Analyze the World Happiness Report 2021-2022 and predict the 2022 data based on the past year's dataset

**Yifei Zhang1**

*1(1st Affiliation) School of Physics Science, University of Liverpool, Liverpool, United Kingdom.*

|  |  |
| --- | --- |
| *Article history*  Received: 7 July 2014  Revised: 28 August 2014  Accepted: 2 September 2014  \*Corresponding Author:  Zhang, Yifei  School of Physics Science, University of Liverpool, Liverpool, United Kingdom  Email:zhangyifei0728@outlook.com | ***Abstract:*** **Happiness scores are considered appropriate indicators to measure the progress of social development. This work presents two Linear Regression models to predict happiness scores across countries in 2022. Data is sourced from the World Happiness Report dataset from 2015 to 2021, available open-source. Preliminary exploratory data analysis was carried out to select the most appropriate variables to include in the models. The models’ accuracy was tested by comparing the output values to the true 2022 World Happiness Report data. The experiment results show that the Linear regression achieved an RMSE = 0.236, and MSE = 0.056 for 2022.**  **Keywords:** World Happiness Report, Linear Regression, Data Analysis, Machine Learning |

# **Introduction**

Measuring Happiness is challenging due to varied definitions of Happiness. The world has suffered from the global COVID-19 Pandemic for three years, producing unusual results in recent happiness measurements – especially the 2021-2022 Happiness Scores, which represent pandemic struggles, and are not indicative of previous global Happiness Trends.

Motivated by happiness scores of this author’s homeland, we sought to understand how is happiness quantified and how the UN measures happiness, and what countries have a higher score.

The first World Happiness Report was published in 2012 by the United Nations, which compared the happiness of people in 156 countries and regions around the world [1]. According to the world happiness official website, there are six variables: Economic production (gross domestic product, or GDP); Social Support; Life Expectancy; Freedom; Absence of Corruption; and Generosity, all of which contribute to each country’s Happiness Score also referred to as a Ladder Score.

The World Happiness Report is essential for both governments and the public as it provides information that is difficult to quantify in the real world. For the government, happiness scores can provide evidence of their population’s wellbeing. For the civilization, it provides understanding of their actual life quality for many aspects including health care, work environment and education. Hence it is necessary to study and analyze the data. We aim to analyze the data about the World Happiness Report from 2021 and 2022, the distribution of the region for these countries, and the relationship between different variables on the Ladder Score. By using Linear Regression, this study will predict Ladder Scores using each country’s GDP and provide the best fit linear equation. These analyses will be carried out in Python.

# **Literature Review**

In the work of [2], the authors have used nine different ways to predict the Quality of Life through the World Happiness Index including Lasso Regression, Multiple Linear Regression, LSTM, Random Forest Regressor, Support Vector Regression, Gradient Boosting Regressor, XGboost Regressor, MLP Regressor and AdaBoost Regressor. The performance to evaluate the models is MAE, MSE, RMSE, and (the definition of which, is in section 4). The data from 2015 to 2021 has been collected and divided for training (2015-2020) and testing (2021). Their results show that the best performance is achieved using the Lasso Regression with a 0.8954 score and 0.0656 RMSE.

However, this paper is mainly focused on the overall happiness scores but not on how the variables such as GDP, and health influence the World Happiness Scores.

In research presented in [3], the authors identified the essential issues in using data from the importance of the variables included in the dataset. The research uses various Machine Learning Methods such as NN (Neural Network), RF (Random Forest), and GB (XGboost) to classify the GDP as one of the primary indicators of life happiness scores. Also, the insight gained from the study is that high life expectancy may lead to a higher Happiness Score by classifying as the first rule, while the use of OneR classification methods and its results by evaluating different performance indicators increases the finding reinforcement.

Researchers in [4] use various approaches based on the scope of Machine Learning to analyze Global Happiness. The Principal Component Analysis (PCA) was used to analyze gender equality and life satisfaction. In the feature, selection trees were used as well as for life satisfaction prediction. The findings of this study are key characteristics of life expectancy, incoming distribution and freedom summarized using permutation tests. The results show happiness in life in the form of a visual map.

