Harnessing AI in Quantitative Finance: Predicting GDP using Gradient Boosting, Random Forest, and Linear Regression Models

Farid Soroush, PhD
soroushfarid@gmail.com

2023

1 Introduction

1.1 Background

Predicting key macroeconomic indicators such as Gross Domestic Product (GDP) is a critical task in quantitative finance and economics. Precise forecasts of GDP can help in policy-making, investment decisions, and understanding the overall economic health of a country [8]. Machine learning has emerged as a powerful tool in this domain, offering sophisticated techniques for modeling complex systems and making predictions.

This project presents a comparative analysis of three machine learning models – Gradient Boosting Regressor, Random Forest Regressor, and Linear Regression – for predicting GDP. Our aim is to assess their performance and identify the model that provides the most accurate forecasts.

1.2 Machine Learning Models for Economic Forecasting

Machine learning models can capture complex non-linear relationships between variables, making them suitable for economic forecasting [5]. Among these, Gradient Boosting, Random Forest, and Linear Regression are widely used due to their versatility and robustness.

- **Gradient Boosting Regressor:** A powerful ensemble machine learning algorithm that builds a strong model by combining the predictions of several weak models, typically decision trees. It iteratively adds new models to correct the errors made by existing models [3].

- **Random Forest Regressor:** A type of ensemble machine learning model that fits a number of decision tree regressors on various subsamples of the dataset and averages the predictions to improve the predictive accuracy and control over-fitting [1].

- **Linear Regression:** A fundamental statistical and machine learning method. It assumes a linear relationship between the input variables (independent variables) and the single output variable (dependent variable). More specifically, that output (Y) can be calculated from a linear combination of the input variables (X) [6].

2 Methodology

2.1 Data Acquisition and Preprocessing

We obtained data from the Federal Reserve Economic Data (FRED) API, which included time series data on GDP, unemployment rate, inflation rate, federal funds rate, and the consumer price index (CPI) (Figure 1).

The data was normalized and time-aligned. NaN values, if any, were handled appropriately to ensure the quality of the data for the machine learning models [4]. The normalization process is essential because it scales the variables to a specific range, allowing for a fair comparison between different variables [9].

