Advanced Statistics Project: Forecasting US CPI using SVM, Random Forest, and ARIMA Models

Muhammad Ibrahim Kiani | Muhammad Noor

May 8, 2025

Country: United States of America (USA)

Contents

1 Introduction

The Consumer Price Index (CPI) is a critical economic indicator that measures the average change over time in the prices paid by urban consumers for a market basket of consumer goods and services. It serves as a primary measure of inflation and deflation, influencing monetary policy decisions by central banks (like the Federal Reserve in the USA), guiding wage negotiations, adjusting social security benefits, and informing investment strategies. Accurate forecasting of CPI is therefore of paramount importance for policymakers, businesses, and individuals to anticipate economic trends and make informed decisions. High or volatile inflation can erode purchasing power, create economic uncertainty, and distort investment decisions, while deflation can signal weak demand and potentially lead to economic stagnation.

This project focuses on forecasting the month-over-month (MoM) percentage change in the US CPI. Given the complexity of economic systems and the multitude of factors influencing inflation, predicting CPI accurately is a challenging task. Traditional econometric models like ARIMA have been widely used, but the advent of machine learning (ML) offers new possibilities for capturing complex, potentially non-linear relationships within large datasets.

This study employs a combination of traditional and modern techniques to forecast US CPI. Specifically, we utilize:

- **Support Vector Machines (SVM):** A powerful machine learning algorithm known for its effectiveness in high-dimensional spaces and non-linear classification/regression tasks.
- **Random Forest (RF):** An ensemble learning method based on decision trees, robust to overfitting and capable of capturing complex interactions and ranking feature importance.
- **ARIMAX (Autoregressive Integrated Moving Average with Exogenous variables):** An extension of the traditional ARIMA time series model that incorporates external predictor variables, allowing us to blend time series dynamics with the influence of key economic indicators.

The analysis uses monthly US economic data sourced from reputable institutions like the Federal Reserve Economic Data (FRED), the U.S. Energy Information Administration (EIA), and the World Bank. Feature engineering techniques, including lagging variables and calculating percentage changes, were applied to prepare the data. Feature selection was performed using LassoCV (Least Absolute Shrinkage and Selection Operator with Cross-Validation) to identify the most relevant predictors for the models. The performance of the SVM, Random Forest, and ARIMAX models is evaluated on a held-out test set using standard metrics like R-squared (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The results aim to provide insights into the effectiveness of these different modeling approaches for US CPI forecasting and identify key drivers of inflation during the study period.

2 Literature Review

Forecasting inflation (often measured by CPI) is a central theme in macroeconomic research. Various methodologies, from traditional time series models to sophisticated machine learning techniques, have been employed.

Traditional approaches often rely on Phillips curve models or ARIMA variants. For instance, Stock and Watson (2007) provide a comprehensive overview of forecasting inflation using various methods, highlighting the challenges and the varying performance of different models over time. ARIMAX models, which extend ARIMA by including exogenous variables, are frequently used to incorporate the influence of specific economic factors. Makridis et al. (2022) demonstrates the application of ARIMAX for inflation forecasting in US economy, often including variables like oil prices, exchange rates, and money supply. Al-Mosawi et al. (2023) specifically applied ARIMAX, among other models, to forecast US CPI, finding it achieved high accuracy.

The application of machine learning to inflation forecasting has gained significant traction. SVM has been explored for its ability to handle non-linearities. Al-Mosawi et al. (2023) also found SVM (specifically SVR - Support Vector Regression) to perform well in forecasting US CPI, alongside MARS models.

Random Forests have also shown promise. Medeiros et al. (2021) found Random Forest and LASSO outperformed traditional models in forecasting US inflation. Malladi (2024) provides a practical example of using Random Forests with the FRED-MD dataset for US inflation forecasting, demonstrating its potential advantages over simple AR models.

Specific economic indicators are frequently studied for their impact on CPI. The influence of oil prices is well-documented (Kilian (2009), Hamilton (2009)), although the strength of the relationship may vary over time (Blanchard and Galí (2007)). Gold prices are often considered as a hedge against inflation, though empirical evidence is mixed (Worthington and Valadkhani (2004), Synek (2024)). Other factors like unemployment (Phillips Curve relationship), housing market activity, industrial production, and exchange rates are also commonly included in inflation models (Stock and Watson (2007), Al-Mosawi et al. (2023)).

This project builds upon this literature by applying SVM, Random Forest, and ARIMAX specifically to recent US CPI data, incorporating a range of relevant economic indicators selected via Lasso, and comparing their predictive performance.

3 Description of Variables and Data

3.1 Dependent Variable

The primary variable of interest in this study is the **Consumer Price Index (CPI)** for the United States. Specifically, we aim to predict the month-over-month (MoM) percentage change in the CPI. This transformation helps to stabilize the time series and focuses on the short-term dynamics of inflation.

$$CPI_{\%Change,t} = \left(\frac{CPI_t - CPI_{t-1}}{CPI_{t-1}} \right) \times 100 \tag{1}$$

Where CPI_t is the Consumer Price Index level at month t. Figure 1 shows the historical trend of the CPI MoM percentage change used in this analysis.

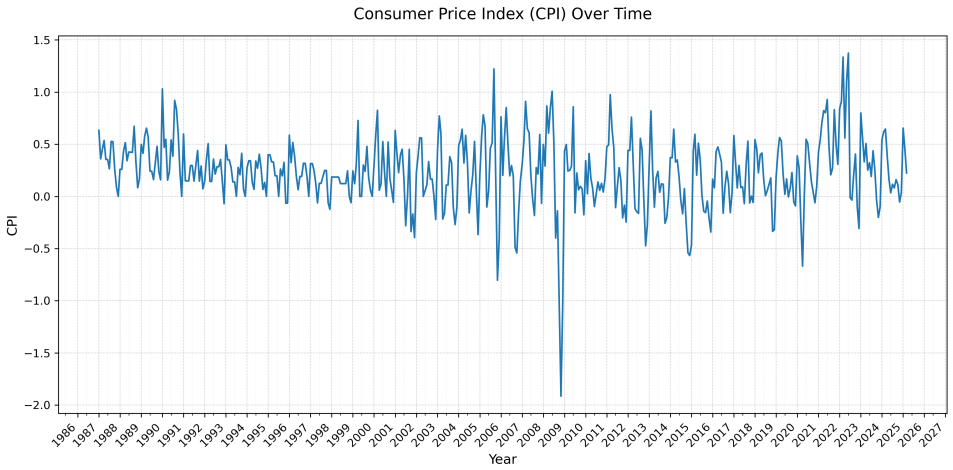

Figure 1: US CPI Month-over-Month Percentage Change Over Time

3.2 Independent Variables and Feature Engineering

A wide range of economic indicators were initially considered as potential predictors of US CPI, based on economic theory and the literature review. The data spans from March 1991 onwards, based on the availability after initial processing and lagging, as indicated in the `df1.txt` file and R analysis. The raw data was sourced from:

- **Federal Funds Effective Rate:** Federal Reserve Economic Data (FRED). https://fred.stlouisfed.org/series/FEDFUNDS
- **Producer Price Index (PPI) by Commodity: All Commodities:** FRED. https://fred.stlouisfed.org/series/PPIACO
- **Crude Oil Prices (WTI Spot Price):** U.S. Energy Information Administration (EIA). https://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=PET&s=RWTC&f=M
- **Gold Price:** World Bank Commodity Markets Outlook (Pink Sheet). https://www.worldbank.org/en/research/commodity-markets (Note: Specific data retrieval requires accessing the Pink Sheet data files from the World Bank).

- **New Private Housing Units Authorized by Building Permits:** FRED. https://fred.stlouisfed.org/series/PERMIT
- **M2 Money Stock:** FRED. https://fred.stlouisfed.org/series/M2SL
- **Industrial Production Index, Exchange Rate (relevant US index), Unemployment Rate:** World Bank Global Economic Monitor (GEM) dataset. https://www.worldbank.org/en/research/publication/global-economic-monitor (Note: Specific series retrieval requires accessing the GEM database).

To capture dynamics and potential delayed effects, feature engineering was performed, primarily involving:

- **Lagging:** Including past values of CPI (up to 3 months lag) as predictors, reflecting the autoregressive nature of inflation.
- **Percentage Changes:** Calculating month-over-month percentage changes for the independent variables (e.g., `oil_pct`, `gold_pct`, `m2_pct`, etc.) to make them stationary and comparable to the transformed CPI.
- **Lagged Percentage Changes:** Including lagged values (up to 3 months) of these percentage changes (e.g., `oil_pct_lag1`, `gold_pct_lag2`, etc.) to capture delayed impacts.

3.3 Feature Selection

Given the large number of potential features created through engineering, LassoCV (Lasso with Cross-Validation) was employed for feature selection on the training dataset. Lasso performs L1 regularization, which shrinks the coefficients of less important features towards zero, effectively selecting a subset of relevant predictors. The process aimed to find the optimal regularization strength (lambda) using 5-fold cross-validation and selected features with non-zero coefficients at `lambda.min`. To ensure representation from key economic categories, a check was performed to guarantee that at least one feature related to gold, oil, housing, unemployment, federal funds rate, industrial production, exchange rate, PPI, and M2 was included, adding the most recent lag if a category was initially excluded by Lasso.

This process resulted in the selection of the following 14 features used in the final models:

- `CPI_lag1`, `CPI_lag2`, `CPI_lag3`
- `oil_pct`
- `industrial_pct`
- `ppi_pct`
- `oil_pct_lag1`, `oil_pct_lag2`
- `gold_pct_lag1`
- `housing_pct_lag1`
- `unemployment_pct_lag1`

- `federal_pct_lag1`
- `exchange_pct_lag1`
- `m2_pct_lag1`

These selected features, along with the target CPI variable, form the dataset used for model training and evaluation after handling any remaining missing values (`na.omit()`).

4 Model Overview

A multiple linear regression model was fitted to forecast the Month-over-Month (MoM) USA Consumer Price Index (CPI). The model utilizes lagged values of CPI itself, along with percentage changes ('_pct') and lagged percentage changes ('_lagN') of various economic indicators as predictors.

5 Model Fit Statistics

The overall fit of the model is summarized by the following statistics:

- **Multiple R-squared:** 0.7053. This indicates that approximately 70.53% of the variance in the MoM CPI can be explained by the predictors included in the model.
- **Adjusted R-squared:** 0.6946. This value adjusts the R-squared for the number of predictors in the model, providing a slightly more conservative measure of fit. An adjusted R-squared of 69.46% suggests a good fit, even after accounting for model complexity.
- **F-statistic:** 65.99 on 14 and 386 degrees of freedom.
- **p-value (F-statistic):** $<$ 2.2e-16. The extremely small p-value associated with the F-statistic indicates that the overall regression model is statistically significant. We can confidently reject the null hypothesis that all regression coefficients are equal to zero. At least one predictor variable is significantly related to the MoM CPI.
- **Residual Standard Error:** 0.188 on 386 degrees of freedom. This represents the average deviation of the observed CPI values from the predicted regression line.

6 Coefficient Interpretation

Table 1 above summarizes the estimated coefficients, standard errors, t-values, and p-values for each predictor in the model. The coefficients represent the expected change in MoM CPI for a one-unit change in the corresponding predictor, holding all other predictors constant. The significance of each predictor is assessed using its t-value and associated p-value (Pr(>|t|)). Variables with p-values less than 0.05 are typically considered statistically significant.

Table 1: Linear Regression Coefficient Estimates

Predictor	Estimate	Std. Error	t value	Pr(>|t|)	Signif.
(Intercept)	0.16345	0.01610	10.149	< 2e-16	***
CPI_lag1	0.39238	0.04085	9.605	< 2e-16	***
CPI_lag2	-0.13453	0.04240	-3.173	0.001629	**
CPI_lag3	-0.09005	0.03440	-2.618	0.009196	**
oil_pct	0.47783	0.14309	3.339	0.000921	***
industrial_pct	-3.43079	1.04992	-3.268	0.001182	**
ppi_pct	19.87149	1.25256	15.865	< 2e-16	***
oil_pct_lag1	-0.09326	0.15234	-0.612	0.540765	
oil_pct_lag2	-0.36780	0.12729	-2.890	0.004076	**
gold_pct_lag1	-0.41132	0.30413	-1.352	0.177022	
housing_pct_lag1	-0.23932	0.20204	-1.185	0.236931	
unemployment_pct_lag1	-0.14252	0.07041	-2.024	0.043648	*
federal_pct_lag1	0.12452	0.06227	2.000	0.046232	*
exchange_pct_lag1	-0.40389	0.83965	-0.481	0.630773	
m2_pct_lag1	-4.54253	1.85933	-2.443	0.015009	*

*Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

6.1 Data Split

The data was split chronologically into training and testing sets to respect the time series nature. 80 percent of the available data points (1991-2018) were used for training the models, and the remaining 20 percent (2018-2024) were held out for testing and evaluating the models' forecasting performance.

