RAMAIAH ISSN (Print) : 0975-7988

MS RAMAIAH MANAGEMENT REVIEW
January - March 2025, Vol 16(01), 39 – 50
DOI: 10.52184/msrmr.v16i01.099

Forecasting India's GDP Growth Using Sectoral Data: A Comparative Study of Statistical Models and Machine Learning Approaches

Gayatri Sah and Swet Chandan

Department of Economics,
M S Ramaih University of Applied Sciences, MSR Nagar, Bangalore
Gayatri.sah13@gmail.com

Article Type: Research Article

Article Citation: Gayatri Sah and Swet Chandan, Forecasting India's GDP Growth Using Sectoral Data: A Comparative Study of Statistical Models and Machine Learning Approaches. M.S. Ramaiah Management Review. 2025; 16(01), 39-50. DOI: 10.52184/msrmr.v16i01.099

Received date: October 10, 2024

Accepted date: December 10, 2024

***Author for correspondence:**
Swet Chandan swetchandan@gmail.com Professor, School of Computer Science and Application, D Y Patil International University, Akurdi, Pune

Abstract

India is the world's seventh-largest economy by nominal GDP and the third largest by purchasing power parity (PPP), this has been possible due to the important transformations in the economic development. This paper showcases a complete analysis of India's GDP by focusing on its two critical periods: the pre-liberalization era (1947-1991) and the post-liberalization era (1991-2008). Each of which is signified with its own distinct economic methods.

Over time, India's economy has observed a prominent shift. In the pre-liberalization period, a large share of GDP is because of agriculture. However, in the post-liberalization period, it decreased significantly, whereas the services sector, especially the IT and finance expanded. This played a key role in the rapid economic growth, making India a global leader in the IT sector.

Several machine learning and statistical models such as Ordinary Least Squares (OLS) regression for analyzing linear relationships, Multilayer Perceptrons (MLP) and Autoregressive Integrated Moving Average (ARIMA) models are used to analyse India's GDP growth.

Gradient Boosting, Elastic Net, and Random Forest techniques are also utilized to enhance the accuracy of predictions and to better understand the dynamics of India's GDP.

Keywords: GDP Growth, Sectoral Data, Statistical Models, Machine Learning Approaches

I. INTRODUCTION

Today, the current landscape requires that policy makers, financial institutions as well as market participants be able to make time sensitive decisions in real time, sometimes even when they have limited or late information available to them. It is worth noting that GDP data, an economic indicator that is among the primary focuses of any economy, are subject to release delays, a wider concern in Emerging Market Economies like India, which reports it's GDP figures with a two month lag. This lag creates uncertainty amongst the policymakers, banks and other financial participants as they deal with critical time sensitive decisions in the absence of the most current information of

© 2021 The Author(s). This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

the economy. As a solution to this problem, nowcasting has focussed on shorter-horizon forecasts, and, more generally, tries to measure or analyse the present, near future or very recently concluded periods of economic time series (Banbura et al, 2010).

Nowcasting models have mechanisms to enhance their forecasting abilities by employing high-frequency data that is accessible prior to the stated metrics being officially announced. Target variables with high-frequency data are hardly used in economic forecasting, predominantly due to traditional time series and bridge models which are ready available and quite popular, however, these types of models tend to be constrained by the assumptions regarding functional forms and distribution. Recent years have also seen an improvement in the success of nowcasting by machine learning (ML) approaches due to their ability to effectively manage large datasets and a range of different types of data sets, hence applying ML techniques in economic prediction is justified (Richardson et al., 2021)

The underlying difficulty of nowcasting GDP in EMEs is now the availability of high-frequency data regarding various economic sectors. Specifically with reference to India, earlier studies employed the bridge equation as well as Dynamic Factor Model (DFM) in estimating GDP growth (Bhattacharya et al., 2011; Iyer & Gupta, 2019). Heavily reliant upon high-frequency indicators, these models are nevertheless hamstrung by the availability and qualitative needs of the data. This study aims to add to the existing literature by examining the combination of traditional econometric models and machine learning techniques in forecasting the India's GDP more precisely.

Specifically, our research expands the potential of machine learning techniques of Random Forest, Gradient Boosting, and Multilayer Perceptron (MLP) with Ordinary Least Squares (OLS) and other regression-based techniques. Employing a hybrid methodology enables us harness the advantages of both conventional and ML models and provides a greater degree of certainty and control in our forecasts. Furthermore, we use sectoral data with reference to agriculture, manufacturing, transport, and mining, thus giving a broader economic view and at the same time aiding the comprehension of the GDP growth geological wrinkles.

