India’s GDP Prediction

Authors:

Satyajit Ghana, SHIKHAR SINGH

2019

Table of Contents

[Project Aim 1](#_Toc14049540)

[Project Data 1](#_Toc14049541)

[About the Data: 2](#_Toc14049542)

[Graphical Visualization 3](#_Toc14049543)

[Model Structure 6](#_Toc14049544)

[Model 1 (MLP Regressor): 6](#_Toc14049545)

[Model 2 (Time Series): 7](#_Toc14049546)

[Outcomes and Results: 7](#_Toc14049547)

[Project Members: 9](#_Toc14049548)

India’s State GDP Prediction

# Project Aim

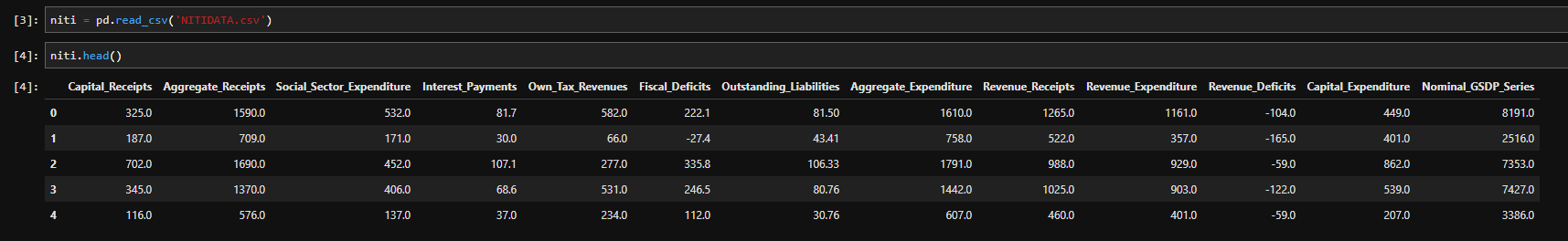
The main aim of this project was to create a ML model that would help in predicting the expected GDP of each state of India with minimum margin of error for any given year in the future such as 2020 and beyond. This is very beneficial since by this data we could predict he growth of different states of the country and it will be able to predict how the GDP must be used in an effective way for the benefit of the state.

# Project Data

The Data used in the project was taken from <https://niti.gov.in/> . The RAW Data was taken from the financial section that contained the following documents: Aggregate Expenditure, Aggregate Receipts, Capital Expenditure, Capital Receipts, Fiscal Deficits, Interest Payments, Outstanding Liabilities, Own Tax Revenues, Revenue Deficits, Revenue Receipts, Social Sector Expenditure and Nominal GSDP Series. The RAW data was in XLSX format which was converted to a well-known CSV format that could be later easily read by python pandas library. The unknown values that were originally represented by a ‘-’ was replaced by NA’s by pandas.

The Data was given for 29 states and 2 Union Territories, from the year 1981 to 2016, some of the data was missing for some of the states, due to the fact that the data was started to be collected at a later period of time, or maybe the state wasn’t even formed in that year, such years were removed from the data by using remove NA’s in python pandas, and the clean and preprocessed data was dumped to a CSV file. The preprocessing done to create the final CSV file that contained the list of features and the corresponding GSDP was created by fetching the features of every state and every year and matching them to their corresponding GSDP of that state in that year. By using this and removing those series which had NA’s a total of 660 observations were made in the final CSV file.

