**Abstract**

Macroeconomic indicators let countries focus on goods, services, and other entities that boost their GDP. Nations frequently struggle to identify these categories of indicators. The study evaluated a typical data collection with two primary goals. First, six machine learning algorithms were used to forecast GDP based on macroeconomic indicators: Multiple linear Regression (MLR), Ridge Regression (RR), Support Vector Regression (SVR), Random forest Regression (RFR), Adaboost Regression, and Gradient Boosting Regression. Second, identify significant macroeconomic variables that may impact GDP growth. The approaches were examined by random grid cross-validation. The Adaboost regression method outperformed previous methods, predicting GDP to macroeconomic data with 81% accuracy and a root mean square error of 0.056. Some macroeconomic indicators impacted GDP positively, while others did not. The study’s main contribution is the use of machine learning regularization algorithms to predict GDP, rather than standard statistical approaches. It identified additional macroeconomic variables for the calculation of real GDP.

**1. Introduction:**

The computation of market value for total goods and services produced in a country for a given period to determine the size and strength of her economy represents the real Gross Domestic product (GDP). It is expressed as a nation’s five macroeconomic indicators: consumption, investment, government spending, and export and import rates (WikiBooks, 2021).

A country’s progress relies heavily on its GDP growth rate. Predicting GDP can assist governments in making informed decisions to advance the country’s economic development. To make the best policy judgments, policymakers, governments, and economists must understand the economic cycle. To make those decisions, they must consider both prior and current economic situations. Gross domestic product (GDP) is the primary metric for determining the performance of our economy. It assists our economists and policymakers in determining whether the economy is increasing or declining, allowing them to make policy decisions.

This paper aims to improve GDP prediction accuracy using machine learning algorithms. We investigate which model of Machine Learning algorithm is best suited to predicting our GDP growth rate. We analyze 43 years of data to determine the best Machine Learning model for predicting GDP growth and identifying the most relevant independent data, features, and parameters. The study attempts to predict the GDP to macroeconomic indicators by applying six machine learning models: Multiple Linear Regression (MLR), Ridge Regression (RR), Support Vector Regression (SVR), Random Forest Regression (RFR), Gradient Boosting Regression, and Adaboost Regression.

We use machine learning algorithms to predict GDP growth in Bangladesh, taking into account other variables to determine the best independent variable and ML algorithm. Our goal is to better understand how other variables impact GDP growth. There is a constant flow of information and research about economic growth. Lack of data analysis prevents economists from providing effective guidance. We want to assist economists and governments in promoting GDP growth. Our goal is to introduce economists to machine learning and demonstrate its potential for developing economic theory. Using machine learning, we can test the limitations of its forecasting power and make economics more policy-conscious. This initiative will assist economists and economic authorities in analyzing economic growth by forecasting GDP. Bangladesh’s economy has been impacted by several events, including political conflicts, natural catastrophes, and budget imbalances. Bangladesh is a developing country, which requires consistent economic growth to reach its aim of becoming a developed country. Economists can assist the government in ensuring economic stability by analyzing present conditions and forecasting future growth. However, there is always an abundance of information and research on economic growth.

2. **Methods and Techniques:**

Machine learning is a type of artificial intelligence in which computers may learn from data and improve their performance without being explicitly programmed. It focuses on building systems that learn or improve performance based on the data they consume.

In general, GDP is influenced by many macroeconomic indicators which have been featured under gross Domestic Product Current Price, Consumption Expenditure, Private Consumption Expenditure, Government Expenditure, Gross Fixed Capital Formation, Change in Stocks, Export of Goods and Services, Import of Goods and Services, Agriculture, Industry, Services, Reserve, External Debt, Foreign Direct Investment, Consumer Tendency Index, Domestic Investment, Foreign Investment, Inflation Rate, Central Government Debt and Unemployment Rate.

