Forecasting real oil returns using an deep forest ensemble approach

Xingfu Xu

Email: 12031232@mail.sustech.edu.cn

Weihan Liu

Email: weihanliu2002@yahoo.com

Department of Finance, Southern University of Science and Technology

1088 Xueyuan Avenue, Shenzhen 518055, P.R. China

This version: July 26, 2023

Forecasting real oil returns using a deep forest ensemble approach

Abstract: We made an application of a cutting-edge machine learning method, the deep forest ensemble approach (DFEA, Zhou and Feng (2017, 2019)), to empirically predict crude oil prices. We used a large data set with 36 explanatory variables to compare the predictability of the DFEA with seven popular machine learning models and their mean combination method. The out-of-sample forecasting results showed that the DFEA statistically and economically outperforms all the competing models in terms of out-of-sample R square and success ratio. The DEFA also displayed sizable certainty equivalent return (CER) gains for a mean-variance investor in practice. Furthermore, we found that the predictive power of the DFEA mainly stems from technical indicators, especially momentum predictors. In terms of economic value, the DFEA can also deliver substantial average returns and Sharpe ratios from a market timing perspective. Our results survived in various robustness checks.

Keywords: machine learning methods; deep forest ensemble approach, support vector machine; LASSO

JEL Classifications: C53, E37, Q47

1 Introduction

Known as the “lifeblood of the industry,” crude oil plays its vital role in transportation, chemistry, and many other industries. The changes in oil prices directly impact the real economy and global financial markets enormously (Kilian and Park 2009, Hou, Mountain, and Wu 2016, Boubaker, Liu, and Zhang 2022). Governments and central banks must monitor and evaluate macroeconomic risks and develop policies as responses to the major oil prices changes (He et al. 2021). Investors and portfolio managers also need to accordingly adjust their asset allocations (Zhang, Ma, and Wang 2019). Thus, accurate oil price forecasts are of vital interest to the industry, governments, and individual investors (Sun et al. 2022).

To enhance accurate forecasting performance, we need to consider two fundamental issues: the driving factors and the modeling techniques. We notice that we generally have two major categories of driving factors, technical indicators and macroeconomic variables (Guidolin and Pedio 2020, He et al. 2021), exerting their prominent influence on crude oil price levels. Several previous studies (Yin and Yang 2016, Zhang, Ma, and Wang 2019) documented that technical indicators (including three categories, moving average indicators, momentum indicators, and volume indicators) exhibit stronger predictive ability than macroeconomic variables. In dealing with the variety of technical indicators, one should specify the significant items at work. Meanwhile, due to the increasing complexity in the changing international political economy and the macro-and micro-economic environment, it is a highly challenging task to predict oil prices (Zhao, Li, and Yu 2017).

Previous studies relied on two major categories of modeling: econometric techniques and machine learning techniques (Wang, Li, et al. 2018). The major econometric techniques include the vector autoregressive model (Baumeister and Kilian 2014), forecast combinations (Zhang, Ma, and

Wang 2019, He et al. 2021), the LASSO (the Least Absolute Shrinkage and Selection Operator) regression (Miao et al. 2017, Zhang, Ma, and Wang 2019). The econometric models sometimes failed to concurrently model the time series with complex, irregular, time-varying, and non-linear features (Zhao, Li, and Yu 2017). However, some machine learning models are confirmed to emerge as surpassing alternatives and outperform econometric models in forecasting complex tasks (Li, Zhu, and Wu 2019, Creamer and Lee 2019, Boubaker, Liu, and Zhang 2022). Several studies reported the excellent forecasting performance of the classical machine learning methods, such as random forest, SVM, and others (Guo, Li, and Zhang 2012, Wang, Athanasopoulos, et al. 2018, Alshater et al. 2022). Especially, Hornik, Stinchcombe, and White (1989) reported that the deep neural networks (DNNs) evidently surpassed all other machine learning methods because DNNs can theoretically and universally approximate any function. The outperformance of DNNs is confirmed in various empirical analyses. Among the various applications of DNNs, Zhao, Li, and Yu (2017) used a deep learning ensemble approach to forecast crude oil prices, and their approach exhibited even better forecasting performance. However, DNNs commonly require many plots to tune hyper-parameters for prediction, such as the learning rate, the number of hidden layers, choice of activation function, and so on. DNNs are inevitably computationally costly. We thus need to improve DNNs for a better tradeoff between prediction accuracy and computational cost.

Although computational resource is not a major concern nowadays, we prefer a more straightforward and efficient alternative for more timely predictions. We thus turned to the cutting-edge machine learning algorithm which can achieve better prediction performance with few hyper-parameters without entangling with the aforementioned plots. We employed the deep forest ensemble approach (DFEA) proposed by Zhou and Feng (2017, 2019) to forecast crude oil prices.

Inspired by the structure of DNNs, the DFEA uses the cascade forest structure and multi-grained scanning procedure to enhance its representational learning ability in each process step. It is noteworthy that the DFEA can succulently and completely extract internal features in complex data sets and achieve predicting performance highly comparable to those of other more complex DNNs. More importantly, if we compare with other DNNs, the DFEA has much fewer hyper-parameters to estimate, but the DFEA can provide more robust estimates based on almost the same hyper-parameter settings. A strand of recent literature confirmed the outperformance of the DFEA in prediction based on various datasets of different domains (Ma et al. 2020, Xia et al. 2020, Zeng et al. 2020).

However, to the best of our knowledge, none of these noteworthy studies used the DFEA to predict oil prices. We are motivated to investigate whether the DFEA can further improve the forecasting performance of crude oil prices as compared with these conventional machine learning techniques. Considering the outperformance of the DFEA in various prediction problems, we expected to shed light on the forecasting of the highly volatile oil prices (Regnier 2007). We further included a large data set used in previous literature to test DFEA. We extended our scope of included variables as wide as possible and included 36 explanatory variables: 18 technical indicators and 18 macroeconomic variables. These two sets of indicators are commonly used to predict oil prices (Yin and Yang 2016, Zhang et al. 2018, Zhang, Ma, and Wang 2019). While we acknowledge this highly challenging task to model the complex relations between oil price returns and these predictors, we need a powerful and efficient forecast method for timely predictions.

To illustrate the superiority of the DFEA, we conducted a horserace as comprehensive as possible to compare the predictive power of DFEA with those of conventional machine learning

methods, including random forest (Breiman 2001), extreme gradient boosting regression (XGBoost; Chen and Guestrin 2016), Gaussian process regression (GPR; Rasmussen and Williams 2006), SVM (Vapnik 1999), the ridge regression (Hoerl and Kennard 1970), the LASSO regression (Tibshirani 1996) and k-nearest neighbors regression (KNN; Altman 1992). We selected these three machine learning methods because existing studies confirmed their strong predictive power in forecasting crude oil prices (Zhang, Ma, and Wang 2019, Costa et al. 2021). We also considered the mean combination method for performance comparison as suggested in several previous literature (Rapach, Strauss, and Zhou 2010, Zhu and Zhu 2013). In terms of the out-of-sample performance comparison, we focus to examine two fundamental issues: the out-of-sample model fit and directional accuracy of forecast. We thus employed the out-of-sample R square statistics introduced by Campbell and Thompson (2008) and the directional accuracy test of Pesaran and Timmermann (1992) to compare the success ratio of various forecasting models with that of random walk. These two criteria help us confirm if our prediction performance is better than that of a random walk as a benchmark. Moreover, we investigated if DFEA can yield economic gains and computed the certainty equivalent return (CER) gains of various models for a mean-variance investor from an asset allocation perspective. Finally, we conducted the forecast encompassing test (Harvey, Leybourne, and Newbold 1998) to compare the information content of the forecast based on the DFEA to those forecasts based on the various competing models (Rapach, Strauss, and Zhou 2010, Zhang, Ma, and Wang 2019).

Our empirical results showed that the DFEA outperforms all the competing models in terms of out-of-sample R square and success ratio. The results of the forecast encompassing tests showed that the DFEA encompasses the competing models, whereas the competing models generally cannot

encompass the DFEA forecast. This implies that the DFEA forecast provides more useful information for forecasting oil price returns as compared with the competing forecasts. Furthermore, we explored the driving factors of the strong predictive power of the DFEA forecast. We found that technical indicators have a stronger predictive power as compared with macroeconomic variables. Momentum predictors are the principal source of the strong predictability of technical indicators. We noticed that the presence of late-informed investors can explain the outperformance of momentum predictors. Finally, the various robustness checks corroborated the excellent performance of the DFEA.

We also investigate from the perspective of the economic value and used the trades of crude oil futures from a market timing perspective to demonstrate the strong predictive power of the DFEA. Specifically, we selected the two well-known crude oil futures as the trading assets: the West Texas Intermediate (WTI) crude oil futures in the New York Mercantile Exchange (NYMEX) and the Brent crude oil futures in the Intercontinental Exchange (ICE).

Specifically, we designed the market timing strategies based on the DFEA and various competing models by taking a long position in crude oil futures in the next month if the forecasting oil price return in that month is positive and taking a short position otherwise. We also employed the *Always Long* strategy as the benchmark, which takes a long position of oil futures from the beginning of the next month and closes it at the end of the next month. The outcomes of the market timing strategy indicated that the DFEA led to the largest economic gains in terms of average return and Sharpe ratio when we traded the NYMEX WTI crude oil futures. For trading the ICE Brent crude oil futures, the DFEA-based strategy also generated the largest economic gain in both average return and Sharpe ratio. In sum, the market timing strategy based on the DFEA can yield substantial

economic gains as compared with the strategies based on the competing models and the *Always Long* strategy.

Our paper related to the studies of Yin and Yang (2016) and Zhang, Ma, and Wang (2019) and made two new contributions. First, we dug deeper into the independent variables and identified the significant driving factors in our empirical analysis. Existing studies (Yin and Yang 2016, He et al. 2021) demonstrated that technical indicators delivered larger predictive power than macroeconomic variables but did not specify which kind of technical indicators contributed the most to the forecasting performance. We reported a new finding that the momentum predictors were the major source of the strong predictive power of technical indicators. Second, we confirmed the outstanding forecasting power of the DFEA as compared with various machine learning methods. Yin and Yang (2016) and Zhang, Ma, and Wang (2019) focused on econometric methods, such as the LASSO regression and the elastic net model. However, these linear frameworks were not designed to capture the complex and nonlinear relations between crude oil prices and various driving factors. Our empirical analysis outcomes indicated that the modeling performance of those econometric methods cannot be highly expected.

