Machine Learning Engineer Nanodegree

Capstone Proposal

P.T.V.KRISHNA

June 27th, 2018

Life Expectancy of a country By WHO

Domain Background

- Although there have been lot of studies undertaken in the past on factors affecting life expectancy considering demographic variables, income composition and mortality rates, it was found that effect of immunization and human development index was not taken into account in the past.
- Also, some of the past research was done considering multiple linear regression based on data set of one year for all the countries. Hence, this gives motivation to resolve both the factors stated previously by formulating a regression model based on mixed effects model and multiple linear regression while considering data from a period of 2000 to 2015 for all the countries. Important immunization like Hepatitis B, Polio and Diphtheria will also be considered.
- In a nutshell, this study will focus on immunization factors, mortality factors, economic factors, social factors and other health related factors as well. Since the observations this dataset are based on different countries, it will be easier for a country to determine the predicting factor which is contributing to lower value of life expectancy. This will help in suggesting a country which area should be given importance in order to efficiently improve the life expectancy of its population.

Problem Statement

- By accurately predicting the 'Life expectancy' of a country based on various factors which are correlated.
- By using the data-set on Life Expectancy, the task is to predict the accuracy of 'Life Expectancy' of a country based on various factors mentioned in the dataset. The model utilizes the important characteristics of the data to develop models that can predict the target variable.
- I decided to implement Machine-Learning techniques to predict the 'Life Expectancy' of a country by using various factors like BMI, infant and adult mortality rate, diseases, health care improvement rate etc..
- This Project constitutes Data Exploration and Visualizations, Data Preprocessing and finally testing various algorithms and techniques.

Datasets and Inputs

- The data was collected from WHO and United Nations website with the help of Deeksha Russell and Duan Wang. It was collected from kaggle website.
- The data-set constitutes a single CSV files which is obtained from Kumarrajarshi from Kaggle
- Link :- https://www.kaggle.com/kumarajarshi/life-expectancy-who/home
- The data has 22 columns i.e., they are Country, Year, Status, BMI, Mortality rate, Hepatitis etc. and the target variable being 'Expectancy' and has 2400 rows. • We are comprised with a good amount of features which are potentially utilized to estimate good result.

Solution Statement

- The brighter solution to this problem, is drawn from various features of the individual countries and the data, that can be efficiently used obtain the life expectancy in a country based on its input features.
- This potential data can be helpful for testing the supervised machine learning models to predict or estimate the life expectancy of countries.
- I may use models such as SVM's,Decision Trees and Random Forests etc., along with GridSearchCV for Model optimization.

Benchmark Model

- Since the given problem expects to predict a continuous output,it is more obvious to determine a metric value(R2_score) that will potentially help us to establish a comparison between the performances of the Bench Mark Model and the Optimal Model thus selected.
- The BenchMark model is a base model that potentially has no intelligence, hence we start with finding the r2_score that provides an indication of the goodness of fit of a set of predictions to the actual values.

Evaluation Metrics

- The current problem is a regression problem, since it takes certain features as inputs and tries to figure out a life expectancy(in years) that will be helpful for an individual country to determine the health care expenditure by its government to keep up its standards in healthcare.
- Hence, I've decided to use 'Co-Efficient Of Determination' as the performance metric that could be applied to check the performance of the life expectancy obtained from the Bench Mark Model and the Optimal Model considered.

Project Design

- Data Acquisition :- Indeed the first step of the project is acquiring data which I collected from the Kaggle.
- Data Exploration and Visualization :- Exploring and visualizing the dataset, cleaning and observing the relevance of every feature.
- Data Pre-Processing :- Data Preprocessing is the key factor in Machine-Learning, to clean and remove any irrelevant data.
- Model Evaluation and Validation :- Model is tested for the performance in it's BenchMark State and the Optimized State respectively.
- Optimization :- The naive model is optimized by the application of GRIDSEARCHCV which helps in tuning the parameters optimal for the model if I take any one of them