New Approaches to Forecasting Growth and Inflation: Big Data and Machine Learning

Sabyasachi Kar
Amaani Bashir
Mayank Jain

2021

IEG Working Paper No. 446

सत्यमेव परमो धर्मः

October 2021

New Approaches to Forecasting Growth and Inflation: Big Data and Machine Learning

Sabyasachi Kar[1]

Amaani Bashir[2]

Mayank Jain[3]

JEL Classification Numbers: C14, C45, C52, C53, C55, E17, E37

Keywords: Forecasting, Big Data, Machine Learning, Supervised Learning, Meta-analysis, Growth, Inflation

Abstract

The use of big data and machine learning techniques is now very common in many spheres and there is growing popularity of these approaches in macroeconomic forecasting as well. Is big data and machine learning really useful in the prediction of macroeconomic outcomes? Are they superior in performance compared to their traditional counterparts? What are the trade-offs that forecasters need to keep in mind, and what are the steps they need to take to use these resources effectively? We carry out a critical analysis of the existing literature in order to answer these questions. Our analysis suggests that the answer to most of these questions are nuanced, conditional on a number of factors identified in the study.

[1] RBI Chair Professor, Institute of Economic Growth (IEG)
[2] Junior Professional, New Development Bank
[3] Independent Researcher

The authors are grateful to Dr. Dweepobotee Brahma and Dr. Archana Aggarwal for their helpful comments and discussions. This work was initiated while the authors were affiliated to the National Institute of Public Finance and Policy (NIPFP)

1. Introduction

Timely forecasts of key macroeconomic indicators, such as GDP growth and inflation, are important for policymakers and market participants to gauge the health of the economy, form expectations about the future and calibrate their actions. A central bank's decision on whether to increase or decrease the monetary policy rate, a government's expectations about the tax revenue it may obtain over the coming year, and decisions regarding investments - all require forming expectations about the inflation or GDP growth in a country, which rest on a foundation of good forecasts.

In recent times, the use of big data and machine learning (to be used interchangeably with the term ML in some places in this paper) techniques are increasingly gaining popularity in various spheres[4]. A limited but growing literature is also emerging that uses these new approaches for macroeconomic forecasts. While most of the research in this field is conducted in the U.S and other developed countries' central banks, work has been undertaken recently in developing countries as well. However, a broad-based understanding about these approaches is yet to emerge. In this paper, we attempt to contribute to a better understanding of these issues by undertaking a critical review of the relevant literature.

What is big data and machine learning? As far as big data is concerned, there exists no single definition, with different studies emphasising different characteristics (Doornik and Hendry, 2015; IBM). However, it is best understood by drawing a distinction between 'designed' and 'organic' data (Rigobon, 2018). Designed data refers to surveys and administrative records: instruments that are specifically designed to capture information about an entity or a group of entities. Data on vehicular and industrial production, tax records, census information, labour surveys — all refer to designed data. Organic data on the other hand "*...is the data that is generated by individuals without them noticing they are being surveyed. It is the data in the GPS of your phone, your searches on the web, the friends in your network, the things you purchase, etc. Every time any individual acts through any of these channels it provides information about herself*" (Rigobon, 2018).

For example, typing a search query in Google and swiping a credit card can provide the owner of that data information about your consumption preferences. These preferences can be

[4] Chakraborty and Joseph (2017), for example, provide a review of ML methods and their applicability in financial and economic contexts.

aggregated, categorised, and made amenable to statistical analysis. Since this data is a by-product of an electronic transaction or activity, it is generated at a much higher frequency (or 'Velocity') than designed data, and is typically orders of magnitude bigger in terms of size, requiring special storage and processing tools. It can thus be described as high in 'Volume'. Organic data is generated from a multitude of sources, like internet searches, credit card purchases, mobile phone data, GPS systems and so on. Such data, therefore, has high 'Variety'. However, not all aspects of an organic data source may be useful for a given purpose: say, macroeconomic prediction. The greater the share of useful observations, or signals, in an organic data source, the higher its 'Veracity', and vice-versa. Together, Velocity, Volume, Variety and Veracity are termed the four "Vs" that distinguish big data from its 'smaller' counterparts (IBM). It can therefore step in and complement more traditional, designed sources of data when the latter are unavailable or delayed. These characteristics are particularly useful in carrying out macroeconomic 'nowcasts', which are an increasingly popular subset of forecasts that involves making predictions at higher-than-monthly frequencies.

Machine learning, on the other hand, is more well-defined. Very simply, it is the "*study of computer algorithms that improve automatically through experience*" (Mitchell, 1997). Statistical techniques for predicting a target variable, such as GDP growth or inflation, fall within the domain of "supervised learning". This domain of statistics is primarily concerned with relating observations of a set of inputs or predictor variables represented by X, to a supervising output/response/target variable Y, through a function f — with some associated error $\boldsymbol{\varepsilon}$ (James et. al., 2013) :

$$Y = f(X) + \boldsymbol{\varepsilon}$$

The function f determines how X and Y are related. This relationship can take many shapes — linear or non-linear — and is used to predict future values of Y, for each given value of X. For instance, if past values of GDP or inflation (X) predict their future values (Y) reasonably well and are related through a linear function f, then using the same function, we can predict Y as more data on X becomes available. As Jung et. al. (2018) note, machine learning methods do not make assumptions about the functional form of f. Macroeconomic predictions using such methods are thus dependent on first estimating the shape or the 'fit' of the function f, i.e. how

does it relate X and Y (on 'training' or in-sample data), and checking whether this function does a good job of relating X and Y in the presence of previously-unseen data ('test' or out-of-sample data).

Technically, both non-ML (regressions, for example) and ML techniques involved in macroeconomic predictions fall within the domain of supervised statistical learning. What makes ML techniques more attractive relative to their econometric or non-ML counterparts is their ability to more efficiently handle datasets with Velocity, Volume and Variety that is orders of magnitude higher than designed datasets. This can be particularly useful in economic contexts where structured data and surveys are either infrequent or do not exist, for example in a Least Developed Country (LDC) with a low-capacity statistical apparatus.

In this paper, we attempt to investigate whether the existing literature on macroeconomic forecasting has been able to establish any advantages of using big data and machine learning approaches over more traditional alternatives. In the case of big data, this is done through an analytical survey of the literature. In case of machine learning, we conduct a meta-analysis of the literature on forecasting GDP growth and inflation using ML techniques. The aim is to understand whether ML techniques are better than their 'standard' or non-ML counterparts in providing more accurate forecasts. To the best of our knowledge, such a meta-analysis is an original contribution to the literature. Finally, based on our understanding of the strengths and limitations of these new approaches, we attempt to provide guidance on how to use them for macroeconomic forecasting.

The structure of the paper is as follows. Section 2 deals with the advantages and limitations of using big data for macroeconomic forecasting, relative to traditional data. Section 3 provides an introductory review of the ML techniques most commonly used in the existing literature. Section 4 describes the dataset constructed for our meta-analysis and discusses the empirical results of the analysis. A guide for using big data and ML techniques for macroeconomic forecasting is provided in Section 5. Section 6 concludes.

2. Use of big data for macroeconomic forecasting: advantages and limitations

Regular or 'designed' data that is used in macroeconomic predictions, especially of GDP growth, comes with certain limitations. Among these, non-synchronous and lagged data releases are particularly important. The former refers to the different release schedules of different indicators. In India, for instance, hard data on monthly production of commercial

vehicles is released in the middle of the month, whereas data on monthly production of coal and crude oil is released early in the month. Similarly, data for an indicator for a given month may be released next month with a lag, as is the case with data on steel and fertiliser production. Together, non-synchronous and lagged data releases lead to a 'jagged-panel' of data or 'jagged-edge' data (Bhadury et al., 2018). This implies that at the time of making a (typically, quarterly) forecast, the forecaster will not have all the data available for all the relevant indicators, with respect to the quarter in question.

Issues with standard data used for inflation forecasts are slightly different, as price data that is part of the Consumer Price Index (CPI) and its variants are all released at the same time on a monthly basis. Lags and non-synchronicity in such data releases, therefore, do not exist. However, non-price variables that go into inflation forecasts such as liquidity, industrial measures, net exports and extraneous environmental measures like rainfall and crop yields, can be non-synchronous and lagged, leading to the aforementioned 'jagged edge' problems in inflation forecasting as well.

Different techniques such as Bridge Equation frameworks and Kalman Filter algorithms (used in Dynamic Factor Models) or Multiple Imputations as a form of stochastic iterations (Rubin, 1977) attempt to deal with the problem of jagged-edge data in different ways. Nonetheless, a key issue that a jagged-panel creates is to limit the information available to the forecaster at the time of the forecast. In other words, with regular ordered data being used from a multitude of sources, non-synchronicity and delays are inevitable. The use of big data is very effective in this context. It can help in plugging the data gaps created by these problems, along with possibly increasing the accuracy of the macroeconomic predictions being made.

How exactly does big data help? This question can be answered in terms of the four "Vs" discussed earlier. Since big data is organic and a by-product of everyday activities, the higher Velocity of its generation allows it to step in where there are issues of delays or non-synchronicity with smaller, designed data. 'Google Trends', a service offered by Google is a popular source of such high-velocity data. It shows a time series of the frequency with which a particular keyword was searched on Google, relative to all searches. The user can specify the geographical region and language for this exercise. Buono et. al. (2017) provide a good overview of the studies utilising Google Trends for predictive purposes in varied fields such as economics, finance, health and politics, among others. Woloszko (2020) constructs a weekly tracker of GDP growth forecasts for 46 countries from both the OECD and G20 blocs, covering

a wide range of economies. On average, his model utilising Google Trends performs well on a weekly basis. However, it does not outperform models that use more 'standard indicators', once they are released. Nevertheless, obtaining reasonably accurate estimates of GDP growth at a higher-than-quarterly frequency may be very useful from a policy perspective. Similarly, using satellite data on nightlights allows Bhadury et. al. (2018) to predict GDP growth estimates before all the standard high-frequency indicators become available, illustrating the utility of incorporating big data in macroeconomic forecasts.

Besides high Velocity, the higher Variety of big data allows the incorporation of information that may not be captured in standard datasets, potentially offsetting measurement biases. Take, for instance, scanner and online prices. Scanner price data refers to the information encoded in the bar codes of retail products that is generated when the barcode is scanned. This information can be utilised to create baskets of prices. With the increased popularity of online shopping, prices collected over the internet can also help construct such price baskets. Changes in the price data collected can be used to measure or predict inflation (while keeping in mind competitive and retail market behaviours online and how it affects pricing).

Besides the obvious advantage of high Velocity, price data collected from these sources can improve upon the coverage and accuracy of the surveys traditionally used to collect such data. For example, scanner data, by virtue of being 'organic', can provide a wider geographical coverage while collecting prices since it is more labour intensive to expand the reach of traditional surveys. Scanner data can also offer a more granular insight into price changes for different products (Buono et. al., 2017). Additionally, online prices can more accurately account for prices of products that are bought relatively more frequently online, such as smartphones and other consumer electronics (Rigobon, 2018). The utility of these sources of price data is well-established by now. 'The Billion Prices Project', started in 2008, collects daily price data from online retailers around the world and computes alternative measurements of inflation. It is one of the most well-known examples of using such data to measure and predict inflation. The official statistical agencies of New Zealand and Netherlands are at different stages of formally utilising scanner price data to measure price inflation (Buono et. al., 2017).

While sources of big data may serve as reliable, high-frequency proxies of indicators that help predict macroeconomic series like GDP growth and inflation, it is useful to keep their potential limitations in mind. The biggest limitation of big data is that it is not (yet) a representative

sample of different types or different parts of economies (Rigobon, 2018; Buono et. al., 2017). To illustrate, the intuition behind using Google Trends to predict unemployment is that individuals are likely to search for jobs online if they have been laid off. A high volume of such searches would show up in Google Trends and could indicate the early stages of a recession. However, the implicit assumption here is that the search data is capturing a representative sample of the geography in question. People are more likely to utilise the internet to search for jobs in developed economies and there may be a selection bias within such countries too. A younger, more educated demographic in urban areas, for instance, may be more likely to search for jobs using the internet as opposed to their lesser educated peers in rural areas.

Similarly, using scanner and online prices to construct price indices is likely to capture the consumption baskets of those who have a certain level of purchasing power, financial literacy, and access to the internet. Existing price collection methods for CPI take care to judiciously survey 'Working-Class' markets and 'Middle-Class' households. Complex stratification ensures that a representative sample is selected, making it more statistically driven than big data's 'behaviour driven' approach. The latter approach is much more likely to be biased towards those households, individuals and products which constitute the digital markets.

