SmartIntern Internship Project Report

Predicting life expectancy using Machine Learning

Name – Kanishk Gupta
Email kanishkgupta2000@gmail.com

1. Introduction

1.1 Overview

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. For example. by predicting life expectancy, we can analyze the aggressiveness of any disease.

The main challenge of the project is the deployment of the model onto an interactive web portal, For this purpose IBM Watson Machine Learning Service has been used along with the Node-Red integration provided by IBM cloud, we review our dataset, find what needs to be taken care off in regards to it’s cleaning, then we introduce required pipelines for the dataset that involve imputing, scaling and normalization on the dataset as well as when input is passed through an application to this model as a scoring service.

1.2 Purpose

The purpose of this project is to find the life expectancy of a person of a person from a given country, the prediction is done on various socio-economic factors such as GDP, Education, spread of diseases, expenditure and so on and so forth. As a part of the model development we first dive into correlations between these factors and Life Expectancy and see as a beginner machine learning engineer how data makes logical coherence and the correlations seen in the data are somewhat coherent to our expectations. Although the end result for demonstration purposes is a Web Page with a form to input those features, since we have built an endpoint this machine learning model, which has been enabled as a service by IBM Watson Machine Learning Service can also be used in Geographic applications and Medical Research Institutions. We have treated it as a classic regression problem.

2. Literature Survey

2.1 Existing problem

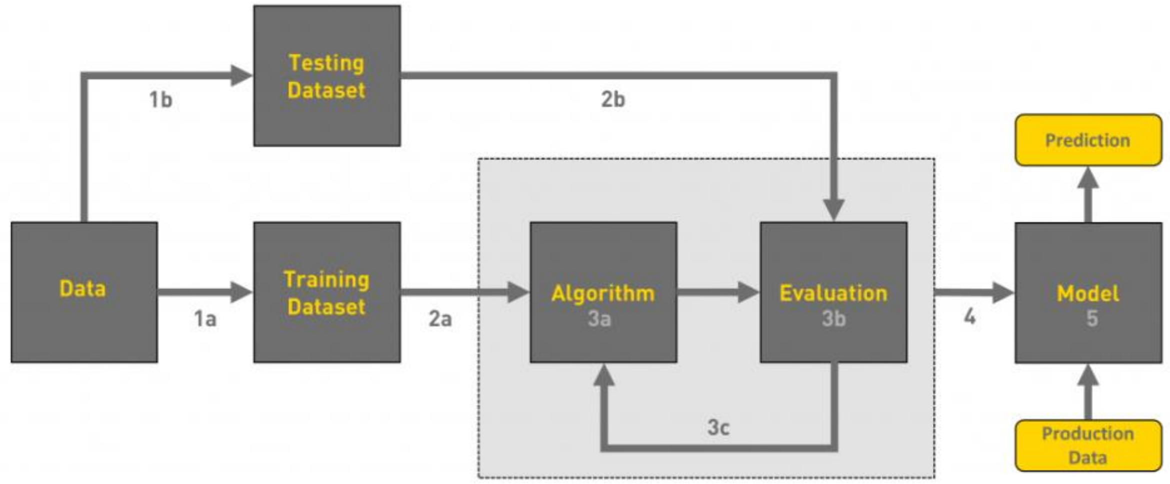
While healthcare has advanced rapidly in the past few decades , it is a strong believe that the recent rapid developments in the arena of Machine learning has paved way for Potential development in Health Services, Life Prediction is not only a hobby project, it’s a project that gives us an introduction to the power of Machine learning and what all is possible with further development.

2.2 Proposed Solution

My proposed solution is to provide a user-friendly platform to enable any person to interact with a Life expectancy model. The focus of the project will be first preparing a dataset by cleaning it, replacing null values and convert into numerical features completely. This should be done via building Pipelines that employ column transformers. The next step is the model metric selection followed by comparison of different Regression models by training them on the prepared dataset. This is followed by Deployment of the model as a service on IBM Watson Client. Then we develop a Node- Red Application using IBM cloud integration and use the scoring endpoint obtained by model deployment as an API on our Node-Red App

