Name: Divij Mahajan
Class: 2K22/CO/171

Project Report: Predicting House Prices with the Boston Housing Dataset

1. Introduction

1.1. Objective

This project's objective is to develop a predictive model that can be used to estimate Boston real estate prices based on a range of criteria, including location, size, number of bedrooms, and other pertinent characteristics. The Boston Housing dataset is utilised by the model to attain precise price forecasts.

1.2. Dataset Overview

The Boston Housing dataset includes various features related to housing prices, such as:

- **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
- **LSTAT**: Percentage of lower status of the population
- **MEDV**: Median value of owner-occupied homes in $1000s (target variable)

2. Data Exploration

2.1. Statistics

To comprehend the dataset's structure and feature distribution, a preliminary analysis was done. The dataset's overall distribution, dispersion, and central tendency were all revealed by summary statistics.

2.2. Data Visualization

Visualising relationships between features and the target variable (MEDV) was a key component of exploratory data analysis, or EDA. Important visuals comprised:

- Use scatter plots to investigate the connection between certain attributes and home values.
- A correlation heatmap to show the direction and strength of features' relationships.

3. Methodology

3.1. Data Preprocessing

- **Handling Missing Values**: No missing values were found in the dataset.
- **Data Splitting**: The dataset was divided into training and testing sets to evaluate model performance.

3.2. Model Selection

Linear Regression was chosen for this project as it is well-suited for predicting continuous outcomes and is a standard choice for regression problems.

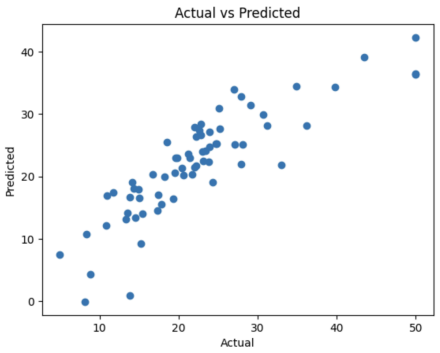
4. Results

4.1. Model Evaluation

The performance of the linear regression model was assessed using several metrics:

- **Mean Absolute Error (MAE)**: Indicates the average magnitude of prediction errors.
- **Root Mean Squared Error (RMSE)**: Measures the average squared differences between predicted and actual values.
- **R-squared (R^2)**: Reflects the proportion of variance in the target variable that is predictable from the features.

```
Model performance
RMSE :4.838564025266022
MAE: 3.659485816027954
r2: 0.7398653051224335
```



5. Conclusion

Based on the given features, the house price prediction model efficiently estimated values. The analysis indicated areas for further improvement and highlighted key factors influencing housing prices. Subsequent endeavours may encompass increasingly sophisticated methodologies to enhance the precision of forecasts and integrate a wider assortment of data sets.

References

- Boston Housing Dataset - Kaggle
- Scikit-learn Linear Models Documentation
- Pandas Documentation