This 'digital divide' and how it creates sampling biases across and within economies must be kept in mind (Buono et. al., 2017) while using big data. The existence of a digital divide does not mean, however, that such data is not useful for macroeconomic forecasting in, say, a developing economy. As discussed earlier, big data can be useful for filling-in missing data and providing proxies for lower frequency or unattainable data. The user, however, must be conscious of the nature of its sampling bias and use it to complement, rather than substitute 'smaller', designed data. "big data hubris", therefore, must be avoided (Lazer et. al., 2014).

Another limitation of big data arises due to the fact that most of it is generally proprietary and made available to the public, licensee or official statistical agency after some pre-treatment and/or aggregation. Lack of knowledge about the pre-treatment of the raw data, along with a lack of access to the same, introduces some opacity about the measurement quality of the data (Kapetanios and Papailias, 2018). Moreover, when data is proprietary, the nature of access may also change over time. Google Trends may stop being available as a free service or may be stopped entirely. Retailers may forbid web scraping algorithms from compiling prices over the internet. Relatedly, the nature of measurement may also change over time. Google keeps tinkering with its search algorithm in order to provide better search results and user experience.

These 'algorithm dynamics' may affect the search queries that end up being correlated with an event or series of interest. Algorithm dynamics may also be differentiated along axes of geography, showing different search results and query-prompts to individuals in different locations. Change in the nature of access and measurement can disrupt the forecaster's ability to use a big data source and to make accurate predictions (Buono et. al., 2017; Kapetanios and Papailias, 2018; Lazer et. al., 2014).

Finally, since the start date of many big data sources is relatively recent, the temporal dimension of the source may be limited for robust parameter estimations and out-of-sample evaluations. This issue is compounded by the fact that the reference series — inflation or GDP growth — are usually of a lower frequency (quarterly or annual) than the big data source. Thus, the latter will have to be aggregated to match the frequency of the target variable: offsetting the gains that may have been derived from the higher frequency and volume of the big data being used (Buono et. al., 2017).

3. Machine learning techniques

ML methods most commonly used in the literature on forecasting growth and inflation can be broadly divided into 3 types: Penalisation, Tree-based and Neural Networks. We introduce and describe each of these types below.

3. 1 Penalisation

Parameter estimation used in most standard linear prediction methods follows the Ordinary Least Squares (OLS) procedure. Take a standard linear regression as an example:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

Estimating β using OLS requires minimising the sum of squared residuals (henceforth RSS):

$$RSS = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2$$

This estimation works well in many situations. However, if the number of predictors is close to, or greater than, the number of observations, then estimation via OLS can lead to overfitting, wherein the estimated parameters have low bias but high variance. This means that they might

fit the relationship between the predictors and the response variable closely in the available sample (low bias), but are sensitive and prone to vary with the introduction of new observations — they will perform well in-sample but not out-of-sample (high variance), which is not good for a forecasting model.

However, with some modifications to the OLS procedure, some bias in parameter estimation can be introduced in order to reduce their variance, thereby improving their predictive performance. These modifications are provided by 'Regularisation' or 'Shrinking' techniques such as Ridge regression, Least Absolute Shrinkage and Selection Operator (LASSO), and Elastic net.

3.1.1 Ridge Regression

Ridge regression introduces a penalty in the estimation of parameters β_j. This penalty is a simple addition to the OLS estimation procedure seen earlier:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2$$

The penalty $\lambda\sum_{j=1}^{p}\beta_j^2$, known as 'shrinkage penalty' is small when $\beta_1, \beta_2, \ldots, \beta_j$ are

close to zero. Therefore, minimising the RSS subject to the constraint imposed by the shrinkage penalty involves 'shrinking' the coefficients towards zero. The tuning parameter (or 'hyperparameter') λ determines the relative importance of the penalty term. As is evident from the equation, if $\lambda = 0$, this term has no effect and coefficients are estimated using the standard OLS procedure. As λ tends towards infinity, the effect of the penalty term grows and coefficient estimates β_j shrink towards zero. A different set of coefficient estimates are produced for each value of λ. Therefore, it is important to carefully select a value for the tuning parameter. This can be done using cross-validation, where ridge regressions with varying shrinkage penalty values can be modelled to see where the bias-variance tradeoff is minimized, leading to the lowest MSE.

3.1.2 Least Absolute Shrinkage and Selection Operator (LASSO)

While ridge regressions regularize parameters estimates with high variance in order to improve their predictive ability, in datasets where the number of potential predictors is large, challenges still remain in interpreting models. For example, many predictors will have negligible coefficients, i.e. they will not add value in terms of predictive content. In such situations, it can

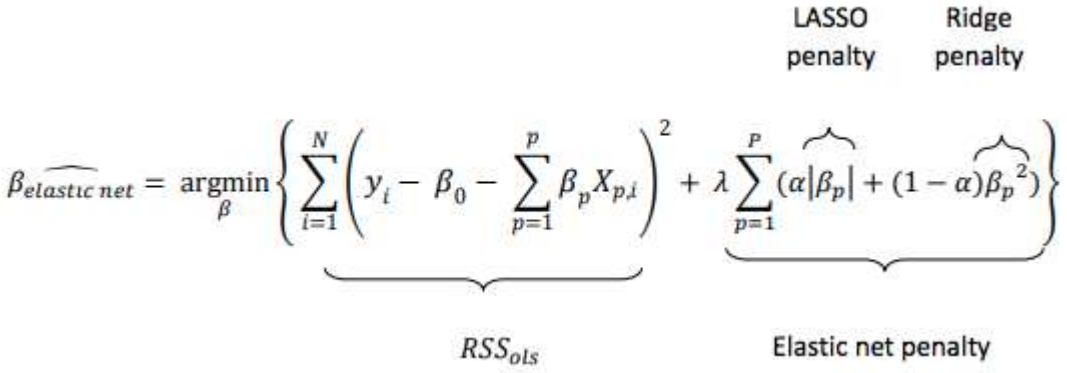
be difficult to parse which predictors are the most useful. The LASSO technique, much like ridge regression, shrinks the parameters towards zero to reduce their variance. However, it lets some parameters shrink exactly to zero when the tuning parameter λ is large enough, so that those variables drop out of the model. This creates more sparse and more interpretable models with only those variables included that are relevant for predicting the response variable. This is achieved by changing the shrinkage penalty from $\lambda \sum_{j=1}^{p} \beta_j^2$ to $\lambda \sum_{j=1}^{p} |\beta_j|$:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p}|\beta_j| = RSS + \lambda \sum_{j=1}^{p}|\beta_j|$$

Like ridge regression, minimising the above equation entails shrinking the parameter estimates towards zero. Due to the nature of the constraint set by the penalty term in LASSO, some variables are allowed to shrink to zero, given a large enough value of the tuning parameter. As before, cross-validation should be used to determine the optimum value of the tuning parameter to balance the bias-variance trade-off.

3.1.3 Elastic Net (ElNet)

A combination of LASSO and ridge regressions are Elastic nets, which allow for a model to not be too dependent on the selection made by LASSO, yet allow for model interpretability by limiting the number of variables from a large predictor space. This is done by minimizing the RSS subject to both ridge and LASSO penalties as shown below.



Source: Mahajan and Srinivasan (2020)

Here, we have two hyperparameters λ and α. λ, as before, ranges from 0 to infinity and increases the degree to which parameter estimates are shrunk. α, the new hyperparameter, determines the

relative importance of the LASSO and ridge penalties. It ranges from 0 to 1. If $\alpha = 0$, then the LASSO penalty drops out and we have a ridge regression. If $\alpha = 1$, the ridge penalty drops out and we have a LASSO formulation. This combination allows for relatively stable coefficients (with low variance) in the presence of 'Fat'[5] data and multicollinearity (Mahajan and Srinivasan, 2020).

3.2 Tree-Based methods

Tree-based methods involve partitioning the available predictor-space into a number of regions based on pre-specified rules, and making a prediction for the response variable in each region. Since this partitioning can be graphically represented in a 'decision-tree', such methods are called 'tree-based' methods or Classification and Regression Trees (CART). Popular examples include Random Forests and its 'Bagged' or 'Boosted' implementations, which involve 'growing' multiple trees. In order to understand them, however, we first need to understand how a single tree is grown.

First, the predictor-space composed of predictors $X_1, X_2, \ldots, X_j$is divided into distinct, non-overlapping regions $R_1, R_2, \ldots, R_j$. This is commonly done via recursive binary splitting. This entails taking all predictor observations as part of a single region at first. This predictor-space is then split on the basis of the predictor and a 'cutpoint' that minimises the RSS across the two resulting regions. In other words, this partitioning rule searches across all predictors R_j to find an appropriate cutpoint s, such that we have two regions:

$$R_1(j,s) = \{X|X_j < s\} \; and \; R_2(j,s) = \{X|X_j \geq s\}$$

where all values of X_j belonging to R_1 are less than s, and the rest belong to R_2. The goal behind this is to find the predictor $X_j$and cut-point s, that minimise the RSS across all regions:

$$min \; (\Sigma \; (y_i - \hat{y}_{R_1})^2 + \Sigma (y_i - \hat{y}_{R_2})^2)$$

[5] See section 5.1 below

Here, $\hat{y}_{R_1}$ is the mean of the response variable for the training observations in region R_1 and $\hat{y}_{R_2}$ is the mean of the response variable for observations in region R_2.

This process is then repeated in the two regions R_1 and R_2, with the aim to find the predictor and cut-point that would minimise the overall RSS once more, thereby creating a third region. Subsequent regions are created in the same manner. This binary splitting continues until a stop-criterion is reached i.e. a maximum number of nodes or splits, or a minimum number of observations in each region. Once the splitting concludes and we have regions $R_1, R_2, \ldots, R_j$, we predict the value of the response variable in each region to be its mean[6], and a tree is thus grown.

With recursive binary splitting, it is easy to have a very high number of regions. Theoretically, it is possible to have a region for each training observation of the predictors, resulting in a perfectly fitted tree. However, overfitting leads to high variance and thereby, poor out-of-sample predictions. The solution is 'Pruning' of trees. Pruning techniques complemented by appropriate cross-validation allow the user to select sub-trees that balance in-sample and out-of-sample prediction accuracy[7].

Tree-based methods have many advantages. They are useful in modelling interactions between different predictors in a better manner than standard techniques, as the latter tend to create more regressors relative to data points (Mahajan and Srinivasan, 2020). The relative importance of different predictors is also more intuitively understood in a decision-tree format. Nonetheless, while single trees can be intuitive to understand, their predictive performance can be poor. The solution to this problem is a group of models that grows multiple trees and are known as 'Bagging', 'Random Forests' and 'Boosting' respectively.

Growing multiple trees and aggregating the predictions that they provide, improve upon the performance of the tree-based methods significantly. If all the predictors are available as candidates for splitting and the average prediction of multiple trees grown using bootstrapped samples is considered, it is called 'Bagging'[8]. Bagging reduces the variance that accompanies a single tree. However, if the predictors are highly correlated, the reduction in variance will not

[6] The mode of the response variable can also be chosen as the de-facto prediction. Mean, however, appears to be the default choice in macroeconomic prediction.
[7] For details on these techniques, see James et. al. (2013), pg. 307.
[8] 'Bagging' is short for 'bootstrap aggregation'. It is a general purpose technique and can be used for other statistical learning methods too, such as regressions.

be large since most trees will be grown in a similar order. Hence, growing multiple trees with different variables is useful in order to reduce prediction variance further. This objective is achieved by the group of models called 'Random Forests'.

Random Forests 'decorrelate' the trees by modifying the Bagging methodology. At each split in a tree, the learning algorithm is not allowed to consider all the available predictors. Using bootstrapping principles, multiple trees are then grown using multiple draws from the training sample. This is useful as this leads to different trees created out of different variables, which lowers the out-of-sample variance of the model[9].

Random Forests or Bagging techniques develop multiple trees independently of each other and then average them. A third technique called Boosting[10] is slightly different from these two approaches as it grows trees sequentially. The advantage of this sequential process is that the residuals of the past trees are used to provide information that is used to grow newer trees. This allows the algorithm to learn relatively 'slowly' and improve the fit of each tree where its residuals are large. This sequential growing of trees lowers the forecasting error.

While techniques like random forests, bagging and boosting can improve forecast accuracy over the predictions of single trees, it can be difficult to interpret the resulting models. For example, it is much easier to visually delineate important splits (variables) in a single tree, but an average of predictions over multiple trees cannot be as clearly visualised. Hence, tree-based methods that grow multiple trees improve predictions at the expense of interpretability. But, one can use prediction measures such as RSS and Gini indices to obtain summaries of the importance of each predictor, even if the relationship is not visually clear.

