Ensemble methods for Stock Market Prediction

Jorge Miguel Bravo

NOVA IMS, Universidade Nova de Lisboa, Portugal &
Université Paris-Dauphine PSL &
MagIC & CEFAGE-UE & BRU-ISCTE-IUL

Torino, 22 September 2023

Ensemble methods for Stock Market Prediction

Jorge Miguel Bravo[1]

NOVA IMS

MIDAS 2023, Torino 2023

[1]Universidade Nova de Lisboa NOVA IMS & Université Paris-Dauphine PSL (jbravo@novaims.unl.pt)

Background

- Forecasting stock market prices and returns is particularly difficult, given the nonlinearity, non-stationarity, volatility, time-evolving complexity of the time series and the generally accepted, semi-strong form of market efficiency (Fama, 1965; Malkiel and Fama, 1970)
- Financial markets are impacted by many highly interconnected systematic and specific factors including economic variables, industry/sector-specific factors and company-specific variables
- However, some researchers (and many in the industry) believe financial markets are inefficient and financial variables can be, to some extent, predictable
- This is because of the influence of psychological factors, information asymmetries between market participants, and market inefficiencies generating arbitrage opportunities (Cervelló-Royo et al. 2015)
- The adoption of plain vanilla and complex structured stock market investment and risk management strategies relies on the estimation and forecasting of vast amounts of data

Previous research

- Univariate or multivariate techniques considering the lagged values of the time series, fundamental analysis factors, technical indicators, or intermarket indicators as feature variables
- Forecasting techniques
  - ARIMA models
  - Stochastic volatility (e.g., GARCH) models and extensions (Franses & Ghijsels, 1999)
  - Linear regression, linear discriminant analysis, quadratic discriminant analysis
  - SVM, ANNs, Random Forest, fuzzy systems, genetic algorithms (Tsai et al., 2011; Toochaei and Moeini, 2023)
  - deep learning methods,...
- Dynamic model combinations (Bravo et al., 2021, 2023); Challenges:
  - Identifying and selecting the model set
  - Model weighting and aggregation
  - Model diversity
  - Meta-learning strategies: Stacking, (Wolpert, 1992), Arbitrating (Ortega et al., 2001),...

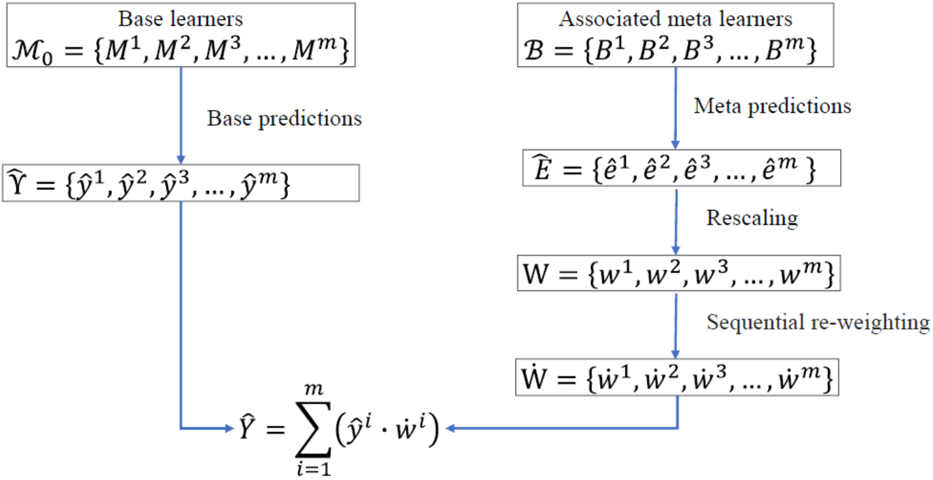
This paper...

