Influential factors for Life Expectancy (March 2022)

Fenghua An, Peiyao Zhang

Abstract*—The gap of life expectancy between developing and developed countries has existed for decades. Due to the economic, political and medical conditions, the evaluation and forecast of life expectancy are necessary to undertake research. Using data from WHO, three algorithms are built: linear regression, logistic regression and KNN model. Comparing their predictability, we finally conclude that linear regression is better performed. The correlation between variables and life expectancy and the rough forecast in this model illustrate the information from the developing countries.

INTRODUCTION

Everything has an expiration date; humans are no exception. We are in unprecedented times, and while humans are living longer and longer, life expectancy varies widely across the globe. The term "life expectancy" refers to the life length that a person expects. Life expectancy is based on an estimate of the average age at which members of a particular population group die. As machine learning and data science continue to advance, we will focus on parameters that have a large impact on an individual's longevity. We can also accurately predict a person's remaining lifespan given basic parameters.

To determine which indicators are statistically significant and predict life expectancy, we tried different models. We found this data to be a perfect example of regression, which determines the relationship between a dependent variable (y: Life Expectancy) and multiple independent variables (x). Linear regression is a relatively simple and interpretable algorithm. The deployment of linear regression requires minimal effort, but conversely, it lacks accuracy when the data is non-linear. Complex algorithms perform better on nonlinear datasets, but the models lack interpretability. After learning deep data exploration and many other tools on Python, it's time to go a step further in regression. Machine learning helps us have many models of varying degrees and options. In order to make regression models, in addition to common models such as Pandas, NumPy in Python, we can also use many libraries and tools, such as sklearn and statsmodels for train-test splits.

DATA DESCRIPTION

1. Source: The dataset is from WHO

2.Life expectancy by age is the dependent variable(the one we are trying to predict)

3. No. observation: 64,658

4. Attributes:

a. ICOR (INCOME COMPOSITION OF RESOURCES)

ICOR measures a country's ability to utilize its resources.

The ICOR is rated between 0 and 1, with a higher ICOR indicating the best use of available resources.

b. INFANT DEATHS: INFANT DEATHS PER 1,000 PEOPLE

c. ALCOHOL CONSUMPTION PER CAPITA (LITERS)

d. ADULT MORTALITY

Adult mortality is shown as the probability (per 1,000 people) that a 15-year-old will die before the age of 60.

e. EXPENDITURE

Health expenditure as a percentage of GDP per capita (%)

f. HEPATITIS B

Hepatitis B (HepB) immunization coverage in 1-year-olds (%)

g. MEASLES

Measles cases reported per thousand people

h. BMI

Average body mass index for the entire

i. POPULATION

j. DEATH UNDER THE AGE OF FIVE

Deaths of children under 5 per 1,000

k. POLIO

1-year-old polio (Pol3) immunization coverage (%)

l.TOTAL EXPENDITURE

General government health spending as a percentage of total government spending (%)

m. DIPHTHERIA

Diphtheria, tetanus, toxoid and pertussis (DTP3) immunization coverage in 1-year-old children (%)

n. HIVAIDS

HIV/AIDS deaths per 1,000 live births (ages 0-4)

o. GDP PER CAPITA (USD)

p. SLIMMING 10-19 YEARS OLD

Prevalence of thinness (%) in children and adolescents aged 10 to 19 years

q. Slimming 5-9 years old

Slimming rate of children aged 5 to 9 (%)

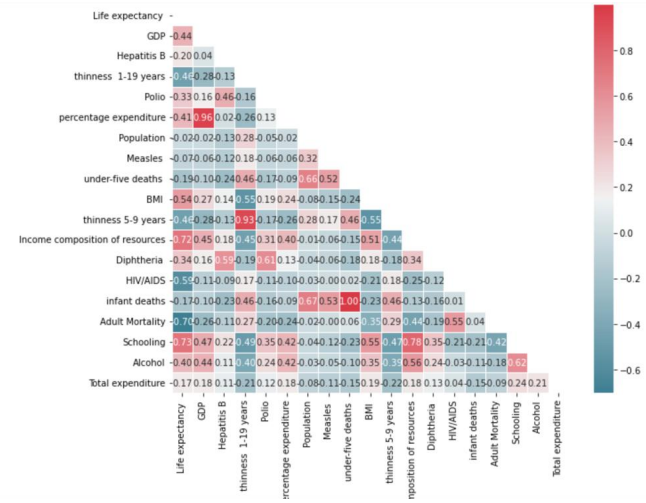
r. Composition of Resource Income

Human Development Index expressed as resource income composition (index range from 0 to 1)

