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 and background

Life expectancy is a statistical measure of the average time an organism is expected to live based various factors such as Adult Mortality, BMI, GDP, Income etc. Over the past 20 years the life expectancy has increased by more than 6 years, from 67 to 73 approximately.

Through this work we seek to find the various health factors influencing the life expectancy in different countries. The dataset consists of health factors for 193 countries and has been obtained from WHO data repository website. It has been observed that in the past 15 years, there has been a huge development in the health sector, which has resulted in an improvement in life expectancy, especially in the developing nations. Therefore, for this dataset we have considered data from the years 2000-2015 for 193 countries. As the datasets were from WHO, there were no evident errors. The missing data was from less known countries like Vanuatu, Cabo Verde 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 '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
   <u>Summary:</u>
   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, physicians per 1000 individuals as well as the number of hospital beds per 1000, and the adequacy of social protection (Social Security). However, to ensure these things, one must first understand what factors may affect them and to what degree.

   Although a great many factors can be said to affect the health and wellbeing of a population, it is only realistic to cover a comparatively small number of such factors for the sake of statistical analysis. Regressions between these variables show whether the variables are correlated as well as degree of correlation.

2. Analysis of Life Expectancy using various Regression Techniques, by Anshu Pandey, Rita Chhikara
   <u>Summary:</u>
   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 analysed 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 neighbours in the training set.
- Decision Tree Regression: A Decision Tree 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 explains more of the variability of the response data around its 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

Summary:

This paper aims to analyse 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 analyse 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

Summary:

Regional disparities in life expectancy in Indonesia have been there for a long and now have become a public health policy challenge. A systematic clustering of provinces can be a valuable alternative for organizing cooperation that aims to increase life expectancy and reduce disparities. Here we aim to identify determinants of life expectancy and designate clusters of Indonesian provinces with similar 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. We then perform structural equation modelling (SEM) 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

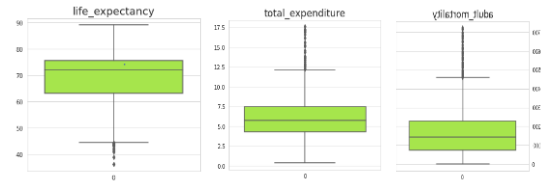
Pre-Processing:

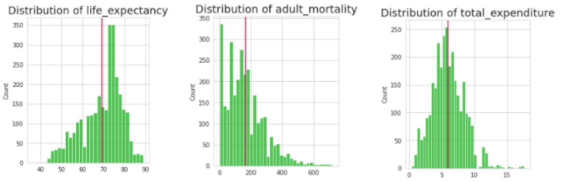
A. Data cleaning:

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 which is an approach of analysing data sets to summarize their main characteristics, 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.

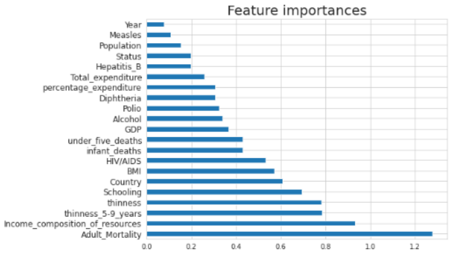
These include:

1. The measure of central tendency (Mean, Median and Mode).
2. Measure of Spread (Variability, variance, and standard deviation).
3. Shape of the distribution
4. The existence of outliers.

Now, we perform Feature engineering which is the process of using the domain language to extract features from raw data. This here will be useful to impute mean values with the mean of the column they belong to.

We then perform Label Encoding which is the process of converting labels into numeric form to convert them into machine-readable form.

Finally, we perform feature selection which is the process of reducing the number of input variables when developing a model.



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.

Root mean square error is the standard deviation of the residuals (observed subtracted by estimated values), it tells you how concentrated 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.

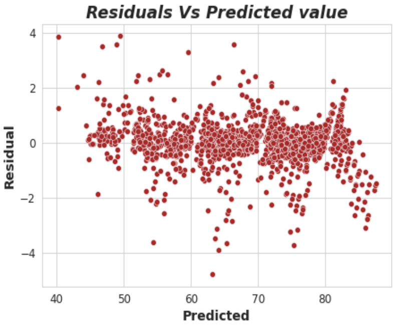
In our model we get a test accuracy of 98.25% (which represents a very high level of precision between the observed and expected output).

The final error finding method used is Mean absolute percentage error, which is a measure of prediction accuracy of forecasting method in statistics.

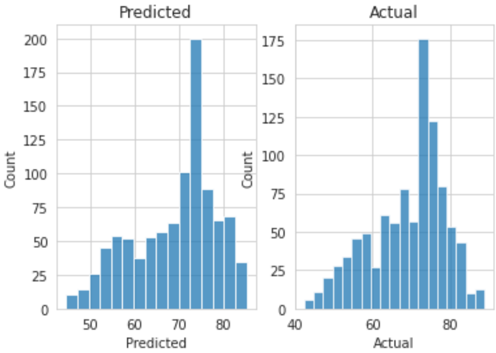
Our result for this test stands at 2% which shows very low error rate.

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 well-behaved plot which will form a roughly horizontal band around the residual = 0 line and no data points will stand out from the basic random pattern of the other residuals.



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 will handle the missing values and maintain the accuracy of a large proportion of data. It is faster to train than decision trees because we are working only on a subset of features in this model, so we can easily work with hundreds of features.

ACKNOWLEDGMENT

I would like to express my special thanks of gratitude to our teacher Prof. Bharathi Saravanan who gave us this golden opportunity to do this wonderful project on the topic of data analytics, which also helped us in doing a lot of research and we have come to know about so many 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.