Concepts and Technologies of AI 5CS037

Regression Analysis Report

Uni ID: 2408485

Submitted by: Aryush Khatri

Lecturer: Ms. Sunita Parajuli

Tutor: Mr. Siman Giri

Submitted on: 2025/02/11

Abstract

The purpose of this analysis is to predict life expectancy using regression models. The chosen dataset underwent preprocessing and visualization to facilitate the analysis. The process involves doing EDA, building models from scratch and from sklearn using linear regression and random forest regression, optimizing hyperparameter tuning, and performing feature selection after that we build a final model using the best parameters.

The evaluation of the models was conducted utilizing R-squared and mean-squared error (MSE). These models demonstrated their capacity to accurately forecast

Life expectancy.

In general, the regression models showed good performance. Essential insights highligh t the importance of feature selection and hyperparameter tuning for enhancing model accuracy.

Table of Contents

1. Introduction

The goal of this project is to predict Life Expectancy using regression models based on the dataset used for the task.

The dataset utilized for this analysis is the Life Expectancy Data .csv, sourced from Kaggle. It holds organized data that is appropriate for regression analysis. This dataset supports the United Nations Sustainable Development Goals (UNSDG) by enhancing health and well-being.

The goal of this analysis is to create a predictive regression model that predicts life expectancy using the attributes found in the dataset.

2. Methodology

- Many missing values were found in the dataset, which was addressed by using the median to fill these gaps. No duplicate values were detected. The categorical variable was encoded through list comprehension.

In Exploratory Data Analysis visualizations like
scatter plots, histograms, and summary statistics are
to comprehend the data distribution and correlations between variables
Two regression models were considered for this task:

- Linear Regression
- Random Forest Regression

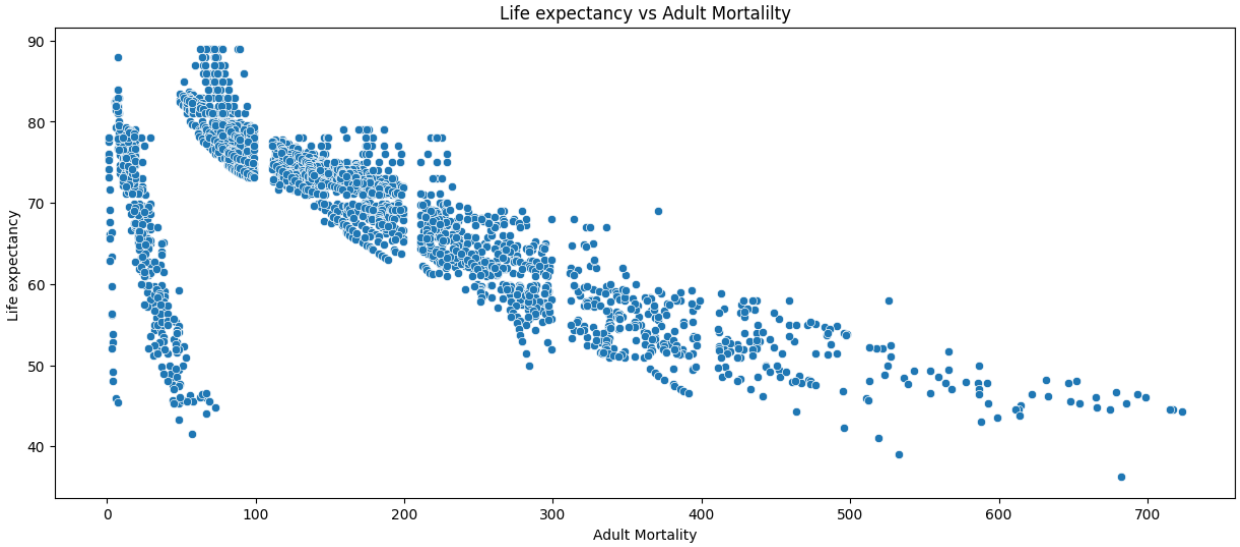
The models were developed by:

• Split the dataset into training and testing sets.

