Predicting India's GDP Using Linear Regression and Random Forest Methods

A Debasish Patra, Ashutosh Pattnaik, Dr. Manjushree Nayak

*Department of Computer Science and Engineering, NIST Institute of Science and Technology (Autonomous), Berhampur Odisha- 761008*

# **ABSTRACT**

Machine learning is still a relatively new field in economics. Furthermore, there is still a shortage of knowledge about machine learning. What machine learning is, how it differs from conventional econometrics, and how analysts and organizations may best use it. Our research aims to evaluate prediction models employing modern ML methods. We test if ML algorithms can enhance performance and identify characteristics that promote economic recovery or forecast economic recessions. Some indicators can provide insight into the status of the economy while others cannot. In this paper, we will study India's GDP and make predictions with the help of Machine Learning.

**Keywords:** Economics, GDP, Machine Learning, Forecasting, Linear Regression, Random Forest Regressor.

# **1. INTRODUCTION**

In the dynamic landscape of India's economic growth, accurate forecasting of the Gross Domestic Product (GDP) stands as a linchpin for informed decision-making, strategic planning, and policy formulation. As the nation experiences rapid economic evolution, the utilization of advanced predictive modeling techniques becomes imperative to navigate the complexities inherent in shaping its financial trajectory. This study undertakes a comprehensive exploration into the realm of economic forecasting, employing two powerful methodologies:

* Linear Regression (LR)
* Random Forest (RF)

India, characterized by its diverse economic drivers and resilient growth, demands sophisticated forecasting tools that can capture the multifaceted influences shaping its economic landscape. Linear Regression, a conventional statistical approach, has historically provided insights into linear relationships between GDP and key indicators. However, in an era marked by unprecedented data volumes and intricate non-linear relationships, the integration of advanced machine learning techniques, such as Random Forest, becomes essential. This study seeks to discern the comparative effectiveness of these methodologies in providing nuanced insights into India's economic dynamics.

Economic measures in India play a pivotal role not only in steering commercial enterprises but also in guiding governmental decisions. The accuracy of economic forecasting emerges as a critical factor for business executives, policymakers, and investors alike, influencing the quality of strategic choices. As a result, the ability to make precise projections about economic measures becomes an increasingly significant aspect of navigating the intricacies of India's economic landscape.

The formula governing India's GDP is expressed as:

**GDP = C + I + G + (X- M)**

where various components contribute to the nation's economic output. Consumption (C) reflects personal consumer expenditure, Gross Private Domestic Investment (I) represents private sector investments, Government Spending (G) denotes public sector expenditures, Total Exports (X) and Total Imports (M) contribute to the balance of trade, and the difference between exports and imports (X-M) signifies total net exports.

India's economic composition encompasses agriculture, handicrafts, industrial plants, and various other sectors. Although two-thirds of the population is engaged in agriculture, the service sector currently propels economic development. Notably, India has emerged as a global leader in information technology, driven by a large pool of educated individuals proficient in English. The nation has undergone significant economic reforms since the early 1990s, gradually opening its markets to international commerce and investment by reducing government control. This period marked the ascent of India's economy on the global stage.

Economic forecasting in India extends beyond capturing historical patterns; it involves deciphering the influence of global and domestic factors on the nation's economic health. With agriculture, services, and manufacturing contributing significantly to the GDP, the task becomes inherently complex. Moreover, India's economic ties to global markets necessitate forecasting models that can accommodate a wide array of variables and dynamic influences.

In the pursuit of understanding and enhancing the forecasting process, this research transcends conventional econometrics and embraces the capabilities of machine learning. By applying Linear Regression and Random Forest Regression, the aim is to unravel the potential of these methodologies in predicting India's GDP. This exploration not only contributes to the academic discourse on economic forecasting but also holds practical implications for those steering economic policies, business strategies, and investments in the vibrant and evolving economy of India.

# **2. LITERATURE SURVEY**

The assessment of a country's economic stability often hinges on the trajectory of its Gross Domestic Product (GDP). Various research endeavors have explored diverse methods to predict economic growth.

