Boston Housing Price Prediction

Machine Learning Project Report

CONTENT

1. INTRODUCTION

The real estate market is one of the most dynamic and influential sectors of the economy.
Accurately predicting housing prices can help:

Buyers make informed purchasing decisions.

Sellers set competitive prices.

Investors & Developers evaluate market opportunities.

This project focuses on predicting Boston housing prices using machine learning techniques. The aim is to build a model that learns from historical housing data and predicts the median value of homes based on socio-economic and geographical features.
The workflow covers data exploration, preprocessing, model development, evaluation, and deployment as an interactive web application.

2.DATASET EXPLANATION

The dataset used in this project is the Boston Housing dataset, sourced from Kaggle.

It contains information collected by the U.S. Census Service about housing in the Boston, Massachusetts area.

Rows: 506

Columns: 13 numerical features + 1 target variable (MEDV)

COLUMN DESCRIPTIONS:

1. CRIM – Per capita crime rate by town.

2. ZN – Proportion of residential land zoned for lots over 25,000 sq.ft.

3. INDUS – Proportion of non-retail business acres per town.

4. CHAS – Charles River dummy variable (1 if tract bounds river; 0 otherwise).

5. NOX – Nitric oxide concentration (parts per 10 million).

6. RM – Average number of rooms per dwelling.

7. AGE – Proportion of owner-occupied units built prior to 1940.

8. DIS – Weighted distances to five Boston employment centers.

9. RAD – Index of accessibility to radial highways.

10. TAX – Full-value property tax rate per $10,000.

11. PTRATIO – Pupil–teacher ratio by town.

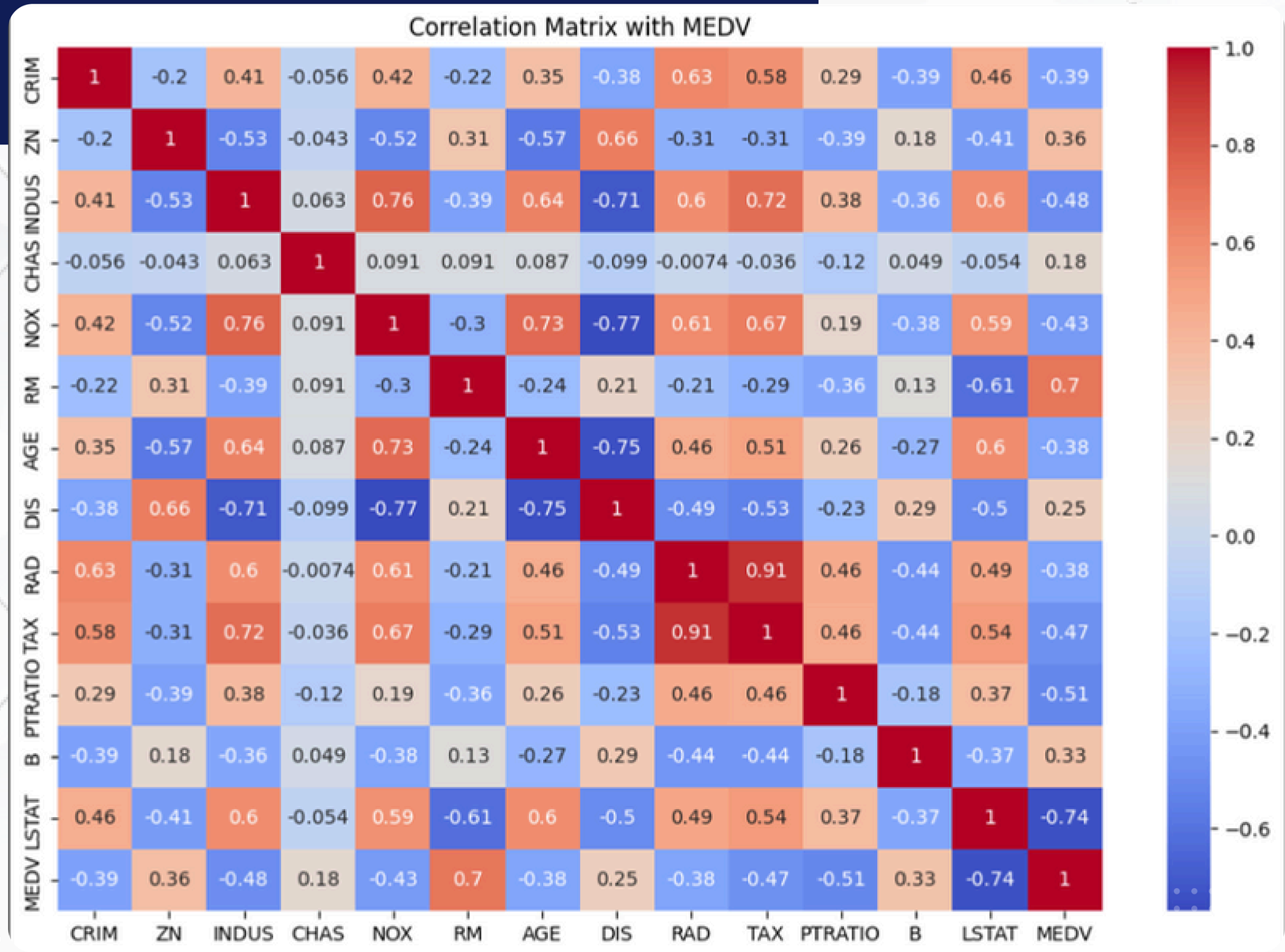
12. B – $1000(Bk – 0.63)^2$, where Bk is the proportion of Black residents by town.

