Predicting Federal Interest Rates Based on Economic Factors

Nadun Kulasekera Mudiyanselage

1 Introduction

1.1 Problem Statement

This project aims to develop a model that can predict federal interest rates in the USA based on macroeconomic factors. The federal interest rate is a critical tool that drives U.S. monetary policy and affects every person's life. Interest rates play a crucial role in deciding everything from the rate of return on savings accounts to the rates of borrowing, for example, credit card debt. Figure 1 shows historical federal interest rates.

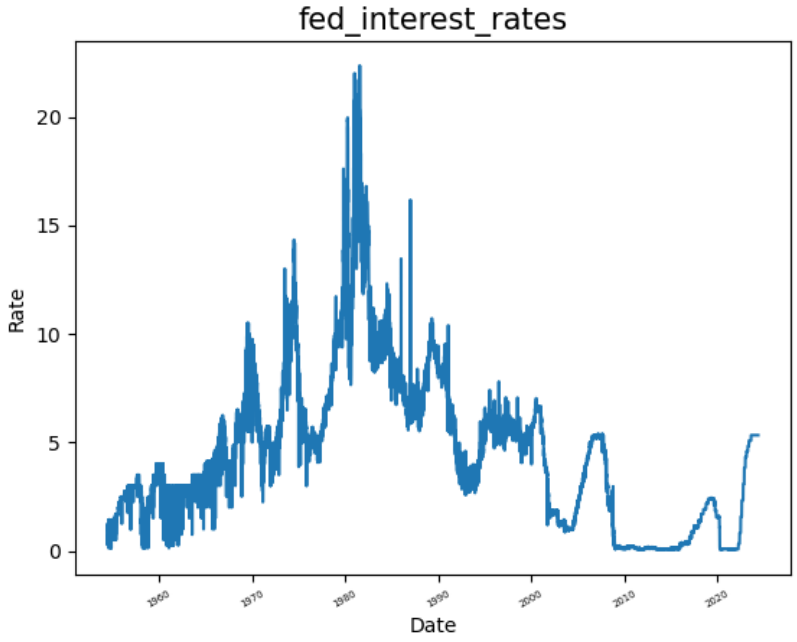

Figure 1: Federal Interest Rates

The federal rates depend on many factors, such as unemployment, GDP growth, and US dollar rates. However, US debt has been rising for years and has been affecting macroeconomic

factors for an extended period. Therefore, I incorporated US public debt into the analysis to see if it has an effect.

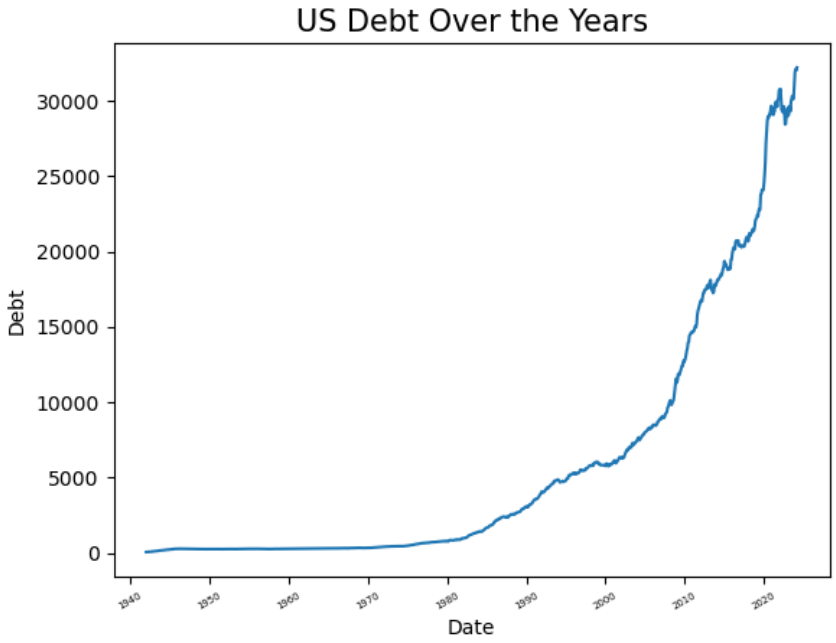

Figure 2: US Debt

1.2 Datasets

The data for this study, crucial for understanding the impact of US debt on the federal interest rates, are sourced from a reliable and widely used database: Federal Reserve Economic Data (FRED). This database has been providing comprehensive data since the 1920s.

2 Data Wrangling and EDA

2.1 Cleaning Data

The dataset contains time series data of the following features.