II. LITERATURE

The time-series approach has administered the analysis of economic growth forecasting, but the recent trends in incorporating machine learning offer a new perspective to global GDP dynamics. Economies that are stable and forecastable have absorbed traditional techniques such as ARIMA and ordinary least squares regression in a bid to study and predict linear dependencies on key economic variables. However, the trends are changing due to increasing data complexity, especially in emerging economies such as India, rendering these models highly simplistic and driving experts to seek better data analysis methods.

The past several decades have also witnessed a noticeable improvement in using Machine learning (ML) methods to complement conventional GDP estimating techniques as they have enabled the incorporation of non-linear relationships with high dimensionality. Studies have shown

that traditional AR and regression models often struggle to capture the complex relationships between economic indicators, whereas more advanced methods like Gradient Boosting and Random Forest are better suited for this task. Barnett et al. (2016) highlighted this limitation, noting that while basic models can handle sectoral data, they lack the depth needed for more intricate economic forecasting. Neural networks, particularly Multi-Layer Perceptron (MLP) models, have shown strong accuracy in GDP forecasting, especially in more dynamic economic settings (Bhattacharya et al., 2011).

Recently, there's been significant progress in using machine learning for GDP forecasting, especially in advanced economies. Researchers are increasingly turning to dynamic factor models (DFMs) to improve nowcasting and forecasting accuracy, as seen in the work of Banbura et al. (2013) and Giannone et al. (2008). High-frequency indicator models have also proven particularly effective for forecasting in complex economies like Germany (Marcellino & Schumacher, 2010), France (Barhoumi et al., 2011), and the USA (Lahiri & Monokroussos, 2011).

According to the new publications for India, the machine learning tools can be effectively used to model the transitions in the country's economy. Bhadury et al. (2021) and Bragoli & Fosten (2018), for example, have developed such models that simultaneously use several HFIs and combine them with GDP nowcasting in order to overcome data-lagging and data frequency issues. These studies highlight the relevance of applying models incorporating both time-series financial and cross-section agriculture indicators such as rainfall deviation which is majorly affected India due to strong reliance on rain-fed agriculture (Iyer & Gupta, 2019). Bragoli and Fosten (2018) go a step further in suggesting the employment of the nominal and the international series as proxies in bridging the gaps created by the lack of data concerning the Indian economy.

Once these challenges are noted, our study then extends this focus to include not only the traditional but also more complex machine learning approaches with the goal of forecasting India's GDP paying attention to problems that emerging economies face like the selection of appropriate indicators, rough edges in economic HFIs, and data frequency mismatch between quarterly time series of GDP and monthly time series of indicators. Our intention is to enhance the Indian forecasting framework by these findings which include ideas pertaining to complexity of the Indian economy and its implications on economic forecasting and policymaking in other emerging markets.

