Analysis and Forecasting of Indian Economic Data

Shubham Lohiya
Department of Mechanical Engineering
IIT Bombay
shubhlohiya@cse.iitb.ac.in

Swarada Bharadwaj
Department of Energy Science and Engineering
IIT Bombay
18D170031@iitb.ac.in

***Abstract*—In this project, we have performed exploratory and predictive analyses on macroeconomic indicator data for India. The data in use covers wide-ranging economic indicators related to banking, production output, inflation, interest rates, imports and exports. We aim to forecast inflation using this data, and use CPI (Consumer Price Index) as a measure of inflation. After examining CPI's correlation to the other variables, we perform feature engineering and selection for our forecasting framework. We consider various statistical and Machine Learning models for forecasting CPI, and achieve very good results, i.e. forecasts with very low dispersion error.**

***Index Terms*—Indian Economy, Time Series Forecasting, Time Series Analysis**

I. Introduction

In this project, we perform exploratory and predictive analyses on macroeconomic indicator data for India to get an accurate historical and modern picture of the Indian economy. The Reserve Bank of India (RBI) has begun to openly publish datasets relating to the Indian economy in areas including those of banking, production, inflation, interest rates, foreign trade, currency exchange and money reserves. These datasets have been made available for the use of data science applications and analysis. Furthermore, while papers can be found relating to the application of machine learning to forecast economic outcomes for the world or for many other countries (such as Brazil and the United States), not as much research of this kind which is centered around the Indian economy has been made openly accessible. The Indian economy is a unique case - India is the second-most populous country in the world, and economic outcomes in India affect a disproportionately large percentage of the world's population. India has made enormous economic strides over the past 60 years, having had to build many industries from the ground up in this short period of time. Due to these circumstances, until 1991, the Indian economy was characterised by state intervention and regulation. Our analysis starts from 1996, only a short while after the Indian economy's liberalization. Data from the RBI's archives is available in many different formats and timeseries, so the need is felt to gather and pre-process the data in order to match the timeseries of all variables and make it usable for an ML task. The RBI data is also augmented with Indian macroeconomic timeseries data from other official sources. After this is done, exploratory data analysis is done to show correlations among variables as well as their movements along the timeseries consisting of several decades. After an accurate picture of these variables is obtained, an attempt is made to apply ML prediction techniques to forecast Indian economic indicators, particularly inflation, inspired by similar attempts made for other counties' economic data. Inflation is among the most important indicators for assessing the health of the national economy, and is easily represented through many indexes.

II. Background and Prior Work

Background: The pre-processing stage of economic data often includes the de-seasonalisation of the timeseries data. Economic data tends to contain certain seasonalities, wherein data from a certain season has characteristics of its own. In order to treat each row of data as an independent datapoint and to assess the correlation of predictors without any dependency of the season of the datapoint, all the data obtained in its raw form has to be de-seasonalised.
The Consumer Price Index (**CPI**) is a measure that examines the weighted average of prices in a basket of consumer goods and services, such as transportation, food, and medical care. Changes in the CPI are used to assess price changes associated with the cost of living. CPI is one of the most frequently used statistics for identifying periods of inflation or deflation.
We use **Shapley Values** [11] of features for a particular trained forecasting model to see the impact of each feature on model prediction. It is a solution concept from cooperative game theory, and has gained recent popularity as an **Explainable AI** technique. In the current context, can be understood as a black-box method that analyses the prediction impact of each input feature.

Prior Work: There are many examples of timeseries analysis done on Indian economic data for different purposes. In [1], Guha Deb and Mukherjee examined the relationships between stock market growth and economic growth in India over a decade by looking at timeseries data. Timeseries analysis was also used by Sehrawat and Giri [2] to create a financial development index for the Indian economy and examine the effect of trade liberalisation on economic growth in India. Baybuza [3] used basic Machine Learning methods such as LASSO [6] and Ridge regression [7] and Random Forest [8] to forecast inflation in Russia, confirming the possibility of more accurate forecasting capabilities in Machine Learning. For this, the author used macroeconomic series reflecting the conditions of business activities, industrial production, foreign

trade and the financial market, among other indicators. The Consumer Price Index (CPI) was used here has the measure of inflation. In [4], Hellwig explored the use of Machine Learning to predict fiscal crises - disruptive economic events causing a fall in economic output which has long-lasting repercussions. The models the author selected identified predictors which reflect high correlation with fiscal crises, rather than causation. In [5], the authors used Deep Learning models for inflation prediction, employing recurrent neural networks with a gated recurrent unit (GNU-RNN) [9] [10] and obtaining a superior performance in comparison to traditional methods.

