An Analysis of What Variables Contribute to a Country's GDP

Report

Student Name: Mondli Gina

Student ID: 0424012

COSC-4426EL Data Mining

Supervised by: Mr Sujay Kalakala

Bharti School of Engineering and Computer Science, Laurentian University

Table of Contents

Abstract 1

Introduction 1

Methodology 2

- **Datasets** 2
- **Exploratory Data Analysis (EDA)** 2

Results 2

- **Correlation Matrix** 2
- **Heatmap** 3
- **Logistic Regression** 3
  - **GDP** 3
  - **HDI** 4
  - **High Population Growth** 4
- **XGBoost** 5
  - **XGBoost Regressor** 5
- **Mean Squared Error** 6

Discussion 7

Conclusion 8

Abstract

This report sought to determine the relationship between GDP, HDI, GDI, population size and MYS. The report demonstrates the difficulties in determining GDP and offers potential explanations for this. Correlations between GDP and other metrics were also weaker than hypothesized. Interestingly, composite metrics such as HDI and GDI had stronger correlations with each other and predictive models for these metrics were also much more accurate. In summary, the report reaffirmed the literature by demonstrating that there are intricate causal factors behind a country's GDP. Future research may focus on the significance of GDP in quality of life instead.

Introduction

Many factors go into measuring a country's economic prosperity or the quality of life that its citizens experience. The number of features that can be considered when measuring each metric makes the task highly involved and potentially subjective. This is the reason why economists usually focus on a single metric or a small number of metrics when assessing the state of a country. The most popular of these metrics is Gross Domestic Product (GDP). This is a standardized metric that measures the total value of goods and services that are produced by a country within a given year. This is a fiscal metric, meant to be more comprehensive than similar micro metrics like Gross National Product (GNP) and Gross National Income (GNI). Human progress and human development are typically measured using the composite metric Human Development Index (HDI). HDI considers life expectancy at birth, Mean Years of Schooling (MYS) and income per capita. Economists and social scientists may also be interested in a country's level of unemployment. The immediate assumption is that unemployment is negatively correlated with GDP. One might think that low unemployment levels mean higher productivity and ultimately a higher GDP. As this report demonstrates, this is only usually true.

GDP, although an important metric, does not measure every aspect of a country's performance. HDI, GDI and population levels more specifically are more relevant to individual citizens. Having a high-value output and exporting many goods to the rest of the world may benefit a government specifically, but the revenue may not benefit a country's citizens proportionally. Still, this is a question that this report seeks to answer. I hypothesize that HDI is strongly correlated with GDP. The former may even be a causal factor in GDP. Population size may be a causal factor in GDP and GDP growth as well. A growing population at least increases the potential for a greater GDP output due to a potentially increasing workforce. The report aims to ascertain the relationships between each of the aforementioned metrics and what balance of metric figures yields the highest GDP.

Methodology

Datasets

I retrieved the world_population.csv dataset from Kaggle.com. The dataset contained the populations of each country in every year from 1960 – 2022. I also retrieved the dataset GDP_Per_Country.csv which contained the GDP of every country in the world in every year from 1960 – 2020 from Kaggle.com. Also from Kaggle.com, I retrieved the Unemployment_Analysis.csv and Human_Development_Index.csv datasets. The Human_Development_Index.csv dataset contained information on the other metrics of interest that were used in this report, namely, GDI and MYS.

Exploratory Data Analysis (EDA)

The dataframes were cleaned, melted and merged to form a single super dataframe to facilitate the EDA. Once the super dataframe was prepared, various techniques were used to gather insights into the relationships between HDI, GDI, MYS, Population size and GDP. The methods and their purposes are listed below.

Measures of Central Tendency: Mean, median. Applied to GDP and population

Measures of Dispersion: Q1, median, Q3, IQR. Applied to GDP and population

Correlations: Pearson Correlation with correlation matrices. Applied to GDP and population, GDP and GDI, GDP and HDI, GDP and MYS, HDI and population, HDI and MYS. A heatmap was also used to visualize these correlations.

