Healthy Life Expectancy based on Happiness Indicators

Fjolle Gjonbalaj

May 13, 2021

Abstract

This paper explores the topic of Healthy Life expectancy based on happiness indicators. A special focus is provided towards indicators such as perception of corruption, social support, freedom to make life choices, generosity and log GDP per capita. Statistical models such as the linear hand-built model, lasso model, bagging model and random forest are considered in order to come up with the best performing model for the happiness data considered in the paper. While the bagging and the random forest model perform similar to one-another, lasso is the best performing model among the four.

Keywords: Average healthy Life Expectancy, Social Security, Generosity, log GDP per capita, Freedom to make life choices, Perception of corruption

1 Introduction

The perception of corruption in a country is a crucial component when it comes to the happiness of the people and their healthy life expectancy (Achim, 2019). However, corruption is not the only indicator of happiness. Other variables such as social support, freedom to make life choices, generosity as well as the log GDP per capita also play a key role on the happiness of a nation (WorldHappinessReport, 2021).

In order to construct a statistical model that accurate predicts happiness levels across the world I apply four different statistical models based on the variable Healthy life expectancy: The linear model, Lasso, Random Forest and Boosting. Predictions using each one of the above models are then compared to observe which one provides us with the best performance. I analyze different healthy life expectancy broken-down by region of the world in order to observe the score of the factors adding to the population's happiness level. Hence, this paper is structured in the following way: In the first section I show Healthy life expectancy based on different happiness indicators broken down by individual regions of the world, and then compare findings. In the second section I build four different statistical models: The hand-built linear model, the lasso model, bagging and random forest. In the third section I then compare the four models and based on my findings conclude which one of the four statistical models is the most appropriate for analysis. Lastly, I draw conclusions in Section four.

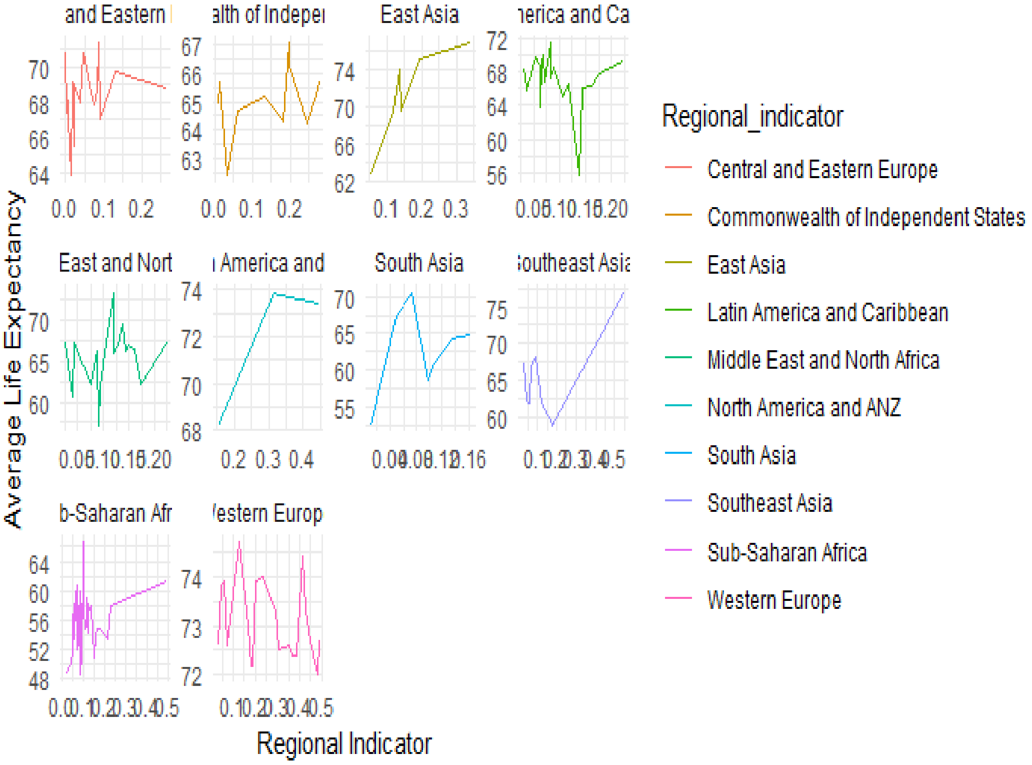
2 Methodology

The data-set used in this project is data from the Gallup Poll and The World Happiness Report 2021. The data is based on answers to the main life evaluation questions which are asked on the poll, also referred to as the Cantril ladder. The poll asks respondents to rank their lives on a 0 to 10 scale, with 10 being the best possible life while 0 being the worst possible life (WorldHappinessReport, 2021). The typical annual sample for each country is a minimum of 1000 people and can get as high as 3000 respondents. The data-set has its shortcomings. That is, the data-set is based on the year 2021 in the midst of the COVID-19 pandemic. All the variables in the data-set have the potential of being affected by this historic event in comparison to other "normal" years without a pandemic. For instance, the unfortunate death of many seniors may cause the costs of Social Security to drop in the short-run. However, the elderly contribute a very small percentage of social security in comparison to the entire working population. As more and more people lost their jobs, they stopped contributing to the Social Security Program through payroll taxes, which is a major funding source for the program. Thus, in the longer run this may actually pivot into the opposite direction (Burling, 2021). As a result, the healthy life expectancy based on social security will also be affected. Moreover, many other components such as social relationships have a crucial impact on a healthy life expectancy and longevity for people, yet are not included in this data set(J,R, 2021). In addition, the data is based on the perception of the indicators in the study rather than the indicators themselves. Although survey data is not always the most reliable source, especially when it comes to subjective matters such as happiness which varies from person-to-person, it is the only metric we have for this type of analysis. With this in mind, the data-set is sufficient to enable me to show statistical model results.

After data cleaning the following variables are used for the analysis on happiness:

- Regional Indicator
- Logged GDP per capita
- Social Support
- Healthy life expectancy
- Freedom to make life choices
- Generosity
- Perception of Corruption
- Explained by Log GDP per capita
- Explained by Social Support
- Explained by Healthy life expectancy
- Explained by Freedom to make life choices
- Explained by Generosity
- Explained by Perception of corruption

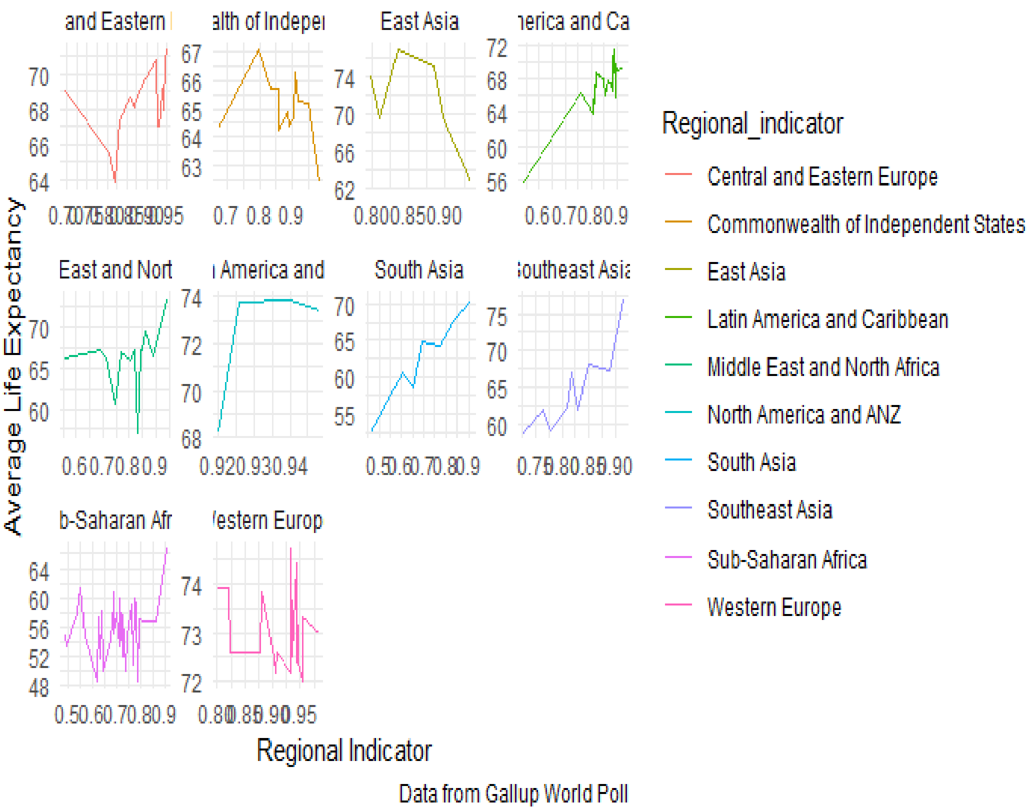
3 Visual Representation



Data from Gallup World Poll

Figure 1: Healthy Life expectancy based on the Perception of Corruption

Figure 1. is a representation of the Average Life Expectancy based on the perception of corruption for each Region specified. It is evident from the line plots for each region that perception of corruption has a distinct effect on the Average life expectancy. For the Central and Eastern European graph we see that as the Perception of Corruption index increases, the Average Life Expectancy decreases. A similar result is observed for Western Europe and to some degree for Latin America and the Caribbean. However, the results are not as clear for the other regions of the world. The results from this graph are somewhat puzzling given that we know from previous evidence that corruption has a bigger impact on the poorer regions of the world than it does on the richer (Nadpara, Samanta, 2015). An additional factor to consider is the way the corruption as well as the Average Life Expectancy are calculated in the happiness dataset. The dataset used for this project is entirely based on the survey scores, using the Gallup weights which are used to make the estimates representative (WorldHappinessReport, 2021). On the other hand, the paper by Nadpara and Samanta is based on panel data estimations of 30 countries look at corruption and its effect at birth. Another factor to consider is whether or not the degree of cultural differences within regions contributes to this variation.



Data from Gallup World Poll

Figure 2: Healthy life expectancy based on Social Support

Figure 2. shows the Average Life Expectancy based on Social support within a region. From the line graphs it can be observed that for the majority of the regions, an increase in the Social Security index leads to an increase in the Average Life Expectancy, with the exception of Commonwealth of Independent States, East Asia and Western Europe. The results from the graph are rather consistent with what we would also expect to see. That is, in the US, for instance, social support in entirely based on the level of income or the number of dependents. In other parts of the world, such as Central and Eastern Europe the social support is need based and also depends on other factors, such as gender.

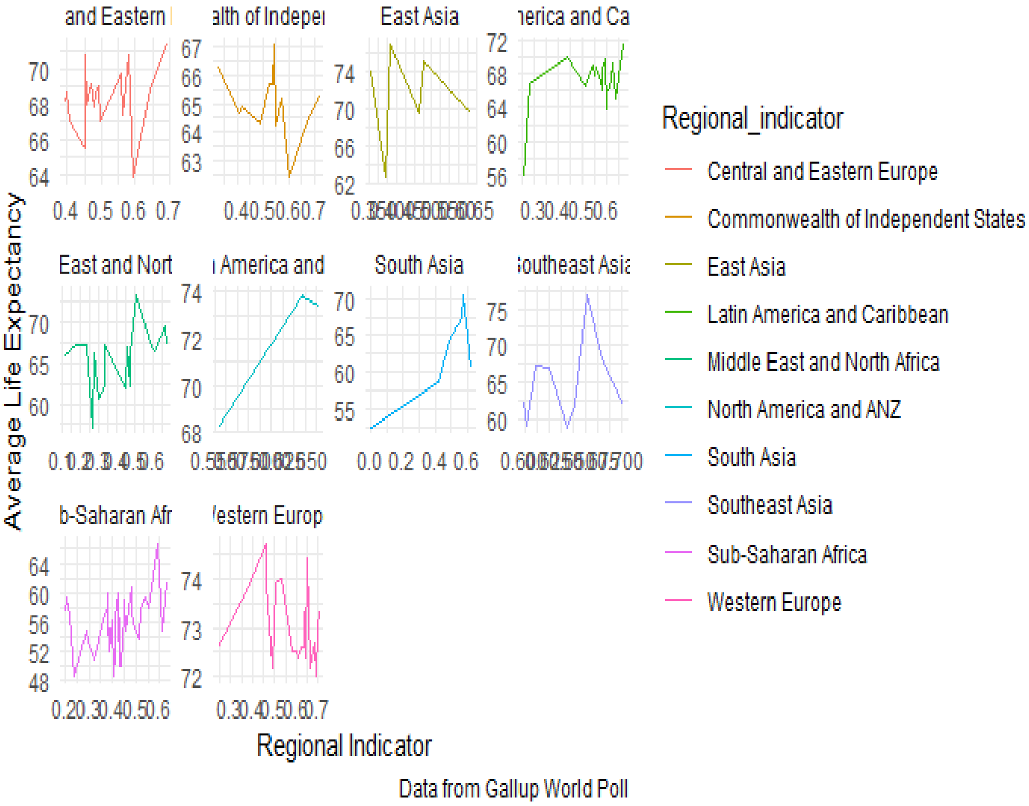
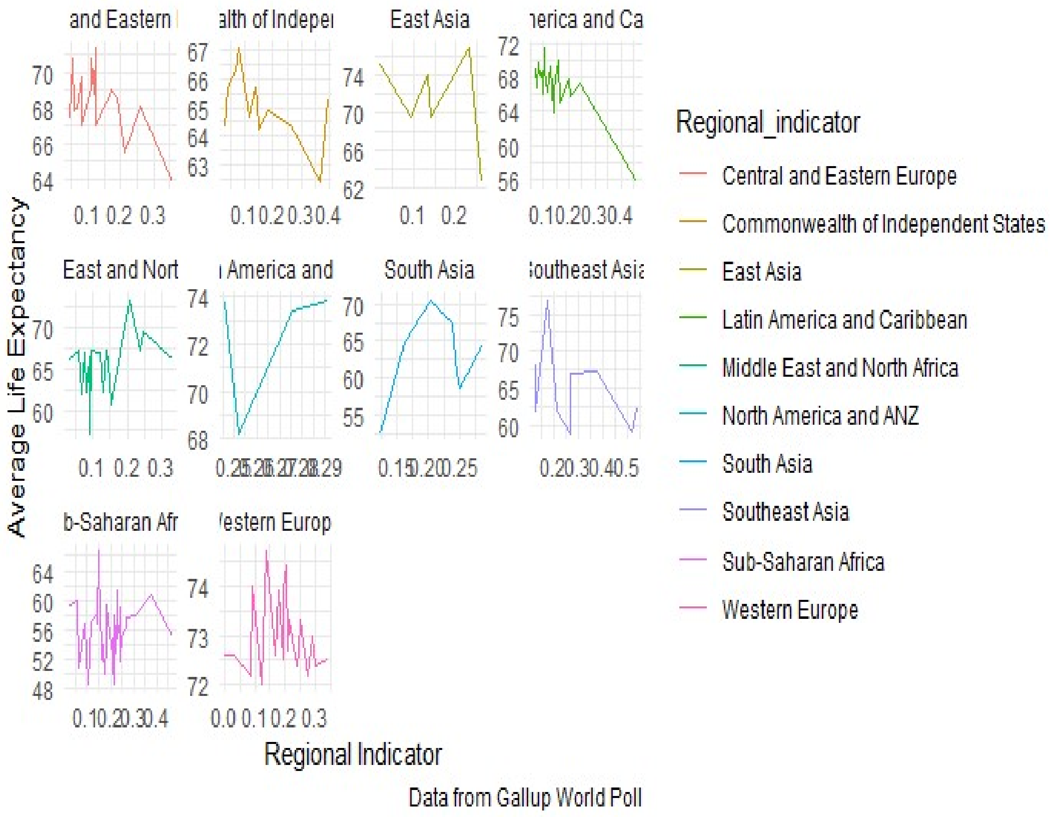


Figure 3: Healthy Life expectancy explained by the freedom to make life choices

Figure 3. shows the Average Life Expectancy as a result of the level of freedom to make life choices within a region. Here we can observe that for Central and Eastern Europe there is an upward trend on the line graph, indicating that as the perception by the people on the freedom to make life choices increases, the average life expectancy increases in tandem. This trend is observed for other regions such as Latin America and Caribbean, Middle East and North America, North America and ANZ, South Asia and Sub-Saharan Africa. Some of the more puzzling results are observed for East Asia, Southeast Africa and Western Europe.



Data from Gallup World Poll

Figure 4: Healthy Life expectancy explained by generosity

Figure 4. Shows the Average Life Expectancy based on the perception of generosity in the region. The results from this graph convey that in Central and Eastern Europe, as the perception of generosity increases, life expectancy declines. This is not a result we would expect to see. However, going back to one of the main shortcomings of the data set used in this project, more often than not people tend to rank their countries as not being enough generous towards them. This, in turn, would have affected the shape of the line and its direction. This problem does not seem to affect regions such as Latin America and Caribbean and to a certain extent South Asia and Sub-Saharan Africa.

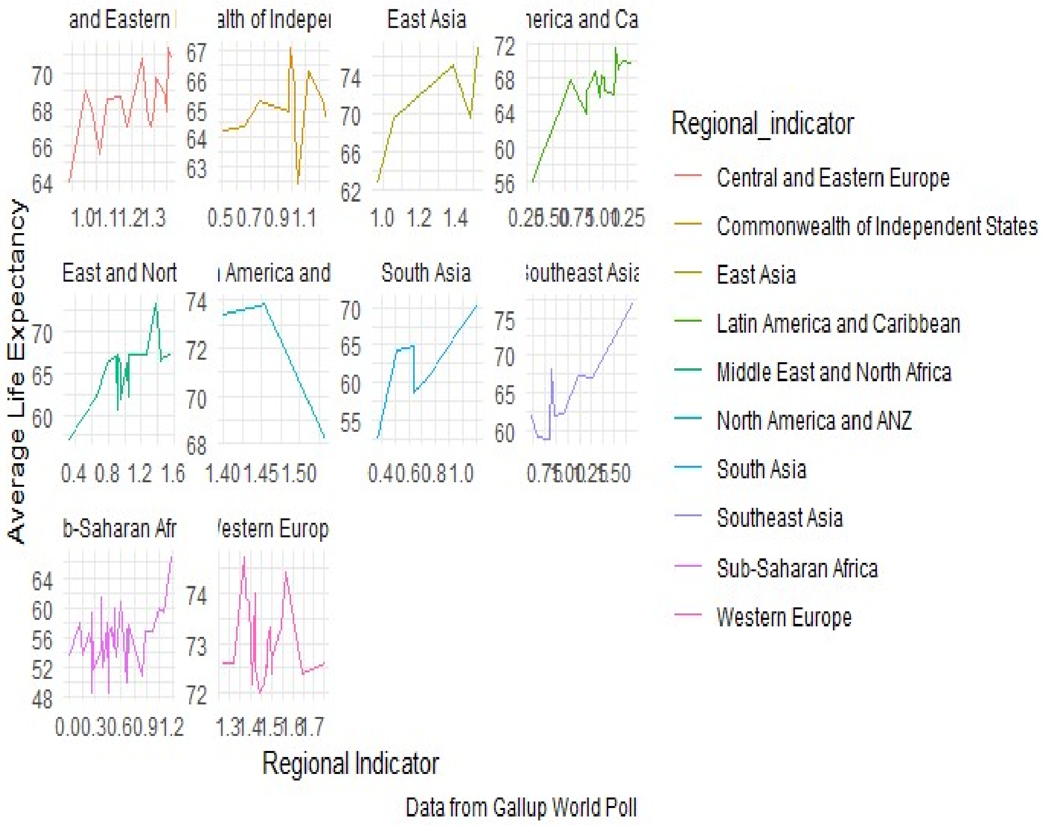


Figure 5: Healthy Life expectancy explained by log GDP per capita

Figure 5. displays the line graphs for the Average Life Expectancy based on the log GDP per capita. All of the regions, but North America and ANZ show an upward trend when it comes to the log GDP per capita and its effect on the Average Life Expectancy.

4 Statistical models

Figure 6. shows the Lasso model trace plot for all the coefficients of the model. In this plot, each colored line represents the value of a coefficient in the model. L1 Norm is the weight provided to the regularization term, also referred to as lambda. Hence, as lambda approaches zero, the loss function in the model approaches the Ordinary Least Square loss function. As lambda increases, the regularization term will have a larger effect and fewer variables in the model are observed. This is due to more of the coefficients being valued at zero. The hand-built model was used as a base model from which comparisons can be drawn. In this case it can be observed that the lasso model performs slightly better than the hand-built linear model, even though the difference between the two is

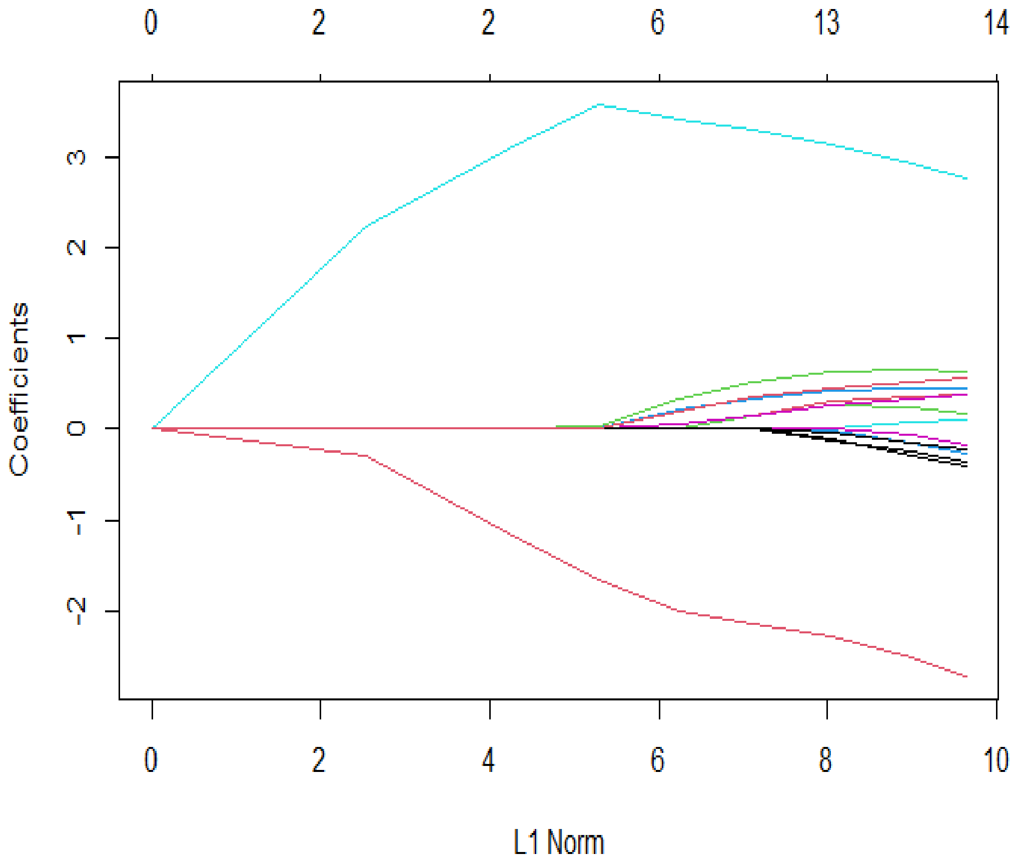
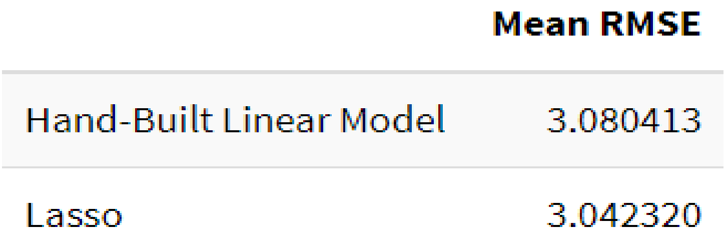