- Investigates the predictive accuracy of alternative model combination approaches in financial time series forecasting
- The set of methods includes a meta-learning strategy called Arbitrated Dynamic Ensemble (ADE, Cerqueira et al. 2019)
  - Based on Arbitrating, dynamically combines heterogeneous learners by creating an embedded meta-learner for each base algorithm that specializes them across the time series
  - The model space (base learning algorithms) includes both statistical learning and machine learning methods (e.g., SVR, Gaussian processes, Random Forest, ...)
  - Different parameter specifications are considered for each of the individual forecasters, adding up to 52 different models
- The forecasting performance is compared with
  - Univariate time series models (ARIMA, the Exponential Smoothing State Space Model (ETS), Seasonal Naïve (SNAÏVE)) and
  - Meta-learning strategies (Stacking, Arbitrating, Average of trimmed base forecasters,...)

Arbitrated Dynamic Ensemble

- Let $Y = \{y_1, ..., y_t\}$ denote a numerical time series (e.g., stock prices) with components $y_t \in \mathbb{R}$ observed at times $t = 1, \ldots, t$
- The time series forecasting problem is framed as a **regression task**, with the past K observations of the time series (**embedding vector**) and corresponding summary statistics as attributes in the learning of the experts
- Each observation comprises
  - a feature vector $x_i \in X \in \mathbb{R}^{(K+N)}$ including the past K values and N summary statistics on the embedding vector
  - a target vector $y_i \in Y \in \mathbb{R}$ representing the variable we want to predict
- The goal is to estimate the approximation $\hat{F}(x)$ of the function $F(x)$ mapping the unknown functional dependence $x \xrightarrow{F} y$, that minimizes the expected value of some loss function $\mathcal{L}(y, F(x))$ over the distribution of all $Y-$values, where F denotes the regression function

Arbitrated Dynamic Ensemble: Workflow



Arbitrated Dynamic Ensemble

- The ADE methodology for time series forecasting is implemented in 3 steps:
  - Step 1: Train the $m \in \mathcal{M}_0$ heterogeneous base learners included in the initial model space $\mathcal{M}_0$ to forecast future values of $\hat{Y} = \left\{\hat{y}^1, ..., \hat{y}^m\right\}$
  - Step 2: Train the associated meta-learners $B = \left\{B^1, ..., B^m\right\}$ using the same feature vector to model the error of $\mathcal{M}_0$, $\hat{E} = \left\{\hat{e}^1, ..., \hat{e}^m\right\}$
    - The Random Forest model was used as a meta-learner
    - the error is used to form a committee of best forecasters $\mathcal{M}_0^{\Omega} \in \mathcal{M}_0$ according to their relative forecasting accuracy, trimming and suspending a percentage of recently poor forecasters
    - ADE selects the $\Omega = 50\%$ base forecasters $\mathcal{M}_0^{\Omega}$ with the lowest mean absolute error (MAE) in the last $\lambda = 50$ observations
    - The predictions of the meta-level models (B^{Ω}) are used to weigh $W = \left\{w^1, ..., w^m\right\}$ the selected forecasters
    - To boost model diversity, a sequential re-weighting procedure $W \longrightarrow \dot{W} = \left\{\dot{w}^1, ..., \dot{w}^m\right\}$ is adopted considering the Pearson's correlation among the output of the forecasters in a window of recent observations
  - Step 3: Compute the ensemble prediction using the weighted average of the individual forecasters $\hat{y}^m_{t+1}$ relative to their re-weighted importance $\dot{w}^m_{t+1}$

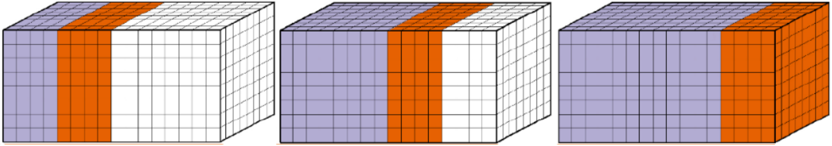
$$\hat{y}_{t+1} = \sum_{j \in B^{\Omega}} \dot{w}^j_{t+1} \cdot \hat{y}^j_{t+1}$$

