Housing Price Forecast findings

### I/ Introduction

This is the documentation of the findings gathered from the evaluation of three machine learning models (Linear Regression, Random Forest, and Gradient Boosting) and their accuracy (calculated by the mean absolue error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Square Error (RMSE)\_ trained a dataset of housing information, containing only numerical variables (being postcode, room count, small room count, property area, balcony area, ad view count, active days and number of rooms) and forecasting the housing price of a test dataset.

It is important to note that during Data Preparation, due to the lack of some data in certain areas, some data was fabricated using the mean or median of given datasets such that the distribution curve stays relatively constant.

### II/ Results

After running the Linear Regression, Random Forest, and Gradient Boosting models three times, the results of the models is documented as such:

|  |  | Test 1 |  |  | Test 2 |  |  | Test 3 |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | MAE | MAPE | RMSE | MAE | MAPE | RMSE | MAE | MAPE | RMSE |
| Linear Regression | 5.35 | 29.39 | 7.40 | 5.35 | 29.34 | 7.40 | 5.36 | 29.39 | 7.41 |
| Random forest | 3.91 | 21.53 | 5.95 | 3.89 | 21.36 | 5.92 | 3.89 | 21.36 | 5.93 |
| Gradient Boosting | 4.42 | 24.34 | 6.37 | 4.39 | 24.15 | 6.33 | 4.39 | 24.10 | 6.33 |

Table 1.0: The results after running each model 3 times

|  | MAE | MAPE | RMSE |
| --- | --- | --- | --- |
| Linear Regression | 5.35 | 29.39 | 7.40 |
| Random forest | 3.89 | 21.36 | 5.92 |
| Gradient Boosting | 4.39 | 24.1 | 6.33 |

Table 1.1: The average of the erros after running each model 3 times

Key**:**

**MAE**: Represents the average absolute difference between predicted and actual values. Lower values indicate better performance.

**MAPE**: Indicates the average percentage error; it's useful to understand the accuracy of predictions in relative terms. A lower MAPE suggests better performance.

**RMSE**: Penalizes larger errors more than MAE. It's useful for identifying outliers; lower RMSE values indicate better accuracy.

### III/ Observations

Random Forest consistently outperformed both Linear Regression and Gradient Boosting in all three runs, with the lowest MAE, MAPE, and RMSE. This suggests that the Random Forest model captures the underlying patterns in the data effectively.

Gradient Boosting had performance metrics between those of Linear Regression and Random Forest, indicating that it provides a good trade-off but does not match the performance of Random Forest.

Gradient Boosting had performance metrics between those of Linear Regression and Random Forest, indicating that it provides a good trade-off but does not match the performance of Random Forest.

The values for each metric across the three runs were relatively stable, especially for the Random Forest model, indicating consistent performance. The small variations suggest that the model is robust to changes in the test set.

### IV/ Conclusions

Based on the evaluation metrics, the Random Forest Regressor would be the best choice for predicting housing prices in the dataset compared to Linear Regression and Gradient Boosting. It demonstrates better accuracy and a lower error rate compared to the other models. However, it is important the average accuracy of these are still quite low and unoptimized being only about 70 - 80% making it unreliable for accurate forecasts.