ROBERT H. SMITH
SCHOOL OF BUSINESS

BUDT704: DATA PROCESSING AND ANALYSIS IN PYTHON

Univariate Time Series Forecasting and Analysis of Unemployment Rate

GROUP O - 507

Jasvinder Singh

Mentor

Prof. Peng Huang

INTRODUCTION

In the pursuit of unraveling the complex fabric of global economic trends, our project harnesses the power of cutting-edge technology, specifically employing univariate time series forecasting using ARIMA (Auto Regressive Integrated Moving Average) methodologies. By leveraging these advanced techniques, we aim to bring a higher level of precision and accuracy to our analysis of macroeconomic indicators, with a primary focus on forecasting unemployment rates.

Univariate time series forecasting allows us to distill intricate patterns from historical data, providing a lens into the evolving dynamics of unemployment over time. The ARIMA forecasting method, renowned for its effectiveness in modeling time-dependent data, enhances our predictive capabilities by capturing trends, seasonality, and underlying patterns that may elude traditional analyses.

In essence, our project stands at the intersection of economic analysis and technological innovation, utilizing univariate time series forecasting and ARIMA methodologies to illuminate the economic landscape and empower decision-makers with foresight in the face of global economic uncertainties.

BACKGROUND

Why Forecast unemployment rate?

The unemployment rate not only gives a measure of joblessness, but also is an indicator of economic growth. This is a lagging indicator and therefore is used to measure the impact of recession whether it is just beginning or is in the decline. It also provides confirmation of the state of the economy when evaluated in combination with other macroeconomic variables. When the unemployment rate increases, as it did during the last financial crisis in 2008, with an unemployment rate of 7 to 8%, which peaked at 10%, the government intervenes by stimulating the economy through a myriad of policy implementation including adding unemployment benefits, adding liquidity into the economy — 'quantitative easing', lowering interest rates, and, lowering tax rates to allow access to capital for households and businesses as well as introducing other government spending programs to increase employment opportunities.

This project will consume data from U.S. Bureau of Labor Statistics, Unemployment Rate https://fred.stlouisfed.org/series/UNRATE, [UNRATE], retrieved from FRED, Federal Reserve Bank of St. Louis; December 6, 2023, to churn out rich insights. We plan to use exploratory data analysis to take a deep dive into the world and regional economy. Eventually, we are aiming to devise recommendations for various stakeholders.

RESEARCH METHOD SUMMARY

Utilizing exploratory data analysis, our objective is to extract significant insights from the datasets. To achieve this, we will embark on a structured five-step methodology:

1. Data Wrangling Journey with World Bank's Unemployment Data:
   Embark on a detailed process of sorting and refining the global unemployment data sourced from the World Bank. Utilize Python's data manipulation capabilities to structure and prepare the data for in-depth analysis.
2. Preliminary Statistical Analysis of Unemployment Trends:
   Perform an initial statistical examination of the unemployment data. Use Python's statistical tools to extract basic insights, focusing on global unemployment patterns and variations.
3. Visualizing Unemployment Trends and Identifying Anomalies:
   Apply Python's graphical libraries to chart the trends in global unemployment. Analyze these visualizations to detect outliers and anomalies that deviate from typical unemployment patterns.
4. Preparing Models for Unemployment Forecasting:
   Develop predictive models tailored to forecast unemployment rates using Python. Fine-tune these models to ensure they accurately reflect and predict changes in global unemployment scenarios.
5. Summarizing Policy Recommendations Based on Unemployment Forecasts:
   Leverage the insights obtained from your analysis to formulate strategic recommendations aimed at addressing global unemployment issues. Highlight these recommendations, emphasizing their potential effectiveness based on your Python-driven economic forecasting.

DELIVERABLES

Addressing the complex and dynamic challenge of global unemployment is essential for economic stability and growth. Accurate forecasting and in-depth analysis are crucial in developing effective strategies to tackle unemployment. By employing advanced analytical algorithms, we aim to make significant strides in this endeavor.

