Illinois Institute of Technology

Statistical Analysis on factors influencing Life Expectancy

Team Members:
Jiang Cheng (JC) · Rahul Nair (RN) · Yurun Liu (YL)

Introduction

Lifespan is a topic that accompanies the individual's life. With the development of science and technology and the passage of time, more and more attentions are paid to human lifespan. In the US, we have average 2,839,205 people death per day.[1] The average life expectancy is 78.7 for people in the US.[1] According to The Global Health Observatory (GHO), 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. If we can find demographic factors associated with life expectancy, we can expect certain people for lifespan.

Data from World Health Organization (WHO) can be used to achieve this. Link to the dataset is here: https://www.kaggle.com/kumarajarshi/life-expectancy-who. 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. The sample is globally representative. Among all categories of health-related factors only those critical factors were chosen which are more representative. For example, we have immunization related factors, mortality factors, economic factors, and social factors.

For this project, I would like to explore people's life expectancy, and the demographic factors associated with life expectancy for a model to expect people's lifespan.

Methods

Life Expectancy data

As described before, The Global Health Observatory (GHO) data repository under World Health Organization (WHO) have considered data from year 2000-2015 for 193 countries. Variables included in this study were life expectancy in age, country status, some disease cases, GDP, and so on. The life expectancy in age is outcome and other variables are demographics of interest. These variables were obtained from the Life Expectancy files.

First, we did exploratory data analysis (EDA) for the dataset. We use several graph ways to deal with the data set. For the boxplot, we use it for checking the outlier and use winsorizing and log transformation to clean the outlier. For the histogram, we can check it is normal distribution or not. Also, we use it for compare the developing and developed country's life expectancy. For pie plot, we use it to see the percentage for developing and developed country in this dataset. In addition, we build the correlations matrix for the predict values to select the correlations between the income variables. We also draw a line plot and scatter plot for the data and see the relation with life expectancy.

After the EDA, we start to think about the models. First, we change the country status to dummy variables. And then we recheck and clean the data again to make sure there is no other issues from data. After above, we splitted the data to be 70% for training and 30% for testing dataset. Since the predictor data is all affect life expectancy. So, we include all variables into the multiple linear regression, ridge regression, Lasso regression, decision tree, and random forest model to check the adjusted R square and the root mean square error score for the data. We will depend the overfitting and score to decided which model is better.

Results and Discussion

A total of 193 countries were included in the current study. We can see the developed countries life expectancy are higher than developing countries' life expectancy (Figure 1 and 2). The factors like income composition of resources which is giving us a good picture to show HIV/AIDS are highly correlated to the target variable (Life_Expectancy). Apart from that, we see some high correlations among the variables as well such as DiphtheriaPolio (0.88) and Thinness_1-19_yearsThinness_5-9_years (0.97) which is very high. We also check about the life expectancy with year. We can easily to see that life expectancy is on the rise with time (Figure 3). We list the top 10 countries for the life expectancy and most of countries are developed country.

First, We splitted dataset for 70% training dataset and 30% testing dataset. We use training dataset to build multiple linear regression model. We can see the table 1 (table 1) which shows the p values are less for pretty much all the predictors except a few. The coefficients value make sense. For example, we can see the life expectancy with variable yeas. The coefficient is positive which means it is increasing by year. The adjusted r square is 0.95 which is pretty good. The F statistic is 236.7 that proofs our predict value are significant. For the checking the residuals. We draw a residuals Vs fits plot (Figure 4). The line is kind of horizontal which proofs linear regression model is appropriate for the data. Also, the model generated is good enough as the points are getting clustered towards the center. However, when we calculate evaluation metrics for testing data, we got root mean square error (RMSE) is 2.157, r square is 0.948, and Adjusted R square is 0.948. but we got evaluation metrics for training data RMSE is 1.808, r square is 0.963, and Adjusted R square is 0.959. The test RMSE is greater than training RMSE

little bit which means the model is a little bit overfitting. So, we try to use shrinking methods like lasso regression and ridge regression and see if we can reduce that.

