

**Forecasting GDP and inflation affects**

**Research Question:**Inflation affects GDP growth, as rising inflation can slow down economic growth by reducing purchasing power and increasing production costs. However, moderate inflation is often seen as a sign of a growing economy. **How can we accurately predict whether a country will experience inflation in the coming year?**

** Executive Summary**

**Rational:**

Gross Domestic Product (GDP) and Consumer Price Index (CPI) inflation are critical indicators of the economic health and stability of the United States. GDP measures the total value of goods and services produced, representing overall economic growth. In contrast, the CPI inflation rate reflects changes in consumer prices over time, indicating shifts in cost of living and purchasing power.

Historically, the U.S. economy has experienced robust growth, with occasional recessions—such as the 2008 financial crisis and the COVID-19 pandemic—that temporarily disrupted GDP growth. Nonetheless, recoveries followed these downturns, showcasing the economy’s resilience. Conversely, inflation has fluctuated, with high inflation in the 1970s and 1980s due to oil crises, followed by more stable rates in subsequent decades. Recently, the country has seen a resurgence of inflation due to supply chain disruptions, heightened demand, and pandemic-related factors, leading to inflation levels unseen in decades.

Given the impact of inflation on purchasing power and production costs, accurately forecasting inflation is essential for policymaking and business planning. Understanding the interplay between GDP and inflation can help maintain economic stability and prosperity.

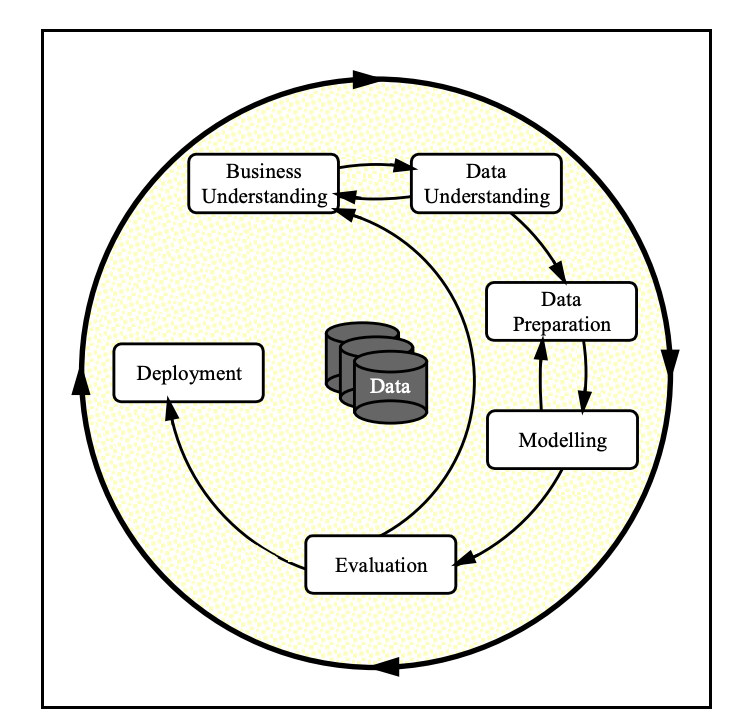
**Importance of Forecasting**:

* Governments need accurate inflation forecasts to adjust monetary policies and mitigate adverse effects.
* Businesses must incorporate GDP and inflation projections into budgeting, pricing, and wage negotiations, as unanticipated inflation can affect profitability.

**CRISP-DM Framework for Forecasting**

We follow the **CRISP-DM (Cross-Industry Standard Process for Data Mining)** framework, a widely used methodology, to predict inflation. The phases are:

1. **Business Understanding**: Forecast inflation rates for different countries and identify key features influencing inflation.
2. **Data Understanding**: Analyze World Bank datasets on inflation.
3. **Data Preparation**: Clean and preprocess the data for analysis.
4. **Modeling**: Build and test models for forecasting.
5. **Evaluation**: Assess model performance.
6. **Deployment**: Implement the model for future forecasts.



** Business Understanding**

**Objective**:

1. Review the inflation rates for different countries from Year 1970 till today.
2. Identify the key features influencing GDP in US. Build a model to forecast GDP rates in US and future trends.

** Data Understanding**

The dataset was sourced from the World Bank and Federal Reserve Economic Data. It covers inflation data from 1970 to 2023 for 186 countries and includes various sheets for different categories of inflation-related data.

