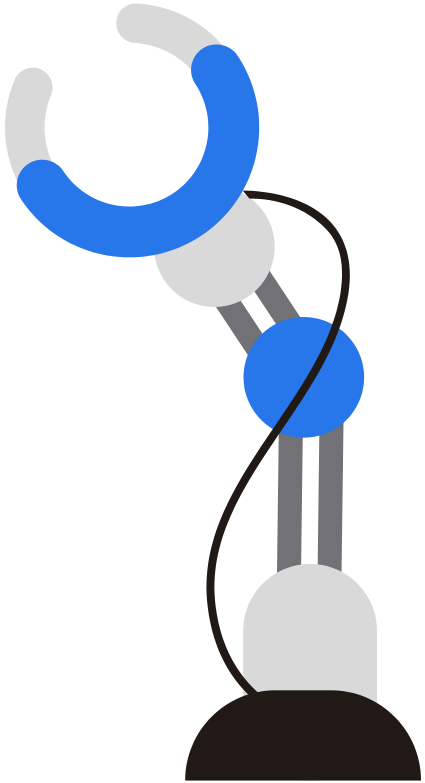
Machine Learning Models to Predict Life Expectancy

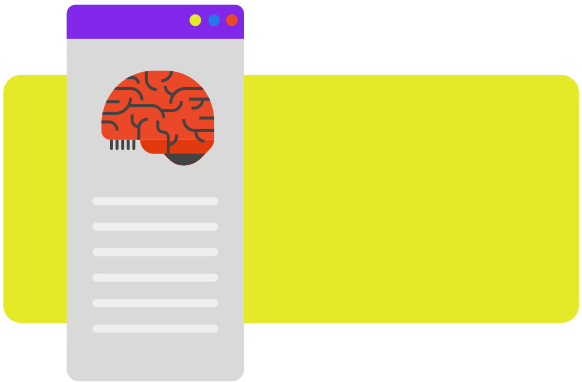
By Alex Cortes

Project Overview

This project explores methods to predict life expectancy using machine learning models. It aims to identify the best-performing approach by improving the R^2 score through feature engineering and model optimization.



Problem statement



To predict the life expectancy of people in a country based on various immunization factors and the country in which they are based so that necessary action can be taken to increase the life expectancy.

Dataset Information

The dataset “Life Expectancy (WHO)", publicly available on Kaggle, compiled from the Global Health Observatory (GHO) data repository under the World Health Organization (WHO) and supplemented with corresponding economic data from the United Nations.

It tracks life expectancy and related health factors for 193 countries between 2000 and 2015.

Key Steps

01 **Data Exploration**

Understand the dataset structure and variable distributions.

02 **Exploratory Data Analysis (EDA)**

Handle missing values using mean/mode imputation.
Perform one-hot and label encoding for categorical variables.

03 **Feature Engineering**

Feature scaling using normalization techniques.
Feature extraction: Identify the top features impacting life expectancy.

04 **Model Building**

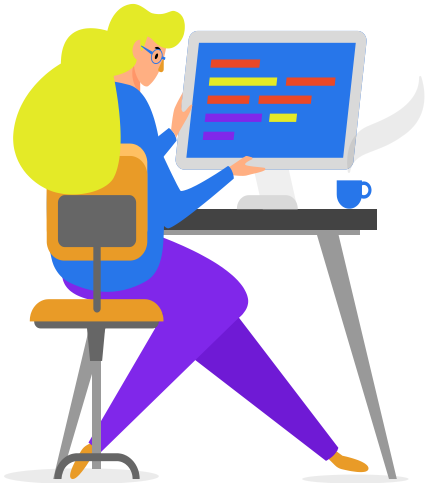
Train and evaluate regression models including:

- Linear Regression
- Lasso Regression
- Advanced feature selection using PCA

05 **Evaluation**

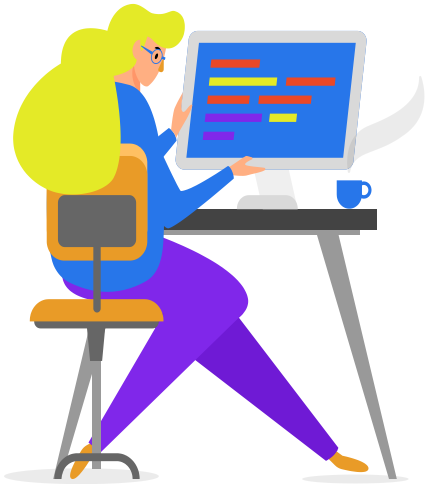
Evaluate models using metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and (R^2).

