Analyzing the Impact of Socioeconomic Factors on GDP Using Machine Learning Models

1 Data Description

This exercise is based on data downloaded from Eurostat that cover most of the socioeconomic and environmental indicators. The aim is to develop a model mapping a relationship between GDP and these indicators. The summary of the data used for this preliminary analysis is as follows:

1. **GDP**: Gross Domestic Product at market prices, measured in chain-linked volumes (index 2010=100) to show changes in economic output.
2. **Inflation Rate**: Annual average change in consumer prices, measured by the Harmonized Index of Consumer Prices (HICP).
3. **Population Growth**: Annual total population change, comprising natural growth and migration.
4. **Unemployment Rate**: Proportion of the labor force (ages 15-74) actively looking for a job.
5. **Investment**: Investment in infrastructure, equipment, and other fixed assets, measured in chain-linked volumes (index 2010=100).
6. **Export & Import**: Annual value in money terms of goods exported and imported, measured in million ECU/EURO.
7. **Energy Consumption**: Final energy consumption, measured in thousand tonnes of oil equivalent, for industry, households, and other consumers.
8. **Climate Losses**: Annual losses from climate-related impacts including extreme weather events and natural disasters; measured in million euros.
9. **Gini Coefficient**: A measure of income inequality. The coefficient of 0 expresses perfect equality, while 100 expresses maximum inequality.
10. **Consumption**: Household expenditure on goods and services, in chain-linked volumes (index 2010=100).
11. **Government Expenditure**: Total public sector expenditure, in million euros.

A Note on the Model

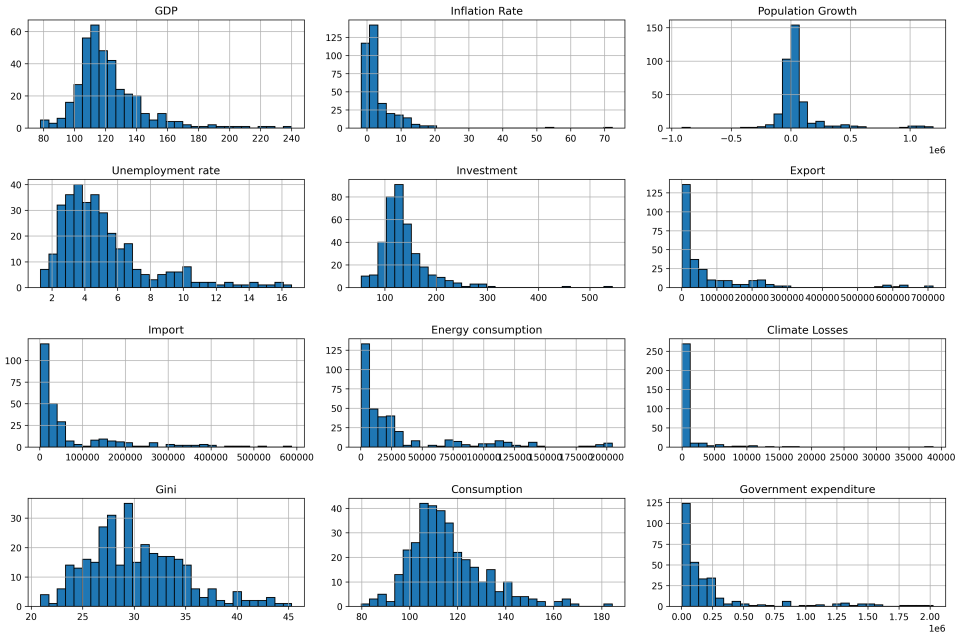
This model is a first draft to comprehend the impact of these variables on GDP and to get preliminary feedback. It is work in progress and an effort is being done to collect more data to further modify and enhance it.

2 Data Import and Preliminary Exploration

The dataset used in this project is downloaded from Eurostat in the form of an Excel file. The dataset has been given a MultiIndex structure with indices `Time` and `Country`, which gives an easier way to manipulate and analyze the data in periods and across regions. It contains information on socio-economic indicators for many countries over the years 2014 to 2023.

