

**SCHOOL OF INFORMATICS & IT**

**Diploma in Information Technology**

# Machine Learning for Developers (CAI2C08)

# AY2024/2025 October Semester

**Project Report**

Student Name (Matric Number) : Mohamed Farook Mohamed Farviz (2301186I)

Tutorial Group : P02

Jenny LING (TP)

Tutor : Jin LEE (TP)

Submission Date : 24/1/2025

**Introduction**

The topic is housing, and the data set focuses on analysing the resale prices of HDB flats in Singapore from 1990 to 1999. By examining factors such as flat type, location, and floor area, the goal is to build a model that predicts resale prices based on these attributes.

Source: [data.gov.ag](https://data.gov.sg/datasets/d_ebc5ab87086db484f88045b47411ebc5/view)

**Dataset Description:**

The dataset includes the following attributes:

**Month**: Month of sale.

**Town**: Location of the flat.

**Flat Type**: Classification by room count.

**Block**: Block number.

**Street Name**: Name of the road.

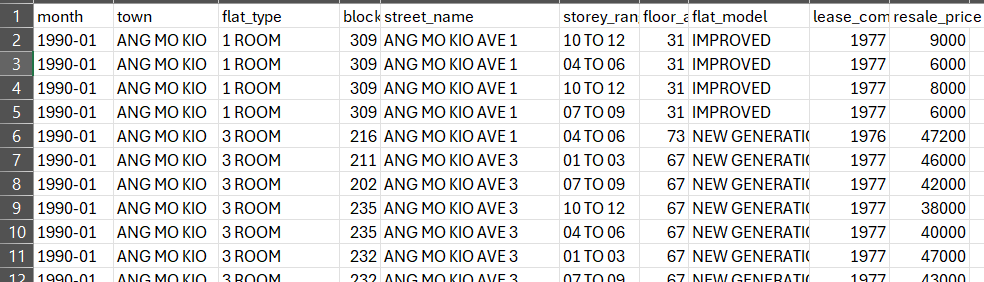
**Storey Range**: Floor range.

**Floor Area**: Interior space in square meters.

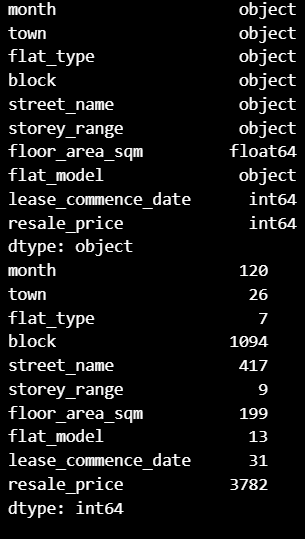
**Flat Model**: Model of the flat.

**Lease Commence Date**: Starting date of lease.

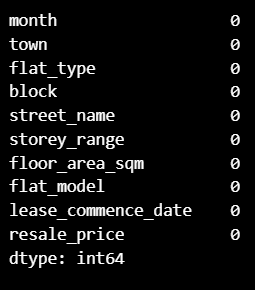
**Resale Price**: Price of the flat.



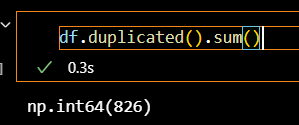
**Data Exploration and Pre-processing of data**



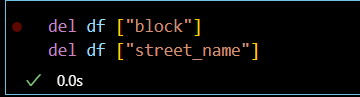
After importing the necessary packages and reading the dataset CSV file, the structure of the data was explored using df.dtypes to check the data types of each column, df.nunique() to identify the number of unique values in each column, and df.shape to examine the dimensions of the dataset. This exploration revealed the dataset contains multiple categorical variables such as town, flat\_type, and block, as well as numerical variables like floor\_area\_sqm and resale\_price. The insights helped in understanding the overall composition of the dataset and the diversity within each feature.