Predictions: Logistic Regression was applied to GDP, HDI and Population Growth Rate. XGBoost was applied to HDI and GDP. An XGBoost Regressor was applied to GDP.

Primary Target Attribute: GDP

Python Libraries: Pandas, Numpy, Seaborn, Matplotlib, Statsmodels, Scikit-learn

The entire data analytics project was done in Jupyter Notebook. Jupyter Notebook is a web-based computing platform.

Results

Correlation Matrix

	Year	HDI	GDI	MYS	Population	GDP
Year	1.000	0.180	0.190	0.190	0.019	0.065
HDI	0.180	1.000	0.680	0.900	-0.029	0.230
GDI	0.190	0.680	1.000	0.670	-0.090	0.120
MYS	0.190	0.900	0.670	1.000	-0.053	0.210
Population	0.019	-0.029	-0.090	-0.053	1.000	0.460
GDP	0.065	0.230	0.120	0.210	0.460	1.000

Heatmap

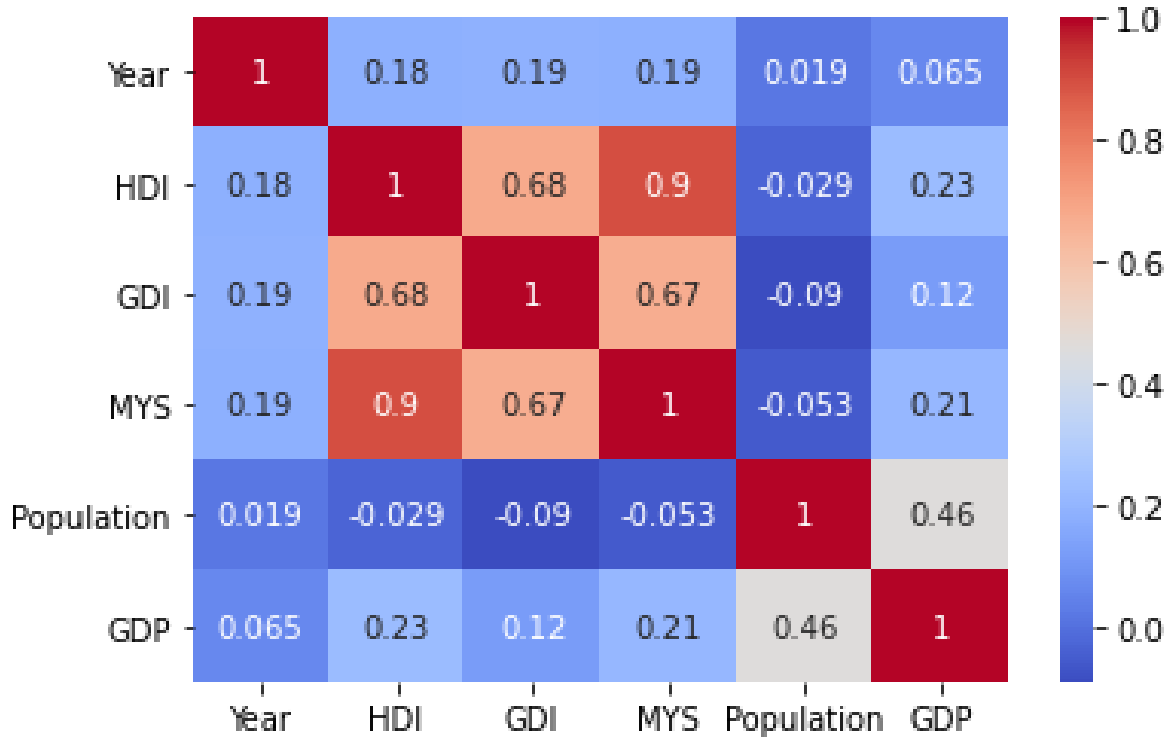

Figure 1. Heatmap of correlations between country variables.

