Predicting World Happiness

CSE 146: Final Report

Maximilian Grieser
University of California, Santa Cruz
Santa Cruz, California
mgrieser@ucsc.edu

Bryan Lopez
University of California, Santa Cruz
Santa Cruz, California
blopez24@ucsc.edu

Matthew Oey
University of California, Santa Cruz
Santa Cruz, California
moey@ucsc.edu

Abstract

The purpose of this final report is to explain and describe the process that has taken place in predicting world happiness. The project is meant to predict a country's happiness score using gross domestic product (GDP) per capita as a feature.

The World Happiness Report is a survey of the state of global happiness that ranks countries by how happy their citizens perceive themselves to be. Factors that play a role in the total happiness are: economy, family, life expectancy, trust, etc. In the report, the economy factor is referred to as the GDP weight. The GDP weight is how much citizens contribute GDP to their happiness. This project only focuses on the GDP weight and total happiness score from the World Happiness Report.

The World Bank contains the GDP per capita of the world for the last 60 years. The project uses this dataset to cross reference the World Happiness Report GDP weight. The datasets gathered are used to train the linear regression model to predict world happiness.

The following pages contain the process of the project from beginning to end, starting with the motivation and objective. It answers questions about how and why we decided on predicting happiness and the process of cleaning, sorting, and matching datasets. This report answers the ethical and societal issues that arose from using the dataset. Which model and algorithms we used are described in detail, followed by results and analysis that came from it. To end the report, we included how each member of the team contributed to the final project and how we foresee the project progressing if work were to continue in the future.

1. Motivation and Objective

When first starting this project, our team was struggling to find data that we could use that also held an ethical perspective. We told ourselves that we wanted to touch on the topic of causality, as we felt that data we were searching on sites like Kaggle could best fit that domain mold. Once we found the data, we had the intention of bringing awareness of the country's economies and how each country contributes their economy to their happiness.

Moreover, our project illustrates the value of money. We asked ourselves foundational questions when first starting off like "How would a person living in one of the richest GDP countries rank the GDP contribution to the calculation of the Happiness Score?". From there, we examined the relationship of money and happiness illustrated through our graphs, which we hope brings light to how our world characterizes their economy differently to their overall happiness.

Another important aspect that should be touched on is the definition of happiness. Happiness is defined, according to vocabulary.com, as the "state of well-being characterized by emotions ranging from contentment to intense joy". Though, it is a very formal definition and somewhat binary way of looking at this emotion.

Happiness through our lens is very narrow as our group filtered out various external factors (life expectancy, freedom, family, government trust, and generosity), so judging and determining happiness with GDP per capita of a country is by far no means the sole factor of happiness. (Our group was originally planning on also looking at the life expectancy of countries and exploring the connection between that and how it also contributes to happiness of countries, but couldn't due to time constraints).

2. Datasets

For this project, we worked with five datasets sourced from two different websites. Four of those five datasets were obtained from the World Happiness Report on kaggle.com, as we used data from the years 2015 to 2018. Each dataset contains happiness data on over 155 countries, with the 2015 dataset being the largest at 158. For each of those countries, there is a corresponding happiness score and ranking that was determined by data from the Gallup World Poll. The scores are based on the individual weights of six factors: GDP per capita, family, life expectancy, freedom, generosity, and trust (government corruption). These weights signify how much a country values each of those factors in their contribution to the aggregate happiness score.

The fifth dataset, GDP per capita, was sourced from the World Bank on the theworldbank.org. This dataset contains GDP per capita data on over 250 countries from the years 1960 to 2019 and is denominated by the current US$. Unfortunately, some data is missing for countries in certain years. Also included within the dataset is the country code for each country.

In order to properly work with our data, many cleaning tasks were carried out. We began by removing all columns that weren't vital to the training of the model for both the happiness and GDP per capita datasets. All columns dating before and after the time period of 2015 to 2018 in the GDP per capita dataset were removed, as well as all columns that were not the country name, happiness score, or weight of GDP per capita from the happiness dataset. Additionally, countries that had inconsistent naming conventions across all datasets were renamed manually within each .csv file. To conclude our cleaning tasks, we removed all countries that didn't appear in all five datasets, leaving us with 131 countries across a four year timespan.

Once data cleaning was completed, we proceeded to perform various preprocessing tasks before the training phase. First, we sorted all datasets alphabetically by ascending country name in order to ensure accurate plotting of the data in our figures. From there, we combined the 2015, 2016, and 2017 columns from the GDP per capita dataset into one pandas dataframe to use as our input vector. However, in order to properly use this input vector for training our model, we had to transform the dataframe into a numpy array and reshape it. This was done because we were only training on one feature vector. We then combined the 2015, 2016, and 2017 happiness score columns from the World Happiness Report dataset into one dataframe for our target labels. To keep the shape consistent with our input vector, we also transformed that dataframe into a numpy array and reshaped it. Finally, we stored test sets for both the feature vector and target values in separate dataframes and reshaped them to be consistent with the shape of the training data.