3. Theoretical Analysis

3.1 Block Diagram

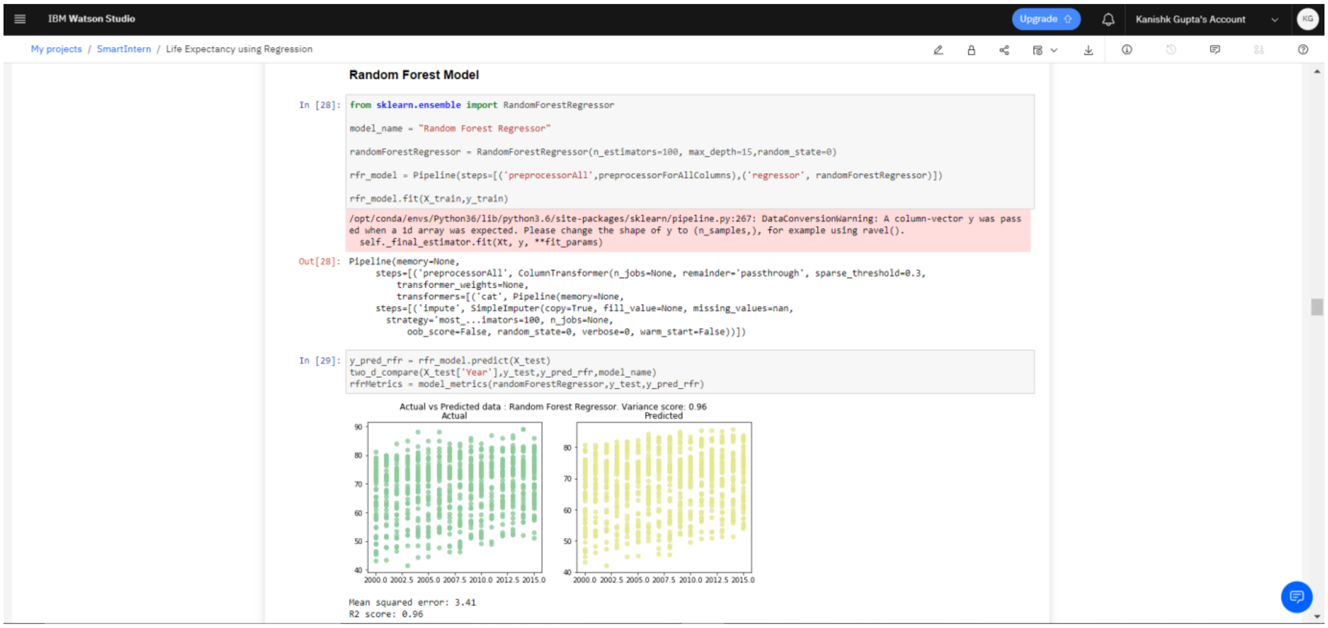


Overview of the Workflow of ML

3.2 Hardware/Software designing

1. Python notebook containing all the code.
2. A node red application which can input data and outputs a prediction for life
3. expectancy.
4. A json file containing the architecture of node red project.
5. URL of the node red application.

Jupyter notebook :

