Predicting Life Expectancy over the World Using Data From 2000-2015

Xiner Zhao
DSI, Brown University
Github: Life-Expectancy-Prediction

1 Introduction

(1) Motivation

Forecasting life expectancy is more and more crucial with demographic shifts and an aging population. Accurate predictions guide policymakers, healthcare professionals, and researchers in intervention planning and resource allocation. Addressing evolving healthcare needs and challenges tied to an aging society, life expectancy prediction informs decisions on healthcare and pension obligations, serving as a roadmap for meeting diverse population needs amid changing demographics.

Furthermore, this project aims to answer the following key questions:

Is there a discernible difference in life expectancy between developing and developed countries?

Which factors exhibit strong correlations with life expectancy, and which of these factors significantly influence it?

Can machine learning models effectively predict life expectancy? How do their performances compare to a baseline model?

(2) Dataset Description

This project utilizes a Kaggle-sourced[1] dataset from the World Health Organization (WHO), offering comprehensive information on life expectancy. Encompassing demographic indicators, socio-economic factors, healthcare metrics, and lifestyle elements, the dataset is regularly updated for accuracy. It spans life expectancy, health, immunization, economic, and demographic data for 179 countries from 2000-2015, featuring 21 variables and 2864 datapoints.

(3) Current Research

Up to 2023, A lot of studies on life expectancy prediction have been carried out. For example, the results of a study done on the Asian countries showed that the higher levels of socioeconomic advantage and more excellent healthcare resources of the people were more likely to enhance life expectancy[2]. Nowadays, several machine learning methods have been used in previous studies on life expectancy, including Decision Tree, Naive Bayes, k-Nearest Neighbor and Support Vector Machine[3].

2 Exploratory Data Analysis

To enhance comprehension of the data's structure and distribution and gain deeper insights into the correlation between the target variable and features, I conducted exploratory data analysis.

Based on Figure 1, which visualized the distribution of life expectancy, we observe a relatively balanced distribution of life expectancy, albeit slightly skewed. Most countries exhibit life expectancies clustered around 70-75 years.

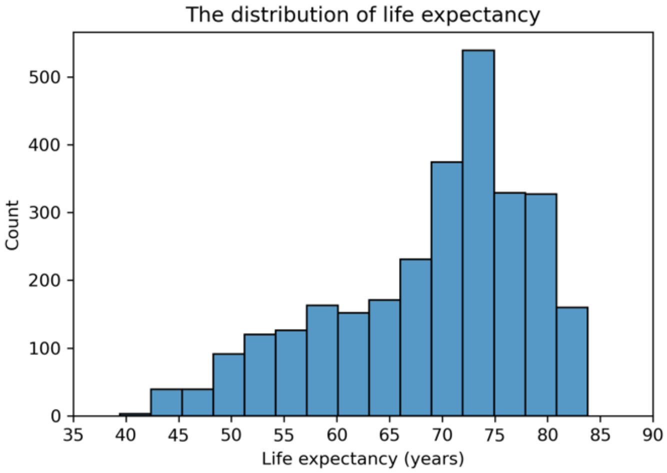


Figure 1. The distribution of life expectancy

Subsequently, a correlation matrix heatmap (Figure 2) was generated to visually explore the correlations between each feature and the correlation between the target variable and each feature.