**Dataset URLs**:

<https://www.worldbank.org/en/research/brief/inflation-database>

<https://fred.stlouisfed.org/series/GDP>

**Data Description:**

The dataset contains multiple sheets representing different categories of inflation-related data. Here are the sheet names:

1. **Intro** (introductory or metadata information)
2. **top** (top-level inflation data)
3. **hcpi\_m**, **hcpi\_q**, **hcpi\_a** (harmonized consumer price index in monthly, quarterly, and annual formats)
4. **ecpi\_m**, **ecpi\_q**, **ecpi\_a** (emerging consumer price index in monthly, quarterly, and annual formats)
5. **fcpi\_m**, **fcpi\_q**, **fcpi\_a** (forecasted consumer price index in monthly, quarterly, and annual formats)
6. **ccpi\_m**, **ccpi\_q**, **ccpi\_a** (core consumer price index in monthly, quarterly, and annual formats)
7. **ppi\_m**, **ppi\_q**, **ppi\_a** (producer price index in monthly, quarterly, and annual formats)
8. **def\_q**, **def\_a** (inflation deflator data in quarterly and annual formats)
9. **Aggregate** (summary or aggregate data)

**Relevant Sheets for Analysis**:

* **hcpi\_a**: Harmonized Consumer Price Index (annual)
* **ecpi\_a**: Emerging Consumer Price Index (annual)
* **fcpi\_a**: Forecasted Consumer Price Index (annual)
* **ccpi\_a:** Core Consumer Price Index (annual)
* **ppi\_a:** Producer Price Index (annual)
* **def\_a**: Inflation deflator (annual)

**Key Columns**:

* **Country**: Country name.
* **Country Code**: Country abbreviation.
* **Inflation Data**: Monthly CPI data from 1970 to 2023.

** Part 1 Data Preparation (inflation1.ipynb)**

Data preparation is crucial for accurate model predictions. The following steps were implemented:

**Data Cleaning:**

* + Removed irrelevant columns.
  + Handled missing values and outliers.
  + Normalized or scaled numerical features, such as GDP and inflation rates.

**Data Transformation**:

* + Renamed columns for clarity.

'Headline Consumer Price Index': 'HCPI',

'Energy Price Index': 'ECPI',

'Food Price Index': 'FCPI',

'Official Core Consumer Price Index': 'OCPI',

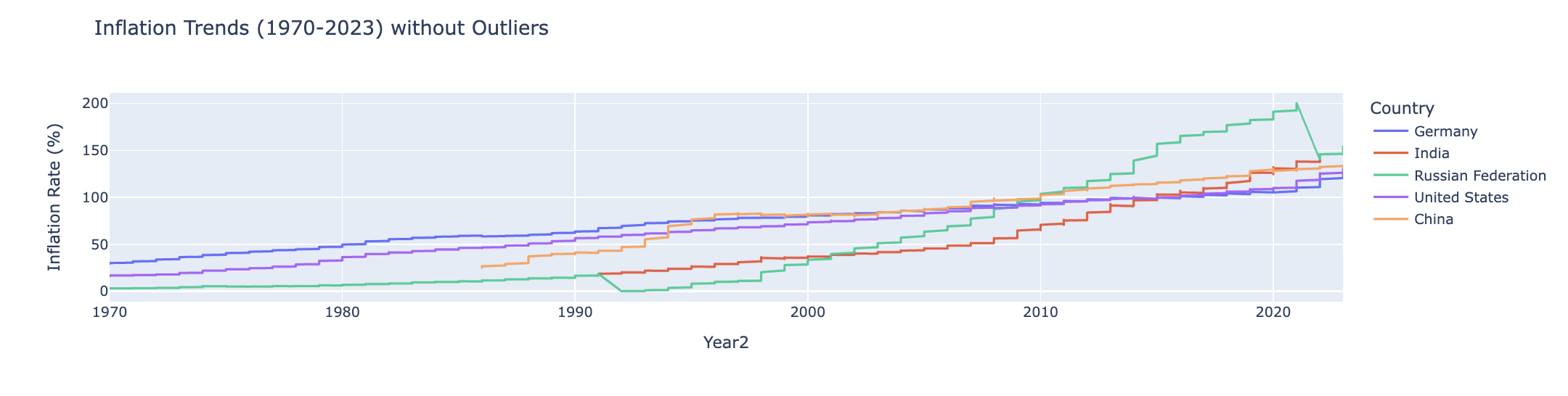
'Producer Price Index': 'PPI'

* + Checked for and removed duplicates.

** Exploratory Data Analysis (EDA)**

In the Data Exploration phase, conducted a thorough analysis to uncover patterns, relationships, and insights within the dataset.