The remainder of the paper is organized as follows. Section 2 briefly presents the methodology of the DFEA and its competing models. Section 3 describes the dataset, including crude oil prices and the predictors. Section 4 summarizes the out-of-sample forecasting outcomes and its discussion. Section 5 demonstrates the robustness checks. Section 6 reports the economic value of the DFEA from a market timing perspective. Finally, Section 7 concludes.

2 Methodology

This section summarizes the DFEA originated by Zhou and Feng (2017) and briefly overviews eight competing models: random forest, XGBoost, GPR, SVM, Ridge, LASSO, KNN and the mean combination method. These competing models were noticed for their prediction performance and selected for performance comparison.

We first considered a general task to forecast the log of spot oil price return in a month $t+1$ using information up to a month t. The predictive model for the return at $t+1$ can be expressed as

$$r_{k,t+1} = f_{kt}(X_t) + \epsilon_{k,t+1}, \quad (1)$$

where r_{kt} is the log return of model k on spot oil prices at month t; X_t is a vector of explanatory variables, including technical indicators and macroeconomic variables; and ϵ_{kt} denotes the error term of model k in a month t. f_{kt} is the selected k-th predictive model at month t. This study selected nine predictive models, including the DFEA and the eight competing models (random forest, XGBoost, GPR, SVM, Ridge, LASSO, KNN and the mean combination method).

2.1 Deep Forest Ensemble Approach

The DFEA is a novel machine learning method inspired by the DNNs (Zhou and Feng 2017, 2019). The key features of the DFEA indicate that this method is not based on neural networks and does not depend on the backpropagation algorithm. The DFEA is the first deep model that avoids rigorous differentiability condition for the backpropagation algorithm and can function as an alternative to DNNs in many challenging machine-learning tasks. Unlike DNNs which require lots of tricks to refine many hyper-parameters, the DFEA has much fewer hyper-parameters to deal with and makes it much easier to train. To understand the DFEA, we first illustrate the cascade forest structure and then the procedure of multi-grained scanning.

Figure 1 illustrates the cascade forest structure motivated by the structure of DNNs and

considers a classification problem predicting three classes. The cascade forest structure starts with the input feature vector on the very left, and each level (Level 1, Level 2, …, Level *N*) receives feature information processed by its proceeding level. An ensemble of tree forests constitutes the element of each level. These tree forests are diversified to achieve better ensemble construction. Each forest yields a three-dimensional vector and is concatenated as an input feature vector for the next level. At Level *N*, we aggregate the prediction results for all forests by taking their respective average value (Ave.) to obtain one 3-dimensional aggregated vector. The final prediction is obtained by taking the class with the largest value (Max) in the aggregated vector. This design helps depict high level of feature information.

DFEA is initially designed for image recognition. Its multi-grained scanning, inspired by DNNs and recurrent neural networks (RNNs), can further enhance the performance of cascade forests by detecting the spatial relations on image data. That is, RNNs can capture sequential relations on sequence data. Specifically, this multi-grained scanning uses sliding windows of various sizes to scan the input feature vector at each level and thus produces diverse grained feature vectors. We aggregate these grained feature vectors into a high-dimensional feature vector. Like the function of the scanning procedure in DNNs and RNNs, the high-dimensional feature vector can facilitate capturing the spatial and sequential relations in the data set by the DFEA.

In short, the DFEA is based on the ensemble produced by tree forests. It uses a cascade structure to represent feature information and efficiently process raw features. A series of experiments showed that the DFEA achieved excellent performance with the default setting of hyper-parameters and made the training of the DFEA more convenient (Zhou and Feng, 2017). The multi-grained scanning can further enhance its representation learning ability.

2.2 The competing models

2.2.1 Random Forest

The random forest (Breiman 2001) is a powerful machine-learning method for high-dimensional regression and classification. Constructed by many decision trees, the random forest employs a bagging ensemble strategy to reduce prediction variance, generally outperforming a single decision tree. This method is applied extensively to predict energy prices, especially crude oil prices (Ferrari, Ravazzolo, and Vespignani 2021). The algorithm of random forest is outlined as follows (Costa et al. 2021):

Given a dataset of explanatory variables and oil price returns (X_t, r_{t+1}) with n observations, we used repeatedly bagging (B times) to select a random sample from the dataset with replacement and to fit decision trees:

For b = 1, …, B:

(1) Sample with replacement of n training samples from (X_t, r_{t+1}); call these sampled (X_b, r_{b+1}) the training sample.

(2) Train a decision tree model $f_b()$ on the training sample (X_b, r_{b+1});

(3) The random forest is constructed by averaging the prediction from all the decision trees. For a new sample X_{new}, the prediction $\hat{f}$ on X_{new} can be evaluated as:

$$\hat{f} = \frac{1}{B}\sum_{h=1}^{B} f_b(X_{new}) \tag{2}$$

Due to its randomness, the subsampling procedure can significantly affect the performance of the random forest. We notice that the random forest will be very likely to generate inconsistent estimates with either no subsampling or over-subsampling, as reported in several previous studies (Tang, Garreau, and von Luxburg 2018).

2.2.2 XGBoost

XGBoost (Chen and Guestrin 2016) provides a powerful and efficient implementation of gradient boosting framework and has achieved cutting-edge results on many machine learning competitions. The distinguishing characteristic of this algorithm lies in its utilization of the second-order Taylor approximation within the loss function. The generic algorithm of XGBoost model can be described as follows:

Algorithm: The Implementation of Extreme Gradient Boosting Model

Input: training set $\{(X_t, r_{t+1})\}_{t=1}^{n}$, a differentiable loss function $L(r, f(X))$ where $f(X)$ can be seen as the parameters of the model, a learning rate α and M classification and regression tree (CART) models as base learners.

Output: the prediction $\hat{f}(X)$.

1.Use a constant value to initialize model

$\hat{f}_0(X) = \arg\min_{\theta} \sum_{t=1}^{n} L(r_{t+1}, \theta)$, θ denotes the parameters of the model.

2.**for** $m = 1$ to M **do**

(1) Calculate the first derivative and the second derivative with respect to $f(X_t)$:

$$\hat{g}_m(X_t) = \left[\frac{\partial L(r_{t+1}, f(X_t))}{\partial f(X_t)} \right]_{f(X)=\hat{f}_{(m-1)}(X)}$$

$$\hat{h}_m(X_t) = \left[\frac{\partial^2 L(r_{t+1}, f(X_t))}{\partial f(X_t)^2} \right]_{f(X)=\hat{f}_{(m-1)}(X)}$$

(2) Train a base learner using the training set $\{(X_t, -\frac{\hat{g}_m(X_t)}{\hat{h}_m(X_t)})\}_{t=1}^{n}$ by resolving the following optimization problem:

$$\hat{\phi}_m = \arg\min_{\phi \in \Phi} \sum_{t=1}^{n} \frac{1}{2} \hat{h}_m(X_t) \left[-\frac{\hat{g}_m(X_t)}{\hat{h}_m(X_t)} - \phi(X_t) \right]^2$$

$\hat{f}_m(X) = \alpha \hat{\phi}_m(X)$.

(3) Update the model:

$\hat{f}_m(X) = \hat{f}_{m-1}(X) + \hat{f}_m(X)$.

end

3. Output $\hat{f}(X) = \hat{f}_M(X) = \sum_{m=0}^{M} \hat{f}_m(X)$.

2.2.3 GPR

GPR is a probabilistic supervised machine learning framework that has been widely used for

regression tasks (Rasmussen and Williams 2006, De Spiegeleer et al. 2018). It is a Bayesian method that starts from a prior Gaussian process to incorporate initial knowledge (kernels) and then combines with the observed data set, resulting in a posterior distribution of functions. Considering a training set of observations $\{(X_t, r_{t+1}\}_{t=1}^{n}$ where X_t is the explanatory variables of dimension k at month t and r_{t+1} are the oil price returns at month $t+1$. The GPR model assumes that

$$r_{t+1} = F(X_t) + \epsilon_t, \tag{3}$$

where $F(\cdot)$ is a Gaussian process and $\epsilon_t \sim N(0, \sigma_n^2)$ denote independent and identically distributed Gaussian random variables. The process $F(X_t)$ is only determined by its mean function $m(X)$ and a covariance or kernel function $c(X_s, X_t)$:

$$F \sim N(0, C(X, X)), \tag{4}$$

where $C(X, X)$ represents the covariance matrix, denoted as

$$C(X, X) = \begin{bmatrix} c(X_1, X_1) & \cdots & c(X_1, X_n) \\ c(X_2, X_1) & \cdots & c(X_2, X_n) \\ \vdots & \ddots & \vdots \\ c(X_n, X_1) & \cdots & c(X_n, X_n) \end{bmatrix}. \tag{5}$$

Consider a test set of explanatory variables X_* and denote the corresponding vector of functional value by F_*. Then the joint distribution of r and function values F_* is Gaussian,

$$\begin{bmatrix} r \\ F_* \end{bmatrix} \sim N(0, \begin{bmatrix} C(X, X) + \sigma_n^2 I & C(X, X_*) \\ C(X_*, X) & C(X_*, X_*) \end{bmatrix}). \tag{6}$$

To compute the Gaussian process posterior for the test set of explanatory variables, the joint distribution is conditioned on r is given by

$$\begin{aligned} F_* \mid X_*, X, r \sim N(&C(X_*, X)\left[C(X, X) + \sigma_n^2 I\right]^{-1} r, \\ &C(X_*, X_*) - C(X_*, X)\left[C(X, X) + \sigma_n{}^2 I\right]^{-1} C(X, X_*)) \end{aligned} \tag{7}$$

and predictions for the test set are calculated as the mean values of this distribution.

2.2.4 Support Vector Machine

The SVM, or support vector machine, is another most popular machine learning techniques to predict electricity prices (Papadimitriou, Gogas, and Stathakis 2014) and crude oil prices (Wang et al. 2020, Li, Zhu, and Wu 2019). Suppose $\{(X_t, r_{t+1})\}_{t=1}^{n}$ is a given set of training data, where $X_t = (x_{1t}, x_{2t}, ..., x_{kt})$ is the k explanatory variables at month t, and r_{t+1} is the oil price returns at month $t+1$. We aim to define a linear or nonlinear model that explains this data set in the best way. For a linear model of f for the input X can be written as

$$f(X) = < \omega, X > + b, \tag{8}$$

where ω denotes a weight vector, b is a bias term and $< \omega, X >$ represents the dot product of ω and X, i.e., $< \omega, X >= \omega^{\mathrm{T}} X$. For nonlinear regression problems, we first use a nonlinear kernel $g(.)$ to transform X and then the corresponding model of f is

$$f(X) = < \omega, g(X) > + b. \tag{9}$$

In order to have a good generalization performance, the weight vector ω needs to be as flat as possible, indicating that the norm ($\|.\|$) of ω should be minimized. We follow Smola and Schölkopf (2004) and minimize the Euclidean norm of ω. The optimization problem can be expressed as:

$$min \quad \frac{1}{2} \| \omega^2 \| + C \sum_{t=1}^{n} (\xi_t + \xi_t^*), \tag{10}$$

$$\text{subject to} \begin{cases} r_{t+1} - < \omega, X_t > - b \le \epsilon + \xi_t \\ < \omega, g(X_t) > + b - y_t \le \epsilon + \xi_t^* \\ \xi_t, \ \xi_t^* \ge 0 \end{cases}, \tag{11}$$

where ξ_t and ξ_t^* are two slack variables. The regularization parameter C determines the trade-off between the flatness of f and the penalizing error that exceeds a certain tolerance level ϵ.

