Analyzing Economic Growth: Factors and Predictive Trends

Sherali Ozodov
School of Information
University of Arizona
sheraliozodov@arizona.edu

1 Introduction

This project aims to advance the understanding of economic growth by deploying some machine learning techniques to analyze a wide array of economic indicators. Economic growth, traditionally assessed through GDP per capita, is an essential but often oversimplified indicator of a country's economic health. By incorporating a richer set of variables such as agricultural, industrial, and service sector contributions to GDP, health and education expenditures, and trade dynamics, this study seeks to present a more nuanced view of what drives economic progress across nations.

The use of linear regression and k-means clustering allows for both linear modeling and clustering of countries into various development stages. This approach not only predicts future economic trends but also provides a deeper understanding of the intricate factors that influence these trends. Initial exploratory data analysis has helped identify key variables that correlate strongly with GDP, which are further explored through detailed models.

The methodologies and findings of this project are intended to offer actionable insights that could influence policy-making and strategic economic planning. By identifying the significant drivers of economic growth and classifying countries by their development profiles, this research provides a foundational tool for policymakers and economic strategists to tailor their approaches to the specific needs and strengths of their economies.

2 Methods

Description of Algorithms

Linear Regression

Linear regression predicts a quantitative response, expressed mathematically as:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

where $\beta_0, \beta_1, \ldots, \beta_p$ are coefficients, $X_1, \ldots, X_p$ are predictors, and ϵ represents the error term.

Polynomial Regression

Enhances linear models by introducing polynomial terms to capture non-linear relationships:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \cdots + \beta_n X_1^n + \epsilon$$

Ridge Regression

Applies a shrinkage penalty to the coefficients to prevent overfitting, expressed as:

$$\hat{\beta}_{ridge} = \min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

where λ is the regularization parameter.

K-Means Clustering

Aims to partition n observations into k clusters, formulated as:

$$S_i^* = \arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2$$

where S represents different clusters and μ_i is the centroid of cluster S_i.

Description of the Dataset

The dataset is a comprehensive collection of global economic indicators, collected annually from the World Bank. It includes a wide range of features such as GDP, GDP Per Capita, Health and Education Expenditure, and sector contributions. The dataset spans from 2000 to 2022 and includes data for 215 countries.

Key Features

The dataset includes several economic indicators which are crucial for analyzing the economic health and development of a country. Here are the key features used in this study:

- **GDP**: Total economic output of a country, indicative of economic health.
- **GDP Per Capita**: Average economic output per person, a measure of average income and living standards.
- **Health Expenditure (% of GDP)**: Reflects the investment in health relative to the economy's size.
- **Education Expenditure (% of GDP)**: Indicates investment in education.
- **Inflation Rate (%)**: Annual percentage change in the cost of living.
- **Net Trade (% of GDP)**: Difference between export and import percentages of GDP, indicating trade surplus or deficit.
- **Population**: Total number of people in the country.
- **Agriculture, Industry, and Services (% of GDP)**: Economic sector contributions to GDP.

Country Name	Year	Agriculture (% GDP)	Ease of Doing Business	Education Expenditure (% GDP)
Afghanistan	2000	27.50	40.72	13.67
Afghanistan	2001	27.50	40.72	13.67
Afghanistan	2002	38.63	40.72	13.67
Afghanistan	2003	37.42	40.72	13.67
Afghanistan	2004	29.72	40.72	13.67

Table 1: Sample of the dataset

Evaluation Metrics

- **Root Mean Squared Error (RMSE)**: Measures the average magnitude of the errors.

- **R-Squared (R^2)**: Indicates the goodness of fit, measuring how well observed outcomes are replicated by the model.

Validation Method

We employ a train-test split method, partitioning the data into 75% training and 25% testing subsets. Additionally, 5-fold cross-validation is used during model training to ensure robustness and generalizability, effectively utilizing different subsets of the training data as validation sets to tune and evaluate the model before the final testing phase.

Hyperparameter Tuning

For ridge regression, we use a grid search with 10-fold cross-validation to find the optimal value of λ. This process helps in balancing bias and variance.

K-Means Clustering

To determine the optimal number of clusters (k), we apply the elbow method by analyzing the inertia plot to find a point where increasing k yields diminishing returns.