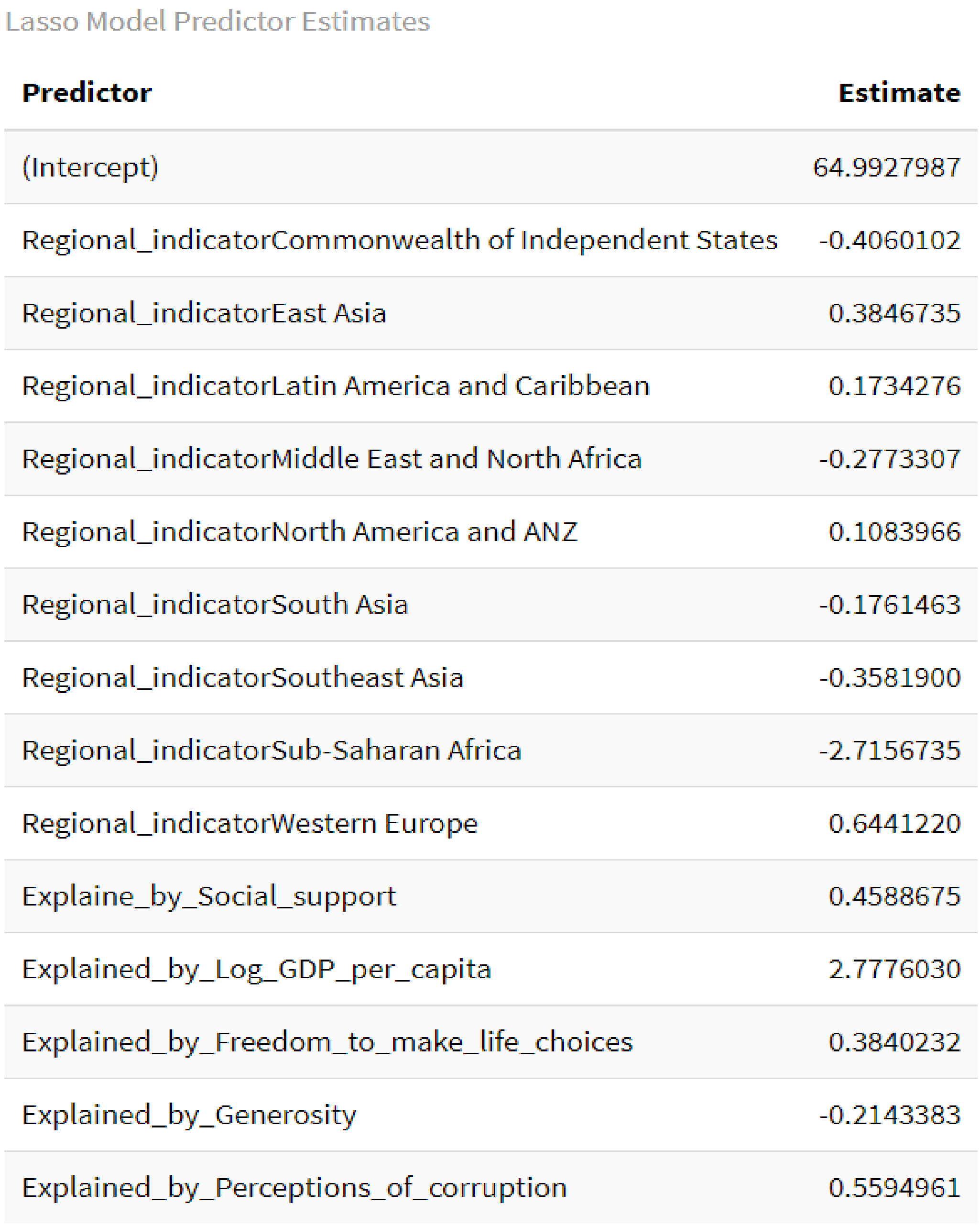
Figure 6: Lasso model trace plot

very small, with the hand-built model giving a mean RMSE of 3.080413 and the Lasso model a mean RMSE of 3.042320.



	Mean RMSE
Hand-Built Linear Model	3.080413
Lasso	3.042320

Figure 7: Mean root mean squared error



Lasso Model Predictor Estimates

Predictor	Estimate
(Intercept)	64.9927987
Regional_indicatorCommonwealth of Independent States	-0.4060102
Regional_indicatorEast Asia	0.3846735
Regional_indicatorLatin America and Caribbean	0.1734276
Regional_indicatorMiddle East and North Africa	-0.2773307
Regional_indicatorNorth America and ANZ	0.1083966
Regional_indicatorSouth Asia	-0.1761463
Regional_indicatorSoutheast Asia	-0.3581900
Regional_indicatorSub-Saharan Africa	-2.7156735
Regional_indicatorWestern Europe	0.6441220
Explaine_by_Social_support	0.4588675
Explained_by_Log_GDP_per_capita	2.7776030
Explained_by_Freedom_to_make_life_choices	0.3840232
Explained_by_Generosity	-0.2143383
Explained_by_Perceptions_of_corruption	0.5594961

Figure 8: Lasso model predictor estimates

The third statistical model performed is Bagging. A random sample of 6 explanatory variables are chosen from the entire group of predictors. Averaging the predictors is utilised to improve on the accuracy of the predictors and control over-fitting. Due to each split using only 1 out of the 6 predictors, a sample of 6 predictors is considered for each individual split. The mean of the squared residuals is just below 9, while the percentage of the variation explained by the model is approximately 80.

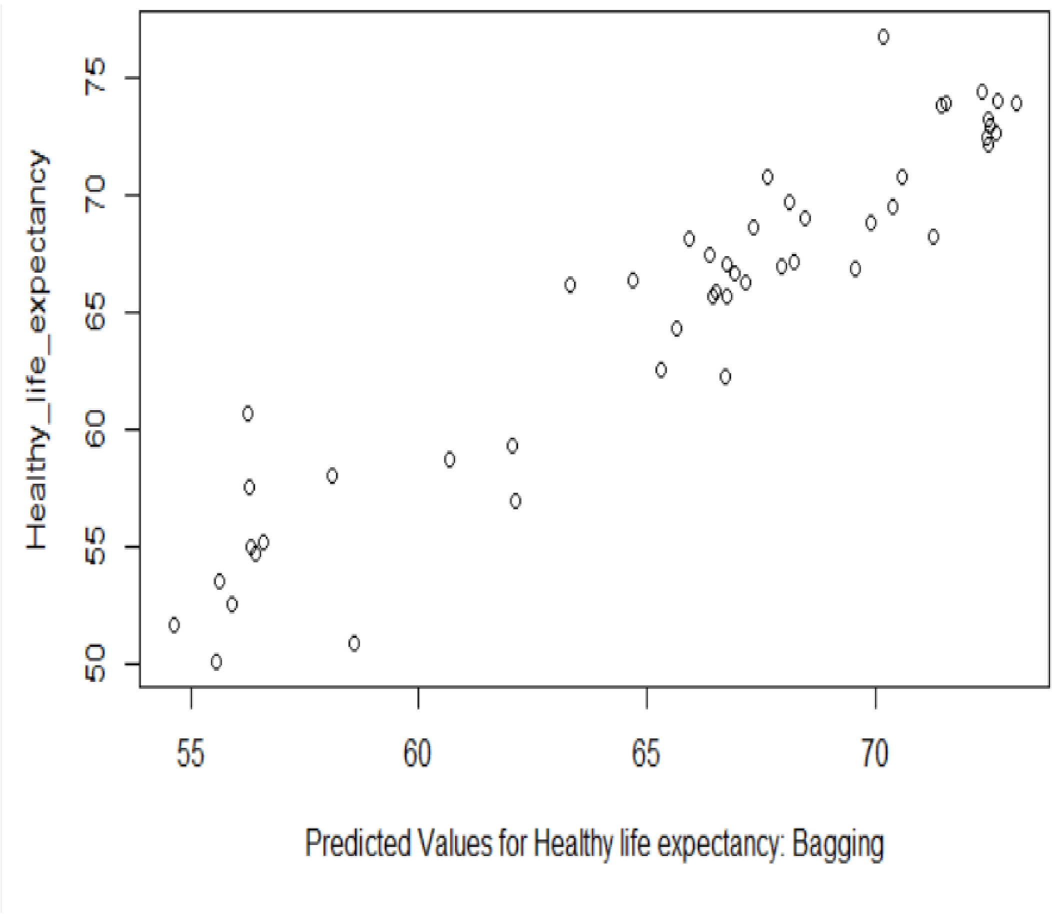
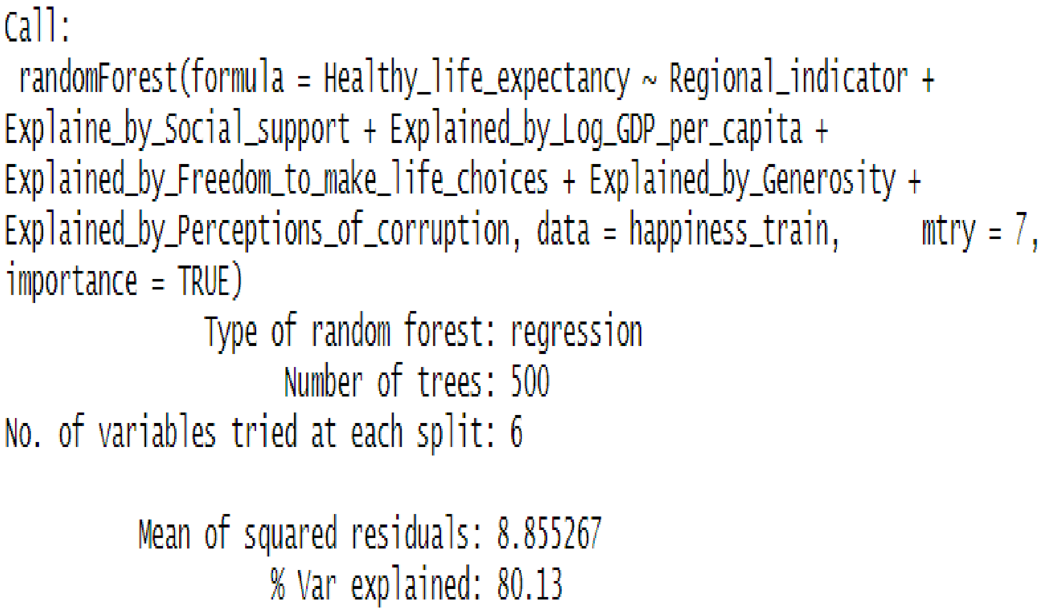


