CSE351 Final Project Report

What makes people in a country happy?

Author 1: Shengnan You 112361646

Author 2: Boren Wang 111385010

Dataset

We have been always curious about what are the most important factors that contribute to people's happiness, so we decided to do a data analysis to find out them. The data we used were the World Happiness Report published by the United Nations Sustainable Development Solutions Network. It ranks the countries by their happiness scores and tries to explain their happiness scores using factors such as GDP per capita, life expectancy, social support, freedom, trust on government, and so on. We had the world happiness report data from year 2015 to 2019, and we used the data from 2015 to 2018 as the training data and the remaining as test data.

Data Pre-processing

The first thing we do is to clean and merge our dataset.

The datasets from 2018 and 2019 have 7 same attributes: 'Country or region', 'GDP per capita','Social support', 'Healthy life expectancy', 'Freedom to make life choices', 'Generosity', 'Perceptions of corruption' and 2 outcome variables: 'Overall rank', which is also completely determinedx by 'Score'. We will use attributes in 2018 and 2019 as standard attributes during the training and testing.

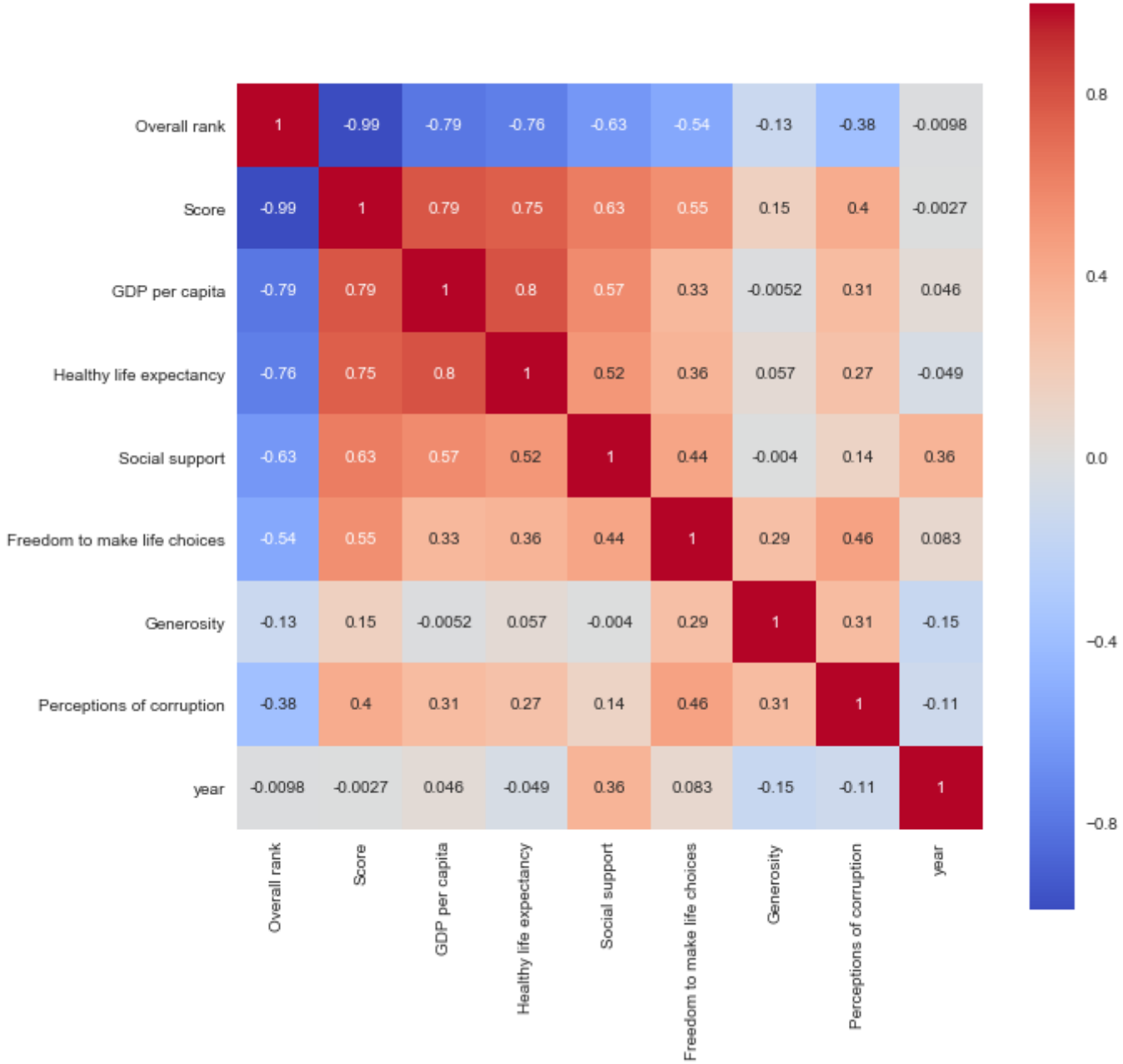
The datasets from 2015 to 2017 contains attributes that do not belong in the standard attributes, so we will remove them. In addition, while some attributes in 2015 and 2017 belong in the standard attributes, they do not have the same names. So, we will rename those columns.