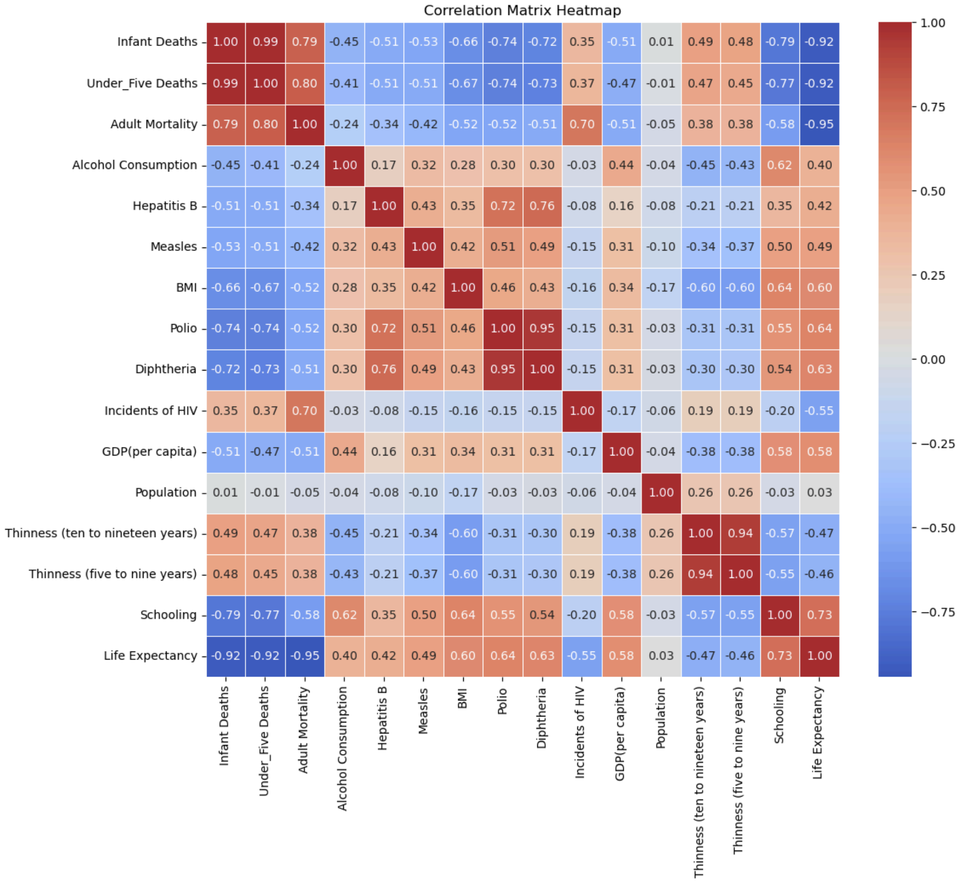


Figure 2. Correlation matrix heatmap

From the correlation heatmap, it becomes evident that life expectancy shows strong correlations with adult mortality, infant deaths, under-five deaths, and schooling.

Given the strong correlation of adult mortality and schooling with life expectancy, I created two visualizations (Figure 3 and Figure 4). Developing countries are represented by blue dots, while developed countries are depicted by orange dots in the plots.

