MODELING MODERN CONTRIBUTIONS TO GLOBAL LIFE EXPECTANCY

JAPHETH CARLSON AND NATHAN LONGHURST

Abstract. Recent developments have revolutionized the medical field and access to quality healthcare across the globe, transforming the improvement, and disparity, of global average life expectancy. We seek to answer the question, *What are the most important features impacting global life-expectancy in the 21st-century, and how would life-expectancy improve if these issues were mitigated?* We analyze a comprehensive global dataset incorporating myriad features to identify which have the greatest impact on modern life expectancy, as well as predict the potential benefits of mitigating the influence of the most detrimental features.

1. Research question and overview of the data

It is well-documented that life-expectancy varies significantly by one's home country (e.g., [1], [2] and [3]), but also that life expectancy can be greatly improved through combined individual and government efforts.[4][5] Recent efforts expand this to project future life expectancy given current trends using standard linear regression [6], clustering [7], and Bayesian model ensembling.[8][9] We note a clear gap for modeling future life expectancy given concerted efforts to improve targeted health categories.

Recently, statistics compiled by the World Health Organization (WHO) detail the life expectancy of 193 recognized nations spanning the years 2000-2015. [10] Alongside life expectancy (by year) are attributes in several categories, including immunization, mortality, national economy, and lifestyle. We analyze this dataset by feature, then model life expectancy improvements by targeting specific changes in one feature. This allows us to not only find the features with strongest correlation to mortality, but also targeted approaches for each nation that would maximize their respective increase in life expectancy. Through this analysis, we seek to answer the all-important question: *What features impact global life-expectancy most in the 21st-century, and how would life-expectancy improve if they were mitigated?*

We anticipate verifying that developing nations have significantly lower life expectancies than developed nations. Moreover, we anticipate lifestyle factors, such as alcohol consumption and average body mass index (BMI), to play a large role in a nation's life expectancy; however, we expect factors

Date: February 1, 2026.

influenced primarily by one's government (e.g., GDP, immunization rates, etc.) to have the greatest impact in a given population's life expectancy.

We selected this dataset because it comes directly from the WHO, with data from nearly all recognized nations around the world. This offers strong confidence in the data's consistency, completeness, and reliability, even for countries without dependable data reporting. The most notable weakness of this dataset includes various data entries that appear inconsistent or missing. We correct for weaknesses and describe our methods in Section 2.

Overall, this dataset represents the most reliable and complete global health dataset for the years 2000-2015, allowing us to analyze global health and separate the impact of both government expenditure and personal lifestyle. We expect to find the impact of common health practices (e.g., immunization) and model the potential outcomes of populations in the future should specific changes in government spending and/or personal lifestyle be made. As a result, we hope to also answer the reader's most pressing question, *How much of a role do my choices have in my life expectancy and how much of that depends on the country I live in?*

2. Data Cleaning / Feature Engineering

2.1. **Data Cleaning.** Our initial analysis of the dataset revealed several glaring problems. For instance, 212,183 cases of measles were recorded per 1,000 people in Nigeria for the year 2000.[10] While most reported values seemed plausible, it was clear that record keeping was inconsistent at best. We have many attempts of throwing out missing data and attempting to scale data by a factor of 10 where it seems a decimal place was shifted in recording (e.g., GDP was recorded in Australia with values of 36118, 4991, and 49664, over three consecutive years). Values beyond reasonable bounds (0-100 for percentages) were also thrown out or scaled based on the same shift in the decimal place where this error seemed relevant. We do not recommend use of this dataset by anyone; however, in our attempts we tried to keep our code as robust as possible.[1] Due to the erroneous dataset, we used another publicly available dataset with the permission of Dr. Grant that seeks to correct missing values by calculating the nearest three-year average. [11] Much of the data was also re-recorded and gathered from public sources. This new dataset has no values outside the reasonable range or jumps in orders of magnitude from year-to-year for any country.

After cleaning our dataset, we noticed several features that directly affect the calculation of life expectancy; however, these features may not be as useful if we hope to prescribe how a country can improve their life expectancy. Specifically, we ignore adult mortality, infant deaths, and under five deaths, as no additional analysis needs to connect these features with life expectancy.

[1] Our project code is hosted at: https://github.com/nlong1/vol3-project-semester1

	Random Forest	XGBoost	LR (AIC)	LR (BIC)
R^2	0.9174	0.9918	0.8267	0.8156
MSE	6.8531	0.6794	14.3807	15.2991

Table 1. Summary statistics comparing quality of prediction for each model (LR = Linear Regression).

0.5630	0.1791	0.0887	0.0376	0.0365
HIV Cases	GDP (Log)	Schooling	BMI	Developed Economy

Table 2. Most correlated features to life expectancy, found by XGBoost (higher values indicate stronger correlation).

2.2. **Feature Engineering.** From the original dataset, we engineered three new features which we included in our analysis.

2.2.1. *Log of GDP.* Because the relative value of money decreases as one's wealth increases, we engineered a new feature by taking the logarithm (base-10) of each nation's per-capita GDP to more closely approximate the relative impact felt by the individuals of the population based on their nation's GDP.

