STAT 5630 Final Project

James Lee, Yifan Li, Tianrui Zhu, Shulei Yang

World Happiness Report

Purpose

Can we create a model that accurately predicts the average happiness of countries around the world? Which predictors are the most important in determining happiness?

Data

Overview

The World Happiness Report (Helliwell, Layard, & Sachs, 2019) is an annual report of 156 countries that tracks how happy the citizens of that country perceive themselves to be. The report was first published in 2012 and continues to gain global recognition. The data used in the report are from the Gallup World Poll, which is a survey that has been tracking important issues worldwide since 2005. The survey and interview methods used in the Gallup World Poll are rigorous and consistent and track issues such as “food access, employment, leadership performance, and well-being.” Careful measures are taken to ensure that the trends are representative and comparable across time and across countries around the world. For detailed information on how the survey is conducted, see Gallop, 2019.

The data in our analysis was used in the World Happiness Report and is sourced from the World Gallup Poll. It contains the survey results for each country each year, starting from 2006. Not all countries have data every year. The quality and availability of data generally increases over time, but there are occasionally gaps in country data when conflicts or other issues prevent the survey from being conducted safely.

Key Variables

Life Ladder

This is the national average response to the question of life evaluation, also known as the Cantril life ladder (score from 0 to 10). “Please Imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?”

GDP Per Capita

GDP per capita in purchasing power parity at constant 2011 international dollar prices are from the November 14, 2018 update of the World Development Indicators.

Healthy Life Expectancy

Data is drawn from the World Health Organization’s Global Health Observatory data repository which contains information for the years 2000, 2005, 2010, 2015, and 2016. Interpolation and extrapolation were used for remaining years from 2005 to 2018. Some territories/countries used different sources for their data.

Social Support

National average of the binary responses (0 or 1) to the survey question: “If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?”

Freedom to Make Life Choices

National average of the binary responses (0 or 1) to the survey question: “Are you satisfied or dissatisfied with your freedom to choose what you do with your life?”

Generosity

Residual of regressing national average of response to the question “Have you donated money to a charity in the past month?” on GDP per capita.

Corruption Perception

National average of the binary responses (0 or 1) to two survey questions about corruption. Larger value indicates higher corruption.

Positive effect

National average response of three positive effect measures: happiness, laughter and enjoyment. Larger value indicates more positive effect.

Negative effect

National average response of three negative affect measures: worries, sadness and anger. Larger value indicates more negative effect.

Transformations

The original data set has yearly observations for each country. However, we were not interested in doing a time series analysis and decided to average all a country’s observations for all its year it had data. First of all, we removed all observations of Cuba, North Cyprus, Oman, Somalia, Somaliland region, Swaziland. These countries/regions, for various reasons, contain too many missing data and are not helpful in analysis. Additionally, we removed the following parameters:

- *Year*
- *Perceptions.of.corruption*
- *Confidence.in.national.government*

Year was removed since we would be averaging data across all the years. Although *Perceptions.of.corruption* and *Confidence.in.national.government* seem important, some

countries we wanted to include in the analysis did not have data on those variables available. Furthermore, we found that these two predictors are highly correlated with other variables in the dataset, and decided to remove the two variables rather than remove more countries. Finally, for all our models, we used *Life.Ladder as* the response variable. The terms 'Life-Ladder' and 'Happiness' will be used interchangeably in this report.

Methods

Linear regression with all the variables

We used our training data set to perform a linear regression model with all the variables with 'Life.Ladder' be the response (results shown below). From the result, we can see that R-squared is 0.8255 which means that 82.55% of the variance found in the response variable (*Life.Ladder*) can be explained by the predictor variables. The model p-value is smaller than 2.2e-16, which suggests that our model is statistically significant. By examining each variable, we find that '*Log.GDP.per.capita*', '*Healthy.life.expectancy.at.birth*', '*Social.support*' and '*Positive.affect*' are particularly significant.