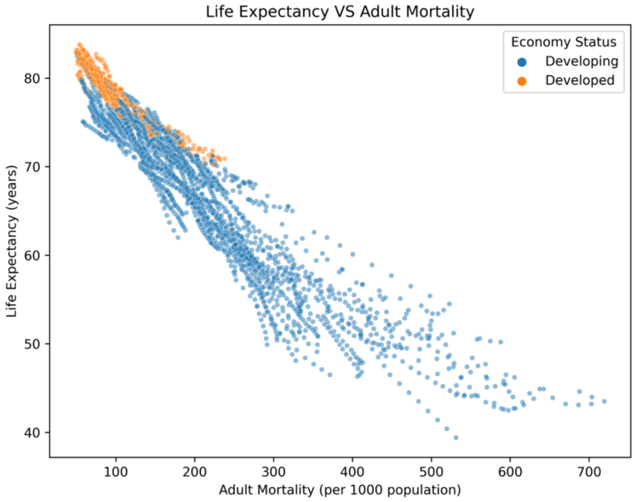


Figure 3. Life expectancy VS Adult Mortality

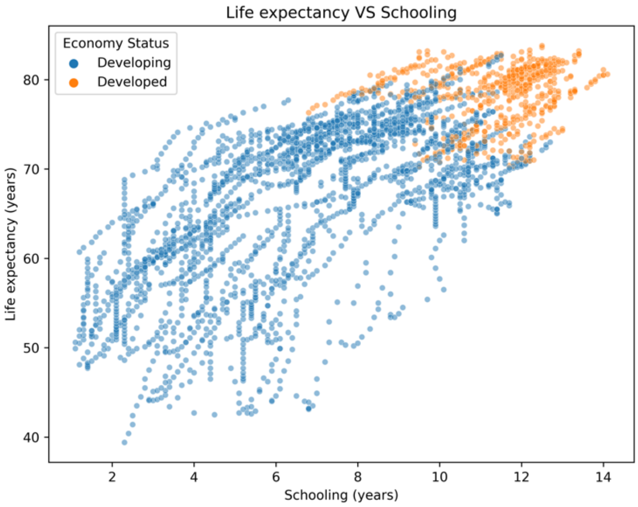
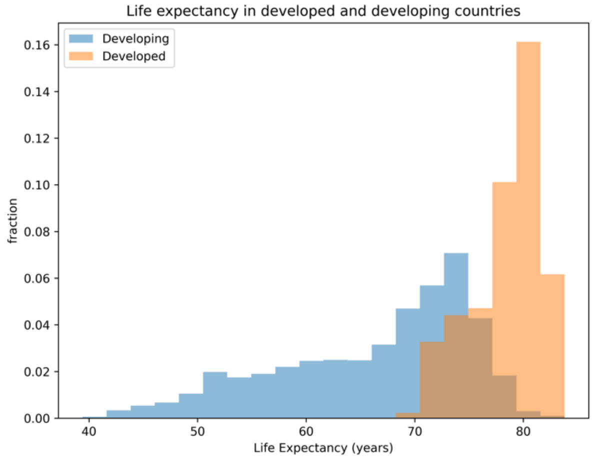


Figure 4. Life expectancy VS Schooling

Figure 3 provides a clear depiction: as adult mortality rises, life expectancy decreases. Moreover, at equivalent levels of adult mortality, developed countries generally exhibit higher life expectancies than developing ones. In Figure 4, the scatter plot of life expectancy with schooling in developed and developing countries reveals that higher schooling years in a country tend to correspond with a longer life expectancy.



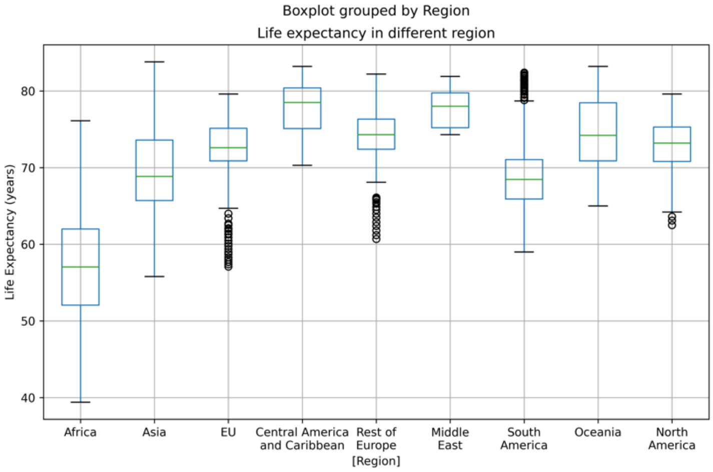


Figure 5. Life expectancy in developed and developing countries

Figure 6. Life expectancy in different regions

Furthermore, Figure 5 explicitly reveals that overall life expectancy is higher in developed countries compared to developing ones. Additionally, Central America and the Caribbean exhibit the highest life expectancy among nine regions, whereas Africa has the lowest. Notably, the dataset has no missing values, eliminating the need for data imputation methods.

3 Methods

Figure 7 illustrates the overall pipeline and methodologies employed in this machine learning-driven project.

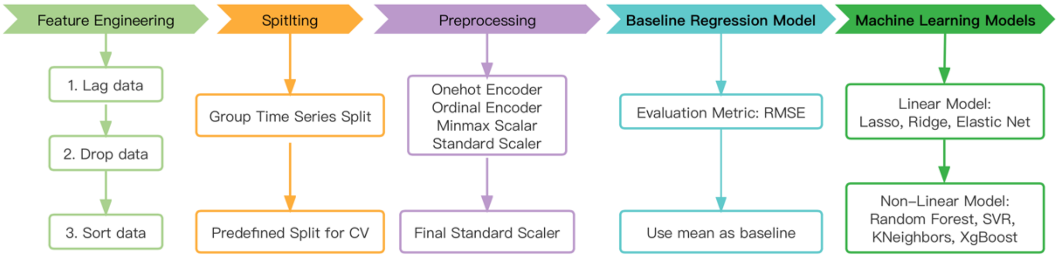


Figure 7. Pipeline

(1) Feature Engineering

As the dataset is a time series spanning from 2000 to 2015, the initial step involves feature engineering before we split it. To achieve this, we lagged the target variable and all features, excluding Country, Region, Year, and Economy Status, for 1-5 years, respectively. However, this resulted in some datapoints having a substantial number of missing values, rendering them meaningless. To enhance computational efficiency, these datapoints were excluded. The dataset was then sorted in ascending order by years to facilitate the splitting of the time series dataset.