Schooling: years of education

EXPERIMENTS

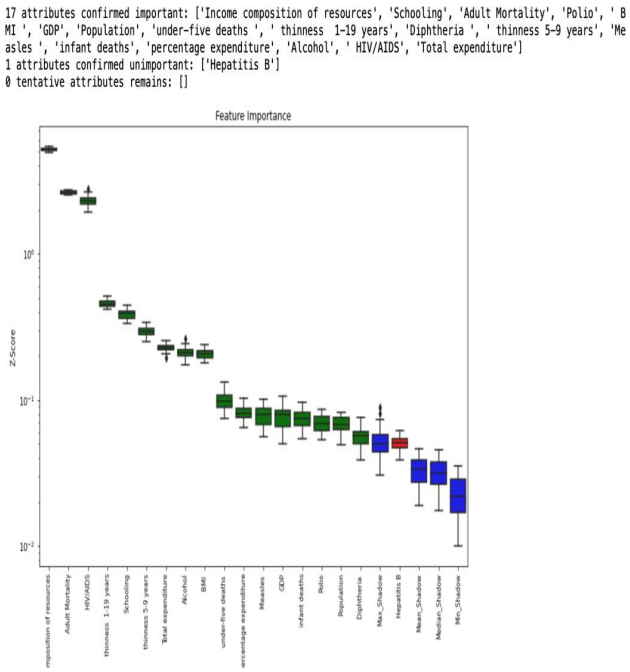
CORRELATION GRAPH



Variables Selection

We are analyzing the factors influencing life expectancy. After variable selections through Boruta shape, we select 18 important variables to evaluate and predict the life expectancy in developing countries, which are Schooling, GDP, Polio, thinness

1-19 years, under-five deaths, Diphtheria, Measles, Total expenditure, infant deaths, Population, HIV/AIDS, Hepatitis B, thinness 5-9 years, Alcohol', 'percentage expenditure, BMI, Income composition of resources, Adult Mortality.



*We drop ['Hepatitis B'] that confirmed uni mportant in the Boruta feature selection test from data frame to avoid bias and noise.

Model Design& Selection

Linear Regression Model

Linear regression is a regression algorithm that uses a linear approach. This is a supervised regression algorithm where we try to predict continuous values for a given data point by generalizing to the data we have at hand. The linear part represents a linear approach to data generalization.

From the OLS result summary, we can notice that life expectancy was significantly associated with adult mortality, BMI, educational attainment, HIV/AIDS, ICOR, and GDP.: As expected, life expectancy and adult mortality are highly inversely correlated. BMI is positively correlated with life expectancy. GDP is also positively correlated with life expectancy, and it can be inferred that as a country's GDP increases, so the quality-of-life increase and then does life expectancy. Not surprisingly, years of education are highly positively correlated with life expectancy. Higher education level leads to healthy habits and discipline.

The evaluation metrics chosen for this model are mean absolute error and R square score.

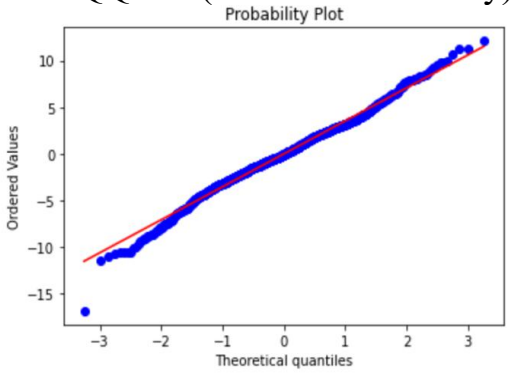
Mean absolute error measures the average error between the predicted value and the actual value. The lower the mean absolute error, the better.