The paper [5] aimed to find the most accurate model to predict Ladder Scores based on the dataset for the top 30 countries and regions in the world. The traditional linear regression model has been used to show that it is risky to be used because of the correlation between various explanatory variables. Ridge regression LASSO regression and elastic net based on Machine Learning are applied to get an accurate prediction and the elastic net model has been found to be the most accurate model in this paper.

Authors in [6] mainly focus on three methods: ANN (Artificial Neuron Network), SVM (Support Vector Machine), and RT (Regression Tree). The most significant prediction method is the ANN # 3 6-20-1 model which has an accuracy of 83.68%, and the significant test for SVM2 is found as R 0.15 and RMSE 0.5454. The regression of the Tree2 model has significant testing of error RMSE 0.57815 that the lowest and enclosed 0. The author has indicated that SVM is simpler to apply than ANN but the first choice of option for the prediction of the World Happiness Score is ANN.

The paper [7] studies various factors to determine their importance in GDP growth and develops a forecasting model to forecast the future using Gaussian Process, Decision Table, Random Tree, Multilayer Perception, and Random Tree and achieved an MAE of 1.801% using Linear Regression.

We can conclude that research about world happiness prediction applied various Machine Learning methods to analyze the results. Most used ML algorithms are LR, SVM, RF, and NN such as LR is used in the context. However, each paper is based on a unique dataset and applies the different methods to predict which might need a high-level knowledge about Machine Learning, this paper is focused on the basic Linear Regression methods in Machine Learning to help beginners in Machine Learning study to analyze the Happiness Score data.

# **Methodology**

*Data Availability*

The database for this research is from Kaggle World Happiness Report [8] which is an essential tool to analyze global happiness. This dataset ranks 155 countries and regions in their Ladder Scores. According to the different survey respondents, their happiness scores were rated on a scale from 0 to 10 marks to show their best possible life. Besides the dataset includes several insights from variables that will influence the scores.

*Describing the Dataset*

The dataset features rely on daily life experiences from respondents based on their considerations of the most persuasive lives and the worst lives. These features are described in detail below.

Table 1: Data sample 10 rows.

We will use the World Happiness Report 2021 as an example to show the top 10 countries' specific data in table 1.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| RANK | Country | Happiness Score | Dystopia Residual | Economy  (GDP per Capita) | Family | Health (Life Expectancy) | Freedom | Generosity | Trust (Government Corruption) |
| 1 | Finland | 7.821 | 2.518 | 1.892 | 1.258 | 0.775 | 0.736 | 0.109 | 0.534 |
| 2 | Denmark | 7.636 | 2.226 | 1.953 | 1.243 | 0.777 | 0.719 | 0.188 | 0.532 |
| 3 | Iceland | 7.557 | 2.32 | 1.936 | 1.32 | 0.803 | 0.718 | 0.27 | 0.191 |
| 4 | Switzerland | 7.512 | 2.153 | 2.026 | 1.226 | 0.822 | 0.677 | 0.147 | 0.461 |
| 5 | Netherlands | 7.415 | 2.137 | 1.945 | 1.206 | 0.787 | 0.651 | 0.271 | 0.419 |
| 6 | Luxembourg\* | 7.404 | 2.042 | 2.209 | 1.155 | 0.79 | 0.7 | 0.12 | 0.388 |
| 7 | Sweden | 7.384 | 2.003 | 1.92 | 1.204 | 0.803 | 0.724 | 0.218 | 0.512 |
| 8 | Norway | 7.365 | 1.925 | 1.997 | 1.239 | 0.786 | 0.728 | 0.217 | 0.474 |
| 9 | Israel | 7.364 | 2.634 | 1.826 | 1.221 | 0.818 | 0.568 | 0.155 | 0.143 |
| 10 | New Zealand | 7.2 | 1.954 | 1.852 | 1.235 | 0.752 | 0.68 | 0.245 | 0.483 |

Before proceeding with the analysis of the dataset, we provide some brief intuition to the reader, to understand the world happiness report on 2021-2022. Figure 2, it has shown the Happiness Score 2022 for the countries and regions in the dataset we used through a world map.