(2) Splitting

In handling our time series dataset, the initial approach is to employ Time Series Split, ensuring the train set contains the earliest data and the test set comprises the most recent. However, Time Series Split is designed for situations where each time period has only one datapoint, and our dataset contains data for 179 countries per year. This means that, if applied, Time Series Split might scatter datapoints from the same year across different sets. To address this problem, we used the Group Time Series Split method, guaranteeing that all datapoints from the same year remain within the same set.

(3) Preprocessing

During the preprocessing step, Onehot Encoder, Ordinal Encoder, Minmax Scalar, and Standard Scalar were used to preprocess the whole dataset, which could help improve the performance and stability of machine learning models in the next few steps.

In addition, for certain linear models, Standard Scaler was applied after Onehot Encoder to normalize coefficients so as to make sure all features' mean is 0 and standard deviation is 1, which is important for later analysis of feature importance.

(4) Evaluation Metric

Since the target variable life expectancy is a continuous variable, the project is a regression machine learning model. Thus, we chose Root Mean Square Error (RMSE) as the evaluation metric.

(5) Baseline Regression Model

The mean of the life expectancy in the test set serves as the baseline for this regression problem. Subsequently, we computed the root mean square error (RMSE) of the life expectancy in the test set relative to the baseline, resulting in approximately 7.9233. This value was designated as the baseline test score. Any model with a test score (RMSE) lower than the baseline score indicates superior performance compared to the baseline regression model. Conversely, a higher score implies poorer performance, suggesting limited predictive power.

(6) Machine Learning Regression Models

To enhance the precision of life expectancy predictions, seven machine learning models were implemented. Specifically, linear regression models—Lasso Regression, Ridge Regression, and Elastic Net Regression—were employed, alongside non-linear models, including SVR, Random Forest, KNeighbors, and XGBoost, for the prediction of life expectancy.

(7) Hyperparameter Tuning and Cross-Validation

Hyperparameter tuning aims to discover the best parameter values that improve model performance, with a focus on the chosen evaluation metric RMSE. Iterating through potential parameter configurations on the training set, the best values are determined by evaluating performance on the cross-validation set. Table 1 presents the parameters and their corresponding values for each machine learning model during the tuning process. Additionally, a Predefined Split was employed for cross-validation, considering the chronological order predefined across the train, validation, and test sets.

Table 1. Hyperparameter Tuning

Model	Hyperparameter	Values
Lasso Regression	alpha	29 proportional values between $10^{-7}$and 10^{0} (np.logspace(-7,0,29))
Ridge Regression	alpha	51 proportional values between $10^{-10}$and 10^{0} (np.logspace(-10,0,51))
Elastic Net Regression	alpha l1_ratio	[0.1, 0.5, 1, 2, 5, 10] [0.1, 0.3, 0.5, 0.7, 0.9]
Random Forest	max_depth max_features	[None, 1, 3, 10, 30, 100] [None, 0.5, 0.75, 1.0]
Support Vector Regressor	C gamma	$[10^{-1}, 10^{0}, 10^{1}]$ $[10^{-3}, 10^{-1}, 10^{1}, 10^{3}, 10^{5}]$
KNeighbors	n_neighbors	[1,3,10,30]
XGBoost	max_depth reg_alpha reg_lambda	[1,3,10,30,100] $[0, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}]$ $[0, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}]$

Once the best values for each parameter in every model were determined, we predicted life expectancy in the test set and computed the RMSE for each model using their best parameters. To address uncertainties stemming from splitting and non-deterministic models, we conducted model training and testing across 10 different random states.

4 Results

(1) Model Comparison

With the best parameters for each model, the mean RMSEs were computed over 10 different random states for each model and each dataset (in various lags), as illustrated in Figure 8. As depicted in the results, all models exhibited test scores lower than the baseline test score of 7.9233, as mentioned earlier. Notably, Lasso Regression in the 4-year lag had the smallest RMSE at 0.1978, indicating the best performance in predicting life expectancy. Conversely, KNeighbors in the 1-year lag had the largest RMSE at 1.1255, signifying the least predictive power.