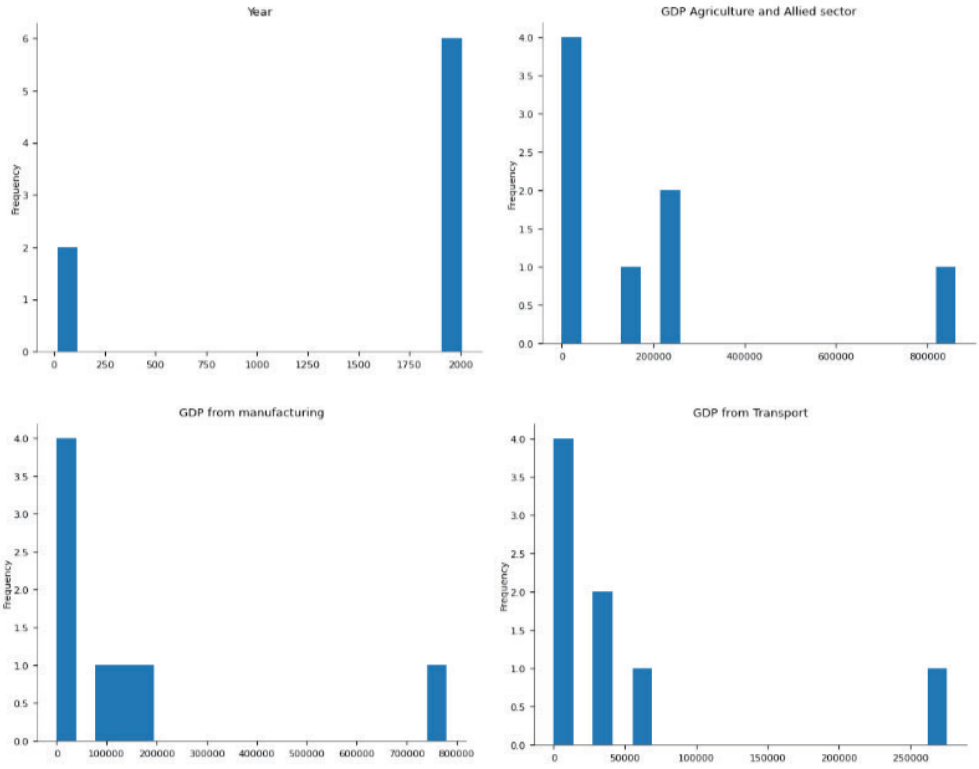
III. DATA

Year	GDP Agriculture and Allied sector	GDP from manufacturing	GDP from Transport	GDP from mining & quarrying	GDP Total
1950	5080	1056	327	75	6537
1951	5245	1171	370	84	6871
1952	5110	1103	362	86	6660
1953	5630	1243	377	87	7337
1954	4789	1289	399	91	6568
1955	4644	1325	427	93	6489
1956	5900	1578	480	113	8070
1957	0	1680	567	131	2378
1958	0	1794	621	140	2555
1959	0	2009	629	149	2787
1960	7090	2339	678	177	10284
1961	7343	2580	749	187	10859
1962	7497	2865	855	224	11441
1963	8823	3267	928	243	13261
1964	10781	3593	999	256	15628
1965	10751	3829	1081	296	15957
1966	12506	4188	1189	321	18205
1967	15650	4475	1316	374	21815
1968	16132	4843	1489	399	22863
1969	17644	5578	1588	444	25254
1970	18192	6088	1671	465	26416
1971	18584	6754	1814	487	27639
1972	20440	7505	2028	525	30497
1973	26936	9081	2110	600	38727
1974	29509	11670	2551	856	44586
1975	29249	12139	2989	1088	45465
1976	29882	13432	3592	1245	48151
1977	35380	15046	3964	1376	55765
1978	36361	17326	4535	1496	59717
1979	37616	19840	4649	1867	63972
1980	47312	22159	4978	2327	76777
1981	53327	26062	6238	4150	89777

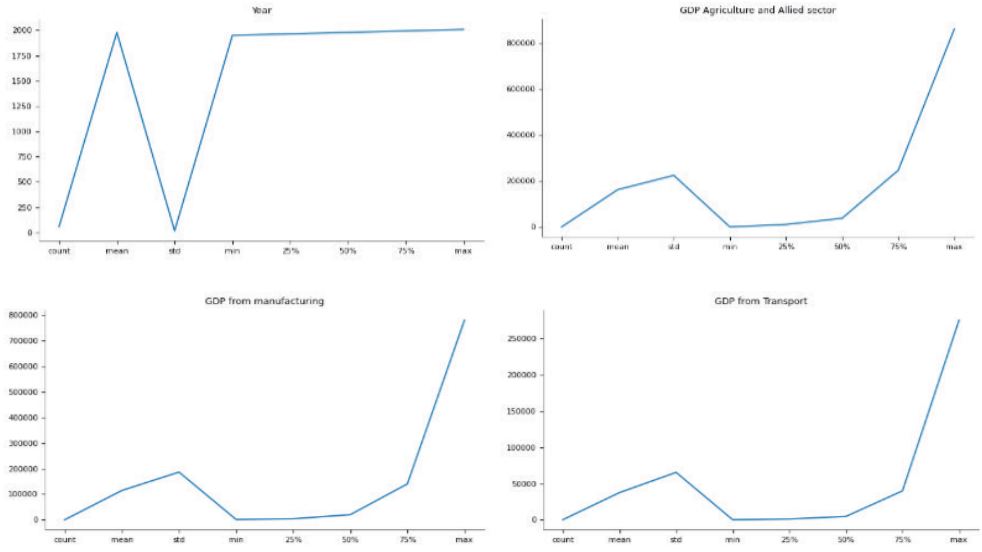
The dataset contains 59 rows x 6 columns

- **Year:** The specific year for which GDP data is recorded.
- **GDP Agriculture and Allied sector:** Value of goods and services produced by agriculture, including crops, livestock, forestry, and fishing.
- **GDP from manufacturing:** Value of goods produced by manufacturing industries like machinery, electronics, and textiles.
- **GDP from Transport:** Value of services provided by transportation industries such as air, rail, road transport, and logistics.
- **GDP from mining & quarrying:** Value of minerals, ores, crude petroleum, and natural gas extracted from the earth.
- **GDP Total:** Sum total of economic output from all sectors, reflecting the overall GDP for the year.

