Goh Chee Wei Christopher

The dataset that I have selected is HDB resale prices 2021. This data is taken from open sources (Data.gov.sg). This dataset is of personal interest to me as my mom works at HDB so this was an excellent opportunity for me to play with data related to her workplace. Additionally, I find it fascinating how I can use the concepts that I learned in DSE1101 and apply them to real-world data. Predicting resale prices involves building predictive models, which is a fundamental skill in data science. The models that I will be using are Decision Trees, Multiple Linear Regression, and K-nearest Neighbours (KNN). Since this is a regression problem, the models are evaluated based on the out-of-sample mean squared error (MSE) using a test set.

Data Preparation:

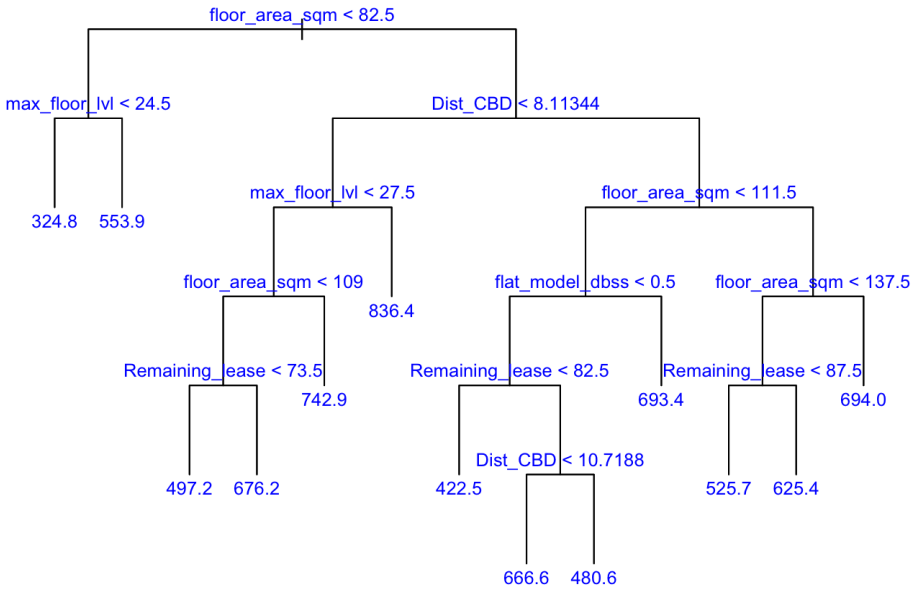
I checked for missing values and handled them by omitting the respective rows. The resale prices are scaled by dividing them by 1000 for ease of interpretation. I performed a 50/50 train-test split, with 3000 observations in the training set and the rest in the testing set. I also set a seed for reproducibility.

Variable Selection:

I first deduced a few variables that would be important in my study. The variable that I felt was the most important is the "floor_area_sqm" variable. This is a no-brainer as floor area is the first thing that shows up in every advertising brochure/website pertaining to a resale flat. It is so crucial that many resale websites such as propertyguru.com.sg have indicated a dollar-per-floor area metric, for buyers to judge the value of the house. Based on my understanding, with a larger floor area, homeowners would have a larger area to play around with, in terms of interior design and furniture and therefore, house valuation would be higher. Based on my economics knowledge, I also considered "Remaining_lease" and "Dist_nearest_station" to be crucial as well. When deciding which HDB to buy, rational buyers would consider the opportunity cost in making the decision. They would prefer to choose flats closer to MRT stations to cut down the implicit cost incurred which is the time taken to walk to the nearest MRT station. Similarly, they would choose HDBs with a higher remaining lease value so that they can cut down the time taken to search for a new flat.

Additionally, I used Decision Trees for variable selection. It is a non-parametric method which can perform regression tasks. I choose trees for many reasons. It is computationally efficient, has automatic interaction detection, handles both categorical and numeric X well and performs variable selection. I first fitted the entire data set and generated a big tree, using a small value mindev=0.0001 as a stopping criterion. The optimal tree size (13 leaves) was determined through 10-fold cross-validation. The original tree was pruned to the optimal size and the pruned tree is visualised (as shown in Figure 1) and assessed on the test data. Ultimately, the out-of-sample MSE is 6496.478.

