Organisation for Economic Co-operation and Development

ECO/WKP(2020)1

Unclassified **English - Or. English**

10 January 2020

ECONOMICS DEPARTMENT

ADAPTIVE TREES: A NEW APPROACH TO ECONOMIC FORECASTING

ECONOMICS DEPARTMENT WORKING PAPERS No. 1593

By Nicolas Woloszko

All Economics Department Working Papers are available at www.oecd.org/eco/workingpapers

JT03456542

ABSTRACT/RÉSUMÉ

Adaptive Trees: a new approach to economic forecasting
The present paper develops Adaptive Trees, a new machine learning approach specifically designed for economic forecasting. Economic forecasting is made difficult by economic complexity, which implies non-linearities (multiple interactions and discontinuities) and unknown structural changes (the continuous change in the distribution of economic variables). The forecast methodology aims at addressing these challenges. The algorithm is said to be "adaptive" insofar as it adapts to the quantity of structural change it detects in the economy by giving more weight to more recent observations. The performance of the algorithm in forecasting GDP growth 3- to 12-months ahead is assessed through simulations in pseudo-real-time for six major economies (USA, UK, Germany, France, Japan, Italy). The performance of Adaptive Trees is on average broadly similar to forecasts obtained from the OECD's Indicator Model and generally performs better than a simple AR(1) benchmark model as well as Random Forests and Gradient Boosted Trees.
JEL codes: C01, C18, C23, C45, C53, C63, E37.

Keywords: forecasting, machine learning, interpretable AI, concept drift, structural change, GDP growth, business cycles, short-term forecasts, feature engineering.

« Adaptive Trees » : une nouvelle méthode de prévision économique
Cet article introduit les « Adaptive Trees », une nouvelle méthode de machine learning spécifiquement adaptée à la prévision économique. La prévision économique est difficile en raison de la complexité de l'économie, qui recouvre des non-linéarités (interactions multiples et discontinuités) ainsi que le changement structurel (le changement dans la distribution des variables au cours du temps). La présente méthodologie de prévision vise à répondre à ces problématiques. L'algorithme proposé est dit « adaptif » dans la mesure où il s'adapte à la quantité de changement structurel détectée dans l'économie en donnant plus de poids aux observations les plus récentes. La performance de l'algorithme pour la prévision de la croissance du PIB de 3 à 12 mois à l'avance est évaluée par des simulations en pseudo-temps réel pour six grands pays (les États-Unis, le Royaume Uni, l'Allemagne, la France, le Japon et l'Italie). La performance des « Adaptive Trees » est en moyenne peu ou prou similaire à celle du Modèle d'Indicateurs de l'OCDE, et meilleure qu'un modèle auto-régressif d'ordre 1, qu'une « Random Forest » et qu'un « Gradient Boosted Trees ».
Codes JEL : C01, C18, C23, C45, C53, C63, E37.

Mots clés : prévision, apprentissage statistique, interprétabilité, changement structural, croissance du PIB, cycle des affaires, prévision de court-terme, *feature engineering.*

Table of contents

Tables

Figures

Boxes

Adaptive Trees, a new approach to economic forecasting

By Nicolas Woloszko[1]

1. Introduction and main findings

1. Machine learning was born in the 1960s, as a set of techniques designed to extract information from data. It gained wider currency around the early 2000s thanks to the advent of Big Data and improvements in computer processing. Big Data provided both the data and computational power to experiment with more sophisticated algorithms[2]. Since then, machine learning has become ubiquitous in industry and is at the core of the artificial intelligence revolution.

2. The focus of this paper is on predictions made with Adaptive Trees and how they can be explained. Adaptive Trees is a new algorithm based on regression trees that addresses non-linearities and structural change in the macroeconomic data. A prediction made with Adaptive Trees can be additively decomposed into variable contributions, thus ensuring model interpretability. In doing so, the paper contributes to the growing literature on the application of machine learning to macroeconomic forecasting (Jung, Patnam and Ter-Martirosyan, 2018[1]; Chakraborty and Joseph, 2017[2]; Gogas et al., 2015[3]).

3. Macroeconomic forecasting is a challenging task and existing techniques have some limitations. Moreover, failing to anticipate the 2008 crisis has called for a renewal of the forecasting methods (Romer, 2016[4]; Blanchard, 2014[5]). The paper contends that machine learning techniques can be well suited to address non-linearities, a challenge to macroeconomic modelling that is particularly conspicuous around crises.

4. There have been a series of attempts to apply machine learning to macroeconomic forecasting (Biau and D'Elia, 2009[6]; Chakraborty and Joseph, 2017[2]; Tiffin, 2016[7]). Most experiments have applied off-the-shelf machine learning algorithms to economic data and obtained reasonable results. The algorithmic approach in this paper is tailored to address the specific challenges of macroeconomic forecasting.

5. Macroeconomic forecasting has long relied heavily on econometric estimation. This raises the question of the resemblance and differences between machine learning and econometrics (see Box 1). Both disciplines share the purpose of learning from data. Machine learning differs from econometrics as it does not require prior domain knowledge (Breiman, 2001[8]) in terms of model or statistical assumptions. Machine learning is a field of statistics and computer sciences that provides methods to deal with large information sets and allows general forms of non-linearities and interactions between variables. As distinct from Bayesian econometrics, machine learning does not rely on probabilistic

[1] This paper was produced with the support of the OECD Innovation LAB. The author wishes to thank Catherine L. Mann for her trust, Nicolas Ruiz, Orsetta Causa, Mohammed Benlaldj, Dorothée Rouzet, Sebastian Barnes for their strong support and sharp comments, as well as Alain de Serres and Luiz de Mello for their useful comments and for their support of this research projects. The author also thanks David Turner and Pierre-Alain Pionnier for their thorough reviews.

[2] See glossary in Annex D

beliefs about the data generating processes and performs model selection on the sole basis of out-of-sample goodness-of-fit.

6. Linear econometrics may be particularly challenged where economic complexity, which may play an important role in macroeconomics (Kirman, 2010[9]), is concerned. Complexity implies among other things non-linearities in macroeconomic behaviour. Non-linear relations can be specified even in a linear model, including by using polynomials or simple interaction terms. However, such specific interactions in a model may fail to capture multiple interactions and multiple discontinuities. Complexity may also imply structural changes, as the economy is an ever-changing complex system where the probability distributions may change over time. Standard econometric models suppose stable relations and make the hypothesis that the distribution of data remains the same across history, as long as structural breaks are not explicitly specified. For instance, it is well documented that the Philipps Curve changed in nature around the 1990s thanks to new frameworks for monetary policy that tamed inflation and inflation expectations. Complexity may also imply the context-dependence of economic relationships.

7. The flexibility of machine learning and richness of non-linear and time-varying processes that can be modelled make it potentially well-suited to capture complex economic relationships. Multiple interactions, discontinuities and structural breaks are particularly conspicuous around turning points and recessions. A telling example is housing bubbles. Growing house prices may signal strong GDP growth up until a given threshold, beyond which the bubble bursts and the economy may decelerate. Complexity is related to the emergence of crises, and that is why non-linearities and structural breaks may be considered to be "where the danger lurks" (Blanchard, 2014[5]).

8. The more complex a machine learning algorithm, the more accurately it can fit a complex reality. There is a potential trade-off between accuracy and interpretability in machine learning, often referred to as Occam's dilemma. More complex and accurate models may be harder to interpret, making them challenging for users and providing less confidence in a forecasting context that models are capturing meaningful relationships.

9. There is a growing literature on "interpretable machine learning" (Lipton, 2016[10]), and a number of methods aim at proving interpretability *ex post* to trained models (Lundberg and Lee, 2017[11]; Renard et al., 2019[12]). Among these, Tree Interpreter (Saabas, 2014[13]) has been developed to shed light on the predictions made by tree-based models and is compatible with the method developed in this paper.

10. This paper develops a new forecasting method specifically tailored to deal with non-linearities, structural changes and the detection of tipping points. The Adaptive Trees algorithm is based upon Gradient Boosted Trees, a widely used machine learning non-linear algorithm. Adaptive Trees results from an incremental modification of its functioning that aims at addressing structural changes. The algorithm is said to be "adaptive" insofar as it adapts to the quantity of structural changes it detects in the economy by giving more weight to more recent observations. It also relies on feature[3] engineering in order to enhance tipping point detection.

11. The forecasting algorithm is assessed through a "horse race" that benchmarks its forecasts against alternative standard econometric forecasts. Performance is compared using pseudo-real time simulations of the forecasting of GDP growth in G6 countries (US, UK, France, Germany, Japan and Italy). The information sets comprise the same set of

[3] See glossary in Annex D

variables used for the OECD Indicator Models, a series of forecast models run by the OECD Economics Department (Turner, 2016[14]; Ollivaud et al., 2016[15]; Pain et al., 2014[16]). The simulations provide an assessment of forecast accuracy benchmarked against the OECD's Indicator model and a baseline AR(1) model.

12. The paper first introduces the data, then presents the forecast methodology, and reports the results it achieves in forecast simulations.

13. The main findings are:

- Adaptive Trees-based forecasts can be compared with benchmark models using the exact same set of variables. At M+3 and M+6, the Adaptive Trees algorithm performs better for the United Kingdom, equally for France, Japan and the United States, and not as well for Italy and Germany compared to the OECD Indicator Model. It generally performs better than an AR(1) model.
- Adaptive Trees consistently outperforms the Random Forest and Gradient Boosted Trees in the pseudo-real time simulations.
- Adaptive Trees forecast perform well in the short run, but become uninformative past 12 months. No long-term forecast simulations were available for the Indicator Model.
- The Adaptive Trees machine learning algorithm can handle data in high dimensions, that is a large number of variables compared to the number of observations. This capability has not been exploited in the like-for-like comparison used in this paper based on few variables. However, future research experimenting with larger data sets could increase the forecasting performance and advantage of the Adaptive Trees algorithm.

Box 1. Supervised machine learning methods focus on maximising out-of-sample error by striking a optimal bias-variance trade-off

14. The objective of supervised machine learning algorithms is to maximise the out-of-sample predictive accuracy for a target variable, for example quarterly GDP. By contrast, econometrics is more concerned with unbiased parameter inference. Machine learning includes estimators that achieve better generalisability at the expense of some amount of bias.

15. While aiming at minimising out-of-sample error, usually measured using a loss function (such as mean square error), machine learning relies heavily on numerical optimization techniques (including gradient descent, an iterative algorithm using the gradient of the loss function to incrementally converge towards an optimum). Machine learning relies on techniques meant to prevent overfitting and adjust the level of model complexity in order to maximise out-of-sample goodness-of-fit. Such techniques include cross-validation[4].