地图

描述已自动生成

Fig. 2: 2022 Happiness Score distribution in a world map vision

In Figure 3, it has shown the Happiness Score 2021 for the countries and regions in the data set we used through a world map.

地图

描述已自动生成

Fig. 3: 2021 Happiness Score distribution in a world map vision

It seems like no countries and regions have dramatic changes between 2021-2022. The happiness scores are distributed to similar scores in different states. Figure 4 shows the details about different regions on the happiness scores.

图表, 条形图

描述已自动生成

Fig. 4: Happiness Score for different regions

Next, we explore the dataset to search for drastic changes between these two years, shown in Figure 5.

图表, 直方图

描述已自动生成

Figure 5: Density distribution for Happiness Score between 2021-2022, blue representing 2021 and orange 2022.

From Figure 5, what we can conclude is that the basic distribution line is not changed in a dramatic way. The mean happiness score for 2022 is higher than the 2021 Happiness Score. The score intersection between 5-7 has a noticeable increase.

For the 6 variables Economic production (GDP), Social Support, Life Expectancy, Freedom, Absence of Corruption, and Generosity, the next step is to figure out which will influence the Ladder Scores more rapidly. Thus, we construct a heat map to get a clear view of the relationship between these 6 variables and the Ladder Score which will be shown in Figure 5.

图表

描述已自动生成

Fig. 6: The heat map of each parameter’s correlations

As we can see, the most essential variable that will influence the Happiness Score is the Economy (GDP per Capital) and the Family and it has the highest correlation with Happiness Score which is 0.78 and 0.76. The least important variable is generosity which only has a correlation with happiness score of 0.18. To have a clearer view, we construct the plot diagram (Fig. 7) for every two variables.

图表

描述已自动生成

Fig. 7: Scatter diagram of the distribution between each parameter

So, we choose the Economy (GDP per Capital) to be the key variable to predict the Happiness Scores.

The Machine Learning algorithm we choose to predict the data is Linear Regression. Various python packages were used to accomplish different purposes during the research.

NumPy [9] (Numerical Python) is an extensive library of the Python language, which supports a large number of dimensional array and matrix operations, and also provides a large number of mathematical functions libraries for array operations.

Pandas [10] is an open source, BSD-licensed library that provides high-performance, easy-to-use data structures, and data analysis tools.

Sklearn [11] (Scikit-Learn) is a powerful Machine Learning library that covers everything from data preprocessing to model training.

Seaborn [12] is a data visualization library based on Matplotlib. It builds on Matplotlib with a higher level of API encapsulation to make drawing easier and refined without a lot of tweaking.

# **Results**

In order to evaluate the performance of these models we used several metrics, which are:

* Mean Absolute Error (MAE) is a measure of errors that corresponds to standard L1 and measures the average of the absolute difference between the actual and predicted values [8].
* Mean Squared Error (MSE) is a criterion that measures the mean square error of the mismatch between predicted and real values [8].
* Root-Mean-Square Error (RMSE) is the square root of Mean Squared Error [8].

We used these metrics to evaluate the proposed models. They are defined as follows:

Where represents the total number of elements of the test data,  is the predicted value and  is the corresponding true value of the sample.

For the Linear Regression, we set our target equation to be where is the GDP value and is the Happiness Score of the sample. Then we selected 75% data as the training set and the remaining 15% for testing. The resulting model is shown in Figure 8.

图表, 散点图

描述已自动生成

Fig. 8: Distribution of the training model

Through our training we obtain the results that:

= 0.28465782843416604

= -0.6227407645545371

The target equation will be:

Subsequently, the next 15% of the data, used for testing, is shown (the green points) in Figure 9. It can be verified that the test data is in fact evenly distributed on both sides of the target line.

图表, 散点图

描述已自动生成

Figure 10: The training and test data distribution.

Finally, we let the remaining 10% of the dataset is used to predict the Happiness Score through the target line. The results of the validation are shown in Fig. 10, where we show that the data is appropriately fitted by the Linear Regression model we generated.

