Predicting Happiness from International Indicators

Matthew Moravec
Marquette Department of Math and Statistical Sciences
Marquette Department of Computer Science
Milwaukee, Wisconsin 53233
Matthew.Moravec@marquette.edu

Rajib Alam
Marquette Department of Computer Science
Milwaukee, Wisconsin 53233
rajib.alam@marquette.edu

Thomas Florian
Marquette Department of Computer Engineering
Milwaukee, Wisconsin 53233
thomas.florian@marquette.edu

Abstract

This paper attempts to predict the happiness score of a nation based on various international indicators using a variety of modeling techniques. Building a relationship between happiness and the various qualities of a nation could help provide insight on what aspects lead to a generally content population, and this data could be used to improve quality of life. The models can then be used to identify which international indicators are most important to having a happy population. Various models are analyzed throughout the paper, each with its own advantages and disadvantages: linear regression models, K-nearest neighbor classification model, and a non-linear neural network.

1 Introduction

Happiness itself is hard to define, and it is not always clear what makes someone happy as it is dependent on a multitude of variables; however, it is commonly accepted that everyone wants to be happy. This project has the goal predicting a country's happiness based on different international indicators and additionally aims to evaluate the relevance of those indicators to build an understanding of the aspects of life are more important for happiness. A literature review on happiness was conducted before and after starting the project. A psychology textbook attempted to note some factors that are connected to happiness, mentioning that a person with a happy family, happy marriage, great friends, and other social relationships are usually happier, and that older people tend to be more satisfied with life. Moreover, the textbook notes that money and economy are related to happiness and that well-being rises with the increase of annual income, but both only up to a certain level. Happiness is also described as positively correlated with education and employment as happy people are more likely to graduate from college and get a satisfying job. The authors of the textbook also identified some factors that are not generally correlated with happiness, including gender, intelligence, parenthood, and attractiveness.[1] A different analysis of various datasets on happiness revealed that there is a strong correlation between income and happiness, and that countries where people tend to live longer also tend to be happier. Also, there was a positive correlation between the sense of freedom and happiness. The other factors that correlated with life satisfaction and happiness are smiling frequency, frequent verbal expression of positive emotions, sleep quality, active involvement in religion, and recent positive change of circumstances (for example, marriage).[2] A very similar research to ours in 2017 concluded that GDP per capita is the most important factor for happiness, and family and health are important factors as well.[3] Another similar study analyzing the World Happiness Report concluded saying, “Economic GDP tends to have the greatest impact on happiness with Health following close by.” [4] Again, another study on the happiness score pointed out that happiness had a strong positive correlation with GDP per capita, social support, health life expectancy, positive correlation with freedom to make life choices, fragile linear relationship with generosity, and a very weak linear relationship with corruption of a country.[5] For the most part, our research agrees with the findings of the other researches. However, most of these studies tried to predict the most important factors of happiness only, whereas, we also wanted to predict the happiness of country that is not part of our training dataset as well as the most important features in predicting that happiness level.

2 Methodology

2.1 Dataset

After setting our research goal, the next step was to retrieve, process, and prepare the required data. To that end, the first was to obtain the happiness score dataset from the World Happiness Report from Kaggle. We used the report published in 2015. We initially added the following features: Household size, Income per capita, Percent urban, Births per 1000, Deaths per 1000, Life expectancy at birth, Percent young (<15), Percent old (>65), Infant mortality rate, and fertility rate. These datasets were collected from Population Reference Bureau (PRB). Three things were crucial in processing our data – cleansing the datasets, transforming them, and combining them with the happiness score dataset to obtain a single dataset for model development. To clean our data, we looked for values that are typo, outliers, or impossible. We also paid attention to missing values. To remove these missing values from different datasets, we used *DataFrame.dropna*() function. Some of our datasets had data of multiple years for the same country. Therefore, we used the functions *groupby()* and *max()* to get the desired value for each country. In some cases, we used the function *set_index()* to set the index of dataset to the name of the countries which eventually came in handy when we were combining different datasets. Again, we noticed that the same countries were recorded with different names in different datasets. Therefore, we used *DataFrame.rename()* method to change those names. We also modified a few things directly using the spreadsheet for some datasets. For example, we used the TRIM function in excel to remove the whitespaces in the country names. Once we had the separate datasets ready, it was fairly easy to combine them using the function *DataFrame.join()*. Later, we added more features listed as follows: Gross enrollment ratio (tertiary education), School life expectancy, Human freedom index, Economic freedom index, and Press freedom index. Similar to the process before, we cleaned, and transformed the datasets before combining them with the main dataset using *pandas.concat()*. It is also to be noted that we dropped quite a few columns from different datasets since they were not relevant for our purpose. However, after we combined these datasets with the happiness score dataset, we ended up losing a significant amount of datapoints because these datasets either have information of less number of countries, or the countries are different from the countries that we had in the happiness score dataset. Altogether, though we started with 158 countries in the happiness score dataset, in the end, we had 125 countries with 16 features. Then, we divided the datasets into two parts – training dataset comprising of 92 countries and test set with 33 countries. We separated the test set from the training set to remove any sort of bias while predicting the happiness of the countries in the test set.

