Resale hdb prices Report

Justin Cheong Zekang

CONTENT

Intro

The Singapore property market has recently seen a boom amid the Covid-19 pandemic. Analysis on the HDB secondary market would be useful in helping us understand some of the factors that drive the resale flat market. In this report, I will first discuss the transformations that I have made on the dataset then briefly give some summary statistics of the data to help with the us with its understanding. Next, I will introduce some interesting stats that was discovered over the course of my analysis. I will then analyze the dataset using a few methods described below and compare them to my benchmark Linear regression model. My main goal will be to create a model that efficiently predicts the resale housing price given a new set of data.

Dataset Transformations

The dataset is a subsample from Nathanael Lam Zhao Dian's Honours Thesis dataset. It was taken from open sources (such as data.gov.sg), following which, I have imported into Rstudio and cleaned it by getting rid of duplicated values. I then normalized the HDB resale prices by diving it by 1000 to make it more manageable during the analysis. The Boxplot method was used to get rid of outliers and a Random Seed was set to ensure replicability. Of the remaining 5857 observations, approximately half (2928) was used as training dataset and the rest(2929) was used as the testing set.

Summary statistics and interesting findings

Min.	1st Quantile	Median	Mean	3rd Quantile	Max.
180.0	380.0	466.8	488.2	581.9	890.0

From domain knowledge, floor area of a HDB and its remaining lease will be directly proportional to its resale price. Other factors are less clear cut. We will explore the extent of their relationships later.

Although a simple linear model that regresses the HDB resale price using distance to nearest station produces a model that only explain 1.5% of the data (R^2 = 0.01402), figure 2 shows a clear distinction between the resale prices of HDBs near different lines. Notably, the notch of NSL does not overlap with notches of every other line, telling us that its median differs from the rest. This could be due to NSL tend to be far from the CBD. The max value of TEL is also significantly lower than the rest, possibly due to the smaller sample size because TEL is relatively new and less developed.

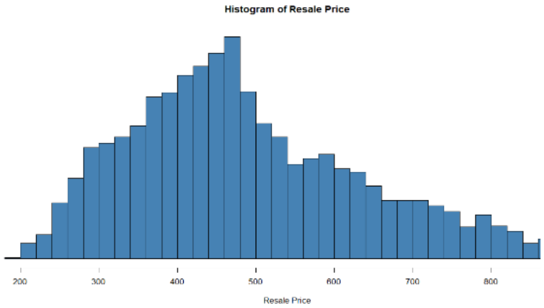


Figure 1: Histogram of Resale Price. Training data.
Single mode, skewed right, not normal(asymmetric)

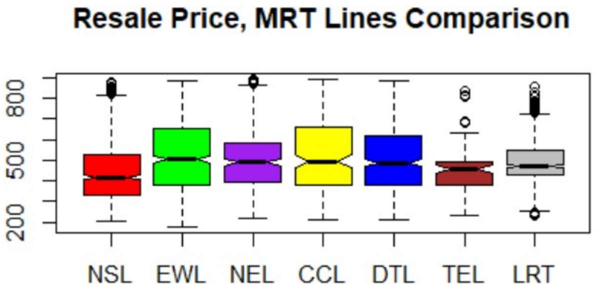
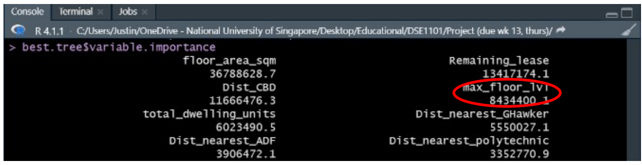


Figure 2: Boxplot comparing Resale Price between houses that are close to different MRT Line. Training data.

Linear Model (Benchmark)

To attain the best linear model, a few simple linear fits based on domain knowledge was explored and floor area(sqm), remaining lease and distance from CBD are important predictors of resale price.

A decision tree was built to check and find other useful predictors. From figure 3, the maximum floor level is also a notable predictor. It was also included in the model.


Console
Terminal
Jobs
R 4.1.1 · C:/Users/Justin/OneDrive - National University of Singapore/Desktop/Educational/DSE1101/Project (due wk 13, thurs)/

```
> best.tree$variable.importance
             floor_area_sqm            Remaining_lease
                 36788628.7                 13417174.1
                   Dist_CBD                max_floor_lvl
                 11666476.3                  8434400.1
        total_dwelling_units      Dist_nearest_GHawker
                  6023490.5                  5550027.1
          Dist_nearest_ADF   Dist_nearest_polytechnic
                  3906472.1                  3352770.9
```

Figure 3: Importance of Variables ranked

Then, by looking at the plots of predictors against resale price, it seems that floor area and remaining lease could potentially be better modeled by polynomials. Thus, I performed a 10-fold cross validation to both. Results show a 0.68% increase in data explained for floor area and a 0.87% increase for remaining lease.

