GDP Prediction and Forecasting using Machine Learning

Tanvi Gharte[1], Himani Patil[2], Soniya Gawade[3]

[1,2,3] *Students, Dept. of Computer Science and Technology, Usha Mittal Institute of Technology, Maharashtra, India*

---***---

Abstract - *The topic of GDP has become of high importance among the indicators of economic variables. GDP prediction is a crucial job in the economy and growth analysis of a country. The goal of the paper is to give a different approach concerning the classical econometric techniques, and to show how Machine Learning techniques may improve calculating the Gross Domestic Product accurately. The GDP of countries is impacted by various social, economic, and cultural parameters. We have analysed those parameters from 1970 to 2018 and used supervised learning methods to build our models. Finally compared the performance of the model using 3 algorithms and therefore the best prediction performance is achieved by Gradient Boosting, then Random Forest and Linear Regression. And finally, the model is deployed into a web application which estimates and forecasts GDP of a country just by giving some attribute as input for that country.*

***Key Words*: Gross Domestic Product, Machine Learning, Linear Regression, Random Forest, Gradient Boosting**

1. INTRODUCTION

GDP is an important parameter to know the health and condition of a country compared to other countries. Therefore, knowing beforehand about GDP helps in knowing whether a country is progressing or its economic health is declining. Gross domestic product is a measure to assess overall economic performance of a country, it includes all products and services created by the economy as well as personal consumption, government purchase, etc.

The economic growth of the country depends on different factors like Social, Economic and Cultural Environment. In our system, we have considered parameters like population, area, population density, coastline, net migration, literacy, phones, infant mortality, arable land, crop land and other land, birth rate, death rate, region and climate to calculate the GDP per capita of the country. We have thus built a prediction model by including all such factors as independent variable and world GDP as dependent variable.

Our aim is to predict and forecast GDP per capita for a country using linear regression, random forest and gradient boosting machine learning algorithms. Prediction of GDP involves application of applied mathematics and mathematical models to predict future developments within the economy. It permits us to review previous economic movements and predict however current economic changes can modify the patterns of previous trends. Hence, more accurate prediction would provide a significant facilitation to the government in setting up economic development goals. Consequently, a correct Gross Domestic Product prediction presents a number one insight that associates an understanding for future economic trends.

2. LITERATURE SURVEY

GDP not only helps in diagnosing the economic problem but also helps in correcting the problem. Keeping all these points in consideration, we have chosen a topic of predicting GDP which can be used by any normal citizen who is willing to know the GDP of their country. Shelley G. L. and Wallace F. H (2004) [1] studied the relation between M1 money, real GDP and inflation in Mexico for the period 1944 to 1991. Co-integration relationships existed between the inflation and money hence study suggested that reduction in money growth may have been resulted in reduction of inflation in Mexico. The variations in inflation were divided into two components namely predictable and unpredictable components. In differenced inflation, predictable increases resulted in having a negative effect on GDP growth. Unpredictable increases resulted in having a positive effect on real GDP growth. This paper considered only money as an economical factor and neglected other important indicators of economic health.

The paper titled "GDP Prediction by Support Vector Machine Trained with Genetic Algorithm" was published in 2010 by Gang Long [2]. In this study, a support vector machine trained with a genetic algorithm is applied in GDP forecasting. Author concluded that the genetic algorithm can get optimal solutions in a short time, which is an excellent method in parameters selection of support vector machine. Then, a genetic algorithm is introduced to simultaneously optimize the SVM parameters. The GDP data from 1989 to 2002 used for training and 2003 to 2007 for testing. But the limitation of this project is that other Machine Learning algorithms can achieve better performance than SVM.

Another paper titled "Predicting Gross Domestic Product Using Autoregressive Models" published in 2017 by J. Roush, K. Siopes and G. Hu [3]. They used autoregressive models and then constructed a vector autoregressive model to predict GDP. The predicted result matches historical GDP data and predicts consistent future growth. Restriction of this approach is that it fails to overcome historic economic recession. This paper also didn't account for the other parameters such as trade, economic, geographical to predict the GDP growth.

