# Report – Prediction of happiness

Lucie Schaynová

2021-10-31

## Contents

# 1 Introduction

Machine learning is about creating and using models. Our goal is to use existing data to develop models that we can use to predict various outcomes for new data. Depending on the type of data, prediction can be accomplished through classification models, random forest, k-nearest neighbors or many other machine learning algorithms.

Our data set used in this project contains data that can be found here: https://www.kaggle.com/ajaypalsinghlo/world-happiness-report-2021?select=world-happiness-report-2021.csv

```
## 'data.frame':    149 obs. of  9 variables:
##  $ Country.name                : chr  "Finland" "Denmark" "Switzerland" "Iceland" ...
##  $ Regional.indicator          : chr  "Western Europe" "Western Europe" "Western Europe" "Western Eu
##  $ Ladder.score                : num  7.84 7.62 7.57 7.55 7.46 ...
##  $ Logged.GDP.per.capita       : num  10.8 10.9 11.1 10.9 10.9 ...
##  $ Social.support              : num  0.954 0.954 0.942 0.983 0.942 0.954 0.934 0.908 0.948 0.934 ..
##  $ Healthy.life.expectancy     : num  72 72.7 74.4 73 72.4 73.3 72.7 72.6 73.4 73.3 ...
##  $ Freedom.to.make.life.choices: num  0.949 0.946 0.919 0.955 0.913 0.96 0.945 0.907 0.929 0.908 ...
##  $ Generosity                  : num  -0.098 0.03 0.025 0.16 0.175 0.093 0.086 -0.034 0.134 0.042 ..
##  $ Perceptions.of.corruption   : num  0.186 0.179 0.292 0.673 0.338 0.27 0.237 0.386 0.242 0.481 ...
```

Our data set contains 149 observations (rows) and 9 variables (columns).

`Country.name` or `Regional.indicator` mean country or region, respectively, of respondents. `Ladder.score` is our predicted variable and means happiness score or subjective well-being. The top of the ladder (number 10) represents the best possible life and the bottom of the ladder (number 0) represents the worst possible life. `Logged.GDP.per.capita` are statistics of GDP per capita in purchasing power parity at constant international dollar prices. `Social.support` is the national average of the binary responses (either 0 or 1). The question for respondents was: "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?" Data in the column `Healthy.life.expectancy` were extracted from the World Health Organization's Global Health Observatory data repository. Each row of `Freedom.to.make.life.choices` means the national average of responses to the question: "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?" `Generosity` is the residual of regressing national average of response to the question: "Have you donated money to a charity in the past month?" The `Perceptions.of.corruption` is the average of the survey responses to two questions: "Is corruption widespread throughout the government or not" and "Is corruption widespread withing businesses or not?"

# 2 Data analysis

```
##  Country.name       Regional.indicator  Ladder.score   Logged.GDP.per.capita
##  Length:117         Length:117         Min.   :2.523   Min.   : 6.635
##  Class :character   Class :character   1st Qu.:4.852   1st Qu.: 8.551
##  Mode  :character   Mode  :character   Median :5.534   Median : 9.585
##                                        Mean   :5.513   Mean   : 9.456
##                                        3rd Qu.:6.255   3rd Qu.:10.382
##                                        Max.   :7.842   Max.   :11.647
##  Social.support   Healthy.life.expectancy Freedom.to.make.life.choices
##  Min.   :0.4630   Min.   :48.48           Min.   :0.3820
##  1st Qu.:0.7500   1st Qu.:59.96           1st Qu.:0.7190
##  Median :0.8320   Median :66.60           Median :0.8000
##  Mean   :0.8138   Mean   :64.97           Mean   :0.7913
##  3rd Qu.:0.9050   3rd Qu.:69.50           3rd Qu.:0.8760
##  Max.   :0.9830   Max.   :76.95           Max.   :0.9700
##    Generosity       Perceptions.of.corruption
##  Min.   :-0.28800   Min.   :0.082
```

```
##  1st Qu.:-0.13900   1st Qu.:0.682
##  Median :-0.04100   Median :0.789
##  Mean   :-0.02677   Mean   :0.735
##  3rd Qu.: 0.06700   3rd Qu.:0.847
##  Max.   : 0.50900   Max.   :0.939
```