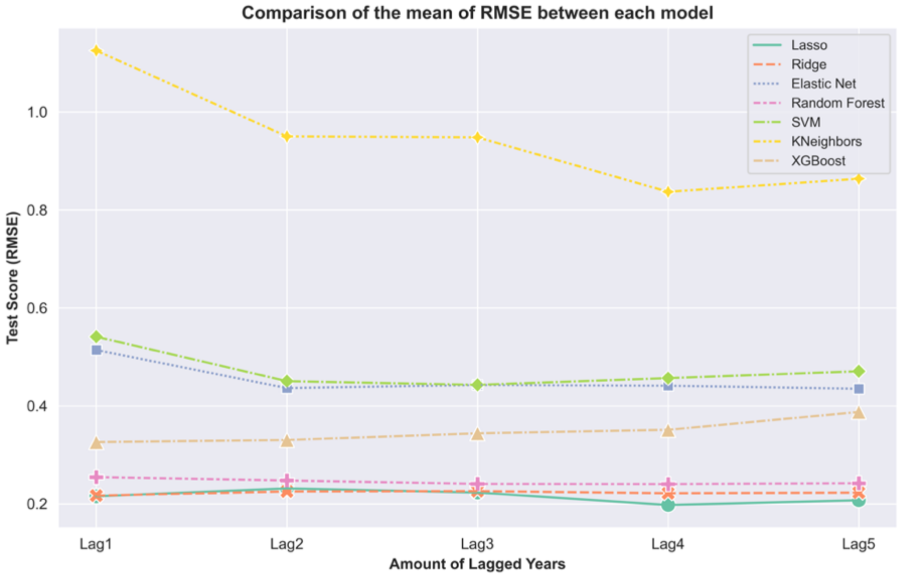


Figure 8. Comparison of the mean of RMSE between each model

Figure 9 illustrates the comparison of the standard deviation of RMSE among each model, revealing that the standard deviation for all models is consistently below 0.005. While the standard deviation of Random Forest is slightly higher than that of other models, it remains within a tiny range.

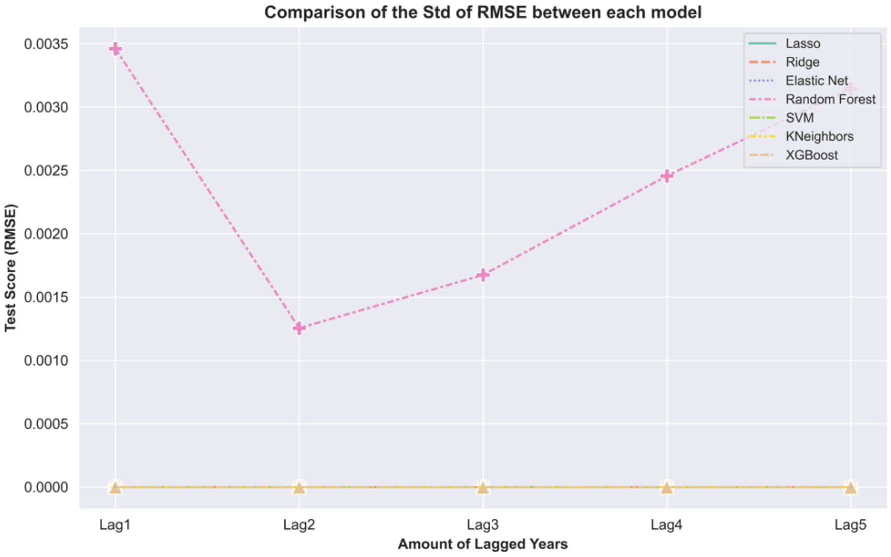


Figure 9. Comparison of the Std of RMSE between each model

(2) Global Feature Importance

Global feature importance aids in assessing the overall contribution of each feature to a model's predictive performance and facilitates the identification of key features. Particularly, three types of metrics were employed to gauge feature contribution in the best model, Lasso Regression in 4-year lag.

Firstly, permutation importance was applied to identify the top 10 important features, as depicted in Figure 10. The figure highlights that life expectancy in the 1-year lag emerged as the most crucial feature in predicting overall life expectancy.

