Remote Summer Internship Program 2020
Machine Learning, Career Basic Program
Smartinternz

Predicting Life Expectancy using Machine learning

Report

By,

Prem Patel

19/5/20-18/6/20

INDEX

1. INTRODUCTION

1.1 Overview

A typical Regression Machine Learning project leverages historical data to predict insights into the future. This problem statement is aimed at predicting the Life Expectancy rate of a country given various features. This problem can be considered as a supervised machine learning problem where previous data can be used for predicting future behaviors and patterns.

1.2 Purpose

Life expectancy is a statistical measure of the average time a human being is expected to live, Life expectancy depends on various factors: Regional variations, Economic Circumstances, Sex Differences, Mental Illnesses, Physical Illnesses, Education, Year of their birth and other demographic factors. This problem statement provides a way to predict the average life expectancy of people living in a country when various factors such as year, GDP, education, alcohol intake of people in the country, expenditure on the healthcare system, and some specific disease-related deaths that happened in the country are given.

2. LITERATURE SURVEY

Mortality rates and measures of life expectancy are widely used by planners and policymakers to compare and monitor health across communities. For example, mortality statistics have demonstrated both a shift in disease patterns to chronic diseases and a recent decrease in cardiovascular deaths in the world. At the same time, the possibility of extending life expectancy is increasingly constrained by the biological limits of the natural life span. Thus, there is a need to add dimensions to population health other than mere survival based on certain different aspects.

Consider a fact, female life expectancy has risen for the last 160 years at a constant 3 month per year rate. In 1840, a record was held by a Swedish woman who lived on average a little more than 45 years, which is very less in comparison to 85 average life in Japan. This 4-decade rise in the last 16 decades is quite extraordinary linear. Due to the current advances in developed countries, the average life expectancy has increased drastically, while in the developing and underdeveloped countries, the expectancy is still low.

Mortality research has exposed the empirical misconceptions and specious theories that underlie the pernicious belief that the expectation of life cannot rise much further. First, experts have repeatedly asserted that life expectancy is approaching a ceiling: these experts have repeatedly been proven wrong. Second, the apparent leveling off of life expectancy in various countries is an artifact of laggards catching up and leaders falling behind. Third, if life expectancy were close to a maximum, then the increase in the record expectation of life should be slowing. It is not. For 160 years, best-performance life expectancy has steadily increased by a quarter of a year per year.

2.1 Existing Problem

Predicting the lifespan of a specific human being can be a really challenging task, hence this problem can be formulated as predicting an average life expectancy of a country based on the external factors like Sanitation, diseases, GDP etc.

2.2 Proposed Solution

The proposed solution uses a linear regression model for finding life expectancy by considering various parameters. The data that is taken into consideration for training the model is taken from the Kaggle website (link is given in bibliography) that contains 19 numerical and 2 categorical parameters. So, the pipeline is created that contains two steps, first is to transform the categorical columns with an imputer and one-hot encoding

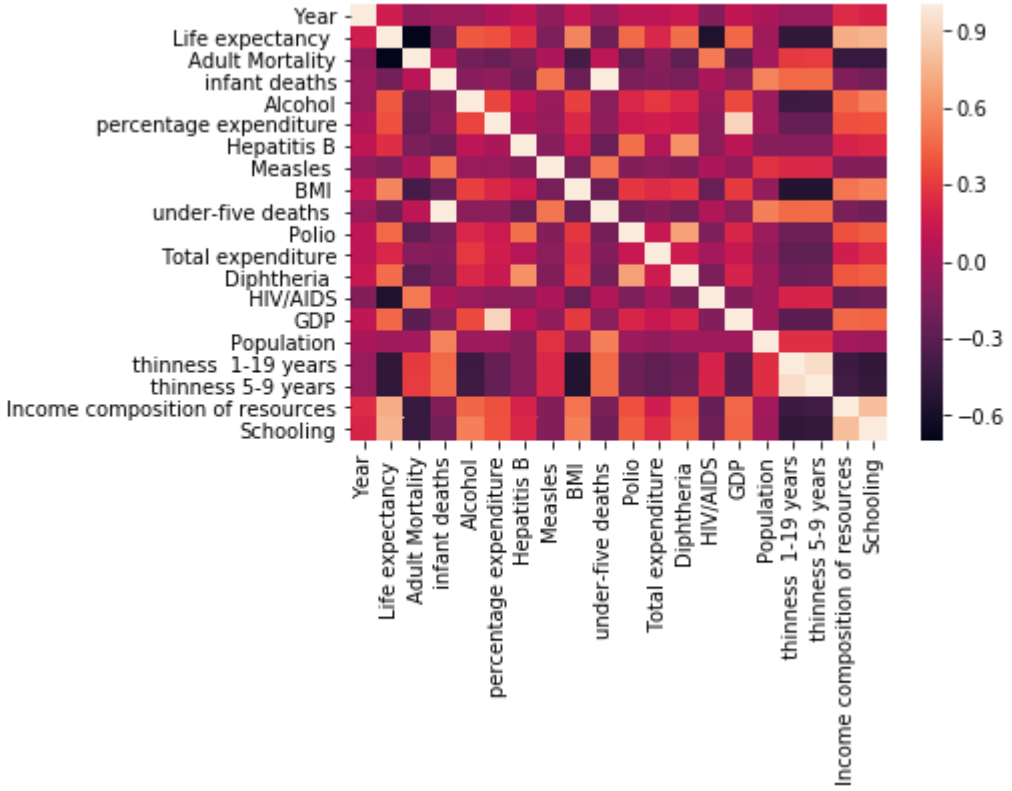
and second is the linear regression.

