AISE 4010 Final Project Report

Group 1: Dev Patel, Chris Latosinsky

Introduction

Economic forecasting is a critical tool for policymakers, investors, and business to navigate uncertainties in economic cycles. Especially in today's economic climate, where inflation and the cost of goods are rising, the ability to understand and predict economic shifts is more important than ever. Accurate forecasting provides insights that enable proactive decision-making, helping to mitigate risks and prepare for potential economic recession. This project focuses on predicting key U.S. economic indicators, including GDP percent change, inflation percent change, unemployment percent change, and various yield data features, to allow for early classification of a recession.

Datasets

The datasets used in this project were carefully selected to capture key economic indicators essential for predicting recessions. These include GDP data [1], inflation data [2], unemployment data [3], and yield curve data [4]. Together, these datasets provide a good sense of the economic state in the US. GDP measures economic output and growth, while inflation tracks changes in purchasing power. Unemployment data reflects labor market health, and yield curve data, particularly the slope, has historically served as a reliable predictor of economic downturns according to many economists. Finally, a recession dataset [5], used as the target for classification, enables early identification of economic recessions. Together, these datasets contribute to the project's goal of predicting and classifying recessions.

Data Preprocessing

Preprocessing was a crucial step in ensuring the data was ready for modeling. All five datasets had missing values, which could negatively impact model performance. Additionally, the datasets had different sampling frequencies, making it challenging to align them for analysis and modeling. To address these issues, all datasets were resampled to a common monthly frequency and linear interpolation was used to fill all missing values while maintaining data trends.

The yield curve data required additional preprocessing to fill the first ten years of the dataset for the 10-year zero coupon yield (ZCY) and instantaneous forward rate (IFR). These values are important for understanding long-term economic expectations. The ZCY reflects investor sentiment and long-term interest rate predictions, while the IFR provides insights into the expected short-term interest rate at any given future point. Filling these missing values using linear interpolation, forward fill, or backward fill would not accurately capture the intricate relationships and temporal trends in the data. Instead, a variational autoencoder (VAE) with temporal convolutional network (TCN) layers was used to generate values based on the observed patterns in the existing data. This was achieved by flipping the dataset to position the missing values at the end, enabling the VAE to learn from the available data, and training the model on the relationships between the existing 10-year ZCY and IFR data and the other features in the yield data. From there, the VAE could generate new data points for the missing decade of 10-year ZCY and IFR to complete the dataset.

Exploratory Data Analysis and Feature Selection

After preprocessing all the datasets, feature selection was the next step to reduce the dimensionality of the data while preserving the most relevant information for modeling. If all the datasets were combined without filtering, the resulting dataset would include over 70 features, making it highly complex and computationally intensive to process. High-dimensional data increases the risk of overfitting and could reduce the model's ability to generalize to unseen data. To address this, a combination of manual evaluation and statistical analysis was used.

First, irrelevant or redundant features were removed, such as unemployment rates broken down by gender and age groups, which were deemed unnecessary for this project. After refining the GDP, inflation, and unemployment datasets, a correlation matrix was applied to mainly analyze the yield data (Figure 1). This analysis revealed that many yield features were highly correlated. As a result, the slope, long-term rate, 1-year ZCY, and 10-year ZCY were selected as they exhibited the most meaningful variation and the least redundancy.