Logistic Regression

GDP

```
Optimization terminated successfully.
         Current function value: 0.292927
         Iterations 11
                           Logit Regression Results
==============================================================================
Dep. Variable:                High_GDP   No. Observations:                 4095
Model:                          Logit   Df Residuals:                     4090
Method:                           MLE   Df Model:                            4
Date:                Tue, 19 Nov 2024   Pseudo R-squ.:                  0.5774
Time:                        20:56:30   Log-Likelihood:                -1199.5
converged:                       True   LL-Null:                       -2838.4
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const        -13.2265      0.974    -13.585      0.000     -15.135     -11.318
HDI           21.3410      0.944     22.599      0.000      19.490      23.192
GDI           -2.3824      1.024     -2.327      0.020      -4.389      -0.376
MYS           -0.2017      0.036     -5.566      0.000      -0.273      -0.131
Population  1.982e-07   8.07e-09     24.556      0.000    1.82e-07    2.14e-07
==============================================================================

Possibly complete quasi-separation: A fraction 0.11 of observations can be
perfectly predicted. This might indicate that there is complete
quasi-separation. In this case some parameters will not be identified.
```

Figure 2: Logistic Regression to Determine Probability that Country Has High GDP

HDI

```
Optimization terminated successfully.
         Current function value: 0.262791
         Iterations 10
                           Logit Regression Results
==============================================================================
Dep. Variable:               High_HDI   No. Observations:                 4095
Model:                          Logit   Df Residuals:                     4090
Method:                           MLE   Df Model:                            4
Date:                Tue, 19 Nov 2024   Pseudo R-squ.:                  0.6199
Time:                        20:56:29   Log-Likelihood:                -1076.1
converged:                       True   LL-Null:                       -2831.3
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const        -15.4024      1.189    -12.958      0.000     -17.732     -13.073
GDI            9.2006      1.278      7.197      0.000       6.695      11.706
MYS            0.8396      0.033     25.211      0.000       0.774       0.905
Population -4.168e-08   3.81e-09    -10.938      0.000   -4.91e-08   -3.42e-08
GDP         8.571e-12   7.24e-13     11.831      0.000    7.15e-12    9.99e-12
==============================================================================
```

Figure 3: Logistic Regression to Determine Probability that Country Has High HDI

High Population Growth

```
Optimization terminated successfully.
         Current function value: 0.619431
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:     High_Population_Growth   No. Observations:                 4095
Model:                              Logit   Df Residuals:                     4090
Method:                               MLE   Df Model:                            4
Date:                    Tue, 19 Nov 2024   Pseudo R-squ.:                  0.1064
Time:                            20:56:33   Log-Likelihood:                -2536.6
converged:                           True   LL-Null:                       -2838.4
Covariance Type:                nonrobust   LLR p-value:                 2.405e-129
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.1731      0.511      2.296      0.022       0.172       2.175
HDI           -2.2049      0.512     -4.304      0.000      -3.209      -1.201
GDI           -0.0400      0.646     -0.062      0.951      -1.307       1.227
MYS           -0.0103      0.024     -0.430      0.667      -0.057       0.037
GDP         3.273e-12   2.31e-13     14.145      0.000    2.82e-12    3.73e-12
==============================================================================
```

Figure 4: Logistic Regression to Determine Probability that Country will have a High Population Growth

XGBoost

```
[[547  16]
 [ 20 646]]
              precision    recall  f1-score   support

           0       0.96      0.97      0.97       563
           1       0.98      0.97      0.97       666

    accuracy                           0.97      1229
   macro avg       0.97      0.97      0.97      1229
weighted avg       0.97      0.97      0.97      1229
```

Figure 5: Trained XGBoost model to predict whether a country will have a high HDI or not.

