Boston Housing Dataset Analysis and Regression Modeling

1. Introduction

This report summarizes the exploratory data analysis and regression modeling performed on the Boston Housing dataset. The goal is to predict house prices (`medv`) based on various housing and environmental features.

2. Dataset Overview

The dataset consists of 506 instances with 14 attributes:

- `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 centres
- `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 blacks by town
- `lstat`: % lower status of the population
- `medv`: Median value of owner-occupied homes in $1000's (target variable)

3. Data Exploration

A quick statistical summary of the dataset:

Feature	Count	Mean	Std Dev	Range
`crim`	506	3.61	8.60	0.0063 – 88.98
`zn`	506	11.36	23.32	0.0 – 100.0
`indus`	506	11.14	6.86	0.46 – 27.74
`chas`	506	0.07	0.25	0 / 1
`nox`	506	0.55	0.12	0.39 – 0.87
`rm`	501	6.28	0.71	3.56 – 8.78
`age`	506	68.57	28.15	2.9 – 100
`dis`	506	3.80	2.11	1.13 – 12.13
`rad`	506	9.55	8.71	1 – 24
`tax`	506	408.24	168.54	187 – 711
`ptratio`	506	18.46	2.16	12.6 – 22
`b`	506	356.67	91.29	0.32 – 396.9
`lstat`	506	12.65	7.14	1.73 – 37.97
`medv`	506	22.53	9.20	5.0 – 50.0

No missing values were found, except for 5 missing values in `rm`, which were dropped prior to modeling.

4. Data Preparation

The features (`X`) were taken as all columns except `medv`, the target variable (`y`) being the median home value. The dataset was split into training and testing sets with an 80-20 ratio, random state fixed for reproducibility.

5. Modeling

Linear Regression

A linear regression model was trained to predict housing prices.

Performance metrics on test set:

- Root Mean Squared Error (RMSE): 4.55
- R^2 Score: 0.72

Random Forest Regression

A Random Forest regressor with 100 trees was also trained, improving performance.

Performance metric on test set:

- R^2 Score: 0.88

6. Conclusion

The Random Forest regression model showed substantial improvement over linear regression for predicting Boston housing prices. With an R^2 of 0.88, it explains a significant proportion of the variance in the dataset, making it a suitable choice for this regression problem.

7. Appendix

The trained linear regression model was saved as `boston_house_price_model.pkl` for future use.