Fundamentals of Data Science

CSC 4780, 24SPRING

Final Project

Title

Macro-economic Trends Prediction

Group Members

Tsion Yetwale, Morgan Williams, Jason Kim

In today's rapidly changing economic environment, the ability to predict macroeconomic trends with a high degree of accuracy is of paramount importance for governments, financial institutions, and businesses worldwide. These trends serve as vital indicators for economic health, influencing policy-making, investment decisions, and strategic planning. Traditional economic forecasting methods, while useful, often fall short in capturing the complexity and dynamism of the modern economy. This project aims to analyze past macroeconomic data and predict future trends using advanced data science technology and Python programming. By leveraging state-of-the-art machine learning algorithms, we aim to develop a predictive model that not only surpasses traditional forecasting accuracy but also uncovers deeper insights into the factors driving macroeconomic indicators.

Project Objective and Problem Statement:

The project's primary objective is to utilize essential economic indicators, including personal consumption expenditure, net exports, private domestic investment, and government total expenditure, to predict Gross Domestic Product (GDP). GDP serves as a fundamental measure of a country's economic performance and overall well-being, making accurate forecasting crucial for policymakers, businesses, and investors. By developing a predictive model capable of accurately forecasting GDP, we aim to provide valuable insights into economic trends, facilitate informed decision-making, and contribute to a deeper understanding of macroeconomic dynamics.

To achieve this objective, we have chosen regression techniques as they are well-suited for predicting continuous target variables like GDP based on multiple predictor variables. Regression models allow us to establish relationships between GDP and its economic determinants, enabling us to quantify the impact of factors such as consumption, investment, exports, and government spending on overall economic output. By leveraging these techniques, we can identify key drivers of economic growth, assess their relative importance, and anticipate future economic trends.

Furthermore, the alignment between our project objective and the chosen data science techniques is evident in the nature of the problem. GDP prediction inherently involves analyzing the relationships between economic variables and their collective influence on overall economic performance. Regression techniques provide a systematic framework for modeling these relationships, allowing us to capture complex interactions and dependencies within the data. Through careful model selection, training, and evaluation, we aim to develop a robust predictive model capable of generating accurate GDP forecasts.

Data Acquisition and Understanding:

In acquiring the necessary data indicators for GDP prediction, we utilized the Fred API, recognized as a reliable and comprehensive economic data source. The Federal Reserve Economic Data (FRED) API offers access to a vast array of economic datasets, providing up-to-date information on various economic indicators, including personal consumption expenditure, net exports, private domestic investment, and government total expenditure – all essential components for GDP analysis. Leveraging the FRED API ensured that we had access to high-quality and reliable data, crucial for conducting robust analysis and model development.

By accessing the FRED API, we were able to retrieve individual datasets for each economic indicator of interest. These datasets were then combined into a unified table, leveraging a primary key ID to facilitate further analysis. The consolidation of data into a single table allowed for a streamlined and cohesive approach to data management, ensuring transparency and reproducibility in the data collection process. Additionally, this unified dataset provided a comprehensive overview of the economic indicators, enabling us to explore their relationships and interactions effectively.

A thorough understanding of the dataset's structure, variables, and characteristics was essential for conducting effective analysis and model development. Each independent variable, including personal consumption expenditure, net exports, private domestic investment, and government total expenditure, represents a distinct economic factor that contributes to overall GDP. Understanding the significance of each variable and its potential impact on GDP allowed us to formulate hypotheses and guide our modeling approach.

Data Preparation and Cleaning:

In the process of preparing and cleaning the data for analysis, several crucial steps were undertaken to ensure the quality and reliability of the dataset. The initial phase involved addressing any inconsistencies or imperfections present in the data, which could potentially compromise the accuracy of our analysis and modeling efforts.

One of the primary tasks during data preparation was to identify and handle missing values and duplicates within the dataset. Missing values, if left unaddressed, can introduce bias and inaccuracies into our analysis, potentially leading to erroneous conclusions. Duplicates, on the other hand, can distort our understanding of the data distribution and skew the results of our analysis. To mitigate these issues, we systematically identified and removed any missing values and duplicates, ensuring that the dataset was clean and free of any irregularities.

Furthermore, in anticipation of potential issues such as heteroscedasticity – a condition where the variance of the dependent variable differs across levels of one or more independent variables – we employed data transformation techniques to preprocess the data effectively. One such technique involved log-transformation on the GDP values, which aims to stabilize the variance and achieve a more consistent spread of data points. By applying log-transformation, we were able to address heteroscedasticity and improve the overall performance of our models.