**Visualize inflation trends** over time for 5 selected countries: United States, Germany, Russian Federation, China and India.



**Key findings from data exploration include:**

1. **Overall Trend**:
   * All countries show an upward trend in inflation over time, indicating that inflation rates have generally increased across the observed period from 1970 to 2023.
2. **Country-Specific Observations**:
   * **China (Orange)**: Shows a steady increase in inflation over the years but remains relatively stable compared to others.
   * **Germany (Blue)**: Has one of the lowest and most stable inflation rates over the years, indicating better control over inflation.
   * **India (Red)**: Has a gradual increase in inflation over time, staying within a moderate range compared to other countries.
   * **Russian Federation (Green)**: Displays sharp fluctuations, especially around the late 1980s and early 1990s, which might correspond to major economic changes like the dissolution of the Soviet Union. Afterward, inflation stabilized but still remains relatively high.
   * **United States (Purple)**: Displays a relatively smooth and steady increase in inflation over the years, showing consistent growth without sharp fluctuations.
3. **Significant Events**:
   * The sharp dips and spikes in the Russian Federation's data around the late 1980s to early 1990s suggest significant economic disruptions, possibly due to political changes and economic reforms.
   * Other countries like Germany and the United States exhibit steadier trends, indicating more stable economies over the years.
4. **Inflation Control**:
   * Among these countries, Germany appears to have the most effective control over inflation, as evidenced by its stable and lower inflation trend.
   * China and India show consistent growth in inflation, suggesting rising prices but with fewer extreme variations compared to the Russian Federation.

**Conclusion**

The inflation trends indicate that while inflation has generally risen across all these countries from 1970 to 2023, some countries (e.g., Germany and the United States) have experienced more stable increases, while others like the Russian Federation have had more volatility due to economic and political changes.

## **Part 2- US GDP Trend(inflation2.ipynb)**

**Explore US Data:**

A group of blue bars

Description automatically generated

A group of green and white boxes

Description automatically generated

A screenshot of a graph

Description automatically generated

A graph of a graph of a graph

Description automatically generated with medium confidence

**Observations and Conclusions:**

**Histograms:**

* Most of the variables are skewed right, indicating that their values tend to be lower overall with some higher outliers.
* The distribution patterns suggest that data normalization or transformation might be needed before using these variables in a machine learning model.

**Boxplots :**

* There are no extreme outliers shown in the boxplots, but there is a noticeable range of variability, with GDP having the widest range.
* Most variables exhibit moderate to high dispersion, indicating variability in the dataset.

**Correlation Matrix:**

* There is a very high positive correlation between several features, such as:
  + HCPI and FCPI (correlation = 1)
  + OCPI and FCPI (correlation ≈ 1)
  + PPI and HCPI (correlation ≈ 0.98)
* GDP also shows strong correlations with other variables like HCPI (0.97), suggesting that many of these variables could be good predictors of GDP but may suffer from multicollinearity.

**Pair Plot :**

* There is a strong linear relationship between many of the variables (e.g., HCPI vs. FCPI, OCPI vs. GDP), suggesting that these features might be highly correlated.
* Some of the points are tightly clustered along linear paths, indicating multicollinearity (high correlation) between features.
* The different colors representing the GDP Trend (0 and 1) show overlapping distributions, suggesting that the features might not fully distinguish between the two trends.

**Key Conclusions:**

1. **Multicollinearity Issue**: The high correlations among several features (e.g., HCPI, FCPI, OCPI) suggest that multicollinearity is present, which can impact the performance and interpretability of models like linear regression. Dimensionality reduction techniques such as PCA or feature selection could be useful.
2. **Feature Importance**: Given the strong correlation between GDP and features like HCPI, FCPI, and PPI, these variables are likely to be important predictors in a GDP-related model.
3. **Data Transformation**: Since many of the features show skewness, data normalization or transformation might be necessary to improve model performance.

**Data preparation and transformation**



[This Photo](https://commons.wikimedia.org/wiki/File:Fire.JPG) by Unknown Author is licensed under [CC BY-SA](https://creativecommons.org/licenses/by-sa/3.0/)

* Define GPD Trend as Boolean as Targe variable.

