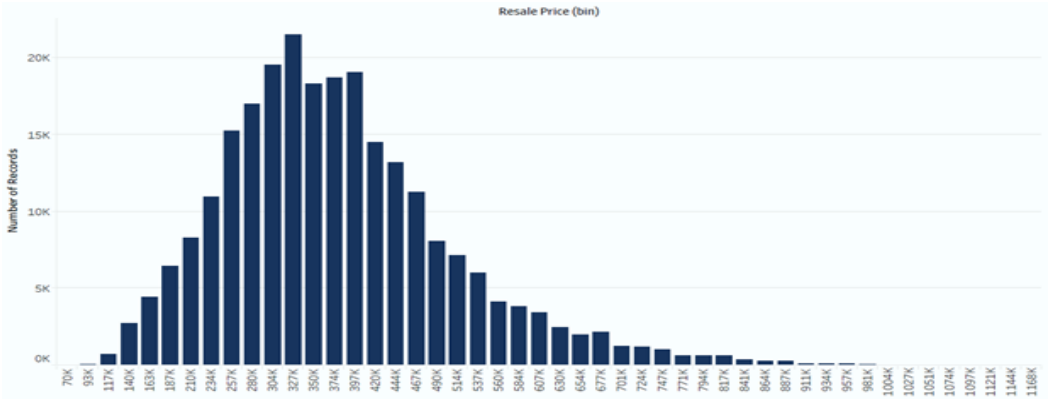
- Government keeps 2-room flat prices low to provide affordable housing for low-income households
- Premium apartment loft has only 4 records; and there is not much variance among those 4

For flat type, it is interesting to note that the lower tail does not increase much with every better flat type (except for the Multi Generation type). However, the upper tails do vary a lot—there are even many outlier for 3, 4, and 5 Room flat types. One can assume that such price variability is accountable by the other variables.





	
The above graph shows how resale prices have progressed over the years. Price changes have generally been smooth for most of the models, but 'Multi Generation', 'Terrace' and 'Adjoined Flat' have unstable average resale prices over the years.	The above graph shows that across flat types, the average prices are generally stable. Only Multi-generational flats see higher fluctuation among all flat types.



From the histogram above, it can be seen that the resale prices are positively skewed. Therefore, we performed log transformation on the skewed data to approximately conform to normality.

4.3. Model Building

We considered 3 models primarily: namely **linear regression**, **random forest**, and **XGBoost**.

As mentioned under 4.2 Exploratory Data Analysis, we faced a data sparsity issue with our categorical variables—especially flat_model. We considered 2 options:

1. To group data-sparse categories as a single category, or
2. To leave them as-is

We attempted both methods for our models, but ultimately decided to leave majority of the categories alone. While they performed less accurately for linear regression, they performed more optimally for ensemble models like random forest and XGBoost—the ensemble models are better able to reduce the variance of these sparse categories through aggregation. The flat models removed were:

- Type S1, Type S2: perfectly correlated with other variables

- Multi Generation: perfectly correlated with the “MULTI GENERATION” flat type

Also, all Premium Apartment Loft flat models were renamed to be Premium Apartment flat models, as they were significantly sparse (only 4 records). lease_commence_date was near perfectly correlated with date_year; we resolved the issue by computing a lease_age variable. We also initially explored “dist_to_cbd” as one of the attributes, but found it of little predictive power or interpretability, especially when the “town” variable is present.

Train-test split

As our data is time series data, we took a sequential split instead of randomly sampling. We aimed for an 80-20 train-test split by recursively going back month by month; subsetting and evaluating if the data proportion is above 20%. We ended up with the following split:

Role	Proportion	No. of records	Timeframe
Train	79.81%	197.212	Before 2015

Linear regression

Linear regression is typically used in a hedonic housing price model, and also provides intuitive interpretations—which would be useful in assessing the relationships of the variables. It also doubles as a baseline model for the random forest and XGBoost.

