Housing Price Prediction Model

Housing Price Prediction Project Report

1. Introduction

This project aims to predict housing prices using a machine learning model. The dataset used is Boston Housing Dataset, which contains various features that might influence the housing prices in a particular area. This report details the process of data preprocessing, model selection, training, evaluation, and the interpretation of results.

2. Data Exploration and Preprocessing

2.1 Data Overview

The dataset consists of several features including:

- CRIM: Per capita crime rate by town
- ZN: Proportion of residential land zoned for lots over 25,000 sq. ft.
- INDUS: Proportion of non-retail business acres per town
- CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- NOX: Nitric oxides concentration (parts per 10 million)
- RM: Average number of rooms per dwelling
- AGE: Proportion of owner-occupied units built prior to 1940
- DIS: Weighted distances to five Boston employment centers
- RAD: Index of accessibility to radial highways
- TAX: Full-value property tax rate per $10,000
- PTRATIO: Pupil-teacher ratio by town
- B: 1000(Bk - 0.63)^2 where Bk is the proportion of black people by town
- LSTAT: Percentage of lower status of the population
- MEDV: Median value of owner-occupied homes in $1000s (target variable)

2.2 Data Cleaning and Preparation

- Handled missing values:
  Missing values are handled using the `SimpleImputer` with a mean strategy.
- Outliers Removal
  Outliers are removed based on the Interquartile Range (IQR) method.
- Feature Engineering
  New features are created by squaring `RM` and `LSTAT`.

- Feature Selection
  Highly correlated features (correlation > 0.95) are dropped to avoid multicollinearity.
- Feature Scaling
  Features are scaled using `StandardScaler`.
- Splitting the dataset into training and testing sets (80/20 split).

3. Model Training

3.1 Model Selection

We chose the `Random Forest Regressor` for its ability to handle complex interactions between features and robustness to overfitting.
Grid search with cross-validation is used to find the best hyperparameters.

3.2 Training the Model

The model was trained on the training dataset using cross-validation to ensure generalizability and prevent overfitting.