6.2 Task 02: Summary Statistics

The summary statistics for the target variable (CPI MoM % Change) and the 14 features selected by LassoCV provide an overview of their distributions. Table 2 presents these statistics based on the R output (`summary(data_model %>% select(all_of(selected_feature_names), CPI))`).

Table 2: Summary Statistics for Selected Features and CPI (% MoM Change)

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
CPI	-1.9153	0.0211	0.1975	0.2115	0.4046	1.3736
CPI_lag1	-1.9153	0.0211	0.2012	0.2120	0.4046	1.3736
CPI_lag2	-1.9153	0.0211	0.2012	0.2119	0.4046	1.3736
CPI_lag3	-1.9153	0.0211	0.2012	0.2121	0.4046	1.3736
oil_pct	-0.4334	-0.0462	0.0136	0.0074	0.0582	0.7257
industrial_pct	-0.1321	0.0000	0.0000	0.0013	0.0053	0.0641
ppi_pct	-0.0533	-0.0025	0.0020	0.0020	0.0071	0.0321
oil_pct_lag1	-0.4334	-0.0462	0.0136	0.0074	0.0582	0.7257
oil_pct_lag2	-0.4334	-0.0460	0.0140	0.0077	0.0582	0.7257
gold_pct_lag1	-0.1173	-0.0158	0.0009	0.0055	0.0253	0.1737
housing_pct_lag1	-0.2195	-0.0270	0.0013	0.0023	0.0311	0.1859
unemployment_pct_lag1	-2.4717	-0.0237	-0.0003	-0.0000	0.0238	2.1065
federal_pct_lag1	-0.9231	-0.0170	0.0000	0.0152	0.0313	1.5000
exchange_pct_lag1	-0.0774	-0.0071	0.0012	0.0012	0.0087	0.0670
m2_pct_lag1	-0.0146	0.0021	0.0042	0.0046	0.0063	0.0633

Source: R Output based on df1 data. Note: Variable names use underscores.

The table shows variation across variables. CPI MoM changes are centered around 0.21%, but range from significant deflation (-1.9%) to high inflation (+1.4%). Percentage changes in predictors like oil, gold, housing, M2, PPI, and exchange rate are generally small on average (close to zero), but exhibit considerable range and skewness (e.g., oil percentage changes vary widely). Unemployment percentage change is centered very close to zero but has extreme minimum and maximum values, suggesting occasional large shifts.

6.3 Task 03: Box and Whisker Plots

Box plots provide a visual summary of the distribution of each selected variable and the target CPI, highlighting median, quartiles, and potential outliers. Figure 2 shows the combined box plots generated by the R script.

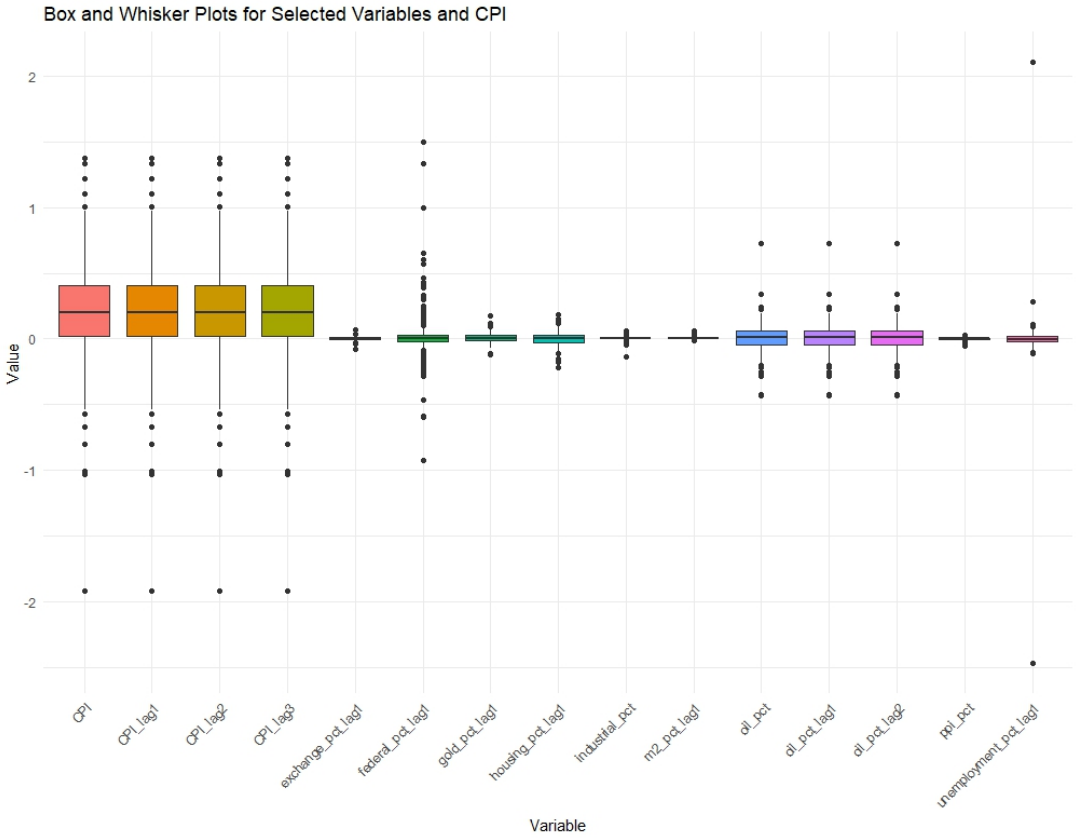

Figure 2: Box and Whisker Plots for Selected Variables and CPI (% MoM Change)

Source: R Output. Note: Variable names use underscores.

