House Price Prediction Using Linear Regression and Random Forest

1. What I Implemented

In this assignment, I developed a machine learning approach to predict house prices using two different regression models: Linear Regression and Random Forest Regressor.

I began by loading a cleaned housing dataset and preparing the data by separating the target variable (Price) from the input features. All columns except Price and LogPrice were used as predictors. The dataset was then split into training data (80%) and testing data (20%) using a fixed random state to ensure reproducibility.

After preprocessing, I trained a Linear Regression model using the training dataset. I then trained a Random Forest Regressor with 100 decision trees. To evaluate both models, I calculated R^2, MAE, MSE, and RMSE on the test dataset. I also performed three sanity checks by selecting individual rows from the test set and comparing the predicted prices to the actual prices.

2. Comparison of Models

During the three sanity checks, I observed clear differences between the predictions of the two models. Linear Regression sometimes produced predictions that were farther from the actual values, particularly in cases where the relationships between features and price were complex.

Random Forest predictions were generally closer to the real prices. This suggests that the model was able to capture patterns that Linear Regression could not represent effectively. Linear Regression assumes a linear relationship between variables, which can be unrealistic in housing markets where factors such as size, location, and amenities interact in complex ways.

3. Understanding Random Forest

Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve prediction accuracy and stability.

Instead of training a single decision tree, Random Forest builds many trees using different random subsets of the training data and features. Each tree makes its own prediction, and the final result is obtained by averaging all predictions. This process reduces overfitting and makes the model more robust.

In simple terms, Random Forest works by combining many weak models to create a stronger overall model.

4. Metrics Discussion

To compare model performance, I examined R^2, MAE, and RMSE.

R^2 measures how well the model explains the variation in house prices, while MAE and RMSE measure prediction error. Lower MAE and RMSE indicate more accurate predictions.

In my results, Random Forest achieved a higher R^2 and lower error values than Linear Regression. This indicates that Random Forest captured the structure of the data more effectively.

However, this does not mean Linear Regression is useless. Linear Regression is faster, easier to interpret, and useful as a baseline model. Random Forest, while more accurate, requires more computation and is harder to interpret because it is made up of many trees rather than a single equation.

5. My Findings

Based on the results, Random Forest appears to be the more suitable model for house price prediction. Housing prices are influenced by many interacting variables, and these relationships are often non-linear. Random Forest handles these complexities better and produces predictions that are closer to real values.

That said, Linear Regression still has value. It provides a simple and interpretable model and is useful for understanding general trends in the data. In practice, it can be helpful to start with Linear Regression as a baseline and then use more advanced models like Random Forest when higher accuracy is required.