BANK OF ENGLAND

Staff Working Paper No. 923
Forecasting with machine learning methods and multiple large datasets

Nikoleta Anesti, Eleni Kalamara and George Kapetanios

September 2024
This is an updated version of the Staff Working Paper originally published on 28 May 2021

Staff Working Papers describe research in progress by the author(s) and are published to elicit comments and to further debate. Any views expressed are solely those of the author(s) and so cannot be taken to represent those of the Bank of England or to state Bank of England policy. This paper should therefore not be reported as representing the views of the Bank of England or members of the Monetary Policy Committee, Financial Policy Committee or Prudential Regulation Committee.

BANK OF ENGLAND

Staff Working Paper No. 923

Forecasting with machine learning methods and multiple large datasets

Nikoleta Anesti,[1] Eleni Kalamara[2] and George Kapetanios[3]

Abstract

The usefulness of machine learning techniques for forecasting macroeconomic variables using multiple large datasets is considered. The predictive content of surveys is compared with text-based indicators from newspaper articles and a standard macroeconomic dataset, extending the evidence on the contribution of each dataset in predicting economic activity. Among the linear models, the Ridge regression and the Partial Least Squares models report the largest gains consistently for most of the forecasting horizons, and among the non-linear machine learning models, Support Vector Regression performs better at shorter horizons compared to the Neural Networks and Random Forest that yield more accurate forecasts up to two years ahead. Text-based indicators have similar informational content to surveys, albeit combining the two datasets provides with more accurate forecasts for most of the forecast horizons. The largest forecasting gains are overwhelmingly concentrated at the shorter horizons for the majority of models and datasets and they decrease significantly after one year. Non-linear machine learning models appear to be mostly useful during the Great Financial Crisis and perform similarly to their linear counterparts in more normal periods.

Key words: Forecasting, survey data, text indicators, machine learning.

JEL classification: C53, C55.

(1) Bank of England. Email: Nikoleta.Anesti@bankofengland.co.uk
(2) King's Business School, King's College London. Email: Eleni.Kalamara@kcl.ac.uk
(3) King's Business School, King's College London. Email: George.Kapetanios@kcl.ac.uk

This paper was previously circulated under the title, 'Forecasting UK GDP growth with large survey panels'. The views expressed in this paper are those of the authors, and not necessarily those of the Bank of England or its committees. We are grateful to the editor, Erricos John Kontoghiorghes, and two anonymous referees for many insightful comments, which have substantially improved the paper. All errors are our own.

The Bank's working paper series can be found at www.bankofengland.co.uk/working-paper/staff-working-papers

Bank of England, Threadneedle Street, London, EC2R 8AH
Email enquiries@bankofengland.co.uk

© Bank of England 2024
ISSN 1749-9135 (on-line)

1 Introduction

Extensive research focuses on identifying optimal predictors for monitoring and forecasting macroeconomic conditions, crucial for informed economic and policy decisions and enhancing forecast resilience during crises and recessions. Economic forecasts are regularly published, such as the Bank of England's quarterly Monetary Policy Reports, which assess economic conditions and project key macroeconomic variables for interest rate decisions.

Most traditional and widely used forecasting models for economic variables rely on fitting data to a pre-specified relationship between the input variables (indicators) and the output (target) variable. These models thereby make an implicit assumption of a stochastic process underlying the true relationship between the target variable and the indicators. Machine Learning (ML) methods present an alternative to traditional forecasting techniques. ML models can outperform traditional forecasting methods because they focus on out-of-sample (rather than in-sample) performance and better handle nonlinear interactions among a large number of predictors. ML methods are specifically designed to learn complex relationships from past data while resisting the tendency of traditional methods to over-extrapolate historical relationships into the future. Indeed, an empirical literature has emerged which suggests that ML methods often outperform traditional linear regression-based methods in terms of both accuracy and robustness.

Traditional forecasting models for economic variables rely on fitting data to a pre-specified relationship between input indicators and the output variable, assuming an underlying stochastic process. Machine learning (ML) methods present an alternative, frequently surpassing traditional techniques by emphasizing out-of-sample performance and effectively handling nonlinear interactions among numerous predictors. Designed to learn complex relationships from past data, ML methods avoid over-extrapolating historical patterns. As a result, a number of studies have applied machine learning methods in the context of macroeconometrics and especially macroeconomic forecasting. Examples include Ahmed et al. (2010), Stock and Watson (2012), Ng (2014), Smeekes and Wijler (2018), Diebold and Shin (2019), Coulombe et al. (2020), Medeiros et al. (2021) and Joseph et al. (2021). Although ML models have faced the "black-box critique" regarding interpretability, studies, such as Joseph (2019) and Zhao and Hastie (2021) have contributed towards the interpretability of ML models.

The contribution of this paper is three-fold. First, it extends the empirical research on machine learning models in macroeconomic forecasting, specifically predicting monthly UK GDP growth from one month ahead up to two years, using data from 2000 to 2018. Among linear models, the Partial Least Squares (PLS), Ridge, and Elastic Net (EN) based forecasts show better predictive performance compared to benchmark specifications, such as Principal Component Regression (PCR) and a simple AR(1) model, for up to one year ahead. Among non-linear models, Random Forests and Support Vector Regression (SVR) yield the most accurate predictions in the out-of-sample period.

Second, the role of the information set is examined, testing the predictive content of survey data, text-based indicators, and official statistics to predict economic activity. Large panels of survey data for the UK economy from businesses and consumers are collected, including questions about the current and future state of the economy, orders, labor market prospects, and consumer financial outlook. Despite central banks widely using survey data due to their timeliness compared to official releases, this paper collects more disaggregated information. This approach takes advantage of the richer information set usually accompanying survey releases, rather than focusing solely on headline balances, recognizing that different indicators might be more or less useful at different times or phases of the business cycle.

There exists a trade-off between the precision of signals extracted from indicators and their timeliness. Real activity indicators, such as the Industrial Production Index, directly enter Gross Domestic Product (GDP) calculations and are highly correlated with GDP, making them core macro indicators for GDP forecasting. However, most macro data are released with substantial delays compared to the reference month. Conversely, business and consumer surveys are usually available within or at the end of the reference month, offering increased timeliness at the expense of lower precision. Surveys are generic, often referring to broad concepts like "business situation," with respondents typically indicating the direction of change (improvement, no change, deterioration). Despite this, survey data have useful properties for forecasting: they are subject to minimal revisions, offer broader sectoral coverage than typical macro data, and include respondents' views on future developments (e.g., sales expectations), which may have leading properties beneficial for short-term forecasting.

Recent research explores text as an alternative high frequency data source to answer economic and policy related questions (see Gentzkow et al. (2019); Bholat et al. (2015), for a review). This study compares the predictive content of the disaggregated survey balances with 15 text-based indicators suggested by Kalamara et al. (2020) that aim to capture uncertainty and sentiment in the UK economy. These metrics are calculated by applying existing text analytics methods to newspaper articles and modified appropriately to obtain valuable information in real time. As such, we create an alternative dataset that contains 'soft' information extracted from newspaper articles and conduct an extensive out-of-sample evaluation exercise to provide further empirical evidence on the use of surveys and text indicators for predicting economic activity by exploring different linear and non-linear machine learning techniques.

The majority of forecasting gains are concentrated in the short run, with models like Elastic Net (EN) and Lasso regressions maintaining some forecasting superiority over the benchmark for more than one year ahead. The usefulness of "soft" information from business and consumer surveys is examined in comparison to text indicators and a broader macroeconomic and financial time series, following the spirit of the Stock and Watson (2002c) datasets. It is found that the richer information set marginally improves predictions for longer horizons, whereas for shorter horizons (one and three steps ahead), the surveys-only models outperform both text and macro-based forecasts. Additionally, combining survey and text indicators yields the most accurate predictions, with gains up to 30% at one month ahead for the Ridge regression forecasts. However, for longer horizons (one to two years ahead), all models, except the Neural Network (NN) and Support Vector Regression (SVR), perform better using only macro series information.

The final contribution is an examination of the Great Financial Crisis as a potential source of unusually large forecast errors in the sample period. The performance of the models is significantly affected by this period, with non-linear specifications better capturing downturns in the data compared to linear counterparts. Similar to Coulombe et al. (2020), it is found that the forecasting gains of non-linear techniques are associated with high macroeconomic uncertainty and financial stress, suggesting that machine learning is particularly useful for macroeconomic forecasting by capturing important nonlinearities in these environments.

The remainder of the paper is organized as follows: Section 2 describes the different models and explains the tuning parameter selection process. Section 3 summarizes the datasets compiled. Section 4 outlines the main features of the out-of-sample forecasting exercises and reports the empirical results. Appendices A and B provide a detailed description of the datasets and additional forecasting results for robustness.

2 Machine Learning Methods

We briefly review the machine learning methods we use to forecast UK GDP growth. We refer the reader to Kapetanios and Papailias (2018) for a more comprehensive review of all the methodologies. Given the complexity of the economic system and the vast availability of the data, the linearity assumption is potentially restrictive, hence a variety of methods originating from the machine learning literature are becoming increasingly popular. As such, we review some supervised machines algorithms here i.e, random forest regressions, neural networks and support vector regressions that allow us to incorporate non-linearities but also to exploit the entire span of the independent variables without imposing that they all carry useful information for the prediction of the target variable. We conclude this section by summarising the approach we apply on how to pin down the tuning parameters required for each model.

