Hiew Shani A0259557L **DSE1101 Individual Project**

Pre-analysis

My chosen dataset is the HDB resale prices dataset. With 6000 observations and 230 variables, I started by looking through all the variables and determining which I felt were more important based on prior knowledge and created some data frames with certain variables as well.

I split the 6000 observations into the training and testing data in a 2:1 ratio. I felt that it would be beneficial to have more training data as many different factors affect HDB resale prices and the model should be trained to recognise more patterns in the data at once. I scaled the resale prices down by 1000 for readability for all my models, as the resale prices are very large in magnitude. I also rounded all my found values throughout the report to 3 significant figures for simplicity.

Gaining insights into data

Given the wide range of variables and many observations available in this dataset, I decided to first use unsupervised learning to gain some insights into the data.

Using the Kernel Density Estimate with a bandwidth minimising the mean squared error, the plot as shown in figure 1 is obtained. From this plot, one insight we can obtain is that the most common HDB resale price lies around $470000, where a vertical line has been drawn in figure 1.

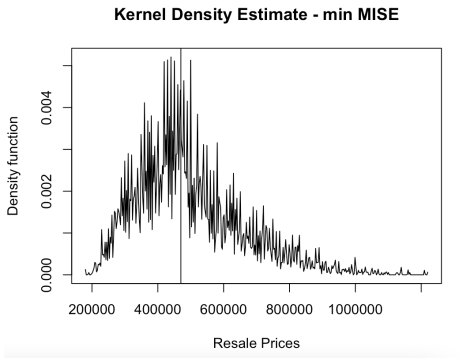

Figure 1: Kernel Density Estimate Plot for HDB Resale Prices

Plotting a decision tree (figure 2), I managed to get more insight into which variables are more important in determining HDB resale prices. As shown in figure 2, floor area seems to be the top predictor, followed by max floor level and distance from the CBD.

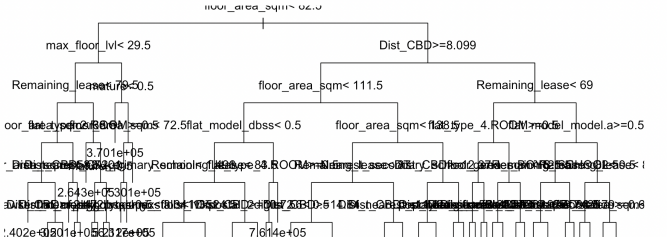

Figure 2: Part of a decision tree plotted using all variables

Thereafter I used simple and multiple linear regression to gain more insights into how the more significant variables as identified by the decision tree affect HDB resale prices.

Performing a multiple linear regression with the data of the town the HDBs are situated in, it was found that a HDB flat situated Bishan, Bukit Timah, Central Area, Queenstown would see the highest increase in resale prices of more than $200000.

Performing a simple linear regression between the HDB remaining lease and resale prices, it can be deduced that a 1-year increase in the remaining lease increases the flat resale price by $3940.

Performing a simple linear regression between the HDB flat floor area and resale prices, it can be deduced that a 1 square metre increase in floor area increases the flat resale price increase by $4300.

Performing a simple linear regression between the distance to CBD and resale prices, it can be deduced that for every 1 km increase in the distance from the flat to the CBD, HDB resale price decreases by $10100.

Performing a simple linear regression between the maximum floor level and resale prices, it can be deduced that for every 1 storey increase in the maximum floor level of a HDB block the flat is situated in, HDB resale price increases by $11200.

Predicting HDB Prices

Multiple Linear Regression

As the categorical variables are dummy coded in the dataset, to reduce the number of variables included in my models, I have only included continuous variables. I first included the variables that were already determined to be more important in the previous section, namely the remaining lease, floor area and maximum floor level and distance from CBD. However, when only these few variables are included, the adjusted r squared is 0.7946.

As such I decided to include more variables, one of which is mature[1]. When mature is added, the adjusted r squared increases by quite a bit to 0.8157, indicating that mature is a significant variable to be included in the model. I decided it was also good to add in as a proxy for all the towns a flat could be located, which requires a lot more variables to be included.

I also decided to investigate whether the distance to the nearest MRT and distance to the nearest CC[2] were significant. Including distance to the nearest MRT only increased the adjusted R squared to 0.8183, an increase of 0.0026, while including distance to the nearest CC increased the adjusted R squared to 0.8235 an increase of 0.0078. This was to my surprise as I thought having a flat near a HDB would be valued more than it being near a CC. As such, I added distance to the nearest CC in the model.

Something I was curious about was also whether distance to the nearest mall and distance to the nearest primary school (which property agents often include in listings) greatly affect the HDB resale price. After adding both separately to the model, I figured they were not that significant, with both only increasing adjusted R squared by around 0.001.

My final multiple linear regression model is given by the following equation:

-258.96836 + 4.44513 Remaining_lease + 4.68429 floor_area_sqm - 10.99279 Dist_CBD + 5.27188 max_floor_lvl + 72.36966 mature - 50.40291 Dist_nearest_CC

K Nearest Neighbours (KNN)

To create an appropriate KNN model, I used Leave one out cross-validation on the test set to identify the best K to use and used the same predictors as I investigated above for multiple linear regression.

Decision Tree

As the decision tree will select more significant variables on its own with the CART algorithm, this was the simplest to implement.

Model Comparison

Mean squared error for *multiple linear regression*: 4774.318 thousand SGD

Mean squared error for *K nearest neighbours*: 2528.39 thousand SGD

Mean squared error for *Decision tree*: 3377.469 thousand SGD

Comparing the mean squared error (MSE) between the 3 models, multiple linear regression with the largest MSE performs the worst, while KNN performs the best with the lower MSE.

[1] Mature = 1, means that HDB flat is located in a mature estate, and otherwise if mature = 0

[2] Community Club/Centre

Testing models on actual data

I obtained details for 2 HDB flats currently being resold from PropertyGuru. Given the lack of complete information available, I could not use the decision tree which requires input for all the variables to predict the prices.

For the Kallang/Whampoa HDB flat (117B Jalan Tenteram), the actual resale price currently being listed is $840000. My multiple linear regression model predicted a resale price of $743000 (3s.f.), while my KNN model predicted a resale price of $837000.

For the Clementi HDB flat (440A Clementi Avenue 3), the actual resale price currently being listed is $750000. My multiple linear regression model predicted a resale price of $661000 (3s.f.), while my KNN model predicted a resale price of $777000.

It can be seen that the KNN model definitely works better, both in theory as based on the MSE, and in practice as the predicted prices were much closer to the actual price than using linear regression. This is likely because not all the variables included have a close linear relationship with resale prices, and there would be more bias introduced in the linear regression model as compared to the KNN model.

Limitations

There are indeed many other factors which would affect resale prices. Regarding the flat itself, whether the flat is furnished and renovated well will have an impact on resale prices, which would be why listings often have multiple photos of the interior and state how well-furnished the flat is.

Additionally, for my actual predictions, the data is from 2022, while the data my models are trained on come from 2021. As such, there could definitely be changes in the statistics since a year ago which could be significant but not taken into account. There has also been inflation, meaning that the current resale prices are not equivalent to the resale price of the same price a year ago, which my predictions with real data did not account for. A more accurate analysis of how well the model predicts would be done by comparing the current resale price in 2021 Singapore dollars to the predicted price in 2021 Singapore dollars. Resale prices also depend on market conditions, like the demand and supply of flats which also depend on factors not included in the dataset, such as the expectation of future prices which will be hard to determine even with additional information.

Definitely, knowledge of how to create more plots would also help with better visualisation of the data, given a large number of variables in this dataset.