The R square score accounts for linear dependencies between features. It tells how much variance in the dependent variable (target) is a consequence of the independent variable (feature). It is a fraction between 0 and 1. A higher score means a higher correlation between the feature and the target. The higher the score the better

Heteroscedasticity by using Breusch-Pagan (BP) Test

With such a low p-value ('p-value', 1.134014645267 0996e-17) in the OLS model, we would reject the null h ypothesis and try to account for the heteroscedasticity in our model. I try to Address heteroscedasticity by using Robust Standard Errors method and add HC0 in ne w model then also reject the null hypothesis with als o low p value ('p-value', 1.833409234946076e-29) means there is heteroscedasticity in new modelhc0; After checking the QQ plot which is looks quite good t hen I just continue to this model.

QQ Plot(Normal Probability)



A normal distiribution, half of the data is above or below the median (0)

In most cases where confidence intervals or significance levels are calculated, we assume normality If this assumption fails, our confidence intervals, and significance tests may be very misleading. There are formal tests of normality (see Kolmogorov-Smirnov and Shapiro-Wilk tests). These may fail in certain cases. Visual inspection is a fast way to validate this assumption via a quantile-quantile (QQ) plot.

Variance Inflation Factor

Adult Mortality	4.388668
infant deaths	210.860851
Alcohol	3.882615
percentage expenditure	15.916591
Measles	1.577663
BMI	8.827290
under-five deaths	193.433809
Polio	23.207676
Total expenditure	8.348256
Diphtheria	26.214184
HIV/AIDS	1.622873
GDP	17.708910
Population	2.483985
thinness 1-19 years	13.895255
thinness 5-9 years	13.895438
Income composition of resources	34.274832
Schooling	53.714148
dtype: float64	

We want a small VIF, usually less than 10 is prefer; the infant deaths and under-five deaths have extremely high VIF number close to 200 because the under-five deaths overlap with the infant deaths. However, considering under-five group have wider range than only infant and they all pass feature selection test, we decide to keep both features in the model.

Model Specification

The Ramsey RESET test is a general test of functional form misspecification. Now let represent the fitted values estimated from running the above regression. We can then add polynomials in the to test for nonlinear functions of our predictors. We can use the expanded equation; to test whether any quadratic combinations of our predictors may be appropriate -RESET is the F statistic from testing

Linear regression model with scaled features

In addition to OLS model, I also try the linear regression model with scaled features, means transform all x variables to normalize it. The best R square score of OLS with scaled features is 0.85 with MSE 12.61. I also all L1 and L2 regulation and combination of L1&L2 to reduce less important features and can avoid model overfitting issue. I try Ridge regression, Lasso regression and elastic net. Then finally, we try Logistic function and KNN model.

Ridge Regression Model

While the method of least squares determines the values of the parameters in the equation, it minimizes the residual sum of squares. Ridge regression, on the other hand, minimizes the sum of squared residuals plus lambda, and squares the slope of the regression line. Also, to make sure I didn't overfit my training data, I also checked the Ridge model. After checking several hyperparameters, the best R square score of Ridge model is 0.84.

Lasso (Least Absolute Shrinkage and Selection Operator) Regression Model

Ridge regression minimizes the residual sum of squares plus lambda and squares the slope of the regression line, while lasso regression minimizes the residual sum of squares plus lambda and the absolute value of the slope of the regression line. After checking several hyperparameters like learning rate, the best R square score of Lasso model is 0.85.

Ridge shrinks the parameters by keeping all parameters, while Lasso Regression eliminates and creates a simpler model to explain. Therefore, I would also like to get results from this model in order to provide broad elements to my predictions.

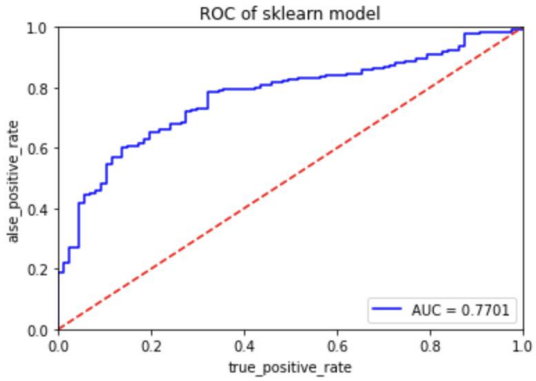
Elastic Net Regression Model

If we have a huge dataset, Elastic Net Regression is the best choice because the model preserves both Ridge regression and Lasso regression lambdas.

