PROJECT TITLE

Predicting Life Expectancy using Machine Learning

FINAL PROJECT REPORT

NAME : SHREYA SINGH
COMPLETED UNDER SMARTBRIDGE
DATE : 5/06/2020

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 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. It compares three machine learning models to determine the life span of an individual.

- **DATASET**

The data-set related to life expectancy, health factors for 193 countries have been collected from the same World Health Organization 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, such as the Gross Domestic Product, illnesses, Schooling, percentage expenditure etc. Therefore, in this project we have considered data from year 2000-2015 for 193 countries for further analysis.

1.2 PURPOSE

The prediction of Life expectancy has become easier with the development of prediction algorithms in recent years which can determine the Life span of an individual affected by various factors. This project serves to accomplish the task of calculating Life span of individuals by produced continued value outputs by means of regression, and determining the best model to complete this task.

1.3 ROLES AND RESPONSIBILITY

This project was undertaken by me solely till the completion of the given problem statement which was to calculate the Life expectancy of an individual using regression techniques of Machine Learning.

2 LITERATURE SURVEY

2.2 PROPOSED SOLUTION

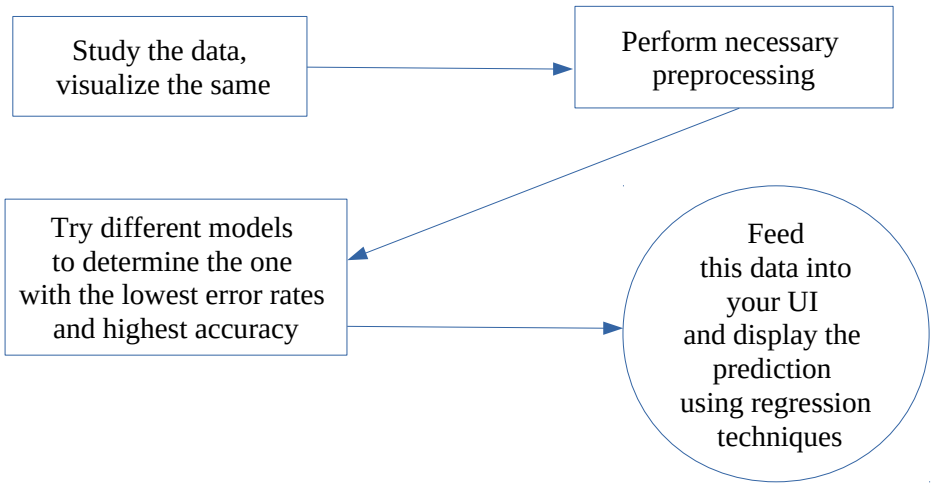
Four Machine Learning Models were tested in the making of this project

1. Model 1 - Multivariate Linear Regression
2. Model 2 – Random Forest Regression
3. Model 3 – XGBoost
4. Model 4 – Neural Networks

Observing the performance, the Random Forest Regression model provided the best output among the models tested with the lowest error rates.

3. THEORETICAL ANALYSIS

3.1 BLOCK DIAGRAM



3.2 HARDWARE/SOFTWARE DESIGNING

- Python
- IBM Cloud
- IBM Watson Studio
- Regression models
- Visualization Library - Matplotlib, Node Red.
- Github
- Nodered
- Model Integration and Deployment

4. EXPERIMENTAL INVESTIGATION

SCOPE

The outcome of the project is the prediction of the Life expectancy of an individual based on certain factors such as Schooling, diseases, GDP of the country, population, Alcohol consumption, percentage expenditure etc.
It was found that factors such as the GDP, illnesses such as Diptheria, Alcohol consumption and Schooling, significantly impacted the overall regression output. Other factors that less impacted the outcome were Country, Hepatitis B, Adult Mortality and Status i.e. whether a country was developing or developed. 70 percent data was trained upon and 30 percent kept for testing purposes.
The model successfully calculated the Life expectancy with a mean absolute error as small as 3.4468 and a root mean square error of 5.4594, for the multivariate linear regression model,
Random forest validation MAE = 1.306961451247165, RMSE = 1.952089292965296 and XGBoost validation MAE = 1.5699614537816469
XGBoost validation RMSE = 1.952089292965296
Neural Networks were also tested but did not give good regression predictions when the number of features to be learned were increased despite increasing the number of hidden layers and epochs/

