The Impact of Trade and Expenditure on GDP

Table of Contents

The Impact of Trade and Expenditure on GDP

Abstract

This study examines the effect of Trade and Expenditure on GDP per capita. For this purpose, data is taken from the world bank indicator (WBI) Website which consists of GDP, Trade, Expenditure, and nine variables which include Foreign Direct Investment, Net Inflows, General Government Final Consumption Expenditure, Gross Capital Formation, Real Interest Rate, Exports of Goods and Services, Gross Domestic Savings, Net Income, Population Total, and Total Revenue, and each variable has 468 observations. The data is taken from 1970 to 2021 for nine developed countries the United States, the United Kingdom, Italy, France, Portugal, Germany, Australia, New Zealand, and India. There are five Machine Learning Regression Models which include Linear Regression, Random Forest, Lasso Regression, Ridge Regression, and Net Elastic Regression Models are used in this study, and all data analysis is performed by R software. According to the R result, we conclude that Random Forest is a more preferred model than the other models, and also conclude that there has a positive effect of Trade and a negative effect of Consumption Expenditure on GDP per capita.

Introduction

Machine Learning is the study of teaching computers to learn and act like people, and to enhance their learning over time in an independent manner, using data and information from observations and real-world interactions.

Machine learning is made up of three parts:

Making conclusions relies heavily on computational calculations. The selection is based on several factors and highlights. The framework is empowered (prepared) to learn based on the base knowledge for which the right reaction is realized initially, the model is concerned with

boundary data for which an acceptable response is known. The computation is then repeated, with revisions made as necessary, until the yield (Learning) agrees with the known solution. Expanding information measures now support the framework in learning and cycling higher computational options. It's anything, but an investigation of how to many PCs do the things which at present people can improve. Machine Learning is a sort of artificial insight that can be characterized is "AI is customizing PCs to streamline an exhibition model utilizing model information or Experience."

The framework ought to gain naturally from given information. "AI is a technique for information investigation that robotizes scientific model structure. It's anything but a part of man-made brainpower dependent on the possibility that frameworks can gain from information, distinguish examples, and settle on choices with negligible human intervention"." ML enables PCs to learn without being modified". It is firmly identified with information mining (analytics of information) and Statistics (prediction making/probabilities). If a PC program's presentation on T, as judged by P, improves with experience, it is said to acquire E for some assignment T and some exhibition measure P. Experience E is the train data. A specific execution measure is called exactness.

A few focuses that show the contrast between ML and Traditional techniques are given beneath ML calculations don't rely upon rules characterized by human specialists. They measure information in a simple structure. For instance, text, messages, records, web-based media content, pictures, voice, and video. An ML framework is a learning framework, off chance that it's anything but customized to play out an assignment. However, it is modified to figure out how to play out the undertaking ML is additionally more expectation situated, though Statistical Modeling is essentially understanding-focused. Not a strong differentiation, particularly as the orders merge. Most ML models are uninterpretable, and therefore, they are generally unacceptable when the reason for existing is to get connections or even causality. They, for the most part, function admirably where one just necessity expectations.

ML learns from previous computations or data and produces reliable and repeatable results for new data. ML produces perfect and efficient results for big data. ML provides efficient results for a large amount of data in very little time. ML is best for complex problems for which there is no good solution by using all traditional approaches. ML computation is very cheaper and more powerful and affordable storage. ML provides precise findings and aids enterprises in the development of statistical models based on real-time data.

The Impact of Trade and Expenditure on GDP

Supervised Machine Learning consists of a dependent variable which is also called the Target or Outcome variable is to be estimated and predicted from a given set of independent variables which is also called Predictor or Feature variables. We create a function that translates inputs to desired outputs using these sets of variables, including Dependent and Independent. The training process in Supervised Machine Learning continues until the ML model achieves the target degree of accuracy on the training dataset. Previous/Past dataset is used to make predictions for testing data in supervised machine learning.

The supervised Machine Learning method is concerned with teaching the ML model with the knowledge that allows it to understand and then predict future value using that knowledge. Supervised Machine Learning deals with labeled data. The model is trained using the labeled data so it can predict the future values concerning the sample data. The task of inferring a function from the training data is known as inference. The training data is made up of a set of observations and their outcomes. There are two main types of Supervised Machine Learning methods, Regression and Classification. Regression ML consists of Simple Linear Regression Model, Multiple Linear Regression Model, Polynomial Regression Model, Decision Tree, Random Forrest, Support Vector Machine, etc. Classification ML consists of Binary Logistic Regression Model, Multinomial Linear Regression Model, Ordinal Logistic Regression Model, SVM, DR, RF, XGB, etc.

Literature Review

In particular, a gradient boosting model and a random forest model are presented in this study as a strategy for building machine learning models to anticipate real GDP growth. Forecasts are provided for the years 2001 through 2018 in this study, which focuses on the real GDP growth of Japan. As a baseline, the predictions made by the Bank of Japan and the International Monetary Fund are employed. The cross-validation procedure, which is made to select the best hyperparameters, is used to enhance out-of-sample prediction. Mean absolute percentage error and root squared mean error serve as indicators of forecast accuracy. The findings of this research demonstrate that the forecasts produced by the gradient boosting model and random forest model are more accurate than the benchmark forecasts for the period of 2001–2018. The gradient boosting model proves to be more accurate when compared to the random forest model. The use of machine learning models in macroeconomic forecasting should be increased, according to this study.

The Impact of Trade and Expenditure on GDP

A study on Predicting Economic Recessions Using Machine Learning Algorithms was undertaken by Nyman and Ormerod in 2016. The third estimate of real GDP growth in the pertinent quarter serves as the dependent variable in the analysis. We mostly select explanatory variables from the financial markets. We examine the outcomes of two alternative estimating methods: random forest machine learning and ordinary least squares regression. To perform the random forest analysis, we download the random Forest package and utilize the statistical program R. The findings indicate that although there are little correlations between the actual and anticipated data, they do differ significantly from zero. We draw the conclusion that the algorithm never predicts a recession when one doesn't take place based on the results. In the instance of the United Kingdom, we get even better results using random forest machine learning approaches.

Cogoljevic et al., (2017) conducted a study on a machine-learning approach for predicting the relationship between energy resources and economic development. This study's objective is to create and implement a machine learning method for predicting gross domestic product (GDP) based on the mix of available energy sources. Our findings suggest that using a machine learning approach can somewhat increase the forecast accuracy of GDP. Extreme learning machines (ELM) and back propagation algorithms are the study's key inputs. This research note's major objective is to use machine learning to address the data's significant nonlinearity. We examine the connection between a nation's mix of energy resources and GDP-based economic expansion. As a result, we draw the conclusion that there is considerable interest in the relationship between the mix of energy resources and economic progress, especially given the circumstances of today.