The linear function for expressing linear regression is as follows, with *f(x)* being the prediction given the explanatory variables *x(i)* and weight coefficients w_i.

$$\hat{f}(\mathbf{x}) = w_1 x(1) + w_w x(2) + \ldots + w_n(n) = \mathbf{w} \cdot \mathbf{x}$$

The objective function in linear regression is to minimize the sum of squared errors by choosing a series of w_i for each explanatory variable, thus projecting a linear relationship between the target variable and the explanatory variables. The formula for the sum of squared errors is shown below:

$$\sum_{i=1}^{m} \left(\hat{f}(x_i) - y_i \right)^2 = \sum_{i=1}^{m} (\mathbf{w} \cdot \mathbf{x_i} - y_i)^2$$

From the linear regression results, we observe the p-value for the F-statistic is less than 0.1%, indicating the model is statistically significant at α = 99.9%. Furthermore, p-values for values other than flat_model variables are less than 0.1%, indicating individual variable statistical significance at α = 99.9%.

It should be acknowledged that flat_model variables suffer from a high p-value, and this is presumably because the models are correlated with other variables such as flat_type and floor_area_sqm. To list some examples:

- 2-room flat model is always 2 room flat type
- Improved or Improved-Maisonette flat model is always 5 room flat type

While this should not affect the predictive power of the model, it should be acknowledged that interpretability of the regression coefficients is compromised, as they affect each other in one way or another. Therefore, one should take caution when interpreting the variables. Based on the summary alone, we concluded the following:

- Town contributes to an approximate 0.1 - 0.3% change in resale price

- Transportation accessibility, in fact, plays a key role in determining resale price, contributing to an approximate 0.1% change in resale price
- A higher mortgage rate leads to lower resale prices, presumably because demand falls

We can also make the following interpretations, albeit less reliably (due to their underlying correlation):

- Each storey contributes to an approximate 0.01% change in housing price
- Flat type is key in determining resale price; but increments fall with every better flat type
- There are certain flat models with significantly higher coefficients than the other models (e.g. 2-room, Premium Maisonette, Improved Maisonette). This indicates considerable variability in resale prices with a different flat model.
    - For a 2-room flat model, the relatively high positive coefficient (0.2751) is likely compensating for the lack of other factors within the apartment

A summary of the regression results is shown below:

OLS Regression Results

Dep. Variable:	resale_price	R-squared:	0.815
Model:	OLS	Adj. R-squared:	0.815
Method:	Least Squares	F-statistic:	1.612e+04
Date:	Sun, 12 Nov 2017	Prob (F-statistic):	0.00
Time:	02:38:35	Log-Likelihood:	1.0017e+05
No. Observations:	197252	AIC:	-2.002e+05
Df Residuals:	197197	BIC:	-1.997e+05
Df Model:	54		
Covariance Type:	nonrobust		

