International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 06 | June 2021 www.irjet.net p-ISSN: 2395-0072

Gross Domestic Product Prediction using Machine Learning

Vaishnavi Padmawar[1], Pradnya Pawar[2], Akshit Karande[3]

[1]*Student, Dept. of Computer Engineering, Vishwakarma Institute of Technology, Maharashtra, India*
[2]*Student, Dept. of Computer Engineering, Vishwakarma Institute of Technology, Maharashtra, India*
[3]*Student, Dept. of Computer Engineering, Dr. Babasaheb Ambedkar Technological University, Maharashtra, India*

Abstract - *Gross Domestic Product is cited as vital and most widely accepted economic indicator which not only helps in diagnosing the problems related to the economy but also correcting it. The usage of the gross domestic product as a measure of the market price of ultimate services and product that are produced over a selected amount of time will definitely continue to owe an abundant to the producing age. To policy makers and statisticians especially, gross domestic product helps in conveying data about the economy in particular and thereby notifying about country's economic health. This paper makes an attempt to expedite the process of prediction of Gross Domestic Product. Machine Learning algorithms such as Linear Regression and Random Forest are used for prediction. The proposed method using machine learning model proves to be fruitful for financial management.*

Key Words: **Economy, Expenditures, Gross Domestic Product, Linear Regression, Random Forest**

1. INTRODUCTION

Gross Domestic Product (GDP) is the market price of all product and services that area unit made inside the country's national borders in a year. Gross domestic product could be a measure to assess overall economic performance of a country, it includes all product and services made by the economy as well as personal consumption, government purchase, non-public inventories, paid in construction prices and therefore the foreign trade gap. The topic of GDP became of high importance among the indicator of economy variables. Information on Gross Domestic Product is thought to be a crucial indicator for evaluating the national economic development and growth of entire macro economy.

GDP aggregates the complete economic motion. It is frequently used as a finest measure to calculate the performance of the economy. GDP is mostly measured in one of a 3 approaches. First, the Expenditure approach, it involves the worth of all domestic expenditures created on final product and services of the year, beside consumption expenditures, investment expenditures, government expenditures, and web exports. Second, the assembly approach, it's involving the summation of all additional activities at each a part of production by all industries inside the country, taxes and product's subsidies of the amount. Third is that the financial gain approach, it's the summation of all aspect of the financial gain created by production inside the economy as remuneration of workers, capital financial gain, and gross in operation surplus of enterprises i.e., profit, taxes on production and imports less grants of the quantity.

The aim of this study is to predict GDP, using linear regression and random forest for a particular period. Prediction of GDP involves application of applied mathematics and mathematical model to predict future developments within the economy. It permits to review previous economic movements and predict however current economic changes can amend the patterns of previous trend; therefore, a more accurate prediction would provide a significance facilitate to the government in setting up economic development goals, ways and policies. Consequently, a correct Gross Domestic Product prediction presents a number one insight associate an understanding for future economics' trend.

2. LITERATURE REVIEW

Gross Domestic Product's growth rate is treated as a sign of the economic health of the country. A number of studies demonstrates the factors for prediction of GDP using various methodologies. The GDP data ranging from the year 1989 to 2007 of Anhui region in particular was studied by Gang Long [1]. The method depicts the comparative performance of the GA-SVM and RBF neural networks respectively. Jaehyun Yoon [2] explored the Gross Domestic Growth of Japan from the year 2001 to 2018.The data is collected from International Monetary Fund and Bank of Japan. The author worked with gradient boosting and random forest machine learning classifier. MAPE and RMSE method are taken into consideration for the purpose of measuring accuracy of the model. Further, cross validation and hyper parameter tuning are used for the creation of more accurate models. The vector machine was trained with genetic algorithm and henceforth used for GDP forecasting. Relative error method was used to evaluate the model performance. The author concluded that in SVM, optimal solution in short time was acquired by genetic algorithm which worked as a better approach in parameters selection of SVM. For optimizing the support vector machine's parameters, Genetic algorithm was introduced. Various Economic Indicators play a vital role in Gross Domestic Product prediction. Consumption is normally the largest GDP component in the economy. John

[3] coined Real Government Consumption Expenditures, Real Personal Consumption Expenditures and Gross Private Domestic Investment as more vital indicators for predicting GDP. Autoregressive approach predicts consistent future growth in terms of factors related to GDP but fails to overcome historic economic recession. Shelly and Wallace [4] studied the relation between M1 money, real GDP and inflation in Mexico. Annual data from the year 1944 to 1991 is studied. This work indicates that a positive effect on real Gross Domestic Product growth is obtained by unpredictable increases in differenced inflation while predictable increases in differenced inflation results in negative impact on real Gross Domestic Product growth.

