ORIGINAL PAPER

Short term forecasting of base metals prices using a LightGBM and a LightGBM - ARIMA ensemble

Konstantinos Oikonomou[1] · Dimitris Damigos[1]

Abstract

Base metals are key materials for various industrial sectors such as electronics, construction, manufacturing, etc. Their selling price is important both for the profitability of the mining and metallurgical companies that produce and trade them, as well as for the countries whose economies rely on their exports or tax revenues as a means for national income. Prices are also critical for companies that use base metals as inputs to fabricate end products. The prediction of prices' future movements can serve as a tool for risk mitigation and better budget planning. In this study, the logarithmic returns of base metals are forecasted using an autoregressive Light Gradient Boosting Machine (LightGBM) as well as an ensemble comprising the aforementioned algorithm and a classical time series forecasting model (i.e., ARIMA). The two models are then compared to three simpler benchmark models, namely a global mean model, an exponential smoothing model and an ARIMA model. When comparing using RMSE, the autoregressive LightGBM model outperformed the three univariate benchmark models (and the ensemble) for forecasting 6 months ahead for aluminum and nickel returns, while copper and zinc returns were forecasted better by the ensemble. Neither of the proposed models performed better than an ARIMA model when it comes to forecasting lead and tin returns.

Keywords Base metals · Prices · Forecasting · Gradient boosting · ARIMA

Introduction

Due to their diverse properties, base metals and their alloys have been extensively used in a multitude of applications. For example, aluminum is used in applications ranging from modern aeronautics to packaging, while copper is used in sectors spanning from electronics to architecture. Among others, nickel is used as an alloying element in steels, zinc as a galvanizing agent, and lead-tin alloys are used for soldering applications. Moreover, non-ferrous metals play a crucial role to decarbonisation and clean energy transition, as they are essential components in a wide array of related technologies, ranging from electric vehicles (Dhar et al. 1997; Wang et al. 2023a) and solar/wind based power generators (Lacal-Arántegui 2015) to modern batteries (Zhou et al. 2022) and energy storage systems. But, precisely because of their infiltration into modern technology and infrastructure, base metals remain critical elements for their preservation and further development, as confirmed by the US Department of Energy in their critical materials assessment (Bauer et al. 2023), which in turn spurs their future demand (Backman 2008; Watari et al. 2021).

The prices of base metals are forged in exchange markets such as the London Metal Exchange (LME) and are mainly determined by market dynamics (Dooley and Lenihan 2005). On the supply side, disruptions in the supply chain, such as natural disasters or commodity-specific unforeseen events, may cause an increase in base metals prices, given the same level of demand. On the demand side, in times of economic downturn and reduced overall industrial production, the need for base metals may be reduced, leading to a price decrease. Due to this sensitivity to market dynamics and global economic conditions, base metals prices have historically exhibited periods of volatility, especially during major socio-economic global events (e.g., the mortgage crisis in the US and the COVID-19 pandemic) (Ahmed and Sarkodie 2021; M.-H. Chen 2010).

✉ Konstantinos Oikonomou
koikonomou@metal.ntua.gr

1   School of Mining & Metallurgical Engineering, National Technical University of Athens, Zografou Campus, 9, Iroon Polytechniou str, 15780 Athens, Greece

Base metals prices and the volatility they exhibit are critical for many of those involved in their production, consumption, and trading, and the knowing, or at least having an estimate, of their future behaviour in advance is of utmost importance. First, countries with ore-rich deposits base a large part of their annual income on exports or tax revenues of these primary base metals (Alam et al. 2022; Sánchez Lasheras et al. 2015). Zambia (Chipili 2016), Chile (Medina and Soto 2007), and Peru (Lust 2019) have been reliant on copper, Mozambique on aluminum (Castel-Branco and Goldin 2003) and Indonesia on nickel (Krustiyati et al. 2022), just to name a few. Likewise, mining and metallurgical companies anticipate sales prices sufficient to cover the initial investment and generate profits, particularly when the primary production of base metals requires substantial capital and operating expenditures for exploration or operational activities (Dooley and Lenihan 2005; Du et al. 2020). Furthermore, the prices of base metals are critical for industries that use these goods as inputs in their production processes, as the cost of purchasing raw materials is proportional to them (Rossen 2015). Last, base metals prices are important to institutional investors that may enrich their portfolio with base metals futures (Watkins and McAleer 2004).

Hence, accurate forecasts can be used by governments of countries whose revenues are heavily dependent on the export of these metals to better plan their country's budget and fiscal policies. For instance, in cases where prices are temporarily at a higher level but are projected to fall, governments may decide to save the surplus revenues to support economic activity when this occurs (Kabundi et al. 2022). In addition, accurate forecasts can prevent phenomena whereby over-optimism about future prices of a commodity can lead to over-taxation of the industry that may result in its contraction in the long term (Radetzki & Wårell, 2020). Similarly, accurate price forecasts are important for mining and metallurgical companies to better estimate future revenues, as prices directly affect them. Thus, in cases where prices are expected to fall, these companies may decide to implement cost reduction policies, control their debt levels (MacDiarmid et al. 2018), or contemplate investment decisions (Foo et al. 2018). As for the companies using base metals for their end products, accurate forecasts can facilitate more precise forward cost estimation, sound inventory management, and better overall production planning. Finally, investors may use forecasts either for speculative purposes, portfolio optimization, or risk mitigation.

Due to the wide range of factors that can affect the dynamics of the base metals market, forecasting their prices is challenging, especially in the long term (Dooley and Lenihan 2005). Both univariate and multivariate forecasting methods have been proposed in the literature, with mixed results. Examples of univariate methods include work from Dooley and Lenihan 2005 that employed an ARIMA model to forecast zinc and lead prices and found that it performed marginally better than a lagged forward forecasting method. Kahraman and Akay 2023 experimented in forecasting base metals prices with variations of models belonging to the exponential smoothing family and Sánchez Lasheras et al. 2015 used artificial neural networks to forecast copper spot prices, showing better results compared to the ones obtained by a standalone ARIMA model. Kriechbaumer et al. 2014 used wavelet decomposition in combination with ARIMA to show significantly better results in base metals prices forecasting compared to an ARIMA benchmark. Chen et al. 2016 examined the forecasting performance of a modified grey wave method when forecasting aluminum and nickel prices and found that it outperformed the random walk and the ARMA benchmarks. In addition, univariate forecasting work has been extended to ferrous markets, for example Xu and Zhang 2023 utilized Gaussian process regressions to forecast price indices for ten major steel products in China, outperforming traditional econometric models and machine learning approaches.

Examples of the use of multivariate models are the work of Liu et al. 2017 who, by using the price lags of other commodities and indices as explanatory variables in regression trees, created forecasts at different horizons for copper prices, or the work of Díaz et al. 2020 who used the same variables in more complex models such as random forests and gradient boosted trees and achieved better results than simple decision trees though without being competitive to a random walk baseline for short and medium forecasting horizons. Khoshalan et al. 2021 used a number of parameters such as the prices of aluminum, crude oil, gold etc., as inputs into different artificial intelligent models in order to forecast copper prices and concluded that a neural network model outperformed the rest of the proposed models. Finally, Pincheira Brown and Hardy 2019 used Chilean exchange rates to forecast LME returns of base metals.