A study on forecasting GDP growth using classical time series regression models and machine learning algorithms was conducted out by Premraj, (2019). This thesis compares predictions provided by ML algorithms and conventional forecasting methods to estimate GDP growth for 10 different economies. R is used for data preparation, data handling, and performance analysis. We employ the Bayesian Additive Trees Regression Trees (BART), Elastic-Net Regularized Generalized Linear Models (GLMNET), Stochastic Gradient Boosting (GBM), and eXtreme Gradient Boosting (XGBoost) in this study, as opposed to the traditional time series regression techniques of Autoregressive (AR) models, Autoregressive Integrated Moving Average (ARIMA) models, and Vector Autoregressive (VAR) models. Our goal is to determine whether using this technology produces better forecast performance than conventional time series techniques, as well as whether decision-makers in central banks and other pertinent institutions could be able to use it. Additionally, we wish to compare ML with economic theory to see if the ML approaches provide any additional predictors for GDP forecasting. The findings support the claim that multivariate VAR models are preferable, demonstrating the applicability of the

selected variables and models for forecasting GDP growth and the superiority of TS regression models over ML algorithms. As a result, we draw the conclusion that applying ML algorithms to forecasting macroeconomic variables is a young and developing topic that has already proven to be useful in the banking, healthcare, and retail sectors.

Maulud et al. (2020) performed a study to examine a Review of Linear Regression Comprehensive in Machine Learning. The optimal method to maximize prediction and precision is used in this paper to compare the performance of linear regression and polynomial regression. A statistical approach frequently used in research is regression modeling, particularly for observational studies. The findings show how models were estimated using data sets, their accuracy was assessed, and how crucial it is for a method to be able to predict outcomes. To assess how well a regression approach is working, a comparison between predicted and sample values was conducted. We conclude that a model's effectiveness must be associated with the actual values obtained for the explanatory variables.

Methodology

To explain the machine learning methods, first, describe the dataset using descriptive statistics and a graphical representation.

Data Description

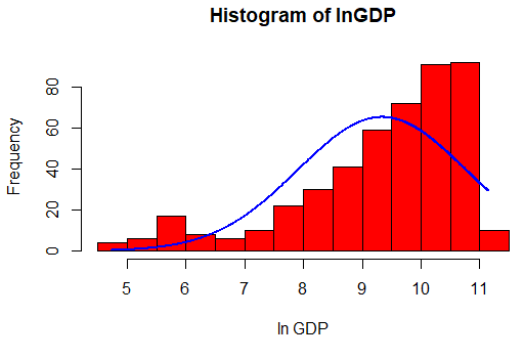
The data is taken from the world bank indicator (WBI) Website which consists of 12 variables and each variable has 468 observations. The data is taken from 1970 to 2021 for nine developed countries the United States, the United Kingdom, Italy, France, Portugal, Germany, Australia, New Zealand, and India. The variables are described is given below.

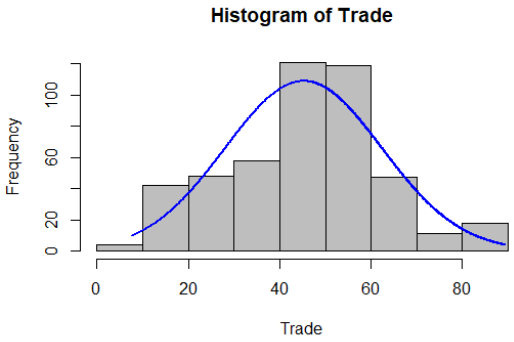
Variables	Description
GDP	Gross Domestic Product Per Capita
LnGDP	Log of Gross Domestic Product Per Capita
Trade	Trade
EXP	Final Consumption Expenditure
FDI	Foreign Direct Investment, Net Inflows
GCE	General Government Final Consumption Expenditure
GCF	Gross Capital Formation
IR	Real Interest Rate
EGS	Exports of Goods and Services

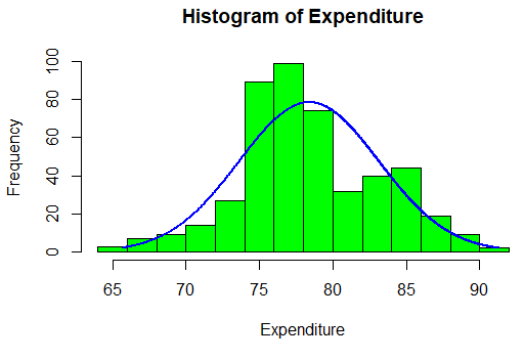
The Impact of Trade and Expenditure on GDP

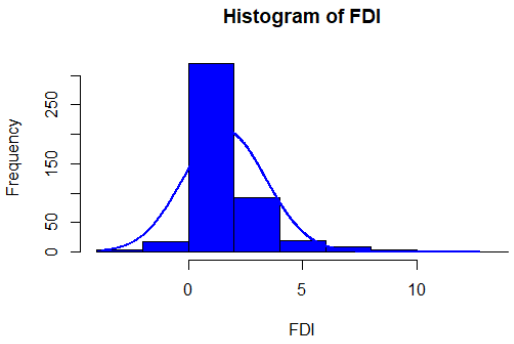
GDS	Gross Domestic Savings
NI	Net Income
TP	Population Total
TR	Total Revenue

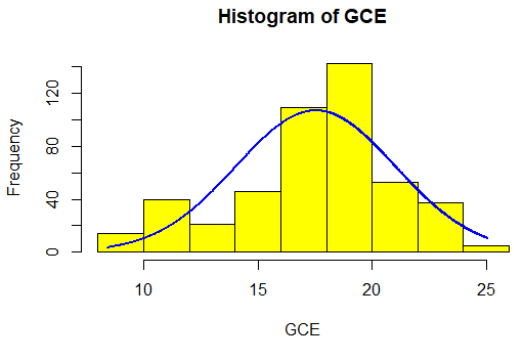
To check the normality of the data using histograms which are given bellows.