13. LSTAT – Percentage of lower-status population.

14. MEDV (Target) – Median value of owner-occupied homes in $1000s.

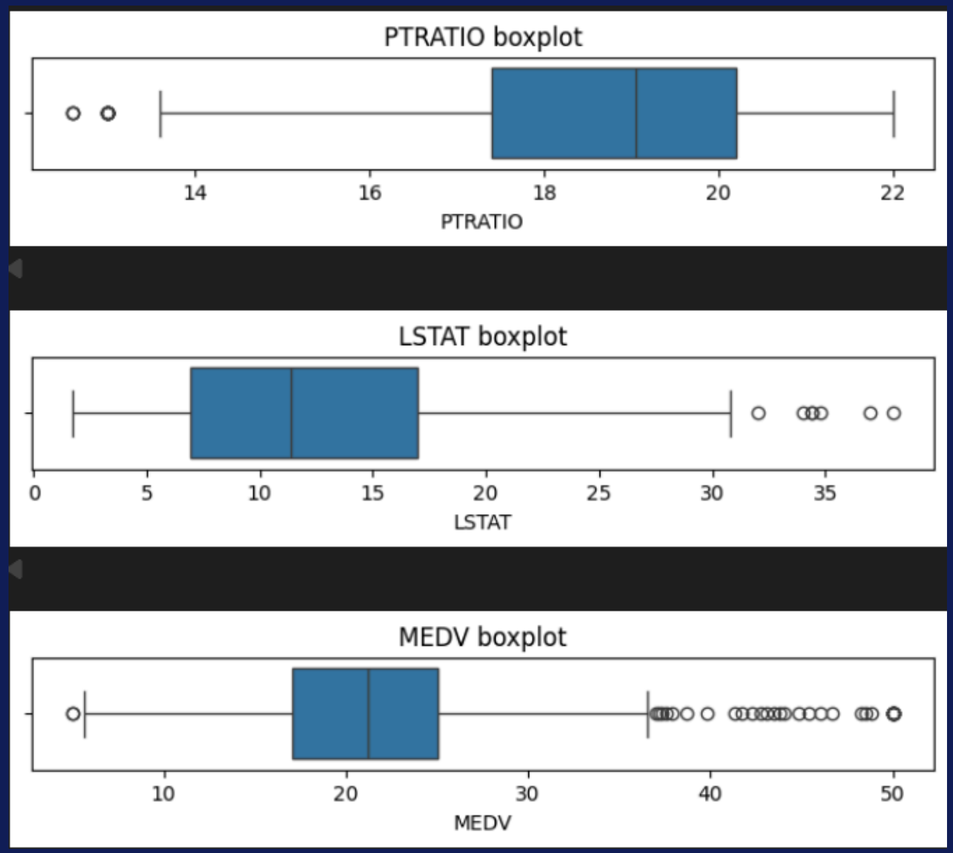
3.EXPLORATORY DATA ANALYSIS (EDA)