The user interface contains a form that will take input of parameters from the user and call the machine learning model API with HTTP request and show the output to the user.

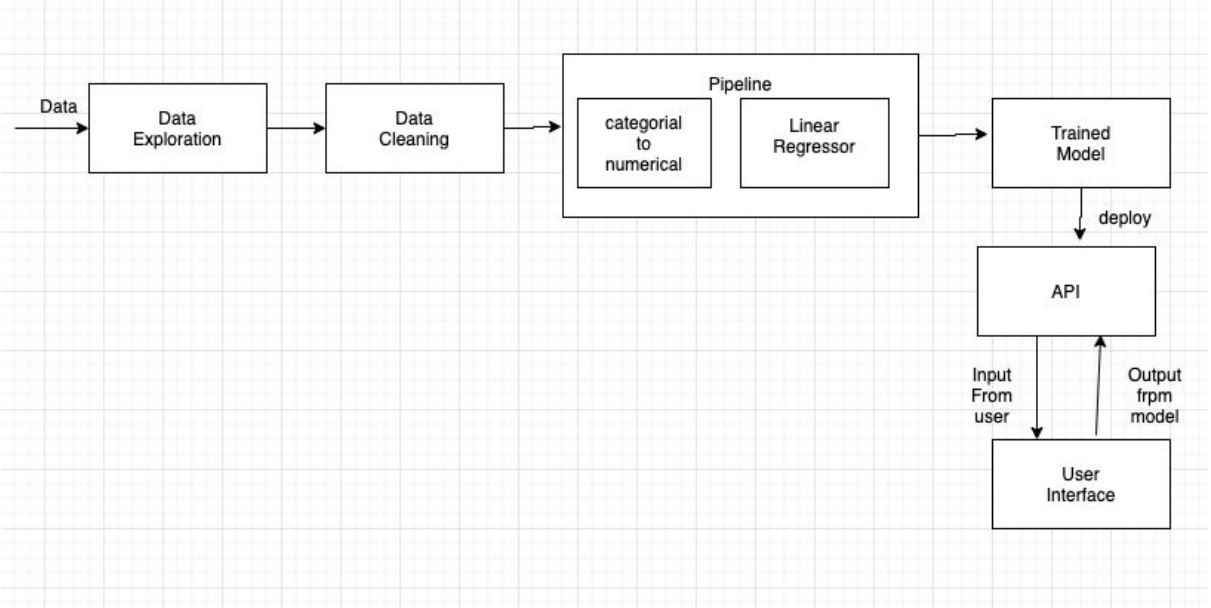
3. ANALYSIS

By the analysis of the data, we can gain various useful insights into the problem and guidance to the appropriate solutions. The following heatmap demonstrates correlations of parameters to each other and with the output. It also describes positive and negative correlations. As we can see Life expectancy is highly dependent on parameters like HIV/AIDS, Adult mortality, thinness, and schooling.

Also, we can see schooling and income composition of resources has a high positive correlation same as GDP-percentage expenditure and polio-infant deaths.