Country	Household Size	Income Per Capita	Percent Urban	Births Per 1000	Deaths Per 1000	Life Expectancy at Birth	Percent Young (<15)	Percent Old (>65)	Infant Mortality Rate	Fertility Rate	Gross Enrollment Ratio	Literacy Rate	School Life Expectancy	Human Freedom Score	Economic Freedom Summary Index	Press Freedom Index
Angola	0.811268	-0.698781	0.127181	2.150234	0.449983	-1.460536	1.84121	-1.1136	2.054723	2.376442	-0.982624	-0.643233	-0.820364	-1.284213	-2.433957	0.203097
Argentina	-0.357364	0.031567	1.41248	-0.295362	0.073634	0.633211	-0.30865	0.34167	-0.62941	-0.27873	1.405405	0.720988	1.188372	-0.135582	-1.311504	-0.191336
Armenia	-0.201547	-0.400404	0.171502	-0.74825	0.449983	0.386888	-0.66697	0.34167	-0.7608	-0.76784	0.417472	0.805321	-0.249225	0.272609	1.020585	-0.205149
Australia	-0.980635	1.286983	1.146556	-0.657672	-0.679066	1.495343	-0.75654	0.92378	-0.89689	-0.69797	1.629263	0.775556	2.172729	1.411747	1.358411	-0.848982
Austria	-1.214361	1.625857	-0.050102	-0.929405	0.449983	1.125858	-1.20443	1.36036	-0.91566	-0.83772	1.419041	0.725948	0.838957	1.269355	0.889814	-1.188932

Figure 1: Dataset visualization

2.2 Models

Three types of models were used to predict the happiness scores of countries based on the 16 different international indicators: linear models, classification models, and a non-linear neural network. All models were trained using the same training set, any hyperparameter optimization was performed using a leave-one-out cross validation of that training set, and all models were evaluated with the same test set.

2.2.1 Linear Regression

When analyzing the effect different features of the dataset had on the model, we utilized three different linear regression models, the ordinary least squares regression, the ridge regression and the lasso regression. The ordinary least squares regression is our simplest model, making it the center of comparison. The lasso and ridge regressions are both regularizing, making them ideal for showing the effect a feature has on the model. The reason for including both the lasso and ridge, is to show the difference is coefficients for significant features when the insignificant features can have coefficients equal to zero compared to when they cannot have coefficients equal to zero. We also performed these regressions with a 10-fold cross validation to see the effect cross validation has on the model.

2.2.2 Classification Model

In our attempt to predict the happiness of a country more accurately, we implemented K-nearest neighbor algorithm on our training dataset. In the happiness report, the countries are scored from 1 to 10, and the final scores are continuous values. Through classification, we can label countries more generally as a class. After labeling the countries, we plotted Andrews curves and Parallel coordinates to visualize the multivariable data, and it was very evident that the datapoints of different classes are clustering together for the most part. There were also discrepancies which suggested that some of the features are not good indicators of happiness of a nation. We fit the data to KNN classifier initially with a random value for K, then through optimization, we came up with the best K-value for our dataset.

2.2.3 Neural Network Architecture

Prediction of the happiness score with the 16 international indicators was conducted using a multilayer convolutional neural network. A leave-one-out cross validation of the training set was used to determine the best hyperparameter values as well as the best network architecture. The convolutional architecture was preferred as it provided a better convergence of the loss function compared to a fully connected architecture. The final model had six 1D convolution layers with a rectified linear unit (ReLU) activation function, 8 filters, and a filter size of 1. The model was then flattened into two dense layers with 32 neurons each followed by a dropout layer with a 0.1 dropout rate and finally into a single output. The model was compiled with a mean squared error loss function and Adam as the optimizer. The model was trained with 200 epochs, a batch size of 32, and a learning rate of 0.002.

