Forecasting Life Expectancy Through Machine Learning Models

PROJECT REPORT
JULY 2021

Life Expectancy Estimator (WEB APP)

Renzo Bejarano Varela | Arron Nguyen | Callum Linnegan | Carmen Jiaman Li | Tealiie Le | Nicklaus Ng

CONTENTS:

1. INTRODUCTION

A key metric for assessing a country's population health is by its life expectancy value. This value is derived from focusing on a broader set of conditions that captures different causes of mortality across the course of a lifetime. The end result is a figure based on the average age of death in each country's population.
Estimates suggest that the average life expectancy in the pre-modern era was 30 years old. As the onset of industrialisation improves the lives of people across industrialised countries, it remains low in the rest of the world.
In 2019 the country with the lowest life expectancy was the Central African Republic at 53 years. In contrast, the top of the list was Palau with a high life expectancy of 97 years. This highlights the growing divides in-terms of life expectancy between developed and developing countries.

1.1 CONTEXT

One of the most used measures of a population's overall health is through its life expectancy. By examining causes, patterns and trends in death, it is possible to help explain differences and changes in the health of a population. Therefore, life expectancy data is important as it contributes to the evaluation of health strategies, planning and policymaking.

1.2 OBJECTIVE

Using the group's diverse field of specialisations, we sought to collate a more rounded dataset which considers 19 social, economic, mortality and health-related indicators in which we considered to be essential for the calculation of life expectancy. Furthermore, using this updated dataset, we will create a machine learning model to forecast a more accurate prediction of life expectancy based on the indicators.

2. EXTRACTION, TRANSFORMATION, LOADING (ELT) PROCESS:

2.1 EXTRACTION

The group originally found a dataset on Kaggle that was expected to consist of a multitude of data regarding factors affecting life expectancy. However this data only considered a handful of demographic variables, income composition and mortality rates which was not sufficient for model predictions. To supplement the deficiency within the data, the group sought to create a new dataset. We found a wide-range of data on social, economic, mortality and health-related indicators from the World Bank website.
As the data consisted of hundreds of indicators, it was filtered down by identifying the top 19 key indicators most applicable for predicting life expectancy, as well as the range of years to analyse from 1999 to 2019. This data was then downloaded into three separate Comma Separated Value (CSV) files.

2.2 TRANSFORMATION

The World Bank dataset was unique as it consisted of the 'Indicators' as rows, and 'Countries' and 'Years' as columns. This made the dataset very long as there was a row for each different year in which data was collected for a particular country. For the purpose of machine learning, the dataset needed to be reshaped so that it could be fed into the machine learning model.

Using Pandas pivot-table function, we changed the shape of the data by putting the 'Indicators' as columns and the 'Countries' as rows. This made the dataset cleaner and easier to work with. Additionally, whilst the dataset was fairly clean, several countries were missing data for some years, as well as data for some of the indicators.
To circumvent the missing data, we filled the 'NaN' values with the latest year data that was available. Moreover, for 'NaN' data that did not have the latest year data available, we took an average of the region and used it as the value.

2.3 LOADING

The finished CSV files were loaded into Amazon Web Services (AWS) S3 buckets. We chose to utilise AWS as it is flexible and allows the data to be called through our used platforms of Python, Pandas, Tableau and JavaScript. Finally, as our data took on the form of a non-relational database, we used a noSQL database for this project.

3. MACHINE LEARNING MODEL /APP:

Using the updated dataset, we created a machine learning model to forecast a more accurate prediction of life expectancy based on these indicators.

As the data consisted of over a thousand indicators, we used Linear Regression as an exploratory starting point to analyse the original data (1960-2019). The idea was to infer causal relationships between the independent variable of 'life expectancy' and 21 mostly economic dependable variables.

We then filtered it further by identifying 19 more rounded (Social, Economic, Mortality and Health-Related) indicators which we felt were most correlated with life expectancy. Moreover, for the purpose of our analysis, we narrowed our data for the time period of 1999 - 2019.

Finally, we built and deployed an interactive Gradient Boosted Regressor (GBR) machine learning model to predict life expectancy based on our 19 key indicators, during the period of 1999-2019.

3.1 Gradient Boosted Regressor (GBR)

The GBR is a machine learning model that produces a prediction model in the form of an ensemble of weak learner models stacked to make a more powerful Learning Boosting tool. GBR is based on decision trees. The key idea is to set the target outcomes for each proceeding model to minimize the error outputs and errors in prediction.
With GBR, we can also change the various parameters that controls the model. We can also play with various combinations and select the best ones that yield a good cross-validation. This however requires good understanding of parameters and implications and can take more time to run.