Figure 9: Bagging



```
Call:
 randomForest(formula = Healthy_life_expectancy ~ Regional_indicator +
Explaine_by_Social_support + Explained_by_Log_GDP_per_capita +
Explained_by_Freedom_to_make_life_choices + Explained_by_Generosity +
Explained_by_Perceptions_of_corruption, data = happiness_train,     mtry = 7,
importance = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 6

          Mean of squared residuals: 8.855267
                    % Var explained: 80.13
```

Figure 10: Lasso model

The fourth model I perform is Random Forest. Using this model it can be examined how the Average life expectancy changes based on a sample of variables used in this model. Although conveniently computed, the random forest model does not come without any shortcomings. Interpreting the results from the random forest model can be challenging. This is because this model does not account for all the variables in the model. To overcome this, it might be useful to explicitly specify variables that we want to include in the model.

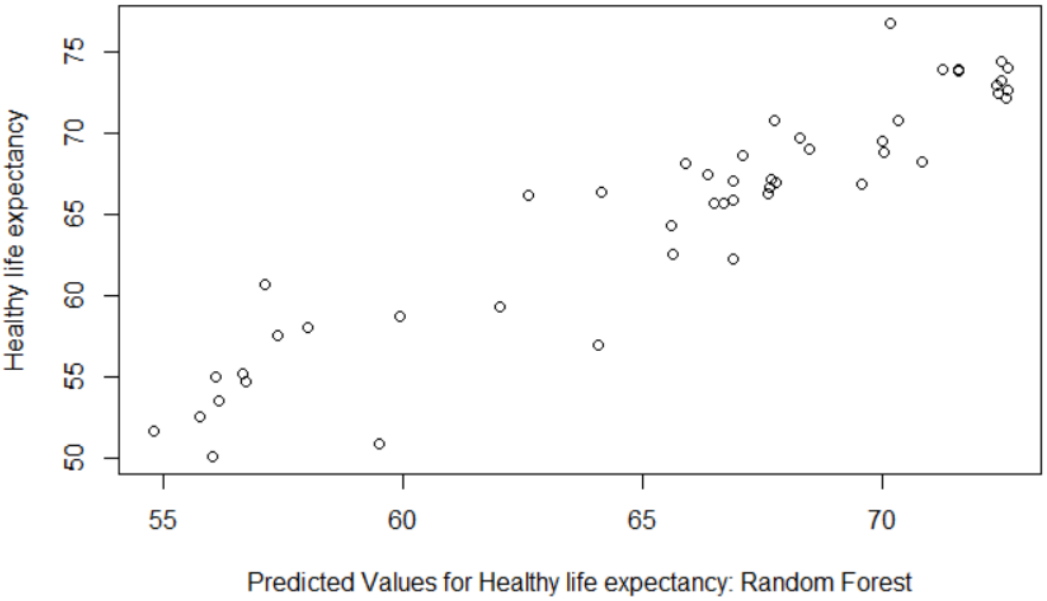
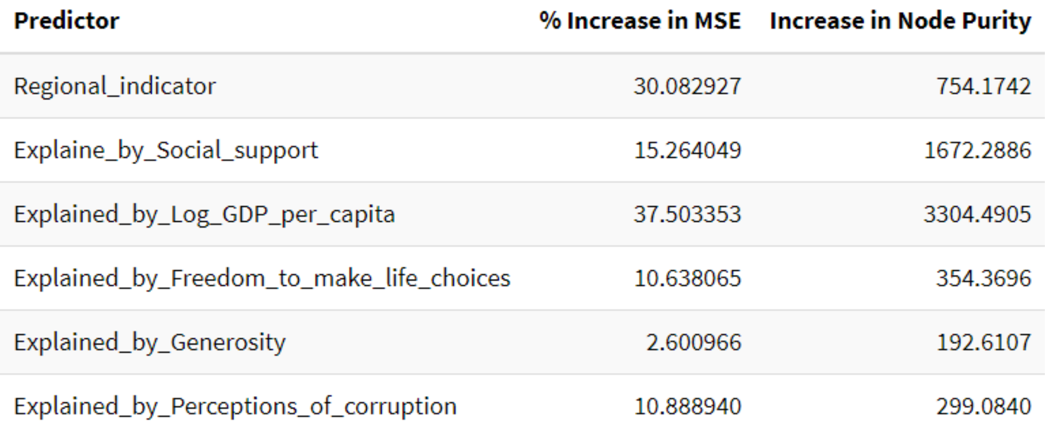