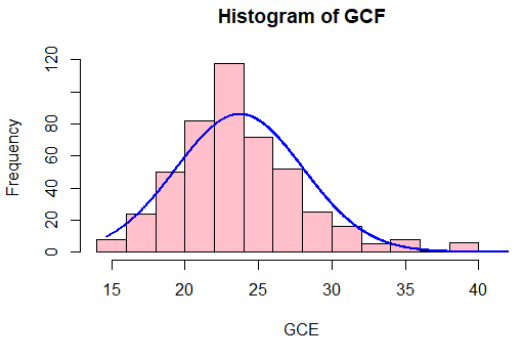


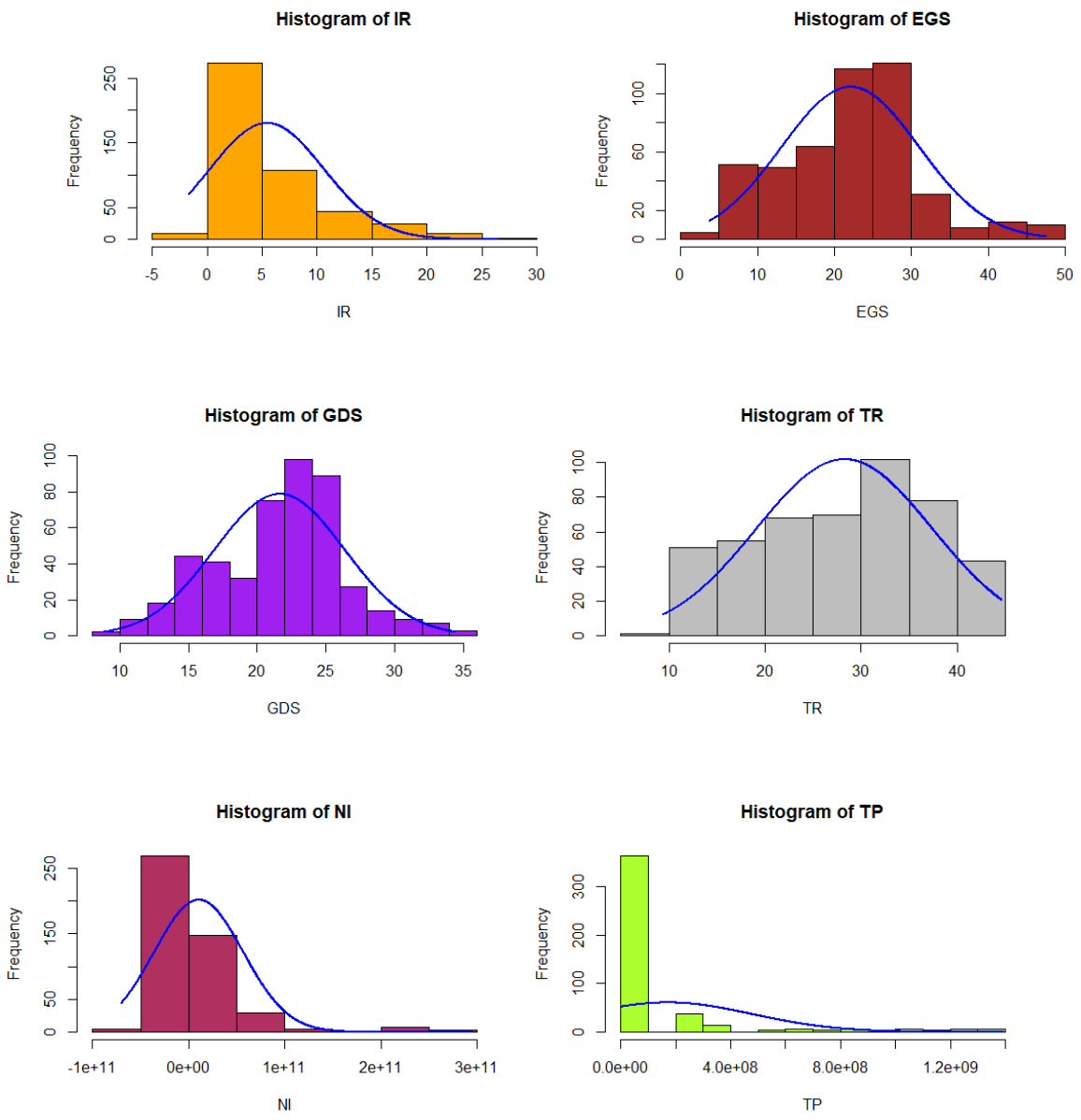












The histograms show that trade, expenditure, GCE, GCF, EGS, GDS, and TR are symmetrical which means that these variables follow a normal distribution. IR, NI, and TP show positive, log GDP is negatively skewed and FDI shows leptokurtic which means that these all variables do not follow a normal distribution. Now describe the main feature of these all variables by using descriptive statistics using R. The descriptive statistics in the table form are given below.

Statistic	***N***	***Mean***	***St. Deviation***	***Minimum***	***Maximum***
GDP	468	20685.17	16722.4	112.434	69287.54
lnGDP	468	45.228	1.433	4.722	11.146
Trade	468	45.228	17.163	7.662	89.386
EXP	468	78.378	4.762	65.623	91.269

FDI	468	1.59	1.806	-3.812	12.732
GCE	468	17.532	3.488	8.407	25.07
GCF	468	23.721	4.339	14.632	41.951
IR	468	5.439	5.181	-1.649	26.4
EGS	468	22.073	8.949	3.667	47.459
GDS	468	21.629	4.752	8.731	34.377
NI	468	10603001985	46366957088	-70503886262	2.9455E+11
TP	468	170273003	306908540	2810700	1393409033
TR	468	28.258	9.17	9.422	44.609

The average value of GDP is 20685.17 and standard deviation of GDP is 16722.4 which means that each observation of GDP has large deviation from their mean. The minimum and maximum values of GDP are 112.434 and 69287.54 respectively which means that GDP data consist of extreme values. The average value of log of GDP is 9.335 and standard deviation of log of GDP is 1.433 which means that each observation of log of GDP are very close with their mean. The minimum and maximum values of log of GDP are 4.722 and 11.146 respectively, which means that log of GDP data has not any extreme value. The average value of trade is 45.228 and standard deviation of trade is 17.163 which means that each observation of trade has large deviation from their mean. The minimum and maximum values of trade are 7.662 and 89.386 respectively which means that trade data consist of extreme values.

The average value of EXP is 78.378 and standard deviation of EXP is 4.762 which means that each observation of EXP is very close with their mean. The minimum and maximum values of EXP are 65.623 and 91.269 respectively, which means that EXP data has not any extreme value. The average value of FDI is 1.59 and standard deviation of FDI is 1.806 which means that each observation of FDI is very close to their mean. The minimum and maximum values of FDI are -3.812 and 12.732 respectively, which means that FDI data has not any extreme value. The average value of GCE is 17.532 and standard deviation of GCE is 3.488 which means that each observation of GCE is very close with their mean. The minimum and maximum values of GCE are 8.407 and 25.07 respectively, which means that GCE data has some extreme values.

The average value of GCF is 23.721 and standard deviation of GCF is 4.339 which means that each observation of GCF is very close with their mean. The minimum and maximum values of GCF are 14.632 and 41.951 respectively, which means that GCF data has some extreme values.

The average value of IR is 5.439 and standard deviation of IR is 5.181 which means that each observation of IR is very close with their mean. The minimum and maximum values of IR are -1.649 and 26.4 respectively, which means that IR data has some extreme values. The average value of EGS is 22.073 and standard deviation of EGS is 8.949 which means that each observation of EGS is very close with their mean. The minimum and maximum values of EGS are 3.667 and 47.459 respectively, which means that EGS data has some extreme values.

The average value of GDS is 21.629 and standard deviation of GDS is 4.752 which means that each observation of GDS is very close with their mean. The minimum and maximum values of GDS are 8.731 and 34.377 respectively, which means that GDS data has some extreme values. The average value of NI is 10603001985 and the standard deviation of NI is 46366957088 which means that each observation of NI has a large deviation from their mean. The minimum and maximum values of NI are -70503886262 and 294552000000 respectively which means that NI data consist of extreme values. The average value of TP is 170273003 and the standard deviation of TP is 306908540 which means that each observation of TP has a large deviation from their mean. The minimum and maximum values of TP are 2810700 and 1393409033 respectively which means that TP data consist of extreme values. The average value of TR is 28.258 and the standard deviation of TR is 9.17 which means that each observation of TR is very close to their mean. The minimum and maximum values of TR are 9.422 and 44.609 respectively, which means that TR data has some extreme values.

Model's Specification

Machine learning is an application of artificial intelligence (AI) that allows systems to learn and evolve without having to be explicitly programmed. The goal of machine learning is to create computer systems that can access and utilize data on their own. This phase of learning starts with observations or data, such as examples, direct experience, or training, to look for patterns in the data and form better future judgments based on the examples we provide. The major goal is to allow computers to learn on their own and change assignments without human intervention.

Algorithms for machine learning are typically considered to be supervised, unsupervised and reinforcement machine learning.

