Comparing Model Accuracy and Identifying Pain Points in the Indian Healthcare System

Group No-25
Team behind the project

Ikshu Manjan Kumbhar - 18IM30009
Nanisetty Sai Shailesh - 18IM30013
Oqais Tanvir - 18IM30015
Suman Kumar - 18IM30022
Shashank Srivastava - 18IM3FP14
Abhyudaya Nilosey - 18IM3FP27

Abstract - The healthcare sector around the world has been a topic of quite some deliberation, with healthcare policies swaying elections in many countries of the developing and the developed world. This project aims at creating a score for life expectancy of people in a country using regression models applied over some defining features. The chosen features are used to build a multiple linear regression based model. On encountering the problem of multicollinearity, which was expected due to the high correlation between individual features, ridge and LASSO regression models were applied to gain insights on the importance of various features on the final model. This study ends with highlighting the critical pain points for the Indian Healthcare system and points out the key focus areas to revive from the status quo.

Introduction - The study chooses life expectancy as a measure of quality of healthcare in a country. Life expectancy is the average period that a person may expect to live. The project aims at analyzing the features that can lead to a higher or lower life expectancy, and quantifying the degree of sensitivity various factors might have on the average life expectancy of individuals. The dataset contains all 249 countries and their corresponding average life expectancy. In this dataset we are specifically studying the effect of dengue on the life expectancy of the country, although there are some other factors as well.

The various features present in the dataset are: -

1. Birth Rate
2. Cancer Rate
3. Dengue Cases
4. Environmental Performance Index (EPI)
5. Gross Domestic Product (GDP)
6. Health Expenditure
7. Heart Disease Rate
8. Population
9. Area
10. Population Density
11. Stroke Rate

In this project we will develop a regression model using various algorithms like Ridge, Lasso and Multiple Linear Regression and also compare the results of the algorithms. We will also demonstrate the amount of dependency of our dependent variable (i.e. Life Expectancy) on each of our independent variables. We will check for correlation and multicollinearity among the variables and would demonstrate that as well. We will demonstrate the distribution of variables. In Ridge and Lasso regression we will also demonstrate the effect of the change in regressor variable on the result of the regression.

Literature and Theoretical Review -

The study begins with deploying a Multiple Linear Regression model to the given dataset. Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable.
Formula and Calculation of Multiple Linear Regression
$yi=\beta 0+\beta 1xi1+\beta 2xi2+...+\beta pxip+\epsilon$
where, for $i = n$ observations:
yi = dependent variable, xi = *explanatory* variables, $\beta 0$ = y-intercept (constant term), βi = slope coefficients for each explanatory variable, ϵ = model's error term (also known as the residuals)
The "RESIDUAL" term represents the deviations of the observed values y from their means $\boldsymbol{\mu}_y$, which are normally distributed with mean 0 and variance σ^2. In the least-squares model, the best-fitting line for the observed data is calculated by minimizing the sum of the squares of the vertical deviations from each data point to the line (if a point lies on the fitted line exactly, then its vertical deviation is 0). Because the deviations are first squared, then summed, there are no cancellations between positive and negative values.

Problem with MLR model: Multicollinearity can create inaccurate estimates of the regression coefficients, inflate the standard errors of the regression coefficients, deflate the partial t-tests for the regression coefficients, give false, nonsignificant, p values, and degrade the predictability of the model (and that's just for starters).

In ridge regression, the cost function is altered by adding a penalty equivalent to square of the magnitude of the coefficients.

$$\sum_{i=1}^{M}(y_i-\widehat{y_i})^2=\sum_{i=1}^{M}(y_i-\sum_{j=0}^{p}w_j*x_{ij})^2+\lambda\sum_{j=0}^{p}w_j{}^2$$

- Cost function for ridge regression

Thereby ridge regression puts constraint on the coefficients *(w)*. The penalty term (λ) regularizes the coefficients such that if the coefficients take large values the optimization function is penalized. So, ridge regression shrinks the coefficients and it helps to reduce the model complexity and multi-collinearity. The regularization function used is the L2 Norm of the coefficient vector w.