Martin Schneider, Martin Spitzer [4] developed a framework for short-term forecasting of real GDP for Austria using the generalized dynamic factor model.

Vaishnavi Padmawar, Pradnya Pawar and Akshit Karande [5] predicted GDP, using linear regression and random forest. "Random Forest" utilized during this study worked well and got 86% accuracy. But this model can be improved by using better machine learning algorithms to acquire more accuracy.

In the literature discussed above, none of the work considered the holistic view of GDP dependence on Social, Economic, Geographical, Environmental impacts to predict world GDP. Most of the work focused on building mathematical or time series models. Therefore, in our proposed model we are trying to include all such comprehensive parameters. We will look at what parameters impact the GDP, what parameters are correlated.

3. METHODOLOGY

3.1 Design and Framework

In our proposed system, the objective is to build a GDP Estimation Tool with a higher accuracy ML model than the existing one. And our system can be used globally by anyone and is not restricted to any certain group of users. Fig.1 shows the machine learning approach of the implementation of the project.

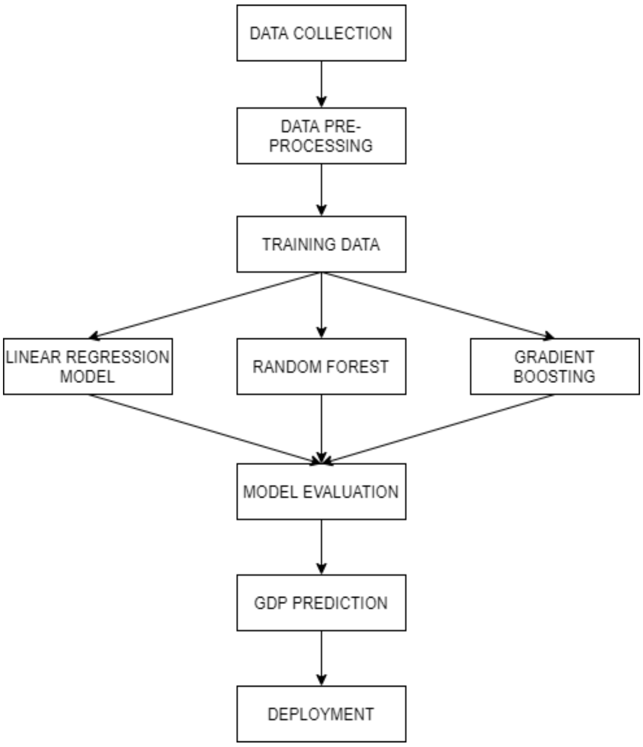


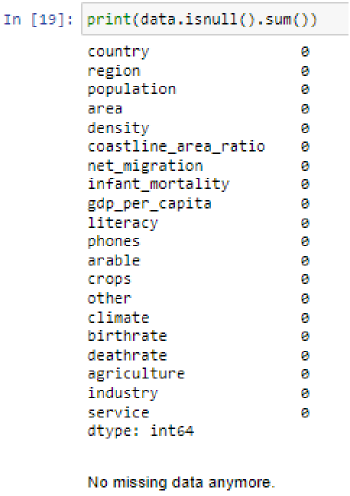
Fig -1: Architecture Diagram

A. Data Collection:

We will be using Kaggle for getting data to predict factors influencing the growth of GDP. The dataset consists of 227 countries with 20 different parameters. The parameters that are taken in consideration while predicting GDP are Literacy, Net migration, Population, Infant mortality, Agricultural economy, Industrial economy, Services economy, etc.