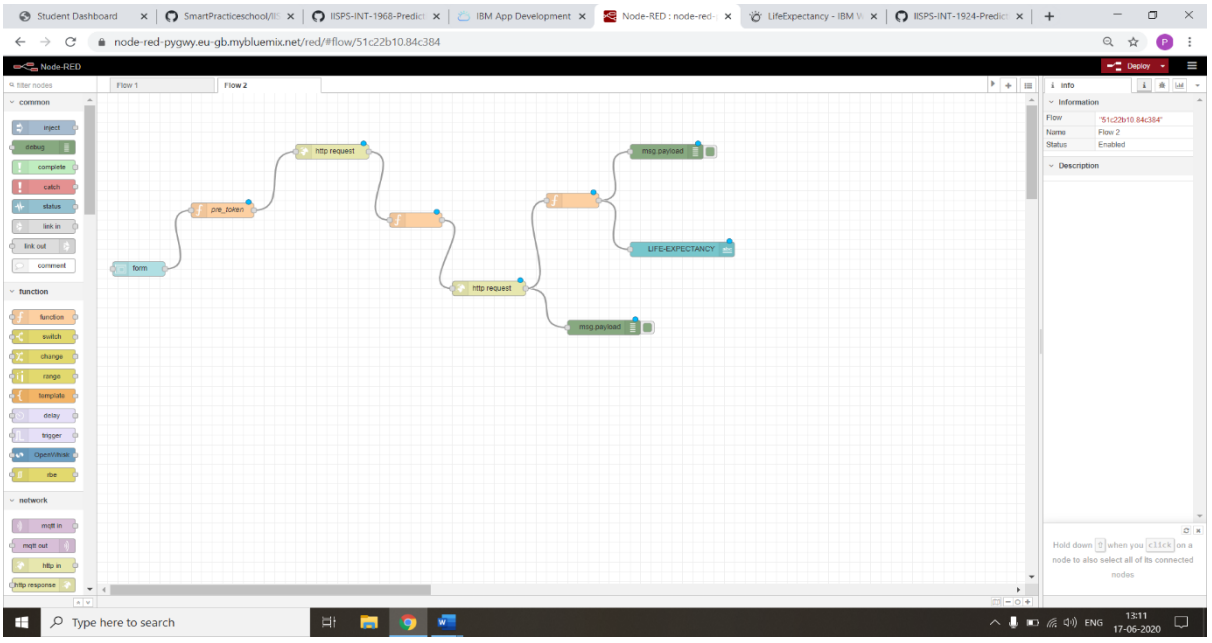


3.1 Block Diagram

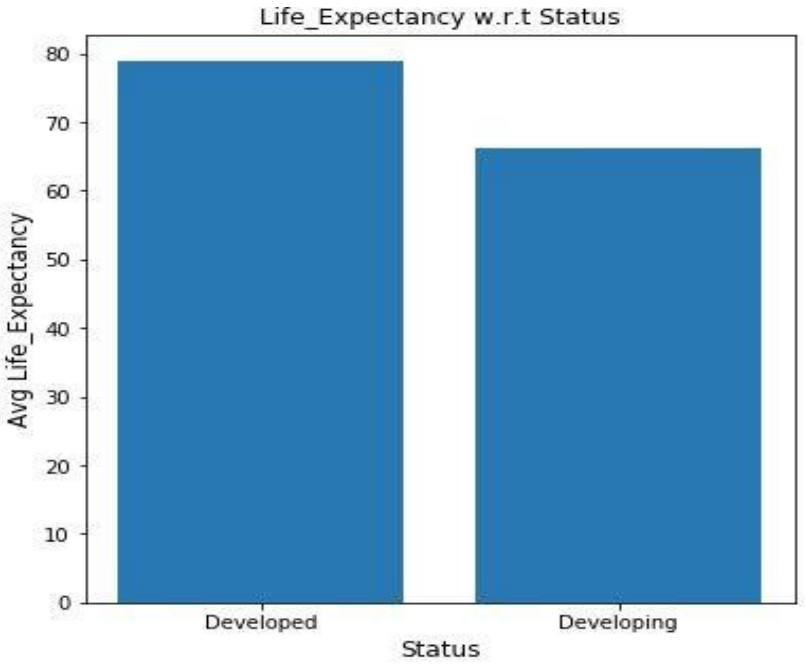


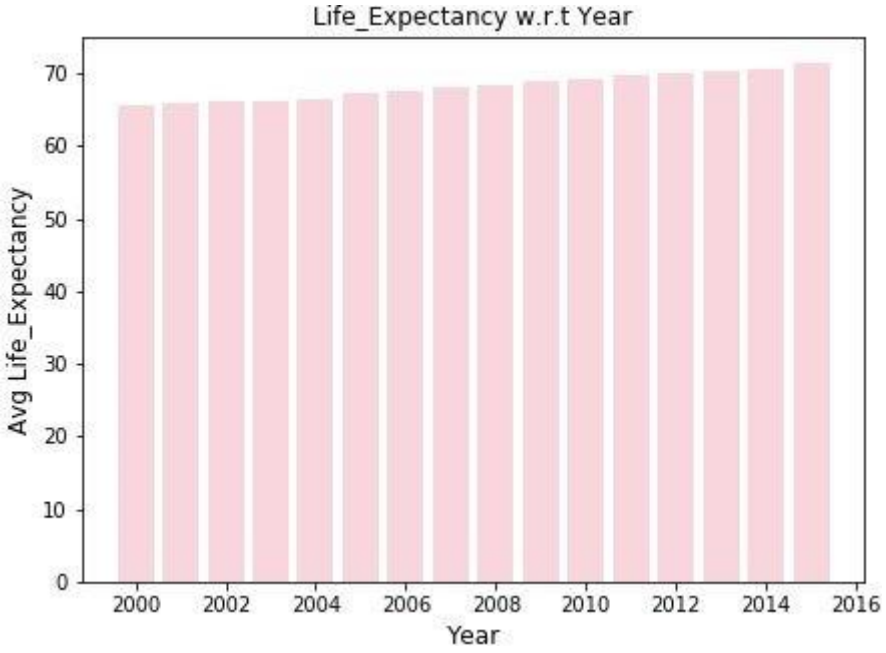
3.2-Hardware/Software Designing

For the implementation of the proposed approach, we have used IBM Watson Studio and IBM Node-red app services. IBM Watson Studio will help us to build backend API service by deploying our machine learning model while the node-red app will give a platform to build a user interface by designing the flow of different nodes given. The following screenshot demonstrates the node-red flow for this app.