- Federal interest rates
- Volatility index
- Budget surplus or deficit
- Global price index
- Personal consumer expenses
- Total construction spend
- Unemployment
- CPI
- Inflation Adjusted CPI
- Bond yield
- US dollar index
- S&P 500
- US debt
- GDP
- M2 money supply

First, data was checked for missing, duplicated, and data types. There were no missing data,

but some numeric data was recorded with string characters and converted to numeric data. Furthermore, the frequency of data was different, as data was recorded weekly, monthly, or quarterly. To be consistent, all data was converted to monthly data (the first day of each month). If the first day of the month is a holiday, the nearest data value was taken. The value of the nearest month was used for missing months in quarterly data. The data starts on 01 - 01 - 1990 and ends on 03 - 01 - 2024.

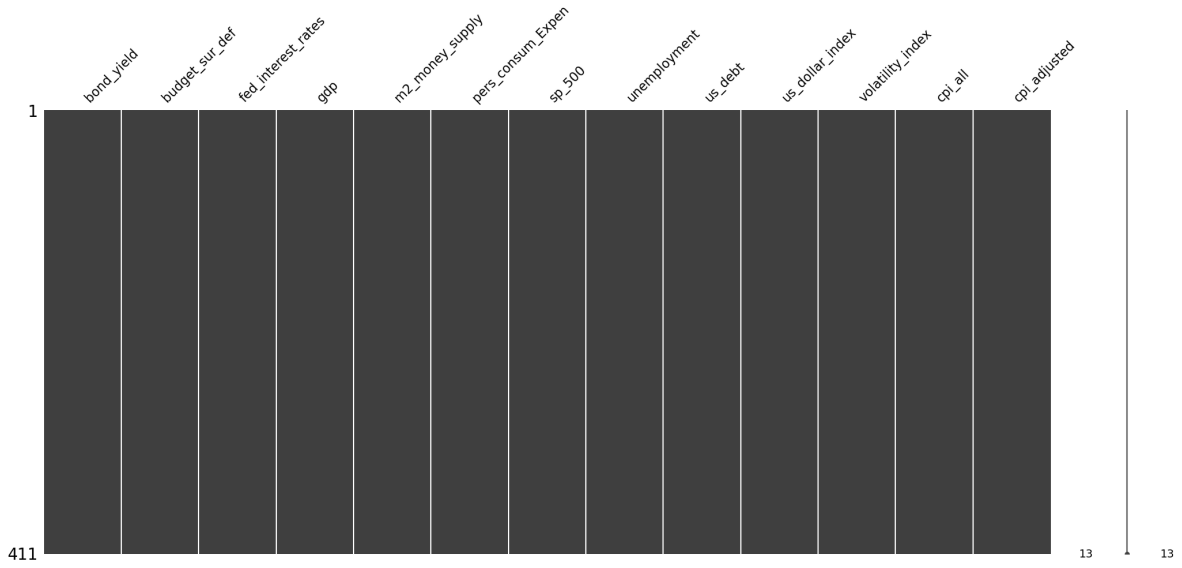

Figure 3: Looking for missing data with missingno

I also created two new features from existing features. One is the U.S. debt to GDP ratio, and the other is adjusted S&P500, which is S&P500 divided by CPI.

2.2 Exploratory Data Analysis

First, the dataset was inspected for correlations. Below is a heatmap showing the Pearson correlation coefficient of the features.