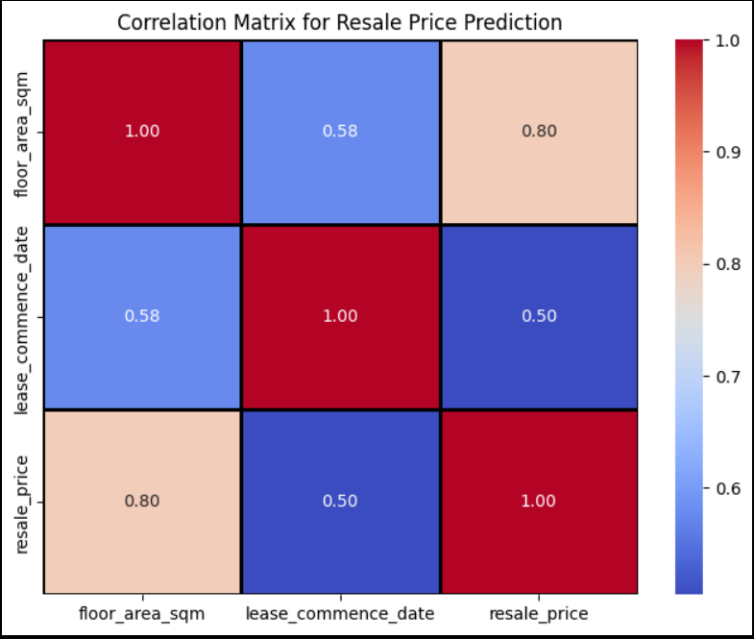
The dataset was checked for missing values using df.isnull().sum(), which showed that there are no missing values in any of the columns. This indicates that the dataset is complete and no further imputation or removal of missing data is required.

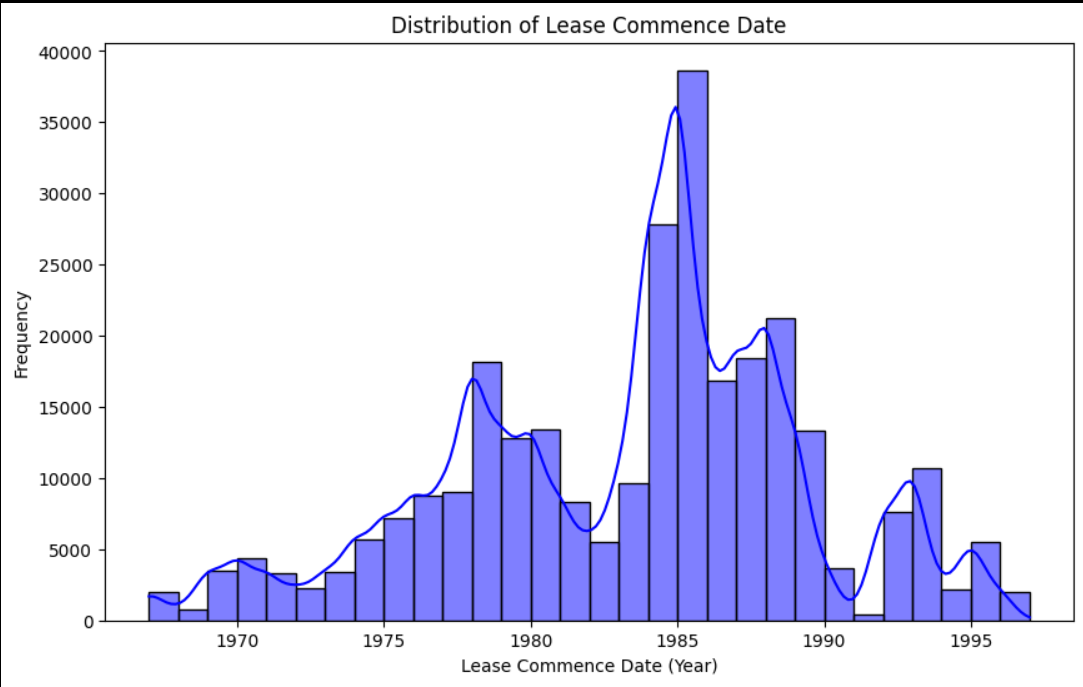


it was found that the dataset contained 826 duplicate rows, identified using df.duplicated().sum(). These duplicate rows were removed using the df.drop\_duplicates() method to ensure the dataset contains only unique records for further analysis.



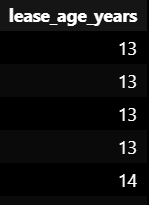
The block and street\_name columns were removed as they don’t significantly influence resale price predictions. The block column lacks predictive value, while street\_name is redundant because the town column already captures the location's effect on resale prices.

The correlation matrix shows that floor\_area\_sqm has a strong correlation (0.80) with resale\_price. While lease\_commence\_date has a moderate correlation (0.50) with resale\_price, it remains important as the lease age, calculated as 99 minus the lease commencement year, can influence the resale price.