Our insights have wide-ranging applications for various stakeholders:
Economic Analysts: Enhance understanding of unemployment trends for informed economic evaluations.
Government Agencies: Provide data-driven insights for effective policy making and labor market interventions.
International Organizations (like the World Bank and IMF): Offer a macroeconomic perspective for global economic planning and support.
Business Leaders and Entrepreneurs: Aid in strategic planning by understanding labor market dynamics.
Educational Institutions and Researchers: Supply foundational data for academic research and economic studies.
Public Policy Advocates: Inform advocacy and lobbying efforts with accurate unemployment data and forecasts.
Job Seekers and Workforce Development Agencies: Enable better career planning and workforce training programs through trend analysis.

OUR CHOSEN DATASET - EXPLANATION

This database encompasses four key attributes, each offering a unique perspective on the economic health of various countries. Unlike medical datasets, our focus is on macroeconomic indicators, specifically examining the relationship between unemployment rates and GDP across different nations and years.

Attribute Information:

1. Country Name: Identifies the country to which the data pertains, enabling a geographical analysis of economic trends.
2. Year: Indicates the year of the recorded data, allowing for a temporal analysis of economic changes over time.
3. Unemployment Rate: Expressed as a percentage, this attribute represents the proportion of the labor force that is unemployed but actively seeking employment. It is a critical indicator of economic health and labor market dynamics.
4. GDP (Gross Domestic Product): Measured in USD, it reflects the total monetary value of all goods and services produced within a country's borders in a specific year. GDP is a primary indicator of economic activity and growth.

By analyzing these attributes, we aim to uncover patterns and correlations between unemployment rates and GDP, providing valuable insights into the economic conditions of different countries over various years.

DATA SCRAPING AND CLEANING

For our project, we initiated the data extraction process by employing BeautifulSoup, a powerful Python library, for scraping relevant economic data. This approach enabled us to efficiently gather comprehensive unemployment and GDP data across different countries and years.

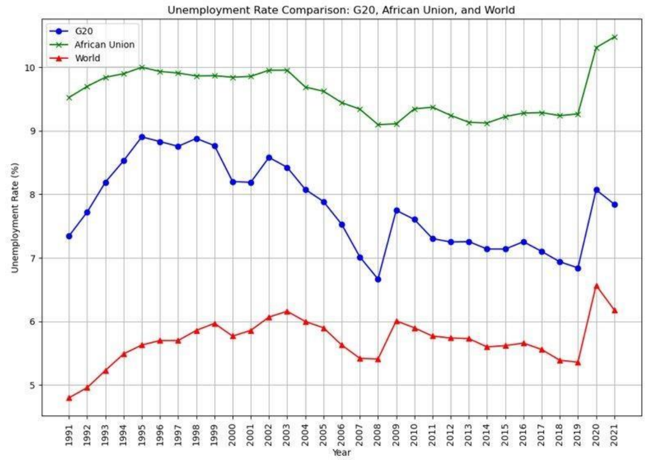
During the data cleaning phase, we took deliberate steps to enhance the dataset's accuracy and relevance. Recognizing the atypical economic conditions in certain years, we removed outliers that could skew our analysis. Specifically, we excluded unemployment data from the year 2008, acknowledging the global economic depression's impact, and from 2020, due to the unprecedented economic disruptions caused by the COVID-19 pandemic.

Additionally, we conducted a meticulous check for null values in the dataset. After identifying and removing these null values, we ensured the dataset's integrity and completeness. This rigorous cleaning process significantly contributes to the reliability of our analysis, allowing us to proceed with a dataset that accurately reflects consistent economic trends, unaffected by extraordinary global events.