Lasso Regression : The cost function for Lasso (least absolute shrinkage and selection operator) regression can be written as:

$$\sum_{i=1}^{M}(y_i-\widehat{y_i})^2=\sum_{i=1}^{M}(y_i-\sum_{j=0}^{p}w_j*x_{ij})^2+\lambda\sum_{j=0}^{p}|w_j|$$

- Cost function for Lasso regression

The only difference is instead of taking the square of the coefficients, magnitudes are taken into account. This type of regularization (L1) can lead to zero coefficients i.e. some of the features are completely neglected for the evaluation of output. So Lasso regression not only helps in reducing overfitting but it can help us in feature selection.The regularization function used is the L1 Norm of the coefficient vector w.

MLR	Ridge Regression	Lasso Regression
No Regularization	Regularization with L2 Norm	Regularization with L1 Norm
Minimum Mean Square Error	For given λ, Error<Lasso Error	For given λ, Error>Ridge Error

Methodology

On observing the dataset it was found that while some features like GDP had values of the order of 10^7, others such as the EPI index took very small values. Thus, the need for normalisation was felt. The dataset was first scaled using MinMaxScaler() function.This function scales the data as follows:

$$x_{scaled}=\frac{x-min(x)}{max(x)-min(x)}$$

Here, x is an individual feature vector and x_scaled is the scaled feature vector.
Max(x) and min(x) give the maximum and minimum value of the feature vector.

The whole dataset is then split randomly into a test and a train set with a ratio of 25:75 of the original dataset using the train_test_split function of scikit_learn library of python. Then, the train dataset as well as the test dataset is separated into feature matrix 'X' and target vector 'y'.

$\widehat{\beta}$ is the best estimated coefficient vector and ε is the error vector. The pearson correlation method was used to find correlation among features and a heatmap function was deployed for the sake of visualisation. The formula used to find pearson correlation is:

$$r_{xy}=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}*\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}$$

Here, n =size of feature vector, x_i *and* y_i are individual data points of two features, $\overline{x}$ is the mean of data points of one feature vector, $\overline{y}$ is the mean of data points of another feature vector.
To check if there is multicollinearity in our dataset, it is checked whether X^TX is invertible or not, so, determinant of X^TX is calculated.The value of the determinant came out to be in the order of 10^{-15} which is approximately zero.This indicates that the features in the dataset have multicollinearity. $\widehat{\beta}$ is estimated using the formula $\widehat{\beta}=((X^TX)^{-1}X^T y$.But as there is multicollinearity $\widehat{\beta}$ was not feasible because the variability of β_hat distribution was large.The predicted and mean square error values(ε) are calculated using following formula:

$$\text{Predicted matrix}(\widehat{y})=X*\widehat{\beta}+\varepsilon$$

$$\text{Mean square error}=\sum_{i=1}^{n}(1/n)(\widehat{yi}-yi)^2$$

Ridge Regression

The dataset is trained using Ridge Regression. The variability of β_hat was checked by changing λs.A tuned λ is chosen from all the λs and then further work is done.

$$\widehat{\beta}\sim N(\beta,\ \sigma^2*(X^TX+\lambda*I)^{-1})$$

Here, Ridge Regression has an additional factor called lambda(λ) which is called the penalty factor(also referred to as the regularisation parameter, in this project) which is added while estimating β coefficients.The λ penalizes the higher β coefficients leading to shrinkage of it and thereby reducing the mean squared error. This facilitates the overall

objective of minimizing least square condition in a way that β hat has a bounded norm.

$$\sigma^2 = (\sum_{i=1}^{n} (\widehat{y}_i - y_i)^2)/\text{n-k-1}$$

where n-k-1 is the degree of freedom.

$$\text{Predicted matrix}(\widehat{y}) = X * \widehat{\beta} + \varepsilon$$

$$\text{Mean square error} = \sum_{i=1}^{n} (1/n)(\widehat{yi} - yi)^2$$

The distribution of absolute error is plotted using the seaborn library. The distribution function of seaborn library plots the estimated PDF over the error data.
A line graph is plotted between mean squared error and λs using matplotlib library. The value of mean squared error is increasing on increase in the value of λs(penalty factor).