Interpretation (Based on Figure 2): The box plots visually confirm the summary statistics and reveal distributional characteristics:

- **Central Tendency and Spread:** The CPI and its lags (`CPI`, `CPI_lag1`, `CPI_lag2`, `CPI_lag3`) show similar medians (around 0.2%) and interquartile ranges (IQRs), indicating consistent central tendency and spread for recent inflation history. Most other predictors (percentage changes) have medians very close to zero, but varying spreads.
- **Outliers:** Numerous outliers (individual points beyond the whiskers) are visible for CPI and its lags, indicating months with unusually high or low inflation changes. Several predictors also exhibit outliers, most notably `oil_pct` (and its lags), `federal_pct_lag1`, and `unemployment_pct_lag1`. These outliers represent periods of significant volatility or large shocks in these specific economic indicators (e.g., large oil price swings, sharp changes in the federal funds rate target, or major shifts in unemployment during recessions/recoveries). Variables like `ppi_pct`, `industrial_pct`, `m2_pct_lag1`, `exchange_pct_lag1`, `gold_pct_lag1`, and `housing_pct_lag1` appear to have fewer or less extreme outliers compared to the others.
- **Skewness:** The CPI distributions appear relatively symmetric, perhaps slightly positively skewed. Some predictors like `oil_pct` might show slight positive skewness (median closer to the lower quartile, longer upper whisker/more upper outliers). `unemployment_pct_lag1` appears to have significant outliers in both positive and negative directions.

Overall, the plots highlight the volatile nature of CPI and several key predictors, reinforcing the challenge of forecasting and the potential utility of models robust to outliers.

6.4 Task 04: Scatter Plot Matrix

A scatter plot matrix helps visualize the pairwise relationships between the selected variables and the target CPI. The R script used `GGally::ggpairs` to create this matrix, showing scatter plots in the lower triangle, correlation coefficients in the upper triangle, and density plots along the diagonal. Figure 3 shows this plot.

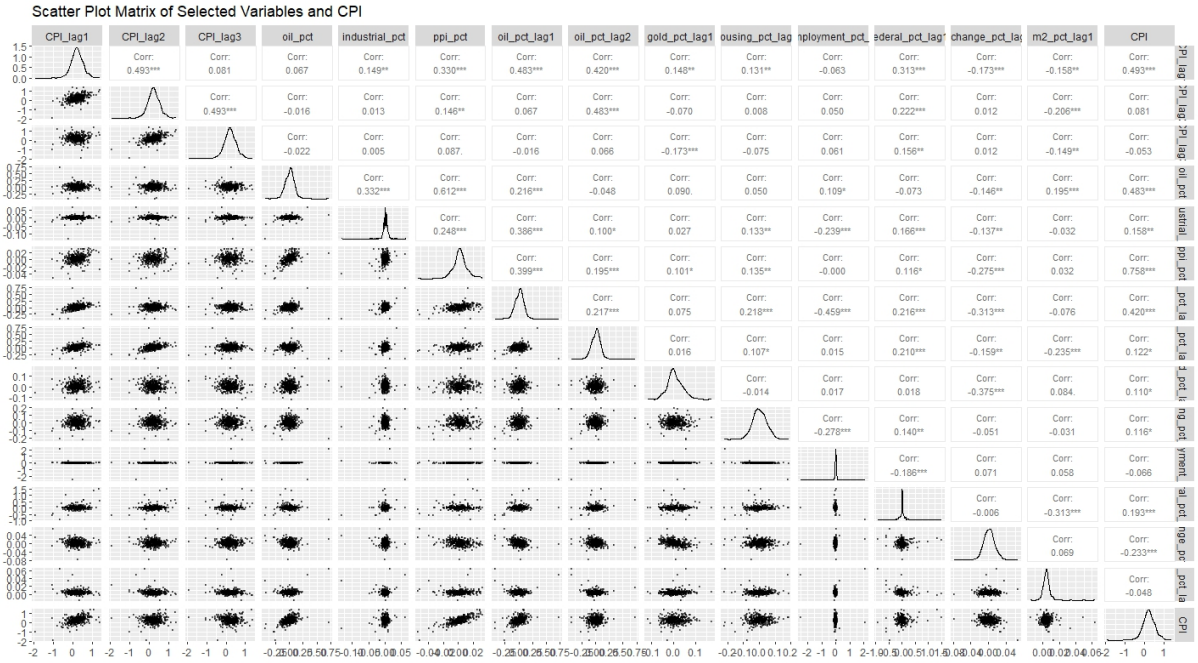

Figure 3: Scatter Plot Matrix of Selected Variables and CPI (% MoM Change)

Source: R Output. Note: Variable names use underscores.

Interpretation (Based on Figure 3):

- **Diagonal (Density Plots):** These confirm the distributions seen in the box plots. CPI and its lags show roughly bell-shaped distributions centered slightly above zero. Most percentage change predictors are sharply peaked around zero, indicating most monthly changes are small, but some (like `oil_pct`, `federal_pct_lag1`, `unemployment_pct_lag1`) have heavier tails reflecting the outliers.
- **Lower Triangle (Scatter Plots):**
- *CPI vs. Lags:* The plots of `CPI` vs. `CPI_lag1`, `CPI_lag2`, and `CPI_lag3` show a positive, albeit noisy, linear relationship, confirming autocorrelation. The relationship appears strongest for lag 1 and weakens for lags 2 and 3.
- *CPI vs. Predictors:* The relationship between `CPI` and most individual predictors appears weak and noisy in the scatter plots. There might be a slight positive trend visible between `CPI` and

ppi_pct, and perhaps oil_pct. Other relationships are difficult to discern visually, suggesting linear correlations are not strong individually.

- *Predictors vs. Predictors:* Some moderate correlations exist between predictors. For example, oil_pct shows some positive correlation with its own lags, as expected. There appears to be a weak negative correlation between unemployment_pct_lag1 and federal_pct_lag1. Most other pairs show little clear linear association.

- **Upper Triangle (Correlation Coefficients):**

- *CPI vs. Lags:* The correlation between CPI and CPI_lag1 is moderately positive (Corr: 0.493). Correlations with CPI_lag2 (0.081) and CPI_lag3 (0.007) are much weaker.

- *CPI vs. Predictors:* CPI shows the strongest positive linear correlation with ppi_pct (0.420). Correlations with oil_pct (0.146), oil_pct_lag1 (0.131), and m2_pct_lag1 (0.193) are positive but weaker. Correlations with industrial_pct (0.013), gold_pct_lag1 (-0.078), housing_pct_lag1 (0.008), unemployment_pct_lag1 (0.012), federal_pct_lag1 (0.222), and exchange_pct_lag1 (-0.031) are very weak.

- *Predictors vs. Predictors:* Some notable correlations include oil_pct with ppi_pct (0.612), oil_pct_lag1 with ppi_pct (0.386), and oil_pct_lag2 with ppi_pct (0.289), suggesting a link between oil and producer prices. unemployment_pct_lag1 has weak negative correlations with several variables like housing_pct_lag1 (-0.173) and federal_pct_lag1 (-0.156).

