|| Volume 6 || Issue 5 || May 2021 || ISSN (Online) 2456-0774
INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH
AND ENGINEERING TRENDS

MACHINE LEARNING BASED GENERIC GDP ANALYSIS AND PREDICTION SYSTEM

Nikhil Vyas[1], Jay Patel[2], Darshit Vala[3] , Devansh Patel[4] , Rohit Patel[5]

Student[1-4], Assistant Professor[5], *Information and Communication Technology Department,*
Adani Institute of Infrastructure, Ahmedabad, India
nikhilvyas.ict17@gmail.com[1], drrohitpatel.aiie@gmail.com[5]

---------- ***----------

Abstract: - One of the key aspects of sustainability goals is self-reliance. The Gross Domestic Product (GDP) is one of the metrics to ensure self-sustained growth for any country. The total monetary value of goods and services flowing through an economy over time is measured by GDP. GDP, along with other economic data points, is an indicator of the health of any nations' economy. Measuring and predicting the GDP is one of the major concerns for researchers across the globe. A generic technique to predict the GDP values from the customized dataset for Gujarat State is proposed in this work. Models based on various machine learning techniques like ARIMA and Random Forest Regressor are proposed in this work. Regression and time – series analysis models are created for GDP analysis and visualization.

Keywords: - *Gross Domestic Product (GDP), Machine Learning, Data Analysis, LASSO Regression, ARIMA Model, Random Forest Regressor.*

---------- ***----------

I INTRODUCTION

GDP is the most closely monitored and crucial economic indicator for economists as well as investors. GDP is the monetary value of all the manufactured commodities produced within a country's borders in a specific period by the country's citizens and foreigners. It is basically used to determine a country's economic health. [1]

Regression analysis is a set of statistical methods used for determining the correlation between a dependent variable and one or more independent variables. Regression model [2] is used for estimating the relationship between a GDP and affecting factors (dependencies). The Random Forest algorithm is a supervised learning algorithm that incorporates ensemble learning methods[3]. Hence, prediction for test data or new data is done by the regression model.

Time series analysis [4] consists of techniques for analysing time series data to extract meaningful information and other attributes of the data. Time-series forecasting is the process of using a model to predict possible trends based on previously observed values. Here the ARIMA(Aautoregressive integrated moving average model) model[5] is used for better forecasting. ARIMA has 3 hyper parameters p, d, and q. The values/order of these parameters have been determined by the ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) plots. We have fitted a 2nd order differencing method for data stationarity. The visuals like trends, seasonality and stationery will be given by the ARIMA model.

The proposed work uses a feature selection algorithm to determine the factors that affect GDP of a country. Using these factors, GDP of a country will be predicted with the help of previous year data (1951-2012). Objectives of the work is to make a model which can predict the GDP of the country and make any changes with less time compared to the recent paper-based system of calculating GDP.

This work uses a new analytic model named pandas-visual-analysis. It is used for auto exploratory data analysis and it provides the solution to the problem of outliers. Random Forest Regression [3], one of the most accurate learning algorithms is used. It generates a highly accurate classifier for several datasets. On large databases, it performs well. We also have Lasso Regression [6] for structural changes. Since shrinking and eliminating coefficients can minimize variance without significantly increasing bias, it can provide good prediction accuracy. Government can reduce manpower for calculating GDP with the help of this model. [7]

The work performs various seasonality and stationarity changes which were impossible earlier. Structural changes are made in the model. So, the software will be used for economical calculations. Government realizes the factors influencing the investment or consumption of resources that indirectly affect the GDP metrics of the country. Thereby, suitable measures and policies are proposed to improve the economy of the country.

Organization of the paper is as mentioned. Section I introduces

the idea of work; section II discusses the literature survey. Section III elaborates the idea behind the Proposed Approach, section IV shows the Implementation results and Section V concludes the work showcased herewith.

II LITERATURE REVIEW

D. Cielen (2016) [8] Data Science is the study of concepts to in order to fulfil the fundamental tasks data scientists are responsible for. Proposing a suitable data science process, data visualization, graph databases, the use of NoSQL are the main ingredients of data science. Languages like python and R are basically used in dealing with data. Discover how Python enables us to derive judgment from large datasets that require various machines to store them, or from moving data at such a fast rate that no single machine can handle it. This work gives a hands-on experience with the help of various Python data science libraries, Scikit-learn and Stats Models.

