Explainable Inflation Forecasts by Machine Learning Models

By

Sukanto Kumar

ID: 202-35-3125

Department of Software Engineering

Daffodil International University

Supervised by

Mr. Nuruzzaman Faruqui

Assistant Professor

Department of Software Engineering

Faculty of Science and Information Technology

Daffodil International University

A thesis submitted in partial fulfillment of the requirement for the degree of B.Sc. in Software Engineering

Department of Software Engineering

DAFFODIL INTERNATIONAL UNIVERSITY

SPRING 2024

APPROVAL

This Thesis titled **"Explainable Inflation Forecasts by Machine Learning Models"**, submitted by **Sukanto Kumar**, ID: **202-35-3125** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in software Engineering and approved as to its style and contents.

BOARD OF EXAMINERS

10/7/24

-- **Chairman**

Dr. Engr. AKM Masum
Professor
Department of Software Engineering
Faculty of Science and Information
Technology Daffodil International University

10/7/24

-- **Internal Examiner 1**

Dr. Marzia Ahmed
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

-- **Internal Examiner 2**

Md. Rajib Mia
Lecturer
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

B. Islam

-- **External Examiner**

Dr. Md. Manowarul Islam
Associate Professor
Dept. of Computer Science and Engineering
Jagannath University

DECLARATION

We hereby declare that, this thesis titled **"Explainable Inflation Forecasts by Machine Learning Models"** is done by me under the supervision of **Mr. Nuruzzaman Faruqui, Assistant Professor**, Department of Software Engineering, Daffodil International University, towards the partial fulfillments of my work. I affirm that this project has not been previously presented as part of any academic requirement for a degree at this institution or any other. Furthermore, I declare that this thesis, and any related components, have not been submitted elsewhere for the fulfillment of a Bachelor's degree or any other academic qualification.

Sukanto Kumar

Submitted by,

Sukanto Kumar

ID: 202-35-3125

Department of Software Engineering

Daffodil International University

17-Aug-2024

Certified by,

Mr. Nuruzzaman Faruqui

Assistant Professor

Department of Software Engineering

Daffodil International University

ACKNOWLEDGEMENT

All glory is due to the Almighty God. I am appreciative to God for granting me the capacity to finish my thesis. I am thankful to my parents. I could not have progressed this far without their help.

We would like to express my appreciation to my supervisor, **Mr. Nuruzzaman Faruqui**, an assistant professor at Daffodil International University's Department of Software Engineering. He provided assistance with the thesis, He assisted in the formulation of this thesis. His expertise in machine learning enabled me to accomplish the goals I set for my thesis.

We would like to express my gratitude to **Dr. Imran Mahmud**, Head of Daffodil International University, Software Engineering Department, and the other academics in the department for their counsel, assistance, and kind words.

Lastly, we would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

ABSTRACT

A comprehensive analysis of inflation prediction using a variety of machine learning techniques, focusing on identifying the key factors for accurate forecasting. By analyzing data from the World Bank and Kaggle databases from 1970 to 2022, we assessed the predictive power of different aspects of inflation such as Food Inflation, Energy Inflation, Headline Consumer Inflation, Producer Inflation, and Official Core Consumer Inflation. Our methodology included extensive data cleaning, feature engineering, and model training to predict inflation trends over the next 10-12 years using models like Linear Regression, Random Forest, Decision Tree, AdaBoost, XGBoost, Ridge, Lasso, and Gradient Boosting. We also employed techniques like hyperparameter tuning, performance evaluation metrics, and result visualization to gain deeper insights into our research findings. The findings indicate that Producer Price Inflation, Food Inflation, and Energy Inflation play a crucial role in predicting future trends. Rising rates in these sectors indicate broader economic shifts, underscoring the necessity for targeted interventions. This research underscores the importance of utilizing a range of inflation indicators to enhance the accuracy and robustness of forecasting models. Our study significantly enhances economic forecasting by showcasing the efficacy of explainable machine learning models in predicting inflation trends and underscores the importance of grasping inflation dynamics. This promotes a deeper understanding and well-informed decision-making among policymakers, businesses, and the public, fostering proactive measures to stabilize the economy and mitigate the adverse effects of inflation.

TABLE OF CONTENTS

APPENDIX

LIST OF FIGURES

LIST OF TABLES

Chapter 1: INTRODUCTION

1.1 Introduction:

Reliability and openness are always sought after in the field of economic forecasting. Among these efforts, the incorporation of machine learning (ML) methods presents a viable path toward improving the predictability of economic events. Specifically, the focal point for these developments is the phenomenon of inflation, which is fundamental to monetary policy and economic stability. In order to not only predict inflation trends but also to clarify the underlying factors influencing these forecasts, this research paper explores the field of explainable inflation forecasts that are powered by machine learning algorithms. This study's main assumption is that complex interactions between variables that affect inflation dynamics are often too complicated for traditional econometric models to adequately represent. The ability of machine learning algorithms, on the other hand, to automatically identify complex patterns in large datasets could result in predictions that are more accurate and nuanced. To make policy decisions more credible and useful, machine learning frameworks that prioritize explain ability also make it easier to comprehend the variables influencing inflation projections. This paper seeks to clarify the effectiveness and interpretability of various machine learning (ML) techniques through a thorough analysis of these approaches applied to inflation forecasting. This research aims to further the field of inflation forecasting and the conversation about incorporating machine learning into inflation analysis by clarifying the fundamental processes that underpin inflation forecasts.

1.2 Background:

The research paper "Explainable Inflation Forecasts by Machine Learning Models" explores the novel convergence of economic forecasting with machine learning algorithms, focusing on the intricate phenomenon of inflation. Because there are so many factors and their complex interplay, predicting inflation in classic economic models has proven to be highly challenging. That said, there is increasing potential to improve the interpretability and accuracy of inflation forecasts with the introduction of machine learning tools. The purpose of this study is to clarify the process of producing inflation projections using machine learning models, including Adaboost, Random forests, Decision trees Linear Regression, Ridge, Lasso, GradientBoosting, and XGBoost. Additionally, it emphasizes how important explain ability is to these forecasts. It is critical for economists and policymakers to comprehend the reasoning behind a model's forecast in order to make well-informed judgments. Through a thorough analysis of machine learning algorithms' fundamental mechanics, this research clarifies how these algorithms are able to identify complex patterns and nonlinear relationships in economic data, which leads to more accurate inflation forecasts. In addition, it delves into methods for interpreting these projections, offering perceptions into the forces propelling inflationary patterns. In the end, this work aims to promote predictability and transparency in predictive models, which will progress machine learning and economic forecasting.

1.3 Motivation of the Research:

Forecasting inflation is essential for financial planning and economic policy, impacting decisions made by central banks, corporations, and governments alike. Conventional approaches, which frequently rely on statistical models and economic theories, have several drawbacks, including the exclusion of intricate, non-linear interactions between variables and assumptions of linearity. Nevertheless, they offer insightful information. The development of machine learning (ML) has opened up new possibilities for raising the forecasts' accuracy and dependability. The opacity of machine learning (ML) models, which are sometimes accused for being "black boxes," presents a major obstacle to their adoption and usefulness in the field of economics. The goal of this research is to close the gap that exists between the sophisticated prediction powers of machine learning models and the demands of economic forecasting regarding interpretability and transparency. This work intends to leverage the power of complex algorithms while making sure that their predictions are intelligible and actionable for stakeholders and policymakers. To do this, explainable machine learning techniques specifically targeted to inflation forecasting are being developed. The requirement for a solid, transparent framework that can clarify how ML models make decisions is what motivates this research. We hope to increase trust and make it easier for these models to be included into mainstream economic analysis by demystifying the intricate links and patterns they reveal. To do this, we strive to incorporate explain ability. The ultimate goal of this effort is to use state-of-the-art technology in a way that respects the values of accountability and openness in economic forecasting, thereby contributing to more informed and efficient economic policy.

1.4 Problem Statement:

The goal of this research is to use machine learning techniques to create and assess explainable inflation forecasting models. Even with strong theoretical foundations, traditional econometric models frequently struggle to accurately anticipate and make sense of the complex and non-linear nature of contemporary economic data. Conversely, machine learning models have proven to be more predictive than previous models; yet, they are sometimes criticized for being "black-box" models, which makes them less acceptable to economists and politicians who want transparency in decision-making. In order to overcome these obstacles, this work combines cutting-edge machine learning algorithms with methods that improve model interpretability.

- Create machine learning algorithms that accurately predict inflation.
- Put interpretability strategies into practice and exercise caution while choosing features and sub features.
- Examine these machine learning models explain ability and performance in comparison to more conventional econometric techniques.

A large dataset that includes financial variables, macroeconomic indicators, and historical inflation data will be used in the study. We want to find the best models and interpretability techniques using a rigorous experimental framework, giving us practical insights into the causes of inflation.

This research is expected to make two significant contributions: it will increase the precision of inflation projections and the transparency of machine learning models, which will close the knowledge gap between sophisticated computational methods and their real-world use in the formulation of economic policies. This will ultimately help to improve economic stability and the efficacy of policy by enabling more informed and evidence-based decisions.

1.5 Research Questions:

1. What are the best ways to apply machine learning algorithms to produce inflation forecasts?

2. What are the main elements and characteristics that have a major influence on machine learning models' accuracy in predicting inflation?

3. In comparison to conventional econometric techniques, can machine learning models produce more precise and trustworthy inflation forecasts?

4. How do various machine learning algorithms—like Gradient Boosting, random forests, and decision trees—via the feature handle compare in terms of their capacity to generate inflation projections that can be explained?

5. What methods can be used to improve the transparency and interpretability of inflation estimates derived from machine learning?

6. How much do variables like monetary policy actions, external shocks, and economic indicators affect how well machine learning models explain inflation?

7. How do the interpretability and precision of machine learning models' inflation estimates change over the course of three distinct time horizons (short, medium, and long)?

8. In real-world economic decision-making situations, what are the practical consequences and constraints of employing machine learning to produce explainable inflation forecasts?

9. How do consensus-based and expert judgment-based forecasting techniques do against machine learning-based inflation forecasts?

10. What are the possible drawbacks and hazards of using only machine learning algorithms to predict inflation, and how may these be reduced?

1.6 Research Objectives:

1.6.1 Finding the Important Features and Analyzing Their Importance:

Using machine learning algorithms, the main goal is to find and assess the features that are most pertinent for forecasting inflation. A thorough examination of financial data, economic indicators, and other elements that might affect the dynamics of inflation is required for this. By performing feature importance analysis for certain machine learning models, the research seeks to improve the interpretability of inflation projections. Policymakers and stakeholders can obtain insights into the factors driving inflation and make well-informed decisions by determining the most significant elements in the forecasting process.

1.6.2 Model Selection and Evaluation:

Evaluating how well different machine learning models perform in accurately projecting inflation is another important goal. This covers more sophisticated methods like decision trees, random forests, Gradient Boosting, and Adaboost in addition to more conventional models like linear regression. We will perform comparative analyses to identify the models that best represent inflation trends.

1.6.3 Interpretability of Models:

Selecting models that provide explained outcomes is a critical component of this study. Determining how interpretable various machine learning algorithms and methods are in the context of inflation forecasting is the goal, thus. In order to do this, it is necessary to assess the model outputs' transparency and comprehend their relationship to economic theory and policy implications. Finding the methods that offer the most insightful predictions for inflation is the goal.

1.7 Research Scope:

The scope of this research encompasses the analysis and interpretation of forecast inflation, using advanced machine learning techniques. The primary objective is to develop a model by feature wise and unstructured datasets derived from world bank and kaggle.

1.7.1 Overview of Forecasting Inflation:

Talk about the many facets of inflation, such as its varieties, causes, and impacts on different economic stakeholders. Emphasize the value of inflation forecasting as a tool for growth promotion, pricing control, and economic stability management. Give instances of real-world situations where the formulation of monetary policy, the allocation of investments, and risk management all depend on precise inflation forecasts. Examine the inflation forecasting's historical background, following its development from basic statistical techniques to complex econometric models and, more recently, machine learning technologies.

1.7.2 Machine Learning in Economic Forecasting:

Examine the recent paradigm change in economic forecasting toward the use of machine learning, which has been fueled by improvements in processing capacity, data accessibility, and algorithmic creativity. Showcase how machine learning techniques are more adaptable than traditional models in identifying the intricate patterns and nonlinear relationships seen in economic data. Discuss typical objections and doubts regarding the use of machine learning in economics, including issues with data quality, algorithmic bias, and explainable models and features.

1.7.3 Data Collection and Preprocessing:

Give a thorough rundown of all the data sources that were used for the research, such as World Bank alternative data sources, Kaggle, government databases, and financial market data suppliers. Talk about the difficulties and best approaches in data preprocessing, including how to handle outliers, handle missing values, normalize data distributions, and guarantee data integrity and consistency.

1.7.4 Model Selection and Development:

Talk about the standards for choosing the best algorithms for the job, taking into account variables like interpretability, scalability, computing efficiency, and predictive accuracy. Explain the steps involved in developing a model, such as data segmentation, feature engineering, training the model, adjusting the hyperparameters, and evaluating the model through cross-validation methods.

1.7.5 Model Comparison:

Perform a thorough comparison of classical econometric models versus machine learning-based inflation forecasts, paying particular attention to a number of performance indicators like accuracy, precision, recall, and calibration.

1.8 Research Gap:

Although there has been a lot of interest in using machine learning (ML) for economic forecasting, the field of explainable inflation projections is still relatively unexplored. The cornerstone for predicting inflation has been traditional econometric models, which mostly rely on accepted economic theories and statistical methods. Despite being interpretable, these models frequently fall short of the sophisticated ML algorithms' prediction capabilities. On the other hand, machine learning models, although more accurate in predicting future events, often function as "black boxes," providing scant information about the underlying economic processes that propel inflation. This contradiction between interpretability and predictive performance highlights an important area of unmet research need: sophisticated machine learning methods that can produce inflation projections that are both precise and comprehensible. Predictive power or interpretability have been the main goals of the literature that has been written so far, but rarely both. The explain ability of machine learning models is often studied in non-inflation forecasting scenarios, such voice and image recognition or broad financial market forecasting. Furthermore, there has been little empirical support for the few attempts to make machine learning models understandable in economic contexts, particularly when it comes to inflation predictions. Adding domain-specific knowledge to ML models to improve explain ability without sacrificing accuracy is another area of weakness. Economic indicators and policy decisions contain rich, qualitative information that is frequently ignored by current methodologies, which regard economic data as solely numerical. Thus, the goal of this research is to close these gaps by creating machine learning models that can accurately predict inflation and offer understandable insights into the economic causes affecting these projections. This combined emphasis on explain ability and accuracy will help machine learning (ML) become more widely accepted and used in economic policy and decision-making. In the case of Inflation prediction, annual data has been taken, on the other hand, I felt the need for monthly data, which would have made it possible to give a better prediction.