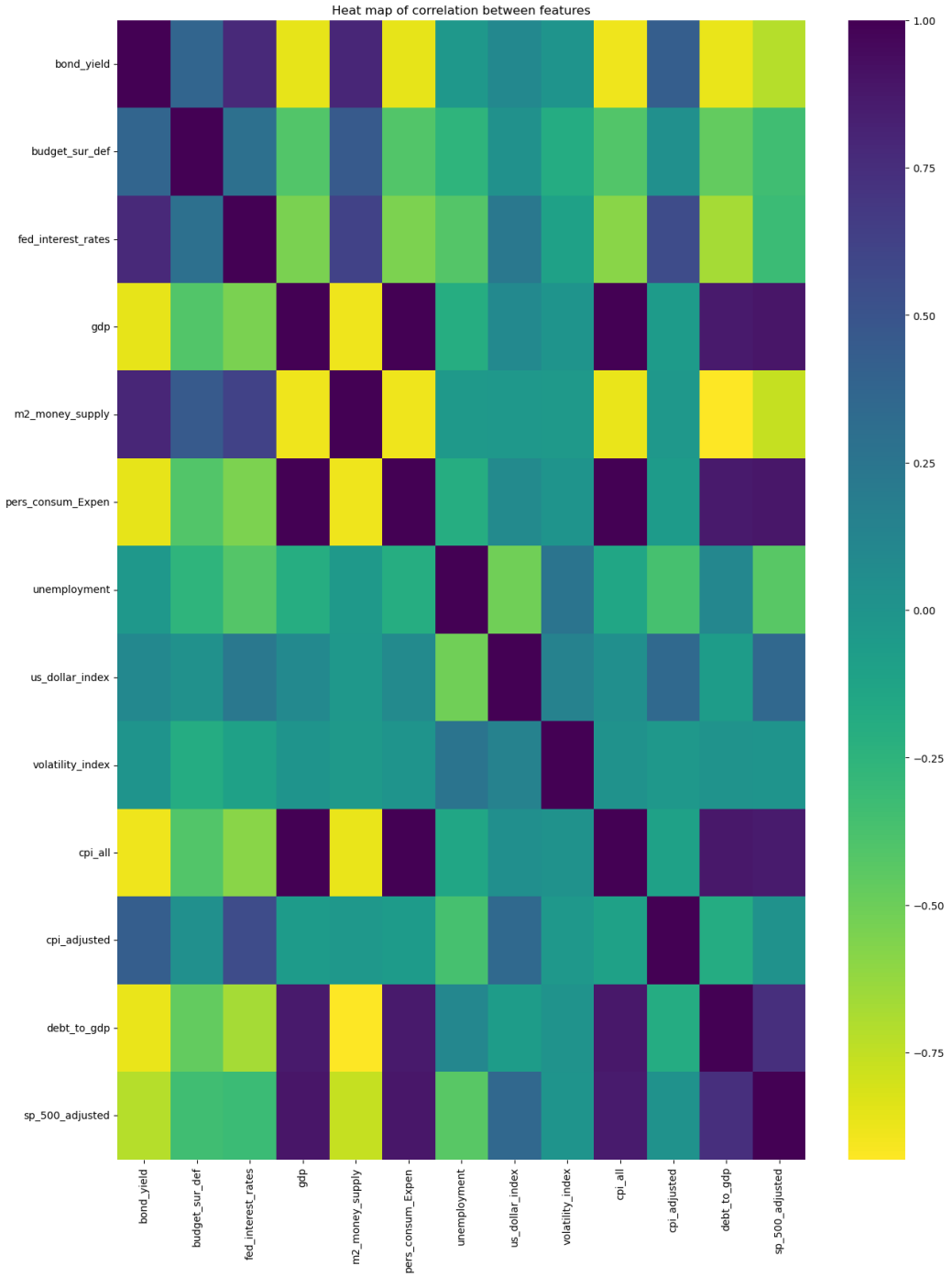

Figure 4: The heatmap - Pearson correlation coefficient of the features

There are a few notable correlations. GDP, personal consumption expenditure, and S&P500 are highly correlated with each other and also with other variables such as bond yield, m2 money supply, our debt-to-GDP ratio, and CPI. Since GDP is already considered with the debt-to-GDP ratio, we can drop it. Personal consumption expenditure is also highly correlated with CPI and some other features. Therefore, considering all these factors, I decided to drop GDP and personal consumption expenditure and adjusted S&P500.

2.2.1 Principal Component Analysis

I used principal component analysis on the dataset to identify the most important features.