But for this study, the Export rates, Import rates, Final Consumption Expenditure (FCE), General government final consumption expenditure (G\_F\_C\_E), Gross Domestic Savings (D\_Savings), Gross Savings (G\_Savings), Inflation rates, Remittances, Unemployment rate, Industry, Reserve are the chosen macroeconomic variables as a percentage of GDP growth, making up 11 predictors while GDP is the response variable. Goods and services imports and exports have an impact on GDP growth rates. A nation has a trade deficit if it imports more than it exports, as opposed to a trade surplus otherwise. General government final consumption expenditure (formerly general government consumption) includes all government current expenditures for purchases of goods and services (including compensation of employees). It also includes most national defense and security expenditures but excludes government military expenditures that are part of government capital formation (World Bank). Gross Domestic Saving is GDP minus final consumption expenditure. It is expressed as a percentage of GDP (The Economics Times). Gross savings are calculated as gross national income less total consumption, plus net transfers (World Bank). The rate at which prices rise over a specified period time is known as inflation. The term ‘remittance’ more widely describes the money migrants, who work and reside overseas, send to their family back home. When someone is employable and actively looking for work but is not able to find one, they are said to be unemployed.

**2.1. Machine learning methods for the study:**

The dataset used in this study includes quantitative (numerical) information that influenced the use of machine learning regression, namely: MLR (Multiple Linear Regression, RR (Ridge regression), LR (Lasso Regression), SVR (Support vector regression), Random Forest Regression, Gradient Boosting regression, and Adaboost regression. A statistical method called multiple linear regression (MLR), or just multiple regression, uses many explanatory variables to forecast a response variable.

**2.1.1 Linear Regression model:**

Linear regression is widely used to model variable relationships. The model assumes a linear relationship between independent and dependent variables. The model uses a linear equation to express the relationship between the independent and dependent variables. The intercept value represents the point on the Y-axis where the regression line crosses. This method is commonly employed in simple regression when a single independent variable predicts the dependent variable’s numerical value.

The formula for simple linear regression is as follows:

…………… (1)

Where Y is the dependent variable, is the intercept and is the coefficient for the independent variable X. In multiple linear regression, the formula is expanded to include multiple independent variables:

…………….. (2)

Where,, …., are the independent variables and,, … are their corresponding coefficients.

**2.1.2 Ridge regression:**

Ridge regression is one of the shrinkage methods that is very similar to ordinary least squares (OLS). This method was designed to overcome the instability of the least square’s estimator by penalizing the coefficients on -norm. Its coefficient estimates are the values that minimize:

Where is a tuning parameter, which controls the strength of the penalty term P is the number of features used in modeling. The , called shrinkage penalty, is small when, …, are close to zero. The effect of the shrinkage penalty is shrinking the estimates of towards zero.

**2.1.3 Support Vector Regression:**

Support vector regression (SVR) is a sort of support vector machine (SVM) that performs regression problems. It seeks to identify a function that best predicts a continuous output value given an input value. The hyperplane is a linear function of the form:

Where w is the weight vector, x is the input vector, and b is the bias. To maximize the margin, we need to minimize:

Where *l* is the sum of training points, C > 0 is the regularization parameter that constrains/regularizes or shrinks the coefficient estimates towards zero. The first term in the error function is a penalty term that increases as the model becomes more complex. The second term is the -insensitive loss function that penalizes errors that are greater than, allowing flexibility to the model.

