GDP Growth Prediction Using Macroeconomic Indicators

Author: Muhammad Taha Raees
Course: Data Science Tools & Techniques

FAST-NUCES

1. Introduction

Gross Domestic Product (GDP) growth is one of the most important indicators of a country's economic performance. Accurately predicting GDP growth can help governments, policymakers, and financial institutions make informed decisions regarding economic planning, investment strategies, and risk management.

This project uses macroeconomic indicators from the **World Bank** to build a machine-learning model capable of forecasting GDP growth. The goal is to understand which features contribute most to economic performance and evaluate how well a Random Forest model can predict a country's annual GDP growth.

(This project only uses data for the country Pakistan)

2. Methodology

The project followed a standard machine-learning pipeline:

Data Preprocessing

- Loaded macroeconomic indicators from the World Bank dataset.
- Selected numerical indicators such as:
  - Inflation
  - Unemployment

  - Lending rate
  - Interest rate

- Scaled numerical features using **StandardScaler**.
- Extracted **year** and performed a **train-test split**, using:
  - **Data before 2018 for training**
  - **Data from 2018 onward for testing**

Models Implemented

Two models were implemented for comparison:

- **Random Forest Regressor (primary model)**
- **Ridge Regression (baseline linear model)**

Hyperparameter tuning was conducted for the Random Forest model using:

- **RandomizedSearchCV**

The best model was identified based on **RMSE, R2, MAE** on the test set.

---

3. Results

The following summarizes the comparative results of each method using test MSE:

Method	R2	MAE	RMSE
Random Forest	~-0.300	~2.337	~3.033
Ridge Regression	~-0.320	~2.393	~3.056

1. **Overall Poor Performance**: Both the tuned Random Forest and Ridge Regression models exhibited poor predictive performance, as indicated by their negative R2 scores. A negative R2 suggests that neither model was able to explain any of the variance in GDP growth, performing worse than a simple baseline model that would predict the average GDP growth every time. This is a strong indicator that the models are not capturing the underlying relationships in the data effectively.

2. **Similar Error Magnitudes**: The MAE and RMSE values are relatively similar for both models (MAE ~2.3-2.4 and RMSE ~3.0-3.1). This suggests that, on average, both models' predictions were off by about 2.3-2.4 percentage points, with larger errors being somewhat penalized by RMSE

Visual Analysis

Key findings from visualizations:

- **Correlation Heatmap:**
  No macroeconomic indicator had a really strong correlation with GDP growth. **‘lending_rate’** and **‘inflation’** show the strongest relationships with **‘gdp_growth’**, both exhibiting negative correlations. This suggests that as lending rates and inflation increase, GDP growth tends to decrease.

- **Scatterplots:**
  Indicators like **‘unemployment’** and **‘lending_rate’** showed weak but noticeable relationships, although none were strongly predictive on their own.

- **Actual vs Predicted Plots:**

  - Random Forest predictions follow the actual GDP trend more closely than Ridge Regression.

  - Ridge regression fails to capture nonlinearity, producing smoother, less accurate predictions.

- **Feature Importances:**
  The Random Forest model identified the most influential features such as:

  - Lending rate

  - Inflation indicators

  - Unemployment rate

4. Discussion

The project demonstrates that predicting GDP growth from a limited set of World Bank indicators is challenging, primarily because:

1. **GDP growth is influenced by many qualitative and global factors** (politics, trade shocks, global recessions) that may not be represented in the dataset.
2. **Most indicators individually show weak correlation**, making linear models (like Ridge) unsuitable.
3. **Tree-based models perform better**, indicating nonlinear relationships between features and GDP growth.

Despite this, the Random Forest model provided moderately accurate predictions and meaningful feature importance insights. The model successfully captured complex interactions in the data, though overall prediction accuracy remains limited by the dataset's depth and economic complexity.

5. Conclusion

This project successfully explored macroeconomic data to forecast GDP growth using machine-learning techniques. The **Random Forest** model outperformed **Ridge Regression** and proved more capable of handling nonlinear relationships between macro indicators and economic outcomes.

Key takeaways:

- GDP prediction requires models that handle complex, nonlinear dependencies.
- Feature importance analysis highlights which macro indicators historically contributed most to GDP growth.
- While correlated features were identified, it's possible that the selected indicators do not fully capture all critical drivers of GDP growth

Future improvements could include:

- Considering the impact of global economic events, political stability, and natural disasters, which are not captured in the current indicators, could also be crucial for economic forecasting.

- Using more advanced models such as XGBoost or LSTMs

- Expanding to global data

References

- World Bank API