The purpose of our study is twofold. First, to fill the gap found in the literature regarding pure autoregressive tree-based algorithms used to short-term forecast base metals prices and their combination with classical time series models. This will be done by investigating the performance of an autoregressive LightGBM (henceforth, AutoRegLightGBM), trained to only use lags of the time series to produce forecasts both as a standalone and as part of an ensemble. We chose LightGBM, an algorithm that leverages gradient boosting trees to approximate (possibly non-linear) temporal relationships present in the time series, motivated by the fact that this algorithm has proved its performance in both classification and regression tasks for various problems (Rufo et al. 2021; Tang et al. 2020; Wang et al. 2022), but most importantly because it was the algorithm that won

the M5 forecasting competition (Makridakis et al. 2022). In addition, we combine its forecasts with the ones of an ARIMA model, a model that is able to produce forecasts by approximating linear temporal dependencies in a time series. Ensembling models that employ diverse methods to approximate the underlying structure of a time series has been found to be a successful strategy (Bates and Granger 1969; Petropoulos and Svetunkov 2020), and this can be attributed to the fact that this approach reduces the uncertainty arising from model selection and parameter estimation (X. Wang et al. 2023a). Second, in contrast to the research done so far, we evaluate the forecasting accuracy through a more sophisticated evaluation methodology (i.e., evaluation on a rolling forecasting origin) that is more robust compared to the single train-test split approach. This is because the single train-test split approach may present good results just for a single validation period, and the model may not generalize well for other periods (Tashman 2000). We show that when comparing RMSE scores, the AutoRegLightGBM exhibited better performance in forecasting aluminum and nickel returns 6 months ahead. In addition, the ensemble approach demonstrated better accuracy for copper and zinc returns as it outperformed the global mean, the exponential smoothing, the ARIMA model, and the AutoRegLightGBM. Neither of the proposed methods performed better than the ARIMA benchmark when forecasting lead and tin returns.

Methods

Data collection and transformations

Monthly LME price data were obtained from the World Bank database (https://www.worldbank.org/en/research/commodity-markets) for the following commodities: aluminum, copper, lead, tin, nickel and zinc, from January 1990 to August 2023. The timeseries were transformed into log returns using the formula:

$$r_t = ln\left(\frac{p_t}{p_{t-1}}\right)$$

Where p_t is the price of a given base metal at time t and r_t is the return at time t. Log transforming the initial time series is important for our analysis as it stabilizes the variance. In addition, converting the prices into returns detrends the series. Both transformations were chosen in order to induce stationarity, which is essential not only when applying ARIMA models but also when using tree-based algorithms for time series forecasting due to their inherent incapacity to extrapolate trends beyond the training period (Joseph

2022). This resulted in each time series having a length of 403 observations.

Global mean, exponential smoothing, ARIMA, LightGBM and Ensemble

In the present study, 5 models will be used to produce point forecasts of the returns of the base metals for 6 months ahead. The first three, namely global mean, exponential smoothing, and ARIMA will serve as the benchmark models, and the other two, the AutoRegLightGBM and the AutoRegLightGBM-ARIMA ensemble, will be compared to them.

Global mean method

There are many variations of simple forecasting models found in the literature. One of them that is applied to time series that do not exhibit trend or seasonality, as is the case for log returns, is the global mean method. This method assumes that all historical data are equally important in helping forecast the future (Hyndman and Athanasopoulos 2018). Therefore, forecasts are nothing but a simple average of all the observations:

$$\widehat{r}_{T+h} = \bar{r} = \frac{1}{T}\sum_{i=1}^{T} r_i$$

where $\{y_1, y_2, \cdots y_T\}$ is the training data and h the forecasting horizon.

The global mean method is appropriate for white noise processes (Kolassa et al., 2023), that is, time series that do not exhibit autocorrelation and whose values move randomly around a long-term average.

Exponential smoothing

Developed by Brown 1959; the simple exponential smoothing (ES) model is based on the notion that recent observations will serve better in generating forecasts than those from the distant past (Pankratz 2009). As such, the forecasts created by ES are a weighted average of all observations, with the coefficients decreasing exponentially as we move away from the forecast origin. The smoothing rate is determined by the α parameter, and the forecasts generated are of the form:

$$\widehat{r}_{T+h} = \alpha r_T + (1 - \alpha)\widehat{r}_T$$

where $0 \leq \alpha \leq 1$.

Higher values of α indicate a constantly changing level, and as α approaches 0, the level resembles the long-run

average of the time series (Kolassa et al., 2023). The estimation of the parameter α is done by minimizing the sum of squared errors (SSE) between past values and past one-step ahead forecasts in the training set.

In its simplest form, the ES model is appropriate for modelling time series that do not exhibit trend or seasonality (Hyndman and Athanasopoulos 2018) and the forecasts created by the model are constant, meaning that the forecast will have the same value for $h = 1, \ldots, n$. ES has been extended so that it can be used for time series that have trend or seasonality by Holt 2004 and Winters 1960. In our analysis, we will use the simple ES approach since the log returns do not exhibit trend or seasonality.

ARIMA

Autoregressive Integrated Moving Average (ARIMA) models are univariate time series models that exploit the autocorrelation of past values and errors in order to forecast future values. In its general form, a non-seasonal ARIMA (*p, d,q*) model can be defined as (Pankratz 2009):

$$\varphi(B)\nabla^d r_t = C + \theta(B)\alpha_t$$

where B is the backshift operator, α_t is the error term, $\varphi(B) = (1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p)$ the *p*-order autoregressive operator, $\theta(B) = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)$ the *q*-order moving average operator, $\nabla^d = (1 - B)^d$ the *d*-order differencing operator, and C a constant term.

As suggested by Box et al. 2015 the ARIMA methodology is implemented in three stages: identification, estimation, and residual diagnostic checking. In the identification stage, the orders of an ARIMA model (or a set of tentative ARIMA models) are identified by visually examining the sample autocorrelation (acf) and sample partial autocorrelation (pacf) functions and mapping them with the theoretical acf and pacf of known processes. According to the pacf and acf, the analyst may determine the order of the differencing as well as the orders of the autoregressive and moving average parts of the model. During estimation, estimates of the model's coefficients identified previously are obtained through the maximum likelihood criterion. Last, diagnostic checking is performed to evaluate whether the residuals of the fitted model(s) are statistically adequate.

In our analysis, instead of examining the acf and pacf functions, we use an algorithmic approach to identify the orders of the ARIMA model proposed by Hyndman and Khandakar 2008. The algorithm performs unit root tests to identify the order of differencing, and in a repetitive fashion, searches for a model that minimizes the corrected Akaike Information Criterion (AICc):

$$\text{AICc} = -2\log(L) + 2(p + q + k + 1)\left[1 + \frac{(p + q + k + 2)}{T - p - q - k - 2}\right]$$

Where L is the likelihood, p, q are the orders of the autoregressive and moving average part of the model, respectively, $k = 1$ if there is a constant term otherwise $k = 0$ and T is the length of the time series. The idea behind the minimizing the AICc in order to select the orders of the ARIMA model is that information criteria seek to optimize the trade-off between goodness of fit and the number of estimated parameters (i.e., selecting a parsimonious model).

After model identification and estimation, the examination of the residuals is performed in order to evaluate if they exhibit two important statistical properties. First, the mean of the residuals is approximately zero, ensuring that the forecasts are unbiased and second, if there is any residual autocorrelation. For the latter, we check for independence using the Ljung-Box test (Ljung and Box 1978) that is based on the Q statistic:

$$Q = T(T + 2)\sum_{k=1}^{l} \frac{\widehat{\rho}_k^2}{T - k}$$

Where $\widehat{\rho}_k$ is the autocorrelation at lag k, and l is the maximum lag under consideration. Q follows a χ^2 distribution and for a significance level of $\alpha = 0.05$ we can obtain the *p*-values and either reject the null hypothesis, stating that the residuals are independent or accept it.