3.2 Linear Regression

We used linear regression to predict life expectancy (the continuous variable) against a number of independent variables (development indicators).
Linear regression takes on the form y i = β0 + β 1X 1 + ... + β iX i.
y is the dependent variable (Life Expectancy) which we will predict.
X is the regressor (development indicator)
B_0 is the intercept, and βi as the regression coefficients.
Linear regression models assume that the relationship between the dependent variable (y) and the regressors (X) are linear.
The model fits the data in ways in which errors are identified and minimised. For this case, we use the least squares method.
Our analysis observes a few model variations:
a. Linear regression on all 21 predictors of life expectancy.
b. Multivariate regression with a few select indicators.
c. Regression with a single predictor.

3.3 Web Application

We created an app using Streamline. This app applies the Machine Learning algorithms to the dataset and estimates the life expectancy for a chosen set of variables. The data variables used for these calculations are based on the nineteen hand-picked features obtained from World Bank Open Data.

The GBR is designed for economists, policymakers and governments to estimate life expectancy based on the 19 World Development Indicators and improve the wellbeing of their citizens. The purpose of the app is to determine the most important factors affecting life expectancy and highlights focus points for individual countries.

4. LIMITATIONS:

4.1 Data only covers up to 2019

One of the limitations of the dataset is that the data is only available up until 2019. As soon as the data is updated, we would be able to do some analysis in regards to COVID-19 and its effects on each regions' social, economic, mortality and health related indicators and their overall life expectancy.

4.2 Insufficient time

As we only had a 2 week timeframe, time became a limitation. There was more than enough data available for us to engage in more detailed analysis. Furthermore, if we had more time, we would have liked to have run more features into more machine learning models. That would have given us a more accurate prediction of life expectancy.

4.3 Dataset was not recorded for in-depth gendered analysis

Despite the multitude of data available on the features contributing to life expectancy, the dataset that shows the differences in women and men was only measured based on overall life expectancy. Therefore in-depth analysis on gender differences on other features such as alcohol consumption, suidcide and accidents rate was not possible. Our analysis only shows that women live longer than men but cannot explain why and in which area due to the limitations in the dataset.

4.4 Expert, non-expert, expertise

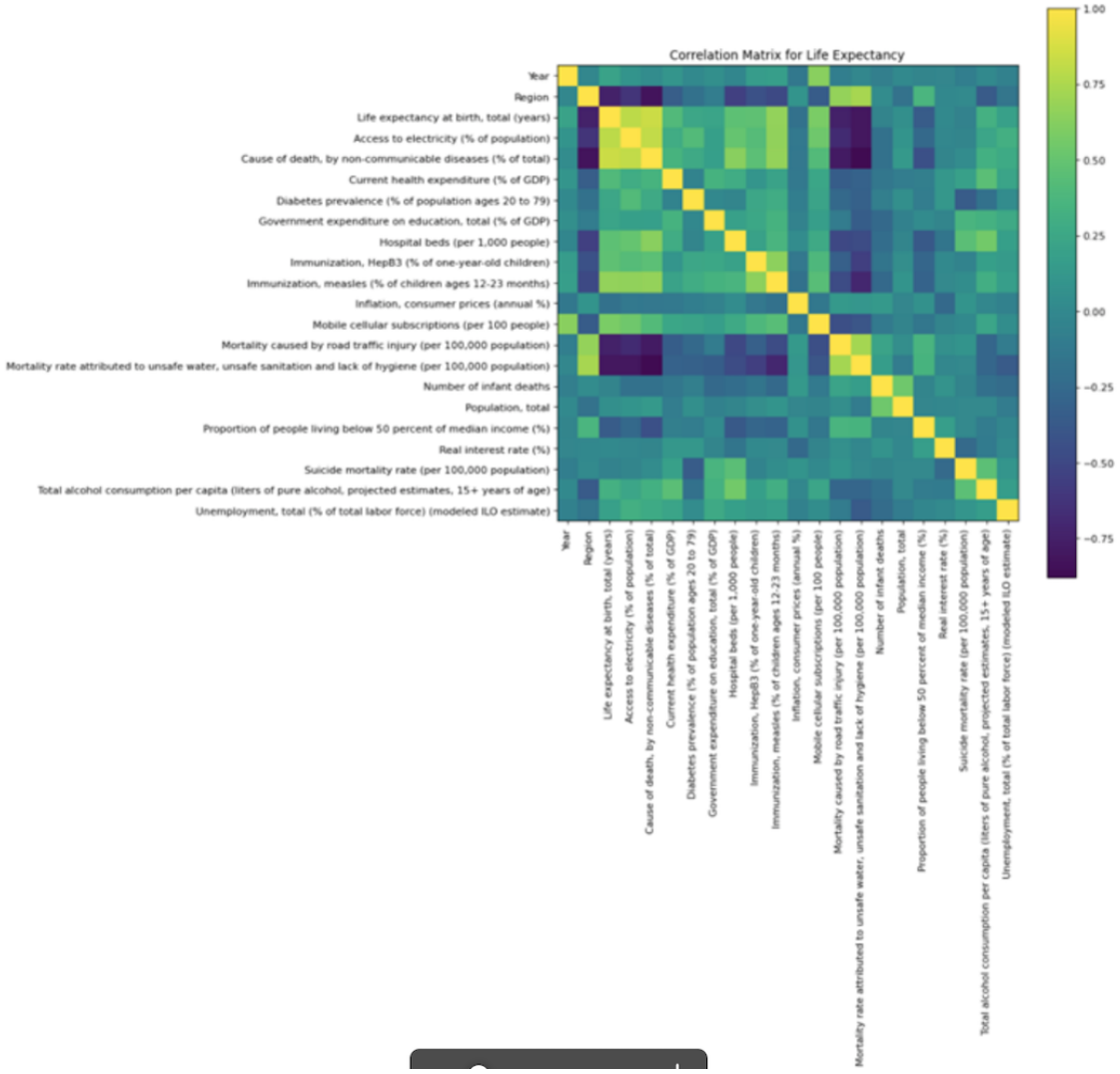
Although we picked data that our grid analysis showed were correlated to life expectancy, we are cautious with the fact that none of us are experts in the field of the analysis of life expectancy. Therefore, we would advise that our findings be considered along with our expert, non-expert, expertise in mind.