Methodology



1. **Data Preprocessing**
    a. Imputation of missing values
    b. Label /One Hot encoding.
    c. Format text.

Methodology



2. Feature Selection

a. Address multicollinearity using Variance Inflation Factor (VIF).

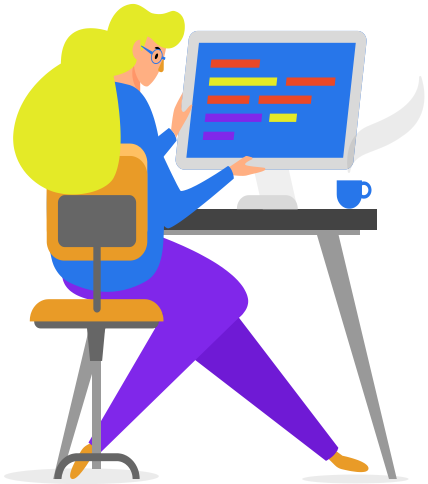
b. Identify highly correlated features to life expectancy.

c. Handle positive and negative skewness with Yeo-Jonson and Square methods.

d. MinMax Scaling to normalize data.

e. Perform Principal Component Analysis (PCA) to reduce dimensionality.

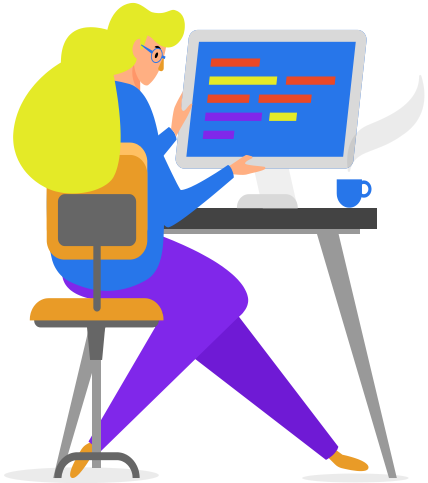
Methodology



3. Model Training

- **a.** Use 60/40 and 80/20 splitted data to compare performance.
- **b.** Build a simple linear regression model over 60/40 data.
- **c.** Build a multilinear regression model with both 60/40 and 80/20 data.
- **d.** Plot the best fit line for each independent feature
- **e.** Compare results from multiple regression models to maximize R^2.