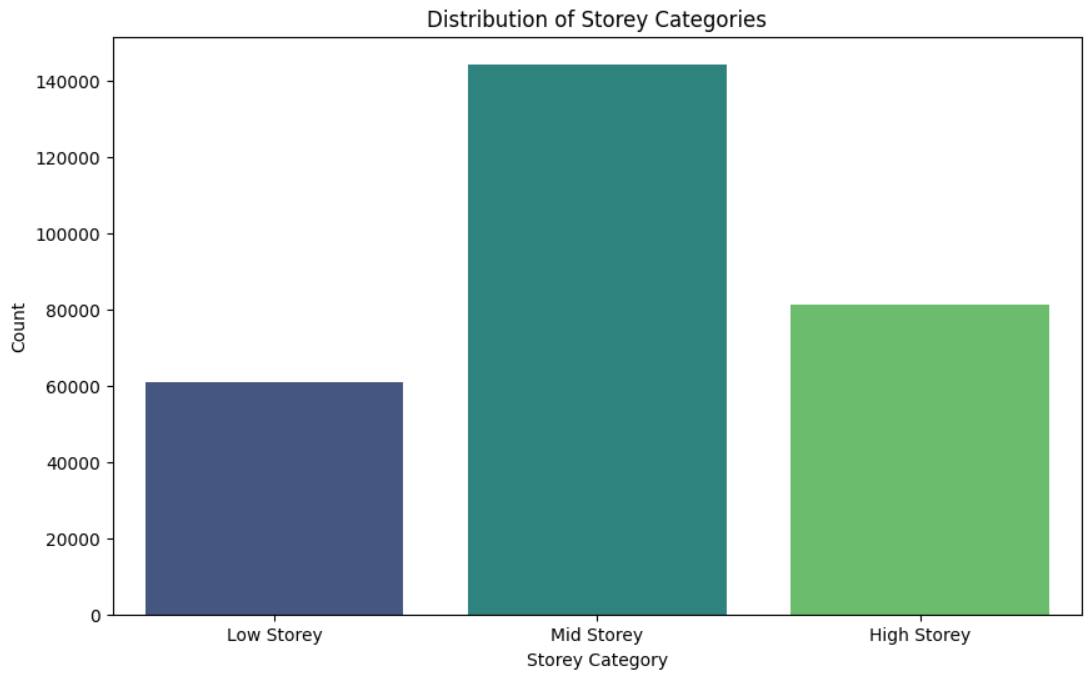




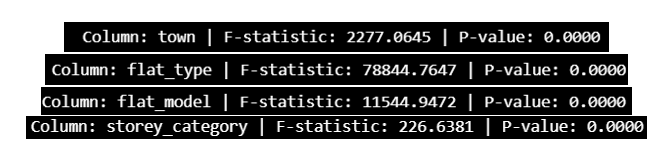




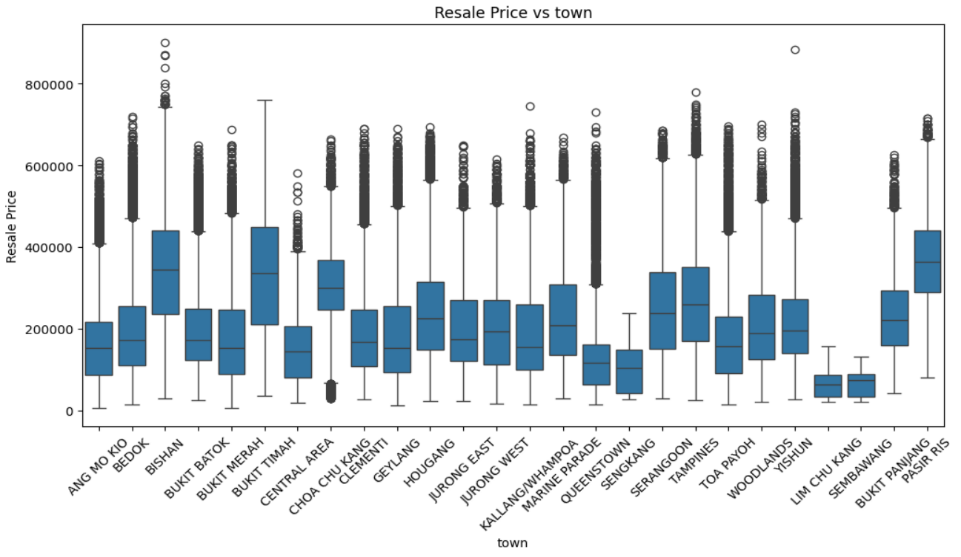
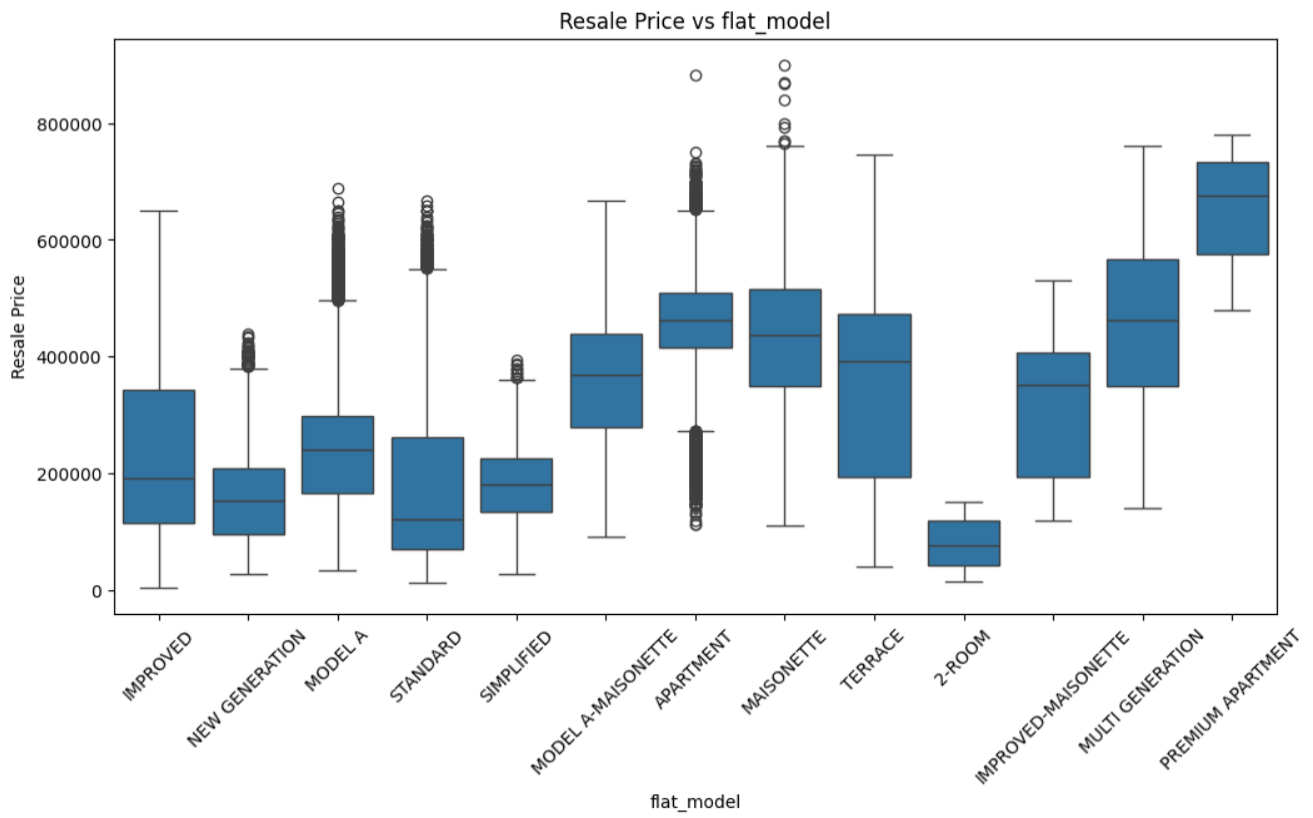
The distribution of the lease\_commence\_date is skewed, with most leases starting in the 1980s and fewer in the later years. This skewness can affect model performance, especially for algorithms like linear regression. To address this, the lease\_commence\_date was combined with the month column to create a new feature, lease\_age, which focuses on the age of the flat. This simplifies the data and removes the bias from the skewed distribution, providing a more effective feature for predicting resale prices. The original columns were removed.

