GLIS-692-001
Special Topics 2
Data Science Report

World Happiness Report Dataset

Aryan Gupta
260807226

I. INTRODUCTION

a. About the Dataset

The dataset utilized in this report is called World Happiness Report for the year of 2017. The numeric values and rankings have been collected from the Gallup World Poll. The answers to the survey are based on real-life evaluation questions. These types of questions, known as Cantril ladder, ask the person to imagine a ladder and scale it on the basis on their understanding of Best Possible Life as 10 and Worst Possible Life as 0. Based on the parameters, FAMILY, LIFE EXPECTANCY, ECONOMY, GENEROSITY, TRUST IN GOVERNMENT, FREEDOM and DYSTOPIA RESIDUAL, a Happiness Score for the whole country is calculated and ranking is done.

b. Research Question

The research question for this report is to find how numeric value of Economy GDP of a country is affected by the parameters of the dataset namely FAMILY, LIFE EXPECTANCY, GENEROSITY, TRUST IN GOVERNMENT, FREEDOM and DYSTOPIA RESIDUAL.

c. Target Audience

The target audience for this data science report is the people who have formed the major governing bodies in the country. By using this analysis, they can come out with efficient plans for governing the country. Let's say they encounter a parameter that is heavily affecting the happiness score of the country. They can take measures to fix the factors that decrease the value of that parameter. This will increase the happiness score of the country.

II. METHODOLOGY

a. Parameters of the Dataset

Parameters of the Happiness Score Report Dataset are economic production, social support, life expectancy, freedom, absence of corruption, and generosity

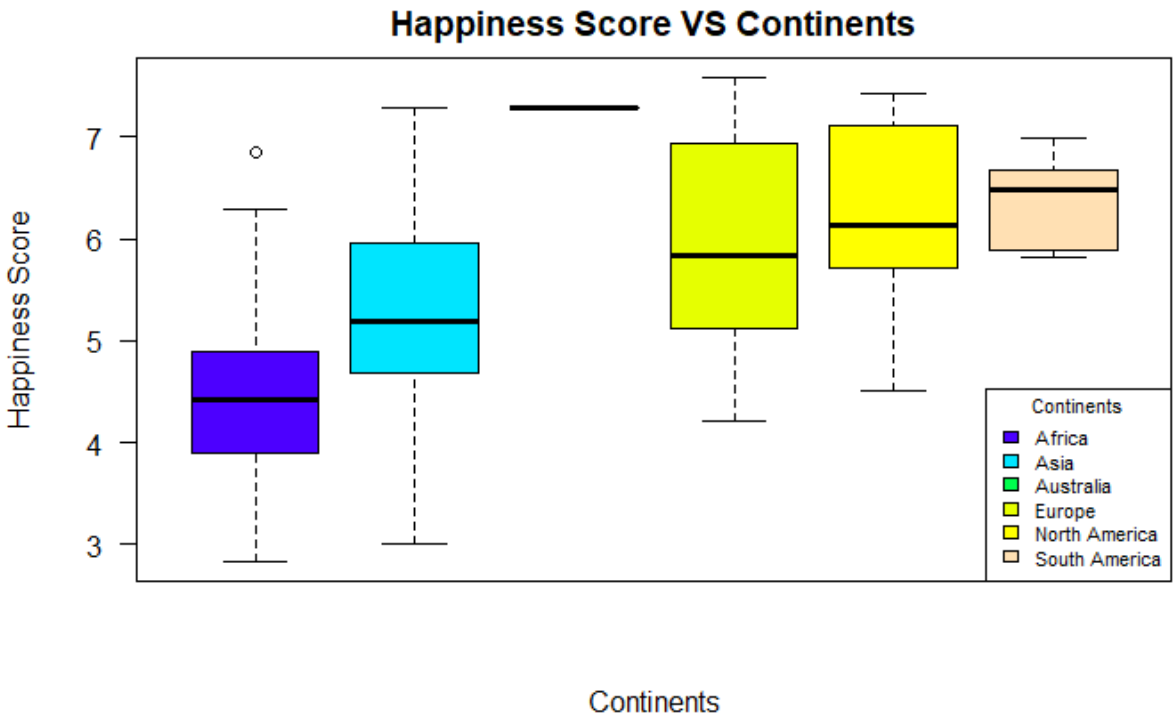
1. Economy (GDP per Capita) [# NUMERIC]: Degree of Happiness score affected by Economic GDP.
2. Family [# NUMERIC]: The degree to which Family contributes to the calculation of the Happiness Score
3. Health (Life Expectancy) [# NUMERIC]: The degree to which Life expectancy affects the calculation of the Happiness Score
4. Freedom [# NUMERIC]: The amount of Happiness Score calculations are affected by Freedom.
5. Trust (Government Corruption) [# NUMERIC]: The degree to which Perception of Corruption contributes to Happiness Score.
6. Generosity [# NUMERIC]: The extent of Happiness Score calculations affected by Generosity value.
7. Dystopia Residual [# NUMERIC]: The above 6 parameters contribute towards making life evaluations higher as compared to these evaluation values in dystopia, a hypothetical country where the values of the 6 parameters are among the lowest global average. Dystopia is used as a benchmark since no country would be worse than dystopia, where the happiness score is the least. The Dystopia Residual metric is the Dystopia Happiness Score (1.85) + the Residual value (unexplained value for each country).