III. Data and Methodology

A. Data Description

Data is obtained through two sources: the RBI Database on the Indian Economy (RBI-DBIE) and the Federal Reserve Bank of St. Louis.' Economic Data (FRED). Data regarding the components of the national money stock and India's foreign trade in US Dollars is obtained from DBIE. From FRED, data related to goods manufacture, CPI, GDP (Gross Domestic Product), government expenditure, gross fixed capital, interest rates, share prices and industry prices is obtained. The DBIE data is available in different formats, and the FRED dataseries all match in terms of format but are formatted differently from the DBIE data. Furthermore, around half of the FRED data needs to be de-seasonalised. This de-seasonalisation is carried out in Excel.

Data is available as both monthly and quarterly time-series, with different starting and ending dates. The data formats are aligned and all datasets are combined using an inner join with the date as the index. In this way, the greatest common timespan is acquired (July'96 to April'21) for which quarterly data is available for all variables. This leaves the final dataset with all predictors containing 100 rows of data. The predictors present in this final table are: Exports (Million Dollars), Imports (Million Dollars), Trade Balance (Million Dollars), Currency in Circulation (Crore Rupees), Cash with Banks (Crore Rupees), Currency with the Public (Crore Rupees), Consumer Goods Manufacture (Index 2015=100), CPI (Index 2015=100), GDP (OECD Index), Government Expenditure (Indian Rupees), Gross Fixed Capital (Indian Rupees), Immediate Rate (percent), Interest Rates (percent), Share Prices (growth rate from the previous year) and Wholesale Industry Prices (Index 2015=100).

B. Forecasting Approach

Consumer Price Index (CPI) is one of the most frequently used statistics for identifying periods of inflation or deflation. Due to CPI's high importance as an economic indicator, we build a framework to forecast it to predict inflation or deflation. For forecasting, we develop a set of features as predictors with CPI as our target variable. The developed features fall into two categories:

- Features based on historical data of CPI itself, these include lags of CPI (data from previous fiscal quarters) and trends of CPI over various intervals (indicated by slope).
- Exogenous features based on historical data of other economic indicators as mentioned in data description. These features also include lags and trends of each feature as described above.

1) Feature Selection: Lags of features were selected based on auto-correlation analysis for CPI and cross-correlation analysis for other exogenous features. For Example, (Fig. 1) CPI auto-correlation analysis shows that CPI's correlation with lags of itself seems to be monotonically decreasing. The partial auto-correlation plot further shows that previous lag captures all important information, so further lags are redundant for forming historical trend/seasonal features. This is why we drop rolling averages from consideration as features for our forecasting models. For more feature selection analyses, refer to the EDA notebook in our code-base.

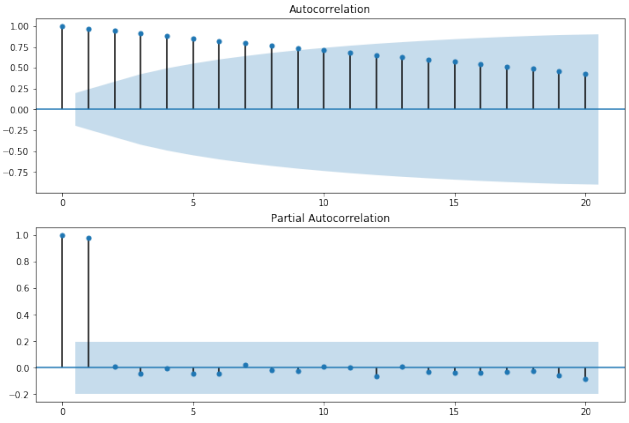

Fig. 1. Auto-correlation and Partial Auto-correlation of CPI

We also use Shapley Values, an Explainable AI technique, to get interpretable insights from our forecasting models. These Shapley analyses can be used to determine redundant features, which can be removed to reduce noise. Shapley analyses on retrained models and feature pruning can be iteratively done to arrive at the best results with the smallest subset of features. We leave this for future work.