Generally, the SVM is a robust and accurate method of regression, and its performance is

highly contingent on the setting of kernel function. However, so far there is no generally accepted procedure to find the proper kernel type (Beyca et al. 2019).

2.2.5 Ridge Regression and LASSO

The ridge regression (Hoerl and Kennard 1970) and LASSO (Tibshirani 1996) are broadly used in economic and financial prediction (Li, Tsiakas, and Wang 2014, Li and Tsiakas 2017). Both the ridge regression and LASSO uses a shrinkage method for linear regression. Mathematically, the ridge regression aims to minimize the loss function of the classical ordinary least squares plus the sum of squared coefficients while LASSO penalizes based on the sum of the absolute values of the coefficients.

For a linear regression of oil price returns r_{t+1} on explanatory variables X_t, the estimator $\hat{\beta}$ of ridge regression and LASSO is given by

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{t=1}^{n} (r_{t+1} - \beta_0 - \sum_{j=1}^{k} x_{jt}\beta_j)^2 + \alpha \sum_{j=1}^{k} |\beta_j|^p \right\} \tag{12}$$

where β_0 is the constant coefficient, x_{jt} is the j-th predictor of the explanatory vector X_t, β_j is the coefficient for the predictor x_{jt}, k is the number of predictors, and $\alpha \geq 0$ is the shrinkage parameter which is usually determined by the cross-validation method. When $p = 1, 2$, we get the estimator $\hat{\beta}$ of ridge regression and LASSO, respectively.

The ridge regression is mainly used in scenarios where the independent variables are highly correlated. By contrast, the LASSO model can retain the good features of variable selection and ridge regression because it can shrink some coefficients to exactly 0. However, both the ridge regression and LASSO model are essentially linear models, and they are not designed to capture the complex and non-linear relations between oil price returns and the explanatory variables.

2.2.6 KNN

KNN (Altman 1992) is a useful machine learning technique for both classification and regression. Its fundamental objective is to identify a group of close observations in the training set to make accurate predictions. In the case of regression, the output is the average value of its nearest neighbors. The concept of "nearest" is a metric which is based on a distance measure L_p:

$$L_p(X_s, X_t) = (\|X_s - X_t\|^p)^{\frac{1}{p}} = (\sum_{i=1}^{k} |x_{is} - x_{it}|^p)^{\frac{1}{p}}, \quad (13)$$

where $X_s = (x_{1s}, x_{2s}, ..., x_{ks})$ and $X_t = (x_{1t}, x_{2t}, ..., x_{kt})$ are two explanatory variables at month s and t, respectively. The parameter p ($p \geq 1$) is a positive real number. When $p = 2$, we have the frequently used Euclidean norm or the L_2 norm.

2.2.7 Combination Methods

The combination method is a popular technique that uses aggregation techniques to enhance predictive performance. To avoid model uncertainty and instability of individual predictive models, Rapach, Strauss, and Zhou (2010) claimed that combining the models' forecasts can consistently show statistically and economically significant predictive power relative to the historical average over time. We thus included the combination methods as another competing model. Assuming we have N individual forecasts, the combination forecast is given as follows,

$$\hat{r}_{c,t+1} = \sum_{i=1}^{N} \omega_{i,t} \hat{r}_{i,t+1}, \quad (14)$$

where $\hat{r}_{c,t+1}$ is the combination forecast at month $t+1$. $\hat{r}_{i,t+1}$ is the individual forecast at month t + 1, and $\omega_{i,t}$ is the ex-ante combining weight of the i-th individual forecast formed at the end of month t.

We follow Rapach, Strauss, and Zhou (2010), Zhu and Zhu (2013), and Zhang, Ma, and Wang (2019) and used the popular mean combination method. The mean combination method is computed

as the mean value of the N individual forecasts, implying that $\omega_{i,t} = 1/N$.

3 Data

This section describes the data sources for crude oil prices and predictor variables. Our included predictors consist of macroeconomic variables and technical indicators.

3.1 Oil prices

Crude oil prices were highly volatile and widely used to test the predictive power of various models (Regnier 2007, Zhang, Ma, and Wang 2019). We employed one of the dominant proxies of crude oil price, the monthly spot price of WTI crude oil, because it is widely used to forecast crude oil prices (Baumeister and Kilian 2014, Zhang, Ma, and Wang 2019). We also included two other popular proxies of crude oil prices, Europe Brent spot prices (Brent hereafter) and the U.S. crude oil imported acquisition cost by refiners (RAC hereafter) for robustness checks (He et al. 2021). All the three nominal monthly crude oil prices were retrieved from the U.S Energy Information Administration (EIA) website.[1] We deflated the crude oil prices with the U.S. consumer price index (CPI) to construct the real oil price series. The CPI data series was downloaded from the Federal Reserve Economic Data (FRED) of St. Louis[2].

In line with He et al. (2021), Rapach et al. (2010), Yin and Yang (2016), and Zhang et al. (2019), we generated the out-of-sample forecasts of real oil returns using an expanding estimation window. Our entire sample period ranged from January 1986 to December 2019. We select the period from January 1986 to December 2000 for the in-sample estimation period, and the out-of-sample forecast started from January 2001.

[1] https://www.eia.gov/
[2] https://fred.stlouisfed.org/

3.2 Technical indicators

We considered two categories of predictors in this study: technical indicators and macroeconomic variables. We first included 18 popular technical indicators to forecast spot oil price returns since these technical indicators were documented to show significant in-sample and out-of-sample predictive power (Yin and Yang 2016, He et al. 2021). These 18 technical indicators are equally divided into three groups based on their respective rules: the moving average rule, the momentum rule, and the on-balance volume rule.

The first group of 6 technical indicators is based on the moving average (MA) rule which constructs a buy or sell signal ($S_{i,t}=1$ or $S_{i,t}=0$, respectively) by comparing two moving averages at the end of month t.

$$S_{i,t}=\begin{cases}1 \text{ if } MA_{s,t}\geq MA_{l,t}\\ 0 \text{ if } MA_{s,t}<MA_{l,t}\end{cases} \tag{15}$$

where

$$MA_{j,t}=\frac{1}{j}\sum_{i=0}^{j-1}P_{t-i} \text{ for } j=s,l, \tag{16}$$

P_t is the crude oil spot price at the end of month t. s and l are the time length in month of short and long MAs, respectively. The buy (sell) signal is released when the short-term MA is larger (less) than the long-term MA. The MA rule captures the variation of trends in the oil price time series. We considered six (3*2) technical indicators with $s=1,2,3$ and $l=9,12$.

The second group with six indicators employs a momentum (MOM) rule that produces a trading signal by comparing the current oil price and its level in the past m months,

$$S_{i,t}=\begin{cases}1 \text{ if } P_t\geq P_{t-m}\\ 0 \text{ if } P_t<P_{t-m}\end{cases}. \tag{17}$$

We considered $m=1,2,3,6,9$, and 12 in the six momentum indicators. These momentum indicators exhibit their respective market condition of the past few months.

The third group with six indicators is based on the on-balance volume (OBV) rule. In conjunction with past prices, the volume data were widely used to identify market trends. A trading signal of OBV is given as follows,

$$S_{i,t}=\begin{cases}1 \text{ if } MA_{s,t}^{OBV}\geq MA_{l,t}^{OBV}\\ 0 \text{ if } MA_{s,t}^{OBV}< MA_{l,t}^{OBV}\end{cases}, \tag{18}$$

where

$$MA_{j,t}^{OBV}=\frac{1}{j}\sum_{i=0}^{j-1}OBV_{t-i} \text{ for } j=s,l \tag{19}$$

$$OBV_t=\sum_{k=1}^{t}VOL_k D_k . \tag{20}$$

VOL_k denotes the trading volume during the month k, and D_k is a binary variable that takes a value of 1 if $P_k-P_{k-1}\geq 0$ and -1 otherwise. Particularly, we used six (3*2) OBV indicators with $s=1,2,3$ and $l=9,12$.

We followed Yin and Yang (2016) and He et al. (2021) and used the closing prices and the trading volumes of crude oil futures Contract 1 (with 1-month maturity) traded on NYMEX to construct 18 technical indicators.

3.3 Macroeconomic variables

We followed Neely et al. (2014), combined technical indicators with macroeconomic variables to provide complementary information over the business cycle, and improved predictive power. We followed Yin and Yang (2016) and considered 18 macroeconomic variables and updated data series in our study. These 18 variables are divided into three groups. The first set of 8 macroeconomic variables are highly related to stock return predictability (Welch and Goyal 2008): (1) Book-to-market ratio (BM), (2) Treasury bill rate (TB), (3) Long-term yield (GB), (4) Term Spread (TS), (5) Inflation (CPI), (6) Dividend-price ratio (DP), (7) Dividend yield (DY), and (8) Earnings-price ratio

(EP). The series of DP, DY, and EP are downloaded from the personal website of Amit Goyal[3] and taken log treatment.

We employed the second set of 6 macroeconomic variables to reflect the general condition of the economy: (9) the monthly unemployment rate in the USA (UER), (10) the growth of M2 money stock (M2), (11) the monthly growth in the U.S. industrial production index (GIP), (12) log of the capacity utilization in manufactory index (CUM), (13) log of the monthly purchasing managers' index in the USA (PMI), and (14) Chicago Fed's national activity index (NAI). The PMI data were retrieved from the Wind database[5], while the other five macroeconomic variables are downloaded from the website of the FRED of St. Louis (FRED).

The third set of 4 macroeconomic variables were to detect variation in demand and supply pressures in crude oil markets. We included two variables to capture the demand for commodities in global markets: (15) Kilian's real global economic activity index (KI) and (16) log of the U.S. field production of crude oil (FPO). The KI data can be obtained from FRED, while FPO data is downloaded from the website of EIA. Another two incorporated variables reflected the fluctuations of financial markets: (17) log of NYSE Arca oil index (OI), obtained from Yahoo Finance, and (18) log of real U.S. trade-weighted real exchange rate (TWI), available on FRED.

