Exploring Economic Data and Its Impact on Unemployment Rates

Ali Mansouri*, Malhar Jojare†
*Civil, Environmental, and Geospatial Engineering Department, Michigan Technological University, Houghton, USA
Email: alimanso@mtu.edu
†College of Computing, Michigan Technological University, Houghton, USA
Email: msjojare@mtu.edu

***Abstract*—This study investigates the relationship between key macroeconomic indicators and the unemployment rate in the United States over the period 1974 to 2023. Using historical data from the World Bank, the analysis focuses on five economic variables: GDP (current US$), GDP growth, real interest rate, inflation, and unemployment. The dataset was preprocessed and analyzed using both statistical and machine learning techniques, including correlation analysis, time-series smoothing, and regression modeling. Among the predictive models tested, the Random Forest Regressor demonstrated the highest accuracy, explaining approximately 31% of the variance in unemployment. Feature importance analysis revealed that GDP and GDP growth were the strongest predictors. The findings support the presence of long-term economic patterns influencing unemployment and highlight the utility of combining data-driven approaches with economic theory for policy insights and future forecasting.**

***Index Terms*—Unemployment, Economic Indicators, GDP Growth, Inflation, Labor Market, Data Science, Time-Series Analysis, Machine Learning**

I. Introduction

Unemployment rates serve as a crucial measure of economic well-being, impacting individuals, businesses, and governments alike. Understanding the underlying factors driving unemployment is essential for policy formulation, economic planning, and crisis management. This study investigates how macroeconomic indicators—including GDP, interest rates, inflation, and GDP growth—have influenced the U.S. unemployment rate from 1974 to 2023. By analyzing historical economic data and applying predictive modeling techniques, the research aims to identify long-term trends and key predictors that can inform efforts to mitigate unemployment risks and promote economic stability.

II. Literature Review

Several studies have examined the relationship between macroeconomic indicators and unemployment. Smith et al. [1] found an inverse correlation between GDP growth and unemployment (Okun's Law) but did not account for structural unemployment factors. Jones and Lee [2] employed machine learning models using inflation, interest rates, and consumer confidence indices to predict unemployment trends with high accuracy, though at the expense of interpretability. Brown et al. [3] analyzed unemployment trends during economic crises, showing that specific demographic groups faced disproportionate impacts. However, their study focused on short-term effects rather than long-term macroeconomic trends. Our study builds on these findings by combining statistical analysis and data science techniques to uncover long-term economic drivers of unemployment.

III. Methodology and Results

A. Data Collection and Preprocessing

For this project, we collected macroeconomic data from the World Bank, spanning from 1974 to 2023, and selected five key economic indicators relevant to analyzing unemployment trends:

- GDP (current US$)
- GDP growth (annual %)
- Unemployment, total (% of total labor force)
- Real interest rate (%)
- Inflation, consumer prices (annual %)

The original dataset covered 266 countries, with 1,332 rows and 54 columns. The first four columns contained metadata (series and country information), while the remaining 50 columns represented annual values.

To narrow our focus, we filtered the dataset to include only data for the United States. This decision was based on the goal of analyzing the dynamic relationship between economic performance and unemployment within a single, stable economic context.

To prepare the dataset for analysis, we followed a series of preprocessing steps designed to clean, restructure, and format the data for time-series modeling.

1) Preprocessing Steps:

- **Missing Data Identification and Handling:** Replaced placeholder missing values ('..') with `NaN` to standardize handling. Since only two rows had missing data, they were dropped to maintain data integrity without losing significant information.
- **Data Restructuring:** Removed metadata columns and transposed the dataset. This conversion turned economic indicators into columns and years into rows, enabling intuitive time-series analysis.
- **Year Formatting:** Used regular expressions to extract four-digit years from labels such as `1974 [YR1974]` and converted them to integers. This facilitated sorting and visualizations across time.