# Split data into features and target

X = df2.drop('GDP Trend', axis=1)

y = df2['GDP Trend']

* Dataset is imbalanced:

GDP Trend

0 443

1 207

Name: count, dtype: int64

* Handling class imbalance with SMOTE
* Feature scaling

scaler = StandardScaler()

X\_train = scaler.fit\_transform(X\_train)

X\_test = scaler.transform(X\_test)

** Model Training and Evaluation**

Following models are applied to the US data:

* 'Logistic Regression'
* 'Decision Tree'
* 'Random Forest'
* 'SVM'
* 'Gradient Boosting'
* 'k-NN'

**Performance Evaluation**:  
Performance was measured using precision, recall, F1-score, and accuracy.

**No Model Stands Out:** None of the models provide a satisfactory classification performance. The highest accuracy is only 52% (Logistic Regression), which is barely better than random guessing.

**Possible Data Issues**:

* The low performance across all models suggests that the features might not be informative or discriminative enough for this classification task. There could be issues like:
* Poor feature selection or lack of relevant features.
* High noise or overlapping data between classes.

**Lack of diverse information:**

Features are highly positively correlated, it leads to multicollinearity, meaning that the model has difficulty discerning the individual impact of each feature on the target variable (The issue will be addressed in Project 2)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Classification Report** | **precision** | **recall** | **f1-score** | **support** |
| **Logistic Regression:** | 0 | 0.53 | 0.58 | 0.55 | 137 |
| 1 | 0.51 | 0.47 | 0.49 | 129 |
| macro avg | 0.52 | 0.52 | 0.52 | 266 |
| weighted avg | 0.52 | 0.52 | 0.52 | 266 |
| accuracy |  |  | **0.52** | 266 |
| **Decision Tree** | 0 | 0.51 | 0.45 | 0.48 | 137 |
| 1 | 0.48 | 0.55 | 0.51 | 129 |
| macro avg | 0.5 | 0.5 | 0.5 | 266 |
| weighted avg | 0.5 | 0.5 | 0.49 | 266 |
| accuracy |  |  | 0.5 | 266 |
| **Random Forest:** | 0 | 0.49 | 0.47 | 0.48 | 137 |
| 1 | 0.45 | 0.47 | 0.46 | 129 |
| macro avg | 0.47 | 0.47 | 0.47 | 266 |
| weighted avg | 0.47 | 0.47 | 0.47 | 266 |
| accuracy |  |  | 0.47 | 266 |
| **SVM** | 0 | 0.53 | 0.41 | 0.46 | 137 |
| 1 | 0.49 | 0.61 | 0.55 | 129 |
| macro avg | 0.51 | 0.51 | 0.5 | 266 |
| weighted avg | 0.51 | 0.51 | 0.5 | 266 |
| accuracy |  |  | 0.51 | 266 |
| **Gradient Boosting** | 0 | 0.48 | 0.46 | 0.47 | 137 |
| 1 | 0.46 | 0.48 | 0.47 | 129 |
| macro avg | 0.47 | 0.47 | 0.47 | 266 |
| weighted avg | 0.47 | 0.47 | 0.47 | 266 |
| accuracy |  |  | 0.47 | 266 |
| **KNN** | 0 | 0.43 | 0.34 | 0.38 | 137 |
| 1 | 0.43 | 0.52 | 0.47 | 129 |
| macro avg | 0.43 | 0.43 | 0.43 | 266 |
| weighted avg | 0.43 | 0.43 | 0.42 | 266 |
| accuracy |  |  | 0.43 | 266 |

A group of blue squares with white text

Description automatically generated

**ROC Curve and AUC Curve**

A graph of a curve

Description automatically generated with medium confidence

**Key Conclusions:**

* **Poor Model Performance:** All models exhibit weak discriminatory power, as evidenced by their AUC values, which hover around 0.5. This suggests that the models are not effectively capturing the underlying patterns in the data to make accurate predictions.
* **SVM Performs Slightly Better:** Although the SVM model has the highest AUC (0.55), this is still not a strong performanceand indicates only marginal improvement over random guessing.

**Model Accuracy Comparison**

A graph of a graph showing different colored bars

Description automatically generated with medium confidence

**Overall Analysis:**

All models exhibit low accuracy, indicating that they are not effectively distinguishing between the classes in the dataset.

Logistic Regression, despite being the best performer, is only marginally better than random guessing, which suggests that the features being used are not providing clear predictive power for the classification task.

**Recommendation:**

**Re-evaluate the Dataset**: Since all models are performing poorly, consider re-evaluating the features being used or gathering more informative data.

** Time Series Model**

An **ARIMA(5, 1, 0)** model was used to forecast GDP trends. Key insights include:

* **Significant Predictors**: Lagged values (AR terms) at ar.L1 and ar.L2 were significant in predicting GDP trends.
* **Model Fit**: While the model captures time dependencies well, residual analysis suggests non-normality, indicating room for improvement.