To fix ideas, let y_t, $t = 1, \dots, T$ be the target variable and $x_t = (x_{1t}, \dots, x_{Nt})\prime$ be the set of predictors with N potentially large. Without any assumption for the underlying data generating process y_t can written in the form

$$y_t = \alpha + g(x_{1t}, \dots, x_{Nt}) + u_t. \tag{1}$$

The aim is to provide estimates for future values of y_t using the function g obtained through training a machine learning model. We focus on the Ridge regression, Lasso regression, and the EN for which we consider a linear model of the form:

$$y_t = \alpha + \sum_{i=1}^{N} \beta_i x_{it} + u_t. \tag{2}$$

Additionally, we explore the SVR, a random forest and a NN. We finally include in our review some dimensionality reduction methods, like the PCR and PLS that have been used as a standard approach in the macroeconomic forecasting literature to produce a smaller set of generated regressors.

2.1 Ridge Regression

Shrinkage methods have been suggested to produce effective estimates using different penalisation schemes. The main idea is that all the coefficients of the variables which are not part of the true model approach or become 0. Ridge regression penalises the residual sum of squares with the sum of squared coefficients. This forces the coefficients with a minor contribution in the model to shrink substantially and approach zero, but never become exactly zero. Under this framework, $\hat{\beta}^{\text{Ridge}}$ is computed following the optimisation problem below:

$$\min_{\beta_N} \left\{ \sum_{t=1}^{T} (y_t - a - \beta_N' x_{t,N})^2 + \lambda \sum_{i=1}^{N} \beta_i^2 \right\}. \tag{3}$$

2.2 Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO regression penalises the sum of squared residuals with the L1 norm, i.e. the sum of absolute coefficients. Some of the coefficients are set exactly to 0 which tends to give more parsimonious results. The LASSO estimators $\beta^{L\hat{A}SSO}$ are then computed by solving the following optimisation problem:

$$\min_{\beta_N}\left\{\sum_{t=1}^{T}(y_t - a - \beta_N' x_{t,N})^2 + \lambda\sum_{i=1}^{N}\mid\beta_i\mid\right\}. \tag{4}$$

2.3 Elastic Net

A slightly different approach, called Elastic Net Zou and Hastie (2005) combines the L1 and L2 norms penalties. This method enjoys both the shrinkage of the coefficients (Ridge Regressions) and the variable selection (LASSO). The idea of implementing two penalty schemes is based on the aim to include all the true regressors in the model, even if they are strongly correlated. The "naive" estimators of elastic net, $\beta^{naiveEN}$ are computed by solving the problem:

$$\beta^{naiveEN} = \min_{\beta_N}\left\{\sum_{t=1}^{T}(y_t - a - \beta_N' x_{t,N})^2 + \lambda_1\sum_{i=1}^{N}\mid\beta_i\mid + \lambda_2\sum_{i=1}^{N}\beta_i^2\right\}. \tag{5}$$

2.4 Support Vector Regression (SVR)

The Support Vector Machine technique was originally introduced as a classification method based on the idea of using support vectors to represent the class boundaries in the classification problems Vapnik (1998).

Support vector machine regression solves

$$\min_{w,b,\vec{\xi},\vec{\xi^*}} \frac{1}{2}\|w\|^2 + C\sum_t \xi_t + C\sum_t \xi_t^*, \tag{6}$$

subject to

$$\begin{cases} y_t - \vec{w}^T\phi(\vec{x}_t) - b \leq \epsilon + \xi_t^*, \\ \vec{w}^T\phi(\overrightarrow{x_t}) + b - y_t \leq \epsilon + \xi_t, \\ \xi_t, \xi_t^* \geq 0 \quad \forall t. \end{cases} \tag{7}$$

where $K(\vec{x}_t, \vec{x}_{t\prime}) = \phi(\vec{x}_t)^T\phi(\vec{x}_{t}\prime)$ is a kernel function.

2.5 Tree Models and Random Forests

Tree models is a non-parametric method tailored for both regression and classification problems. Tree models based on the idea to divide consecutively split the in-sample dataset until an assignment criterion with respect to the target variable into a "data bucket" (leaf) is reached. The aim is to minimise the objective function within areas of the target space (buckets) conditioned on the input X.

A disadvantage of regression trees is that they are not identically distributed: they are built adaptively to reduce the bias. Growing decision trees in the above form may lead to severe over-fitting. An alternative modelling set up that overcomes this problem is the so called "Random Forest" (see Breiman (2001) for details). Random forest "grows" a set of uncorrelated trees which are estimated separately. Hastie et al. (2009) provides an algorithm for growing a random forest from a specific number of trees.

2.6 Artificial Neural Networks

Neural networks are similar to linear and non-linear least squares regressionS and can be viewed as an alternative statistical approach to solving the least squares problem. Both neural networks and conventional regression analysis attempt to minimise the sum of squared errors.

The architecture of the neural net can vary with respect to the hidden layers and nodes included on each layer. For example, note that each layer can have a different number of nodes. Determining the architecture of the neural network is in fact, a model selection problem. To select the neural net specification we investigate the predictive performance of different network architectures through extensive cross-validation. We start from a shallow network with one hidden layer and expand the layers up to five, that is four hidden layers and the output layer. Similarly, we select the nodes on each layer setting up a grid from one to four nodes. We find that using five layers with 4 nodes each yield the best predictive performance.

2.7 Partial Least Squares (PLS)

Partial least squares regression makes use of the dependent variable y_t to identify new features that not only approximate the old features well, but also that are related to the response as first introduced by Wold (1985). PLS is a dimensionality reduction technique that estimates multiple regressions under a large but finite number of regressors. PLS is similar to PCR in that we estimate factors that are linear combinations of the x_t covariates and then the obtained factors are used in the regression instead of the x_t. A significant difference is that PLS estimates factors are estimated by maximising both the variability of the y_t and the covariates x_t, while PCR only considers the variability of the covariates. PLS approach seeks to find linear combinations that help explain both the dependent variable and the regressors.

Table 1: Summary table of the tuning parameters

Models	**Tuning Parameters**
Ridge	$\lambda = 1$
Lasso	$\lambda = 1$
Elastic Net	$\lambda_1 = 0.8$ and $\lambda_2 = 0.7$
SVR	$\epsilon = 0$, $C = 700$
Random Forest	300 trees, maximum depth of 8, minimum sample split of 2
Neural Network	5 layers with 4 nodes each

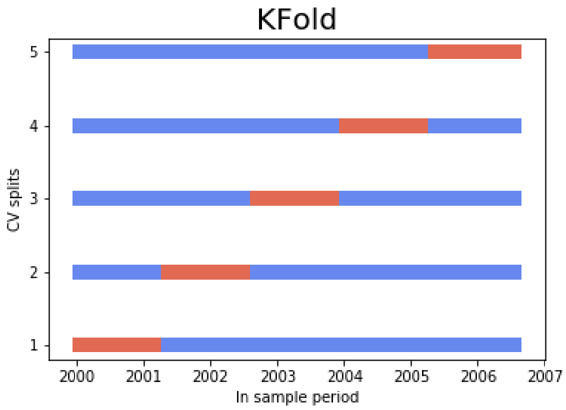


Figure 1: Schematic representation of a 5-fold cross validation. The figure depicts the splitting strategy we perform to the in-sample period at each point in time t. The model is estimated on 80 % of the data and tested on the remaining 20%. The red area shows the test set of observations for each split.

2.8 Selecting the tuning parameters

We discuss how we select the tuning parameters for the above mentioned methods using cross validation that has recently become a popular way to pin them down as they can be used to any model, including those for which the derivation of information criteria is not feasible as noted in Arlot and Celisse (2010), Bergmeir et al. (2014) and Coulombe et al. (2019). All of our models involve some kind of parameter selection prior to estimation. The details for each model are summarised in the following Table 1.

Cross validation techniques consist of estimating a particular specification over the training sample and computing the forecast performance over the validation sample. In a time-series forecasting setting, a cross-validation exercise is conducted only in-sample to avoid information leakage Kalamara et al. (2020). We opt to apply the standard 5-fold cross validation to select the tuning parameters required for each model specification. Coulombe et al. (2019) report that K-fold provides the best performance compared to a set of different cross validation procedures and information criteria.

The method is based on a resampling scheme; we use 5 folds, i.e. the in-sample set is randomly split into five disjoint subsets as shown in Figure 1. For each one of the 5 subset and set of the tuning parameters considered, 4 subsets are used for estimation (training set) and the remaining corresponding observations of the in-sample set is used as a test subset (validation set) to generate forecasting errors. As a performance metric and condition to select the best tuning parameter, we consider the average mean squared error in the validation set. Once the tuning parameters are selected, each model is estimated using the entire in-sample period and used to generate out-of-sample predictions.