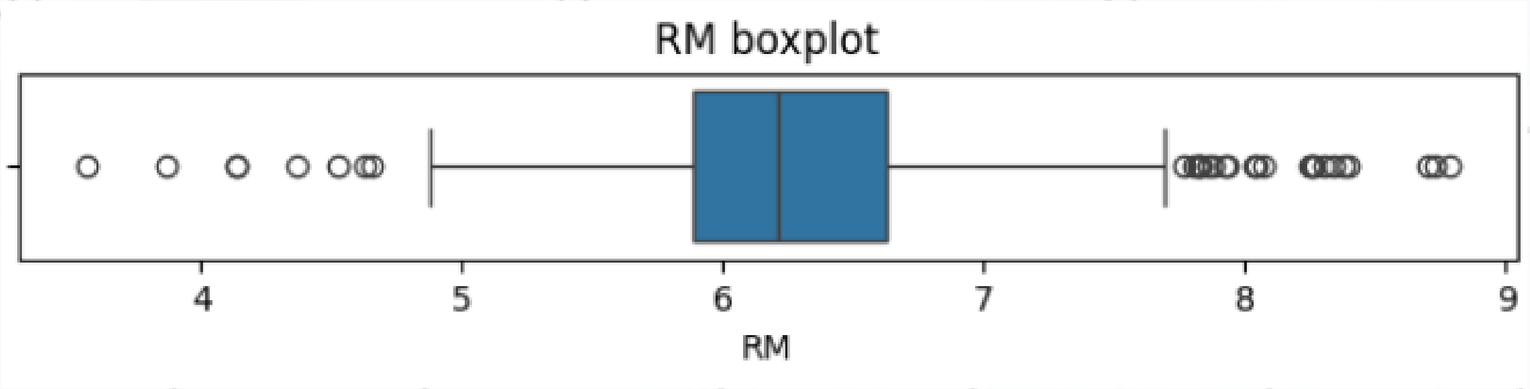
- Displayed the dataset to understand its structure and the data types.
- Checked unique values in each column to ensure data consistency.
- Created a heatmap to visualize correlations between features and the target variable (MEDV).
- Identified that RM has the strongest positive correlation with MEDV, while LSTAT has the strongest negative correlation & PTRATIO has the second strongest negative correlation.
- Noted that CHAS and B have very weak correlations with MEDV.
- Used boxplots and other plots to detect outliers and understand feature distributions



4. DATA PREPROCESSING

- Verified that there were no missing values or duplicate rows in the dataset.
- Removed features with weak correlation to the target variable (CHAS and B).
- Detected significant outliers in numerical features and removed them to improve data quality.
- Split the data into features (X) and target (y - MEDV).
- Selected top features based on correlation: RM, PTRATIO, LSTAT.
- Applied Min-Max Scaling to normalize feature values for better model performance.
- No encoding was necessary since all features are numerical.





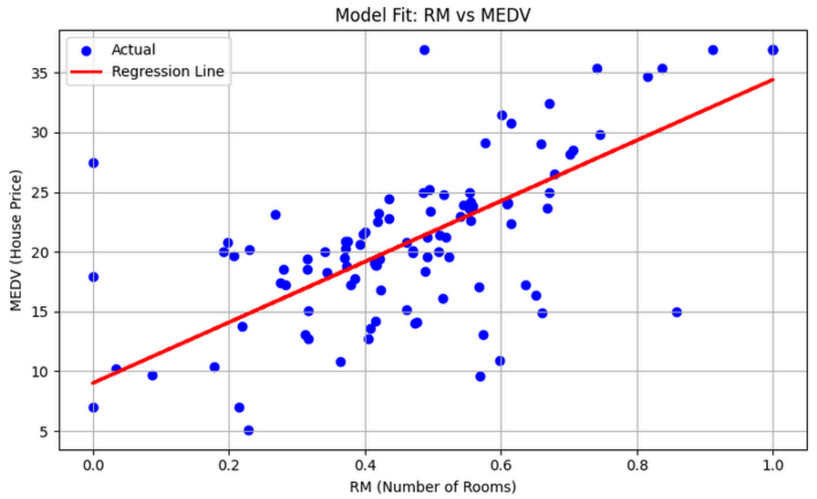
5. FEATURE PREPARATION

- Split data into X (features) and y (target: MEDV).

