LIFE EXPECTANCY PREDICTION

AIM:
To construct a life expectancy prediction model and build a UI to display the results.

ABOUT THE DATASET

The World Health Organization (WHOGlobal)'s Health Observatory (GHO) data repository monitors each nation's health status as well as a number of other relevant variables. For the goal of analysing health data, the datasets are made available to the general public. The same WHO data repository website served as the source for the dataset relating to life expectancy, health factors for 193 nations, and its corresponding economic statistics. Only the most representative critical factors from each category of health-related factors were selected.

DATA PREPROCESSING

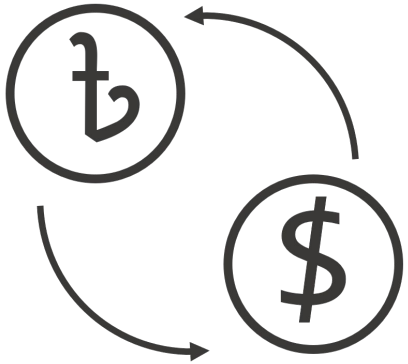
- This turned out to be the most time consuming component of our analysis due to the veracity of the data.
- The dataset had numerous null values and intercorrelated attributes which were fixed using packages and tools in python.

1. MISSING VALUE IMPUTATION

- The first problem that was faced during the EDA was the numerous cells that had no values.
- All the null cells were then filled by mean values when it came to numerical variables and mode, when it came to the categorical variables

2. Converting Categorical Variables

- One significant step when it comes to predictive and statistical analysis using machine learning is to Encode the categorical variables into numeric values
- We do this to ensure that the model doesnt get confused while dealing with the large amounts of categorical cells.



3. MULTICOLINIEARITY REMOVAL

The issue of multicollinearity was dealt with by taking Correlation between the independent variables and VIF (Variance inflation factor) into consideration.

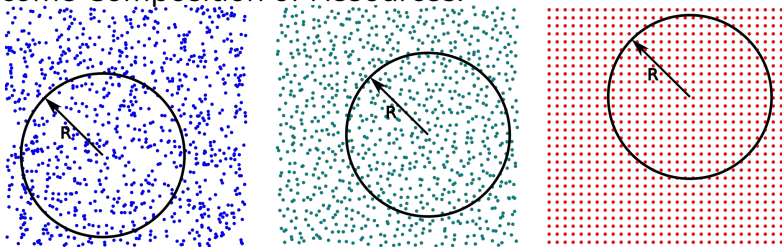
1. Checking the Correlation Matrix

   This was a significant step in order to determine how much intercorrelated the attributes of this dataset were. We constructed a Correlation Matrix for all the Xi's and built a Heatmap which gave us a visual representation for the same to check which attributes are highly correlated among themselves. Those variables which had a high correlation amongst themselves (Pearson's Correlation Coefficient >0.60) were taken forward as candidates for their removal from the dataset.

3. MULTICOLLINEARITY REMOVAL

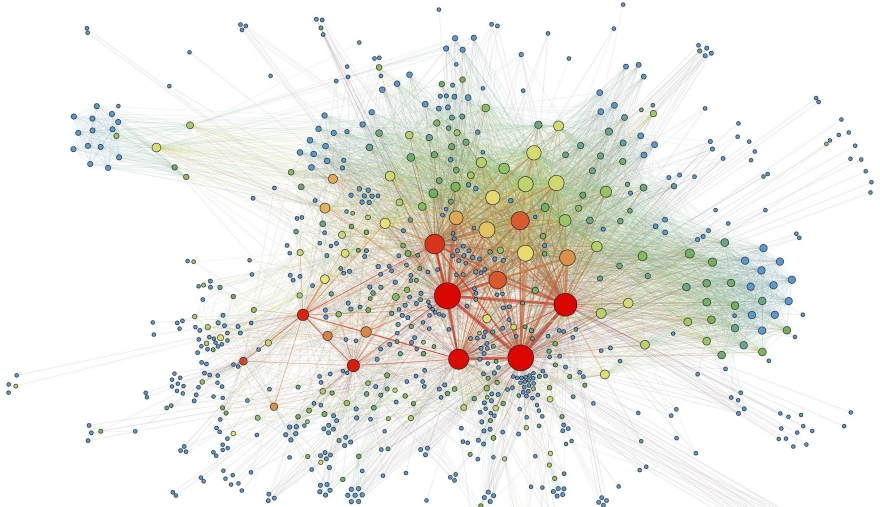
2. Checking Variance Inflation Error

Furthermore, to support our argument of removing the attributes based on the correlation matrix, we also took into account the Variance Inflation Factor of each continuous numerical attribute except the categorical variables (Countries and Status) within the dataset. Nearly all the variables which we were aiming to remove, displayed a VIF score of more than 10, which is why we could assertively remove the most intercorrelated columns. These included variables namely Hepatitis B, Diphtheria, Polio, Schooling, Infant Deaths, Thinness (5-9 years) and Income Composition of Resources.



DATA ANALYSIS

- Once our preprocessing was over, we moved on to our full-fledged data analysis
- We incorporated tens of packages in python to achieve the most accurate analysis of our dataset



1. Panel Data Regression

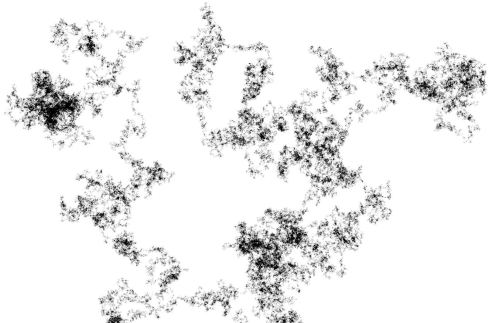
In statistics and econometrics, panel data and longitudinal data are both multi-dimensional data involving measurements over time. Panel data is a subset of longitudinal data where observations are for the same subjects each time. We observed that the data in the problem statement was a strong example of Panel Data, which is why we chose to apply two regression models that correspond to Panel Data Regression.

i. **Pooled OLS Model**

The first model that we considered under Panel Data Regression is the basic Pooled Ordinary Least Squares (Pooled OLS) model, which does not take into account the time variability within the dataset and considers the data as plain cross-sectional data.