3.3 Artificial Neural Networks

Artificial Neural Networks, also known simply as Neural Networks, are a set of techniques that mimic biological decision-making processes using a set of inputs, analysing their relative importance, and using that information to determine an output. The earliest models of such nodal learning were known as 'perceptrons' (Rosenblatt, 1958), that used inputs $(X_1, X_2, X_3 \ldots X_i)$, and their weights $(w_1, w_2, w_3 \ldots w_i)$, to determine the output, as the weighted sums of the inputs. Neural Networks however, deal with more complex modelling objectives,

[9] See James et. al. (2013), pg. 320.
[10] As with Bagging, Boosting too is a general purpose technique applicable on other statistical learning methods (see footnote 8). However, it is most popularly used on Tree-based methods.

and these are achieved by building sequential layers of multiple perceptrons, rather than a single one. These intricate network of perceptrons can be categorized into three parts, i.e., an input layer, an output layer, and one or more hidden layers in between the input and output layers, each consisting of multiple nodes or 'neurons'.

There are two more ways in which Neural Networks improve over simple perceptrons. The first is to introduce the possibility of non-linearity in the impact that weighted inputs have on output. This works through the 'activation function' that imposes a non-linear transformation on the weighted inputs before they provide information to the next neuron in the chain. Some of the commonly used non-linear functions used include sigmoid, Tanh and Rectified Linear Unit (ReLU) functions. The second way that a Neural Network improves its performance is by a process known as 'backword propagation'. This is a feedback mechanism from the output error in any particular iteration (difference between the estimated output in that iteration and its actual value) back to the inputs, enabling an adjustment in their weights in the next iteration, so that more and more accurate predictions become possible.

Figure 1 below is a schematic representation of a single iteration of the process described above. There is an input layer that has two neurons (independent variables), a single hidden layer with three neurons, and an output layer with a single neuron (target or dependent variable). The box highlights the first step in the process of running the model. The importance of each of these two input neurons (X_1, X_2) to the next neurons in the hidden layer (here 3, 4 or 5) is determined by their weights (w_{13}, w_{14}, w_{15}) and (w_{23}, w_{24}, w_{25}) respectively. Conversely, the information that can be received by any particular neuron in the hidden layer is a weighted summation of X_1 and X_2. Apart from the weighted inputs, another term that is added in this transformation is a 'bias', which is a constant, and similar in concept and function to the intercept term in a linear regression. These are represented by ($\theta_3, \theta_4, \theta_5$) for neurons 3, 4 and 5 respectively. Finally, the chosen non-linear activation function is imposed on the weighted inputs (plus bias), before the results are transferred to the next neuron.

Next, we look at Figure 2 below. The box now highlights the next step in running the model. The transformations described in the last paragraph create derived variables for Node 3, 4 and 5 respectively. The process in the first step is now repeated using these derived variables as inputs. This involves first assigning weights and biases to these inputs and then imposing the

activation function on this transformation. The result is the output for the model for this round of iteration[11].

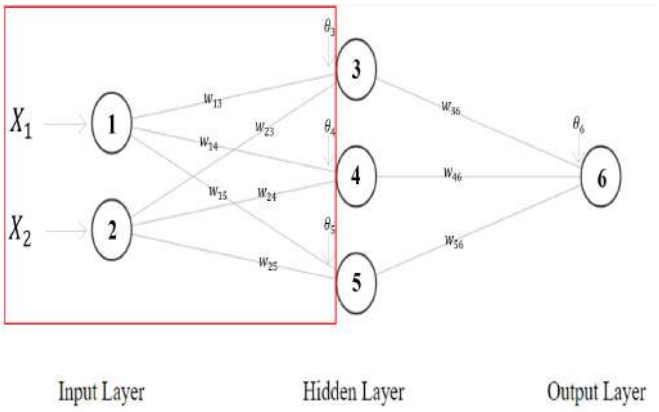


Figure 1: Input to Hidden Layers

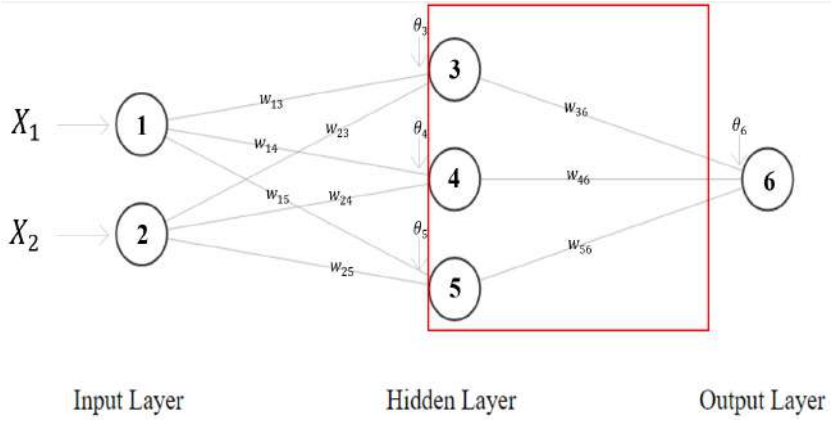


Figure 2: Hidden Layers to Output

Next, as described earlier, a process of 'Backward Propagation' introduces circularity to the process in order to let the network learn which connections to emphasise, and thus bring the final output closer to a desired value. This process uses a technique called 'gradient descent'

[11] Iterations are called epochs in the neural network literature.

to efficiently find the direction and magnitude of change needed in the weights and biases used in all the steps of the second and subsequent iterations[12]. This repeated iteration helps the model learn more and more about the data and provide an accurate prediction of the output or target variable.

While Neural Networks were initially designed to analyse cross-sectional data, they are increasingly being used in time series analysis by modifying the model to accommodate time dimensions as well. Deep Neural Networks (DNNs) build on this concept by having an increased number of hidden layers, leading to more complex transformations and functional relationships. However, DNNs alone cannot deal with time series data adequately, as the data is revealed to each node in the hidden layer simultaneously, as opposed to sequentially, to account for temporal dimensions. This is why using only DNNs in this iteration may not necessarily help to discern complex relationships that are shaped over time. Recurrent Neural Networks (RNNs), on the other hand, allow the incorporation of information sequentially, which is suitable for time series analysis. The structure of RNNs is the same as regular neural networks, with layers of neurons feeding-forward information. The basic difference is that RNNs create multiple 'copies' of the same network with each copy corresponding to each observation sequentially. The output from the first copy informs how the second one will determine its output, which in turn determines the third, the fourth, and so on. Needless to say, a drawback of such complicated versions of neural networks is that the computational power required to run them goes up substantially.

4. A meta-analysis of the literature

In this section, we carry out a meta-analysis of the literature that uses machine learning for macroeconomic forecasts. Our goal is to understand whether ML techniques tend to perform better than their non-ML counterparts in forecasting growth and inflation. Further, we want to find out the conditions under which they perform better: longer or shorter forecast horizons, emerging or advanced economies, and so on.

4.1 Data

In order to conduct this meta-analysis, we assemble a dataset of papers which utilise ML techniques in their forecasts of growth and inflation. Another criterion for the selection of

[12] The weights and biases in the first iteration do not have any feedback and thus have to be assigned arbitrarily.

papers was that their forecast period was more recent, definitely after the year 2000. While this criterion is inherently arbitrary, the aim is to ensure that our data is relevant, since the global economy and inflation rules by central banks changed significantly close to and after the year 2000[13].

In order to compile our dataset, we traced the thread of citations that met our criteria, beginning with the most recently available papers. In other words, a 2021 paper meeting our criteria is included, its citations are searched to find more relevant papers, and this process is repeated for each of the papers identified through this process. Through this search process, we have been able to identify sixteen papers that use at least one type of ML technique in forecasting either growth or inflation[14]. Each of these papers also have one quarter and one year ahead forecasts that belong to the post-2000 period. Amongst the sixteen papers, there is a relatively even split within four groupings: growth in advanced economies, inflation in advanced economies, growth in emerging economies and inflation in emerging economies.

Within advanced economies, Jahn (2018), Biau & D'elia (2010), Jung et. al. (2018) and Richardson et. al. (2019) look at growth forecasting in EU countries, Japan, USA and New Zealand – Jung et. al. is the only paper in the entire set that looks at both advanced and emerging economies (i.e. Mexico, Philippines and Vietnam). The papers in this set use Univariate, Multivariate, Factor-based, Penalisation, Tree-based and Neural Network methods on their samples. While Jahn (2018) uses regular Neural Networks, Jung et al. (2018) use Recurrent Neural Networks.

Terasvirta et al. (2004), Medeiros et al. (2019), Marcellino (2007), Nakamura (2005), and Baybuza et al. (2018) look at inflation forecasting in G7 countries and Russia. Moreover, Nakamura (2005), Terasvirta et al. (2004), Medeiros et al. (2019) and Marcellino (2007) all look at the United States over different time periods using different methods. The papers in this subset use Univariate, Multivariate, Penalisation, Tree-based and Neural Network methods. Terasvirta et al. (2004) use Neural Networks, whereas Marcellino (2007) looks at Autoregressive Neural Networks.

[13] The Taylor Rule, for example, was first proposed around 1992-93 (Taylor, 1993) and modified in 1999 (Taylor, 1999)

[14] The ML techniques used in the identified papers are classified under Penalisation, Tree-based or Neural Network methods. Non-ML techniques are classified under Univariate, Multivariate or Factor-based methods

Within emerging economies, a range of countries are covered. For growth, Sanyal & Roy (2014) and Roy et al. (2016) look at India whereas Tiffin (2016), Jung et al. and Chuku et al. (2017) look at Lebanon, Mexico, Philippines, Vietnam, South Africa, Nigeria and Kenya. The papers in this subset use Univariate, Multivariate, Factor-based, Penalisation and Neural Network methods.

With respect to inflation in emerging economies, the subset contains forecasts for Chile (Leal et al., 2020), India, South Africa, China (Mahajan & Srinivasan, 2019) and Brazil (Garcia et al., 2018). The papers in this subset use Penalisation, Univariate, Factor-based, Tree-based and Neural Network methods. While Jung et al. (2018) uses Recurrent Neural Networks, Sanyal & Roy (2014) and Roy et al. (2016) use regular Neural Networks. For inflation, within both advanced and emerging economy subsets, adaptive iterations of Penalisation techniques have also been considered (Leal et al. (2020), Medeiros et al. (2019), Garcia et al. (2018)).

Several papers such as Richardson et al. (2019), Sanyal & Roy (2014), Jung et al. (2018), Tiffin (2016) and Terasvirta (2004) use Combination Forecasts, which have not been included in the meta-analysis since they cannot be strictly categorized under the six classes of ML and non-ML techniques identified in the literature. Table C1 in the Appendix summarises the salient characteristics of the papers included in our analysis.

The papers selected for our study provide us with a sample of 313 forecasts. Table 1 provides summary statistics of "adjusted RMSEs"[15] of the forecasts of various subgroups of this full sample. As seen in Table 1, the number of observations or forecasts for ML (155) and non-ML (158) techniques are almost the same, with the latter being marginally higher. While adjusted RMSEs belonging to ML techniques have a smaller mean, non-ML techniques have a smaller standard deviation, implying a smaller dispersion in their forecasts. There are more observations for shorter horizon (1-quarter-ahead) forecasts (165) than for longer ones (4-quarters-ahead) (148), with shorter horizon forecasts displaying smaller adjusted RMSEs, on average. Shorter horizon forecasts also have a smaller dispersion, indicating lower volatility in their predictions. The number of advanced and emerging economies in our dataset are almost equal, with the former group showing smaller adjusted RMSEs on average. Emerging economies, however, display a smaller dispersion in their forecast errors.

[15] For a definition of RMSE and "adjusted RMSE", see section 4.2 below.

There are 93 more observations available for inflation relative to growth, in the collected dataset. Growth forecasts appear more accurate, with a smaller average adjusted RMSE. Inflation forecasts have a smaller dispersion, however, indicating less volatility in its predictions. Within growth forecasts, there are 8 more observations belonging to ML techniques relative to non-ML techniques. While ML techniques display a higher average accuracy due to a smaller average, non-ML techniques show lower volatility through a smaller dispersion. Within inflation forecasts, there are 11 more observations for non-ML techniques relative to ML techniques. Here too, ML techniques have a smaller average but higher dispersion.

Table 1: Summary Statistics of Adjusted RMSEs by Groups of Forecasts

Group	n	Mean	SD	Min	Max
Non-ML	158	0.016	0.399	-1.108	2.008
ML	155	-0.07	0.568	-1.988	1.993
Horizon					
4-quarter	148	0.0206	0.413	-1.988	1.315
1-quarter	165	0.002	0.411	-1.472	2.008
Country type					
Advanced	157	-0.074	0.545	-1.988	1.993
Emerging	156	0.0211	0.425	-1.472	2.008
Indicator type					
Inflation	203	-0.038	0.332	-0.963	1.242
Growth	110	-0.004	0.695	-1.988	2.008
Growth					
Non-ML	51	0.158	0.581	-1.108	2.008
ML	59	-0.145	0.758	-1.988	1.993
Inflation					
Non-ML	107	-0.051	0.249	-0.917	0.854
ML	96	-0.024	0.408	-0.964	1.242
Method					
Univariate	101	0.006	0.365	-0.917	2.008
Multivariate	43	0.083	0.483	-1.108	1.315
Factor	14	-0.118	0.325	-0.997	0.118
Penalisation	62	0.020	0.346	-0.947	0.885
Trees	21	0.012	0.507	-0.825	1.185
Neural Networks	71	-0.173	0.715	-1.988	1.993
SVM	1	-0.0348	-	-0.0348	-0.0348

Source: Author's calculations.

For the purposes of the meta-analysis, we also club each ML and non-ML technique used in the collected forecasts into six broader classes. Non-ML techniques can be clubbed under

Univariate, Multivariate and Factor-based methods, and ML techniques can be clubbed under Penalisation methods, Tree-based methods, and Neural Networks. Among these six classes of methods that encompass the ML and non-ML techniques used in the collected forecasts, the highest number of observations belong to Univariate techniques. This is expected, since autoregressive techniques are commonly used as benchmarks in macroeconomic forecasts. The lowest number of observations correspond to a single forecast that uses the Support Vector Machine (SVM) method, followed by Factor-based methods. While the former method is an ML technique, it cannot be strictly categorised under any of the three classes of ML techniques identified. There are also an insufficient number of forecasts in the dataset that use this method. Thus, while it is included in the meta-analysis as an ML technique wherever ML techniques as a group are compared with their non-ML counterpart, it is excluded in analyses regarding the classes of methods.

Excluding the single SVM forecast therefore, the lowest number of observations correspond to Factor-based methods. Among the six classes, the lowest average adjusted RMSEs belongs to Neural Networks followed by Factor-based methods. Neural networks, however, also display the highest volatility in their predictions, evident by the group's standard deviation. Factor-based methods, on the other hand, display the lowest volatility among the six classes. Initial results thus suggest that a growth or inflation forecaster would do well to use Factor-based methods due to their low volatility of results and low average forecast error.

4.2 Methodology

From the papers included, we extract the Root Mean Squared Error (RMSE) of their respective forecasts, by each forecasting method using:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Predicted_i - Actual_i)^2}{n}}$$

The RMSE is a commonly used measure to indicate the precision of a forecast —the lower the difference between the predicted and actual observation, the better the forecast. To keep our analysis focussed, we only extracted RMSEs for the 1-quarter-ahead (1Q-ahead) and 4-quarters-ahead (4Q-ahead) forecast horizons. These represent short-term and long-term predictions of growth and inflation. Once extracted, the RMSEs are organised by forecast

method and country (Appendix Tables A1-A4). We also specify whether each country is an 'advanced' or an 'emerging' economy.

It may be noted here that the RMSEs for different countries may have country fixed effects as a strong underlying factor. Before making comparisons between the RMSEs produced by ML and non-ML technique based forecasts, these effects need to be controlled. To do this, the mean RMSE of each country (across all studies) is subtracted from the RMSE of each forecast based on that country. The recalculated RMSEs are termed "adjusted RMSEs". These are presented in Tables A5 to A8 in the Appendix.

On the adjusted RMSEs, we conduct tests of statistical significance in order to gauge the relative performance of ML and non-ML techniques. It may be noted that while our dataset is fairly exhaustive in terms of including relevant papers, it is still a small sample (345 forecasts) because the literature that uses ML techniques in forecasting growth and inflation is still in nascent stages. As is well known, a small sample from an unknowable population violates the assumptions that go into the more popularly used T-test (Ch-13.3, Freund, 2000). Non-parametric tests like the Mann-Whitney, on the other hand, are useful in this scenario because they make few, if any, assumptions about the distribution of the population and are relatively more suitable for smaller samples. We find that a non-parametric approach is used in other meta-reviews of ML techniques as well, for instance in the context of financial market forecasting (Ryll & Seidens, 2019). For these reasons, we use the Mann-Whitney U-test as the first step in determining statistical significance.

The U-test provides us a first layer of inference with respect to statistical significance. However, being a two-tailed test, it only tell us whether the ML and the non-ML forecasts have statistically significantly different performance levels or not. This does not tell us clearly whether the ML forecasts are better than the non-ML ones. We get some indication on which of these two has a better performance from the p-order, which is the probability, using random draws, that the average (adjusted) RMSE of the first group in our sample (non-ML forecasts) is greater than that of the second (ML forecasts). Next, for more robustness, we complement the U-test with the T-test.

We conduct a total of 82 tests. However, only those results that have a significant p-value for both the U-Test and the T-test are considered significant or conclusive. These results are described in Section 4.3.

4.3 Results

The first set of statistical tests described in Section 4.2 are conducted on the 'superset' or the entire sample. The superset contains the adjusted RMSEs for both growth and inflation, for both advanced and emerging economies, and both one quarter and four quarter ahead horizons. Subsequently, tests are conducted separately on growth and inflation, and within those two, separately for types of countries and for forecast horizons. Results can be seen in Table 2. The results highlighted in green have statistically significant U and T tests.

As seen from the rows highlighted in green, ML based forecasts have superior performance for the sample for growth forecasts as a whole, as well as the sample for longer run growth forecasts (with a horizon four quarters ahead) and the sample for growth forecasts in emerging economies. Additionally, the sample for growth forecasts in advanced economies at the four quarter ahead horizon also shows superior ML performance.

While the results for the superset are not conclusive, they do indicate better performance by ML techniques, since the p-order implies a 55% chance that a random draw from the entire sample would yield higher (adjusted) RMSEs for non-ML methods than for ML methods. The T-test in the final column also shows that the average non-ML adjusted RMSE is greater than its ML counterpart. The relatively better performance of ML techniques in forecasting growth, appears to be driving the result for the full sample including growth and inflation.

Next, we dive deeper in order to gauge the performance of different types of ML techniques. In order to do this, we compare each of the three classes of ML techniques (Penalisation methods, Tree-based methods and Neural Networks) against each class of non-ML techniques (Univariate, Multivariate, Factor-based methods) to obtain more detailed results about their relative performance. These comparisons yield nine pairs, seen in Table 3. U-Tests and T-tests are conducted first on these samples, and we find that the only significant result here is that Neural Network based forecasts have done better than Multivariate methods.

We finally partition these nine samples (each having one class of ML and one class of non-ML based forecasts respectively) further, in three alternative ways. The first partition separates those having growth forecasts from those with inflation forecasts. The second partition separates developing from developed countries. The third partition separates shorter horizon forecasts from those with longer horizon. We repeat the statistical tests for these new samples. For brevity, only the significant results from these tests are shown in Table 4.

Table 2: Significance tests: ML vs non-ML

	# of observations		Mann-Whitney U test	Student's T-test			
	Non-ML	ML	p-value	p-order (non-ML > ML)	p value (H1≠0)	p value (H1<0)	p value (H1>0)
Superset	158	155	0.146	0.548	0.122	0.9390	0.0610*
Inflation	107	96	0.801	0.490	0.4696	0.285	0.7152
Inflation - 1Q	52	44	0.749	0.481	0.7552	0.3776	0.6224
Inflation - 4Q	55	52	0.890	0.492	0.6217	0.3108	0.6892
Inflation – EM	51	52	0.331	0.556	0.3177	0.8412	0.1588
Inflation – AE	56	44	0.131	0.412	0.1373	0.0687*	0.9313
Inflation – AE & 1Q	28	21	0.332	0.418	0.3008	0.1504	0.8496
Inflation – AE & 4Q	28	23	0.869	0.486	0.2756	0.1378	0.8622
Inflation – EM & 1Q	24	23	0.708	0.533	0.4781	0.7610	0.2390
Inflation – AE & 4Q	27	29	0.363	0.572	0.4873	0.7564	0.2436
Growth	51	59	0.0241***	0.625	0.9901	0.0197	0.0099***
Growth – 1Q	24	28	0.569	0.454	0.67	0.335	0.665
Growth – 4Q	27	31	0.0009***	0.754	0.0012***	0.9994	0.0006***
Growth – EM	19	38	0.0004***	0.787	0.001***	0.999	0.0001***
Growth – AE	32	21	0.414	0.432	0.6497	0.3248	0.6752
Growth – AE & 1Q	4	15	0.841	0.533	0.531	0.2658	0.7342
Growth – AE & 4Q	15	23	0.0004***	0.842	0***	1	0***
Growth – EM & 1Q	20	13	0.396	0.412	0.7726	0.3898	0.6102
Growth – EM & 4Q	12	8	0.969	0.495	0.7240	0.3620	0.6380

*** = 1%, **=5%, *=10%; H0 (null hypothesis) is that group means are equal [mean (non-ML) – mean (ML) = 0]; Green cells represent ML outperforming non-ML. Red cells represent the converse.

Table 3: Significance tests (pairwise): superset

	# of observations		Mann-Whitney U test	Students T-test			
	Non-ML	ML	p value	p order (non-ML > ML)	p value (H1≠0)	p value (H1<0)	p value (H1>0)
Penalised vs Univariate	101	62	0.512	0.531	0.8161	0.4080	0.5920
Penalised vs Multivariate	43	62	0.804	0.514	0.4648	0.7678	0.2324
Penalised vs Factor	14	62	0.703	0.533	0.1714	0.0857*	0.9143
Tree-based vs Univariate	101	21	0.353	0.565	0.9649	0.4824	0.5176
Tree-based vs Multivariate	43	21	0.607	0.540	0.5953	0.7023	0.2977
Tree-based vs Factor	14	19	0.606	0.554	0.362	0.181	0.819
Neural Network vs Univariate	101	71	0.2104	0.556	0.0548**	0.9726	0.0272**
Neural Network vs Multivariate	43	71	0.1185*	0.587	0.0247**	0.9876	0.012**
Neural Network vs Factor	14	71	0.722	0.531	0.655	0.6725	0.3275

*** = 1%, **=5%, *=10%; H0 (null hypothesis) is that group means are equal [mean (non-ML) – mean (ML) = 0]; Green cells represent ML outperforming non-ML. Red cells represent the converse.

Table 4: Significance tests (pairwise) : significant results by indicator, country-type and forecast horizon

		# of observations		Mann Whitney U test	Students T-test			
Pairwise	Condition	Non-ML	ML	p-value	p-order (non-ML > ML)	p value (H1≠0)	p value (H1<0)	p value (H1>0)
Penalisation by Univariate methods	Growth	23	18	0.0554**	0.676	0.0314**	0.9843	0.0157***
Penalisation by Multivariate methods	EM	19	33	0.0795*	0.352	.032**	0.0161**	0.9839
Tree by Univariate methods	AE	48	10	0.0235**	0.727	.1095	0.9453	0.0547**
Tree by Multivariate methods	EM	19	11	0.0051**	0.196	.0178**	0.0089***	0.9911
	AE	24	10	0.0026***	0.821	0.0009***	0.9995	0.0005***
Neural Networks by Univariate methods	Growth	23	37	0.0355*	0.662	.0229*	0.9885	0.0115**
	EM	53	29	0.0883*	0.615	.0506*	0.9747	0.0253**
	4Q	49	44	0.0221**	0.638	.02049**	0.9875	0.0125**
Neural Networks by Multivariate methods	Growth	25	37	0.0492**	0.648	.0407**	0.9796	0.0204***
	AE	24	42	0.0233***	0.668	.0085***	0.9958	0.0042***
	1Q	16	27	0.0390**	0.310	.2590	0.1295*	0.8705
	4Q	27	44	0.0022***	0.714	0.0017***	0.9991	0.0009***

*** = 1%, **=5%, *=10%; H0 (null hypothesis) is that group means are equal [mean (non-ML) – mean (ML) = 0]; Green cells represent ML outperforming non-ML. Red cells represent the converse.

Pairwise results from Table 4 tests show that:

1. Penalisation methods provide better forecasts than univariate methods in forecasting growth. Multivariate methods perform better than penalisation methods in the context of emerging economies.
2. Tree-based methods provide better forecasts than Univariate methods for advanced economies. The same holds true when Tree-based methods are compared with Multivariate methods. However, Multivariate methods also show relatively superior performance in emerging economy contexts.
3. Neural Networks provide better forecasts than Univariate methods for emerging economies, longer forecast horizons (four quarters ahead) and in forecasting growth. When compared with Multivariate methods, the same conditions hold, with the exception of the country type — Neural Networks do better in advanced economies when compared against Multivariate methods. Multivariate methods also outperform Neural Networks at shorter forecast horizons (one quarter ahead).