6. MODEL DEVELOPMENT & EVALUATION

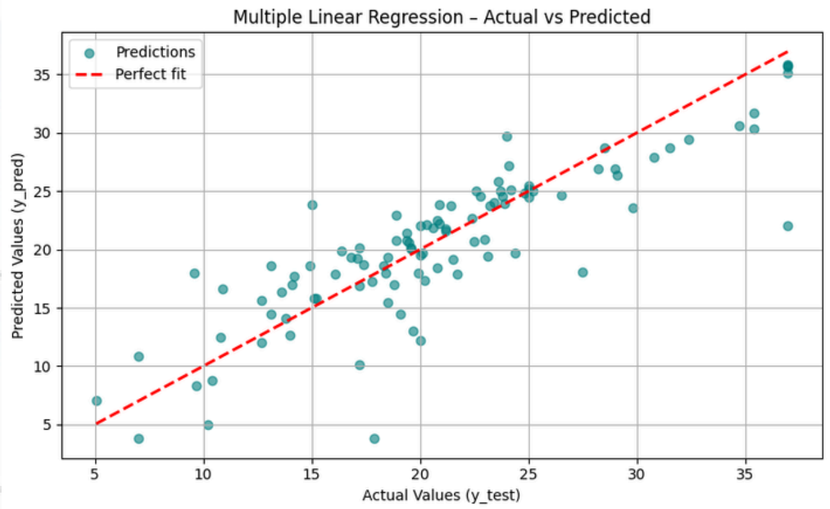
a) Simple Linear Regression

- Used RM as the only feature.
- MSE: 28.92
- Plotted regression line vs actual values.



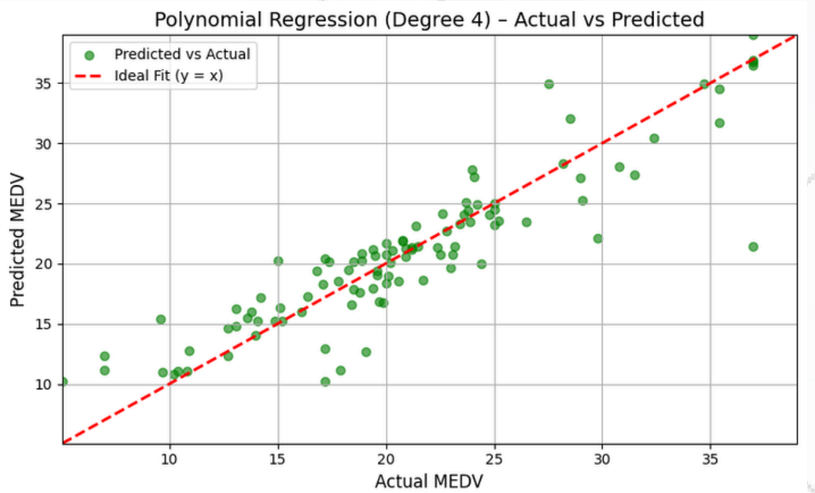
b) Multiple Linear Regression

- Used RM, PTRATIO, LSTAT
- MSE: 13.82
- Plotted predicted vs actual values.



c) Polynomial Regression (Degree = 4)

- Used RM, PTRATIO, LSTAT.
- MSE: 9.29 (Best performance).
- Plotted predicted vs actual values.