4 Empirical analyses and discussions

This section discussed the out-of-sample performance of the DFEA and the competing models. We employed three evaluation criteria: out-of-sample R square, success ratio, and CER gain. We also conducted the forecast encompassing test to examine whether the information contained in the

[3] https://sites.google.com/view/agoyal145/?redirpath=/
[5] https://www.wind.com.cn/

DFEA dominates the information found in the competing models or not. We finally identified the factors contributing the most to the predictive power of the DFEA.

4.1 Out-of-sample performance

Welch and Goyal (2008) underlined that an out-of-sample test is more reliable in assessing predictive power because in-sample predictability is probably caused by over-fitting. We thus focused more on out-of-sample performance. Specifically, we used the data starting from January 1986 up to time t to construct the out-of-sample forecast at time $t+1$. The out-of-sample evaluation period ranged from January 2001 to December 2019, including business cycles of busts and booms. We used the out-of-sample test to compare the prediction performance of the DFEA with those of three classical machine learning methods and their combination methods discussed in Section 2.

We followed He et al. (2021) and Yin and Yang (2016) and employed the widely used out-of-sample R square (R_{OS}^2) statistics by Campbell and Thompson (2008) to evaluate the out-of-sample performance. In terms of the benchmark, we first considered the no-change forecast which simply considers the current oil return at month *t* as oil return of month *t*+1. This is a popular benchmark model used by Baumeister and Kilian (2012), Wang et al. (2017), Zhang et al. (2018), and others. We also employed another prevailing benchmark model, the historical average benchmark forecast, for robustness checks for a more subjective comparison. The out-of-sample R square is given as follows,

$$R_{OS}^2 = 1 - \frac{MSPE_M}{MSPE_B}, \tag{21}$$

where

$$MSPE_M = \frac{1}{q}\sum_{i=1}^{q}(r_{m+i} - \hat{r}_{m+i})^2, MSPE_B = \frac{1}{q}\sum_{i=1}^{q}(r_{m+i} - \hat{r}_{B,m+i})^2 \ , \tag{22}$$

$MSPE_M$ is the mean squared prediction error (MSPE) of the forecasting model of interest, and $MSPE_B$ is the MSPE of the benchmark model. While r_{m+i}, $\hat{r}_{m+i}$ and $\hat{r}_{B,m+i}$ denote the actual oil

return, the oil return predicted by the DFEA and eight competing models in Section 2, and the oil return predicted by the benchmark model at month $m+i$, respectively. The parameters m and q denote the length of the in-sample period and out-of-sample period, respectively. Intuitively, the R_{OS}^2 measures the proportional change in MSPE for the forecasting model of interest relative to the benchmark model. Specifically, a positive or negative value of R_{OS}^2 statistics suggests the forecasting model outperforms or underperforms the benchmark model, respectively.

We proceeded to ascertain whether the predictive model of interest significantly improves MSPE relative to the benchmark model. We adopted the $MSFE-adjusted$ statistics definition by Clark and West (2007)[7] to compare the MSPE of the benchmark model with that of the predictive model. Specifically, the $MSFE-adjusted$ statistics is used to test the null hypothesis that the MSPE of the benchmark model is less than or equal to that of the predictive model against the alternative hypothesis that the MSPE of the benchmark model is greater than that of the predictive model. That is, $H_0: R_{OS}^2 \le 0$ against $H_A: R_{OS}^2 > 0$. However, Leitch and Tanner (1991) argued that forecasting the market movement is more important than an accurate point forecast since an investor's profit has a stronger relationship with directional accuracy than point accuracy, as measured by the estimation statistics such as MSPE and R_{OS}^2. Besides, Vrontos, Galakis, and Vrontos (2021) documented that machine learning methods can deliver substantial improvement to the detection of directional patterns. We need to empirically verify this advantage. Thus, we computed success ratios to test the accuracy of the predictive models and conducted the directional accuracy test of Pesaran and Timmermann (1992). The null hypothesis states that the success ratio of the predictive model of interest is less or equal to the success ratio of random walk.

Moreover, we also considered a representative investor who relies on mean-variance framework and determines asset allocation decisions between oil futures and risk-free bills based on various forecasts of oil returns. We calculated the CER for each forecasting model to compare their economic gains. The investor determines the optimal ratio of the portfolio to oil futures at the

[7] The Diebold-Mariano test (Diebold and Mariano 2002) is another popular test to compare the predictive accuracy of various forecasts. In terms of the predictions of crude oil prices, the test of Clark and West (2007) is a more popular alternative (Zhang, Ma, and Wang 2019, He et al. 2021).

end of month t as

$$\omega_t = \frac{1}{\gamma}\frac{\hat{r}_{t+1}}{\hat{\sigma}^2_{t+1}}, \tag{23}$$

where γ is the degree of risk aversion of the investor, $\hat{r}_{t+1}$ is the out-of-sample forecast of oil returns, and $\hat{\sigma}^2_{t+1}$ denotes the estimate of the oil return variance at month $t+1$. We followed Campbell and Thompson (2008) and adopts a 5-year fixed-length moving window of past oil returns to estimate oil return variance. The ratio of risk-free bills is $1-\omega_t$, and thus the realized portfolio return (r^P_{t+1}) at month $t+1$ is given as

$$r^P_{t+1} = \omega_t r_{t+1} + r^f_{t+1}, \tag{24}$$

where r_{t+1} is oil returns and r^f_{t+1} is the risk-free return at month $t+1$. The investor can realize a CER of

$$\text{CER} = \bar{r}^P - \frac{1}{2}\gamma\sigma^2(r^P), \tag{25}$$

where $\bar{r}^P$ and $\sigma^2(r^P)$ denote the mean and variance of portfolio returns over the out-of-sample period, respectively. The CER gain for the mean-variance investor is defined as the difference of the CER of forecasts generated by various predictive models and those of the benchmark forecasts. We multiply the CER difference by 12 and annualize CER gain.

The CER is the risk-free rate of return that an investor would be willing to accept in lieu of holding up the risky portfolio. This CER gain can be interpreted as the annual percentage portfolio management fee that an investor is willing to pay to access the various forecasts in place of the benchmark forecasts. We followed Rapach, Ringgenberg, and Zhou (2016) and exhibited the results for $\gamma = 3$ and limited the weight ω_t to -0.5 and 1.5, imposing pragmatic portfolio restrictions and adopting well-constructed portfolio weights.

In terms of estimation, we employed the typical five-fold cross-validation method to determine the key parameters in each competing model and Table 1 illustrated the hyperparameter settings.

We use the *scikit-learn* package[8] (Pedregosa et al. 2011) in Python to implement these machine learning algorithms[9]. For the DFEA, we adopted the default settings by the *deepforest* module[10] in Python. We did not finetune the hyperparameters in the DFEA, which is a well-designed machine learning algorithm and confirmed to achieve performance with its default settings (Zhou and Feng 2017).

Table 2 reported the out-of-sample forecasting results and demonstrated three important observations. First, the DFEA forecast outperformed all the competing models in terms of R_{OS}^2 statistics. The full-sample R_{OS}^2 statistics for all predictive models were significantly above zero, indicating that all the forecasting models outperformed the no-change benchmark forecast. Among these models, the DFEA performs the best with its R_{OS}^2 of 40.04%.

Second, the DFEA delivered significantly larger success ratios than its competing models. The DFEA reported a remarkable directional accuracy of 64.76%, a success ratio that exceeded those of all the competing models. This finding is crucial for an investor who can take only a long or a short position of oil futures from a market timing perspective because a market timing strategy responds solely to the sign of market movement in oil prices. Section 6 further discussed the economic value of the DFEA from the market timing perspective.

Third, we observed that all the forecasting models can deliver positive annualized CER gains, except for GPR and KNN. Specifically, DFEA delivered the second largest CER gain of 20.33%, suggesting an investor is willing to pay an annual fee of up to 20.33% in order to access DFEA forecast. The best CER gain of 22.25% is generated by the Ridge forecast, which is 1.9% larger than

[8] https://github.com/scikit-learn/scikit-learn
[9] The XGBoost algorithm is implemented using the *xgboost* module (https://xgboost.readthedocs.io/en/stable/).
[10] https://github.com/LAMDA-NJU/Deep-Forest

that of the DEFA forecast. The mean combination method delivers the fourth best CER gain of 15.75%.

We followed Yin and Yang (2016), Zhang, Ma, and Wang (2019), and He et al. (2021), and summarized the time series plots of the differences between the cumulative squared forecast errors for the benchmark model and those of the competing models in Figure 2. The rising curves in each panel of Figure 2 indicated that all the forecasting models showed strong predictability for most out-of-sample periods. Five forecasts, XGBoost, GPR, SVM, LASSO, and KNR, unveiled a significant decrease during the global financial crisis since 2008, reflecting a large reduction of predictive power during recessions. That is, market turmoil critically challenged the predictive power of those five methods which are not designed to predict market downfalls. On the other hand, the DFEA exhibited a continuous upward trend despite a slight drop around 2009. It indicated that the DFEA lost a limited degree of predictive power during major crisis but soon recovered and accumulated its stronger predictive ability over time.

4.2 Forecast encompassing test

The forecast encompassing test (Harvey, Leybourne, and Newbold 1998) compares and examines whether the information content of one forecast dominates the other competitor (Zhang, Ma, and Wang 2019, He et al. 2021). The basic idea is to form an optimal composite forecast of r_{t+1} as a convex combination of the forecasts based on model j and model h, which can be expressed as follows,

$$r_{t+1} = (1-\delta)\hat{r}_{j,t+1} + \delta\hat{r}_{h,t+1}, \tag{26}$$

where $0 \le \delta \le 1$. Subscripts j and h index the competing model and the DFEA, respectively. Specifically, if $\delta(1-\delta) = 0$, model j (h) encompasses the model h (j). That is, model j

(h) contains all of the useful information found in the model h (j) to form the optimal composite forecast. In contrast, if $\delta(1-\delta)>0$, then the model j (h) forecast does not encompass the model h (j) forecast. Thus, the model h (j) does contain relevant information to form the optimal composite forecast. Essentially, rejecting the null hypothesis ($H_0:\delta(1-\delta)=0$) of encompassing test implies that combining forecasts from models j and h would be helpful, rather than relying solely on the model j or h. The alternative hypothesis ($H_1:\delta(1-\delta)>0$) is that model $j(h)$ forecast does not encompass the model h (j).