B. Data Pre-processing and Cleaning:

Data pre-processing is required for cleaning the data and making it suitable for a machine learning model. It helps to increase the accuracy and efficiency of a model. Identifying and removing errors and duplicate data, in order to create a reliable dataset is the main aim of data cleaning. Our dataset consists of some missing data points, but it is not extensive. 14/20 of our columns have missing data points. We have imputed some missing values from past observations. After data cleaning, there are no more missing values in the dataset.



```
In [19]: print(data.isnull().sum())

         country               0
         region                0
         population            0
         area                  0
         density               0
         coastline_area_ratio  0
         net_migration         0
         infant_mortality      0
         gdp_per_capita        0
         literacy              0
         phones                0
         arable                0
         crops                 0
         other                 0
         climate               0
         birthrate             0
         deathrate             0
         agriculture           0
         industry              0
         service               0
         dtype: int64


         No missing data anymore.
```

Fig -2: Dataset after pre-processing

Above Fig.2 shows the data set after the data pre-processing.

- Exploratory Data Analysis (EDA)

We have performed EDA that uses a range of methods to gain a deeper understanding of a data set and helps to find outliers. We have plotted the correlation heatmap which is used to visualize correlation between different features of a dataset. Given Fig. 3 presents the matrix for the correlation between the dependent variable and the independent variables. We can see that there exist expected and unexpected correlations between the parameters like strong correlation between infant mortality and birthrate, strong correlation between GDP per capita and phones and unexpected strong correlation between birthrate and phones, etc.

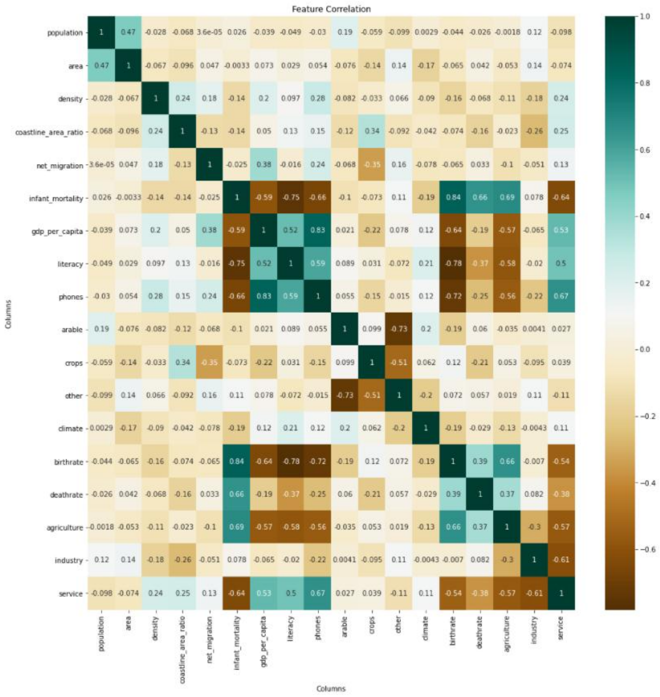


Fig -3: Exploratory Data Analysis (EDA)

C. Model Training:

The different algorithms are used to train and test the split data. We have applied Linear Regression, Random Forest and Gradient Boosting algorithms.

D. Model Evaluation:

After completion of model training, we have checked the accuracy of all the three machine learning algorithms and considered one with the highest accuracy. Performance of these algorithms is evaluated using r2 score.

E. GDP Prediction and Deployment:

Once model evaluation is done, we will be able to predict the GDP per capita for any country. And finally, the model is deployed into a web application which will estimate and forecast GDP of a country just by giving some attribute as input for that country.

3.2 Modelling

A. Linear Regression

It is a supervised machine learning algorithm that performs a regression task. Basically, it is the mathematical model that analyses the linear relationship between a dependent variable with a given set of independent variables. In the project we will use simple linear regression 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 the remaining 20% was used to test the dataset.

B. Random Forest

It is one of the well-known machine learning algorithms that belongs to the supervised learning technique category. Random Forest can be used both for regression and classification problems in machine learning. It is basically a classifier that consists of decision trees of the given dataset on numerous subsets. Further, the algorithm takes the average in order to improve the forecasting accuracy. Predictions from each 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.

Then we have first tried random forest with our data splits (With and without feature selection). Scaling is not going to be tested for Random Forest, since it should not affect this algorithm's performance. We have used grid search in order to obtain good parameters for our RF regressor. And then we optimized the parameters like n-estimators, min samples leaf, max features, bootstrap.