In order to produce short-term forecasts of real Austrian GDP, Schneider M. and Spitzer M [5] utilized a generalized dynamic factor model. Macro-Economic Variables has a great influence in country's GDP. Amongst factors like service, agricultural and livestock sector, business sector and industrial sector proves to be dominating one as far as contribution to the GDP is concerned [6]. The influence of small medium enterprises was described by author Maciej Woźniak [7] stating small medium enterprises plays key social role as they reduce unemployment. Carlos Encinas-Ferrer [8] stated why Foreign Direct Investment does not show as an independent variable since FDI has small proportion within the national investment in the countries like Brazil, China, Peru, Mexico and therefore lead to low multiplier effect on the national economy.

3. METHODOLOGY

3.1 Design and Framework

The current study aims to model factors behavior, forecast Gross Domestic Product for a specific amount of your time. A machine learning algorithm, namely, linear regression, is utilized to model and analyze the information provided for Gross Domestic Product prediction. The block diagram depicts the machine learning approach of the implementation of project in Fig.1.

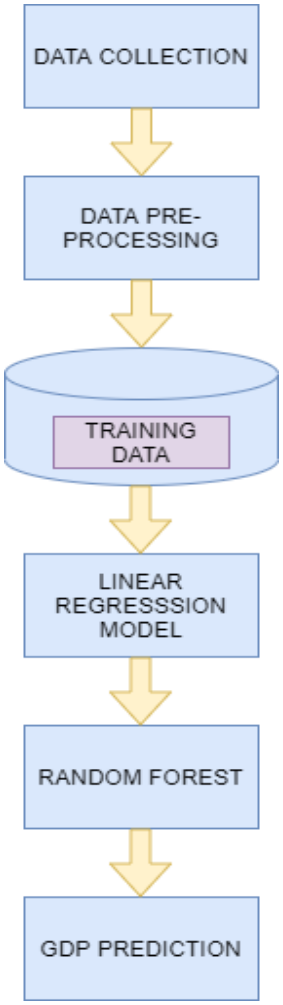

Fig -1: Block diagram

Steps followed for implementation of this project:

A. Data Collection

Data for predicting factors influencing the growth of GDP is taken from Kaggle. The dataset consists of 227 countries data with 20 different factors namely literacy, net migration, population etc.

B. Preprocessing of Data

The goal is to pre-process the data for consecutive step. Therefore, a method is used to handle missing values and outliers. The feature selection was done manually on the basis of literature survey and research. Data cleaning was done by removing the null values followed by removal of outliers by using Interquartile Range (IQR) method. Fig.2. shows the data set after the data pre-processing.

Fig -2: Data set after pre-processing

C. Modelling

During this step, various error metrics are used in order to calculate the accuracy of the forecasting model.

D. Output and Usage

This step presents the accuracy graph of the model.

3.2 Data Analysis and Modelling

A. Linear Regression

It is supervised machine learning algorithm, the most basic type of regression. Basically, it is the mathematical model that analyses the linear relationship between a dependent variable with given set of independent variables(s). In the project the simple linear regression was used to predict the individual attribute of the dataset. For this 80% of the dataset was the training dataset i.e., used for training the model and remaining 20% was used to test the dataset.

Equation of simple linear regression,

$$y = b_0 + b_1 x \quad (1)$$

Y - dependent variable

X - independent variable

b_1 - slop of the regression line

b_0 – Y-intercept

B. Random Forest

Random Forest is one of the well-known machine learning algorithms that belongs to supervised learning technique. Random Forest is used both for regression and classification problems in machine learning. Ensemble Learning is a concept in which multiple classifiers are integrated in order to resolve a complicated drawback and hence it improves the performance of the model. Random Forest relies on the concept of ensemble learning.