**2.1.4 Random Forest Regression:**

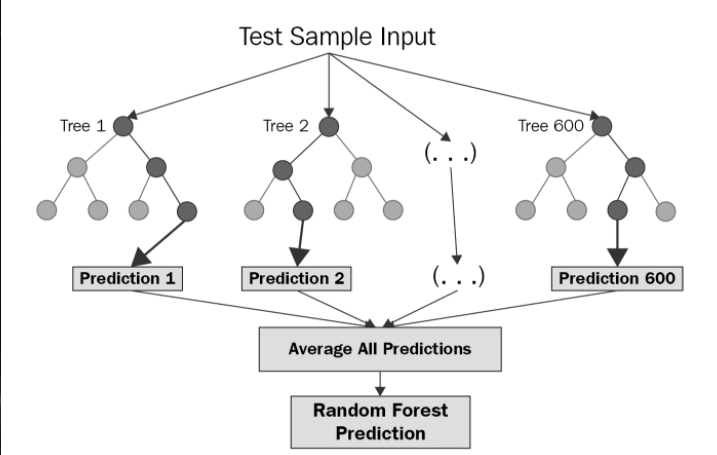
Random Forest (RF) is a set of tree predictors in which each tree is determined by the value of a random vector collected separately and uniformly across the forest. It is a significant departure from the bagging technique in that there is no interaction between trees while contracting the random forest. RF can be applied to classification and regression modeling. If RF is used for regression, it takes the average of all tree forecasts. Figure 1 illustrates the Random forest algorithm

Random Forest combines many binary regression trees, built using several bootstrap samples on the dataset that consists of a response and p inputs, for each of N observations. Let us consider a learning set L consisting of for, with. To grow a regression tree, the algorithm needs to decide on the splitting variables and splitting points, and also what shape the tree should have. The steps carried out by the algorithm are as follows:

1. Suppose first we have M regions that partitioned the dataset into
2. The model of the response is a constant value in each region.
3. The predicted value of obtained by averaging in region :

The optimum binary partition in the regression tree was discovered by experimenting with various threshold values and selecting the threshold with the lowest sum of squares. For example, for region and, we seek the splitting variable j and splitting point s that solve:

After finding the best split, the dataset is partitioned into two resulting subsets. Then the process continues until each node reaches a user-specified minimum node size and becomes a terminal node.



**Figure 1: Random Forest Algorithm Illustration.**

**2.1.5 Gradient Boosting Regression:**

Friedman (2001) created the gradient boosting model, a type of ensemble machine learning. The gradient boosting model works by combining numerous weak learners to increase the final model’s accuracy and robustness.

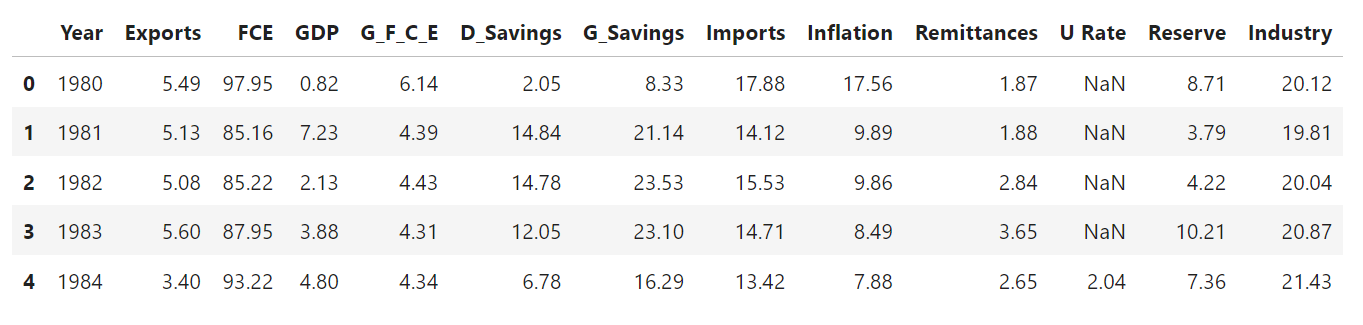
The gradient boosting model begins by generating a single leaf and constructing regression trees. A regression tree is a sort of decision tree that is intended to estimate a continuous real-valued function rather than a classifier. The regression tree is built iteratively by dividing data into nodes or branches, which are then broken into smaller groups. Initially, all observations are classified into the same group. The data is divided into two partitions based on all available predictors. The predictor that splits the tree is the one that most clearly separates the observations into two distinct groups and minimizes the residual error. The study is quantified using the Friedman MSE, which was introduced by Friedman (2001).

