**Introduction**

Housing and Development Board (HDB) resale prices has always been a hot conversation topic. The common HDB prices we see are largely anchored on a sales comparison approach, driving the question of whether we can determine the flat’s intrinsic value based on various factors. This report aims to analyse a curated dataset of HDB prices in 2021, a subsample of Nathanael Lam Zhao Dian’s Honours Thesis dataset, to unravel key factors that play a role in determining the resale prices of HDB flats. The performance, advantages and limitations of different models will be discussed and compared to determine the best price predictor model. The data is normalised by dividing the resale prices by 1000, and the dataset will be split 80-20 for the train-test set (4800/1200).

**Key descriptive statistics and visualisations**

To first investigate the importance of different variables in affecting HDB resales prices, we can find the correlation values of each input variable with the output variable (resale prices). The top few variables are summarised below.

|  |  |  |  |
| --- | --- | --- | --- |
| Physical properties | Flat type/model | Amenities/Qualities | Others |
| Floor area (sqm): 0.631  Max floor level: 0.507  Remaining lease: 0.351 | 3 room flat: -0.503  Executive: 0.347  5 room flat: 0.341 | Distance from CBD: -0.269  Maturity: 0.219 | X3 room sold: -0.419  X5 room sold: 0.342 |

From the relatively higher positive correlation between physical properties and resale prices, we see that buyers tend to focus on floor area, maximum floor level and remaining lease . This could be explained by the practical mindset of buyers to preferring to pay more for a larger area and longer holding period, as it possibly translates to larger gains when it is resold. Interestingly, the importance of proximity to amenities like hospitals, schools, stations and CBD seems to have lower impact on resale prices. This can be seen from the much gentler slopes on simple linear regression models (LR) for proximity to station and CBD, with lower adjusted R-squared values of 0.011 and 0.0724 as compared to the model for floor area. However, for LR of floor area, a large distribution of scatter plots above the regression line where floor area is between 75-150 sqm indicates that there are many other factors attributing to higher prices for smaller flats.

|  |
| --- |
|  |

A box plot of different flat types show a positive correlation between better flat types and higher resale prices. However, there are many outliers for flats with 3 to 5 rooms, indicating that under certain conditions, buyers are willing to pay much higher for the same type of flats. A possible condition is the maturity of the estate, shown by the higher resale prices at the 75th percentile for mature estates compared to non-mature estates.

|  |
| --- |
|  |

**Predictive modelling – Multiple Linear Regression**

When all input variables in the dataset was fitted onto a regression model, an adjusted R-squared value of 0.9295 is observed. However, this could be an overfit for the model, which is not desirable.

The benchmark model for predictive modelling of resale prices used is Multiple Linear Regression. The initial tuning parameters selected were floor area, maximum floor levels, remaining lease, maturity of estate, and distance to amenities. The adjusted R2 value was 0.8479. Upon the previous analysis, this value was improved to 0.8501 after I added in two new variables - flat type and MRT lines, and reduced the amenities to only proximity to CBD and stations. Flat type and MRT variables are a collation of discrete variables of different flat types and location according to MRT lines. This suggests a good fit of the model.

**The fitted model:**

Resale Price = – 161.0 + 5.2 (Remaining lease) + 4.2 (Floor area) + 4.1 (Max Floor Level) + 45.6(Mature) – 13.3 (Distance to CBD) – 46.9 (Distance to nearest station) – 130.7 (1 room flat) – 138.2 (2 room flat) – 113.4 (3 room flat) – 113.4 (4 room flat) – 94.4 (5 room flat) – 55.0 (Executive flat) + 78.3 (NSL) + 86.6 (EWL) + 58.6 (NEL) + 82.5 (CCL) + 44.5 (DTL) + 68.8 (TEL)