ADE: Summary of base forecasters

ID	Algorithm	Parameter	Value
GP	Gaussian processes (Karatzoglouet al., 2004)	Kernel	{Linear, RBF, Polynomial, Laplace}
SVR	Support Vector Regression (Karatzoglouet al., 2004)	Kernel	{Linear, RBF Polynomial, Laplace} Cost ∈ {1, 5, 10} ε ∈ {0.001, 0.01}
PPR	Projection pursuit regression (R Core Team, 2022)	No. terms Method	{2, 5, 10} {Super smoother, spline}
MLP	Multi-layer perceptron (Venables andRipley, 2002)	Hidden units Decay	{5, 10, 15, 30} {0.01, 0.05}
MARS	Multivariate Adaptive Regression Splines (Milborrow, 2012)	Degree No. Terms Forward thresh.	{1, 3, 5} {7, 15, 30} {0.001}
GLM	Generalised linear regression models (Friedman et al., 2010)	Penalty mixing Distribution	{0, 0.25, 0.5, 0.75, 1} {Gaussian}
GBR	Generalized boosted regression (Ridgeway, 2022)	Depth Distribution Shrinkage No. Trees Learning rate	{5, 10, 15} {Gaussian, Laplace} {0.1, 0.01} {500, 1000} {0.1}
RF	Random Forest (Wright 2023)	No. trees	{500, 1000}
PCR	Principal components regression (Mevik et al., 2023)	Default	
PLS	Partial least squares regression (Mevik et al., 2023)	Method	{Kernel PLS, Sijmen de Jong's SIMPLS, Principal Component

Arbitrated Dynamic Ensemble

- The ADE methodology uses the training set to generate out-of-bag predictions which are subsequently considered to compute an unbiased estimate of the loss of each base-learner
- The ADE methodology uses a blocked prequential procedure with equally sized (10-folds) and time-sequential blocks of contiguous observations with a growing lookback window approach to producing out-of-bag samples



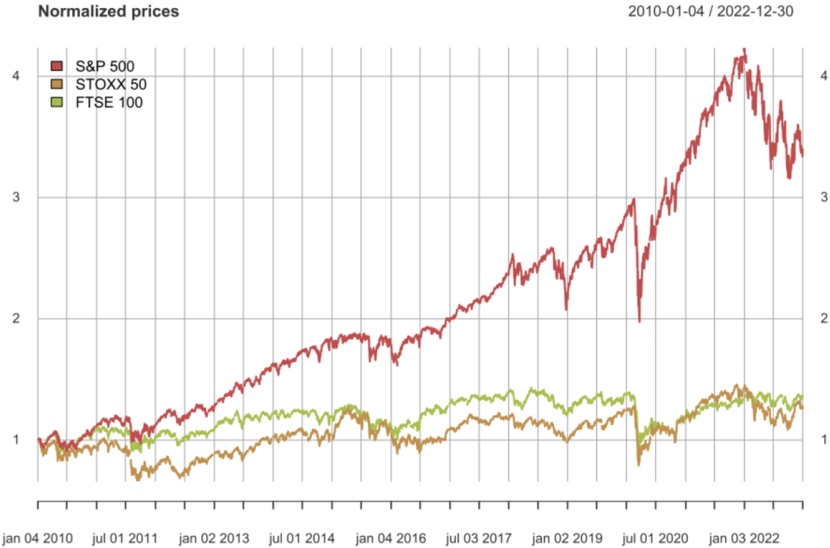
- The estimation method considers a repeated holdout procedure (learning plus testing cycle) in 25 randomly selected testing periods using different but overlapping observations
  - Each repetition uses three years of data for training and a lookforward window of 21 trading days for testing