3. Ethical and Societal Issues

When we were initially discussing the idea of this project, we came to the conclusion that no one dataset or factor would be able to accurately convey the happiness of the world, as everyone has different ideas of happiness. Despite this, we also concluded that in order to train a linear regression model to predict a vague idea such as happiness, we had to quantify it in some way. Now, happiness is not binary, as no feeling is, yet we essentially had to determine binary happiness due to the limitation of our linear regression model.

In order to quantify happiness in some way, we decided to use solely GDP per capita to predict happiness, although we believe that this will prove to be insufficient in the future, as each country, and every individual, values the contribution of GDP per capita to their happiness differently. This raised the age-old question that we sought to answer with our work: does more money equate to greater happiness?

4. Models and Algorithms

To predict a country's total happiness score, we as a team decided to use a Linear Regression Model. We decided on a linear regression model due to the nature of our variables, GDP per capita and total happiness score. The linear regression attempts to model the relationship between the two by fitting a linear equation. We made sure to check the relationship between the two variables. We concluded that GDP per capita is a factor to the GDP weight, which is a factor to the total happiness score; therefore, there is some significant association between the two variables.

The data used to train and test the linear regression model was split 75/25. It was trained on the years 2015, 2016, and 2017 of GDP per capita for its feature and on the years 2015, 2016, 2017 for its total happiness score as its labels. The model was tested on the year 2018. The reason behind this was due to the amount of years we were able to obtain; the World Happiness Report began from 2015 and the World Bank ended at 2018.

After splitting the data, we proceeded to fit the model with the training set and obtain the predicted values of the test set. Next, we moved onto comparing the predicted values with the correct values: we obtained a 100% error rate. The reason behind this was due to the predicted values and correct values being non-binary numbers. Even if they were off by 0.01, it was considered incorrect. To solve the issue, we had to come up with a solution to convert the values into something more simple, like zeros and ones based on some threshold.

Using panda's describe function we obtained the mean for the total happiness score of each year. The threshold we applied was based on the sum mean for all four years divided by four. We decided that any country that had a total happiness score greater than or equal to the threshold would be considered a happy country and would return 1. The countries below the threshold would be considered unhappier countries and return 0. We proceeded to make a method that applied the thresholds to the predicted values and actual test values.

Comparing the new binary values we obtained with the threshold, we received a 77.1% accuracy rate. This meant that we were able to predict for the most part whether a country would be considered happy or unhappier.

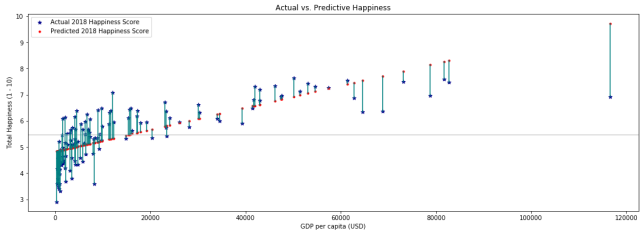

Fig. 4.1. Dots above the horizontal line (threshold) are considered happier countries. Anything below is considered an unhappier country.

Figure 4.1 displays the original predicted (o) and actual total happiness score values (*) from the test set. The teal vertical line connects them to represent the margin of error with its length. The threshold line separates them from a happier country to an unhappier country. In the end, we finished using the model to predict whether it would be a happier one or an unhappier one, rather than relying solely on the total happiness score.

5. Results and Analysis

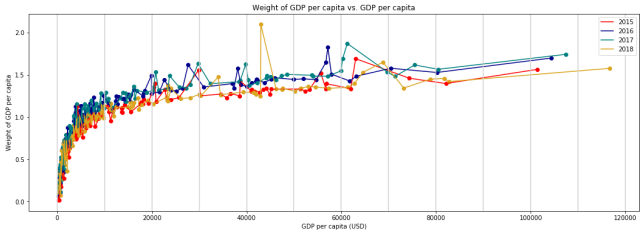

Fig. 5.1. GDP per capita (how much each country has valued their GDP to their happiness) versus GDP per capita (sorted by gdp)

Figure 5.1 portrays the correlation between the weight of GDP per capita (how much each country has valued their GDP to their happiness) versus GDP per capita (sorted by gdp). Through 2015-2018, there is an overall trend of increase in the weight of the GDP as GDP per capita increases. Though with this increase, the weight of GDP does seem to flatten out as GDP reaches a certain value - around the 70k-80k GDP score. Moreover, this flattening out can mean that there is some GDP limit that has country's valuing their GDP to their happiness. For example, Luxembourg had a GDP of ~107,000 in 2017, but they only valued their GDP weight at around 1.74 while Qatar had a GDP of ~61,000 but they valued their GDP weight at 1.87, the highest out of all the countries in 2017.

