FORECASTING OF LIFE EXPECTANCY

by Adithya M

Submission date: 07-Dec-2021 04:40PM (UTC+0530)
Submission ID: 1723289712
File name: Why_Axis_Final_Report.pdf (269.5K)
Word count: 2420
Character count: 12874

FORECASTING OF LIFE EXPECTANCY

Adithya M, Aishwarya Harinivas, Chetan A Gowda, R Sharmila

Department of Computer Science and Engineering, PES University
100 Feet Ring Road, BSK III Stage, Bengaluru-560 085

PES2UG19CS015@pesu.pes.edu
PES2UG19CS019@pesu.pes.edu
PES2UG19CS097@pesu.pes.edu
PES2UG19CS309@pesu.pes.edu

I. Introduction

The statistical measure of the average time an organism is expected to live based on various factors such as Adult Mortality, BMI, GDP, Income, etc. is called life expectancy. Over the past 20 years, life expectancy has increased by more than 6 years, from67 to 73 approximately.

Through this work, we seek to find the various health factors influencing life expectancy in different countries. The dataset consists of health factors for 193 countries and has been obtained from the WHO data repository website. We can see that in the past 15 years, the health sector has had a massive development, which has resulted in an improvement in life expectancy, especially in developing nations. Therefore, for this dataset, we have considered data from the years 2000-2015 for 193 countries. The dataset we are using is taken from the WHO so we can assume that there are no evident errors. The data which couldn't be found (missing) are from lesser-known countries like Cabo, Verde, Vanuatu, etc. As finding all the data for these countries was difficult, these countries' data was excluded from the final dataset.

The importance of life expectancy comes from the fact that there is no better indicator of a country's socio-economic development than having a long and healthy life.

The specific problem we choose to solve is Predicting life expectancy in different countries using the 'Random Forest Regression Model'.

II. Previous Work

We have investigated multiple models from different research papers on the topic of life expectancy to find the best way to build a model.

These models also helped us understand what components we need to mainly consider and make sure never gets omitted.

They also showed us what different methods we can use and how you can combine multiple regression models and what are the things we must be caution with and what are the things we can relax on.

Below we have mentioned a few models which we think are the most relevant along with their authors and a summary about the model: -

1. Factors Explaining Average Life Expectancy, by Maity, Akansha Rhenman, Emelie Sanders, Elijah
   Summary:
   The data sets contain variables for nation population, GNI per capita (PPP), poverty headcount ratio at $1.00, life expectancy at birth for males and females as well as the averages between the two, the expenditure on health per capita, the completion rate of secondary education, hospital beds per 1000, physicians per 1000 individuals as well as the number of and the adequacy of social protection (Social Security). However, to make sure that these things occur, we must first have a full understanding as to what factors affect them and to what level (degree).

   When we consider the wellbeing and health of a population there are many factors and dependencies that are at play but only a small no of these factor must be chosen for the ease and implementation of statistical analysis. By performing a Regression test between these variables which are selected from all the factors that affect health we can see if these variables are correlated and if yes the level of correlation.

2. Analysis of Life Expectancy using various Regression Techniques, by Anshu Pandey, Rita Chhikara
   Summary:
   In this study they examined trends in life

expectancy and provided an analysis through data visualization of how it will change according to the country, income, education, epidemic, infant death, and sexes. Different regression techniques were applied and compared to develop a predictive model.

Around 193 countries data was analyzed through visualization techniques to bring out the relationship between different parameters which have an impact on life expectancy.

Data set includes attributes such as Country, Year, Status, Life Expectancy, Adult Mortality, Infant deaths, Alcohol, Percentage expenditure, GDP, Population, etc.

Following regression techniques were used:

In the study, Life expectancy is the dependent variable, and all the other factors are the independent variable.