Figure 11: Random Forest

Figure 12. shows the importance of each variable in the random forest prediction model. Through random forest it is possible to quantify the Mean Squared Error increase if we were to assume that the variable we are interested in was omitted. Based on the percentage increase in MSE, we can observe that the most important variable in the model is log GDP per capita. In other words, omitting this variable would result in the highest increase in the root mean squared error. Similarly, looking at the increase in node purity, which measures how homogeneous a node is, we can see, again, that the most important variable in the random forest model is log GDP per capita, as this variable is associated with the highest value in the 'Increase in Node Purity' chart.

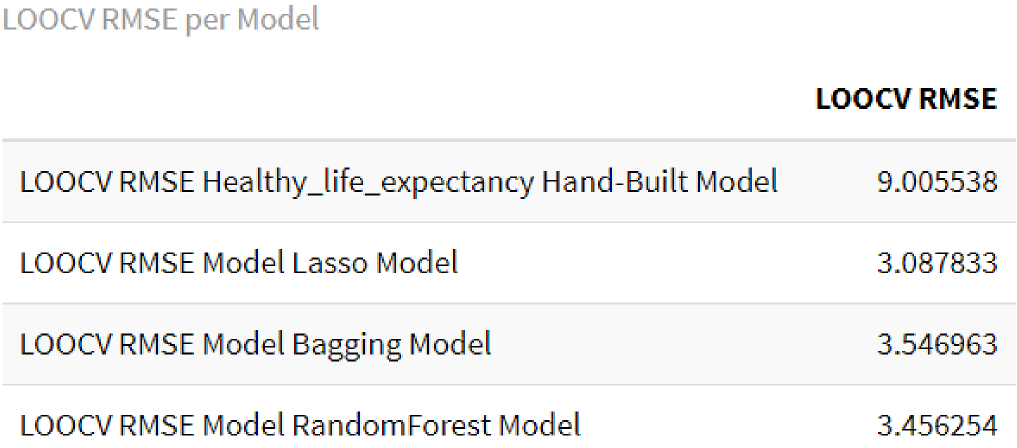
The leave-one out cross-validation RMSE method provides us with the accuracy measures for each one of the individual models considered in this paper. It can be seen from



Predictor	% Increase in MSE	Increase in Node Purity
Regional_indicator	30.082927	754.1742
Explaine_by_Social_support	15.264049	1672.2886
Explained_by_Log_GDP_per_capita	37.503353	3304.4905
Explained_by_Freedom_to_make_life_choices	10.638065	354.3696
Explained_by_Generosity	2.600966	192.6107
Explained_by_Perceptions_of_corruption	10.888940	299.0840

Figure 12: Percentage increase in MSE and node purity

Figure 12. that the Lasso model provides us with the lowest leave-one out cross-validation root mean squared error.



LOOCV RMSE per Model

	LOOCV RMSE
LOOCV RMSE Healthy_life_expectancy Hand-Built Model	9.005538
LOOCV RMSE Model Lasso Model	3.087833
LOOCV RMSE Model Bagging Model	3.546963
LOOCV RMSE Model RandomForest Model	3.456254

Figure 13: RMSE

5 Conclusion

In brief, the Leave-one out cross-validation RMSE Random Forest model performs slightly better than the Leave-one out cross-validation bagging model. However, it does not perform nearly as good as the leave-one out cross-validation lasso model. This can be seen by the lower RMSE in the lasso model compared to the other models considered. The linear hand-built model is the model that performs the worst. However, it is considered as a base model from which comparisons can be drawn.

References

[1] Achim, M., Văidean, V., amp; Borlea, S. *(2019, October 05). Corruption and health outcomes within an economic and cultural framework. Retrieved May 6, 2021, from https://link.springer.com/article/10.1007/s10198-019-01120-8*

[2] Burling, S. *(2021, March 01). Social security and Medicare may experience their OWN Covid-19 side EFFECTS, experts say. Retrieved April 20, 2021, from https://www.inquirer.com/health/coronavirus/how-will-covid-19-affect-social-security-medicare-disability-20210301.html*

[3] J, R. (n.d.). *Family relationships, social support and subjective life expectancy. Retrieved April 18, 2021, from https://pubmed.ncbi.nlm.nih.gov/12664677/*

[4] Nadpara, N., and Samanta, S. *(2015). An empirical examination of the effect of corruption on health outcomes. The College of New Jersey, 1-26.*

[5] WorldHappinessReport. *(n.d.). Retrieved April 26, 2021, from https://worldhappiness.report/*