CHEMICAL AND BIOMOLECULAR ENGINEERING

CB0494 – Introduction to Data Science & Artificial Intelligence

Problem Statement: Predicting Resale Prices for Informed Decision-Making

Name:	Eugene Chia Kai Jun
	Ian Joshua Sainani
Group:	C22 (Group 14)
Date:	26/03/2024

Table of Contents

List of Tables

List of Figures

1. Background

For individuals exploring resale houses, important considerations such as the storey of the house, location, and property size are crucial. Understanding how these factors affect resale prices and having an estimated price enables informed decision-making when purchasing resale houses. Given the substantial investment that real estate represents for individuals, these understandings are crucial for securing properties aligned with their preferences and financial objectives.

Therefore, is it possible to predict the resale price of a flat based on the prospective owner's preferences (such as location, storey range, floor square area, etc.), enabling them to make informed decisions about their expected expenditure?

2. Objectives

This study aims to predict resale prices, providing valuable insights for individuals and enabling informed choices.

3. Setup/Technique Used

The data frame undergoes initial cleaning before any coding or analysis begins. During this process, columns such as "street name" and "lease commence date" were removed, as they were deemed unnecessary. Additionally, the remaining lease is expressed as months.

3.1 Correlation Matrix

Firstly, the analysis aims to determine the correlation between factors such as floor square area, remaining lease duration, and the storey level against resale prices. To visualize these correlations, a correlation matrix was used. Based on the values of the correlations, it was then deduced which factors should be used to create a bi-variate joint plot to assess their suitability as predictors.

3.2 Bi-variate Joint Plot

Outliers were removed using the LocalOutlierFactor(LOF) function before plotting. Subsequently, the R^2 values were examined to assess whether the model sufficiently accounted for the variance in the response data around its mean. This evaluation allows one to determine the effectiveness of the model as a predictor.

3.3 Box and Whisker Diagram

To determine if location of the house has an impact on the resale price, a box and whisker plot comparing prices across various towns was employed. This clearly visualizes the distribution of prices, highlighting disparities, if any, in median prices among different towns. Variations in the median suggest that the town could be a contributing factor affecting resale prices.

3.4 Transforming Categorical Variables into Numerical Format

Since the machine learning algorithm requires numerical input variables, it is essential to convert categorical variables (town, flat model, flat type etc.) into a numerical format. The

"pd.get_dummies" function in pandas was employed for this purpose. It generates dummy numerical variables by creating new columns for each unique value present in each original categorical column. Each observation is then represented by a combination of "1" and "0" across the newly created columns, where "1" indicates the presence of the specific category and "0" indicates its absence. This transformation allows categorical data to be effectively incorporated into the machine learning algorithm.

3.5 Multi-Variate Prediction

Once the categorical data were converted into numerical format, they can be used in a multi-variate linear regression model. This expands the model's capabilities to capture complex relationships and improve predictive accuracy. Pre-processing steps such as feature scaling and selection were crucial in preparing the dataset for modelling. Without scaling, features with larger magnitudes or ranges may dominate the learning process, resulting in biasness towards those features. Scaling ensures that all features contribute proportionally to the model's learning process, thereby promoting a fair representation, and preventing dominance by any particular feature.

4. Results and Discussion

4.1 Analysis of results

Upon examination of the correlation matrix shown in Figure 1, it can be observed that the floor square area and resale prices exhibit the highest correlation coefficient of 0.6. While this value may not indicate a strong correlation, it signifies a moderate relationship between the two variables. Given that the correlation between floor square area and resale prices is the highest, a bivariate joint plot was employed to assess its suitability for predicting resale prices based on floor square area.

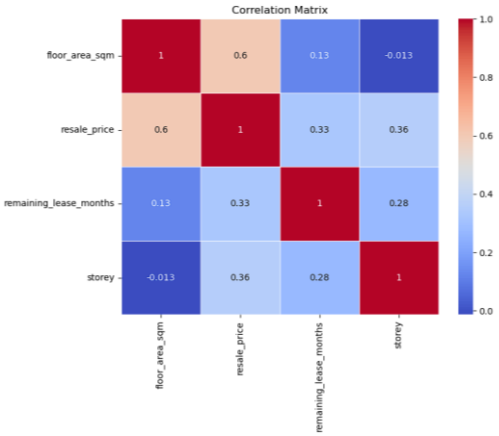

Figure 1: Correlation Matrix

Upon employing a bivariate joint plot, the R^2 was observed to be approximately 0.369, suggesting that the model inadequately explains the variability of the response data around its mean, despite removing the outliers. Thus, it can be concluded that floor square area may not be a reliable predictor of the resale price.

Subsequently, a box and whisker plot comparing resale prices across various towns was employed. It is observed in Figure 2 that there are noticeable variations in median prices among different towns, with "Bukit Timah" exhibiting the highest median prices, followed by "Queenstown". On the other hand, "Ang Mo Kio" and "Yishun" emerge as the towns with the lowest median prices. This indicates that resale prices are influenced by the location of the house.