The decision to employ log-transformation was based on its proven effectiveness in addressing heteroscedasticity and enhancing the performance of regression models. By transforming the GDP values using the natural logarithm function, we achieved a more homoscedastic distribution of data points, thereby reducing the mean squared error and improving the overall accuracy of our predictions.

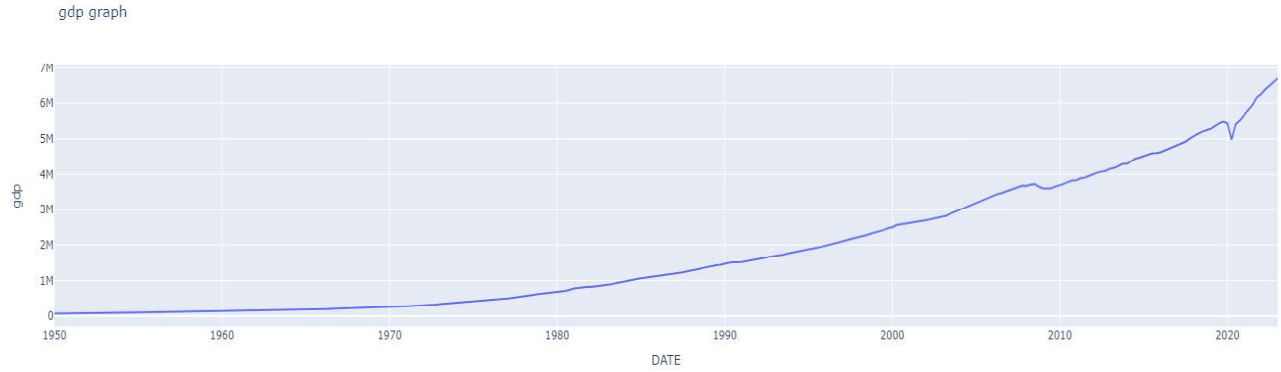
Exploratory Data Analysis (EDA):

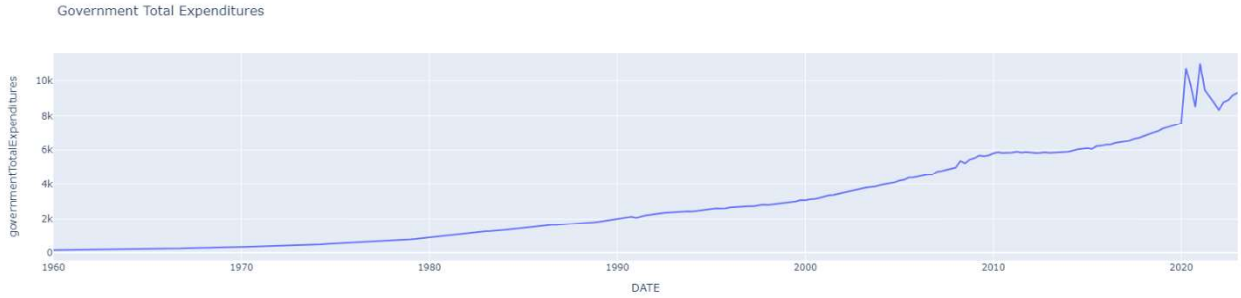
Exploratory Data Analysis (EDA) was pivotal in our project, providing essential insights into the dataset and uncovering patterns crucial for shaping our modeling strategy. We used simple visualization techniques and summary statistics to understand the distribution of GDP and its relationships with other variables. These observations helped us hypothesize and guide further analysis, such as identifying heteroscedasticity and the need for data transformation.

We started by examining the trends of GDP using line plots, which showed how it changed over time. This helped us understand the fluctuations and patterns in GDP dynamics.Additionally, we looked at how GDP related to other key variables like personal consumption expenditure, net exports, private domestic investment, and government total expenditure..

One notable finding was the impact of the COVID-19 pandemic on economic indicators. Graphs showing GDP, personal consumption expenditure, Private domestic investment and net exports revealed a significant decline during the pandemic, reflecting the broader

economic downturn. On the other hand, government total expenditure increased, likely due to measures aimed at mitigating the pandemic's effects.





