

**Forecasting GDP in United States**

** Executive Summary**

Forecasting GDP (Gross Domestic Product) is crucial for guiding economic policies and decision-making. Accurate GDP forecasts help governments, businesses, and investors anticipate future economic conditions and plan accordingly. Historical examples demonstrate the impact of GDP forecasts on shaping major U.S. policy responses:

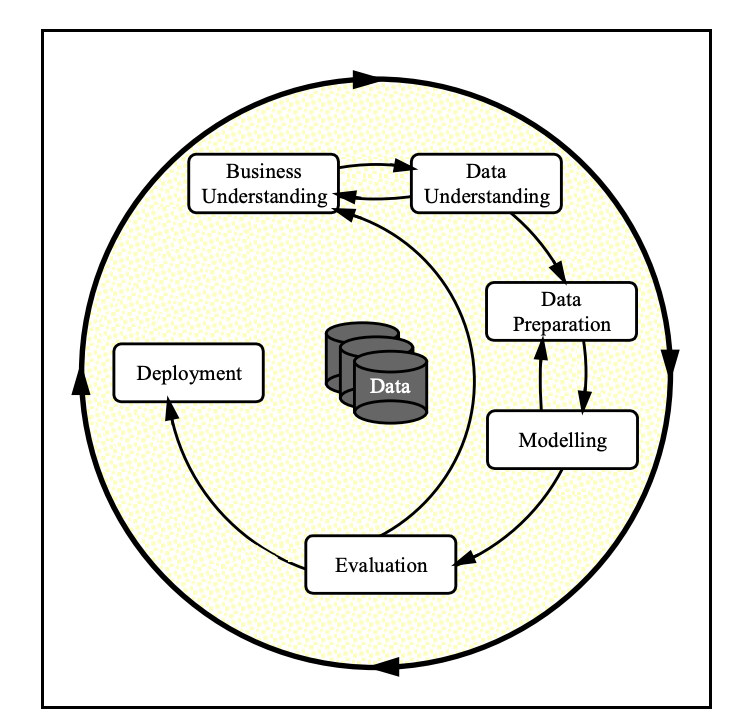
* **2008 Financial Crisis**: Early GDP forecasts showed a sharp economic contraction, prompting the U.S. government to implement the **$700 billion Troubled Asset Relief Program (TARP)** to stabilize the banking system and prevent a prolonged recession.
* **Post-World War II Economic Boom**: Positive GDP forecasts predicted strong post-war growth, leading to transformative policies like the **G.I. Bill** and large-scale infrastructure projects, which fueled decades of economic expansion.
* **COVID-19 Pandemic**: As GDP forecasts in early 2020 predicted significant economic declines due to lockdowns, the U.S. government passed the **CARES Act**, a $2.2 trillion stimulus package, to mitigate the economic fallout.

These cases illustrate the strategic importance of GDP forecasts in shaping timely and impactful policy interventions, underscoring the need for reliable economic predictions.

**CRISP-DM Framework for Forecasting**

We follow the **CRISP-DM (Cross-Industry Standard Process for Data Mining)** framework, a widely used methodology, to predict inflation. The phases are:

1. **Business Understanding**: Forecast inflation rates for different countries and identify key features influencing inflation.
2. **Data Understanding**: Analyze World Bank datasets on inflation.
3. **Data Preparation**: Clean and preprocess the data for analysis.
4. **Modeling**: Build and test models for forecasting.
5. **Evaluation**: Assess model performance.
6. **Deployment**: Implement the model for future forecasts.



** Business Understanding**

**Objective:** The goal of this project is to forecast the future trends of the U.S. GDP (Gross Domestic Product) and understand its implications for strategic decision-making across various sectors. GDP is a key economic indicator that provides insights into the health of a nation’s economy, and accurate forecasting can help guide both public and private sector planning.

Accurate GDP forecasts enable governments to assess future market conditions, make informed fiscal and monetary policy decisions, and mitigate economic risks. For businesses, these forecasts inform strategic investments, resource management, and growth planning by providing insight into future demand trends. Investors, too, rely on GDP predictions to gauge economic stability, identify risks, and optimize their portfolios for maximum returns.

Overall, by predicting GDP movements, stakeholders across sectors can align their decisions with expected economic shifts, leading to more effective resource allocation and long-term planning.

** Data Understanding**

The datasets were sourced from the Federal Reserve Economic Data:

<https://fred.stlouisfed.org/series/CPIAUCSL>

The Consumer Price Index for All Urban Consumers: All Items (CPIAUCSL) is a price index of a basket of goods and services paid by urban consumers.

<https://fred.stlouisfed.org/series/GDP>

Gross domestic product (GDP), the featured measure of U.S. output, is the market value of the goods and services produced by labor and property located in the United States.

<https://fred.stlouisfed.org/series/UNRATE>

The unemployment rate represents the number of unemployed as a percentage of the labor force. Labor force data are restricted to people 16 years of age and older, who currently reside in 1 of the 50 states or the District of Columbia, who do not reside in institutions (e.g., penal and mental facilities, homes for the aged), and who are not on active duty in the Armed Forces.

<https://fred.stlouisfed.org/series/DFF>

The Effective federal funds rate is the interest rate at which depository institutions trade federal funds (balances held at Federal Reserve Banks) with each other overnight. When a depository institution has surplus balances in its reserve account, it lends to other banks in need of larger balances. It represents the actual interest rate at which **depository institutions** (like banks) lend reserve balances to other depository institutions overnight on an uncollateralized basis.

<https://fred.stlouisfed.org/series/INDPRO>

The Federal Reserve's monthly index of industrial production and the related capacity indexes and capacity utilization rates cover manufacturing, mining, and electric and gas utilities.

<https://fred.stlouisfed.org/series/FEDFUNDS>

The Target Federal funds rate is the interest rate **set by the Federal Open Market Committee (FOMC)** as part of U.S. monetary policy. The FOMC sets a **target range** for the Federal Funds Rate to influence economic activity (e.g., to control inflation or stimulate growth). It is a policy rate set by the Federal Reserve to guide the economy. The actual DFF moves toward this target but may vary from day to day.

**Data Description:**

The above datasets show the different features from year 1960 till 2024. All of them are merged and manipulated under one dataset with the Date as index. Here is the structure of merged data:

* DATE : Index
* CPIAUCSL : Consumer Price Index
* DFF : Effective Federal Funds. DFF had daily data and it is aggregated to monthly(similar to other dataset date)
* FEDFUNDS: The Target federal funds rate
* GDP : Gross domestic product. GDP data was in quarters and the other datasets were in monthly date. The data from the quarter is populated for every 3 respective months
* INDPRO: The Federal Reserve's monthly index of industrial production
* UNRATE: Unemployment rate

** Data Preparation**

Data preparation is crucial for accurate model predictions. The following steps were implemented:

**Data Cleaning:**

* + Removed irrelevant columns.
  + Handled missing values and outliers.
  + Convert the 'DATE' column to datetime and set it as the index

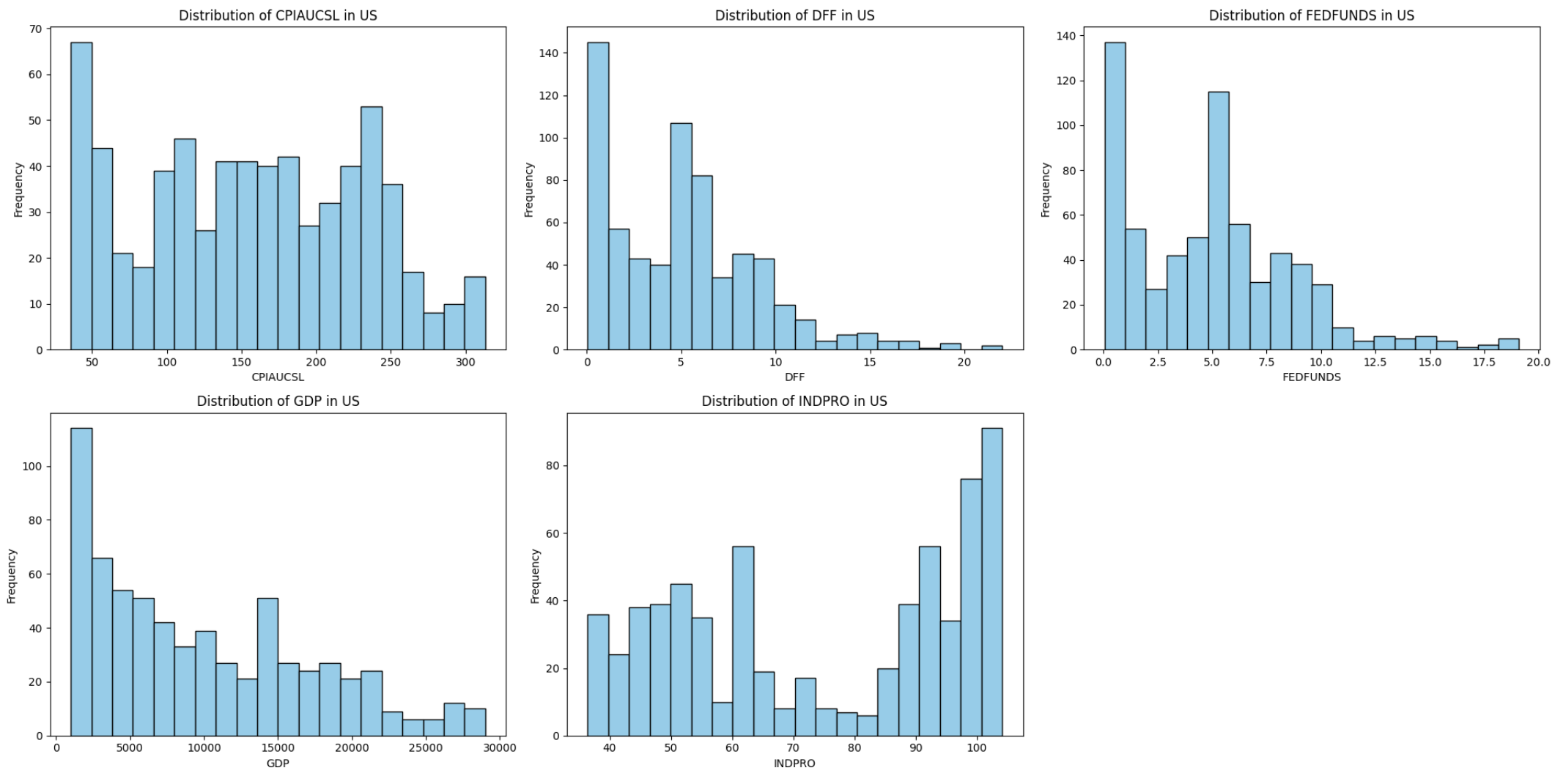
**Data Transformation**:

* + Checked for and removed duplicates.
  + Merge all data sets in one DataFrame-Using Outer Join
  + Categorial Encoding was not relevant for the dataset since all columns were numerical

** Exploratory Data Analysis (EDA)**

In the Data Exploration phase, conducted a thorough analysis to uncover patterns, relationships, and insights within the dataset.

**Visualize numerical columns** by histogram plots, correlation heatmap, and Box Plots

****

**Histogram Findings:**

CPIAUCSL (Consumer Price Index for All Urban Consumers):

The distribution is right-skewed, with a large concentration of values around 50 and a significant decline in frequency as the CPI increases. This suggests that for most of the dataset, the CPI was relatively low.

DFF (Discount Rate):

The distribution is right-skewed, with most values clustered around the lower end (0-5), and very few instances above 10. This indicates that the discount rate was low for most of the time period covered by the data.

FEDFUNDS (Federal Funds Rate):

Similar to DFF, this is also right-skewed with the majority of the data points concentrated between 0 and 5, which reflects low interest rates over much of the period.

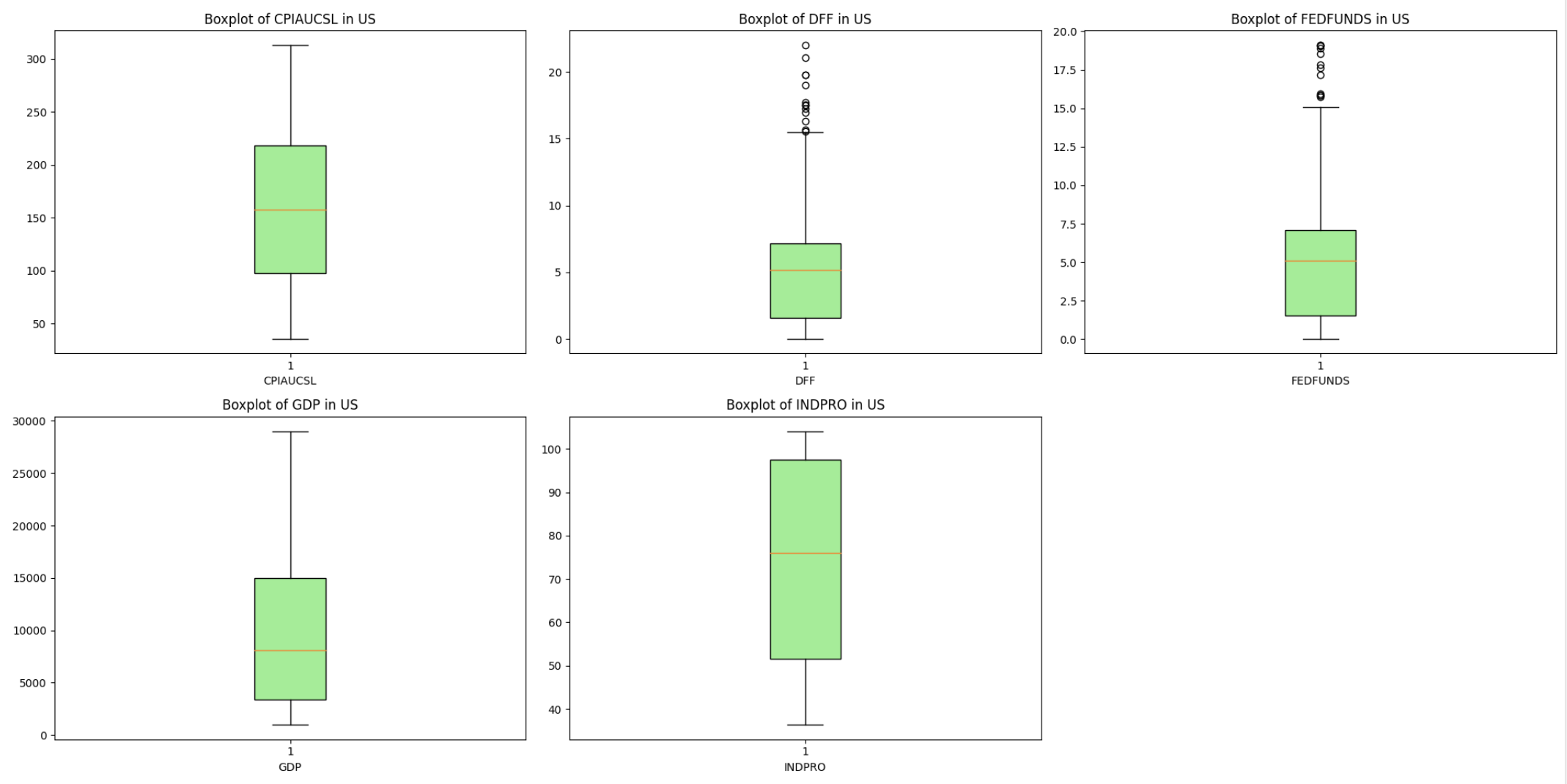
GDP (Gross Domestic Product):

The GDP distribution is heavily right-skewed, with most values clustered toward the lower end and fewer data points at higher GDP values. This indicates that, historically, the GDP was lower during a large portion of the dataset, but there have been periods of significant GDP growth as well.

INDPRO (Industrial Production Index):

The data shows a bimodal distribution, with two noticeable peaks around 40 and 100, suggesting two distinct periods with significantly different levels of industrial production.

These visualizations suggest that the U.S. economy has generally experienced low inflation, low interest rates, and moderate to high industrial production in the period that we have data.

****

**Box Plot Findings:**

CPIAUCSL (Consumer Price Index for All Urban Consumers):

The boxplot shows a relatively symmetrical distribution of values, with no outliers. The interquartile range (IQR) is between approximately 100 and 200, indicating that most CPI values fall within this range. The median is slightly above 150, which reflects the central tendency of the data.

DFF (Federal Funds Effective Rate):

The boxplot for DFF shows a right-skewed distribution with several outliers beyond 15. The bulk of the data is between 0 and 7, with the median around 5. The presence of outliers indicates that while the discount rate was generally low, there were some periods where it spiked significantly.

FEDFUNDS (Federal Funds Rate):

This boxplot also shows a right-skewed distribution, with many outliers above 10, though most of the data falls between 0 and 7.5, similar to the DFF. The median value is around 5, and the outliers indicate some higher federal fund rates during certain periods.