The gradient boosting model trains a new tree based on the previous one’s error and continues until the desired number of trees cannot be improved. To prevent overfitting, the gradient boosting model scales the contribution from the new tree based on the learning rate.

**2.1.6 Adaboost regression:**

**2.2 The study’s data:**

The data set of interest is Bangladesh’s macroeconomic indicators, which were obtained and extracted from DataBank World Development indicators. It contains (43) instances from 1980 and 2022, with (13) attributes that have missing values. Where our dataset has 40x13 example vectors. Bangladesh’s macroeconomic data from 1971 to 1979 were not included in the experiment due to their negative values. And there are 11 features in dataset and these are: ‘Year’, ‘GDP’, ‘Exports’, ‘FCE’, ‘G\_F\_C\_e’, ‘D\_Savings’, ‘G\_Savings’, ‘Imports’, ‘Inflation’, ‘Remittances’, ‘U Rate’, ‘Reserve’, ‘Industry’. Since, we want to predict ‘GDPGrowth’ rate thus, ‘GDPGrowth’ is our dependent feature and others are independent features.



**Figure 2: Dataset Example**

**3. Calculation:**

The experiment was carried out using Python 3.11.5 on Jupyter Notebook which is an explore–execute environment containing a large library of machine learning algorithms for data analysis. Estimated coefficients of the variables were computed using some calculus built inside Python machine learning.

**3.1 Numerical variables:**

We verify the number of numerical values in our dataset. We categorize all features into two categories: discrete and continuous.

All features are numerical, with 12 being continuous ‘GDP’, ‘Exports’, ‘FCE’, ‘G\_F\_C\_e’, ‘D\_Savings’, ‘G\_Savings’, ‘Imports’, ‘Inflation’, ‘Remittances’, ‘U Rate’, ‘Reserve’, ‘Industry’ and left feature is discrete (‘Year’).

**3.2 Replacing Missing value:**

Finally, we calculate the mean and median of the ‘Reserve’ and ‘U rate’ features respectively, and replace null values with the mean and median.

After replacing all missing values with mean and median the output of the dataset:

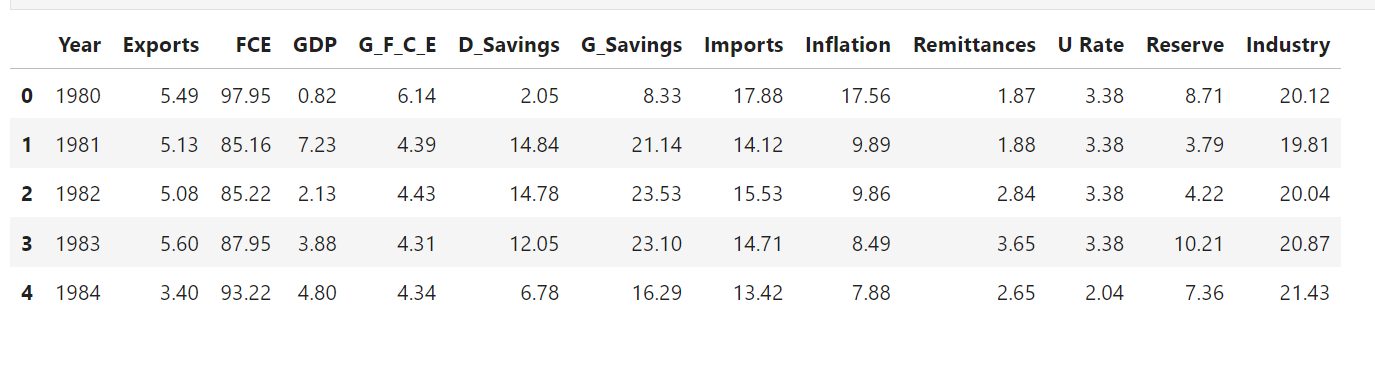
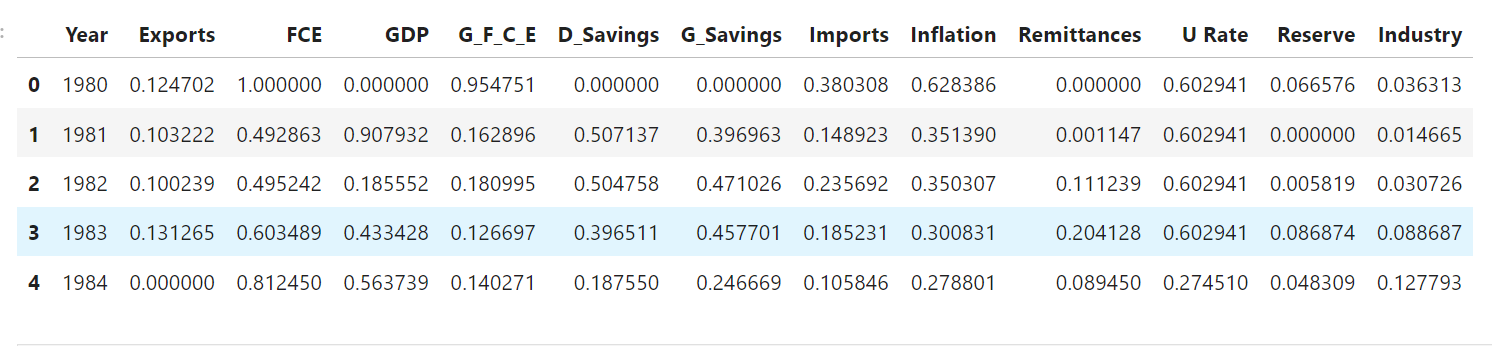


Figure 3: Dataset without missing values

**3.3 Features Scaling:**

We utilize a fixed range due to its significant variability to standardize the independent parameters in our dataset. We use the Min-Max Normalization approach to scale our data.

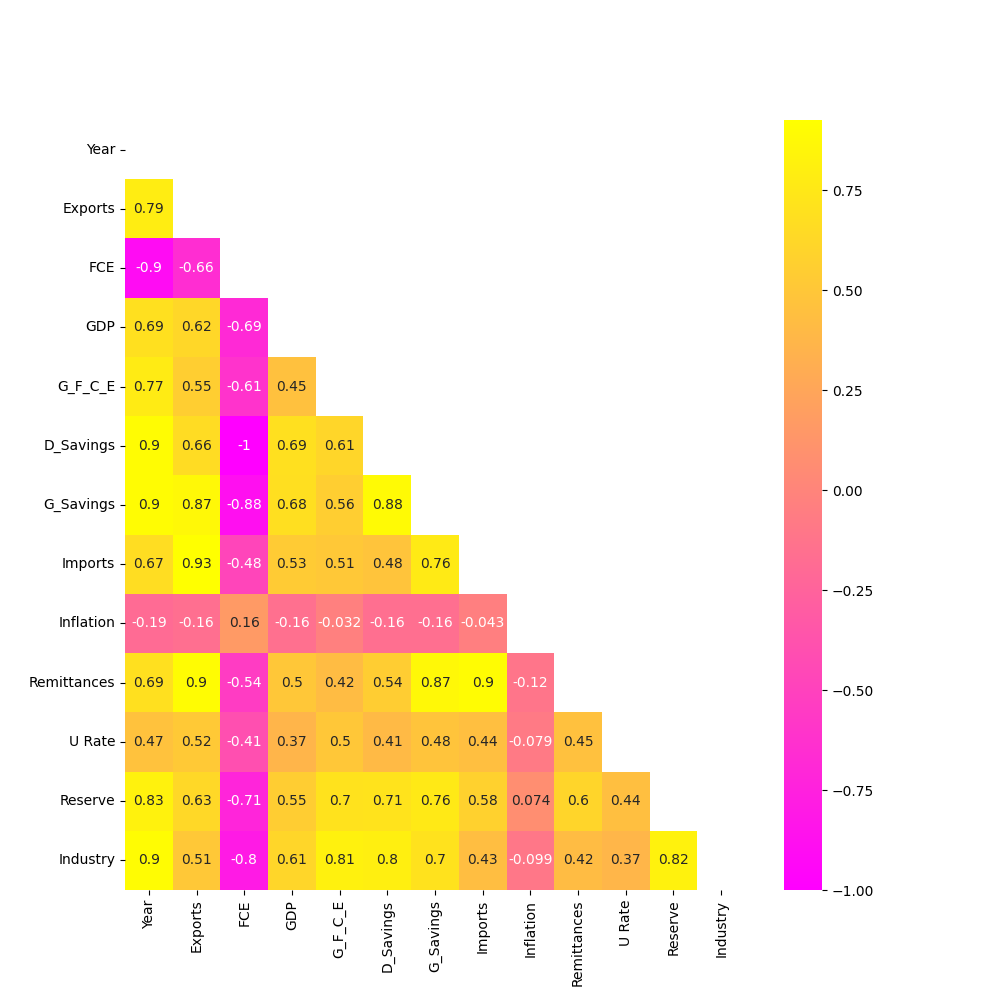
After feature engineering, our dataset is ready for training. We scaled our dataset from 0 to 1.