We add a column to each dataset called 'year' to specify which year that data belongs in.

After we processed the data from different years. We now merge datasets from 2015 to 2019 into one dataset, which is our training data. We use the dataset in 2019 as our testing data.

EDA

To get an idea about what features are strongly correlated with people's happiness, we did an exploratory data analysis with lots of visualization using matplotlib and seaborn. We used a heatmap to see the correlation coefficients between happiness score and different variables, and we found out that GDP per capita, Life expectancy, Social support, and Freedom to make life choices have strong correlations with people's happiness. As a result, we decided to use those four variables as features to build models on for future prediction.



Modelling

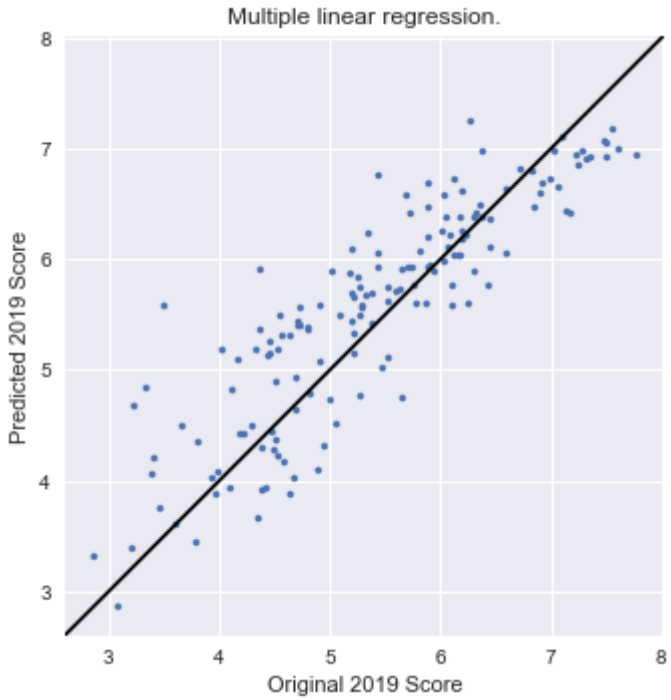
After examining the correlations, we decide to use GDP per capita, Life expectancy, Social support, and Freedom to make life choices (They have the highest correlations) in our modeling process. We built three models for predicting people's happiness scores, and the models use multiple linear regression, neutral network, and support vector machine (SVM) respectively.

Model 1: Multiple Linear Regression:

The first model we built is a multiple linear regressor. We imported linear_model.LinearRegression from the sklearn library and used GDP per capita, Life expectancy, Social support, and Freedom to make life choices as features. We trained the model based on the data from 2015 to 2018, and the linear relation we got is:

happiness_score = 2.29937623994713 + 1.14588287 *GDP_per_capita* + *1.16559883* social_support + 0.5446503 *life_expectancy* + *1.85021572* freedom

