Received: 22 October 2022 | Revised: 6 November 2023 | Accepted: 24 November 2023

DOI: 10.1002/jae.3088

WILEY Journal of APPLIED ECONOMETRICS

RESEARCH ARTICLE

The benefits of forecasting inflation with machine learning: New evidence

Andrea A. Naghi[1] | **Eoghan O'Neill**[2,3] | **Martina Danielova Zaharieva**[4]

[1]Queen Mary University of London, London, UK

[2]Erasmus University Rotterdam, Rotterdam, The Netherlands

[3]Tinbergen Institute, Amsterdam, The Netherlands

[4]CUNEF University, Madrid, Spain

Correspondence
Andrea A. Naghi, Queen Mary University of London, London, UK.
Email: a.naghi@qmul.ac.uk

Funding information
EU Horizon 2020, Marie Skłodowska-Curie individual grant, Grant/Award Number: 797286; Dutch Research Council (NWO), Grant/Award Number: EINF-3246; Spanish State Research Agency (Agencia Estatal de Investigación - Ministerio de Ciencia e Innovacíon), Grant/Award Number: PID2022-138289NB-I00

Summary

Medeiros et al. (2021) (Journal of Business & Economic Statistics, 39:1, 98–119) find that random forest (RF) outperforms US inflation forecasting benchmarks. We replicate the main results in Medeiros et al. (2021) and (1) considerably expand the set of machine learning methods, (2) analyse the predictive ability of both the initial and extended sets of methods on Canadian and UK data, (3) add results on coverage rates and widths of prediction intervals and (4) extend the sample from January 2016 to October 2022. Our narrow replication confirms the main findings of the original paper. However, the wider replication results suggest that other methods are competitive with RF and often more accurate. In addition, RF produces disappointing results during the coronavirus pandemic and subsequent high inflation of 2020–2022, whereas a stochastic volatility model and some gradient boosting methods produce more accurate forecasts.

KEYWORDS

forecasting, inflation, machine learning, random forest, tree-based methods

1 | INTRODUCTION

Medeiros et al. (2021) is a landmark contribution in the application of machine learning methods to inflation forecasting. The paper showed that standard inflation forecasting methods, such as the random walk (RW), autoregressive (AR) and unobserved components stochastic volatility (UCSV) models, can be outperformed by machine learning methods, such as LASSO or random forest (RF). This challenged the claim that it is "exceedingly difficult" to improve on simple models (Faust & Wright, 2013; Stock & Watson, 2007). Medeiros et al. (2021) find that among the ML methods under consideration, RF is the one that consistently performs the best. Its good performance is attributed to its variable selection properties and to its ability to explore interactions and nonlinearities.

Given the importance of these findings, in light of the emerging literature on models for data-rich environments, and coupled with the current worldwide increasing inflation rates, we replicate the paper by Medeiros et al. (2021) in both a narrow and a wide sense. For the replication in a narrow sense, we revisit and verify the computational validity of the main results for the FRED-MD data (McCracken & Ng, 2016).[1]

[1]For the replication in a narrow sense, we use a *Github* fork of the *R* code written by Medeiros et al. (2021) and edited in October 2021 subsequent to publication, to correct a look ahead bias in the original code – —found by Medeiros et al. (2021) and explained in email correspondence. We also fixed

The study is replicated in a wider sense as follows. First, given the promising results returned by ML methods, especially RF, we investigate over 30 other ML methods. We employ the following groups of ML methods: RF variations, gradient boosted trees (GBTs), Bayesian additive regression trees (BART), support vector machines (SVMs), neural networks (NNs) and penalized linear regressions. We add to the analysis of Medeiros et al. (2021) by including prediction interval widths and coverage rates. Then, we explore whether ML methods outperform standard benchmarks in forecasting CPI inflation for Canada and the UK. Finally, we extend the original sample to October 2022.

Our replication in a narrow sense confirms the main findings of Medeiros et al. (2021). RF outperforms standard benchmark methods and shrinkage methods. The wider replication results for the Canadian data suggest that RF provides limited gains in forecast accuracy relative to linear methods. Moreover, for the Canadian data, most of the gains in accuracy of RF result from variable selection as opposed to detection of nonlinearities. For the UK data, our results mostly confirm the findings of Medeiros et al. (2021). For all datasets, we find that machine learning methods that forecast the conditional median can outperform standard conditional mean forecasts,[2] even in terms of RMSE. The improvements in RMSE from use of conditional median forecasts instead of conditional mean forecasts are much larger for the US data than for Canada or the UK. In addition, we find that simple forecast combinations of machine learning methods consistently produce among the most accurate forecasts. In the extended US data, RF is one of the best methods up to 2019. However, from 2020 to 2022, RF is less accurate than RW forecasts, while UCSV models and some boosted tree methods outperform a RW by a similar margin to pre-pandemic times. Overall, the results show that an AR model is a more appropriate benchmark. Across all datasets, RMSE improvements are generally much smaller relative to an AR model than relative to RW.[3]

Following the notation of Medeiros et al. (2021), the goal is to forecast inflation π_{t+h} at h-periods ahead conditional on the information set available up to time t, using a large set of predictors denoted by $\mathbf{x_t} = (x_{1t}, \dots, x_{nt})'$: $\pi_{t+h} = G_h(\mathbf{x_t}) + u_{t+h},\ t = 1, \dots, T,\ h = 1, \dots, H$. The function $G_h(\cdot)$ takes predictors $\mathbf{x_t}$ as input and u_{t+h} is the prediction error. The direct forecast is given by $\hat{\pi}_{t+h|t} = \hat{G}_{h,t-R_h+1:t}(\mathbf{x_t})$ where $\hat{G}_{h,t-R_h+1:t}$ is the estimated target function and R_h is the window size. Following Medeiros et al. (2021) we forecast in a *pseudo* out-of-sample fashion using the January 2016 vintage due to data availability.

The narrow replication in Table 1 confirms the findings of Medeiros et al. (2021). For brevity, the complete replication study and discussion are deferred to Appendix A in the Supporting Information.