Lasso Regression

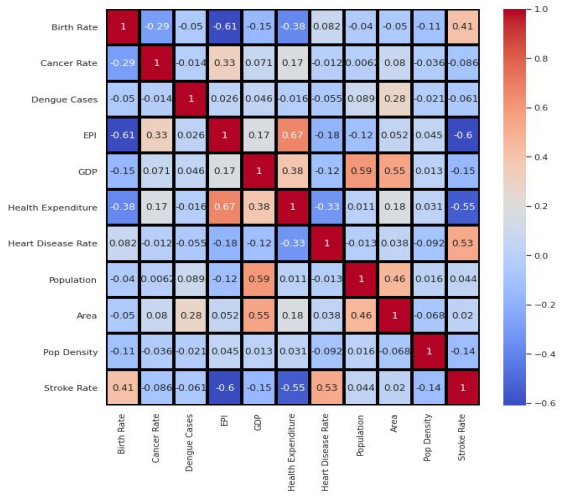
The Lasso Regression is applied after Multiple linear Regression. The penalty parameter λ is tuned for the lasso part. The variability of β coefficients are checked by changing λs. It is inferred that some of the β coefficients of features are tending to zero very early(meaning have less variability) while others are not tending to zero early(meaning having high variability). Then after, the distribution of absolute error is plotted which resembles the normal distribution plot. A plot is drawn between mean squared error and λs(penalty factors).

$$\text{Predicted matrix}(\widehat{y}) = X * \widehat{\beta} + \varepsilon$$

$$\text{Mean square error} = \sum_{i=1}^{n} (1/n)(\widehat{yi} - yi)^2$$

Results

The obtained heatmap enlists the pearson correlation coefficient between each pair of features used in the regression model. As is evident from the correlation matrix, health expenditure and EPI are two features that give a high positive correlation value of 0.67, while population and GDP also are highly correlated with a pearson correlation coefficient of 0.59. Stroke rate and EPI have a high negative correlation with a magnitude of 0.6. This measure is deemed intuitive as the countries that spent more on healthcare have a higher EPI score and lower stroke, and a larger area yields a higher population.



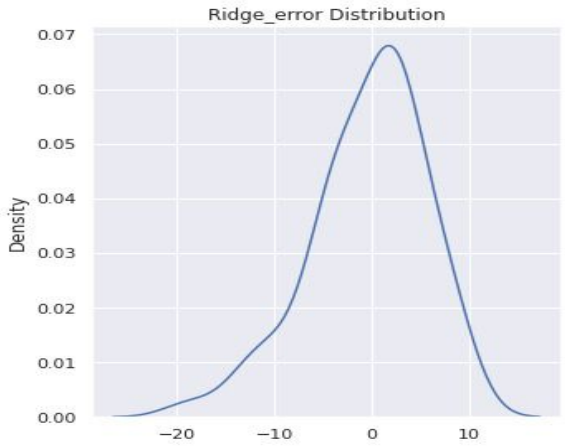
The multiple linear regression model was deployed to perform preliminary analysis. While the obtained values of β vector did manage to generate a model in python, the code of estimating $\widehat{\beta}$ vector showed error in running. The determinant $|X^T X|$ was evaluated and found to be small, and approaching zero, suggestion that $X^T X$ can be a non singular matrix. It was concluded that the data had unbounded variance, attributed to the high similarity between chosen features. Hence, the Ridge and Lasso regression models were chosen to gain further insights from the dataset.