AutoRegLightGBM

Gradient boosting algorithms (GBM) use weak learners (typically decision trees) in order to approximate a function to be used to solve a regression or a classification problem. More specifically, given a training set $\{(x_i, y_i)\}_{i=1}^n$ consisting of explanatory variables x and a response variable y, the objective of GBM is to estimate an approximation function $F(x)$ that minimizes the expected value of a loss function $L(y, F(x))$ (common functions for regression are the squared error and the absolute error):

$$\widehat{F} = \underset{F}{argmin}\, E_{x,y}[L(y, F(x))]$$

This is achieved through an iterative process where decision trees are used to predict the pseudo-residuals that were created in the previous steps. The pseudo-code explaining the algorithm is shown below (Friedman 2002):

Step 1: $F_0(x) = \underset{\gamma}{argmin}\sum_{i=1}^{N} L(y_i, \gamma)$
Step 2: For m = 1 to M do:

i) Calculate pseudo-residuals $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$ for $i = 1,\ldots, n$.

ii) Fit a regression tree to r_{im} and create terminal nodes R_{jm} for $j = 1,\ldots, J_m$.

iii) For $j = 1,\ldots, J_m$ calculate $\gamma_{jm} = \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma)$

iv) Set $F_m(x) = F_{m-1}(x) + v \bullet \gamma_{jm} 1(x \in R_{jm})$ where v is the learning rate

Step 3: Output $F_M(x)$

As the name suggests, LightGBM is a variant of algorithms belonging to the gradient boosting family that enhances the performance of standard gradient boosting through several innovations. First, it uses a histogram-based method to compute the gradients, improving the speed and efficiency of the algorithm. Furthermore, the trees grow leaf-wise instead of depth wise, targeting more informative splits. Lastly, it employs the gradient based one side sample approach (GOSS) which helps sampling efficiency, and the exclusive feature bundling, (EFB) which reduces feature space sparsity (Ke et al. 2017).

LightGBM, as many other machine learning algorithms, can be further optimized through what is called hyperparameter tuning. This process involves the systematic search and selection of the hyperparameters (e.g., number of trees, learning rate, etc.) from a hyperparameter space and the evaluation of each using a cross-validation strategy. Common methods of performing the search include the grid search, in which all the combinations of a given set of hyperparameters are evaluated, and the random search, in which random combinations are selected given the number of iterations provided by the analyst. In this analysis, we employed random search to perform the hyperparameter tuning, mainly to avoid the higher computational time required for grid search for the same hyperparameter space. During tuning, we not only searched for appropriate hyperparameters but also looked for the number of lags to be included as input variables (a minimum of 1 lag and a maximum 6 lags). Thus, each set of hyperparameters and lags generated forecasts on a rolling forecasting origin (Hyndman and Athanasopoulos 2018) in a recursive manner for 6-months-ahead over a 100-month period. Forecast errors were then calculated using the mean squared error (MSE) averaged across the tuning period:

$$MSE = \frac{1}{h} \sum_{i=1}^{h} (r_{T+i} - \widehat{r}_{T+i})^2$$

The hyperparameters and lags selected for the validation set were the ones of the model exhibiting the smallest MSE in the tuning period.

Ensemble

After fitting the ARIMA and the tuned LightGBM models to the training data, the ensemble model produced forecasts as the average of the forecasts of the two models for 6 months ahead:

$$\widehat{r}_{T+h_{Ensemble}} = \frac{1}{2}\widehat{r}_{T+h_{ARIMA}} + \frac{1}{2}\widehat{r}_{T+h_{LightGBM}}$$

The use of equal weights ensures the equal contribution of two models with diverse characteristics: one that is parsimonious and able to capture linear relationships, and another that is able to approximate well non-linear relationships that may exist in the time series. Even though averaging the forecasts of the two estimators may seem like a simple approach when combining their forecasts and weight optimization could potentially give better results, in practice it has been shown that simple operators such as the mean or the median of forecast combinations tend to yield equally or even more accurate forecasts (Petropoulos & Svetunkov, 2020; Spiliotis, 2023).

Framework and models' evaluation

In this present study, the accuracy of 5 univariate models to point forecast 6 months ahead was evaluated and compared. The main reasons for limiting the horizon to 6 months have to do with the fact that it coincides with the maximum window of lags that the LightGBM algorithm will consider during the hyperparameter tuning, and because we expect that some of the models under study will not provide additional information for longer horizons. For example, in the case of a stationary time series, an AR(1) model will converge to the mean after a few forecast steps and subsequent forecasts will remain constant. For horizons shorter than the selected horizon, the models under consideration may still be useful, as the future points of a shorter horizon are included in the horizon under study.

The price series of each base metal was divided into training-fitting data and test data, with the latest observations belonging to the test set and the earliest to the training set. For the ARIMA model, the training data was used to find the model's order through the automated algorithm described earlier, while for the AutoRegLightGBM, the training data were re-divided into training and tuning data (in the same fashion as before) in order to search and select a good set of hyperparameters and lags through the rolling

forecasting origin evaluation strategy. Given their simplicity and widely acknowledged success we employed the global mean, exponential smoothing, and ARIMA as benchmark models. We then compared the performance of the two proposed models against these benchmarks to ascertain their effectiveness (Fig. 1).

All models were evaluated using the rolling forecasting origin strategy (as in hyperparameter tuning) by refitting each model to the new window data and creating forecasts recursively for 6 months ahead. The total evaluation period was 80 months, leading to the generation of 75 forecasts in the testing period. The evaluation metrics used in this analysis were the root mean squared error (RMSE), the mean absolute error (MAE), as well as the scaled version of RMSE (RMSSE) proposed by Hyndman and Koehler 2006; averaged for the testing period.

$$RMSE = \sqrt{\frac{1}{h}\sum_{i=1}^{h}\left(r_{T+i} - \widehat{r}_{T+i}\right)^2}$$

$$MAE = \frac{1}{h}\sum_{i=1}^{h}\left|r_{T+i} - \widehat{r}_{T+i}\right|$$

$$RMSSE = \sqrt{\frac{1}{h}\sum_{i=1}^{h}\frac{\left(r_{T+i} - \widehat{r}_{T+i}\right)^2}{\overline{\Delta}_r^2}}$$

$$\text{where} \overline{\Delta}_r^2 = \frac{1}{T-1}\sum_{J=2}^{T}\left(\Delta_{r_J}\right)^2 \text{while} \Delta_{r_J} = r_J - r_{J-1}.$$

The selection of the metrics we used in our analysis was done in such a way as to highlight the different aspects of the error each one of them measures. We used two scale dependent metrics (RMSE and MAE), one that gives more weight to larger errors (RMSE) by squaring them, and one that gives the same weight to all errors (MAE) (Willmott and Matsuura 2005). In addition, we evaluated all models using the RMSSE, a scale-independent metric that compares the RMSE of a proposed model to the average RMSE a simpler model (in our case, a lag forward-naive model) would have when forecasting one step ahead in the training period. Thus, in cases RMSSE ≥ 1, the predictions of the proposed model are worse than the average one-step forecasts of a naïve model in the training period, and when RMSSE ≤ 1 the forecasts are better.

