Happiness Score Prediction Using Linear Regression and Support Vector Regression Approaches

Tzu Wang
Computing and Information Systems, Queen Mary University of London, London, United Kingdom

ec23460@qmul.ac.uk

Abstract

Happiness score is a crucial indicator for human well-being, reflecting the overall satisfaction of a country's population. The ability to predict the happiness score is therefore beneficial for policy-making and policy adjustments. For instance, by accurately predicting happiness scores, governments can better allocate resources, tailor social welfares, and make plans for improving citizens' lives, overall developing a better country.

Two datasets were used in this research, the 2018 and 2019 happiness reports, downloaded from Kaggle [1]. This research presents the application of two regression models, Linear Regression and Support Vector Regression, to predict happiness scores. The following sections of the paper will provide detailed information on the steps used for this prediction and evaluate each of the results.

Keywords: Happiness Score, Linear Regression, Support Vector Regression, Kaggle

1. Introduction

The world ranking for happiness scores uses data from the Gallup World Poll. The happiness score is rated from 0 (the unhappiest) to 10 (the happiest), based on six factors: GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity and perceptions of corruption. [1]

GDP per Capita

Gross domestic product (GDP) refers to the value of all the goods and services a country produces on a yearly basis. GDP per capita is the GDP divided by the total population of a country. [2]

Social Support

Social support is calculated by a 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?". People can only answer with a yes or no. A yes would equal 1 and a no would equal 0. The average of all answers results in a single value that represents the amount of social support that is present in a country. [2]

Healthy Life Expectancy

Healthy life expectancy represents the average number of years a healthy child is estimated to live, based on the calculation by the World Health Organisation (WHO). [2]

Freedom to Make Life Choices

The freedom to make life choices is calculated by a question: "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?". The average of all answers determines the result of this factor per country. [2]

Generosity

Generosity is calculated by a question: "Have you donated money to a charity in the past month?". The average of all responses (yes being 1 and no being 0) determines the output of this key factor. [2]

Percentage of Corruption

Percentage of corruption is calculated by two questions: "Is corruption widespread throughout the government or not?" and "Is corruption widespread within businesses or not?". The average of all answers determines the outcome of this factor per country. [2]

This research aims to estimate happiness scores with the above six features, using Linear Regression and Support Vector Regression.

1.2. Research Objectives

- Use python libraries to preprocess the data, including data combination, data cleaning, data visualisation and data standardisation.
- Assess the applications of Linear Regression and Support Vector Regression models in predicting happiness scores based on the datasets.
- Check the coefficients between the happiness score and each feature.
- Evaluate the results of both models with cross validation and the calculation of Mean Squared Error (MSE).

2. Literature Review

The first source is a research paper that uses machine learning to analyse the happiness index with Vector Quantization Methodology. The research paper is called "An AI-based Model to Measure World Happiness using Vector Quantization Methodology" by Tanvi Nautiyal, Shruti Agarwal, Shagun Gupta, Garima Sharma and Ankita Nainwal (2023), who used K-means clustering to analyse the happiness index of various countries. This method allowed them to identify significant clustering trends within the happiness data. In their research, they had three clusters of countries: cluster 0 contains the countries with high happiness scores, cluster 1 are countries with low happiness scores, and cluster 2 countries with medium happiness scores. They also performed graphical analysis to investigate the relationships between various global factors that affect the happiness index. [3]

The second source is the study "Understanding World Happiness using Machine Learning Techniques" by Fabiha Ibnat, Jigmey Gyalmo, Zulfikar Alom, Md. Abdul Awal, and Mohammad Abdul Azim (2021) [4]. This study uses machine learning methods to predict the life happiness scores of different countries, showing a key understanding of the factors which influence the scores. The methodology includes three classification models: Decision Table, Random Forest and SMOreg. The Random Forest method performed the most accurately.

The third source is the paper "Predicting Quality of Life using Machine Learning: case of World Happiness Index" by Ayoub Jannani, Nawal Sael and Faouzia Benabbou (2021) [5]. The paper used various machine learning and deep learning models to predict the happiness score, including Lasso Regression, Linear Regression and LSTM. The result of the Lasso Regression model is the most accurate, achieving 89.9% accuracy and 0.0656 RMSE, and the Linear Regression model was also 89% with 0.066 RMSE. Overall, this paper showed the effectiveness of Linear models in predicting happiness score.

