Remote Summer Internship Program 2020
Machine Learning, Career Basic Program
Smartinternz, SmartBridge

Predicting Life Expectancy using Machine Learning

Internship Report
by
Vinny Chamoli

15/5/2020-15/6/2020

Contents

1. Program Information

Summer Internship Program by Smartbridge is an annual initiative taken up by them to teach and prepare students across the globe for industry experience. They believe that experiential learning and development in a professional like environment can only bridge the gap between students and industries opening ways for both of them to achieve better results. This initiative enables students to better their resume to embark upon a successful industrial journey.

They provide various roles for internship according to the possible aptitude of the students such as Artificial Intelligence, Machine Learning, Internet of Things etc.A project is assigned to students individually with access to the platform wherein the students code and deploy their model just like in industries. This not only introduces them to environments like IBM cloud but also helps them in understanding the industrial environment better.

Classes are organised according to the technologies to be taught and doubt sessions are taken up by the mentors to help students complete their projects too.

2. Internship Description - Problem Statement

Predicting Life Expectancy using Machine Learning

My role at the internship was of a Machine Learning Engineer and the project given to me was to generate a Regression based Machine Learning model.

'Life Expectancy' refers to the number of years a person can be expected to live which is governed by a number of factors ranging from demographic to genetic to habits.

This project involved building a machine learning (ML) model to predict the life expectancy rate of a country given various features like year, GDP, education, alcohol intake of people in the country, expenditure on healthcare system and some specific disease related deaths that have happened.

The developed model works on the data set provided by the World Health Organization(WHO) to evaluate the life expectancy in different years of a country. The time frame offered in the data set is from the year 2000 to 2015. Regression based algorithm that has been used to predict the life expectancy for the data : - Random Forest Regression

The model that involves Python to code these regression techniques uses the Jupyter Notebook in IBM Watson Studio to import data and automate the ML model otherwise the IBM Watson Studio services are used to auto AI the experiment. A Node - RED flow is built to integrate the ML services or Auto AI.

3. Design of Project

Methodology used

work flow for this project can be divided into three sub-tasks.These include acquiring the data and understanding various features of the data, preprocessing the data set to align it with our requirements and remove any inconsistency, and finally analyzing the data using regression based prediction algorithm with the key performance index being accuracy of prediction.

Figure 3.1: Workflow of the project

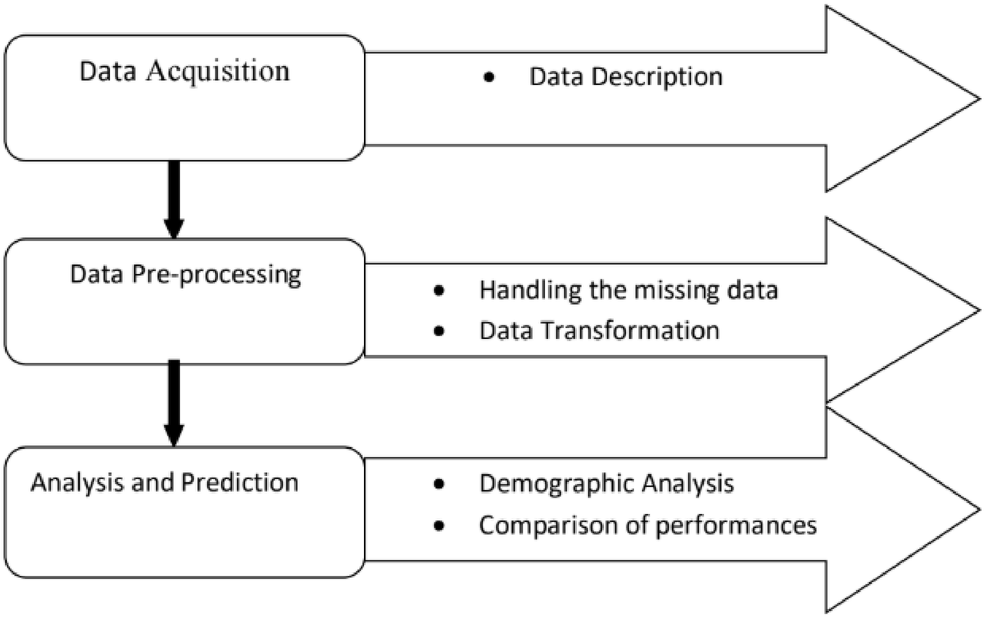


Fig 1. Work flow of the project

4. Machine Learning Model

4.1 Algorithm Formulated to solve the given Problem Statement

*A*lgorithm steps:
Step 1: Import the Data set
Step 2: Read and Understand the data
Step3: Explore the Data set
Step4: Decide the amount of data for training data and testing data
Step5: Give 70 percent data for training and remaining data for testing.
Step6: Assign train data set to the models
Step7: Choose the algorithm and create the model
Step8: Make predictions for test data set.
Step9: Calculate accuracy for the algorithm
Step 10: Apply the model for further predictions.