We can look at average `Ladder.score` per `Regional.indicator`:

```
## # A tibble: 10 x 2
##    Regional.indicator                  Average.score
##    <chr>                                       <dbl>
##  1 South Asia                                   4.44
##  2 Sub-Saharan Africa                           4.47
##  3 Middle East and North Africa                 5.26
##  4 Southeast Asia                               5.42
##  5 Commonwealth of Independent States           5.44
##  6 East Asia                                    5.77
##  7 Latin America and Caribbean                  5.91
##  8 Central and Eastern Europe                   5.95
##  9 Western Europe                               6.83
## 10 North America and ANZ                        7.11
```

We can see that the highest average happiness is in North America and ANZ, the lowest is in South Africa.

```
##     Country.name Ladder.score
## 115     Botswana        3.467
## 116     Zimbabwe        3.145
## 117  Afghanistan        2.523
```

```
##   Country.name Ladder.score
## 1      Finland        7.842
## 2      Denmark        7.620
## 3      Iceland        7.554
```

Particularly, the highest happiness is in Finland, the lowest in Afghanistan.

Determining the correlation of each factor to each other is our utmost interest now. According to the correlation, we can discuss which of the factors may have influence on our forecast. We can prefer the variables with higher correlation and include them into our models. We will not take into account variables with low correlation. As `Country.name` and `Regional.indicator` are not numeric, we need to exclude them from correlation. At the moment, all the remaining variables are potential predictors of `Ladder.score`.

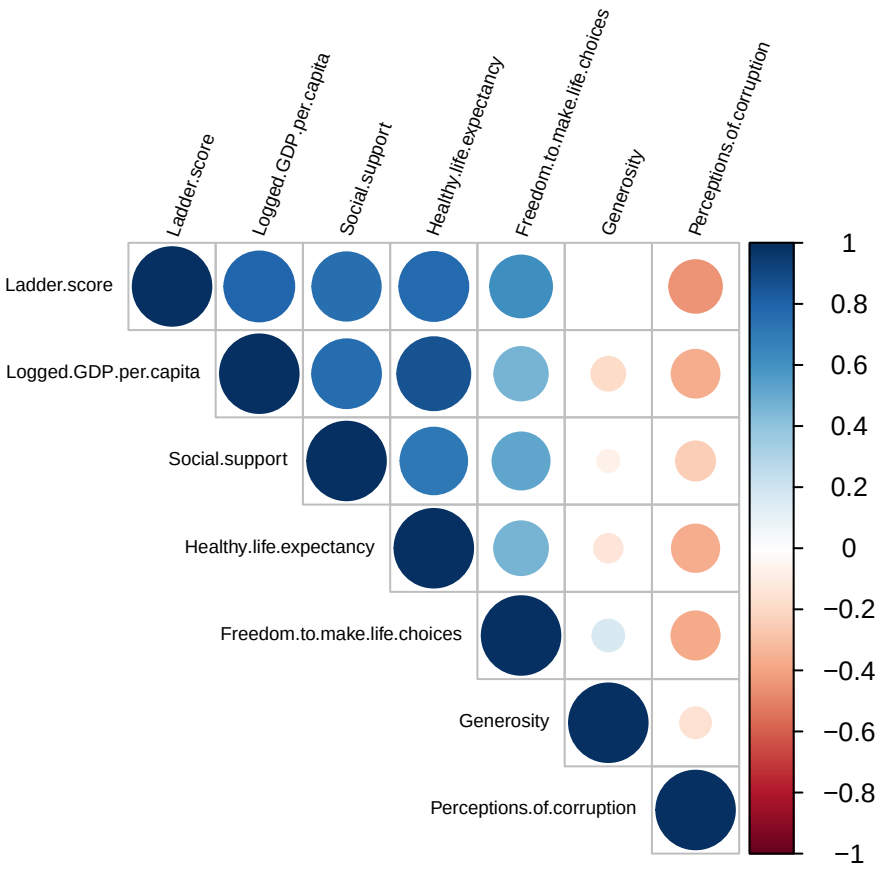



Figure 1: Correlation between variables

We can see that all factors are at least somewhat correlated. The strongest positive correlation with `Ladder.score` has `Logged.GDP.per.capita`, `Social.support`, `Healthy.life.expectancy`, and `Freedom.to.make.life.choices`. Negative correlation is with `Perceptions.of.corruption`.