GDP (Gross Domestic Product):

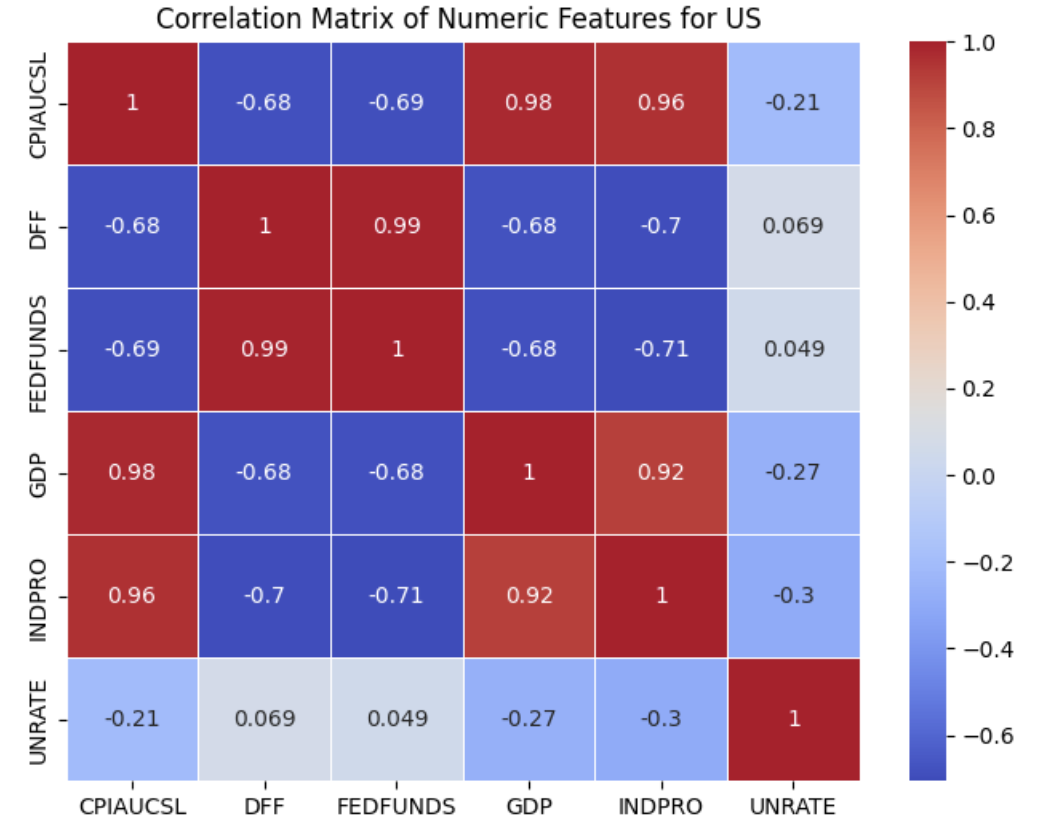
The boxplot shows that GDP values have a wide range, with the IQR spanning from about 5000 to 15000. There are no outliers, and the distribution appears relatively symmetrical. The median GDP value is around 10000, indicating a steady central tendency in GDP values over time.

INDPRO (Industrial Production Index: Total Index):

The distribution of INDPRO is relatively symmetrical as well, with the IQR ranging from approximately 40 to 80. The median is around 60, suggesting a moderate level of industrial production, with no visible outliers.

Summary:

CPIAUCSL, GDP, and INDPRO have more symmetrical distributions without outliers, suggesting that the data points are fairly consistent over time. DFF and FEDFUNDS have right-skewed distributions with many outliers, indicating that while these rates were generally low, there were periods of significant spikes. The outliers in the interest rates suggest that there have been occasional periods of elevated rates, while the overall trends have been towards lower values.



**Correlation Heatmap Findings:**

**High Correlation with GDP:**

Strong Positive Correlations:

CPIAUCSL vs. GDP (0.98) and CPIAUCSL vs. INDPRO (0.97): There is a very strong positive correlation between the Consumer Price Index (CPI) and both GDP and industrial production. This suggests that as inflation rises (as measured by CPI), GDP and industrial production tend to increase, indicating a relationship between inflation and economic growth. GDP vs. INDPRO (0.92): There is also a strong positive correlation between GDP and industrial production, which is intuitive since economic growth is often associated with higher production outputs.

Strong Negative Correlations:

CPIAUCSL vs. DFF (-0.50) and CPIAUCSL vs. FEDFUNDS (-0.52): There are moderately strong negative correlations between the CPI and both the Discount Rate (DFF) and the Federal Funds Rate (FEDFUNDS). This suggests that when inflation (CPI) rises, interest rates tend to decrease, or vice versa. This could reflect monetary policy where rates are adjusted to control inflation. GDP vs. DFF (-0.55) and GDP vs. FEDFUNDS (-0.56): There is a similar negative relationship between GDP and interest rates, indicating that higher economic growth tends to coincide with lower interest rates.

Interest Rates (DFF and FEDFUNDS):

These two rates have a near-perfect positive correlation (0.99). This is expected, as both represent closely related measures of interest rates that tend to move together in response to monetary policy changes.

Weak or Insignificant Correlations:

UNRATE (Unemployment Rate): This feature has weak correlations with most other variables, suggesting that unemployment does not have a strong linear relationship with inflation, interest rates, or GDP in this dataset. For example, the correlation between unemployment and CPI is only -0.045, and with GDP, it is -0.13. This could be due to the fact that unemployment and GDP often have a **nonlinear** or **lagged relationship**. While GDP growth generally signals a healthy economy and higher employment, the **unemployment rate** does not always adjust immediately in response to changes in GDP. For example, during periods of economic recovery, businesses may delay rehiring even as GDP begins to improve, leading to a temporary disconnect between the two indicators. Additionally, structural factors like labor market policies, automation, and changes in workforce participation can further weaken the direct correlation between GDP and unemployment. Therefore, UNRATE may not exhibit a strong immediate correlation with GDP, even though they are conceptually linked over the long term.

Summary:

Strong correlations exist between inflation (CPI), GDP, and industrial production, indicating that these factors tend to rise together. Interest rates (DFF, FEDFUNDS) are negatively correlated with inflation and GDP, which aligns with standard economic theory. The unemployment rate appears to be relatively uncorrelated with the other variables in this dataset, suggesting that its dynamics may be influenced by different or more complex factors not captured in these features.

**Z-Score**

Z-Score results shows outliers in columns: ['DFF', 'FEDFUNDS', 'UNRATE']. I capped outliers at the 5th and 95th percentiles for these columns.

** Data preparation and transformation**

* Define GPD as Regression Targe variable.

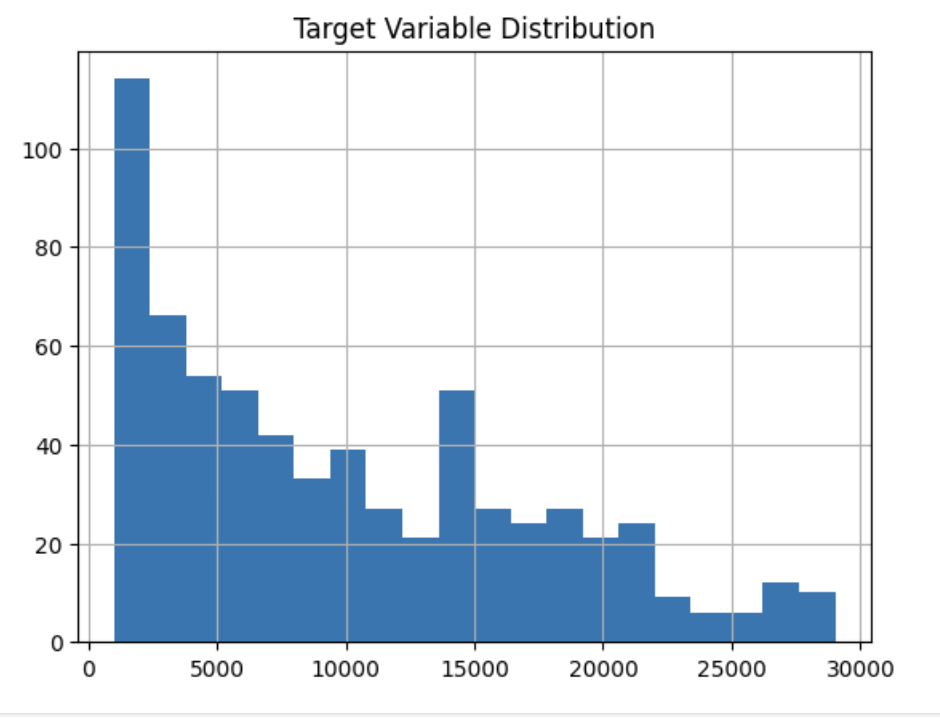
X = df.drop(columns=['GDP'], errors='ignore')

y = df['GDP']

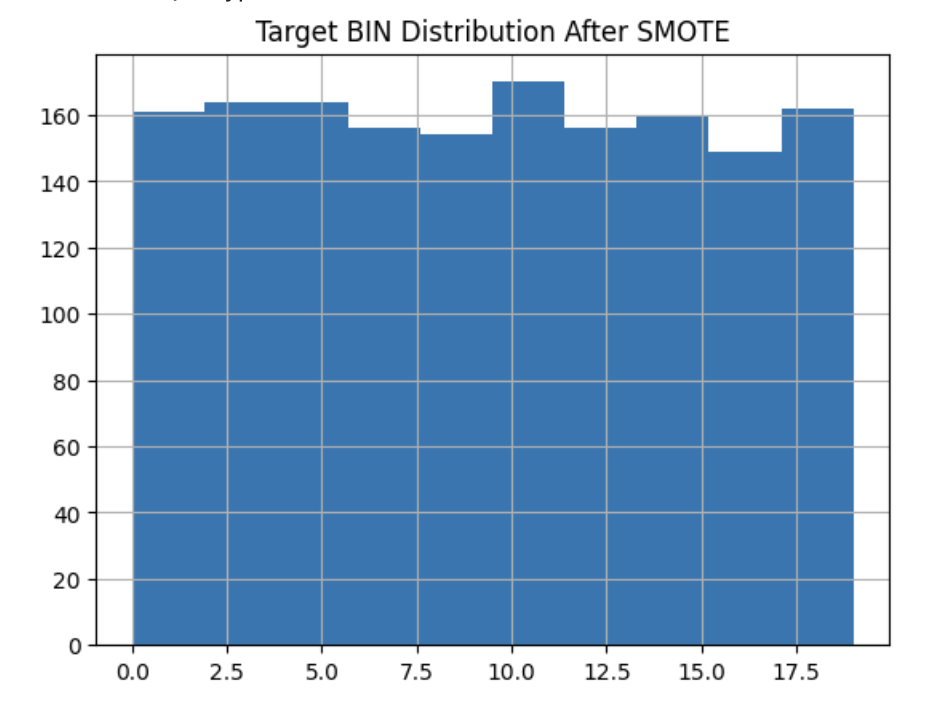
* Split data into features and target

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=42)