2 | REPLICATION IN A WIDE SENSE

2.1 | US dataset

The original data is the FRED-MD database of monthly US macroeconomic variables, January 2016 vintage. The number of variables is 122, to which four principal components, four lags of all variables, and four AR terms are added, totalling to 508 predictors. The data covers from January 1960 to December 2015 (672 observations), the out-of-sample part being from January 1990 to December 2015. The first 132 rolling windows each contain 359 observations, and the next 180 rolling windows each contain 491 observations.

2.2 | Extended set of methods

We consider over 30 methods in addition to those considered by Medeiros et al. (2021). See Appendices C and D in the Supporting Information for detailed descriptions of the chosen methods and parameter settings. We group the new methods into the clusters described below.

We consider variations of *RFs*: (honest) local linear forest (LLF) (Friedberg et al., 2020), quantile RF (Athey et al., 2019), extremely randomized trees (ERTs) (Geurts et al., 2006), targeted RF (Borup et al., 2022), macroeconomic RF (Coulombe, 2020) (MacroRF) and tuned RF (Probst et al., 2019).

The *GBT* methods include extreme gradient boosted trees (XGBoost) (Chen & Guestrin, 2016), light gradient boosted machines (LightGBM) with default parameters, ERT (LightGBM ERT) learners and quantile loss (Ke et al., 2017)

PCA calculations that used data up to time $t + h - 1$ when forecasting time $t + h$ inflation at time t. The forked repository is available at https://github.com/EoghanONeill/ForecastingInflation.

[2]The conditional median of Y given X is the 0.5th quantile of the conditional distribution of Y given X. "Quantile forecasts" and acronyms preceded by Q refer to conditional median forecasts throughout.

[3]Tables of results relative to the AR model are available on request.

TABLE 1 Results for the US full sample, CPI index and initial and new sets of methods.

Initial methods	RMSE	MAE	MAD	MCS unif.	MCS avg.	New methods	RMSE	MAE	MAD	MCS unif.	MCS avg.
RW	1.00	1.00	1.00	0.00	0.00	ERT	0.76	0.77	0.70	0.23	0.02
AR	0.85	0.86	0.78	0.00	0.02	Q ERT	0.74	0.74	0.68	0.91	0.27
UCSV	0.85	0.86	0.90	0.00	0.02	LLF	0.80	0.80	0.76	0.00	0.01
LASSO	0.80	0.81	0.76	0.09	0.01	LLF Honest	0.79	0.81	0.80	0.15	0.03
adaLASSO	0.80	0.80	0.78	0.01	0.02	RF	0.77	0.77	0.73	0.05	0.01
ElNet	0.80	0.82	0.76	0.09	0.02	Q RF	0.74	0.74	0.68	0.54	0.11
adaElNet	0.81	0.81	0.77	0.00	0.02	RF Tuned	0.78	0.78	0.73	0.01	0.01
RR	0.86	0.90	0.79	0.05	0.02	Targeted RF	0.80	0.81	0.76	0.04	0.01
BVAR	0.79	0.81	0.79	0.58	0.06	MacroRF	0.79	0.80	0.76	0.00	0.02
Bagging	0.84	0.89	0.94	0.05	0.00	GBM	0.80	0.81	0.81	0.05	0.00
CSR	0.85	0.85	0.84	0.00	0.02	GBM Tuned	0.80	0.80	0.75	0.00	0.01
JMA	0.86	0.92	1.01	0.01	0.00	CatBoost	0.76	0.75	0.75	0.15	0.01
Factor	0.85	0.88	0.88	0.00	0.00	LightGBM	0.79	0.80	0.80	0.01	0.01
T.Factor	0.84	0.89	0.91	0.00	0.00	LightGBM ERT	0.78	0.79	0.80	0.02	0.00
B.Factor	0.86	0.93	1.03	0.01	0.02	Q LightGBM	0.76	0.76	0.72	0.21	0.01
RF	0.77	0.77	0.73	1.00	1.00	CV XGBoost	0.74	0.75	0.73	0.32	0.03
Mean	0.78	0.78	0.77	0.89	0.07	XGBoost	0.76	0.78	0.76	0.22	0.01
T.Mean	0.78	0.78	0.76	0.89	0.12	BART Dirichlet	0.81	0.84	0.85	0.00	0.00
Median	0.78	0.78	0.76	0.84	0.02	BART	0.79	0.83	0.88	0.00	0.00
RF/OLS	0.79	0.81	0.84	0.89	0.06	BART-BMA	0.81	0.85	0.83	0.00	0.00
adaLASSO/RF	0.78	0.79	0.75	0.18	0.04	HBART	0.76	0.78	0.81	0.09	0.01
						MOTR-BART	0.88	0.86	0.85	0.00	0.01
						XBART	0.82	0.88	0.72	0.22	0.03
						RVM	0.82	0.85	0.93	0.32	0.01
						SVR Tuned	0.96	1.13	0.82	0.00	0.00
						SVR-LS	0.79	0.84	0.86	0.20	0.02
						Q SVR-LS	0.77	0.81	0.86	0.44	0.11
						SVR	0.81	0.84	0.90	0.32	0.01
						NN 3 layers	0.85	0.88	0.91	0.00	0.00
						NN 5 layers	0.81	0.85	0.91	0.05	0.00
						NN 8 layers	0.78	0.81	0.87	0.22	0.01
						EN Tuned	0.80	0.82	0.78	0.00	0.01
						RLASSO	0.78	0.79	0.78	0.44	0.01
						Post-RLASSO	0.80	0.81	0.83	0.22	0.01
						Mean	0.74	0.73	0.72	1.00	1.00
						T.Mean	0.74	0.73	0.72	0.44	0.08
						Median	0.74	0.74	0.71	0.44	0.05

Note: The table reports the average RMSE, MAE, MAD and the uniform and average multiple horizon MCS test *p*-values (Quaedvlieg, 2021). SoftBART is not included as it produces an error for at least one window. See Appendix B in the Supporting Information for method descriptions.
Abbreviations: adaElNet, adaptive elastic net; adaLASSO/RF, adaLASSO selection + RF; adaLASSO, adaptive LASSO; B.Factor, boosted factor; Bagging, bagged OLS, variable selection; BVAR, Bayesian vector autoregression; CSR, complete subset regression; ElNet, elastic net; JMA, jackknife model averaging; LASSO, Least absolute selection and shrinkage operator; RF/OLS, RF selection, then OLS; RR, ridge regression; T.Factor, *t*-statistic targeted factor; T.Mean, trimmed mean.

