GDP Growth Prediction Model

World Bank

ML_Project_TheBridge

Marta Gil Antunano

PROBLEM

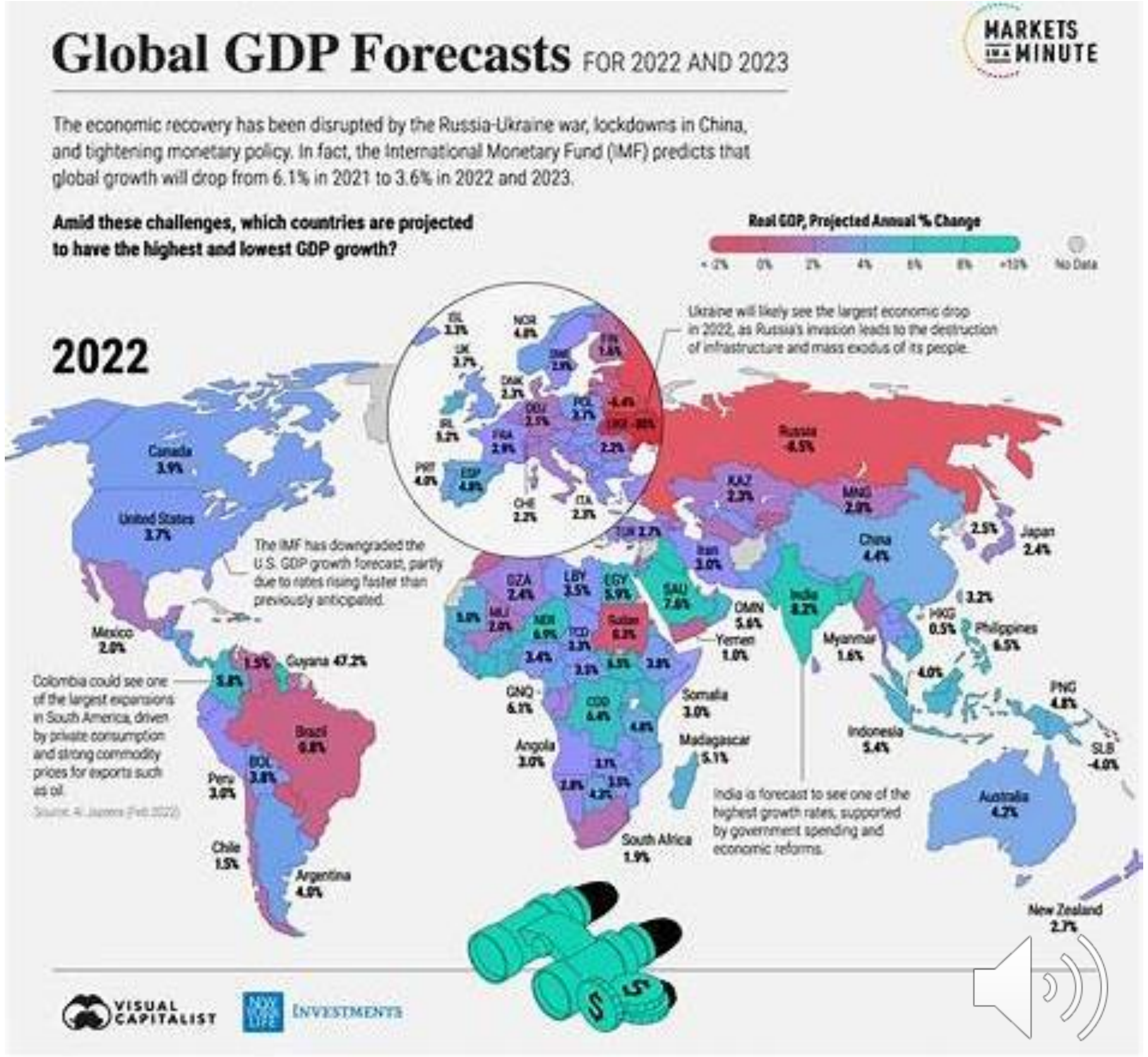
A private equity firm, we want to be able to understand which countries should our clients invest in, and continuously reappraise as new data comes in.

HOW WE WILL SOLVE THIS

Projected GDP growth is a key marker of a country´s economic health and performance.

We can track GDP changes globally to best determine which countries are opportunities for investment, and continuously monitor our investments as new data comes in, optimizing our investments.