Methodology

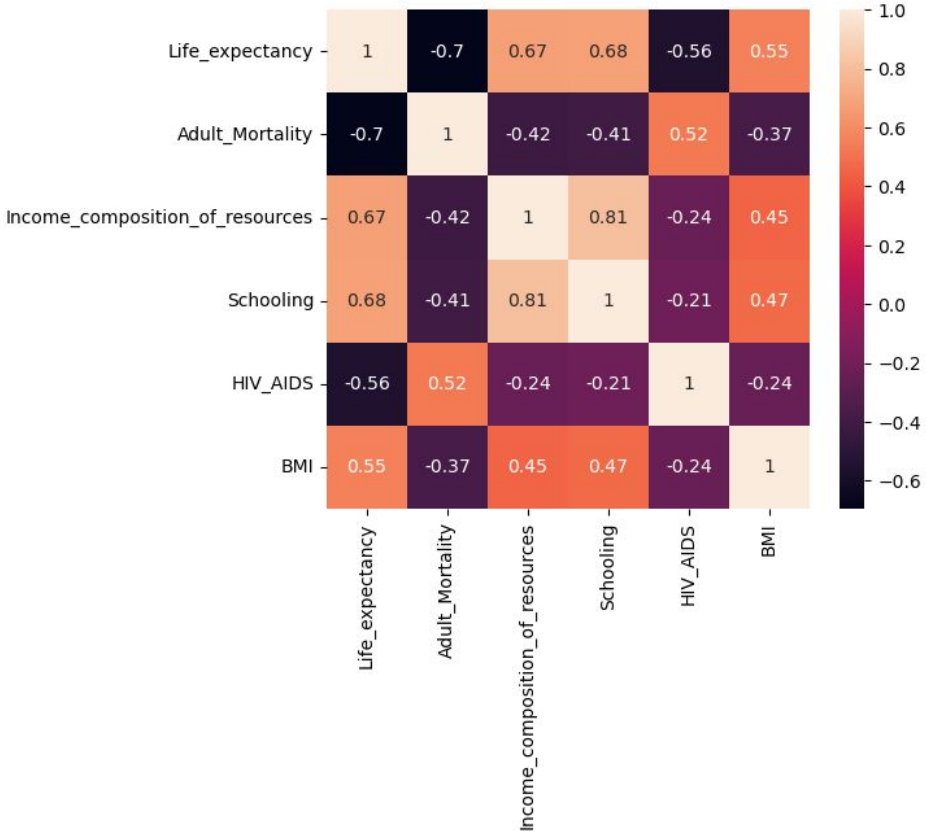


4. **Model evaluation and tuning**
   a. Display test data predictions for each model.
   b. Evaluate models using metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2 Score.
   c. K-fold Cross Validation and Lasso Regression

Results

Top 5 correlated features

Identify top highly correlated features to life expectancy.



Results

Models built over splitted 60% train data and 40% test data

- Multiple Linear Regression R^2 Score for test data: 0.7982
  - Top 5 correlated features.
- Simple Linear Regression R^2 Score for test data: 0.5519
  - Income_composition_of_resources was chosen as the key feature in a simple linear regression model.

Results

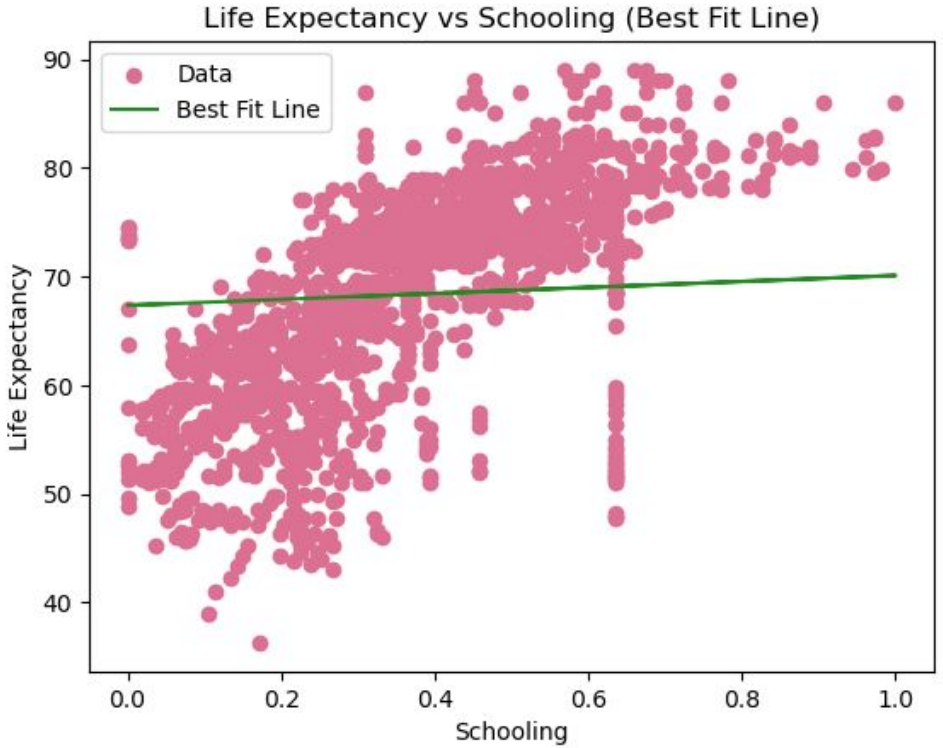
Best fit line for each feature in multilinear regression model

As Adult Mortality increases, Life Expectancy slightly decreases. Higher adult mortality typically suggests poorer health conditions. The best-fit line shows a weak negative trend while the spread in the data indicates Adult Mortality it's not the only factor impacting life expectancy

Results

Best fit line for each feature in multilinear regression model

Even though life expectancy may increase slightly with schooling, it's not enough to create a noticeable slope in a linear regression