As the name itself suggests, "Random Forest is basically a classifier that consists of decision trees of the given dataset on varied subsets. Further, the random forest takes the average in order to improve the forecasting accuracy." Predictions from every tree that is formed are taken into consideration instead of just relying on a single decision tree and after that; based on majority votes of prediction, output is predicted.

The classification of the classes in the project are done on the basis of the predicted data from linear regression. The results of this implementation and its analysis are mentioned further.

4. RESULT AND ANALYSIS

True GDP per capita was plotted against the prediction in order to evaluate model using linear regression. The prediction includes features that encompasses co-relation score more than 0.3 with respect to GDP per capita such as net migration. The Fig. shows the depiction of linear regression model for true GDP per capita prediction.

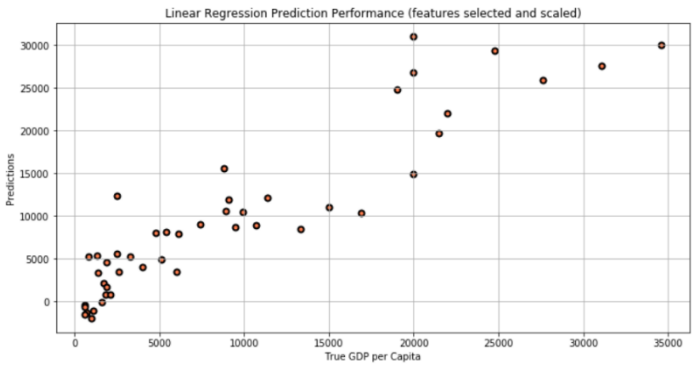

Fig -3: Linear Regression Prediction Performance

Evaluation metrics like MAE (Mean Absolute Error), RMSE (Root Mean Square Error) and R-squared score (R2 score) are used for purpose of Accuracy performance of the model.

Table -1: Linear Regression Accuracy Performance

Data Splitting Criteria	MAE	RMSE	R2_SCORE
All features, No scaling	330350.858	1570337.545	-29843.120
All features, with scaling	569019.468	1283170.821	-19925.990
Selected Features, No scaling	2965.935	4088.794	0.797
Selected Features with Scaling	2879.521	3756.436	0.829

From the metrics analysis, it was clear that feature selection is essential for linear regression model training, in order to get acceptable results on this dataset. On the other hand, feature scaling contains a small positive result on LR's prediction performance. We got decent prediction performance from LR with feature selection and scaling.

For random forest with our data splits (with and without feature selection). Scaling was not tested for Random Forest, since it should not affect the algorithm's performance.

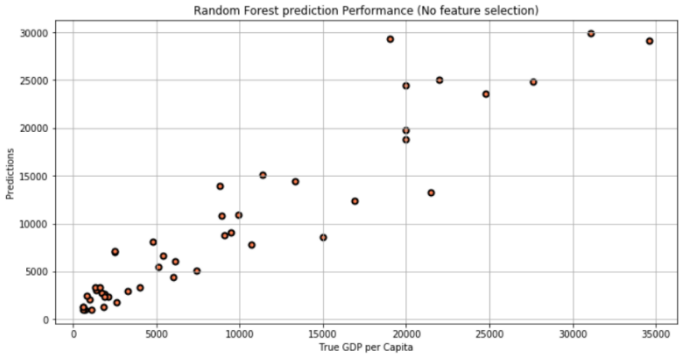

Fig-4: Random Forest Prediction Performance

Table -2: Random Forest Model Accuracy Performance:

Data Splitting Criteria	MAE	RMSE	R2_SCORE
All features, No scaling	2142.130	3097.194	0.883
All features, with scaling	2416.065	3533.590	0.848

1. Optimization

We used grid search in order to obtain good parameters for Random Forest. The parameters we optimized were:

a) n-estimators
b) min_samples_leaf
c) max_features
d) bootstrap

Optimization performance:

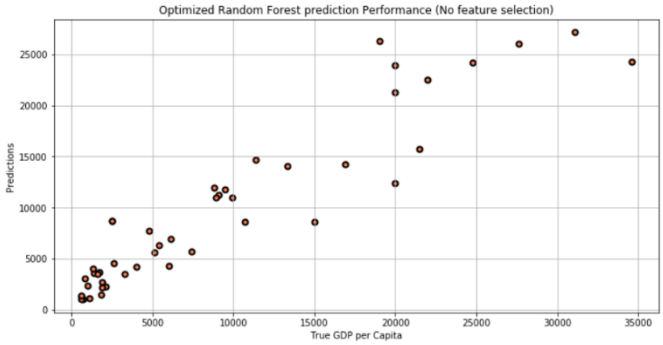

Fig-5: Optimized Random Forest Prediction Performance

Table -3: Random Forest Optimization Accuracy Performance

Data Splitting Criteria	MAE	RMSE	R2_SCORE
Optimized Random Forest prediction (No feature selection)	2356.153	3302.030	0.868

The optimization process on Random Forest has not changed the performance in a noticeable manner, yet the slight change was actually to the worst, that is probably because our initial parameters were already very close to the optimum ones.

CONCLUSIONS

The resultant study encourages the utilization of machine learning classifiers namely linear regression and Random Forest in macroeconomic data forecasting. On the basis of optimization process, the machine learning algorithm "Random Forest" utilized during this study worked well with the accuracy 86 percentage in order to predict the true GDP per capita. Random Forest Classifier produces more accurate forecasts as compared to the linear regression. Accuracy is measured by MAE and RMSE

evaluation metrics. The main focus of traditional economics models is mainly on explanations of relationships whereas machine learning classifiers target predictions. Though it may seem as Machine learning models do not turn out to be good performers while discovering the impact of independent variable on the dependent variable or analyzing a causal relationship. However, as described in this paper and in previous studies as well, machine learning models often tend to convey and exhibit high prediction power. In future this model can be improved by using better machine learning algorithm which may result in even better performance.

REFERENCES

[1] Long Gang, "GDP Prediction by Support Vector Machine Trained with Genetic Algorithm," in *2nd International Conference on Signal Processing Systems (ICSPS)*, 2010.

[2] Jaehyun Yoon, "Forecasting of Real GDP Growth Using Machine Learning Models: Gradient Boosting and Random Forest Approach," *Springer Science+Business Media, LLC, part of Springer Nature,* 2020.

[3] John Roush, Keith Siopes, Gongzhu Hu, "Predicting Gross Domestic Product Using Autoregressive Models," in *IEEE SERA*, London, UK, June 7-9, 2017.

[4] Gary L. Shelley, Frederick H. Wallace, "Inflation, money, and real GDP in Mexico: a causality analysis," *Applied Economics Letters,* vol. 11, no. 4, p. 223–225, 2004.

[5] Martin Schneider, Martin Spitzer, "Forecasting Austrian GDP using the generalized dynamic factor model," 17 September 2004.

[6] Anwar Ali Shah G.Syed, Faiz Muhammad Shaikh, "Effects of Macroeconomic Variables on Gross Domestic Product (GDP) in Pakistan," in *International Conference on Applied Economics (ICOAE)*, 2013.

[7] Maciej Woźniak, Joanna Duda, Aleksandra Gąsior, Tomasz Bernat, "Relations of GDP growth and development of SMEs in Poland," in *23rd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems*, 2019.

[8] Carlos Encinas-Ferrer, Eddie Villegas-Zermeño, "Foreign direct investment and gross domestic product growth," in *International Conference on Applied Economics, ICOAE* , Kazan, Russia, 2-4 July 2015.

[9] Luca Coscieme, Lars F. Mortensen , Sharolyn Anderson, James Ward , Ian Donohue, Paul C. Sutton, "Going beyond Gross Domestic Product as an indicator to bring coherence to the Sustainable Development Goals," *Journal of Cleaner Production,* vol. 248, 2019.

[10] Gourav Kalbalia, Vivek Tambi, "Forecasting GDP: A Linear Regression Model," *DU Journal of Undergraduate Research and Innovation,* vol. 2, no. 2, pp. 41-46, 2016.