2.1 Feature Distributions

A distribution plot was made for each feature to see how spread out the values of each feature are across the dataset. These histograms show statistical properties of the data and illustrate the necessity of preprocessing in some instances.

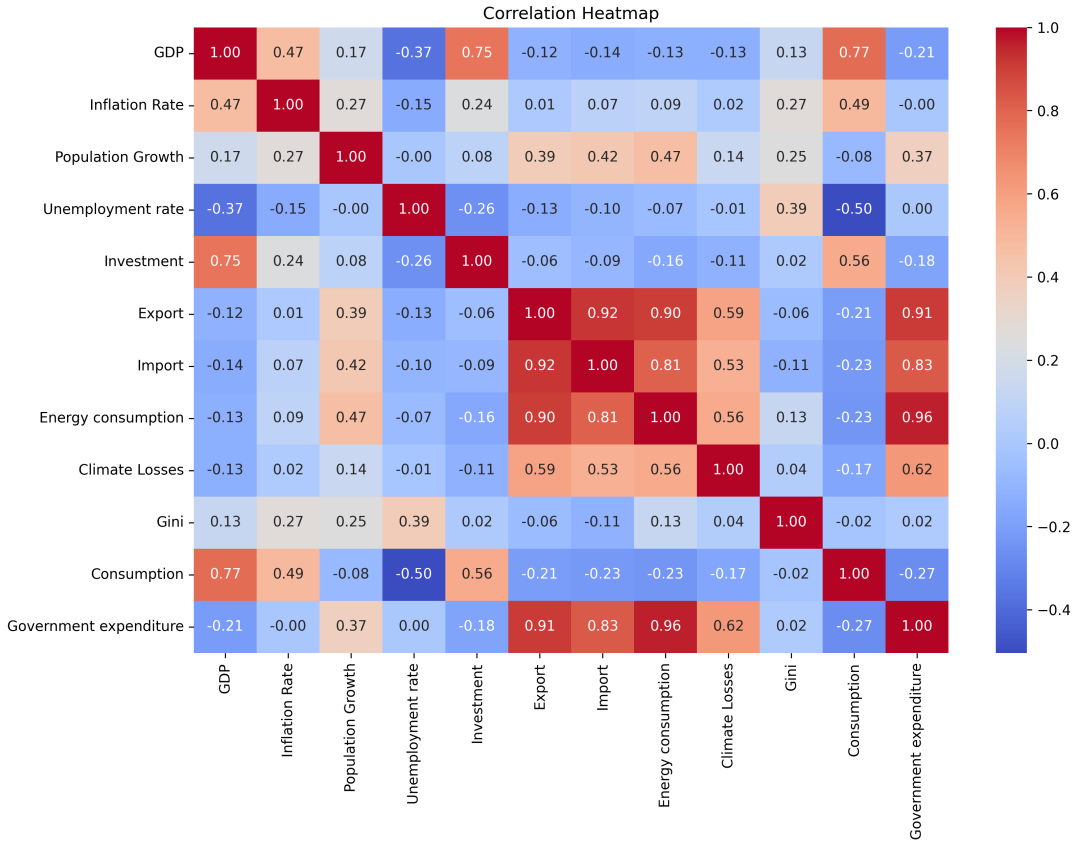


2.2 Correlation Heatmap

A heatmap for correlations was generated to understand the interrelation between the features in the dataset. The heatmap reveals several significant correlations:

- **GDP and Investment**: Strong positive correlation (0.75), which means higher Investments are most probably associated with higher GDP.
- **Export and Import**: Very high correlation of 0.92, indicating the interdependent nature of trade.

- **Climate Losses and Government Expenditure**: Moderately positive correlation of 0.62, indicating a relationship between climate-related impacts and public spending.



2.3 Summary Statistics

The `describe` method was used to generate statistical summaries for each feature. Key observations include:

- **GDP**: Ranges between 78.07 and 240.03, its mean value is 121.85, while its standard deviation equals 22.00.
- **Losses due to Climate Change**: Highly variable, between 0 and 38,744 million euros in a single year. This justifies the high variability, hence also potential impacts on GDP.
- **Population Growth and Energy Consumption**: Display rather large ranges as well; this reflects diverse economic conditions among countries and years.

