The Health of Nations

Jordan Levy
jdlevy@ucsd.edu
University of California, San Diego

Shiyin Liang
ssliang@ucsd.edu
University of California, San Diego

Website Link: https://jordanlevy99.github.io/DataHacks_2021/Intro_Page.html
GitHub Link: https://github.com/JordanLevy99/DataHacks_2021

Introduction to the Data

Prosperity is defined as the condition of being successful or thriving. But can prosperity be quantified? Which countries would be the most prosperous then? How would the prosperities of countries change over time?

Thanks to Legatum Institute, an organization that compiles an annual prosperity ranking of every country, prosperity can indeed be quantified. In Legatum's data, prosperity scores are calculated from 9 pillars: Economic Quality, Business Environment, Natural Environment, Governance, Education, Health, Safety and Security, Personal Freedom, and Social Capital. In turn, the scores for the pillars are determined by their subcategories. For our project, we can use Legatum's data to determine the most prosperous countries as well as how prosperity scores can change over time. Furthermore, we can utilize the data collected over 2007 to 2014 of the prosperity scores and prosperity ranks of 149 countries to predict a prosperity score and rank for each of these countries in 2015 and 2016.

Data Cleaning Method

The datasets we received were already pretty clean so there were only a few tasks that needed to be done to prepare the datasets for use. First, we omitted the columns where all the values were "***" since we were not granted permission to use them. Secondly, while there were some missing values in the form of "--" in the datasets, these missing values only happened in the columns listing the years in which the subcategories of each pillar was collected. The years which the subcategories of each pillar was collected is inconsequential to our project so we opted to just omit these columns instead of imputing them. Each dataset corresponded to a pillar and was separated further into training and test data. Since each dataset contained records for the same countries and the same years, we were able to simply merge all the training datasets into one dataset where each row now contains information data for each of the 9 pillars and their subcategories for each country of a given year without any further wrangling .

Trends

For visualizations of trends, please go to:
https://jordanlevy99.github.io/DataHacks_2021/Intro_Page.html

Results & Analysis for Prompts

Determining Top 5 Countries with the Most Growth in Prosperity Overall

In order to determine which countries grew the most in terms of prosperity from 2007 to 2014, we look at compound annual growth rate, or CAGR. Typically used in banking/investing, compound annual growth rate can be used to compute how much a value grows on average year to year. The equation is as follows:

$$\mathrm{CAGR} = \left(\frac{V_{\mathrm{final}}}{V_{\mathrm{begin}}} \right)^{1/t} - 1$$

CAGR = compound annual growth rate
V_{begin} = beginning value
V_{final} = final value
t = time in years

Figure 1: *Formula for Compound annual growth rate (CAGR)*

Using this equation and with the values for 2007 prosperity and 2014 prosperity as V_{begin} and V_{final}, we get the following countries and their respective CAGR values as the top growth countries from this time period:

Country	**Compound Annual Growth Rate (CAGR)**
Chad	0.019655
Togo	0.019244
Zimbabwe	0.017231
Ivory Coast	0.013437
Georgia	0.012872

Table 1: *Top Five Growing Countries and CAGR Values*

Determining Pillars with the Most Impact on Prosperity

Looking at the pillars of each of the top 5 growing countries, we see the following pillars have the most impact on prosperity. Here, impact is defined as the correlation between pillar value and prosperity value for the years 2007 through 2014.

Country	Pillar with Most Impact	Correlation to Prosperity
Chad	Health	0.9318
Togo	Natural Environment	0.9228
Zimbabwe	Economic Quality	0.9947
Ivory Coast	Governance	0.9483
Georgia	Governance	0.9718

Table 2: *Pillars most correlated to Prosperity for Top Five Growing Countries (2007-2014)*

Machine Learning and Feature Importance

Each of the 9 pillars represent how each sector of society is functioning for a country. We want to understand how each pillar gets its score, so that we can better understand how prosperity of a country is calculated. With the data we are given, we are able to predict each pillar's score with fairly high certainty. For example, in the plot below, we are able to predict Governance values that have 0.9901 r-squared correlation with the true Governance values.

