COMPUTATIONAL MACHINE LEARNING

ASSIGNMENT - 1

Name - Ayaz Aziz Mujawar
Student Number - s3751555

Predicting Life Expectancy of NewBorn

Abstract

The aim of this report is to find a suitable regression model to predict the life expectancy of the new born baby based on certain socio-economic and health conditions with accuracy and efficiency. Regression models namely Multiple Linear Regression, Polynomial Regression, Ridge Regression and Lasso Regression were adopted and trained using 2071 samples data. The report concludes that Ridge Regression is the better performing model in terms of accuracy and error rate and gives better predictions for life expectancy of new born babies. For better accuracy and reduced error rate it is recommended to normalize the data variables, remove the outliers and select the significant features to get the best results.

Methodology

In this section there are two main goals: First understanding the characteristics about the type of values a feature can take, the ranges into which these values fall and how the values in datasets are distributed across the range. The second goal is to determine the relationship between the independent and dependent variables by plotting the scatterplot, heatmap and calculating the correlation coefficients.

Firstly I tried to check the assumptions on Linear Regression by using different conditions -

1. **Linearity** - Linearity depicts the relationship between the independent variables and the target variable. As shown in the scatter plot (Fig 1.6 and 1.7) not all independent variables possess a linear relationship with the target variables.So the output variable will not change in direct proportion to change in the input.

2. **Lack of multicollinearity** - Lack of multicollinearity means there should not be a strong relationship between two or more independent variables. Generally the magnitude of correlation coefficient should be less than 0.80. But as shown in the Heatmap (fig 1.4), a lot of variables are heavily correlated with each other. For example columns like Thinness 1-9 year are strongly correlated with Thinness 5-9 years.Also percentage expenditure are strongly correlated with GDP and columns like Under five Deaths are highly correlated with Infant's death.Due to multicollinearity it is very difficult to tell which independent variable has an effect on the response variable.

3. **P-value** - I also calculated the p-value. A low p-value(<-.05) indicates that the predictor variable is a more meaningful(significant) addition to the model. Contrarily, a large p-value (>0.05) suggests that changes in the predictor variable does not affect the response variable.
   In the output figure below(fig 1.8), you can see that the predictor variables like 'Year', 'Status', 'AdultMortality', 'InfantDeaths', 'Alcohol','BMI', 'UnderFiveDeaths', 'Polio', 'Diphtheria', 'HIV-AIDS', 'GDP', 'IncomeCompositionOfResources', 'Schooling' has p-value less than 0.05 So this variables are statistically significant for the model.

However p-values of variables such as 'Country','AdultMortality-Male', 'AdultMortality-Female', 'InfantDeaths', 'Alcohol', 'PercentageExpenditure', 'Measles', 'BMI', 'UnderFiveDeaths', 'Polio', 'TotalExpenditure', 'Diphtheria', 'HIV-AIDS', 'GDP', 'Population', 'Thinness1-19years', 'Thinness5-9years' are greater than the common alpha value of 0.05, which indicates that these variables are not statistically significant.

4. **Multivariate Normality** - Multivariate Normality means residuals should be normally distributed. As shown in figure 1.5 we have used qq-plot to visualize the error term in normalized form.

5. **Homoscedasticity** is the state where all the error terms have the same finite variance. But as per our analysis as shown in the fig. 1.4 the error terms are not distributed.

Then I checked for outliers in the data points that are not significant and deceived the training process. I tried to detect the outliers in the dataset using boxplots(as shown in fig: 1.1) and found that there were a lot of extreme values(outliers). As per the requirements I was not supposed to remove the outlier But from my analysis I would suggest to cap the outlier data to the maximum or minimum arbitrary set value. (for example upper extreme value replace with upper limit of 3rd quartile and lower extreme value replace with lower limit of 1st quartile). Although there might be some chances of bias but it will degrade the effect of outliers to a great extent.

Further as seen in our datasets most of the features are varying in magnitudes, units and ranges. For example in our dataset ‘GDP’ and ‘BMI’ have different magnitudes and units.To overcome this issue I used Feature Scaling where firstly I applied a standard scalar to transform the data where it has mean of 0 and standard deviation of 1. But due to non-normality and skewness in the distribution of data it was not a good fit in this case. Also I tried Min-Max scaler to scale the data, but it also fails due to the large number of outliers in our dataset. Finally I used the Robust Scalar method because it was scaling all the values of the features to the same scale and also reducing the effects of outlier unlike min-max scalers.