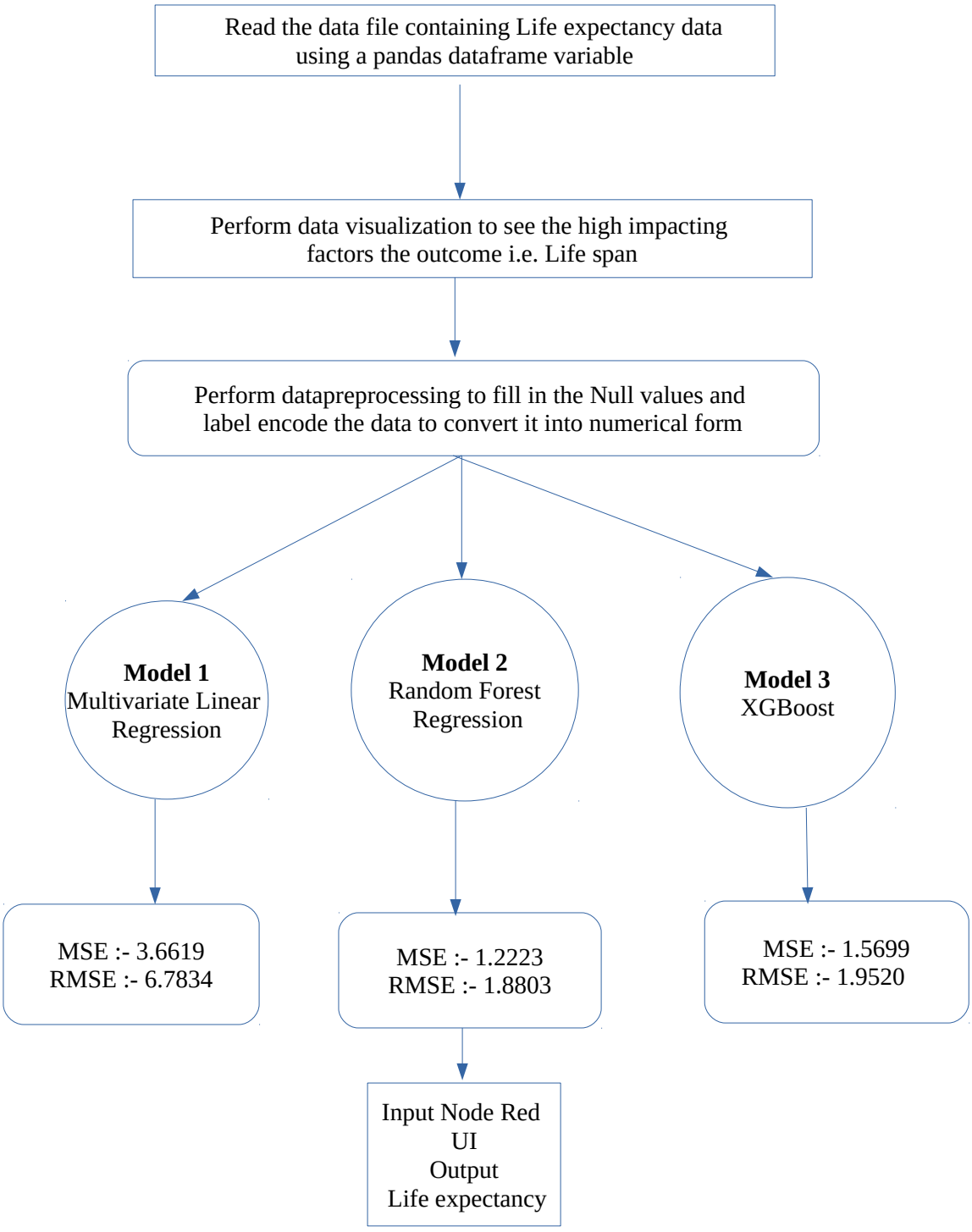
SCHEDULE :- 1 month

- Project planning and kickoff
- Setting up development environment -
- Creation of IBM cloud account
- Creating a basic Node Red application
- Exploring IBM Watson Studio and use cases
- Studying various algorithms such as classification, regression and clustering to choose best algorithm for predicting the life expectancy
- Building the model using Multivariate Linear Regression, Random Forest and XGBoost
- Comparing the model with the performance of Neural Networks on the same data
- Building the Node Red flow and integrating it with the model