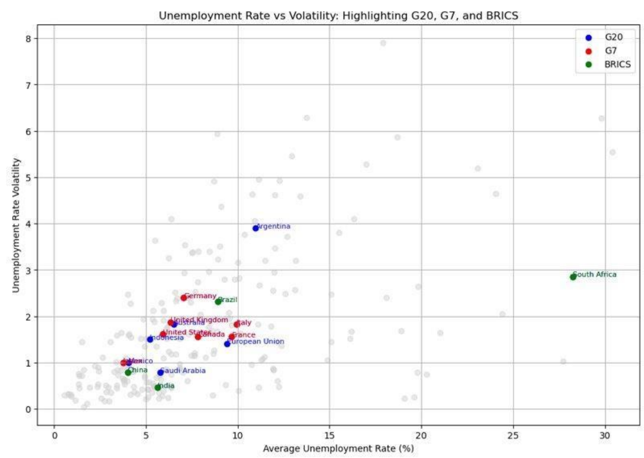
EXPLORATORY DATA ANALYSIS

1. We employed a line plot to compare the unemployment rates among the G20, the African Union, and the global average. This line plot effectively highlighted the varying unemployment trends across these economic groups. The data revealed a consistent pattern where the unemployment rates in the African Union were notably higher than those in the G20 countries, with both exceeding the global average.

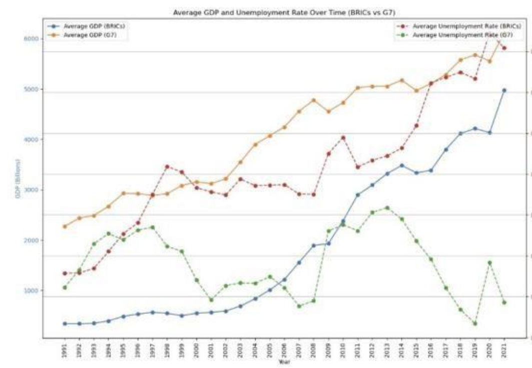
This trend underscores the significant, yet underleveraged, economic potential within the African Union. While acknowledging this, our analysis primarily focuses on the G20 nations due to their crucial role as key drivers in the global economy. By centering our attention on the G20, we aim to unearth deeper insights and understand the broader implications these trends may have on worldwide economic dynamics.



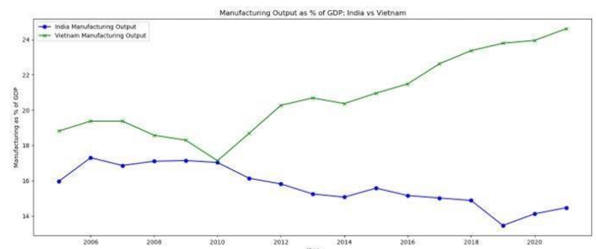
2. Our visualization comparing unemployment rates and volatility among G20, G7, and BRICS nations revealed a key insight: the G20 is representative of global trends. This implies that findings from the G20 can generally be extended to a global context. The analysis highlighted the varied economic landscapes of these groups, offering a comprehensive view of global unemployment patterns and their dynamics.

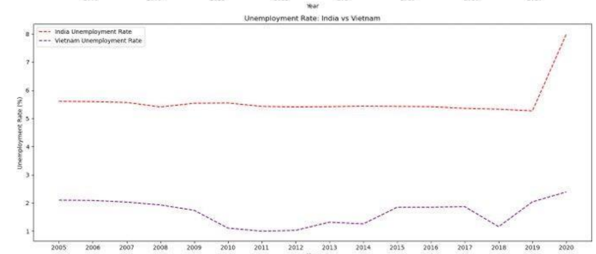


3. Our analysis of average GDP and unemployment rates over time in BRICS and G7 countries presents distinct economic narratives. In the G7, we observed a negative correlation between GDP and unemployment, suggesting that economic growth is accompanied by job creation. Conversely, BRICS nations exhibited a positive correlation, indicating that despite their robust economic expansion, job growth has lagged behind. This contrast underscores the necessity for tailored employment strategies in rapidly developing economies.

