**Part 1: Introduction**

Within the past decade, the Singapore property market has been on a rise, particularly the Housing and Development Board (HDB) which motivates me to explore the driving factors behind HDB resale prices. The HDB 2021 by data.gov.sg includes HDB data from 2021, including information like resale prices and other characteristics of those flats. I will primarily be using the dataset to model the resale prices against the characteristics in the dataset by selecting the best suited variables and creating a model that will predict the resale prices of these flats.

**Part 2: Data Description**

The data I have chosen to use is an open source dataset by data.gov.sg, curated by Nathanael Lam Zhao Dian in the subsample of his Honours Thesis dataset. Before analysis of the data, preprocessing was first done to the data. These steps included: 1) normalizing the resale prices by dividing it by $1,000 to make the data more interpretable. 2) Removing variables that are linearly dependent and have near-zero variance to prevent collinearity. 3) Applying a 80-20 train-test split to test how the model would perform against unseen data. A seed number of 100 is also set to ensure that results can be replicated.

**Part 3: Data Analysis**

3.1 - Interesting Insights

A graph of a high floor category

Description automatically generated with medium confidence

Fig. 1 – Average Resale Price of HDB units by Floor Category

In Fig.1, I categorized the “storey\_range” variables into a new categorical variable ‘floor\_category’ with 3 classes-‘High Floor’ (floors 19-51), ‘Mid Floor’ (10 to 20), and ‘Low Floor’ (1 to 9). I was able to draw some insights on how floor range affects resale price. As shown in the blue bar chart above, higher floors yield higher average resale prices, with ‘High Floor’ units representing highest average resale price and ‘Low Floor’ units representing the lowest average resale price.

A screenshot of a computer code

Description automatically generated

Fig. 2 – Average Resale Price and Average Floor Area by Flat Types

Following this, in Fig. 2, I categorized the ‘flat\_type’ variables into a new categorical variable ‘flat\_type’, representing the different unit types (1 Room, 2 Room, 3 Room, 4 Room, 5 Room, Executive and Multi Generation), followed by grouping the average resale price and floor area by unit types. From this, I could infer a positive correlation between flat type and average floor area, with executive flats having the largest average floor area of 145 sqm. Additionally, there appears to also be a positive correlation between average floor area and average resale price, with Multi-Generation flats being an exception. This led to an exploration that maturity of estates these Multi-Generation units are located in could be the factor influencing resale prices.

A screenshot of a computer program

Description automatically generated3.2 – Base Model Prediction with Multiple Linear Regression

A screenshot of a computer

Description automatically generated  
Fig. 3 – Variables Selected in the Linear Model

A computer screen shot of a computer

Description automatically generated

Fig. 4 – Results of Base model

Firstly, I fitted the full linear model with all it variables after the pre-processing step, followed by applying backwards propagation, iteratively removing the least important variable, from this, I determined the model that yield the greatest Adjusted R-Squared value (0.9006071). Fig. 4 shows all the variables selected for this model. After extracting the variables selected above and fitting it again into a new linear model, I tested the performance of this model against the test dataset. From Fig. 3, this linear model yielded a Mean Squared Error (MSE) of 3173.5 and Root Mean Squared Error (RMSE) of 56.3.

3.3 - Prediction with KNN

A number and a number

Description automatically generated with medium confidence

Fig 5. Results of KNN Model

Next, KNN was tested against the test data, to determine the optimal K value to be used, Leave One Out Cross Validation (LOOCV) was used. K = 4 was the best value. This value of K was then used in the KNN model. This model yielded an MSE of 7251.3 and an RMSE of 85.2. This indicates that the KNN model does not perform as well as the multiple linear regression.

3.4 - Prediction with Decision Tree

A diagram of a floor

Description automatically generated

Fig. 6 – Plot of Decision Tree Used

A number of trees with numbers

Description automatically generated with medium confidence

Fig. 7 – Results of Decision Tree Model

Finally, a decision tree model was used, by first fitting the entire linear model with all the similar variables, the best Complexity Parameter (CP) was determined to be 0.001. This value was then used to prune the tree to form the new model to be used in prediction. As shown in Fig. 6, the root node being “floor\_area\_sqm < 82.5” indicates that this is the most important predictor. When this model is used to predict the test data, it yielded a MSE of 4292.2 and RMSE of 65.5. This indicated that the decision tree model performed better than the KNN model but worse than the base model.

4 – Conclusion

4.2 – Performance of Models

A greater MSE and RMSE value indicates that there’s greater deviation between the prediction and actual test values, indicating that the model performs worse on the test data. Since these models are fitted with the purpose of predicting values on unseen data, or our test data, we can gauge a model’s performance based on how it performs on the test data. Hence, it is fair to say that a model with a lower MSE and RMSE value is a better model. Based on the results, the base model performs the best with the lowest MSE and RMSE value, followed by the decision tree and the KNN model. KNN performs the worst as it is a simplistic approach of using the distance between data points which could be affected by noise. The base model works the best as the HDB resale prices follow a linear price movement, followed by the decision tree as it handles complex relationships but may overfit and produce less accurate predictions.

4.1 – Limitations of This Method

There are some constraints when we compare the models this way, primarily in feature selection. By selecting the best variables to use for the base linear model and using it for the KNN and decision tree model, the linear would perform better since it is the best selected variables for this model, however, if all predictors were to be used, the high-dimensional data would cause the models to perform worse as well. Therefore, although there might be some bias in variable selection, this method ensures consistency among the models in variable selection which allows for a fair comparison of the effectiveness of these models.

**Appendix**

I acknowledge the use of ChatGPT to debug the summary and grouping of data in my code that was subsequently included in my report. I entered the following prompts: “how do I create a new variable that categorizes flat\_type”, “how do I group the average resale\_price and floor\_sqm by flat\_type”, “how do I create a new variable that categorizes storey\_range”, “how do I extract the variables used in this model”, “why does this return an error” -> ChatGPT then subsequently returned a debugged code that removed the near zero variance and collinear variables for this prompt.

I acknowledge the use of ChatGPT to debug my graph. I entered the following prompts: “why does the graph return this error”