In line with many similar studies (Y. Chen et al., 2016; Díaz et al., 2020; Dooley & Lenihan, 2005), we do not test for statistical significance of the results, mainly for the reasons described by Armstrong, 2007; Kostenko & Hyndman, 2008. In addition, while we acknowledge that economic evaluation criteria could be useful as alternative measures for evaluating the proposed models, we have chosen to exclude them from the scope of this study.
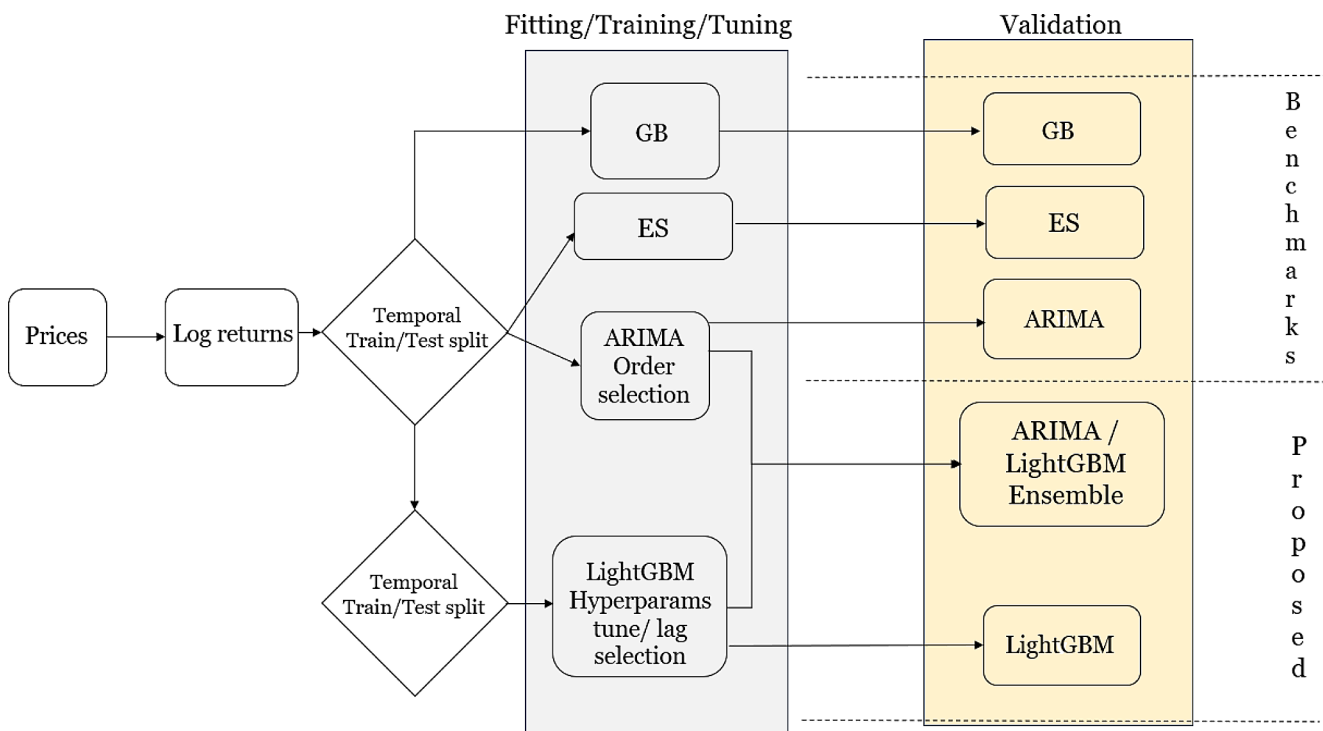

Fig. 1 Methodology of the current study

Table 1 Results of ARIMA model identification, estimation and diagnostic checking

Metal	Model	Lag	Coef	Log Likelihood	AICc	Mean resid	LBQ	LBQ p-val
Alum	AR (1)	1	0.26	523.63	-1043.23	~ 0	8.22	0.51
Copp	AR (1)	1	0.39	459.37	-914.71	~ 0	11.16	0.26
Lead	AR (1)	1	0.21	399.56	-795.09	~ 0	4.62	0.87
Tin	MA (2)	1	0.23	486.37	-966.68	~ 0	9.62	0.29
		2	0.18					
Nick	AR (1)	1	0.30	364.64	-725.26	~ 0	5.31	0.81
Zinc	AR (1)	1	0.31	450.74	-897.45	~ 0	5.41	0.80

Table 2 Results of hyperparameter tuning and lag selection using validation on a rolling forecasting origin

Metal	Lags	N trees	Max depth	Learning rate	MSE
Aluminum	6	50	20	0.01	0.0027
Copper	6	200	5	0.01	0.0039
Lead	3	100	3	0.01	0.0048
Tin	3	300	30	0.001	0.0046
Nickel	5	100	9	0.01	0.0063
Zinc	1	300	5	0.001	0.0040

Table 3 Root mean squared error averaged across the testing period for all models

Metal	RMSE				
	GM	ES	ARIMA	LightGBM	Ensemble
Aluminum	0.04474	0.04474	0.04411	**0.04400**	0.04402
Copper	0.04400	0.04505	0.04396	0.04399	**0.04330**
Lead	0.04477	0.04477	**0.04461**	0.04499	0.04472
Tin	0.05807	0.06228	**0.05711**	0.05825	0.05754
Nickel	0.07487	0.07487	0.07427	**0.07376**	0.07393
Zinc	0.05986	0.06226	0.05982	0.05990	**0.05976**

Table 4 Mean absolute error averaged across the testing period for all models

Metal	MAE				
	GM	ES	ARIMA	LightGBM	Ensemble
Aluminum	0.03816	0.03816	0.03772	**0.03755**	0.03758
Copper	0.03509	0.03608	0.03476	0.03547	**0.03466**
Lead	**0.03806**	**0.03806**	0.03813	0.03857	0.03822
Tin	0.04816	0.05105	**0.04743**	0.04824	0.04772
Nickel	0.06186	0.06186	0.06123	**0.06079**	0.06093
Zinc	0.05029	0.05315	**0.04996**	0.05030	0.05006

Results

Models' identification, estimation and tuning

The algorithmic approach of ARIMA order identification selected an ARIMA(1,0,0) without constant term as the most appropriate model for all the time series except the one of tin, for which the algorithm identified an ARIMA(0,0,2) without constant. Since the algorithm outputted $d = 0$ for all the time series, it means that the log returns are already stationary and no additional differencing is required. The estimated autoregressive coefficients of the first model fitting were all positive, indicating a positive relationship between the current price and the previous price/shocks. As for the diagnostic checking of the residuals, the mean was approximately zero, ensuring unbiased forecasts, and no autocorrelation was identified through the Ljung-Box test for the first 10 lags, indicating that the residuals series are essentially white noise. Additional information regarding the maximum likelihood and AICc can be found in Table 1.

The tuning of hyperparameters and lags selection through validation on a rolling forecast origin revealed that the set of hyperparameters found through this process are different for each time series, possibly due to the unique characteristics of each time series, and that the inclusion of additional lags in a non-linear approximator may be beneficial for their forecasting. Specifically, as shown in Table 2, for aluminum returns, the algorithm used 6 lags as input variables, 50 trees, with a maximum tree depth of 20 and a learning rate of 0.01 having MSE = 0.0027. A LightGBM using 6 lags, 200 trees with a maximum depth of 5 and a learning rate of 0.01 was found to be the most appropriate for copper returns (MSE = 0.0039). For lead and tin returns, 3 lags

were selected and the estimator had an MSE = 0.0048 and MSE = 0.0046, respectively. Last, a LightGBM of 100 trees using 5 lags was chosen for nickel returns (MSE = 0.0063) and a LightGBM of 300 trees using 1 lag was chosen for zinc returns (MSE = 0.0040).

