LIFE EXPECTANCY PREDICTION OF A GIVEN POPULATION USING MULTILINEAR REGRESSION MODEL

Business Understanding

Life expectancy prediction is the act of trying to estimate the average age that an individual of a given age is expected to live. Certain factors can either increase or reduce the lifespan of the said individual.

In this research, we seek to build a predictive model that will help us in estimating that age based on certain criteria such as their BMI, number of schooling years, whether they are from developed or developing countries and if immunisation also plays a key role in affecting life expectancy.

Research Question

Build a multilinear regression model predicting the life expectancy of the given population.

Objectives

Main Objective

To build a multilinear regression model that will predict the life expectancy of a given population.

Specific Objectives

- To investigate if immunisation has an impact on life expectancy
- To determine if schooling affects life expectancy
- To determine whether adult mortality rates affect life expectancy

Data Understanding

Data Source

Data is downloaded from Kaggle website.

Data Description

1. Country: Name of the country listed
2. Status: Developed or Developing status
3. Life expectancy: Life Expectancy in age
4. Adult Mortality: Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
5. Infant Deaths: Number of Infant Deaths per 1000 population
6. Alcohol: Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
7. Percentage Expenditure: Expenditure on health as a percentage of Gross Domestic Product per capita(%)
8. Hepatitis B: Hepatitis B (HepB) immunisation coverage among 1-year-olds (%)
9. Measles: number of reported cases per 1000 population
10. BMI: Average Body Mass Index of entire population
11. Under-five Deaths: Number of under-five deaths per 1000 population
12. Polio: Polio (Pol3) immunisation coverage among 1-year-olds (%)
13. Total expenditure: General government expenditure on health as a percentage of total government expenditure (%)
14. Diphtheria, tetanus toxoid, and pertussis (DTP3) immunisation coverage among 1-year-olds (%)
15. HIV/AIDS: Deaths per 1000 live births HIV/AIDS (0-4 years)
16. GDP: Gross Domestic Product per capita (in USD)
17. Population: Population of the country
18. Thinness 1-19 years: Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
19. Thinness 5-9 years: Prevalence of thinness among children for Age 5 to 9(%)

20. Income composition of resources: Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
21. Schooling: Number of years of Schooling(years)
22. Year: Year listed in the dataset

Data Preparation

Loading Libraries

Load the libraries necessary for cleaning and analysis

Loading the Data

Load the dataset from the CSV file. The name of the CSV file is "Life Expectancy Data.csv".
The shape of the dataset is 2938 by 22 (2938 rows and 22 columns)

Cleaning the Data

Data cleaning is the process of preparing data for analysis by weeding out information that is irrelevant or incorrect. This type of information typically reinforces a false belief, which might have a negative effect on the model or algorithm it is given into.

The steps for the data cleaning process is:

1. Consistency - Ensure there are no duplicates in the data.

2. Uniformity - Ensure the data types for the datasets are accurate.

3. Completeness - Ensure the dataset has no missing values.

Modelling

In model building the target variable to be used is life expectancy. The predictor variables are:

1. Schooling: since it is the one highly correlated with life expectancy, it is going to be the base model.
2. Adult Mortality: this feature has a high negative correlation with life expectancy and we would like to see how it affects the predictor variable.
3. Diphtheria: we want to see the effect of immunisation on life expectancy.
4. Status: this will serve as our categorical column. We want to compare how the two statuses differ from each other in terms of life expectancy.
5. BMI(Basal Metabolic Index): we also want to check the effect of one's BMI on their life expectancy.

Baseline model was constructed using simple linear regression and the output displayed.

The aforementioned characteristics were used to create a multilinear regression model. In order to prevent multicollinearity, the categorical column (status) is encoded using binary encoding and one column is removed.

Because the fitted model's condition number was too high, linear transformation was necessary. This technique combines centering and scaling by standardising (zero centering). A final model will be fitted, and the results will be shown as well. The intercept and coefficients are statistically significant and the condition number has greatly reduced from the previous model, thus we can say it is a better model compared to the first. The RMSE (Root Mean Square Error) is also calculated to measure average error of the model.

Conclusion and Recommendation

Conclusion

- The number of years one spends in school has a positive increase in their life expectancy
- Getting immunised increases one's life expectancy
- The BMI range of a person also directly affects their life expectancy. People who are in the middle range, meaning neither too thin nor too fat have longer life spans.
- Developed Nations have better health care expenditure and thus their life expectancy is higher compared to the developed countries.
- Developing countries have a lower life expectancy as compared to developed countries. This can be attributed to civil wars and economic crises.

Recommendation

1. Being immunised against various diseases through vaccination greatly increases the life expectancy of an individual.
2. Going to school directly affects an individual because they are more prone to take vaccines and immunisation shots for them and their families due to the knowledge they gain in school.
3. Health expenditure directly affects the life expectancy of people since they have access to medical care and can be given medication when they are sick.