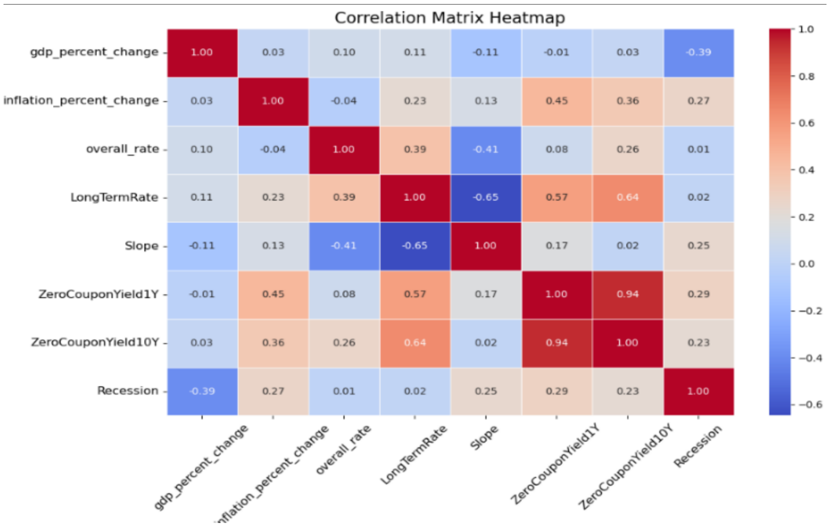

Figure 1: Correlation Matrix

Once the features of interest were selected, it was necessary to align the datasets' date ranges to ensure consistency across all data points. Each dataset started and ended at different dates, which would have created gaps in the final dataset if not addressed. To resolve this, the datasets were clipped to a common period based on their overlapping date ranges, resulting in a start date of 1967-11-30 and an end date of 2023-06-30. The clipped datasets had their features of interest extracted and combined into a new merged dataset ready for modeling.

Seasonal decomposition was done on the selected features to separate them into trend, seasonal, and residual components. This helped identify seasonal fluctuations, long-term trends, and random variations in the data. The complex temporal dependencies and periodic patterns indicated that a deep learning model would be more suited for handling the data than a traditional time series machine learning model such as ARIMA.

Economic Indicator Prediction

A Long Short-Term Memory (LSTM) network was selected to perform multivariate prediction on the merged dataset due to its strong performance at modeling long term dependencies in complex data. The merged dataset was normalized using MinMaxScaler to have all features on an equal scale. The recession column was also dropped since it will only be used in the classification task. A sliding window of size 12 was used to get a full year's worth of data before predicting the next month. The merged data was split into training and testing sets. The training dataset was set from 1967 to the end of 2004 and the testing dataset from 2005 to 2023. This split would allow for observation of the model's performance during notable recessions which occurred in 2008 and 2020.

The LSTM structure contains two LSTM layers, two dropout layers, and an output layer. Hyperparameter tuning using RandomSearch was done to find the optimal units of 256, dropout rate of 0.1, and learning rate of 0.001. Early stopping with a patience of 15 was also used to prevent the model from overfitting and saving computational resources. The trained model produced a training loss of 0.0048 and a validation loss of 0.0075 which indicated the model was able to model the data and generalize well (Figure 2).

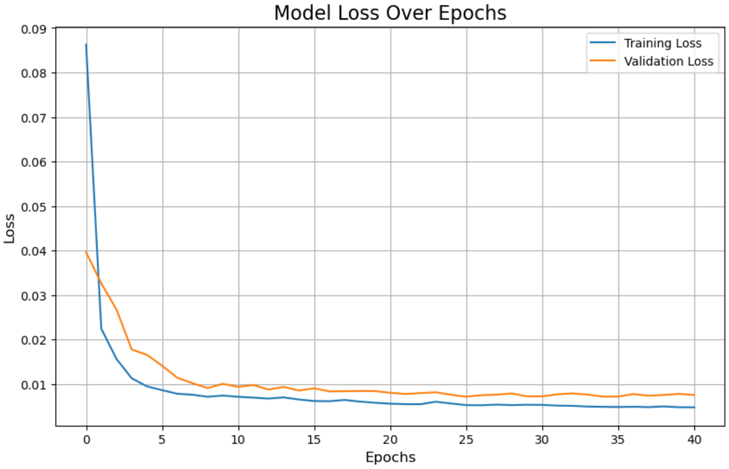

Figure 2: Training and Validation Loss Over Epochs for LSTM Prediction

LSTM Prediction Results and Model Improvement

The resulting graphs (Figure 6 in Appendix) show the plots of the actual data values and the predicted values over time. The model demonstrates the ability to capture the complex trends and relationships between the features as shown by the alignment of the curves. However, the model seems to deviate in highly volatile regions, most notably in 2020, which indicates that it struggles to account for sudden spikes and dips. This behaviour makes sense considering the model does not have information on real world events. This could potentially be addressed by incorporating additional data features that influence economic volatility, such as indicators derived from global economic news or major policy changes. For instance, combining the LSTM with a natural language processing (NLP) model trained to extract sentiment or key economic events from news articles could provide real-time insights into external factors driving volatility. This ensemble approach could enhance the overall predictive capabilities of the model, especially during period of rapid change.