Kumar (2014) [9] Models that predict future values based on existing(past) data are created using time series forecasting. A prediction is a measurement or estimate that uses data from past events coupled with current trends to predict the outcome of a future event. A prediction, on the other hand, is the act of indicating that something will happen in the future, with or without prior knowledge. Gdp data is nonlinear data, Differencing is a technique for converting a non-stationary time series to one that is stationary. This is a critical step in the data preparation process for an ARIMA model. The parameters of ARIMA are determined by ACF (Autocorrelation function) and PACF (Partial Autocorrelation function) plots.

M. Patel (2020) [10] Various Machine Learning algorithms are covered in the work which gave a better idea about regression models. A regression analysis is a form of predictive modelling in which the connection between a dependent (target) and independent variable(s) is investigated (predictor). It is a significant device for dissecting and modelling of data. In this method, we attempt to fit the line/curve to the data focuses to minimize the differences between distances of data focuses from the curve or line. There are different sorts of regression investigation like linear, calculated and polynomial. Univariate regression analysis is used when there is only one independent variable, while multivariate regression analysis is used when there are several independent variables.

Fonti V. (2017) [6] Feature selection is one of the most important and difficult tasks in statistical modelling. This is because the desired performance differs for various sets of data, and it is difficult to find a model that fits for every type of problem. In this paper LASSO (Least Absolute Shrinkage and Selection Operator) method is used for feature selection problem. This method was tested in different setups, two type of statistical models: Linear model, generalized linear model are used here. LASSO regression provides very good accuracy and minimizes prediction error.

III PROPOSED APPROACH

Seasonality and stationarity changes are implemented here, which were not possible earlier. Due to the said implementation, related manual, time consuming and inaccurate tasks will reduce with the use of forecasting. Structural modifications are possible now easily as the software can be used for economical calculations.

Design Methodology

Figure 1 shows the design methodology to implement a machine learning based GDP Prediction. In the said methodology there are five steps [8]

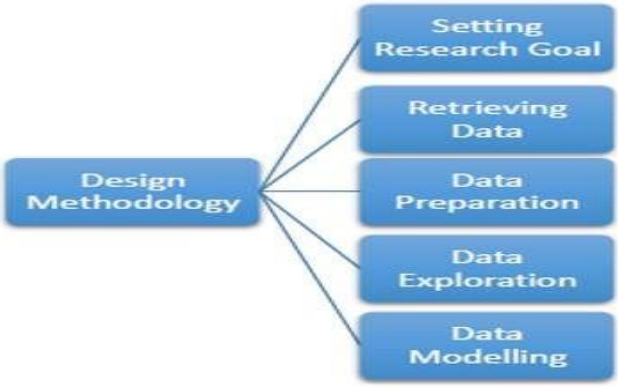


Figure 1 Design Methodology

1. Identifying the research goal

- A clear research goal.
- The project aims and context
- Algorithms for Data Analysis
- Resources to use.
- Feasibility analysis
- Deliverables and a measure of success
- A timeline

2. Retrieving data

The initial step in data science is to gather the required data. Sometimes field study is useful to perform the data, but most of the time, researchers are directly not involved in the data compilation step. Data can be stored in a variety of formats, from plain text files to database tables. The objective now is to acquire the related data. Getting access to data is also a challenging task. Organizations realize the importance and sensitivity of data, and often have norms in place to ensure that everyone has access to just what they want. [11]

3. Data Preparation

Data Cleansing integrating, and transforming data are the main steps in data preparation. However, the clean and validated information is gathered from open source. The dataset is clean and hence can be used as it is.

4. Data exploration

Study of a broad data set in an unstructured manner is conducted to find initial trends, features, and points of interest. When information is presented in the form of an image, it is much easier to comprehend. As a result, to obtain a better understanding of the data and the relationships between variables, graphical techniques are used. This stage focuses on data exploration. Extra Trees Regressor class is used to implement a Meta estimator that fits several decision trees on various sub samples. We are also using Pandas visual analysis tool for detailed data exploration.