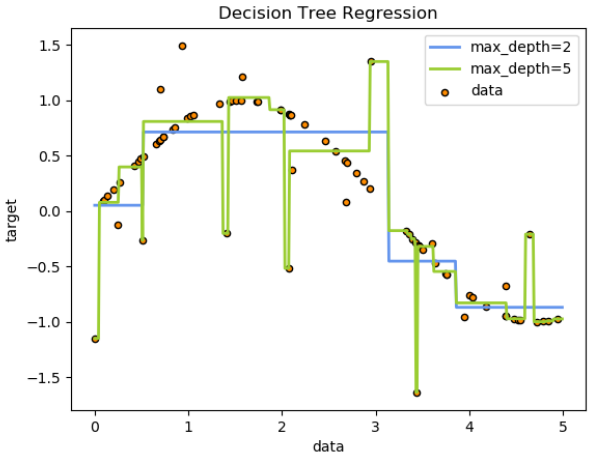
	coef	std err	t	P>|t|	[0.025	0.975]
const	13.1267	0.147	89.558	0.000	12.839	13.414
floor_area_sqm	0.0041	5.93e-05	69.089	0.000	0.004	0.004
storey_normalized	0.0085	7.13e-05	119.558	0.000	0.008	0.009
Stations within 500m	0.0891	0.001	111.652	0.000	0.088	0.091
date_month	0.0028	9.8e-05	28.115	0.000	0.003	0.003
mortgage_rate	-0.1670	0.000	-501.179	0.000	-0.168	-0.166
pmi	-0.0100	0.000	-57.048	0.000	-0.010	-0.010
town_BEDOK	-0.0497	0.002	-24.533	0.000	-0.054	-0.046
town_BISHAN	0.0919	0.003	31.961	0.000	0.086	0.098
town_BUKIT BATOK	-0.1288	0.002	-57.350	0.000	-0.133	-0.124
town_BUKIT MERAH	0.1573	0.002	65.803	0.000	0.153	0.162
town_BUKIT PANJANG	-0.2202	0.002	-88.089	0.000	-0.225	-0.215
town_BUKIT TIMAH	0.2324	0.007	33.249	0.000	0.219	0.246
town_CENTRAL AREA	0.1362	0.004	30.784	0.000	0.128	0.145
town_CHOA CHU KANG	-0.2438	0.002	-103.691	0.000	-0.248	-0.239
town_CLEMENTI	0.0458	0.003	17.990	0.000	0.041	0.051
town_GEYLANG	-0.0148	0.003	-5.656	0.000	-0.020	-0.010
town_HOUGANG	-0.1280	0.002	-58.510	0.000	-0.132	-0.124
town_JURONG EAST	-0.1179	0.003	-45.133	0.000	-0.123	-0.113
town_JURONG WEST	-0.2122	0.002	-102.667	0.000	-0.216	-0.208
town_KALLANG/WHAMPOA	0.0386	0.003	14.910	0.000	0.034	0.044
town_MARINE PARADE	0.2866	0.004	69.293	0.000	0.279	0.295
town_PASIR RIS	-0.1468	0.003	-57.823	0.000	-0.152	-0.142
town_PUNGGOL	-0.1383	0.003	-48.912	0.000	-0.144	-0.133
town_QUEENSTOWN	0.1305	0.003	48.316	0.000	0.125	0.136
town_SEMBAWANG	-0.2587	0.003	-92.624	0.000	-0.264	-0.253
town_SENGKANG	-0.1663	0.002	-70.308	0.000	-0.171	-0.162
town_SERANGOON	-0.0570	0.003	-20.659	0.000	-0.062	-0.052
town_TAMPINES	-0.0784	0.002	-37.197	0.000	-0.082	-0.074
town_TOA PAYOH	0.0311	0.003	12.265	0.000	0.026	0.036
town_WOODLANDS	-0.2700	0.002	-130.980	0.000	-0.274	-0.266
town_YISHUN	-0.2054	0.002	-97.803	0.000	-0.210	-0.201
flat_type_2 ROOM	0.2400	0.013	17.931	0.000	0.214	0.266
flat_type_3 ROOM	0.4615	0.013	35.269	0.000	0.436	0.487
flat_type_4 ROOM	0.6333	0.013	47.067	0.000	0.607	0.660
flat_type_5 ROOM	0.7652	0.014	54.495	0.000	0.738	0.793
flat_type_EXECUTIVE	0.8090	0.015	54.595	0.000	0.780	0.838
flat_type_MULTI-GENERATION	0.9071	0.147	6.161	0.000	0.619	1.196
flat_model_2-room	0.2751	0.154	1.790	0.073	-0.026	0.576
flat_model_Adjoined flat	0.0157	0.146	0.107	0.915	-0.270	0.302
flat_model_Apartment	0.0320	0.146	0.219	0.826	-0.254	0.318
flat_model_DBSS	0.0714	0.147	0.487	0.626	-0.216	0.359
flat_model_Improved	-0.0226	0.146	-0.155	0.877	-0.308	0.263
flat_model_Improved-Maisonette	0.2116	0.149	1.422	0.155	-0.080	0.503
flat_model_Maisonette	0.0264	0.146	0.181	0.856	-0.259	0.312
flat_model_Model A	0.0330	0.146	0.226	0.821	-0.253	0.319
flat_model_Model A-Maisonette	0.0430	0.146	0.295	0.768	-0.243	0.329
flat_model_Model A2	0.0168	0.146	0.115	0.908	-0.269	0.302
flat_model_New Generation	0.0054	0.146	0.037	0.970	-0.280	0.291
flat_model_Premium Apartment	0.0462	0.146	0.317	0.751	-0.239	0.332
flat_model_DBSS	0.0714	0.147	0.487	0.626	-0.216	0.359
flat_model_Improved	-0.0226	0.146	-0.155	0.877	-0.308	0.263
flat_model_Improved-Maisonette	0.2116	0.149	1.422	0.155	-0.080	0.503
flat_model_Maisonette	0.0264	0.146	0.181	0.856	-0.259	0.312
flat_model_Model A	0.0330	0.146	0.226	0.821	-0.253	0.319
flat_model_Model A-Maisonette	0.0430	0.146	0.295	0.768	-0.243	0.329
flat_model_Model A2	0.0168	0.146	0.115	0.908	-0.269	0.302
flat_model_New Generation	0.0054	0.146	0.037	0.970	-0.280	0.291
flat_model_Premium Apartment	0.0462	0.146	0.317	0.751	-0.239	0.332
flat_model_Premium Maisonette	0.1383	0.149	0.929	0.353	-0.154	0.430
flat_model_Simplified	-0.0380	0.146	-0.261	0.794	-0.324	0.248
flat_model_Standard	-0.0687	0.146	-0.471	0.637	-0.354	0.217
flat_model_Terrace	0.6289	0.146	4.303	0.000	0.342	0.915
lease_age	-0.0035	6.55e-05	-54.042	0.000	-0.004	-0.003