There is also positive correlation between `Logged.GDP.per.capita` and `Healthy.life.expectancy`.

The relationships between specific variables can be visualized by plotting them and determining the line of the best fit.

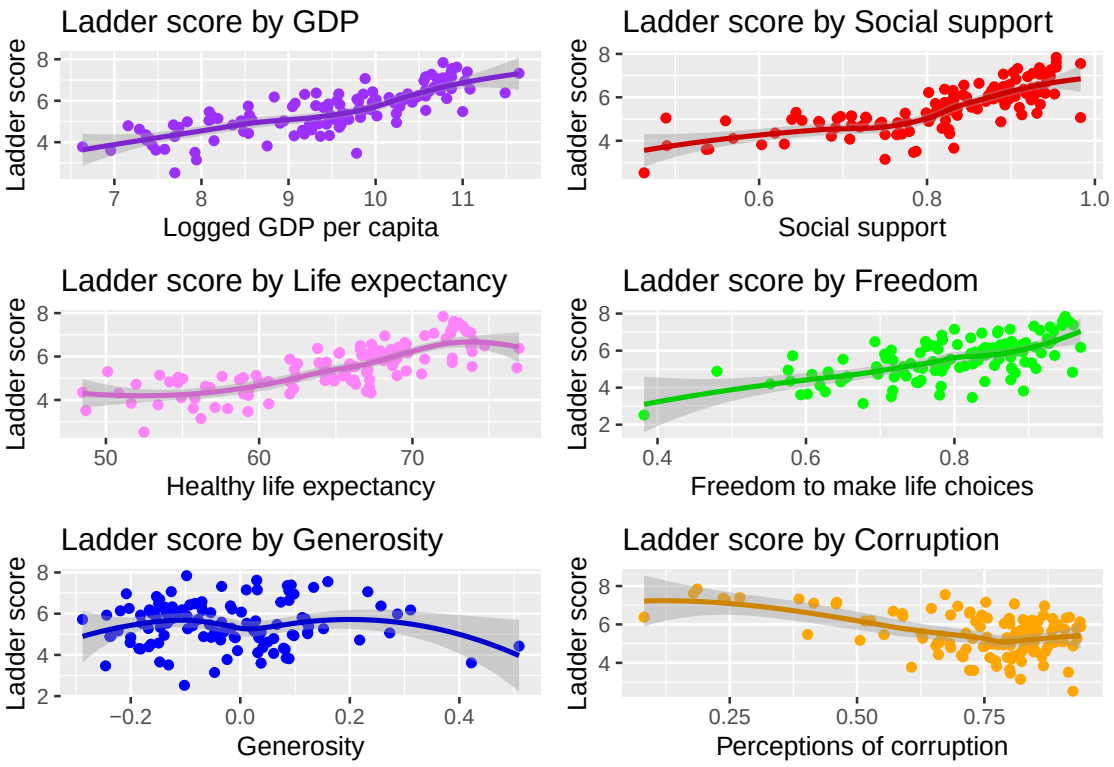



Figure 2: Ladder.score relation with the rest of variables

The plot of Ladder score by GDP clearly shows a positive correlation: when GDP increases, Ladder score also increases. Negative correlation can be seen between Corruption and Ladder score.

# 3 Machine learning methods

In this section, we will use some machine learning methods and focus on their performance on our data set. Based on the performance, we will choose the best method for prediction.

There are many methods we can use, hovewer, we will use Multivariate regression (**MVR**), k-nearest neighbors (**KNN**), Neural networks (**NN**), and Generalized linear model (**GLM**). The question is: which one is the best to provide the most accurate predictions? We will allow cross-validation across all the models and observe their average performance. Our steps will be as follows:

- Divide original data to training (80 %) and testing (20 %) data set.
- Use training data to train all models 5 times using cross-validation.
- Calculate average RMSE for each method.
- Use the method with the lowest RMSE on original testing data, validate and calculate RMSE.