16. As machine learning focuses on predictive performance and out-of-sample goodness-of-fit, the bias-variance trade-off is a major issue. The out-of-sample mean square error can be decomposed into bias, variance and noise:

$$E\left[\left(y - \hat{f}(x)\right)^2\right] = Bias\left(\hat{f}(x)\right)^2 + Var\left[\hat{f}(x)\right] + \sigma^2$$

Where $y = f(x) + \sigma$ is a noisy set of observations, and $\hat{f}$ is an estimator of f. The bias term is equal to :

$$Bias\left(\hat{f}\right) = E[\hat{f} - f]$$

And the variance term is:

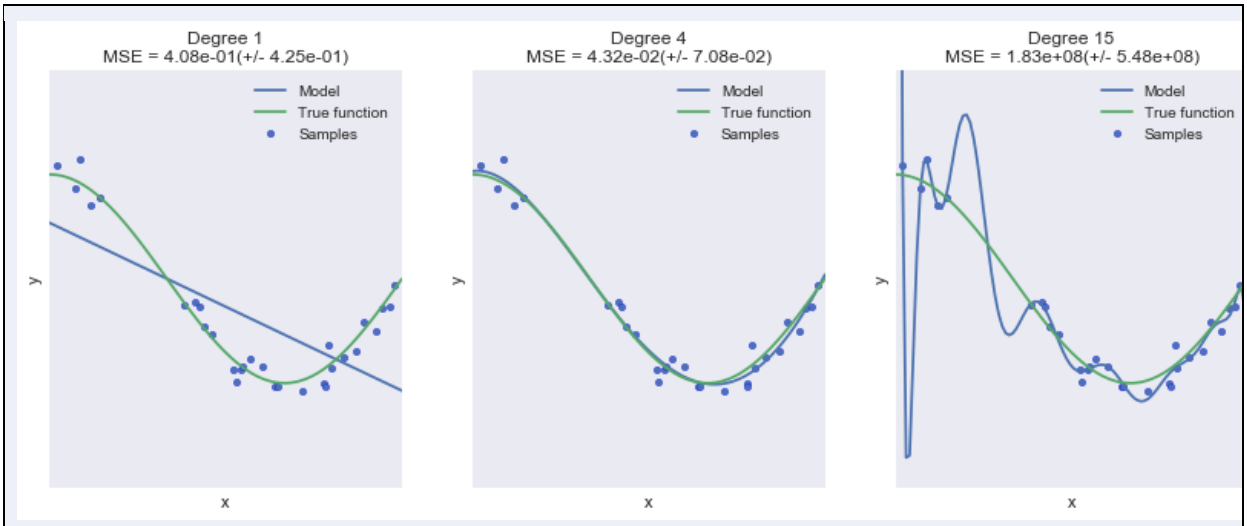
$$Var\left[\hat{f}\right] = E[\left(\hat{f} - E(\hat{f})\right)^2]$$

The last term $\sigma^2$is the variance of the observation noise and is the irreducible part of the error.

17. There is a trade-off between bias and variance as a very simple model will have low variance and high bias (underfitting), whereas very complex models may have a high variance and a low bias (overfitting). In Figure 1, a sinusoid is fitted with a polynomial. In the left pane, a degree 1 polynomial underfits the data. In the right pane, a degree 15 polynomial overfits the data. In the centre panel, the degree 4 polynomial provides a good fit and low mean square error.

Figure 1. Underfitting, right fit, overfitting

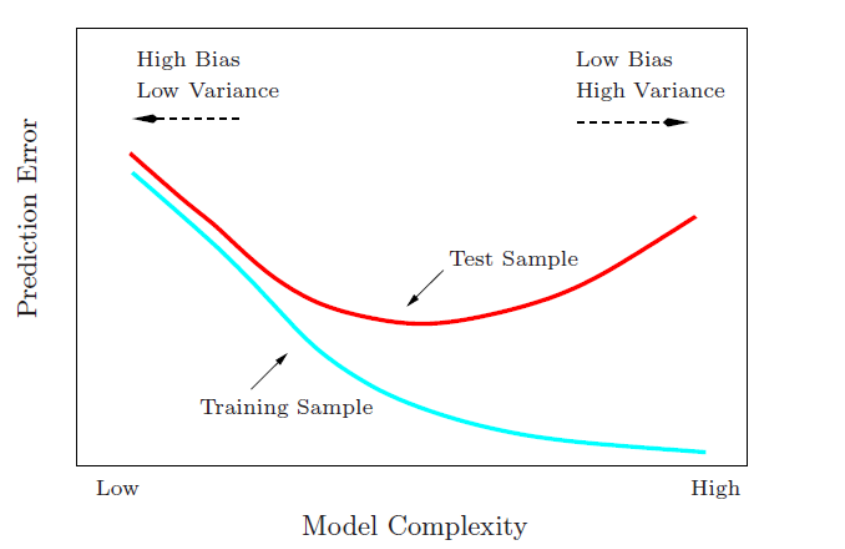
[4] See glossary in Annex D



Source: Pedregosa et al., 2011

18. As shown in Figure 2, the more complex a model, the better the in-sample goodness-of-fit. Past a certain threshold, however, out-of-sample goodness-of-fit starts decreasing due to overfitting. In order to find the right degree of model complexity and minimise out-of-sample error, cross-validation can be used.

Figure 2. Bias-variance trade-off



Source: J. Friedman et al., 2001

19. Cross validation (Efron, 1983[17]; Schneider, 1997[18]) is a model evaluation method that is often used in predictive settings. Some of the data is removed before the training[5] begins. When training is done, the data that was removed can be used to test the performance of the learned model on “new" data. Cross-validation is an out-of-sample goodness-of-fit evaluation method that can be used to choose the most appropriate model.

2. Data sources

20. This paper deals with forecasting GDP growth in all G7 countries but Canada.[6] For each country, the data used is the same set of leading indicator variables (Table 1) as used for the OECD Indicator Model (Ollivaud et al., 2016[15]; Sédillot and Pain, 2003[19]). This is to facilitate comparisons and to see whether the Adaptive Tree-based forecasts can outperform the leading indicator model even when applied in a constrained manner.

Table 1. Variables used by the Indicator Model

Italy	USA	France
Industrial production	Industrial Production	Industrial production
Car registrations	Consumption	Household consumption
PMI (manufacturing)	Employment	Output trend
Houshold confidence	Construction	Business survey
PMI (services)	Inventories	Order book and demand
	Exports	Household confidence
	PMI	
	Housing permits	
	Housing prices	

UK	Germany	Japan
Industrial production	Industrial production	Industrial production
Retail sales	Business surveys expectations	Inventory ratio
Housing prices	Exports	Living expenditure
Business confidence	Manufacturing orders	Job offers to applicants ratio
Economic sentiment indicator	Business survey	Small business sentiment sales
PMI	PMI (manufacturing)	Business sentiment financial position
	PMI (services)	Tankan
	Consumer confidence	PMI
	Vacancies	

21. All base variables are monthly series, whereas GDP growth is quarterly. In each country, the target variable is the quarter-on-quarter growth (Q/Q) of GDP in volume. All variables (including GDP) come with release delays that were carefully taken into account during the simulations[7].

22. The OECD **Error! Bookmark not defined.**Indicator Model (see Annex 1) is a series of short-term forecasts for major economies based on leading indicators and using

[5] See glossary in Annex D

[6] Canada is not included as GDP estimates are released on a monthly rather than a quarterly basis (Mourougane, 2006[57]).

[7] As in (Ollivaud et al., 2016[15]) in the case of bridge equations, staggered indicator releases are dealt with by re-aligning monthly indicators before estimation of model. Before estimation, those monthly indicators which are not available for a given month are shifted forward until all indicators are aligned.

Bayesian VAR methods. It builds on the work of (Turner, 2016[14]; Pain et al., 2014[16]; Lewis and Pain, 2015[20]) in using short term economic indicators to predict quarterly movements in GDP by exploiting all available monthly and quarterly information. These models typically combine information from both "soft" indicators, such as business tendency and consumer surveys, and "hard" indicators, such as industrial production, retail sales and house prices, and use is made of different frequencies of data. The Indicator Model forecast is a simple average across forecasts from a pure hard indicator model, a survey indicator model and a mixed indicator model. The Adaptive Tree forecast is made using the exact same data as the Indicator Model.

23. The forecast simulations from the Adaptive Trees and benchmark models are made in pseudo-real time, using the latest vintage of estimates from statistical agencies rather than data that was available in real-time.

3. Method

24. The algorithmic approach used in this paper ("Adaptive Trees") draws on recent machine learning techniques and an original contribution to the field. The methodology is tailored for macroeconomic forecasting, and aims at tackling three main challenges in economic forecasting: non-linearities (1), structural change (2), and tipping points (3). This section provides the main intuitions of the Adaptive Trees framework. A detailed description is provided in Annex A.

3.1. A tree-based approach to tackle non-linearities

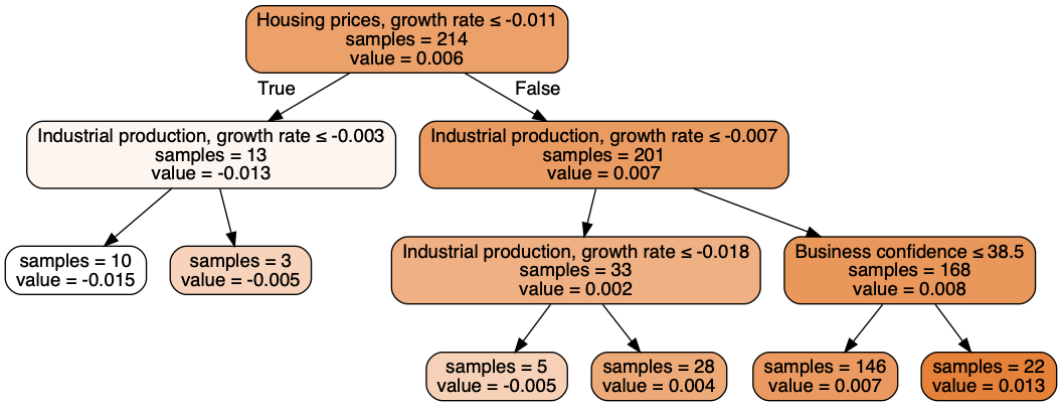
25. The machine learning literature includes a number of methods designed to perform non-linear modelling. The core estimator used in the Adaptive Trees framework is XGBoost (Chen and Guestrin, 2016[21]), an efficient implementation of the Gradient Boosting Trees algorithm. Gradient Boosting Trees build upon the widely-used Regression Tree algorithm, whose functioning is presented below. This sub-section provides insights about the three concepts (Regression Trees, Gradient Boosting Trees, and XGBoost) and how the approach is relevant to macroeconomic modelling.

