World Happiness Report Analysis

Rohan R. Deshmukh

Harrisburg University of Science & Technology

Table of Contents

Introduction

One of the fundamental research areas in social sciences and psychology involves defining the state of happiness and the factors that affect it; both positively and negatively. The analysis of happiness is a way to assess welfare which combines the techniques usually used by psychologists. It relies on surveys of the reported wellbeing of hundreds of thousands of individuals across countries and continents.

There are pre-conceived notions revolving around human happiness and prevailing ideas that involve many factors such as material wealth, location, weather, personal life, etc. Psychologists have studied the qualitative aspects of happiness from time to time in the past decades and have published relevant analyses which I have tried to cover in my final paper.

This study is directed towards a specific and focused quantitative experimental analysis of the survey conducted by the Sustainable Development Solutions Network. The World Happiness Report is a benchmark survey of the state of global happiness. The first report was published in 2012, followed by the second in 2013, the third in 2015, and the fourth in the 2016 and the latest being in 2019.

Background and Literature Review

World Happiness 2017 report ranks 155 countries by their happiness levels. It was released at the United Nations during an event celebrating International Day of Happiness on March 20th. The report is supported by many national governments, non-profit organizations and civil groups and is utilized this study as happiness indicators to help them take better policy decisions. Experts in the fields of economics, survey data analysis, national statistics, health, public policy, psychology describe the correlation between a nation’s well-being and its progress. The reports review the

state of happiness in the world today and show how the new science of happiness explains personal and national variations in happiness.

The happiness scores and rankings use data from the Gallup World Poll. The poll consists of questions that relate to various domains of human life. The survey enlists answers to these evaluation questions. This question, known as the Cantril ladder, asks respondents to think of a ladder with the best possible life for them being a 10 and the worst possible life being a 0 and to rate their own current lives on that scale. The scores are from nationally representative samples for the years 2013-2016 and use the Gallup weights to make the estimates representative. The columns following the happiness score estimate the extent to which each of six factors – economic production, social support, life expectancy, freedom, absence of corruption, and generosity – contribute to making life evaluations higher in each country than they are in Dystopia, a hypothetical country that has values equal to the world's lowest national averages for each of the six factors. They have no impact on the total score reported for each country, but they do explain why some countries rank higher than others.

Dystopia

One of the variables in the dataset is Dystopia. Dystopia is a hypothetical country that is supposed to have the world's least-happy population. The intentions of having this variable is establishing a benchmark against which all countries can be equally compared. No country can perform more poorly than Dystopia. One way to see this is a minimum scale of values. The lowest scores observed for the six key variables, therefore, characterize Dystopia. As life would be very difficult and unfavorable in a country with the world's lowest incomes, lowest life expectancy, lowest generosity, most corruption, least freedom and least social support, it is referred to as "Dystopia".

Residuals

The residuals, or unexplained components, differ for each country, reflecting the extent to which the six variables either over- or under-explain average 2014-2016 life evaluations. These residuals have an average value of approximately zero over the whole set of countries.

Data Analysis

For this project, I have used Happiness data. The dataset has been collected and analyzed during a research collaboration of United Nations and Gallup World Poll. This dataset has been listed as World Happiness Report in Kaggle.

Data exploration

The dataset contains following columns: GDP per Capita, Family, Life Expectancy, Freedom, Generosity, Government Trust, Happiness Score and Happiness Rank. The higher the happiness score the lower the happiness rank. The Dystopia Residual metric actually is the Dystopia Happiness Score (1.85) + the Residual value or the unexplained value for each country. Following figure summarizes the structure of our dataset.

```
'data.frame':   155 obs. of  12 variables:
 $ Country                       : Factor w/ 155 levels "Afghanistan",..: 105
 $ Happiness.Rank                : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Happiness.Score               : num  7.54 7.52 7.5 7.49 7.47 ...
 $ Whisker.high                  : num  7.59 7.58 7.62 7.56 7.53 ...
 $ Whisker.low                   : num  7.48 7.46 7.39 7.43 7.41 ...
 $ Economy..GDP.per.Capita.      : num  1.62 1.48 1.48 1.56 1.44 ...
 $ Family                        : num  1.53 1.55 1.61 1.52 1.54 ...
 $ Health..Life.Expectancy.      : num  0.797 0.793 0.834 0.858 0.809 ...
 $ Freedom                       : num  0.635 0.626 0.627 0.62 0.618 ...
 $ Generosity                    : num  0.362 0.355 0.476 0.291 0.245 ...
 $ Trust..Government.Corruption.: num  0.316 0.401 0.154 0.367 0.383 ...
 $ Dystopia.Residual             : num  2.28 2.31 2.32 2.28 2.43 ...
```