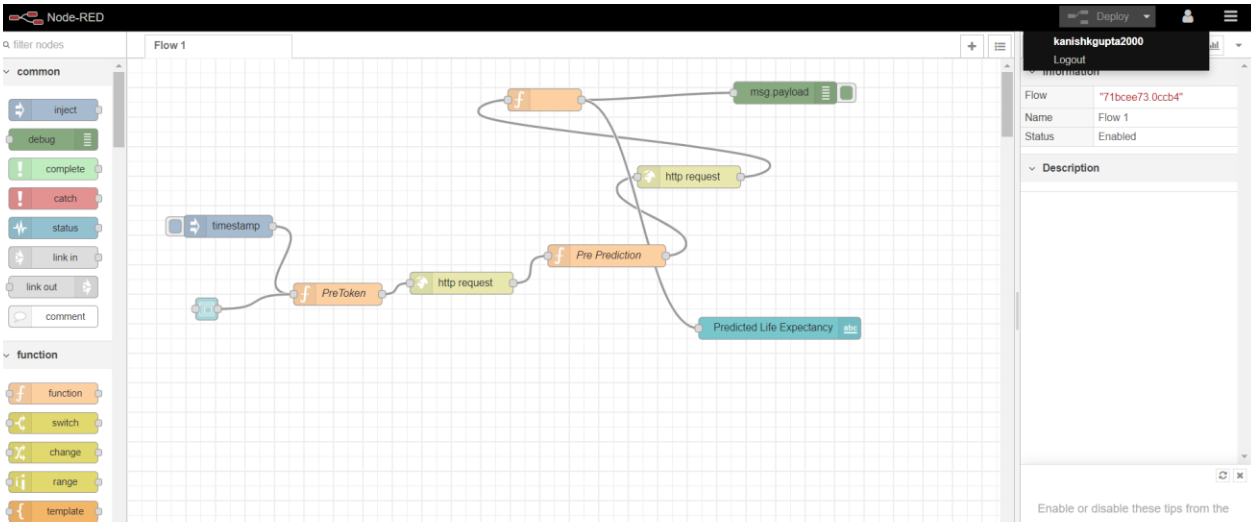

IBM Watson Studio
Upgrade
Kanishk Gupta's Account
KG
My projects / SmartIntern / Life Expectancy using Regression

Random Forest Model

```
In [28]: from sklearn.ensemble import RandomForestRegressor

         model_name = "Random Forest Regressor"

         randomForestRegressor = RandomForestRegressor(n_estimators=100, max_depth=15,random_state=0)

         rfr_model = Pipeline(steps=[('preprocessorAll',preprocessorForAllColumns),('regressor', randomForestRegressor)])

         rfr_model.fit(X_train,y_train)

         /opt/conda/envs/Python36/lib/python3.6/site-packages/sklearn/pipeline.py:267: DataConversionWarning: A column-vector y was pass
         ed when a 1d array was expected. Please change the shape of y to (n_samples,), for example using ravel().
           self._final_estimator.fit(Xt, y, **fit_params)

Out[28]: Pipeline(memory=None,
              steps=[('preprocessorAll', ColumnTransformer(n_jobs=None, remainder='passthrough', sparse_threshold=0.3,
                  transformer_weights=None,
                  transformers=[('cat', Pipeline(memory=None,
              steps=[('impute', SimpleImputer(copy=True, fill_value=None, missing_values=nan,
                strategy='most_...imators=100, n_jobs=None,
                    oob_score=False, random_state=0, verbose=0, warm_start=False))])

In [29]: y_pred_rfr = rfr_model.predict(X_test)
         two_d_compare(X_test['Year'],y_test,y_pred_rfr,model_name)
         rfrMetrics = model_metrics(randomForestRegressor,y_test,y_pred_rfr)
```

```
Mean squared error: 3.41
R2 score: 0.96
```

Node RED :