Happiness score for a country is calculated by adding numeric values of all these factors. Since Happiness score is directly related to all these factors, thus all of them have a high correlation with the Happiness Score. Thus, instead of making an analysis on Happiness Score, I have chosen to make a prediction on Economy GDP of a country. The Happiness Score column is not used in the coming analysis and only FAMILY, LIFE EXPECTANCY, TRUST IN GOVERNMENT, FREEDOM and DYSTOPIA RESIDUAL are used.

b. Descriptive Graphs

Shown below are some graphs that describe the dataset

i. Boxplot for Happiness Score VS Continents



The boxplot above shows the range on Happiness Score of the countries in the respective continents. It shows that Africa has the lowest happiness score range which is because they have low Economy GDP and have the highest Corruption. These factors heavily affect the Happiness Score of the country.

ii. Scatter Plot with regression line of Economy GDP VS Happiness Score

The above scatter plot reveals the relationship between Economy and Happiness Score. It shows that they are directly proportional and if one increases, the other increases as well.

c. Analytical Methods used for Predictions

Analytical Methods used in the report for making predictions are Multiple Linear Regression and Support Vector Regression.

Multiple Linear Regression was used since

- This is used since it is a regression problem and Linear Regression cannot be used until the problem is converted to a classification problem.
- This model acts as a baseline to find a correlation between ECONOMY with the feature space.

Support Vector Regression was used since

- This model is implemented to outperform the Multiple Linear Regression Model to predict the dependent variable (Economy) by using the independent variables.

- SVR is a good choice since it has faster execution times since it only stores the support vectors and works well with small feature space.