Then we used the testing data to fit the model and predicted the value of 'Life.Ladder'. We calculated an MSPE = 0.1282466 of our model.

```
Call:
lm(formula = Life.Ladder ~ ., data = train_avg[, -1])

Residuals:
     Min       1Q   Median       3Q      Max
-1.48694 -0.26380  0.03846  0.32478  1.06896

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      -2.72119    0.68912  -3.949 0.000135 ***
Log.GDP.per.capita                0.26959    0.08434   3.196 0.001789 **
Social.support                    1.98312    0.63713   3.113 0.002331 **
Healthy.life.expectancy.at.birth  0.02622    0.01159   2.262 0.025575 *
Freedom.to.make.life.choices      0.40988    0.52988   0.774 0.440767
Generosity                        0.21793    0.32420   0.672 0.502773
Positive.affect                   2.60139    0.70237   3.704 0.000326 ***
Negative.affect                   0.92424    0.75342   1.227 0.222388
Democratic.Quality               -0.04938    0.10363  -0.477 0.634582
Delivery.Quality                  0.17806    0.11943   1.491 0.138668
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4757 on 117 degrees of freedom
Multiple R-squared:  0.8255,    Adjusted R-squared:  0.812
F-statistic: 61.48 on 9 and 117 DF,  p-value: < 2.2e-16
```

Ridge regression

We first used the *cv.glment* function and 5 fold cross-validation to find the optimal tuning parameter (0.2094127). Then we performed a ridge regression on the training dataset with *glmnet* (Result shown below) and predicted the value of ‘*Life.Ladder*’ with the testing dataset. Finally, we calculated an MSPE of 0.2122564.

```
> coef(lm_ridge)
10 x 1 sparse Matrix of class "dgCMatrix"
                                            s0
(Intercept)                         -0.47782112
Log.GDP.per.capita                   0.26803427
Social.support                       1.29314871
Healthy.life.expectancy.at.birth     0.01920617
Freedom.to.make.life.choices         .
Generosity                           .
Positive.affect                      1.63405546
Negative.affect                      .
Democratic.Quality                   .
Delivery.Quality                     0.07680004
```

Lasso regression

We used *glmnet* package with a 5 fold cross-validation to perform a Lasso regression on our training dataset (Result shown below). The optimal tuning parameter selected was 0.0114387. Then we used the testing data to fit the model and calculated an MSPE of 0.1254919.

```
> CF_lasso
                                            1
(Intercept)                         -0.85205599
Log.GDP.per.capita                   0.27269053
Social.support                       1.35630593
Healthy.life.expectancy.at.birth     0.02080650
Freedom.to.make.life.choices         0.07767280
Generosity                           0.00000000
Positive.affect                      1.81105662
Negative.affect                      0.00000000
Democratic.Quality                   0.00000000
Delivery.Quality                     0.08487567
```

Akaike information criterion (AIC)

We used the *step* function on the linear model and let direction = 'both' to perform the AIC model selection on training dataset. The result is shown below. We then used the training data to fit the model and calculate the MSPE (0.1303666)

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -2.34101    0.63773  -3.671 0.000361 ***
Log.GDP.per.capita               0.29311    0.08008   3.660 0.000375 ***
Social.support                   1.64015    0.58233   2.817 0.005670 **
Healthy.life.expectancy.at.birth 0.02673    0.01129   2.367 0.019507 *
Positive.affect                  2.89553    0.51506   5.622 1.23e-07 ***
Delivery.Quality                 0.13857    0.07324   1.892 0.060898 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4748 on 121 degrees of freedom
Multiple R-squared:  0.8202,    Adjusted R-squared:  0.8128
F-statistic: 110.4 on 5 and 121 DF,  p-value: < 2.2e-16
```

Bayesian Information Criterion (BIC)

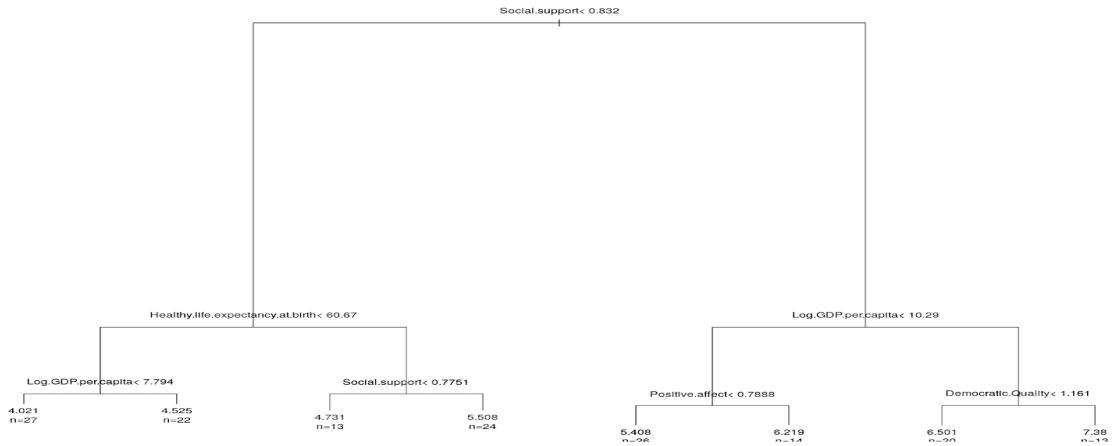
We used the *step* function on the linear model again and let direction = 'both' , k = log(number of rows of training dataset) to perform the BIC model selection. The result is shown below. We then used the training data to fit the model and calculate an MSPE of 0.1392965.

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -3.29453    0.39485  -8.344 1.29e-13 ***
Log.GDP.per.capita               0.35353    0.07420   4.764 5.28e-06 ***
Social.support                   1.56001    0.58689   2.658  0.00891 **
Healthy.life.expectancy.at.birth 0.03174    0.01109   2.861  0.00497 **
Positive.affect                  3.10312    0.50852   6.102 1.28e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4798 on 122 degrees of freedom
Multiple R-squared:  0.8149,    Adjusted R-squared:  0.8088
F-statistic: 134.3 on 4 and 122 DF,  p-value: < 2.2e-16
```