## 3.1 Multivariate Regression (MVR)

We will try to find all of the features that have influence on happiness (`Ladder.score`). Let us add features one by one and observe how RMSE, Multiple R-squared values (or Adjusted R-squared) and p-values change. We start with `Logged.GDP.per.capita` feature.

```
## average RMSE:
```

```
## [1] 0.7716585
```

```
## Coefficients of the best model:
```

```
##                          Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)           -1.6395606 0.55795020 -2.938543 4.168096e-03
## Logged.GDP.per.capita  0.7507675 0.05933513 12.653001 7.531044e-22
```

```
## Multiple R-squared:
```

```
## [1] 0.6350632
```

Now we will add `Social.support` feature:

```
## average RMSE:
```

```
## [1] 0.7249397
```

```
## Coefficients of the best model:
```

```
##                          Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)           -1.8707219 0.53117232 -3.521874 6.720608e-04
## Logged.GDP.per.capita  0.5148138 0.08794406  5.853878 7.511732e-08
## Social.support         3.0522617 0.87670191  3.481527 7.678855e-04
```

```
## Multiple R-squared:
```

```
## [1] 0.6779585
```

We can see that RMSE decreased and multiple R-squared slightly increased. p-values are below 0.05 which is statistically significant. Let us add `Healthy.life.expectancy` feature:

```
## average RMSE:
```

```
## [1] 0.7300257
```

```
## Coefficients of the best model:
```

```
##                            Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)             -2.53560588 0.60055352 -4.222115 5.774241e-05
## Logged.GDP.per.capita    0.32901512 0.12022522  2.736657 7.478166e-03
## Social.support           2.71390715 0.87196499  3.112404 2.487423e-03
## Healthy.life.expectancy  0.04151594 0.01874571  2.214690 2.930643e-02
```

```
## Multiple R-squared:
```

```
## [1] 0.6946022
```

R-squared slightly increased and also RMSE. p-values are still statistically significant. Now we add `Freedom.to.make.life.choices` feature:

```
## average RMSE:
```

```
## [1] 0.6470034
```

```
## Coefficients of the best model:
```

```
##                                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)                 -3.18053266 0.58705958 -5.417734 5.092294e-07
## Logged.GDP.per.capita        0.31499810 0.11242205  2.801924 6.232672e-03
## Social.support               1.91169417 0.84257053  2.268883 2.569313e-02
## Healthy.life.expectancy      0.03481654 0.01761031  1.977054 5.113063e-02
## Freedom.to.make.life.choices 2.36840523 0.63204942  3.747184 3.172245e-04
```

```
## Multiple R-squared:
```

```
## [1] 0.7362185
```

Add `Generosity` feature:

```
## average RMSE:
```

```
## [1] 0.6507993
```

```
## Coefficients of the best model:
```

```
##                                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)                  -3.16948851  0.5866425 -5.402760 5.531753e-07
## Logged.GDP.per.capita         0.33225339  0.1134676  2.928177 4.339947e-03
## Social.support                1.89366386  0.8420099  2.248980 2.701010e-02
## Healthy.life.expectancy       0.03481045  0.0175951  1.978417 5.100931e-02
## Freedom.to.make.life.choices  2.17680372  0.6562082  3.317245 1.322826e-03
## Generosity                    0.50682391  0.4718007  1.074233 2.856550e-01
```

```
## Multiple R-squared:
```

```
## [1] 0.7396328
```

`Perceptions.of.corruption` feature:

```
## average RMSE:
```

```
## [1] 0.6724573
```

```
## Coefficients of the best model:
```

```
##                                 Estimate Std. Error    t value    Pr(>|t|)
## (Intercept)                  -1.40394538 0.85089204 -1.6499689 0.102555333
## Logged.GDP.per.capita         0.27047212 0.11161381  2.4232853 0.017453076
## Social.support                2.26638418 0.82265489  2.7549635 0.007147884
## Healthy.life.expectancy       0.02945948 0.01706941  1.7258644 0.087921830
## Freedom.to.make.life.choices  1.72554967 0.65307429  2.6421951 0.009766489
## Generosity                    0.33456924 0.45898974  0.7289253 0.468005090
## Perceptions.of.corruption    -1.10606376 0.39831666 -2.7768453 0.006720988
```

```
## Multiple R-squared:
```

```
## [1] 0.7608306
```

`Perceptions.of.corruption` feature increased Multiple R-squared and RMSE. Now we can observe that `Healthy.life.expectancy` and `Generosity` are not statistically significant (are greater than 0.05).