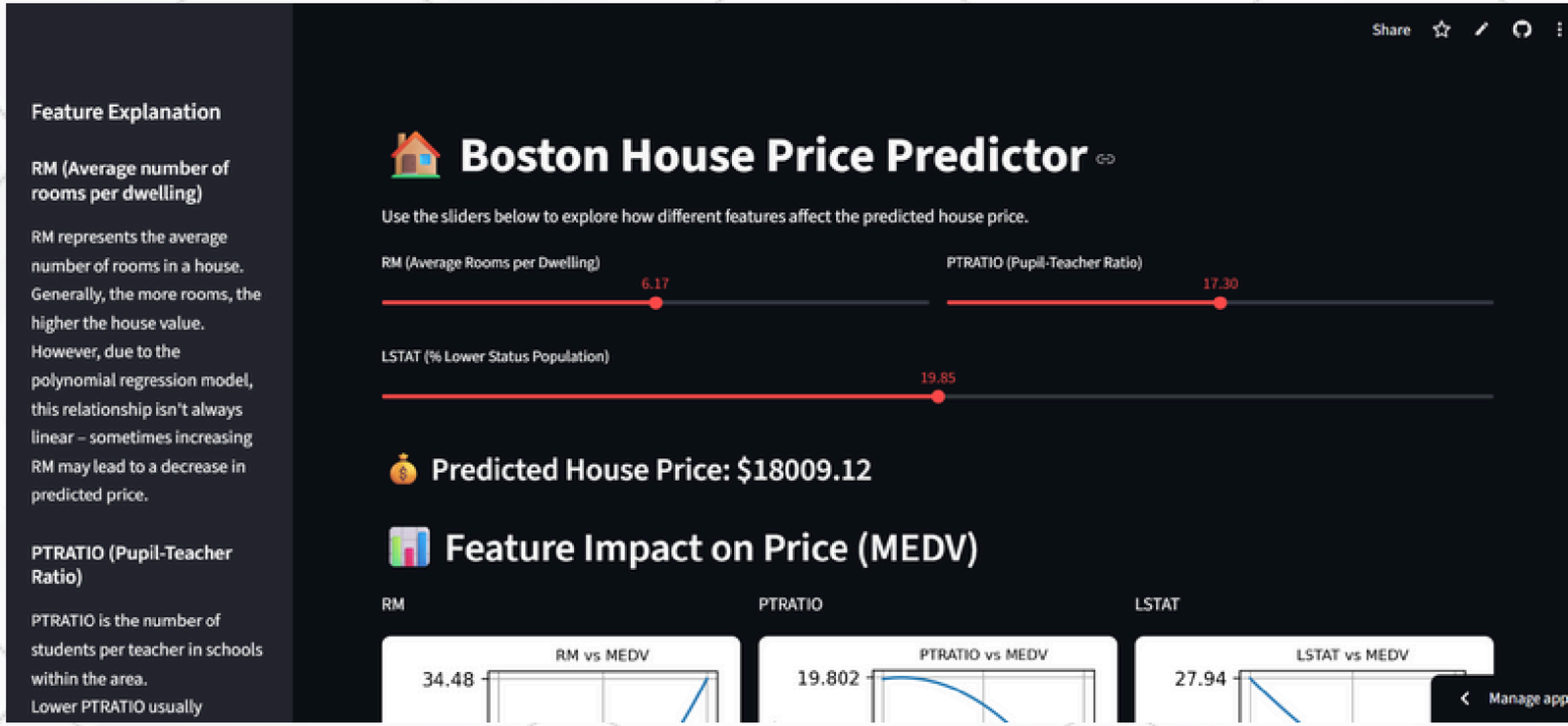


7. MODEL SELECTION & DEPLOYMENT

- Chose Polynomial Regression as the final model due to lowest MSE.
- Created a prediction function using the selected 3 features.
- Saved the model using Joblib for deployment in the web application.

8. WEB APPLICATION

- Developed an interactive app using Streamlit.
- Allows users to input the 3 key features (RM, PTRATIO, LSTAT) and get price predictions instantly.
- Includes a sidebar explaining dataset columns and their relationship with the target variable.
- Displays plots showing the correlation between selected features and MEDV.



- https://bostonhousing-price-prediction.streamlit.app/

9. CONCLUSION & INSIGHTS

Removing irrelevant features (such as CHAS and B) and outliers significantly improved the model's accuracy and robustness.

Polynomial Regression demonstrated superior performance compared to both simple and multiple linear regression models, capturing non-linear relationships more effectively.

The three most influential features for predicting Boston housing prices were identified as RM (average number of rooms), PTRATIO (pupil-teacher ratio), and LSTAT (% lower status population).

The deployed web application offers an intuitive and user-friendly interface for real-time Boston housing price predictions, facilitating practical use and accessibility.