House Price Prediction: Summary Report

Abhinav Singh

March 29, 2025

1 Introduction

The goal of this analysis is to predict house prices using a dataset containing various features such as the size of the house, number of bedrooms, age of the house, and its location. Two different machine learning models, namely Linear Regression and Decision Tree Regressor, were trained and evaluated for performance.

2 Approach

The approach to model training involved the following steps:

1. **Data Preprocessing:**
   - The dataset was loaded, and categorical variables, such as location, were encoded using Label Encoding.
   - The features were standardized using StandardScaler to ensure that all features were on the same scale.
2. **Feature Selection:** The features selected for training included:
   - Size in square feet
   - Location (encoded)
   - Number of bedrooms, bathrooms, and garage
   - House age
   - Pool availability
   - Distance to the city center
3. **Data Split:** The data was split into training (80%) and testing (20%) sets using the `train_test_split` function.

3 Insight

We can see that the Price column is very strongly correlated to the 'Size_sqft' column and the 'Location' column. This suggests that Price could be linearly related to these attributes. Therefore, we will first apply LinearRegression and then compare it other models.

4 Model Training

Two regression models were trained and evaluated:

- **Linear Regression:** The first model used was Linear Regression. It was trained on the training dataset, and the performance was evaluated on the test set.
- **Decision Tree Regressor:** The second model was a Decision Tree Regressor. Hyperparameters such as maximum depth, minimum samples split, and minimum samples leaf were tuned to optimize performance.

5 Results

The evaluation of the models on the test set yielded the following results:

- **Linear Regression:**
  - R^2: 0.99
  - RMSE: 39797.35
  - MAE: 31989.14
- **Decision Tree Regressor:**
  - R^2: 0.96
  - RMSE: 70520.86
  - MAE: 56379.41

Linear regression outperformed the Decision Tree model in R^2, MAE and RMSE, indicating a better fit to the data.

6 Cross-validation for Linear Regression

To further validate the Linear Regression, cross-validation was performed using 5-fold cross-validation. The results were:

- **Cross-validation R^2:** Mean : 0.99
- **Cross-validation RMSE:** Mean = 41281.55
- **Cross-validation MAE:** Mean = 32566.67

Cross-validation showed consistent performance across different subsets of the data, further reinforcing the robustness of Linear Regression.

7 Conclusion

The Linear Regression model outperformed the Decision Tree Regressor model in predicting house prices. It provided better evaluation, and the cross-validation results confirmed its robustness. The feature importance analysis revealed that the size of the house is the most important factor in predicting the price, followed by the location and the number of bedrooms.