Again, I checked with a different range of alpha values, the best R square score of Lasso model is -0.03, means Elastic Net Regression doesn't work well for our dataset.

Logistic and KNN

For the life expectancy, it varies from 36.3 to 89. I prefer to analyze the performance of each life expectancy in each country, compared with that of the rest of the world. I set 63.2 years old as a standard for all the countries in this dataset. If a life expectancy is greater than 63.2, it means the life expectancy is satisfactory, otherwise it is not. The logistic regression facilitates my analysis in this dataset.



Last step is to evaluate the predictability in this regression. I applied ROC/AUC and accuracy into it. The score from accuracy is 0.7523. In order to a comprehensive understanding of the predictability, I applied KNN into this dataset. The similar metric of a satisfactory life expectancy is set down. Through adjusting the n_neighbors from 20 to 800, I got the predictability to be about 0.73. After 6 adjustments in this parameter, I found predictability is increasing with an increase in k, but when k reaches to 200, an increase in k slightly impacts on the predictability. I think it is due to the characteristic of the KNN, k is the number of the top similar items in one class. If k is too large, the model would be underfitting. If k is too small, the model would be overfitting, so an appropriate k is required to avoid underfitting and overfitting.

Finally, after comparing the performances between logistic model and KNN, logistic model is better performed to predict life expectancy based on the 18 important variables.

Challenges and Solutions

There are some confusions impeding our research.

Firstly, in the data cleaning part, the null values account for more than 1/3 in the whole database. We were thinking it would not be robust to analyze the correlation and predictability if there are so many missing values. However, after replacing the missing values with the mean, the correlations between variables and life expectancy is strong enough to dig into analysis further.

Secondly, in order to predict life expectancy, we classify this experiment into 2 parts, one is to forecast numerical outcomes, another is to evaluate the outcomes which we forecast based on the metric of the performance of the life expectancy. So, we were eager to have a deeper understanding of the metric of a satisfactory life expectancy. The values of life expectancy vary from about 30 to 90, and there is no strict definition to define what a great life expectancy is. I researched the world bank and the value description, so I set the top 25% of smallest life expectancy in the world as a metric to evaluate the performance of a life expectancy.

Thirdly, we were having different opinions on life expectancy's data processing. One thought considers it as a numeric value and forecast the future's trend in numbers' format; one thought the key is to evaluate and predict the performance of life expectancy, so a number cannot tell us whether this life expectancy is satisfactory or not. Based on this discrepancy, we finally make an agreement that even if the model outputs the numeric values, we can transfer these values into dummy values with a certain metric.

MODEL EVALUATION

Model Evaluation:

Comparison of models

	R square(best one)	MSE
Linear regression	0.85	12.61
OLS with scaled features	0.85	12.61
Ridge	0.84	13.69
Lasso	0.85	NA
Elasctic net	-0.03	NA
Logitisc function	0.135	NA
KNN	-0.266	NA

From the evaluation metrics, linear regression model is the best fit for this dataset. We prefer higher R squared and lower MSE.

CONCLUSION

Due to the comparison between three models above, we finally conclude that the linear model has a higher predictability in this dataset. Life expectancy depends on several variables such as government spending, health status, etc. The trend is smoothly increasing, additionally based on machine learning, it is reasonable to predict life expectancy will increase in these developing countries.

FUTURE WORK

Although we use predict/forecast in machine learning, we think a detailed and comprehensive forecast is necessary. How the related variables would change in the next 10 years and what the trend of the life expectancy would be based on the changes of the variables are critical. Especially, the covid-19 plays an important role to the health status in recently years. It might flip the coin, when other variables reach a good result of health standards. Therefore, our database will be required to supplement and update the forecast result.