Empirical strategy

- Datasets: S&P 500, Euro Stoxx 50, and FTSE 100 stock index daily data from January 1, 2010, to December 31, 2022 (3272 daily observations of the adjusted closing price)
- Source: Nasdaq Data Link
- To account for trend and check for stationarity, we use the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) unit root test
- The optimal embedding dimension (K) is estimated using the method of False Nearest Neighbours (Kennel et al., 1992) setting the tolerance of false nearest neighbours to 1%
- The feature set that includes the embedding vector and several characteristics summarising the overall structure of time series:
  - Local trend
  - Skewness
  - Mean
  - Standard deviation
  - Serial correlation
  - Long-range dependence using a Hurst exponent estimation with wavelet transform
  - Chaos, using the maximum Lyapunov exponent
- The approach can easily be extended to include other (external) attributes

Empirical strategy

Stock market index normalized prices 2010-2022



Empirical strategy

- The performance of the ADE methodology is compared with:
  - **Stacking** for times series (Wolpert, 1992)
  - **Arbitrating** (Ortega et al., 2001)
  - **Simple**: model ensemble with individual forecasters averaged using an arithmetic mean (Timmermann, 2008)
  - **SimpleTrim**: Simple average of base forecasters with model trimming, with $\Omega = 50\%$ of the best past performing models are selected to take part in the ensemble committee
  - **LossTrain**: static weighted average of forecasters, with weights defined according to the performance of experts in the training set
  - **BestTR**: forecasting approach that selects the best performing model in the training data
  - **EWA**: forecast combination method based on an exponentially weighted average of experts
  - **FixedShare**: the fixed share approach adapted for identifying the best forecaster across a time series (Cesa-Bianchi and Lugosi, 2006)
  - **MLpol**: the polinomially weighted average forecast combination (Cesa-Bianchi and Lugosi, 2006)
  - **OGD**: An approach based on online gradient descent (Zinkevich, 2003)
  - **ETS**: The exponential smoothing state space model (Hyndman and Athanasopoulos, 2021)

Empirical strategy

- In addition, we investigated the performance of four variants of ADE:
  - **ADE-SB**: a variant of ADE in which at each time point the best performing model (the one with the lowest predicted loss) is selected to make a prediction
  - **ADE-ALLM**: A variant of ADE without the formation of a committee, i.e., with $\Omega = 100\%$
  - **ADE-v0**: variant of ADE with *linear re-weighting* of the output of the arbiters instead of the *softmax-type function* and no sequential re-weighting
  - **ADE-noSR**: A variant of ADE in which there is no sequential reweight of the experts
- The forecasting accuracy of base-learners and model combination approaches is evaluated using the RMSE
- To draw inferences about the differences in model forecasting performance, we use two post-hoc non-parametric tests: i) **Friedman test** (1940); ii) **The Nemenyi test**

Empirical findings

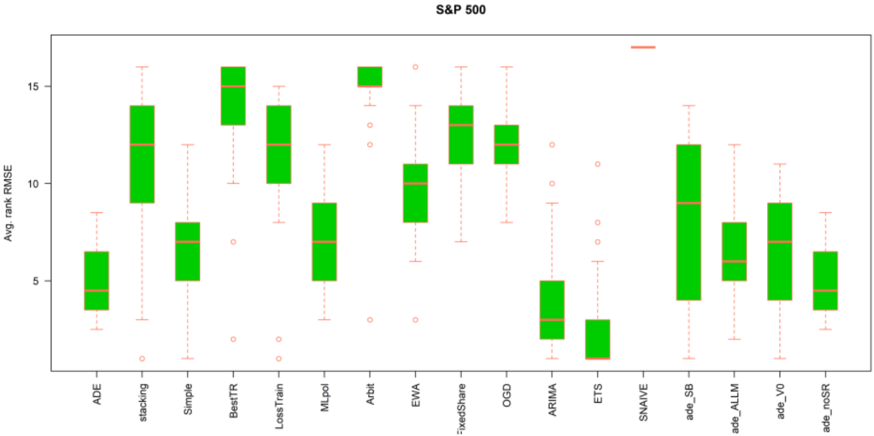
Table 2. Average Root Mean Squared Error of state-of-the-art forecasters