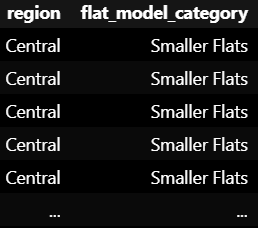


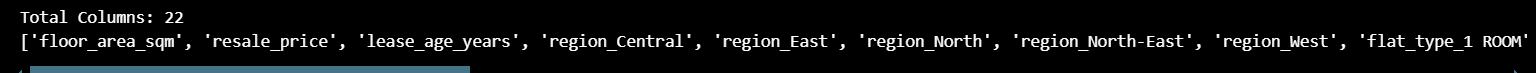
The storey\_range column originally contained raw floor ranges (e.g., "01 TO 03", "04 TO 06"), making it difficult to analyze and use in modeling. The issue was that the data was in a range format, not useful for prediction. To address this, the highest floor from each range was extracted, and percentiles (25th and 75th) were calculated to define thresholds. These thresholds were used to categorize the storey data into three categories: "Low Storey", "Mid Storey", and "High Storey". This transformation simplifies the raw range data into structured categories, making it more suitable for model training.



ANOVA was used to check the correlation between categorical variables (town, flat\_type, flat\_model, and storey\_category) and resale\_price. The results showed that all the categorical variables have a very low p-value (< 0.05), indicating strong statistical significance. This suggests that these features are important for predicting resale prices accurately. However, variables like town and flat\_model contain a relatively high number of unique values, which could impact the model's efficiency.





The town and flat\_model columns initially had too many unique values, complicating analysis. To simplify, town was mapped into broader regions (e.g., "ANG MO KIO" to "Central") and flat\_model was grouped into categories like "Smaller Flats", "Maisonettes", and "Larger Flats". This reduced complexity and made the data easier to analyze. The original columns were then removed and the new columns were added such as region and smaller flats.

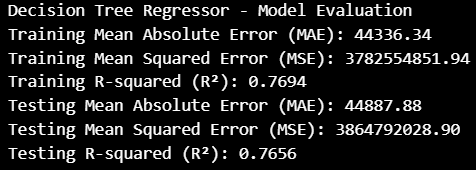


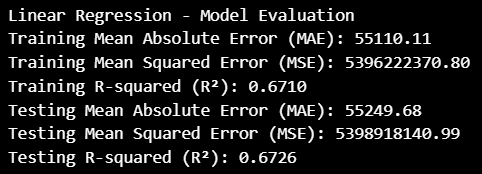
After cleaning and preparing the dataset, One-Hot Encoding (OHE) was applied to the categorical columns region, flat\_type, flat\_model\_category, and storey\_category. This transformation converted each unique category into separate binary columns, allowing the categorical data to be represented numerically for machine learning models. These columns are crucial as they capture key factors influencing resale prices, such as location, flat type, and storey category. The resulting dataset is now fully encoded and ready for predictive modeling.

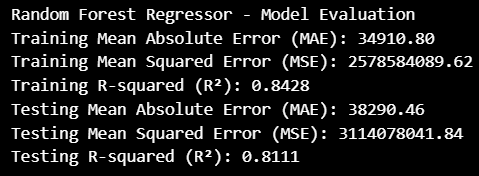
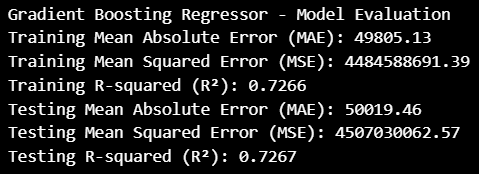
**Methods and Improvements**

The dataset is split into x and y. The features include all columns except for resale\_price, which serves as the target. Then, using train\_test\_split from sklearn, the data is divided into training and testing sets. 80% of the data is used for training the model, while the remaining 20% is used for testing. The random\_state is set to ensure consistent splits across runs.

The models used in this evaluation were Linear Regression, Decision Tree Regressor, Random Forest Regressor, and Gradient Boosting Regressor. These models were selected for their effectiveness in regression tasks, where the goal is to predict a continuous output, such as resale price in this case. Linear Regression is a simple model used for its interpretability and ease of implementation. Decision Tree Regressor captures non-linear relationships in the data, making it useful for complex datasets. Random Forest Regressor, an ensemble method, builds multiple decision trees to improve accuracy and reduce overfitting. Gradient Boosting Regressor also combines multiple weak learners to provide accurate predictions. These models were chosen to ensure comparability in predicting resale price, as they vary in complexity and performance, allowing for a thorough evaluation of their effectiveness for this specific task.