Out-of-sample forecasts

Tables 3 and 4 presents the RMSE and MAE for the test period for the rolling point forecasts produced by the different methods, respectively. As expected, the RMSE was larger compared to the MAE counterpart. In addition, Table 5 summarizes the RMSSE of the different models showing that, in fact, all models perform better than a naïve method would in the training period. RMSE and RMSSE suggest the same models as more appropriate, which is expected since the latter is analogous to the former. The best performing model for aluminum log returns was the AutoRegLightGBM model, which had a RMSE = 0.04400 and a MAE = 0.03755. As far as copper returns is concerned, the AutoRegLightGBM-ARIMA ensemble outperformed the GM, ES and AR(1) models as well as the AutoRegLightGBM (RMSE = 0.04330

Table 5 Root mean squared scaled error across the testing period for all models

Metal	RMSSE				
	GM	ES	ARIMA	LightGBM	Ensemble
Aluminum	0.75767	0.75767	0.74701	**0.74496**	0.74531
Copper	0.64523	0.66095	0.64475	0.64479	**0.63493**
Lead	0.51198	0.51198	**0.51026**	0.51466	0.51154
Tin	0.87620	0.93942	**0.86167**	0.87883	0.86810
Nickel	0.78483	0.78483	0.77863	**0.77320**	0.77497
Zinc	0.81199	0.84470	0.81147	0.81253	**0.81066**

and MAE = 0.03466). Lead and tin returns were forecasted better by the ARIMA model, which had a RMSE = 0.04461 and RMSE = 0.05711, respectively, even though the MAE of the GM model was smaller (0.03806) for lead returns. The AutoRegLightGBM had a smaller RMSE and MAE for nickel returns (0.07376 and 0.06079, respectively). Last, the ensemble model forecasted better in terms of RMSE, zinc returns (0.05976), although the MAE of the AR(1) benchmark was slightly smaller (MAE = 0.04996).

It is worth noting that the ensemble model was still more accurate than the three benchmark models for the base metals that had lower RMSE and MAE with the standalone AutoRegLightGBM. Moreover, for aluminum, lead, and nickel returns, the alpha parameter estimated through the minimization of SSE of the in-sample predictions for the ES model led to a model that produced forecasts nearly identical to the global mean method. This explains why the error metrics are essentially the same for the two models.

The fact that different error metrics suggest different models is common in this type of analysis and it is the preference of researchers to choose which is the most appropriate for their applications. Furthermore, the fact that LightGBM was tuned using MSE as the error metric does not necessarily mean that the set of parameters selected minimizes the MAE. Finally, as for the tuning of the LightGBM hyperparameters, it is possible that this could have led to even better results if a finer hyperparameter space was explored or a different tuning approach was selected (i.e., grid search).

Figure 2 shows the rolling forecasts for the test period of the AutoRegLightGBM model. It is interesting to note the fact that the first few steps ahead resemble the ones an ARIMA (as the ones identified for each time series before) would produce and that the generated forecasts of all base metals, except copper, seem to converge to the mean, again as one would expect from a stationary ARIMA model. The phenomenon is particularly evident in the returns of lead, tin, nickel, and zinc. Aluminum forecasts also follow this behaviour but seem to produce additional patterns. Finally, copper forecasts do not seem to mean revert (at least on a 6-month forecasting horizon) but produce patterns that often coincide with reality. This phenomenon may be attributed to possible longer memory non-linear effects present in

the time series (note that the number of lags selected during tuning was 6 for copper returns).

Figure 3 shows the rolling forecasts of the AutoReg-LightGBM-ARIMA ensemble for the test period. The addition of the ARIMA model forecasts appears to impart two additional features to the LightGBM predictions. The first is that it enhances the direction of the first few steps ahead forecasts and the second is that it intensifies the mean reversion. The former is particularly evident in the aluminum, nickel, and zinc returns, while the latter is evident in the copper returns. Enriching LightGBM forecasts with the aforementioned ARIMA forecast characteristics appears to be beneficial when forecasting copper and zinc returns. In contrast, it does not seem to help in improving LightGBM forecasts of aluminum and nickel returns.

Discussion

The prices of base metals have a significant impact on the participants in their market. Thus, a significant part of the literature related to mineral economics has been devoted to the development of tools designed to forecast base metals prices. A variety of different models have been employed, from univariate ARIMA models (Dooley and Lenihan 2005) to multivariate decision trees (Liu et al. 2017). Although the use of independent variables in forecasting models can be useful in understanding how these variables affect the dependent variable, more often than not, future values of these variables are unknown, and in order to generate authentic forecasts of the dependent variable, one needs to first obtain forecasts for the independent variables or resort to scenario forecasting (Hyndman and Athanasopoulos 2018). In addition, if lags of the independent variables are used for forecasting, then this may raise constraints on the horizon we can forecast (or it still requires forecasting models to be built for the independent variables for time steps beyond the horizon covered by the lags). Thus, the complexity and uncertainty of such a problem are likely to increase. On the other hand, univariate forecasting models in their basic forms (ARIMA, exponential smoothing, etc.) do not take into account external information and work by learning systematic patterns of the time series (autoregression, seasonality, trend, etc.) and reproducing them into the future.