COSTS

Rs 0.00 (No cost was incurred during the making of this project).

5. FLOWCHART

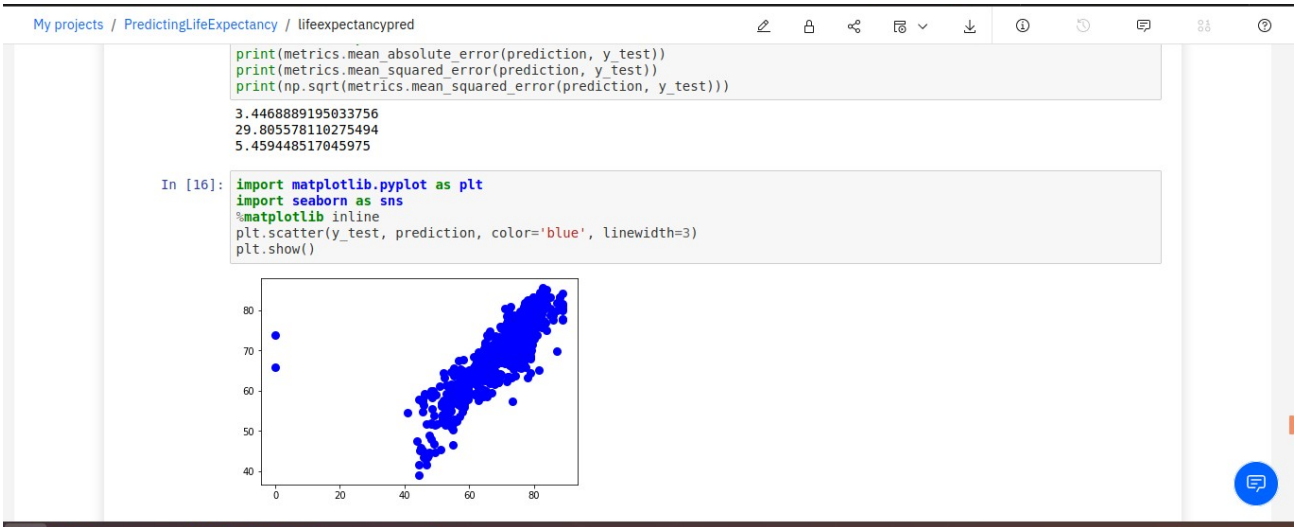

Read the data file containing Life expectancy data using a pandas dataframe variable
Perform data visualization to see the high impacting factors the outcome i.e. Life span
Perform datapreprocessing to fill in the Null values and label encode the data to convert it into numerical form
Model 1
Multivariate Linear Regression
Model 2
Random Forest Regression
Model 3
XGBoost
MSE :- 3.6619
RMSE :- 6.7834
MSE :- 1.2223
RMSE :- 1.8803
MSE :- 1.5699
RMSE :- 1.9520
Input Node Red UI
Output
Life expectancy