Table 3 listed the p-values of the forecast encompassing tests of the out-of-sample forecasts. We set the significance level at 10%. For all the competing models, if $\delta=0$ was significant at 10% level, we should reject the null hypothesis that the competing models encompass the DFEA forecast. If we cannot reject the null hypothesis that $1-\delta=0$ for all competing models, we accept the null hypothesis that the forecast based on the DFEA encompasses all the competing models. Overall, the DFEA was empirically confirmed to contain significantly more relevant information for forecasting oil returns than the other competing models.

4.3 What does the predictive power originate from?

We further explored the group of explanatory variables contributing the highest predictability of the DFEA. To this end, we used the technical indicators and macroeconomic variables separately to generate out-of-sample forecasts for the DFEA, and Table 4 summarized the outcomes. First, the DFEA with all predictors yielded the largest R_{OS}^2, indicating that the macroeconomic variables can deliver additional information and enhance the predictive ability. Second, the DFEA with technical indicators generated substantially larger R_{OS}^2 than the one with the macroeconomic variables. This finding coincided with the existing studies, such as Yin and Yang (2016) and Zhang, Ma, and Wang

(2019). The success ratio outcomes provided further support that technical indicators exhibit stronger predictive power than macroeconomic variables. In general, the oil prices are more predictable with technical indicators. Based on predictability, it implies an informationally inefficient market, coinciding with Neely et al. (2014).

Furthermore, we investigated which group of technical indicators provided the principal source of the strong predictability of technical predictors: MA, MOM, and OBV. Table 5 reported the forecasting results of the DFEA with different groups of technical indicators. Among all the three groups of technical indicators, the DFEA with MOM showed an R_{OS}^2 of 34.59%, very close to the R_{OS}^2 of the DFEA with all technical indicators. This suggested that the MOM dominated the source of the strong predictive power of technical indicators in terms of R_{OS}^2 estimates. More noteworthy, we found that the DFEA displayed a directional accuracy of 68% only with MOM, and this success ratio was 4% larger than that of the DFEA with all technical indicators. This observation coincides with several recent studies that justify the effectiveness of MOM in stock markets (Moskowitz, Ooi, and Pedersen 2012, Neely et al. 2014). Specifically, the outperformance of MOM can be attributed to the presence of late-informed investors (Gao et al. 2018). Considering good news is released at month t, some investors can react at month t while others process the news more slowly since several previous studies showed that investors can react even to month-old information (Baker and Wurgler 2006, Hong, Torous, and Valkanov 2007, Cohen and Frazzini 2008). These observations consolidated significant number of lag in information and the consequent market inefficiency. Trading in the same direction as month t can generate a correlated positive return in month $t+1$.

5 Robustness checks

We reexamined the predictability of the DFEA and compared its performance from the prospective of neural networks (5.1), change of benchmark forecast (5.2), various forecasting window sizes (5.3), two other prevailing proxies of crude oil prices (5.4), and nominal prices of crude oil (5.5).

5.1 Comparison with neural networks

Zhou and Feng (2019) employed several neural networks to compare their performance with DFEA. We followed their analysis design and compared the forecasting performance of neural networks with different architectures in Table 6. In terms of architectures of neural networks in financial predictions, we followed Gu, Kelly, and Xiu (2020) and considered architectures with depth of hidden layers ranging from 1 to 5, denoting them by NN1 to NN5. We noticed that NN2 generally performed the best of all the five neural networks with R_{OS}^2, success ratio, and CER gain of 38.75%, 67.84%, and 22.62, respectively. However, DFEA beat NN2 with an R_{OS}^2 of 40.04%. The success ratio and CER gain of DFEA were slightly lower than those of NN2. We thus concluded that DFEA is able to achieve highly competitive performance to those of neural networks in forecasting crude oil prices.

5.2 Alternative choices of benchmark forecast

We followed Rapach, Strauss, and Zhou (2010) and employed the historical average benchmark to test the robustness of the out-of-sample forecasting performance. Table 7 showed that the DFEA had stronger predictive power than the historical average benchmark and the other competing models. Evidently, the DFEA performed the best in terms of R_{OS}^2 statistic and success ratio of 14.6% and 64.76, respectively. The DEFA also displayed the second largest CER gain of 57.34%.

5.3 Different forecasting window sizes

Rossi and Inoue (2012) documented that a forecast window size plays its role in determining the out-of-sample performance. To reduce the forecasting bias stemming from different forecasting windows, we followed Zhang, Ma, and Wang (2019) and He et al. (2021) and used another two out-of-sample forecasting windows for the detailed evaluation. Specifically, the first one forecasting window ranged from January 1996 to December 2019, and the second one spanned from January 2006 to December 2019. Table 8 reports the out-of-sample forecasting performance based on these two forecasting windows. The DFEA ranks among the top two in terms of R_{OS}^2 estimate, success ratio, and CER gain.

5.4 Alternative proxies of crude oil prices

In addition to WTI, Brent and RAC are two other popular proxies of crude oil prices. We further used their prices of Brent and RAC, deflated by the U.S. CPI, to examine the predictive ability of various models. Table 9 reported the out-of-sample forecasting results based on real prices of Brent and RAC. When we predicted the real prices of Brent, the DFEA yielded the largest out-of-sample R square value of 41.1%. For the real prices of RAC, we found that the out-of-sample R squares of the DFEA is 27.26%, which is close to the best out-of-sample R squares of 27.57% generated by the mean combination method (Mean). The DFEA also demonstrated an impressive capacity to attain remarkably competitive performance in both success ratio and CER gain when predicting the real prices of Brent and RAC. We concluded that our out-of-sample forecasting results are robust to alternative proxies of crude oil price, and the DFEA generally presents excellent predicting performance.

5.5 Nominal prices of crude oil

While most existing literature focused on the forecasting performance of real crude oil prices,

some other studies targeted on the nominal oil price, such as Alquist, Kilian, and Vigfusson (2013), Zhang, Ma, and Wang (2019) and He et al. (2021). We also considered the nominal prices of WTI and Brent to compare the forecasting ability as robustness check. Table 10 reported the main out-of-sample results based on the two nominal prices. We found that the DFEA showcases either the best or near-best performance with respect to all the three evaluation criteria. In summary, we confirmed the general outperformance of the DFEA robust to predicting nominal crude oil prices.

6 The economic value of DFEA for market timing

Our empirical analyses confirmed that the DFEA yielded the largest success ratios when we forecasted log real returns on WTI, Brent, and RAC crude oil prices. This empirical evidence indicated that the DFEA exceled in detecting the market movement of oil prices. We thus took advantage of the remarkable directional accuracy of the DFEA forecast and applied this method to crude oil futures markets. We thus formed market timing strategies because these strategies relied exclusively and heavily on timing or directional signals of oil prices at month $t+1$ when we decide to buy or sell crude oil futures at the end of month t. We used the oil prices of futures Contract 1 traded on the NYMEX and the Brent crude futures contract traded on ICE, instead of the spot prices of WTI and Brent, to gauge economic gains. These two crude oil futures prices are more tractable than their oil spot prices.

In line with Gao et al. (2018) and Zhang, Ma, and Zhu (2019), we implemented the market timing test as follows. Denoting the current month by t, we used the forecasting log return on crude oil at a month $t+1$ as a timing signal to trade NYMEX WTI crude futures or ICE Brent crude oil futures. Specifically, we entered into a long position of oil futures at the beginning of month $t+1$ if a positive timing signal is given (that is, the forecasting log return at month $t+1$ is positive), or

took a short position otherwise. The long or short position is closed at the end of month $t+1$. The market timing strategy based on the forecasting log return for month $t+1$ can generate a return as,

$$\lambda(t+1)=\begin{cases} \mathrm{r}_{t+1}, & \text{if } \hat{r}_{t+1}>0 \\ -r_{t+1}, & \text{if } \hat{r}_{t+1}\le 0 \end{cases}, \tag{27}$$

where $\hat{r}_{t+1}$ denotes the forecasting log return using the DFEA or the competing models at month $t+1$, r_{t+1} is the simple return of taking a long position for oil futures at month $t+1$, and $\lambda(t+1)$ is the realized return of month $t+1$ and the time subscript t started from December 2000. To measure the performance of various market timing strategies, we employed the *Always Long* strategy as a benchmark, taking a long position of oil futures from the beginning of the month $t+1$ and closing it at the end of the month $t+1$. We did not consider transaction costs, which is another issue to explore.

Table 11 reported the market timing performance of nine market timing strategies and the *Always Long* strategy. We found that the DFEA market timing strategy yielded the largest average return of 77.1% per year if we traded the NYMEX WTI crude oil futures. We measured the risk by standard deviation and used the Sharpe ratio to gauge the strategy performance. The DFEA strategy generated a Sharpe ratio of 1.02 and outperformed all other competing strategies and the *Always Long* benchmark strategy. The results of Brent crude oil futures also showed the same conclusion that the DFEA strategy delivered a remarkable annual average return and a Sharpe ratio of 71.10% and 0.93, respectively. The DFEA market timing strategy outperformed all the competing market timing strategies and the *Always Long* strategy. In brief, we recommended investors taking a long position at the beginning of month $t+1$ if the DFEA forecast displayed a positive timing signal, or taking a short position otherwise. The DFEA market timing strategy delivered considerable economic gains in terms of average return and Sharpe ratio.

7 Conclusions

We used the DFEA by Zhou and Feng (2017) to empirically forecast crude oil prices with a large data set containing 36 explanatory variables. We employed three evaluation criteria (the out-of-sample R squares, the success ratios, and CER gains) to compare the out-of-sample performance of the DFEA and the eight competing models. The out-of-sample forecasting results showed that the DFEA generally outperformed all the competing models. The forecast encompassing test results showed that the DFEA contained more relevant information than its competing models in forecasting oil returns.

In addition, we found that the technical indicators demonstrated superior predictive ability than macroeconomic variables. This conclusion coincided with some existing studies (Zhang, Ma, and Wang 2019, He et al. 2021). More importantly, the momentum predictors contributed the most to the strong predictive power of the technical indicators, in contrast with moving average indicators and volume indicators. The outperformance of momentum indicators can be attributed to the presence of late-informed investors. These conclusions were robust with respect to neural networks, choices of benchmark forecast, forecasting window sizes, proxies of crude oil prices, and nominal crude oil prices.

Moreover, in terms of practical application, we exploited the more accurate predictive performance of the market movement of the DFEA and realized substantial economic gains from a market timing perspective. To prove our market timing strategies, we traded the two major crude oil futures, NYMEX WTI crude oil futures and ICE Brent crude oil futures. We found that the market timing strategy based on the DFEA yielded the largest average return and Sharpe ratio for oil futures.