* Dataset is imbalanced



* Handling class imbalance with Bin-SMOTE- Initially, the dataset presented an imbalance issue, as the GDP values clustered within certain ranges. To address this, the continuous GDP variable was **binned** into categories, converting the regression task into a **classification problem**. This allowed for the application of **Bin-SMOTE (Synthetic Minority Over-sampling Technique)**, which oversampled the minority classes by creating synthetic samples to balance the dataset. This step improved the model’s ability to learn patterns from all GDP categories.



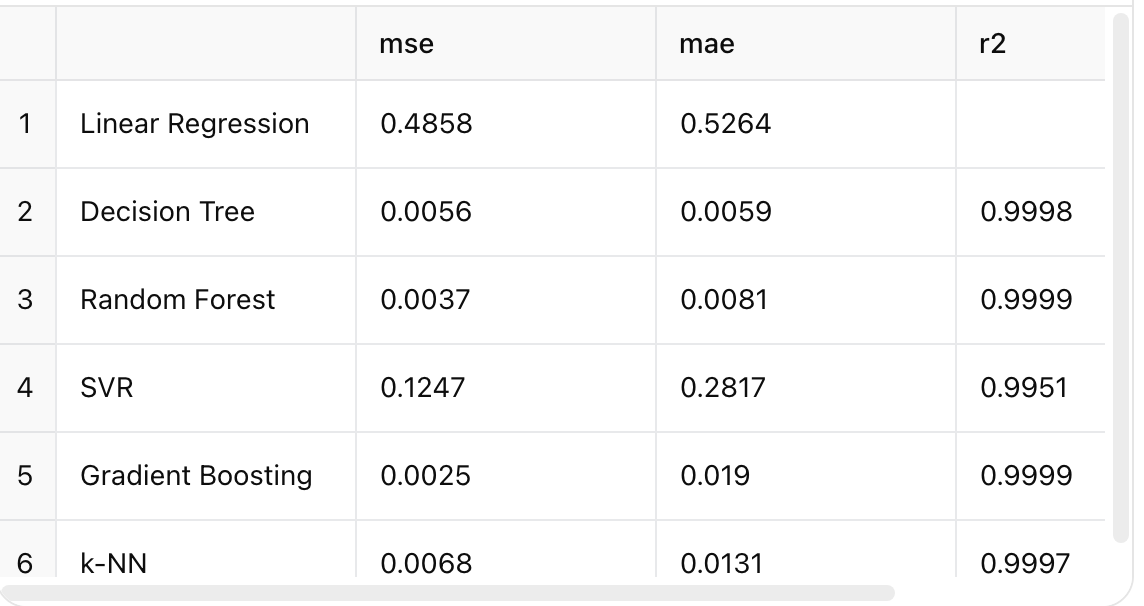
* Convert Bin to Regression- However, while Bin-SMOTE helped with classification tasks, the goal of forecasting future GDP values remained. After applying Bin-SMOTE, the models were reverted to focus on **regression tasks** to predict continuous GDP values accurately.

** Model Training and Evaluation**

Following models are applied to the merged data:

* 'Logistic Regression'
* 'Decision Tree'
* 'Random Forest'
* 'SVM'
* 'Gradient Boosting'
* 'k-NN'

**Model Results**

****

Decision Tree has the best performance with the lowest Mean Squared Error (MSE) and Mean Absolute Error (MAE), and the highest R² value. This means the Decision Tree model is almost perfectly fitting the data, explaining nearly 99.99% of the variance.

Random Forest and Gradient Boosting also perform very well, with MSE and MAE values close to those of the Decision Tree. Their R² values are also extremely high (0.9998). These ensemble methods tend to generalize better than individual models like Decision Trees, making them excellent choices for your problem.

k-NN (K-Nearest Neighbors) also shows excellent performance, with a high R² and low errors. However, it performs slightly worse than the Decision Tree and the ensemble methods (Random Forest and Gradient Boosting).

Linear Regression performs reasonably well but is significantly outperformed by the other models, with a noticeably higher MSE (0.5212) and a lower R² (0.9819). While it's still a good model, it isn't as effective as the more complex models like Decision Trees and Random Forest.

SVR (Support Vector Regression) performs decently but not as well as the others. While the R² is relatively high (0.9957), it has a higher MSE (0.1250) and MAE compared to the other models, indicating that it doesn't fit the data as well as the tree-based models.

### **Conclusion:**

Decision Tree, Random Forest, and Gradient Boosting are the top-performing models, almost perfectly fitting the data, with extremely low errors and near-perfect R² values. You may want to choose one of these models, depending on your specific use case. Linear Regression and SVR show reasonable performance but are significantly outperformed by the tree-based models and k-NN.

**Model Tuning- Hyperparameter**

Each model was optimized using hyperparameter tuning, resulting in near-perfect R² values across all models, indicating the strong predictive power of the features in the dataset.

**Tuning Hyperparameters**

|  |  |  |
| --- | --- | --- |
| **Model** | **Best Params** | **R2** |
| Decision Tree | 'min\_samples\_split': 20, 'min\_samples\_leaf': 1, 'max\_features': None, 'max\_depth': 40 | 0.9998 |
| Random Forest | 'n\_estimators': 200, 'min\_samples\_split': 2, 'min\_samples\_leaf': 1, 'max\_features': 'log2', 'max\_depth': 20 | 0.9997 |
| SVR | 'kernel': 'rbf', 'gamma': 0.01, 'epsilon': 0.1, 'C': 1 | 0.9979 |
| Gradient Boosting | 'subsample': 0.6, 'n\_estimators': 100, 'min\_samples\_split': 2, 'max\_depth': 10, 'learning\_rate': 0.2 | 0.9998 |
| k-NN | 'weights': 'distance', 'n\_neighbors': 5, 'metric': 'manhattan' | 0.9999 |

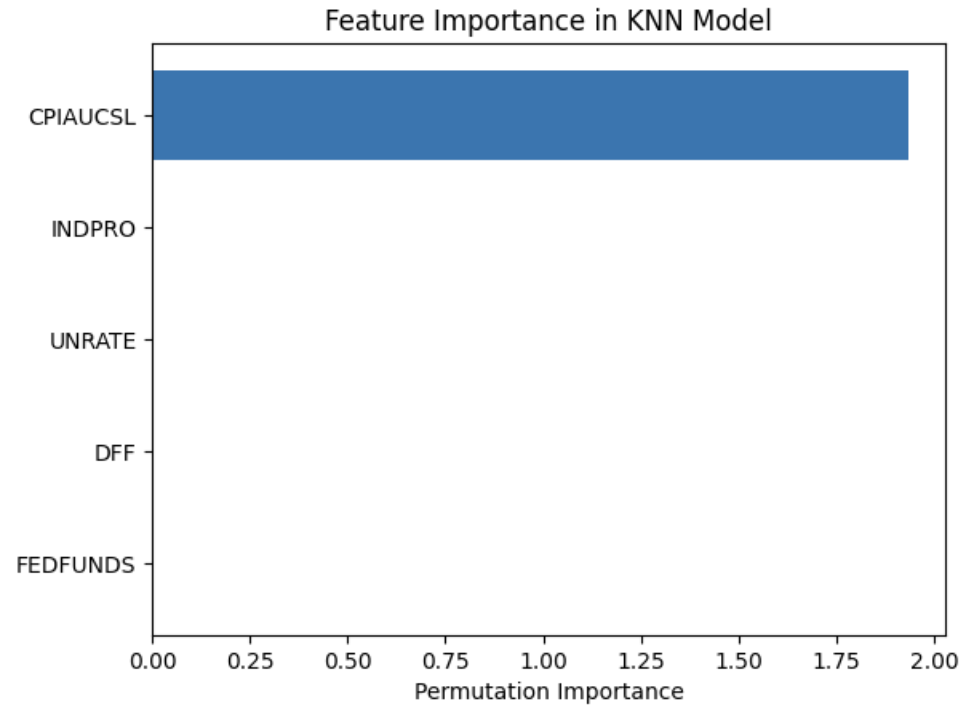
All models perform remarkably well, with R² values extremely close to 1. This suggests that the dataset has patterns that can be effectively captured by a wide variety of models. k-NN has the highest R² score, followed by Gradient Boosting and Decision Tree, indicating that these models might be best suited for your GDP forecasting task. The SVR model performs slightly lower but is still very accurate. If computational efficiency is important, models like k-NN (due to lazy learning) or SVR (which can be computationally intensive) may not be as efficient as Random Forest or Gradient Boosting.