Supervised Machine Learning

Using labeled examples to apply what has been learned to new data in the past, supervised machine learning algorithms can predict future episodes. By studying a well-known training data set, the learning process generates an inferred function to predict the output values. After proper training, the system may set goals for any new input. The learning algorithm can also compare

its findings to the correct intended output to make necessary changes to the model. In machine learning, the most popular model has supervised learning. It is simple to understand and implement. A young person's education is comparable to flashcards.

Based on information in the label form, these label pairs may be added one by one to a learning algorithm that enables the system to forecast each label and provide comments on whether a correct answer has been expected. With time, the algorithm learns the exact character of the connection between samples and their brands. Thus, a new and unheard example can be seen, and a good title forecast if the study's algorithm is trained correctly.

Random Forest

Random Forest is a commonly utilized learning algorithm in controlled learning technology. In ML, both classification and regression issues can be used. The idea of ensemble learning combines several classifications to resolve a complex problem and improve model performance. As its name states: "Random Forest is a classification that uses the average of the data set to increase predictability by including numerous decision-making bodies in diverse subsets of the given dataset. A random forest, rather than a single decision tree, accumulates forecasts from each tree and is based on majority voting predicting and final output. The more trees in the forest, the better the accuracy and the less crowded.

Regression

Regression analysis is a method of statistical modeling to interact the dependent (goal) and independent (predictor) components with one or more independent variables. Regression analyses, in particular, help us understand how, if other separate variables are maintained, the value of the response variable transforms into a regressor. It forecasts current and ongoing statistics such as temperature, age, revenue, and price. In this approach, we have no variable predicting/estimating target/result. It is used in groups for the grouping population that individual customers often operate in different groups for segmentation or particular intervention.

Linear Regression

Linear regression is one of the most popular and well-known machine learning techniques. Predictive testing is carried out using a statistical technique. Sales, wages, age, commodity prices, and other continuous/real or quantitative parameters are forecasted using linear returns. A linear regression technique shows a linear link and hence a linear regression between one or more separate variables (y). Because of linear regression, it detects how the value of the variable depends on the independent variable value.

Polynomial Regression

Polynomial Regression is a technique of regression that models as nth grade polynomial the connection between a dependent(s) and an independent variable(s). The following is the equation of polynomial regression:

$$y= b_0+b_1x_1+ b_2x_1^2+ b_2x_1^3+..... b_nx_1^n$$

The unique condition of multiple linear regression in ML is also mentioned. Because the multiple linear regression equation is transformed into polynomial regression by adding additional polynomial terms. It is a linear model with several improvements to accuracy. In polynomial regression, the dates of training are nonlinear. For complicated and non-linear data and functions, a linear regressive model is used. "The original characteristics are therefore converted in a Polynomial Regression into a Polynomial characteristic of the required degree (2, 3....n) and modeled on a linear model".

Regularization

Regularization is one of the core principles of machine learning. It provides extra information and is a method to prevent overfitting. The machine learning model sometimes is good for the training data, but not suitable for the test data. This shows that the model cannot forecast the result by adding noise into the output while dealing with unseen data; hence the model is referred to as overfitting. This difficulty can be addressed using a regularization technique. This procedure can be employed so that the magnitude of the variables may be reduced to all variables or features of the model. The model is therefore accurate and general. The coefficient of characteristics is largely regularized or reduced to zero. Simply said, "We lower the size of the features by maintaining the same number of features in the regulating approach."

Ridge Regression

Ridge regression is one of the linear regression forms with low prejudice so that we can predict the longer term better. Ridge regression is a regularization approach that minimizes the complexity of the model. Sometimes L2 is called regularization. By adding the penalty time in this technique, the cost function is changed. The quantity of information provided to the model is a penalty of Ridge Regression. We may calculate this by raising the lambda to the square weight of each feature.

Lasso Regression

The Lasso regression is another way to regulate the models to reduce complexity. It is the operator with the smallest absolute value as well as the selection operator. The punishment period only

includes absolute weights instead of a plot. The same goes with Ridge regression. This is true. It therefore may decrease to 0 as it demands absolute quantities, but Ridge regression may decrease only near to 0.

Result and Discussion

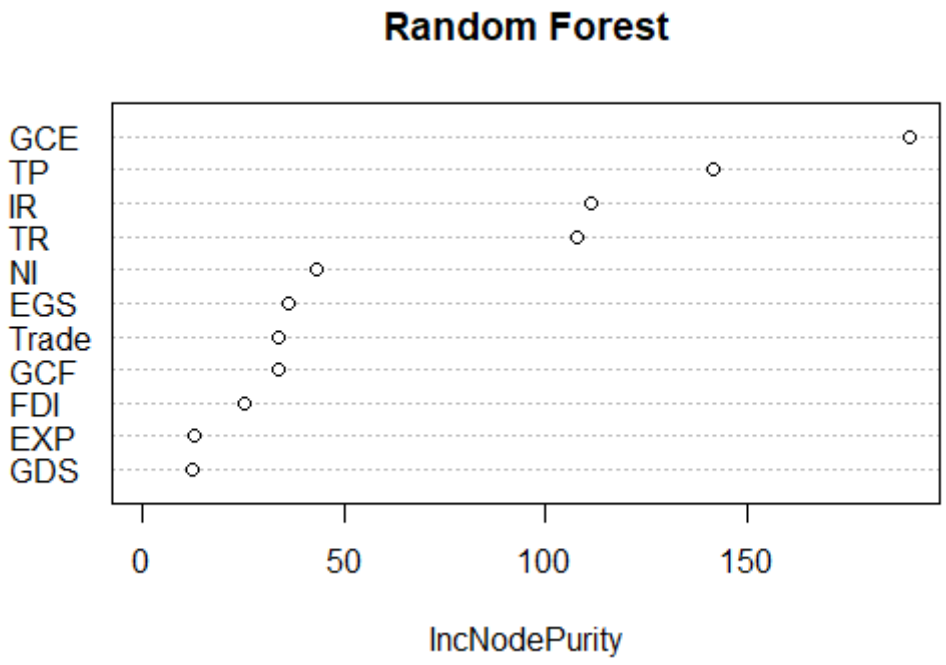
Machine Learning Models

Five machine-learning regression models are estimated to test the effect of expenditure and trade on Economic growth (GDP per capita), including the Multiple Linear Regression Model, Random Forest, Lasso Regression, Ridge Regression, and Elastic Net Regression. The result of R is given below.

Random Forest Regression Model

The output and plot of the random forest machine learning method for a regression problem is given below.

Number of trees	500
No. of variables tried at each split	3
Mean of squared residuals	0.07426909
% Var explained	96.28



According to the random forest result, there will be created 500 trees and 3 variables tried at each split. The random forest plot indicated that the GCE variable has the most important rule in the GDP prediction, the second important variable is TP, 3rd one is IR, then TR, NI, EGS, Trade, GCF, and FDI. The two less important variables in the GDP prediction are EXP and GDS.

Multiple Linear Regression Model

Estimate the multiple linear regression model by using the ordinary least square estimation technique in which log GDP per capita is used as a dependent, Final Consumption Expenditure and Trade are used as explanatory variables, Foreign Direct Investment, Net Inflows, General Government Final Consumption Expenditure, Gross Capital Formation, Real Interest Rate, Exports of Goods and Services, Gross Domestic Savings, Net Income, Population Total, Total Revenue are used control variables. The R output of the multiple linear machine learning regression model is given below.

