Predicting Life Expectancy using Machine Learning

Submitted By: K.Anirudhan

As part of the SmartInternz Remote Summer Internship Program

Project ID: llSPS-INT-1908

Contents

Introduction 3
Overview 3
Purpose 3
Literature Survey 3
Existing Problem 3
Proposed Solution 3
Theoretical Analysis 4
Block Diagram 4
Software Designing 4
Experimental Investigations 5
Flowchart 6
Result 7
Advantages and Disadvantages 8
Advantages 8
Disadvantages 8
Applications 8
Conclusion 8
Future Scope 8
Bibliography 8
Appendix 8

Introduction

Overview

Life expectancy is a statistical measure of the average time a human being is expected to live. It is dependent on several factors like regional variations, economic circumstances, sex differences, mental illnesses, etc. The primary goal of this project is to predict the life expectancy given related features like expenditure on healthcare system, GDP, education and some specific disease related data.

Purpose

The purpose of this project is to design a machine learning based solution to predict the life expectancy given features like GDP, healthcare expenditure of a country, disease related deaths, and other factors on which the life expectancy is dependent.

Literature Survey

Existing Problem

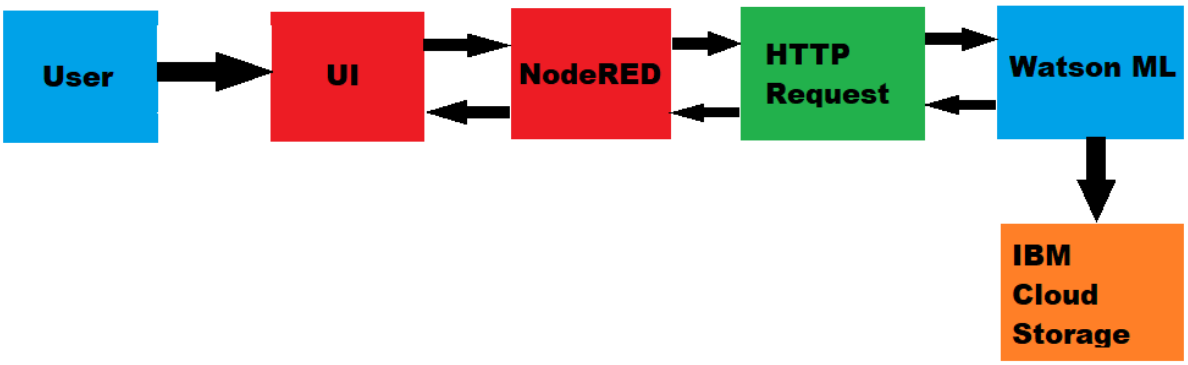
The problem of predicting the life expectancy given certain features is a regression problem. After pre-processing and feature engineering the data, existing regressions models can be trained on the data and then used to predict the life expectancy. The models can be further tuned to improve accuracy of predictions. The dataset is sourced from [1].

Proposed Solution

The proposed solution uses IBM Watson's AutoAI service, which builds and trains machine learning models based on the given data. The reason for using an automated service instead of a conventional method is that it is relatively easier to use and takes care of feature engineering, etc., which is usually time consuming. IBM's NodeRED is used to implement the UI and to integrate it with the model.

Theoretical Analysis

Block Diagram



Software Designing

The project primarily uses IBM's Cloud offerings and IBM's Watson services. NodeRED, which is based on node.js is used to design the UI (front-end) and connect it to the machine learning model (back-end).

The dataset is stored in IBM Cloud storage. IBM Watson Machine Learning service is used to create, train and deploy the machine learning model. AutoAI is used to create and train the model. The deployed model has endpoints to which the NodeRED sends HTTP requests to get the prediction based on the data fed by the user.