Overall, these detailed investigations confirmed that the DFEA contributed significantly to the accuracy of the crude oil price forecasting and delivered substantial economic gains from a market timing perspective. These conclusions have their solid contribution in investment decision on crude oil price based on their rigorous empirical examination of the DFEA.

Bibliography

Alquist, Ron, Lutz Kilian, and Robert J Vigfusson. 2013. "Forecasting the price of oil." In *Handbook of Economic Forecasting*, 427-507. Amsterdam, North-Holland: Elsevier.

Alshater, M. M., I. Kampouris, H. Marashdeh, O. F. Atayah, and H. Banna. 2022. "Early warning system to predict energy prices: the role of artificial intelligence and machine learning." *Annals of Operations Research*:1-37.

Altman, Naomi S. 1992. "An introduction to kernel and nearest-neighbor nonparametric regression." *The American Statistician* 46 (3):175-185.

Baker, Malcolm, and Jeffrey Wurgler. 2006. "Investor sentiment and the cross-section of stock returns." *Journal of Finance* 61 (4):1645-1680.

Baumeister, Christiane, and Lutz Kilian. 2014. "What central bankers need to know about forecasting oil prices." *International Economic Review* 55 (3):869-889.

Beyca, Omer Faruk, Beyzanur Cayir Ervural, Ekrem Tatoglu, Pinar Gokcin Ozuyar, and Selim Zaim. 2019. "Using machine learning tools for forecasting natural gas consumption in the province of Istanbul." *Energy Economics* 80:937-949.

Boubaker, S., Z. Liu, and Y. Zhang. 2022. "Forecasting oil commodity spot price in a data-rich environment." *Annals of Operations Research*:1-18.

Breiman, Leo. 2001. "Random forests." *Machine Learning* 45 (1):5-32.

Campbell, John Y., and Samuel B. Thompson. 2008. "Predicting excess stock returns out of sample: Can anything beat the historical average?" *Review of Financial Studies* 21 (4):1509-1531.

Chen, Tianqi, and Carlos Guestrin. 2016. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.

Clark, Todd E, and Kenneth D West. 2007. "Approximately normal tests for equal predictive accuracy in nested models." *Journal of Econometrics* 138 (1):291-311.

Cohen, Lauren, and Andrea Frazzini. 2008. "Economic links and predictable returns." *Journal of Finance* 63 (4):1977-2011.

Costa, Alexandre Bonnet R., Pedro Cavalcanti G. Ferreira, Wagner P. Gaglianone, Osmani Teixeira C. Guillén, João Victor Issler, and Yihao Lin. 2021. "Machine learning and oil price point and density forecasting." *Energy Economics* 102:105494.

Creamer, Germán G., and Chihoon Lee. 2019. "A multivariate distance nonlinear causality test based on partial distance correlation: A machine learning application to energy futures." *Quantitative Finance* 19 (9):1531-1542.

De Spiegeleer, Jan, Dilip B Madan, Sofie Reyners, and Wim Schoutens. 2018. "Machine learning for quantitative finance: fast derivative pricing, hedging and fitting." *Quantitative Finance* 18 (10):1635-1643.

Diebold, Francis X, and Robert S Mariano. 2002. "Comparing predictive accuracy." *Journal of Business & Economic Statistics* 20 (1):134-144.

Ferrari, Davide, Francesco Ravazzolo, and Joaquin Vespignani. 2021. "Forecasting energy commodity prices: A large global dataset sparse approach." *Energy*

Economics 98:105268.
Gao, Lei, Yufeng Han, Sophia Zhengzi Li, and Guofu Zhou. 2018. "Market intraday momentum." *Journal of Financial Economics* 129 (2):394-414.
Gu, Shihao, Bryan Kelly, and Dacheng Xiu. 2020. "Empirical asset pricing via machine learning." *Review of Financial Studies* 33 (5):2223-2273.
Guidolin, Massimo, and Manuela Pedio. 2020. "Forecasting commodity futures returns with stepwise regressions: Do commodity-specific factors help?" *Annals of Operations Research* 299 (1-2):1317-1356.
Guo, Xiaopeng, DaCheng Li, and Anhui Zhang. 2012. "Improved support vector machine oil price forecast model based on genetic algorithm optimization parameters." *AASRI Procedia* 1:525-530.
Harvey, David I., Stephen J. Leybourne, and Paul Newbold. 1998. "Tests for Forecast Encompassing." *Journal of Business & Economic Statistics* 16 (2):254-259.
He, Mengxi, Yaojie Zhang, Danyan Wen, and Yudong Wang. 2021. "Forecasting crude oil prices: A scaled PCA approach." *Energy Economics* 97:105189.
Hoerl, Arthur E., and Robert W. Kennard. 1970. "Ridge regression: biased estimation for nonorthogonal problems." *Technometrics* 12 (1):55-67.
Hong, Harrison, Walter Torous, and Rossen Valkanov. 2007. "Do industries lead stock markets?" *Journal of Financial Economics* 83 (2):367-396.
Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. 1989. "Multilayer feedforward networks are universal approximators." *Neural Networks* 2 (5):359-366.
Hou, Keqiang, Dean C. Mountain, and Ting Wu. 2016. "Oil price shocks and their transmission mechanism in an oil-exporting economy: A VAR analysis informed by a DSGE model." *Journal of International Money and Finance* 68:21-49.
Kilian, Lutz, and Cheolbeom Park. 2009. "The impact of oil price shocks on the U.S. stock market." *International Economic Review* 50 (4):1267-1287.
Leitch, Gordon, and J Ernest Tanner. 1991. "Economic forecast evaluation: profits versus the conventional error measures." *American Economic Review* 81 (3):580-590.
Li, J., I. Tsiakas, and W. Wang. 2014. "Predicting exchange rates out of sample: Can economic fundamentals beat the random walk?" *Journal of Financial Econometrics* 13 (2):293-341.
Li, Jiahan, and Ilias Tsiakas. 2017. "Equity premium prediction: The role of economic and statistical constraints." *Journal of Financial Markets* 36:56-75.
Li, Jinchao, Shaowen Zhu, and Qianqian Wu. 2019. "Monthly crude oil spot price forecasting using variational mode decomposition." *Energy Economics* 83:240-253.
Ma, Chao, Zhenbing Liu, Zhiguang Cao, Wen Song, Jie Zhang, and Weiliang Zeng. 2020. "Cost-sensitive deep forest for price prediction." *Pattern Recognition* 107:107499.
Miao, Hong, Sanjay Ramchander, Tianyang Wang, and Dongxiao Yang. 2017. "Influential factors in crude oil price forecasting." *Energy Economics* 68:77-88.

Moskowitz, Tobias J, Yao Hua Ooi, and Lasse Heje Pedersen. 2012. "Time series momentum." *Journal of Financial Economics* 104 (2):228-250.

Neely, Christopher J., David E. Rapach, Jun Tu, and Guofu Zhou. 2014. "Forecasting the equity risk premium: The role of technical indicators." *Management Science* 60 (7):1772-1791.

Papadimitriou, Theophilos, Periklis Gogas, and Efthimios Stathakis. 2014. "Forecasting energy markets using support vector machines." *Energy Economics* 44:135-142.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. 2011. "Scikit-learn: machine learning in Python." *Journal of Machine Learning Research* 12:2825-2830.

Pesaran, M Hashem, and Allan Timmermann. 1992. "A simple nonparametric test of predictive performance." *Journal of Business & Economic Statistics* 10 (4):461-465.

Rapach, David E, Matthew C Ringgenberg, and Guofu Zhou. 2016. "Short interest and aggregate stock returns." *Journal of Financial Economics* 121 (1):46-65.

Rapach, David E., Jack K. Strauss, and Guofu Zhou. 2010. "Out-of-sample equity premium prediction: combination forecasts and links to the real economy." *Review of Financial Studies* 23 (2):821-862.

Rasmussen, C, and C Williams. 2006. "Gaussian processes for machine learning." *the MIT press*.

Regnier, Eva. 2007. "Oil and energy price volatility." *Energy Economics* 29 (3):405-427.

Rossi, Barbara, and Atsushi Inoue. 2012. "Out-of-sample forecast tests robust to the choice of window size." *Journal of Business & Economic Statistics* 30 (3):432-453.

Smola, Alex J, and Bernhard Schölkopf. 2004. "A tutorial on support vector regression." *Statistics and computing* 14:199-222.

Sun, W., H. Chen, F. Liu, and Y. Wang. 2022. "Point and interval prediction of crude oil futures prices based on chaos theory and multiobjective slime mold algorithm." *Annals of Operations Research*:1-31.

Tang, Cheng, Damien Garreau, and Ulrike von Luxburg. 2018. "When do random forests fail?" *Proceedings of the 32nd International Conference on Neural Information Processing Systems* 31:2987–2997.

Tibshirani, Robert. 1996. "Regression shrinkage and selection via the Lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1):267-288.

Vrontos, Spyridon D., John Galakis, and Ioannis D. Vrontos. 2021. "Implied volatility directional forecasting: A machine learning approach." *Quantitative Finance* 21 (10):1687-1706.

Wang, Jue, George Athanasopoulos, Rob J. Hyndman, and Shouyang Wang. 2018. "Crude oil price forecasting based on internet concern using an extreme learning machine." *International Journal of Forecasting* 34 (4):665-677.

Wang, Jue, Xiang Li, Tao Hong, and Shouyang Wang. 2018. "A semi-heterogeneous

approach to combining crude oil price forecasts." *Information Sciences* 460:279-292.

Wang, Jue, Hao Zhou, Tao Hong, Xiang Li, and Shouyang Wang. 2020. "A multi-granularity heterogeneous combination approach to crude oil price forecasting." *Energy Economics* 91:104790.

Welch, Ivo, and Amit Goyal. 2008. "A comprehensive look at the empirical performance of equity premium prediction." *Review of Financial Studies* 21 (4):1455-1508.

Xia, Min, Namei Tian, Yonghong Zhang, Yiqing Xu, and Xu Zhang. 2020. "Dilated multi-scale cascade forest for satellite image classification." *International Journal of Remote Sensing* 41 (20):7779-7800.

Yin, Libo, and Qingyuan Yang. 2016. "Predicting the oil prices: Do technical indicators help?" *Energy Economics* 56:338-350.

Zeng, X., Y. Zhong, W. Lin, and Q. Zou. 2020. "Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods." *Brief Bioinform* 21 (4):1425-1436.

Zhang, Yaojie, Feng Ma, Benshan Shi, and Dengshi Huang. 2018. "Forecasting the prices of crude oil: An iterated combination approach." *Energy Economics* 70:472-483.

Zhang, Yaojie, Feng Ma, and Yudong Wang. 2019. "Forecasting crude oil prices with a large set of predictors: Can LASSO select powerful predictors?" *Journal of Empirical Finance* 54:97-117.

