Linear Regression using Scikit-Learn on Boston Housing Dataset

Vaibhav Chourasia

October 2, 2025

Abstract

This report presents the implementation of linear regression using Scikit-Learn to predict housing prices on the Boston Housing dataset. The model is trained and tested using an 80/20 split, evaluated using Mean Squared Error (MSE) and R^2 score, and visualized with Actual vs Predicted and Residual plots.

1 Introduction

Linear regression is a fundamental supervised machine learning algorithm that models the relationship between a dependent variable (y) and independent variables (X) by fitting a linear equation:

$$y = w^T X + b \tag{1}$$

where w represents the coefficients and b is the intercept. This study uses Scikit-Learn's `LinearRegression` implementation to predict median house prices.

2 Dataset Description

The Boston Housing dataset contains 506 instances with 13 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
- 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: proportion of blacks by town
- LSTAT: % lower status of the population
- MEDV: Median value of owner-occupied homes (target)

3 Methodology

3.1 Train/Test Split

The dataset was split into training (80%) and testing (20%) sets to evaluate model performance on unseen data.

3.2 Linear Regression

Scikit-Learn's `LinearRegression` was used to fit the model on training data and make predictions on both training and test sets. Scikit-Learn calculates the coefficients (w) and intercept (b) using inbuilt analytical methods. This eliminates the need for manual gradient descent or iterative updates.

3.3 Residuals

Residuals are calculated as:

$$r_i = y_i - \hat{y}_i \tag{2}$$

where y_i is the actual price and $\hat{y}_i$ is the predicted price.

4 Results

4.1 Regression Equation

The regression equation obtained from Scikit-Learn is:

$$\begin{aligned}\text{Price} =& (-0.1131 \cdot \text{CRIM}) + (0.0301 \cdot \text{ZN}) + (0.0404 \cdot \text{INDUS}) + (2.7844 \cdot \text{CHAS}) \\ &+ (-17.2026 \cdot \text{NOX}) + (4.4388 \cdot \text{RM}) + (-0.0063 \cdot \text{AGE}) + (-1.4479 \cdot \text{DIS}) \\ &+ (0.2624 \cdot \text{RAD}) + (-0.0106 \cdot \text{TAX}) + (-0.9155 \cdot \text{PTRATIO}) + (0.0124 \cdot \text{B}) \\ &+ (-0.5086 \cdot \text{LSTAT}) + 30.2468\end{aligned} \tag{3}$$

4.2 Evaluation Metrics

- Training MSE: 21.641
- Training R^2: 0.751

- Testing MSE: 24.291
- Testing R^2: 0.669

4.3 Predictions Example

The first 10 predicted prices (test set) in $1,000 are :

```
[28.997, 36.026, 14.817, 25.032, 18.770, 23.254, 17.663, 14.341, 23.013, 20.632]
```

These predictions correspond to the test set, as shown in the scatter plots.

4.4 Plots

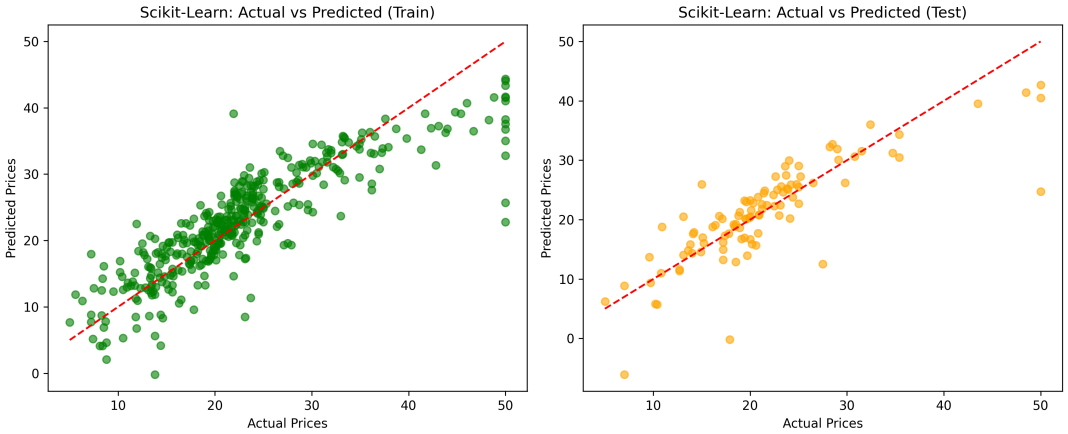

Figure 1: Actual vs Predicted Prices (Training and Test sets)

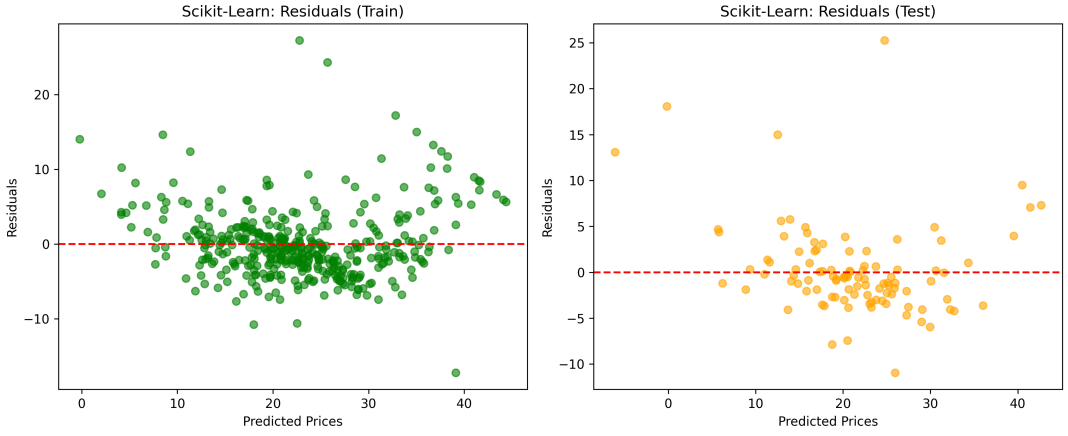

Figure 2: Residuals (Training and Test sets)

5 Discussion

The Actual vs Predicted plots show that the model predicts both training and test data reasonably well, with points distributed around the ideal $y = x$ line. Residual plots

indicate that the errors are randomly scattered around zero, confirming the assumptions of linear regression. Unlike the scratch implementation, this model uses analytical solutions via Scikit-Learn's inbuilt functions, eliminating the need for gradient descent.

6 Conclusion

Linear regression using Scikit-Learn successfully predicted median housing prices in the Boston dataset. The model achieved good performance on both training and test sets. The regression equation provides interpretable relationships between features and the target. Visualizations further confirm that the model is appropriate.