Omnibus:	5239.443	Durbin-Watson:	0.468
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5668.435
Skew:	-0.408	Prob(JB):	0.00
Kurtosis:	3.153	Cond. No.	2.15e+05

Random Forest Regression

Having determined that all variables are useful in predicting resale prices, we experimented with random forest, an ensemble learning method bagging multiple decision tree regressors. Our rationale for selecting random forest regression as a model is due to our data's inherent price variability across different combinations of explanatory variables—random forest regression is able to identify and model such relationships.

While decision trees are typically used in the context of classification, it can segment the data and local regressions are performed on these data segments. An illustration by *sklearn* is shown below (albeit overfit):



Random forest regression iteratively does the following for *n_estimators*:

1. Take a randomized subsample (with replacement; i.e. a bootstrap sample)
2. Fit a decision tree regression

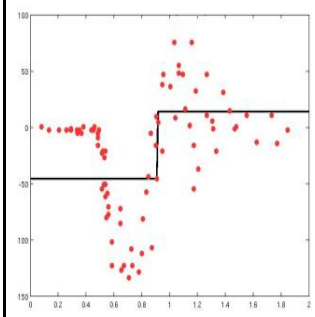
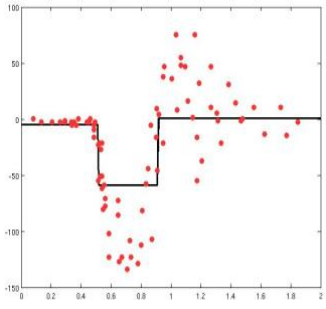
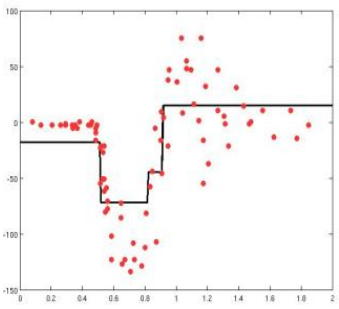
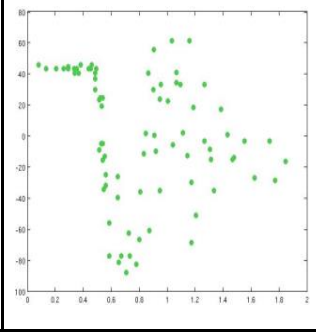
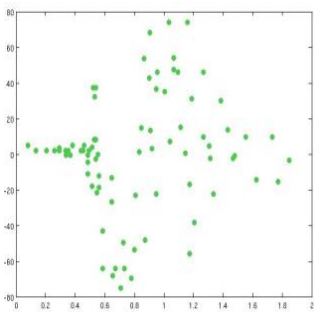
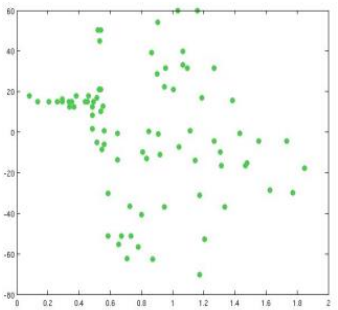
The regressors are then aggregated and averaged to provide the final model, and prediction. The objective function is to minimize the mean squared error, which is simply the average of the sum of squared errors. The following parameters have been tuned to achieve better predictive performance:

Parameter	**Adjusted value**	**Comments**
n_estimators	200	No. of underlying decision trees (default=10) More trees will lead to decreased variance, but this improvement slows with larger numbers of trees
max_features	0.8	No. of features considered when splitting (default="auto") Not many features, hence taking the *log* or *sqrt* of number of variables will heavily limit tree depth
max_depth	50	Maximum tree depth (default=None) A max depth of 50 allows the trees to grow sufficiently, but not overfit
min_samples_split	0.0001	Minimum number of samples required to split (default=2)

		With 0.01% chosen, then each node will require 25 or more samples before it can be further split—an ideal balance between avoiding overfitting and underfitting.
min_samples_leaf	3	Minimum number of samples in a node (default=1) While we did not want a leaf to only contain 1 value, we understood there are certain variables with sparse data. Hence, we accommodated and balanced that with the value of 3.

XGBoost

XGBoost is an extremely popular boosted tree algorithm within the machine learning community, known for both its speed and predictive power. It is an optimized implementation of stochastic gradient boosting (also known as multiple additive regression trees).

Iteration	**1**	**2**	**3**
Fit the data			
Fit the residuals			

Stochastic gradient boosting works by iteratively fitting the data and the residuals for *max_depth*:

1. Fit the data using a decision tree regressor
2. Fit the residuals using a decision tree regressor
3. Aggregate the data fit and residual fit, and use this aggregated regressor in step 1

The following parameters have been tuned to achieve higher performance:

Parameter	Adjusted value	Comments
max_depth	50	No. of underlying decision trees (default=6) Maximum depth of each tree; increasing it introduces more information, but has a higher tendency to overfit.
learning_rate	0.05	Step size shrinkage used in update to prevent overfitting (default=0.3)

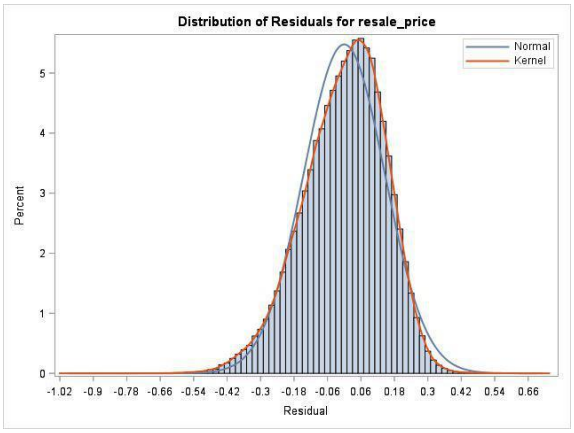
4.4. Model Validation:

For our prediction model, we will not use confusion matrix to validate the model accuracy because it is used for classification model while ours is a regression model for a continuous target variable. Therefore, our approach is utilizing the graphical residual analysis and the predicted vs actual regression plot.