Modelling

The regression approach for modelling is used and the dataset is split into training and test sets at the ratio of 80:20. For this split, four regression models are used: Multiple Linear Regression, Polynomial Regression, Ridge Regression and Lasso Regression. The performance of the model is recorded based on the accuracy R-Squared and Adjusted R-Square while the errors measures were Mean Absolute Error(MAE), Mean Squared Error(MSE) and Root Mean Squared Error(RMSE).

Evaluation Metrics

The two metrics that I am choosing to evaluate algorithms are -

1. **Adjusted R2**
   Adjusted R2 helps in measuring the effect of independent variables on the dependent variables. The main reason for choosing Adjusted R2 over .R-square is Adjusted Adjusted R-square value will increase/decrease only when independent variables share a significant relationship with the dependent variable. Unlike R-square everytime you add a new feature to the model irrespective of the significance R-square value will increase.

2. **Mean Absolute Error(MAE)**
   MAE defines the average of absolute difference between predicted and the observed values. The main reason to use MAE as an evaluation metric instead of MSE is because in the provided dataset there are a lot of outliers and MAE is more robust to outliers.

Multiple Linear Regression

Multiple linear regression is a statistical technique in which several predictors variables are used to predict the outcome of the response variable. In our case multiple linear regression was developed using 22 predictors variables and 1 response variable. As shown in fig 1.6 not all the variables show the moderate linear relationship with the dependent variables. Fig 1.4 shows the multicollinearity in the datasets as a lot of independent variables share the moderate positive relationship among each other. This model also possesses high bias on the data which leads to high error rate. From table 1.1 we can tell that the mean absolute error (mae) is 3.71. Also from the residual plot (fig 1.4) we can tell that the plot is not evenly distributed and they have an outlier.

Polynomial Regression

As for some of the variables in (fig 1.4) we can see that data is correlated but the relationship is not linear(For e.g Relationship between HIV-AIDS and Target_Life Expectancy). To overcome those drawbacks we use Polynomial regression. Polynomial regression works on non-linear data. For this model I have tried 1st,2nd and 3rd degree of polynomials. Out of the three 2nd degrees polynomial provides the highest adjusted R2 square value with 0.79 and lowest mean absolute error of 2.85.

Ridge Polynomial Regression

By adding regularization of type ridge in the regression we tried to keep the number of features the same but reduce the magnitude of the coefficients to avoid overfitting. By applying ridge regularization to our polynomial regression we have slightly improved our model because the accuracy of R-square has been slightly increased. To set the hyperparameter of the ridge that is to select the value of alpha, we used k-fold cross validation technique. Throughout the process I also noticed that as we increase the value of alpha the value of coefficients decreases and also

R-square is maximum at alpha value of 0.05. But to reduce the error term and keep the accuracy intact I have decided to go with alpha value of 0.03. As it is giving the Adjusted R-square of 0.80 and MAE of 2.90.
Some of the other reasons I have considered are that ridge regression helps to decrease the coefficient values so it mostly prevents multicollinearity issues and ultimately helps in reducing model complexity and avoid overfitting.

Lasso Polynomial Regression

In Lasso Regression some of the variables that are not significant directly shrinks to 0. From my analysis even the small amount of change in the alpha value tends to make coefficient 0. This Lasso regression does feature selection on the variables. It selects the important features while reducing the coefficients of other features to 0.The main issue with lasso regression is when two or more variables are correlated then this technique will keep only one feature and discard all other features. This can lead to information loss and also resulting in lower accuracy.
As we can see in the table for our model alpha value is taken 0.03 and we are getting the accuracy of Adjusted R-squared of 0.68 and MEA of 3.73.

Conclusion

After applying all the regression models to the dataset at the ratio of 80:20, it turns out that the Ridge Regression gives the highest adjusted r-square value and lower error rate. Also the complexity of this mode is simple. However the optimal parameter or optimal model in this research may not be applicable due to following constraints of not removing the outlier and not doing the feature selection on this highly imbalanced dataset.

But If we normalize the data and perform feature selections based on multicollinearity and p-value then it would increase the accuracy of the model by reducing the residuals.