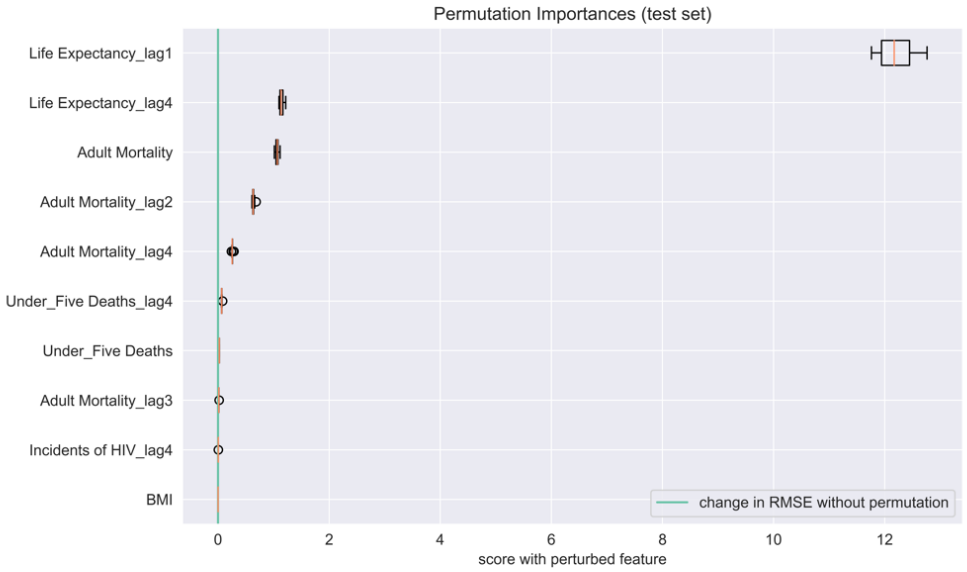


Figure 10. Top 10 important features by using permutation importance

Secondly, we used the value of coefficients to explain the feature importance in Lasso Regression. As illustrated in Figure 11, life expectancy in the 1-year lag retains its position as the most crucial feature. Notably, the second most important feature shifts to adult mortality, while the third becomes life expectancy in the 4-year lag. Furthermore, the coefficient of life expectancy in the 1-year lag is five times greater than that of adult mortality.

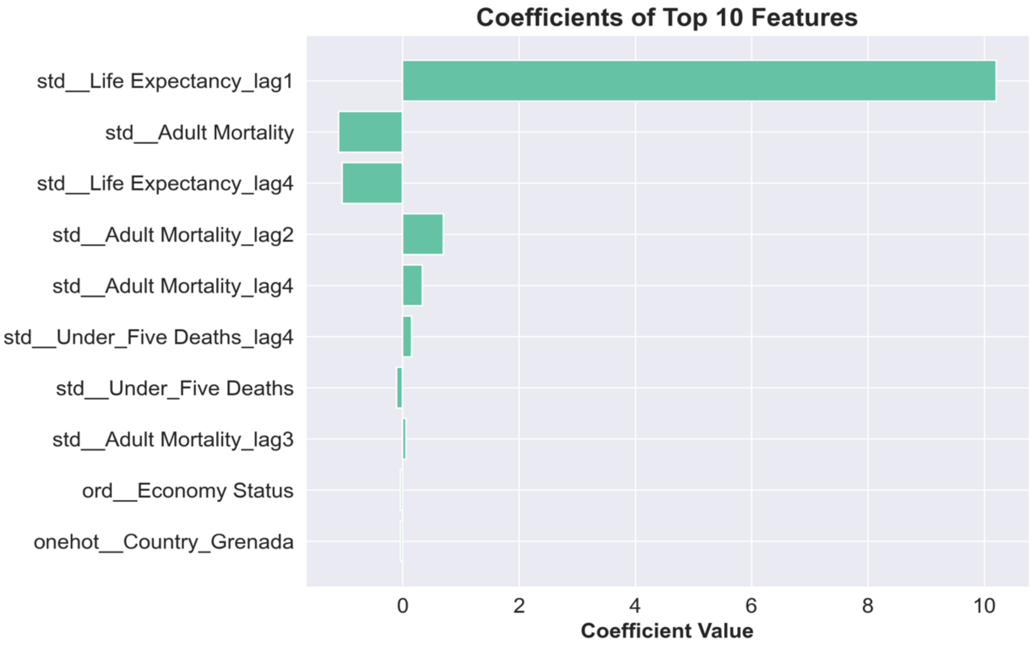


Figure 11. Feature importances by comparing coefficients

Lastly, we employed the mean of SHAP values to visualize the contribution of each feature, as illustrated in Figure 12, and we could again see that life expectancy is the most important feature in Lasso Regression with a 4-year lag.