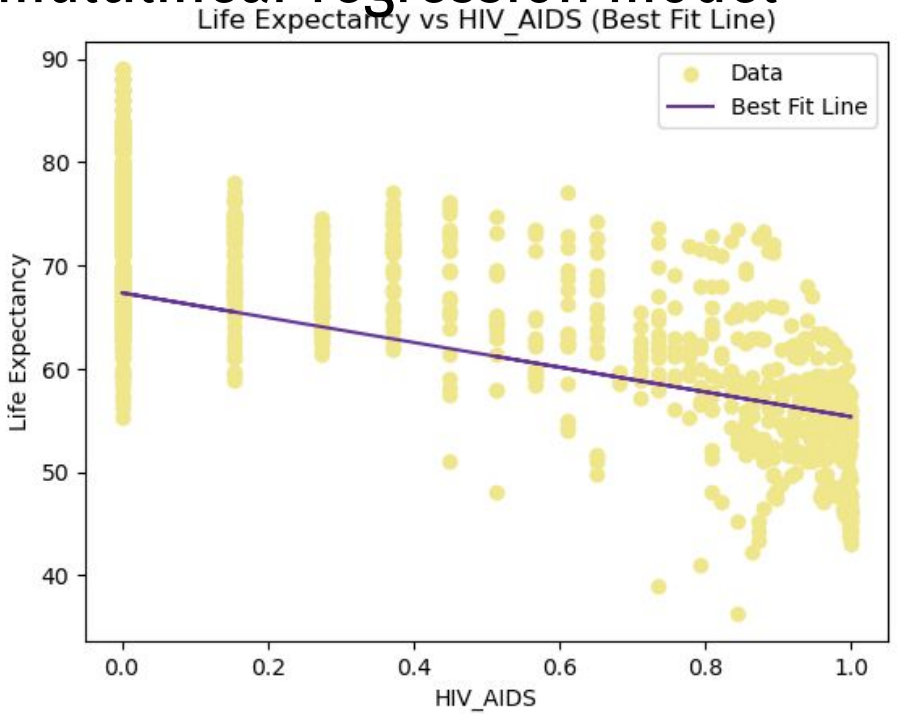


Results

Best fit line for each feature in multilinear regression model

The downward slope of the best-fit line indicates a negative relationship, meaning that as HIV/AIDS prevalence increases, life expectancy tends to decrease.

At low HIV/AIDS rates, life expectancy becomes less predictable based on this feature alone



Results

Best fit line for each feature in multilinear regression model

Since the fit line has a slight positive slope, it indicates a weak positive correlation between BMI and life expectancy. As BMI increases, life expectancy also increases slightly, The spread of the data means BMI is not a dominant predictor by itself.

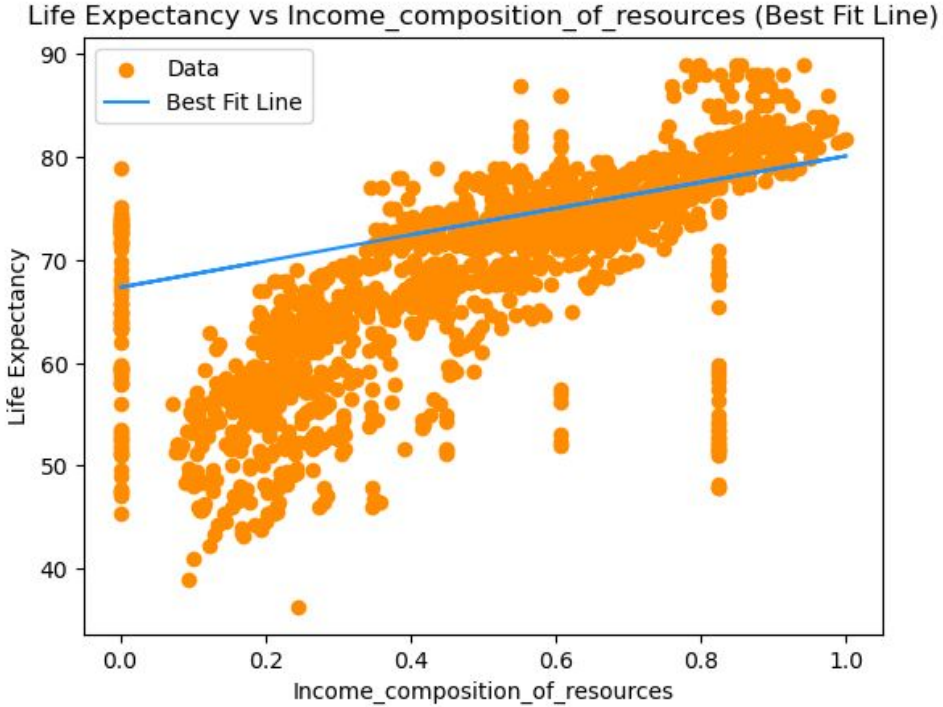
Results

Best fit line for each feature in multilinear regression model

The positive slope of the best-fit line suggests a stronger relationship, as Income Composition of Resources increases, Life Expectancy generally increases.

