Supervised Learning To Predict Happiness Index Of Countries during CoVID-19 Pandemic

Project: DS203 Programming for Data Science

Jaideep Singh Chawla
Roll Number: 190110030
Department of Metallurgical Engineering and Material Science
IIT Bombay
jdchawla@iitb.ac.in

Raghav Gupta
Roll Number: 190040083
Department of Civil Engineering
IIT Bombay
190040083@iitb.ac.in

***Abstract*— Well-being of citizens can be used effectively to assess a nation's progress. The World Happiness Report is published by the United Nations Sustainable Development Solutions Network every year that ranks national happiness based on how citizens rate their own lives. We study the various factors involved in determining the so-called happiness index and use regression analysis to build a model that predicts happiness index of countries affected by the 2021 pandemic.**

Keywords—component, formatting, style, styling, insert (key words**)**

I. Introduction and Background

The annual report on World Happiness is publicly available. The source for the report is based on Gallup World Poll. Various life-factors are correlated with life evaluations of the respondents of the survey. Their lives are considered between a ladder from 0-10. It should be noted that the variables used in the analysis denote important correlations, not casual estimates. So, data like inequality or unemployment is left out due to lack of availability.

In 2020, especially, the survey results were incomplete and almost 1/3rd were ranked using data available from previous years. We, therefore, develop our own methodology to predict the happiness indices for these countries.

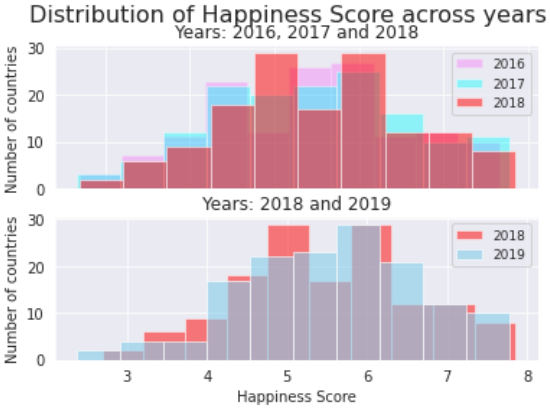
II. Methodology

We started with some Exploratory Data Analysis on the data. We visualized the distribution of Happiness Index (variable Life Ladder) across the years. This was followed by studying the correlation between the Happiness Index and all the features contributing to it by plotting scatter plots between the variable Life Ladder and every feature. A Pearson correlation matrix was prepared. We fitted regression lines for each scatter plot. Through that, we recognized essentially that three features are linearly correlated with the Happiness Index and can be approximated with a straight line to reasonable extents. However, a multiple regression model incorporating all features may be more representative. We moved forward with model selection and employed numerous supervised learning algorithms and trained it with the dataset to find out which algorithm would give the best fit and accuracy.
The hyperparameters were finetuned to maximize R^2 score on 5-fold Cross Validation score for better fit and accuracy respectively.
The various algorithms used were:

- Multiple Linear Regression (all features and top three features)
- Lasso, Ridge and Elastic Net Regression
- Support Vector Regression with kernels - rbf, linear and polynomial
- Neural Network with optimizers - lbfgs, sgd and adam solvers

After having chosen the 'best' model, it is employed to predict the 2020 happiness index for countries that whose data was not recorded in the said period.

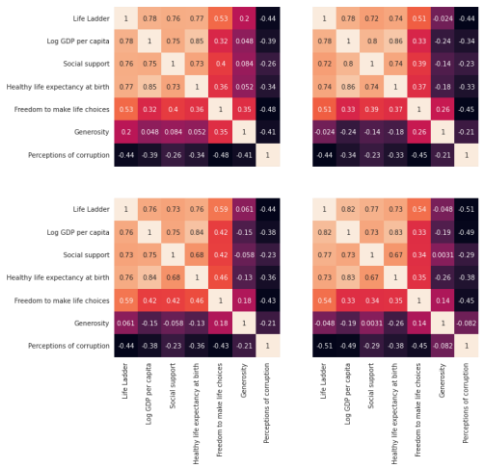
III. Experiment and Results



There was not much variation in the distribution of happiness index across successive years.
A representative ranking for countries is shown as below (2019):

Finland is at the top while Afghanistan is at the bottom of the life ladder.

