Journal of Computing and Communication Vol.2 , No.1 , PP. 1-8 , 2023

Nowcasting Egypt GDP Using machine learning Algorithms

Mohamed F. Abd El-Aal [a] , Diaa Salama AbeElminaamb [b,c,*], Alaa-Eldin Abd-Elatif [d]

[a] *Economics department, faculty of commerce, Arish university*

[b]*Department of Computer Science, Misr International University, Cairo, Egypt*

[i]*nformation systems department; Faculty of Computers and artificial intelligence; Benha university*, *Egypt*

[d]*Higher Institute of Commercial Sciences, Mahla, Egypt*

[*]Corresponding Author: Diaa Salama [***diaa.salama@miuegypt.edu.eg***]

ARTICLE DATA

Article history:
Received 30 Sep 2022
Revised 12 Jan 2023
Accepted 12 Jan 2023
Available online

Keywords:
Gross domestic product,
Inflation rate,
Unemployment rate,
Machine learning algorithms,
Gradient boosting.

ABSTRACT

This paper aims to determine the most accurate algorithm for predicting the Egyptian gross domestic product (GDP) and not only and also to determine the relative weight of the effect of the components of the output on it to assist decision-makers in making good economic policies. It turned out that the gradient algorithm is the most accurate and highly efficient algorithm for predicting the Egyptian GDP. It also became clear that government investment is the biggest influence on the Egyptian GDP at 21%, followed by the consumer spending of the family sector at 19%, followed by investment spending by the private sector and imports by the same 15%, then exports and government spending by 14% and 13%, respectively. Thus, to stimulate and maximize the size of the Egyptian GDP, the decision-maker must focus on stimulating government investment spending and consumer spending for the household sector and investment spending for the private sector.

1. Introduction

Policymakers, corporations, and financial market participants are all interested in the economy's condition to make future decisions. Gross domestic product (GDP) shows that, so the economist working to expect and predict its values to put macroeconomic policies. GDP doesn't show the economic volume of the country only but determines the inflation and unemployment rate and predicts it put Either a state of pessimism or optimism among local and foreign investors. And GDP prediction makes us determine the strong and weak points in our economy based on this status, and we make future policies to address the problem before it happens.

The Egyptian GDP achieved decreasing growth rates from 2000 to 2020. The growth rate in output reached about 6% in the year 2000. It decreased to 4.4% in 2005, then increased again to 5.1%, but quickly decreased in 2015 and 2020 to 4.3% and 3.5%, respectively. It's important to note that the Egyptian GDP only saw a 7.1% growth spike between 2007 and 2008. It reached its lowest growth rate in 2011 due to the political turmoil resulting from the January 25 revolution "world bank data. [1]"

As for the sectoral distribution of GDP from 2000 to 2020, the service sector came with the most significant contribution in that period, as it contributed 46% in 2000 and increased to 51.6% in 2020. The industrial sector followed a contribution rate of 30.7% in 2000 and increased to 31.8% in 2020. Finally, the agricultural sector came with a contribution rate of 15.5% in 2000, which decreased to 11.1% in 2020 in favor of the service and industrial sectors "world bank data. [1]"

To evaluate which technique is more accurate and apply it to our mission, we will utilize machine learning techniques(ML) specializing in Gradient boosting algorithms (GB), logistic regression (LR), support vector machine (SVM), naive Bayes (NB), k-nearest neighbors (KNN), and Random forest (RF).

In section 2, similar studies are discussed, and relevant results are presented. In Section 3, the methodology followed in this study is explained. Section 4 discusses the techniques used and the output of each one after testing on all data sets. Section 5 provides an evaluation of the results presented in Section 4. Lastly, the conclusion is conducted in Section 6.

2. Related Work

In [2], Biau and D'Elia (2010) utilized a random forest model to estimate the GDP data for the euro area, and they discovered that the predictions made by the machine learning model were more precise than those made by a conventional autoregressive model. Using the elastic net, recurrent neural network, and super-learner models,

In[3], Jung et al. (2018) forecasted real GDP growth in the United States, the United Kingdom, Germany, Spain, Mexico, the Philippines, and Vietnam. Elastic net and random forest models were used by Tifn (2016) to anticipate GDP growth in Lebanon, which only releases official GDP growth statistics after a two-year wait.

In[4], Emsia and Coskuner (2016) employed support vector regression to forecast Turkey's GDP growth. However, the prior literature has not sufficiently studied Japan's real GDP growth projection. Second, over a significant length of time, the machine learning approach used in this work outperforms the predictions given by the BOJ and the IMF in terms of predicting yearly real GDP growth in Japan. Last but not least, this study introduces a cross-validation and hyperparameter tuning approach to handle forecasting concerns, like overfitting problems and provides the precise parameters utilized in the prediction models, which can serve as a useful reference for pertinent research in the future. Also, other related work in [5-20] has been proposed in recent years to address machine learning and its application in different fields.

3. Research Methodology

This paper used RF, SVM, LR, NB, KNN, and GB models. All algorithms have monitored ML to study training data and produce a data prediction function.

Six algorithms are tested for accuracy, with the accuracy model utilized to estimate GDP and extract the GDP value-influencing factors.

The world bank dataset was used for my prediction from 1990 to 2019, and I depend on the expenditure approach to assess and predict GDP {GDP=C+I+G+(X-M)}. The Scikit-Learn package was used to create the ML methods used in this work in Python.

4. The ML techniques :

4.1. Support Vector Machine (SVM)

When utilizing a special machinery learning method, classification applications uncover an independent and identically distributed data set (iid). A discriminating categorization algorithm is given a data point called x. This discriminating function reliably foresees new occurrence labels. When performing classification tasks, it places it in one of the many classes instead of generative ML techniques, including computing probability distributions. When outlines are necessary, discriminatory approaches, which are less effective and are frequently used, need fewer resources, especially in multidimensional fields. When just later opportunities are required, a multidimensional surface equation that best separates several classes must be found. As the convex optimization problems are analytically resolved, SVM always offers the same optimum space value, unlike evolutionary algorithms like perceptrons, widely employed in ML classification. Perceptrons have significant setup and termination needs. "El-Aal, A., Mohamed, F, et al., 2021" [5].

Vapnik proposed the SVM regression model as a non-parametric technique (1995). The SVM linear function looks like this:

$$f(x) - \langle w, x \rangle + b \tag{1}$$

We indicate to weight vector x is the input or feature vector, and b denotes the bias, intending to keep the function as flat as feasible, i.e., a small WW. Minimizing the usual, i.e., w2, is one technique to do this "Richardson and Mulder et al., 2018" [6]. Defined the function as a convex optimization problem:

$$0.5 \parallel w \parallel^2 + C \sum_{i=1}^{l} \left|(y_l - f(x_l))\right|_{\epsilon} \tag{2}$$

4.2. Random Forest (RF) :

An RF algorithm is composed of many different decision trees. We forecast a binary outcome variable using a classification-style decision tree instead of a serial number. These two decision trees similarly divide the data into two groups at each decision point. At every node, a yes-or-no choice is made. For instance, is it true or false that x is greater than 5? The data is then divided based on the response. The data is then divided again, but more explanatory factors are included this time. The one that can explain the most substantial data separation is the first explanatory variable that is selected. The mean value of the data in the separated bucket serves as the model's prediction for that smaller bucket. Because the model was trained too closely to the in-sample data, overfitting can happen when a decision tree has an excessive number of partitions, which leads to poor performance in out-of-sample predictions. Restricting the number of variables and decision nodes is advisable when out-of-sample prediction is a severe problem. "Rajkumar, 2017". [7]
The RF method seeks to avoid overfitting without constricting the size of the tree or the permitted number of divisions. By cultivating many trees for many people. To reduce the forecast variation, the trees' findings are averaged. In addition, the RF partitions the data at each node using a random variable from a subsample of variables. As a result, each tree's nodes cannot access the same variables. In-sample data overfitting rarely results in problems. "Tiffin, 2016". [8]
The following Equation is the basic RF model:

$$F_0(x) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \gamma)^2 \tag{3}$$

Where γ indicates the expected value, and y_i indicates the observed value

4.3. Logistic Regression (LR)

When a binary dependent variable is present, the LR provides interpretation probabilities limited to zero to one, just like linear regression. Additionally, it implies that the return to explanatory factors decreases when the likelihoods approach zero or one. This rise in the separate variable will cause a far larger change in the output when it is close to half as opposed to zero or one end. "Rajkumar,2015". [7] The linear function's transformed log undergoes a modification log $\left(\frac{p}{1-p}\right)$:

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x \tag{4}$$

β_0 indicates to intercept, and β_1 indicates the effect of explanatory variables (C, I, G, (X-M) on the dependent variable (GDP).

When utilizing a special machinery learning method, classification applications uncover an independent and identically distributed data set (iid). A discriminating categorization algorithm is given a data point called x. This discriminating function reliably foresees new occurrence labels. When performing classification tasks, it places it in one of the many classes instead of generative ML techniques, including computing probability distributions. When outlines are necessary, discriminatory approaches, which are

4.4. Naïve Bayes

Reverend Thomas Bayes, a British scientist, invented the NB classifier using Probability and statistical approaches. The NB is a widely used model in ML applications because of how easily all attributes can influence the final decision. In many complex real-world circumstances, the NB works far better than one may assume. The NB technique is appealing and suitable for various fields due to its simplicity, equating to computing efficiency prior, posterior, and class conditional probability, which make up the NB Classification's three primary parts. "Nugraha,2019". [9]

The formula for the Bayes Theorem is given by:

$$\mathrm{P}(\Phi \mid \mathrm{X}) = \frac{\mathrm{P}(\mathrm{X}|\Phi)\cdot \mathrm{P}(\Phi)}{\mathrm{P}(\mathrm{X})}, \tag{5}$$

X refers to Unknown class information, Φ refers to hypothesis (X) as a specific class, $\mathrm{P}(\Phi \mid \mathrm{X})$ refers to The Probability of the (Φ) hypothesis refers to (X), $\mathrm{P}(\mathrm{X} \mid \Phi)$ refers to Probability (X) in the hypothesis (Φ), $\mathrm{P}(\Phi)$ refer to Probability of the hypothesis (Φ), and P(X) refers to Probability (X).

To know the Naive theorem Bayes, it is essential to recognize that the classification process uses several indications to identify the sample-based class (Nugraha,2019). [9] This transformed the theorem of Bayes into:

$$\mathrm{P}(\Phi \mid \mathrm{X}_1 \ldots\ldots \mathrm{X}_n) = \frac{\mathrm{P}(\Phi)\mathrm{P}(\mathrm{X}_1 \ldots\ldots \mathrm{X}_n|\Phi)}{\mathrm{P}(\mathrm{X}_1 \ldots\ldots \mathrm{X}_n)}, \tag{6}$$

The *Φ* variable represents class, or variable X1 … Xn indicates the features of the required instructions for the classification process.

4.5. K-nearest neighbors (K-NN)

One of the most often used algorithms in master's study research is the closest Neighbor (kNN). Given that the label of an instance matches its kNN instances, KNN is based on labeling the k examples closest to the data. It could also be referred to as a unique case. The fundamental idea behind kNN's prediction accuracy is that it's an easy-to-create, obvious technology. K-NN makes no presumptions regarding the distribution of the data. These benefits make incremental learning simply because it is based on examples without training to produce predictions. KNN is typically used in supervised learning tasks involving classification and regression. "Kang, 2021". [10]

4.6. Gradient Boosting (Gb model)

Various low-quality models can provide an advanced preview using gradient enhancement. These techniques often start by applying a loss function to a basic target variable model. A new model will be shown after the leftovers from the earlier models are subjected to the loss function. This process still has some momentum. At a high level, we're iterating through the stages below: "Richardson, Mulder, et al., 2019". [6]

$$F_m(x) = F_{m-1}(x) + v\Delta_m(x), \qquad (7)$$

In Fm(x), when the new mapping x shows the target, Fm−1(x) specifies the preceding model. The term $\Delta_m(x)$ signifies the low learner, and ν represents the reduction parameter.

5. Empirical Results:

The key findings of our analysis, which determined the accuracy model, are discussed in this section. Table (1) below demonstrates this:

Table(1): Performance of ML algorithms valuation (dataset 1990-2019)

algorithms	AUC	CA	F1	Precision	Recall
RF	0.997	0.967	0.967	0.970	0.967
SVM	0.997	0.967	0.967	0.970	0.967
L R	1	0.933	0.931	0.939	0.933
NB	0.995	0.933	0.931	0.939	0.933
K-NN	0.975	0.933	0.931	0.939	0.933
GB	1	1	0.99	0.99	O.99

Source: python results by author.

According to the table, the GB model is the most accurate with a percentage of 100%, the RF and SVM with a percentage of 0.967%, and then logistic regression, naive Bayes, and KNN with a percentage of 0.933%. I shall rely on the GB model for GDP prediction despite the model's overall great accuracy, as demonstrated in the table (2):

Table (2): the GB prediction for actual GDP value
Mil dollar

year	GB prediction value (Mil dollar)	GDP actual value (Mil dollar)
1990	42984	42979
1991	37400	37388
1992	41863	41856
1993	46576	46579
1994	51902	51898
1995	60159	60159
1996	67619	67630
1997	78447	78437
1998	84834	84829
1999	90704	90711
2000	99843	99839
2001	96687	96685
2002	85163	85146
2003	80278	80288
2004	78790	78782
2005	89600	89601
2006	107427	107426
2007	130438	130438
2008	162818	162818
2009	189144	189147

2010	218985	218984
2011	235978	235990
2012	279120	279117
2013	288432	288434
2014	305589	305595
2015	329360	329367
2016	332432	332442
2017	235741	235734
2018	249704	249713
2019	303087	303092

Source: python results by author

The table clearly shows that the gradient boosting-predicted values are extremely similar to the actual values of the Egyptian GDP, demonstrating the forecast's accuracy and high caliber.

This paper used RF, SVM, LR, NB, KNN, and GB models. All algorithms have monitored ML to study training data and produce a data prediction function.
Six algorithms are tested for accuracy, with the accuracy model utilized to estimate GDP and extract the GDP value-influencing factors.

Using **the feature importances of the GB algorithm** might help you identify which of your variables has the greatest influence on these models, even though these methods are most frequently employed for prediction. And the results of this code are displayed in the following table:

Table (3): Independent variables' effect on GDP (feature importance indicators)

Independent variables	Percentage
gross fixed capital formation (public sector)	0.21
Households consumption expenditure(C)	0.19
General government consumption expenditure (G)	0.13
gross fixed capital formation (private sector)	0.15
Total exports (X)	0.14
Total imports (M)	0.15

Source: python results by author

From the table, we find that the most important independent variables used to explain the dependent variable are gross fixed capital formation public sector (21%), then the Household's Final consumption expenditure. And the attempt to reduce imports negatively affects the gross domestic product (15%) compared to the positive effect of Egyptian exports (14%). Therefore, the decision-maker should focus in the coming period on these two variables to ensure the achievement of significant growth rates in the gross domestic product.

We can state the results of the paper accurately in the next points:

- The GB algorithm is the most accurate for predicting the Egyptian GDP, and the decision-maker can rely on it to predict the outcome in the future.
- The most important independent variables used to explain the dependent variables are gross fixed capital formation public sector and the Household's Final consumption expenditure. So increase in those variable values lead to an increase in the GDP value.

6. Conclusion

The GDP is one of the indicators that decision-makers always seek to analyze and predict. Suppose the GDP is a significant determinant of the state's economic strength and an indicator for comparing its economic size with other countries. In that case, it is also a determinant of the standard of living of its people. Its growth translates to growth in the level of employment in it, and its stability translates to the stability of the level of inflation in the country. As a result, we can put up the appropriate economic policy for the future by predicting it accurately.

The results of this study proved that using ML techniques in forecasting macroeconomics has high prediction power. And we evaluate the performance of ML algorithms, especially RF, SVM, LR, NB, KNN, and GB, by now casting Egypt's GDP values. And the GB approved that it is an accurate model for GDP prediction than other algorithms. And we also approved the most important independent variables used to explain the dependent variables: gross fixed capital formation public sector and the Household's Final consumption expenditure.

References

[1] World bank dataset, world development indicators, "https://databank.worldbank.org/source/world-development-indicators," Dec 2022. [Online]. Available: https://databank.worldbank.org/source/world-development-indicators

[2] B.Olivier, and A. D'Elia." *A non-balanced survey-based indicator to track Industrial Production"*. No. 259600028. EcoMod, 2010.

[3] J. Jin-Kyu, M.Patnam, and A. Ter-Martirosyan. "An algorithmic crystal ball: Forecasts-based on machine learning". *International Monetary Fund, 2018.*

[4] E.Elmira, and C. Coskuner. "Economic growth prediction using optimized support vector machines." *Computational Economics 48, no. 3 (2016): 453-462.*

[5] Abd Elminaam, D.S., El Tanany, A., Salam, M.A. and Abd El Fattah, M., 2022, May. CPSMP_ML: Closing price Prediction of Stock Market using Machine Learning Models. *In 2022 2nd International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC) (pp. 251-255).* IEEE.

[6] Ai, M.A., Shanmugam, A., Muthusamy, S., Viswanathan, C., Panchal, H., Krishnamoorthy, M., Elminaam, D.S.A. and Orban, R., 2022. Real-time facemask detection for preventing COVID-19 spread using transfer learning based deep neural network. *Electronics, 11*(14), p.2250.

[7] Neggaz, N. and AbdElminaam, D.S., 2021, May. Automatic sport video mining using a novel fusion of handcrafted descriptors. *In 2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC) (pp. 387-394).* IEEE.

[8] Salam, M.A., Ibrahim, L. and Abdelminaam, D.S., 2021. Earthquake Prediction using Hybrid Machine Learning Techniques. *International Journal of Advanced Computer Science and Applications, 12(5), pp.654-6652021.*

[9] Mahmoud, E., Kader, H.A. and Minaam, D.A., 2019, October. Fuzzy knowledge base system for floating car data on SUMO. *In 2019 29th International Conference on Computer Theory and Applications (ICCTA) (pp. 38-42).* IEEE.

[10] AbdElminaam, D.S., ElMasry, N., Talaat, Y., Adel, M., Hisham, A., Atef, K., Mohamed, A. and Akram, M., 2021, May. HR-chat bot: Designing and building effective interview chat-bots for fake CV detection. *In 2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC) (pp. 403-408).* IEEE.

[11] AbdElminaam, D.S., Fahmy, A.G., Ali, Y.M., El-Din, O.A.D. and Heidar, M., 2022, May. DeepECG: Building an Efficient Framework for Automatic Arrhythmia classification model. *In 2022 2nd International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC) (pp. 203-209).* IEEE.

[12] AbdElminaam, D.S., Fahmy, A.G., Ali, Y.M., El-Din, O.A.D., Aly, A.R. and Heidar, M., 2022, May. ESEEG: An Efficient Epileptic Seizure Detection using EEG signals based on Machine Learning Algorithms. *In 2022 2nd International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC) (pp. 210-215).* IEEE.

[13] AbdElminaam, D.S., Ahmed, N., Yasser, M., Ahmed, R., George, P. and Sahhar, M., 2022, May. DeepCorrect: Building an Efficient Framework for Auto Correction for Subjective Questions Using GRU_LSTM Deep Learning. *In 2022 2nd International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC) (pp. 33-40).* IEEE.

[14] Ali, M.A., Orban, R., Rajammal Ramasamy, R., Muthusamy, S., Subramani, S., Sekar, K., Rajeena PP, F., Gomaa, I.A.E., Abulaigh, L. and Elminaam, D.S.A., 2022. A Novel Method for Survival Prediction of Hepatocellular Carcinoma Using Feature-Selection Techniques. *Applied Sciences, 12(13), p.6427.*

[15] M.F. El-Aal, Abd , A. Algarni, A. Fayomi, R. Abdul Rahman, and K.Alrashidi. "Forecasting Foreign Direct Investment Inflow to Egypt and Determinates: Using Machine Learning Algorithms and ARIMA Model." *Journal of Advanced Transportation 2021 (2021).*

[16] A.Richardson , and M. Thomas. "Nowcasting New Zealand GDP using machine learning algorithms." (2018).

[17] Rajkumar, Ved. "Predicting surprises to GDP: a comparison of econometric and machine learning techniques." *PhD diss., Massachusetts Institute of Technology, 2017.*

[18] Tiffin, Mr Andrew. "Seeing in the dark: a machine-learning approach to nowcasting in Lebanon." *International Monetary Fund, 2016.*

[19] Wibawa, Aji Prasetya, Ahmad Chandra Kurniawan, Della Murbarani Prawidya Murti, Risky Perdana Adiperkasa, Sandika Maulana Putra, Sulton Aji Kurniawan, and Youngga Rega Nugraha. "Naïve Bayes Classifier for Journal Quartile Classification." *Int. J. Recent Contributions Eng. Sci. IT 7, no. 2 , PP 91-99 - (2019).*

[20] Kang, Seokho. "K-nearest neighbor learning with graph neural networks." *Mathematics 9, no. 8, 830, (2021).*