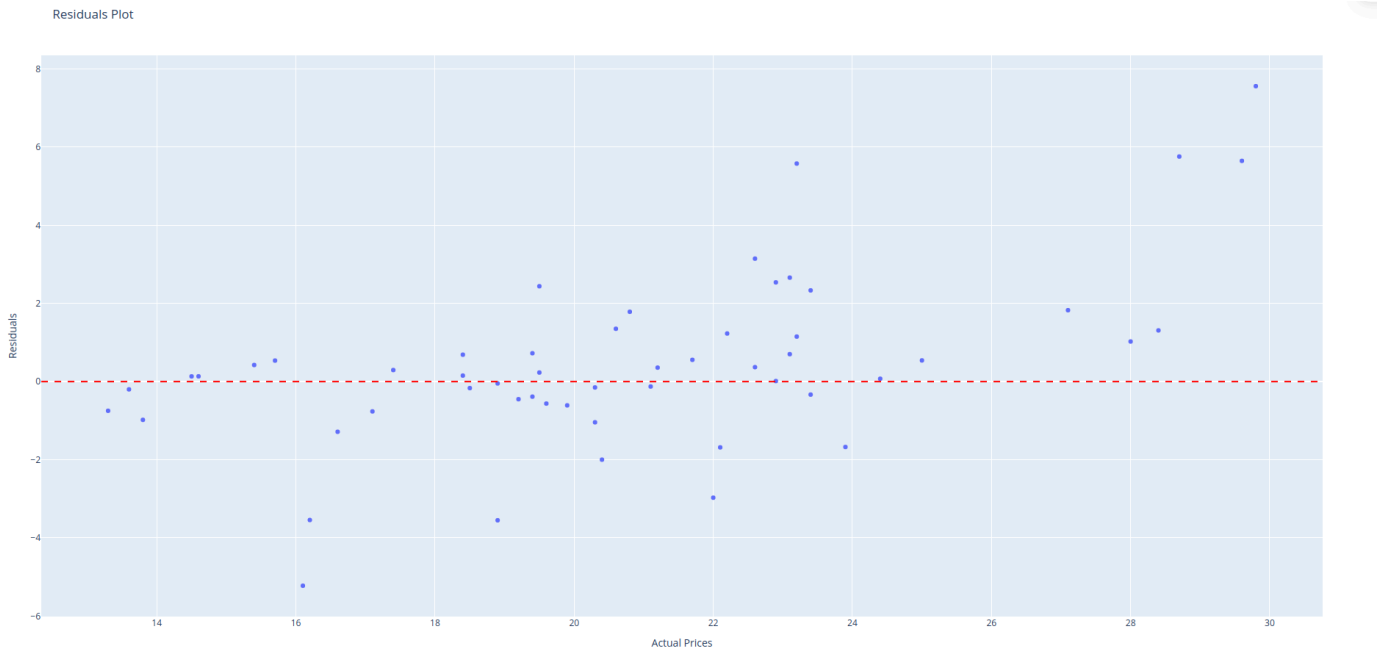
4. Model Evaluation

4.1 Cross-Validation and Metrics

- **Cross-Validated Mean Absolute Error (MAE):** 1.8660918369743946
- **Mean Absolute Error on Test Set:** 1.5143129276844538
- **Mean Squared Error on Test Set:** 5.2250967221739995