3 Datasets

We describe various aspects of the datasets we explore for forecasting UK GDP growth. We start by describing the disaggregated business and consumer survey balances used for the forecasting exercises and then we move to explain how to expand the dataset to include text based indicators and other macroeconomic variables related to economic activity, prices, labour market statistics and some financial market data and explore if there's additional predictive content.

3.1 Survey Dataset

Collection and publication of official data are subject to substantial processing delays; For example, in the United Kingdom the monthly index of industrial production (including manufacturing output) is published at least 40 days after the end of the month to which it refers to. Therefore, significant resources are devoted to exploit alternative sources of information in order to gauge the continuously evolving state of the real economy. As such, the use of more timely and reliable information about current economic conditions is crucial for policy makers and expectations formation. Quantitative survey balances from businesses and firms are prime candidates to this end and they have proven very useful for short term forecasting Lahiri and Monokroussos (2013); Bańbura et al. (2013); Giannone et al. (2008a) due to their timeliness (they are usually published at a monthly frequency and just a couple of days after the end of the reference month) and their high correlation with GDP growth. Many policymakers and market participants take recourse to survey evidence to measure current economic conditions. This is widely evidenced by monetary policy communications, which frequently point to survey evidence when describing the current macroeconomic situation.

Their publication is usually accompanied by some discussion of what can be learnt from them about the most recent movements and short-term expected future movements in economic activity, at least in the sector to which the surveys relate to. These business surveys ask *inter alia* whether, after adjusting for normal seasonal movements, output has risen, stayed the same or fallen in recent months, what are the firms' expectations for the next quarter or the year ahead or what are the employment prospects in their business. At the same time, 'soft' information from the consumers' perspective is contained in the consumer confidence indicators published by the European Commission, where questions about the financial situation of the households or the general economic situation are included and can provide with a more timely assessment of consumers' perception of current and expected economic conditions. They have also been found to be empirically useful for forecasting movements in economic conditions.

More details on the exact definition and wording of the survey balances questions can be found in the Appendix Table A.1.

3.2 Text Dataset

As an additional check, we explore whether other types of soft information add any significant value to our results, in terms of improving the forecasting performance. Recently, extensive research has been undertaken that explores text as an alternative data source and its usefulness to answer macroeconomic and policy related questions. To this end, we make use of the 15 text-based indicators suggested by Kalamara et al. (2020) that aim to capture uncertainty and sentiment in the economy. The metrics are created by applying existing text

Positive and negative dictionary	Boolean	Computer science-based
Financial stability Correa et al. (2017)	Economic Uncertainty Alexopoulos et al. (2009)	VADER sentiment Gilbert (2014)
Finance oriented Loughran and McDonald (2013)	Monetary policy uncertainty Husted et al. (2017)	'Opinion' sentiment Hu and Liu (2004); Hu et al. (2017)
Afinn sentiment Nielsen (2011)	Economic Policy Uncertainty Baker et al. (2016)	Punctuation economy (this paper)
Harvard IV (used in Tetlock (2007))		
Anxiety-excitement Nyman et al. (2018)		
Single word counts of "uncertain" and "economy"		
tf-idf applied to "uncertain" and "economy"		

Table 2: The three broad categories of algorithm-based text metrics used.

analysis methods to newspaper articles and modified appropriately to obtain valuable information in real time.

In a forecasting setting, Kalamara et al. (2020) convert UK newspaper articles to time series text-metrics applying a range of text analysis methods with the raw text dataset comprising of the three mostly circulated newspapers in the UK, i.e. the Daily Mail, the Guardian and the Daily Mirror. Then, the generated indicators are used as predictors to forecast key economic indicators. Their findings indicate that text significantly improves the out of sample forecasting performance of GDP growth relative to popular benchmarks of the forecasting literature and this is particularly relevant during periods of stress. This is an important finding because it suggests that text-based data can act as a strong complement to high frequency financial market data and to less timely, surveys or macroeconomic data.

The text analytics models used to create the indices can be found in Table 2, and are the most commonly used in the literature to extract signal from text. The dataset includes a wide range of methods from simple counts of specific term occurrences such as the word 'uncertain' and 'economy' in each article divided by the number of words in the article, to more sophisticated approaches adopted from text-specific algorithms. The numerical scores for a particular month are found from the mean of the scores of the articles that were published in that month.

Broadly speaking, the methods fall into three main categories: dictionary-based which associate specific scores (e.g. positive or negative sentiment) and count the net score per article, boolean methods which provide a count of articles only if the terms in an article satisfy some logical condition and computer-science based which follow different algorithmic procedures drawn from computational linguistics.

3.3 Macro Dataset

We are also interested in evaluating the informational content of a wider dataset that includes not only "hard" indicators as published by the official statistical agencies, like data on production, services, prices and labour market statistics, but also some financial market data taken as end-of-month averages. We use 49 additional series that are selected to represent broad categories of macroeconomic time series: real output and income, employment and hours, real retail, manufacturing and sales data, international trade, labor costs, price indexes, interest rates, stock market indicators, and foreign exchange measures for the UK economy (further details on the extended dataset can be found in the Appendix Table A.3). This type of datasets has been widely used in the literature Stock and Watson (2002a,b) for the US Economy and has been empirically proven to correlate fairly well with future GDP growth. Therefore, we employ a battery of linear and non-linear models to understand the predictive content of these series and ask whether they have any additional predictive ability over and above the panel of survey balances and if they do for which forecast horizons and which phases of the business cycle.

The variables used in all the exercises are considered at a monthly frequency, starting in January 2000 until August 2018, purely due to data availability constraints. Although we do not explicitly take into account the real time release calendar of the input series, in the design of the out-of-sample forecasting evaluation exercise we only consider the series that were available at every point in time. Our target variable for this exercise is the monthly GDP estimate as published by the Office of National Statistics, transformed to three month-on-three month growth rates. All series are finally seasonally adjusted.

4 Empirical Results

We first outline the details of the out of sample forecast evaluation exercise In Section 4 including the models we use as benchmark specifications and then we present and discuss the main empirical findings. We design the evaluation exercise in such a way so that we can assess the usefulness of various data sources individually and collectively.

We proceed with testing the informational content of a more extensive dataset, first using text based indicators only and then more standard macroeconomic time series, related to activity, prices, financial market data, labour market in the spirit of Stock and Watson (2002c) and report results in terms of relative Root Mean Square Forecast Errors (RMSFE) estimated using the same information set every time. We then combine all the "soft" information from the surveys and text based indicators and examine whether there's additional predictive content in a larger information set or simply the use of survey balances is sufficient to forecast economic activity from one month up to two years ahead.

Subsequently, we perform an additional exercise in which we compare across different models rather than datasets to understand better the models' forecasting performance. We conclude this part by asking whether during recessionary episodes, like the Great Financial Crisis, linear and non-linear models perform similarly or non-linear models can be more useful in an out-of-sample forecasting exercise.

4.1 Evaluation Design

We use the data from January 2000 until August 2006 as training sample (estimation sample), as a result the origin of the forecast exercise is September 2006 and we end up with 145 observations for the out-of-sample evaluation period that ends in August 2018. We set a maximum of 24 months (two years) ahead forecasts generated at every month and we compute **direct** forecasts from models estimated with expanding samples over the out-of-sample period, that is, at each forecast origin we re-estimate each model and we use all observations available up to the forecasting origin.

We use the relative RMSFEs as a measure of forecasting performance against two different benchmarks, an AR(1) and a PCR, and the Diebold and Mariano (1995) t statistics to test for equal accuracy with the Newey-West estimator with maximum order increasing with the horizon. The use of principal components analysis for the estimation of factor models is, by far, the most popular factor extraction method. It has been popularised by Stock and Watson (2002a,b) and has been widely used in the macroeconomic forecasting literature. Additionally, AR(1) models despite being simple, there's substantial empirical literature noting that, overall, they are particularly tough benchmarks to beat using both linear and non-linear models as noted in Carriero et al. (2019).

Table 3: Relative RMSFE with Survey Data

relative to PCR	(1)	(3)	(6)	(9)	(12)	(24)
Lasso	0.991	0.977	0.961	0.940	0.950	0.990
Ridge	**0.726*****	**0.824***	0.846	0.891	0.935	1.104
Elastic	**0.817*****	**0.843***	0.843	0.868	0.908	1.065
PLS	**0.826*****	**0.841***	0.840	0.861	0.899	1.068
Random Forest	**0.879*****	0.901	0.856	0.885	0.914	1.045
SVR	**0.722*****	**0.859***	0.872	0.889	0.920	1.074
NN	0.768	0.874	0.869	0.900	0.937	1.103
relative to AR(1)	**(1)**	**(3)**	**(6)**	**(9)**	**(12)**	**(24)**
Lasso	**0.720*****	**0.826***	0.846	0.896	0.937	1.109
Ridge	**0.702*****	0.820	0.890	0.969	**1.011***	0.975
Elastic	**0.726*****	**0.716***	0.758	0.907	0.993	0.965
PLS	**0.792*****	0.789	0.830	0.939	0.998	0.976
Random Forest	**0.751*****	0.770	0.835	0.877	0.972	0.972
SVR	**0.625*****	0.850	0.891	0.970	1.010	0.985
NN	**0.687*****	0.895	0.923	0.986	1.011	0.968

Note: Top panel: relative RMSFEs across different specifications against a **PCR** model using only *Survey* data for $h = 1$ up to $h = 24$ months ahead. Bottom panel: relative RMSFEs across different specifications against a **AR(1)** model using only *Survey* data. *,** & *** denote rejection at 10%, 5% and 1% level of the null hypothesis of equal forecasting method accuracy of the Diebold and Mariano (1995) test against the respective benchmark models.