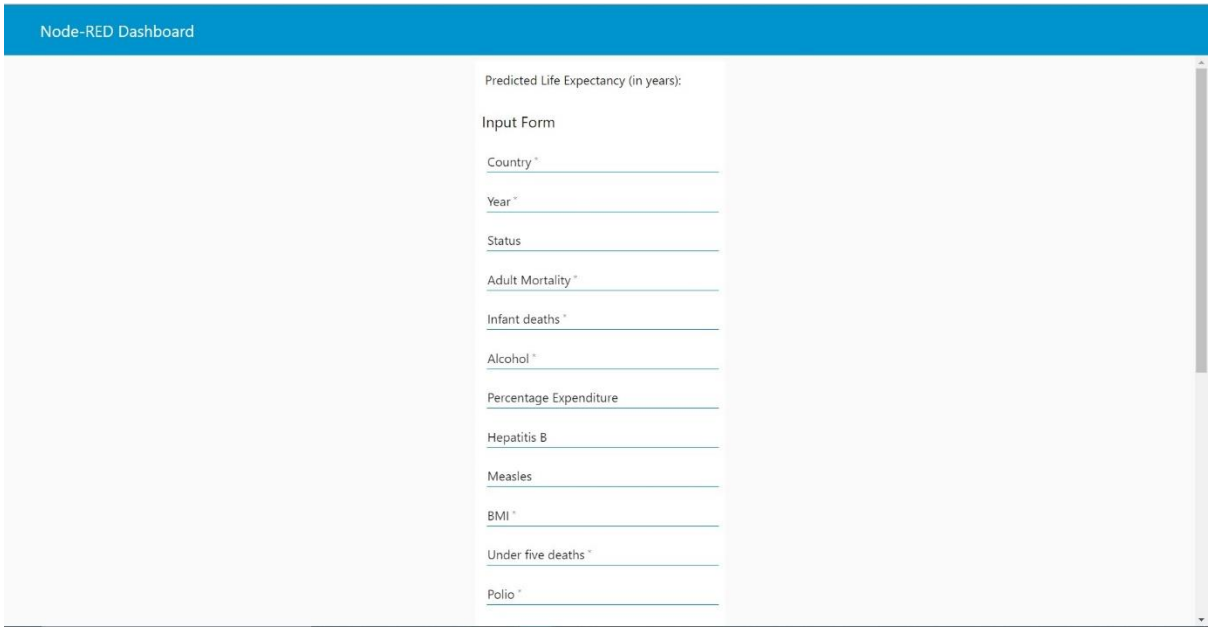


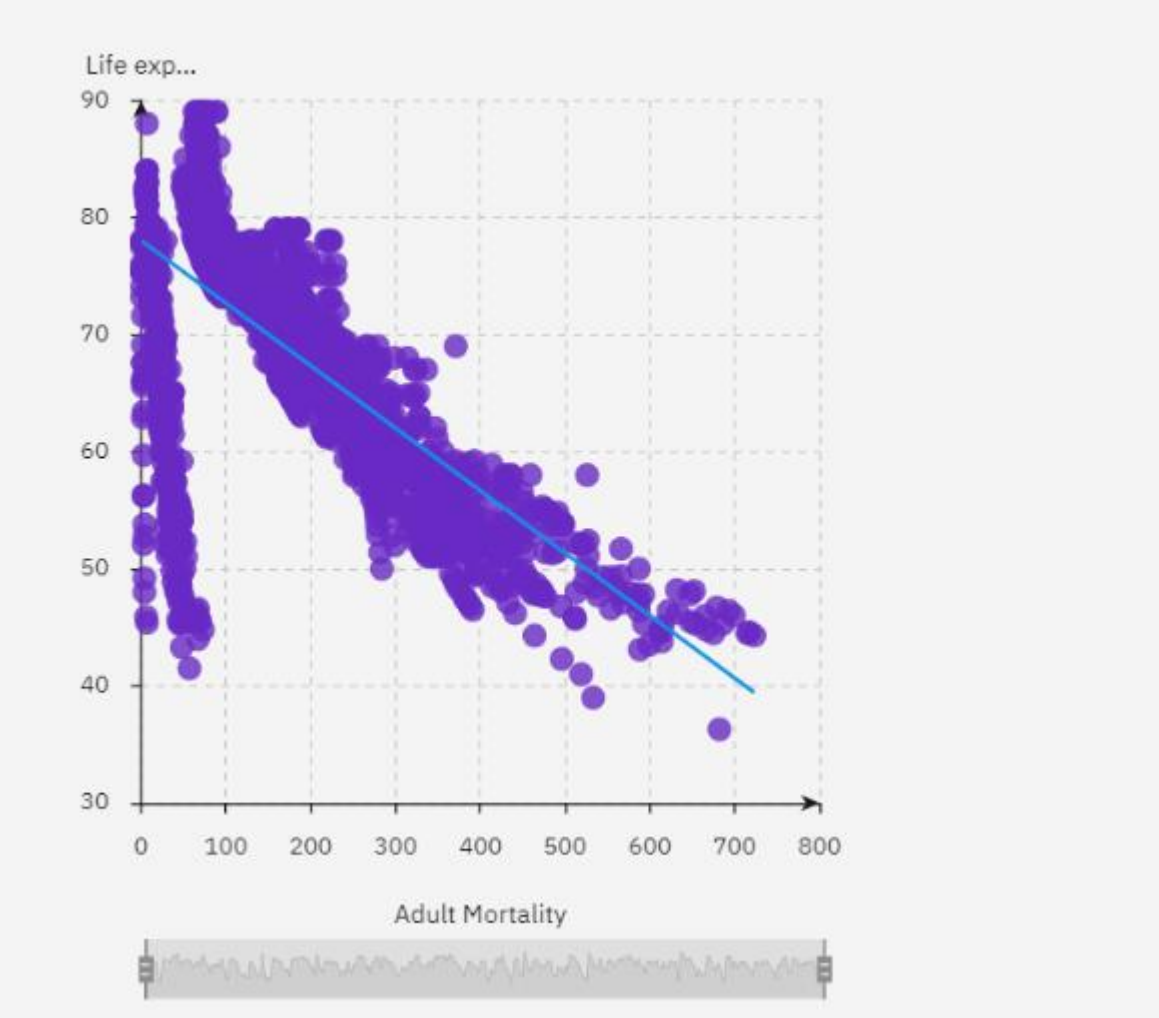
Figure shows the nodeRED webpage that the user uses to input the data.

Experimental Investigations

While the front-end can be implemented quite easily using NodeRED's interactive flow designer, the back-end can be implemented in two ways. The first is the traditional, manual way of doing it, i.e., using a Jupyter Notebook deployed in Watson Machine Learning. The other is using AutoAI, which takes care of everything from pre-processing the data, feature engineering to training and optimizing the model.

Initially, the conventional way was tried and then AutoAI was used. The results of both very similar, but the AutoAI model gave slightly better results and hence it was chosen for the final deployment.

After some pre-processing of the data, it is observed that certain features affect life expectancy more than others. One such feature is adult mortality, whose correlation graph is shown below. This is also seen by the importance given to this particular feature by the AutoAI model.



The image shows a scatter plot with a line to show the strong negative correlation between adult mortality and life expectancy.

Flowchart

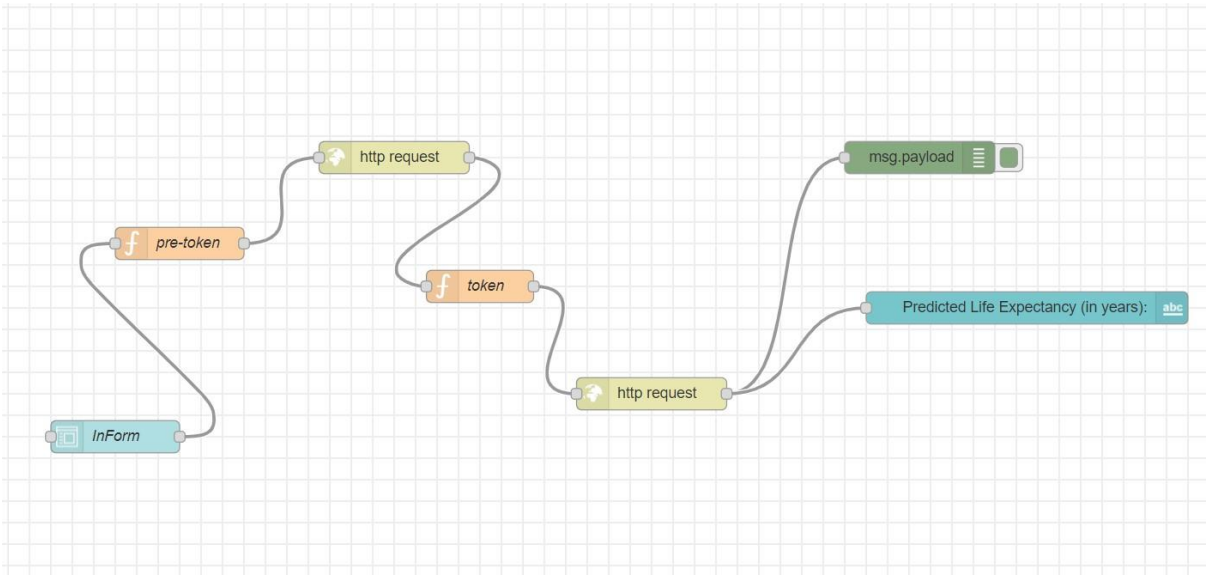


Figure shows the nodeRED flow, which is pretty much the same as the data flow. The user feeds in the data through the HTML form, which then sends a HTTP request to access the Machine Learning Service. Then, the user's input is fed to the model and the prediction thus obtained is displayed in a textfield in the webpage.

Result

	Holdout Score	Cross Validation Score
Root Mean Squared Error (RMSE)	1.830	2.010
R^2	0.961	0.956
Explained Variance	0.961	0.956
Mean Squared Error (MSE)	3.347	4.057
Mean Squared Log Error (MSLE)	0.001	0.001
Mean Absolute Error (MAE)	1.182	1.282
Median Absolute Error (MedAE)	0.740	0.747
Root Mean Squared Log Error (RMSLE)	0.028	0.031

The table shows the evaluation metrics' values of the best performing model generated by AutoAI.

Advantages and Disadvantages

Advantages

1. The use of AutoAI takes a pretty considerable burden off the shoulders of the developer.
2. NodeRED is interactive and really easy to use, abstracting the coding part and keeping it to a minimum.

Disadvantages

1. The UI of the webpage created is slightly tedious to use because of the sheer number of features to be given as input to the model.

Applications

Since the model was created and trained for the sole purpose for predicting the life expectancy given certain related features, it can be used anywhere where there is a need to predict the life expectancy.

Conclusion

The model performs the intended function really well. IBM's AutoAI made the whole process of create and training the model extremely easy and NodeRED made the daunting task of integrating and deploying the model on the web very easy, particularly for machine learning novices who don't have much experience in web development.

Future Scope

The model could possibly be improved by manually pre-processing the data and doing feature engineering. Also, with a better dataset, a more complete one, the model can be trained better and will give more accurate predictions.

Bibliography

[1] - https://www.kaggle.com/kumarajarshi/life-expectancy-who

Appendix

Source code for the model (Jupyter notebook) and the NodeRED flow can be found in the project's repository.