The fourth source is the study "Analysis of World Happiness Report Dataset Using Machine Learning Approaches" by Moaiad Ahmad Khder, Mohammad Adnan Sayfi and Samah Wael Fujo (2022). [6] The study analyses the World Happiness Report Dataset using machine learning approaches. It also aims to classify critical variables which affect the happiness score. They used supervised learning ways, such as Neural Network Classifier and OneR Classifier models to analyse the dataset. The findings showed that GDP per capita has the most impact on the happiness score. Classification metrics and confusion matrices were then also used to evaluate the results. The neural network model was shown to have a higher accuracy in classifying and predicting happiness scores.

The last source is the study "Analyzing Determinants of Happiness Score: A Comparison Based on Machine Learning Approaches" by Yuxuan Xiong (2023) [7]. It predicts happiness scores with different machine learning models, including K-Nearest Neighbors, Random Forest, Linear Regression and an ensemble model. The ensemble approach has the highest accuracy 0.3256 MAE, 0.2180 MSE and 0.4669 RMSE. The ensemble model is a model which combines the individual predictions from KNN, RF and LR models, integrating the diverse capabilities of each of them. The ensemble model helps balance the variancebias trade-off, and ensure that the model is neither overfitting nor underfitting.

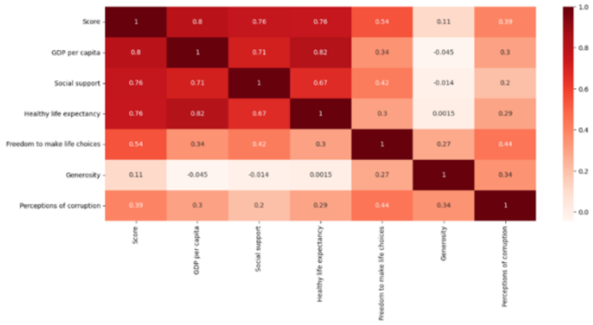
3. DATA PROCESSING

3.1 Datasets

The primary datasets were the 2018 and 2019 happiness reports, downloaded from Kaggle [1]. These two datasets were combined into one dataset containing nine features: Overall rank, Country or region, Score, GDP per capita, Social support, Healthy life expectancy, Freedom to make life choices, Generosity, Perceptions of corruption.

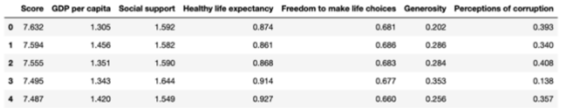
3.2 Data Cleaning and Feature Selecting

The columns "Overall Rank" and "Country or Region" were both dropped in this analysis due to being irrelevant to the machine learning models. The former also directly derives from the "Score" column, making it unnecessary to repeat. We then used heatmap to make clear the correlation between each column which, as seen in Figure 1, with the exception of the Generosity column, all showed significant influence on the Happiness score. We decided to keep the Generosity column for the following steps due to the significant correlations it has with the Freedom to make life choices and Perceptions of corruption columns.



	Score	GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
Score	1	0.8	0.76	0.76	0.54	0.11	0.39
GDP per capita	0.8	1	0.71	0.82	0.34	-0.045	0.3
Social support	0.76	0.71	1	0.67	0.42	-0.014	0.2
Healthy life expectancy	0.76	0.82	0.67	1	0.3	0.0015	0.29
Freedom to make life choices	0.54	0.34	0.42	0.3	1	0.27	0.44
Generosity	0.11	-0.045	-0.014	0.0015	0.27	1	0.34
Perceptions of corruption	0.39	0.3	0.2	0.29	0.44	0.34	1

Figure 1: Correlations Heatmap



	Score	GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
0	7.632	1.305	1.592	0.874	0.681	0.202	0.393
1	7.594	1.456	1.582	0.861	0.686	0.286	0.340
2	7.555	1.351	1.590	0.868	0.683	0.284	0.408
3	7.495	1.343	1.644	0.914	0.677	0.353	0.138
4	7.487	1.420	1.549	0.927	0.660	0.256	0.357

Figure 2: Dataset after feature selecting

The dataset has no duplicates, and there is only one missing value in the column "Perceptions of Corruption". We checked the median and mean of that column, and decided, given that the mean is affected by outliers, to replace this with the median value. In this case, the mean was calculated as 0.111 and the median as 0.082.