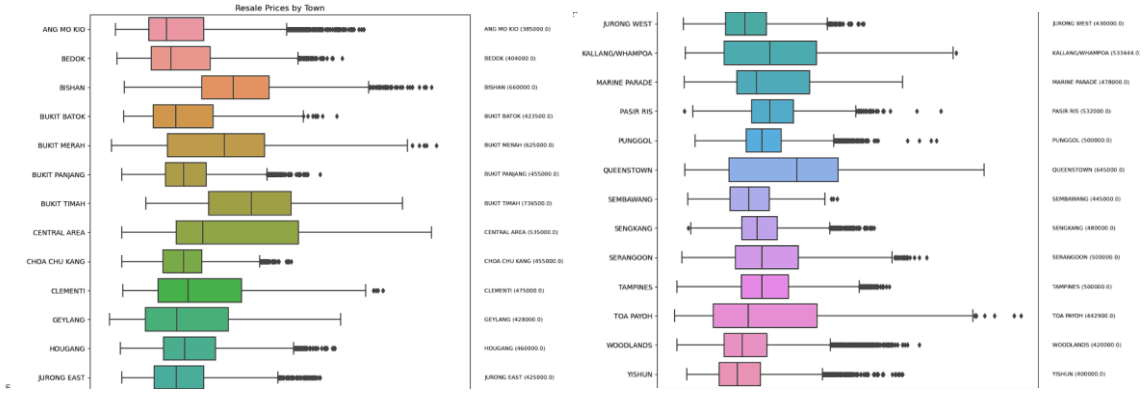

Figure 2: Box and Whisker diagram of Resale Prices in Different Towns

Subsequently, a multiple linear regression model was employed to predict the resale prices. Figure 3 illustrates a graph depicting the actual and predicted resale prices for the initial 250 samples from the multivariate prediction model. Additionally, a comparison graph was plotted to compare the predicted prices against the actual prices. In this graph, the red-dotted line represents a perfect prediction, where the predicted prices equal the actual prices(y=x), as shown in Figure 4.

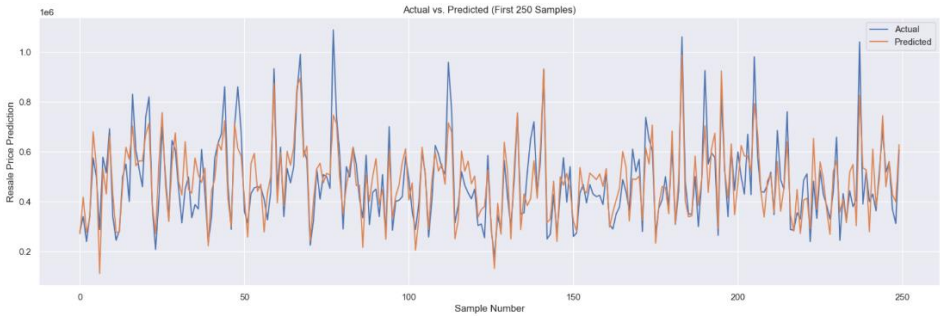

Figure 3: Comparing Actual and Predicted Resale Prices for the Initial 250 Samples

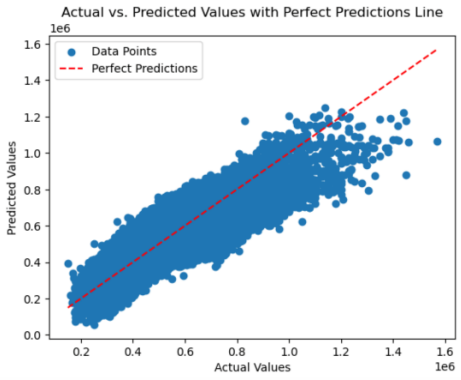

Figure 4: Linear Regression Model of Predicted Values vs Actual Values

The multiple linear regression model developed achieved a R^2 value of 0.75, indicating a reasonably good fit of data. Additionally, the root mean square error (RMSE) of approximately $85,000 suggests that, on average, the model's predictions are within this margin of error from the actual resale prices. Thus, the model can be considered to be fairly accurate in its prediction of the resale prices. Furthermore, it should be noted that RMSE are sensitive to outliers.[1] These outliers may be due to houses with abnormally high resale prices.

4.2 Error Analysis

To analyse the shortcomings of the linear regression model, it is crucial to understand the features of the datasets that exhibited the largest absolute error between the predicted and actual resale price.

It was found that among the top 100 datasets with the largest absolute error, in terms of flat type, 58 of which were 5-room flats, and 23 were executive flats. In terms of town, 21 belonged to “Bukit Merah”, 13 to “Ang Mo Kio” and 12 to “Toa Payoh”.

Referring to Figure 2 and Figure 5, these features consist of many outliers, and exhibit a box plot of large spread, or rather, variance.