2.2.4 Neural Network Feature Selection

A common drawback of using neural network models is their lack of interpretability, often serving as just a black box. This issue was addressed with Layer-wise Relevance Propagation,[6] which was used in order to calculate a mean relevance score for each feature with a backwards propagation of the prediction in the neural network. This score provides a metric to rank the relevance of the feature to the model prediction. Features with a mean relevance score near zero were deemed as irrelevant to the model prediction. The training and test sets were merged in order to determine the important features of the dataset.

3 Results and Analysis

We use three different approaches to predicting a countries happiness. One approach is using linear models to predict the exact value of a country's happiness score. This approach also shows the weight a feature has on the model. Another approach is a classification method that predicts whether a country is generally unhappy, neutral or happy. The third approach is a neural network to predict a countries happiness.

3.1 Linear Models

The overall effectiveness of the linear models is not very good. The R^2 score of the models on the training data was significantly higher than the R^2 score of the model with the testing data. This generally signifies overfitting and suggests a need to alter the alpha value in the lasso and ridge regressions to alter the regularization in the model. In figures 2 and 3 it is evident that even with regularization optimization the models are not good fits with the test data. The 10-fold models performed better than the regular models, indicating that our model. However, it is still not a great predictor. The linear models are not very good predictors of a country's happiness; however, they can still tell us a lot about the features that are indicative of a country's happiness.

Table 1: R^2 Scores for the OLS linear regression

	Linear Model R^2 Score
Train Data	.777
Test Data	.461

Table 2: R^2 Scores for the lasso regression

	Lasso Model R^2 Score
Train Data alpha = .1	.72
Test Data alpha = .1	.46
Train Data alpha = .05	.749
Test Data alpha = .05	.492
Train Data alpha = .01	.770
Test Data alpha = .01	.495
Train Data alpha = .001	.776
Test Data alpha = .001	.470

Table 3: R^2 Scores for the ridge regression

	Ridge Model R^2 Score
Train Data alpha = 1	.775
Test Data alpha = 1	.486
Train Data alpha = 10	.763
Test Data alpha = 10	.502
Train Data alpha = 20	.754
Test Data alpha = 20	.497
Train Data alpha = 50	.735
Test Data alpha = 50	.487
Train Data alpha = 100	.714
Test Data alpha = 100	.456

Table 4: R^2 Scores for the lasso regression 10-Fold

Linear Model R^2 Score 10-Fold Average	.594
Lasso Model R^2 Score 10-Fold Average	.595
Ridge Model R^2 Score 10-Fold Average	.603

The significant features differ in each model. The ordinary least squares regression has the most significant features, as expected, because there is no regularization. The number of significant features in the lasso and the ridge regressions is less than the ordinary least squares linear regression, due to the regularization in those models. The difference between the models comes from the lasso regression being able to reduce some coefficients to zero while the ridge cannot reduce coefficients to zero. Looking at the features with p-values less than .05, we can see that the most common significant feature is income per capita. This makes logical sense for a certain range of income. If the average person in the country is not in poverty, then overall happiness will be greater than if the average person is in poverty. However, once the average income reaches a certain value (not important in this context) then marginal happiness from the additional income will decrease. The second most common significant feature is the press freedom index. This is also a logical indicator of happiness, as freedom of the press is typically indicative of a country where people are not oppressed. Knowing the significant features is important as it shows the features influence the model, however, it is not the only indication of feature importance. By looking at the coefficients of the significant features we can see the impact they have on the model.

Linear Regression	Income Per Capita: .002 Deaths Per 1000: .018 Infant Mortality Rate: .019 Gross Enrollment Ratio: .047 Press Freedom Index: .010
Lasso Regression	Income Per Capita: .002 Press Freedom Index: .031
Ridge Regression	Income Per Capita: .029

Figure 2: P-values for regression coefficients

What the three models have in common is the significance of income per capita in the model. The difference in the model is the coefficients of the common significant features. The only significant feature for all three linear models is income per capita. In the ordinary least squares linear model the coefficient for income per capita is relatively high, but not as high as infant mortality. Both features are significant in the model, so the higher coefficient is indicative that the infant mortality has a large impact on this model than income per capita. The other models contradict this interpretation, as infant mortality is not a significant feature. By this we can see the regularization's effect on the lasso and ridge regression models. The main difference between the lasso and ridge regression models is the inclusion of the press freedom index as a significant feature. The coefficient for income per capita in the lasso regression is larger than in the ridge regression, indicating that income per capita is more influential in the lasso regression compared to the ridge. However, the lasso also has a feature with a negative coefficient, the press freedom index. Interestingly the coefficient for income per capita in the ridge regression is very close the coefficient for income per capita in the lasso regression minus the coefficient for the press freedom index in the lasso regression. While the press freedom index and income per capita to not change at the same rate, the increased weight I of income per capita in the lasso model could be influenced by the other significant coefficients. Overall, the linear models show that the income per capita is an important indicator of predicting the overall happiness of a country and the coefficients seen in figure 8 represent the weight each significant feature has on the model.