(Q LightGBM), CatBoost (Prokhorenkova et al., 2018) and a tuned and untuned standard gradient boosted machine (GBM) (Friedman, 2001; Lundell, 2023).

BART (Chipman et al., 2010) have been applied to macroeconomic forecasting both in a direct forecasting framework (Behrens et al., 2020; Prüser, 2019) and within vector autoregressions (Huber & Rossini, 2022). We include results for standard BART, BART with a hyperprior on splitting variable probabilities (Dirichlet BART) (Linero, 2018), BART with soft splitting rules (softBART) (Linero & Yang, 2018), BART with model trees (Prado et al., 2021) (MOTR-BART), BART with heteroscedastic errors (HBART) (Pratola et al., 2020), accelerated BART (XBART) (He & Hahn, 2021) and BART with Bayesian model averaging (BART-BMA) (Hernández et al., 2018).

Support vector regression (SVR) is an adaptation of the SVM classification method to the task of predicting continuous outcomes (Drucker et al., 1996). We include results for tuned and untuned SVR, least-squares SVR (Suykens &

Vandewalle, 1999) with squared error (SVR-LS) and quantile loss (Q SVR-LS) and a Bayesian variation known as relevance vector machines (RVM) (Bishop & Tipping, 2013; Tipping, 1999, 2001).

Forecasts are obtained from NNs with a ReLU activation function and three, five or eight layers. Medeiros et al. (2021) implemented a three-layer network.

Medeiros et al. (2021) considered *regularized linear regression* methods LASSO, ridge regression and elastic net. We add to the set of penalized linear regression results by evaluating heteroscedastic LASSO with data-driven penalty (RLASSO), OLS applied to the variables selected by RLASSO (Post-RLASSO) (Belloni et al., 2012) and tuned elastic net (Lundell, 2023). We also present results for the mean, median and 10% trimmed mean (T.Mean) of forecasts from the extended set of methods, as in Medeiros et al. (2021).

2.3 | Results: US data, extended set of methods

Table 1 (right panel) presents the results for the full sample of US data from 1990 to 2015, using our new set of methods. It contains the criteria included by Medeiros et al. (2021): root mean squared error for model m, horizon h, $RMSE_{m,h} = \sqrt{\frac{1}{T-T_0+1}\sum_{t=T_0}^{T}\hat{e}_{t,m,h}^2}$, the mean absolute error, $MAE_{m,h} = \frac{1}{T-T_0+1}\sum_{t=T_0}^{T}|\hat{e}_{t,m,h}|$ and the median absolute deviation from the median error, $MAD_{m,h} = \text{median}[|\hat{e}_{t,m,h} - \text{median}(\hat{e}_{t,m,h})|]$, all relative to RW, where $\hat{e}_{t,m,h}$ denotes the forecast error and T_0 is the first out- of- sample observation. Following Medeiros et al. (2021), these criteria are averaged over all 15 forecasts, including 1 to 12 months horizons and 3, 6 and 12 months accumulated forecast horizons, giving the reported numbers $\frac{1}{15}\sum_{\forall h}RMSE_{m,h}$, $\frac{1}{15}\sum_{\forall h}MAE_{m,h}$ and $\frac{1}{15}\sum_{\forall h}MAD_{m,h}$. The table includes uniform and average multiple horizon model confidence set (MH MCS) p-values (Quaedvlieg, 2021) for the set of forecasts over all 12-month horizons.

The uniform MH MCS test entails superior performance at each horizon. The average MH MCS test allows inferior performance at some horizons to be compensated at other horizons.[4] Let $\boldsymbol{d}_{ij,t}$ denote the vector of loss differences between model i and model j at time t for all horizons, with elements $\boldsymbol{d}_{ij,t}^h$. The null hypothesis of the uniform (average) multihorizon MCS test is that the minimum (arithmetic mean) across horizons of the expected loss differential is zero for all pairs of models in the set. For a set $\mathcal{M}$, $H_{\mathcal{M},uSPA} : \min_h \mathbb{E}(\boldsymbol{d}_{ij,t}^h) \leq 0 \forall i,j \in \mathcal{M}$ or $H_{\mathcal{M},aSPA} : \frac{1}{15}\sum_{h=1}^{15}\mathbb{E}(\boldsymbol{d}_{ij,t}^h) \leq 0 \forall i,j \in \mathcal{M}$.[5]

In terms of RMSE, MAE and MAD averaged over all months and accumulated forecasts, median forecasts produced by CV XGboostXGBoost, Q ERT and Q RF perform best. The uniform and average MH MCS p-values suggest that the mean forecast combination, Q ERT and the Q RF produce the most accurate forecasts. Some methods in all clusters are competitive with RF. HBART performs well in terms of 95% forecast interval coverage and width. Prediction interval results are presented in Table 4 and Appendices E.3 and E.4 in the Supporting Information. RMSE, MAE and MCS p-values for each horizon (Hansen et al., 2011) are reported in Appendices E.1 and E.2 in the Supporting Information. Q RF is more accurate than RF at all horizons except $h = 1$ (Table 4). There are smaller gains relative to a RW for the first horizon than for other horizons.

We examine variable importance using Shapley values (Lundberg & Lee, 2017).[6] Let S denote a subset of all covariates, F, excluding ℓ ($S \subseteq F \backslash \{\ell\}$). The Shapley value of variable ℓ for observation i is a weighted average, over possible subsets S, of differences in forecast values with and without the variable ℓ added to the set S, denoted by $f_{S\cup\{\ell\}}(\boldsymbol{x}_{S\cup\{\ell\},i})$ and $f_S(\boldsymbol{x}_{S,i})$ respectively. i.e. $\phi_{shap}(\ell,\boldsymbol{x}_i) = \sum_{S\subseteq F\backslash\{\ell\}} \frac{|S|(|F|-|S|-1)}{|F|}(f_{S\cup\{\ell\}}(\boldsymbol{x}_{S\cup\{\ell\},i}) - f_S(\boldsymbol{x}_{S,i}))$. The global Shapley value is defined as $\frac{1}{N}\sum_{i=1}^{N}\phi_{shap}(\ell,\boldsymbol{x}_i)$.[7]

Figure 1 displays the variables with the highest absolute Shapley values for RF, LightGBM and XGBoost for $h = 1$ and $h = 12$, and the subperiods defined by Medeiros et al. (2021). For $h = 1$, CPI and CPI Less Medical Items are two of the most important variables. Other important variables include lags of price indices and interest rate variables. For $h = 12$, more importance is given to variables pertaining to the labour and housing markets, including all employees: financial

[4]MCS results presented in this paper involve squared losses. See Appendix E.2 in the Supporting Information for implementation details.