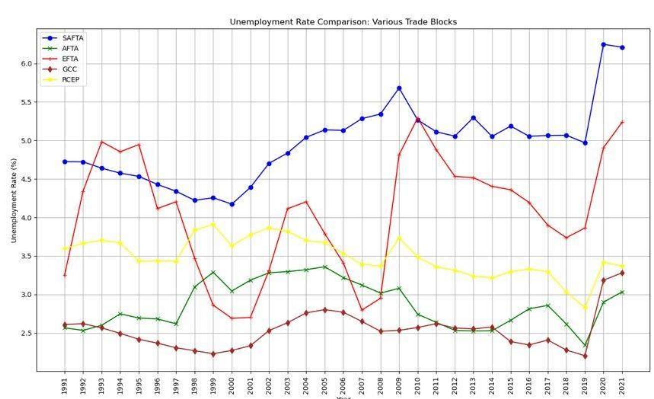


4. Vietnam's manufacturing output, a vital growth driver, is propelled by its low labor costs and strategic location near key South Asian trade routes. This synergy has led to a positive correlation between the growth of its manufacturing sector and employment stability, highlighting the sector's crucial role in bolstering the nation's economic and labor market health.



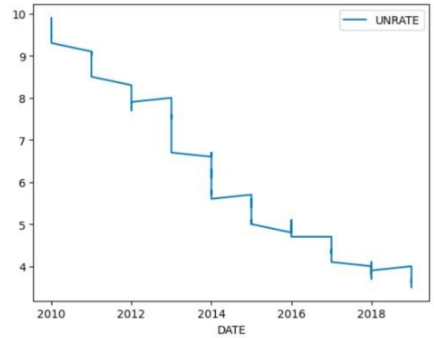


5. Our analysis of unemployment rates in different trade blocks underscores how regional trade acts as an engine for economic growth. It creates an ecosystem where businesses can explore and capitalize on new market opportunities. This surge in trade activities leads to an upswing in production and, consequently, employment. Such dynamics are instrumental in driving holistic economic development within these regions.

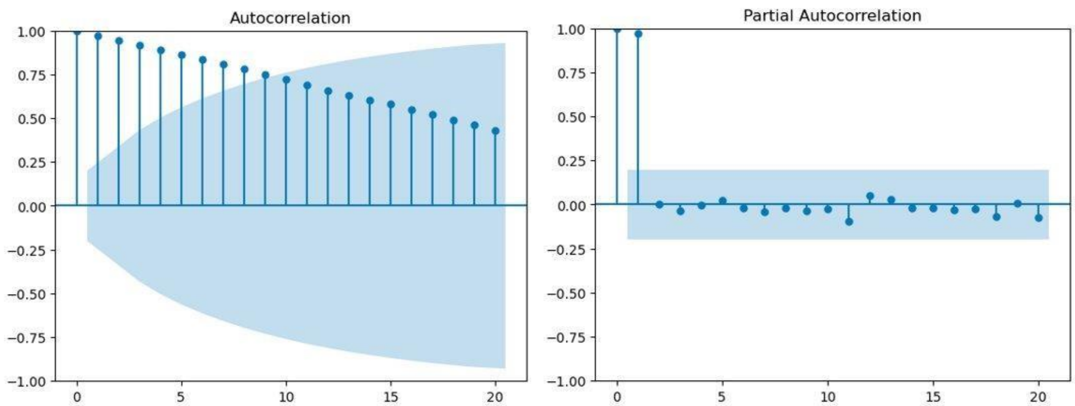


ARIMA

- Exploratory Data Analysis (EDA): We examine the time series data to understand trends, seasonality, and any other patterns. Our data set has a discernible downward trend.



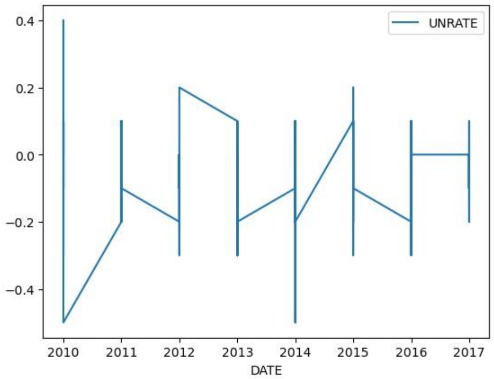
- Stationarity Check: We ensure the time series is stationary by checking for trends and seasonality. If non-stationary, perform differencing to make it stationary. We use autocorrelation and partial autocorrelation plot to check for stationarity in data set. ACF plot shows that correlation with the lags is high and positive with very slow decay. While PACF shows partial autocorrelations have a single spike at lag 1. This is sign of non-stationary data.