In the study titled "Forecasting of Real GDP Growth Using Machine Learning Models: Gradient Boosting and Random Forest Approach" by Jaehyun Yoon, machine learning models, specifically gradient boosting and random forest models, are employed to forecast real GDP growth in Japan from 2001 to 2018. The forecasts are compared with those from the International Monetary Fund and the Bank of Japan, serving as benchmarks. To enhance prediction accuracy, a cross-validation process is utilized to optimize hyperparameters. Evaluation metrics such as mean absolute percentage error and root squared mean error gauge the accuracy of the forecasts. The random forest model employs bootstrapped data and averages the output of independently trained regression trees to generate predictions.

In the paper titled "Gross Domestic Product Prediction using Machine Learning" by Vaishnavi Padmawar, Pradnya Pawar, and Akshit Karande, it is highlighted that inflation variability exerts a negative influence on real GDP growth. Machine learning classifiers like random forest and linear regression are advocated for macroeconomic data forecasting. Unlike traditional economic models primarily focused on explaining relationships, machine learning models excel at producing accurate predictions. The authors leverage a methodology involving a linear regression model followed by random forest to achieve optimal accuracy.

# **3. RELATED WORKS**

Figure:1.1 Machine-Learning Process

## 3.2 ALGORITHM

### 3.2.1 Linear Regression

Linear regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The goal of linear regression is to find the best-fitting line that minimizes the sum of the squared differences between the observed and predicted values.

The linear regression equation is typically represented as:

[ Y = mX + b \]

Here, \(Y\) is the dependent variable, \(X\) is the independent variable, \(m\) is the slope of the line, and \(b\) is the y-intercept.

The process of linear regression involves estimating the values of \(m\) and \(b\) based on the given data. This is often done using the method of least squares, which minimizes the sum of the squared residuals (the differences between observed and predicted values).

Linear regression is widely used in various fields, including economics, finance, biology, and engineering, for tasks such as predicting trends, understanding relationships between variables, and making forecasts. Simple linear regression involves one independent variable, while Linear Regression can handle multiple independent variables. The simplicity and interpretability of linear regression make it a fundamental tool in statistical modeling and analysis.

### 3.2.2 Random Forest

Random Forest (RF) is a powerful machine learning algorithm employed for predictive modeling and analysis. In the context of forecasting the Indian Gross Domestic Product (GDP), RF offers a sophisticated approach to capture intricate relationships among economic variables. Unlike traditional statistical methods like LR, RF is a non-linear ensemble learning technique known for its flexibility and effectiveness, making it well-suited for complex and dynamic economic landscapes.

The essence of RF lies in its ability to build an ensemble, or a "forest," of decision trees. Each tree independently contributes to the overall prediction, and the final forecast is an aggregate of these individual predictions. This ensemble approach enhances predictive accuracy and generalizability, making RF particularly adept at handling non-linear relationships, interactions, and capturing patterns that might elude linear models.

The key features of RF include:

1. Non-linearity and Flexibility:

RF does not assume a linear relationship between the dependent variable (GDP) and the independent variables. It can capture complex, non-linear patterns within the data, providing a more nuanced understanding of the economic dynamics in India.

1. Variable Importance:

RF can automatically assess the importance of each independent variable in predicting the GDP. This feature is invaluable for identifying which economic indicators significantly contribute to the forecast.

1. Robustness and Adaptability:

RF exhibits robustness against outliers and noise in the data. It adapts well to changing conditions and is less prone to overfitting, a common concern in predictive modeling.

1. Handling Multicollinearity:

RF can effectively handle multicollinearity; a situation where independent variables are correlated. This is crucial in economic forecasting where certain economic indicators may be interrelated.

While RF has its strengths, it is essential to recognize its limitations, such as loss of structure in data, overfitting of noisy data, computational intensity, lack of interpretability

# 4. METHODOLOGY

The whole approach is depicted by the following flowchart:

Figure:4.1 Methodology

## 4.1 Data Collection

The first step is to determine what sort of data is needed to address a specific problem and whether or not there are any privacy problems associated with the data. It is the initial stage in the process of developing a machine learning model. To conduct our research, we used World Bank [5] open-source data.

## 4.2 Data Pre-processing

Big data requires cleaning and preparation before feeding into the model.

This involves:

1. Data Cleaning: Removing inconsistencies, duplicates, and errors.
2. Missing Value Handling: Strategies like imputation or deletion can be used.
3. Feature Engineering: Creating new features that might be helpful for training.
4. Data Normalization: Scaling numerical features to a common range.