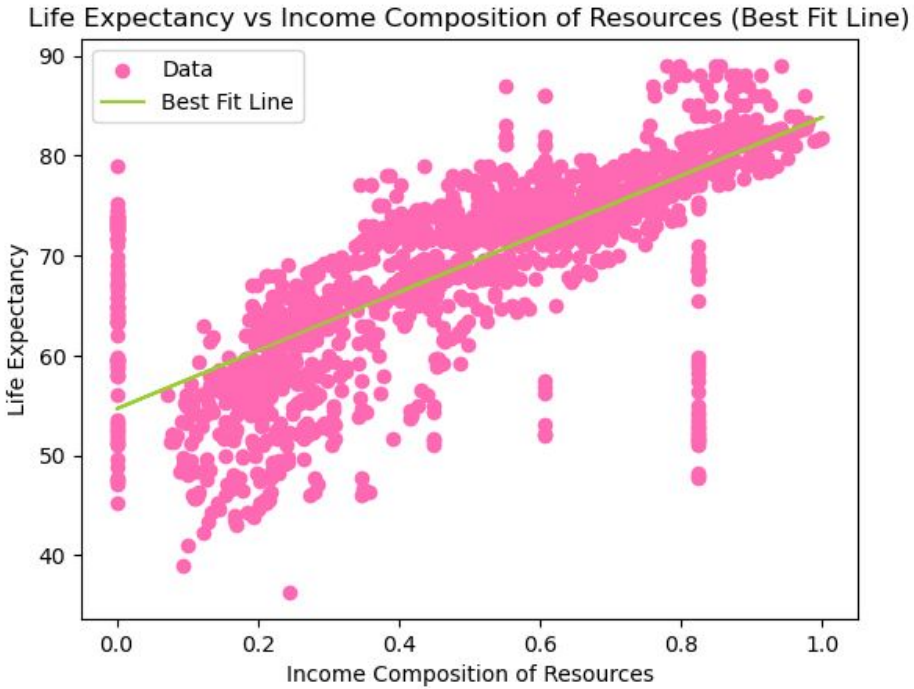
Higher values in this feature are associated with better resource access, often translating to better health care, education, and overall living conditions, which positively affect life expectancy.

This feature might be a key predictor for this model



Results

Best fit line for key predictor feature simple linear regression model (60/40)



Income composition of resources was chosen as the key feature to predict life expectancy in a simple linear regression model.

Model Evaluation and Tuning

Comparing predicted values

Multilinear regression 60/40

	Predicted Value	Actual Value
2402	56.582810	54.0
203	67.749405	67.3
2325	76.232059	79.3
1744	65.510064	73.5
1093	54.467481	57.6
...	...	...
2544	71.033074	73.8
1408	76.418564	73.2
124	79.336922	83.0
2452	71.720805	69.1
1196	64.592463	64.4

Simple linear regression 60/40

	Predicted Value	Actual Value
2402	66.687899	54.0
203	62.473818	67.3
2325	76.423791	79.3
1744	54.646892	73.5
1093	60.265321	57.6
...	...	...
2544	68.405997	73.8
1408	74.704231	73.2
124	81.237323	83.0
2452	70.783285	69.1
1196	63.629443	64.4

Results

Comparing metrics

Multilinear regression 60/40

Metrics
- Mean Absolute Error: 3.10
- Mean Squared Error: 17.70
- R^2 coefficient: 0.7982

Simple linear regression 60/40

Metrics
- Mean Absolute Error: 4.23
- Mean Squared Error: 39.31
- R^2 coefficient: 0.5519

Overall, the model explains a significant proportion of the variance in life expectancy, which suggests that the selected features (Adult Mortality, Income Composition of Resources, Schooling, HIV/AIDS, and BMI) are relevant predictors of life expectancy and given the metrics, we can conclude, life expectancy relies not only in one key factor, but in a bunch of it.

