Project Report

Title:

PREDICTING LIFE EXPECTANCY USING MACHINE LEARNING

In Python

By:

Shouvit Pradhan

Introduction

Problem Description:

A typical Regression Machine Learning project leverages historical data to predict insights into the future. This problem statement is aimed at predicting Life Expectancy rate of a country given various features. 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 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 healthcare system and some specific disease related deaths that happened in the country are given.

About:

The project relies on the accuracy of data. The Global Health Observatory (GHO) data repository under the World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries the data-sets are made available to the public for the purpose of health data analysis. The data-set related to life expectancy, health factors for 193 countries have been collected from the same WHO data repository website and its corresponding economic data was collected from the United Nations website. Among all categories of health-related factors, only those critical factors were chosen which are more representative.

Existing Problem:

This problem statement provides a way to predict 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 healthcare system and some specific disease related deaths that happened in the country are given. 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

My Take

Approach:

Having a limited dataset was a problem for this project as the dataset only had about 2.8k values which are very less compared to the 20+ features available.
To tackle this problem since this was a regression task, correlation values were used to fill the null values in each column. The logic is that instead of filling nan values with random mean or median values why not use the values of other columns with high correlation since all were continuous values.
Taking this approach as all the columns seemed to be correlated in some way or other, all the nan values were filled and there was no loss of data.
The non-numeric values were changed to corresponding dummy values so that the model could learn something from it.
Initially, the model was built using all the features and the accuracy was pretty high. This would've been fine only if the aim of the project was to achieve high accuracy but the model had to be deployed.

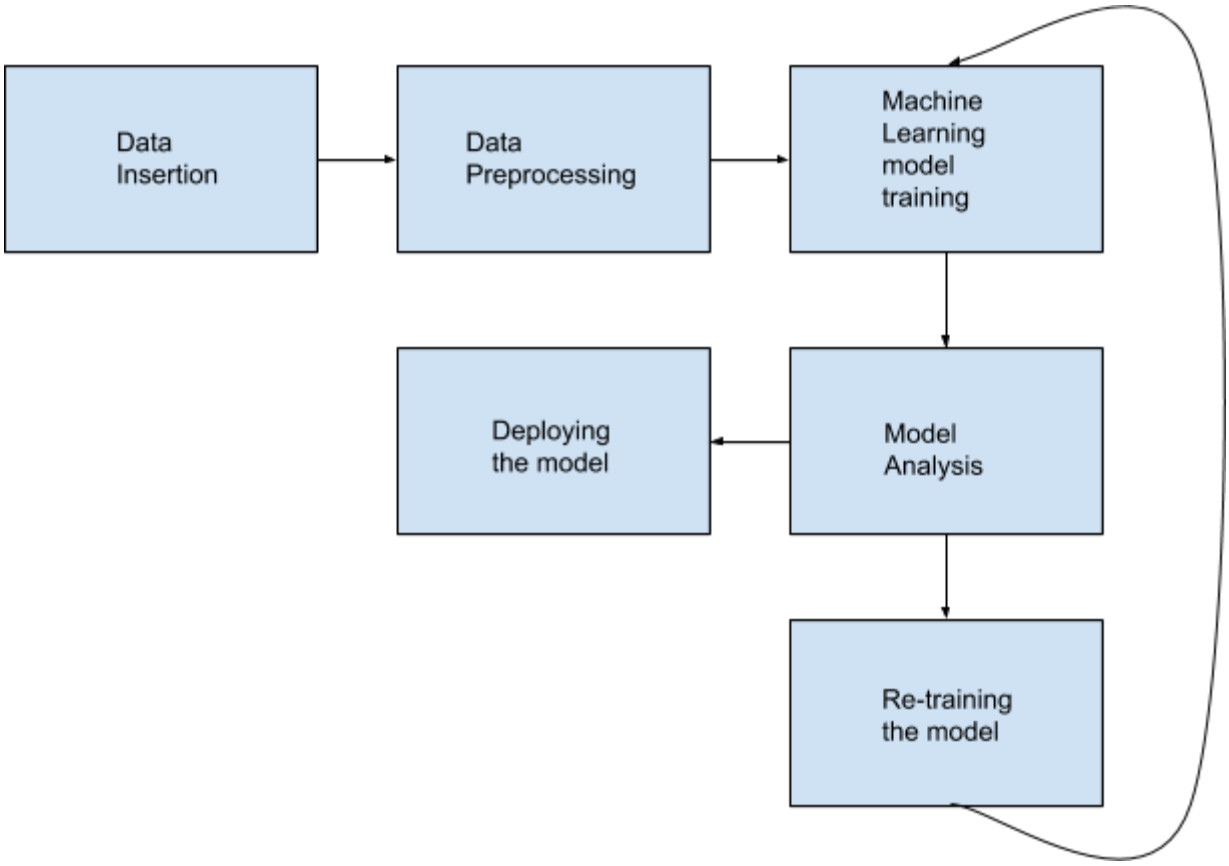
.