6. PROJECT PERFORMANCE SUMMARY / RESULT

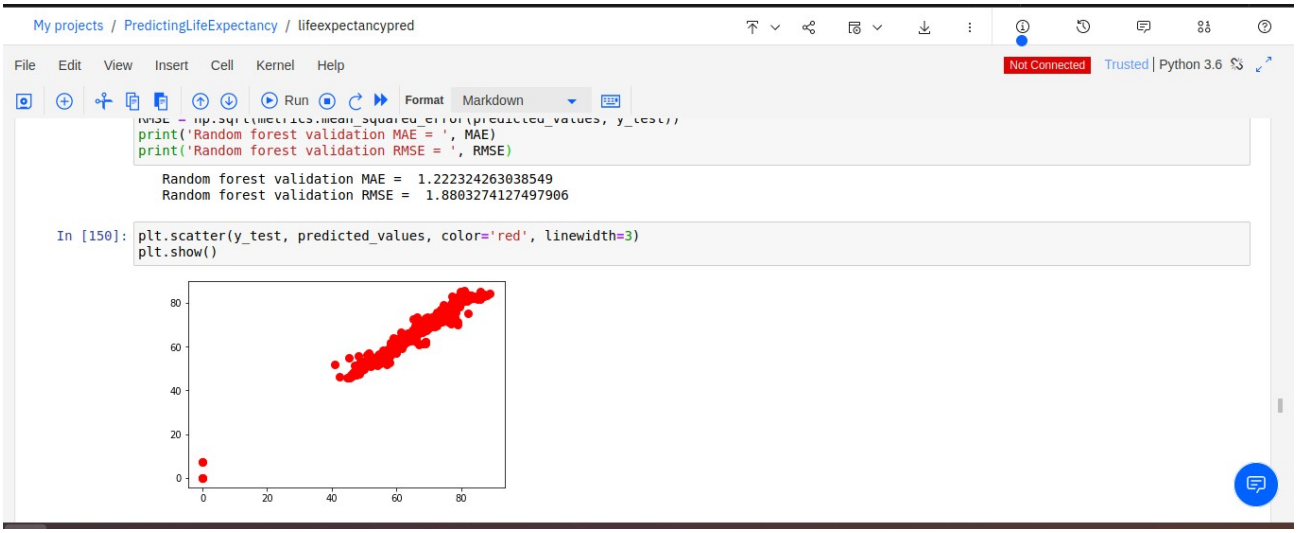
The model successfully calculated the Life expectancy with a mean absolute error as small as 3.4468 and a root mean square error of 5.4594, for the multivariate linear regression model,
Random forest validation MAE = 1.306961451247165, RMSE = 1.952089292965296 and XGBoost validation MAE = 1.5699614537816469, XGBoost validation RMSE = 1.952089292965296

Below are the scatter plots of predicted values vs the true values for the multivariate linear regression model and random forest model to show the accuracy of the predictions.

1. Scatter plot of Multivariate Linear Regression Model



2. Scatter plot of the best performing model : Random Forest Model



7. ADVANTAGES AND DISADVANTAGES

Advantages : We can calculate the life span of individuals from over 193 countries with this model and obtain nearly accurate solutions by taking into account only a few factors. It can be estimated at any age depending on the impact factors which are provided in the dataset.
We can study the trends that impact Life expectancy and the correlation factors by means of simple data visualization.

Disadvantages : Younger deaths / premature mortality are hard to determine given the existing dataset , factors such as infant mortality are found to be rare occurrences and hence do not affect the overall Life expectancy to the extent that it should. At smaller geographies, due to smaller numbers more subject to random variation year on year. These numbers are harder for the machine learning model to capture.

8. APPLICATIONS / CONCLUSION

As mentioned previously, we can calculate the life span of individuals from over 193 countries with this model and obtain nearly accurate solutions by taking into account only a few factors. This model has been generalized for the above purposes.

9. FUTURE SCOPE

The model can be used in the future to determine average Life span, with the necessary impact factors the accuracy of the model can be improved further.

10. BIBLIOGRAPHY

- https://developer.ibm.com/tutorials/how-to-create-a-node-red-starter-application/
- https://nodered.org/
- https://developer.ibm.com/technologies/machine-learning/series/learning-path-machine-learning-for-developers/
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4502577/
- Dataset reference :https://www.kaggle.com/kumarajarshi/life-expectancy-who

11. APPENDIX

A. Source Code

Can be found in the git repository

https://github.com/SmartPracticeschool/llSPS-INT-1651-Predicting-Life-Expectancy-using-Machine-Learning/blob/master/Lifeexpectancyprediction.ipynb