4.5.1 Residual Analysis

Interpreting the distribution plot of residuals

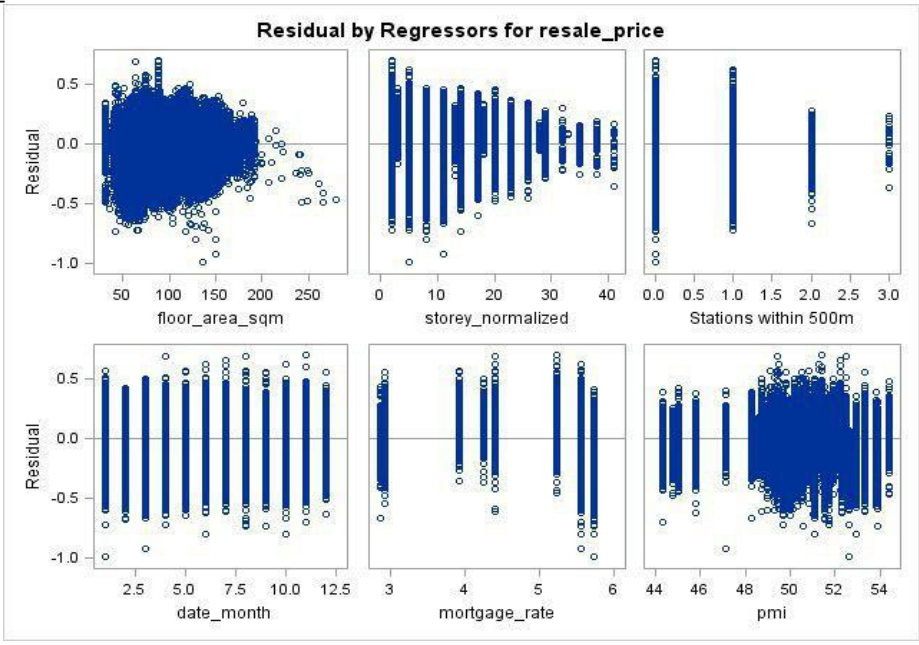
The residual (or error) is the difference between the actual and predicted values at each combination values of the explanatory variables. If the data fits in the model, the residuals would show random errors that suggest the relationship between explanatory variables and response variable a statistical relationship.



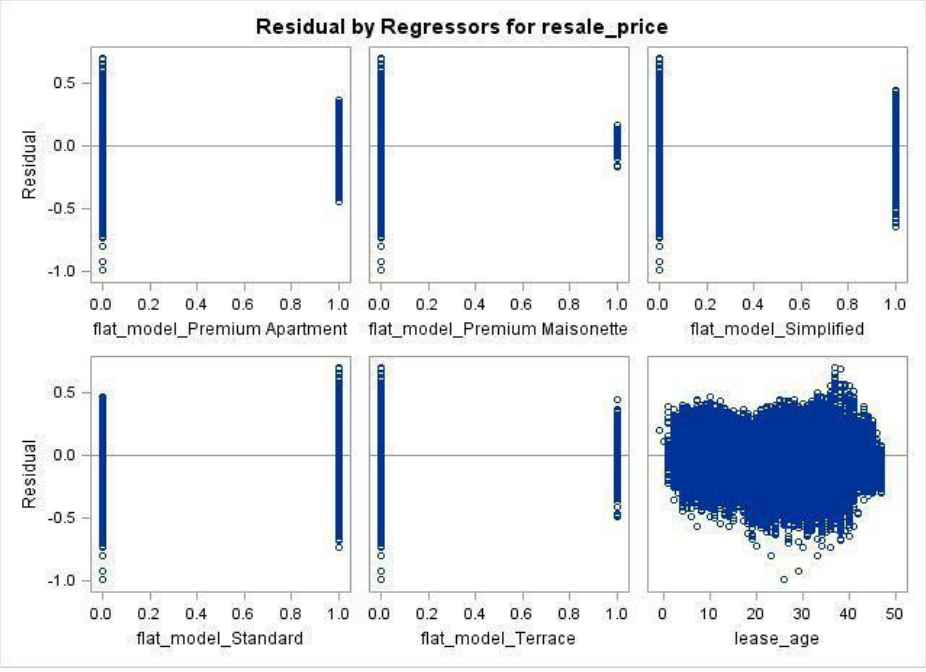
In our situation, according to the distribution plot of the residuals above, we can see that the residuals are normally distributed, which means the residuals behave randomly (random residuals do not follow some pattern induced by bias within the model). This finding suggests the model fits the data well.

Graphical residual analysis

We also use graphical residual analysis to evaluate the relationships between the residuals and each explanatory variable. However, this method is difficult to apply to binary variables. Therefore, we will not assess the residual plots for binary explanatory variables such as town_YISHUN or flat_model_Standard.