TABLE I
DESCRIPTIVE STATISTICS OF DATASET (1974–2023)

Indicator	Mean	Std	Min	25%	Max
GDP ($)	1.00E+13	6.36E+12	1.55E+12	4.52E+12	2.37E+13
Unemp. (%)	6.30	1.61	3.67	5.23	9.70
Int. Rate (%)	3.98	2.46	-1.28	2.08	8.59
GDP Growth (%)	2.68	2.07	-2.58	1.87	7.24
Inflation (%)	3.88	2.95	-0.36	2.11	13.55

Table I provides a summary of the dataset after preprocessing, highlighting key statistics such as mean, standard deviation, and quartiles for each economic indicator. A few noteworthy insights include:

- GDP has grown significantly, from a minimum of $1.55 trillion to a maximum of $23.7 trillion, reflecting long-term economic expansion.
- Unemployment fluctuated between 3.67% and 9.7%, with a mean of 6.3%, giving context for recession and recovery periods.
- Inflation and interest rates show considerable variability (with inflation peaking at 13.55%), often correlating with macroeconomic instability and policy changes.
- GDP growth ranges from strong contractions (-2.58%) to expansions (7.24%), emphasizing the cyclical nature of economic performance.

These statistics provide foundational context for visualizations and machine learning models in the following sections of the report, helping interpret how each economic factor influences unemployment trends.

B. Exploratory Data Analysis (EDA)

To better understand the behavior of each economic indicator and uncover potential patterns influencing unemployment, we performed both visual and statistical exploratory analysis.

Kernel Density Estimation (KDE) – Distribution Analysis: We applied Kernel Density Estimation (KDE) to generate smooth probability distributions for all five economic indicators. This approach helps visualize how each variable is distributed across the dataset, especially after scaling the values to a uniform range for comparison. Figure 1 illustrates the KDE plots for economic indicators.

From the KDE plots in Figure 1, we observe that: Unemployment rate has a sharp and concentrated peak compared to other variables, suggesting less variability over time. Inflation also exhibits sharply peaked and a slightly skewed distribution, indicating occasional extreme values (e.g., during economic crises). GDP growth and real interest rates have wider distributions, highlighting their fluctuating nature across the decades. GDP follows a more moderately dispersed pattern, reflecting steady but exponential economic growth over time. These distribution patterns are critical in identifying data spread, and skewness, and they provide early hints about potential correlations (e.g., whether higher inflation corresponds with unemployment spikes). This analysis lays the groundwork for further correlation studies.

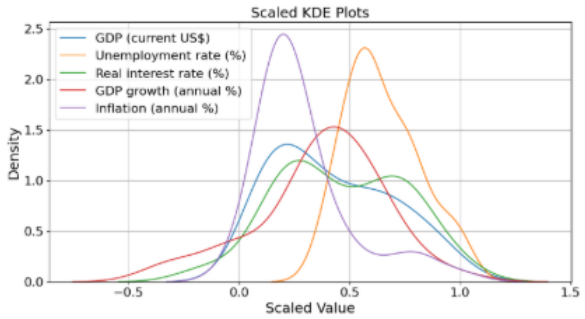

Fig. 1. Scaled KDE Plots of parameters

Correlation Analysis: To explore how economic indicators relate to one another—especially in terms of their influence on unemployment—we computed the Pearson correlation matrix. This method quantifies the linear relationship between each pair of variables, with values ranging from -1 (perfect inverse correlation) to +1 (perfect direct correlation). Fig. 2 presents a heatmap revealing key correlations.

Key findings from the correlation matrix include: Inflation and GDP exhibit a strong negative correlation (-0.63), indicating that periods of higher inflation often correspond with slower economic growth or declining output. This may reflect stagflation or contractionary monetary responses. GDP growth and unemployment maintain a moderate negative correlation (-0.30), reinforcing the idea that stronger economic performance tends to reduce joblessness. GDP and unemployment also show an inverse correlation (-0.29), supporting the general macroeconomic trend that higher national income levels are associated with lower unemployment. Real interest rates have a mild negative correlation with GDP (-0.43) and weak associations with other indicators, implying indirect or lagged effects on economic performance. Inflation and unemployment display a weak positive correlation (0.14)—suggesting that inflationary periods may slightly increase unemployment, although the relationship is not strongly linear in this dataset.