We are looking at a supervised regression problem.



Global GDP Forecasts FOR 2022 AND 2023

The economic recovery has been disrupted by the Russia-Ukraine war, lockdowns in China, and tightening monetary policy. In fact, the International Monetary Fund (IMF) predicts that global growth will drop from 6.1% in 2021 to 3.6% in 2022 and 2023.

Amid these challenges, which countries are projected to have the highest and lowest GDP growth?

Ukraine will likely see the largest economic drop in 2022, as Russia's invasion leads to the destruction of infrastructure and mass exodus of its people.

The IMF has downgraded the U.S. GDP growth forecast, partly due to rates rising faster than previously anticipated.

Colombia could see one of the largest expansions in South America, driven by private consumption and strong commodity prices for exports such as oil.

Source: Al Jazeera (Feb 2022)

India is forecast to see one of the highest growth rates, supported by government spending and economic reforms.

(Example of global GDP projection in 2022)

OUR PROCESS

- Our dataset is imported from kaggle, and is an overview of World Bank Indicators from 1960 to the present moment.

- Target is: GDP growth (annual %)

- Some indicators were phased out due to the size of our dataset and their related importance towards GDP changes

(please refer to World Bank Indicators on the main.ipynb for more information)

- Features that were analyzed and considered necessary for this model were:

  - Lagged GDP growth - Past growth strongly predicts future growth

  - Investment rate (Gross fixed capital formation % of GDP)

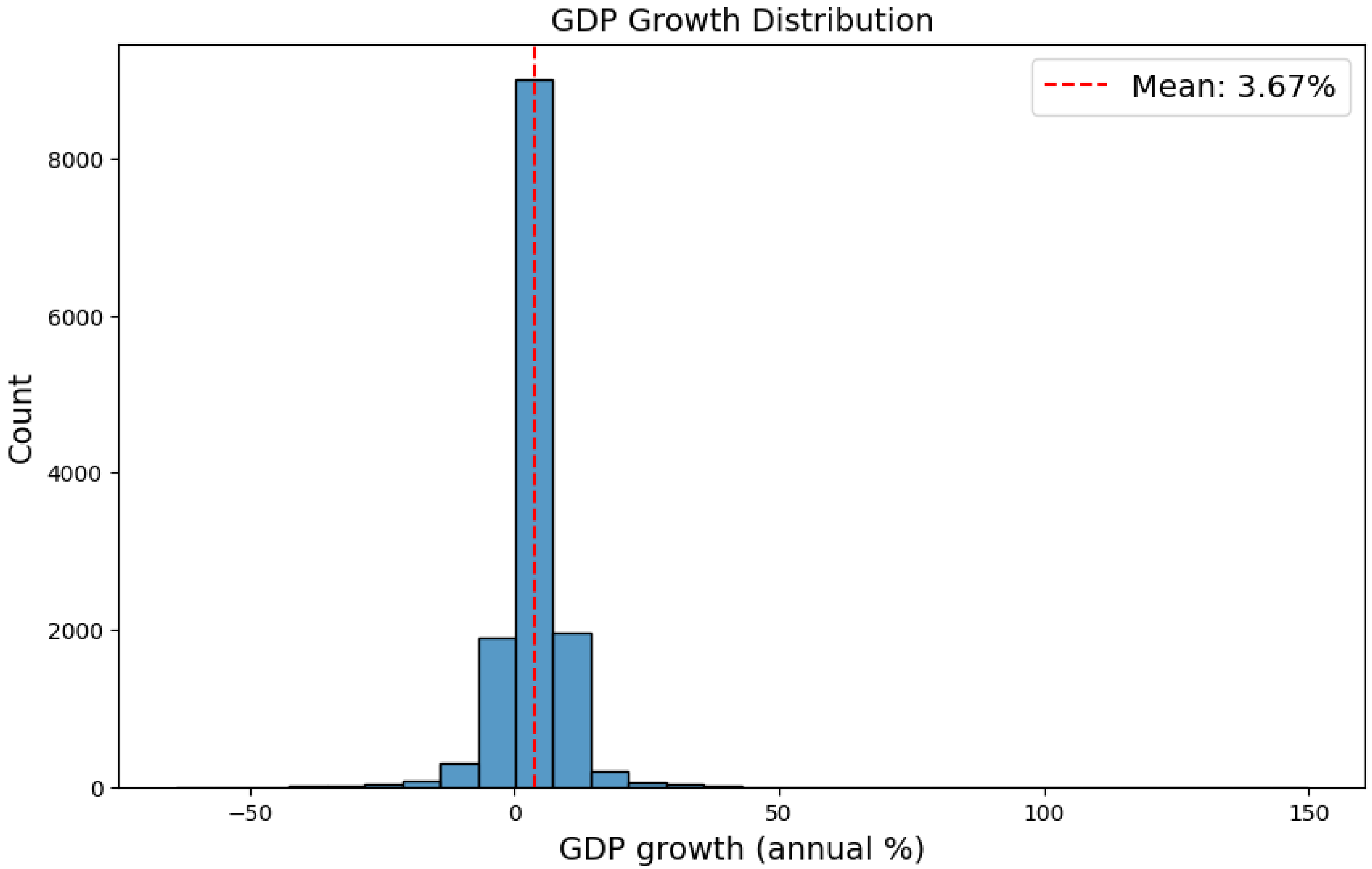
  - Savings rate - Funds future investment