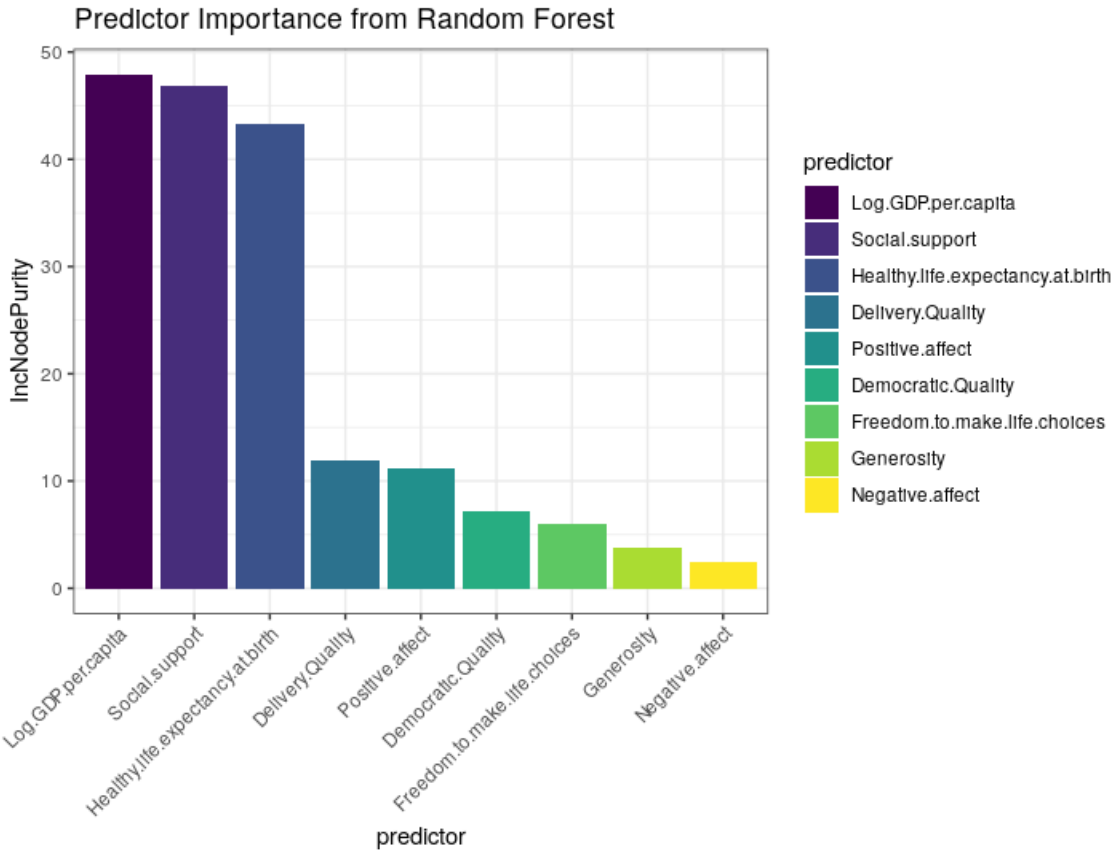
Decision Tree

We created a decision tree using the *rpart* and *caret* package. 10 fold cross-validation was used for training, and the optimal *cp* (complexity parameter) selected was cp = 0.015. The decision tree splits are shown in the graph below.