Proposed Solution:

Designed a Regression model to predict Life Expectancy on some features such as Adult Mortality, GDP, BMI, expenditure, disease-related deaths, thinness, income, schooling in any country.
The above features were selected based on their correlation to 'Life Expectancy'.
Also since the output can never be 100% correct, the output displayed in the node-red is a range, which is obtained by adding and subtracting the rmse value to the output of the model.

Theoretical Analysis

Block Diagram:



Hardware and Software designing:

- Hardware : Desktop / Laptop, Internet Connectivity
- Software : IBM Cloud, IBM Watson Studio, Node-Red App

Experimental Investigation

1) Choose a Project :

Predicting Life Expectancy of a person based on values that the user inputs.

2) Collection of Dataset :

https://www.kaggle.com/kumarajarshi/life-Expectancy-who

3) Hypothesis :

Based on our study and information gathered we can predict the average age of a person.

4) Design :

Construct various Machine Learning Models and finally selecting the model with maximum accuracy.

5) Conclusion :

Model will be able to predict the life expectancy of a person with maximum accuracy.

Node-Red Flowchart

A flowchart is a diagram that depicts a process, system or computer algorithm. They are widely used in multiple fields to document, study, plan, improve and communicate often complex processes in clear, easy-to-understand diagrams. Flowcharts, sometimes spelled as flow charts, use rectangles, ovals, diamonds and potentially numerous other shapes to define the type of step, along with connecting arrows to define flow and sequence.

Node-RED is a programming tool for wiring together hardware devices, APIs and online services in new and interesting ways. It provides a browser-based editor that makes it easy to wire together flows using the wide range of nodes in the palette that can be deployed to its runtime in a single-click.