Second, we use ridge regression and got the evaluation metrics for training RMSE is 1.808, r square is 0.963, and Adjusted R square is 0.963. The ridge regression evaluation metrics for testing RMSE is 2.154, r square is 0.948, and Adjusted R square is 0.948. It still has overfitting problem and the Adjusted R square get worse. So, we try Lasso regression now. We got the evaluation metrics for training RMSE is 1.808, r square is 0.963, and Adjusted R square is 0.963. The ridge regression evaluation metrics for testing RMSE is 2.158, r square is 0.948, and Adjusted R square is 0.948. We got same result as ridge regression. So, we are thinking about other models now. There are decision tree and random forest. By decision tree, we got the decision tree evaluation metrics for training RMSE is 1.28, r square is 0.982, and Adjusted R square is 0.982. The decision tree evaluation metrics for testing RMSE is 2.65, r square is 0.921, and Adjusted R square is 0.921. The Random Forest evaluation metrics for training RMSE is 0.906, r square is 0.991, and Adjusted R square is 0. 991. The Random Forest evaluation metrics for testing RMSE is 1.97, r square is 0.956, and Adjusted R square is 0. 957. They both has over fitting problem. Decision Tree which is performing the worst of all.

Figure 1: country status VS life expectancy

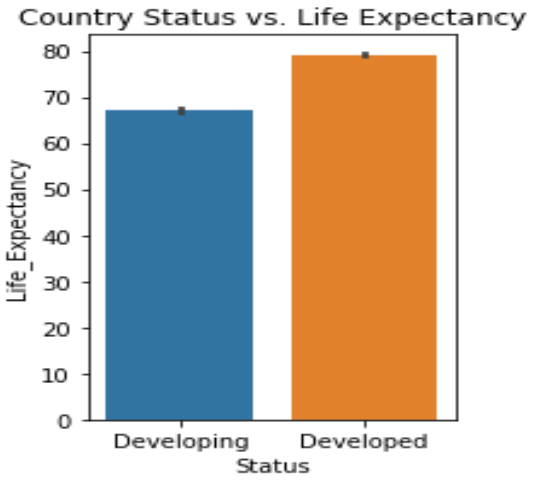

Figure 2: country status VS life expectancy

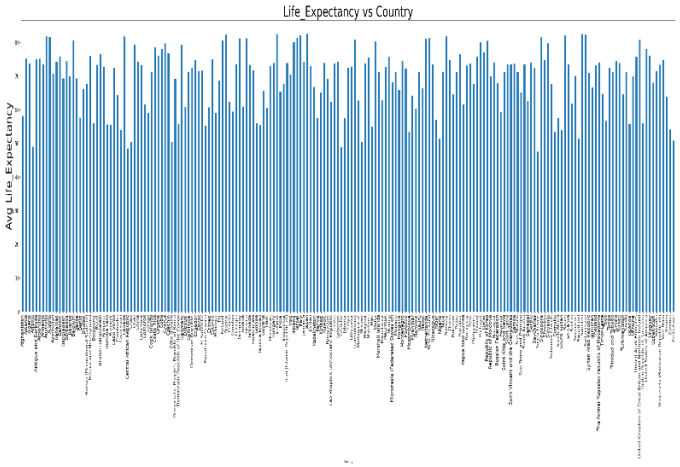

Figure 3: life expectancy by year

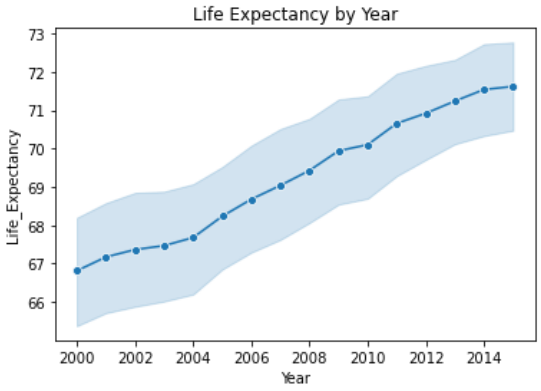

Figure 4: residual Vs fits

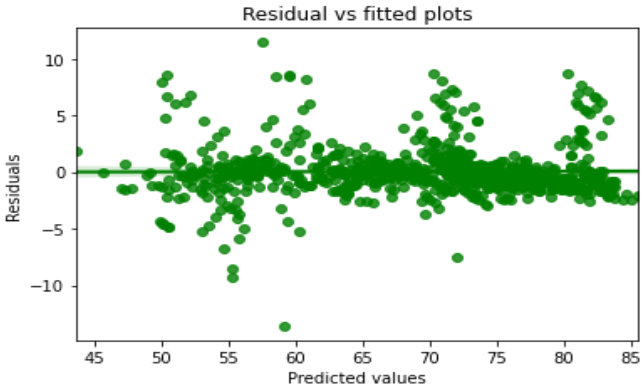

Table 1 Result and part of summary for multiple regression model

OLS Regression Results

Dep. Variable:	Life_Expectancy	R-squared:	0.963
Model:	OLS	Adj. R-squared:	0.959
Method:	Least Squares	F-statistic:	236.7
Date:	Tue, 01 Dec 2020	Prob (F-statistic):	0.00
Time:	14:30:16	Log-Likelihood:	-4135.4
No. Observations:	2056	AIC:	8685.
Df Residuals:	1849	BIC:	9850.
Df Model:	206		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-331.7414	19.395	-17.105	0.000	-369.779	-293.704