[5]We do not report arithmetic averages of p-values across horizons, as reported by Medeiros et al. (2021), because the p-values pertain to tests of different hypotheses. Moreover, a simple arithmetic average of p-values without scaling or transformation is possibly erroneous (Birnbaum, 1954; Vovk & Wang, 2020).

[6]We implemented Shapley values using the `R` package `fastshap` and 100 samples of covariate subsets. RF, LightGBM and XGBoost are among the best performing methods for which Shapely values can be obtained from the package.

[7]The Shapley value is estimated by sampling S and plugging in estimates $\hat{f}_{S\cup\{\ell\}}(\boldsymbol{x}_{S\cup\{\ell\},i})$ and $\hat{f}_S(\boldsymbol{x}_{S,i})$. For an alternative approach, see Buckmann et al. (2022).

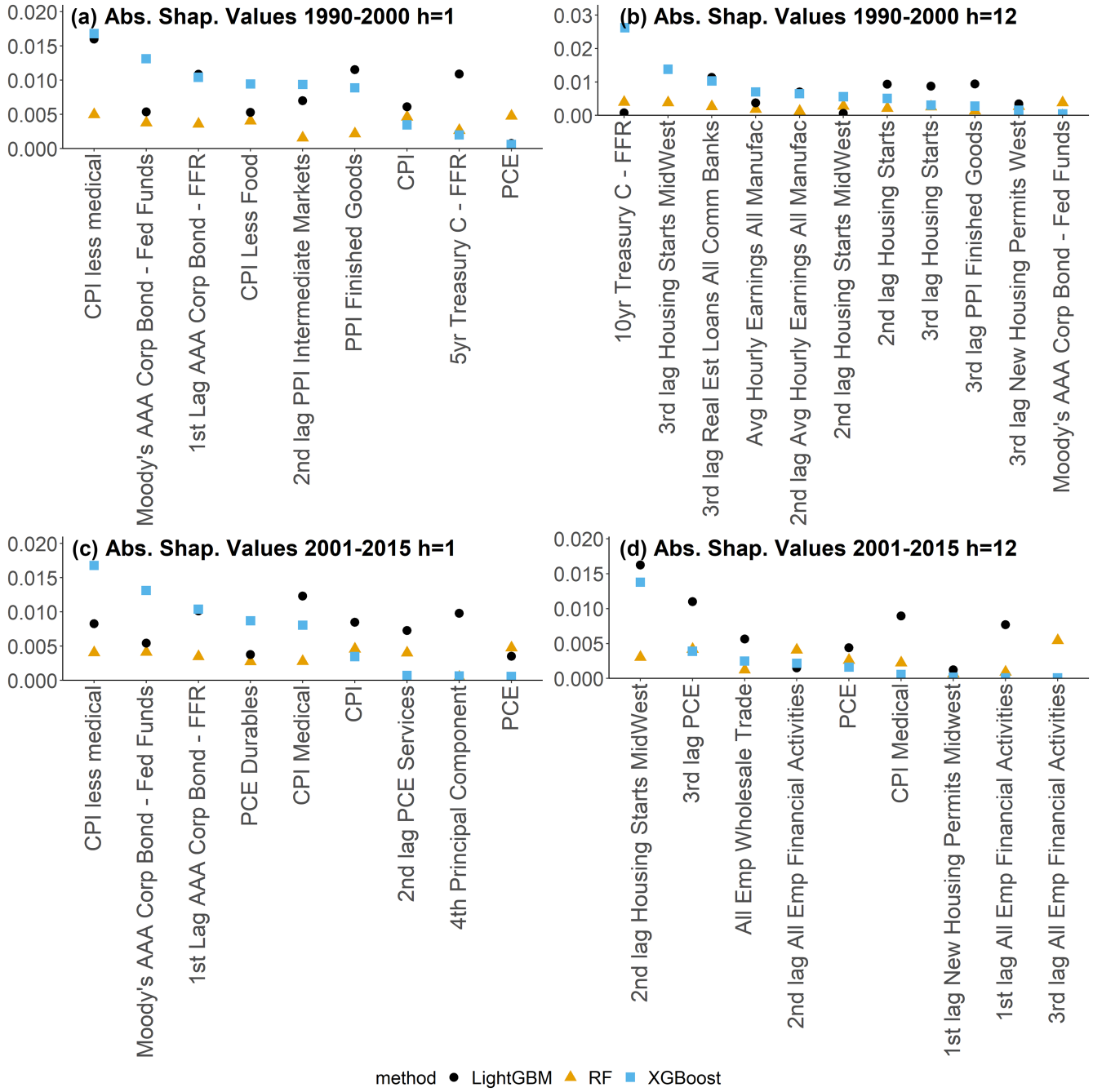

FIGURE 1 Absolute Shapley values for RF, LightGBM and XGBoost. US data, subsamples 1990–2000 and 2001–2015, horizons $h = 1$ and $h = 12$. The graphs include the union across each method of the variables with the five highest absolute Shapley values.

activities and housing start variables. These results are in line with economic intuition, which strengthens trust in the interpretability of the models.

2.4 | Canada and UK datasets

We consider a UK dataset (with a complete data description given in tab. A7 in Goulet Coulombe et al., 2021) and a Canadian dataset (with a detailed description given in Fortin-Gagnon et al., 2018) of monthly data, both of which are comparable in size and structure to the FRED-MD data.

The Canadian data features 114 macroeconomic variables recorded from January 1981 to February 2022 and includes four principal component factors, four lags of all variables and four AR terms of the dependent variable CPI. The

TABLE 2 Results for Canadian, UK and recent US data, CPI index and initial set of methods.