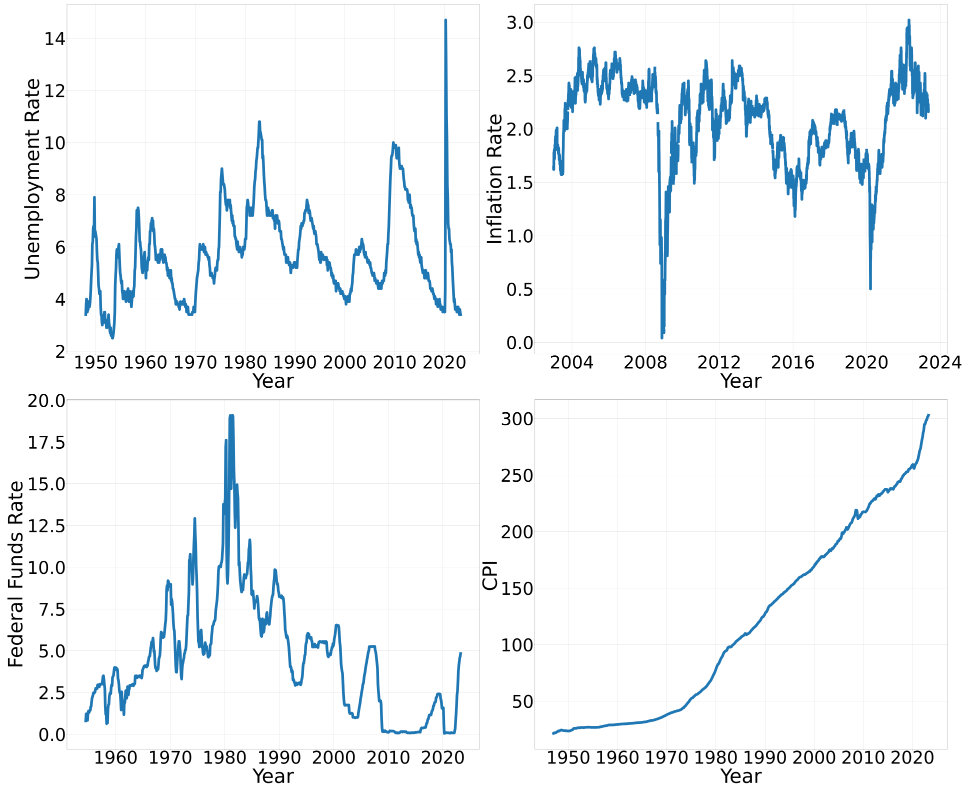

Figure 1: Economic Indicators Over Time. (a) Unemployment Rate, (b) Inflation Rate, (c) Federal Funds Rate, (d) Consumer Price Index.

2.2 Feature Engineering

The process of feature engineering involved creating new variables from existing ones to better represent the underlying data structure and improve model performance [10]. For instance, the interaction between the inflation rate and CPI may provide valuable information about the overall economic situation, which might not be captured by the individual variables.

2.3 Model Training and Parameter Tuning

Each of the three models were trained on the preprocessed dataset. The performance of these models was evaluated using the mean absolute error (MAE) metric, which provides a straightforward quantification of prediction error magnitudes.

The hyperparameters for the Gradient Boosting and Random Forest models were tuned using a grid search approach, where we tested various combinations of hyperparameters and selected the one that resulted in the smallest MAE.

3 Results and Analysis

Upon prediction (Figure 2), the performance of the Gradient Boosting Regressor, Random Forest Regressor, and Linear Regression models were evaluated. In comparing the different models, the mean absolute errors (MAEs) were calculated for each. The Gradient Boosting Machine (GBM) model had an MAE of approximately 447.54, the Random Forest (RF) model resulted in an MAE of about 409.36, while the Linear Regression (LR) model displayed a slightly higher MAE at approximately 538.04. These values provide a quantitative measure of each model's performance, with lower MAEs indicating more accurate predictions. In this analysis, the RF model proved to be the most accurate, as indicated by its lowest MAE. The MAE was calculated using the following formula:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{1}$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of samples.