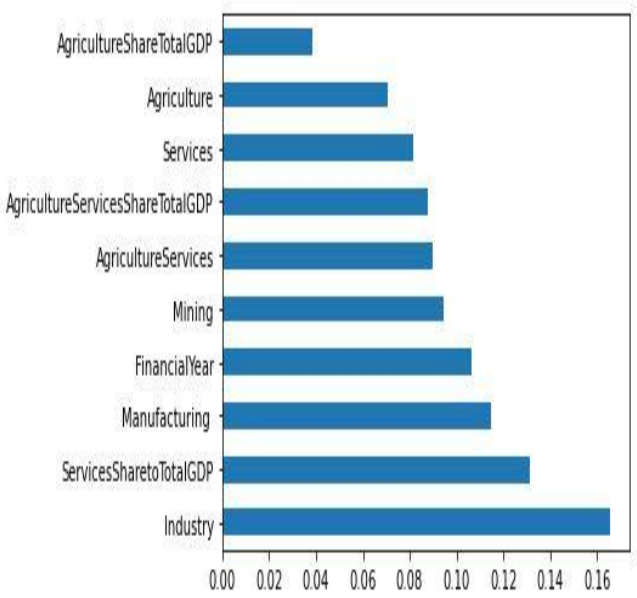


Figure 2 Major Affecting Factors

5. Data modelling

Models can be designed with clean data and a clear understanding of the content in order to make better predictions, identify objects, or gain an understanding of the system to model. Since the researchers would be able to tell what they are searching for and what the result should be, this phase is far more concentrated than the exploratory analysis stage. In the dataset, the dependent variable is the regression type. In data modelling, type of the model is to be selected by the dataset; so, random forest regression is used. Both classification and regression tasks will benefit from the Random Forest algorithm. It provides higher accuracy through cross validation.

B. Design Analysis

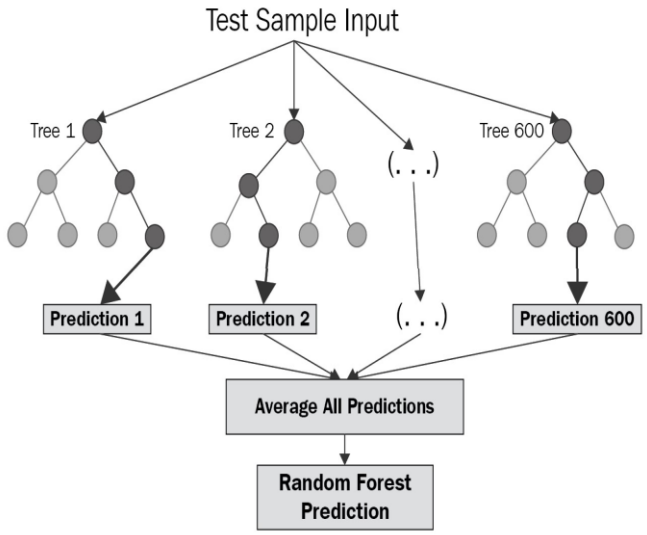


Figure 3 Random Forest Regressor

The systematic procedure of developing a design, which includes all knowledge discovery, preparation, and communication, is known as design analysis. This can be applied to any kind of architecture, including physical objects like buildings as well as intangible entities like software, data, and processes.

Data is being split into two parts: train and test. The data is being processed by random forest regressor algorithm. The basic concept is to use a combination of decision trees to determine the final outcome rather than relying on independent decision trees. The R2 score method is used to validate the project. [12]

IV IMPLEMENTATION AND RESULTS

Tools and dependencies required the work are listed as under:

1. Integrated Development Environment (Jupyter Lab is used)
2. Python3 (Language used for code)
3. Flask / Django (API for python to get data)
4. Matplotlib (Python Library used for Machine Learning) [13]

Following libraries for time series forecasting and regression are used:

- numpy
- pandas
- matplotlib
  - rcParams
- statsmodels
  - adfuller
  - acf & pacf
  - ARMIA
- sklearn
  - Random Forest Regressor

