Reflection Paper: House Price Prediction

What I Implemented

For this project, I built a **house price prediction system** using two machine learning models: **Linear Regression (LR)** and **Random Forest (RF)**. The dataset used was already cleaned and included engineered features that helped the models learn more effectively.

The first step was splitting the dataset into **features (X)** and the **target variable (y)**. The target was the house price, while the features included other property-related attributes. After splitting the data into training and testing sets (80% for training and 20% for testing), I trained both models.

- **Linear Regression**: A simple model that assumes a straight-line relationship between the features and house prices. It is easy to understand and gives a good baseline.
- **Random Forest**: An advanced model that combines many decision trees to make predictions. It is better at handling complex and non-linear relationships.

After training, I tested both models using evaluation metrics such as **R^2, MAE, MSE, and RMSE**. These metrics allowed me to measure accuracy and compare the two approaches.

Comparison of Models

When I compared the models, I looked at both the overall metrics and the **sanity check with a single row** of data.

For the sanity check, the predictions were different. Linear Regression often gave results that were noticeably off from the actual price, sometimes too high or too low. On the other hand, Random Forest's prediction was usually much closer to the true value. This difference shows that Random Forest can capture the hidden relationships in the data that Linear Regression cannot.

In real-world housing markets, many factors interact in non-linear ways (for example, location and house size might interact differently depending on the neighborhood). Random Forest is able to account for these interactions, while Linear Regression cannot.

From these tests, I found that **Random Forest provided more realistic results** and would be more reliable for real-world predictions.

Understanding Random Forest

Random Forest is a type of **ensemble model**, which means it combines multiple models to make a stronger overall prediction. It works by creating many **decision trees**, each trained on a random subset of the data.

Each decision tree gives its own prediction, and then the Random Forest takes the **average of all the predictions** (for regression tasks like this). By averaging, the model reduces errors and avoids overfitting, which is a common problem in single decision trees.

In simple words, Random Forest is like asking the opinion of many experts instead of relying on just one. Each expert might make small mistakes, but when their answers are combined, the overall result is usually much more accurate.

Metrics Discussion

To evaluate the models, I used four metrics:

- **R^2 (coefficient of determination):** Measures how much of the variation in prices the model can explain. Higher values mean better performance.
- **MAE (Mean Absolute Error):** Shows the average size of the errors in dollars. Lower is better.
- **MSE (Mean Squared Error):** Squares the errors before averaging, which makes larger mistakes more noticeable. Lower is better.
- **RMSE (Root Mean Squared Error):** The square root of MSE, in the same units as price. Lower is better.

From the results, Random Forest had a **higher R^2** and **lower MAE and RMSE** than Linear Regression. This means Random Forest explained the data better and made more accurate predictions. Linear Regression, while still useful, could not capture the complexity of the housing data.

This shows the trade-off between the two models:

- **Linear Regression**: Simple, easy to understand, but less accurate.
- **Random Forest**: More complex, less interpretable, but much stronger at prediction.

Findings

Based on my results, I believe **Random Forest is the better choice for house price prediction**. It consistently produced predictions closer to the actual values and performed better across all evaluation metrics. This makes sense because housing prices depend on many factors that interact in non-linear ways, and Random Forest is designed to handle that complexity.

However, Linear Regression is still valuable in some cases. It is much easier to explain to people who want to understand how the model works, since it shows the direct relationship between each feature and the price. If interpretability is the most important goal, Linear Regression may still be useful.

In conclusion, I prefer **Random Forest** for practical house price prediction tasks because accuracy is usually more important than simplicity. It captures the real-world complexity of housing markets, reduces errors, and produces predictions that are closer to reality. Linear Regression remains a good baseline, but Random Forest provides the reliability needed in real-world applications.