3 Preprocessing

3.1 Creating a New Feature: Export-to-Import Ratio

In order to enrich the dataset, a new feature was generated and named *Export-to-Import Ratio*. This ratio gives an insight into the trade balance of each country and how it may

potentially impact GDP. This feature is calculated by dividing the value of exports by the value of imports.

This new feature brings additional insight into the trade balance of each country, which is expected to have an effect on the prediction of GDP. Further preprocessing steps will refine the dataset for model development.

3.2 Scaling the Data

To standardize the features and to have them on the same scale, the dataset was scaled using `StandardScaler` from the `sklearn.preprocessing` library. It transforms each feature to have a mean of 0 and a standard deviation of 1. Standardization is critical for models that depend on distance computation or gradient-based optimization methods since it prevents features with large magnitudes from dominating the model.

Having scaled the dataset, numerical values were transformed and made suitable for further analysis and training of machine learning models. The scaled dataset ensures that all features contribute equally to the learning of the model.

3.3 Handling Missing Values

Missing values in the dataset were handled using `KNNImputer` from the `sklearn.impute` module. This method replaces the missing values with the average values of the closest neighbors, ensuring the data is consistent and unbiased. The steps included:

- First step is the computation of the missing values for each column:
  - **GDP**: 4 missing values
  - **Inflation Rate**: 22 missing values
  - **Population Growth**: 11 missing values
  - **Unemployment Rate**: 43 missing values
  - **Investment**: 4 missing values
  - **Export**: 110 missing values
  - **Import**: 110 missing values
  - **Energy Consumption**: 24 missing values
  - **Climate Losses**: 70 missing values
  - **Gini**: 52 missing values
  - **Consumption**: 33 missing values
  - **Government Expenditure**: 80 missing values
  - **Export-to-Import Ratio**: 110 missing values
- Using the `KNNImputer` with 5 neighbors to impute the missing data.

After imputation, missing values were handled successfully, making the whole dataset complete for further analysis and modeling.

4 Clustering

The `K-Means` algorithm was implemented to find patterns in the data. The `K-Means` algorithm groups the features in data points into clusters, depending on their similarities.

4.1 Elbow Method for Optimal Clusters

The elbow method was used to determine the optimal number of clusters, which is denoted as k. Inertia values, defined as the sum of squared distances of samples to their closest cluster center, were plotted against various values of k. From the plot, there is a visible bend at k=2, indicating that two clusters are a good choice.

4.2 K-Means Clustering

Using k=2 clusters, the `K-Means` algorithm was run on the dataset, and a new feature, *Cluster*, was created to indicate the cluster assignment for each data point. The clustering process provides insight into group-level patterns and similarities in data.

4.3 Silhouette Score

The silhouette score was computed in order to assess the quality of clustering. It measures how similar each data point is to its cluster compared to other clusters. With a score of 0.739, scaled between 0 and 1, the resulting score indicates that the clusters are well identified and separate.

5 Data Splitting

After preprocessing the dataset, the next step was to split the data into a training and a test dataset to build and evaluate models more efficiently.

5.1 Train-Test Split

Proper model evaluation process, the data was divided as follows:

- **Features**: All columns except *GDP*.
- **Target Variable**: The *GDP* column, the variable to be predicted.
- **Training Set**: 80% of the data, used for training machine learning models.
- **Testing Set**: 20% of the data, held out to test model performance on unseen data.

This ensures that the model is trained on an independent subset of the entire dataset and later tested on unseen data. It prevents overfitting and provides an unbiased estimate of predictive performance.

6 Regression Model

This section explores modeling GDP and some socioeconomic features through regression analysis using a polynomial regression model.

6.1 Polynomial Regression

At this step, a degree-1 polynomial regression model was used. The following procedure was followed:

- **Feature Transformation**: The `PolynomialFeatures` method was used to transform features into terms of polynomial degree one.
- **Model Training**: A linear regression model was fitted on the transformed training dataset.
- **Predictions**: The fitted model was used to make predictions on the testing dataset for GDP values.