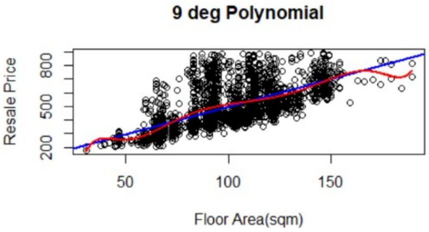
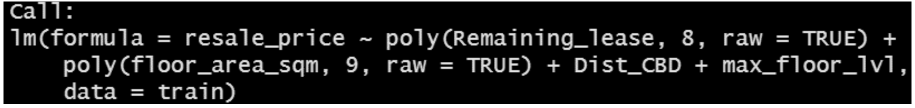


Figure 4: Effects of using polynomial to fit linear model

Blue line: Degree 1 Red line: Degree 9

This is my benchmark linear model:



```
Call:
lm(formula = resale_price ~ poly(Remaining_lease, 8, raw = TRUE) +
    poly(floor_area_sqm, 9, raw = TRUE) + Dist_CBD + max_floor_lvl,
    data = train)
```

It explains roughly 80% of the data given(since $R^2 = 0.7981$) and gives me a benchmark MSE of 4391.854 which is what I will be comparing the rest of my models against. It will be stored in the variable called mse_comparison in my R script.

Surprisingly, Distance from MRT stations is not a very important predictor for resale prices given that prices differ when comparing the which MRT station the HDB is close to.

Hierarchical clustering and K-Means clustering

Firstly, we try to uncover some meaningful relationships within the data using unsupervised learning since we are still unsure what patterns exist within the dataset. We use Hierarchical clustering first to gain some insights about the data before choosing K in K-Means clustering.

To do this, we create a subset of train dataset which contains only the continuous variables, so that dissimilarity measure(Euclidean distance) between each point is meaningful. Only 100 datapoints will be used (greater readability) in the dendrogram since results will be similar. Average Linkage is used over Single Linkage since it is more balanced and over Complete Linkage since it produces a dendrogram that is cleaner to cut. From Figure 5, it seems natural to cut the tree at height = 3, forming 4

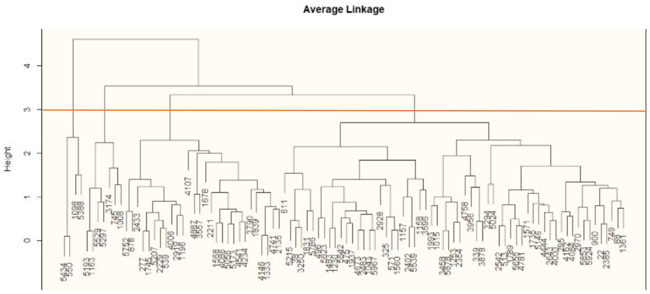


Figure 5: Hierarchical Clustering Dendrogram

Numbers at leaf node are indexes of data, not significant in analysis

Cut at Height = 3

groups. However, this is not very informative by itself since we cannot ascetain what basis we are cutting on.

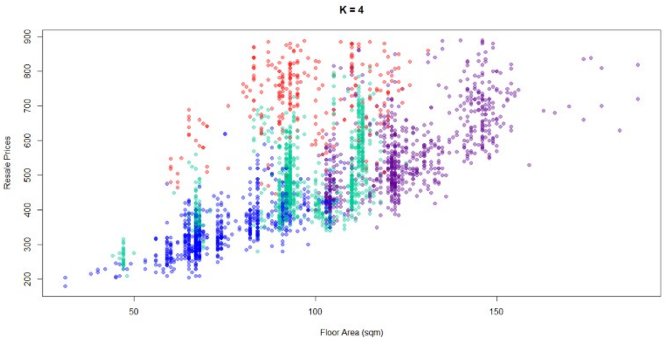


Figure 6: K-means Clustering of 4 clusters
Transparency indicates low density and vice versa

For K-means clustering, we look at both plots with center, k = 19(based on AICC and BIC) and k = 4 (based on Hierarchical Clustering) and found that k = 19 gives too many groups with no clear basis to intepret them. Thus k = 4 is used since it is able to be intepreted.

From figure 4, it seems that HDBs with small floor area tend to have low resale price(blue). There is also a group in the middle of the plot(green). These could be the average HDB. Those with high floor area generally have higher resale prices(purple), but some have lower resale prices that is near the median value. There is also a group with high resale price without as much floor area(red). This could be due to red being in prime locations such as being very near CBD while purple may be in non-mature/less desired locations.

Decision Tree

To find the best model for prediction, we look at the tree that we built earlier. To produce the tree, I create a big tree based on all quantitative data. Then we attain the best complexity parameter by eliminating leaf if it cutting it only returns us a small increase in loss.

This gives us a MSE of 3245.252, outperforms the benchmark model although a tree is easy to build. Tree produced is too complex to be interpretable and is thus not displayed here but it is still very useful in predicting data.