Distributions



Values



IV. METHODOLOGY

MODELS:

A. **Ordinary Least Squares (OLS)**

Ordinary Least Squares estimates model variables by minimizing sum of squared differences between observed and predicted values. With Key assumptions of linearity, independence, homoscedasticity and normality, it creates a linear relationship between dependent variable (Y) and independent variable (X).

B. **Multilayer Perceptron (MLP)**

Multilayer Perceptron is a type of ANN composed of multiple layers of nodes(perceptron), organised in input, one or more hidden and output layer.

The key components are:

- **Input Layer**
  Each perceptron in Input layer, represents a feature of the dataset.
- **Hidden Layer**
  Here the model learns complex pattern in the data.
- **Output Layer**
  It generates the final predictions.

Working

- **Feedforward:** Input data includes features for a neural net, propagates through all the layers of the net, where each node of the net performs a weighted summation of its inputs followed by the addition of some bias followed by an activation function.
- **Backpropagation:** MLP employs backpropagation for the modification and update of weights and biases. Its working entails making a prediction and then getting a corresponding actual value which produces an error also called the "loss." This loss is then sent backwards through the models, updating the weights accordingly to minimise the loss.

C. **Gradient Boosting**

Gradient Boosting is an algorithm which combines weak learners into strong learners. Using gradient descent, each model is trained to minimize loss function (mean-square error or cross entropy of previous model). To improve accuracy, each model is fitted to residual of current model. This process is incremental. It can be mathematically represented as:

$$G_m(x) = G_{m-1}(x) + \nu\Delta_m(x)$$

Here, $G_m(x)$ = Updated model; $G_{m-1}(x)$ = Previous Model; $\Delta_m(x)$ = weak learner and ν = learning rate (regularization factor)

Generalization is improved by lower learning rate. In this case, learners are regression trees utilizing the LS_TreeBoost algorithm, Friedman (2001).

D. **Ridge, Lasso Regression and Elastic Net**

In case of large number of features in a dataset, Ridge, Lasso Regression and Elastic Net are commonly used models for minimizing model complexity.

- **Ridge Regression**
  Ridge Regression (L2 regularization) improves OLS by using a penalty term, comprising of the tuning factor multiplied by squared sum of magnitude of coefficient values. It can be represented as:

$$\propto = argmin[\sum_{i=1}^{l}(y_i - \propto_0 - \sum_{j=1}^{p} x_{ij} \propto_j)^2 + \lambda \sum_{j=1}^{p} \alpha_j^2]$$

Here,

hyperparameter determining the severity of the shrinkage applied

This helps reduce the model's sensitivity to predictor variable changes and provides mitigation of multicollinearity by penalizing large coefficients.

- **Lasso Regression**

  Lasso Regression (L1 regularization) creates sparse models by adding a penalty term to the coefficients equal to their absolute values. Lasso Regression can be used for feature selection and generation of sparse models as some of the coefficients of the variables can be set to zero during the regularization process. It can be represented as:

$$\propto = argmin[\sum_{i=1}^{l}(y_i - \propto_0 - \sum_{j=1}^{p} x_{ij} \propto_j)^2 + \lambda \sum_{j=1}^{p}|\alpha_j|]$$

- **Elastic Net**

  Elastic Net is hybrid version of both ridge and lasso regressions, as it involves convex sum of penalties of both regressions and relative weights of the two penalties are determined by tuning variable . The equation is as follows:

  It deals easily with both multicollinearity and feature selection; hence it is used commonly with datasets where both feature simultaneously.

$$\propto = argmin[\sum_{i=1}^{l}(y_i - \propto_0 - \sum_{j=1}^{p} x_{ij} \propto_j)^2 + \lambda \sum_{j=1}^{p}(1 - \beta)\alpha_j^2 + (\beta)|\alpha_j|]$$

E. **Random Forest**

Random Forest is a supervised ML algorithm constructed from decision tree algorithms. It uses different dataset samples and creates multiple decision trees. For regressions, it either used a majority vote or average of the samples.

F. **ARIMA model**

ARIMA (AutoRegressive Integrated Moving Average) is widely used for analyzing and forecasting time series data. Any data which displays certain level of trend or seasonality integrates ARIMA model.

ARIMA model is defined with three main parameters:

1. **AutoRegressive(AR):** It refers to relation between a current observation and a certain number of previous observations. **p** represents number of previous observations.
2. **Integrated(I):** It remove the trends and seasonality to make the series stationary by differencing. **d** indicates the number of times the data values are differenced to achieve stationarity.
3. **Moving Average(MA):** It represents the relation between observation and residual error from a moving average model to previous observations. **q** represents the size of moving average window.