3.3 Normality Checking

Before training the model, we checked the normality of the dataset with Q-Q plot. Q-Q plot is a tool used to compare two probability distributions by plotting their quantiles against each other. [8] If the distributions are similar, the points on the Q-Q plot will lie approximately on the line. In Figure 3, even though most of the blue points are close to the red line, there are still some relatively far away, which means our dataset is not normally distributed. Thus, we performed standardisation before we trained our data in the following steps.

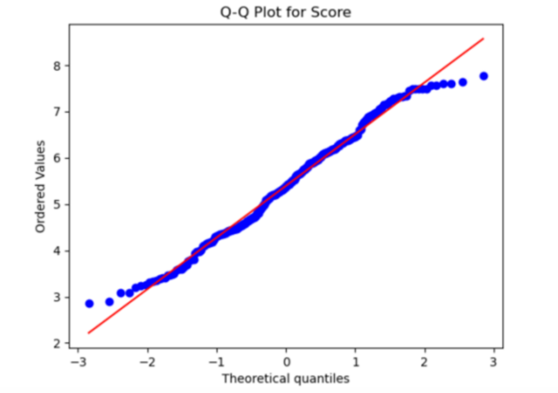


Figure 3 Q-Q plot for Score

4. METHODS

In this research, we first standardised the dataset by using the function imported from sklearn.preprocessing library to change the scale and balance feature contributions. After standardisation, we used two supervised learning models to train with the dataset and make the predictions. Supervised learning is a category of machine learning where an algorithm is trained on labeled data, which means the algorithm learns to make predictions from the input labelled data. In this case, our label is the happiness score. We used two types of supervised learning models: a Linear Regression model and a Support Vector Regression model. In a regression model, the output is a continuous value, like our happiness score from 1 to 10, and the aim is to make the prediction as accurate as possible. [9]

Standardisation

Standardising features involves adjusting their values so that they have a mean of 0 and a standard deviation of 1, which is calculated with the formula z = (x – u) / s, where x is the feature value, u is the feature mean, and s is the standard deviation. This process makes each feature contribute equally. [10]

Linear Regression

Linear regression is a supervised learning algorithm which is used to model the relationship between dependent variables by fitting a linear equation to observed data. It aims to find the best fitting line for the data points in Figure 4. Linear regression is widely used for making predictions.

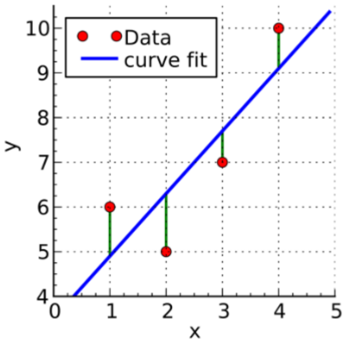


Figure 4 Example of Linear Regression [11]

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p \,,$$

Figure 5 Multiple Linear Regression Equation

Figure 5 shows the formula of multiple linear regression, where Y hat is the predicted value of the dependent value, X1 – Xp are the independent variables, b0 is the intercept, b1 – bp are the estimated coefficients. In our case, Y hat is the Happiness score and X1 to Xp are the other six features. [12]

Support Vector Regression

Support Vector Regression is a supervised learning algorithm used to make predictions. The principle behind it is the same as the Support Vector Machine. The aim for Support Vector Regression is to find the best fit line where the hyperplane has the maximum number of points in Figure 6 [13]. Hyperplanes are decision boundaries that help in making predictions about continuous outputs. [14]

There are different types of kernels for Support Vector Regression. The linear kernel is a simple dot product between two input vectors, and the non-linear (rbf) captures the more intricate patterns in the data. [15] We used both of them to predict our happiness score.