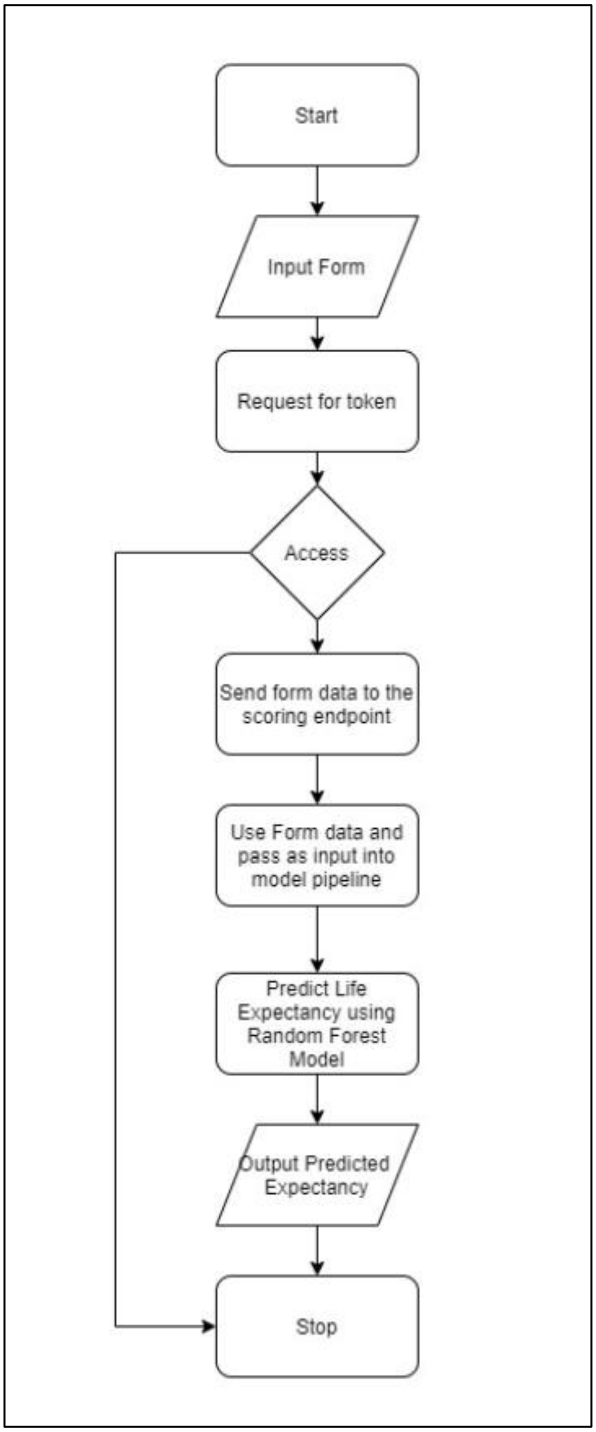


kanishkgupta2000
Logout

Flow "71bcee73.0ccb4"
Name Flow 1
Status Enabled

Enable or disable these tips from the

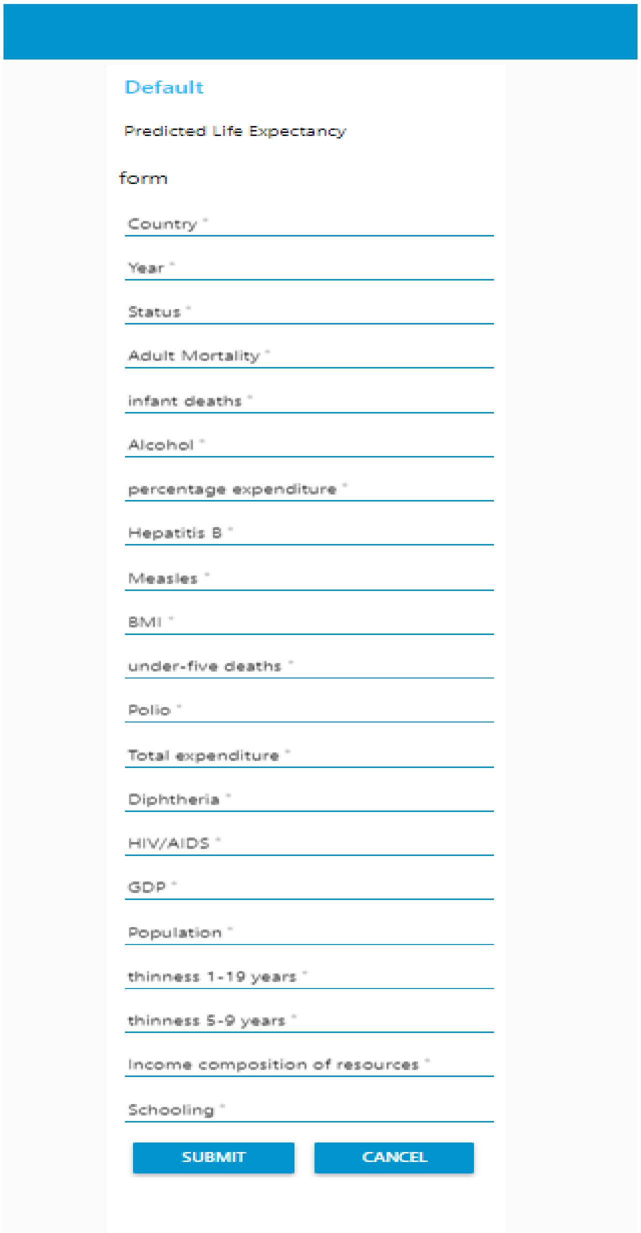
4. Flow Chart


Start
Input Form
Request for token
Access
Send form data to the scoring endpoint
Use Form data and pass as input into model pipeline
Predict Life Expectancy using Random Forest Model
Output Predicted Expectancy
Stop