Model	S&P 500	STOXX 50	FTSE 100
ADE	38.375	25.490	60.667
Stacking	38.722	26.349	62.009
SimpleTrim	38.441	25.764	60.880
BestTR	39.570	26.053	62.640
LossTrain	38.936	25.823	61.667
MLpol	38.453	26.020	60.834
Arbitrating	40.384	26.518	62.723
EWA	38.778	29.852	61.726
FixedShare	39.192	28.614	61.936
OGD	38.881	26.030	61.557
ARIMA	38.184	25.737	60.873
ETS	38.121	25.442	60.393
SNAIVE	53.443	34.203	84.319
ade_SB	38.514	25.715	61.491
ade_ALLM	38.429	25.615	60.812
ade_V0	38.416	25.570	61.177
ade_noSR	38.375	25.490	60.667

Note: Average RMSE values over 25 randomly selected testing periods.

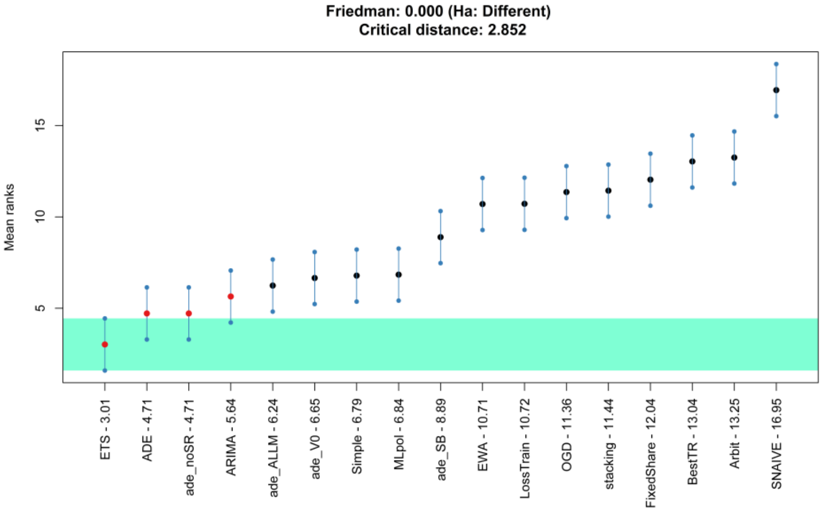
Empirical findings

Distribution of RMSE of state-of-the-art forecasting approaches



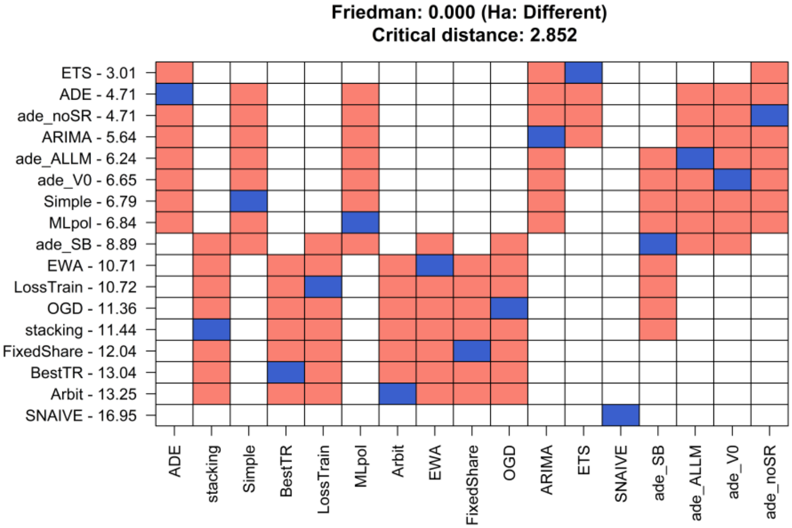
Empirical findings

Distribution of rank of forecasting approaches across datasets and testing periods