- Multiple Linear Regression: A multiple linear regression model involves more than one independent variable to find out dependent variable.
- Polynomial Regression: polynomial regression is used to draw relationships between variables which are having nonlinear relations.
- KNN Regression: The dependent value is predicted by local interpolation of the dependent value associated with the nearest neighbors in the training set.
- Decision Tree Regression: It is a flowchart-like tree Structure and non- parametric supervised learning method.
- Gradient Boosting Regression: It is an ensemble Method combining k learned model with aim of creating an improved composite model.
- R square was used as a performance measure. Higher values indicate that the model is used to explain more about the variability of the response on the data around the data's mean.

3. A research study on the variables affecting Life Expectancy Descriptive and inferential statistics with Excel and R, by Suresh Kumar Karna and Elisa D'Odorico

<u>Summary:</u>

This paper aims to analyze how various factors such as GDP, traffic accidents, mortality rates affect life expectancy. It aims to achieve this by performing descriptive, as well as inferential statistics on the data. The dataset has 14 variables, of which 3 are qualitative, and 11 quantitative. Descriptive statistics analysis to analyze the strength of the relationship between life expectancy and the variables is performed by plotting graphs between life expectancy and one of the quantitative variables. One such example is the plotting of a line graph between life expectancy and health expenditure. From the graph it was inferred that there exists a positive relationship between the two variables. To perform inferential statistics, a multiple linear regression has been used. MLR allows us to obtain more precise insight into how all the variables affect life expectancy and draw conclusions. Using MLR, the values of the variables are estimated, R2 is calculated to check how much variance is explained by the model and the residual error is also calculated. It also helps to estimate p-value and perform F-statistics to decide whether to accept or reject the null hypothesis which states that there is a significant relationship between the variables and life expectancy.

4. Determinants of life expectancy and clustering of provinces to improve life expectancy: an ecological study in Indonesia, by Sekar Ayu Paramita, Chiho Yamazaki & Hiroshi Koyama

<u>Summary:</u>

Regional disparities in life expectancy in Indonesia have been there for a long time and have now become a challenge to public health policy. We can find a valuable alternative by using systematic clustering of provinces for organizing cooperation that aims to increase life expectancy and reduce disparities. Our goal here is to identify the factors affecting life expectancy and find clusters of Indonesian provinces which are allocatable and have matching characteristics. We will also see if there is an alternative method that can be implemented.

We carefully select variables that impact life expectancy and gather 2015 data from Indonesia's Ministry of health which will form the dataset. We then perform structural equation modeling to select domains that needed to work on from these theoretical models. Then from the results we get from the SEM, we perform cluster analysis to arrange cooperation groups.

III. Proposed Solution

We started the preparation of our model by

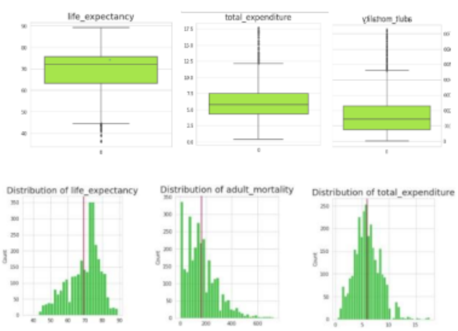
Pre-Processing:

A. Data cleaning:

Initially in the pre-processing phase we wanted to see some insight on the data set we were working on, and on doing that we found out that the number of columns in our dataset was 22, most of attributes in our dataset are skewed few of them are similarly distributed and few of them even had multiple values which were 0 and few had a constant length (of 1).

B. Exploratory data analysis.

We proceeded by performing exploratory data analysis (EDA), which is a process to be followed to analyze data sets to summarize the main characteristics of the data, by using graph and other data visualization methods. For detecting outliers, we decided to plot all the attributes on histograms and box plots. We also plotted the mean in the histograms to check the distribution of the data, that is if the data is symmetric or skewed.



Then comes the non-graphical EDA. This preliminary data analysis step had four main mechanisms that we must examine.

These include:

1. The measures of central tendency: Median, Mean,and Mode.
2. Measures of Spread: Variance, Variability, and standard deviation.
3. Shape of the distribution
4. The existence of outliers.