2) Forecasting Models: To learn to forecast CPI for future fiscal quarters based on historical information, we consider the following models: (1) XG-Boost Regression (2) Random Forest Regression (3) XG-Boost Random Forest Regression (4) Multi-Layered Perceptron Regression. The first three are decision tree-based models whereas the last one is a deep learning model. In our initial analyses, we also considered statistical forecasting models like SARIMA and SARIMAX, but didn't pursue them as our Machine Learning models outperformed them easily.

3) Forecasting Evaluation: We use two metrics of evaluation:

- Mean Dispersion Error: The mean dispersion error for forecasts is defined as

$$\frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}}{y}-1\right|$$

where $\hat{y}$ and y are the forecasted CPI and actual CPI values respectively, and n is the total number of datapoints/fiscal quarters over which the mean dispersion Error is being calculated. This metric will hereon be referred to with the shorthand *Dispersion*.

- Root Mean Squared Error (RMSE): The RMSE for forecasts is defined as

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}-y)^2}$$

with symbols as previously defined.

IV. Experiments and Results

A. Time Series Analysis of Economic Data

First basic time-series plots are created to track the variations of all predictors over time. The first plots are made after dividing all normalized predictors by GDP:

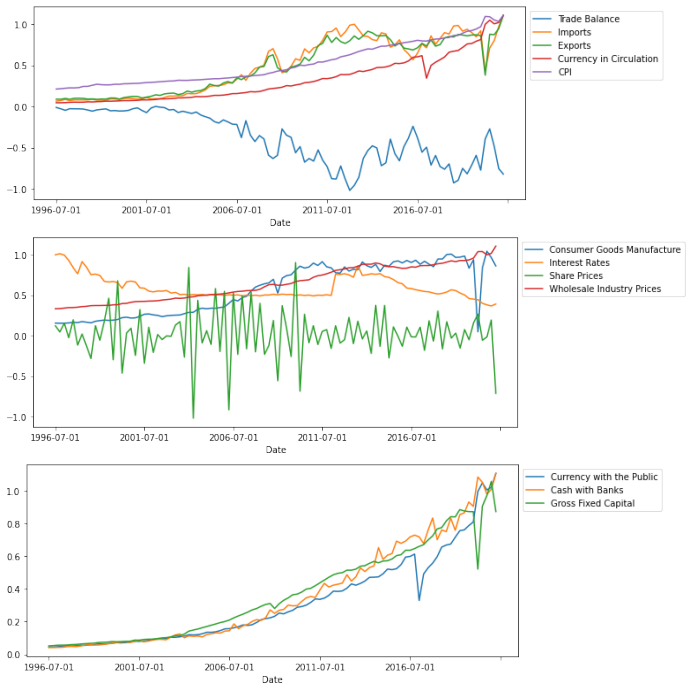

Fig. 2. Trends seen in all predictors as a proportion of GDP

In 2 it is observed that overall the money in circulation as well as the CPI (the measurement of inflation) have both grown with respect to GDP. This is what we would expect, as one is almost causal to the other. Consumer goods manufacture as a proportion of GDP has also risen despite one drastic plunge. Share prices (or the financial markets as a proportion of GDP) displayed great volatility, and have plunged in recent times, presumably due to the pandemic. Wholesale Industry Prices have steadily increased. Both imports and exports as a proportion of GDP have risen, while their difference has dropped drastically, indicating that India has become an importer overtime. Interest rates (the cost of loans) has remained stable - possibly due to fiscal regulatory measures.