4.2 Comparison across datasets

Table 3 summarises the results of the benchmark specification. In the upper panel of the table we report the relative RMSFEs against the PCR benchmark and in the lower panel we compare our results against an AR(1) model using only the disaggregated survey balances as indicators. As a result, values larger than 1 indicate that the benchmark is more accurate and values less than 1 indicate that the model under consideration is more accurate. Overall, based on our empirical results, the PCR is a harder benchmark to beat compared to an AR(1) and as such we will use it as our main model to benchmark our results for all the subsequent exercises to comment on the results.

Among the linear models, Ridge, PLS and the elastic net are among the ones reporting the largest gains consistently for all forecasting horizons with that go up to 27% for the Ridge regression-based forecasts against the PCR benchmark at 1 month ahead horizon. At longer horizons, the PLS and the EN regressions perform better with gains up to 10% against the PCR at one year ahead forecasts and the Ridge based forecasts report gains ranging from 6% to 15% for medium term forecasts (from 6 to 12 months ahead).

The non-linear models do not appear to perform substantially better than the linear ones for most of the forecasting horizons, when using just survey balances as input variables. Within this class of models, however, SVR reports the largest gains for shorter horizons compared to the NN and Random Forest with forecasting gains 23% and 12% at one month ahead forecasts respectively.

Another interesting feature appears in Table 3: although the majority of the models are able to maintain the forecasting gains against the benchmark specifications for longer forecasting horizons (one year ahead), it seems that these gains are getting smaller after the first 6 months. Notably, there's no other model, apart from the Lasso regression, able to outperform the PCR specification at the two-years horizon. This finding is consistent with the predictive content of business and consumer surveys in terms of forecasting economic activity: 'soft' information from survey data is mostly relevant for short term forecasting, whereas for longer term one needs to explore official releases further. We will explore this last finding further in the next section.

Table 4: Relative RMSFE Across Different Datasets

Text Indicators	(1)	(3)	(6)	(9)	(12)	(24)
Lasso	**0.902***	0.914	0.896	0.898	0.918	1.033
Ridge	**0.904***	0.914	0.897	0.898	0.918	1.033
Elastic	0.986	0.975	0.959	0.940	0.950	0.990
PLS	**0.937*****	0.932	0.924	0.916	0.932	1.012
Random Forest	**0.930**$*{*}*$	0.941	0.936	0.927	0.940	1.014
SVR	**0.823*****	**0.882***	0.904	0.912	0.939	1.040
NN	**0.879*****	0.904	0.950	0.903	0.924	1.056

Macro Indicators	(1)	(3)	(6)	(9)	(12)	(24)
Lasso	**0.959*****	0.956	0.940	0.931	0.945	0.997
Ridge	**0.855*****	0.884	0.863	0.882	0.896	1.065
Elastic	**0.946*****	0.947	0.932	0.925	0.940	1.003
PLS	**0.885*****	0.901	0.876	0.888	0.899	1.045
Random Forest	**0.933*****	0.947	0.920	0.921	0.938	1.020
SVR	**0.722*****	0.886	0.872	0.896	0.929	1.118
NN	**0.802*****	**0.895***	0.928	0.897	0.934	1.125

Survey & Text	(1)	(3)	(6)	(9)	(12)	(24)
Lasso	0.992	0.977	0.960	0.940	0.950	0.990
Ridge	**0.698*****	**0.831***	0.855	0.901	0.946	1.097
Elastic	**0.805*****	**0.837***	0.844	0.873	0.913	1.062
PLS	**0.818*****	**0.834***	0.840	0.867	0.909	1.057
Random Forest	**0.794*****	0.882	0.873	0.882	0.916	1.044
SVR	0.663	0.874	0.865	0.895	0.936	1.125
NN	0.667	0.851	0.878	0.903	0.952	1.148

Note: Top panel presents relative RMSFEs across different specifications using only *Text* data. Middle panel presents relative RMSFEs across different specifications using only *Macro* data. Bottom panel presents relative RMSFEs across different specifications combining *Survey* and *Text* indicators. The results are reported against a **PCR** model as a benchmark estimated using the same data source *,** & *** denote rejection at 10%, 5% and 1% level of the null hypothesis of equal forecasting method accuracy of the Diebold and Mariano (1995) test against the benchmark specification.

4.2.1 Text and other macro variables

As explained in Section 3, we want to understand whether other types of 'soft' information could have predictive value for GDP forecasting. An alternative data source that we explore in this content is text-based indicators capturing uncertainty and sentiment about the economy. Text based indicators of sentiment and uncertainty might contain some of the forward looking information that is crucial for policymakers in the decision making process. As such, we are interested in exploring their predictive content and compare it with this of surveys. We explained in Section 3.2 how the measures are constructed and we now use them in an identical forecasting environment to test their predictive content. The results are summarised in the upper panel of Table 4. The performance of the models is evaluated in terms of relative RMSFE with the benchmark specification being a PCR model estimated using the same indicators for all the models.

Overall, their predictive content is fairly similar to this of survey balances: their reporting gains are mainly concentrated in the shorter horizons (1-6 months ahead) and they gradually decrease as the forecast horizon increases. Within that, on average the non linear methods perform better that the linear ones, with predictive gains ranging from 5% up to 14% for the NN that is the best performing specification among both the linear and non-linear models that is in line with what Kalamara et al. (2020) find in their paper. SVR also performs better than the the PCR benchmark with predictive gains from 6% in the one-year ahead forecasts to 18% for one-month ahead. In terms of predictive content, while the surveys and the text based forecasts are pretty similar, the surveys-only models perform marginally better in shorter horizons and the text-only models are able to maintain some of their gains for longer horizons up to one-year ahead. All models, both linear and non-linear, perform better than the benchmark specification, a PCR model, and the gains range between 1% to 17%. Similarly to what Ferrara and Simoni (2019) highlight, information contained in alternative data sources, Google data in their case, is predominantly useful for predicting the shorter horizons. As explained in Section 3.3, we also consider a macroeconomic dataset in the spirit of Stock and Watson (2002c) in attempt to evaluate the informational content of additional official data sources like economic activity indicators, prices and financial market data that have been traditionally used for macroeconomic forecasting published from the statistical agency. We repeat the same out of sample evaluation exercise with a different dataset and the results are summarised in the middle panel of Table 4.

The overall message is that the relative RMSFEs based on the out of sample exercise suggest that information from macroeconomic time series is only marginally more useful for some non-linear methods like the SVR and the NN for up to one year ahead forecasts compared to the alternative of using text based indicators, although the macro-only forecasts from all the models except from the Lasso regression cannot maintain their predictive gains for longer than one-year ahead against the benchmark specification. On average, the macro-only forecasts do not outperform the forecasts of text indicators suggesting that the information contained in newspaper articles is comparable, if not superior, to the traditional macroeconomic dataset that has been widely used for forecasting.

In a central bank environment, policymakers need to make decisions and weight all the available information when publishing a set of forecasts about the future path of the economy. With official data published with a substantial delay, they need to make decisions in real-time with incomplete information. As such, combining information from surveys and text indicators might provide them with a better understanding of the current and future economic conditions and this is in practice how policy making is performed in real time.

Additionally, as all these methods are designed for handling large datasets, we performed the same exercise combining all the 'soft' information at hand, i.e. the surveys and the text information to understand whether we can achieve better forecasting performance with larger information set. The results are reported in the bottom panel of Table 4 against a PCR model that also uses both surveys and text indicators. The combination of the survey and the text indicators does not seem to offer any substantial improvement on the results: the general pattern of the relative RMSFEs is closer to that of the forecasting exercise when using just the survey balances suggesting that enlarging the information set of survey balances with text indicators does not add much predictive content. For robustness, we also consider a combined dataset with the surveys, the text and the macroeconomic indicators and the overall narrative doesn't change in any substantial way. The results are summarised in Appendix B for completeness.