4. No ML class performed significantly better than their factor-based non-ML counterparts.

From these results, there seems to be some indication that ML methods tend to outperform their non-ML counterparts in advanced economies, for longer forecast horizons, and in predicting growth. Is there some explanation behind these results? In other fields, there is some evidence to suggest that ML methods tend to outperform their non-ML counterparts in the presence of strong non-linearity in the relationship between independent and target variables. Are non-linearities more accentuated in forecasting growth, over longer horizons or particularly for advanced economies? This is outside the scope of this paper and a deeper analysis would be needed to examine the underlying causal implications.

5. Using big data and machine learning for economic forecasting: a user's guide

The previous sections aimed to provide a comparative analysis of the use of big data and ML techniques in economic forecasting, against their standard counterparts. This section will discuss some of the practical issues that forecasters, who wish to make use of big data and ML techniques for macroeconomic forecasting, have to keep in mind. This includes describing the choices one has to make at each stage of the forecasting process.

The first step involves some basic pre-checks on the data. These include tests for stationarity (eg. ADF test), and structural breaks (eg. Zivot-Andrews test), that are standard practices when regular, traditional databases are used for forecasting. Testing for stationarity is important since a stationary series would largely behave consistently over time, making it easier to predict the future using past observations. Moreover, many forecasting models assume stationarity of the data (Sanyal and Roy, 2014). Based on these tests, appropriate data transformations can be conducted, such as making the data stationary.

Once the basic pre-checks and necessary data transformations have been conducted, the forecaster needs to select the relevant predictors from the set of the available *n*, also known as 'feature selection' in the machine learning literature. One way of doing this is through cross-correlation matrices, cointegration tests, and similar procedures to identify the variables that co-move with the target variable. Another way to do this is through penalisation methods (see Section 3.1) like LASSO, ridge regressions and elastic nets as these methods minimise the effects of less relevant predictors. The forecaster can even use LASSO strictly as a feature

selection method, by taking the predictors selected within LASSO and using them in a different predictive model.

Ideally, the forecaster should balance 'automatic' feature selection techniques such as penalisation methods with some theoretical scrutiny, i.e. the selected predictors' co-movement with the target variable should make some theoretical or intuitive sense. Thus, an unintuitive predictor revealed through LASSO need not be rejected outright, but should be judged on the basis of its possible theoretical link and predictive strength with respect to the target variable. Statistical techniques such as 'stability selection' applied to penalisation methods, can also help make this process easier by guarding against the inclusion of spurious predictors[16]. Variants of the LASSO, such as hierarchical LASSO and group LASSO can also be utilised for this purpose after careful consideration of the modelling strategy they bring to the table in the context of macroeconomic forecasting[17].

The next step involves further understanding the data generating process (DGP) of the target variable in order to get a sense of the way it moves. In particular, tests for checking whether the target variable is mean-reverting or has a long memory are important. If a series is mean-reverting, it has a greater tendency to oscillate: lower observations follow higher observations (and vice-versa). On the other hand, a series with a long memory has observations generally moving in the same direction as previous observations: higher (lower) observations follow higher (lower) observations. The Hurst Exponent can be used to determine this property (Sanyal and Roy, 2014).

Mean-reverting series are non-linear, whereas those with a long memory are relatively linear. Knowing which category the target variable belongs to, should direct the forecaster to a particular class of models. ML techniques in general are well-suited for predicting non-linear series since they do not make assumptions about the functional form of the 'fit' (see Introduction). However, non-ML techniques that allow for changes in parameters (such as time-varying parameter regressions) are also suitable for tracking non-linear series.

While an examination of the DGP should guide the forecaster towards a class of models that is expected to perform well, it should not restrict their attention solely to that class. As our meta-analysis shows, forecast performance is conditional on factors such as country-type, and

[16] See Meinshausen & Bühlmann (2010) for further reading on stability selection
[17] See Bien et al. (2013) and Yuan & Lin (2006) for further reading on hierarchical and group LASSO respectively

forecast horizon. Moreover, the literature on macroeconomic forecasting that uses ML methods is still nascent. At this stage, therefore, it is difficult to unambiguously know the exact class of models that will provide the best results, given the circumstances in which the user finds themselves. Thus, a good approach would be to formulate a hypothesis about the class of models that one would expect to perform well, given the nature of the DGP. The forecaster should then test multiple types of classes (ML, non-ML, linear, non-linear) so as to not exclude a class that may end up performing well, and see whether the hypothesis is accepted or rejected.

Finally, it is worth noting that regular k-fold cross-validation may not be appropriate for optimising the ML techniques used in macroeconomic forecasting. Cross-validation is important for finding the optimum value of the hyperparameters of ML techniques in order to find the right balance between bias and variance (see Section 3). The popularly used k-fold cross validation assumes that observations are independent and identically distributed, which allows observations to be interchangeably assigned to different folds within the series. Since macroeconomic time-series usually have at least some degree of correlation among the observations, k-fold cross validation can sever that correlation. However, variations such as rolling window cross validation or block cross validation exist, which preserve this correlation and are more appropriate for optimising ML techniques in the context of time series[18].

The choices available to a forecaster with respect to technique selection also depend on the nature of the data available. To illustrate this point, we use the classification by Doornik and Hendry (2015), who divide big data into three types: 'Fat Data', 'Tall data' and 'Huge Data'.

5.1 Fat Data

Fat Data are characterised by a large number of predictors (n), indicating a large cross-sectional dimension, but a limited temporal dimension (t). In other words, if $n > t$, data is said to be 'Fat'. Examples of this include large cross-sectional data, such as census data or single-period price collection data.

The forecaster has to first undertake the pre-checks discussed above, including stationarity and structural break tests. It must be noted, however, that these are only possible if the temporal dimension in a Fat dataset is sufficiently large. Next, since Fat Data has a very large number of potential predictors, selection of an optimal number of predictors is an important step in this

[18] See Schnaubelt (2019) for further reading

case. Techniques like LASSO are very useful here for gauging the most relevant predictors, since it excludes those that are not. In general, the penalization-based methods are useful here as they minimise the effects of less relevant predictors. In some sense they combine the 'feature selection' and 'predictive modelling' steps into one, which is particularly useful when n is very large. Panel estimation and factor-based methods such as Dynamic Factor Models, Factor Augmented VARs or methods that build on Principal Component Analysis can also be used in their 'sparse' versions (such as a sparse Factor-Augmented Vector Autoregressive (FAVAR) models), so that they account for the small temporal dimension in a Fat dataset. Bayesian methods are also preferred over classical estimation methods as they handle a fixed *t* and large cross-section with appropriate priors (Buono et. al., 2017).

Fat datasets are usually not particularly useful for macroeconomic predictions due to their limited temporal content, unless either *t* is large enough or variables are homogeneous enough over time (i.e., they are, or can be aggregated to, the same data frequency) to allow the use of panel techniques. However, both non-ML and ML techniques can be used for making forecasts with fat data, given that careful processing and pre-checks are conducted before choosing a predictive model.

5.2 Tall Data

Tall Data has a limited number of predictors and a big temporal content ($t > n$). Data such as daily cash withdrawals from ATM machines or second-by-second keyword internet search volumes fall in this category. Reconciling high and low frequency data in this instance is important, since predicting macroeconomic variables like growth or inflation may require aggregating high-frequency Tall data to a lower frequency (quarterly or annual) that matches that of the target macroeconomic series in question.

Working with Tall data should start with pre-checks similar to those mentioned above regarding stationarity and structural breaks, using the appropriate statistical tests. Unlike Fat data, since Tall data does not have a big *n* dimension, feature selection is relatively easier than in Fat data. It is therefore not necessary to use penalisation techniques to select features or factor-based methods to reduce many features into fewer factors.

Since Tall data are characterised by their big temporal content, tests to further understand the DGP of the series beyond stationarity or structural breaks, such as its mean-reverting properties, remain important to account for time variation in the data. Based on this information

and the necessary data-transformations, one can formulate a hypothesis on the class of model–– linear, non-linear, non-ML or ML—that will best fit the data. Classical time series econometric models can be used with tall data once the appropriate data transformations have been conducted and mean-reverting properties of the data have been examined.

5.3 Huge Data

Huge datasets are characterised by having both a large number of predictors and a large temporal dimension (big *n* and big *t*). All activity records of mobile subscribers or all Point-of-Sale (POS) transactions in a country in a time period are considered relevant types of Huge data. Since such data is rich in terms of its temporal content and the potential predictors that can be used, it is most desirable from a macroeconomic forecasting perspective. However, only a few information aggregators like Google (eg. Google Trends) have successfully managed to make such data publicly available.

The first step to working with Huge data for macroeconomic forecasting again involves conducting the necessary pre-checks for stationarity and structural breaks in the data, and conducting appropriate transformations.

Feature selection becomes important with Huge data, much like Fat data, since it can be computationally taxing to find the most relevant predictors from a large set of n. Penalisation techniques, therefore, have a natural advantage in this scenario. More complex ML techniques, such as neural networks or tree-based methods, which automate the process of finding relationships between the predictors and target series can also be used with Huge data. As is the case for Fat data, it is useful to balance 'automatic' feature-selection with theoretical and intuitive scrutiny. Using strict priors on a general regression could also lead to well-performing Bayesian estimates instead of selecting or summarising indicators (Buono et al. 2017).

Once the subset of relevant predictors has been identified after the necessary pre-checks on the data, the steps remain the same: further understand the DGP of the reference series, formulate a hypothesis on the class of models that are expected to provide the best forecasts given the DGP, and test different kinds of models (linear, non-linear, ML and non-ML).

6. Conclusion

Are big data and machine learning based forecasts better than their traditionally used, standard counterparts? With respect to data, we find that there are both advantages and limitations to using big data. Because of its organic nature, it provides a variety of high-frequency data that can help with issues of jagged-panels and measurement biases that accompany standard data. However, it also suffers from sampling biases stemming from the 'digital divide' that remains even in an increasingly connected and electronic world. Big data, therefore, should be considered a complement and not a substitute of its standard counterparts, and 'big data hubris' needs to be avoided.

With respect to forecasting techniques, the results are less straightforward. Our meta-analysis suggests that ML techniques do not necessarily dominate their non-ML counterparts in every condition, but are definitely superior under certain conditions: for longer forecast horizons, advanced economies and for growth forecasting. Non-ML Factor-based methods are especially competitive against ML methods and we see no significant results when the former is compared against different classes of ML techniques. On the other hand, all ML techniques did significantly better than Univariate methods. This is important because Univariate methods are still commonly used as benchmarks in forecasts of growth and inflation, effectively setting an artificially low standard for comparing the performance of other techniques. Future studies using ML techniques for forecasting growth and inflation should thus use more competitive benchmarks in order to evaluate the performance of ML techniques more accurately. Our dataset also suggests that for growth, ML techniques are used far more in advanced economies than in emerging countries, and this may be driving some of the insignificant results in case of developing countries. For inflation, while ML and non-ML techniques are used equally in emerging economies, non-ML techniques are used more in advanced economies. Therefore, there is significant scope for growth forecasters in emerging economies and inflation forecasters in advanced economies to experiment more with ML techniques in the future.

Macroeconomic forecasting is not an exact science and there are parts of the process where choices have to be made based on the data available and the model-builders intuition. Hopefully, the meta-analysis of the predictive performance of these state-of-the-art forecasting techniques, and the user's guide provided in this paper, will help macroeconomic forecasters have a better understanding about the choices available to them.