The model predicts modest GDP recovery and growth over the next few years, but it may not account for unprecedented future events.

**A screenshot of a computer

Description automatically generated**

**Key Conclusions:**

* **Significance of Parameters:** The model appears to rely mainly on the third lag (ar.L3), as other AR terms are not statistically significant.
* **Diagnostic Concerns:** The model's residuals are not normally distributed and exhibit heteroskedasticity, indicating potential issues with the model's fit. These issues could imply that the model is not capturing all the complexities of the data.
* **Forecast Reliability:** The predicted values indicate a steady increase in GDP, but the issues with residual normality and heteroskedasticity suggest that the forecast may not be entirely reliable.

**Recommendations:**

* **Model Refinement:** Consider refining the model by re-evaluating lag terms or trying different ARIMA configurations. You might also consider adding exogenous variables or using models that can handle non-stationarity or heteroskedasticity.
* **Further Analysis:** Investigate why the residuals are not normally distributed—there may be structural breaks, seasonal patterns, or external factors that are not accounted for in the model.

A graph of growth of gdp

Description automatically generated

The ARIMA model forecasts a modest recovery and gradual growth in GDP over the next few years but doesn’t predict a dramatic upward or downward trend.

Given that ARIMA models are based on historical patterns, the forecasted values reflect past fluctuations but may not account for unprecedented future events or changes (e.g., policy changes, global events).

A graph of different colored lines

Description automatically generated with medium confidence

* The long-term upward trend in GDP is the most dominant feature, and the seasonal component is relatively stable.
* The residual component highlights the impact of major economic events like the financial crisis and the pandemic, but the trend indicates that GDP continues to recover and grow.

**Overall Conclusion**

* The time series analysis confirms that GDP has experienced steady growth over the past decades, with predictable seasonal fluctuations. The major deviations from this trend are associated with significant global economic events (e.g., financial crisis, pandemic).
* The SARIMA model's forecast shows a continued upward trajectory for GDP, although at a more moderate rate, suggesting that while growth will continue, it may not reach the rapid increases observed in earlier periods.
* The seasonal component's stability implies that the model captures the regular cyclical patterns in the data well, but attention should be given to potential outliers or disruptions that could affect future GDP trends.

** Conclusion and Next Steps**

**Conclusion:**

* The analysis reveals that the **long-term upward trend in GDP** is the most dominant feature, indicating sustained economic growth over time. Despite periods of volatility, such as the 2008 financial crisis and the 2020 pandemic, the GDP trend demonstrates resilience, consistently returning to an upward trajectory.
* The **seasonal component is relatively stable**, suggesting that GDP experiences regular and predictable fluctuations throughout the year. These fluctuations, while present, do not significantly alter the overall trend, indicating that seasonality is a smaller but consistent factor in GDP variations.

**Next Steps (Will be implemented in Capstone Project#2)**

1. **Enhance Model Accuracy with Additional Data**:
   * To improve the robustness and accuracy of the GDP forecasting model, it’s crucial to **gather more comprehensive datasets**. Including **exogenous variables** such as interest rates, unemployment rates, inflation rates, trade balances, and other macroeconomic indicators could provide a more holistic view of the factors influencing GDP changes, especially during periods of economic uncertainty.
2. **Compare Forecasts with Alternative Models**:
   * To ensure the reliability of the GDP forecasts, it's essential to **compare the results with other forecasting models** .This comparison can help validate the accuracy and robustness of the SARIMA model's predictions and identify if any alternative methods perform better in capturing the underlying patterns.
3. **Optimize Model with Hyperparameter Tuning**:
   * To achieve better predictive performance, **fine-tune the model’s hyperparameters** (e.g., ARIMA order (p, d, q), seasonal parameters) using grid search or other optimization techniques. Proper hyperparameter tuning can significantly enhance the model’s ability to capture the complexities of the GDP time series data, leading to more accurate forecasts.
4. **Transition from Classification to Regression Analysis**:
   * The current analysis might involve a binary classification task with GDP Trend as a Boolean variable (growth/no growth). However, for a more nuanced understanding and prediction, **switching the target variable from a Boolean (GDP Trend) to a continuous regression target (actual GDP values)** can allow for more detailed and precise forecasts. This change will enable the model to predict the exact GDP value rather than just identifying whether it’s increasing or decreasing, providing more actionable insights for policymakers and analysts.