In our study, we followed the second approach to create forecasts of the future values of the prices of base metals. We chose this approach mainly because the univariate methodology reduces complexity and does not require additional forecasts of the independent variables. Our findings demonstrate that autoregressive tree-based algorithms, in this case LightGBM, are capable of producing forecasts at a level of accuracy equal to or better than the one obtained
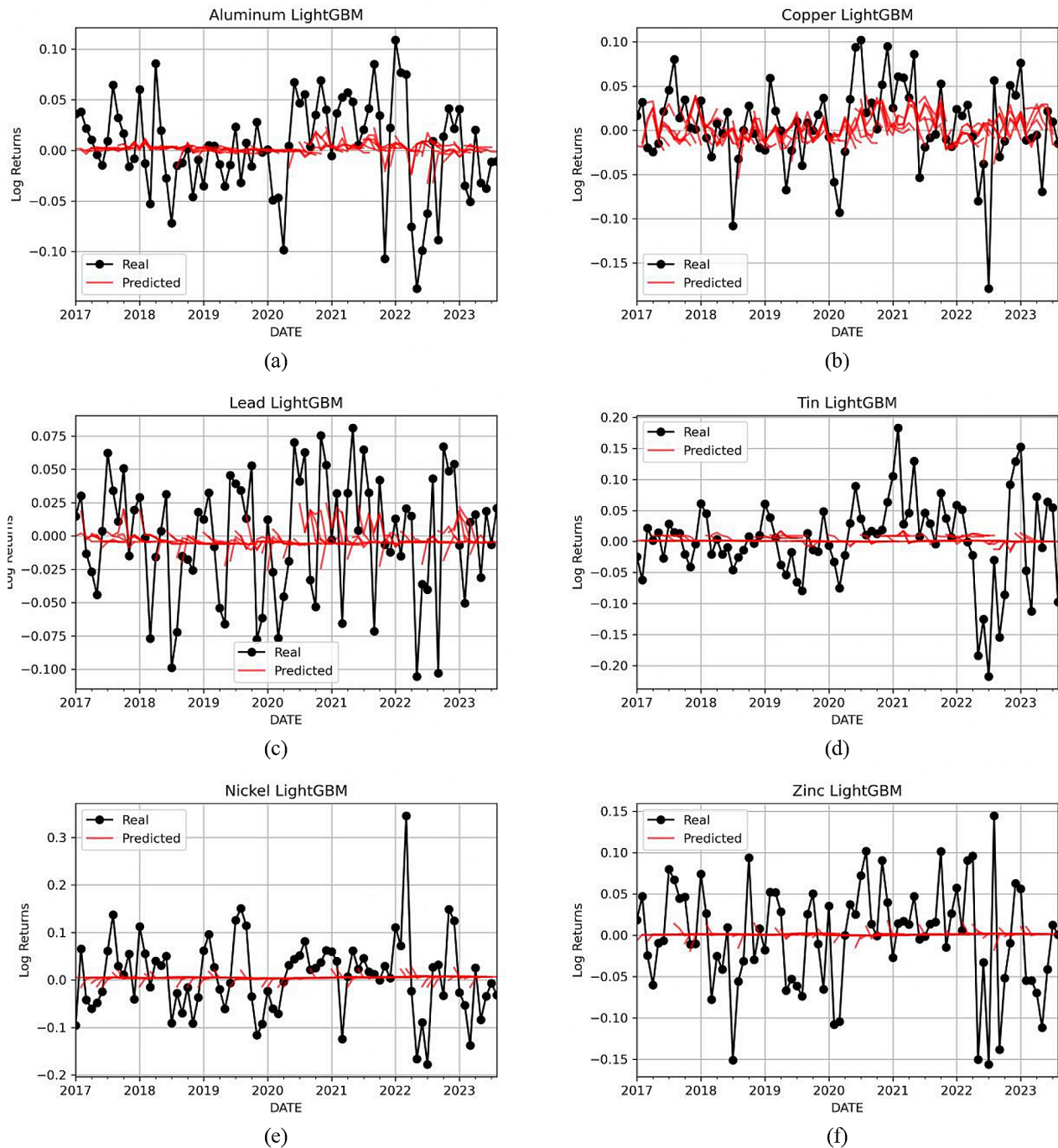
Fig. 2 Test data (black) and rolling forecasts (red) of the tuned AutoRegLightGBM algorithm (**a**) aluminum, (**b**) copper, (**c**) lead, (**d**) Tin, (**e**) Nickel, (**f**) Zinc

by established time series models, given the appropriate transformations. This can be attributed to the fact that the LightGBM algorithm can approximate longer memory, non-linear temporal relationships that are likely to exist in the time series of base metals prices in contrast to linear models such as ARIMA. In addition, it highlights that the combination of forecasts produced by different methods

can be beneficial as it reduces the uncertainty associated with the assumptions for the models selected (X. Wang et al. 2023a). Finally, it confirms the notion that higher model complexity is not necessarily associated with better out-of-sample performance (Makridakis et al. 2018a, b) since for lead and tin returns, the most accurate method was found to be a simple ARIMA model.
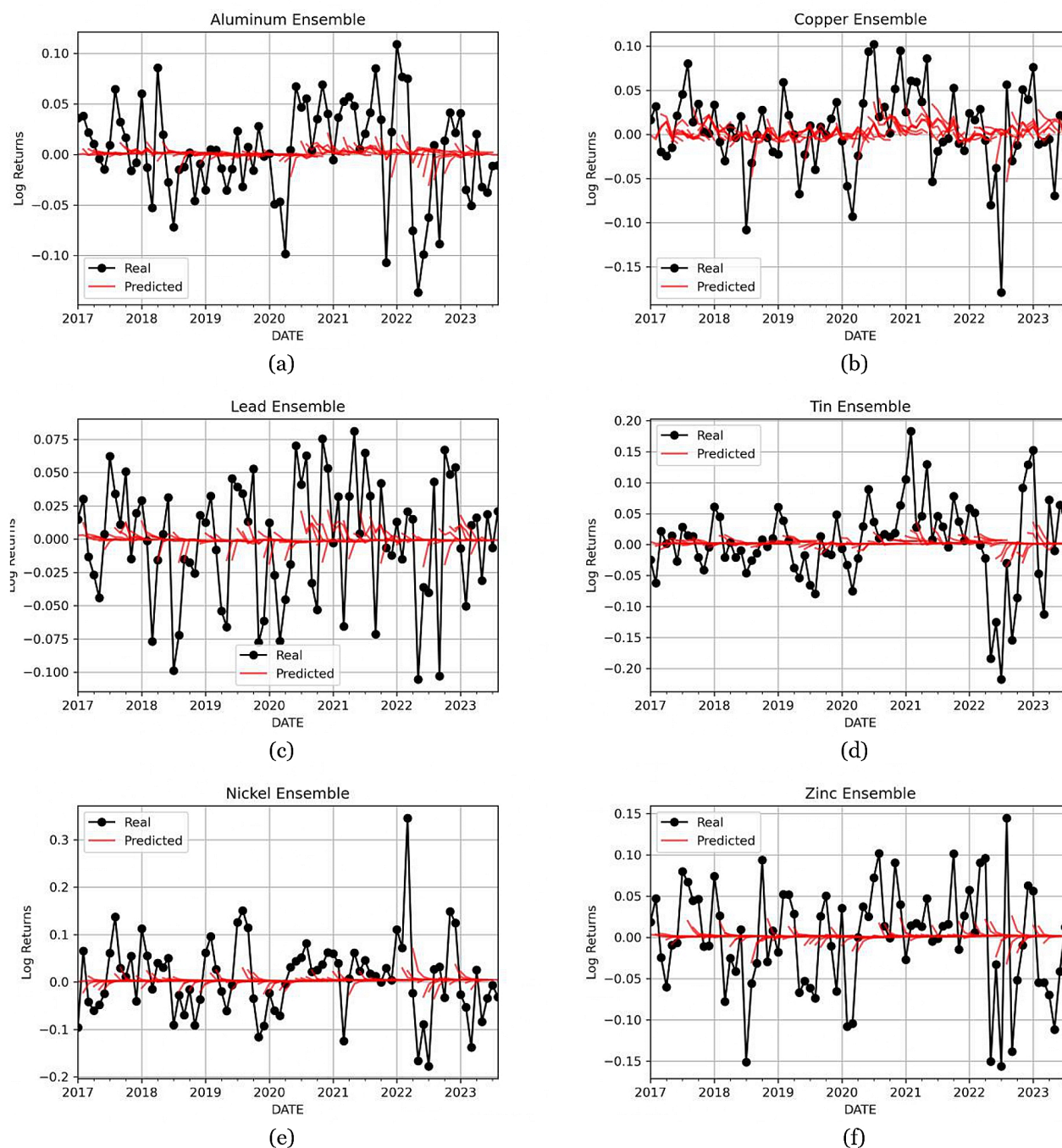
Fig. 3 Test data (black) and rolling forecasts (red) of the tuned LightGBM-ARIMA ensemble (**a**) aluminum, (**b**) copper, (**c**) lead, (**d**) Tin, (**e**) Nickel, (**f**) Zinc

Nevertheless, the study comes to support the idea that forecasting the prices of base metals is a challenging task (Dooley and Lenihan 2005) and that, in fact, only a fraction of the future can be explained from the past. The dynamic nature of the base metals markets and the stochasticity that characterizes these systems underscore the limitations of models to accurately forecast, especially at longer horizons. That being said, tools like the ones developed under this study, should not necessarily be used as standalones. Their forecasting capability can be further enhanced by expert judgement as this has been found to be a promising strategy in the literature (Franses and Legerstee 2011; Lawrence et al. 1986). Thus, the forecasts generated by the combination of exogenous and endogenous information can prove useful for stakeholders to reduce the uncertainty arising from base metals prices.