Appendices

Table 1.1

Models	R-Square	Adjusted R-Square	Mean Absolute Error(MAE)	Mean Square Error(MSE)	Root Mean Square Error(RMSE)
Linear Regression	0.74	0.74	3.71	22.96	4.79
Polynomial Regression	0.82	0.79	2.85	15.71	3.96
Ridge Regression	0.83	0.80	2.90	13.98	3.73
Lasso Regression	0.73	0.68	3.73	4.74	22.52

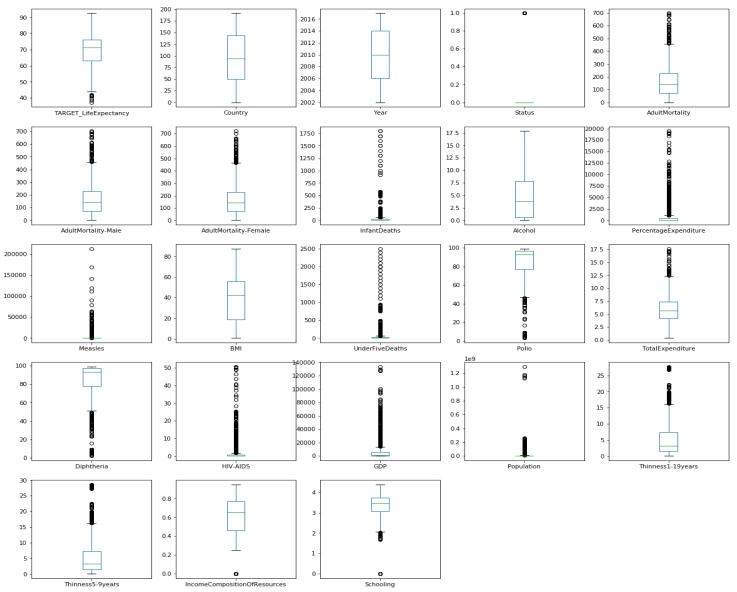

Fig 1.1

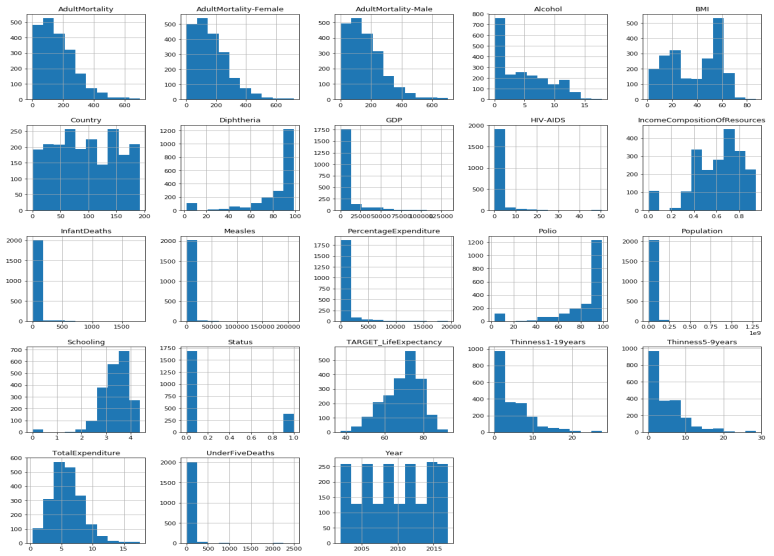

Fig 1.2

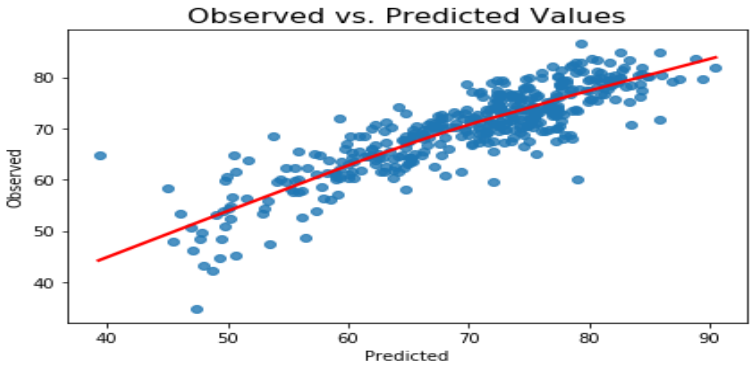

Fig 1.3

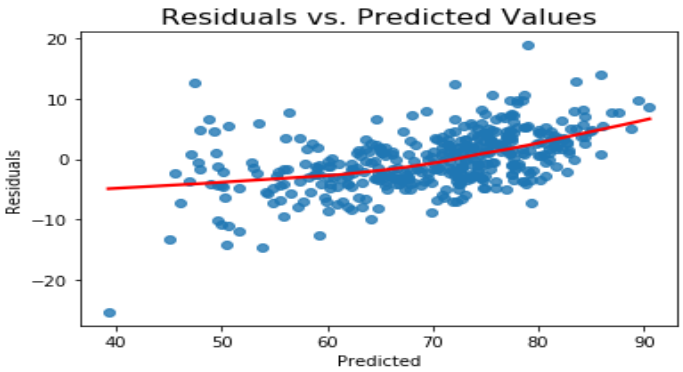

Fig 1.4

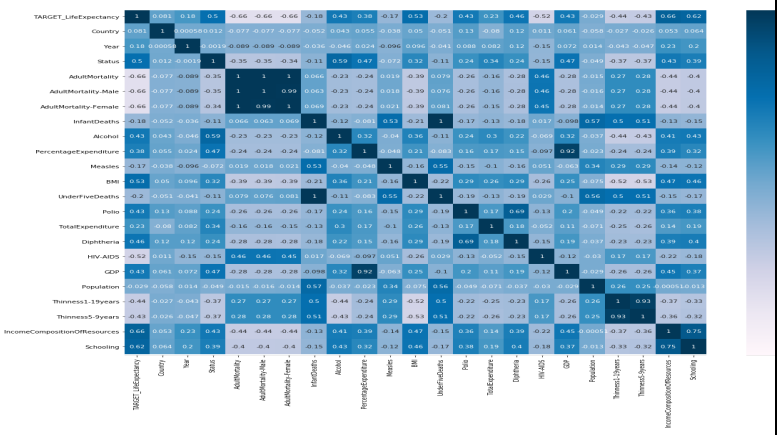

Fig 1.4

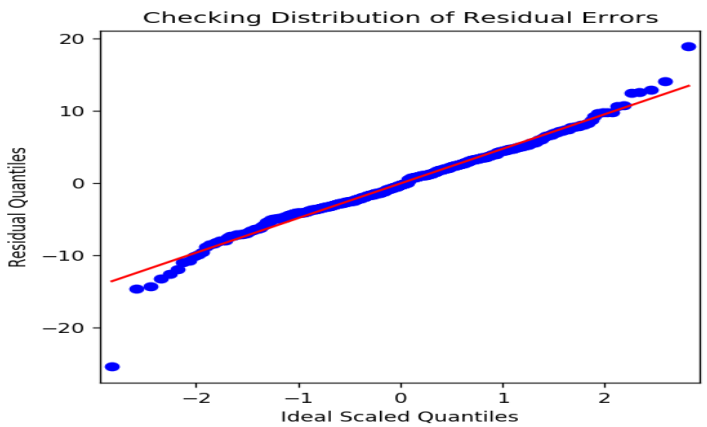

Fig 1.5

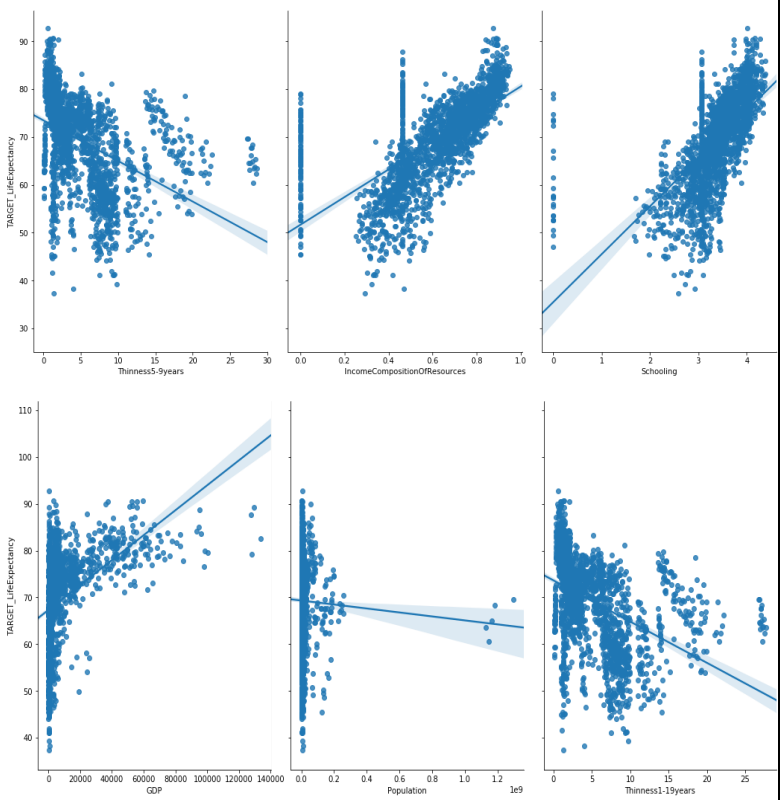