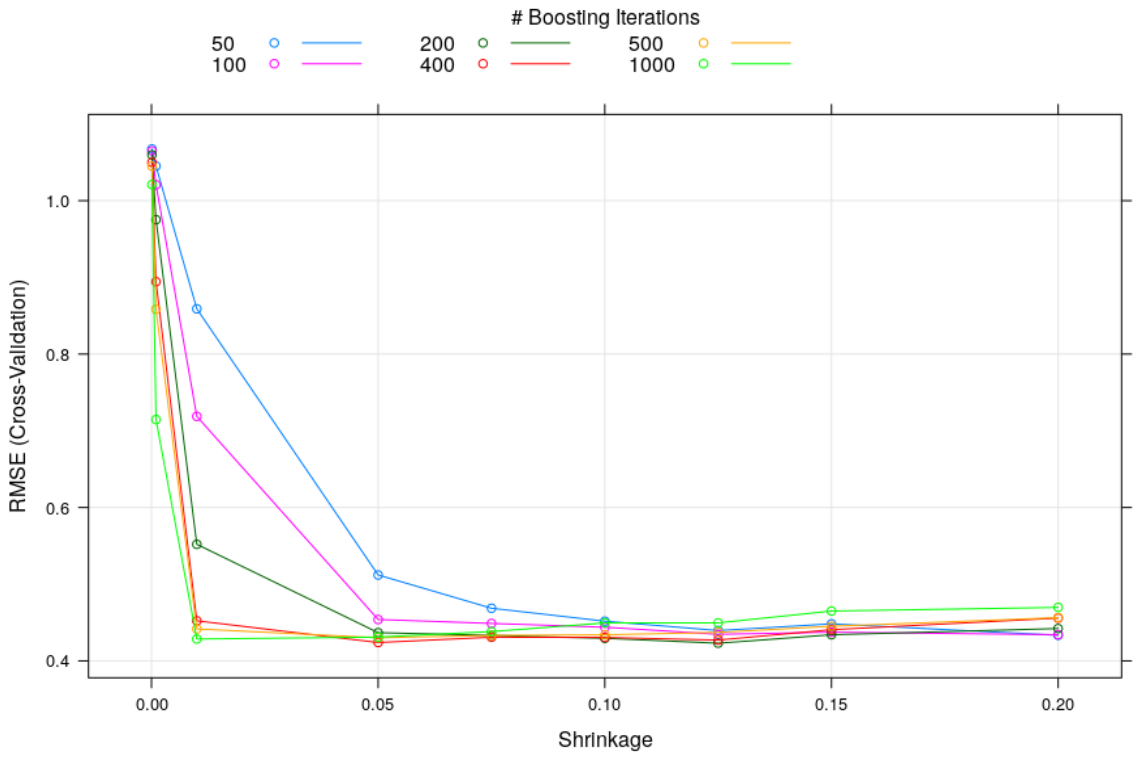
Random Forest

The random forest model was created using the *randomForest* and *caret* packages. We used 10 fold cross-validation for parameter selection, and found that the optimal number of randomly selected parameters, *mtry,* was mtry = 5. Five hundred trees were used in our model. The plot below shows the variable importance from the random forest model.



Gradient Boosting Trees

The Gradient Boosting Trees model was created using the *gbm* and *caret* packages. 10 fold cross-validation was used, and the final model had *interaction.depth* = 1, *n.minobsinnode* = 10, *n.trees* = 400, and *shrinkage* = 0.05. The plot below shows the training process.