Bibliography

1. Baybuza, I. 'Inflation Forecasting Using Machine Learning Methods'. December 2018. Russian Journal of Money and Finance, 77(4), pp. 42–59. Available at: https://rjmf.econs.online/upload/iblock/ca3/RJMF_77-04_ENG_Baybuza.pdf
2. Biau, O., & D'Elia, A. Euro area GDP forecasting using large survey datasets: a random forest approach. 2009. Directorate General for Financial and Economic Affairs - European Commission. Available at: https://ec.europa.eu/economy_finance/db_indicators/surveys/documents/workshops/2010/ec_meeting/biau_delia_gdp_forecasting.pdf
3. Bien, J., Taylor, J., & Tibshirani, R. 2013. 'A lasso for hierarchical interactions'. Annals of statistics, 41(3), 1111. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4527358/
4. Buono, D., Mazzi, G., Kapetanios, G., Marcellino, M., Papailias, F. 'Big Data Types for Macroeconomic Nowcasting'. June 2017. Eurostat Review On National Accounts And Macroeconomic Indicators. Available at: https://www.researchgate.net/publication/318340545_Big_data_types_for_macroeconomic_nowcasting
5. Chakraborty, C. and Joseph, A. 'Staff Working Paper No. 674: Machine Learning at Central Banks'. September 2017. Bank of England. Available at: https://www.bankofengland.co.uk/working-paper/2017/machine-learning-at-central-banks
6. Chuku, C., Oduor, J., & Simpasa, A. Intelligent forecasting of economic growth for African economies: Artificial neural networks versus time series and structural econometric models. 2017. Washington Research Program on Forecasting, 1-28. Available at: http://unassumingeconomist.com/wp-content/uploads/2017/04/Intelligent-forecast-using-ANN.pdf
7. Coulombe, P. G., Leroux, M., Stevanovic, D., & Surprenant, S. How is machine learning useful for macroeconomic forecasting? 2020. arXiv preprint arXiv:2008.12477. Available at: https://arxiv.org/pdf/2008.12477.pdf
8. Doornik, J. A. and D.F. Hendry (2015), 'Statistical Model Selection with Big Data', Cogent Economics & Finance, 3(1), 2015. Available at: https://www.economics.ox.ac.uk/materials/papers/13640/paper735.pdf
9. Duncan, R. and Garcia, E.M. 'Working Paper 338: New Perspectives on Forecasting Inflation in Emerging Market Economies: An Empirical Assessment'. January 2018. Federal Reserve Bank of Dallas Globalization and Monetary Policy Institute Working Paper Series. Available at: https://www.dallasfed.org/~/media/documents/institute/wpapers/2018/0338.pdf
10. Freund, J. E. Mathematical Statistics. 2000. Prentice-Hall.
11. IBM. 'The Four V's of Big Data'. Available at: https://www.ibmbigdatahub.com/infographic/four-vs-big-data#:~:text=IBM%20data%20scientists%20break%20big,%2C%20variety%2C%20velocity%20and%20veracity.

12. Jahn, M. Artificial neural network regression models: Predicting GDP growth. 2018. HWWI Research Paper (No. 185). Available at: https://www.econstor.eu/bitstream/10419/182108/1/1030440859.pdf
13. James, G., Witten, D., Hastie, T. and Tibshirani, R. 2013. 'An Introduction to Statistical Learning with Applications in R'. Springer Texts in Statistics
14. Jung, J. K., Patnam, M., & Ter-Martirosyan, A. An algorithmic crystal ball: Forecasts-based on machine learning. 2018. International Monetary Fund. Available at: https://www.elibrary.imf.org/view/journals/001/2018/230/article-A001-en.xml
15. Kapetanios, G., & Papailias, F. 2018. Big data & macroeconomic nowcasting: Methodological review. Economic Statistics Centre of Excellence (ESCoE) Discussion Papers ESCoE DP-2018-12, Economic Statistics Centre of Excellence (ESCoE). Available at: http://escoe-website.s3.amazonaws.com/wp-content/uploads/2020/07/13161005/ESCoE-DP-2018-12.pdf
16. Lazer, D., Kennedy, R., King, G., & Vespignani, A. 2014. The parable of Google Flu: traps in big data analysis. Science, 343(6176), 1203-1205. Available at: https://dash.harvard.edu/bitstream/handle/1/12016836/The%20Parable%20of%20Google%20Flu%20%28WP-Final%29.pdf?sequence=1
17. Leal, F., Molina, C., Zilberman, E. 'No. 860: Inflation Forecast in Chile with Machine Learning Methods'. January 2020. Working Papers of the Central Bank of Chile. Available at: https://www.bcentral.cl/documents/33528/133326/dtbc860.pdf/3509186b-a98b-0885-1153-8fd2847dacc2?t=1606836344547
18. Mahajan, K. and Srinivasan, A. 'Inflation Forecasting in Emerging Markets: A Machine Learning Approach'. February 2020. Available at: https://www.cafral.org.in/Research-Publication.aspx
19. Marcellino, M. 2007. A comparison of time series models for forecasting GDP growth and inflation. Bocconi University, Italia. Available at: https://www.researchgate.net/profile/Niels_Haldrup/publication/228650389_A_comparison_of_time_series_models_for_forecasting_GDP_growth_and_inflation/links/0c96051b6b0d0e7951000000/A-comparison-of-time-series-models-for-forecasting-GDP-growth-and-inflation.pdf
20. Medeiros, M., Vasconcelos, G., Veiga, A., Zilberman, E. July 2019. 'Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods'. Journal of Business & Economic Statistics. 39. 1-45. Available at: https://www.researchgate.net/publication/334155178_Forecasting_Inflation_in_a_Data-Rich_Environment_The_Benefits_of_Machine_Learning_Methods
21. Meinshausen, N., & Bühlmann, P. 2010. 'Stability selection'. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72(4), 417-473. Available at: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2010.00740.x
22. Mitchell, T. M. Machine Learning. 1997. McGraw-Hill
23. Nakamura, E. 'Inflation forecasting using a neural network'. 2005. Economics Letters. Volume 86, Issue 3.Pages 373-378. Available at: https://www.sciencedirect.com/science/article/pii/S0165176504003088

24. Reserve Bank of India. October 2020. 'Monetary Policy Report'. Available at: https://www.rbi.org.in/Scripts/PublicationsView.aspx?id=19980#I2
25. Richardson, A., van Florenstein Mulder, T., & Vehbi, T. 2019. Nowcasting GDP using machine-learning algorithms: A real-time assessment. Discussion Paper 2019-03, Reserve Bank of New Zealand. Available at: https://www.rbnz.govt.nz/-/media/ReserveBank/Files/Publications/Discussion%20papers/2019/dp2019-03.pdf?revision=6505b3a5-634e-45bd-9938-a7af3b4317fb
26. Rigobon, R. 2018. 'Big Data and Measurement: From Inflation to Discrimination'. Second Suresh Tendulkar Memorial lecture delivered at the Reserve Bank of India. Available at: https://m.rbi.org.in/scripts/bs_viewcontent.aspx?Id=3526
27. Roy, I., Sanyal, A., & Ghosh, A. K. 2016. Nowcasting Indian GVA Growth in a Mixed Frequency Setup. Reserve Bank of India Occasional Papers, Vol. 37, No. 1&2. Available at: https://rbidocs.rbi.org.in/rdocs/Publications/PDFs/RBIOP37020520185C7D2C48AFA9427993B773BD194D95A2.PDF#page=93
28. Rubin, D.B. 1977. The design of a general and flexible system for handling nonresponse in sample surveys. Report prepared for the U.S. Social Security Administration. Available at: www.jstor.org/stable/27643587
29. Ryll, L., & Seidens, S. Evaluating the performance of machine learning algorithms in financial market forecasting: A comprehensive survey. 2019. arXiv preprint arXiv:1906.07786. Available at: https://arxiv.org/pdf/1906.07786.pdf
30. Sanyal, A., & Roy, I. 2014. Forecasting major macroeconomic variables in India: Performance comparison of linear, non-linear models and forecast combinations. RBI Working Paper Series No. 11, Tech. Rep. Available at: https://www.rbi.org.in/scripts/PublicationsView.aspx?id=16144
31. Schnaubelt, M. 2019. 'A comparison of machine learning model validation schemes for non-stationary time series data'. (No. 11/2019), FAU Discussion Papers in Economics. Available at: https://www.econstor.eu/bitstream/10419/209136/1/1684440068.pdf
32. Taylor, J.B. Discretion versus Policy Rules in Practice. 1993. Carnegie-Rochester Conference Series on Public Policy 39, December, 195-214. Available at: http://opendata.dspace.ceu.es/bitstream/10637/2345/1/p%20195_214.pdf
33. Taylor, J.B. (Ed). Monetary Policy Rules. 1999. Chicago: University of Chicago. Available at: https://www.nber.org/system/files/chapters/c7419/c7419.pdf
34. Teräsvirta, T., Van Dijk, D., & Medeiros, M. C. 2005. Linear models, smooth transition autoregressions, and neural networks for forecasting macroeconomic time series: A re-examination. International Journal of Forecasting, 21(4), 755-774. Available at: https://www.econstor.eu/bitstream/10419/175972/1/td485.pdf
35. Tiffin, A. 2016. Seeing in the dark: a machine-learning approach to nowcasting in Lebanon. IMF Working Paper WP/16/56. Available at: https://www.researchgate.net/profile/Andrew_Tiffin/publication/301741469_Seeing_in_the_Dark_A_Machine-Learning_Approach_to_Nowcasting_in_Lebanon/links/573201b608ae9ace84047701.pdf

36. Woloszko, N. 2020. "A Weekly Tracker of activity based on machine learning and Google Trends", OECD Economics Department Working Papers 1634, Paris: OECD Publishing. Available at: https://www.oecd-ilibrary.org/economics/tracking-activity-in-real-time-with-google-trends_6b9c7518-en
37. Yuan, M., & Lin, Y. 2006. 'Model selection and estimation in regression with grouped variables'. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(1), 49-67. Available at: http://control.dei.unipd.it/public/Schenato/PSC/2010_2011/gruppo4-Building_termo_identification/Bibliografia%20Suggerita/2_%5B49%5D_glar_tr1095.pdf

Appendix

A. RMSEs and adjusted RMSEs

Table A1: 1Q-ahead RMSEs (Growth)

	Sanyal & Roy	Roy	Richardson	Jung et al							Biau & D'elia	Chuku et al.		
	India	India	New zealand	U.S	U.K	Germany	Spain	Mexico	Philippines	Vietnam	Euro area	South Africa	Nigeria	Kenya
No change														
AR(1)			0.497											
AR(p)											0.8			
Factor-augmented AR														
Random Walk														
OLS regression												0.963	1.903	1.089
MIDAS														
ARTV														
ARTV-VAR		0.8												
LSTAR	1.5	1.2												
SETAR	1.25	1.2												
Holt-Winters	1.5	1.7												
Bridge-equation											0.374			
Bridge (w/ factors)														
AAR	1.5	1.6												
BVAR														
PCA/PCF														
ARIMA	1.1	1.6										0.985	4.447	1.359
LASSO			0.425											
Ridge			0.483											
Elastic Net			0.426	0.46	0.43	0.73	0.27	0.87	1.09	1.08				
Random Forest											0.655			
Boosting			0.384											
Neural Network	0.8	1.5	0.397									0.851	0.967	1.825
RNN				0.77	0.71	1.55	3.21	1.46	1.83	1.15				
SVM			0.393											
Dynamic Factor Models		0.25	0.46											
Combination Forecasts	1.2		0.385	0.33	0.25	0.37	0.17	0.62	0.54	0.55				
Average	1.264285714	1.23125	0.427777778	0.52	0.463333333	0.883333333	1.216666667	0.983333333	1.153333333	0.926666667	0.609666667	0.933	2.439	1.424333333

Table A2: 1Q-ahead RMSEs (Inflation)

	Baybuza et al.	Leal et al.		Mahajan & Srinivasan	Terasvirta et al.							Marcellino	Nakamura	Garcia et al.	Medeiros et al.
	Russia	Chile (CPI)	Chile (CPIEFE)	India	Canada	France	Germany	Italy	Japan	UK	USA	US	US	Brazil	US
AR(1)	1				0.004	0.003	0.004	0.004	0.006	0.005	0.004				
CNN+LSTM				0.689											
Median		0.78	0.74												0.72
Mean		0.78	0.74												0.73
adaENet		0.79	0.74												0.73
adaLASSO		0.78	0.75											0.00243	0.74
Bagging															0.76
CSR															0.76
Complete Subset Regressions		0.81	0.74											0.00223	
CNN				0.8											
PCA/PCF		0.86	0.77												
BVAR		0.77	0.74												0.75
AR(p)	0.9878	0.8	0.73									0.82		0.00221	0.79
BMA															0.84
Logistic Smooth Transition												0.84			
UCSV		0.85	0.91												0.8
Encoder Decoder				0.854											
Random Walk	1.1227	1	1											0.0041	1
PC															
VAR(1)															
LASSO	1.4668	0.85	0.79											0.00321	0.73
ARMA															
STAR					0.004	0.003102	0.004	0.003676	0.005874	0.00506	0.004				
Ridge	1.7934	0.8	0.76												0.72
Neural Network AR												1			
MA				1.329											
Neural Network		0.94	0.98		0.004576	0.00297	0.004104	0.003847	0.00711	0.006285	0.00368		1.715		
Random Forest	0.8402	0.79	0.75	2.211										0.00254	0.71
AR(1)/LASSO	1.0976														
Boosting	0.7574			1.585											
Elastic Net	1.5674	0.79	0.74	2.419											0.73
SARIMA				2.388											
Top 5														0.00248	
FOCUS survey														0.00248	
Factor Model														0.00248	