Figure 1: Tree diagram



From the tree, the variables shown are:

1. floor_area_sqm 2. max_floor_lvl 3. Dist_CBD 4. Remaining_lease 5. flat_model_dbss

These variables will be used as inputs for Multiple Linear Regression and K-nearest Neighbors (KNN). The HDB_resale_readme.txt file in the HDB_resale_prices folder mentioned that the main variables are: resale_price, year, month, postal_2digits, town, remaining_lease, floor_area_sqm, max_floor_lvl, storey_range and flat_model. It was interesting to see that my tree model was able to capture 4 out of the main variables mentioned. This also shows how Decision Trees did a good job in variable selection.

Model fitting:

Decision Trees:

As mentioned previously, it was fitted using all the variables in the train data.

The out-of-sample MSE is **6496.478.**

Multiple Linear regression:

(1) Fitted using the selected variables from the tree model on train data and examined the diagnostic plots as shown in Figure 2a.

The red line in the diagnostic plots was generally horizontal, indicating linearity.

The out-of-sample MSE is **4964.732.**

(2) Fitted using the variables that I initially thought were crucial based on domain knowledge using the train data: (floor_area_sqm, Remaining_lease, Dist_nearest_station)

The out-of-sample MSE is **13697.36.**

(3) As an exploration exercise, I went to find out what factors affected the market value of BTO flats in Singapore. From this HDB website* I gathered a few factors: Proximity to city or town centre (Dist_CBD), Accessibility to key transport nodes(Dist_nearest_station), Accessibility to key amenities, such as supermarkets, food centres or malls(Dist_nearest_mall), Larger flat sizes(floor_area_sqm) and Higher storey heights(max_floor_lvl). Logically these factors should also affect the resale price of HDBs and I wanted to test that. I mapped those factors to their corresponding variables in my dataset and I fitted those variables using the train data.
The out-of-sample MSE is **6885.307.**

(4) Lastly, I fitted all the variables in the train data to compare their performance to my model.
The out-of-sample MSE is **425400.6.**

K-nearest Neighbours (KNN)

(1) Fitted using the selected variables from the tree model on train data. The optimal K = 4 is determined through cross-validation (as shown in Figure 2b) The final KNN model is evaluated on the test set.
The out-of-sample MSE is **3129.765.**

(2) Lastly, I fitted all the variables in the train data to compare their performance to my model.
The out-of-sample MSE is **5818.863.**

Figure 2a

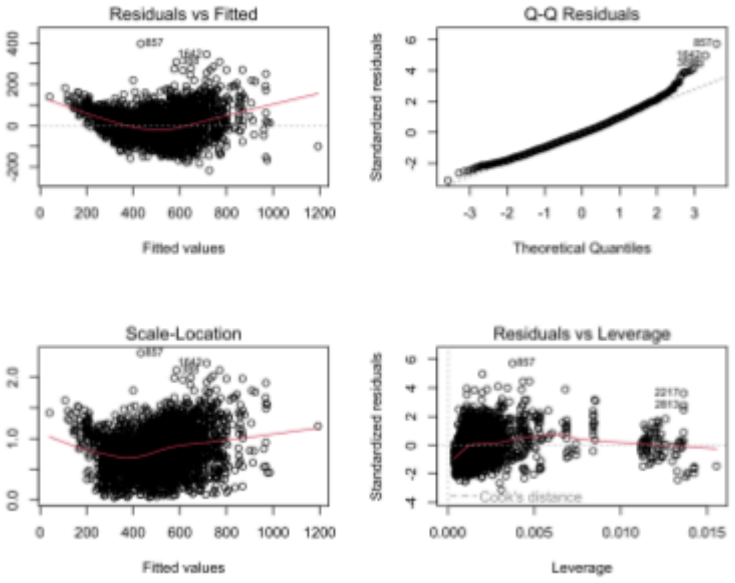
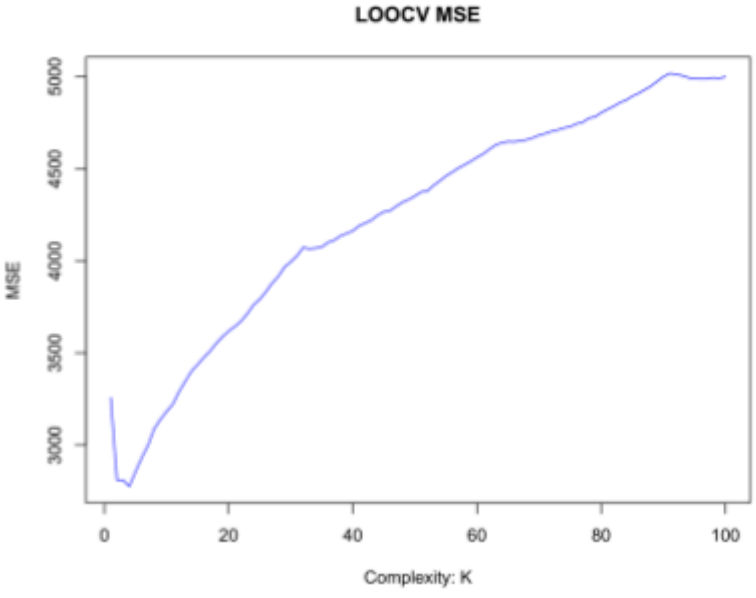