Zhang, Yaojie, Feng Ma, and Bo Zhu. 2019. "Intraday momentum and stock return predictability: Evidence from China." *Economic Modelling* 76:319-329.

Zhao, Yang, Jianping Li, and Lean Yu. 2017. "A deep learning ensemble approach for crude oil price forecasting." *Energy Economics* 66:9-16.

Zhou, Zhi-Hua, and Ji Feng. 2017. "Deep Forest." *International Joint Conferences on Artificial Intelligence*:3553-3559.

Zhou, Zhi-Hua, and Ji Feng. 2019. "Deep forest." *National Science Review* 6 (1):74-86.

Zhu, Xiaoneng, and Jie Zhu. 2013. "Predicting stock returns: A regime-switching combination approach and economic links." *Journal of Banking & Finance* 37 (11):4120-4133.

Table 1: Hyperparameter settings for the selected models

Models	Tuning parameters
Random Forest	The number of trees in the forest: $\{80, 90, 100, 110, 120\}$
	The minimum number of samples required to split an internal node: $\{1, 2, 3\}$
	The number of features when looking for the best split: {'sqrt','log2','auto'}
XGBoost	L1 regularization term on weights: $\{0, 0.5, 1\}$
	L2 regularization term on weights: $\{0, 0.5, 1\}$
	Step size shrinkage used in update to prevents overfitting: $\{0.01, 0.1\}$
	Maximum depth of a tree: $\{4, 5, 6, 7, 8\}$
	Minimum sum of instance weight(hessian) needed in a child: $\{0, 1, 10\}$
GPR	Value added to the diagonal of the kernel matrix during fitting: $[10^{-10}, 0.1]$
	The number of restarts of the optimizer: $\{0, 1, 2, 3, 4\}$
SVM	The kernel type used in the algorithm: {'linear','poly','rbf','sigmoid'}
	Degree of the polynomial kernel function: $\{2, 3, 4\}$
	Regularization parameter: $\{0.1, 0.5, 1\}$
Ridge	Constant that multiplies the L2 term and controls regularization strength: [0.01,1]
	The maximum number of iterations: $\{10^3, 10^4, 10^5, 10^6\}$
LASSO	Constant that multiplies the L1 term and controls regularization strength: [0.01,1]
	The maximum number of iterations: $\{10^3, 10^4, 10^5, 10^6\}$
KNN	Number of neighbors: $\{3, 4, 5, 6, 7\}$
	Weight function used in prediction: {'distance','uniform'}
	Leaf size: $\{20, 30, 40\}$
	Power parameter for the Minkowski metric: $\{1, 2, 3\}$

Note: We use the default settings for other parameters in each machine learning model. For further details, refer to the module documentation (https://scikit-learn.org/).

Table 2: Out-of-sample forecasting performance

Forecasting model	R^2_{OS} (%)	Success ratio (%)	CER gain (%)
DFEA	40.04***	64.76***	20.33
	(8.99)	(4.11)	
Random Forest	38.55***	63***	15.69
	(8.61)	(3.75)	
XGBoost	28.41***	61.23***	8.86
	(8.46)	(2.92)	
GPR	27.38***	47.58	-5.02
	(6.47)	(-0.92)	
SVM	23.36***	55.95**	2.67
	(7.32)	(2.04)	
Ridge	36.22***	63.88***	22.25
	(8.40)	(4.29)	
LASSO	33.93***	63.88***	18.6
	(7.01)	(3.99)	

KNN	21.79***	51.1	-16.15
	(6.11)	(0.08)	
Mean	38.58***	63.88***	15.75
	(8.22)	(3.99)	

Notes: This table reported the out-of-sample R-squares, success ratios, and CER gains of the selected models using the spot returns of WTI crude oil. The statistics in parentheses corresponding to out-of-sample R-squares and success ratios are their respective test statistics based on Clark and West (2007) and Pesaran and Timmermann (1992). The asterisks ***, **, and * denote the significance levels at the 1%, 5% and 10%, respectively. The mean combination method is computed as the mean value the N individual forecasts. The sample period ranged from January 1986 to December 2019, while the out-of-sample period covered from January 2001 to December 2019.

Table 3: Forecast encompassing test results

Forecasting models	DFEA	
	δ	1-δ
Random Forest	0.03	0.33
XGBoost	0	0.79
GPR	0	0.41
SVM	0	0.74
Ridge	0	0.05
LASSO	0.01	0.29
KNR	0	0.6
Mean	0.05	0.35

Notes: This table reported the p-values for the Harvey, Leybourne, and Newbold (1998) statistics. The p-values in the δ column are used to test the null hypothesis that a competing model in the very left column encompasses the DFEA against the alternative hypothesis that a competing model does not encompass the DFEA. The p-values in the $1-\delta$ column are used to test the null hypothesis that the DFEA encompasses a competing model against the alternative hypothesis that the DFEA does not encompass a competing model.

Table 4: DFEA forecasting results using technical and macroeconomic predictors separately

Evaluation criterion	All predictors	Technical indicators	Macroeconomic variables
R^2_{OS} (%)	40.04***	36.19***	25.32***
	(8.99)	(6.79)	(7.43)
success ratio (%)	64.76***	63.88***	49.34
	(4.11)	(3.93)	(-0.64)

Notes: This table reported the out-of-sample forecasting performance using DFEA with technical and macroeconomic indicators separately. The statistics in parentheses for out-of-sample R-squares and success ratios are their respective test statistics based on Clark and West (2007) and Pesaran and Timmermann (1992). The asterisks ***, **, and * denote the significance levels at the 1%, 5%, and 10%, respectively. The sample period ranged from January 1986 to December 2019, while the out-of-sample period ranged from January 2001 to December 2019.

Table 5: DFEA forecasting results with three different groups of technical indicators

Evaluation criterion	Technical indicators	MA	MOM	OBV
R^2_{OS} (%)	36.19***	20.49***	34.59***	30.72***
	(6.79)	(5.96)	(6.61)	(6.93)
success ratio (%)	63.88***	51.1	67.84***	53.74
	(3.93)	(-0.22)	(5.14)	(0.61)

Notes: This table reported the out-of-sample forecasting performance using DFEA with moving average (MA), momentum (MOM), and on-balance volume (OBV) indicators. The statistics in parentheses for out-of-sample R-squares and success ratios are the test statistics based on Clark and West (2007) and Pesaran and Timmermann (1992). The asterisks ***, **, and * denote the significance levels at the 1%, 5%, and 10%, respectively. The sample period ranged from January 1986 to December 2019, while the out-of-sample period ranged from January 2001 to December 2019.

Table 6: Out-of-sample result of the comparison with various neural networks

Forecasting model	R^2_{OS} (%)	Success ratio (%)	CER gain (%)
DFEA	40.04***	64.76***	20.33
	(8.99)	(4.11)	
NN1	33.3***	63***	19.79
	(7.78)	(3.94)	
NN2	38.75***	67.84***	22.62
	(7.35)	(5.26)	
NN3	34.2***	62.56***	14.57
	(6.61)	(3.47)	
NN4	31.38***	57.71*	-17.9
	(6.72)	(1.54)	
NN5	29.97***	55.07	-33
	(6.63)	(-0.03)	

Notes: This table reported the out-of-sample R-squares, success ratios, and CER gains of DFEA and several neural networks. The statistics in parentheses for out-of-sample R-squares and success ratios are the test statistics based on Clark and West (2007) and Pesaran and Timmermann (1992). The asterisks ***, **, and * denote the significance levels at the 1%, 5%, and 10%, respectively. The sample period ranged from January 1986 to December 2019, while the out-of-sample period ranged from January 2001 to December 2019.

Table 7: Out-of-sample results based on the historical average benchmark

Forecasting model	R^2_{OS} (%)	Success ratio (%)	CER gain (%)
DFEA	14.6***	64.76***	57.34
	(3.37)	(4.11)	
Random Forest	12.48***	63***	52.69
	(3.11)	(3.75)	
XGBoost	-1.97***	61.23***	45.87
	(2.71)	(2.92)	
GPR	-3.44	47.58	31.99
	(-0.6)	(-0.92)	
SVM	-9.15**	55.95**	39.68

	(2.19)	(2.04)	
Ridge	9.16***	63.88***	59.25
	(4.12)	(4.29)	
LASSO	5.89***	63.88***	55.61
	(4.62)	(3.99)	
KNN	-11.4	51.1	20.86
	(0.73)	(0.08)	
Mean	12.03***	64.32***	55
	(3.57)	(4.18)	

Notes: This table reported the out-of-sample R-squares, success ratios, and CER gains based on the historical average benchmark. The statistics in parentheses for out-of-sample R-squares and success ratios are the test statistics based on Clark and West (2007) and Pesaran and Timmermann (1992). The asterisks ***, **, and * denote the significance levels at the 1%, 5%, and 10%, respectively. The mean combination method is computed as the mean value the N individual forecasts. The sample period ranged from January 1986 to December 2019, while the out-of-sample period covered from January 2001 to December 2019.

Table 8: Out-of-sample results for different window sizes.

Forecasting models	Out-of-sample period: 1996:01-2019:12			Out-of-sample period: 2006:01-2019:12		
	R^2_{OS} (%)	Success ratio (%)	CER gain (%)	R^2_{OS} (%)	Success ratio (%)	CER gain (%)
DFEA	41.03***	63.41***	15.3	39.85***	64.67***	20.9
	(10.21)	(4.15)		(7.63)	(3.58)	
Random Forest	39.27***	61.67***	11.98	37.82***	62.87***	11.26
	(9.58)	(3.72)		(7.19)	(3.23)	
XGBoost	32.05***	61.32***	4.61	24.75***	61.68***	18.52
	(9.39)	(3.25)		(7.05)	(2.73)	
GPR	32.5***	47.39	-11.81	26.54***	46.71	1.77
	(7.96)	(-1.05)		(5.39)	(-0.76)	
SVM	26.98***	52.96*	-2.04	18.74***	55.69**	9.24
	(8.84)	(1.3)		(5.91)	(1.83)	
Ridge	36.63***	63.07***	22.75	35.84***	60.48***	21.43
	(9.74)	(4.39)		(7.07)	(2.94)	
LASSO	37.97***	64.11***	13.12	31.59***	65.87***	20.73
	(8.5)	(4.61)		(5.68)	(3.89)	
KNN	26.58***	51.57	-18.09	20.2***	50.9	-10.91
	(7.45)	(0.24)		(5)	(0.13)	
Mean	40.33***	62.37***	15.07	37.67***	62.28***	16.69
	(9.63)	(3.91)		(6.8)	(3.1)	

Notes: This table reports the out-of-sample R-squares, success ratios, and CER gains of various models using different window sizes. The statistics in parentheses for the out-of-sample R-squares and success ratios are the test statistics based on Clark and West (2007) and Pesaran and Timmermann (1992). The asterisks ***, **, and * denote the significance levels at the 1%, 5%, and 10%, respectively. The mean combination method is computed as the mean value the N individual forecasts. The sample period ranged from January 1986 to December 2019, while the out-of-sample periods ranged from January 1996 to December 2019 and from January 2006 to December 2019.