6.2 Model Evaluation

Model performance was tested with the R^2*: coefficient of determination*, which tells how much the independent variables explain the variance in GDP. The model scored 0.867 R^2, meaning that it fit the data very well and captured the underlying relationships of features to GDP effectively. This is a good baseline.

7 AutoML for Model Selection

The ultimate algorithm would be found using AutoML. AutoML automates the selection of models and hyperparameter optimization, constituting an efficient approach to evaluating various algorithms. This methodology aims to identify a specific algorithm that predicts GDP most effectively.

7.1 AutoML Setup

This is how the `flaml.AutoML` was applied:

- **Algorithm Candidates**: Models like XGBoost, Extra Trees, and Random Forest were considered.
- **Evaluation Metric**: Model performance was evaluated using R^2.
- **Cross-validation**: A five-fold cross-validation technique was implemented to ensure robust evaluation.

7.2 Results

After running AutoML with a time budget of 300 seconds, the highest-performing algorithm was identified as **Extra Trees Regressor**. The following were the optimal hyperparameters:

- **Number of Estimators**: 154
- **Maximum Features**: 1.0
- **Maximum Leaves**: 151

7.3 Performance of Models

Using the optimized Extra Trees Regressor, an R^2 score of 0.927 was achieved on the test dataset. This score is significantly higher than the polynomial regression model, demonstrating how AutoML improves model selection and quality.

8 Hyperparameter Tuning

Hyperparameter tuning of the Extra Trees Regressor was performed to further improve its performance using exhaustive grid search. This method evaluates all possible parameter combinations to identify the best configuration.

8.1 Configuration for Grid Search

The following hyperparameters were considered for tuning:

- **Number of Estimators**: [50, 100, 150]
- **Maximum Features**: [0.5, 0.6, 0.7, 0.8]
- **Maximum Leaf Nodes**: [100, 120, 140]
- **Minimum Samples Split**: [2, 5, 10]
- **Minimum Samples Leaf**: [1, 2, 5]

A five-fold cross-validation was used to ensure accurate and unbiased evaluation of each parameter combination.

8.2 Results

After evaluating 324 parameter combinations, the optimal configuration for the Extra Trees Regressor was identified as:

- **Number of Estimators**: 150
- **Maximum Features**: 0.5
- **Maximum Leaf Nodes**: 140
- **Minimum Samples Split**: 2
- **Minimum Samples Leaf**: 1

With these parameters, the Extra Trees Regressor achieved an R^2 score of 0.915 on the testing dataset.

9 Feature Importance Analysis

Feature importance provides insights into the contribution of each feature to the model's predictions. Using the optimized Extra Trees Regressor, the following key findings were identified:

- **Consumption**: Most influential feature, contributing 30.45% to the predictions.
- **Investment**: Second most influential feature, contributing 27.22%.
- **Inflation Rate**: Contributed 9.94%, indicating a significant impact.
- **Export and Unemployment Rate**: Contributed 5.00% and 4.80%, respectively, reflecting moderate influence.
- **Gini**: Contributed 4.80%.
- **Climate Losses**: Contributed 0.93%, providing complementary insights.

This analysis underscores the importance of a holistic approach to analyzing economic indicators for GDP prediction.