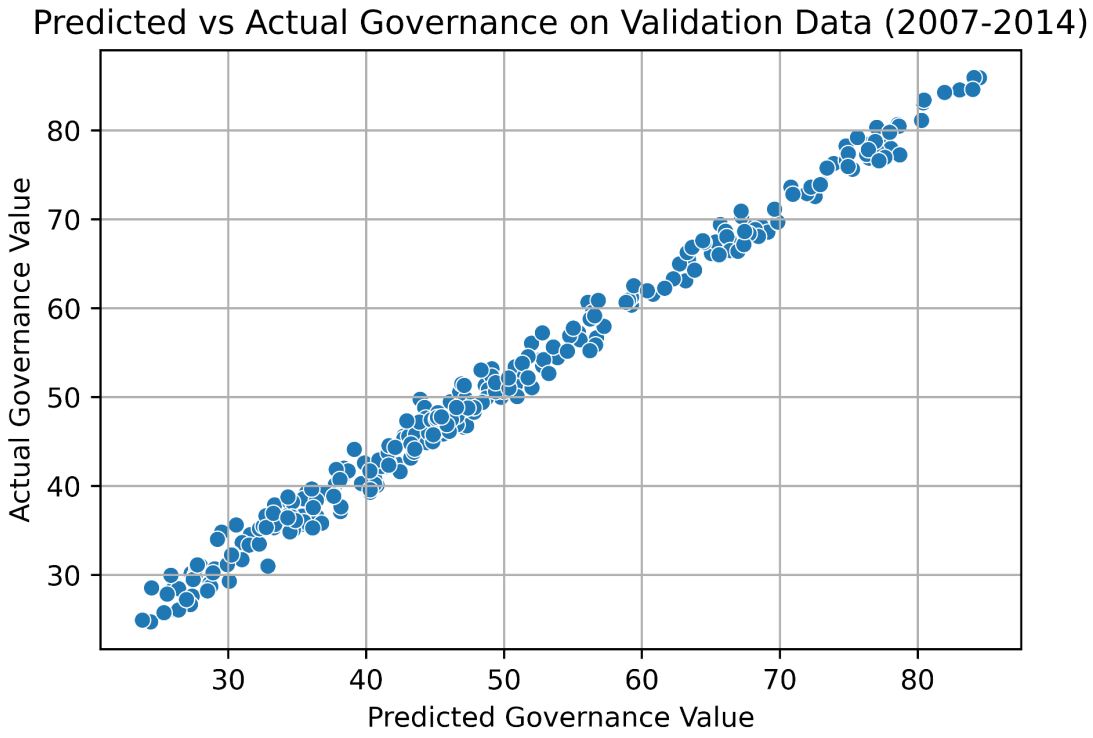

Figure 2: *Predicted vs Actual Governance Values Plot (on validation set)*

We have two main predictive tasks that we set out to accomplish: predict the values/ranks for each pillar, and predict the overall prosperity value/rank for each country for 2015 and 2016 using the 2007-2014 data. In practice, our first predictive task led us directly to our answer for the second, as overall prosperity is simply an average of all the pillar values for a given country/year. Therefore, all we need to do is predict the value of each pillar, then compute the predicted prosperity value accordingly.

To prepare our data to be entered into a machine learning model was a fairly straightforward process. We first select all quantitative, continuous variables; that is, our sub-categories within a given pillar. We then standardize our data column-wise, so that the different scales of variables don't affect the output, and the model parameters can have similar magnitude across columns. Lastly, we partition our data into training and validation splits with a 0.75, 0.25 split, respectively.

After preparing our data, we employ the following three machine learning models on our training data: Lasso Regression, ElasticNet Regression, and Support Vector Regression (SVR). Each of these models carry their own set of assumptions, but all perform reasonably well on the dataset. Lasso regression is Linear Regression with L1 penalty, where alpha is a regularization constant. The higher alpha is, the higher the L1 penalty (absolute loss) is, which encourages the model to shrink it's parameter coefficients or eliminate certain parameters entirely. ElasticNet is a more robust version of Lasso regression that combines L1 and L2 penalty together and should in theory perform at least as good as Lasso regression with proper hyperparameter tuning. Lastly, we use Support Vector Regression, which is the regression form of Support Vector Machine, a classification algorithm that aims to separate different classes as much as possible by introducing a slack regularization penalty. The regression form gets the raw output of our Wx+b to output a continuous value.

All three models perform exceedingly well on our validation data, as shown in the table below.

	Busi	Econ	Educ	Envi	Gove	Heal	Pers	Safe	Soci
Lasso	0.9743	0.9786	0.9809	**0.9657***	0.9901	0.9565	**0.9533***	0.9017	0.2239
Elastic Net	**0.9743***	0.9786	**0.9876**	0.9657	**0.9901***	0.9565	0.9533	0.9017	0.2239
SVR	0.9551	**0.9789**	0.9809	0.9444	0.9848	**0.9591**	0.9522	**0.9034**	**0.3200**

Table 3: *R-Squared Cross-Validation Values of 3 Models Tested*
* ties are randomly broken

This table gives the average r-squared score across 5-fold cross-validation. We optimize our prediction of pillar values by performing hyperparameter tuning on the three models listed above, as well as only selecting the models that produce the highest scores through a dictionary of pillar name and optimized model.