XGBoost Regressor

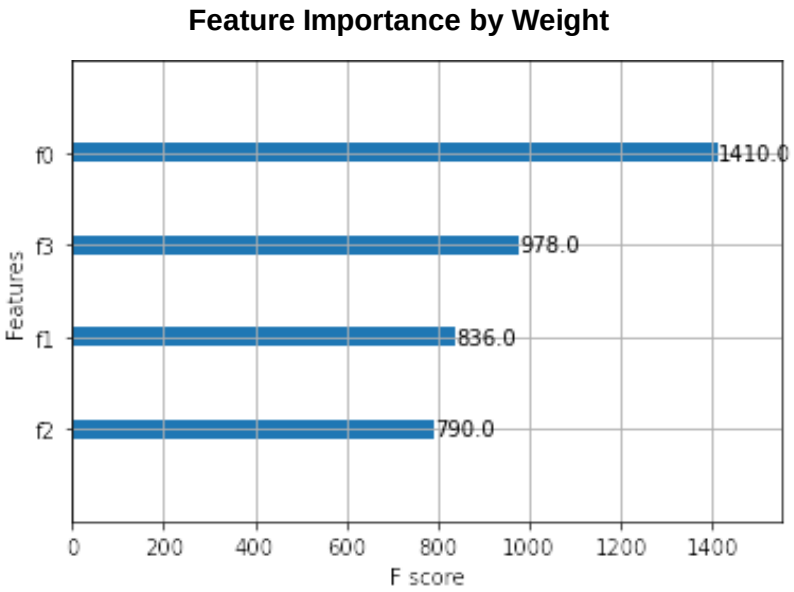

Figure 6: XGBoost Regressor plotting feature importance by weight

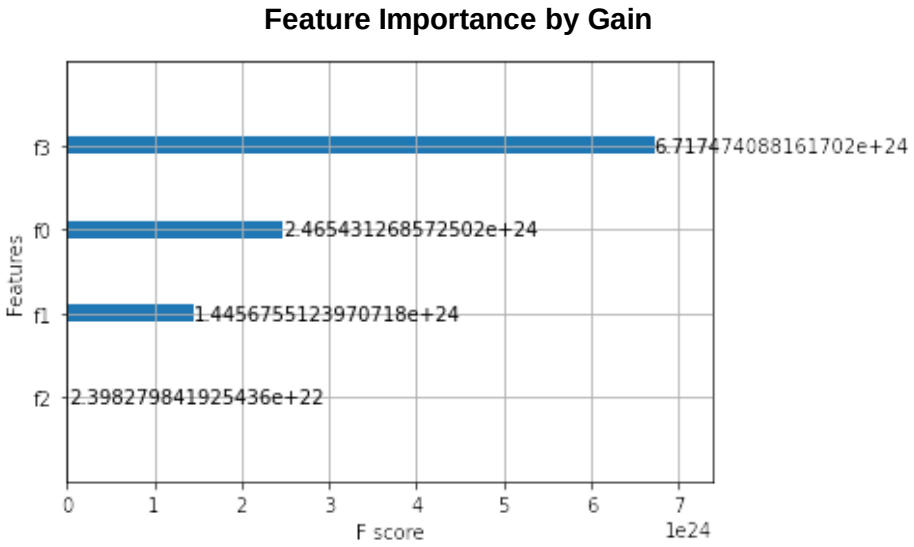

Figure 7: XGBoost Regressor plotting feature importance by gain

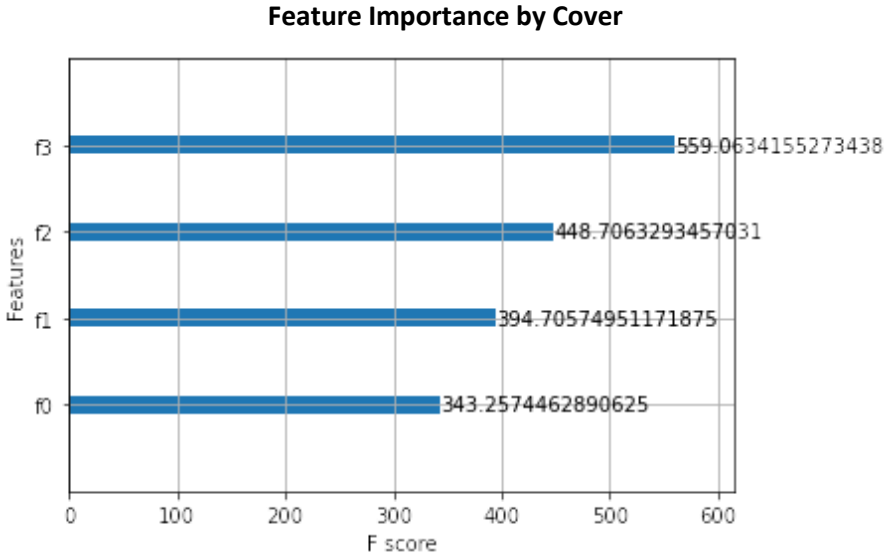

Figure 8: XGBoost Regressor plotting feature importance by cover

Mean Squared Error

Mean Squared Error (MSE) for GDP using XGBoost: $1.558316685153844 \times 10^{22}$

Discussion

This paper demonstrates the different relationships interplaying between GDP, HDI, GDI, MYS and population size. These metrics all measure some aspect of a country's condition. The HDI is measured by calculating a country's life expectancy at birth, mean years of schooling for adults, expected years of schooling for children and Gross National Income (GNI) per capita. GDI is the ratio of a country's female HDI to male HDI. The report found that all measures of GDP central tendency varied enormously between countries and continents. These discrepancies in GDP were apparent even between countries with similar population sizes and population growth rates. This finding further supports the suggestions made by the XGBoost model – that population size cannot be used to predict GDP.

The logistic regression model had a current function value of 0.292927. This is encouraging as generally, lower values indicate better model fit. Another measure that indicated a reasonably good fit was the Pseudo R-squared value of 0.5774. The model suggests that HDI and Population size have a significant effect on GDP. The likelihood of a country having a high GDP appears to increase as these variables increase. Conversely, both GDI and MYS seem to be negatively related to GDP. As each of these variables increase, the likelihood of a country having a high GDP decreases. These negatively correlated relationships are weaker than the positive ones between GDP, HDI, and GDI, but they should not be ignored. Factors such as a cheap labor force and low adherence to human rights may explain the negative relationship.