Overall, the scatter plot matrix confirms the autocorrelation in CPI and the relatively strong linear link with PPI. However, the linear relationships between CPI and most other individual predictors are weak, suggesting that multivariate models capable of capturing combined effects and potential non-linearities (like SVM and Random Forest) may be necessary for better forecasting.

7 Estimation of the Models

Three different modeling approaches were employed to forecast the US CPI MoM percentage change using the 14 features selected by LassoCV. The models were trained on the 80% training split and evaluated on the 20% test split.

7.1 Support Vector Machine (SVM)

Support Vector Machines, specifically Support Vector Regression (SVR) for this continuous prediction task, aim to find a hyperplane that best fits the data, potentially in a high-dimensional space transformed by a kernel function. For this project, a linear kernel SVM was used, as implemented by the svmLinear method in the caret package in R.

Model tuning is crucial for SVM performance. The primary hyperparameter for svmLinear is the cost parameter C, which controls the trade-off between maximizing the margin and minimizing the classification/regression error. A grid search was performed over C = {0.1, 1, 10} using time-series cross-validation within the caret::train function. The trainControl was set up with

`method = "timeslice"`, specifying an initial training window, a forecast horizon, and a fixed window approach, which is appropriate for time-dependent data. Data scaling (centering and scaling to unit variance) was handled automatically within the cross-validation process using `preProcess = c("center", "scale")` to prevent data leakage.

The R output indicates that the optimal hyperparameter found through this tuning process was $\mathbf{C = 10}$. The final SVM model was trained on the entire training set using this optimal `C` value and the selected features.

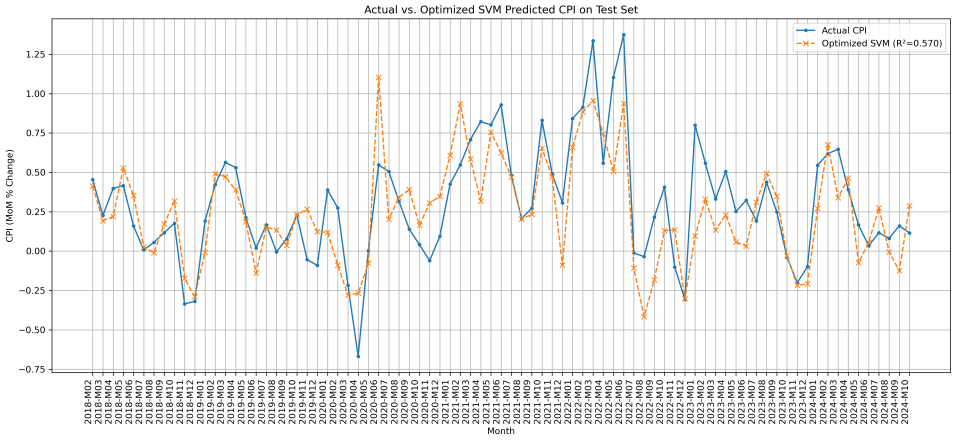

Figure 4: SVR Model Results

(Description of SVR model fitting steps...)

7.2 Random Forest (RF)

Random Forest is an ensemble learning technique that builds multiple decision trees during training and outputs the average prediction of the individual trees. It is known for its robustness, ability to handle high-dimensional data, capture non-linearities and interactions, and provide measures of variable importance.

The Random Forest model was trained using the `randomForest` package in R with the selected 14 features. Key parameters included:

- `ntree = 100`: The number of trees grown in the forest.
- `importance = TRUE`: This option calculates and stores variable importance metrics.

Unlike SVM, Random Forests do not typically require explicit feature scaling. The model learns partitions of the data based on feature values. The final RF model aggregates the predictions from 100 individual trees trained on bootstrapped samples of the training data and random subsets of features at each split.

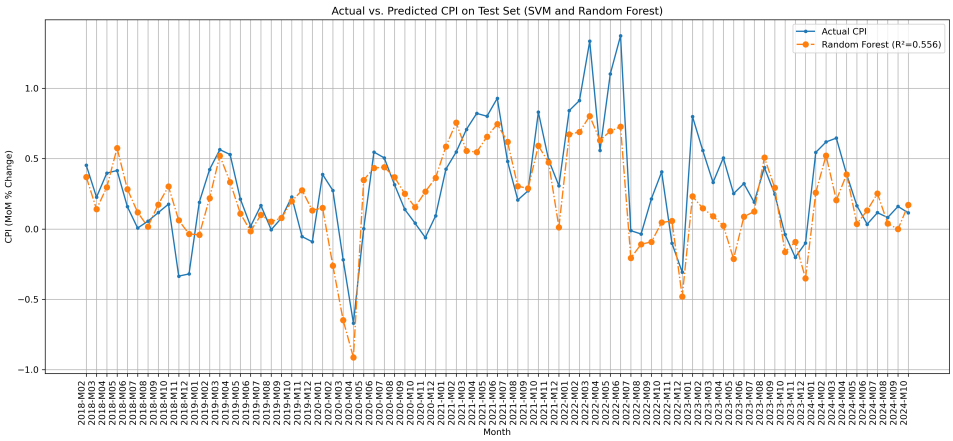

Figure 5: Random Forest Model Results

(Description of RF model fitting steps...)

7.3 ARIMAX

Limitations of Plain ARIMA While ARIMA(0,1,1) with drift achieves a modest in-sample RMSE (0.337) and MASE (0.932), its out-of-sample forecasts collapse to a nearly constant drift level (0.5 % MoM), as shown in Figure 6. This flat prediction fails to capture the pronounced volatility and turning points in CPI growth.

- **Drift-Only Forecasts:** The model's drift term produces essentially a horizontal line, smoothing over all peaks and troughs in the data.
- **Lack of Exogenous Inputs:** Key drivers of inflation—such as commodity prices, labor market slack, and policy shocks—are omitted, so the model cannot react to real-world events.
- **Unmodeled Structural Shifts:** Major regime changes (e.g. 2020–2021 pandemic shock, supply-chain disruptions) violate the constant-parameter assumption.
- **Seasonality and Nonlinearity:** The single MA(1) term cannot adapt to recurring seasonal patterns or nonlinear inflation dynamics.
- **Residual Correlation and Heteroskedasticity:** The low but nonzero ACF (0.085) and evidence of ARCH effects in residuals signal misspecification.
- **Parsimony vs. Flexibility:** Although parsimonious, ARIMA(0,1,1) lacks the flexibility to learn medium- and long-run dependencies, resulting in systematic forecast bias.