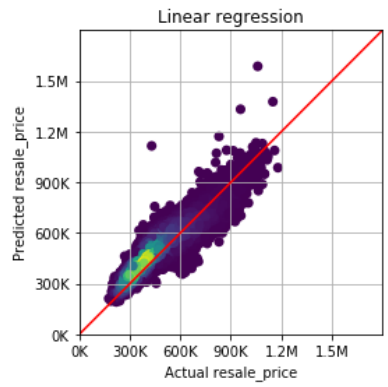
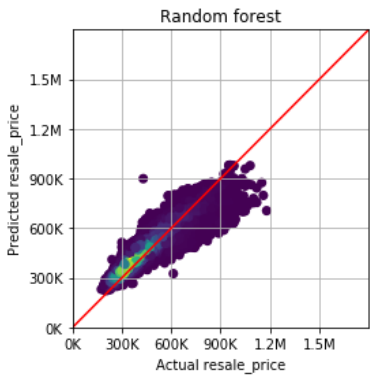
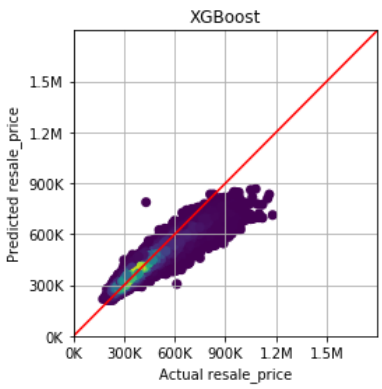


floor_area_sqm suffers from minor heteroskedasticity, as the variance of its residuals decrease with increasing *floor_area_sqm*. There are also several negative residual outliers within the 125 - 175 range—these values are more challenging to predict, but do not pose a significant problem because they are not too far off.



The residual plot for *lease_age* looks healthy, with residuals randomly distributed across the *lease_age* variable.

4.5.2 The Relationship Between Actual vs Predicted Resale Prices

		
R^2: 0.75012 RMSE: 0.14529	R^2: 0.84026 RMSE: 0.11617	R^2: 0.88875 RMSE: 0.09695

We applied 2 methods of evaluating the best model:

1. By inspecting the actuals vs. predicted graph
2. By evaluating the actual vs. predicted R^2 and root mean square error (RMSE) values
    - R^2 calculated with the population average as a baseline model; hence it can be below zero (performs worse than a constant)

Admittedly, random forest regression and XGBoost performed better than multivariate linear regression, in terms of predictive power on average. This can be seen from the higher R^2 and lower RMSE, indicating that the predicted values on average fit the actuals better. While this is usually sufficient to substantiate a model selection, we decided to dig deeper and analyse the actuals vs. predicted plots.

From the actuals vs. predicted plots, we observed that linear regression underpredicted for a segment between 600 - 700K, as seen from our actual vs. predicted plots. For higher prices, linear regression failed to predict a number of the resale prices (presumably due to higher price variability at those prices). However, random forest and XGBoost were able to capture these differentiating details, and were better able to predict at higher prices, albeit underfit a significantly higher amount on average.

Ultimately, we selected the linear regression model because:

1. Resale prices follow a largely linear relationship with the explanatory variables, and linear regression was able to capture succinctly
2. Difference in RMSE is only 0.04834, which means the linear regression output will on average output resale prices that are 0.04834% more off than XGBoost
3. For the purposes of our target audience (government, investors, developers), linear regression is able to provide higher levels of interpretability and a model that represents the fair value of any given apartment with underlying reasons—which would be useful for future developments

5. Experiments

Many components of our planned prediction model require experimentation to decide on what methods work best what do not. Experimentation areas:

- **Data aggregation method:**

  The original HDB Resale Price Dataset contains many variables which, if used without merging and transformation, could potentially lead to a dimensionality issue. In order to reduce the number of dimensions in the model, we experimented with data aggregation methods ranging from weighted sum and unweighted sum of variable values to arrive at a composite score for internal HDB factors.
  Data aggregation using weighted sum and unweighted sum both yielded results that proved less accurate than the results yielded by inputs generated by variables transformed in the fashion described above in Section 4.1.1. Dimensions were reduced by dropping variables that posed multicollinearity issues as well as data merging.