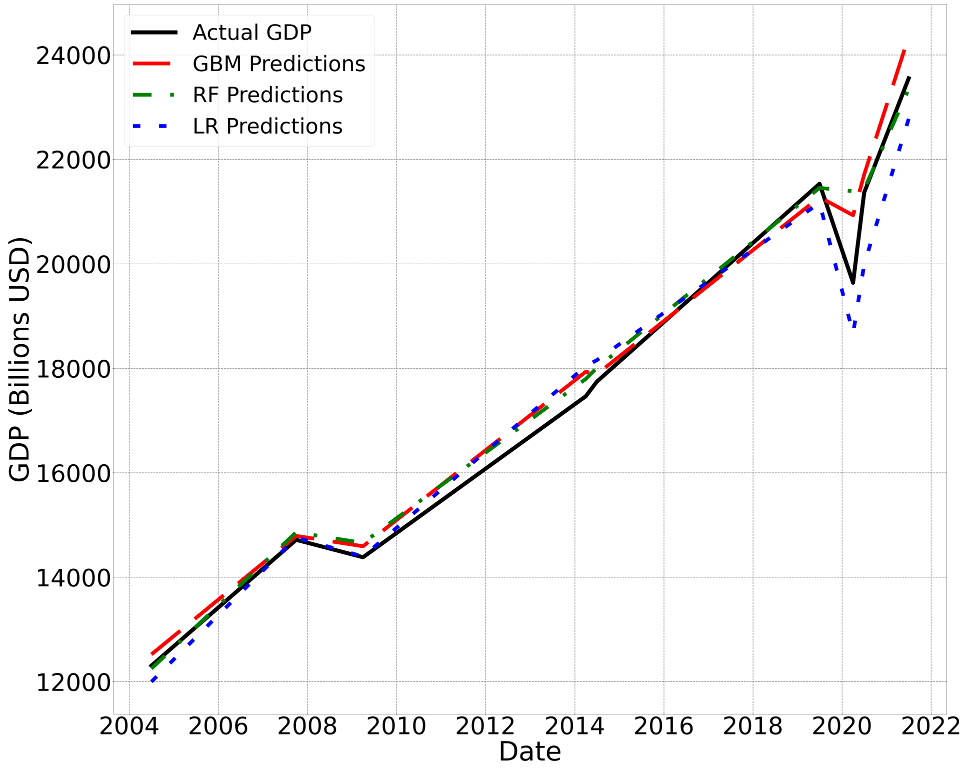

Figure 2: Actual GDP vs Predicted GDP.

4 Discussion

The results show that all three models performed well in predicting GDP. However, the Random Forest (RF) model outperformed the other two models in terms of MAE, which suggests that it managed to better capture the complex relationships between the economic indicators and GDP.

Despite these promising results, there are several limitations that need to be considered. One of the main challenges in economic forecasting is the unpredictability of economic events. Unexpected events such as a global pandemic or financial crisis can have a significant impact on GDP and are difficult to predict using historical data alone [2].

5 Conclusion and Future Work

Our study offers a valuable comparison of different machine learning models for predicting GDP. These models provide us with a robust toolset for making informed forecasts, but there are still many opportunities for improvement.

Future work could involve exploring more sophisticated models, such as deep learning or recurrent neural networks, which have shown promising results in time series forecasting [7]. Additionally, incorporating additional economic indicators or external data sources, such as news sentiment or social media data, could potentially improve prediction accuracy.

6 Appendix: PythonCode

The Python code used in this project is available on GitHub at the following link:

`https://github.com/FaridSoroush/Harnessing-AI-in-Quanti...`

6.1 Pseudocode for the Machine Learning Pipeline

Here is the pseudocode representing the main steps of the machine learning pipeline used in this study:

Algorithm 1 Machine Learning Pipeline for GDP Prediction

```
procedure GDPPrediction
    data $\leftarrow$ Load data from FRED API
    data $\leftarrow$ Preprocess data (normalization, handle NaN values)
    X, y $\leftarrow$ Split data into features and target
    X_train, X_test, y_train, y_test $\leftarrow$ Split data into training and test sets
    for model $\in$ [Linear Regression, Random Forest, Gradient Boosting] do
        model.fit(X_train, y_train)
        predictions $\leftarrow$ model.predict(X_test)
        mae $\leftarrow$ calculate MAE(y_test, predictions)
        print(model name, mae)
    end for
end procedure
```

References

[1] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[2] Olivier Coibion, Yuriy Gorodnichenko, and Rupal Kamdar. Forecasting in the presence of instabilities: How do we know whether models predict well and how to improve them. 2020.

[3] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[4] Salvador Garcia, Julian Luengo, and Francisco Herrera. Data preprocessing in data mining. 2016.

[5] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. Statistical and machine learning forecasting methods: Concerns and ways forward. *PloS one*, 13(3):e0194889, 2018.

[6] NIST/SEMATECH. *e-Handbook of Statistical Methods*. NIST/SEMATECH, 2003.

[7] Yao Qin, Dongjin Song, Haifeng Chen, Wei Cheng, Guofei Jiang, and Garrison Cottrell. A dual-stage attention-based recurrent neural network for time series prediction. *arXiv preprint arXiv:1704.02971*, 2017.

[8] James H Stock and Mark W Watson. Forecasting inflation. *Journal of Monetary Economics*, 44(2):293–335, 1999.

[9] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Elsevier, 2011.

[10] Alice Zheng and Amanda Casari. Feature engineering for machine learning: Principles and techniques for data scientists. 2018.