Life Expectancy Prediction Using Various Regression Models With Immunization and HDI As Focus Area

Sampreeth Nadig
Dept. of Computer Science Engineering
PES University
Bengaluru, India

Shruti M Karande
Dept. of Computer Science Engineering
PES University
Bengaluru, India

Shamith V S
Dept. of Computer Science Engineering
PES University
Bengaluru, India

***Abstract*—Life expectancy is one of the most used summary indicators for the overall health of a population. Our project aims to predict life expectancy based on several regression models and compare the models to find out which model is the best. We will also find out which factors have the most significant effect on life expectancy.**

I. Introduction

Life Expectancy is the average age that the members of a particular population group will be when they die. Life expectancy varies with developed and developing countries, ratio of birth to death, mortality rates of different countries and ratio of literate to illiterate population, all affect the survival time in one way or the other . The country's growth, advancements and accessibility of resources all are the factors of affect living rate of population. The life expectancy is calculated as the average survival time which indicates the median age of population where some might live till then, some might live more time span, some might live less but on an average the predicted value is the lifetime of that continent .The Human Development Index is a statistic composite index of life expectancy, education, and per capita income indicators, which is used to rank countries into four tiers of human development. Our aim is to calculate the effect of various factors on life expectancy and find out which factors contribute the most to increasing/decreasing life expectancy and predict life expectancy.It will help in suggesting a country which area should be given importance in order to efficiently improve the life expectancy of its population.

II. Previous Works

A. Early Works

Auster, Leveson, and Sarachek (1969) were the first researchers to study a population production function for health: a regression of state-level mortality rates on medical care and environmental variables. Since Auster, Leveson, and Sarachek, several economic studies have attempted to answer pertinent questions regarding life expectancy using data from the United States or multiple countries. While the empirical results were mixed, the general consensus was that population life expectancy (or mortality) is a function of environmental measures (e.g., wealth, education, safety regulation, infrastructure), lifestyle measures (e.g., tobacco or alcohol consumption), and health care consumption measures (e.g., medical or pharmaceutical expenditures). However, the appropriate econometric methodology for disentangling these effects and its meaning for the relative importance(statistical or economic) of the estimated effects has been very contentious.

B. Present Day Research

Frech and Miller(2000) partitioned OECD data into age strata (life expectancy at birth, age 40 years, and age 60 years) and estimated separate life expectancy regressions for each stratum (pooling data for males and females). The determinants in each stratum regression were wealth, some lifestyle variables (alcohol, tobacco, and animal fat consumption), and pharmaceutical and nonpharmaceutical medical expenditures. Few studies have focused on pharmaceutical expenditures as a separate input to life expectancy. These include Peltzman (1987),Babazono and Hillman (1994) and Lichtenberg (1996, 1998).C.H [8] et.al proposed a simple linear regression technique with the logit model -transformed survival ratio between the cohort, gender and age combination referents through simulation from the national life table.Palak Agarwal et.al proposed machine learning regression and classification models for predicting life expectancy. They applied multiple linear regression and random forest regression techniques and achieved good results.

C. Shortcomings in Present Day Research

Most of the present day studies on the subject focus on developed countries. Very few studies focus on factors relevant to developing countries like India or the world in general. It was found that affect of immunization and was not taken into account in the past.

III. METHODOLOGY

A. Data Source

We have sourced our dataset from Kaggle which was collected from WHO and United Nations website with the

help of Deeksha Russell and Duan Wang. The dataset consists of 15 years data from 2000-2015 for 193 countries across 21 parameters.

B. Data Preprocessing

The data consisted of lot of missing values since data could not be collected from many countries due to various reasons such as governmental hindrance, small sized countries etc. To handle the missing values, we grouped the data according to the status of each country(Developed/Developing) and imputed the mean value of the variable according to its status. If the missing value belonged to a developing country, then it was imputed with the mean value of that variable for all developing countries. There were no duplicated rows in the database.Next,we decided to detect outliers. We first drew boxplots to visually see the outliers. We considered data points that lie 1.5 times of IQR above Q3 and below Q1 as outliers. We then calculated the number and percent of data that are outliers. We used winsorization technique to deal with the outliers. Winsorizing or winsorization is the transformation of statistics by limiting extreme values in the statistical data to reduce the effect of possibly spurious outliers.

C. Exploratory Data Analysis

After cleaning the data , EDA was done to get a better understanding of the data.

A lineplot was plotted to study how life expectancy has evolved over time in developing and developed countries.

A violinplot was plotted to see the distribution of life expectancy among developed and developing nations.