R^2	0.804	Residual Std. Error (367)	0.637
Adjusted R^2	0.798	*F (11,367)*	136.461***

Dependent Variable: ***GDP***

Variables	*Estimate*	*Standard Error*	*t-value*	*P-value*
Constant	72.529	19.798	3.663	0.000
Trade	0.152	0.018	8.444	0.000
EXP	-0.650	0.198	-3.283	0.000
FDI	0.093	0.021	4,429	0.000
GCE	0.174	0.018	9.667	0.000
GCF	-0.157	0.018	8.722	0.000
IR	-0.061	0.008	7.625	0.000
EGS	-0.287	0.035	8.2	0.000
GDS	-0.512	0.198	2.586	0.007
NI	0.0007	0.000	0.000	0.000
TP	-0.0004	-0.000	-0.000	0.000
TR	-0.023	0.008	2.875	0.000

As one unit increases in Trade, ln GDP also increases by 15.2% which means that there has a positive effect of Trade on GDP. The P-value of the Trade estimate is less than the critical value

0.05, which means that there has a statistically significant effect of Trade on GDP. As one unit increases in EXP, ln GDP decreases by 65% which means that there has a negative effect of EXP on GDP. The P-value of the EXP estimate is less than the critical value 0.05, which means that there has a statistically significant effect of EXP on GDP. As one unit increases in FDI, ln GDP also increases by 9.3% which means that there has a positive effect of FDI on GDP. The P-value of the FDI estimate is less than the critical value 0.05, which means that there has a statistically significant effect of FDI on GDP.

When one unit increases in GCE, ln GDP also increases by 17.4% which means that there has a positive effect of GCE on GDP. The P-value of the GCE estimate is less than the critical value 0.05, which means that there has a statistically significant effect of GCE on GDP. When one unit increases in GCF, ln GDP decreases by 15.7% which means that there has a negative effect of GCF on GDP. The P-value of the GCF estimate is less than the critical value 0.05, which means that there has a statistically significant effect of GCF on GDP. As one unit increases in IR, ln GDP decreases by 6.1% which means that there has a negative effect of IR on GDP. The P-value of the IR estimate is less than the critical value 0.05, which means that there has a statistically significant effect of IR on GDP.

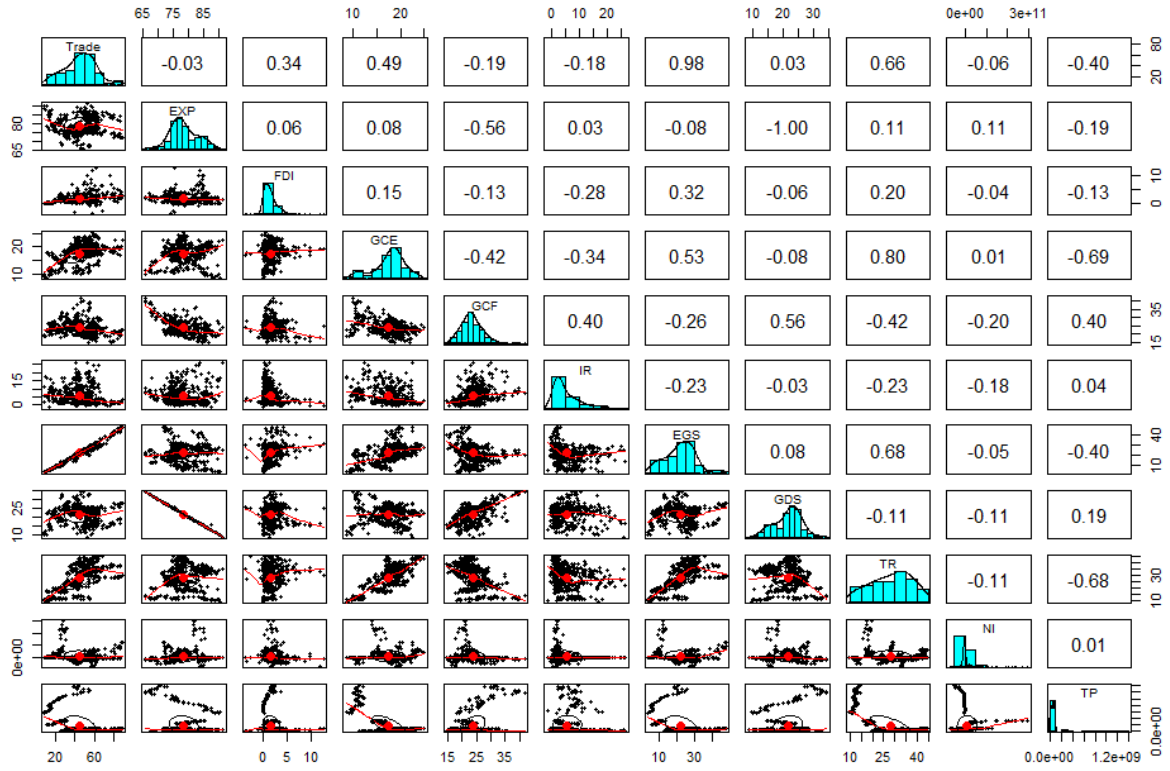
As one unit increases in EGS, ln GDP decreases by 28.7% which means that there has a negative effect of EGS on GDP. The P-value of the EGS estimate is less than the critical value 0.05, which means that there has a statistically significant effect of EGS on GDP. As one unit increases in GDS, ln GDP decreases by 51.2% which means that there has a negative effect of GDS on GDP. The P-value of the GDS estimate is less than the critical value 0.05, which means that there has a statistically significant effect of GDS on GDP.

When one unit increases in NI, ln GDP also increases by 0.0000005% which means that there has a positive effect of NI on GDP. The P-value of the NI estimate is less than the critical value 0.05, which means that there has a statistically significant effect of NI on GDP. When one unit increases in GCE, ln GDP decreases by 0.0000001% which means that there has a negative effect of GCE on GDP. The P-value of the GCE estimate is less than the critical value 0.05, which means that there has a statistically significant effect of GCE on GDP. As one unit increases in TR, ln GDP decreases by 2.3% which means that there has a negative effect of TR on GDP. The P-value of the TR estimate is less than the critical value 0.05, which means that there has a statistically significant effect of TR on GDP.

The P-value of the F-test is less than the critical value 0.05 which means that the overall model is statistically significant. The coefficient of determination (R^2) value is 0.80, which means that

80% of the variation in GDP is explained by the variation in Trade, Expenditure, and all other control variables, so we conclude that the overall model is well fitted for future prediction.

Now check the multicollinearity problem in the independent variables by using a correlation matrix with their graphical representations. If any one of them will highly be correlated with another one so it will create the problem of overfitting, so for this purpose, we will need to go regularization which includes Ridge, Lasso, and Elastic Net Regression Models. The correlation matrix with their plots is given bellows.



