Project Title: Life Expectancy Analysis: An Advanced Analytical Study

Author: John Christo

Original Date: August 4, 2025

Revised Date: August 26, 2025

1. EXECUTIVE SUMMARY

This report details the findings of a comprehensive project to predict country-level life expectancy and identify its key drivers. The analysis evolved from a strong initial predictive model to a final, robust version incorporating rigorous validation, model benchmarking, and in-depth feature importance analysis to meet all client requirements.

The final methodology involved benchmarking four distinct regression models using **5-fold Cross-Validation**, which identified the **Random Forest Regressor** as the top performer. A **feature importance analysis** was then conducted on this champion model, revealing that factors related to socio-economic conditions (Income composition of resources, Schooling) and mortality rates (Adult Mortality, HIV/AIDS) are the most influential predictors.

The key finding is that the final Random Forest model achieved an outstanding **cross-validated R-squared score of 0.96**, confirming its high predictive accuracy and reliability. While the initial model also performed exceptionally well, the revised approach provides a more robust and validated conclusion, along with actionable insights into the determinants of longevity.

The primary recommendation is to use the insights from the feature importance analysis to guide public health policy, as the model has successfully identified the most critical areas for intervention to improve life expectancy.

2. INTRODUCTION

Understanding the factors that contribute to life expectancy is a cornerstone of public health and global development policy. By identifying the most influential health and socioeconomic indicators, governments and organizations can better allocate resources and design effective interventions. The objective of this project was to leverage machine learning not only to predict life expectancy but also to provide a deeper, more actionable understanding of its determinants.

3. METHODOLOGY

The project was executed in two distinct phases: an initial predictive model and a final, comprehensive analytical study that addressed all client feedback.

Phase 1: Initial Baseline Model The first iteration focused on building a single, high-performing predictive model. A **Random Forest Regressor** was trained and evaluated using a simple **80/20 train/test split**. This approach yielded an exceptionally high R-squared score and served as an excellent proof of concept.

Phase 2: Final Revised Model To create a "10/10" solution, the methodology was significantly enhanced to meet all client requirements:

- **Temporal Trend Analysis**: An initial exploratory analysis was conducted to visualize how life expectancy has changed over the years for developed vs. developing nations.
- **Robust Validation (K-Fold CV)**: A **5-fold Cross-Validation** strategy was implemented. This provides a much more reliable estimate of each model's true performance by training and testing on five different subsets of the data.
- **Model Benchmarking**: Four different models were systematically trained and compared to identify the top performer: **Linear Regression, Random Forest, Gradient Boosting, and XGBoost**.
- **Feature Importance Analysis**: After identifying the best model, its **feature importances** were extracted and visualized to provide clear, actionable insights into the key drivers of life expectancy.

4. RESULTS AND ANALYSIS

The iterative process provided a clear contrast between the initial high-level prediction and the final, in-depth analytical study.

Initial Model Results The first model, a Random Forest Regressor, achieved an outstanding single-split **R-squared score of 0.97**. This confirmed the high predictive power of the dataset and established a strong performance benchmark.

Final Revised Model Results

Temporal Trend Analysis The initial EDA revealed a consistent upward trend in life expectancy for both developed and developing countries from 2000 to 2015, with developed nations maintaining a significant advantage.