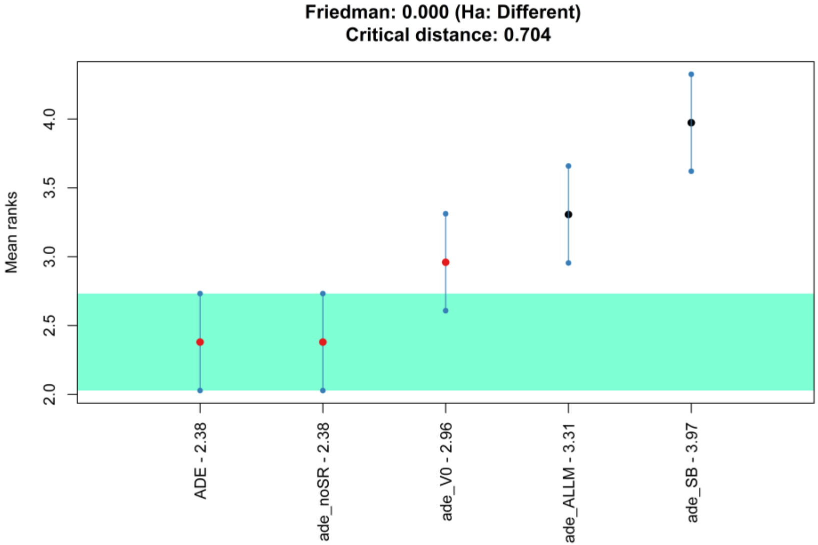
Empirical findings

Statistical significance of forecasting accuracy differences



Empirical findings

Distribution of rank of ADE and its variants



Empirical findings: summary

- The Exponential Smoothing State Space (ETS) model exhibits the lowest RMSE error across the three stock market indices...
  - followed closely by ARIMA model in the S&P 500 series and by the ADE meta-learning approach in the STOXX 50 and FTSE 100 series
- The ETS model exhibits the best average rank, followed closely by ADE
- The ADE method presents a better average rank compared to competing widely used model combination approaches, including the original Arbitrating approach, Stacking, the polynomially weighted average forecast combination, Simple averaging of individual learners, Fixed Share, or OGD
- Except for the ETS method (although within the test confidence intervals), the ADE meta-learning approach exhibits superior short-term forecasting accuracy when compared to competing state-of-the-art methods
- Compared to Arbitrating, ADE represents a considerable improvement in short-term stock price prediction
  - The average rank of ADE is 4.71 against 13.25 for Arbitrating
  - This suggests that using the training set to generate out-of-bag predictions, using them to compute an unbiased estimate of the loss of each base learner, and adopting a blocked prequential procedure contributes to improving the forecasting performance

Empirical findings: summary

- Among the model combination approaches, the Simple averaging approach preceded by a selection of a percentage of recent best past-performing models exhibits an interesting predictive performance
- The alternative dynamic ensemble method MLpol, a polynomially weighted average forecaster based on regret minimization exhibits also an interesting per-formance when compared with Stacking or Arbitrating
- The results of the Friedman test for the null hypothesis that the forecasters have similar performance a 95% confidence level show that the ADE approach, its variant ADE_noSR and the standard ARIMA method (by a small margin) are the only three forecasting approaches for which we cannot reject the null hypothesis that they exhibit similar forecasting performance when compared to the top forecaster ETS
- The results on the performance of ADE against its variants show that
  - the ADE methodology benefits from combining the base learners as opposed to selecting the best forecasting model at each iteration (ADE_SB)
  - The ADE approach benefit from model selection, i.e., improves ADE-ALLM
  - Sequentially re-weighting the experts according to the recent correlation does not improve or deteriotate the ADE accuracy

Ensemble methods for Stock Market Prediction

THANK YOU

Jorge Miguel Bravo
(jbravo@novaims.unl.pt)

Scopus ID: 57193745638; ORCID: 0000-0002-7389-5103

Universidade Nova de Lisboa, NOVA IMS & Université Paris-Dauphine PSL & MagIC & BRU-ISCTE-IUL, CEFAGE-UE ;