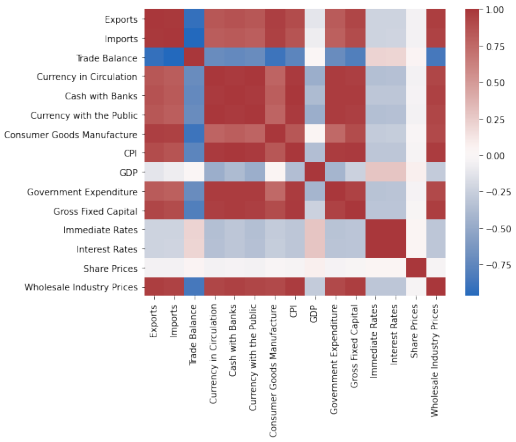

Fig. 3. Correlation heatmap of all predictors

3 is a heatmap showing the correlations between each variable.

B. Time Series Forecasting Setup

After feature engineering, we have usable data from 99 fiscal quarters. We split this data into train-validation-test splits in a 70:20:10 ratio.
We conduct hyperparameter tuning for all our models on the Validation data, and then retrain the model on the Train + Validation data before evaluating it on our Test data.

C. Forecasting Experiments

In this section, we describe the hyperparameter tuning setup for each model, and report the best observed parameters. Hyperparameter tuning was conducted using the Grid Search method.
The results for all models are indicated in Table 1.

1) XGBoost Regression: For this model, hyperparameter tuning setups is as follows: (1) n_estimators is tuned from [50, 100, 250, 500, 1000] (2) max_depth is tuned from [5, 10, 20, 30].
The best observed parameters are n_estimators = 100 and max_depth = 20.

2) Random Forest Regression: For this model, hyperparameter tuning setups is as follows: (1) n_estimators is tuned from [500, 1000, 5000, 10000] (2) max_depth is tuned from [5, 10, 20, 30].
The best observed parameters are n_estimators = 5000 and max_depth = 5.

TABLE I
PERFORMANCES OF FORECASTING MODELS, TABULARIZED

	XGBoost Regressor		**Random Forest Regressor**		**XGBRF Regressor**		**MLP Regressor**	
	Dispersion	RMSE	Dispersion	RMSE	Dispersion	RMSE	Dispersion	RMSE
Test	0.0897	12.4333	0.1257	17.913	0.1064	15.6583	0.00996	1.3552
Train + Val	6.7300e-6	0.0005	0.00632	0.6705	0.00528	0.39033	0.01199	0.96139

3) XGBoost Random Forest Regression: For this model, hyperparameter tuning setups is as follows: (1) n_estimators is tuned from [10, 25, 50, 75, 100] (2) max_depth is tuned from [3, 5, 10, 20].
The best observed parameters are n_estimators = 100 and max_depth = 20.

4) Multi-Layered Perceptron Regression: For this model, hyperparameter tuning setups is as follows: (1) Number of hidden layers is ranged for 1 to 3 (2) Hidden layer sizes are tuned from [256, 512, 1024, 2048].
The best observed parameters are n_layers = 2 and hidden_size = 1024.

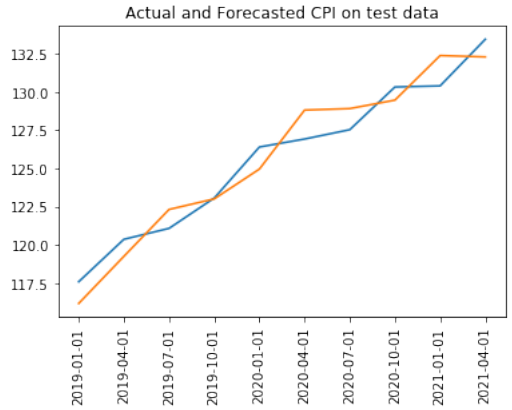

Fig. 4. Plot of forecasted and actual CPI on test data

D. Explainable AI Analysis: Shapley Value

As previously discussed, we use Shapley Values to interpret our model and the relative impact of various input features to our model. This analysis can also be used for feature selection by pruning low impact variables and re-training the model iteratively. We defer this to future work. Refer to (Fig. 5) for some insights from our Shapley analysis.

(Fig. 5) indicates the relative importance of various features of economic indicators for two models. Further discussion on Shapley analysis can be found in the forecasting notebook in the codebase. Also note that the XGBoost predictions rely mainly on only 2-3 features, but Random Forest is able to make use of more features to make predictions.