	Canada 2011–2022				UK 2010–2021				US 2016–2022			
	RMSE	**MAE**	**MCS unif.**	**MCS avg.**	**RMSE**	**MAE**	**MCS unif.**	**MCS avg.**	**RMSE**	**MAE**	**MCS unif.**	**MCS avg.**
RW	1.00	1.00	0.59	0.00	1.00	1.00	0.41	0.44	1.00	1.00	0.60	0.09
AR	0.74	0.73	0.74	0.01	0.84	0.83	0.41	0.48	0.91	0.91	0.85	0.09
UCSV	0.79	0.79	0.93	0.01	1.13	1.11	0.41	0.44	0.86	0.85	0.94	1.00
LASSO	0.73	0.73	0.59	0.01	1.49	1.27	0.41	0.48	0.91	0.89	0.87	0.09
adaLASSO	0.72	0.72	0.59	1.00	1.23	1.09	0.41	0.48	0.88	0.87	1.00	0.30
ElNet	0.73	0.74	0.47	0.01	0.96	0.92	0.41	0.39	0.91	0.88	0.82	0.09
adaElNet	0.72	0.73	[a]		0.86	0.84	0.41	0.48	0.89	0.87	0.94	0.26
RR	0.77	0.75	0.59	0.01	0.91	0.84	0.41	0.48	0.98	0.97	0.69	0.09
BVAR	0.74	0.73	0.93	0.01	0.81	0.79	0.67	0.48	0.97	0.90	0.40	0.09
Bagging	0.82	0.78	0.59	0.01	[b]				1.31	1.12	0.40	0.09
CSR	0.72	0.71	0.93	0.30	0.81	0.80	0.41	0.48	0.92	0.89	0.82	0.09
JMA	1.17	1.08	0.59	0.01	2.52	1.90	0.30	0.47	1.71	1.41	0.40	0.09
Factor	0.77	0.75	0.59	0.01	0.93	0.87	0.41	0.43	0.95	0.93	0.82	0.09
T.Factor	0.83	0.83	0.59	0.00	1.20	0.97	0.16	0.48	0.97	0.95	0.40	0.09
B.Factor	0.74	0.75	0.66	0.01	0.89	0.85	0.41	0.48	0.96	0.93	0.69	0.09
RF	0.72	0.71	0.93	0.35	0.80	0.78	1.00	1.00	0.98	0.96	0.69	0.09
Mean	0.72	0.72	0.92	0.01	0.85	0.82	0.41	0.48	0.90	0.87	0.69	0.09
T.Mean	0.72	0.72	1.00	0.23	0.83	0.80	0.41	0.48	0.89	0.87	0.94	0.45
Median	0.72	0.72	1.00	0.13	0.81	0.79	0.92	0.48	0.90	0.87	0.87	0.10
RF/OLS	0.72	0.72	0.93	0.40	0.84	0.80	0.41	0.47	0.96	0.92	0.40	0.09
adaLASSO/RF	0.78	0.78	0.59	0.00	0.83	0.82	0.41	0.48	0.90	0.79	0.69	0.09

Note: The table reports the average RMSE, MAE and the uniform and average multiple horizon MCS test *p*-values (Quaedvlieg, 2021).

[a] Adaptive elastic net produced the same forecasts as elastic net for some horizons. Therefore, the MCS test produced an error.

[b] The `HDeconometrics` package bagging function produced the error "`Error: The pre-testing was not able to reduce the dimension to N<T.`"

out-of-sample period is March 2011–February 2022, and it is predicted by 132 rolling windows. The UK data contain 110 macroeconomic variables combined with the same generic predictors for the time period January 1998–July 2021. The predicted variable for the UK is CPI index and the out-of-sample period consists of 132 rolling windows from August 2010 to July 2021.

2.5 | Results: Canada and UK and initial set of methods

Table 2 presents loss measures and MCS test statistics for the models considered by Medeiros et al. (2021) applied to the Canadian and UK inflation data.[8] Overall, the results for Canadian inflation challenge some of the main findings of Medeiros et al. (2021) for the US data. First, when comparing the complete forecasting results for Canada with those for the United States, we find that the set of proposed methods again outperforms the benchmarks, although not as consistently as in the US case. For instance, the AR model's loss values and UCSV's MCS test results are quite comparable to those of most alternative methods.

Second, while RF is among the best performing methods, it is outperformed by the adaptive LASSO in terms of average MCS test results. Medeiros et al. (2021) also state that the shrinkage methods are the most serious competitors to RF. Furthermore, we observe that contrary to the findings of Medeiros et al. (2021), among shrinkage methods, sparse methods outperform RR, indicating that a subset of variables are predictive of inflation.

Third, the RF/OLS combination achieves better (average MCS test) or equal (RMSE) predictive ability in comparison to RF. Thus, we cannot state that the relationship between Canadian inflation and its predictors is highly nonlinear and, following the reasoning of Medeiros et al. (2021), RF's performance might be mostly the result of variable selection. CSR and the three forecast combination methods also perform well.

Turning now to the UK data, we note that Table 2 confirms most of the findings of Medeiros et al. (2021): RF performs best, although the MCS test does not reject any models at the 10% level.[9] In addition, RF performs better then both RF/OLS and adaLASSO/RF combinations, which supports the thesis that RF produces more accurate predictions than other methods because it both detects nonlinearities and selects important variables. However, we notice that for many methods, we do not reject the MCS test null hypothesis at the 5% level, which is not the case for the US inflation forecasts. Therefore, we cannot conclude that RF is the best method with the same level of confidence as for the US data.

2.6 | Results: Canada and UK, extended set of methods

The results obtained with the new set of machine learning methods applied to the Canadian CPI inflation dataset are presented in the left panel of Table 3, and the results for UK CPI are in the middle panel. The results confirm that forecasts of conditional median inflation are competitive with conditional mean forecasts. The MCS test p-values generally favour quantile-based forest methods over standard forest methods.[10]

For Canada, GBT methods show considerable improvement relative to RF, with CatBoost, LightGBM ERT, and CV XGBoost being superior in all five measures. For the UK, Quantile LightGBM, and Quantile ERT provide competitive forecasts according to multiple measures. The forecast combination methods generally outperform individual methods. No method outperforms RF by the same extent as for the US dataset.