The ridge regression model was then deployed to provide a bound to the estimation of β hat vectors and obtain a model. The model obtained from running Ridge Regression was as follows -

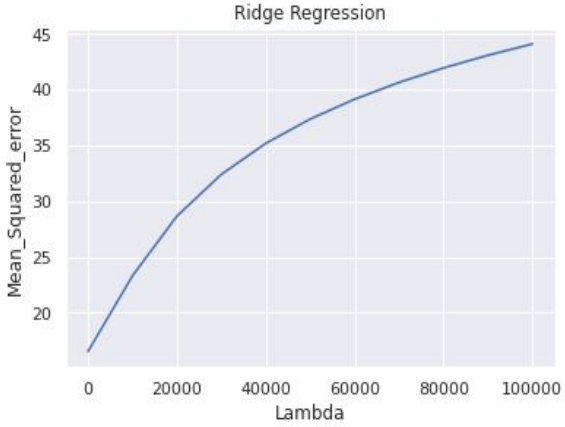
$y=(-3.611*10^{-1})x1+(3.941*10^{-3})x2+(-3.400*10^{-6})x3+(1.557*10^{-1})x4+(3.375*10^{-7})x5+(-1.749*10^{-4})x6+(2.085*10^{-3})x7+(-1.244*10^{-9})x8+(-1.253*10^{-8})x9+(6.746*10^{-4})x10+(-6.796*10^{-2})x11+75.41936073246083$

The value of λ is fixed at 10 for the sake of comparing plots and to be able to grasp the difference between the functioning and the results obtained from the two deployed models.

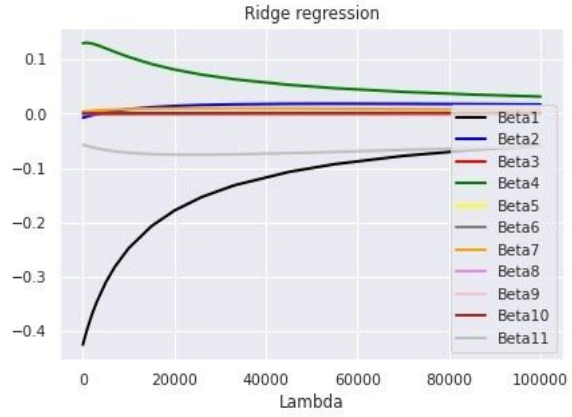
The plot for the mean squared error resembles the gaussian bell curve, but has a non-negative mean, is positively skewed and slightly distorted at the left tail.



The ridge regression model works on a fundamental trade off between the regularisation term and the linear regression term. This establishes a consequent trade off between the regularisation parameter and the least square method. This means that with increasing contribution of the regularisation parameter on the model, the mean squared error is bound to increase. This theory is validated by examining the graph below. As is evident from the increasing curve, at high values of the regularisation parameter the predicted values deviate largely from the target variable, thereby increasing the mean squared error.



Further, a plot between regression parameter β and regularisation parameter λ is obtained and examined. The graph suggests that the values of β decrease as λ is increased. This provides validation to the trade off theory between regression and regularisation components of the model. The quicker the plot reaches zero, the lesser is the sensitivity and hence, the significance of the feature in the model at higher values of regularisation parameter.

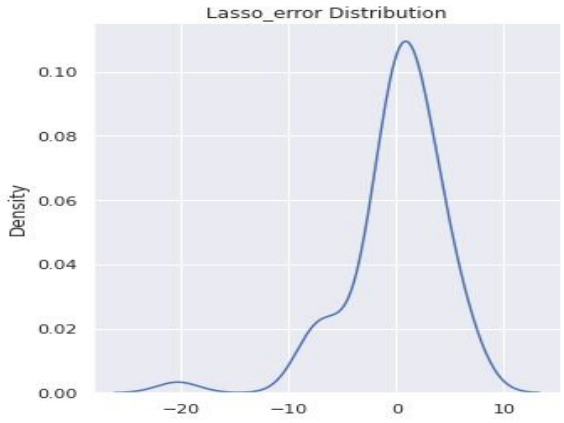


A plot between the regularizing parameter and the β variables of the ridge regression is generated. Each β variable corresponds to a specific variable of the model and the magnitude of each β variable influences the impact of the corresponding variable on the outcome of the model. A positive β variable suggests a positive relationship between the corresponding variable and the dependent variable while a negative β variable suggests a negative relationship between the same. Some of the β variables are observed to take zero value even when the regularization parameter is zero, hence it is concluded that these variables are redundant and do not contribute much to the output of the model. Thus only those variables which have a β vector value not very close to zero can only contribute to the output. It is concluded that β1, β4 and β11 are the prominent regression parameters and thus the variables that correspond to these parameters are the most important variables in the prediction. This helps us in developing an intuitive understanding of the model, and its sensitivity to the selected features.

