DSE Final Project

HDB Resale Prices

1. Introduction

Housing has always been high on people's list of concerns. In Singapore, 80% of residents live in HDB flats, which are quality and affordable public housing built by the Singapore government since the 1960s to provide homes for Singaporeans. While the government heavily subsidises first-hand purchases, resale prices of HDB flats have varied greatly across the country. This paper will investigate the relationships between the resale price of HDB flats and their structural and environmental characteristics, and attempt to build an adequate model to predict resale prices.

2. Overview of Data

This project would use HDB flat resale transactions available from 2021 taken from data.gov.sg and managed by the Housing and Development Board. The data set contains 6000 observations taken between January and May, each with 230 variables accounting for characteristics of the HDB flat being transacted. The data is split into training and test sets based on a 7/3 ratio using seed 123456. The training set contains 4200 observations while the test set contains 1800.

3. Initial Analysis

The data consists of 52 continuous and 178 dummy variables. As the data used is a subset of the full data set, many dummy variables, such as 'month_Jun' record 0 variance throughout the data due to the small sample of observations.

Examination of the variables presented in the data set reveals numerous sets of correlating variables. Firstly, there is a strong correlation between the floor area of the HDB flat and its flat type, with larger flat types such as Executive having greater floor area on average than smaller flat types such as 3-room (Figure 1). Additionally, a clear positive correlation is also found between the resale price and the storey range of HDB flats. On average, higher-level flats would sell for higher prices (Figure 2).

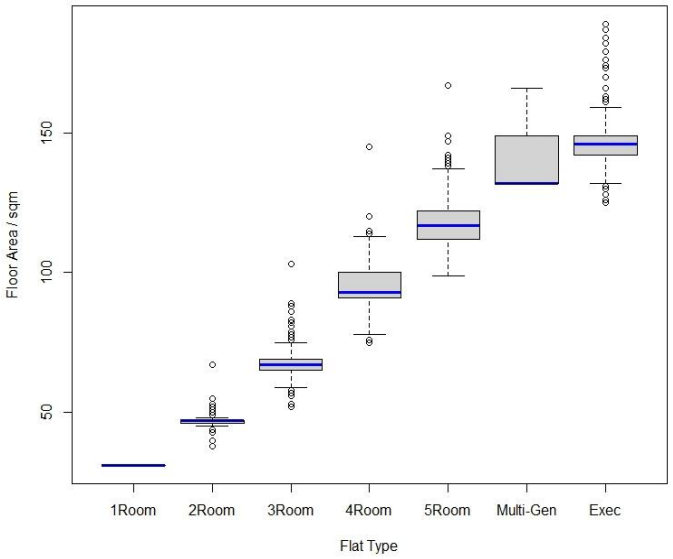

Figure 1: Box plot of floor area of different flat types (Blue line indicates median floor area)

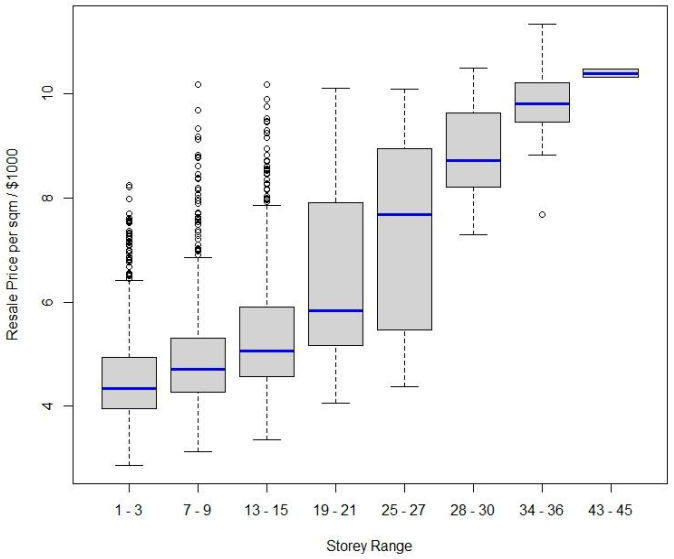

Figure 2: Resale price per sqm of floor area for 8 storey ranges (Blue line indicates median resale price per sqm)

4. Predictive Modelling

Considering the large number of categorical predictors in the data, multiple linear regression (MLR), decision tree and principal component regression (PCR) would be used to model resale price. Other models such as K-nearest neighbours and K-means clustering would not be applicable due to their use of Euclidean distance, which is not meaningful when applied to categorical variables. Resale prices would also be normalized by dividing by 1000 for better interpretability.

To start, an MLR model was fitted on the training set with all variables. The resultant model contained 63 undefined coefficients due to singularities caused by a lack of variation in the variables or collinearity with other variables. Additionally, many variables have p-values > 0.05, which suggests that they are statistically insignificant in influencing resale price at a 5% significance level. As such, variable selection should be done to improve the model.

To do this, variables with a lack of variation and those with p-values < 0.05 were removed. Then, sets of potentially collinear variables were examined and only the most representative variables were selected within each set. For example, based on the previous finding of correlations between flat types and floor area, the dummy variables ‘flat_type_X’ and ‘flat_model_X’ can be removed as the information can be represented through the variable ‘floor_area_sqm’.