Scatterplots were plotted to find out the relationship between various parameters like BMI, alcohol consumption etc. and life expectancy

A heatmap was plotted to see how various parameters correlate with each other

Python libraries like matplotlib and seaborn were used to plot the graphs.

D. Building the Model

Firstly, the most important features were extracted using the ANOVA method. We then decided to adopt two methods to predict life expectancy: neural networks and Random Forest Regressor.Random Forest regressor makes use of multiple decision trees to predict the output.

In our model, we used a neural network with four hidden layers and each layer using the relu activation function.

For the Random Forest Regressor, randomizedsearchcv was used to find out the best parameters Both the models were implemented with the help of python libraries like scikit-learn.

E. Evaluation

Mean Squared error and R squared score were used as the evaluation metrics for the model. Mean Squared Error is simply the average of the square of the difference between the original values and the predicted values. In statistics, the coefficient of determination, denoted R2 or r2 , is the proportion of the variation in the dependent variable that is predictable from the independent variable(s).

IV. Results

A. EDA Results

The graphs provided us deep insight about the data and the following conclusions were drawn from the graphs-

1)As infant deaths decrease, life expectancy increases. Hence ,they follow an inverse relationship.

2) The violinplot for developing countries has a longer tail than the plot for developed countries. This can be because the term developing countries encompasses a large number of countries with low to middle income levels.

3) The heatmap shows positive correlation between Life expectancy and polio vaccination coverage, Hepatitis B vaccination coverage and Diphtheria vaccination coverage. This shows that vaccination has gone a long way in improving life expectancy across both developed and developing countries.

4) The heatmap shows negative correlation between life expectancy and adult mortality, measles cases, HIV/AIDS and thinness.

5) Countries with higher Income Composition of Resources(HDI) have higher life expectancy. This shows that boosting HDI goes a long way in increasing life expectancy.

B. Model Results

The neural network model returned an R2 score of 0.91 while the Random Forest Regressor returned an R2 score of 0.95 and mean squared error of 4.42.

C. Interesting Results

Life Expectancy and population seem to follow an inverse relationship. More the population less is the life expectancy.This can either be because smaller countries are indeed more efficient in last mile delivery of healthcare services or because the most populous countries in the world are considered developing(4 out of the top 5 most populous countries are developing) and hence the populatrion is negatively correlated with life expectancy. Prima Facie, the latter seems to be the reason for this correlation.

Alcohol consumption doesn't show any significant effect on life expectancy. This is surprising given that alcohol is generally associated with several diseases like liver cirrohis. More research needs to be done on this topic.

The life expectancy for both developed and developing countries has increased at more or less the same pace in the period under study. This is surprising given that most of the economic growth during the period has come from the developing world. One of the possible reasons for this is that health expenditure hasn't kept pace with growing GDP.

V. Conclusions

As can be seen from the graphs, a nation should focus on bringing down adult and infant mortality and reduce thinness among its population to increase its life expectancy. This means that direct foodgrain distribution to the poor, such as the PDS system in India should be prioritised in order to reduce malnutrition and thinness and increase life expectancy.

Measles vaccination should be prioritised since increase in measles cases brings down life expectancy. In general, vaccination for all diseases have greatly increased life expectancy.

Fighting HIV/AIDS should be of prime importance since it is a major cause of decreased life expectancy.

More research needs to be conducted on how alcohol affects the human body and hence life expectancy.

Random Forest Regressor is a better model to predict life expectancy than neural networks with a higher R squared score.

VI. References

[1] Auster, Richard D., Irving Leveson, and Deborah Sarachek. 1969. The production of health: An exploratory study. Journal of Human Resources 4:411–36.

[2] Frech, H. E., III, and Richard D. Miller, Jr. 1999. The productivity of healthcare and pharmaceuticals: An international comparison. Washington, DC: American Enterprise Institute.

[3] Babazono, Akira, and Alan L. Hillman. 1994. A comparison of international health outcomes and healthcare spending. International Journal of Technology Assessment in Health Care 10:40–53.

[4] Cremer, Helmuth, Jean-Marie Lozachmeur, and Pierre Pestieau. 2004. Social security, retirement age and optimal income taxation. Journal of Public Economics 88:2259–81.

[5] Palak Agarwal, Navisha Shetty, Kavita Jhajharia, Gaurav Aggarwal, Neha V Sharma, Machine Learning for Prognosis of Life Expectancy and Diseases, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-10, August 2019.

[6] Michael B Schultz, Alice E Kane1, Sarah J Mitchell, Age and life expectancy clocks based on machine learning analysis of mouse frailty