26. Regressions Trees are appealing in the context of macroeconomic modelling, where non-linearities may occur. A given variable's effect may depend upon its own value (threshold effects) or upon the value of a set of other variables (interactions). For instance, house price increases growth may signal wealth effects up until a certain threshold, past which they only reflect a housing bubble. The economy is characterized by complex patterns (if… then…) that linear models cannot capture unless these patterns are specified. Regression trees can be used to uncover such patterns and help better specify linear models, introducing *ad hoc* interactions or threshold effects.

27. Regression trees predict the value of a target variable by learning simple "if-then" decision rules from the data. Regression trees recursively divide the sample of observations into sub-groups to minimise the within-group variance of the predicted variable, say the growth of GDP. Figure 3 provides an example of a regression tree trained on the UK dataset. At first, the algorithm selects the splitting variable (house prices growth rate) and the splitting point (-1.1%) that minimises the variance of the target variable (GDP growth) within the two resulting sub-groups based on considering all possible variables and splits. It repeats this procedure at each node until reaching final nodes (called "leaves").

Figure 3. A single regression tree

Single regression tree trained on UK data for an M-0 GDP growth forecast



Note: In each node, "value" indicates the average value of GDP growth within the node, and "samples" the number of observations falling in the node. The first node "contains" the full sample, and the full-sample mean is 0.6%. Each node contains a condition, and make a split between observations that validate the condition (left) and observations that do not (right). At first, the algorithm splits observations between quarters where the growth of house prices is inferior to -1.1 % and observations where it is not. For the former (left), the algorithms picks the growth rate of industrial production to make a split. For observations where house price growth is lower than -1.1% and the growth of industrial production less than -0.3%, the algorithm predicts that GDP growth will be equal to -1.5%.
Source: OECD Economic Outlook databases, and OECD calculations.

28. A prediction is made following a path in the tree and computing the average target value of the past observations that fall in the same leaf. Intuitively, regression trees may capture multiple interactions and threshold effects. Whereas predictions made with linear regressions are a weighted mean of the covariates (weighted by the regression coefficients), a regression tree introduces a logical structure (if…, and if…, then…). This structure may capture complex non-linear patterns that are to be found in the economy.

29. A regression tree can be arbitrarily deep, in terms of the number of nodes on a path in the tree. The depth of the tree in Figure 3 is equal to 3 as there are three series of splits. The deeper, the more splitting variables and splitting points come into play, the less observations per leaf. Deeper trees are more likely to overfit. In Figure 3, the prediction made in the left-most case is based upon only 10 observations. In turn, too shallow trees are likely to miss patterns in the data. The tree depth is a parameter that can be optimised using cross-validation.

30. Regression trees are an appealing albeit weak approach to non-linear regression, as they have a tendency to overfit. "Boosting" techniques can overcome this tendency by building an "ensemble" of trees, i.e., an array of regression trees whose predictions are averaged. The series of trees is built iteratively. A first tree is trained on the sample. Then each subsequent tree is trained on the residual from the predictions made by the average of all previous trees. Up to a few thousand trees can be added to the ensemble. It follows that observations that are difficult to predict and that yield large prediction errors receive an ever-increasing 'attention' by the model. An earlier version of the algorithm called AdaBoost (Freund and Schapire, 1997[22]) iteratively gives more weight to observations harder to predict.

31. More recent Boosting algorithms generalise on this intuition and use Gradient Boosting (see Box 2). XGBoost is an implementation of Gradient Boosed Trees (GBT).

XGBoost is an optimized distributed gradient boosting library. It has gained widespread currency in the machine learning community and has emerged as the main challenger to deep learning approaches when it comes to structured data (see Annex A for more detail).

32. An interesting feature of this implementation is that XGBoost can be trained on a dataset that includes missing values (on the X side). When the data are sparse, an instance is classified in the optimal default direction. At every tree node, there are two possible direction: left or right. The optimal default direction is learnt from the data. This characteristic is particularly relevant to macroeconomic forecasting as the variables often have different historical depth. Using XGBoost thus allows to include a larger training set. For instance, among variables used to predict GDP growth in the UK, the Purchasing Manager Index starts in 1992. All data before that date would have to be done away with if it were not for the use of XGBoost[8].

33. One of the reasons for choosing tree-based approaches is their intrinsic interpretability. Tree-based methods provide specific interpretability tools (Saabas, $2014_{[13]}$). Even though ensemble methods such as Gradient Boosted Trees involve a large number of trees, one may compute each variable's contribution to a given prediction. Determining each variable's contribution is done by computing the average of the target variable (*e.g.*, the GDP growth) in each intermediary node along the prediction path. For instance, the predictions made for observations falling in the left-most leaf node in Figure 3 can be decomposed as follows. At the origin of the tree, the prediction is equal to the population mean, 0.6%. At the next step, the house prices shunt towards the sub-region where the target mean is equal to -1.3%. The contribution of house prices so far is thus $-1.3\% - 0.6\% = -1.9\%$. And so on. In the end, the final prediction in the final leaf is the sum of the contributions of all the variables that intervened along the prediction path plus the population average:

$$\hat{Y} = 0.6\% + \sum Feature\ Contributions$$

[8] The possibility to use datasets containing missing values directly (without the need for prior imputation) is a distinctive feature of certain machine learning algorithms, as distinct from linear regressions that require complete datasets (and to drop observations with missing values or impute them). Subsequently, Adaptive Trees can take advantage of larger datasets and a deeper history in the data.

Box 2. Gradient Boosted Trees (GBT)

34. The Gradient Boosted Trees algorithm[9] (Friedman, 2002[23]) uses boosting, an iterative procedure. Boosting is an ensemble of regressions trees, i.e., a series of simple trees whose predictions are averaged.

35. Gradient Boosted Regression Trees consider additive models of the following form:

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x)$$

36. Where $h_m(x)$ are the simple regression trees (weak predictors). Gradient Boosted Trees build the additive model in a forward stagewise fashion:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

$h_m(x)$ is a regression tree grown in order to minimize a given loss function, usually least square error, and γ_m is calculated using numerical optimization.

37. Two additional features: shrinkage (J. H. Friedman, 2001) and subsampling (J. H. Friedman, 2002).

38. Shrinkage is a simple regularization strategy that scales the contribution of each weak learner by a factor ν:

$$F_m(x) = F_{m-1}(x) + \nu\gamma_m h_m(x)$$

39. The parameter ν is also called the learning rate. It helps reduce the risk of overfitting by reducing the impact of each extra weak predictor. When ν is very low, it takes more time and more predictors to reach sufficient accuracy. When ν is high, the risk of overfitting becomes more important. In the Adaptive Trees framework, ν is optimised by cross-validation.

40. Subsampling consists in selecting only a random subsample of available observations when growing each weak predictor $h_m(x)$. Gradient Boosting combined with subsampling becomes Stochastic Gradient Boosting (J. H. Friedman, 2002). At each iteration, a given fraction η of all available data is drawn at random. This randomly selected subsample is used instead of the full sample to fit the weak learner. Introducing some randomness has proved to improve the overall quality of the algorithm. η is another parameter to be set by the user that can be optimised using cross-validation.

3.2. From regression trees to adaptive trees: dealing with structural change

41. The economy may be seen as a complex ever-changing system. Standard models that are trained on large training window suppose that the economy abides by stable rules. However, structural breaks may occur (often around crises) and change the nature of the relations between the covariates and the target variable. Standard models, whose

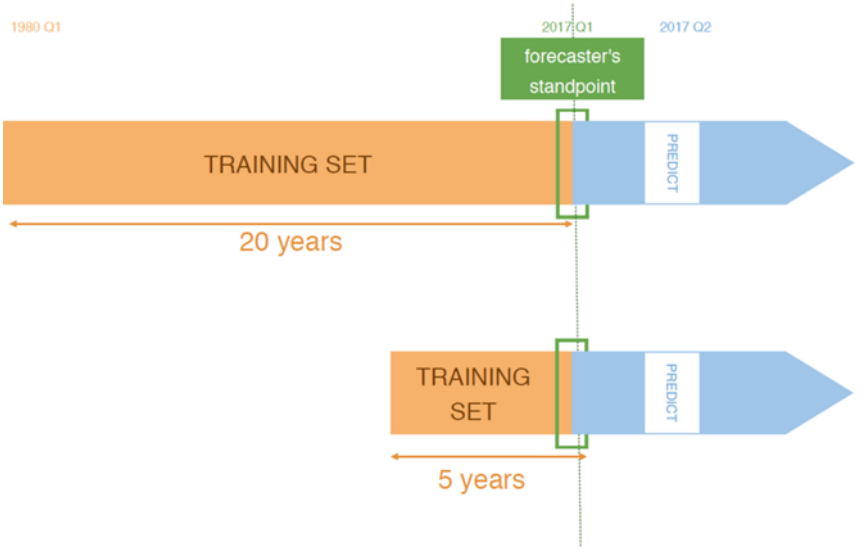
[9] This box relies on (Friedman, Hastie and Tibshirani, 2001[55]), where the reader will find a more complete introduction to the Gradient Boosting Trees algorithm.

coefficients are supposed to endure for extended periods, may fail to capture structural changes (unless the latter are explicitly specified). There can be sudden structural breaks or long-standing structural change, resulting for instance from technological changes.

42. Structural change is a problem known as "concept drift"[10] in the machine learning literature (see (Žliobaitė, 2010[24]) for a review). Concept drift refers to the idea that the distribution of data (the distribution of the target Y, the distribution of the variables X, and the joint distribution (X,Y)) may change over time, be it suddenly, incrementally, or through "reoccurring contexts".

43. Because of structural change, using a small training window may yield more accurate predictions than with a larger training window, as illustrated in Figure 5. Concept drift implies that the most recent past can be more informative about the near future than a more distant past.

Figure 4. Choosing the size of the training sample



Source: OECD calculations.

44. However, a forecast that would only use very recent history as a guide might be short-sighted. Recognizing a pattern from a distant past may improve accuracy in the case where such pattern would reoccur. For instance, one may want the forecast algorithm to detect a housing bubble even if one had only previously occurred in the series 20 years earlier.

45. The "Adaptive Trees" algorithm developed in this paper aims to optimise over this trade-off by evaluating what combination of past and recent data provides the best forecast at each point in time. Adaptive Trees are an extension of Gradient Boosted Trees. Adaptive Trees build on the fact that Gradient Boosted Trees give more importance to observation harder to predict in order to obtain an adaptive behaviour. The adaptive behaviour results from the use of increasing *ex ante* observation weights during the training of the Gradient Boosted Trees. Instead of initialising the observation weights to be equal to each other, as

[10] In the machine learning parlance, the word "concept" refers to the mechanism acting as the data generating process (in this case, the economy).

it is usually the case with GBT, more recent observations receive greater initial weights. Adaptive Trees thus adjust to structural change by giving more weight to the recent past when the more remote past becomes less informative about the future, as structural change makes latest observations harder to predict from the more remote past.

