Global Economy Indicator

A Project Report

Submitted By

Heyt Gala

Illinois Institute of Technology Chicago

Masters in Computer Science

CS584 Machine Learning
Professor: Oleksandr Narykov

Introduction:

The "Global Economy Indicators" dataset on Kaggle is a comprehensive collection of economic indicators across countries and years. This dataset encompasses a wide range of economic factors, including but not limited to Gross Domestic Product (GDP), population, inflation rates, unemployment rates, and more. Compiled from various reliable sources, this dataset provides a valuable resource for researchers, analysts, and data enthusiasts interested in exploring and understanding the dynamics of the global economy.

Key Features of the Dataset:

Diverse Economic Indicators: The dataset covers a broad spectrum of economic indicators, allowing for a holistic analysis of global economic trends.

Temporal and Geographical Coverage: Spanning multiple years and countries, the dataset facilitates time-series analysis and cross-country comparisons, enabling users to identify patterns, trends, and outliers.

Missing Value Handling: The dataset has undergone preprocessing, including the handling of missing values, ensuring the reliability of the data for analytical purposes.

Data Types: Both numeric and categorical variables are present, allowing for a variety of analytical approaches, including regression, clustering, and classification.

The overarching goal is to glean insights into global economic trends, with a specific emphasis on GDP dynamics across countries and time periods. The analytical approach involves data preprocessing, exploratory data analysis (EDA), and the implementation of cross-entropy strategies utilizing various regression models.

Data Loading and Exploration:

The initial phase involves loading the dataset into a Pandas data frame, allowing for a preliminary understanding of its structure and content. Subsequently, any missing values are imputed with zeros, ensuring a comprehensive dataset for subsequent analyses.

Data Preprocessing:

To facilitate effective modeling, categorical variables within the dataset are transformed into numeric representations through label encoding. This step is pivotal for the application of machine learning algorithms, enabling them to discern patterns in the data.

Non-numeric columns are selected, and label encoding is applied to convert categorical variables into numeric format. The correlation matrix of the entire dataset is calculated and visualized using a heatmap.

Exploratory Data Analysis:

4.1 Global GDP Over Time

A temporal analysis of global GDP provides valuable insights into its trajectory since 1970. A line plot is employed to visually depict this evolution, offering an overarching perspective on global economic trends.

4.2 Top and Bottom Countries by GDP in 2021

The identification of the top and bottom countries based on GDP in the year 2021 is crucial for understanding global economic disparities. Horizontal bar plots succinctly portray these rankings, shedding light on the economic standing of nations.

Cross Entropy Strategies Implementation:

The subsequent section delves into the application of cross-entropy strategies, leveraging different regression models to predict GDP trajectories for both the top and bottom countries.

5.1 Linear Regression:

5.1.1 Top and Bottom Countries

Linear Regression serves as the foundational model for predicting GDP trajectories. The evaluation metrics, including Mean Squared Error (MSE) and R-squared, provide a quantitative assessment of model performance.

5.1.2 Visualization

Scatter plots visually compare the predicted GDP values against actual values, elucidating the efficacy of the Linear Regression model.

5.2 Random Forest:

5.2.1 Top and Bottom Countries

The Random Forest Regressor, a more complex ensemble model, is introduced to capture nuanced relationships within the data. Performance metrics, such as MSE and R-squared, offer a comprehensive evaluation.

5.2.2 Visualization

Scatter plots extend the comparative analysis, encompassing both Linear Regression and Random Forest predictions for enhanced interpretability.

5.3 Gradient Boosting:

5.3.1 Top and Bottom Countries

Gradient Boosting, a sequential ensemble method, further refines the predictive capabilities. Evaluation metrics and visualizations contribute to a holistic understanding of model performance.

5.3.2 Visualization

Scatter plot comparing actual GDP values and predicted values for Linear Regression, Random Forest, and Gradient Boosting.