Recession Classification

The first step is to preprocess the data to better suit this task. Since GDP, Inflation Rate, and Unemployment Rate were all converted to percent change, these values exhibited stability and had a somewhat normal distribution, so these values were left unchanged. The Slope and LongTermRate were standardized since we cared about negative values. Finally, the ZCY1 & ZCY10 were scaled using MinMaxScaler using the global min and max values, this is because we cared about the relative difference in these two values. Also, scaling was performed on training sets to prevent data leakage.

The first approach to this problem was using a Hidden Markov Model (HMM) to predict the hidden states (peak, contraction, trough, and expansion), however since we were unable to find labeled data for these 4 states, there was no way to tell if these predictions were accurate. Instead, we switched to 2 states identifying recession or no recession [5]. We used a GaussianHMM with 2 states and full covariance to allow each feature to have its own covariance matrix. However, after fitting this model the F1 score was calculated to be 50%.

The next approach was to train an LSTM to identify recessions using labeled data. After various tests, we determined a sliding window size of 4 performed the best. The Y value was chosen as the recession label for the very last month in the window. The training data spans from 1967-2000, and the testing data was from 2000 onwards. This allows the model to train of 5 examples of a recession, and test on 3 recessions that occurred after the year 2000. After running hyper parameter tuning the best model was chosen and trained further upon. The model had an overall f1 score of 0.76 and was chosen to be used for the final design (Shown in Figure 7 of the appendix).

Final Ensemble

The final version of this project involves predicting if there will be a recession in the future. To do so, we combined the predictive future value model with our classification model to predict probability of recession in the next 4 months. We start off by creating the sequences where X contains 12 months of data with 7 features each. The target y value is the recession label after 4 months. Next, we create a temporary array containing 12 months of data, we start from the very first month and predict the values for the next month, this is then appended to a temporary array and the very first month is removed. This ensures that there will always be 12 months in this temp array. The predictions are also stored in a predictions array to be used later. This process is repeated 4 times, using a combination of current month values as well as predicted month values in the prediction model. Once all 4 predicted months are generated, the next step is to inverse the original transformations applied, and scale the data using the same method used for the classification step. The 4 future months are then passed to the classification model which outputs the probability of recession at the last month (Shown in Figure 5 of the Appendix). This process is repeated for each sequence in X and eventually generates an array containing the probabilities of recession in the next 4 months. While we also could use our target y variable to further train our models this step was left out as our results proved to be working. The combined model had an accuracy of 83% and an f1 score of 56%, however we were happy with this model since it was able to predict a recession in 2001, 2007 and 2021 (Figure 4). An interesting fact to note is that currently there is roughly an 80% chance of recession in the US.

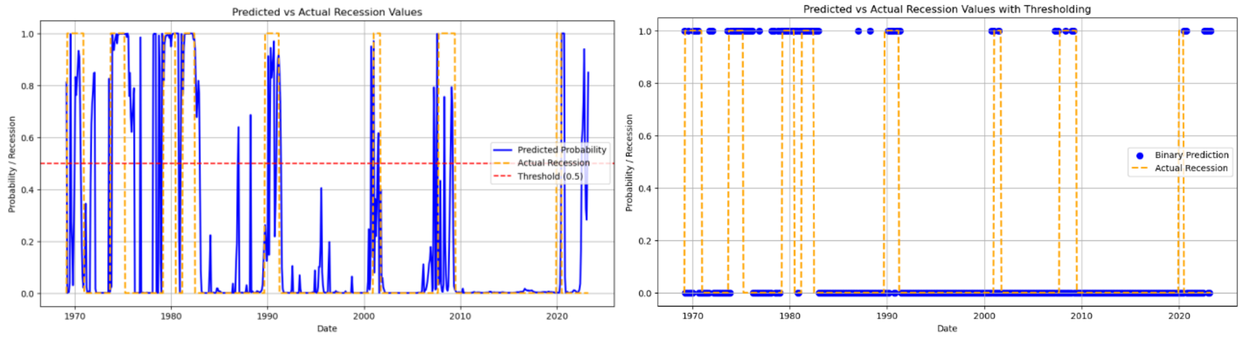

Figure 4: Final Recession Predictions

Conclusion

In conclusion, we were able to achieve a successful result from preprocessing, exploratory data analysis, our feature prediction, classification models, and future variable prediction. We were able to predict recessions and key economic shifts. The ensemble model effectively combined predictive and classification techniques, achieving an 83% in identifying economic states. These results demonstrate the potential of leveraging our findings in the course to navigate economic uncertainties with greater precision.