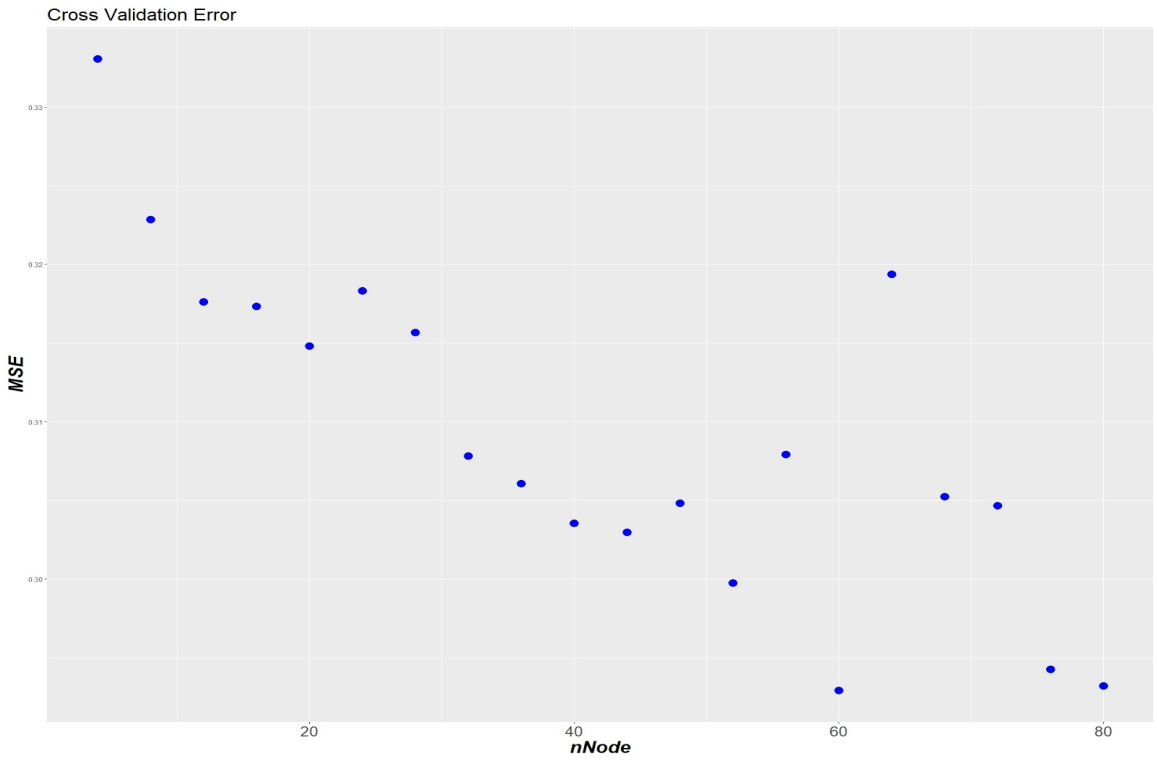


Neural Network

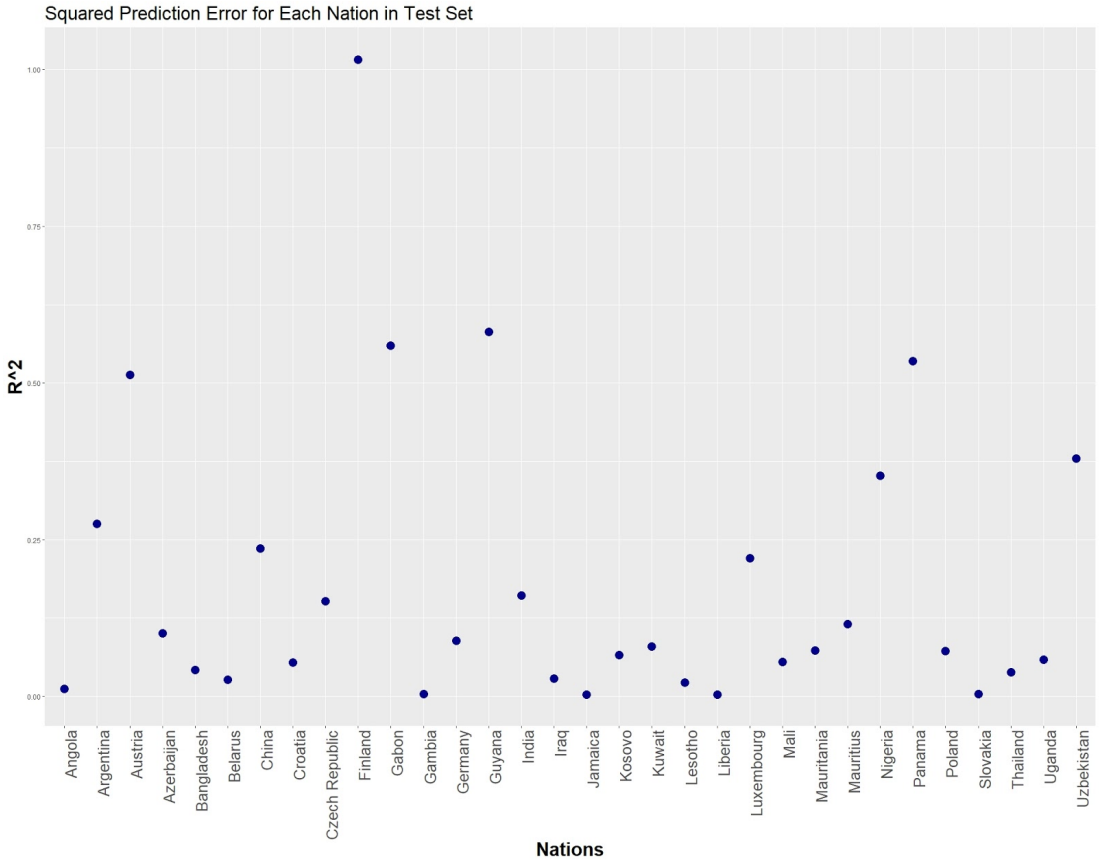
A Neural Network model was created by utilizing the R package *nnet*. Due to the relative small size of our dataset, we cannot expect to fit an elaborate deep Neural Net and hence we did not expect the results to outperform the best of the above-mentioned methods. We are also fully aware of the potential drawbacks of Neural Network which make interpretation of coefficients rather impossible. Nevertheless, we believe this would be an appropriate demonstration of what Neural Network could achieve given additional data or with supplementary information.

Then we discuss our choice of some of the commonly tuned hyperparameters in our instance, including learning rate, number of hidden layers, and number of nodes in each hidden layer. Given the size of our dataset, we restricted ourselves to a Neural Network to only one hidden layer. Since the *nnet* package utilizes a quasi Newton-Raphson algorithm

(namely the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm) rather than a more commonly implemented stochastic descent search, the hyperparameter ‘Learning rate’ does not apply. Note that the BFGS algorithm inherits some of the drawbacks of any quasi Newton-Raphson method such as it is difficult to apply on a large dataset. Fortunately, the nature of our dataset made us fully utilize the fast convergence rate of the BFGS algorithm without dealing with some of its limitations.This leaves the number of nodes in the hidden layer as the only tunable hyperparameter in our application. After utilizing a five-fold cross validation, we arrived at the conclusion that a hidden layer with 60 connected nodes would be optimal, as shown below. Note that since we are dealing with a regression analysis rather than a classification problem, we did not implement any softmax function in the output layer, but rather normalized our output to be between 0 and 1 for prediction before scaling them back to match the actual spread.



Implementing this simple Neural Network with one fully connected hidden layer and 60 nodes on the test set gave us a Mean Square Prediction Error of 0.1852038 with the following spread:



Conclusions

Final Model

After creating all the models, we tested each of them against the test data set. The MSPE for each model is shown in the table below. In the end, the model that performed the best on the test set was the Random Forest model.