2.2.2. *Vaccination Score.* To succinctly summarize the overall impact of vaccination for each population, we averaged the vaccination rates for all listed diseases into a single *Vaccination Score* for each given year and nation.

2.2.3. *Lifestyle Index.* To interpret the cumulative importance of lifestyle, we define the lifestyle index. For each lifestyle category given (BMI and alcohol consumption), we assigned a score of 0 (ideal), 1 (elevated risk), and 2 (significant risk) based on the national average for each respective category. The sum over both categories defines the nation's overall lifestyle index score. Note that for BMI, we defined a 0 as those in the "healthy range," 1 as those "overweight," and 2 as those "obese," as described by the CDC.[12] Current research is inconclusive on the risks of alcohol consumption, so we assigned the category score based on the average number of drinks per day (0, 1, or 2), which yielded scores consistent with our rating of BMI.

3. Data Visualization and Basic Analysis

3.1. **Data Summary.** Table 1 summarizes the accuracy scores for each of the four regression models we employed. From the table, we find XGBoost clearly outperformed the other methods, which meets initial expectation when compared to a random forest or linear regression.

From our XGBoost model, we summarize the five most important features for predicting life expectancy (in order) as Table 2 . From this table, it is clear HIV is by far the most predictive feature in life expectancy, followed by the (log) GDP, with all other features showing minimal correlation.

3.2. **Basic Analysis.** Figure 1 demonstrates the predicted change in life expectancy from each model when reducing HIV incidence rate by four per 1,000 people aged 15-49. This illustrates both the predictive ability of our model and the nuance needed to identify which parameters impact life expectancy most. As the HIV incidence rate decreases, we find diminishing returns provided by reducing the rate even further. To find the ideal parameter(s) for a country to change, and by how much, a simple grid search could be performed to find the optimal candidate(s) and the theoretical impact.

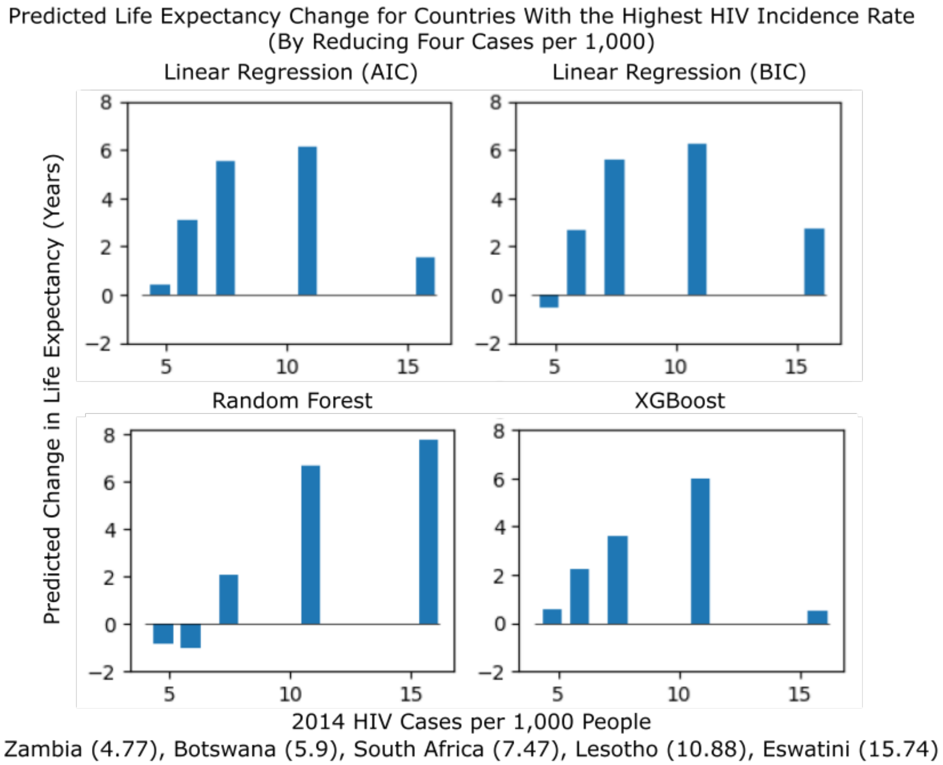

Figure 1. Modeled change in life expectancy for five countries if each reduced their HIV incidence rate by four per 1,000 people aged 15-49 (leaving all other features constant).

4. Learning Algorithms and In-depth Analysis

We performed a grid search and found the models with the best mean squared error (MSE) for both random forest and XGBoost. With linear regression, we optimized for AIC and BIC scores and saved both models to use in predictions. In each model, HIV incidence and log of GDP per capita were among the three most influential factors for predicting the national life expectancy. Schooling ranked third for both random forest and XGBoost.

We note the considerable improvement of MSE for the random forest and XGBoost models when evaluated with our test set in Table 1. We draw from this that governmental factors seem much more important than the available individual factors for life expectancy (BMI and alcohol consumption), though lifestyle factors (BMI and, at least partially, HIV incidence) provide notable contribution as well.