5.4 XGBoost:

5.4.1 Top and Bottom Countries

XGBoost, a sophisticated gradient boosting algorithm, represents the pinnacle of the cross-entropy strategies. Comprehensive evaluations and visualizations provide nuanced insights into its predictive power.

5.4.2 Visualization

Scatter plot comparing actual GDP values and predicted values for Linear Regression, Random Forest, Gradient Boosting, and XGBoost.

Few Results:

Figure 1: Correlation Heat Map

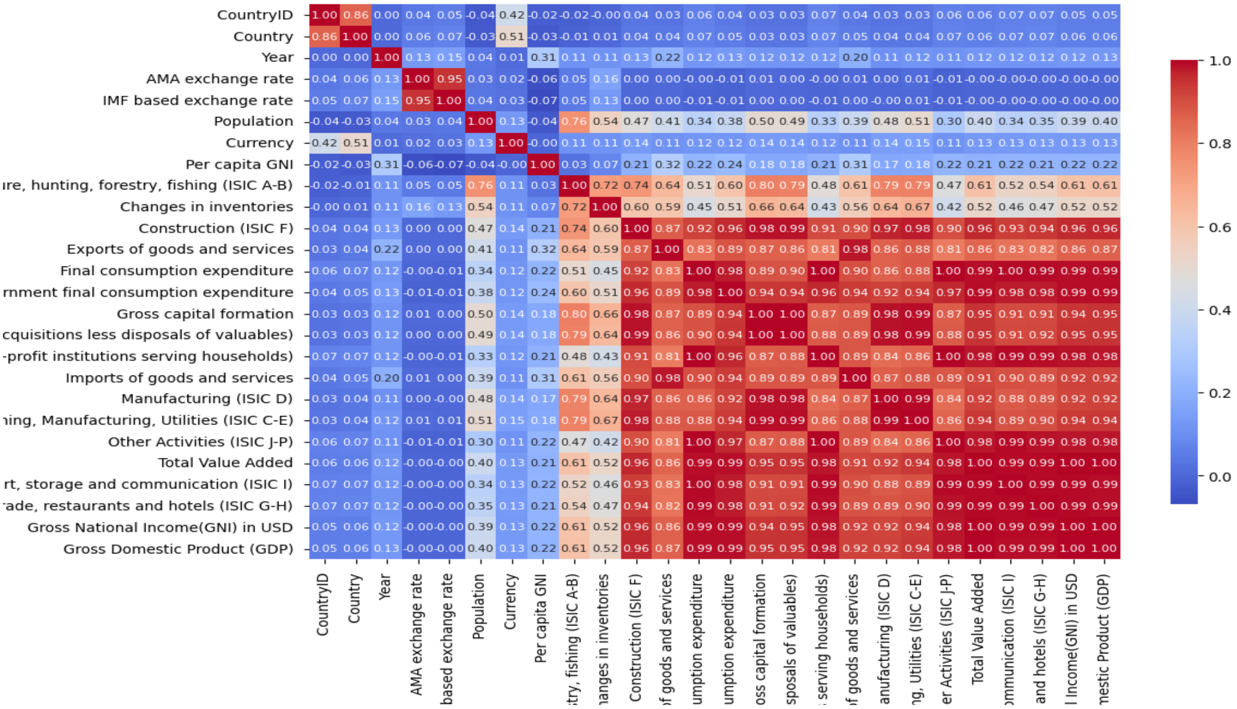


Figure 2: Global GDP

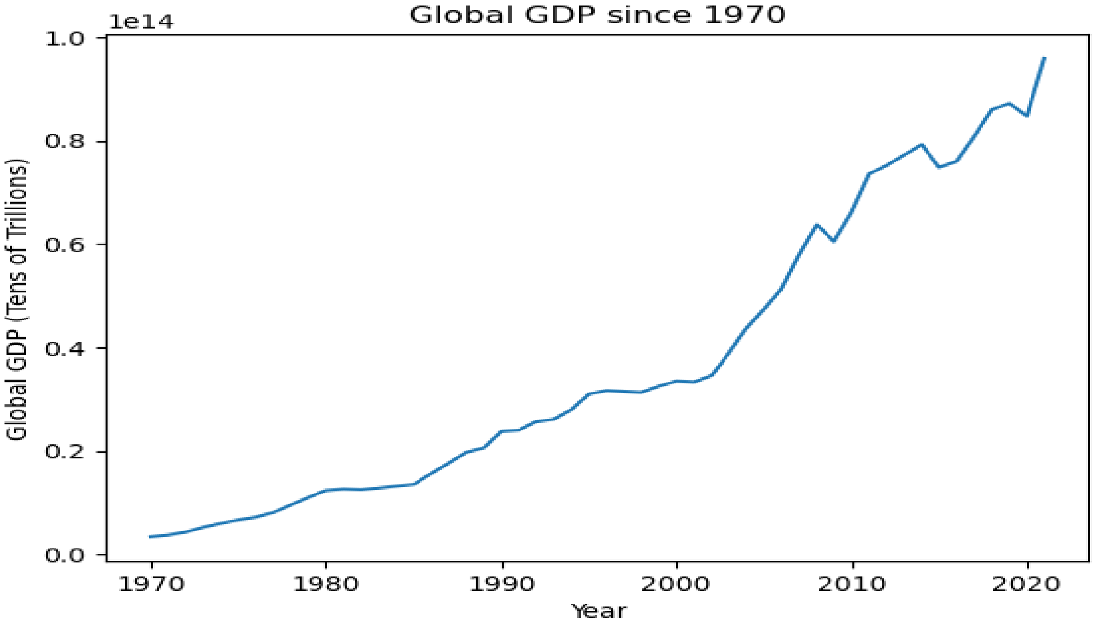
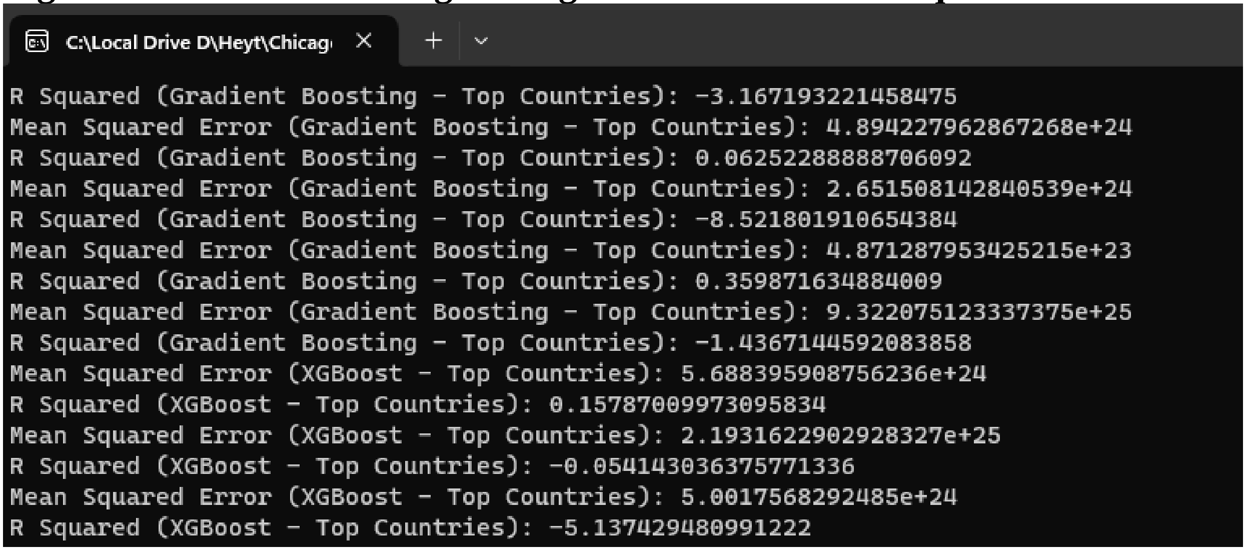
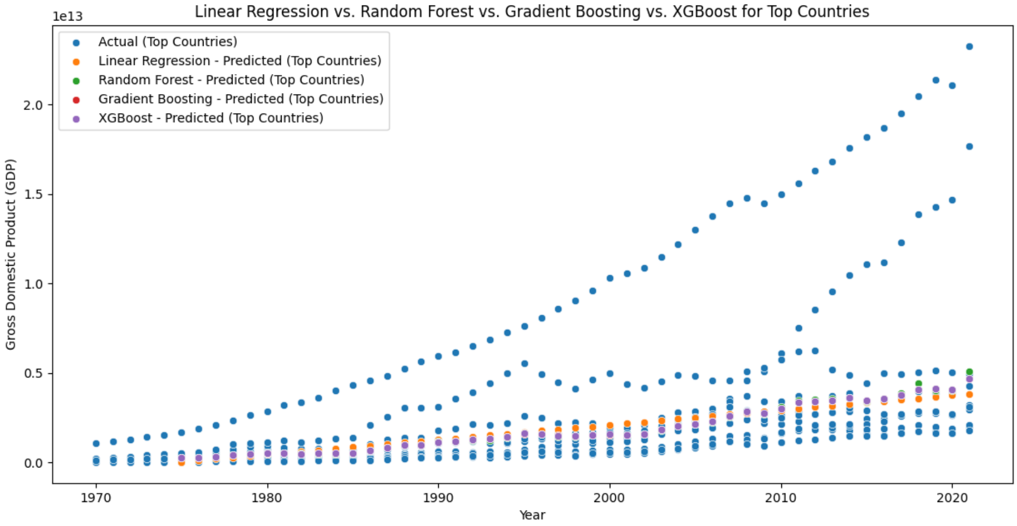