Figure 1 summarizes the structure of the data along with number of observations and variables.

The “Whisker.high” and “Whisker.low” represent the upper and lower confidence interval of happiness score. We can exclude it from the analysis and exploratory process. Furthermore, I have renamed the columns from the dataset for reader’s readability. The updated structure of the dataset is as follows.

```
'data.frame':   155 obs. of  10 variables:
 $ Country         : Factor w/ 155 levels "Afghanistan",..: 105
 $ HappinessRank   : int  1 2 3 4 5 6 7 8 9 10 ...
 $ HappinessScore  : num  7.54 7.52 7.5 7.49 7.47 ...
 $ Economy         : num  1.62 1.48 1.48 1.56 1.44 ...
 $ Family          : num  1.53 1.55 1.61 1.52 1.54 ...
 $ LifeExpectancy  : num  0.797 0.793 0.834 0.858 0.809 ...
 $ Freedom         : num  0.635 0.626 0.627 0.62 0.618 ...
 $ Generosity      : num  0.362 0.355 0.476 0.291 0.245 ...
 $ Trust           : num  0.316 0.401 0.154 0.367 0.383 ...
 $ DystopiaResidual: num  2.28 2.31 2.32 2.28 2.43 ...
```

Figure 2 summarizes the updated structure of the dataset

As part of the exploratory process, I have analyzed the relationship between the numeric variables within the dataset. For this, I have used corrplot () function. The following figure shows us the correlation plot for the dataset.

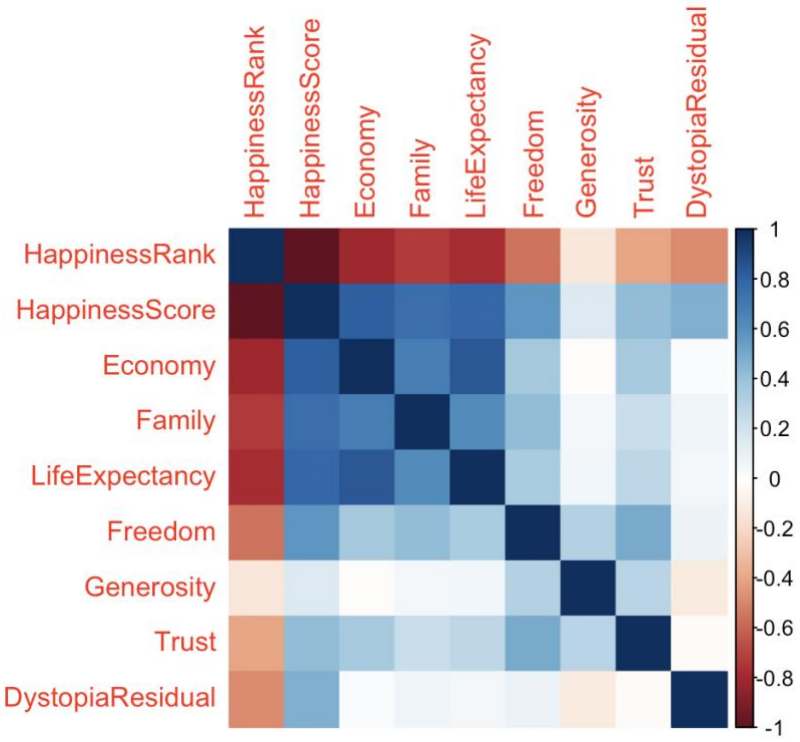

Figure 3 displays the correlation plot for the numeric variables of dataset

As we can see for the above plot, Happiness Rank has an inverse ratio with all the other variables. This was expected as Happiness Rank is lower for countries with higher happiness score. Also, it depicts that Happiness Score has strong positive correlation with Economy, Family and LifeExpentency, Freedom, Trust and DystopiaResidual. Generosity variable doesn't seem to have much effect on the happiness score of the country.