4.3 Comparison across models

A final exercise summarised in Table 5, is to compare the forecasting performance across different models. We compute relative RMSFEs using exactly the same benchmark, a PCR estimated using only survey data as shown in Table 3 for all datasets to evaluate the predictive ability across models. In this way, we could get some additional insights in terms of which models work best for specific datasets. We extract the same forecast errors from Table 4 but this time we use as a benchmark the same PCR model estimated with survey balances

Table 5: Relative RMSFE Across Different Models

	Text Indicators					
	(1)	(3)	(6)	(9)	(12)	(24)
Lasso	**0.906*****	0.916	0.897	0.898	0.918	1.033
Ridge	**0.908*****	0.916	0.898	0.898	0.918	1.033
Elastic	0.990	0.977	0.961	0.940	0.950	0.990
PLS	**0.941*****	0.934	0.925	0.916	0.932	1.011
Random Forest	0.934	0.943	0.937	0.927	0.940	1.013
SVR	**0.827*****	**0.884***	0.906	0.912	0.939	1.040
NN	**0.883*****	0.906	0.952	0.903	0.923	1.056
	Macro Indicators					
	(1)	(3)	(6)	(9)	(12)	(24)
Lasso	**0.970*****	**0.965****	0.947	0.937	0.952	0.953
Ridge	**0.866*****	**0.892****	0.869	0.887	0.903	1.018
Elastic	**0.957*****	**0.956****	0.939	0.931	0.948	0.959
PLS	**0.896*****	**0.909****	0.883	0.893	0.906	0.999
Random Forest	**0.944*****	**0.956****	0.926	0.927	0.945	0.975
SVR	**0.730*****	**0.894****	0.878	0.902	0.936	1.068
NN	**0.812*****	**0.904****	0.935	0.903	0.942	1.075
	Survey & Text					
	(1)	(3)	(6)	(9)	(12)	(24)
Lasso	**0.990*****	0.977	0.961	0.940	0.950	0.990
Ridge	**0.696*****	**0.831***	0.855	0.900	0.946	1.097
Elastic	0.804	**0.837***	0.844	0.873	0.913	1.062
PLS	**0.816*****	**0.833***	0.840	0.867	0.909	1.057
Random Forest	**0.792*****	0.882	0.873	0.882	0.916	1.044
SVR	0.662	0.874	0.866	0.895	0.936	1.125
NN	**0.666*****	0.851	0.879	0.903	0.952	1.148

Note: Top panel presents relative RMSFEs across different specifications using only *Text* data. Middle panel presents relative RMSFEs across different specifications using only *Macro* data. Bottom panel presents relative RMSFEs across different specifications combining *Survey* and *Text* data. **All** results are reported against a **PCR** model as a benchmark estimated using on the survey balances dataset from Section 3.1. *,** & *** denote rejection at 10%, 5% and 1% level of the null hypothesis of equal forecasting method accuracy of the Diebold and Mariano (1995) test against the benchmark specification.

only from Section 3.1 to compute the relative RMSFEs. While each panel of theTable 5 corresponds to different data sources, Text, Macro and Survey combined with Text, the benchmark against which all the models are assessed is now the same model estimated with the same input variables to compare the predictive content across all the models at hand. Lasso appears to perform better when using text indicators only compared to macro (and a combination of text and survey) series for all forecast horizons, whereas for the PLS, EN and the Ridge regression a combination of text and survey indicators produces more accurate predictions up to 2 years ahead. The picture is broadly similar when looking at the results from the Random Forest, SVR and NN: the majority of forecasting gains is when information from both survey and text based indicators is combined instead of only using each dataset separately. This finding provides with some further empirical support for the use of data rich environments when forecasting GDP growth, as it is also noted in Kotchoni et al. (2019).

Another interesting insight for the use of macro indicators emerges from the results in Table 5 that is worth highlighting: on average as explained above for the majority of models we consider here, a combination of survey and text indicators yields the most accurate predictions with gains up to 30% at $h = 1$ for the Ridge regression forecasts. However, especially for longer horizons, i.e. between 1-2 years ahead, all models, except from the NN and the SVR, perform better when using information only from macro series. This finding provides some additional empirical evidence for the informational content of macro series that appear to be more relevant for forecasting longer than shorter term.

Overall, we think that the empirical evidence from this comprehensive pseudo out of sample evaluation exercise provides with some insights and guidance to practitioners for forecasting UK GDP growth at monthly frequency. Survey data especially at disaggregate level contain really useful information for predicting economic activity in the short run and text based indicators are also a good alternative to surveys as they appear to carry similar informational content. This implies that in periods when timeliness is a priority, for example during the Covid crisis, text based indicators can be used for a real time assessment of the evolution of the economy.

At the same time, as the number of indicators considered in this exercise is quite large, ML techniques are proven to be useful for this signal extraction problem. In fact, the combination of the survey and text based indicators yield more accurate predictions compared to each dataset individually. This is particularly relevant as so far we have been considering text based indicators as a substitute dataset for surveys, when in fact the combination of the two datasets results in higher forecasting gains.

Some further discussion in the following section provides with additional insights on the use of non linear models during periods of distress.

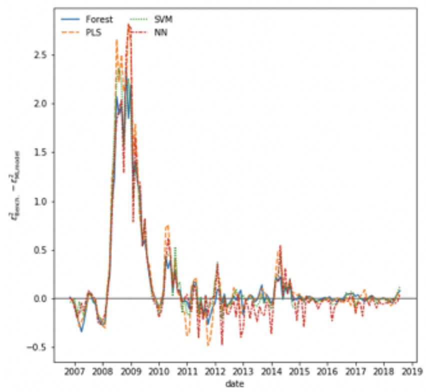
4.4 Are machine learning models more useful around recessionary episodes?

In Section 4, we look at the overall out of sample evaluation results across different datasets and models to assess their ability in tracking UK GDP growth. We focus on the behaviour of the forecast errors of linear and non linear models and whether they can be more/less useful during specific phases of the business cycle in this section. We report results against the PCR benchmark model which uses as input only survey indicators (the corresponding evaluation results are summarised in Table 5). An episode of special interest remains the Great Financial Crisis as it would allows us to test whether highly non-linear models are able to capture turning points more timely and accurately than their linear counterparts.

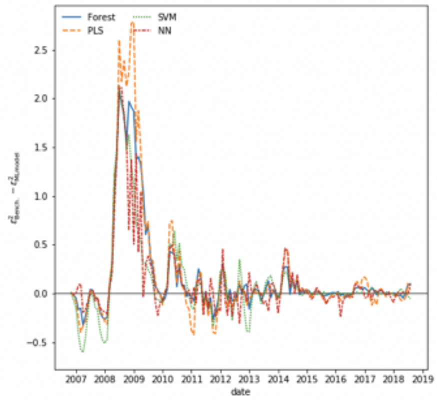
Figure 2 plots the mean squared error differences over the out of sample evaluation period from September 2006 until August 2018. We calculate the error differences between the SVR, NN, Random Forest and PLS models with the benchmark model under consideration, PCR. The errors plotted here are averaged over the h-month ahead out-of-sample forecasts, where h = 1, 3, 6, 9,12 ,24. For the squared errors differences, a point above zero means that the model produces smaller errors than the benchmark model PCR and a point below zero implies that the benchmark is more accurate.

A couple of interesting features stand out in Figure 2. First of all, the largest forecast errors both in absolute and in relative terms (against the benchmark specification) are reported during the Great Financial Crisis and this is to be expected, given the magnitude of the episode and the lack of a similar event in the estimation part of the sample that could have been used to train the models appropriately. However within that, the Machine Learning models perform overall better than the PLS (and the PCR) specification, chosen here specifically as it is on average the best performing linear model. Consequently, the majority of improvement in terms of forecasting errors (highest forecast gains) happens during the Great Financial Crisis. This is in fact a more general finding that is true for all the datasets and their combination that we consider in the paper, lending some empirical evidence to the fact that the machine learning models might be more appropriate to capture large non-linearities in the data that arise during crises compared to their linear counterparts.

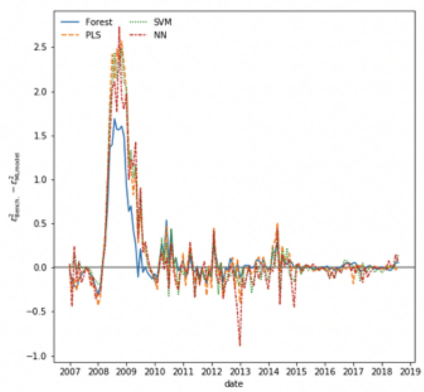
Another interesting point in terms of the relative forecasting performance of the different types of datasets we examine is the direction of the forecast errors difference during the Great Financial Crisis: in all cases, GDP squared error differences comparison



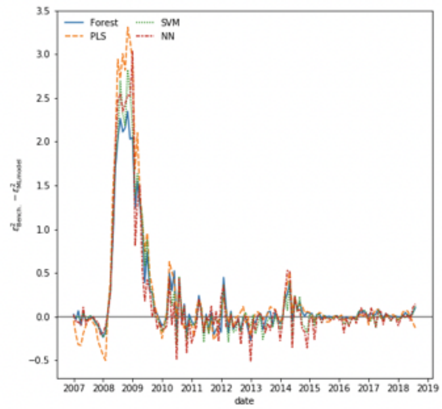
(A) Surveys



(B) Text



(C) Macro



(D) Combination of Surveys, Text and Macro data

FIGURE 2: Mean squared error differences between the PCR benchmark model and other models. Errors are the average of h-month ahead out-of-sample forecasts, where h = 1, 3, 6, 9,12 ,24. Each panel represents a different dataset.

both the linear and non linear models perform substantially better than the benchmark specification. The latter finding is broadly in line with the evidence found in Kalamara et al. (2020), who argue that text in combination with a machine learning model predicts better during stressed times. While their exercise is based on a rather different setting and their forecasts are produced by feeding the machine learning model with high dimensional term frequency vectors, the overall message about the use of Machine Learning models during recessionary episodes is rather consistent. Finally, It is equally interesting to highlight here that during more normal periods the models' forecasting performance is broadly similar against the benchmark specification and as a consequence it appears that using models that account for non linearities becomes relevant during periods of financial and/or macroeconomic uncertainty.