Table 5: Coefficients for significant features in each model

	Coefficient
Linear Regression	Income Per Capita: 0.347 Deaths Per 1000: -0.327

	Infant Mortality: 0.493 Gross Enrollment: 0.2233 Press Freedom Index: -0.253
Lasso Regression	Income Per Capita: 0.352 Press Freedom Index: -0.252
Ridge Regression	Income Per Capita: 0.197

3.2 Classification Model

The first step for the classification purpose was to standardize the data using the function StandardScaler(). Since, KNN is a supervised algorithm, the training data had to be labeled or classified. We decided to have three classification for the countries – High-range, Mid-range, and Low-range. Now, we had different options in terms of how we define these ranges. The low-range was defined as less than 5 ($x<5$), the mid-range as greater than or equal to 5 and less than 6 ($5\leq x<6$), and the high-range as greater than equal to 6 ($x\geq 6$) on the happiness score. Once the dataset was ready for the model, using KneighborsClassifer( ) a KNN classifier was created setting the number of neighbors equal to 5. Similar to the training data, the test dataset was also processed for classification. Afterwards, using the trained model, the class labels were predicted using the predict( ) method of the classifier. Then, the classification report using classification_report() and confusion matrix using confusion_matrix( ) were printed out. They are as follows:

	precision	recall	f1-score	support
0	0.50	0.67	0.57	3
1	0.68	0.88	0.77	17
2	1.00	0.54	0.70	13
accuracy			0.73	33
macro avg	0.73	0.70	0.68	33
weighted avg	0.79	0.73	0.72	33

Figure 3: Initial classification report

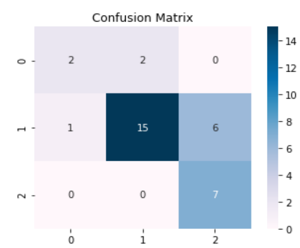

Figure 4: Initial confusion matrix

In order to improve the model, the error rates were plotted along with number of neighbors ranging from 0 to 40. It was then found that 21 is the optimal number of neighbors for our dataset.

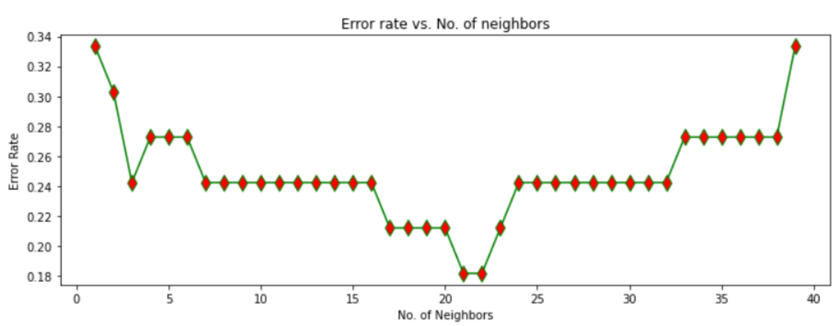

Figure 5: Error rate vs no. of neighbors – KNN classification

The KNN classifier was then modified using the number of neighbors as 21, and the class labels were predicted again. The metrics report for the modified classifier is as follows:

	precision	recall	f1-score	support
0	1.00	0.67	0.80	3
1	0.74	1.00	0.85	17
2	1.00	0.62	0.76	13
accuracy			0.82	33
macro avg	0.91	0.76	0.80	33
weighted avg	0.87	0.82	0.81	33

Figure 6: Final classification report

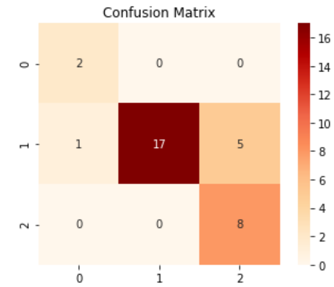

Figure 7: Final confusion matrix

The model now has weighted average precision score of 0.87, recall score of 0.82 and f1-score of 0.81. The accuracy rate is 82%. The confusion matrix reveals that out of 33 countries, the model was able to predict the correct class label of 27 countries. However, it goes without saying that due to the limitation of our dataset in terms of the number of countries, the datapoints for training the model was very limited. With more datapoints, we believe that we can further improve the model. Using the model, we can predict the happiness level of a country based on the features i.e. if we are given a new country, and we have the required international indicators of that country, this model can predict where it stands on the happiness score – high-range, mid-range, or low-range. However, part of our research goal was to evaluate the importance/relevance of the features in determining the happiness of a country, but we did not examine that using the KNN model.