Table 9: Out-of-sample results for the real oil prices of Brent and RAC

Forecasting models	Real Brent			Real RAC		
	R^2_{OS} (%)	Success ratio (%)	CER gain (%)	R^2_{OS} (%)	Success ratio (%)	CER gain (%)
DFEA	41.1***	62.74***	12.55	27.26***	65.64***	15.3
	(7.16)	(3.26)		(6.93)	(4.37)	
Random Forest	38.15***	56.6*	4.88	26.72***	64.76***	12.55
	(6.99)	(1.46)		(6.78)	(4.22)	
XGBoost	26.86***	55.19	-11.32	19.08***	64.76***	8.52
	(6.69)	(0.96)		(6.51)	(3.85)	
GPR	30.84***	43.87	-20.42	9.93***	53.74	-16.4
	(5.62)	(-1.84)		(4.85)	(0.89)	
SVM	24.26***	52.36	-28.65	5.95***	50.22	-11.32
	(5.38)	(0.4)		(5.2)	(0.58)	
Ridge	36.69***	63.68***	14.18	21.32***	65.64***	16.68
	(6.7)	(3.96)		(5.97)	(4.88)	
LASSO	38.26***	61.79***	6.65	21.39***	63.88***	10.12
	(6.09)	(3.26)		(5.38)	(3.99)	
KNN	23.81***	48.11	-42.3	1.67***	52.86	-21.32
	(5.49)	(-1.22)		(4.52)	(0.46)	
Mean	39.4***	60.38***	10.6	27.57***	63.44***	13.77
	(6.53)	(2.77)		(5.93)	(3.75)	

Notes: This table summarized the out-of-sample R-squares, success ratios, and CER gains of various models using the real oil prices of Brent and RAC. The statistics in parentheses for the out-of-sample R-squares and success ratios are the test statistics based on Clark and West (2007) and Pesaran and Timmermann (1992). The asterisks ***, **, and * denote the significance levels at the 1%, 5%, and 10%, respectively. The mean combination method is computed as the mean value the N individual forecasts. The sample period ranged from January 1986 to December 2019, while the out-of-sample period ranged from January 2001 to December 2019.

Table 10: The out-of-sample results for the nominal prices of WTI and Brent

Forecasting models	Nominal WTI			Nominal Brent		
	R^2_{OS} (%)	Success ratio (%)	CER gain (%)	R^2_{OS} (%)	Success ratio (%)	CER gain (%)
DFEA	38.64***	66.96***	20.55	41.42***	63.68***	13.76
	(8.92)	(4.57)		(7.08)	(3.47)	
Random Forest	37.22***	61.67***	12.39	37.35***	56.13*	4.06
	(8.59)	(2.94)		(6.86)	(1.39)	

XGBoost	27.58***	63***	10.93	29.41***	57.55**	-6.79
	(8.22)	(3.23)		(7.03)	(1.65)	
GPR	26.15***	49.34	-10.75	29.72***	44.81	-21.15
	(6.36)	(-0.71)		(5.55)	(-1.65)	
SVM	22.02***	52.86	-4.2	19.22***	50.47	-34.95
	(7.15)	(1.03)		(5.36)	(-0.26)	
Ridge	35.7***	63***	21.73	36.13***	64.15***	15.38
	(8.37)	(4.08)		(6.66)	(4.1)	
LASSO	33.48***	63***	14.1	37.78***	62.74***	7.45
	(6.9)	(3.55)		(6.01)	(3.47)	
KNN	20.43***	53.3	-18.66	23.08***	51.42	-36.75
	(5.95)	(0.44)		(5.4)	(-0.36)	
Mean	37.88***	64.32***	18.37	38.81***	62.26***	13.83
	(8.13)	(3.93)		(6.5)	(3.26)	

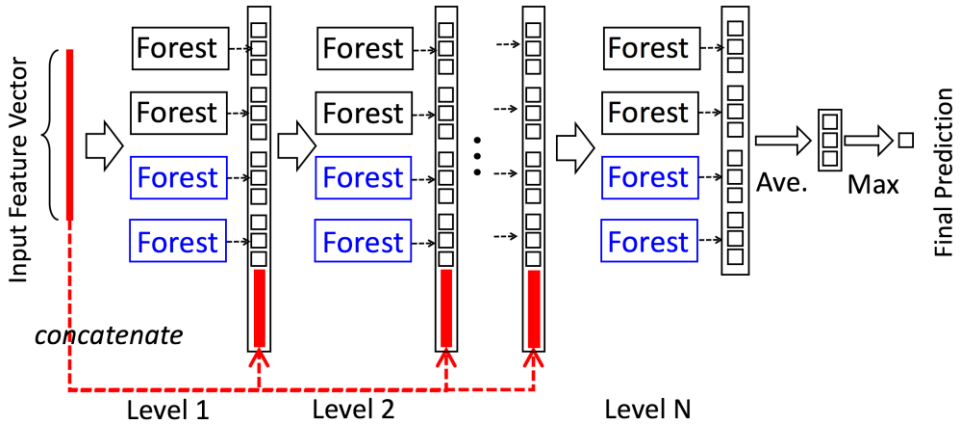
Notes: This table reports the out-of-sample R-squares, success ratios, and CER gains of various models using the nominal oil prices of WTI, Brent, and RAC. The statistics in parentheses for the out-of-sample R-squares and success ratios are their respective test statistics based on Clark and West (2007) and Pesaran and Timmermann (1992). The asterisks ***, **, and * denote the significance levels at the 1%, 5% and 10%, respectively. The mean combination method is computed as the mean value the N individual forecasts. The sample period ranged from January 1986 to December 2019, while the out-of-sample period covered from January 2001 to December 2019.

Table 11: Market timing results

Market timing strategy	NYMEX WTI oil futures		ICE Brent oil futures	
	Average return (%)	Sharpe ratio	Average return (%)	Sharpe ratio
DFEA	77.1	1.02	71.1	0.93
Random Forest	67.03	0.79	62.16	0.74
XGBoost	59.83	0.67	54.24	0.61
GPR	20.28	0.19	18.99	0.19
SVM	33.6	0.33	28.65	0.29
Ridge	71.69	0.89	63.35	0.77
LASSO	73.5	0.93	66.16	0.82
KNR	26.26	0.25	21.72	0.21
Mean	75.8	0.99	70.84	0.93
Always Long	8.73	0.08	9.68	0.09

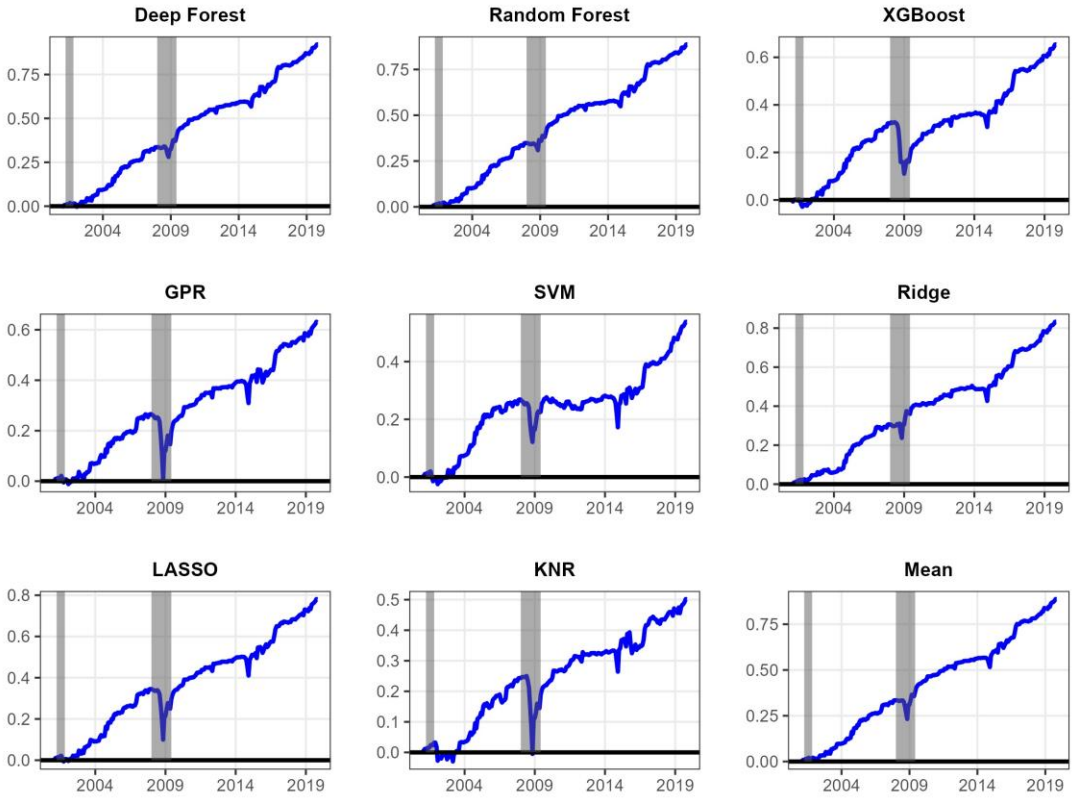
Notes: The table reported the market timing results for both WTI and Brent oil futures in average return and Sharpe ratios. The *Always Long* benchmark strategy involves taking a long position for oil futures at the beginning of the subsequent month and closing it at the end of that month. The average return is annualized in percentage. The sample period ranged from 1986:01 to 2019:12, while the out-of-sample period ranged from 2001:01 to 2019:12.

Figure 1. Schema of the cascade forest structure (Zhou and Feng 2017)



Note: This figure depicts the cascade forest structure of the DFEA. The process starts from the Input feature Vector on the very left and ends with "Final Prediction" on the very right. Each level (vertical layer) consists of an ensemble of tree forests. 'Ave.' takes the average value of prediction results. 'Max' obtains the largest value of the aggregated vector.

Figure 2: Cumulative squared prediction error of the benchmark model minus that of the models of interest



Notes: The rugged line depicted the differences between the benchmark model and the models of interest. Vertical shade bars represent NBER-dated recessions. The sample period ranged from 2001:01 to 2019:12.