46. Let $w(t)$ be the weight of the observation at time $t, t \in [1, N]$, N being the number of observations. The Adaptive Trees methodology defines w as follows:

$$w(t) = e^{-\gamma(1-\frac{t}{N})}$$

47. The parameter γ is optimised using cross-validation and is usually close to 15. That means that in a sample of 10 years (120 months), the last observation $(t = N)$ will have weight 1, the last but one $(t = N - 1)$ will have weight 0.88, at the middle of the sample $(t = N/2)$, the weight is 0.0005 and at $t = 0$ the weight is 3e-7. In the forecast simulations, cross-validation is done once per country/forecast horizon at the start.

48. When structural breaks arise, the newest observations, that are more heavily weighted, will receive ever-increasing weights along the boosting steps given that they will be inaccurately predicted. The adaptive algorithm will thus give a higher importance to newest observations as soon as their distribution differs, thus signalling concept drift. Adaptive Trees will thus rely more heavily on the recent past when structural change is at play, while also exploiting information from a more remote past. More details on the methodology and insights on the adaptive property are provided in Annex A.

3.3. Feature engineering and feature selection to better detect tipping points

49. Feature engineering has become standard practice in the machine learning community when dealing with time series (Christ et al., 2018[25]). A feature is an attribute of a time-series variable over a given time window, such as its min, max, standard deviation, a lag, and so on. Feature engineering consists in extracting features from time series variables. In the Adaptive Trees framework, ten features are extracted from each variables for eight time windows. The features are described in Table 5. Each feature is extracted over the 3, 6, 9, 12, 18, 24, and 36 past months. Should the original data include six variables, 6 * 7 * 10 = 420 features would be added to the training data. The large number of resulting features calls for the subsequent use of feature selection.

Table 2. Feature engineering

Features extracted from time series variables.

Feature	Definition
Moving average	Average in the time window
Cumulated growth	$\prod_t(1 + x_t) - 1$ where x_t is a growth rate (only for variables expressed as a growth rate)
Volatility	Standard deviation over the time window
Change over the period	Last value minus first value
Spread Min Max	Max value minus min value over the time window
Max	Max over period
Min	Min over period
Mean second derivative	Average of the twice-differenced series over the time window
Mean absolute change	$\frac{1}{n}\sum_{t=1}^{n}\lvert x_t - x_{t-1}\rvert$
Trend projection	Prediction by a linear regression of y on the time vector at a M+6 horizon

Source: OECD.

50. Feature engineering may enhance tipping point detection. A literature on complex systems highlights that certain statistical properties of time series variables may be leading indicators of critical transitions (see (Dakos et al., 2012[26]) for a review). For instance, strong increase in volatility, or large overall increases or decreases in a variable over a given period of time may signal medium-run dynamics better than the current or lagged value of the variable (Carpenter and Brock, 2006[27]). Features such as the second derivative or the spread between min and max over a given time window may provide information of the instability of a system and enhance the detection of turning points.

51. The resulting high number of variables calls for the use of feature selection. Feature selection reduces the feature space by removing features that have low predictive power or high noise and may thus increase predictive accuracy (Guyon and Elisseeff, 2003[28]; Chandrashekar and Sahin, 2014[29]; Cai et al., 2018[30]). There is a large number of feature selection methods in the literature. The Adaptive Trees framework relies on model-based feature selection, as it uses the feature importance scores produced by the XGBoost predictor. In a single tree model, feature importance corresponds to the number of splits that use this feature and their height in the tree. The higher a feature intervenes in a tree, the more important it is. In an ensemble of trees such as XGBoost, the feature importance score averages the feature importances computed for each tree. All features with null importance are removed. The resulting training set is then used for the training and prediction. Annex A provides more detail on the feature selection scheme.

4. Results

52. The forecast performance of the Adaptive Trees model and IM and AR(1) benchmarks is assessed using a pseudo real-time approach. For each quarter T between 2007Q1 and 2017Q1, the models are trained and forecasts are derived. This ensures that no information is used to make projections which is dated after the time the forecast is made. The Adaptive Trees parameters are optimised using gridsearch at the onset of each

simulation. The simulations are *pseudo*-real time insofar as current rather than historical vintages are used.

53. Forecast simulations are made with data available at the 15th day of the month according to Datastream release dates. Release delays are taken into account by re-aligning the data. An M-3 forecast simulation will thus mimick the conditions of, say, the 15th of January 2007 to forecast the GDP growth released in April the same year.

54. Forecast performance is evaluated based on the root mean squared error (RMSE) of the different approaches over the period 2007Q1 to 2017Q1. The following paragraphs present the forecast results and performance for all G6 countries, at five time horizons: three months before the quarterly GDP release date (M+3), six (M+6), nine (M+9), twelve (M+12) and twenty-four (M+24). Forecast's RMSE are provided as a ratio of a baseline AR(1)'s RMSE. It should be noted that the data sample covers a relatively short period of time so this evaluation is based on a small number of observations and covers only a few macroeconomic episodes. The results may therefore depend to a degree on the specific nature of this sample period.

4.1. Forecast simulations

55. Table 3 displays summary statistics on forecast performance for the Adaptive Trees, the Indicator Model as well as a Random Forest and a Gradient Boosted Tree. Short-run forecasts made with Adaptive Trees are comparable to the results of the Indicator Model with broadly similar performance in most cases. At terms larger than M+6, the Adaptive Trees forecasts perform similarly to a simple AR(1) model. Forecasts at longer time horizons than 2 quarters are barely informative, thus concurring the findings of Breitung and Knüppel (2018[31]). At M+3 and M+6, the Adaptive Trees outperform the Indicator Models for the United Kingdom, yields comparable performances for France, Japan and the United States, and compares poorly in the cases of Germany and Italy. The Adaptive Trees consistently outperforms both the Random Forest and Gradient Boosted Trees by a large margin.

56. The off-the-shelf machine learning methods (Random Forest and Gradient Boosted Trees) underperform the AR(1) in most cases. This result suggests that standard machine learning approaches do not always give satisfactory results. This may be due to the numerous problem-specific issues of macroeconomic forecasting. Moreover, time series forecasting is not a standard supervised learning problem, as observations are not *independent and identically distributed*, and off-the-shelf supervised learning algorithms may not be well suited to address it.

Table 3. Forecasts accuracy over 2007Q1-2017Q1

	Adaptive Trees	Indicator Model	Random Forest	Gradient Boosted Trees	AR(1)
UK, M+3	0.63	0.96	1.22	1.37	0.0067
UK, M+6	0.83	1.06	1.37	1.48	0.0070
UK, M+12	0.97		1.23	1.31	0.0077
UK, M+24	1.12		1.08	1.16	0.0074
USA, M+3	0.72	0.76	0.99	0.94	0.0070
USA, M+6	0.89	0.89	1.02	1.38	0.0073
USA, M+12	0.94		0.97	1.08	0.0076
USA, M+24	0.99		1.12	1.24	0.0070
France, M+3	0.72	0.70	1.00	1.02	0.0057
France, M+6	0.82	0.81	1.00	1.08	0.0062
France, M+12	0.93		0.85	0.87	0.0064
France, M+24	0.89		0.86	0.83	0.0064
Japan, M+3	0.93	0.90	1.13	1.16	0.0126
Japan, M+6	0.98	1.08	1.04	1.08	0.0126
Japan, M+12	0.97				0.0127
Japan, M+24	1.02				0.0123
Germany, M+3	0.81	0.65	1.15	1.21	0.0094
Germany, M+6	0.97	0.97	1.13	1.13	0.0096
Germany, M+12	1.10		0.99	1.46	0.0096
Germany, M+24	0.88				0.0109
Italy, M+3	0.69	0.62	1.01	1.12	0.0085
Italy, M+6	0.90	0.80	0.92	1.04	0.0090
Italy, M+12	0.99		1.05	1.52	0.0084
Italy, M+24	1.00				0.0089

Note: The Adaptive Trees, Indicator Model, Random Forest and Gradient Boosted Trees columns provide forecast RMSE as a ratio of the RMSE obtained with the baseline AR(1) model. The AR(1) columns provides RMSE of q-o-q quarterly GDP growth forecasts made with the baseline forecast. Simulations of the IM were only available for short-term horizons. The Random Forest is run with 500 trees. The max depth of trees is grid-searched[11] at the onset of the simulation among 9 candidates. Other parameters are sklearn defaults. GBT is trained with 500 trees as well. The learning rate is grid-searched among 10 candidates, and other parameters are sklearn defaults.
Source: OECD Economic Outlook databases, and OECD calculations.

[11] See glossary in Annex D

Table 4. Forecasts accuracy over 2007Q1-2010Q1

	Adaptive Trees	Indicator Model
UK, M+3	0.84	1.51
UK, M+6	1.34	1.70
USA, M+3	0.98	1.06
USA, M+6	1.30	1.29
France, M+3	0.99	0.97
France, M+6	1.19	1.25
Japan, M+3	1.32	1.14
Japan, M+6	1.38	1.50
Germany, M+3	1.22	0.96
Germany, M+6	1.56	1.54
Italy, M+3	1.05	0.99
Italy, M+6	1.36	1.31

Note: The Adaptive Trees and Indicator Model columns provide forecast RMSE as a ratio of the RMSE obtained with the baseline AR(1) model.
Source: OECD Economic Outlook databases, and OECD calculations.

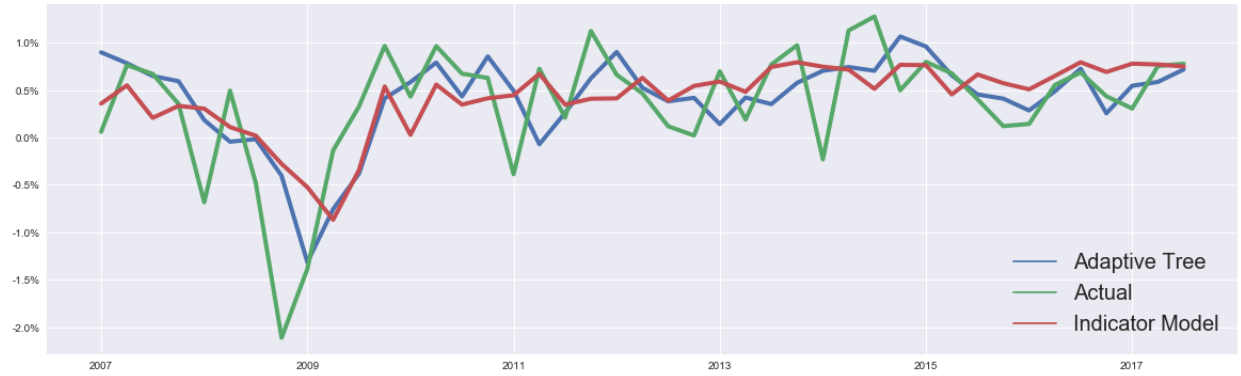
57. Table 4 displays results for the period around the global financial crisis (GFC). The Adaptive Trees perform better than the Indicator Models for the UK and the US, by a larger margin at M+3. For France and Italy, the results are comparable. The Indicator Model performs better in the cases of Japan and Germany.