Feature Engineering:

Feature engineering was a critical step in preparing the dataset for modeling. We carefully selected or created relevant predictors for GDP prediction based on domain knowledge and experimentation. By leveraging feature engineering techniques, we aimed to enhance the predictive power of our models and improve their ability to capture the underlying relationships within the data.

Model Selection and Training:

For GDP prediction, we employed a variety of regression models, each offering distinct advantages suited to different aspects of our dataset. Linear regression, a fundamental and interpretable model, assumed a linear relationship between predictors and GDP, providing insights into the direct impact of each variable. Support Vector Regression (SVR), adaptable to nonlinear relationships, aimed to find the best-fitting hyperplane while allowing for flexibility in kernel selection. Random Forest Regression, an ensemble method, leveraged multiple decision trees to capture complex interactions among

predictors and yielded feature importance measures for interpretation. Gradient Boosting Regression, another ensemble technique, sequentially built decision trees to minimize errors and achieve high predictive accuracy while offering robustness against overfitting. By employing this diverse ensemble of models, we aimed to effectively capture the nuanced dynamics of the dataset and improve the accuracy of GDP prediction through comprehensive modeling and analysis.

Model Evaluation and Validation:

To ensure the reliability of our models, we employed a technique called cross-validation, which involved splitting our data into smaller subsets. We then trained the models on some subsets and tested them on others, repeating this process multiple times to gauge their performance on new data.

Our evaluation primarily focused on two key metrics: mean squared error (MSE) and R-squared (R2). These metrics provided insights into how close our predictions were to the actual GDP values and how well the models explained the variation in GDP, respectively.

Additionally, we conducted visual checks, analyzing graphs and scrutinizing the differences between predicted and actual values, to ensure our models' performance. Overall, through testing and validation, we ensured that our final predictions were accurate and reliable for GDP prediction.

Moreover, by transforming the dependent variable (GDP) using a logarithmic function to address heteroscedasticity, we observed significant improvements in MSE across all models. For example, in the linear regression model without log transformation, the MSE was exceedingly high at 1,033,808,428.26, indicating substantial prediction errors. However, after applying the log transformation, the MSE decreased substantially to 0.10, signifying improved prediction accuracy. This reduction in MSE suggests that the log transformation helped stabilize the variance of GDP values, leading to more reliable predictions.

Similarly, the SVR model also experienced a notable decrease in MSE transformation to 0.01 after log transformation. Although the R-squared value for the SVR model after log transformation was negative (-0.414), indicating a poor model fit, the improvement in MSE suggests better prediction performance in terms of reducing errors.

Furthermore, both the Random Forest Regression and Gradient Boosting Regression models demonstrated significant MSE improvements after log transformation, with the MSE dropping to 0.00 for both models. This near-zero MSE indicates that the log transformation effectively addressed heteroscedasticity, resulting in highly accurate predictions with minimal errors.

After evaluating all the models, we found that Random Forest Regression and Gradient Boosting Regression performed better than Linear Regression and SVR. They had lower MSE values, meaning they made predictions closer to the actual GDP values, and their R2 values were closer to 1, indicating they explained more of the variation in GDP.

Overall, the project highlights the significance of leveraging data science techniques to forecast macroeconomic indicators, providing valuable insights for policymakers, investors, and other stakeholders. Accurate predictions of GDP and other economic trends enable better decision-making, resource allocation, and risk management, ultimately contributing to economic stability and growth.

Member Contribution

Tsion Yetwale: Led data extraction from Fred using the Fred API, ensuring the acquisition of crucial economic indicators.

Morgan William: data integration by merging individual datasets into a unified table, contributing to the project's architectural foundation.

Jason Kim: Played a key role in data cleaning, eliminating NAN values and duplicates, ensuring the dataset's integrity within the project's architecture.

In predictive modeling, Tsion handled dataset partitioning, Jason led model construction and training, and Morgan contributed to model evaluation using metrics like mean squared error and r squared. Moving forward, the team anticipates continued contributions from all members to enhance predictive accuracy, incorporating additional factors like inflation into the model for a more comprehensive project architecture.

Reference

https://fred.stlouisfed.org/ https://www.investopedia.com/terms/g/gdp.asp https://www.vexpower.com/brief/homoskedasticity

https://people.duke.edu/~rnau/411log.htm#:~:text=Logging%20a%20series%20often%20has,which%20price%20index%20to%20use)

https://www.imf.org/en/Publications/fandd/issues/Series/Back-to-Basics/gross-domestic-product-GDP