**Cross Validation for KNN (highest R2 score model):**

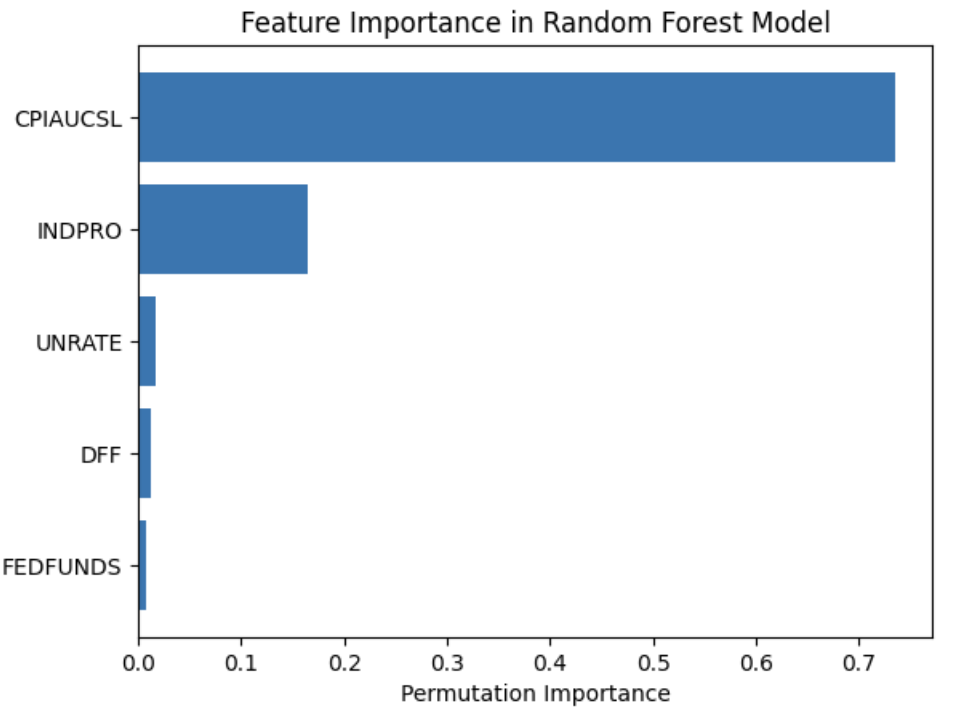
The results of cross-validation: mean R² score of 0.9999 and a very low standard deviation of 0.00008. This suggests that KNN model is performing well and consistently across all folds, with minimal variance between the different subsets of the data.

The R² score on the test set of 0.99987 is very close to the cross-validation mean R² score of 0.99985, indicating that the model is generalizing very well to unseen data. This suggests that:

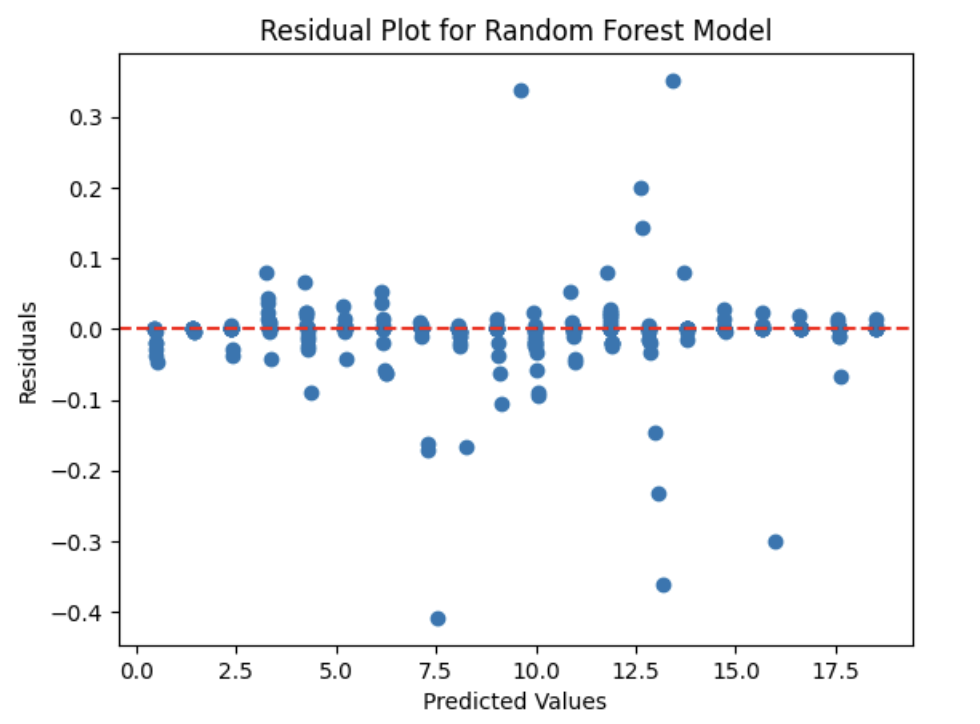
**Overfitting is not a concern.** The model is performing consistently across the training, validation (cross-validation), and test sets. With both cross-validation and test set scores being nearly identical, the current hyperparameter setup is effective.



KNN model is relying too heavily on one feature, it might be useful to experiment with other models that can capture complex interactions between features. Based on Permutation Importance of other models, Random Forest showed a better result, so it was selected as the optimize model.



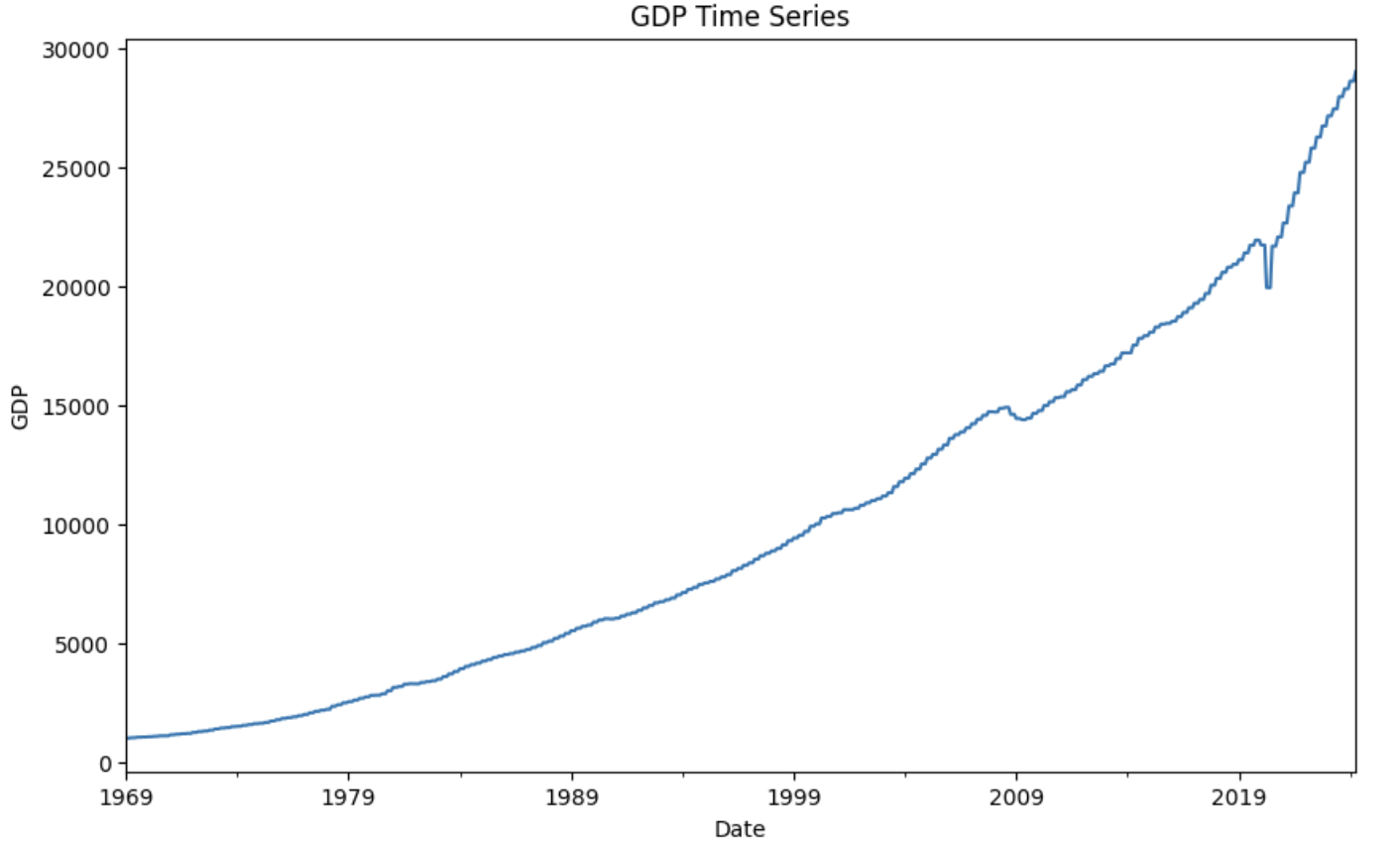
**Residual Plot for Random Forest:**



**Overall Performance:** Despite some outliers and heteroscedasticity, the model appears to be performing quite well, especially in the mid-range of predicted values, where residuals are very close to zero.