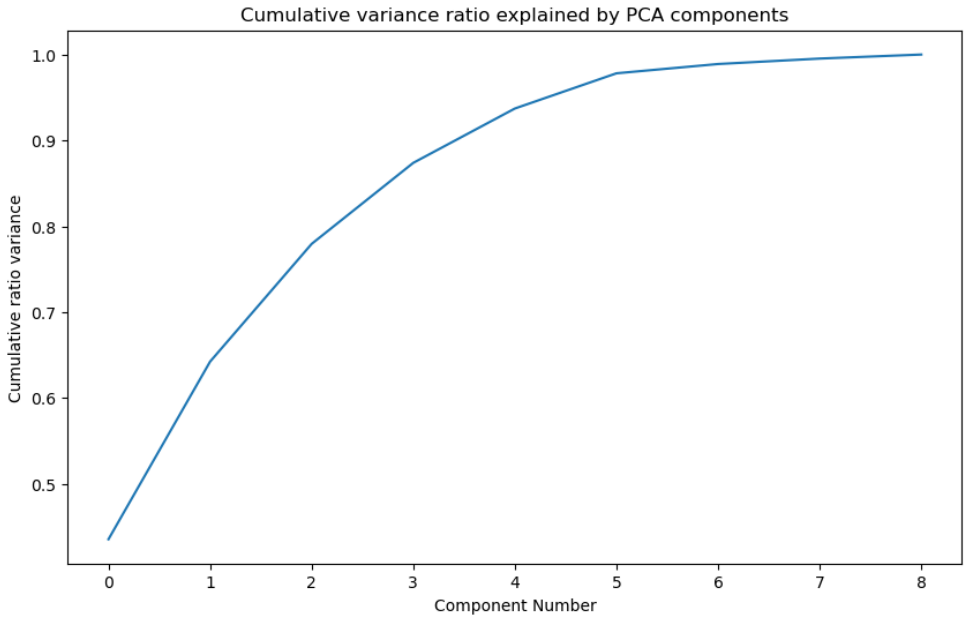

Figure 5: Cumulative variance ratio explained by PCA components

Figure 5 shows that three principal components account for 78% of the variance. Figure 6 shows the first two principal components, and the colormap depicts federal interest rates. It is clear that about four clusters of interest rates are somewhat separable from each other.

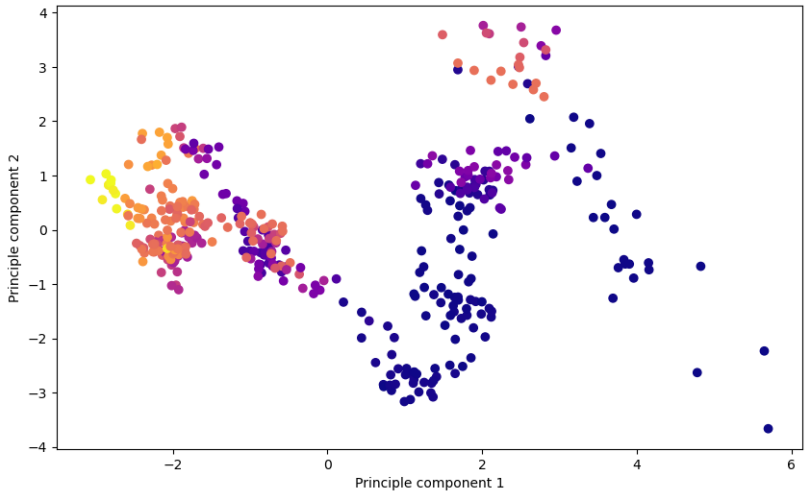

Figure 6: First two principal components and federal interest rates

Finally, Figure 7 shows that bond yield, m2 money supply, CPI, and the debt to GDP ratio are significant features related to the first principal component. Unemployment, the US dollar index, and adjusted CPI are the most represented in the second principal component. The volatility index and the budget surplus/deficit are the highlights of the third principal component.

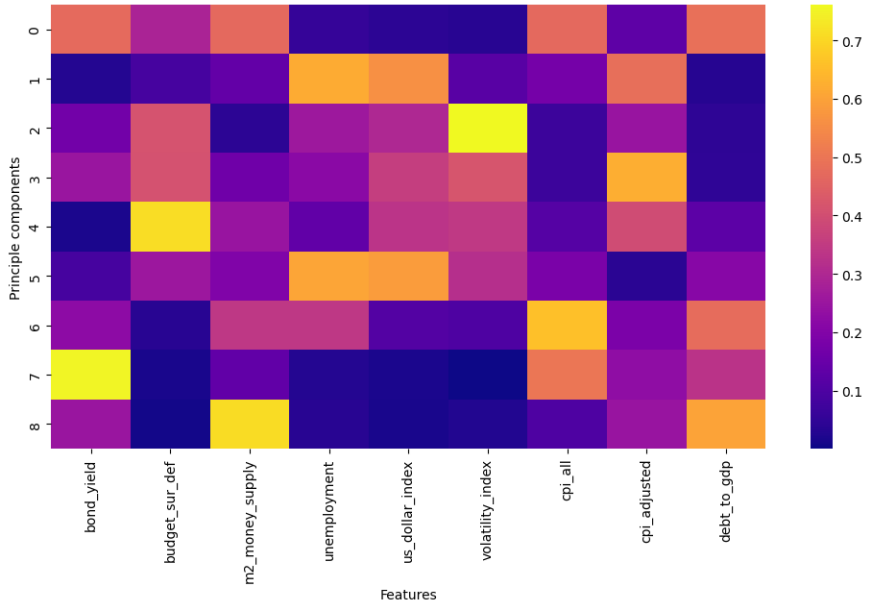

Figure 7: principal components and features

3 Modeling

3.1 Modeling choices

For modeling, I used linear regression and XGBoost. For the metrics of success, I used mean squared error and R-squared. The dataset was split into train and test sets (80-20 split). Standard scaling was then used on the train-test split. The train test was used with both models via cross-validation, and the test set was held until the better model was picked. Five-fold cross-validation was used with both models. Overall, XGBboost was the best model.