Also, the top ten and least ten happiest countries are as below.

	Country <fctr>	HappinessRank <int>	HappinessScore <dbl>
1	Norway	1	7.537
2	Denmark	2	7.522
3	Iceland	3	7.504
4	Switzerland	4	7.494
5	Finland	5	7.469
6	Netherlands	6	7.377
7	Canada	7	7.316
8	New Zealand	8	7.314
9	Sweden	9	7.284
10	Australia	10	7.284

Figure 4 displays the top ten happiest countries

	Country <fctr>	HappinessRank <int>	HappinessScore <dbl>
146	Yemen	146	3.593
147	South Sudan	147	3.591
148	Liberia	148	3.533
149	Guinea	149	3.507
150	Togo	150	3.495
151	Rwanda	151	3.471
152	Syria	152	3.462
153	Tanzania	153	3.349
154	Burundi	154	2.905
155	Central African Republic	155	2.693

Figure 5 displays the least ten happiest countries

Model Formulation and Testing

As part of this study, I will be using various machine learning algorithms to analyze the happiness survey data and create a predictor to accurately predict happiness score. For this purpose, we have split our dataset into training and test set. In this scenario, First, we should split our dataset into training and test set. Our dependent variable is happiness score, and the independent

variables are family, economy, life expectancy, trust, freedom, generosity, and dystopia residual. based on the seven independent variables. To be specific, I will be using Multiple Linear regression, Support Vector Machines, Decision tree, Random Forest and Neural Net predictor.

Multiple Linear Regression

As we have multiple variables that affect the happiness score of a country, it is interesting to see whether linear regression model is a good fit for our dataset. I have used the lm () method from R to create a linear regression model. This is followed by testing the model using the test split set we created earlier. To verify the effectiveness of the dataset, I have plotted the actual happiness score and the score predicted by the linear regression model. Following figure shows us the plot and helps us understand the model in a better way.

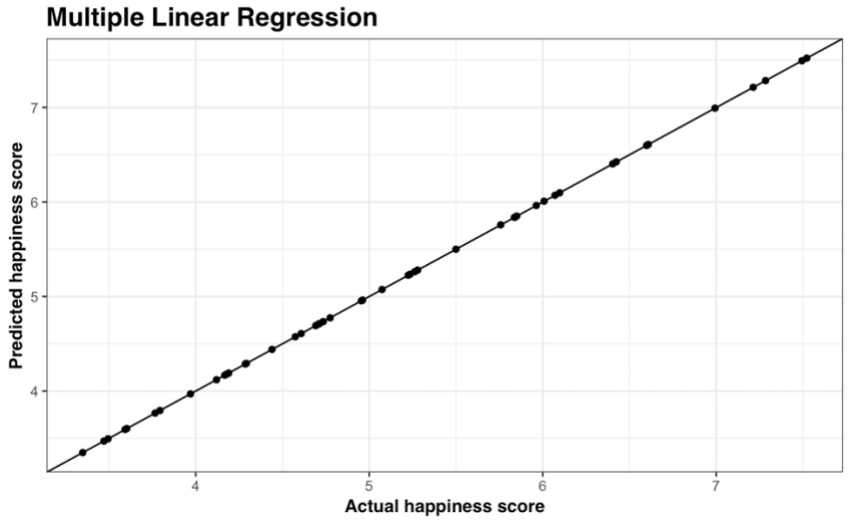

Figure 6 plots actual happiness score vs. predicted score for multiple linear regression model