58. Figure 5 displays the forecasts made with Adaptive Trees and the Indicator Model for selected countries and horizons. It shows the two sets of forecasts and the "actual" value of GDP growth. In the case of the United States and the United Kingdom, the gain in performance displayed in Table 3 seems to stem from a quicker adaptation to the decrease in GDP growth resulting from the crisis, although the Adaptive Trees forecast does not predict the crisis. The German forecast made by the Indicators Model is very accurate and the Adaptive Trees does no better.

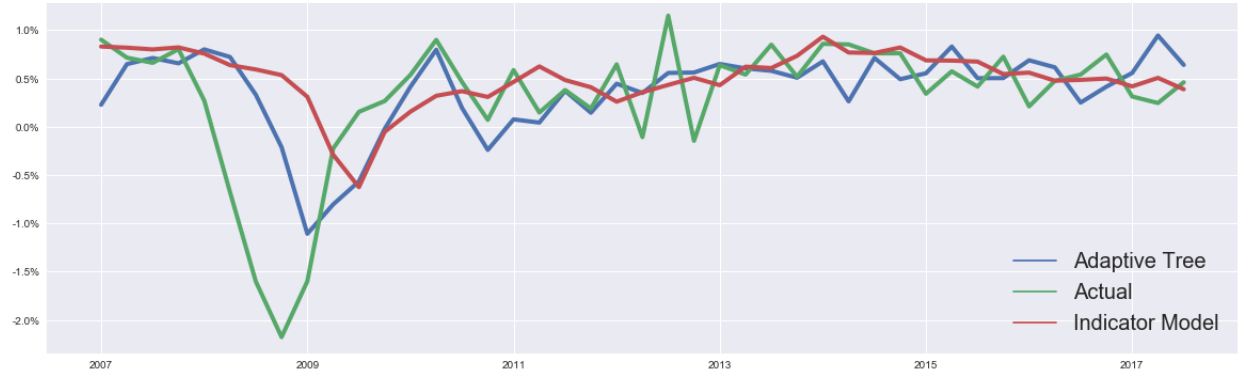
59. A reason why the Adaptive Trees forecast performs better in the case of the UK, and to a lesser extent in the case of the US might be that these forecasts involve house prices whose relationship with GDP growth is likely to be non-linear. The leverage cycle theory (Geanakoplos, 2014[32]) has it that crises often result from a sudden drop following a steep increase in asset prices, in most cases houses. The non-linear nature of Adaptive Trees could be particularly well suited to capture such phenomenon. A better assessment of how the Adaptive Trees capture this specific type of non-linearity would require more research.

Figure 5. Forecast simulations, GDP growth (Q on Q), selected countries and forecast horizons, 2007Q1-2017Q1

Panel A: United States, M+3



Panel B: United Kingdom, M+6



Panel C: Germany, M+3

Source: OECD Economic Outlook databases, and OECD calculations.

4.2. Interpretation

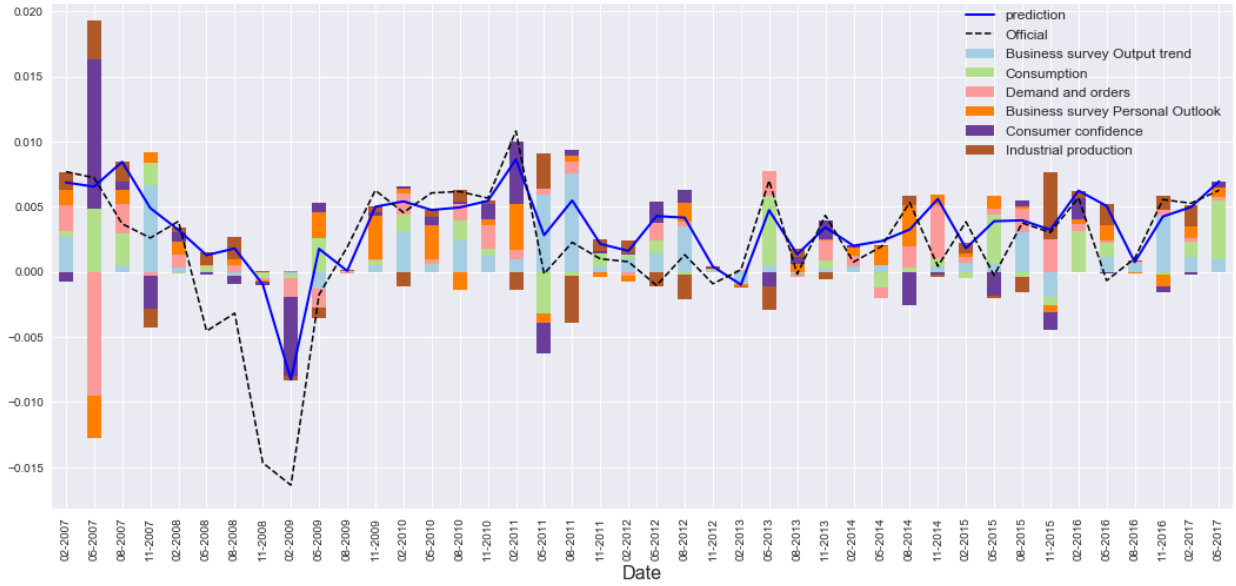
60. One advantage of ensemble of trees, including Adaptive Trees, is that it is possible to decompose each variable's contribution to a given prediction. This provides a way to interpret each prediction[12]. As argued above, a prediction made with Adaptive Trees can

[12] Decomposing predictions into variables' contributions provide "local interpretability", that is, an interpretation for the behaviour of the algorithm in the vicinity of a given observation. Local

be additively decomposed into a sample average and each variable's contribution (that explains the deviation from the sample average). Using variable contributions helps understand the patterns identified by the algorithm. Decomposition into variable contributions provides a window of interpretability on tree-based predictions. As the algorithm uses a multiplicity of lags and features, the contributions of each lag and feature of a given variable are summed up. Working with aggregated variable contributions enhances readability.

61. Figure 14 displays variable contributions to the France M+3 forecast. The business survey variables and consumer confidence play a critical role in forecasts for recessions, thus proving the importance of soft indicators in this model.

Figure 6. Aggregated variable contributions, France, M+3



Note: The bars represent the aggregated feature contributions. For instance, the blue bar represents the sum of the contributions of all the lags of Business survey, output trend used as covariates. The arithmetic sum of feature contributions (which can be positive or negative) sum to the prediction.
Source: OECD Economic Outlook databases, and OECD calculations.

5. Conclusion

62. Adaptive Trees resorts to an array of machine learning techniques to address specific issues arising in macroeconomic forecasting. The approach aims at addressing non-linearities and structural change in macroeconomic data, in particular by extending the Gradient Boosted Trees algorithm to weight more recent observations to better reflect structural changes. This research adds to the existing body of evidence that machine learning brings relevant new items to the forecaster's toolbox.

interpretability is distinct from global interpretability, that provides a rationale for all possible predictions (Renard et al., 2019[12]).

63. Forecast performance was assessed in a contest exercise: using the same data as the OECD Indicator Model, and comparing to the forecasts obtained using other techniques. The respective performances of the Adaptive Trees and the Indicator Model were assessed using pseudo-real time simulations, independently from one another. The Adaptive Trees performs broadly in line with the Indicator Model: it performs better for the United Kingdom forecasts, equally for France, Japan and the United States, and not as well for Germany and Italy.

64. The evaluation exercise in this paper has compared models using the same relatively small dataset. However, one advantage of Adaptive Trees and other similar algorithms is the ability to draw on a wide source of variables, allowing for higher dimensionality than the number of observations. Linear bridge equations cannot do this, while Dynamic Factor Models (DFM) often used for macroeconomic forecasts can bring in a large number of variables and then reduces them to a small number of factors. As distinct from DFMs, some machine learning techniques can capture a wide array of non-linearities in a data rich environment (Goulet Coulombe et al., 2019[33]).

65. Further application and development of Adaptive Trees could widen the set of input variables considerably, which may increase predictive power. The leading indicators used by the Indicator Model are supposed to be linearly correlated with GDP growth. It would be useful to apply the Adaptive Trees to broader sets of indicators, possibly non-linearly correlated to GDP, yet possibly highly informative about its future growth, such as financial data, policy data, or more granular data (including big data). Adding the Adaptive Trees forecast to the toolbox used for forecasting could yield interesting insights, especially when it diverges from linear models.

66. Interpretability is a major concern when it comes to forecasting. The proposed interpretation method yields local interpretation, i.e. explanation of given predictions by feature contributions. It is specific to tree-based predictive algorithms. New local interpretability methods have emerged recently (Lundberg and Lee, 2017[11]) that extend the possibility to decompose predictions in feature contributions to a larger class of algorithms. This innovation paves the way for experiments with other algorithms, such as deep learning or Gaussian processes, that could also be relevant to GDP growth forecasting.

67. If Adaptive Trees, or other machine learning algorithms, can capture non-linearities and structural change in the economy, there may be other applications than forecasts. In particular, there is a growing body of literature about the estimation of heterogeneous treatment effects of economic policies (Wager and Athey, 2018[34]; Athey and Imbens, 2016[35]; Athey, Tibshirani and Wager, 2019[36]). Predictive algorithms may be used to predict counterfactuals in order to perform causal inference. Powerful machine learning algorithms may be very useful in this regard (Alaa and Van Der Schaar, 2019[37]; Ertefaie, Asgharian and Stephens, 2018[38]; Alaa and Schaar, 2019[39]; Denton, 1995[40]; Chernozhukov et al., 2016[41]; Woloszko and Mavroeidi, 2019[42]). Even though linear models facilitate interpretability, the possibility to uncover non-linearities and structural changes provides supplementary arguments in favour of the use of machine learning algorithm for analysing economic policies.