3 Results

This section presents the outcomes of the multivariate linear regression model developed to assess the impact of various factors on economic growth. The analysis leveraged several economic indicators, including Health Expenditure, Education Expenditure, Export, Import, Population, RD, Land, and Net Trade.

Linear Regression Analysis

The linear regression model achieved a high coefficient of determination (R^2), indicating an excellent fit to the data:

- Test RMSE: 56,770,177,062.45
- Test R^2: 0.99743

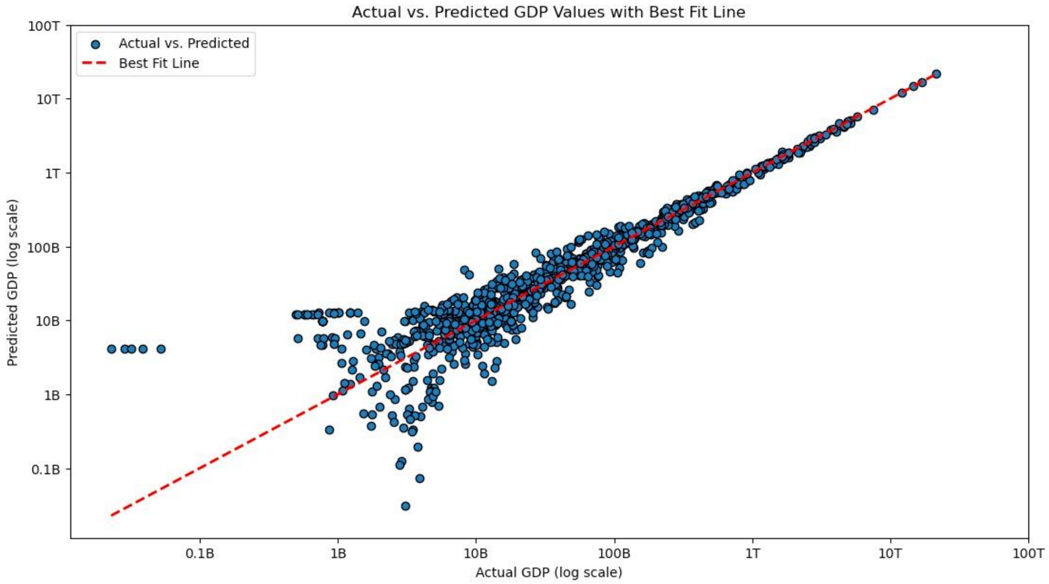

Figure 1: Scatter plot of actual versus predicted GDP values with best fit line, showing a strong linear relationship on a logarithmic scale.

Despite a seemingly high RMSE value, when compared to the scale and variability of the GDP data, the model's predictions are quite accurate. The RMSE percentage relative to the mean GDP stands at 18.19%, which is reasonable given the GDP's standard deviation of 1,405,806,885,993.53.

Regression Coefficients Interpretation

The model provided several significant coefficients, indicating the importance of RD, trade balance, and education and health expenditures on GDP.