The following figure shows the summary of the linear regression model.

```
Call:
lm(formula = HappinessScore ~ ., data = hapdat2)

Residuals:
       Min         1Q     Median         3Q        Max
-5.516e-04 -2.363e-04 -1.223e-05  2.559e-04  4.735e-04

Coefficients:
                  Estimate Std. Error   t value Pr(>|t|)
(Intercept)      1.581e-04  1.350e-04     1.171    0.243
Economy          1.000e+00  1.175e-04  8513.535   <2e-16 ***
Family           9.999e-01  1.158e-04  8637.656   <2e-16 ***
LifeExpectancy   9.999e-01  1.845e-04  5419.875   <2e-16 ***
Freedom          1.000e+00  1.973e-04  5069.247   <2e-16 ***
Generosity       1.000e+00  1.907e-04  5245.010   <2e-16 ***
Trust            9.998e-01  2.773e-04  3605.630   <2e-16 ***
DystopiaResidual 1.000e+00  4.704e-05 21259.291   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.000286 on 147 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 3.442e+08 on 7 and 147 DF,  p-value: < 2.2e-16
```

Figure 7 shows summary for the linear regression model

From the above summary, it is evident that all the variables have a significant effect on the happiness score ($p<0.05$). In addition, both Multiple R-squared and Adjusted R-squared is equal to 1. Therefore, it is evident that there is a linear correlation between dependent and independent variables. As per the background of the dataset, the happiness score is a derivative of all the variables in the dataset. This is the justification for having an adjusted R-squared equals to 1. As a result, Multiple Linear Regression predicts happiness scores with 100 % accuracy.

Support Vector Regression

For the purpose of having a dimension-based model in our analysis, I trained a Support Vector regressor on our training data. The svm () function from e1071 package is utilized in this method. I have used “radial” kernel and “eps-regression” in this case. The following figure outputs the actual score vs. predicted score comparison.

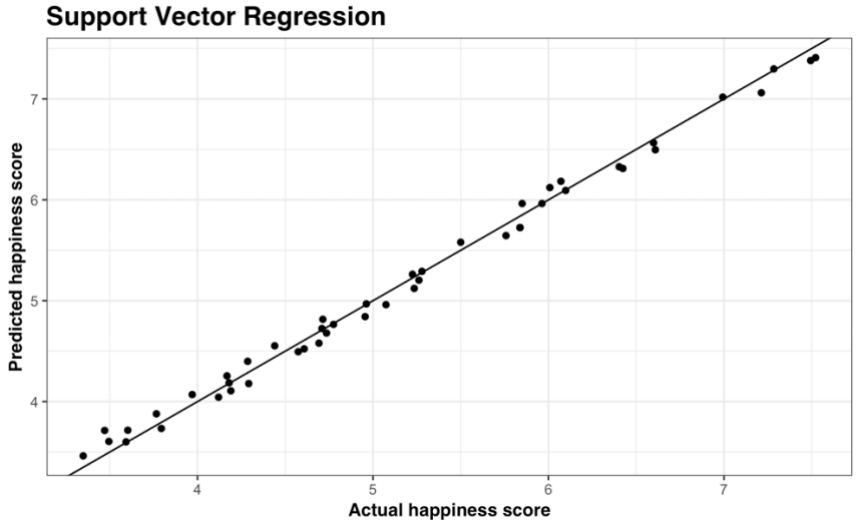

Figure 8 plots actual happiness score vs. predicted score for SVR model

As we can see from the plot, the points line up with the regression line quite nicely. This is an evidence of the accuracy of the model.

Decision Tree

We have created decision tree to predict the happiness score. I have used 10 minimum instances per node and "anova" method for creating the decision tree. The following figure displays the output of the decision tree.