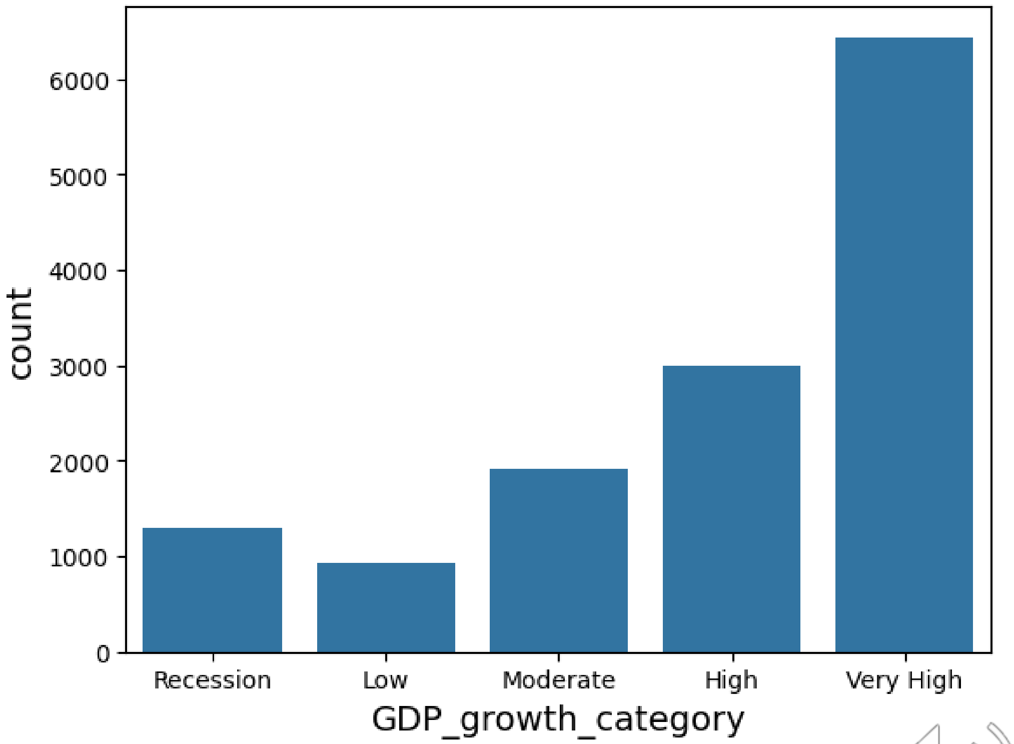
  - Trade balance - External sector health

  - Inflation - Macroeconomic stability indicator

- With some initial analysis we can see that globally GDP skews positively however there are outliers in both extremes.

- There is an imbalance in categories for GDP growth globally, probably due to different events across continents and within each country.