4.2 Setting Up the Environment

An IBM cloud account was set up to access various services to create and deploy the model.

The following services have been used in the project:

1. Watson Studio - This is where the notebook has been created in a project to write the regression code along with the data set.

2. Node Red - Node Red is the front end application that uses interconnecting nodes to interact with machine learning services of the cloud and the model to show predictions when inputs are given

4.3 Data Acquisition

4.4 Model Requirements

4.4.1 Python

Python is a multi-paradigm, general purpose, high level programming language, which focuses on code readability. It has a large library, which provides tools for many tasks and has a wide support base. This project uses python 3.5.

4.4.2 Python Libraries

4.4.2.1 Pandas

Pandas is used for data manipulation and analysis through operations and data structures on numerical tables and time series.

4.4.2.2 Numpy

It adds support as well as contains high-level mathematical functions to operate on large multidimensional arrays and matrices.

4.4.2.3 Matplotlib

It is a plotting library that that enables 2d diagramming and begetting of bar charts, histograms and so forth.

4.4.2.4 Sci-kit learn

It is a free software machine learning library that features various regression, clustering and classification algorithms. It works in conjunction with numPy and python scientific library sciPy.

4.5 Data Preprocessing

Data preprocessing is an essential step in order to increase the accuracy of machine learning models. It involves handling inaccurate and missing data, noisy data in the form of outliers, and inconsistent data in the form of duplication and others.

4.6 Data Cleaning

Data was often not consistent; missing values or values out of range was common. The methods used for cleaning is to replace the missing or noisy values by forward filling them using mean of the feature.

4.7 Analysis and Prediction

Random Forest algorithm has been applied to the data set to train the model and increase the accuracy for prediction of the the life expectancy of any given country.

4.7.1 Data Transformation

4.7.1.1 Scaling

Scaling is required to standardize the independent feature in the dataset to a fixed range. Primarily, two types of feature scaling methods:

1. Min-max scaling (Normalization) (value - min)/(max - min) Sklearn provides a class called MinMaxScaler for this

2. Standardization (value - mean)/std Sklearn provides a class called StandardScaler for this

4.7.2 Evaluation Metrics

Evaluation metrics calculated are:

1. Mean Cross validation score

2. Score without cv

3. R$^2 score$

4.7.3 Random Forest Regression

4.7.4 Introduction

Random forests is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is. Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of the feature importance.

It technically is an ensemble method (based on the divide-and-conquer approach) of decision trees generated on a randomly split dataset. This collection of decision tree classifiers is also known as the forest. The individual decision trees are generated using an attribute selection indicator such as information gain, gain ratio, and Gini index for each attribute. Each tree depends on an independent random sample. In a classification problem, each tree votes and the most popular class is chosen as the final result. In the case of regression, the average of all the tree outputs is considered as the final result. It is simpler and more powerful compared to the other non-linear classification algorithms.

4.7.5 Working

It works in four steps:

1) Select random samples from a given dataset.

2) Construct a decision tree for each sample and get a prediction result from each decision tree.

3) Perform a vote for each predicted result.

4) Select the prediction result with the most votes as the final prediction.

4.7.6 Node Red Flow

A Node RED starter application was created to implement the front end of the project. In the starter application, nodes are dragged and dropped to create a flow to integrate the application with the machine learning model.

The following form appears after deployment of the app wherein the user can input values and life expectancy prediction is displayed acccording to the inputs.