Finally, the results below tell us that the happiness is tied more to the combined feature set of GDP, Social support, and Freedom to make life choices than to the Healthy life expectancy, Generosity and Corruption. We removed the Corruption feature because this gives us higher RMSE. We can look at results without the insignificant features:

```
## average RMSE:
```

```
## [1] 0.6376093
```

```
## Coefficients of the best model:
```

```
##                                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)                  -2.6632328 0.53395049 -4.987790 2.953826e-06
## Logged.GDP.per.capita         0.4684578 0.08262897  5.669413 1.710407e-07
## Social.support                2.1495538 0.84730577  2.536928 1.290523e-02
## Freedom.to.make.life.choices  2.4952678 0.63886390  3.905789 1.813599e-04
```

```
## Multiple R-squared:
```

```
## [1] 0.7246337
```

The corresponding results have the lowest RMSE from all the observations. So our final linear equation can look like this:

$$\text{Ladder.score} = -2.66 + 0.47\ \text{Logged.GDP.per.capita} + 2.15\ \text{Social.support} + 2.50\ \text{Freedom.to.make.life.choices}$$

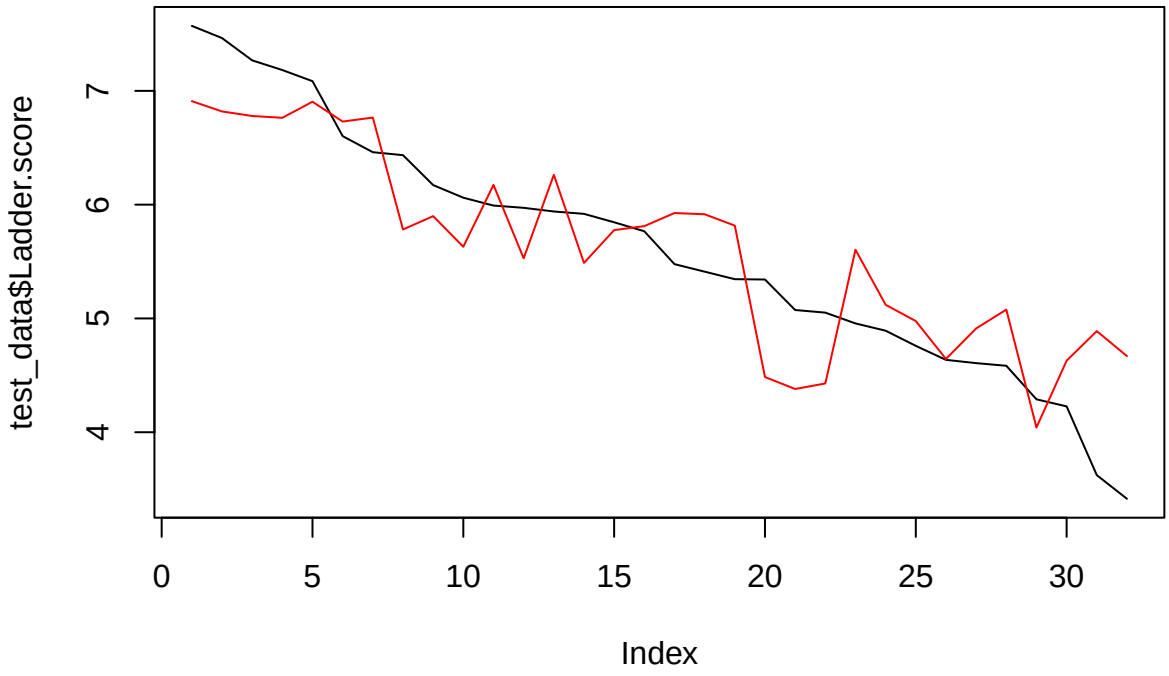
We tried to increase degree of polynomial in our model for each feature, but the results did not improve at all.