V. LEARNING, CONCLUSIONS AND FUTURE WORK

Through this project, we have gained first hand experience dealing with huge quantities of raw data, and learned how to organize, clean and filter it. To get usable data for forecasting from data of various macro-economic indicators, we also gained knowledge of these indicators, their influence in the economy, and how they relate to each other. We have also learned fundamentals of time series analysis and forecasting, and have successfully applied various statistical and machine learning models to forecast Consumer Price Index which is an important indicator for identifying inflation or deflation in the economy.

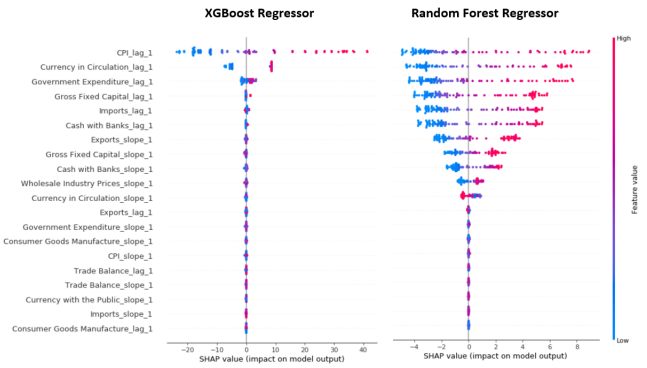

Fig. 5. Comparison of Shapley Insights for XGBoost and Random Forest models

In conclusion, we have gained important economic insights by jointly analysing historical macro-economic data. We have also successfully demonstrated how an economic indicator can be predicted using historical data of itself and other relevant indicators. In our experiments with forecasting CPI, we have achieved a very low forecasting dispersion error of about 1%.

In the future, this analysis can be extended to other macro-economic indicators. The Shapley analysis can be further used to perform feature selection in an iterative manner as described before. Such work can provide important insights and lend foresight to economists and policy makers.

CONTRIBUTION OF TEAM MEMBERS

Swarada Bharadwaj contributed to the consolidation of the datasets, the data preprocessing, formatting and merging, the exploratory data analysis and report writing.
Shubham Lohiya contributed to the data analysis for feature engineering and selection, development and evaluation of forecasting frameworks, and to the report writing.

REFERENCES

[1] Soumya Guha Deb and Jayadeep Mukherjee, "Does stock market development cause economic growth? A time series analysis for Indian economy," International Research Journal of Finance and Economics, ISSN 1450-2887 Issue 21, pp. 142–149, 2008.

[2] Madhu Sehrawat and A.K. Giri, "Financial structure, interest Rate, trade openness and growth: time series analysis of Indian economy," Global Business Review, 2017, pp.1278–1290.

[3] Ivan Baybuza, "Inflation forecasting using Machine Learning methods," Russian Journal of Money and Finance, December 2018

[4] Klaus-Peter Hellwig, "Predicting fiscal crises: a Machine Learning approach," upublished, IMF Working Paper WP/21/150, 2021

[5] Cheng Yang and Shuhua Guo, "Inflation prediction method based on Deep Learning," Hindawi Computational Intelligence and Neuroscience, August 2021

[6] Robert Tibshirani, "Regression shrinkage and selection via the Lasso," Journal of the Royal Statistical Society: Series B (Methodological), 1996

[7] Arthur E. Hoerl, Robert W. Kennard, "Ridge Regression: biased estimation for nonorthogonal problems," Technometrics, American Statistical Association and the American Society for Quality, 1970

[8] Leo Breiman, "Random Forests," Machine Learning volume 45, pp.5-32, 2001

[9] J. Yuan, H. Wang, C. Lin, D. Liu, and D. Yu, "A novel GRU-RNN network model for dynamic path planning of mobile robot," IEEE Access, vol. 7, pp. 15140–15151, 2019.

[10] J. T. Connor, R. D. Martin, and L. E. Atlas, "Recurrent neural networks and robust time series prediction," IEEE Transactions on Neural Networks, vol. 5, no. 2, pp. 240–254, 1994

[11] Kalai, E. and Samet, D., 1987. On weighted Shapley values. International journal of game theory, 16(3), pp.205-222.