A novelty found was that the United Arab Emirates (UAE) values their GDP weight to their happiness the highest out all countries in 2018 (with a score of just above 2 compared to the rest of the countries where they had barely hit the 1.5 mark), but they did not have the highest GDP. In fact, the UAE had less than half of Luxembourg' GDP, the country with the highest GDP in 2018, with a GDP of ~116,000.

So in conclusion for figure 5.1, the GDP weight and GDP per capita increase until reaching $80,000, in which then GDP weight decreases. And in 2018 United Arab Emirates valued their GDP weight the highest although having half of the richest GDP per capita country.

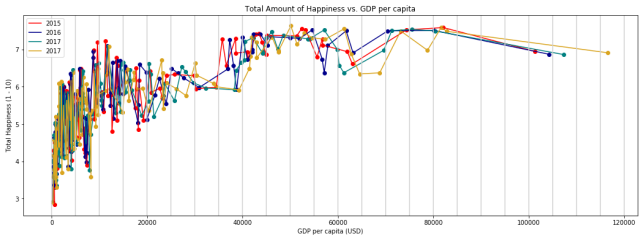

Fig. 5.2. Total amount of happiness (on a scale from 1-10) vs. GDP per capita (sorted by GDP)

Figure 5.2 illustrates the relationship between the total amount of happiness (on a scale from 1-10) vs. GDP per capita (sorted by GDP). Between the 70k - 80k GDP per capita, the total happiness score seems to flatten and begin to fall after 80k. Between the 0k - 20k GDP per capita, the total happiness differs greatly. In more detail, on a scale from 1 - 10, few countries with a GDP per capita of 10k seem to be very happy with a score of about 7. At the same times countries with a GDP per capita less than 10k differ greatly (3-6). In addition, after reaching a GDP per capita of 30k, the total happiness score is no lower 5.5 and after reaching a GDP per capita of 55k, the total happiness score is no lower 6.25.

We also noticed that countries that surpass a GDP per capita of 80k see a decrease of total happiness score. Countries with a GDP per capita of 80k have a total happiness score 7.5 (the highest scores) while countries with a GDP per capita over 80k have a lower total happiness score.

So in conclusion for figure 5.2, countries with a GDP per capita lower than $20,000 have significantly different perspectives on how they see GDP weight, with the lowest being slightly below 3 and the highest being above 7. These fluctuations could stem from a variety of external factors that were not taken into consideration during our data (mentioned later).

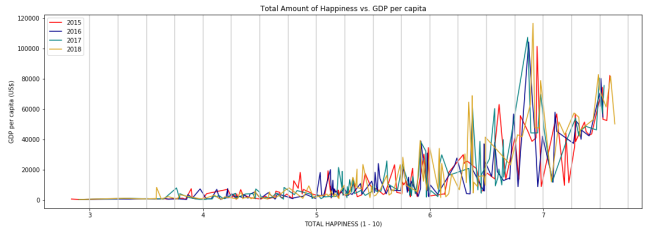

Fig. 5.3. Total amount of happiness vs. GDP per capita (sorted by total happiness)

Figure 5.3 illustrates the relationship between the total amount of happiness vs. GDP per capita (sorted by total happiness). Most countries with a GDP per capita lower than 10k, tend to have a total happiness score lower than 5 (1-10 scale). Countries with total happiness scores in the range (5-7) seem to significantly differ in GDP per capita. For example, 5 - 6 total happiness scores have a GDP per capita from ~5k to 40k, 6 - 6.5 total happiness scores have a GDP per capita from ~5k to 70k, 6.5 - 7 total happiness scores have a GDP per capita from ~10k to 120k. Also interestingly, countries with the highest happiness scores (+7) aren't the ones with the highest GDP per capita, but aren't the poorest either. Countries had a minimum GDP of 10k and maximum of 80k and the happiest country of 2018 had a GDP per capita of 50k with a total happiness score (+7.5)

When comparing figure 5.1 and figure 5.2, we see in both graphs that around the GDP scores of around 70k - 80k that there is a slight decrease in weight of GDP and total happiness. This can mean that while a higher GDP can contribute some value to the country's happiness, it does not necessarily mean that this pattern will continue. As an example of that, Luxembourg had a GDP of ~107,000 in 2017, but they only valued their GDP weight at around 1.74 and was the 19th most happiest country (along with other factors) while Qatar had a GDP of ~61,000 but they valued their GDP weight at 1.87, the highest out of all the countries in 2017 and was the 36th most happiest country.