4.5 Lower population sizes can distort life expectancy results

Our dataset shows that over the past two decades Palau and Dominica have consistently been ranked as the countries with overall highest life expectancy, at 97 years and 92 years respectively. However we have to take into account that due to these countries having much smaller population sizes, the result might not be the most representative.

5. THE SOCIAL, ECONOMIC, MORTALITY AND HEALTH-RELATED INDICATORS:

By creating a correlation matrix, we are able to show the levels of correlation between all indicators for life expectancy. The lighter the colour is between indicators, the more correlated it is to the corresponding indicator. For example, 'Current Health Expenditure' has around a 0.6 point correlation to 'Life Expectancy at Birth'.



5.1 Population, total

Drastic increases in population impacts the populations' access to resources. Especially in developing countries, large populations place pressures on a country's economy and its government's ability to deliver vital infrastructures and services needed for higher life expectancy. Generally, with some exceptions, a country with a smaller population should expect a higher life expectancy in contrast to larger populations.

5.2 Proportion of people living below 50 percent of median income (%)

The percentage of people in the population who live in households whose per capita income or consumption is below half of the median income or consumption per capita.
The higher the proportion of the population living under 50% of the median income, indicates a possibility of an uneven distribution of income within the population. A majority of the population living under the median income may highlight a significant proportion of the population's inability to afford basic necessities such as food, water and essential medical supplies, which are essential for a higher standard of living.

5.3 Government expenditure on education, total (% of GDP)

Government expenditure has an indirect effect on health levels. A higher Government expenditure on education was linked to higher completion rates and lower levels of poverty.
This generally translates into more affluent and healthier communities, due to their access to health care services.

5.4 Hospital beds (per 1,000 people)

The amount of available hospital beds serves as a measure for a population's access to inpatient services. It should only be used as a general indication, as inpatient services required for individual countries depend on several factors - such as demographic issues and the burden of diseases. Therefore, 2 beds per 1000 people may be sufficient for one country, however the same figure may be inadequate for another country.

5.5 Access to electricity (% of population)

Energy is necessary for creating the conditions for economic growth. Access to electricity is particularly crucial to human development as electricity is indispensable for certain basic activities, such as lighting, refrigeration, and the running of medical appliances. Energy is important for improving a population's standard of living. But electricity generation also can be detrimental to the environment and people's health. The damage depends largely on how electricity is generated. For example, burning coal releases twice as much carbon dioxide - a major contributor to global warming and pollution.