4. INVESTIGATIONS



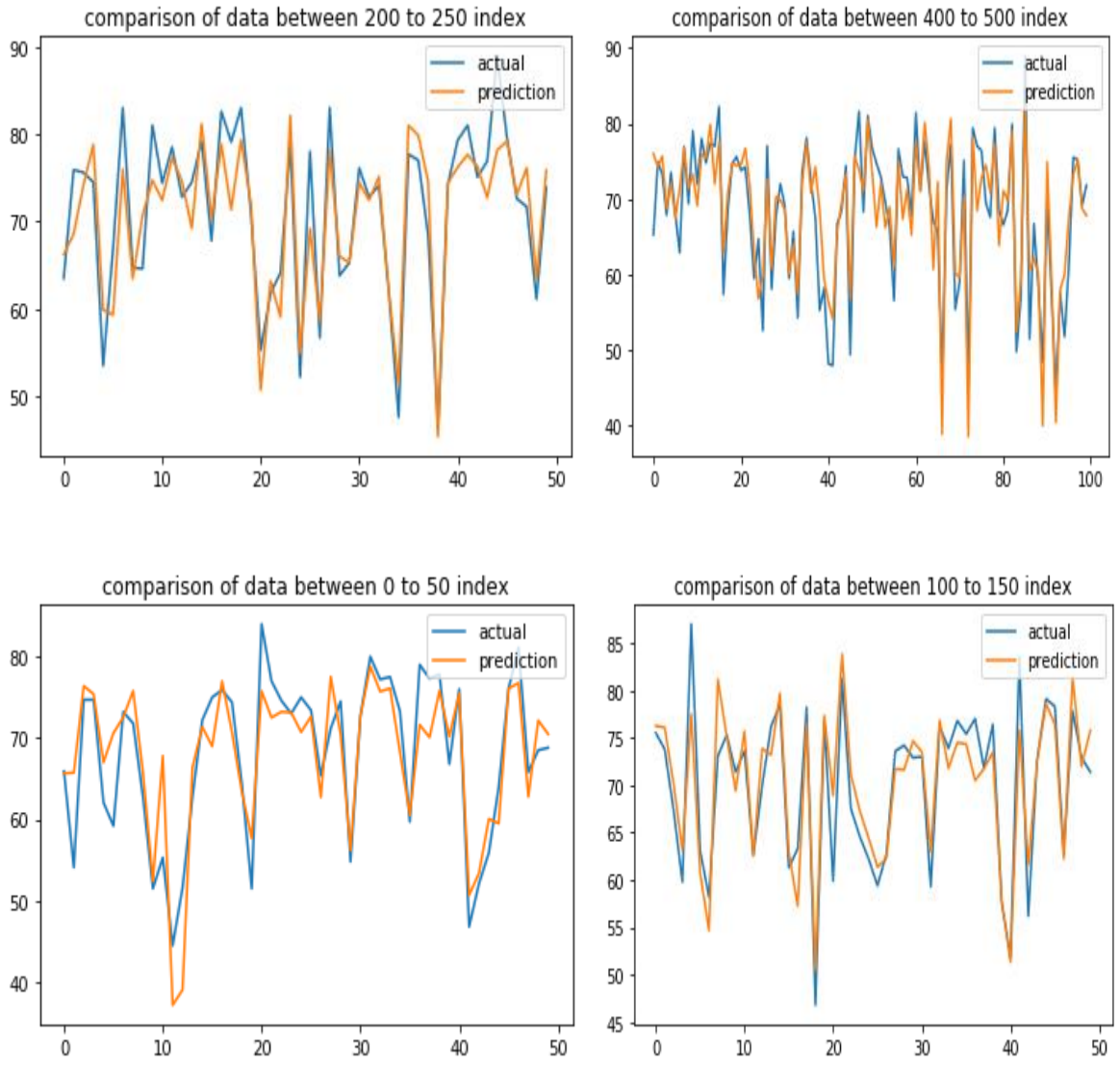


5. RESULTS

The following metrics describe the result of the trained model's efficiency to predict the result.

- Mean absolute error~3.358
- Root mean squared error~4.404
- R2 score~0.779

The following visualizations show the comparisons between test data and actual predicted data at different slices of the data.



6. ADVANTAGES AND DISADVANTAGES

As we have used a linear regression approach, the proposed approach will have its advantages and disadvantages.

Advantages of Linear Regression:

1. Linear Regression performs well when the dataset is linearly separable. We can use it to find the nature of the relationship among the variables.
2. Linear Regression is easier to implement, interpret and very efficient to train.
3. Linear Regression is prone to over-fitting, but it can be easily avoided using some dimensionality reduction techniques, regularization (L1 and L2) techniques and cross-validation.

Disadvantages of Linear Regression:

1. Main limitation of Linear Regression is the assumption of linearity between the dependent variable and the independent variables. In the real world, the data is rarely linearly separable. It assumes that there is a straight-line relationship between the dependent and independent variables which is incorrect many times.
2. Prone to noise and overfitting: If the number of observations are lesser than the number of features, Linear Regression should not be used, otherwise it may lead to overfit because is starts considering noise in this scenario while building the model.
3. Prone to outliers: Linear regression is very sensitive to outliers (anomalies). So, outliers should be analyzed and removed before applying Linear Regression to the dataset.
4. Prone to multicollinearity: Before applying Linear regression, multicollinearity should be removed (using dimensionality reduction techniques) because it assumes that there is no relationship among independent variables.

7. APPLICATIONS