**Figure 4: Scaled dataset**

**3.4 Relation between Independent and Dependent Feature:**

In this step, we try to capture the best parameters for our machine-learning model. First, we try to determine which parameters are highly related to our GDP growth rate. GDP is moderate to strongly positively correlated with most economic indicators, including exports, domestic savings, gross savings, imports, remittances, reserves, and industry. It has a moderate negative association with total consumer spending and a mild negative correlation with inflation. The positive link with unemployment rates may be unique to this dataset or context.



**Figure 5: Presentation of correlation**

**3.5 Splitting training from the testing dataset:**

Typically, the data necessary to test the results of predictive models may not be available. The dataset can be divided into training and testing sets. To gather the training and testing data for this study, we divided the dataset into an 8:2 ratio and used 80% of it to train the model and 20% to test it. Variable names were assigned to specific regions of the divided dataset. Eighty percent of the macroeconomic predictors (exports, domestic savings, gross savings, imports, remittances, reserves, industry, unemployment rate, and inflation) were collectively assigned to xtrain variable along with the corresponding 80% instances of response (GDP) assigned to ytrain variable for training the model. The remaining 20% instances of the macroeconomic predictors were collectively assigned to the xtest variable along with the corresponding 20% instances of the GDP assigned to the ytest variable for testing the model.

**3.6 Development and evaluation techniques for the model:**

The techniques for developing the predictive models for the MLR, RR, SVR, RFR, Gradient Boosting regression and Adaboost methods that were used in the study were from machine learning scikit-learn libraries in Python. For example, the following tools from scikit-learn libraries were used for developing the RR model. The “ridge.fit (xtrain, ytrain)” method was used for building the RR model, and the coefficients of the model were computed using “ridge.intercept\_” and “ridge.coef” and testing the prediction for the GDP was done using” ridge.predict (xtest)”.

Evaluation of the various machine learning models was done using Random Search CV to find out the best model. We do perform on six algorithms.