When we think about the feature `Healthy.life.expectancy`, one would expect that this feature will affect the happiness. But what we saw in computations is that it does not have any effect. The data were collected from 117 countries so respondents were from developing or industrially advanced countries, men or women, etc. This might have established some noise in data.

Now we use original testing data to test our best linear model, calculate RMSE, and visualize the results in plot of actual (black) and predicted (red) values.

```
## RMSE:
```

```
## [1] 0.5349595
```




The RMSE tells us that our predicted values are 0.53 units far from observed (real) values on average.

## 3.2 k-Nearest Neighbors (KNN)

Now we will use k-Nearest neighbors method. This method takes all available cases in our data into account and provides prediction based on distance measure. It takes a baseline of data and measure the distance between all the points. Then it compares other data with it.

We will run our model for 5 times using cross-validation.

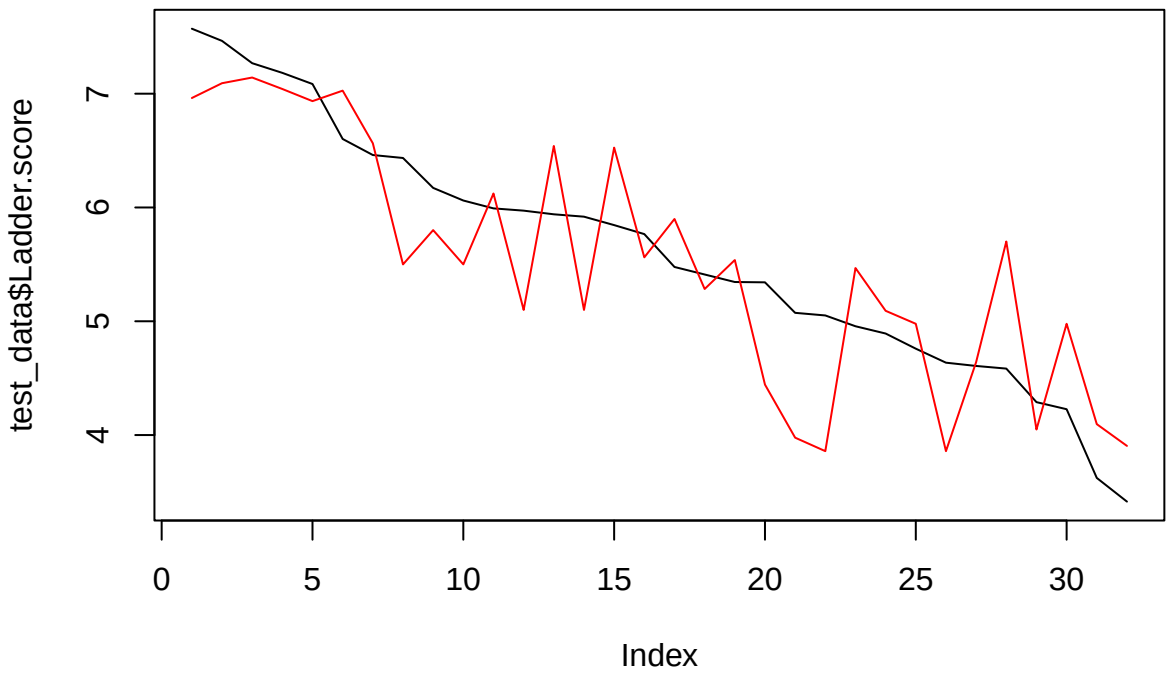
```
## average RMSE:
```

```
## [1] 0.8626899
```

Now we can use original testing data to test our best linear model, calculate RMSE, and visualize the results in plot of actual and predicted values.

```
## RMSE:
```

```
## [1] 0.5957027
```