Github Repository: https://github.com/ChrisLato/AISE-4010-Project/tree/main

Appendix

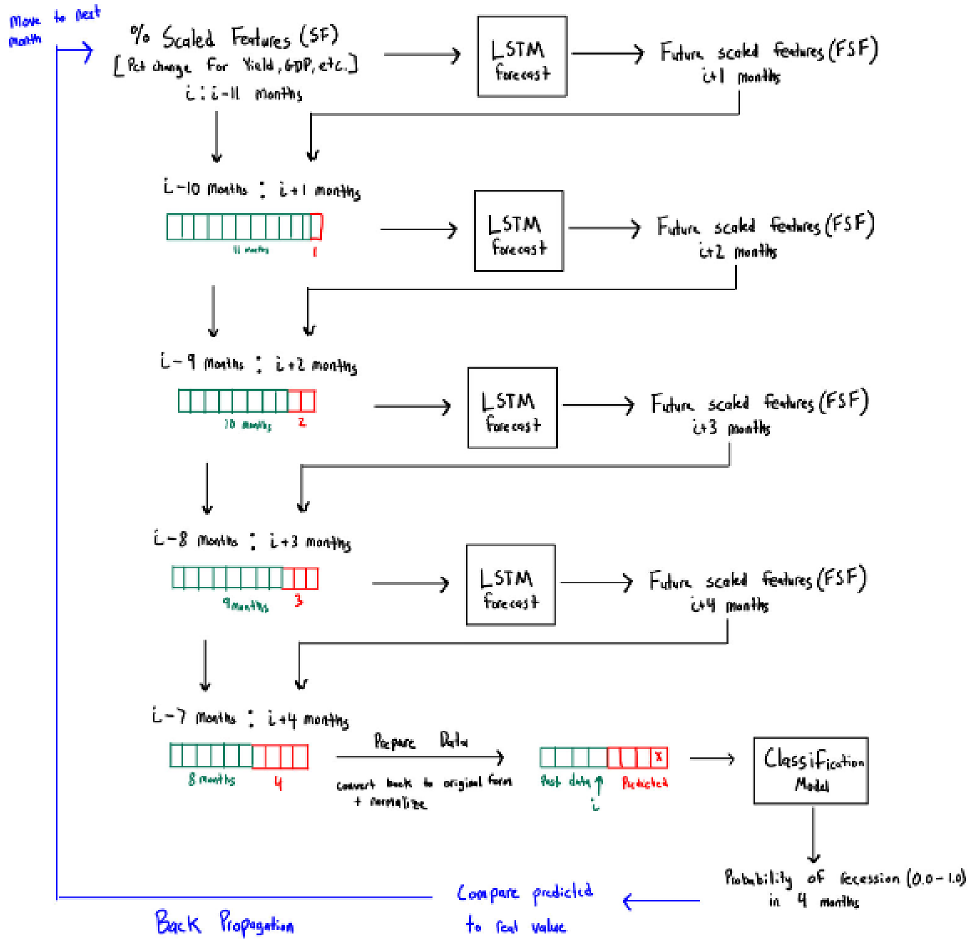

Figure 5: Initial Design Concept

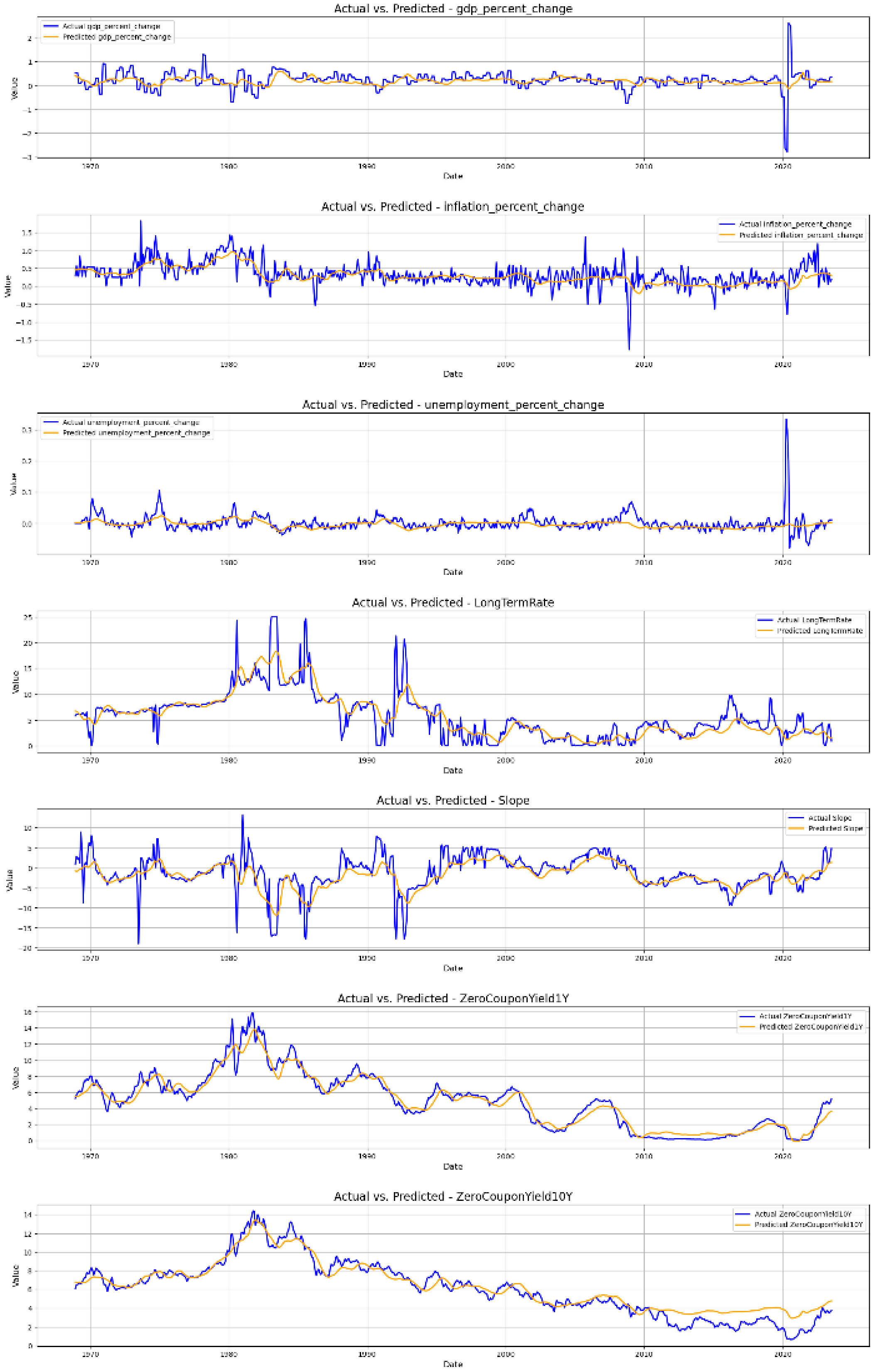

Figure 6: Enlarged LSTM Predictions

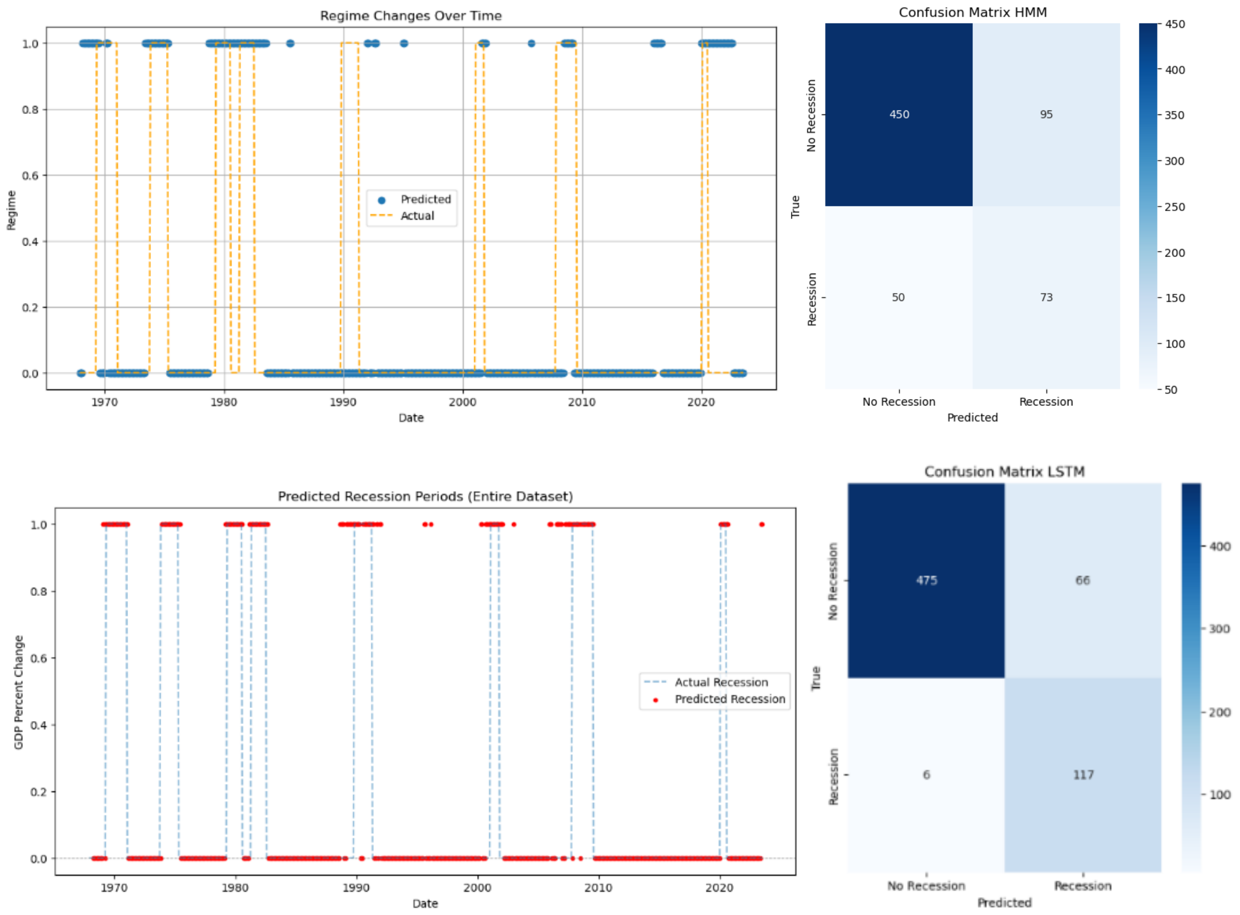

Figure 7: HMM vs LSTM classification

References

[1]R. S. Gurkaynak, B. Sack, and J. H. Wright, “The U.S. Treasury Yield Curve: 1961 to the Present,” Federalreserve.gov, Dec. 2011. https://www.federalreserve.gov/econres/feds/the-us-treasury-yield-curve-1961-to-the-present.htm (accessed Nov. 23, 2024).

[2]“U.S. Unemployment Rates,” www.kaggle.com. https://www.kaggle.com/datasets/guillemservera/us-unemployment-rates

[3]U.S. Bureau of Economic Analysis, “Real Gross Domestic Product,” Stlouisfed.org, 2024. https://fred.stlouisfed.org/series/GDPC1

[4]“US Inflation Dataset (1947 - 2023),” www.kaggle.com. https://www.kaggle.com/datasets/pavankrishnanarne/us-inflation-dataset-1947-present

[5]Hamilton, James, “Dates of U.S. recessions as inferred by GDP-based recession indicator,” FRED, Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/JHDUSRGDPBR