Report on Predicting Life Expectancy Using Machine Learning

-by Siddhant Meshram

INDEX

I. Introduction

1. Overview

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly [1].

2. Purpose

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. The dataset is provided by WHO. The data offers a timeframe from 2000 to 2015. The data is originated from https://www.kaggle.com/kumarajarshi/life-expectancy-who/data [2].

II. Literature Survey

1. Existing Problem

Life expectancy plays an important role when decisions about the final phase of life need to be made, but there's always been a big challenge in collecting data for predicting life using machine learning due to some laws related to privacy or other government policies [3]. Also, previously the medical data required for predicting life expectancy were not available digitally. So, predicting life expectancy using machine learning was a tedious task. As we know that life expectancy depends on many other factors like, Regional variations, Economic Circumstances, Sex Differences, Mental Illnesses, Physical Illnesses, Education, Year of their birth and other demographic factors. If we want to know life expectancy of a particular country then, 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 should be present with us. Although now-a-days medical records are increasingly available in the

form of electronic medical records (EMRs), they remain underutilized for developing clinical decision support systems, and improving health care in general. EMRs are characterized by irregularly-sampled time-series data, missing values, long-term dependencies involving symptoms, diagnoses and interventions, and are prone to documentation errors. Moreover, they contain important information in the form of unstructured, textual data, from which information cannot be extracted straightforwardly. These challenges lead to suboptimal use and even waste of large portions of data, especially when the data is unstructured and noisy. Free texts make up a significant and important part of EMR data, but their ambiguous and noisy character and the and lack of canonical forms for medical concepts and the relations between them make it difficult to 'mine' these texts effectively [3].

2. Proposed Solution

The proposed solution for the task of predicting life expectancy is to use a supervised machine learning model. The life expectancy data provided by WHO, was trained & tested in order to develop model using a simple regression. IBM Watson's Machine Learning service was used to train & test the model through its Auto AI function. Moreover, the Node Red application was created using IBM Cloud Services & integrated our model with Node-Red flow in order generate a user interface (UI).

III. Theoretical Analysis

1. Block Diagram

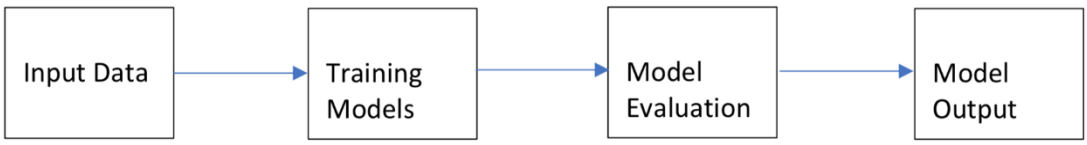

Fig 1. Basic Machine Learning Process [4]

2. Hardware/ Software Designing

A User Interface of Predicting Life Expectancy using ML was created in software designing using following softwares:

a. IBM Cloud Services
b. IBM Watson
c. IBM Node Red Application
d. IBM Machine Learning
e. IBM Auto AI
f. GitHub
g. Zoho Writer

IV. Experimental Investigations

The IBM Watson's Machine Learning service was used in order to train & test our model. The Auto AI function, which is a part of IBM Watson's ML service, was used to automate data preparation, model development, feature engineering & hyper-parameter optimization.

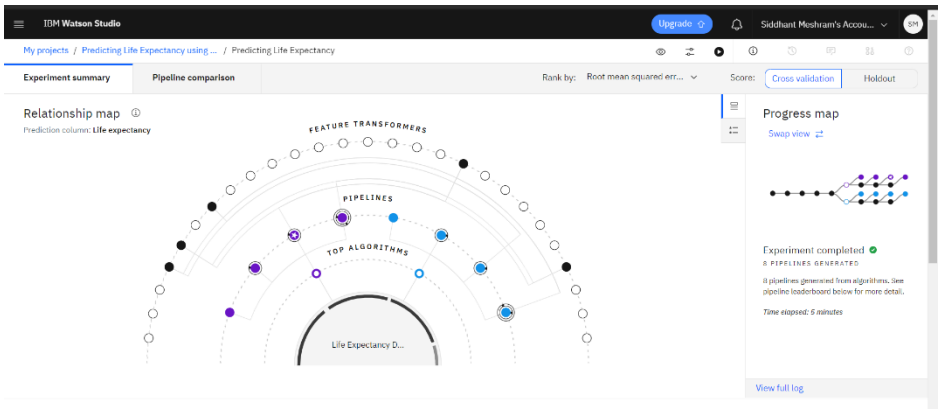

Fig 2. Relationship Map (Auto AI)