## 4.3 Data Visualization

In data visualization, we can observe how the data appears and what type of association exists between its many characteristics. It's the quickest approach to verify if the attributes match the output. We've utilized Python packages like Matplotlib and Seaborn to show data in an appealing way.

## 4.4 Data Analysis

For GDP prediction, we have made use of Linear Regression algorithm and Random Forest Regression algorithm.

## 4.5 Model Training

When developing algorithms, model training is the stage of the data science development lifecycle in which researchers attempt to find the optimal mix of weights and biases for the methodology to limit a loss function throughout the predicted range. There are 70% training and 30% testing datasets in our research.

## 4.6 Model Evaluation

Model evaluation is the method of assessing a machine learning model's progress and identify any flaws. Our model compares predicted data to actual data. Among the metrics we utilized for our assessments were R-squared and Root Mean Squared Error on Prediction (RMSE/RMSEP).

# 5. RESULT

This study presents a method for predicting the annual GDP of India from 1960 to 2020. The prediction includes features that encompass correlation with GDP such as services, manufacturing and agriculture having a significant contribution. We performed regression using Multiple Linear Regression with manufacturing, value added; services, value added; industry (including construction), value added; population, total as independent variables while GDP (in current USD) as dependent variable.

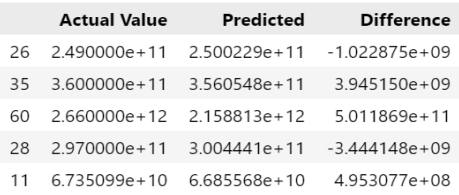


Fig 5.1: Difference in Actual vs Predicted Values

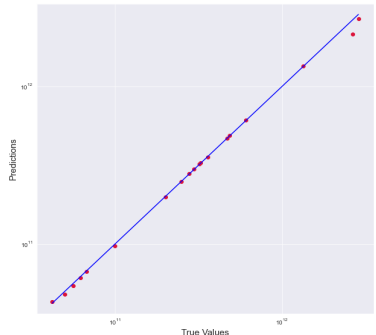


Fig 5.2: Plot of True Values vs Predicted Values using Multiple Linear Regression

# 5. CONCLUSION

In conclusion, our study utilized both linear regression and random forest models to forecast India's GDP based on data spanning from 1960 to 2022. The performance of both models demonstrated promising results, showcasing robust accuracy and precision metrics.

The linear regression model offered valuable insights into the linear relationship between the year and GDP, while the random forest model adeptly captured the complexities within the dataset, thereby enhancing predictive accuracy. Together, these models provide a holistic view of GDP trends, facilitating informed decision-making.

Nevertheless, it is crucial to acknowledge certain limitations, such as the potential influence of external factors on GDP fluctuations. Future improvements could entail integrating additional economic indicators for a more comprehensive analysis.

Our findings underscore the importance of precise GDP forecasts for effective economic planning and policymaking. This study contributes significantly to understanding India's economic dynamics and sets the stage for refining forecasting methodologies. We are confident that these models will continue to play a pivotal role in shaping future predictions and assisting in strategic decision-making processes.

## References

1. Aiyar, SA. “Subramanian fails the smell test no less than India’s GDP data.” The Times of India, 16 June 2019,

timesofindia.indiatimes.com/blogs/Swaminomics/subramanian-fails-the-smell-test-noless-than-India’s-gdp-data/.

1. Bhalla, Surjit S. “The Most Off-Track of Them All.” The Indian Express, 21 June 2019, indianexpress.com/article/opinion/columns/gdp-economy-narendra-modi-arvind-subramanianfinance-5793864/.
2. J.Banget al., New Tools for Predicting Economic Growth Using Machine Learning: A Guide for Theory and Policy, 2015.
3. V.Padmawaret al., “Gross domestic product prediction using machine learning,” vol. 08, pp. 2395-0056, 2021.
4. "Simple Linear Regression", *Wiley Series in Probability and Statistics*, pp. 19-46, Jan. 2005.
5. J. Yoon, “Forecasting of real GDP growth using machine learning models: Gradient boosting and random forest approach,” Comput. Econ.,v ol. 57, no. 1, pp. 247-265,2021[doi:10.1007/s10614-020-10054-w].
6. "The Linear Regression Model (LRM)", *Linear Regression: A Mathematical Introduction*, pp. 1-21, 2019
7. Patnaik, Prabhat. “India’s GDP Growth in the Recent Period.” Peoples Democracy, XLIII, 26 June 2019, peoplesdemocracy.in/2019/0623\_pd/india’s-gdp-growth-recent-period.
8. MEAC. “GDP Estimation in India- Perspectives and Facts.” 19 June 2019, eacpm.gov.in/wpcontent/uploads/2019/06/EAC-Paper\_GDP-estimation\_19-June-2019.pdf.