Fig 1.6

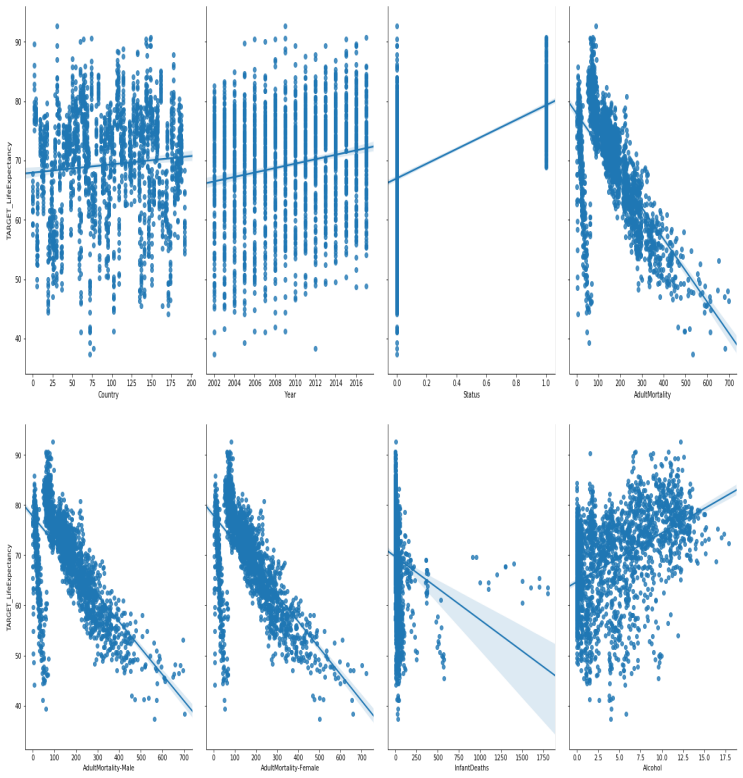

Fig 1.7

	coef	std err	t	P>|t|	[0.025	0.975]
const	-61.6791	47.730	-1.292	0.196	-155.284	31.925
Country	0.0030	0.002	1.584	0.113	-0.001	0.007
Year	0.0582	0.024	2.443	0.015	0.011	0.105
Status	2.4749	0.375	6.606	0.000	1.740	3.210
AdultMortality	-0.0068	0.000	-18.327	0.000	-0.008	-0.006
AdultMortality-Male	-0.0028	0.007	-0.420	0.674	-0.016	0.010
AdultMortality-Female	-0.0108	0.007	-1.601	0.110	-0.024	0.002
InfantDeaths	0.0950	0.011	8.535	0.000	0.073	0.117
Alcohol	0.1874	0.034	5.540	0.000	0.121	0.254
PercentageExpenditure	-5.143e-06	0.000	-0.039	0.969	-0.000	0.000
Measles	-8.069e-06	1.32e-05	-0.612	0.540	-3.39e-05	1.78e-05
BMI	0.0313	0.007	4.559	0.000	0.018	0.045
UnderFiveDeaths	-0.0706	0.008	-8.636	0.000	-0.087	-0.055
Polio	0.0199	0.006	3.174	0.002	0.008	0.032
TotalExpenditure	-0.0034	0.046	-0.074	0.941	-0.093	0.086
Diphtheria	0.0314	0.006	4.915	0.000	0.019	0.044
HIV-AIDS	-0.5029	0.025	-20.244	0.000	-0.552	-0.454
GDP	4.333e-05	1.87e-05	2.318	0.021	6.67e-06	8e-05
Population	-1.487e-09	2.03e-09	-0.733	0.464	-5.47e-09	2.49e-09
Thinness1-19years	-0.0762	0.061	-1.249	0.212	-0.196	0.043
Thinness5-9years	-0.0349	0.061	-0.575	0.565	-0.154	0.084
IncomeCompositionOfResources	6.2988	0.810	7.773	0.000	4.710	7.888
Schooling	2.2667	0.280	8.086	0.000	1.717	2.816

Fig 1.8