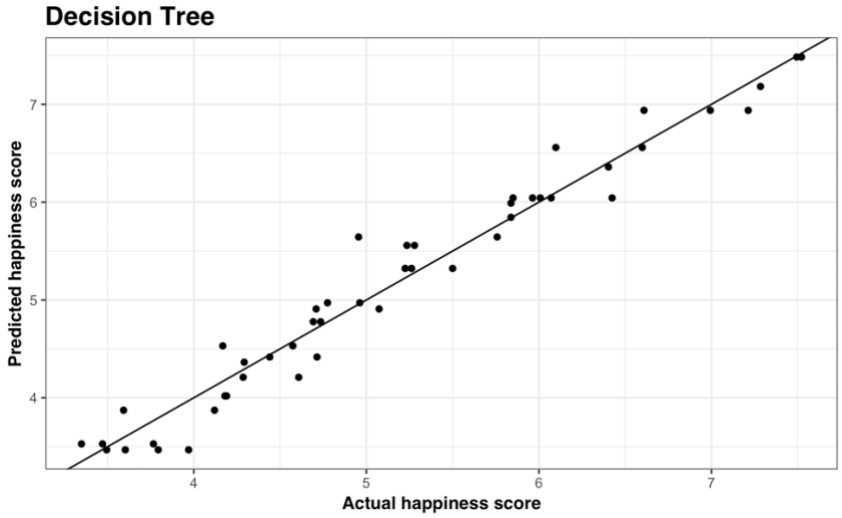

Figure 9 plots actual happiness score vs. predicted score for decision tree model

From the above plot, we can conclude that the decision tree model is not best option for fitting and predicting our data. Most of the data points don't fall on the regression line. Even, the mean squared error is quite high in this case.

Random Forest

We have used random forest algorithm to predict the happiness score. I have used 10 minimum instances per node and "anova" method for creating the decision tree. The following figure displays the output of the decision tree.

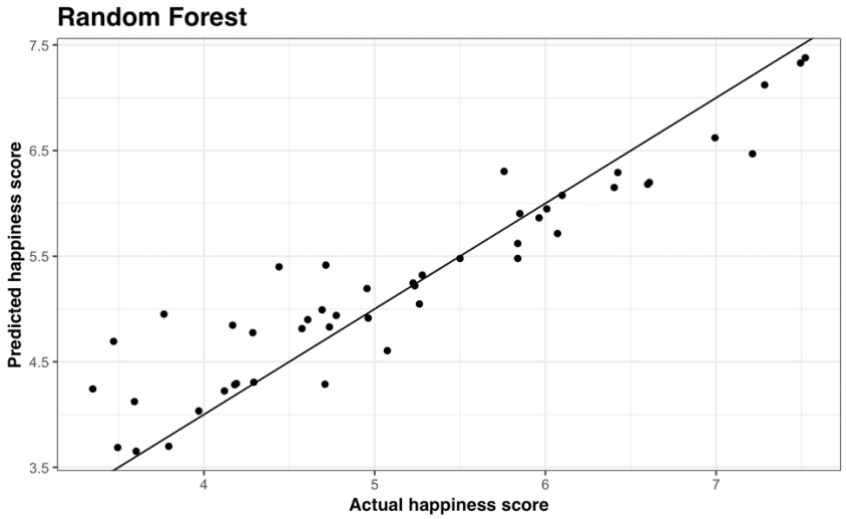

Figure 10 plots actual happiness score vs. predicted score for random forest model

From the above, we can conclude that Random Forest regression is not as good as SVR and decision tree regarding predicted happiness scores.

Neural Net

I have experimented with neural networks as well. The neuralnet R package is a good option in this scenario. I have used 10 hidden neurons along with linear output enabled. The following figure displays the resulting neural network built as part of the modelling process.