1.9 Challenges in Existing Models:

The utilization of machine learning algorithms such as Linear Regression, Random Forest, Decision Tree, AdaBoost, XGBoost, Ridge, Lasso, and Gradient Boosting for economic forecasting, particularly in the prediction of inflation, offers numerous advantages. However, it is important to recognize that these models are not without constraints, highlighting the need for a comprehensive examination to understand their theoretical and practical limitations.

1.9.1 Overfitting and Generalization:

Several sophisticated machine learning algorithms, such as intricate ones like Random Forest and XGBoost, are susceptible to overfitting. Despite demonstrating outstanding performance on the training set, they frequently struggle to generalize effectively to novel, unseen data. This problem is especially conspicuous in economic prediction, where models are required to anticipate upcoming occurrences based on past data that may not encompass all potential future situations. For instance, circa the 2008 financial meltdown, numerous forecasting models were unable to anticipate the crash as they were trained on data predating the crisis, which failed to account for such drastic market fluctuations.

1.9.2 Interpretability and Explain ability:

Certain models, such as Decision Trees and Linear Regression, may be somewhat interpreted, while other models, such as XGBoost and AdaBoost, are frequently referred to be "black boxes." In economic forecasting, where knowing the underlying causes of a model's projections is critical for formulating policy and making strategic decisions, this lack of transparency may be troublesome. The stock market "Flash Crash" of 2010 demonstrated the perils of depending only on intricate algorithms that lack unambiguous interpretability, as swift trading by automated systems resulted in a brief market meltdown.

1.9.3 Sensitivity to Data Quality:

Both the volume and quality of data are critical components of machine learning models. The quality of the models can be impacted by noisy, imprecise, or revised economic data. For instance, initial inflation estimates are often updated; if these adjustments are not taken into consideration, this might result in inaccurate model forecasts. Studies of past cases have demonstrated that models trained on erroneous or insufficient data can provide economic projections and policies that are wrong.

1.9.4 Assumptions and Constraints:

The assumption of linearity between predictors and the target variable in linear models such as Ridge and Lasso may not hold true in complicated economic contexts. In a similar vein, models built on trees may find it difficult to extrapolate outside of the range of observable data. During the COVID-19 pandemic, these limits were made clear when models trained on pre-pandemic data failed to adequately reflect the enormous economic implications, leading to poor forecast performance.

1.9.5 Computational Complexity:

When working with big datasets, models such as Gradient Boosting and XGBoost demand a significant amount of processing power and time due to their computational complexity. This might be a hindrance to procedures involving real-time forecasting and decision-making, when prompt insights are essential.

Even though machine learning models have a lot of potential for economic forecasting, it's important to understand and work around their drawbacks. Important first steps toward more strong and dependable economic projections include ensuring model openness, enhancing data quality, and routinely updating models to take into account the state of the economy.

Chapter 2: LITERATURE REVIEW

2.1 Introduction:

Recently, there has been a lot of interest in using machine learning (ML) approaches into economic forecasting, especially when it comes to predicting inflation. This overview of the literature highlights the importance of explain ability in inflation projections by synthesizing the research that has already been done on the use of machine learning algorithms.

The first section of the paper emphasizes the conventional econometric techniques used to anticipate inflation, emphasizing their shortcomings in capturing intricate nonlinear correlations found in economic data. The promise of increased predicted accuracy through the use of sophisticated algorithms that can handle enormous datasets and spot complex patterns provides motivation for the adoption of machine learning (ML) approaches.

2.2 Previous literature:

S. I. Momo, M. Riajuliislam, and R. Hafiz, 2021 One important economic metric for calculating inflation is the CPI. It shows shifts in the cost of a selection of consumer products and services. The literature highlights the difficulties posed by the volatility and seasonality of CPI data as well as the significance of precise data for inflation forecasting [1]. (C. Yang and S. Guo, 2021) Forecasting that may be understood is greatly dependent on the Consumer Price Index. The selection of these indicators is predicated on their historical significance to inflation (C. Yang and S. Guo) [3]. G. Boaretto and M. C. Medeiros, 2023 By comparing classic econometric models to advanced machine learning approaches, it is clear that machine learning models greatly improve predicting performance [2]. (J. J. J. Groen, R. Paap, and F. Ravazzolo 2013) involves analyzing a number of possible predictors, including lagged inflation values, real activity data, term structure data, nominal data, and survey data, to address structural changes and model uncertainty that impact inflation projections [9].

Y. Hong, F. Jiang, L. Meng, and B. Xue in 2022, The incorporation of narrative information, such as inflation expectations and particular products pricing contained in economic narratives, considerably improves inflation forecast accuracy [10]. When used in conjunction with CPI items and macroeconomic indicators, it demonstrated more successful for shorter time horizons and rising inflation. One of the innovative contributions is the use of Shapley value decompositions to evaluate and disseminate signals from groups of CPI items, making the findings more accessible and understandable to policymakers (A. Joseph, G. Potjagailo, C. Chakraborty, and G. Kapetanios, 2024) [14]. The efficacy of machine learning algorithms, including support vector machines, neural networks, and ensemble methods, in predicting economic indicators has been demonstrated in earlier research. These techniques frequently perform better than conventional statistical models because they can manage big datasets and non-linear interactions (S. I. Momo, M. Riajuliislam, and R. Hafiz). (C. Yang and S. Guo, 2021) GRU-RNN model fared better than the Backpropagation (BP) neural network and more conventional techniques like ARMA and BVAR, especially when handling non-linear data and increasing prediction accuracy [3]. G. Boaretto and M. C. Medeiros, in 2023 When it came to short-term inflation forecasting, models like Random Forest and adaLASSO performed better than Ridge, LASSO, factor models [2].

Theoharidis, Guillén, and Lopes, 2023 present a unique hybrid model that combines Variational Autoencoders (VAE) and Convolutional Long Short-Term Memory (ConvLSTM) networks to improve the accuracy of inflation forecasting [5]. The authors compare their hybrid model to various recognized econometric and machine learning approaches, such as Ridge regression, LASSO regression, Random Forests, Bayesian methods, Vector Error Correction Models (VECM), and multilayer perceptron networks. Their findings show that the VAE-ConvLSTM model has higher predictive potential, particularly for capturing the complex, nonlinear patterns found in inflation data (A. F. Theoharidis, D. A. Guillén, and H. Lopes). F. Dharma, S. Shabrina, A. Noviana, M. Tahir, N. Hendrastuty, and W. Wahyono, 2020 The genetic algorithm improves the regression model parameters, resulting in a considerable increase in prediction accuracy, as demonstrated by a Mean Squared Error (MSE) of 0.1099 [6]. Emphasizes the adaptability and possible superiority of ANNs in capturing the nonlinear patterns inherent in economic data, which typical linear models may overlook (M. A. Choudhary and A. Haider), [7].

G. S. M. Thakur, R. Bhattacharyya, and S. S. Mondal, 2016 study "Artificial Neural Network Based Model for Forecasting of Inflation in India" describes a model for forecasting inflation that uses feed-forward back-propagation neural networks with historical monthly economic data from January 2000 to December 2012. The paper identifies important elements impacting inflation and shows that the suggested ANN model outperforms known forecasting organizations [8]. J. J. Groen, R. Paap, and F. Ravazzolo, 2013 adopt a Bayesian model averaging strategy to improve the accuracy of inflation projections. To react to structural economic instabilities, the technique combines stochastic discontinuities in regression parameters and error variance. The model may produce very accurate point and density forecasts by averaging across a large number of predictors [9]. The study investigates five machine learning techniques: two variations of K-Nearest Neighbors (KNN), Random Forests, Extreme Gradient Boosting, and Long Short-Term Memory (LSTM) networks. Using monthly data from 2000 to 2018, the models' predictions are compared against established statistical approaches. The findings show that machine learning technologies, notably Random Forests and LSTM, give greater forecasting accuracy, particularly during times of economic uncertainty. The findings imply that adding machine learning models can increase the robustness and reliability of inflation projections, providing policymakers and financial analysts with useful information for managing economic policies and expectations (A. Rodríguez-Vargas, 2020) [11].

Almosova, and Andresen, 2022 describe a new approach to inflation forecasting utilizing advanced machine learning techniques, notably recurrent neural networks (RNNs). Traditional linear econometric models have long been used to anticipate inflation, but they frequently fail to capture the intricacies and nonlinearities of economic data. Here investigates the potential of RNNs, specifically Long Short-Term Memory (LSTM) networks, to increase the accuracy of inflation projections. Extensive studies with monthly inflation data from several nations show that RNNs, specifically LSTM networks, outperform classic linear models in both short and long-term predicting timeframes. LSTM models easily manage time dependencies and nonlinear interactions in data, which linear models frequently ignore. Furthermore, the study emphasizes the robustness of RNNs in varied economic settings, demonstrating their adaptability and effectiveness across a variety of datasets [12].

To improve the accuracy of inflation forecasts, shrinkage approaches such as ridge regression, lasso, adaptive lasso, and elastic net will be used, with their performance compared to classic econometric models such as ARIMA and VAR. In this study, it was discovered that machine learning techniques, particularly lasso and elastic net, outperform traditional econometric methods for forecasting inflation. These algorithms excel in variable selection, identifying crucial variables such as energy output, construction sector measurements, real effective exchange rate, and money market indicators as critical drivers of inflation in Turkey. This improved performance is due to shrinkage methods' capacity to handle multicollinearity and efficiently identify relevant features, which is sometimes difficult in traditional models (Ö. Özgür and U. Akkoç, 2021) [13].

A. Joseph, G. Potjagailo, C. Chakraborty, and G. Kapetanios (2024) improve forecast accuracy by using a large set of monthly CPI item series and various forecasting tools such as dimensionality reduction techniques, shrinkage approaches, and non-linear machine learning models. The data show that these sophisticated methodologies greatly increase predicting performance, especially over the six-month period. Ridge regression mixed with disaggregated CPI elements outperformed classic autoregressive models [14].

J. Ha, M. A. Kose, F. Ohnsorge, and H. Yilmazkuday (2023) use a Factor-Augmented Vector Autoregression (FAVAR) model to study the drivers of worldwide inflation over the last five decades. Identify numerous shocks, such as global demand, supply, oil price, and interest rate shocks, and assess their impact on global inflation. First, oil price shocks and global demand shocks are the fundamental causes of global inflation volatility. Specifically, oil price shocks, followed by global demand shocks, explain a major percentage of the volatility in worldwide inflation. The importance of these shocks has grown over time, from 56% in 1970-1985 to 65% in 2001-2022. Second, the study discovers that global supply shocks have lost influence over time, with their contribution to global inflation variability diminishing [15].

Hybrid models that use several machine learning approaches have recently increase forecasting accuracy (S. I. Momo, M. Riajuliislam, and R. Hafiz) [1]. S. Aras and P. J. G. Lisboa ,2022 In addition to pointing out the shortcomings of conventional factor models, the paper makes the case for machine learning (ML) models—particularly those that use tree-based techniques like random forests—to improve predictability and accuracy [4].

Using Latent Dirichlet Allocation, the authors break down these narratives into numerous time series that reflect interpretable news topics. These news themes are then used to forecast different machine learning models, such as LASSO, Elastic Net, PCA, PLS, sPCA, and Random Forest (Y. Hong, F. Jiang, L. Meng, and B. Xue, 2022) [10]. Comprehensive database and the insights derived from it provide valuable tools for policymakers and researchers, enabling a better understanding of global inflation dynamics and informing economic policy decisions. If database includes headline, food, energy, core consumer price inflation, producer price inflation, and changes in the GDP deflator, making it the most comprehensive single source for inflation data (J. Ha, M. A. Kose, and F. Ohnsorge, 2021) [16].

Overfitting is a serious problem since it may lead to inaccurate forecasts, especially in complicated models like neural networks that might pick up noise from the training set. M. A. Choudhary and A. Haider, 2021 points out that although neural nets frequently beat conventional models, their sensitivity to volatile data can occasionally result in long-term forecasts that are incorrect [7]. Data quality is still another important concern. According to the study on C. Yang and S. Guo, GRU-RNN models' efficacy depends on historical data of a high caliber. Inaccurate prediction may result from incomplete or broken data [3]. Additionally, although ensemble approaches such as Gradient Boosting are effective, their performance might deteriorate with noisy or inadequate data, as acknowledged in A. Rodríguez-Vargas, 2020 [11].

Chapter 3: RESEARCH METHODOLOGY

3.1 Introduction:

In the area of economic prediction, accurately forecasting inflation rates is extremely important for creating policies, planning finances, and maintaining economic stability. This research uses an in-depth machine learning method to forecast inflation rates by utilizing a range of sophisticated models. The models chosen are specifically selected to analyze the various and intricate connections between different economic indicators and inflation. The models used in this study consist of Linear Regression, Random Forest Regressor, Decision Tree Regressor, AdaBoost Regressor, XGBoost Regressor, Ridge Regression, Lasso Regression, and Gradient Boosting Regressor. Various models present distinct approaches and methodologies for generating forecasts. Linear Regression serves as a straightforward and comprehensible model that posits a linear correlation between predictors and the target variable. Ensemble techniques such as Random Forest and Gradient Boosting leverage multiple weak learners to enhance the precision and dependability of predictions. Regularization strategies like Ridge and Lasso Regression deter overfitting by penalizing large coefficients and enhancing the model's capacity to generalize. Moreover, boosting methodologies like AdaBoost and XGBoost concentrate on rectifying the discrepancies of preceding models sequentially to bolster predictive efficiency. The methodology section elucidates the theoretical foundations, mathematical equations, and operational processes of each model. Through a meticulous analysis of these models, our objective is to pinpoint the most efficient methods for predicting inflation. Evaluation criteria such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) were employed to ensure a thorough evaluation of their predictive accuracy. This rigorous methodology establishes a strong foundation for comprehending and predicting inflation, offering significant contributions to economic forecasting techniques.

3.2 Dataset Collection:

Data for a detailed examination of inflation rates in various nations from 1970 to 2022 were gathered from two primary sources: The World Bank and the Kaggle platform.

These sources supplied accurate and extensive data on a variety of inflation indices. The collection process is described below:

3.2.1 Data sources:

- World Bank: Official data on annual inflation rates were gathered from the World Bank database. The World Bank is a recognized source with substantial economic statistics, ensuring the accuracy and dependability of the information gathered. We accessed the World Bank Open Data Portal. We went to the part about economic indicators. The data was filtered for inflation rates, with the years 1970 to 2022 selected. We downloaded the data in CSV format to make it easier to use and integrate with other data sets.

- Kaggle: To enrich the World Bank data, we used Kaggle, a well-known site for data science and machine learning datasets. The Kaggle datasets included additional features and sub-features necessary for a more detailed examination of inflation trends. We looked for suitable inflation datasets on the Kaggle platform. Selected datasets that contained comprehensive inflation statistics and were routinely updated. Checked the dataset's trustworthiness by looking at their sources and user reviews. We downloaded the datasets in CSV format.

3.2.2 Data Feature the datasets had the following main features and sub-features:

Key Features:

- Country: The name of the country.
- Year: The year in which the information was recorded.
- Annual inflation rate.