Figure 4.1: Workflow of the Radom Forest Regression

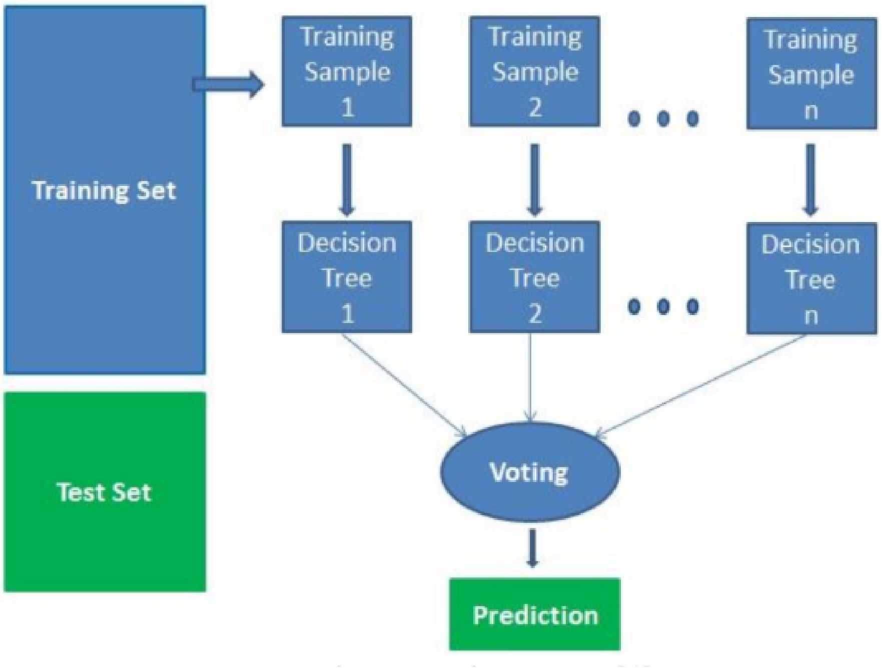


Figure 4.2: Node-RED flow

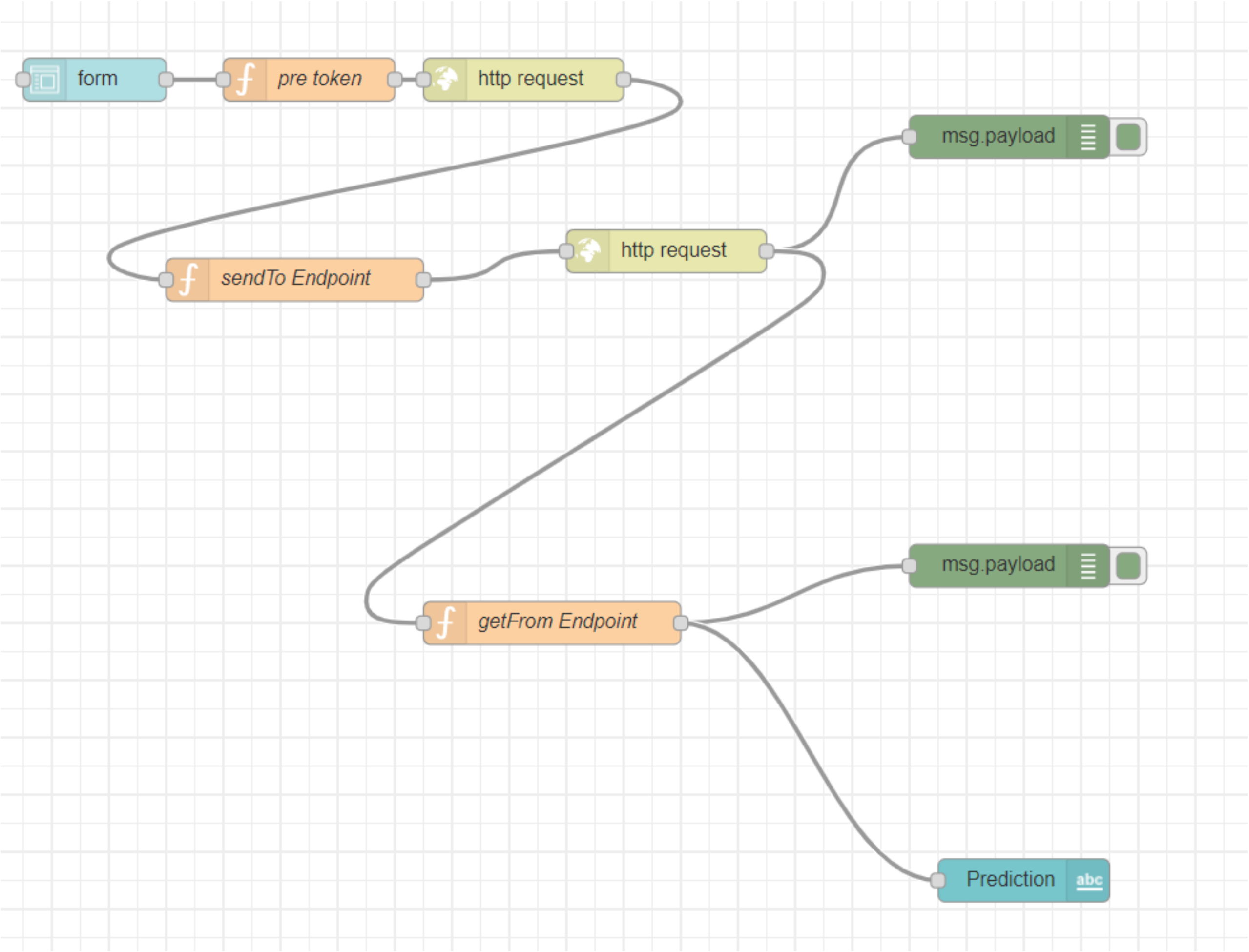
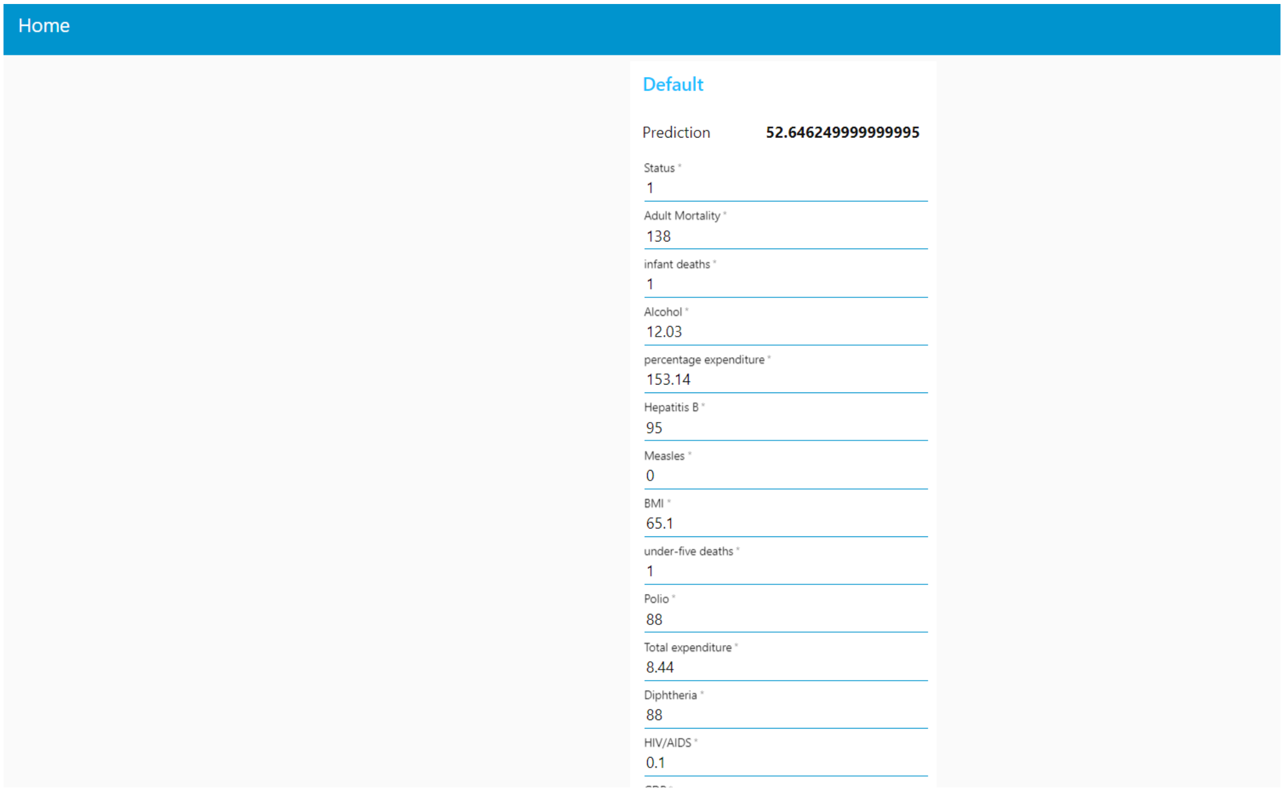


Figure 4.3: Form of the ML Model



Default

Prediction **52.646249999999995**

Status *
1

Adult Mortality *
138

infant deaths *
1

Alcohol *
12.03

percentage expenditure *
153.14

Hepatitis B *
95

Measles *
0

BMI *
65.1

under-five deaths *
1

Polio *
88

Total expenditure *
8.44

Diphtheria *
88

HIV/AIDS *
0.1

GDP *

5. Overview of Internship Experience

*D*uring my internship experience with Smart Internz, I was able to develop my Machine Learning skills. It was an enriching experience as I got to work in a professional like environment. The mentors were very helpful with the webinars they conducted on how to proceed with the project.Our doubts were solved on the slack channel regularly.

I particularly found the IBM cloud experience new and useful in improving my industrial skills. Although I found the Node RED service quite challenging, I found it to be valuable in developing my front end integration skills.

This internship has given me a clearer idea on how to proceed with improving my machine learning skills and I am grateful to Smart Internz platform for letting me be a part of this initiative.