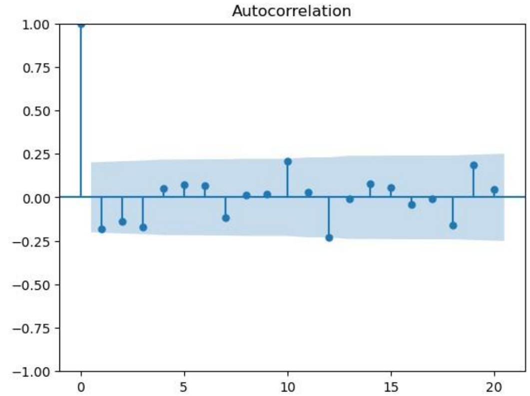


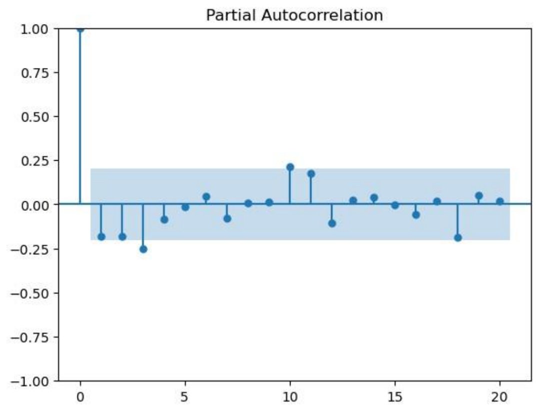
ADF test = 0.676812124847063

This further confirms that the data is non-stationary.

- Differencing: To make data ready for forecasting, we must remove trend. This can be done by differencing. Differencing our data once makes it stationary.







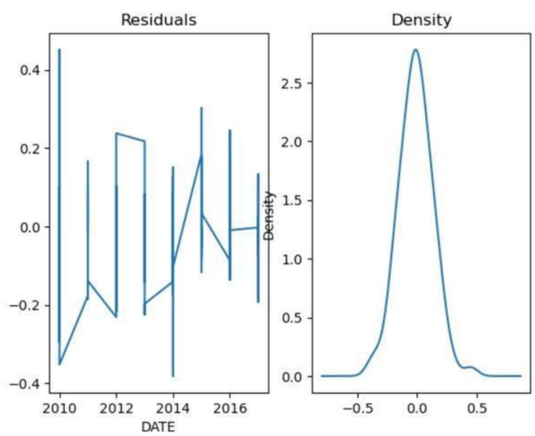
ADF test = 1.9204588733955935e-13

ACF, PACF and ADF test confirm that the trend from data is removed and now it is stationary.

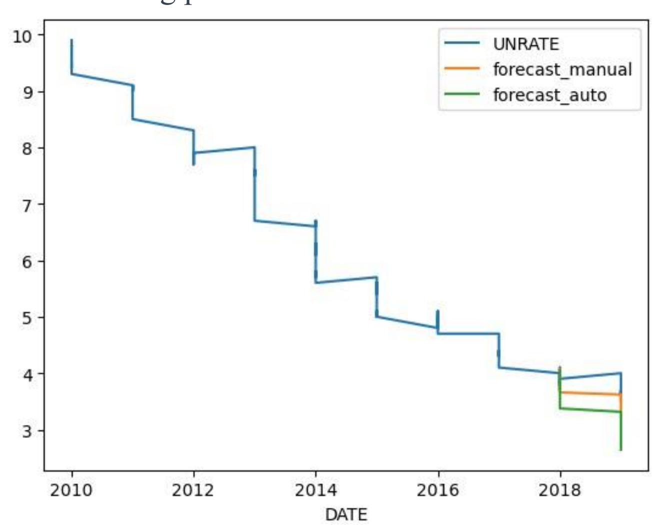
- **Estimation and Diagnostics**:
  1. **Parameter Selection**:
- From differencing, we determine the value of d = 1
- We determine the hyperparameters p, q using goodness of fit measure AIC. We choose the combination of p and q which minimizes AIC.
  2. **Data Splitting**: We reserve a portion (20%) of the time series data for testing the accuracy of the ARIMA model.
- **Building the model:** We use the identified parameters to train the ARIMA model on the training set.