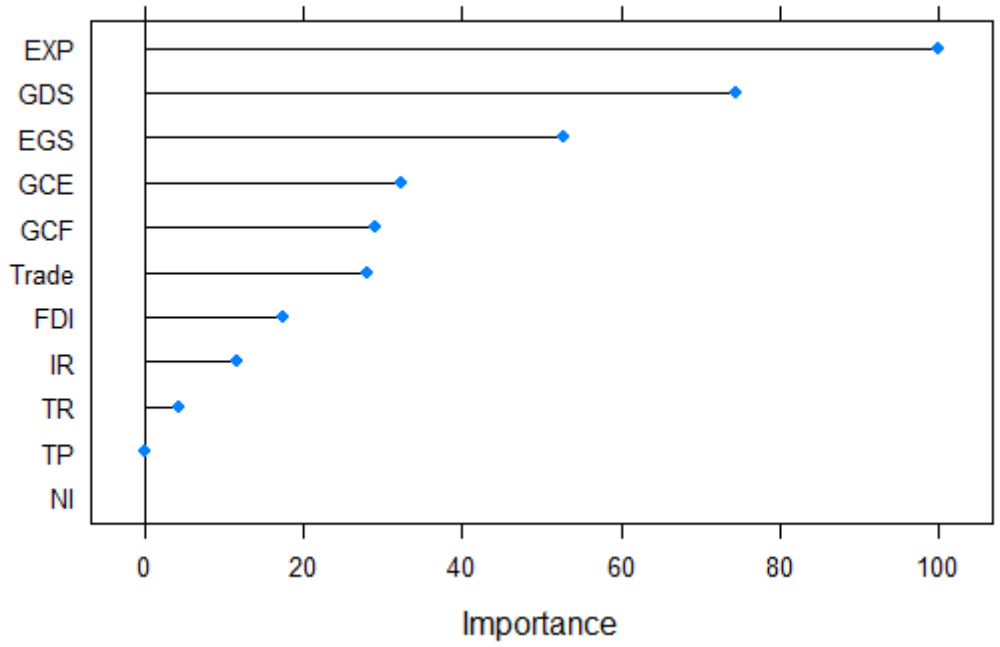
The scatter plot shows that there is a very high positive correlation between Trade and EGS, TR, and GCE, and a very high negative correlation between GDS and EXP, which means that there has a problem of multicollinearity in the model and the model is overfitted. Due to the multicollinearity problem, the coefficient of the linear model remains unbiased but the variances of the estimator are very large which means that the estimators are less efficient, so we need to go to regularization methods which are given bellows.

Lasso Regression Model

To shrink the coefficients of the model for making it efficient, use the Ridge, Lasso, and Net Elastic Regression Model.

Less significant characteristics in a dataset are removed when penalizing because coefficients would shrink towards a mean of zero. As input variables are effectively eliminated, the shrinkage of these coefficients based on the tuning alpha value results in some sort of automatic feature selection.

lambda	RMSE	Rsquared	MAE
0.0001	0.651855	0.789568	0.535163
0.250075	0.864512	0.68698	0.65009
0.50005	1.075717	0.589476	0.830099
0.750025	1.252466	0.499793	0.984752
1	1.40366	0.349075	1.0861
The tuning parameter 'alpha' was held constant at a value of 0			
RMSE was used to select the optimal model using the smallest value.			

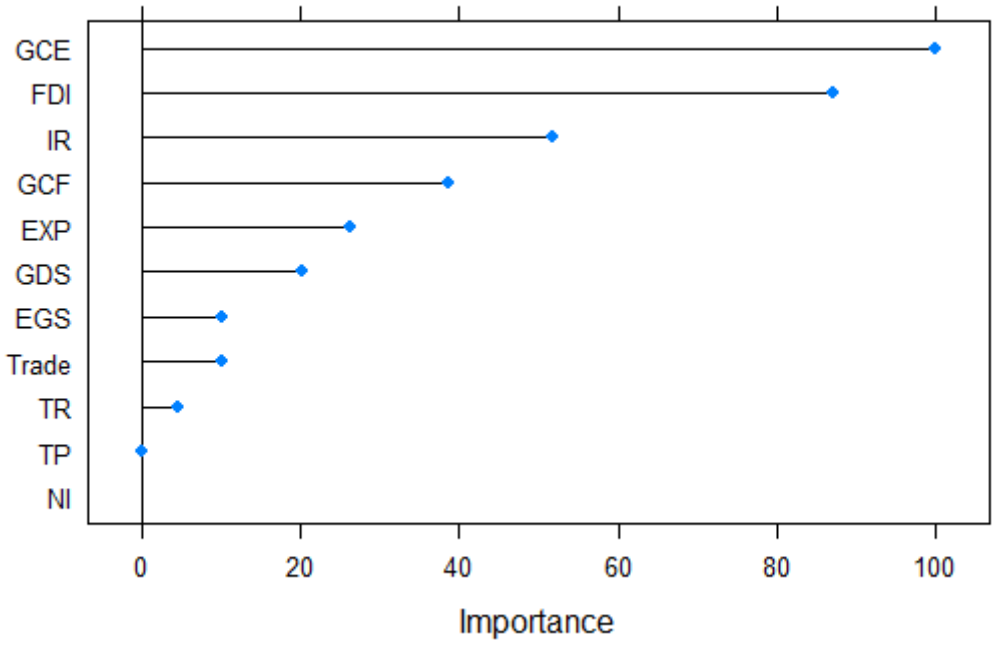


According to the lasso regression model, EXP is the most important factor in the model, the second important factor is GDS, the third one is EGS, the fourth one is GCE, the fifth one is GCF, the sixth and seventh one is Trade, and FDI. The less important variables in the model are IR, TR, TP, and NI. According to lasso regression results when we remove IR, TR, TP, and NI so we will get efficient estimators of the model.

Ridge Regression Model

By incorporating a penalty component, ridge regression constrains the coefficients in a manner similar to lasso regression. Ridge regression, on the other hand, uses the square of the coefficients whereas lasso regression uses their magnitude.

lambda	RMSE	R squared	MAE
0.0001	0.698775	0.761664	0.562028
0.250075	0.712343	0.754986	0.567709
0.50005	0.757057	0.746678	0.575604
0.750025	0.757057	0.739049	0.583005
1	0.779446	0.731965	0.592086
The tuning parameter 'alpha' was held constant at a value of 0			
RMSE was used to select the optimal model using the smallest value.			



According to the ridge regression model, GCE is the most important factor in the model, the second important factor is FDI, the third one is IR, the fourth one is GCF, the fifth one is EXP, and the sixth one is GDS. The less important variables in the model are EGS, Trade, TR, TP, and NI. According to ridge regression results when we remove EGS, Trade, TR, TP, and NI so we will get efficient estimators of the model.