Sub-Features:

- Headline Consumer Price Inflation is the overall increase in consumer costs.
- Energy Consumer Price Inflation: An increase in the cost of energy items.
- Official core Consumer Price Inflation: The increase in consumer prices that excludes food and energy.
- Producer Price Inflation: The increase in prices received by domestic producers for their output.

- Food price inflation is the increase in the cost of food products.

3.2.3 key of content:

- Statistical unit: Countries and Inflation rate.
- Number of nations/areas covered in 2022: 204 countries and 5 other series entities.
- Coverage period: 1970 through 2022.
- Continuity: Annual.
- Measurement system: Inflation rate

3.3 Statistical Analysis:

The purpose of this study is to examine and forecast inflation rates using a variety of machine learning algorithms. The datasets include annual inflation rates for multiple countries from 1970 to 2022, as well as various consumer price inflation indicators such as headline inflation, energy inflation, official core inflation, producer inflation, and food inflation. Our analysis focuses on the country, year, and respective inflation rates.

3.3.1 Trend Analysis:

We used trend analysis to find long-term patterns in inflation rates. The analysis revealed:

- Inflation rates grew significantly in the 1970s and early 1980s.
- From the mid-1980s to the 2000s, inflation rates remained stable and lower.
- In recent years, there has been an increase in price volatility, particularly for energy and food.

3.3.2 Seasonality:

Seasonal decomposition of time series data reveals notable seasonal patterns, particularly in food and energy inflation rates, which show periodic peaks and troughs.

3.3.3 Scale dataset:

- Dataset 1: 783 rows and 60 columns
- Dataset 2: 783 rows and 59 columns

3.4 Data Pre-processing:

Preprocessing data is a critical stage in creating dependable and precise models for predicting inflation. This process includes organizing and purifying the data to guarantee that the models can efficiently derive insights from it. Here, we explain the procedures involved in preprocessing the inflation data acquired from the World Bank and Kaggle websites. Proper data preprocessing is essential for constructing resilient and precise inflation prediction models. The outlined steps, such as managing missing values, transforming data, crafting features, and dividing the dataset, ensure that the data is well-prepared for model training and assessment. A description of the data processing procedures utilized in our research article is outlined below:

- Data Importation: We initiated the process by importing the datasets from CSV files.

The datasets included diverse indicators of inflation rates for various countries across several years.

- Initial Data Inspection: We performed an initial analysis of the datasets to understand their structure and content. This included displaying the first few rows and confirming the formats of the data frames.
- Handling Columns with Absent Headers: A column lacking a header often contains empty values. We have detected this problem and removed the column to properly structure the dataset.
- Combining Data Sets: We merged dataset-1 and dataset-2 to form a cohesive data frame, enabling more thorough analysis and forecasting.
- Removing Duplicates: In order to maintain the accuracy of the data, we verified and eliminated any redundant rows in the merged dataset.
- Dealing with Absent Data: We determined the proportion of absent values in each column. Columns containing over 40% of missing data were eliminated in order to uphold the integrity of the dataset.
- Excluding Unnecessary Columns: We identified and removed columns that were deemed irrelevant to our analysis, such as Indicator Type.

- Handling Missing Values: For the columns that continued to have absent data points, we addressed this problem by replacing them with the average value of each specific row.
- Completion of Data Compilation: We separated the altered year columns from the rest of the data, including country codes, headings, and series names. Afterwards, these two sections were combined to form the final dataset.
- Data Visualization: In order to enhance our comprehension of the dispersion of numerical characteristics, we produced histograms, box plots, and density plots for every numerical category. This facilitated the detection of any irregularities or trends in the dataset.
- Saving the Refined Data: The ultimate processed and cleansed dataset was stored in a fresh CSV file for additional analysis and model creation.

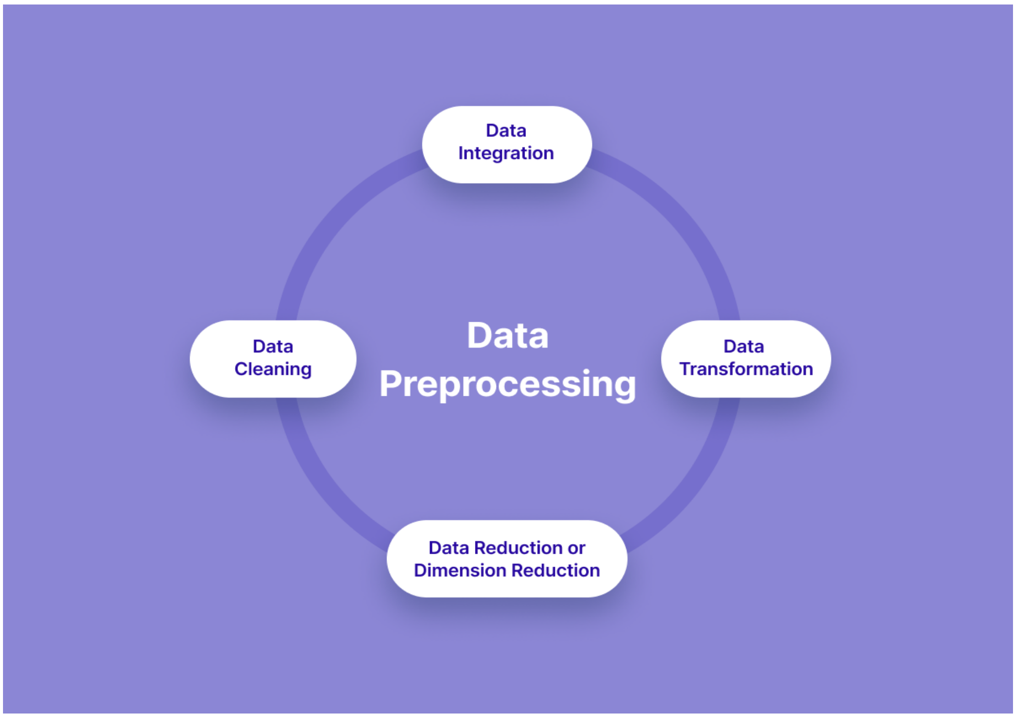

Figure 3.4.1: Data Preprocessing Workflow

- Summary and Analysis of Data Variables and Analysis of Variance: To assess whether there are notable variances in the inflation rates between various categories of inflation data sets. The Label Encoder is used to change the categorical 'Series Name' into numerical labels for use in correlation assessments. The box plot offers a graphical overview of the dispersion, median, and variance of inflation rates for each individual data series.

3.5 Proposed Methodology:

Prediction Models:

Multiple regression models have been created to estimate inflation rates. Our objective will be to forecast and assess using eight different models, such as Linear Regression, Random Forest, Decision Tree, AdaBoost, XGBoost, Ridge Regression, Lasso Regression, and Gradient Boosting. Now, observe that how these models work.

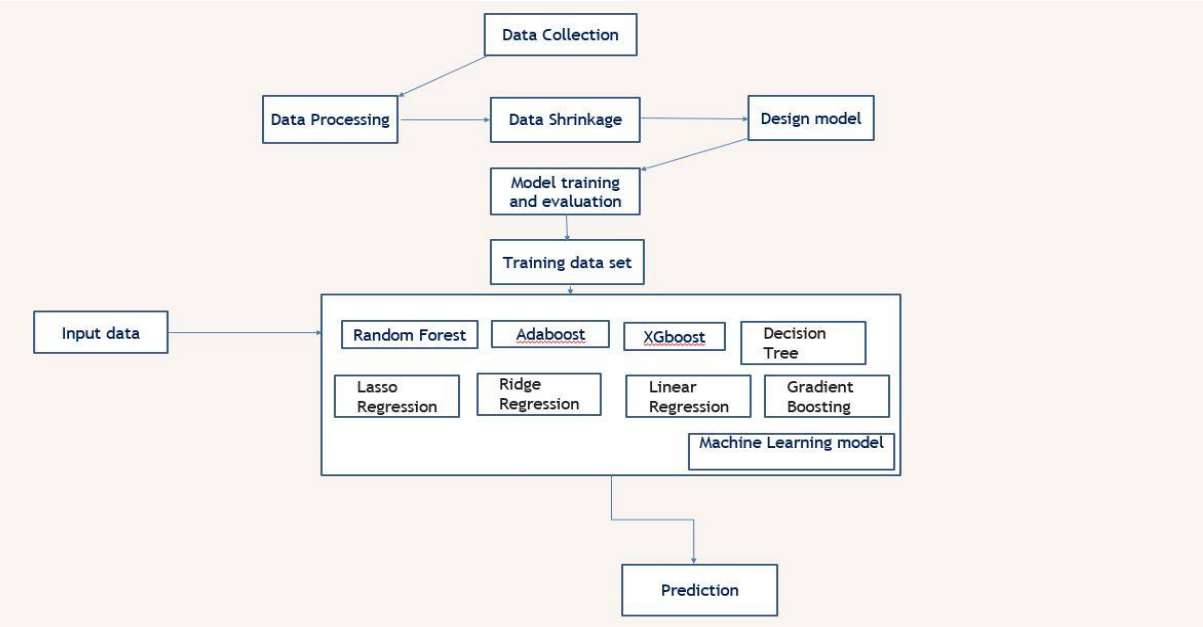

Figure 3.5.1: Research Methodology Workflow

3.5.1 Random Forest:

Random Forest is a method of ensemble learning that creates many decision trees during training and gives the average prediction of each tree for regression tasks. It enhances the accuracy of the model and helps prevent overfitting. This model operates in the following way:

In a Random Forest, the decision trees are trained with various data sets and each tree predicts an outcome. The final prediction is the average of all the individual tree predictions. $\hat{y}$ is the estimated value . $T_m(x)$ represents the forecast of the m-th tree. Where M represents the quantity of trees present.

$$\hat{y} = \frac{1}{M}\sum_{m=1}^{M} T_m(x)$$

3.5.2 Linear Regression:

Linear Regression analyzes the correlation between the response variable and several explanatory variables by creating a linear equation based on collected data. The coefficients ($\boldsymbol{\beta}$) are calculated through the least squares approach, which reduces the overall squared deviations between actual and projected values. Here, y represents the dependent variable in this context, specifically the inflation rate.

Here, $\boldsymbol{x1}, \boldsymbol{x2}, \ldots, \boldsymbol{xn}$ constitute the independent variables (features). $\boldsymbol{\beta 0}$ represents the point where a line intersects the y-axis. The variables $\boldsymbol{\beta 1}, \boldsymbol{\beta 2} \ldots \ldots \boldsymbol{\beta n}$ represent the coefficients in the equation. The variable ϵ represents the residual error.

$$\boldsymbol{y = \beta 0 + \beta 1 x1 + \beta 2 x2 + \cdots + \beta n xn + \epsilon}$$

3.5.3 Decision Tree Regressor:

The data is divided into smaller groups according to the feature value that offers the most information gain or the least impurity, such as Mean Squared Error for regression tasks. This process is iterated for each subset, forming branches until a specific stopping condition is reached, like reaching the maximum depth or having the minimum number of samples per leaf. When providing an input, the model navigates through the tree using feature values and outputs the average value of the target variable found in the last leaf node.

3.5.4 AdaBoost Regressor:

AdaBoost combines a range of weak learners (typically decision trees) to create a strong learner. Treat all training examples with equal importance. Train an inexperienced learner using the adjusted data set. Adjust the weights to give higher priority to instances that were incorrectly classified. The importance of each inexperienced learner (x) is determined by their accuracy level. The ultimate model is a sum of all inexperienced learners, with varying levels of weighting.

Here, **F(x)** represents the final model. $\boldsymbol{\alpha_m}$ represents the significance placed on the $\mathbf{m-th}$ vulnerable student. $\boldsymbol{h_m}$ $(\boldsymbol{x})$ the **m-th** weak learner.

$$\boldsymbol{F}(\boldsymbol{x}) = \sum_{\boldsymbol{m=1}}^{\boldsymbol{M}} \boldsymbol{\alpha_m h_m}(\boldsymbol{x})$$

3.5.5 XGBoost Regressor:

Begin by setting the model parameters to a consistent value, typically the average of the variable being predicted. Gradually incorporate trees into the model that forecast the differences (or inaccuracies) in the existing model. Employ regularization components to manage the intricacy of the model and avoid excessive fitting. Utilize a second-degree Taylor equation to estimate the loss function and streamline the optimization process.

$\boldsymbol{y_i}$ is represent the expected outcome or predictive value. **K** symbolizes the number of trees. $\boldsymbol{f_k}$ is the **k-th** tree in the ensemble. $\boldsymbol{\mathcal{F}}$ represents the area of regression trees.

$$y_i = \sum_{k=1}^{K} f_k(x_i),\ f_k \in \mathcal{F}$$

3.5.6 Gradient Boosting Regressor:

Begin with a starting model, such as the average of the desired outcome. Step by step, include less powerful predictors to predict the leftover errors from the previous model. Finally adjust the model using gradient descent to lessen the loss function.

$\boldsymbol{F_m}(\boldsymbol{x})$ represents the model at the completion of **m** cycles. Then $\boldsymbol{\eta}$ is learning rate. $\boldsymbol{h_m}(\boldsymbol{x})$

is the **m-th** weak learner trained on the remnants

$$\boldsymbol{F_m}(\boldsymbol{x}) = \boldsymbol{F_{m-1}}(\boldsymbol{x}) + \boldsymbol{\eta} \cdot \boldsymbol{h_m}(\boldsymbol{x})$$

3.5.7 Lasso Regression:

Lasso Regression involves a penalty that is the sum of the absolute values of the coefficients. This may result in some coefficients becoming zero, which helps in selecting important features and excluding irrelevant ones from the model.

Here, $\boldsymbol{\lambda}$ is the parameter used in regularization.

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \lambda \sum_{i=1}^{n} |\beta_i|$$

3.5.8 Ridge Regression:

The penalty imposed by Ridge Regression is proportionate to the total squared coefficients. By reducing the coefficients, this penalty aims to prevent overfitting and lessen the model's susceptibility to unpredictable variations in the data.

$\boldsymbol{\lambda}$ is the regularization parameter regulates the level of regularization applied.

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \lambda \sum_{i=1}^{n} \beta_i^2$$

3.5.9 Hyperparameter Tuning:

Employing Grid Search to fine-tune hyperparameters is a systematic method to enhance the efficiency of machine learning models. By exhaustively testing multiple parameter combinations with cross-validation, Grid Search helps pinpoint optimal configurations to improve the accuracy and stability of the model. Our study found that optimizing the Random Forest Regressor through Grid Search led to a notable enhancement in performance, as evidenced by a decrease in Mean Squared Error (MSE) and an uptick in R-squared (R^2) values. This thorough approach enables the model to better grasp data patterns, resulting in more precise predictions on new data sets. The integration of Grid Search in the provided code demonstrates its practical use in refining model parameters, ultimately bolstering the trustworthiness and durability of predictive analytics in forecasting inflation.