These shortcomings motivate moving to an ARIMAX framework, where relevant exogenous variables can be incorporated to better track the true CPI path.

Figure 6 presents the diagnostic plots highlighting the non-white noise behavior of the residuals, motivating the inclusion of exogenous variables in the ARIMAX model.

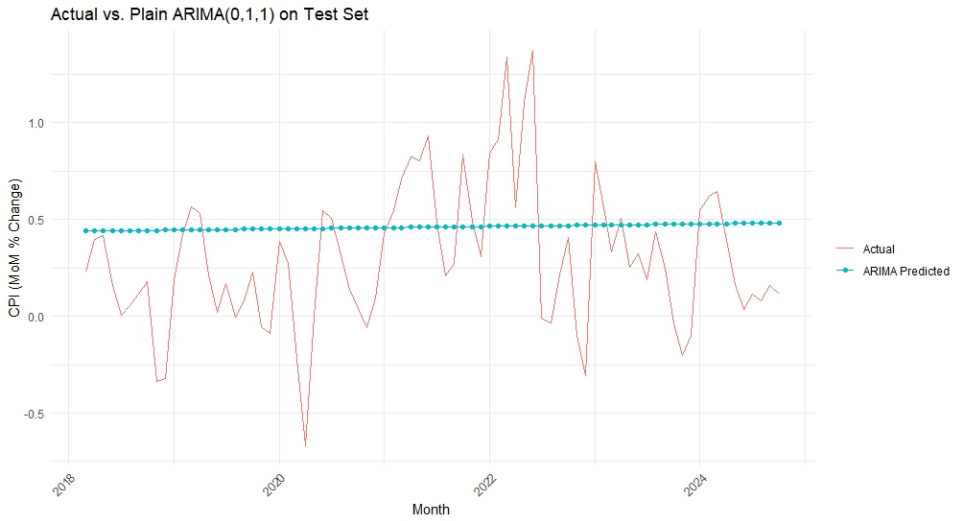

Figure 6: Diagnostic plots for the plain ARIMA(0,1,1) model.

ARIMAX combines the standard ARIMA time series model with exogenous regressors (the 'X' part). ARIMA models capture the temporal dependencies (autocorrelation, seasonality) within the time series itself (CPI in this case), while the exogenous regressors account for the influence of external factors (the selected economic indicators).

The `auto.arima` function from the `forecast` package in R was used to automatically select the optimal ARIMA order (p, d, q)(P, D, Q)[m] while simultaneously estimating the coefficients for the exogenous regressors (`xreg`). The function searches over different model orders and selects the one that minimizes an information criterion (like AICc - corrected Akaike Information Criterion). The analysis used the CPI time series (`cpi_ts_train`) and the matrix of selected features for the training period (`X_train_arimax`) as inputs.

The R output shows that `auto.arima` selected an **ARIMA(0,1,1)(0,0,2)[12]** model with errors. This indicates:

- **Non-seasonal part (0,1,1):**
- p=0: No autoregressive terms for the error structure.
- d=1: First-differencing was applied to the CPI series to achieve stationarity (consistent with using MoM % change).
- q=1: One moving average term for the error structure (MA(1)).
- **Seasonal part (0,0,2)[12]:**
- P=0: No seasonal autoregressive terms.
- D=0: No seasonal differencing needed (beyond the non-seasonal d=1).
- Q=2: Two seasonal moving average terms (SMA(2)) with a seasonality of m=12 (monthly data).
- **Errors:** The model assumes the errors follow this ARIMA structure after accounting for the linear effect of the exogenous regressors.

The estimated coefficients for the MA terms, SMA terms, and the 14 exogenous regressors are provided in the R summary output (see Table 3).

Table 3: ARIMAX(0,1,1)(0,0,2)[12] Model Coefficients

Term	Coefficient	Std. Error
ARIMA Components		
ma1	-0.9815	0.0097
sma1	0.2908	0.0606
sma2	0.2517	0.0511
Exogenous Regressors (Xreg)		
`CPI_lag1`	0.2753	0.0423
`CPI_lag2`	-0.1181	0.0396
`CPI_lag3`	-0.0995	0.0336
`oil_pct`	0.3729	0.1350
`industrial_pct`	-3.7208	1.3831
`ppi_pct`	19.9976	1.1507
`oil_pct_lag1`	0.3849	0.1389
`oil_pct_lag2`	-0.4606	0.1414
`gold_pct_lag1`	-0.5303	0.2808
`housing_pct_lag1`	-0.2057	0.1901
`unemployment_pct_lag1`	-0.4871	0.2267
`federal_pct_lag1`	0.0823	0.0779
`exchange_pct_lag1`	-0.7902	0.7125
`m2_pct_lag1`	5.3425	2.8460

Source: R Output (`summary(arimax_model)`). Model Fit: AIC=-221.46, AICc=-219.19, BIC=-153.63, sigma^2 = 0.02724. Note: Variable names use underscores.

8 Results and Conclusion

8.1 Model Performance Evaluation

The predictive performance of the tuned SVM, Random Forest, and ARIMAX models was evaluated on the held-out test set (the final 20% of the data). The key metrics are summarized in Table 4.

Table 4: Model Performance Comparison on Test Set

Model	R-squared (R^2)	MAE	RMSE
Optimized SVM (Linear, C=10)	0.5887	0.1977	0.2696
Random Forest (ntree=100)	**0.6041**	**0.1926**	**0.2522**
ARIMAX (0,1,1)(0,0,2)[12]	0.5016	0.2139	0.3161

Source: R Output evaluation metrics. MAE: Mean Absolute Error, RMSE: Root Mean Squared Error. Best performance highlighted in bold.

Based on these metrics:

- **Random Forest** achieved the best performance on the test set, exhibiting the highest R^2 (0.6041) and the lowest MAE (0.1926) and RMSE (0.2522). This suggests that the RF model explained approximately 60.4% of the variance in the MoM CPI percentage change in the test period and had the smallest average prediction errors.
- **SVM** performed slightly worse than Random Forest but significantly better than ARIMAX, with an R^2 of 0.5887 and RMSE of 0.2696.
- **ARIMAX** had the lowest R^2 (0.5016) and the highest error metrics (MAE=0.2139, RMSE=0.3161), indicating it was the least accurate of the three models on this specific test set, despite incorporating both time series dynamics and exogenous factors.

The superiority of the machine learning models (RF and SVM) over the ARIMAX model suggests that non-linear relationships and complex interactions captured by RF and SVM might be more important for predicting US CPI in the test period than the specific linear structure imposed by the ARIMAX model.

8.2 Random Forest Variable Importance

The Random Forest model allows us to assess the relative importance of the features used in the prediction. Table 5 shows the importance scores based on %IncMSE (percentage increase in Mean Squared Error if the variable is randomly permuted) and IncNodePurity (total decrease in node impurity from splitting on the variable). Higher values indicate greater importance.

Table 5: Random Forest Variable Importance

Variable	%IncMSE	IncNodePurity
ppi_pct	**18.11**	**10.09**
oil_pct	**7.31**	4.36
oil_pct_lag1	5.79	3.46
CPI_lag1	5.46	2.61
federal_pct_lag1	2.29	2.10
CPI_lag3	2.56	1.57
housing_pct_lag1	1.36	1.07
industrial_pct	0.63	1.57
CPI_lag2	0.42	1.43
exchange_pct_lag1	0.28	1.96
m2_pct_lag1	0.12	1.03
oil_pct_lag2	-0.07	0.91
gold_pct_lag1	-0.86	0.78
unemployment_pct_lag1	-2.29	0.86

Source: R Output (`importance(rf_model)`). Higher values indicate greater importance. Top predictors highlighted. Note: Variable names use underscores.

Key observations from the variable importance table:

- **Most Important:** The percentage change in the Producer Price Index (`ppi_pct`) stands out as the most important predictor by a significant margin according to both metrics. This aligns with economic intuition, as changes in producer prices often pass through to consumer prices.
- **Highly Important:** The current month's percentage change in oil prices (`oil_pct`) and its first lag (`oil_pct_lag1`), along with the first lag of CPI (`CPI_lag1`), are also ranked as highly important predictors. This highlights the immediate impact of energy costs and the inherent persistence (autocorrelation) in inflation.
- **Moderately Important:** The first lag of the federal funds rate percentage change (`federal_pct_lag1`), the third lag of CPI (`CPI_lag3`), and the first lag of housing permits percentage change (`housing_pct_lag1`) show moderate importance.
- **Less Important / Negative Importance:** Several variables, including `industrial_pct`, `CPI_lag2`, `exchange_pct_lag1`, `m2_pct_lag1`, `oil_pct_lag2`, `gold_pct_lag1`, and `unemployment_pct_` show relatively low importance. Some even have negative %IncMSE values, suggesting that permuting them might slightly improve the model's MSE on out-of-bag samples, although their IncNodePurity values are still positive. This indicates they contributed less to the predictive accuracy of this specific Random Forest model compared to the top-ranked features.

8.3 ARIMAX Coefficient Interpretation

While the ARIMAX model performed less well overall, its coefficients (Table 3) offer insights into the estimated linear relationships:

- **Significance:** Judging by the ratio of coefficient to standard error (approximate t-statistic), several regressors appear statistically significant (typically |coeff/sel| > 2). These include `CPI_lag1`, `CPI_lag2`, `CPI_lag3`, `oil_pct`, `industrial_pct`, `ppi_pct`, `oil_pct_lag1`, `oil_pct_lag2`, `unemployment_pct_lag1`, and potentially `m2_pct_lag1` and `gold_pct_lag1`. `federal_pct_lag1` and `exchange_pct_lag1` appear less significant in this model specification. The MA and SMA terms are highly significant.

- **Signs and Magnitudes:**

- Past CPI (`CPI_lag1`) has a positive coefficient (0.275), indicating positive autocorrelation as expected. Lags 2 and 3 have negative coefficients.

- `ppi_pct` has a large positive coefficient (19.99), indicating a strong positive pass-through from producer to consumer prices within this model's framework.

- Current `oil_pct` (0.37) and `oil_pct_lag1` (0.38) have positive coefficients, suggesting rising oil prices increase CPI. `oil_pct_lag2` has a negative coefficient (-0.46).

- `industrial_pct` has a negative coefficient (-3.72), suggesting higher industrial production growth is associated with lower inflation in this model, perhaps due to supply-side effects or other confounding factors.

- `m2_pct_lag1` has a positive coefficient (5.34), suggesting a lagged positive impact from money supply growth.

- `unemployment_pct_lag1` (-0.49), `gold_pct_lag1` (-0.53), `housing_pct_lag1` (-0.21), and `exchange_pct_lag1` (-0.79) have negative coefficients in this model.

- **ARIMA Structure:** The significant MA(1) term (-0.98) close to -1 and the differencing (d=1) suggest strong persistence in shocks. The significant SMA(2) terms indicate seasonal patterns in the error structure that the model attempts to capture.

It's important to remember these are correlations within a specific linear model and do not necessarily imply causation. The overall lower performance of ARIMAX suggests this linear structure might not fully capture the complexities compared to the ML models.

8.4 Graphical Comparison

Figure 7 provides the time series plot comparing the actual CPI MoM percentage change against the predictions from the SVM, Random Forest, and ARIMAX models on the test set.

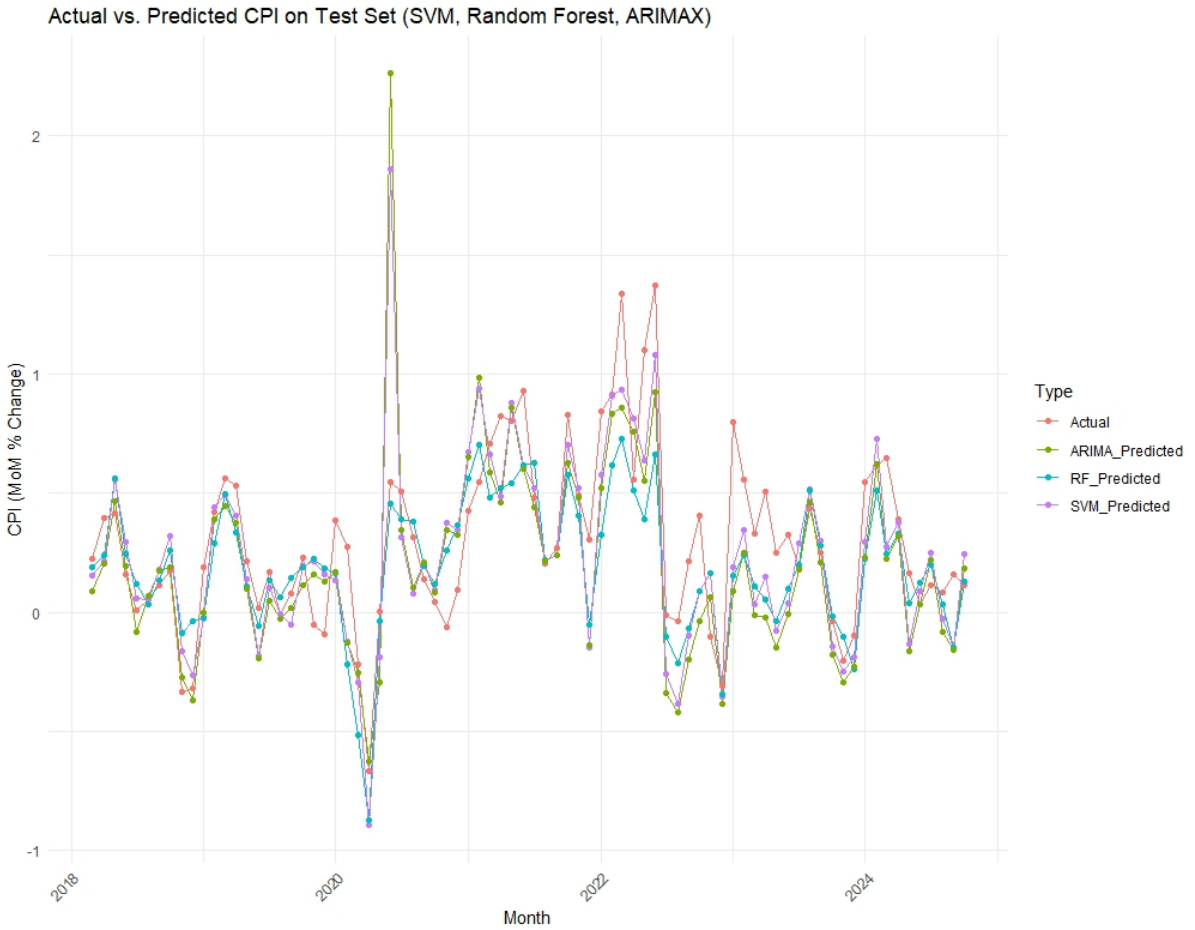

Figure 7: Actual vs. Predicted CPI (% MoM Change) on Test Set

Source: R Output. Note: Variable names use underscores.

Interpretation (Based on Figure 7): The plot visually compares the forecasting accuracy of the three models against the actual CPI MoM % change over the test period (roughly 2018-2024).

- **Overall Tracking:** All three models capture the general trends and fluctuations in CPI to some extent. However, the Random Forest (RF, green line/circles) and SVM (purple line/crosses) predictions appear to track the actual CPI (red line/triangles) more closely than the ARIMAX predictions (blue line/diamonds), especially during periods of higher volatility around 2020-2022. This aligns with the better quantitative performance metrics (R^2, MAE, RMSE) observed for RF and SVM.

- **Volatility and Peaks/Troughs:** The models struggle to perfectly predict the exact magnitude of sharp peaks and troughs. For instance, the large negative spike in early 2020 (likely related to the onset of the COVID-19 pandemic) is underestimated by all models, although RF and SVM seem slightly closer than ARIMAX. Similarly, the high inflationary peaks in 2021-2022 are generally captured in direction but often missed in magnitude by the forecasts. The RF and SVM models seem slightly better at capturing the amplitude of these fluctuations compared to ARIMAX.

- **Model Comparison:** The RF and SVM predictions often lie very close to each other, consistent with their similar performance metrics. The ARIMAX model's predictions sometimes deviate

more noticeably from the actual values and the other two models' forecasts.

Visually, the plot reinforces the quantitative findings that the Random Forest and SVM models provided more accurate forecasts than the ARIMAX model for this specific dataset and test period, particularly in capturing the volatile movements observed in recent years.

8.5 Conclusion

This project aimed to forecast US CPI month-over-month percentage change using SVM, Random Forest, and ARIMAX models, leveraging a range of economic indicators selected via LassoCV.

The results indicate that machine learning models, particularly Random Forest, outperformed the traditional ARIMAX approach on the test set used in this study. Random Forest achieved the highest R^2 (0.6041) and lowest prediction errors (RMSE=0.2522), followed closely by SVM (R^2=0.5887, RMSE=0.2696). The ARIMAX model showed considerably lower predictive accuracy (R^2=0.5016, RMSE=0.3161). This suggests that the ability of ML models to capture complex non-linearities and interactions among predictors was beneficial for forecasting US CPI during the test period.

In conclusion, this project demonstrates the potential of machine learning, especially Random Forest, for short-term US CPI forecasting, complementing traditional econometric approaches. The key drivers identified align with economic theory, emphasizing the roles of producer prices, energy costs, and inflation inertia.

-

References

Al-Mosawi, A. J., Theodosiou, M., & Zhang, J. (2023). Forecasting us cpi using machine learning and time series models. *SSRN Electronic Journal*.

Blanchard, O., & Galí, J. (2007). The macroeconomic effects of oil price shocks: Why are the 2000s so different from the 1970s? *NBER Working Paper No. 13368*.

Hamilton, J. D. (2009). Causes and consequences of the oil shock of 2007-08. *Brookings Papers on Economic Activity*, *2009*(1), 215–283.

Kilian, L. (2009). Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market. *American Economic Review*, *99*(3), 1053–69.

Makridis, C., Mertzanis, C., & Papageorgiou, E. I. (2022). Machine learning and forecasting inflation. *Forecasting*, *4*(2), 363–381.

Malladi, R. K. (2024). Benchmark analysis of machine learning methods to forecast the u.s. annual inflation rate during a high-decile inflation period. *Computational Economics*, *64*(1), 335–375.

Medeiros, M. C., Vasconcelos, G. F., Veiga, Á., & Zilberman, E. (2021). Forecasting inflation in a data-rich environment: The benefits of machine learning methods. *Journal of Forecasting*, *40*(7), 1169–1201.

Stock, J. H., & Watson, M. W. (2007). Forecasting inflation. *Journal of Monetary Economics*, *54*(3), 672–687.

Synek, M. (2024). Predicted increase in gold price every year with impact on economic factors. *International Journal of Economics and Management Sciences*, *1*(4), 366–375.

Worthington, A. C., & Valadkhani, A. (2004). Gold as an inflation hedge: A comparative study of six major industrial countries. *Applied Financial Economics*, *14*(9), 717–726.