2.7 | Extended US sample

We now extend the results for the US data to October 2022. Results for 2016–2022 are included for the initial methods in Table 2 and for the new methods in Table 3. Appendix F in the Supporting Information contains tables of results for the full extended sample 1990–2022, while Appendix G in the Supporting Information contains results for many sub-periods. For 1990–2022, the general pattern of results is similar to that observed in all the tables above. RF is the best of the initial

[8]MCS test results by horizon are in Appendix E.2 in the Supporting Information. The RMSE and MAE for each horizon are presented in Appendix E.1 in the Supporting Information.

[9]The MCS test results by horizon are presented in Appendix E.2 in the Supporting Information.

[10]The RMSE, MAE and MCS test results by horizon for the new methods applied to the Canadian and UK datasets are in Appendices E.1 and E.2 in the Supporting Information. Similar conclusions can be made from these tables.

TABLE 3 Results for Canadian, UK and recent US data, CPI index and new set of methods.

	Canada 2011–2022				UK 2010–2021				US 2016–2022			
	RMSE	MAE	MCS unif.	MCS avg.	RMSE	MAE	MCS unif.	MCS avg.	RMSE	MAE	MCS unif.	MCS avg.
ERT	0.71	0.71	0.84	0.19	0.81	0.78	0.70	0.32	0.89	0.86	0.98	0.61
Q ERT	0.71	0.71	0.84	0.29	0.80	0.78	0.98	1.00	0.87	0.84	0.94	0.61
LLF	0.71	0.70	0.96	0.29	0.85	0.82	0.01	0.32	1.01	0.91	0.04	0.61
LLF Honest	0.73	0.73	0.16	0.17	0.86	0.84	0.56	0.32	0.90	0.86	0.97	0.61
RF	0.72	0.71	0.10	0.20	0.80	0.78	0.82	0.32	0.98	0.96	0.21	0.61
Q RF	0.71	0.71	0.66	0.29	0.80	0.78	0.92	0.32	0.89	0.85	0.98	0.61
RF Tuned	0.72	0.71	0.08	0.20	0.80	0.79	0.98	0.32	0.93	0.91	0.17	0.61
Targeted RF	0.73	0.72	0.84	0.27	0.81	0.81	0.92	0.15	0.89	0.87	0.98	0.61
MacroRF	0.87	0.78	0.00	0.29	0.85	0.83	0.16	0.01	0.96	0.92	0.15	0.61
GBM	0.72	0.71	0.84	0.20	0.84	0.83	0.17	0.01	0.99	0.95	0.21	0.61
GBM Tuned	0.71	0.70	0.82	0.29	0.83	0.81	0.17	0.28	0.98	0.94	0.14	0.61
CatBoost	0.70	0.70	0.96	0.29	0.81	0.79	0.73	0.32	0.86	0.84	0.99	0.82
LightGBM	0.72	0.70	0.84	0.29	0.84	0.83	0.26	0.09	0.97	0.95	0.89	0.61
LightGBM ERT	0.71	0.69	0.96	0.29	0.83	0.82	0.26	0.08	0.91	0.89	0.98	0.54
Q LightGBM	0.71	0.70	0.88	0.29	0.80	0.78	1.00	0.32	0.86	0.83	1.00	1.00
CV XGBoost	0.71	0.69	0.96	0.29	0.80	0.79	0.63	0.32	0.89	0.87	0.98	0.61
XGBoost	0.72	0.71	0.84	0.29	0.86	0.81	0.16	0.32	0.95	0.92	0.21	0.61
BART Dirichlet	0.75	0.75	0.29	0.03	0.83	0.82	0.26	0.03	0.95	0.92	0.21	0.44
BART	0.73	0.72	0.29	0.02	0.82	0.79	0.17	0.15	0.98	0.94	0.21	0.56
BART-BMA	0.79	0.79	0.07	0.01	0.92	0.94	0.16	0.00	1.03	0.98	0.06	0.56
HBART	0.71	0.70	0.84	0.29	0.81	0.78	0.63	0.32	0.91	0.88	0.98	0.61
MOTR-BART	0.78	0.74	0.01	0.20	0.94	0.87	0.01	0.32	1.31	1.06	0.04	0.61
SoftBART	0.72	0.70	0.84	0.24	0.81	0.78	0.73	0.32	0.90	0.89	0.98	0.61
XBART	0.80	0.80	0.29	0.20	1.11	1.18	0.26	0.32	1.19	1.18	0.21	0.61
RVM	0.81	0.81	0.09	0.10	0.98	1.02	0.16	0.07	1.21	1.20	0.21	0.58
SVR Tuned	0.74	0.75	0.71	0.20	0.86	0.84	0.56	0.32	1.00	1.00	0.21	0.58
SVR-LS	0.74	0.75	0.29	0.18	0.87	0.85	0.16	0.26	1.09	1.09	0.11	0.37
Q SVR-LS	0.74	0.74	0.29	0.20	0.86	0.84	0.26	0.30	1.07	1.04	0.07	0.54
SVR	0.83	0.79	0.02	0.23	0.91	0.82	0.01	0.32	1.15	1.00	0.04	0.61
NN 3 layers	0.92	0.91	0.06	0.08	0.99	0.96	0.01	0.09	1.17	1.08	0.06	0.55
NN 5 layers	0.87	0.84	0.03	0.13	0.93	0.89	0.01	0.28	1.11	1.05	0.06	0.61
NN 8 layers	0.87	0.86	0.06	0.11	0.89	0.88	0.15	0.29	1.17	1.09	0.04	0.61
EN Tuned	0.73	0.72	0.72	0.22	0.86	0.82	0.01	0.32	0.94	0.90	0.06	0.61
RLASSO	0.72	0.72	0.71	0.26	0.83	0.81	0.16	0.30	0.87	0.86	0.98	0.61
Post-RLASSO	0.73	0.72	0.02	0.26	0.84	0.82	0.01	0.32	0.88	0.88	0.98	0.61
Mean	0.70	0.70	1.00	0.29	0.80	0.77	0.26	0.32	0.91	0.87	0.21	0.61
T.Mean	0.70	0.70	0.96	1.00	0.80	0.77	0.70	0.32	0.90	0.87	0.71	0.61
Median	0.70	0.70	0.96	0.29	0.80	0.77	0.98	0.32	0.89	0.86	0.95	0.61

Note: The table reports the average RMSE, MAE and the uniform and average multiple horizon MCS test *p*-values (Quaedvlieg, 2021).