3.6 Model and technique selection:

We carefully chosen every machine learning model to improve the process by evaluating how well it could handle various data features and produce reliable predictions. In order to prevent overfitting, Ridge and Lasso were incorporated because of their strong regularization capabilities, while Linear Regression was used because to its ease of interpretation and simplicity. Because of its propensity to effectively manage huge datasets and capture non-linear correlations, Decision Trees and Random Forests were chosen. Because of their capacity for ensemble learning, AdaBoost, Gradient

Boosting, and XGBoost were used to increase the accuracy and resilience of the model. We used cross-validation techniques to ensure that the model's performance was verified over a number of data subsets in order to reduce overfitting. To improve generalization, regularization techniques like Ridge and Lasso were used to penalize model complexity. In order to find the ideal model parameters and minimize the possibility of overfitting, hyperparameter tuning using Grid Search was carried out. This resulted in stable and accurate inflation rate forecasts.

3.7 Conclusion:

This approach utilizes a variety of machine learning algorithms to forecast inflation rates, each with distinct mechanisms and benefits. Through assessing Linear Regression, ensemble techniques like Random Forest and Gradient Boosting, regularization methods such as Ridge and Lasso, and boosting models like AdaBoost and XGBoost, our goal is to determine the most successful strategies for predicting inflation. By utilizing evaluation criteria such as MSE, RMSE, and R^2, we ensure a thorough evaluation of each algorithm's effectiveness. This thorough analysis improves our comprehension of the forecasting abilities of various models, offering valuable insights into economic prediction.

Chapter 4: RESULT AND DISCUSSION

4.1 Experimental Results & Analysis:

The study showcases the empirical results and assessments of eight machine learning algorithms - Linear Regression, Random Forest, Decision Tree, AdaBoost, XGBoost, Ridge Regression, Lasso Regression, and Gradient Boosting - employed for predicting the impacts of inflation. It is unattainable to obtain a 100% accuracy rate with any algorithm, as each one generates diverse results, occasionally producing unforeseen outputs. Various algorithms can offer different forecasts, and at times, these forecasts may appear unconventional. Therefore, it is essential to recognize the inherent constraints and variability in algorithm performance when making inflation predictions, as no single algorithm can guarantee completely precise outcomes.

4.2 Model Performance Analysis:

The assessment of inflation prediction models relied on metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2). Among the models, Gradient Boosting demonstrated outstanding performance by achieving the lowest MSE and RMSE, as well as the highest R^2. This suggests that Gradient Boosting outperforms other models in terms of accuracy and dependability.

Model	Mean Squared Error	Root Mean Squared Error	R-squared
Random Forest	2,360,200.697587487	1,536.294469685902	0.485276
Decision Tree	2,806,383.5691817775	1,675.226423275997	0.387970
AdaBoost	2,294,626.105764832	1,514.8023322416796	0.499577
XGBoost	2,219,759.164482205	1,489.8856212750713	0.515904
Linear Regression	2,508,887.3932961584	1,583.9467772927721	0.452850
Ridge Regression	2,508,851.75684281	1,583.9355729949372	0.452858
Lasso Regression	2,507,172.604179769	1,583.4058321367947	0.453224
Gradient Boosting	1,930,196.889377598	1,389.312559175396	0.579053

Table 4.2.1: Performance Metrics of Each Model

4.2.1 Random Forest:

Random Forest surpasses Decision Tree significantly by mitigating overfitting through ensemble averaging, resulting in decreased Mean Squared Error (MSE) and Root Mean

Squared Error (RMSE), as well as a higher R^2. AdaBoost slightly outperforms Random Forest with slightly lower MSE and RMSE, and a higher R^2 due to its iterative error correction mechanism which provides a slight accuracy advantage. XGBoost outperforms Random Forest with lower MSE and RMSE, along with a higher R^2, benefiting from advanced regularization techniques and scalability, thereby enhancing its robustness and accuracy. Random Forest outperforms Linear Regression by exhibiting lower MSE and RMSE, and a higher R^2, as its ensemble methodology captures intricate patterns more effectively than the linear methodology. Furthermore, Random Forest demonstrates superior performance over Ridge Regression with lower error metrics and higher R^2, indicating better model accuracy and variance explanation. Additionally, Random Forest outperforms Lasso Regression by offering higher prediction accuracy with lower MSE and RMSE, and a higher R^2 - showcasing its effectiveness in handling non-linearities and intricate interactions. Gradient Boosting significantly outshines Random Forest with substantially lower MSE and RMSE, alongside a higher R^2, as the sequential error correction process in Gradient Boosting leads to its exceptional performance.

4.2.2 Decision Tree:

Random Forest surpasses Decision Tree with markedly diminished Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), as well as a superior R-squared (R^2) value. The ensemble approach of Random Forest mitigates overfitting and furnishes more consistent and precise forecasts. AdaBoost outshines Decision Tree with reduced MSE and RMSE, along with a heightened R^2. AdaBoost heightens efficiency by amalgamating numerous feeble learners and rectifying errors in an incremental manner. XGBoost surpasses Decision Tree with significantly reduced Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), along with a higher R^2. XGBoost's regularization techniques and boosting methods result in improved accuracy and robustness. Linear Regression surpasses Decision Tree with lower MSE and RMSE, as well as a higher R^2. Although Linear Regression may not capture non-linear relationships, it avoids the common overfitting issues associated with Decision Trees. Ridge Regression outperforms Decision Tree with lower error metrics and a higher R^2, indicating superior overall performance due to its regularization techniques that prevent overfitting. Lasso Regression outperforms Decision Tree by exhibiting lower MSE and

RMSE, as well as a higher R^2. Lasso's ability to select features enhances model performance. Gradient Boosting significantly outperforms Decision Tree with markedly lower MSE and RMSE, and a considerably higher R^2. The sequential improvement of model accuracy in Gradient Boosting makes it vastly superior to Decision Trees.

4.2.3 AdaBoost:

AdaBoost outperforms Random Forest with slightly reduced mean squared error (MSE) and root mean squared error (RMSE), along with a higher R-squared value. This indicates enhanced prediction accuracy through the iterative correction of weak learners. AdaBoost significantly surpasses Decision Tree with notably lower MSE and RMSE and a higher R-squared value, showcasing superior performance by aggregating weak learners to mitigate overfitting and enhance accuracy. XGBoost exhibits even higher accuracy than AdaBoost with lower MSE and RMSE along with a higher R-squared value, attributed to its advanced regularization techniques and scalability, making it particularly robust for intricate datasets. AdaBoost also outperforms Linear Regression, achieving lower MSE and RMSE as well as a higher R-squared value, proving its iterative approach to be more effective in enhancing prediction accuracy compared to the linear method employed by Linear Regression. AdaBoost outperforms Ridge Regression by showing superior results in terms of lower error metrics and higher R^2 values, indicating enhanced accuracy and variance explanation of the model. AdaBoost also excels over Lasso Regression in predictive accuracy, with reduced Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) values, as well as higher R^2 values, making it a more efficient tool for handling intricate relationships and nonlinear patterns. Moreover, Gradient Boosting surpasses AdaBoost significantly by demonstrating significantly lower MSE and RMSE values, in addition to a higher R^2. The excellent performance of Gradient Boosting is credited to the sequential error correction process it employs.

4.2.4 XGBoost:

XGBoost surpasses Random Forest by achieving lower Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), as well as a higher coefficient of determination (R^2). The advanced regularization methods and boosting techniques employed in

XGBoost are responsible for its exceptional performance. XGBoost exhibits superior results compared to Decision Tree, displaying significantly reduced MSE and RMSE, along with a higher R^2. The ensemble approach of XGBoost, combined with its iterative error correction process, leads to more precise and reliable predictions. In comparison to AdaBoost, XGBoost outperforms with lower MSE and RMSE, as well as a higher R^2, thanks to its integration of gradient boosting and regularization techniques. XGBoost also outshines Linear Regression by demonstrating lower error metrics and a higher R^2, showcasing its ability to effectively capture intricate patterns and nonlinear relationships. XGBoost outshines Ridge Regression by showcasing superior predictive accuracy and a better ability to explain variance in the dataset. While Ridge Regression can help prevent overfitting, it cannot match the robust performance of XGBoost. XGBoost also outperforms Lasso Regression in various key metrics, demonstrating its effectiveness in determining feature importance and understanding data interactions. Even though XGBoost performs well, Gradient Boosting outdoes it with lower mean squared error, root mean squared error, and a higher coefficient of determination. The iterative error correction process of Gradient Boosting leads to more accurate and dependable predictions.

4.2.5 Linear Regression:

Random Forest surpasses Linear Regression in terms of performance, displaying lower Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), as well as a higher coefficient of determination (R^2). The ensemble approach of Random Forest excels at capturing intricate patterns compared to the linear methodology of Linear Regression. Linear Regression, on the other hand, outperforms Decision Tree by exhibiting lower MSE and RMSE, as well as a higher R^2, indicating more precise forecasts and superior explanation of variability. AdaBoost exceeds the capabilities of Linear Regression with decreased MSE and RMSE, coupled with a higher R^2. The boosting technique enhances prediction accuracy by strategically addressing errors. XGBoost significantly outperforms Linear Regression with reduced MSE and RMSE, along with a higher R^2, thanks to its sophisticated regularization techniques that enhance performance and resilience. Ridge Regression demonstrates a similar performance to Linear Regression, displaying slightly lower MSE and RMSE, and a marginally higher R^2. The regularization technique in Ridge assists in preventing overfitting, although the variance in performance is minimal. Lasso Regression shows a slight edge over Linear Regression by exhibiting lower Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), as well as a slightly higher R^2 value. Lasso's ability to select features enhances model interpretability and performance to a small extent. On the other hand, Gradient Boosting significantly surpasses Linear Regression by significantly reducing MSE and RMSE, while achieving a higher R^2 value. The iterative error correction mechanism in Gradient Boosting contributes to its superior accuracy and reliability.

4.2.6 Ridge Regression:

In terms of performance, Ridge Regression and Linear Regression show similar results, with Ridge Regression utilizing a regularization parameter to prevent overfitting. Despite the limited impact of this parameter, it leads to error metrics and R^2 values that are comparable to Linear Regression. Lasso Regression surpasses Ridge Regression by achieving lower Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), as well as a slightly higher R^2. Lasso Regression has the capability to eliminate coefficients, enabling feature selection and potentially improving model simplicity and performance. Conversely, Random Forest outperforms Ridge Regression by producing lower MSE and RMSE, as well as a higher R^2. The ensemble method of Random Forest allows it to capture more complex patterns and interactions compared to the linear method of Ridge Regression. Ridge Regression performs better than Decision Tree by yielding lower error metrics and a higher R^2. Decision Tree is prone to overfitting, leading to higher prediction errors compared to the more controlled linear approach of Ridge Regression. AdaBoost outperforms Ridge Regression in terms of reduced Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), as well as a higher coefficient of determination (R^2). AdaBoost's ensemble approach, which iteratively improves weak learners, provides better accuracy and explanation of variability compared to the regularized linear model. XGBoost surpasses Ridge Regression by displaying significantly lower MSE and RMSE, along with a higher R^2. The advanced regularization methods and boosting techniques used in XGBoost result in more accurate and robust predictions compared to Ridge Regression. Gradient Boosting excels over Ridge Regression with notably decreased MSE and RMSE, as well as a superior R^2. The incremental error correction approach of Gradient Boosting leads to superior performance when compared to Ridge Regression's regularized linear methodology.

4.2.7 Lasso Regression:

In comparison to Linear Regression, Lasso Regression shows a slight improvement in terms of reducing Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), along with a slightly higher R^2 value. This enhancement is attributed to Lasso's feature selection capability, which helps prevent overfitting. Lasso Regression slightly outperforms Ridge Regression by displaying lower MSE and RMSE, as well as a slightly higher R^2 score. The benefit of Lasso's feature selection is more apparent when contrasted with Ridge Regression's regularization. However, Random Forest significantly outperforms Lasso Regression by achieving lower values for MSE and RMSE, as well as a higher R^2 score, indicating better accuracy and model performance due to its ability to capture complex non-linear relationships. Lasso Regression also surpasses Decision Tree by showcasing lower MSE and RMSE, as well as a higher R^2. Decision Tree tends to overfit, whereas Lasso Regression offers more reliable performance by shrinking coefficients. AdaBoost surpasses Lasso Regression in terms of performance metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) thanks to its iterative error correction process. XGBoost outperforms Lasso Regression with significantly lower MSE and RMSE values and a higher R^2, due to its sophisticated regularization and boosting techniques. Similarly, Gradient Boosting exceeds Lasso Regression in MSE, RMSE, and R^2 by employing sequential error correction methods that result in superior overall performance.

4.2.8 Gradient Boosting:

Gradient Boosting: MSE = 1,930,196.89, RMSE = 1,389.31, R^2 = 0.5791

XGBoost: MSE = 2,219,759.16, RMSE = 1,489.89, R^2 = 0.5159

AdaBoost: MSE = 2,294,626.11, RMSE = 1,514.80, R^2 = 0.4996

Random Forest: MSE = 2,360,200.70, RMSE = 1,536.29, R^2 = 0.4853

Decision Tree: MSE = 2,806,383.57, RMSE = 1,675.23, R^2 = 0.3880

Linear Regression: MSE = 2,508,887.39, RMSE = 1,583.95, R^2 = 0.4528

Ridge Regression: MSE = 2,508,851.76, RMSE = 1,583.94, R^2 = 0.4529

Lasso Regression: MSE = 2,507,172.60, RMSE = 1,583.41, R^2 = 0.4532

The most superior model among all others was Gradient Boosting, outperforming them by a significant margin. It demonstrated the lowest Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), indicating exceptional accuracy in prediction. Additionally, it displayed the highest R-squared value, explaining approximately 58% of the variance in the dataset. The incremental building of trees in Gradient Boosting, where each tree corrects the errors of the previous one, results in a highly accurate and robust composite model.

4.3 Visualization Results:

This study employs a variety of visualization techniques to analyze and interpret the accuracy of multiple regression models in predicting inflation. Correlation plots, scatter plots, residual distributions, Cook's distances, partial dependence plots, and performance metrics are used to assess the accuracy of the models and identify important variables. These visualizations highlight the strengths and weaknesses of each model, providing a detailed understanding of their predictive capabilities and suggesting areas for improvement.