### 3.3 Neural Networks (NN)

A neural network is basically a set of equations. We use the equations to calculate an outcome.

```
## average RMSE:
```

```
## [1] 0.6850706
```

```
## a 3-5-1 network with 26 weights
## options were - linear output units  decay=0.1
##  b->h1 i1->h1 i2->h1 i3->h1
##  -0.48   0.02   0.25   1.01
##  b->h2 i1->h2 i2->h2 i3->h2
##  -0.48   0.02   0.25   1.01
##  b->h3 i1->h3 i2->h3 i3->h3
##  -6.42   0.36   1.88   0.97
##  b->h4 i1->h4 i2->h4 i3->h4
```

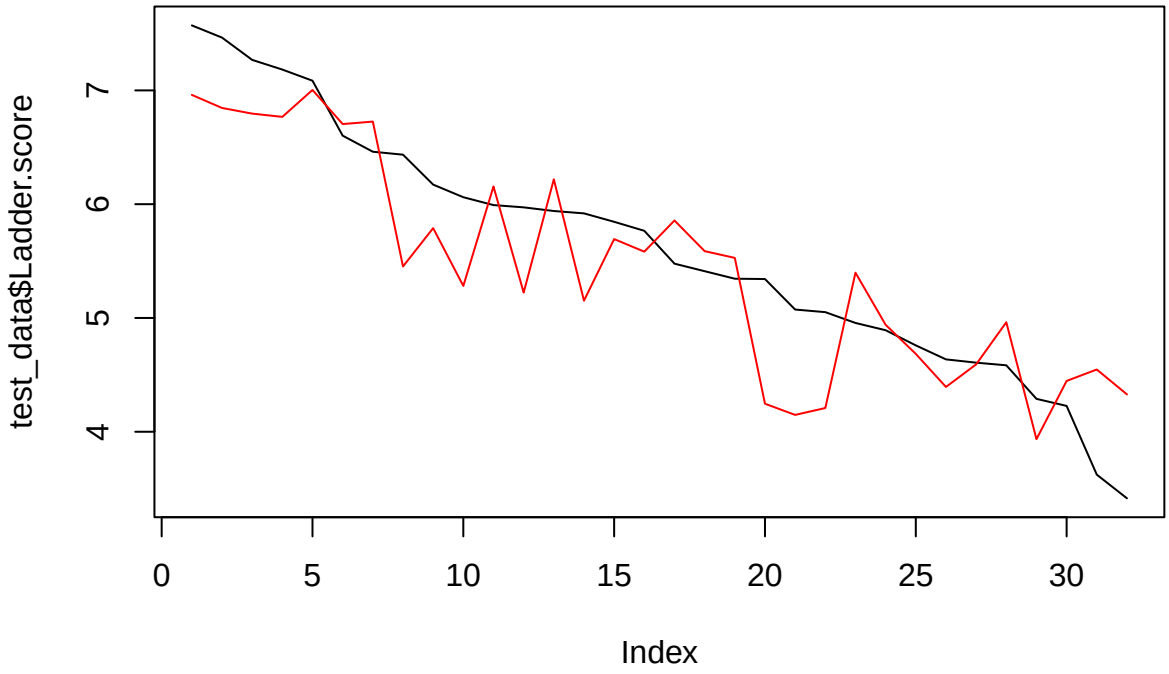
```
##   0.07  -0.09  -0.06  -0.23
##  b->h5 i1->h5 i2->h5 i3->h5
##  -0.49   0.02   0.26   1.01
##  b->o h1->o h2->o h3->o h4->o h5->o
##  0.51  1.51  1.51  5.81 -0.36  1.51
```

Our neural network with the best RMSE has 3 layers, 10 neurons, and 26 weights.

Now we use testing data to test our best linear model, calculate RMSE, and visualize the results in plot of actual and predicted values.

```
## RMSE:
```

```
## [1] 0.5451614
```




### 3.4 Generalized Linear Model (GLM)

Generalized linear model generalizes linear regression model. It allows the linear model to be related to the response variable via a link function. It unifies various other statistical regression models (logistic, Poisson, etc.).