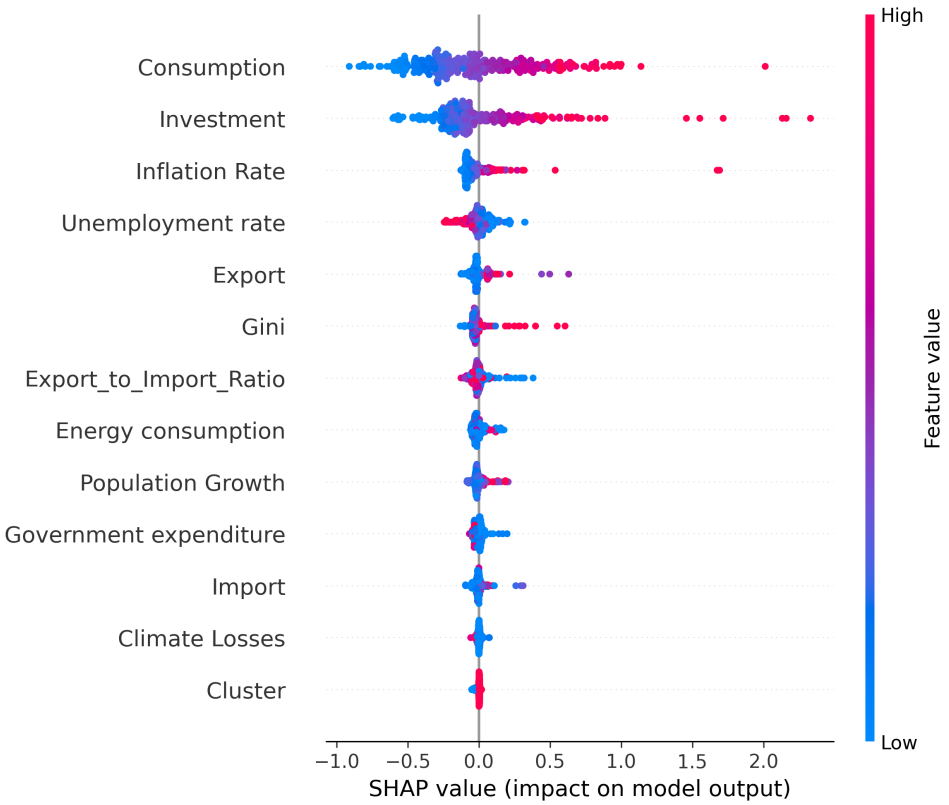
10 SHAP-Based Feature Importance Analysis

To get an understanding of how each feature affects the model's predictions, SHAP values were used. SHAP is a powerful interpretability technique that can be used to understand the contribution of each feature in the model's decision-making process, both globally and locally.

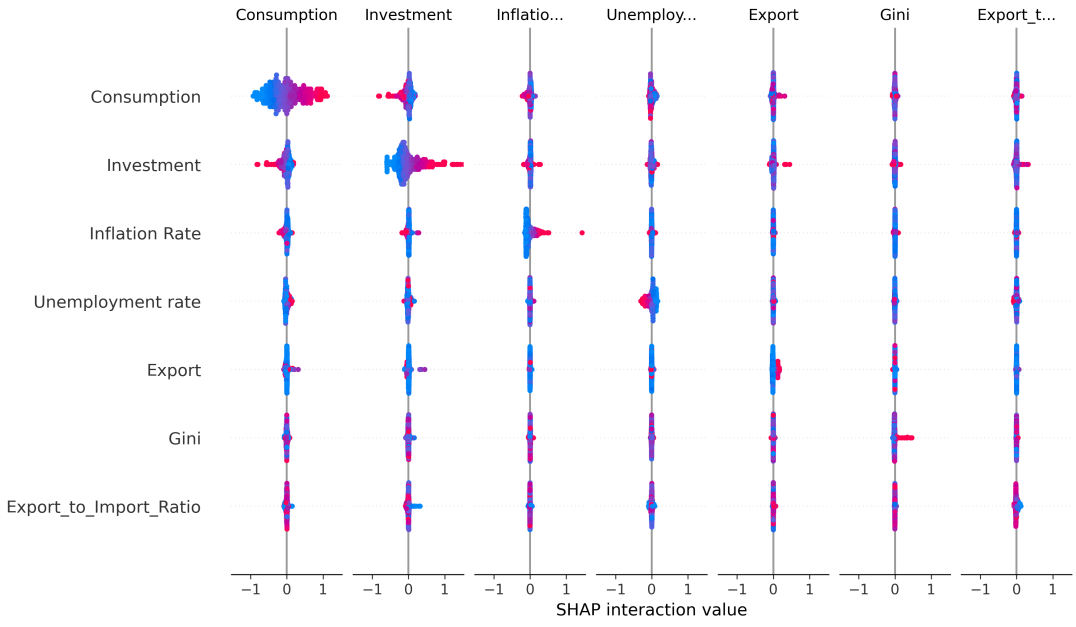
Key Insights

- Consumption and Investment continued to show the highest impact on GDP predictions, which is consistent with the earlier feature importance analysis.
- SHAP values provided insights into how features like Inflation Rate and Export interact with the predictions of GDP, showing their detailed relationships.
- Environmental factors such as Climate Losses highlighted subtle interactions that could influence GDP under specific conditions.