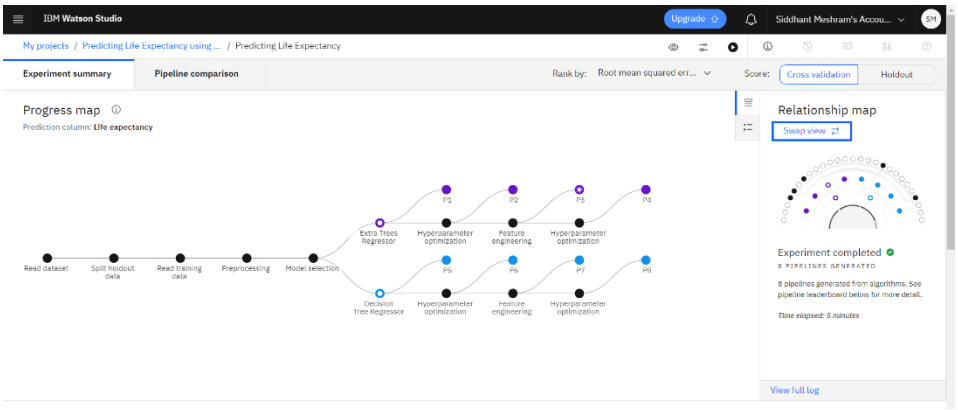

Fig 3. Progress Map (Auto AI)

The Auto AI ran different algorithms & ranked them according to its performance (as shown in Fig 4). The pipeline 3 which used Extra Trees Regressor algorithm turned out to be performing best. Also, Root Mean Squared Error (RMSE) model had the best evaluation measures (as shown in Fig 5).

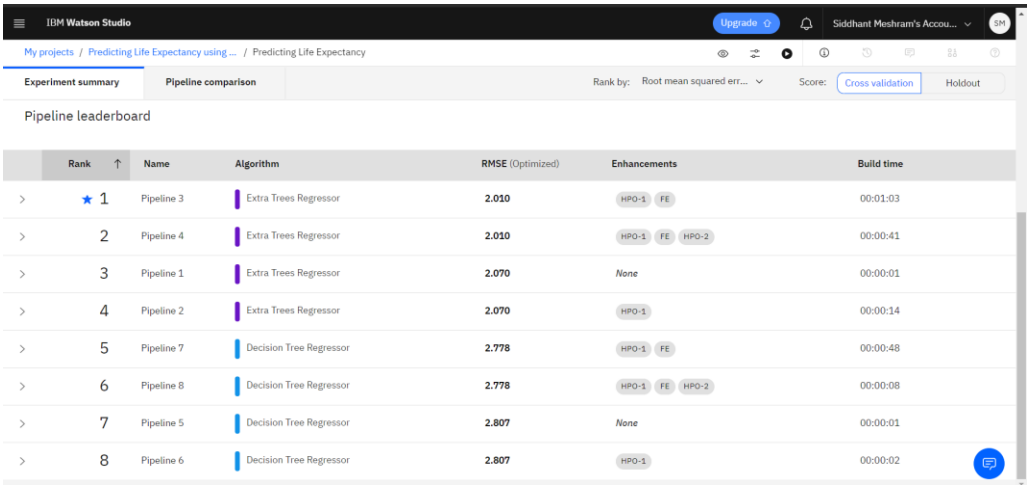

Fig 4. Pipeline Leaderboard

Model Evaluation Measures

TARGET : LIFE EXPECTANCY

	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

Fig 5. Model Evaluation Measures

In pipeline 3 model, the most important feature is Income Composition of Resources (as shown in Fig 6).