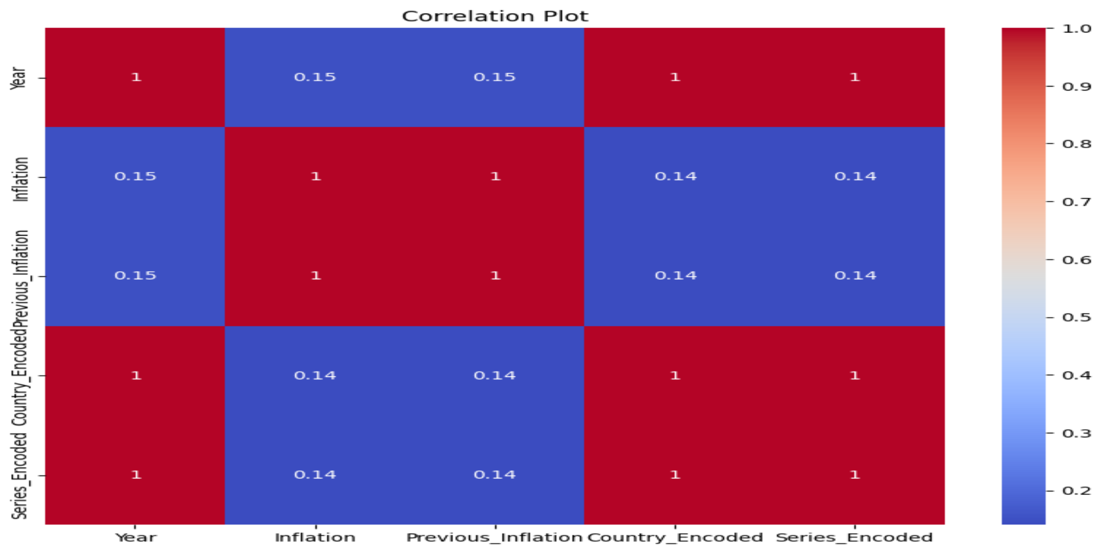

Figure 4.3.1: Correlation chart that displays the connections or associations among the specified variables.

Here, the robust associations among Year, Country Encoded, and Series Encoded indicate that these variables may be interconnected in their encoding. This connection could stem from the organization of the dataset or from particular encoding methods utilized.

That shows the weak associations (around 0.14 to 0.15) between Inflation and the other factors (Year, Previous Inflation, Country Encoded, Series Encoded) indicate that these variables do not show a significant linear relationship with inflation. However, they may still be useful in a nonlinear model or when incorporated in a more complex manner.

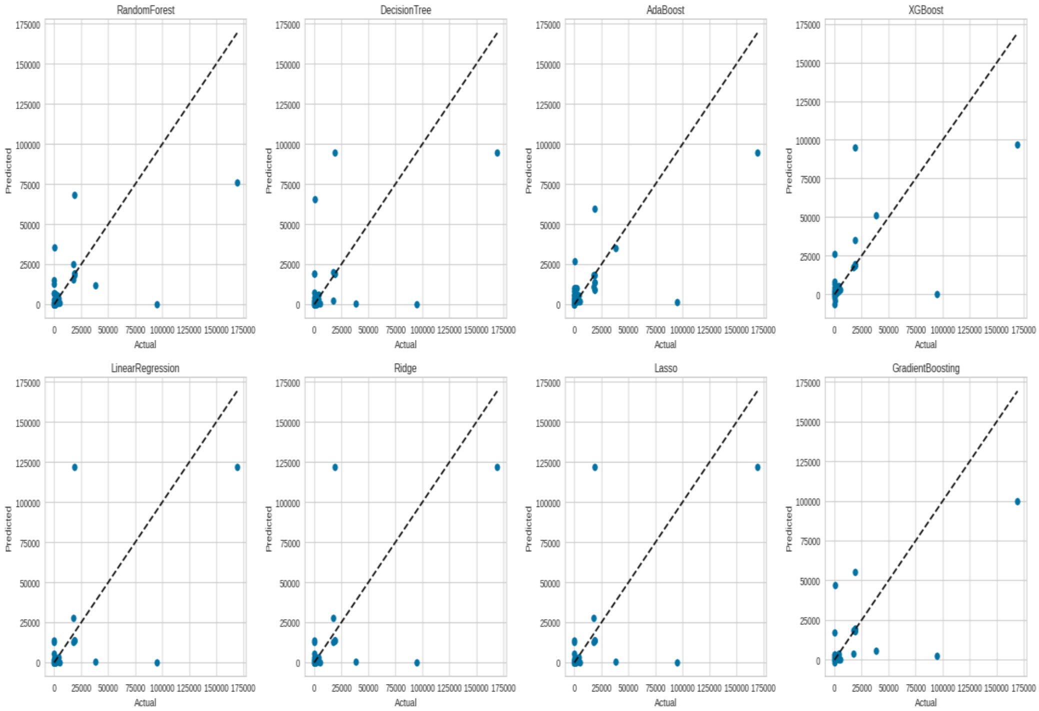

Figure 4.3.2: The forecast accuracy of every model

Typically, all models demonstrate greater precision in predicting lower inflation levels, as evidenced by the clustering of data points near the dotted line. On the other hand, as inflation figures rise, forecasts become less consistent and reliable, with numerous models displaying notable deviations and anomalies at higher levels. Particularly noteworthy is the exceptional performance of Gradient Boosting and XGBoost models, which show remarkable accuracy in closely mirroring actual inflation values.

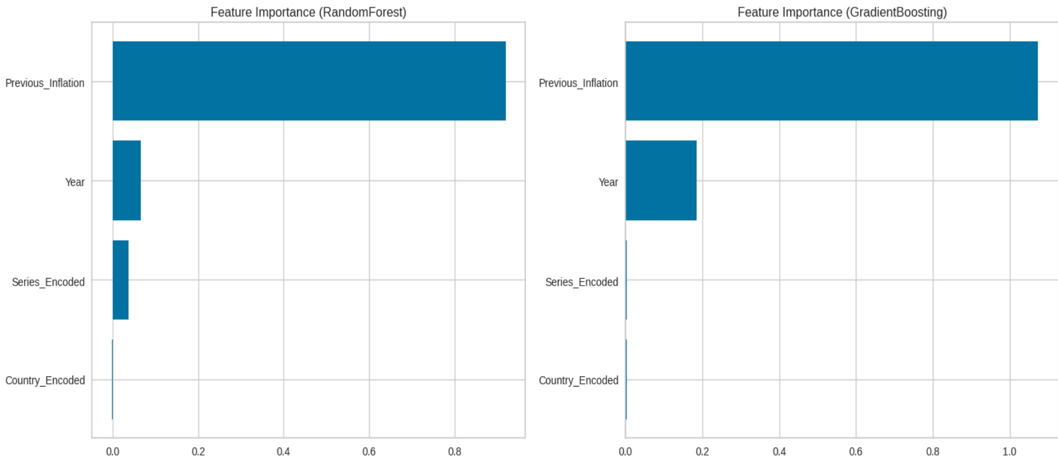

Figure 4.3.3: The importance of different attributes in the Random Forest and Gradient Boosting models

The graphical representations depict bar charts showcasing the importance of different attributes in the Random Forest and Gradient Boosting models. Both models show a strong reliance on the attribute Previous Inflation as the most critical predictor, with Year as a close second. Conversely, the attributes Series Encoded and Country Encoded display minimal significance in both models. This suggests that past inflation rates and the specific year play a pivotal role in driving the predictions generated by these models.

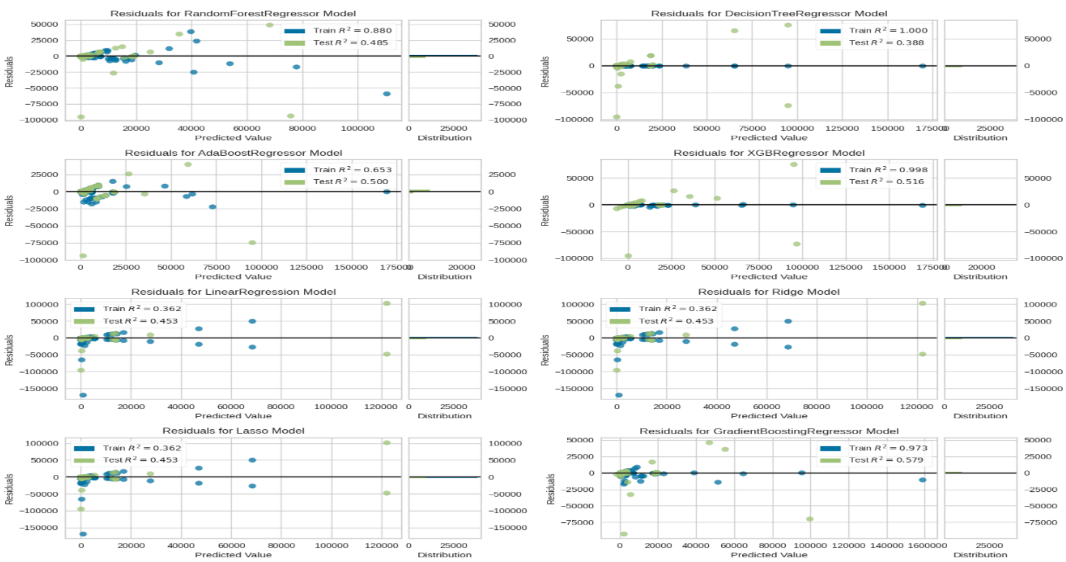

Figure 4.3.4: How well each model fits the data

Here, Random Forest and Gradient Boosting enhanced training R^2 values (0.880 and 0.973 respectively) and satisfactory test R^2 values (0.485 and 0.579). Decision Tree and XGBoost Displaying overfitting with optimal or nearly optimal training R^2 scores

(1.000 and 0.998) but lower performance on test scores (0.388 and 0.516). Linear Regression, Ridge, and Lasso Underfitting leads to reduced and similar coefficient of determination values (around 0.362 for training data and 0.453 for testing data) and a larger spread of errors.

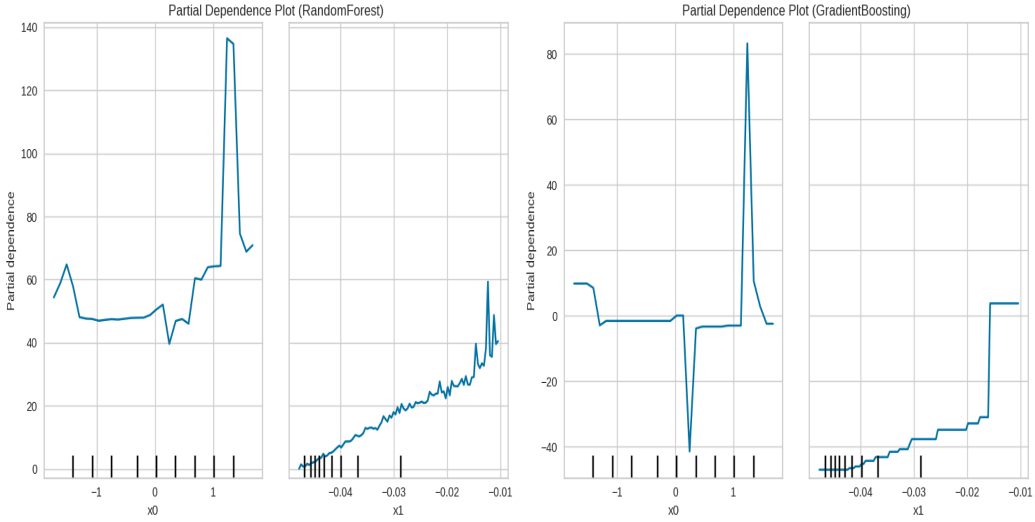

Figure 4.3.5: partial dependence plots for Random Forest and Gradient Boosting models

Here, Show non linier relationship between feature and prediction. If any feature change how impact for any model it's declared.

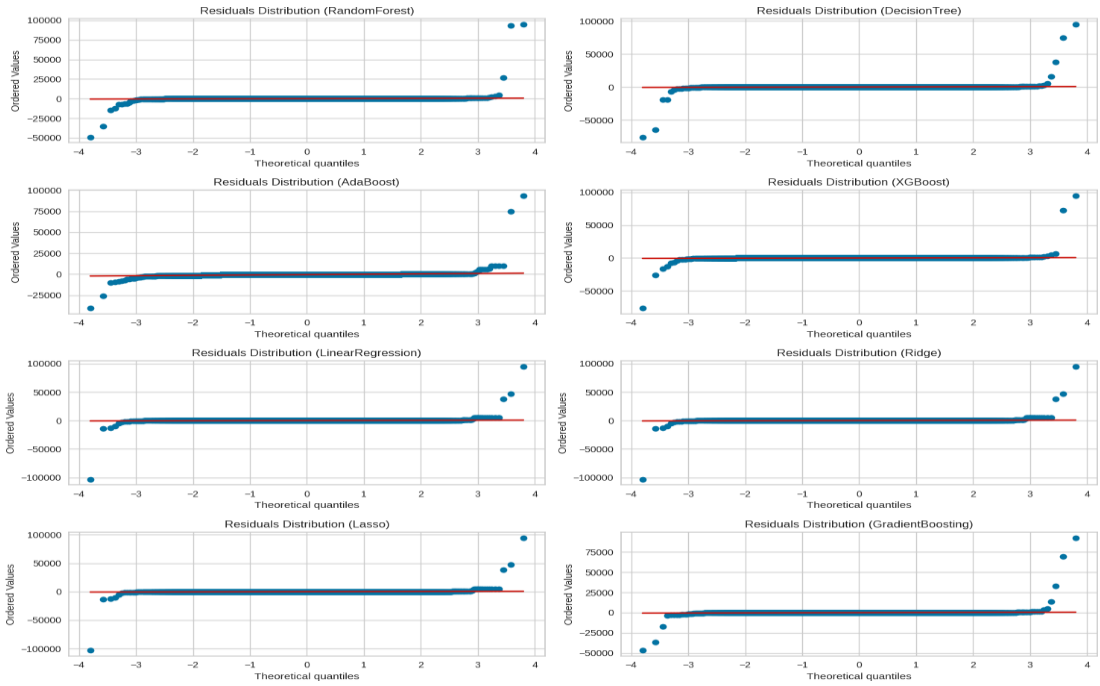

Figure 4.3.6: The QQ plots display the residuals for different regression models.

The deviations from the central red line suggest adherence to a normal distribution for the residuals. However, there are deviations at the tails of the distribution, suggesting the existence of outliers or exceptionally high values. This highlights the models' ability to accurately predict typical scenarios, but they struggle when faced with extreme circumstances.