Figure 3: Gradient Boosting and XgBoost – MSE and R squared



```
R Squared (Gradient Boosting - Top Countries): -3.167193221458475
Mean Squared Error (Gradient Boosting - Top Countries): 4.894227962867268e+24
R Squared (Gradient Boosting - Top Countries): 0.06252288888706092
Mean Squared Error (Gradient Boosting - Top Countries): 2.651508142840539e+24
R Squared (Gradient Boosting - Top Countries): -8.521801910654384
Mean Squared Error (Gradient Boosting - Top Countries): 4.871287953425215e+23
R Squared (Gradient Boosting - Top Countries): 0.359871634884009
Mean Squared Error (Gradient Boosting - Top Countries): 9.322075123337375e+25
R Squared (Gradient Boosting - Top Countries): -1.4367144592083858
Mean Squared Error (XGBoost - Top Countries): 5.688395908756236e+24
R Squared (XGBoost - Top Countries): 0.15787009973095834
Mean Squared Error (XGBoost - Top Countries): 2.1931622902928327e+25
R Squared (XGBoost - Top Countries): -0.0541430363757713 36
Mean Squared Error (XGBoost - Top Countries): 5.0017568292485e+24
R Squared (XGBoost - Top Countries): -5.137429480991222
```

Figure 4: Comparison of Cross entropy strategies algorithms using Time Series data concept



Conclusion:

In summation, this analysis combines data-driven methodologies with machine learning techniques to unravel the intricate tapestry of global economic indicators. Theoretical considerations underpin each stage, from data preprocessing to model implementation, fostering a nuanced understanding of the complex interplay between time, economic variables, and predictive models. The findings herein provide a foundation for deeper inquiries into the dynamics of global economic evolution.

GitHub Repository Link:

https://github.com/Heytgala/GlobalEconomyIndicators/