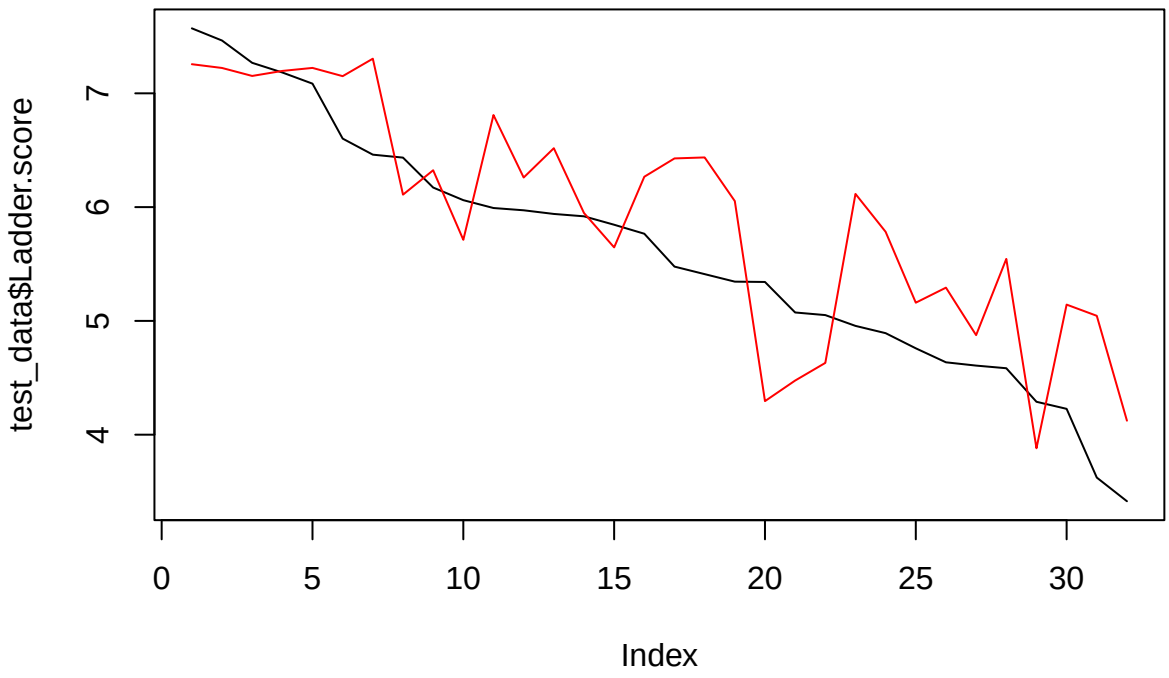
```
## average RMSE:
```

```
## [1] 1.28792
```

Now we use testing data to test our best linear model, calculate RMSE, and visualize the results in plot of actual and predicted values.

```
## RMSE:
```

```
## [1] 0.6654519
```




# 4 Results

Our first goal was to reduce number of regressors as much as possible. Final model should be the most accurate and the simplest, i.e. not overfitted. Lower number of regressors allows faster training.

We used Multivariate regression for this. According to our correlation table, the feature `Healthy.life.expectancy` seemed to be significant but our later observations did not confirm that.

From our results follows that people are happier with higher GDP, social support and freedom to make life choices.

The best method seems to be Multivariate regression but results of the remaining methods are comparable. We can see results in the following table:

| Method | MLR | KNN | NN | GLM |
|---|---|---|---|---|
| RMSE | 0.54 | 0.60 | 0.55 | 0.67 |

Neural networks method provides similar results as Multivariate regression. Final predictions are close to real values.

# 5 Conclusion

Data were selected well. Our outcomes offer good prediction which is close to real values. Each row corresponds to one country. More data from more countries could provide better or worse predictions.

All our models performed well. Results of all methods were very close to real values.

In the report we used Multivariate regression, k-Nearest neighbors, Neural networks, and Generalized linear model methods.

Multivariate regression has the best results. At the beginning, we did research of relationships between regressors. We used the MLR method to eliminate insignificant regressors because their number played big role.

The second best method was Neural networks. This method provided good results and did not have any speed issues with our data so we did not need to rely on tuning our training objects or optimization of parameters.

It would be interesting to compare our methods on larger set of data. Maybe the Neural networks method could become a winner.

There would be interesting additional project which would identify a continent based on our data of happiness, i.e. use classification methods of machine learning.