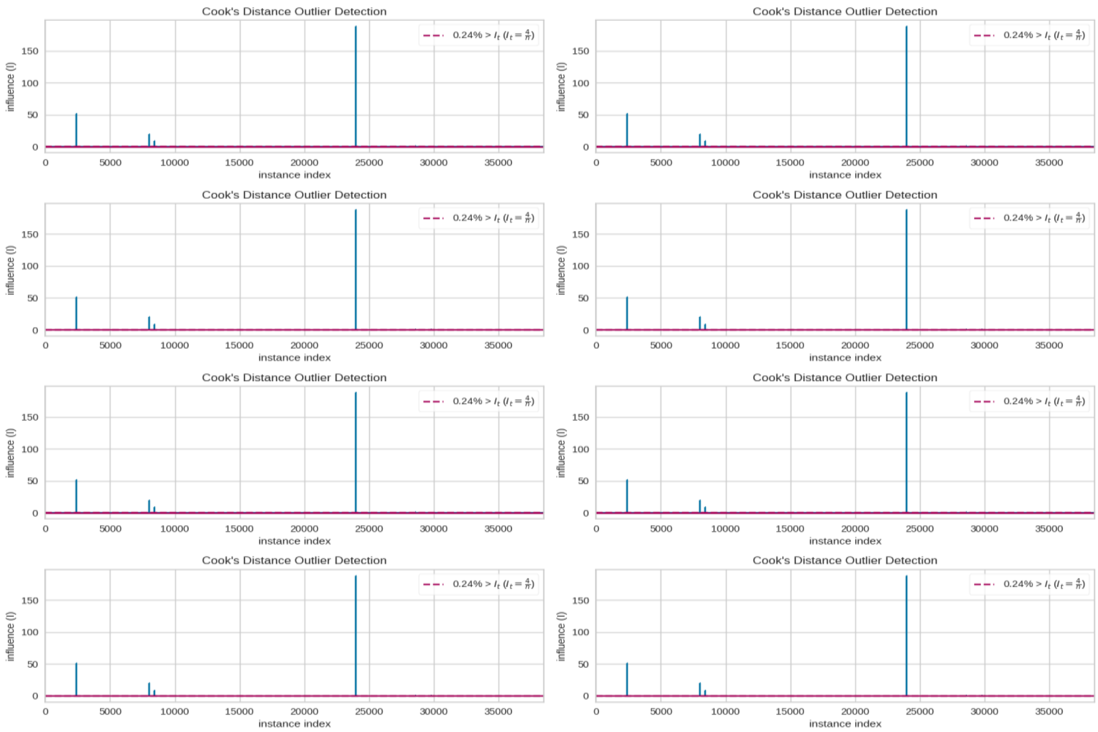

Figure 4.3.7: Cook's Distance plots for outlier detection in different regression models

The Cook's Distance plots identify influential abnormalities in regression analyses. While most data points have little impact, a small group shows substantial influence, indicating potential outliers. This consistent trend of influential points in different models highlights the importance of specific data points in determining model accuracy.

4.4 Feature handle:

Accurately controlling variables is crucial for forecasting inflation, as it ensures model precision, captures vital economic trends, and boosts forecast reliability, ultimately aiding policymakers and economists in making well-informed decisions.

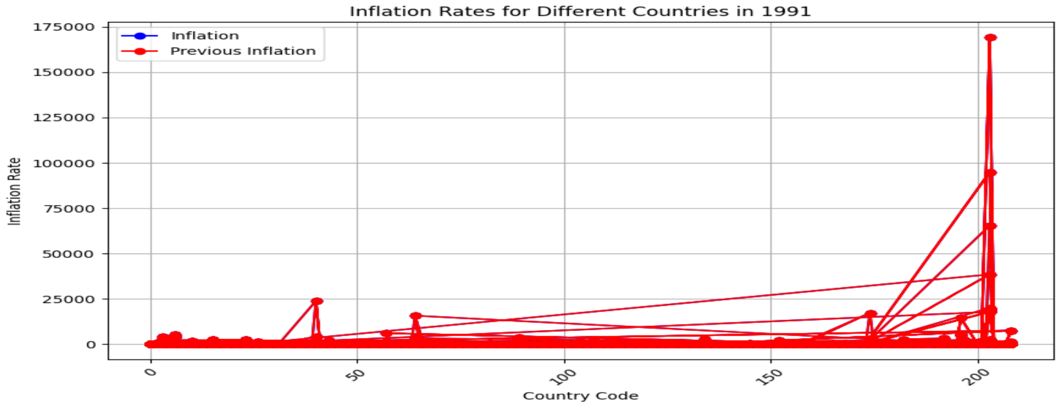

Figure 4.4.1: A visual depiction depicting the rise in prices and past inflation rates for different countries in the year 1991

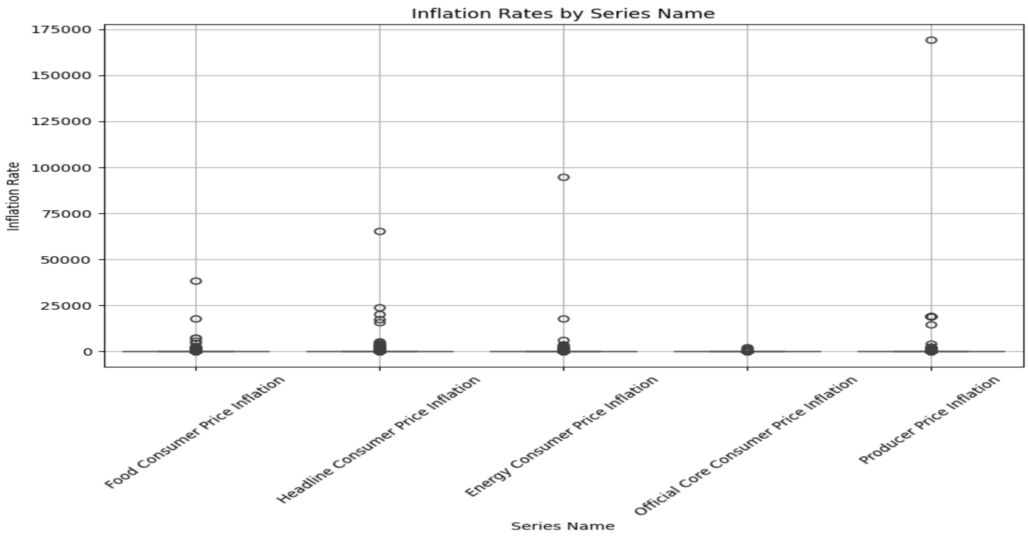

Figure 4.4.2: Inflation rate by series Name

Here, the distribution of inflation rates for several series categories, such as producer price inflation, food consumer price inflation, headline consumer price inflation, energy consumer price inflation, and official core consumer price inflation, are displayed in a box plot. A few extreme outliers with noticeably greater inflation rates are scattered throughout the data, with the bulk of the data points being closer to the lower end of the range. As a result, even while most inflation rates are generally moderate, there are sporadic cases of extremely high inflation rates in every category. At first shows that Producer Price Inflation, then Headline Consumer Price Inflation over time Food Consumer Price Inflation, Energy Consumer Price Inflation, Official Core Consumer Price Inflation. Here shows that Which feature has caused inflation to be high, and will have an impact on future inflation. That represent Explainable Inflation Forecast, where feature will detect that Which factor has contributed to the high rate of inflation and will influence inflation going forward.

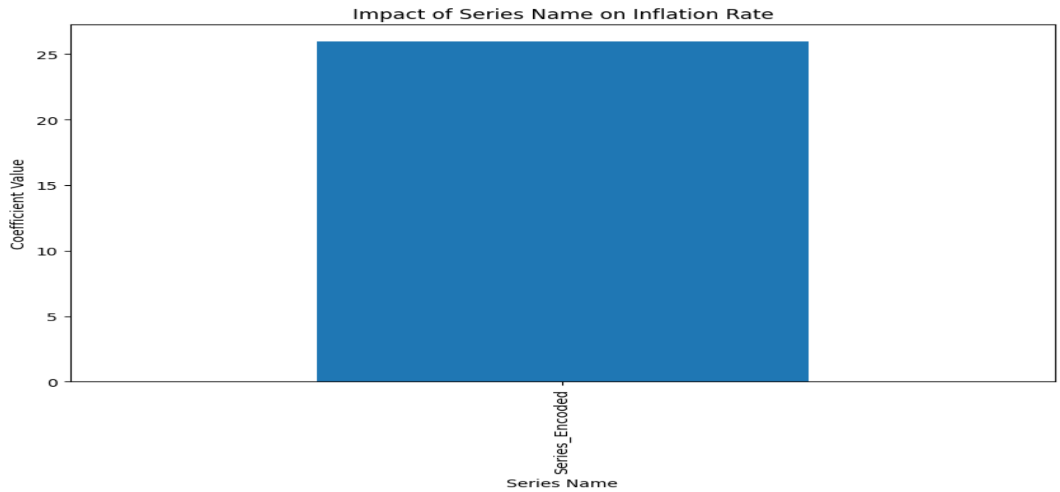

Figure 4.4.3: Impact of Series Name on Inflation Rate

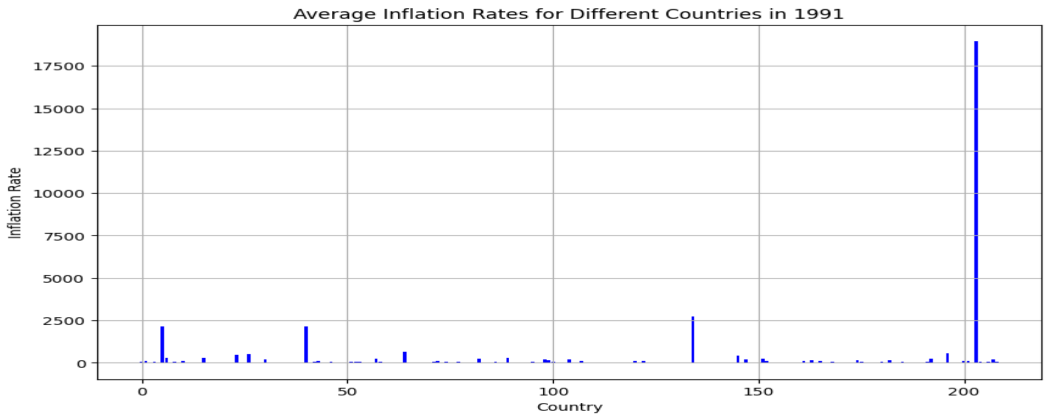

Figure 4.4.4: Average inflation Rates For different Counties 1991

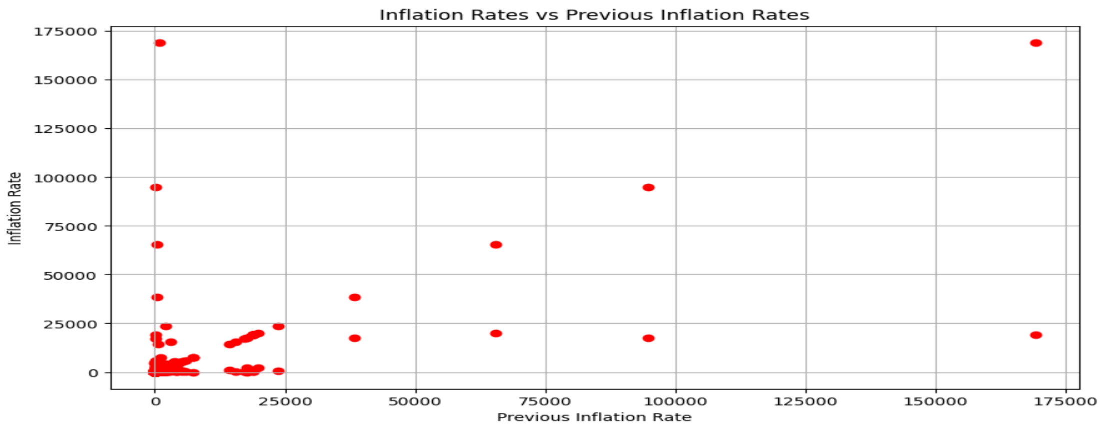

Figure 4.4.5: Inflation vs Previous Inflation Rate

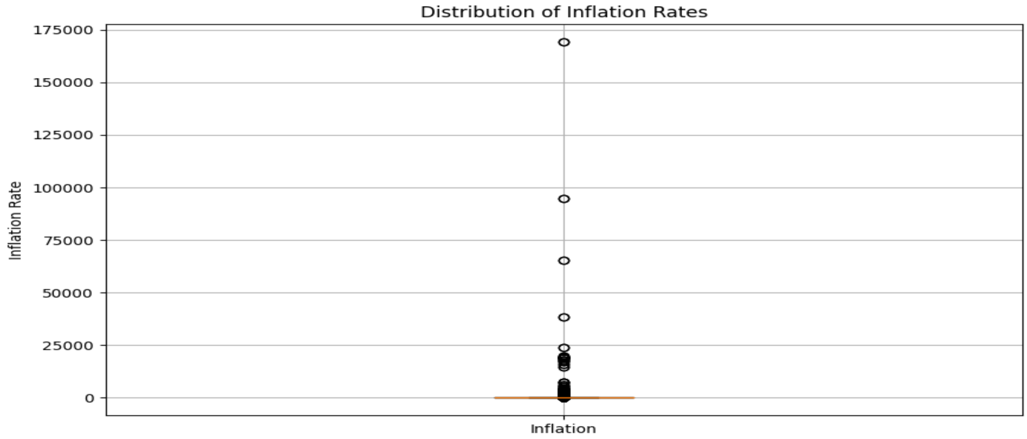

Figure 4.4.6: Distribution of inflation Rate

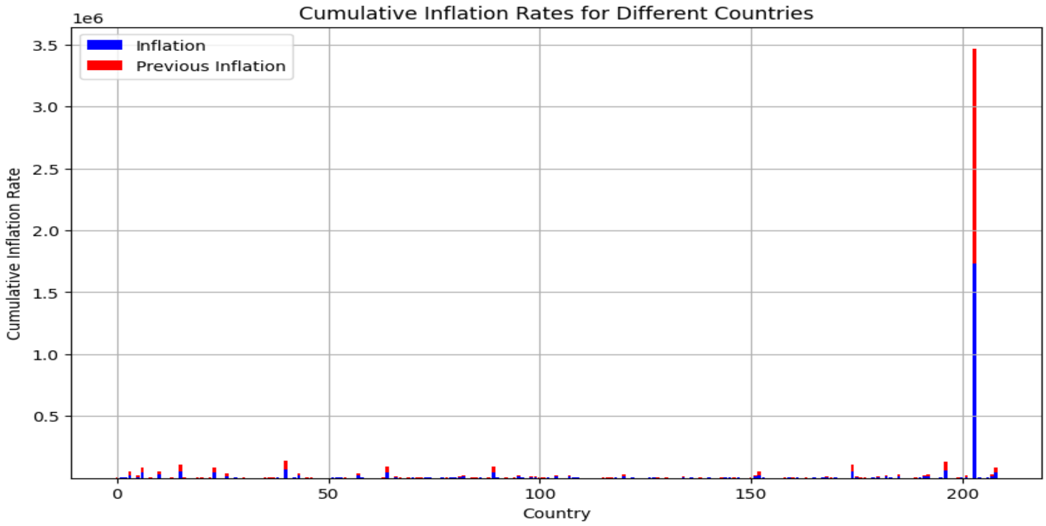

Figure 4.4.7: Cumulative Inflation Rates for Different Countries

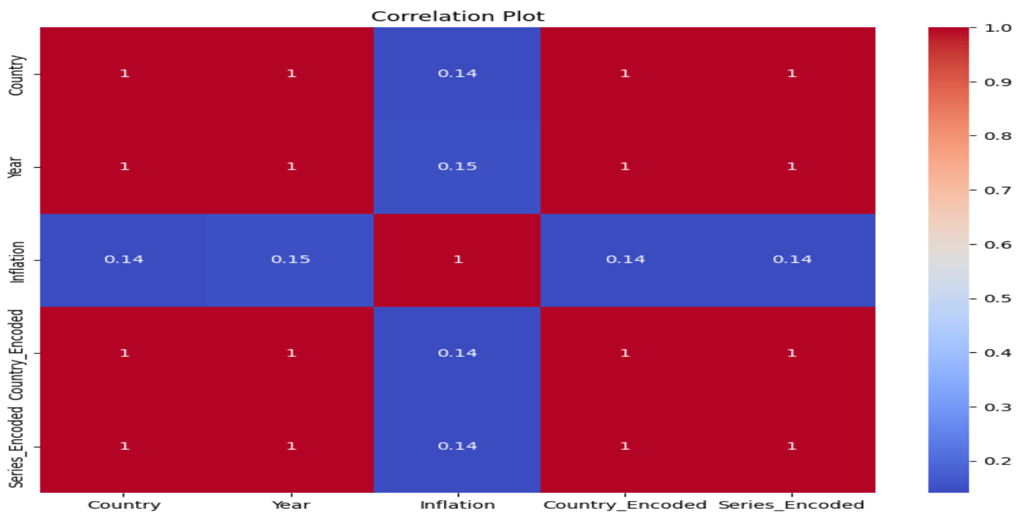

Figure 4.4.8: Correlation Matrix of Different Variables in Dataset

The correlation plot illustrates strong correlations between "Country," "Year," and their encoded counterparts, pointing towards their probable categorical nature. On the other hand, "Inflation" displays minimal correlations with the remaining variables, indicating limited linear associations within the dataset. In the end, correlation matrices can aid in the diagnosis of model problems, assessment procedures, and feature selection, resulting in a more comprehensible and trustworthy model. In the end, correlation matrices can aid in the diagnosis of model problems, assessment procedures, and feature selection, resulting in a more comprehensible and trustworthy model.