5.6 Inflation, consumer prices (annual %)

Inflation measures the increase in prices of goods and services. As inflation affects the cost of foods such as grains, eggs, fish, and meat. It becomes an important indicator for the ability of the population to afford nutritious food needed for a healthy balanced diet.
Access to food, especially those needed for a healthy balanced diet will inevitably lead to a healthier lifestyle and a greater life expectancy.

5.7 Mobile cellular subscriptions (per 100 people)

Access to mobile cellular subscriptions includes both prepaid and postpaid phone services. Cellular subscriptions are used as an indicator for development and a nation's quality of economic infrastructure. Communication is seen as an important tool for both foreign and domestic investors. Therefore, a higher level of Mobile cellular subscription, should therefore translate into a more developed economy.

5.8 Real interest rate (%)

Real interest rate is the lending interest rate adjusted for inflation. Whilst a higher interest rate might indicate that disposable income is readily available for a large proportion of the population. A high interest rate could also indicate that there is too much money in the market. (1930's Germany). However, a decrease in interest rate indicates a contraction within the economy. This would affect the populations' ability to purchase essential goods and services associated with higher life expectancy.

5.9 Unemployment, total (% of total labour force) (modelled ILO estimate)

Unemployment refers to the share of the labour force that is without work but available for and seeking employment. High and sustained unemployment indicates serious inefficiencies in resource allocation. A high and sustained unemployment rate, especially in developing countries, would most likely hinder a large proportion of the population's ability to afford necessities such as food, water, and medical supplies, adversely affecting life expectancy rate.

5.10 Cause of death, by non-communicable diseases (% of total)

Data on cause of death is compiled by the WHO, based mainly on data from the International Statistical Classification of Diseases and Related Health Problems, 10th revision. The data has been carefully analysed to consider incomplete coverage of vital registration and the likely differences in cause of death patterns that would be expected in under-covered and often poorer subpopulations. It also considers cardiovascular diseases, cancer, injuries, and general ill-defined categories.

5.11 Mortality caused by road traffic injury (per 100,000 population)

Road traffic injuries and deaths are a major global health problem. Traffic crashes are currently the leading cause of death for children and young adults in the world. There is a strong correlation between the risk of road traffic death and the income levels of a country. It is disproportionately high among low- and middle-income countries. This could be due to a government's inability to provide safe road infrastructure, or a lack of access for the population to afford safe, modern cars.

5.12 Mortality rate attributed to unsafe water, unsafe sanitation and lack of hygiene (per 100,000 population)

Unsafe drinking water, sanitation and lack of hygiene are important causes of death. Most cases of diarrheal deaths in the world are a result of unsafe water, sanitation, and hygiene. These diarrheal diseases could be easily prevented if adequate sanitation and clean water standards are provided. Therefore, high mortality rates caused by a lack of hygiene is a strong indicator of poverty and low-socioeconomic statuses.

5.13 Number of infant deaths

In 2019 alone, 7.4 million children, youths and adolescents died mostly of preventable or treatable diseases.
Early childhood mortality is synonymous with developing countries due to the lack of access to food, clean water, and sanitation. Furthermore, populations that belong to developing countries experience more diseases due to the lack of vaccines, adequate medical treatment, and facilities.

5.14 Current health expenditure (% of GDP)

Current Health Expenditure describes the share of spending on health in each country relative to the size of its economy. It includes expenditures corresponding to the final consumption of health care goods and services and excludes investment, exports, and intermediate consumption.
A higher health expenditure generally means a more developed economy and a healthier population.

5.15 Diabetes prevalence (% of population ages 20 to 79)

Long-term effects of diabetes include damage to large and small blood vessels, which leads to heart attack, strokes, and problems with vital organs. The prevalence of diabetes within a population is an indication of the general health within the population. A high prevalence in diabetes might see a decrease in life expectancy, especially within less developed countries, with populations without proper access to healthcare.

5.16 Immunisation, HepB3 (% of one-year-old children)

The percentage of one-year-olds who have received three doses of hepatitis B vaccine each year. Immunization is an essential component for reducing under-five mortality. Immunization coverage estimates are used to monitor coverage of immunization services and to guide disease eradication and elimination efforts. It is a good indicator of health system performance.