References

Alaa, A. and M. Schaar (2019), "Validating Causal Inference Models via Influence Functions", *Proceedings of Machine Learning Research*, pp. 191-201, http://proceedings.mlr.press/v97/alaa19a.html (accessed on 27 June 2019). [39]

Athey, S. and G. Imbens (2016), "Recursive partitioning for heterogeneous causal effects", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 113/27, pp. 7353-7360, http://dx.doi.org/10.1073/pnas.1510489113. [35]

Athey, S., J. Tibshirani and S. Wager (2019), "Generalized random forests", *The Annals of Statistics*, Vol. 47/2, pp. 1148-1178, http://dx.doi.org/10.1214/18-AOS1709. [36]

Biau, O. and A. D'Elia (2009), *Euro area GDP forecasting using large survey datasets*, http://millenniumindicators.un.org/unsd/nationalaccount/workshops/2010/moscow/AC223-S73Bk4.PDF. [6]

Blanchard, O. (2014), "Where Danger Lurks", *Finance & Development*, https://www.imf.org/external/pubs/ft/fandd/2014/09/blanchard.htm. [5]

Breiman, L. (2001), "Random Forests", *Machine Learning*, Vol. 45/1, pp. 5-32, http://dx.doi.org/10.1023/A:1010933404324. [53]

Breiman, L. (2001), "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)", *Statistical Science*, Vol. 16/3, pp. 199-231, http://dx.doi.org/10.1214/ss/1009213726. [8]

Breitung, J. and M. Knüppel (2018), "How far can we forecast? Statistical tests of the predictive content", *Discussion Papers*. [31]

Cai, J. et al. (2018), "Feature selection in machine learning: A new perspective", *Neurocomputing*, http://dx.doi.org/10.1016/j.neucom.2017.11.077. [30]

Carpenter, S. and W. Brock (2006), "Rising variance: a leading indicator of ecological transition", *Ecology Letters*, Vol. 9/3, pp. 311-318, http://dx.doi.org/10.1111/j.1461-0248.2005.00877.x. [27]

Chakraborty, C. and A. Joseph (2017), "Machine Learning at Central Banks", *SSRN Electronic Journal*, http://dx.doi.org/10.2139/ssrn.3031796. [2]

Chandrashekar, G. and F. Sahin (2014), "A survey on feature selection methods", *Computers and Electrical Engineering*, http://dx.doi.org/10.1016/j.compeleceng.2013.11.024. [29]

Chaudhuri, K. and R. Salakhutdinov (eds.) (2019), *Validating Causal Inference Models via Influence Functions*, PMLR, Long Beach, California, USA, http://proceedings.mlr.press/v97/alaa19a.html. [37]

Chen, T. and C. Guestrin (2016), *XGBoost*, http://dx.doi.org/10.1145/2939672.2939785. [21]

Chernozhukov, V. et al. (2016), "Double machine learning for treatment and causal parameters", *arXiv preprint arXiv:1608.00060*, https://arxiv.org/abs/1608.00060. [41]

Christ, M. et al. (2018), "Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package)", *Neurocomputing*, Vol. 307, pp. 72-77, http://dx.doi.org/10.1016/J.NEUCOM.2018.03.067. [25]

Christ, M., A. Kempa-Liehr and M. Feindt (2016), "Distributed and parallel time series feature extraction for industrial big data applications", http://arxiv.org/abs/1610.07717 (accessed on 6 April 2019). [50]

Clements, M. and A. Galvão (2008), "Macroeconomic Forecasting With Mixed-Frequency Data", *Journal of Business & Economic Statistics*, Vol. 26/4, pp. 546-554, http://dx.doi.org/10.1198/073500108000000015. [46]

Cook, R. (1977), "Detection of Influential Observation in Linear Regression", *Technometrics*, Vol. 19/1, pp. 15-18, http://dx.doi.org/10.1080/00401706.1977.10489493. [54]

Craven, M. and J. Shavlik (1996), "Extracting tree-structured representations of trained networks", *Advances in neural information processing systems*, http://papers.nips.cc/paper/1152-extracting-tree-structured-representations-of-trained-networks.pdf (accessed on 30 January 2019). [56]

Dakos, V. et al. (2012), "Methods for detecting early warnings of critical transitions in time series illustrated using simulated ecological data.", *PloS one*, Vol. 7/7, p. e41010, http://dx.doi.org/10.1371/journal.pone.0041010. [26]

Denton, J. (1995), "How good are neural networks for causal forecasting?", *The Journal of Business Forecasting*, Vol. 14/2, p. 17, http://search.proquest.com/openview/0e806ff6c62b7986147ed84a85b38e28/1?pq-origsite=gscholar&cbl=28144. [40]

Dietterich, T. (2000), "Ensemble Methods in Machine Learning", Springer, Berlin, Heidelberg, http://dx.doi.org/10.1007/3-540-45014-9_1. [52]

Efron, B. (1983), "Estimating the error rate of a prediction rule: improvement on cross-validation", *Journal of the American Statistical Association*, Vol. 78/382, pp. 316-331, http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1983.10477973. [17]

Ertefaie, A., M. Asgharian and D. Stephens (2018), "Variable Selection in Causal Inference using a Simultaneous Penalization Method", *Journal of Causal Inference*, Vol. 6/1, http://dx.doi.org/10.1515/jci-2017-0010. [38]

Foroni, C. and M. Marcellino (2013), "A Survey of Econometric Methods for Mixed-Frequency Data", *SSRN Electronic Journal*, http://dx.doi.org/10.2139/ssrn.2268912. [58]

Freund, Y. and R. Schapire (1997), "A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting", Vol. 55, pp. 119--139, https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.32.8918 (accessed on 8 April 2019). [22]

Friedman, J. (2002), "Stochastic gradient boosting", *Computational Statistics & Data Analysis*, Vol. 38/4, pp. 367-378, http://dx.doi.org/10.1016/S0167-9473(01)00065-2. [23]

Friedman, J., T. Hastie and R. Tibshirani (2001), *The elements of statistical learning*, http://statweb.stanford.edu/~tibs/book/preface.ps (accessed on 8 April 2019). [55]

Fulcher, B. (2018), "Feature-Based Time-Series Analysis", in *Feature Engineering for Machine Learning and Data Analytics*, CRC Press, http://dx.doi.org/10.1201/9781315181080-4. [49]

Geanakoplos, J. (2014), "Leverage, Default, and Forgiveness: Lessons from the American and European Crises", *Journal of Macroeconomics*, Vol. 39, pp. 313-333, http://dx.doi.org/10.1016/J.JMACRO.2014.01.001. [32]

Ghysels, E. and J. Wright (2009), "Forecasting Professional Forecasters", *Journal of Business & Economic Statistics*, Vol. 27/4, pp. 504-516, http://dx.doi.org/10.1198/jbes.2009.06044. [45]

Giannone, D., L. Reichlin and D. Small (2008), "Nowcasting: The real-time informational content of macroeconomic data", *Journal of Monetary Economics*, Vol. 55/4, pp. 665-676, http://dx.doi.org/10.1016/J.JMONECO.2008.05.010. [48]

Gogas, P. et al. (2015), "Yield Curve and Recession Forecasting in a Machine Learning Framework", *Computational Economics*, http://dx.doi.org/10.1007/s10614-014-9432-0. [3]

Goulet Coulombe, P. et al. (2019), "How is Machine Learning Useful for Macroeconomic Forecasting?". [33]

Guyon, I. and A. Elisseeff (2003), "An Introduction to Variable and Feature Selection", *Journal of Machine Learning Research*, Vol. 3/Mar, pp. 1157-1182, http://www.jmlr.org/papers/v3/guyon03a.html (accessed on 6 April 2019). [28]

Jung, J., M. Patnam and A. Ter-Martirosyan (2018), "An Algorithmic Crystal Ball: Forecasts-based on Machine Learning", *IMF Working Paper*, https://books.google.com/books?hl=en&lr=&id=7KV6DwAAQBAJ&oi=fnd&pg=PP1&dq=An+Algorithmic+Crystal+Ball:+Forecasts-based+on+Machine+Learning&ots=p8JkLItoDC&sig=fHIqdmKdVH6WVYQ9sTOH0E_hZLQ (accessed on 8 April 2019). [1]

Kirman, A. (2010), *Complex economics: individual and collective rationality*, https://www.taylorfrancis.com/books/9781136941689 (accessed on 17 May 2019). [9]

Kotsiantis, S., D. Kanellopoulos and P. Pintelas (2006), "Data preprocessing for supervised leaning", *International journal of computer sciences*, Vol. 1/1, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.104.8413&rep=rep1&type=pdf (accessed on 5 April 2019). [43]

Lam, S., A. Pitrou and S. Seibert (2015), "Numba: A LLVM-based python JIT compiler", *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC - LLVM '15*, http://dx.doi.org/10.1145/2833157.2833162. [59]

Lewis, C. and N. Pain (2015), *Lessons from OECD forecasts during and after the financial crisis by*, https://www.oecd.org/eco/growth/Lessons-from-OECD-forecasts-during-and-after-the-financial-crisis-OECD-Journal-Economic-Studies-2014.pdf (accessed on 7 April 2019). [20]

Lipton, Z. (2016), "The Mythos of Model Interpretability", http://arxiv.org/abs/1606.03490 (accessed on 7 April 2019). [10]

Lundberg, S. and S. Lee (2017), *A Unified Approach to Interpreting Model Predictions*, http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions (accessed on 7 April 2019). [11]

Mariano, R. and Y. Murasawa (2010), "A Coincident Index, Common Factors, and Monthly Real GDP", *Oxford Bulletin of Economics and Statistics*, Vol. 72/1, pp. 27-46, http://dx.doi.org/10.1111/j.1468-0084.2009.00567.x. [47]

Mourougane, A. (2006), "Forecasting Monthly GDP for Canada", *OECD Economics Department Working Papers*, No. 515, OECD Publishing, Paris, https://dx.doi.org/10.1787/421416670553. [57]

Ollivaud, P. et al. (2016), "Forecasting GDP during and after the Great Recession: A contest between small-scale bridge and large-scale dynamic factor models", *OECD Economics Department Working Papers*, No. 1313, OECD Publishing, Paris, https://dx.doi.org/10.1787/5jlv2jj4mw40-en. [15]

Pain, N. et al. (2014), "OECD Forecasts During and After the Financial Crisis: A Post Mortem", *OECD Economics Department Working Papers*, No. 1107, OECD Publishing, Paris, https://dx.doi.org/10.1787/5jz73l1qw1s1-en. [16]

Pedregosa, F. et al. (2011), "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research*, Vol. 12/Oct, pp. 2825-2830, http://www.jmlr.org/papers/v12/pedregosa11a.html (accessed on 6 April 2019). [51]

Renard, X. et al. (2019), "Concept Tree: High-Level Representation of Variables for More Interpretable Surrogate Decision Trees", http://arxiv.org/abs/1906.01297 (accessed on 9 September 2019). [12]

Romer, P. (2016), "The Trouble with Macroeconomics", *The American Economist*, http://dx.doi.org/10.1177/ToBeAssigned. [4]

Saabas, A. (2014), *Interpreting random forests*, http://blog.datadive.net/interpreting-random-forests/ (accessed on 7 April 2019). [13]

Schneider, J. (1997), "Cross validation", *A Locally Weighted Learning Tutorial Using Vizier*, Vol. 1. [18]

Sédillot, F. and N. Pain (2003), "Indicator Models of Real GDP Growth in Selected OECD Countries", *OECD Economics Department Working Papers*, No. 364, OECD Publishing, Paris, https://dx.doi.org/10.1787/275257320252. [19]