C. Gradient Boosting

It is a kind of machine learning algorithm that can be used to solve classification or regression predictive modelling problems. The gradient boosting model starts by creating one leaf and building regression trees. Based on the error made by the previous tree, the gradient boosting model trains another tree, and it continues to make additional trees. It will work on previous errors and boost the performance. Gradient Boosting is a method which allows us to combine all the weak models. And after the combination of various weak models, we get a single model, which will improve the accuracy of our model. In this project, we will first train the GBM regressor with the default parameter values, then we will try optimizing its parameters. The parameters we have optimized are Learning rate, n-Estimator, Min sample leaf, max depth, Min sample split, subsample, max features.

4. RESULT AND ANALYSIS

True GDP per capita was plotted against the prediction in order to evaluate the model using linear regression.

Fig. 4 shows the depiction of linear regression model for true GDP per capita prediction. True GDP per capita was plotted against the prediction in order to evaluate the model using linear regression.

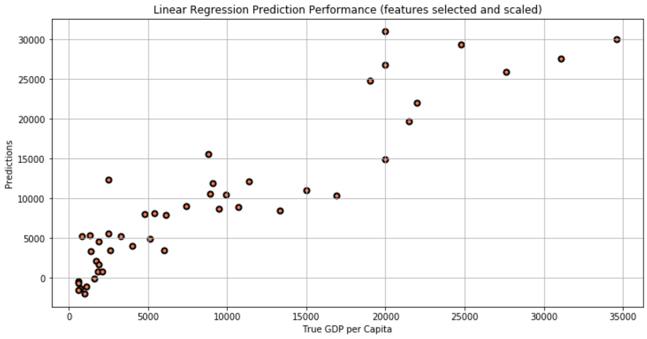


Fig -4: Linear Regression Prediction Performance

Table -1: Linear Regression Accuracy Performance

Data Splitting Criteria	MAE	RMSE	R2 SCORE
All features, No scaling	330350.858	1570337.545	-29843.12
All features, with scaling	569019.468	1283170.821	-19925.99
Selected Features, No scaling	2965.935	4088.794	0.797
Selected Features with Scaling	2879.521	3756.436	0.829

It is clear that feature selection is necessary for linear regression, in order to get acceptable results on this dataset. On the other hand, feature scaling has a small effect on LR's prediction performance. We got satisfactory prediction performance from Linear Regression with feature selection and scaling.

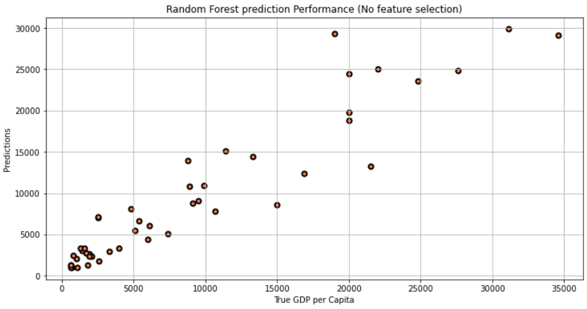


Fig -5: Random Forest Prediction Performance

Table -2: Random Forest Accuracy Performance

Data Splitting Criteria	MAE	RMSE	R2 SCORE
All features, No scaling	2142.13	3097.194	0.883
Selected Features, No scaling	2416.065	3533.59	0.848

Below diagram represents the optimization process on Random Forest regressor.

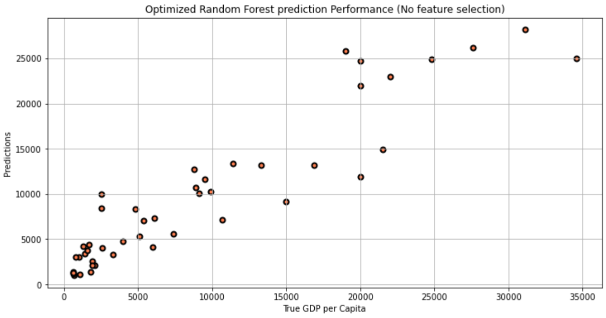


Fig -6: Optimized Random Forest Prediction Performance

Table -3: Random Forest Optimized Accuracy Performance

Data Splitting Criteria	MAE	RMSE	R2 SCORE
Optimized Performance	2360.747	3219.924	0.878

Here also we have plotted true GDP per capita against the prediction. And it is clear that Gradient Boosting gave us nearly same performance as that of random forest.