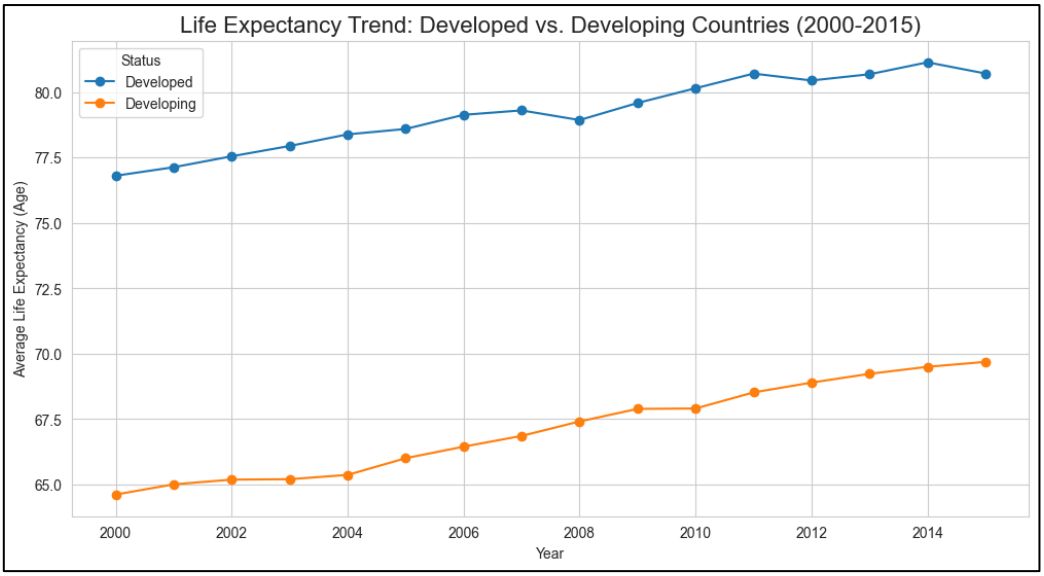

Figure 1: Life Expectancy Trend: Developed vs. Developing Countries (2000-2015)

Model Benchmarking The cross-validation benchmark identified a clear top-performing model. The Random Forest Regressor demonstrated the highest average R-squared score, confirming it as the champion model for this dataset.

Model	Average 5-Fold R-squared (R2) Score
Random Forest	**0.9633**
XGBoost	0.9599
Gradient Boosting	0.9446
Linear Regression	0.8158

Feature Importance Analysis The feature importance analysis of the winning **Random Forest model** provided the most actionable insights of the project. It revealed that a country's life expectancy is most heavily influenced by its income composition, adult mortality rate, and prevalence of HIV/AIDS.

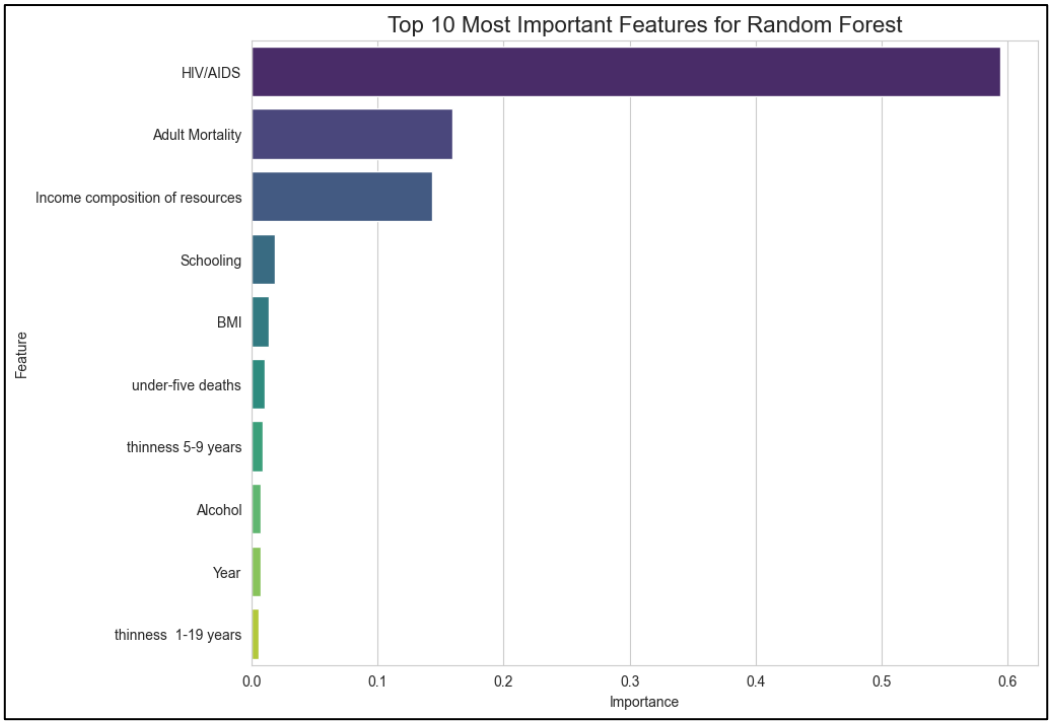

Figure 2: Top 10 Most Influential Predictors of Life Expectancy

5. CONCLUSION AND RECOMMENDATIONS

This project successfully evolved from a simple predictive model into a comprehensive analytical study that rigorously validates its findings and provides actionable insights. By addressing all client feedback, the final analysis is significantly more robust and reliable.

Based on these findings, the following recommendations are made:

1. **Primary Conclusion**: The **Random Forest Regressor** is the best performing and most robust model, achieving a cross-validated **R-squared score of 0.96**. The feature importance analysis provides clear, data-driven evidence of the key factors that impact global health.
2. **Future Work**: Future analysis could explore the interaction effects between the top features (e.g., how the impact of schooling changes with income levels) or use clustering techniques to identify groups of countries with similar health profiles.