Now, we perform Feature engineering. The process of obtaining raw data using domain language is called Feature engineering.
This here will be useful to impute mean values with the mean of the column they belong to.
We then perform Label Encoding. The process of obtaining machine-readable form by converting labels into numerical (number) form is called Label Encoding.

Finally, we perform feature selection to reduce the number of input features when creating our model. The below graph shows the importance of the various features in our dataset.

C. Model Building and evaluation.

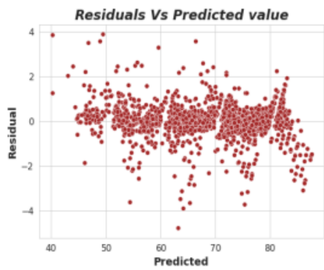
Now we move onto training of the model:
After evaluating the accuracy of different models such as Adaboost Regressor, gradient boosting regressor, Random Forest Regressor, etc. we decided to use the Random Forest Regressor as our model since it had the highest accuracy among all the models.

We end it by checking if the model we created is accurate or simply put, is a good model that represents our data and how it will work in the future for predictions. This part is called Model Evaluation. A model can be evaluated by multiple methods like mean squared error, mean root squares error, absolute error, mean absolute percentage error, etc. Here we will find the test models root mean square error, accuracy, and the mean absolute percentage error (MAPE).
The Root mean square error (RMSE) is the standard deviation of the residuals (observed subtracted by estimated values), it tells you how densely populated the data is around line of best fit.

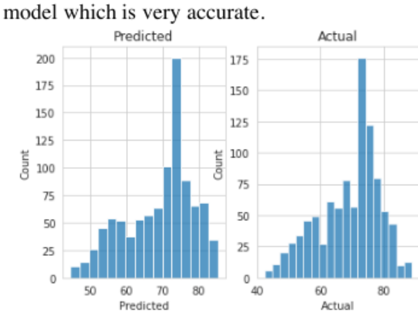
IV. Experimental Result and Explanation

In our model, we get the Test root mean squared error as 1.9, and the closer that value is to 0 the better. This is because zero specifies that all the values lie on the regression line and therefore there is no error. So, our Model passes the Root mean Squared test.
Next up is the accuracy test. Accuracy measures the quality or state of the model being correct or precise and is usually represented as percentage.
In our model we get a test accuracy of 98.25% (which represents a very high level of precision between the observed and expected output).
The final error finding method used is Mean absolute percentage error, Mean absolute percentage error is the process of measuring prediction accuracy of forecasting method in statistics.
Our result for this test stands at 2% which shows very low error rate.
To finish off we make three visual graphs, one is residual vs predictor plot and the other two are histogram of predicted and actual value.

The residual vs predictor plot is a very well-behaved plot. In this plot we see that the residuals are centered around residual = 0 line, which indicates that the predicated value is very close to the actual value.



The other two histograms are for visual representation purpose of how close the original value is to the values our model predicted. By all this testing we conclude that we have a good model which is very accurate.



V. Conclusions

In conclusion, it is possible to predict human life expectancy in advance by making use of dataset and the correlation between different attributes with the life expectancy. Through this paper we discovered that the Random Forest regressor can predict life expectancy with more accuracy by making use of cross validation. It also reduces the overfitting of datasets and increases precision. Random forest classifier (RFC) is also known for keeping the accuracy of large proportion of data high and it is also known for handling missing value. It is good in handling variance. Training of Random Forest classifier is faster than that of the decision tree as we will be working only on subset of data in this model, so working with hundreds of features will be easy.

Acknowledgment

We would like to express our thanks of gratitude to our teacher Prof. Bharathi Saravanan who gave us her support, guidance, and the opportunity to do this great project on the topic of data analytics, which also helped us in doing a lot of research and explore and learn a lot of new things.