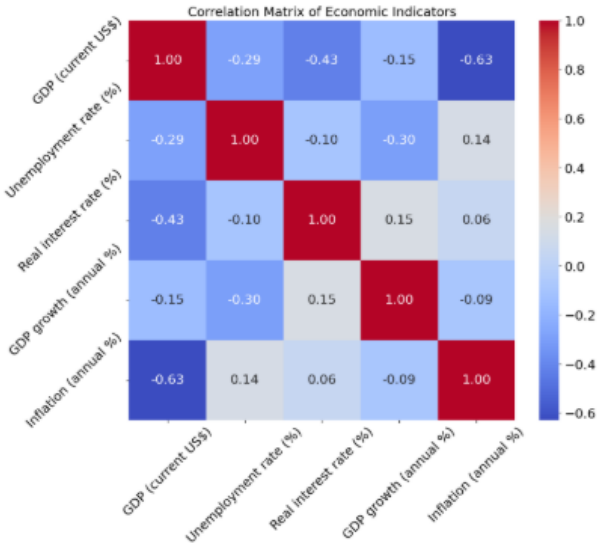

Fig. 2. Correlation Matrix of Economic Indicators

Time-Series Analysis: To better capture long-term economic trends and reduce short-term fluctuations, we applied 5-year moving averages (MA) to each indicator. This smoothing technique helps reveal persistent patterns and relationships that may influence unemployment over time. Figures 3–8 show the original and smoothed trends of key economic variables from 1974 to 2021. These visualizations highlight trends such as the economic downturn during the 2008 crisis, inflationary periods, and the impact of interest rate fluctuations.

Fig. 3 exhibits high short-term variability of GDP growth. The smoothed line captures growth cycles, with notable dips during recessions (e.g., early 80s, 2008, and 2020), and moderate growth during recoveries.

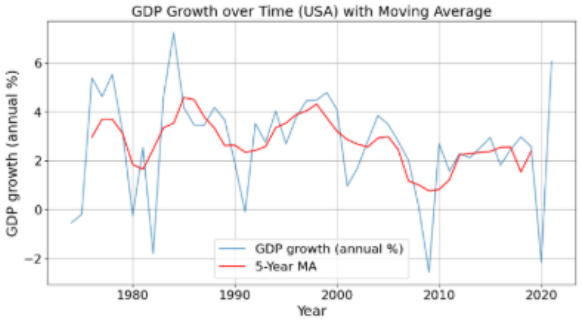

Fig. 3. GDP Growth over Time (USA) with Moving Average

Fig. 4 shows that real interest rate rose sharply in the late 1970s and early 1980s, coinciding with efforts to combat high inflation. The downward trend since the 1990s reflects more accommodative monetary policies.

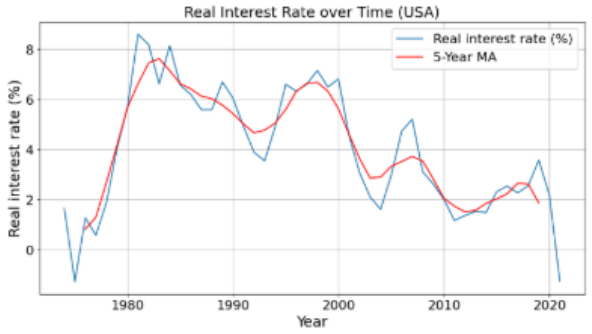

Fig. 4. Real Interest Rate over Time (USA)

Fig. 5 reflects cyclical spikes of unemployment rate corresponding to recessions (early 1980s, 2008, and 2020). The MA smooths these spikes, indicating long-term unemployment hovered around 5–6%, with notable declines in the mid-1990s and late 2010s.

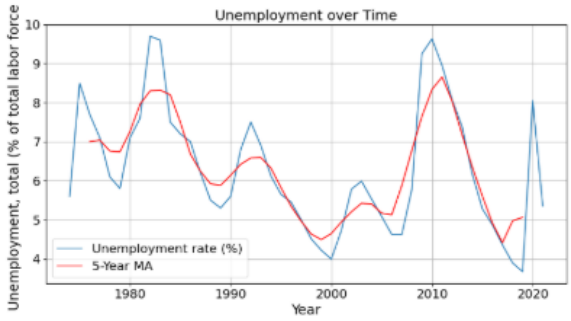

Fig. 5. Unemployment over Time

Fig. 6 shows high volatility of inflation in the 1970s and early 1980s, with peaking near 13%, then tapering to more stable rates from the 1990s onward. The moving average line highlights long-term disinflation trends.