Tiffin, A. (2016), "Seeing in the Dark: A Machine-Learning Approach to Nowcasting in Lebanon", *IMF Working Papers*, Vol. 16/56, p. 1, http://dx.doi.org/10.5089/9781513568089.001. [7]

Turner, D. (2016), *The use of models in macroeconomic forecasting at the OECD*, http://www.oecd.org/eco/workingpapers (accessed on 7 April 2019). [14]

Wager, S. and S. Athey (2018), "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests", *Journal of the American Statistical Association*, Vol. 113/523, pp. 1228-1242, http://dx.doi.org/10.1080/01621459.2017.1319839. [34]

Wallis, K. (1986), "Forecasting with an econometric model: The 'ragged edge' problem", *Journal of Forecasting*, Vol. 5/1, pp. 1-13, http://dx.doi.org/10.1002/for.3980050102. [44]

Woloszko, N. and E. Mavroeidi (2019), "Analysing non-linear relationships between structural policies and growth with Double-Post-Lasso", OECD (Forthcoming). [42]

Žliobaitė, I. (2010), "Learning under Concept Drift: an Overview", http://arxiv.org/abs/1010.4784 (accessed on 9 September 2019). [24]

Annex A. Detailed methodology

1. This annex aims at providing a detailed technical description of the components of the Adaptive Trees algorithms, used to perform GDP short-term forecasts. It also provides insights on the contribution of each component to overall forecast performance. The complete description of the algorithm includes data pre-processing steps and estimation. The proposed algorithm combines multiple existing machine learning techniques (rescaling, feature engineering, feature selection, gridsearch, early stopping, ensemble learning) and innovative approaches to the field (namely predictive interpolation and adaptive boosting). This annex aims at guaranteeing replicability of the proposed method. It first describes the four pre-processing steps, and then covers the characteristics of the training and prediction.

Pre-processing

2. Data pre-processing is often a pivotal element of predictive machine learning (Kotsiantis, Kanellopoulos and Pintelas, 2006[43]). The Adaptive Trees approach uses a combination of both existing machine learning pre-processing techniques (scaling, feature engineering and selection), and introduces predictive interpolation as novel method to deal with mixed-frequency data.

First pre-processing steps

5.1.2. Time alignment

3. In real-time, publication delays cause missing values for some of the variables at the end of the sample (Wallis, 1986[44]). This so-called "ragged-edge" problem is dealt with an alignment procedure that consists in shifting columns by an amount of time equal to the release delay. Should industrial production has a publication delay of two months, it will be lagged by two months in order to ensure time consistency. Subsequently, contemporaneous values of the variables coming with release delays are made unavailable during the training and simulation.

5.1.3. Standardisation

4. All features are standardised using rescaling. This simple pre-processing sets each feature's mean to 0 and standard deviation to 1. When dealing with XGBoost, standardisation is not neutral. Standardisation has an impact on feature selection as well.

Predictive interpolation

5. Predicting quarterly GDP growth with monthly indicators (such as employment, industrial production, retail sales) raises a mixed frequency problem. The literature proposes various approaches to address this issue, including Mixed Frequency Sampling (MIDAS, see for instance (Ghysels and Wright, 2009[45]; Clements and Galvão, 2008[46]))), mixed frequency VARs (Mariano and Murasawa, 2010[47]), and mixed frequency factor models (Giannone, Reichlin and Small, 2008[48]) [13]. The approach used in the Adaptive

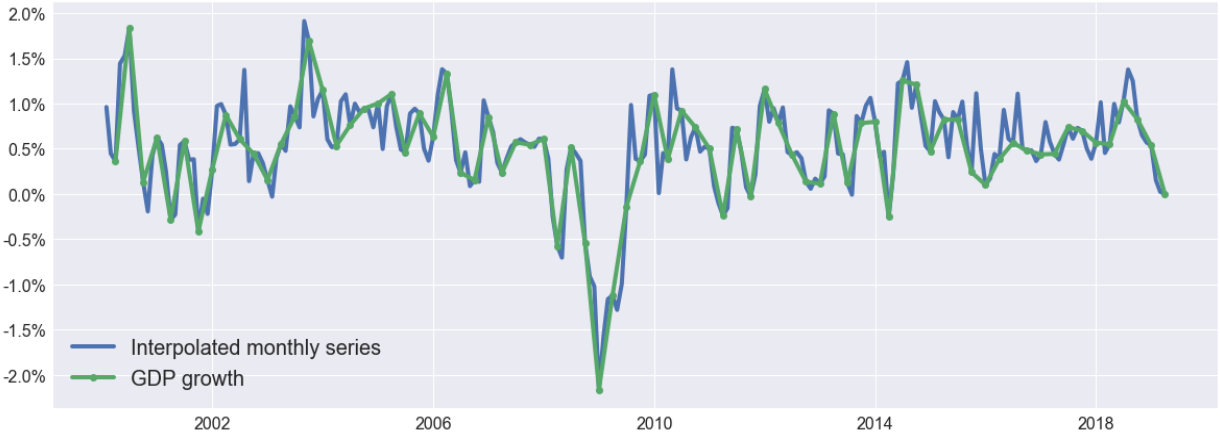
[13] See (Foroni and Marcellino, 2013[58]) for a literature review.

Trees framework relies on two steps: first, creating an estimated monthly GDP growth series by interpolating missing values using a machine learning predictor, and second performing the estimation of the Adaptive Trees model on the complete set of monthly observations. In the first step, the "predictive interpolation" method trains a predictive algorithm on available observations, and predicts monthly estimates of GDP growth. The predictive interpolation algorithm uses XGBoost, feature selection, and is fine-tuned using gridsearch (thus sharing the main features of the second step predictor). Gridsearch is performed using 5-fold cross-validation, and the learning rate and level of regularization of the XGBoost estimator. Feature selection is achieved using the same methodology as in the second step, that is described at length below.

6. Predictive interpolation allows to make use of the full dataset, and yields better performance than naive interpolation (by the mean or linear interpolation) or no interpolation. The in-sample contemporaneous estimation of monthly GDP values in the first step is accurate enough to enhance the forecasting of out-of-sample leading GDP growth values in the second step. The resulting interpolated monthly series in the case of the United States are shown in Figure 7.

Figure 7. Predictive interpolation

GDP growth and monthly interpolated series for the United States



Source: OECD calculations.

Feature engineering

7. Feature engineering has become standard practice in the machine learning community when dealing with time series (Fulcher, 2018[49]; Christ, Kempa-Liehr and Feindt, 2016[50]; Christ et al., 2018[25]). A feature is an attribute of a time-series variable over a given time window, such as its min, max, standard deviation, a lag, and so on. Feature engineering consists in extracting features from time series variables. In the Adaptive Trees framework, ten features are extracted from each variables for eight time windows. The features are described in Table 5. Each feature is extracted over the 3, 6, 9, 12, 18, 24, and 36 past months. Should the original data include six variables, 6 * 7 * 10 =

420 features would be added to the training data[14]. The large number of resulting features calls for the subsequent use of feature selection.

Table 5. Feature engineering

Features extracted from time series variables.

Feature	Definition
Moving average	Average in the time window
Cumulated growth	$\prod_t(1 + x_t) - 1$ where x_t is a growth rate (only for variables expressed as a growth rate)
Volatility	Standard deviation over the time window
Change over the period	Last value minus first value
Spread Min Max	Max value minus min value over the time window
Max	Max over period
Min	Min over period
Mean second derivative	Average of the twice-differenced series over the time window
Mean absolute change	$\frac{1}{n}\sum_{t=1}^{n}\lvert x_t - x_{t-1}\rvert$
Trend projection	Prediction by a linear regression of y on the time vector at a M+6 horizon

Source: OECD.

Feature selection

8. Feature selection reduces the feature space by removing features and may thus increases predictive accuracy (Guyon and Elisseeff, 2003[28]; Chandrashekar and Sahin, 2014[29]; Cai et al., 2018[30]). There is a large number of feature selection methods in the literature. The Adaptive Trees framework relies on model-based feature selection, as it uses the feature importance scores issues by the XGBoost[15]. All features with null importance are removed. The resulting training set is then used for the training and prediction.

9. Resorting to feature selection significantly reduces training time. The training time of the XGBoost algorithm increases with the number of features. Selecting a smaller number of features considerably decreases the time taken by the grid search, that trains the algorithm for each hyper-parameters combination, thus allowing for finer parameter gridsearching.

Training and prediction

10. Pre-processing, feature engineering and selection yield a training data set that is used for training the forecast component of the Adaptive Trees framework and make a GDP growth prediction. Three performance-enhancing techniques are used around the key XGBoost predictor: grid search, ensemble learning and adaptive boosting. The following

[14] For the sake of efficieny, features are computed using Python and jit-compiler Numba (Lam, Pitrou and Seibert, 2015[59]).

[15] XGBoost has three feature importance score. The default "gain" measure is used.

paragraphs provide a detailed description of the XGBoost predictor and these three techniques.

XGBoost: a fast and powerful predictive algorithm

11. XGBoost is an implementation of the Gradient Boosting Trees. A detailed presentation of the XGBoost algorithm is beyond the scope of this paper, and can be found in (Chen and Guestrin, 2016[21]). XGBoost is an optimized distributed gradient boosting library. It has gained widespread currency in the machine learning community and has emerged the main challenger to deep learning approaches when it comes to structured data. The XGBoost implementation of the gradient boosting algorithm differs from the standard scikit-learn implementation in three ways. First, it can be parallelized and proves much faster. Second, it implements regularization (both in the L1 and L2 norms), in order to prevent overfitting.

12. Third, XGBoost can be trained on a data set that includes missing values (on the X side). When the data is sparse, an instance is classified in the default direction. At every tree node, there are two possible direction: left or right. The optimal default direction is learnt from the data. This characteristic is particularly relevant to macroeconomic forecasting as the variables often have different historical depth. Using XGBoost thus allows to include a larger training set. For instance, among variables used to predict GDP growth in the UK, the Purchasing Manager Index starts in 1992. All data before that date would have to be done away with if it were not for the use of XGBoost.

13. The XGBoost framework includes the option to define the number of trees in the ensemble with Early Stopping. The algorithms adds regression trees to the ensemble until performance on a user-defined test set has not improved after a fixed number of training iterations. Early stopping contributes to further reducing the training time. In the Adaptive Trees forecast, the early stop test set is randomly drawn from the training data and represents 10% of the data.

Gridsearching hyperparameters

14. XGBoost involves a number of parameters that require fine-tuning. The Adaptive Trees framework resorts to cross-validated gridsearch to do so. It cross-validates each possible parameter combination in a user-given "parameter grid" and retains the one that yields the best out-of-sample accuracy (Pedregosa et al., 2011[51]). The next paragraph provides information on the parameter grid and the cross-validation scheme.