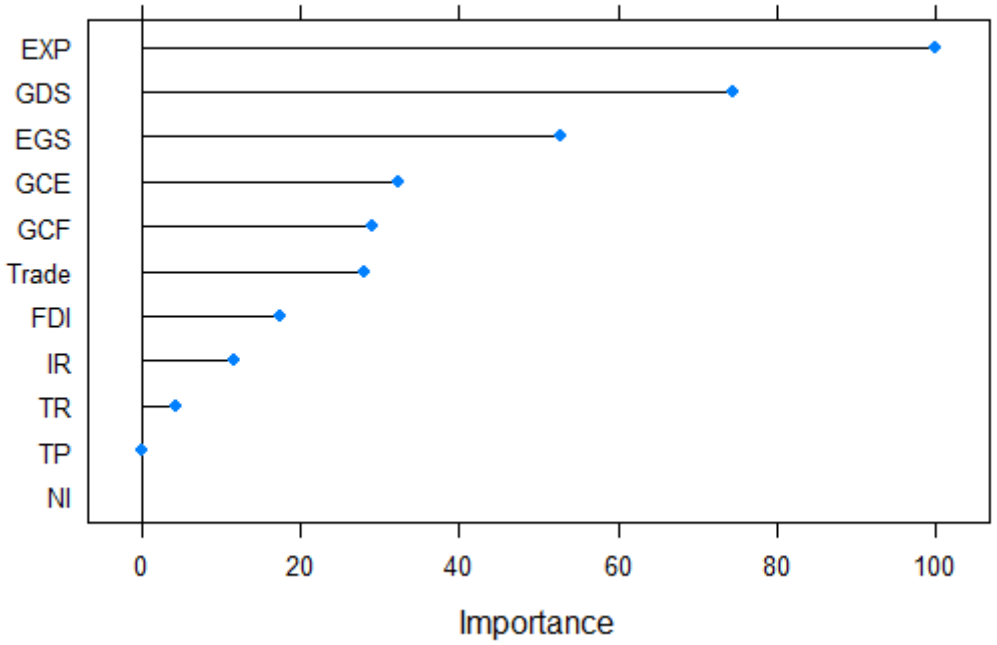
The Impact of Trade and Expenditure on GDP

Net Elastic Regression Model

After lasso was criticized for having a changeable selection that could be overly dependent on data and unstable, elastic net first came to be. To achieve the best of both worlds, the approach is to mix the penalties of ridge regression and lasso.

alpha	lambda	RMSE	Rsquared	MAE
0	0.0001	0.697309	0.761469	0.561394
0	0.250075	0.710811	0.755226	0.56653
0	0.50005	0.732731	0.747382	0.574217
0	0.750025	0.755556	0.740095	0.581791
0	1	0.777973	0.733293	0.590573
0.111111	0.0001	0.648483	0.790809	0.531681
0.111111	0.250075	0.718091	0.754314	0.568603
0.111111	0.50005	0.756978	0.743244	0.580947
0.111111	0.750025	0.800882	0.730417	0.60052
0.111111	1	0.845707	0.715936	0.627533
0.222222	0.0001	0.648168	0.790956	0.531832
0.222222	0.250075	0.729713	0.751287	0.572326
0.222222	0.50005	0.794318	0.730516	0.600553
0.222222	0.750025	0.857958	0.712554	0.636968
0.222222	1	0.919536	0.69743	0.68467
0.333333	0.0001	0.648116	0.790985	0.532009
0.333333	0.250075	0.746522	0.744743	0.580329
0.333333	0.50005	0.830671	0.718891	0.621246
0.333333	0.750025	0.917177	0.694545	0.683626
0.333333	1	1.008378	0.656129	0.759427
0.444444	0.0001	0.648119	0.790987	0.532123
0.444444	0.250075	0.768168	0.734139	0.59411
0.444444	0.50005	0.867894	0.707296	0.647008
0.444444	0.750025	0.985318	0.658426	0.740527
0.444444	1	1.078847	0.635617	0.827975
0.555556	0.0001	0.64803	0.791062	0.53216
0.555556	0.250075	0.785504	0.727073	0.604497

0.555556	0.50005	0.91246	0.68443	0.681203
0.555556	0.750025	1.037831	0.639132	0.790143
0.555556	1	1.15314	0.599118	0.898509
0.666667	0.0001	0.648074	0.79104	0.532223
0.666667	0.250075	0.801074	0.72186	0.61247
0.666667	0.50005	0.95698	0.657001	0.716736
0.666667	0.750025	1.093873	0.614169	0.844951
0.666667	1	1.234021	0.509691	0.97185
0.777778	0.0001	0.648006	0.791094	0.532235
0.777778	0.250075	0.819166	0.714553	0.622732
0.777778	0.50005	0.994121	0.640079	0.74995
0.777778	0.750025	1.159228	0.556519	0.907612
0.777778	1	1.285203	0.495139	1.008581
0.888889	0.0001	0.648054	0.791067	0.532284
0.888889	0.250075	0.839485	0.704835	0.634752
0.888889	0.50005	1.031712	0.622545	0.78753
0.888889	0.750025	1.213513	0.496746	0.957705
0.888889	1	1.342592	0.495139	1.047049
1	0.0001	0.647964	0.791131	0.532264
1	0.250075	0.861197	0.692958	0.647714
1	0.50005	1.075342	0.589614	0.830347
1	0.750025	1.251349	0.495139	0.98506
1	1	1.40208	0.385497	1.086394
RMSE was used to select the optimal model using the smallest value.				
The final values used for the model were alpha = 1 and lambda = 1e-04.				



According to the net elastic regression model, EXP is the most important factor in the model, the second important factor is GDS, the third one is EGS, the fourth one is GCE, the fifth one is GCF, the sixth and seventh one is Trade, and FDI. The less important variables in the model are IR, TR, TP, and NI. According to net elastic and lasso, both regressions result in we conclusion that when we remove IR, TR, TP, and NI so we will get efficient estimators of the model.

Comparison

To compare these five machine learning models on the basis of Mean Absolute Error, Root Mean Square Error, and Coefficient of Determination (R^2). The values of MAE, RMSE, and, R^2 of each model is given below in the table form.

Model	*MAE*	*RMSE*	*R^2*
Random Forest	0.187	0.236	0.960
Linear Regression	0.514	0.687	0.804
Lasso Regression	0.535	0.653	0.790
Ridge Regression	0.562	0.699	0.762
Net Elastic Regression	0.532	0.648	0.791

According to the MAE, RMSE, and $R^{2,}$ the random forest model is one of the best regression models as compared to the other 4 models due to minimum MAE, RMSE, and maximum R^2. If the model has a minimum value of mean absolute error, or minimum value of root mean square

error than the other models then the model will be the best model the other models or if the model has a maximum coefficient of determination (R^2) value then the model is well fitted than the other models for future prediction. The Random Forest regression model has minimum values of mean absolute error, root mean square error, and maximum value of the coefficient of determination (R^2), which concludes that the Random Forest Regression Model is a well-fitted model for future prediction than Linear Regression, Lasso, Ridge, and Net Elastic Regression Models.

Conclusion

According to these best regression models which are Random Forest, and Linear Regression Models we get some conclusions about the effects of trade, expenditure, and other control variables which includes Foreign Direct Investment, Net Inflows, General Government Final Consumption Expenditure, Gross Capital Formation, Real Interest Rate, Exports of Goods and Services, Gross Domestic, Savings Net Income, Population Total, and Total Revenue. According to the best model which is called Random Forest Model, we conclude the General Government Final Consumption Expenditure variable has the most important rule in GDP prediction, the second important variable is the Total number of People, 3rd one is the Real Interest Rate, then Total Revenue, Net Income, Exports of Goods and Services, Trade, Gross Capital Formation, and Foreign Direct Investment. The two less important variables in the GDP prediction are Final Consumption Expenditure and Gross Domestic Savings.