For seasonal time series, a variant called SARIMA (Seasonal ARIMA) includes seasonal components.

V. RESULTS

In this study, using several machine learning models we tried to forecast India's GDP. Models like Ordinary Least Squares (OLS), Ridge Regression, Lasso Regression, Elastic Net, Random Forest, Gradient Boosting, and Multi-Layer Perceptron (MLP) were trained and tested on the historical GDP data and its predictions were compared to the actual values to assess accuracy.

We achieved a remarkable accuracy of 100% using OLS and Ridge regression model which perfectly aligned with the current GDP values present in dataset. This proposes that OLS and Ridge models can help in effectively capturing the fundamental trend in the growth of GDP based

on the features selected. Also, other models such as Elastic Net, Lasso Regression, and Gradient Boosting attained an accuracy of 99%. Due to these remarkable results, it made it possible to indicate the robustness and reliability of the models in GDP prediction.

Figure 1 displays the actual GDP values alongside the predicted values obtained from each model. As observed, only a slight deviation is seen in a few cases across different time points of actual GDP values. Especially, in most instances, the actual values closely align with predictions from the Gradient Boosting and Random Forest models.

Compared to OLS and Ridge regression, Multi-Layer Perceptron (MLP) while less precise still achieved great results in demonstrating the neural network models potential in economic forecasting tasks. Still along with the minimal difference in the accuracy between MLP and OLS and Ridge Regression, MLP with added complexity may not be able to deliver remarkable advantages in this application.

In summary, the effectiveness of both the linear and ensemble models in GDP forecasting can be validated with each model offering its unique strengths. In terms of accuracy and computational efficiency, OLS and Ridge Regression are highly accurate, whereas Gradient Boosting and Random Forests provide strong predictions by using the power of ensemble learning.

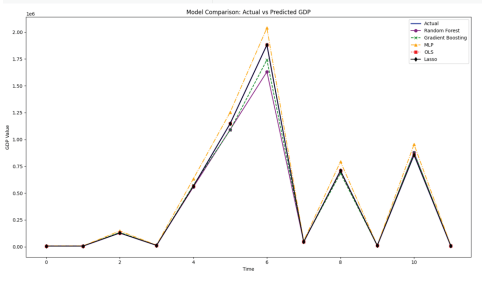


FIGURE 1

To better understand the interdependencies among different economic sectors and their impact on the overall GDP, we computed a correlation matrix, as shown in Figure 2. The matrix highlights the pairwise correlations between GDP from various sectors, including Agriculture and Allied, Manufacturing, Transport, and Mining & Quarrying, as well as the total GDP.

The results indicate that most sectors have a high positive correlation with the total GDP, suggesting that changes in sectoral GDP values strongly affect the overall GDP. Notably:

The Agriculture and Allied sector is an important contributor to the overall economic performance as there is a very high correlation with the total GDP (0.99), also strong interlinkages are indicated among the sectors like Mining & Quarrying (0.98).

There is a high correlation with the total GDP (0.99) as seen in Manufacturing sector and other sectors, mainly in Agriculture and Mining (0.99 and 1.0 approx. respectively). This strong relationship implies that growth in manufacturing likely drives or is driven by other key sectors.

The Transport sector shows a moderately high correlation with total GDP (0.89), though it is relatively less correlated with other sectors compared to Agriculture and Manufacturing. This suggests that while Transport plays an important role, its contributions to GDP might be more independent.

The Mining & Quarrying sector also exhibits a high correlation with total GDP (0.99) and very strong connections with Manufacturing and Agriculture (both approximately 1.0), indicating this sector's close relationship with primary and secondary production activities.

These high correlation values suggest that certain sectors, particularly

Agriculture, Manufacturing, and Mining & Quarrying, are highly interdependent and collectively contribute significantly to the GDP. Therefore, policy changes or growth in these sectors are likely to have substantial impacts on the overall economy.

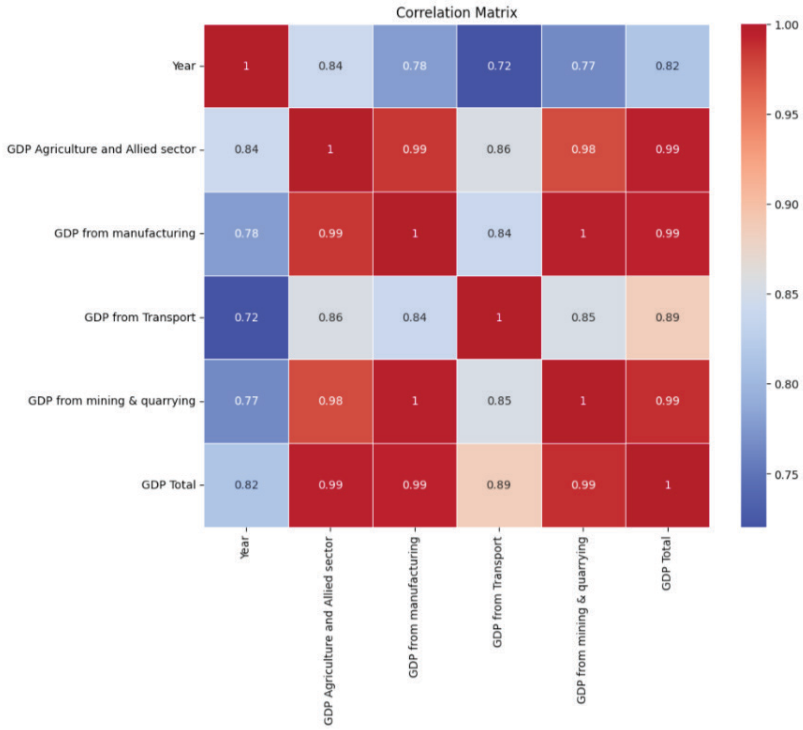
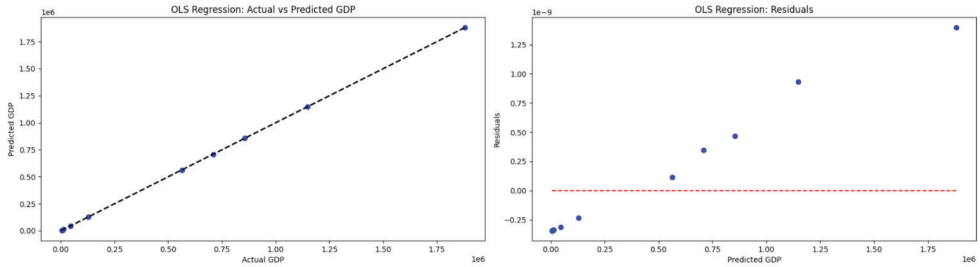
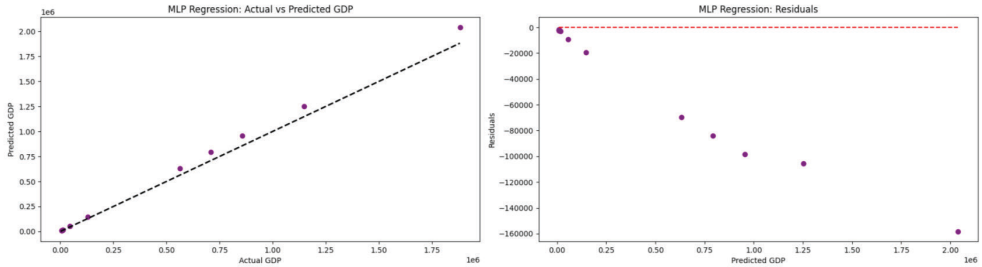