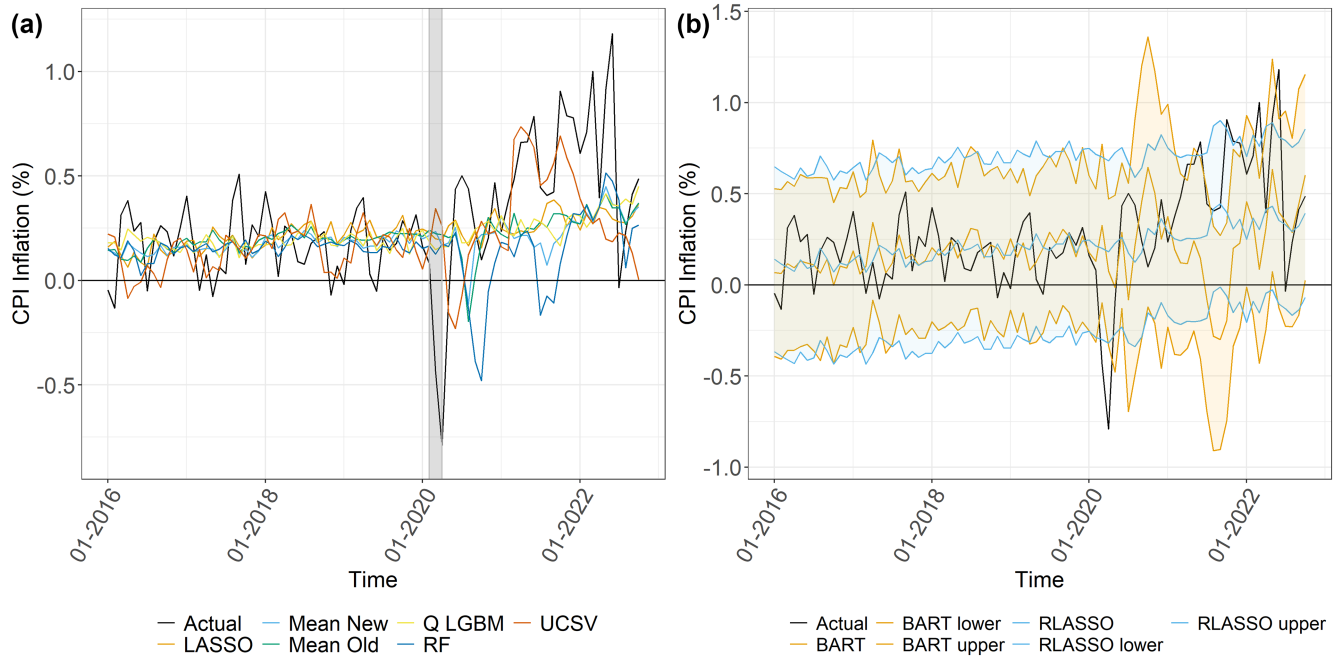

FIGURE 2 CPI monthly inflation and forecasts $h = 4$, US 2016–2022. (a) RF, LASSO, UCSV, Q LightGBM, mean of forecasts from initial (Mean Old) and extended (Mean New) sets of methods. (b) Forecasts with 95% forecast intervals for BART and RLASSO.

TABLE 4 Results by horizon, US 1990–2015, CPI index and selected methods.

													Accumulated		
RMSE	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**3m**	**6m**	**12m**
RF	0.85	0.75	0.72	0.75	0.73	0.74	0.74	0.73	0.74	0.78	0.80	0.73	0.76	0.80	0.94
Q RF	0.89	0.73	0.71	0.73	0.70	0.71	0.72	0.71	0.72	0.76	0.77	0.69	0.75	0.75	0.84
Mean	0.83	0.74	0.71	0.73	0.71	0.72	0.72	0.70	0.71	0.75	0.76	0.68	0.74	0.74	0.81
95% Coverage															
BART	0.932	0.97	0.947	0.97	0.955	0.97	0.962	0.962	0.962	0.955	0.97	0.947			
RLASSO	0.962	0.97	0.97	0.977	0.977	0.97	0.985	0.985	0.992	0.985	0.977	0.985			

Note: RMSE relative to random walk for all 12 horizons and the three accumulated forecasts and 95% interval coverage.

set of methods, whereas among the extended set of methods some boosted tree and quantile-based methods perform best. Overall trimmed means of forecasts from all methods are most accurate.

When we compare the performance of methods from 2016 onward, we obtain very different conclusions. The best performing method from the initial set is UCSV, while RF forecasts are somewhat disappointing. A small number of new methods CatBoost, Q LightGBM, Q ERT and RLASSO are competitive with UCSV. Analysis of subperiods reveals that some new methods strongly outperform UCSV until 2020, and then from 2020 onward, UCSV is the most accurate method by a large margin. We caution that the sample sizes for these subperiods are relatively small, resulting in low statistical power.

Figure 2 displays the 4- month- ahead forecasts and month-on-month CPI inflation for a range of methods in 2016–2022. It can be observed that no method predicted the temporary deflation at the beginning of the coronavirus pandemic, and after this period, most methods failed to accurately forecast the notable rise in inflation, with RF performing particularly poorly. UCSV is an exception, which produced relatively accurate inflation forecasts throughout 2021. Figure 2 also displays 95% forecast intervals for two of the methods with the best coverage, RLASSO and BART. The intervals are accurate (see Table 4), although both methods unsurprisingly fail to predict the deflation at the beginning of the coronavirus pandemic, and BART's intervals fluctuate more in 2020 and 2021. Appendix E.3 in the Supporting Information includes prediction interval results for all datasets and subsamples.[11]

[11] Appendices H and I in the Supporting Information contain similar figures for $h = 1, 4, 12$.