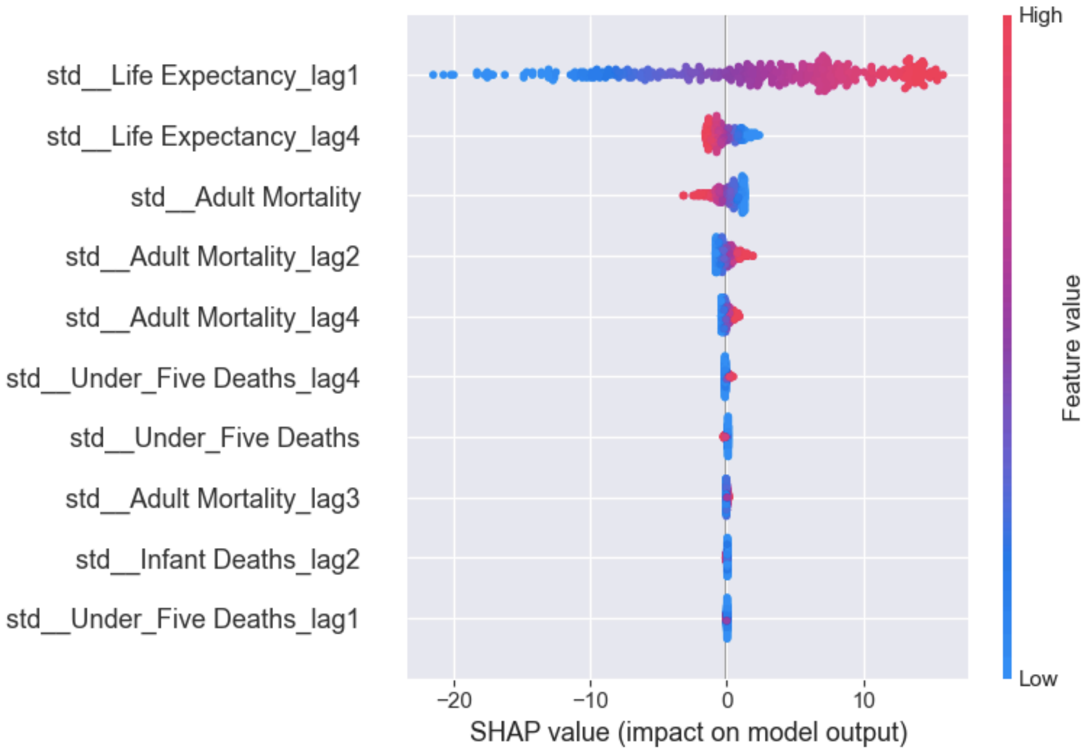


Figure 12. Feature importances by using shap value

In conclusion, despite the varied values and methodologies employed across the three feature importance metrics, the top 8 important features remain consistent. Compared to the correlation heatmap we created before in the EDA part, adult mortality, under-five deaths, and infant deaths are still three important features that would influence life expectancy. However, what surprised us most was that schooling, which was highly correlated with life expectancy as we saw in the correlation heatmap before, was not the important feature anymore here.

(3) Local Feature Importance

Each feature's contribution to the model's prediction relative to the baseline is depicted in SHAP force plots (Figure 13, 14, 15). Using the datapoints for India, South Africa, and China in 2015, positive values in red signify an increase in prediction, while negative values in blue indicate a decrease. For example, in Figure 13, life expectancy in the 1-year lag exhibits the most negative contribution, whereas adult mortality contributes the most positively—a pattern similarly observed in Figure 14 and 15.