FIGURE 2

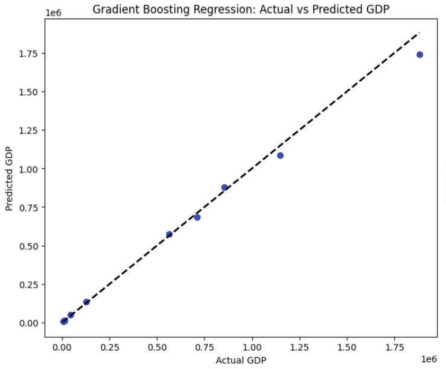
OLS Result

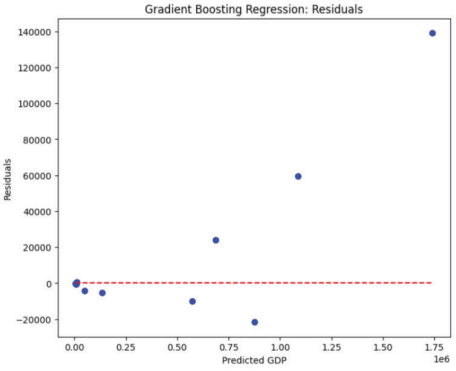


MLP Result

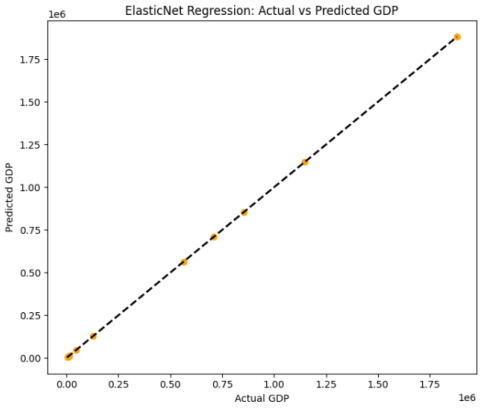


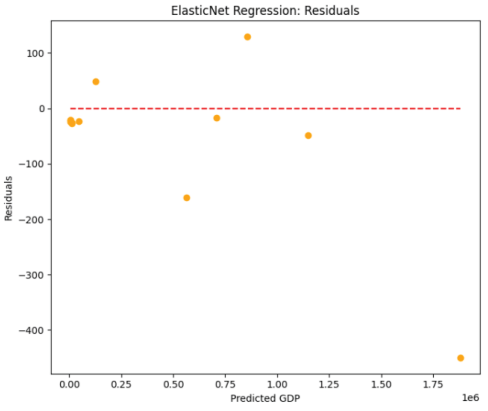
Gradient Boosting



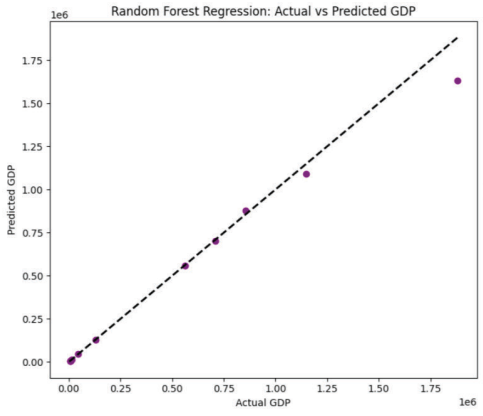


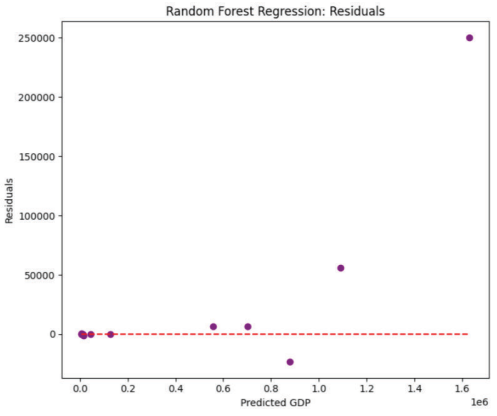
Elastic Net



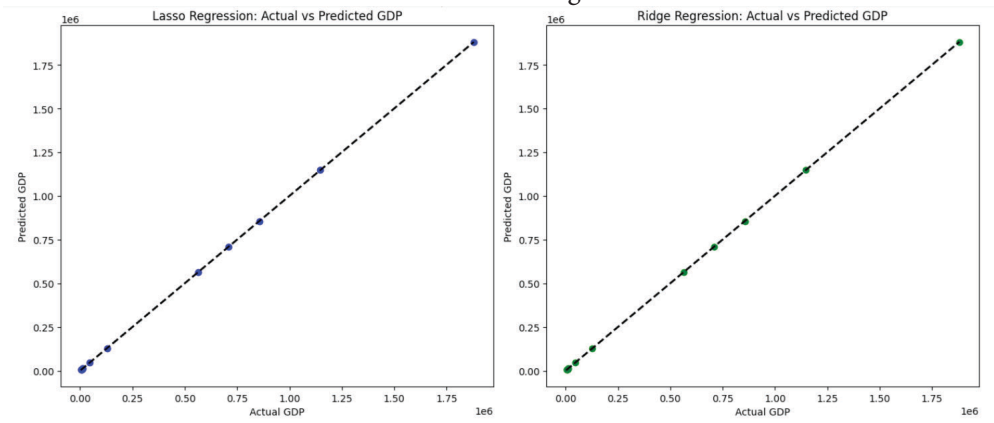


Random Forest

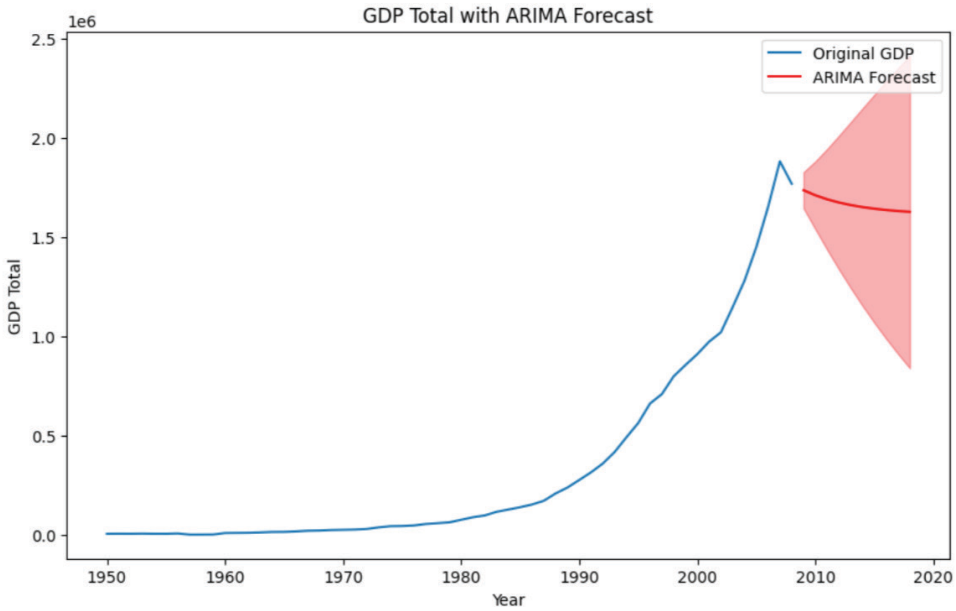




Lasso and Ridge



ARIMA



VI. CONCLUSION

In this paper, we have compared the forecasting performance between the traditional statistical models and modern machine learning algorithms for predicting India's GDP growth keeping in factor the unique economic shifts before and after liberalization. We have evaluated the predictive capabilities of various models like OLS, regression, ARIMA, MLP, Gradient Boosting and Random Forest with the sectoral data of agriculture, industry and various services. Our findings revealed that there are notable improvements in forecasting accuracy when using machine learning technique compared to traditional statistical methods.

This comparative analysis suggested that we can forecast GDP using various machine learning approaches and can help policymakers by providing more precise and timely insights into economic trends. This can support better policy formulation specifically in dynamic economic context where traditional models may be limited. Our results highlight the importance of machine learning to understand economic changes which will surely benefit India's ongoing economic planning and growth strategies.

VII. REFERENCES

Aastveit, K., & Trovik, T. (2012). Nowcasting Norwegian GDP: The role of asset prices in a small open economy. *Empirical Economics, 42*(1), 95–119. https://doi.org/10.1007/s00181-010-0425-6