Conclusions

Base metals prices are a major source of uncertainty for producing countries, the companies involved in the mining and metallurgical sectors, and the industries that rely on base metals to produce their final products. The development of tools to forecast future prices of these commodities is important for stakeholders as it can facilitate forward planning and enable risk mitigation. In this direction, this study contributes to the literature by exploring how the accuracy of a competition-winning machine learning algorithm compares to that of classical time series models (GM, ES and ARIMA) in base metals prices forecasting. It also examines whether performance improves when LightGBM forecasts are combined with those generated by an ARIMA model. The results, when using RMSE and RMSSE as evaluation metrics, showed that the AutoRegLightGBM outperformed the three benchmarks (and the ensemble) when forecasting aluminum and nickel returns. Similarly, the ensemble AutoRegLightGBM-ARIMA model outperformed its benchmark counterparts when it came to the returns of copper and zinc. In contrast, neither of the two models under consideration appear to forecast better than the ARIMA benchmark when it comes to lead or tin returns, at least with the hyperparameter setting selected through random search. This finding suggests that complex models are not a panacea for all forecasting problems and that in certain cases the forecast accuracy of classical time series models is very difficult to be surpassed even by state-of-the-art algorithms.

Univariate forecasting models are based on finding systematic phenomena that occur historically in a time series and extrapolating them into the future. The limitations of choosing such models to forecast a time series lie in the fact that external information is not inherently included in them. At the same time, the assumptions made when selecting a model are crucial when it comes to the final accuracy. Thus, in our view, future research on base metals prices forecasting could be headed in two directions. First, to explore whether forecasts created by experts or organizations studying base metals markets can be used to adjust or enhance the forecasts of these models (for instance, the World Bank regularly publishes price projections for base metals). In this way, the final forecasts will incorporate information from the past (derived from the statistical/computational models) and estimates for the present and future outlook of the metals markets (from expert judgement). Secondly, future researchers could introduce more models in such an ensemble, taken both from the field of machine learning (neural networks, k-nearest neighbours, etc.) and from the field of classical time series analysis (theta method, etc.), given the fact that each of them makes different assumptions and has different advantages when applied to such problems. Once such ensembles are created, it would be of value to study the strategy of selecting and optimizing the weights of the models that make up the ensemble in order to investigate how weighting affects the out-of-sample accuracy.

Declarations

References

Ahmed MY, Sarkodie SA (2021) COVID-19 pandemic and economic policy uncertainty regimes affect commodity market volatility. Resour Policy 74:102303. https://doi.org/10.1016/j.resourpol.2021.102303

Alam MR, Forhad MAR, Sah NB (2022) Consumption- and speculation-led change in demand for oil and the response of base metals: a Markov-switching approach. Finance Res Lett 47:102783. https://doi.org/10.1016/j.frl.2022.102783

Armstrong, J. S. (2007). Significance tests harm progress in forecasting. International Journal of Forecasting, 23(2), 321–327. https://doi.org/10.1016/j.ijforecast.2007.03.004

Backman C-M (2008) Global supply and demand of metals in the future. J Toxicol Environ Health Part A 71(18):1244–1253. https://doi.org/10.1080/15287390802209582

Bates JM, Granger CWJ (1969) The combination of forecasts. J Oper Res Soc 20(4):451–468. https://doi.org/10.1057/jors.1969.103

Bauer DJ, Nguyen RT, Smith BJ (2023) Critical materials Assessment. U.S Department of Energy

Box GE, Jenkins GM, Reinsel GC, Ljung GM (2015) Time series analysis: forecasting and control. Wiley

Brown RG (1959) Statistical forecasting for inventory control. (No Title). https://cir.nii.ac.jp/crid/1130282269492751488

Castel-Branco CN, Goldin N (2003) Impacts of the Mozal aluminium smelter on the Mozambican economy. Final Report Presented to Mozal. https://www.iese.ac.mz/~ieseacmz/lib/cncb/Mozal_and_economic_development.pdf

Chen M-H (2010) Understanding world metals prices—Returns, volatility and diversification. Resour Policy 35(3):127–140. https://doi.org/10.1016/j.resourpol.2010.01.001

Chen Y, He K, Zhang C (2016) A novel grey wave forecasting method for predicting metal prices. Resour Policy 49:323–331. https://doi.org/10.1016/j.resourpol.2016.06.012

Chipili JM (2016) Copper Price and Exchange Rate Dynamics in Zambia. J Int Dev 28(6):876–886. https://doi.org/10.1002/jid.3183

Dhar SK, Ovshínsky SR, Gifford PR, Corrigan DA, Fetcenko MA, Venkatesan S (1997) Nickel/metal hydride technology for consumer and electric vehicle batteries—A review and update. J Power Sources 65(1):1–7. https://doi.org/10.1016/S0378-7753(96)02599-2

Díaz JD, Hansen E, Cabrera G (2020) A random walk through the trees: forecasting copper prices using decision learning methods. Resour Policy 69:101859. https://doi.org/10.1016/j.resourpol.2020.101859

Dooley G, Lenihan H (2005) An assessment of time series methods in metal price forecasting. Resour Policy 30(3):208–217. https://doi.org/10.1016/j.resourpol.2005.08.007

Du P, Wang J, Yang W, Niu T (2020) Point and interval forecasting for metal prices based on variational mode decomposition and an optimized outlier-robust extreme learning machine. Resour Policy 69:101881. https://doi.org/10.1016/j.resourpol.2020.101881

Foo N, Bloch H, Salim R (2018) The optimisation rule for investment in mining projects. Resour Policy 55:123–132. https://doi.org/10.1016/j.resourpol.2017.11.005

Franses PH, Legerstee R (2011) Combining SKU-level sales forecasts from models and experts. Expert Syst Appl 38(3):2365–2370. https://doi.org/10.1016/j.eswa.2010.08.024

Friedman JH (2002) Stochastic gradient boosting. Comput Stat Data Anal 38(4):367–378

Holt CC (2004) Forecasting seasonals and trends by exponentially weighted moving averages. Int J Forecast 20(1):5–10. https://doi.org/10.1016/j.ijforecast.2003.09.015

Hyndman RJ, Athanasopoulos G (2018) Forecasting: principles and practice. OTexts

Hyndman RJ, Khandakar Y (2008) Automatic Time Series forecasting: the forecast Package for R. J Stat Softw 27:1–22. https://doi.org/10.18637/jss.v027.i03

Hyndman RJ, Koehler AB (2006) Another look at measures of forecast accuracy. Int J Forecast 22(4):679–688

Joseph M (2022) Modern Time Series forecasting with Python: explore industry-ready time series forecasting using modern machine learning and deep learning. Packt Publishing Ltd.