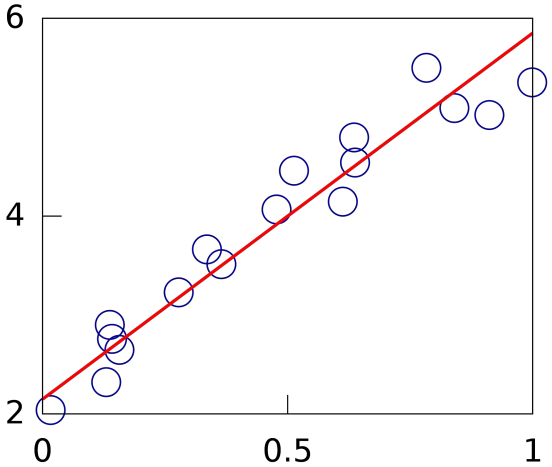
1. Panel Data Regression

Another type of Panel Data Regression model is the Random (Mixed) Effects model, which takes into account both the time series factor as well as the cross-sectional or longitudinal feature of the panel dataset. Due to this salient feature of the model, we wished to fit this model to our dataset. We tried implementing the code for this model in Python. However, due to the lack of domain expertise in deploying the model in Python as well as due to the lack of sufficient time, we weren't able to successfully complete its deployment. This can be considered as one of the limitations of the project. However, our team strongly suggests this model as one of the best alternatives for fitting the type of data (panel data) at hand.



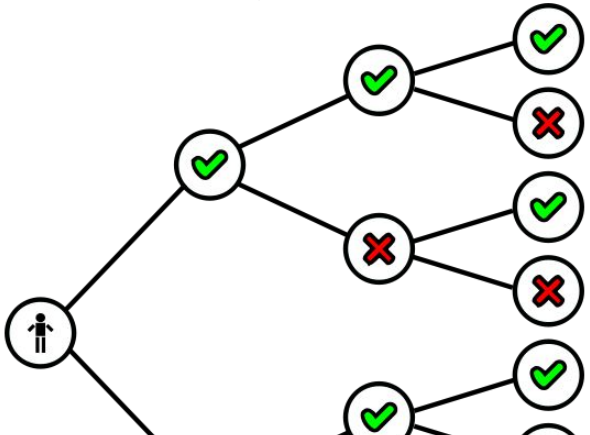
2. Linear Regression Model

We implemented the linear regression model as well, which gave us the accuracy of R^2=0.96 along with Root Mean Square Error (RMSE) of greater than 1.80 . However, we strongly suspect overfitting of the data here, as simple linear regression is not one of the best approaches to be applied on panel data, which is both cross-sectional as well as time-series.



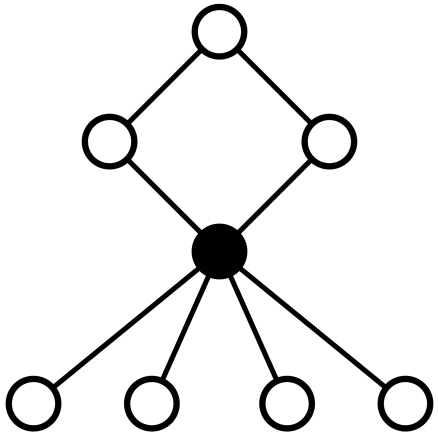
3. Decision Tree Regression Model

Next, we tried a couple of Machine Learning models to fit our dataset, out of which the first one was the Decision Tree Regression model. The model gave an accuracy of R^2= 0.92 and RMSE of 0. 2.73. However, we wanted to explore further with other different ML model types which provided higher accuracy. Hence, we went forward and finally applied the Random Forest Regression Model.



4. Random Forest Regression Model

Random Forest Regression Model is considered to provide very highly accurate results, as it takes the decision while taking into account multiple Decision Trees. This model gave the RMSE of 1.75 (which is quite low) and the accuracy of R^2=0.97, which was the highest amongst all our other models that we applied to fit our data. Hence we went forward and finalized the Random Forest Regression Model for predicting Life Expectancy.



RESULTS AND DISCUSSION

The World Health Organization (WHOGlobal)'s Health Observatory (GHO) data repository monitors each nation's health status as well as a number of other relevant variables. For the goal of analysing health data, the datasets are made available to the general public. The same WHO data repository website served as the source for the dataset relating to life expectancy, health factors for 193 nations, and its corresponding economic statistics.

We eliminated the Baby Deaths variable and kept the Under 5 Deaths variable because the data in the Under 5 Deaths variable would also account for infant mortality. The relationship between income composition and education was likewise extremely highly significant (0.80). Once more, we attempted to compute the VIF score while keeping one variable and eliminating the other. We had to eliminate both variables because the other variable (turnwise), even with one removed, still had a high VIF score. Likewise, the variable Thinness (5–9 years) was eliminated from the dataset along with the variables Thinness (1–19 years) and Thinness (5–9 years).

Coming to the Data Analysis section, the Pooled OLS Panel Data Regression Model's model accuracy was pretty considerable (R2=0.70). We did not, however, finish this model because it does not account for the time variability element.

The R2 value for the Random Forest Regression Model, which we used to fit our data, was 0.97, which was the highest of all the models we tested. The Random Forest Regression Model for Predicting Life Expectancy was thus completed.