Hyperparameter Tuning

To perform hyperparameter tuning, we simply do a grid search over a variety of hyperparameters. For Lasso regression, we look at 10 values of alpha equally spaced between 0.1 and 2, inclusively. For ElasticNet, we look at the same values of alpha, and in addition, look at 10 values of L1 ratio equally spaced between 0 and 1. This allows us to incorporate various amounts of L1 and L2 penalties in order to create a more robust model. Lastly, for SVR, we are looking at 'rbf' and 'poly' kernel, as well as the following values for our slack coefficient C: 0.1, 0.25, 0.5, 1, 2. The following image shows our optimized model dictionary once we finish our search of hyperparameters and model selection based on average cross-validation score:

```
{'busi': ElasticNet(alpha=0.1, l1_ratio=1.0),
 'econ': SVR(C=2),
 'educ': ElasticNet(alpha=0.1, l1_ratio=0.8888888888888888),
 'envi': Lasso(alpha=0.1),
 'gove': Lasso(alpha=0.1),
 'heal': SVR(C=2),
 'pers': ElasticNet(alpha=0.1, l1_ratio=1.0),
 'safe': SVR(C=2),
 'soci': SVR(C=1)}
```

Figure 3: *Pillar name and Corresponding Optimal Model and Hyperparameters*

As you can see, our model picks a variety of hyperparameters and different models, suggesting that it was worth searching through different hyperparameters and models to optimize our predictions. Our predictions can be found in our GitHub repository under 'Processed_Data/predictions_ensemble.csv'.

SHAP Model Explainer

Now that we've predicted values for pillars and overall prosperity, we can get to the root of our question: what makes a country prosperous? To help answer this question, we use the model explainer SHAP, which essentially tells us the marginal contribution of each feature to the model's output, and whether that contribution had a positive or negative impact. The following three plots show examples of our feature importance values for 3 of our 9 pillars. The rest of the plots can be found on the GitHub repository under 'Figures/shap_plots'.