Year	0.2850	0.015	19.176	0.000	0.256	0.314
Adult_Mortality	-0.0010	0.001	-1.711	0.087	-0.002	0.000
Infant_Deaths	0.1729	0.052	3.329	0.001	0.071	0.275
Alcohol	-0.0983	0.029	-3.396	0.001	-0.155	-0.042
Percentage_Expenditure	0.0006	0.000	2.244	0.025	7e-05	0.001
HepatitisB	-0.0093	0.004	-2.306	0.021	-0.017	-0.001
Measles	-0.0002	0.000	-1.046	0.296	-0.001	0.000
BMI	-0.0054	0.004	-1.504	0.133	-0.012	0.002
Under_Five_Deaths	-0.2524	0.036	-7.020	0.000	-0.323	-0.182
Polio	0.0122	0.006	1.973	0.049	7.02e-05	0.024
Total_Expenditure	-0.1208	0.135	-0.898	0.369	-0.385	0.143
Diphtheria	0.0256	0.007	3.682	0.000	0.012	0.039
HIV/AIDS	-0.3945	0.273	-1.443	0.149	-0.931	0.142
GDP	-5.305e-05	1.78e-05	-2.984	0.003	-8.79e-05	-1.82e-05
Thinness_1-19_Years	0.0268	0.040	0.674	0.500	-0.051	0.105

Thinness_5-9_Years	-0.0877	0.039	-2.226	0.026	-0.165	-0.010
Income_Composition_Of_Resources	1.2464	0.982	1.270	0.204	-0.679	3.172
Schooling	0.0869	0.067	1.299	0.194	-0.044	0.218

Table 2 evaluation metrics for training data

	Model R_square	R_square	Root Mean_square	Adjusted R_square
0	Multiple Linear Regression	0.963470	1.982585	0.959400
1	Ridge Regression	0.963470	1.808408	0.963488
2	Lasso Regression	0.963455	1.808773	0.963473
3	Decision Tree	0.981688	1.280393	0.981697
4	Random Forest	0.990821	0.906523	0.990825

Table 3 evaluation metrics for testing data

	Model R_square	R_square	Root Mean_square	Adjusted R_square
0	Multiple Linear Regression	0.947973	2.156721	0.948032
1	Ridge Regression	0.948102	2.154049	0.948161
2	Lasso Regression	0.947901	2.158218	0.947901
3	Decision Tree	0.921199	2.654278	0.921288
4	Random Forest	0.956497	1.972153	0.956546

Conclusion

In conclusion, we can see the adjusted r square is kind of same for each model. but we conclude the multiple linear regression model is better base on the RMSE because there is not too much difference. So, the over fitting will be small. It will be useful for life expectancy.

Bibliography

1. https://www.cdc.gov/nchs/fastats/deaths.htm
2. https://www.kaggle.com/kumarajarshi/life-expectancy-who