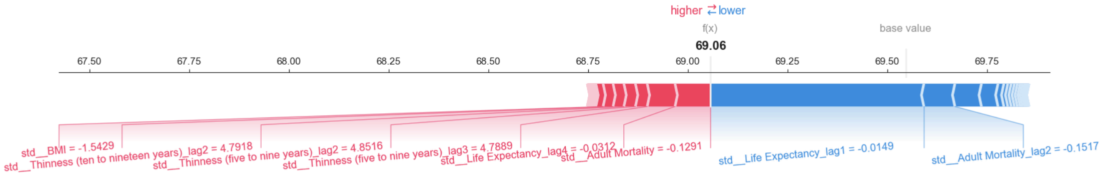


Figure 13. Shap force plot for the datapoint of India in 2015

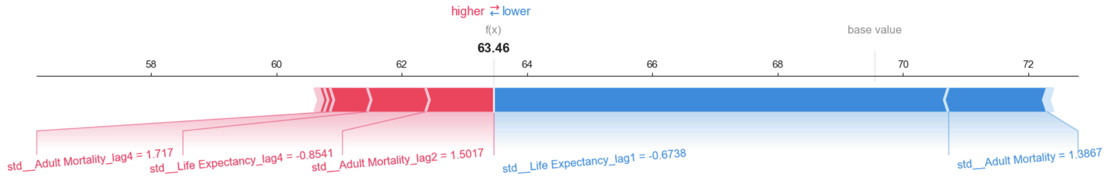


Figure 14. Shap force plot for the datapoint of South Africa in 2015

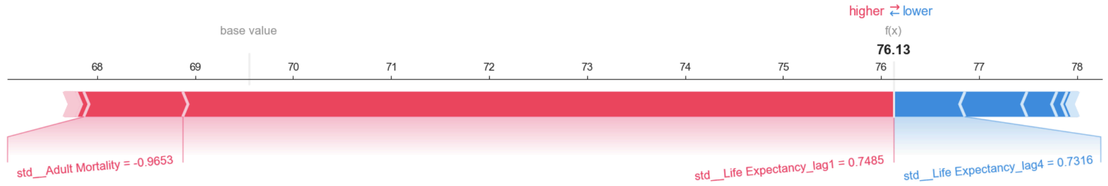


Figure 15. Shap force plot for the datapoint of China in 2015

5 Outlook

Nothing is perfect, so is this project. Despite the progress made in this project over several months, there are areas that could be enhanced, considering the following aspects:

Firstly, the dataset covers life expectancy data only from 2000 to 2015. Including more recent years' data could bolster the models' credibility.

Additionally, while RMSE is employed as the evaluation metric for this regression problem, exploring alternative metrics such as R2 score and Mean Absolute Error might yield diverse insights.

Furthermore, broadening the parameter range and experimenting with a wider array of models could potentially enhance the accuracy of model predictions.

Lastly, XGBoost offers five different metrics for measuring feature importance. Integrating additional methods for assessing feature importance might provide a more comprehensive interpretation of models from varied perspectives.

6 References

[1] https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated?rvi=1

[2] Chan MF, Devi MK. Factors affecting life expectancy: evidence from 1980-2009 data in Singapore, Malaysia, and Thailand. Asia Pac J Public Health. 2015 Mar;27(2):136-46. doi: 10.1177/1010539512454163.

[3] Mohammad Suhaimi, Nurafifah, et al. "Predictive model of graduate-on-time using machine learning algorithms." *Soft Computing in Data Science: 5th International Conference, SCDS 2019, Iizuka, Japan, August 28–29, 2019, Proceedings 5*. Springer Singapore, 2019.

[4] Pisal, Nurul Shahira, et al. "Prediction of life expectancy for Asian population using machine learning algorithms." *Malaysian Journal of Computing* 7.2 (2022): 1150-1161.

[5] Agarwal, Palak, et al. "Machine learning for prognosis of life expectancy and

diseases." *Int J Innov Technol Explor Eng (IJITEE)* 8.10 (2019): 1765-1771.

[6] Bali, Vikram, et al. "Life Expectancy: Prediction & Analysis using ML." *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*. IEEE, 2021.

[7] Ji, Sukwon, Bumho Lee, and Mun Yong Yi. "Building life-span prediction for life cycle assessment and life cycle cost using machine learning: A big data approach." *Building and Environment* 205 (2021): 108267.