Exploring & Predicting Life Expectancy to Improve Policy Decisions

Project Report

Group VIII

Introduction:

* We worked on the Life expectancy dataset related to life expectancy, health factors for 138 countries from year 2000 to 2014.
* Life expectancy is one of the key measures of a population’s health, and an indicator used widely by policymakers and researchers to complement economic measures of prosperity, such as GDP per capita.
* The dataset focuses on immunization factors, mortality factors, economic factors, social factors and other health-related factors as well.

Data Manipulation:

* The dataset downloaded from Kaggle had missing data, corrupted data - data for few data points were copied incorrectly. So, we decided to create dataset from the scratch using the same features from the Kaggle dataset. We collected data from World Bank and United Nations Development Programme and prepared a dataset.
* Few lesser known countries like Sudan, Virgin Islands (U.S.) etc. had missing data for 2000 to 2014 so we had to remove them from the dataset.
* Then we used *SimpleImputer* from the *sklearn* library to replace NaN values using the "mean" strategy. Each country from 2000 to 2014 with the column where null values are present is given as a parameter to the "fit\_transform(...)" function. Mean of each column is calculated and replaced in place of NaN values.
* In the end, we got a cleansed dataset which was stored as a pickle file.

Data Visualization:

* We used different visualization techniques and got the following insights:
  + Life expectancy depends on important predictors such as adult mortality, GDP, total expenditure, Polio vaccination.
  + We can conclude that low-income countries have the least life expectancy and as the status of a country increases, the life expectancy also increases.
  + Predictors like Polio, BMI, GDP are highly correlated with Life Expectancy where as Adult Mortality is negatively correlated with Life Expectancy.
* Through the project we learned how to choose the best visualization technique to represent the different dataset and how important the visualization understand the data. Understand the importance of interactive plots and tried to implement in my project.

MultiLinear Regression:

* After checking for VIF, multicollinearity issue we removed few predictors in the final regression model such as Thinness\_10-19\_years, Thinness\_5-9\_years, GDP, Diphtheria and Measles which do not have statistically significant effects on Life Expectancy (p-values > 0.05)
* We also removed Infant Deaths, High income and BMI 18+ years which have multicollinearity problem.

Prediction Models:

* We predicted Life expectancy using different prediction models and chose the best model by observing the performance metrics.
* By using this model, we can explain or predict the Life expectancy from the other factors it is dependent on.
* We took the performance metrics( mean absolute error, mean squared error and r-squared error) to check the accuracy of the model.
* For prediction and performance measurement, we used regressor models instead of classification models because our variable ‘Life expectancy’ to be predicted is a continuous response variable.
* We used 6 different regressor models to predict life expectancy. They are Decision tree, Random forest, XGBooster, Catboost regressor, K-NN regressor, and Gradient Boosting regressor.
* We have tuned hyperparameters for each model to get the best performance metrics.
* Our best model with the best performance metrics is the XGbooster regressor model with hyperparameter tuning.
* The performance metrics for the best model are:

MAE value = 0.387982

MSE value = 0.287131

R2 error value = 0.535846