References

[1] Factors Explaining Average Life Expectancy, Maity, Akansha Rhenman, Emelie Sanders, Elijah

[2] Life expectancy across high income countries: retrospective observational study, Jessica Y Ho, Arun S Hendi

[3] Assessing the potential impact of COVID-19 on life expectancy, Guillaume Marois, Raya Muttarak, Sergei Scherbov

[4] Life Expectancy and Mortality Rates in the United States, 1959-2017, Steven H. Woolf, Heidi Schoomaker

[5] Quantifying impacts of the COVID-19 pandemic through life-expectancy losses: a population-level study of 29 countries, Jose´ Manuel Aburto ,Jonas Scholey , Ilya Kashnitsky ,Luyin Zhang ,Charles Rahal ,Trifon I Missov, Melinda C Mills Jennifer B Dowd and Ridhi Kashyap

[6] Analysis of Life Expectancy using various Regression Techniques, Anshu Pandey, Rita Chhikara

[7] A research study on the variables affecting Life Expectancy Descriptive and inferential statistics with Excel and R. *FLEXChip Signal Processor (MC68175/D)*, Motorola, 1996, Suresh Kumar Karna and Elisa D'Odorico

[8] Trends in life expectancy and healthy life years at birth and age 65 in the UK, Claire E.Welsh, Fiona E.Matthews, and CarolJagger.

[9] A. Karnik, "Performance of TCP congestion control with rate feedback: TCP/ABR and rate adaptive TCP/IP," M. Eng. thesis, Indian Institute of Science, Bangalore, India, Jan. 1999, Khalaf Alsalem, Alyx Steinmetz, Nayawiyyah Muhammad, Danielle Frierson and Michael Nashed.

[10] Determinants of life expectancy and clustering of provinces to improve life expectancy: an ecological study in Indonesia, Sekar Ayu Paramita, Chiho Yamazaki & Hiroshi Koyama.

[11] The impact of increasing education levels on rising life expectancy: a decomposition analysis for Italy, Denmark, and the USA, Marc Luy, Marina Zannella, Christian Wegner-Siegmundt, Yuka Minagawa, Wolfgang Lutz & Graziella Caselli.

[12] Global age-sex-specific fertility, mortality, healthy life expectancy (HALE), and population estimates in 204 countries and territories, 1950–2019: a comprehensive demographic analysis for the Global Burden of Disease Study 2019, The Author(s). Published by Elsevier Ltd.

[13] Global Health Observatory (GHO) data repository under World Health Organization (WHO) data from 2000-2015 for 193 countries.

FORECASTING OF LIFE EXPECTANCY

ORIGINALITY REPORT

16%	12%	12%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.kaggle.com Internet Source	2%
2	toc.proceedings.com Internet Source	2%
3	dro.deakin.edu.au Internet Source	1%
4	Anshu Pandey, Rita Chhikara. "Analysis of Life Expectancy using various Regression Techniques", 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020 Publication	1%
5	genus.springeropen.com Internet Source	1%
6	jaisankarg.synthasite.com Internet Source	1%
7	doaj.org Internet Source	1%

8	Submitted to University College London Student Paper	1%
9	en.wikipedia.org Internet Source	1%
10	Submitted to University of York Student Paper	1%
11	Submitted to University of Huddersfield Student Paper	1%
12	pubmed.ncbi.nlm.nih.gov Internet Source	1%
13	fdotwww.blob.core.windows.net Internet Source	1%
14	www.dataquest.io Internet Source	<1%
15	bmcpublichealth.biomedcentral.com Internet Source	<1%
16	Claire E. Welsh, Fiona E. Matthews, Carol Jagger. "Trends in life expectancy and healthy life years at birth and age 65 in the UK, 2008–2016, and other countries of the EU28: An observational cross-sectional study", The Lancet Regional Health - Europe, 2021 Publication	<1%
17	Sekar Ayu Paramita, Chiho Yamazaki, Hiroshi Koyama. "Determinants of life expectancy and	<1%

clustering of provinces to improve life expectancy: an ecological study in Indonesia", BMC Public Health, 2020

Publication

Exclude quotes Off
Exclude bibliography On
Exclude matches Off