**Table 1: Model Parameters for Tunning**

|  |  |
| --- | --- |
| **Model Name/Algorithm names** | **Parameters for Algorithm** |
| Multiple Linear Regression | Normalize: [True, False] |
| Ridge Regression model | Alpha:[1e-15,1e-10,1e-8,1e-3,1e-2,1,5,10,20,30,35,40,45,50,55,100] |
| Support Vector Regression model | C: [0.1, 1, 10, 100, 1000],  Gamma: [1, 0.1, 0.01, 0.001, 0.0001],  kernel: ['rbf'] |
| Adaboost Regression Model | n\_estimators: [50, 100,150,200,500],  learning\_rate : [0.01,0.05,0.1,0.3,1],  loss : ['linear', 'square', 'exponential'] |
| Gradient Boosting Regression | n\_estimators: [ 100, 200, 300, 400,500, 600, 700, 800, 900, 1000,1100, 1200]  min\_samples\_split: [2,4,8],  learning\_rate: [0.01,0.1,1],  max\_features: [ 'auto', 'sqrt', 'log2'],  max\_depth: [5, 10, 15, 20, 25, 30] |
| Random Forest Regression Model | n\_estimators: [25, 50, 100, 150],  max\_features: ['sqrt', 'log2', None],  max\_dept': [3, 6, 9,10,15,20],  max\_leaf\_nodes: [3, 6, 9] |

Then after tunning all models, the output looks like this:

**Table 2: Best Parameters**:

|  |  |  |
| --- | --- | --- |
| **Model Name/Algorithm names** | **Best\_Score** | **Best parameter** |
| Multiple Linear Regression | 0.809 |  |
| Ridge Regression model | 0.788 | Alpha=100 |
| Support Vector Regression model | 0.667 | C=10, gamma=0.001 |
| Adaboost Regression Model | 0.807 | learning\_rate=0.05,  loss='square' |
| Gradient Boosting Regression | 0.854 | learning\_rate=1,  max\_depth=15,  max\_features='sqrt',  min\_samples\_split=4,  n\_estimators=400 |
| Random Forest Regression Model | 0.591 | max\_depth=6,  max\_features='log2',  max\_leaf\_nodes=3,  n\_estimators=25 |

**4. Result and discussion:**

It is necessary to establish the macroeconomic indicators responsible for GDP growth and the specific nature of their influences on GDP with a more reliable machine learning approach, to achieve a high level of economic development. If these relationships are established, they will enable economic policymakers to redesign the influencing variables to achieve the desired GDP growth rate. The contribution of this study comes in two main dimensions. The first is to use machine learning methods to build predictive models that would be suitable for predicting whether GDP would grow or not, given macrocosmic data. In this regard, the study applies and compares MLR, RR, SVR, RFR, Gradient Boosting regression, and Adaboost methods. The second goal is to improve the economy’s performance by identifying key macroeconomic indicators that impact GDP. We want to determine the relationship between GDP and macroeconomic factors by minimizing two terms: RSS and the cost function (Lambda). Adding the cost function allows for a trade-off between bias and variance, eliminating the linear link between GDP and macroeconomic factors. This flexibility aims to reduce model overfitting and mean square error (MSE) while improving forecast accuracy.

Also, our target is to predict the GDP growth next five years. Using six models our predicted values are in table format.

**4.1. Predictive Accuracy and Result Compression:**

The accuracy of MLR, RR, SVR, RFR, Gradient Boosting Regression, and Adaboost regression was 80.9%, 78.8%, 66.7%, 59.1%, 85.4%, and 80.7% respectively, with no missing values, accounting for all 43 observations.

**Table 3: Comparison between three models:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model Name | Sore on training data | Score on testing data | RMSE | MAE (Mean Absolute error) | MAE (median absolute error) | Accuracy |
| Multiple linear Regression | 0.548 | 0.809 | 0.1499 | 0.115 | 0.094 | 0.809 |
| Adaboost regression | 0.94 | 0.81 | 0.056 | 0.0408 | 0.031 | 0.81 |
| Gradient Boosting Regression | 1.00 | 0.85 | 1.346 | 1.807 | 0.0 | 0.85 |

Though gradient boosting regression has greater accuracy it has comparatively greater RMSE, MAE, and Median Absolute Error. On the other hand, multiple linear regression and Adaboost regression have almost equal accuracy but in Adaboost regression model has comparatively lower RMSE, MAE, and Median Absolute Error thus, we select the Adaboost regression model as the best model for our research.