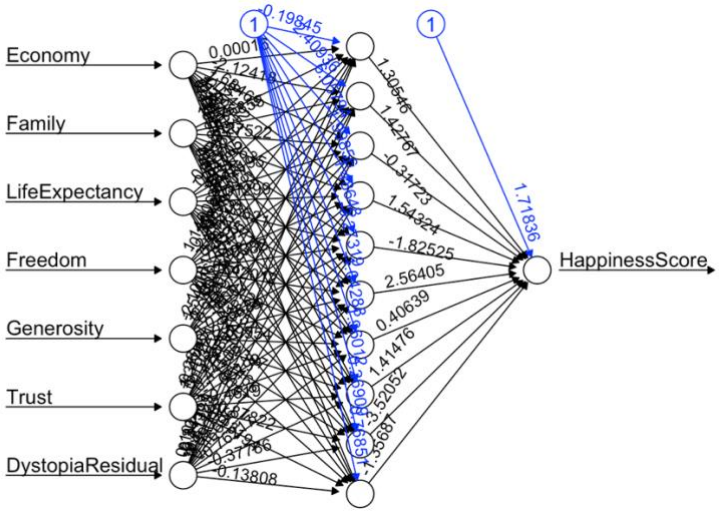

Figure 11 shows the neural network for the neural network model

Also, the following figure displays the output of the neural network model.

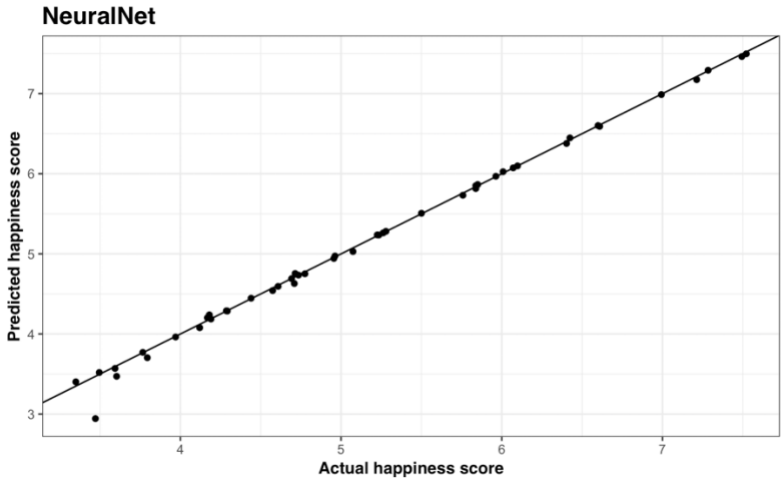

Figure 12 plots actual happiness score vs. predicted score for neural network model

From the above plot, we can see that neural net model really fits the data well and the prediction has also great accuracy. It is the next best model after multiple linear regression.

Conclusion

Multiple Linear Regression and neural net did the best job and predicted approximately the same. SVR and Random Forest stood in the second place regarding accuracy in prediction. And finally, Decision Tree was the worst algorithm to predict happiness scores. The goals of this study are to analyze the survey data compiled in the World Happiness Report published in 2017. One way of analyzing the data would involve finding the correlation between various factors in the data and find out which factors are more important to live a happier life. The conclusion will help people and countries focus on more crucial aspects of day-to-day life and ignore the ones that don't contribute towards living a happy life. On a personal level, I was interested in understanding the results of this survey and comprehend the essentials of living a stress-free life with more contentment and happiness. Following are some of the conclusion from our above study

1. The predictions seem to be reliable. The metadata mentions that the happiness score is a linear derivative of the other variables in the dataset. Hence, it is expected to see multiple regression perform the best among all the models.
2. Most of the variables in the dataset were positively correlated with the Happiness Score.
3. Neural network-based model is better than all of the models except for the multiple linear regression.
4. The Nordic region seems to be the happiest region in the world followed by the Australia continent. While, on the other hand, African continent is the least happiest region in the world.

References

[1] World Happiness Report, Sustainable Development Solutions Network. (2017). Retrieved from https://www.kaggle.com/unsdsn/world-happiness

[2] Helliwell, J., Layard, R., & Sachs, J. (2017). World Happiness Report 2017, New York: Sustainable Development Solutions Network.

[3] Akerlof, G., & Shiller, R. (2010). Animal spirits: How human psychology drives the economy, and why it matters for global capitalism. Princeton: Princeton University Press.

[4] Altindag, D. T., & Xu, J. (2017). Life satisfaction and preferences over economic growth and institutional quality. Journal of Labor Research.

[5] Michy A. (2015). Fitting a neural network in R; neuralnet package. Retrieved from https://www.r-bloggers.com/fitting-a-neural-network-in-r-neuralnet-package/