**Predictive modelling – K-nearest Neighbours Regression (KNN)**

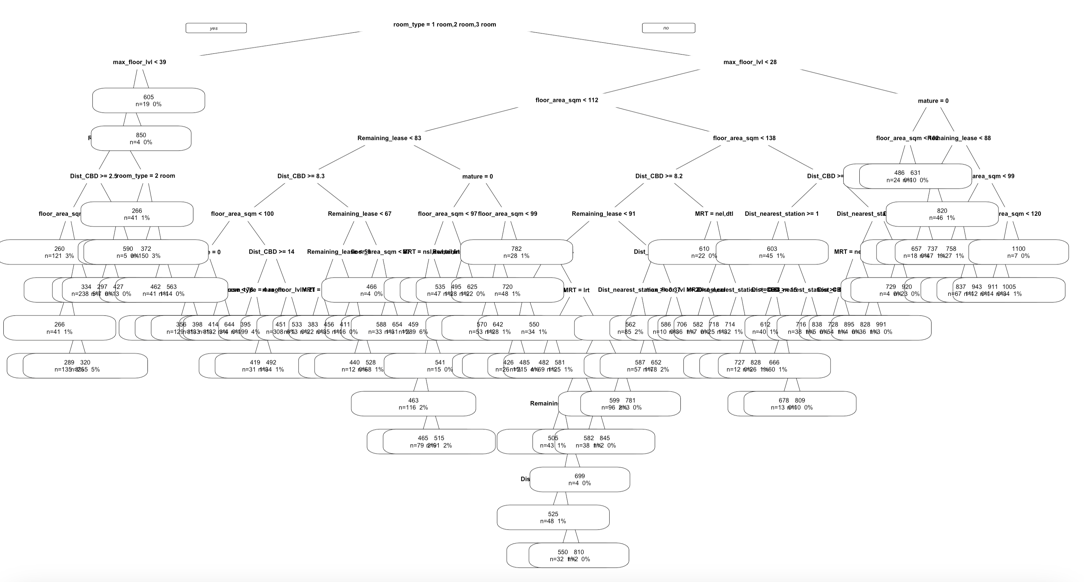
The KNN model is first fitted using all predicators with kmax=100 using the test set. After cross validation (CV), the best K value is found to be 2 for both, and kernel = ‘gaussian’ performed better with a Root Mean Squared Error (RMSE) of 70.41 compared to kernel = ‘rectangular’ with RMSE of 70.69. Gaussian Kernel assigns closer points a higher weight, which allows it to handle spatial dependencies better such as how spatial arrangement of neighbourhoods influence resale prices, with certain locations having a higher influence. Gaussian kernel is also less sensitive to noise, which can be more suitable for this dataset since there may be some low significance and noisy datapoints.

After fitting the KNN model with specified predictors from the previous step, RMSE for ‘gaussian’ is 45.04 and RMSE for ‘rectangular’ is 46.22. The relatively lower RMSE suggests a better fit of the model to the underlying patterns in the data, and a better performance on out of sample data.

**Predictive modelling – Decision Tree Regression**

The decision tree is first fitted using all predictors, with method = ‘anova’ and starting complexity parameter (CP) = 0.0005. The tree is then pruned using the best CP found to be 5e-04 using CV. The RMSE using rpart package is 53.13, significantly lower than using tree package which gives RMSE of 81.12. Upon specifying predictors to those previously chosen, the RMSE value for rpart package is 55.53 and that of tree package is 81.12.

This suggests that Classification and Regression Trees (CART) used in rpart package is more suited to capture relationships in dataset. By creating complex decision boundaries to capture how combination of factors could interact and influence the resale price, it can partition dataset into different segments and better identify unique trends. For instance, different segments of the resale market have different characteristics and economic influences, like how flats near CBD or in mature estates are able to sell at higher prices even if its floor area is smaller. CART could also have identified interactions between predictor variables, such as how the combinations of floor area, location and property type will interact to influence resale price. From the decision tree, the most important variable is type flats. For the subgroup of better flat types (more than 3 rooms or executive), buyers will be more sensitive to floor area and maturity of the estate, as pricing may not be a concern and they are willing to pay much more for better living conditions.



**Predictive modelling – Principal Component Analysis (PCA)**

PCA was attempted with all the predicators to reduce dimensions of the dataset. By plotting a scree plot, it is observed that each Principal Component (PC) accounts for a very small variability. 69PCs are required to explain 80% of the variability, suggesting that PCA is not a great predictive model type as the data has high dimensionality and is quite complex. The model might be capturing noise instead of underlying patterns. This can be justified as there are complex factors like size, amenities, locations in analysing resale prices.

**Performance**

|  |  |  |  |
| --- | --- | --- | --- |
|  | MLR | KNN - gaussian | DT - rpart |
| All predictors | 46.29 | 70.41 | 53.13 |
| Specified predictors | 65.63 | 45.04 | 55.33 |

The respective RMSE values for different predictive models are summarised above. After comparison, the best fit for the model is KNN using gaussian kernel using specified predictors, slightly better than MLR, the basic benchmark model, using all predictors. MLR is not recommended as the model may appear to fit the complex training data well, but not generalize well to new data. The good performance might be due to overfitting, and once it is tuned to specified predictors, there is a large drop in performance. Predictive model KNN is recommended, as it better reflects non-linear relationships between predictors and resale prices, which might be the case in real life as certain predictors like proximity to amenities only impacts the prices if the distance is less than 0.5km, and that improvement of flat types increases resale prices at a non-constant rate.

**Potential limitations**

This dataset is limited to resale prices of HDB flats in the year of 2021, which recorded a record growth since 2010. This was heavily due to renewed market confidence in the low interest rate environment after Covid-19, boosting resale flats demand. Hence, this predictive model may not apply to future samples as economic policies are also important predictors that could be used in the predictive models. For DT, the model is too complex to read and follow through, and the best CP obtained from CV was unable to produce a well fitted model. Alternative feature selection methods could have been attempted to simplify tree produce better results. While KNN could perform well, it is difficult to interpret and might not be as useful for flat buyers. The attempt for PCR failed as there were errors arising, and it might have captured too much noise that reduced its predictive ability. Overall, key parameters identified are still effective in fitting KNN model to identify non-linear trends and predict HDB resale prices, but domain knowledge is also necessary when making a judgement.