15. The XGBoost predictor used to produce the GDP growth forecasts results from a gridsearch of the learning rate (over 10 values), the observation weight parameter (over 8 values), and the gamma regularization parameter (over 10 values). For each of the 800 possible parameter combinations, a predictor is cross-validated using "forward looking cross-validation". The sample is split 12 times into a training set (all observations until a given date T) and a test set (the 4 observations following T). The process thus implies 800*12= 9600 estimations.

Ensemble

16. Ensemble methods consist in averaging the predictions of a series of models in order to reduce overall prediction standard deviation (Dietterich, 2000[52]). Examples of ensemble learning algorithms include random forests, that aggregate the predictions of a number regression trees trained on bootstrap samples drawn from the training set (Breiman,

2001[53]). In the Adaptive Trees framework, the GDP growth prediction resorts to ensemble learning. 50 XGBoost predictors are trained and their predictions are averaged. The 50 estimators have the same parameters based on the gridsearch described above. They differ in two respects. First, each has a different seed parameter, thus introducing randomness in the prediction. Second, the training and test sets used to perform early stopping are drawn randomly, thus introducing some randomness in the resulting number of trees. The use of an ensemble contributes to reduce the effect of these two factors of randomness and decreases the variance of the prediction.

Adaptive Boosting

17. The Adaptive Trees framework introduces an innovative feature called "adaptive boosting" in order to tackle concept drift (or structural change, i.e., the changing joint distribution of target and predictors over time). Adaptive boosting consists in introducing increasing *ex ante* observation weights to a gradient boosting algorithm (such as XGBoost) that give more weight to hard-to-predict observations over the course of boosting iterations. This innovation results under certain conditions on *ex post* observation weights that reflect shocks and structural breaks by putting greater weight on more recent observations. The next paragraphs provide more details about the functioning of adaptive boosting and some insights on the *ex post* observations weights.

18. The *ex ante* observation weights are user-defined and can thus be optimised with parameter gridsearch. Let $w(t)$ be the weight of the observation made at time $t, t \in [1, N]$, N being the number of observations. The Adaptive Trees methodology defines w as follows:

$$w(t) = e^{-\gamma(\frac{t}{N}-1)}$$

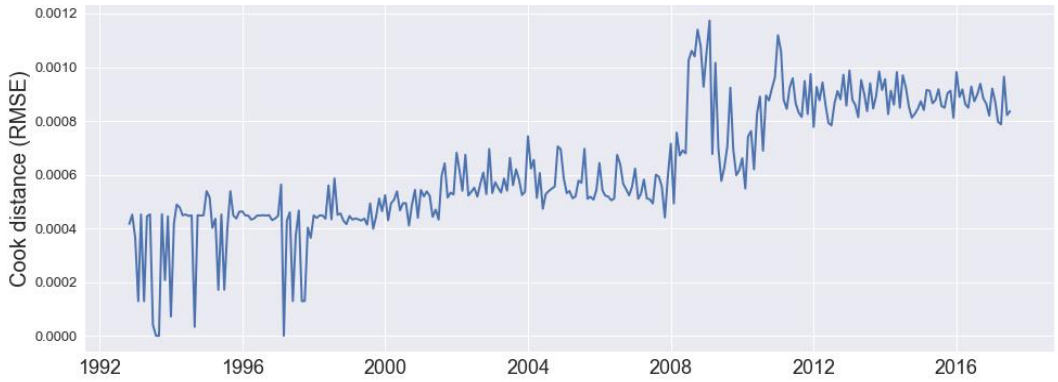
19. The first observation thus has an *ex ante* weight equal to $e^{-\gamma}$when the last observation has an *ex ante* weight of 1. The γ parameter defines the steepness of the *ex ante* weights curve and is cross-validated. It is often close to 15. The value of γ might be considered as a measure of structural change in the training set.

20. The *ex post* weights result from the training. They are neither user-defined nor directly observable. Although this is not necessary to perform forecasts, the ex post weights can be analysed using the Cook distance (Cook, 1977[54]). The Cook distance measures the importance of a given observation by evaluating how different are predictions when that observations is removed from the training set. Formally, it is equal to the mean squared difference between a set of predictions made using the whole training sample and the same set of predictions after a given observation has been removed. It thus measures the impact of a given observation in the training of a predictive algorithm. It was originally developed with linear regressions, but can be applied to machine learning predictors as well.

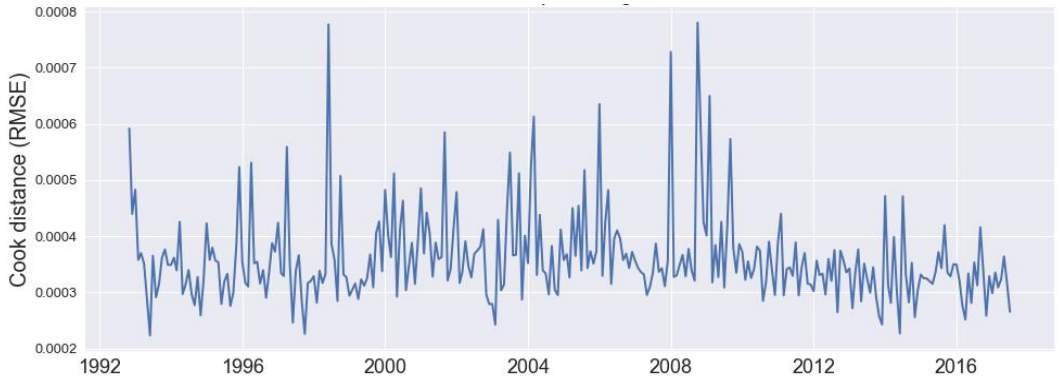
21. Figure 2 shows the Cook distance for each observation for a US GDP forecast at a M+6 horizon. In panel A, the predictions are made using Adaptive Trees (with $\gamma = 12$), while in panel B $\gamma = 0$. Although *ex ante* observation weights follow an exponential increase across observations, *ex post* observation weights in panel A display two clear plateaus and are higher after the GFC. This pattern seems to underline some structural change happening around the GFC, thus making subsequent observations more informative about future GDP growth. In panel B, *ex ante* observation weights are constant ; there is no trend in *ex post* observation weights. Some observations are particularly important, especially during the GFC.

Figure.8. *Ex post* observation weights

Panel A. Cook distance from Adaptive Trees.



Panel B. Cook distance from Gradient Boosted Trees.



Note: The y-axis measures how different predictions over the whole sample are when removing the observation defined on the x-axis from the training sample.
Source: OECD Economic Outlook databases, and OECD calculations.

Annex B. The OECD Indicator Model

For the euro area and individual G7 economies, the Indicator Model is a suite of statistical models using high-frequency indicators to provide estimates of near-term quarterly GDP growth, typically for the current and next quarter or so. This analysis builds on the work of Sédillot and Pain (2003) and Mourougane (2006) in using short term economic indicators to predict quarterly movements in GDP by efficiently exploiting all available monthly and quarterly information. These models typically combine information from both "soft" indicators, such as business sentiment and consumer surveys, and "hard" indicators, such as industrial production, retail sales, house prices etc. and use is made of different frequencies of data and a variety of estimation techniques. The procedures are relatively automated and can be run whenever major monthly data are released, allowing up dating and choice of model according to the information set available.

The most important gains from using the indicator approach are found to be for current-quarter forecasts made at or immediately after the start of the quarter in question, where estimated indicator models appear to outperform autoregressive time series models, both in terms of the size of error and directional accuracy. The main gains from using a monthly approach arise once one month of data is available for the quarter being forecast, typically two to three months before the publication of the first official outturn estimate for GDP. For one-quarter-ahead projections, the performance of the estimated indicator models is only noticeably better than simpler time series models once one or two months of information become available for the quarter preceding that being forecast. Modest gains are nonetheless to be made in terms of directional accuracy from using the indicator models.

Statistical indicator models are nonetheless limited in their ability to forecast quarterly GDP growth. Even with a complete set of monthly indicators for the quarter, the 70 per cent confidence bands around any point estimate for GDP growth in that quarter lie in the range from 0.4 to 0.8 percentage points, depending on the country or region and the degree of uncertainty is found to widen as the forecast horizon lengthens. Forecasting errors can also arise for a variety of reasons, including revisions to the initial published data and inaccuracies in the projections of the incoming monthly data.

Regular indicator model-based estimates of GDP now feed into both routine Economic Outlook assessment exercises and interim analyses and forecast updates released to the press on a routine basis.

Annex C. Data description

Table A C.1. Data availability per country and number of variables

Country	Start	End	Number of observations	Number of variables
Japan	07-2002	01-2017	58	8
Germany	03-1998	01-2017	75	9
USA	10-1985	01-2017	125	9
Italy	10-1998	01-2017	73	5
UK	10-1992	01-2017	97	6
France	10-1987	01-2017	117	6

Source: OECD Economic Outlook databases, and OECD calculations.

Annex D. Glossary

Feature. A feature is a characteristic of an observation. It is a column of the dataset, where observations correspond to rows. When dealing with time series variables, features may include the variable itself, a lead, a lag, or a more complex characteristic such as the rolling mean, standard deviation, and so on.

An algorithm. In machine learning, an algorithm refers to a piece of software that learns from the data and produces an set of decision rules used to predict values (regression) or categories (classification). The set of decision rules can also be labelled an algorithm. In other words, the training algorithm produces the predictive algorithm, that produces the predictions.

Training. The training is the phase when the (training) algorithm learns from the (training) data. It ushers in the production of the predictive algorithm. The training data has to be labelled: it requires both X and y.

Test set. Data removed from the training set, used to assess out-of-sample accuracy.

Ensemble. An ensemble is an algorithm composed of a plurality of algorithms. It consists in a series of predictive algorithm and a rule to decide how the array of learners "vote" to establish a prediction. In regression settings, the ensemble's prediction can be the mean, or the median (or else) of the predictions made by the series of learners.

Overfitting. An algorithm overfits when it learns characteristics of the sample instead of characteristics of the underlying population. Overfitting is diagnosed by the coincidence of good prediction accuracy on the training data and poor accuracy on the test data.

Cross-validation. A method designed to avoid overfitting and underfitting. It consists in splitting the sample in a series of 2-fold partitions, one fold being used as a training set and the other being used a the test set.

Grid search. A method designed to chose optimal hyperparameters. The user specify a grid: possible values for each parameter. Grid search consists in applying cross-validation to every possible parameter combination and select the one whose average performance on the test sets is the best.

Annex E. Full charts

Figure E.1. USA

Source: OECD Economic Outlook databases, and OECD calculations.

Figure E.2. UK

Source: OECD Economic Outlook databases, and OECD calculations.

Figure E.3. France

Source: OECD Economic Outlook databases, and OECD calculations.

Figure E.4. Japan

Source: OECD Economic Outlook databases, and OECD calculations.

Figure E.5. Germany

Source: OECD Economic Outlook databases, and OECD calculations.

Figure E.6. Italy

Source: OECD Economic Outlook databases, and OECD calculations.