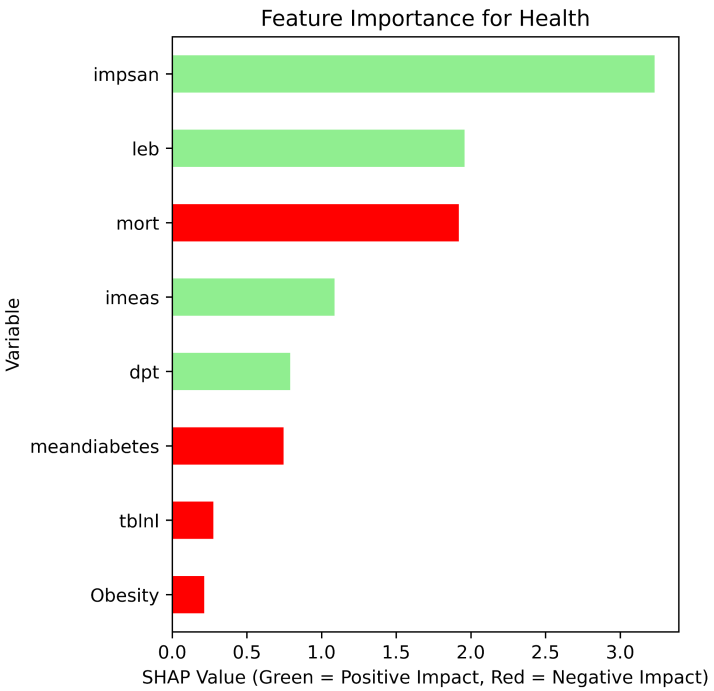

Figure 3: *Most Important Features for Health*

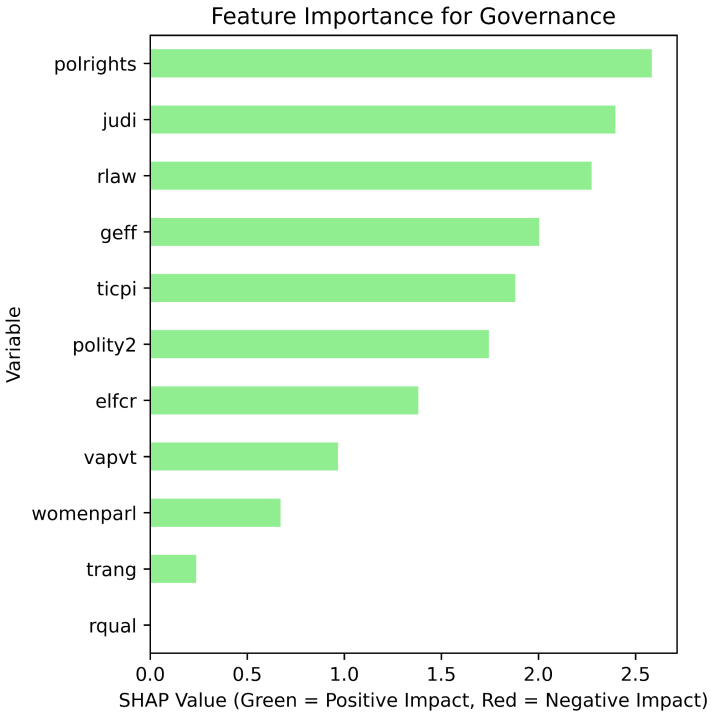

Figure 4: *Most Important Features for Governance*

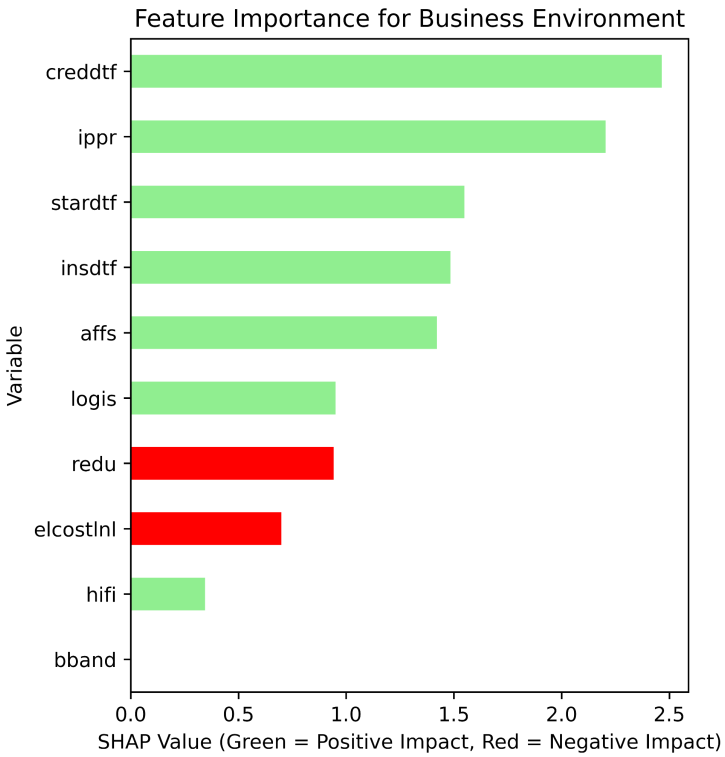

Figure 5: *Most Important Features for Business*

In Figures 3, 4, and 5 above, we get a glimpse at how important each of the features is for 3 of our 9 pillars. While we're not entirely sure what some of these features mean, there are some interesting features such as 'mort' (mortality), 'meandiabetes', and 'Obesity' that all have a negative impact on the output for the Health score, which intuitively makes sense. We can see a variety of other attributes that are relevant to each pillar's score, and thus, overall prosperity as well.

Conclusion

To summarize some of our more important findings, the top 5 countries with the highest growth rate in prosperity are Chad, Togo, Zimbabwe, Ivory Coast, and Georgia. Between the Ivory Coast and Georgia, governance was the key driver in prosperity growth. In Zimbabwe, economic quality was the main influence in prosperity growth. In Togo and Chad, the natural environment and health respectively was the main influence.

We were able to predict with fairly high certainty the values for each of the pillars, as we consistently got above 0.9 r-squared cross-validation score for every pillar except 'Soci' which only provides one feature. This score was optimized through k-folds cross-validation, grid search hyperparameter tuning, and model selection between Lasso Regression, ElasticNet Regression, and Support Vector Regression. We have forecasted the values and ranks for both pillars and overall prosperity, which can be found on our GitHub.

In addition to predicting the values and ranks of pillars and overall prosperity, we are able to determine the most important categories that make up each pillar. We do this with our SHAP model explainer, which assesses marginal contributions of each category to figure out which are most significant to the model's output.

For future work, we would like to know what the subcategories of each pillar refers to so that our work for finding the most impactful subcategories for each pillar can be interpreted more meaningfully. In terms of modelling approaches, we could have utilized more time series approaches, such as rolling average or assigning different weights on past data to account for time decay. However, these approaches can be a bit limited because the granularity of our data is year to year, and we are only looking at 8 individual years of data.

Overall, we can use the information acquired from this project to gain a better understanding of the world around us, and the issues we have faced in the past, the present, and will face in the future. We see this as a starting point for gaining insight into how the world operates, and will ultimately need to dive deeper into the data and see which problems are at the forefront of our society.