PROJECT REPORT

PREDICTING LIFE EXPECTANCY USING MACHINE LEARNING

INTRODUCTION

1.1 OVERVIEW

Life Expectancy of a person is a statistical measure of the average time he/she is expected to live, based on several factors. Life Expectancy depends upon factors such as: Regional Variations, Economic Circumstances, Sex Differences, Mental Illnesses, Physical Illnesses, Education, Year of Birth and various other Demographic Factors.

Calculating Life Expectancy will require a sophisticated system that will take into account all these factors. With Machine Learning and services such as IBM Watson this has now become possible.

The goal of this project is to predict the Average Life Expectancy of people living in a country given certain factors consisting of Social, Economical, Immunization, Mortality Factors.

1.2 PURPOSE

The purpose of the project is prediction of Life Expectancy of people living in a country given the features: Year, Country Status, Adult Mortality, Infant Deaths,Under-five Deaths, Alcohol, Percentage Expenditure, Total Expenditure, Income Composition of Resources, Measles, Polio, Diphtheria, HIV/AIDS, GDP, Population, BMI, Thinness 1-19 years, Thinness 5-9 years, Schooling.

Life Expectancy Prediction will have certain benefits like policy making, optimizing individual's health etc. This will motivate countries and their citizens to make healthier lifestyle decisions.

LITERATURE SURVEY

2.1 EXISTING PROBLEM

Life expectancy is the key metric for assessing population health. Broader than the narrow metric of the infant and child mortality, which focus solely on mortality at a young age, life expectancy captures the mortality along the entire life course.

There is a need of an application which will predict the Life Expectancy given Social, Economical, Mortality and Immunization Factors. The application must give prediction accurately, within fractions of second.

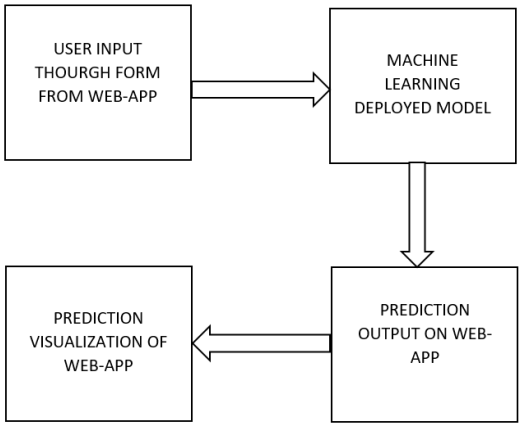
2.2 PROPOSED SOLUTION

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.

For the Life Expectancy prediction, user will have to give input values of various features and after analyzing these features, the Machine Learning model will pass its prediction to the user and also indicate whether the predicted value is Safe Life Expectancy Rate, or the country needs to improve its factors.

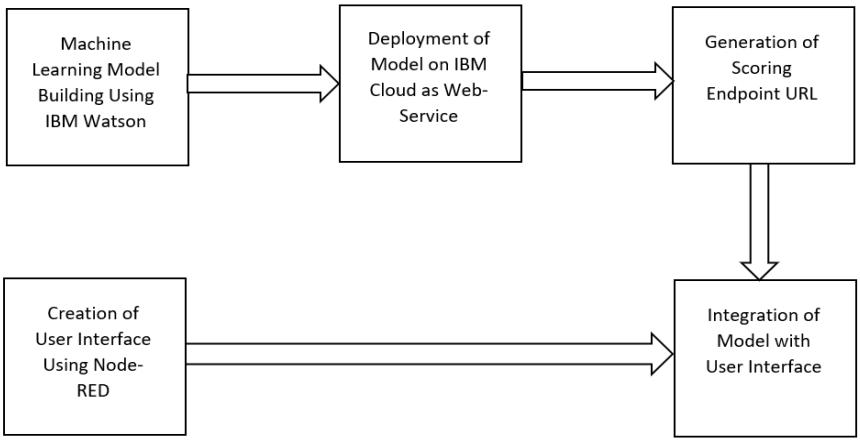
THEORETICAL ANALYSIS

3.1 BLOCK DIAGRAM



3.2 HARDWARE / SOFTWARE DESIGNING

There is no Hardware involved as such. The web-app has IBM Watson Machine Learning as its backend service and Node-RED for the User Interface development. The web-app has been deployed on IBM Cloud as a Web Service. Basic Flow is as follows:



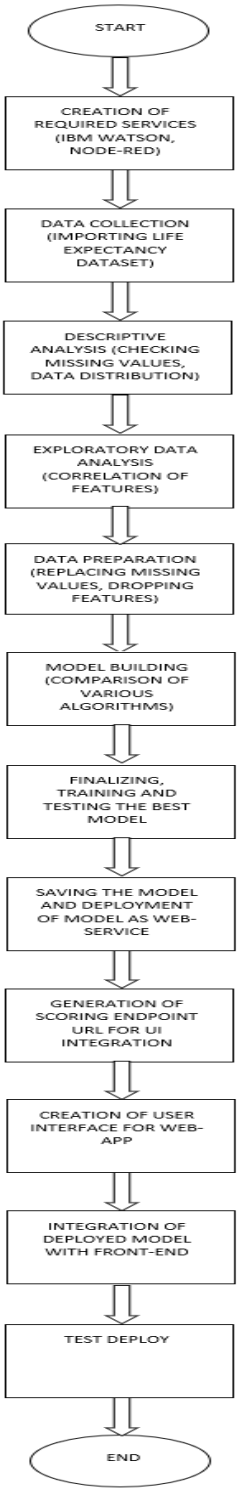
EXPERIMENTAL INVESTIGATIONS

After the collection of data, a Descriptive Analysis was performed which revealed that the data set had many missing values in the features. There were some missing values for the output column (Life Expectancy) as well. The rows having missing values from the outcome variable column were dropped from the data frame as the model will not be able to learn from these missing values.

The missing values from input variables were filled by the mean values of the respective feature for each country. The reason behind this approach is that there were many missing values in the features and if all values would have been replaced by the mean of the whole column, there would not have been any significance of that column. Thus, the mean values for features for each country replaced these missing values. The countries which did not have any value present for that feature were replaced by mean value of the whole feature.

While doing Exploratory Data Analysis, it was noticed that the feature named Population did not have a strong correlation with the target variable. Hence, that feature was dropped from data frame. While experimenting with different algorithms in search of the one which gives better R2 score and minimum RMSE (Root Mean Square Error), ExtraTreesRegressor out-performed all the algorithms. Further study was done by using Scaling Techniques such as MinMaxScaler and StandardScaler resulting use of Standard Scaler with ExtraTreesRegressor outputted the best outcome.

FLOWCHART


START
CREATION OF REQUIRED SERVICES (IBM WATSON, NODE-RED)
DATA COLLECTION (IMPORTING LIFE EXPECTANCY DATASET)
DESCRIPTIVE ANALYSIS (CHECKING MISSING VALUES, DATA DISTRIBUTION)
EXPLORATORY DATA ANALYSIS (CORRELATION OF FEATURES)
DATA PREPARATION (REPLACING MISSING VALUES, DROPPING FEATURES)
MODEL BUILDING (COMPARISON OF VARIOUS ALGORITHMS)
FINALIZING, TRAINING AND TESTING THE BEST MODEL
SAVING THE MODEL AND DEPLOYMENT OF MODEL AS WEB-SERVICE
GENERATION OF SCORING ENDPOINT URL FOR UI INTEGRATION
CREATION OF USER INTERFACE FOR WEB-APP
INTEGRATION OF DEPLOYED MODEL WITH FRONT-END
TEST DEPLOY
END

RESULT

Removal of feature Population yielded better results. Replacement of missing values enabled the model to learn more examples and generalize well.

ExtraTreesRegresssor was the top-performer in all the experiments. The R2 score obtained was 0.964 and RMSE was 1.771.

The results were tallied with the model created using AutoAI experiment. The AutoAI model gave R2 score as 0.961 and RMSE of 1.830.

Metrics	Machine Learning Model	AutoAI Model
R2 Score	0.9635483479199352	0.96050013469328
RMSE	1.7714090404600051	1.829561183885059

The testing results convinced that the trained model has generalized well and can give better predictions to a great extent.

ADVANTAGES & DISADVANTAGES

The following can be the Advantages:

- Predicting Life Expectancy has the potential to benefit individuals, health service providers and governments.
- The web-app provides user-friendly interface, it also displays the necessary information about the app so that the user understands what he has to do and how the app will work accordingly.
- The prediction is returned to user in a short-time and user is intended about the prediction through notification. The prediction is also outputted in the web-app and a visual representation about the same is shown.

The following can be the Disadvantages:

- The web-app is not handy to use for users as it is not a mobile app.
- The user has to input all the fields and only then the prediction will be given to the user. However, inputting all these values is a tiresome job, and the user cannot input these values through speech.