# About the Data:



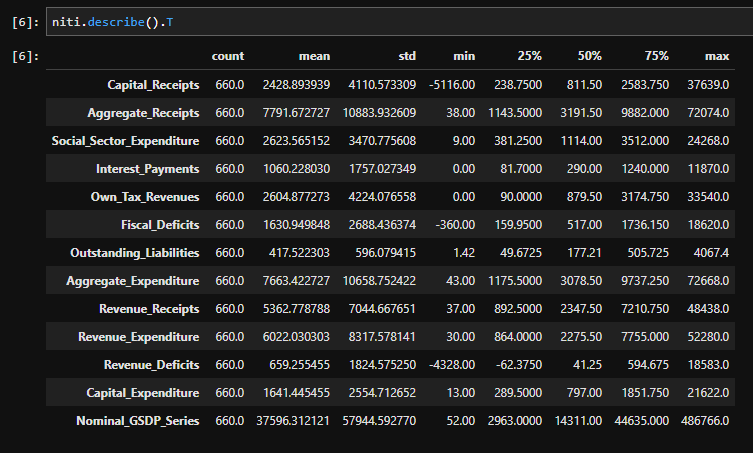


Figure ‑ Data Description

# Graphical Visualization

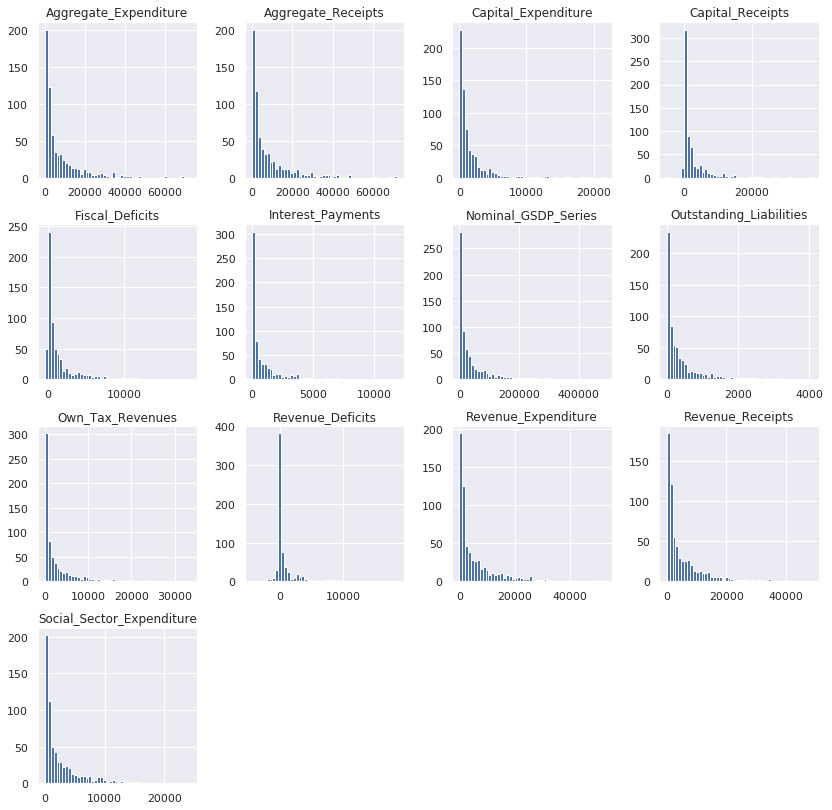


Figure ‑ Histogram Plot



Figure ‑ Pair Plot

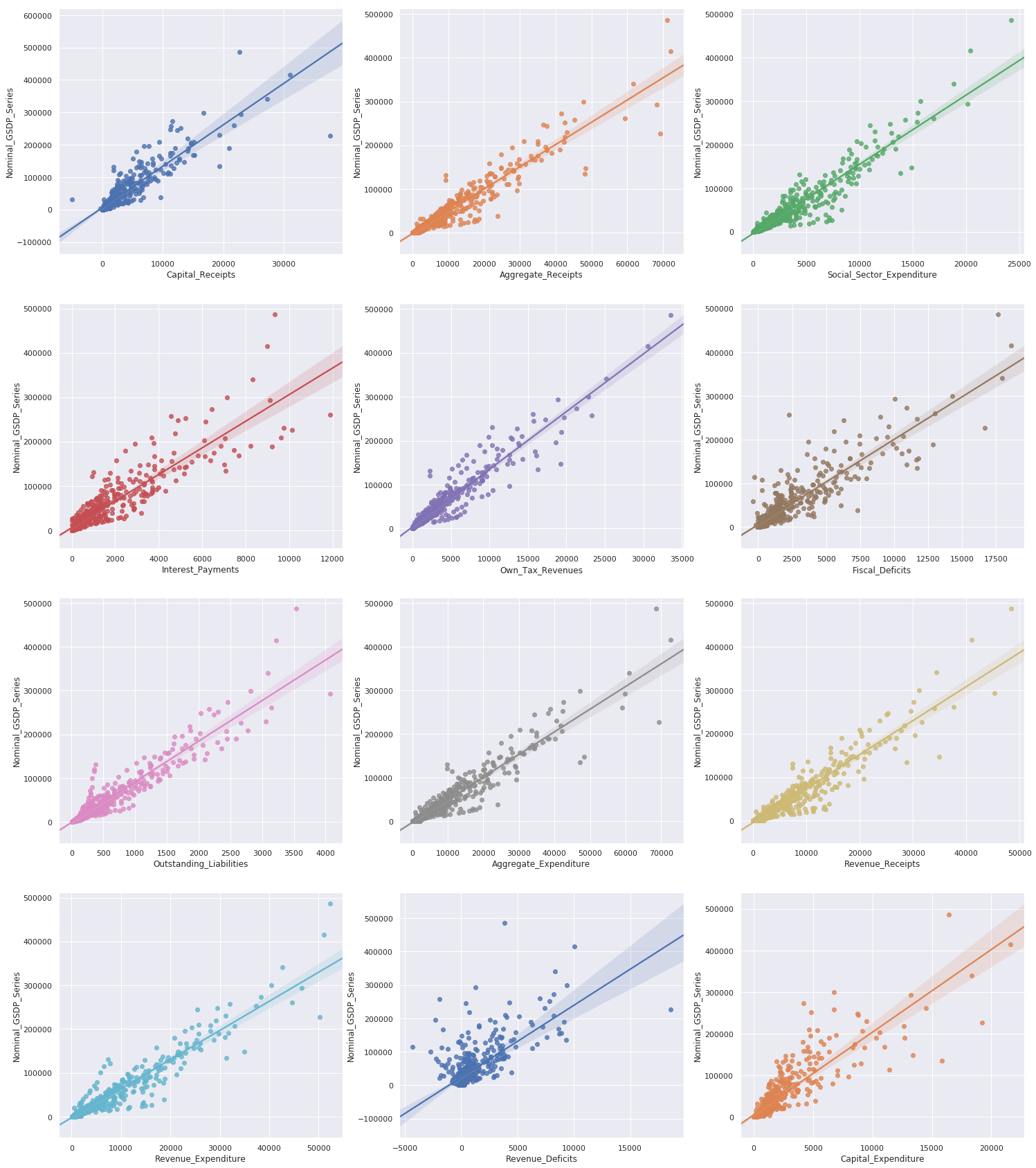


Figure ‑ Regression Plot

# Model Structure

The Model is split into two parts, one model predicts the Nominal GSDP from the features mentioned in the Project Data such as the Aggregate Expenditure, Aggregate Receipts, etc. and the other model is a time series Model that predicts these features for any given year, now the idea is to use the already given data from NITI Aayog and train the model to predict the GSDP series, then feed the predicted data for the year 2020 to get the Nominal GSDP values for the year 2020 and beyond.

## Model 1 (MLP Regressor):

The first model is a Multi-Layer Perceptron Regressor, a type of neural network, which is used here to predict the Nominal GSDP for the given data. The features were scaled before being used in the regressor model.