A. Steps for Data Analysis and Visualization are as specified below:

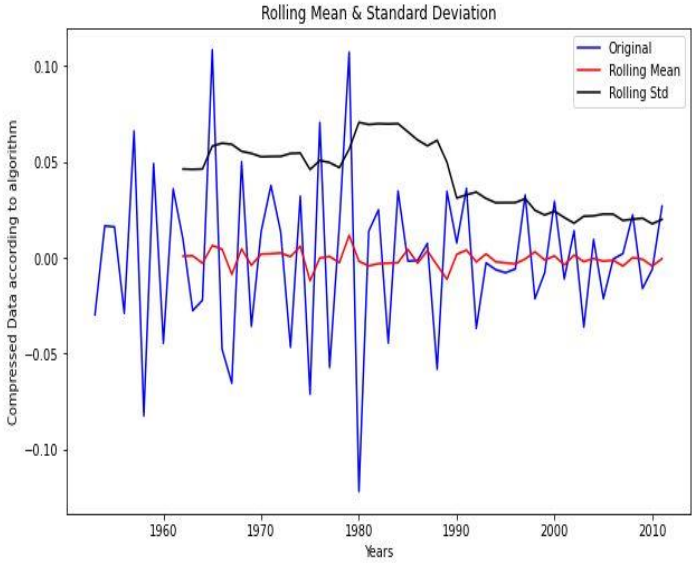
- Read the dataset from the csv file
  - .df = pd.read_csv("filename.csv")
- Do the feature engineering and data exploration part.
- Divide the data into training and testing section.
  - X_train, X_test, y_train, y_test=train_test_split(x, y, random_state=2, test_size=0.25)
- Fit the model into suitable regression technique.
- Make Predictions and check the validity.
  - Predictions=regg.predict(X_test)

B. Steps for Time Series Forecasting:

- Make the data stationary.
- Perform Dickey Fuller Test
- Determine ARIMA parameters by using acf and pacf plots.
- Retrieve the form of original data
  - Results ARIMA.plot predict(x,y)

C.Output/Testing/Results

Figure 4 shows the original, rolling mean and standard deviation of data. In the result of the Dickey fuller test, P-value is shown. If the value of P is close to 0 then it is good for the model. And if we get three critical values close to test stats then it is better for the model. [14]




```
Results of Dickey Fuller Test:
Test Statistic                  -6.636233e+00
p-value                          5.553744e-09
#Lags Used                       6.000000e+00
Number of Observations Used      5.200000e+01
Critical Value (1%)             -3.562879e+00
Critical Value (5%)             -2.918973e+00
Critical Value (10%)            -2.597393e+00
dtype: float64
```

Figure 4. Augmented Dickey-Fuller Test

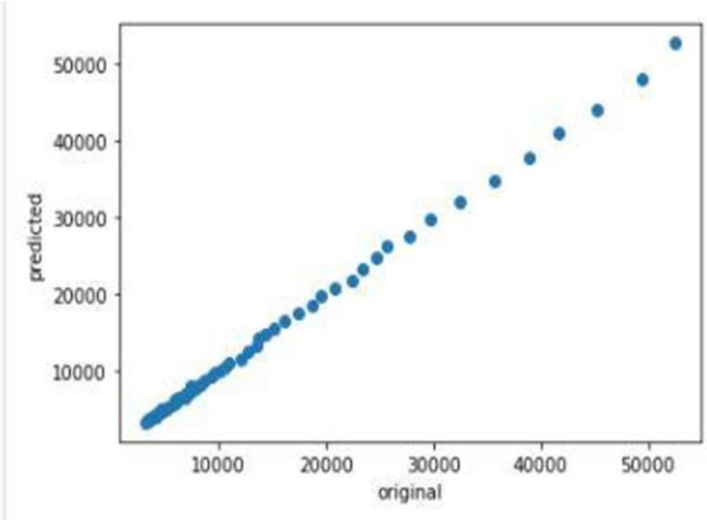


Figure 5 Original versus Predicted values

Figure 5 shows the scatter plot graph. An R2 score is used for checking the model's accuracy. Input data is taken as predicted values and 25% data is separated for testing and accuracy is computed as shown in Figure 5. The graph shows a linear scatter plot which shows that the prediction accuracy carries 95% confidence and is a satisfactory model for the GDP Prediction. [15]

