Intelligent Systems with Applications 14 (2022) 200082

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.elsevier.com/locate/iswa

Predicting gross domestic product to macroeconomic indicators

S.C. Agu[a,*], F.U. Onu[b], U.K. Ezemagu[c], D. Oden[a]

[a] Department of Computer Science, Madonna University Nigeria, Elele Campus
[b] Department of Computer Science, Ebonyi State University Abakiliki, Nigeria
[c] Department of Anatomy, Alex Ekwueme Federal University Ndufu Alike, Ikwo, Ebonyi State, Nigeria

ARTICLE INFO

Article history:
Received 31 December 2021
Revised 31 March 2022
Accepted 20 April 2022
Available online 1 May 2022

Keywords:
Machine learning
Regression techniques
Coefficient estimates
Macroeconomic indicators
Gross domestic product

ABSTRACT

Macroeconomic indicators enable countries to concentrate on goods, services, and other entities that grow their Gross Domestic Product (GDP). Often, identifying these groups of indicators poses a challenge to nations. The study considered a typical data set with two main objectives. First, to predict GDP to macroeconomic indicators by applying four machine learning methods namely, Principal Component Regression (PCR), Ridge Regression (RR), Lasso Regression (LR), and Ordinary Least Squares (OLS). Second, identify the most likely key macroeconomic variables that could affect the growth of GDP. The methods were evaluated using 5-fold cross-validation, and the estimated coefficients associated with the macroeconomic indicators were computed. The results revealed that PCR method with an accuracy of 89% and a mean square error of -7.552007365635066e+21 predicted GDP to macroeconomic indicators accurately, more than other methods. Some macroeconomic indicators did affect GDP positively, while others did not. The major contribution of the study is the use of machine learning regularization methods to predict GDP instead of the traditional statistical methods. It also identified additional macroeconomic variables to compute 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, export and import rates (WikiBooks, 2021).

Purchase Power Parities (PPP) is a measure of what an economy's local currency can buy in another economy (World Bank, 2020). The ranking of PPP GDP per capita over World PPP GDP per capita reveals that only 76 out of 190 countries crossed the benchmark of $17,100. Interestingly, it excluded the big names as China, Iraq, South Africa, and India (Worldometer, 2017). Unfortunately, for past decades, some countries' economies had been unstable and several programs have been embarked upon to improve the economy, but without desired success. For instance, Nigeria has implemented Structural Adjustment Programs (SAP), National Poverty Eradication Program (NAPEP), and National Economic Empowerment and Development Strategies (NEEDS) in a bid to grow her economy but all to no avail (Oden, Ibeto & Agu, 2020). Instead, the GDP of Nigeria between 2008 and 2015 declined from 6.9% - 1.58% and the average growth rate by 2021 was 0.31% (Ukpe, 2021).

Notably, Picardo (2021) suggested that GDP is a strong indicator of the growth of a nation's economy. Hence, countries try to maximize GDP when making fiscal planning to achieve a high level of economic growth (Divya & Rama, 2014). In the cause of this study, we observed that many researchers and economists use traditional statistical Ordinary Least Squares (OLS), Time Series (TS), as well as other statistical approaches to explore the phenomenon of predicting GDP via several macroeconomic indicators. The result of changes in GDP of China-specific macroeconomic TS data, and a large number of predictor variables which were predicted using Principal Component Analysis (PCA) and applying Dynamic Factor Model (DFM), revealed a significant improvement in degree of prediction. It also performed better than univariate auto regression analysis (Asger & Miha, 2020). However, the recent work of some authors (Giovanni, Giacomo & Sara, 2021), revealed that machine learning algorithm; K-Nearest Neighbour (KNN) model performs better than traditional TS analysis to self-predict U.S. GDP. Furthermore, it has been established that KNN does not work well

* Corresponding author.
E-mail address: sndyaguu@gmail.com (S.C. Agu).

https://doi.org/10.1016/j.iswa.2022.200082

with high dimensional and large datasets (Genesis, 2018; Naresh, 2020). Hence, the need for machine learning regularization methods which this study attempt to consider.

Recently, Shaobo (2021) combined back propagation (BP) neural network and the ARIMA model to predict the nonlinear residual of GDP. He added the predicted values of the two models to obtain a model which could predict the daily price of nonlinear residual of GDP. However, it was a good work, but anchored on PPP-based household consumption per capita instead of PPP GDP per capita (World Bank, 2020). Authors (Divya & Rama, 2014) predicted India GDP through macroeconomic indicators, and Oden et al. (2020) and Yua, Adoms, Okaro and Ogbonna (2017) applied OLS on a dataset of variable indicators to predict Nigeria GDP. Patrick & Sebastian, 2009 predicted the GDP growth of Baltic States, Estonia, Latvia, and Lithuania using a reduced Vector Autoregressive (VAR) method. Syrkri (2020) determined the relationship between GDP and some macroeconomic variables in Indonesia using the granger causality test, and VAR which is a TS method. In contrast, our study aim to develop a predictive model for PPP GDP per capita that accounts for real GDP upon which countries' economies are rated over World PPP GDP per capita, to enable decision-makers to identify and design the indicators influencing real GDP for economic growth.

Consequently, the study attempt to predict the GDP to macroeconomic indicators by applying three machine learning models; Principal Component Regression (PCR), Ridge Regression (RR) and Lasso Regression (LR). It will also apply one classic statistical model and predictive method; Ordinary Least Squares (OLS), revealing strategies which perform better than others. The second objective is to identify the key macroeconomic variables that will influence the growth of GDP, and achieve prediction with high accuracy to explain a best practice for building trust between machine learning and decision-makers. The notion is that economic policymakers should accept with responsiveness and espouse machine learning as a potent tool to predict the GDP to macroeconomic indicators.

2. Methods and techniques

Machine learning is part of artificial intelligence that allows software applications to predict outcomes accurately, without being programmed explicitly, and its algorithms use historical data as input to predict new output values (Ed Burns, 2021; Nilesh, 2021). It focuses on building systems that learn or improve performance based on the data they consume (Oracle, 2021; Prateek, 2021, Kung and Huang, 2018).

Generally, GDP is influenced by many macroeconomic indicators which have been featured under agriculture and rural development, climate change, economy and growth, education, energy and mining, environment, external debt, financial sector, public and private sectors, science and technology, and so on (world bank, 2022). Each of these categories has many indicating variables ranging from population, mortality rate, poverty headcount ratio, renewable energy consumption, central government debt, gross capital formation, labor force, inflation rate and unemployment rate.

However, the selected macroeconomic variables for this study are the export rates, import rates, population, Foreign Direct Investment (FDI), federal government expenditure, oil revenue, and foreign exchange rates data, making up 7 predictors while GDP is the response variable. Exports and imports of goods and services influence the level of growth in GDP (Tejvan, 2017). If a country imports more than it exports, it runs a trade deficit as opposed to a trade surplus if otherwise. When population growth and per capita GDP growth are completely independent, higher population growth rates would lead to higher economic growth rates (World Bank, 2021a). FDI is made to establish effective management control over an enterprise in another country. The IMF suggests that investments should account for at least 10 percent of voting stock to be counted as FDI. In practice, many countries set a higher threshold (World Bank, 2021b). The real GDP seems to be inelastic or respond at small range to the increase in government current and capital expenditure (Ahmad & Malak, 2017). Earnings from natural resources, especially from fossil fuels and minerals economic rents, account for a sizable share of GDP (Word bank, 2021c). Good implementation of foreign exchange policies is supposed to make a significant impact on economic growth (World Bank, 2021d).

2.1. Machine learning methods for the study

The dataset for this study consists of quantitative (numerical) data that informed the adoption of machine learning regression methods namely: PCR, RR, LR, and OLS methods. The PCR curbs the problems of high dimensional and collinear dataset by reducing the number of predictor attributes (Tom, Michael, Marie-Louise & Alan, 2005). It increases the accuracy of interpreting large dataset and minimizes information loss, and also reduce eigen value/vector problem (Jolliffe & Cadima, 2016). Furthermore, it obtains the transformed predictors Z_1, Z_2 . . . Z_m representing $M < P$ linear combinations of the original predictors p and apply least squares to fit the linear regression model using the M predictors (Gareth, Daniela, Trevor & Robert, 2017). The OLS makes a prediction using the estimated coefficients that only minimizes the Residual Sum of Squares (RSS). New types of ridge regression estimators: one parameter ridge-type estimator (Kibria & Lukman, 2020), and rank regression (Arashi, Roozbeh, Hamzah & Gasparini, 2021) have been developed to curb the outlier and multiple collinear dataset issues that is associated with the OLS estimator. The ridge regression coefficient estimates; $\hat{\beta}_{\lambda}^{R}$, use the values that minimize two terms; the RSS and the cost function (Gareth et al., 2017). Like ridge, lasso regression coefficient estimates $\hat{\beta}_{\lambda}^{L}$ also use values that minimize the RSS and the cost function. The difference however is that lasso regression has the effect of forcing some of the coefficient estimates to be exactly zero when the tuning parameter λ is sufficiently large (Gareth et al., 2017).

2.2. The data set of the study

The data set of interest concerns Nigeria's macroeconomic indicators retrieved and extracted from World Bank macroeconomic indicators repository containing (39) instances between 1981 and 2019 with (7) attributes without missing values. Nigeria's macroeconomic data from 1970 to 1980 were not included in the experiment because they were not captured by the World Bank and the data for 2020 and 2021 were also not used because they contained missing values. Table 1 shows a parts of Nigeria's macroeconomic dataset extracted from World Bank and cleaned.

3. Theory/calculation

The experiment was carried out using Python 3 on Jupyter notebook which is an explore-execute environment containing a vast collection of machine learning algorithms for data analysis tasks as adopted by McKinney (2017). Estimated coefficients of the variables were computed using some calculus built inside the python machine learning library.

3.1. Splitting training from testing dataset

Most times, the data that would be used for testing the result of predictive models may not be available. Then, dividing the dataset into two (training and testing datasets) becomes the available option (Gareth et al., 2017). To obtain the training and testing data for this study, we split the dataset in the ratio of 8:2

Table 1
Part of Nigeria's Macroeconomic Dataset from World Bank.

	Year	POPULATION	EXCHANGE RATE	FEDERAL GOVERNMENT EXPENDITURE	FOREIGN DIRECT INVESTMENT	IMPORT RATE	EXPORT RATE	OIL REVENUE	GDP
0	1981	75440502.0	0.617708	1.776406	5.423273e+08	1.575943e+10	1.412856e+10	3.228661	1.644750e+11
1	1982	77427546.0	0.673461	2.106435	4.306113e+08	1.011682e+10	9.556563e+09	1.513785	1.427690e+11
2	1983	79414840.0	0.724410	2.077117	3.644346e+08	4.380360e+09	5.372793e+09	4.080556	9.709491e+10
3	1984	81448755.0	0.766527	2.091336	1.891648e+08	2.226400e+09	4.666830e+09	8.947623	7.348436e+10
4	1985	83562785.0	0.893774	1.939839	4.855813e+08	2.430040e+09	5.233610e+09	9.571146	7.374582e+10
5	1986	85766399.0	1.754523	1.929236	1.932149e+08	2.130169e+09	2.876809e+09	4.306789	5.480585e+10
6	1987	88048032.0	4.016037	1.632709	6.105521e+08	3.501734e+09	6.767637e+09	9.566623	5.267604e+10
7	1988	90395271.0	4.536967	1.552698	3.786671e+08	2.865443e+09	5.545311e+09	8.105433	4.964847e+10
8	1989	92788027.0	7.364735	1.315222	1.884250e+09	3.930284e+09	1.111111e+10	20.085615	4.400306e+10
9	1990	95212450.0	8.038285	1.220141	5.878830e+08	5.376544e+09	1.133389e+10	21.857766	5.403580e+10
10	1991	97667632.0	9.909492	1.220982	7.123734e+08	6.274751e+09	1.190968e+10	13.874440	4.911843e+10

and used 80% of the dataset for training the model while 20% of the dataset was used for testing the model. Various portions of the split dataset were assigned to the various variable names. Eighty percent instances of the macroeconomic predictors (population, exchange rate, federal government expenditures, federal direct investment, import rate, export rate, and oil revenue) were collectively assigned to *x_train* variable along with the corresponding 80% instances of the response (GDP) assigned to *y_train* variable for training the model. The remaining 20% instances of the macroeconomic predictors were collectively assigned to *the x_test* variable along with the corresponding 20% instances of the GDP assigned to the *y_test* variable for testing the model.

3.2. Development and evaluation techniques for the model

The techniques for developing the predictive models for the OLS, RR, LR, and PCR methods that were used in the study were from machine learning scikit-learn libraries in python (sci-kit, 2022). For example, the following tools from scikit-learn libraries were used for developing the RR model. The "*ridge*.fit(*x_train*, *y_train*)" 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(*x_test*)".

Evaluation of the various machine learning models was done using *the k*-fold (5 fold) Cross-Validation technique. The PCR was built using 2 principal components. The ridge regression and lasso regression were built with optimal lambda values of 1.00 and 0.001 respectively

3.3. Conventions used in the study

The words in *italics* indicate the names the authors choose to assign to the variables while the words in "double quotation marks" indicate the names from the tools and techniques used in this study. These conventions are used in Section 3.1 & 3.2.

3.4. Empirical/Theoretical framework of the study

The GDP measurement of an economy could be computed based on market value by applying the model in Eq. (1), where the price of a product n is p, and the quantity of product n purchased is Q.

$$GDP = (P_1 \cdot Q_1) + (P_2 \cdot Q_2) + \cdots + (P_{n+1} \cdot Q_{n+1}) + (P_n \cdot Q_n) \quad (1)$$

Eq. (1) assumes that only the current year production is considered. Such an assumption makes the above calculation a nominal GDP (WikiBooks, 2021). Unfortunately, change in nominal GDP cannot always signify a future change. Hence, the real GDP compares the current year GDP with a base-line GDP. The model for the real GDP of an economy is shown in equation 2, where $P_{B,n}$ is the base year price of product n and $Q_{C,n}$, the quantity of good n purchased in the current year.

$$Real\ GDP = (P_{B,\ 1} \cdot Q_{C,1}) + (P_{B,2} \cdot Q_{C,2}) + \cdots + (P_{B,n+1} \cdot Q_{C,n+1}) + (P_{B,n} \cdot Q_{C,n}) \quad (2)$$

On the hand, the national expenditure accounting model of the GDP is given in Eq. (3), where C, I, G, E, and I stand for Consumption, Investment, Government spending, Export and Import.

$$GDP = C + I + G + (X - M) \quad (3)$$

Regression is a machine learning technique for solving quantitative problems (Pavan, 2020). The relationship between the predictor variables, *Xs* and the response variable, *Y* for a multiple linear regression is given in Eq. (4), where the coefficient β_0 is the intercept term - that is, the expected value of *Y* when $Xs = 0$, and the coefficient β_1 through β_p is the average effect on *Y* associated with a one-unit increase in each of the *X*. The error term $\in$ is a catch-all for what we miss with this model (Gareth et al., 2017)

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \in \quad (4)$$

OLS computes the estimated coefficients $\hat{\beta}_0,\ \hat{\beta}_1,\ \ldots,\ \hat{\beta}_p$ by minimizing RSS given in Eq. (5), where n is the number of instances, p the number of predictor variables, and i the individual instances.

$$RSS = \sum_{i-1}^{n} \left(y_i - \hat{y}_i\right)^2$$

$$= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip}\right)^2 \quad (5)$$

RR coefficient estimates $\hat{\beta}_{\lambda}^{R}$ use the values that minimize two terms given in Eq. (6), where λ is a tuning parameter, the second term, $\lambda \sum_{i=1}^{n} \beta_j^2$ is the regularization penalty, also known as the cost function. The cost function is small when the coefficient estimates $\beta_0,\ \beta_1,\ \ldots,\ \beta_p$ are close to zero (Gareth et al., 2017).

$$RSS + \lambda \sum_{i=1}^{n} \beta_j^2 \quad (6)$$

LR computes the estimated coefficients $\hat{\beta}_0,\ \hat{\beta}_1,\ \ldots,\ \hat{\beta}_p$ by minimizing two terms as shown in Eq. (7), where the regularization

Table 2
Model Accuracy.

Methods	Accuracy	MSE	Best λ	nC
OLS	88.5%	−3.000312454006231e+22		
RR	88.5%	−2.7828793032300693e+22	100	
LR	87.3%	−3.0003128434166187e+22	0.001	
PCR	88.9%	−7.552007365635066e+21		2

MSE denotes Mean Square Error; a measure of relative error of regression method.

penalty, $\lambda \sum_{i=1}^{n} |\beta_j|$ has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large.

$$RSS + \lambda \sum_{i=1}^{n} \left|\beta_j\right| \tag{7}$$

PCR, first, obtains the transformed predictors $Z_1, Z_2, \ldots, Z_m$ representing $M < P$ linear combinations of the original predictors p, given in Eq. (8), for some constants $\phi_{1m}, \phi_{2m}, \ldots \phi_{pm}$, $m = 1, \ldots, M$

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j \tag{8}$$

Second, it applies OLS to fit the linear regression model using the M predictors given in Eq. (9), where the regression coefficients are given by $\theta_0, \theta_1, \ldots, \theta_M$.

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m Z_{im} + \epsilon_i, \quad i = 1, \ldots, n, \tag{9}$$

4. Results and discussion

It is necessary to establish the macroeconomic indicators responsible for growth of GDP and the specific nature of their influences on GDP with a more reliable machine learning approach, to achieve a high level of economic growth. If these relationships are established, they would enable economic policymakers to redesign the influencing variables to achieve a 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 which would be suitable for predicting whether GDP would grow or not, given macroeconomic data. In this regard, the study applies and compares PCR, RR, LR, and OLS methods. The second objective is to enhance the economy's effectiveness by determining the key macroeconomic indicators that influence the GDP. We intend to identify the relationship between the GDP and the macroeconomic variables by minimizing two terms – the RSS and cost function (Lambda). The addition of the cost function would allow for a bias-variance trade-off that eliminates the rigid linear relationship between the GDP and the macroeconomic variables. This flexibility is hoped to reduce the over-fitting and the mean square error (MSE) of the model, and improve its predictive accuracy.

The result revealed that PCR had the highest predictive accuracy of 88.9%, with the number of missing values being 0 making 100% of the 39 observations (Table 2). Table 3 shows that estimated coefficients $\hat{\beta}_j$ associated with each of the macroeconomic indicators for each of the four models that were built with the four different methods. The four methods that were adopted for building the models were visualized for actual GDP against predicted GDP and PCR had the best predictive accuracy and minimal Mean Square Error (MSE) relative to its counterparts (Figs. 1–4). The results of the study were summarized in Tables 2 and 3, and Figs. 1–4.

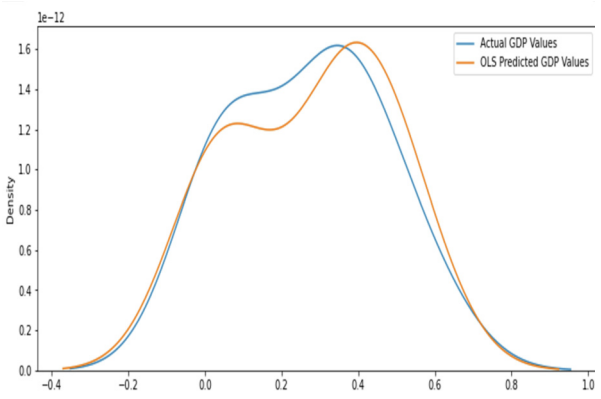

Fig. 1. OLS: Actual GDP vs. Predicted GDP.
Figs. 1–4: The four methods that were adopted for building the models were visualized for Actual GDP against Predicted GDP. The PCR has the best predictive accuracy and minimal Mean Square Error (MSE) relative to its counter parts.

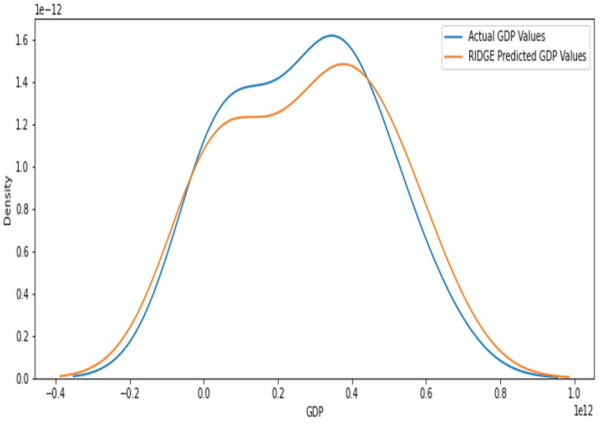

Fig. 2. Ridge: Actual GDP vs. Predicted GDP.
Figs. 1–4: The four methods that were adopted for building the models were visualized for Actual GDP against Predicted GDP. The PCR has the best predictive accuracy and minimal Mean Square Error (MSE) relative to its counter parts.

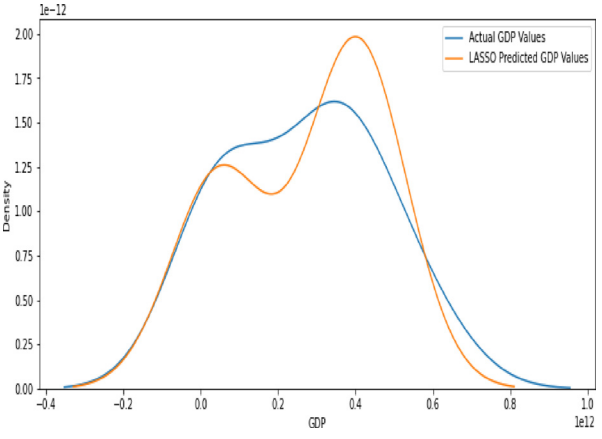

Fig. 3. Lasso: Actual GDP vs. Predicted GDP.
Figs. 1–4: The four methods that were adopted for building the models were visualized for Actual GDP against Predicted GDP. The PCR has the best predictive accuracy and minimal Mean Square Error (MSE) relative to its counter parts.

4.1. Predictive accuracy

The result revealed that OLS, LR, RR, and PCR produced an accuracy of 88.5%, 87.3%, 88.5%, and 88.9% respectively with the number of missing values being 0 making 100% of the 39 observations. Generally, the best regression method would have

Table 3
Estimated Coefficients $\hat{\beta}_j$ of Macroeconomic Indicators for the 4 Models.

Pred	OLS $\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}$	Ridge $\hat{\beta}_j^R$	Lasso $\hat{\beta}_j^L$	PCR $\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}$
Cons	−31,982,197,633.217712	3,140,795,154.2847595	−119,720,953,530.28549	−119,720,953,530.28549
Pop	1.00575940e+03	6.82614338e+02	1.91987796e+03	–
Xcr	−2.12583178e+08	−1.68328489e+08	−1.18367329e+08	–
Fge	1.65897780e+10	2.54062299e+09	2.93740470e+10	–
Fdi	−8.38115960e+00	−5.44116207e-01	−6.21229837e+00	–
Imr	3.64293149e+00	4.74395349e+00	1.01301290e+00	–
Exr	7.82089949e-01	7.90344390e-01	4.67938757e-01	–
Oil	−4.17033197e+09	−3.77813055e+09	−5.08831835e+09	–
Z_1	–	–	–	1.91987796e+03
Z_2	–	–	–	−1.18367329e+08

Table 2 shows that estimated coefficients $\hat{\beta}_j$ associated with each of the macroeconomic indicators for each of the four models that were built with the four different methods. Constant $\hat{\beta}_0$ is the intercept term 3,140,795,154.2847595, showing the expected GDP when the macroeconomic indicators are not considered.

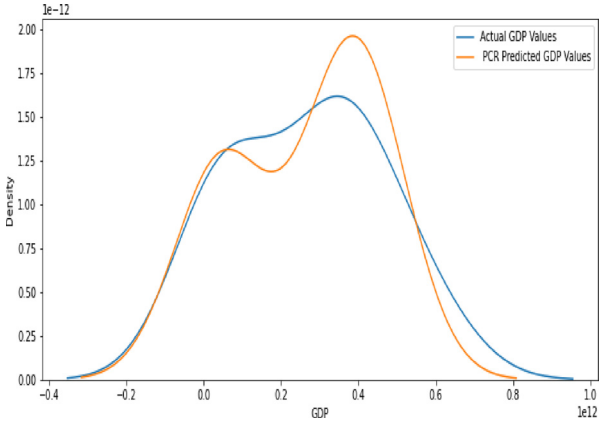

Fig. 4. PCR: Actual GDP vs. Predicted GDP.
Figs. 1–4: The four methods that were adopted for building the models were visualized for Actual GDP against Predicted GDP. The PCR has the best predictive accuracy and minimal Mean Square Error (MSE) relative to its counter parts.

an MSE value being very close to the irreducible error $Var(e)$ (Gareth et al., 2017). The irreducible error is the lowest achievable MSE among all possible methods. On the other hand, when a regression method is farther away from $Var(e)$, it has a relatively higher amount of error. From the result, the PCR method achieved the lowest MSE of −7.552007365635066e+21 followed by ridge regression with MSE of −2.7828793032300693e+22 marginally outpacing the least-squares method which produced an MSE of −3.0003124540062 31e+22. Finally, the least performed regression method is the lasso regression with an MSE of −3.0003128434166187e+22. Hence, the PCR is the most accurate among the four regression methods with 88.9 percent accuracy.

Furthermore, each of the four methods that were adopted for building the models was visualized for actual GDP against predicted GDP, as shown in Figs. 1–4. It also indicates that PCR has the best predictive accuracy and minimal Mean Square Error (MSE) relative to its counterparts.

The prediction of the GDP results could have been developed by using the PCR method if we were considering the model for inference "white-box" where the goal would be concluding and confirming the outcome of the GDP result (Gareth et al., 2017). Similarly, Giovanni et al. (2021), stated that PCA does not provide the economic interpretation of the results with regards to transforming and reducing the number of macroeconomic variables. Hence, we suggest building the model using RR that ranked second in predictive accuracy and optimal mean square error. The idea behind choosing the RR model is that it would provide results that would be reliable estimates of future levels of GDP. Presumably, the model is hoped to perform well for other countries than Nigeria.

4.2. Coefficient analysis of the macroeconomic indicators

In line with the second objective of the study which is to identify the characteristics of macroeconomic variables that are most likely to influence the GDP, coefficient analysis was performed on all the macroeconomic variables by estimating the coefficients of each of the macroeconomic indicators. Table 3 displays the estimated coefficients of macroeconomic variables.

The following variables stand for their respective macroeconomic indicators. *Pop*: population, *xcr*: foreign exchange rate, *fge*: federal government expenditure, *fdi*: federal direct investment, *imr*: import rate, *exr*: export rate, and *oil*: oil revenue

Ridge regression explained the influence of the macroeconomic indicators on GDP as follows: for a given value of other macroeconomic indicators, GDP will approximately increase by 6.82614338e+02 increase in the number of population; GDP will approximately decrease by −5.44116207e-01 units increase in foreign direct investment; GDP will approximately decrease by −3.77813055e+09 units increase in oil revenue. It also demonstrated that when other macroeconomic indicators are fixed, GDP will approximately decrease by −1.68328489e+08 units increase in the exchange rate; GDP will approximately increase by 2.54062299e+09 units increase in federal government expenditures; GDP will approximately increase by 4.74395349e+00 units increase in import rate; GDP will approximately increase by 7.90344390e-01 units increase in export rate.

Conversely, the findings agree with the study conducted by Giovanni et al. (2021), revealing that machine learning methods capture predictive ability and perform better than traditional OLS and TS analysis that tend to overestimate the GDP predictions. In contrast to the work of Divya and Rama (2014) which revealed that exchange rate had a strong positive influence on GDP. This study suggested otherwise. This could be attributed to the machine learning method adopted in this study against the OLS approach used by Divya and Rama (2014). Yua et al. (2017) suggested that various macroeconomic indicators could not affect GDP in Nigeria economy, using Granger Causality Technique. Moreover, the current research indicated that population is another strong indicator influencing GDP. The study revealed oil revenue as a negative indicator of GDP in Nigeria, using the prediction as a preliminary point, and inherently, it may be likewise for other oil-producing countries. Thus PPP GDP per capital economies evaluation of GDP growth is suggested not to consider oil revenue variable, especially in Nigeria.

Summarily, using RR model and the macroeconomic variables that influence GDP, positively, we suggest the following equation to predict GDP;

$GDP = 3, 140, 795, 154.2847595 + pop$ x $6.82614338e + 02 + fge$ x $2.54062299e + 09 + imp$ x $4.74395349e + 00 + exr$ x $7.90344390e - 01$

4.3. Motivations

Regrettably, the disappointment over measured GDP in recent years has spurred widespread concerns about whether the statistical systems are efficiently capturing economic variables (Karen and Louise, 2018) and the use of TS data sometimes results in a linear regression which necessitates having a static root test to see the mean, variance, and covariance consistency (Syrkri, 2020). Hence, this research seeks to help support the statistical agencies and decision-makers by building alternative machine learning models using PCR, RR, and LR that would curb the limitations of classical OLS (Gareth et al., 2017) as well as Time Series techniques. Interestingly, OLS assumes a linear relationship between the macroeconomic variables and GDP which in most cases are violated. This has the effect of reducing the predictive accuracy of the model. Second, OLS has no capability of regularizing the model it builds. Hence, it has no support for tuning parameters also known as cost function or penalty term without which OLS will not allow for bias-variance trade-offs. It makes the relationship between the predictors and the response variable rigid, thus subjecting the method to over-fitting the model and invariably yielding low prediction accuracy and high mean square error.

Furthermore, high dimensional data: when the number of observations is not much more than the number of predictors ($n >> p$) as it is in this study, the variability of the OLS approach increases which tends to overestimate the model leading to inaccurate predictions. When the number of observations is lesser than the number of predictors ($n < p$), the OLS will not have a unique solution. Apparently, using R^2 statistics, F-statistics, and correlation as the metrics to measure the fitness of OLS and TS models will perform well on the training data set but not on test data set, since they cannot use unseen data for prediction (Gareth et al., 2017).

Although there are many machine learning methods for solving different kinds of problems, each predictive method has its scope of application and also has its weaknesses caused by the characteristics of its model (Brownlee, 2019; Katrina, 2021; Shaobo, 2021). Hence, this study considered PCR, RR, and LR because their algorithms are specifically designed for training and solving quantitative, continuous numerical data that the dataset of this study consists of. These alternative methods are expected to cover the above enumerated gaps inherent in OLS.

5. Conclusion

The PCR predicts macroeconomic indicators accurately, with minimal MSE better than RR, LR and OLS. Most likely, increase in population, federal government expenditure, import rate, and export rate lead to increase in GDP, while that of foreign direct investment, exchange rate, and oil revenue reduced GDP. These insights and revelations would help the economic policymakers to understand the influence of these macroeconomic indicators on GDP and in turn enhance the nation's economy.

Recommendations

We recommend that future work should apply non-parametric methods and compare their predictive accuracies with the parametric methods that were used in this study, and also use many predictors and apply feature selection techniques to choose the variables that will most influence the GDP.

Data availability statement

The data analyzed in this study are freely available at the World Bank Macroeconomic Dataset repository websites, World Bank (2021a), World Bank (2021b), World Bank (2021c), and World Bank (2021d). The interested reader may use the provided link at the references to explore the data. Any further inquiries can be directed to the corresponding author Agu S. C. or Oden, D.

Authors contributions

Agu S C: Conceptualization, Methodology, Software, Data curation, Writing- Original draft preparation. **Onu F U:** Conceptualization, Validation, Project administration. **Ezemagu U K:** Validation, Project administration, Writing- Reviewing and Editing. **Oden D:** Software, Investigation, Visualization. Approval to submit the manuscript to your journal; all authors

ORCID information

The ORCID Numbers are yet to be obtained and would be submitted when we acquire them.

Author statement

The authors agree to Intelligent Systems with Applications terms and policies.

Declaration of conflict of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Ahmad, A., & Malak, A. (2017). The Relationship between Government Expenditure and GDP: Empirical Testing of Wagner's Law in the Jordanian Economy. https://www.yu.edu.jo/econconf9/New/shares/-. Retrieved 26th April 2021

Arashi, M., Roozbeh, M., Hamzah, N. A., & Gasparini, M. (2021). Ridge regression and its applications in genetic studies. *PloS one, 16*(4), Article E0245376 https://doi.org/10.1371/journal. pone.0245376.

Asger, L., & Miha, T. (2020). Including news data in forecasting macro economic performance of China. *Computational Management Science, 17*(4) No 6, 585-611.

Brownlee, J. (2019). Why One-Hot Encode Data in Machine Learning? https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/ Retrieved February 6, 2020

Burn, E(2021).s, d.In-depth Guide to Machine Learning in the Enterprise. https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML. Retrieved 29th April 2022.

Divya, K. H., & Rama, D. V. (2014). A Study on Predictors of GDP: Early Signals. *Elsevier Journal of Procedia Economics and Finance, 11*, 375–382.

Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2017). *An introduction to statistical learning*. New York (ISL): Springer.

Genesis. (2018). Pros and Cons of K-Nearest Neighbors. https://www.fromthegenesis.com/pros-and-cons-of-k-nearest-neighbors/ Retrieved March 13, 2022.

Giovanni, M., Giacomo, M., & Sara, S. (2021). GDP Forecasting: Machine Learning, Linear or Autoregression? *Frontiers in Artificial Intelligence*. https://doi.org/10.3389/frai.2021.757864.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Phil. Trans. R. Soc. A.* 3742015020220150202 http://dx.doi.org/10.1098/rsta.2015.0202.

Katrina, W. (2021). A Guide to the Types of Machine Learning Algorithms and their Applications. https://www.sas.com/en_gb/insights/articles/analytics/machine-learning-algorithms.html Retrieved 29hJanuary 2021.

Kibria, B.M.G., & Lukman, A.F. (.2020). A New Ridge-Type Estimator for the Linear Regression Model: Simulations and Applications. Hindawi Scientifica, 2020, Article 9758378. https://doi.org/10.1155/2020/9758378

McKinney, W. (2017). *Python for data analysis, data wrangling with pandas, numpy, and IPython*. O'Reilly Second Edition.

Naresh, K. (2020). Advantages and Disadvantages of KNN Algorithm in Machine Learning. http://theprofessionalspoint.blogspot.com/2019/02/advantages-and-disadvantages-of-knn.html Retrieved March 13, 2022.

Oden, D., Ibeto, A., & Agu, S. C. (2020). *Predictive model for nigeria's domestic product, BSc project*. Department of Computer Science, Madonna University of Nigeria.

Oracle (2021). What is Machine Learning? https://www.oracle.com/ng/data-science/machine-learning/what-is-machine-learning/ Retrieved 29th January 2022.

Patrick, P., & Sebastian, P. (2009). Forecasting GDP Growth – The Case of The Baltic States. https://www.diva-portal.org/smash/get/diva2:229044/FULLTEXT01.pdf. Retrieved 9th March 2022.

Pavan, V. (2020). 6 Types of Regression Models in Machine Learning You Should Know About.https://www.upgrad.com/blog/types-of-regression-models-in-machine-learning/ Retrieved 21st April 2021

Picardo, E. (2021). The Importance of GDP. https://www.investopedia.com/articles/investing/121213/gdp-and-its-importance.asp Retrieved August 8, 2020.

Shaobo, L. (2021). Research on GDP Forecast Analysis Combining BP Neural Network and ARIMA Model. *Computational Intelligence and Neuroscience, 2021*(Article ID 1026978). https://doi.org/10.1155/2021/1026978.

Syrkri, A. U. (2020). The Relationship between Gross Domestic Products with International Balance of Payment: Empirical Evidence from Indonesia. *Journal of Developing Economies, 5*(2), 103–119.

Tejvan, P. (2017). Importance of exports to the economy. https://www.economicshelp.org/blog/7164/trade/importance-of-exports-to-the-economy/ Retrieved 26th April 2021.

Tom, H., Michael, G. M., Marie-Louise, O. C., & Alan, G. R. (2005). The Effect of Principal Component Analysis on Machine Learning Accuracy with High Dimensional Spectral Data. In *Conference paper presented at the meeting of the Specialist Group on Artificial Intelligence*.

Ukpe, W. (2021). GDP Performance of Nigeria's Presidents since 1999. https://nairametrics.com/2021/05/21/gdp-performance-of-nigerias-presidents-since-1999/ Retrieved 10th August 2021.

WikiBooks (2021). Principles of Economics/GDP https://en.wikibooks.org/wiki/Principles_of_Economics/GDP. Retrieved August 8, 2020.

World Bank (2020). International Comparison Program (ICP) https://www.worldbank.org/en/programs/icp Retrieved 11th March 2022.

World Bank. (2021a). Population, total – Nigeria https://data.worldbank.org/indicator/SP.POP.TOTL?locations=NG&view=chart Retrieved, 26th April 2021.

World Bank. (2021b). Foreign direct investment, net inflows (BoP, current US$) – Nigeria. https://data.worldbank.org/indicator/BX.KLT.DINV.CD.WD?locations=NG&view=chart Retrieved 26th April 2021.

World Bank. (2021c). Oil rents (% of GDP) – Nigeria. https://data.worldbank.org/indicator/NY.GDP.PETR.RT.ZS?locations=NG&view=chart Retrieved 26th April 2021.

World Bank. (2021d). Official exchange rate (LCU per US$, period average) – Nigeria. https://data.worldbank.org/indicator/PA.NUS.FCRF?locations=NG&view=chart Retrieved 26th April 2021.

World Bank (2022). Indicator. https://data.worldbank.org/indicator?tab=featured. Retrieved 9th March 2022.

Worldometer (2017). GDP per capital. https://www.worldometers.info/gdp/gdp-per-capita/#:~:Text=Gross%20Domestic%20Product%20(GDP)%20per,divided%20by%20its%20total%20population. Retrieved 9th March 2022.

Yua, H., Adoms, F. U., Okaro, C. S., & Ogbonna, K. S. (2017). Causal Relationship between Financial Structure and Economic Growth in Contemporary African Economy: A Case Study of Nigeria from 1990 to 2018. *European – American Journal, 8*(3), 59–68.