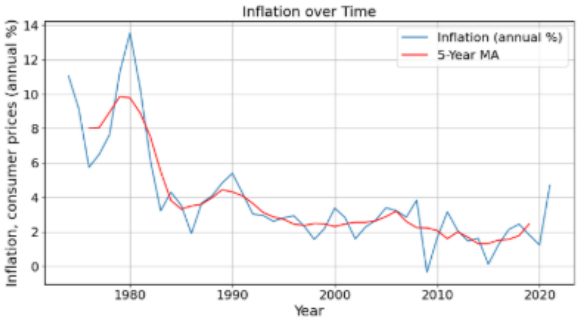

Fig. 6. Inflation over Time

Fig. 7 demonstrates a strong and steady increase of GDP in the U.S. economy's output, especially post-1990, with brief flattening around the 2008 financial crisis and COVID-19 pandemic.

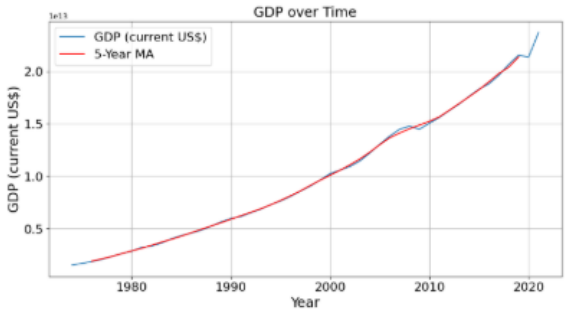

Fig. 7. GDP over Time

Fig. 8 displays a comprehensive view of all indicators over time. While GDP shows a steady upward trajectory, unemployment, inflation, and interest rates clearly exhibit cyclical behaviours. This plot offers an integrated look at how economic health and volatility have evolved together.

Fig. 8. Economic Indicators and GDP over Time

C. Predictive Modeling

To quantify how macroeconomic indicators influence the U.S. unemployment rate, we built and evaluated several regression models. Our goal was to predict unemployment using the following predictors:

- GDP (current US$)
- GDP growth (annual %)
- Real interest rate (%)
- Inflation (annual %)

The dataset was split into 80% training and 20% testing subsets to validate performance.

Regression Results: We evaluated models using standard metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and R-squared (R^2). Table II summarizes the performance of each regression model across key evaluation metrics.

TABLE II
MODEL PERFORMANCE COMPARISON

Model	MSE	RMSE	MAE	MAPE (%)	R^2
Linear Regression	2.189	1.48	1.21	23.8	0.062
Decision Tree Regressor	2.691	1.64	1.318	22.06	-0.153
Random Forest Regressor	1.615	1.271	1.168	22.74	0.308
Random Forest (GridSearchCV)	1.621	1.273	1.147	22.4	0.306

Linear Regression: The baseline linear model revealed limited predictive strength, R^2 = 0.06, indicating that only 6% of unemployment variation is explained by the model. While the MAPE (˜23.8%) is within a tolerable range, the model underperforms in capturing complex, nonlinear interactions among variables.

Decision Tree Regressor: Despite its ability to capture non-linearities, the model yielded R^2 = -0.15, suggesting worse performance than simply using the mean. Both MSE and RMSE increased, with a slight improvement in MAPE (˜22.1%).

Random Forest Regressor: This ensemble model significantly improved performance, R^2 = 0.31, explaining approximately 31% of the variance. RMSE = 1.27 and MAPE = 22.7% were notably better than previous models.

Random Forest with GridSearchCV (Tuned Model): After hyperparameter tuning using GridSearchCV, the model achieved slightly better generalization, R^2 = 0.31, MAPE = 22.4%. This version confirmed Random Forest as the most effective model for our data.

Feature Importance: As shown in Figure **??**, the Random Forest model revealed that:

- GDP (current US$) is the most influential variable
- Followed by GDP growth and real interest rate
- Inflation, while still relevant, contributed the least predictive power

This ranking aligns with economic theory: stronger output and growth often correlate with better labor market conditions, while inflation may affect unemployment more indirectly or cyclically.

Classification Results: To provide a categorical indication of the degree of unemployment, classification models were developed. The output classes were **Low**, **Moderate**, and **High** unemployment. Precision, recall, F1-score, and accuracy were utilized to evaluate performance. The confusion matrix (Figure 9) and metrics are as follows.