3.2 Neural Network Model

The neural network yielded a R^2 value of 0.511 when evaluated on the test set, as shown in figure 14. The non-linear model should be able to better represent the complex relationship between certain indicators and happiness; however, the R^2 is not a significant improvement from the linear models. The neural network is a data hungry model that ideally should be trained on thousands of samples, so the small nature of the training set limits the robustness of the neural network. Increasing the size of the dataset would most likely yield a significant increase in the R^2 score.

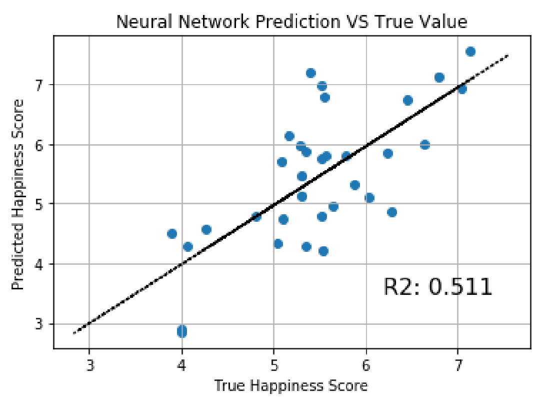

Figure 8: Neural Network Test Set Performance

Important features were selected using LRP from the entire dataset, the results can be seen below in table 15. The LRP relevance score represents the impact the feature had on the model's prediction of happiness, so features with a near-zero relevance score had little impact on the predictions as a whole. This process should help answer which international indicator is most impactful on a nation's happiness. As determined by LRP, the most important features were life expectancy at birth, income per capita, and literacy rate, while household size, human freedom score, and press freedom index were not impactful to the model. Feature reduction was attempted to improve the model results by dropping irrelevant features; however, the technique did not result in better performance in the training set cross validation.

Table 6: Feature LRP relevance scores

Feature	LRP Relevance Score
Life Expectancy at Birth	0.148568
Income Per Capita	0.112832
Literacy Rate	0.109983
Percent Urban	0.085142
School Life Expectancy	0.083921
Deaths Per 1000	-0.066391
Percent Old (>65)	-0.047102
Births Per 1000	-0.028149
Gross enrolment ratio, tertiary	0.022539
Fertility Rate	-0.020714
Percent Young (<15)	-0.020529
Economic Freedom Summary Index	0.019768
Infant Mortality Rate	0.017982
Press Freedom Index	-0.016181
Human Freedom Score	0.012904
Household Size	0.00502

4 Conclusion

We used three different methods to predict happiness from the 16 features of our data set. We used linear models (ordinary least squares regression, lasso regression, ridge regression), a classification method (K-nearest neighbor) and a neural network. The K-nearest neighbor can predict the class label of the countries with a good accuracy, but it cannot determine the importance of a particular feature. The linear regressions show the impact different features have on the models. The income per capita is the most common significant feature, and it is present in all linear models. The neural network yielded comparable conclusions to the linear regression models with its relevance scores for the different features. Our analysis found statistical correlations that match psychological research into human happiness. The leading indicator/cause of happiness is income per capita. Continuing this type of analysis, it would be beneficial to break each country down into smaller pieces such as states/provinces or counties to increase the sample size. Our study included 125 countries, which is small. A combination of related features and the addition of new unique features could also improve the accuracy of the analysis.

References

[1] "Factors Connected to Happiness." *Psychology*, by Rose M. Spielman et al., OpenStax, Rice University, Houston, TX, 2017.

[2] Ortiz-Ospina, Esteban, and Max Roser. "Happiness and Life Satisfaction." *Our World in Data*, 14 May 2013, ourworldindata.org/happiness-and-life-satisfaction.

[3] Chou, Clementine. "World Happiness Analysis." *Stanford*, 3 Nov. 2017, web.stanford.edu/~kjytay/courses/stats32-aut2018/projects/world_happiness_analysis-1.html.

[4] “Data Analysis of World Happiness Report.” *Kaggle*, Kaggle, 9 June 2018, www.kaggle.com/koki25ando/data-analysis-of-world-happiness-report.

[5] Nguyen, XuanKhanh. “Happiness and Life Satisfaction.” *Medium*, Towards Data Science, 9 Aug. 2020, towardsdatascience.com/happiness-and-life-satisfaction-ecdc7d0ab9a5.

[6] Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W (2015) On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLoS ONE 10(7): e0130140. https://doi.org/10.1371/journal.pone.0130140