A SHAP summary plot was created to visualize these insights, showing the magnitude and direction of each feature's contribution to the predictions.



Further, an investigation of SHAP interaction values was done to determine the interaction between features. These interactions reveal complex dependencies, offering deeper insights for future model improvements and decision-making processes.

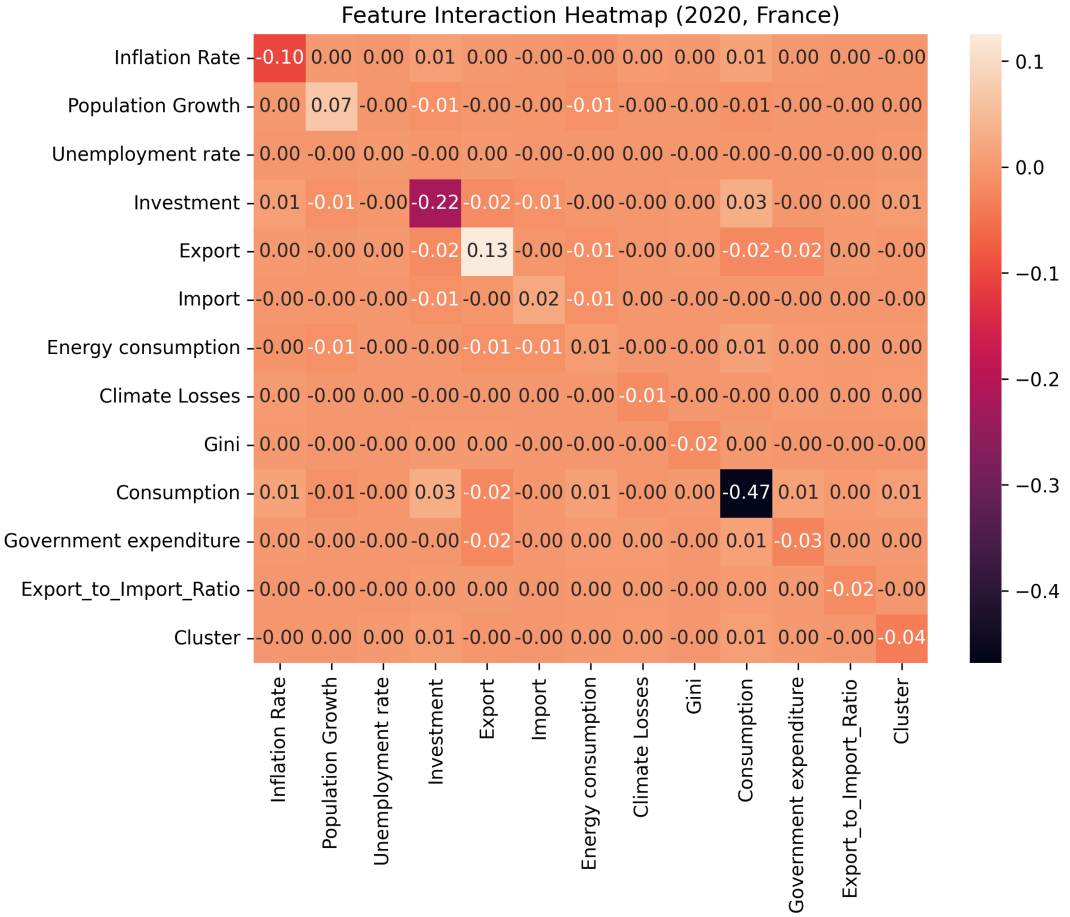


This analysis reinforced the earlier feature importance findings, showcasing the robustness of the Extra Trees Regressor in selecting features that drive GDP. Moreover, SHAP provided actionable insights for future model improvement and policy decisions.