Table A3: 4Q-ahead RMSEs (Growth)

	Sanyal & Roy	Roy	Jorda														Jung et al.						
	India	India	Austria	Belgium	Denmark	Finland	France	Germany	Greece	Italy	Japan	Netherlands	Portugal	Spain	Sweden	UK	United States	U.S	U.K	Germany	Spain	Mexico	Philippines	Vietnam
No change																								
AR(1)																								
AR(p)																								
Factor-augmented AR																								
Random Walk																								
OLS regression			2.64	1.34	1.73	2.56	1.34	1.7	1.92	1.43	1.05	2.6	2.3	2.14	0.84	1.11	1.35							
MIDAS																								
ARTV																								
ARTV-VAR		1.7																						
LSTAR	1.8	1.8																						
SETAR	1.76	2.1																						
Holt-Winters	1.8	2.4																						
Bridge equation																								
Bridge (w/ factor)																								
AAR	1.9	3.2																						
BVAR																								
PCA/PCF																								
ARIMA	1.7	3																						
LASSO																								
Ridge																								
Elastic Net																		1.49	1.55	1.26	2.51	1.28	2.34	0.49
Random Forest																								
Boosting																								
Neural Network	2	2.8	0.46	0.36	0.34	0.17	0.66	0.1	0.71	0.81	0.47	0.75	0.2	1.13	0.38	0.25	0.04							
RNN																		0.65	0.55	1.26	3.73	1.36	1.96	1.17
SVM																								
Dynamic Factor Models		1.25																						
Combination Forecasts	1.9																	1.91	1.65	4.49	2.03	1.8	3.31	0.91
Average	1.835714286	2.10025	1.55	0.85	1.039	1.393	1	0.9	1.353	1.12	0.79	1.605	1.25	1.439	0.76	0.67	0.695	1.35	1.25	2.346666667	2.423333333	1.54	2.55	0.856666667

Table A4: 4Q-ahead RMSEs (Inflation)

	Baybuza et al.	Leal et al.		Mahajan & Srinivasan			Terasvirta et al.							Medeiros et al.	Marcellino	Nakamura	Garcia et al.
	Russia	Chile (CPI)	Chile (CPIEFE)	India	China	South Africa	Canada	France	Germany	Italy	Japan	UK	USA	USA	US	US	Brazil
AR(1)	1						0.004	0.004	0.005	0.004	0.007	0.005	0.004				
CNN-LSTM				0.846	0.407	0.59											
VAR(1)																	
Logistic Smooth Transition															0.67		
PC																	
CNN				0.888	0.527	0.635											
BVAR		0.78	0.760003333											0.74			
Median		0.77	0.740003333											0.71			
Bagging														0.74			
Mean		0.78	0.740003333											0.71			
AR(p)	1.0154	0.74	0.740003333											0.75	0.57		0.00284
adaENet		0.79	0.740003333											0.73			
CSR														0.76			
adaLASSO		0.8	0.790003333											0.72			0.00246
Combination Forecasts							0.004088	0.003992	0.004825	0.00386	0.007021	0.00503	0.003768				
ARIMA																	
JMA														0.8			
Encoder Decoder				1.088	0.526	0.809											
Complete Subset Regressions		0.86	0.800003333														0.00226
UCSV		0.86												0.78			
PCA/PCF		0.9	0.820003333														
MA				1.46	0.54	1.418											
STAR							0.004	0.004244	0.005	0.003944	0.007126	0.00513	0.004				
Neural Network AR															1.05		
Ridge	2.1443	0.87	0.820003333											0.7			
Random Forest	0.4937	0.82	0.740003333	2.995		0.9338								0.68			0.00246
Neural Network		1.07	0.940003333				0.004276	0.00416	0.00509375	0.003878	0.010857	0.007695	0.003628			2.1625	
LASSO	1.8262	1.29	1.060003333											0.73			0.00333
AR(1)LASSO	1.2635																
Boosting	0.5149			2.203													
Elastic Net	1.8262	0.93	0.730003333	2.61										0.73			
SARIMA				2.388	1.012	1.217	0.004091	0.004099	0.00509375	0.003878	0.008001	0.00571375	0.003849				
Random Walk	1.7878	1	1.000003333											1			0.00275
Top 5																	0.00255
FOCUS																	0.0026
Factor Model																	0.00244

Table A5: 1Q-ahead adjusted RMSEs (Growth)

Adjusted RMSE	Sanyal & Roy	Roy	Richardson	Jung et al							Biau & D'elia	Chuku et al.		
	India	India	New zealand	U.S	U.K	Germany	Spain	Mexico	Philippine	Vietnam	Euro area	South Afri	Nigeria	Kenya
No change														
AR(1)			0.069222222											
AR(p)											0.190333333			
Factor-augmented AR														
Random Walk														
OLS regression												0.03	-0.536	-0.33533
MIDAS														
ARTV														
ARTV-VAR		-0.44667												
LSTAR	0.25333	-0.04667												
SETAR	0.00333	-0.04667												
Holt-Winters	0.25333	0.45333												
Bridge-equation											-0.235666667			
Bridge (w/ factors)														
AAR	0.253330	0.35333												
BVAR														
PCA/PCF														
ARIMA	-0.14667	0.35333										0.052	2.008	-0.06533
LASSO			-0.002777778											
Ridge			0.055222222											
Elastic Net			-0.001777778	-0.06	-0.03333	-0.15333	-0.94667	-0.11333	-0.06333	0.153333				
Random Forest											0.045333333			
Boosting			-0.043777778											
Neural Network	-0.44667	0.25333	-0.030777778									-0.082	-1.472	0.400667
RNN				0.25	0.246667	0.666667	1.993333	0.476667	0.676667	0.223333				
SVM			-0.034777778											
Dynamic Factor Models		-0.99667	0.032222222											
Combination Forecasts	-0.04667		-0.042777778	-0.19	-0.21333	-0.51333	-1.04667	-0.36333	-0.61333	-0.37667				

Table A6: 1Q-ahead adjusted RMSEs (Inflation)

Adjusted RMSE	Baybuza et al.	Leal et al.		Mahajan &	Terasvirta et al.							Marcellino	Nakamura	Garcia et al.	Medeiros et al.
	Russia	Chile (CPI)	Chile (CPIEFE)	India	Canada	France	Germany	Italy	Japan	UK	USA	US	US	Brazil	US
AR(1)	-0.181478				-0.000192	-2.4E-05	-3.5E-05	0.000159	-0.000328	-0.000448	-0.839223				
CNN+LSTM				-0.84538											
Median		-0.029	-0.05200												-0.123223
Mean		-0.029	-0.05200												-0.113223
adaElNet		-0.019	-0.05200												-0.113223
adaLASSO		-0.029	-0.04200											-0.000254	-0.103223
Bagging															-0.083223
CSR															-0.083223
Complete Subset Regressions		0.001	-0.052											-0.000454	
CNN				-0.73438											
PCA/PCF		0.051	-0.022												
BVAR		-0.039	-0.052												-0.093223
AR(p)	-0.193678	-0.009	-0.062									-0.023223		-0.000474	-0.053223
JMA															-0.003223
Logistic Smooth TRansition												-0.003223			
UCSV		0.041	0.118												-0.043223
Encoder Decoder				-0.68038											
Random Walk	-0.058778	0.191	0.208											0.001416	0.156777
PC															
VAR(1)															
LASSO	0.285322	0.041	-0.002											0.000526	-0.113223
ARMA															
STAR					-0.000192	7.8E-05	-3.5E-05	-0.000165	-0.000454	-0.000388	-0.839223				
Ridge	0.611922	-0.009	-0.032												-0.123223
Neural Network AR												0.156777			
MA				-0.20538											
Neural Network		0.131	0.188		0.000384	-5.4E-05	6.9E-05	6E-06	0.000782	0.000837	-0.839543		0.871777		
Random Forest	-0.341278	-0.019	-0.042	0.676625										-0.000144	-0.133223
AR(1)/LASSO	-0.083878														
Boosting	-0.424078			0.050625											
Elastic Net	0.385922	-0.019	-0.052	0.884625											-0.113223
SARIMA				0.853625											
Top 5														-0.000204	
FOCUS survey														-0.000204	
Factor Model														-0.000204	

Table A7: 4Q-ahead adjusted RMSEs (Growth)

Adjusted RMSE	Senyel & Roy	Roy	Jahn														Jung et al.						
	India	India	Austria	Belgium	Denmark	Finland	France	Germany	Greece	Italy	Japan	Netherlan	Portugal	Spain	Sweden	UK	USA	USA	U.K	Germany	Spain	Mexico	Philippines	Vietnam
No change																							
AR(1)																							
AR(p)																							
Factor-augmented AR																							
Random Walk																							
OLS regression			1.09	0.49	0.695	1.195	0.34	-0.068	1.315	0.31	0.29	0.905	1.05	0.032	0.18	-1.108	0.442							
MIDAS																							
ARTV																							
ARTV-VAR		-0.28																					
LSTAR	-0.180	-0.58																					
SETAR	-0.23	0.12																					
Holt-Winters	-0.18	0.42																					
Bridge-equation																							
Bridge (w/ factors)																							
AAR	-0.08	1.22																					
BVAR																							
PCA/PCF																							
ARIMA	-0.28	0.02																					
LASSO																							
Ridge																							
Elastic Net																		0.582	-0.668	-0.508	0.402	-0.86	-0.21	-0.30667
Random Forest																							
Boosting																							
Neural Network	0.02	0.82	-1.09	-0.49	-0.695	-1.195	-0.34	-1.668	1.315	-0.31	-0.29	-0.905	-1.05	-0.978	-0.18	-1.988	-0.868							
RNN																		-0.258	-1.668	-0.478	0.622	1.2	-0.59	0.313333
SVM																							
Dynamic Factor Models		-0.73																					
Combination Forecasts	-0.08																	0.102	5.432	2.722	-0.078	-0.34	0.8	0.053333

Table A8: 4Q-ahead adjusted RMSEs (Inflation)

Adjusted RMSE	Baybuza et al.	Leal et al.		Mahajan & Srinivasan			Terasvirta et al.							Medeiros et al.	Marcellino	Nakamura	Garcia et al.
	Russia	Chile (CPI)	Chile (CPIEFE)	India	China	South Africa	Canada	France	Germany	Italy	Japan	UK	USA	USA	USA	USA	Brazil
AR(1)	-0.31911111						-9E-05	-0.0001	-2.5E-06	9E-05	-0.001	-0.00071	-0.91642				
CNN+LSTM				-0.96375	-0.1954	-0.3438											
VAR(1)																	
Logistic Smooth TRansition															-0.25042		
PC																	
CNN				-0.92175	-0.0754	-0.2988											
BVAR		-0.0695	-0.099496667											-0.18042			
Median		-0.0795	-0.109496667											-0.21042			
Bagging														-0.18042			
Mean		-0.0695	-0.109496667											-0.21042			
AR(p)	-0.30371111	-0.1095	-0.109496667											-0.17042	-0.35042		3E-05
adaElNet		-0.0595	-0.109496667											-0.19042			
CSR														-0.16042			
adaLASSO		-0.0495	-0.059496667											-0.20042			-0.00015
Combination Forecasts							-2E-06	-0.000108	-0.0001775	-5E-05	-0.00098	-0.00068	-0.91665				
ARMA																	
JMA														-0.12042			
Encoder Decoder				-0.72175	-0.0764	-0.1248											
Complete Subset Regressions		0.0105	-0.049496667														-0.00035
UCSV		0.0105												-0.14042			
PCA/PCF		0.0505	-0.029496667														
MA				-0.34975	-0.0624	0.4842											
STAR							-9E-05	0.000144	-2.5E-06	3.4E-05	-0.00087	-0.00058	-0.91642				
Neural Network AR															0.12958		
Ridge	0.82518889	0.0205	-0.029496667											-0.22042			
Random Forest	-0.82541111	-0.0295	-0.109496667	1.18525										-0.24042			-0.00015
Neural Network		0.2205	0.090503333				0.000186	6E-05	9.125E-05	-3.2E-05	0.002857	0.001985	-0.91679			1.24208	
LASSO	0.50708889	0.4405	0.210503333											-0.19042			0.00072
AR(1)/LASSO	-0.05561111																
Boosting	-0.80421111			0.39325													
Elastic Net	0.50708889	0.0805	-0.119496667	0.80025										-0.19042			
SARIMA				0.57825	0.4096	0.2832	1E-06	-1E-06	9.125E-05	-3.2E-05	1E-06	3.75E-06	-0.91657				
Random Walk	0.46868889	0.1505	0.150503333											0.07958			0.00014
Top 5																	-6E-05
FOCUS																	-0.00001
Factor Model																	-0.00017

B. Pair-wise results

Table B1: Significance tests (Penalisation methods vs Univariate methods)

	# of observations		Mann Whitney U test	Students T-test			
Penalisation methods by Univariate methods	Non-ML	ML	p-value	p-order (non-ML > ML)	p value (H1≠0)	p value (H1<0)	p value (H1>0)
Growth	23	18	0.0554**	0.676	0.0314**	0.9843	0.0157***
Inflation	78	44	0.594	0.471	.0179**	0.0090**	0.9910
EM	53	33	0.430	0.551	.2276	0.8862	0.1138
AE	48	29	0.9064	0.508	.1269	0.0634*	0.9366
1Q	52	32	0.3553	0.561	.8380	0.5810	0.4190
4Q	49	30	0.9799	0.498	.6201	0.3101	0.6899

*** = 1%, **=5%, *=10%; H0 (null hypothesis) is that group means are equal [mean (non-ML) – mean (ML) = 0]; Green cells represent ML outperforming non-ML. Red cells represent the converse.