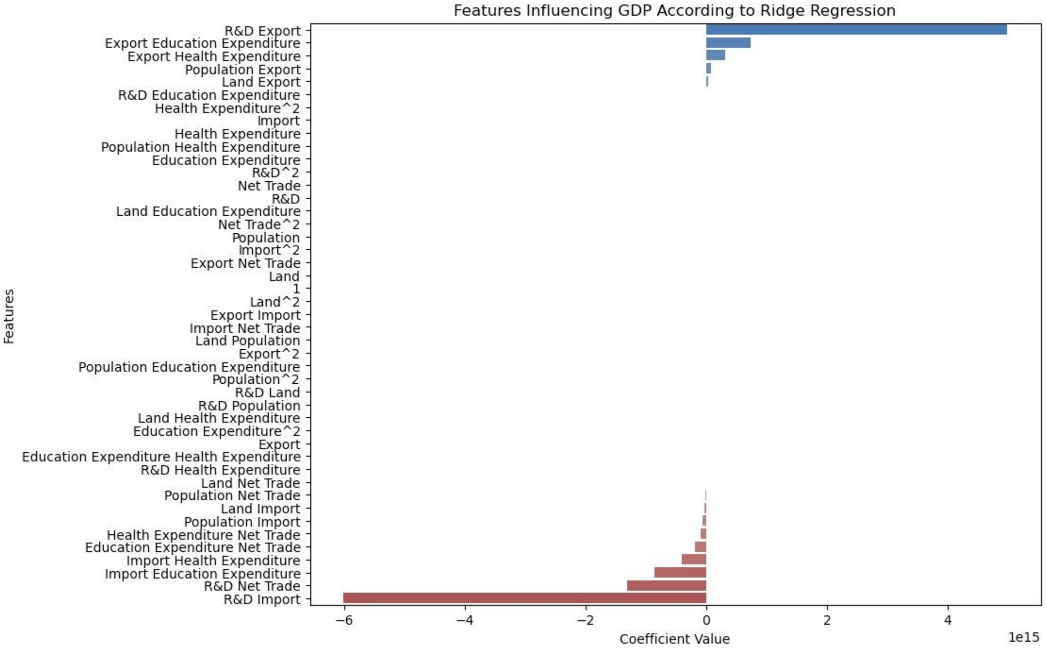

Figure 2: Features influencing GDP according to ridge regression analysis.

Key Observations from Regression Coefficients

- **R&D and GDP:** The model reveals a very strong positive influence from terms like ‘R&D Export‘, with an exceptionally high coefficient (approximately 4.985×10^{15}). This suggests that investments in R&D, particularly those linked to export activities, have a potentially massive impact on GDP. This observation supports economic theories that posit innovation as a crucial driver of economic growth.

- **Trade Balance (Net Trade):** The interaction terms involving Net Trade (exports minus imports) such as ‘Net Trade^2‘ and ‘Education Expenditure × Net Trade‘ indicate that a positive trade balance generally benefits GDP. This result is consistent with economic models where export-driven strategies bolster national economic performance.

Clustering Analysis

The K-Means clustering algorithm categorized countries into groups with similar economic characteristics, unveiling intrinsic groupings in the data. The PCA of economic indicators visualized the clustering result.

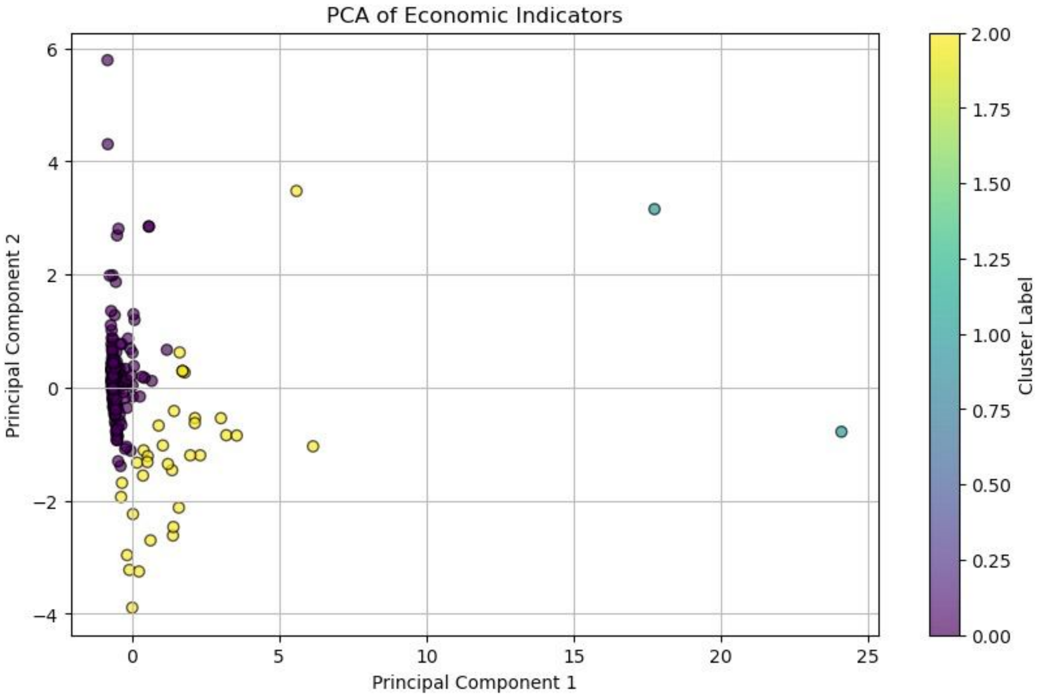

Figure 3: PCA of economic indicators showing the clustering of countries.