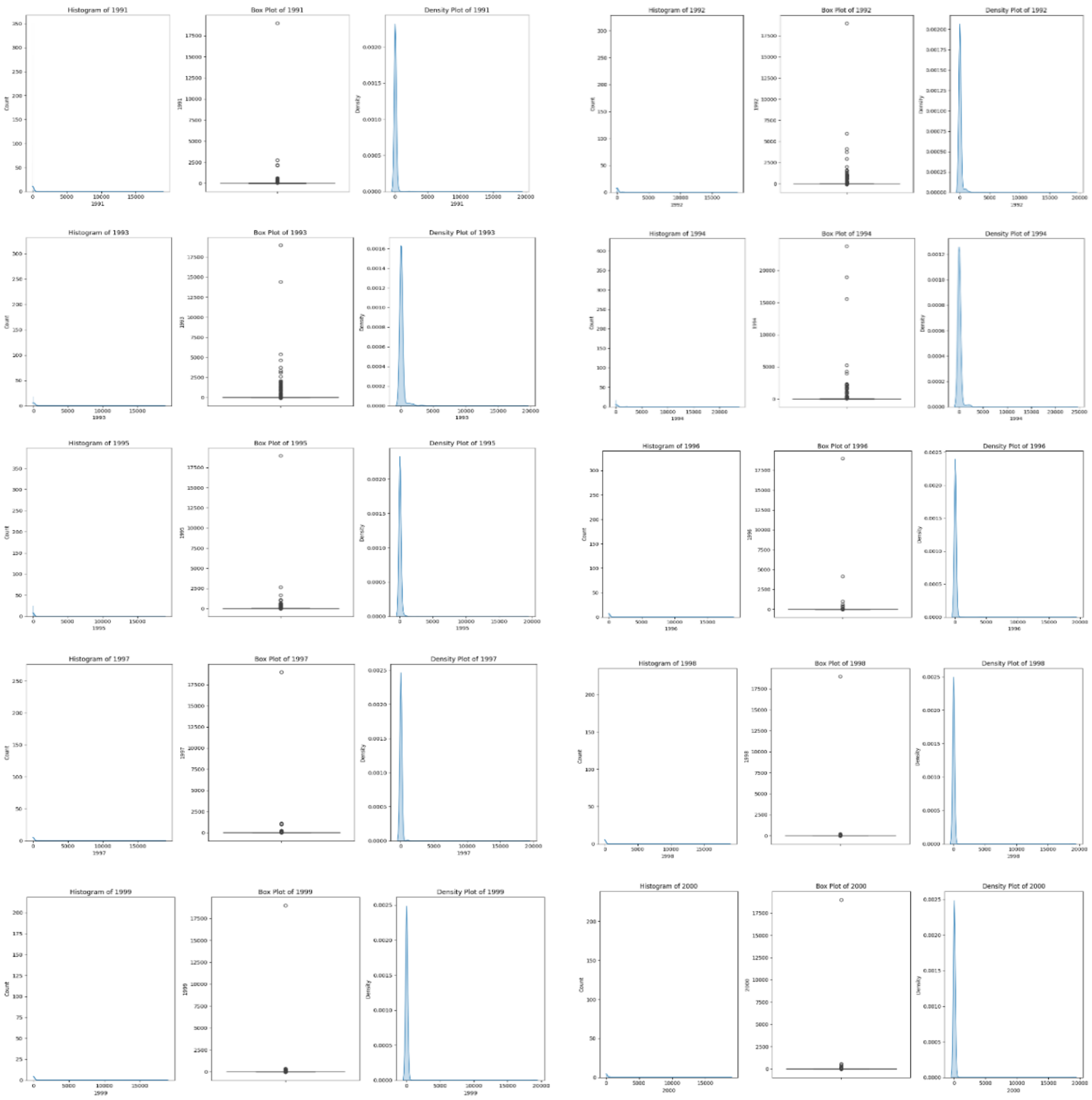

Figure 4.4.9: Plot distribution for each numerical column (1991 to 2000)

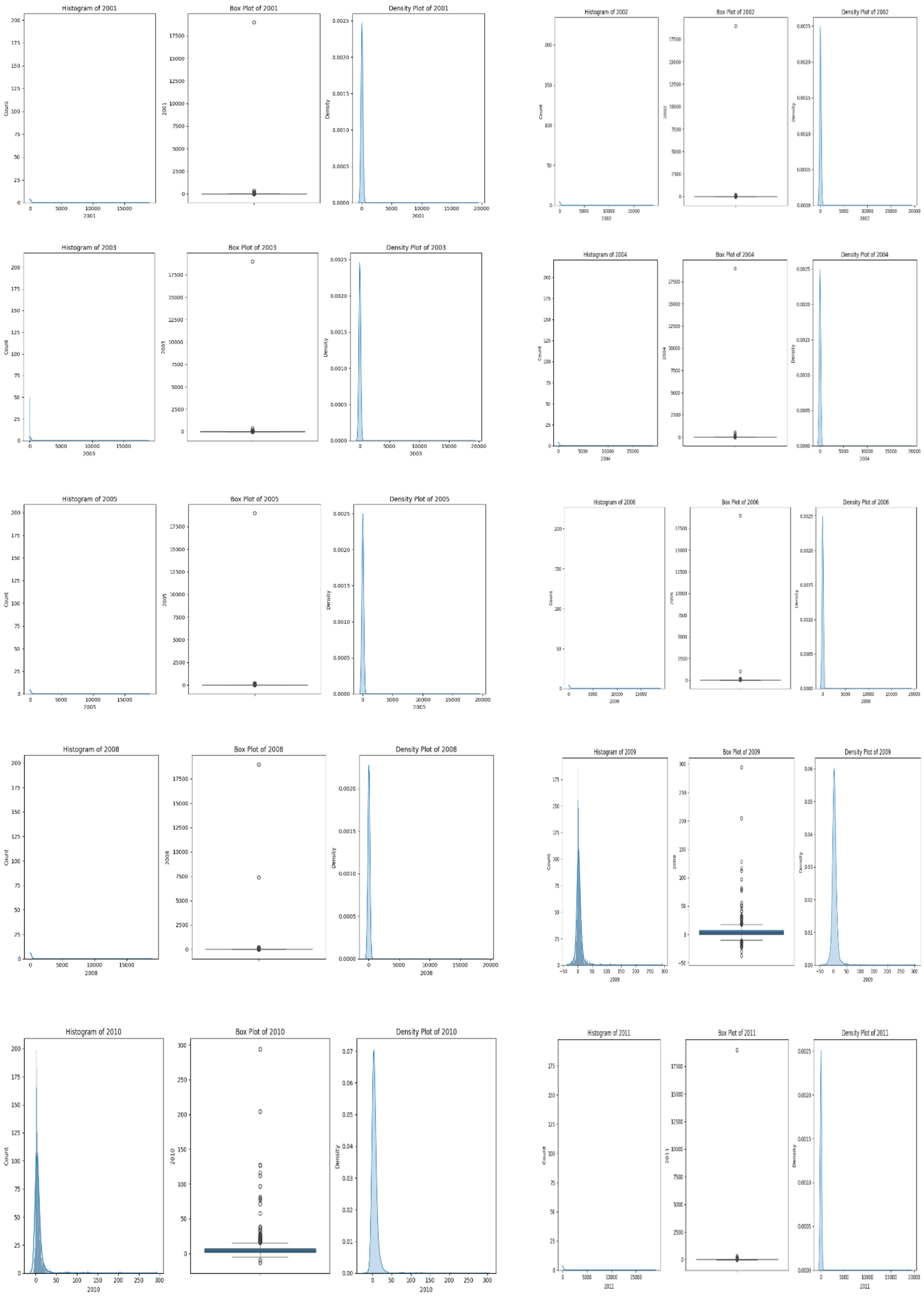

Figure 4.4.10: Plot distribution for each numerical column (2001 to 2011)

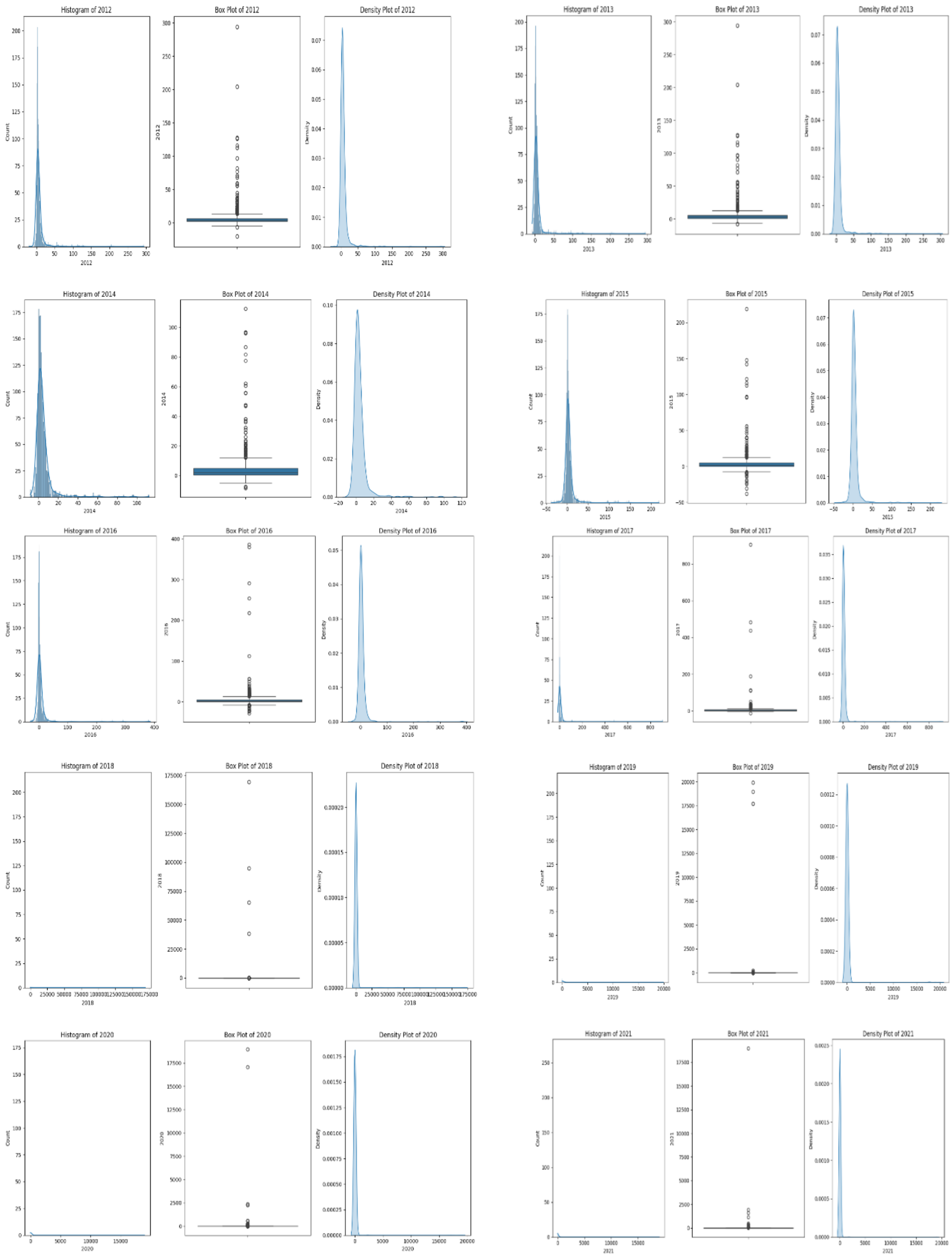

Figure 4.4.11: Plot distribution for each numerical column (2012 to 2021)

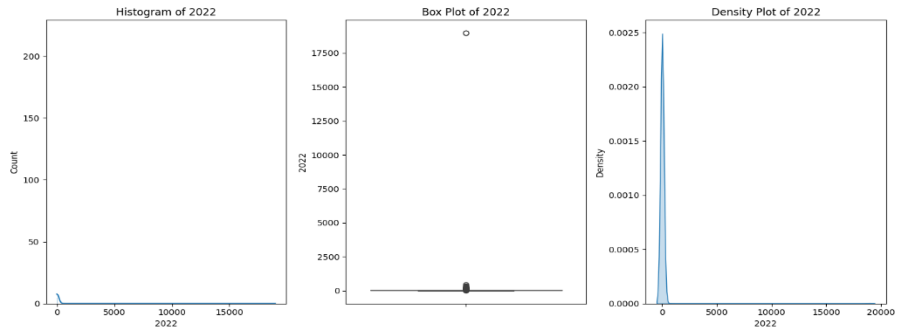

Figure 4.4.12: Plot distribution for each numerical column (2022)

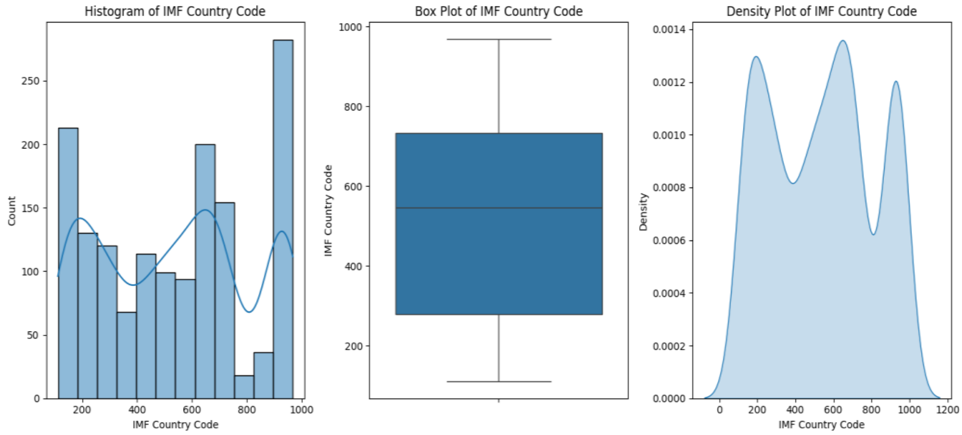

Figure 4.4.13: Plot distribution for each numerical column with country code

4.5 Multi-collinearity and non-stationarity handle:

In order to manage multi-collinearity, we compute the variance inflation factor (VIF) for every feature and eliminate those with a high VIF. We utilize the Augmented Dickey–Fuller test (ADF test) and differencing to check for non-stationarity in our time series data and make sure it is stationer. Because of this, the Augmented Dickey-Fuller (ADF) test and the Variance Inflation Factor (VIF) test help to improve a model's interpretability by addressing non-stationarity and multicollinearity, respectively.