Table B2: Significance tests (Penalisation methods vs Multivariate methods)

	# of observations		Mann Whitney U test	Students T-test			
Penalisation methods by Multivariate methods	Non-ML	ML	p-value	p-order (non-ML > ML)	p value (H1≠0)	p value (H1<0)	p value (H1>0)
Growth	25	18	0.1130	0.644	.0616**	0.9692	0.0308**
Inflation	18	44	0.1303	0.376	.003***	0.0015***	0.9985
EM	19	33	0.0795*	0.352	.032**	0.0161**	0.9839
AE	24	29	0.1174	0.626	.0809*	0.9595	0.0405**
1Q	16	32	0.2288	0.393	.0759*	0.0380**	0.9620

4Q	27	30	0.2341	0.593	0.1657	0.9171	0.0829*

*** = 1%, **=5%, *=10%; H0 (null hypothesis) is that group means are equal [mean (non-ML) – mean (ML) = 0]; Green cells represent ML outperforming non-ML. Red cells represent the converse.

Table B3: Significance tests (Penalisation methods vs Factor methods)

	# of observations		Mann Whitney U test	Students T-test			
Penalisation methods by Factor methods	Non-ML	ML	p-value	p-order (non-ML > ML)	p value (H1≠0)	p value (H1<0)	p value (H1>0)
Growth	3	18	0.2617	0.278	.2798	0.1399	0.8601
Inflation	11	44	0.3734	0.589	.1343	0.0672*	0.9328
EM	11	33	0.5882	0.556	.1919	0.0960*	0.9040
AE	3	29	0.9472	0.625	0.5245	0.2622	0.7378
1Q	8	32	0.4591	0.588	.367	0.1836	0.8196
4Q	6	30	0.8919	0.481	.3175	0.1587	0.8413

*** = 1%, **=5%, *=10%; H0 (null hypothesis) is that group means are equal [mean (non-ML) – mean (ML) = 0]; Green cells represent ML outperforming non-ML. Red cells represent the converse.

Table B4: Significance tests (Tree-based methods vs Univariate methods)

	# of observations		Mann Whitney U test	Students T-test			
Tree methods by Univariate methods	Non-ML	ML	p-value	p-order (non-ML > ML	p value (H1≠0)	p value (H1<0)	p value (H1>0)
Growth	23	3	0.7546	0.435	.7017	0.3508	0.6492
Inflation	78	18	0.2614	0.586	.9760	0.4880	0.5120
EM	53	11	0.2759	0.394	.290	0.1450	0.8550
AE	48	10	0.0235**	0.727	.1095	0.9453	0.0547**
1Q	52	12	0.6012	0.550	.8508	0.4254	0.5746
4Q	49	9	0.3646	0.598	.877	0.5614	0.4386

*** = 1%, **=5%, *=10%; H0 (null hypothesis) is that group means are equal [mean (non-ML) – mean (ML) = 0]; Green cells represent ML outperforming non-ML. Red cells represent the converse.

Table B5: Significance tests (Tree-based methods vs Multivariate methods)

	# of observations		Mann Whitney U test	Students T-test			
Tree methods by Multivariate methods	Non-ML	ML	p-value	p-order (non-ML > ML)	p value (H1≠0)	p value (H1<0)	p value (H1>0)
Growth	25	3	0.6221	0.400	.6691	0.3345	0.6655
Inflation	18	18	0.8329	0.522	.8759	0.4375	0.5621
EM	19	11	0.0051**	0.196	.0178**	0.0089***	0.9911
AE	24	10	0.0026***	0.821	0.0009***	0.9995	0.0005***
1Q	16	12	0.3528	0.393	.2185	0.1093*	0.8907
4Q	27	9	0.2217	0.640	.2778	0.8611	0.1389

*** = 1%, **=5%, *=10%; H0 (null hypothesis) is that group means are equal [mean (non-ML) – mean (ML) = 0]; Green cells represent ML outperforming non-ML. Red cells represent the converse.

Table B6: Significance tests (Tree-based methods vs Factor methods)

	# of observations		Mann Whitney U test	Students T-test			
Tree methods by Factor methods	Non-ML	ML	p-value	p-order (non-ML > ML	p value (H1≠0)	p value (H1<0)	p value (H1>0)

Growth	3	3	0.2000	0.111	.1309	0.0654**	0.9346
Inflation	11	18	0.1338	0.672	.6315	0.6843	0.3157
EM	11	11	0.3653	0.380	.0301**	0.0151**	0.9849
AE	3	10	0.1608	0.800	.0394**	0.9803	0.0197**
1Q	8	12	0.6784	0.563	.3489	0.1744	0.8526
4Q	6	9	0.6070	0.693	.765	0.3827	0.6173

*** = 1%, **=5%, *=10%; H0 (null hypothesis) is that group means are equal [mean (non-ML) – mean (ML) = 0]; Green cells represent ML outperforming non-ML. Red cells represent the converse.

Table B7: Significance tests (Neural network methods vs Univariate methods)

	# of observations		Mann Whitney U test	Students T-test			
Neural Network methods by Univariate methods	Non-ML	ML	p-value	p-order (non-ML > ML	p value (H1≠0)	p value (H1<0)	p value (H1>0)
Growth	23	37	0.0355*	0.662	.0229*	0.9885	0.0115**
Inflation	78	34	0.5912	0.468	.3534	0.8233	0.1767
EM	53	29	0.0883*	0.615	.0506*	0.9747	0.0253**
AE	48	42	0.8390	0.487	.5147	0.7426	0.2574
1Q	52	27	0.1599	0.403	.9588	0.4794	0.5206
4Q	49	44	0.0221**	0.638	.02049**	0.9875	0.0125**

*** = 1%, **=5%, *=10%; H0 (null hypothesis) is that group means are equal [mean (non-ML) – mean (ML) = 0]; Green cells represent ML outperforming non-ML. Red cells represent the converse.

Table B8: Significance tests (Neural network methods vs Multivariate methods)

	# of observations		Mann Whitney U test	Students T-test			
Neural Network methods by Multivariate methods	Non-ML	ML	p-value	p-order (non-ML > ML	p value (H1≠0)	p value (H1<0)	p value (H1>0)
Growth	25	37	0.0492**	0.648	.0407**	0.9796	0.0204***
Inflation	18	34	0.3323	0.417	.4180	0.7910	0.2090
EM	19	29	0.9127	0.490	.9514	0.4757	0.5243
AE	24	42	0.0233***	0.668	.0085***	0.9958	0.0042***
1Q	16	27	0.0390**	0.310	.2590	0.1295*	0.8705
4Q	27	44	0.0022***	0.714	0.0017***	0.9991	0.0009***

*** = 1%, **=5%, *=10%; H0 (null hypothesis) is that group means are equal [mean (non-ML) – mean (ML) = 0]; Green cells represent ML outperforming non-ML. Red cells represent the converse.

Table B9: Significance tests (Neural network methods vs Factor methods)

	# of observations		Mann Whitney U test	Students T-test			
Neural Network methods by Factor methods	Non-ML	ML	p-value	p-order (non-ML > ML	p value (H1≠0)	p value (H1<0)	p value (H1>0)
Growth	3	37	0.4569	0.360	.3694	0.1847	0.8153
Inflation	11	34	0.5092	0.568	.1042	0.9475	0.0521
EM	11	29	0.9050	0.486	.9274	0.4637	0.5363
AE	3	42	0.7791	0.556	.2417	0.8792	0.1208
1Q	8	27	0.2862	0.370	.3998	0.1999	0.8001
4Q	6	44	0.3742	0.616	.3087	0.8456	0.1544

*** = 1%, **=5%, *=10%; H0 (null hypothesis) is that group means are equal [mean (non-ML) – mean (ML) = 0]; Green cells represent ML outperforming non-ML. Red cells represent the converse.

C. Descriptive characteristics of included literature

Table C1: Papers by country, time-period examined, forecast horizon and indicator forecasted

Author	Country	Time period	Forecast horizon	Indicator
Richardson et. al.(2019)	New Zealand	1995-2008	1-quarter	Growth
Jung et. al. (2018)	U.K, U.S., Germany, Spain, Mexico, Philippines, Vietnam	1987-2016	1-quarter, 4-quarter	Growth
Jahn (2018)	Austria, Belgium, Denmark, Finland, France, Germany, Greece, Italy, Netherlands, Portugal, Spain, Sweden, U.K., U.S., Japan	1996-2016	4-quarter	Growth
Biau & D'elia (2010)	Euro Area	1995-2009	1-quarter	Growth
Sanyal & Roy (2014)	India	1997-2011	1-quarter, 4-quarter	Growth
Roy et. al. (2016)	India	2006-2016	1-quarter, 4-quarter	Growth
Tiffin (2016)	Lebanon	1996-2013	1-quarter	Growth
Chuku et. al. (2017)	South Africa, Nigeria, Kenya		1-quarter	Growth
Baybuza et al (2018)	Russia	2008-2018	1-quarter, 4-quarter	Inflation
Nakamura (2005)	U.S.	1960-2003	1-quarter, 4-quarter	Inflation
Terasvirta et al(2004)	Canada, France, Germany, Italy, UK,.U.S., Japan	1960-1999	1-quarter, 4-quarter	Inflation
Medeiros et al. (2019)	U.S.	1990-2015	1-quarter, 4-quarter	Inflation
Marcellino (2007)	U.S.	1959-2004	1-quarter, 4-quarter	Inflation
Garcia et al. (2018)	Brazil	2003-2015	1-quarter, 4-quarter	Inflation
Mahajan & Srinivasan (2019)	India, South Africa, China	2003-2019	1-quarter (India only), 4-quarter	Inflation
Leal et al. (2020)	Chile	2003-2019	1-quarter, 4-quarter	Inflation

Recent IEG Working Papers:

Mohapatra, Sonali Madhusmita (Oct. 2021). Export Performance: A Study of Labour and Capital Intensive Manufacturing Industries in India, Working Paper Sr. No.: 445

Priyadarshini, D. and Kar, Sabyasachi (Sep. 2021). Central Bank Digital Currency (CBDC): Critical Issues and the Indian Perspective, Working Paper Sr. No.: 444

Garg, Bhavesh and Sahoo, Pravakar (Sep. 2021). DO DIFFERENT TYPES OF CAPITAL INFLOWS HAVE DIFFERENTIAL IMPACT ON OUTPUT? Evidence from Time series and Panel Analysis, Working Paper Sr. No.: 443

Garg, Sandhya and Gupta, Samarth (Sep. 2021). Financial Access and Gender Gap in Entrepreneurship and Employment: Evidence from Rural India, Working Paper Sr. No.: 442

Sahoo, Pravakar and Bishnoi, Ashwani (Sep. 2021). Drivers of Corporate Investment Slowdown in India: A Firm Level Analysis, Working Paper Sr. No.: 441

C.S.C. Sekhar and Thapa, Namrata (Sep. 2021). AGRICULTURAL MARKET IMPERFECTIONS AND FARM PROFITABILITY IN INDIA, Working Paper Sr. No.: 440

Sahoo, Pravakar and Bishnoi, Ashwani (Aug 2021). Investment Slowdown in India: Role of Fiscal-Monetary policy and Economic Uncertainty, Working Paper Sr. No.: 439

Ghosh, Nilabja; Rajeshwor, R. and Narayan, Hrishabh(Aug. 2021). Economy's response to Rainfall and Economic factors, Working Paper Sr. No.: 438

IEG Working Paper No. 446

INSTITUTE OF ECONOMIC GROWTH

University Enclave, University of Delhi
(North Campus) Delhi 110007, India
Tel: 27667288/365/424
Email: system@iegindia.org