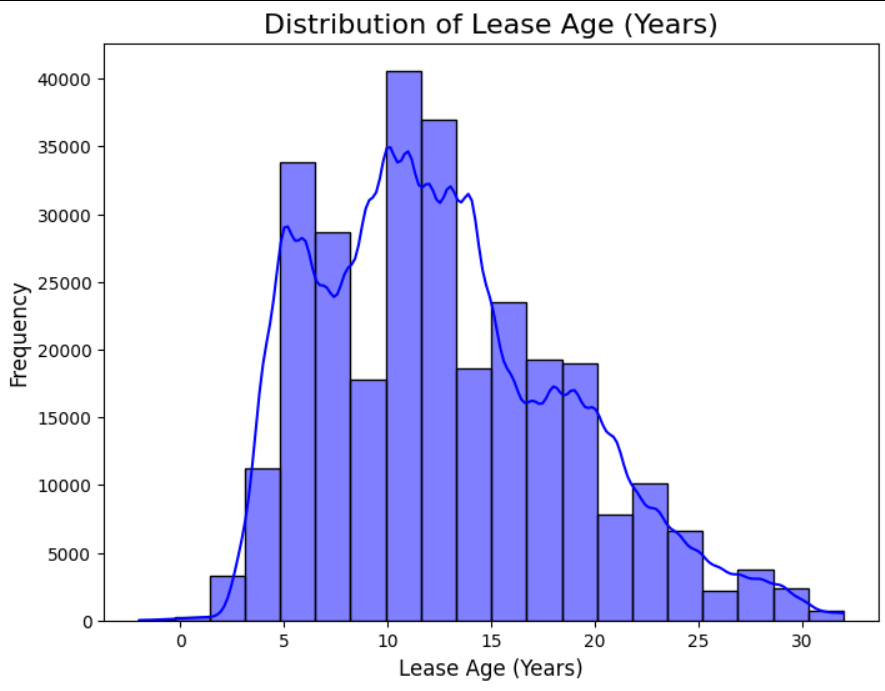
.

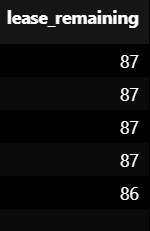


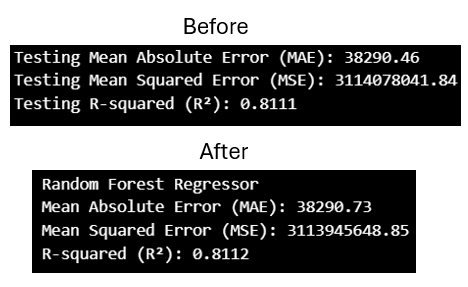


In model evaluation, Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R²) were used to assess the model's accuracy. MAE represents the average magnitude of errors between predicted and actual values, while MSE emphasizes larger errors by squaring the differences. R² indicates the proportion of variance explained by the model, with higher values (closer to 1) reflecting a better fit to the data. These metrics collectively help in determining the accuracy of the model’s predictions.