Cluster analysis revealed distinct economic groupings:

- **Cluster 0**: Emerging Economies
- **Cluster 1**: Dominant Economies
- **Cluster 2**: Developed Economies

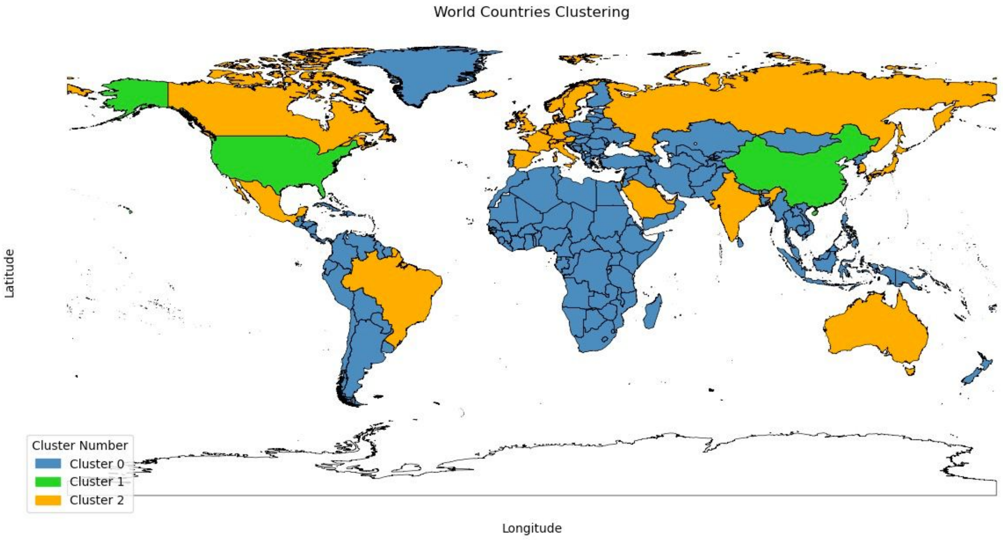

Figure 4: World map showing countries color-coded by cluster, highlighting global economic divisions.

Each cluster exhibits unique characteristics, such as GDP size, inflation rate, health and education expenditures, and trade dynamics.

4 Discussion

The multivariate linear regression analysis has revealed that R&D investments, along with health and education expenditures, are significant predictors of a country's GDP. The strong positive correlation between R&D spending and economic output supports the theory that innovation is a key driver of economic growth. This underscores the importance of strategic investments in research and development as a means to foster economic advancement.

In addition, the balance of trade, as captured through the Net Trade variable, emerged as an influential factor. Countries with a positive trade balance tend to have higher GDP figures, suggesting that policies encouraging exports over imports can be beneficial to economic health.

The clustering analysis added an extra dimension to our understanding by identifying distinct groups of countries with similar economic characteristics. For instance, Cluster 0 represents emerging economies with significant variability in GDP size and higher inflation rates, indicating a phase of rapid growth or transition. Cluster 1, comprising dominant economies, showed high GDP and GDP per capita, alongside substantial investments in health and education, suggesting a link between such investments and higher living standards. Cluster 2 encompasses developed economies with well-established industrial and technological sectors.

These findings have profound implications for economic policy. For emerging economies, there is a clear indication that policies fostering R&D, alongside educational and health improvements, may accelerate their growth trajectory. For developed economies, maintaining the balance between exports and imports could be key to sustaining economic stability.

Overall, the study highlights the multifaceted nature of economic growth and the need for a nuanced approach to economic policymaking that takes into account a wide range of factors, including R&D, health and education, and international trade dynamics.

Implications for Economic Policy:

1. **Invest in R&D:** The data strongly supports enhancing national research and development initiatives. Policy measures should focus not only on boosting domestic R&D investments but also on ensuring these investments translate into export opportunities which the model suggests have significant positive effects on GDP.
2. **Refine Trade Policies:** The contrasting coefficients on R&D-linked imports and exports underscore the need for nuanced trade policies. It's essential to strike a balance that supports the import of innovation-enhancing goods while promoting the export of domestically produced high-tech products and services. This strategy can help optimize the benefits of global value chains.
3. **Monitor and Support Strategic Sectors:** The model highlights significant roles for *Education Expenditure* and *Health Expenditure*. Policies that support these sectors may not only improve national welfare but also boost economic performance, as evidenced by their positive coefficients in the model.