PROJECT TITLE

Predicting Life Expectancy using Machine Learning

FINAL PROJECT REPORT

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

INTRODUCTION

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.

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.

FINAL PROJECT SUMMARY

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.

1.1 CONTENT SUMMARY

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).

1.2 LESSONS LEARNED

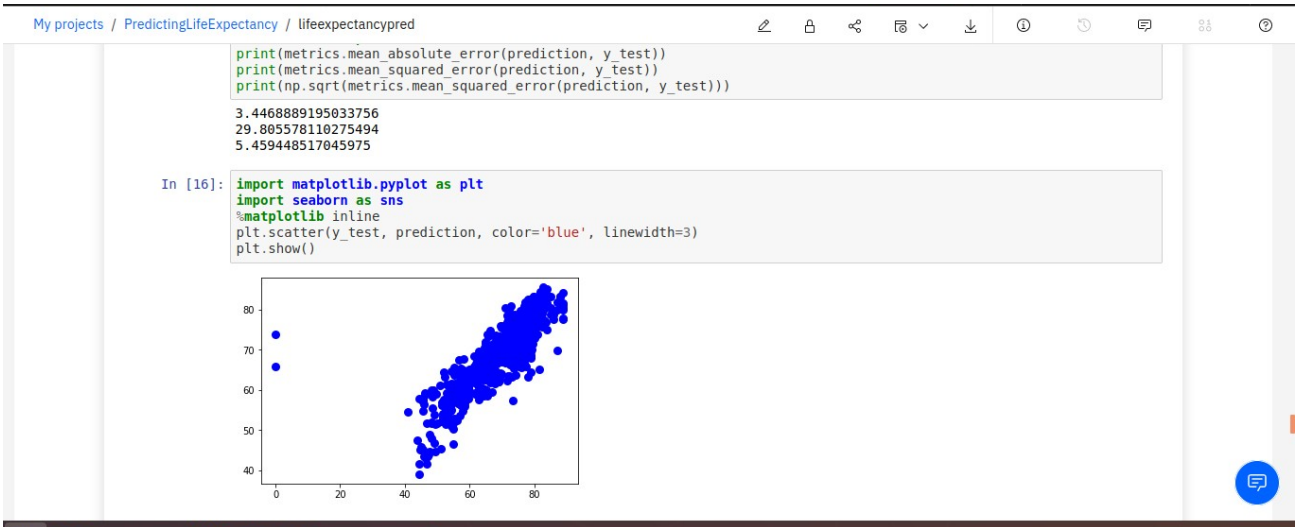
1. IBM Cloud -Watson services
2. Node Red flow editor
3. Regression Models
4. Data Visualization
5. Model Integration and deployment
6. Github
7. Auto AI
8. Zoho Writer

1.3 PROJECT PERFORMANCE SUMMARY

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