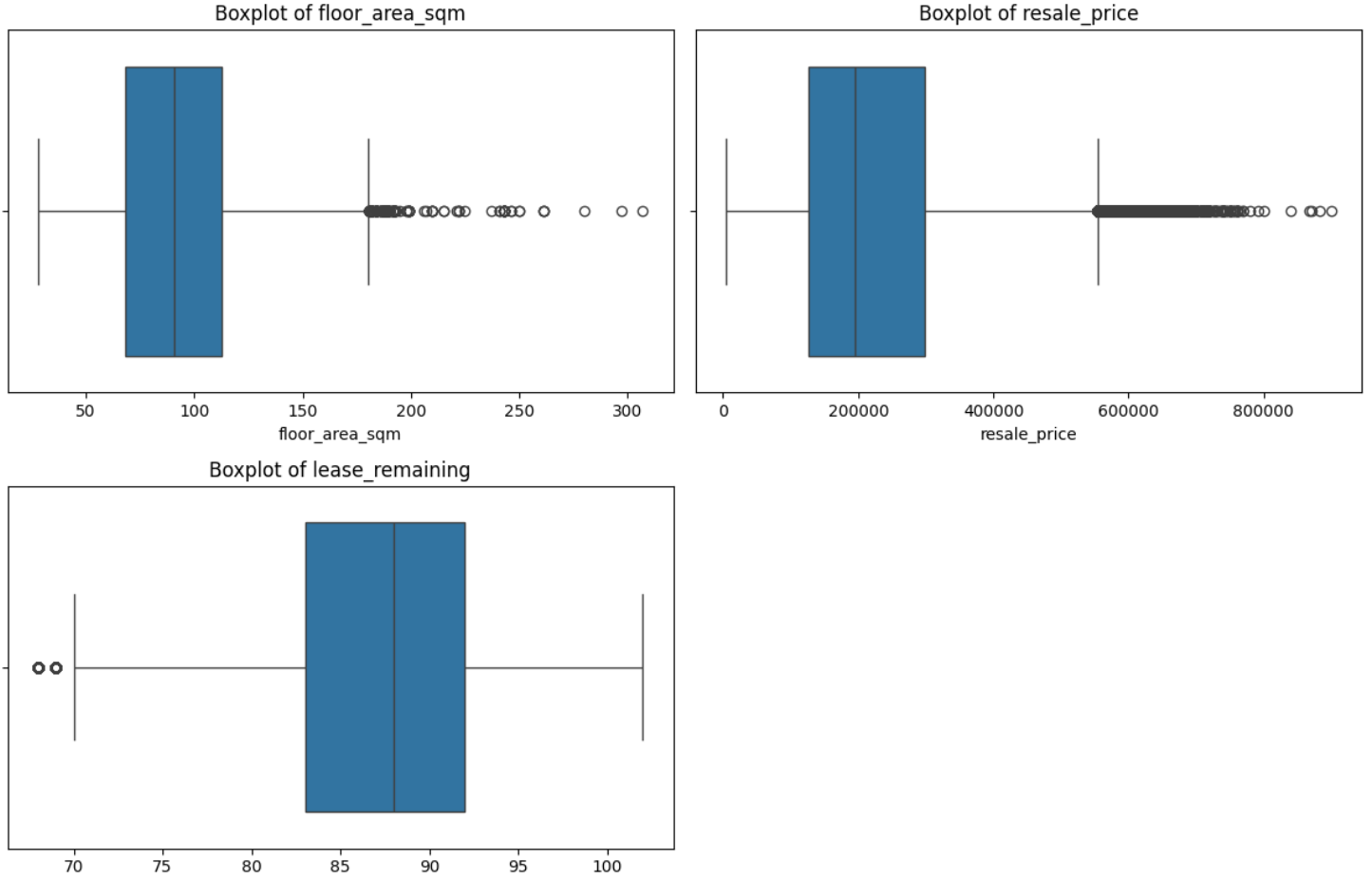
From the comparison of the models, the Random Forest Regressor performs the best with the lowest MAE (38,290.46) and MSE (31,140,780,841.84), along with a highest R² value (0.8111). The Training and Testing R² values are quite close (0.8428 vs 0.8111), indicating that the model is not overfitting or underfitting. Overall, the Random Forest Regressor gives the most accurate predictions without any significant overfitting or underfitting and so it’s a optimal fit.

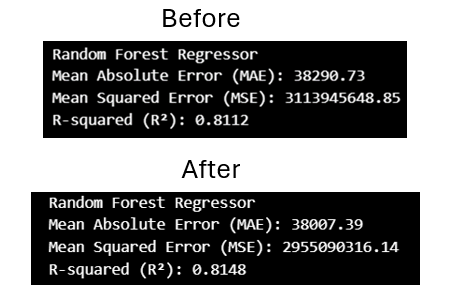


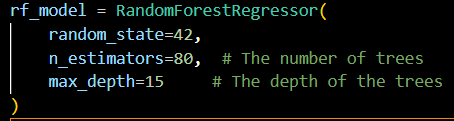




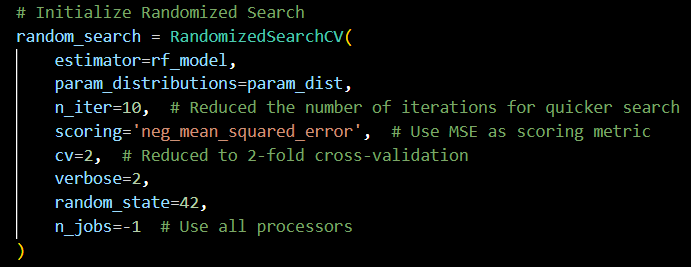
For improving the model, the lease\_age\_years feature was transformed into lease\_remaining, which represents the percentage of the lease left. This was done by subtracting the lease\_age\_years from the maximum HDB lease of 99 years and calculating the remaining years as a percentage. The lease\_remaining feature was then rounded to the nearest integer for clarity. After applying this transformation, the model's performance showed a very slight improvement, with the R-squared value increasing from 0.8111 to 0.8112. Though the improvement was marginal, this change made the data more interpretable and could help enhance the model's overall performance in the long run.