图表, 散点图

描述已自动生成

Fig. 11: Testing the model with comparison between predicted and actual values for 2022 Happiness Scores

After applying the algorithms to the data, we used the three performance equations mentioned earlier to evaluate our model. The results obtained are the following:

MAE: 0.195

MSE: 0.056

RMSE: 0.236

**Discussion:**

From the above result, MAE, RMSE, and MSE are all expected to be as small as possible. In [6], we get the other authors' results about the RMSE, MSE, and MAE which have been cited above which is RMSE=0.5454.

By comparing these two models, our model's RMSE is smaller than the ANN model in [6].

The results we obtain and 6 are of a satisfactory range.

However, it remains uncertain whether GDP is the best variable to predict the Happiness Score. From the heat map in Figure 6, it can be observed that the variable of Family also displays a high correlation to Happiness Score. Thus, we construct another Linear Regression model utilizing family as the predictive feature and assess how this model performs in comparison to the one made using GDP.

Figure 12 shows the distribution between happiness scores and Family. The gradient of this target line is positive. Thus, with the increase of the Family Score, the Happiness Score should increase with the Family Score which is the same as the GDP. However, the distribution for the Family Score exhibits more dispersion than the GDP.

图表, 散点图

描述已自动生成

Fig. 12: The training data distribution for Family Score

The gradient of the Linear Regression model is and the y-intersection is equal to which will give the equation of the target line:

Subsequently, 15% of the data will be used to test the model. The green points will be the test points and give a clear information that the test points have evenly distributed on both sides of the line which is also similar to the results of GDP which will be shown in Figure 13.

图表, 散点图

描述已自动生成

Fig. 13: Test points distribution for Test points.

Last but not the least, we got RMSE, MSE, and MAE to evaluate the accuracy of the prediction, obtaining the following results:

Compared to the numbers above in the paper, three numbers are all increased which means the prediction for the Happiness Score is not as accurate as the prediction by GDP. In that way, using GDP gives us the best variable to predict the Happiness Scores.

During the research, the key question is to determine which of the six variables influences the Happiness Scores most or which of the parameter can predict the Happiness Score most accurately. Through the methodology, heat maps are drawn to show the correlation between each parameter and the Happiness Score. Two parameters that have a higher correlation coefficient will be considered as the target parameters. However, a higher correlation coefficient doesn’t mean having the ability to predict the Happiness Score better. Through our results, the Family variable has the highest correlation coefficient 0.78. However, when utilizing this variable to build a Linear Regression model, the RMSE for the Family model is greater than the RMSE for the GDP.

In conclusion, GDP is the best parameter to predict the Happiness Score when using a Linear Regression model.

However, the limitation of Linear Regression is also obvious. For a given target line, if one country’s GDP is greater than another country our model will predict its Happiness Score must be greater than another country which is not a certain result. Because there are another 5 variables to determine the Happiness Score it might have some special circumstances for example Singapore and Hong Kong. In 2022, Singapore and Hong Kong are third and ninth rank in GDP but only get twenty-seven and eighty-one in Happiness rank. Their Dystopia Residuals are not high enough to support them to get higher marks though the GDP has contributed a lot.

Thus, a machine learning method that includes more variables should be considered for further research to get a more accurate result.

# **Conclusion**

Happiness research remains a challenge for researchers concerned with this aspect of societal development. The use of data science tools to model and analyse happiness predictions can be very useful in addressing the challenges associated with this subject matter, and aid in future research.

The limitations of using machine learning methods to predict the happiness score also exist. A number of outstanding events have not been taken into account during the process of the predictions, which will undoubtedly have a significant influence on the happiness scores of specific countries. For example, in February 2022 Russia-Ukraine war broke out. This war is highly likely to influence the Happiness Score for both Ukraine and Russia. Our model can only simulate a trend between GDP and Happiness Score, but events like these will not be taken into consideration, thus leading to cases where the model fails to accurately predict happiness scores.