5 Conclusions

We have studied machine learning methods as devices for jointly using multiple large datasets to forecast macroeconomic variables. In particular, we reviewed machine learning techniques to forecast UK GDP growth at monthly frequency for horizons from one month up to two years ahead. To this end, we gather large panels of disaggregated survey data from UK businesses and consumers and explore their informational content. We also consider text based indicators from newspaper articles and a more standard macroeconomic dataset as potential predictors for the target variable and run a battery of out of sample evaluation exercises to assess the predictive ability across models and datasets.

We provide with a comprehensive assessment of the usefulness of the three different datasets, surveys, text based indicators and macroeconomic series, frequently used in macroeconomic forecasting, to predict UK GDP growth at a monthly frequency for up to two years ahead. We find that survey and text-based forecasts perform similarly, however survey-based models are somewhat better at shorter horizons and the text-only models can maintain some of the forecasting gains for longer horizons, namely up to one year ahead. The macro based forecasts are more useful for some non-linear methods like the SVR and the NN for up to one year ahead forecasts compared to the alternative of using text based indicators, although the macro-only forecasts cannot maintain their predictive gains for longer than one year ahead against the benchmark specification.

We also extend the out of sample evaluation exercise so that we can compare across different models and we find that Lasso appears to perform better when using text indicators only compared to macro series for all forecast horizons, whereas for PLS, EN and the Ridge regression a combination of text and survey indicators produces more accurate predictions up to 2 years ahead. The picture is broadly similar when looking at the results from Random Forest, SVR and NN: the majority of forecasting gains is when information from both survey and text based indicators is combined instead of only using each dataset separately.

Finally, we compare linear to non-linear models and we find that PLS is among the best performing linear specification and the SVR performs better on average among the machine learning models. We also bring some additional empirical evidence on the usefulness of machine learning models in forecasting economic activity during turbulent times and their ability to capture non-linearities in the data in a more accurate way.

Acknowlegments

The authors wish to thank the editor, Erricos John Kontoghiorghes, and two anonymous referees for many insightful comments, which have substantially improved the paper. All errors are our own. The views expressed are those of the authors and do not represent those of the Bank of England or any of its Committees.

References

Ahmed, Nesreen K., Amir F. Atiya, Neamat El Gayar, and Hisham El-Shishiny (2010) "An Empirical Comparison of Machine Learning Models for Time Series Forecasting," *Econometric Reviews*, Vol. 29, No. 5-6, pp. 594–621.

Alexopoulos, Michelle, Jon Cohen et al. (2009) "Uncertain times, uncertain measures," *University of Toronto Department of Economics Working Paper*, Vol. 352.

Anesti, Nikoleta, Simon Hayes, Andre Moreira, and James Tasker (2017) "Peering into the present: the Bank's approach to GDP nowcasting," *Bank of England Quarterly Bulletin*, Vol. 57, No. 2, pp. 122–133.

Arlot, Sylvain and Alain Celisse (2010) "A survey of cross-validation procedures for model selection," *Statist. Surv.*, Vol. 4, pp. 40–79.

Baker, Scott R, Nicholas Bloom, and Steven J Davis (2016) "Measuring economic policy uncertainty," *The Quarterly Journal of Economics*, Vol. 131, No. 4, pp. 1593–1636.

Bańbura, Marta, D. Giannone, M. Modugno, and L. Reichlin (2013) "Now-casting and the Real-Time Data Flow," in *Handbook of Economic Forecasting, volume 2A*: Elsevier, Chap. 4, pp. 195–237.

Bergmeir, Christoph, Mauro Costantini, and José M Benítez (2014) "On the usefulness of cross-validation for directional forecast evaluation," *Computational Statistics & Data Analysis*, Vol. 76, pp. 132–143.

Bholat, David M, Stephen Hansen, Pedro M Santos, and Cheryl Schonhardt-Bailey (2015) "Text mining for central banks."

Breiman, Leo (2001) "Random forests," *Machine learning*, Vol. 45, No. 1, pp. 5–32.

Carriero, Andrea, Ana Beatriz Galvo, and George Kapetanios (2019) "A comprehensive evaluation of macroeconomic forecasting methods," *International Journal of Forecasting*, Vol. 35, No. 4, pp. 1226 – 1239.

Correa, Ricardo, Keshav Garud, Juan M Londono, Nathan Mislang et al. (2017) "Constructing a Dictionary for Financial Stability,"Technical report, Board of Governors of the Federal Reserve System (US).

Coulombe, Philippe Goulet, Maxime Leroux, Dalibor Stevanovic, and Stphane Surprenant (2019) "How is Machine Learning Useful for Macroeconomic Forecasting?," CIRANO Working Papers 2019s-22, CIRANO.

——— (2020) "How is Machine Learning Useful for Macroeconomic Forecasting?."

De Mol, Christine, Domenico Giannone, and Lucrezia Reichlin (2008) "Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components?" *Journal of Econometrics*, Vol. 146, No. 2, pp. 318–328. Honoring the research contributions of Charles R. Nelson.

Diebold, Francis X. and Roberto S. Mariano (1995) "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, Vol. 13, No. 3, pp. 253–263.

Diebold, Francis X. and Minchul Shin (2019) "Machine learning for regularized survey forecast combination: Partially-egalitarian LASSO and its derivatives," *International Journal of Forecasting*, Vol. 35, No. 4, pp. 1679–1691.

Ferrara, Laurent and Anna Simoni (2019) "When are Google data useful to nowcast GDP? An approach via pre-selection and shrinkage,"Technical report.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy (2019) "Text as data," *Journal of Economic Literature*, Vol. 57, No. 3, pp. 535–74.

Giannone, Domenico, Lucrezia Reichlin, and Saverio Simonelli (2009) "NOWCASTING EURO AREA ECONOMIC ACTIVITY IN REAL TIME: THE ROLE OF CONFIDENCE INDICATORS," *National Institute Economic Review*, No. 210, pp. 90–97.

Giannone, D., L. Reichlin, and D. Small (2008a) "Nowcasting GDP and Inflation: The Real-Time Informational Content of Macroeconomic Data Releases," *Journal of Monetary Economics*, Vol. 55, pp. 665–676.

Giannone, Domenico, Lucrezia Reichlin, and David Small (2008b) "Nowcasting: The real-time informational content of macroeconomic data," *Journal of Monetary Economics*, Vol. 55, No. 4, pp. 665–676.

Gilbert, CJ Hutto Eric (2014) "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*.

Girardi, Alessandro, Andreas Reuter, and Christian Gayer (2014) "The role of survey data in nowcasting euro area GDP growth," European Economy - Economic Papers 2008 - 2015 538, Directorate General Economic and Financial Affairs (DG ECFIN), European Commission.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009) "Random forests," in *The elements of statistical learning*: Springer, pp. 587–604.

Hu, Minqing and Bing Liu (2004) "Mining and Summarizing Customer Reviews," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ?04, p. 168?177, New York, NY, USA: Association for Computing Machinery.

Hu, Guoning, Preeti Bhargava, Saul Fuhrmann, Sarah Ellinger, and Nemanja Spasojevic (2017) "Analyzing users' sentiment towards popular consumer industries and brands on Twitter."

Husted, Lucas F, John H Rogers, and Bo Sun (2017) "Monetary Policy Uncertainty."

Joseph, Andreas (2019) "Parametric inference with universal function approximators," Bank of England working papers 784, Bank of England.

Joseph, Andreas, Eleni Kalamara, Galina Potjagailo, and George Kapetanios (2021) "Forecasting UK inflation bottom up," *Unpublished Manuscript*.

Kalamara, Eleni, Arthur Turrell, Chris Redl, George Kapetanios, and Sujit Kapadia (2020) "Making text count: economic forecasting using newspaper text," *Bank of England Staff Working Papers*, Vol. 865.

Kapetanios, George and Fotis Papailias (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).

Kotchoni, Rachidi, Maxime Leroux, and Dalibor Stevanovic (2019) "Macroeconomic forecast accuracy in a data?rich environment," *Journal of Applied Econometrics*, Vol. 34, No. 7, pp. 1050–1072, November.

Lahiri, Kajal and George Monokroussos (2013) "Nowcasting US GDP: The role of ISM business surveys," *International Journal of Forecasting*, Vol. 29, No. 4, pp. 644–658.

Loughran, Tim and Bill McDonald (2013) "IPO first-day returns, offer price revisions, volatility, and form S-1 language," *Journal of Financial Economics*, Vol. 109, No. 2, pp. 307–326.

Marcellino, Massimiliano (2008) “A linear benchmark for forecasting GDP growth and inflation?,” *Journal of Forecasting*, Vol. 27, No. 4, pp. 305–340.