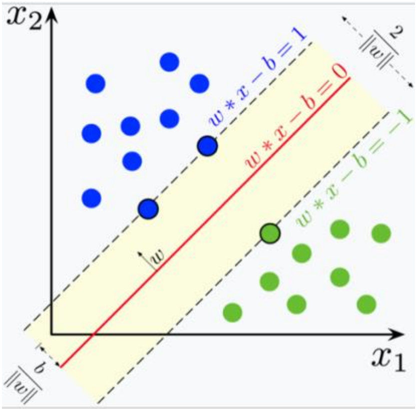
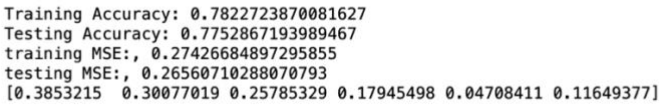


Figure 6 Support Vector Regression Graph

5. Testing and Results

After data pre-processing and standardization, we started to split our dataset into training and testing datasets. The dataset was split into 70% for training and the other 30% for testing, because our data samples size is quite small (312 rows). We also set a random state for training to ensure the results of our models are reproducible and consistent. Thus, we could produce a fair comparison and evaluation between the models. We also used five-fold cross validation and mean squared error to make sure our results were fair and not overfitting. Mean squared error tells us the average number of the prediction errors.

For our linear regression model, we obtained 78% accuracy on training with a 0.27 mean squared error, and a 77.5% accuracy on testing with a 0.26 mean squared error. The coefficients for each column show that the column Generosity, where the coefficient is only 0.047, has the least influence on our Happiness score, while the other columns have significant impact on our Happiness score prediction Figure 7.

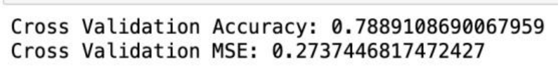


```
Training Accuracy: 0.7822723870081627
Testing Accuracy: 0.7752867193989467
training MSE:, 0.27426684897295855
testing MSE:, 0.26560710288070793
[0.3853215  0.30077019 0.25785329 0.17945498 0.04708411 0.11649377]
```

Figure 7. Linear regression model results

The cross-validation results for our linear regression model have a 78% accuracy and

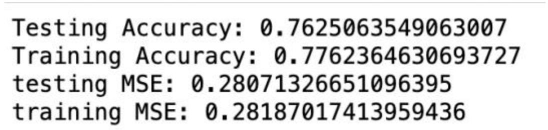
0.27 mean squared error on average, as seen in Figure 8. By comparing the results of our training, testing and cross validation, we can say our model performed in a fair way without overfitting or underfitting, and the accuracy of prediction is reliable.



```
Cross Validation Accuracy: 0.7889108690067959
Cross Validation MSE: 0.2737446817472427
```

Figure 8. Cross validation for Linear regression model

In regards to the Support Vector Regression model, we initially tried to use the linear kernel and the results had a 76% accuracy with 0.28 mean squared error (see Figure 9 below).

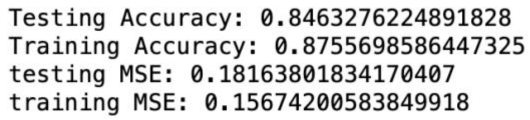


```
Testing Accuracy: 0.7625063549063007
Training Accuracy: 0.7762364630693727
testing MSE: 0.28071326651096395
training MSE: 0.28187017413959436
```

Figure 9. SVR linear kernel results

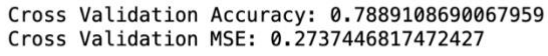
The rbf kernel is extremely reliable for non-linear data, and also allows the model to map out our input features onto higher-dimension spaces. For these reasons, we decided to use the rbf kernel to train our model. The results showed that the testing accuracy was 84.6% with 0.18 mean squared error, and the training accuracy was 87% with 0.156 mean squared error. We also performed cross validation with 79% accuracy and 0.27 mean squared error in average.

The results showed that our SVR rbf kernel model has a higher prediction rate and lower mean squared error than our linear regression model (Figure 10), and the cross validation process confirm our model is fair (Figure 11).



```
Testing Accuracy: 0.8463276224891828
Training Accuracy: 0.8755698586447325
testing MSE: 0.18163801834170407
training MSE: 0.15674200583849918
```

Figure 10. SVR rbf kernel results



```
Cross Validation Accuracy: 0.7889108690067959
Cross Validation MSE: 0.2737446817472427
```

Figure 11. SVR cross validation result

6. Conclusion

This research used machine learning methods to predict the Happiness score based on certain features. In order to obtain better results, we performed data preprocessing, data visualisation and data standardization before training and testing our models. We then trained two machine learning models, the Linear Regression model and the Support Vector Regression model. Both of our models displayed precise prediction rates on the Happiness score, which are 78% and 84% accurate (respectively). We also used five-fold cross validation methods to evaluate our results, showing that our models were fair and there were no overfitting problems.