We can predict life expectancy by giving different parameters and thus we can analyze and improve different solutions to improve the statistics and ultimately it will be beneficial to people and the government. By knowing the probable statistics, we can plan and set the path for that country at that time accordingly to serve the people and the country in the best possible way.

We can also make different analyses of the contribution of different parameters to life expectancy. It will give us the path for setting our priorities to the plans of the government so that it will give the country an optimistic path towards the health and wellbeing of people as well as economical and sustainable growth of the country and the world ultimately.

8. CONCLUSION

This project has provided me with the opportunity to implement my data science skill set to predict the life expectancy of a human being based on the factors around which he/she revolves around. I.e. This project has helped me get insights into multiple aspects like GDP, sanitation, etc can have a significant impact on average human life expectancy. This platform provided me the medium where I can use Machine learning techniques for predicting future average life expectancy using several external factors.

9. FUTURE SCOPE

We can improve the efficiency, flexibility, and accuracy of the current approach by following different strategies and logics for feature engineering, training, and visualizations. We can choose a more custom and relatable approach in data cleaning and preparation. We can do different variations to find the best possible approach to the problem. We can consider various algorithms and models like decision tree regression, polynomial regression, and neural networks.

10. BIBLIOGRAPHY

Data:
https://www.kaggle.com/ life-expectancy-who

IBM Watson studio:
https://bookdown.org/caoying4work/watsonstudio-workshop/jn.html

IBM Machine Learning:
https://developer.ibm.com/technologies/machine-learning/series/learning-path-machi ne-learning-for-developers/

IBM Node-red Application:
https://developer.ibm.com/tutorials/how-to-create-a-node-red-starter-application/

11.APPENDIX

GitHub Repository :

https://github.com/SmartPracticeschool/llSPS-INT-1993-Predicting-Life-Expectancy-using-Machine-Learning

Node-red Application:

https://node-red-pygwy.eu-gb.mybluemix.net/ui/#!/0?socketid=va7ZqUh8NvviyHpJAAAG