11 Feature Interaction Analysis for 2020, France

This analysis accounts for different factors influencing the GDP forecast of France during the year 2020. The selected features include Inflation Rate, Investment, and Consumption, among others.

A Feature Interaction Heatmap using SHAP interaction values illustrates the interaction effect on GDP. For instance, investment and consumption demonstrate a positive interaction (0.03).

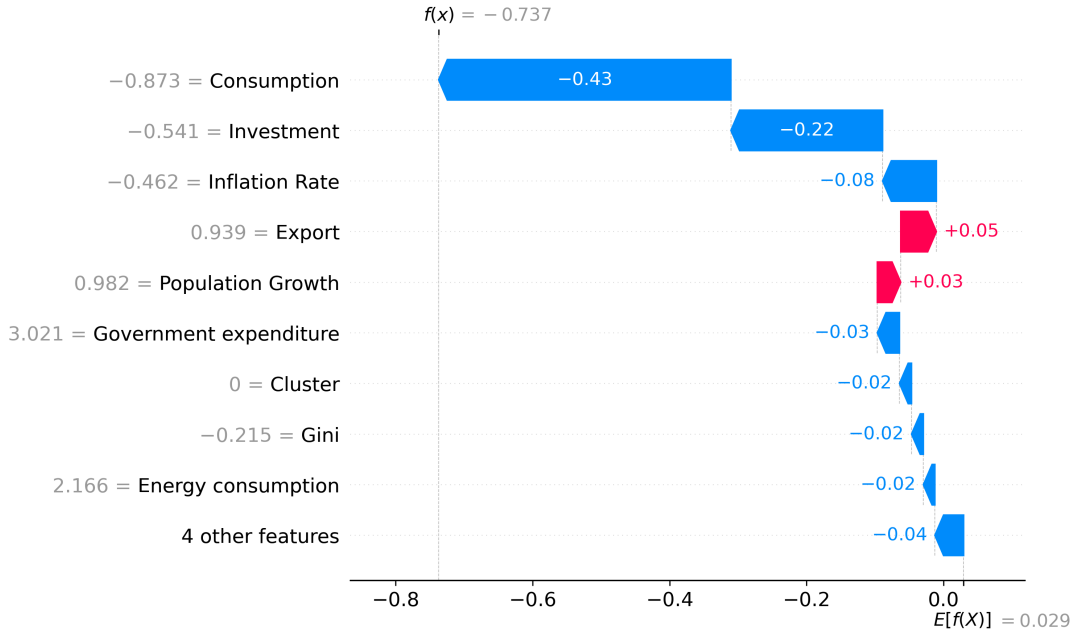


In addition, the SHAP waterfall plot details how each feature contributes to the predicted deviation of GDP for France in 2020. This visualization shows the cumulative impact of features, from the base value which is the average prediction (0.029) to the final model output(-0.737).

Key Observations

- **Consumption (-0.43):** The most negative contributor, indicating that lower household spending significantly impacts GDP.
- **Investment (-0.22):** A crucial factor showing how decreased investments negatively affect GDP.
- **Export (+0.05)** and **Population Growth (+0.03):** Positive contributors, reflecting their supportive roles in GDP predictions.

- **Government Expenditure (-0.03):** A moderate negative impact, emphasizing its complex relationship with GDP.



The waterfall plot provides nuanced insights into the interplay of features and their contributions, enhancing the understanding of socioeconomic factors shaping France's GDP in 2020.

Conclusion

It gives evidence that machine learning models can predict the impact of different factors on GDP, with accuracy over 90%. The results will enable us to judge how changes in features such as investment or consumption of any country and year affect GDP.

By gathering further data and concentrating on features that impact the most, with machine learning methods like SHAP, I want to bring the model's accuracy well above 95%. With improved predictions, this should also yield better tools by which to understand and cope with economic trends.