Results

Model built over splitted 80% train data and 20% test data

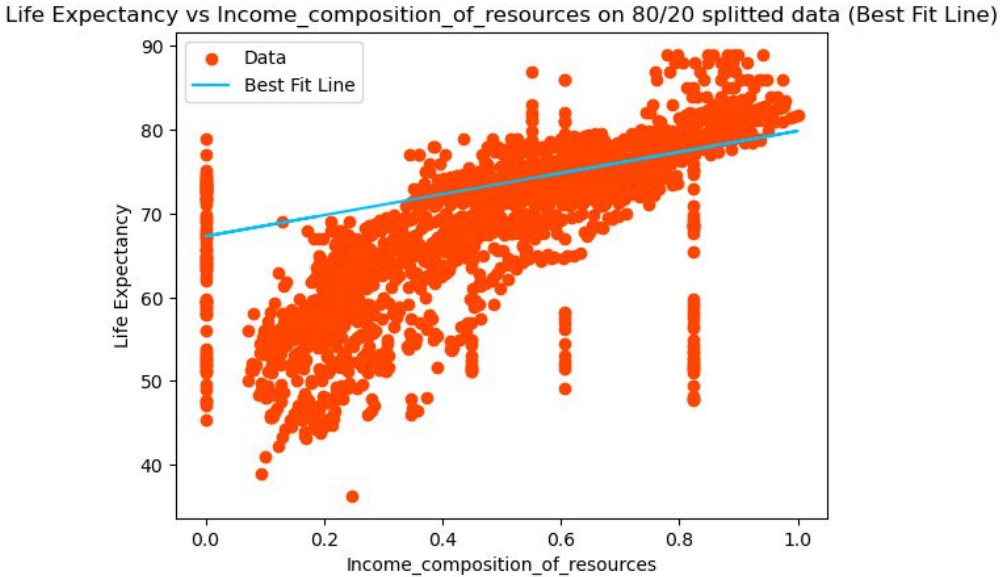
- Multiple Linear Regression R^2 Score for test data: 0.8054
  - Top 5 correlated features.

The most fitable model to capture life expectancy variations is multilinear regression.

After increasing the train data by splitting 80/20, the multilinear model improved significantly.

Results

Best fit line for key predictor feature in a multilinear regression model (80/20)



As Income composition of resources was chosen as the key feature before, it was used to find the best fit line in a multi linear regression model.

The data in slightly less spread than in the fit line with simple linear regression.

Model Evaluation and Tuning

Comparing predicted values (80/20)

	Predicted Value	Actual Value
2402	56.682376	54.0
203	67.681383	67.3
2325	76.167788	79.3
1744	65.567808	73.5
1093	54.608398	57.6
...	...	...
2445	72.800379	74.5
370	77.670208	77.1
1791	61.341972	63.5
975	54.732963	56.6
1954	66.990155	63.5

Results

Comparing metrics

Multilinear regression 80/20

Metrics

- Mean Absolute Error: 3.02
- Mean Squared Error: 17.31
- R^2 coefficient: 0.8054

After increasing the train data by splitting 80/20, the multilinear model showed a significant improvement on the metrics.

Metric results are sign that the model fits the data well and explains a significant portion of the variability in life expectancy.

Results

Comparing metrics

Multilinear regression 80/20

- MAE: On average, the model's predictions deviate from the actual values by about 3.02 years of life expectancy. This indicates that the model provides predictions that are quite close to the actual values, and a lower MAE suggests improved accuracy compared to your previous model.
- MSE: This value represents the average squared difference between the predicted and actual values. An MSE of 17.31 indicates that the model's predictions are relatively precise.
- r2 coefficient: This r2 value means that about 80.54% of the variance in life expectancy can be explained by the model using the selected features.

Cross Validation: Lasso Regression

Since Multicollinearity was already handle by VIF method, K-fold cross validation was performed for Lasso regularization.

```
Lasso
Lasso(alpha=0.01, tol=0.0925)
```

Metrics

- R^2 coefficient after lasso: 0.8047

The drop in R^2 score after Lasso regression is likely due to unnecessary regularization in an already well-performing model. Lasso might not be suitable in this specific case

- The model doesn't show significant signs of overfitting.
- The dataset and feature set may already be optimized.