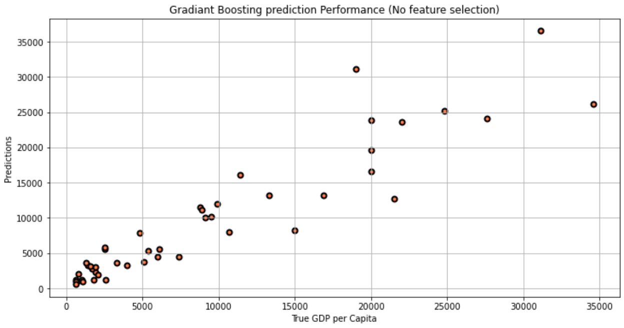


Fig -7: Gradient Boosting Prediction Performance

Table -4: Gradient Boosting Accuracy Performance

Data Splitting Criteria	MAE	RMSE	R2 SCORE
All features, No scaling	2280.462	3413.635	0.858
Selected Features, No scaling	2467.208	3789.297	0.826

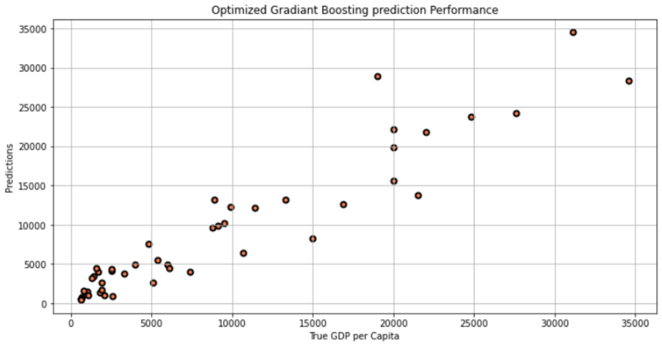


Fig -8: Optimized Gradient Boosting Prediction Performance

Table -5: Gradient Boosting Optimized Accuracy Performance

Data Splitting Criteria	MAE	RMSE	R2 SCORE
Optimized Performance	2362.935	3469.360	0.889

5. CONCLUSIONS AND FUTURE WORK

We explored all the supervised regression models in order to get the best fitting models. We have trained the model using Linear Regression, Random Forest and Gradient Boosting machine learning algorithms and also estimated the performance of these models. Evaluation is done using MAE and RMSE techniques and then compared all three models to get a clear overview of performance. The accuracy obtained by linear regression algorithm is 82% and by random forest is 87%. On the basis of the optimization process, the machine learning algorithm "Gradient Boosting" utilized during this project worked well with the accuracy 89% in order to predict the true GDP per capita. Finally deployed the highest accuracy model to "GDP Estimation Tool" which estimates and forecasts GDP of a country just by giving some attribute as input for that country.

REFERENCES

[1] Gary L Shelley and Frederick H Wallace. Inflation, money, and real gdp in mexico: a causality analysis. Applied Economics Letters, 11(4):223–225, 2004.

[2] Gang Long. Gdp prediction by support vector machine trained with genetic algorithm. In 2010 2nd International Conference on Signal Processing Systems, volume 3, pages V3–1. IEEE, 2010.

[3] J. Roush, K. Siopes and G. Hu, "Predicting gross domestic product using autoregressive models," 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA), 2017, pp. 317-322.

[4] Martin Schneider, Martin Spitzer, et al. Forecasting austrian gdp using the generalized dynamic factor model. Technical report, 2004.

[5] Vaishnavi Padmawar, Pradnya Pawar, and Akshit Karande. Gross domestic product prediction using machine learning.