Positive correlations:

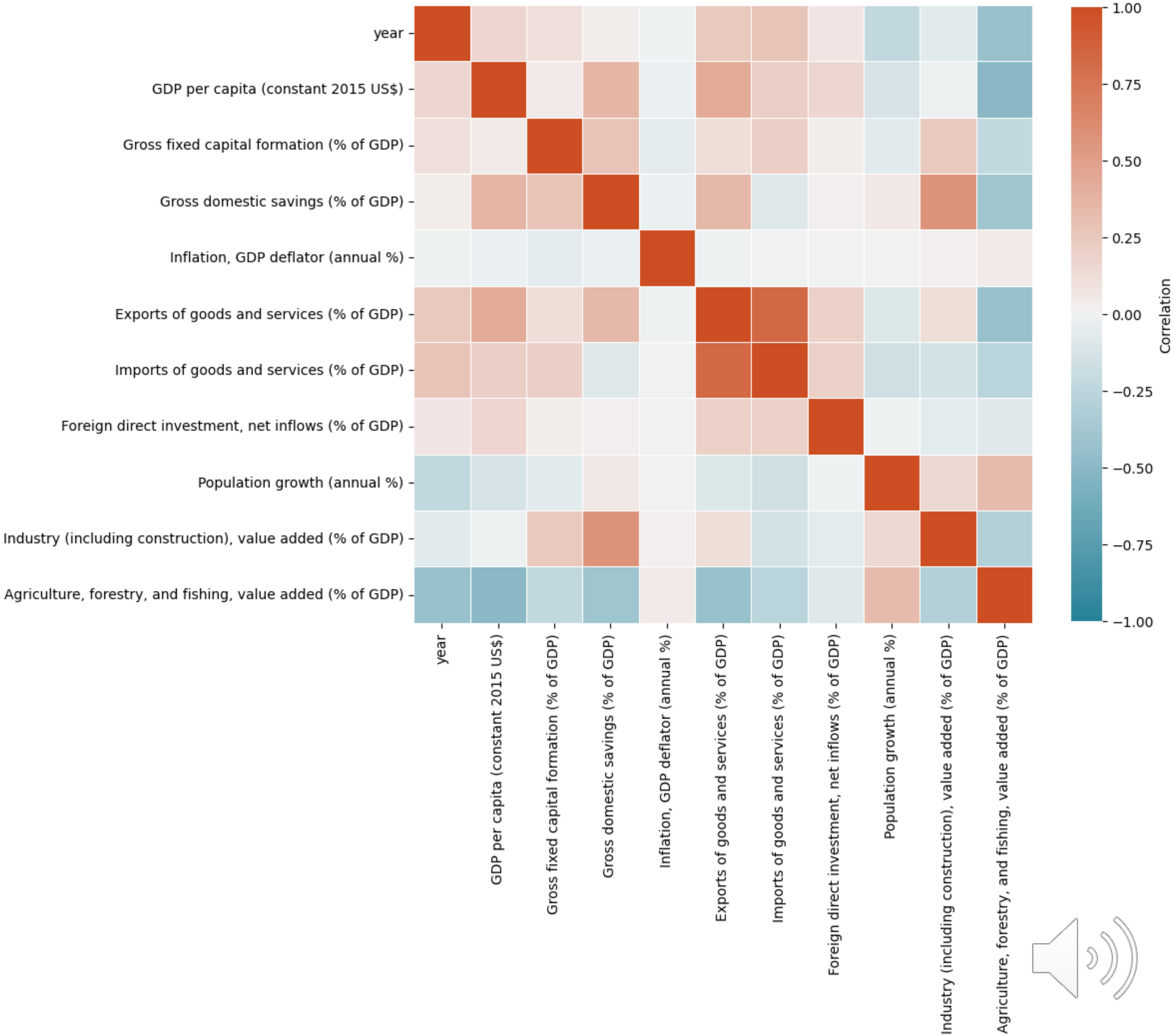
- Exports and Imports of goods and services (% of GDP) = 0.83