Although there are some lifestyle factors that promote an increase in HIV incidence, nations with high HIV incidence tend to have poor medical and sanitation practices, indicating at least some environmental/government causes for this feature. We model the life expectancy shift in five different countries should they each reduce their HIV incidence rate in Figure 1, and previously explored this in Section 3.2. However, to truly understand what personal factors play into global life expectancy, a more thorough analysis, including more detailed features with reliable data recording, must be obtained. Exercise levels, nutrition, and many other factors may provide valuable insights. Next semester, we hope to explore these features and find how they influence life expectancy, since BMI and alcohol consumption do not form a conclusive (or, evidently, very predictive) lifestyle index.

5. Ethical Implications and Conclusions

Because all used data is publicly published by the WHO, we do not find any ethical implications concerning the data itself, its processing, or use.

Our original question asks whether a nation's life expectancy is driven by its government or cultural lifestyle. As such, our results and conclusions are inherently political and personal. Ethically, national leaders and individuals alike must understand single results cannot generalize to all circumstances. As Figure 1 illustrates, what works for one nation may not work as well (or at all) for another. Thus, the answer to our question is highly dependent upon the specific circumstances of each country and no single factor sufficiently describes their outcome. Therefore, we recommend each nation (and individual) to carefully examine *all* relevant factors – including **both** governmental and cultural – when addressing better health outcomes for its people (or self). Ethically, the clear solution is to work together and improve as many features together to find the right combination of lifestyle and government assistance to maximize life expectancy.

Our models, particularly with XGBoost, provide an accurate framework for contextualizing each feature and predicting the life expectancy change. We found the most important global feature is HIV incidence, which is influenced by both governmental and lifestyle factors. Moreover, our model provides the necessary framework to perform a grid search to predict future improvements in life expectancy based on improving specific features. We found the most effective, and ethical, means of improving national life expectancy is by improving both lifestyle and governmental assistance in parallel, based on the unique feature make-up of each group of people.

References

[1] Galvani-Townsend, S., Martinez, I., & Pandey, A. (2022, November 12). *Is life expectancy higher in countries and territories with publicly funded health care? Global Analysis of Health Care Access and the social determinants of health.* Journal of global health. https://pmc.ncbi.nlm.nih.gov/articles/PMC9653205/

[2] Steel, N., Bauer-Staeb, C. M. M., & Ford, J. A. (2025, March). Changing life expectancy in European countries 1990–2021: A subanalysis of causes and risk factors from the global burden of Disease Study 2021 - the lancet public health. https://www.thelancet.com/journals/lanpub/article/PIIS2468-2667(25)00009-X/fulltext

[3] Freeman, T., Gesesew, H. A., & Bambra, C. (2020, November 10). Why do some countries do better or worse in life expectancy relative to income? an analysis of Brazil, Ethiopia, and the United States of America - International Journal for equity in health. https://link.springer.com/article/10.1186/s12939-020-01315-z

[4] Crimmins, E. M. (2015, December). *Lifespan and Healthspan: Past, present, and promise.* The Gerontologist. https://pmc.ncbi.nlm.nih.gov/articles/PMC4861644/

[5] Fadnes, L. T., Celis-Morales, C., & Økland, J.-M. (2023, November 20). *Life expectancy can increase by up to 10 years following sustained shifts towards healthier diets in the United Kingdom.* Nature News. https://www.nature.com/articles/s43016-023-00868-w

[6] Cao, X., Hou, Y., & Zhang, X. (2020, December). *A comparative, correlate analysis and projection of global and regional life expectancy, healthy life expectancy, and their gap: 1995-2025.* Journal of global health. https://pmc.ncbi.nlm.nih.gov/articles/PMC7568920/

[7] Levantesi, S., Nigri, A., & Piscopo, G. (2023, January 14). *Multi-country clustering-based forecasting of healthy life expectancy - quality & quantity.* SpringerLink. https://link.springer.com/article/10.1007/s11135-022-01611-6

[8] Cai, J., Hu, W., & Yang, Y. (2023, December 27). *Healthy life expectancy for 202 countries up to 2030: Projections with a Bayesian model ensemble.* Journal of global health. https://pmc.ncbi.nlm.nih.gov/articles/PMC10750449/

[9] Kontis, V., Bennett, J. E., Mathers, C. D., Li, G., Foreman, K., & Ezzati, M. (2017). Future life expectancy in 35 industrialised countries: projections with a Bayesian model ensemble. *The Lancet*, *389*(10076), 1323–1335. https://doi.org/10.1016/s0140-6736(16)32381-9

[10] KumarRajarshi. (2018, February 10). *Life expectancy (WHO).* Kaggle. https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

[11] Lasha. (2023, March 30). *Life expectancy (WHO) fixed.* Kaggle. https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated

[12] *Adult BMI categories.* Centers for Disease Control and Prevention. (2024a, March 19). https://www.cdc.gov/bmi/adult-calculator/bmi-categories.html