Features: Aggregate Expenditure, Aggregate Receipts, Capital Expenditure, Capital Receipts, Fiscal Deficits, Interest Payments, Outstanding Liabilities, Own Tax Revenues, Revenue Deficits, Revenue Receipts, Social Sector Expenditure

Output: Nominal GSDP Series

Learning Rate: 0.001

Hidden Layers Sizes: 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12

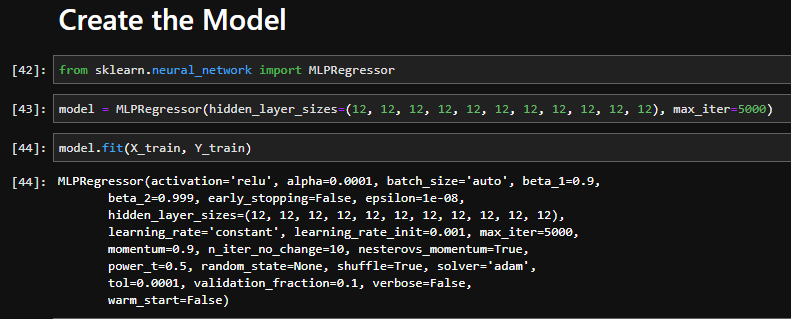


Figure ‑ MLP Model

The Model was trained for the training data and the following score was obtained on the test data.

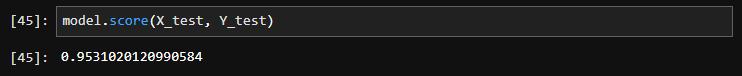


Figure ‑ Model Score

The Model was analyzed using the standard sklearn metrics

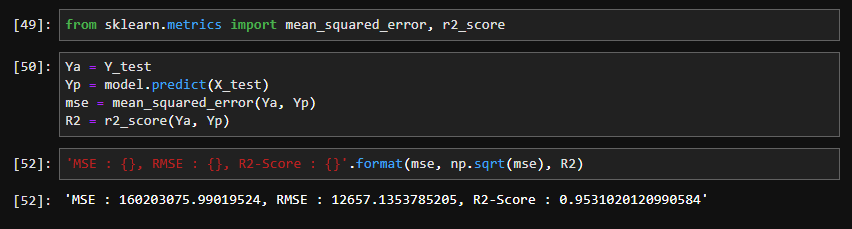


Figure ‑ Model Metrics

The R2 score here determines the accuracy of the model, and 95.31% is considered to be a good score.

## Model 2 (Time Series):

The second model was to be a time series model that predicts the features for a given year in the future, for this there are options such as ARIMA, which is considered to be a really good model, but since the data points are very few for us, Holt Winters model was used. Holt-Winters forecasting is a way to model and predict the behavior of a sequence of values over time—a time series. Holt-Winters is one of the most popular forecasting techniques for time series.

For each of the states, the features for the year 2020 was predicted, and this was stored in a CSV file.

# Outcomes and Results:

The data predicted for the year 2020 from the Model 2 was fed into Model 1 to predict the Nominal GSDP for the individual states and also all the states, to get the following results:

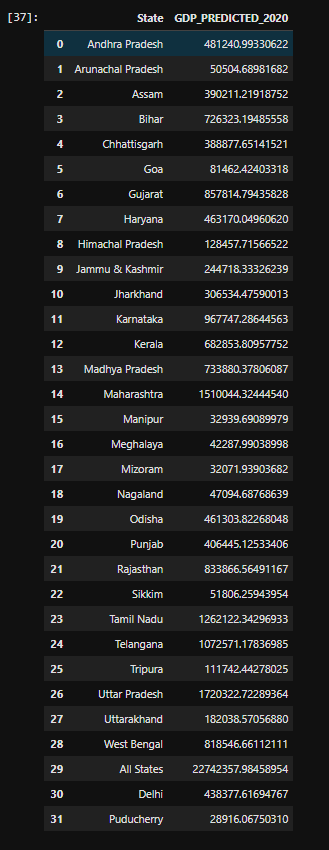


Figure ‑ Model Results

# Project Members:

Satyajit Ghana

Shikhar Singh