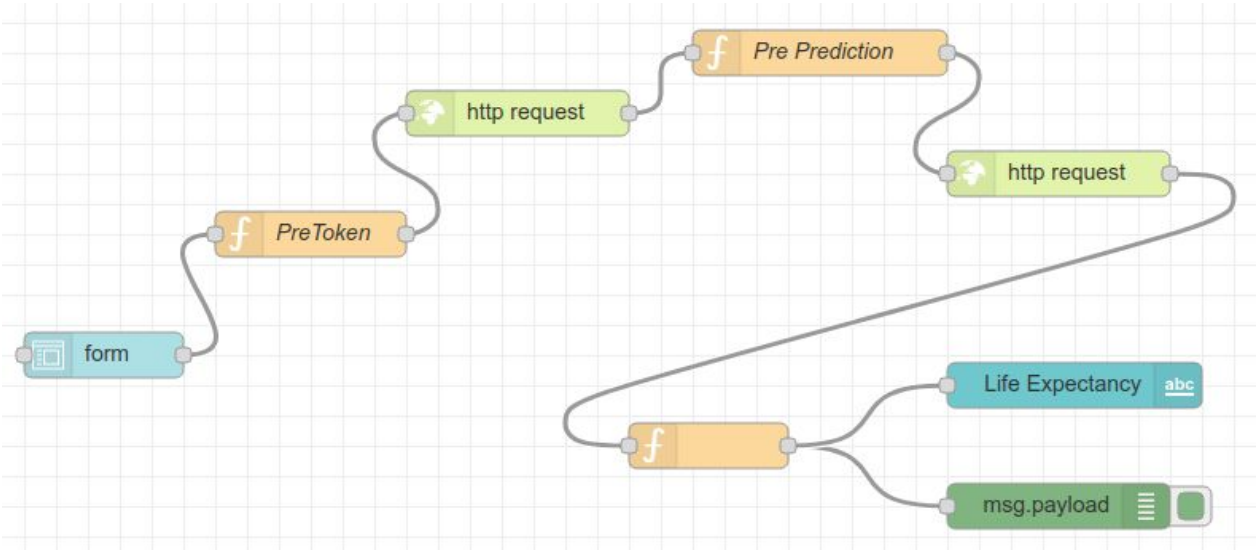


Fig: Node-Red Flow

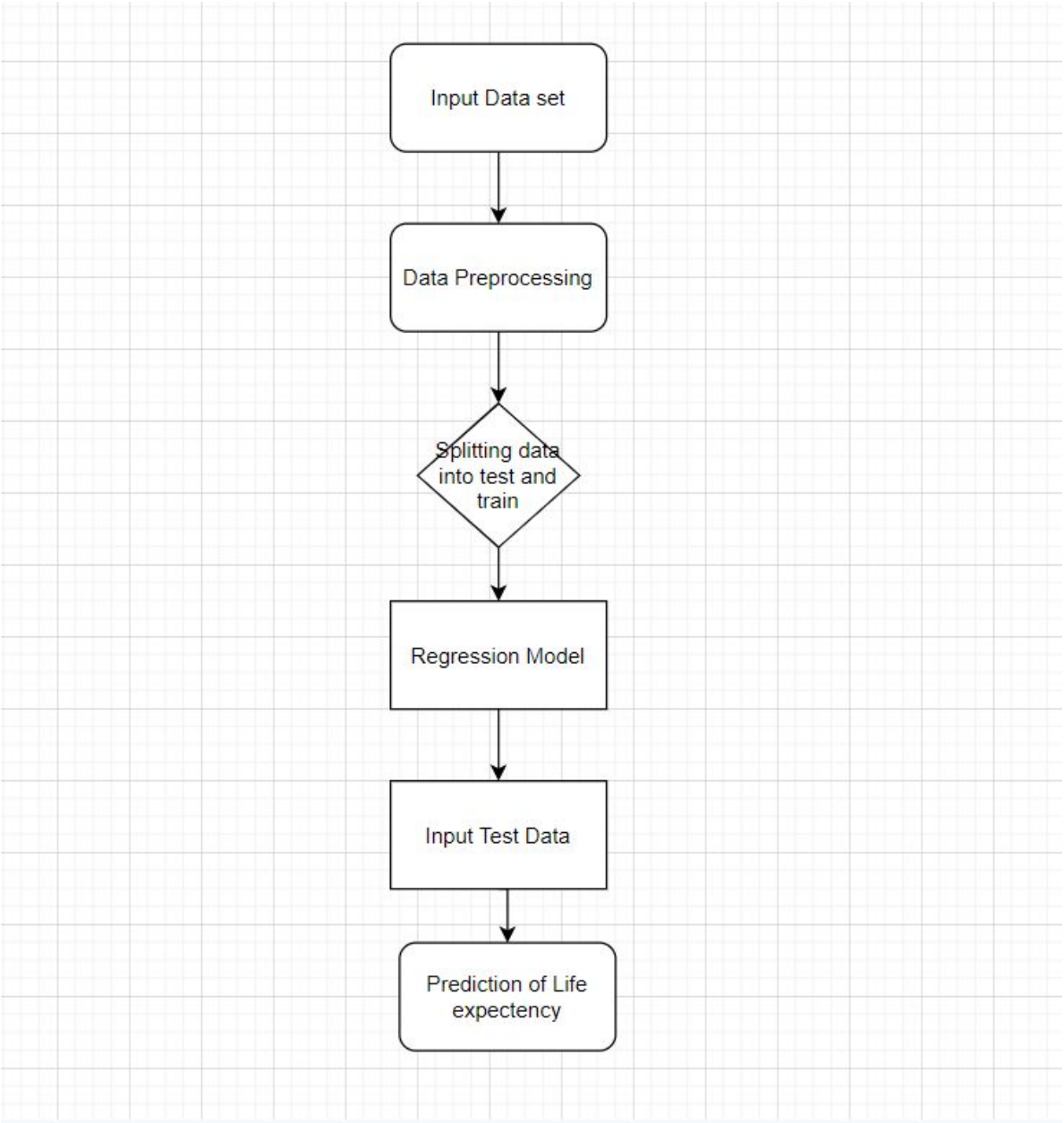
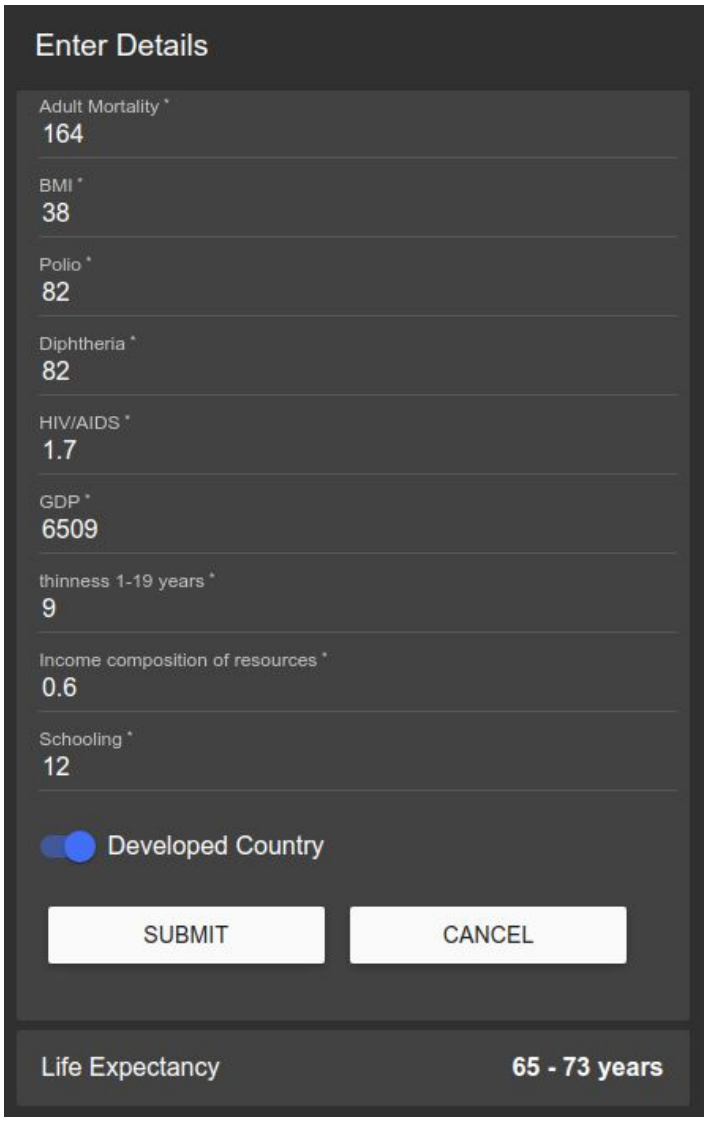