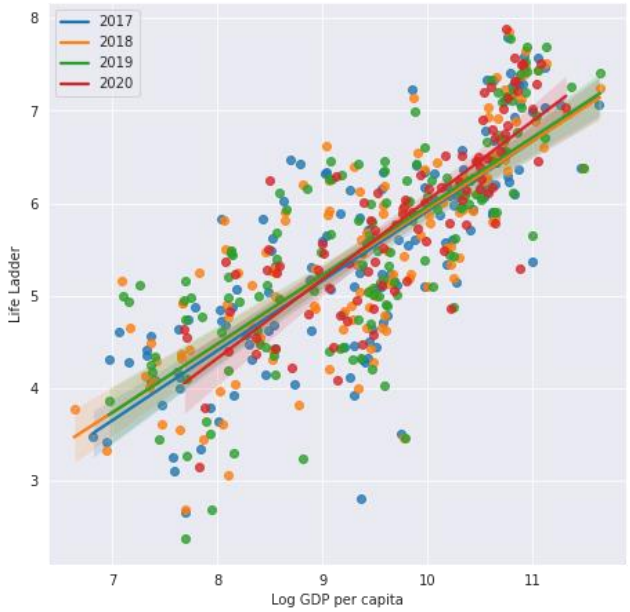
A correlation heatmap shows the graphical representation of correlation matrix representing the correlation between different variables.

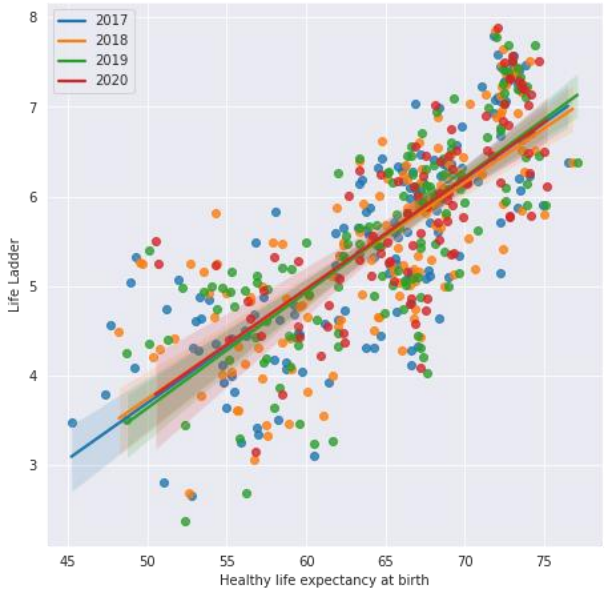


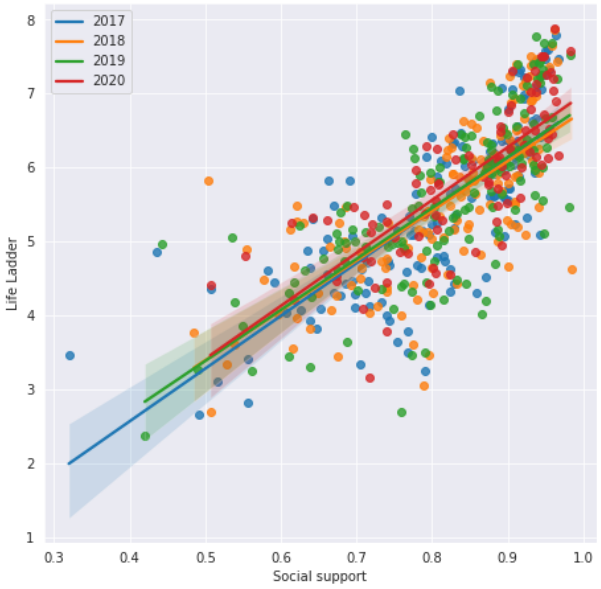
The correlations can be better visualized through scatter plots.

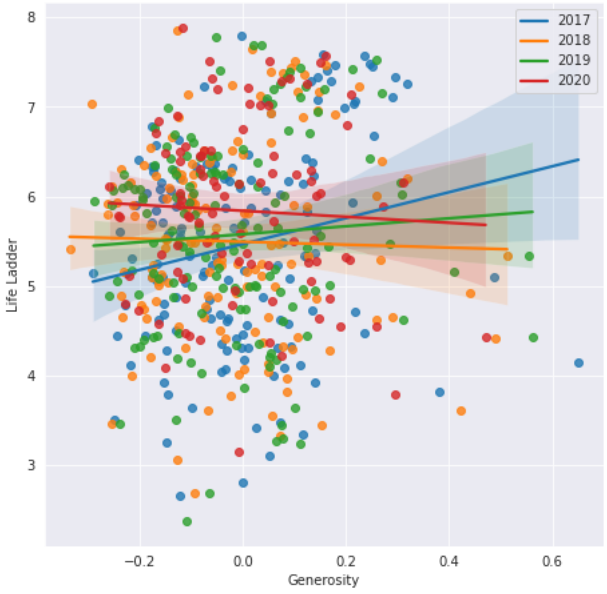
Three features viz. Log GDP per capita, Social Support and Healthy Life expectancy at birth have reasonably good correlation coefficients with Life Ladder.