An XGBoost regressor model was ran to determine the likelihood of a country having a high GDP. The model yielded a mean squared error (MSE) of 1.558316685153844e+22. This extremely high MSE value may have been due to improper feature selection or suboptimal hyperparameters. Another likely reason the GDP XGBoost learning model made such poor predictions is the low correlation values between GDP and other variables. Correlations, or lack thereof, will reflect in the models performance. HDI is a more straightforward prediction than GDP. As the results show, the two strongest correlations between different variables involve HDI. HDI and MYS has a correlation of 0.9 and HDI and GDI have a correlation of 0.68. These are considered strong and moderately strong positive correlations respectively. GDI and MYS also had a moderately strong positive correlation of 0.67.

An interesting and relevant detail about the correlation between HDI and GDI is that these are manufactured metrics. That is to say, their definitions and the features that contribute to them have been determined by people and institutions that are run by people. It is difficult to consider these composite metrics as dependent variables in any study because they are not determined by unknown, organic variables. One can easily manipulate the HDI or GDI of a country by simply redefining what features contribute to these metrics and how significantly. As described in the first paragraph of this discussion, HDI and GDI both have education and human potential as contributing features. This common ground means that there is bound to be some correlation between the two, because a feature that affects one is likely to affect the other. Education alone is a significant contributor to both indices, even though its exact percentage contribution will vary from year to year and country to country.

Conclusion

GDP proved to be difficult to predict using the other variables in the dataset. This finding is consistent with the literature. To date, there is no single variable or even combination of variables that shows a strong correlation with GDP. Interestingly, however, HDI was much more predictable. Future studies and reports may focus on HDI as a target attribute and even question the significance of GDP on a population's prosperity.