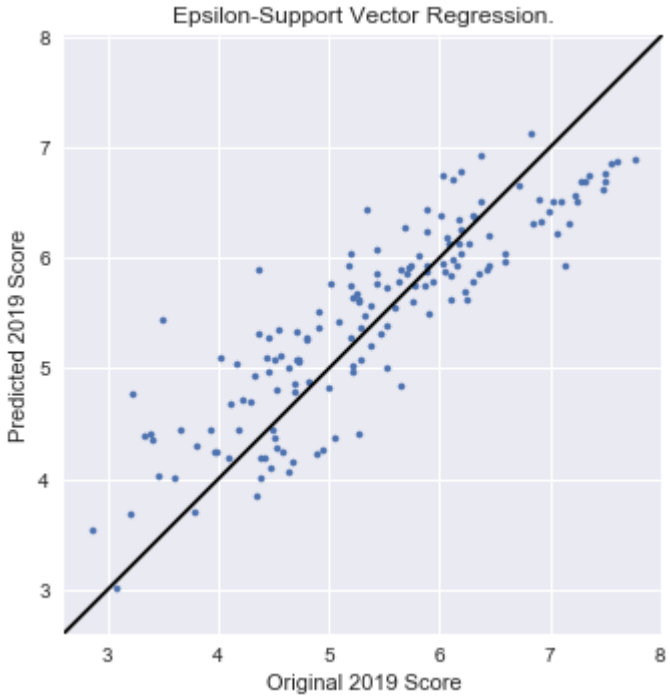
We tested the performance of the model by predicting the happiness score for year 2019 and comparing the predicted scores & ranks with the original scores & ranks and found out that the RMSE for the predicted score was 0.5628852790041101 while the mean of happiness scores was 5.407096153846153. The RMSE is quite small, so we can say the multiple model did a pretty good job at predicting the happiness scores.



Model 2: Epsilon-Support Vector Regression.

The second model we built is a multiple linear regressor. We imported StandardScaler from the sklearn library and to standardize our variables. And then we imported SVR from sklearn.svm lirabry to build our model. The features we use are: GDP per capita, Life expectancy, Social support, and Freedom to make life choices as features. We trained the model based on the data from 2015 to 2018.

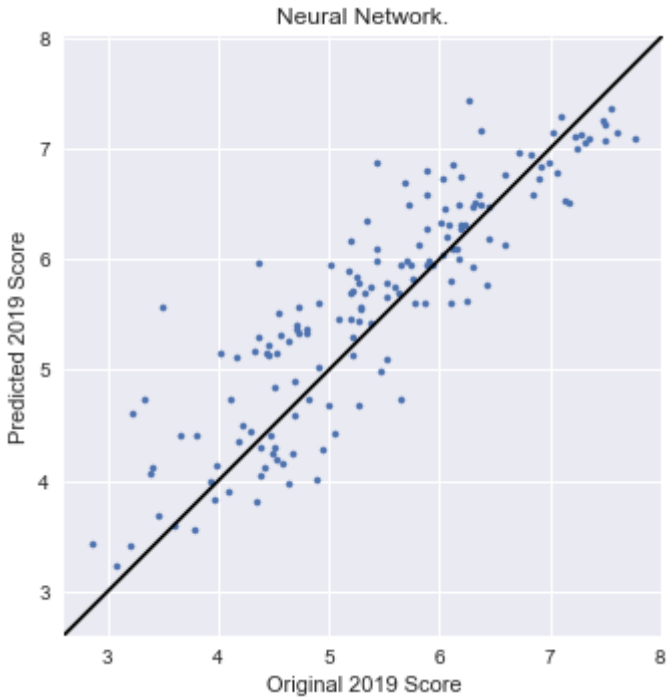
We tested the performance of the model by predicting the happiness score for year 2019 and comparing the predicted scores & ranks with the original scores & ranks and found out that the RMSE for the predicted score was 0.5591065502163332 while the mean of happiness scores was 5.407096153846153. The RMSE is quite small, so we can say the multiple model did a pretty good job at predicting the happiness score



Model 3: Neural Network

The third model we built is a neural network model. We imported MLPRegressor from sklearn.neural_network lirabry to build our model. The features we use are: GDP per capita, Life expectancy, Social support, and Freedom to make life choices as features. We trained the model based on the data from 2015 to 2018.

We tested the performance of the model by predicting the happiness score for year 2019 and comparing the predicted scores & ranks with the original scores & ranks and found out that the RMSE for the predicted score was 0.5473220858274861 while the mean of happiness scores was 5.407096153846153. The RMSE is quite small, so we can say the neural network model did a pretty good job at predicting the happiness score



Conclusion

In conclusion, the most important factors that contribute to people's happiness are GDP per capita, Life expectancy, Social support, and Freedom to make life choices. As a result, in order to make citizns of a country happier, the leader should develop the economy (GDP per capita), build a strong health care system & encourage people to do exercise frequently (Life expectancy), encourage people to support the needed people (social support), and create a free environment where people can make life dicision freely in order to make my citizens happier. In addition, all three models did a good job at predicting a country's happiness score. However, the neural network model is the best among them with the smallest RSME of 0.5473.