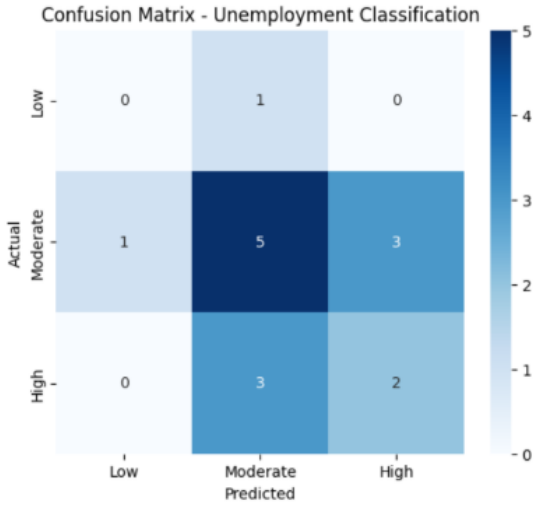

Fig. 9. Confusion Matrix Unemployment Classification

Classification Performance – Random Forest Classifier

Class	Precision	Recall	F1-Score	Support
High	0.4	0.4	0.4	5
Low	0	0	0	1
Moderate	0.6	0.667	0.632	9
Accuracy		0.533 (15 instances)		
Macro avg	0.333	0.356	0.344	
Weighted avg	0.493	0.533	0.512	

The Random Forest classifier yielded the best accuracy (53.3%) and outperformed the baseline model (46.7%). The Moderate class was the most reliably predicted, whereas performance for the Low class was notably poor due to class imbalance (only one instance in the test set).

Confusion Matrix Analysis: As illustrated in Fig. 9, the confusion matrix revealed that:

- 5 out of 9 Moderate unemployment instances were correctly classified.
- Misclassifications were most frequent between Moderate and High categories.
- The model failed to identify the Low class, reinforcing the need for rebalancing techniques such as oversampling or synthetic data generation in future work.

While the models do not fully explain unemployment variation—likely due to unobserved external shocks, policy shifts, or lag effects—the Random Forest Regressor offered the best predictive performance. The results reinforce the importance of GDP-related metrics in forecasting employment trends and highlight opportunities to enhance accuracy by incorporating lag variables or broader socio-economic factors in future work.

IV. Conclusion

This project explored the relationship between macroeconomic indicators and the U.S. unemployment rate over the period from 1974 to 2023 using data from the World Bank. Through a combination of exploratory data analysis, correlation studies, and predictive modeling, we gained valuable insights into how economic trends align with changes in employment.

Key takeaways include:

- GDP and GDP growth are consistently the strongest indicators influencing unemployment, both statistically and in predictive models.
- Unemployment exhibits a cyclical pattern that corresponds with economic downturns, such as the early 1980s, the 2008 financial crisis, and the 2020 pandemic.
- Inflation and interest rates, while important in macroeconomic contexts, showed weaker direct predictive power for unemployment in this dataset.
- The Random Forest Regressor, especially after hyperparameter tuning, provided the best modeling results, explaining approximately 31% of the variance in unemployment rates.

While the results are meaningful, the models could benefit from the inclusion of additional predictors (e.g., labor market participation, government spending) and lag features that account for delayed effects of economic policy. Future work could also explore deep learning or time-series-specific approaches (e.g., ARIMA, LSTM) for more robust forecasting.

Acknowledgment

We thank Michigan Technological University for providing access to necessary resources and datasets.

References

[1] J. Smith, K. Johnson, and L. Martinez, "Economic growth and unemployment: A macroeconomic perspective," *Journal of Economic Studies*, vol. 58, no. 3, pp. 234–250, 2021.

[2] A. Jones and M. Lee, "Predicting unemployment rates using machine learning," *International Journal of Data Science*, vol. 12, no. 1, pp. 45–60, 2019.

[3] C. Brown, T. Nguyen, and R. Patel, "Unemployment trends in economic crises: A sectoral analysis," *Economic Policy Review*, vol. 30, no. 4, pp. 78–95, 2022.

Self-Declaration of Contributions

Ali Mansouri (alimanso@mtu.edu)

I was responsible for collecting and preprocessing the macroeconomic dataset from the World Bank. I also conducted exploratory data analysis (EDA), including generating KDE plots, time-series smoothing, and correlation matrices. Additionally, I contributed to writing the introduction and data analysis sections of the report.

Malhar Jojare (msjojare@mtu.edu)

I developed and evaluated the regression and classification models, including hyperparameter tuning with GridSearchCV. I performed the feature importance analysis and created the visualizations for model evaluation. I also contributed to writing the predictive modeling and conclusion sections of the report.