Kabundi A, Nagle P, Ohnsorge F, Yamazaki T (2022) Causes and Consequences of Industrial Commodity Price Shocks. Commodity Markets: Evolution, Challenges, and Policies. https://thedocs.worldbank.org/en/doc/b4ff84b2d5dc4d0963a5074102460cc1-0350012022/related/Commodity-Markets-Chapter-4.pdf

Kahraman E, Akay O (2023) Comparison of exponential smoothing methods in forecasting global prices of main metals. Mineral Econ 36(3):427–435. https://doi.org/10.1007/s13563-022-00354-y

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. https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

Khoshalan HA, Shakeri J, Najmoddini I, Asadizadeh M (2021) Forecasting copper price by application of robust artificial intelligence techniques. Resour Policy 73:102239. https://doi.org/10.1016/j.resourpol.2021.102239

Kolassa, S., Rostami-Tabar, B., & Siemsen, E. (2023). Demand forecasting for executives and professionals. CRC Press. https://books.google.com/books?hl=el&lr=&id=sKPMEAAAQBAJ&oi=fnd&pg=PP1&dq=Demand+Forecasting+for+Executives+and+Professionals&ots=0AlQ-CSXRt&sig=gRIt8JLusUOqYJknD76NyeFIzLo

Kostenko, A. V., & Hyndman, R. J. (2008). Forecasting without significance tests. Manuscript, Monash University, Australia. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=f9e981ce476390a452f699e4f36f4acdf1450c16

Kriechbaumer T, Angus A, Parsons D, Rivas Casado M (2014) An improved wavelet–ARIMA approach for forecasting metal prices. Resour Policy 39:32–41. https://doi.org/10.1016/j.resourpol.2013.10.005

Krustiyati JM, Christine N, Al-Fatih S, Jaelani AK (2022) Analyzing the lawsuit of the European Union over nickel ore export regulation in Indonesia. Croatian Int Relations Rev 28(89):120–135

Lacal-Arántegui R (2015) Materials use in electricity generators in wind turbines– state-of-the-art and future specifications. J Clean Prod 87:275–283. https://doi.org/10.1016/j.jclepro.2014.09.047

Lawrence MJ, Edmundson RH, O'Connor MJ (1986) The Accuracy of combining judgemental and statistical forecasts. Manage Sci 32(12):1521–1532. https://doi.org/10.1287/mnsc.32.12.1521

Liu C, Hu Z, Li Y, Liu S (2017) Forecasting copper prices by decision tree learning. Resour Policy 52:427–434. https://doi.org/10.1016/j.resourpol.2017.05.007

Ljung GM, Box GE (1978) On a measure of lack of fit in time series models. Biometrika 65(2):297–303

Lust J (2019) The rise of a capitalist subsistence economy in Peru. Third World Q 40(4):780–795. https://doi.org/10.1080/01436597.2018.1529540

MacDiarmid J, Tholana T, Musingwini C (2018) Analysis of key value drivers for major mining companies for the period 2006–2015. Resour Policy 56:16–30. https://doi.org/10.1016/j.resourpol.2017.09.008

Makridakis S, Spiliotis E, Assimakopoulos V (2018a) Statistical and machine learning forecasting methods: concerns and ways forward. PLoS ONE 13(3):e0194889. https://doi.org/10.1371/journal.pone.0194889

Makridakis S, Spiliotis E, Assimakopoulos V (2018b) The M4 competition: results, findings, conclusion and way forward. Int J Forecast 34(4):802–808. https://doi.org/10.1016/j.ijforecast.2018.06.001

Makridakis S, Spiliotis E, Assimakopoulos V (2022) M5 accuracy competition: results, findings, and conclusions. Int J Forecast 38(4):1346–1364. https://doi.org/10.1016/j.ijforecast.2021.11.013

Medina JP, Soto C (2007) Copper price, fiscal policy and business cycle in Chile 458. Citeseer. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=d423019060bfac558e1caa6bb33393d2f0408d18

Pankratz A (2009) Forecasting with univariate Box-Jenkins models: concepts and cases. Wiley

Petropoulos F, Svetunkov I (2020) A simple combination of univariate models. Int J Forecast 36(1):110–115. https://doi.org/10.1016/j.ijforecast.2019.01.006

Pincheira Brown P, Hardy N (2019) Forecasting base metal prices with the Chilean exchange rate. Resour Policy 62:256–281. https://doi.org/10.1016/j.resourpol.2019.02.019

Radetzki, M., & Wårell, L. (2020). A handbook of primary commodities in the global economy. Cambridge University Press. https://books.google.com/books?hl=el&lr=&id=ifwAEAAAQBAJ&oi=fnd&pg=PR7&dq=A+handbook+of+primary+commodities&ots=_J6_WXYwQL&sig=bpLVbj23e8QPRW997-H3-kwgkEA

Rossen A (2015) What are metal prices like? Co-movement, price cycles and long-run trends. Resour Policy 45:255–276. https://doi.org/10.1016/j.resourpol.2015.06.002

Rufo DD, Debelee TG, Ibenthal A, Negera WG (2021) Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM). Diagnostics 11(9):1714

Sánchez Lasheras F, de Cos Juez FJ, Suárez Sánchez A, Krzemień A, Riesgo Fernández P (2015) Forecasting the COMEX copper spot price by means of neural networks and ARIMA models. Resour Policy 45:37–43. https://doi.org/10.1016/j.resourpol.2015.03.004

Spiliotis E (2023) Time Series Forecasting with Statistical, Machine Learning, and Deep Learning Methods: Past, Present, and Future. In M. Hamoudia, S. Makridakis, & E. Spiliotis (Eds.), Forecasting with Artificial Intelligence: Theory and Applications (pp. 49–75). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-35879-1_3

Tang M, Zhao Q, Ding SX, Wu H, Li L, Long W, Huang B (2020) An improved lightGBM algorithm for online fault detection of wind turbine gearboxes. Energies 13(4):807

Tashman LJ (2000) Out-of-sample tests of forecasting accuracy: an analysis and review. Int J Forecast 16(4):437–450

Wang D, Li L, Zhao D (2022) Corporate finance risk prediction based on LightGBM. Inf Sci 602:259–268. https://doi.org/10.1016/j.ins.2022.04.058

Wang B, Zhang Z, Xu G, Zeng X, Hu W, Matsubae K (2023a) Wrought and cast aluminum flows in China in the context of electric vehicle diffusion and automotive lightweighting. Resour Conserv Recycl 191:106877. https://doi.org/10.1016/j.resconrec.2023.106877

Watari T, Nansai K, Nakajima K (2021) Major metals demand, supply, and environmental impacts to 2100: a critical review. Resour Conserv Recycl 164:105107. https://doi.org/10.1016/j.resconrec.2020.105107

Watkins C, McAleer M (2004) Econometric modelling of non-ferrous metal prices. J Economic Surveys 18(5):651–701. https://doi.org/10.1111/j.1467-6419.2004.00233.x

Willmott CJ, Matsuura K (2005) Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Res 30(1):79–82

Winters PR (1960) Forecasting sales by exponentially weighted moving averages. Manage Sci 6(3):324–342. https://doi.org/10.1287/mnsc.6.3.324

Xu X, Zhang Y (2023) Price forecasts of ten steel products using gaussian process regressions. Eng Appl Artif Intell 126:106870. https://doi.org/10.1016/j.engappai.2023.106870

Zhou D, Guo X, Zhang Q, Shi Y, Zhang H, Yu C, Pang H (2022) Nickel-based materials for Advanced Rechargeable batteries. Adv Funct Mater 32(12):2107928. https://doi.org/10.1002/adfm.202107928