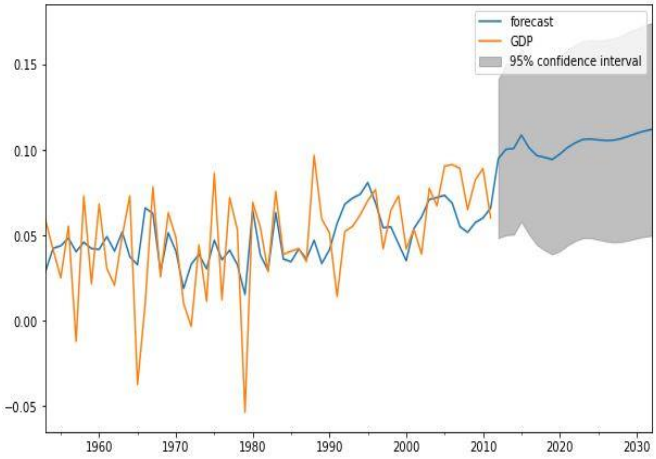


Figure 6 GDP Forecasting

In Figure 6, X-axis = years, Y-axis = values of compressed data after applying a differencing method. The dataset consists of 61 records. Here, ".plot_predict(1,80)" method is implemented so that trend or forecasting for the next 20 years of GDP can be observed

V CONCLUSIONS

By Implementing this model, the task of GDP calculation can be made less tedious as well as paperwork can be decreased. The chances of error are very decreased because combination of various algorithms gives very high accuracy. Except for

unpredicted circumstances like global pandemic, world war, global economic crisis the model is highly accurate. Government organizations can directly be helped by this model, because any modification to the model and calculation parameters is very easy. The model will be deployed on a website, which will be accessible to public.

REFERENCES

[1] D. REID, "Combining three estimates of gross domestic product," *Economica*, vol. 35, no. 140, 1968.

[2] R. H. Myers, Classical and modern regression with applications, Belmont, CA: Duxbury Press, 1990.

[3] M. R. Segal, Machine Learning Benchmarks and Random Forest Regression, 2003.

[4] C. H. X. Chatfield, The analysis of time series: an introduction with R, CRC Press, 2019.

[5] P. Newbold, "ARIMA model building and the time series analysis approach to forecasting," *Journal of Forecasting* , vol. 2, no. 1, 1983.

[6] E. B. Valeria Fonti, "Feature Selection using LASSO," *VU Amsterdam Research Paper in Business Analytics*, p. 25, 2017.

[7] S.-s. A. C. J. S. S. L. Wang, "Application of the combination prediction model in forecasting the GDP of China," *Journal of Shandong University (Natural Science)*, no. 10, 2009.

[8] A. D. B. M. a. M. A. Davy Cielen, Introducing Data Science, Manning Publications Co., 2016.

[9] A. M. Kumar Manoj, "An Application Of Time Series Arima Forecasting Model For Predicting Sugarcane Production In India," *Studies in Business and Economics*, p. 14, 2014.

[10] A. Priyadarshi, M. Patel, "An analysis of various machine learning algorithms and their implementations in the contemporary world" International Journal of Advance Scientific Research and Engineering Trends, vol. 5, no. 12, 2020.

[11] DATA SOURCE LINK "https://data.gov.in/".

[12] W. M. Liaw A., "Classification and regression by randomForest," R news, 2002.

[13] P. Lemenkova, "Processing oceanographic data by python libraries numpy" Aquatic Research , no. 2, 2019.

[14] Y.-W. C. &. K. S. Lai, "Lag Order and Critical Values of the Augmented Dickey–Fuller Test," Journal of Business & Economic Statistics, 1995.

[15] V. K. F. Z. R. MICHAEL, "PSEUDO-R2 MEASURES FOR SOME COMMON LIMITED DEPENDENT VARIABLE MODELS," *Journal of Economic surveys* , vol. 10, no. 3, 1996.

[16] V. Agrawal, "GDP modelling and forecasting using ARIMA: an empirical study from India.," *Central European University*, 2018.

[17] S. F. Chandrasekhar G., "A survey on feature selection methods," *Internationa Journal on Computers and Electrical Engineering*, vol. 40, no. 1, 2014.

[18] G. A., Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems, O'Reilly Media.,2019.

[19] Dedić, Fuad, Elmir Babović, Sanja Dizdar-Kapetanović, and Srđan Nogo. "Regression Analysis of Dependency Between Related Courses on 1 st Year of Study on Faculty of Information Technologies." In *2021 20th International Symposium INFOTEH-JAHORINA (INFOTEH)*, pp. 1-6. IEEE, 2021.