4.5.1 Variance Inflation Factor (VIF):

Variables/ Feature	VIF Result
Year	5.382093
Previous Inflation	1.003468

Country Encoded	3.855376
Series Encoded	2.724489

Table 4.5.1: Summarizing the Variance Inflation Factor (VIF) for each variable

Moderate multicollinearity is indicated by the year's VIF of 5.382093. This implies a certain degree of correlation between the 'Year' variable and the other variables in the model. It's not very troublesome, though, as it falls below the crucial level of 10. With a relatively low VIF of 1.003468 for previous inflation, there is very little multicollinearity. 'Previous Inflation' is thus essentially independent of all other model predictors. The low to moderate multicollinearity of the country encoded is shown by its VIF of 3.855376. It appears from this that there is some association, however not enough to be concerned about, between the 'Country Encoded' variable and other predictors. The Series Encoded VIF of 2.724489 suggests a low to moderate level of multicollinearity. Although there is some link between "Series Encoded" and other predictors, it is not very strong.

It appears that multicollinearity is not a major problem in our present model because none of the variables have VIF values more than 10. According to the present VIF values, the coefficient estimates of the model appear to be stable and trustworthy. In light of this, the model's interpretability and the accuracy of the coefficient estimations are improved.

4.5.2 Augmented Dickey–Fuller test (ADF test):

ADF Statistic	p-value	Critical Values (1%)	Critical Values (5%)	Critical Values (10%)
-6.8459	1.742600018857429e-09	-3.5628	-2.9189	-2.5973

Table 4.5.2: Summarizing the results of the Augmented Dickey-Fuller (ADF) test for the original series

We reject the null hypothesis of a unit root since the ADF test statistic is much smaller than the crucial levels and the p-value is very small. The time series data is therefore said to be stationary. Our time series data appears to be steady, as indicated by the results of the ADF test. This suggests that the series is appropriate for time series modeling and forecasting since its mean and variance are expected to remain consistent throughout time. Any patterns or associations found by our models are more likely to be consistent and dependable over time because of the series' stationarity.

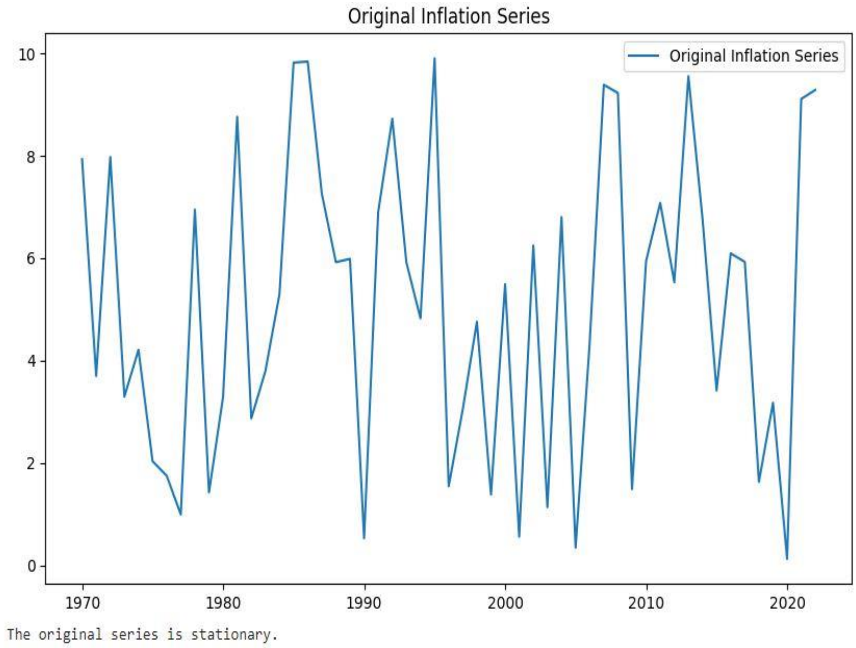

Figure 4.5.1: The original inflation series stationary

The "Original Inflation Series" plot provides visual evidence for the stationary nature of the series, which is the result of the ADF test. It may be used for additional time series analysis and modeling since it indicates that the inflation rates vary across the measured period around a consistent mean and variance.

4.5.3 Implications of Multicollinearity and Non-Stationarity for Model Forecast Reliability:

Consequences for Model Forecast Reliability for accurate model forecasts, multicollinearity and non-stationarity must be taken into account. The variation of coefficient estimates is inflated by multicollinearity, which makes it challenging to assess each predictor's individual contribution. We guarantee more stable and comprehensible coefficients by controlling multicollinearity with the Variance Inflation Factor (VIF). The time series data's statistical characteristics are guaranteed to remain constant across time when non-stationarity is detected and rectified using the Augmented Dickey-Fuller (ADF) test. Because the associations found in stationary data are more likely to be true in subsequent periods, it improves the predicted accuracy and dependability of the model. These actions enhance the validity and robustness of the model together.

4.6 Discussion:

This research assessed the effectiveness of eight machine learning algorithms in predicting inflation using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) as evaluation criteria. Among these algorithms, Gradient Boosting emerged as the top performer, displaying the lowest MSE and RMSE, as well as the highest R^2, indicating superior predictive precision and consistency. In the feature importance analysis conducted on Gradient Boosting, it was found that headline consumer price inflation, energy price inflation, and food price inflation were the most influential predictors. This observation underscores the significant role of specific components of inflation in shaping overall inflation trends. In contrast, simpler models such as Decision Tree and Linear Regression did not perform as well, likely due to their inability to capture intricate relationships within the data. Ensemble methods, notably Gradient Boosting and XGBoost, surpassed other models by effectively minimizing prediction errors through their robust learning mechanisms. Variance Inflation Factor (VIF), Partial Dependence, Cook's Distance, Correlation Matrix, Augmented Dickey-Fuller (ADF), and Residual Distribution offer valuable insights into the interrelationships among variables, model estimates, and robustness. These insights aid in the precise forecasting of inflation and enhance the interpretability of the model. In conclusion, these results emphasize the significance of utilizing advanced ensemble techniques in economic forecasting to achieve more precise and reliable outcomes.

Chapter 5: CONCLUSION AND RECOMMENDATION

5.1 Conclusion:

In the prediction of inflation rates, Gradient Boosting surpassed other models in performance. Collective techniques like Gradient Boosting and XGBoost demonstrated superior efficiency when compared to individual models like Decision Tree and Linear Regression. Within the collective methods, Gradient Boosting emerged as the standout due to its exceptional accuracy in prediction, solidifying its reputation as the most reliable model for forecasting inflation in this analysis. A comprehensive evaluation of metrics such as MSE, RMSE, and R^2 conclusively determined Gradient Boosting as the best-performing model, showcasing minimal errors in prediction and maximum explanatory power. But we can choose XGBoost as an alternative of the Gradient Boosting model. Models using Decision Trees and Linear Regression have the most restrictions. Because it relies on linearity, which may not capture complicated connections, linear regression is extremely sensitive to outliers. Our research highlights the paramount significance of unique indicators for inflation - such as Food Inflation, Energy Inflation, Headline Consumer Inflation, Producer Inflation, and Official Core Consumer Inflation - in accurately forecasting overall inflation trends. Variables provide essential insights, facilitating the development of more precise and explainable predictive models. As inflation within these particular sectors rises, it signals impending economic shifts, necessitating customized approaches to mitigate inflationary pressures and align supply with demand. This proactive capability is essential for policymakers, businesses, and individuals, fostering heightened awareness and preemptive measures. By grasping the intricacies of these various elements contributing to inflation, stakeholders can more effectively predict economic changes and devise methods to maintain stability. This study not only adds to the knowledge of predicting economic trends but also emphasizes the significance of openness and comprehensibility in inflation patterns. Educating the public about these factors gives people and communities the knowledge to make informed decisions and build resilience against economic ups and downs. In the end, our research emphasizes the importance of incorporating a range of inflation indicators into forecasting models to enhance the ability to manage inflation and its effects on society.

5.2 Implication for Future Research:

The research results, utilizing a wide range of machine learning models (Linear Regression, Random Forest, Decision Tree, AdaBoost, XGBoost, Ridge, Lasso, and Gradient Boosting), provide several avenues for further investigation. Adding more sophisticated ensemble techniques or hybrid models might improve prediction accuracy and robustness, especially when it comes to resolving the shortcomings of individual models. Incorporating neural network topologies, such Long Short-Term Memory (LSTM) networks or Recurrent neural networks (RNNs), might further enhance forecasting ability by identifying intricate temporal patterns in inflation data. Further studies should examine the effects of adding a wider range of exogenous variables and macroeconomic indicators, such as world commodity prices, monetary policy indicators, and geopolitical events, in order to enhance the model's inputs and offer a more comprehensive understanding of inflation dynamics. Adding more precise, high-frequency data to the dataset may also enhance model responsiveness and provide insights into short-term inflationary patterns. Finally, improving the transparency and reliability of machine learning-based inflation projections should help policymakers make well-informed decisions. This could be achieved by investigating the interpretability of complicated models and creating explainable AI frameworks.

References:

[1] S. I. Momo, M. Riajuliislam, and R. Hafiz, "Forecasting of inflation rate contingent on Consumer Price Index: Machine Learning approach," in *Lecture notes in networks and systems*, 2021, pp. 137–144. doi: 10.1007/978-981-16-3153-5_17.

[2] G. Boaretto and M. C. Medeiros, "Forecasting inflation using disaggregates and machine learning," *arXiv.org*, Aug. 22, 2023. https://arxiv.org/abs/2308.11173.

[3] C. Yang and S. Guo, "Inflation Prediction Method based on Deep learning," *Computational Intelligence and Neuroscience*, vol. 2021, pp. 1–6, Aug. 2021, doi: 10.1155/2021/1071145.

[4] S. Aras and P. J. G. Lisboa, "Explainable inflation forecasts by machine learning models," *Expert Systems with Applications*, vol. 207, p. 117982, Nov. 2022, doi: 10.1016/j.eswa.2022.117982.

[5] A. F. Theoharidis, D. A. Guillén, and H. Lopes, "Deep learning models for inflation forecasting," *Applied Stochastic Models in Business and Industry*, vol. 39, no. 3, pp. 447–470, Apr. 2023, doi: 10.1002/asmb.2757.

[6] F. Dharma, S. Shabrina, A. Noviana, M. Tahir, N. Hendrastuty, and W. Wahyono, "Prediction of Indonesian Inflation Rate Using Regression Model Based on Genetic Algorithms," *5. Prediction of Indonesian Inflation Rate Using Regression Model Based on Genetic Algorithms*, vol. 5, no. 1, pp. 45–52, Jul. 2020, doi: 10.15575/join.v5i1.532.

[7] M. A. Choudhary and A. Haider, "Neural network models for inflation forecasting: an appraisal," *Applied Economics*, vol. 44, no. 20, pp. 2631–2635, Jun. 2011, doi: 10.1080/00036846.2011.566190.

[8] G. S. M. Thakur, R. Bhattacharyya, and S. S. Mondal, “Artificial neural network-based model for forecasting of inflation in India,” *Fuzzy Information and Engineering*, vol. 8, no. 1, pp. 87–100, Mar. 2016, doi: 10.1016/j.fiae.2016.03.005.

[9] J. J. J. Groen, R. Paap, and F. Ravazzolo, “Real-Time inflation forecasting in a changing world,” *Journal of Business & Economic Statistics*, vol. 31, no. 1, pp. 29–44, Jan. 2013, doi: 10.1080/07350015.2012.727718.

[10] Y. Hong, F. Jiang, L. Meng, and B. Xue, “Forecasting Inflation with Economic Narratives and Machine Learning,” *Social Science Research Network*, Jan. 2022, doi: 10.2139/ssrn.4175749.

[11] A. Rodríguez-Vargas, “Forecasting Costa Rican inflation with machine learning methods,” *Latin American Journal of Central Banking*, vol. 1, no. 1–4, p. 100012, Jan. 2020, doi: 10.1016/j.latcb.2020.100012.

[12] A. Almosova and N. Andresen, “Nonlinear inflation forecasting with recurrent neural networks,” *Journal of Forecasting*, vol. 42, no. 2, pp. 240–259, Sep. 2022, doi: 10.1002/for.2901.

[13] Ö. Özgür and U. Akkoç, “Inflation forecasting in an emerging economy: selecting variables with machine learning algorithms,” *International Journal of Emerging Markets*, vol. 17, no. 8, pp. 1889–1908, Feb. 2021, doi: 10.1108/ijoem-05-2020-0577.

[14] A. Joseph, G. Potjagailo, C. Chakraborty, and G. Kapetanios, “Forecasting UK inflation bottom up,” *International Journal of Forecasting*, Feb. 2024, doi: 10.1016/j.ijforecast.2024.01.001.

[15] J. Ha, M. A. Kose, F. Ohnsorge, and H. Yilmazkuday, what explains global inflation. 2023. doi: 10.1596/1813-9450-10648.

[16] J. Ha, M. A. Kose, and F. Ohnsorge, One-Stop Source: A global database of inflation. 2021. doi: 10.1596/1813-9450-9737.

202-35-3125

ORIGINALITY REPORT

16%	12%	10%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
2	www.mdpi.com Internet Source	1%
3	Submitted to Higher Education Commission Pakistan Student Paper	<1%
4	ijcs.stmikindonesia.ac.id Internet Source	<1%
5	Submitted to Daffodil International University Student Paper	<1%
6	Submitted to Taylor's Education Group Student Paper	<1%
7	Leon, Javier. "Forecasting Imported Fruit Prices in the United States Using Neural Networks", Northcentral University, 2024 Publication	<1%
8	Nawazish Mirza, Syed Kumail Abbas Rizvi, Bushra Naqvi, Muhammad Umar. "Inflation	<1%