Medeiros, Marcelo C., Gabriel F. R. Vasconcelos, lvaro Veiga, and Eduardo Zilberman (2021) “Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods,” *Journal of Business & Economic Statistics*, Vol. 39, No. 1, pp. 98–119.

Nakamura, Emi (2005) “Inflation forecasting using a neural network,” *Economics Letters*, Vol. 86, No. 3, pp. 373–378.

Ng, Serena (2014) “Viewpoint: Boosting Recessions,” *Canadian Journal of Economics*, Vol. 47, No. 1, pp. 1–34.

Nielsen, Finn Årup (2011) “A new ANEW: Evaluation of a word list for sentiment analysis in microblogs,” *arXiv preprint arXiv:1103.2903*.

Nyman, Rickard, Sujit Kapadia, David Tuckett, David Gregory, Paul Ormerod, and Robert Smith (2018) “News and narratives in financial systems: exploiting big data for systemic risk assessment,” *Bank of England Staff Working Papers*, Vol. 704.

Smeekes, Stephan and Etienne Wijler (2018) “Macroeconomic forecasting using penalized regression methods,” *International Journal of Forecasting*, Vol. 34, No. 3, pp. 408–430.

Stock, James and M. Watson (1999) *A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series*, pp. 1–44, Oxford: Oxford University Press.

Stock, James and M.W. Watson (2002a) “Forecasting Using Principal Components from a Large Number of Predictors,” *Journal of the American Statistical Association*, Vol. 97, p. 1167–1179.

——— (2002b) “Macroeconomic Forecasting Using Diffusion Indexes,” *Journal of Business and Economic Statistics*, Vol. 20, No. 2, p. 147–162.

Stock, J.H. and M.W. Watson (2002c) “Forecasting using principal components from large numbers of predictors,” *Journal of the American Statistical Association*, Vol. 97(460), pp. 1167–1179.

Stock, James H. and Mark W. Watson (2012) “Generalized Shrinkage Methods for Forecasting Using Many Predictors,” *Journal of Business & Economic Statistics*, Vol. 30, No. 4, pp. 481–493.

Swanson, Norman R. and Halbert White (1997) “A Model Selection Approach to Real-Time Macroeconomic Forecasting Using Linear Models and Artificial Neural Networks,” *The Review of Economics and Statistics*, Vol. 79, No. 4, pp. 540–550.

Tersvirta, Timo, Marcelo Medeiros, and Gianluigi Rech (2006) “Building neural network models for time series: a statistical approach,” *Journal of Forecasting*, Vol. 25, No. 1, pp. 49–75.

Tetlock, P (2007) “Giving content to investor sentiment: the role of media in the stock market,” *Journal of Finance*, Vol. 62(3), pp. 1139–1168.

Vapnik, Vladimir (1998) "Statistical Learning Theory. John Wiley&Sons," *Inc., New York*.

Varian, Hal R. (2014) "Big Data: New Tricks for Econometrics," *Journal of Economic Perspectives*, Vol. 28, No. 2, pp. 3–28, May.

Wold, Hellman (1985) *Partial Least Squares*: Encyclopedia of statistical sciences.

Zhao, Qingyuan and Trevor Hastie (2021) "Causal Interpretations of Black-Box Models," *Journal of Business & Economic Statistics*, Vol. 39, No. 1, pp. 272–281.

Zou, H and T Hastie (2005) "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society*, Vol. Series B 101(476), pp. 301–320.

A Details of the Dataset

Table A.1: Panel of survey balances

Code	Variable Name	Source	Transf
1	CBI Distributive trades total-SALES	CBI	L
2	CBI Distributive trades total-ORDERS	CBI	L
3	CBI Distributive trades total-SALES FOR TIME OF YEAR	CBI	L
4	CBI Distributive trades total-STOCKS	CBI	L
5	CBI Distributive trades Retailing-SALES	CBI	L
6	CBI Distributive trades Retailing-ORDERS	CBI	L
7	CBI Distributive trades Retailing-SALES FOR TIME OF YEAR	CBI	L
8	CBI Distributive trades Retailing-STOCKS	CBI	L
9	CBI Distributive tradesWholesaling-SALES	CBI	L
10	CBI Distributive tradesWholesaling-ORDERS	CBI	L
11	CBI Distributive tradesWholesaling-SALES FOR TIME OF YEAR	CBI	L
12	CBI Distributive tradesWholesaling-STOCKS	CBI	L
13	CBI Distributive trades Motor Trades-SALES	CBI	L
14	CBI Distributive trades Motor Trades-ORDERS	CBI	L
15	CBI Distributive trades Motor Trades-SALES FOR TIME OF YEAR	CBI	L
16	CBI Distributive trades Motor Trades-STOCKS	CBI	L
17	CBI Monthly Trends-Export Orders	CBI	L
18	CBI Monthly Trends-FG Stocks	CBI	L
19	CBI Monthly Trends-Expected Output	CBI	L
20	CBI Monthly Trends-Average Prices	CBI	L
21	CBI Monthly Trends-Reported Output	CBI	L
22	CIPSManufacturing-PMI	HIS/MARKIT CIPS	L
23	CIPSManufacturing- new orders	HIS/MARKIT CIPS	L
24	CIPSManufacturing-New export orders	HIS/MARKIT CIPS	L
25	CIPSManufacturing-Output	HIS/MARKIT CIPS	L
26	CIPSManufacturing-Employment	HIS/MARKIT CIPS	L
27	CIPSManufacturing-Suppliers' deliver times	HIS/MARKIT CIPS	L
28	CIPSManufacturing-Stock of purchases	HIS/MARKIT CIPS	L
29	CIPSManufacturing-Input prices	HIS/MARKIT CIPS	L
30	CIPSManufacturing-Quantity of purchases	HIS/MARKIT CIPS	L
31	CIPSManufacturing-Stock of finished goods	HIS/MARKIT CIPS	L
32	CIPSManufacturing-Output prices	HIS/MARKIT CIPS	L
33	CIPSManufacturing-Work backlogs	HIS/MARKIT CIPS	L
34	CIPSConstruction-Total activity	HIS/MARKIT CIPS	L
35	CIPSConstruction-Commercial activity	HIS/MARKIT CIPS	L
36	CIPSConstruction-Civil engineering activity	HIS/MARKIT CIPS	L
37	CIPSConstruction-Suppliers delivery times	HIS/MARKIT CIPS	L
38	CIPSConstruction-Employment	HIS/MARKIT CIPS	L
39	CIPSConstruction-Future business activity	HIS/MARKIT CIPS	L
40	CIPSConstruction-Housing activity	HIS/MARKIT CIPS	L

Code	Variable Name	Source	Transf
41	CIPSConstruction-Input Prices	HIS/MARKIT CIPS	L
42	CIPSConstruction-New orders	HIS/MARKIT CIPS	L
43	CIPSConstruction-Quantity of purchases	HIS/MARKIT CIPS	L
44	CIPSServices-Business activity	HIS/MARKIT CIPS	L
45	CIPSServices-Incoming new business	HIS/MARKIT CIPS	L
46	CIPSServices-Outstanding business	HIS/MARKIT CIPS	L
47	CIPSServices-Employment	HIS/MARKIT CIPS	L
48	CIPSServices-Average prices charged	HIS/MARKIT CIPS	L
49	CIPSServices-Average input prices	HIS/MARKIT CIPS	L
50	CIPSServices-Business expectations	HIS/MARKIT CIPS	L
51	GfK CC-Financial situation of households over last 12 months	European Commission	L
52	GfK CC-Financial situation of households over next 12 months	European Commission	L
53	GfK CC-General economic situation over last 12 months	European Commission	L
54	GfK CC-General economic situation over next 12 months	European Commission	L
55	GfK CC-Price trends over last 12 months (LHS)	European Commission	L
56	GfK CC-Price trends over next 12 months	European Commission	L
57	GfK CC-Unemployment over next 12 months	European Commission	L
58	GfK CC-Major purchases at present	European Commission	L
59	GfK CC-Major purchases over next 12 months	European Commission	L
60	GfK CC-Savings at present	European Commission	L
61	GfK CC-Savings over next 12 months	European Commission	L
62	GfK CC-Current financial situation of households (saving a lot v running into debt)	European Commission	L
63	CBI SSS-BusPro-Present level of Business	CBI	L
64	CBI SSS-BusPro-Volume business past 3 months	CBI	L
65	CBI SSS-BusPro-Volume business next 3 months	CBI	L
66	CBI SSS-BusPro-Number of employed past 3 months	CBI	L
67	CBI SSS-BusPro-Number of employed next 3 months	CBI	L
68	CBI SSS-BusPro-Average selling prices next 3 months	CBI	L
69	CBI SSS-Consumer-Present level of Business	CBI	L
70	CBI SSS-Consumer-Volume business past 3 months	CBI	L
71	CBI SSS-Consumer-Volume business next 3 months	CBI	L
72	CBI SSS-Consumer-Number of employed past 3 months	CBI	L
73	CBI SSS-Consumer-Number of employed next 3 months	CBI	L
74	CBI SSS-Consumer-Average selling prices next 3 months	CBI	L
75	CBI SSS-Business-Present level of Business	CBI	L
76	CBI SSS-Business-Volume business past 3 months	CBI	L
77	CBI SSS-Business-Volume business next 3 months	CBI	L
78	CBI SSS-Business-Number of employed past 3 months	CBI	L
79	CBI SSS-Business-Number of employed next 3 months	CBI	L
80	CBI SSS-Business-Average selling prices next 3 months	CBI	L
81	CBI SSS-Professional-Present level of Business	CBI	L
82	CBI SSS-Professional-Volume business past 3 months	CBI	L
83	CBI SSS-Professional-Volume business next 3 months	CBI	L
84	CBI SSS-Professional-Number of employed past 3 months	CBI	L
85	CBI SSS-Professional-Number of employed next 3 months	CBI	L
86	CBI SSS-Professional-Average selling prices next 3 months	CBI	L
87	CBI SSS-Total Sector-Present level of Business	CBI	L
88	CBI SSS-Total Sector-Volume business past 3 months	CBI	L
89	CBI SSS-Total Sector-Volume business next 3 months	CBI	L
90	CBI SSS-Total Sector-Number of employed past 3 months	CBI	L
91	CBI SSS-Total Sector-Number of employed next 3 months	CBI	L
92	CBI SSS-Total Sector-Average selling prices next 3 months	CBI	L
93	NIESR-Monthly GDP	NIESR	LD
94	Lloyds Business Barometer-Overall Business Confidence	Lloyds Bank	L
95	Lloyds Business Barometer-Business Activity next year	Lloyds Bank	L
96	Lloyds Business Barometer-Economic Optimism	Lloyds Bank	L
97	Monthly GDP	ONS	LD