However, there are some exceptions to this rule. The variables ‘storey_range_X’ for instance, should be kept despite a large p-value, as people’s preference for better views of

scenery would translate to higher prices for higher storeys. Conversely, variables with a lack of meaningful relationships to resale price, such as ‘dist_nearest_GAI_jc’ would be omitted despite statistical significance as there is no logical link that there is a higher preference for living near government-aided JCs and therefore higher ‘resale_prices’. The higher prices are likely due to the coincidence that such JCs are located in more mature and popular HDB neighbourhoods.

With the above considerations, 58 variables were selected and fitted with MLR. It was also discovered that taking logarithm on the y-variable ‘resale_price’ improved the linearity of the model (Figure 3). The final model produced an adj. R^2 value of 0.91, which is comparable to the benchmark model of resale price against all variables with adj. R^2 of 0.93.

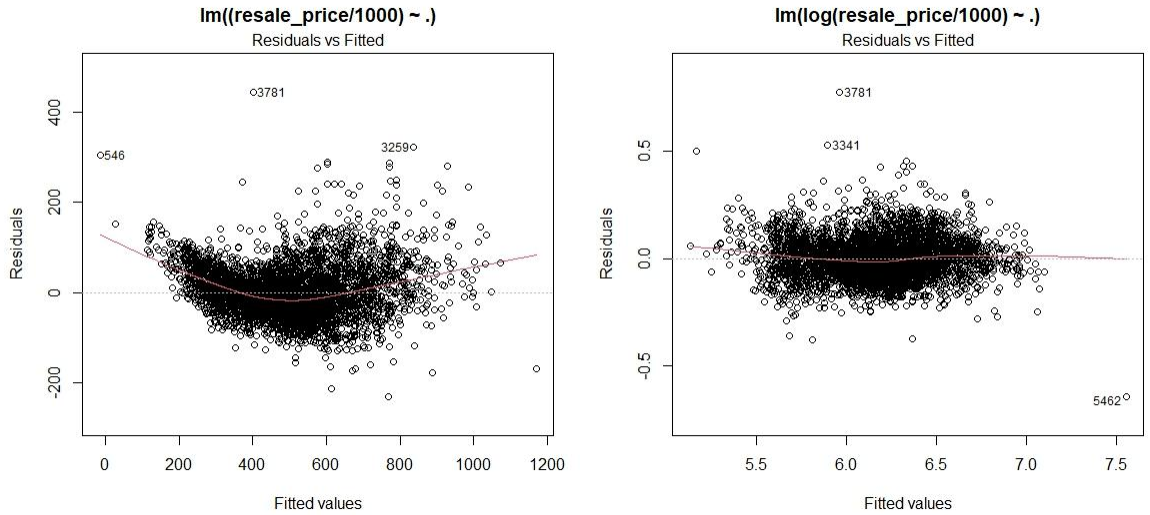

Figure 3: Residuals vs Fitted Plots of models (Red line became more horizontal when log(resale_price) is taken)

Decision trees were then fitted on the raw data and variable selection was done automatically. With MSE as the loss function, cross-validation in rpart package produced a more complex but less interpretable tree with 97 nodes. The tree package produced a smaller tree with 12 nodes after cross-validation, which is a subset of the larger tree. The 2 packages produced contrasting trees due to differing methods and conditions used for pruning in each package.

5. Model Performance

The models produced previously were then used to predict the resale prices on the test set. The performance of each model is evaluated based on the RMSE value of predicted values against the actual value. The tabulated results are shown below.

S/N	Model	RMSE (rounded to 2 d.p.)
1	MLR with all var.	45.51
2	MLR with 58 var.	60.14
3	MLR with 58 var. and log(resale_price)	57.04
4	Decision Tree using rpart() with 97 leaves	57.03
5	Decision Tree using tree() with 12 leaves	79.90
6	PCR with CV-selected no. of components (164)	45.44
7	PCR with 164 components and log(resale_price)	41.56
8	PCR with 25 components and log(resale_price)	65.68
9	PCR with 58 components and log(resale_price)	56.44

The initial MLR models after variable selection performed relatively well, this suggests that most of the critical information was retained in the set of variables selected. Transformation done on 'resale_price' also improved model performance as expected. The performance of decision trees however varied greatly, with the more complex tree produced from rpart producing a smaller error than the simpler tree.

It is also clear that machine learning methods such as decision trees and PCR were able to generate very good predictive models automatically, with little input from the user, especially in doing variable selection. PCR (S/N. 7) in particular was able to effectively reduce the number of dimensions (from 239 to 164) and deliver better performance than MLR models. This reinforced the importance of machine learning in data analysis. Ideally, a combination of unsupervised and supervised learning methods should be employed, coupled with some human input to develop an optimal model.

6. Conclusions

This project allowed for a deeper appreciation of the various modelling methods, as well as how they can be applied to analyse real-world data. With an error of $41,000 for the best-performing model, it could certainly be optimised further using more advanced methods and be applied in a real-world scenario.

The project is however bound by limitations. Firstly, the small sample of data meant that many variables lacked variation throughout the data. The model would thus not be able to account for factors such as the month of transaction when predicting resale prices and a larger sample could be obtained to mitigate this problem. Secondly, variable selection for MLR was mostly done through intuition to determine the relative importance of the variables. A more technical and quantitative method would be more appropriate for this purpose.