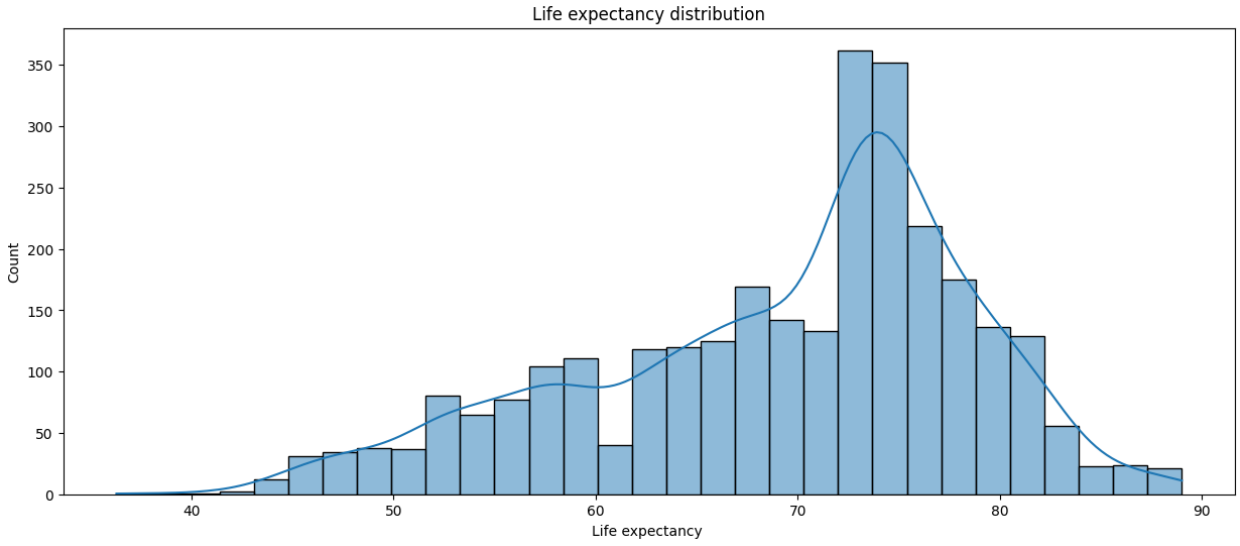
• Train the models using relevant features.

• Perform feature selection to significantly enhance model performance.

• Optimize hyperparameters with RandomizedSearchCV for the best results.



The above scatterplot shows the relationship between Adult Mortality and Life expectancy. We can observe a strong negative correlation between the two variables. As the Adult Mortality rate increases, the life expectancy can be observed significantly decreasing. This shows us that life expectancy is affected byu Adult Mortality rate.



The above plot shows the distribution of our target column. We can observe that it is left skewed with its median being greater than mean.

- Hyperparameter tuning was conducted with RandomizedSearchCV to identify the best parameters that improve model performance. Hyperparameter tuning was conducted with RandomizedSearchCV to identify the best parameters that improve model performance.

Optimal parameters for Random Forest Regressor:

- n_estimators: 100 → The model will use 100 decision trees.
- min_samples_split: 6 → A node must have at least 6 samples to be split.
- min_samples_leaf: 2 → Each leaf node must have at least 2 samples.
- max_features: 'sqrt' → The model will use the square root of the total number of features when considering splits.
- max_depth: 10 → The maximum depth of each decision tree is 10.
- criterion: 'absolute_error' → The model uses Mean Absolute Error (MAE) as the criterion for splitting nodes.

Optimal parameters for Linear Regression:

- fit_intercept=True → The model estimates the intercept (b) in y = mx + c

- Feature selection was conducted using SelectFromModel to identify the most significant features for each model:

Feature selection for RandomForest Regressor: The feature selection process reduced the number of features from 20 to 3, selecting the most important ones based on feature importance scores. These selected features ['Adult Mortality', ' HIV/AIDS', 'Income composition of resources'] contribute the most to the model's predictive performance, improving efficiency while maintaining accuracy.

Feature selection for Linear Regression: SelectKBest selects the top k features based on statistical tests while f_regressor helps to identify features that are most relevant for predicting the target variable in a regression context. The feature selection reduced features from 20 to 5, selecting ['Adult Mortality', ' BMI ', ' HIV/AIDS', 'Income composition of resources', 'Schooling'].

3. Conclusion

The model's performance was evaluated using R-squared and MSE, demonstrating reasonable predictive accuracy. The results showed:

The initial results using Linear Regression:

RMSE:4.2269724848436425

MAE:3.1712162105324135

R2 Score:0.8077570548140709

The initial results using Randomforest regressor:

RMSE:2.1010289340303374

MAE:1.2622312925170092

R2 Score:0.9525041530803975

The Randomforest regressor was the most effective at predicting Life Expectancy, with r2 score 0.95

- Selecting the right models for this dataset
- Finding the best parameters using RandomizedSearchCV was computationally time-consuming

Future improvements may involve using advanced regression techniques, increasing dataset size, and refining feature engineering.

4. Discussion

The model's performance was evaluated using standard regression metrics, demonstrating that Random Forest Regression outperformed Linear Regression across all metrics. Implementing hyperparameter tuning and feature selection significantly enhanced other metrics, such as RMSE and MAE, although the R^2 score experienced a slight decrease.

The chosen features played a crucial role in the model's performance, and the results aligned with our expectations. However, there were certain limitations, including constraints related to the dataset and specific assumptions made by the model.

Future research could focus on exploring different regression algorithms, expanding the dataset, and applying more advanced feature engineering techniques.