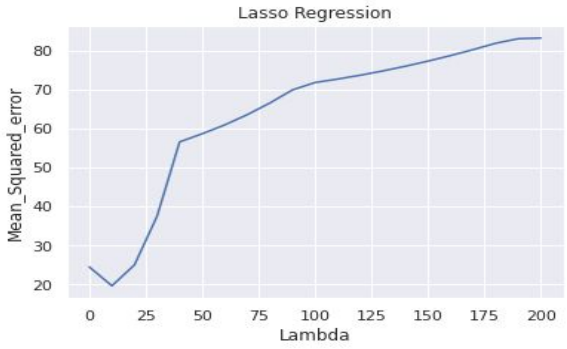
A similar analysis is done on the regularised model obtained from LASSO regression. The model obtained for ($\lambda = 10$) was -

$y=(-2.778*10^{-1})x1+(1.181*10^{-2})x2+(-1.629*10^{-7})x3+(2.177*10^{-2})x4+(9.397*10^{-7})x5+(5.032*10^{-4})x6+(2.464*10^{-3})x7+(-4.984*10^{-9})x8+(-3.268*10^{-7})x9+(7.139*10^{-4})x10+(-7.782*10^{-2})x11+80.75678838445178$

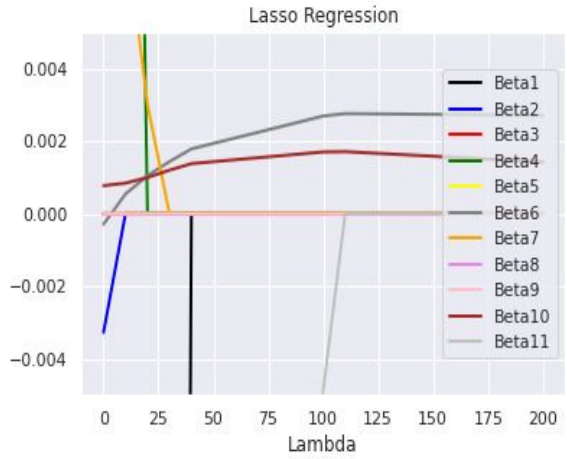
The plot for the mean squared error resembles the gaussian bell curve, but has a non-negative mean, is positively skewed and slightly distorted at the left tail. This distortion is more in comparison to those observed for the ridge plot.



Further, the plot between mean squared error and λ was obtained. As is evident from the increasing curve, at high values of the regularisation parameter the predicted values deviate largely from the target variable, thereby increasing the mean squared error for the deployed LASSO model. This curve is seen to have several non-differentiable points which is expected due to the application of the L1 norm in the model. It can be clearly observed that MSE rises more steeply with increasing lambda as compared to Ridge regression. This is expected, as the L1 norm regularizes the beta coefficients more "strictly"



The plot between regression parameter β and the lasso regularisation parameter λ is obtained and examined. The steepness of these plots are observed to be really high in comparison to those obtained for the ridge regression. The quicker the plot reaches zero, the lesser is the sensitivity and hence, the significance of the feature in the model is low.
This plot reinforces the conclusion that the variables x1,x4 and x11 corresponding to birth rate, EPI, and stroke rate turn out to be variables of primary importance in the model. It also suggests that x2 (cancer rate) has relatively less importance, while x6 (Health Expenditure) also has substantial importance.



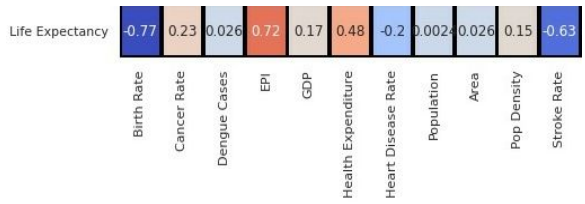
The application accuracy of both models was then evaluated. As is evident from the table below, the value of R squared decreases with increase in ridge parameter, suggesting that model accuracy decreases with increase in λ. This is in line, with our previous inferences from the plots. The table also suggests that for small values of regularisation parameter, LASSO performs better than Ridge on this dataset, while for greater values of λ, a Ridge based model offers better R squared value, hence better model accuracy.

