HackVerse 2022

Team BRO_CODE

(Abhik Dey, Amrita Veshin, Atharva Vetal, Rahil Khan)

1. **Aim:** The aim of this project is to construct a Life Expectancy Prediction Model and build a UI to display our results.
2. **About The DataSet:**

   The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries The datasets are made available to public for the purpose of health data analysis. The dataset related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from United Nation website. Among all categories of health-related factors only those critical factors were chosen which are more representative. It has been observed that in the past 15 years , there has been a huge development in health sector resulting in improvement of human mortality rates especially in the developing nations in comparison to the past 30 years. Therefore, in this project we have considered data from year 2000-2015 for 193 countries for further analysis. The individual data files have been merged together into a single dataset. On initial visual inspection of the data showed some missing values. As the datasets were from WHO, we found no evident errors. Missing data was handled in R software by using Missmap command. The result indicated that most of the missing data was for population, Hepatitis B and GDP. The missing data were from less known countries like Vanuatu, Tonga, Togo,Cabo Verde etc. Finding all data for these countries was difficult and hence, it was decided that we exclude these countries from the final model dataset. The final merged file(final dataset) consists of 22 Columns and 2938 rows which meant 20 predicting variables. All predicting variables was then divided into several broad categories:Immunization related factors, Mortality factors, Economical factors and Social factors.

3. **Data PreProcessing:**
   a. **Missing Value Imputation**

      The first factor we had to deal with while cleaning the data is missing values. We observed that the dataset displayed numerous null values. For us, the factor of

empty values was a big problem so we imputed those values and they were replaced by the means of their respective columns.

b. Converting The Categorical Variables:

Another drawback that we usually face while fitting statistical or ML models to the data sometimes get 'confused' when it comes to categorical variables. Hence we replaced them with numerical values which could be easily translated by the models that we planned to apply.

c. Removing Multicollinearity From The Dataset

Another issue that is necessary to deal with is the presence of multicollinearity within the dataset. While considering linear or multilinear regression models, one of the assumptions is that the features or the independent variables (Xi's) are independent of each other. If there is any correlation between any of the Xi's, this is known as multicollinearity, which significantly affects the model and the results that are produced are biased. We dealt with multicollinearity within our dataset by dropping all those independent variables which had high correlations amongst each other and had high Variance Inflation Factor (VIF) score.

i. **Checking The Correlations Between Xi's**

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.

ii. **Checking The Variance Inflation Factor (VIF) of the Xi's:**

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.

4. **Data Analysis:**

   a. **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.

      ii. **Random Effects Model**

      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.

   b. **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.

c. **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.

d. **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.

5. Results And Discussion

Starting with the correlation matrix of the independent variables (Xi's), variables Hepatitis B and Polio displayed high correlations (0.61 and 0.67 respectively) with Diptheria. We tried removing only selective variables out of these three and keeping the rest within the dataset. However, the VIF score was still significantly high (>>>10) in all the cases, which is why we were forced to remove all of these 3 variables. Similarly, the variable Infant Deaths showed a perfect positive correlation with the Under 5 Deaths variable (Pearson's coefficient=1). Since the data within the Under 5 Deaths would also account for the infant mortality, hence we removed the Infant Deaths variable and kept the Under 5 Deaths variable. The correlation of Income Composition with Schooling was also quite significantly high (0.80). Again, we tried removing only one of the variables and keeping the other and then computing the VIF score. Since the VIF score of the other variable (turnwise) was still coming high despite the removal of one variable, hence we had to remove both. Similarly, out of the variables Thinness (1-19 years) and Thinness (5-9 years), the latter was removed from the dataset.

Coming to the Data Analysis part, the model accuracy produced by the Pooled OLS Panel Data Regression Model was quite significant ($R^2=0.70$). However, since the model does not take into account the time variability factor, hence we didn't finalize this model.

For the Random Forest Regression Model, $R^2=0.97$, 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.