- More industry, higher additions to a nation's capital stock = 0.57

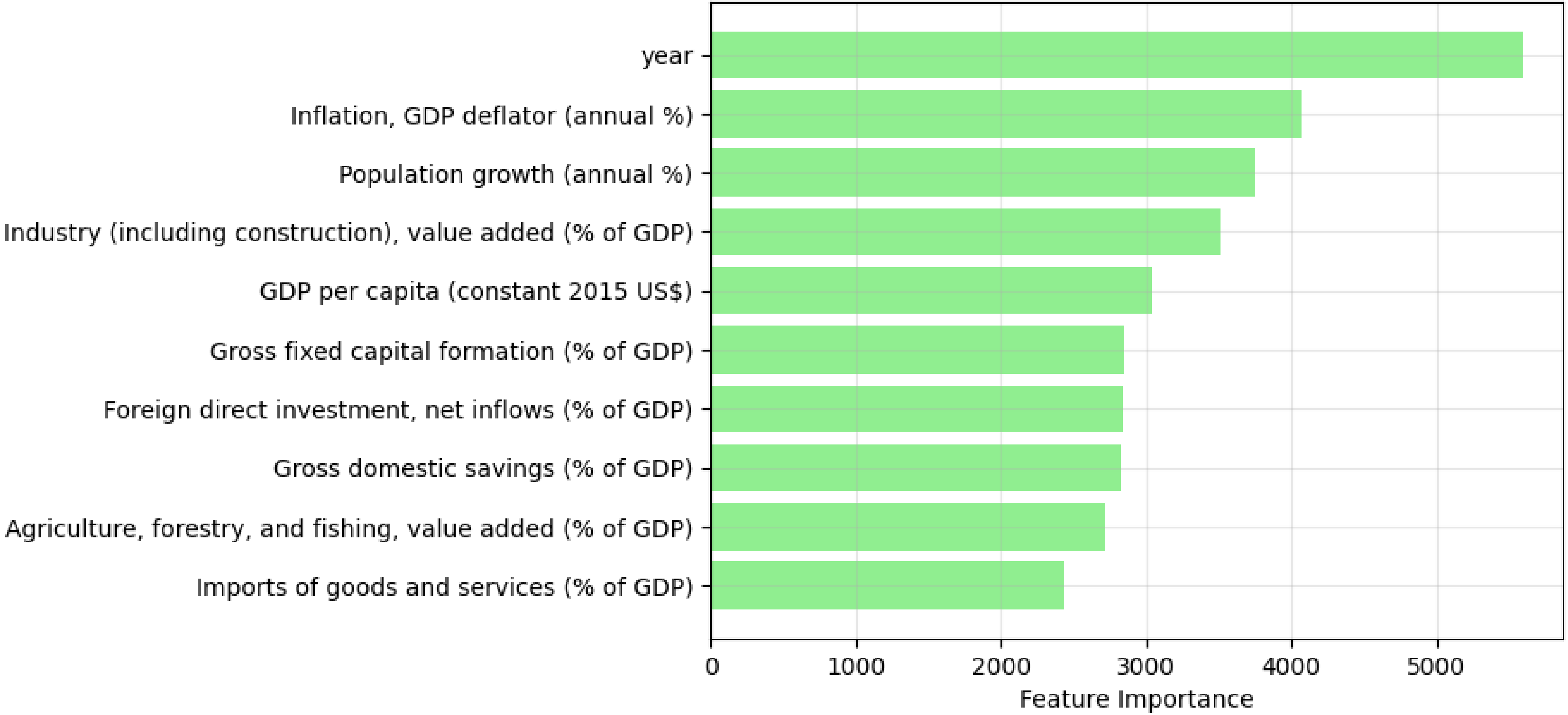
Negative correlations:

- Economies that rely on agriculture may have lower per capita income = -0.50

- Faster population growth can correlate with lower GDP per capita = -0.13



Top 10 Feature Importance



MODELLING & RESULTS

- Models tested: LighGBM, RandomForest, Linear Regression, XGBoost, Elastic Net and Gradient Boosting.

- We cared accurately capturing volatile economic events and RMSE penalizes larger errors more heavily, therefore lower RMSE is better.

- RMSE across models :

Random Forest: 5.87
Gradient Boosting: 5.95
ElasticNet: 6.27
Linear Regression: 6.18
LightGBM: 5.78
XGBoost: 5.92

- LightGBM was chosen due to RMSE of **5.78**

- After hyperparameter optimization with Random Search:
  - Best CV RMSE: 5.3904
  - RMSE after test performance: **5.7536**

- Overfitting detected (R^2 diff: 0.378)