[10] The Online Retail Market Analysis for Social Development with Machine Learning M Nayak, B Narain – SPAST Abstracts, 2021.

[11] On-Line Big Data Analysis Using K-Mean and Modified K-Mean Algorithm with Machine Learning Techniques\* B Narain 2021.

[12] Big Data Mining Algorithms for Predicting Dynamic Product Price by Online Analysis\* M Nayak, B Narain – Computational Intelligence in Data Mining, 2020.

[13] D Ather et al.“Selection of Smart Manure Composition for Smart Farming Using Artificial Intelligence Technique” J. Food Qual., vol. 2022, article ID 4351825, 2022. doi:[10.1155/2022/4351825](https://doi.org/10.1155/2022/4351825).

[14] P. K. Jena et al., “A novel approach for diabetic retinopathy screening using asymmetric deep learning features,” Big Data Cogn. Comput., vol. 7, no. 1, p. 25, 2023. doi:[10.3390/bdcc7010025](https://doi.org/10.3390/bdcc7010025).

[15] B. Narain et al., “Impact of emotions to analyze gender through speech,” 4th International Conference on Signal Processing, Computing and Control. (ISPCC), Solan, India, 2017, 2017, pp. 31-34. doi:[10.1109/ISPCC.2017.8269645](https://doi.org/10.1109/ispcc.2017.8269645).

[16] M. Nayak and B. Narain, “Predicting dynamic product price by online analysis: Modified K-means cluster” in Computational Intelligence in Pattern Recognition. Advances in Intelligent Systems and Computing, vol. 1120, A. Das, J. Nayak, B. Naik, S. Dutta and D. Pelusi, Eds. Singapore: Springer, 2020, 1-15. doi:[10.1007/978-981-15-2449-3\_1](https://doi.org/10.1007/978-981-15-2449-3_1).

[17] M. Nayak et al., “An AI-based efficient model for the classification of traffic signals using convolutional neural network” in Building Secure Business Models Through Blockchain Technology: Tactics, Methods, Limitations, and Performance, S. Dewangan, S. Kshatri, A. Bhanot and M. Shah, Eds. IGI Global, 2023, pp. 20-35. doi:[10.4018/978-1-6684-7808-0.ch002](https://doi.org/10.4018/978-1-6684-7808-0.ch002).

[18] M. Nayak and A. K. Dass, “GSM and Arduino based Smart Home Safety and Security System,” Recent Trends Inf. Technol. Appl., vol. 6, no. 1, pp. 1-6, 2023. doi:[10.5281/zenodo.7610756](https://doi.org/10.5281/zenodo.7610756).

[19] Intelligent analytics for Industry 4.0 applications, 2023. (n.p.): CRC Press.

[20] M. Nayak and A. Barman, “A real-time cloud-based healthcare monitoring system” in Computational Intelligence and Applications for Pandemics and Healthcare, S. Kshatri, K. Thakur, M. Khan, D. Singh and G. Sinha, Eds. IGI Global, 2022, pp. 229-247. doi:[10.4018/978-1-7998-9831-3.ch011](https://doi.org/10.4018/978-1-7998-9831-3.ch011).

[21] O. Pattnaik et al., “Empowering safety: A deep dive into AI and machine learning solutions for women’s Security” in AI Tools and Applications for Women’s Safety, S. Ponnusamy, V. Bora, P. Daigavane, S. Wazalwar, Eds. IGI Global, 2024, pp. 121-131. doi:[10.4018/979-8-3693-1435-7.ch007](https://doi.org/10.4018/979-8-3693-1435-7.ch007).