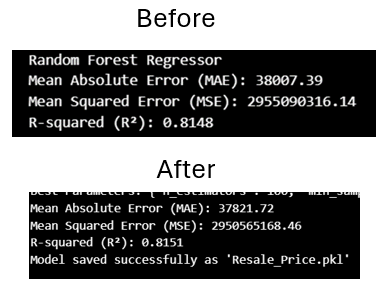






Another improvement was done as the outliers in the dataset were capped to minimize their impact on the model's performance, and the Random Forest Regressor was fine-tuned by manually adjusting its parameters. By limiting the number of trees and the depth of the trees, the model's accuracy improved slightly. After these adjustments, the Mean Absolute Error (MAE) dropped to 38,007.39, the Mean Squared Error (MSE) decreased to 295,509,316.14, and the R-squared value increased to 0.8148, showing better model performance compared to earlier results.





The model performance was further improved using RandomizedSearchCV to fine-tune the Random Forest Regressor. The hyperparameters, such as n\_estimators, min\_samples\_split, and max\_depth, were optimized to find the best combination. After running the search, the model was able to achieve a lower Mean Absolute Error (MAE) of 37,821.72, a Mean Squared Error (MSE) of 295,565,168.46, and an R-squared (R²) value of 0.8151. These improvements make the model more reliable and increase its prediction accuracy, with the model being saved for deployment as 'Resale\_Price.pkl' .

However, after the model was trained, it was found that the saved file was too large to be uploaded to GitHub for deployment. To resolve this, the best parameters found through Randomized Search were used to retrain the model with slight adjustments. After the training, the model was compressed and saved again. This process resulted in a minute improvement in R-squared, increasing it from 0.8111 to 0.8112. This change indicates that the model performed slightly better, but the effect was minimal. The final model was successfully saved in a compressed format.



.

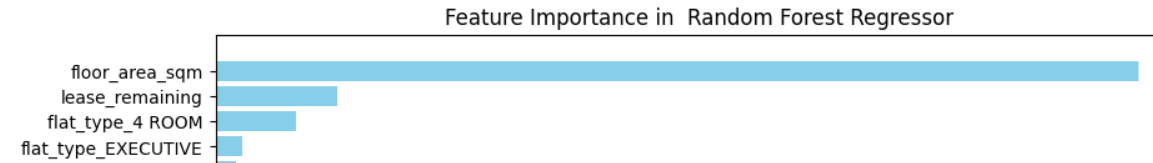
**Results And Analysis**

The Random Forest Regressor performed the best as compared to the other models such as Linear Regression, Decision Tree Regressor and Gradient Boosting Regressor . The model was trained using features such as floor area, region, and lease remaining, among others. After feature engineering, handling outliers, and tuning the hyperparameters using RandomizedSearchCV, the model achieved a final R-squared value of 0.8152, indicating that it explains 81.52% of the variance in the data.

The Mean Absolute Error (MAE) and Mean Squared Error (MSE) for the optimized Random Forest model were 37974.31 and 2948424955.89, respectively. These metrics show that the model performs well in predicting resale prices with relatively low error.

A notable improvement was observed after fine-tuning the model's parameters, such as reducing the complexity by limiting the number of trees and adjusting the tree depth. The minor adjustment of changing the lease age to lease remaining had only a marginal improvement in the model's accuracy, further supporting that the Random Forest model's initial setup was already well-optimized.

Additionally, a feature importance analysis in the Random Forest Regressor revealed that the **floor area** and **lease remaining** were the most influential features in predicting resale prices. This shows that these two features play a significant role in determining the price of flats, which aligns with the domain knowledge of the housing market.



**Conclusion**

In conclusion, the Random Forest Regressor emerged as the most effective model for predicting resale prices of flats. After addressing issues such as outliers and improving the model with feature engineering and hyperparameter tuning, the model achieved a strong R-squared value of 0.8151, indicating its ability to explain a significant portion of the variance in the dataset. The improvements made, including the transformation of the lease age feature and optimization through RandomizedSearchCV, contributed to the model's enhanced accuracy.

Despite a marginal change in performance with the adjusted lease age, the Random Forest model remains the most reliable and robust option for this task. Future work can focus on exploring additional feature selection techniques or further refining the model to potentially achieve even better performance.

Overall, the model provides a solid foundation for accurately predicting resale prices and could be applied in practical real-world scenarios.