Figure 2b



Analysis and limitations:

Overall, KNN with a K size of 4 had the lowest out-of-sample MSE of **3129.765**, followed by Logit with **4964.732**. This indicates KNN's superior predictive performance over Logit and Decision Trees. In Week

8's Lecture, KNN performed slightly worse than logit, but in my case, KNN performed better. However, that was a classification, not a regression problem and I did not use 1 fewer predictor this time. Both KNN and Logit performed better using the variables that I have selected using trees compared to fitting all the variables. This is because, upon analysis of the Logit(1) summary, all the t-values of the variables are more than 1.96 or less than -1.96, indicating they are all statistically significant and thus helpful for my analysis.

Limitations include:

1. Trees usually lose out to other methods of prediction as they are highly dependent on training data. A small change in data can result in a large change in the estimated tree. This would ultimately affect the variables that I chose for my model.
2. For Logit(3): I ignored factors such as "More ideal orientation with desirable views" and "Prevailing market conditions" from the HDB website as there were no variables in my dataset that I could map those factors to. These factors are not included in my analysis and could improve MSE, which was otherwise not represented.

As an exercise, I used my Logit(1) model to predict the price of a resale flat: 418 Serangoon Central**. The inputs are as follows: floor_area_sqm = 146.0436, max_floor_lvl = 12(I live nearby so I checked that the max floor is 12), Dist_CBD = 9.4(Used the shortest distance to Downtown Core on Google Maps), Remaining_lease = 64(Used HDB Emap*** to input the address and find remaining lease), flat_model_dbss = 0(Not a DBSS flat). The predicted price is $705,932.2, whereas the actual price is $ 1,150,000. This discrepancy could be due to benefits not accounted for such as the fact that it is a maisonette and its closeness to schools, transport and amenities.

Conclusion

It was interesting to find out how in Logit(2), the initial variables that I chose turned out to not perform well in fitting the model. This made me reflect on my understanding and made me read up on the many other factors that affect resale prices. These insights gained will aid in my decision-making when buying a resale flat in the future. Overall, this was an eye-opening experience to be able to explore handling large datasets using the models that were taught in class.

References:

*https://www.hdb.gov.sg/about-us/news-and-publications/publications/hdbspeaks/How-BTO-Flats-are-Priced#:~:text=First%2C%20HDB%20establishes%20the%20market,individual%20attributes%20of%20the%20flats.&text=Projects%20with%20locational%20advantages%20will,to%20city%20or%20town%20centre

**https://www.propertyguru.com.sg/listing/hdb-for-sale-418-serangoon-central-24087219

***https://services2.hdb.gov.sg/web/fi10/emap.html