Fig: Model Workflow

Result

The model appears to the user in the form of an interface as Shown. The user has to fill in the inputs and click on the “SUBMIT” button at the end of the form. On clicking the “SUBMIT” button, the user will be displayed the predicted life expectancy as a range, based on the inputs provided, at the bottom of the page as shown.



Relative Merits & Consequences

Advantages :

1. IBM Watson:
    a. Process Unstructured Data.
    b. Fulfill Human Limitations.
    c. Improves performance and abilities.
    d. Handles large amount of data.
    e. Easy node-red integration.
    f. Easy access to other IBM services.
2. Easy User Interface (UI).
3. User – friendly.
4. Easy to predict.
5. Don’t require storage space.
6. Real-time prediction.

Disadvantages :

1. IBM Watson:
    a. Only in English Language.
    b. Maintenance.
2. Requires Internet Connectivity.

Applications:

Life expectancy is the primary factor in determining an individual's risk factor and the likelihood they will make a claim. Insurance companies consider age, lifestyle choices, and several other factors when determining premium rates for individual life insurance policies. It can be used by researchers to make meaningful researches out of it and thus, bring about something that will help increase the expectancy consider the impact of a specific factor on the average lifespan of people in a specific country.

Conclusion

Thus, we have developed a model that will predict the life expectancy of a specific demographic region based on the inputs provided. Various factors have a significant impact on the life span such as Adult Mortality, Population, Under 5 Deaths, Thinness 1-19 Years, Alcohol, HIV, Hepatitis B, GDP, Percentage Expenditure, and many more. Users can interact with the system via a simple user interface which is in the form of a form with input spaces which the user needs to fill the inputs into.

Future Scope:

As a future scope, we can connect the model to the database to have a record of predictions. This will help us analyze the trends in the life span. A model with country wise bifurcation can be made, which will help to segregate the data demographically.

Bibliography:

Source Code
Node-red Flow
Project Provider
Author
Test the Project