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                 UNRATE   No. Observations:                   96
Model:                 ARIMA(1, 1, 3)   Log Likelihood                  51.488
Date:                Thu, 07 Dec 2023   AIC                            -92.976
Time:                        03:29:23   BIC                            -80.207
Sample:                             0   HQIC                           -87.816
                                 - 96
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9934      0.012     83.020      0.000       0.970       1.017
ma.L1         -1.2849      0.087    -14.753      0.000      -1.456      -1.114
ma.L2          0.1090      0.150      0.728      0.467      -0.185       0.403
ma.L3          0.2213      0.111      1.997      0.046       0.004       0.439
sigma2         0.0194      0.003      7.513      0.000       0.014       0.025
===================================================================================
Ljung-Box (L1) (Q):                   0.06   Jarque-Bera (JB):                 1.33
Prob(Q):                              0.81   Prob(JB):                         0.52
Heteroskedasticity (H):               0.60   Skew:                             0.14
Prob(H) (two-sided):                  0.15   Kurtosis:                         3.50
===================================================================================
```

- **Model Validity:** Our model's reliability hinges on two critical assumptions: firstly, the data should exhibit stationarity, and secondly, the residuals should adhere to randomness and a normal distribution. A thorough examination of the residual plots confirms that these assumptions are satisfied, validating the model's appropriateness for predictive purposes.



- **Forecasting:** We use trained ARIMA model to make predictions on unseen test data. While employing the 'auto_arima' feature for automatic parameter determination in ARIMA models is common, we observed superior outcomes by manually selecting parameters for our trained ARIMA model when making predictions on unseen test data.



- **Model Evaluation:** We assess the model's performance on the test set using metrics like Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE).

```
mae - manual: 0.1706671111922954
mape - manual: 0.04568193844131607
rmse - manual: 0.2078108340934406

mae - auto: 0.44591402296272875
mape - auto: 0.12003643089765366
rmse - auto: 0.5335821307744302
```

The results indicate that the manually derived predictions exhibit lower errors and better accuracy when compared to the automated predictions, emphasizing the effectiveness of the manual approach.

Limitations

1. **Limited Feature Set:** The selected dataset uses only historic unemployment rate data for forecasting. Going forward other macroeconomic variables such as GDP, Inflation, Federal Funds Rate, Money supply, etc will be included for multivariate forecasting to achieve better results.

2. **Exclusion of extreme events:** While restricting the analysis to the period between 2010 and 2019 provides a stable context for ARIMA forecasting, it inherently excludes extreme economic events such as the 2008 financial crisis and the 2020 COVID-19 pandemic. This approach assumes that the underlying economic conditions during this period are representative of future scenarios. However, unforeseen events can significantly impact the unemployment rate, and the model may not fully capture the dynamics associated with such extreme occurrences. Users should interpret forecasts with the awareness that the model's predictive accuracy might be less reliable during unprecedented economic disruptions.

Future Scope

- Sectoral Analysis: Different sectors of the economy may experience varying levels of unemployment. Forecasting unemployment allows for a more granular analysis of the impact on specific industries and occupations.
- While ARIMA models may not directly predict the occurrence of a recession, they can be valuable in forecasting economic indicators (such as GDP, unemployment rates) and understanding trends in the data.
- Multivariate Forecasting: Going forward other macroeconomic variables such as GDP, Inflation, Federal Funds Rate, Money supply, etc will be included for multivariate forecasting to achieve better results.