The biggest limitation for this research is that we had quite a small sample data size (312 rows). If we were able to incorporate more samples, we believe that our models would perform even more accurately.

There is also an alternative method of doing this research. This would be to classify the countries into high happiness countries, mid happiness countries and low happiness countries, similar to the research done by Tanvi Nautiyal, Shruti Agarwal, Shagun Gupta, Garima Sharma and Ankita Nainwal. This can be done by converting the happiness score in the dataset into labels. For instance, score 0 to 4 is low happiness, 4 to 6 is mid happiness and 7+ is high happiness. Instead of using regression models like our studies, we could use classification models like logistic regression to classify countries.

This analysis and research will hopefully help governments in creating better national policies by consulting happiness scores. To enhance the model’s accuracy, we could include more diverse features such as healthcare systems, education levels and the quality of environment etc, revealing deeper insights into the happiness score.

7. Reference

[1]World Happiness Report, Sustainable Development Solutions Network (Owner), 2020 URL:https://www.kaggle.com/datasets/unsdsn/world-happiness

[2]Happiness Index: What is it and How does it work?, Hugo, 2023

URL:https://www.trackinghappiness.com/happiness-index-2018/

[3] An AI-based Model to Measure World Happiness using Vector Quantization Methodology, Tanvi Nautiyal, Shruti Agarwal, Shagun Gupta, Garima Sharma and Ankita Nainwal, 2023

URL:https://ieeexplore.ieee.org/document/10112293

[4]Understanding World Happiness using Machine Learning Techniques by Fabiha Ibnat, Jigmey Gyalmo, Zulfikar Alom, Md. Abdul Awal, and Mohammad Abdul Azim, 2021

URL:https://ieeexplore.ieee.org/document/9768407

[5] Predicting Quality of Life using Machine Learning: case of World Happiness Index by Ayoub Jannani, Nawal Sael and Faouzia Benabbou, 2021

URL:https://ieeexplore.ieee.org/document/9668429

[6]World Happiness Report Dataset Using Machine Learning Approaches, Moaiad Ahmad Khder, Mohammad Adnan Sayfi and Samah Wael Fujo, 2022 URL:https://www.i-csrs.org/Volumes/ijasca/2022.1.2.pdf

[7]Analyzing Determinants of Happiness Score: A Comparison Based on Machine Learning Approaches, Yuxuan Xiong, 2023

URL:http://creativecommons.org/licenses/by-nc/4.0/

[8]Q-Q plot WIKIPEDIA, WIKIPEDIA, 2024

URL:https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot

[9]Supervised Learning, Wikipedia, 2024

URL:https://en.wikipedia.org/wiki/Supervised_learning

[10]Scikit Learn, 2024

URL:https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

[11]Linear Regression WIKIPEDIA, WIKIPEDIA, 2024

URL:https://en.wikipedia.org/wiki/Linear_regression

[12]Multivariable Methods, Wayne W. LaMorte, 2016

URL:https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704-ep713_multivariablemethods/bs704-ep713_multivariablemethods2.html

[13]Support vector machine WIKIPEDIA, WIKIPEDIA, 2024

URL:https://en.wikipedia.org/wiki/Support_vector_machine

[14] Unlocking the True Power of Support Vector Regression, Ashwin Raj, 2020

URL: https://towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0

[15] Support Vector Regression (SVR) using Linear and Non-Linear Kernels in Scikit Learn, Geeksforgeeks, 2023

URL:https://www.geeksforgeeks.org/support-vector-regression-svr-using-linear-and-non-linear-kernels-in-scikit-learn/

[16] Numpy Library

URL:https://numpy.org/

[17] Pandas Library

URL:https://pandas.pydata.org/

[18] Seaborn Library

URL: https://seaborn.pydata.org/