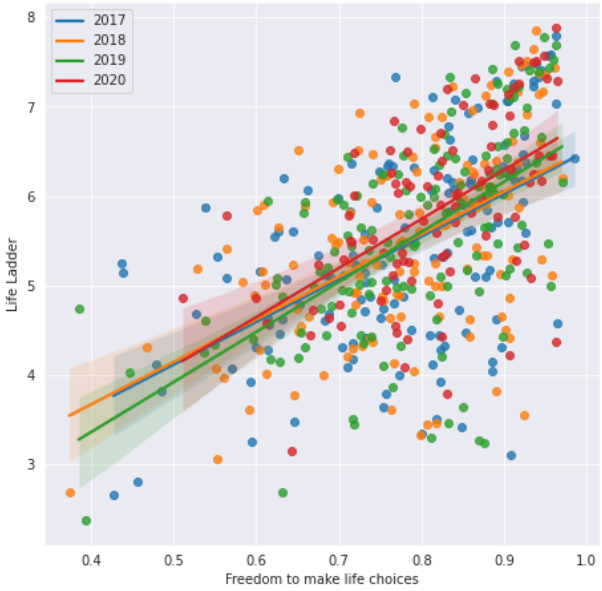
The plots of Life Ladder vs Log GDP per capita, Social Support and Healthy Life expectancy at birth suggest that they are not only well correlated but this correlation tends remain the same for all the years i.e., there is not much variation in the regression line obtained. However, Freedom to make Life Choices, Generosity and Perceptions of Corruption are not well correlated to the Life Ladder and the relation varies a lot for every year.

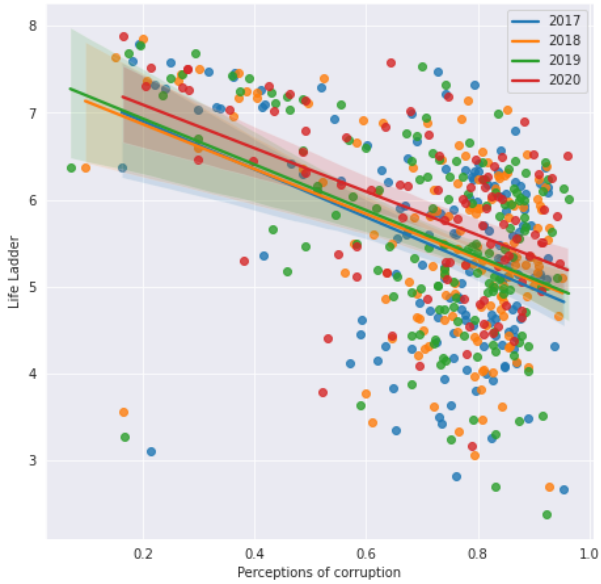












Model Selection:

With the visualization part done, next is the model selection which involves scikit learn library for importing models. The dataset is first split the dataset into 80% training and remaining test data.
Results of the various models used are described in detailed:

I. Multiple Linear Regression:
Multiple regression model was applied with all the features being used. We achieved fair values of RMSE, R2 as well 5-Fold cross validation.

II. Multiple Linear Regression with features most correlated with Life Ladder:
Multiple regression model was applied with the top three features being Log GDP per capita, Social Support and Healthy Life expectancy at birth. We achieved poorer values of RMSE, R2 as well 5-Fold cross validation. Hence this model was not as useful.

III. Lasso, Ridge and ElasticNet Regression:
We applied regularization with a list of learning rates and selected the best learning rate on the basis of cross validation score for each model. We saw that none of the models could give better results than our previous models.

IV. Support Vector Regression with kernels: rbf, linear and polynomial:
For each of the three kernels, we applied regularization with a list of learning rates and selected the best learning rate on the basis of cross validation score for each model as done earlier.
We observe that rbf kernel with C= 10000 gives the best cross validation score and outperforms all the previous models.

V. Neural Network with lbfgs, sdg and adam solvers:
Taking hidden layer size as (10,5) We achieved lower RMSE values and fair R2 and cross validation scores.

Evaluation Table

Model	Details	RMSE	R-squared-training	R-squared-test	5-Fold Cross Validation
Multiple Linear Regression	All features	0.5396	0.7286	0.7889	0.7214
Multiple Linear Regression	Top 3 features	0.5947	0.6738	0.7436	0.674
Lasso Regression	alpha = 0.001	1.1747	0.7285	0.7883	0.7217
Ridge Regression	alpha = 1	0.9071	0.7284	0.7883	0.722
ElasticNet Regression	alpha = 0.001	1.1747	0.7283	0.7881	0.722
SVR(kernel=rbf)	C = 10000	1.1747	0.7576	0.8163	0.7447
SVR(kernel=linear)	C = 0.1	1.1747	0.7239	0.7858	0.7217
SVR(kernel=poly)	C = 100	1.1747	0.745	0.8013	0.742
Neural Net	lbfgs	0.5182	0.7444	0.8053	0.7299
Neural Net	adam	0.5851	0.6978	0.7518	0.6898

Final Prediction Results, visualized: