Inflation Forecasting for Pakistan

Using Time Series and Machine Learning Techniques

Abubakar Shahid (22i-1833)

Eesha Khurram (23L-2641)

Department of Data Science
FAST NUCES Islamabad

May 5, 2025

Abstract:

This project compares time series forecasting methods and modern machine learning models for predicting Pakistan's inflation from 2019 to 2023. We implement ARIMA, Ridge, LASSO, Elastic Net, Random Forest, and XGBoost models using macroeconomic variables sourced from the World Development Indicators (WDI) and the International Monetary Fund (IMF). Models are trained on data from 1983–2018 and tested on data from 2019–2023. We evaluate forecast accuracy using MSE. All analyses are performed in R, and results are reported in this document.

Contents

1 Introduction

Inflation: Inflation is defined as the rate at which the general price level of goods and services rises over a period of time, leading to a decrease in the purchasing power of money. It reflects how much more expensive the general cost of living becomes over time.

Example: If inflation is high, the same amount of money buys fewer items than before. For example, if a loaf of bread costs 100Rs this year and 110Rs next year, that's 10 percent inflation.

Causes Of Inflation

There are several factors that contribute to inflation:

- Demand-pull inflation: Occurs when the demand for goods and services exceeds their supply, pushing prices higher.
- Cost-push inflation: Happens when the cost of production rises (e.g., due to increased wages or material costs), leading to higher prices for final goods and services.
- Built-in inflation: Also known as wage-price inflation, occurs when workers demand higher wages to keep up with cost of living increases, which in turn causes businesses to raise prices.

Problems arising from Inflation

Inflation can cause significant problems for both individuals and economies, including:

- The cost of living rises, especially for essentials like food, fuel, and healthcare. Housing costs, including rent and mortgages, also increase.
- Businesses face uncertainty due to inflation. Rising production costs can disrupt supply chains, leading to product shortages and delays.
- Inflation often leads to higher interest rates as central banks try to control prices.
- Inflation can cause the currency to depreciate, making imports more costly. This can weaken the currency.

- A wage-price spiral can occur when workers demand higher wages, leading businesses to raise prices. This worsens inflation.
- Governments may struggle with rising debt costs during inflation.
- High inflation can make exports more expensive and reduce competitiveness. Consumers may turn to cheaper imports, hurting domestic production.

Inflation in Pakistan

Inflation in Pakistan has been a persistent problem for several years. The country faces high inflation rates due to factors like political instability, volatile commodity prices, external shocks (like oil price fluctuations), and exchange rate instability. Inflation in Pakistan can lead to social unrest, increased poverty levels, and further economic disparity, particularly affecting the middle and lower-income groups.

Inflation Forecasting

Accurately forecasting inflation is critical for economic planning and policy formulation. It helps governments and central banks design strategies to control inflation and avoid adverse economic conditions. However, inflation forecasting is a challenging task due to its dependence on both internal and external factors, including Government policies, exchange rates, commodity prices, and global economic conditions. The objective of this study is to compare and evaluate the performance of time series model and machine learning models in forecasting inflation for Pakistan. The study aims to:

- Discuss summary statistics of variables used to predict inflation.
- Fit ARIMA, Ridge, LASSO, Elastic Net, Random Forest, and XGBoost models to forecast inflation.
- Compare models based on predictive accuracy and robustness.
- Provide recommendations for effective inflation prediction.

2 Literature Review

A survey of contemporary studies shows that researchers have explored various strategies for forecasting inflation, blending traditional econometric models with modern machine learning techniques:

- Ahmed et al. (2022) employed ARIMA models to predict Pakistan's inflation. The study utilized CPI data as the primary variable to model inflation trends, noting robust performance during stable periods but limitations during abrupt economic changes.
- Khan and Javed (2021) adopted a Vector AutoRegression (VAR) approach to model the dynamic relationship between inflation and interest rates. Their model included variables such as global crude oil prices, imports, money supply, government borrowing, exchange rates, and interest rates. These were selected to capture the multifaceted determinants of inflation.
- Raza and Mehmood (2024) provided evidence that while machine learning models are sensitive to data variability, they outperform ARIMA in turbulent economic periods. The study compared classical and machine learning models, showing that ensemble methods like Random Forest frequently deliver lower RMSE values.
- Hussain and Akram (2018) applied a Vector Error Correction Model (VECM) to study the impact of exchange rate shocks on inflation. Their model focused on the relationship between exchange rates and inflation over time.
- Shabbir and Iqbal (2021) reviewed deep learning models and confirmed their effectiveness in time series forecasting in emerging markets. They highlighted the strengths of models like LSTM and GRU in capturing complex inflation trends.
- Tariq and Noor (2020) compared Ordinary Least Squares (OLS) and Ridge Regression for lag selection in forecasting models. The study addressed multicollinearity, a key issue when dealing with multiple inflation predictors.

- Malik et al. (2023) proposed a hybrid ARIMA-XGBoost model combining the interpretability of ARIMA with the predictive strength of XGBoost, aiming to reduce bias and variance in forecasts.

- Fatima et al. (2023) compared classical forecasting methods with machine learning techniques, highlighting that ensemble methods like Random Forest often yield lower RMSE values. Their model incorporated variables such as CPI, food prices, and oil prices.

- Siddiqui and Hamid (2020) demonstrated that XGBoost effectively handles nonlinear data, improving accuracy under volatile conditions. They used macroeconomic indicators including exchange rates, fuel prices, and CPI components.

- Ali et al. (2019) showed that Long Short-Term Memory (LSTM) networks can capture long-term dependencies, enhancing the forecasting of inflation trends. Their model relied on sequential data inputs like monthly CPI, interest rates, and energy prices.

These studies provide a strong foundation for the current research, which aims to refine and extend their methodologies by directly comparing time series with advanced machine learning techniques in the context of Pakistan's inflation forecasting.

3 Variables and Data Description

We use annual data from 1983 to 2023 sourced from WDI and IMF. The variables include:

GDP Growth (annual %) Measures the yearly increase in a country's total economic output. A higher GDP growth rate indicates stronger economic activity, which can influence inflation through increased consumer and business demand.
https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG

Oil Prices (USD per barrel) refer to the international price of crude oil. Changes in oil prices directly affect transportation and production costs, often leading to overall price

level changes in the economy.

https://www.eia.gov/dnav/pet/hist/LeafHandler.ashxn=pets=f000000$_{3f=m}$

Exchange Rate (PKR/USD) indicates the value of the Pakistani rupee against the US dollar. A depreciating exchange rate increases the cost of imported goods and services, contributing to inflationary pressure.

https://data.imf.org/en/Data-Explorer?datasetUrn=IMF.RES:WEO(6.0.0)

Foreign Direct Investment (USD) represents investment by foreign entities in domestic markets. Higher FDI can enhance productivity and supply but may also increase money flow and demand, impacting inflation.

https://data.worldbank.org/indicator/BX.KLT.DINV.CD.WD

Imports of Goods and Services (% of GDP) show the proportion of goods and services a country imports relative to its GDP. High import dependence exposes the economy to external price shocks that can influence inflation.

https://data.worldbank.org/indicator/NE.IMP.GNFS.ZS

External Debt Stocks (% of GNI) measure the total debt owed to foreign creditors as a share of gross national income. Large external debt may strain fiscal capacity and affect inflation through borrowing-related policy measures.

https://data.worldbank.org/indicator/DT.DOD.DECT.CD

Broad Money (% of GDP) includes the total supply of money available in the economy relative to GDP. An increase in broad money without a corresponding increase in goods and services can lead to inflation.

https://data.worldbank.org/indicator/FM.LBL.BMNY.GD.ZS

IMF Credit (Million USD) refers to financial assistance provided by the International

Monetary Fund. It reflects economic instability and often comes with policy adjustments that influence inflation control.
https://data.imf.org/en/DataExplorer?datasetUrn=IMF.STA:FA(8.0.0)INDICATOR=CLIMF$_XDR$

Unemployment Rate (%) is the percentage of the labor force actively seeking work but unable to find employment. High unemployment typically reduces demand-driven inflation but may signal a sluggish economy.
https://data.imf.org/en/Data-Explorer?datasetUrn=IMF.RES:WEO(6.0.0)

The dataset is cleaned, free of missing values, and organized appropriately for modeling.

4 Estimation of the Models

This section outlines the statistical and machine learning models employed to predict the target variable.

4.1 ARIMA Model

The Autoregressive Integrated Moving Average (ARIMA) model is widely used for forecasting univariate time series data. It captures temporal dependencies using autoregressive terms (AR), moving average terms (MA), and differencing (I) to ensure stationarity.

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t \tag{1}$$

where:

- y_t is the observed value at time t,
- c is a constant term,
- ϕ_i are the autoregressive coefficients,
- θ_i are the moving average coefficients,
- ϵ_t is white noise.

The ARIMA(p, d, q) model is selected based on the autocorrelation and partial autocorrelation plots along with AIC/BIC criteria.

4.2 Ridge Regression

Ridge Regression is a regularized linear regression method that penalizes large coefficients using an L_2 norm. This helps reduce multicollinearity and overfitting, especially when the number of predictors is high.

$$\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{2}$$

Key Features:

- Shrinks coefficients but does not set them exactly to zero.
- Useful when all predictors contribute and multicollinearity exists.

4.3 LASSO Regression

LASSO introduces L_1 regularization, which can shrink some coefficients exactly to zero, performing both variable selection and regularization.

$$\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{3}$$

Key Features:

- Performs feature selection by eliminating less important variables.
- Helps build interpretable models with fewer predictors.

4.4 Elastic Net

Elastic Net is a compromise between Ridge and LASSO, incorporating both L_1 and L_2 penalties. It is especially useful when predictors are highly correlated or when the number of predictors exceeds the number of observations.

$$\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\} \tag{4}$$

Advantages:

- Encourages both sparsity and grouping of correlated variables.
- Addresses limitations of LASSO when variables are highly correlated.

4.5 Random Forest

Random Forest is a robust ensemble learning technique based on decision trees. It constructs a large number of trees on bootstrapped samples and averages the predictions to reduce variance.

Key Concepts:

- Each tree is trained on a random subset of data and features.
- Output is the mean prediction (for regression) or majority vote (for classification).
- Helps reduce overfitting and improves generalization.

4.6 XGBoost

XGBoost (Extreme Gradient Boosting) is a powerful tree-based ensemble algorithm that sequentially builds trees to correct the errors of previous ones. It introduces regularization to control complexity and overfitting.

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{5}$$

where:

- $l(y_i, \hat{y}_i)$ is a differentiable loss function,
- $\Omega(f_k)$ is a regularization term that penalizes model complexity.

Advantages:

- Efficient and scalable.
- Handles missing data internally.
- Regularization avoids overfitting.

5 Results

5.1 Summary Statistics

Summary statistics including mean, median, mode, quartiles for each variable are as follows:

Year	CPI	GDP_growth	oil_price	exchange_rate	FDI
Min. :1983	Min. : 2.529	Min. :1.014	Min. :10.87	Min. : 13.12	Min. :2.946e+07
1st Qu.:1992	1st Qu.: 4.603	1st Qu.:3.607	1st Qu.:15.96	1st Qu.: 24.76	1st Qu.:2.956e+08
Median :2000	Median : 7.645	Median :4.452	Median :26.04	Median : 55.70	Median :7.194e+08
Mean :2000	Mean : 7.907	Mean :4.664	Mean :38.67	Mean : 54.85	Mean :1.244e+09
3rd Qu.:2009	3rd Qu.:10.074	3rd Qu.:6.077	3rd Qu.:57.19	3rd Qu.: 82.58	3rd Qu.:1.774e+09
Max. :2018	Max. :20.286	Max. :7.831	Max. :95.99	Max. :121.82	Max. :5.590e+09

imports_pct	external_debt	broad_money_pct	IMF_loans	unemployment_rate
Min. :11.83	Min. :23.28	Min. :28.69	Min. : 411.4	Min. :3.071
1st Qu.:17.46	1st Qu.:27.38	1st Qu.:39.68	1st Qu.: 972.9	1st Qu.:4.812
Median :19.13	Median :34.99	Median :45.07	Median :1239.6	Median :5.841
Mean :18.59	Mean :38.09	Mean :43.26	Mean :1954.4	Mean :5.539
3rd Qu.:20.44	3rd Qu.:49.10	3rd Qu.:46.81	3rd Qu.:2553.6	3rd Qu.:5.981
Max. :22.60	Max. :55.90	Max. :51.41	Max. :5672.3	Max. :8.270

Figure 1: Summary statistics

5.2 Box and Whisker Plots

Figure shows the distribution and identifies outliers across all variables.

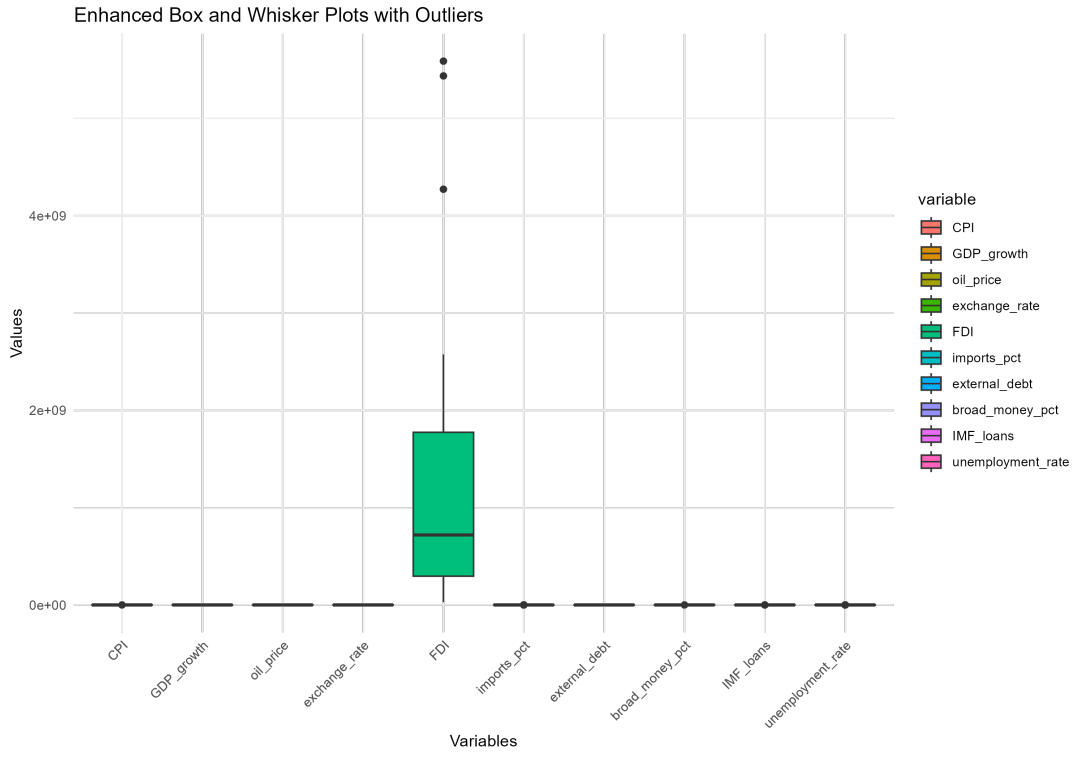

Figure 2: Box and Whisker Plots

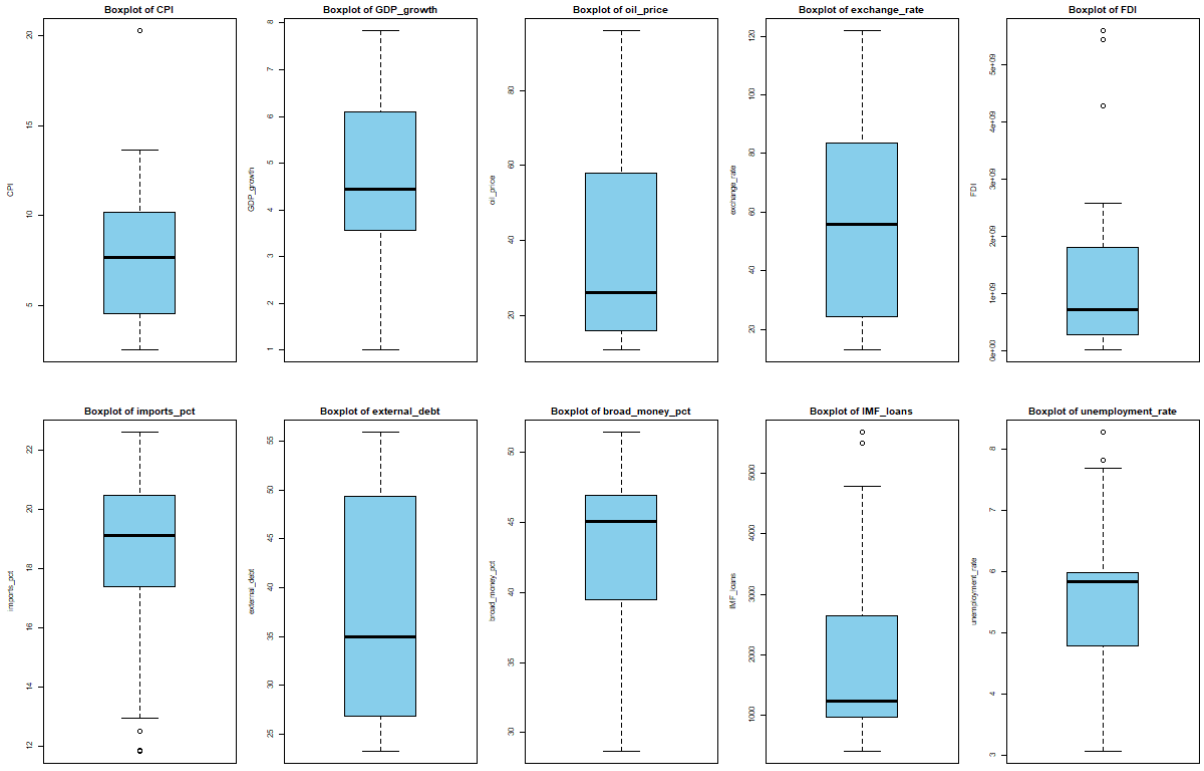

Figure 3: Individual Box and Whisker Plots

Interpretation:

1. **Broad Money (percent of GDP)**

IQR is small, tightly clustered values.

Median is near the upper quartile, suggesting a slightly left-skewed distribution.

2. **FDI**

Narrow IQR, indicating low variability.

Median near the bottom of the box, and several upper outliers.

3. **Imports (percent of GDP)**

Moderate IQR, with a few low-end outliers.

Median is slightly above the center, indicating a slightly left-skewed distribution.

4. **Exchange Rate**

Large IQR, showing high variability.

Median is centered, suggesting symmetric distribution.

5. **GDP Growth**

Medium IQR, relatively symmetric.

Median near the center, so growth is balanced around the median. Range is large

6. **Unemployment Rate**

Small IQR, tight clustering.

Upper outliers present, median near the center.

7. **External Debt**

Large IQR, high variability.

Median near the bottom, implying a right-skewed distribution.

8. **IMF Loans**

Narrow IQR, low variability overall.

Many high outliers, with median near the bottom.

9. **Oil Price**

Large IQR, wide price swings.

Median near the bottom, indicating right-skewed distribution.

10.**Consumer Price Index** Medium IQR, relatively symmetric.

Median near the center, so growth is balanced around the median. Outlier present in upper region. **Summary:**

High IQR variables (like **Exchange Rate**, **External Debt**, **Oil Price**) may be more volatile contributors to inflation.

Low IQR variables (like **FDI**, **IMF Loans**) may have less consistent influence, but outliers matter during economic shocks.

5.3 Scatter Plot Matrix

Relationships between variables are visualized in the scatter plot below.

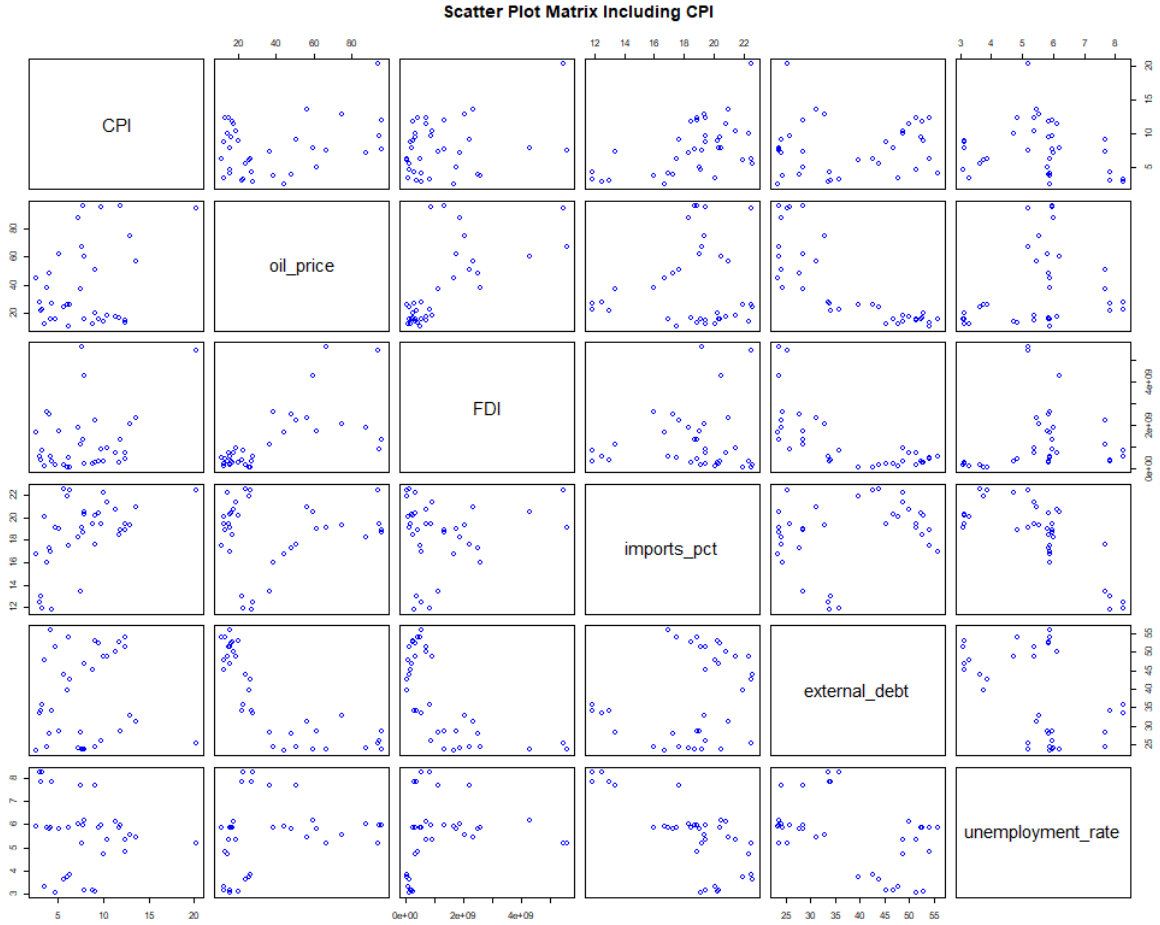

Figure 4: Scatter Plot Matrix

Interpretation:

The scatter plot reveals that CPI does not exhibit strong linear relationships with most of the selected economic indicators—oil price, FDI, imports percentage, external debt, and unemployment rate. Among the variable pairs, CPI and oil price show a slight positive trend, suggesting that higher oil prices may be associated with higher CPI.However, the pattern is weak and scattered. The relationship between CPI and FDI appears to be random, indicating no significant association. Similarly, CPI and imports percentage show no discernible trend, suggesting that import levels, as a share of GDP, may not directly influence inflation in this dataset. The CPI vs. external debt plot also lacks any noticeable structure, implying little to no correlation. A somewhat weak negative relationship is

visible between CPI and unemployment rate, which may suggest that higher inflation is mildly associated with lower unemployment. Overall, the matrix suggests that none of the independent variables strongly predict CPI on their own, highlighting the potential need for more complex modeling.

Corelation between each variable is visualized below

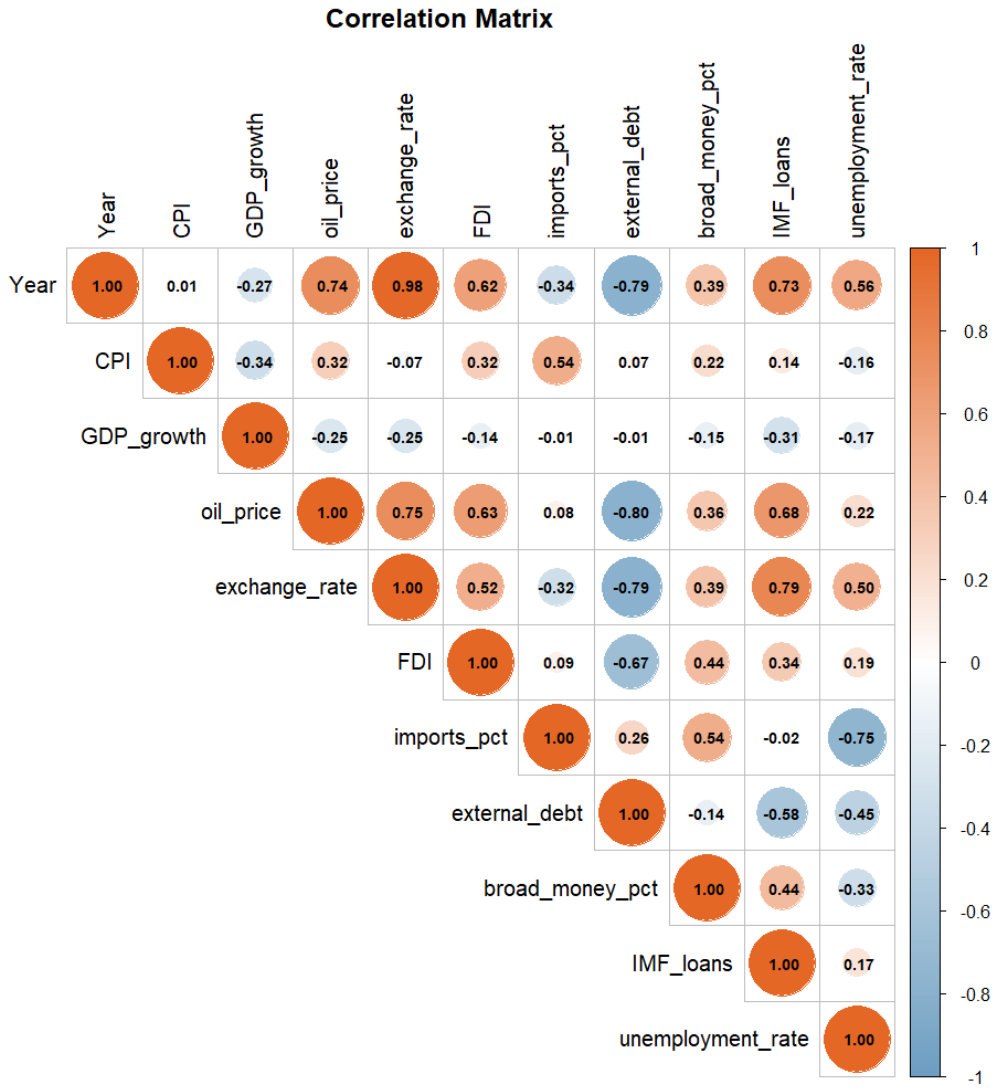

Figure 5: Corelation Matrix

5.4 Variable Importance

Variable importance for Random Forest model is as shown:

	%IncMSE	IncNodePurity
GDP_growth	7.479198564	80.10000017
oil_price	5.788640739	50.59499697
exchange_rate	7.992055014	44.00002806
FDI	4.906209912	56.16013807
imports_pct	14.31117707	106.2205005
external_debt	3.085418186	32.42652025
broad_money_pct	4.065317934	25.57771871
IMF_loans	5.716409539	49.07836869
unemployment_rate	7.428687166	35.6341398

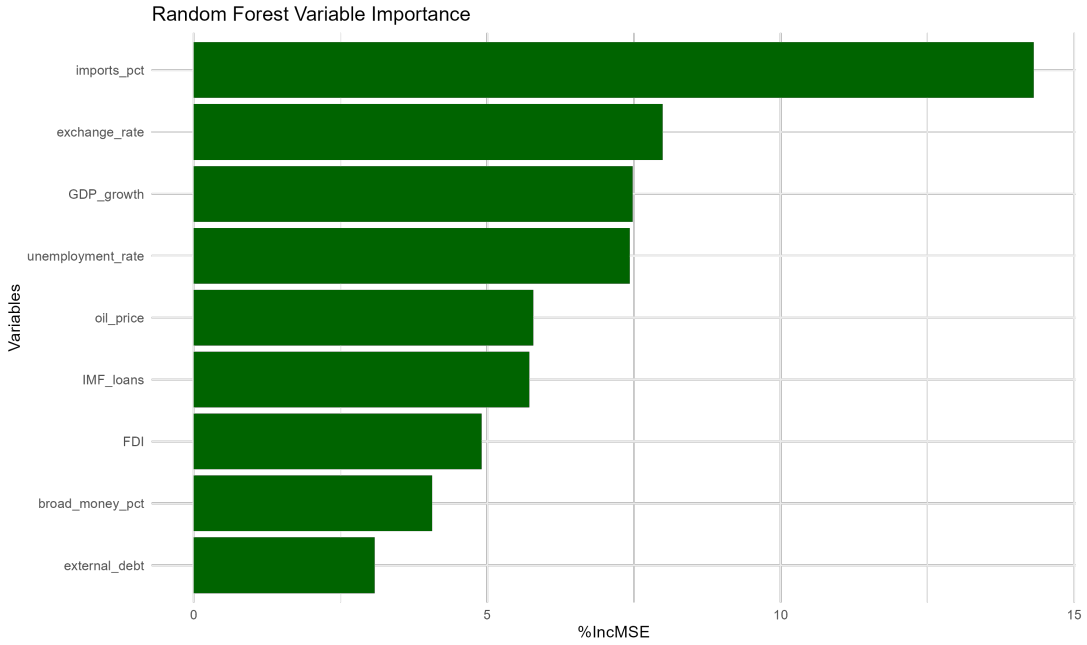

Figure 6: Top Variables by Importance (Random Forest)

Variable importance for XGBoost model is as shown:

	Feature	Gain	Cover	Frequency
1	imports_p	0.36844	0.172437	0.131132
2	FDI	0.183811	0.201921	0.119811
3	exchange_	0.136777	0.09143	0.10283
4	GDP_grow	0.117076	0.158067	0.274528
5	broad_mo	0.060209	0.057702	0.04434
6	external_d	0.054714	0.110193	0.095283
7	oil_price	0.035919	0.077656	0.128302
8	IMF_loans	0.028486	0.078773	0.059434
9	unemployr	0.014568	0.05182	0.04434

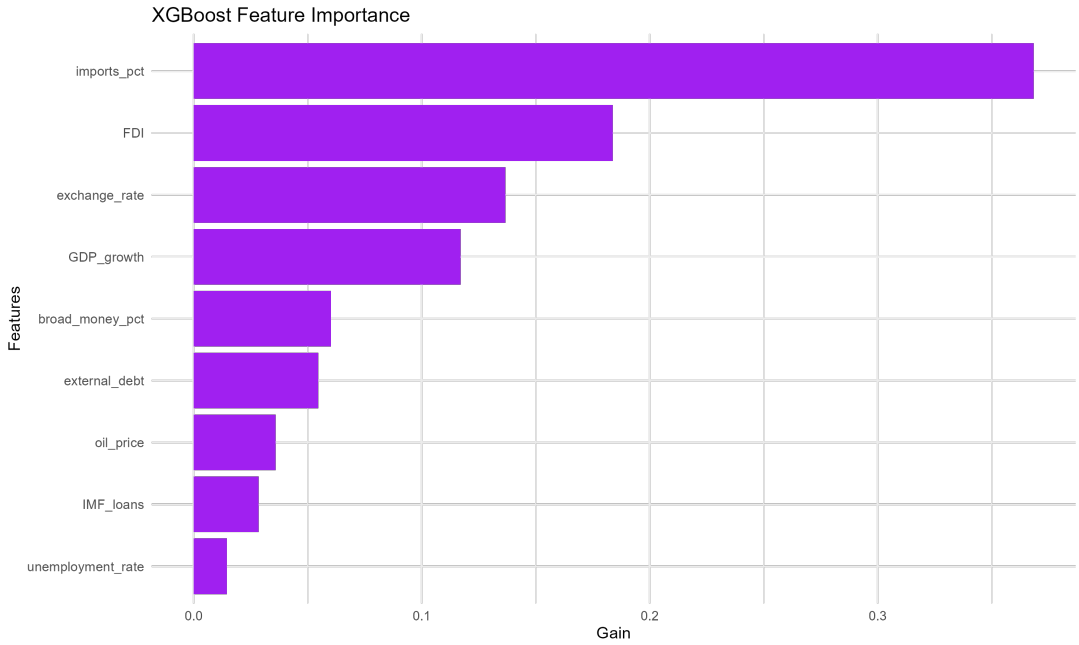

Figure 7: Top Variables by Importance

5.5 Predicted vs Actual

The actual vs. predicted CPI values are plotted in Figure below.

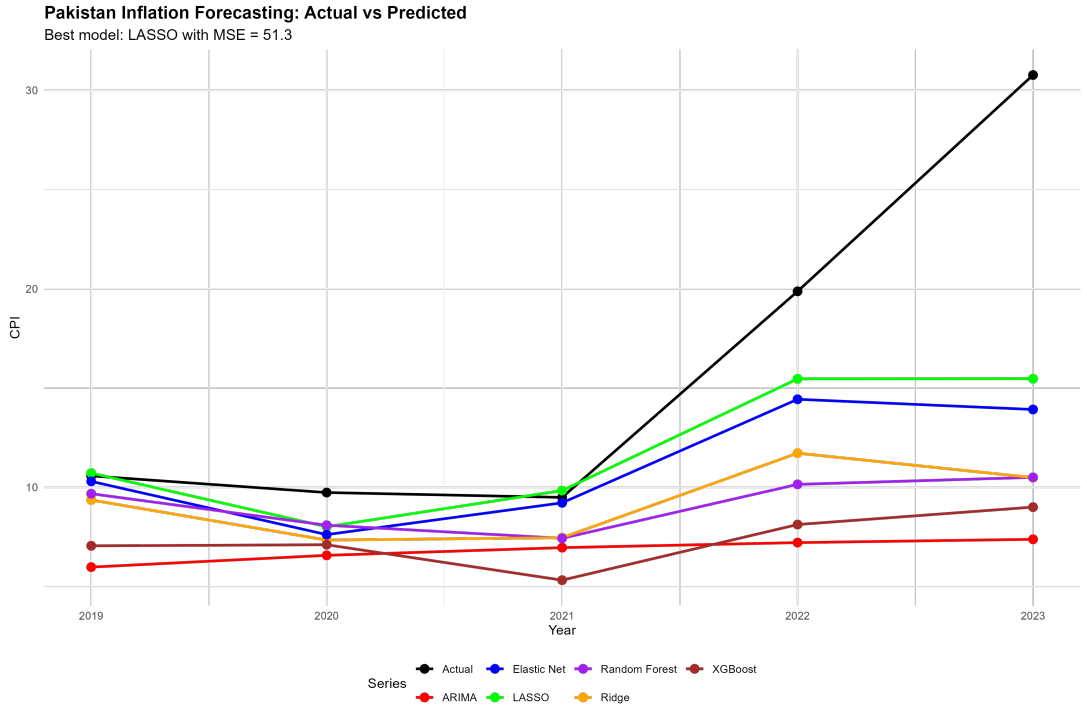

Figure 8: Actual vs Predicted CPI (2019–2023)

Model	MSE	Rank
LASSO	51.30281425	1
Elastic Net	63.58807723	2
Ridge	97.77545539	3
Random Forest	102.5758411	4
XGBoost	129.6521169	5
ARIMA	148.8895364	6

Figure 9: MSE values for each model

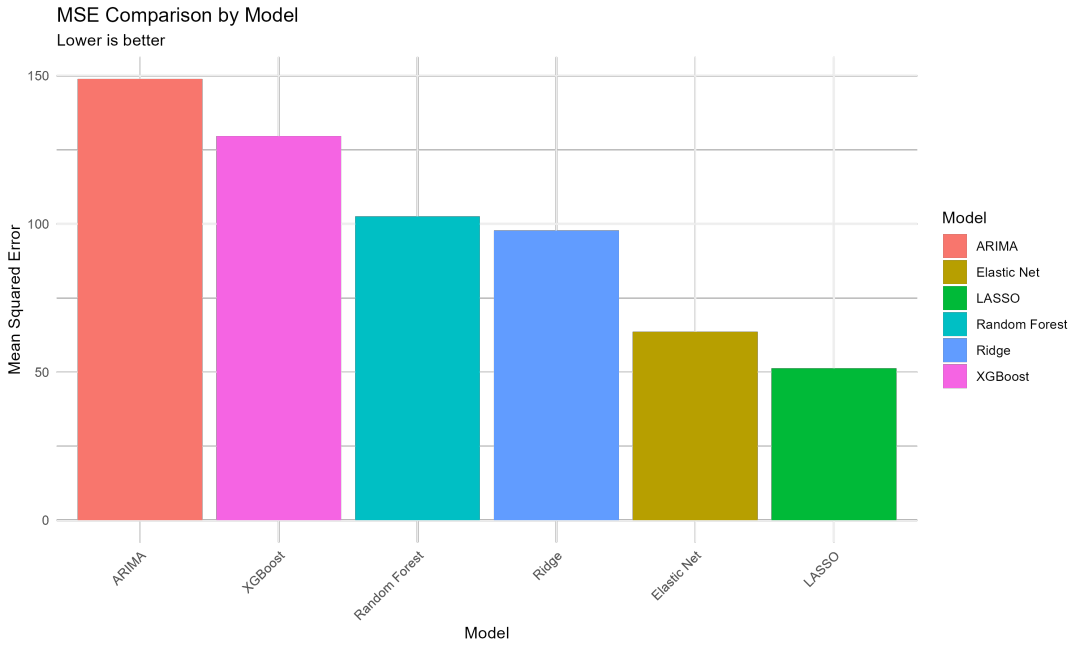

Figure 10: Graph of mse comparison by modal

Interpretation:

The LASSO model performed best with the lowest Mean Squared Error (MSE = 51.3), followed by Elastic Net and Ridge, indicating strong predictive accuracy.

Tree-based models (XGBoost and Random Forest) showed higher errors, while the traditional ARIMA model performed the worst with the highest MSE (148.89).

As seen in the graph, all models predicted CPI values fairly accurately from 2019 to 2022.

However, in 2023, the actual CPI rose sharply to around 30, a significant increase that most models failed to capture accurately.

Despite this, LASSO and Elastic Net still followed the trend more closely than the others.

While all models handled earlier years well, only regularized linear models like Elastic Net and LASSO showed resilience in adjusting to sharp changes, making them more suitable for CPI forecasting in volatile periods.

6 Conclusion

The analysis reveals that while individual macroeconomic indicators such as oil prices, external debt, and exchange rates exhibit high variability and potential influence on inflation, most show weak or no linear correlation with the Consumer Price Index (CPI), as evidenced by the scatter plots and correlation matrices. Box-and-whisker plots highlight the presence of significant outliers and skewness in several variables, suggesting that inflation is driven by complex, non-linear interactions rather than any single factor.

Among the predictive models evaluated, the lasso model demonstrated the highest accuracy, achieving the lowest Mean Squared Error (MSE = 51.3), followed by Elastic net and Ridge regression. In contrast, tree-based models and the ARIMA model performed comparatively worse. While all models predicted CPI trends reasonably well through 2022, they failed to capture the sharp spike in 2023, indicating their limitations in forecasting sudden economic shocks.

Overall, the findings emphasize that a multivariate, regularized approach enhances inflation prediction. However, further refinement is necessary to better account for unexpected macroeconomic disruptions.

Limitations of ARIMA

- **ARIMA assumes linearity and stationarity**, expecting future values to follow past patterns.
- **Sudden jumps or extreme values** (e.g., inflation spikes) violate these assumptions, reducing ARIMA's forecasting accuracy.
- ARIMA tends to **underpredict or oversmooth** during periods of high volatility due to its reliance on recent trends.
- **It does not adapt well to non-linear relationships or abrupt regime changes**, unlike some machine learning models.

References

1. Ahmed, S., Rehman, F., and Saeed, R. (2022). Forecasting CPI inflation in Pakistan using ARIMA models. *Pakistan Economic Review*, **3**(1), 22–35.

2. Khan, M. U., and Javed, A. (2021). Interest rate-inflation nexus: A VAR approach for Pakistan. *Journal of Economic Forecasting*, **8**(2), 45–61.

3. Fatima, H., Rauf, M., and Naeem, A. (2023). Comparing machine learning and classical models in inflation forecasting. *International Journal of Forecasting and Economics*, **12**(1), 17–31.

4. Siddiqui, T., and Hamid, A. (2020). Application of XGBoost in inflation prediction: Evidence from Pakistan. *Computational Economics Review*, **5**(3), 99–110.

5. Ali, Z., Hussain, T., and Ahmed, N. (2019). Forecasting inflation trends using deep learning: An LSTM approach. *Journal of Data Science and AI*, **7**(4), 101–115.

6. Raza, M. A., and Mehmood, F. (2024). Classical vs. machine learning models in times of economic turbulence. *Applied Economic Modelling*, **14**(2), 28–42.

7. Hussain, A., and Akram, W. (2018). Exchange rate shocks and inflation: A VECM approach. *SBP Working Papers*, **72**, State Bank of Pakistan.

8. Shabbir, M., and Iqbal, M. (2021). A review of deep learning models for time series forecasting. *Journal of Emerging Economies and Policy*, **9**(3), 88–100.

9. Tariq, H., and Noor, L. (2020). Lag selection in inflation forecasting: OLS vs. Ridge Regression. *Review of Quantitative Economics*, **6**(1), 39–54.

10. Malik, F., Saeed, K., and Zahid, R. (2023). A hybrid ARIMA-XGBoost approach to inflation forecasting. *Econometrics and Forecasting Journal*, **11**(2), 66–81.