Since the tree outperforms the benchmark model, the data seems to be better modelled by a non-linear model. However, this result is surprising since decision trees does not capture all the information when dealing with continuous variables, which is what we mainly deal with.

K-Nearest Neighbors

To perform KNN regression, a non-parametric method, I first isolate relevant data from my training and testing dataset for easier handing. I then find the best value of K using LOOCV. Best value of K is 2.

Next, KNN regression is performed on the test set to get the out-of-sample MSE to evaluate the model.

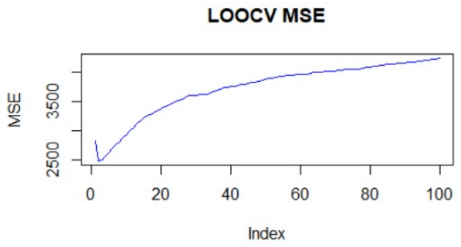


Figure 7: Decision to use K = 2

KNN is much slower as compared to the benchmark model since it keeps track of all existing datapoints and does not summarize them. With a MSE of 41544.604, it is outperformed by both the benchmark model and the tree. Thus, it is rather inefficient model for HDB resale price. This is surprising since the data is low in dimension.

Principal Component Regression

We use 13 components since it is enough to effectively explain 90% of the variance. As seen on figure, each additional PC becomes does not add much value after 13.

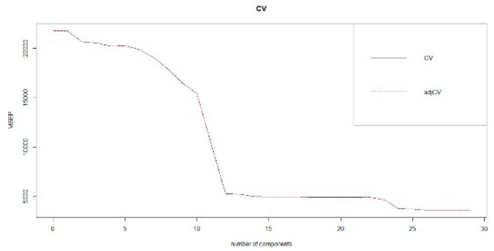


Figure 8: Choosing number of components

 From figure 9, many components see a spike in loading value at certain variables, possibly indicating a greater positive relation of those variables while those with high negative loadings indicate a negative correlation. For example, the black line in figure 9 represents PC 1, which is consists of more negative loadings. This means that there is more predictors that are negatively correlated to resale price. A biplot is not shown due to the large number of variables, making it less readable.

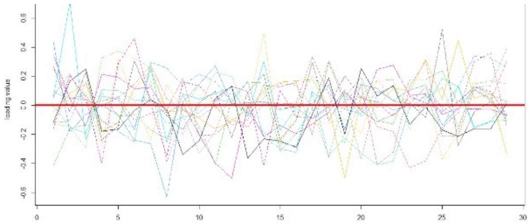


Figure 9: Positive and Negative loadings for components.
Legend found in R script.

The MSE obtained for this method is 5170.118. This is worse than our benchmark and tree model. Possibly due to some variables having low correlation, which makes PCR perform worse. However, PCR allowed us to reduce the dimension by half while still giving us a relatively good prediction.

Naïve Bayes

Aside from the regression problem, to help us to classify which Resale HDBs would be considered more expensive and to supplement our earlier findings, we build a Naïve Bayes Classifier to help us with the classification.

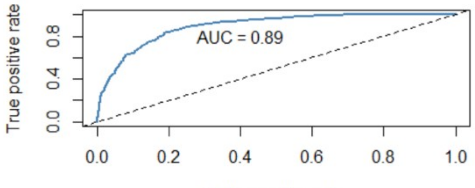


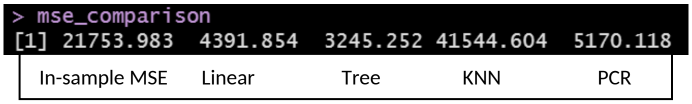
Figure 10: ROC curve

NPred	False	True
False	1251	319
True	249	1110

The model is built to predict if the resale price is higher than the Population median resale price based on the most relevant predictors – floor area sqm, Remaining lease, Distance from CBD and the max floor level. This model is adequate since it classifies more than 80% of the test data correctly. We did not compare this to other classification methods since the main aim is a prediction problem.

Conclusion

To conclude, we have used graphing methods such as the boxplot and clustering to gain additional insights to the data. Then we built some models based on the training set and tested them using the testing set giving us a set of MSE.



```
> mse_comparison
[1] 21753.983  4391.854  3245.252 41544.604  5170.118
```

The best result is the decision tree which is unexpected due to the large proportion of variables being continuous which is unfavorable for the decision tree method.

It is surprising that KNN did not work well since the data used is low in dimension and KNN seem to be a good fit for the problem. However, normality assumption might be a reason for the poor results. Perhaps different weightings of neighbors could be explored and is especially relevant in current market conditions where a record number of million-dollar HDB resale flats were sold.

Another limitation of this study is that categorical variables such as maturity of estate, and geographical variables(except distance from CDB) such as postal codes has been excluded for the most part. A more holistic study would include these since it is known that properties in different locations have varying prices due to the value of the land and the availability of facilities.

For more accuracy, future studies should also look at economic conditions that drive the resale HDB markets such as interest rates and government initiatives such as the The Prime Location Public Housing (PLH) Model announced by HDB recently.