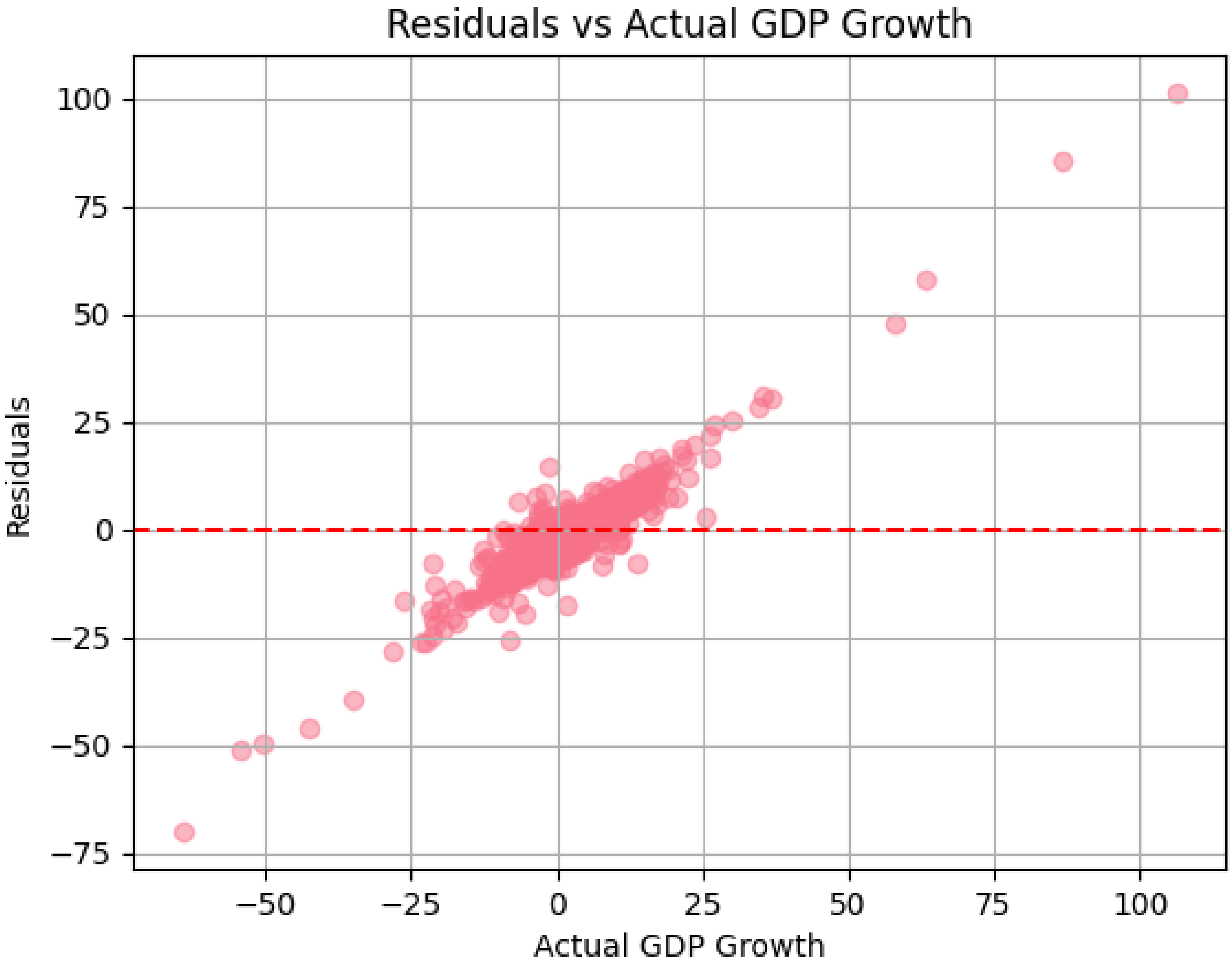
FINAL CONCLUSIONS

Model works reasonably well for the central GDP values (say, -10% to +20%), but fails to generalize to extreme economic scenarios, likely because:

- These are rare in training data

- The model is regularized and can't fit these extremes

- Outliers dominate RMSE and especially MAPE



NEXT STEPS

Further optimization of current model:

- 1. Clipping or Winsorizing Target Values: Limit extreme GDP growth values (e.g., clipping to [-25%, +25%]):
  This is best when outliers are rare and not critical to your use case.

- 2. Log-Transforming (or Box-Cox) the Target:
  This would be best if we ´d want smooth behaviour and have mostly positive or shifted GDP values.

- 3. Remove or Segment Outliers, detect and remove extreme GDP entries or treat them as a separate model task:
  If outliers were rare and clearly different from your core dataset (e.g., conflict states, pandemics).

- 4. Robust Loss Functions (huber, quantile, fair): Change the LightGBM loss function to reduce sensitivity to outliers:
  When the goal iskeep the full data but reduce outlier sensitivity smartly.

Next Step Choice: *Remove or Segment Outliers*