- **Model Accuracy:**

  Multiple models were used in order to leverage on the most accurate method. We ran three prediction algorithms - Multiple Linear Regression, Random Forests and XGBoost. The accuracy of the results was measured using the model validation method as described in Section 4.5. Using this the best method was chosen.

- **Model Validation Method:**

  To assess the accuracy of our model, we performed model validation by building a line of best fit using the method described in Section 4.5. Initially, we hoped to use a confusion matrix but due to the constraints of our prediction model, we experimented with residual analysis and correlation analysis. Since both methods have their merits, we decided to use both validation methods.

- **Incorporating Future Developments:**

  To incorporate future MRT projects, we explored two main approaches - adjusting for the projected increase in prices after the prediction model was run and adjusting prior to model output. The first approach was to adjust for the average change in prices based on training data after an MRT Station was opened in an area. This adjustment would be made after the model outputs a predicted price. The other approach is to append the MRT Station information to the MRT Station Dataset which allows the prediction algorithm to make predictions based on future developments as well as existing MRT Stations. The computational and time efficiency of the second method paved way for it to be chosen as the approach used in the final prediction model.

6. Conclusion

6.1. Summary

6.1.1. Project Findings

Using information from multiple, diverse sources allowed for some interesting inferences to be drawn. We observed that while the flat's location (town) is significant in determining price, its closeness to MRT stations serves a similarly significant importance. Also, we found that mortgage rate affects resale prices inversely.

Flat type and models, on the other hand, contribute to a significantly larger proportion of the price when compared to the above-mentioned factors. That is indeed intuitive: a 3-room flat is pricier than a 2-room flat, a 4-room flat is pricier than a 3-room flat. However, it is interesting to note that there are diminishing returns to price increments with each better flat type. The storey at which the apartment is can change the housing price to a small extent.

We also uncovered a largely linear relationship between our explanatory variables and the resale price: they jointly affect the price in a linear, predictable fashion. However, this relationship does not hold as strongly with more expensive apartments, suggesting that there might be customers buying apartments at a discount—at those prices.

6.1.2. Achievements

- Assessing the impact of selected macroeconomic conditions on the resale prices of HDB Flats
- Calculating and understanding the contribution of MRT connectivity on resale prices
- Incorporating future developments into the prediction model to make better predictions
- Building a holistic prediction model using information from multiple sources

6.2. Recommendations and Future Directions

While our model considers a breadth of factors, it does not consider a large number of detailed factors. Some factors that we considered incorporating but eventually decided against due to time constraints are nearby hospitals, schools and shopping malls. More locational factors such as the ones mentioned, macroeconomic factors as well as internal attributes of HDB apartments will allow models to make more informed predictions.
Future models could potentially deliver far better results by leveraging on newer, more accurate prediction algorithms.

7. References

1. Brownlee, J. (2015, October 9). Regression Tutorial with the Keras Deep Learning Library in Python. Retrieved November 12, 2017, from https://machinelearningmastery.com/regression-tutorial-keras-deep-learning-library-python/
2. Brownlee, J. (2016, March 02). How To Backtest Machine Learning Models for Time Series Forecasting. Retrieved November 12, 2017, from https://machinelearningmastery.com/backtest-machine-learning-models-time-series-forecasting/
3. Menon, P. (2017, October 2). Data science simplified: Understanding logistic regression. Retrieved November 12, 2017, from https://www.techinasia.com/talk/data-science-simplified-11-logistic-regression
4. Data.gov.sg. (n.d.). Retrieved November 12, 2017, from https://data.gov.sg/
5. Scikit-learn. (n.d.). Retrieved November 12, 2017, from http://scikit-learn.org/stable/
6. The Python Tutorial. (n.d.). Retrieved November 12, 2017, from https://docs.python.org/3/tutorial/
7. Engineering Statistics HandBook. (n.d.). Retrieved November 12, 2017, from http://www.itl.nist.gov/div898/handbook/pmd/section4/pmd44.htm
8. Room, R. (n.d.). Monetary Authority of Singapore (MAS). Retrieved November 12, 2017, from http://www.mas.gov.sg/
9. Land Transport & Authority. (n.d.). Retrieved November 12, 2017, from https://www.lta.gov.sg/content/ltaweb/en.html