3 | CONCLUSION

We have reproduced the forecasting exercise of Medeiros et al. (2021) and confirmed that RF outperforms standard benchmark methods. This is in agreement with the finding that RF is a strong benchmark across a range of tasks (Fernández-Delgado et al., 2014; Grinsztajn et al., 2022). The claimed superiority of RF is also confirmed to an extent by the UK data. The Canadian data results cast some doubt on the generalizability of this claim. The accuracy of the Canadian data RF predictions appears to be the result of variable selection and not detection of nonlinearities. We also find that other tree-based methods are competitive with RF. Furthermore, ML methods that forecast conditional median inflation can produce more accurate forecasts. Generally, the best methods are simple forecast combinations of many machine learning methods. In the extended US data, we find similar results to Medeiros et al. (2021) up to January 2020. After this point, RF performs poorly, while UCSV and some boosted tree methods are most accurate.

OPEN RESEARCH BADGES

This article has been awarded Open Data Badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. Data is available at https://doi.org/10.15456/jae.2023340.1218742483.

DATA AVAILABILITY STATEMENT

The data is available at http://dx.doi.org/10.15456/jae.2023340.1218742483.

ORCID

Eoghan O'Neill https://orcid.org/0000-0002-1274-4248
Martina Danielova Zaharieva https://orcid.org/0000-0002-5776-7013

REFERENCES

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148–1178.
Behrens, C., Pierdzioch, C., & Risse, M. (2020). Do German economic research institutes publish efficient growth and inflation forecasts? A Bayesian analysis. *Journal of Applied Statistics*, *47*(4), 698–723.
Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, *80*(6), 2369–2429.
Birnbaum, A. (1954). Combining independent tests of significance. *Journal of the American Statistical Association*, *49*(267), 559–574.
Bishop, C. M., & Tipping, M. (2013). Variational relevance vector machines. arXiv preprint arXiv:1301.3838.
Borup, D., Christensen, B. J., Muhlbach, N. S., & Nielsen, M. S. (2022). Targeting predictors in random forest regression. *International Journal of Forecasting*, *39*, 841–868.
Buckmann, M., Joseph, A., & Robertson, H. (2022). An interpretable machine learning workflow with an application to economic forecasting: Bank of England Technical report.
Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, pp. 785–794.
Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, *4*(1), 266–298.
Coulombe, P. G. (2020). The macroeconomy as a random forest. arXiv preprint arXiv:2006.12724.
Drucker, H., Burges, C. J., Kaufman, L., Smola, A., & Vapnik, V. (1996). Support vector regression machines. *Advances in Neural Information Processing Systems*, *9*, 155–161.
Faust, J., & Wright, J. H. (2013). Forecasting inflation, *Handbook of economic forecasting*, Vol. *2*: Elsevier, pp. 2–56.
Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, *15*(1), 3133–3181.
Fortin-Gagnon, O., Leroux, M., Stevanovic, D., & Surprenant, S. (2018). A large Canadian database for macroeconomic analysis: Document de travail Technical report.
Friedberg, R., Tibshirani, J., Athey, S., & Wager, S. (2020). Local linear forests. *Journal of Computational and Graphical Statistics*, *30*(2), 503–517.
Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*, 1189–1232.
Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42.
Goulet Coulombe, P., Marcellino, M., & Stevanovi, D. (2021). Can machine learning catch the Covid-19 recession? *National Institute Economic Review*, *256*, 71–109.
Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data?. *Advances in Neural Information Processing Systems*, *35*, 507–520.

Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*(2), 453–497.
He, J., & Hahn, P. R. (2021). Stochastic tree ensembles for regularized nonlinear regression. *Journal of the American Statistical Association*, *118*, 551–570.
Hernández, B., Raftery, A. E., Pennington, S. R., & Parnell, A. C. (2018). Bayesian additive regression trees using Bayesian model averaging. *Statistics and Computing*, *28*(4), 869–890.
Huber, F., & Rossini, L. (2022). Inference in Bayesian additive vector autoregressive tree models. *The Annals of Applied Statistics*, *16*(1), 104–123.
Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*, 3146–3154.
Linero, A. R. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association*, *113*(522), 626–636.
Linero, A. R., & Yang, Y. (2018). Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *80*(5), 1087–1110.
Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 4765–4774.
Lundell, J. (2023). EZtune: A package for automated hyperparameter tuning in R. arXiv preprint arXiv:2303.12177.
McCracken, M. W., & Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, *34*(4), 574–589.
Medeiros, M. C., Vasconcelos, G. F. R., Veiga, A., & Zilberman, E. (2021). Forecasting inflation in a data-rich environment: The benefits of machine learning methods. *Journal of Business & Economic Statistics*, *39*(1), 98–119.
Prüser, J. (2019). Forecasting with many predictors using Bayesian additive regression trees. *Journal of Forecasting*, *38*(7), 621–631.
Prado, E. B., Moral, R. A., & Parnell, A. C. (2021). Bayesian additive regression trees with model trees. *Statistics and Computing*, *31*(3), 1–13.
Pratola, M. T., Chipman, H. A., George, E. I., & McCulloch, R. E. (2020). Heteroscedastic BART via multiplicative regression trees. *Journal of Computational and Graphical Statistics*, *29*(2), 405–417.
Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(3), e1301.
Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, *31*, 6638–6648.
Quaedvlieg, R. (2021). Multi-horizon forecast comparison. *Journal of Business & Economic Statistics*, *39*(1), 40–53.
Stock, J. H., & Watson, M. W. (2007). Why has U.S. inflation become harder to forecast? *Journal of Money, Credit and Banking*, *39*, 3–33.
Suykens, J. A. K., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, *9*(3), 293–300.
Tipping, M. (1999). The relevance vector machine. *Advances in Neural Information Processing Systems*, *12*, 652–658.
Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, *1*(Jun), 211–244.
Vovk, V., & Wang, R. (2020). Combining p-values via averaging. *Biometrika*, *107*(4), 791–808.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of the article.

How to cite this article: Naghi A. A., O'Neill E., & Zaharieva M. D. (2024). The benefits of forecasting inflation with machine learning: New evidence. *Journal of Applied Econometrics*, *39*(7), 1321–1331. https://doi.org/10.1002/jae.3088