Angelini, E., Camba-Mendez, G., Giannone, D., Reichlin, L., & Rünstler, G. (2011). Short-term forecasts of euro area GDP growth. *The Econometrics Journal, 14*(1), C25–C44. https://doi.org/10.1111/j.1368-423X.2010.00328.x

Banbura, M., Giannone, D., & Reichlin, L. (2010). Nowcasting. *European Central Bank Working Paper Series, No. 1275.* https://doi.org/10.2139/ssrn.1660460

Bhadury, S., Ghosh, S., & Kumar, P. (2021). Constructing a coincident economic indicator for India: How well does it track gross domestic product? *Asian Development Review, 38*(2), 237–277. https://doi.org/10.1162/adev_a_00154

Bhattacharya, R., Pandey, R., & Veronese, G. (2011). Tracking India's growth in real time. *National Institute of Public Finance and Policy, Working Paper No. 11/90.*

Bragoli, D., & Fosten, J. (2018). Nowcasting Indian GDP. *Oxford Bulletin of Economics and Statistics, 80*(2), 259–282. https://doi.org/10.1111/obes.12194

Cepni, O., Guney, I. E., & Swanson, N. R. (2019). Nowcasting and forecasting GDP in emerging markets using global financial and macroeconomic diffusion indexes. *International Journal of Forecasting, 35*(2), 555–572. https://doi.org/10.1016/j.ijforecast.2018.09.002

Iyer, T., & Gupta, A. S. (2019). Nowcasting economic growth in India: The role of rainfall. *Asian Development Bank Economics Working Paper Series, No. 593.* https://doi.org/10.22617/WPS190336-2

Marcellino, M., & Schumacher, C. (2010). Factor MIDAS for nowcasting and forecasting with ragged-edge data: A model comparison for German GDP. *Oxford Bulletin of Economics and Statistics, 72*(4), 518–550. https://doi.org/10.1111/j.1468-0084.2010.00588.x

Zhemkov, M. (2021). Nowcasting Russian GDP using a forecast combination approach. *International Economics, 168,* 10–24. https://doi.org/10.1016/j.inteco.2021.02.001