3.1.1 Linear Regression

The first model used in the analysis is multiple linear regression. The mean squared error (MSE) and R-squared during the five-fold cross-validation steps were calculated, as shown in the following figures.

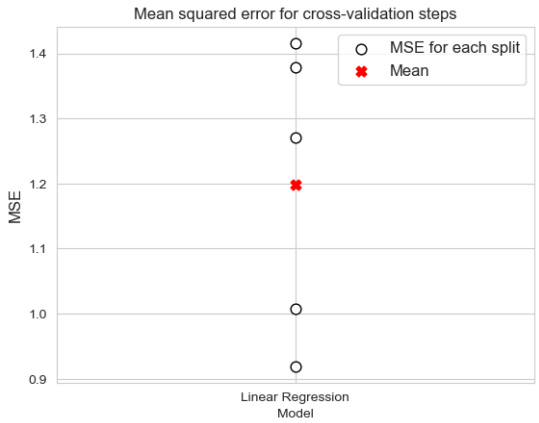

Figure 8: MSE during cross-validation

The mean MSE is around 1.2, and the smallest MSE during cross-validation is 0.92.

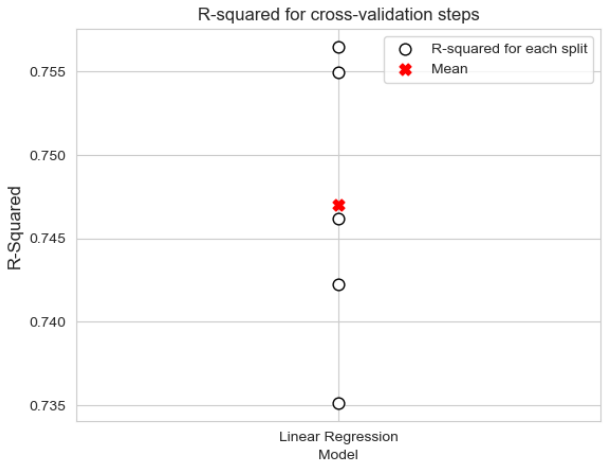

Figure 9: R-squared during cross-validation

The mean R-squared is around 0.745, and the smallest R-squared during cross-validation is 0.735.

3.1.2 XGBoost

The first step with XGBoost was hyperparameter tuning, Which was done using GridSearchCV. The best parameters for max_depth, colsample_bytree, and eta were 5,0.45 and 0.1, respectively. The best MSE during hyperparameter tuning was roughly 0.40, which is better than linear regression. Therefore, XGBoost was chosen as the preferred model.

Finally, the model was used on the test set for the first time; the MSE was 0.08, and the R-squared was 0.98.

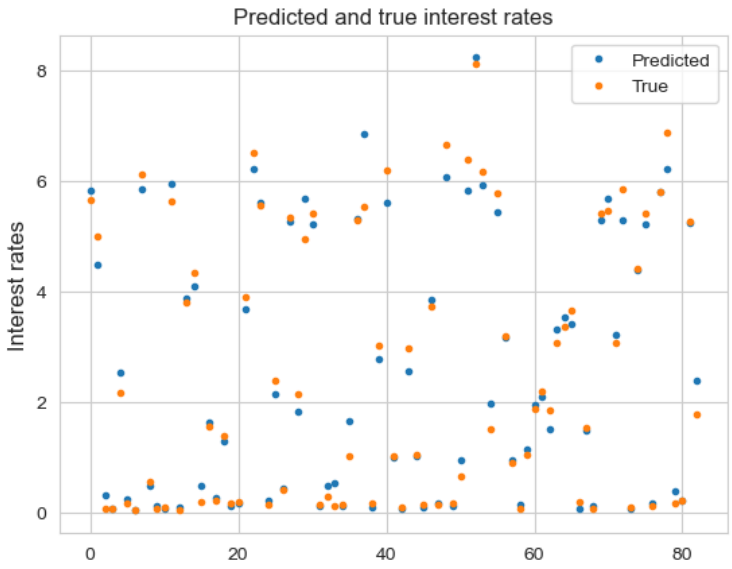

Figure 10: Predicted vs true interest rates

Metrics Table:

Method	Mean Squared Error	R-Squared
Linear Regression	0.40	0.735
XGBoost	0.08	0.98