λ	**R2 Ridge**	**R2 LASSO**
0.1	0.6504656	0.747336
1	0.6504545	0.744257
10	0.6503437	0.582726
100	0.6492298	-1.297219
1000	0.6375917	-2.193998

Negative R2 appears when Residual sum of squares approaches the total sum of squares, that means the explanation towards response is very low or negligible. This happens with the Lasso regression model with higher λ values as with higher λ values most of the important feature coefficients tend to 0 and the remaining features are not able to explain variation in the independent variable y.

With the objective of developing a conclusive theory from the above analysis, a pearson coefficient plot is obtained for the target variable life expectancy and the features under consideration. The plot validates the inferences from the β vs λ plot, suggesting that the variables x1, x4, x6, and x11

that correspond to the features birth rate, EPI, health expenditure and stroke rate, have high correlation with life expectancy.



It is inferred that countries with high birth rate have a negative impact on life expectancy. A plausible argument for this is that the countries with higher birth rates have a highly increasing population, and each individual is not able to avail proper facilities due to limited health resources, resulting in lower life expectancy. India ranks 88th in birth rate amongst the countries in the world, as per a UN report in 2020, with 18.2 births/1000 population. Thus, India should make policies focussing on controlling this high birth rate, in order to provide a longer and better quality of life to its people.

It is also inferred that countries with high Environmental Performance Index have higher average life expectancy. The pearson's correlation coefficient is a staggering 0.72, which suggests that a better and greener environment has a high positive impact on life expectancy. This provides conclusive evidence that climate change needs to be handled carefully by the governments, and under no circumstances, can it be ignored. India ranks 168th in EPI in a study performed on 180 countries of the world. This is exceptionally poor, and can be considered as one of the major challenges that the nation is facing.

The analysis also suggests that countries with higher health expenditure have a fairly high life expectancy. India stands at th position amongst 250 countries as of 2020 on health expenditure index. This calls for health centric policies and greater investment in the health domain.

Conclusion

The ridge and lasso regression models offer a handy solution when the explanatory variables suffer from multicollinearity. Their corresponding model accuracies depend upon the extent of regularisation, which is quantified by the value of the parameter λ. Further the trade off between regularisation and linear regression in Ridge and LASSO models is examined in the context of the global healthcare data. It is observed that for higher values of λ, the mean squared error also increases, thereby decreasing the accuracy of the model. For larger values of λ, Ridge works much better, while for smaller values, LASSO provides a more accurate model. LASSO also happens to be much more sensitive to change in the value of the regularisation parameter, as is evident from the β vs λ plot.

The country stands in a tough spot in the healthcare sector. The major pain points are a high birth rate, degrading environment, and very low healthcare funding. Among other political rights, the right to life stands as a basic amenity that people of this world deserve. But these issues of importance get lost in the sky of politics. Continuing political turmoil would make it impossible for the country to tackle these pressing healthcare issues. The situation is upsetting and demands immediate action. A country cannot achieve greatness if its citizens cannot avail of this fundamental right to a better quality of life.

Scope for further research

A more detailed analysis can be done on life expectancy of the countries. This study takes into consideration eleven features, some of which turn out to be fairly correlated to each other. The LASSO model was noted to have a few discrepancies, and a few variables were found to misbehave in the model. The reason for such errors can be explored to gain further insights into the model.
Further research focussing on detailing how features like gender and social diversity impact healthcare can provide interesting insights in the global healthcare space.

Appendix

Python code -
https://colab.research.google.com/drive/14zOv9PwH8naHUgy_AhhaLZPBf98fhx1h?usp=sharing

Dataset -
https://www.kaggle.com/sumaniit/life-expectancy-data