** Time Series**

* **GDP Time Series**

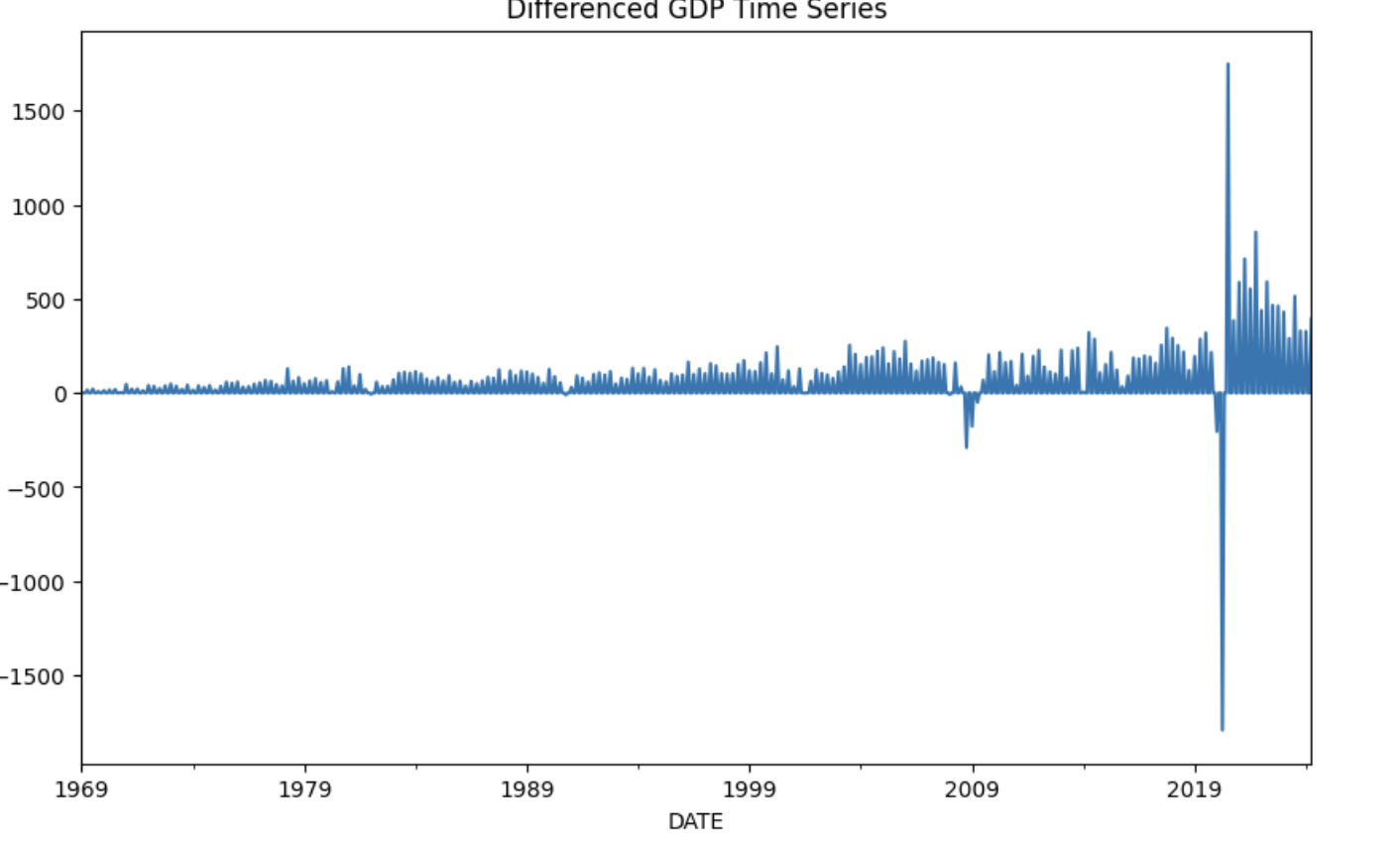
****

* **Checking Stationary**

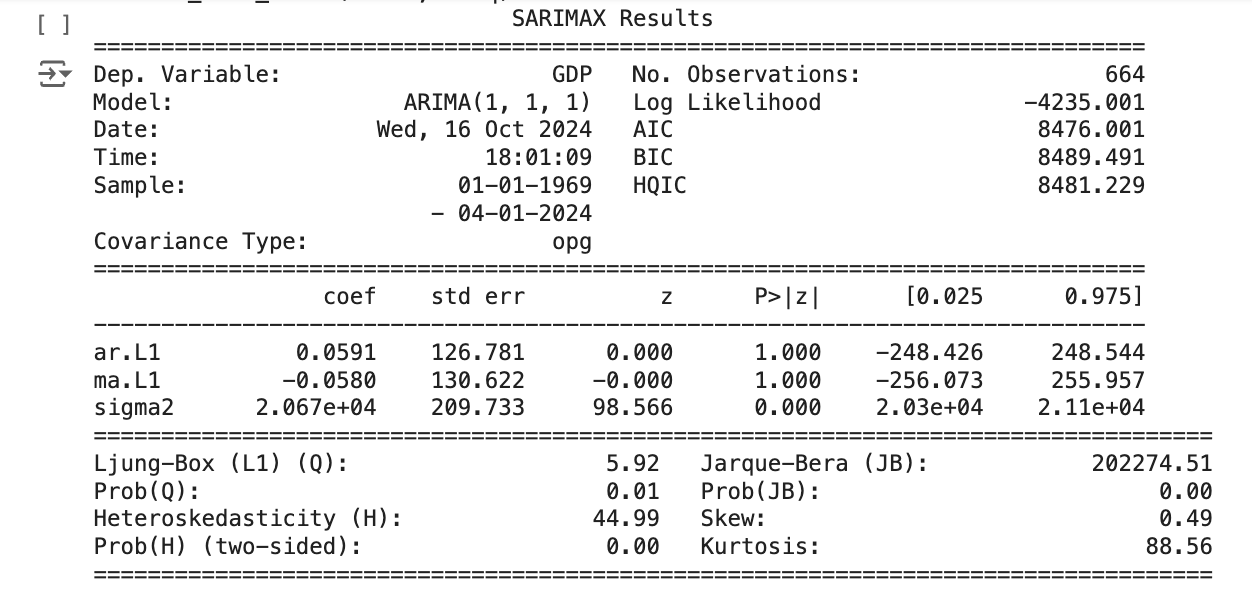
ADF Statistic: 3.58188

p-value: 1.0

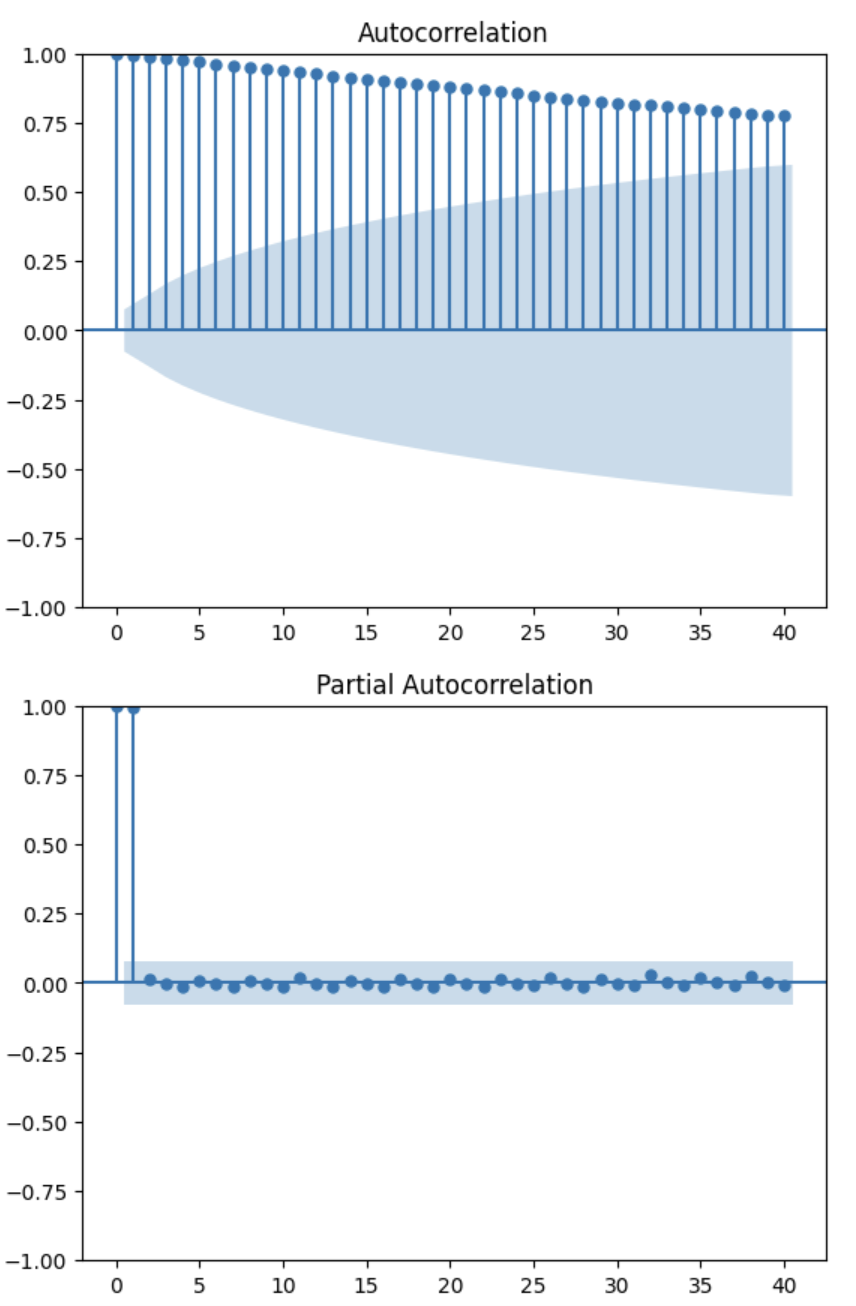
* Making the series Stationary



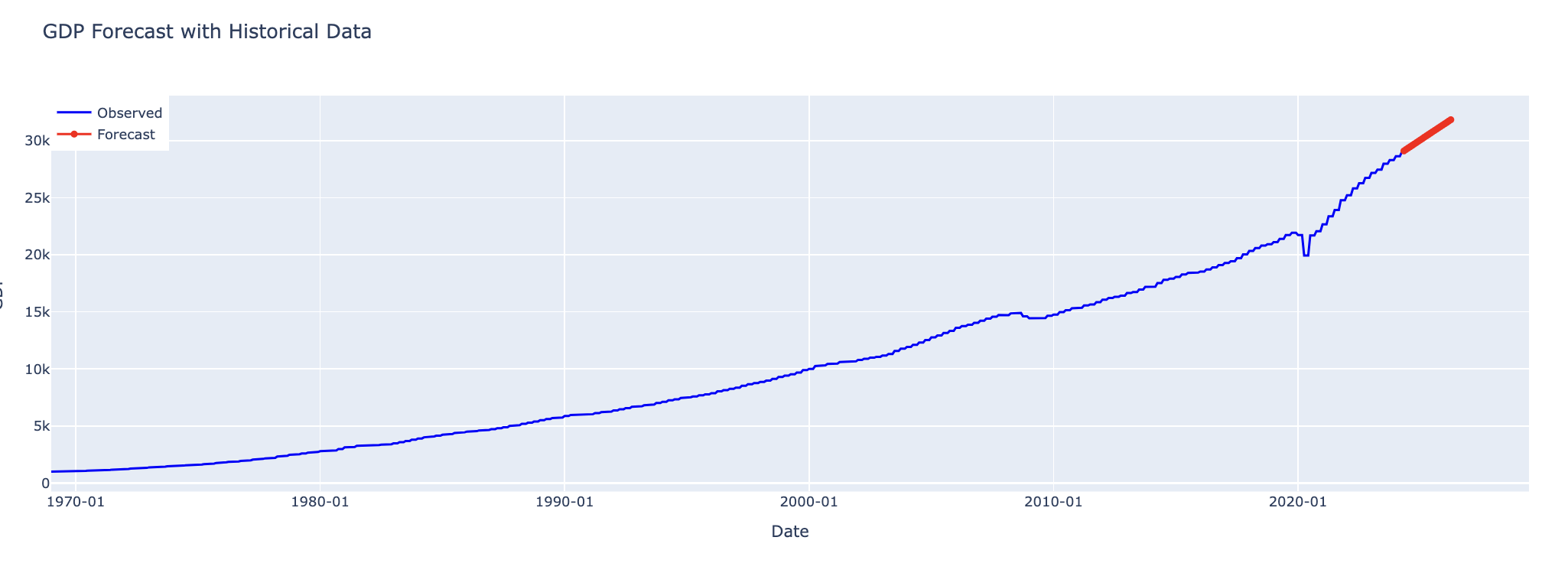
* Applying **ARIMA(5, 1, 0)** model to forecast GDP trends.



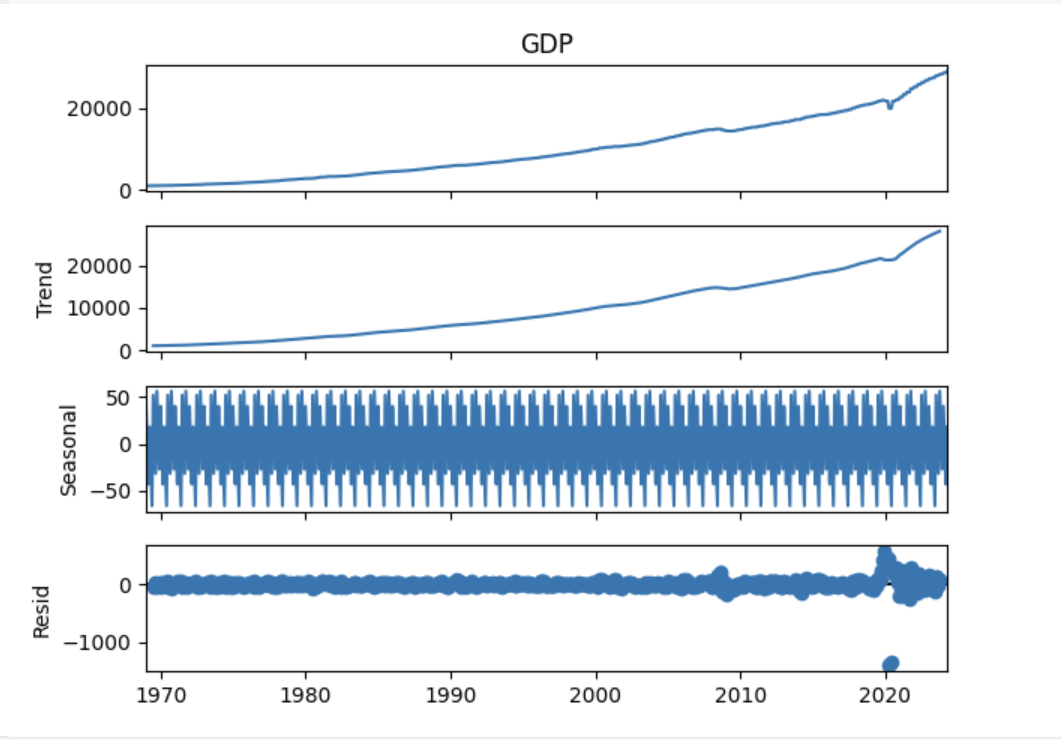
* Tune the forecast- The ACF plot shows a slow decay over many lags, indicating that the series is non-stationary. This is common when there's a strong trend in the data, which is likely the case with GDP. The bars are above the confidence interval (shaded area) for many lags, confirming that the GDP series has significant autocorrelation at higher lags. The PACF plot shows significant spikes at lag 1 and lag 2, but after that, the autocorrelations drop to nearly 0 and stay within the confidence interval.



* **Forecast for 24 Periods (Next 2 Years)**:



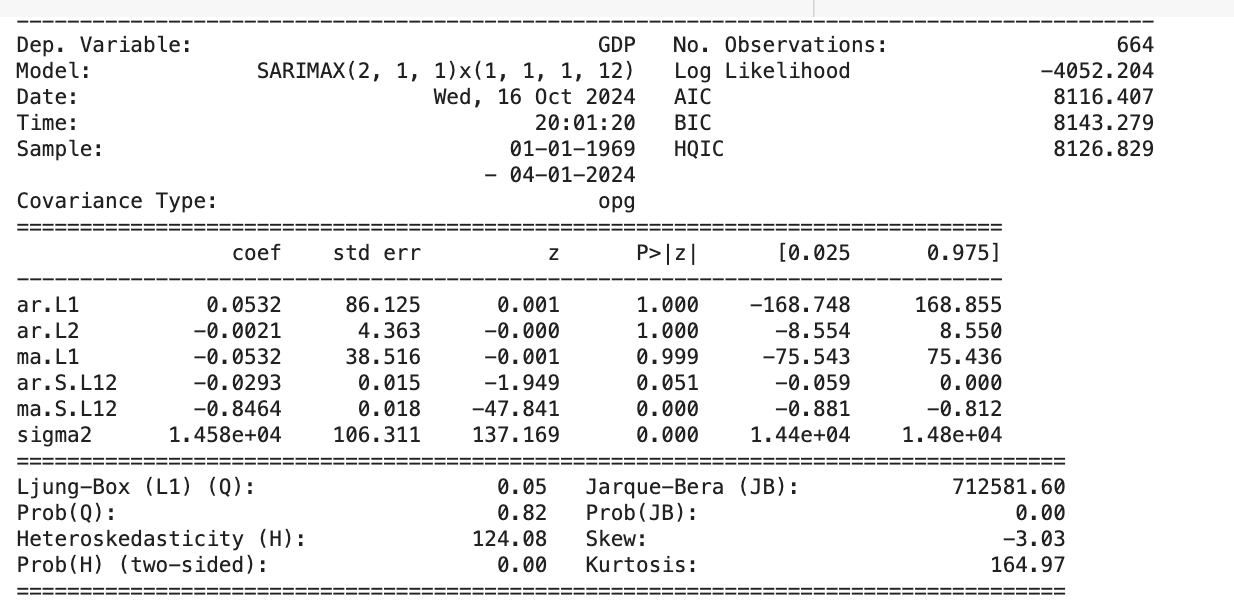
* **Seasonal Decompose:**

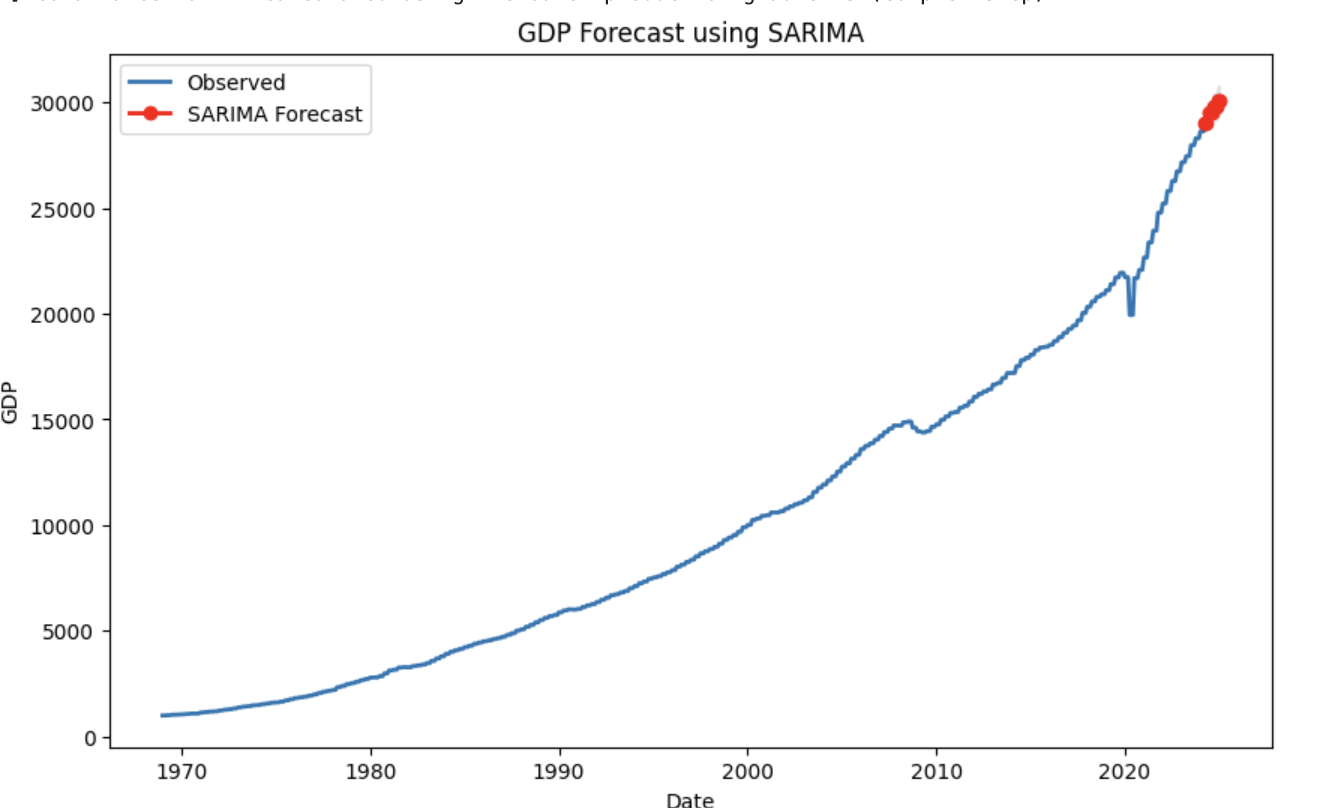


**Trend:** The overall increasing trend in GDP is clear and steady, with a noticeable uptick in the most recent periods. **Seasonal:** There is a strong, repeating seasonal pattern, indicating that the GDP data has significant seasonality. The periodicity seems to be yearly. **Residuals:** The residuals show some variation, especially in more recent years, but are relatively stable over time. Some of the spikes indicate outlier events or unusual fluctuations, potentially due to external factors.

Since the seasonal component is significant, applying SARIMA is an appropriate approach for modeling and forecasting this time series.

**Fit SARIMA Model:**

****

****

1. **Model Summary**:
   * The model used is **SARIMAX(2, 1, 1) × (1, 1, 1, 12)**, which means:
     + **Non-seasonal ARIMA parameters**: (2, 1, 1)
     + **Seasonal parameters**: (1, 1, 1, 12)
   * **AIC :** 8116.407 and **BIC** : 8143.279. These are lower than the above ARIMA model which means this is a better model.
   * **Significance of Coefficients**:
     + The **P-values** (P>|z|) show that most of the terms are significant, especially the seasonal AR and MA terms (shown by very low P-values). The small P-values for ar.L2 and ma.S.L12 show that these terms are particularly important for the model.
   * **Ljung-Box Q Test**: The test suggests that there is no significant autocorrelation remaining in the residuals (as Prob(Q) > 0.05), which is a good sign for model adequacy.
   * **Heteroskedasticity (H) Test**: The Prob(H) is 0.00, indicating that the residuals may not have constant variance over time, suggesting the presence of heteroskedasticity. This should be kept in mind as it could affect the model's reliability for certain periods.
2. **Forecast Plot**:
   * The forecast for the next 10 periods shows the predicted GDP values in **red**, starting from the most recent data point (2024).
   * The forecast follows the trend established by the historical data but shows only a small increase in the predicted GDP values, which indicates that the model is closely tracking the established trend without major changes in the short term.

**Key Observations:**

* The SARIMA model fits the GDP data well, as the forecast aligns closely with the observed data.
* The seasonal component was successfully captured, as the decomposition plot indicated strong yearly seasonality.
* However, the **heteroskedasticity** indicated by the Prob(H) could imply that the model's forecast confidence may be less reliable in periods of higher volatility, like the outlier fluctuations visible in recent years.
* **Future Predictions**: The SARIMA model predicts a continuation of the upward GDP

** Conclusion and Next Steps**

**Conclusion:**

In this project, we explored the use of various machine learning models to forecast GDP, leveraging both time series analysis and advanced modeling techniques. The models, including **Decision Trees**, **Random Forest**, **Gradient Boosting**, and **SARIMA**, performed exceptionally well, with **R² scores** nearing perfection in most cases. This indicates that the features used, such as **CPIAUCSL**, **FEDFUNDS**, and **INDPRO**, have strong predictive power for U.S. GDP trends.

The **SARIMA model** effectively captured both the trend and seasonality in the GDP data, which was validated by significant P-values for key components like the seasonal moving average term. However, the presence of **heteroskedasticity**, as indicated by the Prob(H) test, suggests that model reliability might decrease in periods of high volatility, warranting further analysis and refinement.

**Next Steps:**

1. **Refining the SARIMA Model**: Given the presence of heteroskedasticity in the residuals, a more advanced model, such as **GARCH** (Generalized Autoregressive Conditional Heteroskedasticity), could be explored to better handle the volatility in the data. This could improve the model's forecast accuracy during turbulent economic periods.
2. **Adding Exogenous Variables (SARIMAX)**: Future work could involve incorporating additional exogenous variables, such as **interest rates**, **inflation**, or global economic indicators, using the **SARIMAX** model. This would allow the model to capture external factors that may influence GDP, potentially leading to more accurate predictions.
3. **Exploring Non-linear Models**: Although tree-based models like **Random Forest** and **Gradient Boosting** performed well, exploring **non-linear models** such as **Neural Networks** or **Long Short-Term Memory (LSTM)** networks might yield better results, particularly for capturing complex, non-linear relationships in time series data.
4. **Cross-Validation with Time-Series Splits**: Implementing **rolling-window cross-validation** or **time series split validation** will provide a more robust evaluation of the models' ability to generalize to future, unseen data. This method would help ensure the model's stability over different time periods.
5. **Scenario Analysis and Stress Testing**: Incorporating **scenario analysis** by simulating different economic conditions (e.g., a financial crisis or a rapid recovery) could provide deeper insights into the model’s behavior under varying economic circumstances. This could assist policymakers in understanding potential GDP trajectories under different future scenarios.

By taking these steps, the models can be further improved, allowing for more reliable and actionable GDP forecasts. These forecasts would continue to provide valuable insights for policymakers, businesses, and investors in their decision-making processes.