APPLICATIONS

The following can be the Applications:

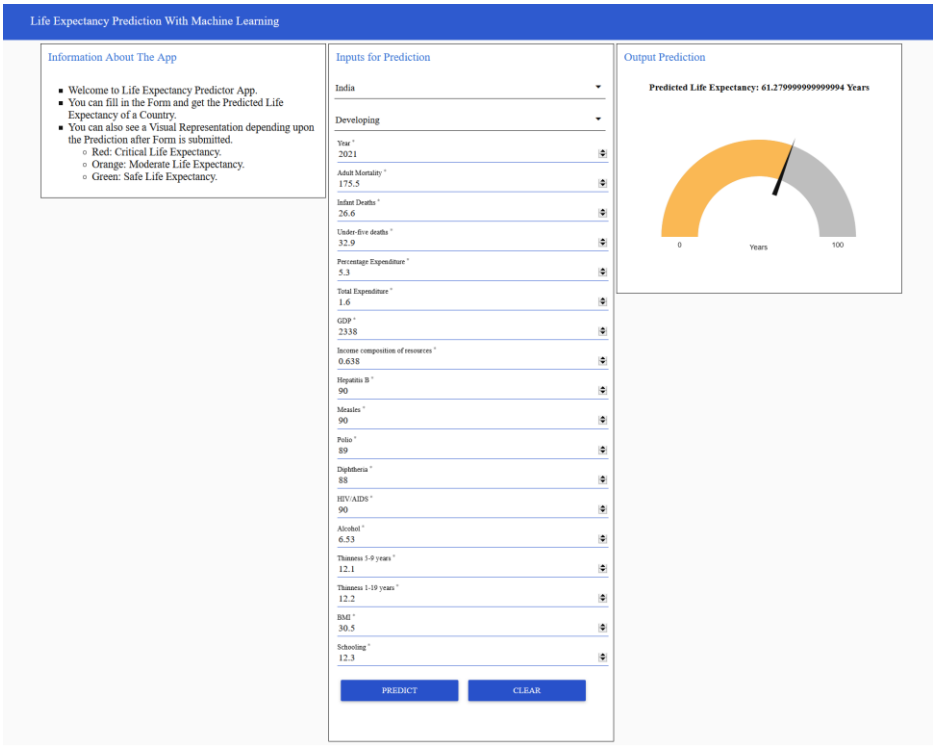
- People can become more aware of their general health and improvise on it. Insurance Companies can use these predictions to provide individualized services.
- Governments can use predictions to allocate limited resources efficiently. Social welfare, Health-care funding to individuals and in areas of greater needs can be assigned effectively.
- It can benefit for policy making, and help optimize an individual’s health, or the services they receive.

CONCLUSION

The web-app has been built successfully yielding good results. It is well-integrated with IBM Watson and Node-RED. The web-app is available at the following link:

https://life-expectancy-prediction-app.mybluemix.net/ui/

There still can be more improvements in the model by doing Feature Engineering, creating a Mobile App instead of a web-app, Enabling speech input to he model for better convenience.



FUTURE SCOPE

The following can be the Improvements and Future Scope:

- Making the model more accurate, doing more relevant feature extraction.
- Creating a mobile app for more convenience.
- Modifying the model so that it will accept speech-driven inputs too.

BIBLIOGRAPHY

- Nick O'Leary, Create a Node-RED starter application, IBM,https://developer.ibm.com/tutorials/how-to-create-a-node-red-starter-application/, 25 May 2020.

- Kumar Rajarshi, Life Expectancy (WHO), https://www.kaggle.com/kumarajarshi/life-expectancy-who, 03 June 2020.

- IBM, Chapter 4 Build and Deploy models in Jupyter Notebooks,https://bookdown.org/caoying4work/watsonstudio-workshop/jn.html, 07s June 2020.

- G D Abhishek, Deploy Machine Learning (scikit-learn) Models in IBM Cloud — Watson Studio, https://medium.com/@gdabhishek/step-by-step-procedure-to-deploy-machine-learning-scikit-learn-models-in-ibm-cloud-watson-de1ff8be92de, 08 June 2020.

- Richard Hagarty, Automate model building in IBM Watson Studio, https://developer.ibm.com/tutorials/watson-studio-auto-ai/, 10 June 2020.

- Misha Ketchell, Don’t die wondering: apps may soon be able to predict your life expectancy, but do you want to know?, https://theconversation.com/dont-die-wondering-apps-may-soon-be-able-to-predict-your-life-expectancy-but-do-you-want-to-know-129068, 13 June 2020.

APPENDIX

A. SOURCE CODE

- The source code for IBM Watson Project and Node-RED Flow can be found at: https://github.com/SmartPracticeschool/llSPS-INT-1934-Predicting-Life-Expectancy-using-Machine-Learning
- The web-app can be used at: https://life-expectancy-prediction-app.mybluemix.net/ui/