Adding to the comparisons, there is an overall increase from 2015-2018 in both weight of GDP per capita and total amount of happiness as GDP per capita increases. At the beginning, there are some fluctuations of countries in terms of their happiness and weight of GDP within the 0 - 20k GDP range. The poorer countries in terms of their GDP value their happiness and weight their GDP differently due to some factors that are beyond the scope of what our data measures - it could be cultural differences, lifestyle, government dynamics, etc that can play a role into these fluctuations.

This can also beg the question: Does more money mean more happiness? According to our data and graph, not necessarily. Though, there are some other external aspects of the data we modified. We did have to filter out some countries that were not consistent within our data from both all the happiness reports and GDP reports. We also created a country filter to see common countries within happiness reports from 2015-2018. We then matched that filter with the GDP filter to make sure all countries were named/assigned correctly with all valid values (some countries could have missing GDP per capita data). We also only focused on GDP and how it relates to happiness rather than GDP and other factors (life expectancy, freedom, government trust, and generosity were among the other factors used in the World Happiness Report that our team did not have time to go through).

6. Contributions

We as a group found the datasets to work with and came up with the objective to predict world happiness. We met weekly to discuss the project and plan the work that needed to be done. We each played a role in looking through the GDP dataset to match the countries with the happiness dataset.

For the project, Bryan initially began by working on cleaning the world happiness dataset by removing the unused column, but later on, Matthew helped adjust the datasets. Afterwards Bryan processed the datasets to make them have the same number of matching countries and sorted the datasets by GDP per capita. Bryan plotted Fig-5.2 and Fig-5.3. The last item Bryan worked on was fixing the training model and coming up with the threshold. At the end of the project, Bryan went back to the code to clean and leave comments for a better description. For the video, Bryan read outloud his part of the script and for the final report, he wrote the Abstract and Models/Algorithms section.

During the beginning of this project, Matthew worked on mounting the Google Drive to Google Colab since we had issues importing data across all devices in Colab (Bryan could import the CSV files since he uploaded the data on his end, but neither Max nor Matthew could import). Matthew and Bryan continued to work on cleaning the world happiness dataset by importing the new 2015-2018 happiness dataset, and cleaning all happiness datasets by removing unnecessary columns to have consistent column values (Country, Happiness Score, GDP/country). Matthew also manually scrubbed through the happiness report csv's from 2015-2018 to reflect the appropriate country filter that we wanted.

Matthew then plotted data in a 3D graph for weight of GDP, total happiness, and GDP per capita. Matthew helped draw conclusions for figure 5.1 as well make general conclusions from all the graphs combined and high-level conclusions from every single graph. In addition, Matthew helped Bryan create the linear regression model by using his template/test model to predict our 2018 data by setting a new threshold to match the mean of all 4 years and then plotting the predictive and actual data with the new threshold.

Lastly, Matthew worked on editing the audio and visual portions of the video. Due to his prior experience with video editing, Matthew took the lead to create the video. The team created the script and made sure it was informative and Matthew edited the video in a way to make it engaging (since it was us speaking as the main audio) through the use of stock footage and showing samples of our code for context, modifying the poster to highlight key parts, and text blocks and background music to aid the pacing of the video.

Lastly, Matthew wrote the Motivation and Objective and Results and Analysis sections.

Throughout the project, Max assisted Bryan and Matthew in cleaning both datasets, removing unnecessary or duplicate data. Once data cleaning was complete, Max plotted figure 5.1, helping Matthew draw conclusions from it as well. After summarizing the findings, Max proceeded to fit the initial linear regression model with the preprocessed data provided by Bryan. Once fitted, Max tested the model with the other portion of the preprocessed data and analyzed the accuracy of the model with a basic error function derived from standard error. Because this model was returning a 100% error rate, Max handed training and testing of the model off to Bryan, who was able to derive a new means of analyzing the accuracy of the model, utilizing a binary threshold. For the poster presentation, Max helped Matthew record for the final video, reading aloud his part of the script and providing feedback during the final editing process. Finally, Max wrote the Datasets, Ethical and Societal Implications, and Future Work sections of the final project report.

7. Future Work

There is much to be done in the foreseeable future for this project. For starters, it would be prudent to develop a model that was able to predict a range of happiness values, rather than just a binary happy or unhappy value. In addition to this, we believe that modeling the other factors included in the calculation of the happiness score both individually and in conjunction with one another would prove to be fruitful, although finding the datasets to accurately model those factors may prove to be challenging. To expand on this idea further, we took inspiration from one of our lab assignments for this class, which had us determine a pair of thresholds that minimized "fairness error." By determining which set of factors to focus on globally, or by country, one could design a model to improve the happiness of the most people possible. This, however, may come into conflict with game theory, which would be a useful application to explore in the pursuit of happiness.