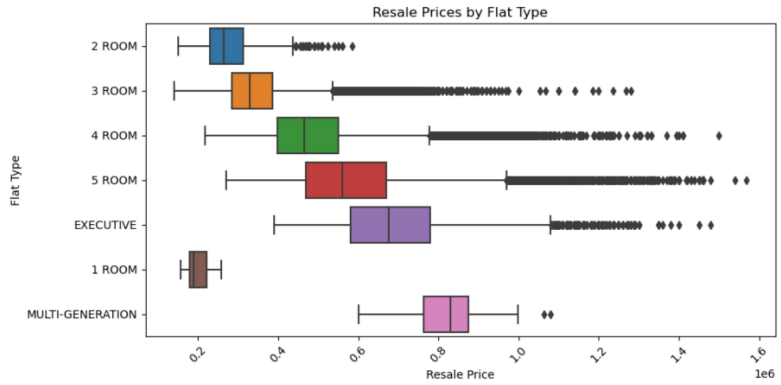

Figure 5. Box and Whisker diagram of Resale Prices of Different Flat Types

As such, anomaly detection is crucial to detect and remove these abnormal behaviours from the dataset.

5. Future Work

To improve the accuracy of the prediction, other algorithms can be explored such as Random Forest and Extreme Gradient Boosting(XGBoost) can be explored. Additionally, outlier detection can be done to reduce the RMSE.

5.1 Anomaly Detection

Multivariate Anomalies occur when the values of various features, taken together seem anomalous even though the individual features do not take unusual values.[2] Some common outlier detection methods such as Local Outlier Factor(LOF) , Isolation Forest and Robust Random Cut Forest can be employed.

5.2 Random Forest

Random Forest is a widely employed machine learning algorithm, developed by Leo Breiman and Adele Cutler. It combines results from multiple decision trees to produce one outcome. Its simplicity and adaptability make it widely used for both classification and regression problems, with a reduced risk of overfitting by tightly fitting all samples within the training data. Moreover, Random Forest provides flexibility, handling both regression and classification tasks with high accuracy. Despite its capability to manage large datasets and deliver precise predictions, the computational process can be slow due to the computation involved for each individual decision tree.[3]

5.3 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting(XGBoost) employs a Gradient Boosting technique with the use of decision trees. It iteratively constructs short, fundamental decision trees, where each tree is referred to as a "weak learner" due to its high bias. Initially, it builds a basic tree that has poor performance. Subsequently, it creates another tree trained to predict what the initial weak learner cannot. This sequential process produces weaker learners, each correcting the errors of the previous tree until a stopping condition is met, such as reaching the specified number of trees (estimators). [4]

It has an outstanding performance across diverse datasets, outpacing conventional gradient-boosting methods in terms of speed, accuracy, and precision.[5] However, it does not perform well on sparse and unstructured data.

6. Conclusion

It can be concluded that factors such as remaining lease duration and the floor level of the house exhibit weak correlations with resale prices, whereas the floor square area shows a moderate correlation (0.6) with resale prices. Furthermore, variations in median prices across different towns suggest that the location of the house is a factor affecting the price, with Bukit Timah having the highest median price of $736,500, while Ang Mo Kio having the lowest median price of $385,000.

Furthermore, a multiple linear regression was employed to forecast resale prices based on the provided training data. A graph for visualisation was subsequently employed to compare the predicted values against the actual values to assess its accuracy, where the red-dotted line represents a perfect prediction(y=x). The analysis reveals that the model exhibits moderate accuracy in forecasting resale prices with a R^2 value of 0.75 and RMSE of $85,000, serving as a valuable information for individuals to make informed decisions with regards to their anticipated expenditures. Nonetheless, other algorithms can be explored to determine if they can offer a more accurate and precise prediction.

7. Contribution of work

Table 1: Contribution of each member

	Contribution of work done	
	Eugene	Ian
Report	50%	50%
Programming	50%	50%
Presentation	50%	50%

References

1. Advanced pattern recognition tools for disease diagnosis - ScienceDirect. https://www.sciencedirect.com/science/article/abs/pii/B9780323905480000115.
2. Anomaly Detection in Python — Part 2; Multivariate Unsupervised Methods and Code | by Nitish Kumar Thakur | Towards Data Science. https://towardsdatascience.com/anomaly-detection-in-python-part-2-multivariate-unsupervised-methods-and-code-b311a63f298b.
3. What Is Random Forest? | IBM. https://www.ibm.com/topics/random-forest.
4. Subasi, A., Panigrahi, S. S., Patil, B. S., Canbaz, M. A. & Klén, R. Advanced pattern recognition tools for disease diagnosis. *5G IoT and Edge Computing for Smart Healthcare* 195–229 (2022) doi:10.1016/B978-0-323-90548-0.00011-5.
5. What are the advantages and disadvantages of xgboost regression? | 5 Answers from Research papers. https://typeset.io/questions/what-are-the-advantages-and-disadvantages-of-xgboost-4q3onp0ddu.