5. Result

As illustrated in the provided Jupyter Notebook, we have prepared different machine learning regression models on the preprocessed clean dataset. We recorded the R2 score and the Mean Square Error of machine learning models such as Multiple Linear Regression, Elastic Net Regression, Random Forest Model and Ridge Regression. Among the four, the R2 score and the Mean Square Error was optimal for Random Forest Model, the results can also be seen via the screenshot of the Jupyter Notebook above.

The screenshot of the Node RED website deployed using Watson Machine learning service:



6. Applications

We can assume various uses of the given model, it can be used to raise awareness among individuals by creating an interactive portal that could show them how their lifestyle choices are correlated to their life expectancy, not only by the virtue of the country they live in but how they respond to their country's circumstances. For instance, our model can provide an average life expectancy of a country, but the feature correlation of alcohol intake depends on personal choices. The kind of education index our country stands cannot define us if we choose to take care of the socio-economic factors that shape our life. This dataset was retrieved via World Health Organization (WHO) , so I believe the core motive was to motivate life expectancy studies and in-depth correlations between these factors. Hence this model is a step forward in raising interest in academic circles.

7. Future Scope

Given the recent times of COVID-19, there are high chances that the infection is here to stay for a long time, in the coming times, the correlations observed today may not be relevant, such is the impact od COVID-19, I believe that in the future scope we could further explore appropriate model deployment and setup pipelines for model retraining in order to improve upon our understanding of correlation factors between the virus spread, the socio-economic factors that might show a strong connection with the spread and the ability to see what age groups is the spread infecting and it's impact on life expectancy.

8. Conclusion

We conclude that Machine learning models like the Random Forest Model is a trusted choice for Regression problems as vouched for by many machine learning experts. We explored as a Machine Learning Engineer the arena of Professional deployment on cloud services , a skill that is very important for every person interested in data science. We understood how Life Expectancy is dependent on a nation's features and how much impact our lifestyle and our way of life can bring on our being.

9. Advantages/Disadvantages

1. The advantage of this project is that it enables us to learn a boilerplate method to build models and to deploy them on industry leading platforms, and help us make sure that the models we build in code actually see their way into meaningful projects.

2. Currently the application built has the flaw that it is a very big form that take all features, it might be good as an academic and an exploration project but practically a better interface could be implemented.

10. Bibliography

1. Project planning and kickoff :

a.

https://www.allbusinesstemplates.com/download/?filecode=2KBA4&lang=en&iuid=9f9faa699fab-40ee-8457-ea0e5df8c8de

b.

https://www.youtube.com/watch?v=LOCkV-mENq8&feature=youtu.be

c.

https://github.com/

d.

https://www.zoho.com/writer/help/working-with-text.html

2.

IBM Cloud :

a. https://my15.digitalexperience.ibm.com/b73a5759-c6a6-4033-ab6b-d9d4f9a6d65b/dxsites/151914d1-03d2-48fe-97d9-d21166848e65/

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

c. https://www.w3schools.com/howto/howto_make_a_website.asp

3. IBM Watson services :

a.

https://www.ibm.com/watson/products-services

b.

https://developer.ibm.com/technologies/machine-learning/series/learning-path-machine-learning-for-developers/

c.

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

4. Dataset:

a.

Life Expectancy (WHO)

Appendix

Notebook: https://eu-gb.dataplatform.cloud.ibm.com/analytics/notebooks/v2/9dedc2e9-2a4d-46c4-80a7-d3cc050e06fd/view?projectid=4b1f34fd-f164-4e2d-9a19-1c88a949d0ae&context=wdp

Node RED: https://node-red-vivjj.eu-gb.mybluemix.net/red/#flow/71bcee73.0ccb4

Form link: https://node-red-vivjj.eu-gb.mybluemix.net/ui/#!/0?socketid=WjsphCuuq8i_4dSOAAAA

Github Link; https://github.com/SmartPracticeschool/llSPS-INT-1921-Predicting-Life-Expectancy-using-Machine-Learning