Report: Predicting Housing Prices Using Machine Learning Models

1. Introduction

This project aimed to predict housing prices using machine learning models. After preprocessing the data, the project evaluated three models: Linear Regression, Decision Tree Regressor, and Random Forest Regressor. The primary goal was to compare model performance using metrics like R-squared, Mean Absolute Error (MAE), and Mean Squared Error (MSE).

2. Data Preprocessing

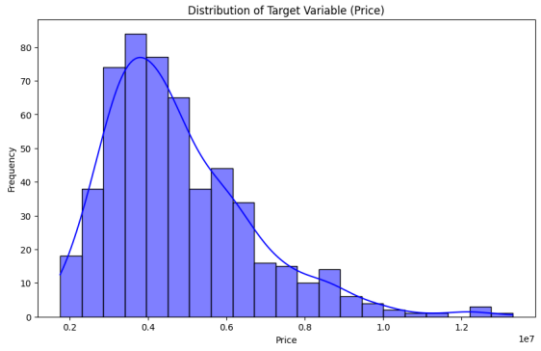
The dataset underwent several preprocessing steps to ensure clean data for modeling:

- Duplicate Rows: Duplicates were removed to avoid bias.
- Missing Data: Missing values in numerical columns were imputed with the median, and categorical columns with the most frequent value.
- Categorical Encoding: "Yes"/"No" columns were mapped to binary values (1 and 0), and other categorical features were encoded numerically.
- Outlier Removal: Outliers were identified using the IQR method and removed to prevent their distortion on model performance.

3. Exploratory Data Analysis (EDA)

Several visualizations were performed:

- Price Distribution: A histogram showed the skewed distribution of housing prices.



- Correlation Heatmap: Identified strong correlations between certain features
- Outliers: Boxplots revealed extreme price values, which were removed based on the IQR method.

4. Model Selection and Evaluation

Three models were trained and evaluated:

```
Linear Regression Evaluation:
R-squared: 0.6489
Mean Absolute Error: 818791.1209
Mean Squared Error: 1215825380211.7498
--------------------------------------------------
Decision Tree Evaluation:
R-squared: 0.4495
Mean Absolute Error: 1076387.3585
Mean Squared Error: 1906306775007.5471
--------------------------------------------------
Random Forest Evaluation:
R-squared: 0.6065
Mean Absolute Error: 910605.2984
Mean Squared Error: 1362403108336.6069
--------------------------------------------------
Cross-validation MSE for LinearRegression: 856980823028.1340
Cross-validation MSE for DecisionTreeRegressor: 1165605677921.3306
Cross-validation MSE for RandomForestRegressor: 937731939248.4008
```

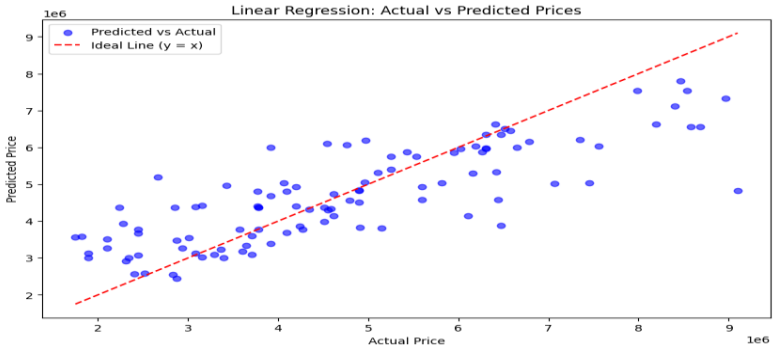
- Linear Regression: Performed best with an R-squared of 0.6489, lowest MSE and a Mean Absolute Error (MAE) of 818,791.

5. Hyperparameter Tuning

Hyperparameter optimization for Decision Tree and Random Forest models was done using GridSearchCV:

- Best parameters were found for both models, but their performance still lagged behind Linear Regression.

Prediction Results



6. Areas of Improvement and Difficulties

Areas of Improvement:

- Model Complexity: More advanced models like XGBoost could yield better results.
- Outlier Removal: Some extreme price points were difficult to categorize as outliers or genuine cases.
- Finding a new dataset: It was difficult to search for a new dataset with the same columns (features), had to device one on my own, it will be best if the new set of data was provided.

7. Conclusion

The project successfully demonstrated the power of machine learning in predicting housing prices. While Linear Regression performed the best, areas for improvement include more advanced models, feature engineering, and tuning techniques. Overcoming challenges such as data quality and Outlier Removal are crucial for improving prediction accuracy in future iterations.