5.17 Immunization, measles (% of children ages 12-23 months)

Immunization against measles is an essential component for reducing under-five mortality. Immunization coverage estimates are used to monitor coverage of immunization services and to guide disease eradication and elimination efforts. It is a good indicator of health system performance. It is estimated that between 2000 – 2018, measles vaccinations prevented 23.2 million deaths. As a highly contagious disease, measles was the cause of 140,000 deaths of mostly children under the age of five in 2018 alone.

5.18 Life expectancy at birth, total (years)

Average number of years a newborn is expected to live if mortality patterns at the time of birth remain constant in the future. High mortality in younger age groups significantly lowers the life expectancy at birth. However, if a person survives a childhood of high mortality, they might live much longer than the expected age.

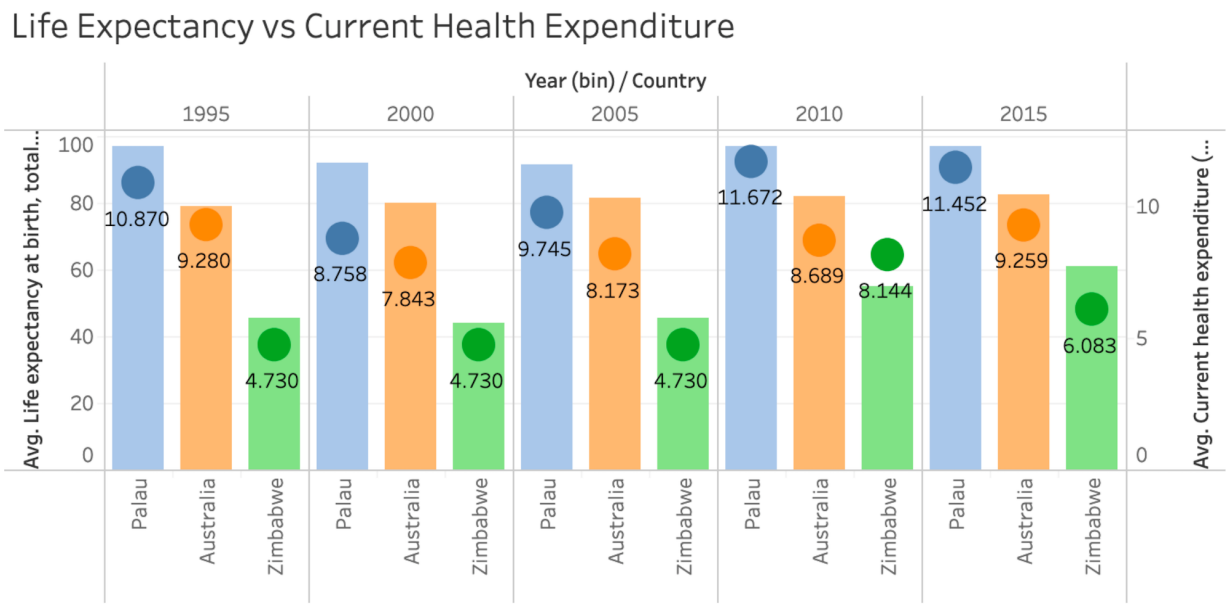
5.19 Total alcohol consumption per capita (liters of pure alcohol, projected estimates, 15+ years of age)

Total alcohol per capita consumption is defined as the total amount of pure alcohol consumed per person over a calendar year. Alcohol consumption is a causal factor of more than 200 diseases and injury conditions such as alcoholism, cancers, cardiovascular diseases as well as injuries from violence and road accidents. Higher levels of alcohol consumption might indicate adverse societal issues and higher rates of alcohol related illnesses, which would impact life expectancy within a region.

6. FINDINGS:

6.1 Wealth & health matters

There is a strong positive relationship between GDP per capita and the length of a population's life expectancy. The wealthier the country, the better access the population has to goods and services. Moreover, wealth allows nations to provide access to healthcare infrastructure. However, bear in mind that the measurement of GDP per capita is strongly affected by the population size, and ignores individuals' actual purchasing power. We also found that safe sanitation improved people's life expectancy. By analysing the data by region, the region with the lowest life-expectancy, Sub-Saharan Africa had the highest mortality rate caused by unsafe sanitation. This was 30% more likely as compared with the rest of the world.



6.2 A small population goes a long way

We found that there was a positive correlation between a small population and a higher standard of living, on the one condition, that the country is developed and wealthy. We believe that a smaller population makes it more manageable for governments to deliver health and social policies as well as the provision of health infrastructure and support for the population. Additionally, it also means that a greater proportion of money can be allocated per person.