The main aim of this study is to check the effect of Trade, and Expenditure on the Gross Domestic Product (GDP) per capita, so the linear regression model is one of the best models for checking this effect, the Linear Regression Model we conclude that there has a positive effect of Trade on GDP per capita, and there is a negative effect of Expenditure on GDP per capita. Linear Model also concludes that the control variables Foreign Direct Investment, General Government Final Consumption, and Net Income on GDP, and there is a negative effect of Gross Capital Formations, Interest Rate, Exports of Goods and Services, Gross Domestic Savings, General Government Final Consumption, and Total Revenue on GDP and all explanatory and control variables have a significant effect on GDP per capita.

For future study, we can suggest that there are a lot of explanatory variables which has a statistically significant effect on GDP per capita but are not included in the model. We just estimate the regression models to check the effect of trade, and expenditure on GDP in this study but did not predict that for the future so we suggest that it can be predicted in the future study.

Appendices

R Codes

```
## Data Import

data=read.csv(file.choose(),sep=",",header=T)

fix(data)


## Install Packages

install.packages("rcompanion")

library(rcompanion)

install.packages("stargazer")

library(stargazer)

install.packages("randomForest")

library(randomForest)

install.packages("caret")

library(caret)

install.packages("ggplot2")

library(ggplot2)

install.packages("caret")

library(caret)

install.packages("glmnet")

library(glmnet)

install.packages("psych")

library(psych)


## Graphical Representation
```

```
plotNormalHistogram(data$lnGDP, prob = FALSE,
                    main = "Histogram of lnGDP",
                    length = 1000,col='red',xlab='ln GDP' )


plotNormalHistogram(data$Trade, prob = FALSE,
                    main = "Histogram of Trade",
                    length = 1000,xlab='Trade' )

plotNormalHistogram(data$EXP, prob = FALSE,
                    main = "Histogram of Expenditure",
                    length = 1000,col='green',xlab='Expenditure' )

plotNormalHistogram(data$FDI, prob = FALSE,
                    main = "Histogram of FDI",
                    length = 1000,col='blue',xlab='FDI' )

plotNormalHistogram(data$GCE, prob = FALSE,
                    main = "Histogram of GCE",
                    length = 1000,col='yellow',xlab='GCE' )

plotNormalHistogram(data$GCF, prob = FALSE,
                    main = "Histogram of GCF",
```

```
        length = 1000,col='pink',xlab='GCE' )


plotNormalHistogram(data$IR, prob = FALSE,
        main = "Histogram of IR",
        length = 1000,col='orange',xlab='IR' )


plotNormalHistogram(data$EGS, prob = FALSE,
        main = "Histogram of EGS",
        length = 1000,col='brown',xlab='EGS' )


plotNormalHistogram(data$GDS, prob = FALSE,
        main = "Histogram of GDS",
        length = 1000,col='purple',xlab='GDS' )


plotNormalHistogram(data$TR, prob = FALSE,
        main = "Histogram of TR",
        length = 1000,col='gray',xlab='TR' )


plotNormalHistogram(data$NI, prob = FALSE,
        main = "Histogram of NI",
        length = 1000,col='maroon',xlab='NI' )


plotNormalHistogram(data$TP, prob = FALSE,
        main = "Histogram of TP",
        length = 1000,col='greenyellow',xlab='TP' )
```

```
## Descriptive Statistics

summary(data)

stargazer(data,type='text',title = "Descriptive Statistics")


## Data Partition

set.seed(1234)

p1=sample(2,nrow(data),replace=TRUE,prob=c(0.8,0.2))

train=data[p1==1,]

test=data[p1==2,]


## Multiple Regression

model1=lm(lnGDP~Trade+EXP+FDI+GCE+GCF+IR+EGS+GDS+NI+TP+TR,data = train)

summary(model1)

stargazer(model1,type = 'text',title = 'Multiple Linear Regression')


MAE1=MAE(train$lnGDP,predict(model1))

MAE1

RMSE1=sqrt(mean((test$lnGDP-predict(model1,test))^2))

RMSE1


## Random Forest

model2=randomForest(lnGDP~Trade+EXP+FDI+GCE+GCF+IR+EGS+GDS+NI+TP+TR,data=train)

print(model2)
```

```
varImpPlot(model2,main='Random Forest')

MAE2=MAE(train$lnGDP,predict(model2))

MAE2

RMSE2=sqrt(mean((test$lnGDP-predict(model2,test))^2))

RMSE2

r2=model2$rsq

mean(r2)


#prediction

prediction1=predict(model2,train)

head(prediction1)

prediction2=predict(model2,test)

head(prediction2)


## Ridge, Lasso and Net Elastic Regression

pairs.panels(data[c(-1,-2)],cex=2)


custom=trainControl(method="repeatedcv",number=10,repeats=5,verboseIter = T)


set.seed(1234)

lm=train(lnGDP~Trade+EXP+FDI+GCE+GCF+IR+EGS+GDS+NI+TP+TR,

      train,method='lm',

      trControl=custom)

lm$results

summary(lm)
```

```
## Ridge Regression

set.seed(1234)

Ridge=train(lnGDP~Trade+EXP+FDI+GCE+GCF+IR+EGS+GDS+NI+TP+TR,

        train,method='glmnet',tuneGrid=expand.grid(alpha=0,

                                        lambda=seq(0.0001,1,length=5)),

        trControl=custom)

print(Ridge)


plot(varImp(Ridge,scale = T))


## Lasso Regression

Lasso=train(lnGDP~Trade+EXP+FDI+GCE+GCF+IR+EGS+GDS+NI+TP+TR,

        train,method='glmnet',tuneGrid=expand.grid(alpha=1,

                                        lambda=seq(0.0001,1,length=5)),

        trControl=custom)

print(Lasso)

plot(varImp(Lasso,scale = T))




## Elastic Net Regression

Elastic=train(lnGDP~Trade+EXP+FDI+GCE+GCF+IR+EGS+GDS+NI+TP+TR,

        train,method='glmnet',tuneGrid=expand.grid(alpha=seq(0,1,length=10),

                                        lambda=seq(0.0001,1,length=5)),

        trControl=custom)
```

```
print(Elastic)

plot(varImp(Elastic,scale = T))



## Compare the Models

models=list(LinearModel=lm,Ridge=Ridge,Lasso=Lasso,ElasticNet=Elastic)

res=resamples(models)

summary(res)
```

References

Yoon, J., 2021. Forecasting of real GDP growth using machine learning models: Gradient boosting and random forest approach. *Computational Economics*, *57*(1), pp.247-265.

Maulud, D. and Abdulazeez, A.M., 2020. A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, *1*(4), pp.140-147.

Cogoljević, D., Alizamir, M., Piljan, I., Piljan, T., Prljić, K. and Zimonjić, S., 2018. A machine learning approach for predicting the relationship between energy resources and economic development. *Physica A: Statistical Mechanics and its Applications*, *495*, pp.211-214.

Nyman, R. and Ormerod, P., 2017. Predicting economic recessions using machine learning algorithms. *arXiv preprint arXiv:1701.01428*.

Paruchuri, H., 2021. Conceptualization of machine learning in economic forecasting. *Asian Business Review*, *11*(2), pp.51-58.

Singh, A., Thakur, N., & Sharma, A. (2016, March). A review of supervised machine learning algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 1310-1315). Ieee.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*(1), 3-24.