Life Expectancy

Regression analysis

Map

Description automatically generated

DSC630-T301 Predictive Analytics

Bellevue University

Sowmya Chavali

**EXECUTIVE SUMMARY**

Life expectancy is based on an estimate of the average age that members of a particular population group will be when they die. It’s no secret, though, that life expectancy varies widely across the globe. I’ve always attributed living longer with increased access to healthcare and socio economic status of the country. To test this assumption and understand what actually best predicts life expectancy, I explored country-level data across the globe.

The primary goal of this project is to build a predictive model which helps improve life expectancy by working on how each factor influence longevity across the countries in the world. This helps the countries work on improving these features for better life span. This is a cross-country study collecting different features from all the countries for the years between 2018 and 2019.

Questions addressed-

* What are the factors affecting life expectancy of a country?
* Do rich countries have better Environmental Performance Index (EPI) ?
* Do people from rich nations live longer?

This task is split into three parts – first data preparation, EDA and building a model that is a best fit for the problem statement.

Linear Regression is the chosen model among the three.80.4% model accuracy is achieved by simple linear regression and ~79.5% with both LASSO and Ridge regression models.

**TECHNICAL REPORT**

**Background**

Recently, I watched a show on Netflix, “Down to Earth” where Zac Efron journeys around the world in search of healthy, sustainable ways to live. One episode in that travelogue explores life expectancy in Sardinia, one of the highest populations of centenarians in the world. They take a closer look at what is increasing Sardinian longevity. Their research has identified three elements of extended life expectancy: genetics, diet, and environment.

After watching this travelogue made me think of what other factors effects the life expectancy of a country and questions like do people in rich nations live longer?

Longevity has been increasing over the past century thanks to medical advances and lifestyle improvements. Not only has the average life expectancy increased since 1900, but a larger number of people are living to older ages, driven in part by a steep decline in the high infant mortality rate that characterized the early 1900s.

Life expectancy ranges from 53 years in Lesotho and the Central African Republic to 84 years in Japan and South Korea, a staggering gap of 31 years.  These extreme health inequities partly reflect wealth inequities between countries. Generally, wealthier countries have a higher average life expectancy than poorer countries, which can be argued to be achieved through higher standards of living, more effective health systems, and more resources invested in determinants of health.

**Data Understanding and preparation**

The data required for this project is life expectancy, demographics, socioeconomic status, health care resources, cancer and cardiovascular disease rates of all the countries. The features are life expectancy, population, area, development status, Environmental Performance Index, GDP, Health expenditure, birth rate, Heart disease rate, stroke rate, cancer rate. There was no available dataset for this project, thus for data collection step, I scraped data from Wikipedia and worldlifeexpectancy.com using requests.

Data Source:

<https://en.wikipedia.org/wiki/List_of_countries_by_population_in_2010>

<https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)>

<https://en.wikipedia.org/wiki/List_of_sovereign_states_and_dependent_territories_by_birth_rate>

<https://en.wikipedia.org/wiki/Environmental_Performance_Index>

<https://en.wikipedia.org/wiki/List_of_countries_by_total_health_expenditure_per_capita>

<https://en.wikipedia.org/wiki/List_of_countries_by_life_expectancy>

<https://www.worldlifeexpectancy.com/cause-of-death/all-cancers/by-country/>

<https://www.worldlifeexpectancy.com/cause-of-death/coronary-heart-disease/by-country/>

<https://www.worldlifeexpectancy.com/cause-of-death/stroke/by-country/>

All the data tables scraped are merged into a continuous dataset. Prior to merging, data cleaning steps were performed on each data set This includes renaming columns to make it more readable,correcting the data by stripping the special characters for few country names, removing duplicate data etc. After merging the data into one unit, outlier checks were performed. For the exploratory analysis I checked the multi collinearity between features. Finding correlation between features and target (life expectancy).

Chart, treemap chart

Description automatically generated

Feature selection is done by removing one feature at a time and checking on its significance in the model. To look for significant features, I drop one feature at a time to see its impact on the simple regression model. I look into the features which will adversely affect the model based on r2 score. I also consider selecting features using the LassoCV feature in SkLearn. Fit a machine learning model to predict the longevity.

Below are the models that could be tested to the trained dataset.

* Linear Regression (a straight line which approximates the relationship between the dependent variables and the independent target variable).
* Ridge Regression (this reduces model complexity while keeping all coefficients in the model, known as L2 penalty)
* LASSO Regression (Least Absolute Shrinkage and Selection Operator reduces model complexity by penalizing model coefficients to zero, i.e., L1 penalty)

**Results**

A screen shot of a game

Description automatically generated with medium confidence

Calendar

Description automatically generated

Countries who spent more on health expenditure are having higher EPI score, when health expenditure is higher, stroke rate is also lower; a larger area yields a higher population. The countries with high standards of living and low birth rates has higher life expectancy rates.

To live a long life, you should have low stroke rate, high health expenditure, take good care of the environment.

Chart, box and whisker chart

Description automatically generated

Developed countries has high life expectancy rate than the developing countries,

Background pattern

Description automatically generated

Background pattern

Description automatically generated

Life expectancy ranges from 53 years in Lesotho and the Central African Republic to 84 years in Japan and South Korea, a staggering gap of 31 years.  These extreme health inequities partly reflect wealth inequities between countries. Wealthier countries have a higher average life expectancy than poorer countries, which can be argued to be achieved through higher standards of living, more effective health systems, and more resources invested in determinants of health.

**Model deployment**

Since I have 250 rows (data limited by the number of countries in the world), I used the entire data set to simulate the test data set. I used KFold Cross Validation with 5 splits to evaluate the model performance.

**Model performance and selection**

* Linear regression R^2 : 0.804 -> Selected model
* Ridge regression R^2 : 0.79584
* Lasso regression R^2 : 0.79552

**Final Model stats and Interpretation**

**Graphical user interface, table

Description automatically generated**

* If your country has low birth rate, add 6 more years to your life.
* If the EPI (Environment Performance Index) is high, add 4 more years to your life.
* If you live in a rich country, add a year to your life.
* Finally for every unit (or rather LOG unit) decrease in stroke rate, 2 and half more years could be added to your life

**Risk**

* Getting the right data will help me arrive at the correct outcome.
* There are chances for high data discrepancy as in many countries most of these details are not correctly documented.

**Conclusion**

In conclusion the analysis aimed at choosing the relevant factors that affects the life expectancy. Japan has the highest life expectancy (84.3 years), Lesotho (52 years) and many countries in the African continent are at the bottom of scale. Taking good care of the environment and investing in health care has the larger impact on the country’s life expectancy.

**Next Steps**

* To collect more data, possible to expand scope to cities instead of countries
* To expand scope to include more features (factors)
* To split into life expectancy for males and females

**References**

1. <https://www.usatoday.com/story/news/world/2017/11/14/sardinia-oldest-people-world-italy/811783001/>
2. <https://www.worldlifeexpectancy.com/>