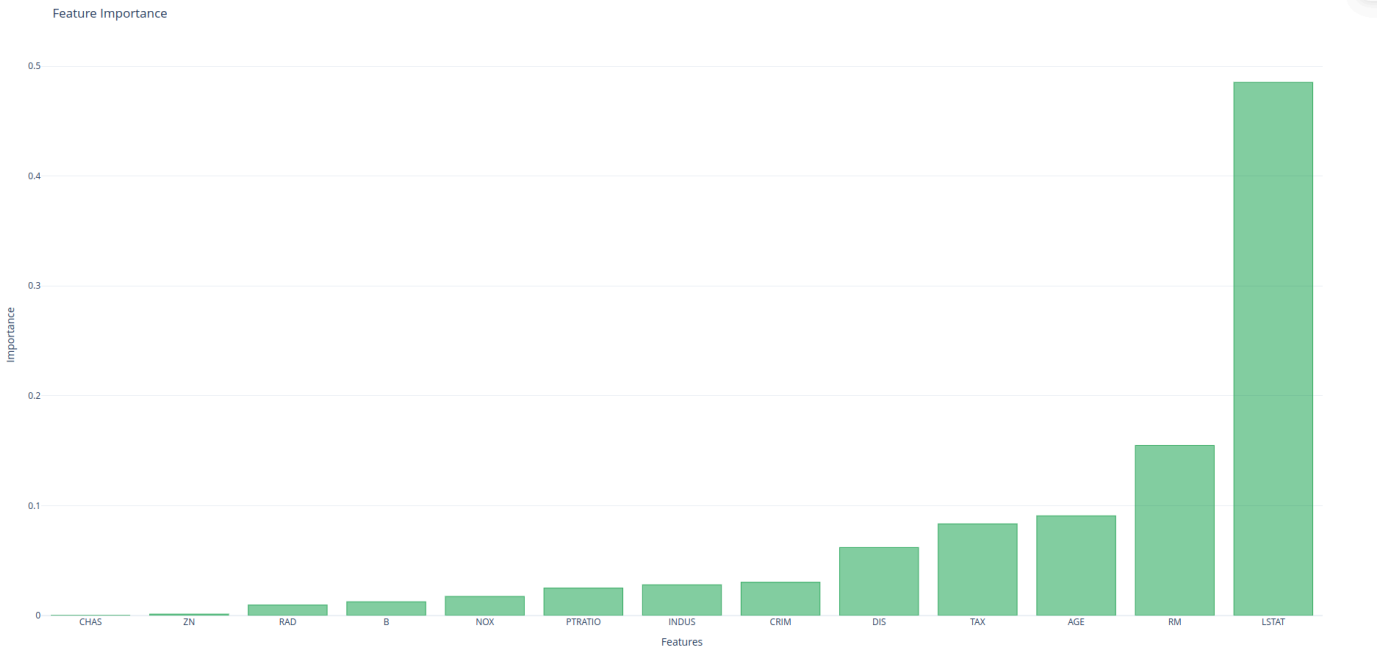
5. Results and Analysis

5.1 Residuals Analysis



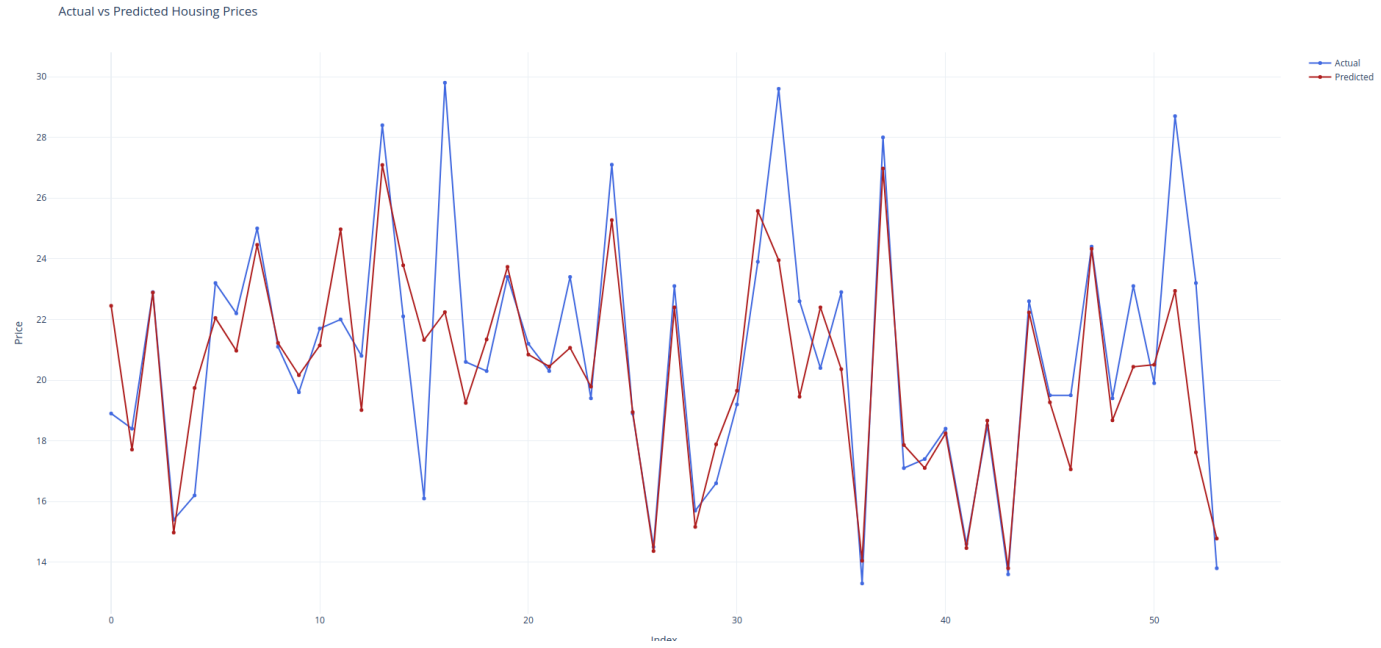
The residuals plot indicates the difference between the actual and predicted prices. The residuals are fairly randomly distributed around zero, suggesting a good fit.

5.2 Feature Importance



The feature importance plot shows that 'LSTAT' and 'RM' are the most significant predictors of housing prices. This insight aligns with domain knowledge, as the lower status of the population and the number of rooms are crucial factors in housing prices.

5.3 Actual vs Predicted Prices



The actual vs. predicted prices plot demonstrates the model's performance. The predictions closely follow the actual values, indicating a high level of accuracy.

6. Conclusion

The **Random Forest Regressor** effectively predicts housing prices with a low mean absolute error. The analysis of feature importance provides valuable insights into the factors influencing housing prices. Further improvements could be made by tuning hyperparameters and exploring other machine learning models.

7. References

- Dataset source: Boston Housing Dataset
- plotly : Link
- Random Forest Regressor : Link

Github Link : Here