Root Mean Square Error (RMSE) =

The result of our model RMSE =0.056

Mean Absolute Error (MAE) =

The result of our model MAE: 0.0408

Median Absolute Error =

The result of our model Median Absolute Error: 0.031

**4.2. Prediction:**

Our goal here is to predict the GDP growth for the next five years, thus we choose Adaboost as the best regression model. Using this model, the next five years 2023, 2024, 2025, 2026, and 2027 will be 6.34, 5.11, 4.35, 5.17, and 5.00 respectively.

**5. Conclusion:**

A country’s GDP is extremely important. It provides further information on the size and sustainability of a country’s economy. The Adaboost regression model predicts macroeconomic indicators accurately, with minimal RMSE better than MLR, RR, SVR, RFR, and Gradient boosting regression. Most likely, an increase in exports, domestic savings, gross savings, imports, remittances, reserves, and industry leads to an increase in GDP growth, while inflation and total consumer spending reduce GDP growth. Various causes, including political disputes, natural catastrophes, and budget imbalances influence Bangladesh’s economic situation. Bangladesh, a developing country, requires consistent economic growth to reach its aim of becoming a developed country. Economists play a crucial role in ensuring economic stability by advising governments on current economic conditions and future growth prospects. This study will determine which characteristics should be prioritized to achieve greater GDP growth, while also assessing the prediction power of machine learning. We will demonstrate how machine learning may advance economic theory.

**Research Gap:**

Investigate the impact of International Flow and Product services on GDP.

**References:**

Agu, S. C., Onu, F. U., Ezemagu, U. K., & Oden, D. (2022). Predicting gross domestic product to macroeconomic indicators. *Intelligent Systems with Applications*, *14*, 82. https://doi.org/10.1016/j.iswa.2022.200082

Borgohain, B. R., & Singh, K. K. (2022). An Empirical Analysis Using Machine Learning to Identify Features Influencing the Green Economy of India. *Jindal Journal of Business Research*, *11*(1), 7–23. https://doi.org/10.1177/22786821221082611

Cicceri, G., Inserra, G., & Limosani, M. (2020). A machine learning approach to forecast economic recessions-an Italian case study. *Mathematics*, *8*(2), 1–20. https://doi.org/10.3390/math8020241

Hasan, M., Sattar, A., Hossain, A., Hossen, M. D., & Ahmudul Hasan, M. (2021). GDP growth prediction of bangladesh using machine learning algorithm. *Ieeexplore.Ieee.OrgA Hossain, M Hossen, MM Hasan, A Sattar2021 Third International Conference on Intelligent Communication, 2021•ieeexplore.Ieee.Org*. https://doi.org/10.1109/ICICV50876.2021.9388593

Lai, H. (2022). A comparative study of different neural networks in predicting gross domestic product. *Journal of Intelligent Systems*, *31*(1), 601–610. https://doi.org/10.1515/jisys-2022-0042

Ozden, E., & Guleryuz, D. (2022). Optimized Machine Learning Algorithms for Investigating the Relationship Between Economic Development and Human Capital. *Computational Economics*, *60*(1), 347–373. https://doi.org/10.1007/s10614-021-10194-7

Soybilgen, B., & Yazgan, E. (2021). Nowcasting US GDP Using Tree-Based Ensemble Models and Dynamic Factors. In *Computational Economics* (Vol. 57, Issue 1). https://doi.org/10.1007/s10614-020-10083-5

Yoon, J. (2021). Forecasting of Real GDP Growth Using Machine Learning Models: Gradient Boosting and Random Forest Approach. *Computational Economics*, *57*(1), 247–265. https://doi.org/10.1007/s10614-020-10054-w