Note: Sources are the Office for National Statistics (ONS), the Bank of England database (BOE), IHS Markit/CIPS, the Confederation of British Industries (CBI), LLoyds Bank, the European Commission. Transformation codes: LDD = log double difference, LD = log difference, L = levels, D = first difference.

Table A.3: Dataset augmented with Macroeconomic Series

Code	Variable Name	Source	Transf
1	IoS: Services, Index	ONS	LD
2	PNDS: Private Non-Distribution Services: Index	ONS	LD
3	IoS: G: Wholesales, Retail and Motor Trade: Index	ONS	LD
4	IoS: 47: Retail trade except of motor vehicles and motorcycles: Index	ONS	LD
5	IoS: 46: Wholesale trade except of motor vehicles and motorcycles: Index	ONS	LD
6	IoS: 45: Wholesale And Retail Trade And Repair Of Motor Vehicles And Motorcycles: Index	ONS	LD
7	IoS: O-Q: PAD, Education and Health Index	ONS	LD
8	IoP:Production	ONS	LD
9	IoP:Manufacturing	ONS	LD
10	Energy output (utilities plus extraction) Pound Sterling (Index	ONS	LD
11	IoP: SIC07 Output Index D-E: Utilities: Electricity, Gas, Water Supply, Waste Management.	ONS	LD
12	IOP: B:MINING AND QUARRYING:	ONS	LD
13	RSI:VolumeAll Retailers inc fuel:All Business Index	ONS	LD
14	Construction Output: Seasonally Adjusted: Volume: All Work	ONS	LD
15	BOP Total Exports (Goods)	ONS	LD
16	BOP:EX:volume index:SA:Total Trade in Goods	ONS	LD
17	BOP Total Imports (Goods)	ONS	LD
18	BOP:IM:volume index:SA:Total Trade in Goods	ONS	LD
19	CPI all items	ONS	LDD
20	RPI all items	ONS	LDD
21	RPI ex Mortgages Interest Payments (RPIX)	ONS	LDD
22	PPI Output	ONS	LDD
23	PPI Input	ONS	LDD
24	Nationwide House Price MoM	BoE database	D
25	RICS House Price Balance	BoE database	D
26	M4 Money Supply	BoE database	LD
27	New Mortgage Approvals	BoE database	LD
28	Bank of England UK Mortgage Approvals	BoE database	LD
29	Average Weekly Earnings	ONS	LD
30	LFS Unemployment Rate	ONS	D
31	LFS Number of Employees (Total)	ONS	LD
32	Claimant Count Rate	ONS	D
33	New Cars Registrations	BoE database	LD
34	Oil Brent	BoE database	LD
35	UK mortgage base rate	BoE database	L
36	3m LIBOR	BoE database	L
37	FTSE all share	BoE database	LD
38	Sterling exchange rate index	BoE database	LD
39	FTSE volatility	BoE database	LD
40	GBP EUR spot	BoE database	LD
41	GBP USD spot	BoE database	LD
42	FTSE 250 INDEX	BoE database	LD
43	FTSE All Share	BoE database	LD
44	UK focused	BoE database	LD
45	S&P 500	BoE database	LD
46	Euro Stoxx	BoE database	LD
47	Sterling ERI	BoE database	LD
48	VIX	BoE database	LD
49	UK VIX - FTSE 100 VOLATILITY INDEX - PRICE INDEX	BoE database	LD

Note: Sources are the Office for National Statistics (ONS), the Bank of England database (BOE), IHS Markit/CIPS, the Confederation of British Industries (CBI), LLoyds Bank, the European Commission. Transformation codes: LDD = log double difference, LD = log difference, L = levels, D = first difference.

B Additional Empirical Results

Table B.1: Relative RMSFE with survey, text and macro data

relative to PCR	(1)	(3)	(6)	(9)	(12)	(24)
Lasso	0.964***	0.958	0.941	0.931	0.944	0.996
Ridge	0.702***	0.860	0.857	0.904	0.953	1.115
Elastic	0.940***	0.939	0.926	0.920	0.937	1.009
PLS	0.837***	0.841	0.827	0.844	0.869	1.045
Random Forest	0.722***	0.874	0.877	0.890	0.927	1.051
SVM	0.702***	0.866	0.856	0.884	0.923	1.111
NN	0.698	0.886	0.850	0.898	0.948	1.116
relative to AR(1)	**(1)**	**(3)**	**(6)**	**(9)**	**(12)**	**(24)**
Lasso	0.957***	0.955	0.940	0.931	0.945	0.997
Ridge	0.696***	0.858	0.856	0.904	0.953	1.115
Elastic	0.933***	0.937	0.925	0.920	0.938	1.010
PLS	0.831***	0.839	0.826	0.844	0.869	1.045
Random Forest	0.809***	0.884	0.879	0.891	0.923	1.078
SVM	0.697***	0.864	0.855	0.884	0.923	1.112
NN	0.779***	0.873	0.887	0.897	0.942	1.143***

Note: Top panel: Relative RMSFEs across different specifications using combined *Survey*, *Text* and *Macro* data against a **PCR** model. Mid panel: RSMFEs relative to an **AR(1)**. *,** & *** denote rejection at 10%, 5% and 1% level of the null hypothesis of equal forecasting method accuracy of the Diebold and Mariano (1995) test against the benchmark models

Table B.2: RMSFE Across Different Models with AR(1) benchmark

relative to PCR	(1)	(3)	(6)	(9)	(12)	(24)
Lasso	0.900***	0.913	0.896	0.898	0.918	1.033
Ridge	0.902***	0.914	0.897	0.898	0.918	1.033
Elastic	0.984	0.974	0.959	0.940	0.950	0.990
PLS	0.935***	0.931	0.924	0.916	0.932	1.012
Random Forest	0.930***	0.941	0.935	0.926	0.941	1.013
SVM	0.821***	0.881*	0.905	0.912	0.939	1.040
NN	0.870***	0.960	0.887	0.904	0.945	1.018

	(1)	(3)	(6)	(9)	(12)	(24)
Lasso	0.957***	0.955	0.940	0.931	0.945	0.997
Ridge	0.854***	0.883	0.862	0.882	0.896	1.065
Elastic	0.944***	0.946	0.932	0.925	0.941	1.003
PLS	0.883***	0.899	0.876	0.888	0.899	1.045
Random Forest	0.931***	0.946	0.920	0.922	0.939	1.020
SVM	0.720***	0.885	0.872	0.896	0.929	1.118
NN	0.858***	0.888	0.901	0.895	0.935	1.126

	(1)	(3)	(6)	(9)	(12)	(24)
Lasso	0.984***	0.974	0.959	0.940	0.950	0.990
Ridge	0.692***	0.828*	0.854	0.901	0.946	1.097
Elastic	0.798***	0.835**	0.843	0.873	0.914	1.062
PLS	0.811***	0.831*	0.839	0.868	0.909	1.058
Random Forest	0.787***	0.880	0.872	0.882	0.917	1.044
SVM	0.657*	0.871	0.864	0.895	0.936	1.125
NN	0.661*	0.849	0.877	0.903	0.952	1.148

Note: Top panel: Relative RMSFEs across different specifications using only *Text* data. Middle panel: Relative RMSFEs across different specifications using only *Macro* data. Bottom panel: Relative RMSFEs across different specifications combining *Survey* and *Text* data. All results are reported against an **AR(1)** model as a benchmark. *,** & *** denote rejection at 10%, 5% and 1% level of the null hypothesis of equal forecasting method accuracy of the Diebold and Mariano (1995) test against the benchmark specification.