In our work, we delve into the literature related to this concept from a data science perspective, where machine learning and deep learning algorithms are used. In this paper, a brief analysis of the comparison of World Happiness Reports between 2021 to 2022 is given. The Happiness Scores do not have a dramatic change between these two years. We proposed an experimental approach to explore the potential of Linear Regression models in predicting happiness Scores and then compared their performances. The performance is achieved with an MAE: 0.195, MSE: 0.056 and RMSE: 0.236. A comparison group is constructed by using another important factor Family to predict the Happiness Score. The result gives that MAE: 0.214, MSE: 0.066, RMSE: 0.258. All three metrics of error are larger in the Family model compared to the GDP model, which asserts that GDP is the best variable to predict the Happiness Score in a Linear Regression model.

Using these techniques, we can study, predict, and model happiness accurately by discovering the most highly correlated variable: GDP. For future research, additional types of models to predict happiness scores such as ANN and SVM can be used. For further study, we can combine the other variables like Social Support, Life Expectancy, Freedom, Absence of Corruption, and Generosity, to contribute to the prediction more accuracy by adding more layers to reduce the dimension.

**Acknowledgments**

Throughout the writing of this work, I have received a great deal of support and assistance.

I would first like to thank my supervisor, Aida Manzano Kharman, a PhD student at Imperial College London, whose expertise was invaluable in formulating the research questions and methodology. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

I would also like to thank my tutors, Tanvika Parlikar, Chloé-Rose Colombero, for their valuable guidance throughout my studies. You provided me with the tools that I needed to choose the right direction and successfully complete my dissertation.

In addition, I would like to thank my parents for their wise counsel and sympathetic ear. You are always there for me.

Finally, I could not have completed this dissertation without the support of my friends, Haoming Guo, who provided stimulating discussions as well as happy distractions to rest my mind outside of my research.

# **References**

[1] J. HELLIWELL, R. Layard and J. Sachs, "WORLD HAPPINESS REPORT 2016," Sustainable Development Solutions Network., New York, 2016.

[2] Jannani, A., Sael, N., & Benabbou, F., Predicting Quality of Life using Machine Learning: case of World Happiness Index. In 2021 4th International Symposium on Advanced Electrical and Communication Technologies (ISAECT) (pp. 1-6). IEEE. 2021

[3] Khder, Moaiad Ahmad, Mohammad Adnan Sayfi, and Samah Wael Fujo. "Analysis of World Happiness Report Dataset Using Machine Learning Approaches." *Int. J. Advance Soft Compu. Appl* 14.1 (2022).

[4] L. Millard, "Data Mining and Analysis of Global Happiness: A Machine Learning Approach," University of Bristol:DEPARTMENT OF COMPUTER SCIENCE, Bristol, 2011.

[5] Tan, Y., Singhapreecha, C., & Yamaka, W., Applying Machine Learning to Predict Happiness: A case study of 20 Countries. Mukht Shabd Journal, 9(6), 3433-3437. 2020

[6] Khemraj, S., Thepa, P. C. A., Chi, H., Wu, W. Y., Samanta, S., & Prakash, J., Prediction of world happiness scenario effective in the period of COVID19 pandemic, by artificial neuron network (ANN), support vector machine (SVM), and regression tree (RT). NVEO-NATURAL VOLATILES & ESSENTIAL OILS Journal| NVEO, 13944-13959. 2021

[7] U. Salma, M. M. Hasan and M. I. Sultana, "A Regression Model-Based Approach to Identifying Determining Factors for GDP Growth in Bangladesh", 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), pp. 1-6, Feb. 2020.

[8] A. E. H. Bouzid Ait-Amir and Philippe Pougnet, "6 - Meta-Model Development", Embed. Mechatron. Syst., vol. 2, pp. 157-187, 2020.

[9] Harris, C.R., Millman, K.J., van der Walt, S.J. et al. *Array programming with NumPy*. Nature 585, 357–362 (2020). DOI: [10.1038/s41586-020-2649-2](https://doi.org/10.1038/s41586-020-2649-2). ([Publisher link](https://www.nature.com/articles/s41586-020-2649-2)).

[10] [Data structures for statistical computing in python](https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf), McKinney, Proceedings of the 9th Python in Science Conference, Volume 445, 2010.

[11] [Scikit-learn: Machine Learning in Python](http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html), Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[12] Waskom, M. L., (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021, https://doi.org/10.21105/joss.03021