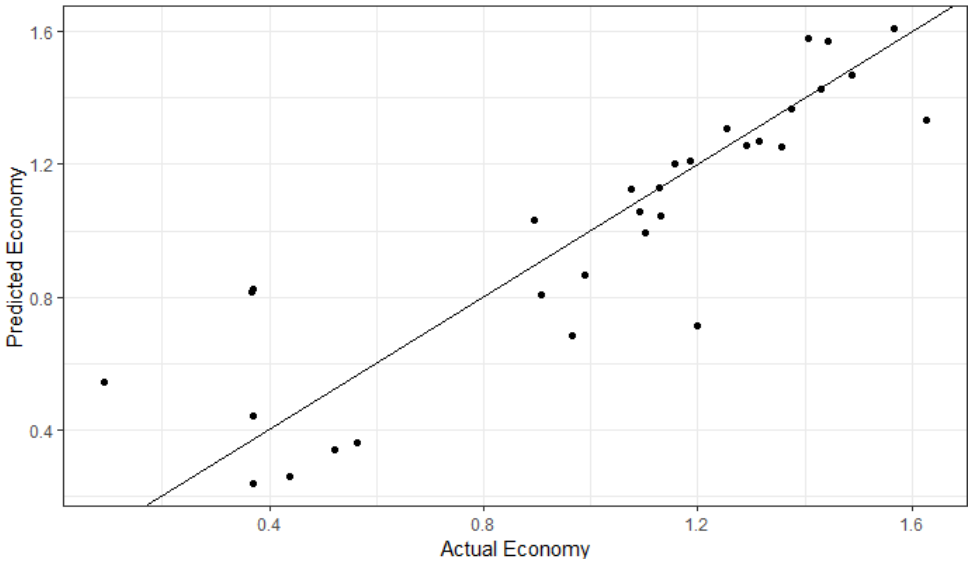
III. IMPLEMENTATION & RESULTS

a. Multiple Linear Regression

i. Implementation and Charts

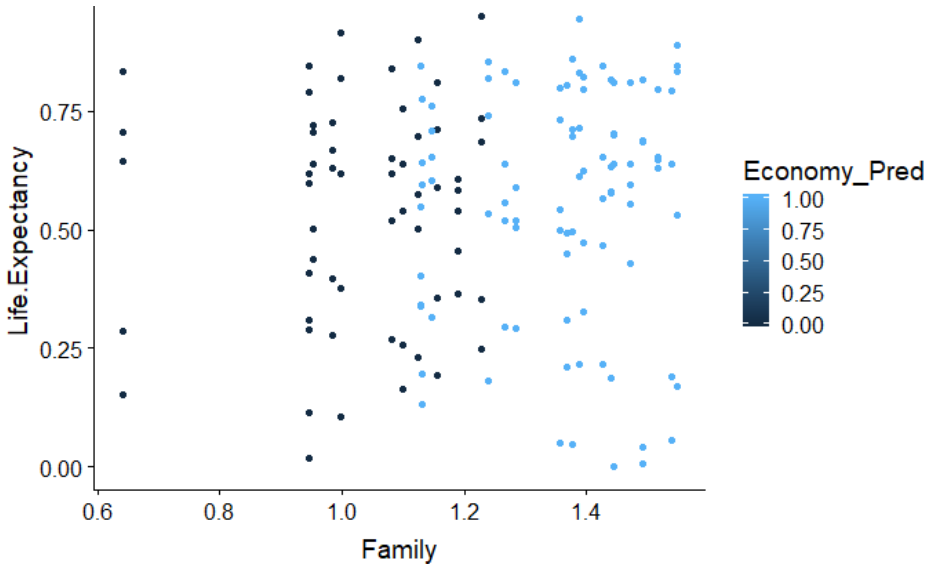
Before implementing the Multiple Linear Regression model to the dataset, feature selection was done to make the best model out of all the parameters. For feature selection, backwards elimination was used where the threshold for P-Value was chosen as 0.001 After backward elimination, the best features for prediction came out to be 'Family' and 'Life Expectancy'. The charts made after implementing Multiple Linear Regression are:

a)



The above graph shows the correlation between the predicted values and the regression line which has been plotted based on True Values. According to the graph, the predicted values for Economy have the highest error around the value 0.4. The cause of this is due to the low values since most of the Economy Values are above the average value i.e. 0.9.

b) For this chart, I had to convert the regression problem to a binary classification problem to plot the predicted Economy GDP Datapoints, color-coded based on 2 classes i.e. ‘Family’ and ‘Life Expectancy’.



The above graph shows the distribution of datapoints color-coded with respect to the ECONOMY GDP value being 1 if more than mean value and ECONOMY GDP value being 0 if less than the mean value. From the above graph, it is obvious that the economy values are low, lower than the average value, where the value of Family is low. Most of the above average economy values are present where ‘Family’ and ‘Life Expectancy’, both, have a high numeric value.

ii. Validation Statistics

- ✓ R Squared is a statistical measure of how well the data is fitted to the regression line. More the R Squared value better is the fit.

R Squared = 0.768565108135112

- ✓ Adjusted R Squared is a modified version of R Squared and is adjusted based on the number of predictors. The values only increase if a new significant predictor is added to the model. More the value better is the fit of datapoints on the regression line.

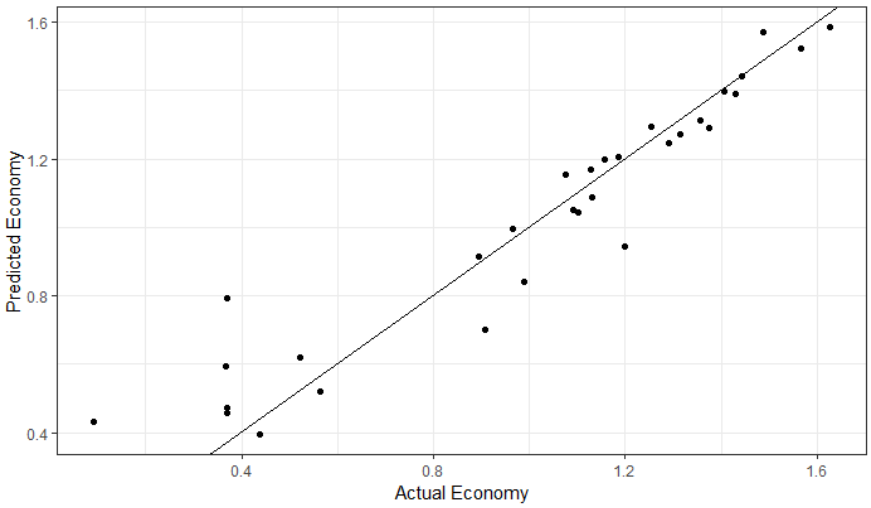
Adjusted R Squared = 0.762779235838489

b. Support Vector Regression (SVR)

i. Implementation and Charts

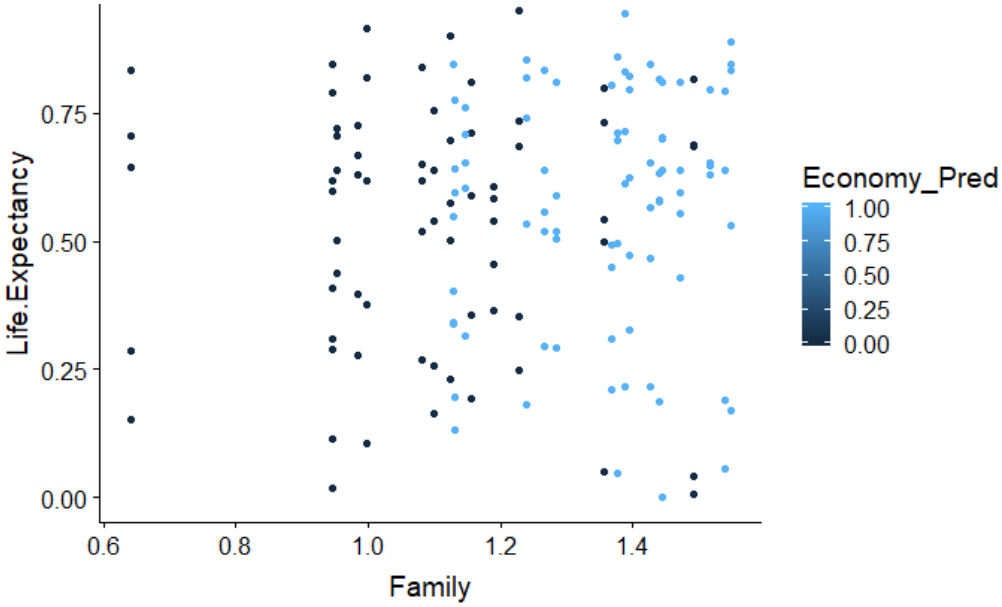
For the SVR model, no feature selection was done. The charts made after implementing Support Vector Regression were

a)



The above graph shows the correlation between the predicted values and the regression line which has been plotted based on True Values. Like the Multiple Linear Regression Correlation Plot, the predicted values for Economy have the highest error in the same region, where Economy values are 0.4. The amount of data available is the cause of this plot, most of the values have values above the mean value i.e. 0.9. But now have less error as compared to the Multiple Linear Regression Correlation Plot.

c) For this chart, I had to convert the regression problem to a binary classification problem to plot the predicted Economy GDP Datapoints, color-coded based on 2 most significant classes i.e. ‘Family’ and ‘Life Expectancy’.



The above graph shows the distribution of datapoints color-coded with respect to the ECONOMY GDP value being '1' if more than mean value and ECONOMY GDP value being '0' if less than mean value. From the above plot, we can conclude that Economy values are below average if Family value is low. Even now, most of the above average Economy values are in the section where the Life Expectancy and Family values are high. Difference between the Multiple Linear Regression graph is, the number of Economy values below average i.e. 0.9 is more in region with high Family value.

ii. Validation Statistics

- ✓ RMSE (Root Mean Squared Error) is a measure of the amount of error between the True and Predicted Values in the regression problem. Less the value, less is the error, better the predictions.

Root Mean Squared Error = 0.133646185220686

IV. RECOMMENDATIONS

From the above analysis and conclusion, some recommendations as to how to utilize them are stated below.

- If a governing body is looking to increase the Happiness of an area, they can use these conclusions to help their cause since it is evident from the descriptive graphs that Happiness is directly proportional to Economy GDP.
- If someone is looking to increase the average Economy GDP of an area, he should for ways to increase the Life Expectancy. This can be done by providing better health plans and providing free medical checkup for kids of age less than 10, to make sure that they all have been vaccinated for known diseases.
- If people are provided with facilities that make them take out time for their families, that can also lead to a rise in average Economy GDP, as concluded by the Support Vector Regression charts. To make it possible, parks and picnic spots can be constructed or cleaned.

V. CONCLUSION

So, form the information and analysis done in this report, we can conclude that happiness is directly related to all the parameters, i.e. FAMILY, LIFE EXPECTANCY, ECONOMY, GENEROSITY, TRUST IN GOVERNMENT and FREEDOM. If you want to increase the average happiness of the country, working to improve any one of the parameters will lead to an increase in its average.

The same cannot be said for increasing the average value of Economy GDP of an area. Not all parameters are equally relevant to increasing the average value. The most significant parameters which bring the highest growth in Economy GDP are LIFE EXPECTANCY and FAMILY.

In the future, we can add more data to the original dataset based on countries like population or average temperature, which may give us more insight as to how the Economy of a country is affected.

BIBLIOGRAPHY

- Sustainable Development Solutions Network. (2017, June 14). World Happiness Report. Retrieved from https://www.kaggle.com/unsdsn/world-happiness
- Helliwell, J., Layard, R., & Sachs, J. (Eds.). (2017, March 20). WORLD HAPPINESS REPORT 2017. Retrieved from https://worldhappiness.report/ed/2017/
- Building Regression Models in R using Support Vector Regression. (n.d.). Retrieved from https://www.kdnuggets.com/2017/03/building-regression-models-support-vector-regression.html