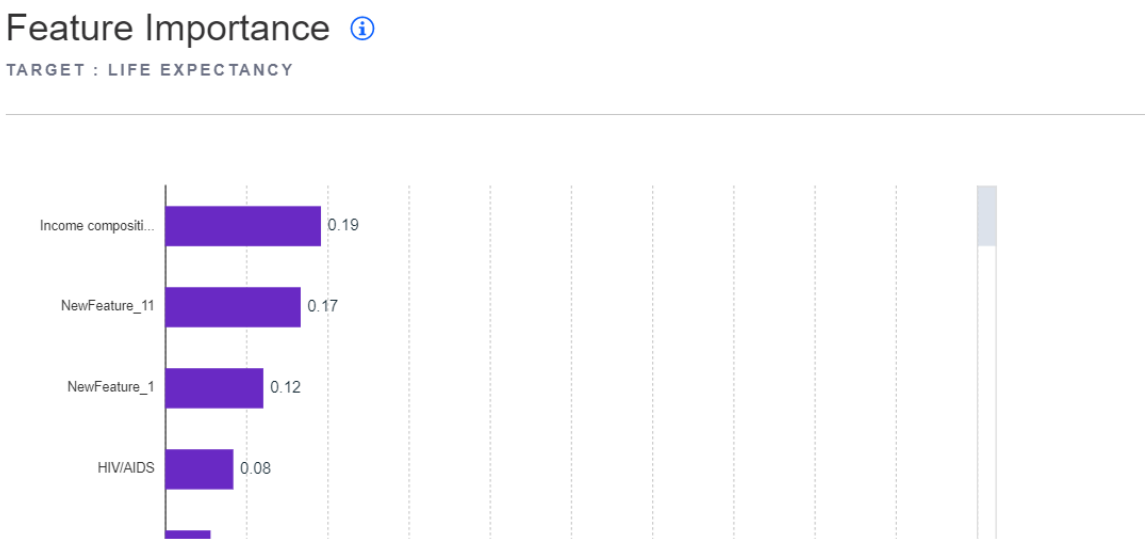

Fig 6. Feature Importance

V. Flowchart

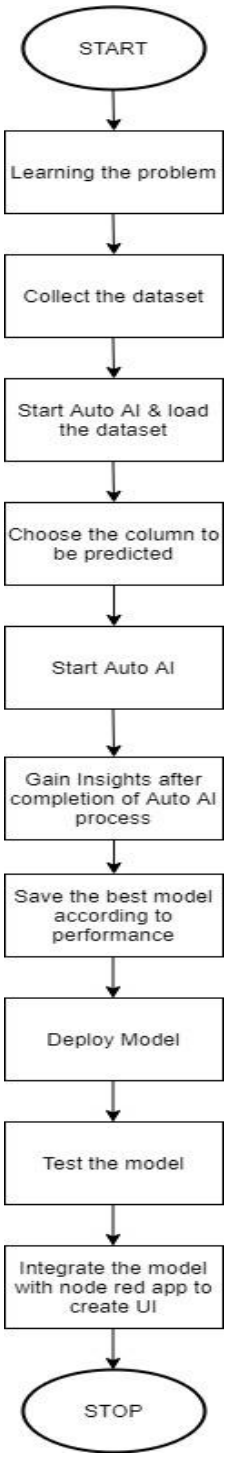

Fig 7. Process of Auto AI & creation of UI

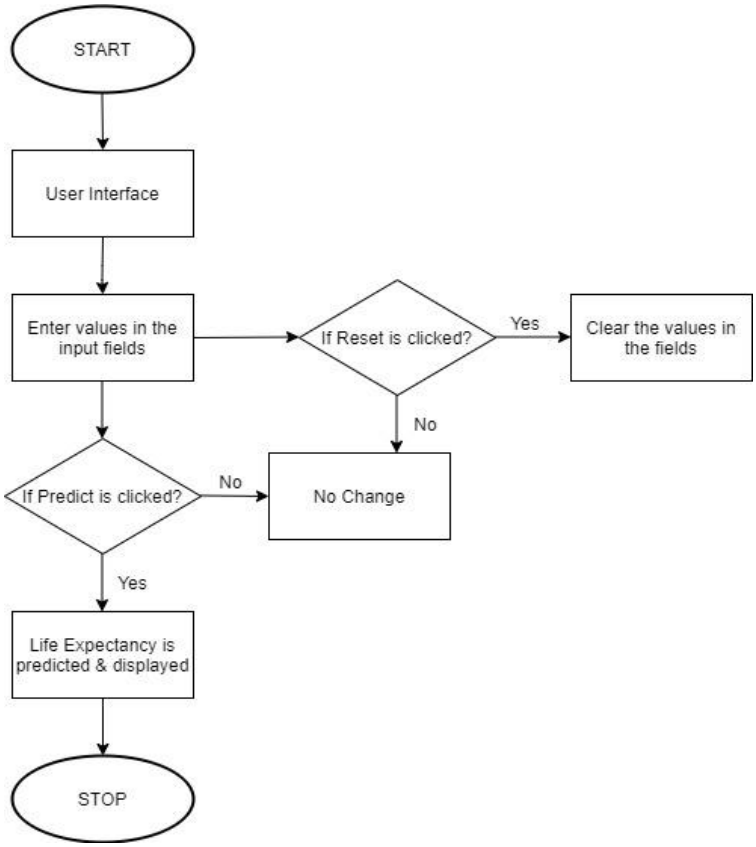

Fig 8. Working of User Interface

VI. Result

The Life Expectancy Model was trained & tested using IBM Watson's ML service. As we implemented the project using Auto AI, the Root Mean Squared Error (RMSE) had the best evaluation measures with value 2.010. Also, the best performing algorithm was Extra Trees Regressor algorithm which was used by pipeline 3. The most important feature on which the life expectancy depended on was 'Income Composition of Resources'. The User Interface was successfully implemented by integrated Node Red flow and Auto AI. The node red flow was then deployed & after filling input fields in the UI, we obtained our predicted value (as shown in Fig 9).

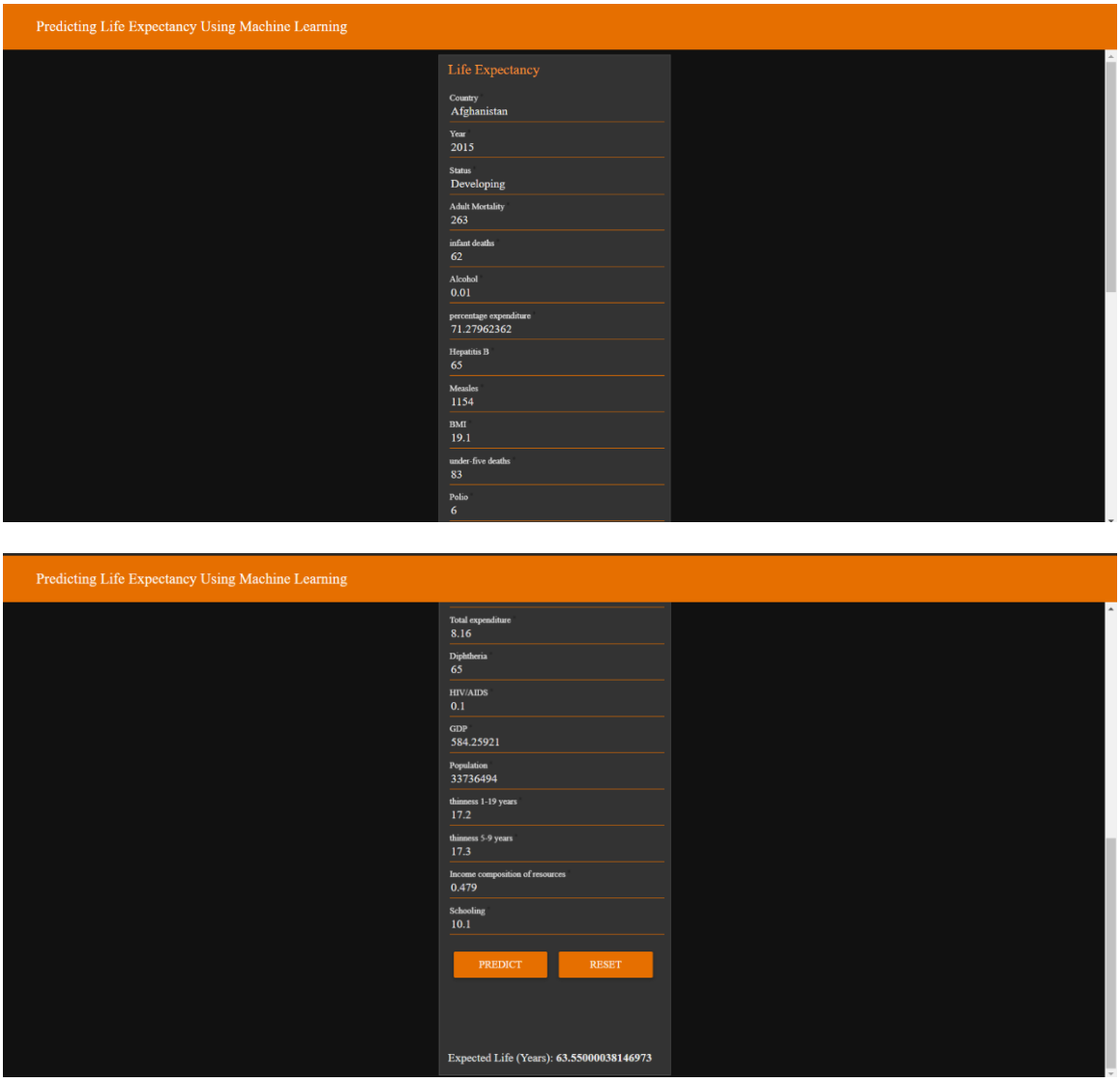

Fig 9. User Interface

VII. Advantages & Disadvantages

1. Advantages

a. Since the factors on which the Life Expectancy depends on are known, therefore by improving these factors in life can help us to improve our life span.
b. The model is integrated with UI; therefore, it is easy to use this interface to predict life expectancy easily.
c. The Auto AI model tells us the most important feature on which the Life Expectancy depended on; therefore, we can use this insight to create awareness about it & this might help people in improving their life spans.

2. Disadvantages

a. Error in data, can result in wrong predictions.
b. Since the insights are purely based on the available data, there may be other factors or features on which the Life Expectancy depends on.
c. Since the model is implemented through Auto AI, so it may require some manual tuning.

VIII. Applications

a. The Model can be implemented in the health care & pharmaceutical organizations.
b. This model will also help us to increase life expectancy considering the impact of a specific factor on the average lifespan of people of specific country.

IX. Conclusion

The Prediction of Life Expectancy using Machine Learning was successfully implemented. The best performing algorithm was Extra Trees Regressor algorithm which was used by pipeline 3. The Root Mean Squared Error (RMSE) had the best evaluation measures with value 2.010. The UI was created using integrating node red flow & Auto AI model. Also, the scoring endpoint has been integrated with UI using Node Red flow. The UI has been tested & deployed successfully.

X. Future Scope

Prediction of Life Expectancy using Machine Learning can be a foundation for predicting life of different species or it can be used to get insights for life expectancy after going through some major diseases (like HIV, COVID-19, etc.). Some additional features can be included by using an enhanced dataset. Also, the UI can be improved by adding some interactive features to it. The analysis of big data using machine learning can provide some good insights to health researchers around the world.

XI. Bibliography

[1] https://expertsystem.com/machine-learning definition/#:~:text=Machine%20learning%20is%20an%20application,use%20it%20learn%20for%20themselves.

[2] https://www.datasciencesociety.net/using-machine-learning-to-explain-and-predict-the-life-expectancy-of-different-countries/

[3] https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0775-2

[4] https://medium.com/@sattiraju/exploring-the-mutually-inclusive-modern-data-architecture-of-machine-learning-and-serving-1f2a078538c4

[5] https://bookdown.org/caoying4work/watsonstudio-workshop/jn.html

[6] https://bookdown.org/caoying4work/watsonstudio-workshop/auto.html

[7] https://developer.ibm.com/tutorials/how-to-create-a-node-red-starter-application/

XII. Appendix

A. **UI Link:** https://node-red-pleml.eu-gb.mybluemix.net/ui/#!/0?socketid=dn7sNLc8duoXFFekAAAN
B. **Dataset Link:** https://www.kaggle.com/kumarajarshi/life-expectancy-who/data
C. **Source Code:** https://github.com/SmartPracticeschool/llSPS-INT-2631-Predicting-Life-Expectancy-using-Machine-Learning.git
D. **Demonstration Video:** https://drive.google.com/file/d/1Gj5cpSiSZzBlkjVZ3r5kMcYegWICRm7T/view?usp=sharing