Model	Hyperparameters	MSPE
Linear Regression	N/A	0.128

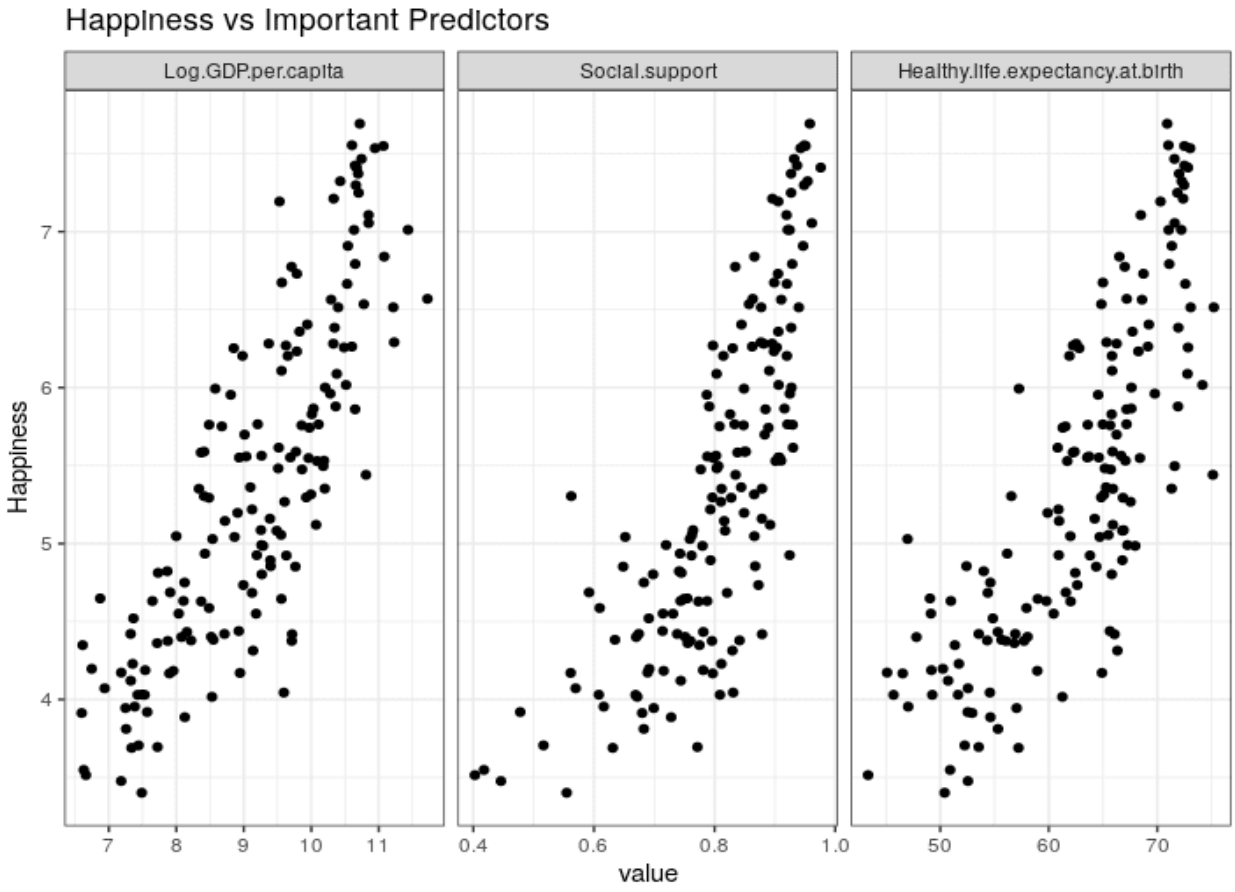
Ridge	λ= 0.209	0.212
Lasso	λ= 0.0114	0.125
Decision Tree	cp = 0.015	0.092
Random Forest	mtry = 5 n.trees = 500	0.018
Gradient Boosting Trees	interaction.depth = 1 n.minobsinnode = 10 n.trees = 400 shrinkage = 0.05	0.044
Neural Network	nNode = 60	0.185

Important Predictors of Happiness

Based on the models we created, there several variables that were consistently considered important in predicting a country's happiness. The most important variables were:

1. Log of GDP per capita
2. Social support
3. Healthy life expectancy at birth

Below is a scatter plot showing how each of the top predictors relate to happiness.



Future Work

For future work, there are a couple of questions that are worth investigating. First of all, it might be interesting to find out whether the correlated variables are related to real world events. Also, since we have longitudinal data, analyzing each country as time series might produce some meaningful results. Some interesting questions are how is the trend of happiness score look like? Why does a particular country has a certain trend and what is the main cause of it? Besides, is there a better way to deal with missing data? At this point we just simply deleted data with missing values. It might be meaningful to impute the missing values using mean or regression if two variables are highly correlated.

Reference

1. Helliwell, J.F., Layard, R., & Sachs, J.D. (2019). *World Happiness Report*. New York: United Nations.

Gallop. (2019). *How Does the Gallup World Poll Work?* Retrieved from https://www.gallup.com/178667/gallup-world-poll-work.aspx